Time and Frequency Measurement

Edited by
Christine Hackman and
Donald B. Sullivan

AAPT American Association of Physics Teachers

Time and Frequency Measurement

American Association of Physics Teachers
One Physics Ellipse
College Park, MD 20740-3845

ISBN 0-917853-67-9

On the Cover
The repeated image appearing on the background of the cover shows mercury ions in a linear electromagnetic trap. This string of trapped 199 Hg ions is being studied as a possible atomic frequency standard. The ions are illuminated with laser radiation at 194 nm and fluoresce to produce this image in an ultraviolet imaging system. Studies involving both microwave and ultraviolet transitions indicate strong potential for higher-accuracy frequency standards. The gaps in the string result from the presence of impurity ions, most likely other isotopes of mercury that do not fluoresce at the same frequency as does 199 Hg.

Photo appears courtesy of the National Institute of Standards and Technology (NIST).

Cover design by Rebecca Heller Rados.

Contents

Characterization

Scientific Measurements

RESOURCE LETTER

Roger H. Stuewer, *Editor*
School of Physics and Astronomy, 116 Church Street
University of Minnesota, Minneapolis, Minnesota 55455

This is one of a series of Resource Letters on different topics intended to guide college physicists, astronomers, and other scientists to some of the literature and other teaching aids that may help improve course content in specified fields. [The letter E after an item indicates elementary level or material of general interest to persons becoming informed in the field. The letter I, for intermediate level, indicates material of somewhat more specialized nature; and the letter A, indicates rather specialized or advanced material.] No Resource letter is meant to be exhaustive and complete; in time there may be more than one letter on some of the main subjects of interest. Comments on these materials as well as suggestions for future topics will be welcomed. Please send such communications to Professor Roger H. Stuewer, Editor, AAPT Resource Letters, School of Physics and Astronomy, 116 Church Street SE, University of Minnesota, Minneapolis, MN 55455.

Resource Letter: TFM-1: Time and frequency measurement

Christine Hackman and Donald B. Sullivan
Time and Frequency Division, National Institute of Standards and Technology, Boulder, Colorado 80303

(Received 26 August 1994; accepted 5 December 1994)

This Resource Letter is a guide to the literature on time and frequency measurement. Journal articles and books are cited for the following topics: frequency standards; methods of characterizing performance of clocks and oscillators; time scales, clock ensembles, and algorithms; international time scales; frequency and time distribution; and applications. [The letter E after an item indicates elementary level or material of general interest. The letter I, for intermediate level, indicates material of somewhat more specialized nature, and the letter A indicates rather specialized or advanced material. The designations E/I and I/A are used to indicate that the article contains material at both levels, so that at least part of the article is written at the lower of the two levels.]

I. INTRODUCTION

Archeological evidence indicates that since prehistoric times man has been devising progressively better means of keeping track of the passage of time. In the earliest stages this involved observation of the apparent motion of the sun, but finer subdivision of the day later involved devices such as water clocks, hourglasses, and calibrated candles. After long development with many variations, mechanical methods for keeping time, in the form of pendulum clocks, achieved excellent precision (a fraction of a second per day). However, with the invention of the two-pendulum clock in 1921 by William Hamilton Shortt, the practical performance limit of such mechanical clocks was reached. The distinction between frequency standard and clock (between frequency and time) is easily recognized in the pendulum clock. The constant frequency of oscillation of the pendulum constitutes a frequency standard. The mechanism used to count the ticks and display their accumulation as seconds, minutes, hours, days, and years converts this frequency standard into a clock.

The modern era of timekeeping began with the development of the quartz crystal oscillator. In a 1918 patent application, Alexander M. Nicholson disclosed a piezoelectric crystal as the control element in a vacuum tube oscillator. The first clock controlled by a quartz crystal was subsequently developed in 1927 by Joseph W. Horton and Warren A. Marrison. Since the introduction of the quartz oscillator, the performance of frequency standards has advanced by many orders of magnitude, and industry and science have come to rely on the timing made possible by them.

Many modern technological applications require that geographically distributed systems have the same time (synchronization) or run at the same rate (syntonization). Thus, an important consideration in time and frequency measurement has been the precise transfer of timing between separated stations. This has led to an interplay between the development of the two key technologies, (1) frequency standards and clocks and (2) methods of time and frequency transfer. Comparisons between early quartz timepieces were accomplished with adequate precision using signals transmitted by terrestrial radio waves.

Quartz crystal oscillators remained at the performance

forefront for only a short time. In 1949, the atomic-timekeeping era began with the construction of the first atomic clock. This standard, based on a resonance in the ammonia molecule, was constructed at the National Bureau of Standards in a project led by Harold Lyons. The ammonia standard was quickly superseded by the cesium-beam frequency standard that forms the current basis for defining the second.

Atomic standards progressed rapidly in accuracy and stability through the 1950s and 1960s. By the mid-1970s the best atomic standards were realizing the definition of the second with an uncertainty of 10^{-13}, allowing timekeeping uncertainty of 10 ns over the period of one day. But the existing methods of time transfer could not easily support comparison of performance between geographically separated devices. At the time, the most convenient method for comparing standards over long distances involved LORAN-C navigation signals. Separated stations could simultaneously monitor the same highly stable LORAN-C broadcast to achieve comparison of the standards, but time comparison errors as large as 500 ns were observed. Clocks could be compared more precisely using portable atomic clocks, but since this involved flying a fully operational clock from one laboratory to another, it was expensive and impractical to do more often than a few times per year.

The most recent thrust forward was fostered by two developments. More precise time transfer using GPS satellites in a "common-view" mode was developed in the early 1980s. This provided the means for precise comparisons of standards constructed in different laboratories. During this same period, physicists were developing methods for using lasers to control the atomic states and the motions of atoms and ions. These new methods offer promise of dramatic advances in accuracy of atomic standards, and evidence of such advances is just beginning to appear. With these new concepts for improved standards and the technology required to compare the performance of separated standards, physicists have gained new motivation to build better atomic clocks, and work is progressing rapidly in many laboratories.

Significant improvements have been made in cesium-beam standards by replacing magnetic state selection and state detection with optical selection and detection. But still greater advances will come with the practical realization of entirely new concepts that have been demonstrated in the laboratory. Radiatively cooled ions stored in electromagnetic traps provide the ultimate answer to Doppler-shift problems, but the atomic-fountain standard that uses laser-cooled neutral atoms offers substantial promise as well.

The pace of improvement in atomic standards, an increase in accuracy by a factor of nearly 10 every 7 years, is expected to continue. The most advanced concepts promise accuracy improvement of three to four orders of magnitude. Frequency is already the most accurately realized unit of measure, so it is surprising to find such great potential for improvement. Since even modest frequency accuracy is often well beyond the accuracy with which other measurements are made, other quantities are often converted (transduced) to frequency to achieve better precision, resolution, and ease of measurement. Length and voltage are examples of units now based on frequency measurements.

The atomic definition of the second has provided the means to improve frequency accuracy, but in many applications the only requirement is for frequency stability. In this case, what matters is that two or more oscillators stay at the same, although not necessarily accurate, frequency. While high accuracy ensures good relative stability, it is not essential for it. Thus, in responding to application requirements, developers of atomic and quartz devices have often emphasized frequency stability rather than frequency accuracy.

Time plays a major role in physical theory, and better clocks have provided physicists with improved means for testing their theories. Over the last 45 years, progress in atomic clock technology has been driven more by scientific than by industrial requirements, but clever engineers have taken good advantage of the advancing technology. A good example is the exceptional navigational accuracy provided by the Global Positioning System (GPS). This accuracy is critically dependent on atomic-clock technology. Other areas benefitting from this technology include telecommunications and electrical-power distribution.

In this resource letter we address time and frequency standards and measurement methods, the special statistical methods used for handling noise in clocks and oscillators, and the distribution of time and frequency signals. Signal distribution is included because a large fraction of measurements made in this field rely on timing signals transmitted from central sources to the points of measurement. Articles in many of the areas covered by this resource letter are published in well-known journals, but in some instances we are forced to refer to papers presented at conferences, the proceedings of which are not widely available in libraries. In these instances, we provide an address where the proceedings can be obtained. In our selection of articles we recognize that biases will be evident, and we apologize for these biases. There are a great number of papers that could have been cited, but are not because of space limitations.

II. JOURNALS

The majority of papers on time and frequency topics are found in the following journals:

IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control
IEEE Transactions on Instrumentation and Measurement
Journal of Applied Physics
Metrologia
Physical Review
Physical Review Letters
Physics Today
Proceedings of the IEEE
Radio Science
Science
Scientific American

Four special issues of IEEE journals (the first four references below) have been devoted exclusively to time and frequency topics. Many of the invited articles in these issues were written as reviews of specialized subtopics and are therefore very readable. Two other special issues on the broader subject of radio measurement methods and standards are also included because they contain a number of useful and easy-to-read articles on time and frequency.

1. **Time and Frequency**, Special Issue of Proc. IEEE **79**(7) (1991).
2. **Frequency Control**, Special Issue of IEEE Trans. UFFC **UFFC-34**(6) (1987).
3. **Time and Frequency**, Special Issue of Proc. IEEE **60**(5) (1972).
4. **Frequency Stability**, Special Issue of Proc. IEEE **54**(2) (1966).

5. **Radio Measurement Methods and Standards**, Special Issue of Proc. IEEE **74**(1) (1986).
6. **Radio Measurement Methods and Standards**, Special Issue of Proc. IEEE **55**(6) (1967).

III. CONFERENCE PROCEEDINGS

Conference proceedings in this field contain a good mix of specialized articles and an occasional review article. The following series of conferences contain the largest concentration of time-and-frequency papers.

A. IEEE International Frequency Control Symposium

The IEEE International Frequency Control Symposium is the best-known conference in the field. It was previously known as the Annual Symposium on Frequency Control. Through 1994, there have been 48 conferences in this series. The proceedings of the last 12 of these conferences are available from IEEE, 445 Hoes Lane, Piscataway, NJ 08854. The most recent eight and three others cited later are listed here.

7. **1994 IEEE Int. Freq. Control Symp.**, IEEE Catalogue No. 94CH3446-2 (1994).
8. **1993 IEEE Int. Freq. Control Symp.**, IEEE Catalogue No. 93CH3244-1 (1993).
9. **1992 IEEE Int. Freq. Control Symp.**, IEEE Catalogue No. 92CH3083-3 (1992).
10. **45th Annu. Symp. Freq. Control**, IEEE Catalogue No. 91CH2965-2 (1991).
11. **44th Annu. Symp. Freq. Control**, IEEE Catalogue No. 90CH2818-3 (1990).
12. **43rd Annu. Symp. Freq. Control**, IEEE Catalogue No. 89CH2690-6 (1989).
13. **42nd Annu. Symp. Freq. Control**, IEEE Catalogue No. 88CH2588-2 (1988).
14. **41st Annu. Symp. Freq. Control**, IEEE Catalogue No. 87CH2427-3 (1987).
15. **38th Annu. Symp. Freq. Control**, IEEE Catalogue No. 84CH2062-8 (1984).
16. **36th Annu. Symp. Freq. Control**, Document No. AD-A130811, available from National Technical Information Service, 5285 Port Royal Road, Sills Building, Springfield, VA 22161 (1982).
17. **35th Annu. Symp. Freq. Control**, Document No. AD-A110870, available from same source as Ref. 16 (1981).

B. European Frequency and Time Forum (EFTF)

This is the major European conference in the field. Through 1994 there have been eight meetings in this series. Printed proceedings are not readily available in libraries, but can be obtained from the EFTF Secretariat at: FSRM, Rue de l'Orangerie 8, CH-2000, Neuchâtel, Switzerland. Three volumes cited later are:

18. **Proc. 7th European Freq. and Time Forum** (1993).
19. **Proc. 5th European Freq. and Time Forum** (1991).
20. **Proc. 4th European Freq. and Time Forum** (1990).

C. Conference on Precision Electromagnetic Measurements (CPEM)

This biennial international conference covers a wide range of measurement topics, but includes a substantial number of useful articles on time and frequency topics. The conference focus on measurements makes this a particularly important reference. Since 1962 the proceedings of this biennial conference have been published in the *IEEE Transactions on Instrumentation and Measurement*. References to the last seven conferences are given below.

21. **Proceedings of CPEM'92**, IEEE Trans. Instrum. Meas. **42**(2) (1993).
22. **Proceedings of CPEM'90**, IEEE Trans. Instrum. Meas. **40**(2) (1991).
23. **Proceedings of CPEM'88**, IEEE Trans. Instrum. Meas. **38**(2) (1989).
24. **Proceedings of CPEM'86**, IEEE Trans. Instrum. Meas. **IM-36**(2) (1987).
25. **Proceedings of CPEM'84**, IEEE Trans. Instrum. Meas. **IM-34**(2) (1985).
26. **Proceedings of CPEM'82**, IEEE Trans. Instrum. Meas. **IM-32**(1) (1983).
27. **Proceedings of CPEM'80**, IEEE Trans. Instrum. Meas. **IM-29**(4) (1980).

D. Symposium on Frequency Standards and Metrology

Although a few other subjects have been covered, this symposium is devoted primarily to the physics of frequency standards. The four conferences in the series have been separated by 5 to 7 years, reflecting the symposium philosophy that very significant progress in the field should occur before each meeting is held. The proceedings of the first and second of these are not widely available, but the 1988 and 1981 proceedings are available in book form.

28. **Frequency Standards and Metrology: Proceedings of the Fourth Symposium**, edited by A. De Marchi (Springer-Verlag, New York, 1989).
29. **Third Symposium on Frequency Standards and Metrology**, J. Phys. **42**, Colloque C-8, Supplement no. 12 (1981).

E. Annual Precise Time and Time Interval Applications and Planning Meeting

The proceedings of these military/NASA planning meetings have been published by NASA in recent years. Through 1993 there have been 25 meetings in the series. Copies of the proceedings are available from the U.S. Naval Observatory, Time Service, 3450 Massachusetts Ave., N.W., Washington, D.C. 20392-5420. The most recent five and three others cited later are noted here. Three of these do not have publication numbers.

30. **25th Annual PTTI Meeting**, NASA Conf. Publ. 3267 (1993).
31. **24th Annual PTTI Meeting**, NASA Conf. Publ. 3218 (1992).
32. **23rd Annual PTTI Meeting**, NASA Conf. Publ. 3159 (1991).
33. **22nd Annual PTTI Meeting**, NASA Conf. Publ. 3116 (1990).
34. **21st Annual PTTI Meeting**, (1989).
35. **18th Annual PTTI Meeting**, (1986).
36. **15th Annual PTTI Meeting**, (1983).
37. **13th Annual PTTI Meeting**, NASA Conf. Publ. 2220 (1981).

F. Other special conferences

Two special conferences, one held in Finland and the other in India, should also be noted.

38. First Open Symposium on Time and Frequency of URSI Commission A, Radio Science **14**(4) (1979).
39. International Symposium on Time and Frequency, J. Inst. Elec. Telecomm. Eng. **27**(10) (1981).

IV. BOOKS

40. The Quantum Physics of Atomic Frequency Standards, 2 Vols., J. Vanier and C. Audoin (Adam Hilger, Bristol, England, 1989). The most comprehensive resource available on atomic standards. (A)
41. Precision Frequency Control, 2 Vols., edited by E. A. Gerber and A. Ballato (Academic, New York, 1985). Very comprehensive coverage of the entire field. (A)
42. From Sundials to Atomic Clocks, J. Jespersen and J. Fitz-Randolph (Dover, New York, 1982). Comprehensive coverage with an easily readable style. Laced with sketches that effectively convey concepts. (E)
43. Frequency and Time, P. Kartaschoff (Academic, New York, 1978). Comprehensive coverage of the entire field. (A)
44. Quartz Crystals for Electrical Circuits, edited by R. A. Heising (Van Nostrand Co., New York, 1946). An old, yet useful, book covering many fundamental concepts. (I)
45. Molecular Beams, N. F. Ramsey (Clarendon, Oxford, 1956). Important background for atomic-beam frequency standards. (A)
46. Systems with Small Dissipation, V. B. Braginsky, V. P. Mitrofanov, and V. I. Panov (University of Chicago, Chicago, 1985). Treats high-Q resonators that are important in several different frequency standards. (I/A)

V. CURRENT RESEARCH TOPICS

A. General review articles

There is no single review paper that covers the field comprehensively. A full review of the field can be found in several of the books cited above. The articles listed below provide good reviews of limited segments of the field. Two of these papers review the history of development of atomic standards and in the process give good introductions to their principles of operation.

47. "Accurate Measurement of Time," W. M. Itano and N. F. Ramsey, Sci. Am. **269**(1), 56–65 (1993). A popular review article. (E)
48. "Time Generation and Distribution," D. B. Sullivan and J. Levine, Proc. IEEE **79**(7), 906–914 (1991). Reviews current trends in the field. (E)
49. "Standard Time and Frequency Generation," P. Kartaschoff and J. A. Barnes, Proc. IEEE **60**(5), 493–501 (1972). An older but concise review. (I)
50. "History of Atomic and Molecular Standards of Frequency and Time," N. F. Ramsey, IEEE Trans. Instrum. Meas. **IM-21**(2), 90–99 (1972). The early history of atomic standards from one of the key contributors to the development of the concepts. (E/I)
51. "A Historical Review of Atomic Frequency Standards," R. E. Beehler, Proc. IEEE **55**(6), 792–805 (1967). Another early history of atomic standards. (E/I)
52. "Time, Frequency and Physical Measurement," H. Hellwig, K. M. Evenson, and D. J. Wineland, Phys. Today **31**(12), 23–30 (1978). A popular review article. (E/I)
53. "Frequency and Time Standards," R. F. C. Vessot, Chap. 5.4 in **Methods of Experimental Physics**, Vol. 12, **Astrophysics**, Part C: *Radio Observations* (Academic, New York, 1976), pp. 198–227. An older, but useful review. (I/A)
54. "Timekeeping and Its Applications," G. M. R. Winkler, in **Advances in Electronics and Electron Physics**, Vol. 44 (Academic, New York, 1977), pp. 33–97. An older, but useful review. (I/A)
55. "Communications Frequency Standards," S. R. Stein and J. R. Vig, **The Froehlich/Kent Encyclopedia of Telecommunications**, Vol. 3, edited by F. E. Froehlich and A. Kent (Marcel Dekker, New York, 1992), pp. 445–500. A good review with emphasis on quartz-oscillator technology. (E)

B. Frequency standards

The different frequency standards (oscillators) are organized into seven categories. These are arranged roughly in ascending order of complexity and cost.

1. Quartz oscillators

Quartz oscillator development has been characterized by slow, steady improvement for many years and this pace of improvement is likely to continue into the future. Their small size, low power consumption, and excellent short-term performance make quartz oscillators suitable for a large number of applications: more than 2×10^9 units are produced annually. However, quartz oscillators exhibit long-term aging and are sensitive to environmental changes, so they are not suitable for some applications.

With special attention to packaging and environmental control (at some expense) quartz oscillators can provide a frequency stability of 10^{-11} to 10^{-12} for averaging times of one day. The most comprehensive collection of articles on this subject are found in the two-volume set edited by Gerber and Ballato (Ref. 41). Reference 55 includes a more concise review of the topic.

56. "Introduction to Quartz Frequency Standards," J. Vig, in Tutorials from the 23rd Annual PTTI Applications and Planning Meeting, pp. 1–49 (1991). (This specific section of the publication is available from the National Technical Information Service, 5285 Port Royal Rd., Springfield, VA 22161.) A very useful introduction. (E/I)
57. "Quartz Crystal Oscillators from Their Design to Their Performances," J.-J. Gagnepain, in Ref. 20, pp. 121–129. A general review paper. (I)
58. "Environmental Sensitivities of Quartz Oscillators," F. L. Walls and J.-J. Gagnepain, IEEE Trans. UFFC **39**(2), 241–249 (1992). A good review including the basic model of quartz oscillators. Contains a good reference list. (I)
59. "Fundamental Limits on the Frequency Instabilities of Quartz Crystal Oscillators," J. R. Vig and F. L. Walls, in Ref. 7, pp. 506–523. A good review including the basic model of quartz oscillators. Contains a good reference list. (I)

60. "Spectral Purity of Acoustic Resonator Oscillators," T. E. Parker and G. K. Montress, in Ref. 9, pp. 340–348. A survey of state-of-the-art oscillators. (I)
61. "Quartz Crystal Resonators and Oscillators, Recent Developments and Future Trends," R. J. Besson, J. M. Groslambert, and F. L. Walls, Ferroelectrics **43**, 57–65 (1982). A review of development trends. (I)
62. "Filters and Resonators—A Review: I. Crystal Resonators," E. Hafner, IEEE Trans. Son. Ultrason. **SU-21**(4), 220–237 (1974). A review of quartz resonators. (I/A)
63. "The Evolution of the Quartz Crystal Clock," W. A. Marrison, Bell Sys. Tech. J. **27**(3), 510–588 (1948). An excellent early history of quartz oscillator development. (E)

2. Rubidium standards

Rubidium frequency standards are the least costly of the atomic standards. In terms of long-term noise and drift, rubidium standards are generally better than quartz, but not nearly as good as other atomic standards. The traditional rubidium standard relies on optical pumping of atomic states by a discharge lamp. Recent research suggests that substantial improvement in performance can be achieved by pumping the states with a spectrally pure source such as a laser. Rubidium standards are typically passive devices, but they can also be operated in an active (masing) mode.

64. "The Optically Pumped Rubidium Vapor Frequency Standard," M. E. Packard and B. E. Swartz, IRE Trans. Instrum. **I-11**(3&4), 215–223 (1962). A useful description of an early rubidium standard. (E/I)
65. "Rubidium Frequency Standards," J. Vanier and C. Audoin, Chap. 7 in Ref. 40, Vol. 2, pp. 1257–1409. Comprehensive review with extensive reference list. (A)
66. "Fundamental Stability Limits for the Diode-Laser-Pumped Rubidium Atomic Frequency Standard," J. C. Camparo and R. P. Frueholz, J. Appl. Phys. **59**(10), 3313–3317 (1986). Basic model of the rubidium standard including projected performance when pumped by a narrow-line source. (A)
67. "Experimental Study of the Laser Diode Pumped Rubidium Maser," A. Michaud, P. Tremblay, and M. Têtu, IEEE Trans. Instrum. Meas. **40**(2), 170–173 (1991). Operation as a maser. (I)

3. Cesium standards

Since 1967, the second has been defined as "the duration of 9 192 631 770 periods of the radiation corresponding to the transition between two hyperfine levels of the ground state of the cesium-133 atom." Thus it is not surprising to find a large body of literature on cesium frequency standards.

Cesium standards are passive devices, that is, the atoms play a passive role wherein an external oscillator scans through a range of frequencies and some detection scheme then indicates when the oscillator is on the atomic resonance. Cesium standards typically use the method of separated oscillatory fields, a special mode of interaction of the atoms with the external oscillating field, to produce an especially narrow resonance. This reduces sensitivity to dc and oscillating field inhomogeneities. Elementary introductions to the principles of operation of cesium-beam frequency standards can be found in many of the general review articles cited above in Sec. V A.

The best traditional cesium frequency standards (based on atomic-beam methods) now realize the definition of the second with a relative uncertainty of about 1×10^{-14}. The primary source of this uncertainty is associated with the motions of the atoms. Substantial improvements in cesium standards will require slowing of the atoms. The most promising avenue for such improvement involves the cesium-fountain standard. In this device atoms are laser cooled and then lofted vertically. The resonance is detected as the atoms first rise and then fall under the influence of gravity. Such an approach increases observation time by two orders of magnitude compared to traditional atomic-beam devices, and also dramatically reduces the Doppler shift. A number of laboratories are working on this concept.

68. "The Method of Successive Oscillatory Fields," N. F. Ramsey, Phys. Today **33**(7), 25–30 (1980). Good description of the state-interrogation concept used in cesium frequency standards. (E/I)
69. "The Caesium Atomic Beam Frequency Standard," J. Vanier and C. Audoin, Chap. 5 in Ref. 40, Vol. 2, pp. 603–947. Comprehensive review with extensive reference list. (A)
70. "Atomic Beam Frequency Standards," R. C. Mockler in **Advances in Electronics and Electron Physics**, Vol. 15 (Academic, New York, 1961), pp. 1–71. Deals primarily with the theory of the cesium-beam frequency standard. (A)
71. "CS2: The PTB's New Primary Clock," A. Bauch, K. Dorenwendt, B. Fischer, T. Heindorff, E. K. Müller, and R. Schröder, IEEE Trans. Instrum. Meas. **IM-36**(2), 613–616 (1987). Description of an excellent primary standard of traditional design. (I)
72. "The NIST Optically Pumped Cesium Frequency Standard," R. E. Drullinger, D. J. Glaze, J. P. Lowe, and J. H. Shirley, IEEE Trans. Instrum. Meas. **40**(2), 162–164 (1991). An optically pumped version of the cesium-beam frequency standard. (I)
73. "Design of an Optically Pumped Cs Laboratory Frequency Standard," E. de Clercq, A. Clairon, B. Dahmani, A. Gérard, and P. Aynié, in Ref. 28, pp. 120–125. Provides good detail on design. (I)
74. "Ramsey Resonance in a Zacharias Fountain," A. Clairon, C. Salomon, S. Guellati, and W. D. Phillips, Europhys. Lett. **16**(2), 165–170 (1991). Description of a fountain frequency standard with potential for use as a primary frequency standard. (I)
75. "Laser-Cooled Neutral Atom Frequency Standards," S. L. Rolston and W. D. Phillips, Proc. IEEE **79**(7), 943–51 (1991). Includes a review of the fountain concept. (I)
76. "Laser-Cooled Cs Frequency Standard and a Measurement of the Frequency Shift Due to Ultracold Collisions," K. Gibble and S. Chu, Phys. Rev. Lett. **70**(12), 1771–1774 (1993). Identifies atomic-collision limit to the accuracy of fountain standards. (I)
77. "Observation of the Cesium Clock Transition Using Laser-Cooled Atoms in a Vapor Cell," C. Monroe, H. Robinson, and C. Wieman, Opt. Lett. **16**(1), 50–52 (1991). Suggests a simple cell concept for a cooled-cesium frequency standard. (I)

4. Hydrogen masers

The most common type of hydrogen-maser frequency standard differs from other atomic standards in that it oscillates spontaneously and therefore exhibits very high signal-to-noise ratio. This results in excellent short-term stability. With proper servo control of the resonance of the microwave cavity, the hydrogen maser can also provide exceptional long-term stability. Hydrogen masers can also be operated in a passive mode wherein a local oscillator is tuned to the peak of the transition.

Hydrogen atoms in the maser cavity are contained within a bulb. The atoms interact numerous times with the walls of the bulb, resulting in a very long interrogation time. Since interaction with the walls produces a small frequency shift, the wall coating of the bulb limits the frequency accuracy of the maser. Much work on masers has focused on polymer (PTFE, Teflon) wall coatings. Recent studies indicate that large performance improvements might be achieved through use of a superfluid liquid-helium wall coating.

78. "The Atomic Hydrogen Maser," N. F. Ramsey, IRE Trans. Instrum. **I-11**(3&4), 177–182 (1962). A short review of early work on hydrogen masers. (I)

79. "The Hydrogen Maser," J. Vanier and C. Audoin, Chap. 6 in Ref. 40, Vol. 2, pp. 949–1256. Comprehensive review with an extensive reference list. (A)

80. "Theory of the Hydrogen Maser," D. Kleppner, H. M. Goldenberg, and N. F. Ramsey, Phys. Rev. **126**(2), 603–615 (1962). A classic paper on hydrogen maser theory. (A)

81. "Hydrogen-Maser Principles and Techniques," D. Kleppner, H. C. Berg, S. B. Crampton, N. F. Ramsey, R. F. C. Vessot, H. E. Peters, and J. Vanier, Phys. Rev. A **138**(4A), 972–983 (1965). A good overview of early design principles. (I/A)

82. "The Active Hydrogen Maser: State of the Art and Forecast," J. Vanier, Metrologia **18**(4), 173–186 (1982). A good general review with an extensive reference list. (I/A)

83. "Frequency Standards Based on Atomic Hydrogen," F. L. Walls, Proc. IEEE **74**(1), 142–146 (1986). A brief review covering both active and passive masers. (I)

84. "Experimental Frequency Stability and Phase Stability of the Hydrogen Maser Standard Output as Affected by Cavity Auto-Tuning," H. B. Owings, P. A. Koppang, C. C. MacMillan, and H. E. Peters, in Ref. 9, pp. 92–103. Demonstrates enhanced long-term stability through servo control of the cavity resonance. (I)

85. "Spin-Polarized Hydrogen Maser," H. F. Hess, G. P. Kochanski, J. M. Doyle, T. J. Greytak, and D. Kleppner, Phys. Rev. A **34**(2), 1602–1604 (1986). One of three pioneering efforts on the development of a cryogenic hydrogen maser. (I)

86. "The Cold Hydrogen Maser," R. F. C. Vessot, E. M. Mattison, R. L. Walsworth, and I. F. Silvera, in Ref. 28, pp. 88–94. One of three pioneering efforts on the development of a cryogenic hydrogen maser. (I)

87. "Performance of the UBC Cryogenic Hydrogen Maser," M. C. Hürlimann, W. N. Hardy, M. E. Hayden, and R. W. Cline, in Ref. 28, pp. 95–101. One of three pioneering efforts on the development of a cryogenic hydrogen maser. (I)

5. Stored-ion standards

A particularly promising approach to the problem of Doppler-shift and interrogation-time limitations encountered in cesium-beam standards involves the use of trapped ions. Positive ions can be trapped indefinitely in electromagnetic traps thus eliminating the first-order Doppler shift. They can then be cooled through collisions with a buffer gas to modest temperatures or laser cooled to extremely low temperatures; even the second-order Doppler shift is thus reduced substantially. Using these methods, the systematic energy shifts in transitions in certain ions can be understood with an uncertainty of 1×10^{-18} implying the potential for a frequency standard with this uncertainty.

The construction of such a stored-ion standard poses a very difficult engineering challenge. A key problem to overcome is the lower signal strength associated with the smaller number of particles (ions) involved in most of these standards. Improvement in signal-to-noise ratio can be achieved by increasing the signal and decreasing the noise. Traps of linear geometry readily provide increased signal strength since they store more ions. A proposal has been made to reduce noise using squeezed-state methods.

88. "Atomic Ion Frequency Standards," W. M. Itano, Proc. IEEE **79**(7), 936–42 (1991). A brief review. (I)

89. "Trapped Ions, Laser Cooling, and Better Clocks," D. J. Wineland, Science **226**(4673), 395–400 (1984). A brief review. (E/I)

90. "A Trapped Mercury 199 Ion Frequency Standard," L. S. Cutler, R. P. Giffard, and M. D. McGuirre, in Ref. 37, pp. 563–578. Description of the first buffer-gas-cooled ion standard. (E/I)

91. "Initial Operational Experience with a Mercury Ion Storage Frequency Standard," L. S. Cutler, R. P. Giffard, P. J. Wheeler, and G. M. R. Winkler, in Ref. 14, pp. 12–19. Performance of a buffer-gas-cooled ion standard. (E/I)

92. "Linear Ion Trap Based Atomic Frequency Standard," J. D. Prestage, G. J. Dick, and L. Maleki, IEEE Trans. Instrum. Meas. **40**(2), 132–136 (1991). Improved signal-to-noise performance through use of linear trap geometry. (I)

93. "A 303-MHz Frequency Standard Based on Trapped Be^+ Ions," J. J. Bollinger, D. J. Heinzen, W. M. Itano, S. L. Gilbert, and D. J. Wineland, IEEE Trans. Instrum. Meas. **40**(2), 126–128 (1991). Description of the first laser-cooled ion standard. (I)

94. "Hg^+ Single Ion Spectroscopy," J. C. Bergquist, F. Diedrich, W. M. Itano, and D. J. Wineland, in Ref. 28, pp. 287–291. Concepts for ultra-high-accuracy standards. (I)

95. "Squeezed Atomic States and Projection Noise in Spectroscopy," D. J. Wineland, J. J. Bollinger, W. M. Itano, and D. J. Heinzen, Phys. Rev. A **50**(1), 67–88 (1994). Suggests a fundamental noise-reduction method that could have impact on atomic clocks. (A)

6. Other oscillators

The superconducting-cavity-stabilized oscillator and the cooled-sapphire oscillator do not fit neatly into previous categories and they are not yet widely used. However, the potential for extremely good short-term stability has been demonstrated for both, and they could play a role in the future, so they are mentioned here.

96. "Development of the Superconducting Cavity Maser as a Stable Frequency Source," G. J. Dick and D. M. Strayer, in Ref. 15, pp. 435–446. Provides design details. (I)
97. "Ultra-Stable Performance of the Superconducting Cavity Maser," G. J. Dick and R. T. Wang, IEEE Trans. Instrum. Meas. **40**(2), 174–177 (1991). Description of performance. (I)
98. "Ultra-Stable Cryogenic Sapphire Dielectric Microwave Resonators," A. G. Mann, A. N. Luiten, D. G. Blair, and M. J. Buckingham, in Ref. 9, pp. 167–71. Design and performance description. (I)
99. "Low-Noise, Microwave Signal Generation Using Cryogenic, Sapphire Dielectric Resonators: an Update," M. M. Driscoll and R. W. Weinert, in Ref. 9, pp. 157–162. Design and performance description. (I)

7. Optical-frequency standards and optical-frequency measurement

Many of the techniques described above can be used with optical (rather than microwave) transitions to produce optical-frequency standards. We include a section on this topic because optical-frequency standards have a special niche within the field: their development burgeoned after researchers first measured the speed of light c by measuring the frequency and wavelength of a visible laser. After the accuracy of measurement of c was improved through many measurements, an international agreement defined the speed of light as a constant and redefined the meter in terms of c and the second.

But the interest in optical-frequency standards goes well beyond the redefinition of the meter. Because of their higher Q (or narrower relative linewidth $\Delta f/f$), frequency standards based on optical transitions have the potential for achieving higher performance than those based on microwave transitions. The key disadvantage in using optical transitions is that most applications require access to a frequency in the microwave or lower range. An optical-frequency standard thus requires an auxiliary frequency-synthesis system to accurately relate the optical frequency to some convenient lower frequency. With current technology this is very difficult. Nonetheless, work on optical-frequency standards proceeds with the assumption that the frequency-synthesis methods will be simplified or that optical-frequency standards can be directly useful in the optical region. In fact, good optical-frequency measurements already contribute to more accurate spectral measurements that support a wide range of important applications.

100. "Resource Letter RMSL-1: Recent Measurements of the Speed of Light and the Redefinition of the Meter," H. E. Bates, Am. J. Phys. **56**(8), 682–687 (1988). A good review of the subject with an excellent reference list. (E)
101. "Speed of Light from Direct Frequency and Wavelength Measurements of the Methane-Stabilized Laser," K. M. Evenson, J. S. Wells, F. R. Petersen, B. L. Danielson, G. W. Day, R. L. Barger, and J. L. Hall, Phys. Rev. Lett. **29**(19), 1346–1349 (1972). First measurement of c using this technique. (I)
102. "Laser Frequency Measurements and the Redefinition of the Meter," P. Giacomo, IEEE Trans. Instrum. Meas. **IM-32**(1), 244–246 (1983). A good brief review. (E)
103. "Documents Concerning the New Definition of the Metre," Metrologia **19**(4), 163–178 (1984). Report of the international agreement redefining the meter. (E)
104. "Microwave to Visible Frequency Synthesis," J. J. Jimenez, Radio Science **14**(4), 541–560 (1979). A good review. (I)
105. "Optical Frequency Measurements," D. A. Jennings, K. M. Evenson, and D. J. E. Knight, Proc. IEEE **74**(1), 168–179 (1986). A comprehensive review. (I)
106. "Infrared and Optical Frequency Standards," V. P. Chebotayev, Radio Science **14**(4), 573–584 (1979). A good description of high-stability lasers. (I)
107. "Optical Frequency Standards," J. Helmcke, A. Morinaga, J. Ishikawa, and F. Riehle, IEEE Trans. Instrum. Meas. **38**(2), 524–532 (1989). A good review focusing on more recent work. (I)
108. "Resolution of Photon-Recoil Structure of the 6573-Å Calcium Line in an Atomic Beam with Optical Ramsey Fringes," R. L. Barger, J. C. Bergquist, T. C. English, and D. J. Glaze, Appl. Phys. Lett. **34**(12), 850–852 (1979). Describes application of optical Ramsey fringes (three standing waves) in resolving a calcium line that is currently considered to be an ideal reference for optical-frequency standards. (I)
109. "Optical Ramsey Fringes with Traveling Waves," C. J. Bordé, C. Salomon, S. Avrillier, A. Van Leberghe, C. Breant, D. Bassi, and G. Scoles, Phys. Rev. A **30**(4), 1836–1848 (1984). Theory for optical Ramsey fringes with four traveling waves. (A)

C. Methods of characterizing performance of clocks and oscillators

1. Estimation of systematic effects and random noise

Oscillators and clocks are subject to both systematic effects and random noise. In evaluating the performance of a particular device, we commonly first estimate and remove systematic effects, and then examine the residuals to assess the magnitude of random noise. For most physical systems, the standard variance is used to characterize the random noise and in such systems we usually find that a longer averaging time leads to a lower uncertainty. Unfortunately, the standard variance cannot be applied to clocks and oscillators because this variance is appropriate only if the noise in the system is white, that is, if the noise power is constant over the Fourier frequency interval to which the system is sensitive. Clocks and oscillators exhibit white noise over some frequency range, but for lower frequencies (or longer averaging times) noise components more often depend on negative powers (f^{-1}, f^{-2}, etc.) of the Fourier frequency. In such cases continued averaging of the data can result in progressively poorer results. While the source of the f^{-1} behavior is partially understood for some devices, there is only speculation that the higher-order, nonwhite noise terms are the result of environmental changes affecting systematic terms.

To handle this nonwhite noise, special statistical-characterization techniques have been developed. In the time domain the two-sample (Allan) variance is used to characterize this type of noise. Modifications of this variance have been developed to deal with special situations, but the different variances now in use are all closely related. In the frequency domain, noise in oscillators is characterized by com-

puting the spectral density of either the phase or frequency fluctuations. Spectral density remains a well-behaved quantity in the face of nonwhite noise processes. The choice of approach (time domain *versus* frequency domain) depends on the physical measuring system (as discussed below) and the application. The time-domain specification is most useful for discussing performance in the long term while the frequency-domain measures are most useful for describing short-term behavior. These measures have been the subject of considerable confusion, so the field has adopted standards for terminology and characterization.

Oscillators and clocks respond to changes in environment, so this aspect of characterization is also important. Performance in the face of temperature change is probably of broadest concern, but some applications demand relative insensitivity to, for example, magnetic field, acceleration, and humidity.

110. "Time and Frequency (Time-Domain) Characterization, Estimation, and Prediction of Precision Clocks and Oscillators," D. W. Allan, IEEE Trans. UFFC **UFFC-34**(6), 647–654 (1987). A review of clock/oscillator models and time-domain methods for characterization. (I)

111. "A Frequency-Domain View of Time-Domain Characterization of Clocks and Time and Frequency Distribution Systems," D. W. Allan, M. A. Weiss, and J. L. Jespersen, in Ref. 10, pp. 667–678. Provides a very useful frequency-domain interpretation of the two-sample variances. Introduces a variance useful for characterizing time-transfer systems. (I)

112. "Characterization of Frequency Stability in Precision Frequency Sources," J. Rutman and F. L. Walls, Proc. IEEE **79**(6), 952–960 (1991). Review of time-domain and frequency-domain techniques and their relationships. (I)

113. "The Measurement of Linear Frequency Drift in Oscillators," J. A. Barnes, in Ref. 36, pp. 551–582. Demonstrates difficulty in estimating linear drift. (I)

114. "Confidence on the Second Difference Estimation of Frequency Drift," M. A. Weiss, D. W. Allan, and D. A. Howe, in Ref. 9, pp. 300–305. Improved method for estimating drift. (I/A)

115. "Characterization of Clocks and Oscillators," NIST Technical Note 1337, edited by D. B. Sullivan, D. W. Allan, D. A. Howe, and F. L. Walls (1990). (Available from the Superintendent of Documents, U.S. Government Printing Office, Washington, D. C. 20402-9325.) A collection of reprints with an introductory guide and errata for all of the reprints. (E/I/A)

116. "Characterization of Frequency Stability," J. A. Barnes, A. R. Chi, L. S. Cutler, D. J. Healey, D. B. Leeson, T. E. McGunigal, J. A. Mullen, Jr., W. L. Smith, R. L. Sydnor, R. F. C. Vessot, and G. M. R. Winkler, IEEE Trans. Instrum. Meas. **IM-20**(2), 105–120 (1971). Until the late 1980s, this served as the *de facto* standard for defining measures of performance for clocks and oscillators. Some nomenclature has since changed. (A)

117. "Standard Terminology for Fundamental Frequency and Time Metrology," D. W. Allan, H. Hellwig, P. Kartaschoff, J. Vanier, J. Vig, G. M. R. Winkler, and N. F. Yannoni, in Ref. 13, pp. 419–425. This paper is identical to IEEE Standard 1139-1988. (I)

118. "IEEE Guide for Measurement of Environmental Sensitivities of Standard Frequency Generators," IEEE Standard 1193. (Available from IEEE, 445 Hoes Lane, Piscataway, NJ 08854.) Describes standard methods for characterizing response to changes in environment. (E)

2. Measurement systems

Measurement systems used to characterize clocks and oscillators can be classified into three general categories. These are (1) direct measurements where no signal mixers are used, (2) heterodyne measurements where two unequal frequencies are mixed, and (3) homodyne measurements where two equal frequencies are mixed. The first are by far the simplest, but lack the resolution of the other methods. Measurement methods can also be categorized as being in the time domain or the frequency domain. The time-domain-measurement systems, on the one hand, usually acquire time-series data through repeated time-interval-counter measurements. Proper analysis of these data yields the performance as a function of data-averaging time τ. Such time-domain analysis is most useful for looking at medium-term to long-term noise processes. Frequency-domain measures, on the other hand, use fast Fourier transforms (FFTs) and spectrum analyzers, and are effective for looking at higher-frequency noise processes. Caution must be used when quantitative results are derived from spectrum analyzers, since the type of measurement window used by each instrument affects the results.

119. "Properties of Signal Sources and Measurement Methods," D. A. Howe, D. W. Allan, and J. A. Barnes, in Ref. 17, pp. A1–A47. A very good tutorial. (E/I)

120. "Frequency and Time—Their Measurement and Characterization," S. R. Stein, Chap. 12 in Ref. 41, Vol. 2, pp. 191–232. A comprehensive review of measurement concepts. (I/A)

121. "Phase Noise and AM Noise Measurements in the Frequency Domain," A. L. Lance, W. D. Seal, and F. Labaar, Chap. 7 in **Infrared and Millimeter Waves**, Vol. 11, edited by K. J. Button (Academic, New York, 1984), pp. 239–289. A comprehensive review. (I)

122. "Performance of an Automated High Accuracy Phase Measurement System," S. Stein, D. Glaze, J. Levine, J. Gray, D. Hilliard, D. Howe, and L. Erb, in Ref. 16, pp. 314–320 A method for high-accuracy phase measurement. (I)

123. "Biases and Variances of Several FFT Spectral Estimators as a Function of Noise Type and Number of Samples," F. L. Walls, D. B. Percival, and W. R. Irelan, in Ref. 12, pp. 336–341. Demonstrates the effect of window shape on FFT spectral measurements. (A)

D. Time scales, clock ensembles, and algorithms

Since all clocks exhibit random walk of time at some level, any two independent clocks will gradually diverge in time. Thus, the operation of several clocks at a single site poses a major question: Which clock should be trusted? Timekeeping, furthermore, requires extreme reliability. When a timekeeping system fails, the time must be reacquired from another source. This can be very difficult for

high-accuracy timekeeping. Thus, methods for increasing reliability have great appeal to those charged with maintaining national time scales. To improve reliability and timekeeping performance, a number of national laboratories combine the data from many standards to form something referred to as a clock ensemble. The problem is how to integrate the data from different clocks so as to produce the best possible ensemble time scale.

Combining data from several clocks requires an algorithm that assigns an appropriate weight to each clock and combines the data from all of the clocks in a statistically sound manner. Such algorithms have evolved over the past two decades, and systems based on these algorithms deliver an ensemble performance that is statistically better and much more reliable than that of any single clock in the ensemble.

124. "A Study of the NBS Time Scale Algorithm," M. A. Weiss, D. W. Allan, and T. K. Peppler, IEEE Trans. Instrum. Meas. **38**(2), 631–635 (1989). A study of one of the earliest time-scale algorithms. (I/A)

125. "Comparative Study of Time Scale Algorithms," P. Tavella and C. Thomas, Metrologia **28**(2), 57–63 (1991). Comparisons of the international algorithm (ALGOS) and the NIST algorithm (AT1). (I/A)

126. "Report on the Time Scale Algorithm Test Bed at USNO," S. R. Stein, G. A. Gifford, and L. A. Breakiron, in Ref. 34, pp. 269–288. Comparison of the performance of two algorithms including description of one of the algorithms. (A)

127. "Sifting Through Nine Years of NIST Clock Data with TA2," M. A. Weiss and T. P. Weissert, Metrologia **31**(1), 9–19 (1994). Study of an improved algorithm. (I/A)

128. "An Accuracy Algorithm for an Atomic Time Scale," D. W. Allan, H. Hellwig, and D. J. Glaze, Metrologia **11**, 133–138 (1975). The addition of frequency-accuracy considerations to a time-scale ensemble. (A)

E. International time scales

Until 1967 the world's timekeeping system was based on the motions of the earth (see Chap. 7 of Ref. 42 for details on the history of this period). After the stability of atomic timekeeping was recognized to be far superior to that defined by the earth's motions, the world redefined the second in terms of the cesium atom as noted in Sec. V B 3. Subsequent decisions then gave the world the atomic-time scales denoted as International Atomic Time (TAI) and Coordinated Universal Time (UTC). TAI is a purely atomic scale derived from data from the best atomic clocks in the world. UTC is similarly derived, but incorporates a provision for addition or deletion of "leap seconds" that are needed to keep the world's atomic-timekeeping system in synchronization with the motions of the earth. UTC has the stability of atomic time with a simple means for adjusting to the erratic motions of our planet. This world time scale replaces the familiar Greenwich Mean Time. The Bureau International des Poids et Mesures (BIPM) in Paris serves as the central agent for this international timekeeping activity.

Many laboratories throughout the world contribute raw clock data to the BIPM for the computation of UTC. Most of these in turn steer their own output time signals to UTC, assuring that their broadcast time signals are in very close agreement with this international scale. Each laboratory designates its output signal as UTC(XXXX) where XXXX designates the laboratory. Thus, UTC(NIST) is NIST's best representation of UTC.

129. "Standards of Measurement," A. V. Astin, Sci. Am. **218**(6), 2–14 (1968). Contains a brief discussion of the atomic definition of the second. (E)

130. "The BIPM and the Accurate Measurement of Time," T. J. Quinn, Proc. IEEE **79**(7), 894–905 (1991). A good review of international timekeeping. (E/I)

131. "Establishment of International Atomic Time," BIPM Annu. Rep., D1–D22 (1988). (Available upon request from the Director, BIPM, Pavillon de Breteuil, F-92312 Sèvres Cedex, France.) Description of how TAI is generated. Other parts of this report give a good description of the process of international timekeeping. Includes extensive data for the year. (I/A)

F. Frequency and time distribution

Once a national laboratory generates its estimate of UTC, any of a number of methods can be employed to deliver it to the user. The complexity of the method selected is determined by the accuracy and precision required by the user. At the highest accuracy, the synchronization of widely separated clocks involves consideration of relativistic effects.

132. "Characterization and Concepts of Time-Frequency Dissemination," J. L. Jespersen, B. E. Blair, and L. E. Gatterer, Proc. IEEE **60**(5), 502–521 (1972). An older, but very useful review of the topic. (E/I)

133. "Synchronization and Relativity," G. M. F. Winkler, Proc. IEEE **79**(7), 1029–1039 (1991). A good review. (I)

134. "Practical Implications of Relativity for a Global Coordinate Time Scale," N. Ashby and D. W. Allan, Radio Science **14**(4), 649–669 (1979). A good review. (I/A)

1. One-way time transfer

In "one-way" time transfer, the user receives a broadcast signal that corresponds to a given time scale and then compares the clock to be set with the received time signal. To obtain higher accuracy, some estimate of the time delay associated with transmission is often factored into the setting of the clock. Short-wave broadcasts of timing signals are but one example of this type of time transfer. Such broadcasts usually include a digital time code so that the clock-setting process is readily automated. Several types of one-way broadcasts are described below.

a. Shortwave and low-frequency broadcasts. In the United States, WWV, WWVH, and WWVB, stations operated by NIST, provide broadcasts in the short-wave and low-frequency (LF) regions. The best achievable uncertainty for such broadcasts is about 10^{-11} for frequency and about 100 μs for time. Similar stations are operated by many other countries, and the broadcasts from these stations are regulated by the International Telecommunications Union (ITU).

135. NIST Time and Frequency Services, NIST Special Publication 432 revised (1990), R. E. Beehler and M. A. Lombardi (1991). (Available from the Superintendent of Documents, U. S. Government Printing Office, Washington, D. C. 20402-9325.) Outlines all NIST time-and-frequency dissemination services. (E/I)

136. "The Role of the Consultative Committee on International Radio (CCIR) in Time and Frequency," R. E. Beehler, in Ref. 32, pp. 321–330. Describes the international coordination of timing-signal broadcasts. (E)

b. Broadcasts from geostationary satellites. Geostationary satellites have been used effectively for broadcasts of timing signals of moderate accuracy. Such broadcasts generally include regularly updated information on the location of the satellite so that the receiver, knowing its location and that of the satellite, can correct for the propagation delay. The timing uncertainty (typically 100 μs) of such systems is limited by the accuracy of the satellite position broadcast with the time signal.

137. "NBS Time to the Western Hemisphere by Satellite," D. W. Hanson, D. D. Davis, and J. V. Cateora, Radio Sci. **14**(4), 731–740 (1979). Describes dissemination of timing signals through weather satellites. (E)

138. "Satellite Time Broadcasting of Time and Frequency Signals," A. Sen Gupta, A. K. Hanjura, and B. S. Mathur, Proc. IEEE **79**(7), 973–981 (1991). Describes a service operated by India. (E/I)

c. Digital time codes by telephone. Clocks in computers can be conveniently set using a digital code transmitted by telephone. The first service of this type was developed in Canada. One U.S. version of this service is the Automated Computer Time Service (ACTS). We include this under one-way methods because such services can be operated in that mode. However, a number of these services include provision for operation in the more-accurate two-way mode as described in the next section. More recently, time services of this type have been added to the INTERNET. For information on one such service see the directory /pub/daytime at the INTERNET address time.nist.gov.

139. "A Telephone-Based Time Dissemination System," D. Jackson and R. J. Douglas, in Ref. 35, pp. 541–553. Description of a Canadian system. (E)

140. "The NIST Automated Computer Time Service," J. Levine, M. Weiss, D. D. Davis, D. W. Allan, and D. B. Sullivan, J. Res. NIST **94**(5), 311–321 (1989). Description of a U.S. service. (E)

141. "Keeping Time on Your PC," M. A. Lombardi, Byte **18**(11), 57–62 (1993). A popular review article that describes telephone and other timing delivery methods. (E)

d. LORAN-C. LORAN-C is a U.S. Coast Guard radionavigation system consisting of many stations located throughout the Northern Hemisphere. Special LORAN-C receivers allow a user to precisely determine frequency relative to that of a specific LORAN-C frequency. Since the LORAN-C frequencies are themselves regulated by atomic clocks and carefully steered in frequency to each other, this represents a very useful and readily available frequency reference. With an averaging period of 1 day, frequency uncertainty approaching 1×10^{-12} can be achieved.

142. "Precise Time and Frequency Dissemination via the LORAN-C System," C. E. Potts and B. Wieder, Proc. IEEE **60**(5), 530–539 (1972). A review of LORAN-C timing. (E)

143. Traceable Frequency Calibrations: How to Use the NBS Frequency Measurement System in the Calibration Lab, NBS Special Publication 250-29, G. Kamas and M. A. Lombardi, (U. S. Department of Commerce, Washington, D. C., 1988). (Available from the Superintendent of Documents, U. S. Government Printing Office, Washington, D. C. 20402-9325.) Description of a service based on LORAN-C. (E)

e. Global Positioning System (GPS). The Global Positioning System (GPS) is a system of satellites operated by the U. S. Department of Defense. These satellites are used for navigation and timing purposes. With a GPS receiver, one can determine local time relative to that of the GPS system clock. The GPS system clock is traceable to a time scale maintained by the U. S. Naval Observatory.

The GPS signal is subject to Selective Availability, an intentional degradation aimed at reducing the real-time accuracy available to civilian users of the system. With the current level of Selective Availability, timing uncertainty may approach 100 to 200 ns under favorable conditions. With the character of degradation (noise) now used, local averaging can improve the accuracy, but there is no guarantee that the character of the degradation will remain the same.

144. "Using a New GPS Frequency Reference in Frequency Calibration Operations," T. N. Osterdock and J. A. Kusters, in Ref. 8, pp. 33–39. GPS for frequency measurement. (E)

145. "A Precise GPS-Based Time and Frequency System," J. McNabb and E. Fossler, in Ref. 18, pp. 387–389. GPS for time and frequency. (E)

146. "Real-Time Restitution of GPS Time Through a Kalman Estimation," C. Thomas, Metrologia **29**(6), 397–414 (1992). Description of a process for improving the accuracy of GPS time signals. (A)

f. Wide Area Augmentation System (WAAS). In the near future, a signal very similar to that used in GPS will be broadcast from a number of commercial geosynchronous communications satellites. Satellites covering the U.S. and adjacent regions will be operated by the U.S. Federal Aviation Administration (FAA). The primary purpose of this satellite system is to provide critical information on the integrity of GPS signals so that GPS can be used for commercial air navigation. Fortunately, among other features of this service, the FAA is including provisions for delivery of highly accurate timing signals. The timing accuracy achievable using these signals, while yet to be demonstrated, should be at least comparable to and perhaps better than that of GPS.

147. "Precise Time Dissemination Using the INMARSAT Geostationary Overlay," A. Brown, D. W. Allan, and R. Walton, in Ref. 8, pp. 55–64. Describes preliminary experiments demonstrating the utility of WAAS for timing applications. (I/A)

2. Common-view time transfer

In common-view time transfer, two sites compare their time and frequency by recording the times (according to the clock of each lab) of arrival of signals emanating from a single source that is in view of both sites.

A well-established method of common-view time-and-frequency transfer uses GPS signals. In common-view GPS time transfer, most of the effects of selective availability are canceled out. This comparison method can yield uncertainties in time comparison of 1–10 ns for an averaging time of 1 day and can compare the frequency of the clocks with an uncertainty on the order of 10^{-14}. The Russian counterpart to GPS is known as GLONASS. Common-view time transfer using GLONASS has also been studied, but only more re-

cently. The availability of this second independent satellite system for use in time transfer should prove useful in the future.

148. "GPS Time Transfer," W. Lewandowski and C. Thomas, Proc. IEEE **79**(7), 991–1000 (1991). Discussion of performance of common-view time transfer. (I).

149. "GPS Time Closure Around the World Using Precise Ephemerides, Ionospheric Measurements and Accurate Antenna Coordinates," W. Lewandowski, G. Petit, C. Thomas, and M. A. Weiss, in Ref. 19, pp. 215–220. An exacting test of the accuracy of GPS common-view time transfer. (I)

150. "Precision and Accuracy of GPS Time Transfer," W. Lewandowski, G. Petit, and C. Thomas, IEEE Trans. Instrum. Meas. **42**(2), 474–479 (1993). A good review. (I)

151. "An NBS Calibration Procedure for Providing Time and Frequency at a Remote Site by Weighting and Smoothing of GPS Common View Data," M. A. Weiss and D. W. Allan, IEEE Trans. Instrum. Meas. **IM-36**(2), 572–578 (1987). A report on the earliest work on the subject. (I)

152. "Comparison of GLONASS and GPS Time Transfers," P. Daly, N. B. Koshelyaevsky, W. Lewandowski, G. Petit, and C. Thomas, Metrologia **30**(2), 89–94 (1993). Comparison of time transfer using the Russian and U.S. systems. (I)

3. Two-way time transfer

The two-way method assumes that the path connecting the two participants is reciprocal, that is, that the signal delay through the transmission medium in one direction is the same as that in the reverse direction. In two-way time transfer, each of two sites transmits and receives signals. The timing of the transmitted signal at each site is linked to the site's time scale. If these transmissions occur at nearly the same time, the transmission delay between the sites cancels when the data recorded at the two sites are compared. The result is a time comparison dependent only on imperfections in the transmission hardware along the path and on second-order effects relating to reciprocity.

Several of the telephone time services described in Sec. V F 1 c above include a provision for two-way operation that allows them to achieve an uncertainty approaching 1 ms rather than the 10 to 30 ms normally encountered in the one-way mode. The two-way method produces much higher accuracy when used with satellite or optical-fiber links.

a. Two-way time transfer using satellites. Commercial communication satellites are typically used for this time transfer. Properly implemented, such time transfer should achieve a time-comparison uncertainty of less than 1 ns and stability of a few hundred picoseconds.

153. "Telstar Time Synchronization," J. McA. Steele, W. Markowitz, and C. A. Lidback, IEEE Trans. Instrum. Meas. **IM-13**(4), 164–170 (1964). An early two-way experiment. (E/I)

154. "Two-Way Time Transfer Experiments Using an INTELSAT Satellite in a Inclined Geostationary Orbit," F. Takahashi, K. Imamura, E. Kawai, C. B. Lee, D. D. Lee, N. S. Chung, H. Kunimori, T. Yoshino, T. Otsubo, A. Otsuka, and T. Gotoh, IEEE Trans. Instrum. Meas. **42**(2), 498–504 (1993). A Japanese–Korean experiment. (I)

155. "NIST-USNO Time Comparisons Using Two-Way Satellite Time Transfers," D. A. Howe, D. W. Hanson, J. L. Jespersen, M. A. Lombardi, W. J. Klepczynski, P. J. Wheeler, M. Miranian, W. Powell, J. Jeffries, and A. Meyers, in Ref. 12, pp. 193–198. A U.S. experiment. (I)

156. "Comparison of GPS Common-view and Two-way Satellite Time Transfer Over a Baseline of 800 km," D. Kirchner, H. Ressler, P. Grudler, F. Baumont, C. Veillet, W. Lewandowski, W. Hanson, W. Klepczynski, and P. Uhrich, Metrologia **30**(3), 183–192 (1993). A comparison of the two methods. (I)

b. Two-way time transfer in optical fiber. The two-way method has also been applied to time transfer in optical fiber. The broad bandwidth of optical fiber and its immunity to pickup of electromagnetic noise make it an attractive option for linking sites that must be tightly synchronized. Furthermore, since optical fiber is used so widely in telecommunications, telecommunications synchronization might one day take advantage of this method.

157. "Characteristics of Fiber-Optic Frequency Reference Distribution Systems," R. A. Dragonette and J. J. Suter, in Ref. 18, pp. 379–382. Local-area timing distribution. (I)

158. "Precise Frequency Distribution Using Fiber Optics," R. L. Sydnor and M. Calhoun, in Ref. 18, pp. 399–407, Local-area timing distribution. (I)

159. "Accurate Network Synchronization for Telecommunications Networks Using Paired Paths," M. Kihara and A. Imaoka, in Ref. 18, pp. 83–87. Description of long-line (2000 km) timing experiments. (I)

VI. APPLICATIONS

In a technological society, the need for accurate time and frequency is ubiquitous. The uses of frequency and time range from the commonplace need to know when it is time to go to lunch to scientifically demanding tests of relativity theory. In fact, science has often been the driving force behind many of the major advances in timekeeping, but industry has quickly commercialized much of this technology for applications involving transportation, navigation, power generation and distribution, and telecommunications. The following references discuss some of the more notable commercial and scientific applications. These are offered without comment.

160. "Time, The Great Organizer," J. Jespersen and J. Fitz-Randolph, Chap. 11 in Ref. 42, pp. 99–109. (E)

161. "Atomic Frequency Standards in Technology, Science and Metrology," J. Vanier and C. Audoin, Chap. 9 in Ref. 40., Vol. 2, pp. 1531–1543. (E/I)

162. "Uses of Precise Time and Frequency in Power Systems," R. E. Wilson, Proc. IEEE **79**(7), 1009–1018 (1991). (E/I)

163. "Field Experience with Absolute Time Synchronism Between Remotely Located Fault Recorders and Sequence of Events Recorders," R. O. Burnett, Jr., IEEE Trans. Power App. Syst. **PAS-103**(7), 1739–1742 (1984). (I)

164. "Synchronization in Digital Communications Networks," P. Kartaschoff, Proc. IEEE **79**(7), 1019–1028 (1991). (E/I)

165. "Network Synchronization," W. C. Lindsey, F. Ghazvinian, W. C. Hagmann, and K. Dessouky, Proc. IEEE **73**(10), 1445–1467 (1985). (A)

166. "Network Timing and Synchronization," D. R. Smith, Chap. 10 in **Digital Transmission Systems** (Van Nostrand Reinhold, New York, 1985), pp. 439–490. (I)

167. "Military Applications of High Accuracy Frequency Standards and Clocks," J. R. Vig, IEEE Trans. UFFC **40**(5), 522–527 (1993). (E)

168. "Very Long Baseline Interferometer Systems," J. M. Moran, Chap. 5.3 in **Methods of Experimental Physics**, Vol. 12, **Astrophysics**, Part C: *Radio Observations* (Academic, New York, 1976), pp. 174–197. (I)

169. "Applications of Highly Stable Oscillators to Scientific Measurements," R. F. C. Vessot, Proc. IEEE **79**(7), 1040–1053 (1991). (A)

170. "Time, Frequency and Space Geodesy: Impact on the Study of Climate and Global Change," P. F. MacDoran and G. H. Born, Proc. IEEE **79**(7), 1063–1069 (1991). (I)

171. "Test of Relativistic Gravitation with a Space-Borne Hydrogen Maser," R. F. C. Vessot, M. W. Levine, E. M. Mattison, E. L. Blomberg, T. E. Hoffman, G. U. Nystrom, B. F. Farrel, R. Decher, P. B. Eby, C. R. Baugher, J. W. Watts, D. L. Teuber, and F. D. Wills, Phys. Rev. Lett. **45**(26), 2081–2084 (1980). (I/A)

172. "Experimental Test of the Variability of G Using Viking Lander Ranging Data," R. W. Hellings, P. J. Adams, J. D. Anderson, M. S. Keesey, E. L. Lau, E. M. Standish, V. M. Canuto, and J. Goldman, Phys. Rev. Lett. **51**(18), 1609–1612 (1983). (A)

173. "Test of the Principle of Equivalence by a Null Gravitational Red-Shift Experiment," J. P. Turneaure, C. M. Will, B. F. Farrell, E. M. Mattison, and R. F. C. Vessot, Phys. Rev. D **27**(8), 1705–1714 (1983). (A)

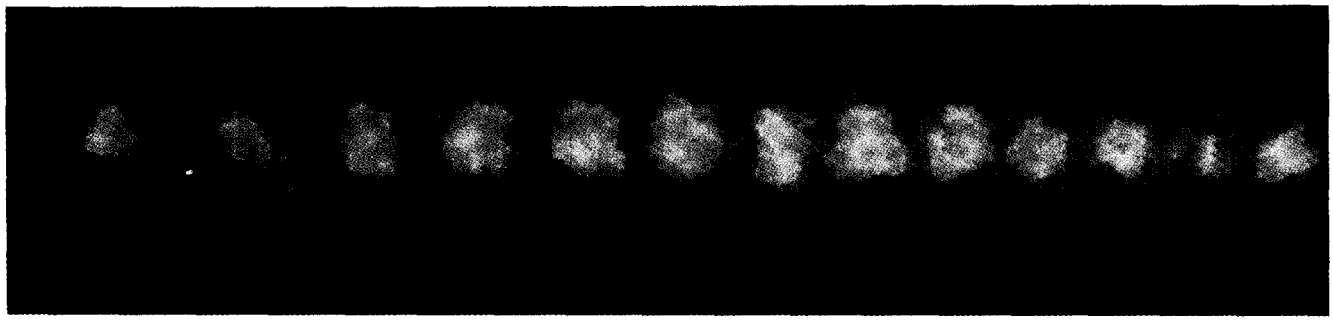
174. "Millisecond Pulsars: Nature's Most Stable Clocks," J. H. Taylor, Jr., Proc. IEEE **79**(7), 1054–1062 (1991). (I)

175. "Time and Frequency in Fundamental Metrology," B. W. Petley, Proc. IEEE **79**(7), 1070–1076 (1991). (I)

Accurate Measurement of Time

Increasingly accurate clocks—now losing no more than a second over millions of years—are leading to such advances as refined tests of relativity and improved navigation systems

by Wayne M. Itano and Norman F. Ramsey



Few people complain about the accuracy of modern clocks, even if they appear to run more quickly than the harried among us would like. The common and inexpensive quartz-crystal watches lose or gain about a second a week—making them more than sufficient for everyday living. Even a spring-wound watch can get us to the church on time. More rigorous applications, such as communications with interplanetary spacecraft or the tracking of ships and airplanes from satellites, rely on atomic clocks, which lose no more than a second over one million years.

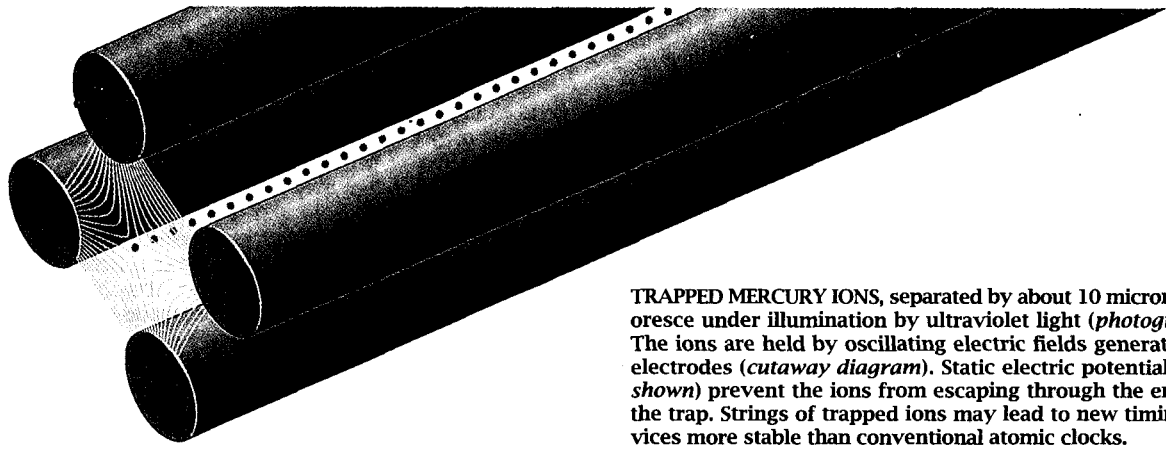
WAYNE M. ITANO and NORMAN F. RAMSEY have collaborated many times before writing this article: Itano earned his Ph.D. at Harvard University under the direction of Ramsey. Itano, a physicist at the Time and Frequency Division of the National Institute of Standards and Technology in Boulder, Colo., concentrates on the laser trapping and cooling of ions and conducts novel experiments in quantum mechanics. He is also an amateur paleontologist and fossil collector. Ramsey, a professor of physics at Harvard, earned his Ph.D. from Columbia University. He has also received degrees from the University of Oxford and the University of Cambridge, as well as several honorary degrees. A recipient of numerous awards and prizes, Ramsey achieved the highest honor in 1989, when he shared the Nobel Prize in Physics for his work on the separated oscillatory field method and on the atomic hydrogen maser.

There might not seem to be much room for the improvement of clocks or even a need for more accurate ones. Yet many applications in science and technology demand all the precision that the best clocks can muster, and sometimes more. For instance, some pulsars (stars that emit electromagnetic radiation in periodic bursts) may in certain respects be more stable than current clocks. Such objects may not be accurately timed. Meticulous tests of relativity and other fundamental concepts may need even more accurate clocks. Such clocks will probably become available. New technologies, relying on the trapping and cooling of atoms and ions, offer every reason to believe that clocks can be 1,000 times more precise than existing ones. If history is any guide, these future clocks may show that what is thought to be constant and immutable may on finer scales be dynamic and changing. The sundials, water clocks and pendulum clocks of the past, for example, were sufficiently accurate to divide the day into hours, minutes and seconds, but they could not detect the variations in the earth's rotation and revolution.

A clock's accuracy depends on the regularity of some kind of periodic motion. A grandfather clock relies on the sweeping oscillation of its pendulum. The arm is coupled to a device called an escapement, which strikes the teeth of a gear in such a way that the gear moves in only one direction. This gear, usually through a series of additional gears, transfers the motion to the hands of the clock. Efforts to improve clocks are directed for the most part toward finding systems in which the oscillations are highly stable.

The three most important gauges of frequency standards are stability, reproducibility and accuracy. Stability is a measure of how well the frequency remains constant. It depends on the length of an observed interval. The change in frequency of a given standard might be a mere one part per 100 billion from one second to the next, but it may be larger—say, one part per 10 billion—from one year to the next. Reproducibility refers to the ability of independent devices of the same design to produce the same value. Accuracy is a measure of the degree to which the clock replicates a defined interval of time, such as one second.

Until the early 20th century, the most accurate clocks were based on the regularity of pendulum motions. Galileo had noted this property of the pendulum after he observed how the period of oscillation was approximately independent of the amplitude. In other words, a pendulum completes one cycle in about the same amount of time, no matter how big each sweep is. Pendulum clocks became possible only after the mid-1600s, when the Dutch scientist Christiaan Huygens invented an escapement to keep the pendulum swinging. Later chronometers used the oscillations of balance wheels attached



TRAPPED MERCURY IONS, separated by about 10 microns, fluoresce under illumination by ultraviolet light (*photograph*). The ions are held by oscillating electric fields generated by electrodes (*cutaway diagram*). Static electric potentials (*not shown*) prevent the ions from escaping through the ends of the trap. Strings of trapped ions may lead to new timing devices more stable than conventional atomic clocks.

to springs. These devices had the advantage of being portable.

Considerable ingenuity went into improving the precision of pendulum and balance-wheel clocks. Clockmakers would compensate for temperature changes by combining materials with different rates of thermal expansion. A more radical approach came in the 1920s, when William H. Shortt, a British engineer, devised a clock in which a "slave pendulum" was synchronized to a "free pendulum." The free pendulum oscillates in a low-pressure environment and does not have to operate any clock mechanism. Instead it actuates an electrical switch that helps to keep the slave pendulum synchronized. As a result, the period of the Shortt clock is extremely stable. These clocks had an error of a few seconds in a year (about one part per 10 million) and became the reference used in laboratories.

The next major advance in timekeeping was based on the development of quartz-crystal electronic oscillators. The frequency of such devices depends on the period of the elastic vibration of a carefully cut quartz crystal. The vibrations are electronically maintained through a property of such crystals called piezoelectricity. A mechanical strain on the crystal produces a low electric voltage; inversely, a voltage induces a small strain.

The quartz vibrates at a frequency that depends on the shape and dimensions of the crystal. In some wristwatches, it is cut into the shape of a tuning fork a few millimeters long. In other timepieces, it is a flat wafer. The quartz is connected to an electric circuit that produces an alternating current. The electrical feedback from the quartz causes the frequency of the circuit to match the frequency at which the crystal naturally vibrates (usually 32,768 hertz). The alternating current from the circuit goes to a frequency divider, a digital electronic device that generates one output pulse for a fixed number of input pulses. The divider also actuates either a mechanical or digital electronic display.

In the late 1920s Joseph W. Horton and Warren A. Marrison, then at Bell Laboratories, made the first clock based on a quartz-crystal oscillator. In the 1940s quartz-crystal clocks replaced Shortt pendulum clocks as primary laboratory standards. These clocks were stable to about 0.1 millisecond per day (about one part per billion). Relatively inexpensive, quartz clocks continue to be extensively used. The timekeeping elements of common quartz watches and clocks are simplified and miniaturized versions of quartz frequency standards. Quartz wristwatches became common once the ability emerged to cut the quartz into thin, tuning-fork shapes reliably and to manufacture miniature, low-power digital electronic components.

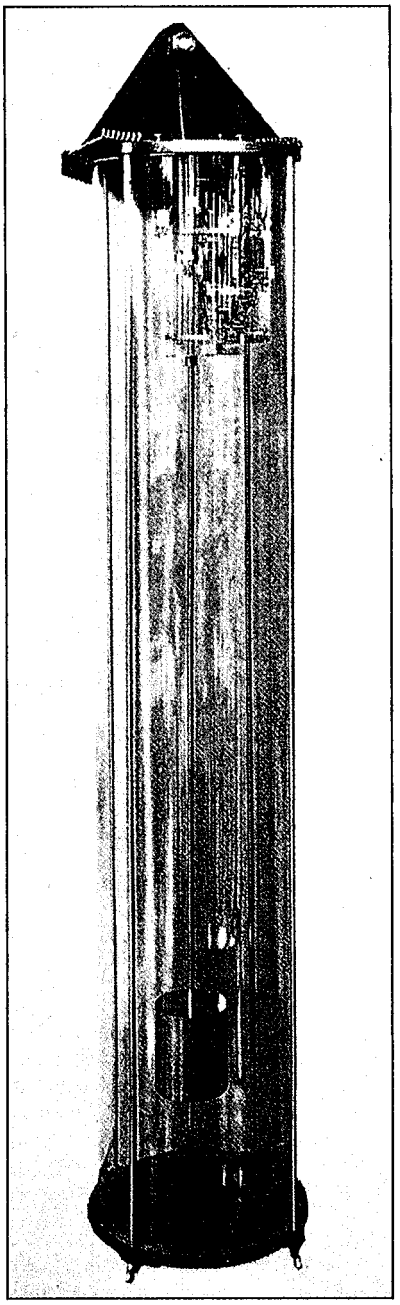
Yet quartz-crystal clocks prove inadequate for many scientific applications, such as tests of relativity. According to Albert Einstein's calculations, gravity distorts both space and time. The difference in gravitational potential causes time to pass more quickly high in the atmosphere than it does on the surface. The difference is slight. Time runs about 30 millionths of a second per year faster at the top of Mount Everest than it does at sea level. Only atomic frequency standards achieve the requisite precision.

The quantized energy levels in atoms and molecules provide the physical basis for atomic frequency standards. The laws of quantum mechanics dictate that the energies of a bound system, such as an atom, have certain discrete values. An electromagnetic field can boost an atom from one energy level to a higher one. The process can also work in reverse. If the atom is in a high energy level, it can drop to a lower level by emitting electromagnetic energy.

The maximum amount of energy is absorbed or emitted at a definite frequency—the resonance frequency, or the difference between the two energy levels divided by Planck's constant. This value is sometimes called the Bohr frequency. Such frequencies make ideal time standards because they are extremely stable. Time can be kept by observing the frequencies at which electromagnetic energy is emitted or absorbed by the atoms. In essence, the atom serves as the master pendulum whose oscillations are counted to mark the passage of time.

Although we have described general quantum properties, the effects exploit-

ed in atomic clocks are slightly more complicated. In most atomic clocks the energy that atoms absorb or release actually results from transitions between so-called hyperfine energy levels. These levels exist because of an intrinsic property of particles known as the magnetic moment. Electrons and the nuclei of most atoms spin about their axes as if they were tops. In addition, they are magnetized, like compass needles oriented along their axes of rotation. These axes can have different orientations with respect to one another, and the energies of the orientations may differ.



MASTER PENDULUM of this 1920s Shortt clock oscillates in an evacuated enclosure. It actuates an electrical switch to synchronize a slave pendulum, which drives the clock mechanism.

These positions correspond to the hyperfine levels. The nomenclature comes about because the levels were first observed in spectroscopy as small splittings of spectral lines.

On paper, standards based on atomic processes are ideal. In practice, perfection is elusive. Atoms do not absorb or emit energy precisely at the resonance frequency. Some energy is spread over a small interval surrounding the frequency—a smearing of frequencies, so to speak. All else being equal, the precision to which the resonance frequency can be measured is inversely proportional to this smearing. The greater the spread, the less precise the measurement. The spread is often expressed in terms of the quality factor, or Q, which is equal to the resonance frequency divided by the frequency spread. In many cases, the higher the resonance frequency, the higher the Q. Furthermore, smearing is often inversely proportional to the time the atom is in the apparatus. In those situations, the Q of the resonance, and hence the precision of the measurement, increases as the measuring time increases.

The motions of the atoms also introduce uncertainty by causing apparent shifts in the resonance frequencies. Such changes appear because of the Doppler effect. The phenomenon can be divided into first- and second-order shifts if the atoms are moving much slower than the speed of light. The first-order Doppler shift is an apparent change in the frequency of the applied electromagnetic wave as seen by a moving atom. The amount of the shift is proportional to the velocity of the atom. If the atom moves in the same direction as the wave does, the shift is to a lower frequency. If the atom's motion is opposed to that of the wave, the shift is to a higher frequency. If the directions are perpendicular, the first-order shift is zero.

The second-order Doppler shift comes about as a consequence of time dilation. According to relativity, time slows down for objects in motion; a moving atom "sees" a slightly different frequency than does a stationary counterpart. The effect on the resonance frequency is usually much smaller than the first-order shift. The second-order shift is proportional to the square of the atomic velocity and does not depend on the relative directions of the atomic motion and the electromagnetic wave.

Several other factors affect the quality of the information. Atoms in the system may collide with one another; the impacts add noise to the signal. The surrounding environment can perturb the resonance frequencies. Defects in the electronic equipment, stray electromagnetic fields and the ever present thermal radiation all introduce errors. Therefore, a good atomic frequency standard not only must establish a steady, periodic signal but also must minimize these potential errors.

One of the earliest and now widely used methods to sidestep many of these difficulties is called atomic beam resonance, pioneered by I. I. Rabi and his colleagues at Columbia University in the 1930s. The atoms emerge from a small chamber, exit through a narrow aperture and then travel as a beam. The entire instrument can be shielded from stray magnetic and electric fields and insulated from external sources of heat. Perhaps more important, collisions of atoms are virtually eliminated, because the entire device is housed in a long, evacuated chamber. The pressure in the chamber is so low that the atoms are unlikely to strike anything before reaching the other end.

In simplified form, atomic beam resonance involves three steps. The first is to select only those atoms in the appropriate energy level. This selection is accomplished by using a specially shaped magnetic field, which acts as a kind of filter. It allows atoms in one energy level to pass and blocks all others by bending the beam. Only atoms in the correct energy level are bent the correct amount to reach and pass through the aperture that serves as the entrance to the cavity.

The second and crucial step is to send the selected atoms into another energy level. The task is accomplished by passing the atoms through an oscillating microwave field inside a cavity. The atoms will go to another energy level only if the frequency of the applied oscillating microwaves matches their Bohr frequency.

The third step is to detect those atoms that have changed energy levels. At this point, the beam of atoms passes through another magnetic field filter, which allows only atoms in the correct energy level to strike a detector that records the atoms as current flow. An abundance of such atoms will exist if the frequency of the applied oscillating microwaves precisely matches their natural frequency. If the frequency of the applied microwave field is off the mark, fewer atoms change their energy

levels, and so fewer will strike the detector. One knows, therefore, that the applied microwaves match the natural frequency of the atoms if the number of atoms striking the detector is maximal. An electronic feedback mechanism, called a servo loop, keeps this value constant. If it finds that the current from the detector is falling off, it changes the frequency of the applied field until the current reaches a maximum again.

By keeping the current from the detector at a maximum, the servo loop maintains the frequency of the applied microwave field at the natural frequency of the atoms. To measure time, one couples the applied field to a frequency divider, which generates timing pulses. By analogy, the atoms represent the quartz crystal in a watch or the master pendulum in a Shortt clock. The applied microwave field is the oscillating circuit or the slave pendulum, which actually drives the clock mechanism.

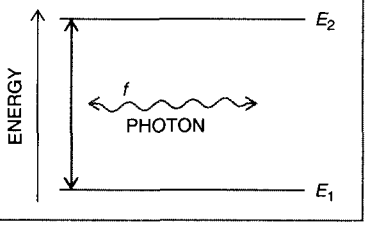
Minor variations of the atomic beam standard exist. For example, in some devices the atoms that undergo a change in energy level are made to miss, rather than strike, the detector. Not much difference in accuracy exists, however. Rather all the versions to some extent represent trade-offs in terms of size, cost and complexity.

A more important modification of the atomic beam came in 1949, when one of us (Ramsey) invented the so-called separated oscillatory field method. Instead of irradiating the atoms with a single applied field, this technique relies on two fields, separated by some distance along the beam path. Applying the oscillating field in two steps has many benefits, including a narrowing of the resonance and the elimination of the first-order Doppler shift. Jerrold R. Zacharias of the Massachusetts Institute of Technology and Louis Essen and John V. L. Parry of the National Physical Laboratory in Teddington, England, adapted this method to working frequency standards in the mid-1950s.

Currently the separated oscillatory field method provides the most reproducible clocks. The best ones are located at a few national laboratories, although smaller and less accurate versions are commercially available. The clocks rely on cesium, which has several advantages over other elements. It has a relatively high resonance frequency—about 9,192 megahertz—and low resonance width, which lead to an excellent Q. Cesium can also be detected readily and efficiently; all that is needed is a hot metal filament. When a cesium atom strikes the filament, it ionizes and becomes observable as electric current.

Resonance Frequency

Atomic frequency standards depend on the quantization of the internal energies of atoms or molecules. A pair of such energy levels, shown here as levels E_1 and E_2, is associated with an atomic resonance. The resonance frequency f, at which it absorbs or emits electromagnetic radiation, is $f = (E_2 - E_1)/h$, where h is Planck's constant. The radiation, however, is not precisely f but instead is spread over a range near f, called Δf. The precision to which f can be measured is proportional to the quality factor, Q, defined by $Q = f/\Delta f$. The higher the Q, the more stable the clock.



The Qs of these standards are about 100 million, exceeding the Q of quartz wristwatches by a factor of several thousand. The greatest reproducibilities are about a part per 10^{14}. The best cesium frequency standards are so much more reproducible than the rate of rotation and revolution of the earth that in 1967 the second was defined as 9,192,631,770 periods of the resonance frequency of the cesium 133 atom.

One of the most promising improvements in cesium atomic-beam standards is the use of optical pumping to select the atomic states. Beginning in the 1950s optical-pumping techniques were developed by Francis Bitter of M.I.T., Alfred Kastler and Jean Brossel of the École Normale Supérieure and others. In this method, light, rather than a magnetic field, selects atoms in the desired states. Before the atoms are subjected to the microwave field, radiation from a laser is used to drive (or pump) the atoms from one energy level into another. In fact, one can control the number of atoms in energy levels by tuning the frequency of the light.

After the atoms have been irradiated by the microwave field, they pass through a second light beam. Only atoms occupying the correct energy level absorb this light, which they quickly reemit. A light-sensitive detector records the reemissions and converts them into a measurable current. As in atomic beam resonance that relies on magnetic selection, one knows that the applied microwave field matches the natural frequency of the atoms if the current from the detector is at a maximum.

Using light instead of magnets has many advantages. Perhaps the most crucial is that, with the right optical-pumping techniques, all the atoms in the beam can be put into the desired energy level. Magnetic selection merely filters out those that are in the other energy levels. Hence, the signal strength from optical pumping is much higher than it is from magnetic selection. Researchers at various laboratories are developing optically pumped cesium atomic-beam clocks. One such clock, at the National Institute of Standards and Technology (NIST) in Boulder, Colo., has recently become the primary frequency standard for the U.S. Designated NIST-7, it has an expected error of one second in about one million years, making it many times more stable than its predecessor.

There is an optically pumped atomic clock that is available commercially. Such a clock is based on the 6,835-megahertz, hyperfine resonance of rubidium 87. Rather than moving through the apparatus as a beam, the rubidium atoms are contained in a glass cell. The cell also houses a mixture of gases that prevents the rubidium atoms from colliding with the cell walls. A discharge lamp containing rubidium vapor, rather than a laser, irradiates the atoms. A photovoltaic sensor on the opposite side of the cell detects changes in the amount of light absorbed by the atoms. The atoms are prepared, the microwaves applied and the light detected in one cell. As a result, rubidium clocks can be made to fit in a cube about 10 centimeters on a side. In contrast, cesium beam clocks can extend from about 50 centimeters to more than five meters. Rubidium clocks are also much less expensive than are cesium ones.

The drawback is that the rubidium devices are generally less accurate and less reproducible. The Q of rubidium standards is about 10 million, a factor of 10 less than the cesium beam's quality factor; their reproducibility is only about a part per 10^{10}. Shifts in the resonance frequency mostly account for the poor reproducibility. The frequent collisions of the rubidium atoms with other gas molecules cause the shifts. But the rubidium standards' short-term stabilities are good—in fact, better than those of some cesium atomic beams.

The atomic clocks described thus far work in a rather roundabout way—by

detecting a change in some signal, such as the number of atoms striking a detector, as the frequency of the applied oscillatory field shifts. One way to make use of the radiation emitted by the atoms more directly relies on the principle of the maser (an acronym for microwave amplification by stimulated emission of radiation). In 1953 Charles H. Townes and his associates at Columbia invented the first maser, which was based on ammonia. Beginning in 1960, Ramsey, Daniel Kleppner, now at M.I.T., H. Mark Goldenberg, then at Harvard University, and Robert F. C. Vessot, now at the Harvard-Smithsonian Center for Astrophysics, developed the atomic hydrogen maser, the only type that has been used extensively as an atomic clock.

In this instrument, a radio frequency discharge first splits hydrogen molecules held in a high-pressure bottle into their constituent atoms. The atoms emerge from a small opening in the bottle, forming a beam. Those in the higher energy level are focused by magnetic fields and enter a specially coated storage bulb surrounded by a tuned, resonant cavity.

In the bulb, some of these atoms will drop to a lower energy level, releasing photons of microwave frequency. The photons will stimulate other atoms to fall to a lower energy level, which in turn releases additional microwave photons. In this manner, a self-sustaining microwave field builds up in the bulb—thus the name "maser." The tuned cavity around the bulb helps to redirect photons back into the system to maintain the stimulated emission process. The maser oscillation persists as long as the hydrogen is fed into the system.

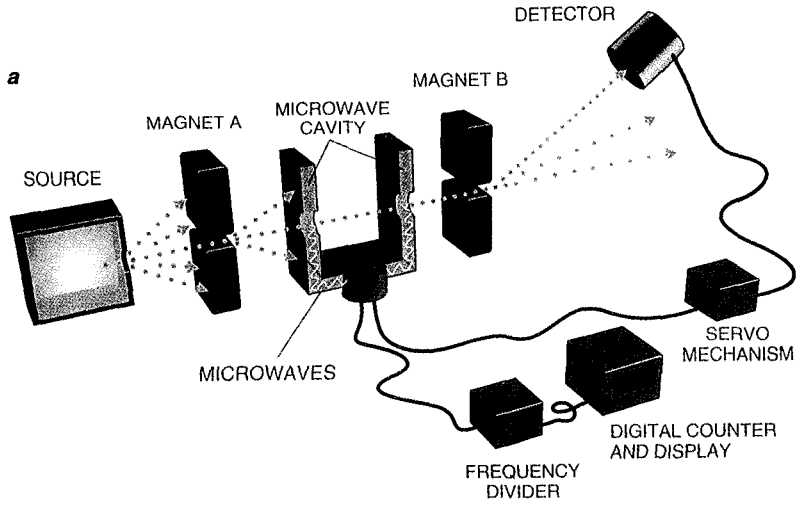
A loop of wire in the cavity can detect the oscillation. The microwave field induces a current in the wire, which leads out of the cavity to a series of circuits. The circuits convert the induced current to a lower frequency signal suitable for generating timing pulses.

The resonance frequency in the hydrogen maser is about 1,420 megahertz, which is much lower than the resonance frequency of cesium. But because the hydrogen atoms reside in the bulb much longer than cesium atoms do in a beam, the maser's resonance width is much narrower. Consequently, the Q of a hydrogen maser standard is about 10^9, exceeding the Q of the cesium atomic clock by an order of magnitude. In addition, a hydrogen maser has the highest stability of any frequency standard, better than one part per 10^{15}.

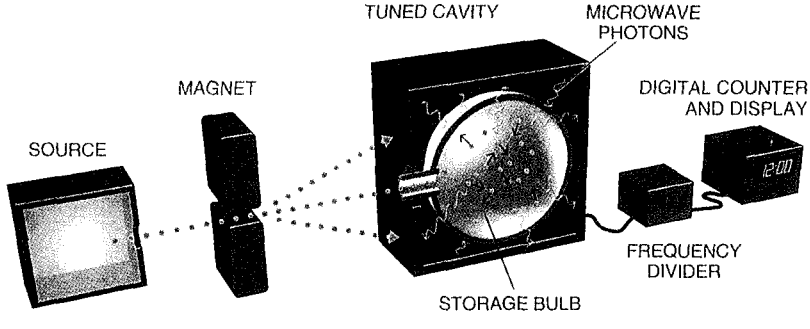
Unfortunately, the maser's superior attributes last just for a few days. Beyond that, its performance falls below that of cesium beams. The stability decreases because of changes in the cavity's resonant frequency. Collisions between the atoms and the bulb shift the frequency by about one part per 10^{11}.

One way to overcome the problem is to operate the hydrogen maser at low temperatures. This condition allows more atoms to be stored (thus resulting in a stronger signal) and reduces electronic noise. Coating the inside of the bulb with superfluid liquid helium also enhances performance. This substance acts as a good surface against which the hydrogen atoms can bounce. More effective magnets, better coating substances and servo loop techniques that keep the cavity resonance centered on the atomic resonance are other approaches now being taken to improve maser stability.

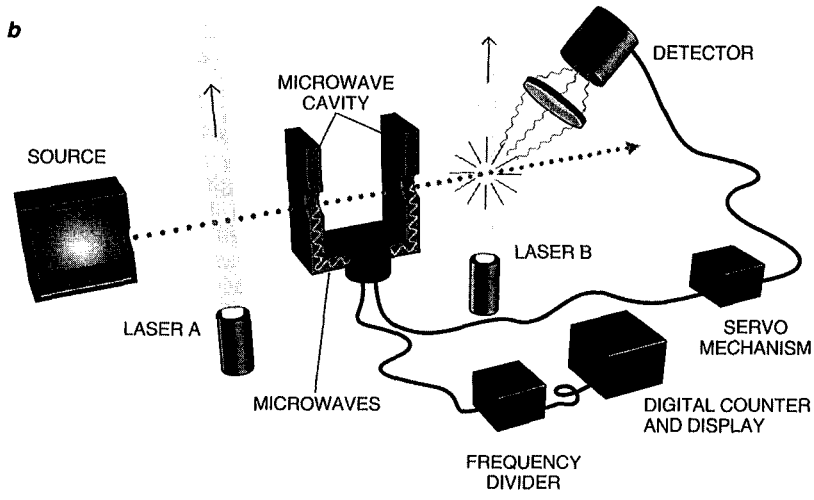
Although the cesium atomic-beam frequency standard is the most accurate, long-term standard we have, several breakthroughs have indicated that it is possible to fabricate even more precise clocks. One of the most promising depends on the resonance frequency of trapped, electrically charged ions. Trapped ions can be suspended in a vacuum so that they are almost perfectly isolated from disturbing influences. The ions themselves stay well separated from one another

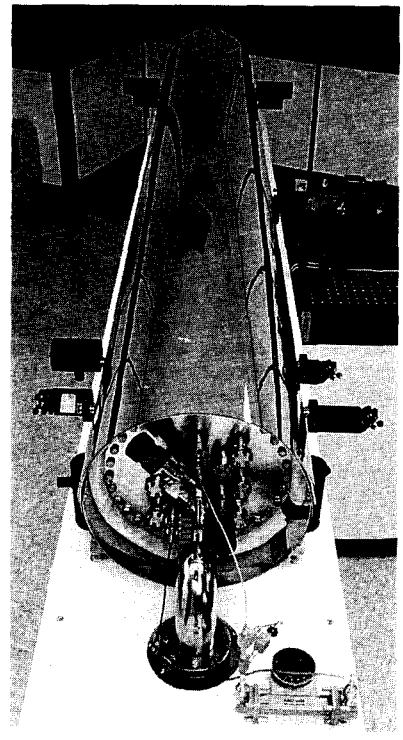


ATOMIC-BEAM frequency standards provide the most accurate, long-term timekeeping. Conventional atomic clocks rely on magnets (*a*). Atoms in the correct energy level are deflected by magnet A through the microwave cavity. Microwave fields oscillating at the resonance frequency of the atoms drive some of them into a second energy level. These atoms are deflected by magnet B so as to strike a detector. The servo mechanism monitors the detector and maintains the frequency of the applied microwaves at the resonance frequency. To keep time, some of the microwaves are



ATOMIC HYDROGEN MASER relies on a self-sustaining microwave field to serve as a frequency standard. Hydrogen atoms in the correct energy level are deflected by a magnet into a storage bulb. Some atoms will drop to a lower level, releasing a microwave photon. The photon stimulates other atoms to drop to a lower level, which produces more photons. The process quickly builds up a microwave field in the bulb. The field induces an alternating current in a wire placed in the cavity. The tuned cavity helps to redirect the photons back into the bulb to maintain the process.





directed to a device that divides the frequency into usable timing pulses. Optically pumped standards (*b*) use light rather than magnets to select atoms. Laser A pumps the atoms into the right energy level, preparing them to be excited by the microwaves. Only atoms placed in the correct energy level by the microwaves absorb light from laser B. They quickly reemit that energy, which is sensed by a photodetector. An optically pumped clock using cesium atoms at the National Institute of Standards and Technology, called NIST-7, now keeps time for the U.S. (*photograph*).

because they have the same electric charge. Hence, they do not suffer collisions with other particles or with the walls of the chamber. Ions can be trapped for long periods, sometimes for days.

Two different types of traps are used. In a Penning trap, a combination of static, nonuniform electric fields and a static, uniform magnetic field holds the ions. In a radio frequency trap (often called a Paul trap), an oscillating, nonuniform electric field does the job. Each type of trap has its own characteristic shortcoming. The strong magnetic fields of Penning traps can alter the resonance frequency. The electric field in Paul traps can create heating effects that cause Doppler shifts. The kind of trap chosen depends on its suitability for a particular experimental setup.

Workers at Hewlett-Packard, the Jet Propulsion Laboratory in Pasadena, Calif., and elsewhere have fabricated experimental standard devices using Paul traps. The particles trapped were mercury 199 ions. This ion was selected because it has the highest hyperfine frequency—40.5 gigahertz—of all the atoms that are appropriate for the trapping technique. A few million such ions are caught between the electric fields generated by electrodes. Then the ions are optically pumped by ultraviolet radiation from a lamp. Subsequent operation resembles that of the optically pumped standards, but the maximum Qs of trapped-ion standards exceed 10^{12}. This value is 10,000 times greater than that for current cesium beam clocks. Their short-term stabilities are also extremely good, although they do not yet reach those of hydrogen masers. The second-order Doppler shift limits the reproducibility to about one part per 10^{13}.

The Doppler shifts can be greatly reduced by laser cooling. In 1975 David J. Wineland, now at NIST, Hans G. Dehmelt of the University of Washington, Theodor W. Hänsch, now at the University of Munich, and Arthur L. Schawlow of Stanford University first proposed such a technique. In essence, a beam of laser light is used to reduce the velocities of the ions. Particles directed against the laser beam absorb some of the laser photon's momentum. As a result, the particles slow down. To compensate for the Doppler shifting as the particle moves against the laser, one tunes the beam to a frequency slightly lower than that produced by a strongly allowed resonance transition.

Many laboratories are developing frequency standards based on laser-cooled ions in traps. A standard based on beryllium 9 ions, laser-cooled in a Penning trap, has been constructed. Its reproducibility is about one part per 10^{13}, limited as it is by collisions of the ions with neutral molecules. Improvements in the quality of the vacuum should significantly increase the reproducibility because the uncertainty of the second-order Doppler shift is only about five parts per 10^{15}.

During the past few years, there have been spectacular developments in trapping and cooling neutral atoms, which had been more difficult to achieve than trapping ions. Particularly effective laser cooling results from the use of three pairs of oppositely directed laser-cooling beams along three mutually perpendicular paths. A moving atom is then slowed down in whatever direction it moves. This effect gives rise to the designation "optical molasses." Several investigators have contributed to this breakthrough, including William D. Phillips of NIST in Gaithersburg, Md., Claude Cohen-Tannoudji and Jean Dalibard of the École Normale Supérieure and Steven Chu of Stanford [see "Laser Trapping of Neutral Particles," by Steven Chu; SCIENTIFIC AMERICAN, February 1992].

Neutral-atom traps can store higher densities of atoms than can ion traps, because ions, being electrically charged, are kept apart by their mutual repulsion. Other things being equal, a larger number of atoms results in a higher signal-to-noise ratio.

The main hurdle in using neutral atoms as frequency standards is that the resonances of atoms in a trap are strongly affected by the laser fields. A device called the atomic fountain surmounts the difficulty. The traps capture and cool a sample of atoms that are then given a lift upward so that they move into a region free of laser light. The atoms then fall back down under the influence of gravity. On the way up

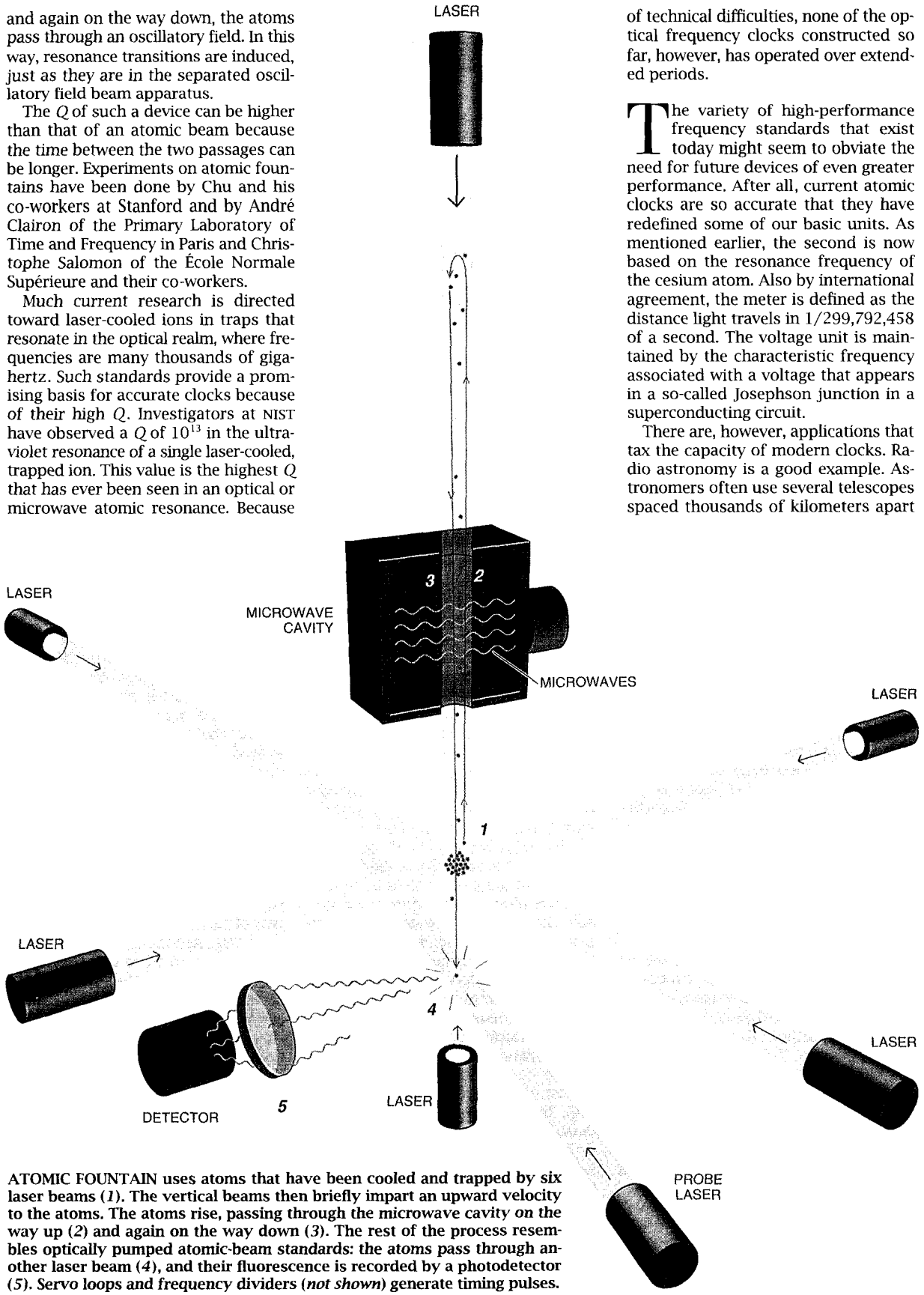
and again on the way down, the atoms pass through an oscillatory field. In this way, resonance transitions are induced, just as they are in the separated oscillatory field beam apparatus.

The Q of such a device can be higher than that of an atomic beam because the time between the two passages can be longer. Experiments on atomic fountains have been done by Chu and his co-workers at Stanford and by André Clairon of the Primary Laboratory of Time and Frequency in Paris and Christophe Salomon of the École Normale Supérieure and their co-workers.

Much current research is directed toward laser-cooled ions in traps that resonate in the optical realm, where frequencies are many thousands of gigahertz. Such standards provide a promising basis for accurate clocks because of their high Q. Investigators at NIST have observed a Q of 10^{13} in the ultraviolet resonance of a single laser-cooled, trapped ion. This value is the highest Q that has ever been seen in an optical or microwave atomic resonance. Because of technical difficulties, none of the optical frequency clocks constructed so far, however, has operated over extended periods.

The variety of high-performance frequency standards that exist today might seem to obviate the need for future devices of even greater performance. After all, current atomic clocks are so accurate that they have redefined some of our basic units. As mentioned earlier, the second is now based on the resonance frequency of the cesium atom. Also by international agreement, the meter is defined as the distance light travels in 1/299,792,458 of a second. The voltage unit is maintained by the characteristic frequency associated with a voltage that appears in a so-called Josephson junction in a superconducting circuit.

There are, however, applications that tax the capacity of modern clocks. Radio astronomy is a good example. Astronomers often use several telescopes spaced thousands of kilometers apart



ATOMIC FOUNTAIN uses atoms that have been cooled and trapped by six laser beams (*1*). The vertical beams then briefly impart an upward velocity to the atoms. The atoms rise, passing through the microwave cavity on the way up (*2*) and again on the way down (*3*). The rest of the process resembles optically pumped atomic-beam standards: the atoms pass through another laser beam (*4*), and their fluorescence is recorded by a photodetector (*5*). Servo loops and frequency dividers (*not shown*) generate timing pulses.

to study a stellar object, a technique that dramatically increases the resolution [see "Radio Astronomy by Very-Long-Baseline Interferometry," by Anthony C. S. Redhead; SCIENTIFIC AMERICAN, June 1982]. Two radio telescopes spaced 10,000 kilometers apart have an effective angular resolution more than one million times better than either telescope alone. But to combine the data from each telescope appropriately, investigators need to know precisely when each telescope received the signal. Present-day hydrogen masers have the stability required for such observations. More stable clocks may be needed for space-borne radio telescopes.

Highly stable clocks are essential for the best tests of relativity. Timing measurements of millisecond pulsars, some of which are as stable as the best atomic clocks, offer evidence for gravity waves. In 1978 Joseph H. Taylor, Jr., and his associates at Princeton University found that the period of a binary-pulsar system has been slowly varying by just the amount that would be expected for the loss of energy by gravitational radiation, as predicted by general relativity. Greater precision can be achieved if measurements are taken over many years, so clocks with better long-term stability would be useful.

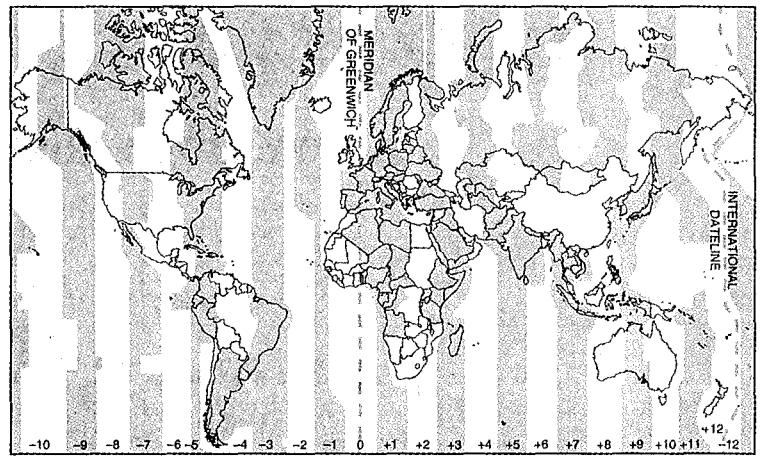
In other tests of relativity, Vessot and his colleagues confirmed the predicted increase in clock rates at high altitudes. They sent on board a rocket a hydrogen maser and measured the small, relativistic clock shift to within an accuracy of 0.007 percent at an altitude of 10,000 kilometers. Highly stable clocks have also been used by Irwin I. Shapiro, now at the Harvard-Smithsonian Center for Astrophysics, to observe the relativistic delay of a light signal passing by the sun.

Ultraprecise timekeeping has more practical applications as well—most notably, for navigation. The location of *Voyager 2* as it sped by Neptune was determined by its distance from each of three widely separated radar telescopes. Each of these distances in turn was obtained from accurate measurements of the eight hours it took for light to travel from each telescope to the spacecraft and return.

Navigation is, of course, also important on the earth. One of the latest applications of precise clocks is the satellite-based assemblage called the Global Positioning System, or GPS. This system relies on atomic clocks on board orbiting satellites. The GPS enables anyone with a suitable radio receiver and computer to determine his or her position to approximately 10 meters and the correct time to better than 10^{-7} second. Two observers monitoring the same satellite can synchronize their clocks to within a few nanoseconds.

It is expected that the GPS will have widespread practical applications, such as pinpointing the positions of ships, airplanes and even private automobiles. The GPS was used during the 1991 Persian Gulf War to enable troops to determine their positions on the desert. Commercial receivers can be purchased for less than $1,000, although these civilian versions are limited to an accuracy of about 100 meters because of deliberate scrambling of the signals transmitted from the satellites. A full complement of 24 satellites would give 24-hour, worldwide coverage. The system is nearly complete.

These and other applications show the importance of time and frequency standards. The anticipated improvements in standards will increase the effectiveness of the current uses and open the way for new functions. Only time will tell what these uses will be.



Coordinating Time Scales

In the article, we discuss the measurement of an interval of time, such as a second or a minute. This process requires only a good clock. But to be able to state that an event happened at a particular time, say, 22 seconds after 12:31 P.M. on July 5, 1993, requires synchronization with a clock that is, by mutual agreement, the standard. The world's "standard clock" exists on paper as an average of the best clocks in the world. The International Bureau of Weights and Measures in Sèvres, France, is responsible for coordinating international time. This coordinated time scale is called International Atomic Time, or TAI.

Many users require a time scale that keeps pace with the rotation of the earth. That is, averaged over a year, the sun should be at its zenith in Greenwich, England, at noon. The day as determined by the apparent position of the sun is irregular but on the average longer than the 24 hours as defined by TAI. To compensate, another time scale, called Coordinated Universal Time, or UTC, is specified by occasionally adding or subtracting a whole number of seconds from TAI. These seconds, or leap seconds, are inserted or deleted, usually on December 31 or June 30, to keep UTC within 0.9 second of the time as defined by the rotation of the earth. The record of leap seconds must be consulted to determine the exact interval between two stated times.

FURTHER READING

FROM SUNDIALS TO ATOMIC CLOCKS: UNDERSTANDING TIME AND FREQUENCY. J. Jespersen and J. Fitz-Randolph. Dover, 1982.

HISTORY OF ATOMIC CLOCKS. N. F. Ramsey in *Journal of Research of the National Bureau of Standards*, Vol. 88, No. 5, pages 301–320; September/October 1983.

PRECISE MEASUREMENT OF TIME. N. F. Ramsey in *American Scientist*, Vol. 76, No. 1, pages 42–49; January/February 1988.

TIME AND FREQUENCY. Special issue of *Proceedings of the IEEE*, Vol. 79, No. 7; July 1991.

Reprinted with permission from *Physics Today* **33** (7), 25–30, ©1980 American Institute of Physics.

The method of successive oscillatory fields

An extension of Rabi's molecular-beam resonance method, originally devised for measuring nuclear magnetic moments, is proving useful also for microwave spectroscopy, masers and lasers.

Norman F. Ramsey

In 1949 I was looking for a way to measure nuclear magnetic moments by the molecular-beam resonance method, but to do it more accurately than was possible with the arrangement developed by I. I. Rabi and his colleagues at Columbia University. The method I found[1,2] was that of separated oscillatory fields, in which the single oscillating magnetic field in the center of a Rabi device is replaced by two oscillating fields at the entrance and exit, respectively, of the space in which the nuclear magnetic moments are to be investigated. During the 1950's this method became extensively used in the original form. In the same period more general applications of the method arose, and the principal extensions included:

- Use of relative phase shifts between the two oscillatory fields[3]
- Extension generally to other resonance and spectroscopic devices,[4] such as masers, which depend on either absorption or stimulated emission
- Separation of oscillatory fields in time instead of space[4]
- Use of more than two successive oscillatory fields[4,5]
- General variation of amplitudes and phases of the successive applied oscillatory fields.[5]

Since the 1950's the method has been further extended; it has also been applied to lasers by Y. V. Baklanov, B. V. Dubetsky and V. B. Chebotsev,[6] by James C. Bergquist, S. A. Lee and John L. Hall,[7] by Michael M. Salour and C. Cohen-Tannoudji,[8] by C. J. Bordé,[9] by Theodore Hänsch[10] and by others.[11]

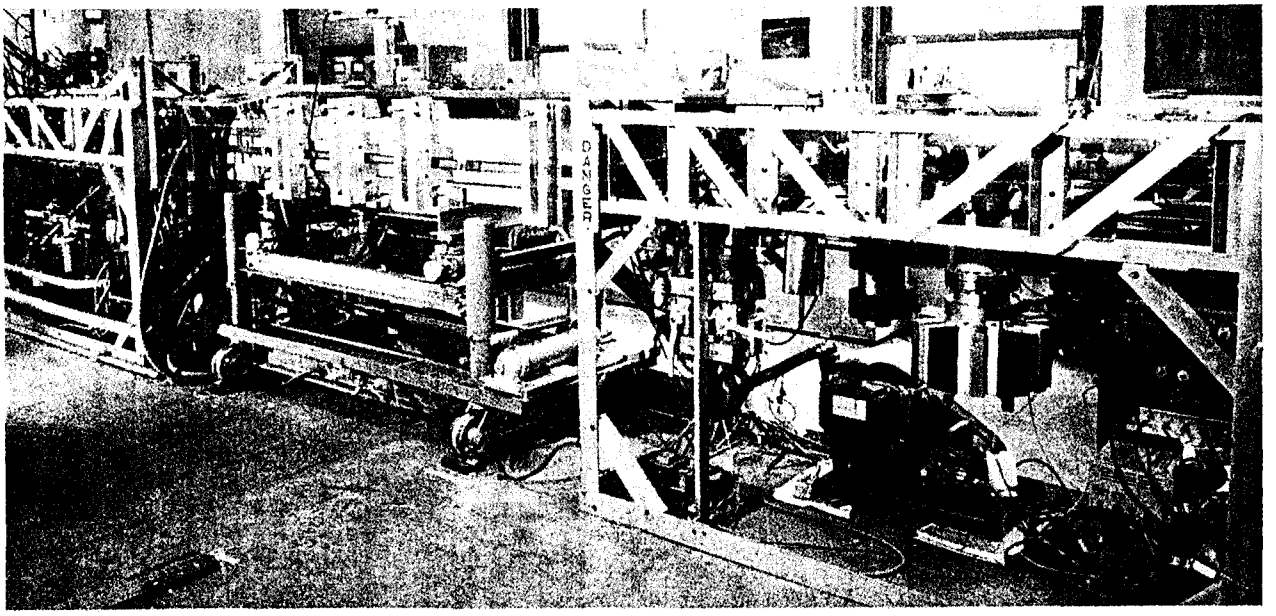
The device shown in figure 1 is a molecular-beam apparatus embodying successive oscillatory fields that has been used at Harvard for an extensive series of experiments in recent years.

Let me now review the successive oscillating field method, particularly as it is applied to molecular beams, microwave spectroscopy and masers.

The method

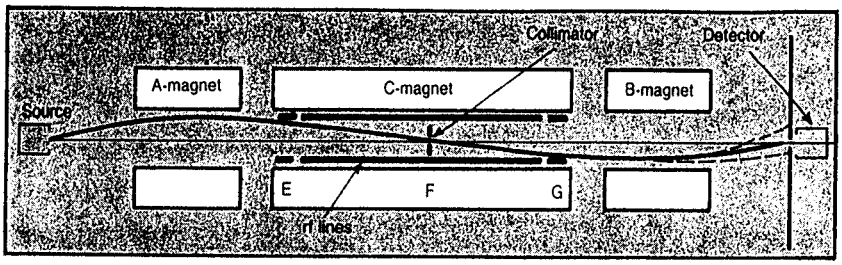
The simplicity of the original application—measurement of nuclear mag-

Norman F. Ramsey is Higgins Professor of Physics at Harvard University.

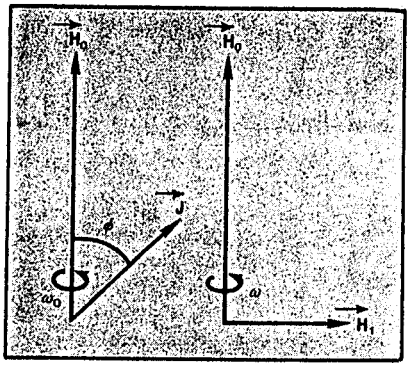


Molecular-beam apparatus with successive oscillatory fields. The beam of molecules emerges from a small source aperture in the left third of the apparatus, is focussed there and passes through the middle third in an approximately parallel beam. It is focussed again in the right third to a small detection aperture. The separated oscillatory electric fields at the beginning and end of the middle third of the apparatus lead to resonance transitions that reduce the focussing and therefore weaken the detected beam intensity. Figure 1

0031-9228 / 80 / 0725-06/ $00.50



Molecular-beam magnetic resonance. A typical detected molecule emerges from the source, is deflected by the inhomogeneous magnetic field A, passes through the collimator and is deflected to the detector by the inhomogeneous magnetic field B. If, however, the oscillatory field in the C region induces a change in the molecular state, the B magnet will provide a different deflection and the beam will follow the dashed lines with a corresponding reduction in detected intensity. In the Rabi method the oscillatory field is applied uniformly throughout the C region as indicated by the long rf lines F, whereas in the separated oscillatory field method the rf is applied only in the regions E and G. Figure 2



Precession of the nuclear angular momentum **J** (left) and the rotating magnetic field H_1 (right) in the Rabi method. Figure 3

netic moments—provides one of the best ways to explain the method of separated oscillatory fields, so my discussion will be, at first, in terms of this simple model. Extension to the more general cases is then straightforward.

First let us remember, in outline, Rabi's molecular-beam resonance method (figure 2). Consider classically a nucleus with spin angular momentum $\hbar\mathbf{J}$ and magnetic moment $\boldsymbol{\mu} = (\mu/J)\mathbf{J}$. Then in a static magnetic field $\mathbf{H}_0 = H_0\hat{\mathbf{k}}$ the nucleus, due to the torque on the nuclear angular momentum, will precess like a top about $\mathbf{H}_0$ with the Larmor angular frequency

$$\omega_0 = \frac{\mu H_0}{\hbar J} \qquad (1)$$

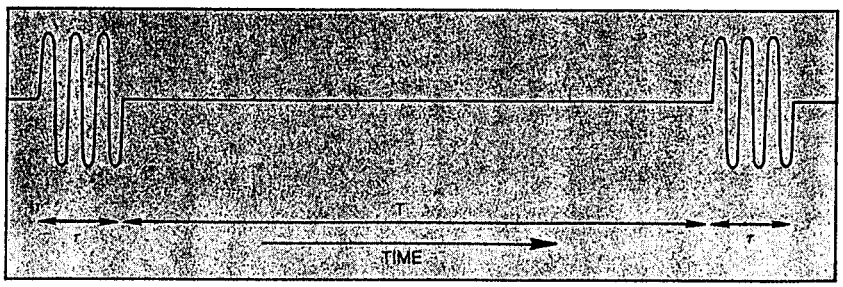
as shown in figure 3. Consider an additional magnetic field $\mathbf{H}_1$ perpendicular to $\mathbf{H}_0$ and rotating about it with angular frequency ω. Then, if at any time $\mathbf{H}$ is perpendicular to the plane of $\mathbf{H}_0$ and $\mathbf{J}$, it will remain perpendicular to it provided $\omega = \omega_0$. In that case $\mathbf{J}$ will also precess about $\mathbf{H}_1$ and the angle ϕ will continuously change in a fashion analogous to the motion of a "sleeping top"; the change of orientation can be detected by allowing the molecular beam containing the magnetic moments to pass through inhomogeneous fields as in figure 2. If ω is not equal to ω_0, $\mathbf{H}$ will not remain perpendicular to $\mathbf{J}$; so ϕ will increase for a short while and then decrease, leading to no net change. In this fashion the Larmor precession frequency ω_0 can be detected by measuring the oscillator frequency ω at which there is maximum reorientation of the angular momentum. This procedure is the basis of the Rabi molecular-beam resonance method.

The separated oscillatory field method is much the same except that the rotating field $\mathbf{H}_1$ seen by the nucleus is applied initially for a short time τ, the amplitude of $\mathbf{H}_1$ is then reduced to zero for a relatively long time T and then increased to $\mathbf{H}_1$ for a time τ, with phase coherency being preserved for the oscillating fields as shown in figure 4. This can be done, for example, in a molecular-beam apparatus in which the molecules first pass through a rotating field region, then a region with no rotating field and finally a region with a second rotating field driven phase coherently by the same oscillator.

If the angular momentum is initially parallel to the fixed field (so that ϕ is equal to zero initially) it is possible to select the magnitude of the rotating field so that ϕ is $\pi/2$ radians at the end of the first oscillating region. While in the region with no oscillating field, the magnetic moment simply precesses with the Larmor frequency appropriate to the magnetic field in that region. When the magnetic moment enters the second oscillating field region there is again a torque acting to change ϕ. If the frequency of the rotating field is exactly the same as the mean Larmor frequency in the intermediate region there is no relative phase shift between the angular momentum and the rotating field.

Consequently, if the magnitude of the second rotating field and the length of time of its application are equal to those of the first region, the second rotating field has just the same effect as the first one—that is, it increases ϕ by another $\pi/2$, making $\phi = \pi$, corresponding to a complete reversal of the direction of the angular momentum. On the other hand, if the field and the Larmor frequencies are slightly different, so that the relative phase angle between the rotating field vector and the precessing angular momentum is changed by π while the system is passing through the intermediate region, the second oscillating field has just the opposite effect to the first one; the result is that ϕ is returned to zero. If the Larmor frequency and the rotating field frequency differ by just such an amount that the relative phase shift in the intermediate region is exactly an integral multiple of 2π, ϕ will again be left at π just as at exact resonance.

In other words if all molecules have the same velocity the transition probability would be periodic as in figure 5. However, in a molecular beam resonance experiment one can easily distinguish between exact resonance and the other cases. In the case of exact resonance the condition for no change in the relative phase of the rotating field and of the precessing angular momentum is independent of the molecular velocity. In the other cases, however, the condition for integral multiple of 2π relative phase shift is velocity dependent, because a slower molecule is in the intermediate region longer and so experiences a greater shift than a faster molecule. Consequently, for the molecular beam as a whole, the reorientations are incomplete in all except



Two separated oscillatory fields, each acting for a time τ, with zero field acting for a time T. Phase coherency is preserved between the two oscillatory fields. Figure 4

the resonance cases. Therefore, one would expect a resonance curve similar to that shown in figure 6, in which the transition probability for a particle of spin $^1/_2$ is plotted as a function of frequency.

From a quantum-mechanical point of view, the oscillating character of the transition probability in figures 5 and 6 is the result of the cross term $C_{iif}C_{iff}^*$ in the calculation of the transition probability; C_{iff} is the probability amplitude for the nucleus to pass through the first oscillatory field region with the initial state i unchanged but for there to be a transition to state f in the final field, whereas C_{iff} is the amplitude for the alternative path with the transition to the final state f being in the first field with no change in the second. The interference pattern of these cross terms for alternative paths gives the narrow oscillatory pattern of the transition probability curves in figures 5 and 6. Alternatively the pattern can in part be interpreted in terms of the Fourier analysis of an oscillating field, such as that of figure 4, which is on for a time τ, off for T and on again for τ. However, the Fourier-analysis interpretation is not fully valid since with finite rotations of **J**, the problem is a non-linear one.

For the general two-level system, the transition probability P for the system to undergo a transition from state i to f can be calculated quantum mechanically in the vicinity of resonance and is given[2] by

$$P = 4\sin^2\theta \sin^2{}_{\frac{1}{2}}a\tau\,[\cos{}_{\frac{1}{2}}\lambda T \cos{}_{\frac{1}{2}}a\tau - \cos\theta \sin{}_{\frac{1}{2}}\lambda T \sin{}_{\frac{1}{2}}a\tau]^2 \quad (2)$$

where

$$\cos\theta = (\omega_0 - \omega)/a, \quad \sin\theta = 2b/a \quad (3)$$

$$a = [(\omega_0 - \omega)^2 + (2b)^2]^{1/2}$$

$$\omega_0 = (W_i - W_f)/\hbar \quad (4)$$

$$\lambda = [(\overline{W}_i - \overline{W}_f)/\hbar] - \omega \quad (5)$$

and where W_i is the energy of the initial state i, W_f of the final state f, and b is the amplitude of the perturbation V_{if} inducing the transition as defined[2] by

$$V_{if} = \hbar b e^{i\omega t} \quad (6)$$

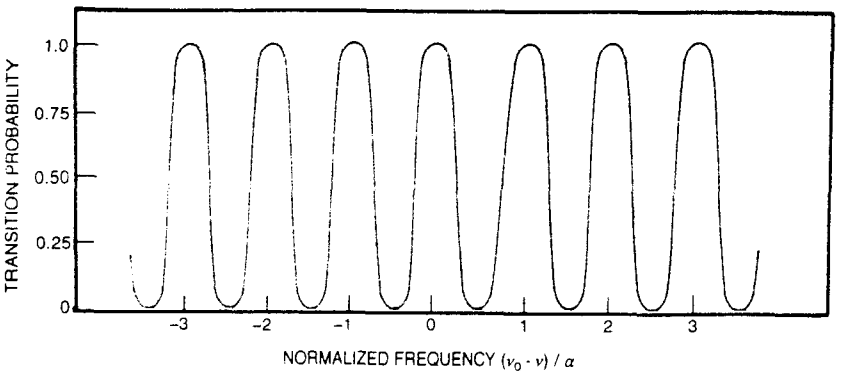
and the averages in $(\overline{W}_i - \overline{W}_f)$ are over the intermediate region. For the special case of spin $^1/_2$ and a nuclear magnetic moment, ω_0 is given in equation 1 and

$$\lambda = \overline{\omega}_0 - \omega, \; 2b = \omega_0 H_1/H_0 \quad (7)$$

The results for spin $^1/_2$ can be extended to higher spins by the Majorana formula.[12]

If equation 2 is averaged over a Maxwellian velocity distribution the result in the vicinity of the sharp resonance is given in figure 6.

The separated oscillatory field method possesses a number of advantages, which we will discuss next.



Transition probability that would be observed in a separated oscillatory field experiment if all the molecules in the beam had a single velocity. Figure 5

Advantages

Among the benefits are:

▶ The resonance peaks are only 0.6 times as broad as the corresponding ones with the Rabi method and the same length of apparatus. This narrowing corresponds to the peaks in a two-slit optical interference pattern being narrower than the central diffraction peak of a single wide slit aperture whose width is equal to the separation of the slits.

▶ The sharpness of the resonance is not reduced by non-uniformities of the constant field since from both the qualitative description and from equations 2 and 5 it is only the space-average value of the energies along the beam path that is important. This advantage often permits increases in precision by a factor of 20 or more.

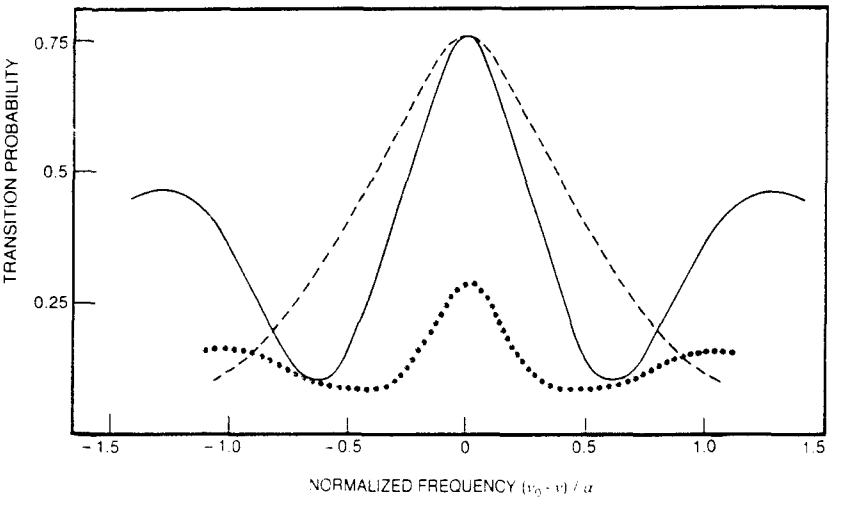
▶ The method is often more convenient and effective at very high frequencies where the wavelength may be comparable to the length of the region in which the energy levels are studied.

▶ Provided there is no unintended phase shift between the two oscillatory fields, first-order Doppler shifts are eliminated.

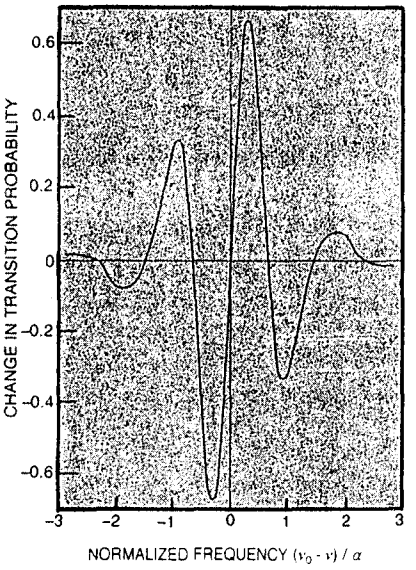
▶ The method may be applied to study energy levels in a region into which an oscillating field cannot be introduced; for example, the Larmor precession of neutrons can be measured while they are inside a magnetized iron block.

▶ The lines can be narrowed by reducing the amplitude of the rotating field below the optimum, as shown by the colored curve in figure 6. The narrowing is the result of the low amplitude favoring slower-than-average molecules.

▶ If the atomic state being studied decays spontaneously, the separated oscillatory field method permits the observation of narrower resonances



When the molecules have a Maxwellian velocity distribution, the transition probability is as shown by the black line for optimum rotating field amplitude. (L is the distance between oscillating field regions, α is the most probable molecular velocity and ν is the oscillator frequency.) The colored line shows the transition probability for an oscillating field of one-third strength, and the dashed line represents the single oscillating field method when the total duration is the same as the time between separated oscillatory field pulses. Figure 6



Theoretical change in transition probability on reversing a $\pi/2$ phase shift. This resonance shape gives the maximum sensitivity by which to detect small shifts in the resonance frequency. Figure 7

than those anticipated from the lifetime and the Heisenberg uncertainty principle provided the two separated oscillatory fields are sufficiently far apart; only states that survive sufficiently long to reach the second oscillatory field can contribute to the resonance. This method, for example, has been used effectively by Francis Pipkin, Paul Jessop and Stephen Lundeen[13] in studies of the Lamb shift.

These advantages have led to the extensive use of the method in molecular and atomic-beam spectroscopy. One of the best-known of these uses is in atomic cesium standards of frequency and time.

As in any high-precision experiment, care must be exercised with the separated oscillatory field method to avoid obtaining misleading results. Ordinarily these potential distortions are more easily understood and eliminated with the separated oscillatory field method than are their counterparts in most other high-precision spectroscopy. Nevertheless the effects are important and require care in high-precision measurements. I have discussed the various effects in detail elsewhere[13,14,15] but I will summarize them here.

Advice from Rabi

The first advice I received from Rabi in 1937 when I applied to him to begin my research was that I should not go into the field of molecular beams since the interesting problems amenable to that technique had already been solved and there was little future to the field. I have often wondered how I . . . had the temerity to disregard this bit of advice from the master. However, I am grateful that I did since the advice was given only a few months before Rabi's great invention of the molecular beam resonance method . . . which led to such fundamental discoveries as the quadrupole moment of the deuteron, the Lamb shift, the anomalous magnetic moment of the electron and [to] numerous other discoveries and measurements.

(Quoted from Norman Ramsey's contribution to A Tribute to Professor I. I. Rabi, *Columbia University, New York, 1970.)*

Precautions

Variations in the amplitudes of the oscillating fields from their optimum values may markedly change the shape of the resonance, including the replacement of a maximum transition probability by a minimum. However, symmetry about the exact resonance frequency is preserved, so no measurement error need be introduced by such amplitude variations.[14,15]

Variations of the magnitude of the fixed field between (but not in) the oscillatory field regions do not ordinarily distort a molecular beam resonance provided the *average* transition frequency (Bohr frequency) between the two fields equals the values of the transition frequencies in each of the two oscillatory field regions alone. If this condition is not met, there can be some shift in the resonance frequency.[14,15]

If, in addition to the two energy levels between which transitions are studied, there are other energy levels partially excited by the oscillatory field, there will be a pulling of the resonance frequency as in any spectroscopic study and as analyzed in detail in the literature.[13,14,15]

Even in the case when only two energy levels are involved, the application of additional rotating magnetic fields at frequencies other than the resonance frequency will produce a net shift in the observed resonance frequency, as discussed elsewhere.[13,14,15] A particularly important special case is the effect identified by Felix Bloch and Arnold Siegert,[16] which occurs when oscillatory rather than rotating magnetic fields are used. Since an oscillatory field can be decomposed into two oppositely rotating fields, the counter-rotating field component automatically acts as such an extraneous rotating field. The theory of the Bloch–Siegert effect in the case of the separated oscillatory field method has been developed by Ramsey,[17,19] J. H. Shirley,[18] R. Fraser Code[19] and Geoffrey Greene.[20] Another example of an extraneously introduced oscillating field is that which results from the motion of an atom through a field $\mathbf{H}_0$ whose direction varies in the region traversed.

Unintended relative phase shifts between the two oscillatory field regions will produce a shift in the observed resonance frequency.[13,14,15] This is the most common source of possible error, and care must be taken to avoid it either by eliminating such a phase shift or by determining the shift—say by measurements with the molecular beam passing through the apparatus first in one direction and then in the opposite direction.

In some instances of closely overlapping spectral lines the subsidiary maxima associated with the separated oscillatory field method can cause confusion. Sometimes this problem can be eased by using three or four separated oscillatory fields as we shall see, but ordinarily when there is extensive overlap of resonances and the resultant envelope is to be studied, it is best to use a single oscillatory field.

Extensions of the method

Although, as we discussed above, unintended phase shifts can cause distortions of the observed resonance, it is often convenient to introduce phase shifts deliberately to modify the resonance shape.[21] Thus, if the change in transition probability is observed when the relative phase is shifted from $+\pi/2$ to $-\pi/2$ one sees a dispersion curve shape[21] as in figure 7. A resonance with the shape of figure 7 provides maximum sensitivity for detecting small shifts in the resonance frequency, say by applying an external electric field.

For most purposes the highest precision can be obtained with just two oscillatory fields separated by the maximum time, but in some cases it is better to use more than two separated oscillatory fields.[4] The theoretical resonance shapes with two, three, four and infinitely many oscillatory fields are given in figure 8. The infinitely many oscillatory field case, of course, by definition becomes the same as the single long oscillatory field if the total length of the transition region is kept the same and the infinitely many oscillatory fields fill in the transition region continuously as we assumed in figure 8. For many purposes this is the best way to think of the single oscillatory field method, and this point of view makes it apparent that the single oscillatory field method is subjected to complicated versions of all the distortions discussed in the previous section. It is noteworthy that, as the number of oscillatory field regions is increased for the same total length of apparatus, the resonance width is broadened; the narrowest resonance is obtained with just two oscillatory fields separated the maximum distance apart. Despite this advantage, there are valid circumstances for using more than two oscilla-

tory fields. With three oscillatory fields the first and largest side lobe is suppressed, which may help in resolving two nearby resonances; for a larger number of oscillatory fields additional side lobes are suppressed, and in the limiting case of a single oscillatory field there are no sidelobes. Another reason for using a large number of successive pulses can be the impossibility of obtaining sufficient power in a single pulse to induce adequate transition probability with a small number of pulses.

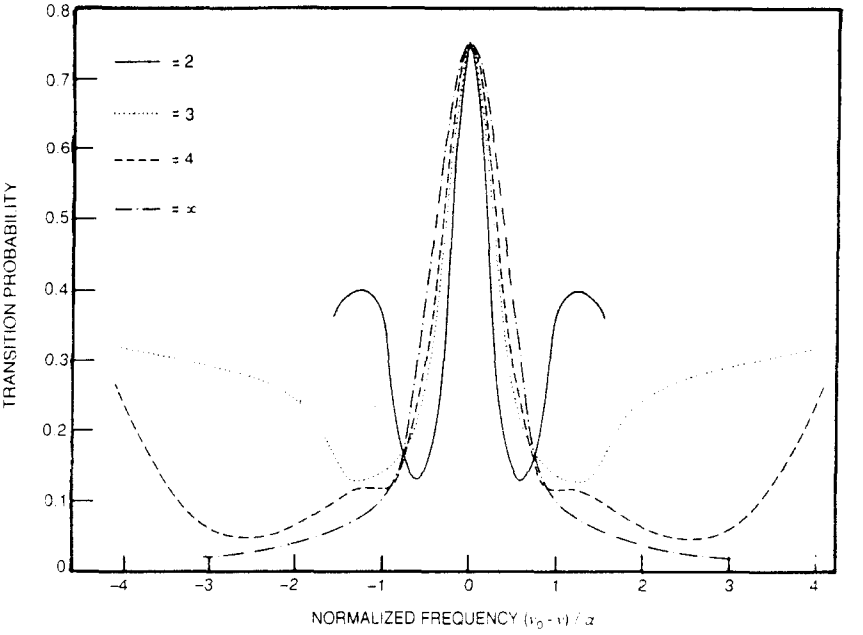
The earliest use of the successive oscillatory field method involved two oscillatory fields separated in space, but it was early realized that the method with modest modifications could be applied to the separations being in time, say by the use of coherent pulses.[4]

If more than two successive oscillatory fields are utilized it is not necessary to the success of the method that they be equally spaced in time;[4] the only requirement is that the oscillating fields be coherent— as is the case if the oscillatory fields are all derived from a single continuously running oscillator. In particular, the separation of the pulses can even be random,[4] as in the case of the large box hydrogen maser[22] shown in figure 9. The atoms being stimulated to emit move randomly into and out of the cavities with oscillatory fields and spend the intermediate time in the large container with no such fields.

The full generalization of the successive oscillatory field method is excitation by one or more oscillatory fields that vary arbitrarily with time in both amplitude and phase.[5]

More recent modifications of the successive oscillatory field method include the following three applications.

V. F. Ezhov and his colleagues,[23] in a neutron-beam experiment, used an inhomogeneous static field in the region of each oscillatory field region such that initially when the oscillatory field is applied conditions are far from resonance. Then, when the resonance condition is slowly approached, the magnetic moment that was originally aligned parallel to $\mathbf{H}_0$ will adiabatically follow the effective magnetic field on a coordinate system rotating with $\mathbf{H}_1$ until at the end of the first oscillatory field region the moment is parallel to $\mathbf{H}_1$ and thereby at angle $\phi = \pi/2$. As in the ordinary separated oscillatory field experiment, the moment precesses in the intermediate region, and the phase of the precession relative to that of the oscillator is detected in the final region by an arrangement where the fields are such that the effective field in the rotating coordinate system shifts from parallel to $\mathbf{H}_1$ to parallel to $\mathbf{H}_0$. This arrangement has the theoretical advantage that the maximum transition

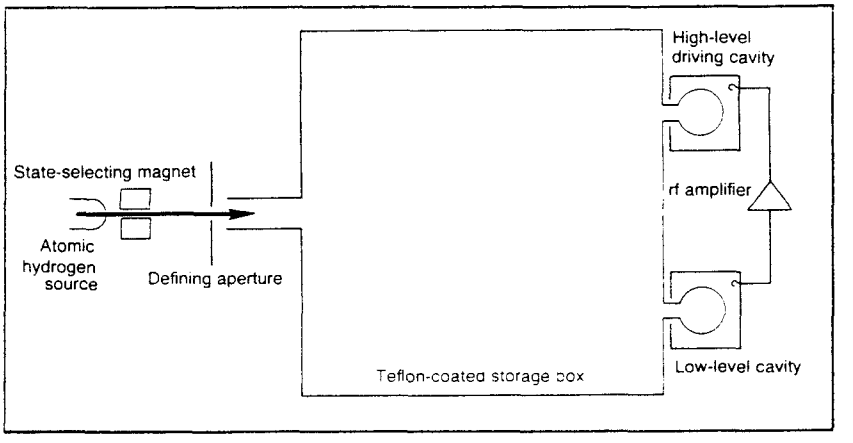


Multiple oscillatory fields. The curves show molecular-beam resonances with two, three, four and infinitely many successive oscillating fields. The case with an infinite number of fields is equivalent to a single oscillating field extending throughout the entire region. Figure 8

probability can be unity even with a velocity distribution, and the resonance is less dependent on the value of the field in the oscillatory field region. On the other hand, the method may be less well adapted to the study of complicated spectra, and it requires a special nonuniform magnetic field.

I emphasized in an earlier section that one of the principal sources of error in the separated oscillatory field method is that which arises from uncertainty in the exact value of the relative phase shift in the two oscillatory fields. David Wineland, Helmut Hellwig and Stephen Jarvis[24] have pointed out that this problem can be overcome with a slight loss in resolution by driving the two cavities at slightly different frequencies so that there is a continual change in the relative phase. In this case the observed resonance pattern will change continuously from absorption to dispersion shape. The envelope of these patterns, however, can be observed and the position of the maximum of the envelope is unaffected by relative phase shifts. Since the envelope is about twice the width of a specific resonance there is some loss of resolution in this method, but in certain cases this loss may be outweighed by the freedom from phase-shift errors.

Finally, the extension of the method to laser spectroscopy has aroused a great deal of interest. These applications of course require some modifica-



A large box hydrogen maser shown schematically. The two cavities on the right constitute two separated oscillatory fields; atoms move randomly in and out of these cavities but spend the greater part of their time in the large field-free container. Figure 9

tions from the form of the technique most widely applied to molecular beams,[6–10] due to differences in wavelengths, state lifetimes and so on. However, the necessary modifications have been made and the method continues to be of real value in laser spectroscopy—particularly in cases where the extra effort required is justified by the need for multiple pulses or high precision.

* * *

This article is an adaptation of a paper presented at the Tenth International Quantum Electronics Conference, held in Atlanta, Georgia, in May 1978.

References

1. N. F. Ramsey, Phys. Rev. **76**, 996 (1949).
2. N. F. Ramsey, Phys. Rev. **78**, 695 (1950).
3. N. F. Ramsey, H. B. Silsbee, Phys. Rev **84**, 506 (1951).
4. N. F. Ramsey, Rev. Sci. Instr. **28**, 57 (1957).
5. N. F. Ramsey, Phys. Rev. **109**, 822 (1958).
6. Y. V. Baklanov, B. V. Dubetsky, V. B. Chebotsev, Appl. Phys. **9**, 171 (1976) and **11**, 201 (1976).
7. J. C. Bergquist, S. A. Lee, J. L. Hall, Phys. Rev. Lett. **38**, 159 (1977) and Laser Spectroscopy **III**, 142 (1978).
8. M. M. Salour, C. Cohen-Tannoudji, Phys. Rev. Lett. **38**, 757 (1977), Laser Spectroscopy **III**, 135 (1978), Appl. Phys. **15**,119 (1978) and Phys. Rev. **A17**, 614 (1978).
9. C. J. Bordé, C. R. Acad. Sci. Paris **284B**, 101 (1977).
10. T. W. Hänsch, Laser Spectroscopy **III**, 149 (1978).
11. V.P Chebotayev, A.V. Shishayev, B.Y. Yurshin, L.S. Vasilenko, N.M. Dyuba, M.I. Skortsov, Appl. Phys. **15**, 43, 219 and 319 (1987).
12. S. R. Lundeen, P. E. Jessop, F. M. Pipkin, Phys. Rev. Lett. **34**, 377 and 1368 (1975).
13. N. F. Ramsey, *Molecular Beams*, Oxford University Press (1956).
14. N. F. Ramsey, Le Journal de Physique et Radium **19**, 809 (1958).
15. N. F. Ramsey, in *Recent Research in Molecular Beams* (I. Estermann, ed.) Academic Press, New York (1958); page 107.
16. F. Bloch, A. Siegert, Phys. Rev. **57**, 522 (1940).
17. N. F. Ramsey, Phys. Rev. **100**, 1191 (1955).
18. J. H. Shirley, J. Appl. Phys. **34**, 783 (1963).
19. R. F. Code, N. F. Ramsey, Phys. Rev. **A4**, 1945 (1971).
20. G. Greene, Phys. Rev. **A18**, 1057 (1978).
21. N. F. Ramsey, H. B. Silsbee, Phys. Rev. **84**, 506 (1951).
22. E. E. Uzgiris, N. F. Ramsey, Phys. Rev. **A1**, 429 (1970).
23. V. F. Ezhov, S. N. Ivanov, I. M. Lobashov, V. A. Nazarenko, G. D. Porsev, A. P. Serebrov, R. R. Toldaev, Sov. Phys.-JETP **24**, 39 (1976).
24. S. Jarvis Jr., D. J. Wineland, H. Hellwig, J. Appl. Phys. **48**, 5336 (1977). □

26 October 1984, Volume 226, Number 4673 **SCIENCE**

Trapped Ions, Laser Cooling, and Better Clocks

D. J. Wineland

Summary. Ions that are stored in electromagnetic "traps" provide the basis for extremely high resolution spectroscopy. By using lasers, the kinetic energy of the ions can be cooled to millikelvin temperatures, thereby suppressing Doppler frequency shifts. Potential accuracies of frequency standards and clocks based on such experiments are anticipated to be better than one part in 10^{15}.

In a recent experiment (*1*) at the National Bureau of Standards (NBS) in Boulder, Colorado, the frequency of a particular hyperfine transition in the ground state of beryllium atomic ions was measured with an inaccuracy of only about one part in 10^{13}. In this experiment, the ions were confined or "trapped" in a small region of space by using static electromagnetic fields and their kinetic temperature was lowered to less than 1 K by a process sometimes called "laser cooling." In all of physics, only a few measurements can boast a higher accuracy; those experiments measure similar transitions in neutral cesium atoms.

Experimental spectroscopy (the study of the interaction between radiation and matter) has traditionally provided a means of checking the theory of quantum mechanics which predicts the internal energy structure of atoms and molecules. In the case of hyperfine structure, which represents the magnetic coupling between the nucleus and atomic electrons, the theory quickly becomes very complicated. For alkali-like ions, the most sophisticated calculations (*2*) agree with experiment at a level of only a few tenths of a percent. On the experimental side, about 8 years ago a measurement (*3*) of beryllium ion hyperfine structure to an accuracy of 3 ppm (parts per million) was made. Since the accuracy of this previous measurement is presumably good enough to satisfy the theorists for quite some time in the future, one can logically ask why anyone would want to make a better measurement. Briefly, I will give three of the reasons.

1) The primary driver in several laboratories (including NBS) is to provide better clocks and frequency standards. The principal use of atomic clocks is in navigation and communications, where requirements have continued to press the state of the art. The way an atomic clock works is perhaps apparent in the internationally agreed on definition of the second: "The second is the duration of 9,192,631,770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium-133 atom." A simplified model for the practical realization of the second involves making a device (an atomic beam apparatus) which allows one to induce and detect transitions between the two ground-state hyperfine energy levels. When the frequency of the radiation for maximum transition probability is attained, the cycles are electronically counted; when 9,192,631,770 cycles have occurred, 1 second has passed. Several laboratories are trying to apply this same idea to the internal energy levels of ions; there is reason to believe that the inaccuracy of a time standard based on stored ions can eventually be much smaller than that of the cesium clock, which can have an accuracy of about one part in 10^{13} or less (*4*).

2) With the extreme accuracy attained with stored ion techniques, it may be possible to measure various small effects which would otherwise be masked by measurement imprecision. As an example, it should be possible to measure nuclear magnetic susceptibility as a small perturbation to atomic hyperfine structure (*5*). Nuclear magnetic susceptibility, which has not been measured previously, could give a new kind of information about nuclear structure. Trapped ions may provide a way to make the required precise measurements.

3) The system itself, laser-cooled stored ions, is intrinsically interesting and may provide the basis for other experiments which are only peripherally related to spectroscopy. An example is the study of strongly coupled three-dimensional plasmas (*6*).

For brevity, this article will only touch on some aspects of stored ion spectroscopy where laser cooling is employed and how they are related to better clocks and frequency standards. A recent review article (*7*) includes many interesting trapped ion experiments which are not discussed here.

Trapped Ions

The principal attraction of the stored ion technique is that charged particles, including electrons and atomic ions, can be stored for long periods of time (days are not uncommon) without the usual perturbations associated with confinement [for example, the frequency shifts associated with the collisions of ions with buffer gases in a more traditional optical pumping experiment (*8*)].

Storage has principally been accomplished in four types of "traps": the RF (radio frequency) or Paul trap, the Pen-

The author is a research physicist in the Time and Frequency Division of the National Bureau of Standards, Boulder, Colorado 80303.

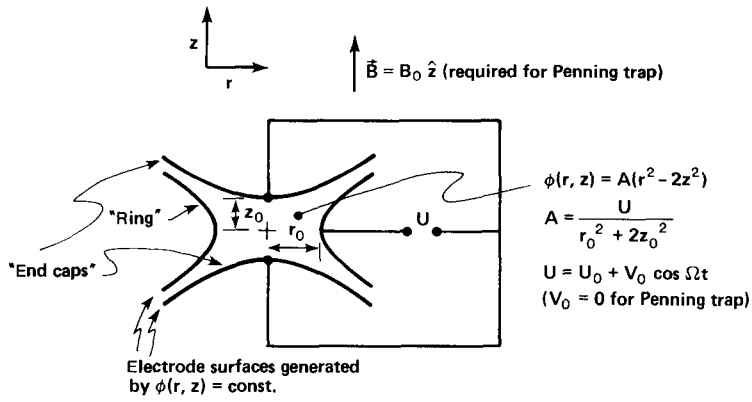


Fig. 1. Schematic representation of the electrode configuration for the "ideal" Paul (RF) or Penning trap. Electrode surfaces are figures of revolution about the z axis and are equipotentials of $\phi(r,z) = A(r^2 - 2z^2)$. (Cylindrical coordinates are used with the origin at the center of the trap.) Typical dimensions are $\sqrt{2}\ z_0 = r_0 \cong 1$ cm. Typical operating parameters are: for the Paul trap, $V_0 \cong 300$ V/cm, $\Omega/2\pi \cong 1$ MHz; for the Penning trap, $U_0 \cong 1$ V, $B \cong 1$ T.

ning trap, the Kingdon (electrostatic) trap, and the magnetostatic trap ("magnetic bottle"). Magnetic bottles have had limited use in high-resolution work because the trapping relies on spatially inhomogeneous magnetic fields, which can cause shifts and broadening of magnetic field–dependent lines. A notable exception is the electron g factor measurements at the University of Michigan by H. R. Crane, A. Rich, and their colleagues (*9*). The Kingdon trap (*10*) is perhaps the simplest, using only static electric fields for trapping. Since an electrostatic potential minimum cannot exist in a charge-free region, the Kingdon trap relies on a dynamical equilibrium for trapping (ions orbit around an attractive wire). Kingdon traps have been used in spectroscopic experiments by Prior and his colleagues (*11*), but so far neither Kingdon traps nor magnetic bottles have been used in laser cooling experiments, and therefore they will not be discussed further here.

The Paul (*12*) or RF trap uses inhomogeneous RF electric fields to provide confinement in a pseudopotential well (*7*, *13*). It is the three-dimensional analog of the Paul quadrupole mass filter. To see how it works, we first note that in a (homogeneous) sinusoidal RF electric field, ion motion is sinusoidal but is 180° out of phase with respect to the electric force. If the field is somewhat inhomogeneous, it is easy to show that the force on the ion averaged over one cycle of the driven motion is toward the region of weaker field. Since an electric field minimum can exist in a charge-free region, stable trapping can be accomplished. Such a trap is shown schematically in Fig. 1, where the three trap electrodes are shaped to provide an electric potential of the form $(r^2 - 2z^2)$ inside the trap. For this "ideal" trap shape, an ion is bound in a nearly harmonic well; detection of the ion well frequencies can then be used to perform mass analysis.

The "ideal" Penning (*14*) trap uses the same electrode configuration as in Fig. 1 but uses static electric and magnetic fields. A harmonic potential well is provided along the z axis by static electric fields. This, however, results in a radial electric field, which forces the ions toward the "ring" electrode. This effect can be overcome if a static magnetic field **B** is superimposed along the z axis. In this case the x-y motion of the ions is a composite of circular cyclotron orbits (primarily due to the **B** field) and a circular $\mathbf{E} \times \mathbf{B}$ drift magnetron motion about the trap axis.

Both the Paul and Penning traps can provide long-term confinement. Storage times of days are not uncommon; in the first single-electron (Penning trap) experiments (*15*) the same electron was used in experiments for several weeks! This long-term storage is important in spectroscopy because (i) "transit time" broadening (the broadening of transitions due to the time-energy uncertainty relation associated with the time the ion stays in the trap) can be made negligible and (ii) the average velocity $\langle \mathbf{v} \rangle$ of the ions approaches zero. The latter is important because it can make averaged first-order Doppler frequency shifts negligible. Suppression of such first-order Doppler shifts is perhaps the chief advantage over the atomic beam method; for example, residual first-order Doppler shifts (due to the net velocity of the atomic beam) are the main limitation to accuracy for the cesium beam frequency standard. (Actually, even though $\langle \mathbf{v} \rangle \to 0$ for the traps, the first-order Doppler effect is proportional to $\langle \mathbf{k} \cdot \mathbf{v} \rangle$, where $\mathbf{k}$ is the wave vector of the radiation. For spatially inhomogeneous radiation fields, shifts can occur. If the lifetime of the ion's internal transition is long compared to the periods of ion motion, these effects result in asymmetric sidebands at the ion motional frequencies; however, the "pulling" caused by these sidebands can be extremely small.)

The perturbation of the ions' internal structure due to trapping can be extremely small. Perhaps the most troublesome frequency shifts are caused by electric fields. Shifts which are linearly dependent on the electric field are absent because the average electric field $\langle \mathbf{E} \rangle = 0$. (We know that because if $\langle \mathbf{E} \rangle \neq 0$ the ions would leave the trap.) Second-order shifts can be quite small. As an example, the shift of the ground-state hyperfine frequency of Hg^+ ions, ν_0 (Hg^+), has been calculated (*16*) as $\delta\nu/\nu_0$ (Hg^+) $\cong -1.4 \times 10^{-18} E^2$, where E is in volts per centimeter. Radio-frequency electric fields in an RF trap may be as high as 300 V/cm; this could give a fractional shift of 10^{-13}. However, for small samples of laser-cooled ions electric fields can be smaller than 1 mV/cm (*17*), yielding negligible shifts. (The restoring electric forces become less as the ion's kinetic energy is reduced.) Of course, there are also shifts associated with electric fields due to ion-ion collisions, but these are expected to be smaller than the trapping fields (*17*).

In most cases the magnetic field (**B**) of the Penning trap (typically about 1 T or 10^4 gauss) causes large frequency shifts to the ions' internal structure. The energy separation of the beryllium hyperfine transition mentioned in the introduction goes to zero at zero magnetic field (*1*); therefore, in some sense, the energy separation is entirely due to the external magnetic field. This might be regarded as a severe disadvantage, except that this transition and some others become independent of magnetic field to first order at certain magnetic fields. Second-order field dependence can be small; for the beryllium example $\Delta\nu/\nu_0 = -0.017\ (\Delta B/B)^2$, so if the field is held to 10^{-8} of the nominal field (this can be done with a superconducting magnet), the fractional shifts are only 1.7×10^{-18}. Thus, magnetic fields can strongly perturb the internal structure but these perturbations may only be a philosophical disadvantage. For the clock application we do not care about their existence if they can be made reproducible and stable; more-

over, these perturbations are physically interesting since they can sometimes be calculated with high precision.

A third feature of the traps may be regarded as either a disadvantage or advantage. Typically, the number of ions that can be trapped is rather small. Densities in the range of 10^5 to 10^7 cm^{-3} are typical; therefore the total number of ions may be quite small. Fortunately, atomic ions can be sensitively detected; this is apparent in the single-ion experiments discussed below. In many spectroscopic experiments it is anticipated that the signal-to-noise ratio can approach the theoretical limit (*18*)—that is, that it can be limited only by the statistical fluctuations in the number of ions that make the transition. We note that if we could obtain much higher densities by using larger confining fields then we would lose one of the advantages of the technique because electric field frequency shifts would become troublesome in very high resolution work.

The small sample sizes can actually be regarded as an advantage in a couple of ways. First, the small numbers imply that the ions can be confined to a small region of space—down to dimensions on the order of 1 μm or less for single ions (*19, 20*). This means that field imperfections—deviations from the quadratic electric potential or deviation from uniformity of the magnetic field in Penning traps—can be quite small over the ion sample. Therefore accuracy in magnetic field–dependent studies (mass spectroscopy, g factor measurements) can be extremely high. Second, in single-photon absorption spectra it is desirable to satisfy the Lamb-Dicke criterion (*21*)—that is, confinement to dimensions $\lesssim\lambda/2\pi$, where λ is the wavelength of the radiation. When the Lamb-Dicke criterion is satisfied, first-order Doppler effects (broadening or sideband generation) are suppressed. For optical wavelengths this condition can be met only for single confined ions.

For high-resolution spectroscopy, Paul and Penning traps have many desirable features in common, but they differ in some important respects. The magnetic field of the Penning trap may be a disadvantage in some experiments but it may also be the clear choice for magnetic field–dependent studies. The RF trap is able to provide tighter spatial confinement and may be the best choice for optical frequency standards (where it is desirable to satisfy the Lamb-Dicke criterion), although the Penning trap can nearly satisfy the Lamb-Dicke criterion on optical transitions in certain cases. Heating mechanisms in the Paul trap (due to the large RF fields) are typically more severe than in the Penning trap. This can be an important problem for large numbers of stored ions.

Laser Cooling

The experiments of Dehmelt and collaborators (*13*) in the 1960's showed the ability of the stored ion technique to obtain very high resolution in atomic spectra. Prior to 1970, the ground-state hyperfine transition in the $^3He^+$ ($\nu_0 \cong 8.7$ GHz) ion was measured in an RF trap with a line width of only 10 Hz (*22*), but the accuracy was limited by the second-order Doppler shift ($\Delta\nu/\nu_0 = -1/2\langle v^2\rangle/c^2$) to 10 Hz ($c$ = speed of light). The second-order Doppler effect is due to relativistic time dilation. Because the atoms are moving, their time proceeds slower than a laboratory observer; this effect must be accounted for. The relatively large second-order Doppler frequency shifts imposed by the high velocities of the stored ions and the difficulty of measuring their velocity distribution have historically been the main limitation to achieving high accuracy in stored ion spectra.

In 1975 proposals were made (*23, 24*) to get around this general problem of the second-order Doppler shift by a process commonly called laser (or sideband) cooling. The idea is outlined for the case of an atom with internal (optical) transition frequency ν_a having natural (radiative) line width $\Delta\nu_a$. Assume that the atom is constrained to move in a harmonic well along the z axis (one-dimensional model of the ion trap). Therefore its velocity is given by $v_z = v_0 \cos 2\pi\nu_v t$, where ν_v is its oscillation frequency in the well and we assume $\nu_v >> \Delta\nu_a$. When observed along the direction of the motion, the spectrum in the laboratory contains the central resonance line (at frequency ν_a) with sidebands generated by the first-order Doppler effect at frequencies $\nu_a + n\nu_v$ having intensity $J_n^2(v_0\nu_a/c\nu_v)$ (with n a positive or negative integer); here, J_n is the Bessel function of order n. This spectrum is a simple frequency modulation (FM) spectrum where the frequency modulation is supplied by the first-order Doppler effect (*21*). If we irradiate the atom with photons of frequency $\nu_L = \nu_a + n\nu_v$, the frequencies of the resonantly scattered photons occur at ν_a and nearly symmetrically around ν_a at the sideband frequencies $\nu_a \pm \nu_v$, $\nu_a \pm 2\nu_v$, ... Therefore, although photons of energy $h(\nu_a + n\nu_v)$ are absorbed, on the average photons of energy $h\nu_a$ are reemitted; when n is negative, this energy difference causes the kinetic energy of the atom to decrease by $nh\nu_v$ per scattering event.

This explanation in terms of sidebands (*24*) is easily visualized for $\nu_v >> \Delta\nu_a$. For all experiments done so far $\Delta\nu_a >> \nu_v$; however, the above conclusion is still valid. An alternative explanation (*23*) for this limit (note that when $\nu_v \to 0$ the atoms are free) is that when $\nu_L < \nu_a$, the atom predominantly interacts with the incident radiation when it moves toward the source of radiation and Doppler shifts the frequency into resonance—that is, when $\nu_L(1 + v_z/c) = \nu_a$. In the absorption process, the photon momentum is first transferred to the atom, causing its momentum to change by h/λ, where h is Planck's constant. Since the remission occurs nearly symmetrically in the $\pm z$ directions, the net effect is to change the velocity of the atom by $\Delta v_z = h/M\lambda$ (M = mass of the atom). If $\Delta v_z << v_z$, then the kinetic energy of the absorber decreases by an amount $Mv_z\Delta v_z = nh\nu_v$. The cooling process is weak in that it takes about 10^4 scattering events to do substantial cooling below room temperature; it is strong in that, for allowed transitions, the resonant scattering rate can approach 10^8 sec^{-1}. Clearly, however, we require a situation where a two-level system is approached in order to have 10^4 scattering cycles; hence we usually think in terms of simple atomic systems.

Laser cooling was first observed in 1978 in experiments at both NBS (Boulder) and Heidelberg University. In the NBS experiments (*25*), the temperature of Mg^+ ions was monitored by observing the induced currents (*26*) in the trap electrodes of a Penning trap. At magnetic fields of about 1 T a two-level system is formed in $^{24}Mg^+$ (the most abundant isotope) by driving the $3s^2S_{1/2}$ ($M_J = -1/2$) $\leftrightarrow$ $3p^2P_{3/2}$ ($M_J = -3/2$) transition with linearly polarized light (see inset in Fig. 2). From the selection rule $\Delta M_J = 0, \pm 1$, the ion must fall back to the original ground state. In addition, because other allowed transitions are driven weakly, about 16/17 of the ions are pumped into the $M_J = -1/2$ ground state (*27*). In these experiments, cooling to ≅40 K was observed and was limited by the noise in the induced current detection.

In the first Heidelberg experiments (*28*), cooling was observed through the increased storage time of Ba^+ ions in a miniature Paul trap. (Separation between the end caps $2z_0 \cong 0.5$ mm.) The line

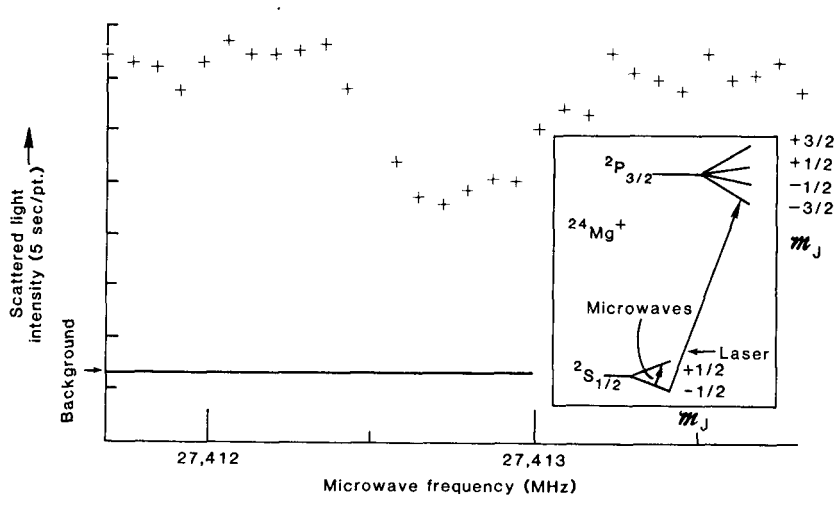
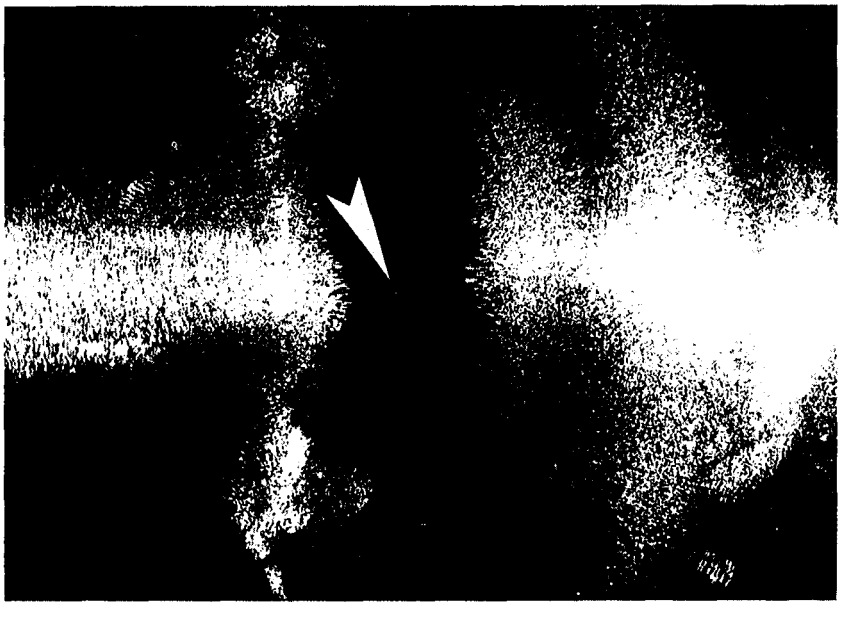


Fig. 2. Microwave/optical double resonance spectrum of $^{24}Mg^+$. Inset shows relevant energy levels of $^{24}Mg^+$ in a magnetic field. With the laser tuned to the transition shown, the ions are pumped into the $^2S_{1/2}$ ($M_J = -1/2$) ground state and a two-level system is formed with this ground state and the excited $^2P_{3/2}$ ($M_J = -3/2$) state. When incident microwaves are tuned to the ($M_J = -1/2$) $\leftrightarrow$ ($M_J = +1/2$) ground-sate Zeeman transition, these levels are nearly equally populated, which causes a decrease in fluorescence scattering from the ions. Transitions in other ions are detected in a similar way to this example. [From (*27*)]



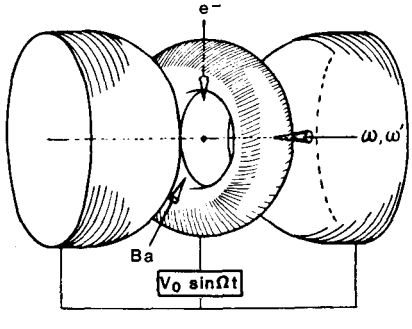


Fig. 3. Photographic image of a single Ba^+ ion (indicated by arrow) localized at the center of a miniature RF trap ($z_0 \cong 0.25$ mm). The lower part of the figure is a drawing of the trap electrodes in the same orientation as in the photograph. [From (*19*)]

that was used for cooling was the $6s^2S_{1/2} - 6p^2P_{1/2}$ transition ($\lambda = 493$ nm). A second laser ($\lambda = 650$ nm) was required to empty the metastable $^2D_{3/2}$ state to which atoms could decay from the $^2P_{1/2}$ state. More recently, laser cooling experiments on trapped ions have been carried out at Seattle (*20*) and Orsay (*29*). In addition, laser cooling on neutral atomic beams has also been observed (*30*).

In both kinds of traps, it has become customary to describe the resulting ion kinetic energy in terms of temperature; however, this must be interpreted with caution. For a cloud of ions in a Penning trap it is theoretically possible (*31*) to cool the cyclotron and axial energy to approximately $h\Delta\nu_a/2$. Equating this energy to k_BT where k_B is Boltzmann's constant, we get $T \cong 1$ mK for $\Delta\nu_a = 43$ MHz (Mg^+). However, the kinetic energy in the magnetron motion depends on the space charge density and size of the cloud (*6*) and can be much larger than this for more than a few ions in the trap. In a similar vein, for a cloud of ions in an RF trap, the energy of motion in the pseudopotential well can be cooled to the same limit as the cyclotron or axial motion in a Penning trap (*28*). However space charge repulsion tends to push the ions toward the edge of the cloud where the energy in the driven motion can be much larger than this. These problems (*17*) which can cause undesirable second-order Doppler shifts can be suppressed in both traps by going to very small numbers of ions—down to one.

Single Ions

In subsequent Heidelberg experiments (*19*) single ions were observed in an RF trap by laser fluorescence scattering. Figure 3 is a photograph of a single Ba^+ ion. The size of the image determined the extent of the ion motion; therefore the temperature in the pseudopotential well was measured to be between 10 and 36 mK. [The driven motion "temperature" will be equal to or larger than this (*13, 17*).] Laser cooling of single Mg^+ ions in Penning (*32*) and RF (*20*) traps has also been accomplished. The lowest temperatures attained are those of the Seattle group (*20*), where the temperature in at least two directions of the pseudopotential motion was determined to be less than 20 mK. [Cooling in all directions will be straightforward (*19*).] In both of the magnesium experiments, the temperature was determined from the Doppler broadening on the optical cooling

transition. For temperatures below about 0.1 K this Doppler broadening contributes only a small part of the total line width, which is now primarily due to radiative decay. Therefore, very low temperature becomes difficult to measure. In the future this problem may be circumvented by probing narrow optical transitions as described below. In any case, the amount of cooling that has already been achieved gives a significant reduction in the second-order Doppler shift correction. If we can assume, for example, that magnesium ions have been cooled to 10 mK, then the second-order Doppler shift correction is about one part in 10^{16}.

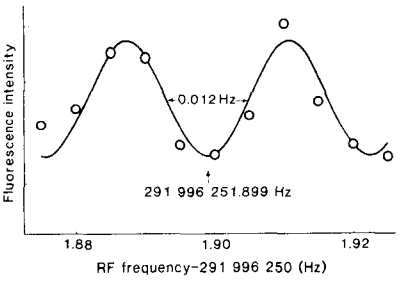


Fig. 4. Graph of a hyperfine resonance of trapped $^{25}Mg^+$ ions. The oscillatory line shape results from the use of the Ramsey resonance method, implemented by applying two coherent RF pulses 1.02 seconds long, separated by 41.4 seconds. The solid curve is a theoretical line shape [From (*32*)]

Spectra

Strongly allowed transitions, which are desirable for the laser cooling, are perhaps not so interesting for high-resolution spectroscopy since the resolutions are limited by the radiative line width (as in the case of Mg^+ above). For high-resolution spectroscopy we usually think in terms of optical pumping/double-resonance detection schemes. A simple example which is characteristic of the method is shown in Fig. 2. In this case, the object was to detect the $(M_J = -1/2) \leftrightarrow (M_J = +1/2)$ ground-state Zeeman transition. The ions are both laser-cooled and pumped into the $^2S_{1/2}$ $(M_J = -1/2)$ ground state as discussed above. The fluorescence (scattered) light intensity is monitored while a microwave generator whose output is directed at the ions is frequency swept through the Zeeman transition. When the resonance condition is satisfied the ground-state populations are nearly equalized; this causes a decrease in the scattered light, which is then the signature of the microwave resonance. This example is illustrative of the various detection schemes used but it is not so interesting for high-resolution spectroscopy since here the line width of the transition was limited by magnetic field fluctuations.

A more interesting example is given by the ground state $(M_I = -3/2, M_J = +1/2) \leftrightarrow (M_I = -1/2, M_J = +1/2)$ nuclear spin flip hyperfine transition of $^{25}Mg^+$. At a field of about 1.24 T, the first derivative of this transition frequency with respect to magnetic field goes to zero; therefore the transition frequency becomes highly insensitive to magnetic field fluctuations. At this field, the resonance shown in Fig. 4 was measured (*33*) with a line width of only 0.012 Hz. The oscillatory line shape results from the use of the Ramsey method (*34*) in the time domain: two phase coherent RF pulses 1.02 seconds long separated in time by 41.4 seconds were used to drive the transition. In order to avoid light shifts, the laser was shut off while the RF transition was driven.

A similar transition was used for the beryllium "clock" mentioned in the introduction. For the beryllium case, the most important limitation to accuracy was caused by a second-order Doppler shift. This resulted because when the laser was off during the RF transition period (about 20 seconds), the ions were heated slightly due to background gas collisions. This problem can be suppressed in the future by using cryogenic pumping.

Frequency Standards and Clocks

In a frequency standard or clock, measurement imprecision $(\delta\nu_{error}/\nu_0)$ is approximately equal to $(Q\ S/N)^{-1}$, where $Q \equiv \nu_0/\Delta\nu_0$ and S/N is the signal-to-noise ratio for detecting the number of ions that have made the transition. If the radiative line width is small enough, then the experimentally observed line width $(\Delta\nu_0)$ need only be limited by the length of time taken to induce the transition. Because of this, $\Delta\nu_0$ is probably independent of the species of trapped ion used. Therefore, we would like to use as high a frequency (ν_0) as possible in order to increase Q and reduce measurement imprecision. This is the single disadvantage of either Mg^+ or Be^+ ions, since the interesting "clock" transitions are only around 300 MHz $(Q \geq 10^{10})$. A better ion for a laser-cooled microwave clock is perhaps Hg^+ (*18*) $(\nu_0 \cong 40$ GHz for $^{199}Hg^+)$. Very important frequency standard work has already been accomplished with this ion (*35*), but laser cooling is much harder to achieve than for Be^+ or Mg^+ (partly because the 194-nm cooling radiation is difficult to produce), and has not been done yet.

A logical extension of this idea is to go to much higher frequency—for example, to use a narrow optical transition. The anticipated Q in this case can be extremely high, 10^{15} or more. A number of transitions in various ions have been proposed (*7*); Dehmelt (*36*) was the first to suggest that such extremely high resolution spectroscopy could be carried out by using single-photon transitions in, for example, single group IIIA ions. For instance (*36*), the $^1S_0 \leftrightarrow {}^3P_0$ transition in Tl^+ ($\lambda = 202$ nm) has a $Q \cong 5 \times 10^{14}$. For such single-photon optical transitions, it is desirable to approximately satisfy the Lamb-Dicke criterion; this is most easily accomplished with single trapped ions. Others (*37*) have proposed using two-photon Doppler-free transitions, for example the $^2S_{1/2} \rightarrow {}^2D_{5/2}$ transitions in Hg^+ ($\lambda = 563$ nm, $Q = 7 \times 10^{14}$). Two-photon optical transitions with equal frequency photons have the advantage of eliminating the first-order Doppler effect for a cloud of many ions, where it is impossible to satisfy the Lamb-Dicke criterion. They ultimately have the disadvantage that the rather large optical fields necessary to drive the transition cause undesirable a-c Stark shifts.

Already, in experiments at Heidelberg and Washington (*38*), the two-photon $^2S_{1/2} \rightarrow {}^2P_{1/2} \rightarrow {}^2D_{3/2}$ stimulated Raman transition in Ba^+ has been observed. In these transitions, the first-order Doppler effect is present; its magnitude (for co-propagating beams) is the same as that of a single laser beam at the difference frequency. Present results (*38*) are limited by laser line width broadening, but such transitions should also be extremely narrow; the lifetime of the $^2D_{3/2}$ state in Ba^+ is 17.5 seconds (*39*), which would give an intrinsic Q of 1.6×10^{16}.

For single ions, optical spectra should give precise temperature information through the intensity of the motional sidebands generated by the Doppler effect (*35, 40*). In the near future, the resolution of the sideband structure in the transitions noted above will probably be limited by the laser line widths. This problem might be alleviated by driving a stimulated Raman transition between two nearly degenerate levels in the electronic ground state of the ion. In $^{24}Mg^+$, for example, the $^2S_{1/2}$ $(M_J = -1/2) \rightarrow {}^2P_{1/2}$ $(M_J = -1/2) \rightarrow {}^2S_{1/2}$ $(M_J = +1/2)$ transition could be driven by using two

laser beams separated in frequency by the $M_J = -1/2$, and $M_J = +1/2$ ground-state frequency difference. The effects of the laser line width would be suppressed by generating the two laser lines with a phase modulator (*41*); the line width of the overall transition would be limited by the ground-state lifetime. The intensity of the motional sidebands would depend on $\mathbf{v} \cdot (\mathbf{k}_1 - \mathbf{k}_2)$, where $\mathbf{k}_1$ and $\mathbf{k}_2$ are the wave vectors of the two laser beams and $\mathbf{v}$ is the ion velocity. Thus the angle between the beams could be chosen to optimize the temperature information.

Conclusions

The projected accuracy for optical frequency standards is extremely high. As an example, in In^+, the line width of the $^1S_0 \rightarrow {}^3P_1$ "cooling" transition is about 1.3 MHz; this implies a second-order Doppler shift of 10^{-19} or lower. Other systematic shifts can occur (*1, 7, 13, 16, 18, 33, 35–38, 42*) but it is not unreasonable to think they will be controllable to this level. These extreme accuracies make important the problem of measurement imprecision, since the signal-to-noise ratio on a single ion will be about one for each measurement cycle. Practically speaking, this means that a long averaging time will be required to reach a measurement precision equal to these accuracies. In fact, for a while, the accuracy and resolution may be limited by laser line width characteristics (line width and line width symmetry); however, the potential for extremely narrow lasers also exists (*43*).

With the great potential for the optical frequency standards, one can logically ask why we bother thinking about RF or microwave frequency standards, where the desired large numbers of ions (to increase signal-to-noise and measurement precision) causes unwanted second-order Doppler effects (*17*). At present, the answer concerns the utility of optical frequency standards within current technological limitations. To use such devices as clocks as in communications and navigation, one must count cycles of the radiation. At microwave frequencies this is straightforward. At optical frequencies it is technically feasible but very hard (*44*); it has not been done yet. To illustrate further, one might also have asked why we do not push the frequency even higher, that is, make a clock based on narrow Mössbauer transitions in bound nuclei. Here, however, the technological problems become even more apparent. In spite of the technical problems of making an optical "clock," optical frequency standards will, of course, find many immediate uses. An obvious class of experiments are cosmological in nature; for example, more precise measurements of the gravitational red shift will arise. In any case, the potential accuracy for stored ion spectroscopy in all spectral regions seems extremely high. Frequency standards and clocks with inaccuracy of one part in 10^{15} appear very reasonable; eventually they could be orders of magnitude better than this.

References and Notes

1. J. J. Bollinger, W. M. Itano, D. J. Wineland, *Proc. 37th Annu. Symp. Freq. Control* (1983), p. 37 (copies available from Systematics General Corporation, Brinley Plaza, Route 38, Wall Township, N.J. 07719).
2. S. Garpman, I. Lindgren, J. Lindgren, J. Morrison, *Z. Phys. A* **276**, 167 (1976); S. Ahmad *et al.*, *Phys. Rev. A* **27**, 2790 (1983).
3. J. Vetter, H. Ackermann, G. ZuPutlitz, E. W. Weber, *Z. Phys. A* **276**, 161 (1976).
4. A. G. Mungall and C. C. Costain, *IEEE Trans. Instrum. Meas.* **IM-32**, 224 (1983); G. Becker, *Proc. 10th Int. Congr. Chronometry* **2**, 33 (1979); D. J. Wineland, D. W. Allan, D. J. Glaze, H. Hellwig, S. Jarvis, Jr., *IEEE Trans. Instrum. Meas.* **IM-25**, 453 (1976).
5. D. J. Larson, *Bull. Am. Phys. Soc.* **19**, 1193 (1974).
6. J. J. Bollinger and D. J. Wineland, *Phys. Rev. Lett.* **53**, 348 (1984).
7. D. J. Wineland, W. M. Itano, R. S. Van Dyck, Jr., *Adv. At. Mol. Phys.* **19**, 135 (1983).
8. E. W. Weber, *Phys. Rep.* **32**, 123 (1977).
9. A. Rich and J. C. Wesley, *Rev. Mod. Phys.* **44**, 250 (1972).
10. K. H. Kingdon, *Phys. Rev.* **21**, 408 (1923).
11. M. H. Prior and E. C. Wang, *Phys. Rev. A* **16**, 6 (1977).
12. E. Fisher, *Z. Phys.* **156**, 1 (1959); R. F. Wuerker, H. Shelton, R. V. Langmuir, *J. Appl. Phys.* **30**, 342 (1959).
13. H. G. Dehmelt, *Adv. At. Mol. Phys.* **3**, 53 (1967); *ibid.* **5**, 109 (1969).
14. F. M. Penning, *Physica* **3**, 873 (1936).
15. P. Ekstrom and D. Wineland, *Sci. Am.* **243** (No. 2), 104 (1980); D. J. Wineland, P. Ekstrom, H. Dehmelt, *Phys. Rev. Lett.* **31**, 1279 (1973).
16. W. M. Itano, L. L. Lewis, D. J. Wineland, *Phys. Rev. A* **25**, 1233 (1982).
17. D. J. Wineland, *Natl. Bur. Stand. (U.S.) Publ. 617*, in press.
18. D. J. Wineland, W. M. Itano, J. C. Bergquist, F. L. Walls, *Proc. 35th Annu. Symp. Freq. Control* (1981), p. 602 (copies available from Electronic Industries Association, 2001 Eye Street, NW, Washington, D.C. 20006).
19. W. Neuhauser, M. Hohenstatt, P. Toschek, H. Dehmelt, *Phys. Rev. A* **22**, 1137 (1980).
20. W. Nagourney, G. Janik, H. Dehmelt, *Proc. Natl. Acad. Sci. U.S.A.* **80**, 643 (1983).
21. R. H. Dicke, *Phys. Rev.* **89**, 472 (1953).
22. H. A. Schuessler, E. N. Fortson, H. G. Dehmelt, *ibid.* **187**, 5 (1969).
23. T. W. Hänsch and A. L. Schawlow, *Opt. Commun.* **13**, 68 (1975).
24. D. J. Wineland and H. G. Dehmelt, *Bull. Am. Phys. Soc.* **20**, 637 (1975).
25. D. J. Wineland, R. E. Drullinger, F. L. Walls, *Phys. Rev. Lett.* **40**, 1639 (1978).
26. H. G. Dehmelt and F. L. Walls, *ibid.* **21**, 127 (1968); D. A. Church and H. G. Dehmelt, *J. Appl. Phys.* **40**, 3421 (1969).
27. D. J. Wineland, J. C. Bergquist, W. M. Itano, R. E. Drullinger, *Opt. Lett.* **5**, 245 (1980).
28. W. Neuhauser, M. Hohenstatt, P. Toschek, H. Dehmelt, *Phys. Rev. Lett.* **41**, 233 (1978).
29. F. Plumelle, Abstracts for the 15th European Group on Atomic Spectroscopy Conference, Madrid, 5 to 8 July 1983.
30. S. V. Andreev, V. I. Balykin, V. S. Letokhov, V. G. Minogin, *JETP Lett.* **34**, 442 (1981); J. V. Prodan, W. D. Phillips, H. J. Metcalf, *Phys. Rev. Lett.* **49**, 1149 (1982); J. L. Hall, W. Ertmer, R. Blatt, in preparation; W. D. Phillips, Ed., "Laser-cooled and trapped atoms," *Natl. Bur. Stand. (U.S.) Spec. Publ. 653* (1983).
31. W. M. Itano and D. J. Wineland, *Phys. Rev. A* **25**, 35 (1982).
32. D. J. Wineland and W. M. Itano, *Phys. Lett. A* **82**, 75 (1981).
33. W. M. Itano and D. J. Wineland, *Phys. Rev. A* **24**, 1364 (1981).
34. N. F. Ramsey, *Molecular Beams* (Oxford Univ. Press, London, 1956).
35. F. G. Major and G. Werth, *Phys. Rev. Lett.* **30**, 1155 (1973); M. D. McGuire, R. Petsch, G. Werth, *Phys. Rev. A* **17**, 1999 (1978); M. Jardino *et al.*, *Appl. Phys.* **24**, 107 (1981); L. S. Cutler, R. P. Giffard, M. D. McGuire, *Proc. 37th Annu. Symp. Freq. Control* (1983), p. 32 (copies available from Systematics General Corporation, Brinley Plaza, Route 38, Wall Township, N.J. 07719); M. Jardino, F. Plumelle, M. Desaintfuscien, J. L. Duchêne, *Proc. 38th Annu. Symp. Freq. Control* (Philadelphia, 1984), in press.
36. H. Dehmelt, *IEEE Trans. Instrum. Meas.* **IM-31**, 83 (1982).
37. P. L. Bender *et al.*, *Bull. Am. Phys. Soc.* **21**, 599 (1976).
38. W. Neuhauser, M. Hohenstatt, P. E. Toschek, H. Dehmelt, in *Spectral Line Shapes*, B. Wende, Ed. (deGruyter, Berlin, 1981); W. Nagourney, *Bull. Am. Phys. Soc.* **29**, 815 (1984).
39. R. Schneider and G. Werth, *Z. Phys. A* **293**, 103 (1979).
40. H. A. Schuessler, *Appl. Phys. Lett.* **18**, 117 (1971); F. G. Major and J. L. Duchêne, *J. Phys. (Orsay)* **36**, 953 (1975); H. S. Lakkaraju and H. A. Schuessler, *J. Appl. Phys.* **53**, 3967 (1982); M. Jardino, F. Plumelle, M. Desaintfuscien, in *Laser Spectroscopy VI*, H. P. Weber and W. Lüthy, Eds. (Springer-Verlag, New York, 1983), p. 173.
41. J. E. Thomas *et al.*, *Phys. Rev. Lett.* **48**, 867 (1982).
42. D. J. Wineland and W. M. Itano, *Bull. Am. Phys. Soc.* **27**, 864 (1982).
43. J. L. Hall, L. Hollberg, Ma Long-Shen, T. Baer, H. G. Robinson, *J. Phys. (Orsay)* **42**, C8-59 (1981); A. Yariv and K. Vahala, *IEEE J. Quant. Electron.* **QE-19**, 889 (1983); R. W. P. Drever *et al.*, *Appl. Phys. B* **31**, 97 (1983); J. Hough *et al.*, *ibid.* **33**, 179 (1984).
44. See Proceedings of the Third Symposium on Frequency Standards and Metrology, *J. Phys.* **42**, Colloque C-8, December 1981; V. P. Chebotayev *et al.*, *Appl. Phys. B* **29**, 63 (1982); D. A. Jennings *et al.*, *Opt. Lett.* **8**, 136 (1983); K. M. Baird, *Phys. Today* **36** (No. 1), 52 (1983).

45. I gratefully acknowledge the support of the Office of Naval Research and the Air Force Office of Scientific Research. I also thank J. J. Bollinger, J. C. Bergquist, J. Cooper, W. M. Itano, J. D. Prestage, and S. R. Stein for helpful comments on the manuscript.

PHYSICAL REVIEW VOLUME 138, NUMBER 1A 17 MAY 1965

Hydrogen-Maser Principles and Techniques*

D. KLEPPNER,† H. C. BERG,‡ S. B. CRAMPTON,§ AND N. F. RAMSEY
Harvard University, Cambridge, Massachusetts

AND

R. F. C. VESSOT, H. E. PETERS, AND J. VANIER
Varian Associates, Beverly, Massachusetts

(Received 17 December 1964)

Techniques and design principles relevant to the construction and operation of a hydrogen maser are presented in detail. These include methods for the generation of atomic hydrogen, state selection, design of the microwave cavity, production of very low magnetic fields, coating the hydrogen storage bulb, and tuning the maser. A figure of merit is introduced which indicates the optimum choice of parameters.

I. INTRODUCTION

ALTHOUGH the hydrogen maser has proved useful both as a spectroscopic tool and as a frequency standard, only a portion of its theory has so far been described.[1] Most of the description of experiments with the maser have given the results but have omitted technical details.[2–8] In the present paper, the interrelation of the various physical effects governing the maser's behavior will be discussed along with relevant operational and technical considerations.

The hydrogen maser operates between the ground-state hyperfine levels of atomic hydrogen. For use as a frequency standard, the maser oscillates on the transition $(F=1, m_F=0) \rightarrow (F=0, m_F=0)$ at a frequency of approximately 1420 Mc/sec. Figure 1 is a schematic diagram. Molecular hydrogen is dissociated in the source and is formed into an atomic beam which passes through a state-selecting magnet. The emergent beam contains only atoms in the states $(F=1, m=1)$ and $(F=1, m=0)$. The beam passes into a storage bulb which has a specially prepared surface and in which the atoms remain for approximately 0.3 sec before escaping. The bulb is located in a cavity tuned to the hyperfine transition frequency. Stimulated emission occurs if the beam flux is sufficiently high and a signal is produced in the cavity. This signal is detected by means of a small coupling loop. The cavity is surrounded by magnetic shields to reduce the ambient field and a small uniform field is produced at the storage bulb by a solenoid.

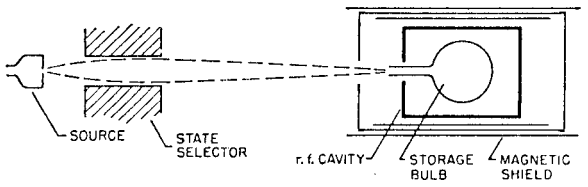


FIG. 1. Schematic diagram of the hydrogen maser.

In Sec. II, some formulas are presented which govern the choice of design parameters. Subsequent sections are: III. Source, IV. State Selector, V. Vacuum System, VI. Cavity, VII. Magnetic Shields, VIII. Storage Bulb, IX. Electronics, and X. Tuning Methods.

II. OPERATING CONDITIONS

Basic formulas governing the operation of the hydrogen maser have been published by Kleppner, Goldbenberg, and Ramsey[1] hereafter referred to as (KGR). Their results will be extended here in order to illustrate more clearly the interdependence of the design parameters. In particular, by including the effect of spin-exchange relaxation in the governing equations, it will be shown that for oscillation to occur on the transition $(F=1, m=0) \rightarrow (F=0, m=0)$ there is not only a minimum beam flux but also a *maximum* permissible flux. The analysis leads to a constraint condition governing the cavity geometry, storage-bulb size, and various relaxation times. Unless the constraint is satisfied, the maser will not oscillate, and the constraint equation leads to a useful figure of merit which predicts how suitable a given configuration will be.

Our starting point is KGR Eq. (7),

$$P=\tfrac{1}{2}I\hbar\omega\frac{x^2}{(1/T_b)^2+x^2+(\omega-\omega_0)^2}, \qquad (1)$$

where P is the power radiated by the atoms, I is the

* Work supported by the National Science Foundation and the Office of Naval Research (Harvard) and by the National Aeronautics and Space Administration and the Office of Naval Research (Varian).

† Alfred P. Sloan Foundation Fellow.

‡ Junior Fellow, Society of Fellows, Harvard University.

§ National Science Foundation Postdoctoral Fellow.

[1] D. Kleppner, H. M. Goldenberg, and N. F. Ramsey, Phys. Rev. **126**, 603 (1962).

[2] H. M. Goldenberg, D. Kleppner, and N. F. Ramsey, Phys. Rev. Letters **5**, 361 (1960); Appl. Opt. **1**, 55 (1962).

[3] S. B. Crampton, D. Kleppner, and N. F. Ramsey, Phys. Rev. Letters **11**, 338 (1963).

[4] H. C. Berg, D. Kleppner, and N. F. Ramsey, Bull. Am. Phys. Soc. **8**, 379 (1963).

[5] S. B. Crampton, D. Kleppner, and H. G. Robinson, Bull. Am. Phys. Soc. **8**, 351 (1963).

[6] S. B. Crampton and D. Kleppner, Bull. Am. Phys. Soc. **9**, 451 (1964).

[7] H. G. Robinson, H. C. Berg, and S. B. Crampton, Bull. Am. Phys. Soc. **9**, 564 (1964).

[8] E. N. Fortson, D. Kleppner, and N. F. Ramsey, Phys. Rev. Letters **13**, 22 (1964).

net input flux, i.e., the difference in the flux of atoms entering in the state ($F=1$, $m=0$) and the state ($F=0$, $m=0$), ω_0 is the resonance frequency, and $x=\mu_0 H_z/\hbar$, where H_z is the oscillating field amplitude. T_b is the mean storage time of the bulb (denoted by γ^{-1} in KGR). Under the assumptions discussed below, this equation can be generalized to the case in which other relaxation processes occur. In particular, we shall distinguish between processes which relax the population difference between the two states of interest and those which relax the oscillating moment. In analogy with nmr terminology, we designate the decay times for the processes by T_1 and T_2, respectively. Then it can be shown that Eq. (1) becomes[9,10]

$$P=\tfrac{1}{2}I\hbar\omega\frac{x^2}{1/(T_1T_2)+x^2+(T_2/T_1)(\omega-\omega_0)^2}. \quad (2)$$

If the only relaxation mechanism is escape of atoms from the storage bulb, then $T_1=T_2=T_b$. However, we must also allow for hydrogen-hydrogen spin exchange (s.e.). For this process, $T_2=2T_1$,[11,12] where

$$(1/T_1)_{\text{s.e.}}=n\sigma\bar{v}_r. \quad (3)$$

Here σ is the hydrogen "spin-flip" cross section,[13] estimated by Mazo[14] to be 2.85×10^{-15} cm^2, $\bar{v}_r$ is the average relative hydrogen velocity [$\bar{v}_r=4(kT/\pi m)^{1/2}=3.58\times10^5$ cm/sec at $T=308$°K], and n is the hydrogen density given by

$$n=I_{\text{tot}}T_b/V_b, \quad (4)$$

where I_{tot} is the total flux of atoms entering the storage bulb, and V_b is the storage-bulb volume. Normally, the states ($F=1$, $m=0$) and ($F=1$, $m=1$) are focused, so that $I_{\text{tot}}=2I$. If the state selection is imperfect, other states may be present so that I_{tot}/I can have a large value.

There are a variety of other possible relaxation processes,[11] some of which are discussed in KGR. For these, T_1 and T_2 are not in general the same. We shall allow for them by letting T_1' and T_2' stand for the total relaxation times due to all processes other than escape from the bulb and hydrogen spin exchange. Then

$$\begin{aligned}1/T_1&=1/T_1'+(1/T_1)_{\text{s.e.}}+1/T_b,\\ 1/T_2&=1/T_2'+(1/T_2)_{\text{s.e.}}+1/T_b.\end{aligned} \quad (5)$$

Here the subscript s.e. indicates that the relaxation mechanism is spin exchange.

We assume, for the present, that T_1' and T_2' are constants, insofar as they do not depend on the hydrogen atom flux or on the oscillation level of the maser. This is not necessarily true, since the structure of the storage-bulb wall may depend on the hydrogen density and the state populations, as will be discussed below.

To analyze the conditions for stationary oscillation of the maser, we proceed as in KGR by equating the radiated power to the dissipated power $\omega W/Q$ where W is the stored energy in the cavity. By substituting Eqs. (3), (4), and (5) in Eq. (2) with $\omega=\omega_0$ and rearranging, using $x^2=(\mu_0/\hbar)^2(8\pi W/V_c)\eta$, we arrive eventually at the following relation between the power radiated and the beam flux:

$$P/P_c=-2q^2(I/I_{\text{th}})^2+(1-cq)I/I_{\text{th}}-1. \quad (6)$$

P_c is defined by the following:

$$P_c=\omega\hbar^2V_c/8\pi\mu_0^2QnT_t^2, \quad (7)$$

where

$$\begin{aligned}1/T_t^2&=(1/T_1'+1/T_b)(1/T_2'+1/T_b),\\ \eta&=\langle H_z\rangle_{\text{bulb}}^2/\langle H^2\rangle_{\text{cavity}},\end{aligned} \quad (8)$$

and

$$V_c=\text{cavity volume}.$$

I_{th} is defined by

$$I_{\text{th}}=2P_c/\hbar\omega. \quad (9)$$

[Physically, I_{th} is the net threshold flux for oscillation to occur providing spin exchange can be neglected. In the presence of spin exchange, the net flux for oscillation is somewhat larger, and is given by $I_{\min}$, as shown in Eq. (13) below.]

Also,

$$c=\left[\frac{1/T_b+1/T_1'}{1/T_b+1/T_2'}\right]^{1/2}+2\left[\frac{1/T_b+1/T_2'}{1/T_b+1/T_1'}\right]^{1/2}. \quad (10)$$

The quantity q is an important quality parameter which has the following value:

$$q=\frac{\sigma\bar{v}_r\hbar}{8\pi\mu_0^2}\frac{T_b}{T_t}\frac{V_c}{\eta V_b}\frac{1}{Q}\frac{I_{\text{tot}}}{I}. \quad (11)$$

From Eq. (6) it is apparent that the power radiated by the maser oscillator is a quadratic function of the beam flux. If we require that both the flux and the power be positive quantities, we obtain the following condition:

$$q<(c-2\sqrt{2})/(c^2-8). \quad (12)$$

It also follows that the maximum and minimum net beam fluxes for oscillation are given by

$$\underset{\min}{I_{\max}}/I_{\text{th}}=\frac{1-cq\pm[1-2cq+(c^2-8)q^2]^{1/2}}{4q^2}. \quad (13)$$

Because of spin exchange, the resonance linewidth $\Delta\nu$ depends on the flux. It is given by the following

[9] S. B. Crampton, Ph.D. thesis, Harvard, 1964 (unpublished).
[10] P. L. Bender, Phys. Rev. 132, 2154 (1963).
[11] H. C. Berg, Ph.D. thesis, Harvard, 1964 (unpublished). Also H. C. Berg, Phys. Rev. 137, A1621, (1965).
[12] J. P. Wittke and R. H. Dicke, Phys. Rev. 103, 620 (1956).
[13] L. C. Balling, R. J. Hanson, and F. M. Pipkin, Phys. Rev. 132, 2154 (1964).
[14] R. M. Mazo, J. Chem. Phys. 34, 169 (1961).

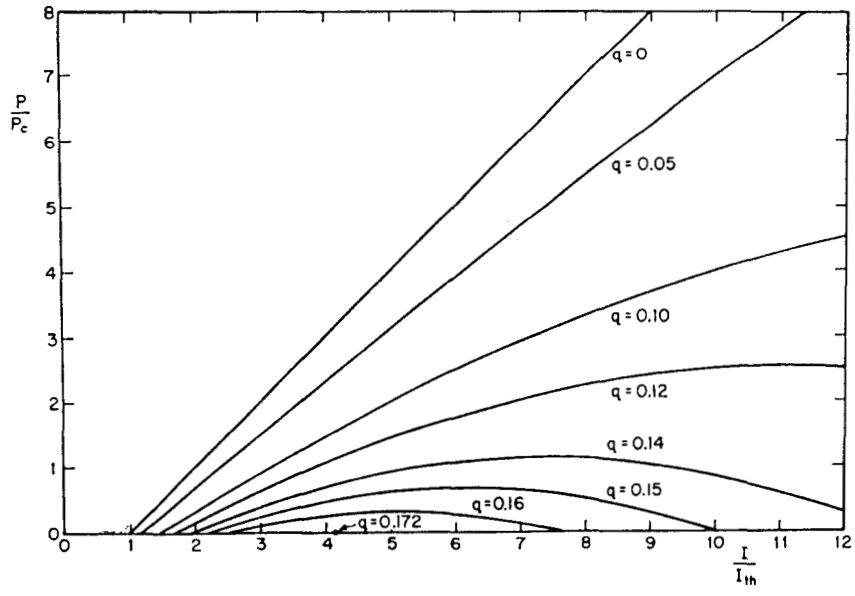


FIG. 2. P/P_c versus I/I_{th} for different values of the parameter q. This family of curves shows the strong influence of q on the operating conditions. If spin exchange is neglected, $q=0$ and radiated power increases monatonically with beam flux. For $q>0$, there is an upper limit to the flux for oscillation to occur and above a certain value ($q=0.172$) the maser cannot radiate at any beam flux. q is defined by Eq. (11).

expression:

$$\Delta\nu=1/\pi T_2=(1/\pi)[(1/T_2'+1/T_b)+(I/I_{th})q/T_t]. \quad (14)$$

To proceed further, we must choose a particular ratio for T_1' and T_2'. In the event that $T_1'=T_2'$, $c=3$, and we have

$$P/P_c=-2q^2(I/I_{th})^2+(1-3q)I/I_{th}-1, \quad (15)$$

$$q<3-2\sqrt{2}=0.172, \quad (16)$$

$$I_{\substack{\max\\ \min}}/I_{th}=\frac{1-3q\pm(1-6q+q^2)^{1/2}}{4q^2}, \quad (17)$$

$$\Delta\nu=(1/\pi T_t)(1+qI/I_{th}). \quad (18)$$

For purposes of tuning, to be discussed in Sec. X, it is essential that the ratio of the resonance widths at the extremes of permissible flux be large. This ratio is

$$r=\frac{\Delta\nu\ \max}{\Delta\nu\ \min}=\frac{1+q+(1-6q+q^2)^{1/2}}{1+q-(1-6q+q^2)^{1/2}}. \quad (19)$$

To illustrate the behavior of the maser as the factor q varies, Eq. (15) is plotted in Fig. 2 for a few values of q. Equations (17) and (19) are plotted in Fig. 3, which shows both r and the permissible extremes of flux as a function of q. It is evident from the figures that a relatively small decrease in q can take the maser from a state of critical dependence on flux, or even no oscillation at all, to a favorable region of oscillation.

In order for the maser to operate as an oscillator, it is essential that the inequality in (16) be satisfied. If we substitute Eq. (11) in Eq. (16), we obtain

$$T_b/T_t\times V_c/(\eta V_b)\times 1/Q\times I_{tot}/I<3.47\times 10^{-4}. \quad (20)$$

As an example, let us insert the following typical values:

$$V_c/V_b=6.7,\ \eta=3,\ I_{tot}/I=2,\ Q=3\times 10^4.$$

Then, for oscillation,

$$T_b/T_t=1+T_b/T_1<2.32. \quad (21)$$

If unwanted relaxation mechanisms are present, the bulb storage time must be decreased so that this inequality holds. Conversely, if a given bulb geometry and storage time are required, Eq. (21) can be used to set an upper limit on permissible relaxation.

If we assume $V_c=1.3\times 10^4$ cm^3, $T_b/T_t=1.3$, and $T_b=0.3$ sec, reasonable values in practice, then we have further

$$P_c=5.9\times 10^{-13}\ \text{W},$$
$$I_{th}=1.3\times 10^{12}\ \text{particles/sec},$$
$$q=0.097,$$
$$I_{min}=1.9\times 10^{12}\ \text{particles/sec},$$

and

$$r=4.$$

The time T_2 can be measured by observing the decay of radiation when the maser is pulsed while operating below oscillation threshold.[11] In this condition, spin-exchange is usually negligible, so that if the relation between T_1' and T_2' is known, T_t can be determined from Eq. (8). The other quantities involved in q can be calculated in principle, though the quantity η is difficult to compute precisely, except for the case of simple geometry, such as a cylindrical bulb in a cylindrical cavity, which will be discussed in Sec. VIII. Alternatively, the cavity Q can be decreased by increasing the coupling until the maser ceases to operate at any flux. At this point, the inequality (12) becomes an equality. By measuring Q at this cutoff condition and

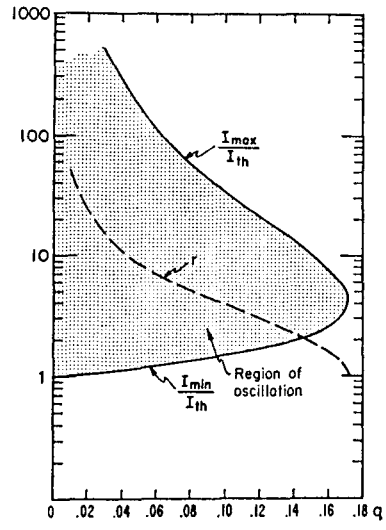


FIG. 3. I_{max}/I_{th} and I_{min}/I_{th} versus q. The maser can oscillate if the flux obeys $I_{min} \leqslant I \leqslant I_{max}$, indicated by the shaded region. The curve r gives the ratio I_{max}/I_{min} and is a useful measure of the amount by which the resonance line can be broadened by spin exchange.

at the normal operating condition, q can be determined for the latter case.

It should be pointed out that care must be used in applying this analysis. In some cases, T_1 is known once T_2 has been measured. For instance, in the case of spin-exchange collisions between hydrogen and a spin-$\frac{1}{2}$ or spin-1 background gas, $T_2 = \frac{4}{3}T_1$.[11] However, unless the relaxation mechanisms are understood, the ratio of T_1 to T_2 must be separately measured. Any mechanism which converts atoms to molecular form in effect behaves like a hole through the wall and decreases the storage time of the bulb T_b. In addition, two assumptions have been made which do not necessarily always hold. In the first place, the maser has been treated as a simple two-level system. This is not necessarily true if there is coupling with the remaining two hyperfine levels. As an example, relaxation due to motion of the atoms through an inhomogeneous magnetic field couples the three upper hyperfine states. The effect of this is, on the one hand, to relax the oscillating moment, and on the other hand, to feed atoms from the (1,1) and (1, −1) states into the (1,0) state as this is depleted by radiation to the ground state. A full analysis of this can be carried out using the density matrix formalism,[9] but for many purposes the process can be neglected by ensuring that magnetic relaxation is not large.

The second assumption is that the remaining relaxation processes are simply characterized by constant rates. In fact, there is evidence to indicate that part of the relaxation at the storage-bulb surface can occur at a rate proportional to the hydrogen density.[11] One effect of such a process is the enhancement of the value of σ entering Eq. (1), and this must be taken into account.

III. SOURCE

Of the three time-honored devices for dissociating molecular hydrogen—Wood's discharge, rf (or microwave) discharge, and thermal dissociator—the rf discharge has so far proven to be the most convenient for use in the hydrogen maser. The Wood's tube is cumbersome to build and operate, and the thermal dissociator produces a relatively hot beam which is difficult to focus. A simple rf discharge is described here.

The discharge takes place in a spherical Pyrex bulb, approximately $2\frac{1}{2}$ cm diam. The source aperture is a hole typically $\frac{1}{2}$ mm diam, though holes up to 1 mm diam function satisfactorily. The aperture can be connected to the bulb by a short length of Pyrex tubing or led directly through the wall of the bulb, as shown in Fig. 4. If desired, the flux to the source chamber pump can be reduced by using a multitube glass collimator instead of a simple hole.[15–17] Pressure in the discharge tube is typically 0.3 mm Hg, though the discharge operates well at pressures between 0.05 and 0.8 mm Hg. Power for the discharge is supplied by a single-tube oscillator operating at a frequency of 200 Mc/sec or higher. A variable link is used to couple the power to the discharge. A circuit for the oscillator and details of the coupling link are shown in Fig. 5. Tuning is facilitated by using a directional power meter in the transmission line. Matching must be done with the tube lit. Under typical conditions the oscillator tube will draw 150 mA at 400 V, and the input power to the discharge tube is 10 W. At higher power levels, the discharge tube should be cooled with forced air.

It is possible to treat the discharge tube walls to reduce recombination using phosphoric acid or Dri Film[18] (but not Teflon, which is decomposed by the discharge). The most satisfactory operation has been

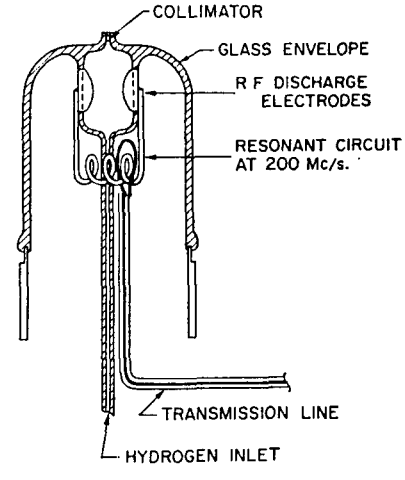


FIG. 4. Atomic hydrogen source.

[15] R. F. C. Vessot and H. E. Peters, IRE (Inst. Radio Engrs.), Trans. Instr. 11, 183 (1962); J. Vanier, H. E. Peters, and R. F. C. Vessot, *ibid.* (to be published).

[16] P. Grivet and N. Bloembergen, *Quantum Electronics III* (Columbia University Press, New York, 1964), pp. 333–347; also pp. 409–417.

[17] The collimators described in Refs. 16 and 17 are effective at low source pressures but are not useful if high flux is desired. Finer collimators which are useful over a wider pressure range can be supplied by Permeonics, Inc., Southbridge, Massachusetts.

[18] Type SC-02 (dimethyldichlorosilane) General Electric Company, Silicone Products Department, Waterford, New York.

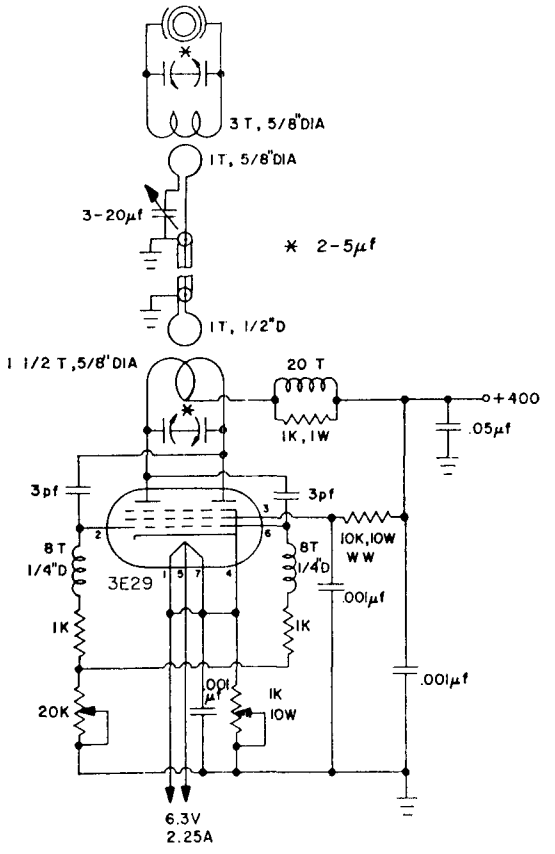


FIG. 5. Hydrogen discharge power oscillator.

obtained with untreated Pyrex. Normally, the tube must age for a few hours before the discharge acquires its characteristic red color. After long operation, the discharge tube becomes discolored by the decomposition of the glass; black borosilicate products appear on the inside surface of the tube. This does not appear to have a serious effect on the operation of the discharge.

Hydrogen may be introduced to the discharge by means of a conventional gas-handling system or through a combination purifier, pressure reducer, and flow controller consisting of a simple palladium leak. (Conventional regulators with variable leaks are difficult to reset and usually will not regulate pressure to better than 10%.) Both the leak and a servo system for controlling the source pressure are shown in Fig. 6. The source pressure is monitored by a Pirani gauge consisting of a thermistor bead suspended in the hydrogen inlet line. The gauge is housed in a thermally controlled box which also houses the resistance bridge and servo preamplifier. Flux through the palladium leak is controlled by varying its temperature with a heating coil.

IV. STATE SELECTOR

The hexapolar field proposed by Friedburg and Paul[19] is ideally suited to the hydrogen maser. This type of state selector has a large acceptance angle, but in general yields an atomic beam of relatively large area. Fortunately, the entrance to the storage bulb can be made large enough to accept a substantial part of the beam.

The design of a simple permanent six-pole magnet has been given by Christensen and Hamilton.[20] A magnet similar to their design has been constructed and has the following properties: gap diameter, $\frac{1}{8}$ in.; length, 3 in.; maximum field at pole tip (as measured by a rotating coil magnetometer), 9900 Oe. The magnet can be bolted or glued together with epoxy and is magnetized by passing approximately 500 A through 33 turns of wire wound on the poles or by using a pulsed magnetizer.[21]

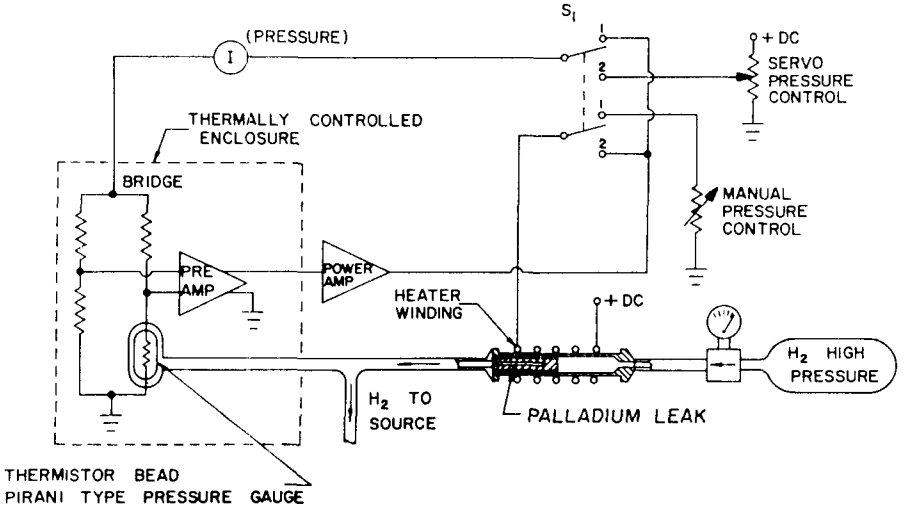


FIG. 6. Hydrogen gas-handling system and flow regulator. The palladium leak is made from an alloy pellet 70% palladium, 30% silver, approximately $\frac{1}{2}$-in. long, $\frac{3}{16}$-in. o.d., $\frac{1}{16}$-in. i.d. The Pirani-gauge preamplifier and bridge are similar to those in Fig. 8.

[19] H. Friedburg and W. Paul, Naturwiss. **38**, 159 (1951).
[20] R. L. Christensen and D. R. Hamilton, Rev. Sci. Instr. **30**, 356 (1959).
[21] These magnets can be obtained commercially from Varian Company, Bomac Division, Beverly, Massachusetts.

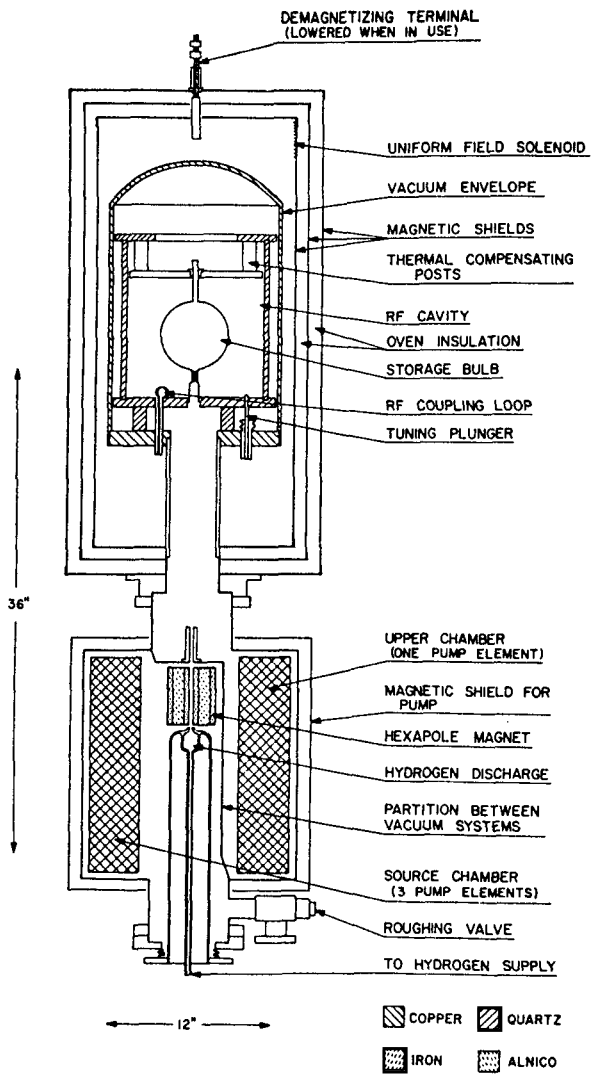


FIG. 7. Details of hydrogen maser.

Typical dimensions for the beam optics of a hydrogen maser are indicated in Fig. 7. The total flux from a collimated source is 10^{16} atom/sec with a source pressure of 0.1 mm Hg. The collimater is composed of 400 glass tubes forming a cylinder 1 mm diam and 1 mm long. The net flux of atoms to the bulb in the desired state is approximately 2×10^{12} atoms/sec.

V. VACUUM SYSTEM

Because of the possibility of contamination of the storage bulb surface, the hydrogen maser demands a relatively clean vacuum system. Mercury diffusion pumps are satisfactory, providing they are properly baffled. Oil diffusion pumps can lead to appreciable contamination of the surface, even when there is a very low residual pressure, and are not recommended. The most satisfactory pump appears to be an ion sputtering pump. These pumps have both high speed and high capacity for hydrogen. For example, a 75 liter/sec VacIon pump can absorb well over $1\frac{1}{2}$ mole of H_2, providing care is taken to prevent the cathode from shorting to the anode as it expands due to the absorbed hydrogen.

In general, it is desirable to pump differentially the source and storage bulb chambers in order to prevent excessive beam scattering. Fortunately, a relatively high background pressure of molecular hydrogen (up to 10^{-5} mm Hg) does not interfere with the maser's operation since H_2 is a good buffer gas. The background pressure with the beam turned off should be 10^{-8} mm Hg, or less, to prevent broadening of the resonance line due to spin exchange collisions with O_2.

Conservative design of the vacuum system calls for a pumping speed of 200 liter/sec (air) for the source chamber, and perhaps half as much for the rest of the system. Under conditions of extremely high flux, the source chamber pressure will be 3×10^{-5} mm Hg (ion gauge, uncorrected for H_2), while under normal operating conditions the pressure is a factor of 10 lower.

Figure 7 shows a system employing a titanium sputtering pump enclosing the source structure. Here, a VacIon 250-liter/sec pump is used with the modification that a partition is installed which allows one of the four pumping elements to pump the upper chamber. The state-selecting magnet is mounted on the partition. Flanges at the upper and lower ends of the pump connect to the bell jar and source assemblies, respectively. Stray fields from the pump magnets are reduced by a shield of $\frac{1}{16}$-in. Armco iron surrounded by a second shield of $\frac{1}{16}$-in. Mu Metal.

The microwave cavity must be evacuated to prevent atmospheric disturbance of the tuning.[15,16] This may be done by a separate roughing system, or, as shown in Fig. 7, by connecting it directly to the high-vacuum system. To afford thermal isolation from the bell jar, the cavity supports are three thin-walled quartz tubes. The bell jar of $\frac{1}{4}$-in. copper provides a thermal enclosure for the cavity. The copper bell-jar base plate is connected to the pump manifold by a thin-walled stainless steel neck section to minimize heat transfer from the base plate to the maser frame. Further aspects of the thermal control of the cavity enclosure are dealt with in the next section.

VI. CAVITY

Stability of the microwave cavity is critical to the operation of the maser. The problem is simply summarized by noting that the fraction of the atomic resonance linewidth by which the maser is "pulled" by a mistuned cavity is identical to the fraction of the cavity linewidth by which the cavity is mistuned. This is apparent from the following expression for the shift of the oscillator frequency ν, from the true resonance frequency ν_0, by a cavity tuned to a frequency ν_c.[1]

$$\frac{\nu-\nu_0}{\nu_0}=\frac{\nu_c-\nu_0}{\nu_0}\frac{\Delta\nu_r}{\Delta\nu_c}, \tag{22}$$

where $\Delta\nu_r$ and $\Delta\nu_c$ are the resonance and cavity linewidths, respectively.

In practice, the major source of drift of the maser is frequency pulling due to cavity drift. Although a particular cavity design is described here, it should be emphasized that other approaches are possible. In particular, it should be possible to stabilize the cavity frequency with a servo system. However, at the time of writing, such a system has not been put into operation.

The most convenient cavity mode is the cylindrical TE_{011} mode, since this has only azimuthal wall currents so that good contact between the cylinder and end plates is not required. The corresponding spherical mode should also be favorable, though it is difficult to construct.

The electrical properties of the TE_{011} mode are well described in the literature,[22] and only a few points will be summarized here. The unloaded cavity resonates with length and diameter both equal to 27.6 cm. The theoretical Q is 87 000 for silver-plated walls. In practice, $Q=60\,000$ is obtainable. The quartz storage bulb does not appreciably affect Q, but a bulb 15 cm in diameter with walls 1 mm thick will decrease the resonant length by about 5 cm.

Power is coupled from the cavity by a loop mounted in an end plate near the position of maximum magnetic field. A loop area ≈ 1 cm^2 will couple the cavity critically to a 50-Ω line.

Since thermal and mechanical stability are of great importance, it is desirable to make the cavity tube out of quartz. This is done commercially by grinding quartz pipe to the correct i.d. and then fusing a silver film to the inner surface.[23] A wall thickness of $\frac{5}{16}$ in. gives adequate mechanical strength, while the silver coating should be at least 0.001 in. to ensure minimum wall loss.

Because loading by the storage bulb changes the cavity length appreciably, the position of one end plate must be adjustable. For maximum stability the end plate should be mounted rigidly on spacers which are cut to bring the cavity within the range of a fine tuning control. For short term experiments involving frequent storage bulb changes, it is convenient to mount the end plate on a threaded drum with at least a 6-in. diam and a micrometer thread. After adjustment the end plate must be clamped in position.

Fine tuning can be accomplished by a tuning plunger or by coupling reactance to the cavity. A $\frac{1}{4}$-in.-diam plunger protruding through the end plate near the electric field maximum has a tuning sensitivity of approximately 7 kc/sec/in. The tuning range of the plunger is limited to about 70 kc/sec by the fact that it eventually couples in the TM_{111} mode. Other plungers have a tuning coefficient roughly proportional to their cross-sectional areas.

Reactance tuning is a very convenient fine-tuning method, since it is accomplished without mechanical adjustment. If the load coupled to the cavity by a line of impedance Z_0 is varied from Z_0 (i.e., a matched load) to $Z_0(1+jX)$, where $X\ll 1$, then the cavity will be detuned by an amount $\delta\nu_c$ given by[24]

$$\delta\nu_c=(-\beta/2Q)X. \tag{23}$$

β is the coupling coefficient.[25] A simple method for introducing the reactance is by coupling a fraction of the power from the cavity line with a directional coupler whose output is terminated by a crystal diode or a varactor. The reactance of the diode is adjusted by a biasing current. Although this presents a resistive as well as a reactive mismatch to the line, the coupling is sufficiently small so that the resultant power loss is negligible.

Preliminary tuning of the cavity to an accuracy of about 500 cps can be accomplished by conventional reflection techniques, providing care is taken not to vary the load reactance coupled to the cavity during the process. Final tuning is done with the maser itself, as described in Sec. X.

In general, it is necessary to regulate thermally the cavity to limit drift. Thermal sensitivity of the cavity can be reduced by mounting an end plate on metal spacers with a length chosen to shorten the cavity the necessary amount to compensate for the effect of expansion of the cylinder. A typical thermal coefficient for a partially compensated cavity is 1 kc/sec °C.

The system in Fig. 7 uses a cavity designed to provide as much thermal contact as possible among its components in order to reduce the effect of thermal gradients. Insofar as possible, the cavity is thermally isolated from the base structure. The base structure and bell jar assembly form an isothermal enclosure which is mounted so as to provide as little heat conduction as possible to the lower manifold. The bell jar is enclosed in an oven by a thermally insulated aluminum cylinder nesting outside the inner magnetic shield. A second similar oven encloses the first. Thermal control of the two ovens and the neck temperature results in the control of the bell jar temperature to about 0.01°C. Sensing and temperature correction are provided by the circuit shown in Fig. 8.

A word of caution about materials. Due to the requirements for very low magnetic fields, discussed in the next section, great care must be taken to avoid any ferromagnetic materials in the cavity. Stainless steels should not be used since at machined edges magnetically hard spots develop which cannot be removed by annealing. Even nickel plating over a nonmagnetic base material can cause difficulty. All cavity parts should be checked for magnetism during assembly.

[22] C. G. Montgomery, *Techniques of Microwave Measurement* (McGraw-Hill Book Company, Inc., New York, 1947), p. 297.
[23] Syncor Products Company, Malden, Massachusetts.
[24] Reference 22, p. 291.
[25] E. L. Ginzton, *Microwave Measurements* (McGraw-Hill Book Company, Inc., New York, 1957), p. 290.

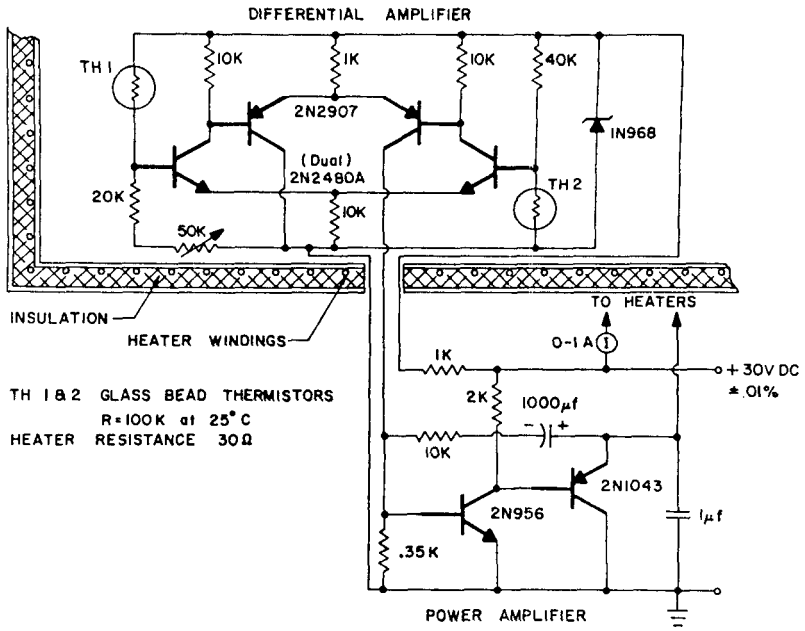


FIG. 8. Temperature sensing and servo systems.

VII. MAGNETIC SHIELDS

As a result of the quadratic field dependence of the transition $(F=1, m=0)\rightarrow(F=0, M=0)$, it is desirable to reduce the ambient field at the storage bulb to a very low value, preferably 1 mOe or less. The most satisfactory way to accomplish this is through the use of magnetic shields. Ideally, the shields completely eliminate the magnetic field, and then a small uniform field is applied with a solenoid.

In practice, the low-field limit at which the maser operates is determined by either (a) the magnitude of the residual field which may lie in an undesirable direction, or (b) the gradients in the residual field which can cause prohibitive relaxation as the Zeeman frequency is reduced. In the first case, the applied field must be large enough so that the oscillating moment is substantially parallel to the oscillating field. However, even then, owing to (b), gradients of the residual transverse field can cause relaxation owing to the random motion of the atoms. As discussed in KGR, this relaxation rate is uniform at low fields and drops rapidly when the Zeeman frequency exceeds the mean "rattle frequency" of the atoms in the bulb. Consequently, large gradients necessitate a relatively large uniform field. In poorly demagnetized shields, this means a field of 5 mOe or more.

The magnetic field dependence of the transition of interest is

$$\nu=\nu_0+2750\,H^2 \text{ cps}, \tag{24}$$

where H is the applied field in oersted.

The fractional shift of the frequency due to a small change in field is

$$\delta\nu/\nu=3.9\times10^{-6}\,H^2\,(\Delta H/H). \tag{25}$$

The maser has been operated at fields as low as 6×10^{-5} Oe, where field stability is no longer a significant problem.

In principle, it is possible to shield external fields to a high degree by the use of successive concentric shields. For the case of a uniform field at right angles to a series of three infinitely thin concentric cylinders, the shielding factor, i.e., the ratio of the applied external field H_0 to the net internal field H_i, is given[25a] by

$$\frac{H_0}{H_i}=\frac{1}{2}\left[\frac{\mu_1 t_1}{r_1}\,\frac{\mu_2 t_2(s_{12}-s_{12}^2/2r_2)}{r_2^2}\,\frac{\mu_3 t_3(s_{23}-s_{23}^2/2r_3)}{r_3^2}\right], \tag{26}$$

where μ is the permeability of the shields, and the dimensions are as shown in Fig. 9. It is apparent from this that with moderately high permeability, very high shielding factors are obtainable.

A shield configuration which has proven successful in use consists of the following: The shields are composed of three coaxial cylinders fabricated from 0.025-in. Moly Permalloy. The innermost is 14 in. in diameter and 30 in. long; the middle shield is 16 in. in diameter and 32 in. long; and the outer shield is 18 in. in diameter and 36 in. long. The seams are spot-welded over $1\frac{1}{2}$-in. laps. (Mu Metal has also been used but is not as satisfactory as Moly Permalloy.) The inner two shields have end caps spun from Moly Permalloy. The shields were annealed by the manufacturer.[26] A 4-in.-diam hole through the end caps at one end does not appreciably deteriorate their performance. The static field is applied by means of a solenoid on an aluminum form just inside the inner shield. It is useful to have taps on the solenoid,

[25a] H. P. Wills, Phys. Rev. **9**, 208 (1899).

[26] Allegheny Ludlum Steel Corporation, Brackenridge, Pennsylvania.

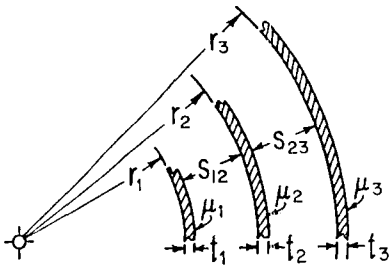


FIG. 9. Geometry of concentric cylindrical magnetic shields.

so that a few turns at each end can be energized separately. These coils can be used to reduce end effects. They are also useful for providing a small gradient for the magnetic-quenching tuning technique which will be described in Sec. X.

Proper demagnetization is critical to the shield's performance. A 60-cps demagnetizing current is passed through the center of the shields, using the vacuum can as a conductor. The resulting field is azimuthal, so that the demagnetizing flux lies completely in the shields. Current is obtained from a transformer made by winding seven turns of No. 4 copper cable around the toroidal core of a 20-A autotransformer. For fine current control, two Variacs in series are used to drive the transformer. Best results are obtained by slowly raising the current to its maximum value and then decreasing it to zero, taking about 1 min altogether.

Current up to 10^3 A is required to saturate the shields, but such strong demagnetizing current is seldom necessary. The best results are obtained by demagnetizing with maximum current of about 150 A. The residual fields appear to be caused primarily by spots of remnant magnetization in the shields. These spots vary in position and intensity from one demagnetization to another but are most commonly found along the seam of the innermost shield. The residual transverse field is generally quite nonuniform. Its maximum value is typically 10^{-5} to 10^{-4} Oe, based on measurements of the low-field radiation lifetime. Lower values are obtained from time to time.

By the demagnetization procedure described, very large external fields can be shielded. For instance, a powerful horseshoe magnet placed directly on the shields will cause a serious disturbance, but the disturbance is almost completely eliminated by demagnetization. Likewise, fields due to nearby large magnets on VacIon pumps can be shielded without undue difficulty. Apparently, the chief function of demagnetization is to raise the flux level to a point where the shield magnetization can favorably reorient itself.

Although large permanent fields are effectively reduced by the shields, they are much less effective at reducing small variations in the field, since the incremental permeability drops off rapidly at low fields. Addition of merely another shield does not help matters, since its shielding contribution is counterbalanced by the decrease in permeability of the inner shields which now operate at a lower field. Measurements made with a magnetometer indicate a shielding factor for small changes in the ambient field of about 600. (With the innermost shield made of Mu Metal, this figure is about 100.) During operation at favorable times, the field fluctuations at the storage bulb are about 10^{-7} Oe, as measured with the maser when it is operating on the field-dependent transition with a 10-sec sampling time. The rms fluctuations are almost 10 times smaller.

VIII. STORAGE BULB

A. Bulb Design

A number of considerations are involved in designing the storage bulb. The dimensions must be chosen so that the restriction on q given by Eq. (12) is satisfied, and preferably so that q is as small as possible. For this reason, it is desirable to maximize the filling factor[27] η' given by

$$\eta' = \frac{\langle H_z \rangle_{\text{bulb}}^2}{\langle H^2 \rangle_{\text{cavity}}} \frac{V_b}{V_c}. \tag{27}$$

A spherical bulb is the easiest to construct, and η' for that case may be determined using the plot of η in KGR (note the correction in footnote 27). For a cylindrical bulb of radius r and length l in a cylindrical cavity of radius R and length L, the following expression can be derived:

$$\eta' = \frac{32}{\pi^2} \frac{J_1^2(kr)}{J_0^2(kR)} \frac{\sin^2(\pi/2)(l/L)}{[1+(\pi/kL)^2]} \frac{L}{lk^2R^2}. \tag{28}$$

$J_n(x)$ is the nth-order Bessel function, $J_1(kR)=0$. From this it follows that η' is maximized by the choice $r=0.52R$, $l=0.74L$. Although the above formula neglects distortion of the fields in the cavity by the storage bulb, more detailed calculations show that η' is not significantly altered, providing the dimensions inserted in (28) refer to the actual dimensions of the cavity when it is loaded by the storage bulb. Two other considerations which enter the bulb design are the desired storage time and the collision rate. In order to avoid increasing q, the storage time must not be made long compared to the radiation lifetime. Because the wall shift is proportional to the collision rate, it is desirable to keep the bulb dimensions as large as possible. For this reason, a bulb somewhat larger than needed to maximize Eq. (7) may be desirable. The storage time T_b and mean distance between wall collisions λ are given by

$$T_b = 4V_b/\bar{v}A_a, \tag{29}$$

$$\lambda = 4V_b/A_b. \tag{30}$$

Here V_b=storage bulb volume, A_b=surface area of

[27] The symbol η' is chosen since η is the conventional nmr symbol for the filling factor. Unfortunately, η was used in KGR for the ratio of the field averages. The vertical scale of the plot of η in KGR has an error and should be divided by two.

bulb, A_a = area of the exit aperture, v = average velocity $= (8kT/\pi m)^{1/2}$. These formulas are derived in Appendix A. If a collimating tube is used instead of a simple aperture, then flow is reduced by a factor K, so that

$$T_b = (1/K)(4V_b/\bar{v}A_a). \quad (31)$$

A table of values of K for tubes of various sizes is given by Dushman.[28] A collimating tube will also slightly alter λ, and a correction for this is presented in Appendix A.

The storage bulb is generally made of fused quartz with a diameter between $3\frac{1}{2}$ and $6\frac{1}{2}$ in. and a wall thickness of about 0.040 in. Atoms enter the bulb through a tube whose diameter and length are chosen for the desired lifetime. If it is important to make efficient use of the beam flux, the entrance aperture can be enlarged and several thin-walled tubes introduced to make a simple collimator.

B. Wall Coating

A number of materials have been used successfully for coating the storage bulb wall. At room temperature, both long-chain paraffin[2] and Dri-Film[15] surfaces have been used and yield times T_2' of about 0.3 sec in a 16-cm-diam bulb. The limit appears to be set by the chemical reaction of hydrogen with the wall. This problem is diminished by the use of Teflon, a fluorocarbon which also has the advantage of an appreciably smaller wall shift than paraffin or Dri-Film. At 35°C, a 16-cm-diam Teflon-coated storage bulb has a time T_2' of abut 3 sec and a fractional wall shift[3] of -2.1×10^{-11}. At higher temperatures, there is a decrease both in the wall shift and in the lifetime. At 100°C, T_2' is reduced by a factor of approximately 2, and the fractional wall shift is reduced by a factor of 3. The nature of the wall relaxation is discussed in Ref. 11.

The method for applying a Teflon film has been previously reported.[29] However, new techniques have been developed since then which simplify the procedure. The coating technique described in Ref. 29 is for TFE Teflon (du Pont TFE clear finish 852-201). The procedure described is improved by the use of hot white fuming nitric acid as the cleaning agent, rather than glass-cleaning solution. However, the entire coating process is simplified by the use of FEP Teflon (du Pont FEP Teflon product Code 120), since this forms a more uniform coat and several coats can be applied successively.[11] The FEP suspension is applied as described in Ref. 29. However, during the fusing process clean air is circulated through the bulb. Decomposition and waste products are thereby oxidized and removed as gases. The fusing oven temperature is brought up to 360°C during the course of an hour, held at that temperature for about 20 min, and then cooled, during the course of another hour. The film is strong and transparent and is inert to hot fuming nitric acid; however, it can be removed bodily by heating in the bulb a solution of 20% HF and 20% HNO_3.

The collimating tube in the neck of the bulb can be machined from solid Teflon, or it can be made by coating a pyrex plug. The latter procedure is preferable, since solid Teflon tends to outgas for several days.

Although the wall-coating procedure described here is quite reliable, the surfaces are not entirely inert, and it is possible that this is due to a contaminant in the Teflon. New surfaces are currently being investigated.

IX. ELECTRONICS

Systems for processing the maser signal will not be described in detail, since the procedure depends on the type of measurement to be made, and often to a large extent on the available equipment. However, there are a few points of general interest.

It is important to keep the maser operating into a constant load since reactance changes will cause frequency pulling, as discussed in Sec. VI. An isolator is invaluable for this purpose. The isolator is also helpful in decoupling two masers which are operating into converters powered by a common local oscillator. Isolation requirements are stringent, if frequency locking of the masers is to be avoided. This can be seen by the following argument: The initial phase of a maser is random, since the maser originally turns on due to noise signals. The phase of the maser is constantly perturbed by thermal noise power lying within the resonance bandwidth $kT\Delta\nu_r$. A coherent signal of less power cannot lock the maser, since random fluctuations due to noise are sufficient to randomize the maser phase with respect to the incident signal. Therefore, isolation sufficient to reduce the unwanted signal to less than $kT\Delta\nu_r$ will completely prevent locking. This required 80 to 90 dB of isolation. Fortunately, isolators with less than 1 dB of insertion loss and up to 60 dB of isolation are available.[30] It is usually possible to obtain the remaining 30 dB of isolation through the use of balanced mixers and balanced power dividers. If the maser frequencies are offset, the isolation requirements are considerably decreased. A number of systems for processing the maser signals are described in Refs. 15, 16, 31, and 32.

X. TUNING METHODS

Tuning the maser involves setting the magnetic field to a given value, and tuning the cavity. The magnetic field is easily measured by applying a small audio signal to the bulb by a single turn of wire placed

[28] S. Dushman, *Scientific Foundations of Vacuum Technique* (John Wiley & Sons, Inc., New York, 1962), p. 93.
[29] H. C. Berg and D. Kleppner, Rev. Sci. Instr. 33, 248 (1962).
[30] Ferrotec, Inc., Newton, Massachusetts; also Mel. Labs., Palo Alto, California.
[31] Frequency 1, 28 (1963).
[32] R. F. C. Vessot, H. Peters, and J. Vanier, Frequency 2, 33 (1964).

around the cavity for that purpose. When the signal is at the Zeeman frequency, the oscillation level is markedly changed; usually the power is decreased. Alternatively, resonance of the Zeeman signal may be detected by using the fact that the hyperfine frequency is "pulled" when the Zeeman signal is close to resonance. The pulling is zero at resonance. Because the Zeeman line is as narrow as 1 cps under favorable conditions, it is convenient to apply a standard Zeeman frequency and then trim the magnetic field to resonance.

Cavity tuning is a more involved procedure. Normally, it is done using two masers, one acting as a frequency reference while the other is tuned. The most accurate tuning methods make use of the frequency pulling effect described in Eq. (21). The resonance width is varied by a method described below, and the cavity is tuned until the maser frequency is unperturbed by variations in the linewidth.

The following two-part method has proven useful:

(1) The resonance *width* is altered by operating the maser under two separate conditions (which we will denote by A and B), and the ratio $R=\Delta\nu_A/\Delta\nu_B$ is determined.

(2) The difference in pulling of the maser *frequency* under conditions A and B is measured. Knowledge of this figure, along with R, is sufficient to set the cavity to the correct frequency.

Here are details of the method:

(*1*) *Determination of R.* This is accomplished by offsetting the cavity frequency by some amount, typically 1 kc/sec. (The actual value of the offset need not be determined.) One simple way of accomplishing this is by switching an extra current through the diode of the reactance tuner described in Sec. VI. The amount by which the frequency of the maser changes as the cavity frequency is offset is measured. We denote this quantity by δ. δ is measured for each of the two operating conditions, A and B. From Eq. (22), it is easy to show that

$$R=\delta_A/\delta_B. \tag{32}$$

(*2*) *Setting the cavity.* The cavity tuning is now left undisturbed, and the amount by which the frequency is shifted when the maser's condition is changed from A to B is measured. We denote this quantity by Δ.

Next, the maser is returned to condition A, and the frequency of the cavity is trimmed so that the maser frequency is increased by an amount

$$\delta_{\text{trim}}=\Delta/(R-1). \tag{33}$$

At this point, the cavity is tuned. This can be checked by confirming that there is no frequency shift in going from A to B.

Any method for varying the resonance linewidth without shifting the resonance frequency can be used for the above procedure. However, for good sensitivity, the linewidth must be changed appreciably; R (or $1/R$) must be large compared to 1.

One method for varying the linewidth is to use spin exchange broadening. States A and B correspond to the maser operating near maximum and minimum flux, respectively. In this case, R is close to the quantity r introduced in Sec. II. [Since R is known from the result of Eq. (32), this method also allows us to find the quantity parameter q.] Another method for broadening the line is to apply an inhomogeneous magnetic field, providing care is taken not to shift the average field appreciably. The latter method has the advantage of convenience, but does not afford as much sensitivity as the former method.

A third method for cavity tuning is to plot ν versus ν_c for different values of $\Delta\nu_r$, as given in Eq. (21). The result is a family of straight lines intersecting at the point $\nu_c=\nu_0$.[31,33]

An important feature of all the above methods for cavity tuning is that the effect of a spin exchange frequency shift is eliminated; the cavity will be mistuned by just the right amount to cause the maser to oscillate on the true hyperfine frequency, regardless of flux.[3,9,32]

How well the maser can be tuned by these or other methods depends on how far q is below the critical value. Unless q is substantially below the maximum permissible value, the operating range is small, and the results will not be satisfactory.

ACKNOWLEDGMENTS

The authors wish to thank the following for their contributions to the work described in this paper: A. O. McCoubrey, H. G. Robinson, E. Recknagel, E. N. Fortson, B. S. Mathur, and L. Mueller.

APPENDIX: LIFETIME AND MEAN COLLISION DISTANCE OF ATOMS IN A STORAGE BULB

We assume that we have an enclosed region of space (i.e., a storage bulb) in which atoms are introduced and from which they effuse through a small aperture. It is assumed for the present that the mean free path is sufficiently long and the aperture is sufficiently small so that the density of atoms throughout the bulb is constant. Let V_b = volume of the bulb, A_b = total area of the enclosed surface, A_a = area of aperture, $\bar{v}$ = mean velocity of atoms, N = total number of atoms in the bulb. Then the rate of loss of atoms from the bulb is $\frac{1}{4}(N/V_b)\bar{v}A_a$, and this can be shown to be equal to N/T, where T is the mean lifetime of atoms in the bulb. Hence

$$T=4V_b/\bar{v}A_a. \tag{29}$$

It follows that the mean distance traveled by an atom before escape is $L=T/\bar{v}=4V_b/A_a$. However, we also

[33] J. Vanier and R. F. C. Vessot, Appl. Phys. Letters 4, 122 (1964).

have $L=n\lambda$, where n is the mean number of collisions an atom makes before escape, and λ is the mean distance between collisions. It can be shown that $n=(A_b+A_a)/A_a \simeq A_b/A_a$. Equating these two expressions for L then yields

$$\lambda = 4V_b/A_b. \tag{30}$$

Application of this formula to a few simple geometries gives:

Sphere of radius R: $\lambda=4R/3$.

Cylinder of radius R and length L: $\lambda=2R/(1+R/2L)$.

Cube of side L: $\lambda=2L/3$.

The use of a collimating tube at the bulb entrance will increase the lifetime as described by Eq. (31) in the text. It also modifies λ, since the derivation above assumes a constant density throughout the bulb. In the neck, the density falls uniformly from the equilibrium density to 0, so that the average density is one-half that in the bulb. Since the rate of wall collision is proportional to the density, the mean wall collision rate is increased by a factor $1+\frac{1}{2}(V_t/V_b)$, where V_t is equal to the volume of the collimating tube and we have assumed $V_t \ll V_b$. The distance between collisions is then given by

$$\lambda = \frac{4V_b}{A_b} \frac{1}{1+\frac{1}{2}(V_t/V_b)}.$$

IEEE TRANSACTIONS ON ULTRASONICS, FERROELECTRICS, AND FREQUENCY CONTROL, VOL. 39, NO. 2, MARCH 1992

Environmental Sensitivities of Quartz Oscillators

Fred L. Walls and Jean-Jacques Gagnepain

Abstract—**The frequency, amplitude, and noise of the output signal of a quartz oscillator are affected by a large number of environmental effects. The paper examines the physical basis for the sensitivity of precision oscillators to temperature, humidity, pressure, acceleration and vibration, magnetic field, electric field, load, and radiation. The sensitivity of quartz oscillators to radiation is a very complex topic and poorly understood. Therefore only a few general results are mentioned. The sensitivity to most external influences often varies significantly both from one oscillator type to another and from one unit of a given type to another. For a given unit, the sensitivity to one parameter often depends on the value of other parameters and on history. Representative sensitivity to the above parameters will be given.**

I. Introduction

QUARTZ OSCILLATORS are a fundamental element in many areas of frequency metrology affecting applications such as communication and navigation. Their frequency, output level, amplitude noise, and phase noise are generally critical parameters that determine the overall performance of a system. In many applications their performance is significantly less than that obtained under ideal environmental conditions. In this paper we will briefly outline the physical basis and representative values for the sensitivity of quartz oscillators to environmental parameters. The most important of these are temperature, humidity, pressure, acceleration and vibration, magnetic field, electric field, load, and radiation. The sensitivity of quartz oscillators to radiation is a very complex and poorly understood topic and therefore only a few general results are mentioned. For a given oscillator the sensitivity to one parameter often depends on the value of other parameters and on the history of the device. It is often difficult to separate the influence of one parameter from that of another. Several methods for characterizing the environmental sensitivities of both quartz resonators and oscillators are discussed in [1]. Much more effort is needed in this area.

II. Model of a Quartz Oscillator

Fig. 1 shows a simplified diagram of a quartz oscillator. The well known oscillation conditions are that the phase around the loop be an integer multiple of 2π and that the loop gain be 1 [2]–[5]. Small changes in the loop phase, $d\phi(f)$, are compensated by a change $\Delta\nu$ in oscillation frequency of

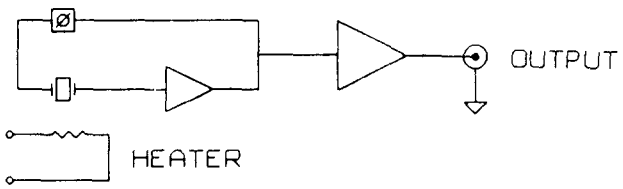


Fig. 1. Simplified block diagram of a quartz-oscillator. Loop gain=1 and loop phase = $n2\pi$ $n = 0, 1, 2 \ldots$

$$\Delta\nu/\nu_0 = \frac{1}{2Q}\left(1 + (2fQ/\nu_0)^2\right)^{-1/2} d\phi(f) \qquad (1)$$

where f is the Fourier frequency offset from the oscillation frequency ν_o, and Q is the loaded Q-factor of the resonator. Small changes in the frequency of the resonator are directly translated into changes in the frequency of the oscillator since the rest of the electronics is generally very broadband compared to the resonator so that the complete oscillator dependence is given by

$$\Delta\nu/\nu_{0_{\text{oscillator}}} \approx d\nu/\nu_{0_{\text{resonator}}} + \frac{1}{2Q}\left(1 + (2fQ/\nu_0)^2\right)^{1/2} d\phi(f). \qquad (2)$$

Equation (2) provides the basis for understanding how the various environmental conditions affect both the short-term and the long-term frequency stability of the oscillators.

III. Changes Within the Resonator

Environmental effects that change the frequency of the resonator have been investigated by many people in much more detail than can be described in this paper. The most important environmentally driven changes within the resonator are driven by changes in temperature, level of excitation (RF amplitude), stress, adsorption, and desorption of material on the surfaces, acceleration and vibration, radiation, electric field, and magnetic field.

The frequency shifts due to these effects are universal to the extent that the loop is broadband and unchanged by the environmental factors. Different oscillator designs have an influence on the apparent isolation of the resonator from the environmental parameters.

A. Temperature and Temperature Changes

Temperature variations change the value of the elastic constants and the dimensions of the resonator. The resulting change in resonator frequency with temperature varies greatly

Manuscript received March 5, 1991; revised and accepted September 13, 1991.

F. L. Walls is with the National Institute of Standards and Technology, Time and Frequency Division, 325 Broadway Boulder, CO 80303.

J.-J. Gagnepain is with the Laboratoire de Physique et Metrologie des Oscillateurs du C.N.R.S. associé à l'Université de Franche-Comte-Besancon 32, Avenue de l'Observatoire, 25000 Besançon, France.

IEEE Log Number 9105587.

0885-3010/92$03.00 © 1992 IEEE

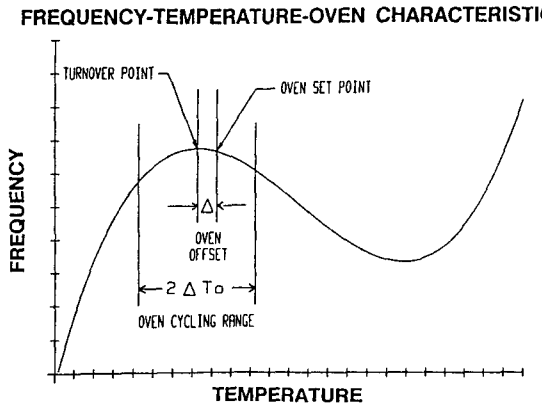


Fig. 2. Idealized frequency versus temperature curve for a quartz resonator. The turnover point and the oven offset from turnover are indicated. From [16].

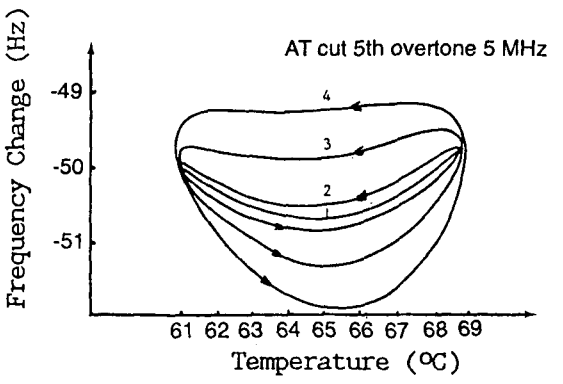


Fig. 3. Frequency versus temperature for a 5-MHz AT-cut resonator. The static temperature curve is shown in curve 1. The other curves show the response with ±4°C sinusoidal temperature cycling at a sweep frequency of 9.2×10^{-5} Hz-curve 2, 3.7×10^{-4} Hz-curve 3, and 7.4×10^{-4} Hz-curve 4. A model fit to the curves yields a dynamic temperature coefficient of -1.3×10^{-5} s/°C. From [10].

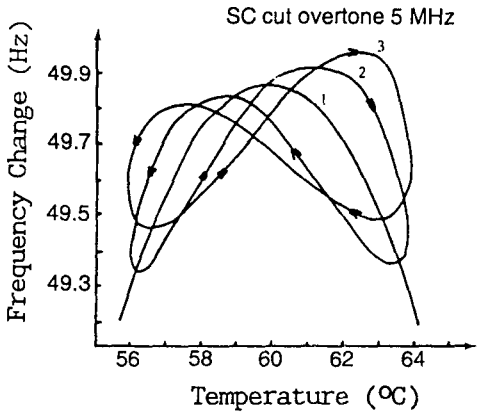


Fig. 4. Frequency versus temperature for a 5-MHz SC-cut resonator. The static temperature curve is shown in curve 1. The other curves show the response with ±4°C sinusoidal temperature cycling at a sweep frequency of 9.1×10^{-4} Hz-curve 2, and 1.8×10^{-3} Hz-curve 3. A model fit to the curves yields a dynamic temperature coefficient of $3 \times 10 - 7$ s/°C. From [10].

with crystallographic cut and orientation [6]–[21]. A number of different cuts compensate for the dependence of frequency on temperature. Fig. 2 shows a typical static frequency versus temperature curve for precision quartz resonators [16]. The actual values depend on the resonator cut, overtone, frequency, diameter, and mounting technique. Temperature changes and temperature gradients often cause frequency changes that are large compared to the slope of the static curve [6]–[22]. Typical coefficients for the frequency–temperature effect for 5-MHz resonators are

$$\Delta\nu/\nu_0 = 10^{-9}\Delta T^2 - 10^{-5}\, dT/dt \qquad (3)$$

for fifth-overtone AT-cuts, and

$$\Delta\nu/\nu_0 = 10^{-9}\Delta T^2 + 10^{-7}\, dT/dt \qquad (4)$$

for third-overtone SC-cuts. Here ΔT is the temperature difference in K from turnover and dT/dt is the rate of change of temperature in K/s.

The dynamic temperature effect leads to hysteresis in the experimental measurements of temperature coefficients as shown in Figs. 3 and 4. This effect makes it difficult to locate the exact turnover point. This in turn leads to oscillators with finite frequency changes even for slow, small temperature excursions, especially for AT-cut resonators. Table I shows typical static temperature coefficients for an AT-cut resonator as a function of the error in setting the oven to the exact turnover point. Table II shows typical static temperature coefficients for an SC-cut resonator as a function of the error in setting the oven to the exact turnover point. The actual change in temperature may be driven by changes in atmospheric pressure and/or humidity, which change the thermal conductance and temperature gradients within the oscillator package [22], [23]. Deposited radiative energy can also change the temperature of the resonator. The performance potential of AT-cut resonators is extremely difficult to attain due to their very high dynamic temperature coefficient. Sulzer was the first to attain a performance $\sigma_y(\tau) \sim 3 \times 10^{-13}$ from AT-cut resonators, and it was many years before it was generally realized that the reason for the excellent performance was an oven design with extremely low thermal transients [25]. The reduction in the dynamic temperature coefficient for SC-cut resonators as compared to the earlier AT- and BT-cut resonators represents a major advance in the practical application of quartz oscillators in nonideal environments using only simple ovens. Much better thermal performance can be obtained using multiple ovens [26], aged high-performance thermistors [27], and compensated oven designs [28]. The trade-off is increased complexity, size, weight, and cost.

Temperature generally does not affect the phase noise or short-term frequency stability of an oscillator (except in cases where activity dips are involved, see below.) However, in cases when the sustaining electronics is not temperature controlled, slight changes in gain and noise figure with temperature will be reflected in the output phase noise. Reviews of the correlation of output phase noise with the noise performance of the sustaining stage and output amplifier are given in [2]–[5], [29].

Another important effect, especially with AT-cut resonators, is activity dips. Activity dips are due to the accidental overlap

TABLE I
TYPICAL STATIC FREQUENCY VERSUS TEMPERATURE COEFFICIENTS FOR AN AT-CUT RESONATOR WITH A TURNOVER TEMPERATURE OF 85°C AS A FUNCTION OF OVEN PARAMETERS[a]

$\Delta\nu/\nu_o = 3 \times 10^{-8}\Delta T^2$				
Oven Offset	Oven Change (mK)			
(mK)	±100	±10	±1	±0.1
0	3×10^{-10}	3×10^{-12}	3×10^{-14}	3×10^{-15}
1	3×10^{-10}	5×10^{-12}	2×10^{-13}	6×10^{-15}
10	5×10^{-10}	2×10^{-11}	6×10^{-13}	6×10^{-14}
100	2×10^{-9}	6×10^{-11}	6×10^{-12}	6×10^{-13}

[a] Frequency stability of 1×10^{-13} requires $\Delta T/dt < 10$ nK/s due to the dynamic temperature effect given in (3).

TABLE II
TYPICAL STATIC FREQUENCY VERSUS TEMPERATURE COEFFICIENTS FOR AN SC-CUT RESONATOR WITH A TURNOVER TEMPERATURE OF 85°C AS A FUNCTION OF OVEN PARAMETERS[a]

$\Delta\nu/\nu_o = 3 \times 10^{-8}\Delta T^2$				
Oven Offset	Oven Change (mK)			
(mK)	±100	±10	±1	±0.1
0	4×10^{-11}	4×10^{-13}	2×10^{-15}	4×10^{-17}
1	4×10^{-11}	6×10^{-13}	2×10^{-14}	8×10^{-16}
10	6×10^{-11}	2×10^{-12}	8×10^{-14}	8×10^{-15}
100	2×10^{-10}	8×10^{-12}	8×10^{-13}	8×10^{-14}

[a] Frequency stability 1×10^{-13} requires $\Delta T/dt < 330$ nK/s due to the dynamic temperature effect given in (4).

of some other vibration mode of the resonator with the main resonance mode and usually occurs over a narrow range of temperatures. This coupling to the unwanted mode leads to increased losses, an increase in motional resistance, and hence a reduction in the oscillator amplitude. The frequency of the resonator is also pulled by the coupling to the other mode, which usually has much higher sensitivity to temperature than the primary mode of oscillation. This leads to temperature–frequency coefficients that vary rapidly with temperature over very narrow ranges in temperature of the resonator [30]–[34]. Fig. 5 shows the frequency-versus-temperature performance of the primary clock in the Ginga satellite [33]. The nonlinear thermal coefficient near 18°C make it very difficult to model and accurately recover clock timing. SC-cut resonators and some types of lateral field resonators show much reduced incidence of activity dips and their associated quirks in the temperature coefficients [34]–[36].

The dynamic temperature effect and the possibility of an activity dip significantly complicates the specification and measurement of oscillator temperature coefficients. For critical applications the frequency must be measured over the entire operating temperature range using a model of the actual temperature profiles.

B. RF Excitation Level

The frequency of the resonator is also a function of amplitude of the signal level as shown in Fig. 6 [16], [18]–[20], [34], [37]–[39]. The sensitivity to this effect, usually called the amplitude–frequency effect, is a function of a large number of

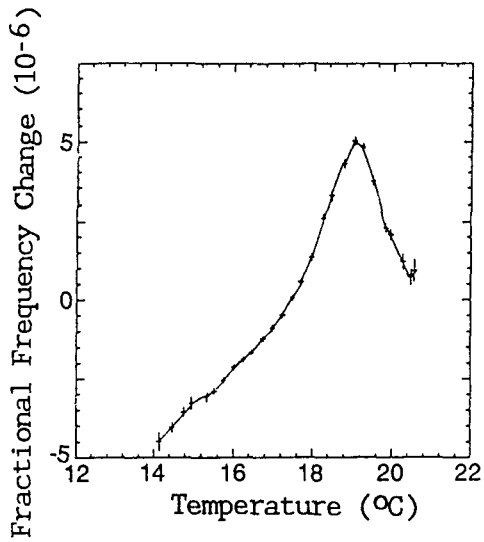


Fig. 5. Fractional frequency versus temperature for the primary timing clock of the Ginga satellite [33].

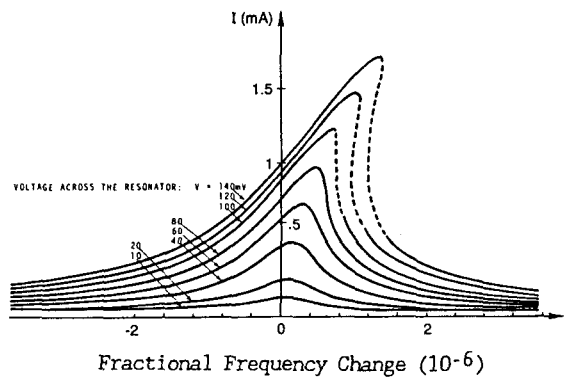


Fig. 6. Fractional frequency versus the RF excitation level for a 5-MHz AT-cut resonator [19].

resonator fabrication details. Fig. 7 shows the dependence on blank curvature with most other parameters held constant [34]. There is also a dependence on the electrical size since both the radius of curvature and electrical dimension define the energy trapping in the resonator. Typical sensitivities to this parameter range from approximately $10^{-9}/\mu$W for fifth-overtone AT- or BT-cut resonators to parts in $10^{-11}/\mu$W for third-overtone, SC-cut resonators at 5-MHz. This effect scales approximately as $1/Q$ or $1/\nu_o$. The primary environmental drivers for this effect are temperature, humidity, or radiation changing the excitation level through interaction with the automatic gain control (AGC) and the gain of the sustaining stage. Changes in the dc supply voltage can also affect the amplitude of oscillation by changing the AGC circuitry.

C. Stress (Force)

Stress on the resonator blank changes the resonance frequency through the nonlinear elastic behavior of the crystal [11]–[21], [34], [40]–[44]. Stress is transmitted to the

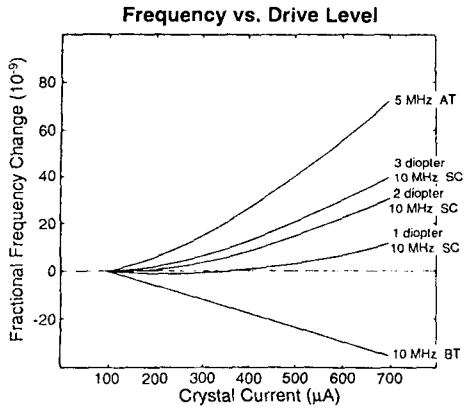


Fig. 7. Fractional frequency versus the RF excitation level for AT-, BT-, and several SC-cut resonators [16], [33].

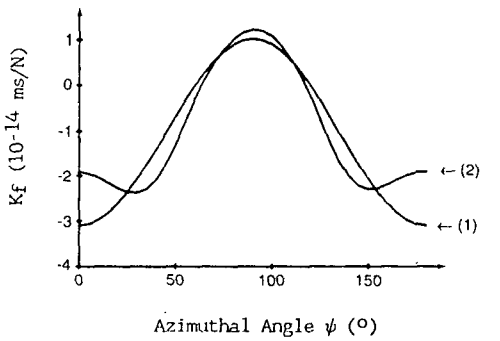


Fig. 8. Force-frequency coefficient for diametrically opposed forces in the plane of an AT-resonator as a function of the angle between the force and the azimuthal axis. Curve 1 shows the result of the classical isotopic model, while curve 2 shows the result of an analytical anisotropic model: $K_f = ((2td/V)(\delta\nu_0/\nu_0))$ where t is the thickness, d the diameter of the disk, V the acoustic wave velocity, and f the magnitude of the force. From [24].

resonator through the mounting structure. It can originate from temperature-driven dimensional variations in the vacuum enclosure, changes in the pressure surrounding the vacuum enclosure, changes in the magnetic field causing a change in the mounting force due to the use of magnetic components, or from changes in the body forces due to acceleration and vibration as described in Section III-C. The stress on the resonator, due to the use of electrodes directly plated onto the resonator, can change with resonator drive, temperature, radiation exposure, and time. The stress due to the electrode has been estimated [11]–[13], [20], [21], but the changes with environmental effects such as vibration and temperature cycling are very difficult to estimate. Fig. 8 shows the change in frequency of a traditional AT-cut plate due to diametrically opposed forces in the plane of the resonator as a function of the angle between the applied force and the x axis [24]. The effect of electrode deposition is absent from the BVA electrodeless resonators [20], [21]. The influence due to the mounting stress can be minimized in traditional resonators by the proper choice of the mounting angle. Mounting stress in BVA resonators is minimized by both the choice of mounting angle and by the use of a supporting ring machined from the same monolithic piece of quartz as the resonator [20], [21]. The sensitivity to mounting force scales approximately as the reciprocal of the resonator plate cross section or ν_o^2.

D. Adsorption–Desorption

Changes in the quantity or distribution of molecules on the surface of the resonator can lead to very large changes in the frequency of the resonator [6], [16], [46]–[48]. One monolayer added to a 5-MHz resonator amounts to roughly 1 ppm change in the frequency [16]. Major drivers of changes in the background pressure and in the movement of adsorbed gasses are temperature changes and enclosure outgassing or leaks. This effect scales approximately as the reciprocal of the plate thickness or ν_o.

The background pressure of helium inside the resonator enclosure can significantly increase if the vacuum enclosure is glass and the resonator is operated in an environment with large amounts of helium. A typical helium leak rate for a glass enclosure operating at 80°C in a pure helium environment is 5×10^{-3} Pa/s. Even in air at 80°C the helium builds up at a rate of approximately 2×10^{-8} Pa/s or about 0.7 Pa/yr [49]. The use of metal enclosures greatly reduces the helium leak rate, but most metals outgas significant amounts of hydrogen and may contribute to the drift of some oscillators [29], [49]. A typical sensitivity for an AT-cut resonator to a nonreactive gas is 10^{-7}/Pa (0.7×10^{-10}/Torr). Ceramic enclosures that reduce the helium and hydrogen leak rates to negligible values have been developed [46]. There also is some question as to what portion of the residual phase noise in quartz resonators is due to time varying rates of collisions with the background gas [48].

E. Acceleration and Vibration

Although a large shock and/or vibration can change the long-term frequency of the resonator, the dominant effect is usually the instantaneous change in the frequency of the resonator due to changes in the stress applied to the resonator through the mounting structure [6], [14], [16], [18]–[21], [50]–[57]. The frequency change depends on orientation and is linear with applied acceleration or vibration up to approximately 50 g (g is the acceleration due to gravity) ([19]. See Fig. 9. The maximum sensitivity is typically of order 2×10^{-9}/g. Significant effort has been expended in minimizing this effect through compensation [52]–[54], mounting techniques [20], [21], [49], [50], [52], and resonator fabrication techniques [20], [21]. The net sensitivity for specially compensated or fabricated oscillators ranges from approximately 10^{-11} to 3×10^{-10}/g. The change in frequency due to inversion "2g tipover test" is often biased due to changes in the temperature gradient as shown in Fig. 10. Although not commonly mentioned, magnetic field sensitivity can also bias the test since the magnetic field effect also is a function of position and motion [58]–[60].

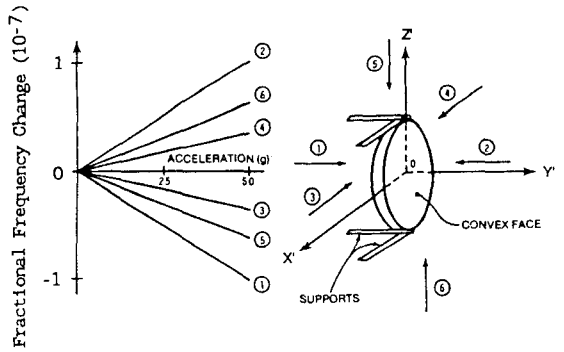


Fig. 9. Typical sensitivity of a AT-cut resonator to acceleration as a function of direction [19].

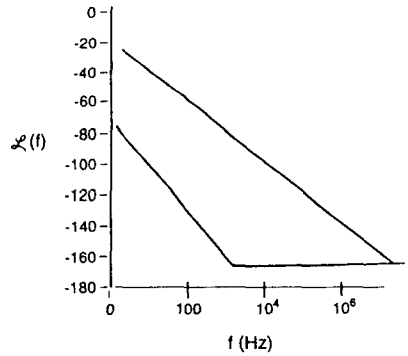


Fig. 11. Quiescent phase noise of a high quality 100-MHz oscillator and the induced phase noise from the application of sinusoidal acceleration at amplitude 2 g. An acceleration sensitivity of 2×10^{-9}/g has been assumed.

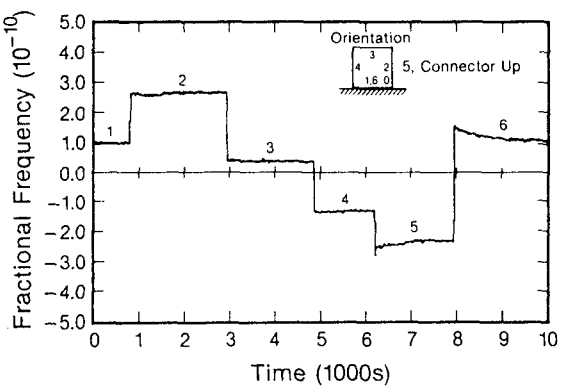


Fig. 10. Fractional frequency of a precision oscillator as a function of orientation. Position 1 and 6 are the same. Note the slow change of frequency immediately after the rotation. This is likely due to small changes in the temperature gradient within the oscillator [23].

The single sideband phase noise, $\mathcal{L}(f)$, of a resonator subjected to vibration is increased by an amount

$$\mathcal{L}(f) = \frac{1}{4}(\nu^2/f^2)\Gamma^2 A^2 \tag{5}$$

where Γ is acceleration sensitivity and A is the applied acceleration. Fig. 11 shows the quiescent phase noise of a 100-MHz oscillator and that obtained with $\Gamma = 2 \times 10^{-9}$/g and $A = 2\cos 2\pi ft$. The increase in phase noise over that obtained under quiescent conditions is approximately 70 dB at a Fourier frequency of 10 kHz. Even with $\Gamma = 1 \times 10^{-11}$/g the degradation would be about 24 dB.

F. Radiation

Radiation interacts with the resonator in many ways. Although not fully characterized for each resonator type and oscillator, it is possible to list some common aspects. A more detailed summary is found in [16].

Pulse Irradiation Results

1) For applications requiring circuits hardened to pulse irradiation, quartz resonators are the least tolerant element in properly designed oscillator circuits.
2) Resonators made of unswept quartz or natural quartz can experience a large increase in series resistance, R_s, following a pulse of radiation; the radiation pulse can even stop the oscillation.
3) Resonators made of properly swept quartz experience a negligible change in R_s when subjected to pulsed ionizing radiation (the oscillator circuit does not require a large reserve of gain margin).

Steady-State Radiation Results

1) At doses < 100 rad (1 rad = 10^{-2} Gy) frequency change is not well understood. Radiation can induce stress relief. Surface effects such as adsorption, desorption, dissociation, polymerization, and charging may be significant. The frequency change is nonlinear with dose.
2) At doses >1 krad, frequency change is quartz impurity dependent. The ionizing radiation produces electron-hole pairs; the holes are trapped by the impurity Al sites while the compensating cation (Li or Na, for example) is released. The freed cations are loosely trapped along the optic axis. The lattice near the Al is altered, and the elastic constant is changed; therefore, the frequency shifts. Ge impurities are also troublesome.
3) At 10^6 rad, frequency change ranges from 10^{-11}/rad for natural quartz to 10^{-14}/rad for high quality swept quartz.
4) Frequency change is negative for natural quartz; it can be positive or negative for cultured and swept cultured quartz.
5) Frequency change saturates at doses > 10^6 rad.
6) The Q degrades if the quartz contains a high concentration of alkali impurities; the Q of resonators made of properly swept cultured quartz is unaffected.
7) Frequency change anneals at $T > 240$°C in less than 3 h.
8) Preconditioning (for example with doses > 10^5 rad) reduces the high dose radiation sensitivities upon subsequent irradiations.
9) High dose radiation can also rotate frequency-versus-temperature characteristic.

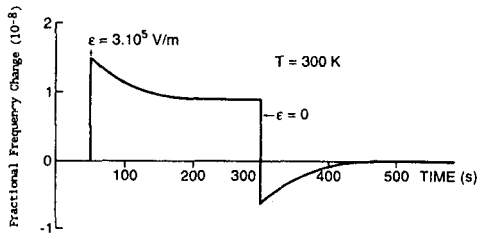


Fig. 12. Fractional frequency changes as a function of applied voltage. The slow variation after the change in voltage is due to the movement of ions within the resonator. From [1].

G. Electric Field

The frequency of certain resonator cuts are directly affected by the application of even small electric fields through changes in dimension and effective mass and through interaction with the nonlinear coefficients [61]–[65]. The application of electric fields also tends to cause ions within the crystal to move which changes the frequency. The result is that the change in frequency generally has a fast component due to piezoelectricity and the interaction with the crystal constants and one or more slower components associated with the movement of ions as shown in Fig. 12. The shorter time constants depend exponentially on temperature. This electric field effect has been used to vibration compensate SC-cut resonators [54] and to create an ultralinear phase modulator [61]. The sensitivity to this effect is highly dependent on resonator cut, material, and electrode configuration. Coefficients range from approximately 10^{-11} to 10^{-8}/V applied across the resonator and scale approximately as the reciprocal of the plate thickness or ν_o. For a resonator to exhibit a sensitivity, the dc electric field must have a component along the x axis of the crystal. AT-cuts with regular electrodes therefor exhibit very little sensitivity to electric fields and doubly rotated cuts, such as the SC-cut, generally show significant sensitivity to dc electric fields. Large electric fields and elevated temperatures are sometimes used to "sweep" ions out of the quartz bar prior to resonator fabrication [16], [62]–[64]. This is most often used on resonators for radiation environments.

H. Magnetic Field

The inherent magnetic field sensitivity of quartz resonators is probably smaller than 10^{-11}/T [1], [58]–[60]. Most resonators are, however, constructed with magnetic holders. As the magnetic field changes the force on the various components of the resonator changes. This causes a frequency shift through the force-frequency coefficient discussed above and the circuit phase shifts discussed below. Generally the magnetic field effect is unchanged under inversion and maximized for rotation through 90° [58], [59].

IV. Changes Within the Loop Electronics

Many environmental parameters cause a change in the phase around the oscillator loop. The most important are temperature, humidity, pressure, acceleration and vibration, magnetic field, voltage, load, and radiation. This environmental sensitivity often leads to increases in the level of wide-band phase noise in the short-term, random-walk frequency modulation in the medium-term, and drift in the long-term. The values of loop phase shift are not universal, but critically depend on the circuit design and the loaded Q-factor of the oscillator.

A. Tuning Capacitor

The frequency of the oscillator is generally fine-tuned using a load capacitor, C_L, of order 20 to 32 pF. In practice this capacitor is often made up of a fixed value (selected at the time of manufacture), a mechanically tuned capacitor for coarse tuning, a varactor for electronic tuning, and a contribution from the input capacitance of the sustaining stage and matching networks. The frequency change for small changes in C_L is

$$\Delta\nu/\nu_0 = \frac{C_1/2}{C_0 + C_L}\frac{dC_L}{C_0 + C_L} \tag{6}$$

where C_0 is the parallel capacitance and C_1 is the motional capacitance. It is not uncommon for the first term in (6) to be of order 10^{-3} [16]. In this case a change in C_L of only 10^{-6} results in a frequency change of 10^{-9}. Primary environmental parameters which change C_L are temperature, humidity, pressure, and shock or vibration. Humidity and pressure change the ratio between convection and conduction cooling. See Figs. 10 and 13. Pressure changes in a hermetically sealed oscillator can also be driven by temperature. This changes the temperature gradients and thereby the temperature of the tuning capacitor (and also the temperature of the resonator). Shock and vibration can change the mechanical capacitor, if present. Changes in the supply voltage, AGC, and temperature all change the input capacitance of the sustaining stage and the varactor diode. Humidity can change the value of the dielectric coefficient and losses in the capacitors and even the circuit board. Radiation can change the gain and offsets of the AGC and the sustaining stage and thereby change the effective input capacitance.

B. Mode Selection and Tuned Circuits

Most oscillators use matching circuits and filters to adjust the loop phase to approximately 0π and, especially with SC-cut resonators, to suppress unwanted modes. The phase shift across a transmission filter is given approximately by (1) with the Q in this case being that of the tuned circuit, Q_c. The fractional change in output frequency due to small changes in either circuit inductance L_c or circuit capacitance C_c is approximately given by

$$\Delta\nu/\nu \sim d\phi/2Q \sim \frac{Q_c}{Q}\left(\frac{dC_c}{C_c} + \frac{dL_c}{L_c}\right). \tag{7}$$

Far from resonance the change in loop phase with change in

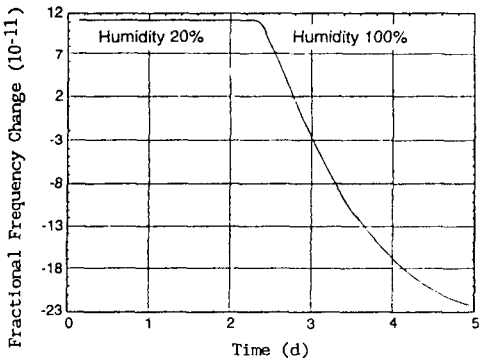


Fig. 13 Fractional frequency change of quartz controlled oscillator due to a change in humidity. Fractional frequency stability is approximately 2×10^{-13} at 1 s. From [22].

filter capacitance or inductance is generally much less than that given by (1). This suggests that the use of notch filters to suppress unwanted modes is probably superior to the use of narrowband transmission filters.

The values of most resistors, inductors and capacitors, and even parameters associated with the active junctions, are a function of temperature, humidity, current, or voltage. Significant improvements in the medium-term frequency stability can often be obtained merely by sealing an oscillator to prevent changes in the humidity and pressure [22], [23]. Fig. 13 shows the frequency change of a high performance oscillator due to a change in relative humidity from approximately 20% to 100% [22], [23]. This fractional change of frequency is about 1000 times the normal 1 s frequency stability of 3×10^{-13}.

C. External Load

If the external load of the oscillator changes there is a change in the amplitude and/or phase of the signal reflected back into the oscillator. The portion of this reflected signal that reaches the oscillating loop changes the phase of the oscillation and hence the output frequency by an amount given by (1) and expressed once more in (8). In this case we can estimate the maximum phase change as just the square root of the reverse isolation of the output circuit. For example, 40-dB reverse isolation corresponds to a maximum phase deviation of 10^{-2} rad:

$$\Delta\nu/\nu_0 \sim \left(\frac{d\phi}{2Q}\right) \sim \left(\frac{1}{2Q}\right)\sqrt{\text{isolation.}} \qquad (8)$$

For $Q \sim 10^6$ and isolation of 40 dB 10^{-4}, the maximum pulling is approximately $\Delta\nu/\nu_0 = 5 \times 10^{-9}$. This is an approximate model for most 5-MHz oscillators. As the frequency increases, the problem of load pulling becomes much worse because both the Q-factor and the reverse isolation decrease.

D. Acceleration, Vibration, and Magnetic Fields

Acceleration and vibration can distort the circuit substrate and the position of components leading to changes in the stray inductance and/or capacitance and thereby changes in the frequency. Virtually none of the presently available oscillators have magnetic shielding. The presence of a magnetic field complicates matters significantly since changes in orientation within the magnetic field lead to frequency shifts and motion dependent effects. Therefore movement, acceleration, and or vibration certainly leads to induced electric and magnetic fields that disturb the quiescent performance of the oscillator. In most cases such effects are very difficult to separate from other motion induced effects. Acceleration and vibration of circuit elements in the absence of a magnetic field lead to phase/frequency modulation. These effects are difficult to separate from those due to changes of stress applied to the resonator, but are no doubt present and may in some cases prevent accurate measurements on some low-g-sensitivity resonators [1], [56]–[60].

V. Conclusion

The frequency stability of quartz oscillators has been refined to the point that small changes in a wide variety of environmental parameters are now significant. The most important environmental parameter is probably acceleration and vibration at frequencies less than approximately 100 Hz. All of the other environmental drivers can be significantly reduced by appropriate attention to circuit design and/or shielding. Among the remaining environmental drivers, temperature is probably the most significant. Most oscillator ovens could be significantly improved using better circuits and better insulation. Magnetic shielding may be necessary to actually realize the full performance of low-g-sensitivity resonators. Pressure and humidity effects can be very serious in open oscillators. Fortunately these effects can be eliminated by hermetically sealing the oscillator. The sensitivity to low doses of radiation is not well understood. More work needs to be done to refine the characterization of sensitivity to various environmental parameters

Acknowledgment

We are particularly grateful to A. Ballato, R. Besson, R. Filler, and J. Vig for many discussions of systematic effects in quartz oscillators.

References

[1] J.-J. Gagnepain "Sensitivity of quartz oscillators to the environment: Characterization methods and pitfalls," *IEEE Trans. Ultrason. Ferroelec. Freq. Contr.*, vol. 37, pp. 347–354, 1990.
[2] D. B. Leeson, "A simple model of feedback oscillator noise spectrum," *Proc. IEEE*, vol. 54, pp. 329–330, 1966.
[3] T. E. Parker, "Characteristics and sources of phase noise in stable oscillators," in *Proc. 41st Annu. Symp. Freq. Contr.*, 1987, pp. 99–110.
[4] G. S. Curtis, "The relationship between resonator and oscillator noise, and resonator noise measurement techniques," in *Proc. 41st Annu. Symp. Freq. Contr.*, 1987, pp. 420–428.
[5] B. Parzen, *Design of Crystal and Other Harmonic Oscillators*. New York: Wiley, 1983.
[6] E. A. Gerber and A. Ballato, Eds., *Precision Frequency Control*. New York: Academic, 1985.
[7] A. Ballato and J. Vig., "Static and dynamic frequency-temperature behavior of singly and doubly rotated, oven-controlled quartz resonators," in *Proc. 32nd Annu. Symp. Freq. Contr.*, 1978, pp. 180–188.
[8] R. Holland, "Non uniformly heated anisotropic plates: I—Mechanical distortions and relaxation," *IEEE Trans. Sonics Ultrason.*, vol. SU-21,

1974.
[9] R. Holland, "Non uniformly heated anisotropic plates: II—Frequency transients in AT and BT quartz plates," *Ultrasonics Symp. Proc.*, pp. 592–598, IEEE Cat. # 74 CHO 896-ISU 1974.
[10] G. Théobald, G. Marianneau, R. Prétot, and J. J. Gagnepain, "Dynamic thermal behavior of quartz resonators," in *Proc. 33rd Annu. Symp. Freq. Contr.*, 1979, pp. 239–246.
[11] J. Kusters, "Transient thermal compensation for quartz resonators," *IEEE Trans. Sonics Ultrason.*, vol. SU-23, pp. 273–276 1976.
[12] J. A. Kusters and J. G. Leach, "Further experimental data on stress and thermal gradient compensated crystals," in *Proc. IEEE*, 1977, pp. 282–284.
[13] E. P. EerNisse, "Quartz resonator frequency shifts arising from electrode stress," in *29th Proc. Annu. Symp. Freq. Contr.*, 1975, pp. 1–4.
[14] A. Ballato, "Static and dynamic behavior of quartz resonators," *IEEE Trans. Sonics Ultrason.*, vol. SU-26, p. 299, 1979.
[15] B. K. Sinha and H. F. Tiersten, "Transient thermally induced frequency excursions in doubly-rotated quartz thickness mode resonators," in *Proc. 34th Annu. Symp. Freq. Contr.*, 1980, pp. 393–402.
[16] J. R. Vig, "Quartz crystal resonators and oscillators for frequency control and timing applications, a tutorial," Mar. 1991. Available from US Army Electronics Technology and Devices, SLCET-EQ, Ft. Monmouth, NJ 07703–5000, and from NTIS, AD A231 604..
[17] J. P. Valentin, G. Théobald, and J. J. Gagnepain, "Frequency shifts arising from in-plane temperature gradient distribution in quartz resonators," in *Proc. 38th Annu. Symp. Freq. Contr.* 1984, pp. 157–163.
[18] J. J. Gagnepain, R. Besson, "Nonlinear effects in piezoelectric quartz crystal," *Physical Acoustics*, Vol XI, W.P. Mason, Ed. Academic, 1975, pp. 245–288.
[19] J. J. Gagnepain, "Nonlinear properties of quartz crystal and quartz resonators: A review," in *Proc. 35th Annu. Symp. Freq. Contr.*, 1981, pp. 14–30.
[20] R. Besson, "A new piezoelectric resonator design," in *Proc. 30th Annu. Symp. Freq. Contr.*, 1976, pp. 78–83.
[21] R. Besson and U. R. Peier, "Further advances on B.V.A. quartz resonators," in *Proc. 34th Annu. Symp. Freq. Contr.*, 1980, pp. 175–182.
[22] F. L. Walls, "The influence of pressure and humidity on the medium and long-term frequency stability of quartz oscillators," in *Proc. 42nd Annu. Symp. Freq. Contr.*, 1988, pp. 279–283.
[23] F. L. Walls, "Environmental effects on the medium and long term frequency stability of quartz oscillators," in *Proc. 2nd Eur. Freq. and Time Forum*, Neuchatel, Switzerland, Mar. 1989, pp. 719–727.
[24] S. Ballandras, "Sensitivity of BAW devices to radial in-plane stress distribution, comparison between analytical and finite element results," in *Proc. 5th Eur. Freq. and Time Forum*, Besançon, Mar. 1991.
[25] M. B. Bloch, J. C. Ho, C. S. Stone, A. Syed, and F. L. Walls, "Stability of High Quality Quartz Crystal Oscillators: An Update," in *Proc. 43rd Annu. Symp. Freq. Contr.*, 1989, pp. 80–84.
[26] G. Marianneau and J.-J. Gagnepain, "Digital temperature control for ultrastable quartz oscillators," in *Proc. 34th Annu. Symp. Freq. Contr.*, 1980, pp. 52–57.
[27] J. F. Schooley, *Thermometry*. Boca Raton, FL: CRC Press, 1986.
[28] F. L. Walls, "Analysis of high performance compensated thermal enclosures," in *Proc. 41st Annu. Symp. Freq. Contr.*, 1988, pp. 439–443.
[29] F. L. Walls and J. J. Gagnepain, "Special applications, precision frequency control," in *Precision Frequency Control*, (E. A. Gerber and A. Ballato, Eds. New York: Academic Press, 1985, ch. 15, pp. 287–296.
[30] A. F. B. Wood and A. See, "Activity dips in AT-cut crystals," in *Proc. 21st Annu. Symp. Freq. Contr.*, 1967, pp. 420–435.
[31] C. Franx, "On activity dips of AT Crystals at high levels of drive," in *Proc. 22nd Annu. Symp. Freq. Contr.*, 1967, pp. 436–454.
[32] J. Birch and D. A. Weston, "Frequency/temperature, activity/temperature anomalies in high frequency quartz crystal units," in *Proc. 30th Annu. Symp. Freq. Contr.*, 1978, pp. 32–39.
[33] J. E. Deeter and H. Inoue, "Temperature dependence of the ginga clock rate," ISAS Res. Note 430, Jan. 1990.
[34] J. A. Kusters, "The SC-cut crystal—An overview," in *Proc. IEEE Ultrason. Symp.*, 1981, pp. 402–409.
[35] A. W. Warner, Jr. and B. Goldfrank, "Lateral field resonators," in *Proc. 39th Annu. Symp. Freq. Contr.*, 1985, pp. 473–474.
[36] A. W. Warner, "Measurement of plano-convex SC quartz blanks using lateral field excitation," in *Proc. 42nd Annu. Symp. Freq. Contr.*, 1988, pp. 202–204.
[37] J. J. Gagnepain, J. C. Poncot, C. Pégeot, "Amplitude-frequency behavior of doubly rotated quartz resonators," in *Proc. 31st Annu. Symp. Freq. Contr.*, 1977, pp. 17–22.
[38] R. L. Filler, "The amplitude-frequency effect in SC-cut resonators," in *Proc. 39th Annu. Symp. Freq. Contr.*, 1985, pp. 311–316.
[39] R. J. Besson, J. M. Groslambert, and F. L. Walls, "Quartz crystal resonators and oscillators, recent developments and future trends," *Ferroelectrics* vol. 43, pp. 57–65, 1982.
[40] J. M. Ratajski, "Force-frequency coefficient of singly-rotated vibrating quartz crystals," *IBM J. Res. Dev.* vol. 12, p. 92, 1968.
[41] C. R. Dauwalter, "The temperature dependence of the force sensitivity of AT-cut quartz crystals," in *Proc. 26th Annu. Symp. Freq. Contr.*, 1972, pp. 108–112.
[42] A. Ballato, E. P. Eernisse, and T. Lukaszek, "The force-frequency effect in doubly rotated-quartz resonators," in *Proc. 31st Annu. Symp. Freq. Contr.*, 1977, pp. 8–16.
[43] A. Ballato, "Force-frequency compensation applied to four-point mounting of AT-cut resonators," *IEEE Trans. Sonics Ultrason.*, vol. SU-25, 233, 1978.
[44] E. P. EerNisse, "Temperature dependence of the force frequency effect for the rotated X-cut," in *Proc. 33rd Annu. Symp. Freq. Contr.*, 1979, pp. 300–305.
[45] J. J. Gagnepain, M. Olivier, and F. L. Walls, "Excess noise in quartz crystal resonators," in *Proc. 37th Annu. Symp. Freq. Contr.*, 1983, pp. 218–225.
[46] R. L. Filler, L. J. Keres, T. M. Snowden, and J. R. Vig, "Ceramic flatpack enclosed AT and SC-cut resonators," in *Proc. IEEE Ultrason. Symp.*, 1980.
[47] J. R. Vig, "UV/Ozone cleaning of surfaces," *J. Vac. Sci. Technol.* vol. A3, pp. 1027–1034, 1985.
[48] Y. K. Yong and J. R. Vig, "Resonator surface contamination—A cause of frequency fluctuations," in *Proc. 42nd Annu. Symp. Freq. Contr.*, 1988, pp. 397–403.
[49] W. G. Perkins, "Permeation and outgassing of vacuum materials," *J. Vac. Sci. Technol.*, vol. 10, pp. 543–556, 1973.
[50] A. Ballato, "Resonators compensated for acceleration fields," in *Proc. 33rd Annu. Symp. Freq. Contr.*, 1979, pp. 322–336.
[51] R. L. Filler, "The acceleration sensitivity of quartz crystal oscillators: A review," *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.*, vol. 35, pp. 297–305, 1988.
[52] F. L. Walls and J. R. Vig, "Acceleration insensitive oscillator," U. S. Pat. No. 4,575,690, 1986.
[53] J. M. Pryjemski, "Improvement in system performance using a crystal oscillator compensated for acceleration sensitivity," in *Proc. 32nd Annu. Symp. Freq. Contr.*, 1978, pp. 426–430.
[54] V. J. Rosati and R. L. Filler, "Reduction in the effects of vibration on SC-cut quartz oscillators," in *Proc. 35th Annu. Symp. Freq. Contr.*, 1981, pp. 117–121.
[55] Y. S. Zhou and H. F. Tiersten, "On the influence of a fabrication imperfection on the normal acceleration sensitivity of contoured quartz resonators with rectangular supports," in *Proc. 44th Annu. Symp. Freq. Contr.*, 1990, pp. 452–453. (See also H. F. Tiersten and Y. S. Zhou, "An analysis of the in-plane acceleration sensitivity of contoured quartz resonators with rectangular supports," same proceedings, pp. 461–467.)
[56] J. R. Vig, C. Audoin, M. Driscoll, E. P. EerNisse, R. L. Filler, M. Garvey, W. Riley, R. Smythe, and R. D. Weglein, "The effects of acceleration on precision frequency sources," LABCOM R+D Tech. Report SLCET-TR-91-3, Mar. 1991, available from U.S. Army, Electronics Technol. and Devices Lab., Labcom, Ft. Monmouth, NJ.
[57] M. H. Watts, E. P. EerNisse, R. W. Ward, and R. B. Wiggins, "Technique for measuring the acceleration sensitivity of SC-cut quartz resonators," in *Proc. 42nd Annu. Symp. Freq. Contr.*, 1988, pp. 442–446.
[58] R. Brendel, C. El Hassani, M. Brunet, and E. Robert, "Influence of magnetic field on quartz crystal oscillators," in *Proc. 43rd Annu. Symp. Freq. Contr.*, 1989, pp. 268–274.
[59] F. Deyzac, "Magnetic sensitivity of quartz oscillators," in *Proc. 4th Annu. Eur. Forum on Time and Freq.*, Mar. 1990, pp. 255–258.
[60] A. Ballato, T. J. Lukaszek, and G. J. Iafrate, "Subtle effects in high stability vibrators," in *Proc. 34th Annu. Symp. Freq. Contr.*, 1980, pp. 431–444
[61] J. Lowe and F. L. Walls, "Ultralinear small angle phase modulator,"in *Proc. 45th Annu. Symp. Freq. Contr.*, 1991.
[62] R. Brendel, and J.-J. Gagnepain, "Electroelastic effects and impurity relaxation in quartz resonators," in *Proc. 36th Annu. Symp. Freq. Contr.*, 1982, pp. 97–107.
[63] J. J. Martin, "Electrodiffusion (sweeping) of ions in quartz," *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.*, vol. 35, 288–296, 1988.
[64] J. G. Gualtieri, "Sweeping quartz crystals," in *Proc. 1989 IEEE Ultrason. Symp.*, 1989, pp. 381–391.
[65] C. K. Hruska and M. Kucera, "the dependence of the polarizing effect on the frequency of quartz resonators," *J. Canadian Ceramic Society*, vol. 55, pp. 39–41, 1986.
[66] R. Brendel, J.-J.Gagnepain, J. P. Aubry, "Impurity migration study in quartz crystal resonators by using electroelastic effect," in *Proc. 40th Annu. Symp. Freq. Contr.*, 1986, pp. 121–126.

Fred L. Walls was born in Portland, OR, on October 29, 1940. He received the B.S., M.S., and Ph.D. degrees in physics in 1958, 1964, and 1970, respectively, from the University of Washington, Seattle. His Ph.D. thesis was on the development of long-term storage and nondestructive detection techniques for electrons stored in Penning traps and the first measurements of the anomalous magnetic (g-2) moment of low energy stored electrons.

Since 1973, he has been a Staff Member of the Time and Frequency Division of the National Institute of Standards and Technology (formerly the National Bureau of Standards), Boulder, CO. He is presently engaged in research and development of hydrogen maser devices and cesium beam standards for use as ultrastable clocks, quartz oscillators with improved short- and long-term stability, low-noise microwave oscillators, frequency synthesis from RF to infrared, low-noise frequency stability measurement systems, and accurate phase-noise metrology. From 1970 to 1973, he was a Postdoctoral Fellow at the Joint Institute for Laboratory Astrophysics, Boulder, CO. This work focused on developing techniques fro long-term storage and nondestructive detection of fragile atomic ions stored in Penning traps for low-energy collision studies.

Dr. Walls is a member of the American Physical Society. He has publisher more than 70 scientific papers and holds three patents.

Jean-Jacques Gagnepain was born in Montbeliard, France, in 1942. He received the Ph.D. degree in electrical engineering in 1972 from the University of Besançon, Besançon, France.

He joined the Centre National de la Recherche Scientifique in 1976 and he is presently Director of the Laboratoire de Physique et Métrologie des Oscillateurs. He was a Guest Worker at the National Bureau of Standards, Boulder, CO, in 1975–1976.

Optical Frequency Measurements

DONALD A. JENNINGS, KENNETH M. EVENSON, AND DAVID J. E. KNIGHT

Invited Paper

This paper is a review of the history of the measurement of coherent optical frequencies. Since coherent optical frequency implies a laser device, this is therefore a review of laser frequency measurement. The development of frequency measurement from the Cs frequency standard to the visible is traced. Two related aspects of optical frequency measurements, the speed of light and the redefinition of the meter, are also discussed.

I. Introduction

The absolute frequency measurement of radiation at frequencies above 500 GHz was made possible by the laser: a source of "coherent" radiation [1]. The discovery of continuous-wave (CW) laser oscillation in the infrared regions of the electromagnetic regions in He–Ne gas mixture by Javan *et al.* in 1961 [2] was soon followed by the discovery of CW laser oscillation in the visible [3] and far-infrared regions [4] of the spectrum. The coherence of laser radiation brought about the possibility of the measurement of the frequency of this "coherent" radiation. The first such measurement of a laser frequency (rather than the wavelength) was reported by Hocker *et al.*, in 1967 [5]. The frequencies of single-mode emissions at 890 and 964 GHz (337 and 311 μm) were measured using submillimeter harmonic mixing techniques with silicon-point contact diodes, similar to those developed [6] to generate submillimeter radiation by harmonic generation from millimeter radiation.

These frequency measurements made it possible to use lasers for measuring the speed of light in vacuum c_0. The most reliable measurement of the speed of light at that time was considered to be that made by Froome in 1958 [7]

Manuscript received November 19, 1984; revised August 14, 1985.
D. A. Jennings and K. M. Evenson are with Time and Frequency Division, National Bureau of Standards, Boulder, CO 80303, USA.
D. J. E. Knight is with the Division of Quantum Metrology, National Physical Laboratory, Teddingtin, Middlesex, TW11 0LW, United Kingdom.

0018-9219/86/0100-0168$01.00 © British Crown Copyright 1986

in which the frequency f and the vacuum wavelength λ_0 were measured for stabilized microwave radiation in the 4-mm band. The result was given by $c_0 = f\lambda_0$. The main source of uncertainty in this experiment was in the wavelength determination, due to diffraction in the interferometer apertures and from the correction for the refractive index of the air. Froome's one-sigma uncertainty was ± 3 parts in 10^7, but there was disagreement with other results, particularly long-baseline ones (Bergstrand) [8]. The speed of light is used for electromagnetic distance measurement, especially in critical applications such as geodetic-survey baseline measurements, which were at that time limited by the uncertainty of c.

It was apparent by the mid-1960s that a more accurate determination of λ could be made at the shorter wavelength afforded by laser radiation; hence a more accurate value of c could be obtained. This possibility enabled the development of laser frequency measurement programs at various national standards laboratories such as the National Bureau of Standards at Boulder, CO, USA, the National Physical Laboratory, Teddington, UK, and the National Research Council, Ottawa, Canada.

Important steps in the development of laser frequency measurements were the introduction of the metal–insulation–metal (MIM) diode and the demonstration of high-order mixing (12th harmonic of the HCN laser against an H_2O laser at 28 μm) by Evenson *et al.* [9]. Cryogenic Josephson junctions were proving useful for very-high-order mixing with submillimeter lasers [10], [11]. In 1972 accurate frequency measurements were extended to 3.39 μm [12], with a chain of five lasers using MIM diode harmonic mixing devices to generate the laser harmonics.

For accurate frequency measurements it was necessary to have sufficiently stable "targets" to measure. These were provided by narrow atomic or molecular absorption or emission features used to stabilize ("lock") the frequency of the laser. Narrow references were provided by saturated-absorption features observed in passive gas cells at low pressure. These were free of the first-order Doppler effect. Particularly important were the methane-stabilized He–Ne laser at 3.39 μm [13] and the CO_2 laser locked to a saturated fluorescence absorption in CO_2 itself [14]. The methane-stabilized laser was notable for its reproducibility ($\pm$ a few parts in 10^{11}) and compactness, and the stabilized CO_2 laser was a very useful device because of the large number of closely spaced emissions that could be accurately characterized. The I_2-stabilized He–Ne laser oscillating at 633 nm (red) [15] was used as a wavelength standard for laboratory length measurement; but the measurement of its higher optical frequency was foreseen as a much more challenging problem.

"Optical" frequency measurements, therefore, consist of the measurements of these narrow spectral features used to stabilize the lasers, and the metrological landmarks consist of accurate measurements of these features. Most of the early measurements were aimed at determining the speed of light. The first of these highly accurate determinations of the speed of light was made in 1972 at the National Bureau of Standards in Boulder by measuring the frequency and wavelength of the 88-THz (3.39-μm) methane transition. A value for the speed of light (299 792 458 m/s) was recommended the following year by the Consulative Committee for the Definition of the Meter (CCDM) [16]. This value was eventually to become the "fixed" value c_0 used in the redefinition of the meter. It originally had an uncertainty of ± 4 parts in 10^9 which arose almost entirely from difficulty in comparing wavelengths with the ^{86}Kr emission used as the length standard. The recommended value was based on the NBS Boulder speed of light measurement together with values of the wavelength from three other laboratories. The CO_2 frequency measurement used to measure methane was also confirmed by an independent measurement at NPL, Teddington [17].

The 1973 recommendation for c_0 was made to meet the requirements of astronomers using lasers to measure astronomical distances. They used the speed of light to convert electromagnetic wave transit times into distance. (The time to the moon and back was being measured to about 1 part in 10^{10}.) Various speed of light determinations then in progress were completed and the methane frequency at 88 THz was directly checked by a measurement at NPL [18]. There was excellent agreement in all these measurements.

After the initial speed of light determinations, work on optical frequency measurement followed three broad directions. These were first to extend frequency measurement to higher frequencies than 88 THz, particularly towards the visible region, in the hope of unifying the standards of time and length. Second, the more accurate stabilized lasers were of interest as potential frequency standards and it is desirable to know their frequencies to accuracies close to their reproducibilities. A third direction was the use of laser frequency measurements for ultra-high-resolution, high-accuracy spectroscopy. The technique is extremely important in understanding complicated molecular spectra such as those from SF_6 and OsO_4 [19]–[21], and SiF_4 [22].

The extension of optical frequency measurements to 260 THz (1.15 μm) was performed at NBS Boulder through a sequence of gas lasers at 2.03, 1.52, and 1.15 μm [23]. Harmonic mixing was not obtained in the MIM diode above 200 THz (1.5 μm) consequently, phase-matched second-order effects in nonlinear crystals were used for frequency addition or doubling above 200 THz. An I_2 transition at 520 THz (576 nm) in the yellow–green region was shown to be an excellent stabilization reference for frequency-doubled light from the 1.5-μm He–Ne laser at the National Research Council (NRC), Ottawa [24]. This was followed by a joint NBS–NRC experiment in 1979 which measured the frequency of this transition to provide the first frequency measurement of a visible radiation [25].

A series of measurements of the methane transition used to stabilize the He–Ne laser at 88 THz was achieved between 1978 and 1981 at NPL [26], the Laboratoire Pour Temps et Fréquences (LPTF) [27] Paris, France, and in two laboratories in the USSR [28]–[30]. The methane-stabilized laser was used as a standard at NBS Boulder in making a more accurate measurement of the visible I_2 transition at 520 THz (576 nm) and a red I_2-stabilized He–Ne laser at 474 THz (634.3 nm) [31], [32]. The uncertainty in these measurements was ($\pm 1.6 \times 10^{-10}$), and provided frequency values 10 times more accurate than the corresponding wavelength could be determined from the Krypton-86 length standard. The latter experiment made use of excited Ne atoms for resonant three-frequency addition of 88, 125, and 260 THz (3.39, 2.39, and 1.15 μm, respectively) [33]. The NBS experi-

ments used for the first time a tunable color-center laser as a transfer oscillator, to provide a source at half the frequency of the 1.15-μm He–Ne laser (2.3 μm). This same laser was also used instead of a 2.39-μm He–Ne laser to provide sufficient power for making the 633-nm measurement.

The accurate frequency measurements permitted the adoption of a new definition of the meter by the General Conference of Weights and Measures in Paris in 1983 [34]. The new definition adopted an exact fixed value for the speed of light: the value recommended in 1973. The "Mise en pratique," recommends a list of stabilized-laser frequencies to be used to realize the meter [35]. The most accurate values in the "Mise en pratique" from the visible region have 1-sigma uncertainties of 4 parts in 10^{10}, over an order of magnitude better than the 4 parts in 10^9 uncertainty of the previous length standard, the ^{86}Kr lamp. The "Mise en pratique" can be revised whenever better measurements become available.

The following text summarizes some of the important techniques of laser frequency measurement, describes some illustrative frequency-synthesis chains, lists accurate measurements of the speed of light, and discusses laser frequency standards. Recent proposals for more efficient harmonic generation and for laser-based frequency standards are also discussed. No replacement of the cesium frequency standard is imminent, but a methane-stabilized laser "clock" [36] has been demonstrated; it provides a less accurate, but interesting alternative to a cesium clock at a much higher frequency.

II. Techniques of Optical Frequency Measurement

A. Nonlinear Devices for Harmonic Generation and Mixing

Up to about 150 THz (2 μm) almost all frequency measurements have been made using essentially wide-band nonlinear point-contact devices, notably the MIM, the metal to semiconductor, Josephson-junction, and Schottky diodes. This is a perhaps surprising extension of the range of the early RF and microwave cat-whisker diode. Above about 150 THz, bulk nonlinear dielectric crystals have been used for second-order effects, frequency doubling, and addition or subtraction of frequencies. In one experiment, excited Ne atoms in the gaseous state served for the summation of three frequencies [32], [33]. Mixing (difference-frequency generation) above 150 THz has been achieved with photodiodes, although most commercial diodes are limited to difference frequencies substantially less than 100 GHz.

Reviews concentrating on nonlinear devices for laser frequency measurement have been given by Knight and Woods [37] and by Klementev *et al.* [38]. Thus opportunities for frequency synthesis are, at present, more restricted above 150 THz than below.

Information on the highest harmonic order, highest frequency reached, and the largest beat frequency observed in the photo-mixing regime is given for the principal kinds of mixing device in Table 1.

Most of the devices used operate through a resistive rather than reactive nonlinearity, and none of these is highly efficient for harmonic generation. Each has a useable frequency range determined mainly by the physical characteristics of the devices, and in practice, this is more important than ease of harmonic generation. Thus Josephson junctions generate high-order harmonics easily, to mix with frequencies up to 1 or 2 THz but become much less effective at 4 THz. For niobium, the most practical superconductor so far, the paired-electron band-gap energy corresponds to a frequency of 0.8 THz. The MIM diode by comparison with a semiconductor diode has only about 1/100th of the spreading resistance in the post material, thus offering a 100-times higher cutoff frequency, even though the voltage–current characteristic is markedly less nonlinear than that of most semiconductor diodes. The GaAs Schottky diode is an improvement on the silicon point-contact diode. It exhibits a higher mobility of the semiconductor, and a rugged and highly "ideal" voltage–current characteristic of the barrier formed with the evaporated metal contact. Both MIM and Schottky diodes have recently proved useful at visible-laser frequencies for obtaining large difference frequency signals [47], [48]. Mixing in phase-matched bulk crystals all generate much larger difference frequencies but there are problems in phase-matchability and transparency of the available crystals. The individual crystal also has a limited tuning range, obtained by varying its temperature or the beam angles.

Since the main characteristics of MIM diodes have been

Table 1

Device	Wide-Band Action—Highest Harmonic n_{MAX}	f/THz	Ref.	Harmonic Mixing Highest Frequency f/THz	n	Ref	Second-Order Action High-Frequency Beat Δf_{MAX}/GHz	f_0/THz	Ref
Semiconductor PC diode (W–Si)	23	1.58	[39]	3.56	4	[42]			
Josephson junction PC (Nb–Nb)	825	0.89	[11]	4.25	43	[43]	109	30	[46]
Schottky diode (GaAs)	401	3.8	[10]						
Metal–Insulator–Metal (MIM) diode (W–Ni)	33	2.5	[40]	5.3	13	[44]	900	528	[47]
Avalanche photodiode (GaAs)	12	10.7	[9]	197	$f_1 + f_2 \sim 2$	[45]	2500	$\sim$ 500	[48]
Three-terminal FET (GaAs)	16	26	[41]	130	$3f_1 + 2f_2 \sim 5$	[31]			
							100	$\sim$ 500	[49]
							300	474	[50]

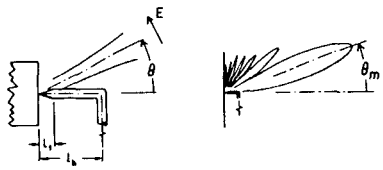


Fig. 1. The MIM diode optical antenna. (a) Showing both the conical antenna part ℓ_r and the long wire antenna part ℓ_b. (b) The calculated antenna pattern for $\ell_r = 7\lambda$.

published [51]–[53] brief notes will follow on investigations relevant in an engineering sense.

The usual room-temperature mixer, a MIM or a Schottky diode, has an open structure (without waveguide enclosures) and radiation is coupled to it with lenses or mirrors. The whisker is usually in the horizontal plane so that the face of the post is vertical. The whisker acts as a long antenna for laser radiation up to about 30 THz (Fig. 1), having a length of several wavelengths, and a right-angle bend serves to define the end of the antenna. Above about 30 THz, coupling to the diode occurs at the conical end of the antenna. The cone is made by electrochemically etching the tungsten, and this conical antenna then produces the electrical currents at the tip.

The long antenna has a series of lobes, the first and largest having a maximum at an angle θ_m to the antenna, given by $\theta_m = \cos^{-1}(1 - 0.371\lambda/L)$, where L is the effective antenna length. The lobes are axially symmetrical about the whisker. The conical antenna exhibits a broader, simpler, but somewhat similar pattern.

In practice, the laser radiation is focussed from a Gaussian beam onto the antenna and the angle is adjusted to the peak of the first lobe. With the whisker horizontal, all laser radiation applied is horizontally polarized and each wavelength is arranged at an angle near its own θ_m in the horizontal plane. To achieve optimum coupling, the focal length and beam diameter should be chosen to match the converging beam to the half-width of the antenna lobe. Corner reflectors have also been used to improve coupling to the diode [54].

Frequencies from dc up to perhaps 10 GHz can be coupled in or out of the diode directly by coaxial lines or with short antennas on the coax. Care is necessary to isolate the extremely delicate point contact (especially in the case of MIM diodes) from mechanical forces on the coaxial lead. Millimeter-wave frequencies, typically 20–100 GHz, are applied to the whisker from an open waveguide, placed above the diode, with the E-vector parallel to the whisker.

In some situations there may be three laser beams and microwave radiation on the diode, and the heterodyne frequency of the order of 10–600 MHz extracted. It is not easy for all radiations to be optimally coupled at the same time. Usually, higher power and better coupling are required for the radiation having the highest harmonic. Excessive power from the radiation at the fundamental frequency generally increases the diode noise and diminishes the signal obtained. With laser radiation applied to MIM diodes, increases in noise of the order of 10 dB are seen at low frequencies, decaying into the white noise at frequencies above about 30 MHz. Laser powers of the order of a microwatt (in the FIR region) produces an excellent signal-to-noise ratio in low-order mixing experiments on the MIM diode; higher powers can be beneficial up to about 200 mW in the IR region, at which point destructive whisker heating takes place.

B. Laser Frequency Measurement and Frequency-Synthesis Chains

In general, when various frequencies of electromagnetic radiation are incident on a nonlinear detector-mixer, such as the MIM diode, Schottky diode, or the Josephson junction, all harmonics and sums and differences are generated. A low-frequency difference, Δf (usually between 30 and 1500 MHz), is used to measure unknown frequencies as will be shown in the following discussion.

1) Mixing Microwave Harmonics with a Laser: The simplest scheme for measuring a laser frequency f_L is by beating directly against a harmonic of a known microwave frequency f_M using a mixer. The value of f_L is calculated from measuring the beat frequency Δf, its sign, and the harmonic order n, in the expression

$$f_L = nf_M \pm \Delta f. \tag{1}$$

f_L is the frequency of a laser which, in the case of a gas laser, is tunable by about ± 2 parts per million (ppm). The measurement accuracy can exceed 1 part in 10^{11}. However, such reproducibility and accuracy are achieved only when a laser stabilized to sub-Doppler atomic or molecular absorption is being used. Both n and the sign of Δf can be established by varying f_M, and for this purpose a wide-band spectrum analyzer of order 1 GHz is useful.

The microwave oscillator (usually a klystron oscillating at less than 100 GHz) can be phase-locked to a quartz crystal. In this case, the crystal phase noise is multiplied right up to the laser frequency and the beat signal is broadened, thus the resolution of Δf is reduced [55].

Alternatively, the microwave oscillator can be phase-locked to the laser being measured. In this case, the phase-lock signal originates from the phase angle between the heterodyne signal Δf and a crystal-controlled IF reference. Quartz-crystal frequency multiplication is not then required beyond the microwave frequency, although signals have been shown to be clean enough (16-dB signal/noise in 100-kHz bandwidth) from a simple 120-MHz quartz-crystal oscillator when multiplied to 2.5 THz [56]. The phase-lock loop (PLL) bandwidth must be wide enough to suppress the klystron-produced noise at the laser frequency. In a simplified methane-frequency chain, a 99-GHz reflex klystron has been phase-locked to a laser at 4.25 THz (70 μm) by 43rd-harmonic mixing in a Joseph junction using a servo-loop with an IF bandwidth of 6 MHz and an overall natural frequency near 300 kHz [26], [46].

2) Two-Laser Mixing: The lasers used are usually gas lasers whose frequencies are selected from those available. A microwave frequency f_M is often needed to produce a final beat frequency Δf in the RF region. Thus the two-laser harmonic-mixing experiment is of the form

$$f_{L2} = n_1 f_{L1} \pm f_M \pm \Delta f \tag{2}$$

where f_{L2} is the unknown frequency and f_{L1} is the known one.

In such a measurement f_M can be phase-locked to a quartz-crystal oscillator. It is sometimes useful to phase-lock f_{L1} to f_{L2}, or to lock f_M to the laser pair via Δf. In order to obtain a fast means of controlling a laser frequency for phase locking, a long-range relatively slow PZT mirror controller is used in conjunction with a high-speed device such as a discharge-current controller [57], a small PZT [58], Stark cell [59], or intra-cavity electrooptic crystal [60], [61]. Fast frequency control by extracavity devices has also been demonstrated [62].

3) Three-Laser Mixing: When two suitable lasers do not exist for two-laser mixing, a third laser may be introduced according to the scheme

$$f_{L3} = n_1 f_{L1} \pm n_2 f_{L2} \pm f_M \pm \Delta f \qquad (3)$$

where f_{L3} is the unknown frequency and f_{L1}, f_{L2}, and f_M are known. This has recently been used to measure a color-center laser's frequency with two stabilized CO_2 lasers [31].

Three lasers are also used when one laser's frequency corresponds to the difference between the frequencies of two lasers needing to be compared. Then n_1 and n_2 are one and minus one, respectively. This has recently been demonstrated with MIM mixers for spectrally narrow visible dye lasers separated by 2.5 THz [51]. The "high-frequency beat" section of Table 1 shows the achievements with different mixers for this kind of mixing.

4) Laser Frequency-Synthesis Chains Involving Several Stages of Mixing: To attain the highest frequencies in the processes just described, a sequence of frequency measurements are combined in a so-called chain of lasers. The construction of these laser frequency-synthesis chains depends upon a combination of the available lasers and the obtainable properties of the nonlinear mixing elements summarized in Table 1. CW gas lasers have mainly been used because of their simplicity and spectral purity. Power output deficiencies have been overcome by building gain tubes up to 8 m long [9]. The set of several hundred CO_2 frequencies (from seven isotopic species) [63], covering the spectrum from 25 to 30 THz (9 to 11 μm), is widely used in synthesis from the microwave region to the 100–150-THz (3–2-μm) region.

Above 100–150 THz, the average spacing between gas-laser frequencies increases and the bandwidths of available frequency-mixing devices become smaller. Thus the CW tunable-laser systems have a role in simplifying frequency-synthesis schemes, in spite of their added complexity. An important example is the use of a color-center laser at 2.3 μm by Pollock *et al.* [31] to increase the measurement accuracy of the previously mentioned chain to 520 THz [32].

In the submillimeter region it is a fortuitous accident that three discharge lasers: the HCN laser at 337 μm, the D_2O laser at 84 μm, and the H_2O laser at 28 μm, produced convenient harmonically related frequencies connecting to a CO_2-laser line at 32 THz (9.3 μm). Only a few lines were available from these discharge lasers; however, this soon changed with the discovery of optically pumped far-infrared (FIR) lasers (pumped by CO_2 laser radiation) in 1970 [64]. By 1980 there were 1200 such CW emissions [65], [66] and by now there have been perhaps 1000 more reported. It is important to note, however, that comparatively few of these are powerful (10 mW or more) and that again comparatively few emit at frequencies above 3 THz (100 μm). Thus some current synthesis chains use three-laser mixing to access specific CO_2 frequencies [67], [68]. Also a chain at NRC, Ottawa by Whitford [69], [70] uses only CO_2 lasers and MIM diodes and utilized the difference frequencies between CO_2 lasers as transfer oscillators to multiply from the microwave region.

Some examples of frequency-measurement chains are shown in the following figures. The first chain to methane in 1973 [12] is shown in Fig. 2. Note the use of the HCN

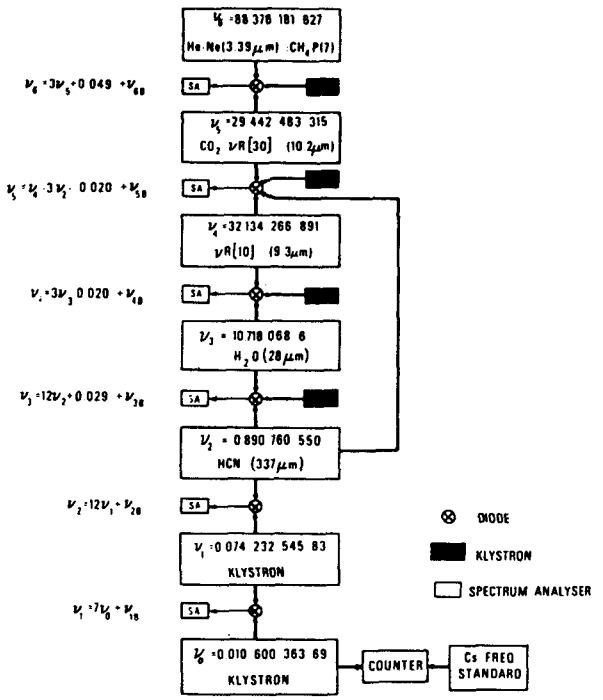


Fig. 2. The first laser frequency synthesis chain to frequency measure the He–Ne CH_4-stabilized laser. (All frequencies in THz.)

laser twice, in the chain to the H_2O laser, and separately to measure the difference frequency between the CO_2 lasers. Stabilized-CO_2 lasers were used as transfer standards in the chain. For more accurate measurement of the methane frequency the chain shown at Fig. 3 was used [26]. This chain used laser transfer oscillators at only two intervening frequencies and used simultaneous counting of beat frequencies to eliminate the effect of their frequency fluctuations. The frequency mixing range of the Josephson junction was extended for this experiment to 4.25 THz.

A third accurate chain to methane [30] is shown at Fig. 4. Here the various oscillators were phase-locked together and counting was performed at only one point in the chain: at 29.5 THz.

The extension of accurate measurement to a visible frequency, already mentioned, was performed in 1983 [31] using the scheme in Fig. 5. The target was an I_2 transition at twice the frequency of the 1.15-μm (260-THz) He–Ne line. A methane-stabilized laser was used to measure the frequencies of the two well-controlled stabilized CO_2 lasers

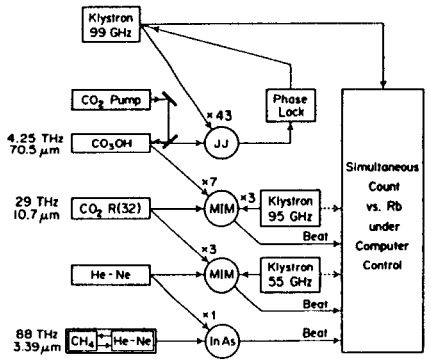


Fig. 3. A simultaneous-counting scheme to frequency measure the He–Ne CH_4-stabilized laser.

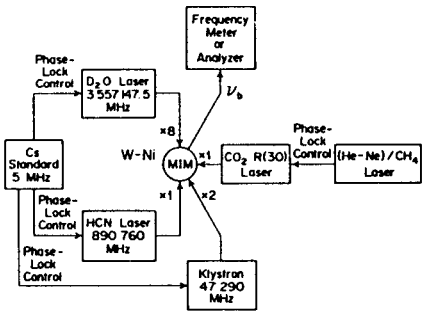


Fig. 4. A phase-locked method for the He–Ne CH_4-stabilized frequency measurement.

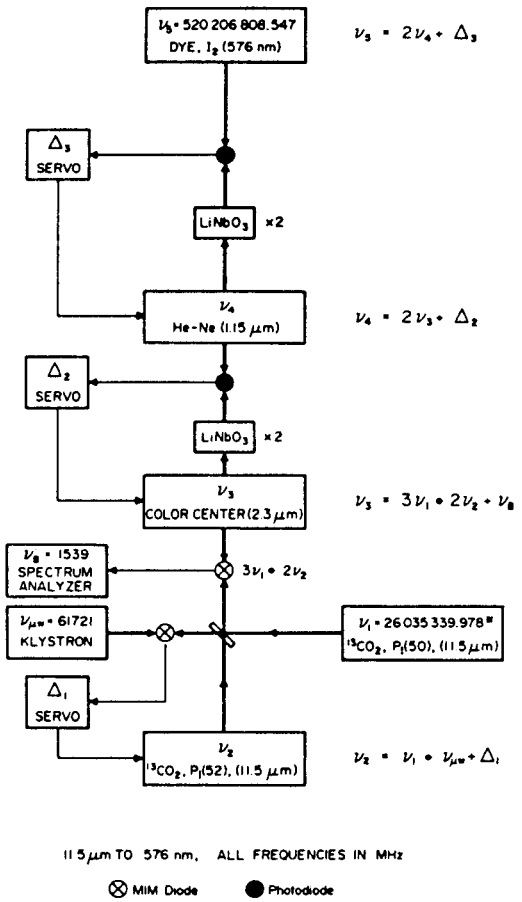


Fig. 5. The first laser frequency chain to the visible. The target was a molecular iodine transition at 520 THz.

used at the bottom of the chain. This scheme is more complicated than that for methane in Fig. 3 in spite of its smaller frequency-multiplication range, because of the limitation to frequency-doubling stages above 130 THz (2.3 μm).

The scheme previously mentioned that made use of the energy levels of the Ne atom to sum the 3.39-, 2.39-, and 1.15-μm emission of the He–Ne laser [32] is shown in Figs. 6 and 7. Radiation at exactly the sum of the three radiations traversing the tube is emitted and corresponds to the well-known 633-nm red emission. It is, however, a sum frequency due to resonant four-wave mixing in Ne, and is not a laser emission [33]. This scheme was used by the group at NBS, Boulder to transfer from their frequency measurement of an I_2 transition at 576 nm to the I_2-stabilized He–Ne laser at 633 nm. The 1.15-μm laser was doubled and stabilized to the previously measured I_2 transition at 576 nm; the color-center laser was used instead of He–Ne at 2.39 μm, and was measured with respect to methane as is shown in Fig. 5; and the 3.39-μm laser was directly compared with methane. Thus the 633-nm emission produced was known in terms of the methane frequency. The frequency of the iodine line used to stabilize a He–Ne laser was measured by measuring the RF beat produced by the two beams in a photodiode.

III. Results

A. Measurements of the Speed of Light

A summary of values for the speed of light obtained with lasers is shown in Fig. 8. Apart from the first experiment listed [71], which was notable for the economy of apparatus, all others involved the measurement by harmonic generation and mixing techniques of the frequency of a stabilized laser. The laser's wavelength was separately measured by comparing its wavelength with the emission from a ^{86}Kr lamp by which the meter was realized either by direct interferometry [88] or by a technique involving upconversion [73], [75], [98]. It was recognized in 1973 that the main uncertainty arose from the imprecision of measurements based on the ^{86}Kr lamp so that the CCDM recommended value was accompanied by the expression that it would be desirable to keep this value of c_0 for a possible future redefinition of the meter. A new definition of the meter was adopted in 1983 by retaining and fixing this value, after the intervening measurements (Fig. 8) showed i) that there was no serious discrepancy in the value adopted in 1973 and ii) that the frequency values of the visible laser radiations in common use as laboratory length standards had been established more accurately than their corresponding wavelength values.

B. Absolute Frequency Measurements of Stabilized Lasers

Early frequency measurements of metrological importance were those on the CO_2 fluorescence-stabilized CO_2 laser [90], [91]. These were used to calibrate accurate interline measurements of CO_2 and the associated determination of its molecular constants. The CO_2 frequencies then

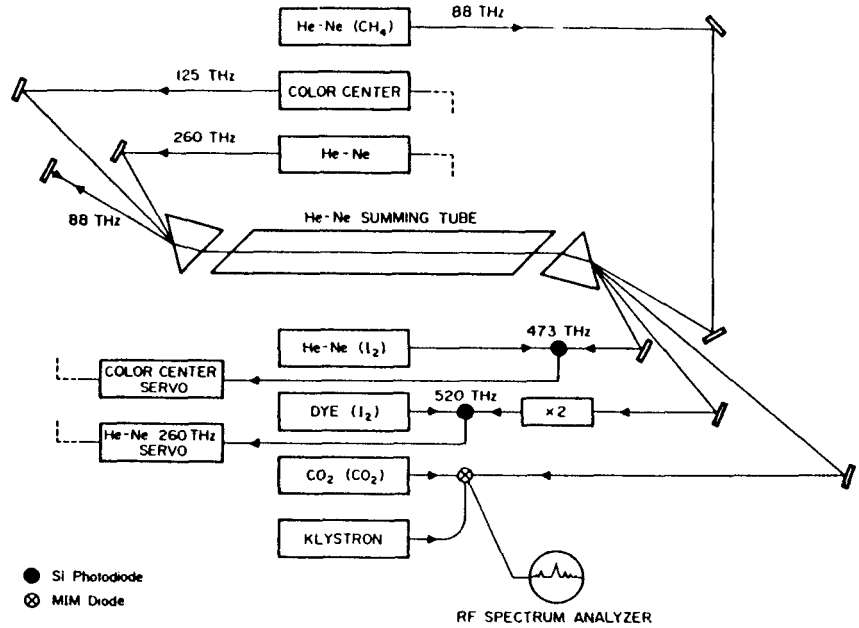


Fig. 6. A schematic diagram of the method used to frequency measure the molecular iodine transition used to stabilized the visible He–Ne laser at 473.6 THz (633 nm).

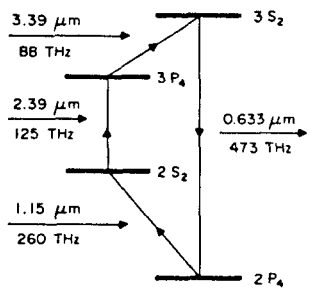


Fig. 7. A partial energy level diagram of Ne showing the energy levels involved in summing to obtain output at 473.6 THz.

obtained provided a grid of frequency references between 25 and 33 THz (9 to 11 μm) and were accurate to about 1 part in 10^9.

Fluorescence-stabilized CO_2 lasers were used as transfer oscillators to obtain measurements of the frequency of the He–Ne laser locked to the $F_2^{(2)}$ transition of methane at 88 THz (3.39 μm). These measurements [12], [18] were accurate to about 5 parts in 10^{10}, and agreed within 2 parts in 10^{10}.

Numerous measurements have been made on gas lasers in the submillimeter region in more than a dozen laboratories around the world. About one third of the 2000 or more emissions known in 1984 had been characterized by frequency measurement. These measurements mostly refer to the center of the laser tuning profile and are, therefore, comparatively less accurate, about 2 parts in 10^7. However, they are better than any wavelength measurements and are of considerable value for far-infrared spectroscopy.

A list of the lines ordered by frequency is available for submillimeter laser transitions with wavelengths larger than 12 μm [65], and a list of frequency measurements of optically pumped far-infrared laser lines is in preparation [66].

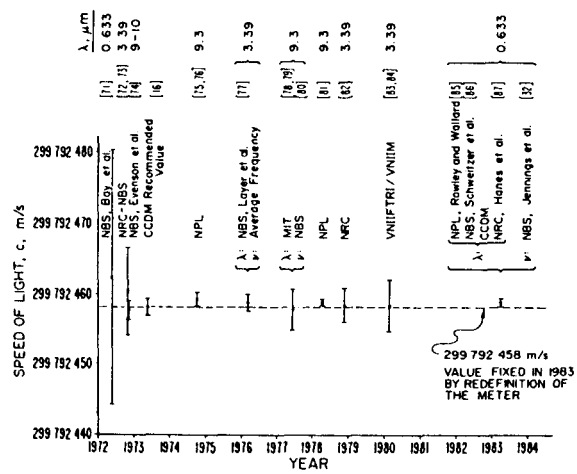


Fig. 8. Laser speed of light measurements from 1972 to 1984.

After the first CO_2 and methane frequency measurements, the speed of light was satisfactorily determined, and several laboratories made more accurate measurements of the frequency of the methane transition used to stabilize the 3.39-μm He–Ne laser. From the beginning, this laser had shown promise of accuracy close to 1 part in 10^{11} [13], but later attempts to improve this pointed to an estimated limit of about 1 part in 10^{13} [92]. Nonetheless, this is so far the most accurate stabilized laser in current use. Eight measurements from four laboratories have been reported since 1979 at an accuracy better than about 1 part in 10^{10}. A comparison of the various measurements since 1973 is shown in Table 2.

The Soviet Laboratories, VNIIFTRI, Moscow [29], [93] and the Institute of Thermophysics, Novosibirsk [28], followed

Table 2 Frequency Measurements of the $F_2^{(2)}$ Methane-Stabilized He–Ne Laser at 88 THz (3.39 μm)

Day	Date Published Mo.	Yr	Reference	Institution	ν − 88 376 181 000 kHz	σ kHz
11	Nov.	72	[72]	NBS	627	50
2	Feb.	73	[12]			
17	Apr.	75	[18]	NPL	608	43
5	Sept.	79	[29]	VNIIFTRI (GS)	586	10
12	Dec.	80	[93]			
	Dec.	79	[82]	NRC	570	200
12	Dec.	80	[27]	LPTF	618	14
12	Dec.	80	[26]	NPL	616	3
	Dec.	81	[94]	ITP	603	3.0
	May–June	83	[36]			
12	Dec.	81	[35]	LPTF	612	11
20	Aug.	81	[30]	VNIIFTRI (GS)	603.4	1.4
	Oct.	83	[35]	[Recommended value	608	3.9]
11	Nov.	83	[95]	ITP	602.9	1.2
	Jan.	85	[96]			
6	Jun.	85	[67]	LPTF	600.0	3.4

broadly the original scheme of Fig. 2 but improved the accuracy by extensive use of phase locking. This was carried so far in Novosibirsk as to provide the downward phase-locking methane clock previously mentioned [36]. This clock is now said to operate typically for 10–20 min without interruption. This is the only methane $F_2^{(2)}$ frequency measurement to have been made with the hyperfine structure resolved. The second measurement at VNIIFTRI [30] was made with the scheme of Fig. 4.

The measurement at LPTF in France [27] also broadly followed the scheme of Fig. 2, and exhibited improved laser stability, and phase-locked operation of the lowest frequency laser. A measurement at NPL [26] achieved improvement in resolution by using the much-simplified synthesis scheme in Fig. 3, and by simultaneously counting the beat frequencies (for 1 s) to eliminate the instability effects of the transfer lasers. A recent measurement at LPTF [67] makes use of fully phase-locked counting of an OsO_4-stabilized laser at 10.3 μm with ±50-Hz resolution.

The accuracy of these experiments demonstrates that the techniques of laser frequency measurement are capable of approaching the reproducibilities (accuracies) of the stabilized lasers being measured. This accuracy compares with that of the available rubidium standards and approaches that of even the cesium standard used.

The advances made at NBS Boulder in the extension of frequency measurement beyond 88 THz, to 260 THz (1.15 μm) and to the visible region of the spectrum, are shown in Figs. 5 and 6 [31], [32]. The scheme involved frequency multiplication from stabilized CO_2 lasers. These were calibrated against a methane-stabilized laser at 88 THz (the JILA telescope laser), assuming, from foregoing measurements (Table 2), that

$$f_{CH_4} = 88\,376\,181.609 \pm 0.009\ \text{MHz}.$$

The frequency of the ^{20}Ne Lamb-dip stabilized He–Ne laser at 1.15 μm was found to be

$$f_{1.15} = 260\,103\,249.26\ \text{MHz}$$

with a total fractional one-sigma uncertainty of 3.1 parts in 10^{10}. The frequency of the "o" hyperfine component of the $^{127}I_2$ 17-1 P(62) transition at 576 nm was found to be

$$f_{576\text{"o"}} = 520\,206\,808.547\ \text{MHz}$$

with a total fractional one-sigma uncertainty of 1.6 parts in 10^{10}.

The "o" transition of I_2 thus measured was used to measure the "i" and "g" hyperfine components of I_2 at 633 nm by the mixing process portrayed in Figs. 6 and 7 [32]. The frequency was likewise referred to the methane-stabilized laser assuming the value just given. The results for the $^{127}I_2$ 11-5 R(127) transition was

$$f_{633\text{"i"}} = 473\,612\,214.830 \pm 0.074\ \text{MHz}$$

where the I_2-stabilized laser was adjusted to the CIPM-recommended operating conditions [35].

These measurements in the visible part of the spectrum have accuracies of ±1.6 parts in 10^{10} which is within a factor of three of the best accuracy of the stabilized lasers used in the lower part of the chain.

A list of recommendations of the best known values for the frequencies of stabilized lasers of interest as length standards was prepared by the Consultative Committee for the Definition of the Meter in 1983 [35], and is shown in Table 3. The list accompanies the new definition of the meter.

The list takes account of the preceding frequency mea-

Table 3 CIPM Recommended Stabilized Laser Frequencies [35]

Laser	λ(nm)	Stabilization Ref.				ν(THz)	$\Delta\nu/\nu$ (3-sigma)
He–Ne	3392	CH_4	ν_3	P(7)	$F_2^{(2)}$	88. 376 181 608	$\pm 1.3 \times 10^{-10}$
Dye	576	$^{127}I_2$	17-1	P(62)	0	520. 206 808 51	$\pm 6 \times 10^{-10}$
He–Ne	633	$^{127}I_2$	11-5	R(127)	i	473. 612 214 8	$\pm 1 \times 10^{-9}$
He–Ne	612	$^{127}I_2$	9-2	R(47)	0	489. 880 355 1	$\pm 1.1 \times 10^{-9}$
Ar^+	515	$^{127}I_2$	43-0	P(13)	a_3	582. 490 603 6	$\pm 1.3 \times 10^{-9}$

surements discussed above and will be seen to assign a greater uncertainty than those of individual frequency measurements. The reasons are that i) a three-sigma uncertainty is given, rather than the usual one-sigma value, ii) the measurements of methane disagreed by about three times one-sigma, and iii) for some visible frequencies additional data were taken into account.

The uncertainties in Table 3 are about an order of magnitude larger than the accuracies claimed for the lasers themselves, so that improved measurements can be expected in the future.

Schemes to measure the frequencies of stabilized lasers in the visible spectral region are being explored at various laboratories. A suitable stabilized laser at 576 nm has been evaluated at NPL [97] and a frequency synthesis experiment has been described [68].

A summary of stabilized-laser frequency measurements traceable to cesium, and with accuracies exceeding 3 parts in 10^9, has been compiled [98]. Recent measurements of high accuracy have been made on an OsO_4 transition near 29 THz (10.3 μm) [67] and on the methane E line at 88.373 THz (3.39 μm) [95], [96].

IV. Discussion

A. Sources of Error in Laser Frequency Measurement

1) Inexactitude in Frequency-Mixing Devices: When the nonlinear properties of systems with well-defined energy levels are being used for frequency mixing it is important to guard against quanta characteristic of the mixing device (as opposed to those of the input frequencies) contributing to the output frequency. Checks can be made, by varying the input frequencies, for example.

On devices used for harmonic generation, some experimental checks have been made for differences of this type. An indirect check on the agreement between Josephson-junction harmonic mixing and the use of entirely room-temperature harmonic mixers is afforded by the results for the 88-THz methane frequency in Table 2. The measurements from NPL alone used the Josephson junction. In a specific test at NPL, see [76, pt. I, Appendix], frequency multiplication to 0.89 THz in a Josephson junction was found to agree with that in an MIM diode within a 1-sigma uncertainty of ± 1.2 parts in 10^{10}. (The test was primarily to compare frequency measurement by frequency division and counting with that by observation of a beat on a spectrum analyzer, but the harmonic mixers were also different.)

A comparison of frequency doubling in a bulk nonlinear crystal and in an MIN diode has been made by Edwards and colleagues at NPL [61]. A CO_2 laser's emission was doubled in a crystal of CdGeAs, and was also doubled in an MIM diode. Then the two frequencies were compared. Any net difference arising in the two frequency-doubling processes was found to be less than 7 parts in 10^{16}.

2) Effect of Phase Noise in Sources Used for Frequency Multiplication: The problem of spectral broadening in harmonics as a result of phase noise in the fundamental-signal oscillator showed at an early stage in laser frequency measurement. Lasers provided clean local oscillators with which to observe harmonic signals multiplied from quartz crystals, for example, and from microwave oscillators.

In the beginning, phase noise from quartz crystals used for phase-locking klystrons was a problem, but by using simplified high-frequency quartz-crystal oscillators (near 100 MHz) and straight frequency multipliers (avoiding some kinds of synthesizers) clean signals could be observed to 2.5 THz [56].

Studies on the special requirements of laser frequency measurement and on the properties of multiplied quartz-crystal noise [55] have helped to clarify the problem, specifically as to the point of catastrophic spectral broadening.

The use of downward phase-lock loops has diminished demands on upward frequency multiplication. High loop bandwidths of order 300 kHz have permitted downward phase locking a 99.6-GHz reflex klystron to a 4.25-THz laser via 43rd-harmonic mixing in a Josephson junction [46].

3) Doppler Effect and Time-Dependent Phase Shifts: At high accuracies it is necessary to guard against the effects of relative motion of different parts of the apparatus which might not cancel for the different methods used to observe signals, such as stored spectrum analyzer displays. Another potential source of error is from phase shifts in signal paths or filters which drift monotonically with time, as from a slowly changing temperature. Discrepancies at the level of 1 part $\sim 10^{12}$ have been observed [26, Appendix 1].

4) Phase-Lock Loop Errors: Phase-lock errors are actual mistakes and can usually be corrected: however, we will mention them anyway. A possible source of error is to lock on a loop-oscillation sideband. The least offset likely in this way would be comparable with the natural frequency of the loop, typically from 10 kHz upwards. Such an error is relatively easily resolved in the more accurate experiments and thus will generally be detected by the experimenter. Another problem arises from spurious modulation of weak signals used to phase-lock oscillators. For example, silicon-controlled rectifer spikes can cause intermittent unlock for say half-a-line-frequency period which could alter the mean locked frequency over a typical 1-s count-averaging time [46]. Careful spectrum analyzer checks are required in such circumstances. The direction of the error is, however, usually random so that accuracy increases with repeated measurement.

V. Summary

Many hundreds of laser frequencies and stabilization reference frequencies have now been measured, and accurate measurements have even been extended as high as the visible, yellow–green, portion of the spectrum.

This has permitted the recent (1983) redefinition of the meter fixing the value of c_0 exactly, so that length and wavelength are based on the same physical standard as time and frequency. Thus length is now referred to the most accurate physical standard available—the cesium clock. Substantial immediate improvements in realizing the meter have been obtained. Also, in such activities as time-of-flight ranging in space, the conversion from time to length does not increase the uncertainty.

Laser frequency measurement spans a spectral range from about 0.5–500 THz. These three orders of magnitude are comparable with the entire frequency span of microwave measurements 0.5–500 GHz; or, in a linear scale, frequency measurements represent a thousand-fold increase. This is extremely significant for communications purposes. The extension of frequency measurements to higher frequencies and improved accuracies is especially important for future experiments to seek frequency standards better than cesium.

The methane clock, current fundamental work towards new frequency standards using heavy molecules in the 10-μm band, and work on trapped ions at visible or higher frequencies [99] all point towards possible frequency standards with reproducibilities and accuracies exceeding that of the current Cs frequency standard.

ACKNOWLEDGMENT

One of the authors (DJEK) would like to thank W. R. C. Rowley for helpful comments.

REFERENCES

[1] T. H. Maiman, "Stimulated optical radiation in ruby," *Nature*, vol. 187, pp. 493–494, 1960.
[2] A. Javan, W. R. Bennett, Jr., and D. R. Harriott, "Population inversion and continuous optical maser oscillation in a gas discharge containing a He–Ne mixture," *Phys. Rev. Lett.*, vol. 6, pp. 106–110, 1961.
[3] A. D. White and J. D. Rigden, "Continuous gas maser operation in the visible," *Proc. IRE.*, vol. 50, p. 1697, 1962.
[4] A. Crocker, H. A. Gebbie, M. F. Kimmitt, and L. E. S. Mathias, "Stimulated emission in the far infrared," *Nature*, vol. 201, pp. 250–251, 1964.
[5] L. O. Hocker, A. Javan, D. Ranochandra Rao, L. Frenkel, and T. Sullivan, "Absolute frequency measurement and spectroscopy of gas laser transitions in the far infrared," *Appl. Phys. Lett.*, vol. 10, pp. 147–149, 1967.
[6] G. Jones and W. Gordy, "Extension of submillimetre wave spectroscopy below a half millimetre wavelength," *Phys. Rev.*, vol. 135A, p. 295, 1964.
[7] K. D. Froome, "A new determination of the free-space velocity of electromagnetic waves," *Proc. Roy. Soc., Ser. A*, vol. 247, pp. 109–122, 1958.
[8] E. Bergstrand, "A determination of the velocity of light," *Arkiv for Fysik*, vol. 2, 119–150, 1950.
[9] K. M. Evenson, J. S. Wells, L. M. Matarrese, and L. B. Elwell, "Absolute frequency measurements on the 28 and 78 μm cw water vapour laser lines," *Appl. Phys. Lett.*, vol. 16, pp. 159–162, 1970.
[10] D. G. McDonald, A. S. Risley, J. D. Cupp, K. M. Evenson, and J. R. Ashley, "Four-hundredth-order harmonic mixing of microwave and infrared laser radiation using a Josephson junction and a maser," *Appl. Phys. Lett.*, vol. 20, pp. 296–299, 1972.
[11] T. G. Blaney and D. J. E. Knight, "Direct 825th harmonic mixing of a 1 GHz source with an HCN laser in a Josephson junction," *J. Phys. D*, vol. 7, pp. 1882–1886, 1974.
[12] K. M. Evenson, J. S. Wells, F. R. Petersen, B. L. Danielson, and G. W. Day, "Accurate frequencies of molecular transitions used in laser stabilization: the 3.39 μm transition in CH_4 and the 9.33- and 10.18-μm transitions in CO_2," *Appl. Phys. Lett.*, vol. 22, pp. 192–195, 1973.
[13] R. L. Barger and J. L. Hall, "Pressure shift and broadening of methane line at 3.39 μm studied by laser-saturated molecular absorption," *Phys. Rev. Lett.*, vol. 22, pp. 4–8, 1969.
[14] C. Freed and A. Javan, "Standing-wave saturation resonances in the CO_2 10.6 μm transitions observed in a low-pressure room temperature absorber gas," *Appl. Phys. Lett.*, vol. 17, pp. 53–56, and 541, 1970.
[15] G. R. Hanes and C. E. Dahlstrom, "Iodine hyperfine structure observed in saturated absorption at 633 nm," *Appl. Phys. Lett.*, vol. 14, 362, 1969.
[16] Comite Consultatif pour la Definition du Metre, 5th Session, BIPM, Sevres, France, 1973.
[17] T. G. Blaney, C. C. Bradley, G. J. Edwards, D. J. E. Knight, P. T. Woods, and B. W. Jolliffe, "Absolute frequency measurement of the R(12) transition of CO_2 at 9.3 μm," *Nature*, vol. 244, p. 504, Aug. 24, 1973.
[18] T. G. Blaney, G. J. Edwards, B. W. Jolliffe, D. J. E. Knight, and P. T. Woods, "Absolute frequencies of the methane-stabilized He–Ne laser (3.39 μm) and the CO_2, R(32) stabilized laser (10.17 μm)," *J. Phys. D.*, vol. 9, pp. 1323–1330, 1976, and Nature, vol. 254, no. 5501, pp. 584–585, Apr. 17, 1975.
[19] Ch. J. Bordé, M. Ouhayoun, A. Van Lerberghe, C. Salomon, S. Avrillier, C. D. Cantrell, and J. Bordé, "High resolution saturated spectroscopy with CO_2 lasers. Application to the ν_3 bands of SF_6 and OsO_4," in *Laser Spectroscopy IV* (Series in Optical Sciences, vol. 21), H. Walther and K. W. Rothe, Eds. Berlin: Springer-Verlag, 1979, pp. 142–153.
[20] A. Clairon, A. Van Lerberghe, C. Salomon, M. Ouhayoun, and Ch. J. Bordé, "Towards a new absolute frequency reference grid in the 28 THz range," *Opt. Commun.*, vol. 35, no. 3, pp. 368–372, Dec. 1980.
[21] J. Bordé and Ch. Bordé, "Superfine and hyperfine structures in the ν_3 band of $^{32}SF_6$," *Chem. Phys.*, vol. 71, pp. 417–441, 1982.
[22] R. S. McDowell, C. W. Patterson, N. G. Nereson, F. R. Petersen, and J. S. Wells, "CO_2 laser coincidences with ν_3 of SiF_4 near 9.7 μm," *Opt. Lett.*, vol. 6, no. 9, pp. 422–424, Sept. 1981.
[23] D. A. Jennings, F. R. Petersen, and K. M. Evenson, "Frequency measurement of the 260 THz (1.15 μm) He–Ne laser," *Opt. Lett.*, vol. 4, pp. 129–130, 1979.
[24] G. R. Hanes, "Doubly-resonant intracavity generation of second harmonic of 1153 nm radiation with stabilization on h.f.s. components of $^{127}I_2$," *Appl. Opt.*, vol. 18, pp. 3970–3974, 1979.
[25] K. M. Baird, K. M. Evenson, G. R. Hanes, D. A. Jennings, and F. R. Petersen, "Extension of absolute frequency measurements to the visible: frequencies of ten hyperfine components of iodine," *Opt. Lett.*, vol. 4, pp. 263–264, 1979.
[26] D. J. E. Knight, G. J. Edwards, P. R. Pearce, and N. R. Cross, "Measurement of the frequency of the 3.39 μm methane-stabilized laser to ±parts in 10^{11}," *IEEE Trans. Instrum. Meas.*, vol. IM-29, pp. 257–264, Dec. 1980.
[27] A. Clairon, B. Dahmani, and J. Rutman, "Accurate absolute frequency measurements on stabilized CO_2 and He–Ne infrared lasers," *IEEE Trans. Instrum. Meas.*, vol. IM-29, pp. 268–272, 1980.
[28] V. P. Chebotayev, "Optical time scale," *J. Phys. (Colloq. C8)*, supplement to vol. 42, no. 12, pp. C8-505–C8-512, Dec. 1981. V. G. Gol'dort, V. F. Zakhar'yash, V. M. Klement'ev, M. V. Nikitin, B. A. Timchenko, and V. P. Chebotaev, *Pis'ma Zh. Eksp. Fiz.*, vol. 8, pp. 157–161, 1982, *transl:* "Development of an optical time scale," *Sov. Tech. Phys. Lett.*, vol. 8, no. 2, pp. 68–69, 1982.
[29] Yu. S. Domnin, N. B. Koshelyaevskii, V. M. Tatarenkov, and P. S. Shumyatskii, *Pis'ma Zh. Eksp. Teor. Fiz.*, vol. 30, no. 5, pp. 273–275, Sept. 5, 1979, *transl:* "Absolute measurements of IR laser frequencies," *JETP Lett.*, vol. 30, no. 5, pp. 253–255, Sept. 5, 1979.
[30] Yu. S. Domnin, N. B. Koshelyaevskii, V. M. Tatarenkov, and P. S. Shumyatskii, *Pis'ma Zh. Eksp. Teor. Fiz.*, vol. 34, no. 4, pp. 175–178, Aug. 20, 1981, *transl:* "Measurement of the frequency of a He–Ne/CH_4 laser," *JETP Lett.*, vol. 34, no. 4 pp. 167–170, Aug. 20, 1981.
[31] C. R. Pollock, D. A. Jennings, F. R. Petersen, J. S. Wells, R. E. Drullinger, E. C. Beaty, and K. M. Evenson, "Direct frequency measurements of transitions at 520 THz (576 nm) in iodine, and 260 THZ (1.15 μm) in neon," *Opt. Lett.*, vol. 8, pp. 133–135, 1983.
[32] D. A. Jennings, C. R. Pollock, F. R. Petersen, R. E. Drullinger, K. M. Evenson, J. S. Wells, J. L. Hall, and H. P. Layer, "Direct frequency measurement of the I_2-stabilized He–Ne 473 THz (633 nm) laser," *Opt. Lett.*, vol. 8, pp. 136–138, 1983.
[33] V. M. Klementyev, Yu. A. Matyugim, and V. P. Chebotayev, *Pis'ma Zh. Ekp. Teor. Fiz.*, vol. 24, no. 1, pp. 8–12, July 5, 1976, *transl.*: "Mixing of the frequencies 88.37 THz (λ = 3.39 μm), 125.13 THz (λ = 2.39 μm), and 260.1 THz (λ = 1.15 μm) in a gas, and production of continuous coherent emission with combined frequency 473.6 THz (λ = 0.63 μm)," *JETP Lett.*, vol. 24, no. 1, pp. 5–8, July 5, 1976.
V. P. Chebotayev, V. M. Klementyev, Yu. G. Kolpakov, and Y. A. Matyugin, "Frequency synthesis of 0.63 nm radiation by mixing optical frequencies in gas and nonlinear crystals," in *Proc. 2nd Frequency Standards and Metrology Symp.* (Copper Mountain, CO, July 5–7, 1976. Boulder, CO: Nat. Bur. Stand., Frequency and Time Standards Section, Nov. 1976), pp. 368–384.
[34] *Comptes Rendus des Séances de la 17e CGPM*, BIPM, Sevres,

France, 1983.
[35] BIPM, "Documents concerning the new definition of the metre," *Metrologia*, vol. 19, pp. 163–177, 1984.
[36] S. N. Bagayev, B. D. Borisov, V. G. Gol'dort, A. Yu. Gusev, A. S. Dychkov, V. F. Zakharyash, V. M. Klementyev, M. U. Nikitin, B. A. Tinchenko, V. P. Chebotayev, and V. V. Yumin, USSR Academy of Sciences, Siberian Branch, Institute of Thermophysics, Preprint 78-82, Novosibirsk, 1982, publ. "Optical time standard," *USSR Acad. Sci. Avtometriya*, no. 3, pp. 37–58, May–June 1983, *transl*: Automatic Monitoring and Measurement (9B).
[37] D. J. E. Knight and P. T. Woods, "Application of nonlinear devices to optical frequency measurement," *J. Phys. E.*, vol. 9, no. 11, pp. 898–916, 1976.
[38] V. M. Klement'ev, Yu. A. Matyugin, and V. P. Chebotaev, *Kvant. Elecktron.*, vol. 5, no. 8, pp. 1671–1681, Aug. 1978, *transl*: "Frequency measurements in the optical range: Current status and prospects," *Sov. J. Quantum Electron.*, vol. 8, no. 8, pp. 953–958, Aug. 1978.
[39] L. O. Hocker, D. R. Rao, and A. Javan, "Absolute frequency measurement of the 190 μm and 194 μm gas laser transitions," *Phys. Lett.*, vol. A-24, pp. 690–691, 1967.
[40] H. R. Fetterman, B. J. Clifton, P. E. Tannenwald, and C. D. Parker, "Submillimeter detection and mixing using Schottky diodes," *Appl. Phys. Lett.*, vol. 24, pp. 70–72, 1974.
[41] K. M. Evenson, unpublished.
[42] L. O. Hocker, J. G. Small, and A. Javan, "Extension of absolute frequency measurements to the 84 μm range," *Phys. Lett.*, vol. A-29, pp. 321–322, 1969.
[43] T. G. Blaney, N. R. Cross, D. J. E. Knight, G. J. Edwards, and P. R. Pearce "Frequency measurement at 4.25 THz (70.5 μm) using a Josephson harmonic mixer and phase-lock techniques," *J. Phys. D: Appl. Phys.*, vol. 13, pp. 1365–1370, 1980.
[44] C. O. Weiss and A. Godone, "Harmonic mixing and detection with Schottky diodes up to the 5 THz range," *IEEE, J. Quatum Electron*, vol. QE-20, pp. 97–99, 1984.
[45] K. M. Evenson, D. A. Jennings, F. R. Petersen, and J. S. Wells, "Laser frequency measurements: A review, limitations, extension to 197 THz (1.5 μm)," in *Laser Spectroscopy III*, J. L. Hall and J. L. Carlsten, Eds. Berlin, Germany: Springer-Verlag, 1977, pp. 56–58.
[46] D. G. McDonald, F. R. Petersen, J. D. Cupp, B. L. Danielson, and E. G. Johnson, "Josephson junctions at 45 times the energy-gap frequency," *Appl. Phys. Lett.*, vol. 24, pp. 335–337, 1974.
[47] H-U. Daniel, B. Maurer, and M. Steiner, "A broadband Schottky point contact mixer for visible laser light and microwave harmonics," *Appl. Phys.*, vol. B-30, pp. 189–193, 1983.
[48] R. E. Drullinger, K. M. Evenson, D. A. Jennings, F. R. Petersen, J. C. Bergquist, L. Burkins, and H. U. Daniel, "2.5 THz frequency difference measurements in the visible using metal-insulator-metal diodes," *Appl. Phys. Lett.*, vol. 42, pp. 137–138, 1983.
[49] S. W. Wang, *Lasers and Applications*, vol. III, no. 10, p. 6 (late news), Oct., 1984; see also; S. Y. Wang, D. M. Bloom, and D. M. Collins, "20-GHz bandwidth GaAs photodiode," *Appl. Phys. Lett.*, vol. 42, no. 2, pp. 190–192, Jan. 15, 1983.
[50] A. Chu, H. R. Fetterman, D. D. Peck, and P. E. Tannerwald, "Heterodyne experiments from millimetre wave to optical frequencies using GaAs MESFETs above f_T," in *IEEE Conf. Preprint* (Dallas MTT, 1982), pp. 25–28, 1982.
[51] K. J. Siemsen and H. D. Riccus, "Experiments with point-contact diodes in the 30-130 THz frequency region," *Appl. Phys.*, vol. A35, pp. 177–187, 1984.
[52] H. D. Riccus and K. J. Siemsen, "Point-contact diodes," *Appl. Phys.*, vol. A35, pp. 67–74, 1984.
[53] K. M. Evenson, M. Inguscio, and D. A. Jennings, "Point-contact diode at laser frequencies," *J. Appl. Phys.*, vol. 57, pp. 956–960, 1985.
[54] H. R. Fetterman, P. E. Tannenwald, B. J. Clifton, C. D. Parker and W. D. Fitzgerald, "Far-IR heterodyne radiometric measurements with quasioptical Schotty diode mixers," *Appl. Phys. Lett.*, vol. 33, pp. 151–154, 1978.
[55] F. L. Walls, and A. DeMarchi, "RF spectrum of a signal after frequency multiplication; measurement and comparison with a simple calculation," *IEEE Trans. Instrum. Meas.*, vol. IM-24, pp. 210–217, 1975.
[56] T. G. Blaney, N. R. Cross, and D. J. Knight, "Harmonic mixing and frequency measurements at 2.5 THz using Josephson junctions," *J. Phys. D*, vol. 9, pp. 2175–2180, 1976.
[57] J. S. Wells, "A stabilized HCN laser for infrared frequency synthesis," *IEEE Trans. Instrum. Meas.*, vol. IM-22, pp. 113–118, 1973.
[58] V. G. Gol'dort, V. F. Zakhar'yash, and B. A. Kurnevich, *Pribory i Tekhnika Eksp.*, no. 2, pp. 244–248, Mar.–Apr. 1979, *transl.*: "Wideband unit for phase-frequency locking of lasers," *Instrum. and Exper. Technol.*, no. 2, pp. 557–562, Mar.–Apr. 1979.
[59] S. R. Stein, A. S. Risley, H. Van de Stadt, and F. Strumia, "High speed frequency frequency modulation of far-infrared lasers using the Stark effect," *Appl. Opt.*, vol. 16, pp. 1893–1896, July 1977.
[60] W. R. Leeb, H. K. Phillipp, A. L. Scholtz, and E. Bonek, "Frequency synchronization and phase locking of CO_2 lasers," *Appl. Phys. Lett.*, vol. 41, no. 7, pp. 592–594, Oct. 1, 1982.
[61] D. J. E. Knight, "Optical frequency synthesis at NPL," *J. Phys. (Colloq. C8)*, supplement to vol. 42, no. 12, pp. C8-495–C8-503, Dec. 1981.
[62] J. L. Hall and T. W. Hänsch, "External dye-laser frequency stabilizer," *Opt. Lett.*, vol. 9, no. 11, pp. 502–504, Nov. 1984.
[63] C. Freed, L. C. Bradley, and R. G. O'Donnell, "Absolute frequencies of lasing transitions in seven CO_2 isotopic species," *IEEE J. Quantum Electron.*, vol. QE-16, pp. 1195–1206, 1980.
[64] T. Y. Chang and T. J. Bridges, "Laser action at 452, 496 and 541 μm in optically pumped CH_3F," *Opt. Commun.*, vol. 1, no. 9, pp. 423–426, Apr. 1970.
[65] D. J. E. Knight, "Ordered list of far-infrared laser lines (continuous, lambda > 12 μm)," NPL Rep. QU, 45, 1982, National Physical Laboratory Teddington, Middlesex. TW11 OLW, UK, Feb. 1981, reprinted with added pump-ordered list Sept. 1982.
_____, "Far-infrared lasers," in *Gas Lasers* (vol. II of *Handbook Series of Laser Science and Technology*), M. J. Webber, Ed. Boca Raton, FL: CRC Press, 1982, pp. 421–491. The latter list is ordered separately by lasing atom or molecule.
[66] M. Inguscio, G. Moruzzi, K. M. Evenson, and D. A. Jennings, a review paper on frequency measurements of optically pumped FIR laser lines, in preparation, for *Appl. Phys. Rev.*
[67] A. Clairon, B. Dahmani, A. Filimon, and J. Rutman, "Precise frequency measurements of CO_2/OsO_4 and He–Ne/CH_4 stabilized lasers," *IEEE Trans. Instrum. Meas.*, vol. IM-34, pp. 265–268, June 1985.
[68] D. J. E. Knight, G. J. Edwards, P. R. Pearce, K. I. Pharaoh, G. P. Barwood, N. R. Cross, Fan Xi-Sheng, and J. A. Golby, "Progress in an experiment to measure the frequency of an I_2-stabilized dye laser at 520 THz (576 nm)," presented at the VIIIth Vavilov Conf., Novosibirsk, USSR, July 26–28, 1984, submitted to *Izvestia Akad. Nauk SSSR* (Seriya Fizicheskaya).
[69] B. G. Whitford, "Simultaneous phase-lock of five CO_2 lasers to a primary Cs frequency standard," *Appl. Phys.*, vol. B-35, pp. 119–122, 1984.
[70] _____, "Measurement of the absolute frequencies of CO_2 laser transitions by multiplication of CO_2 laser difference frequencies," *IEEE Trans. Instrum. Meas.*, vol. IM-29, no. 3, pp. 168–176, Sept. 1980.
[71] Z. Bay, G. G. Luther, and J. A. White, "Measurement of an optical frequency and the speed of light," *Phys. Rev. Lett.*, vol. 29, pp. 189–192, 1972.
[72] K. M. Baird, H. D. Riccius, and K. J. Siemsen, "CO_2 wavelengths and the velocity of light," *Opt. Commun.*, vol. 6, pp. 91–95, 1972.
[73] J. D. Cupp, B. L. Danielson, G. W. Day, L. B. Elwell, K. M. Evenson, D. G. McDonald, L. O. Mullen, F. R. Petersen, A. S. Risley, and J. S. Wells, "The speed of light: Progress in the measurement of the frequency of the methane stabilized He–Ne laser at 3.39 μm," in *Proc. Conf. on Precision Electromagnetic Measurements* (Boulder, CO, USA), pp. 79–80, 1972.
[74] K. M. Evenson, J. S. Wells, F. R. Petersen, B. L. Danielson, G. W. Day, R. L. Barger, and J. L. Hall, "Speed of light from direct frequency and wavelength measurements of the methane stabilized laser," *Phys. Rev. Lett.*, vol. 29, p. 1346, 1972.
[75] T. G. Blaney, C. C. Bradley, G. J. Edwards, B. W. Jolliffe, D. J. E.

Knight, W. R. C. Rowley, K. C. Shotton, and P. T. Woods, "Measurement of the speed of light," *Nature*, vol. 251, no. 5470, p. 46, Sept. 6, 1974.

[76] T. G. Blaney, C. C. Bradley, G. J. Edwards, B. W. Jolliffe, D. J. E. Knight, W. R. C. Rowley, K. C. Shotton, and P. T. Woods, "Measurement of the speed of light I. Introduction and frequency measurement of a carbon dioxide laser," and "II. Wavelength measurement and conclusion," *Proc. Roy. Soc. Lond. A*, vol. 355, pp. 61–88, and 89–114, 1977.

[77] H. P. Layer, R. D. Deslattes, and W. G. Schweitzer, Jr., "Laser wavelength comparison by high resolution interferometry," *Appl. Opt.*, vol. 15, pp. 734–743, 1976.

[78] J. P. Monchalin, M. J. Kelly, J. E. Thomas, N. A. Kurnit, A. Szoke, A. Javan, F. Zermike, and P. H. Lee, "Determination of the speed of light by absolute wavelength measurement of the R(14) line of the CO_2 9.4 μm band and the known frequency of the line," *Opt. Lett.*, vol. 1, pp. 5–7 and 140, 1977.

[79] J. P. Monchalin, M. J. Kelly, J. E. Thomas, N. A. Kurnit, A. Szoke, F. Zernike, P. H. Lee, and A. Javan, "Accurate laser wavelength measurement with a precision two-beam scanning Michelson interferometer," *Appl. Opt.*, vol. 20, no. 5, pp. 736–757, Mar. 1, 1981.

[80] F. R. Petersen, D. G. McDonald, J. D. Cupp, and B. L. Danielson, "Laser spectroscopy," in *Proc. Vail (CO) Conf.*, R. G. Brewer and A. Mooradian, Eds. New York: Plenum, 1973, pp. 555–569.

[81] P. T. Woods, K. C. Shotton and W. R. C. Rowley, "Frequency determination of visible laser light by interferometric comparison with upconverted CO_2 laser radiation," *Appl. Opt.*, vol. 17, no. 7, pp. 1048–1054, Apr. 1, 1978. *Note*: The speed of light uncertainty excludes a contribution that arises from the practical variation between laboratories in realizing the meter from its definition using the ^{86}Kr lamp. With this contribution, about ± 4 parts in 10^9, the uncertainty becomes ± 1.2 m/s.

[82] K. M. Baird, D. S. Smith, and B. G. Whitford, "Confirmation of the currently accepted value 299 792 458 metres per second for the speed of light," *Opt. Commun.*, vol. 31, pp. 367–368, 1979.

[83] V. M. Tatarenkov, V. G. Il'in, V. I. Kiparenko, V. K. Korobov, and S. B. Pushkin, *Izmeritel'naya Teknika*, no. 2, pp. 15–19, Feb. 1980, *transl.*: "Current state and prospects for frequency measurement in the optical range," *Meas. Techniques (USSR)*, vol. 23, no. 2, pp. 108–114, Feb. 1980.

[84] V. P. Kapralov, G. M. Malyshev, P. A. Pavlov, V. E. Privalov, Ya. A. Fovanov, and I. Sh. Etsin, *Opt. Spektrosk.*, vol. 50, pp. 67–72, Jan. 1981, *transl.*: "Measurement of the wavelength ratio of lasers stabilized by saturated absorption in iodine and methane," *Opt. Spectrosc. (USSR)*, vol. 50, no. 1, pp. 34–37, Jan. 1981.

[85] W. R. C. Rowley and A. J. Wallard, "Wavelength values of the 633 nm laser, stabilized with $^{127}I_2$ saturated absorption," *J. Phys.*, vol. E6, pp. 647–652, 1973.

[86] W. G. Schweitzer, Jr., E. G. Kessler, Jr., R. D. Deslattes, H. P. Layer, and J. R. Whetstone, "Description, performance, and wavelengths of iodine stabilized lasers," *Appl. Opt.*, vol. 12, pp. 2927–2938, 1973.

[87] G. R. Hanes, K. M. Baird, and J. DeRemigis, "Stability, reproducibility, and absolute wavelength of a 633 nm He–Ne laser stabilized to an iodine hyperfine component," *Appl. Opt.*, vol. 12, pp. 1600–1605, 1973.

[88] R. L. Barger and J. L. Hall, "Wavelength of the 3.39 μm laser-saturated absorption line of methane," *Appl. Phys. Lett.*, vol. 22, pp. 196–199, 1973.

[89] K. M. Baird, D. S. Smith, and W. E. Berger, "Wavelengths of the CH_4 line at 3.39 μm," *Opt. Commun.*, vol. 7, pp. 107–109, 1973.

[90] F. R. Petersen, D. G. McDonald, J. D. Cupp, and B. L. Danielson, "Rotational constants for $^{12}C^{16}O_2$ from beats between Lamb-dip-stabilized lasers," *Phys. Rev. Lett.*, vol. 31, pp. 573–576, 1973.

[91] F. R. Petersen, D. G. McDonald, J. D. Cupp, and B. L. Danielson, "Accurate rotational constants, frequencies, and wavelengths from $^{12}C^{16}O_2$ lasers stabilized by saturated absorption," in *Laser Spectroscopy*, R. G. Brewer and A. Mooradian, Eds. New York: Plenum, 1973, pp. 555–569.

[92] J. L. Hall, "Problems with atomic frequency standards of extreme accuracy and a new atomic beam interaction geometry," *Opt. Commun.*, vol. 18, no. 1, pp. 62–63, July 1976.

[93] Yu. S. Domhin, N. B. Koshelyaevskii, V. M. Tatarenkov, and P. S. Shumyatsky, "Precise frequency measurements in submillimetre and infrared region," *IEEE Trans. Instrum. Meas.*, vol. IM-29, pp. 264–267, 1980.

[94] V. P. Chebotayev, "Optical time scale," *J. de Phys. (Colloq. C8)*, supplement to vol. 42, no. 12, pp. C8-505–C8-512, Dec. 1981.

[95] V. F. Zakhar'yash, V. M. Klement'ev, M. V. Nikitin, B. A. Timchenko, and V. P. Chebotaev, *Zh. Tekh. Fiz.*, vol. 53, no. 11, pp. 2241–2244, Nov. 1983, *transl.*: "Absolute frequency measurement of the E-line of methane," *Sov. Phys. Tech. Phys.*, vol. 28, no. 11, pp. 1374–1375, Nov. 1983.

[96] V. P. Chebotayev, V. M. Klementyev, M. V. Nikitin, B. A. Timchenko, and V. F. Zakharyash, "Comparison of frequency stabilities of the Rb standard and of the Ne–Ne/CH_4 laser stabilized to the E line in methane," *Appl. Phys. B*, vol. 36, pp. 59–61, 1985.

[97] G. P. Barwood and W. R. C. Rowley, "Characteristics of a $^{127}I_2$-stabilized dye laser at 576 nm," *Metrologia*, vol. 20, pp. 19–23, 1984.

[98] D. J. E. Knight, "A tabulation of absolute laser frequency measurements," Preprint, Aug. 1985 (submitted to *Metrologia*).

[99] D. J. Wineland, W. M. Itano, J. C. Bergquist, J. J. Bollinger, and H. Hemmati, "Frequency standard research using stored ions," *Progr. Quant. Electron.*, vol. 8, pp. 139–142, 1984.

Time Generation and Distribution

DONALD B. SULLIVAN AND JUDAH LEVINE

Invited Paper

This paper presents a broad overview of time and frequency technology, particularly those trends relating to the generation and distribution of time and frequency signals. The characterization of components and systems is also addressed.

I. Introduction

Recent progress in time and frequency technology indicates that developments taking place today are more far reaching than any of those of the last two decades. The development of laser methods for manipulating the states and motions of atoms should provide dramatic advances in the performance of atomic standards. Advances in satellite time transfer now provide for transfer accuracy and stability well beyond the performance of most atomic frequency standards. Further progress in time transfer should support the application of even the most advanced frequency standards. In parallel with this activity, there is greatly renewed interest in substantially improving the synchronization of telecommunication networks, navigation networks, and electrical power networks. This paper provides a general look at these trends and, when practical, suggests where they are leading. We refer the reader to other papers, particularly to those in this issue, for greater detail.

For brevity, we will use the shorthand approach of simply referring to time measurement or time transfer to imply frequency measurement or frequency transfer as well. Similarly, we loosely use the term synchronization (same time) to imply syntonization (same frequency). Clearly, the requirements for each differ, and a distinction must be made when addressing a specific application.

II. Background

A. Accuracy and Stability

We first emphasize the difference between accuracy and stability. Consider the performance of a cesium-beam frequency standard. The accuracy of the standard describes its ability to generate a frequency where the systematic uncertainties (frequency shifts) relative to the ideal (the model) are known. An accuracy statement involves an upper and lower limit for deviations of the standard from the model. In simple terms, the frequency stability of the standard is a measure of its ability to stay within specific frequency limits for some sampling time, τ. The evaluation of the accuracy of a standard is based on a physical model of the standard. The accuracy statement involves a proper combination of all of the errors derived from independent measurements and the theoretical predictions of the model. Such evaluations are inevitably checked through comparisons among the independently developed primary standards of the world. If a standard is highly accurate, it obviously has very good long-term stability. In the world's standards laboratories many take the unproven, but intuitively appealing, position that the best approach to long-term stability is to improve accuracy. The basis for this position is the idea that long-term variations in output are caused by variations in the systematic offsets. By reducing and controlling the offsets, we thus improve both accuracy and long-term stability. Clearly, the accuracy of a standard can be no better than its long-term stability.

Outside the standards laboratory, in practical situations, stability is often the key consideration. For example, if several nodes in a telecommunications system must be properly timed for synchronous communication, it matters little whether the time delivered to the nodes is accurate. All that really matters is that all nodes measure the same time. If, however, the network is very large (many nodes) and synchronization is acquired from several alternate sources, then it may be necessary to require accuracy as well.

Kartaschoff and Barnes [1], in the last special issue of this journal covering time and frequency, present a broader discussion of accuracy and stability. The measures of stability used in this field have matured to the point where they are now the subject of an IEEE standard [2], [3]. The precursor to this standard is a highly referenced paper by Barnes *et al.* [4]. In the historical development of these measures, time-domain methods have dominated. In this issue, two papers are devoted to the statistical characterization of frequency standards. Rutman and Walls [5] describe both measurement methods and the conceptual framework in which they are made. Percival [6] focusses on the characterization of frequency stability in the frequency domain.

Manuscript received November 1, 1990; revised March 18, 1991.

The authors are with the Time and Frequency Division, National Institute of Standards and Technology, Boulder, CO 80303.

IEEE Log Number 9101204.

B. Frequency Standards

The major advances in timekeeping in this century have been the development of the quartz-crystal oscillator and the development of the atomic clock. Quartz oscillators are used in almost all timing systems. The number of quartz oscillators produced annually is measured in millions of units, far more than any other type of oscillator. Quartz-oscillator technology has been evolving steadily over a long period. The likelihood is the technology will continue this steady improvement as better crystals and clever compensation schemes are developed (see, for example, Vig [7]).

In order of increasing cost, the four main classes of currently used standards are quartz oscillators, rubidium frequency standards, cesium-beam frequency standards, and hydrogen masers. In general, this is the same as the ascending order of performance, except that the cesium standards provide the best accuracy and long-term stability while hydrogen masers provide the best short-term stability. The volume of atomic standards produced annually is measured in thousands of units.

The atomic clocks of today share two common characteristics: 1) The atoms in all of them move at thermal velocities commensurate with room or higher temperatures. This means that Doppler shifts (and corrections for them) are extremely important. 2) The required difference in the population of atoms in the excited and ground states in clocks is achieved through a pumping or selection process which itself adds complications. For example, the Stern-Gerlach magnets used in cesium standards introduce a transverse velocity dispersion which, through the cavity phase shift, produces a frequency error. In another example, the discharge-lamp pumping of rubidium standards involves excess, broad-spectrum light which complicates the pumping process, and the aging of the discharge lamp contributes to drift and instability. The various types of standards are discussed in greater detail by Lewis [8] in this issue.

We are now on the threshold of a revolution in the design of atomic frequency standards. At the heart of this revolution lies laser control of the motions and atomic states of ions and atoms. Some very general aspects of this revolution are described in Section III.

C. Time Transfer Systems

Ten years ago, the methods for time comparison of widely separated clocks involved uncertainties which were greater than the performance of the best primary standards. In this environment, the development of better standards was largely an academic exercise. Signals could not be reliably transferred to other locations and accuracy claims could not be checked by comparison with other standards. The development of accurate satellite time transfer systems, which has taken place over the past two decades, has changed this. Common-view time transfer using the Global Positioning System (GPS) satellites, pioneered by NIST [9], has proven to be extremely accurate offering global coverage with an uncertainty on the order of 10 ns. The two-way method for time transfer [10], [11], which uses standard satellite communication channels, promises even higher performance. These systems put the accuracy and stability of time transfer (or time comparison) ahead of time generation. The development of better standards is thus no longer academic.

D. Network Synchronization

One of the key practical challenges to time and frequency technology is the synchronization of nodes in major networks such as computer networks, telecommunications networks, electrical power networks, and navigation networks. In the transfer of information in a communications system, synchronous operation provides for higher throughput. Synchronization in electrical power networks supports fault location, event recording, and control of system stability. Better synchronization in navigation networks obviously results in more accurate navigation. In this issue Kartaschoff *et al.* [12] address the synchronization of digital communications networks, Wilson [13] discusses timing for electrical power networks, and MacDoran and Born [14] deal with some aspects of ranging/navigation systems. There are a few general network synchronization issues which underlie all of these systems.

Systems for network synchronization can be divided into two broad categories: 1) peer organizations in which groups of nominally identical nodes exchange time data so as to establish a single self-consistent time scale, and 2) stratified, client-server arrangements in which the time of each node is obtained from a relatively small number of sources or possible a single primary source. By using time servers, the time distribution problem boils down to the determination of the errors introduced by the individual delays between the source and the network nodes. The success of the method will depend on how well these errors can be determined. Although groups of peers must also determine the delays between the nodes, the increased symmetry of this configuration may simplify the problem somewhat since the time of a single node is generally compared to the time of several other nodes connected by different paths with different delays. At least in principle, it is possible to perform a dynamic least-squares adjustment which minimizes the average errors over all the nodes. In either configuration, the reliability and short-term stability of the system is substantially improved if the central timing signal is used to steer a local clock at the node rather than being used to directly control the operations at the node. If the central synchronization signal is interrupted, the local clock can carry the system at that node for some period before synchronization is completely lost. Furthermore, the higher frequency noise inherent in time transfer systems substantially compromises the short-term stability of the received signal. Virtually all network synchronization relies on this approach (the control of a local clock by the synchronization signal). The quality of the local clock and the sophistication of the steering are choices dictated by the reliability required in the application.

Where reliability of synchronization is paramount, a second (or even third) independent distribution system (with or without a second independent reference clock) can be used. With two synchronization signals delivered to each node, the loss of one distribution system (or central clock) simply forces each node to rely on the alternate source.

III. Time Generation

Current studies at a number of laboratories are demonstrating principles which, when applied to standards and oscillators, will certainly result in improved performance. For example, Doppler shifts can now be minimized by reducing velocities of beams of atoms and the thermal motions of trapped ions. Newer methods for optical state selection and detection eliminate some of the negative side effects of current methods. In this issue, Ramsey [15] presents a brief history of the development of atomic standards and projects the future development of these standards. Also in this issue, Itano [16] and Rolston and Phillips [17] describe progress toward cooled-ion and cooled-atom standards. In the following sections, we present brief discussions of the various types of clocks and oscillators, indicating the status and directions of work. Finally, Lewis [8], in this issue presents a comprehensive description of the various types of atomic standards and their characteristics.

A. Cesium-Beam Standards

Primary cesium-beam standards are approaching certain practical limitations. For example, in a typical primary standard the correction for the second-order Doppler shift is about 2×10^{-13}. For a system with a design accuracy of 1×10^{-13}, this correction is not too difficult to make, but at a level of 1×10^{-14} it becomes a substantial problem. The narrowest linewidths for the cesium clock transition ($\sim$ 9.2 GHz) are typically tens of hertz, so at an accuracy of 1×10^{-14} the clock servo system must find line center with an accuracy approaching 1×10^{-6}. These are but a few of the reasons we cannot expect to see the performance of these standards go much beyond 1×10^{-14}. Drullinger [18], [19], Ohshima *et al.* [20], and de Clerq *et al.* [21] are building thermal beam systems which use optical state selection and detection rather than the conventional magnet selection. This approach, as well as other variations on the conventional technology, should allow achievement of the 10^{-14} accuracy, but they do not really avoid the problems noted previously.

The real solution to these limitations lies in slowing the atoms so that the Doppler shift is reduced and longer observation times can yield narrower resonance linewidths. Itano [16], in this issue, describes a most direct solution, the trapping and cooling of positive ions. This concept has matured to the point where high-performance prototype standards have been demonstrated. Phillips [17], also in this issue, discusses an approach involving the slowing of neutral atoms. The lack of a suitable trap for neutral atoms limits the achievable linewidth. One proposal for a cooled-atom standard involves a fountain where slowed atoms are lofted vertically and interrogated as they rise and then fall under the influence of gravity. With slowed neutral-atom standards the potential signal-to-noise ratio is better than with trapped ions. Furthermore, the definition of the second is now based on neutral cesium, so primary standards based on slowed cesium atoms would continue to be favored unless other standards prove to be greatly superior.

The two areas in which new concepts will likely affect practical cesium standards (field standards) in the next ten years involve optical state selection/detection and closed-cell standards in which cesium atoms are cooled sufficiently to reduce transition linewidth and Doppler shift. A key advantage of optical state selection and detection is that it can support the use of all atoms in the atomic beam. In conventional magnetic selection, 15/16 of the beam is in the wrong atomic state and has to be discarded. The more efficient use of the beam atoms means that signal-to-noise ratio can be greatly increased (at the same beam flux), or the lifetime of the standard can be increased by operating at lower oven flux (while maintaining respectable signal-to-noise performance). The cesium-cell concept developed by Monroe *et al.* [22] uses slowed cesium atoms contained in a closed envelope similar to that used in rubidium standards (see the discussion of rubidium standards by Lewis [8] in this issue). The cell approach is likely to result in a simpler overall system (once the laser diodes are sufficiently simplified) of very good medium-term stability.

B. Stored Ion Standards

The most readily understood advantages of using trapped ions as frequency standards are that 1) the first-order Doppler shift is eliminated because the ions remain fixed in position and 2) the transition linewidth is dramatically reduced by long observation times. Furthermore, trapped ions can be readily laser cooled minimizing the effect of the second-order Doppler shift. The use of radiation pressure to cool trapped ions was first demonstrated in 1978 by Wineland *et al.* [23] and Neuhauser *et al.* [24]. The cooling of neutral atoms, first demonstrated in 1981 [25], [26], also uses radiation pressure. In these ion and neutral-atom experiments, the fundamental cooling concept is the same, but the implementations are quite different. The ion-storage technology is clearly more mature, and demonstrations of high accuracy and high stability performance already push the performance of conventional cesium standards.

Bollinger *et al.* [27] demonstrated the first high-accuracy prototype ion standard using Be^+ ions. The accuracy was equal to that of NBS-6, the present U.S. primary standard. Cutler *et al.* [28] have demonstrated a very high stability ion standard (not laser cooled) which outperforms cesium in the medium term. Prestage *et al.* [29] have further developed this type of standard. But the real promise for ion standards still lies in the future. The NIST group [30] in studies of an optical-frequency transition in single trapped Hg^+ ions anticipate that systematic frequency shifts can be determined to 1×10^{-18}. The linewidth of this optical transition is 2 Hz providing an inherent Q factor of 10^{15}. Bergquist *et al.* [31] have locked a laser to this transition

and achieved a linewidth of 80 Hz limited by the linewidth of the laser. This represents the narrowest laser linewidth and the highest-Q atomic or molecular transition ever observed. Optical transitions are appealing for frequency standards because they offer the highest Q factors, but there is a fundamental problem in working in the optical region. This is the difficulty of precisely relating the output frequency to a frequency in the microwave frequency region where the performance can be applied to practical electronic metrology. Frequency synthesis (or division) between the microwave and optical regions is extremely difficult. It must be dramatically simplified before optical frequency standards are widely accepted. The Hg^+ ion also has a clock transition at 40.5 GHz. This can be used in standards that do not look too different (electronically) from conventional rubidium standards. Using laser cooling to narrow the linewidth, a clock based on this transition should achieve an accuracy of better than 1×10^{-15}.

One disadvantage of ion standards is that the highest-accuracy, single-ion systems exhibit rather poor signal-to-noise ratios. Prestage *et al.* [29] have shown that a modification of the usual trap to one of linear geometry allows trapping of a larger number of ions without a substantial decrease in the potential accuracy. The larger number of ions provides for a higher output signal and a better signal-to-noise ratio. Preliminary experiments suggest that ion standards based on a linear trap might even challenge active hydrogen masers (see the following) in short term stability.

C. Other Atomic Standards

Two other classes of atomic standards, hydrogen masers and rubidium standards, play a significant role in science and technology. Hydrogen masers are the common choice where very high short-term stability is required, and rubidium standards now provide a cost-effective performance which is better than that of quartz oscillators, although below that of cesium. (Superconducting-cavity-stabilized oscillators [32] actually show the best short-term stability, but they are not yet in common use.)

There are two types of hydrogen masers. Active hydrogen masers are distinguished by the fact that they oscillate spontaneously. Passive masers use an external oscillator to probe the hydrogen resonance. Active masers are critical to very long-baseline interferometry (VLBI), where observations of radio-emitting stars must be time tagged with very high short-term stability. Passive hydrogen masers developed during the last decade are smaller (comparable to the size of a cesium standard) and less expensive than active masers. Their niche is intermediate-term stability, which is generally better than that of cesium standards.

Much of the research surrounding present maser development involves the wall coatings of the bulbs which contain the hydrogen within the microwave cavity. A key characteristic of masers is the very long interrogation time of the hydrogen transition which results from the fact that individual hydrogen atoms can suffer many collisions with the walls in the bulb without disturbing the atomic state in a major way. There is, nevertheless, an energy (frequency) shift associated with the wall interaction. The stability of this shift has long been a source of study and discussion. Wall coating, often more an art than a science, has played a key role in maser performance, and future improvement in maser performance is critically dependent upon it. Masers developed recently in the Soviet Union [33] have performances which suggest a substantial improvement in wall coating. Another approach to the wall problem has been the development of masers which use liquid ^3He as the wall coating [34], [35]. Preliminary experiments with these cryogenic masers look promising. Yet the technology is still very difficult, and it will probably be many years before such masers are ready for practical applications.

Conventional rubidium standards, optically pumped by special discharge lamps, use a passive buffer gas to slow the diffusion of rubidium atoms. This increases the lifetime of the atomic states in the interrogation region resulting in a narrower linewidth. Because these standards perform better than quartz oscillators and sell for only a fraction of the price of cesium standards, they fill an important gap in the technology. However, they do suffer aging effects which give rise to substantial drift, and the excess spectrum of light from the pump lamp adversely affects the signal-to-noise ratio.

Recent research [36], [37] suggests that some of these problems might be overcome through replacement of the discharge lamp with a suitably controlled diode laser operating at the rubidium transition frequency. This appears to minimize (or even eliminate) both the lamp-aging problem and the excess light which affects the signal-to-noise ratio. Some people suggest that the rubidium standard might some day achieve a short term performance challenging that of today's active hydrogen masers.

D. Quartz Oscillators

The state of development of quartz-crystal oscillators is reviewed by Vig [7]. Steady improvements over many years have involved improvement in the quality of materials, introduction of different cuts of crystals, better mounting and packaging, and improvements in temperature control and vibration isolation. Clever schemes for temperature compensation have given rise to the Temperature-Compensated Crystal Oscillator (TCXO), an advance which significantly reduces drift. More recently, microprocessors have been combined with quartz oscillators to yield the Microcomputer-Compensated Crystal Oscillator (MCXO). The microprocessor in this device stores the frequency-versus-temperature characteristics of the crystal, and, using a second mode of the crystal as a thermometer, makes temperature measurements to arrive at more accurate compensation.

Considering the present state of development of quartz oscillators, there is no reason to expect anything other than steady, small improvements in performance. Unlike the atomic standards, there are no new physical principles offering promise of order-of-magnitude advances.

IV. Time Distribution

The transfer medium plays a critical role in time distribution. For example, short wave (HF) distribution of time signals is very cost effective, but with the unpredictable bouncing of the signals between the earth and ionosphere, it is an inappropriate method for high accuracy dissemination. In another example, the transmission delay for satellite distributed time signals is strongly dependent on the carrier frequency. Ionospheric variations in delay can be tens of nanoseconds at the 1 GHz GPS frequencies. These variations are lower by a factor of 100 in the 12–14 GHz band. Clearly, the selection of the linking medium is an important consideration in network synchronization.

A. General Concepts

The general concepts for time transfer are easily sorted into three categories, one-way, common-view, and two-way time transfer. These are shown schematically in Fig. 1. The problem in every case is how to best deal with the delay introduced by the transmission medium, be it a radio-wave path, an optical fiber, or a coaxial cable.

The one-way delivery of a time signal from a source to a user (Fig. 1(a)) is typified by shortwave radio broadcasts such as those emanating from WWV in Colorado. The delays in such broadcasts, which can be tens of milliseconds, are highly variable because the transmission path often involves multiple reflections between the ionosphere and the surface of the earth. The accuracy of one-way transfer can be improved by characterizing the delay of the medium, but in the case of shortwave broadcasts, this characterization involves a large uncertainty. On the other hand, one-way signal delivery through a coaxial cable or from a satellite involves a much more predictable path, and delay corrections can be quite stable. Cable delays are affected, for example, by temperature and stress on the cable, while satellite delays are affected by dispersion in the ionosphere and troposphere and, of course, changes in position of the satellite. In general, one-way systems are limited by the difficulty in obtaining a complete characterization of delay variations.

The common-view method of time transfer (see Fig. 1(b)) can be used to advantage where the path from some reference source to each of two receivers involves a path delay with common characteristics. For nearly simultaneous observations of the common source, the two sites record time differences which are $A-(R+\tau_{ra})$ and $B-(R+\tau_{rb})$, where A, B, and R are the readings of clocks at site A, site B, and the Reference. τ_{ra} and τ_{rb} are the respective delays between the Reference and sites A and B. If the two sites then share their readings and take the difference between them, the difference, $A - B - (\tau_{ra} - \tau_{rb})$, does not contain the common reference, but leaves only the differential delay, $(\tau_{ra} - \tau_{rb})$, as a correction. Clearly, the more common the characteristics of the path, the better will be the time transfer. Furthermore, the accuracy of the time transfer does not depend (to first order) on the quality of performance of the common reference clock. Allan *et al.* [38] have used this technique with the GPS satellite clocks as references. While time directly received from these space-borne clocks is accurate to no better than about 200 ns, they achieve a time transfer accuracy on the order of ±10 ns. Careful error correction and averaging of at least a day's observations are needed to achieve this performance.

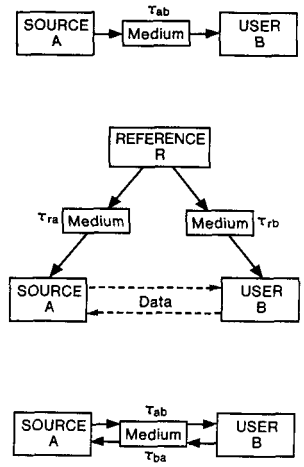


Fig 1. Three general categories for time transfer. (a) ONE-WAY. In the one-way system the timing signal delivered from the source A to the user B is delayedτ_{ab} by the medium. An estimate of this delay can be applied to correct the user clock. (b) COMMON-VIEW. If the source A and the user B make simultaneous observations of the time of a common reference relative to that of their own clocks, then, in subsequently exchanging observations, taking the difference between the two differences removes the reference leaving only the differential delay,$\tau_{ra} - \tau_{rb}$, as the error. Where the two paths share common properties, the differential delay can be substantially smaller than either individual delay. (c) TWO-WAY. In a simultaneous exchange of signals, the source A and user B obtain data which, when compared, yield the difference in readings between the clocks at the two sites plus the difference between the two delays, τ_{ab} and τ_{ba}. If the medium through which the exchange takes place is common, τ_{ab} can almost exactly equal τ_{ba} thus canceling the delay term. This amounts to an accurate calibration of the delay between the two sites.

The common-view method can be extended to the case of more than two receivers, and there may be advantages to such an arrangement. The data from each receiver can be combined with many others to yield simultaneous common-view estimates of the time differences of all of the members of the ensemble, and a least-squares adjustment can then be performed to minimize the error at each node. The power of this procedure will depend on the correlations among the various differential delays.

The two-way method for time transfer (see Fig. 1(c)) provides the best opportunity to accurately determine the transmission delay between separated clocks. When sites A and B simultaneously exchange time signals through the same medium and compare their received readings with their own clocks, they each record the respective differences, $A - (B + \tau_{ba})$ and $B - (A + \tau_{ab})$, where τ_{ba} is the transmission delay from B to A and τ_{ab} is the reverse delay. Taking the difference between these two sets of readings we obtain $2(A - B) - (\tau_{ba} - \tau_{ab})$. Now, if the transmission path is fully reciprocal, that is, if $\tau_{ba} = \tau_{ab}$,

then the difference, $A-B$, is known perfectly and the value of the transmission delay is also determined. The two-way method should provide for near-real-time synchronization of clocks at nanosecond or greater accuracy.

The common-view and two-way methods impose different requirements on the transmission medium. The common-view method depends on the relatively high correlation between the paths from the transmitter to the two receivers. The paths may include arbitrary delays and need not be reciprocal. The two-way method, on the other hand, places a premium on reciprocity and does not depend on the correlation between the delays in different portions of the path. Although there are areas of overlap, each technique also has a domain in which it can provide better performance. The common-view method, for example, is likely to be more robust in computer networks, which have large transmission delays and poor reciprocity. The two-way method is probably the method to choose when signals follow cable or line-of-sight paths.

B. Applications of the Time Transfer Concepts

In the discussion above we used several examples to illustrate the different time transfer methods. But the concepts are general, and each can be applied to a variety of transmission schemes. The discussion in this section is not meant to be comprehensive, but rather to indicate a few of the real and potential applications of the concepts.

Jackson and Douglas [39] and Levine *et al.* [40] describe two-way methods for synchronizing remote clocks through the telephone system. They both use the following variation of the two-way idea. Rather than accomplish the two-way exchange through simultaneous transmissions, these systems send a time signal from site A to site B where it is echoed (reflected) back to site A. Site A then has a measurement of the round-trip delay which, assuming reciprocity in the lines (an assumption good to better than 1 ms for many systems), is divided by 2 in order to provide a direct measure of the delay. The main difference between the methods described in the two publications above is that one takes care of the correction at each receiver and the other makes the corrections at the transmitting end.

Using a similar scheme in cable-connected systems, two-way time transfer can push the stability and accuracy of synchronization into the picosecond regime, even for systems with very long cables. Where the demand for accuracy/stability of distribution of time are high, the use of this method along with optical fiber connections [41] is especially appealing since the fiber transmission provides excellent immunity to noise introduced by ground-isolation problems.

Navigation or ranging systems which are based upon time of flight of radio or optical signals can often be used for time transfer. For ranging applications these systems assume that the time references are known and solve for relative positions of base stations. For time transfer, the known locations of the base stations are given and the relative time at the stations is then determined. This is the basis for the GPS Common-View Method of time transfer described earlier [38]. The common-view technique of VLBI, used to obtain high-resolution images of distant radio stars or relative positions of observation sites, provides another means for highly stable time transfer [42]. In this technique, signals received at two sites from a set of common radio stars are cross correlated (in a computer) in the same fashion that optical signals are combined in an optical interferometer. A key product of this computation is a very precise measure of the relative phase of arrival of signals at the two sites. This can be used to provide an extremely stable measure of the relative stability of clocks operating at the two sites. With the very high cost of radio telescopes and peripheral equipment and the need for extensive computer processing of data, this is an extremely expensive approach, and it is generally used only as a research tool. MacDoran *et al.* [43] have developed a short-baseline interferometer which relies on the same concept. Their scheme uses common-view observations of signals from GPS satellites. Receiver costs are reasonable, but, as with the VLBI technique, this interferometric method still involves substantial computation.

For satellite time transfer, the one-way method remains attractive because of the simplicity of equipment at the receiving site [44]. Two-way time transfer requires broadcasting to a satellite from each station, a process which requires more equipment and special government licensing. The common-view methods call for exchange of information between sites as well as substantial averaging to obtain good accuracy and stability. A one-way satellite dissemination system which promises very good performance has been proposed by Hanson and Howe [45]. They suggest use of commercial communication satellites in the 12–14 GHz region. At these frequencies the ionospheric variation in delay is only a few nanoseconds. The technology for these geostationary satellites has been widely developed for direct-broadcast television, so receivers should be particularly simple and inexpensive. A key facet of their proposal involves the tracking of the satellite (geostationary satellites do move about somewhat) from three sites using the two-way time transfer method. The two-way exchange of signals between all three stations provides ranging information which can establish the position of the satellite. This position is then broadcast with the time signal so that the receiver can correct for variations in the path between the receiver and the satellite. The accuracy of such time transfer should be better than 100 ns and the stability might well be better than 10 ns.

C. Practical Synchronization Limits

Given adequate clocks, the synchronization of two nodes or a network is likely to be limited by the transmission medium and/or the time-transfer method. In many cases it is possible to gain auxiliary information which might allow for a refinement of the estimate of time delay. For example, knowledge of the temperature of a coaxial cable along with an understanding of the delay-temperature relationship can provide for an improvement in one-way synchronization through coaxial cables. In another example, a model of

ionospheric delays is currently used to improve upon GPS common-view time transfer. Furthermore, if variations in delay have suitable statistical properties, these can be used to place a confidence on a particular measurement or, with a proper time constant, to better steer a remote clock to a central standard.

For the one-way method, practical limitations simply involve the variations in path delay and methods used to estimate them. Practical limitations of the two-way method are also readily understood. These limitations relate primarily to the assumption of reciprocity of the signal path through the medium and the transmitting and receiving equipment as well as the simultaneity of the exchange. The current approach to two-way time transfer through communication satellites involves different frequencies for the up links and down links with the satellite. Since the delay is frequency dependent, the path is not fully reciprocal, although the errors introduced appear to be very small [46]. A general analysis of the limitations of common-view time transfer is more difficult for a number of reasons. The common-view cancellation of delay error is usually only partial, and a variety of additional methods are used to reduce the uncertainty in the determination of the remaining differential delay. These methods include straightforward averaging over many observations, cross correlation of the data obtained in observing a number of independent reference sources, and modeling or independent characterization of the transmission medium. Because of this complexity, the limitations of this transfer method involve more detailed consideration.

The synchronization limits imposed by the time transfer process are, in a certain sense, very general limits on time and frequency technology. The development of atomic clocks with performances beyond the time transfer limits makes sense only if such clocks are to be used in isolation. Applications for such isolated clocks are fewer, probably involving only scientific studies.

V. Characterization Of Components And Systems

The wide acceptance of a standard approach (IEEE standard 1139-1988 [2], [3]) to characterization of clocks and oscillators provides the designer with a consistent means for projecting system performance and specifying certain components. A recent edited volume from NIST [47] includes description of the performance definitions along with a collection of papers devoted to measurement methods supporting the definitions.

Unfortunately, there are no consistent standards for characterization of the time transfer links. This is a key problem which will plague systems engineers until it is resolved. Consider the simple hierarchical synchronization shown in Fig. 2. If the four nodes at the bottom represent nodes in a synchronous communications network, then it is the relative times at these four nodes which is important to the system designer. The performance of the individual clocks at each level can be readily characterized, but the links between them have only been treated in isolated ways. Abate *et al.* [48] describe their approach to synchronization of a real telecommunications network. Characterization of the links connecting system clocks is an important implicit consideration in their system. In another paper, Allan [49] has done a preliminary analysis of the properties of GPS common-view links. Clearly these efforts represent only a beginning. We really need to see a conceptually broader approach to this problem. In the long term the technology will require standard definitions and measures of performance which can be easily combined with the measures of clock performance to arrive at an overall estimate of system performance.

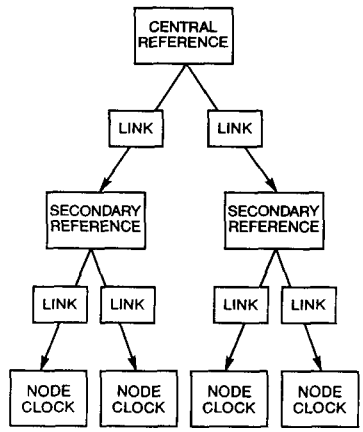


Fig. 2. A simple hierarchical scheme for synchronizing four clocks at four network nodes. The synchronization of such a network is clearly limited by the links between the clocks.

VI. Discussion/Conclusions

A. Clocks and Oscillators

The development of methods for manipulating the states and motions of atoms and ions will certainly result in substantially improved atomic clocks. Since there do not appear to be any new principles to apply to quartz oscillators, we should expect performance improvement to be much less dramatic. There is a strong motivation for improving the performance of clocks only as long as time transfer techniques lead the performance of the clocks. Should clock development outpace transfer systems, then such high performance will be available only in the laboratory. Accuracy which cannot be transferred to another site is probably important only in specialized applications in science (for example, tests of relativity).

B. Time Transfer

The one-way, common-view, and two-way time transfer concepts have been around for many years. The dramatic improvements which we are seeing in time transfer are more closely related to general technological developments which make the implementation of some of these concepts economically feasible. The areas of development which have had impact on time transfer include those in satel-

lite communications, computers and microprocessors, and optical fibers. The focal problem in time transfer continues to be the characterization of the delay through the time transfer medium.

C. Characterization of Components

Because of the importance to network synchronization, we can expect to see efforts directed toward the development of consistent methods for describing the performance of time transfer links. Without standard means for characterization of these links, network synchronization will have to rely on excessive engineering margins, and synchronization levels will remain below the potential of the technology.

REFERENCES

[1] P. Kartaschoff and J. A. Barnes, "Standard time and frequency generation," *Proc. IEEE*, vol. 60, pp. 493–501, May 1972.

[2] *IEEE Standard Definitions of Physical Quantities for Fundamental Frequency and Time Metrology* IEEE Standard 1139-1988, 1988. (Available from IEEE Service Center, 445 Hoes Lane, P. O. Box 1331, Piscataway, NJ 08855–1331.)

[3] D. Allan, H. Hellwig, P. Kartaschoff, J. Vanier, J. Vig, G. M. R. Winkler, and N. Yannoni, "Standard terminology for fundamental frequency and time metrology," in *42nd Ann. Symp. on Frequency Control*, IEEE Catalogue No. 88CH2588-2, June 1988, pp. 419–425.

[4] J. A. Barnes, A. R. Chi, L. S. Cutler, D. J. Healey, D. B. Leeson, T. E. McGunigal, J. A. Mullen, Jr., W. L. Smith, R. L. Sydnor, R. F. C. Vessot, and G. M. R. Winkler, "Characterization of frequency stability," *IEEE Trans. Instrum. Meas.*, vol. IM-20, pp. 105–120, May 1971.

[5] J. Rutman and F. L. Walls, "Characterization and measurement of frequency stability," *Proc. IEEE*, vol. 79, this issue.

[6] D. B. Percival, "Characterization of frequency stability: frequency domain estimation of stability measures," *Proc. IEEE*, vol. 79, this issue.

[7] J. Vig, "Quartz crystal oscillators and resonators for frequency control and timing—a tutorial," AD A218090, 1990. (Available from National Technical Information Service, 5285 Port Royal Rd., Springfield, VA 22161.)

[8] L. L. Lewis, "An introduction to frequency standards," *Proc. IEEE*, vol. 79, this issue.

[9] M. A. Weiss and D. W. Allan, "An NBS calibration procedure for providing time and frequency at a remote site by weighting and smoothing of GPS common view data," *IEEE Trans. Instrum. Meas.*, vol. IM-36, pp. 572–578, June 1987.

[10] D. W. Hanson, "Fundamentals of two-way time transfers by satellite," in *43rd Ann. Symp. on Frequency Control*, IEEE Catalogue No. 89CH2690-6, June 1989, pp. 174–178

[11] Y. Saburi, M. Yamamoto, and K. Harada, "High-precision time comparison via satellite and observed discrepancy of synchronization," *IEEE Trans. Instrum. Meas.*, vol. IM-25, pp. 473–477, Dec. 1976.

[12] P. Kartaschoff, "Synchronization in digital communications networks," *Proc. IEEE*, vol. 79, this issue.

[13] R. E. Wilson, "Use of precise time and frequency in power systems," *Proc. IEEE*, vol. 79, this issue.

[14] P. F. MacDoran and C. H. Born, "Time, frequency and space geodesy: impact on the study of climate and global change," *Proc. IEEE*, vol. 79, this issue.

[15] N. F. Ramsey, "The past, present and future of atomic time and frequency," *Proc. IEEE*, vol. 79, this issue.

[16] W. M. Itano, "Atomic ion frequency standards," *Proc. IEEE*, vol. 79, this issue.

[17] S. L. Rolston and W. D. Phillips, "Laser-cooled neutral atom frequency standards," *Proc. IEEE*, vol. 79, this issue.

[18] R. E. Drullinger, "Frequency standards based on optically pumped cesium," *Proc. IEEE*, vol. 74, pp. 140–142, Jan. 1986.

[19] R. E. Drullinger, D. J. Glaze, J. P. Lowe, and J. H. Shirley, "The new NIST optically pumped cesium frequency standards," *IEEE Trans. Instrum. Meas.*, vol. 40, pp. 162–164, Apr. 1991.

[20] S. Ohshima, Y. Nakadan, T. Ikegami, Y. Koga, R. Drullinger, and L. Hollberg, "Characteristics of an optically pumped Cs frequency standard at the NRLM," *IEEE Trans. Instrum. Meas.*, vol. IM-38, pp. 533–536, Apr. 1989.

[21] E. de Clercq, A. Clairon, B. Dahmani, A. Gérard, and P. Aynié, "Design of an optically pumped Cs laboratory frequency standard," in *Frequency Standards and Metrology, Proc. of 4th Symp.*, A. De Marchi, Ed. New York: Springer-Verlag, 1989, pp. 120–125.

[22] C. Monroe, H. Robinson, and C. Wieman, "Observation of the cesium clock transition using laser-cooled atoms in a vapor cell," *Optics Lett.*, vol. 16, pp. 50–52, Jan. 1991.

[23] D. J. Wineland, R. E. Drullinger, and F. L. Walls, "Radiation-pressure cooling of bound resonant absorbers," *Phys. Rev. Lett.*, vol. 40, pp. 1639–1642, June 1978.

[24] W. Neuhauser, M. Hohenstatt, P. Tsohek, and H. Dehmelt, "Optical-sideband cooling of visible atom cloud confined in a parabolic well," *Phys. Rev. Lett.*, vol. 41, pp. 233–236, July 1978.

[25] S. V. Andreev, V. I. Balykin, V. S. Letokhov, and V. G. Minogin, "Radiative slowing and reduction of the energy spread of a beam of sodium atoms to 1.5K in an oppositely directed laser beam," *JETP Lett.*, vol. 34, pp. 442–445, Oct. 1981.

[26] W. D. Phillips and H. Metcalf, "Laser deceleration of an atomic beam," *Phys. Rev. Lett.*, vol. 48, pp. 596–599, Mar. 1982.

[27] J. J. Bollinger, D. J. Heinzen, W. M. Itano, S. L. Gilbert, and D. J. Wineland, "A 303-MHz frequency standard based on trapped Be+ ions," *IEEE Trans. Instrum. Meas.*, vol. 40, pp. 126–128, Apr. 1991.

[28] L. S. Cutler, R. P. Giffard, P. J. Wheeler, and G. M. R. Winkler, "Initial operational experience with a mercury ion storage frequency standard," in *Proc. 41st Ann. Symp. Frequency Control*, IEEE Catalog No. CH2427–3, May 1987, pp. 12–17.

[29] J. D. Prestage, G. J. Dick, and L. Maleki, "Linear ion trap based atomic frequency standard," *IEEE Trans. Instrum. Meas.*, vol. 40, pp. 132–136, Apr. 1991.

[30] D. J. Wineland, J. C. Bergquist, J. J. Bollinger, W. M. Itano, D. J. Heinzen, S. L. Gilbert, C. H. Manney, and M. G. Raizen, "Progress at NIST toward absolute frequency standards using stored ions," *IEEE Trans. Ultrason., Ferroelect., Freq. Contr.*, vol. 37, pp. 515–523, Nov. 1990.

[31] J. C. Bergquist, F. Dietrich, W. M. Itano, and D. J. Wineland, "Hg+ Single Ion Spectroscopy," in *Proc. 9th Int. Conf. Laser Spectroscopy*, Bretton Woods, NH, M. S. Feld, J. E. Thomas, and A. Mooradian, Eds. San Diego, CA: Academic, June 1989, pp. 274–277

[32] G. J. Dick and R. T. Wang, "Ultra-stable performance of the superconducting cavity maser," *IEEE Trans. Instrum. Meas.*, vol. 40, pp. 174–177, Apr. 1991.

[33] A. A. Belyaev, N. A. Demidov, B. A. Sakharov, V.Yu. Maksimov, M.Yu. Fedotov, and A. E. Yampol'skii, "Ch1-76 small-size passive hydrogen frequency and time standard," *Meas. Tech.*, vol. 30, pp. 767–772, Aug. 1987 and B. A. Gaigerov, L. P. Elkina, and S. B. Pushkin, "Metrological characteristics of a group of hydrogen clocks," *Meas. Tech.*, vol. 25, pp. 23–25, Jan. 1982.

[34] R. F. C. Vessot, E. M. Mattison, R. L. Walsworth, and I. F. Silvera, "The cold hydrogen maser," in *Frequency Standards and Metrology*, A. De Marchi, Ed. New York: Springer-Verlag, 1989, pp. 88–94.

[35] M. C. Hürlimann, W. N. Hardy, M. E. Hayden, and R. W. Cline, "Performance of the UBC cryogenic hydrogen maser," in *Frequency Standards and Metrology*, A. De Marchi, Ed. New York: Springer-Verlag, 1989, pp. 95–101.

[36] J. C. Camparo and R. P. Frueholz, "Fundamental stability limits for the diode-laser-pumped rubidium atomic frequency standard," *J. Appl. Phys.*, vol. 59, pp. 3313–3317, May 1986.

[37] M. Hashimoto and M. Ohtsu, "Improvements in short-term and long-term frequency stabilities of diode laser pumped rubidium atomic clock," in *Frequency Standards and Metrology*, A. De Marchi, Ed. New York: Springer-Verlag, 1989, pp. 434–435.

[38] D. W. Allan, D. D. Davis, M. Weiss, A. Clements, B. Guinot, M. Granveaud, K. Dorenwendt, B. Fischer, P. Hetzel, S. Aoki, M. K. Fujimoto, L. Charron, and N. Ashby, "Accuracy of International time and frequency comparisons via global positioning system satellites in common-view," *IEEE Trans. Instrum. Meas.*, vol. IM-34, pp. 118–125, June 1985.

[39] D. Jackson and R. J. Douglas, "A telephone-based time dissemination system," in *Proc. 18th Precise Time and Time Interval Meeting*, pp. 541–547, Dec. 1986. (Available from Time Ser-

vice, U. S. Naval Observatory, 34th and Massachusetts Ave., N. W., Washington, DC 20392-5100.)
[40] J. Levine, M. Weiss, D. D. Davis, D. W. Allan, and D. B. Sullivan, "The NIST automated computer time service," *J. Res. NIST*, vol. 94, pp. 311–321, Sept.-Oct. 1989.
[41] L. E. Primas, R. T. Logan, Jr., and G. F. Lutes, "Applications of ultra-stable fiber optic distribution systems," in *43rd Ann. Symp. Frequency Control*, IEEE Catalogue No. 89CH2690-6, June 1989, pp. 202–211.
[42] W. K. Klemperer, "Long-baseline radio interferometry with independent frequency standards," *Proc. IEEE*, vol. 60, pp. 602–609, May 1972.
[43] P. F. MacDoran, R. B. Miller, L. A. Buennagel, H. F. Fliegel, and L. Tanida, "Codeless GPS systems for positioning of offshore platforms and 3-D seismic surveys," *Navigation, J. Inst. Nav.*, vol. 31, pp. 57–69, 1984.
[44] A. Sen Gupta, A. K. Hanjura, and B. S. Mathur, "Satellite broadcasting of time and frequency signals," *Proc. IEEE*, vol. 79, this issue.
[45] D. W. Hanson and D. A. Howe, "Industrial time service study," National Bureau of Standards IR 86-3042, 1986. (Available from the Superintendent of Documents, U. S. Government Printing Office, Washington, DC 20402-9325.)
[46] J. Jespersen, "Impact of atmospheric nonreciprocity on satellite two-way time transfers," *43rd Ann. Symp. Frequency Control*, IEEE Catalogue No. 89CH2690-6, pp. 186–192, June 1989.
[47] D. B. Sullivan, D. W. Allan, D. A. Howe, and F. L. Walls, Eds., "Characterization of clocks and oscillators," NIST Tech. Note 1337, March 1990. (Available from the Superintendent of Documents, U.S. Government Printing Office, Washington, DC 20402-9325.)
[48] J. E. Abate, E. W. Butterline, R. A. Carley, P. Greendyk, A. M. Montenogro, C. D. Near, S. H. Richman, and G. P. Zampetti, "AT&T's new approach to the synchronization of telecommunication networks," *IEEE Commun. Mag.*, vol. 27, pp. 35–45, Apr. 1989.
[49] D. W. Allan, "Remote time and frequency comparisons now and in the future," in *4th European Frequency and Time Forum*, Mar. 1990, pp. 619–629.

Donald B. Sullivan was born in Phoenix, AZ on June 13, 1939. He received the B.S. degree in physics from the University of Texas at El Paso, TX in 1961 and the M.A. and Ph.D. degrees in physics from Vanderbilt University, Nashville, TN in 1963 and 1965, respectively.

From 1965 to 1967 he served as an officer in the U.S. Army at Edgewood, Arsenal, MD and in 1967 he joined the National Institute of Standards and Technology (NIST) as a National Research Council Postdoctoral Fellow. In 1969, he joined the permanent staff of the Institute's Boulder, Colorado Laboratories to work on applications of superconducting devices to high accuracy and high speed measurements. After 1977, he gradually moved into program management serving as the Leader of the Cyroelectronics Group for 5 years, as a Program Analyst in the Office of the Director of NIST and as Acting Deputy Director of the Center for Electronics and Electrical Engineering. In 1984, he made a major shift in technical fields, becoming the Chief of the NIST Time and Frequency Division. He has published more than 40 articles on electronic devices and measurement instruments.

Dr. Sullivan was recognized by NIST with the Samuel Wesley Stratton Award, the Institute's highest award for research, for accomplishments that established the feasibility of a fundamental improvement in Josephson voltage standards.

Judah Levine was born in New York City in 1940. He received the B.A. degree from Yeshiva College, New York, in 1960 and the Ph.D. degree in physics from New York University in 1966.

Since 1969, he has been a physicist in the Time and Frequency Division of NIST. His research interests include the application of precision measurement techniques to problems of geophysical interest, especially the design and construction of instruments to measure long-period changes in the positions of points on the surface of the earth. In addition, his recent work has dealt with improving the software that is used to implement the NIST time-scale algorithms. He has also been involved in developing new techniques for disseminating time and frequency information using telephone and computer networks and GPS satellites

Dr. Levine is a member of the American Physical Society, the American Association of Physics Teachers, and the American Geophysical Union.

The BIPM and the Accurate Measurement of Time

TERRY J. QUINN

Invited Paper

Accurate measurements in science and technology require worldwide agreement on standards and units. The role of the Bureau International des Poids et Mesures (BIPM) in assuming the uniformity and reliability of the world's measurement system is described. A brief discussion follows on the meaning of the term accurate measurement. The development of atomic time scales is outlined with particular reference to the Bureau International de l'Heure and the origins of TAI and UTC. The article ends with a brief description of the work currently being carried out at the BIPM, which includes the regular establishment of TAI, its diffusion with UTC, and the associated research work aimed at improving these scales and their accessibility.

I. Introduction

The ability to make precise measurements is one of the prerequisites for the development of advanced technology. While few people would seriously dispute this assertion, few people also would be aware of how precise measurements are made or the extent of the human and other resources that are necessary to enable them to be made. In this introductory article to the special issue on time and frequency, a description will first be given of the organizational structure within which the world's measurement system is maintained and then what it means to make a precise measurement will be briefly examined. This will be followed by an outline of the role played by the Bureau International de l'Heure (BIH) in the establishment of international time scales, and the origins of International Atomic Time (TAI) and Coordinated Universal Time (UTC). Finally a brief description will be given of the present day work undertaken at the Bureau International des Poids et Mesures (BIPM) to establish and improve TAI.

Trade, both national and international, in high technology products, the establishment of communication and navigation networks, the exchange of scientific information, as well as a multitude of pure and applied scientific and technological projects carried out on an international basis, are all highly dependent upon precise measurement. The manufacture of high technology products is itself dependent upon the existence of a traceable chain of precise measurements from the factory floor to national and international standards.

All of this has long been recognized and the increasing demands for national and international uniformity in measurement standards was behind the establishment of the national standards laboratories in the then advanced industrialized countries at the end of the 19th and beginning of the 20th centuries. Principal among these were the Physikalisch-Technische Reichsanstalt (Berlin) founded in 1887, the National Physical Laboratory (Teddington) in 1900, and the Bureau of Standards (Washington) in 1901. Preceding the establishment of these institutes, however, was the setting up of an international agreement on measurement standards and the founding of the Bureau International des Poids et Mesures (BIPM) in 1875 under the Convention du Mètre. Before returning to the question of the meaning of the term "precise measurement" which is fundamental to the whole enterprise of metrology, we first look briefly at the organizational structure established by the Convention du Mètre to ensure worldwide uniformity of measurement.

Manuscript received December 13, 1990; revised February 6, 1991.

The author is with the Bureau International des Poids et Mesures, Pavillon de Breteuil, France.

IEEE Log Number 9100243.

II. The Convention Du Mètre

The Convention du Mètre was signed in Paris in 1875 by representatives of seventeen nations. As well as founding the BIPM and laying down the way in which the activities of the BIPM should be financed and managed, the Convention du Mètre established a permanent organizational structure for member Governments to act in common accord on all matters relating to units of measurement (see Fig. 1). The Convention, modified slightly in 1921, remains the basis of all international agreement on units of measurement. There are now forty-six member nations of the Convention which includes all the major industrialized countries. Under the terms of the Convention, the BIPM operates under the exclusive supervision of the Comité International des Poids et Mesures (CIPM) which itself is elected by and comes under the supervision of the

0018–9219/91/0700–0894$01.00 © 1991 IEEE

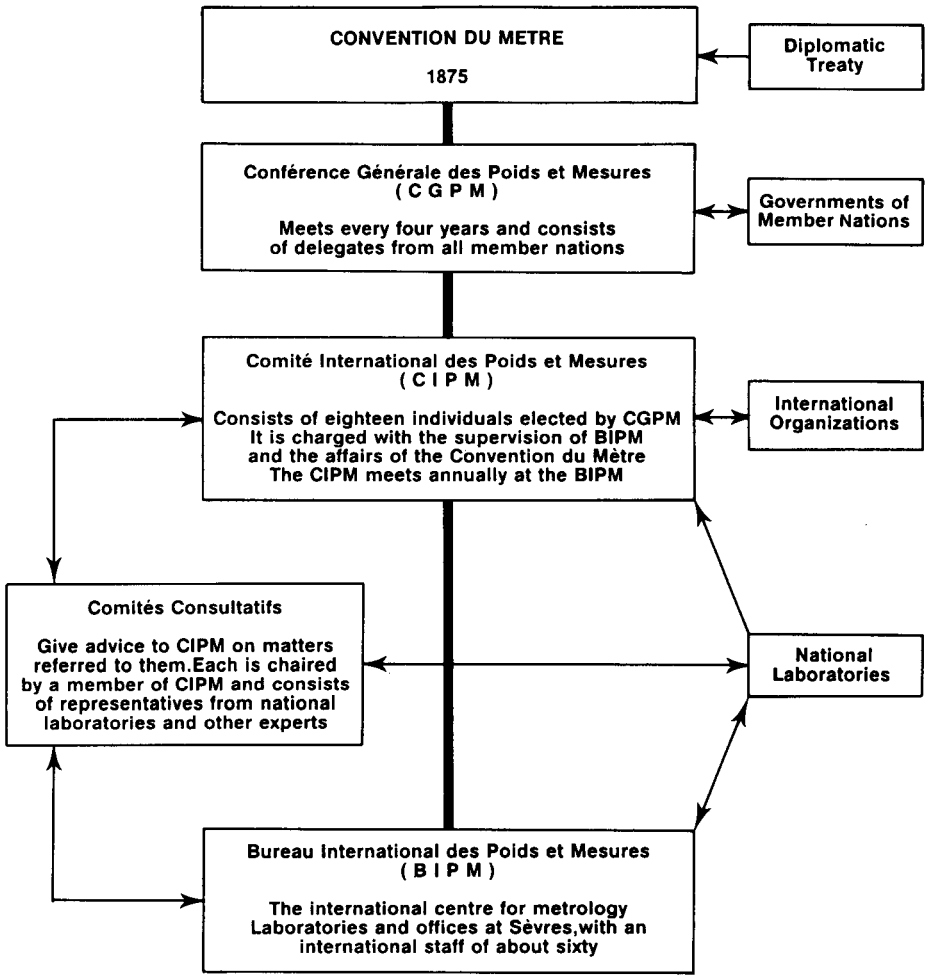


Fig. 1. The organs of the Convention du Mètre. The thick line indicates the line of responsibility, the thin lines indicate interactions between organizations.

Conférence Générale des Poids et Mesures (CGPM). The CGPM meets in Paris once every four years (the nineteenth CGPM will meet in October 1991) and is made up of representatives of the Governments of the member nations. Each CGPM has before it a report by the CIPM on work accomplished since the last Conference, proposals for any changes that may be required in the definitions of units of the SI (International System of Units, the modernized form of the metric system) or structure of the SI and it adopts a program of work and budget for the next four years for the BIPM.

The CIPM, which is made up of eighteen individuals, each a national of a different member state, meets annually at the BIPM. It issues an annual report to member governments on the administrative and financial position of the BIPM and discusses reports presented to it by its consultative committees. The consultative committees are set up by the CIPM to bring together the world's experts in the specified fields as advisers on scientific and technical matters. Among the tasks of the consultative committees are the detailed consideration of advances in physics that directly influence metrology, the preparation of recommendations for discussion at the CIPM, the instigation of international comparisons of standards, and the giving of advice to the CIPM on the scientific work carried out in the laboratories of the BIPM. There are, at present, eight consultative committees; for electricity, photometry and radiometry, thermometry, the definition of the meter, the definition of the second, mass and related quantities, standards of ionizing radiations, and for units. Reports of the meetings of the CGPM, the CIPM and all the consultative committees are published by the BIPM. In addition the BIPM edits and publishes *Metrologia*, the journal of pure and applied metrology now in its twenty-sixth year of publication. A brief report on the activities of the BIPM appears in *Metrologia* in the first issue of each year under the title "News from the BIPM." The scientific work of the BIPM is divided into six sections: length, mass, time, electricity, photometry and radiometry, and ionizing radiations. The work carried out in the time

section is described below; for the rest, the interested reader is referred to [1].

III. The Need For Precise Measurement

Having briefly laid out the organizational structure within which the industrialized nations of the world work to ensure uniformity of measurement standards we now come back to the question of why such work is necessary and what is implied by the term "precise measurement." At one level the problem is a very simple one and can be sketched out in the following way. Suppose that in one locality at a certain time someone makes a measurement of a physical quantity. In a different place at a different time someone else makes a measurement of what is thought to be the same physical quantity and the need arises for the results of these two measurements to be compared. What assurance can we give that these measurements should be the same and how do we find out whether or not they are? To give any answer to these questions we must first be clear as to what it is that we are trying to compare.

In the usual formalism of quantity calculus [2], whose origin is to be found in the work of Maxwell, we write that a quantity having a symbol Q is expressed as a numerical value $\{Q\}$ multiplied by a unit $[Q]$

$$Q = \{Q\} \times [Q]. \tag{1}$$

We can then write

$$Q_1/Q_2 = \{Q_1\}/\{Q_2\} \tag{2}$$

or

$$Q_1 - Q_2 = (\{Q_1\} - \{Q_2\})[Q] \tag{3}$$

and so on. In terms of the questions posed above we can see that for any useful statement to be made about the results of the two measurements, two conditions must be fulfilled. The first is that the physical quantities being measured must be of the same nature and the second is that the units in which the result of the measurements are expressed must also be the same.

At a trivial level, both of these conditions are easily fulfilled, but we do not have to look very far in the real world to find that neither is straightforward. For example, the measurement of the diameter of apertures for radiometric purposes poses particular problems because it is difficult to characterize the quantity that is being measured. The best apertures from the optical point of view should be made of an opaque material having a sharp edge, i.e., zero thickness. Apertures having very thin edges can be made but by their very nature thin edges have no strength and so the position of the edge can only be measured by a noncontact method. The only noncontact method that seems feasible is an optical one, but as is well known the apparent position of an edge when measured optically can change depending upon the conditions of illumination and wavelength of the light used to observe it. The quantity we are trying to measure, namely the diameter of an aperture, thus appears to be a function of one of the parameters of the measuring system. Not only that, but if the aperture is to be used to define a beam of electromagnetic radiation its cross sectional area is given by $\pi[R(\lambda)]^2$ where its radius $R(\lambda)$ is indeed a function of the wavelength of the radiation. In fact the accuracy of the most precise optical radiometry is at present limited by the accuracy of the measurement of the cross-sectional area of apertures [3]!

Suppose that we wish to determine the ratio of the magnitude of the two quantities (assumed now to be the same) expressed by (2). This requires that the unit $[Q]$ is the same when the two quantities Q_1 and Q_2 are being measured. If Q_1 is about the same size as Q_2 this may not be too difficult to check, but if $Q_1 \approx 10^6 \times Q_2$ it is by no means easy to verify that the units are the same. The present limitation to the accuracy of our knowledge of the Rydberg constant R_∞ is set by just this problem. The most accurate measurement of R_∞ is that obtained by Biraben *et al.* [4] using Doppler-free two-photon spectroscopy of Rydberg levels $(2S - nD,\ \mathrm{n} = 8, 10)$ in hydrogen and deuterium atoms. The accuracy of the spectroscopy in the visible is set by the accuracy in the knowledge of the frequency of an iodine stabilized laser at a wavelength of about 633 nm namely 1.6 parts in 10^{10}. The estimated accuracy in the practical realization of the SI second is about three orders of magnitude better than this but in a frequency domain itself some five orders of magnitude different. The establishment of the frequency chain between the caesium clock at a frequency of about 9×10^9 Hz and the frequency of certain transitions in molecular iodine at about 10^{15} Hz has been, and remains, one of the major challenges of fundamental metrology.

Thus in order to make what we have loosely termed precise measurements we have first to specify closely the quantity to be measured and second establish a convenient unit whose multiples and submultiples can be obtained over the whole range of the values of the quantity normally encountered. The definition of a unit of mass in terms of the mass of the proton despite its many attractions, is of little practical use as we do not yet know how to relate such a mass ($m \approx 10^{-27}$ kg) to that of everyday objects ($m \approx 1$ kg) with a sufficient accuracy.

We have so far considered only the problem of making measurements of one physical quantity. We require much more than this before we can discuss what is meant by the term "precise measurement." One of the basic postulates underpinning our understanding of the external world is that nature is internally self-consistent and continuous in space and time. For our theories of physics to be consistent with nature, the results of observations or measurements made of one quantity in one domain must be consistent with observations and measurements made of other quantities in other domains. The system of measurement through which this consistency is confirmed is based upon the notion that a small number of base quantities may be considered dimensionally independent and may be described in terms of a set of the same number of arbitrary chosen base units. The latter together with their multiples and submultiples, the various derived units and supplementary units and rules

for their use make up the International System of Units (SI) [5]. The SI units form a coherent set of units in the sense normally attributed to the word "coherent," i.e., a system of units mutually related by rules of multiplication and division without any numerical factor. The set of seven SI base units are by convention regarded as dimensionally independent.

The requirement for mutual consistency among the practical realizations of the units of the SI can easily be demonstrated. Suppose that we set up a blackbody at a thermodynamic temperature T and measure the total radiant exitance $M(T)$, which has the units of watt per square meter, using an electrically calibrated radiometer based upon laboratory realizations of the volt and the ohm. We can also calculate the total radiant exitance of the blackbody from the Stefan-Boltzmann Law

$$M(T) = \frac{2\pi^5 k^4}{15h^3c^2}T^4. \tag{4}$$

To do this, however, we must use measured values of the Boltzmann constant k [$(1.380\ 658 \pm 0.000\ 012) \times 10^{-23}$ J K^{-1}] and the Planck constant h [$(6.626\ 075\ 5 \pm 0.000\ 004) \times 10^{-34}$ J s]. The mutual consistency of the practical realization of our system of units can be checked by carrying out such an experiment, since the measured values of k and h rely upon quite different combinations of constants and units from that of (4).

The extent to which measured values of different quantities involving different units and fundamental physical constants are in agreement is a measure, and the only one, of the accuracy, i.e., the closeness to the real physical world of our measurement system. This is what we mean by the term "accuracy."

Estimates of accuracy based upon judgements of the uncertainty of all the components of a measurement must always be made, of course, but experience shows that these estimates have a tendency of giving optimistic estimates of accuracy, when finally verified by experiments of the sort mentioned above.

A less demanding requirement, which under certain conditions may be all that is necessary, is that of reproducibility. A measurement made under particular conditions by instruments of similar design can often appear to be very reproducible. It is often claimed that for most practical purposes in industry reproducibility is all that is needed. In the short term and within a limited locality, this may well be true. The reliance upon reproducibility alone, however, can never guarantee long-term stability or worldwide uniformity. It is only by firmly linking measurement standards through physics to the invariant natural world that these can be assured and therefore we must ensure that our measurement standards are "accurate" in the strict sense defined above.

The higher the accuracy that we demand, the more complete must our understanding be of that part of the physical world we are dealing with. In few areas of measurement is this more evident than in the one of time and frequency with which we deal. It is striking that not so many years ago the general theory of relativity was considered by most physicists as abstruse and academic, far from the world of practical experience. Of course, it explained the anomaly in the precession of the perihelion of the planet Mercury, but its predictions of other measurable effects were few indeed. The development of atomic clocks, their use in earth satellites and, above all, in the GPS for time transfer have brought the time and frequency community face to face with the consequences of both special and general relativity in a way undreamed of before. For example, the question of reference frames, central to the relativistic view of the world, is by no means a simple one in the context of high-accuracy time-transfer and is now the subject of detailed study [6]. In the particular field of accurate time measurement, however, the use of quantity algebra as represented by (2) and (3), for example, is by no means straightforward. As we shall see, International Atomic Time, Coordinated Universal Time, and Ephemeris Time, and pulsar time are all different physical quantities, which we need to relate to the same unit, the SI second. But, the SI second is a unit whose value depends upon the reference frame within which it is established. Much of the conceptual, and often real, difficulty in accurate time measurement thus stems from the problem of ensuring that the quantities are properly specified and are indeed being measured in the same units. For a more extended discussion on the needs for accurate measurement in science and industry see [7] and [8].

IV. The Bureau International De L'Heure And The First Atomic Time Scales

The Bureau International de l'Heure (BIH) has played a central role in the establishment of the world's time scales since it was founded before the first World War. The initiative for its creation came from the French Bureau des Longitudes which in 1912 called for an international conference to consider the worldwide unification of time-keeping. This conference took place in 1912 and agreement was reached on the principle of establishing an international organization for time. This scientific conference was quickly followed by a diplomatic conference in 1913 at which an international convention was adopted by thirty-two nations creating an Association Internationale de l'Heure under which would be established a Bureau International de l'Heure. This diplomatic convention, however, was never ratified by the participating governments owing to the outbreak of the first World War.

Nevertheless, the BIH, foreseen in the Convention, began operations at the Paris Observatory. When, in 1919, the International Astronomical Union was created a Time Commission was immediately set up within the IAU to give these operations a more formal setting. The BIH was thus finally established on an international basis on January 1, 1920. From then until the 1950's, the international time scale was based wholly upon astronomical observations of the rotation of the Earth. Correlation of observations of the rotation of the Earth was, therefore, one of the main

activities of the BIH. The unit of time, the second, was also defined in terms of astronomical data and distributed by observatories. After the development of operational caesium clocks, it very quickly became apparent to some that the world's time scale would in the future be based upon atomic clocks. Even before the caesium clock, however, long and difficult studies of orbital motions in the solar system had led to the conclusion that the rotation of the Earth is not uniform in the long term. In the 1930's, pendulum clocks had just about reached the level at which the seasonal variations in the Earth's rotation rate could be observed. Such variations became very clear just before the second World War with the advent of quartz clocks. Even with the development of the first atomic clocks, and much improved quartz clocks in the 1950's, efforts continued to maintain the astronomical basis for time. By changing in 1960 from a definition in terms of the rotation of the Earth about its axis to one in terms of the orbit of the Earth about the Sun (ephemeris time), it was hoped to overcome problems set by the nonuniformity of the rotation of the Earth. This was to no avail, however, and only seven years later at the thirteenth CGPM the SI unit of time was redefined in terms of the frequency of a hyperfine transition of the ground state of the caesium 133 atom.

By this time the BIH was already involved with the comparison of atomic time scales. Since 1955 it had been providing a mean atomic time scale based upon frequency comparisons using the very low frequency carriers (3 to 30 kHz) used for long distance communications and radio navigation. This scale was disseminated by issuing corrections to the times of emission of radio time signals. The accuracy of reading was about 1 ms and the scale was, in the beginning, first designated AM (1955–1957) and then A3 (1958–1968).

V. International Atomic Time And Its Transfer To The BIPM

The accuracy with which clocks could be compared on a routine basis improved considerably with the introduction in 1969 of the Loran C system across the Atlantic Ocean and the calibration of propagation delays by means of clock transportation. It became possible to establish an accurate atomic time scale based upon an average of clock readings. From 1969 to 1972, the BIH time scale was the average of between three and seven laboratory atomic time scales, each based upon several caesium clocks. This time scale was called Temps Atomique (TA) or TA(BIH) until 1970 and thenceforth Temps Atomique International (TAI) or International Atomic Time. This scale could be read to a few tenths of a microsecond. In 1973 the BIH began processing the data of individual clocks using a new algorithm designed to give the best possible long-term stability. This algorithm, ALGOS, includes a general data processing scheme and an adaptive statistical treatment which has been regularly updated to match the performances of the clocks. It remains today the algorithm with which TAI is derived. Soon after ALGOS was introduced, improvements in the performance of clocks showed that the TAI frequency was too high by about one part in 10^{12} and on January 1, 1977 this was corrected by a reduction of exactly one part in 10^{12} in the TAI frequency. At the same time a procedure (known as "steering") was introduced for maintaining the TAI frequency accuracy.

Although the IAU had effectively taken responsibility for the world's time scale when it took responsibility for the BIH in 1920, the change from an astronomical basis to an atomic clock basis which began in the 1950's would inevitably lead to a change in the responsibility for the world's time scale. This began with the establishment in 1956 of the Comité Consultatif pour la Définition de la Seconde (CCDS), a committee which includes representatives of national standards laboratories and the scientific unions, principally the IAU. From the beginning of atomic time, the CCDS has provided the technical guidance first to the BIH and now to the BIPM through a working group of the CCDS specially established for this purpose. Finally in 1987 when the CGPM took responsibility for TAI after the whole of the work had been transferred from the Paris Observatory to the BIPM, the IAU also adopted a resolution transferring responsibility to the BIPM. On January 1, 1988, the BIH ceased to exist, and a new International Earth Rotation Service (IERS) was established based in part upon the section of the BIH at the Paris Observatory previously concerned with Earth rotation.

VI. Coordinated Universal Time (UTC)

So much for the development and formal aspects of TAI. Before coming to the technical aspects of the present day work at the BIPM on TAI we must mention Coordinated Universal Time (UTC).

When caesium clocks began to run on a continuous basis they had, like any other clocks, to be set to time. It was agreed that the new atomic time scales should be set so that atomic time was approximately equal to UT2 (Universal Time UT1 which is derived from transit times of stars, corrected for seasonal variations) at 0 hour UT on January 1, 1958. The scale unit, the second, of the new atomic time was chosen to be equal to the second of ephemeris time which was in fact the value of the second calculated from astronomical observations averaged over the eighteenth and nineteenth centuries! The rate of rotation of the Earth since about the beginning of the nineteenth century had already decreased significantly and thus the scale unit of the new atomic time was already significantly different from the second calculated from the rate of Earth rotation in 1958. This explains the relatively large divergence between atomic time and universal time since 1958 illustrated in Fig. 2. In addition to this large divergence due to the initial offset, smaller changes in the Earth rotation rate have occurred since 1958 and these account for the slope changes visible in Fig. 2. Those users of Universal Time needing precise real time values of UT1–TAI, for space navigation for example, can be kept fully informed by means of rapid data processing and fast communication

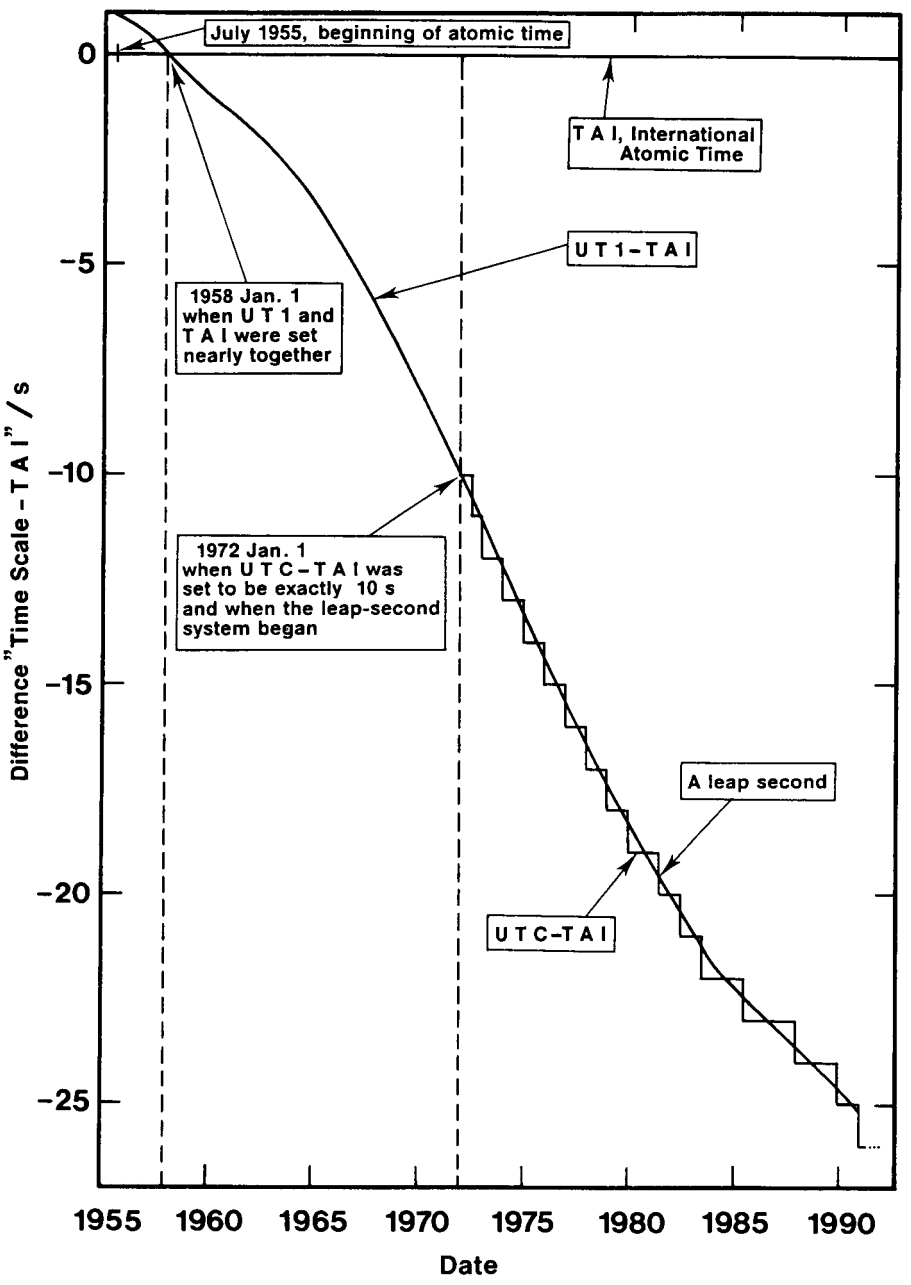


Fig. 2. The divergence between TAI and UT1 since 1955 and the insertion of leap seconds to produce UTC.

systems. For the vast majority of users, however, no such urgency exists, but instead they need an approximation to UT1 in real time. This need was recognized by the definition of UTC which is close to UT1 (which is in turn approximately the same as the old, more loosely defined GMT) but has an exact relation to TAI. Since January 1, 1972 the International Radio Consultative Committee (CCIR) has defined

$$\mathrm{TAI} - \mathrm{UTC} = n \text{ seconds}, \; n \text{ being an integer or zero}$$

and

$$|\mathrm{UT1} - \mathrm{UTC}| < 0.9 \text{ s}.$$

UTC is thus TAI with the addition from time to time of positive or negative leap seconds. The decision to insert a leap second is taken by the IERS and it normally takes place either on July 1 or January 1. Figure 2 shows the departure of UTC from TAI and the insertion of leap seconds. UTC is now the scale used for all radio time signals and is the basis of public time throughout the world after taking

into account the differences of integral numbers of hours required by the time zone system.

VII. TAI And The Work Of The BIPM

Coming now to the work of the BIPM, we can divide the activities into two parts, the first is the establishment of TAI and the diffusion of TAI and UTC to users through the monthly Circular T, and second is the research work which supports and contributes to TAI.

A. The Production of TAI

TAI is a time scale established on the basis of clock comparison data supplied to the BIPM by participating laboratories and a particular algorithm for treating the data known as ALGOS. TAI is a coordinate time scale defined in a geocentric reference frame with the SI second as realized on the rotating geoid as the scale unit. This statement that TAI is a coordinate as opposed to a proper time scale was made by the CCDS in 1980 and provides the necessary information for evaluating the relativistic terms in the establishment of TAI and for its use in nonterrestrial reference frames.

The main aim of TAI is to provide a time scale that is accurate, i.e., having its unit as close as possible to the SI second, reliable and above all stable in the long term. In order to optimize the long-term stability of the scale, at the expense of short-term accessibility, TAI is calculated using data covering an extended period. Clock-comparison data are sent to the BIPM every ten days, on the Modified Julian Dates ending with the figure 9 (01.01.91 is MJD 48257) but blocks of data covering 60 days are taken for the calculation of the scale. The period of 60 days was chosen to place the effective integration time of the scale at the transition between the flicker floor and the random walk frequency modulation of caesium clocks. Stability would not therefore be improved by a longer integration time (see Fig. 3). A period of 60 days is long enough, however, to smooth out the noise coming from the time links (Loran C and GPS) and the white frequency modulation of the clocks. The monthly BIPM Circulars T thus alternate between provisional, based upon only 30 days' data, and definitive based upon a full period of 60 days' data.

The establishment of TAI takes place in three steps:

- the calculation using a post-processing and iterative procedure of an intermediate time scale, known as Échelle Atomique Libre (EAL) or Free Atomic Scale, using the clock comparison data and ALGOS;
- the evaluation of the duration of the scale unit of EAL using data from primary frequency standards and an optimum filter;
- the production of TAI from EAL by applying, if necessary, a correction to the scale interval of EAL to give a value as close as possible to the SI second. This correcting of the scale interval is known as "steering" and is in fact only done at infrequent intervals.

The following gives a brief outline of how TAI is established and is based upon a more detailed description given in [9], but see also [10] and [11].

1) Structure of TAI: The time on EAL at a time t of some reference time scale, EAL$f(t)$, is defined in terms of the readings of $h_i(t)$ of a group of N clocks, H_i, by

$$\mathrm{EAL}(t) = \frac{\sum_{i=1}^{N} p_i[h_i(t) + h_i'(t)]}{\sum_{i=1}^{N} p_i} \tag{5}$$

where p_i is the statistical weight assigned to clock H_i and $h_i'(t)$ is a time correction designed to ensure time and frequency continuity of the scale when either the weightings of individual clocks or the total number of clocks is changed [11].

However, (5) cannot be used directly because the measured quantities that provide the basic data for the calculation of the scale are not the readings of individual clocks but are the results of comparisons between pairs of clocks. At time t, the slowly varying differences $\zeta_{ij}(t)$ between the readings of clocks H_i and H_j are written

$$\zeta_{ij}(t) = h_i(t) - h_j(t). \tag{6}$$

We let the output of the calculation of EAL be a set of N values of the differences $x_i(t)$ defined by

$$x_i(t) = \mathrm{EAL(t)} - \mathrm{h_i(t)} \tag{7}$$

where x_i are the differences between the readings of the individual clocks and the time defined by EAL. Thus (6) can be written

$$x_i(t) - x_j(t) = -\zeta_{ij}(t) \tag{8}$$

and the defining (1) can be transformed into

$$\sum_{i=1}^{N} p_i x_i(t) = \sum_{i=1}^{N} p_i h_i'(t). \tag{9}$$

In the practical computation, a nonredundant system of $N - 1$ time links is employed and (8) and (9) have a strict solution.

2) Weighting procedure: The weight assigned to each clock is calculated in such a way as to favor the long-term stability of the resulting scale and to minimize the annual fluctuations and the drift in its frequency with respect to primary frequency standards. An important feature of ALGOS is that the evaluation of the weight of the clock, although based upon data covering a whole year, takes into account the 60 days of data for which EAL is being computed. It is thus possible to judge clocks on their actual performance during the interval of time during which EAL is being established. It is also possible to take account of any abnormal behavior observed in an individual clock by adjusting its weight, if necessary to zero. This has proved useful on many occasions. The weight is normally based on the variance $\sigma_i^2(6, \tau)$ of mean rate (see (12)) with respect to EAL calculated over six two-monthly samples. This variance, instead of the usual pair variance, was chosen because it gives a greater reduction in the weights of clocks

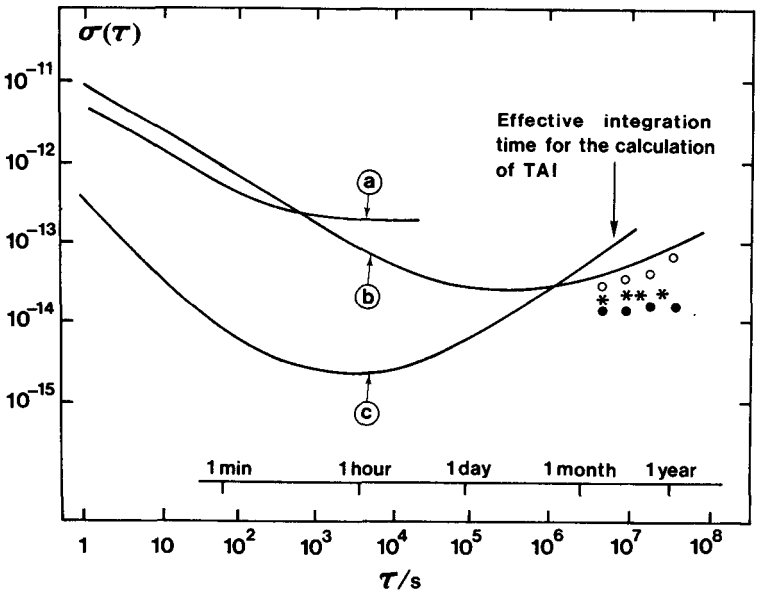


Fig. 3. The stability of various atomic clocks or frequency standards: (a) rubidium clock; (b) average industrial caesium clock, o long-term stability of the best industrial caesium clocks, • the long-term stability of the best laboratory caesium clock; (c) hydrogen masers, * long-term stability of a hydrogen maser with automatic tuning.

showing a frequency drift. The weights are obtained directly from

$$p_i = \frac{1000}{\sigma_i^2(6,\tau)} \tag{10}$$

(σ_i being expressed in ns/day) provided that over the current period of 60 days no abnormal behavior is apparent. In the case of abnormal behavior a weight of zero is assigned. A maximum weight of 100 is assigned to the 15% or so of clocks having $\sigma_i(6,\tau) \leq 3,16$ ns/day. The maximum weight is chosen to ensure that the scale is heavily biased in favor of the best clocks without allowing any one to become predominant, no clock having a participation greater than about 2%.

3) Rate prediction: The time correction term $h_i'(t)$ is made up of two components

$$h_i'(t) = a_i(t_0) + B_{ip}(t) \cdot (t - t_0) \tag{11}$$

where $a_i(t_0)$ is simply the time difference between the clock H_i and EAL at a time t_0, which is the time at the beginning of the 60 day period, and $B_{ip}(t)$ is the predicted difference in rate between the clock H_i and EAL for the period between t_0 and t where the rate of clock H_i, for example, is defined by

$$\text{rate} = \frac{a_i(t_0 - t) - a_i(t_0)}{(t - t_0)} . \tag{12}$$

The prediction of $B_{ip}(t)$ is obtained by a one step linear prediction based upon the previous value. This is justified by the fact that the period of 60 days is such that the predominant clock noise is random walk for which the most probable estimate for the value over the next period is simply that over the immediately preceding period.

Having established the best estimate of EAL the transformation to TAI is made by the decision whether or not the rate of EAL differs sufficiently from the rate of the best primary standards to warrant any correction or "steering" [12]. From 1984 to 1989, no steering was necessary and thus during this period TAI was in fact simply EAL with a constant frequency offset. Since June 1989, however, six frequency changes, each of five parts in 10^{15} have been found necessary.

Finally, the output of these calculations is presented in the monthly Circular T, an example of part of one of these is shown in Table 1, which is distributed to participating laboratories. Remember that this Circular is alternately provisional or definitive depending upon whether it is issued in the middle or at the end of a 60 day calculation period.

B. Research Work on Time Scales and Time Comparisons

The research work at the BIPM which is carried out to back up the regular production of TAI is both theoretical and practical. A general study of the properties of algorithms aims to find improved algorithms for time scales. Several time scale algorithms used in national timing centers and based on the Kalman filter have been analyzed. Because the Kalman filter is linear, optimum in the least squares sense and recursive, it is an efficient and well-adapted tool for smoothing out clock noise and for predicting clock behavior. Its use, together with a definition of the ensemble time and an optimized determination of weights for long-term stability, could help in improving short-term stability of the resulting scale. For the time being, however, this has been tested only on simulated data [13]. Various other studies are underway concerning

Table 1(a) Extracts from BIPM Circular T 34 (December 3, 1990)

UTC—Computed values of UTC - UTC(k)				
(From January 1, 1990, OhUTC, to January 1, 1991, OhUTC, TAI - UTC = 25 s) (From January 1, 1991, OhUTC, until further notice, TAI - UTC = 26 s)				
Date 1990 (OhUTC)	September 25	October 5	October 15	October 25
MJD	48159	48169	48179	48189
Laboratory k		UTC-UTC(k)	(Unit = 1 μs)	
AOS (Borowiec)	−1.36	−2.80	−4.31	−5.80
APL (Laurel)	−0.31	−0.35	−0.40	−0.44
AUS (Canberra)	−0.26	−0.23	−0.18	−0.12
BEV (Wien)	3.80	3.01	2.26	1.48
CAO (Cagliari)	4.20	3.93	3.50	3.16
CH (Bern)	0.24	0.27	0.28	0.34
CRL (Tokyo)	0.61	0.68	0.76	0.86
CSAO (Lintong)	−7.63	−7.35	−7.20	−7.13
DPT (Pretoria)	−21.83	–	–	−22.14
FTZ (Darmstadt)	16.69	16.86	17.05	17.19
TAI and local atomic time scales TA(k). Computed values of TAI - TA(k)				
Date 1990 (OhUTC)	September 25	October 5	October 15	October 25
MJD	48159	48169	48179	48189
Laboratory k		TAI-TA(k)	(Unit = 1 μs)	
NIM (Beijing)	−10.09	−10.30	−10.53	−10.71
NISA (Boulder) (3)	−45067.61	−45067.85	−45068.09	−45068.32
NIST (Boulder)	−45150.75	−45151.38	−45152.00	−45152.63
NRC (Ottawa)	16.91	16.85	16.77	16.73
PTB (Braunschweig)	−359.67	−359.67	−359.68	−359.65
RC (Habana) (4)	−269.00	−269.54	−270.19	−270.89
SO (Shanghai)	−44.31	−44.41	−44.62	−44.59
SU (Moskva) (1)	2827259.02	2827259.01	2827259.00	2827258.90
USNO (Washington DC) (5)	−34609.75	−34610.37	−34611.01	−34611.65

correlations among contributing clocks and the optimization of the weighting procedures.

In the field of time links, the work has been dominated by the consequences of improvement of GPS time comparisons [14]. The introduction of GPS for regular time comparisons between clocks participating in TAI has led to an improvement of at least an order of magnitude, to about 10 ns, in the accuracy of such time comparisons and the potential accuracy is well below 1 ns. To achieve such an accuracy, however, many improvements in technique have yet to be made. Nevertheless, the use of GPS has already had a major influence on the work of the BIPM. The BIPM now issues tracking schedules to laboratories for time comparisons using GPS and is a member of the GPS Civil Users Steering Committee. One of the unexpected results of the BIPM study of the new high-accuracy clock comparisons via

Table 1(b) Extracts from BIPM Circular T 34 (December 3, 1990)

UTC - GPS time and TAI - GPS time

UTC - GPS time = −6 s + Co, TAI - GPS time = 19 s + Co

The GPS data taken at the Paris Observatory, from Block I satellites only, are usually corrected for the measured ionospheric delays and then smoothed to obtain daily values, at OhUTC, of UTC(OP) - GPS time. Co is derived from them using linear interpolation of UTC - UTC(OP).

The DC values are the residuals to the smoothed data for the middle of the 13-minute tracking period. This time is given for the first date of the table. The DC values are reported here only to show the quality of the synchronization.

UTC may be derived from observation from observation at any site of any listed satellite, by interpolating Co. The quality of the access to UTC mainly depends upon local conditions of observation.

		Co		DC(ns)	Block I		
		ns	PRN 6	PRN 9	PRN 13	PRN 12	PRN 3
Date			NAV 3	NAV 6	NAV 9	NAV 10	NAV 11
1990	MJD	OhUTC	Oh36 m	3 h 32 m	4 h 4 m	4 h 20 m	7 h 32 m
September 25	48159	−138	−13	−4	−15	20	2
September 26	48160	−137	−2	−1	−11	–	6
September 27	48161	−140	−2	−3	−3	32	6
September 28	48162	−149	−6	−2	7	6	−4

UTC - GLONASS time

UTC - GLONASS time = C1

From his current observations of both the GPS and GLONASS satellite systems Prof. Daly, University of Leeds, establishes and reports GPS time-GLONASS time at ten-day intervals, together with the standard deviation SD of his daily GLONASS data. C1 is then derived using UTC - GPS time of Section IV.

UTC(USNO) - UTC(SU) is also provided by Prof. Daly.

Date	MJD	C1	SD	UTC(USNO)-UTC(SU)
1990		(μs)	(μs)	(μs)
(OhUTC)	(OhUTC)			
September 25	48159	7.28	0.06	9.28
October 5	48169	6.69	0.08	9.24
October 15	48179	6.05	0.06	9.18
October 25	48189	5.49	0.06	9.02

GPS was the discovery that the geographical coordinates of the antennas of many of the contributing laboratories were significantly in error. Coordinate corrections were calculated for twelve European, four North American, one Middle Eastern and six Far Eastern laboratories. These new coordinates, in the reference frame ITRF 88 of the IERS, were adopted and came into operation on June 12, 1990. The estimated accuracy of the new coordinates in the ITRF 88 is about 50 cm and their adoption corrected errors which in some cases amounted to many meters.

Direct measurement of ionospheric delays can also lead to significant improvements in the accuracy of time comparisons. Experiments are being carried out at the BIPM using different types of dual-frequency receivers. Another activity is the use of improved ephemerides to describe satellite positions. It may become necessary to use these improved ephemerides in current work when, as is likely, the broadcast ephemerides are deliberately degraded for

Table (c) Extracts from BIPM Circular T 34 (December 3, 1990)

Measurement of UTC(j) - UTC(k)

Date	MJD	Time comparisons	Uncertainty source method
1990		(Unit:1 μs)	
June 22	48064.00	UTC(OP) - UTC(SU) = 10.661	0.019 OP report (1)

(1) GPS receiver transportation and common-view method

Duration of the TAI scale interval

The following table gives the departure D of the duration of the TAI scale interval from the SI second at sea level as realized by a given primary frequency standard occasionally evaluated or continuously operating as a clock. In the later case the chosen two-month period of observation is also indicated. The last communicated estimate of the inaccuracy of the standard provides the uncertainty s of the D value.

D and s are expressed in 1*10**−14 second

Standard	Observation period	D	s
NRC-CsV	48129-48189	+10.3	10.0
PTB-CS1	48129-48189	+0.1	3.1
PTB-CS2	48129-48189	+4.8	1.5

The estimate of the duration of the TAI scale interval, computed by the BIPM, from all the available measurements of the TAI frequency, obtained by comparison with primary frequency standards continuously observed or occasionally evaluated (CRL, NIST, NRC, PTB, SU), is:

$$1 + 4*10**-14 \quad +\text{or}- \quad 2*10**-14$$

in SI second at sea level, for the two-month interval 48129-48189.

security reasons. To maintain practical experience in time comparison the BIPM has begun the establishment of a time station using commercial caesium clocks and GPS receivers. We are grateful in this respect to the U.S. Naval Observatory for the loan of a caesium clock.

The time of arrival of pulses emitted by "millisecond pulsars" is measured in terms of terrestrial atomic time scales. It is important to understand the relationship between atomic time and the time scales associated with pulsars [15]. We do not know, at present, what is the ultimate intrinsic frequency stability of pulses emitted by pulsars but it now appears to surpass even that of the best atomic clocks. To obtain the full benefit of millisecond pulsar timing in various fields of astronomy, the uncertainties due to the terrestrial reference time scales must be minimized. This has led the BIPM to produce atomic time scales calculated retrospectively over a number of years. These time scales, known as TT(BIPMxx) where "xx" refers to the year in which they are established, are obtained by reevaluating all the data since 1977 up to the year xx. Thus with the benefit of hindsight, a scale can be made over the whole period from 1977 to 19xx which is more stable and more accurate than was TAI over the same period [16].

In carrying out the regular establishment of TAI and its diffusion with UTC, close contacts are maintained with the principal contributing laboratories. The coordinating role of the BIPM, particularly as regards time transfer, is increasing in importance. This is reflected in the growing need for close coordination among time laboratories if the full potential of GPS for high accuracy time transfer is to be achieved.

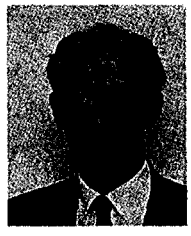
VIII. The Future Of International Time Keeping

What of the future? Time and frequency are domains in which demands for accuracy in measurement for both scientific and technological applications continually approach or even exceed what is possible. For the BIPM, this means that we are often close to or involved in projects that require the ultimate in time and frequency measurement. Studies of the frequency of millisecond pulsars, advances in accurate positioning, for example, are based upon the most accurate time and time transfer. These and others are ultimately limited by the accuracy of the primary frequency standards. The accuracy of primary caesium standards has not improved significantly over the past few years and the number of such standards remains very small. The fact that one laboratory's clocks, those of the PTB, are recognized as being significantly more accurate than those anywhere else is, on the one hand, a great credit to that laboratory but it is, on the other, a considerable worry for the time and frequency community. In the short and medium terms, all that can be done is to encourage other laboratories to try and improve their primary standards. In the long term, the prospects for new primary frequency standards of novel design using, for example, single-ion spectroscopy are very exciting and offer the possibility of accuracies approaching parts in 10^{15} or 10^{16} (see later papers in this issue). The

consequences for time transfer techniques and for TAI of such improvements are difficult to imagine. But what is sure is that if such clocks can be developed, practical uses for the new ultraprecise time will very quickly be found and the demand will exist for the establishment and diffusion of a correspondingly more accurate TAI and UTC. The problems raised by relativity will then become more important and much more difficult, especially for applications in space research and astronomy. Nevertheless, it should be possible to compare, even at this level of accuracy, terrestrial clocks with pulsar clocks in far-distant parts of the universe.

REFERENCES

[1] The BIPM and the Convention du Mètre, BIPM 1987.
[2] J. de Boer, "Group properties of quantities and units," *Amer. J. Phys.*, vol. 47, pp. 818–819, 1979.
[3] T. J. Quinn and J. E. Martin "Cryogenics radiometry, prospects for the future," *Metrologia*, vol. 28, 1991, to be published.
[4] Biraben *et al.*, "Determination of the Rydberg constant by Doppler-free two-photon spectroscopy of hydrogen Rydberg states," *Europhys. Lett.*, vol. 2, no. 12, pp. 925–932, 1986.
[5] BIPM "Le Systé me international d'unités," 6th edition.
[6] G. M. R. Winkler "Synchronization and relativity," *Proc. IEEE*, vol. 79, this issue.
[7] T. J. Quinn "Precision measurement and advanced technology" *IEEE Trans. Instrum. Meas.*, vol. 38, pp. 156–160, 1989.
[8] H. Hellwig "The importance of measurement in technology-based competition" *IEEE Trans. Instrum. Meas.*, vol. 39, pp. 685–688, 1990.
[9] B. Guinot and C. Thomas "The establishment of International Atomic Time" in *BIPM Annual Report of Time Section*, vol. 1, part D, 1989.
[10] P. Tavella and C. Thomas "Comparative study of time-scale algorithms," *Metrologia*, vol. 28, pp. 57–63, 1991.
[11] B. Guinot "Some properties of algorithms for atomic time scale," *Metrologia*, vol. 24, pp. 195–198, 1987.
[12] J. Azoubib, M. Granveaud, B. Guinot, "Estimation of the scale unit duration of time scales," *Metrologia*, vol. 13, pp. 87–93, 1977.
[13] P. Tavella and C. Thomas, "Time scale algorithms: role of the definition of the ensemble time and possible uses of the Kalman filter," in *Proc. 22nd Annual Precise Time and Time Interval (PTTI) Applications and Planning Meeting*, 1990.
[14] W. Lewandowski and C. Thomas "GPS time transfer," Proc. IEEE, vol. 79, July 1991.
[15] B. Guinot and G. Petit, "Atomic time and the rotation of pulsars," *Astron. and Astrophys.*, submitted for publication.
[16] B. Guinot, "Atomic timescales for pulsar studies and other demanding applications," *Astron. and Astrophys.*, vol. 192, pp. 370–373, 1988.
[17] B. Guinot, "Atomic Time" in *Reference Frames in Astronomy and Physics* J. Kovalevsky, I. I. Mueller, B. Kolaczek, Eds. New York: Kluwer, pp. 379–415, 1989.



Terry J. Quinn was born in England in 1938. He received the B.Sc. degree in physics from the University of Southampton, England in 1959 and the D. Phil. degree from the University of Oxford, England in 1963.

In 1962, he joined the National Physical Laboratory, Teddington, England as a Junior Research Fellow and worked on radiation thermometry, temperature scales, and on acoustic determination of the gas constant. From 1967 to 1968 he was a Visiting Scientist at the National Bureau of Standards, Washington, DC, working on radiation and acoustic thermometry. He subsequently became head of the NPL Temperature Section and, in 1975, head of the Mass Section. From 1978 until 1988, he was Deputy Director of the Bureau International des Poids et Mesures (BIPM), Sèvres, France, where he developed his interests in mass and balances. He also continued collaboration in work at NPL on the development of the cyrogenic radiometer which led to determinations of the Stefan-Boltzmann constant and thermodynamic temperatures. From 1984 to 1985 he was a Royal Society Visiting Fellow at the Cavendish Laboratory, Cambridge, England where he worked on the anelastic properties of a Cu-Be alloy used as the flexure element in a flexure strip balance. This balance has been used in experiments to search for a 5th force and anomalous mass in spinning rotors. Since 1988 he has been Director of the BIPM.

GPS Time Transfer

WLODZIMIERZ LEWANDOWSKI AND CLAUDINE THOMAS

Over the past ten years GPS satellites have become the principal tool for national and international comparisons of atomic clocks. Using GPS, time transfer is now ten times more accurate than it used to be using LORAN-C. For the first time, the best frequency standards in the world have been compared in a way which exploits their full level of performance. The present accuracy of GPS time transfer is of the order 10–20 ns for intercontinental distances and 2–3 ns within one continent. Although many improvements have been introduced during the past decade, the next should bring still further improvements leading to an ultimate accuracy at the subnanosecond level.

LIST OF ACRONYMS AND ABBREVIATIONS

BIPM	Bureau International des Poids et Mesures.
CCDS	Comité Consultatif pour la Définition de la Seconde.
CRL	Communications Research Laboratory, Tokyo.
CV	Common view.
DMA	Defense Mapping Agency.
GPS	Global Positioning System.
IERS	International Earth Rotation Service.
ITRF	IERS Terrestrial Reference Frame.
MJD	Modified Julian Day.
NIST	National Institute of Standards and Technology, Boulder, CO.
NBS	National Bureau of Standards (now NIST).
NRC	National Research Council, Ottawa, Canada.
NSWC	Naval Surface Warfare Center.
OP	Observatoire de Paris.
SA	Selective Availability of GPS.
TAI	International Atomic Time.
USNO	U.S. Naval Observatory, Washington, DC.
UTC	Coordinated Universal Time.
UTC(i)	Coordinated Universal Time as realized by laboratory i.
VLBI	Very Long Base Interferometry.
WGS	World Geodetic System.

Manuscript received November 10, 1990; revised March 1, 1991.
The authors are with the Bureau International des Poids et Mesures, 92312 Sèvres Cedex, France.
IEEE Log Number 9100237.

I. INTRODUCTION

The excellence of worldwide unification of time realized by the establishment of International Atomic Time (TAI) and Coordinated Universal Time (UTC), depends on the quality of the participating atomic clocks and the means of time comparison. In the pre-GPS era (until the beginning of the 1980's) the technology of atomic clocks was always ahead of that of time transfer. The uncertainties of the long-distance time comparisons, by LORAN-C, were some hundreds of nanoseconds, and large areas of the earth were not covered. The introduction of GPS has led to a major improvement of world-wide time metrology in precision, accuracy and coverage. With GPS, time comparisons are performed with an accuracy of a few nanoseconds for short baselines (up to 1000 km) and 10–20 ns for intercontinental distances. This makes it possible to compare the best standards in the world at their full level of performance: for integration times of only 10 days, the frequency differences between atomic clocks are measured at the level one part in 10^{14}.

But this is by no means the limit to the possibilities of GPS. The inaccuracy of long-distance time comparisons performed with *C/A-code* (see Section II-A-2) receivers can be reduced to the level of a few nanoseconds by the reduction of some remaining systematic errors. Some of the actions required for this improvement will, in any case, be necessary when voluntary degradation known as *Selective Availability (SA)* (see Section V) is put into operation. On the other hand, geodesists, using a new generation of receivers, expect to measure pseudoranges with uncertainties of 10 cm or less and then to reduce ephemeride errors to less than 10 cm. These developments bring the hope that time comparisons may be achieved with uncertainties of 300 ps or less. Such an accuracy exceeds the present needs but could meet the challenge of the new generation of time and frequency technology.

In this paper, after a brief introduction to GPS, the principles of GPS time transfer are described with emphasis being placed on the so-called *common view* method. The sources of error during GPS time transfer are discussed and

0018–9219/91/0700–0991$01.00 © 1991 IEEE

the various possibilities of reducing them are investigated. Some possibilities for overcoming *SA* are also discussed. GPS is additionally shown to be an outstanding tool for the dissemination of UTC.

II. Background

A. GPS Outline

The NAVSTAR Global Positioning System, GPS, is a military satellite navigation system based on satellite ranging using on-board atomic clocks (cesium or rubidium). In its final stage it will provide position, velocity and time instantaneously and continuously anywhere on or above the earth. The system consists of three major segments: a space segment (satellites transmitting radio signals), a control segment (a control station and several monitor stations) and a user segment which extracts the desired information (position, velocity, time) from the received satellite signals.

1) The Satellites: The first GPS spacecraft was launched in February 1978. By the end of 1990, 15 satellites were operating. Six satellites belong to the *old Block I* and nine to the *new Block II* which is subject to the voluntary degradation known as *Selective Availability* (see Section V). When fully operational in 1993 the system will consist of 21 satellites equally spaced in 6 orbital planes inclined at 55° to the equator. There will also be 3 active spares. The spacecrafts orbit the earth at about 20 183 km above its surface and have sidereal period. This has a practical impact on the organization of GPS time comparisons (see Section III).

2) Transmitted Signals: The space vehicle clock frequency is 10.23 MHz, from which all other frequencies are derived. Each satellite transmits two navigational signals, namely an $L1$ signal at a frequency of 1575.42 MHz (10.23 MHz × 154) and an $L2$ signal at a frequency of 1227.6 MHz (10.23 MHz × 120) [1, 2]. Signal $L1$ is modulated with the *P-code* (*Precise* or *Protected code*) and the *C/A-code* (*Clear/Acquisition code* also known as "*civilian code*"); signal $L2$ is modulated with *P-code* only; both codes are of pseudorandom noise (PRN) type. The *C/A-code* with a chip-rate of 1.023 MHz and a repetition rate of one millisecond has a fast acquisition time, but it does not provide the high accuracy of the *P-code* which has a chip-rate of 10.23 MHz and a repetition rate of 267 days.

3) Navigation Message: Signals $L1$ and $L2$ are continuously modulated with a navigational data stream at 50 b/s in frames of 1500 b lasting 30 s and formatted into five sub-frames of 6 s each. A complete data message requires the transmission of 25 frames and takes 12.5 min. The navigation message generated by the control segment and uploaded to the satellites every day contains all the data required by a user for the computation of position. The message stream includes information on the status of the space vehicle, the parameters for the clock correction needed to compute *GPS time* (see Section II-B), the ephemerides of the satellite and the parameters for computing signal delay due to the ionosphere. The relationship between *GPS time* and UTC(USNO) (see Sections II-B and IV) is also contained in the navigation message.

Future GPS signals will be voluntarily degraded (*SA*) in order to deny unauthorized users access to the full accuracy of instantaneous positioning (see Section V). Authorized users will be able to perform instantaneous positioning with uncertainties of several meters while unauthorized users are limited to uncertainties of 100 m.

To determine a position in three space dimensions and time, a user must receive the signal from four satellites. However, if a user has a known position it is necessary only to track one satellite in order to determine time. In this way, GPS can be used as a time distribution system.

B. GPS Time

GPS time is a time scale established by the Control Segment and related to UTC(USNO). *GPS time* is a continuous time (not corrected by leap seconds) measured in weeks from the *GPS time* zero point defined as 0 h UTC(USNO), January 5, 1980. At the end of 1990, *GPS time* was ahead of UTC(USNO) by 6 s. The Control Segment should control *GPS time* to be within 1 μs of UTC(USNO) (modulo 1 s).

The navigation message contains a correction to be added to the readings of the on-board satellite clocks in order to obtain *GPS time* (see Section II-C). This realization of *GPS time* through the GPS constellation shows a peak-to-peak discrepancy, at a given instant, of up to 30 ns.

The navigation message also contains data giving access to UTC(USNO) in real time with an accuracy of 100 ns.

C. GPS Time Receivers

Most GPS time receivers are one channel *C/A-code* devices because such devices are simple and of reasonable cost. In addition *C/A-code* is accessible to everyone while *P-code* only to authorized users.

The time difference between the user clock connected with GPS time receiver and the *GPS time* is given by the following relationship:

$$DT = pr - (r/c + dS + dI + dT) - dD + dc \quad (1)$$

where

- pr measured "pseudorange" or an apparent propagation time, which is determined by measuring the difference of two identical codes, one generated by the receiver, the other one contained in the signal emitted by the satellite and received by the receiver; each of these codes is synchronized by its own clock,
- r range between the satellite and the station computed from broadcast ephemerides and the known coordinates of the receiver antenna,
- c speed of light,
- dS correction due to the rotation of the Earth (*Sagnac effect*) [3],
- dI propagation delay due to the ionosphere; since the *C/A* code is transmitted on only one frequency ($L1$)

there is no possibility of determining this delay by measurement so one has to rely on a model,

dT propagation delay by the troposphere (model),

dD receiver delay,

dc difference between the on-board clock and *GPS time* determined by a polynomial whose coefficients are contained in the navigation message; these coefficients do not include correction for relativistic effects which is computed by receiver software.

A typical GPS time receiver performs the above process during a 13-min track as follows.

a) The receiver processes short-term raw pseudorange measurements, smoothing them over a period of seconds (typically 6 or 15) through use of a second degree fit or phase accumulation (depending of manufacturer).

b) These short-term smoothed pseudoranges are corrected by the geometrical range (r) and the corrections detailed in (1).

c) A linear fit of the short-term data is used to deduce the time difference between the satellite and laboratory clocks over the 13-min track in terms of a slope, an intercept and a standard deviation.

GPS time receivers are usually equipped with a positioning capability called the *navigation solution* which gives antenna coordinates in the GPS reference frame (WGS 84 in 1990) with an uncertainty of about 5 m. This is enough for many applications but not for high accuracy time transfer (see Section III). The receiver delays are usually calibrated by the manufacturing company using simulated signals. It is generally assumed that the receiver software is equivalent, but for 1 ns accuracy that might not be true.

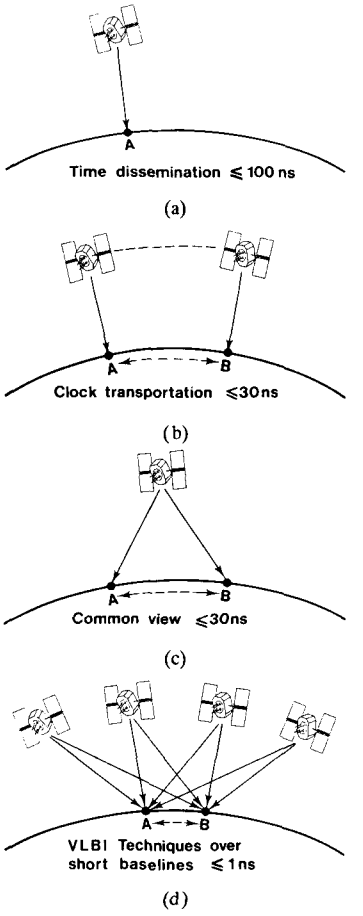


Fig. 1. Four methods of GPS time transfer and their approximate accuracies. (a) Time dissemination $\leq$ 100 ns. (b) Clock transportation $\leq$ 30 ns. (c) Common view $\leq$ 30 ns. (d) VLBI techniques over short baselines $\leq$ 1 ns.

D. GPS Time Transfers

There are four interesting ways of exploiting GPS for time and frequency comparisons (see Fig. 1) [4].

1) Time Dissemination: Station A can deduce *GPS time* from measurements according to (1). This method is the simplest and least accurate, but has a global coverage and requires no other data than those provided by the receiver. The accuracy of access to *GPS time* depends on local conditions of observation, mainly on the quality of the receiver antenna coordinates and on the amount of the acquired data. With antenna coordinates having an uncertainty of 10 m this accuracy ranges from 100 ns, for one 13-min track, to a few tens of nanoseconds for averaging times of one day or more. The receivers used for this method could be competitive on a mass production basis, and could service an unlimited audience in the dissemination of time. As *GPS time* closely follows UTC(USNO), GPS appears to be an outstanding tool for the dissemination of UTC (see Section IV).

2) Clock Transportation: Clock transportation is of interest for high accuracy time comparisons (10 ns or less). Clocks A and B, at different locations anywhere on the earth can be compared by making successive observations of the same satellite, or of the group of satellites, with delayed view times of less than 12 h. This gives the advantage of observing satellites at their maximum of elevation and so decreasing the sensitivity to models of refraction and to the satellite ephemerides (on-track component of the ephemeride error is usually the largest one). This method is however subject to the satellite clock instabilities, which are now small (about 5 ns over 12 h), but will be voluntary degraded when *SA* is implemented.

3) GPS Common-View Method [4]: Stations A and B or more stations receive the signals of the same satellite at the same time and communicate the data to each other to compare their clocks. The main advantage of this method is that satellite clock error contributes nothing (*GPS time* disappears in the difference) so it is of utmost interest during implementation of *SA*. The ultimate accuracy of common-view mode is about 1 ns. The error budget of the common-view time transfer method is discussed in detail in Section IV.

4) VLBI Techniques Over Short Baselines: A method of differential positioning by GPS satellites developed by the Jet Propulsion Laboratory [5] has an uncertainty at

the centimetric level over baselines of the order 100 km. This method, using the techniques of very long base interferometry (VLBI), can be applied for time comparisons with subnanosecond accuracy.

III. Accurate Time Comparisons

There are two interesting modes for accurate time transfer by GPS: *common view* and *clock transportation* (see Section II-D). The accuracy of the clock transportation mode will be limited during the implementation of *SA* (see Section V): for this reason we do not discuss it here.

The common-view method was suggested in 1980 by the NBS [4] and since 1983 has been used by an increasing number of national timing centers for accurate time comparisons of atomic clocks. The nature of their orbits is such that satellites are observed every sidereal day at nearly the same location on the sky, so *scheduled common views* are repeated every *23 h 56 m.* The common-view schedule, established and distributed by the BIPM to the national laboratories is kept without change for about 6 months, then a new schedule is issued. The method requires data exchange between laboratories, which is realized by electronic mail. The data are processed by the BIPM for the computation of international time links directly involved in the establishment of TAI and UTC.

One problem with the use of GPS for time transfer is that it is a one-way system. In addition, most laboratories use only the $L1$ frequency. This affects the propagation delay in many ways. This is so even if the common-view method, in some cases, diminishes ephemeride and ionospheric errors. Here we evaluate these perturbations and describe possible solutions.

A. Uncertainties of GPS Time Comparisons

The typical error budget of Table 1 is given for distances of 1000 and 5000 km, and for the usual tracking durations of 13 min, when applying the common-view mode [6]. Two cases are considered: a single common view and a daily average of 10 common views. This budget is established for normal operating conditions. Much larger errors may occur in the case of a defective receiver, lack of delay calibration, poor environment of the antenna, adoption of wrong antenna coordinates, etc.

B. Possible Improvements of GPS Time Comparisons

1) Antenna Coordinates: It has been found that inaccurate antenna coordinates (reaching sometimes several tens of meters) are the cause of large errors in GPS time transfer.

There are two different requirements on the coordinates of antennas.

a) Coordinates should be known accurately in a global terrestrial reference frame. The accuracy should be for example, of the order of 30 cm for 1-ns accuracy in time comparisons.
b) Satellite positions should be expressed in the same global frame as the antenna coordinates.

Table 1 Typical error budget of GPS time comparisons in common view (CV), at distance d, *C/A-code.* (Unit: 1 ns)

	For a single CV		For 10 CV, average over 1 day[1]	
$d =$	1000 km	5000 km	1000 km	5000 km
Satellite clock error (cancels in CV mode)	0	0	0	0
Antenna coordinates[2]	20	20	7	7
Satellite coordinates	2	8	1	3
Ionosphere (day time, normal solar activity, elevation > 30°)	6	15	1	3
Troposphere (elevation >30°)	2	2	0.7	0.7
Instrumental delay (relative)	2	2	2	2
Receiver software	2	2	2	2
Multipath propagation	5	5	2	2
Receiver noise (13-min average)	3	3	1	1
Total	22	27	8	10

[1]The noise of the laboratory clocks and the rise time of reference pulses bring nonnegligible contributions, which are not considered here.
[2]Assuming uncertainties of the order of 3 m. In practice, errors of coordinates can sometimes reach 30–40 m.

Concerning a), the CCDS has recommended [7] the use of the IERS Terrestrial Reference Frame (ITRF). In practice this frame is established by the IERS publication of the coordinates of about 100 primary sites, with uncertainties of a few centimeters [8], [9].

To ensure b), the ITRF coordinates should be converted in the WGS 84 reference frame used by GPS. But WGS 84 and ITRF frames do not differ more than 1.5 m and the realization of WGS 84 is of the order of 1 m (WGS 84 is based on the Doppler positioning and does not take into account the motions of continental plates), so it is acceptable to keep the coordinates of antennas expressed in ITRF and at the same time ensure b).

It is usually possible to obtain antenna coordinates, by relative positioning between the antenna and the nearest IERS site, with uncertainties of a few centimeters, using geodetic methods. On the other hand, the BIPM has developed a method of differential positioning between GPS antennas using the data of the time comparisons themselves [10]. The consistency of the coordinates by this method is within 30 cm for distances up to 1000 km. Using GPS geodetic differential positioning [11] and the BIPM technique, over the last few years, all national time laboratories equipped with GPS have been linked to IERS sites (four of them, in U.S., Japan, Germany, and Austria, with uncertainty below 10 cm). On June 12, 1990 at 0 h UTC, as suggested by the BIPM, these corrections were introduced into the time receivers ensuring worldwide homogenization of the coordinates in the ITRF. This is not the end of the efforts of the time community to improve the antenna coordinates. Further work is required at those laboratories having antenna coordinates expressed with uncertainties of less than 10 cm.

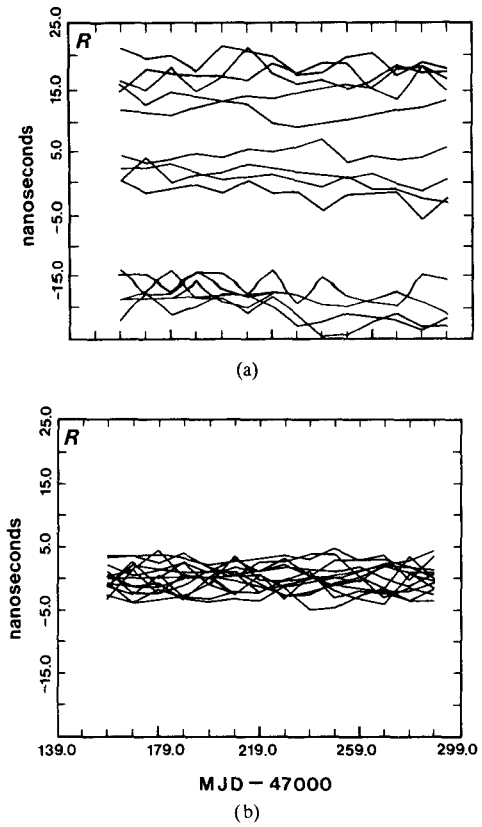


Fig. 2. Residuals (R) with respect to the mean of UTC(NRC)-UTC(USNO), as given by individual tracks. (a) Before coordinate corrections, (b) With differential coordinate corrections derived by BIPM from time comparisons ($dx = 9.47$ m, $dy = -2.96$ m, $dz = 4.05$ m).

Figure 2 illustrates the impact of the error in differential coordinates on the time transfer between two North American laboratories.

2) Satellite Ephemerides: The GPS precise ephemerides and clocks were computed at the NSWC from the beginning of 1986 to July 29, 1989. Since then they have been produced by the DMA. These ephemerides are received on a regular basis at the BIPM. Their estimated accuracy is of the order of 3 m.

Comparisons with broadcast ephemerides, for 1988, show a significant improvement in some cases (for satellites of Block I equipped with rubidium clocks, during the eclipse season, see Fig. 3) [12], [13]. But we can expect future improvements of precise ephemerides to satisfy the needs of geodesists. A global tracking system in support of *missions to the planet Earth,* is being established [14], aiming at a fully reliable, continuously operating, service. GPS satellite tracking is now one of the fundamental techniques of the IERS, together with VLBI and satellite/Moon laser ranging. In these activities, the satellite position errors should drop below 10 cm contributing less than 300 ps in time comparison errors for the longest terrestrial distances.

The main disadvantage of using techniques based on the precise ephemerides provided by external bodies is the delay of access which, at present, is about 3 months. This

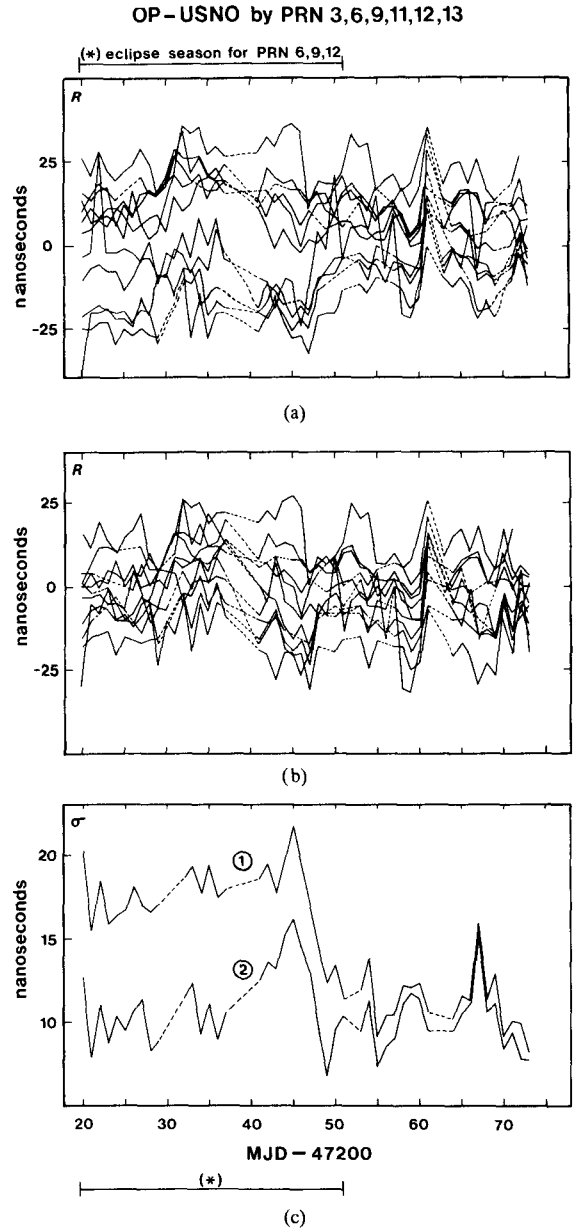


Fig. 3. Residuals (R) with respect to the mean UTC(OP)—UTC(USNO) as given by individual tracks (March–April 1988). (a) With broadcast ephemerides. (b) With precise ephemerides. (c) 1) standard deviations of residuals of (a); 2) standard deviations of residuals of (b).

delay prevents the operational use of precise ephemerides.

In practice, computations with precise ephemerides require knowledge of the broadcast ephemerides used by the receiver software in order to apply differential corrections. A difficulty is the possible change of ephemeride parameters during the usual 13-min tracking period (especially with *SA*) which could lead to the exchange of an enormous amount of short term smoothed pseudorange data. A modification of receiver software is needed so that a single set of parameters is retained for the full duration of the 13-min trackings. The SERCEL company has already solved this

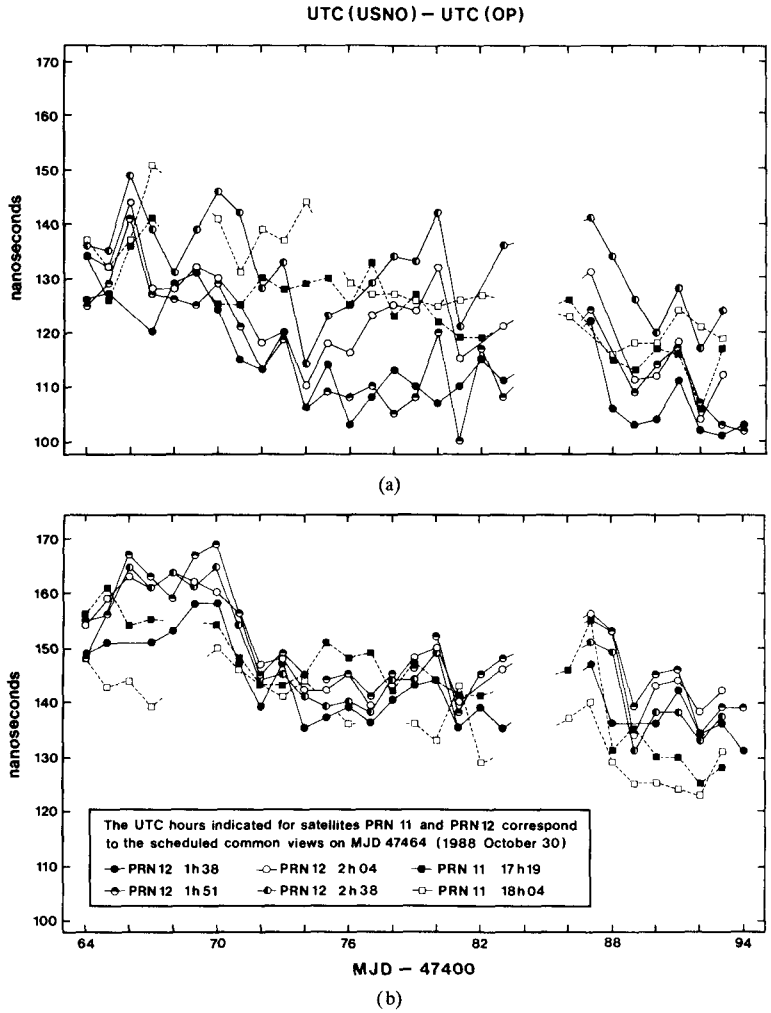


Fig. 4. Time comparison between Paris Observatory (Paris, France) and United States Naval Observatory (Washington, DC) for a 30-day period (November 1988). The two branches of the satellite simultaneous trackings are corrected using (a) ionospheric values derived from a model, (b) ionospheric values obtained from measurements.

question for all its commercialized units and the NIST has done the same on an experimental basis. The BIPM started the regular collection of GPS broadcast ephemerides in May 1990 and the NIST (Boulder, CO) in October 1990.

3) Ionospheric Refraction: Dual frequency receivers, which do not need a knowledge of the *P-code,* have recently been developed. They give measurements of the ionospheric delay along the line of sight of satellites with uncertainties of about 1 ns. A first prototype was built at the BIPM by Michito Imae, a Japanese guest visitor from the CRL [15] and operates routinely at the BIPM. After improvement at the CRL, an industrial version of this instrument has been made available commercially. Another type of *ionospheric measurement system* is being developed at the NIST [16] and prototypes are already operating at the NIST and the OP. This new system will be included in the usual GPS time receivers. The improvement of time comparisons using ionospheric measurements rather than model values is evident over long distances (see Fig. 4) [17]–[19]. A study of the closure around the world, obtained by combination of time transfers OP-NIST, NIST-CRL, and CRL-OP, has shown an improvement of accuracy through the use of ionospheric measurements [20]. At the BIPM, measured ionospheric delays have been used on a regular basis since November 1989 for the link OP-CRL.

4) Tropospheric Refraction: At radio frequencies the troposphere is a nondispersive medium and its effect on pseudoranges and time comparisons cannot be estimated from dual frequency measurements as is done for the ionosphere [21]. Instead, models are used for the estimation of the tropospheric delay. It has been assumed that for the needs of GPS time transfer at the level of 1–2 ns and for observations performed at elevation angles above 30°, a simple global model is sufficient. However, in the practice of common view time transfer over long distances (9000 km), elevations of 20° are sometimes unavoidable. We have also observed that different types of receivers use different tropospheric models. For example, a comparison

of two receivers has shown differences of 1.0 ns at 60° elevation, 1.8 ns at 30°, and 3.2 ns at 20°. To obtain an accuracy of a few hundreds of picoseconds in GPS time transfer, more sophisticated models of the troposphere with the inclusion of local meteorological measurements will be necessary.

5) Instrumental Delays, Receiver Software: Several experiments on the relative calibration of receiver delays have been performed by moving a GPS receiver, used as transfer standard [22]–[24], between sites. The resolution is of the order of 1 ns for 1–2 days of simultaneous tracking. However, only few of these receivers have been checked. Some received a single visit and very few received two or more visits. Our experience concerning the long-term stability of receiver delays is limited and drifts or steps of several tens of nanoseconds could occur without being noticed. Recently a sensitivity to the external temperature of a type of GPS time receiver was discovered [25].

In these relative calibrations, as well as in current GPS time comparisons, the receiver software, the adopted reference frames and the constants should be identical. Unfortunately, differences have been found, as noted in Section III-B-4) and shown in [26]. An important advance would be to agree on standard data processing. It should also be possible to extract from receivers sufficient information to check the conformity of formulae and constants with the adopted standards. A *Group of Experts on GPS Standards for Time Comparisons* is now being set up, its task will be to prepare standards which can be adopted by receiver designers and users.

6) Receiver Noise: With the averaging performed in time services, receiver noise does not seem to be a serious limitation when aiming at an accuracy of 1–2 ns, as long as there are no systematic effects such as temperature dependence. For better performance there exist new advanced GPS receivers which have demonstrated an accuracy of 10 cm on smoothed *P-code* pseudoranges over averages of 2 min and better than 5 cm on averages of 30 min [27].

7) Multipath Propagation: Multipath propagation arises from reflections at objects located around and under a GPS antenna. Resulting instantaneous errors can be as large as several tens of nanoseconds [28], [29]. Fortunately, these errors are partially averaged over 13-min tracks. Still, special care should be taken in the installation of a GPS antenna. An ideal installation would provide total isolation of the antenna from its environment. One approach to this ideal situation would be to install the antenna on the top of a high tower [30]. Another option is to locate the antenna directly on the ground in an open flat field with no obstacle within a radius of several tens of meters. In practice, good locations are difficult to find. In any event GPS antennas should be located in open areas and equipped with protection planes to eliminate reflection from below.

C. Improved GPS Time Comparisons

Table 2 gives a revised error budget for GPS time links, using receivers which are presently in operation, based on the following suppositions:

Table 2 Possible error budget, with optimum operation, in common view (CV), at distance d, *C/A-code*. (Unit: 1 ns)

	For a single CV		For 10 CV, average over 1 day	
$d =$	1000 km	5000 km	1000 km	5000 km
Satellite clock error (cancels in CV mode)	0.0	0.0	0.0	0.0
Antenna coordinates	0.6	0.6	0.2	0.2
Satellite coordinates	1.0	4.0	0.3	1.3
Ionosphere (measures)	1.0	1.0	0.3	0.3
Troposphere (elevation >30°)	2.0	2.0	0.7	0.7
Instrumental delay (relative)	1.0	1.0	1.0	1.0
Receiver software	0.0	0.0	0.0	0.0
Multipath propagation	1.0	1.0	0.3	0.3
Receiver noise (13-min average)	3.0	3.0	1.0	1.0
Total	4.1	5.7	1.7	2.1

error of antenna coordinates of 10 cm (the figures for a single CV correspond to the worst direction at both sites),
error of satellite coordinates of 5 m (in the worst direction for a single CV),
measured ionospheric delay with existing ionospheric measurement systems,
modeled tropospheric delay (as in Table 1),
measured relative instrumental delays by receiver transportation,
identical and correct receiver software,
good shielding of the antennas against multipath propagation,
receiver noise as in Table 1.

Among these suppositions, the compatibility of receiver software is not realized at present. However, the largest improvements arise from the adoption of accurate coordinates for the antennas and measurement of ionospheric delay. The estimates of errors in Table 2 are conservative, nevertheless we observe that the contribution of GPS to the uncertainties of time comparisons at 5000-km distance, on daily averages of 10 common views, can be of the order of 2 ns, using *C/A-code* only.

IV. Dissemination of UTC

As *GPS time* is referred to UTC(USNO) within 1 μs, and UTC(USNO) follows UTC also within 1 μs, in conformity with the recommendation of CCDS, it turns out that GPS system time does not differ more than a few microseconds (modulo 1 s) from UTC. This has some practical consequences [31].

A. Dissemination of UTC In Real Time

It appears that GPS is continuously disseminating UTC all over the world in real time with an uncertainty of a few microseconds. This requires only that the user receiver be calibrated and the user position be known with an

uncertainty of the order of a few hundreds of meters: such a coordinate determination is easily obtained from the receiver itself (see Section II-C).

There is no international commitment that *GPS time* will always remain close to UTC so it is appropriate to check in the publications quoted in Section IV-B that no major offset has occurred.

B. Delayed Dissemination of UTC

Access to UTC is possible with a greater accuracy in deferred time from the following publications.

1) BIPM Circular T: Each monthly issue of BIPM Circular T gives for 0 h UTC every day

$$Co = \text{UTC} - GPS\ time \text{ (modulo 1 s)}.$$

These values are derived from smoothed GPS data, taken at Paris Observatory, from Block I satellites only (no affected by *SA*). UTC may be computed from the observation at any site of any listed satellite and from the interpolation of Co.

2) USNO Series 4: Weekly bulletin USNO Series 4 gives for each satellite of Block I and Block II for every day, values of UTC(USNO)—*GPS time*. The access to UTC is possible via UTC–UTC(USNO) as given in BIPM Circular T.

In both cases the quality of the access to UTC will rely on local conditions of observation. The ultimate uncertainty ranges from 20 to 40 ns, but it requires an accurate positioning of the user (uncertainty better than 5 m) and an accurate calibration of the receiver delay.

V. GPS Time Transfers During Voluntary Degradation *(SA)*

Selective Availability (SA) is an intentional degradation of GPS signals and navigation messages designed to limit the full accuracy of the system to unauthorized users such as the international community of time metrology (most authorized users of GPS belong or are affiliated to the US Department of Defense).

The issue of *SA* is closely linked to the history of GPS development. It was elaborated in the 1970's and may be subject to change in the rapidly evolving international environment of the 1990's. Two events have given some idea about what *SA* looks like: first was an experiment on the degradation of the data of the 6 *Block I* satellites (September 29–October 2, 1989), which might have been a test of *SA*. This showed the following.

1) A phase jitter of the satellite clocks, the effect of which is removed by a strict common view.
2) A bias which changes frequently in the ephemerides of about 100 m, the effect of which, in common view, is roughly proportional to the distance [32].

The second event was the official implementation of *SA* on the *Block II* satellites, which began on March 25, 1990 (see Fig. 5) and was removed on August 10, 1990. Surprisingly the studies of the data of *Block II* satellites have shown only the degradation of the clock readings and had no effect on the ephemerides [32]. This means that the effect of *SA* can be entirely removed by a strict common view as illustrated in Fig. 6.

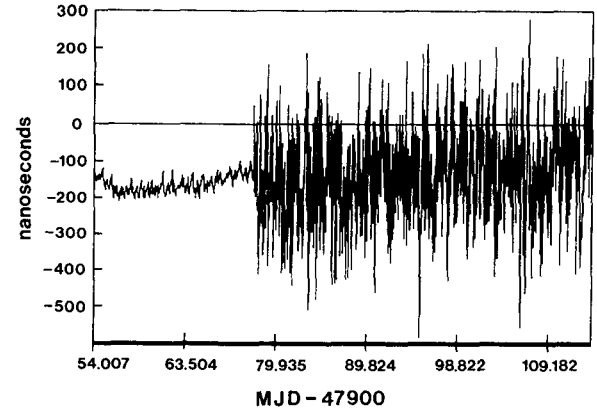


Fig. 5. UTC(USNO)-GPS time by Block II. Implementation of *SA* on Block II satellites on March 25, 1990.

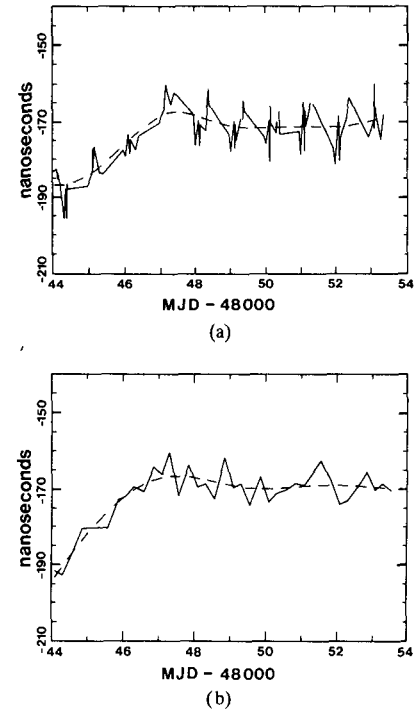


Fig. 6. Values of UTC(OP)-UTC(USNO) as obtained in common view mode by satellites of (a) Block I (*SA off*) and (b) Block II (*SA on*).

The dissemination of time is greatly affected by the *SA*. Access to *GPS time* by a single observation is limited to several hundreds of nanoseconds. However with numerous observations and averaging time of several days or more, one can reduce the noise of *SA* to several tens of nanoseconds [32]. This gives the possibility of a delayed access to UTC with the corresponding uncertainty (see Section IV-B).

Although the *March–August 1990* implementation of *SA* did not include ephemerides degradation, this does not mean that such a degradation will not occur in the

future. To overcome the problem of the possible degradation of ephemerides, various approaches are being studied [13]. The corrections to broadcast ephemerides affected by *SA* can be derived, as noted previously in Section III-B-2), from precise ephemerides. But another possibility can also be considered: this is the availability of differential corrections representing the bias introduced for *SA*. Such corrections could be made available within a few days and would restore the quality of the broadcast ephemerides without *SA*.

VI. Conclusions

The use of GPS for time and frequency control has already demonstrated the outstanding potential of this system. The permanent operation, worldwide coverage, low equipment cost and fully automatic reception makes GPS the most effective system of time dissemination. With *SA off*, GPS is permanently and in real time disseminating *GPS time* with an uncertainty of several tens of nanoseconds, UTC(USNO) with a hundred nanoseconds and UTC with a few microseconds.

In the domain of time metrology the GPS technique is 10–100 times better than the LORAN-C ground wave technique. Its present performance of 10 ns uncertainty for long distance links is sufficient to compare the best clocks at their full level of performance for integration times of 10 days and longer. This is enough for most applications.

The implementation of *SA* would be a severe drawback in terms of direct access to time, but for delayed dissemination of time and for accurate time comparisons, the effects of *SA* can be partially or completely removed by appropriate techniques.

The improved use of the current time receivers can reduce the uncertainty of long-distance time transfer to 2 ns. The next breakthrough would be the use of accurate geodetic receivers and the future ultra-precise ephemerides leading to an uncertainty of 300 ps in time transfer. GPS thus provides a means of meeting the challenge of the new generation of time and frequency technology expected to develop over the next decade.

References

[1] J. J. Spilker, Jr., "GPS signal structure and performance characteristics," *Navigation,* vol. 25, no. 2, p. 121, Summer 1978.

[2] R. J. Miliken and C. J. Zoller, "Principle of operation of NAVSTAR and System Characteristics," *Navigation,* vol. 25, no. 2, p. 121, Summer 1978.

[3] N. Ashby and D. W. Allan, "Practical implications of relativity for a global coordinate time scale," *Radio Sci.,* vol. 14, no. 4, pp. 649–669, 1979.

[4] D. W. Allan and M. Weiss, "Accurate time and frequency transfer during common-view of a GPS satellite," in *Proc. 34th Ann. Symp. on Frequency Control,* pp. 334–346, May 1980.

[5] P. F. MacDoran, "Satellite emission radio interferometric earth surveying, SERIES—GPS geodetic system," *Bull. Géodésique,* 1979.

[6] B. Guinot, W. Lewandowski, and C. Thomas, "A review of recent advances in GPS time comparisons," in *Proc. 4th European Frequency and Time Forum,* pp. 307–312, 1990.

[7] Comité Consultatif pour la Définition de la Seconde, presented at the Rapport de la 11e session, Apr. 1989.

[8] C. Boucher and Z. Altamimi, "The initial IERS terrestrial reference frame," IERS Technical Note 1, June 1989.

[9] IERS, Annual report for 1988, *Observatoire de Paris,* June 1989.

[10] B. Guinot and W. Lewandowski, "Improvement of the GPS time comparisons by simultaneous relative positioning of the receiver antennas," *Bull. Géodésique,* vol. 63, pp. 371–386, 1989.

[11] W. Lewandowski, R. J. Douglas, W. J. Klepczynski, W. Strange, J. Suter, and M. A. Weiss, "Positioning of GPS antennas in timekeeping laboratories of North America," in *Proc. 43rd Symp. on Frequency Control,* pp. 218–224, May 1989.

[12] B. W. Remondi and B. Hofmann-Wellenhof, "GPS broadcast orbits versus precise orbits," in *Proc. Meeting of Int. Ass. of Geodesy,* Aug. 1989.

[13] W. Lewandowski and M. A. Weiss, "The use of precise ephemerides for GPS time transfer," in *Proc. 21st PTTI,* pp. 95–106, 1989.

[14] R. E. Keilhan and W. G. Melbourne, "GPS Global tracking systems in support of missions to the planet earth," *CSTG Bull.,* pp. 129–140, 1989.

[15] M. Imae, W. Lewandowski, C. Thomas, and C. Miki, "A dual frequency GPS receiver measuring ionospheric effects without code demodulation and its application to time comparisons," in *Proc. 20th PTTI,* pp. 77–86, 1988.

[16] D. Davis, M. A. Weiss, and M. Vidmar, "A codeless ionospheric calibrator for time transfer applications," in *Proc. 2nd Int. Meeting Inst. of Navigation,* Sept. 1989.

[17] M. Imae, M. Miranian, W. Lewandowski, and C. Thomas, "A dual frequency codeless GPS receiver measuring ionospheric effects and its application to time comparison between Europe and USA," in *Proc. 3rd European Frequency and Time Forum,* pp. 89–93, Mar. 1989.

[18] M. Imae, C. Miki, and C. Thomas, "Improvement of time comparisons results by using GPS dual frequency codeless receivers measuring ionospheric delay," in *Proc. 21st PTTI Meeting,* pp. 199–204, Nov. 1989.

[19] W. Lewandowski, G. Petit, C. Thomas, and M. Weiss, "The use of precise ephemerides, ionospheric data and corrected antenna coordinates in a long distance GPS time transfer," in *Proc. 22nd PTTI Meeting,* pp. 547–558, 1990.

[20] M. Weiss, T. Weissert, C. Thomas, M. Imae, and K. Davies, "The use of ionospheric data in GPS time transfer," in *Proc. 4th European Time and Frequency Forum,* pp. 327–333, 1990.

[21] A. H. Dodson and J. C. Hill, "The effects of atmospheric refraction on GPS measurements," in *Proc. Seminar on the GPS,* Nottingham, U.K., 12–14 April 1989.

[22] J. A. Buisson, O. J. Oaks, and M. J. Lister, "Remote calibration and time synchronization (R-CATS) between major European time observatories and the US naval observatory using GPS," in *Proc. 17th Ann. PTTI Meeting,* pp. 201–222, 1985.

[23] W. Lewandowski, M. A. Weiss, and D. Davis, "A calibration of GPS equipment at time and frequency standards laboratories in the USA and Europe," in *Proc. 18th PTTI Meeting,* pp. 265–279, 1986; also in *Metrologia,* vol. 24, pp. 181–186, 1987.

[24] M. A. Weiss and D. Davis, "A calibration of GPS equipment in Japan," in *Proc. 20th PTTI Meeting,* pp. 101–106, 1988.

[25] W. Lewandowski and R. Tourde, "Sensitivity to the external temperature of some GPS time receivers," in *Proc. 22nd PTTI Meeting,* pp. 307–316, 1990.

[26] D. Kirchner, H. Bessler, and S. Fassl, "Experience with two collocated C/A code GPS receivers of different type," in *Proc. 3rd European Time and Frequency Forum,* pp. 94–103, Mar. 1989.

[27] J. M. Srinivasan, C. Duncan, G. Purcell, L. Young, S. DiNardo, E. Hushbeck, T. Munson, T. Meehan, and T. Yunck, "Measurement of Aircraft Position, Velocity and Orientation Using Rogue GPS receivers, A Preliminary report," presented at the NASA CDP Fall Meeting, Munich, FRG, 1988.

[28] D. Wells *et al.,* "Guide to GPS positioning," Canadian GPS Associates, Fredericton, NB, Canada (see 9.11 and 9.12), Dec. 1986.

[29] G. J. Bishop and J. A. Klobuchar, "Multipath effects on the determination of absolute ionospheric time delay from GPS signals," *Radio Sci.,* vol. 20, pp. 388–396, 1985.

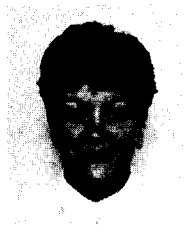
[30] G. Nard, J. Rabian, and R. Gounon, "Utilisation des signaux du GPS en mode différentiel instantané pour les applications temps-fréquence de haute précision," in *Proc. 1st European Time and Frequency Forum,* pp. 237–248, 1987.

[31] B. Guinot, "Time transfers by GPS," to be published in *Proc. GPS Workshop, European Centre for Geodynamics and Seismology,* 20–21 Nov. 1989, to be published.

[32] D. W. Allan, M. Granveaud, W. J. Klepczynski and W. Lewandowski, "GPS time transfer with implementation of selective availability," in *Proc. 22nd PTTI meeting,* pp. 145–156, 1990.

Wlodzimierz Lewandowski was born in Lower Silesia, Poland, in 1949. He received the masters degree in geodetic science from the Warsaw Polytechnic in 1976, and the doctorate degree in geographical science from the French Geographical Institute (IGN) in 1980.

He joined the staff of the Earth Rotation Section of the Bureau International de l'Heure (BIH) in Paris in 1982. In 1985 he moved to the newly created Time Section of the Bureau International des Poids et Mesures (BIPM) in Sèvres, following the transfer of International Atomic Time (TAI) from the BIH to the BIPM. Since then he has participated in the elaboration and diffusion of the TAI and the UTC international time scales. He has also engaged in research on the use of space techniques for time transfer, especially on the use of the GPS. In 1989 he spent six months with the Time and Frequency Division of the National Institute of Standards and Technology in Boulder, Colorado. Since 1988, he has participated in the work of the Civil GPS Service Steering Committee.



Claudine Thomas was born in 1955 in Lugny Champagne, France. She studied fundamental physics at the Ecole Normale Supérieure and the Pierre et Marie Curie University in Paris from 1975 to 1979. She received the Agrégation de Physique in 1979.

In October 1980, she took a position as Assistant at the Ecole Normale Supérieure where her teaching duties were mainly classical mechanics and electromagnetism. At that time she was involved in a research project on atomic physics at the Laboratoire Aimé Cotton in Orsay near Paris, where she completed her Thèse de Troisième Cycle in 1981. She received the doctorate degree in 1987, her dissertation described the behavior of Rydberg atoms in the presence of a strong magnetic field. She joined the Time Section of the Bureau International des Poids et Mesures as a Physicist in March 1987. In October 1990, she became Principal Physicist and Head of the Time Section where she is now responsible for the elaboration and diffusion of the international time references TAI and UTC. Her personal research is concerned mainly with time scale algorithms and time comparisons.

FORTY-FIFTH ANNUAL SYMPOSIUM ON FREQUENCY CONTROL

A FREQUENCY-DOMAIN VIEW OF TIME-DOMAIN CHARACTERIZATION OF CLOCKS AND TIME AND FREQUENCY DISTRIBUTION SYSTEMS

David W. Allan, Marc A. Weiss, and James L. Jespersen

Time & Frequency Division
National Institute of Standards & Technology
Boulder, Colorado 80303

Abstract

An IEEE standard (No. 1139-1988) now exists for "Standard Terminology for Fundamental Frequency and Time Metrology. As defined in this standard, the time-domain stability measure, $\sigma_y(\tau)$ has evolved into a useful means of characterizing a clock's frequency stability. There exists an ambiguity problem with $\sigma_y(\tau)$ for power - law spectral densities, $S_y(f)$, proportional to f^{α}, where $\alpha \geq +1$. For example, white noise phase modulation (PM) and flicker noise PM appear the same on a $\sigma_y(\tau)$ plot. Because of this ambiguity, Mod$\sigma_y(\tau)$ was developed.

More recently, it has become apparent there is no accepted measure for the performance of time and frequency distribution systems. At the current time, there is an important need for a good method for characterizing time and frequency transfer links in telecommunication networks.

Last year at this symposium suggestions were given for ways to characterize time and frequency distribution systems. Because of the above ambiguity problem, $\sigma_y(\tau)$ was shown to be a less useful measure than Mod$\sigma_y(\tau)$ for such systems. It was shown that $\sigma_x(\tau) = \tau \cdot \text{Mod}\sigma_y(\tau)/\sqrt{3}$ is a useful measure of time stability for distribution systems. For the case of white noise PM, $\sigma_x(\tau)$ is simply equal to the standard deviation for τ equal to the data spacing, τ_o and is equal to the standard deviation of the mean for τ equal to the data length (T).

In this paper, we recast these measures into the frequency-domain. We treat each of these measures as a digital filter and study their transfer functions. This type of measure is easily related to the passband characteristics of a given system, a particularly useful engineering approach.

Introduction

This paper concerns the characterization of frequency standards, clocks and associated systems. These associated systems may include: time and frequency measurement systems, time and frequency transmissions systems, time and frequency comparison systems, and telecommunication networks. As we shall see no single characterization is suitable. However, in this paper we discuss three statistical characterizations which cover most of the situations encountered in actual practice.

The characterizations that we require can be approached from two points of view: the time domain and the frequency domain. In this paper we describe in general terms the time and frequency domain approaches and then explain in some detail how the time domain approaches can be interpreted from the frequency domain point of view. We feel it is important to make this interpretation because of the significance of frequency domain approaches--particularly in engineering environments.

We will next explain why the mean and the standard deviation don't work for frequency standards. The simple mean, the standard deviation, and its square the classical variance, are well known statistical measures of a set of data points. It seems natural therefore that we should apply such measures to characterize various kinds of clock associated processes. However, application of these quickly reveals some significant problems.

Let's begin by considering the computation of the mean frequency output of a frequency standard. Normally, when we compute the mean of some process we suppose that including more data points in the computation brings us ever closer to the true mean of the process and that an infinite number of data points yields the true mean. Of course in the real world we must be content with a finite number of points always leaving some uncertainty in our attempt to find the mean. Nevertheless we assume that we can approach the true mean as nearly as we like if we are willing to collect enough data points.

Is this true of frequency standards? The answer surprisingly enough is "No!". How can this be? The answer is simple enough. Frequency standards do not generate a constant frequency output contaminated only by white noise. If they did we could average the output to get rid of the noise. That is, white noise is the kind of noise that can be averaged away. This is because for every phase advance it produces in the output of the frequency

standard, it eventually produces a compensating phase retardation, so that the two cancel in the averaging process.

Much research has shown that a number of noise processes, in addition to white noise, afflict frequency standards. we can qualitatively say that these other kinds of noise represent trends--not necessarily linear trends--in the output frequency of a standard. In later sections of this paper, we will quantitatively identify these "trends" but for now we will stay with the generic "trend" since this notion is enough to demonstrate why standard statistical methods don't work for frequency standards.

As an example, consider a frequency standard whose frequency output is contaminated with white noise and also increases linearly with time--a very simple kind of trend. What can we say about the mean output frequency of such a standard? Not much, because there is no average; we have defined the standard as producing a monitonically increasing frequency. The frequency we find by averaging is a function of when we start the measurement and the length of time over which the average is made.

What can we do when confronted with such a situation? An obvious answer is to remove the trend, by whatever means, and then compute the mean in the normal way from the modified data. In this way we would expect to converge on the true mean as we average away the white noise by using ever more data points.

The process we have just described is the essence of a number of approaches that have been developed to produce useful statistical measures for the output signals of frequency standards and related devices.[1] We can think of it as a two step process: First we remove the offending trends, and then we compute the statistical measures in the standard way. This process is not always evident when we look at the statistical measures in common use. More often than not, the two steps are combined into one obscuring the underlying process.

We will next describe the frequency- and time-domain measures of frequency and time stability. The standard deviation and the mean are examples of what are called time domain measures. That is, we collect a number of data points, one after the other, and then use these points to construct some useful statistical measure. There are also frequency domain measures. The power spectral density of a set of data points is an example. It provides us with a picture of the deviations in the data having a particular (Fourier frequency) spectral component.

As you might suspect, frequency- and time-domain measures are related. In later sections of this paper we explore in some detail the relationships between time- and frequency-domain measures. However, as a simple example consider the following. Suppose we want to remove a long term trend from the output signal of a frequency standard. There are a number of ways we might proceed. We might, for example, fit a polynomial to the data and subtract this polynomial from the data to generate a new set of data which we could then treat in standard statistical fashion. If the trend were strictly linear, then the polynomial curve would simply be a straight line whose slope revealed the magnitude of the frequency drift.

Another approach is to pass the data through a high pass digital filter to remove the low frequency components--which is what a trend looks like to such a filter.[2]

A particularly simple high pass filter can be constructed by taking what are called "first differences" of the data. That is we subtract the kth data point from the kth + 1 data point, for all data points. This process effectively removes long term trends from the data.

As we shall see later the "first differencing" process corresponds to a digital filter in the frequency domain whose characteristics can be defined precisely. We should also add that the polynomial fitting procedure also corresponds to a particular digital filter. However the emphasis in this paper is on "difference" type procedures since they are easy to implement and are commonly used by the time and frequency community.

We now describe three different variances. One is particularly useful for characterizing the frequency stability of clocks and oscillators. The next is most useful for characterizing the frequency stability of time and frequency measurement systems, distribution and comparison systems as well as for distinguishing between white noise PM and flicker noise PM. The last is most useful for characterizing the time stability of any of the above as well as for network synchronization; e.g. telecommunications network. All three of these variances are built upon taking finite differences of the data.

The notion of taking differences to remove trends in data is an old one. We quote von Neuman et. al. from a 1942 paper [3]:

"There are cases, however, where the standard deviation may be held constant, but the mean varies from one observation to the next. If no correction is made for such variation of the mean, and the standard deviation is computed from the data in the conventional way, then the estimated standard deviation will tend to be larger than the true population value. When the variation in the mean is gradual, so that a trend (which need not be linear) is shifting the mean of the population, a rather simple method of minimizing the effect of the trend on dispersion is to estimate the standard deviation from differences."

Perhaps the most important part of this quote is the parenthetical "which need not be linear." As it turns out taking first differences and second differences--repeating the differencing process twice--is sufficient to remove most of the kinds of noise that one encounters in clocks. Much of the work in the last few decades in the statistical characterization of frequency standards has been directed toward understanding in detail the implications of using the differencing approach. Today this approach is the mainstay of time domain approaches to characterizing time related processes. In this paper we will focus on three such characterizations with their associated interpretations in the frequency domain. We briefly introduce them here with details following in later sections.

The first statistical measure we want to introduce, also historically the oldest, is called the "two-sample variance," the "pair variance" or the "Allan Variance" [1, 4-6] It is denoted $\sigma_y^2(\tau)$ and referred to herein as AVAR. It is defined as follows:

$$\sigma_y^2(\tau) = \frac{1}{2} <(\Delta y)^2>, \qquad (1)$$

where the brackets "< >" denote expectation value, Δ is the first finite difference operator and y is the relative frequency offset as defined below.

AVAR was developed to address the problem of finding a suitable measure of the variability of the output frequency of a frequency standard. As we know, the computation of the standard variance will not work when applied to frequency standards because they contain noise processes which cannot be averaged out. The core idea of the Allan variance has already been introduced--the differencing method. Here the data to be differenced consists of a number of samples of the frequency of the standard taken over some period of time. The differencing procedure filters the noise processes that make the normal variance computation unsuitable. The actual formula, discussed later, accomplishes both the filtering and the variance computation in the same step.

Before we leave AVAR in this introductory section we should point out one thing. The output of a frequency standard is actually a signal whose phase advances in time with respect to some reference. We use phase or time difference almost interchangeably. This is so because they are directly proportional: $x(t) = \phi(t)/2\pi\nu_o$, where $\phi(t)$ is the phase difference reading in radians between two standards. The dimensions of x(t) are time. In practice, the frequency is derived by measuring the time or phase difference x(t) of the signal between the standard in question and the reference at two different times say t and $t+\tau$ giving us phases x(t) and $x(t+\tau)$.

Let $\nu(t)$ be the output frequency of the standard in question, and let ν_o be the frequency of the reference. We will assume, without loss of generality, that ν_o is perfect. The average relative frequency offset, $y(t) = (\nu(t)-\nu_o)/\nu_o$ of the standard in question over the time interval t to $t+\tau$ is then

$$y(t) = \frac{(x(t+\tau) - x(t))}{\tau} \qquad (2)$$

If we think of AVAR from the point of view of phase measurements x(t) instead of frequency measurements y(t), then AVAR is constructed in terms of the second differences in phase but first differences in frequency since frequency by definition is obtained from first differences in phase. An alternative and very useful definition of AVAR is as follows:

$$\sigma_y^2(\tau) = \frac{1}{2\tau^2} <(\Delta^2 x)^2>, \qquad (3)$$

where "Δ^2" is the second finite-difference operator.

The second statistical measure we want to introduce is called the modified Allan variance or from here on "MVAR." [5, 7-9] It is defined as follows:

$$\sigma_y^2(\tau) = \frac{1}{2\tau^2} <(\Delta^2 \bar{x})^2>, \qquad (4)$$

where $\bar{x}$ denotes phase averages being used in the second difference. We note that equations (3) and (4) are identical except for the phase averages. The three sequential phase averages are each taken over an interval τ. As τ changes, this changes the bandwidth in the software in just the right way to remove the ambiguity problem in AVAR. In other words, MVAR can distinguish between white noise PM and flicker noise PM, whereas AVAR cannot. In a later section this distinction will be more evident when we compare MVAR and AVAR from the frequency domain point of view.

Both AVAR and MVAR are particularly suited to characterizing the frequency instabilities of frequency standards. However there are situations where the emphasis is not on frequency but on time measurements. This brings up the final measurement we want to introduce, TVAR, where the "T" emphasizes the fact that we are focusing on time rather than frequency measurements. It is defined as follows:

$$\sigma_x^2(\tau) = \frac{1}{6} <(\Delta^2 \bar{x})^2>. \qquad (5)$$

We see that $\sigma_x(\tau)$ is just $\tau \mathrm{Mod}\sigma_y(\tau)/\sqrt{3}$ and has many of the advantages of $\mathrm{Mod}\sigma_y(\tau)$, but is now a time stability measure.

How does the change in emphasis come about? [10-11] The notion of studying frequency instabilities has a local flavor to it in the sense that frequency is defined by a certain resonance frequency of the Cesium atom or quartz

resonator while epoch time is an arbitrary manmade concept requiring coordination over time and space. Thus if we want to compare the frequencies of two remotely located standards we need to introduce some communication link which allows us to compare the phase or time difference between our two clocks. By measuring the change in the time or phase difference between these two standards over time we can determine the frequency offset between the two clocks.

Furthermore we might also want to determine the actual time offset between the clocks which again leads us to making time or phase difference measurements. Both of these examples point up the need for some statistical characterization where the focus is on time or phase rather than frequency, hence TVAR. So there is no confusion, we should point out that the time or phase difference between two clocks is a measurement. Whereas, the finite-difference operators, as in the above variance definitions, operates on a time series of measurement data. The advantages and disadvantages of these three variances will be more apparent later. [11]

We shall also look at the TVAR transfer function from the frequency domain point view This view and some other considerations reveal why TVAR is a more suitable measure for time related measurements than AVAR and MVAR.

Transfer Function Approach to Variances

A variance can be viewed from either the time domain or the frequency domain. [12-14] We intend to look at variances from both perspectives to aid in understanding what certain variances measure. We begin by describing the relationship between the variances in the two domains.

A. Variances in the Time Domain: Convolution

If, in the time domain, we have a time series of observables, $x(t)$, we may define a particular variance with the help of a convolving function, $h(t)$, where $h(t)$ is the impulse response function. We use the convolution, g, of x and h:

$$g(\lambda) = \int x(t) \cdot h(t-\lambda)\, dt. \qquad (6)$$

The variance corresponding to h, of the time series $x(t)$ is the infinite time average of g squared,

$$\sigma^2 = \lim_{T\to\infty} \frac{1}{T} \cdot \int_{\lambda=-T/2}^{T/2} g(\lambda)^2 d\lambda. \qquad (7)$$

The convolving function h, here, is the important definition. It determines how the variance selects data in the time domain, which is then squared and averaged. For example, the convolving function, h, for the classical variance is the step pulse: [2]

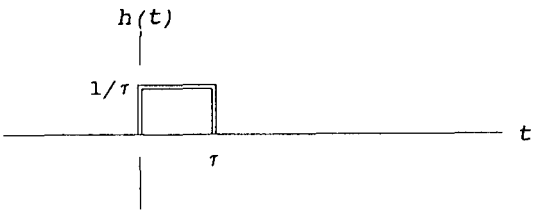


Figure 1. Impulse response function for classical variance. Each sample is taken over with an averaging time τ. Each sample is differenced with the mean, squared and averaged to obtain the classical variance.

This makes

$$g(t) = \overline{x(t)}_\tau , \qquad (8)$$

the average value of x from t to $t+\tau$. Thus the variance here is simply the second moment of $\overline{x}$.

For the Allan Variance (AVAR), h is the double pulse:

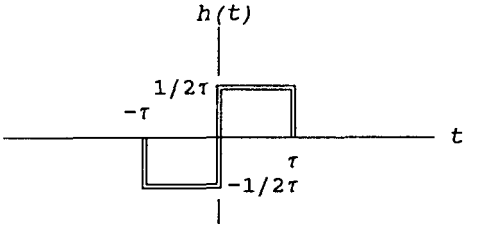


Figure 2. Impulse response function for the Allan or two-sample variance. Adjacent measurements--each averaged over an interval τ--are differenced. This change from one interval τ to the next is squared and averaged across the data, then divided by 2 for an AVAR estimate.

This makes the AVAR the expected value of the first difference squared. The normal use of AVAR is to characterize frequency stability. Thus, if $x(t)$ is a time series of clock time differences, the first difference of these divided by τ, $y(t)$, is the corresponding time series of frequency differences, averaged over the interval τ. Then convolution for the usual AVAR is

$$g(t) = \int y(\lambda) \cdot h_\tau(\lambda - t)\, d\lambda, \qquad (9)$$

giving a first difference of frequencies averaged over a time interval τ, or a second difference of time values. The integral of the square of this g results in the Allan, or two-sample variance

$$\sigma_y^2(\tau) = \lim_{T\to\infty} \frac{1}{T} \cdot \int_{\lambda=-T/2}^{T/2} g(\lambda)^2 d\lambda = \frac{\langle (\overline{y}_t - \overline{y}_{t-\tau})^2 \rangle}{2}. \qquad (10)$$

For the Modified Allan Variance (MVAR), we first note that in the Allan variance the time interval for averaging frequency, τ, is a multiple of the basic sampling

interval τ_0. Thus, it is possible to make n shifts of the pulses in figure 2 by τ_0, where $\tau=n\cdot\tau_0$. The modified Allan variance averages several first differences of frequency in this way, thus adjusting the software bandwidth to exploit the bandwidth dependence of white phase noise. As an example we show the convolving function for MVAR as the sum of two functions, as in figure 2, displaced by $\tau/2$.

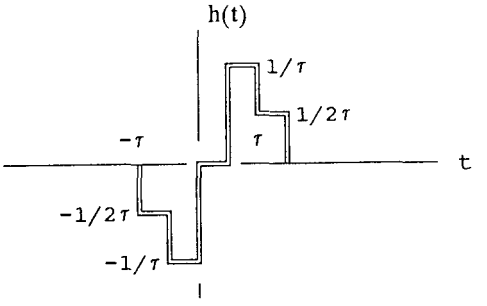


Figure 3. Impulse response function for the modified Allan or two-sample variance. This results from forming a second finite-difference from three contiguous intervals. Each interval contains the phase or time averaged over an interval τ. These second differences are squared and averaged across the data, then divided by $2\tau^2$ for an MVAR estimate.

The new variance, $\sigma_x^2(\tau)$ (TVAR), is simply $\tau^2\cdot\mathrm{Mod}\sigma_y^2(\tau)/3$. Thus, the convolving function has the same shape as in figure 3, but the vertical scaling needs to be multiplied by $\tau^2/3$.

B. Variances in the Frequency Domain: Transfer Function

We now examine the impulse response functions of these variances as transformed into the frequency domain. We will see that the convolving function h again is the important definition. In this domain, the Fourier transform, H of h, becomes a kind of transfer function for defining the variance.

There are two steps to understanding the passage from the functions we've discussed in the time domain to the frequency domain. First we use the fact that an infinite time average of a function squared equals the integral of the spectrum of that function:

$$\lim_{T\to\infty}\int_{\lambda=-T/2}^{T/2} g(\lambda)^2 d\lambda = \int_0^\infty S_g(f)\,df \qquad (11)$$

Since the variance is an infinite time average of the square of the function g, it also equals the area under the power spectral density of g, the square of the Fourier transform of g.

Second, we use the mathematical relation that the Fourier transform of a convolution is the product of the Fourier transforms. Let us put these two facts together. For the general variance σ^2, defined as the integral of the square of the time series $x(t)$ convolved with $h(t)$, we have

$$\sigma^2 = \int_0^\infty S_x(f)\cdot|H(f)|^2\,df, \qquad (12)$$

where $S_x(f)$ is the spectrum of x, and $H(f)$ is the Fourier transform of h.

This $H(f)$ is the transfer function of the variance. This differs from the usual use of the term "transfer function" in that instead of producing a signal sculpted by the shape of H, we produce a variance which is sensitive to frequencies according to the shape of H. [2]

This last equation, then, gives the relationship between the definitions of variance in the time and frequency domains. We see that the convolving function h that defines how the variance selects data in the time domain, also, via its Fourier transform, defines which frequencies the variance is sensitive to.

C. Transfer Functions of AVAR, MVAR, and TVAR

The transfer function for AVAR is shown in figure 4, a linear plot, for two different values of n, where $\tau=n\cdot\tau_0$. We see that the variance selects a band of frequencies for a given τ, and that the width of this band decreases as τ increases. Also note that the function goes to 0 at the origin. Indeed it goes to 0 fast enough that it remains integrable when multiplied by an f^α spectrum with α greater than -3. [6]

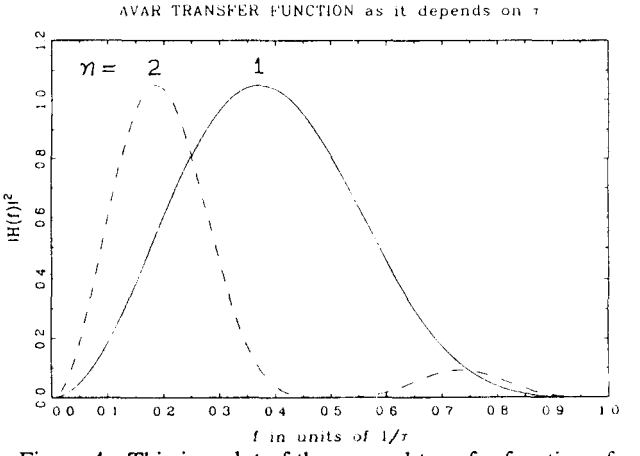


Figure 4. This is a plot of the squared transfer function of the impulse response function shown in figure 2. It is the function that multiplies the spectrum of the frequency deviations to obtain $\sigma_y^2(\tau)$. It is plotted hear for two values of τ (τ_0 and $2\tau_0$ where $1/\tau_0 = 1$). Note, the abscissa is linear and that the bandwidth of H(f) decreases as τ increases.

In figure 5, we see the transfer functions plotted on a logarithmic horizontal axis for n taking on the first eight powers of 2. We see that logarithmically, the bandwidth remains constant, and that the power-of-two transfer functions scan more-or-less independent frequency bands. Figure 6 shows the sum of these transfer functions. We see here that this sum yields a flat band-pass filter. The interpretation here is that this band pass represents the sum of information presented in $\sigma_y(\tau)$ versus τ plot. That is, the Allan variance plot of points chosen with n equal to a range of powers-of-two shows the stability of the data due to a certain band of frequencies. The sensitivity of AVAR to the different frequencies in this band is nearly constant. The band extends from $1/(2n\tau_o)$ to $1/(2\tau_o)$, where n is here the highest power of 2 chosen for the $\sigma_y(\tau)$ plot.

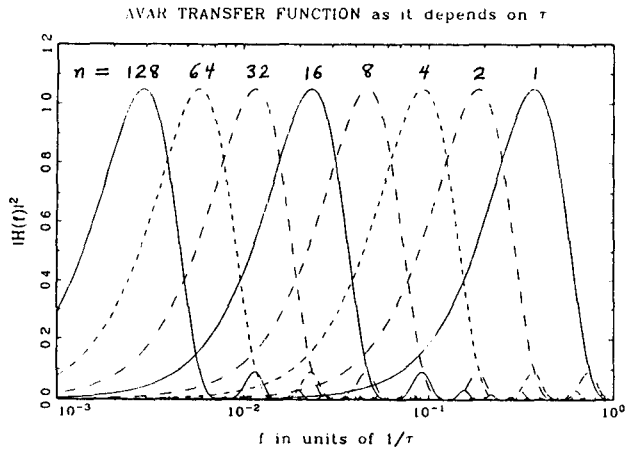


Figure 5. This is a plot of squared transfer functions of the impulse response function shown in figure 2 for 8 values of τ (1, 2, 4, 8....128 x τ_o). Note, the abscissa is logarithmic, and the apparent width of each transfer function is the same. They also appear distributed uniformly across a certain span of Fourier frequencies.

Next we look at the transfer function for MVAR. Analogous to AVAR, we see in figure 7 the MVAR transfer function in a linear plot for two different τ values. Here also, we see that the bandwidth decreases as τ increases, but in addition, the amplitude decreases also. This comes from the additional software filter in MVAR, the phase averaging, allowing MVAR to distinguish white phase noise from flicker phase noise. In figure 8 the MVAR transfer functions for powers-of-two τ values are summed as in figure 6; again, we see a flat band-pass. Thus, an MVAR power-of-two plot also presents the stability information due to a range of frequencies, with nominally equal sensitivity to frequencies within the range. For the same range of τ values, the MVAR cumulative transfer function is a little wider than that of AVAR and the high frequency end is slightly steeper.

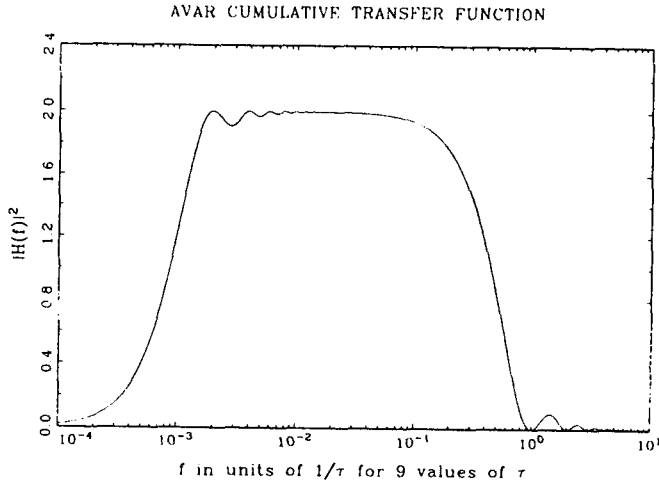


Figure 6. This figure shows the sum of the squared transfer functions for 9 values of τ (1, 2, 4, 8....256 x τ_o) for $\sigma_y^2(\tau)$. We conclude that a $\sigma_y(\tau)$ plot for such a set of τ values gives a nearly constant response to Fourier frequency over about two decades.

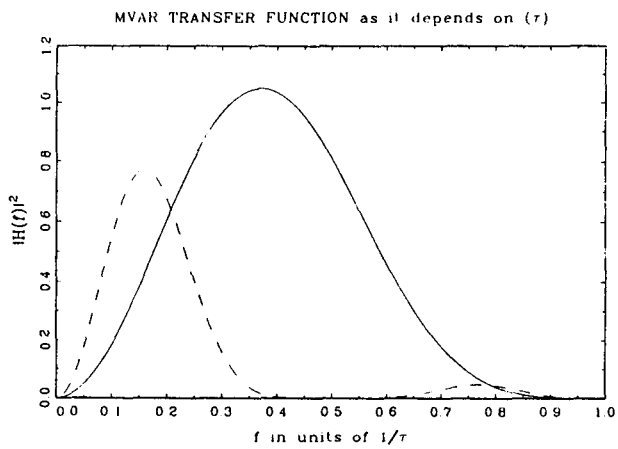


Figure 7. This is a plot of the squared transfer function of the impulse response function shown in figure 3. It is the function that multiplies the spectrum of the frequency deviations to obtain Mod$\sigma_y^2(\tau)$. It is plotted hear for two values of τ (τ_o and $2\tau_o$ where $1/\tau_o = 1$). Note, the abscissa is linear and that the bandwidth of H(f) decreases as τ increases. Notice also, that the amplitude decreases with increasing τ. This is due to the software band-width change brought about by phase averaging.

In figure 9, we see the TVAR transfer function. Note here that the function "rings" forever. That is, neighboring sinusoidal lobes do not die out. Of course with finite data sampling, there is always a high-frequency

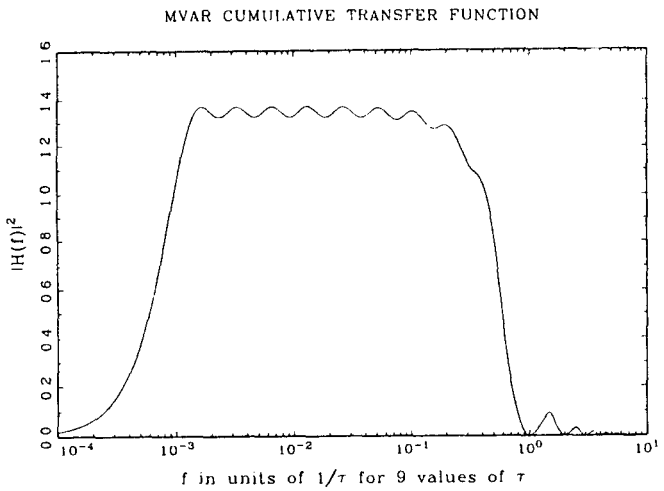


Figure 8. This figure shows the sum of the squared transfer functions for 9 values of τ (1, 2, 4, 8....256 x τ_o) for Mod$\sigma_y^2(\tau)$. We conclude that a Mod$\sigma_y(\tau)$ plot for such a set of τ values gives a nearly constant response to Fourier frequency over slightly more than two decades.

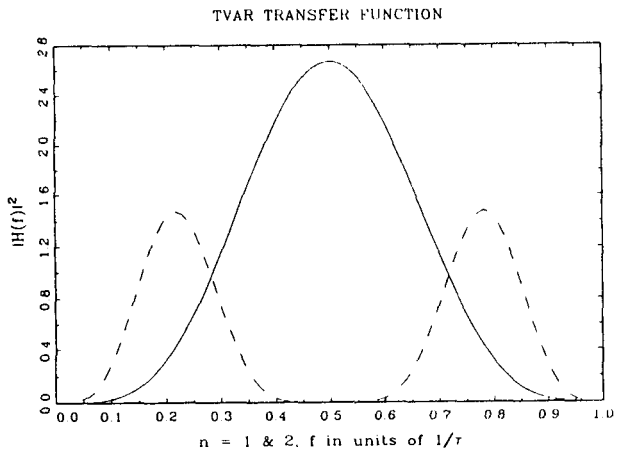


Figure 9. This is a plot of the squared transfer function of the impulse response function for TVAR. It is the function that multiplies the spectrum of the time deviations to obtain $\sigma_x^2(\tau)$. It is plotted hear for two values of τ (τ_o and $2\tau_o$ where $1/\tau_o = 1$). Note, the abscissa is linear and that the bandwidth and the amplitude of H(f) decrease as τ increases--similar to MVAR. As with MVAR, this is due to the software band-width change brought about by phase averaging. This transfer function, for a given τ, has repeat lobes into the higher Fourier frequencies indefinitely.

cut-off given by the Nyquist frequency $1/(2\tau_o)$. The sum of powers-of-two transfer functions, figure 10, shows a fairly flat band-pass up to the high-frequency end where there is greater frequency sensitivity--peaked at the Nyquist frequency, f_{Nyq}. Frequencies higher than the Nyquist frequency can be aliased into a TVAR computation up to the cut-off frequency, f_h. This points out the value of the general rule to have the sampling period equal to or less than $1/(2f_{Nyq})$. Then aliasing will not be a problem.

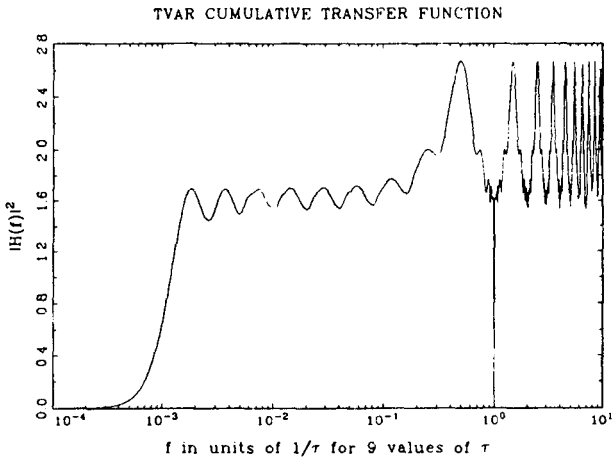


Figure 10. This figure shows the sum of the squared transfer functions for 9 values of τ (1, 2, 4, 8....256 x τ_o) for TVAR. We conclude that a $\sigma_x^2(\tau)$ plot for such a set of τ values gives a nearly constant response to Fourier frequency over slightly two decades but with increased sensitivity to Fourier frequencies at the Nyquist frequency, f_{Nyq}, and at 1/2 f_{Nyq}. Frequencies higher than $2f_{Nyq}$ can be aliased into the computation of TVAR. Frequencies will be aliased up to the measurement system cut-off frequency f_h. Maximum aliasing occurs at 3/2, 5/2, 7/2...x f_{Nyq}. A null occurs at twice the Nyquist frequency. Sensitivities at the non-aliased frequencies above $2f_{Nyq}$ are about the same as they are below the Nyquist frequency.

If the data sampling rate, τ_o, is greater than $1/(2f_{Nyq})$, then a TVAR plot will include Fourier energy from $1/(2n\tau_o)$ up to f_h with about equal sensitivity, except at the aliased values (3/2, 5/2, 7/2... X f_{Nyq} up to f_h). If f_h is equal to f_{Nyq}, then the TVAR cumulative transfer function looks very much like the MVAR cumulative transfer function.

Applications and Discussion

In the previous sections we have learned of three time domain statistical measures that are particularly appropriate for applications involving frequency standards, clocks and their associated measurement, comparison and distribution systems. We have also seen in some detail how these three time domain measures can be interpreted in the frequency domain. In this section we consider where each time domain measure is most appropriately applied. As we shall see, the selection of the appropriate time domain measure is a function of the types of noise which are characteristic of the process we are investigating as well as whether we want to study time stability or the

frequency stability. We want to choose that measure which most clearly reveals the types and levels of noise involved in a particular application.

As we stated in the introduction, AVAR and MVAR were developed first, while TVAR is the newest member of our triad of statistical measures. In general terms, AVAR and MVAR are the measures to use when we are primarily interested in systems and devices where frequency is the quantity of interest, while TVAR is more appropriate where the quantity of interest is primarily time or phase.

Before we begin our discussion of specific applications, let's briefly review the five kinds of noise processes we are likely to encounter for the systems discussed in this paper. Although there are many ways to inventory these noise processes it is common in the time and frequency literature to list them as follows:

1. white noise PM (phase modulation)
2. flicker noise PM
3. white noise FM (frequency modulation)
4. flicker noise FM
5. random walk FM

Mathematically these noise processes have the power-law spectral density relationships shown in the Table 1. Table 2 shows the appropriate mathematical expression for each of the three time domain measures. Table 3 gives the coefficients needed to translate from the time domain to the frequency domain.

Figure 11 displays illustrative examples of the time variation, x(t), of these noise processes. As we proceed from type 1 through type 5 noise we notice that the amplitude variation with time grows increasingly more slowly. Generally speaking, for our applications, the physical explanation for this trend is as follows. The time variations with a f^{-3} and f^{-4} spectrum, for example, are often related to environmental factors such as temperature variations, mechanical shock, and path delay variations while the faster variations represented by f^{0} and f^{-1} processes are more likely related to internal characteristics of the device itself. Here, for example, we think of the noisy electronic components that make up the amplifying stages in a frequency standard.

As we have learned in previous sections AVAR, MVAR and TVAR have different characteristics in the frequency domain so it is not surprising that one time domain measure is better suited for one kind of noise process than another.

AVAR and MVAR are frequency-stability measures, and AVAR is particularly suited for measuring the intermediate to long-term stability of clocks and oscillators. MVAR is generally more suited for electronically

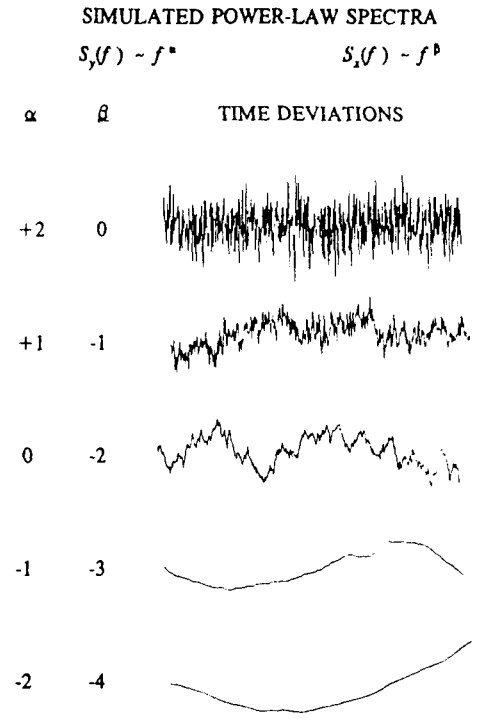


Figure 11. This is a display of the commonly occurring five power-law spectral density processes. These are often used as models for the time and/or frequency deviations in precision frequency sources.

generated noise processes and short-term frequency stability measurements. TVAR is a time-stability measure and is also suited for electronically generated noise processes. Both MVAR and TVAR are also sensitive to low frequency components often determined by environmental factors. TVAR is particularly suited for measuring the stability of time dissemination, comparison or measurement systems. It is also well suited as a measure of synchronization stability in telecommunication's networks.

We can see this from a different perspective by considering how AVAR, MVAR and TVAR vary with τ for our five dominant noise processes. Figures 12, 13 and 14 display the τ dependance for our three time domain measures. If we look at figure 12, we see that AVAR does not discriminate between white PM and flicker PM. This, as we said earlier, was one of the primary reasons for introducing MVAR, which as figure 13 shows, does discriminate between white PM and flicker PM. If we look at figure 14, we see that TVAR displays unambiguously the five noise types, as does MVAR, but that it also more clearly reveals the presence of white PM and flicker PM than does MVAR. This is, of course, the reason that TVAR was introduced since it "focuses" on the noise processes that are of most interest when we are making phase or time measurements.

Table 1. The power-law spectral density relationships for the five kinds of noise processes we are likely to encounter for the systems discussed in this paper.

NOISE TYPE	α	β	μ	μ'	η
White PM	2	0	-2	-3	-1
Flicker PM	1	-1	-2	-2	0
White FM	0	-2	-1	-1	1
Flicker FM	-1	-3	0	0	2
Random Walk FM	-2	-4	1	1	3

Where: $\sigma_y^2(\tau) = a_\mu \tau^\mu$ $\quad S_y(f) = h_\alpha f^\alpha$

$\text{Mod}\ \sigma_y^2(\tau) = b_{\mu'} \tau^{\mu'}$ $\quad S_y(f) = h_\alpha f^\alpha$

$\sigma_x^2(\tau) = c_\eta \tau^\eta$ $\quad S_x(f) = h_\beta f^\beta$

$\sigma_x^2(\tau) = \frac{\tau^2}{3} \text{Mod}\ \sigma_y^2(\tau)$ $\quad S_y(f) = (2\pi f)^2 S_x(f)$

Table 2. The appropriate mathematical expression for each of the three time domain measures.

ABBREVIATION	NAME	EXPRESSION
AVAR	ALLAN VARIANCE	$\sigma_y^2(\tau) = \frac{1}{2}\langle(\Delta y)^2\rangle = \frac{1}{2\tau^2}\langle(\Delta^2 x)^2\rangle$
MVAR	MODIFIED ALLAN VARIANCE	$Mod\sigma_y^2(\tau) = \frac{1}{2\tau^2}\langle(\Delta^2 \bar{x})^2\rangle$
TVAR	TIME VARIANCE	$\sigma_x^2(\tau) = \frac{1}{6}\langle(\Delta^2 \bar{x})^2\rangle$

Table 3. The coefficients needed to translate from the time domain to the frequency domain.

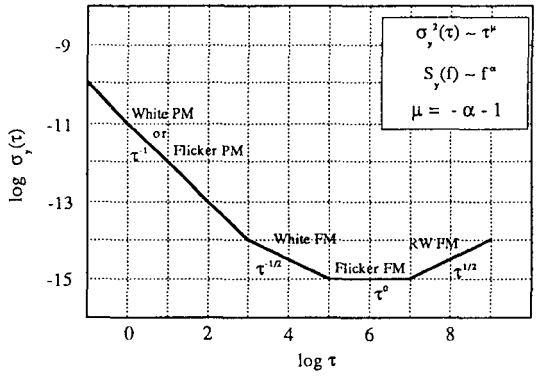
NOISE TYPE	$S_y(f)$	$S_x(f)$
White PM	$\frac{(2\pi)^2}{3fh^2}\ [\tau^2\sigma_0^2(\tau)]\ f^2$	$\frac{1}{\tau_0 f_h}\ [\tau\ \sigma_x^2(\tau)]\ f^0$
Flicker PM	$\frac{(2\pi)^2}{A}\ [\tau^2\sigma_y^2(\tau)]\ f^1$	$\frac{3}{3.37}\ [\tau^0\sigma_x^2(\tau)]\ f^{-1}$
White FM	$2\ [\tau^1\sigma_y^2(\tau)]\ f^0$	$\frac{12}{(2\pi)^2}\ [\tau^{-1}\sigma_x^2(\tau)]\ f^{-2}$
Flicker FM	$\frac{1}{2\ell n2}\ [\tau^0\sigma_y^2(\tau)]\ f^{-1}$	$\frac{20}{(2\pi)^2 9\ell n2}\ [\tau^{-2}\sigma_x^2(\tau)]\ f^{-3}$
Random Walk FM	$\frac{6}{(2\pi)^2}\ [\tau^{-1}\sigma_y^2(\tau)]\ f^{-2}$	$\frac{240}{(2\pi)411}\ [\tau^{-3}\sigma_x^2(\tau)]\ f^{-4}$

$A = 1.038 + 3\ell n(2\pi f_h \tau)$

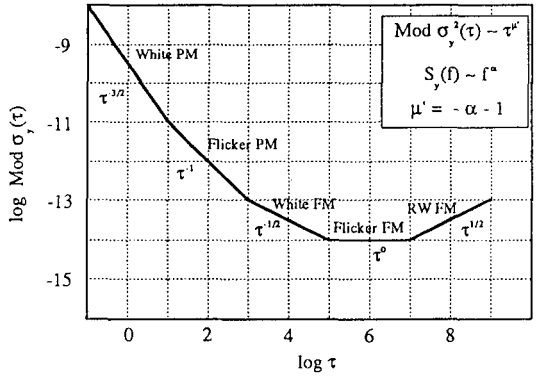
Table 4 shows in some detail the kinds of noise processes that are associated with various areas of application. The table also shows which of the three time domain measures is most appropriate for each particular application. We show the five different types of power-law-spectra and the ranges of applicability. These ranges include those for precision oscillators, for time distribution, network and comparison systems. From this table it is easy to see why MVAR and TVAR are better measures than AVAR for time distribution, network and comparison systems. On the other hand, AVAR estimator of frequency changes for white noise FM processes. This is important for commercial rubidium, cesium and for passive hydrogen masers. AVAR is also simpler to compute and is typically more intuitive than the other two variances. It nicely covers the range of applicability for precision oscillators except for the ambiguity problem in differentiating between white PM and flicker PM. This is only a problem for short-term stability in the case of active hydrogen masers and quartz crystal oscillators. In addition, $k\tau\sigma_y(\tau)$ is unbiased and useful measure of time error of prediction over the interval τ. The constant k depends on the power-law noise type, but is nominally equal to 1. [15]

The five power-law processes for the most part provide adequate modeling for time and frequency metrology. The higher values of alpha typically are used as models for the short-term stability of clocks and oscillators. The lower ends of the ranges are often appropriate models for the long-term stability of clocks and oscillators as well as for the time distribution, comparison, network and measurement systems. These lower values of alpha are often contaminated with diurnal and annual variations in these systems causing them to appear low-frequency dispersive. Some time comparison systems, such as GPS used in the common-view mode, are well modeled by white-noise PM in the day-to-day deviations. The bottom end of the variance ranges are those points where these variances are no longer convergent. If models were needed with lower values of alpha than those shown, then variances with higher order differences could be used, such as the Kolmogorov structure functions. [16] These three measures are convergent for the upper ranges of α and β ($\alpha > +2$ or $\beta > 0$), but only some of the Fourier transform relationships have not been worked out. [2] This is only because these models are not usual.

Sigma Tau Diagram



Modified Allan Variance
distinguishes White PM.



TVAR optimally estimates
time instability with White PM
and distinguishes other noise types.

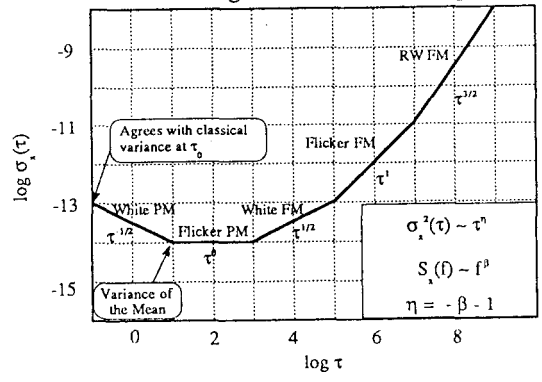
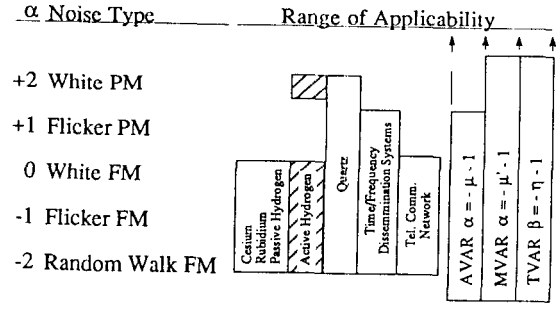


Figure 12, 13 and 14. These three figures are example plots for the square root of AVAR, MVAR, and TVAR, respectively. The sensitivity of these variances to the five power-law spectral density processes illustrated in figure 11 are depicted. Notice, that the slope on the $\sigma_y(\tau)$ plot is the same for white-noise PM as for flicker-noise PM (-1; ζ = -2). Notice also, that it is easier--visually--to distinguish between white phase, flicker phase and random walk phase on a $\sigma_x(\tau)$ plot than from a Mod$\sigma_y(\tau)$ plot. The slope changes are more dramatic to the eye. These three noise processes are particularly useful models for systems where time measurements are important.

Table 4. The kinds of noise processes that are associated with various areas of application. We also see which of the three time domain measures is most appropriate for each particular application.



Conclusion and Summary

Over the last few years, the need for a measure of time stability has become apparent. A search of the literature reveals that the classical measures (standard deviation, mean and variance) have lead to confusing and often misinterpreted conclusions. A measure, TVAR, is shown to have the attributes needed for characterizing the random processes in systems where time stability or phase stability is important. TVAR was compared with and contrasted to the other two previously developed time-domain statistical measures. The need for the three measures is also explained.

In particular, we have discussed how these three measures are appropriate for frequency standards, clocks and their associated measurement and distribution systems. We have shown how these measures can be used to determine the five noise processes that dominate most of the systems of interest in this paper. However, we have also examined these measures from a frequency-domain point of view and have shown how each measure corresponds to a particular transfer function. This procedure reveals the way time domain measures treat the various noise processes from a frequency domain point of view. It also provides a link, for those more accustomed to working in the frequency domain, to the time domain measures which are frequently employed by the time and frequency community.

Acknowledgements

The authors are particularly grateful to the readers, David A. Howe and James A. Barnes, of this paper. We are also very appreciative to the telecommunication industries for their support of the research associated with this paper. We are also grateful for the continued support and encouragement in this project by Donald B. Sullivan.

References

[1] J. Rutman and F. L. Walls, "Characterization of Frequency Stability in Precision Frequency Sources," submitted to Proc. of the IEEE, 1991.

[2] L. S. Cutler and C. L. Searle, "Some Aspects of the Theory and Measurement of Frequency Fluctuations in Frequency Strandards," Proceedings of the IEEE, Vol. 54, No. 2, pp. 136-154, Feb. 1966.

[3] J. von Neuman, R. H. Kent, H. R. Bellinson, and B. I. Hart, "The Mean Square Successive Difference," Ann. Math. Stat. 12, 153-162 (1942).

[4] D. W. Allan, "Statistics of Atomic Frequency Standards," Proc. of the IEEE, 54, 221-231 (1966).

[5] "Characterization of Clocks and Oscillators," Eds. D. B. Sullivan, D. W. Allan, D. A. Howe, and F. L. Walls, NIST Tech Note 1337, 1990.

[6] J.A. Barnes, A.R. Chi, L.S. Cutler, D.J. Healey, D.B. Leeson, T.E. McGunigal, J.A. Mullen, Jr., W.L. Smith, R.L. Sydnor, R.F.C. Vessot, and G.M.R. Winkler, Characterization of Frequency Stability, IEEE Transactions on Instrumentation and Measurement, IM-20, No. 2, 105-120, 1971.

[7] D.W. Allan and J.A. Barnes, A Modified "Allan Variance" with Increased Oscillator Characterization Ability, Proceedings of the 35th Annual Frequency Control Symposium,, 470-475, 1981.

[8] P. Lesage and T. Ayi, "Characterization of Frequency Stability: Analysis of the Modified Allan Variance and Properties of Its Estimate," IEEE Trans. on Instrum. & Meas., IM-33, 332-336 (1984).

[9] L. G. Bernier, "Theoretical Analysis of the Modified Allan Variance," Proc. of the 41st Annual Frequency Control Symposium, 1987, pp. 116-121.

[10] D. W. Allan, D. D. Davis, J. Levine, M. A. Weiss, N. Hironaka, and D. Okayama, "New Inexpensive Frequency Calibration Service from NIST," Proc. of the 44th Annual Frequency Control Symposium, 1990, pp. 107-116.

[11] D. W. Allan, "Time and Frequency Metrology: Current Status and Future Considerations," submitted to Proc. of 5th European Frequency and Time Forum, March 13-15, Bescanson, France, 1991.

[12] D. B. Percival, "Reappraisal of Frequency Domain Techniques for Assessing Stability Measures," Proc. of the PTTI, 1987, pp. 69-80.

[13] J. Rutman, "Characterization of Phase and Frequency instabilities in Precision Frequency Sources: Fifteen Years of Progress," Proc. of the IEEE, Vol. 66, No. 9, pp. 1048-1075, September 1978.

[14] J. Rutman, "Characterization of Frequency Stability: A Transfer Function Approach and Its Application to Measurements via Filtering of Phase Noise," IEEE Trans. on I & M, Vol. IM-23, No. 1, March 1974, pp. 40-48.

[15] D.W. Allan, Time and Frequency (Time-Domain) Characterization, Estimation, and Prediction of Precision Clocks and Oscillators, IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, UFFC-34, 647-654, 1987.

[16] W. C. Lindsey and C. M. Chie, "Theory of Oscillator Instability Based Upon Structure Functions," IEEE 64, 1652-1666 (1976)..

Synchronization and Relativity

GERNOT M. R. WINKLER, FELLOW, IEEE

The use of concepts and formulae of the Special and of the General Theory of Relativity has become a necessity in electronic systems that use high precision timing, especially space systems. The principles involved will be reviewed with emphasis on practical applications and operational standards used in timekeeping and remote synchronization.

I. Introduction

In high precision synchronization over extended distances, certain corrections to the time measurements have to be applied in order to allow for metric effects that are mainly due to the relative speed and gravitational potential difference between the two locations to be linked. These effects are predicted by the Theory of Relativity [1]. The Special Theory of Relativity (SR) is concerned with the transformations that are needed to relate the measurements between two systems that move relative to each other at constant velocity. Velocity is to be understood as a vector, and no gravitational field is assumed to exist. In this case, we speak of two inertial systems. Any system that moves at constant velocity in respect to an inertial system is also an inertial system. It is obvious that the distance measurements will be affected by the relative motion, and it seems simple to account for this effect if we express the measurements of one system in the coordinates of the other. However, contrary to "common sense," and to the principles of classical physics, time must also be transformed; this has given rise to many misunderstandings. The well known notorious twin paradox (shown later) is based on such a misunderstanding.

The basis for the coordinate transformations in the SR is the principle that all laws of physics, including the propagation of light or electromagnetic signals, must be identical for two observers that are in uniform motion relative to each other. While the SR deals with the transformations between two inertial systems, the General Theory of Relativity (GR) goes further and includes all possible motions in the transformations, not just the uniform velocity between two inertia systems as in SR. It includes the effects of gravity. In fact, it has been accepted by many as the modern theory of gravity [2] that produces more accurate predictions than Newtonian gravity.

While the SR is somewhat counter-intuitive and abstract, the GR is even more abstract. It is mathematically sophisticated; it requires for its full appreciation a mastery of tensor analysis. This is why there are differences in details in the interpretation among the experts. In addition, there are several alternative metric theories of gravity, collectively known as Post-Newtonian theories of gravity. Will [3] has given a thorough discussion of these developments and particularly of the experiments that can, or could, decide which one is the most realistic description of nature. However, the practitioner, especially the engineer concerned with timing measurements, at the present state of the art, will hardly have to deal with these further developments within the group of relativity theories, even if the utmost accuracy is required in space electronics. In terrestrial applications, none of these different formulations differ appreciably from the classical, Einsteinian formulation of the GR. The effects that are important for the timing user have been experimentally verified in many different tests, in some cases down to 0.01% of the predicted effects [4].

II. The Principle of Relativity and Some of its Consequences

The principle of relativity recognizes that position and motion can be specified only in respect to some other material bodies, and particularly, that space has no absolute significance in regard to uniform motion. Classical mechanics, of course, uses the concept of inertial systems and the classical relativity principle recognizes that the mechanical laws must be identical in all inertial systems. However, an extension of this principle to electromagnetic phenomena, especially the propagation of light, created a dilemma because the experimental evidence (e.g., the Michelson-Morley experiment), the actual behavior of light as it propagates, contradicts our intuitive notions of a simple addition of the signal and system velocities.

For a while, a reconciliation of the dilemma was attempted by looking for a mechanical explanation of the propagation of light by means of the hypothetical ether, which required additional assumptions of an ether wind, physical contractions of bodies under this wind, etc. How-

Manuscript received November 1, 1990; revised January 16, 1991.

G. M. R. Winkler is with the U.S. Naval Observatory, Washington, DC 20392-5100.

IEEE Log Number 9100238.

ever, Einstein recognized in 1905 that a most elegant resolution of the dilemma was possible by including the propagation of light in the postulate, and thereby modifying our basic concepts of classical mechanics, because this step requires the transformation of the time measure between inertial systems. In the meantime, the postulate of perfect relativity (which includes the propagation of light) has been established as correct in countless experiments and it is now known as Einstein's Principle of Relativity, in contrast to the mechanical principle mentioned before. Again, the difference between the two is that Einstein's version includes all phenomena, particularly the propagation of light. The reason for the failure of our intuition is simply that the speed of light (c) is so great that the "relativity effects" do not show up until we deal with speeds that approach c and/or until our measurement precision is very high indeed.

The experience that the physical phenomena are not affected by a uniform motion through space, i.e., that space has no absolute meaning in respect to unaccelerated motion, has as one of the consequences that time also cannot have an absolute meaning. It is an abstract measure of the state of processes in nature, such as motion of bodies or changes in temperature, that we introduce on the basis of standard processes (clock readings) and we have to find out how to extend this measure consistently into distant regions of space. It is clear that this extension requires the use of some signaling and, therefore, the problem of the speed of signaling is inseparable from the problem of transformation.

We need transformation formulae for the coordinates if we want to refer measurements taken in one inertial system to the coordinates of another (primed) system that moves at velocity, v, in respect to the first. This transformation of the space and time coordinates of an event must depend on the velocity v as parameter:

$$(t, x_1, x_2, x_3) - (v) \rightarrow (t', x_1', x_2', x_3').$$

Furthermore, the transformation of the coordinates into the primed coordinates must be done so that each new coordinate is a linear function of the old (unprimed) coordinates. The linearity is needed so that we maintain the linearity in the description of a uniform, i.e., unaccelerated motion. In order to emphasize that time (t) must be part of these transformations, we will consider it as a fourth coordinate.

There are two conventions in use for the choice of the time index: zero or four. In either case, the combination of time and space coordinates leads to the understanding of the four components

$$x_\alpha (\alpha \text{ to go from 0 to 3, or from 1 to 4})$$

as the components of a four-dimensional coordinate vector, x. We call this vector a worldvector in space-time. It should also be noted that we must be prepared to distinguish between upper and lower indexes which designate contravariant and covariant components respectively, a distinction that will be necessary whenever we deal with a nonEuclidean geometry, such as in the Minkowski space and especially in the GR, which uses Riemannian geometry.

A general linear transformation between two vectors is accomplished by multiplication with a transformation matrix, L, (that in our case must be some function of v)

$$x' = Lx \tag{1}$$

where the two coordinate vectors are taken as column vectors. It is remarkable that Einstein's relativity principle, in conjunction with the required linearity, already determines the transformation matrix, L. We can see this in a plausible way (for details see [2] or [5]) if we express the spherical wave front of a light flash in the coordinates of two, initially coincident, inertial systems. This equation is (c is the speed of light in vacuum):

$$\Delta r = c \cdot \Delta t \text{ or in greater detail}$$

$$\text{for both systems :}$$

$$\sum \Delta x_i^2 - c^2 \Delta t^2 = 0$$

$$= \sum \Delta x_i'^2 - c^2 \Delta t'^2 = s^2 (i = 1 \text{ to } 3). \tag{2}$$

In other words, the transformation requires that the flash can be seen as propagating isotropically in both systems. This, of course, is exactly the point where our intuition tells us that this is not possible. However, we have to respect the fact that all experiments conducted so far to test this, or its consequences, have been in agreement with Einstein's relativity principle. An interesting analysis of the testing problem is given in [6]. If we accept Einstein's expression of the relativity principle as an established fact, and we must do so on the basis of all evidence, then the rest is just a necessary mathematical consequence.

The two conventions about time as a fourth (generalized) coordinate are due to different ways of looking at the mathematics. Minkowski pointed out that the simplest way to understand the transformation is to consider it as a rotation in four-dimensional space-time, a transformation that keeps the length, s, of the worldvector x constant. In order to reduce this to the rotation case, we take as the fourth coordinate

$$x_4 = j \cdot c \cdot t \quad \text{with } j = \sqrt{(-1)}. \tag{3}$$

The factor c means that we measure all coordinates in meters; it would be equally legitimate to measure everything in light time, i.e., in seconds, by applying c as a divisor to the space coordinates instead of using it as a factor of the time coordinate. At any rate, the transformation we seek must accomplish that s, as defined previously, remain constant or, equivalently, that s is invariant. The invariance of s under the transformation is a cornerstone of the SR. This is in striking contrast to the classical concepts, where length and time differences are invariant independently. Here, it is only the four-dimensional space-time interval s that is invariant. Using our coordinate convention, the invariance of s is expressed by

$$\sum \Delta x_i^2 = \sum \Delta x_i'^2 = s^2 = \text{constant} \tag{4}$$

with the summation index i going now from 1 to 4. The invariant interval s can be seen as the radius of a four-dimensional sphere, which is formally zero in all legitimate coordinate systems. However, this is true only for the intervals that connect two events which are at the ends of a light ray. We must also remember that this Minkowski space-time is a complex space and its geometry is, therefore, only pseudo-Euclidean. If $s = 0$ in this space, then it does not at all mean that the three-dimensional distance between the two events is zero in our inertial system. One of them could be here and the other on a distant galaxy, but they would be connected by a light ray.

While the mathematical "trick" (3) is most helpful for the purposes of the SR, it is not feasible to extend the Minkowski convention to the problems of general transformations with which the GR has to deal. For this reason, most workers today, including [2], use the other convention in which

$$x_0 = c \cdot t. \tag{5}$$

In this case, the transformation matrix will, of course, be slightly different. In any case, a transformation that includes time will necessarily treat it as a pseudo-geometric coordinate, which allows us to bring the whole power of the general geometrical formulations to bear upon the problems. But conceptually we must always remember that time and space are not equivalent, even though the mathematical union in which they find themselves under the transformation (1) is due to the fact that time and distance measurements cannot be made independently from each other.

We have seen that the interval s of two events connected by a light ray vanishes in every inertial system because of the relativity principle. However, it can be shown that every interval, s, must be a constant (must be an invariant). Let us consider the infinitesimal intervals between two events as expressed in two inertial systems. They must be proportional because they are of the same order:

$$ds^2 = a \cdot ds'^2.$$

The factor a cannot depend on the coordinates because this would violate the homogeneity of space and time. It can only depend on $|v|$. But if we consider more than two such systems, then the respective factors, each depending on the respective velocity differences, can be expressed as ratios of two other a's. While these ratios do not depend on the angles between these velocities, the velocity difference of our last pair would depend on the angle between their velocities, which would make the expression as ratios of the other a's inconsistent. Therefore, the a's must be a constant, and the constant can only be 1. This implies, therefore, that every s must be an invariant under coordinate transformations, a most important result.

For events that are separated by $s \neq 0$, two possibilities must be distinguished:

Case 1: The time component is greater than the space part—this is a time-like interval, and we form the difference:

$$\tau^2 = \Delta t^2 - (1/c^2) \cdot \sum \Delta x_i^2. \tag{6}$$

Case 2: The space component is greater than the time part—we have a space-like interval, and we form the other difference:

$$\sigma^2 = \sum \Delta x_i^2 - c^2 \Delta t^2. \tag{7}$$

We have used the time measure in the first case, and the space measure in the second, as it is advisable according to the needs of the application. Whatever we do, both the magnitude and the kind of interval are invariant and the possible transformations can only change the magnitude of one component at the expense of the other. Moreover, these transformations can only go so far that the smaller of the two components vanishes (as it was observed above, the two components are not equivalent), and not beyond that, i.e., they cannot make a time-like interval into a space-like and vice versa (this limit is due to v being limited to $v < c$). In the case of a space-like interval, it is possible to chose coordinates that make the events simultaneous $(\Delta t^2 = 0)$ at a spatial distance equal to σ. This distance is the shortest possible distance in any frame. A time-like interval allows a transformation to coordinates in which the events occur at the same location $(\Sigma \Delta x_i^2 = 0)$, with a "proper time" (see the following) difference $\tau = \Delta t^2$. This time difference is the briefest time measure in any system. In other words, simultaneity is relative to the system chosen, and there is no simultaneity possible if the interval between the events is time-like. By using the invariance of s, we can derive the connection of the time measure $\Delta t'$ in a moving vehicle (not necessarily assuming uniform motion) as it looks from an inertial reference Δt. The invariable interval between two events is always s; however, as we just saw, this s is composed of the two components, the time part and the space part. If we denote the system in which the clock is at rest as the "proper" system, then an inertial system used as reference is the "coordinate" system. It is customary to denote the proper time as τ. This is always the time that is kept directly by a clock (which is at rest in its own rest frame). The interval between two events at the location of the clock is then expressed in the two systems as

$$\begin{aligned} \Delta s^2 = {} & \Delta \tau^2 - 0 \\ & \text{(proper time)} \\ = {} & \Delta t^2 - (1/c^2) \sum \Delta x_i^2 (\text{sum over 1 to 3}) \\ & \text{(coordinate time).} \end{aligned} \tag{8}$$

The motion of the clock while it traverses the coordinate distance $\sqrt{(\Sigma \Delta x^2)}$ takes place during the coordinate time Δt at speed $v^2 = \Sigma \Delta x^2 / \Delta t^2$, which allows regrouping of our formula

$$\Delta \tau = \Delta t \cdot \sqrt{(1 - v^2/c^2)}. \tag{9}$$

This is the transformation of the increments of the readings

of a stationary clock (the proper time of this clock) in terms of coordinate time. It is the time part of the general coordinate transformation (the matrix L) that is known as the Lorentz transformation. The details of the Lorentz transformation are amply discussed in the references. The moving clock will always have smaller numerical readings (it will be late) as seen from the coordinate frame. On the other hand, a moving proper length l in the direction of the velocity of the "moving" (of course, it is a relative motion) system will always measure less in the coordinate system; that is the Lorentz contraction. Dimensions perpendicular to the motion will not be affected.

One of the experimental demonstrations of the invariance of s under coordinate transformation consists of the observation that a certain type of elementary particle, π-mesons, are so short lived $(10^{-8}\ s)$ that, after their production in the laboratory, they can move only a few meters before they decay. However, they can be observed on the surface of the Earth, even though their natural origin is high in the upper atmosphere. The explanation is found in the change in the space-time measure due to the extreme speed, very close to c. In the meson comoving system, the distance to the surface is Lorentz contracted by the high relative speed of the Earth system; the terrestrial observer, on the other hand, sees the mesons last longer because of the time dilation due to their high speed relative to us. But, the meson always travels the distance $l = v \cdot \Delta t$ as seen in either system; i.e., we can write this equation in unprimed and primed magnitudes: $l' = v \cdot \Delta t'$ is seen in one system, and $l = v \cdot \Delta t$ in the other. Commensurate with the slowing down of the time measure, the distance is also shortened, as measured in the other system. Only v is the same (with opposite signs) in both systems. Atomic frequency standards will, therefore, appear Doppler-shifted by the same amount as observed in the other system. The situation is exactly the same in both systems, in accordance with the principle of relativity.

An observation regarding the time dilation may be useful. Very often its effects are described in a way which suggests that we deal with physical effects upon the moving clock, that the moving clock is slowed down. This is misleading. Clocks always measure proper time. It is not a physical effect upon the clock if its time measures appear to slow down as its speed increases relative to us. It is merely a metric effect due to the transformation from one system to the other. The same clock will, at the same time, appear to operate at different rates if seen from different reference systems; therefore, the effect is not located in the clock but is purely a metric result due to the relative speed of the coordinate systems chosen.

It is expedient to abbreviate factors that appear repeatedly in the transformation as follows:

$$\beta = v/c \quad \text{and} \quad \alpha = 1/\sqrt{(1-\beta^2)}. \tag{10}$$

In the simple case of two systems with parallel axes, coinciding initially, but moving at relative velocity v along the x-axis, the complete transformation in the case of convention (3) becomes;

$$L = \begin{pmatrix} \alpha & 0 & 0 & j\alpha\beta \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -j\alpha\beta & 0 & 0 & \alpha \end{pmatrix}. \tag{11}$$

It can be seen, that in this special case, only the first and the fourth coordinates will be affected. As expected, L is a linear coordinate transformation. It is orthogonal and the inverse is obtained by exchanging v with $-v$.

In the x_0 to x_3 convention, the matrix must, of course, be different:

$$L = \begin{pmatrix} \alpha & -\alpha\beta & 0 & 0 \\ -\alpha\beta & \alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{12}$$

The more general transformations, with v not in the direction of the x-axis, can be derived through multiplication with rotation matrices, as explained in detail in [7]. However, it must be stressed that the appearance of similarity between the two forms is somewhat deceptive. Mathematically they are quite different and the Minkowski convention has been criticized for reasons given in [2], where the authors want to see it discontinued altogether. However, there are many applications where the Minkowski convention of using an imaginary fourth coordinate, even though useless in the GR, produces simpler formulations in the SR. This has been advocated in [7].

We must note a second important result: the velocity of light plays the role of a limiting maximum velocity beyond which our formulae would produce imaginary results. By the same token, velocities do not add algebraically. We assume a speed u' in the primed system in the same direction as v at which the primed system moves away from the unprimed. This speed will then be transformed as

$$u = (u' + v)/(1 + u'v/c^2). \tag{13}$$

This addition of velocities is also limited by the velocity of light. The limiting role of c is a direct consequence of the relativity principle, which includes the propagation of light as an invariant phenomenon. In the limiting case of traveling photon, in its proper frame, there is no time interval between emission and reception, as there is no distance covered, a remarkable result indeed, but only the logical equivalent of Einstein's relativity principle (because we used it as basis for our mathematical derivations).

The next step can easily lead to a fatal error, made by many people, even experts, the error of the clock paradox. We can integrate the increments of the time readings of the previously mentioned formula even for nonuniform motions of the proper clock, where v is a function of t, as long as we use an inertial reference:

$$\tau_1 - \tau_0 = \int_{t_0}^{t_1} \sqrt{(1 - v^2/c^2)} \cdot dt \tag{14}$$

But we are not allowed to reverse the procedure by using the system of the nonuniformly moving clock as reference

for an analogous computation of the coordinate clock as seen by the proper clock. Only inertial systems can be used as a basis for time–space coordinates. The reason is simple: the relativity principle is not valid for accelerated frames of reference. This is almost obvious: we certainly can feel the acceleration, as we also will find that other phenomena do not remain unaffected by the acceleration. The basis for our concepts derived so far is not valid under acceleration. This solid fact, the nonvalidity of the relativity principle under acceleration, invalidates a reversal of the previously mentioned conclusion that the stationary clock's reading would also be slow if seen from the moving system. Claims that "relativity" means that the relative situation is the same, are false. The situation is clearly not symmetric: the clock that moves nonuniformly is always slow compared to an inertial clock. Problems such as a reverse computation, which cannot be done in the SR because it includes the action of accelerations, have contributed to the development of the GR, which includes the treatment of accelerations due to gravity.

Additional general comments are possibly important. The events in the life of a physical particle are points in the four-dimensional space-time world, and these points form a worldline with differential increments, ds, which are time-like. As projected upon an external coordinate system, the increments have time and space components, but the interval itself is invariant. The invariance of s under transformation has the consequence that the splitting of space-time into space and time coordinates depends upon the reference system chosen. The transformation has the group property, i.e., two successive Lorentz transformations correspond to a one-step transformation of the same kind, and a reverse transformation is obtained by changing the sign of the relative velocity, v. However, the transformation group is not commutative, the result of two successive transformations depends on their order (general coordinate rotation is not commutative [7]). Even more important is that what is simultaneous in one system will not be simultaneous in a system that moves in respect to the first. In other words, simultaneity has no absolute significance, it is relative to the system used. A failure to remember this relativity of simultaneity can easily lead to errors in the reduction of signal times coming from satellites. These errors can occur if the reduction formulae use an Earth-centered, fixed reference in which assumed simultaneities of distant events would be different from what the receiver must use for its own data reduction.

The Lorentz transformation of the coordinates also has a most profound and fundamental importance for science: In physical theory, we cannot continue to allow actions at a distance because the position of the distant body that acts upon the phenomenon under investigation depends upon the reference used. Moreover, the action can not be instantaneous because of the limiting role of c. Only field theories can be allowed from now on because in this case, the action is caused by the local field environment. This is also connected with the fact that the Maxwell equations are intrinsically Lorentz-invariant.

III. The General Theory of Relativity

Einstein's principle of relativity is the fundament upon which the SR has been built. The GR enlarges upon this in two ways: the basic geometry is generalized to allow for arbitrary transformations, and second, it includes the principle of equivalence as the basis for the treatment of the effects of gravitation.

The basic formula for the computation of the four-dimensional infinitesimal interval, or line element, was

$$ds^2 = c^2\,dt^2 - \sum dt_i^2. \tag{15}$$

For a treatment of general transformations we must write this equation in a style that treats time and space parts uniformly. We will use affine coordinates with implied oblique axes and distinction between covariant and contravariant components and vectors and we use the real convention (5) for the time coordinate, as used also in [2] or [4]:

$$ds^2 = \sum g_{\alpha\beta} \cdot dx^\alpha\,dx^\beta \tag{16}$$

where $g_{\alpha\beta}$ is the metric tensor. The summation is to go over both, α and β, going from 0 to 3 (latin indexes will go usually from 1 to 3). Einstein introduced the convention to imply summation over indices that repeat in any term. With this convention one writes the equation without the capital sigma. We have included t in one of the four x coordinates as discussed above by using the real value convention, and it is obvious that until now, most of the coefficients g are zero because in a flat space with Cartesian coordinates we have no mixed terms, only sums of squares of the coordinates. Therefore, the metric tensor for the SR is

$$g_{\alpha\beta} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}. \tag{17}$$

It is clear that if we use transformed coordinates, we could also expect to end up with changed coefficients, g, of this quadratic form because the interval is invariant. However, this is not so. The Lorentz transformation, because it is a linear orthogonal transformation in four-space, keeps the metric tensor (the g matrix) invariant. This is just a sophisticated way of saying that the geometry in the space is not affected and remains flat, i.e., pseudo-Euclidean, because the absence of mixed terms maintains the validity of the sums of squares composition of the interval everywhere. The space is pseudo-Euclidean because the quadratic form for the ds is not positive definite. To recapitulate, (16) with the metric tensor given in (17) is valid whether we use original coordinates or coordinates that have been transformed by using (1) with (12).

But, this is only true for inertial frames (for the SR) and if we want to follow Einstein's program to find general transformations that can handle any reference system and any coordinate system, then we must be prepared to give up this restricted form. This is, indeed, necessary and the GR uses a general Riemannian coordinate space where the

g matrix, the metric tensor, can have different values in 10 of the 16 components (it is a symmetric tensor), and they will be functions of the coordinate chosen. Unfortunately, this makes the GR transformations nonlinear because of the mixed terms, a most serious complication. But, having 10 parameters available in every point of the four-space provides for the possibility of using geometry for the representation of physical effects as envisioned by Riemann in his famous lecture in 1854, and Einstein used this in his treatment of gravitation. The distribution of mass is reflected in a tensor field where the metric tensor is a function of the mass distribution.

But why are the nonlinear transformations which necessarily use all available terms in $g_{\alpha\beta}$ needed at all? If, as desired, an accelerated frame is to be used, then the transition between such frames cannot be represented by coordinate transformations that are linear in the time coordinate. And why do we want to use accelerated frames? Because we want to deal with motion in gravitational fields. And here it was Einstein's ingenious insight that in a local area, represented as differentials of the interval, s, a gravitational field is indistinguishable from the effects of a frame that is accelerated.

This is Einstein's principle of equivalence, and its position in the GR is as fundamental as is Einstein's relativity for the SR. Its mathematical meaning is that in a local area we can always find a coordinate transformation that will diagonalize the quadratic form (16). In other words, locally we can transform away a gravitational field with a transformation that makes the coordinates into a local inertial coordinate system. But we cannot hope to do this globally with one and the same transformation because the $g_{\alpha\beta}$ are functions of the location. (However, a diagonalization of the form (16) by itself is not a sufficient condition for reaching a pure inertial system.)

The use of the geometrical formalism leads to the next step. In the absence of nongravitational fields such as electromagnetic fields, a free particle will follow a line of "shortest" interval (in view of the indefiniteness of ds, it is an extreme value), i.e., a geodesic which will be its gravitational worldline. Such a geodesic worldline will satisfy the condition which makes s a maximum:

$$\delta s = \delta \int_A^B ds = 0. \tag{18}$$

Again, s is the interval, and it will not be zero (we only set the variation to zero) unless the events A and B are at the ends of a light ray. The solution of (18) presents extreme difficulties, but a considerable simplification is possible in the case of weak fields and slow (planetary or satellite) motion: We must expect the terms of the metric tensor to be very close to their values in flat space-time because the GR must deviate very little from classical Newtonian solutions. We will also be able to neglect terms with β^2, and mixed, square, and higher terms in the $\Delta g_{\alpha\beta}$. This leads to the following assumptions:

$$\left.\begin{aligned} g_{\alpha\beta} &= 0 \quad \text{for } \alpha \neq \beta \\ g_{\alpha\alpha} &= -(1 + \Delta g_{\alpha\alpha}) \\ g_{00} &= 1 + \Delta g_{00} \end{aligned}\right\} \tag{19}$$

Furthermore, if we assume only a slowly changing field, we can also set the time derivatives of the $g_{\alpha\beta}$ to zero. With these assumptions, (18) reduces to classical Newtonian equations with

$$U = -kM/r = (c^2/2)\Delta g_{00} \tag{20}$$

for a central body M at distance r. Only one $g_{\alpha\alpha}$, g_{00} appears, and this is the reason why the scalar potential U is a sufficient approximation for most problems in celestial mechanics. For our purposes, (16) will now give us the effect of a gravitational potential upon the measurement of proper time τ in relation to a coordinate time t of a comoving free frame without the potential:

$$d\tau^2 = ds^2 = g_{00}\, dt^2 - 0 \tag{21}$$

or

$$dt = d\tau \cdot \left[1 - (u/c^2)\right]. \tag{22}$$

We use only the first-order approximation in addition to (19) and (20). In other words, the potential U will produce a slower rate for the proper-time clock in respect to a clock outside the potential (which is a negative quantity). This result can also be obtained without the apparatus of GR by picturing the photon as having an effective mass $E/c^2 = hf/c^2$, which will suffer an energy loss by climbing the potential wall:

$$\Delta E = m \cdot \Delta U = hf\Delta U/c^2.$$

This produces a relative frequency change (the redshift) of

$$\Delta f/f = \Delta E/E = \Delta U/c^2. \tag{23}$$

There is an essential difference between these two derivations, however. Only the GR predicts a change in the clock rate that will accumulate the reading differences. It is a metric effect that arises because of the potential U distorting the space-time structure, which affects the relative time readings. The classical picture only speaks about a gravitational "Doppler" shift of the individual photons without necessarily saying anything about the clock readings. Experiments, described in [9] and [10], have unequivocally decided for the correctness of the GR model.

The assumptions which led to (19) mean, of course, that we have linearized the problem (in addition to the other restrictions), and that the effects, not only the gravitational potential effect, can be added separately. A good example for this is given in [9]: the estimate of the magnitude of the combined effect of orbital speed and gravitational potential

difference on the frequency received from an Earth satellite (r is the radius of the Earth, h is the satellite altitude):

$$\Delta f/f = 3.5 \cdot 10^{-10}[3r/(r+h) - 2]. \quad (24)$$

For low altitudes, we have a shift to higher frequency because the speed effect predominates; above $h = r/2$, the shift is towards lower frequencies (red shift) because at high altitudes the satellite speed is smaller and the potential difference is larger, and the two effects have the opposite sign.

IV. The Sagnac Effect

A special case of an accelerated frame is particularly important for terrestrial timing applications. On the Earth, we operate on a rotating reference system, and this will affect clock synchronization. Up to now we have assumed that ordinary clock transport will lead to the same result regardless of the route taken as long as we allow for the corrections due to the speed of travel and the height of the clock in the gravitational potential. In an inertial system, the time transfer is independent of how we travel. Moreover, the result is the same if we exchange timing signals by radio if we take the propagation time delay into account. This is done according to the Einstein convention: we send a signal to and from the remote station and assign to the remote clock as time for the reception the midtime between transmission and reception at the home station.

In a rotating system, this independence from the path of synchronization is not assured. In addition, if we synchronize station 2 to station 1 and station 3 to station 2, then station 3 will not necessarily be synchronized with 1. This transitivity of synchronization is, however, absolutely required in any time ordered system. On the Earth, however, if we use a path around the globe, we would discover a timing discontinuity that depends on the location (latitude and direction) of the path. Consistency, i.e., a useful coordinate time system can be obtained, however, if small corrections are applied to the synchronization procedure.

The same effect exists in the propagation of light. In fact, it was with light Sagnac made the original experiments in 1914. He showed that on a rotating system, the velocity of light must be added to (or subtracted from) the speed due to rotation, an effect that produces a time difference for two rays that travel in opposite directions around a closed path. This is, of course, exactly the same effect that allows the operation of LASER gyroscopes.

The reason for all this is that in the rotating system, we cannot ignore the cross terms in our line element ds (15). In other words, accelerations have an effect on timekeeping and on the propagation of light. It is instructive to study this in some detail. The line element in an appropriate coordinate system, ct, r, φ, z (coordinates that correspond with our previous x^0, x^1, etc.), is in primed coordinates for a nonrotating system:

$$ds^2 = c^2\,dt'^2 - dr'^2 - r'^2\,d\varphi'^2 - dz'^2. \quad (25)$$

In the system that rotates around the z axis at angular velocity Ω, we have $r = r'$, $z = z'$, and $\varphi = \varphi' + \Omega t'$. By inserting this in (25) we obtain in unprimed coordinates:

$$ds^2 = \left(c^2 - \Omega^2 r^2\right) dt^2 - 2\Omega r^2\,d\varphi\,dt - dr^2 - r^2\,d\varphi^2 - dz^2. \quad (26)$$

From this we can identify the terms of the metric tensor $g_{\alpha\beta}$

$$\left.\begin{aligned} g_{00} &= c^2 - \Omega^2 r^2 \\ g_{02} &= -2\Omega r^2 \\ g_{11} &= -1 \\ g_{22} &= -r^2 \\ g_{33} &= -1 \end{aligned}\right\} \quad (27)$$

If we remember (8) and its application in (14), we can do the same in the present more general case, except that we use the new expression for the line element (the interval ds) to compute the proper time:

$$\tau = (1/c) \cdot \int \sqrt{(g_{00})}\,dx^0. \quad (28)$$

For the computation of the time delays in a synchronization worldline, one would start with (16) and solve for dx^0. This is a quadratic equation with two solutions that reflect the coordinate time for the beginning and the end of the worldline. After a few algebraic transformations given in [4], one can derive the time delay in coordinate time for a closed path in the same sense as the rotation, where Ω is the angular velocity of the rotating frame, and Æ is the enclosed area, projected upon the plane of rotation:

$$\Delta t = \Delta s + 2 \cdot \Omega/c^2 \cdot \text{Æ}. \quad (29)$$

This effect is independent of the speed of transfer. Therefore, the speed effect on the readings of a portable clock would be in addition to this. We can also see that in the case of the Earth, a synchronization path along the meridians will not produce this effect at all. This is conceptually useful because we can base our standard synchronization procedure on the rotating Earth on a virtual portable clock synchronization that proceeds only along meridians, to and from the pole. Synchronization that has been reduced to this model will produce a consistent coordinate clock system on the rotating Earth without discontinuities. This is the meaning of correcting synchronization measurements according to (29).

V. Standards for Worldwide Synchronization

The main bodies that are concerned with the international coordination of time are the International Consultative Committee for Radio (CCIR), a body working under the International Telecommunications Union (ITU), and the International Conference for Weights and Measures (CIPM), the executive body under the General Conference for Weights and Measures (CGPM), also known as the Metric Convention. The consultative group concerned with

time is the Consultative Committee for the Definition of the Second (CCDS), which reports to the CIPM. Both groups, the CCIR and the CCDS, have issued recommendations and definitions pertinent to the precise meaning of synchronization and global timekeeping, and formulae to account for relativity effects. These recommendations are based to a large degree upon work that was caused by the requirements of precision timing of remote systems, e.g., of new satellite navigation systems, especially the Global Positioning System (GPS) [9], [10], [11], [12], [13], [14], [15], [16], [17].

The CCIR Study Group VII has issued a report [18] that deals in some detail with the cases of a portable clock near the surface of the Earth, and with electromagnetic signals used for remote synchronization.

The CCDS has clarified the principles involved, and has defined the International Atomic Time (TAI) as a coordinate (and coordinated) time. The wording (translated from the French) is:

> "The TAI is the temporal reference coordinate established by the Bureau International de l'Heure (BIH) on the basis of the readings of atomic clocks that operate in various establishments in conformance with the definition of the second, the unit of time of the International System of Units (SI)." (Recommendation S 2, 1970).

TAI is now established by BIH's successor organization, the Bureau International des Poids et Mesures (BIPM). This definition of TAI has been augmented with a declaration in 1980 that is to clarify the meaning in the relativistic context. This became necessary because the above wording only implies the reference time TAI to be coordinated because TAI is defined on the basis of the clock readings in the contributing establishments. The unit of time, as it is defined as a base unit of the SI, however, necessarily refers to proper time. Wherever we operate a Cesium frequency standard, it gives us the time measure as a proper unit, even though this also was not explicitly spelled out in the original definition. This CCDS declaration is as translated [19]:

> "TAI is a coordinate time scale, defined in a geocentric reference frame with the SI second as scale unit as it is realized on the rotating geoid. Therefore, it can be extended to a fixed or moving point in the vicinity of the Earth with sufficient accuracy at the present state of the art by the application of the first order corrections of the General Theory of Relativity; i.e., the corrections for the differences in the gravitational potential and the differences of speed, in addition to the rotation of the Earth."

Following this declaration, a note explains the mathematical formulae for these first-order corrections.

1. For a traveling clock we have:

$$\Delta t = Æ \cdot 2\Omega/c^2 + \int_{\text{trajectory}} \left[1 - \frac{\Delta U(r)}{c^2} + \frac{v^2}{2c^2}\right] ds \quad (30)$$

where r is the vector from the center of the Earth to the clock, ds is the proper time differential of the clock, $\Delta U(r)$ is the potential difference between the clock's location and the geoid (positive for a positive altitude), c is the velocity of light, v is the velocity of the clock in respect to the Earth, Ω is the angular velocity of the Earth, and Æ is the equatorial projection of the area covered by r in an Earth-fixed coordinate system. In the computation, the differential of the area will be taken as positive if the projection of r moves towards East:

$$Æ = \frac{1}{2} \int_{\text{trajectory}} |r| v_E \cdot \cos\varphi \cdot ds \quad (31)$$

where $|r|$ is the length of the vector r, v_E is the east component of v, and φ is the geocentric latitude.

The formula represents the relativistic effects during transport with a relative uncertainty of less than one part in ten to the fourteen. For clock altitudes of less than 24 km, one can take

$$\Delta U(r) = g_n \cdot h \quad (32)$$

where h is the altitude of the clock above the geoid, and g_n is the normal acceleration due to gravity.

The CCIR Report [18] is consistent with the CCDS note, but gives a more complete computation for $\Delta U(r)$ that can be used for greater altitudes and that includes the effects of latitude and the quadrupole moment of the Earth J_2.

2. In case of synchronization by electromagnetic signals, the coordinate time delay between transmission and reception is

$$\Delta t = \frac{2\Omega}{c^2} Æ + \frac{1}{c} \int_{\text{trajectory}} d\sigma \quad (33)$$

where $d\sigma$ is the coordinate length differential along the trajectory, with the other magnitudes the same as in the first case, with the exception of Æ which, in practice, is the equatorial projection of the triangle between the center of the Earth, the transmitter, and the receiver of the signal The area Æ is positive if the equatorial projection of the line between the transmitter and the receiver has an eastward component. It is also noted that in practice, the term due to $d\sigma$ can be eliminated by sending the signal back and forth (the simultaneous two-way method). In this case, the formula leads to uncertainties of less than 1 ns, even for connections with artificial satellites.

3. The standard numerical values for the above quantities are:

$$\begin{aligned} c &= 2.99792458 \times 10^8 \text{ m/s} \\ &\quad (299.792458 \text{ m}/\mu\text{s, in vacuum}) \\ 2\Omega/c^2 &= 1.6227 \times 10^{-21} \text{ s/m}^2 \ (1.6227 \text{ fs/km}^2) \\ g_n &= 9.81 \text{ m/s}^2 \quad \text{This is an approximate} \\ &\quad \text{value for mid} - \text{latitudes.} \\ &\quad \text{The standard formula for the} \\ &\quad \text{acceleration at sea level is:} \\ g(\varphi) &= 9.780 + 0.052 \sin^2\varphi (\text{m/s}^2) \\ &\quad \text{where } \varphi \text{ is the geographical latitude.} \quad (34) \end{aligned}$$

4. Discussion and Examples. The approximation given above treats the metric corrections as linear superposition of the three components: the speed effect, the effect due to the gravitational potential difference, and the Sagnac effect due to the motion, or light propagation, as seen in a rotating frame of reference. Taken each by itself, one can summarize these effects in the following way. A clock that moves at sea level at speed v will appear to be low in frequency compared with the standard frequency (the clock will be slow), i.e., we have to add to its frequency

$$\frac{\Delta f}{f} = -\frac{v^2}{2c^2}. \tag{35}$$

An altitude of h meters above sea level will produce an apparent approximate increase of its frequency of

$$\frac{\Delta f}{f} = \frac{g(\varphi)h}{c^2} = 1.091 \cdot 10^{-16} \cdot h. \tag{36}$$

This latter correction is an important item in the international comparison of fundamental (primary) frequency standards. One of the major contributors, the National Institute of Standards and Technology (NIST) operates a primary frequency standard in Boulder, CO. At the altitude of 1650 m, the frequency will appear to be high in respect to sea level (the TAI reference) by 1.8 parts in ten to the thirteenth, which is more than ten times the uncertainty of the absolute frequency. Even larger are, of course, the offsets that must be included in atomic frequency standards that are to operate in satellites [11].

If a traveling clock has an east component v_E in its motion, then we must add to the time of the moving clock a delay due to the Sagnac effect. This delay is (with latitude φ and altitude h) proportional to distance traveled towards the East D in km:

$$\Delta t \approx 207.4 \cos\varphi \cdot D/(2\pi(6378 + h)) \text{ (ns)}. \tag{37}$$

The 207.4 ns is the delay accumulated by a clock that circumnavigates the Earth on the equator in an easterly direction. In practice, one will compute Æ and use (30). For a signal that is transmitted via satellites, the effect will be greater, of course. The timing receivers for the GPS must have corresponding corrections for the time transfer built into their reduction software in order for time transfers to be useful in the terrestrial coordinate time. A test to this extent was performed by Allan, Weiss, and Ashby [16], with results that indicated that this was indeed the case for the receivers used.

VI. Clock Time Scales for Astronomical and Space Applications

The International Astronomical Union (IAU) has been concerned with time scales, originally as the sole provider of time standards, now in regard to their use in very-long-baseline interferometry (VLBI) and as reference for the prediction of celestial phenomena, such as orbital motions and pulsar timing. The old Ephemeris Time (ET), that was introduced in the 1950's as a uniform time scale and as the time argument for orbital computations, was replaced in 1977 by Dynamical Time, which included relativity considerations with scaling (changing the rate of the clocks to compensate for the gravitational potential at the point of origin for the purpose of avoiding a secular runoff; see [21]). However, the new space applications of precise time, particularly in pulsar research, suggest further evolution. This is very significant because by far the most stringent requirements for long-term clock stability and accuracy in the relativity corrections that must be applied in the reduction of observations come from astrodynamics and astronomical research. Guinot and Seidelmann [22] discuss the history and propose further developments.

Proposals have now been formulated at the IAU Colloquium on Reference Systems [23] for putting the whole system of space-time references on a systematic basis in the explicit framework of the GR. The Conceptual Recommendation G1 recommends

> " . . . the four space-time coordinates ($x^0 = ct, x^1, x^2, x^3$) be selected in such a way that in each coordinate system centered at the barycenter of an ensemble of masses exerting the main action, the interval ds^2 be expressed at the minimum degree of approximation in the form:
>
> $$ds^2 = -c^2 \cdot d\tau^2 = -\left(1 - 2U/c^2\right)\left(dx^0\right)^2 + \left(1 + 2U/c^2\right)\Sigma\left(dx^i\right)^2 \tag{38}$$
>
> where . . . U is the sum of the gravitational potentials (taken here as positive) . . . "

Recommendation G2 specifically mentions two coordinate systems, the (solar) barycentric and a geocentric system; it specifies that the SI (i.e., the second, the meter, c, etc.) should be extended to outer space without scaling factors; and it links the time coordinates to atomic clocks that operate in conformance with the definition of the second.

Recommendations T1 and T2 establish a consistent nomenclature for the various time-like arguments that need to be distinguished as a consequence of the above recommendations. This nomenclature takes into account the previously introduced time-like arguments that are used in the ephemerides: Terrestrial Dynamical Time (TDT), and Barycentric Dynamical Time (TDB), the former originally taken as TAI + 32.184 s, and usually considered as the relativistic successor to ET. TDB is reckoned at the SI rate with scaling (which effectively assigns a number to the Cesium standard frequency that is different from the SI). T1 and T2 also introduce new arguments in conformance with G1 and G2: Geocentric Coordinate Time (TCG), and Barycentric Coordinate Time (TCB). And lastly, it also introduces an ideal Terrestrial Time (TT) that is practically TAI but without the very small errors of implementation.

First of all, it must be noted that the French names determine the abbreviations. And even though this seems like a bewildering profusion of different time scales, the method is clearer by grouping them in the following way (see also Table 1).

Table 1 Relations Between Astronomical Timescales

1976 Recommendations			1990 Proposals
TDT	(on the Geoid) Successor to ET	identical with	TT
		$TDT \equiv TT \approx TAI + 32.184$ s	
	Geocentric Coordinate Time		TCG
		$TCG - TT \approx 6.9693 \cdot 10^{-10} \cdot \Delta T$	
↕	four-dimensional transformation to the barycenter		↕
	secular term $\approx 1.480813 \cdot 10^{-8} \cdot \Delta T$		
TDB			TCB
	$TCB - TDB \approx 1.550505 \cdot 10^{-8} \cdot \Delta T$ ΔT = (date in days-1977 January 1, 0h) · 86400 s determined in TAI		

Dynamical times were conceived for the sole purpose of providing relativistic successors to the ET. This historical origin, together with the urgency with which they were introduced before a systematic position could be reached, as it exists now, together with the 32.184-second offset inherited from ET, explains why better distinctions and definitions were needed. Moreover, the name Dynamical is misleading because it refers to the intended use, while the scale is really an atomic time. The scaling of the rate means that the standard units (e.g., of mass) must also be scaled if they are connected with TDB, a serious complication that is avoided with the introduction of the "Coordinate" times. These are not scaled but adopt the SI at the origins. This is in the spirit of the SI, which is tacitly assumed to be a proper system of units. Another problem we face is that we need to be able to conceptually separate TAI as an established, operational time, from its ideal concept, which can be better approximated after the fact by reprocessing and the inclusion of additional information (possibly, pulsar observations). That is the reasoning behind TT. It is practically identical with TDT except that the separation of the realized from the ideal time was not explicitly included in exactly this sense in the definition of TDT, which was rather vague on this point [21]. In addition, the possibly misleading implications arising from the name Dynamical are now avoided with TT. The only blemish in this otherwise logical scheme is the continuation of the 32.184-second offset on the right side of the scheme. This is confusing and a potential source of error.

Lastly, we must also recognize that TAI is a coordinate time that is not in conformance with G1 and G2 because it is not a "centric" time, but is defined on the geoid. However, all our observations are necessarily referred to it and we must, therefore, establish a connection with it. This is the purpose of transforming to TCG.

It is hoped that this nomenclature will allow unambiguous references in the discussion of work that deals with precision space observations, even though not all of these scales will actually come to be used. The planetary ephemeris tapes that are presently computed to TDB, but published in reference to TDT, will probably continue to use TDB.

References

[1] Relativity Theory has been developed by Albert Einstein during the years 1905 to 1916. The literature on the historical developments is immense and continues to grow because the subject is clearly of fundamental importance for physics but also for the rest of science and philosophy. The human element of the origins is emphasized in: A. Pais, "Knowledge and belief," *American Scientist*, vol. 76, pp. 154–158, Mar./Apr. 1990. References [8], [10], [11], and [12] are excellent introductions to the subject with emphasis on different details from the timing point of view. Reference [2] is a comprehensive, up-to-date standard text that covers the whole field.

[2] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation.* San Francisco, CA: Freeman, 1973.

[3] C. M. Will, *Theory and Experiment in Gravitational Physics.* London, UK: Cambridge University Press, 1981.

[4] R. F. C. Vessot, M. W. Levine, E. M. Mattison, E. L. Blomberg, T. E. Hoffman, G. U. Nystrom, B. F. Farrell, R. Decher, P. B. Eby, C. R. Baugher, J. W. Watts, D. L. Teuber, and F. D. Willis, "Tests of Relativistic Gravitation with a Spaceborne Hydrogen Maser," *Phys. Rev. Lett.*, vol. 45, pp. 2081–2084, Dec. 1980.

[5] L. D. Landau and E. M. Lifshitz, *The Classical Theory of Fields.* NY: Pergamon, 1975.

[6] J. G. Vargas, "Revised Robertson's test theory of special relativity," *Foundations of Physics*, vol. 14, no. 7, 1984.

[7] H. Goldstein, *Classical Mechanics*, second ed. Reading, MA: Addison-Wesley, 1980, ch. 7, paragraph 3.

[8] C. O. Alley, "Introduction to some fundamental concepts of general relativity and to their required use in some modern timekeeping systems," in *Proc. Thirteenth Annual Precise Time and Time Interval (PTTI) Applications and Planning Meeting*, 1981, pp. 687–727 (NASA Conference Publication 2220).

[9] J. C. Hafele and R. E. Keating, "Around the world atomic clocks," *Science*, vol. 177, pp. 166–170, July 1972.

[10] C. O. Alley, "Relativity and clocks," in *Proc. 33rd Annual Symp. Frequency Control*, 1979, pp. 4–39A.

[11] I. I. Shapiro, "Down to Earth relativity," in *Proc. Tenth Annual Precise Time and Time Interval (PTTI) Applications and Planning Meeting*, 1978, pp. 549–568 (NASA Technical Memorandum 80250).

[12] N. Ashby and D. W. Allan, "Practical implications of relativity for a global coordinate time scale," *Radio Sci.*, vol. 14, pp. 649–669, 1979. (This is a particularly important paper for timing applications: it discusses many specific examples, from clock trips to satellite time comparisons.)

[13] N. Ashby and D. W. Allan, "Coordinate time on and near the Earth," *Phys. Rev. Lett.*, vol. 53, p. 1858, 1984.

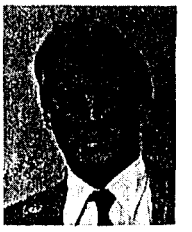
[14] C. O. Alley, *et al.*, "Differential comparison of the one-way speed of light in the east-west and west-east directions on the rotating Earth," in *Proc. Twentieth Annual Precise Time and Time Interval (PTTI) Applications and Planning Meeting*, 1988, pp. 261–286 (available from the U.S. Naval Observatory). For full details of this interesting experiment see [20].

[15] D. W. Allan and N. Ashby, "Coordinate time in the vicinity of the Earth," in Int. Astronomical Union Symp., no. 114: *Relativity in Celestial Mechanics and Astrometry*, J. Kovalevsky and V. A. Brumberg, Eds. Dordrecht, Holland: D. Reidel 1986, pp. 299–313.

[16] D. W. Allan, M. A. Weiss, and N. Ashby, "Around-the-world relativistic Sagnac experiment," *Science*, vol. 228, pp. 69–70, Apr. 1985.

[17] J. B. Thomas, "A relativistic analysis of clock synchronization," in *Proc. Sixth Annual Precise Time and Time Interval (PTTI) Applications and Planning Meeting*, 1974, pp. 425–439 (available from the U.S. Naval Observatory).

[18] Recommendations and Reports of the CCIR, "Relativistic effects in a terrestrial coordinate time system," ITU Geneva, Rep. 439-4, vol. VII, pp. 134–138, 1986. (A slightly modified version will be published in the "Green Book" of the XVIIth Plenary Assembly, Düsseldorf, Germany, 1990.)
[19] Comité Consultatif pour la Définition de la Seconde, "Declaration et Note," 9e Session, Bureau International des Poids et Mesures (BIPM), Pavillon de Breteuil, F-92310 Sèvres, France, 1980.
[20] R.A. Nelson, "Experimental comparison between two methods for synchronization of remote clocks on the rotating earth: The propagation of an electromagnetic signal using laser pulses and the transport of a hydrogen maser atomic clock," Ph.D. dissertation, Univ. Maryland, College Park, MD, 1990.
[21] G.M.R. Winkler and T. van Flandern, "Ephemeris time, relativity, and the problem of uniform time in astronomy," *Astron. J.* vol. 82, pp. 84–92, Jan. 1977.
[22] B. Guinot and P.K. Seidelmann, "Time scales: their history, definition and interpretation," *Astron. Astrophys.* vol. 194, pp. 304–308, 1988.
[23] In I.A.U. Colloquium 127 *Reference Systems,* Virginia Beach, VA, Oct. 14–20, 1990. J.A. Hughes, G. Kaplan, and C. Smith, Eds. U.S. Naval Observatory 1991, in press.



Gernot M.R. Winkler was born in Frohnleiten, Styria, Austria on October 17, 1922. He received the Ph.D. degree in physics and astronomy from the University of Graz, Austria in 1952.

From 1949 to 1953 he was with the University Observatory in Graz, and worked in industry in Austria from 1953 to 1956. From 1953 to 1956 he was a Research Associate of the Institute of Theoretical Physics in Graz, where he developed an NMR apparatus. He joined the staff of the Signal Corps Engineering Laboratories, Fort Monmouth, NJ in 1956, as a consultant in the field of precision frequency control and timing. He contributed to the development, application, and testing of atomic and molecular frequency standards. He, together with Dr. F. Reder of Fort Monmouth, proved the feasibility of transporting precision atomic clock time by aircraft in 1959. He worked in the field of VLF and arctic radio propagation from 1962 to 1966. From 1963 to 1966, he was the Director of a research division in the Institute for Exploratory Research of the Army Electronics Command where he was in charge of a wide variety of research programs in geomagnetics, radiation dosimetry, lunar radar, acoustics, and others.

He joined the staff of the U.S. Naval Observatory in 1966 as Director of the Time Service Department. In this position, he is responsible for the provision of a national time reference as used for precise timing of large electronic systems in navigation and communications. He has served on numerous advisory panels, such as the Jet Propulsion Laboratory Advisory Board for Hydrogen Masers, the National Research Council/National Academy of Science panel on the North American Datum, the National Research Council/National Academy of Science panel advisory to the National Bureau of Standards' Center for Absolute Physical Quantities and on the National Research Council Air Force Studies Board Committee on Accuracy of Time Transfer in Satellite Systems. He has authored more than 80 publications and holds 6 U.S. patents.

Dr. Winkler is a member of the International Astronomical Union (IAU, Commissions 4, 19, and 31) and was president of its Commission 31 (Time). He is a member of URSI (Commission A) and served as delegate to CCIR Study Group VII and the Committee for the Definition of the Second (CCDS) of the International Committee for Weights and Measures. In 1984, he received the rank of Meritorious Executive in the Senior Executive Service.

Applications of Highly Stable Oscillators to Scientific Measurements

ROBERT F. C. VESSOT

Invited Paper

I. Introduction

A. The Technological Basis for Experiments with Clocks

The frequency stability of highly stable oscillators has improved by a factor of about 10 every decade since the 1960 era, when atomic clocks were first introduced. While the frequency accuracy and precision of replication of the most stable oscillators has also improved, the emphasis in this discussion will be on applications of highly stable oscillators.

Most experimental procedures involving oscillators involve measurements of a time varying, or of a modulated, parameter that affects the frequency of the signal from an oscillator. The most useful measures of the error in the measurement process resulting from oscillator instability is the Allan standard deviation, $\sigma_y(\tau)$, and mod $\sigma_y(\tau)$ This is the one-sigma expectation of the fractional frequency difference (designated by the subscript, y) between time-adjacent frequency measurements, each made over time intervals of duration, τ. The functional relationship of $\sigma_y(\tau)$ versus (τ) depends on the Fourier spectrum of of the phase variations [1].

The statistical representation of oscillator performance given by $\sigma_y(\tau)$ or mod $\sigma_y(\tau)$ can provide estimates of the limits imposed on the precision of measurements when comparisons are made of data of some time dependent phenomenon averaged sequentially every τ s. Figure 1 shows sigma versus tau plots for recently developed stored ion devices [2], atomic hydrogen masers [3], and for the binary pulsar [4].

This figure also includes the Allan standard deviation of the disturbance caused by fluctuations in the earth's troposphere and ionosphere on a signal traversing vertically. In the discussion of experimental techniques, the H maser performance data in Fig. 1 will be used as a basis for numerical examples.

Manuscript received January 12, 1991; revised April 4, 1991.

The author is with the Smithsonian Astrophysical Observatory, Cambridge MA 02138.

IEEE Log Number 9101203.

B. Applications of Long-and Short-Term Stability The Long and the Short of It

1) Binary pulsar measurements referenced to terrestrial time scales: Rapidly rotating neutron stars emit bursts of microwave signals that can be detected by radio telescopes [4]. During the past 14 years the frequencies of a number of these pulsars have been monitored. These celestial oscillators have become a new astrophysical laboratory. Measurements of their frequency variations with time have provided information to test the predictions of general relativity. Results from eight years of timing a millisecond pulsar have shown that, once the drift rate has been determined and removed, the frequency stability of this object challenges the long-term stability of the best available ensemble of atomic clocks, as shown in Fig. 1. It is now evident that terrestrial clocks having better frequency stability for intervals beyond one year are needed to resolve a number of cosmological question such as the existence of a Cosmic Gravitational-Wave Background [5] and the possibility of a variability of the gravitational constant [6]. Progress in the new technology of atomic clocks based on trapped and cooled ions, and of levitated atoms, is well under way to fulfill these needs for long term frequency stability.

2) Cryogenic oscillators and short-term performance: The other end of the stability plot, where intervals less than about an hour are of interest, is becoming the domain of cryogenic oscillators. These devices include superconductive cavity stabilized oscillators [7], sapphire dielectric cavity resonators with superconductive coatings [8] and uncoated sapphire disc resonators operating with internal dielectric reflection in the "whispering gallery" mode [9]. Operating at cryogenic temperatures, resonator quality factors in the 10^9 domain have been achieved. These resonators, when combined with a low-noise amplifier such as a ruby laser [10], [11] provide high-level signals of extraordinary spectral purity.

Another oscillator that shows promise is the cryogenically cooled atomic hydrogen maser that operates at about

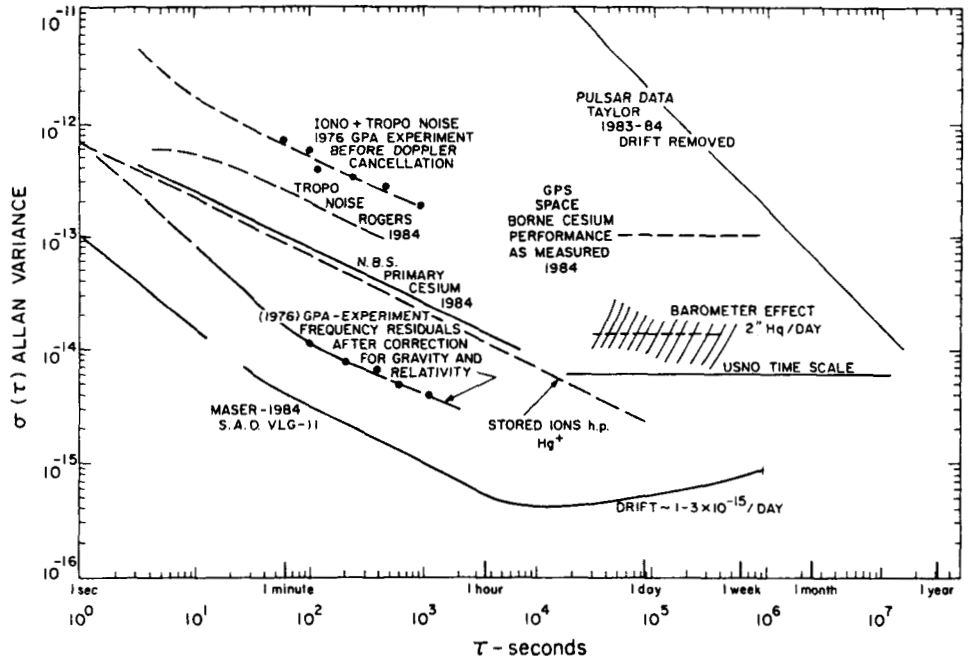


Fig. 1. Sigma versus tau plots of frequency stability of oscillators and of stability degradation owing to signal propagation through the earth's ionosphere and troposphere.

0.5 K and uses films of superfluid helium as the coating for the storage volume that confines the oscillating hydrogen atoms [12]. This oscillator is expected to provide signals with stability in the 10^{-17} to 10^{-18} domain for intervals longer than 10 seconds.

A test of the possible variability in the rate of oscillators that depends on different physical processes was conducted in 1982 between a pair of atomic hydrogen masers and an ensemble of three superconducting cavity stabilized oscillators [13]. The frequency of the hydrogen maser signal depends on (among many other things) the nature of the proton–electron magnetic hyperfine interaction. The frequency of the SCSO depends on the resonators dimensions. The hypotheses for the test is that these two frequency-determining properties might not vary in the same manner in a varying gravitational potential, $\Delta\phi$ [14]. The varying gravitational potential for this test was that of the sun, owing to the motion of the laboratory on the rotating earth and of the earth's eccentricity in its orbit about the sun. Stability data between the SCSO and the H-maser yielded a null result at the 1% level in the variation of $\Delta f/f$ divided by $\Delta\phi/c^2$ over periods of several hours.

C. Measurements using Electromagnetic Signals

1) The effects of oscillator instability on measurements of distance and of Doppler range-rate data:

An estimate of the time dispersion of a clock or oscillator predicted for a future time interval, τ, can be obtained from the relation $\sigma_{\Delta r}(\tau) \sim \tau\,\sigma_y(\tau)$. In the case of distance measurements made with the one-way propagation of light we can obtain an estimate of range dispersion by writing

$$\sigma_{\Delta r}(\tau) = c\tau\sigma_y(\tau) \tag{1}$$

where c is the velocity of light.

The one-way Doppler frequency shift expression of an oscillator transmitting at a frequency, f, moving with velocity v_r toward the receiver is

$$\Delta f = f v_r/c.$$

The contribution of the oscillator to the imprecision of determining range rate, v_r, during a measurement interval, τ, is given by

$$\sigma_{v_r}(\tau) = c\sigma_y(\tau). \tag{2}$$

Figure 2 is a nomograph of range-rate error and range distance error based on the H-maser data in Fig. 1. On the right hand axes are the scales for time dispersion, $\sigma_{\Delta r}(\tau)$, and the corresponding one-way range measurement error.

D. Systems for Cancelling First-Order Doppler and Signal Propagation

1) The three-link system used in the NASA/ SAO gravitational redshift experiment: Precise frequency comparisons can be made between widely-separated clocks in relative motion with a Doppler cancelling system [15], [16]. This technique involves continuously measuring the cycles of phase of a signal that has been transmitted to, and phase-coherently transponded from, a space vehicle. By subtracting one-half the number of these cycles from the phase of the received signal in the one-way microwave link connecting the space

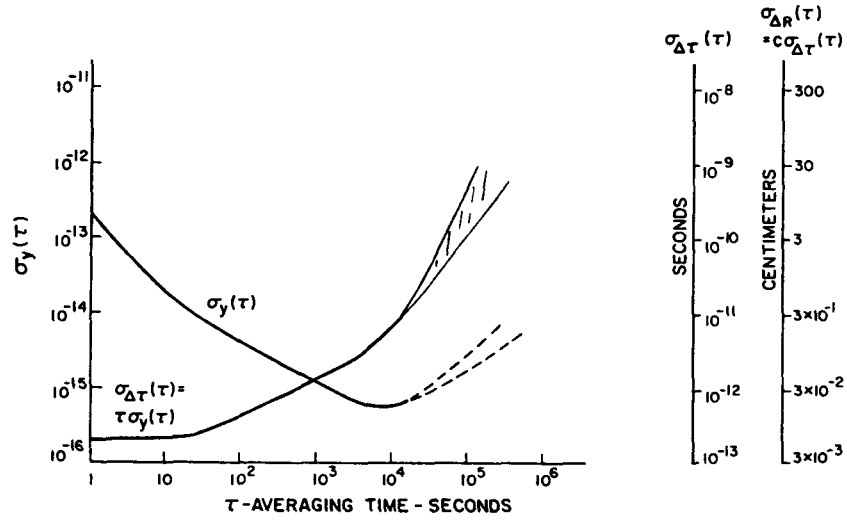


Fig. 2. A nomograph showing estimates of time dispersion and range distance error contributed by an oscillator when frequency stability is described in terms of $\sigma_y(\tau)$.

vehicle clock to the earth station, the propagation effects can be systematically removed.

The cartoon shown as Fig. 3, describes the system that was used in the 1976 SAO-NASA test of the gravitational redshift.

The fractional output frequency variations obtained by subtracting one-half of the two-way Doppler cycles from the one-way cycles received by the earth station is

$$(f_s - f_e)/f_0 = (\phi_s - \phi_e)/c^2 - |\vec{v}_e - \vec{v}_s|^2/2c^2 - \vec{r}_{se}.\vec{a}_e/c^2 . \quad (3)$$

This expression is accurate to order c^2.

Here $(f_s - f_e)/f_0$ is the total frequency shift divided by the clock downlink frequency, f_0, The term $(\phi_s - \phi_e)$ is the Newtonian potential difference between the spacecraft and earth station, $\vec{v}_e$ and $\vec{v}_s$ are the velocities of the earth station and the spacecraft, $\vec{r}_{se}$ is the distance between the spacecraft and earth station and $\vec{a}_e$ is the acceleration of the earth station in an inertial frame. For the 1976 test, an earth centered frame with axes aimed at the fixed stars was sufficiently "inertial" to satisfy the requirements of the the two hour experiment.

The first term is the gravitational redshift resulting from the difference in the Newtonian gravitational potential between the two clocks, the second term is the second-order Doppler effect of special relativity, and the third term is the result of the acceleration of the earth station, during the light time r/c, owing to the earth's rotation. (This term would be zero with an earth station at the earth's poles—a chilling prospect!) During the two-hour near-vertical flight the first-order Doppler shifts were as large as $\pm 2 \times 10^{-5}$ and the noise from ionospheric and tropospheric propagation effects was at a level of about 1×10^{-12} at $\tau \sim 100$ s, as shown in the top left curve of Fig. 1. After the frequency variations predicted in (3) were fitted to the data, we concluded that the error in the fit of the data was within $(+2.5 \pm 70) \times 10^{-6}$ of Einstein's prediction [17]. When the predicted frequency variation over the time of the mission was subtracted from the data and the residuals were analyzed, the resulting Allan standard deviation shown in Fig. 4 was obtained. Here we see that the stability of the frequency comparison made through the 3-link system over signal paths of 10 000 km, in the presence of Doppler shifts of magnitude $\pm 2 \times 10^{-5}$ of the carrier frequency, including ionospheric and tropospheric noise as shown in Fig. 1, is comparable to frequency comparison made between the two reference masers in the same room, reaching 6×10^{-15} stability at about 10^3 s.

A closer look at the system in Fig. 3 is provided in Fig. 5. Phase coherence throughout the system was provided by phase coherent ratio synthesizers. We see that there was a considerable difference in the frequencies owing to the transponders turn around frequency ratio 240/221 that could have caused serious problems from the dispersion caused by the ionosphere.

2) Removal of ionospheric Doppler shifts: The total Doppler frequency shift, f_D, due to the signal path, including the refractive indexes of the propagation medium, is given by

$$f_D = \frac{f}{c}\frac{d}{dt} \int_p n(t) ds \quad (4)$$

where $n(t) = n_A(t) + n_I(t)$ is the sum of the time-varying refractive indexes of the atmosphere and the ionosphere, respectively. The refractive index of the atmosphere, at S-band, has no significant frequency dependence and has no effect on the 3-link Doppler cancellation system when the propagation time is short compared to the measurement interval.

However, propagation through the time varying ionosphere can cause large frequency variations in a received signal. In the 1976 test these shifts were estimated to

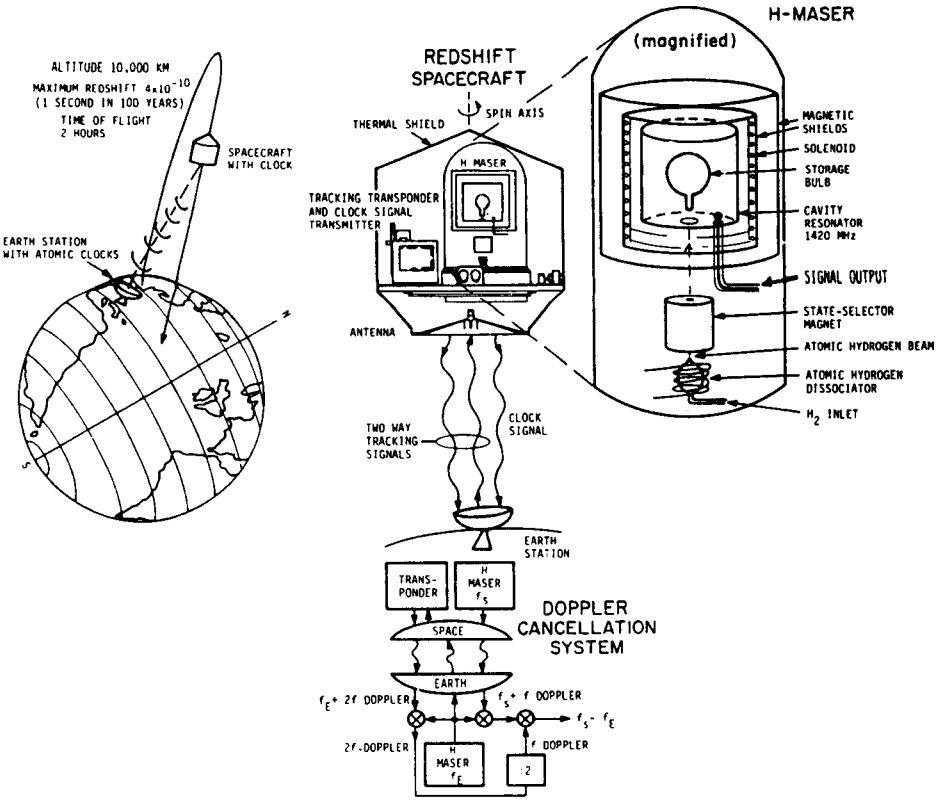


Fig. 3. The 1976 SAO/NASA gravitational redshift experiment.

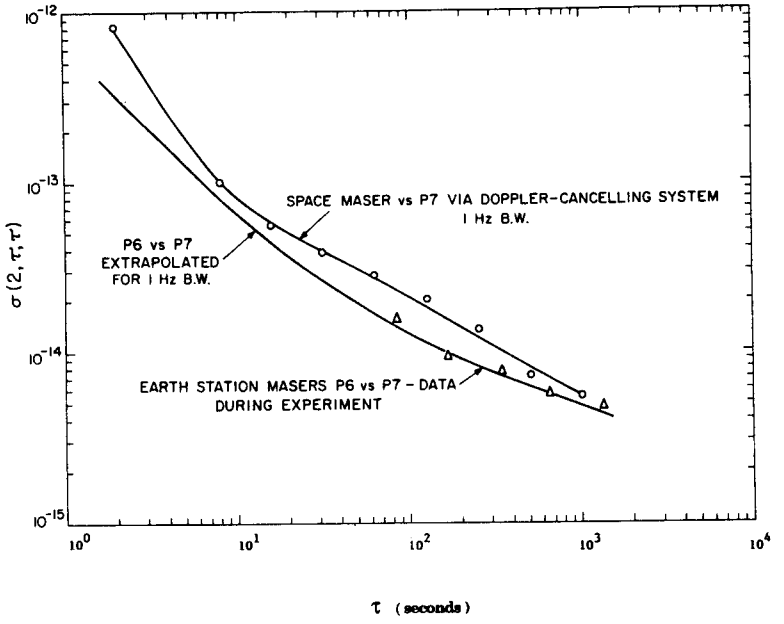


Fig. 4. Allan standard deviation of frequency stability of the two ground station hydrogen masers, P-6 and P-7 made simultaneously with the comparison with the GP-A space maser during the two hour near-vertical mission flown on June 18, 1976. The space maser data are from the frequency residuals after removal of the predicted relativistic and gravitational effects.

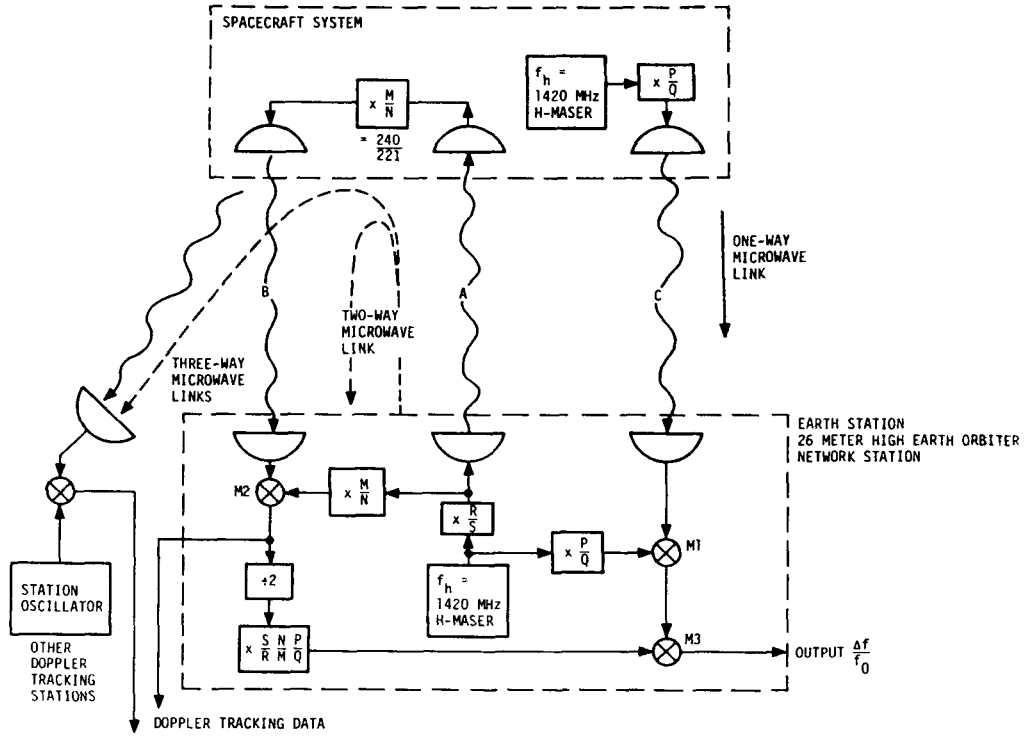


Fig. 5. A schematic diagram of the Doppler cancelling system used to obtain the data shown in Fig. 4. The frequency ratios $R/S = 82/55$, $P/Q = 76/49$, and $N/M = 221/240$ were selected to minimize the effect of ionospheric dispersion.

be several parts in 10^{-10} and would have completely overwhelmed the data from the combined redshift and second-order Doppler effect.

Frequency shifts caused by variations in ionospheric electron density have been analyzed by Tucker and Fannin [18] who write the ionospheric refractive index as follows:

$$n_I = [1 - f_n^2/f^2(1 \pm f_m/f)]^{1/2} \tag{5}$$

where $f_m = \mu_o He/2\pi m = 2.8$ MHz/Oersted, is the effect of the electrons' magnetic interaction (which for S-band signals near earth is small, and in our case could be neglected). The quantity $f_n^2 = \rho e^2/(2\pi)^2 \varepsilon_o m$ is the square of the electron plasma frequency. Here ρ is the electron density, e and m are the electrons charge and mass, and ε_o is the permittivity of free space.

Following Tucker and Fannin, the frequency shift owing to the change in the amount of ionization in the propagation path f_{DI}, is given by including (5) in (4) and expanding, neglecting the magnetic term:

$$\begin{aligned} f_{DI} &= \frac{f}{c}\frac{d}{dt}\int_p \left(1 - \frac{\rho e^2}{8\pi^2 \varepsilon_0 m}\right) ds \\ &= -\frac{e^2}{8\pi^2 f c \varepsilon_0 m}\frac{d}{dt}\int_p \rho(t) ds \\ &= -\frac{40.5}{fc}\frac{d}{dt}\int_p \rho(t) ds. \end{aligned} \tag{6}$$

Here, p is the propagation path over which the integral is taken. The integral represents the columnar electron density in the path. When we follow the ionospheric Doppler frequency shifts through the two transponder links and the clock downlink, we find that the ionosphere contributes a frequency error at the output from mixer M3 of Fig. 5 in the amount

$$\begin{aligned} |f_{1\,error}| = \frac{4.05}{cf_h}\frac{d}{dt}\int_p \rho(t) ds \\ \cdot [Q/P - (1 + N^2/M^2)S^2P/2R^2Q]. \end{aligned} \tag{7}$$

This error can be removed by choosing the ratios so as to make the quantity in the square bracket equal to zero [19]. In the 1976 redshift experiment the transponder ratio, N/M, was 221/240. We chose $P/Q = 76/49$ and $R/S = 82/55$, so that the resulting error was 2.5×10^{-5} of the ionosphere Doppler shift in the one-way link. We estimated that [20] under typical ionospheric conditions $\Delta f_{DI}/f$ could have been as large as 3×10^{-10}, which is comparable to the relativistic and gravitational effects we measured. Without cancellation of the ionospheric effect the experiment would not have been possible.

E. A Symmetrical Four-Link Doppler Cancelling System

By transponding the clock downlink back to the space-

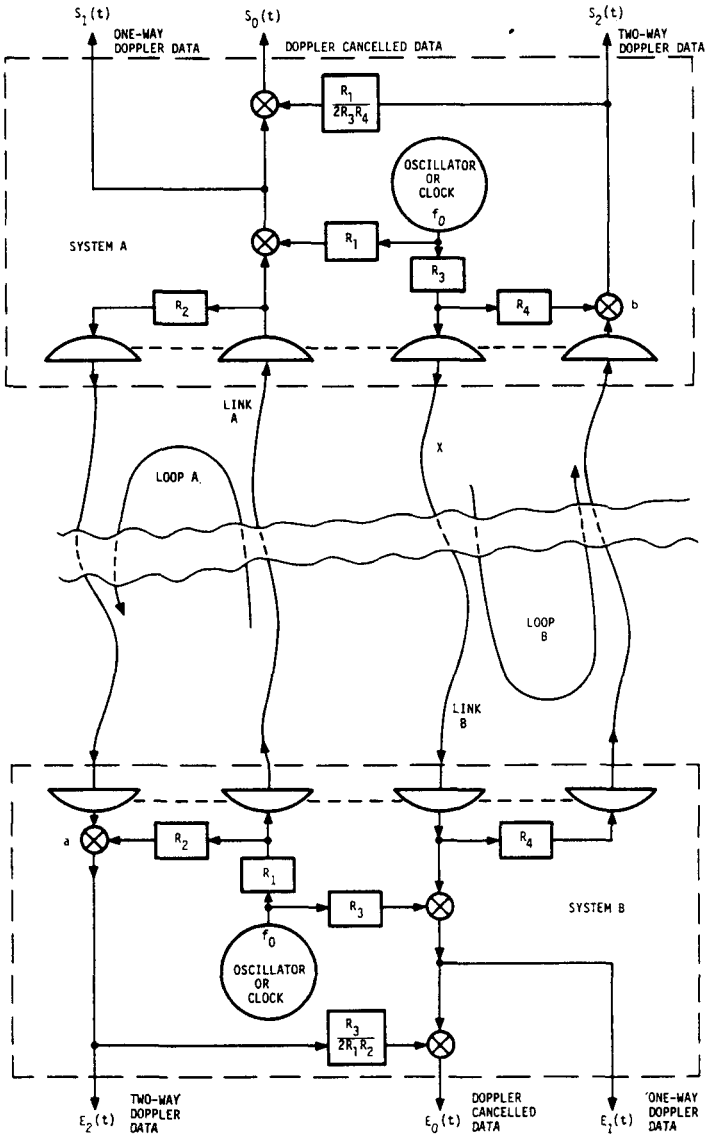


Fig. 6. A four-link Doppler cancelling system that allows time correlated data to be obtained both at the earth and space terminal.

craft we can make the system shown in Fig. 5 symmetrical [21] and record Doppler cancelled data, one-way and two-way data at each end of the system, as shown in Fig. 6. The continuum of space-time paths of the four signals in Fig. 6 is shown in Fig. 7. Here the dots signify the clocks, and the arrows $E_1(t)$ and $E_2(t)$ signify signal outputs representing earth-based one-and two-way data at a particular epoch in the continuum; $S_1(t)$ and $S_2(t)$ represent space-based one-and two-way data.

In addition to providing Doppler cancellation for intervals $\tau < R/c$, the 4-link system also makes possible high precision measurements of Doppler signals over very long distances. By time-correlating the Doppler responses we can systematically cancel a strong localized noise source such as the earth's troposphere and ionosphere [22]. For example, let us consider the frequency variations in the $E_1(t), E_2(t), S_1(t)$, and $S_2(t)$ Doppler outputs shown in Fig. 4, assuming the complete removal of the smoothly varying Doppler shifts owing to relative motion. We see that the iono-tropo noise pattern received from the the spacecraft transmitted at time t_i and received at earth at time $t_i + R/c$, is the same as the noise received at the spacecraft at time $(t_i + 2R/c)$. By advancing $E_1(t)$ by time R/c with respect to $S_1(t)$ and subtracting the two data sets we can systematically remove the noise in the $S_1(t_i + 2R/c) - E_1(t_i + R/c)$ combined data set at the small expense of increasing the random noise in the data by $\sqrt{2}$. In situations where the localized dominant noise is substantially larger than the nonlocalized random noise, this process can be highly effective.

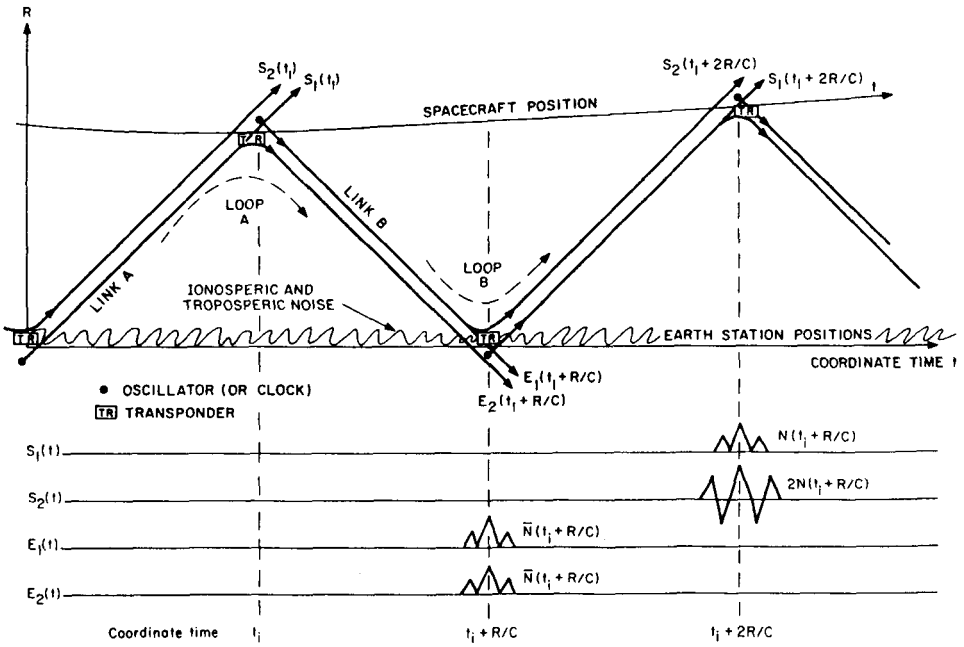


Fig. 7. A space-time diagram of the signals as shown in Fig. 6, showing how time correlation of the data can systematically remove a dominant and localized noise source.

II. Gravitational And Relativistic Measurements With The 4-Link Doppler System

A. Relativistic Doppler Shifts and Redshifts

The Doppler cancelled signal outputs $S_o(t)$ and $E_o(t)$ in Fig. 6 contain relativistic and gravitational information that can be time correlated. Equation 3 is repeated below as $E_o(t)$, along with its counterpart expression at the spacecraft, $S_o(t)$:

$$E_0(t) = \frac{\phi_s - \phi_e}{c^2} - \frac{|\vec{v}_s - \vec{v}_e|^2}{2c^2} - \frac{\vec{r}_{se} \cdot \vec{a}_e}{c^2} \qquad (8)$$

$$S_0(t) = \frac{\phi_e - \phi_s}{c^2} - \frac{|\vec{v}_e - \vec{v}_s|^2}{2c^2} - \frac{\vec{r}_{es} \cdot \vec{a}_s}{c^2}. \qquad (9)$$

By adding the two time ordered data sets we cancel the first term and double the magnitude of the second term, conversely if we subtract the data sets, we double the first-term and cancel the second term. In both instances we must account for the components of acceleration of the earth and space stations along the line of sight in a suitable inertial frame.

B. A Test of Relativistic Gravitation with a Clock in a 24 Hour Eccentric Earth Orbit

The original concept for testing the Gravitational Redshift called for operating a spaceborne clock in a spacecraft placed in a highly eccentric twenty four hour earth orbit [16], [23]. Low inclination orbits with eccentricities as high as 0.6 can produce apogee-to-perigee redshifts of about 4.8×10^{-10}, and still keep the spacecraft in view of an earth station with a lowest elevation angle of about 15 degrees. Accompanying the redshift there is a second-order Doppler shift of comparable magnitude, which produces a combined frequency variation of 9.6×10^{-10} in the Doppler cancelled data described in (8).

Table 1 shows the error analysis of the combined gravitational redshift and second-order Doppler test made in 1976 using a near vertical trajectory. This can be compared with the predicted results of a proposed orbital test. The improvement from 70 parts per million to 2 parts per million results partly from improved clocks, but mostly from the longer averaging intervals and lower estimates of bias errors made available by having time for adjusting and tuning the space maser before taking data.

C. An Extension of the GP-A Experiment to Test General Relativity with a Solar Probe

A series of studies has been conducted [24] of tests of relativistic gravitation using a clock in a space probe going over the poles of the sun within 4 solar radii of the sun's center. The time of travel from pole-to-pole is roughly 14 hours. This test is intended to reveal the behavior of the second order in the redshift, $[\Delta\phi/c^2]^2$, in a varying gravitational potential.

During the 10 hours before and after the perihelion, the value of Φ/c^2 varies from 5.3×10^{-7} at perihelion to 2.0×10^{-7} at 10 hours from perihelion, During the same time interval, the second order redshift $[\Phi/c^2]^2$, varies from 2.81×10^{-13} to 4×10^{-14}, as shown in Figs. 8 and 9. Taking the Allan standard deviation of today's H-masers over 10 hours averaging time as 6×10^{-16}, the inaccuracy of measurement imposed by the maser instability for the

Table 1 Error Analysis of a Proposed Orbiting Redshift Test and Comparison with the 1976 NASA-SAO GP-A Rocket Probe Experiment

	GP-A Test	Proposed Orbiting Test	
	2-Hours	1 Orbit	30 Orbits
Random Noise—$\Delta f/f$			
H-Maser and Microwave System Noise	13×10^{-15}	$< 1 \times 10^{-16}$	$< 1 \times 10^{-16}$
Uncancelled Atmospheric Noise	2.5×10^{-15}	2.5×10^{-15}	0.05×10^{-15}
Uncancelled Ionospheric Noise	5.0×10^{-15}	2.0×10^{-15}	0.37×10^{-15}
RSS	14.2×10^{-15}	3.2×10^{-15}	0.58×10^{-15}
Possible Systematic Bias Errors—$\Delta f/f$			
Ground Station Clocks	3.0×10^{-15}	1.0×10^{-15}	1.0×10^{-15}
Space Clock	4.5×10^{-15}	1.0×10^{-15}	1.0×10^{-15}
Spacecraft Tracking Error	3.3×10^{-15}	1.0×10^{-15}	1.0×10^{-15}
RSS	6.3×10^{-15}	1.7×10^{-15}	1.7×10^{-15}
RSS Combined Random and Systematics—$\Delta f/f$	15.5×10^{-15}	3.6×10^{-15}	1.8×10^{-10}
Variation of $\Delta f/f = \Delta\phi/c^2 + 1/2v_{rel}^2/c^2$	2.2×10^{-10}	9.6×10^{-10}	9.6×10^{-10}
Fractional Accuracy of Test			
$\frac{\text{RSS Combined Random and Systematics}}{\Delta\phi/c^2+1/2v_{rel}^2/c^2} =$	70×10^{-6}	3.8×10^{-6}	1.9×10^{-6}

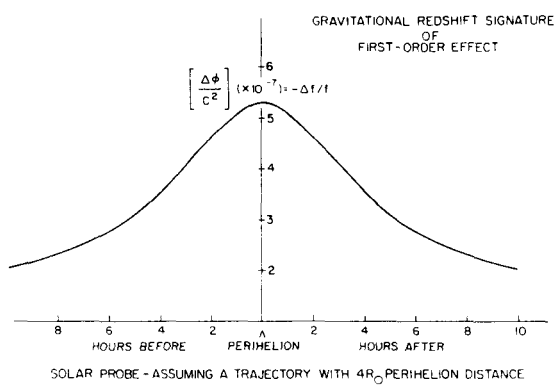


Fig. 8. The gravitational redshift signature at first order in $\Delta\phi/c^2$ for a clock passing over the poles of the sun with 4-solar radius distance from the sun's center when over the equator.

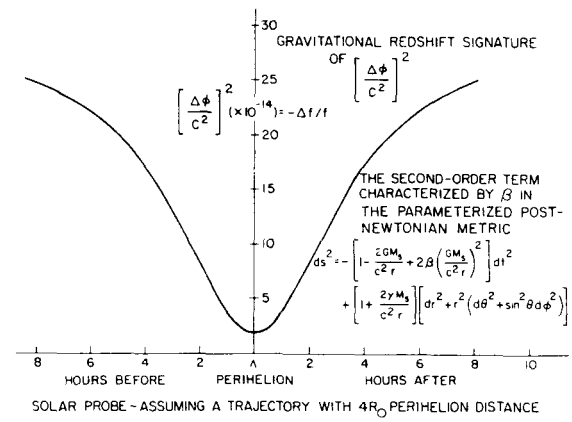


Fig. 9. Second-order redshift signature for a clock passing over the poles of the sun with 4-solar radius distance from the sun's center when over the equator.

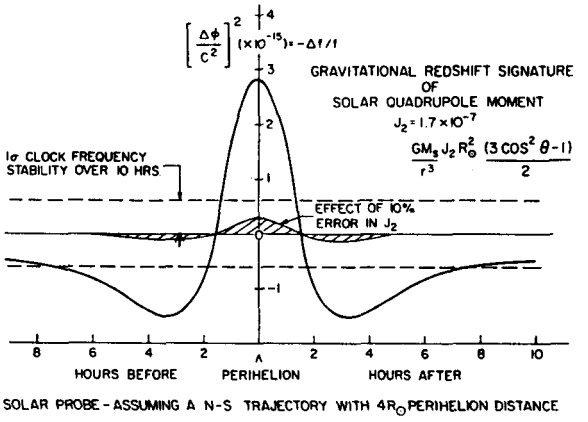


Fig. 10. Gravitational redshift signature at first-order in $\Delta\phi/c^2$ of the solar quadrupole moment, J_2, assumed to be 1.7×10^{-7}.

first-order measurement is 1.8×10^{-9}, the corresponding inaccuracy for the second-order measurement is 2.5×10^{-3}.

The sun's gravitational potential is complicated by having a number of multipole components. The largest of these is the solar quadrupole moment, J_2, which must be accounted for in an accurate measurement of the second order term in the redshift. Measurements of J_2 have been made from solar oscillations [25], and we can confirm the effect of the uncertainty in these measurements from the behavior of the J_2 signature in the data during the 14 hour pole-to-pole passage.

At order c^{-2} the first-order redshift has the following behavior:

$$\Phi/c^2 = \mu/r + \mu/r^3 J_2 R_{\text{sun}}^2 (3\cos^2\vartheta - 1)/2 \quad (10)$$
$$= \Phi_1/c^2 + \Phi_2/c^2 \quad (11)$$

where $\mu = GM_{\text{sun}}/c^2$. Assuming $J_2 = 1.7 \times 10^{-7} \pm 0.17 \times 10^{-7}$, then at perihelion,

$$r = 4R_{\text{sun}}, \quad \vartheta = \pi/2$$

and

$$\Phi_2/c^2 = -2.8 \times 10^{-15} .$$

At about ±7 hours from perihelion,

$$\vartheta = 0, \quad r = 8R_{\text{sun}}$$

and

$$\Phi_2/c^2 = +7.0 \times 10^{-16} . \quad (12)$$

The frequency variation caused by the J_2 contribution to the sun's redshift is shown in Fig. 10 over the time 7 hours before and after perihelion. Its peak-to-peak magnitude is 3.5×10^{-15}. If we have an error of 10% in the estimate of J_2, the uncertainty in its contribution to the redshift over this interval is about 3.5×10^{-16}, comparable to the instability of the clock over the 14-hour passage. The error contribution will have a distinctive $(3\cos^2\vartheta - 1)/r^3$ signature in contrast to the very smooth $1/r$ dependence of the first-order redshift and of the $1/r^2$ dependence of the second-order redshift, as shown in Figs. 8 and 9.

An important feature that makes this experiment possible is the ability to take Doppler cancelled data at the probe. In Fig. 7, the earth's tropo-iono noise having magnitude $N(t_i + R/c)$ occurring at $t = t_i + R/c$ is reported in the spacecraft's two-way Doppler signal at $t = t_i + 2R/c$ with magnitude $2N(t_i + R/c)$. The one-way data, S_1 that is received at the probe at $t_i + 2R/c$ and reported with magnitude $N(t_i + R/c)$ passes through the same tropo-iono conditions as did S_2. The spacecraft Doppler cancellation system can thus *systematically* remove the effect of the tropo-iono noise when it produces the $S_o(t_i + 2R/c)$ data. Note that this is *not* the case for the earth station Doppler cancelled output E_o. In this case there would be about 1000-s delay between the uplink transmission and reception from the transponder at a time $2R/c$ later. The combined atmospheric and ionospheric delay could have varied considerably during this time.

D. The Search for Gravitational Radiations using Doppler Techniques

During the long travel time to Jupiter, the NASA Galileo mission is expected to provide an opportunity to search for gravitational radiation using Doppler techniques with the spacecraft transponder in conjunction with the Deep Space Network Tracking Stations that are equipped with H-maser oscillators [26], [27].

In Einstein's General Theory of Relativity (GRT), gravitational radiation results whenever a massive body is accelerated. Rotating binary stars radiate energy and will eventually collapse together. While evidence for such radiation has not been observed directly, the orbital decay of a binary pulsar has been observed since 1975 and its rate continues to follow very closely the predicted behavior for loss of energy by gravitational radiation [28].

According to GRT, gravitational radiation is described as a wavelike distortion of space-time traveling at the speed of light. When a gravitational wave intercepts an electromagnetic wave, it distorts the frequency of the wave by an amount $h = \Delta f/f$. It should be possible to detect gravitational waves by observing Doppler shifts of a signal that is transmitted by a highly stable microwave (or laser) transmitter and detected by a receiver located at a distance that is greater than about one-half the wavelength of the gravitational wave.

An example of the possible Doppler detection [29] of a pulsed gravitational wave using the 4-link Doppler measurement system is shown in Fig. 11. Here the wavefront of the gravitational pulse is assumed to intercept the earth-probe line at an angle, $\theta = 60$ degrees. With the spacecraft transponder, as in the Galileo experiment, the effect of the pulse would be observed three times in the Earth 2-way Doppler trace:

1) By a Doppler shift of the gravitational-wave disturbing the earth station at $t = t_1$, while it is receiving a signal transmitted earlier by the spacecraft;

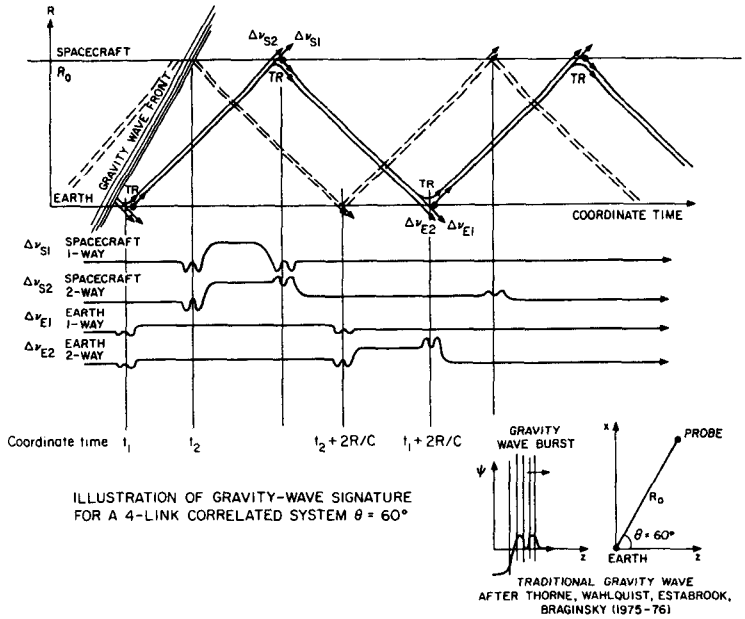


Fig. 11. Illustration of pulsed gravitational wave signatures observed by the 4-link Doppler system, shown in Fig. 6.

2) By its "echo" when the earth station receives a transponded signal at $t = t_1 + 2R/c$;
3) By the disturbance when the gravitational wave arrives at the spacecraft at time t_2 and reported at earth at $t = t_2 + R/c$.

The spacing of the pulses, their time signature designated by the parameter $\Psi(t)$, and the relative magnitude and sign of the signature are described by a single parameter $\mu = \cos\theta$ [30], [31]. Here t_R signifies arrival time at the first station:

$$\frac{df}{f} = \frac{(1-\mu)\psi(t_R)}{2} - \mu\psi\left[t_R - L\frac{(1+\mu)}{c}\right] + (1+\mu)\psi\left(t_R - 2\frac{L}{c}\right). \quad (13)$$

The one-way transmission from the spacecraft would only show the pulses at t_1 and $t_2 + 2R/c$.

A similar set of five observations of the gravitational pulses is available at the spacecraft. In this case, $\mu = \cos(\pi + \theta)$ and another set of five manifestations of the pulse appears in the spacecraft data. While only four of the ten pulses, i.e., those from the two one-way Doppler signals, are unique, the other six are obtained from other paths through the electronics system and offer redundancy to avoid system aberrations masquerading as gravitational wave signals.

If one of the stations is on earth, noise from the earth's troposphere and ionosphere would be the main limitation to the sensitivity of detection. Because of the $1/f^2$ frequency dispersion of the ionosphere, reduction of its noise is possible by operating at higher frequencies than the presently used S-band (2 GHz) and X-band (10 GHz) systems. Future tracking systems are planned to operate at 33 GHz and even higher.

However, tropospheric noise cannot be reduced by such techniques and will substantially degrade the stability of a signal. Studies [32] show that the Allan deviation of the tropospheric noise for signals passing vertically has a $\tau^{-2.5}$ behavior for intervals between 20 and 200 s, with $\sigma_y(100\text{ s}) = 8 \times 10^{-14}$, as shown in Fig. 1. While it is possible to model the tropospheric frequency shifts using other data, such as the columnar water vapor content and the local barometric pressure, tropospheric propagation variations will nevertheless severely limit the detection of gravitational radiation with transponded two-way Doppler signals. Estabrook [22] estimates that the sensitivity with ideal tropospheric conditions at night in the desert will be at a level of $h = \Delta f/f \sim 3 \times 10^{-15}$ for gravitational waves in the millihertz region, which is one to two orders of magnitude above the levels estimated by astrophysicists [31].

As mentioned earlier, in the discussion of Fig. 7, the tropospheric noise can be removed systematically by simultaneously recording of Doppler data from a clock in a spacecraft and at the earth station and combining these data. Computer simulations of this process [22] show nearly complete rejection of such spatially localized sources as the near-earth tropospheric and ionospheric variation and the earth station antenna motion noises. In a future situation where both clock systems are in space, and operating at frequencies where the noise from the solar corona ionization is not significant compared to clock stability, the principal nongravitational noise sources will likely be from the buffeting of the space vehicles by nongravitational forces such as light pressure, particle collisions and sporadic outgassing of the spacecraft. Here, again, since these

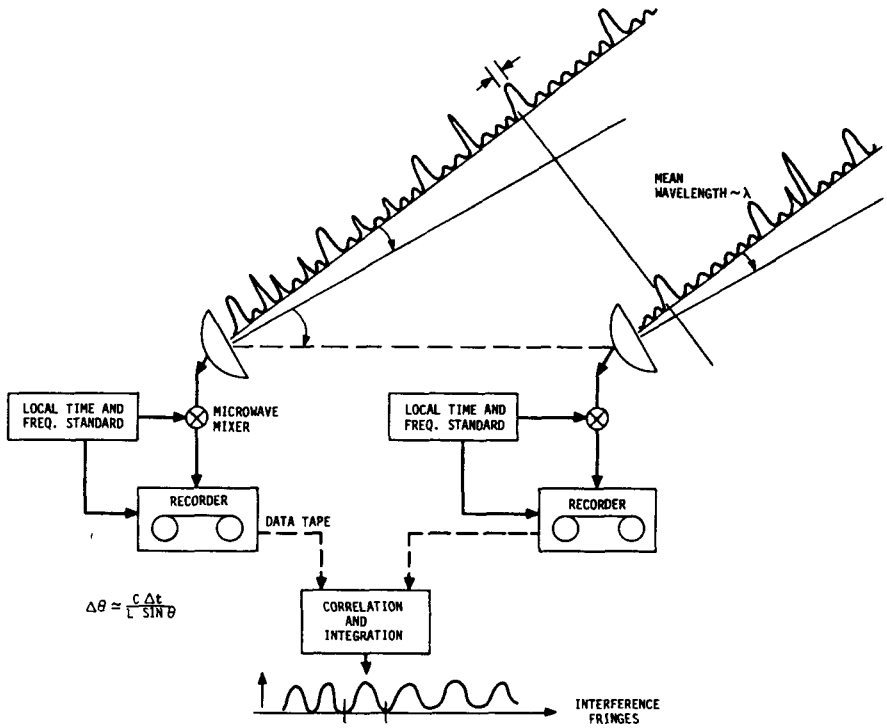


Fig. 12. Schematic of the VLBI technique used in radioastronomy.

disturbances are localized at the ends of the system, the time signatures of the noises are separated by R/c, and can be distinguished from the patterns expected from pulsed gravitational waves, which have signatures that depend on the parameter μ of [13].

III. High Resolution Angular Measurements By Very Long Baseline Interferometry (VLBI)

A. The Effect of Oscillator Instability on the Measurement of Angles

High resolution angular measurements are of interest in astrometry and light deflection tests of relativistic gravitation [33]. Measurement of the angle between the propagation vector of a signal and the direction of a baseline, defined as the line between the phase centers of two widely-separated antenna, can be made with VLBI techniques [34]. In Fig. 12, two radio telescopes (or spacecraft tracking stations), separated by a distance L, each detect the arrival of radio noise signals from a distant radiostar (or a modulated signal from a spacecraft). After heterodyning to a lower frequency the noise signals are recorded as a function of time. This recording is usually made on magnetic tape and the two sets of noise data are subsequently brought to a computer facility to be time correlated. The observable quantities from the correlation process are the correlated amplitude and the relative phase of the signals detected at the widely separated points on the wavefront.

The stability limit on the successive measurements of angle imposed by the oscillator instability on successive measurements of angle taken τ seconds apart is

$$\sigma_{\Delta\theta}(\tau) \sim c\tau\sigma_y(\tau)/L\sin\theta\ , \tag{14}$$

where θ is the angle between the propagation vector and the baseline. The result of correlating the noise data obtained from a common source by the two stations is the production of fringes analogous to those observed from two-slit optical diffraction. The spacing between the fringes is $\lambda/L\sin\theta$, where λ is the average wavelength of the signals arriving at the antennas. The visibility of the fringes depends on the extent to which the signals arriving at the antennas are correlated. The angular resolution of the interferometer in the direction of the source is given by the change of fringe phase, ϕ, with source angle, θ

$$\frac{d\phi}{d\theta} = \frac{2\pi L}{\lambda}\ . \tag{15}$$

The error in successive angular measurements owing to the instability of the clocks in a terrestrial system with $L = 6000$ km, assuming $\sigma_y(10^3\ \text{s}) = 1 \times 10^{-15}$, and $\theta = \pi/2$, is given by

$$\sigma_{\Delta\theta}(10^3\text{s}) = 5 \times 10^{-11} \text{ radians or } 2\mu \text{ arc s.}$$

However, this is far smaller than the realistic limit on angular measurement with terrestrial stations. The effect of tropospheric and ionospheric fluctuations impose limits that are far more serious than clock instability.

To date (December 1990) the best resolution observed has been at the 100 micro-arc-second level over an 8000

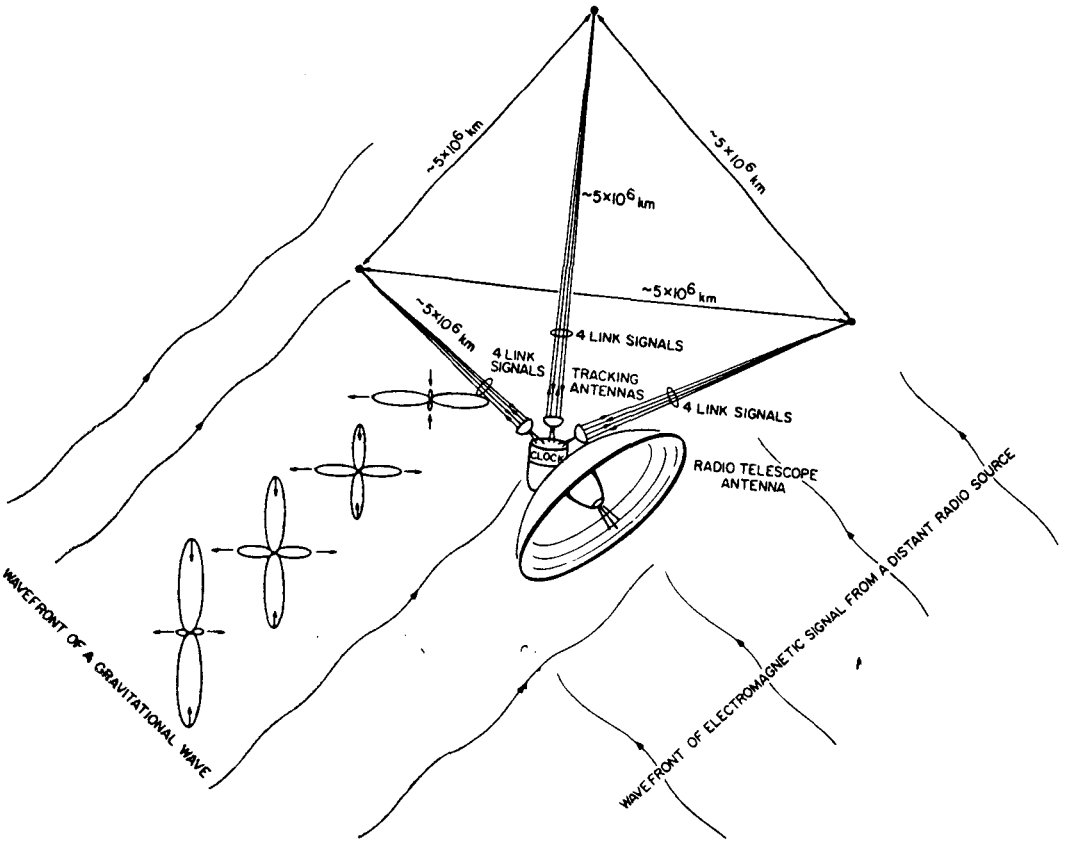


Fig. 13. An array of four spaceborne radio telescopes, each connected to the other by the system shown in Fig. 6.

km baseline in a VLBI experiment operating at 7 mm wavelength [35].

Tropospheric and ionospheric propagation limits and the limits imposed by the size of the earth on the baseline distances can be overcome by operating VLBI stations in space. A successful demonstration of a spaceborne radio telescope operating as a VLBI terminal was made in 1986 [36] using NASA's orbital Tracking and Data Relay Satellite System (TDRSS) system as a spaceborne radiotelescope in conjunction with a number of radiotelescopes on earth.

As an example of the limits that a spaceborne system could achieve, let us consider a spaceborne system where $L = 5 \times 10^6$ km, $\sigma_y(10^4\text{s}) = 4 \times 10^{16}$, and $\theta = \pi/2$. In this case,

$$\sigma_{\Delta\theta}(10^4) = 2 \times 10^{-13} \text{ radians or } 0.05\mu \text{ arc s}.$$

For $\lambda = 1$ mm we have $\lambda/D = 2 \times 10^{-13}$ radians and we see that the limit imposed by clock stability with 10^4 s integration time is capable of resolving fringes at wavelengths, λ, as short as 1 mm in a spaceborne system with baseline distances of 5×10^6 km. The numbers in this example are chosen to be close to present estimates of the limits for having reasonably well correlated flux at the two stations [37] at the distances chosen.

In addition to making astronomical and astrophysical observations, terrestrial VLBI systems are used to record polar motion and rotation of the earth and to monitor the movements of the earth's tectonic plates. Relative positions of radio stars and features of their brightness distribution can be made with a precision of a few tenths of a milli-arc-second, however the absolute directions in space of the baselines between VLBI stations depend on the choice of a frame of reference which is usually taken from the position of very distant radio sources.

B. A Spaceborne Four Terminal (VLBI) Array that Establishes an Inertial Reference Frame, a "Gedanken" System to Exercise our Imaginations

Having already stretched the VLBI technique to baseline distances of 5 million kilometers let us go several steps further and postulate the existence of an array of four such stations in the form of a tetrahedron that defines a three dimensional figure in space that is in an orbit about the sun. (The station separations need not be equal) Fig. 13 shows such an array where each station contains a clock that is synchronized to a coordinate time scale and is connected to its three neighbors by the 4 link system shown in Fig. 6 [38]. The six baselines distances define an object in space whose shape is very precisely known as a function of time

from distance measurements made by measuring the phase of the two-way signals at the stations defining the ends of each baseline.

The orientation of the array of six baselines poses an interesting problem. It should be possible to determine changes in the orientation of the array in terms of an inertial frame defined by the constancy and isotropy of the velocity of light by invoking the Sagnac effect [39]. This effect is the basis for today's laser gyroscopes. To describe the effect, we can visualize the arrival times of light signals sent in opposite senses about a closed path. If the transmitter and receiver are on a surface rotating at Ω rad/s, and their signals go about a path with projected area, A, on that surface, the difference in the arrival times of light signals going around its perimeter is $\Delta\tau = 4\Omega A/c^2$. If we measure the difference in arrival times of signals going in opposite senses about the triangle defining one face of the tetrahedron, we can obtain the component of rotation for that face. From the four triangles that define the tetrahedron we have four rotation components that allow us to measure the rotation vector, and, because the measurement is overdetermined, we can estimate the accuracy of its measurement.

The limit of precision in the determination of the rotation rate imposed by the maser frequency stability performance described in Fig. 1 depends on how the system is operated. If the signals originate from a single clock and are simultaneously transmitted about the three legs in opposite senses and arrive at the point of origin with time difference, τ, the error in the determination of τ, $\sigma_{\Delta\tau}(\tau)$, is given by the time dispersion plot shown in Fig. 2. For intervals up to about 20 s, the white phase noise of the maser oscillator dominates and $\sigma_{\Delta\tau}(\tau)$ is constant at about 2×10^{-13} s. The corresponding error owing to oscillator instability in the determination of in the component of the rotation vector, Ω, normal to plane A of the array, is given by $\sigma_{\Delta\Omega}(\tau) = \sigma_{\Delta\tau}(\tau)c^2/4A = 4 \times 10^{-16}$ rad/s. If the array is in a solar orbit with radius 1 AU, the measurement would have to account for the Einstein-deSitter precession of 2×10^{-2} arc s/yr (3×10^{-15} rad/s) owing to the bending of space-time by the suns gravity.

This array offers possibilities for astrometric observations rather than imaging of radio sources. However, it may have other applications. Essentially it compares the frame of reference defined by the most distant radio sources using VLBI with an inertial frame defined by the local isotropy and constancy of the velocity of light. This comparison opens, in a modern context, the question of inertial frames raised long ago by Bishop Berkeley, Ernst Mach, and others. This system may provide a way to observe some aspects of the behavior of the missing matter in the universe, which is alleged to have mass overwhelmingly greater than the mass of celestial bodies observed by electromagnetic means. We look to theorists to provide scenarios of processes that could interact with this system by considering effects due to various hypothetical types of missing matter.

This system could be extended to detect pulsed gravitational radiation. Here the array of baselines could provide as many as 60 manifestations of the signature from a single pulse and give information about the speed of propagation, the direction, and the polarization of the gravitational wave.

IV. Conclusion

Since the mid-1960's the frequency stability of atomic clocks (or oscillators) has been improving with no end in sight. The units of time and frequency, and the now redefined unit of distance through the velocity of light, are solidly based on atomic frequency standards. This metrology, in a local sense, has been made consistent with the present concepts of gravitation and relativity. Measurements of astronomical and astrophysical quantities near the edges of our universe are now being made in terms of quantum phenomena that occur in particles whose physics encompass staggeringly smaller distance scales. As the performance of atomic clocks improves (and, hopefully, as our national willingness to support fundamental science in space) we should expect some surprises about the nature of our universe, including the relationship between gravitation and the three other known fundamental forces.

There are also new frontiers to explore in spectroscopy, especially in low temperature atomic spectroscopy with the advent of new techniques in cryogenics, laser cooling, and trapping, where the physics of ultralow energy particle interaction will undoubtedly also surprise us.

Acknowledgment

The writer gratefully thanks his colleagues, E. M. Mattison, R. R. Reasenberg, R. L. Walsworth, and J. M. Moran for many helpful discussions and J. Janjigian for her help in preparing the manuscript.

References

[1] J. Rutman and F. L. Walls, "Characterization of frequency stability in precision frequency sources," *Proc. IEEE*, vol. 79, this issue.

[2] D. J. Wineland, J. C. Bergquist, J. J. Bollinger, W. M. Itano, D. J. Heinzen, S. L. Gilbert, C. H. Manney, and M. G. Raizen, "Progress at NIST toward absolute frequency standards using stored ions," *IEEE Trans. Ultrason. Ferroelec. Freq. Contr.*, vol. 37, Nov. 1990.

[3] R. F. C. Vessot, E. M. Mattison, W. J. Klepczynski, G. M. R. Winkler, I. F. Silvera, H. P. Godfried, and R. L. Walsworth, "Present clock stability and realistic prospects for the future," in *Proc. Fourth Marcel Grossman Meeting on General Relativity*, R. Ruffini, Ed., Rome, Italy, June 17–21, 1985.

[4] L. A. Rawley, J. H. Taylor, M. M. Davis, D. W. Allan, "Millisecond pulsar PSR 1937 + 21: A highly stable clock," *Science*, vol. 238.

[5] D. R. Stinebring, M. F. Ryba, J. H. Taylor, and R. W. Romani, "Cosmic gravitational-wave background: limits from millisecond pulsar timing," *Phys. Rev. Lett.*, vol 65, no. 3, pp. 285–288, 1988.

[6] T. Damour, G. W. Gibbons, and J. H. Taylor, "Limits on the variability of G using binary-pulsar data," *Phys. Rev. Lett.*, vol. 61, no. 10, pp. 151–154, 1988.

[7] J. P. Turneaure and S. R. Stein, *Atomic Masses and Fundamental Constants*, J. H. Sanders and A. H. Wapstra, Eds. New York: Plenum, vol. 5, 1976, p. 636.

[8] S. Thakoor, D. M. Strayer, G. D. Dick, and J. Mercereau, "A lead-on-sapphire superconducting cavity of superior quality," *J. Appl. Phys.*, vol. 59, pp. 854–858, 1986.

[9] V. B. Braginsky, V. P. Mitrofanov, and V. I. Panov, *Systems with Small Dissipation*. Chicago, IL: University of Chicago Press, 1985.
[10] G. D. Dick and D. M. Strayer, "Measurements and analysis of cryogenic sapphire dielectric resonators and DRO's," in *Proc. of 41st Annual Freq. Control Symp.*, pp. 487–491, 1987.
[11] R. T. Wang and G. J. Dick, "Improved performance of the superconducting cavity maser at short measuring times," in *Proc. of 44th Symp. Frequency Control*, Baltimore, MD, May 23–25, 1990, pp. 89–93
[12] R. L. Walsworth, I. F. Silvera, H. P. Godfried, C. C. Agosta, R. F. C. Vessot, and E. M. Mattison, "A hydrogen maser at temperatures below 1K," *Phys. Rev A Rapid Communications*, vol. 34, p. 2550, 1896.
[13] J. D. Turneaure, C. M. Will, B. F. Farrell, E. M. Mattison, and R. F. C. Vessot, "Tests of the principle of equivalence by a null gravitational redshift experiment," *Phys. Rev. Lett. D*, vol. 27, no. 8, p. 1009, Apr. 1983.
[14] A. P. Lightman and D. L. Lee, "Restricted proof that the weak equivalence principle implies the Einstein equivalence principle," *Phys.Rev. D*, vol. 8, p. 364, 1973.
[15] R. S. Badessa, R. S. Kent, R. L. Nowell, J. C. and C. L. Searle, "A Doppler-cancellation technique for determining the altitude dependence of gravitational redshift in an earth satellite," *Proc. IRE*, vol. 48, p. 758, 1960.
[16] D. Kleppner, R. F. C. Vessot, and N. F. Ramsey, "An orbiting clock experiment to determine the gravitational redshift," *Astrophys. Space Sci.*, vol. 6, pp. 13–32.
[17] R. F. C. Vessot, M. W. Levine, E. M. Mattison, E. L. Blomberg, T. E. Hoffmann, G. U. Nystrom, B. F. Farrell, R. Decher, P. B. Eby, C. R. Baugher, J. W. Watts, D. L. Teuber, and F. D. Wills, "Tests of relativistic gravitation with a space-borne hydrogen maser," *Phys. Rev. Lett.*, vol. 45, pp. 2081–2084, Dec. 29, 1980.
[18] A. J. Tucker and B. M. Fannin, "Analysis of ionospheric contributions to the Doppler shift of C. W. signals from artificial earth satellites," *J. Geophys. Res., Space Physics*, vol. 73, no. 13, pp. 4325–4334, 1968.
[19] R. F. C. Vessot and M. W. Levine, "Performance data of space and ground hydrogen masers and ionospheric studies for high-accuracy frequency comparison between space and ground clocks," in *Proc. Twenty-eighth Annual Symp. Frequency Control*, U. S. Army Electronics Command, Ft. Monmouth, NJ, 1974, pp. 408–414.
[20] R. F. C. Vessot and M. W. Levine, *NASA Experimental Final Redshift Report*, GPA Project Report, Contract NAS8-27969.
[21] R. F. C. Vessot and M. W. Levine, "A time-correlated four-link Doppler tracking system," in *A Closeup of the Sun*, M. Neugebauer and R. W. Davies, Eds. JPL Publication 78–70, NASA, 1978.
[22] T. Piran, E. Reiter, W. G. Unruh, and R. F. C. Vessot, "Filtering of spacecraft Doppler tracking data and detection of gravitational radiation," *Phys. Rev. Lett. D*, vol. 34, no. 4, p. 984, 1986.
[23] L. L. Smarr. R. F. C. Vessot, C.A Lundquist, R. Decher, and T. Piran, "Gravitational waves and redshifts: A space experiment for testing relativistic gravity using multiple time-correlated radio signals," *General Relativity and Gravitation*, vol. 15, no. 2, pp. 129–163, 1983.
[24] M. Neugebauer and R. W. Davies, *A Closeup of the Sun*. JPL Publications 78–70, NASA, 1978.
[25] T. M. Brown, J. Christensen-Dalsgaar, W. A. Dziembowski, R. Goode, D. O. Gough, and C. A. Morrow, "Inferring the suns internal angular velocity from observed P-mode frequency splitting," *Astrophys. J.*, vol. 343, p. 526, 1989.
[26] F. B. Estabrook, "Gravitational wave searches with ground tracking networks," *Acta Astron.*, vol. 17, no. 5, pp. 585–587, 1988.
[27] J. W. Armstrong, J. D. Anderson, and E. L. Lau, "Application of hydrogen maser technology to the search for gravitational radiation," in *Proc. Twenty-first Annual Precise Time and Time Interval (PTTI) Applications and Planning Meeting*, Redondo Beach, CA, Nov. 28–30, 1989, pp. 259–263.
[28] J. H. Taylor and J. M. Weisberg, "Further experimental tests of relativistic gravity using the binary pulsar PSR 1913 + 16," *The Astrophys. J.*, vol. 345, pp. 434–450, Oct. 1989.
[29] A. J. Anderson, "Probability of long period (VLF) gravitational radiation," *Nature*, vol. 229, pp. 547–548, 1971.
[30] F. B. Estabrook and H. D. Wahlquist, "Response of Doppler spacecraft tracking to gravity waves," *Gen. Rel. Grav.*, vol. 6, pp. 439–447, 1975.
[31] K. S. Thorne and V. B. Braginsky, "Gravitational-wave bursts from the nuclei of distant galaxies and quasars: Proposal for detection using Doppler tracking of interplanetary spacecraft," *Astrophys. J. Lett.*, vol. 204, L1, 1976.
[32] A. E. E. Rogers, A. T. Moffet, D. C. Backer, and J. M. Moran, "Coherence limits in VLBI observations at 3-millimeter wavelength," *Radio Sci.*, vol. 19, no. 6, pp. 1552–1560, 1984.
[33] E. B. Formalont and R. A. Sramek, "Measurement of the solar gravitational deflection of radiowaves in agreement with general relativity," *Phys. Rev. Lett.* vol. 36, p. 1475, 1976.
[34] A. E. E. Rogers and J. M. Moran, "Interferometers and arrays," in *Methods of Experimental Physics*, Astrophysics, M. L. Weeks, Ed. New York: Academic, 1976, vol. 12, ch. 5.
[35] N. Bartel, V. Dhawan, T. Krichbaum, D.A Graham, I. I. K. Pauliny-Toth, A. E. E. Rogers, B. O. Rnn ng, J. H. Spencer, H. Hirabayashi, M. Inoue, C. R. Lawrence, I. I. Shapiro, B. F. Burke, J. M. Marcaide, K. J. Johnston, R. S. Booth, A. Witzel, M. Morimoto, and A. C. S. Readhead, "VLBI imaging with an angular resolution of 100 microarcseconds," *Nature*, vol. 334, no. 6178, pp. 131–135, 1988.
[36] G. S. Levy, C. S. Christensen, J. F. Jordan, R. A. Preston, C. D. Edwards, R. P. Linfield, S. J. DiNardo, L. Skjerve, J. S. Ulvestad, T. Hayashi, T. Nishimura. T. Taskano, T. Yamada, T. Shiomi, H. Kunimori, N. Kawaguchi, M. Inoue, M. Morimoto, H. Hirabayashi, B. F. Burke, A. Whitney, D. L. Jauncey, C. H. Ottenhoff, K. Blaney, and W. Peters, "Results and communications considerations of the very long baseline interferometry demonstration using the tracking and data relay satellite system," *Acta Astron.*, vol. 15, no. 6/7, pp. 481–487, 1986.
[37] V. V. Andrejanov *et al.* "Quasat, A VLBI observatory in space," in *Proc. Workshop* Gross Enzersdorf, Austria, June 18–221984, ESA Publication SP-213.
[38] R. F. C. Vessot, "Clocks and spaceborne tests of relativistic gravitation," R. D. Reasenberg and R. F. C. Vessot, Eds., *Advances in Space Research*. New York: Pergamon, vol. 9, 1989, pp. 21–28.
[39] E. J. Post, "Sagnac effect," *Rev. Mod. Phys.*, vol. 39, p. 475, 1967.

Robert F. C. Vessot received the Ph.D. degree in physics from McGill University, Montreal, Canada in 1956.

From 1951 to 1954 he served in the Royal Canadian Airforce as a Telecommunications Officer. From 1956 to 1960 he was a post-doctoral staff member at the Massachusetts Institute of Technology, Cambridge, MA, working of atomic clocks under the late Professor J. R. Zacharias. From 1960 to 1969 he was manager of the Hydrogen Maser Research and Development of Varian Associates/Hewlett Packard in Beverly, MA. He is presently Senior Physicist at the Harvard-Smithsonian Center for Astrophysics. He has been the P. I. of the Smithsonian Astrophysical Observatory, Hydrogen Maser Laboratory group since 1969. He has worked on the development of atomic hydrogen masers since 1961 and contributed to the success of the Hydrogen Maser as the most stable oscillator now available. The SAO Maser Group has built all the H-masers used in NASA's Deep Space Network and has supported radio astronomical VLBI activity worldwide. He has been the author and coauthor of more than 100 published articles, of which 33 are in refereed journals, and has made many contributions to books.

Dr. Vessot received NASA's medal for "Exceptional Scientific Achievement" for his work as Principal Investigator on the Gravitational Redshift Experiment (NASA's Gravity Probe-A). This test confirmed Einstein's predictions of the effects of relativistic gravitation on the rate of clocks at a precision of 70 parts per million.

Time and Frequency in Fundamental Metrology

BRIAN W. PETLEY

Invited Paper

The role of time and frequency in a wide range of measurements is discussed, particularly those involving the International System of Units (SI) and fundamental physical constants.

I. Introduction

Modern science and technology are placing increasingly stringent demands on our measurement system and associated units. The International System of Units (SI) [1] is almost universally used and other unit systems have generally been defined to have an exact relationship to the SI. The ultimate arbiters for the SI are the General Conference on Weights and Measures (CGPM), the International Committee on Weights and Measures (CIPM), and the Consultative Committee on Units (CCU). The SI embodies seven *base units*: those of time, length, mass, quantity of electricity, thermodynamic temperature, amount of substance, and luminous intensity. There is a second class, the *supplementary units* which provide units of angle and solid angle. These are regarded as dimensioned or dimensionless according to the choice of the user. Along with these there are now some nineteen named combinations of the base units which are known as *derived units*.

There are three distinct time units in the SI, these are:

1) the base unit, the *second* (13th CGPM, 1967) which is the unit of time and time interval;
2) the derived unit known as the *hertz* (11th CGPM, 1960), used to characterize the frequency of periodic phenomena;
3) the derived unit of activity known as the *becquerel* (15th CGPM, 1975) which is used for the stochastic phenomenon of radioactive decay related to the number of disintegrations per second.

Manuscript received November 13, 1990; revised February 26, 1991. This work is based on a talk presented at the 6th European Conference on Time and Frequency, Neuchâtel, Switzerland, March 13–15, 1990.

The author is with the Centre for Basic Metrology, Division of Quantum Metrology, National Physical Laboratory, Teddington, Middlesex TW11 OLW, U.K.

IEEE Log Number 9100246.

Attempts to metricate other time units (happily) failed, and units such as the *minute* (60 s), *hour* (3600 s), and day (86 400 s) were recognized by the CIPM in 1969 as being units used alongside the SI. The *week*, *month*, and *year*, remain in common use.

Although the *meter* and *kilogram* were carefully defined by the CGPM at their first meeting in 1889, the *second* does not appear to have been explicitly defined by them. Everyone "knew" that it was:

⟨⟨*equal to the fraction 1/ 86 400 of the mean solar day*⟩⟩

and left the astronomers to define precisely what was meant by the phrase "*mean solar day*." In 1956 the CIPM followed the decision of the 1955 General Assembly of the IAU, and decided that:

⟨⟨*the second is the fraction 1/31 556 925.9747 of the tropical year for 1900 January 0 at 12 hours ephemeris time*⟩⟩.

This somewhat inaccessible definition was adopted by the 11th CGPM in 1960. This definition only endured until 1967 when the 13th CGPM decided that metrology needed a better unit of time than astronomy could provide and, by its resolution 1, decided that

⟨⟨*The second is the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the caesium– 133 atom*⟩⟩.

At the same time the earlier definition was abrogated.

Metrology is the phrase used to describe the rapidly evolving art and science of accurate measurement. Viewed from a distance it is readily apparent that:

1) accuracy with which we are able to make measurements has been increasing over the centuries;
2) this process has been speeding up, the present rate corresponding roughly to a tenfold improvement every fifteen years.

0018–9219/91/0700–1070$01.00 © 1991 IEEE

All physical quantities show much the same effects, although the accuracy achieved for mass, length, and other physical quantities lags that of time measurement by more than thirty years.

The accuracy with which we may measure dimensioned physical quantities depends ultimately on the accuracy with which the base units are known. At present these range from a fractional accuracy in the region of 10^{-14} for time and frequency measurements, through to the 10^{-4} to 10^{-3} accuracy of luminous intensity measurements. We should not, however, focus on the accuracy of measurement of the unit alone—for most users are almost certainly working well away from the unit. For almost every physical quantity we get a characteristic bath-tub shaped curve (Fig. 1) as we lose accuracy in going to multiples and submultiples of the unit. The well known Allan-variance curve, which is used to characterize the time-dependence of the stability of time and frequency standards and also of lasers (although requiring a slightly different interpretation), is but one example of this type of curve. Mostly, the unit occupies a position somewhere along the flat horizontal portion, which would correspond to the outlet of the bathtub. In the case of time measurements we need to be careful to distinguish between the frequency, or time, that is being measured and the time interval associated with the measurement (which may, for example, be the time interval during which an atom experiences the frequency). It takes rather longer than a second to realize the *second.*

The effect of the overflow to the atomic constants (shown in Fig. 1) is very striking, for we can leap over many orders of magnitude without further loss of accuracy. The result is that today we know the charge on the electron or proton as well as we know the coulomb, or the mass of the electron as well as we know the gram, or the Planck constant as well as we can realize the joule second.

II. Time And Length Measurements: Fringe Counting And Fringe Velocity

The definition of the meter adopted by the 17th CGPM in 1983 reads:

> ⟨⟨*The meter is the length of the path traveled by light in vacuum during a time interval of 1/299 792 458 of a second*⟩⟩.

This invokes the relationship:

$$l = c \cdot t$$

so that our definition of length requires the prior definition of the second. The length, l, of a resonator may be derived from its resonant frequency, f, via $f = c/2l$. This is a useful noncontact method that has been applied by Gallop at the NPL[2] to the measurement of superconducting cavities at 4.2 K with an accuracy of around 10 nm on a 50-mm diameter cylinder (2 parts in 10^7). An increasing number of such relationships are being applied at the highest levels of accuracy, particularly in the visible region where lasers may be slaved to remain tuned to a resonant mode of a cavity.

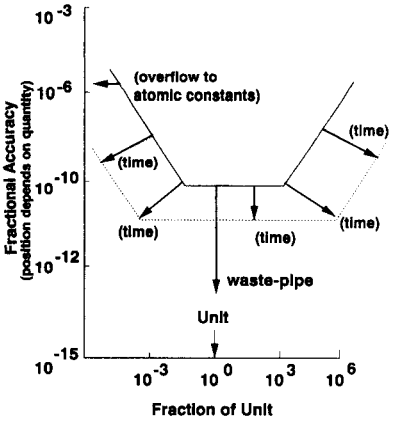


Fig. 1. The metrological bathtub: the reduction in accuracy as one moves away from the unit. The "overflow" pipe effectively takes one across to the fundamental physical constants without further loss of accuracy: the electron mass is known as well as we know the gram, etc. The whole of the bathtub must be moved downwards and outwards as time progresses.

The full implementation of the present definition of the meter awaits the arrival of accurate methods of frequency multiplication and measurement through to the visible region of the spectrum and beyond. At present one measures electromagnetic radiation in the visible region in terms of wavelength rather than frequency—by using cavity wavemeters rather than the digital techniques that are at present available to the submillimeter wavelength region. These limit the present accuracies to between a few parts in 10^{10} or 10^{11} rather than the atomic clock level of accuracy that the meter definition implies.

The measurement of the acceleration due to gravity g, poses an interesting problem for, if we have a cube corner reflector that is part of a Michelson interferometer in free fall, the laser interference fringes have an instantaneous frequency $2gt/\lambda$. That is, the quantity of interest is the rate of change of frequency:

$$g = (\lambda/2) \cdot \frac{df}{dt}.$$

Up to the present the time for the cube to traverse a set number of fringes has been measured instead (the fringe frequencies usually range up to a few MHz). In this way Faller and colleagues at JILA [3], and others elsewhere, have been able to measure the acceleration due to gravity to a few parts in 10^9.

The lunar ranging experiments are an example of very precise length measurement made in terms of a time measurement. The Earth–Moon distance can now be monitored at the centimeter level, corresponding to a resolution of about 2 parts in 10^{11}. Similarly, the orbits of satellites may be tightly mapped by timing laser pulses to provide much useful information about the higher order terms for the earth's gravitational field. Timing accuracies of 0.1 ns or better may be achieved in space experiments.

Gravity wave detectors require the detection of very small strains which are manifested in the form of length changes.

Detectors having strain sensitivities of between 10^{-21} to 10^{-18} are being constructed in several locations. Some of these map length changes into frequency changes and therefore make strong demands on the short-term stability of the frequency discriminator. They require frequency discrimination and sensitivity rather than absolute accuracy.

III. Electrical And Other Measurements And The Radian Per Second

The radian per second is the unit of angular frequency and, although it often occurs in theoretical discussions, there is little necessity to produce and measure exact angular frequencies in radian per second (except as below). Many users prefer to express measurements in terms of alternative units such as degrees per second.

An example of the measurement of small angular displacement rates would be in gyroscopes. The sensitivities of gyroscopes are normally expressed in °/h rather than rad/s, the conversion factor being approximately $1°/h = 50.10^{-6}$ rad/s. Sensitivities of better than 10^{-4} °/h are required for ship navigation, and 10^{-2} °/h in aircraft. For the latter this corresponds to less than one nautical mile of drift error per hour.

The optical gyroscopes [4] make use of the Sagnac (1913) effect to detect small angular velocities. The time difference Δt in an area A and angular velocity Ω is given by:

$$\Delta t = 4A\Omega/c^2 .$$

Michelson and Gale demonstrated this effect in 1925 by using an interferometer to detect the Earth's rotation rate of 15 °/h. For $A = 1$ m^2, and $\Omega = 15$ °/h, $\Delta L(= c\Delta t)$ is about 10^{-12} m.

Fiber optic systems require phase detection sensitivities corresponding to about $10^{-8}\lambda$ in a $10^{9}\lambda$ path length in order to achieve sensitivities of 0.050 °/h. The very best suspended electrical gyroscopes achieve around 10^{-5} °/h, although the present-day active laser, passive laser, and fiber optic ring-gyroscopes have around a 100 times poorer sensitivity they are capable of further refinement. In laser ring gyroscopes, the counter-rotating beat frequencies of < 1 mHz have been measured, enabling long term stabilities of 10^{-3} °/h to be achieved. Performance in some cases is approaching the quantum noise limit of 10^{-4} °/(h)$^{1/2}$.

The transfer across from resistance involves the equations:

$$R = \omega L \text{ and } R = 1/\omega C.$$

Obviously, if C, L, and R for electrical standards are all to be nearly exact multiples of the unit, then ω must also be an integral number of radians per second. Consequently, a rational frequency in rad/s rather than hertz is used experimentally to help make the transfer from resistance to and from capacitance and inductance standards. An angular frequency of 10^4 radians per second is convenient for this. One often finds that specifications for ac electrical measurements require that the measurements are made at, or within 1 Hz, of 1592 Hz. This frequency is regarded as being sufficiently close to the desired angular frequency of 10^4 radian per second. No attempt is made to synthesize the exact angular frequency to much better than this.

A. The Realization of the Volt, Ampere, and Ohm

The accuracy with which the electrical units could be realized in the past was limited by the requirement to measure the linear dimensions of coils of wire very accurately. For many years the ohm, for example, was realized via a precisely constructed mutual inductor until the advent of the Thompson and Lampard calculable capacitor meant that the farad could be realized instead. The capacitance so realized was $(10^7/[4\pi c^2]) \cdot \ln 2$ F/m (about 1.953 pF/m) and was obtained by moving a screening electrode whose displacement was readily measured by counting laser fringes. (Frequency enters the measurement via the 1592-Hz source energizing the ac bridges).

In this way the realizations of the ohm have been improved to yield fractional accuracies of the order of a few parts in 10^8. It ought to be possible to improve the accuracy further, particularly by measuring the length change as a change in frequency, for example, by changing the length of a resonant cavity. The advent of the quantized Hall resistance for MOSFET semiconductors, operated at low temperatures and strong magnetic fields, has enabled the drift of resistance standards to be confirmed as being of the order of 0.05 $\mu\Omega$ per year—the present ability to detect such small resistance drifts and allow for them promises well for the future.

Until recently the volt was realized by some form of attracted electrode electrometer (the most accurate of these being the attracted mercury surface determination of Clothier, Sloggett, and others at the NML[5] in Australia), while the ampere was realized via some traditional form of current balance. These are likely to be replaced by the realization of the watt.

B. The Watt

Kibble at the NPL proposed [6] the method which forms the basis for recent electrical realizations of the watt. Essentially, a coil of wire is suspended from one arm of a mass balance such that it is partially in the field provided by a magnet (both superconducting magnets and permanent magnets are being used).

A current I, whose value is known in terms of the laboratory representation of the ampere, is passed through the coil and the force on the coil balanced by adding or removing mass m, from the scale-pan. The coil is then displaced vertically with a uniform velocity v, and the induced emf E, measured in terms of the laboratory as maintained representation of the volt. The product of the voltage and the current are then given by:

$$E \cdot i = mgv$$

where g is the acceleration due to gravity. Thus the power is known in terms of both the laboratory maintained representation of the watt, K_W watt, and the mechanical watt.

The coil velocity is measured as a laser fringe frequency of about 38 kHz.

The first attempt by Kibble and others at NPL has yielded a fractional accuracy of about a part in 10^7 and one may anticipate a further order of magnitude improvement. At some point it will be possible to use the method to check the stability of the kilogram. This may lead eventually a replacement for the prototype kilogram, which is last of the nineteenth century prototype units. It is too soon to say whether the new definition of the kilogram would be in terms of an atomic mass, or the Planck constant, or the Avogadro constant, or even whether the kilogram might become a derived unit.

IV. Frequency Metrology and the Fundamental Physical Constants

The involvement of the fundamental physical constants with frequency forms the basis for the use of the phrase: *quantum metrology* [7]. This is particularly appropriate in the case of the transition frequency ν between two energy levels E_1 and E_2 such that

$$E_1 - E_2 = h\nu.$$

This is the equation of elementary quantum mechanics which follows from the equation of motion of a wave function ψ :

$$-i\hbar\frac{d\psi}{dt} = H\psi$$

given that E_1 and E_2 are eigenvalues of the Hamiltonian H. The present definition of the second means that the time unit in the wave equation is necessarily of the same character as that for the frequency—which serves to illustrate that the definition of the second has a profound effect on the equations of motion in physics, both at the macroscopic level and at the quantum level.

Other relationships involving frequency are also used in metrology, such as that in synchrotron radiation for luminous intensity measurements (which are at the 0.05% level), or the speed of sound at zero pressure v_o, which, in a gas of molar mass M, at a temperature T, is given by:

$$v_o^2 = \gamma RT/M$$

where γ is the ratio of the specific heats, which is 5/2 for argon. This equation is used in acoustic thermometry and also for measuring the gas constant, R, (and hence the Boltzmann constant k) to a few parts per million. The velocity is obtained from the resonant frequency of a cavity. The accuracy is limited by the need to extrapolate to zero pressures to remove the effects of the virial coefficients. Uncertainties in these and the temperature distribution tend to set the present accuracy limits.

It is often found in metrology that increased accuracy and the increased frequency associated with the measurement go together. One can see this feature particularly with measurements of the fundamental physical constants in Fig. 2. Each time that the fundamental constants are evaluated their

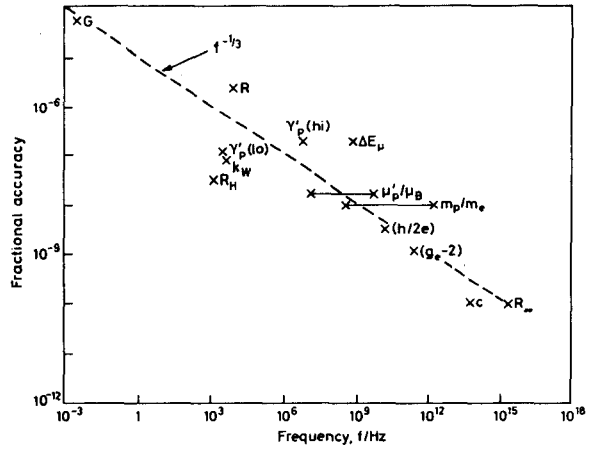


Fig. 2. The accuracy of measurements of the fundamental physical constants which contributed to the 1987 evaluation, plotted as a function of the associated frequency. It is seen that for measurements over a broad range of physics (atomic, electrical, gravitational, optical, thermal, etc.), the accuracy achieved improves with frequency, as $f^{0.3}$.

accuracy improves [8], reflecting the increased involvement of frequency—the fitted line will move downwards and sideways in the future. Increased accuracy also comes through increasing the measurement time. This enables noise and other effects to be averaged, the accuracy often improving as (time)$^{1/2}$. In metrology the observation time is usually limited by other considerations, particularly if the measurements are part of a control process.

A. Magnetometry, Gyromagnetic Ratios, and Cyclotron Frequency

The relationships between the proton spin precession frequency and flux density, and between the electron cyclotron frequency and flux density, involve the equations:

$$\omega_s = \gamma_p' \cdot B$$

and

$$\omega_c = Be/m_e$$

respectively, where the γ_p' is the gyromagnetic ratio of the proton, e is the elementary charge and m_p the rest mass of the proton. The use of the prime symbol indicates that γ_p' is not exactly $2\mu_p/\hbar$, for the spin precession frequency refers to that of protons in a spherical sample of water. There are small temperature and sample shape dependent frequency shifts which amount to some 25.689(15) parts per million.

The gyromagnetic ratio of the proton has been used for some forty years as a transfer constant to enable magnetic flux densities to be measured in terms of frequency by nuclear magnetic resonance (NMR), with an accuracy limited by the accuracy of the realization of the ampere. Other gyromagnetic ratios are known with respect to water and rubidium and ^{3}He and other types of magnetometers are used to measure a wide range of magnetic field strengths in terms of frequency. It is convenient practically to measure

and express the frequencies in hertz and their values are usually tabulated in both hertz and radians per second.

The accuracy of flux density measurements by this method is partly limited by the accuracy of the realization of the volt. The water sample shape, its temperature, the effects of the container, and impurities in the water (dissolved oxygen in particular), limit the accuracy to about a part in 10^8—much depends on the nature of the substitution method involved.

Although not requiring accurate frequency measurement, SQUID magnetometers operate at carrier frequencies up to the microwave region and enable magnetic flux to be measured in terms of the magnetic flux quantum $h/2e$ (around 10^{-15} Wb). These are relative devices since it is difficult to measure the area through which the flux is passing.

B. Ratio Experiments and Mass Spectrometry

There are some important measurements which involve the ratios of ion cyclotron frequencies and/or spin precession frequencies. These include the $g-2$ measurements for the electron, positron and muon, and the ratio of the proton to electron mass. The accuracy of such measurements made a major leap forward through the pioneering work on ion traps by Dehmelt, Van Dyck, and others at the University of Washington and elsewhere. Some of the improvement can also be attributed to the use of higher flux densities, with correspondingly higher frequency discrimination.

Conventional mass spectrometry achieved its 10^{-9} fractional accuracy by setting very stringent geometric requirements on the ion orbit, by using beam defining slits, etc., so that the arrival of the ion at the detector depended critically on its starting phase with respect to the applied cyclotron frequency. Thus in the ion substitution mass synchrometers of Lincoln Smith in Princeton and of Mamyrin in Leningrad, the ions made only two revolutions, taking around a microsecond, before being detected. The timing discrimination (or the equivalent phase discrimination) was effectively of the order of 10^{-14} s for a fractional resolution of a part in 10^8. Ion substitution was achieved by changing both the applied ac and dc potentials, and ac frequencies. Ion traps are beginning to allow the properties of antimatter to be explored and no doubt the findings of these studies will impact on metrology.

C. Josephson Effects

The prediction in 1962 of the Josephson tunneling effects in superconductivity, and their subsequent validation at the part per million level at the University of Pennsylvania, the NPL, and elsewhere by the end of 1969, meant that there was a very precise relation between the potential difference between two closely separated superconductors in the presence of a microwave frequency f given by:

$$2eV = hf$$

where n was an integer, h the Planck constant and e the elementary charge. This equation has been verified in various ways to the very highest levels of precision, including independence of the operating conditions (temperature, material, frequency), and by now must be one of the most accurately verified effects in physics.

Meanwhile, one of the most accurate measurements involving frequency to date has been made via a null measurement by Lukens and others at Stony Brook [9], which involved verifying that the change in the gravitational potential of a microwave photon in traversing a vertical distance of about a centimeter was the same as the change in potential of the electron pair wave. The discrimination achieved was of the order of a part in 10^{18}; needless to say the interpretation of such a measurement is slightly open to debate! We have arrived at the situation in electrical metrology where the volt is maintained in terms of agreed values of the Josephson effect K_J, and the ohm in terms of the quantized Hall resistance von Klitzing constant R_K.[10]

D. Can We Measure h Directly In Terms of the Mechanical Units?

The quantum equation relating energy to frequency:

$$E = h\nu$$

occurs so frequently in modern physics that one wonders whether it can be verified directly in terms of the joule (regarded as a mechanical unit). The equation can and has, of course, been verified for cases where the energy involved is primarily an electrical energy, most frequently as an energy in electron-volts.

If the voltage E_J across a Josephson junction (which is biased onto the nth voltage step when irradiated with a microwave frequency f) is maintained at a value equal to the voltage across a von Klitzing device which is biased onto the second quantized Hall resistance step, then the power in the Hall resistor is:

$$E_J^2/R_H = n^2 f^2 h/2 \, .$$

Today the volt is maintained via the Josephson effect K_J and the ohm via the quantized Hall resistance R_K, and the numerical relationships between these and the units are exactly specified [10]. Consequently, we see that the above realization of the watt by Kibble and others limits our knowledge of h, always provided, of course, that the Josephson and von Klitzing relationships are exact.

V. Metrology And Time Variations Of The Fundamental Constants

Following their interpretation in terms of Dirac's "big numbers" some 55 years ago, there is always considerable interest in the question of whether or not the fundamental constants are constant. The evidence to date is that they are constant on the level of less than one percent over the age of the universe. Some measurements have put even tighter constraints. In each case one must interpret the results with care. This is particularly so for any dimensioned quantity since one cannot say whether it is the unit or the quantity that has varied (equally a null effect may result if they

both vary by the equivalent amount). Turneaure and Stein at Stanford [11] were able to compare the frequency of a caesium clock against that of a high-Q superconducting microwave cavity and make tight inferences about the constancy of the fine structure constant in conjunction with other dimensionless constants. Thus the caesium frequency is represented by:

$$f_{Cs,hfs} = \text{const.} g_p (eh/2m_p)(eh/2m_e)(Za_o)^3$$

where a_o is the Bohr radius, Z the caesium nuclear charge and g_p the proton g-factor. The cavity frequency depends on

$$f_{cavity} = \text{const.} hc/a_o$$

so that

$$f_{cavity}/f_c = \text{const.} g_p \left(\frac{m_e}{m_p}\right)\alpha^3$$

from which Turneaure and Stein obtained:

$$1/\alpha \cdot d\alpha/dt < 3 \cdot 10^{-11} \text{ per year.}$$

The particular significance of their work was that they were able to use laboratory measurements made over a period of 12 days to set constraints on parameters that were expected expected to be constant over cosmological time-scales. That is they were able to set limits on the (almost) instantaneous rate of change of a cosmological quantity. Their work, along with studies of the isotropy of inertia, of Michelson-Morley and Kennedy-Thorndyke experiments, and measurements of the blackbody background radiation, serves to emphasize that cosmology is important in modern metrology and *vice versa*.

There is still some debate about what happens to the fundamental constants at the high energies required by the Grand Unified Theories (GUT) and the "big bang," but in order to discuss such possibilities one needs something else that is more stable to use as an invariant "ruler!" Today, more than ever before, we need to remember that all of our units are inside the universe and affected by events inside it.

VI. Can We Replace The Present Definition Of The Second?

During the next decade or so there is likely to be an extensive exploration of a number of very narrow spectroscopic transitions with the aid of ion traps and atom traps. Many of these transitions will lie well beyond the microwave region. Some may well have Q-values which are 10^8 or more times greater than that of the present generation of caesium beam and hydrogen maser frequency standards. These narrow transitions will require interrogation sources of high spectral purity, and the required millihertz linewidths are being increasingly achieved for the optical (and near optical) laser frequencies. There may well be several transitions of comparable accuracy which are much superior to the accuracies achieved with the next generation of caesium clocks. The advent of these will cause metrologists to consider very carefully what should be done about the present definition of the second.

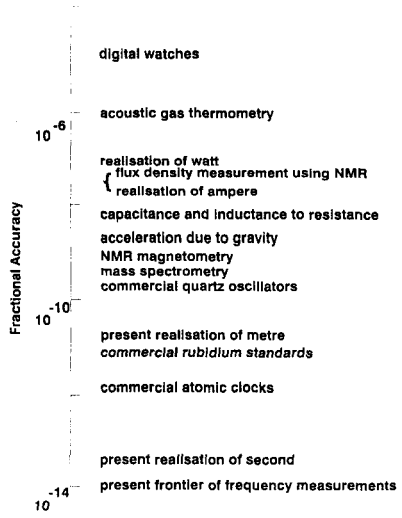


Fig. 3. The present accuracies of some applications of time and frequency metrology.

There are some simple requirements for any new definition of the second, or indeed for any other unit:

1) for continuity the new unit must always lie within the reproducibility of the old;
2) it must be capable of being realized and accessed by users;
3) it should immediately permit an order of magnitude improvement in the accuracy of the realized unit; and
4) it should be capable of further refinement in accuracy.

The most desirable solution would be a definition which was not specific to a particular transition in a particular ion or atom. Metrologists have already adopted this solution for the meter (by invoking a fundamental physical constant rather than a particular laser transition) which thereby avoided our having to redefine the meter each time that a higher performance stabilized laser was developed. There are a number of "natural" candidates for an equivalent definition of our unit of time, ranging from the Planck time, via the more practical, to the age of the universe. There is nothing particularly fundamental about using the caesium-133 atom as a basis for defining the second, but it is up to the experimental metrolgist to demonstrate a practical alternative!

VII. Conclusion

The applications of time and frequency measurements in metrology are wide ranging as shown in Fig. 3, and the accuracy with which we can realize the second continues to run well ahead of the rest of the SI units—a feature that will continue be exploited. It will probably become very important in the future to distinguish between a *time standard* and a possible *frequency standard*, for the latter

need not operate continuously! As we have seen, there are an increasing number of quantum effects which may be produced sufficiently cleanly in the laboratory to enable them to be used for purposes of defining or maintaining the SI units. We can expect in a general way that this process is likely to continue and that the definitions of most of the other base units will rely on time in some way.

As the accuracy of measurement advances it will be necessary to take general and special relativity more into account than before in metrology and elsewhere. That is, it will be more important to take care to use coordinate time, internal time, proper time, and so on, as is appropriate to the particular circumstances of the measurement. The studies of the stability of the millisecond Pulsars (which provide a longer term "flywheel" against which we may look for unexpected secular variations in our time standards) serve as a strong reminder that the whole of our terrestrial physics is modulated by changes in our gravitational potential—such as result from the rotation of the Earth and Moon about one another and about the Sun. Hopefully we will eventually be able to detect such effects solely by laboratory based measurements without recourse to Pulsars. We may expect to have to consider in the future such questions as that of the granularity of time (when measuring the highest frequencies in the X-ray region and beyond), or the reasons for the direction of "*the arrow of time*," and their implications for quantum physics and metrology.

Progress must slow eventually, but there are probably some seven or more orders of magnitude to go before we reach the ultimate accuracy limits for frequency measurement. These may well be reached when the averaging time for the ultimate transition is about the working life of a metrologist! At the present rate of progress that condition may well be reached well before the end of the next century, meanwhile our endeavors to attain such accuracies will provide a lot of technological applications and knowledge, and also provide plenty of fun!

REFERENCES

[1] *Le Systeme International d'Unites (SI)*, 5th ed., Paris: BIPM, 1985.

[2] J. C. Gallop and W. J. Radcliffe, "Shape and dimensional measurement using microwaves," *J. Phys. E: Sci. Instrum.*, vol. 43, pp. 413–416, 1986.

[3] I. Marson and J. E. Faller, "The acceleration of gravity: its measurement and its importance," *J. Phys. E: Sci. Instrum.*, vol. 19, pp. 22–32, 1988.

[4] J. Faucheux, D. Fayoux, and J. J. Roland, "The ring laser gyro," *J. Opt.*, vol. 19, pp. 101–115, 1988.

[5] W. K. Clothier, G. J. Sloggett, H. Bairnsfather, M. F. Currey, and D. J. Benjamin, "A determination of the volt," *Metrologia*, vol. 26, pp. 9–46, 1989.

[6] B. P. Kibble, I. A. Robinson, and J. H. Belliss, "A realization of the SI watt by the NPL moving-coil balance," NPL Report DES 88, 1988.

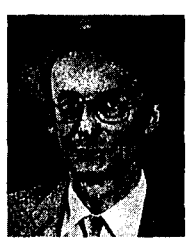
[7] B. W. Petley, *The Fundamental Constants and the Frontier of Measurement*. Adam Hilger: London, U.K., 1988.

[8] E. R. Cohen and B. N. Taylor, "The 1986 evaluation of the fundamental physical constants," *Rev. Mod. Phys.*, vol. 57, pp. 1121–1148, 1988.

[9] A. K. Jain, J. E. Lukins, and J.-S. Tsai, "Test for relativistic gravitational effects on charged particles," *Phys. Rev. Lett.*, vol. 58, pp. 1165–1168, 1987.

[10] B. N. Taylor and T. J. Witt, New international electrical reference standards based on the Josephson and Quantum Hall Effects," *Metrologia*, vol. 26, pp. 47–62, 1989.

[11] J. P. Turneaure and S. R. Stein, "An experimental test of the time variation of the fine-structure constant," in *Atomic Masses and Fundamental Constants 5*, J. H. Sanders and A. H. Wapstra, Eds. London, U.K.: Plenum, 1976, pp. 636–642.



Brian W Petley was born in Dover, England and obtained the Ph.D. degree from Imperial College for work in high energy nuclear physics.

Following a period at the Electrical Research Association he moved to the National Physical Laboratory, where his active research interests and publications include topics in basic metrology, dielectrics, the Josephson effects and SQUID's, laser spectroscopy, ^{3}He magnetometry, atomic hydrogen beam work, and the measurement of fundamental physical constants. He heads the Centre for Basic Metrology at the NPL, is a member of the CODATA Task Group on Fundamental Physical Constants, and is Chairman of the IUPAP Commission C2, the Commission on Symbols, Units, Nomenclature, Atomic Masses, and Fundamental Physical Constants. He has also written a book on *The Fundamental Physical Constants and the Frontier of Measurement* and another on *An Introduction to the Josephson Effects*

Dr. Petley is a member of the Editorial Boards of Measurement Science and Technology, and Metrologia, and a Member of the Editorial Review Committee of IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT.