Following are 15 white papers written by Dr. Robert Nelson. Bob Nelson taught this course for many years prior to his passing in 2013 was regarded an expert in the field of satellites and satellite communications. Chris DeBoy now teaches Satellite Communications Design & Technology and continues the same level of knowledge and teaching dedication.

Via Satellite, July, 1998

A Primer on Satellite Communications

by Robert A. Nelson

In 1945 Arthur C. Clarke wrote an article entitled "The Future of World Communications" for the magazine Wireless World. This article, which the editors renamed "Extra-Terrestrial Relays", was published in the October issue. In it Clarke described the properties of the geostationary orbit, a circular orbit in the equatorial plane of the earth such that a satellite appears to hover over a fixed point on the equator. The period of revolution is equal to the period of rotation of the earth with respect to the stars, or 23 hours 56 minutes 4.1 seconds, and thus by Kepler's third law the orbital radius is 42,164 km. Taking into account the radius of the earth, the height of a satellite above the equator is 35,786 km. Clarke observed that only three satellites would be required to provide communications over the inhabited earth.

As a primary application of such a satellite system, Clarke proposed that satellites in geostationary orbit might provide direct broadcast television service similar to DBS systems like DirecTV -- a remarkable idea at a time when television was still in its infancy and it was not yet known whether radio signals could penetrate the ionosphere. He worked out a simple link budget, assuming a downlink frequency of 3 GHz, and estimated that the required transmitter output power for broadcast service to small parabolic antenna receivers would be about 50 watts. Electric power would be provided by steam generators heated by solar mirrors, but advances in technology might make it possible to replace them by arrays of photoelectric cells. Batteries would be used to provide uninterrupted service during eclipses, which occur in two seasons centered about the equinoxes. Clarke also estimated the mass ratio of a multistage launch vehicle necessary to deploy the satellite. However, he imagined the geostationary satellites to be outposts inhabited by astronauts to whom supplies would be ferried up on a regular basis, much like the Mir space station and the international space station now under construction.

Twenty years later, in his book *Voices from the Sky*, Clarke wrote a chapter entitled "A Short Pre-History of Comsats, Or: How I Lost a Billion Dollars in My Spare Time". For he did not patent the idea of a geostationary orbit and, believe it or not, orbits can and have been patented. (Recall the recent patent controversy between Odyssey and ICO.) However, despite the tongue-in-cheek subtitle, the famous author would not have profited from his idea for two reasons. First, arguably, prior art existed in the literature. In 1929 the Austrian engineer H. Noordwig observed that a satellite at an altitude of 35,786 km in the equatorial plane would appear motionless when viewed from earth (as cited by Bruno Pattan in *Satellite Systems: Principles and Technologies*). Second, had Clarke obtained a patent in 1945, it would have expired in 1962, 17 years after the concept was first disclosed and two years before the first geostationary satellite, *Syncom III*, was successfully launched. Nevertheless, Clarke can rightfully claim credit for the first detailed technical exposition of satellite communications with specific reference to the geostationary orbit. His vision was realized through the pioneering efforts of such scientists as John Pierce of the Bell Telephone Laboratories, head of the Telstar program and co-inventor of the traveling wave tube amplifier, and Harold Rosen of the Hughes Aircraft Company, who was the driving force behind the *Syncom* program.

Since 1964, approximately 265 satellites have been launched into geostationary orbit, of which approximately 185 are operational. Another 67 GEO satellites are presently on order. The majority of these satellites have been used for the traditional fixed satellite service in C- and Ku-band, but also include satellites in the direct broadcast service, digital audio radio service, and mobile satellite service. In addition, numerous nongeostationary systems are in the process of deployment or have been proposed for a variety of consumer services, including mobile telephony, data gathering and messaging, and broadband applications. In May, 1997, 73 new GEO satellites were licensed for broadband services at Ka-band and last September applications for a dozen more systems were submitted to the Federal Communications Commission (FCC) for geostationary, nongeostationary, and hybrid satellite systems to provide broadband services at V-band. The total number of planned new satellites exceeds 1300.

The design of a satellite communications system presents many interesting alternatives and tradeoffs. The characteristics include the choice of orbit, the method of multiple access, the methods of modulation and coding, and the tradeoff between power and bandwidth. In this article, these choices will be briefly described and hopefully a sense of why satellite engineers find this field of endeavor so fascinating will be conveyed.

ORBIT

The system design begins with the choice of orbit. The orbital altitude regimes have been conveniently classified as Low Earth Orbit (LEO), Medium Earth Orbit (MEO), and geostationary orbit (GEO). The altitude of LEO is about 1000 km, or above the atmosphere but below the first Van Allen radiation belt. The altitude of MEO is ten times greater, that is 10,000 km, which lies between the first and second Van Allen belts. The altitude of GEO is uniquely 35,786 km as stated above. A fourth category is High Earth Orbit (HEO), which is at about 20,000 km and is above the second Van Allen belt but below GEO. (The acronym HEO has also been used to mean "highly elliptical orbit"; can we find a new term for this category? The progression LEO, MEO, HEO, GEO is quite appealing.)

Besides altitude, two other important orbital parameters are inclination and eccentricity. The inclination may be selected on the basis of maximizing the

level of multiple satellite coverage. Elliptical orbits may be used with eccentricities designed to maximize the dwell time over a particular region.

The appropriate orbit is often suggested by the nature of the service, the business plan, or the constraints of the communications link. These properties are well illustrated by the variety of satellite mobile telephony systems under construction. Iridium is designed for continuous global coverage. This is a LEO constellation of 66 satellites in polar orbits at an altitude of 780 km. The choice of LEO was dictated by the desire to minimize power in both the satellite and the mobile handset, minimize the satellite antenna size, minimize the time delay, or latency, for a two-way signal, and maximize the angle of elevation. The orbital period is 100 minutes and a given satellite is in view for only ten minutes before handover of a call to a following satellite. An Iridium satellite has extensive onboard processing and a telephone call is routed through the constellation via intersatellite links.

Globalstar employs a constellation of 48 satellites in orbits inclined at 52° at an altitude of 1406 km. This system concentrates coverage over the temperate regions of the earth from 70° S to 70° N latitude. A technique called spatial diversity is used, wherein signals received simultaneously from two satellites are combined in the receiver to mitigate losses due to blockage and multipath effects. Thus an inclined, nonpolar orbit constellation was chosen to ensure that at least two satellites are visible at all times. The Globalstar system uses nonprocessing, or “bent pipe” satellites.

The third major mobile telephony satellite entry is ICO. This system will consist of 10 operational satellites in MEO at an altitude of 10,355 km. (The acronym ICO derives from the term "intermediate circular orbit", a synonym for MEO.) MEO is an excellent compromise between LEO and GEO. The satellite antenna size and power are relatively modest and the latency is still small. Yet the number of satellites required for global coverage is significantly less than LEO and the dwell time is considerably longer. The ICO orbit has a period of revolution of 6 hours and the time a satellite is in view is on the order of two hours.

Other satellite mobile telephony systems include ECCO and Ellipso. ECCO is a circular orbit constellation in the equatorial plane designed for communications in tropical regions. Ellipso employs elliptical orbits to maximize coverage over the northern hemisphere.

There is, nevertheless, a valid geostationary alternative for a mobile telephony satellite. The primary advantage is that the system can be built up on a regional basis. With only one satellite, an entire country or geographical region can be served. Although the two-way time delay can be over a half second and is quite perceptible, this is a defect that a population may be willing to accept if it is underserved by a terrestrial telephony system. An example is the Asia Cellular Satellite system (Aces) that is being built by Lockheed Martin for service to the Pacific Rim. To provide the required cellular coverage, the satellite antennas are about 12 meters across.

MULTIPLE ACCESS

Multiple access refers to the method by which many users share a common satellite resource. There are three primary methods: Frequency Division Multiple Access (FDMA), Time Division Multiple Access (TDMA), and Code Division Multiple Access (CDMA).

With FDMA the available spectrum is divided among all of the users. Each user obtains a dedicated portion of the spectrum. FDMA can be used for either analog or digital signals.

With TDMA each user is assigned a time slot in a repetitive time frame. Data bits are stored in a buffer and are burst to the satellite during the assigned time slot. The signal occupies the entire transponder bandwidth. Because bits are stored during the portion of the time frame not assigned to the user, TDMA is inherently digital.

CDMA is a method in which the signal to be transmitted is modulated by a pseudorandom noise (PRN) code. The code rate is usually several orders of magnitude greater than the information bit rate. Their ratio is called the processing gain. The code spreads the signal over the full bandwidth available (hence CDMA is also called "spread spectrum") and all users share the same spectrum. The receiver modulates the signals from all users simultaneously with a replica PRN code. The desired signal is obtained by autocorrelation, while all of the undesired signals are spread over the full bandwidth and appear as white noise.

Frequency Division Multiple Access is relatively simple both conceptually and in terms of the hardware required. However, a transponder is a nonlinear device. This means that the output power is not merely proportional to the input power, but rather is represented by a curve that can be approximated by a third order polynomial. For multiple carriers, this nonlinearity generates harmonics that produce intermodulation interference among neighboring channels. In order to mitigate this effect, the input power is reduced in order to operate in the linear portion of the transponder output vs. input power characteristic so that intermodulation is reduced to an acceptable level.

The reduction in power is called "backoff". At a typical backoff of 6 dB, the input power is only one fourth the maximum possible value at saturation and the output power is correspondingly less. Backoff is not required when only one carrier occupies the transponder, such as a typical FM video channel, a TDMA carrier, or several channels multiplexed onto a single carrier at the earth station.

A major advantage of TDMA is that backoff is not required, since at any given time a single user occupies the full bandwidth of the transponder. Thus the output power of the transponder is much higher than with FDMA. Another advantage of TDMA is that it is more flexible. User allocations can be changed with relatively simple changes to software.

CDMA offers the potential of greater capacity. However, the theory of CDMA assumes that all users appear to contribute equally to the overall noise. Because users are at different distances with respect to one another, this assumption implies the need for dynamic power control. Another advantage is that the population of users

need not be known in advance. As users are added to the system, the signal quality degrades slowly. Other advantages are that CDMA mitigates interference and enhances data security.

The mobile telephony satellite systems illustrate these alternatives. Both Iridium and ICO use a combination of FDMA and TDMA. The available spectrum is divided into sub-bands and TDMA is used within each sub-band. The capacity per satellite for Iridium is approximately 1100 simultaneous users, whereas each ICO satellite is designed to support at least 4500 telephone channels. Globalstar uses a combination of FDMA and CDMA (channelized CDMA). The available spectrum is divided into 1.25 MHz sub-bands and multiple users simultaneously occupy each sub-band via CDMA.

BANDWIDTH

There are numerous measures of bandwidth and one must be careful to distinguish among them. The noise bandwidth is the bandwidth the noise power would have if it were contained in a rectangle whose height is the peak spectral power density. The noise bandwidth B is the bandwidth relating the thermal noise power N to the system temperature T, such that $N = k\ T B$, where k is Boltzmann's constant.

The occupied bandwidth is the bandwidth required for the signal to pass through a band limited filter. In an FDMA system, it is the occupied bandwidth that determines the system capacity. The occupied bandwidth is about 1.2 times greater than the noise bandwidth. The extra margin is the value of the rolloff in the pulse shaping, which is used to minimize intersymbol interference (ISI). This type of interference is caused when the tails of preceding and following pulses overlap the peak of the observed pulse. Nyquist showed that ISI could be eliminated if the pulses followed a sin x/x function. In practice, this is impossible to achieve and is approximated by raised cosine pulse shaping.

A third measure of bandwidth is the null-to-null bandwidth. This bandwidth is the width between the zeroes of the main spectral lobe. Other measures of bandwidth, such as the half-power bandwidth, are also used.

FREQUENCY

The frequency is chosen on the basis of maximizing the performance of the system and exploiting the portions of the electromagnetic spectrum that are available. One important relation is that the gain of an antenna increases with increasing frequency for a fixed antenna size. On the other hand, the antenna gain is determined by the area of coverage. Thus once the area of coverage is specified, the gain is determined and then for a specified frequency the size of the antenna is determined.

It can be shown that for fixed transmit antenna gain and fixed receive antenna gain, the received carrier power is maximum when the frequency is minimum. These conditions apply to mobile telephony, since the satellite antenna gain is fixed by the terrestrial cell size and the handset antenna gain is fixed by the condition that the antenna must cover the entire sky. Thus L-band (1.6 GHz) is used because it is the lowest practical frequency that is available.

Another factor is the availability of spectrum. Initially, C-band (6/4 GHz) was used exclusively for the fixed satellite service. Later, Ku-band (14/12 GHz) was used, both because it was a frequency domain that was available to expand capacity and because the higher frequency permits the use of smaller earth terminal antennas. However, more power is required to overcome the detrimental effects of rain.

As the frequency increases the effects of rain increase. Rain degrades a satellite communication link in two ways: by attenuating the signal over the signal path and by increasing the system noise temperature of the earth terminal. Attenuation is caused by scattering and absorption of the electromagnetic waves. As the frequency increases, the wavelength decreases. To the extent that the wavelength is comparable to the size of a typical rain drop (about 1.5 mm), the signal becomes more susceptible to scattering and absorption. The system noise temperature increases because the antenna sees the warm rain at room temperature instead of the cold sky.

At C-band (6 GHz) the wavelength is 50 mm (5.0 cm) and the rain attenuation per kilometer of path is about 0.1 dB/km for a maximum rain rate of 22 mm/h, corresponding to an availability of 99.95 percent in Washington, DC. At Ku-band (14 GHz), the wavelength is 21 mm (2.1 cm) and the rain attenuation is 1 dB/km under the same conditions.

New satellite systems for broadband applications are in various stages of development. These new systems will extend the frequency domain into Ka-band and V-band. Rain attenuation increases dramatically at these frequencies. At Ka-band (30 GHz) the wavelength is 10 mm and the attenuation is 5 dB/km for 99.95% availability in Washington. At V-band (50 GHz) the wavelength is only 6 mm and the corresponding attenuation is 9 dB/km. It will thus not be possible to achieve the same availability at Ka-band and at V-band as we are accustomed to achieving at C-band or even Ku-band. Without mitigating techniques, such as spatial diversity and switching to lower frequencies, the availabilities will be in the neighborhood of 98% for any reasonable rain attenuation allowance. Note that in addition to attenuating the signal, the rain also increases the system noise temperature. This contribution to the total system degradation can be comparable in magnitude to the attenuation itself.

MODULATION

A sinusoidal electromagnetic wave has three properties: amplitude, frequency, and phase. Any one of these parameters can be modulated to convey information. The modulation may be either analog or digital. In analog signals, the range of values of a modulated parameter is continuous. In terrestrial radio systems, for example, AM and FM channels represent amplitude and frequency modulation, respectively. In digital signals, the modulated parameter takes on a finite number of discrete values to represent digital symbols. The advantage of digital transmission is that signals can be regenerated without any loss or distortion to the baseband information.

A fundamental parameter in digital communication is the ratio of bit energy to noise density E_b/N_0. This parameter depends on three characteristics: the bit error ratio (BER); the method of modulation; and the method of coding.

By far the most common form of modulation in digital communication is *M*-ary phase shift keying (PSK). With this method, a digital symbol is represented by one of *M* phase states of a sinusoidal carrier. For binary phase shift keying (BPSK), there are two phase states, 0° and 180°, that represent a binary one or zero. With quaternary phase shift keying (QPSK), there are four phase states representing the symbols 11, 10, 01, and 00. Each symbol contains two bits. A QPSK modulator may be regarded as equivalent to two BPSK modulators out of phase by 90°.

For *M*-ary PSK, the noise bandwidth is the information bit rate divided by the number of bits per symbol. Thus for uncoded BPSK modulation, the noise bandwidth is equal to the information bit rate; for uncoded QPSK modulation the noise bandwidth is one-half the information bit rate. The null-to-null bandwidth is twice the noise bandwidth in each case.

QPSK is usually preferred over BPSK because for a given bit rate and BER it requires the same power, yet requires only half the bandwidth. The saving in bandwidth using QPSK instead of BPSK without any greater power is the digital communication equivalent of a "free lunch". The tradeoff is actually added complexity in the modulator, but QPSK modulators are commonplace and the distinction between a QPSK chip and a BPSK chip is comparable to the distinction between a Pentium computer chip and an 80-286 computer chip: the Pentium chip is much more complex, yet it is ubiquitous and inexpensive.

In some situations BPSK might be preferred, such as when sufficient bandwidth is available and it desired to minimize the spectral power flux density to meet a regulatory requirement. BPSK is also used in CDMA systems, in which the basic principle is maximizing the bandwidth.

Higher order PSK modulation schemes are also used, such as 8PSK. With 8PSK the required bandwidth is only one third the bandwidth of BPSK or two-thirds the bandwidth of QPSK. However, the phase states are 45° closer than QPSK, which makes it more difficult for the receiver to distinguish them. Thus for a given BER the required power is higher than that of either BPSK or QPSK. For example, at a BER of 10^{-8}, 8PSK requires about 4 dB more energy per bit.

In *M*-ary PSK, symbols are distinguished from one another by the carrier phase, but the amplitude remains the same. It is possible to modulate both the phase and the amplitude in order to increase the number of bits per symbol and reduce the bandwidth even further. For example, in 16QAM there are twelve phases and four amplitudes. There are four bits per symbol and the bandwidth is one-fourth the bandwidth of BPSK or one-half the bandwidth of QPSK. However, like 8PSK, this method requires more power because it is more vulnerable to transmission impairments. For a BER of 10^{-8} the required E_b/N_0 is about 4 dB more than QPSK. These higher order levels of carrier modulation are being developed in an effort to decrease the required bandwidth and thus increase the bandwidth efficiency of satellite communication systems.

In offset QPSK (OQPSK) and minimum shift keying (MSK), discontinuous phase transitions are avoided to suppress out-of-band interference. These two methods have a constant envelope and are attractive when the intermodulation effects of transponder nonlinearities are to be minimized. Another alternative is frequency shift keying (FSK). With this method of modulation the frequency of the carrier assumes one of a discrete number of frequencies during each bit period.

CODING

The amount of power, as represented by E_b/N_0, can be reduced through the use of forward error correction (FEC) coding. The reduction in the value of E_b/N_0 is called the coding gain. The code rate is the ratio of information bits to the number of coded bits.

Two types of codes are used: block codes and convolutional codes. In a block code a group of information bits are accepted as a block to the encoder and parity bits are added to form a code word. Names associated with this type of code include Hamming, Golay, BCH, and Reed-Solomon. In a convolutional code, bits are added to a shift register continuously and affect the formation of coded symbols over several bit periods. The number of bit periods that a given bit occupies the shift register is called the constraint length. The optimum method of decoding employs the Viterbi algorithm.

It is now becoming common in advanced communications systems to use concatenated coding, involving both an inner convolutional code and an outer Reed-Solomon block code. The Reed-Solomon code detects and corrects bursty type errors. Interleaving is sometimes also used to scramble the bits after coding and unscramble them before decoding so as to cause bursty errors that occur in transmission to be spread out in time and make them appear to be random. However, interleaving introduces an increase in the encoding delay.

Coding reduces power at the expense of increased bandwidth. For example, a rate 1/2 code doubles the required bandwidth. Thus the bandwidth of a rate 1/2 coded signal using QPSK modulation is equal to the bandwidth of an uncoded signal using BPSK modulation. A rate 1/2 coded 8PSK signal requires 2/3 the bandwidth of uncoded BPSK or 2/3 the bandwidth of rate 1/2 coded QPSK.

BIT RATE

The information bit rate R_b is determined by the service or activity to be supported by the communications link. The required carrier to noise density ratio C/N_0 is related to the energy per bit to noise density ratio E_b/N_0 through the fundamental relation $C/N_0 = R_b\, E_b/N_0$. Thus for a specified bit rate -- together with the specified BER, method of modulation, and method of coding -- the required C/N_0 is determined.

On the other hand, the available C/N_0 provided on either the uplink or the

downlink is determined by the transmitter equivalent isotropic radiated power (EIRP), the receiver figure of merit *G/T*, the free space loss, impairments due to rain, any other losses, and various forms of interference. The transmitter EIRP and receiver *G/T* must be designed to achieve the desired bit rate, or conversely, the given EIRP and *G/T* determine the bit rate that the link can support.

As an example, we return to the paradigm of telephony. For standard pulse code modulation (PCM) to convert a baseband analog waveform to a digital signal, the analog signal must be sampled at the Nyquist rate, or twice the highest baseband frequency, and each sample is encoded by n bits to represent one of $2^n - 1$ levels. For a high quality voice channel, the highest baseband frequency is 4000 Hz, and if each sample is encoded by 8 bits to yield 255 levels, the required bit rate is $2 \times 4000 \times 8 = 64{,}000$ bps, or 64 kbps. This is the classic bit rate for a voice channel.

This recipe for PCM actually applies to the analog-to-digital conversion of any waveform without any knowledge of the nature of the signal. In the particular case of human speech, however, it is possible to drastically reduce the required bit rate by modelling speech patterns. In a vocoder (or voice coder), perceptually important parameters describing the pitch, phonetic envelope, and level of vowel sounds are transmitted instead of the full digital representation of the analog waveform. Thus 4.8 kbps or even 2.4 kbps bit rates are possible. Since bandwidth is at a premium, these are the rates that will used in the satellite mobile telephony systems.

CONCLUSION

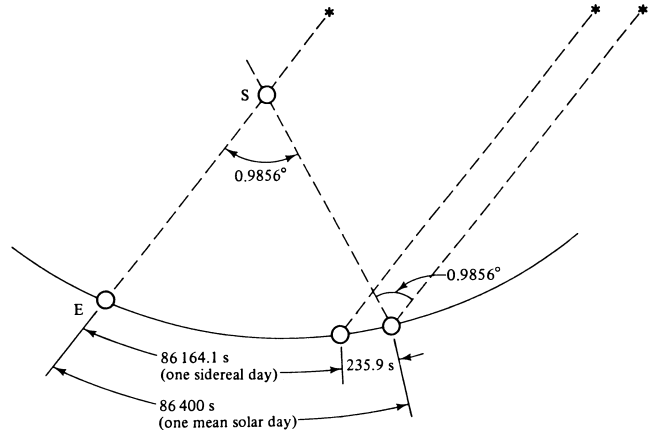
The design of a satellite communication system involves a wide variety of alternatives and tradeoffs. Often a particular set of choices will reflect a particular design philosophy or experience in some other field of communication. The mobile telephony systems illustrate how different designs can be adopted to achieve similar objectives. For example, Iridium is a LEO satellite constellation with polar orbits providing global coverage using FDMA/TDMA. Globalstar is also a LEO constellation but uses inclined orbits for concentration of coverage in mid-latitudes and employs CDMA technology. ICO is an FDMA/TDMA MEO constellation. Aces is a regional system using a single geostationary satellite.

These various possibilities keep the satellite engineer busy. The work, fortunately, is also highly interesting.

Dr. Robert A. Nelson, P.E. is president of Satellite Engineering Research Corporation, a satellite engineering consulting firm in Bethesda, MD. He is *Via Satellite's* Technical Editor.

What Is the Radius of the Geostationary Orbit?

by Robert A. Nelson



Relationship between the sidereal day and the mean solar day.

Most communications satellites operate from the geostationary orbit, since from this orbit a satellite appears to hover over one point on the equator. An Earth station antenna can therefore be pointed at a satellite in a fixed direction and tracking of the satellite across the sky is not required. The basic question to be discussed is, "What is the radius of the geostationary orbit?"

The geostationary orbit must satisfy three conditions: (1) the velocity must be in the direction and sense of the Earth's rotation; (2) the velocity must be constant; and (3) the period of revolution must exactly match the period of rotation of the Earth in inertial space. The first condition implies that the orbit must be a direct orbit in the equatorial plane. The second condition implies that the orbit must be circular. To satisfy the third condition, the radius of the orbit must be chosen to correspond to the required period given by Kepler's third law. According to this law, the square of the orbital period is proportional to the cube of the semimajor axis.[1]

The problem reduces to determining the value of the orbital period. However, it is not simply 24 hours, or one mean solar day. The mean solar day is equal to the average time interval between successive transits of the Sun over a given meridian and is influenced by both the rotation of the Earth on its axis and the motion of the Earth along its orbit. Instead, the appropriate period of the geostationary orbit is the sidereal day, which is the period of rotation of the Earth with respect to the stars. One sidereal day is equal to 23 h 56 m 4.0905 s of mean solar time, or 86 164.0905 mean solar seconds. Using this value in Kepler's third law, we compute the orbital radius as 42 164.172 km.

Yet even this value for the orbital period is not quite correct because the Earth's axis precesses slowly, causing the background of stars to appear to rotate with respect to the celestial reference system. The Earth's axis is tilted by 23.4° with respect to a line perpendicular to the orbital plane and executes a conical motion with a precessional period of about 26 000 years. Therefore, the sidereal day is less than the true period of the Earth's rotation in inertial space by 0.0084 seconds. On this account, the period of the geostationary orbit should be 86 164.0989 mean solar seconds. The corresponding orbital radius is 42 164.174 km.

There is also a correction due to the unit of time itself. The mean solar second is defined as 1/86 400 of a mean solar day. However, in terms of the second of the International System of Units (SI), defined by the hyperfine transition of the cesium atom, the present length of the mean solar day is about 86 400.0025 seconds. The mean solar day exceeds a day of exactly 86 400 seconds by about 2.5 milliseconds due to slowing of the Earth's rotation caused by the Moon's tidal forces on the shallow seas. This extra time accumulates to nearly one second in a year and is compensated by the occasional insertion of a "leap second" into the atomic time scale of Coordinated Universal Time (UTC). Adding this increment to the orbital period, we obtain 86 164.1014 seconds. The corresponding orbital radius is 42 164.175 km.

The analysis so far has assumed that the Earth can be regarded as a perfect sphere. However, in reality the Earth's shape is more nearly oblate. The equatorial radius is 6378.137 km, while the polar radius is 6356.752 km. The gravitational perturbation due to oblateness causes the radius to be increased by 0.522 km.[2] The resulting geostationary orbital radius is 42 164.697 km.

In practice, once the satellite is operational in the geostationary orbit, it is affected by a variety of perturbations that must be compensated by frequent stationkeeping maneuvers using thrusters onboard the spacecraft. These perturbations are caused by the gravitational attractions of the Sun and the Moon, the slightly elliptical shape of the Earth's equator, and solar radiation pressure. Because the orbit is constantly changing, it is not meaningful to define the orbit radius too precisely. By comparison, using recent data for 16 Intelsat satellites, we obtain a semimajor axis with a mean of 42 164.80 km and a standard deviation of 0.46 km.

A perfectly geostationary orbit is a mathematical idealization. Only the distinction between the mean solar day and the sidereal day needs to be taken into account. Therefore, it is customary to quote a nominal orbital period of 86 164 seconds and a radius of 42 164 km. The height above the equator is 35 786 km and the orbital velocity is 3.075 km/s.

---

[1] Mathematically, Kepler's third law may be expressed as $T^2 = (4\pi^2 / GM)\, a^3$, where T is the period, a is the semimajor axis, and GM is the gravitational constant for the Earth, whose value is 398 600.5 km^3 / s^2. For a circular orbit, the semimajor axis a is equal to the radius r.

[2] The correction is $\Delta r = \frac{1}{2} J_2 (R_E / r)^2 r$, where r is the orbital radius, R_E is the Earth's radius, and J_2 is the Earth's oblateness coefficient, 0.001 083.

Via Satellite

Modulation, Power, and Bandwidth

Tradeoffs in Communication Systems Design

by Robert A. Nelson

Modulation is the process by which information is conveyed by means of an electromagnetic wave. The information is impressed on a sinusoidal carrier wave by varying its amplitude, frequency, or phase. Methods of modulation may be either analog or digital.

The power and bandwidth necessary for the transmission of a signal with a given level of quality depends on the method of modulation. There is a classic tradeoff between power and bandwidth that is fundamental to the efficient design of communication systems. This article will identify various methods of analog and digital modulation, describe their characteristics, and analyze their advantages and disadvantages. The scope of the discussion will be restricted to certain common types of modulation systems.

TYPES OF MODULATION

The carrier wave can be represented by the cosine function

$$s(t) = A(t) \cos \theta(t)$$

A sinusoidal carrier wave thus has two fundamental properties: amplitude A and angle θ. Either of these parameters can be varied with time t to transmit information. Frequency and phase modulation are special cases of angle modulation.

In analog modulation the amplitude, frequency, or phase can take on a continuous range of values. The modulated parameter must faithfully follow all of the inflections of the signal to be transmitted. Any variation in this parameter due to propagation losses or interference will result in a distortion of the received demodulated signal.

The principal forms of analog modulation are amplitude modulation (AM) and frequency modulation (FM). These methods are familiar from their application to terrestrial broadcast radio and television.

In digital communication, the modulated parameter takes on only a discrete set of values, each of which represents a symbol. The symbol consists of one or more bits, or binary ones and zeroes. Since the demodulator must merely identify which amplitude, frequency, or phase state is most closely represented in the received signal during each symbol period, the signal can be regenerated without any distortion. Error correction coding is used to reduce bit transition errors caused by interference to meet a specified performance objective.

Two common forms of digital modulation used in satellite communication are phase shift keying (PSK), in which the carrier phase takes on one of a set of discrete values, and frequency shift keying (FSK), in which the frequency may have one of two or more discrete values.

FOURIER PRINCIPLE

A method of representing a time varying function in terms of an infinite trigonometric series was introduced by the eighteenth century French mathematician and physicist Jean Baptiste Fourier (1768 – 1830). According to the Fourier principle, an arbitrary periodic function defined over a specified interval can be represented as the sum of an infinite number of sine and cosine functions whose frequencies are integral multiples of the repetition rate, or fundamental frequency, and whose amplitudes depend on the given function. The frequencies above the fundamental frequency are called the harmonics. The frequency characteristics of a periodic function are determined by the amplitudes of the admixture of harmonics. To a communications engineer, the Fourier principle provides a method of understanding a complicated signal waveform in terms of the amplitudes of the individual harmonics.

For example, the musical sounds produced by a piano, trumpet, or clarinet all performing the tone of concert A (440 Hz) are distinguished by the harmonics that they produce. The fundamental frequency is 440 Hz, but the instruments sound different because they each produce a different set of harmonics. For high fidelity reproduction of these sounds, the range of frequencies should be as high as possible. In the case of the human ear, the frequency range is approximately between 50 Hz and 20,000 Hz. If this range is truncated by the limitations of the recording and reproduction equipment, then the original sound will appear to be distorted and will be easily detected as artificial.

In a typical toll-quality telephone channel, the bandwidth is about 4,000 Hz. This bandwidth is considered to be adequate for the transmission of clear speech. However, since all of the frequencies above 4,000 Hz are filtered out, certain subtle distinctions between similar sounds are lost. That is why, for example, the sounds for *m* and *n* or for *f* and *s* are easily confused over the telephone and we often find it necessary to use phonetics when spelling out a name, even though they are easily distinguished unconscientiously from their higher harmonics when spoken in person.

A mathematical generalization of a Fourier series is the Fourier transform. The Fourier transform permits the conversion of any continuous function in the time domain to a corresponding function in the frequency domain and vice versa. However, the Fourier transform and its inverse involve the use of complex variables. Thus to completely represent the spectrum of a time-dependent function, it is necessary to use the mathematical fiction of both positive and negative frequencies. Using the Fourier transform, one can analyze the frequency spectral content of any time-dependent signal. By a powerful mathematical theorem known as the Wiener-Khintchine Theorem, the power spectral density of a given function of time is the Fourier transform of its autocorrelation function.

FREQUENCY REGIMES

There are three frequency regimes that are involved in the transmission of a signal. These are the baseband frequencies, the intermediate frequency (IF) band, and the radio frequency (RF) band.

The baseband signals are the signals that carry the information, such as from a telephone, microphone, or video camera. The baseband is the range of frequencies generated by the original source of

information. For sound, these frequencies are typically from 0 to a few kilohertz. For video, they may extend to a few megahertz.

The intermediate frequencies are the frequencies present in the signal that are produced after modulation and filtering.

The radio frequency band is the range of frequencies that are transmitted through space. The modulated signal is converted from the intermediate frequency regime to the radio frequency regime by frequency translation. The RF frequencies typically range from a few hundred to a few thousand kilohertz for terrestrial broadcasting and from 1 to 30 gigahertz for satellite communication. These satellite frequencies are in the microwave region, corresponding to wavelengths on the order of a few centimeters, and permit the use of antennas with reasonably sized physical dimensions.

AMPLITUDE MODULATION

With analog amplitude modulation (AM), the message signal $m(t)$ is used to modify the amplitude of the carrier wave. For 100 percent modulation, the amplitude becomes the time-dependent function

$$A(t) = A\,[\,1 + m(t)\,]$$

The angle is given by $\theta = \omega_c\, t + \phi$. The carrier angular frequency ω_c and phase ϕ (which can be taken to be zero) remain constant. Thus the transmitted signal assumes the mathematical form

$$s(t) = A\,[\,1 + m(t)\,]\cos(\omega_c\, t)$$

$$= A\cos(\omega_c\, t) + A\, m(t)\cos(\omega_c\, t)$$

The carrier angular frequency ω_c is related to the frequency f_c by the relation $\omega_c = 2\pi\, f_c$, where ω_c is expressed in radians per second and f_c is expressed in hertz. Multiplication of the cosine function, which is generated in the local oscillator circuit of the modulator, by the message signal produces a spectrum that consists of two sidebands in addition to the frequency of the carrier.

By the Fourier principle, the message signal can be analyzed in terms of its individual sinusoidal components. Thus if the local oscillator generates a carrier $\cos(\omega_c\, t)$ at the intermediate frequency ω_c and it is modulated by one of the components of the message signal represented by $m(t) = \cos(\omega_m\, t)$ at frequency ω_m, then by a trigonometric identity the resulting waveform will be

$$\cos(\omega_c\, t)\cos(\omega\, t) = \tfrac{1}{2}\cos(\omega_c + \omega_m)\, t + \tfrac{1}{2}\cos(\omega_c - \omega_m)\, t$$

The spectrum thus contains the two frequencies $\omega_c + \omega_m$ and $\omega_c - \omega_m$. For example, if the local oscillator generated cosine function at 64 kHz is multiplied by the original baseband signal comprising the set of the four frequencies 1, 2, 3, and 4 kHz, then the resulting spectrum would comprise the frequencies 65, 66, 67, and 68 kHz in the upper sideband and the frequencies 60, 61, 62, and 63 kHz in the lower sideband. Therefore, when the cosine function is multiplied by the message signal, two things happen: the frequencies are translated and the bandwidth is doubled.

In the type of amplitude modulation known as double sideband full carrier (DSB-FC) amplitude modulation, the modulated signal consists of the carrier wave with a time varying amplitude that forms an envelope. The spectrum consists of the carrier frequency, the upper sideband, and the lower sideband. The signal can be easily demodulated simply by passing the modulated signal through a filter to remove the high frequency components contributed by the carrier, leaving the low frequency components of the envelope representing the desired signal.

The transmitted power consists of the carrier power and the power in the sidebands. For 100 percent modulation by a sinusoidal message component, the power in the two sidebands together is one-half the power in the carrier. That is, the total power is three times the power in the sidebands. The sideband power is evenly divided between the two sidebands, giving them each one-fourth the carrier power. For example, full modulation of a 100 watt sinusoidal carrier will add 50 watts to the sidebands, with 25 watts in each sideband, resulting in a total transmitted power of 150 watts.

Since the carrier conveys no information while each sideband contains the same information, this form of modulation is wasteful in both power and bandwidth. The advantage is that only envelope detection is needed to demodulate the signal and the receiver can be built easily and inexpensively. The recovery circuit may be as simple as a diode followed by a low pass filter consisting of a resistor and capacitor in parallel. In US commercial AM radio, the baseband is filtered to 5 kHz and thus the bandwidth per channel is 10 kHz. The AM band extends from 535 kHz to 1705 kHz and the carriers are centered at 540 kHz to 1700 kHz in 10 kHz steps.

In double sideband supressed carrier (DSB-SC) amplitude modulation, both sidebands are transmitted but the carrier is removed. The bandwidth is twice the bandwidth of the baseband signal.

In single sideband (SSB) amplitude modulation the signal is generated by a balanced modulator and filter and the transmitted frequencies consist only of a single sideband. The bandwidth is therefore the same as that of the baseband signal. This method requires only one half the bandwidth as DSB-FC amplitude modulation while transmitting only a fraction of the power.

Envelope detection is not possible in either DSB-SC or SSB. Therefore, the receiver must recover the frequency and phase of the transmitter and is more complex and costly. In DSB-SC a small phase error causes a variation in amplitude, whereas in SSB it affects both amplitude and phase. SSB is thus well suited for voice communication, since the human ear is relatively insensitive to phase distortion, but it is not well adapted to other signals, such as video or digital. It is used in marine and citizens band radio. Before they were replaced by digital circuits, analog telephone channels were combined by frequency division multiplexing using SSB modulation.

FREQUENCY MODULATION

In analog frequency modulation (FM), the message signal is used to vary the frequency of the carrier. The deviation of the instantaneous frequency is directly proportional to the message signal. The amplitude of the carrier remains constant. The range of values of the frequency about the carrier center frequency is called the peak deviation Δf. The instantaneous angular frequency is

$$\omega(t) = d\theta/dt = \omega_c + \Delta\omega\, m(t)$$

where $\Delta\omega = 2\pi\,\Delta f$. For modulation by a sinusoid at the single frequency f_m, the message signal is $m(t) = \cos(\omega_m\, t)$, where $\omega_m = 2\pi\, f_m$. Then $\theta = \omega_c\, t + \beta\sin\omega_m\, t$ and the signal has the mathematical form

$$s(t) = A\cos(\omega_c\, t + \beta\sin\omega_m\, t)$$

where $\beta = \Delta f / f_m$. The parameter β, which is ratio of the peak deviation to the baseband modulation frequency, is a key property called the modulation index.

This expression for $s(t)$ can be expanded into an infinite series of discrete components involving the Bessel function of integral orders, which characteristically occur in mathematical physics when trigonometric functions of trigonometric functions are involved. The resulting spectrum is a distribution of "spikes." (The logo for Cisco Systems is based on this pattern and is also intended to resemble San Francisco's Golden Gate Bridge.) The amplitudes are determined only by the modulation index and become more uniform as the modulation index increases, e.g., for $\beta > 10$. For example, the telemetry, tracking, and control (TT&C) subsystem of a satellite generally uses FM with high modulation index to transmit three tones representing a binary one or zero and execute.

In general, the FM spectrum is a complex function of β consisting of multiple sidebands that occur at integral multiples of the modulating frequency on either side of the carrier rather than, as in AM, consisting of a single pair of sideband frequencies. The spectrum can be analyzed mathematically only in the simplest cases since FM is inherently nonlinear and superposition of individual source signals is not applicable

In principle, the required bandwidth is infinite, but in practice it is given approximately by Carson's rule,

$$B = 2(\beta + 1) f_m = 2(\Delta f + f_m)$$

where f_m is the highest baseband frequency. This well known empirical estimate for determining the practical bandwidth of FM was first suggested in an unpublished memorandum in 1939 by John Renshaw Carson, chief theoretical mathematician at Bell Laboratories. For example, in US commercial FM monaural radio, the highest baseband frequency is 15 kHz and the peak deviation is 75 kHz. Thus the modulation index β is 5 and the required bandwidth B is 180 kHz. Allowing for a 10 kHz guardband on each side, the channel bandwidth is 200 kHz. There are 100 channels, each 200 kHz wide, in the FM band from 88 MHz to 108 MHz.

Two characteristics of FM that are familiar to radio listeners are that the signal quality is much better than AM and that the signal drops out rapidly beyond the nominal range of the transmitter.

The better performance is due to the fact that the signal to noise ratio at the demodulator output is higher for wideband FM than for AM. It was Edwin Howard Armstrong who first recognized the noise-reducing potential of FM for radio broadcasting in the early 1930s. On theoretical grounds Carson had correctly rejected narrowband FM as inferior to AM for the reduction of noise, since he was principally interested in reducing the bandwidth of telephone circuits and hence increasing the system capacity. On the other hand, through experimental measurements Armstrong found that by *widening* the bandwidth the signal to noise ratio could be increased dramatically for radio. He designed and demonstrated the first FM radio circuits.

For a single-frequency sinusoidally modulated signal, the FM output signal to noise ratio at baseband S_b/N_b may be expressed

$$S_b/N_b = 3\beta^2 (B / 2 f_m)(S/N)$$
$$= 3\beta^2 (\beta + 1)(S/N)$$

where S/N is the input signal to noise ratio in the RF channel. Thus after demodulation the output signal to noise ratio is greater than the input signal to noise ratio by the factor $3\beta^2(\beta + 1)$. In contrast, when the same sideband power is transmitted, the output signal to noise ratio is *equal* to the input signal to noise ratio for all types of amplitude modulation. The FM noise density is $N_0 = N/B$. For a double-sideband AM system with the same noise density, the input noise power is $N' = 2 f_m N_0$. If also the AM input signal power is $S' = S$, then

$$S_b/N_b = 3\beta^2 S'/N'$$

Thus after demodulation the FM signal to noise ratio is greater than the corresponding AM signal to noise ratio by a factor of $3\beta^2$. This factor is called the "FM improvement."

From the theoretical relation above, it is seen that as long as $\beta > 0.6$, FM delivers better performance than AM for equal signal power and equal noise power density. However, the FM bandwidth is expanded to $2(\beta + 1)$ times the information bandwidth f_m , whereas the AM bandwidth is $2 f_m$. This is a classic example of trading bandwidth for power. For example, when $\beta = 5$ the FM output signal to noise ratio is 75 times that of an equivalent AM system (19 dB higher), but the bandwidth is 6 times larger. Therefore, the modulation index must be sufficiently high that it provides the desired FM improvement, but it is limited by the need to preserve bandwidth through Carson's rule.

In addition, below a certain threshold input signal to noise ratio that increases somewhat with increasing β, the demodulated signal to noise ratio falls off precipitously. This property is why the range of an FM signal is limited. The existence of a threshold is characteristic of any system that reduces noise in exchange for extra bandwidth and becomes pronounced when the reduction is large. For wideband FM the threshold occurs at roughly 10 dB.

With analog frequency modulation the instantaneous frequency varies directly as the message signal and the phase varies as the integral of the message signal. Analog phase modulation is a closely related form of angle modulation where it is the phase that varies directly as the message signal and where the frequency varies directly as the derivative of the message signal.

TELEVISION

In the United States, the broadcast television standard is the NTSC (National Television System Committee) system. The video signal is modulated by a form of amplitude modulation called vestigial sideband (VSB) amplitude modulation, in which a portion of the lower sideband is transmitted with the upper sideband. VSB AM requires less bandwidth than DSB-SC, overcomes the problem of phase distortion present in SSB and the difficulty of filtering the low frequency content, and permits simple envelope detection. The highest luminance baseband frequency is 4.2 MHz. The upper sideband of the video signal has a bandwidth of 4.2 MHz, while the vestigial lower sideband has a bandwidth of 1.25 MHz. The color signal is transmitted on a separate subcarrier interlaced in the frequency domain with the luminance signal. The audio signal uses frequency modulation, with a highest baseband frequency of 10 kHz and a frequency deviation of 25 kHz. Thus the audio bandwidth is 70 kHz and is centered 4.5 MHz above the video

carrier. The total bandwidth for both the video and audio signals is 6.0 MHz.

For the transmission of a television signal over a satellite, amplitude modulation would be severely affected by losses, various forms of interference, and nonlinearities in the transponder. Therefore, the video signal is frequency modulated along with the audio signal. The peak frequency deviation of video on the main carrier is 12 MHz and the modulation index is 2.86. By Carson's rule, the required bandwidth is 2 (12 MHz + 4.2 MHz) = 32.4 MHz. Thus a bandwidth of 36 MHz was originally chosen for a satellite transponder so that it could safely accommodate one analog FM television channel. Since the FM television channel occupies the entire transponder bandwidth, the transponder can be operated at full power without any intermodulation interference caused by the nonlinear transfer characteristic.

ANALOG TO DIGITAL CONVERSION

In digital communication, information is transmitted in the form of a continuous string of binary ones and zeroes. Thus it is necessary to convert the analog baseband signal, such as a sound or video recording, to a digital signal.

Pulse code modulation (PCM) is a conventional technique that converts an analog waveform into a sequence of binary numbers. The first step is to establish a set of discrete times at which the input signal is sampled. According to a classic theorem of sampling theory stated by Harry Nyquist of Bell Laboratories in 1933, the minimum sampling rate f_s is twice the highest baseband frequency f_m, or $f_s = 2 f_m$. The next step is to represent each analog sample value by a binary number. If there are n bits per sample, then there can be $2^n - 1$ possible levels in each sample. The required bit rate is therefore $R_b = n\ f_s = 2\ n\ f_m$. The original signal waveform is recovered by using a low pass filter. The restriction of each sample to a discrete set of values results in a small amount of quantization noise. This encoding/decoding technique is essentially independent of the nature of the analog signal.

For example, in a conventional toll-quality telephone channel, the practical band extends from about 200 Hz to about 3400 Hz. Rounding up to 4,000 Hz, the Nyquist sampling rate is thus 8,000 samples per second. If 8 bits are allocated for each sample, resulting in 255 possible levels per sample, the required bit rate is 8 × 8,000 bits per second, or 64 kbps, which is the basis of the standard bit rate for a telephone channel. In a digital compact disc (CD) audio recording, the sampling rate is 44,100 samples per second to ensure a perceived bandwidth of more than 20 kHz. With 16 bits per sample for each of two separate stereo channels, the audio data rate is 1.411 Mbps.

In terrestrial cell phone and satellite mobile telephony systems, the bit rate can be as low as 2.4 kbps. This significantly lower bit rate is made possible because the voice coder (vocoder) is designed specifically for speech. The vocoder uses a model of the human vocal tract and synthesizes speech, much as a keyboard musical synthesizer can emulate the sounds of various musical instruments. Only a limited set of perceptually important parameters are transmitted, such as vowel sound, pitch, and level, resulting in fewer bits and smaller bandwidth. Although the speech is intelligible, the quality is below telephone standards. Other forms of sound, such as music, cannot be transmitted.

An NTSC digital television signal following the ITU-R Rec. 601 standard has 30 frames per second, 525 lines per frame, 858 samples for luminance and 429 samples for each of two color differences per line (so-called 4:2:2 component structure), and 8 bits per sample. Theoretically, the required bit rate is 216 Mbps. In practice, there are 480 active lines with 720 samples for luminance and 360 samples for each of two color differences per line, yielding 166 Mbps. The luminance sampling rates for these two formats are 13.5 MHz and 10.4 MHz, repectively, compared with the Nyquist sampling rate of 8.4 MHz for analog video. With compression the bit rate can be reduced to about 8 Mbps (MPEG-2 quality) or 1.5 Mbps (MPEG-1 quality).

PULSE SHAPING

The baseband digital symbols must be represented by a continuous string of pulses of some appropriate form. For example, a "1" may be represented by a positive rectangular pulse and a "0" may be represented by a negative rectangular pulse. This type of pulse train is called "Non-Return to Zero" (NRZ) pulse shaping, because the pulse remains at a constant amplitude over each full bit period. Numerous other pulse shapes are also used, in which "notches" are added to improve synchronization. However, since these pulses require greater bandwidth, the NRZ signal format is generally preferred in satellite communication systems.

Since the pulse train transmits information, each successive pulse is independent of those that came before it. Thus the probability of a given pulse representing a one or zero is random, and the sequence of NRZ pulses is a stochastic process. It can be shown that the autocorrelation function for this case is a triangle function. Thus by the Wiener-Khintchine Theorem, the power spectral density (or power per unit bandwidth at frequency f) is the Fourier transform of the triangle function and happens to be a function that has the form of $(\sin x/x)^2$ centered about 0, where $x = \pi f / R_b$.

In practice, the baseband pulse shapes are not nice, perfect rectangles with right angle corners. To produce such pulses, the bandwidth would have to be infinite. Instead, because of the finite bandwidth of the filter, the pulses are actually rounded "blips." The tails of these blips will tend to overlap, causing a phenomenon known as intersymbol interference (ISI).

Nyquist showed that the pulse shape that required the minimum bandwidth without intersymbol interference is one that in the time domain has the form of the function $\sin(\pi R_b t)/(\pi R_b t)$. For this function, the tails of the preceding and following pulses pass through zero at the peak of the present pulse. In the frequency domain, the Fourier transform looks like a rectangular brick wall. The minium required baseband bandwidth is one half the information bit rate, or $b = R_b / 2$.

But it is impossible to realize this pulse shape in an actual filter. Instead a form of pulse shaping called "raised cosine" filtering is used, characterized by a parameter called the rolloff ρ that is between 0 and 1. A typical value of ρ is 0.2. For zero rolloff, the raised cosine

pulse shape reduces to the ideal sin x/x pulse shape. The actual baseband bandwidth is thus $b = k\ R_b\ /\ 2$, where $k = 1 + \rho$.

DIGITAL MODULATION

In digital communication, the carrier to noise density ratio is given by the relation

$$C/N_0 = R_b\ (\ E_b\ /\ N_0\ )$$

where R_b is the information bit rate. The quantity $E_b\ /\ N_0$ is the ratio of the energy per information bit E_b and the total noise density N_0 (noise power per unit bandwidth) and has fundamental importance. The value of $E_b\ /\ N_0$ is determined by three design factors: the bit error rate, the method of modulation, and the method of forward error correction coding.

If the phase is the parameter that is varied, the modulation is called phase shift keying (PSK). Two common forms of digital modulation used for satellite communication are binary phase shift keying (BPSK) and quaternary phase shift keying (QPSK). If the frequency is varied instead of the phase, the modulation is called frequency shift keying (FSK).

In BPSK modulation the carrier can have one of two phase states, 0° and 180°, which represent a binary one or zero. In a BPSK modulator, the baseband pulse train simply multiplies a cosine function generated by a local oscillator, usually at the intermediate frequency of 70 MHz. Multiplication of cos ($\omega_c\ t$) by a pulse of level + 1 representing a binary 1 leaves the phase of 0 unchanged. On the other hand, multiplication by a pulse of level – 1 representing 0 yields – cos ($\omega_c\ t$) = cos ($\omega_c\ t + \pi$), which changes the phase by 180°. Coherent detection is needed for demodulation. In other words, the receiving circuit must recover the absolute phase of the transmitting circuit. This is usually done by either a Costas loop or a squaring loop.

In QPSK modulation the carrier can assume one of four phase states, consisting of 0°, 90°, 180°, and 270°, which represent the symbols 00, 01, 11, and 10. A QPSK modulator is usually thought of as two BPSK modulators that are out of phase by 90°.

As discussed for AM, forming a product with a cosine function results in a spectrum containing sums and differences of the oscillator frequency and each frequency in the baseband signal. Thus with NRZ pulse shaping, the BPSK spectral density consists of *two* $(\sin\ x/x)^2$ functions, one centered at 70 MHz and the other centered at – 70 MHz in the complex domain. The frequencies are thus translated and the bandwidth is doubled.

In general, the required occupied bandwidth for digital modulation, including forward error correction coding, is

$$B = \ k\ (\ R_b\ /\ m\ )(1\ /\ r\ )$$

where R_b is the bit rate, m is the number of bits per symbol, r is the code rate, and k is the bandwidth expansion factor used to minimize intersymbol interference. For example, if R_b = 64 kbps, m = 2 for QPSK modulation, r = 1/2, and k = 1.2, then B = 76.8 kHz.

For a given bit error rate, the value of $E_b/\ N_0$ required for transmission of both BPSK and QPSK signals is the same and is less than that required for other forms of digital modulation, such as FSK. Hence for a given information bit rate R_b the power is also the same. However, since each QPSK symbol consists of two bits while each BPSK symbol consists of only one bit, the bandwidth required for QPSK modulation is only half that for BPSK. This is the communications equivalent of a "free lunch." (Actually, the tradeoff is in the increased complexity of the QPSK modulator.) Therefore, until recently, QPSK has been the preferred form of digital modulation in satellite communications.

The trend in power and bandwidth does not continue to higher order forms of PSK modulation. For example, in 8-phase PSK (8PSK), there are three bits per symbol, comprising the set 000, 001, 011, 010, 110, 111, 101, and 100. Therefore, the bandwidth for 8PSK is one third that of BPSK and 2/3 that of QPSK. However, since the phase states are closer together and are harder to distinguish, the power required for 8PSK is higher.

The mapping sequences illustrated for QPSK and 8PSK are examples of Gray encoding, in which two symbols represented by neighboring phases differ by only one bit. This method is most often preferred because an error in the demodulator will likely be caused by choosing an adjacent phase state and thus will result in at most one errored bit.

It is possible to vary more than one parameter. In quadrature amplitude modulation (QAM), both the amplitude and phase are modulated. In 16QAM, for example, there are twelve possible phase states and three possible amplitudes. There are four bits per symbol, e.g., 0000, 0001, 0011, etc., and the required bandwidth is one fourth that of BPSK and one-half that of QPSK. However, the required power is much higher. This form of modulation has been used for computer modems and wireless cable television.

SUMMARY

Modulation may be described as the process by which information is impressed on an electromagnetic carrier wave for transmission from one point to another. This article has reviewed several forms of analog and digital modulation. In the design of a communication system, the choice of modulation is of fundamental importance and always involves a tradeoff between power and bandwidth.

In the past, frequency spectrum was relatively plentiful but the power available on a satellite was limited. A satellite typical of the 1980s had a power of less than 1 kW for a payload of 24 transponders. Today, the equation has been reversed. Spectrum is now scarce but a large spacecraft commonly provides 10 to 15 kW for up to 100 transponders. In addition, faster computer processors enable the use of more complex forward error correction coding techniques at high bit rates. Therefore, more spectrum efficient forms of digital modulation such as 8PSK and 16QAM are becoming more attractive, even though the power requirements are higher. Coupled with powerful coding methods such as concatenated Reed Solomon/Viterbi coding, these methods offer the prospect of enhanced spectral efficiency with virtually error-free digital signal transmission.

Dr. Robert A. Nelson, P.E. is president of Satellite Engineering Research Corporation, a satellite engineering consulting firm in Bethesda, Maryland. He is *Via Satellite*'s Technical Editor.

Via Satellite, May 2000

Rain

How It Affects the Communications Link

by Robert A. Nelson

Rain affects the transmission of an electromagnetic signal in three ways: (1) It attenuates the signal; (2) it increases the system noise temperature; and (3) it changes the polarization. All three of these mechanisms cause a degradation in the received signal quality and become increasingly significant as the carrier frequency increases.

At C-band the effects are minor and at Ku-band, while they are noticable, can be accommodated. But at higher frequencies, such as Ka-band or V-band, the degradation can be so great that it simply cannot be be compensated at the level of availability usually expected for lower frequencies. This article will explore the physical mechanisms of rain degradation and will compare the relative effects in various frequency bands used for satellite communication.

ATTENUATION

The first, and most well known, effect of rain is that it attenuates the signal. The attenuation is caused by the scattering and absorption of electromagnetic waves by drops of liquid water. The scattering diffuses the signal, while absorption involves the resonance of the waves with individual molecules of water. Absorption increases the molecular energy, corresponding to a slight increase in temperature, and results in an equivalent loss of signal energy. Attenuation is neglible for snow or ice crystals, in which the molecules are tightly bound and do not interact with the waves.

The attenuation increases as the wavelength approaches the size of a typical raindrop, which is about 1.5 millimeters. Wavelength and frequency are related by the equation $c = \lambda f$, where λ is the wavelength, f is the frequency, and c is the speed of light (approximately 3×10^8 m/s). For example, at the C-band downlink frequency of 4 GHz, the wavelength is 75 millimeters. The wavelength is thus 50 times larger than a raindrop and the signal passes through the rain with relatively small attenuation. At the Ku-band downlink frequency of 12 GHz, the wavelength is 25 millimeters. Again, the wavelength is much greater than the size of a raindrop, although not as much as at C-band. At Ka-band, with a downlink frequency of 20 GHz, the wavelength is 15 millimeters and at V-band, at a downlink frequency of 40 GHz, it is only 7.5 millimeters. At these frequencies, the wavelength and raindrop size are comparable and the attenuation is quite large.

Considerable research has been carried out to model rain attenuation mathematically and to characterize rainfall throughout the world. For example, experimental measurements and methods of analysis are discussed in the book *Radiowave Propagation in Satellite Communications* by Louis J. Ippolito (Van Nostrand, 1986). The standard method of representing rain attenuation is through an equation of the form

$$L_r = \alpha R^{\beta} L = \gamma L$$

where L_r is the rain attenation in decibels (dB), R is the rain rate in millimeters per hour, L is an equivalent path length (km), and α and β are empirical coefficients that depend on frequency and to some extent on the polarization. The factor γ is called the specific rain attenuation, which is expressed in dB/km. The equivalent path length depends on the angle of elevation to the satellite, the height of the rain layer, and the latitude of the earth station.

The rain rate enters into this equation because it is a measure of the average size of the raindrops. When the rain rate increases, *i.e.* it rains harder, the rain drops are larger and thus there is more attenuation. Rain models differ principally in the way the effective path length L is calculated. Two authoritative rain models that are widely used are the Crane model and the ITU-R (CCIR) model.

The original Crane model is the global model. A revision of this model that accounts for both the dense center and fringe area of a rain cell is the so-called two component model. These models are discussed in detail in the book *Electromagnetic Wave Propagation Through Rain* by Robert K. Crane (Wiley, 1996), which is accompanied by spreadsheet add-in software.

In the design of any engineering system, it is impossible to guarantee the perfomance under every conceivable condition. One sets reasonable limits based on the conditions that are expected to occur at a given level of probability. For example, a bridge is designed to withstand loads and stresses that are expected to occur in normal operation and to withstand the forces of wind and ground movement that are most likely to be encountered. But even the best bridge design cannot compensate for a tornado or an earthquake of unusual strength.

Similarly, in the design of a satellite communications link one includes margin to compensate for the effects of rain at a given level of availability. The statistical characterization of rain begins by dividing the world into rain climate zones. Within each zone, the maximum rain rate for a given probability is determined from actual meteorological data accumulated over many years.

Referring to a chart of rain climate zones throught the United States, it might seem inconsistent at first glance that Seattle and San Francisco are in the same rain climate region. Seattle is well known for its rainy climate, whereas San Francisco can justifiably boast of fair weather. However, it is not the annual rainfall that matters, but rather the probability of a given rain rate, since it is the rain rate that determines the average size of a raindrop. Thus in Seattle it rains often but it rarely rains hard. The probability of a cloudburst in Seattle is about the same as that in San Francisco. It is more likely for a heavy rain shower to occur in Washington, DC.

Washington, DC is in rain climate ragion D2. With a probability of 99.95 percent, the maximum rain rate is 22.3 mm/h. Thus if a total rain degradation for this rain rate is compensated by adding sufficient margin to the link budget, there will be a 99.95 percent probability that the signal can be received with the specified system performance objective. That is, there is a probability of only 0.05 percent that the anticipated degradation will be exceeded. This probability translates to a possible total unavailability of 4.38 hours in increments distributed randomly over the entire year.

For a digital signal, the required signal power is determined by the bit rate, the bit error rate, the method of coding, and the method of modulation. The performance objective is specified by the bit error rate. If the maximum allowed rain rate is exceeded, the bit error rate would increase at the nominal bit rate, or else the bit rate would have to decrease to maintain the required bit error rate.

At C-band, the rain attenuation for an elevation angle of 40° and a maximum rain rate of 22.3 mm/h in Washington, DC is 0.1 dB. This is practically a negligible effect. At Ku-band, under the same conditions, the attenuation is 4.5 dB. This is a large but manageable contribution to the link budget. However, at the Ka-band downlink frequency of 20 GHz, the attenuation is 12.2 dB. This would be a significant effect, requiring over 16 times the power as in clear sky conditions. At the uplink frequency of 30 GHz, the attenuation would be 23.5 dB, requiring over 200 times the power. At the V-band downlink frequency of 40 GHz, the attenuation would be 34.6 dB and at the uplink frequency of 50 GHz the attenuation would be 43.7 GHz. These losses simply cannot be accommodated and thus the availability would be much less.

In practice, these high rain attenuations are sometimes avoided by using site diversity, in which two widely separated earth stations are used. The probability that both earth stations are within the same area of rain concentration is small. Alternatively, a portion of spectrum in a lower frequency band may be used where needed. For example, a hybrid Ka-band/Ku-band system might be designed in which Ka-band provides plentiful spectrum in regions of clear weather, but Ku-band is allocated to regions in which the rain margin at Ka-band is exceeded.

SYSTEM TEMPERATURE

In addition to causing attenuation, rain increases the downlink system noise temperature. The figure of merit of the earth station receive antenna is the ratio of the antenna gain to the system temperature G/T. The effect of rain is to increase the system temperature and thus reduce the figure of merit.

The clear sky system temperature is

$$T = T_a + T_e$$

where T_a is the clear sky antenna noise temperature and T_e is the equivalent temperature of the receiver. The antenna temperature is the integrated sky temperature weighted by the antenna gain. At a high angle of elevation, the clear sky temperature is typically about 25 K since the antenna looks at cold space. However, the temperature of liquid water is about 300 K. Thus the rain increases the sky temperature by an order of magnitude. Therefore, the noise admitted to the earth station receive antenna increases and causes further signal degradation. However, rain does not affect the system noise temperature of the satellite because its antenna looks at the warm earth.

The rain layer acts very much like a lossy waveguide. The equivalent temperature of the rain is

$$T_r = (L_r - 1)\, T_0$$

where L_r is the rain loss and T_0 is the physical temperature of the rain. The antenna noise temperature in the presence of rain is given by

$$T'_{\mathrm{a}} = (T_{\mathrm{a}} + T_{\mathrm{r}}) / L_r$$

where T_{a} is the clear sky antenna noise temperature. The system temperature in this case is thus

$$T' = T'_{\mathrm{a}} + T_{\mathrm{e}}$$

where T_e is the equivalent temperature of the receiver, which is the same as before. The increase in system temperature may thus be expressed

$$\Delta T = T' - T = (T_0 - T_{\mathrm{a}})\,(L_r - 1)/L_r$$

The coefficient of the term on the right is about 275 K. The rain causes an increase in system temperature and produces a degradation effect that can be comparable to the attenuation itself. For large attenuation, the limiting ratio of system temperatures is

$$T' / T = (T_0 + T_e) / (T_a + T_e)$$

Thus the antenna temperature approaches the temperature of the rain.

DNPOLARIZATION

Rain also changes the polarization of the signal somewhat. Due to the resistance of the air, a falling raindrop assumes the shape of an oblate spheroid. Wind and other dynamic forces cause the raindrop to be rotated at a statistical distribution of angles. Consequently, the transmission path length through the raindrop is different for different signal polarizations and the polarization of the received signal is altered.

For a satellite communication system with dual linear polarizations, the change in polarization has two effects. First, there is a loss in the signal strength because of misalignment of the antenna relative to the clear sky orientation given by

$$L = 20 \log(\cos \tau)$$

where τ is the tilt angle relative to the polarization direction induced by the rain. Second, there is additional interference noise due to the admission of a portion of the signal in the opposite polarization. The average canting angle with respect to the local horizon can be taken to be 25°.

It is an interesting property of earth-satellite geometry that a linearly polarized signal is not oriented with the local horizontal and vertical directions, even though a horizontally polarized signal is parallel to the equatorial plane and a vertically polarized signal is perpendicular to the equatorial plane when transmitted from the satellite. Thus the optics of the earth station antenna must be correctly rotated in order to attain the appropriate polarization alignment with the satellite. The earth station feed rotation angle θ is given by

$$\tan \theta = G \sin \Delta\lambda / \tan \phi$$

where ϕ is the latitude of the earth station, $\Delta\lambda$ is the difference in longitude, and G is a geometrical factor that for a geostationary satellite is nearly unity. For example, in Washington, DC, at a latitude of 39°, the antenna polarization must be rotated by about 12° if the difference in longitude between the earth station and satellite is 10°. Thus the average effective rain canting angle relative to the polarization direction is about 25° – 12° = 13°. The corresponding polarization loss is 0.2 dB.

CONCLUSION

A variety of new satellite services are being developed in frequency regimes higher than the usual C and Ku bands due to the availability of spectrum. These systems include the broadband services planned for Ka-band. Rain will have a significant impact on the availability. Mitigating techniques such as site diversity or the allocation of spectrum sparingly at lower frequencies where needed will be necessary to ensure uninterrupted service. Alternatively, data rates and bandwidth capacity must be adjusted to maintain the specifed bit error rates.

The mobile satellite service has failed to meet market expectations primarily because of the availability of terrestrial services that are cheaper, have greater signal strength, and require simpler equipment to operate. This paradigm must be avoided if broadband satellite services are to succeed. The competion with fiber and cable will be critically affected by the level of access, the data rates, the complexity of the user equipment, and the availability. The effects of rain will have an important influence on these factors.

Dr. Robert A. Nelson, P.E. is president of Satellite Engineering Research Corporation, a satellite engineering consulting firm in Bethesda, Maryland. He is *Via Satellite*'s Technical Editor.

Effects of Rain Degradation on the Satellite Communications Link

Assumptions	
Rain region	D2
Elevation angle	40°
Earth station latitude	39°
Earth station altitude	0 km
Clear sky temperature	25 K
Receiver equivalent temperature	120 K
Rain temperature	300 K
Polarization	vertical

Availability (percent)		99.99	99.95	99.90	99.50	99.00	98.00	97.00
Unavailability (percent)		0.01	0.05	0.10	0.50	1.00	2.00	3.00
Maximum rain rate (mm/h)		47.1	22.3	15.2	5.3	3.0	1.5	0.9
Attenuation (dB)								
C-band downlink	4 GHz	0.2	0.1	0.1	0.0	0.0	0.0	0.0
C-band uplink	6 GHz	1.3	0.5	0.3	0.1	0.0	0.0	0.0
Ku-band downlink	12 GHz	10.5	4.5	2.9	0.8	0.4	0.2	0.1
Ku-band uplink	14 GHz	13.7	6.1	4.0	1.2	0.6	0.2	0.1
Ka-band downlink	20 GHz	26.4	12.2	8.1	2.5	1.3	0.6	0.3
Ka-band uplink	30 GHz	48.8	23.5	16.1	5.4	2.9	1.4	0.8
V-band downlink	40 GHz	68.8	34.6	24.2	8.6	4.9	2.4	1.5
V-band uplink	50 GHz	83.8	43.7	31.2	11.8	6.9	3.5	1.9
Decrease in G/T (dB)								
C-band downlink	4 GHz	0.4	0.2	0.1	0.0	0.0	0.0	0.0
Ku-band downlink	12 GHz	4.4	3.5	2.8	1.2	0.7	0.3	0.2
Ka-band downlink	20 GHz	4.6	4.4	4.2	2.6	1.8	0.9	0.6
V-band downlink	40 GHz	4.6	4.6	4.6	4.2	3.6	2.6	1.9

Via Satellite, September, 1999

Antennas

The Interface with Space

by Robert A. Nelson

The antenna is the most visible part of the satellite communication system. The antenna transmits and receives the modulated carrier signal at the radio frequency (RF) portion of the electromagnetic spectrum. For satellite communication, the frequencies range from about 0.3 GHz (VHF) to 30 GHz (Ka-band) and beyond. These frequencies represent microwaves, with wavelengths on the order of one meter down to below one centimeter. High frequencies, and the corresponding small wavelengths, permit the use of antennas having practical dimensions for commercial use. This article summarizes the basic properties of antennas used in satellite communication and derives several fundamental relations used in antenna design and RF link analysis.

HISTORY OF ELECTROMAGNETIC WAVES

The quantitative study of electricity and magnetism began with the scientific research of the French physicist Charles Augustin Coulomb. In 1787 Coulomb proposed a law of force for charges that, like Sir Isaac Newton's law of gravitation, varied inversely as the square of the distance. Using a sensitive torsion balance, he demonstrated its validity experimentally for forces of both repulsion and attraction. Like the law of gravitation, Coulomb's law was based on the notion of "action at a distance," wherein bodies can interact instantaneously and directly with one another without the intervention of any intermediary.

At the beginning of the nineteenth century, the electrochemical cell was invented by Alessandro Volta, professor of natural philosophy at the University of Pavia in Italy. The cell created an electromotive force, which made the production of continuous currents possible. Then in 1820 at the University of Copenhagen, Hans Christian Oersted made the momentous discovery that an electric current in a wire could deflect a magnetic needle. News of this discovery was communicated to the French Academy of Sciences two months later. The laws of force between current bearing wires were at once investigated by Andre-Marie Ampere and by Jean-Baptiste Biot and Felix Savart. Within six years the theory of steady currents was complete. These laws were also "action at a distance" laws, that is, expressed directly in terms of the distances between the current elements.

Subsequently, in 1831, the British scientist Michael Faraday demonstrated the reciprocal effect, in which a moving magnet in the vicinity of a coil of wire produced an electric current. This phenomenon, together with Oersted's experiment with the magnetic needle, led Faraday to conceive the notion of a magnetic field. A field produced by a current in a wire interacted with a magnet. Also, according to his law of induction, a time varying magnetic field incident on a wire would induce a voltage, thereby creating a current. Electric forces could similarly be expressed in terms of an electric field created by the presence of a charge.

Faraday's field concept implied that charges and currents interacted directly and locally with the electromagnetic field, which although produced by charges and currents, had an identity of its own. This view was in contrast to the concept of "action at a distance," which assumed bodies interacted directly with one another. Faraday, however, was a self-taught experimentalist and did not formulate his laws mathematically.

It was left to the Scottish physicist James Clerk Maxwell to establish the mathematical theory of electromagnetism based on the physical concepts of Faraday. In a series of papers published between 1856 and 1865, Maxwell restated the laws of Coulomb, Ampere, and Faraday in terms of Faraday's electric and magnetic fields. Maxwell thus unified the theories of electricity and magnetism, in the same sense that two hundred years earlier Newton had unified terrestrial and celestial mechanics through his theory of universal gravitation.

As is typical of abstract mathematical reasoning, Maxwell saw in his equations a certain symmetry that suggested the need for an additional term, involving the time rate of change of the electric field. With this generalization, Maxwell's equations also became consistent with the principle of conservation of charge.

Furthermore, Maxwell made the profound observation that his set of equations, thus modified, predicted the existence of electromagnetic waves. Therefore, disturbances in the electromagnetic field could propagate through space. Using the values of known experimental constants obtained solely from measurements of charges and currents, Maxwell deduced that the speed of propagation was equal to speed of light. This quantity had been measured astronomically by Olaf Romer in 1676 from the eclipses of Jupiter's satellites and determined experimentally from terrestrial measurements by H.L. Fizeau in 1849. He then asserted that light itself was an electromagnetic wave, thereby unifying optics with electromagnetism as well.

Maxwell was aided by his superior knowledge of dimensional analysis and units of measure. He was a member of the British Association committee formed in 1861 that eventually established the centimeter-gram-second (CGS) system of absolute electrical units.

Maxwell's theory was not accepted by scientists immediately, in part because it had been derived from a bewildering collection of mechanical analogies and difficult mathematical concepts. The form of Maxwell's equations as they are known today is due to the German physicist Heinrich Hertz. Hertz simplified them and eliminated unnecessary assumptions.

Hertz's interest in Maxwell's theory was occasioned by a prize offered by the Berlin Academy of Sciences in 1879 for research on the relation between polarization in insulators and electromagnetic induction. By means of his experiments, Hertz discovered how to generate high frequency electrical oscillations. He was surprised to find that these oscillations could be detected at large distances from the apparatus. Up to that time, it had been generally

assumed that electrical forces decreased rapidly with distance according to the Newtonian law. He therefore sought to test Maxwell's prediction of the existence of electromagnetic waves.

In 1888, Hertz set up standing electromagnetic waves using an oscillator and spark detector of his own design and made independent measurements of their wavelength and frequency. He found that their product was indeed the speed of light. He also verified that these waves behaved according to all the laws of reflection, refraction, and polarization that applied to visible light, thus demonstrating that they differed from light only in wavelength and frequency. "Certainly it is a fascinating idea," Hertz wrote, "that the processes in air that we have been investigating represent to us on a million-fold larger scale the same processes which go on in the neighborhood of a Fresnel mirror or between the glass plates used in exhibiting Newton's rings."

It was not long until the discovery of electromagnetic waves was transformed from pure physics to engineering. After learning of Hertz's experiments through a magazine article, the young Italian engineer Guglielmo Marconi constructed the first transmitter for wireless telegraphy in 1895. Within two years he used this new invention to communicate with ships at sea. Marconi's transmission system was improved by Karl F. Braun, who increased the power, and hence the range, by coupling the transmitter to the antenna through a transformer instead of having the antenna in the power circuit directly. Transmission over long distances was made possible by the reflection of radio waves by the ionosphere. For their contributions to wireless telegraphy, Marconi and Braun were awarded the Nobel Prize in physics in 1909.

Marconi created the American Marconi Wireless Telegraphy Company in 1899, which competed directly with the transatlantic undersea cable operators. On the early morning of April 15, 1912, a 21-year old Marconi telegrapher in New York City by the name of David Sarnoff received a wireless message from the Marconi station in Newfoundland, which had picked up faint SOS distress signals from the steamship *Titanic*. Sarnoff relayed the report of the ship's sinking to the world. This singular event dramatized the importance of the new means of communication.

Initially, wireless communication was synonymous with telegraphy. For communication over long distances the wavelengths were greater than 200 meters. The antennas were typically dipoles formed by long wires cut to a submultiple of the wavelength.

Commercial radio emerged during the 1920s and 1930s. The American Marconi Company evolved into the Radio Corporation of America (RCA) with David Sarnoff as its director. Technical developments included the invention of the triode for amplification by Lee de Forest and the perfection of AM and FM receivers through the work of Edwin Howard Armstrong and others. In his book *Empire of the Air: The Men Who Made Radio*, Tom Lewis credits de Forest, Armstrong, and Sarnoff as the three visionary pioneers most responsible for the birth of the modern communications age.

Stimulated by the invention of radar during World War II, considerable research and development in radio communication at microwave frequencies and centimeter wavelengths was conducted in the decade of the 1940s. The MIT Radiation Laboratory was a leading center for research on microwave antenna theory and design. The basic formulation of the radio transmission formula was developed by Harald T. Friis at the Bell Telephone Laboratories and published in 1946. This equation expressed the radiation from an antenna in terms of the power flow per unit area, instead of giving the field strength in volts per meter, and is the foundation of the RF link equation used by satellite communication engineers today.

TYPES OF ANTENNAS

A variety of antenna types are used in satellite communications. The most widely used narrow beam antennas are reflector antennas. The shape is generally a paraboloid of revolution. For full earth coverage from a geostationary satellite, a horn antenna is used. Horns are also used as feeds for reflector antennas.

In a direct feed reflector, such as on a satellite or a small earth terminal, the feed horn is located at the focus or may be offset to one side of the focus. Large earth station antennas have a subreflector at the focus. In the Cassegrain design, the subreflector is convex with an hyperboloidal surface, while in the Gregorian design it is concave with an ellipsoidal surface.

The subreflector permits the antenna optics to be located near the base of the antenna. This configuration reduces losses because the length of the waveguide between the transmitter or receiver and the antenna feed is reduced. The system noise temperature is also reduced because the receiver looks at the cold sky instead of the warm earth. In addition, the mechanical stability is improved, resulting in higher pointing accuracy.

Phased array antennas may be used to produce multiple beams or for electronic steering. Phased arrays are found on many nongeostationary satellites, such as the Iridium, Globalstar, and ICO satellites for mobile telephony.

GAIN AND HALF POWER BEAMWIDTH

The fundamental characteristics of an antenna are its gain and half power beamwidth. According to the reciprocity theorem, the transmitting and receiving patterns of an antenna are identical at a given wavelength

The gain is a measure of how much of the input power is concentrated in a particular direction. It is expressed with respect to a hypothetical isotropic antenna, which radiates equally in all directions. Thus in the direction (θ, ϕ), the gain is

$$G(\theta, \phi) = (dP/d\Omega)/(P_{in}/4\pi)$$

where P_{in} is the total input power and dP is the increment of radiated output power in solid angle $d\Omega$. The gain is maximum along the boresight direction.

The input power is $P_{in} = E_a^2 A / \eta Z_0$ where E_a is the average electric field over the area A of the aperture, Z_0 is the impedance of free space, and η is the net antenna efficiency. The output power over solid angle $d\Omega$ is $dP = E^2 r^2 d\Omega / Z_0$, where E is the electric field at distance r. But by the Fraunhofer theory of diffraction, $E = E_a A / r \lambda$ along the boresight direction, where λ is the

wavelength. Thus the boresight gain is given in terms of the size of the antenna by the important relation

$$G = \eta\,(4\pi / \lambda^2)\,A$$

This equation determines the required antenna area for the specified gain at a given wavelength.

The net efficiency η is the product of the aperture taper efficiency η_a , which depends on the electric field distribution over the antenna aperture (it is the square of the average divided by the average of the square), and the total radiation efficiency $\eta^* = P/P_{in}$ associated with various losses. These losses include spillover, ohmic heating, phase nonuniformity, blockage, surface roughness, and cross polarization. Thus $\eta = \eta_a\,\eta^*$. For a typical antenna, $\eta = 0.55$.

For a reflector antenna, the area is simply the projected area. Thus for a circular reflector of diameter D, the area is $A = \pi D^2/4$ and the gain is

$$G = \eta\,(\pi D / \lambda)^2$$

which can also be written

$$G = \eta\,(\pi D f / c)^2$$

since $c = \lambda f$, where c is the speed of light (3×10^8 m/s), λ is the wavelength, and f is the frequency. Consequently, the gain increases as the wavelength decreases or the frequency increases.

For example, an antenna with a diameter of 2 m and an efficiency of 0.55 would have a gain of 8685 at the C-band uplink frequency of 6 GHz and wavelength of 0.050 m. The gain expressed in decibels (dB) is 10 log(8685) = 39.4 dB. Thus the power radiated by the antenna is 8685 times more concentrated along the boresight direction than for an isotropic antenna, which by definition has a gain of 1 (0 dB). At Ku-band, with an uplink frequency of 14 GHz and wavelength 0.021 m, the gain is 49,236 or 46.9 dB. Thus at the higher frequency, the gain is higher for the same size antenna.

The boresight gain G can be expressed in terms of the antenna beam solid angle Ω_A that contains the total radiated power as

$$G = \eta^*\,(4\pi / \Omega_A)$$

which takes into account the antenna losses through the radiation efficiency η^*. The antenna beam solid angle is the solid angle through which all the power would be concentrated if the gain were constant and equal to its maximum value. The directivity does not include radiation losses and is equal to G / η^*.

The half power beamwidth is the angular separation between the half power points on the antenna radiation pattern, where the gain is one half the maximum value. For a reflector antenna it may be expressed

$$\text{HPBW} = \alpha = k\,\lambda / D$$

where k is a factor that depends on the shape of the reflector and the method of illumination. For a typical antenna, k = 70° (1.22 if α is in radians). Thus the half power beamwidth decreases with decreasing wavelength and increasing diameter.

For example, in the case of the 2 meter antenna, the half power beamwidth at 6 GHz is approximately 1.75°. At 14 GHz, the half power beamwidth is approximately 0.75°. As an extreme example, the half power beamwidth of the Deep Space Network 64 meter antenna in Goldstone, California is only 0.04 ° at X-band (8.4 GHz).

The gain may be expressed directly in terms of the half power beamwidth by eliminating the factor D/λ. Thus,

$$G = \eta\,(\pi k / \alpha)^2$$

Inserting the typical values $\eta = 0.55$ and k = 70°, one obtains

$$G = 27{,}000/\,(\alpha^\circ)^2$$

where α° is expressed in degrees. This is a well known engineering approximation for the gain (expressed as a numeric). It shows directly how the size of the beam automatically determines the gain. Although this relation was derived specifically for a reflector antenna with a circular beam, similar relations can be obtained for other antenna types and beam shapes. The value of the numerator will be somewhat different in each case.

For example, for a satellite antenna with a circular spot beam of diameter 1°, the gain is 27,000 or 44.3 dB. For a Ku-band downlink at 12 GHz, the required antenna diameter determined from either the gain or the half power beamwidth is 1.75 m.

A horn antenna would be used to provide full earth coverage from geostationary orbit, where the angular diameter of the earth is 17.4°. Thus, the required gain is 89.2 or 19.5 dB. Assuming an efficiency of 0.70, the horn diameter for a C-band downlink frequency of 4 GHz would be 27 cm.

EIRP AND *G/T*

For the RF link budget, the two required antenna properties are the equivalent isotropic radiated power (EIRP) and the "figure of merit" G/T. These quantities are the properties of the transmit antenna and receive antenna that appear in the RF link equation and are calculated at the transmit and receive frequencies, respectively.

The equivalent isotropic radiated power (EIRP) is the power radiated equally in all directions that would produce a power flux density equivalent to the power flux density of the actual antenna. The power flux density Φ is defined as the radiated power P per unit area S, or $\Phi = P/S$. But $P = \eta^*\,P_{in}$, where P_{in} is the input power and η^* is the radiation efficiency, and $S = d^2\,\Omega_A$,where d is the slant range to the center of coverage and Ω_A is the solid angle containing the total power. Thus with some algebraic manipulation,

$$\Phi = \eta^*\,(4\pi / \Omega_A)(P_{in} / 4\pi\, d^2) = G\,P_{in} / 4\pi\, d^2$$

Since the surface area of a sphere of radius d is $4\pi\, d^2$, the flux density in terms of the EIRP is

$$\Phi = \text{EIRP} / 4\pi\, d^2$$

Equating these two expressions, one obtains

$$\text{EIRP} = G\,P_{in}$$

Therefore, the equivalent isotropic radiated power is the product of the antenna gain of the transmitter and the power applied to the input terminals of the antenna. The antenna efficiency is absorbed in the definition of gain.

The "figure of merit" is the ratio of the antenna gain of the receiver G and the system temperature T. The system temperature is a measure of the total noise power and includes contributions from the antenna and the receiver. Both the gain and the system temperature must be referenced to the same point in the chain of components in the receiver system. The ratio G/T is important

because it is an invariant that is independent of the reference point where it is calculated, even though the gain and the system temperature individually are different at different points.

ANTENNA PATTERN

Since electromagnetic energy propagates in the form of waves, it spreads out through space due to the phenomenon of diffraction. Individual waves combine both constructively and destructively to form a diffraction pattern that manifests itself in the main lobe and side lobes of the antenna.

The antenna pattern is analogous to the "Airy rings" produced by visible light when passing through a circular aperture. These diffraction patterns were studied by Sir George Biddell Airy, Astronomer Royal of England during the nineteenth century, to investigate the resolving power of a telescope. The diffraction pattern consists of a central bright spot surrounded by concentric bright rings with decreasing intensity.

The central spot is produced by waves that combine constructively and is analogous to the main lobe of the antenna. The spot is bordered by a dark ring, where waves combine destructively, that is analogous to the first null. The surrounding bright rings are analogous to the side lobes of the antenna pattern. As noted by Hertz, the only difference in this behavior is the size of the pattern and the difference in wavelength.

Within the main lobe of an axisymmetric antenna, the gain $G(\theta)$ in a direction θ with respect to the boresight direction may be approximated by the expression

$$G(\theta) = G - 12\,(\theta / \alpha)^2$$

where G is the boresight gain. Here the gains are expressed in dB. Thus at the half power points to either side of the boresight direction, where $\theta = \alpha/2$, the gain is reduced by a factor of 2, or 3 dB. The details of the antenna, including its shape and illumination, are contained in the value of the half power beamwidth α. This equation would typically be used to estimate the antenna loss due to a small pointing error.

The gain of the side lobes can be approximated by an envelope. For new earth station antennas with $D/\lambda > 100$, the side lobes must fall within the envelope $29 - 25 \log \theta$ by international regulation. This envelope is determined by the requirement of minimizing interference between neighboring satellites in the geostationary arc with a nominal 2° spacing.

TAPER

The gain pattern of a reflector antenna depends on how the antenna is illuminated by the feed. The variation in electric field across the antenna diameter is called the antenna taper.

The total antenna solid angle containing all of the radiated power, including side lobes, is

$$\Omega_A = \eta^* (4\pi / G) = (1/\eta_a)\,(\lambda^2 / A)$$

where η_a is the aperture taper efficiency and η^* is the radiation efficiency associated with losses. The beam efficiency is defined as

$$\varepsilon = \Omega_M / \Omega_A$$

where Ω_M is the solid angle for the main lobe. The values of η_a and are ε calculated from the electric field distribution in the aperture plane and the antenna radiation pattern, respectively.

For a theoretically uniform illumination, the electric field is constant and the aperture taper efficiency is 1. If the feed is designed to cause the electric field to decrease with distance from the center, then the aperture taper efficiency decreases but the proportion of power in the main lobe increases. In general, maximum aperture taper efficiency occurs for a uniform distribution, but maximum beam efficiency occurs for a highly tapered distribution.

For uniform illumination, the half power beamwidth is 58.4° λ/D and the first side lobe is 17.6 dB below the peak intensity in the boresight direction. In this case, the main lobe contains about 84 percent of the total radiated power and the first side lobe contains about 7 percent.

If the electric field amplitude has a simple parabolic distribution, falling to zero at the reflector edge, then the aperture taper efficiency becomes 0.75 but the fraction of power in the main lobe increases to 98 percent. The half power beamwidth is now 72.8° λ/D and the first side lobe is 24.6 dB below peak intensity. Thus, although the aperture taper efficiency is less, more power is contained in the main lobe, as indicated by the larger half power beamwidth and lower side lobe intensity.

If the electric field decreases to a fraction C of its maximum value, called the edge taper, the reflector will not intercept all the radiation from the feed. There will be energy spillover with a corresponding efficiency of approximately $1 - C^2$. However, as the spillover efficiency decreases, the aperture taper efficiency increases. The taper is chosen to maximize the illumination efficiency, defined as the product of aperture taper efficiency and spillover efficiency.

The illumination efficiency reaches a maximum value for an optimum combination of taper and spillover. For a typical antenna, the optimum edge taper C is about 0.316, or −10 dB (20 log C). With this edge taper and a parabolic illumination, the aperture taper efficiency is 0.92, the spillover efficiency is 0.90, the half power beamwidth is 65.3° λ/D, and the first side lobe is 22.3 dB below peak. Thus the overall illumination efficiency is 0.83 instead of 0.75. The beam efficiency is about 95 percent.

COVERAGE AREA

The gain of a satellite antenna is designed to provide a specified area of coverage on the earth. The area of coverage within the half power beamwidth is

$$S = d^2\,\Omega$$

where d is the slant range to the center of the footprint and Ω is the solid angle of a cone that intercepts the half power points, which may be expressed in terms of the angular dimensions of the antenna beam. Thus

$$\Omega = K\,\alpha\,\beta$$

where α and β are the principal plane half power beamwidths in radians and K is a factor that depends on the shape of the coverage area. For a square or rectangular area of coverage, $K = 1$, while for a circular or elliptical area of coverage, $K = \pi/4$.

The boresight gain may be approximated in terms of this solid angle by the relation

$$G = \eta'\,(4\pi / \Omega) = (\eta' / K)(41{,}253 / \alpha^\circ\beta^\circ)$$

where $\alpha°$ and $\beta°$ are in degrees and η' is an efficiency factor that depends on the the half power beamwidth. Although η' is conceptually distinct from the net efficiency η, in practice these two efficiencies are roughly equal for a typical antenna taper. In particular, for a circular beam this equation is equivalent to the earlier expression in terms of α if $\eta' = (\pi k / 4)^2 \eta$.

If the area of the footprint S is specified, then the size of a satellite antenna increases in proportion to the altitude. For example, the altitude of Low Earth Orbit is about 1000 km and the altitude of Medium Earth Orbit is about 10,000 km. Thus to cover the same area on the earth, the antenna diameter of a MEO satellite must be about 10 times that of a LEO satellite and the gain must be 100 times, or 20 dB, as great.

On the Iridium satellite there are three main mission L-band phased array antennas. Each antenna has 106 elements, distributed into 8 rows with element separations of 11.5 cm and row separations of 9.4 cm over an antenna area of 188 cm $\times$ 86 cm. The pattern produced by each antenna is divided into 16 cells by a two-dimensional Butler matrix power divider, resulting in a total of 48 cells over the satellite coverage area. The maximum gain for a cell at the perimeter of the coverage area is 24.3 dB.

From geostationary orbit the antenna size for a small spot beam can be considerable. For example, the spacecraft for the Asia Cellular Satellite System (ACeS), being built by Lockheed Martin for mobile telephony in Southeast Asia, has two unfurlable mesh antenna reflectors at L-band that are 12 meters across and have an offset feed. Having different transmit and receive antennas minimizes passive intermodulation (PIM) interference that in the past has been a serious problem for high power L-band satellites using a single reflector. The antenna separation attenuates the PIM products by from 50 to 70 dB.

SHAPED BEAMS

Often the area of coverage has an irregular shape, such as one defined by a country or continent. Until recently, the usual practice has been to create the desired coverage pattern by means of a beam forming network. Each beam has its own feed and illuminates the full reflector area. The superposition of all the individual circular beams produces the specified shaped beam.

For example, the C-band transmit hemi/zone antenna on the Intelsat 6 satellite is 3.2 meters in diameter. This is the largest diameter solid circular aperture that fits within an Ariane 4 launch vehicle fairing envelope. The antenna is illuminated by an array of 146 Potter horns. The beam diameter α for each feed is 1.6° at 3.7 GHz. By appropriately exciting the beam forming network, the specified areas of coverage are illuminated. For 27 dB spatial isolation between zones reusing the same spectrum, the minimum spacing σ is given by the rule of thumb $\sigma \geq 1.4\ \alpha$, so that $\sigma \geq 2.2°$. This meets the specification of $\sigma = 2.5°$ for Intelsat 6.

Another example is provided by the HS-376 dual-spin stabilized Galaxy 5 satellite, operated by PanAmSat. The reflector diameter is 1.80 m. There are two linear polarizations, horizontal and vertical. In a given polarization, the contiguous United States (CONUS) might be covered by four beams, each with a half power beamwidth of 3° at the C-band downlink frequency of 4 GHz. From geostationary orbit, the angular dimensions of CONUS are approximately 6° $\times$ 3°. For this rectangular beam pattern, the maximum gain is about 31 dB. At edge of coverage, the gain is 3 dB less. With a TWTA ouput power of 16 W (12 dBW), a waveguide loss of 1.5 dB, and an assumed beam-forming network loss of 1 dB, the maximum EIRP is 40.5 dBW.

The shaped reflector represents a new technology. Instead of illuminating a conventional parabolic reflector with multiple feeds in a beam-forming network, there is a single feed that illuminates a reflector with an undulating shape that provides the required region of coverage. The advantages are lower spillover loss, a significant reduction in mass, lower signal losses, and lower cost. By using large antenna diameters, the rolloff along the perimeter of the coverage area can be made sharp. The practical application of shaped reflector technology has been made possible by the development of composite materials with extremely low coefficients of thermal distortion and by the availability of sophisticated computer software programs necessary to analyze the antenna. One widely used antenna software package is called GRASP, produced by TICRA of Copenhagen, Denmark. This program calculates the gain from first principles using the theory of physical optics.

SUMMARY

The gain of an antenna is determined by the intended area of coverage. The gain at a given wavelength is achieved by appropriately choosing the size of the antenna. The gain may also be expressed in terms of the half power beamwidth.

Reflector antennas are generally used to produce narrow beams for geostationary satellites and earth stations. The efficiency of the antenna is optimized by the method of illumination and choice of edge taper. Phased array antennas are used on many LEO and MEO satellites. New technologies include large, unfurlable antennas for producing small spot beams from geostationary orbit and shaped reflectors for creating a shaped beam with only a single feed.

Dr. Robert A. Nelson, P.E. is president of Satellite Engineering Research Corporation, a satellite engineering consulting firm in Bethesda, Maryland. Dr. Nelson is instructor for the course *Satellite Communication Systems Engineering: LEO, MEO, GEO* offered by Applied Technology Institute. He is a Lecturer in the Department of Aerospace Engineering at the University of Maryland and is Technical Editor of *Via Satellite* magazine.

Via Satellite May, 1998

Earth Station Technology

The Smarts Behind the Dish

by Robert A. Nelson

The earth station is the link between the terrestrial data sources and the remote satellite resource. Its most familiar component is the earth station antenna, which can be tens of meters in diameter or a small portable dish. In addition, there are numerous, less obvious devices in the chain of devices that transmit or receive the signal. This article will briefly summarize some of the most important aspects of earth station operation.

TRANSMITTER CHAIN

Information to be transmitted is delivered to the earth station via coaxial cable, fiber, terrestrial microwave, or satellite. The devices in the transmitter chain typically consist of the multiplexer, the modulator, the upconvertor, a high power amplifier, and the antenna. The multiplexer combines the individual channels onto a single data stream. The information can be encrypted and encoded with a forward error correction code. The modulator modulates the baseband signal containing the desired information onto an intermediate frequency (IF) carrier, usually at 70 MHz. The upconverter changes the carrier to the radio frequency (RF) signals used to transmit the signal, such as C-band (6 GHz) or Ku-band (14 GHz). The high power amplifier (HPA) amplifies the modulated RF signals from the output of the upconvertors to the required power at the input terminals of the antenna. Finally, the antenna transmits the amplified RF signal to the satellite.

A common form of modulation used in digital satellite communication is M-ary phase shift keying. In this technique, the carrier can assume one of *M* phase states, each of which represents a symbol. In binary phase shift keying (BPSK), there are two phase states, 0° and 180°, representing a binary one or zero. In quaternary phase shift keying (QPSK), there are four phase states that represent the four symbols 11, 01, 00, and 10. A QPSK modulator is equivalent to two BPSK modulators out of phase by 90°. It can be shown that both BPSK and QPSK modulation require the same power per bit for the same bit error rate (BER), but QPSK modulation requires only half the bandwidth. Moreover, all other forms of digital modulation require more power. Thus QPSK is by far the most prevalent form of modulation used in satellite communication and is the industry standard.

Analog frequency modulation (FM) is still commonly used for the transmission of television signals. This has been a convenient mode due to the widespread use of standard equipment. However, there is a slow but deliberate transition to digital technology for television.

The HPA can be either a klystron, a traveling wave tube (TWT), or a solid state power amplifier (SSPA). The bandwidth of a klystron is fairly narrow and is the same as the bandwidth of a transponder, or about 40 MHz at 6 GHz and 80 MHz at 14 GHz. A C-band klystron can have a typical power of 3.3 kW. Although it has a narrow bandwidth, a klystron has relatively high efficiency (about 40 percent) and is generally economical to operate.

A TWT is a broadband device with a bandwidth of about 500 MHz, or about the full bandwidth of a 24 transponder satellite comprising 12 transponders at each polarization. The TWT is more flexible, since it can put the same carrier into all 12 transponders. However, since it is a nonlinear device, it must be backed off to operate in the linear region when multiple carriers are present. A 350 watt Ku-band TWT with 6 dB of backoff has an output power of about 90 watts. The loss can be partially reduced using equalizing devices called linearizers. Helix TWTAs are available at Ku-band with a power of 700 W and at C-band with a power of about 3 kW. Still higher power, at around 10 kW, can be attained with coupled-cavity TWTs.

An SSPA is very efficient and thus does not produce much heat. A typical SSPA power is 2 or 3 watts, but can be as high as 80 or 100 watts.

At Ku-band the HPA must be located near the antenna to minimize losses, but at C-band it can be farther away, such as in the control building, since the loss per unit length of the waveguide diminishes with frequency. For a typical elliptical waveguide, the loss per 30 meters (100 feet) is about 5 dB at Ku-band, compared to about 1 dB at C-band.

RECEIVER CHAIN

The devices in the receiver chain reverse this process. The antenna receives the modulated RF signals from the satellite. The power level at the output terminals of the antenna is about a picowatt. This extremely low power level is comparable to the sound level from a barely audible mosquito. A low noise amplifier (LNA) amplifies the received RF signals. The downconverter changes the received RF signals to IF signals for the demodulators. The information is extracted from the received IF signal by the demodulator and is decoded and decrypted. The demultiplex equipment then distributes the baseband information to the customers through the router and switch after a check of key parameters and rebalancing. Data rates are usually in some standard format, such as a 1.544 Mbps T1 channel or a 45 Mbps DS3 channel, consisting of 28 T1's.

The LNA is mounted on the antenna itself to minimize waveguide loss. This is the first active component and its performance is the primary factor in determining the capability of the receiver. The LNA must have a high gain but contribute very little noise. During the 1980s it was difficult to produce a Ku-band LNA with a noise temperature of 160 K. Today, using field effect transistors, it is possible to reduce this value to around 75 K. Because of the manner in which the noise temperatures combine in a series of devices to produce the overall system temperature, it is essential to place the LNA, with a high gain and low noise temperature, at the head of the receiver chain.

Instead of an LNA, a low noise block downconverter (LNB) may be used. An LNA only amplifies the signal, while an LNB both amplifies the signal and downconverts the frequency to L-band, again to minimize losses. Systems at C-band use both LNA and LNB designs, but Ku-band systems employ LNBs almost exclusively.

ANTENNA

Since electromagnetic energy propagates in the form of waves, the spreading of the energy as it leaves the antenna is described

by the theory of diffraction. The larger the antenna reflector is in comparison with the wavelength, the less spreading there is. The physics of radio waves is identical to the physics of visible light and thus the spreading of radio frequency waves from an antenna reflector is analogous to the transmission of light through an aperture. In fact, a reflector antenna is often referred to as an aperture antenna.

Monochromatic light, such as from a laser, will produce a series of concentric Airy rings when passed through a small circular hole and projected on a screen. The central bright spot is like the main lobe of an antenna pattern. The surrounding dark and bright rings are analogous to the nulls and side lobes of the antenna pattern.

The antenna reflector is usually a paraboloid of revolution. The configuration of the antenna is called a direct feed if the feed horn or low noise amplifier (LNA) is located at the prime focus. Large antennas usually have a subreflector, of either the convex hyperbolic Cassegrain type or the concave ellipsoidal Gregorian type. The subreflector permits the LNA to look into cold space and away from the warm ground, so as to significantly reduce the antenna noise temperture. In an offset antenna, the feed is located to one side. The advantage of the offset design is that it eliminates blockage effects from subreflectors.

Many antennas have tracking capability that permit them to follow a satellite in a geosynchronous, but inclined, orbit. Inclined orbit operation is now a common part of the business plan of satellite operators to extend the useful life of a satellite. The tracking mechanism may be programmed with an ephemeris that determines the look angle as a function of time of day, or it may have an automatic servo loop with a memory that maximizes the received power.

The gain of the antenna is the measure of its ability to concentrate the radio frequency electromagnetic energy in a specified direction, in comparison to a hypothetical isotropic antenna that radiates its energy equally in all directions. It is determined by the size of the physical aperture, the frequency of the radiation, and the efficiency.

The gain is proportional to the square of the antenna diameter and to the square of the frequency. For example, an Andrew 4.6 meter earth station antenna with a Gregorian feed when operated at C-band has a transmit gain of 48.2 dB at 6.175 GHz and a receive gain of 44.4 dB at 4.0 GHz. The same antenna can be used at Ku-band with a transmit gain of 55.1 dB at 14.25 GHz and a receive gain of 53.8 dB at 11.95 GHz.

Factors that affect the efficiency include the geometrical shape of the aperture, the method of illumination (so-called taper), the amount of spillover of energy past the edge of the antenna, surface roughness, blockage, and phase coherence.

Another fundamental parameter is the half power beamwidth. This is the angle between the half power points of the main lobe of the antenna pattern. The half power beamwidth varies in inverse proportion to the frequency and the antenna diameter. For example, the Andrew 4.6 meter antenna at C-band has a transmit half power beamwidth of 0.63° and a receive half power beamwidth of 0.92°, while at Ku-band these values are 0.28° and 0.34°, respectively. On the other hand, a huge 64 m deep space tracking antenna at X-band (8.4 GHz) may have a half power beamwidth of only 0.04°.

Two key parameters are the equivalent isotropic radiated power (EIRP) and the antenna figure of merit. The EIRP is associated with a transmit antenna and is the product of the power P to the input terminals of the antenna and the antenna transmit gain G_t. The figure of merit is associated with a receive antenna. It is the ratio of the antenna receive gain G_r and the system temperature T, which is a measure of the noise power accepted by the antenna and must be as low as possible.

EARTH STATION STANDARDS

Earth stations are characterized by the antenna size, the type of service, the frequency band, the EIRP, and the G/T.

Transmit antennas must conform to international and domestic regulations. The sidelobes must fall within a specified envelope in order to mitigate interference with neighboring satellites and terrestrial systems. The standard international specification for the sidelobe gain of new antennas with diameter to wavelength ratio greater than 100 and operating with a geostationary satellite is given by $G = 29 - 25 \log \theta$ dB, where θ is the off-axis angle. The earth station antenna side lobe pattern is the primary characteristic that determines the minimum spacing between satellites along the geostationary arc.

In addition, the EIRP in a given bandwidth must be within specified values at various bands and the antenna must meet certain radiation hazard constraints. The document governing satellite communications in the United States is Part 25 of the Rules of the Federal Communications Commission (FCC).

Satellite operators also establish standards for their individual systems. For example, INTELSAT has established technical parameters that must be met for acceptance within a particular application.

EARTH STATION FACILITIES

A good example of a commercial earth station facility is the Washington International Teleport, located in Alexandria, Virginia just inside the Washington Capital Beltway. This facility is a hub for voice, data, video, internet, and other services to customers ranging from major television broadcasters to telemedicine and distance learning providers.

Another example is the Hughes Spring Creek earth station, located in southeast Brooklyn, which is the primary TT&C facility for the Hughes C-band and dual payload Galaxy satellites. It also provides backup for the Hughes Ku-band spacecraft. Spring Creek provides uplink access for C-band customers in the New York City area. One of the antennas is used by Hughes Network Systems for a shared VSAT (very small aperture terminal) hub, which supports customers who operate private data networks.

INDUSTRY TRENDS

The legacy of analog video is big transmitters using big antennas. The current trend is to shift the burden of closing the satellite link from the earth station to the satellite, thereby permitting smaller and smaller earth station antennas. Whereas satellites launched during the 1980s were simple repeater ``bent pipe'' satellites, with a typical primary power of 1 to 2 kW, today's generation satellites have extensive onboard processing and a total power of 10 to 15 kW or more.

In addition, there is an emphasis on broadband applications at high frequencies, including Ka-band (30/20 GHz) and the new V-band (50/40 GHz). As noted by Teledesic president Russell Daggatt at the *Satellite 98* Conference, the paradigm for broadband applications used to be video on demand. Today it is internet access via satellite.

There is also changing emphasis on types of services. In the past, satellites have almost entirely provided voice, video, and data connectivity for international and domestic common carriers and operators of television and data networks. Now there is an emphasis on consumer services to meet a global demand for information and a convergence of telephone, data, and video applications.

CONCLUSION

The technology of earth stations has been reviewed and a few illustrative systems have been described. In coming years the number of large earth station facilities that we are accustomed to seeing will continue to grow. However, in addition, there will be an exponential growth of small earth terminals for consumer services. Like a web, the major nodes will be filled in by a dense network of smaller nodes of varies types and sizes.

Dr. Robert A. Nelson, P.E., is president of Satellite Enginering Research Corporation, a satellite engineering consulting firm in Bethesda, MD.

Via Satellite, October, 1998

Earth Station High Power Amplifiers

KPA, TWTA, or SSPA?

by Robert A. Nelson

The high power amplifier (HPA) in an earth station facility provides the RF carrier power to the input terminals of the antenna that, when combined with the antenna gain, yields the equivalent isotropic radiated power (EIRP) required for the uplink to the satellite. The waveguide loss between the HPA and the antenna must be accounted for in the calculation of the EIRP.

The output power typically may be a few watts for a single data channel, around a hundred watts or less for a low capacity system, or several kilowatts for high capacity traffic.

The choice of amplifier is highly dependent on its application, the cost of installation and long term operation, and many other factors. This article will summarize the technologies, describe their important characteristics, and identify some issues important to understanding their differences and relative merits.

TYPES OF AMPLIFIERS

Earth station terminals for satellite communication use high power amplifiers designed primarily for operation in the Fixed Satellite Service (FSS) at C-band (6 GHz), military and scientific communications at X-band (8 GHz), fixed and mobile services at Ku-band (14 GHz), the Direct Broadcast Service (DBS) in the DBS portion of Ku-band (18 GHz), and military applications in the EHF/Q-band (45 GHz). Other frequency bands include those allocated for the emerging broadband satellite services in Ka-band (30 GHz) and V-band (50 GHz). Generally, the frequency used for the earth-to-space uplink is higher than the frequency for the space-to-earth downlink within a given band.

An earth station HPA can be one of three types: a klystron power amplifier (KPA), a traveling wave tube amplifier (TWTA), or a solid state power amplifier (SSPA). The KPA and TWTA achieve amplification by modulating the flow of electrons through a vacuum tube. Solid state power amplifiers use gallium arsenide (GaAs) field effect transistors (FETs) that are configured using power combining techniques. The klystron is a narrowband, high power device, while TWTAs and SSPAs have wide bandwidths and operate over a range of low, medium, and high powers.

The principal technical parameters characterizing an amplifier are its frequency, bandwidth, output power, gain, linearity, efficiency, and reliability. Size, weight, cabinet design, ease of maintenance, and safety are additional considerations. Cost factors include the cost of installation and the long term cost of ownership.

KPAs are normally used for high power narrowband transmission to specific satellite transponders, typically for television program transmission and distribution. TWTAs and SSPAs are used for wideband applications or where frequency agility is required.

Originally, TWTAs provided high power but with poor efficiency and reliability. Compared to a KPA, these disadvantages were regarded as necessary penalties for wide bandwidth. SSPAs first became available about 20 years ago. They were restricted to low power systems requiring only a few watts, such as small earth stations transmitting a few telephone channels.

Within the past decade, however, TWTA and SSPA technologies have both advanced considerably. Today there is vigorous competition between these two technologies for wideband systems.

KPA

The klystron power amplifier (KPA) is a narrowband device capable of providing high power and high gain with relatively high efficiency and stability. The bandwidth is about 45 MHz at C-band and about 80 MHz at Ku-band. Thus a separate KPA is usually required for each satellite transponder.

In a klystron tube an electron beam is formed by accelerating electrons emitted from a heated cathode through a positive potential difference. The electrons enter a series of cavities, typically five in number, which are tuned around the operating frequency and are connected by cylindrical "drift tubes".

In the input cavity the electrons are velocity-modulated by a time-varying electromagnetic field produced by the input radio frequency (RF) signal. The distribution in velocities results in a density modulation further down the tube as the electrons are bunched into clusters when higher velocity electrons catch up with slower electrons in the drift tubes.

Optimum bunching of electrons occurs in the output cavity. Large RF currents are generated in the cavity wall by the density-modulated beam, thereby generating an amplified RF output signal. The energy of the spent electron beam is dissipated as heat in the collector.

The intermediate cavities are used to optimize the saturated power, gain, and bandwidth characteristics. Additional bunching of electrons is provided, yielding higher gain.

The gain is typically 15 dB per cavity, so that a five-cavity klystron can provide a total gain of about 75 dB if synchronously tuned. However, by "stagger tuning" the individual cavities to slightly different frequencies, the bandwidth can be increased with a reduction in gain. A typical gain is on the order of 45 dB.

For a cavity device like a klystron, the bandwidth is a fixed percentage of the frequency of operation. The bandwidth is proportional to the frequency and inversely proportional to the Q (quality) factor, which is defined as 2π times the ratio of the energy stored and the average energy lost in one cycle. Thus at C-band (6 GHz), a typical bandwidth is 45 MHz. But at Ku-band (14 GHz) the bandwidth is about 80 MHz. These bandwidths are well suited for C-band and Ku-band satellite transponders. By adding a sixth, filter cavity the KPA bandwidth can be doubled. Thus 80 MHz KPAs are also available at C-band.

Klystrons can be made with "extended interaction" circuits in one or more cavities that increase the bandwidth substantially. This technology can provide a bandwidth of 400 MHz at 30 GHz. Output powers up to 1 kW can also be achieved at different bandwidths.

Although the bandwidth is relatively small, a conventional klystron can be mechanically tuned over a wide frequency range. A klystron can be capacitively or inductively tuned. All satcom klystrons are inductively tuned because of better efficiency and repeatability. The inductance is varied by moving a wall in the cavity (sliding short).

The output power of a KPA is about 3 kW at C-band and 2 kW at Ku-band. The lowest power KPA offered for commercial satellite communications is around 1 kW, although for certain applications powers under 1 kW are available.

TWTA

The traveling wave tube amplifier (TWTA) consists of the traveling wave tube (TWT) itself and the power supply. The TWT can have either a helix or coupled-cavity design.

The TWT is a broadband device with a bandwidth capability of about an octave, which easily covers the 500 MHz bandwidth typical of satellites in the FSS. It also covers the typical 800 MHz DBS bandwidth requirement, as well as even broader bandwidths in Ka-band and higher bands.

The TWT, like the klystron, is an example of a device based on modulating the flow of electrons in a linear beam, but differs from the klystron by the continuous interaction of the electrons with the RF field over the full length of the tube instead of within the gaps of a few resonant cavities.

The TWT has a heritage of over half a century. The original concept was proposed in 1944 by Rudolf Kompfner, who investigated experimental laboratory microwave tubes while working for the British Admiralty during World War II.

The first practical TWT was developed at the Bell Telephone Laboratories in 1945 by John Pierce and L.M. Field. Bell Labs was interested in the technology for its possible application to communication. By the early 1960s, the TWT was adapted for use in satellite power amplifiers in the Telstar program.

In a TWT, amplification is attained by causing a high density electron beam to interact with an electromagnetic wave that travels along a "slow-wave structure", which usually takes the form of a helical coil. A helix is the widest bandwidth structure available. The electrons are emitted from a heated cathode and are accelerated by a positive voltage applied to an aperture that forms the anode. The electrons are absorbed in a collector at the end of the tube.

The RF signal is applied to the helix. Although the signal travels at nearly the speed of light, its phase velocity along the axis of the tube is much slower because of the longer path in the helix, as determined by the pitch and diameter of the coil, and is nearly equal to the velocity of the electrons. For example, if the electrons are accelerated by a 3,000 volt potential difference on the anode, the speed of the electrons is about one tenth the speed of light. Thus the length of the helix wire should be about ten times the axial length of the tube to bring about synchronism between the RF traveling wave and the electron beam.

The electrons interact with the traveling wave and form clusters that replicate the RF waveform. Midway down the tube, an attenuator, called a "sever", absorbs the RF signal and prevents feedback, which would result in self-oscillation. On the other side of the attenuator, the electromagnetic field of the electron clusters induces a waveform in the helix having the same time-dependence as the original signal but with much higher energy, resulting in amplification. The gain is typically on the order of 40 to 60 dB.

The beam-forming optics are critical parts of the tube. To minimize heat dissipation caused by electrons striking the helix, the beam must be highly focused and the transmission from one end of the tube to the other must be close to 100 percent. When the electrons reach the end of the tube, they impact with the walls of the collector, where most of the heat is generated.

The efficiency of the tube can be improved by applying a negative potential to the collector, which retards the electron beam as the electrons enter it. A collector designed to operate in this way is called a "depressed collector". Less energy is converted to heat as the electron beam impinges on the collector, and consequently less energy is lost as thermal waste.

However, the distribution of electron energies is not uniform. In a multi-stage depressed collector, high energy electrons are directed to stages with high retarding fields and low energy electrons are directed to stages with low retarding fields. This configuration improves the efficiency further, but with greater complexity.

Another means of achieving greater efficiency is through improving beam synchronization. As the electrons travel along the tube and interact with the RF signal, they give up energy and lose velocity. Thus with an ordinary helix, they tend to fall behind the signal. This problem can be mitigated by brute force by increasing the accelerating potential but at the expense of degrading the TWT linearity.

A more elegant method is through the use of a tapered helix, in which the pitch of the helix decreases along the tube. The signal velocity is thus retarded to compensate for beam slowing. The selection of optimum helix configurations has been made possible through advanced computer modeling techniques.

Another type of TWT is a coupled-cavity device, used for high power applications. In this case a series of cavity sections are connected to form the slow-wave structure and is similar to the klystron in this respect. However, in the klystron the cavities are independent, while in the TWT the cavities are coupled by a slot in the wall of each cavity.

The output power of a helix TWTA at C-band ranges from a few watts to about 3 kW, while power levels of 10 kW can be attained with coupled-cavity TWTAs. Helix TWTAs at Ku-band have less power, with a maximum power of around 700 W.

Higher frequency TWTAs are also

available, including those at Ka-band (20 - 30 GHz) and V-band (40 - 50 GHz) where new broadband satellite services are under development. However, because the market is not well established, there are fewer manufacturers of tubes at these frequencies.

The dimensions of the slow-wave structure -- whether a helix, a coupled cavity, or any other type -- are determined by the frequency of operation. The product of wavelength and frequency is equal to the speed of light, so that as the frequency increases the wavelength decreases. The dimensions are proportional to the wavelength. Thus the structure dimensions are approximately inversely proportional to frequency. It is much more difficult to satisfy the criteria for operation at high frequencies such as Ka-band or V-band than at C-band or Ku-band.

The gain of a TWTA can be from 45 dB to 75 dB, depending on the number of active wavelengths in the helix circuit.

SSPA

A solid state power amplifier (SSPA) uses a gallium arsenide (GaAs) metallic semiconductor field effect transistor (FET) as the amplifier gain element. The field effect transistor is a voltage-controlled, unipolar device that conducts only majority carriers and has good thermal stability. In contrast, an ordinary junction transistor is a current-controlled bipolar device, in which both minority and majority carriers participate in conducting an electrical current, and can be thermally unstable. Gallium arsenide FETs can operate at higher frequencies than silicon devices, but the power output is limited by the poor thermal conductivity and lower breakdown voltage.

The maximum continuous output power of a single microwave FET can be from a few watts to several tens of watts. The limiting factor is the generation of heat. At the thermal limit the maximum power is theoretically inversely proportional to the square of the frequency. Thus in the present state of the art, a typical GaAs FET at C-band might have a maximum output power of between 30 W and 45 W, while at Ku-band it is 15 W.

Transistors are combined to form modules. For example, a C-band module containing twelve FETs might be configured with four FETs in parallel in a power-combining output stage, preceded by an intermediate stage with two FETs in parallel and six driver stages in series with one FET per stage. Each FET has a gain of about 8 dB, so that in this case there are eight stages of amplification with a total gain, including losses, of about 60 dB or a factor of 1,000,000. If each of the four FETs in the final stage had an output power of 30 W, the total output power would be 120 W. With a gain of 60 dB, the input power to the first stage would be 0.12 mW.

Higher powers are obtained by assembling modules using standard power combining techniques. The modules are connected in parallel by waveguide elements, such as hybrids or magic tees, to obtain the required total output power. However, the number of parallel modules is limited by combination losses.

SSPAs are readily available with rated powers up to about 500 W at C-band or 100 W at Ku-band.

A new solid state technology is the microwave monolithic integrated circuit (MMIC). This device combines active FETs with passive circuit elements that are deposited on a chip in a single process. The maximum power of a single MMIC is about 20 W at C-band and about 5 W at Ku-band. The total power can be increased by the combination of several MMICs in a series-parallel assembly, but is limited by combination losses which increase as the frequency increases.

Low power MMICs are sometimes used as gain stages to drive high power devices. MMICs can provide higher gain with less space and complexity than discrete low power FETs.

LINEARITY

An important characteristic of any HPA is its linearity. This property is a measure of how well the transfer characteristic of output power vs. input power follows a straight line.

In practice, HPAs are inherently nonlinear devices. Nonlinearity means that the output power is not simply proportional to the input power. Instead, as shown in the figure, the graph representing the output power as a function of input power is more nearly represented by a third order polynomial than by a straight line. Thus there is a region of approximate linearity beyond which the graph curves downward and reaches a plateau.

The output power at this plateau is called the "saturated power" (PS). The saturated power is the maximum power that can be generated. The point of inflection on the curve that is 1 dB below the linear extrapolation is called the "1 dB compression point" (P1).

The transfer characteristic for an SSPA approaches saturation within about 1 dB of the 1 dB compression point, whereas for a TWTA or KPA it bends more gradually, reaching saturation about 3 dB above this point. Therefore, an SSPA has superior linearity to that of a TWTA or KPA over the full range of operation to saturation. However, below the 1 dB compression point, the linearities are similar.

The physical effect of nonlinearity is the generation of harmonics of the fundamental carrier frequency. High frequency harmonics can be eliminated by filtering. For example, at C-band the second harmonic is at 12 GHz and the third harmonic is at 18 GHz, which are well out of band.

For single channel per carrier (SCPC) frequency division multiple access (FDMA) systems, nonlinearity causes intermodulation interference among neighboring channels. The principal source of interference is the third order intermodulation (IM3) product, which comes from the cubic term in the polynomial representation of the transfer characterstic. This contribution to the nonlinearity generates frequencies formed by mixing the second harmonic of one carrier with the fundamental of another. Thus given two carriers with frequencies $f1$ and $f2$, the intermodulation products will have frequencies $2\ f2 - f1$ and $2\ f1 - f2$, which are the same as the frequencies of adjacent channels if they are equally spaced, and cause unacceptable levels of interference. The figure of merit is the so-called two-tone "third order intercept point" (IP3), where the graph of the

intermodulation power intercepts the graph of linear gain.

In this case, the HPA must have a "back off" (BO) to operate at a power (P) in a region that is sufficiently linear where the intermodulation products are within acceptable limits as specified by the maximum carrier to interference power ratio (D3). This ratio may be estimated from the third order intercept and the single carrier ouput power by the relation D3 = 2 (IP3 - P).

Intermodulation interference does not exist if only one carrier occupies the entire bandwidth of the HPA, such as a single 36 MHz analog FM video channel in a KPA or multiple wideband digital time division multiple access (TDMA) channels in a TWTA or SSPA. At any given instant the carrier occupies the full bandwidth of the HPA and there are no neighboring channels with which to interfere. In this case, an HPA can be run at full saturated power.

RATED POWER

The comparison between TWTA and SSPA output power ratings has been obscured by differences in traditional measures of output power. For a TWTA, the rated power is the saturated power, because TWTAs operate at this power for single carrier applications. On the other hand, for an SSPA the rated power is the 1 dB compression point. The "advertised" power of an SSPA is sometimes the saturated power, which is about 1 dB higher. No standards for equal comparison exist in the industry.

Another issue is the distinction between the output power of the TWT and the power at the TWTA output flange, which is about 0.5 to 0.7 dB lower. Allowance must also be made for tube aging. The power delivered to the output flange must be used in system planning. For example, a TWTA with a rated power of 400 W at saturation would actually deliver about 350 W to the antenna waveguide.

For multiple carrier operation, backoff is always referenced with respect to the rated power. A typical output backoff for a TWTA would be about 6 or 7 dB (with respect to saturation). Since every 3 dB corresponds to a factor of 2, a 6 dB backoff would deliver only one-fourth of the rated power. At the same intermodulation specification, an SSPA would require about 2 or 3 dB of backoff (with respect to 1 dB compression), delivering about half the rated power. Thus, as noted by TWTA industry representative Stephan Van Fleteren in *Satellite Online Magazine*, 6 dB of backoff in a TWTA would be roughly equivalent to 3 dB of backoff in an SSPA for the same 1 dB compression point.

For example, in SCPC FDMA applications a C-band TWTA rated at 400 W at saturation would have a practical output power of less than 100 W. On the other hand, an SSPA rated at 175 W at 1 dB compression (or 200 W at saturation) would have a similar practical output power. Therefore, in this situation, an SSPA rated at 175 W would be operationally equivalent to a TWTA rated at 400 W. They would each provide about - 25 dBc separation for two-tone, third order intermodulation performance, which is a standard figure of merit for earth station operation (where dBc refers to the level in decibels of the spurious intermodulation product relative to the carrier).

The same TWTA would have twice the useful power if combined with a linearizer. A linearizer is a network of solid state components that increases gain and phase lead as the input power increases, thus compensating for the gain reduction and phase lag as the TWT approaches saturation. The linearizer reduces the intermodulation level. The output backoff can be reduced by about 3 dB, thereby doubling the output power. Therefore, with a linearizer the traffic capacity could be doubled; alternatively, for a given capacity the required TWTA saturated power could be halved.

If only a single carrier is present, such as in digital TDMA systems, then no backoff is required at all. In this case, the 400 W TWTA without a linearizer would have four times the useful power compared to multicarrier FDMA operation.

In the presence of rain fade, the KPA and TWTA have about 3 dB more margin than an SSPA for extra power when nominally operating in the linear region. There is a tradeoff between increased intermodulation interference and rain attenuation and noise that can be exploited with automatic power control.

EFFICIENCY

The efficiency may be defined as the ratio of the useful output power and the required prime power consumption. Values may differ with different definitions of output power. It is thus best to completely specify the conditions under which the efficiency is calculated. The efficiency depends on the output power and the frequency of operation. A few examples may be illustrative.

In single carrier operation, a typical C-band TWTA rated at 75 W at saturation delivers 70 W to the output flange and has a required prime power consumption of about 350 W. The efficiency is thus 70/350 = 20 percent. A C-band TWTA rated at 400 W delivers 350 W to the flange and requires about 1300 W for an efficiency of 27 percent, and a Ku-band 500 W TWTA delivers 450 W and requires 1900 W for an efficiency of 24 percent. TWTA efficiency has steadily increased, in part due to the development of depressed collector technology and improvements in beam focusing and synchronization.

A representative C-band 100 W SSPA at saturation requires a power of 700 W with an efficiency of about 14 percent. At Ku-band, a typical 100 W SSPA has a power requirement of 1000 W for an efficiency of about 10 percent.

At Ka-band current off-the-shelf TWTA performance is 125 W with a typical efficiency of 20 percent. Current SSPA performance is less than 2 W at about 2 percent efficiency.

As another example, in multiple carrier operation a Ku-band TWTA rated at 125 W at saturation would deliver about 100 W at the output flange. With 6 dB of backoff, the useful power would be 25 W. The maximum prime power consumption would be about 650 W, but in this mode the input power would be about 500 W. The efficiency is thus 5 percent.

This unit would be operationally equivalent to an SSPA rated at 50 W at 1 dB compression, yielding 25 W of useful

power with 3 dB of backoff. The prime power consumption would be approximately 550 W, so the efficiency is about 5 percent.

For the SSPA the power consumption stays the same, regardless of backoff and resulting output power. Until about 10 years ago, this was also true for TWTAs. With multistage depressed collector technology, however, the required input power drops monotonically with output power, albeit not proportionately. Thus in this example, the efficiency of the TWTA is comparable to that of the SSPA.

The efficiency of a KPA is about 40 percent, which is relatively high compared to TWTAs and SSPAs.

RELIABILITY

Reliability is an important consideration in the design of a satellite communication system.

The overall reliability of a TWTA is affected by the failure rates of both the TWT and the power supply. The life-limiting factor of a TWT is cathode depletion. When SSPAs were introduced 20 years ago, TWTs used "B" type cathodes with a relatively short design life of less than 25,000 hours. These are dispenser cathodes made from porous tungsten and filled with metallic compounds of barium, calcium, and aluminum. The operating temperature is about 1000 °C.

Today TWTs employ "M" type cathodes with a design life of over 100,000 hours. These cathodes have a surface layer of osmium, which due to the lower work function enhances electron emission and allows a lower temperature to extend life. The TWT mean time before failure (MTBF) has also improved significantly, from approximately 8,000 hours to approximately 40,000 hours.

The overall TWTA reliability must include the MTBF of the high voltage power supply. The power supplies are susceptible to arcing if they become contaminated. Advances in power supply reliability have in part been the result of a large market for high voltage power supply circuit components with attendant high production and improved quality control. Components used in TWTAs are also used extensively in the consumer product industries to manufacture power supplies for microwave ovens, copiers, and electronic equipment.

SSPAs are not subject to any known life limiting factors. They do not degrade with time, they use low voltage power supplies that are reliable and safe to operate, and they are not affected by vibration. However, SSPAs are sensitive to voltage spikes and fluctuations in temperature.

In redudant 1:1 configurations, the standby SSPA can be inhibited to save power with no penalty in switchover time if the primary SSPA fails. On the other hand, TWTAs have a long warmup time, which requires that the spare be kept in a ready-to-transmit state, consuming full power.

SSPA manufacturers state that SSPAs have a MTBF ten times better than a TWTA's. Additionally, high power SSPAs with multiple FETs in the output stage will continue to operate in the event of a FET failure, although at reduced power.

So far, no authoritative study has been performed on the failure histories of earth station high power amplifiers. The principal data come from studies performed on space-borne satellite power amplifiers. A study of 2400 amplifiers onboard over 70 commercial satellites was reported by R. Strauss in the *International Journal of Satellite Communications* in 1993. Surprisingly, it was concluded that C-band TWTAs had about 33 percent better reliability than C-band SSPAs, while the reliability of Ku-band TWTAs was about the same as that of C-band SSPAs.

The KPA is the most reliable amplifier of all. It has a proven field MTBF of approximately 100,000 hours, or eight years average life.

SUMMARY

There is increasing competition between TWTA and SSPA technologies in C-band and Ku-band. SSPAs compete effectively with TWTAs in efficiency and cost for rated powers up to around 250 W in C-band and 50 W in Ku-band. In these bands TWTAs have several competitive advantages over solid state at higher power levels.

The performance of SSPAs is optimized at lower power levels, where their characteristics include better linearity, lower cost of ownership, and improved safety because of lower voltages. Ease of maintenance is also a consideration, but replacement of the RF module cannot be done easily in the field.

As the power increases, the size and weight of the equipment must increase because of the need for heat sinks. Cooling is accomplished by either conduction or forced air systems.

At high frequencies, TWTAs dominate for high power wideband applications, especially in Ku-band and beyond. At Ka-band and V-band their advantages may become overwhelming. Present wideband amplifiers at Ka-band are all TWTAs. At this time SSPAs are not economically feasible in the DBS band or in Ka-band.

It is often stated that a lower power SSPA can replace a higher power TWTA in multiple carrier FDMA operation due to its superior linearity. However, the comparison may be misleading because of differences in definitions of rated power. In addition, if a linearizer is added, a TWTA will approach the performance of solid state but at higher cost.

When comparing backoffs, power outputs, and efficiencies, the different measures of rated power and any losses in the HPA must be taken into account. The issue of backoff becomes moot for single carrier operation, such as digital TDMA systems, where backoff is not required and the maximum saturated power can be fully utilized.

KPAs have high efficiency and are generally economical to operate. Traditionally, the klystron power amplifier has been a workhorse in the satellite communication industry. For narrowband systems with fixed frequency assignments, especially for television broadcasting, they remain an attractive alternative. The demand continues to grow and contemplated advances in design will further strengthen their role.

Dr. Robert A. Nelson, P.E., is president of Satellite Engineering Research Corporation, a satellite engineering consulting firm in Bethesda, MD, and is Technical Editor of *Via Satellite*.

Satellite 2000 Daily

Advances in Spacecraft Technology

by Robert A. Nelson

Several technical advances over the past decade have resulted in dramatically enhanced spacecraft for the twenty-first century. As a result, power has increased to around 15 kW, the beginning of life mass is in excess of 3000 kg, and payloads have become more complex. These developments have been made possible by improvements in electric power subsystems, propulsion, antennas, on-board processing, and launch vehicles.

Electric power is provided by the solar array during daylight and by batteries during eclipse. The standard solar cell used since the inception of the satellite age has been the silicon cell. This cell has an efficiency of about 15 percent. However, the newest generation satellites use a cell made from gallium arsenide. The efficiency of this cell is 18 percent, but with a dual junction configuration, the efficiency can be increased to 25 percent. Still higher efficiencies are foreseen with triple junction cells. In addition, many spacecraft use concentrators on the solar wings to intensify the incident sunlight. The same power must be provided by the battery during an eclipse. Today, the nickel-hydrogen cell is the chemistry of choice, with a specific energy of 30 W h/kg. However, the lithium ion cell, having an energy density of 110 W h/kg, offers the promise of reducing the battery mass to approximately one-fourth of what is now required for a given power. At this time, the cycle life limits its use for space applications.

There have been several breakthroughs in propulsion. One has been the arcjet technology used in the Lockheed Martin A2100 and Intelsat 8 spacecraft. The specific impulse is 500 seconds, compared to 300 seconds from an integrated propulsion bipropellant system or 220 seconds for the simple catalytic hydrazine thrusters used in the past. However, the most dramatic improvement is in the field of electric ion propulsion. The thrusters have a specific impulse of between 2000 and 4000 seconds and expel ions of xenon. They have a thrust on the order of micronewtons and are operated for periods of about an hour. The xenon ion propulsion system (XIPS) on the HS-702 satellite consumes some 4.5 kW of power from the 15 kW solar array, but requires only 5 kg of xenon per year, and permits a reduction of propellant mass by up to 90 percent for a 12 to 15 year mission life compared to chemical propulsion.

A significant advance in the design of antennas has been the ability to make shaped reflectors. A shaped reflector is a lightweight structure that resembles an oversized “potato chip.” It uses a single feed instead of multiple beams to provide the specified earth coverage. Thus the mass of the antenna subsystem is reduced and the power loss due to the beam forming network is eliminated. This development has been made possible by the fabrication of epoxy graphite materials with extremely low coefficients of thermal expansion that minimize shape distortion and by the creation of sophisticated computer software programs that can analyze the required antenna shape. Another technical advance has been the ability to produce large unfurlable antennas to produce small spot beams, such as those used by the Lockheed Martin Aces satellite for mobile telephony in the Pacific Rim.

As satellites have become larger, they have also become smarter. With on-board computers, the satellites have a large degree of autonomy. Thus instead of stationkeeping and attitude control maneuvers being performed manually, modern spacecraft are capable of maintaining their orbital position and orientation with a minimum amount of intervention from the ground.

It would not be possible to build such large satellites without the ability to put them in space. The Ariane V has a geostationary transfer orbit capability of 5900 kg and is expected to eventually reach 11 000 kg. The Atlas IIIB can put a 4500 kg payload into GTO, while a Zenit Sea Launch rocket has a payload capability of 5000 kg.

In the engineering design of communications satellites, there has been a classic tradeoff between bandwidth and power. In the past, bandwidth was available but the limitations of satellites and launch vehicles constrained the available power. Now the equation has been reversed. There is a tremendous demand for bandwidth, but power is no longer a major problem. As a result, the advances we shall see over the next decade will be in the exploitation of new frequency regimes, such as Ka-band (20 to 30 GHz), Q-band (30 to 40 GHz), and V-band (40 to 50 GHz). The effects of rain at these frequencies will be a challenging obstacle, however. In addition, we can expect to see advances in more spectrum efficient methods of modulation. In the past QPSK has been the industry standard, but other forms of modulation such as 8-PSK and 16-QAM will begin to be used more often. Although these methods require more power, they reduce the bandwidth by factors of 2/3 and 1/2 , respectively, compared to QPSK, thereby requiring less spectrum or permitting higher data rates

Via Satellite, February, 1999

Spacecraft Battery Technology

by Robert A. Nelson

The electrical power subsystem of a spacecraft consists of three basic components: the solar array, the battery, and the power control electronics. The solar array converts light energy from the sun into electrical energy and is the primary source of power. The solar array must also recharge the battery in sunlight. The battery provides electrical power during periods when the sun is eclipsed by the earth and is the secondary source of power. The power control electronics maintain the bus voltage at the desired level.

This article will review the present state of battery technology. The types of batteries available, their physical characteristics, and their advantages and disadvantages will be discussed. In particular, reasons for the trend to use nickel-hydrogen batteries in high power, long lifetime satellite missions will be explained.

ELECTRICAL POWER SUBSYSTEM

In the mid-1980s a typical spacecraft in geostationary orbit had a power of about 1 kW, such as the Hughes HS-376 spin-stabilized spacecraft or the RCA/GE Series 3000 three-axis stabilized spacecraft. By 1990, a power of several kilowatts was common. Beyond 1 kW, three-axis configurations are preferable because they are more mass efficient than spinners.

Today, a typical high performance three-axis stabilized spacecraft has a power between 10 and 15 kW and a nominal lifetime of 15 to 17 years. The Space Systems/Loral Tempo direct broadcast satellite was the first commercial spacecraft in orbit to offer more than 10 kW of power. The Lockheed Martin A2100 Astrolink spacecraft will have 13 kW for broadband services. The Aerospatiale Spacebus 4000 and the Hughes 702 spacecraft will provide 15 kW. Industry analysts predict a power level of 20 kW in the near future. Within a decade, 30 kW satellites may become operational.

The battery must provide this power during each eclipse over the entire satellite lifetime. The battery mass --indeed the entire spacecraft mass -- scales with the total power. Thus the battery must have high reliability with maximum possible energy density.

In geostationary orbit, it has been the practice to design the spacecraft electrical power subsystem as two half-systems, each using one wing of the solar array and one battery. Recently, however, electrical designs using only one battery have been used, due to the proven reliability of nickel-hydrogen batteries and the mass savings that can be realized. For small Low Earth Orbit satellites, a single battery is also advantageous.

The selection of bus voltage is often based on the desire to use proven equipment that has flown on previous satellite programs. In the 1960s, bus voltages of 20 to 30 V were common. By the 1970s and early 1980s, bus voltages had reached 40 to 50 V.

Higher voltages are desirable in order to reduce the required current for a given power, and thus reduce resistive losses and the mass of electric power distribution components. The upper limit of the bus voltage is determined by the power-switching semiconductors. Large spacecraft now in production, such as the Hughes 702 spacecraft, use a bus voltage of around 100 V to handle the increased power.

Achieving high power is not the major problem. Rather, it is managing the heat that is produced as waste. This problem is addressed by designing more efficient components and heat dissipation systems.

ECLIPSES

In geostationary orbit, at an altitude of 35,786 km, the angular radius of the earth is 8.7°. Therefore, the sun is eclipsed by the earth during a portion of the orbit whenever the sun is within 8.7° of the equatorial plane.

There are two eclipse seasons centered about the equinoxes (March 21 and September 21). Each eclipse season lasts 45 days, which is the time the sun takes to move from 8.7° below the equatorial plane to 8.7° above the equatorial plane relative to the earth. Thus in geostationary orbit, there are 90 eclipses per year, requiring 90 charge/discharge cycles of the battery.

The maximum length of an eclipse is 72 minutes (1.2 hours), which occurs at the equinoxes when the sun crosses the equator. The battery must provide power during this time. There are nearly 23 hours available in each revolution to recharge the battery, and typically the battery is recharged in about half that time. Between eclipse seasons, the battery is trickle-charged.

In Low Earth Orbit, at a typical altitude of 1000 km, the orbital period is approximately 100 minutes. The maximum eclipse duration is approximately 35 minutes, which is about one-third of the orbital period, and occurs when the orbital plane is parallel to the earth-sun direction. Only 65 minutes are available to recharge the battery before the next eclipse occurs. For this orbit, there are as many as 14 eclipses per day. Depending on the orbital altitude and inclination, there can be 5000 or more eclipses per year.

BATTERY CHARACTERISTICS

Batteries are either of the primary or secondary type and are classified according to their electrochemistry.

A primary battery is designed for use in lieu of a photovoltaic system. It is discharged to completion and cannot be recharged. It is used for short life missions or for applications that require very little power. A secondary battery is rechargeable and provides power during eclipse periods when the primary source of power, the solar array, is unavailable.

The leading primary battery for spacecraft is the silver-zinc battery. There are also a variety of lithium-based primary batteries, including lithium sulphur dioxide, lithium carbon monofluride, and lithium thionyl chloride. Although lithium has a higher energy density, silver zinc is easier to handle and can be discharged at a much higher rate.

The principal types of secondary (rechargeable) batteries that are designed

for spacecraft use include the nickel-cadmium (NiCd) battery, the nickel-hydrogen (NiH_2) battery, and the super (advanced) nickel-cadmium battery. Silver-zinc (AgZn), lithium ion (Li), and nickel-metal-hydride (NMH) batteries are used for limited applications. The sodium-sulphur (NaS) battery is a technology still in the process of development. Each type of battery has certain applications depending on its performance parameters, such as its energy density, cycle life, and reliability.

The fundamental electrochemical unit is the voltaic cell. A battery consists of several cells connected in series. The bus discharge voltage is equal to the cell voltage multiplied by the number of cells, diminished by the losses.

In each cell, the negative electrode is the source of electrons to the external circuit (oxidation) and thus represents the anode. The positive electrode accepts the electrons from the external circuit (reduction) and thus represents the cathode. The electrolyte is a conducting medium that transfers ions produced at the anode and cathode inside the cell. The separator is a porous material that holds the electrolyte in place and isolates the anode and cathode materials so that electron transfer must occur through the external circuit.

A battery is rated in terms of its capacity. The capacity is the total stored charge. Since charge is the product of the electric current and the time, capacity is measured in ampere hours. The total battery energy, measured in watt hours, is the product of the capacity and the bus voltage. The energy density (specific energy), in watt hours per kilogram, is an important figure of merit for spacecraft applications.

The index of utilization of the battery is the depth of discharge (DoD), defined as the amount of charge drained from the battery expressed as a percentage of its rated capacity.

The charging current, or C-rate, is expressed in the form *C/h*, where *h* is the time in hours to completely charge the battery from its ground state.

The life-limiting property of a spacecraft battery is the number of charge/discharge cycles at a given depth of discharge. The cycle life increases as the depth of discharge decreases. Consequently, a nickel-hydrogen battery rated for 12 years in GEO with 1080 cycles at a depth of discharge of 80 percent might have a life of only 5 years in LEO with 25,000 cycles at a 50 percent DoD.

For example, an *INTELSAT VII* satellite, built by Space Systems/Loral, has two nickel-hydrogen batteries, consisting of 27 cells each. The cells are grouped in two 15 cell modules and two 12 cell modules. The total power requirement during eclipse is approximately 3,100 W at an average discharge voltage of 33.3 V. Each battery has a capacity of 85.5 A h, which provides a total energy of 2847 W h. At 70 percent DoD, the available energy per battery is 1993 W h.

During sunlight operation, the available power from the two solar array wings is 3927 W at autumnal equinox, end of life, and the bus voltage is regulated at 42.0 V. The battery high charge rate is C/13 (6.7 A), and the time to recharge both batteries is about 14 hours.

The total spacecraft dry mass is 1450 kg. The mass budget includes 125 kg for the solar array, 187 kg for the electrical power subsystem, and 62 kg for electrical integration. The batteries alone contribute about 10 percent to the overall spacecraft mass.

NICKEL-CADMIUM

The conventional nickel-cadmium battery was widely used during the first 30 years in the aerospace industry. It consists of four principal components: the nickel positive electrode, the cadmium negative electrode, the aqueous 35 percent potassium hydroxide (KOH) electrolyte, and a nylon cloth separator. Capacities are available in the range of 10 to 40 A h. Nickel-cadmium batteries have high cycle life but have a low energy density of approximately 25 W h/kg.

The cell voltage is approximately constant until it is nearly fully discharged. The temperature is a critical parameter that affects the battery life and must be maintained within a narrow range. In practice, a radiator is used to keep the battery temperature below 24°C, while heaters are used to keep the temperature above 4°C.

Repeated cycling to a deep depth of discharge will cause cracking in the cell plate structures. Over a lifetime of 10 years in geostationary orbit, there will be 900 charge/discharge cycles. Therefore, the depth of discharge is limited to between 50 and 60 percent.

The primary modes of degradation are cadmium migration, hydrolysis and oxidation of the nylon separator material, and electrolyte redistribution. The first two modes are time and temperature dependent, while the third mode is primarily DoD dependent.

In the past, the nylon separator has occasionally posed some difficulties for quality control. In the late 1960s contamination of the Pellon 2505ml nylon material was a problem. A second problem developed in the late 1970s when environmental pollution restrictions caused Pellon to stop producing its 2505ml nylon cloth separator material. Thus substitute materials, such as Pellon 2536, were used that had different physical properties and essentially the nickel-cadmium battery cell had to be redesigned. Stricter environmental laws also increased the cost of working with cadmium, a toxic material, for the negative plate.

NICKEL-HYDROGEN

The nickel-hydrogen battery is now the industry standard. Nickel plates form the positive electrode. Since hydrogen is a gas, the negative electrode contains a platinum catalyst. An aqueous KOH solution is used as the electrolyte. Originally, the separator material was a nonwoven mat of asbestos fibers. Zircar (zirconium oxide) is now commonly used as a separator instead of asbestos.

The nickel-hydrogen battery combines the most stable electrodes of the nickel-cadmium and the oxygen-hydrogen cells. Nickel-hydrogen batteries have fewer inherent failure mechanisms than nickel-cadmium when operated at the same depth of discharge, resulting in higher reliability and longer lifetime in orbit.

The key improvement was the removal of cadmium as the negative electrode. This improvement eliminates cadmium migration as one of the two life-limiting degradation modes within the cell and also circumvents the environmental problems associated with the use of cadmium. The other life-limiting factor, the separator, has also been improved by its replacement first

by asbestos and later by Zircar. Also, the stability of the electrode and the separator strongly reduces electrolyte redistribution. Thus the nickel-hydrogen battery has a considerably longer lifetime than that of nickel-cadmium.

The optimum temperature range for maximum nickel-hydrogen battery capacity is between 10°C and 15°C. On either side of the optimum temperature range, the capacity decreases at the rate of 1 A h per °C of variation.

Three alternative configurations are found in combining cells to form a nickel-hydrogen spacecraft battery: the Individual Pressure Vessel (IPV), which contains one cell per vessel; the Common Pressure Vessel (CPV), which contains two cells per vessel; and the Single Pressure Vessel (SPV), which contains twenty-two cells per vessel.

The Individual Pressure Vessel (IPV) is a widely-used configuration in which each elementary cell is packaged in its own pressure vessel. Each cell generates 1.25 volts. The cells are connected in series to provide the required bus discharge voltage. The mechanical structure required by the high pressure design contributes about 40 percent of the total battery mass.

Nickel-hydrogen cells are manufactured in a wide variety of sizes and capacities. Representative capacities are 5 to 30 A h for a 64 mm cell, 30 to 100 A h for a 90 mm cell, and 100 to 250 A h for a 114 mm cell. The specific energy is approximately 30 W h/kg at 80 percent DoD including packaging.

The Dependent Pressure Vessel (DPV) is a modular IPV type design. The DPV differs from the IPV cell primarily in geometry. The DPV cells are designed to be sandwiched between two endplates.

To reduce mass inefficiency, the Common Pressure Vessel (CPV) design uses two cells in a container. Two cells are connected in series internally within the container and each CPV cell delivers 2.5 volts.

The IPV and CPV cells are typically packaged into multiple cell batteries to provide 28 to 32 V for the spacecraft bus. One additional cell is usually included in an IPV design to allow for a cell failure. The cells are vertically mounted on a lightweight honeycomb baseplate, which provides mechanical structure and a thermal path to remove heat to the radiator.

In the Single Pressure Vessel (SPV) design, all of the cells are packaged in a single container. This design offers the advantages of reductions in mass, volume, and cost. However, the reliability is less because a failure of one cell will result in the failure of the entire battery. Bypass circuits that are generally used in the IPV design cannot be used in this case. The system is designed to operate at internal hydrogen pressures up to 1000 psia.

The trend in communications satellites has been to use nickel-hydrogen in place of nickel-cadmium batteries. There are now well over 5000 nickel-hydrogen cells in over 200 batteries in orbit. This trend in GEO has carried over to LEO. With few exceptions, nearly all GEO and LEO spacecraft are now using or will be using nickel-hydrogen batteries. There is no other chemistry presently available with its unique combination of advantages of energy density, cycle life, and reliability.

SUPER NICKEL-CADMIUM

The super (advanced) nickel-cadmium (S-nickel-cadmium) battery is a proprietary Hughes replacement technology that is now used for some small spacecraft. It consists of nickel plates, cadmium plates, a Zircar separator, and potassium hydroxide electrolyte. The battery is available in capacities ranging from 5 to 50 A h. The specific energy is 31 W h/kg.

The super nickel-cadmium technology has been developed by Hughes as a compromise between the conventional nickel-cadmium and the nickel-hydrogen cells. The goal was to produce a cell with many of the advantages of the nickel-hydrogen cell to prolong lifetime, but retain the packaging advantages offered by the prismatic shape of the conventional nickel-cadmium. They use the same Zircar separator as nickel-hydrogen and have other improvements that are proprietary to Hughes. The few that have been produced and flown are expected to have longer life than the conventional nickel-cadmium batteries. Super nickel-cadmium cells are low pressure, prismatic cells which package as easily as the conventional nickel-cadmium cells. Their use has been mainly on small, LEO missions where they are perceived to have a packaging advantage over nickel-hydrogen. Their disadvantages are that they are both heavier and more expensive than either the conventional nickel-cadmium or the nickel-hydrogen cells.

SILVER-ZINC

The silver-zinc battery is attractive because of its high energy density, which is roughly 110 to 130 W h/kg. Overcharge must be controlled because oxygen that is evolved does not recombine easily. The major disadvantage is low cycle life. It thus has limited application as a secondary battery, but as noted above, it is used widely as a primary battery.

LITHIUM ION

Lithium ion is another high energy density technology. The interest in lithium ion is due to its high specific energy of 85 to 130 W h/kg on a cell basis. It has higher energy density than the nickel-cadmium or nickel-hydrogen technology with fewer hazardous concerns than many other lithium technologies, such as lithium-thyenol-chloride. It can also accept deep discharges, which means more of the available energy can be used.

This technology provides hope that it may eventually be developed to accept a large number of cycles. For these reasons, rechargeable lithium ion development is being watched by all of the prime spacecraft manufacturers for its possible use on selected missions.

Lithium ion does not yet have a competitive cycle life. Typical 20 A h cells have exhibited a 20 percent loss in capacity after less than 200 cycles. At this stage of development, the technology can only be considered for those missions that require very few cycles, such as in sun synchronous orbits or on deep space scientific missions. They are not yet useful for either LEO or GEO orbits.

NICKEL-METAL-HYDRIDE

Nickel-metal-hydride electrochemistry was developed to replace the nickel-cadmium cell with a technology that did not have the problems caused by the cadmium plate. It is seldom used, since its cycle life never approached that of the nickel-cadmium. After significant development by several companies, it was determined that NMH would not have the mass, size, and cycle life that was initially expected.

SODIUM-SULPHUR

The sodium-sulphur battery is still a development technology. It promises to have potentially 50 percent better specific energy than nickel-hydrogen, but is not expected to have as much promise as lithium ion.

Sodium sulphur is a unique technology that must operate at 350°C. Some of the heat required for this high temperature is generated by the battery. But the battery presents a significant impact on the spacecraft thermal design. To minimize this impact, it must be thermally connected to the rest of the spacecraft through a very high, well controlled, thermal resistance.

PROJECTION

There is an enormous market for nickel-hydrogen batteries. These batteries have been demonstrated to be more reliable and mass efficient, with longer cycle life, than their chief competitor, nickel-cadmium. State of the art technologies include lithium ion and sodium sulphur, but these batteries do not have the required cycle life and are difficult to operate.

Spacecraft are being designed with ever higher power and longer lifetimes. Spacecraft powers are now typically around 10 kW and will soon reach 15 to 20 kW. Power levels at 30 kW are foreseen within the next decade.

As these powers increase, so do the satellite lifetimes in orbit. In the 1980s a ten year life was typical. Today, satellites are designed for 15 or more years in geostationary orbit. This lifetime can be extended for another two or three years using inclined orbit techniques, which is becoming standard practice for satellites nearing end of life.

These trends will dictate the use of highly reliable battery technologies, permitting high bus voltages and long life. Nickel-hydrogen will be the likely technology of choice to meet these criteria.

Dr. Robert A. Nelson, P.E., is president of Satellite Engineering Research Corporation, a satellite engineering consulting firm in Bethesda, MD, and is Technical Editor of *Via Satellite*.

Via Satellite, June, 1999

Rocket Science

Technology Trends in Propulsion

by Robert A. Nelson

A satellite is launched into space on a rocket, and once there it is inserted into the operational orbit and is maintained in that orbit by means of thrusters onboard the satellite itself. This article will summarize the fundamental principles of rocket propulsion and describe the main features of the propulsion systems used on both launch vehicles and satellites.

The law of physics on which rocket propulsion is based is called the principle of momentum. According to this principle, the time rate of change of the total momentum of a system of particles is equal to the net external force. The momentum is defined as the product of mass and velocity. If the net external force is zero, then the principle of momentum becomes the principle of conservation of momentum and the total momentum of the system is constant. To balance the momentum conveyed by the exhaust, the rocket must generate a momentum of equal magnitude but in the opposite direction and thus it accelerates forward.

The system of particles may be defined as the sum of all the particles initially within the rocket at a particular instant. As propellant is consumed, the exhaust products are expelled at a high velocity. The center of mass of the total system, subsequently consisting of the particles remaining in the rocket and the particles in the exhaust, follows a trajectory determined by the external forces, such as gravity, that is the same as if the original particles remained together as a single entity. In deep space, where gravity may be neglected, the center of mass remains at rest.

ROCKET THRUST

The configuration of a chemical rocket engine consists of the combustion chamber, where the chemical reaction takes place, and the nozzle, where the gases expand to create the exhaust. An important characteristic of the rocket nozzle is the existence of a throat. The velocity of the gases at the throat is equal to the local velocity of sound and beyond the throat the gas velocity is supersonic. Thus the combustion of the gases within the rocket is independent of the surrounding environment and a change in external atmospheric pressure cannot propagate upstream.

The thrust of the rocket is given by the theoretical equation

$$F = \lambda \dot{m} v_e + (p_e - p_a) A_e$$

This equation consists of two terms. The first term, called the momentum thrust, is equal to the product of the propellant mass flow rate $\dot{m}$ and the exhaust velocity v_e with a correction factor λ for nonaxial flow due to nozzle divergence angle. The second term is called the pressure thrust. It is equal to the difference in pressures p_e and p_a of the exhaust velocity and the ambient atmosphere, respectively, acting over the area A_e of the exit plane of the rocket nozzle. The combined effect of both terms is incorporated into the effective exhaust velocity c. Thus the thrust is also written

$$F = \dot{m} c$$

where an average value of c is used, since it is not strictly constant.

The exhaust exit pressure is determined by the expansion ratio given by

$$\varepsilon = A_e / A_t$$

which is the ratio of the area of the nozzle exit plane A_e and the area of the throat A_t. As the expansion ratio ε increases, the exhaust exit pressure p_e decreases.

The thrust is maximum when the exit pressure of the exhaust is equal to the ambient pressure of the surrounding environment, that is, when $p_e = p_a$. This condition is known as optimum expansion and is achieved by proper selection of the expansion ratio. Although optimum expansion makes the contribution of the pressure thrust zero, it results in a higher value of exhaust velocity v_e such that the increase in momentum thrust exceeds the reduction in pressure thrust.

A conical nozzle is easy to manufacture and simple to analyze. If the apex angle is 2α, the correction factor for nonaxial flow is

$$\lambda = \tfrac{1}{2}(1 + \cos \alpha)$$

The apex angle must be small to keep the loss within acceptable limits. A typical design would be $\alpha = 15°$, for which $\lambda = 0.9830$. This represents a loss of 1.7 percent. However, conical nozzles are excessively long for large expansion ratios and suffer additional losses caused by flow separation. A bell-shaped nozzle is therefore superior because it promotes expansion while reducing length.

ROCKET PROPULSION PARAMETERS

The specific impulse I_{sp} of a rocket is the parameter that determines the overall effectiveness of the rocket nozzle and propellant. It is defined as the ratio of the thrust and the propellant weight flow rate, or

$$I_{sp} = F / \dot{m} g = c / g$$

where g is a conventional value for the acceleration of gravity (9.80665 m/s^2 exactly). Specific impulse is expressed in seconds.

Although gravity has nothing whatever to do with the rocket propulsion chemistry, it has entered into the definition of specific impulse because in past engineering practice mass was expressed in terms of the corresponding weight on the surface of the earth. By inspection of the equation, it can be seen that the specific impulse I_{sp} is physically equivalent to the effective exhaust velocity c, but is rescaled numerically and has a different unit because of division by g. Some manufacturers now express specific impulse in newton seconds per kilogram, which is the same as effective exhaust velocity in meters per second.

Two other important parameters are the thrust coefficient C_F and the characteristic exhaust velocity c^*. The thrust coefficient is defined as

$$C_F = F / A_t p_c = \dot{m} c / A_t p_c$$

where F is the thrust, A_t is the throat area, and p_c is the chamber pressure. This parameter is the figure of merit of the nozzle design. The characteristic

exhaust velocity is defined as

$$c^* = A_t \; p_c \; / \; \dot{m} \;\; = c \; / \; C_F$$

This parameter is the figure of merit of the propellant. Thus the specific impulse may be written

$$I_{sp} = C_F \;\; c^* \; / \; g$$

which shows that the specific impulse is the figure of merit of the nozzle design and propellant as a whole, since it depends on both C_F and c^*. However, in practice the specific impulse is usually regarded as a measure of the efficiency of the propellant alone.

LAUNCH VEHICLE PROPULSION SYSTEMS

In the first stage of a launch vehicle, the exit pressure of the exhaust is equal to the sea level atmospheric pressure 101.325 kPa (14.7 psia) for optimum expansion. As the altitude of the rocket increases along its trajectory, the surrounding atmospheric pressure decreases and the thrust increases because of the increase in pressure thrust. However, at the higher altitude the thrust is less than it would be for optimum expansion at that altitude. The exhaust pressure is then greater than the external pressure and the nozzle is said to be underexpanded. The gas expansion continues downstream and manifests itself by creating diamond-shaped shock waves that can often be observed in the exhaust plume.

The second stage of the launch vehicle is designed for optimum expansion at the altitude where it becomes operational. Because the atmospheric pressure is less than at sea level, the exit pressure of the exhaust must be less and thus the expansion ratio must be greater. Consequently, the second stage nozzle exit diameter is larger than the first stage nozzle exit diameter.

For example, the first stage of a Delta II 7925 launch vehicle has an expansion ratio of 12. The propellant is liquid oxygen and RP-1 (a kerosene-like hydrocarbon) in a mixture ratio (O/F) of 2.25 at a chamber pressure of 4800 kPa (700 psia) with a sea level specific impulse of 255 seconds. The second stage has a nozzle expansion ratio of 65 and burns nitrogen tetroxide and Aerozene 50 (a mixture of hydrazine and unsymmetrical dimethyl hydrazine) in a mixture ratio of 1.90 at a chamber pressure of 5700 kPa (830 psia), which yields a vacuum specific impulse of 320 seconds.

In space, the surrounding atmospheric pressure is zero. In principle, the expansion ratio would have to be infinite to reduce the exit pressure to zero. Thus optimum expansion is impossible, but it can be approximated by a very large nozzle diameter, such as can be seen on the main engines of the space shuttle with $\varepsilon = 77.5$. There is ultimately a tradeoff between increasing the size of the nozzle exit for improved performance and reducing the mass of the rocket engine.

In a chemical rocket, the exhaust velocity, and hence the specific impulse, increases as the combustion temperature increases and the molar mass of the exhaust products decreases. Thus liquid oxygen and liquid hydrogen are nearly ideal chemical rocket propellants because they burn energetically at high temperature (about 3200 K) and produce nontoxic exhaust products consisting of gaseous hydrogen and water vapor with a small effective molar mass (about 11 kg/kmol). The vacuum specific impulse is about 450 seconds. These propellants are used on the space shuttle, the Atlas Centaur upper stage, the Ariane-4 third stage, the Ariane-5 core stage, the H-2 first and second stages, and the Long March CZ-3 third stage.

SPACECRAFT PROPULSION SYSTEMS

The spacecraft has its own propulsion system that is used for orbit insertion, stationkeeping, momentum wheel desaturation, and attitude control. The propellant required to perform a maneuver with a specified velocity increment Δv is given by the "rocket equation"

$$\Delta m \;=\; m_0 \, [\, 1 - \exp(- \Delta v \, / \, I_{sp} \, g) \,]$$

where m_0 is the initial spacecraft mass. This equation implies that a reduction in velocity increment or an increase in specific impulse translates into a reduction in propellant.

In the case of a geostationary satellite, the spacecraft must perform a critical maneuver at the apogee of the transfer orbit at the synchronous altitude of 35,786 km to simultaneously remove the inclination and circularize the orbit. The transfer orbit has a perigee altitude of about 200 km and an inclination roughly equal to the latitude of the launch site. To minimize the required velocity increment, it is thus advantageous to have the launch site as close to the equator as possible.

For example, in a Delta or Atlas launch from Cape Canaveral the transfer orbit is inclined at 28.5° and the velocity increment at apogee is 1831 m/s; for an Ariane launch from Kourou the inclination is 7° and the velocity increment is 1502 m/s; while for a Zenit flight from the Sea Launch platform on the equator the velocity increment is 1478 m/s. By the rocket equation, assuming a specific impulse of 300 seconds, the fraction of the separated mass consumed by the propellant for the apogee maneuver is 46 percent from Cape Canaveral, 40 percent from Kourou, and 39 percent from the equator. As a rule of thumb, the mass of a geostationary satellite at beginning of life is on the order of one half its mass when separated from the launch vehicle.

Before performing the apogee maneuver, the spacecraft must be reoriented in the transfer orbit to face in the proper direction for the thrust. This task is sometimes performed by the launch vehicle at spacecraft separation or else must be carried out in a separate maneuver by the spacecraft itself. In a launch from Cape Canaveral, the angle through which the satellite must be reoriented is about 132°.

Once on station, the spacecraft must frequently perform a variety of stationkeeping maneuvers over its mission life to compensate for orbital perturbations. The principal perturbation is the combined gravitational attractions of the sun and moon, which causes the orbital inclination to increase by nearly one degree per year. This perturbation is compensated by a north-south stationkeeping maneuver approximately once every two weeks so as to keep the satellite within 0.05° of the equatorial plane. The average annual velocity increment is about 50 m/s, which represents 95 percent of the total stationkeeping fuel budget. Also, the slightly elliptical shape of the earth's equator causes a longitudinal drift, which

is compensated by east-west stationkeeping maneuvers about once a week, with an annual velocity increment of less than 2 m/s, to keep the satellite within 0.05° of its assigned longitude.

In addition, solar radiation pressure caused by the transfer of momentum carried by light and infrared radiation from the sun in the form of electromagnetic waves both flattens the orbit and disturbs the orientation of the satellite. The orbit is compensated by an eccentricity control maneuver that can sometimes be combined with east-west stationkeeping. The orientation of the satellite is maintained by momentum wheels supplemented by magnetic torquers and thrusters. However, the wheels must occasionally be restored to their nominal rates of rotation by means of a momentum wheel desaturation maneuver in which a thruster is fired to offset the change in angular momentum.

Geostationary spacecraft typical of those built during the 1980s have solid propellant rocket motors for the apogee maneuver and liquid hydrazine thrusters for stationkeeping and attitude control. The apogee kick motor uses a mixture of HTPB fuel and ammonium perchlorate oxidizer with a specific impulse of about 285 seconds. The hydrazine stationkeeping thrusters operate by catalytic decomposition and have an initial specific impulse of about 220 seconds. They are fed by the pressure of an inert gas, such as helium, in the propellant tanks. As propellant is consumed, the gas expands and the pressure decreases, causing the flow rate and the specific impulse to decrease over the mission life. The performance of the hydrazine is enhanced in an electrothermal hydrazine thruster (EHT), which produces a hot gas mixture at about 1000 °C with a lower molar mass and higher enthalpy and results in a higher specific impulse of between 290 and 300 seconds.

For example, the Ford Aerospace (now Space Systems/Loral) INTELSAT V satellite has a Thiokol AKM that produces an average thrust of 56 kN (12,500 lbf) and burns to depletion in approximately 45 seconds. On-orbit operations are carried out by an array of four 0.44 N (0.1 lbf) thrusters for roll control, ten 2.0 N (0.45 lbf) thrusters for pitch and yaw control and E/W stationkeeping, and two 22.2 N (5.0 lbf) thrusters for repositioning and reorientation. Four 0.3 N (0.07 lbf) EHTs are used for N/S stationkeeping. The nominal mass of the spacecraft at beginning of life (BOL) is 1005 kg and the dry mass at end of life (EOL) is 830 kg. The difference of 175 kg represents the mass of the propellant for a design life of 7 years.

Satellites launched in the late 1980s and 1990s typically have an integrated propulsion system that use a bipropellant combination of monomethyl hydrazine as fuel and nitrogen tetroxide as oxidizer. The specific impulse is about 300 seconds and fuel margin not used for the apogee maneuver can be applied to stationkeeping. Also, since the apogee engine is restartable, it can be used for perigee velocity augmentation and supersynchronous transfer orbit scenarios that optimize the combined propulsion capabilities of the launch vehicle and the spacecraft.

For example, the INTELSAT VII satellite, built by Space Systems/Loral, has a Marquardt 490 N apogee thruster and an array of twelve 22 N stationkeeping thrusters manufactured by Atlantic Research Corporation with a 150:1 columbium nozzle expansion ratio and a specific impulse of 235 seconds. For an Ariane launch the separated mass in GTO is 3610 kg, the mass at BOL is 2100 kg, and the mass at EOL is 1450 kg. The mission life is approximately 17 years.

The Hughes HS-601 satellite has a similar thruster configuration. The mass is approximately 2970 kg at launch, 1680 kg at BOL, and 1300 kg for a nominal 14 year mission.

An interesting problem is the estimation of fuel remaining on the spacecraft at any given time during the mission life. This information is used to predict the satellite end of life. There are no “fuel gauges” so the fuel mass must be determined indirectly. There are three principal methods. The first is called the “gas law” method, which is based on the equation of state of an ideal gas. The pressure and temperature of the inert gas in the propellant tanks is measured by transducers and the volume of the gas is computed knowing precisely the pressure and temperature at launch. The volume of the remaining propellant can thus be deduced and the mass determined from the known density as a function of temperature. Corrections must be applied for the expansion of the tanks and the propellant vapor pressure. The second method is called the “bookkeeping” method. In this method the thruster time for each maneuver is carefully measured and recorded. The propellant consumed is then calculated from mass flow rate expressed in terms of the pressure using an empirical model. The third method is much more sophisticated and is based on the measured dynamics of the spacecraft after a stationkeeping maneuver to determine its total mass. In general, these three independent methods provide redundant information that can be applied to check one another.

NEW TECHNOLOGIES

Several innovative technologies have substantially improved the fuel efficiency of satellite stationkeeping thrusters. The savings in fuel can be used to increase the available payload mass, prolong the mission life, or reduce the mass of the spacecraft.

The first of these developments is the electric rocket arcjet technology. The arcjet system uses an electric arc to superheat hydrazine fuel, which nearly doubles its efficiency. An arcjet thruster has a specific impulse of over 500 seconds. Typical thrust levels are from 0.20 to 0.25 N. The arcjet concept was developed by the NASA Lewis Research Center in Cleveland and thrusters have been manufactured commercially by Primex Technologies, a subsidiary of the Olin Corporation.

AT&T’s Telstar 401 satellite, launched in December 1993 (and subsequently lost in 1997 due to an electrical failure generally attributed to a solar flare) was the first satellite to use arcjets. The stationkeeping propellant requirement was reduced by about 40 percent, which was critical to the selection of the Atlas IIAS launch vehicle. Similar arcjet systems are used on INTELSAT VIII and the Lockheed Martin A2100 series of satellites. INTELSAT VIII, for example, has a dual mode propulsion system incorporating a bipropellant liquid apogee engine that burns hydrazine and oxidizer for orbit insertion and four arcjets that use

monopropellant hydrazine in the reaction control subsystem for stationkeeping.

Electrothermal hydrazine thrusters continue to have applications on various geostationary satellites and on some small spacecraft where maneuvering time is critical. For example, EHTs are used on the IRIDIUM satellites built by Lockheed Martin.

The most exciting development has been in the field of ion propulsion. The propellant is xenon gas. Although the thrust is small and on the order of a few millinewtons, the specific impulse is from 2000 to 4000 seconds, which is about ten to twenty times the efficiency of conventional bipropellant stationkeeping thrusters. Also, the lower thrust levels have the virtue of minimizing attitude disturbances during stationkeeping maneuvers.

The xenon ion propulsion system, or XIPS (pronounced "zips"), is a gridded ion thruster developed by Hughes. This system is available on the HS-601 HP (high power) and HS-702 satellite models and allows for a reduction in propellant mass of up to 90 percent for a 12 to 15 year mission life. A typical satellite has four XIPS thrusters, including two primary thrusters and two redundant thrusters.

Xenon atoms, an inert monatomic gas with the highest molar mass (131 kg/kmol), are introduced into a thruster chamber ringed by magnets. Electrons emitted by a cathode knock off electrons from the xenon atoms and form positive xenon ions. The ions are accelerated by a pair of gridded electrodes, one with a high positive voltage and one with a negative voltage, at the far end of the thrust chamber and create more than 3000 tiny beams. The beams are neutralized by a flux of electrons emitted by a device called the neutralizer to prevent the ions from being electrically attracted back to the thruster and to prevent a space charge from building up around the satellite.

The increase in kinetic energy of the ions is equal to the work done by the electric field, so that

$$\tfrac{1}{2}\, m\, v^2 = q\, V$$

where q, m, and v are the charge, mass, and velocity of the ions and V is the accelerating voltage, equal to the algebraic difference between the positive voltage on the positive grid and the negative voltage on the neutralizer. The charge to mass ratio of xenon ions is 7.35×10^5 C/kg.

The HS-601 HP satellite uses 13-centimeter diameter XIPS engines to perform north-south stationkeeping and to assist the spacecraft's gimballed momentum wheel for roll and yaw control. The accelerating voltage is about 750 volts and the ions have a velocity of 33,600 m/s. The specific impulse is 3400 seconds with a mass flow rate of 0.6 mg/s and
18 mN of thrust. Each ion thruster operates for approximately 5 hours per day and uses 500 W from the available 8 kW total spacecraft power.

The HS-702 spacecraft has higher power
25-centimeter thrusters to perform all stationkeeping maneuvers and to complement the four momentum wheels arranged in a tetrahedron configuration for attitude control. The accelerating voltage is 1200 volts, which produces an ion beam with a velocity of 42,500 m/s. The specific impulse is 4300 seconds, the mass flow rate is 4 mg/s, and the thrust is 165 mN. Each HS-702 ion thruster operates for approximately 30 minutes per day and requires 4.5 kW from the 10 to 15 kW solar array. The stationkeeping strategy maintains a tolerance of ± 0.005° that allows for the collocation of several satellites at a single orbital slot.

The HS-702 satellite has a launch mass of up to 5200 kg and an available payload mass of up to 1200 kg. The spacecraft can carry up to 118 transponders, comprising 94 active amplifiers and 24 spares. A bipropellant propulsion system is used for orbit acquisition, with a fuel capacity of 1750 kg. The XIPS thrusters need only 5 kg of xenon propellant per year, a fraction of the requirement for conventional bipropellant or arcjet systems. The HS-702 also has the option of using XIPS thrusters for orbit raising in transfer orbit to further reduce the required propellant mass budget.

The first commercial satellite to use ion propulsion was PAS-5, which was delivered to the PanAmSat Corporation in August 1997. PAS-5 was the first HS-601 HP model, whose xenon ion propulsion system, together with gallium arsenside solar cells and advanced battery performance, permitted the satellite to accommodate a payload twice as powerful as earlier HS-601 models while maintaining a 15 year orbital life. Four more Hughes satellites with XIPS technology were in orbit by the end of 1998. In addition, Hughes also produced a 30-centimeter xenon ion engine for NASA's Deep Space 1 spacecraft, launched in October 1998.

Another type of ion thruster is the Hall effect ion thruster. The ions are accelerated along the axis of the thruster by crossed electric and magnetic fields. A plasma of electrons in the thrust chamber produces the electric field. A set of coils creates the magnetic field, whose magnitude is the most difficult aspect of the system to adjust. The ions attain a speed of between 15,000 and 20,000 m/s and the specific impulse is about 1800 seconds. This type of thruster has been flown on several Russian spacecraft.

SUMMARY

The demand for ever increasing satellite payloads has motivated the development of propulsion systems with greater efficiency. Typical satellites of fifteen to twenty years ago had solid apogee motors and simple monopropellant hydrazine stationkeeping thrusters. Electrically heated thrusters were designed to increase the hydrazine performance and the principle was further advanced by the innovation of the arcjet thruster. Bipropellant systems are now commonly used for increased performance and versatility.

The future will see a steady transition to ion propulsion. The improvements in fuel efficiency permit the savings in mass to be used for increasing the revenue-generating payloads (with attendant increase in solar arrays, batteries, and thermal control systems to power them), extending the lifetimes in orbit, or reducing the spacecraft mass to permit a more economical launch vehicle.

Dr. Robert A. Nelson, P.E. is president of Satellite Engineering Research Corporation, a satellite engineering consulting firm in Bethesda, Maryland. Dr. Nelson is *Via Satellite's* Technical Editor.

Via Satellite March, 1995

Satellite Constellation Geometry

by Robert A. Nelson

Satellite constellation geometry has been studied as a theoretical problem since the early 1960s. The analysis originally had limited application to photographic reconnaissance and Earth resource missions. However, at present it has achieved particular relevance for the numerous satellite systems under development that offer a variety of new services, including mobile telephony, message and data transfer, and position determination. The problem combines the physics of orbits with optimization of the orbit geometry so as to provide the required Earth coverage while minimizing the number of satellites.

Notable contributions to the theory of constellation geometry have been made by Walker, Draim, Ballard, and Adams and Rider. Walker made an extensive study and found that at least five satellites are required for continuous global coverage from circular orbits at a common altitude and inclination. His method of classifying constellation types with the notation $T/P/F$ is frequently used, where T is the total number of satellites, P is the number of evenly spaced orbital planes and F determines the phase spacing between adjacent planes. Draim found that continuous coverage could be attained by only four satellites in elliptical orbits. Ballard also studied the optimization of satellites in inclined circular orbits, which he called "rosette constellations," using a satellite triad approach. This method minimizes the largest distance between the observation point and any subsatellite point. Adams and Rider deduced the optimum configurations for polar orbit constellations for single or multiple satellite levels of coverage over the entire Earth or above a specified latitude, using a street-of-coverage approach. This method considers a ground swath that is continuously covered.

ALTITUDE

The altitude of the satellite orbit is the primary characteristic of the satellite constellation. It is chosen on the basis of both physical and geometric considerations, including signal propagation delay, signal power, avoidance of the Van Allen radiation belts, time of satellite visibility and coverage area.

The altitude regimes have been divided by convention into Low Earth Orbit (LEO), Medium Earth Orbit (MEO) and Geostationary Orbit (GEO). The altitude of LEO is roughly between 500 krn and 1,500 km. The lower bound is chosen to avoid excessive stationkeeping due to residual atmospheric drag. The upper bound is chosen so as to lie below the first Van Allen radiation belt. The altitude of MEO can be taken to be approximately within the range 5,000 krn to 15,000 krn so as to be within the first and second Van Allen belts. The limits are ten times those of LEO. The altitude of GEO is uniquely 35,786 km in the equatorial plane. At this altitude the period of revolution is exactly equal to the period of rotation of the Earth (23.934 h), so that a satellite appears to remain over a fixed point on the equator. A fourth orbit category is the highly elliptical orbit (HEO), in which the apogee may be beyond the geostationary orbit.

The two principal factors that have created interest in LEO and MEO for satellite communications are the low signal propagation delay and the limitations on gain and power of the Earth terminal. The round trip signal delay for a two-way conversation via satellite at an altitude of 10,000 km is 130 ms, and for a satellite at an altitude of 1,000 km it is only 13 ms. In contrast, the propagation delay from GEO for a two-way conversation is over half a second, which is distracting at best and can be intolerable for digital data transmission using error correcting protocols that require retransmission of blocks with detected errors.

Handheld telephones by their nature must have low gain (on the order of 1 dB) because they must be omni-directional and have fixed power limits (on the order of 350 mW) to safeguard human health. The Earth terminal gain and power determine the required size of the satellite antenna, which must be large enough to provide sufficient link margin. Also, the bandwidth available is limited, so the total coverage area is usually divided into a cellular pattern of spot beams to permit frequency reuse. The cell size is determined by the size of the antenna and the orbit altitude. As the orbit gets higher, it is necessary to use a larger antenna on the spacecraft to achieve a given spot size on the Earth. For example, at L-band (1615 MHz), a 17 meter spacecraft antenna in GEO would be required for the same cell size as a 0.5 meter antenna in LEO. Thus, LEO and MEO are preferable to GEO for mobile hand-held telephony.

Other considerations that affect the choice of altitude are satellite visibility and eclipse time. At Low Earth Orbit the period of revolution is approximately 100 minutes. For a typical pass, the satellite is visible for only about ten minutes. Thus, frequent handover is required for mobile telephony. In addition, during times of the year when the orbital plane is parallel to the direction to the Sun, the satellite is eclipsed for about 30 minutes, or about one third of the orbital period. Consequently, there is a significant demand on battery power, with up to 5,000 charge/discharge cycles per year in Low Earth Orbit. With present nickel-hydrogen battery technology, a battery rated for 10 to 15 years in GEO would have a life of about 5 years in LEO. On the other hand, in Medium Earth Orbit the orbital period is six to eight hours and the time of visibility of a single satellite is over an hour. There are fewer eclipse cycles and battery lifetime is typically seven years.

INCLINATION

The second fundamental parameter of a satellite constellation is its orbital inclination. The choice is governed by the global coverage requirement, the level of coverage, and the minimum angle of elevation. Inclinations of direct circular orbits are generally around 50°. This inclination permits coverage of temperate zones and populated regions of the Earth, while allowing more than one satellite to be visible from a given point for reasonable constellation sizes. Polar constellations have inclinations near 90°, which permits global coverage with the fewest satellites. Retrograde orbits (such

as Sun-synchronous orbits) have inclinations greater than 90°.

A great advantage of inclined or polar LEO and MEO constellations is that they afford high angles of elevation. Elevation angles of from 20° to 40° may be required to avoid blockage from tall buildings in urban areas. These angles are not possible from GEO, even at moderate latitudes of 45°. Many of the capitals of Europe, including Paris, London, Berlin, Warsaw and Moscow, are north of this latitude. Furthermore, a geostationary satellite is below the horizon if the latitude is greater than 81°.

ECCENTRICITY

The third important orbital parameter is the eccentricity, which determines the orbit's shape. For circular orbits, the eccentricity is zero and the satellite moves at uniform speed. For elliptical orbits, however, the eccentricity is between 0 and 1. The satellite moves fastest at perigee, or the point closest to the Earth, and slowest at apogee, or the point farthest from the Earth. By adjusting the position of the apogee, the dwell time of the satellite can be maximized over the region of interest.

Earth oblateness perturbations restrict the inclination of elliptical orbits to 63.4° or 116.6° for satellite communications. These are the only two inclinations at which the major axis remains fixed, so that the apogee remains over the specified latitude. At all other inclinations the gravitational harmonics of the Earth due to its oblate shape cause the major axis to rotate. For example, the Russian 12 hour Molniya orbit is a highly elliptical orbit inclined at 63.4°. The perigee altitude is 1,006 km and the apogee altitude is 39,362 km with apogee over the northern hemisphere. A Molniya satellite spends nearly 11 hours over the northern hemisphere and only 1 hour over the southern hemisphere per revolution.

CONSTELLATION CONFIGURATION

The configuration of the constellation is defined by the number of orbital planes p and the number of satellites per plane s. The values of p and s should be chosen so as to minimize the total number of satellites N that are required to provide the specified level of coverage, where $N = p\,s$.

For a given minimum angle of elevation θ, the angle γ with respect to the Earth's center between the subsate1lite point and edge of coverage is given by

$$\gamma = \arccos\left(\frac{\cos\theta}{1+h/R_E}\right) - \theta$$

where h is the satellite altitude and R_E is the radius of the Earth (6,378 km). The total coverage area may be estimated from the formula

$$S = 2\pi R_E^{\,2}(1-\cos\gamma)$$

Ideally, S should be as large as possible, but it is usually subdivided into an array of cells to permit frequency reuse. It may be limited by the required satellite antenna gain, which is approximately given by $G = 4\pi h^2 (n/S)$, where n is the number of cells. The diameter D of the antenna is then given by $G = \eta(\pi D/\lambda)^2$, where λ is the wavelength and η is the efficiency. These relations imply that the antenna diameter is proportional to the altitude for a given cell size.

The coverage geometry problem is simplest for polar constellations. For global coverage with optimum phasing, the point of intersection of overlapping circles of coverage in one plane coincides with the boundary of a circle of coverage in a neighboring plane. Satellites in adjacent planes revolve in the same direction. However, there is a "seam" in the constellation pattern between the first and last planes, where the satellites revolve in opposite directions. For a given number of planes p and number of satellites per plane s, the Earth central angle γ and ground swath half-width Γ are determined by the equations

$$\cos\Gamma = \frac{\cos\gamma}{\cos(\pi/s)}$$

and

$$(p-1)\,\alpha + \beta = \pi$$

where $\alpha = \Gamma + \gamma$ is the spacing between co-rotating planes and $\beta = 2\,\Gamma$ is the spacing between counter-rotating planes.

For example, the original Iridium constellation, based on a paper by Adams and Rider, consisted of 77 satellites distributed into seven planes with 11 satellites per plane. (The constellation was named after the element iridium, whose atomic structure consists of 77 electrons orbiting the nucleus.) Therefore, the Earth central angle γ was 18.5° and the ground swath half-width Γ was 8.6°. Also, α was 27.1° and β was 17.2°. For a minimum elevation of 10° at edge of coverage, the corresponding altitude was 765 km. This altitude satisfied the constraints that it was sufficiently high that atmospheric drag was negligible, it was sufficiently low that it avoided the Van Allen radiation environment, and the cost of satellite deployment was moderate.

The Iridiurn constellation has been revised by the elimination of one plane to reduce the number of satellites. It now consists of 66 satellites distributed into six planes with 11 satellites per plane at an altitude of 780 km. The plane separation is 31.6° and the orbital inclination has been changed to 86° as a precaution against collisions at the poles. The angle of elevation at edge of coverage on the equator is 8.2°.

The geometric problem for inclined constellations is somewhat more complicated. Walker, Ballard and Rider have examined this problem using a variety of assumptions and techniques. For example, the Globalstar constellation consists of 48 satellites, with 6 satellites in each of 8 orbital planes, at an altitude of 1406 km and inclined at 52°. This constellation was based on the Walker 48/8/1 "delta" pattern and was refined by computer modeling. A basic requirement of this system is that two satellites must be visible from any point. The communications link uses code division multiple access (CDMA) with path diversity. Each mobile telephone receives a signal from each of two satellites at half power to minimize blockage and multipath effects.

EARTH OBLATENESS

Earth oblatness has two important effects on a orbit. First, as mentioned previously, it causes the major axis to rotate. The rate of change of the perigee angle is

$$\frac{d\omega}{dt} = \frac{4.982}{(1-e^2)^2}\left(\frac{R_E}{a}\right)^{3.5}(5\cos^2 i - 1)$$

expressed in degrees per day, where a is the semimajor axis, e is the eccentricity and i is the inclination. This equation

implies that the major axis is stable only for inclinations of 63.4° and 116.6°, which are the only angles that make the right hand side of the equation equal to zero.

Oblateness also causes the ascending node of the orbit to drift. The rate of drift is given by the formula

$$\frac{d\Omega}{dt} = -\frac{9.964}{(1-e^2)^2}\left(\frac{R_E}{a}\right)^{3.5}\cos i$$

expressed in degrees per day. For inclinations less than 90° the ascending node drifts westward, while for inclinations greater than 90° the ascending node drifts eastward. The ascending node does not drift for polar constellations, for which the inclination is 90°. For example, for the Globalstar constellation, the ascending node drifts westward at the rate of 3° per day.

The operational impact of ascending node drift on LEO constellations with intermediate inclinations is the penalty on stationkeeping fuel. In principle, if all the satellites had identical circular orbit altitudes and inclinations, the orbit planes would drift in unison and the relative geometry would remain constant. However, in practice, there are inevitable orbit insertion errors during deployment. In the preceding example, the difference in ascending nodes would accumulate to 0.5° in one year for each kilometer of error in altitude and would accumulate to 2.5° in one year for each 0.1° of error in inclination, compared to the nominal orbit.

SUN-SYNCHRONOUS ORBITS

The drift in ascending node has one important practical application. If the altitude, inclination and eccentricity are chosen so that the ascending node drifts eastward at the same rate as the Earth revolves around the Sun (0.9856° per day), then the Earth–Sun line would maintain a constant orientation with respect to the orbital plane. This type of orbit was first used by the Landsat satellites for Earth photography missions. Landsat-1 was launched in July 1972 into a 910 km altitude orbit inclined at 99°.

If the orbital plane is initially oriented perpendicular to the direction of the Sun, the satellite will always remain illuminated. The solar array would not require a tracking mechanism and batteries would be needed only for contingencies. Another advantage of Sun-synchronous orbits is that the orbital period can be synchronous with the mean solar day instead of the sidereal day over a given point on Earth, so that the satellite maintains the same time-of-day schedule.

The E-Sat satellite system provides an example based on these considerations. This system will provide data messaging and data retrieval services for public utilities and petroleum companies, direct-to-home television broadcast services and the financial services industry. The satellite orbit, is a Sun-synchronous circular orbit with a period of revolution that is a submultiple of a mean solar day. It has thus been given the name of "doubly-synchronous orbit." Since the orbit is Sun-synchronous, the satellite maintains the same time-of-day schedule. The orbital plane is to be oriented perpendicular to the Earth-Sun line and the satellite solar array will be constantly illuminated. The ground trace will repeat itself every day. It was also required that the altitude must be within the range 1,000 km < h < 1,500 km so that the atmospheric drag would be negligible and would not impinge on the first Van Allen radiation belt. Therefore, a circular orbit with an altitude of 1,262 km and an inclination of 100.7° was chosen.

For a minimum elevation angle of 20°, the Earth central angle γ is 18.3°. The corresponding coverage area is 13 million square kilometers, or roughly the size of CONUS. This coverage area implies that the maximum satellite gain must be 1.9 dB at 149.5 MHz. Therefore, for the given Earth terminal power and gain, method of modulation and coding, and various losses, the maximum data rate that can be supported by the communications link is determined. Three satellites will be deployed into one plane to meet the required capacity of the anticipated market. An additional three satellites may be added at a later time. If the latter satellites are deployed into a different orbit plane, they will have the required modifications to the electrical power subsystem to permit solar tracking and accommodate eclipse periods.

A satellite orbit based on similar considerations was proposed in a 1984 NASA-Lewis study for the Voice of America as one of several concepts for a direct broadcast satellite system. In this case, elliptical Sun-synchronous orbits with an integral number of revolutions per mean solar day were investigated. An e1liptical orbit was considered because it would provide a long dwell time over the region to be covered with proper positioning of the apogee. Since the major axis could not rotate and since the inclination of a Sun-synchronous orbit must be greater than 90°, the inclination of 116.6° was required. With this additional level of synchronism, the orbit was given the name "triply-synchronous orbit." The only orbit with an integral number of revolutions per day that does not intersect the Earth is the three hour orbit, with a perigee altitude of 521 km and an apogee altitude of 7,843 km. An identical orbit concept has been adopted by Mobile Communications Holdings, Inc. (MCHI) for thc "Borealis orbit" of its proposed Ellipsat constellation.

NAVIGATION SATELLITES

The Global Positioning System (GPS) is a fully operational satellite system for high precision position determination developed by the U.S. Department of Defense. The GPS constellation consists of 21 operational satellites and three in-orbit spares in circular orbits at an altitude of 20,182 km. The orbital period is one-half a sidereal day, or 11.967 hours. The ground track repeats itself every two revolutions, with the result that a given satellite appears over the same point 4.1 minutes earlier than the previous day. Four satellites are deployed into each of six orbital planes inclined at 55°. At least four satellites are visible at all times from any point on Earth.

Each satellite carries two cesium and two rubidium atomic clocks that maintain a highly stable time and frequency reference. The satellite orbit and clock information is transmitted on each of two L-band carriers (1575.42 MHz and 1227.60 MHz). Two frequencies are used to measure and compensate for the effect of ionosphere and troposphere delay. The baseband signal is modulated by two spread-spectrum pseudorandom noise codes: a precision (P) code at 10.23 Mbps for military use that repeats every 38 weeks and a clear access (C/A) code at 1.023 Mbps for satellite acquisition and civilian use that repeats every 1 ms. Different satellites use different portions

of the same P code. The user's receiver generates an identical code and measures the distance to the satellite by means of an autocorrelation circuit that determines the phase difference needed to align the two codes. The simultaneous measurement of PRN signals from four satellites permits a three-dimensional determination of position with a resolution of better than 10 meters using the P code or between 100 meters and 300 meters with the C/A code. GPS satellites are also used for time comparison between standards laboratories by common view measurements with a precision of a few nanoseconds.

The Russian Global Orbital Navigation Satellite System (GLONASS) is a similar system under development, consisting of 24 satellites at an altitude of 19,132 km evenly distributed into three orbital planes inclined at 64.8°. The orbital period is 11.263 hours, so the ground track repeats itself every eight days. In contrast to GPS, which uses only two frequencies for the entire system, each GLONASS satellite is assigned its own two frequencies within the bands 1240 - 1260 MHz and 1597 -1617 MHz. Satellites are distinguished by radio-frequency channel instead of by pseudorandom noise code. A single code is used, repeating every 1 ms.

CONCLUSION

The basic principles of satellite constellation design have been reviewed and several actual examples have been described. These examples illustrate how various design considerations lead to the choice of orbit, which then drives the choice of link parameters to meet the system requirements.

Via Satellite, September 1998

Iridium: From Concept to Reality

by Robert A. Nelson

On the 23rd day of this month, a revolutionary communication system will begin service to the public. Iridium will be the first mobile telephony system to offer voice and data services to and from handheld telephones anywhere in the world. Industry analysts have eagerly awaited this event, as they have debated the nature of the market, the economics, and the technical design.

As with any complex engineering system, credit must be shared among many people. However, the three key individuals who are recognized as having conceived and designed the system are Bary Bertiger, Dr. Raymond Leopold, and Kenneth Peterson of Motorola, creators of the Iridium system.

The inspiration was an occasion that has entered into the folklore of Motorola. (The story, as recounted here, was the subject of a *Wall Street Journal* profile on Monday, December 16, 1996.) On a vacation to the Bahamas in 1985, Bertiger's wife, Karen, wanted to place a cellular telephone call back to her home near the Motorola facility in Chandler, AZ to close a real-estate transaction. After attempting to make the connection without success, she asked Bertiger why it wouldn't be possible to create a telephone system that would work anywhere, even in the remote Caribbean outback.

Bertiger took the problem back to colleagues Leopold and Peterson at Motorola. Numerous alternative terrestrial designs were discussed and abandoned.

In 1987 research began on a constellation of low earth orbiting satellites that could communicate directly with telephones on the ground and with one another -- a kind of inverted cellular telephone system.

But as they left work one day in 1988, Leopold proposed a crucial element of the design. The satellites would be coordinated by a network of "gateway" earth stations connecting the satellite system to existing telephone systems. They quickly agreed that this was the sought-after solution and immediately wrote down an outline using the nearest available medium -- a whiteboard in a security guard's office.

Originally, the constellation was to have consisted of 77 satellites. The constellation was based on a study by William S. Adams and Leonard Rider of the Aerospace Corporation, who published a paper in *The Journal of the Astronautical Sciences* in 1987 on the configurations of circular, polar satellite constellations at various altitudes providing continuous, full-earth coverage with a minimum number of satellites. However, by 1992 several modifications had been made to the system, including a reduction in the number of satellites from 77 to 66 by the elimination of one orbital plane.

The name Iridium was suggested by a Motorola cellular telephone system engineer, Jim Williams, from the Motorola facility near Chicago. The 77-satellite constellation reminded him of the electrons that encircle the nucleus in the classical Bohr model of the atom. When he consulted the periodic table of the elements to discover which atom had 77 electrons, he found Iridium -- a creative name that has a nice ring. Fortunately, the system had not yet been scaled back to 66 satellites, or else he might have suggested the name Dysprosium.

The project was not adopted by senior management immediately. On a visit to the Chandler facility, however, Motorola chairman Robert Galvin learned of the idea and was briefed by Bertiger. Galvin at once endorsed the plan and at a subsequent meeting persuaded Motorola's president John Mitchell. Ten years have elapsed from this go-ahead decision, and thirteen years since Bertiger's wife posed the question.

In December 1997 the first Iridium test call was delivered by orbiting satellites. Shortly after completion of the constellation in May 1998, a demonstration was conducted for franchise owners and guests. The new system was ready for operation, and Iridium is now on the threshold of beginning service.

REGULATORY HURDLES

In June, 1990 Motorola announced the development of its Iridium satellite system at simultaneous press conferences in Beijing, London, Melbourne, and New York. The Iridium system was described in an application to the Federal Communications Commission (FCC) filed in December of that year, in a supplement of February 1991, and an amendment in August 1992.

At the time, an internationally allocated spectrum for this service by nongeostationary satellites did not even exist. Thus Motorola proposed to offer Radio Determination Satellite Service (RDSS) in addition to mobile digital voice and data communication so that it might qualify for use of available spectrum in the RDSS L-band from 1610 to 1626.5 MHz. A waiver was requested to provide both two-way digital voice and data services on a co-primary basis with RDSS.

Following the submission of Motorola's Iridium proposal, the FCC invited applications from other companies for systems to share this band for the new Mobile Satellite Service (MSS). An additional four proposals for nongeostationary mobile telephony systems were submitted to meet the June 3, 1991 deadline, including Loral/Qualcomm's Globalstar, TRW's Odyssey, MCHI's Ellipsat, and Constellation Communications' Aries. Collectively, these nongeostationary satellite systems became known as the "Big LEOs". The American Mobile Satellite Corporation (AMSC) also sought to expand existing spectrum for its geostationary satellite into the RDSS band.

At the 1992 World Administrative Radio Conference (WARC-92) in Torremolinos, Spain, L-band spectrum from 1610 to 1626.5 MHz was internationally allocated for MSS for earth-to-space (uplink) on a primary basis in all three ITU regions. WARC-92 also allocated to MSS the band 1613.8 to 1626.5 MHz on a secondary basis and

spectrum in S-band from 2483.5 to 2500 MHz on a primary basis for space-to-earth (downlink).

In early 1993 the FCC adopted a conforming domestic spectrum allocation and convened a Negotiated Rulemaking proceeding. This series of meetings was attended in Washington, DC by representatives of the six applicants and Celsat, which had expressed an intention to file an application for a geostationary satellite but did not meet the deadline.

The purpose of the proceeding was to provide the companies with the opportunity to devise a frequency- sharing plan and make recommendations. These deliberations were lively, and at times contentious, as Motorola defended its FDMA/TDMA multiple access design against the CDMA technologies of the other participants.

With frequency division multiple access (FDMA), the available spectrum is subdivided into smaller bands allocated to individual users. Iridium extends this multiple access scheme further by using time division multiple access (TDMA) within each FDMA sub-band. Each user is assigned two time slots -- one for sending and one for receiving -- within a repetitive time frame. During each time slot, the digital data are burst between the mobile handset and the satellite.

With code division multiple access (CDMA), the signal from each user is modulated by a pseudorandom noise (PRN) code. All users share the same spectrum. At the receiver, the desired signal is extracted from the entire population of signals by multiplying by a replica code and performing an autocorrelation process. The key to the success of this method is the existence of sufficient PRN codes that appear to be mathematically orthogonal to one another. Major advantages cited by CDMA proponents are inherently greater capacity and higher spectral efficiency. Frequency reuse clusters can be smaller because interference is reduced between neighboring cells.

In April, 1993 a majority report of Working Group 1 of the Negotiated Rulemaking Committee recommended full band sharing across the entire MSS band by all systems including Iridium. Coordination would be based on an equitable allocation of interference noise produced by each system. The FDMA/TDMA system would be assigned one circular polarization and the CDMA systems would be assigned the opposite polarization. This approach required that each system would be designed with sufficient margin to tolerate the level of interference received from other licensed systems.

Motorola issued a minority report which stated that the Iridium system must have its own spectrum allocation. It proposed partitioning of the MSS L-band spectrum into two equal 8.25 MHz segments for the FDMA/TDMA and CDMA access technologies, with the upper portion being used by the FDMA/TDMA system where it would be sufficiently isolated from neighboring frequencies used by radio astronomy, GPS, and Glonass.

Faced with this impasse, the FCC in January 1994 adopted rulemaking proposals which allocated the upper 5.15 MHz of the MSS L-band spectrum to the sole FDMA/TDMA applicant, Iridium, and assigned the remaining 11.35 MHz to be shared by multiple CDMA systems. However, if only one CDMA system were implemented, the 11.35 MHz allotment would be reduced to 8.25 MHz, leaving 3.10 MHz available for additional spectrum to Iridium or a new applicant.

The response to the Commission's proposals from the Big LEO applicants was generally favorable. Without this compromise, the alternative would have been to hold a lottery or auction to allocate the spectrum. The Iridium system was designed to operate with the full spectrum allocation. However, with 5.15 MHz, the system is a viable business proposition. The additional 3.10 MHz, should it become available, further adds to the system's attractiveness.

The FCC also proposed that the MSS spectrum could be used only by Low Earth Orbit (LEO) and Medium Earth Orbit (MEO) satellite systems. Therefore, the geostationary orbit (GEO) systems of AMSC and Celsat would not be permitted in this band. To qualify for a Big LEO license, the Commission proposed that the service must be global (excluding the poles) and that companies must meet stringent financial standards.

In October, 1994 the FCC issued its final rules for MSS, closely following language of the January proposed rulemaking. However, it allowed the CDMA systems to share the entire 16.5 MHz of downlink spectrum in S-band. The Commission gave the Big LEO applicants a November 16 deadline to amend their applications to conform to the new licensing rules.

On January 31, 1995 the FCC granted licenses to Iridium, Globalstar, and Odyssey but withheld its decision on Ellipsat and Aries pending an evaluation of their financial qualifications. The latter companies finally received licenses in June last year, while in December TRW dropped its Odyssey system in favor of partnership with ICO, the international subsidiary of Inmarsat which entered the competition in 1995.

Outside the United States, Iridium must obtain access rights in each country where service is provided. The company expects to have reached agreements with 90 priority countries that represent 85% of its business plan by the start of service this month. Altogether, Iridium is seeking access to some 200 countries through an arduous negotiating process.

FINANCING

Iridium LLC was established by Motorola in December, 1991 to build and operate the Iridium system, with Robert W. Kinzie as its chairman. In December, 1996 Edward F. Staiano was appointed Vice Chairman and CEO.

Iridium LLC, based in Washington, DC, is a 19-member international consortium of strategic investors representing telecommuni-cation and industrial companies, including a 25 percent stake by its prime contractor, Motorola, Inc.

In August 1993, Motorola and Iridium LLC announced they had completed the first-round financing of the Iridium system with $800 million in equity. The second round was completed in September, 1994, bringing the total to $1.6 billion. In July of last year $800 million in debt financing was completed. Iridium World Communications, Ltd., a Bermuda company, was formed to serve as a vehicle

for public investment in the Iridium system. In June 1997 an initial $240 million public offering was made on the NASDAQ Stock Exchange.

TECHNICAL DESCRIPTION

The Iridium constellation consists of 66 satellites in near-polar circular orbits inclined at 86.4° at an altitude of 780 km. The satellites are distributed into six planes separated by 31.6° around the equator with eleven satellites per plane. There is also one spare satellite in each plane.

Starting on May 5, 1997, the entire constellation was deployed within twelve months on launch vehicles from three continents: the U.S. Delta II, the Russian Proton, and the Chinese Long March. The final complement of five 700 kg (1500 lb) satellites was launched aboard a Delta II rocket on May 17. With a satellite lifetime of from 5 to 8 years, it is expected that the replenishment rate will be about a dozen satellites per year after the second year of operation.

The altitude was specified to be within the range 370 km (200 nmi) and 1100 km (600 nmi). The engineers wanted a minimum altitude of 370 km so that the satellite would be above the residual atmosphere, which would have diminished lifetime without extensive stationkeeping, and a maximum altitude of 1100 km so that the satellite would be below the Van Allen radiation environment, which would require shielding.

Each satellite covers a circular area roughly the size of the United States with a diameter of about 4400 km, having an elevation angle of 8.2° at the perimeter and subtending an angle of 39.8° with respect to the center of the earth. The coverage area is divided into 48 cells. The satellite has three main beam phased array antennas, each of which serves 16 cells.

The period of revolution is approximately 100 minutes, so that a given satellite is in view about 9 minutes. The user is illuminated by a single cell for about one minute. Complex protocols are required to provide continuity of communication seamlessly as handover is passed from cell to cell and from satellite to satellite. The communications link requires 3.5 million lines of software, while an additional 14 million lines of code are required for navigation and switching. As satellites converge near the poles, redundant beams are shut off. There are approximately 2150 active beams over the globe.

The total spectrum of 5.15 MHz is divided into 120 FDMA channels, each with a bandwidth of 31.5 kHz and a guardband of 10.17 kHz to minimize intermodulation effects and two guardbands of 37.5 kHz to allow for Doppler frequency shifts. Within each FDMA channel, there are four TDMA slots in each direction (uplink and downlink). The coded data burst rate with QPSK modulation and raised cosine filtering is 50 kbps (corresponding to an occupied bandwidth of 1.26×50 kbps / 2 = 31.5 kHz). Each TDMA slot has length 8.29 ms in a 90 ms frame. The supported vocoder information bit rate is 2.4 kbps for digital voice, fax, and data. The total information bit rate, with rate 3/4 forward error correction (FEC) coding, is 3.45 kbps ($0.75 \times (8.28$ ms/90 ms$) \times 50$ kbps = 3.45 kbps), which includes overhead and source encoding, exclusive of FEC coding, for weighting of parameters in importance of decoding the signal. The bit error ratio (BER) at threshold is nominally 0.01 but is much better 99 percent of the time.

The vocoder is analogous to a musical instrument synthesizer. In this case, the "instrument" is the human vocal tract. Instead of performing analogue-to-digital conversion using pulse code modulation (PCM) with a nominal data rate of 64 kbps (typical of terrestrial toll-quality telephone circuits), the vocoder transmits a set of parameters that emulate speech patterns, vowel sounds, and acoustic level. The resulting bit rate of 2.4 kbps is thus capable of transmitting clear, intelligible speech comparable to the performance of high quality terrestrial cellular telephones, but not quite the quality of standard telephones.

The signal strength has a nominal 16 dB link margin. This margin is robust for users in exterior urban environments, but is not sufficient to penetrate buildings. Satellite users will have to stand near windows or go outside to place a call. Handover from cell to cell within the field of view of an orbiting satellite is imperceptible. Handover from satellite to satellite every nine minutes may occasionally be detectable by a quarter-second gap.

Each satellite has a capacity of about 1100 channels. However, the actual number of users within a satellite coverage area will vary and the distribution of traffic among cells is not symmetrical.

CALL ROUTING

The Iridium satellites are processing satellites that route a call through the satellite constellation. The system is coordinated by 12 physical gateways distributed around the world, although in principle only a single gateway would be required for complete global coverage. Intersatellite links operate in Ka-band from 23.18 to 23.38 GHz and satellite-gateway links operate in Ka-band at 29.1 to 29.3 GHz (uplink) and 19.4 to 19.6 GHz (downlink).

For example, a gateway in Tempe, Arizona serves North America and Central America; a gateway in Italy serves Europe and Africa; a gateway in India serves southern Asia and Australia. There are 15 regional franchise owners, some of whom share gateway facilities. The constellation is managed from a new satellite network operations center in Lansdowne, Virginia.

As described by Craig Bond, Iridium's vice president for marketing development, the user dials a telephone number with the handset using an international 13 digit number as one would do normally using a standard telephone. The user presses the "send" button to access the nearest satellite. The system identifies the user's position and authenticates the handset at the nearest gateway with the home location register (HLR).

Once the user is validated, the call is sent to the satellite. The call is routed through the constellation and drops to the gateway closest to the destination. There it is completed over standard terrestrial circuits.

For a call from a fixed location to a handset, the process is reversed. After the call is placed, the system identifies the recipient's location and the handset rings, no matter where the user is on the earth.

It is projected that about 95 percent of the traffic will be between a mobile

handset and a telephone at a fixed location. The remaining 5 percent of the traffic represents calls placed from one handset to another handset anywhere in the world. In this case, the call "never touches the ground" until it is received by the handset of the intended recipient.

By comparison, a "bent pipe" satellite system, such as Globalstar, requires that a single satellite see both the user and the nearest gateway simultaneously. Thus many more gateways are needed. For example, in Africa Globalstar will require about a dozen gateways, while Iridium has none at all. Globalstar advocates would counter that this is not a disadvantage, since their system places the complexity on the ground rather than the satellite and offers greater flexibility in building and upgrading the system.

HANDSET

The Iridium handsets are built by Motorola and Kyocera, a leading manufacturer of cellular telephones in Japan. Handsets will permit both satellite access and terrestrial cellular roaming capability within the same unit. The unit also includes a Subscriber Identity Module (SIM) card. Major regional cellular standards are interchanged by inserting a Cellular Cassette. Paging options are available, as well as separate compact Iridium pagers.

The price for a typical configuration will be around $3,000. The handsets will be available through service providers and cellular roaming partners. In June, Iridium finalized its 200th local distribution agreement.

Information on how to obtain Iridium telephones will be advertised widely. Customers will also be actively solicited through credit card and travel services memberships. Distribution of the handsets and setup will typically be through sales representatives who will interface with the customer directly. Rental programs will also be available to give potential customers the opportunity to try out the system on a temporary basis.

MARKET

Iridium has conducted extensive research to measure the market. As described by Iridium's Bond, the intended market can be divided into two segments: the vertical market and the horizontal market.

The vertical market consists of customers in remote areas who require satellites for their communications needs because they cannot access conventional terrestrial cellular networks. This market includes personnel in the petroleum, gas, mining, and shipping industries. It also includes the branches of the U.S. military. In fact, the U.S. government has built a dedicated gateway in Hawaii capable of serving 120,000 users so that it can access the Iridium system at a lower per minute charge.

The horizontal market is represented by the international business traveler. This type of customer wants to keep in contact with the corporate office no matter where he or she is in the world. Although mindful of the satellite link, this customer doesn't really care how the telephone system works, as long as it is always available easily and reliably.

It has been consistently estimated that the total price for satellite service will be about $3.00 per minute. This price is about 25 percent to 35 percent higher than normal cellular roaming rates plus long distance charges. When using the roaming cellular capability, the price will be about $1.00 to $1.25 per minute.

The expected break-even market for Iridium is about 600,000 customers globally, assuming an undisclosed average usage per customer per month. The company hopes to recover its $5 billion investment within one year, or by the fourth quarter of 1999. Based on independent research, Iridium anticipates a customer base of 5 million by 2002.

PROBLEMS

As might be expected for a complex undertaking, the deployment of the constellation and the manufacture of the handsets has not been without glitches. So far, a total of seven spacecraft have suffered in-orbit failures. In addition, Iridium has announced delays in the development of the handset software.

Of the 72 satellites launched, including spares, one lost its stationkeeping fuel when a thruster did not shut off, one was damaged as it was released from a Delta II launch vehicle, and three had reaction wheel problems. In July two more satellites failed because of hardware problems. Delta II and Long March rockets, scheduled to begin a maintenance program of launching additional spares, were retargeted to deploy seven replacement birds to the orbital planes where they are needed in August.

Investors are also nervous about final software upgrades to the handsets. Following alpha trials last month, beta testing of the units was scheduled to commence within one week of the September 23 commercial activation date. The Motorola handsets are expected to be available to meet initial demand, but those made by Kyocera may not be ready until later. [*Note added:* On September 9, Iridium announced that the debut of full commercial service would be delayed until November 1 because more time is needed to test the global system.]

The fifteen gateways have been completed. Equipment for the China gateway, the last one, was shipped recently. Like a theatrical production, the players are frantically completing last minute details as the curtain is about to go up and Iridium embarks upon the world stage.

THE FUTURE

Iridium is already at work on its Next Generation system (Inx). Planning the system has been underway for more than a year. Although details have not been announced, it has been suggested that the system would be capable of providing broadband services to mobile terminals. In part, it would augment the fixed terminal services offered by Teledesic, which Motorola is helping to build, and might include aspects of Motorola's former Celestri system. It has also been reported that the Inx terminal would provide greater flexibility in transitioning between satellite and cellular services and that the satellite power level would be substantially increased.

As customers sign up for satellite mobile telephony service, the utility and competitive advantage will become apparent. Information will flow more freely, the world will grow still smaller,

and economies around the world will be stimulated. There will also be a profound effect on geopolitics and culture. Just as satellite television helped bring down the Berlin Wall by the flow of pictures and information across international boundaries, the dawning age of global personal communication among individuals will bring the world community closer together as a single family.

Dr. Robert A. Nelson, P.E., is president of Satellite Engineering Research Corporation, a satellite engineering consulting firm in Bethesda, MD, and is Technical Editor of *Via Satellite*.

Via Satellite February, 1998

V-Band

Expansion of the Spectrum Frontier

by Robert A. Nelson

The settlement of the American west during the nineteenth century was bounded by a natural frontier: the Pacific Ocean. For pioneers in the satellite industry, there appears to be no analogous frontier in the electromagnetic frequencies used for satellite communications as the upper bound of frequencies is being pushed ever higher.

On September 26, 1997, a dozen companies submitted proposals to the U.S. Federal Communications Commission (FCC) for authorization to build satellite systems that will exploit the frequency bands from 36 GHz to 51.4 GHz, which includes Q-band and V-band. These new systems will supplement the many Ka-band broadband systems now in various stages of development. The proposed constellations span the full range of altitude regimes, including Low Earth Orbit (LEO), Medium Earth Orbit (MEO), and geostationary orbit (GEO).

As historians have noted, the expansion of the American west was made possible through the sudden advance of technology – the steamboat, the telegraph, and the railroad. Similarly, the focus on broadband applications at higher frequencies has been made possible through technological change, including processing satellites, sophisticated switching networks, low-noise amplifiers, modems, codecs, tracking antennas, and intelligent receivers.

High frequencies, together with wide bandwidths, permit the use of small Earth terminals and high data rates and thus make satellite communication available to the home, business, and mobile terminal for diverse applications such as internet access, data retrieval, teleconferencing, and electronic library research.

Bary Bertiger, Corporate Vice President and General Manager of the Motorola Satellite Communication Group, described this capability as "global, instantaneous infrastructure that will be equally available at low cost to consumers in developing countries and industrialized nations. Virtually any intellectual property, such as documents and computer software, can be digitized and delivered via satellite instead of being physically transported by hand or transmitted over wires."

The development of these broadband satellite systems over the next few years will represent a communication revolution, notable both as a new stage of development and as a lens to alter our view of the world. Their existence will affect our very perception of how we communicate and the information resources that we can access, just as the invention of mechanical clocks in the middle ages altered the public perception of time, the growth of high speed travel during the twentieth century altered the perception of geographical distance, and the exploration of space has altered the perception of our place in the universe.

FREQUENCY BANDS

The first band used for commercial satellite communication in the Fixed Satellite Service (FSS) was C-band (6/4 GHz, where the uplink frequency is given first). During the mid-1980s, Ku-band (14/12 GHz) came into use. Due to its higher frequency, this band is sensitive to rain fade but with higher power satellites it has become popular because it permits smaller Earth station antennas.

Mobile telephony systems such as Motorola's IRIDIUM and Loral/Qualcomm's GLOBALSTAR, both in the process of deployment, will use lower frequencies, which are desirable because they maximize the received carrier power for fixed satellite and handset antenna gains. For example, IRIDIUM will use L-band (1.6 GHz) for both uplink and downlink, while GLOBALSTAR will use L-band (1.6 MHz) for the uplink and S-band (2.5 GHz) for the downlink. Satellite systems in the emerging Digital Audio Radio Service (DARS), such as CD Radio and AMRC, will use S-band in the vicinity of 2.3 GHz.

In the early 1990s, a variety of systems were designed for Ka-band (30/20 GHz) for broadband applications, such as Motorola's Millennium, Hughes' Spaceway, SS/Loral's Cyberstar, Lockheed Martin's Astrolink, Echostar, GE*Star, KaStar, Morning Star, Net Sat 28, Orion, and PanAmSat, which are all geostationary constellations, as well as the Teledesic LEO system. The practical use of such high frequencies for communication was first demonstrated by the NASA ACTS program. (The term K-band was originally given to the range 18 – 27 GHz, but after a molecular water vapor absorption resonance was discovered at the center of the band at 22.3 GHz, the terms Ku band (12 – 18 GHz) and Ka band (27 – 40 GHz) were introduced to denote "under" and "above" K-band; however, the regime 20 – 30 GHz for Ka-band is now common usage.)

The new systems will be at even higher frequencies in the so-called Q-band (33 – 50 GHz) and V-band (50 – 75 GHz) as defined by the FCC in its *Bulletin No. 70*, July, 1997. The FCC also defines U-band as 40 – 60 GHz, thus overlapping Q- and V-bands, and W-band as 75 – 110 GHz with additional letter designations all the way up to 220 GHz. However, conventional usage seems to be converging on the definition 40 – 50 GHz for V-band, which has also been called EHF; this trend would suggest designating 30 – 40 GHz as Q-band, 50 – 60 GHz as U-band, and 60 – 70 GHz as W-band if 10 – 20 GHz represents Ku-band, including both the FSS and BSS bands, and 20 – 30 GHz represents Ka-band.

ANTENNAS

Since the product of wavelength and frequency is equal to the speed of light (3×10^8 m/s), the wavelength decreases as the frequency increases. It is significant to note that at V-band (50 GHz), the signal wavelength is only 6 millimeters. By comparison, at C-band (6 GHz) the wavelength is 50 millimeters (5 cm) and at L-band (1.6 GHz) the wavelength is about 200 millimeters (20 cm). It is the very small wavelength that permits the fabrication of a high gain antenna with a small physical aperture.

Earth terminals that communicate with nongeostationary satellites will be required to have tracking and handover capability. A satellite in Medium Earth Orbit at an altitude of 10,000 km is visible for about 2 hours. In Low Earth Orbit at an altitude of 1000 km the maximum time in view is only 15 minutes.

The antennas, with gains on the order of 50 dB, will be either mechanically-steered

reflectors or electronically steered phased arrays. The reflector antennas will be typical for business installations, while the phased arrays will find application in lower capacity systems at residential sites or mobile terminals.

The phased array antennas under development will represent a major achievement in technology. The tracking requirement will depend on advances in microcircuit state of the art more than antenna design and it will be a challenge for the industry to offer them at attractive prices.

RAIN

At high frequencies, rain attenuation is a serious problem. Physically, rain attenuation is due to scattering and absorption of the microwave energy by the rain. As the wavelength decreases and approaches the size of a typical rain drop (approximately 1.5 mm), more scattering and absorption occurs and the attenuation increases. Also, as the rain rate increases during a heavy downpour, the size of the rain drops, and hence the attenuation, increases.

It is the rain rate, and not the annual rainfall, that determines availability. Thus San Francisco and Seattle are in the same rain climate region because the probability of a given rain rate being exceeded is about the same, despite the disparity in the total annual rainfall.

By way of example, for an availability of 99.95 percent or a total outage of 43.8 hours per year in Washington, DC, the maximum rain rate is 22 mm/h. The corresponding specific rain attenuation is approximately 0.05 dB/km at C-band, 1 dB/km at Ku-band, 3 dB/km at Ka-band, and 9 dB/km at V-band. For a given rain attenuation allowance, the availability at V-band is simply not as high as at Ku-band or even Ka-band.

The problem may be mitigated by switching to lower frequencies during periods of heavy rain. Thus dual payload satellites with both Ka-band and V-band steerable beams may be desirable from the point of view of engineering design. Nevertheless, customer awareness of the rain fade issue will be necessary.

For large, high capacity Earth stations, site diversity is used to overcome rain. Earth stations about 10 km apart and connected by terrestrial microwave circuits will see different rain cells, so that at least one Earth station will maintain the satellite link. For example, IRIDIUM uses this technique for its Ka-band mobile telephony gateways. Terrestrial systems can also be used for backup.

PROPOSED V-BAND SYSTEMS

Motorola was the first to explore the use of nongeostationary satellites in the new frequency regime. In September, 1996 the company submitted an application to the FCC for a Low Earth Orbit satellite constellation called M-Star to provide broadband services to businesses in the 40 GHz band. The proposed constellation consists of 72 satellites in circular orbits at an altitude of 1350 km and distributed in 12 planes inclined at 47°. The M-Star system is designed to offer two types of service: voice and data transport to service providers and business customers at 2.048 Mbps and interconnection and backhaul services at up to 51.84 Mbps.

Last June, Motorola submitted an application for a Ka-band LEO system called Celestri. The Celestri LEO system will comprise 63 satellites at an altitude of 1400 km distributed into 7 orbital planes inclined at 48°. Services to be offered include point-to-point symmetric transfer at 64 kbps to 155 Mbps; point-to-point asymmetric transfer with “bandwidth on demand” up to 16 Mbps; broadcast services; and interactive real-time response services.

The market comprises residential consumers seeking work-at-home, entertainment, education, and security capabilities; small businesses purchasing from multimedia outlets; and large multinational corporations seeking improved customer awareness. The Celestri system would augment the recently licensed Millennium Ka-band system of four geostationary satellites and the proposed M-Star system to form a three-tier LEO/GEO FDMA/TDMA communication architecture.

Celestri is presently regarded as an umbrella designation for all three systems. The available data rates for the LEO component is 2 Mbps on the uplink and 16 Mbps on the downlink; the GEO component provides a downlink data rate of 20 Mbps. Motorola has amended its application to request authorization for both V-band and Ka-band payloads on the Celestri satellites, and is considering the incorporation of the M-Star payload on the Celestri bus. Celestri is entirely different from IRIDIUM. Celestri will offer broadband services for high speed data transfer to fixed terminals, while IRIDIUM will provide narrowband services for voice communication and messaging to mobile terminals.

In order to connect the satellite system to end-users, Motorola has developed a range of terminal sizes, which collectively are described by the broad term "Customer Premises Equipment (CPE)". This equipment can be as large as a gateway station for telecommunications carriers and as small as a home unit that can be mounted on a roof. The home unit antenna is a high gain phased array capable of tracking the LEO satellite and providing seamless handover from leading to following satellites. According to Motorola spokesperson Robert Edwards, the projected cost of this unit is about $700, a surprisingly low estimate given the advanced technology it represents.

Three new V-band systems have been proposed by Hughes. “The V-band filings pioneer new spectrum to keep the satellite market strong by advancing technology into the realm of new market demands for mobile connectivity and increased bandwidth,” said Wendy Greene, spokesperson for Hughes.

The first is Expressway, a constellation of 14 geostationary satellites at 10 orbital locations to provide global high-capacity, wideband satellite communications. Expressway will use 3 GHz of uplink and downlink bandwidth in V-band and 500 MHz of uplink and downlink bandwidth in Ku-band. The V-band capacity will be used to serve high data rate users, such as multinational companies, with spot beams that can be activated in response to demand. The Ku-band capacity will be distributed through a series of larger beams. A typical user terminal has a 2.5 meter antenna and a 30 watt HPA.

The satellite architecture uses a piece of proven ACTS technology. On-board TDMA, IF-switched processing facilitates the allocation of “bandwidth on demand” and the satellites are interconnected by optical (laser) links. The data rates vary from T1 (1.544 Mbps) to OC-3 (155 Mbps), a 100:1 ratio on an individual carrier basis. The total capacity is 588,000 equivalent T1 circuits.

Expressway uses a “systems” approach to availability, seamlessly migrating traffic between its V-band and Ku-band capacity

on an individual user basis. With a typical allocation for rain fade of about 3 dB, the V-band availability will be around 98 percent; higher availabilities are provided with the satellite's Ku-band capability.

Expressway has been engineered to optimize capacity. This system is intended for a "leased line" dedicated user. The Ku-band capacity will be used sparingly to enhance availability where needed during periods of rain, and will be allocated depending on user level of service and pricing schedule.

By comparison, the Hughes Ka-band Spaceway system of eight geostationary satellites is optimized to user terminal and availability requirements. It is intended for occasional access, such as to small business and residential consumers, and will be priced by the bit. Spaceway involves substantial processing on the satellite.

The second component of the Hughes system is Spacecast, which will consist of six geostationary satellites. Spacecast will offer video and multimedia services at V-band and Ku-band. Using spot beams, the system will have multitasking capabilities for one-way transmission to small terminals for applications such as corporate training and distance learning. The data rate to a 45 cm terminal would be 26 Mbps and the data rate to a 1 meter terminal would be 155 Mbps.

The third component of the Hughes system is Starlynx. This is a hybrid V-band constellation with four geostationary satellites (two satellites in each of two orbital slots) and 20 Medium Earth Orbit satellites at an altitude of 10,352 km. The MEO constellation consists of four planes inclined at 55° with five satellites per plane. Starlynx will provide two-way data connectivity to portable terminals, such as notebook and desktop computers, using small, flat antennas. The terminals can be either stationary or mobile. For stationary terminals, the antenna size will be about 30 cm × 30 cm and the data rates will be up to 2 Mbps, while for mobile terminals the antenna size will be about 60 cm × 60 cm and the data rates will be up to 8 Mbps.

PanAmSat, an independent company with majority ownership by Hughes, has asked the FCC for approval to launch a 12 satellite geostationary constellation to provide global digital services at V-band. The system, called V-Stream, is to be deployed in 11 orbital slots, from 99° W longitude for North America to 124.5° E longitude for the Pacific Rim. It will use 3 GHz of spectrum in the 50/40 GHz band and will include high powered spot beams with onboard processing and intersatellite links at 33/23 GHz and/or 60 GHz. (The 60 GHz frequency is particularly appropriate for intersatellite links because the atmosphere is opaque in this neighborhood due to resonance absorption by molecular oxygen.)

The V-Stream system will augment PanAmSat's existing network of 16 satellites providing C-band and Ku-band services; in addition, the company has received FCC authorization to operate Ka-band satellites in nine orbital slots.

TRW has requested FCC authorization to launch and operate a system called the TRW Global EHF Satellite Network (GESN). The GESN system space segment consists of a hybrid constellation of four geostationary satellites and 15 MEO satellites that will operate in 6 hour circular orbits at an altitude of 10,355 km. The satellites are distributed in three orbital planes inclined at 50° with five equally spaced satellites per plane to ensure high elevation angle links (greater than 30°).

The MEO component of the constellation has an obvious similarity with TRW's former 12 satellite ODYSSEY system for mobile telephony, which was abandoned in favor of a partnership with ICO, and suggests that TRW may be placing its commercial satellite development emphasis in a new direction.

According to TRW's Director of Telecommunications Policy Peter Hadinger, the requirements of a V-band system certainly complement the experience the company has gained in the satellite arena. "This is really playing to our forté, in terms of the millimeter wave frequency bands and the use of onboard signal processing," Hadinger says. He believes TRW's work on the Milstar project and other military payloads will give the company an advantage on technical development.

The services to be offered on a global basis will be two-way point-to-point wideband data connectivity, multimedia distribution services, and private network services. The GESN system application requests the use 3 GHz of bandwidth in each direction, specifically 47.2 to 50.2 GHz for the uplink and 37.5 to 40.5 GHz for the downlink.

The system will use optical intersatellite links. The uplink supports a standard service link (SSL) of 155.52 Mbps and a wideband service link (WSL) of 1.5552 Gbps The downlink supports a total channel rate, including data rate and overhead, of either 317 Mbps (SSL) or 3.17 Mbps (WSL). The modulation format is OQPSK. These signal structures are used for both the GEO and MEO satellites.

TRW is targeting large businesses and international carriers, not the residential consumer market. The user terminal antenna aperture is 1.5 to 2.2 meters and the RF power is 12 to 30 watts for the SSL, while the antenna aperture is 2.2 to 2.5 meters and the RF power is 100 W for the WSL. Terminals that operate through the MEO constellation will be required to mechanically track the satellites through a 120° arc and will also be required to have dual tracking capability to achieve transparent handovers between leading and following satellites. Recognizing that mechanically steered reflector antennas may be objectionable or impractical for some users, TRW has indicated that it will work with established manufacturers of commercial satellite terminals to develop small, attractively priced, electronically steered, flat phased array antennas using monolithic microwave integrated circuit (MMIC) devices.

Lockheed Martin's proposed Global Q/V-Band Satellite Communications System will consist of 9 geostationary satellites. It requests FCC authorization to provide broadband services requiring 3 GHz for the uplink in the range 47.2 – 50.2 GHz and 3 GHz for the downlink in the range 39.5 – 42.5 GHz.

This system will provide high data rate communication to provide infrastructure to areas not adequately served by terrestrial systems. Through the use of both small and large user terminals, it will provide instant connectivity for the exchange of data at rates up to OC-3 (155 Mbps). This capability will extend the services provided at Ka-band by Astrolink, which will be optimized to provide switched data services at data rates from 64 kbps to 2 Mbps. Astrolink is a Lockheed Martin strategic venture.

The coverage will be composed of 48 transmit and receive beams serving user terminals and 8 transmit and receive beams serving gateway Earth stations. Each beam has a nominal half power beamwidth of 0.3° and occupies 125 MHz of bandwidth.

The transmission scheme utilizes a unique TDM architecture in which redundant data bits are added to user channels experiencing significant rain fading. Each 125 MHz downlink channel contains a single 96.29 Mbps carrier. Ground terminals extract and buffer data addressed only to them. The bi-directional user terminals will have antenna diameters as small as 45 cm with a transmit gain of 44.8 dB. A terminal of this size will have a traveling wave tube amplifier (TWTA) with an output power of 4 watts and for a minimum elevation angle of 30° will support a maximum information uplink data rate of 384 kbps.

Larger antennas will be used for lower elevation angles and high rain rate regions. A 2.4 meter reflector with a transmit gain of 59.3 dB and an output power of 12 watts will support uplink data rates up to 9.216 Mbps. The target availability for the allocated rain margin is 98%. Interference to adjacent satellites is mitigated by interactive power control with the spacecraft.

The Loral Space and Communications system, called CyberPath, consists of 10 geostationary satellites. Loral seeks 1 GHz of spectrum for the uplink and 1 GHz of spectrum for the downlink. Data rates begin at 16 kbps. Higher data rates, such as 6 Mbps, are available on demand for video and data transmission. Trunking data rates up to 90 Mbps are also available. The CyberPath system capacity is 17.9 Gbps.

Each satellite uses on-board demodulation and decoding and ATM-like switching to achieve connectivity among the 100 V-band spot beams and the two inter-satellite links. Data are routed according to the packet header using a TDM/FDM/CDMA format.

The subscriber Earth station, ranging in size from 0.5 to 3.0 meters, is selected to achieve the desired availability in rain. It may be installed at the home, business, or government facility and are expected to initially cost $1500 installed. The Earth station is connected via the subscriber's computer to home or office equipment utilizing the multimedia services. The link is designed to have an availability of 99.5%, reflecting a larger rain allowance than most other systems.

GE American Communications seeks authorization for a constellation of 11 geostationary satellites in nine orbital locations. The global broadband system, called GE*StarPlus, would offer connectivity for data-based applications at rates up to 155 Mbps. The system would use 3 GHz for both uplink and downlink in the 50/40 GHz V-band and 500 MHz within Ku-band. The system will use optical intersatellite links.

Each satellite payload receives uplink signals, demodulates them, and routes them to 20 V-band and 8 Ku-band spot beams and one Ku-band hemispheric beam with dual circular polarizations.

The proposed GE*Starplus system would serve a diverse market for high-data rate communications that previously relied on less suitable telephone network lines, such as for the transport of medical images, desktop publishing, and academic information. Users will be able to change locations easily without requiring connection to wire-based data services. Each satellite will have an estimated capacity of 40,000 equivalent T1 circuits.

Spectrum Astro has designed a 25 satellite, 50/40 GHz V-band system called Aster that will consist of five clusters of collocated geostationary satellites. The cluster approach will enable the company to build up its system in conformance with market demand.

Each of the satellites produces 48 spot beams 0.5° in diameter, 8 elliptical 1° × 1.5° regional beams, and 2 steerable 0.8° beams. The spot beams and regional beams divide the required 2 GHz of bandwidth in each direction. Service is offered at data rates of 155 Mbps and higher through terminals in the range from 4 m to 7 m. Lower data rates from 2 Mbps to 51 Mbps are available through terminals in the 1.2 m to 5 m range. Spectrum Astro’s system will be available to homes, businesses, medical clinics, educational institutions, government agencies, and laboratories.

CAI Satellite Communications, Inc. intends to launch a single V-band geostationary satellite that has the ability to provide high quality two-way video, voice, and data services to business and residential customers in the contiguous United States (CONUS). The satellite would be collocated with a Ka-band satellite proposed by CAI's affiliate, CAI Data Systems, Inc., at 93°, 95°, or 102° W longitude.

The company seeks 1 GHz of spectrum from 49.2 to 50.2 GHz for Earth-to-space and 1 GHz of spectrum from 40.5 to 41.5 GHz for space-to-Earth. This system will complement CAI's existing terrestrial MMDS “wireless cable” system operating in the 2 GHz band to provide subscribers with a greater variety of video and interactive services.

Orbital Sciences is proposing a seven satellite broadband system from MEO called Orblink. The satellites will operate at an altitude of 9000 km in a single equatorial plane and will be equally spaced by 51.4°, forming a "wireless ring" around the Earth.

Two primary services will be offered: service to large gateways for digital trunks and “bandwidth on demand” for high-speed data users. Each satellite will be able to simultaneously accommodate 20 gateway users at 1.244 Gbps each and up to 4000 wideband users at 10 to 51 Mbps each. Orbcomm requests the bands 47.7 to 48.7 GHz for user to satellite, 37.5 to 38.5 GHz for satellite to user, and 65.0 to 71.0 GHz for intersatellite links, all using dual circular polarization.

Pentriad, a system developed by Denali Telecom, LLC, is proposed as an international system to provide broadband multicasting and Direct To Home (DTH) services in the northern hemisphere. The main capacity of the Pentriad satellite system would be utilized for broadband services to telecommunications carriers. It employs a unique constellation of nine operational satellites in highly elliptical orbits distributed into three orbital planes plus three in-orbit spares and one ground spare.

Pentriad proposes to use 2 GHz of V-band (near 50 GHz) for the uplink and 2 GHz of Q-band (near 40 GHz) and 200 MHz of Ku-band (near 12 MHz) for the downlink. The Pentriad satellites have “bent pipe” transponders that relay, but do not process, data from one ground location to another. The basic channel data rate is 155 Mbps, which can be subdivided to provide slower rates down to 10 Mbps or grouped together to provide higher rates up to 3.875 Gbps.

Teledesic is proposing a 72-satellite system called V-Band Supplement (VBS) to augment its already ambitious 288 satellite Ka-band system (down from the original 840 satellite constellation).

LEO One has asked for additional spectrum at 40 GHz for its 48 satellite “Little LEO” constellation for tracking and messaging.

Globalstar plans to launch an 80 satellite V-band constellation called GS-40

to expand its mobile telephony system. In addition, it plans to launch 64 Low Earth Orbit satellites and four geostationary satellites that will operate at 2 GHz.

ANOTHER GOLD RUSH

V-band spectrum was not the only territory on which companies rushed to stake claims. Additional new proposals at 2 GHz meeting the FCC September 26 deadline, contemporaneous with the V-band deadline, include a 16 satellite Medium Earth Orbit constellation proposed by Boeing for navigation services to airlines, a 96 satellite constellation called Macrocell to expand and complement the IRIDIUM 66 satellite constellation, and a 26 satellite constellation to expand the capacity of the 17 satellite Ellipso system.

In addition to the new filings, three letters of intent asking spectrum at 2 GHz were submitted by ICO Global Communications for its 10 satellite MEO mobile telephony system, TMI Communications for CanSat-M3 to supplement its operational MSat-1 satellite that provides two-way voice, tracking, and paging services, and INMARSAT for its 4 satellite Horizons system that will provide data, voice, and videoconferencing capabilities to portable computers.

CROWDED SKIES

The first geostationary satellite, SYNCOM III, was successfully launched in 1964. Since that time approximately 250 satellites have been launched into GEO, of which about 170 are operational. Roughly another 80 satellites are on order to increase or replace services at C-band and Ku-band. About 165 "Big LEO" satellites are planned for mobile telephony and another 200 "Little LEO" satellites are planned for messaging and data gathering. In Ka-band, about 70 geostationary satellites have been proposed in addition to the 288 satellite Teledesic LEO constellation. To these we now add another 250 satellites at S-band and nearly 300 more at Q- and V-bands. The total number of new satellites is staggering and is in excess of 1300 satellites. Not every proposed system will be approved, funded, built, and supported by the market. However, it is clear that there is an enormous growth ahead in satellite hardware and services. The true frontier is nowhere in sight.

Dr. Robert A. Nelson, P.E. is president of Satellite Engineering Research Corporation, a satellite engineering consulting firm in Bethesda, Maryland.

Via Satellite, November 1999

The Global Positioning System

A National Resource

by Robert A. Nelson

On a recent trip to visit the Jet Propulsion Laboratory, I flew from Washington, DC to Los Angeles on a new Boeing 747-400 airplane. The geographical position of the plane and its relation to nearby cities was displayed throughout the flight on a video screen in the passenger cabin. When I arrived in Los Angeles, I rented a car that was equipped with a navigator. The navigator guided me to my hotel in Pasadena, displaying my position on a map and verbally giving me directions with messages like "freeway exit ahead on the right followed by a left turn." When I reached the hotel, it announced that I had arrived at my destination. Later, when I was to join a colleague for dinner, I found the restaurant listed in a menu and the navigator took me there.

This remarkable navigation capability is made possible by the Global Positioning System (GPS). It was originally designed jointly by the U.S. Navy and the U.S. Air Force to permit the determination of position and time for military troops and guided missiles. However, GPS has also become the basis for position and time measurement by scientific laboratories and a wide spectrum of applications in a multi-billion dollar commercial industry. Roughly one million receivers are manufactured each year and the total GPS market is expected to approach $ 10 billion by the end of next year. The story of GPS and its principles of measurement are the subjects of this article.

EARLY METHODS OF NAVIGATION

The shape and size of the earth has been known from the time of antiquity. The fact that the earth is a sphere was well known to educated people as long ago as the fourth century BC. In his book *On the Heavens*, Aristotle gave two scientifically correct arguments. First, the shadow of the earth projected on the moon during a lunar eclipse appears to be curved. Second, the elevations of stars change as one travels north or south, while certain stars visible in Egypt cannot be seen at all from Greece.

The actual radius of the earth was determined within one percent by Eratosthenes in about 230 BC. He knew that the sun was directly overhead at noon on the summer solstice in Syene (Aswan, Egypt), since on that day it illuminated the water of a deep well. At the same time, he measured the length of the shadow cast by a column on the grounds of the library at Alexandria, which was nearly due north. The distance between Alexandria and Syene had been well established by professional runners and camel caravans. Thus Eratosthenes was able to compute the earth's radius from the difference in latitude that he inferred from his measurement. In terms of modern units of length, he arrived at the figure of about
6400 km. By comparison, the actual mean radius is 6371 km (the earth is not precisely spherical, as the polar radius is 21 km less than the equatorial radius of 6378 km).

The ability to determine one's position on the earth was the next major problem to be addressed. In the second century, AD the Greek astronomer Claudius Ptolemy prepared a geographical atlas, in which he estimated the latitude and longitude of principal cities of the Mediterranean world. Ptolemy is most famous, however, for his geocentric theory of planetary motion, which was the basis for astronomical catalogs until Nicholas Copernicus published his heliocentric theory in 1543.

Historically, methods of navigation over the earth's surface have involved the angular measurement of star positions to determine latitude. The latitude of one's position is equal to the elevation of the pole star. The position of the pole star on the celestial sphere is only temporary, however, due to precession of the earth's axis of rotation through a circle of radius 23.5° over a period of 26,000 years. At the time of Julius Caesar, there was no star sufficiently close to the north celestial pole to be called a pole star. In 13,000 years, the star *Vega* will be near the pole. It is perhaps not a coincidence that mariners did not venture far from visible land until the era of Christopher Columbus, when true north could be determined using the star we now call *Polaris*. Even then the star's diurnal rotation caused an apparent variation of the compass needle. *Polaris* in 1492 described a radius of about 3.5° about the celestial pole, compared to 1° today. At sea, however, Columbus and his contemporaries depended primarily on the mariner's compass and dead reckoning.

The determination of longitude was much more difficult. Longitude is obtained astronomically from the difference between the observed time of a celestial event, such as an eclipse, and the corresponding time tabulated for a reference location. For each hour of difference in time, the difference in longitude is 15 degrees.

Columbus himself attempted to estimate his longitude on his fourth voyage to the New World by observing the time of a lunar eclipse as seen from the harbor of Santa Gloria in Jamaica on February 29, 1504. In his distinguished biography *Admiral of the Ocean Sea*, Samuel Eliot Morrison states that Columbus measured the duration of the eclipse with an hour-glass and determined his position as seven hours and fifteen minutes west of Cadiz, Spain, according to the predicted eclipse time in an almanac he carried aboard his ship. Over the preceding year, while his ship was marooned in the harbor, Columbus had determined the latitude of Santa Gloria by numerous observations of the pole star. He made out his latitude to be 18°, which was in error by less than half a degree and was one of the best recorded observations of latitude in the early sixteenth century, but his estimated longitude was off by some 38 degrees.

Columbus also made legendary use of this eclipse by threatening the natives with the disfavor of God, as indicated by a portent from Heaven, if they did not bring desperately needed provisions to his men. When the eclipse arrived as predicted, the natives pleaded for the Admiral's intervention, promising to furnish all the food that was needed.

New knowledge of the universe was revealed by Galileo Galilei in his book *The Starry Messenger*. This book, published in Venice in 1610, reported the telescopic discoveries of hundreds of new stars, the craters on the moon, the phases of Venus, the rings of Saturn, sunspots, and the four inner satellites of Jupiter. Galileo suggested using the eclipses of Jupiter's satellites as a celestial clock for the practical determination of longitude, but the calculation of an accurate ephemeris and the difficulty of observing the satellites from the deck of a rolling ship prevented use of this method at sea. Nevertheless, James Bradley, the third Astronomer Royal of England, successfully applied the technique in 1726 to determine the longitudes of Lisbon and New York with considerable accuracy.

Inability to measure longitude at sea had the potential of catastrophic consequences for sailing vessels exploring the new world, carrying cargo, and conquering new territories. Shipwrecks were common. On October 22, 1707 a fleet of twenty-one ships under the command of Admiral Sir Clowdisley Shovell was returning to England after an unsuccessful military attack on Toulon in the Mediterranean. As the fleet approached the English Channel in dense fog, the flagship and three others foundered on the coastal rocks and nearly two thousand men perished.

Stunned by this unprecedented loss, the British government in 1714 offered a prize of £20,000 for a method to determine longitude at sea within a half a degree. The scientific establishment believed that the solution would be obtained from observations of the moon. The German cartographer Tobias Mayer, aided by new mathematical methods developed by Leonard Euler, offered improved tables of the moon in 1757. The recorded position of the moon at a given time as seen from a reference meridian could be compared with its position at the local time to determine the angular position west or east.

Just as the astronomical method appeared to achieve realization, the British craftsman John Harrison provided a different solution through his invention of the marine chronometer. The story of Harrison's clock has been recounted in Dava Sobel's popular book, *Longitude*.

Both methods were tested by sea trials. The lunar tables permitted the determination of longitude within four minutes of arc, but with Harrison's chronometer the precision was only one minute of arc. Ultimately, portions of the prize money were awarded to Mayer's widow, Euler, and Harrison.

In the twentieth century, with the development of radio transmitters, another class of navigation aids was created using terrestrial radio beacons, including Loran and Omega. Finally, the technology of artificial satellites made possible navigation and position determination using line of sight signals involving the measurement of Doppler shift or phase difference.

TRANSIT

Transit, the Navy Navigation Satellite System, was conceived in the late 1950s and deployed in the mid-1960s. It was finally retired in 1996 after nearly 33 years of service. The Transit system was developed because of the need to provide accurate navigation data for Polaris missile submarines. As related in an historical perspective by Bradford Parkinson, *et al.* in the journal *Navigation* (Spring 1995), the concept was suggested by the predictable but dramatic Doppler frequency shifts from the first *Sputnik* satellite, launched by the Soviet Union in October, 1957. The Doppler-shifted signals enabled a determination of the orbit using data recorded at one site during a single pass of the satellite. Conversely, if a satellite's orbit were already known, a radio receiver's position could be determined from the same Doppler measurements.

The Transit system was composed of six satellites in nearly circular, polar orbits at an altitude of 1075 km. The period of revolution was 107 minutes. The system employed essentially the same Doppler data used to track the *Sputnik* satellite. However, the orbits of the Transit satellites were precisely determined by tracking them at widely spaced fixed sites. Under favorable conditions, the rms accuracy was 35 to 100 meters. The main problem with Transit was the large gaps in coverage. Users had to interpolate their positions between passes.

GLOBAL POSITIONING SYSTEM

The success of Transit stimulated both the U.S. Navy and the U.S. Air Force to investigate more advanced versions of a space-based navigation system with enhanced capabilities. Recognizing the need for a combined effort, the Deputy Secretary of Defense established a Joint Program Office in 1973. The NAVSTAR Global Positioning System (GPS) was thus created.

In contrast to Transit, GPS provides continuous coverage. Also, rather than Doppler shift, satellite range is determined from phase difference.

There are two types of observables. One is pseudorange, which is the offset between a pseudorandom noise (PRN) coded signal from the satellite and a replica code generated in the user's receiver, multiplied by the speed of light. The other is accumulated delta range (ADR), which is a measure of carrier phase.

The determination of position may be described as the process of triangulation using the measured range between the user and four or more satellites. The ranges are inferred from the time of propagation of the satellite signals. Four satellites are required to determine the three coordinates of position and time. The time is involved in the correction to the receiver clock and is ultimately eliminated from the measurement of position.

High precision is made possible through the use of atomic clocks carried on-board the satellites. Each satellite has two cesium clocks and two rubidium clocks, which maintain time with a precision of a few parts in 10^{13} or 10^{14} over a few hours, or better than 10 nanoseconds. In terms of the distance traversed by an electromagnetic signal at the speed of light, each nanosecond corresponds to about 30 centimeters. Thus the precision of GPS clocks permits a real time measurement of distance to within a few meters. With post-processed carrier phase measurements, a precision of a few centimeters can be achieved.

The design of the GPS constellation had the fundamental requirement that at least four satellites must be visible at all times from any point on earth. The tradeoffs included visibility, the need to pass over the ground control stations in the United States, cost, and sparing efficiency.

The orbital configuration approved in 1973 was a total of 24 satellites, consisting of 8 satellites plus one spare in each of three equally spaced orbital planes. The orbital radius was 26,562 km, corresponding to a period of revolution of 12 sidereal hours, with repeating ground traces. Each satellite arrived over a given point four minutes earlier each day. A common orbital inclination of 63° was selected to maximize the on-orbit payload mass with launches from the Western Test Range. This configuration ensured between 6 and 11 satellites in view at any time.

As envisioned ten years later, the inclination was reduced to 55° and the number of planes was increased to six. The constellation would consist of 18 primary satellites, which represents the absolute minimum number of satellites required to provide continuous global coverage with at least four satellites in view at any point on the earth. In addition, there would be 3 on-orbit spares.

The operational system, as presently deployed, consists of 21 primary satellites and 3 on-orbit spares, comprising four satellites in each of six orbital planes. Each orbital plane is inclined at 55°. This constellation improves on the "18 plus 3" satellite constellation by more fully integrating the three active spares.

SPACE SEGMENT

There have been several generations of GPS satellites. The Block I satellites, built by Rockwell International, were launched between 1978 and 1985. They consisted of eleven prototype satellites, including one launch failure, that validated the system concept. The ten successful satellites had an average lifetime of 8.76 years.

The Block II and Block IIA satellites were also built by Rockwell International. Block II consists of nine satellites launched between 1989 and 1990. Block IIA, deployed between 1990 and 1997, consists of 19 satellites with several navigation enhancements. In April 1995, GPS was declared fully operational with a constellation of 24 operational spacecraft and a completed ground segment. The 28 Block II/IIA satellites have exceeded their specified mission duration of 6 years and are expected to have an average lifetime of more than 10 years.

Block IIR comprises 20 replacement satellites that incorporate autonomous navigation based on crosslink ranging. These satellites are being manufactured by Lockheed Martin. The first launch in 1997 resulted in a launch failure. The first IIR satellite to reach orbit was also launched in 1997. The second GPS 2R satellite was successfully launched aboard a Delta 2 rocket on October 7, 1999. One to four more launches are anticipated over the next year.

The fourth generation of satellites is the Block II follow-on (Block IIF). This program includes the procurement of 33 satellites and the operation and support of a new GPS operational control segment. The Block IIF program was awarded to Rockwell (now a part of Boeing). Further details may be found in a special issue of the *Proceedings of the IEEE* for January, 1999.

CONTROL SEGMENT

The Master Control Station for GPS is located at Schriever Air Force Base in Colorado Springs, CO. The MCS maintains the satellite constellation and performs the stationkeeping and attitude control maneuvers. It also determines the orbit and clock parameters with a Kalman filter using measurements from five monitor stations distributed around the world. The orbit error is about 1.5 meters.

GPS orbits are derived independently by various scientific organizations using carrier phase and post-processing. The state of the art is exemplified by the work of the International GPS Service (IGS), which produces orbits with an accuracy of approximately 3 centimeters within two weeks.

The system time reference is managed by the U.S. Naval Observatory in Washington, DC. GPS time is measured from Saturday/Sunday midnight at the beginning of the week. The GPS time scale is a composite "paper clock" that is synchronized to keep step with Coordinated Universal Time (UTC) and International Atomic Time (TAI). However, UTC differs from TAI by an integral number of leap seconds to maintain correspondence with the rotation of the earth, whereas GPS time does not include leap seconds. The origin of GPS time is midnight on January 5/6, 1980 (UTC). At present, TAI is ahead of UTC by 32 seconds, TAI is ahead of GPS by 19 seconds, and GPS is ahead of UTC by 13 seconds. Only 1,024 weeks were allotted from the origin before the system time is reset to zero because 10 bits are allocated for the calendar function (1,024 is the tenth power of 2). Thus the first GPS rollover occurred at midnight on August 21, 1999. The next GPS rollover will take place May 25, 2019.

SIGNAL STRUCTURE

The satellite position at any time is computed in the user's receiver from the navigation message that is contained in a 50 bps data stream. The orbit is represented for each one hour period by a set of 15 Keplerian orbital elements, with harmonic coefficients arising from perturbations, and is updated every four hours.

This data stream is modulated by each of two code division multiple access, or spread spectrum, pseudorandom noise (PRN) codes: the coarse/acquisition C/A code (sometimes called the clear/access code) and the precision P code. The P code can be encrypted to produce a secure signal called the Y code. This feature is known as the Anti-Spoof (AS) mode, which is intended to defeat deception jamming by adversaries. The C/A code is used for satellite acquisition and for position determination by civil receivers. The P(Y) code is used by military and other authorized receivers.

The C/A code is a Gold code of register size 10, which has a sequence length of 1023 chips and a chipping rate of 1.023 MHz and thus repeats itself every 1 millisecond. (The term "chip" is used instead of "bit" to indicate that the PRN code contains no information.) The P code is a long code of length 2.3547 x 10^{14} chips with a chipping rate of 10

times the C/A code, or 10.23 MHz. At this rate, the P code has a period of 38.058 weeks, but it is truncated on a weekly basis so that 38 segments are available for the constellation. Each satellite uses a different member of the C/A Gold code family and a different one-week segment of the P code sequence.

The GPS satellites transmit signals at two carrier frequencies: the L1 component with a center frequency of 1575.42 MHz, and the L2 component with a center frequency of 1227.60 MHz. These frequencies are derived from the master clock frequency of 10.23 MHz, with L1 = 154 x 10.23 MHz and L2 = 120 x 10.23 MHz. The L1 frequency transmits both the P code and the C/A code, while the L2 frequency transmits only the P code. The second P code frequency permits a dual-frequency measurement of the ionospheric group delay. The P-code receiver has a two-sigma rms horizontal position error of about 5 meters.

The single frequency C/A code user must model the ionospheric delay with less accuracy. In addition, the C/A code is intentionally degraded by a technique called Selective Availability (SA), which introduces errors of 50 to 100 meters by dithering the satellite clock data. Through differential GPS measurements, however, position accuracy can be improved by reducing SA and environmental errors.

The transmitted signal from a GPS satellite has right hand circular polarization. According to the GPS Interface Control Document, the specified minimum signal strength at an elevation angle of 5° into a linearly polarized receiver antenna with a gain of 3 dB (approximately equivalent to a circularly polarized antenna with a gain of 0 dB) is - 160 dBW for the L1 C/A code, - 163 dBW for the L1 P code, and - 166 dBW for the L2 P code. The L2 signal is transmitted at a lower power level since it is used primarily for the ionospheric delay correction.

PSEUDORANGE

The fundamental measurement in the Global Positioning System is pseudorange. The user equipment receives the PRN code from a satellite and, having identified the satellite, generates a replica code. The phase by which the replica code must be shifted in the receiver to maintain maximum correlation with the satellite code, multiplied by the speed of light, is approximately equal to the satellite range. It is called the pseudorange because the measurement must be corrected by a variety of factors to obtain the true range.

The corrections that must be applied include signal propagation delays caused by the ionosphere and the troposphere, the space vehicle clock error, and the user's receiver clock error. The ionosphere correction is obtained either by measurement of dispersion using the two frequencies L1 and L2 or by calculation from a mathematical model, but the tropospheric delay must be calculated since the troposphere is nondispersive. The true geometric distance to each satellite is obtained by applying these corrections to the measured pseudorange.

Other error sources and modeling errors continue to be investigated. For example, a recent modification of the Kalman filter has led to improved performance. Studies have also shown that solar radiation pressure models may need revision and there is some new evidence that the earth's magnetic field may contribute to a small orbit period variation in the satellite clock frequencies.

CARRIER PHASE

Carrier phase is used to perform measurements with a precision that greatly exceeds those based on pseudorange. However, a carrier phase measurement must resolve an integral cycle ambiguity whereas the pseudorange is unambiguous.

The wavelength of the L1 carrier is about 19 centimeters. Thus with a cycle resolution of one percent, a differential measurement at the level of a few millimeters is theoretically possible. This technique has important applications to geodesy and analogous scientific programs.

RELATIVITY

The precision of GPS measurements is so great that it requires the application of Albert Einstein's special and general theories of relativity for the reduction of its measurements. Professor Carroll Alley of the University of Maryland once articulated the significance of this fact at a scientific conference devoted to time measurement in 1979. He said, "I think it is appropriate ... to realize that the first practical application of Einstein's ideas in actual engineering situations are with us in the fact that clocks are now so stable that one must take these small effects into account in a variety of systems that are now undergoing development or are actually in use in comparing time worldwide. It is no longer a matter of scientific interest and scientific application, but it has moved into the realm of engineering necessity."

According to relativity theory, a moving clock appears to run slow with respect to a similar clock that is at rest. This effect is called "time dilation." In addition, a clock in a weaker gravitational potential appears to run fast in comparison to one that is in a stronger gravitational potential. This gravitational effect is known in general as the "red shift" (only in this case it is actually a "blue shift").

GPS satellites revolve around the earth with a velocity of 3.874 km/s at an altitude of 20,184 km. Thus on account of the its velocity, a satellite clock appears to run slow by 7 microseconds per day when compared to a clock on the earth's surface. But on account of the difference in gravitational potential, the satellite clock appears to run fast by 45 microseconds per day. The net effect is that the clock appears to run fast by 38 microseconds per day. This is an enormous rate difference for an atomic clock with a precision of a few nanoseconds. Thus to compensate for this large secular rate, the clocks are given a rate offset prior to satellite launch of
- 4.465 parts in 10^{10} from their nominal frequency of 10.23 MHz so that on average they appear to run at the same rate as a clock on the ground. The actual frequency of the satellite clocks before launch is thus 10.22999999543 MHz.

Although the GPS satellite orbits are nominally circular, there is always some residual eccentricity. The eccentricity causes the orbit to be slightly elliptical, and the velocity and altitude vary over one revolution. Thus, although the principal velocity and gravitational

effects have been compensated by a rate offset, there remains a slight residual variation that is proportional to the eccentricity. For example, with an orbital eccentricity of 0.02 there is a relativistic sinusoidal variation in the apparent clock time having an amplitude of 46 nanoseconds. This correction must be calculated and taken into account in the GPS receiver.

The displacement of a receiver on the surface of the earth due to the earth's rotation in inertial space during the time of flight of the signal must also be taken into account. This is a third relativistic effect that is due to the universality of the speed of light. The maximum correction occurs when the receiver is on the equator and the satellite is on the horizon. The time of flight of a GPS signal from the satellite to a receiver on the earth is then 86 milliseconds and the correction to the range measurement resulting from the receiver displacement is 133 nanoseconds. An analogous correction must be applied by a receiver on a moving platform, such as an aircraft or another satellite. This effect, as interpreted by an observer in the rotating frame of reference of the earth, is called the Sagnac effect. It is also the basis for a laser ring gyro in an inertial navigation system.

GPS MODERNIZATION

In 1996, a Presidential Decision Directive stated the president would review the issue of Selective Availability in 2000 with the objective of discontinuing SA no later than 2006. In addition, both the L1 and L2 GPS signals would be made available to civil users and a new civil 10.23 MHz signal would be authorized. To satisfy the needs of aviation, the third civil frequency, known as L5, would be centered at 1176.45 MHz, in the Aeronautical Radio Navigation Services (ARNS) band, subject to approval at the World Radio Conference in 2000. According to Keith McDonald in an article on GPS modernization published in the September, 1999 *GPS World*, with SA removed the civil GPS accuracy would be improved to about 10 to 30 meters. With the addition of a second frequency for ionospheric group delay corrections, the civil accuracy would become about 5 to 10 meters. A third frequency would permit the creation of two beat frequencies that would yield one-meter accuracy in real time.

A variety of other enhancements are under consideration, including increased power, the addition of a new military code at the L1 and L2 frequencies, additional ground stations, more frequent uploads, and an increase in the number of satellites. These policy initiatives are driven by the dual needs of maintaining national security while supporting the growing dependence on GPS by commercial industry. When these upgrades would begin to be implemented in the Block IIR and IIF satellites depends on GPS funding.

Besides providing position, GPS is a reference for time with an accuracy of 10 nanoseconds or better. Its broadcast time signals are used for national defense, commercial, and scientific purposes. The precision and universal availability of GPS time has produced a paradigm shift in time measurement and dissemination, with GPS evolving from a secondary source to a fundamental reference in itself.

The international community wants assurance that it can rely on the availability of GPS and continued U.S. support for the system. The Russian Global Navigation Satellite System (GLONASS) has been an alternative, but economic conditions in Russia have threatened its continued viability. Consequently, the European Union is considering the creation of a navigation system of its own, called Galileo, to avoide relying on the U.S. GPS and Russian GLONASS programs.

The Global Positioning System is a vital national resource. Over the past thirty years it has made the transition from concept to reality, representing today an operational system on which the entire world has become dependent. Both technical improvements and an enlightened national policy will be necessary to ensure its continued growth into the twenty-first century.

Dr. Robert A. Nelson, P.E. is president of Satellite Engineering Research Corporation, a satellite engineering consulting firm in Bethesda, Maryland. He is *Via Satellite*'s Technical Editor.

Via Satellite, February 2000

The International System of Units (SI)

Its History and Use in Science and Industry

by Robert A. Nelson

On September 23, 1999 the Mars Climate Orbiter was lost during an orbit injection maneuver when the spacecraft crashed onto the surface of Mars. The principal cause of the mishap was traced to a thruster calibration table, in which British units instead of metric units were used. The software for celestial navigation at the Jet Propulsion Laboratory expected the thruster impulse data to be expressed in newton seconds, but Lockheed Martin Astronautics in Denver, which built the orbiter, provided the values in pound-force seconds, causing the impulse to be interpreted as roughly one-fourth its actual value. The failure was magnified by the loss of the companion spacecraft Mars Polar Lander due to an unknown cause on December 3.

The incident renews a controversy that has existed in the United States since the beginning of the space program regarding the use of metric or British units of measurement. To put the issue into perspective, this article reviews the history of the metric system and its modern version, the International System of Units (SI). The origin and evolution of the metric units, and the role they have played in the United States, will be summarized. Technical details and definitions will be provided for reference. Finally, the use of metric units in the satellite industry will be examined.

ORIGIN OF THE METRIC SYSTEM

The metric system was one of many reforms introduced in France during the period between 1789 and 1799, known as the French Revolution. The need for reform in the system of weights and measures, as in other affairs, had long been recognized. No other aspect of applied science affects the course of human activity so directly and universally.

Prior to the metric system, there had existed in France a disorderly variety of measures, such as for length, volume, or mass, that were arbitrary in size and variable from one town to the next. In Paris the unit of length was the *Pied de Roi* and the unit of mass was the *Livre poids de marc*. These units could be traced back to Charlemagne. However, all attempts to impose the "Parisian" units on the whole country were fruitless, as they were opposed by the guilds and nobles who benefited from the confusion.

The advocates of reform sought to guarantee the uniformity and permanence of the units of measure by taking them from properties derived from nature. In 1670, the abbe Gabriel Mouton of Lyons proposed a unit of length equal to one minute of arc on the earth's surface, which he divided into decimal fractions. He suggested a pendulum of specified period as a means of preserving one of these submultiples.

The conditions required for the creation of a new measurement system were made possible by the French Revolution, an event that was initially provoked by a national financial crisis. In 1787 King Louis XVI convened the Estates General, an institution that had last met in 1614, for the purpose of imposing new taxes to avert a state of bankruptcy. As they assembled in 1789, the commoners, representing the Third Estate, declared themselves to be the only legitimate representatives of the people, and succeeded in having the clergy and nobility join them in the formation of the National Assembly. Over the next two years, they drafted a new constitution.

In 1790, Charles-Maurice de Talleyrand, Bishop of Autun, presented to the National Assembly a plan to devise a system of units based on the length of a pendulum beating seconds at latitude 45°. The new order was envisioned as an "enterprise whose result should belong some day to the whole world." He sought, but failed to obtain, the collaboration of England, which was concurrently considering a similar proposal by Sir John Riggs Miller.

The two founding principles were that the system would be based on scientific observation and that it would be a decimal system. A distinguished commission of the French Academy of Sciences, including J. L. Lagrange and Pierre Simon Laplace, considered the unit of length. Rejecting the seconds pendulum as insufficiently precise, the commission defined the unit, given the name *metre* in 1793, as one ten millionth of a quarter of the earth's meridian passing through Paris. The proposal was accepted by the National Assembly on March 26, 1791.

The definition of the meter reflected the extensive interest of French scientists in the figure of the earth. Surveys in Lapland by Pierre Louis Maupertuis in 1736 and in France by Nicolas Lacaille in 1740 had refined the value of the earth's radius and established definitively that the shape of the earth is oblate. Additional meridian arcs were measured in Peru in 1735 – 1743 and at the Cape of Good Hope in 1751.

To determine the length of the meter, a new survey was conducted by the astronomers Jean Baptiste Delambre and P.F.A. Mechain between Dunkirk, in France on the English Channel, and Barcelona, Spain, on the coast of the Mediterranean Sea. This work was begun in 1792 and completed in 1798, enduring the hardships of the "reign of terror" and the turmoil of revolution. We now know that the quadrant of the earth is 10 001 966 meters (in the WGS 84 model) instead of exactly 10 000 000 meters as originally planned. The principal source of error was the assumed value of the earth's flattening used in correcting for oblateness.

The unit of volume, the *pinte* (later renamed the *litre*), was

defined as the volume of a cube having a side equal to one-tenth of a meter. The unit of mass, the *grave* (later renamed the *kilogramme*), was defined as the mass of one pinte of distilled water at the temperature of melting ice. In addition, the centigrade scale for temperature was adopted, with fixed points at 0 °C and 100 °C representing the freezing and boiling points of water (now replaced by the Celsius scale).

The work to determine the unit of mass was assigned to Antoine-Laurent Lavoisier, the father of modern chemistry, and Rene-Just Hauy. In a tragedy symbolic of the period, Lavoisier was guillotined by a revolutionary tribunal in 1794. The measurements were completed by Louis Lefevre-Gineau and Giovanni Fabbroni in 1799. However, they found that they could not cool liquid water to exactly 0 °C and that the maximum density occurs at 4 °C, not at 0 °C as had been supposed. Therefore, the definition of the kilogram was amended to specify the temperature of maximum density. We now know that the intended mass was 0.999 972 kg, i.e., 1000.028 cm^3 instead of exactly 1000 cm^3 for the volume of 1 kilogram of pure water at 4 °C.

On August 1, 1793 the National Convention, which by then ruled France, issued a decree adopting the preliminary definitions and terms. The "methodical" nomenclature, specifying multiples and fractions of the units by Greek and Latin prefixes, was chosen in favor of the "common" nomenclature, involving separate names.

A new calendar was established by a law of October 5, 1793. Its origin was designated retroactively as September 22, 1792 to commemorate the overthrow of the monarchy and the inception of the Republic of France. The French Revolutionary Calendar consisted of twelve months of thirty days each, concluded by a five or six day holiday. The months were given poetic names that suggested the prevailing seasons. Each month was divided into three ten-day weeks, or decades. The day itself was divided into decimal fractions, with 10 hours per day, 100 minutes per hour, and 100 seconds per minute. The calendar was politically, rather than scientifically, motivated, since it was intended to weaken the influence of Christianity. It was abolished by Napoleon in 1806 in return for recognition by the Church of his authority as emperor of France. Although the calendar reform remained in effect for twelve years, the new method of keeping the time of day required the replacement of valued clocks and timepieces and was never actually used in practice.

The metric system was officially adopted on April 7, 1795. The government issued a decree (*Loi du 18 germinal, an III*) formalizing the adoption of the definitions and terms that are in use today. A brass bar was made to represent the provisional meter, obtained from the survey of Lacaille, and a provisional standard for the kilogram was derived.

A scientific conference was held from 1798 to 1799 that included representatives of the Netherlands, Switzerland, Denmark, Spain, and the Italian states, as well as France, to validate the computations and design prototype standards. Permanent standards for the meter and kilogram made from platinum were constructed. The full length of the meter bar represented the unit. These standards were deposited in the Archives of the Republic. They became official by an act of December 10, 1799.

During the Napoleonic era, several regressive acts were passed that temporarily revived old traditions. Thus in spite of its auspicious beginning, the metric system was not quickly adopted in France. Although the system continued to be taught in the schools, lack of funds prevented the distribution of secondary standards. Finally, after a three year transition period, the metric system became compulsory throughout France as of January 1, 1840.

REACTION IN THE UNITED STATES

The importance of a uniform system of weights and measures was recognized in the United States, as in France. Article I, Section 8, of the U.S. Constitution provides that Congress shall have the power "to coin money ... and fix the standard of weights and measures." However, although the progressive concept of decimal coinage was introduced, the early American settlers both retained and cultivated the customs and tools of their British heritage, including the measures of length and mass. In contrast to the French Revolution, the "American Revolution" was not a revolution at all, but was rather a war of independence.

In 1790, the same year that Talleyrand proposed metric reform in France, President George Washington referred the subject of weights and measures to his Secretary of State, Thomas Jefferson. In a report submitted to the House of Representatives, Jefferson considered two alternatives: if the existing measures were retained they could be rendered more simple and uniform, or if a new system were adopted, he favored a decimal system based on the principle of the seconds pendulum. As it was eventually formulated, Jefferson did not endorse the metric system, primarily because the metric unit of length could not be checked without a sizable scientific operation on European soil.

The political situation at the turn of the eighteenth century also made consideration of the metric system impractical. Although France under Louis XVI had supported the colonies in the war with England, by 1797 there was manifest hostility. The revolutionary climate in France was viewed by the external world with a mixture of curiosity and alarm. The National Convention had been replaced by the Directory, and French officials who had been sympathetic to the United States either had been

executed or were in exile. In addition, a treaty negotiated with England by John Jay in 1795 regarding settlement of the Northwest Territories and trade with the British West Indies was interpreted by France as evidence of an Anglo-American alliance. France retaliated by permitting her ships to prey upon American merchant vessels and Federalist President John Adams prepared for a French invasion. Thus in 1798, when dignitaries from foreign countries were assembled in Paris to learn of France's progress with metrological reform, the United States was not invited.

A definitive investigation was prepared in 1821 by Secretary of State John Quincy Adams that was to remove the issue from further consideration for the next 45 years. He found that the standards of length, volume, and mass used throughout the 22 states of the Union were already substantially uniform, unlike the disparate measures that had existed in France prior to the French Revolution. Moreover, it was not at all evident that the metric system would be permanent, since even in France its use was sporadic and, in fact, the consistent terminology had been repealed in 1812 by Napoleon. Therefore, if the metric system failed to win support in early America, it was not for want of recognition.

Serious consideration of the metric system did not occur again until after the Civil War. In 1866, upon the advice of the National Academy of Sciences, the metric system was made legal by the Thirty-Ninth Congress. The Act was signed into law on July 28 by President Andrew Johnson.

TREATY OF THE METER

A series of international expositions in the middle of the nineteenth century enabled the French government to promote the metric system for world use. Between 1870 and 1872, with an interruption caused by the Franco-Prussian War, an international meeting of scientists was held to consider the design of new international metric standards that would replace the meter and kilogram of the French Archives. A Diplomatic Conference on the Meter was convened to ratify the scientific decisions. Formal international approval was secured by the Treaty of the Meter, signed in Paris by the delegates of 17 countries, including the United States, on May 20, 1875.

The treaty established the International Bureau of Weights and Measures (BIPM). It also provided for the creation of an International Committee for Weights and Measures (CIPM) to run the Bureau and the General Conference on Weights and Measures (CGPM) as the formal diplomatic body that would ratify changes as the need arose. The French government offered the Pavillon de Breteuil, once a small royal palace, to serve as headquarters for the Bureau in Sevres, France near Paris. The grounds of the estate form a tiny international enclave within French territory.

The first three kilograms were made in 1880 and one was chosen as the international prototype. In 1884 an additional 40 kilograms and 30 meter bars were obtained. They were all manufactured from a an alloy of 90 percent platinum and 10 percent iridium by Johnson, Mathey and Company of London. The original meter and kilogram of the French Archives in their existing states were taken as the points of departure. The standards were intercompared at the International Bureau. A particular meter bar, number 6, became the international prototype. The remaining standards were distributed to the signatories. The work was approved by the First General Conference on Weights and Measures in 1889.

The United States received meters 21 and 27 and kilograms 4 and 20. On January 2, 1890 the seals to the shipping cases for meter 27 and kilogram 20 were broken in an official ceremony at the White House with President Benjamin Harrison presiding. The standards were deposited in the Office of Weights and Measures of the U.S. Coast and Geodetic Survey.

U.S. CUSTOMARY UNITS

The U.S. customary units were tied to the British and French units by a variety of indirect comparisons.

Troy weight was the standard for the minting of coins. Congress could be ambivalent about nonuniformity in standards for trade, but it could not tolerate nonuniformity in its standards for money. Therefore, in 1827 a brass copy of the British troy pound of 1758 was secured by Ambassador to England and former Secretary of the Treasury, Albert Gallatin. This standard was kept in the Philadelphia mint and lesser copies were made and distributed to other mints. The troy pound of the Philadelphia mint was virtually the primary standard for commercial transactions until 1857 and remained the standard for coins until 1911.

The semi-official standards used in commerce for a quarter century may be attributed to Ferdinand Hassler, who was appointed superintendent of the newly organized Coast Survey in 1807. In 1832 the Treasury Department directed Hassler to construct and distribute to the states standards of length, mass, and volume, and balances by which masses might be compared. As the standard of length, Hassler adopted the Troughton scale, an 82-inch brass bar made by Troughton of London for the Coast Survey that Hassler had brought back from Europe in 1815. The distance between the 27th and 63rd engraved lines on a silver inlay scale down the center of the bar was taken to be equal to the British yard. The standard of mass was the avoirdupois pound, derived from the troy pound of the Philadelphia mint by the ratio 7000 grains to 5760 grains. It was represented by a brass knob weight

that Hassler constructed and marked with a star. Thus it has come to be known as the "star" pound.

The system of weights and measures in Great Britain had been in use since the reign of Queen Elizabeth I. Following a reform begun in 1824, the imperial standard avoirdupois pound was made the standard of mass in 1844 and the imperial standard yard was adopted in 1855. The imperial standards were made legal by an Act of Parliament in 1855 and are preserved in the Board of Trade in London. The United States received copies of the British imperial pound and yard, which became the official U.S. standards from 1857 until 1893.

When the metric system was made lawful in the United States in 1866, a companion resolution was passed to distribute metric standards to the states. The Treasury Department had in its possession several copies derived from the meter and kilogram of the French Archives. These included the "Committee" meter and kilogram, which were an iron end standard and a brass cylinder with knob copied from the French prototypes, that Hassler had brought with him when he immigrated to the United States in 1805. He had received them as a gift from his friend, J.G. Tralles, who was the Swiss representative to the French metric convocation in 1798 and a member of its committee on weights and measures. Also available were the "Arago" meter and kilogram, named after the French physicist who certified them. They were purchased by the United States in 1821 through Albert Gallatin, then minister to France. The Committee meter and the Arago kilogram were used as the prototypes for brass metric standards that were distributed to the states.

In 1893, under a directive from Thomas C. Mendenhall, Superintendent of Standard Weights and Measures of the Coast and Geodetic Survey, the U.S. customary units were redefined in terms of the metric units. The primary standards of length and mass adopted were prototype meter No. 27 and prototype kilogram No. 20 that the United States had received in 1889 as a signatory to the Treaty of the Meter. The yard was defined as 3600/3937 meter and the avoirdupois pound-mass was defined as 0.453 592 427 7 kilogram. The conversion for mass was based on a comparison between the British imperial standard pound and the international prototype kilogram performed in 1883. These definitions were used by the National Bureau of Standards (now the National Institute of Standards and Technology) from its founding in 1901 until 1959. On July 1, 1959 the definitions were fixed by international agreement among the English-speaking countries to be 1 yard = 0.9144 meter and 1 pound-mass = 0.453 592 37 kilogram exactly. The definition of the yard is equivalent to the relations 1 foot = 0.3048 meter and 1 inch = 2.54 centimeters exactly.

The derived unit of force in the British system is the pound-force (lbf), which is defined as the weight of one pound-mass (lbm) at a hypothetical location where the acceleration of gravity has the standard value 9.806 65 m/s^2 exactly. Thus, 1 lbf = 0.453 592 37 kg x 9.806 65 m/s^2 = 4.448 N approximately. The slug (sl) is the mass that receives an acceleration of one foot per second squared under a force of one pound-force. Thus 1 sl = (1 lbf)/(1 ft/s^2) = (4.448 N)/(0.3048 m/s^2) = 14.59 kg = 32.17 lbm approximately.

ELECTROMAGNETISM

The theories of electricity and magnetism developed and matured during the early 1800s as fundamental discoveries were made by Hans Christian Oersted, Andre-Marie Ampere, Michael Faraday, and many others. The possibility of making measurements of terrestrial magnetism in terms of mechanical units, that is, in "absolute measure," was first pointed out by Karl Friedrich Gauss in 1833. His analysis was carried further to cover all electromagnetic phenomena by Wilhelm Weber, who in 1851 discussed a method by which a complete set of absolute units might be developed.

In 1861 a committee of the British Association for the Advancement of Science, that included William Thomson (later Lord Kelvin), James Clerk Maxwell, and James Prescott Joule, undertook a comprehensive study of electrical measurements. This committee introduced the concept of a *system* of units. Four equations were sufficient to determine the units of charge q, current I, voltage V, and resistance R. These were either Coulomb's force law for charges or Ampere's force law for currents, the relation between charge and current $q = I\,t$, Ohm's law $V = I\,R$, and the equation for electrical work $W = V\,q = I^2\,R\,t$, where t is time.

A fundamental principle was that the system should be coherent. That is, the system is founded upon certain base units for length, mass, and time, and derived units are obtained as products or quotients without requiring numerical factors. The meter, gram, and mean solar second were selected as base units. In 1873 a second committee recommended a centimeter-gram-second (CGS) system of units because in this system the density of water is unity.

Two parallel systems of units were devised, the electrostatic and electromagnetic subsystems, depending on whether the law of force for electric charges or for electric currents was taken as fundamental. The ratio of the electrostatic to the electromagnetic unit of charge or current was a fundamental experimental constant c.

The committee also conducted research on electrical standards. It issued a wire resistance standard, the "B.A. unit," which soon became known as the "ohm." The

idea of naming units after eminent scientists was due to Sir Charles Bright and Latimer Clark.

At the time, electricity and magnetism were essentially two distinct branches of experimental physics. However, in a series of papers published between 1856 and 1865, Maxwell created a unified theory based on the field concept introduced by Faraday. He predicted the existence of electromagnetic waves and identified the "ratio of the units" c with the speed of light.

In 1888, Heinrich Hertz verified Maxwell's prediction by generating and detecting electromagnetic waves at microwave frequencies in the laboratory. He also greatly simplified the theory by eliminating unnecessary physical assumptions. Thus the form of Maxwell's equations as they are known to physicists and engineers today is due to Hertz. (Oliver Heaviside made similar modifications and introduced the use of vectors.) In addition, Hertz combined the electrostatic and electromagnetic CGS units into a single system related by the speed of light c, which he called the "Gaussian" system of units.

The recommendations of the B.A. committees were adopted by the First International Electrical Congress in Paris in 1881. Five "practical" electrical units were defined as certain powers of 10 of the CGS units: the ohm, farad, volt, ampere, and coulomb. In 1889, the Second Congress added the joule, watt, and a unit of inductance, later given the name henry.

In 1901, the Italian engineer Giovanni Giorgi demonstrated that the practical electrical units and the MKS mechanical units could be incorporated into a single coherent system by (1) selecting the meter, kilogram, and second as the base units for mechanical quantities; (2) expanding the number of base units to four, including one of an electrical nature; and (3) assigning physical dimensions to the permeability of free space μ_0, with a numerical value of $4\pi \times 10^{-7}$ in a "rationalized" system or 10^{-7} in an "unrationalized" system. (The term "rationalized," due to Heaviside, concerned where factors of 4π should logically appear in the equations based on symmetry.) The last assumption implied that the magnetic flux density B and magnetic field H, which are related in vacuum by the equation $B = \mu_0 H$, are physically distinct with different units, whereas in the Gaussian system they are of the same character and are dimensionally equivalent. An analogous situation occurs for the electric fields D and E that are related by $D = \varepsilon_0 E$, where ε_0 is the permittivity of free space given by $c^2 = 1 / \mu_0 \varepsilon_0$.

In 1908, an International Conference on Electrical Units and Standards held in London adopted independent, easily reproducible primary electrical standards for resistance and current, represented by a column of mercury and a silver coulombmeter, respectively. These so-called "international" units went into effect in 1911, but they soon became obsolete with the growth of the national standards laboratories and the increased application of electrical measurements to other fields of science.

With the recognition of the need for further international cooperation, the 6th CGPM amended the Treaty of the Meter in 1921 to cover the units of electricity and photometry and the 7th CGPM created the Consultative Committee for Electricity (CCE) in 1927. By the 8th CGPM in 1933 there was a universal desire to replace the "international" electrical units with "absolute" units. Therefore, the International Electrotechnical Commission (IEC) recommended to the CCE an absolute system of units based on Giorgi's proposals, with the practical electrical units incorporated into a comprehensive MKS system. The choice of the fourth unit was left undecided.

At the meeting of the CCE in September 1935, the delegate from England, J.E. Sears, presented a note that set the course for future action. He proposed that the ampere be selected as the base unit for electricity, defined in terms of the force per unit length between two long parallel wires. The unit could be preserved in the form of wire coils for resistance and Weston cells for voltage by calibration with a current balance. This recommendation was unanimously accepted by the CCE and was adopted by the CIPM.

Further progress was halted by the intervention of World War II. Finally, in 1946, by authority given to it by the CGPM in 1933, the CIPM officially adopted the MKS practical system of absolute electrical units to take effect January 1, 1948.

TEMPERATURE

The concepts of temperature and its measurement have evolved along two parallel paths. On one hand, there has been the steady advance since the early eighteenth century of mercury, alcohol, and resistance thermometers and the development of practical scales of temperature based on arbitrary fixed points. On the other hand, there has been the growth of gas thermometry and the definition of an absolute measure of temperature based on its interpretation in terms of thermodynamic processes.

The first reliable mercury-in-glass thermometers were constructed by the German instrument maker Gabriel Daniel Fahrenheit in the period between 1708 and 1724. He defined the Fahrenheit scale by taking as fixed points the freezing point of water mixed with salt at 0 °F and the normal temperature of the human body at 96 °F (now known to be nearly 3° higher). The resulting freezing and boiling points of pure water were 32 °F and 212 °F , with 180° between them. In 1730, R.A.F. de Reaumer proposed dividing the same interval into 80° using an alcohol thermometer. This scale was widely used in France until the Revolution.

Another mercury thermometer scale was invented by Joseph Delisle in 1732. Delisle took the boiling point of water as 0° and worked downward to 150° as the freezing point. In 1741 the Swedish astronomer Anders Celsius recalibrated the Delisle thermometer with a centigrade temperature scale, having an interval of 100° between the fixed points, again with the boiling point at 0 °C but with the freezing point defined as 100 °C. By 1745, the botanist Carl Linnaeus, a colleague of Celsius, adopted a similar scale, but inverted it so that the freezing and boiling points are at 0 °C and 100 °C, respectively, as is customary today. This centigrade scale of temperature was adopted in France in 1794 during the creation of the metric system.

The notion of an absolute temperature scale based on a thermodynamic process is due to the French physicist Guillaume Amontons, who is credited with the invention of the air thermometer in 1699. According to Amontons, the temperature could be defined as proportional to the pressure of the air.

In 1854 William Thomson (Lord Kelvin) proposed a definition of temperature in terms of the macroscopic notion of heat or work according to the theory of an ideal reversible heat engine, derived by the French engineer Sadi Carnot. The ratio of the thermodynamic temperatures can be defined as the ratio of the heat taken in to that given out by a reversible heat engine operating in a Carnot cyle, so that $T_1/T_2 = Q_1/Q_2$. The definition of thermodynamic temperature is thus independent of the working substance. The research of James Clerk Maxwell, Ludwig Boltzmann, and J. Willard Gibbs provided an equally valid microscopic interpretation of temperature as a measure of the energy distribution of the particles in the system.

The Carnot cycle defines only the ratio of temperatures; to determine the unit of temperature it is also necessary to specify the temperature difference between two fixed points. Historically, these fixed points have been either the freezing and boiling points of water in a relative scale, or the triple point of water with respect to absolute zero in a thermodynamic scale. Such a temperature scale can be realized by means of an ideal gas, whose equation of state is given by $p\ V = n\ R\ T = N\ k\ T$, where p is the pressure, V is the volume, T is the thermodynamic temperature, and R is the universal gas constant. The number of moles is $n = m/M = N/N_0$, where m is the mass, M is the molar mass, N is the number of particles, and N_0 is Avogadro's number. The connection between the macroscopic and microscopic viewpoints is thus made by Boltzmann's constant through the relation $k = R/N_0$.

The First General Conference of Weights and Measures in 1889 adopted the constant volume hydrogen scale based on fixed points at the freezing point (0 °C) and the boiling point (100 °C) of water at standard pressure. The temperature derived from the measured pressure was corrected to thermodynamic temperature by a Joule-Thomson pourous-plug experiment. By extrapolation of the data, it was found that the thermodynamic temperature T, defined by the ideal gas equation of state, was related to the centigrade temperature t_C by the approximate relation $T = t_C + 273$.

The mercury thermometer was selected as a secondary standard. Mercury-in-glass thermometers, made by Tonnelot of Paris of lead-free hard glass and carefully annealed, were distributed to the participants. The United States received six of these thermometers as temperature standards for the range 0 °C to 100 °C to accompany prototype meters 21 and 27 and prototype kilograms 4 and 20. In 1948 the Ninth General Conference on Weights and Measures renamed the centigrade scale as the Celsius scale, with the unit degree Celsius.

INTERNATIONAL SYSTEM OF UNITS (SI)

By 1948 the General Conference on Weights and Measures was responsible for the units and standards of length, mass, electricity, photometry, temperature, and ionizing radiation. At this time, the next major phase in the evolution of the metric system was begun. It was initiated by a request of the International Union of Pure and Applied Physics "to adopt for international use a practical international system of units." Thus the 9th CGPM decided to define a complete list of derived units. Derived units had not been considered previously because they do not require independent standards. Also, the CGPM brought within its province the unit of time, which had been the prerogative of astronomers.

The work was started by the 10th CGPM in 1954 and was completed by the 11th CGPM in 1960. During this period there was an extensive revision and simplification of the metric unit definitions, symbols, and terminology. The kelvin and candela were added as base units for thermodynamic temperature and luminous intensity, and in 1971 the mole was added as a seventh base unit for amount of substance.

The modern metric system is known as the International System of Units, with the international abbreviation SI. It is founded on the seven base units, summarized in Table 1, that by convention are regarded as dimensionally independent. All other units are derived units, formed coherently by multiplying and dividing units within the system without the use of numerical factors. Some derived units, including those with special names, are listed in Table 2. For example, the unit of force is the newton, which is equal to a kilogram meter per second squared, and the unit of energy is the joule, equal to a newton meter. The expression of multiples and submultiples of SI units is

facilitated by the use of prefixes, listed in Table 3. (Additional information is available on the Internet at the websites of the International Bureau of Weights and Measures at http//www.bipm.fr and the National Institute of Standards and Technology at http://physics.nist.gov/cuu.)

METRIC STANDARDS

One must distinguish a unit, which is an abstract idealization, and a standard, which is the physical embodiment of the unit. Since the origin of the metric system, the standards have undergone several revisions to reflect increased precision as the science of metrology has advanced.

The meter. The international prototype meter standard of 1889 was a platinum-iridium bar with an X-shaped cross section. The meter was defined by the distance between two engraved lines on the top surface of the bridge instead of the distance between the end faces. The meter was derived from the meter of the French Archives in its existing state and reference to the earth was abandoned.

The permanence of the international prototype was verified by comparison with three companion bars, called "check standards." In addition, there were nine measurements in terms of the red line of cadmium between 1892 and 1942. The first of these measurements was carried out by A. A. Michelson using the interferometer which he invented. For this work, Michelson received the Nobel Prize in physics in 1907.

Improvements in monochromatic light sources resulted in a new standard based on a well-defined wavelength of light. A single atomic isotope with an even atomic number and an even mass number is an ideal spectral standard because it eliminates complexity and hyperfine structure. Also, Doppler broadening is minimized by using a gas of heavy atoms in a lamp operated at a low temperature. Thus a particular red-orange krypton-86 line was chosen, whose wavelength was obtained by direct comparison with the cadmium wavelength. In 1960, the 11th CGPM defined the meter as the length equal to 1 650 763.73 wavelengths of this spectral line.

Research on lasers at the Boulder, CO laboratory of the National Bureau of Standards contributed to another revision of the meter. The wavelength and frequency of a stabilized helium-neon laser beam were measured independently to determine the speed of light. The wavelength was obtained by comparison with the krypton wavelength and the frequency was determined by a series of measurements traceable to the cesium atomic standard for the second. The principal source of error was in the profile of the krypton spectral line representing the meter itself. Consequently, in 1983 the 17th CGPM adopted a new definition of the meter based on this measurement as "the length of the path traveled by light in vacuum during a time interval of 1/299 792 458 of a second." The effect of this definition is to fix the speed of light at exactly 299 792 458 m/s. Thus experimental methods previously interpreted as measurements of the speed of light c (or equivalently, the permittivity of free space ε_0) have become calibrations of length.

The kilogram. In 1889 the international prototype kilogram was adopted as the standard for mass. The prototype kilogram is a platinum-iridium cylinder with equal height and diameter of 3.9 cm and slightly rounded edges. For a cylinder, these dimensions present the smallest surface area to volume ratio to minimize wear. The standard is carefully preserved in a vault at the International Bureau of Weights and Measures and is used only on rare occasions. It remains the standard today. The kilogram is the only unit still defined in terms of an arbitrary artifact instead of a natural phenomenon.

The second. Historically, the unit of time, the second, was defined in terms of the period of rotation of the earth on its axis as 1/86 400 of a mean solar day. Meaning "second minute," it was first applied to timekeeping in about the seventeenth century when pendulum clocks were invented that could maintain time to this precision.

By the twentieth century, astronomers realized that the rotation of the earth is not constant. Due to gravitational tidal forces produced by the moon on the shallow seas, the length of the day is increasing by about 1.4 milliseconds per century. The effect can be measured by comparing the computed paths of ancient solar eclipses on the assumption of uniform rotation with the recorded locations on earth where they were actually observed. Consequently, in 1956 the second was redefined in terms of the period of revolution of the earth about the sun, as represented by the *Tables of the Sun* computed at the end of the nineteenth century by the astronomer Simon Newcomb of the U.S. Naval Observatory in Washington, DC. The second was defined to be 1/31 556 925.974 7 of the tropical year 1900. The operational significance of this definition was to adopt the linear coefficient in Newcomb's formula for the mean longitude of the sun to determine the unit of time.

The rapid development of atomic clocks soon permitted yet another definition. Accordingly, in 1967 the 13th CGPM defined the second as "the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two ground states of the cesium-133 atom." This definition was based on observations of the moon, whose ephemeris is tied indirectly to the apparent motion of the sun, and was equivalent to the previous definition within the limits of experimental uncertainty.

The ampere. The unit of electric current, the ampere, is defined as that constant current which, if maintained in each of two parallel,

infinitely long wires with a separation of 1 meter in vacuum, would produce a force per unit length between them equal to 2 x 10^{-7} N/m. This formal definition serves to establish the value of the constant μ_0 as 4π x 10^{-7} N/A^2 exactly. Although the base unit for electricity is the ampere, the electrical units are maintained through the volt and the ohm.

In the past, the practical representation of the volt was a group of Weston saturated cadmium-sulfate electrochemical standard cells. A primary calibration experiment involved the measurement of the force between two coils of an "ampere balance" to determine the current, while the cell voltage was compared to the potential difference across a known resistance.

The ohm was represented by a wire-wound standard resistor. Its resistance was measured against the impedance of an inductor or a capacitor at a known frequency. The inductance can be calculated from the geometrical dimensions alone. From about 1960, a so-called Thompson-Lampard calculable capacitor has been used, in which only a single measurement of length is required.

Since the early 1970s, the volt has been maintained by means of the Josephson effect, a quantum mechanical tunneling phenomenon discovered by Brian Josephson in 1962. A Josephson junction may be formed by two superconducting niobium films separated by an oxide insulating layer. If the Josephson junction is irradiated by microwaves at frequency f and the bias current is progressively increased, the current-voltage characteristic is a step function, in which the dc bias voltage increases discontinuously at discrete voltage intervals equal to f / K_J, where $K_J = 2\ e / h$ is the Josephson constant, h is Planck's constant, and e is the elementary charge.

The ohm is now realized by the quantum Hall effect, a characteristic of a two-dimensional electron gas discovered by Klaus von Klitzing in 1980. In a device such as a silicon metal-oxide-semiconductor field-effect transistor (MOSFET), the Hall voltage V_H for a fixed current I increases in discrete steps as the gate voltage is increased. The Hall resistance, or $R_H = V_H / I$, is equal to an integral fraction of the von Klitzing constant, given by $R_K = h / e^2 = \mu_0\, c / 2\,\alpha$, where α is the fine structure constant. In practice, R_K can be measured in terms of a laboratory resistance standard, whose resistance is obtained by comparison with the impedance of a calculable capacitor, or it can be obtained indirectly from α.

A new method to determine the relation between the mechanical and electromagnetic units that has shown much promise is by means of a "watt balance," which has greater precision than an ordinary ampere balance. In this experiment, a current I is passed through a test coil suspended in the magnetic field of a larger coil so that the force F balances a known weight mg. Next the test coil is moved axially through the magnetic field and the velocity v and induced voltage V are measured. By the equivalence of mechanical and electrical power, $F\ v = V\ I$. The magnetic field and apparatus geometry drop out of the calculation. The voltage V is measured in terms of the Josephson constant K_J while the current I is calibrated by the voltage across a resistance known in terms of the von Klitzing constant R_K. The experiment determines $K_J^{\,2} R_K$ (and thus h), which yields K_J if R_K is assumed to be known in terms of the SI ohm.

The Josephson and quantum Hall effects provide highly uniform and conveniently reproducible quantum mechanical standards for the volt and the ohm. For the purpose of practical engineering metrology, conventional values for the Josephson constant and the von Klitzing constant were adopted by international agreement starting January 1, 1990. These values are $K_{J\text{-}90}$ = 483 597.9 GHz/V and $R_{K\text{-}90}$ = 25 812.807 Ω exactly. The best experimental SI values, obtained as part of an overall least squares adjustment of the fundamental constants completed in 1998, differ only slightly from these conventional values.

The kelvin. From 1889 until 1927, the national reference standard of temperature for the United States comprised a set of sixteen mercury-in-glass thermometers. In 1927, the CIPM adopted an International Temperature Scale (ITS-27) based on six reproducible equilibrium states that agreed with thermodynamic temperatures within the limits of measurement. The platinum resistance thermometer, the platinum rhodium/platinum thermocouple, and the optical pyrometer were used for interpolation over three temperature ranges. This scale was modified in 1948 and clarified in 1960.

The Tenth General Conference on Weights and Measures in 1954 adopted the absolute temperature scale with a single fixed point, where the three phases of water (solid, liquid, and gas) coexist, with the unit "degree Kelvin," later renamed simply kelvin. The unit of thermodynamic temperature, the kelvin, is defined as the fraction 1/273.16 of the thermodynamic temperature of the triple point of water. The effect of this definition is to make the temperature of the triple point to be 273.16 K, which corresponds to 0.01 °C. The Celsius scale is defined by the relation $t_C = T - 273.15$ exactly. Although the values of the thermodynamic and Celsius temperatures differ, the units are equivalent. Thus the degree Celsius, with symbol °C, is equal to the kelvin, with symbol K.

A new International Practical Temperature Scale (IPTS-68) with 13 equilibrium states was adopted in 1968 and was amended in 1975. This scale, however, was found to deviate from the thermodynamic temperature in certain regions and

thus was replaced by the International Temperature Scale of 1990 (ITS-90).

By implication, the interval between the freezing and boiling boiling points of water at standard pressure is no longer rigorously 100 °C, since thermodynamic temperature is defined by a single fixed point. Since 1968, when the revised International Practical Temperature Scale was adopted, evidence has indicated that the definition of the kelvin leads to the value 99.975 °C for the boiling point, instead of exactly 100 °C as originally intended. The correct value for the triple point would have been 273.22 K.

METRIC UNITS IN INDUSTRY

The International System of Units (SI) has become the fundamental basis of scientific measurement worldwide. It is also used for everyday commerce in virtually every country of the world but the United States. Congress has passed legislation to encourage use of the metric system, including the Metric Conversion Act of 1975 and the Omnibus Trade and Competitiveness Act of 1988, but progress has been slow.

The space program should have been the leader in the use of metric units in the United States and would have been an excellent model for education. Burt Edelson, Director of the Institute for Applied Space Research at George Washington University and former Associate Administrator of NASA, recalls that "in the mid-'80s, NASA made a valiant attempt to convert to the metric system" in the initial phase of the international space station program. However, he continued, "when the time came to issue production contracts, the contractors raised such a hue cry over the costs and difficulties of conversion that the initiative was dropped. The international partners were unhappy, but their concerns were shunted aside. No one ever suspected that a measurement conversion error could cause a failure in a future space project."

Economic pressure to compete in an international environment is a strong motive for contractors to use metric units. Barry Taylor, head of the Fundamental Constants Data Center of the National Institute of Standards and Technology and U.S. representative to the Consultative Committee on Units of the CIPM, expects that the greatest stimulus for metrication will come from industries with global markets. "Manufacturers are moving steadily ahead on SI for foreign markets," he says. Indeed, most satellite design technical literature does use metric units, including meters for length, kilograms for mass, and newtons for force, because of the influence of international partners, suppliers, and customers.

CONCLUSION

As we begin the new millennium, there should be a renewed national effort to promote the use of SI metric units throughout industry, and to assist the general public in becoming familiar with the system and using it regularly. The schools have taught the metric system in science classes for decades. It is time to put aside the customary units of the industrial revolution and to adopt the measures of precise science in all aspects of modern engineering and commerce, including the United States space program and the satellite industry.

Dr. Robert A. Nelson, P.E. is president of Satellite Engineering Research Corporation, a satellite engineering consulting firm in Bethesda, Maryland. He is *Via Satellite*'s Technical Editor.

Table 1. SI Base Units

Quantity	*Unit*	
	Name	*Symbol*
length	meter	m
mass	kilogram	kg
time	second	s
electric current	ampere	A
thermodynamic temperature	kelvin	K
amount of substance	mole	mol
luminous intensity	candela	cd

Table 2. Examples of SI Derived Units

Quantity	*Unit*		
	Special Name	*Symbol*	*Equivalent*
plane angle	radian	rad	1
solid angle	steradian	sr	1
angular velocity			rad/s
angular acceleration			rad/s^2
frequency	hertz	Hz	s^{-1}
speed, velocity			m/s
acceleration			m/s^2
force	newton	N	kg m/s^2
pressure, stress	pascal	Pa	N/m^2
energy, work, heat	joule	J	kg m^2/s^2, N m
power	watt	W	kg m^2/s^3, J/s
power flux density			W/m^2
linear momentum, impulse			kg m/s, N s
angular momentum			kg m^2/s, N m s
electric charge	coulomb	C	A s
electric potential, emf	volt	V	W/A, J/C
magnetic flux	weber	Wb	V s
resistance	ohm	Ω	V/A
conductance	siemens	S	A/V, Ω^{-1}
inductance	henry	H	Wb/A
capacitance	farad	F	C/V
electric field strength			V/m, N/C
electric displacement			C/m^2
magnetic field strength			A/m
magnetic flux density	tesla	T	Wb/m^2, N/(A m)
Celsius temperature	degree Celsius	°C	K
luminous flux	lumen	lm	cd sr
illuminance	lux	lx	lm/m^2
radioactivity	becquerel	Bq	s^{-1}

Table 3. SI Prefixes

Factor	*Prefix*	*Symbol*	*Factor*	*Prefix*	*Symbol*
10^{24}	yotta	Y	10^{-1}	deci	d
10^{21}	zetta	Z	10^{-2}	centi	c
10^{18}	exa	E	10^{-3}	milli	m
10^{15}	peta	P	10^{-6}	micro	μ
10^{12}	tera	T	10^{-9}	nano	n
10^{9}	giga	G	10^{-12}	pico	p
10^{6}	mega	M	10^{-15}	femto	f
10^{3}	kilo	k	10^{-18}	atto	a
10^{2}	hecto	h	10^{-21}	zepto	z
10^{1}	deka	da	10^{-24}	yocto	y