JOURNAL OF SPACECRAFT AND ROCKETS
Vol. 38, No. 2, March–April 2001

Development of the Quantitative Generalized Information Network Analysis Methodology for Satellite Systems

Graeme B. Shaw,* D. W. Miller,† and D. E. Hastings‡
Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

A generalized analysis methodology for satellite systems has been developed, and it can be used for the analysis of space system architectures addressing any mission in communications, sensing, or navigation. The generalized information network analysis methodology is a hybrid of information network flow analysis, signal and antenna theory, space systems engineering and econometrics and specifies measurable, unambiguous metrics for the cost, capability, performance, and adaptability. The important contribution of the work is that by standardizing the representation of the overall mission objective, in terms of generalized quality of service parameters, the new methodology organizes, prioritizes, and focuses the engineering effort expended in satellite systems analysis. The generalized methodology is thus identified as a valuable tool for space systems engineering, allowing qualitative and quantitative assessment of the impacts of system architecture, deployment strategy, schedule slip, market demographics, and technical risk.

Nomenclature

$\boldsymbol{A}$	=	state transition matrix
$\{Av, I, Is, R\}$	=	capability parameters: availability, integrity, isolation, rate
C_L	=	lifetime cost
C_{launch}	=	launch cost
C_{mfr}	=	manufacturing cost
C_s	=	baseline system cost
D	=	aperture extent, m
d	=	separation of symbols in signal space, V
$E[\,]$	=	expected value
E_b	=	energy per bit, J
E_x	=	elasticity with respect to variable x
f	=	frequency, Hz
$G(s)$	=	interfering signal spectra in Fourier domain
$g(x)$	=	probability density function of a signal
$I(s)$	=	input spectra in Fourier domain
$i(x)$	=	input signal in physical domain
$K_{\min}$	=	number of nearest neighbor symbols
M_c	=	total market capture
m	=	symbol size, bit
N_s	=	noise spectrum in Fourier domain
N_0	=	noise power density, W/Hz
P	=	price
$P(s)$	=	Fourier domain response of system
$\boldsymbol{P}_s$	=	state probabilities vector
$p(x)$	=	physical domain impulse response of system
p_s	=	marginal probability of entering state s
Q	=	quantity of demand
$Q'()$	=	Gaussian complementary distribution function
$R(s)$	=	Fourier domain output of system
$r(x)$	=	physical domain output of system
S	=	mission sensitivity
s	=	Fourier domain variable
$\boldsymbol{s}$	=	signal space symbol vector
$\sin\theta$	=	direction sine, from normal bisector of antenna
t	=	time, s
u	=	spatial frequency, wavelengths
V_f	=	failure compensation costs
v_s	=	compensation cost of state s
v_T	=	threshold voltage, V
W	=	bandwidth, Hz
γ_c	=	coding gain
λ	=	wavelength, m
λ_i	=	failure rate of component i, per year

Introduction

THERE are many different ways to design satellite systems to perform essentially the same task. To compare alternate designs, metrics are required that fairly judge the capabilities and performance of the different systems in carrying out the required task. In today's economic climate, there is also a requirement to consider the monetary cost associated with different levels of performance. Because of the extremely large capital investment required for any space venture, it is especially important for satellite designers to provide the customer with the best value. This hints to the possible benefits of a definable cost per (functional) performance metric. Capability, performance, and cost metrics can be used as design tools by addressing the sensitivity in performance and cost to changes in the system components or by identifying the key technology drivers. This leads to the definition of the adaptability metric that measures the sensitivity to changes in the design or role. Any metric used for comparative analysis should be quantifiable and unambiguous. A measurable metric, therefore, requires a formal definition that leads to a calculable expression. A major goal of this research has been to formally define the three metrics of capability, cost per function, and adaptability, as part of a consistent methodology for the quantifiable analysis of almost all satellite systems, spanning most likely applications: the generalized information network analysis (GINA) methodology.

Satellite Systems as Information Transfer Networks

Many current satellite applications provide some kind of service in communications, sensing, or navigation. The generalization made by Shaw[1] is that these satellite systems are information transfer systems and that ensuring information flow through the system is the overall mission objective. Information transfer systems exist only to serve a market, a demand that specific information symbols be transferred from some set of sources to a different set of, presumably, remote sinks. This origin–destination (O–D) market is distinct from the systems built to satisfy it and is defined by the requirements of the end users (at the sinks).

A satellite system can be represented as a modular information processing network. The satellites, subsystems, and ground stations make up individual modules of the system, each with well-defined

Received 12 May 1999; revision received 20 April 2000; accepted for publication 5 May 2000. Copyright © 2000 by the authors. Published by the American Institute of Aeronautics and Astronautics, Inc., with permission.

*Research Fellow, Department of Aeronautics and Astronautics.

†Associate Professor, Department of Aeronautics and Astronautics. Member AIAA.

‡Professor, Department of Aeronautics and Astronautics. Fellow AIAA.

interfaces (inputs and outputs) and a finite set of actions. This abstraction allows satellite system analysis to be treated as a network flow problem. System analysis is then reduced to characterizing how to "move some entity [information] from one point to another in an underlying network ... as efficiently as possible, both to provide good service to the users ... and to use the underlying transmission facilities effectively.[2]"

The network representation of the satellite system provides the framework for quantitative system analysis, based on the mathematics of information transmission and network flow. If the interaction between each module and the information signal can be estimated, the characteristics of the information arriving at the sinks can be calculated.

Correct representation of the satellite system as an information network requires a functional decomposition of the system into its most important functional modules. The functional modules are those elements of the system that impact the transfer of information from source to sink. Note that functional modules do not necessarily represent system hardware; a rain cloud can assuredly effect radio communication to a satellite, but it is not conventionally considered a system component. In fact, other than for component reliability estimation, the actual hardware configuration of a subsystem is of little interest to the network modeler. Of much greater importance is correctly modeling the functional interaction between a module and the information signals being transferred.

Figure 1 shows a simplified network for a system consisting of a single communication satellite. The system transfers data between a set of users utilizing several spot beams, which are the input and output interfaces for the satellites.

At this most basic level of abstraction, the network is modeled to comprise only the source and sink nodes, the satellite node and the interfaces between them. (Note that provided their interface with the network is similar, the users within each spot beam can be grouped as a single node). This level of detail is probably too simplistic for any useful system analysis. Figure 2 shows the network for the same system, modeled with a finer level of functional decomposition. In this more detailed model, the signal from a source node passes through modules representing the effects of atmospheric rain attenuation, space loss, and cross-channel interference, before being collected at a receiver module on the satellite. For simplicity, only one spot beam is drawn on the uplink. The signal from this receiver is passed (along with the signals from the other spot beams' receivers not shown) through a multichannel module representing the satellite digital signal processor (DSP). This module interprets the information symbols and reroutes them to the correct satellite transmit modules. Again, only one channel is shown for simplicity. The downlink has similar attenuation and interference modules, a user receiver and a DSP, and terminates at a sink node. Clearly, this lower level model is a more accurate representation of the real system.

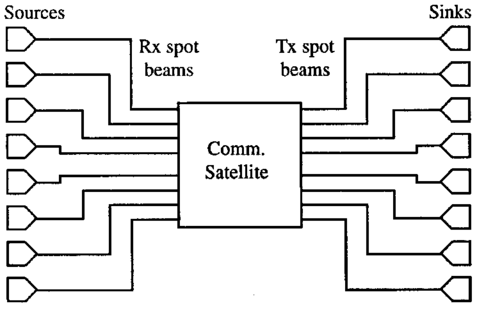


Fig. 1 Top-level network representation of a single communication satellite.

The network model can be further augmented by including additional support modules that are not part of the primary information pathway. For instance, modules representing the power generation system, the propulsion system, or the attitude control system of the satellite could be added. These support modules provide the other primary functional modules with enabling signals (power, propulsion, control, etc.). The functional modules must receive these enabling signals to transfer the information symbols correctly. The inclusion of these support modules in the network adds a further level of detail to the analyses.

This hierarchical nature of the network modeling allows the detail and accuracy of the analyses to be customized depending on the application. For example, at the conceptual design stage, the analyses may only have to predict the feasibility of the architecture. For this level of analysis, only the essential functional modules must be included. Later in the design process, more detail can be added to obtain accurate predictions of the capabilities of the entire system.

Capability Characteristics

The volume of demand served by the system is limited by the market (demographics, capture, and exhaustion) and by the system capabilities. Shaw[1] defined four quality-of-service parameters relating to the quantity, quality, and availability of the information arriving at the sinks as fair measures of the system's capabilities: signal isolation, information rate, information integrity, and information availability.

Signal Isolation

The system's ability to isolate and identify signals from different sources within the field of view is a critical mission driver for many applications. Obviously, a system cannot satisfactorily transfer information between specific O–D pairs unless the individual sources and sinks can be identified and isolated. Various methods are used to isolate the different signals. For communication systems, common isolation schemes separate the signals in frequency [frequency-division multiple access (FDMA)] or time [time-division multiple access (TDMA)]. Also, individual spot beams can be used to access multiple sources that are spatially separated. The same techniques can be applied to radar systems. Doppler frequency shifts are used for identification of the target velocity and clutter rejection, and time gating is used for target ranging. Scanning a small radar beam over a large area allows separate targets to be isolated in space to within a beamwidth. For imaging and remote sensing systems, the same principals apply. Different sources can be identified by detecting in different frequency bands. Spatially separated sources can be isolated using a high-resolution detector. An aperture can distinguish between sources that are separated by a distance no less than the resolution of the aperture. Note the one-to-one correspondence between 1) the resolution of an optic and the beamwidth of an antenna or a radar and 2) the frequency of radiation from a remote sensing pixel, the carrier frequency of a communication signal, and

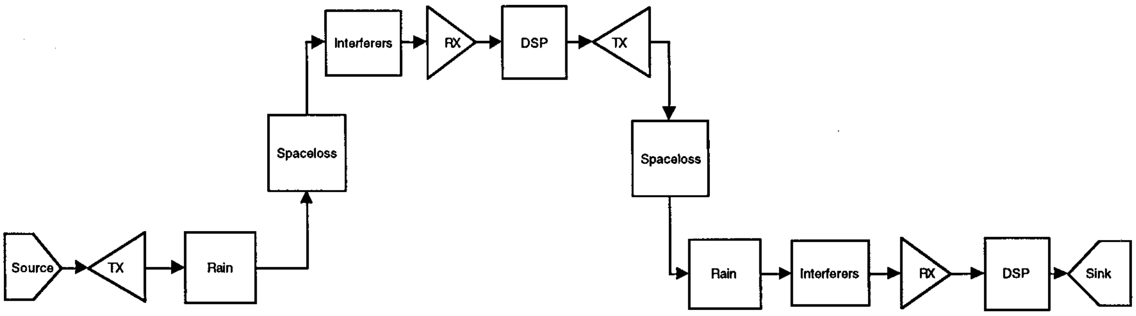


Fig. 2 Detailed network representation of a communication satellite.

the Doppler shifts of a radar signal. The generality exhibited in the mathematics of signal analysis allows these isolation relationships to be formalized.

Generalized Signal Isolation and Interference

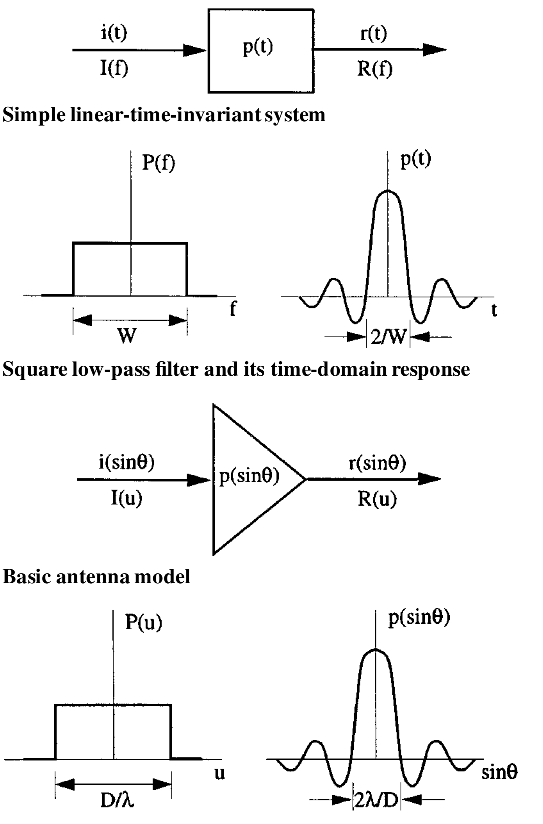
Information transfer systems must be able to isolate a given signal from any others that may be present. If the different signals cannot be distinguished, the cross-source interference will introduce noise that could cause an erroneous interpretation of the information. In general, a signal can be expressed in either of two domains; the physical domain x or the Fourier domain s. These two domains are related by the Fourier transform.

For electrical signals, such as in communications, the physical domain is time t, whereas the Fourier domain is frequency f. Either domain can be used for analysis, although it is often easier to perform the calculations in the frequency domain. For example, consider the simple linear system shown in Fig. 3. The time-domain output $r(t)$ is given by the convolution of the input signal $i(t)$ and the impulse response of the system $p(t)$. Equivalently, in the Fourier domain, the output $R(f)$ is given by multiplying the spectra of the input signal $I(f)$ and the frequency response of the system $P(f)$, such that

$$r(t) = i(t) * p(t) \leftrightarrow R(f) = I(f)P(f) \qquad (1)$$

Note that a square low-pass filter of bandwidth W Hz has a time-domain impulse response equal to a sinc function with a half-width of $1/W$ s, as shown in Fig. 3. This basically means that two time-domain impulse signals passing through the filter can be isolated only if their time separation is greater than this minimum value. The cutoff frequency W effectively limits the filters ability to transfer time-domain information.

There is an exact analogy to these relationships for optics and antenna theory.[3,4] The corresponding Fourier-transform pair is the angle between the propagation direction of the radiation and the normal of the antenna, measured as $\sin\theta$, and a spatial coordinate along the antenna (measured in wavelengths) referred to as the spatial frequency u. It is convenient to consider $\sin\theta$ as the physical variable and u as the Fourier variable, although the choice is arbitrary due to the symmetry of the Fourier transform. The analogy with electrical signal theory allows most of the properties relating to filtering and processing of time-domain electrical signals to be extended to antennas and optics. For example, consider the one-dimensional antenna shown in Fig. 3. The antenna images an unknown object distribution $i(\sin\theta)$ by filtering the object signal with a low-pass filter. An aperture or optic is a spatial filter because it samples only those parts of the signal within its spatial extent. The output image (in the angular domain) is equal to the convolution of the input signal $i(\sin\theta)$ and the impulse response of the aperture, defined as the radiation pattern $p(\sin\theta)$. Equivalently, the Fourier-domain output is given by the product of the input signal $I(u)$ and the aperture (illumination) distribution $P(u)$, such that



Simple linear-time-invariant system

Square low-pass filter and its time-domain response

Basic antenna model

Rectangular aperture distribution and its radiation pattern

Fig. 3 Basic signal and antenna theory.

$$r(\sin\theta) = i(\sin\theta) * p(\sin\theta) \leftrightarrow R(u) = I(u)P(u) \qquad (2)$$

Note that the angular radiation pattern of an aperture is equal to the Fourier transform of the aperture distribution. That is, it is the response of the antenna to uniform illumination over its extent. For a rectangular aperture of size D/λ, this response is a sinc function of half-width $\sin\theta = \lambda/D$, as shown in Fig. 3. The position of this first null in the radiation pattern corresponds to the angular resolution of the aperture because it determines the minimum angular separation of two point sources that can be successfully isolated. The cutoff frequency $u_0 = D/\lambda$ limits the ability of the antenna in transferring angular information. This property corresponds precisely to the earlier-stated isolation capabilities of electrical filters. (This one-to-one correspondence justifies our definition of u being the Fourier-domain variable.)

The similarities between the isolation characteristics of electrical systems and antenna systems are pervasive. The same principles of signal theory apply for both applications, and generalizations can be made about the isolation capabilities of a general system.

Signals can be isolated only if they are distinct and separable. Clearly, two signals that are separated in either the physical or Fourier domain satisfy this condition. For example, two electrical signals with nonoverlapping frequency bands can be isolated using a pair of bandpass filters. Similarly, two time-bounded signals transmitted sequentially can be isolated using a simple time gate. However, the condition that the signals be distinct and separable does not restrict them to exclusive occupation of part of one of the two domains. It is possible for a set of signals to occupy the same parts of the physical and Fourier domains and still be distinguished, albeit with some amount of interference.

To better understand what is meant by distinct and separable, it is helpful to adopt the signal space interpretation of signal analysis. Here, a geometrical linear space is defined by a set of real or complex vectors that represent a set of real or complex signals.[5,6] This approach is useful in signal analysis because it allows a number of mathematically equivalent problems to be treated with a common notation and a common solution.

Signal spaces are Hilbert spaces with a dimensionality defined by the number of orthonormal basis signals. The set $H_{[0,x]}$ of all real or complex L_2 signals with support in a finite physical interval $[0, x]$ is a signal space of countably infinite dimension. This is a statement of the fact that a signal bounded in the physical domain has an infinite number of Fourier components. Similarly, the set H_B of all real or complex signals whose Fourier domain is strictly bandlimited to a band B is a signal space of countably infinite dimension. The orthonormal basis functions here are an infinite set of $(\sin x)/x$ signals in the physical domain.

The detection and isolation process can now be stated in similar geometrical terms. If a particular signal is to be isolated from a set of interfering signals, the detector need only search in the subspace defined by the desired signal. A matched filter optimally designed to isolate a given signal simply projects the input signal space onto a space defined by the desired signal by performing an inner product. According to the Theorem of Irrelevance, this operation results in no loss of information about the desired signal and optimally reduces the interference from other signals. This means that only signals that are orthogonal to each other in signal space can be isolated with zero interference. Signals that are almost orthogonal to each other have a small inner product and can be isolated with a small

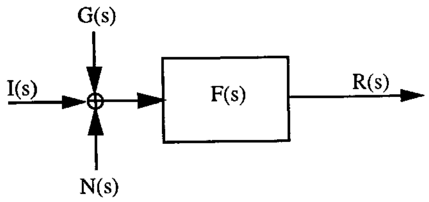


Fig. 4 Basic channel model for a simple system.

amount of interference. This concept of orthogonality is the correct interpretation of distinct and separable.

The amount of interference introduced in the detection process can be quantified. The interference noise power at the output of a matched filter is the integrated squared magnitude of the interfering signals after being projected into the signal space of the matched filter. To understand how this relates to conventional signal analysis, consider the system shown in Fig. 4.

An information signal $I(s)$ and an interfering signal $G(s)$ are the inputs to a system designed to isolate $I(s)$. The system's Fourier response is $P(s)$, and so the output $R(s)$ is given by

$$\begin{aligned} R(s) &= [I(s) + G(s)]P(s) + N(s)P(s) \\ &= I(s) + I(s)[P(s) - 1] + G(s)P(s) + N(s)P(s) \end{aligned} \tag{3}$$

where $N(s)$ is the (thermal) noise spectrum in the Fourier domain. All of the terms except the desired $I(s)$ add noise and distortion to the output of the system. The last term is the noise admittance of the system, but the second and third terms represent the interference outputs.

The isolation capabilities of the system determine the size of these interference outputs. The term $G(s)P(s)$ is the cross-channel interference, and $I(s)[P(s) - 1]$ is the intersymbol interference (ISI) within a signal. To eliminate ISI, the system channel response $P(s)$ must be unity within the bands where $I > 0$ and zero elsewhere. This ISI term is significant if the system involves a sampling of the signal into discrete (digital) components. In this case, $P(s)$ is a periodic, aliased spectrum, and $[P(s) - 1]$ can have positive values. Digital communication systems can be designed to give zero interference at the sampler output by enforcing that the signals satisfy the generalized Nyquist criteria (see Ref. 6). This basically requires each signal to be orthogonal to its translates by multiples of the sampling interval and also to all translates of the other signals. Of course, it is extremely unlikely that this condition be satisfied for remote sensing systems because the signals are externally generated.

The interference power at the output of a system is the squared magnitude of the filtered interfering signals, integrated in the domain in which the desired signal is bounded and over the same limits. For instance, the interference power at the output of a matched filter designed to isolate a signal bounded in the Fourier domain is equal to the power spectrum of all interfering signals, integrated over the bandwidth of the matched filter. Similarly, the interference power at the output of a system designed to isolate a signal bounded to $[0, x]$ in the physical domain is the total power of the filtered interfering signals within the physical limits $[0, x]$.

Information Rate

This is a measure of the rate at which the system transfers information symbols between each O–D pair. This is equivalent to the data rate for communication systems, the revisit rate for imaging systems, or the update rate for navigation systems. The system must deliver information symbols at a rate that matches the characteristic bandwidth of the source or the end user.

Information Integrity

The error performance of data collection and transfer systems is a critical issue in their design and operation.[6] A detector uses an observation of the signal plus noise to make a decision about each information symbol. Generally, the probability of erroneously interpreting an information symbol depends on the energy in the symbol. An error can occur if noise or interference degrades the signal in such a way that an incorrect decision is made about the observation. These errors can be as benign as a single bit error in a communication message or as consequential as a false alarm for an early warning radar system.

The probability of error for a single measurement is the likelihood that the interfering and thermal noise power exceeds some threshold, equal to the difference between information data values. Consider for example, the simplest case of an amplitude modulated binary communication channel [binary pulse amplitude modulation (PAM)]. The two data values {0, 1} are represented by two different voltage levels of the passband carrier wave. The separation between these levels is d V. If the noise component of the signal has a level greater than $d/2$ V, a data symbol {0} can appear in the observation as a {1} or vice versa. The probability of an error of a single bit is then the probability that the noise power is greater than the separation between data symbols. This is equal to the area under the noise probability density function from $[d/2, \infty]$.

For generality, integrity can be placed in the context of the signal space representation of signals introduced earlier. Consider an information transfer system that makes an observation known to be equal to one of two potential symbols, but distorted by noise. The two possible information symbols have signal space representations $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, such that the vector between them is $(\boldsymbol{s}_1 - \boldsymbol{s}_2)$. Define the length of this vector, equal to the separation between the symbols, to be d. The task of the detector is to determine which of the two possible symbols is the correct interpretation of the noisy observation. The decision rule used is based on the position of the observed signal projected into the same signal space. In general, the observation will not be coincident with either the two possible symbols, due to the presence of additive noise. The actual position in the signal space of the observation will be equal to the position of the underlying information symbol, plus the geometrically correct vector representation of the noise, according to standard rules of vector addition. Usually the symbol closest to the observation, among all of those that are possible, is chosen by the detector. For maximum likelihood detection with hard decisions, this corresponds to a decision threshold along the bisector between the two possible signals, at perpendicular distance of $d/2$ from each. A decision error will, therefore, be made if the projection of the noise in the direction of $(\boldsymbol{s}_1 - \boldsymbol{s}_2)$, is greater than $d/2$. For additive noise with a probability density function $g(x)$, the probability of this error [Pr(error)] occurring is

$$\Pr(\text{error}) = \int_{d/2}^{\infty} g(x)\,\mathrm{d}x \tag{4}$$

If there are more than two possible information symbols from which to choose, the net error probability for a given symbol is the sum of the probabilities calculated from Eq. (4) for each value of $d/2$ corresponding to the different pairs of symbols. This can be approximated from the union bound estimate,[6] in which the assumption is made that the closest pairs of symbols dominate the sum. If a given symbol has $K_{\min}$ nearest neighbors at a common distance d, then an estimate for the error probability is

$$\Pr(\text{error}) \approx K_{\min} \int_{d/2}^{\infty} g(x)\,\mathrm{d}x \tag{5}$$

When $g(x)$ is stationary white Gaussian noise with zero mean and variance σ^2, Eq. (5) becomes

$$\begin{aligned} \Pr(\text{error}) &\approx K_{\min} \cdot \frac{1}{\sigma\sqrt{2\pi}} \int_{d/2}^{\infty} \exp\left(\frac{-x^2}{2\sigma^2}\right) \mathrm{d}x \\ &\approx K_{\min} \cdot \frac{1}{2}\operatorname{erfc}\left(\frac{d}{2\sigma\sqrt{2}}\right) \approx K_{\min} \cdot Q'\left(\frac{d^2}{4\sigma^2}\right) \\ &\approx K_{\min} \cdot Q'\left(\frac{d^2}{2N_0}\right) \end{aligned} \tag{6}$$

where $Q'()$ is the Gaussian complementary distribution function, often simply called the q-function. $N_0 = 2\sigma^2$ is the average noise power per hertz. Note that the preceding equations represent the

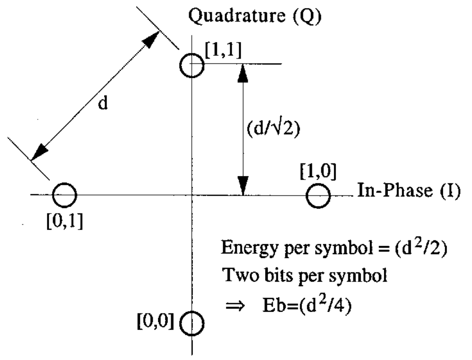


Fig. 5 Signal space representation of QPSK; four information symbols differ in phase, while their amplitude is constant.

symbol error probability. In all but the simplest communication schemes, each symbol represents more than a single bit of information. For example, with the quadrature phase-shift keying (QPSK) modulation scheme used in most satellite communication applications, the phase of the carrier wave is varied to transmit information, such that each of four possible equal-power symbols represents a pair of data values, as shown in Fig. 5. In most well-designed signal sets, adjacent symbols differ only by a single information bit. In these cases, an error in the interpretation of a multibit symbol results in only a single bit error. If each symbol represents m bits, then the probability of bit error, in terms of E_b/N_0, is

$$\text{Pr(bit error)} \approx (K_{\min}/m)\cdot Q'\left[\left(d^2\big/4E_b\right)(2E_b/N_0)\right]$$

$$\approx K_b \cdot Q'[\gamma_c(2E_b/N_0)] \tag{7}$$

where $K_b = K_{\min}/m$ is the average number of nearest neighbors per bit and $\gamma_c = d^2/4E_b$ is defined as the nominal coding gain, a measure of the improvement of a given signal set compared to uncoded binary PAM, in which $\gamma_c = 1$. For QPSK, there is no coding gain since $d^2 = 4E_b$, as shown in Fig. 5. Also $K_b = 1$, and so the bit error rate (BER) is

$$\text{BER} \approx Q'(2E_b/N_0) \tag{8}$$

Additional coding gain can be attained with error-correction coding, which involves further separating the symbols in signal space. For example, QPSK with half-rate Viterbi error correction has $\gamma_c = 2$ such that

$$\text{BER} \approx Q'(4E_b/N_0) \tag{9}$$

Note that $g(x)$ in Eq. (4) is the probability density function of the noise signal at the input to the detector that makes the decisions. This may differ from the density function of the noise at the input to the antenna due to the effects of filters and amplifiers upstream of the detector. For example, consider a simple radar system. A positive radar detection is declared if the envelope (complex amplitude) of the received signal exceeds some predetermined threshold. A radar detector, therefore, includes an envelope detector, to measure the envelope of the signal, and a threshold detector, to actually make the decisions. If the noise entering the envelope detector has a Gaussian probability density function with zero mean and variance σ^2, the probability density function of the noise at output of envelope detector is a Rayleigh distribution[7]:

$$g(x) = (x/\sigma^2)\exp(-x^2/2\sigma^2) \tag{10}$$

In this case, the probability of error, or false alarm, is given by Eq. (5) with $K_{\min} = 1$ and $d/2 = v_T$, the threshold voltage, such that

$$\text{Pr(false alarm)} = \int_{v_T}^{\infty} g(x)\,\mathrm{d}x = \exp\left(\frac{-v_T^2}{2\sigma^2}\right) \tag{11}$$

Information Availability

The availability measures the instantaneous probability that information is being transferred through the network between a given number of known and identified O–D pairs at a given rate and integrity. The availability is a measure of the mean and variance of the isolation, rate, and integrity supportable by the system and as such is sensitive to worst-case scenarios.

Note that availability has a functional definition; it is the probability that the system can instantaneously perform specific functions. In this way, the availability is not a statement about component reliabilities. At any instant, the network architecture is defined only by its operational components, and so all networks are assumed to be instantaneously failure free. Should a component fail, the network changes by the removal of that component. Generally, the capabilities of the new network will be different than those of the previous network.

For a given network, the supportable isolation, rate, integrity, and, hence, the availability, can vary due to the following:

1) There are too many users simultaneously accessing the limited resources of the system. The availability of service to a given user will be poor if the total number of users approaches or exceeds the nominal operating capacity of the system.

2) There is unfavorable viewing geometry and coverage variations. [The spherical error probable (SEP) is the sphere containing 50% of observations.] A system that cannot support continuous coverage of a region will have a low availability for real-time applications. The availability of high-accuracy navigation solutions (SEP $\leq$ 16 m) using the global positioning system (GPS) is dependent on a favorable viewing geometry to several satellites. Spatial and temporal variations in this geometrical dilution of precision dominate the operational availability of GPS. Imaging applications often require specific viewing geometries for each image, effectively limiting the availability of a low-Earth-orbit (LEO) remote sensing system to those times that such a geometry occurs.

3) There are a range variations due to the different elevation angles between the users and the satellites. This is especially true for LEO communication systems in which the range and, hence, free space loss change dramatically as the satellite passes overhead.

4) There is signal attenuation from blockage, rain, or clouds. Clearly atmospheric attenuation can vary geographically and temporally, and the impact on the availability of service can be profound. Visible or ultraviolet imaging is impossible through cloud cover, limiting the availability of such systems. A mobile user of the Big-LEO communication systems will be very susceptible to signal fade from blockage, either by buildings or foliage.

5) There are statistical fluctuations due to noise or clutter. These random variations may be significant if the system is operating close to the limits of its capabilities.

Calculating the Capability Characteristics

These characteristics define the capability of the system, that being the availability of providing an information transfer service between a given number of identified O–D pairs at a given rate and integrity. The capability characteristics are probabilistic measures. The availability is a function of three variables: rate, integrity, and the number of users. For satellite applications, the information rate is usually a deterministic design decision. However, the integrity and the number of simultaneous users can be considered random variables, the former being sensitive any variations in the signal power or noise, and the latter being dependent on the market. Although it is often difficult to predict the statistics of the market, probability distribution functions for the signal power and the noise power can be predicted reasonably well from the statistics of the satellite's orbit and elevation angle, probabilistic blockage or rain attenuation models, and component performance specifications.

Calculating the capability characteristics, therefore, involves tracking the statistics of the information signals delivered to the end users. The network representation of satellite systems provides the framework for these calculations. Statistical distributions can be propagated through a network sequentially, calculating the changes to the distribution functions as a result of the transitions through each node along a path from source to sink.

Consider an arbitrary system component with input signals X and Y and an output signal Z. Treat X and Y as random variables with distribution functions $F_1(x)$ and $F_2(y)$, and probability density functions $f_1(x)$ and $f_2(y)$, such that

$$F_1(x) = \Pr(X \leq x) = \int_{-\infty}^{x} f_1(v)\,\mathrm{d}v \tag{12}$$

$$F_2(y) = \Pr(Y \leq y) = \int_{-\infty}^{y} f_2(v)\,\mathrm{d}v \tag{13}$$

If the output $z = g(x)$ is a function of only one input x, then the random variable Z has a distribution function $F_z(z)$ given by

$$F_z(z) = \Pr(Z \leq z) = F_z[f(x)] = F_1(x) \tag{14}$$

Generally, the output is a function of more than one input, such that $z = g(x, y)$. If X and Y are independent,

$$F_z(z) = \Pr(Z \leq z) = \iint_{g(x,y) \leq z} f_1(x) f_2(y)\,\mathrm{d}x\,\mathrm{d}y \tag{15}$$

Provided the transfer functions $g(\,)$ of each component are known, these equations describe how to propagate the probability distribution functions for the signal power and noise power through the network. The probability distribution function for the integrity of decisions made at a detector can then be evaluated, again using Eq. (15), with the two random variables being E_b and N_0.

Note that some networks include several detectors that make interpretations of the information at intermediate points along the path from source to sink. Any information symbols that are interpreted erroneously by an intermediate detector will be received in error at the next detector before any interpretation is even performed. The net error probability (integrity) is the combination of the errors incurred at each detector. The probability distribution of these errors is once again calculated using Eq. (15), where the random variables are now the error probabilities for the decisions at each detector.

The probability distributions for the integrity of information transfers between a given number of identified O–D pairs at a variety of different rates, thus, can be calculated. These distributions define the availability of providing this information transfer service.

Example Capability Characteristics for a Ka-Band Communication Satellite

Consider the information flow through a typical satellite from one of the proposed Ka-band communication systems. Figure 2 shows a possible network diagram for one such satellite. The modeled system parameters are given in Table 1.

Start at the left-hand side of Fig. 2: Consider first the uplink from the users to the satellite. The modeled satellite employs a TDM/FDMA scheme for each of 48 uplink spot beams. This means that each user transmits information within a specified frequency band, and at specified times, isolating the different users of each spot beam. Note that because the maximum transmitted power of the user terminals is limited, the energy per symbol depends on the user transmission rate.

Each signal then passes through the atmosphere, which attenuates the power (and introduces noise) by varying degrees depending on the local climate, the frequency of the rf carrier, and the elevation angle of the line of sight. The probability distribution for the likely attenuation can be predicted reasonably well using the familiar Crane rain attenuation model.[8] There is additional attenuation from free-space loss, again with a probability distribution due to the distribution of elevation angles for users within the field of view (FOV). The power of the signal arriving at the satellite antenna, therefore, has a statistical distribution. Noise power from thermal noise and cross-source interference (imperfect signal isolation) lead to small average signal-to-noise ratios. The power of each signal entering the DSP is, therefore, weak and varying. The DSP must detect the information symbols and reroute them to their destination. Recall that the integrity of the detection process scales exponentially with E_b/N_0. Therefore, there is a statistical distribution for the BER of each signal leaving the DSP, and the distribution will be different for different user information rates.

Table 1 System parameters for a modeled Ka-band communication satellite

Parameter	Unit	Value
Miscellaneous system parameters		
Mission	Broadband communications	
Market	Western European residential users	
Number of satellites		1
Orbit		25°E GEO
Uplink parameters		
Multiple access scheme	Spot beams + TDM/FDMA	
Modulation	QPSK, 1/2-rate Viterbi error correction	
Frequency	GHz	30
User terminal effective isotropic radiated power (EIRP)	dBW	44.5
Number of uplink spot beams		48
Satellite antenna gain	dB	46.5
System temperature	dBK	27.6
Losses	dB	1.5
Downlink parameters		
Multiple access scheme	Spot beams + TDMA	
Modulation	QPSK, 1/2-rate Viterbi error correction	
Frequency	GHz	20
Number of downlink spot beams	——	48
Channels per beam	——	1
Channel bandwidth	MHz	125
Channel capacity	Mbps	92
Satellite EIRP	dB	59.5
VSAT[a] antenna gain	dB	43
System temperature	dBK	24.4
Losses	dB	1.5

[a] VSAT, Very Small Aperture Terminal.

The downlink involves a single TDM wideband carrier for each of the 48 spot beams. The net information rate of this downlink is the sum of the rates for all users within the beam. This means that the energy per symbol of the downlink stream is a function of both the user information rate and the number of users. A larger numbers of users at a higher rate per user results in a lower energy per symbol.

The downlink signal is also attenuated by the atmosphere, free-space loss, and interference. Individual end users must demultiplex the received signal, extracting only the parts relevant to them. Here, isolation of the correct information signal depends on the stability of the oscillators in the user terminals. Extraction of the wrong information is effectively a multiple-symbol error. The subsequent interpretation of information symbols is sensitive to the received energy per symbol. Recall, however, that some symbols were interpreted erroneously by the satellites. These symbols are received at the user terminals in error before any interpretation is even performed. Therefore, the net symbol error rate is a combination of the errors incurred at the satellite and at the user terminals.

In this example, the rate of information transferred through the system for each O–D pair is a design decision. The integrity of that information, as measured by the symbol error rate, has a statistical distribution depending on the number of users and the rate at which they transmit. The resulting availability of service varies across the range of operating conditions. The capability characteristics for this network are shown in Fig. 6 for two different rates and two different numbers of users. The capability characteristics shown here were calculated using elevation angle statistics for users distributed across western Europe, accessing a geostationary satellite located at 25°E longitude.

These characteristics can be used to determine the maximum number of users that the system can support at a particular rate and integrity. Note that the availability for 3000 users at T-1 rates (1.544 Mbps) is below 95% over all BERs of interest. (Note that it is generally assumed that BERs of 10^{-9} or 10^{-10} are acceptable

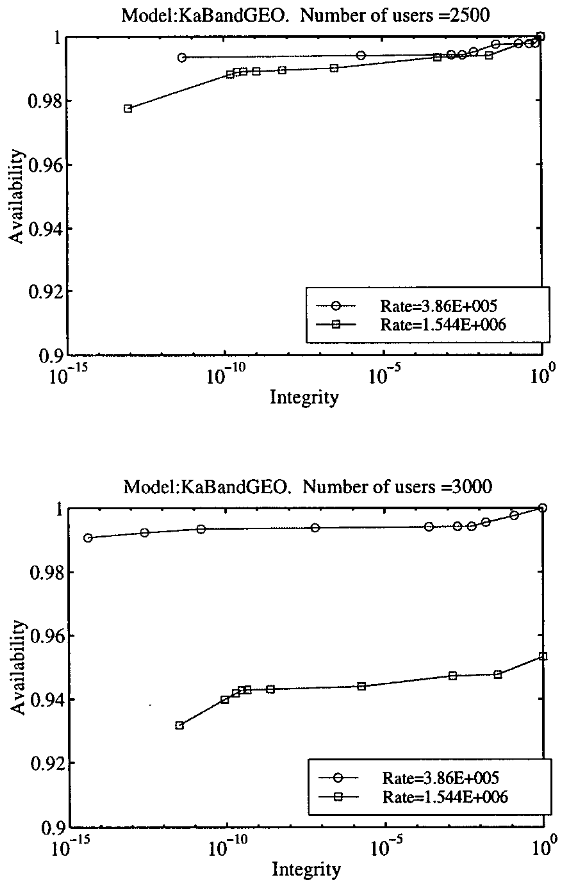


Fig. 6 Capability characteristics for a modeled Ka-band communication satellite.

for broadband services.) This is a result of the demand exceeding the downlink capacity of the satellite. Users must then be queued, reducing their effective availability.

Generalized Performance

The formulation of the capability characteristics allow us to calculate the generalized performance of satellite systems. Performance is perceived in terms of satisfying the demands of a market. This demand is represented by a set of functional requirements, specific to an individual information transfer. The requirements specify minimum acceptable values for 1) signal isolation, 2) information rate, 3) information integrity, and 4) availability of service at the required isolation, rate, and integrity. For instance, consider the market for mobile voice communication. Typically, the requirement is that individual users have at least 95% probability of being able to transmit and receive from small, mobile terminals at a rate of no less than 4800 bps with a maximum BER of 10^{-3}. Note that the isolation requirement enforces that the system be able to address each mobile, individual user within the distributed market. Also note that these functional requirements make no reference to the size of the market being served; they simply specify the quality of service that must be provided to the users.

Performance should always be defined relative to these requirements. To be unambiguous and quantifiable, performance should represent the likelihood that the system can satisfy the functional requirements for a certain number of users from a given market. In short, the performance of a system within a given market scenario is the probability that the system instantaneously satisfies the top-level functional requirements that represent the mission objectives.

It is important to note that performance is distinct from capability, although the two are related. The capability characterizes a particular network's ability to transfer information between a given number of identified users at different rates and integrities. There is no implicit reference to requirements within the definition of capability, and component reliabilities are not reflected. However, a measure of performance should include all likely operating states, and so reliability considerations are necessary. The existence of component failures means that every system has many possible network architectures corresponding to failures in different components. Each network, or system state, is defined only by the components that are operational. Each of these states will have different capabilities. By specifying requirements on isolation, rate, and integrity, the capability characteristics can be used to determine the availability of service offered by each state, for different numbers of users. If the supported availability exceeds the minimum acceptable availability specified by the functional requirements, that system state is deemed operational. The mathematical formulation of the generalized performance follows immediately: The generalized Performance for a given market scenario is simply the probability of being in any operational state.

Therefore, the performance can be improved by either reducing the impact of any component failures that could occur or by improving the component reliabilities so that these failures are less likely. The former approach effectively increases the number of operational states, whereas the latter reduces the probability of transitioning to a failure state. The impact of component failures, blockage, or rain/cloud cover can be reduced if there are redundant information paths. This redundancy can be provided by distributed architectures featuring multifold coverage. For example, a mobile communication user can select, from all of those in view, the operational satellite with the clearest line of sight. This can reduce service outages and improve availability. This concept extends across almost all applications.

Time Variability of Performance

The performance can be quantified for each year over the lifetime of the satellite system to give the performance profile. The performance of the system generally changes in time as a result of three factors:

1) There are typically different rate, integrity, and availability requirements placed on a satellite system at different times within its life. Consequently, the functional requirements are properly specified as an availability profile.

2) System components have finite failure probabilities, and once on orbit, a satellite system is difficult to repair. Thus, there is a higher probability of being in a failed state with a degraded availability late in the lifetime.

3) The number of users targeted by the system will usually change over the lifetime. As shown in preceding sections, the supported availability of a system is a strong function of the number of users.

These trends can compound to give large variations in the performance over the system lifetime.

Calculation of the Generalized Performance

Because the context of its definition includes the notion of state probabilities, the calculation of the generalized performance is well suited to Markov modeling techniques that determine the probability of being in any particular state at a given time. In general, Markov calculations rely the state probabilities $\boldsymbol{P}_s(t_1)$ at some future time t_1 depending only on the current state probabilities $\boldsymbol{P}_s(t_0)$ and on the rate of state transitions[9]:

$$\boldsymbol{P}_s(t_1) = \boldsymbol{A} \cdot \boldsymbol{P}_s(t_0) \tag{16}$$

where $\boldsymbol{A}$ is the state transition matrix. Determination of this matrix requires the characterization of each state as an operational state or a failure state. Herein lies the only complication in calculating the generalized performance compared to conventional Markov modeling. To ascertain whether a state is operational, the capability characteristics of that state must be calculated and compared to the requirements. Because this is nontrivial, generation of the state transition matrix involves a large amount of computation and in most cases dominates over the computations involved in the actual solution of Eq. (16). Therefore, the complexity of the performance calculations grows linearly with the number of possible states because each must be investigated. However, the number of possible states increases geometrically with the number of failure transitions and the number of system components. For this reason, the models usually include fewer than 10 failure transitions from a subset of the most critical system components. State aggregation techniques can also be used to reduce the number of computations.

Example Performance for a Ka-Band Communication Satellite

To illustrate calculation of the generalized performance, return once again to the broadband communication system of Fig. 2 and Table 1. For demonstration purposes, let us assume that 2500 users of the system require availability of at least 98% for communication at a data rate $R = 1.544$ Mbps and a BER of 10^{-9}. By the use of reasonable values for the failure rates of the most critical system components, the failure states corresponding to a violation of these requirements and the associated probabilities can be calculated. For this simple system, there are basically two different types of failure state: those that correspond to degraded payload operations that violate the requirements and those that constitute a total loss of the satellite. These two scenarios can be modeled separately to simplify the analyses. Consider first the failure states corresponding to degraded operation of the satellite payload.

The most failure-prone components along the primary information path through the network are the satellite DSPs and the satellite transmitters. The system shown in Fig. 2 features 48 channels for each of these, one pair for each spot beam, with cross connections to remove serial failure modes. Note that this is not representative of the proposed Ka-band systems that have multiple, redundant DSPs and transmitters. For example, the published plans for the Spaceway system, proposed by Hughes Communications, Inc.,[10–12] include fully redundant cross-strapped DSPs and 64-for-48 redundancy in the transmitters. Such levels of hardware redundancy, mated with technological improvements that reduce the component failure rates, result in a very small probability of payload failure. It can be assumed that a sensible design would feature such redundancy, and at least for this example, we can ignore the effects of degraded payload operations.

The second type of system failure corresponds to a total loss in operational capability of the satellite. This satellite vehicle failure (SVF) can occur when the support modules fail to provide the functional modules with essential resources. For example, the power and propulsion subsystems, the guidance and navigation subsystem (G&N), and the spacecraft control computer (SCC) must all work under normal operations. Calculating the probability of SVF simply involves building a simple model of the satellite bus resources. For this simple example, the spacecraft was modeled to include two parallel SCCs, two G&Ns, and an integrated bus module representing propulsion, power, and structural components. In the G&N and SCC, 1 out of every 10 failures were assumed to be unrecoverable. The equivalent channel failure rates, taken from Ref. 13, are $\lambda_{\mathrm{SCC}} = 0.0246$, $\lambda_{\mathrm{G\&N}} = 0.0136$, and $\lambda_{\mathrm{bus}} = 0.072$, all per year. The resulting probabilities of the SVF modes are shown as a function of time in Fig. 7.

For this example, the failure probability is dominated by the probability of a bus failure. The overall probability of satellite failure exceeds 0.5 after 10 years. The generalized performance of this example satellite system is the complement of the net failure rate, which drops from unity at time 0 to a value just less than 0.5 after 10 years.

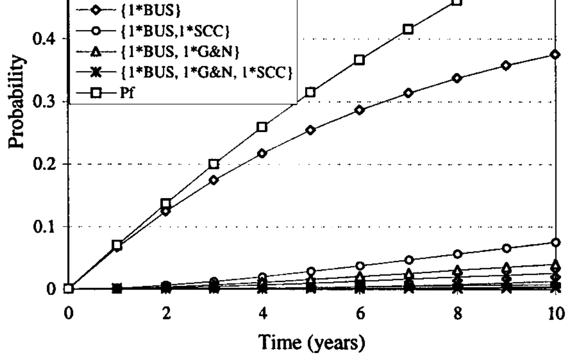


Fig. 7 Failure state probabilities for a modeled Ka-band communication satellite.

Cost Per Function Metric

The cost per function (CPF) metric is perhaps the most important concept introduced within this analysis framework. Its definition is completely generalizable and straightforward: The cost per function metric is a measure of the average cost incurred to provide a satisfactory quality of service to a single O–D pair within a defined market. The metric amortizes the total lifetime system cost over all satisfied users of the system during its life.

The mathematical form of the metric follows immediately from this definition and is the same across all applications:

$$\mathrm{CPF} = \frac{\text{lifetime cost}}{\text{number of satisfied users}} \tag{17}$$

Note that number of users of the system is represented by the number of O–D pairs and the information symbols they exchange. For example, the number of users of a communication service is defined by the total number of bits transferred through the system. Equivalently, the number of users for a space-based search radar is the total area that is searched. However, this alone is insufficient because the definition of a market implicitly includes minimum requirements on the quality of service: the isolation between sources and the rate and integrity of the information being exchanged. Users within the market are only satisfied when the information transfers occur between the correct O–D pairs at the correct rate and with the correct integrity. Therefore, the metric is based on the number of satisfied users, referring to the total number of symbols transferred through the system that satisfy requirements.

Before proceeding, it is helpful to introduce some examples of the CPF for different applications, to concrete understanding of the principal terms. Table 2 summarizes the CPF for a mobile voice communication system,[14] a broadband communication system,[15] a surveillance radar system [ground moving target indicator (GMTI)] for the detection of ground moving targets, and an astronomical telescope.[16,17]

Both of the communication systems must support a quality of service that people will be willing to pay for, a service that is billable. The market for voice requires symbol rates that can support a voice circuit, defined as a full-duplex voice connection of predetermined quality between two users. The quantity of these voice circuits can be measured in minutes. For broadband service, the information rate must be higher, with multimedia applications requiring data rates around T1 (1.544 Mbps).

The surveillance radar must provide a level of service that allows a theater of a given size to be adequately protected. This requires that each square kilometer be safety checked every minute. The total number of protected square kilometers is then total area protected each minute, multiplied by the number of these minute-long intervals in the lifetime of the system. As a result, the time dimension is not explicitly stated in the metric, but is implicit in the definition of protected. Similarly, for the telescope, the concept of the useful image implies a satisfactory resolution, update rate, and image integrity. Again, the time dimension does not appear explicitly, being swallowed by the useful construct.

In every case, the CPF has the dimensions of dollars-per-information symbol. Recall, however, that information symbols represent users of the system. Therefore, although the dimensions of a symbol are strictly bits, a symbol generally has an interpretation, such as a voice circuit or an image. Indeed, by definition, the

Table 2 CPF metrics for example applications

System	Cost per	Satisfied	User
Mobile communications	Cost per	Billable	Voice-circuit minute
Broadband communications	Cost per	Billable	T1 minute
GMTI radar	Cost per	Protected	km^2 of theater
Astronomical telescope	Cost per	Useful	Image

dimensionality of the CPF metric must be equivalent to dollars per user.

Calculating the CPF Metric

To calculate the CPF metric, the impact of improved performance on the cost of a system must be determined. If the value of performance can be quantified, the system cost can be modified to correspond to a common level of performance. The modified system cost should represent the total lifetime cost of a system, where lifetime cost is defined to be the total expenditure necessary to continuously satisfy the top-level system requirements.

System Lifetime Cost

The baseline cost C_s accounts for the design, construction, launch, and operation of the system components. This baseline cost does not, however, account for the expected cost of failures of system components. Because the system must satisfy requirements throughout its design life, expenditure will be necessary to compensate for any failures that cause a violation of this condition. These additional failure compensation costs V_f (Ref. 13) must be added to the baseline system cost to give the total lifetime cost C_L

$$C_L = C_s + V_f \tag{18}$$

As long as it is used consistently, any cost model can be used to calculate the baseline system cost. Note that a premium is paid for more reliable components.

Because some costs are incurred at different times within the lifetime of the system, the cost is actually represented as a cost profile. This profile has to be modified to account for the time value of money. Costs incurred later in the system lifetime have a lesser impact on the overall system cost. A dollar is always worth more today than it is tomorrow; capital expenditure can earn interest if invested elsewhere. Therefore, the yearly costs are discounted according to an assumed discount rate corresponding to an acceptable internal rate of return (IRR). To attract investors to commercial systems, the high risk associated with space ventures necessitates a high IRR of around 30%. For government projects, a discount rate of 10% is often used in costing analysis.[13]

The discounted cost profile $c_s(t)$ must then be integrated over the system lifetime to obtain the total baseline system cost C_s

$$C_s = \sum_{\text{life}} c_s(t) \tag{19}$$

Failure Compensation Cost

The failure compensation cost V_f can be estimated from an expected value calculation:

$$V_f = E[V_f] = \sum_{\text{life}} \left[\sum_{\text{states}} p_s(t) \cdot v_s(t) \right] \tag{20}$$

where $p_s(t)$ is marginal probability of entering failure state s at time t and $v_s(t)$ is the sum of the economic resources required to compensate for the failure. Strictly, this calculation should involve all likely failure states. However, for complex systems this is prohibitive. A reasonable approximation is to truncate the model and include only the states representing the most likely failure modes.

Note that v_s includes the costs of replacement satellites or components, launch costs, and any opportunity costs representing the revenue lost during the downtime of the system. The calculation of v_s is architecture specific and in most cases depends strongly on the nature of the failure mode. A failure mode and effects analysis may be required to estimate the replacement costs. Estimation of the opportunity costs is difficult, requiring a prediction of the failure duration. Despite these problems, v_s can be estimated with reasonable confidence using predictive methods and for simple systems and with simulations for more complex systems. Of course, good market models are also required.

The marginal probabilities $p_s(t)$ of the most likely failure states are the derivatives of the failure state probabilities $P_s(t)$ that are evaluated during the Markov calculation for the generalized performance. It is, therefore, through the failure compensation costs that performance impacts the system lifetime costs. A higher performance system will have a lower probability of transitioning to a failure state and, consequently, a lower expected value of the compensation costs.

System Capture

In a perfect market scenario, the system capture M_c (equivalent to the total number of satisfied users) can be chosen using the capability characteristics and the performance profiles. This would simply be the maximum number of users that the system could satisfy, given a set of requirements. Essentially there is a tradeoff between providing basic service to a large number of users or ensuring high performance to a small number of users. For example, a system that can serve a small number of users with a high probability could instead target a larger number of users at a lower (but still satisfactory) availability. This strategy carries the risk of being more sensitive to component failures, essentially incorporating less performance redundancy. The optimum strategy depends on the expected revenue, the estimated compensation costs, and in particular the opportunity costs associated with dissatisfied customers.

Note, however, that it is usually incorrect to assume a perfect market, and it is then necessary to include a comparison to the size of the expected market. This step is called demand matching and is critical because a system cannot outperform the demand. Extra capacity beyond the market size brings no additional revenue or benefit, but may incur increased costs.

Comparing the design capacity to the size of the local demand and taking the minimum give the achievable capacity of the system. This is defined as the market capture. Because the size of the local demand Q is almost always time and spatially varying, the demand matching calculation involves an integration over the entire coverage region for each year of the satellite lifetime, to give a market capture profile $m_c(t)$,

$$m_c(t) = \sum_{\text{market}} \min\,[\text{design capacity, } Q] \tag{21}$$

Recall that the performance and, hence, failure compensation depended strongly on the number of users addressed. Of course, the opportunity costs associated with lost revenue during downtime is also dependent on the number of addressed users. The market capture profile can be used to determine the maximum number of users that can be served at different times over the lifetime of the satellite.

A further complication arises if the system operation results in monetary income, as is the case for commercial communication systems. In this situation, the time value of money means that there is also a bias in the relative value of market capture, with a weighting toward the start of the systems lifetime. In general, revenue should be earned as close as possible to the time that the associated costs are incurred. For example, revenue earned from the transmission of bits early in the life of a communication satellite is more important than revenue earned late in the lifetime. For this reason, for each year of the lifetime of the satellite, the capture profile $m_c(t)$ must also be correctly discounted according to the same discount rate as was used to discount the costs.

The total number of satisfied users or system capture M_c is then calculated by summing the capture profile over the entire lifetime of the system:

$$M_c = \sum_{\text{life}} m_c(t) \tag{22}$$

Having determined the total system capture, the CPF can now be calculated:

$$\text{CPF} = C_L / M_c \tag{23}$$

Example CPF Calculation for a Ka-Band Communication Satellite

The cost per billable T1 minute is the CPF metric used in the analysis of broadband satellite systems. It is the cost per billable T1

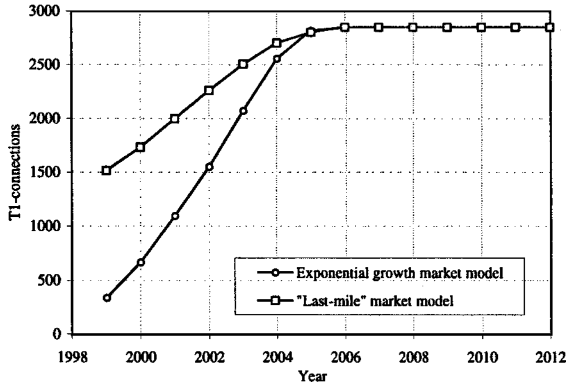


Fig. 8 Market capture profile for a modeled Ka-band communication satellite; two market models represent different projections for the size and distribution of the European residential broadband market.

minute that the company needs to recover from customers through monthly service fees, ground equipment sales, etc., to achieve a specific (30%) internal rate of return.

Once again referring to the example Ka-band system described in Table 1, the cost per billable T1 minute can be calculated from an estimation of the system's market capture and the system costs. The system is assumed to reach initial operating capability (IOC) in 1999 and to be active through the year 2010, requiring a satellite lifetime of 12 years. The calculations are all performed in fiscal year 1996 dollars because this would represent a reasonable project inception date, given an IOC in 1999. All costs are adjusted using the Office of the Secretary of Defense estimates[13] and are discounted back to a present value in 1996 with a 30% discount rate. Consider first the evaluation of the achievable market capture of the system.

The market capture depends on the size of the market accessible to the system and on the system capability characteristics. The limiting effects of market demographics, access, and exhaustion can be quantified only with an adequate market model. For an earlier study, Kelic et al.[15] constructed several reasonable models for the global broadband communications market, based on current and projected internet usage and computer sales growth. Using these market models, computer simulations of several broadband satellite systems have been performed to estimate their market capture. Figure 8 shows the resulting market capture profile for the modeled satellite.

The achievable capacity of the satellite initially grows as the market develops. After 2005, the market capture saturates at around 2800 simultaneous users. If additional users were addressed, the supported availability would drop below requirements, as seen in the capability characteristics of Fig. 6.

The total market capture is the sum over all years of the market capture profile, after discounting at a rate of 30% per year to represent the net value of the revenue stream in 1996. For the exponential market model, the resulting market capture in equivalent fiscal year 1996 T1 users is only 2560. Note that this discounted total is smaller than the true value in any individual year from 2004 onward. This is a direct result of the diminishing value, in real terms, of any revenue earned later in the lifetime of commercial projects. The value of M_c used in the cost per billable T1 minute metric is then simply this number of equivalent T1 users, multiplied by the total number of minutes in a year, so that $M_c = 1.346 \times 10^9$ T1 minutes.

The total baseline cost of the satellite system is estimated including recurring and nonrecurring costs for development, construction, launch, insurance, gateways and control center operations, and terrestrial internet connections. The cost model used for this example is the same as that used by Kelic et al.,[15] which draws on industry experience and observed trends. The theoretical first unit (TFU) cost for communication satellites can be estimated reasonably well assuming $77,000/kg of dry mass. The nonrecurring development costs for commercial systems can be approximated at three to six times the TFU cost, depending on the heritage of the design. For this example, launch costs to geostationary Earth orbit (GEO) can be assumed at $29,000/kg, with insurance at 20%. For linking to the terrestrial network, each OC-3 (155 Mbps) connection costs $8500 installation and $7900 per month. This cost scales with the market capture.

Table 3 System cost profile for a single Ka-band communication satellite

Year	c_s ($\$\times10^6$)	p_f	v_s ($\$\times10^6$)	v_f ($\$\times10^6$)	c_L
1997	264.000	——	——	——	264.000
1998	264.000	——	——	——	264.000
1999	145.000	0.070	0.500	0.035	145.035
2000	1.000	0.067	14.720	0.980	1.980
2001	1.000	0.063	14.490	0.911	1.911
2002	2.000	0.059	14.360	0.852	2.852
2003	2.000	0.056	14.260	0.794	2.794
2004	3.000	0.052	14.320	0.749	3.749
2005	3.000	0.049	14.140	0.693	3.693
2006	3.000	0.046	13.780	0.630	3.630
2007	3.000	0.043	13.250	0.564	3.564
2008	3.000	0.040	8.170	0.324	3.324
2009	3.000	0.037	5.150	0.190	3.190
2010	3.000	0.034	2.440	0.083	3.083

The expected failure compensation costs are calculated from the marginal SVF probability profile $p_f(t)$ (Fig. 7) and the market capture curves (Fig. 8). A satellite failure can be assumed to result in the loss of a single years' revenue, together with the cost of building and launching a replacement satellite. The calculation of the opportunity costs from lost revenue requires an assumption for the average service charge per user. A conservative estimate of $0.05 per T1 connection is used for this example. The baseline system cost and the failure compensation costs can be summed to give c_L, the system cost profile. The baseline costs $c_s(t)$, failure compensation costs $v_f(t)$, and total system costs $c_L(t)$ are shown in Table 3.

Discounting the system cost profile at 30% per year gives the net present value of the costs in fiscal year 1996 dollars. Summing over all years of the discounted profile gives the total lifetime cost, $C_L = \$429$ million. The cost per billable T1-minute metric for this system in an exponentially growing broadband market is, therefore, simply

$$\text{cost per billable T1-minute} = C_L/M_c = \$0.32$$

This implies that the company must be able to charge users at least $0.32/min for this broadband service to obtain a 30% return on the investment.

Utility of the CPF Metric

Part of the utility of this CPF metric is that it permits comparative analysis between different systems with large architectural differences, by scaling their cost according to their performance and market capture. Very large and ambitious systems can be fairly compared to smaller, more conservative systems. The CPF metric can also be used to assess the potential benefits of incorporating new technology in spacecraft designs. New technology should only be included in the design of a new satellite if it can offer reduced cost or improved performance. This can be evaluated with the metric, provided that both the cost and the expected reliability of the new technology can be estimated. Commonly, the largest problem encountered with incorporating new technology in space programs is schedule slip. This can have an adverse effect on the overall success of the program, extending the period of capital expenditure, while delaying operations that bring revenue. These effects can also be captured by the CPF metric. Some typical amount of program slip can be included in the cost profile $c_s(t)$, and the corresponding delay can be applied to the market capture profile. The combined effects of including the new technology will then be apparent, by comparing the CPF metric to those corresponding to designs featuring more established technologies.

Adaptability Metrics

The adaptability metrics judge how flexible a system is to changes in the requirements, component technologies, operational procedures, or even the design mission. It is convenient to define two types of adaptability, different in both their meaning and their mathematical formulation.

1) Type 1 adaptability assesses the sensitivity of the capability, cost, and performance of a given architecture to realistic changes in the system requirements or component technologies. A quantifiable measure of this sensitivity allows the system drivers to be identified and can be used in comparative analyses between candidate architectures. As will be shown in this section, the mathematical form of the type 1 adaptability also makes it entirely compatible with conventional economic analyses of commercial ventures. This adds enormous utility to the metric for investment decision making and business planning.

2) Type 2 adaptability measures the flexibility of an architecture for performing a different mission, or at least an augmented mission set. This is particularly important for government-procured systems. In todays budget-controlled environment, expensive military and civilian space assets must be able to fulfill multiple mission needs cost effectively.

Each of these two types of adaptability has a quantifiable mathematical definition that is a simple extension of the CPF metric.

Type 1 Adaptability: Elasticities

Concisely stated, type 1 adapatabilities represent the elasticity of the CPF metric with respect to changes in the requirements or the component technologies. Elasticity is a mathematical construction most often used in microeconomics. To introduce and formalize notation, it is valuable to briefly summarize the concept of elasticity within the conventional context of microeconomics.

Elasticity is defined as the percentage change that will occur in one variable in response to a 1% change in another variable.[18] For example, the price elasticity of demand measures the sensitivity of the demand for a product to changes in its price and can be written

$$E_p = \frac{\Delta Q/Q}{\Delta P/P} = \frac{P}{Q}\frac{\Delta Q}{\Delta P} \tag{24}$$

where Q is quantity of demand and P is price. Most goods have negative elasticities because price increases result in demand decreases. If the price elasticity is greater than one in magnitude, the demand is termed price elastic because the percentage change in the quantity demanded is greater than the percentage change in price. Consequently, a reduction in the price results in an increase in the total expenditure because disproportionately more goods are sold. An increase in the price results in a reduction of total expenditure because much fewer goods are sold. Conversely, if the price elasticity is less than one in magnitude, the demand is said to price inelastic, and the opposite trends are observed. Finally, a value of unity for the elasticity implies that the total expenditure remains the same after price changes. Any price increase leads to a reduction in demand that is just sufficient to leave the total expenditure unchanged.

Equation (24) specifies that the elasticity is related to the proportional change in P and Q. The relative sizes of P and Q change at different points on the demand curve. Therefore, the elasticity must be measured at a particular point and will usually have very different values at different points along the curve. This, of course, means that the elasticity for a change in price from P_1 to P_2 can be quite different from the elasticity calculated in the other direction, from P_2 to P_1. To avoid this confusion, the arc elasticity represents the average elasticity over a small range:

$$E_p = \frac{\Delta Q/\bar{Q}}{\Delta P/\bar{P}} = \frac{(P_1 + P_2)}{(Q_1 + Q_2)}\frac{\Delta Q}{\Delta P} \tag{25}$$

The choice between using point elasticities and arc elasticities is really the prerogative of the engineer. In general, the arc elasticity is a more consistent measure of sensitivity. For the remainder of this document, the term elasticity is taken to imply arc elasticity, and the overbar is omitted from equations.

Return now to the generalized analysis framework. Analogous to the elasticity of demand, the elasticity of the CPF metric is the percentage change in its value in response to a 1% change in some other relevant variable. The relevant variable here may be a system requirement or a system component parameter. Indeed, it is straightforward to formulate the different requirement elasticities of the CPF at a given design point:

$$E_{Is} = \frac{\Delta \mathrm{CPF}/\mathrm{CPF}}{\Delta Is/Is} \tag{26}$$

$$E_R = \frac{\Delta \mathrm{CPF}/\mathrm{CPF}}{\Delta R/R} \tag{27}$$

$$E_I = \frac{\Delta \mathrm{CPF}/\mathrm{CPF}}{\Delta I/I} \tag{28}$$

$$E_{Av} = \frac{\Delta \mathrm{CPF}/\mathrm{CPF}}{\Delta Av/Av} \tag{29}$$

where E_{Is}, E_R, E_I, and E_{Av} are the isolation, rate, integrity, and availability elasticities of the CPF, with respect to the system requirements Is, R, I, and Av.

Note that ΔCPF is the change in the CPF value as a result of changing a system requirement and is formed by direct subtraction of the CPF values for the two different cases. However, the denominator of the CPF metric carries an implicit reference to these same system requirements, as discussed earlier. Therefore, it is initially tempting to question the validity of simply subtracting two CPF values that have entirely different denominators. The solution to this apparent problem lies in that the CPF metric is defined as the cost per satisfied user. The denominators in all CPF metrics are, therefore, equivalent to a single user, and ΔCPF can be calculated directly. For example, consider the service options that can be provided by a broadband communication system. The cost per billable T1 minute can be compared directly with the cost per one-quarter T1 minute without any modifications. The difference in value ΔCPF represents the difference in cost that must be charged to each broadband user if the data rate provided to them is changed.

In a similar fashion, the technology elasticities can be defined. These can be formed for any particular component of the system that may have an impact on the overall performance or cost. Example technology elasticities are

$$E_{C_{\mathrm{launch}}} = \frac{\Delta \mathrm{CPF}/\mathrm{CPF}}{\Delta C_{\mathrm{launch}}/C_{\mathrm{launch}}} \tag{30}$$

$$E_{C_{\mathrm{mfr}}} = \frac{\Delta \mathrm{CPF}/\mathrm{CPF}}{\Delta C_{\mathrm{mfr}}/C_{\mathrm{mfr}}} \tag{31}$$

$$E_{R_s} = \frac{\Delta \mathrm{CPF}/\mathrm{CPF}}{\Delta R_s/R_s} \tag{32}$$

where $E_{C_{\mathrm{launch}}}$ is the launch cost elasticity, $E_{C_{\mathrm{mfr}}}$ is the manufacture cost elasticity, E_{R_s} is the reliability elasticity, C_{launch} is the budgeted launch cost for the system, C_{mfr} is the manufacturing cost, and R_s is the satellite reliability. In each case, some technology is varied, while the system requirements are held constant. Technology elasticities can be formed for each essential system component, reflecting the likely changes in available technology or the variations in the system parameters that span the design trade space. This allows a quantifiable assessment of design decisions and can identify the most important technology drivers.

Utility of the Elasticities for Economic Analysis

The mathematical form of the elasticities are identical to the conventional elasticities used in econometric analysis. This allows the results from a generalized analysis of a proposed satellite system to be used in the investment decision making process. For example, consider a broadband communication system that had been originally planned to provide users with one-quarter T1 connections. The marketing department then suggests that providing a full T1 connection would give the company a competitive advantage over all others in the marketplace. In addition, they have all of the demand curves to prove it. The system engineer can respond by calculating the rate elasticity of the CPF, as described, for a change from one-quarter T1 to full T1. Because the CPF represents the average cost to provide service to a user, it can be taken to be a surrogate for price. Therefore, the rate elasticity of CPF (or price) can be multiplied by the price elasticity of demand, calculated from the demand curves, to give some number X that represents the change in demand in response to the increase in price associated with improved service. By the comparing of this value to the rate elasticity of demand exhibited by the demand curves, a decision can be made about the rate that maximizes revenue. If X is higher than the rate elasticity of demand, then an increase in the rate results in a disproportionately larger increase in the price, averting more customers than are attracted by the improved service. The marketing departments' idea to offer higher rates must be refused. Alternatively, if X is smaller than the rate elasticity of demand, the engineer can confirm marketings' suggestion with quantitative numbers. Either way, the correct decision can be made.

Type 2 Adaptability: Mission Sensitivity

Type 2 adaptability corresponds to the change in the CPF of a system as the design role is changed or augmented. Recall that a mission is defined by a market and a set of associated derived system requirements. A change in the design mission, therefore, represents a change in the market addressed and all of the system requirements. A classical elasticity formulation that relates a proportional response to proportional variations in the input cannot be constructed because there is no obvious scalar representation of the input variations. Instead, mission sensitivity S is simply defined to be the proportional change in the CPF in response to a particular mission modification:

$$S|_X = \left.\frac{\Delta\mathrm{CPF}}{\mathrm{CPF}}\right|_X \tag{33}$$

where X is just an identifier to specify the mission modification. This is a useful metric for comparing competing designs because it measures just the sensitivity of the CPF to mission modifications, normalizing any differences in the absolute values of the initial CPFs. The sensitivity can be an important factor in deciding between alternate architectures during the conceptual design phase of a program, especially if the mission is likely to change over the lifetime. For example, an architecture that is highly optimized for the baseline mission may have a low CPF but a very high mission sensitivity, implying it is very unsuited to perform any other modified mission. In all but the most predictable markets, a more prudent design choice would be a less optimized system with a lower mission sensitivity, even at the expense of a higher CPF.

Truncated GINA for Qualitative Analysis

For purely qualitative analysis, the GINA methodology can be truncated significantly, while still providing the engineer with valuable insight. In particular, mapping the application into the generalized framework organizes the thought process and allows an unambiguous comparison to be made between competing architectures. The most important discriminators between the systems will be clearly apparent, which allows attention to be focused on the deficiencies or benefits of each architecture. For example, Table 4 shows a qualitative comparison of two very different architectures that have been proposed for a space-based radar to detect ground moving targets.

Discoverer-II, or simply D-2,[19] proposed by the Defense Advanced Research Projects Agency, the National Reconnaissance Office, and the U.S. Air Force, is a constellation of 24 satellites in LEO, each operating independently. The nominal design features satellites in the 1500-kg class, with peak rf power of 2 kW and antenna area of 40 m^2, each costing less than \$100 million. Advanced radar processing techniques, such space–time adaptive processing will be used to cancel clutter for the GMTI mission and principles of synthetic array radar (SAR) will support terrain imaging.

On the other hand, Techsat21,[20] as proposed by the U.S. Air Force Research Laboratory features symbiotic clusters of small satellites (approximately 100 kg, 200-W rf, 1 m^2 of aperture) that form sparse arrays to perform the same mission. The number of clusters is at the moment undecided, depending on the eventual coverage requirements, but for comparison purposes can be taken to be the same as the number of satellites in D-2. Table 4 shows that there are several significant discriminators between these two architectures.

GINA Procedure Steps

The systematic procedure for applying the GINA methodology is summarized.

1) Define the mission objective. What is the real application of the system, in terms of the user needs?

2) Map the user needs into the generalized capability parameters of isolation, rate, integrity, and availability. These define the features of the information transfer that are perceived by the users as quality of service.

3) Construct the network representation from a functional decomposition of the system.

4) Determine functional behavior of each module, in terms of what it does to impact the isolation, rate, integrity, and availability. The modules generally interact with the information via the signal, noise, and interference power.

5) Determine the statistical inputs to each module. Some of the modules require inputs relating to the system characteristics or other parameters, such as elevation angle, coverage, or clutter statistics.

Table 4 Qualitative comparison between Techsat21 and Discoverer-II space-based radar concepts using truncated GINA[a]

Parameter	Discoverer-II	Techsat21
Classification	Collaborative constellation, $n_s = 1$	Symbiotic clustellation, $n_s = 8 - 16$
Isolation		
Clutter compensation	*Clutter cancellation though adaptive clutter processing, nulling, etc.*	*Clutter rejection through sparse aperture synthesis giving narrow main lobe and low sidelobes*
Resolution	*Limited by pulse repetition frequency and antenna dimensions*	*Limited by sparse aperture beamwidth (cluster dimensions)*
Rate (search rate)	*Large aperture has small FOV and so supports a small area search rate (ASR), unless a small dwell time can be tolerated*	*Small apertures have wide FOV that can be filled with multiple receive beams so ASR can be high*
Integrity (P_D)	*High power needed to overcome thermal noise*	*n_s^2 coherent processing gain allows lower power transmitters*
Availability	Dominated by coverage statistics (access, range to target, grazing angle)	Dominated by coverage statistics (access, range to target, grazing angle)
Performance	*Depends on reliability and survivability of single satellite*	*Improved reliability from in-built redundancy and graceful degradation*
CPF	Moderate number of large satellites leads to baseline costs around $\$3\times10^9$ *Poor performance leads to high failure compensation costs*	Large number of small satellites leads to baseline costs around $\$3\times10^9$ *Higher performance leads to smaller failure compensation costs*
Adaptability	*Resolution, rate, and integrity are fixed by power and aperture resources. System can easily support SAR imaging, but cannot perform airborne moving target indication (MTI)*	*Capabilities can be improved by augmenting with more cluster satellites. Imaging is supported, and airborne MTI is possible with more satellites*

[a] Discriminators between architectures in italics.

6) Choose a number of O–D pairs that will be served and determine their isolation characteristics (domain of separation, spacing in that domain, signal spectrum, etc.).

7) For that number of O–D pairs, calculate the integrity of information transfers for a variety of rates. These are the capability characteristics.

8) Set values for the capability parameters corresponding to user requirements for the market.

9) Assign failure rates to each functional module that represents real hardware.

10) Use Markov modeling to calculate the state probabilities corresponding to different combinations of failed components. The sum of the probabilities for those states that satisfy requirements is the generalized performance. Those states that do not satisfy requirements are the failure states.

11) Calculate lifetime cost as the sum of the baseline cost and the failure compensation costs, which are the products of the failure state probabilities and the costs required to compensate for the failures.

12) For a realistic market scenario, calculate the market capture as the maximum number of users that can be addressed satisfactorily.

13) Calculate the CPF as the ratio of the lifetime cost and the market capture.

14) Calculate adaptability metrics by repeating the analysis after changing either a requirement or a technology.

Summary

A generalized analysis methodology has been developed that allows systems with dramatically different space system architectures to be compared fairly on the basis of cost and performance. The framework is very generalizable and can be applied to any satellite mission in communications, sensing, or navigation. The most important concepts of the GINA can be stated concisely.

1) Most satellite systems are information transfer systems that serve O–D markets for the transfer of information symbols.

2) The capabilities of a system are characterized by the isolation, rate, integrity, and availability parameters.

3) Each market specifies minimum acceptable values for these capability parameters. These are the functional requirements placed on the system.

4) Performance is the probability that the system instantaneously satisfies the top-level functional requirements. It is here that component reliabilities make an impact.

5) The CPF metric is a measure of the average cost to provide a satisfactory level of service to a single O–D pair within a defined market. The metric amortizes the total lifetime system cost over all satisfied users of the system during its life.

6) The adaptability metrics measure the CPF sensitivity to changes in the requirements, component technologies, operational procedures, or the design mission.

These concepts extend across almost all applications. In Ref. 1 the methodology is validated by applying it to the existing GPS system, then a comparative analysis of the proposed broadband communication systems, and finally a design study of a military space-based radar.

Acknowledgment

The authors wish to thank Raymond Sedwick for his help preparing this paper.

References

[1] Shaw, G. B., "The Generalized Information Network Analysis Methodology for Distributed Satellite Systems," Ph.D. Dissertation, Dept. of Aeronautics and Astronautics, Massachusetts Inst. of Technology, Cambridge, MA, Feb. 1999.

[2] Ahuja, R. K., Magnanti, T. L., and Orlin, J. B., *Network Flows. Theory, Algorithms and Applications*, Prentice–Hall, Upper Saddle River, NJ, 1993.

[3] Bracewell, R. N., *The Fourier Transform and its Applications*, McGraw–Hill, 2nd ed., 1986.

[4] Drabowitch, S., Papiernik, A., and Griffiths, H., *Modern Antennas*, Chapman and Hall, 1998.

[5] Wozencraft, J. M., and Jacobs, I. M., *Principles of Communication Engineering*, Wiley, New York, 1965.

[6] Lee, E. A., and Messerschmitt, D. G., *Digital Communication*, 2nd ed., Kluwer Academic, Norwell, MA, 1994.

[7] Barton, D. K., *Modern Radar System Analysis*, Artech House, 1988.

[8] Crane, R. K., "Prediction of Attenuation by Rain," *IEEE Transactions on Communications*, Vol. COM-28, No. 9, 1980, pp. 1717–1733.

[9] Babcock, P., "An Introduction to Reliability Modeling of Fault-Tolerant Systems," Charles Stark Draper Lab., TR CSDL-R-1899, Cambridge, MA, 1986.

[10] Fitzpatrick, E. J., "Spaceway. Providing Affordable and Versatile Telecommunications Solutions," *Pacific Telecommunications Review*, Sept. 1995.

[11] Hughes Communications Galaxy, Inc., "Application of Hughes Communications Galaxy, Inc., for Authority to Construct, Launch and Operate Spaceway, a Global Interconnected Network of Geostationary Ka-Band Fixed-Service Communications Satellites," Federal Communications Commission Filing, 26 July 1994.

[12] Hughes Communications Galaxy, Inc., "Application of Hughes Communications Galaxy, Inc., Before the Federal Communications Commission for Galaxy Spaceway, a Global System of Geostationary Ka/Ku Band Communications Satellites—System Amendment," Federal Communications Commission Filing, 29 Sept. 1995.

[13] Larson, W. J., and Wertz, J. R. (eds.), *Space Mission Analysis and Design*, 2nd ed., Microcosm, Inc., and Kluwer Academic, Norwell, MA, 1992.

[14] Gumbert, C., Violet, M., Hastings, D. E., Hollister, W., and Lovell, R. R., "Cost per Billable Minute Metric for Comparing Satellite Systems," *Journal of Spacecraft and Rockets*, Vol. 34, No. 12, 1997, pp. 837–846.

[15] Kelic, A., Shaw, G. B., and Hastings, D. E., "Metric for Systems Evaluation and Design of Satellite-Based Internet Links," *Journal of Spacecraft and Rockets*, Vol. 35, No. 1, 1998, pp. 73–81.

[16] Jilla, C., and Miller, D., "A Reliability Model for the Design and Optimization of Separated Spacecraft Interferometer Arrays," *11th Annual AIAA/USU Conference on Small Satellites*, 1997.

[17] Stephenson, R., Miller, D., and Crawley, E., "Comparative System Trades Between Structurally Connected and Separated Spacecraft Interferometers for the Terrestrial Planet Finder Mission," Space Engineering Research Center, TR SERC 3-98, Massachusetts Inst. of Technology, Cambridge, MA, 1998.

[18] Pindyck, R., and Rubinfeld, D., *Microeconomics*, 4th ed., Prentice–Hall, Upper Saddle River, NJ, 1998.

[19] Discoverer-II: Briefings to Industry. DARPA Tactical Technology Office Presentation, June 1998," URL: http://www.arpa.mil/tto/dis2-docs.htm.

[20] Techsat21-Space Missions Using Satellite Clusters. Air Force Research Laboratory Factsheet, September 1998," URL: http://www.vs.afrl.af.mil/factsheets/TechSat21.html.

A. C. Tribble
Associate Editor