DIGITAL CATV SERVICES

As the migration to digital network and systems infrastructure continues, community antenna television (CATV) network operators are preparing to offer a number of enhanced broadband services. These networks exploit the multimedia delivery capabilities of digital CATV, providing the end user with access to legacy services such as telephony as well as access to advanced services including high-speed Internet access, multiplayer on-line gaming, enhanced pay-per-view, and video-on-demand.

By supporting all of these services and more via a single coaxial wire to the home, digital CATV networks will provide an integral element of the digital information highway. This article will first identify and characterize the key digital CATV system components and services. A treatment of the relevant mathematical framework needed to model such systems and services will follow, presenting the user with a set of tools for analyzing and designing such systems. Included in this treatment is a design sequence identifying an iterative approach to model development and system design.

SYSTEM ARCHITECTURE

While traditional analog CATV systems have utilized several large headends to provide metropolitan area coverage, digital CATV systems are being deployed utilizing a more advanced digital network architecture (Fig. 1). Large service regions are subdivided into interconnected serving offices. These offices are linked via digital fiber ring networks providing redundancy and self-healing capabilities. At the local serving office, the system may be further decomposed into subsystems consisting of the headend delivery components, the distribution facilities, and the subscriber terminal equipment. Depending upon the size of the system, there may be one or more headends and associated distribution facilities (1).

To ready their systems for mass deployment of digital services, operators have been investing in redesign and engineering of their distribution plant. This primarily has involved (1) the upgrading of the plant's frequency passband from typically 450 MHz to 750 MHz and (2) restructuring of the plant's topology. Operators are moving to nodal-based distribution systems in which old tree-and-branch implementations are being replaced with hybrid fiber coax (HFC) systems providing frequency reuse similar to cell phone antenna distribution schemes (1). Because its nodal architecture provides much better immunity to noise funneling and allows better traffic segmentation, an HFC system provides the needed foundation upon which digital services may be successfully deployed.

As digital services are deployed and consumer acceptance grows, the ability to control and manage bandwidth and plant conditions granularly is paramount to meeting quality-of-service objectives. Operators must have the ability to segment plant fault conditions as well as actively limit the number of subscribers sharing a given facility. An HFC system in conjunction with the appropriate digital delivery technology allows this to occur.

The following sections will introduce the major components and subsystems of a digital CATV network prior to describing and detailing its service suite.

Headend

The headend represents the main signal collection, processing, and master distribution facility in a CATV system. It is comprised of a number of components that provide the ability to receive, process, and redistribute analog and digital signals as well as components that provide system management and control capabilities (Fig. 2).

In a digital CATV system, signals may be received either via digital backbone distribution facilities or via digital satellite. Additionally, analog signals may be similarly collected and either retransmitted in analog form or encoded digitally

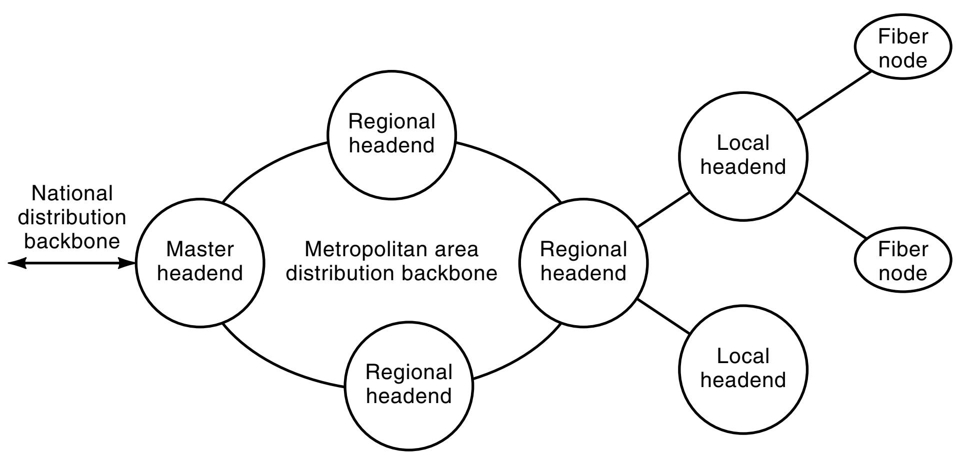

Figure 1. Digital CATV system architecture.

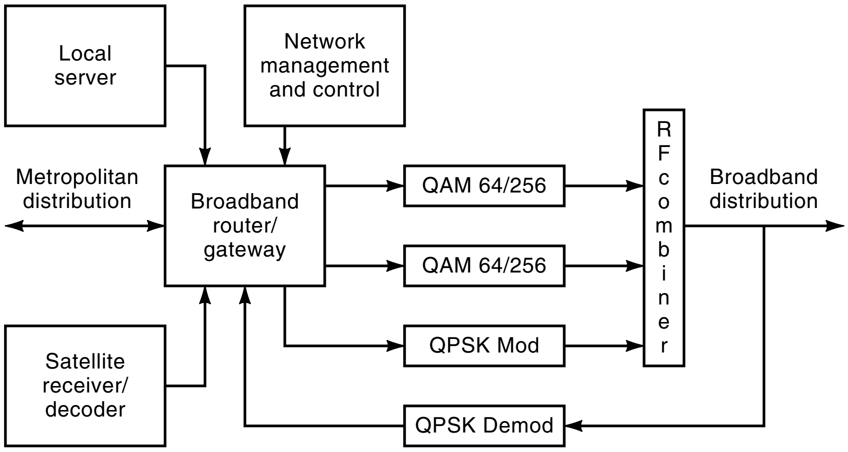

Figure 2. Headend architecture.

prior to distribution to the home. In the case of digital data reception, it also may undergo further processing. For example, digital video received via satellite will be demodulated, potentially transcoded or re-encoded, and then remodulated for transmission to the home.

Data reception is performed by a high-speed multiplexer functioning as a gateway between the backbone distribution network and the broadband-to-the-home CATV distribution plant. This gateway provides the capability to receive data from fiber-based transport facilities such as SONET and provides an interface to the digital modulation equipment used for broadband distribution. Typical interfaces include an OC-3 SONET input interface to the gateway and a set of QAM output interfaces from the associated modulators (2).

Distribution

The distribution architecture utilized for digital CATV services is the aforementioned HFC network. An HFC system utilizes nodal distribution based on a combination of analog fiber and coaxial transport technologies (Fig. 3). Digitally modulated signals, combined electrically in the frequency domain with legacy analog video signals, are input to analog fiber transmission lasers. These lasers utilize optical amplitude modulation to transmit the resulting signal to the destination fiber node. At the destination node, optical-to-electrical conversion occurs and coaxial transmission is used to transport the signal to the home (1).

Subscriber Terminal

The subscriber terminal in a digital CATV system is responsible for processing the collection of signals received at the home and providing the user with access to the range of services offered by the network operator. It is commonly known as a set-top box (STB) (Fig. 4).

The terminal is composed of a number of functional blocks. These blocks consist of network interfaces, application-specific processing units, a general-purpose microprocessor, various memory subsystems providing volatile and nonvolatile storage, and input/output interfaces for command, control and display.

Complete service functionality and access is provided with the addition of downloadable application software. This software provides the graphical user interface, navigation capability, and application-specific functionality.

DIGITAL CATV COMMUNICATION ARCHITECTURE

The communications paradigm for a digital CATV system is based on the integrated use of frequency, time, and packet division multiplexing. This uniquely differentiates its broadband transport architecture from other baseband digital delivery systems.

The network channels resident in the system may be viewed in a hierarchical fashion, with frequency division representing its highest layer. Frequency division is used to par-

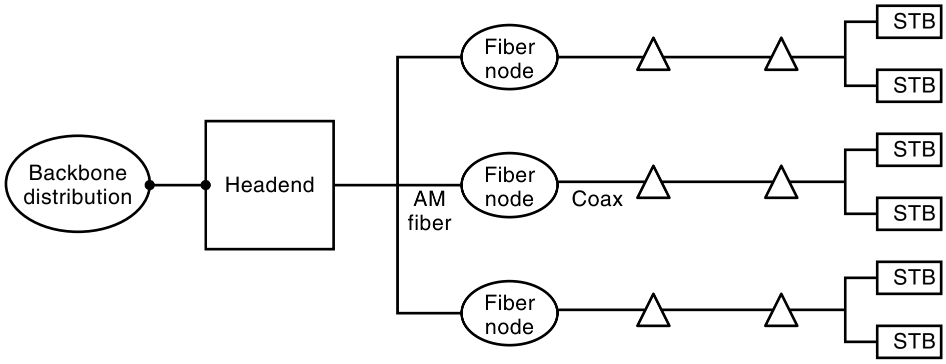

Figure 3. Hybrid fiber coax distribution.

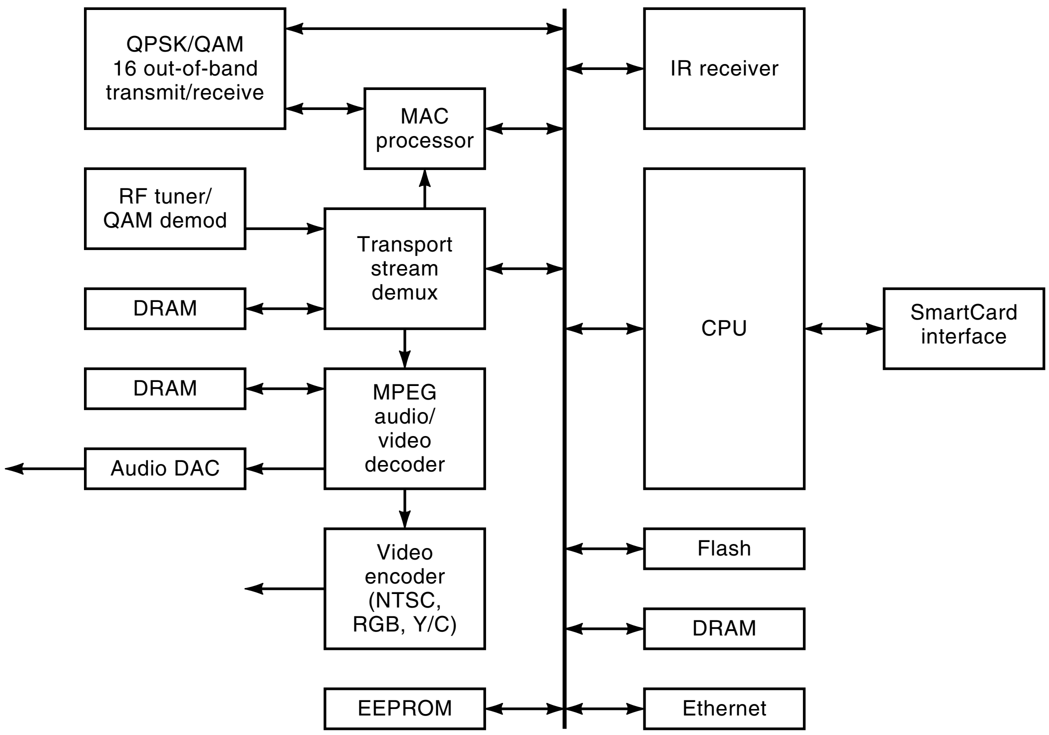

Figure 4. Digital set-top box functional block diagram (3).

tition the broadband spectrum into a number of different service channels. These channels are used not only to segment tiers or classes of service, but also to segment the directions of data transmission.

This segmentation identifies the segments as downstream and upstream. Downstream channels are defined to be those channels providing signal transport from the headend to the home, while upstream channels are those defined to be channels provided transport from the home to the headend.

Within the frequency division, a given service channel or band may use a combination of time division and packet division mechanisms to share usage of its resources. Additionally, because the digital CATV system is based on the use of a shared medium, communication flow from the subscriber to the headend is subject to contention. This introduces the need to provide a medium access control protocol to allow subscribers to efficiently and fairly share the upstream resource (2,4).

Physical Layers

The physical transmission layer varies depending upon the frequency band of interest. Typically, the digital service spectrum utilizes some portion of the bandwidth above 450 MHz to maintain compatibility with existing analog implementations. Within this band, quadrature-amplitude modulation is utilized in either 64 or 256 state mode to provide approximately 30 Mbps to 40 Mbps of transmission capacity per 6 MHz.

Out-of-band control or low-speed data channels are typically implemented utilizing quadrature phase-shift keying (QPSK) to provide 2 Mbps of transmission capacity per 1 MHz of spectrum. In the case of the upstream reverse channel, higher-density modulation schemes such as QAM-16 are also optional modes of transmission (4).

Link Layer and Network Layer

In the downstream channel, the MPEG-2 transport protocol provides the common framework on which additional higher layer protocols are implemented. Some early implementations have examined the use of ATM to the home, but with the growing acceptance of IP and the prevalence of MPEG-2 within the CATV and broadcasting communities, the trend is clearly toward utilizing IP as the common end-to-end network layer. ATM may still be used in the high-speed backbone, where IP over ATM will be utilized to provide interheadend or long-haul connectivity. In such a case, there will be an interworking or gateway function resident in the headend that will reassemble incoming IP packets from ATM cells and then resegment them for transmission via MPEG-2 transport.

Because all traffic downstream is scheduled for transmission via a headend router or multiplexer, a media access control sublayer is not required. In video-only subchannels, MPGEG-2 transport serves as the link layer protocol and provides packet sequencing and error detection and correction. In data or multimedia frequency bands, the MCNS link layer will be used in conjunction with an MPEG-2 sublayer to provide a fixed mapping of variable length frames to fixed length, fixed program identifier (PID) MPEG-2 packets. These frames will consist of MCNS/802.2 encapsulated IP packets. The addressing used within the 802 layer will depend on whether the system is functioning under a bridging or routing paradigm for data transport from the headend to the home. And it should be noted that the MCNS portion of the link layer is not used for addressing but does provide frame typing and security association.

While IP over MPEG-2 is the solution of choice for downstream transport, in the upstream the use of either the DAVIC or MCNS protocols will be used to provide IP services over a shared media channel. Future generations may also see deployment of IEEE 802.14 systems. From the perspective of the headend-to-home subnetwork, data communication will be based on either layer 2 bridging or layer 3 routing (2,4). In all cases, a media access control sublayer (MAC) is used to mediate the shared usage of the upstream channel.

QUALITY-OF-SERVICE FRAMEWORK

Implementing a digital CATV network designed to accommodate a variety of services requires the operator to identify the specific quality objectives associated with each respective offering. These quality objectives are generically referred to as quality-of-service (QOS) parameters. QOS can be interpreted in a variety of ways depending upon the targeted environment. Thus, to ensure a common understanding, a more precise definition of QOS and how it will be used in the context of service development is now provided (5):

> QOS provides a measurement framework in the form of a set of metrics designed to allow objective evaluation and analysis. These metrics are typically reflected in a set of parameters which characterize both the performance of the network as well as the performance requirements of the applications and services.

It is also important to understand that while QOS metrics provide a measurement of the network's performance, their real intent is to provide the operator with metrics such that differentiated class-of-service (COS) may be implemented. COS attempts to provide a framework by which predictability of performance can be introduced. This predictability is introduced by offering prioritized handling of certain types of traffic. Some classes may be given dedicated bandwidth with strict performance bounds while other classes may be processed in a best-effort manner with no guarantee of performance (5).

QOS Metrics

Quality-of-service metrics provide a set of numerical values that are used to evaluate the performance of a network or system. These metrics are defined to characterize the most typical areas of performance and are reasonably simple to calculate (6). Ease of computation becomes an important issue because some of these metrics may be implemented as part of a real-time telemetry system.

- *Average Delay.* Defined as the average of the instantaneous values of elapsed time between the instant a message is ready for transmission and the time until the last bit of the message has been received. Depending upon the subsystem of interest, the delay may be measured between various end-points or across various layers of the communication protocol stack.

$$\mu_{\mathrm{D}} = \frac{1}{N}\sum_{i=1}^{N} d_i \tag{1}$$

where μ_{D} is the average delay, N is the number of samples, and d_i is the ith sample delay.

- *Delay Variation.* Defined as the variance of the instantaneous message delays. This parameter is typically computed on an individual class-of-service basis.

$$\sigma_{\mathrm{D}}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(d_i - \mu_{\mathrm{D}})^2 \tag{2}$$

where σ_{D}^2 is the delay variation, N is the number of samples, d_i is the sample delay, and μ_{D} is the average delay.

- *Delay Coefficient of Variation.* Defined as the ratio of the standard deviation of delay to the mean of the delay

$$\mathrm{CV}_{\mathrm{D}} = \frac{\sigma_{\mathrm{D}}}{\mu_{\mathrm{D}}} \tag{3}$$

where CV_{D} is the delay coefficient of variation, σ_{D} is the standard deviation of delay, and μ_{D} is the average delay.

- *Packet Loss Rate.* Defined as the total number of packets received successfully divided by the total number of packets transferred.

$$P_{lr} = \frac{total\_packets\_input - total\_packets\_received\_successfully}{total\_packets\_input} \tag{4}$$

where *total_packets_received_successfully* represents the total number of packets transmitted successfully to the receiver and *total_packets_input* represents the total number of packets input to the system.

- *Offered Load.* Defined as the actual traffic presented to the network for transmission.

$$G = \frac{total\_packets\_input\_to\_network}{time} \tag{5}$$

- *Throughput.* Defined as the measure of successful traffic being transferred through the network.

$$S = \frac{total\_packets\_transferred\_through\_network}{time} \tag{6}$$

- *Network Saturation.* Defined as the value of offered load at which the access delay experienced in the network tends to infinity. This is a useful measure for evaluating the performance of media access control protocols used to share bandwidth in the upstream digital CATV channel.

SYSTEM MODELS

In constructing a digital CATV network, the system designer will utilize the identified QOS metrics to assess a given implementation. But in order to make this assessment prior to field implementation, a system modeling framework must be provided.

This framework will provide a characterization of the service, data sources, data sinks, and intervening data transmission and processing devices. The granularity of the models will depend upon the type of analysis being formed. If per-packet statistics are required, the models will contain mechanisms for evaluating individual packet transmission times whereas if overall utilization is of interest only aggregate packet counting mechanisms may be deployed.

Systems and their resulting service characteristics may be evaluated using closed-form analysis based on queuing representations or using simulation to more precisely model the system's components and behavior. For existing systems or design validation, measurement methods may be used to capture the behavior of an actual implementation (7).

Measurement

Obviously, a physical system and network must be accessible to allow measurement to occur. Utilizing empirical data has a distinct advantage in that no detail of network operation is excluded (7). Of course, there are constraints in that measurement points need to be introduced carefully and the amount of data collected needs to be manageable and usable. The measurement process must be as unobtrusive as possible to minimize the alteration of the system performance and to minimize any impact on customer service.

Queueing Analysis

Queueing analysis attempts to model the system as a series of interconnected components behaving according to the well-developed theories of known queueing disciplines. Queueing systems are described by the arrival pattern of customers, the service pattern of customers, the number of service channels, and the system capacity. The arrival pattern of customers is the input to the system. If this pattern is stochastic, a probability distribution function is associated with the pattern. Similarly, the service pattern of customers identifies the time associated with serving a given customer and it also may be stochastic (7).

The number of service channels and waiting room are elements of the system capacity and identify the system's ability to provide service.

Queueing systems are described by the Kendall notation, A/B/X/Y/Z, where

A is the interarrival time distribution
B is the service time distribution
X is the number of service channels
Y is the system capacity
Z is the service discipline

Additionally, a number of parameters are associated with describing the queueing system. While not standardized, the notation listed in Table 1 is commonly used (7).

State-Based Modeling

In developing closed-form queueing solutions or models required for simulation, a tool known as state-based modeling

Table 1. Notation Commonly Used in Queueing Analysis

Parameter	Definition
λ	Mean number of arrivals per unit time
μ	Mean service time for each customer
ρ	Utilization
q	Mean number of customers in the system
t_q	Mean time a customer is in the system
w	Mean number of customers waiting
t_w	Mean time a customer waits for service

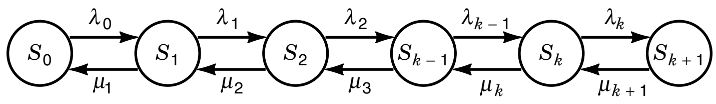

Figure 5. Birth–death Markov process state transition diagram.

is used. A simple example of such a model is the ON/OFF model. This model represents behavior in terms of two states. The ON state represents the sojourn time in which the modeled system is actively transmitting data, while the OFF state represents the sojourn time in which the system is idle. These times may be described stochastically through association with particular probability density functions. Similarly, while in the ON state, an additional distribution may be utilized to characterize the length of each data transmission.

For systems with arbitrary distributions, the process is known as a generally modulated deterministic process (GDMP). If the sojourn times are exponentially distributed, the system is characterized as a Markov modulated deterministic process (MMDP). In such a case, a geometric number of packets are produced. Further generalizations occur by allowing the arrival rate to vary in each state. One example is the Markov modulated poisson process (MMPP) in which the arrival process is Poisson in each state.

A special case is the birth–death Markov process. This process is defined as one in which transitions are permitted only among neighboring states. That is, all past history is captured by the current state, and transitions to the next state only depend on the current state. The Markov process is important because it provides a framework for analyzing a large number of practical problems while remaining mathematically tractable.

As an illustration, let's examine the development of the state transition equations for a birth–death process (Fig. 5). In this case, one can intuitively derive them from an examination of the flow rates in and out of each state. The flow rates into and out of state S_k are specified in Eqs. (7) and (8). To derive the state probabilities, a differential equation is formed representing the difference between the two flow rates. This is shown in Eq. (9). By solving this equation, the state probabilities P_k may be determined (8).

$$S_k = \lambda_{k-1}P_{k-1} + \mu_{k+1}P_{k+1} \tag{7}$$

where $\lambda_{k-1}P_{k-1}$ is the probability flow rate entering state S_k from state S_{k-1} and $\mu_{k+1}P_{k-1}$ is the probability flow rate entering state S_k from state S_{k+1}.

$$S_k = (\lambda_k + \mu_k)P_k \tag{8}$$

where $\lambda_k P_k$ is the probability flow rate departing state S_k for state S_{k+1} and $\mu_k P_k$ is the probability flow rate departing state S_k for state S_{k-1}.

$$\frac{dP_k(t)}{dt} = \lambda_{k-1}P_{k-1}(t) + \mu_{k+1}P_{k+1}(t) - (\lambda_k + \mu_k)P_k(t) \tag{9}$$

As discussed previously, Eq. (9) is a differential equation for the effective rate of probability flow into state S_k. This type of flow balancing can be utilized with other state models as well. By taking the difference between flow rate equations, a differential equation can be derived whose solution specifies the state probabilities P_k (8).

Long-Range Dependent Traffic Models

The previous traffic models are characterized as stationary because they possess an exponentially-decaying correlation structure. Recent research has shown that many networks exhibit aggregate behavior possessing autocorrelation structures with decay rates slower than exponential. This slow decay can be captured mathematically with the notion of long-range dependence and self-similarity. Such characteristics have a more pronounced impact on the network design process which now must accommodate highly variable and "bursty" traffic sources.

A process X_t is said to have long-range dependence if its autocorrelation, ρ_k is not summable (9). This is represented as $\sum_k \rho_k \to \infty$. The power spectral density function is defined as $\sum_k \rho_k e^{-jwk}$. Therefore this density function will be singular near zero. It is also important to note that long-range dependence is based on an asymptotic definition.

A process X_t is said to be exactly self-similar if $\rho_k^{(m)} = \rho_k$ for all m and k; that is, the correlation structure is preserved across different time scales. X_t is said to be asymptotically self-similar if $\rho_k^{(m)} = \rho_k$ for m and k large.

System Identification

The above tools assume the appropriate queueing models and probability distribution functions have been identified. In the case of legacy applications, it is likely that such models have already been derived and exist in the literature. But in the case of new applications or system components, the system designer must be able to follow a process to derive the appropriate models.

System identification refers to the process utilized to derive mathematical models that accurately characterize a particular system or subsystem. This process can be partitioned into data collection, data analysis, and model synthesis phases.

Typical model development techniques are based on the application of standard curve-fitting and statistical analysis tools. These tools include regression testing and analysis, quantile–quantile plots, and hypothesis testing. Once a model has been proposed, it must be validated. This can be done based on goodness-of-fit, with metrics defined to measure the "closeness" to the actual system behavior. Additionally, the number of parameters required in the model and analytical tractability can be evaluated to infer the model's "ease-of-use."

A typical objective is to develop a model which minimizes the sum of squared error, defined as $\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (x_i - \hat{x}_i)^2$, where x is the actual data value and $\hat{x}$ is the modeled or predicted data value (10,11).

Quantile–quantile plots are used to assess the distribution of a set of observed data values. The q_ith quantile is defined as $q_i = F(x_i)$, where $F(x_i)$ is the cumulative distribution function evaluated at the point x_i. This technique plots the observed quantiles versus the assumed theoretical quantile. If the observations do come from the assumed theoretical distribution, the quantile–quantile plot will be linear (10).

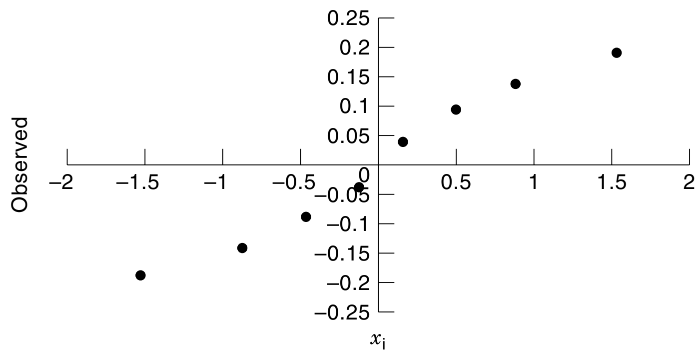

Figure 6. Quantile–quantile plot.

As an example, Fig. 6 illustrates a quantile–quantile plot for a data set with an assumed normal distribution. By examining the plot, one can see the approximate linear characteristic, suggesting that the assumed normal distribution is a reasonable model for the data set (10).

SERVICE CHARACTERIZATION

Digital CATV enables a broad range of services to the end subscriber. These services represent a combination of legacy and emerging applications that leverage the flexible transport capabilities and high bandwidth potential of broadband networks.

The most traditional use of CATV has been to provide broadcast quality video distribution. Digital CATV systems readily support this service while offering the capability to integrate additional services onto the same facility. This provides the end-user an environment in which a single coaxial wire is used to provide a range of services to the home.

To assess the implementation requirements of each service, service definitions will now be developed. These will be followed by a more formal treatment of the mathematical models appropriate for each respective service.

Service Description

It is instructive to consider the most likely service categories and examine their respective data flows and associated quality-of-service requirements. In developing this framework, a service taxonomy will also be constructed in which services are classified according to their QOS requirements for both the downstream and upstream network channels.

Quality-of-Service Framework for Digital CATV

Because of the range of applications envisioned for digital CATV, the ability to provide varying levels of class of service and associated QOS is imperative. This may be accomplished using a variety of mechanisms in both the downstream and upstream environments.

In the downstream, all traffic is inserted on the system via the headend router or gateway. This allows channel and buffering resources to be segmented per class of service under the full control of the headend gateway and its associated management system. Resource management may be done statically via COS/QOS association with known packet addresses or MPEG-2 PIDs or it may be handled dynamically at the IP layer using a mechanism such as RSVP.

In the upstream direction, a future directive of MCNS (version 1.1) is to provide QOS support. This support can also occur statically via the use of a subscription profile to specify data handling requirements per subscriber. Or it may occur dynamically via the use of MAC layer signaling to specify bandwidth requirements on a per session basis.

Digital Video Architecture. The use of digital video and specifically MPEG-2 transport allows for a variety of enhanced video services to be offered. The delivery of such services can be accommodated using either constant bit-rate (CBR) or variable bit-rate (VBR) transmission and encoding.

Early trials have been conducted using CBR encoded and transmitted MPEG-2 for both broadcast and on-demand applications. In these applications, the use of MPEG-2 time stamps allows the receiver to synchronously lock to the source's master clock while network level adaptive buffering is used to synchronize to the transport stream rate (22).

The disadvantage of the CBR approach is the reservation of resources at a fixed peak rate and the variation of quality that is a result of constant rate encoding. The use of VBR transport and encoding can overcome these limitations but alternative models using VBR encoding in conjunction with CBR transport are also possible (22). Such mechanisms generally trade-off network resources for set-top playout buffering resources to allow a constant video quality to be maintained. Further, if the start of playback time can be extended, the use of store-and-forward techniques further reduces the level of network resources required. In such a case, all or a significant portion of the video may be preloaded into the receiver for localized playback streaming.

Digital Video Service Description. Broadcast video provides the end-user with the ability to selectively tune from a prespecified number of scheduled programming material. Traditionally, analog CATV distribution provides a single video program per standard 6 MHz television channel. Typical system implementations may provide up to 50 to 60 channels of programming. Using digital QAM modulation in conjunction with MPEG-2 video compression technology, the number of channels may be significantly increased. And with such increased capacity comes the benefit of an enhanced mode of broadcast video known as near video-on-demand (NVOD) or enhanced pay-per-view (EPPV) (2).

EPPV allows the operator to offer popular content utilizing multiple staggered delivery times. A typical service scenario provides the end-user the ability to select a program from a top 10 list of movies scheduled to start every 15 min during the evening hours. Additional application software provided in the STB allows the user to interactively select the next nearest start time and also provides virtual VCR capabilities by allowing limited pause, rewind and fast-forward capabilities. This requires no additional upstream transmission but rather relies on local synchronization to the next nearest copy of the selected movie's MPEG stream. It is thus characterized as a unidirectional, downstream-only application. For the case of constant bit-rate encoding and transmission (CBRT), data rates are typically in the range of 3 Mbps to 8 Mbps for NTSC quality video. Variable bit-rate (VBR) encoding can be used to lower the average bandwidth requirements at the expense of greater system complexity. VBR also requires the use of more complex stochastic models to represent its traffic characteristics.

For many years the vision of interactive television has been to provide the consumer with the ability to interactively select video programming on demand. Known as video-on-demand (VOD), it has been the focus of a number of major engineering and deployment trials over the last 5 years.

Its basic premise is to provide the user with a STB application that allows easy access to archives of remotely stored video content. The user utilizes the browsing capabilities of the application to select a particular program for viewing. This results in the near-instantaneous scheduling of the program for on-demand playback through the system.

This service is characterized by its generation of traffic in both the upstream and downstream channels. Upstream traffic is due to the interactive video browsing process, while downstream traffic is generated once playback of the program has begun (2).

Internet Access. The emergence of the World Wide Web (WWW) has driven the development of alternative high-speed access architectures designed to overcome the performance limitations of traditional dial-up, analog modem-based services. The CATV community has fostered the development of such an alternative in the form of cable modem technology.

Cable modems are designed to make use of the inherent broadcast nature of the digital CATV transport medium. This allows the network operator to create metropolitan or community area data networks utilizing data transmission equipment deployed in the local headend coupled with cable modem termination equipment resident in the subscriber home.

Engineering a data delivery system requires a detailed characterization of the applications expected to be resident. With the popularity of the Web growing at an exponential rate, we will only consider this characterization in terms of Web-browsing applications. This is reasonable because the Web browser has become the front-end of choice to a multitude of Internet applications. Furthermore, streaming applications accessed through the browser may be characterized in the steady state by their standalone service representations.

WWW applications are identified by their bursty nature and generation of traffic in both the upstream and downstream network channels.

Telephony. One form of digital telephony utilizes voice compression technology to allow low-bit-rate packet voice transmission through a digital CATV system. This is opposed to time-division-based systems which utilize circuit-switched 64/32 kbit/s channels resident with the frequency division multiplex.

Our treatment will focus on the packet-based implementation. This allows telephony applications to be integrated with the same packet division multiplex used to carry other digital services. It also forces the implementation to consider its QOS requirement of low delay.

Telephony applications are characterized by their by bidirectional nature, relatively low bit rate, burstiness, and limited duration.

Gaming. Interactive gaming utilizes the networking capability of the digital CATV system to allow interconnections among multiple remote game players. These applications utilize the CATV network analogously to gaming applications designed to operate in local area network environments.

Gaming applications are characterized by their burstiness and asymmetry in traffic flow. Upstream flows are typically characterized by short data packets representing player game movement, while downstream flows are typically larger in size and represent global game updates sent to all participants (6).

Service Models

Video Services. Broadcast video, enhanced pay-per-view (EPPV), and video-on-demand (VOD) require high-quality video. The standard video compression algorithm used is MPEG-2 which has two modes of operation. Constant bit rate (CBR) and variable bit rate (VBR). CBR mode produces a constant bit stream with a variable video quality. VBR mode produces constant quality with a variable output rate.

The bursty nature and high correlation of VBR video represent a challenging problem. The output bit rate can vary on a small time scale due to the variation within the scene and can vary over a large time scale due to the variation of different scene content. Such traffic significantly increases queue length statistics and makes providing guarantees on delay and delay variation nontrivial.

Recent studies (12,13) suggested that MPEG-2 VBR video can be modeled as a fractionally differenced autoregressive integrated moving average process (F-ARIMA). F-ARIMA was first introduced in Ref. 14. The study in Ref. 12 suggested that the distribution can be best fit by a Gamma/Pareto distribution.

In the case of EPPV and VOD, content is typically stored on a server. The storage capacity required is very large. Storing 200 movies would require a full terabyte of storage. This creates another challenging problem for video server designers, namely, overcoming the bottleneck when reading the data from the storage device and sending it over a network at a speed fast enough to match the playout speed (15).

EPPV requires less bandwidth for the same number of users than VOD. Users demanding the same video content within a time interval can be grouped together and consequently utilize only one video stream.

Internet Access. Based on research conducted in the last 5 years regarding the traffic characteristics of data networks, a model has been developed to characterize the source behavior of Internet users. Specifically, with the advent of the WWW, the model has been developed based on the observed behavior of WWW browsing applications.

The model is a self-similar stochastic model (Fig. 7). In this model the interarrival times of documents requests generated by each source is based on a two-state ON/OFF source model

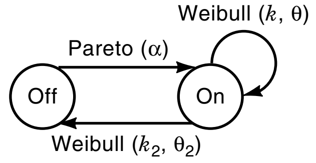

Figure 7. WWW client model (19).

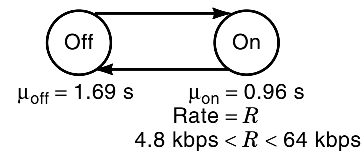

Figure 8. ON/OFF telephony source model. Here μ_{off} is the average duration of the OFF state, μ_{on} is the average duration of the ON state, and R is the rate of voice transmission.

(16,17). The length of each document requested is given by a Pareto distribution (14,18).

In traditional ON/OFF source models, the distributions of the sojourn time spent in both ON and OFF states is assumed to have a finite variance. As a result, the aggregations of large number of such sources will not have a significant correlation, except possibly in the short range. The WWW applications model allows the ON and OFF periods to have infinite variance; the super position of such sources in which the ON/OFF periods have infinite variance produces aggregate traffic that exhibits long-range dependence, or self-similarity (16,19).

In Ref. 20, the authors performed a sensitivity analysis among the parameters α, k, θ, k_2, and θ_2. These parameters control the distribution times in the ON and OFF states as well as the distribution of the interarrival time during the ON states. Nominal values were identified as $\alpha = 0.5$, $k = 0.5$, $\theta = 1.5$, $k_2 = 0.88$, and $\theta_2 = e^{4.5}$.

Telephony. Telephony applications may be modeled as an ON/OFF source, with the OFF state representing periods in which the speaker is silent and the ON state representing periods in which the speaker is active (Fig. 8). From measurement, typical values for the mean ON and OFF periods are 0.96 s and 1.69 s. Depending upon the compression algorithm utilized, traffic rates during the ON period may range from 4.8 kbps to 64 kbps (uncompressed) (7).

Telephony applications require bidirectional bandwidth with bounded delay, as well as delay variation. As the compression ratio is increased, the bounds on packet loss also become more severe.

Gaming. In the upstream channel, gaming applications are characterized by the random arrival of short, minimum-length packets. In the downstream channel, variable-length responses are returned to all game participants.

In the upstream, the interarrival time between individual inputs to the game are modeled by a Poisson process with average time λ (Fig. 9). The length of data generated by an input is modeled as a fixed-length packet of 64 bytes. QOS requirements include minimizing response delay to less than several hundred milliseconds. Packet loss also must be mini-

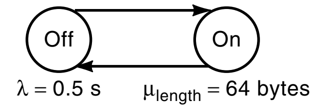

Figure 9. Upstream gaming model. Here λ is the average duration in the OFF state and μ_{length} is the average length of a packet (in this case it represents the length of all packets as they are assumed to be of fixed length) (6).

mized to avoid game play interruption in both the forward and reverse channels.

SYSTEM DESIGN PROCESS

In deploying a set of digital CATV services, a formal process must be followed by which the service requirements are identified and an appropriate system implementation is achieved. With the advent of multimedia applications with varying levels of QOS requirements, the use of ad hoc or back-of-the envelope-style analysis is unlikely to yield optimal results. From the perspective of the network service provider, optimality implies an implementation meeting the customer's quality expectations while minimizing the operator's investment in excess network and system capacity. In practice, this optimality condition is in fact very difficult to achieve.

Using a formal design process, however, a methodology can be constructed such that an operator can design a robust set of services and implement them in a manner that is much closer to the optimum. This methodology must provide the capability of capturing and characterizing the service level requirements as well as provide a set of mechanisms for evaluating the various design alternatives that may exist to implement such services. And lastly it must provide a framework for validation of the system design in light of its expected performance (10).

Design Sequence

A good top-down design process allows the engineer to follow a rigorous development course which begins with the capture of the system requirements and culminates in a validated system design. This process is summarized by the following sequence:

1. Identify service objectives.
2. Identify performance requirements and associated metrics.
3. Develop service models.
4. Develop system component models.
5. Perform analytical characterization (if possible).
6. Develop simulation.
7. Record observed performance.
8. Compare results to objectives.
9. Adjust system parameters and repeat step 6 until objectives are realized.
10. Perform sensitivity analysis.
11. Validate models and recorded data versus measurement of actual system.
12. Update models and system design as needed.

Steps 1 through 4 can be considered the data collection and model development phase. In this phase, the developer must collect information about the system and its intended applications. Information and data must then be collected to allow synthesis of a set of models for the system and its services.

Steps 5 and 6 highlight the analysis process used to capture data regarding the system's modeled (as opposed to actual) performance. This leads to step 9, which specifies an iteration through the simulation phase by adjusting system parameters until the desired design objectives are met.

Steps 10, 11, and 12 represent the last phase of modeling where the developer assesses the model's sensitivity as well validates the models predicted data values versus data collected from an actual working system. At the conclusion of this phase, the system designer should have a high level of confidence that the collection of models are representative of the system's behavior and are therefore a valid tool for analyzing and developing subsequent system design revisions.

Analytical Tools

The tools required to complete a formal systems design offer the engineer the ability to form varying levels of analysis based on the particular implementation objectives. As the network operator moves to full-service deployment, the need to optimize system utilization while maintaining QOS becomes critical.

Queueing Analysis. As discussed previously, simplified "back-of-the-envelope" calculations will not typically yield optimal design results. However, in beginning a formal systems design process, the use of approximations are instructive in highlighting broad system performance issues and providing a baseline characterization of the system's performance. As discussed in the section entitled "System Models," one such mechanism is the use of queueing theory to develop models of sub-system behavior. By mapping the system components and application characteristics into mathematically tractable queueing models, the system may be initially characterized.

This can then be followed with a more detailed analysis and design based on the use of state models and simulation. Simulation allows more behavior to be captured and in fact may be the only mechanism available due to the limitations of deriving closed-form analytical solutions.

Simulation. Using the models developed to characterize the system components and its services (traffic sources/sinks), a simulation can be implemented using either a general-purpose programming language or using one of a number of commercially available simulation tools. These tools provide a framework with a number of common simulation functions already implemented.

Simulation becomes paramount when closed-form queueing solutions are not feasible. For many systems with a complex interconnection of subsystems and a multitude of states, simulation often must be used to obtain more detailed characterization.

Monte Carlo, trace-driven, and discrete-event represent the three main types of simulation commonly used (10). Monte Carlo simulation is used to model probabilistic events that do not depend on time. It is a static simulation technique and does not use a time axis.

Trace-driven simulation uses a trace of time-ordered events captured from a working physical system. Traces are useful in driving system simulations designed to optimize performance or tune different algorithms. They also offer the advantage of not having to derive a representative source model.

Discrete-event simulation utilizes a discrete-state model to represent system dynamics. All discrete-event simulations share a number of common traits. An event scheduler is used to maintain a list of events waiting to happen. A global simulation clock is used to track time. This clock may be advanced either by unit time increments or by the time of the next earliest event. The former approach is a time-driven clock, while the latter is an event-driven clock. Lastly, state variables and event processing routines are used to manipulate the state of the system being modeled (21).

SUMMARY

This article has presented a treatment of digital CATV and its services. The basic components of a representative digital CATV system were identified, followed by a presentation of the relevant mathematical framework needed to model and characterize its performance and services. This included a specific treatment of the current stochastic models used to represent the relevant source models. The article concluded with a discussion of model validation and its role in an iterative model development process.

BIBLIOGRAPHY

1. W. Grant, *Cable Television,* GWG Associates, 1994, pp. 2–15.
2. Digital Audio Visual Council, *Version 1.0 Specification,* 1996, pp. 10–30.
3. *IBM Application Note,* IBM Microelectronics 1998.
4. *MCNS Data-over-Cable Specification,* Version 1.0, 1997, pp. 10–25.
5. P. Ferguson and G. Huston, *Quality of Service: Delivering QOS on the Internet and in Corporate Networks,* New York: Wiley, 1998, pp. 3–4.
6. Limb et al., *Performance Evaluation Process for MAC Protocols,* IEEE 802.14, Document No. 96-083R2, 1996.
7. J. Pitts and J. Schormans, *Introduction to ATM Design and Performance,* New York: Wiley, 1996, pp. 22–24.
8. L. Kleinrock, *Queueing Systems, Vol. I: Theory,* New York: Wiley, 1975, pp. 57–59.
9. N. Adas, *Broadband Traffic Models,* Georgia Tech Document Number GIT-CC-96-01, p. 13.
10. R. Jain, *The Art of Computer Systems Performance Analysis,* New York: Wiley, 1992, pp. 192–199.
11. M. Hayes, *Statistical Digital Signal Processing and Modeling,* New York: Wiley, 1996, pp. 129–131.
12. M. Garrett and W. Willinger, Analysis, modeling and generation of self-similar VBR video traffic, *SIGCOMM '94,* 1994, pp. 269–280.
13. C. Huang et al., Self-similar modeling of variable bit-rate compressed video: A unified approach, *SIGCOMM '95,* 1995.
14. J. Hosking, Fractional differencing, *Biometrica,* **68**: 165–176, 1981.
15. K. Almeroth, *Support for efficient, scalable delivery of interactive multimedia services,* PhD dissertation, Georgia Institute of Technology, 1997.
16. S. Deng, Empirical model of WWW document arrivals at access link, *ICC '96,* 1996.
17. M. Crovella and A. Bestravros, *Explaining World Wide Web self-similarity,* Tech. Rep. TR-95-015, Comput. Sci. Dept., Boston Univ., 1995.
18. S. Jamin et al., A measurement-based admission control algorithm for integrated service Packet networks, *SIGCOMM '96.*
19. W. Willinger et al., Self-similarity through high-variability: Statistical analysis of Ethernet LAN traffic at the source level, *ACM SIGCOMM '95,* 1995.
20. S. Hrastar and A. Adas, Network design of cable modem systems for WWW applications, *IEEE Community Networking Workshop 1997,* pp. 2–3.
21. P. Gburzynski, *Protocol Design for Local and Metropolitan Area Networks,* Upper Saddle River, NJ: Prentice-Hall, 1996, pp. 18–20.
22. J. McManus and K. Ross, *Video on Demand over ATM: Constant-rate Transmission and Transport,* Dept. Syst. Eng., Univ. Pennsylvania, Nov. 1995.

SCOTT HRASTAR
A. ADAS

DIGITAL CATV SERVICES. See CABLE TELEVISION.

DIGITAL CIRCUITS. See INTEGRATED CIRCUITS.

DIGITAL CONTRAST ENHANCEMENT. See IMAGE PROCESSING CONTRAST ENHANCEMENT.