PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Real-time performance analysis of wireless multimedia networks based on partially observed multivariate point processes

William S. Hortos

SPIE.

Real-time performance analysis of wireless multimedia networks based on partially observed, multivariate point processes

William S. Hortos

Florida Institute of Technology, Orlando Graduate Center, 3165 McCrory Place, Suite 161, Orlando, FL 32803

ABSTRACT

Third-generation (3G) wireless networks will support integrated multimedia services based on a cellular extension of a packet-switched architecture using variants of the Internet protocol (IP). Services can be categorized as real-time and delay-sensitive, or non-real-time and delay-insensitive. Each call, arriving to or active within the network, carries demand for one or more services in parallel; each service type with a guaranteed quality of service (QoS). Admission of new calls to the wireless IP network (WIN) from the gateway of a wired network or from a mobile subscriber (MS) is allowed by call admission control (CAC) procedures. Roaming of the MSs among the nodes of the WIN is controlled by handoff procedures between base stations (BSs), or BS controllers (BSCs), and the MSs. Metrics such as the probabilities of call blocking and dropping, handoff transition time, processing latency of a call, throughput, and capacity are used to evaluate the performance of network control procedures. The metrics are directly related to the network resources required to provide the QoS for the integrated services. User mobility, combined with the "bursty" nature of service demands, leads to transient random behavior of the multiple packet flows for integrated services within the WIN. This paper proposes general stochastic models, based on the theory of multivariate point processes (MVPPs) and their representation as semimartingales, to describe the finite-horizon, transient behavior of the packet flows created by integrated multimedia traffic. The point process corresponding to the packet flow of each service type is decomposed into a right-continuous, pure jump process and a predictable, integrated random rate process. The approach leads to predictive models of information flows that can incorporate the measurement-based estimates of probability distributions for voice, video, data, and Internet traffic, as well as the protocol mechanisms for access, routing and switching in wired networks and for blocking, handoff and radio connectivity in the WIN. The effects of CAC and handoff control are also shown in the terms of the integrated random rates for the information flows. Performance metrics and QoS parameters of multimedia services are represented in the MVPP models. Use of the models to develop stochastic filters of the network state, based on partial or incomplete observations of packet flow dynamics, is presented.

Keywords: Wireless multimedia networks; real-time adaptive estimation; call admission control; handoff control; multivariate random point processes; resource allocation; quality of service (QoS); nonstationary probability distributions; self-similar processes; martingale representations

1. INTRODUCTION

The use of adaptive techniques has been proposed to enhance the performance of mobility and portability in current and next-generation wireless multimedia networks. Quality of service (QoS) provisioning in all classes of network services, including voice, data, video, and facsimile, to mobile stations (MSs) leads to distributed control of wireless network resources. The movement to build a third-generation (3G) telecommunications infrastructure to deliver these services using the Internet protocol (IP), for both wired and wireless networks, to parallel and possibly supplant the existing circuit-switched networks has been initiated by the major manufacturers of the infrastructure equipment. An increasing level of multimedia traffic will be packetized and transported via wireless access protocols based on extensions of the IP, known as mobile or wireless IP. Resources in the wireless IP network (WIN) need to be allocated with an acceptably high probability, when a call arrives or a handoff occurs in the wireless domain. In addition, control functions performed through signaling procedures between the MSs and either the base station (BS) controllers (BSCs) or the message control centers (MSCs) of the wireless infrastructure need to be robust to accommodate the random occurrence of service-dependent call events. Markov and Bayesian methods have typically been used to formulate and evaluate adaptive control algorithms. These methods have been considered due to their ability to adapt to changing network conditions in real time, and to enable decentralized procedures to the management of network resources. The combination of these features allows procedures to ensure predefined levels of QoS to different traffic classes in complex hierarchical cellular networks (HCNs) that support multimedia services.

In *Digital Wireless Communication II*, Raghuveer M. Rao, Soheil A. Dianat, Michael D. Zoltowski, Editors, Proceedings of SPIE Vol. 4045 (2000) • 0277-786X/00/$15.00

This paper extends the analytical methods of Markov and Bayesian approaches to wireless multimedia networks to achieve a comprehensive representation of the controllable and observable, real-time call processing events, while overcoming the limitations of those approaches. The proposed analytical models of network behavior are based on the theory of semimartingale representations of multivariate point processes (MVPPs) with randomly modulated rates. The MVPPs are used to represent the transient traffic flows of information packets between nodes in a wireless packet-switched network. In the finite-population network model the set of nodes consists of a maximum M_{max} MSs and the fixed infrastructure of N BSs is controlled by a hierarchy of BSCs and MSCs. Packet entry, routing and switching in the network is regulated by call admission control (CAC) and call handoff (CHO) procedures among the distinct cells and sectors established by the neighboring BSs to support end-to-end connectivity of the MS calls. Inter-event times, and the corresponding counting processes, can follow general non-stationary probability distribution functions (PDFs) for call arrivals, cell occupancy, call holding, integrated service loading, and service completions. Inter-event times are random stopping times, progressively measurable with respect to the σ –algebra generated by the history of network events. Random switching and routing variables, which characterize call admission and handoff, depend on non-stationary path loss distributions due to fading, reflections and user mobility, as well as the observed network history. The extent of network history observable to a node controller depends on the higher layer protocol functions implemented in the node. Limited or partial observations lead to the concept of stochastic filtration of the network state dynamics to effect "best-effort" control at the nodes.

The MVPP models generalize Poisson point processes, exponentially distributed call processing, and other renewal processes commonly assumed to create queueing structures for the evaluation of the asymptotic performance of control schemes for packet networks at or near equilibrium. The actual dynamics of WIN operations, however, are random and transient, and rarely support an assumption of ergodicity or the existence of an equilibrium. Therefore, the models of real-time WIN performance considered here are limited to a finite time interval, $[0,T], T < \infty$. Performance metrics, such as, delay, call completion, call blocking, call dropping, handoff failure, local and global throughput, and system capacity, are indexed by the service types and also have representations in terms of the MVPPs. Limits on network resources, e.g., physical channels, bandwidth, transmit power, receiver sensitivity, and buffer size, that are necessary to achieve QoS requirements, are seen to modulate and constrain the random rates of the MVPPs corresponding to distinct network events.

Over two decades ago, Brémaud [1], Walrand and Varaiya [2] and others established the theoretical foundation for general semimartingale representations of both discrete-space and continuous-space random point processes and their decomposition into both predictable and unobservable components with respect to a probability space $(\Omega; F; \wp)$, where Ω is the set of outcomes for network events, F_t is the σ-subalgebra generated by observations of network events to time t, from the family of σ-subalgebras $F_t = \sigma\{\omega \in \Omega : X_v(\omega), 0 \le v \le t\}, F_t \subseteq F_T = F$, and $\wp$ is the reference probability measure on network events with $\wp(F)=1$. Brémaud has shown that virtually any practical birth-and-death process, such as a queue, can be represented by a unique semimartingale equation in continuous time or between random stopping times.[1] In cases of incomplete or local observations of network behavior, the effects of a real-time adaptive algorithm are examined through stochastic filtration or so-called innovation equations for a generalized likelihood ratio function, based on the MVPPs' integrated random conditional rates constructed from general PDFs for corresponding network events. The latter is based on the author's research [3] and that of Boel, Varaiya and Wong [4] in the control of point processes with incomplete information.

The MVPP approach overcomes limitations of conventional Markov and Bayesian models that assume successive observations, as the MSs move from BS to BS, are independent, thereby permitting the probability of a sequence of observations to be written as the product of probabilities of individual observations. This assumption is clearly precluded by MS mobility and competition for limited network resources over the duration of a call. The Markov limitation occurs when dependencies extend through several successive network events, as in the reservation and allocation of resources to the nodes to support multimedia calls. In Markov models, resources that have been allocated at the designated BS will remain in effect for the specified length of the service interval. When the timer expires, the resources will automatically be returned to the resource pool. For real-time network operations, this limitation results in a sub-optimal allocation of a portion or all of the available resources, so that the QoS of all users will be affected, if some remedial action is not taken by the new BS.

The proposed MVPP models of real-time behavior are mathematically tractable and extendible, even able to encompass self-similar processes of long-range dependence (LRD) that characterize Internet traffic. The objective of the MVPP models is to provide an analytical basis for accurate evaluations of the comparative performance of real-time adaptive algorithms for resource allocation, admission control, handoff, and congestion control in practical multimedia WINs.

2. FEATURES OF WIRELESS MULTIMEDIA NETWORKS

A brief description is first given of WIN operations that handle the flow of multimedia packets during the input, support, and completion of a call. A call can originate within the network from any active MS i as one of $M_{\max}$ mobile nodes. Similarly, calls can originate from the fixed infrastructure at the BSCs or MSCs through the BS in best position to the last known MS location, say BS j, which can be also considered an originating or terminating node. From whatever source calls arrive, the uninterrupted arrival stream of calls to the WIN are most commonly assumed to occur according to a sequence of random F_t-stopping times, $\tau_0^a, \tau_1^a, \ldots, \tau_n^a, \ldots$, such that the corresponding sequence of inter-arrival times, $\tau_1^a - \tau_0^a, \tau_2^a - \tau_1^a, \ldots, \tau_{n+1}^a - \tau_n^a, \ldots$, are independent and identically distributed (i.i.d.) random variables with the common arbitrary right-continuous PDF, $\mathrm{F}^a(t), \tau_n^a < t \le \tau_{n+1}^a$, for every n. In other words, the inter-arrival sequence forms a renewal process. In practical operation, the PDFs can change between stopping times due to transient behavior, i.e., $\mathrm{F}_n^a(t), \tau_n^a < t \le \tau_{n+1}^a$, for every n, violating the renewal assumption. In synchronous network operation or computer simulations, arrival times as well as service times can be slotted and deterministic, i.e., $\tau_n = T_n = nT_0$, for T_0 a known slot time, frame time, or simulation time increment. To represent integrated services in 3G networks, calls are assumed to arrive with a set of simultaneous service loads consisting of a maximum S service types. Each type requires a distinct level of QoS, expressed in terms of network operating parameters. Any of the S integrated services may be active in the call, requiring a different set of network resources to maintain their distinct QoS requirements during call processing. Therefore, the sequence of service completion times $\tau_{i,m}^s$, corresponding to type $s, s = 1,\ldots,S$, at node i and the corresponding sequence of inter-service completion times, $\tau_{i,1}^s - \tau_{i,0,}^s, \tau_{i,2}^s - \tau_{i,1}^s, \ldots, \tau_{i,m+1}^s - \tau_{i,m}^s, \ldots$, may not be i.i.d. random variables and may not share a common PDF with the inter-service time sequences associated with other service types. Superposition of the sequences of the inter-service times corresponding to any two or more of the types will not form a renewal sequence. Lacking the exponential assumption on the inter-service times, superposition of the streams of service completions will not, in general, form a renewal process.[5]

Since World Wide Web (WWW) traffic accounts for more than 25% of Internet traffic and is currently growing more rapidly than any other type, understanding the nature of WWW traffic is increasingly important. The challenge to the modeling assumptions is fitting the self-similar behavior [6] of WWW traffic to MVPP models. Therefore, the only assumptions required by this analytical development are those that support the semimartingale decomposition of the MVPPs describing integrated multimedia traffic, as the sum of (F_t)-predictable, integrated, *non-explosive* rate processes and pure jump martingales, with respect to the probability space $(\Omega; F_t; \wp)$.

2.1. Integrated Service Classes

The services in 3G networks can be classified broadly as either real-time or non-real-time. Real-time services can have distinct constant bit rates (CBRs), such as 8-kilobits per second (kbps) and 13-kbps voice codecs, or variable bit rates (VBRs) such as interactive video. Excessive delay or delay variation noticeably degrades real-time services. In real-time packet modes, a large amount of digitized information is transmitted over a relatively long duration. Non-real-time services, such as file transfers, Internet accesses, e-mail and other delay insensitive services are transmitted by IP networks as high-rate bursts and characterized as "on-off" processes. For packet data services, transmission stops at the end of the data burst, since no information is generated during the unpredictable "off" intervals. Transmission of real-time services is continuously maintained during the call, while packet data services are provided to users with demand for high transmission rates, but short service times. Certain non-real-time packet data services differ in their tolerance of delay variation as opposed to fixed transfer delay. For example, many web pages already include real-time video and audio clips.

While some QoS measures relate to such parameters as receive signal strength (RSS), signal-to-interference ratio (SIR), bit-error ratio (BER), and frame-error ratio (FER), other parameters distinguish service classes by their tolerance to fixed delay and delay variations as shown in Table 1. Three classes of service types are accommodated in the model.[7] Both CBR and VBR services are assumed in each class. Available bit rate (ABR) services are allowed in the third class for "best-effort" QoS in the presence of competing service demands, along with services with undefined bit rate (UBR) requirements.

Class 1 services encompass highly delay-sensitive, real-time connections with very low delay-tolerance, such as voice, interactive video and video conferencing. Real-time multimedia applications, such as videoconferencing, impose these requirements on inter-network gateways, since the traffic they produce must be delivered on a certain temporal sequence or it becomes useless. Class 1 services receive the highest service priority over other classes and require fixed bandwidth or

transmission rates. Terms of service may be negotiated between two or more CBR alternatives based on bandwidth and other radio resource limitations.

Service	Requirements	
	Delay	Delay Variation
File Transfer	Insensitive	Insensitive
Web Browsing	Insensitive	Insensitive
E-mail	Insensitive	Insensitive
Voice over IP	Very low	Very low
Video telecom over IP	Low	Low
Real-time Video	Insensitive	Low

Table 1. Delay-related requirements for typical packet data services

Class 2 services include non-real-time, delay-sensitive, connection-oriented services with limited delay requirements such as remote login, file transfer protocol (FTP), and similar applications associated with the transport control protocol (TCP). This class typically receives lower priority than Class 1. The rate of service can be negotiated as an ABR between a maximum and a minimum acceptable limit, based on QoS latency requirements and resource availability.

Class 3 services are message-oriented and delay-tolerant. Typical services are paging, e-mail, voice mail, facsimile, and data file transfer. They can be packet- or circuit-switched. Class 3 services can be conveyed at the earliest possible time and the rate of transfer can be adjusted continuously based on the available unused bandwidth and other resources, after the QoS requirements of the services from the other two service classes have first been met. For example, File Transfer Protocol (FTP), Simple Mail Transfer Protocol (SMTP), or X Windows, are "best effort" services in which variations in delay often go unnoticed.

The service classes require different service or queueing priorities at the nodes to ensure the QoS delay requirements are met. Since Class 1 services have low tolerance for delay and delay variation, they cannot be stored and forwarded in a long buffer, as can Class 3 services, nor can they be retransmitted with a feedback mechanism, when cumulative errors cannot be corrected by codecs. Unlike Class 2 or Class 3 messages, Class 1 streams cannot be demoted in service priority at the nodes without loss of the voice over IP (VoIP) link.

2.2. Hierarchical Cellular Networks

The radio resource (RR) limits to support a wide range of 3G integrated services depend directly on cell size and MS mobility. These factors, along with transient traffic densities, lead to proposals for a three-tier overlay of macro-, micro- and picocells. Macrocells cover large geographical areas, where MS densities may be low, and can handle fast-moving mobiles. Micro- and picocells serve locations where MS densities are higher. In planned 3G networks, calls requiring rates up to 144 kbps in high- to medium-speed vehicular environments are initially connected to a macrocell BS. Calls requiring user rates between 144 and 384 kbps at low to pedestrian speeds, or in outdoor-to-indoor environments are initially connected to a microcell BS, while calls that require above 384 kbps to 2,048 kbps in indoor and stationary environments will be initially connected to a BS in a picocell. Handoffs and new call admissions depend on three factors: availability of RRs in the active set of cells in the intended layer to meet the QoS needs of the call, the match between the layer and MS mobility, and interference induced in the cell by acceptance of the connection request. It is assumed that cells, not sectors, in all layers of the WIN have the same RRs, but may not be able to accommodate all service connection requests due to power, bandwidth or buffer limitations as well as incompatibilities between required user rates and MS speed.

2.3. Radio Resources and Transport Channels

The 3G radio transmission technology of the International Telecommunications Union (ITU) IMT-2000 requirements have been designed to support mobile multimedia communications. A key feature is the ability to transport multiple parallel services with different QoS requirements on one wireless connection. Another feature is the flexible, dual-mode packet access scheme where packet transmission can occur either on common channels (CCHs) or on dedicated channels (DCHs). CCH packet access is typically used for short infrequent packets. In the dedicated physical data channel (DPDCH) packet

access mode, an initial random access request is used to set up a DPDCH for packet transmission. The DPDCH can be set up for the transmission of a single long packet or for transmission of a sequence of shorter packets.

The 3G physical layer offers information transfer services to the medium access control (MAC) and higher layers of the protocol stack. The physical layer transport services describe how and with what characteristics data are transferred over the radio interface, termed the *transport channel.* Transport channels are classified into two groups: common channels including broadcast CCH (BCCH), paging channel (PCH), forward access channel (FACH), and random access channel (RACH); and dedicated channels (DCHs).

The physical channel bit rates vary in a range from 32 to 2,048 kbps, determined by mobility, service type, and grade of service. Other 3G physical layer features include fast closed-loop power control, used for all DCHs in both uplink and downlink to combat fast fading channels and interference fluctuations. The power control step can be varied adaptively in response to changing radio propagation conditions. An outer control loop is also used to regulate the SIR target, based on required link quality.

The radio link control (RLC) and MAC protocols are responsible for efficiently transferring user content of both real-time and non-real-time services. The transfer of non-real-time data includes the possibility of an optimized low-level automatic repeat request (ARQ) protocol at the RLC layer, offering higher protocol layers reliable data transfer. The MAC layer controls multiplexing of information streams originating from different sources within a subscriber's set of integrated services. The source of the call must explicitly specify its service characteristics and required QoS as part of the connection request. The radio resource manager (RRM) in the radio network controller (RNC), which allocates available resources to the BS, must determine whether this BS or another BS neighboring the MS can meet the needs of a connection request.

2.4. Call Admissions and Handoffs

As an MS moves from one cell to another, RRs in the new cell must continue the QoS for the services still active in the call. A significant part of this mobility support involves allocation of sufficient resources to maintain the QoS of the established connection(s). If sufficient RRs are not allocated or not available, QoS may not be met. This, in turn, may lead to loss of the connection, i.e., handoff failure and subsequent call dropping, if a reduced QoS level cannot be negotiated. Since premature termination of established connections has a more negative impact on perceived QoS than new call blocking, it is common practice to give higher priority to handoff requests than to new call requests and ten times more stringent error requirements on handoff QoS.

Admission control strategies reserve RRs *a priori* in each cell to deal with handoff requests. In single-service networks, where the traffic and QoS of all requests are uniform, reservation of RRs typically occurs in the form of "guard channels." In the multimedia WIN, the complexity of the adaptive scheme increases as the number of service types active in each call, whether new or handoff, require different RR levels to maintain their distinct QoS.

The *active set* or *neighborhood* of an MS is the set of BSs to which an MS is currently connected based upon the RSS or some other signal quality measurements of the pilot or beacon channels in the downlink from the BSs. During a cell search process, the MS searches for the BSs to which it has the lowest path loss.

A *soft handoff* algorithm makes decisions based on some quality measure, e.g., path loss or uplink carrier-to-interference ratio (C/I). The two main parameters in a soft handoff algorithm are the handoff margin (*hm*) and the maximum active set size (*AS*). In the active set the "best" BS is the one with highest value of the signal quality measure, while all other BSs of the set are within the handoff margin, which is defined relative to the best BS value of signal quality. The parameters, *hm* and *AS*, can be used to control the fraction of MSs in soft handoff in the system. In networks with adaptive antenna control and sectored cells, intra-cell soft handoff, called *softer handoff*, is the procedure in which the MS can be connected to more than one sector within the same cell. In *hard handoff*, the connection on the current frequency may be severed as it moves to a new frequency during handoff. WINs support intra-frequency handoff, inter-frequency handoff between HCN layers, and inter-network handoff. For intra-frequency handoff, dedicated circuit-switched channels use soft handoff, DPDCH packet channels can use soft or hard handoff, while CCHs use hard handoff. For inter-frequency and inter-network handoff, a hard handoff procedure is applied, where the RSS measurements on other BS frequencies are performed in slotted-mode, downlink transmission or with a dual receiver.

3. MODELING THE ELEMENTS OF WIRELESS MULTIMEDIA NETWORKS

The following describes specific representations of a wide range of the operational elements of multimedia WINs in terms of the real-time MVPP model construction.

3.1. Traffic Types and Probability Distributions

Calls arriving at the nodes of a 3G network at random times τ_n^a carry one or more of at most S simultaneously active integrated services. These can include voice, video, facsimile, Internet traffic, file transfers, etc. The service load arriving at time τ_n^a is modeled as an embedded, discrete vector-valued, discrete-time process $B_n = (b_{1,n}, b_{2,n}, \ldots, b_{S,n})$, where $b_{s,n}$ is the processing load in packets or information rate corresponding to service type s, $s = 1, 2, \ldots, S$, and can vary from arrival time to arrival time. The condition $b_{s,n} = 0$ represents that service type s is inactive in the call at τ_n^a. Since the load process $(B_n, n \in \mathbf{Z}_+)$ at arrival times τ_n^a and the counting process on new calls, denoted $(N_t^A, t \in [0,T])$, are distinct, with different statistical properties, the combined integrated-service arrival process can be any number of hybrid MVPPs, based on the those properties. For example, if $(B_n, n \in \mathbf{Z}_+)$ is a discrete-time, discrete-space Markov process, and $(N_t^A, t \in [0,T])$ a Poisson process with time-varying rate $\alpha_t, t \in [0,T]$, the combined arrival process is a non-homogeneous, Markov-modulated Poisson process. By selecting appropriate statistical properties of service types and arrival count processes, every random process commonly used in telecommunications, as described by Frost and Melamed, can be constructed.[5]

In general, a multi-server model is appropriate at any node i, with the inter-service events of the processors for each service type s obeying a different PDF, $\mathrm{F}_{i,s,t}^d, t \in [0,T]$. According to the construction in [1, 3], the corresponding conditional random rate for the type s at node i, on the event $\{\tau_{i,n}^s \le t < \tau_{i,n+1}^s\}$, is $\sigma_{i,s,t \wedge \tau_{i,n+1}^s} = -\dfrac{d\,\mathrm{F}^d_{i,s,t \wedge \tau_{i,n+1}^s} / dt}{\left(1 - \mathrm{F}^d_{i,s,t \wedge \tau_{i,n+1}^s}\right)}$, where $\tau_{i,n}^s$ is the n'th service completion time at node i and "$\wedge$" denotes the infimum of two stopping times. As the PDF could also change after each time $\tau_{i,n}^s$, the construction allows a marked renewal sequence with a conditional PDF $\mathrm{F}_{i,s,n,t}^d$ between the n'th and $n+1$'th service completion times. The martingale representation theory for MVPPs, applied to the counting process $\breve{N}_{i,s,t}^D$ of the uninterrupted number of service completions of type s to time t, determines that

$$E\left[\breve{N}_{i,s,t}^D\right] = E\left[\sum_n \int_{\tau_n^s}^{t \wedge t_{n+1}^s} \sigma_{i,s,v}\, dv\right], \text{ and } \breve{N}_{i,s,t}^D - \sum_n \int_{\tau_n^s}^{t \wedge t_{n+1}^s} \sigma_{i,s,v}\, dv \tag{1}$$

is a zero-mean $(F_t, \wp)$-martingale, provided F_t is a σ-algebra of the network events to time t containing the history $\{\breve{N}_{i,s,v}^D, 0 \le v \le t\}$. A similar representation can be provided for the counting process $\breve{N}_{i,s,t}^A$ of the uninterrupted number of new call arrivals to node i of type s to time t in terms of the conditional arrival rate process $\alpha_{i,s,t}$ and the sequence $(\tau_{s,n}^a)$.

Each service type has unique QoS requirements directly related to the RRs at the destination node. For example, the requirements impose service priority disciplines at the nodes to minimize processing delays for delay-sensitive types. Multimedia applications can be divided into three different rates of traffic: CBR, VBR and ABR.

Voice and other constant bit rate applications. Historically, voice traffic and video codecs have injected CBR traffic into networks. These services could not function with less bandwidth or bit rate than some minimum, application-specific requirement, nor benefit from extra bandwidth. Running in circuit-switched wide area network (WAN) environments, these services receive dedicated bandwidth. VoIP services may require some form of flow-based reservation, since effective voice transport requires an application-to-application delay of less than 150 ms and a packet loss of less than 2%. Voice service can be modeled by a non-homogeneous Markov modulated Poisson process (MMPP), with one or two selectable CBRs, α_1 and α_2, as the Poisson intensities. These rates are modulated by a random "on-off" process V_A with a mean "on" time equal to the average talkspurt activity cycle. A single processor for live voice is typical at the MSs, while multiple voice processors in parallel are assumed at the BSs. Live voice service is generally not buffered, so that if the processors at a node are busy, the connection request for live voice is blocked, transferred to a neighboring BS, or dropped.

Available bit rate applications. Data applications, such as file transfers or multimedia mail and notes, can function with a wide range of available bandwidth. These services require little bandwidth to function slowly and operate faster with access to more bandwidth. Packets for these generally connection-less services can be buffered in queues at the nodes. Traditional packet-switched data networks can adequately support ABR services with best-effort QoS guarantees. The arrival and

service completions for ABR applications are best modeled with PDFs with parameters that can adjust to the available RRs of idle signal processors, transport channels, etc., such as those for marked renewal processes.

Variable bit rate applications. Traditional interactive data applications, such as Telnet sessions, and interactive multimedia applications, such as modern codecs and LAN TV, are more "bursty" in nature and fluctuate between low- and high-rate requirements. Researchers have previously chosen MMPPs to model aggregate voice, video and data VBR traffic.[8] However, the significance of LRD in aggregate traffic is under examination in recent studies [6, 9], which reveal that packet loss and delay behavior is radically different in simulations using real traffic data rather than the traditional MMPP models.

Circuit-switched networks are engineered to provide sufficient bandwidth in each circuit or virtual circuit to handle the *peak* rate required by VBR applications. When VBR traffic is below the peak rate, extra bandwidth is unused. Conversely, packet-switched networks provide sufficient bandwidth to handle two to four times the average rate required by the set of active VBR services. Peak demands are handled by statistical sharing of extra bandwidth, a technique known as *predictive* QoS. When traffic gets heavy at a node, express queues are added to expedite a subset of active services in admitted calls. Other services continue to be processed at the same time as those in expedited service lines, but at a slower rate.

3.2. Self-similar Traffic

Recently, self-similar (or fractal) stochastic processes have been proposed as more accurate models of certain categories of traffic in high-speed, high-bandwidth communications networks (e.g., LAN traffic, VBR video traffic, WAN traffic). Studies of LAN traffic [9] and WAN traffic [6] challenge the commonly assumed models for network traffic, e.g., the Poisson distribution and renewal processes. Were traffic to follow a Poisson or Markov arrival process, it would have a characteristic burst length that would tend to be smoothed by averaging over a sufficiently long time scale. Instead, measurements of real traffic indicate that significant variance (burstiness) is present on a wide range of time scales. Such traffic can be described statistically using the notion of self-similarity. Self-similarity is the property associated with fractals, i.e., the object appears the same regardless of the scale, temporal or spatial, at which it is viewed.

A *self-similar* time series has the property that, when aggregated (leading to a shorter time series in which each point is the sum of multiple original points), the new series has the same autocorrelation function as the original. That is, given a stationary time series, $\tau = (\tau_l;\ l = 0,1,2,\ldots)$, the m-aggregated series $\tau^m = (\tau_l^m; l = 0,1,2,\ldots)$ is defined by summing the original series τ over non-overlapping blocks of size m. Then, if τ is self-similar, it has the same autocorrelation function $R(k) = E[(\tau_l - \mu)(\tau_{l+k} - \mu)]$ as the series τ^m for all m. This means that the series is *distributionally* self-similar: the distribution of the aggregated series is the same (except for changes in scale) as that of the original.

As a result, self-similar processes show LRD. A process with LRD has an autocorrelation function $R(k) \sim k^{-\beta}$ as $k \to \infty$, where $0 < \beta < 1$. Thus, the autocorrelation function of such a process decays hyperbolically (as compared to the exponential decay exhibited by traditional traffic models). Hyperbolic decay is much slower than exponential decay, since the sum of the autocorrelation values of such a series approaches infinity. This condition has a number of implications. First, the variance of n samples from such a series does not decrease as a function of n (as predicted by basic statistics for uncorrelated data sets) but rather by the value $n^{-\beta}$. Second, the power spectrum of such a series is hyperbolic, rising to infinity at frequency zero, reflecting the "infinite" influence of LRD in the data.

One of the benefits of using self-similar models for time series, when appropriate, is that the degree of self-similarity of a series is expressed using only a single parameter. The *Hurst* parameter, $H = 1 - \beta/2$, expresses the speed of decay of the series' autocorrelation function. For self-similar series, $1/2 < H < 1$. As $H \to 1$, the degree of self-similarity increases. Thus, the basic test for self-similarity of a series reduces to the question of whether H is significantly different from 1/2.

Heavy-tailed distributions. The PDFs considered for Ethernet and other LAN traffic have the property of being heavy-tailed. A distribution is heavy-tailed if $P[\tau \geq t] \sim t^{-\theta}$, as $t \to \infty$, $0 < \theta < 2$. That is, regardless of the behavior of the distribution for small values of the variable, if the asymptotic shape of the distribution is hyperbolic, it is heavy-tailed.

The simplest heavy-tailed distribution is the *Pareto* distribution. The Pareto distribution is hyperbolic over its entire range; its probability density function is $f(t) = \alpha k^{\theta} t^{-\theta-1}, \theta, k > 0, t \geq k$, and its cumulative distribution function (CDF) is $F(t) = P[\tau \leq t] = 1 - (k/t)^{-\theta}$. The parameter k represents the smallest possible value of the random variable. Pareto distributions have been used for WWW file transfers in the technical analyses of ETSI study groups and other standards

bodies concerned with 3G data traffic. Pareto distributions have a number of properties that are qualitatively different from more commonly encountered ones, such as the exponential, normal, or Poisson. If $\theta \leq 2$, then the distribution has infinite variance; if $\theta \leq 1$, then the distribution has infinite mean. Thus, as θ decreases, an arbitrarily large portion of the probability mass may be present in the tail of the distribution.

Using measurements of WWW user inter-request times, the reasons for the heavy-tailed distribution of quiet times needed for self-similarity can be explained physically, by aggregating a large number of "on-off" renewal processes, whose distributions are heavy-tailed. As the size of the aggregation becomes large, then, after rescaling, the behavior turns out to be the Gaussian self-similar process called fractional Brownian motion. If, however, the rewards, instead of being 0 and 1, are heavy-tailed as well, then the limit is a stable non-Gaussian process with infinite variance and dependent increments. The limit process is not, however, a linear fractional stable motion, but a new type of infinite-variance, self-similar process.

Approximate distributions for self-similar point processes. In order to apply the MVPP approach to self-similar traffic, several useful models have been proposed that capture this behavior, e.g., M/G/∞ model with Pareto service times [10], the superposition of two-state Markov sources [11], the mixture of exponentials to fit the heavy-tail distributions, the superposition of N "on-off" processes with sub-exponential "on" periods [12], deterministic chaotic-maps, and self-similar (fractal) point processes [13]. In each of these models, it has been shown that the number of arrivals over an interval (number of busy servers in an M/G/∞ model) all exhibit an LRD correlation structure. A recent mathematically tractable model has been developed based on a fractal construction of a basic point process (cluster process), where clusters are embedded over an infinite number of time scales.[14] The new model decomposes the self-similar process in a way that is tractable for the accurate characterization and control of packet traffic as well as the efficient synthesis of the process in system simulations.

The point process is constructed recursively as a succession of embedded "on-off" processes that contain m time scales. This process may be viewed as the basic process embedded in the "on" state of the m'th time scale. The time between visits to the "on" and "off" periods in this m time-scale process is exponentially distributed with parameter λq^m, where λ is the underlying Poisson arrival rate and, after each arrival time, a decision is made with probability p to continue generating arrivals with rate λ or with probability $1-p=q$ to turn off for a period of time. The number of arrivals before entering an "off" period is geometrically distributed with a mean q^{-1}. Therefore, the probability density function of inter-arrival times for the corresponding point process is

$$\mathrm{f}_{m,\tau}(t) = \begin{cases} \dfrac{2\lambda e^{-\lambda t} + \sum_{i=1}^{m} (2q)^i \lambda q^i e^{-\lambda t}}{2 + \sum_{i=1}^{m} (2q)^i} & \forall t \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

As the number of time scales approaches infinity, interesting properties of the distribution of the time between clusters (inter-cluster gap) develop. Let $\mathrm{f}_{\tau}(t) \equiv \lim_{m\to\infty} \mathrm{f}_{m,\tau}(t)$, which uniformly converges $\forall t \geq 0$ to the limit probability density function, when the generalized process consists of an infinite number of time-scale embeddings,

$$\mathrm{f}_{\tau}(t) = \begin{cases} \dfrac{2\lambda e^{-\lambda t} + \sum_{i=1}^{\infty} (2q)^i \lambda q^i e^{-\lambda t}}{2 + \sum_{i=1}^{\infty} (2q)^i} & \forall t \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Long-range dependence for finite-horizon and finite-queue models. In some WIN models, the potential problem of LRD may be avoided by observing network behavior over finite horizons and considering practical structures of finite-buffer queues. Grossglauser and Bolot report that the amount of correlation that needs to be considered for performance evaluation depends not only on the correlation properties of the source traffic, but also on the time scales specific to the system under study.[15] For example, the time scale associated with a queueing system is a function of the maximum buffer size. Thus, if finite-buffer queues are used in the WIN, the impact on the loss of the correlation in the arrival process becomes negligible beyond a time scale referred to as the *correlation horizon* (CH).[15] This means that, for performance-modeling purposes, any model among a host of available models, including Markov and self-similar processes, can be chosen as long as the selected model captures the correlation structure of the source traffic up to the CH. This implies that truncated forms of standard heavy-tailed CDFs are sufficient for the development of accurate models.

Consider the special case when the inter-arrival time distribution is a truncated Pareto $F_{T_C}(t)$ is a truncated Pareto PDF defined by

$$F_{T_C}(t) = \begin{cases} \left(\frac{t+k}{k}\right)^{-\theta}, & \text{if } 0 < t < T_C < \infty \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where $1 < \theta < 2$ and the parameter T_C is referred to as the *cutoff lag*. Similarly, truncated versions of (3) can be defined. In any case, the cutoff lag eliminates correlation in the input process beyond a lag equal to T_C. Thus, its impact is similar to that of an "external shuffling" procedure, wherein a time series representing a realization of a process is divided into temporal blocks and the blocks are shuffled. However, the structure of the time series inside a block remains the unchanged. Thus, external shuffling removes correlation from the series or sessions beyond a lag equal to the length of a block.

One of the significant results in [15] is that, in systems with self-similar input traffic, adjusting the marginal scaling factor by statistical multiplexing of several streams or using source traffic control mechanisms is a very efficient way of reducing packet loss rate while keeping utilization high. Another important result is that there exists a CH such that the packet loss rate is not affected if the cutoff lag increases beyond CH. Thus, the CH separates relevant and irrelevant correlation, with respect to the packet loss rate. Intuitively, the CH depends on the correlation structure of the input traffic, and on the system under study, namely, the finite queue. Indeed, the finite queue sets a limit on the local memory of the system, since the buffer "forgets" or discards the past as soon as it begins an idle period or a full one. This is referred to as the resetting effect. Therefore, the CH can be expected to depend on the maximum queue size. Moreover, while correlation on all time scales has an impact on performance for an infinite queue, only the correlation up to the CH has an effect in the finite buffer queue.

3.3. Admission, Handoff, Routing and Connectivity

The handoff feature consists two different functions: handoff resource reservation (HRR) and handoff path switching (HPS). The HRR reserves and activates the new radio and wired resources required for handoff. After HRR is completed, the HPS switches the radio path from the old to the new BS, including any intermediate path combinations required.

For multimedia service in a WIN, handoff raises additional requirements compared to handoff in speech-only networks. Different services have distinct QoS requirements for rate, BER, delay, delay variation, etc. depending on the bearer services in use. Therefore, the HRR has to reserve sufficient resources on the specific communication path for the handoff to meet the QoS requirements. A user may tolerate or negotiate degradation of QoS to an acceptable level when handoff is performed, provided resources are not available in the new cell or sector. For example, a multimedia call, consisting of audio, data and video components, may after handoff consist of only audio and data, or perhaps audio, video, and reduced bit-rate data. This is due to scarce resources in the new cell assigned after handoff. The CAC or CHO algorithms may consider resource conditions in neighboring BSs and probabilities for handoffs between cells before allowing call setups.

The handoff of a multimedia call from BS i to BS j is modeled by a routing vector $\boldsymbol{u}_{ij,t} = \left(u_{ij,1,t}, u_{ij,2,t}, \ldots, u_{ij,S,t}\right)$. The s'th component of $\boldsymbol{u}_{ij,t}$ represents the effect of the handoff on service type s, e.g., a change in processing or bit rate, service cessation, or service interruption due to a lack of required resource (RR) at the intended BS. The vector $\boldsymbol{u}_{0j,t} = \left(u_{0j,1,t}, u_{0j,2,t}, \ldots, u_{0j,S,t}\right)$, i=0, regulates new connections of calls to BS j incoming from gateways of wired packet-switched networks. The components of $\boldsymbol{u}_{ij,t}$ are $\left(F_t\right)$-predictable (more strongly, left-continuous) indicator functions of random events. As such, the expectations of the $\boldsymbol{u}_{ij,t}$ with respect to probability $\wp$ are the probabilities of the call access or handoff events that influence the flow of integrated services through the WIN. The $(N + M_{\max} + 1) \times (N + M_{\max} + 1) \times S$ time-varying array, $U_t = \left(u_{ij,s,t}; i,j = 0,1,\ldots, N + M_{\max}; s = 1,\ldots, S\right)$, for $t \in [0,T]$ describes the random path connectivity granted to integrated services, due to access control, handoff, blocking, routing, resource reservation, as well as radio path distortions, over the period of observed network operation. Note that the sum of the entries of $\boldsymbol{u}_{ij,t}$ over j may be greater than 1 to model point-to-point, point-to-multipoint, and broadcast transmissions among the nodes.

As t varies, the sample paths of the entries $u_{ij,s,t}$ form the temporal evolution of call connectivity enabled by network conditions for both connection-less and connection-oriented services. The exponential random rates $\lambda_{i,t}$ of call duration times are assumed to be much lower than the conditional rates for packet arrivals and service completions for all service types. Then, for a call from an MS moving from node j_1 to node k, the expected values of the entries in an indicator-function

sequence, $(u_{j_1 j_2,s,\tau_1}(\omega), u_{j_2 j_3,s,\tau_2}(\omega),\ldots,u_{j_n k,s,\tau_n}(\omega))$, estimate a history of the routing path that the call bearing that service traverses, given that $\tau_1(\omega) < \tau_2(\omega) < \cdots < \tau_n(\omega)$ are the n arrival or service completion times that occur for service type s during the call at nodes $j_1, j_2, \ldots, j_n$, respectively.

Mobility, defined by the location, speed and direction of MS i at time t, together with the services active in MS's call, determine which BS node j in which cell layer k, (k=1,2,3) of an HCN can be accessed. Cell assignment of a connection request, whether new or handoff, is also modeled by the variables $u_{ijk,s,t}$, by adding a third index to designate the HCN layer. These variables contain factors $1\left(MS_{i,lat,t.} = lat., MS_{i,long,t} = long., MS_{i,vel.,t} \geq vel.\right)$ that indicate mobility conditions at time t used to determine the "best" feasible assignment, according to the criteria of HCN access procedures.

3.4. Network State Processes

A candidate state process of the WIN requires a sufficient number of dimensions to distinguish the services, nodes, sources and destinations of messages. Class 2 and Class 3 services can be queued during periods of deep fading, high interference, handoff blocking, or other channel degradation and to allow preemption by more delay-sensitive services. The queue of service-connected packets, both in service or awaiting service, in a buffer at network node i at time t>0, is represented as the discrete-valued vector of parallel queues, $\boldsymbol{Q}_{i,t} = \left(Q_{i,1,t}, Q_{i,2,t}, \ldots, Q_{i,S,t}\right)$, each component of which has a corresponding birth-and-death equation and a semimartingale representation in terms of a discrete-valued jump, right-continuous, zero-mean, $(F_t, \wp)$-local martingale and an integrated conditional random rate process with respect to the family of σ-algebras $\left(F_t, t \in [0,T]; F_t \subseteq F\right)^1$, generated by evolution of observed network events to time t, i.e.,

$$Q_{i,s,t} = Q_{i,s,0} + \left(Q_{i,s,t} - \int_0^t (\alpha_{i,s,v} - 1(Q_{i,s,v-} > 0)\sigma_{i,s,v})dv \right) + \int_0^t (\alpha_{i,s,v} - 1(Q_{i,s,v-} > 0)\sigma_{i,s,v})dv \tag{5}$$

If there is only one packet processor at the node, the total number of packets of all service types must contend for the single resource. This situation implies that access denial or service pre-emption among the services active in the call(s) to the node may be necessary to satisfy distinct QoS requirements. In the single-server case, the total number of packets at node i at time t is the sum over the number of service types of the components (5) of $\boldsymbol{Q}_{i,t}$.

The integrated conditional arrival and service completion rates in (5) take forms for the mobile nodes different from those for the fixed nodes. If the processing capabilities of the mobile can handle only one active call at a time, the instantaneous (F_t) – progressively measurable conditional rates for the arrivals and departures of service type s at MS i are given by

$$\alpha_{i,s,t} = u_{0i,s,t-}\alpha_{s,t-} + \sum_{j \in \{\text{BSs in the active set of MS } i\}} u_{ji.s.t-} 1\left(Q_{j,s.t-} > 0\right)\sigma_{j,s,t-} \tag{6}$$

and $\sigma_{i,s,t}$, respectively, where the exact formulae for these conditional rates depend on the PDFs in (1) for the underlying random events of packet arrival and service completion for each type s as well as on the extent of observed network history to time t on which the conditional rates are estimated. The indicator functions in (5) and (6) on the random events indicate that uninterrupted packet processing cannot occur when no packets of type s are in the node. For the standard case of non-homogeneous Poisson arrivals with deterministic, time-varying intensities $\alpha_{i,s,t}$, and single-stage exponential processors with deterministic transient rates $\sigma_{i,s,t}$ at the nodes, the form of the rates coincide with equation (6). In general, the structure of the rates in (6) is more complex and is conditioned on the time t between stopping times between $\tau_{i,n}^s$ and $\tau_{i,n+1}^s$, i.e., the event $\{\tau_{i,n}^s \leq t \leq \tau_{i,n+1}^s\}$, for each service type s.

At the fixed nodes of the WIN, there are typically multiple parallel processors that can handle at most L simultaneous calls, each of which may carry up to S active services. Therefore, at most $L \times S$ packet processors or processing modes are assumed available at fixed nodes, such as BSs. Calls arrive to fixed node j originating from the MSs in the coverage area of the BS or from the wired packet-switching infrastructure. At BS j or MSC j of the network, the random instantaneous rates of the arrivals and departures of service type s are given by

$$\alpha_{j,s,t} = \sum_{k \in \{\text{infrastructure connections to BS } j\}} u_{kj,s,t-}\alpha_{k,s,t-} + \sum_{i \in \{\text{MSs in coverage area of BS } j\}} u_{ij,s,t-} 1(Q_{j,s,t-} > 0)\sigma_{i,s,t-} \tag{7}$$

and

$$\sigma_{j,s,t-} = \sum_{l=1}^{L} 1\left(Q^{l}_{j,s,t-} > 0\right)\sigma^{l}_{j,s,t-}, \tag{8}$$

respectively, where $\sigma^{l}_{j,s,t}$ is the random service completion rate of the l'th processor for type s at fixed node j. The terms $u_{ij,s,t-}$ in (7) and (8) are $(F_t)-$predictable (left-continuous) indicator functions of the random events of CHO; packet routing, switching, access, denial or blocking; and packet regeneration that determine the flow of packets over both wired and wireless links. Connectivity between nodes is also determined by less controllable, often unobservable events, created by finite resource constraints. These constraints include finite buffer overflow, processor limitations and queueing delays, as well as radio path distortions due to multipath reflections, fading, and residual channel interference among users. Thus, the general structure of $u_{ij,s,\cdot}$ is a product of the indicators of both controlled, observed events and uncontrollable, indirectly observed events. As indicators of random events, the expected values of the $u_{ij,s,\cdot}$ with respect to $\wp$ and $(F_t, t \in [0,T], F_t \subseteq F)$ are the probabilities of the events. For example, for stationary packet routing at the nodes of a fixed wired network, with $u_{ij,s,t} = 1($ packets of type s from node i are routed to node j at time $t)$, $E[u_{ij,s,t}] = p_{ij,s}$, a fixed probability for any time $t>0$. Depending on the statistical characteristics of network events, i.e., ergodic, stationary, renewal, Markovian, etc., the IP traffic streams in the MVPP model become randomly modulated point processes of the corresponding stochastic type through the terms $u_{ij,s,\cdot}$ that indicate the events.

3.5. Quality of Service and Radio Resources

The QoS requirements of each service type are represented in the MVPP model in terms of parameters, either random or deterministic, that activate the network events corresponding to some of the indicator functions $(u_{ij,s,t})$. Parameters, such as, BER, delay, buffer size, received signal power, and other RRs are variables that determine explicitly the effective conditional rates of call arrivals, service packets processed, and handoffs for each service type at the required QoS. The MVPP models thus encompass *QoS-based routing*.

BER. BER is a major factor of QoS and is related to the average SIR per received bit, or E_b/I_0. The relationship between BER and E_b/I_0 is one-to-one and is determined by the RRs of channel coding, modulation, receiver sensitivity, processing gain, macro-diversity, and transmit power in the terminal equipment at network nodes. BER and E_b/I_0 also depend on channel fading statistics and radio link interference between mobile and fixed nodes. BER can be directly related to the RRs of downlink transmit power from the BSs in the active set to the MS, and uplink transmit powers from the MSs to a BS. Respectively, these output powers also cause mutual interference in the downlink and uplink. BER can also be related to the RR of processing gain due to adaptive spreading codes or, equivalently, to the adaptive number of transport channels used to increase effective bandwidth. Another RR affecting BER is FEC coding gain. At a BER of 10^{-3} in an additive noise environment, the rate ½ coding provides about 4 dB of coding gain. Puncturing the rate ½ code to produce a rate ¾ code reduces the coding gain to about 2.5 dB. Adaptive beamforming within a cell or cell sectorization can simultaneously increase the gain factors of BER and reduce co-channel interference. If the spatial distribution of MSs is assumed uniform in the cell, sectorization reduces interference and increases capacity by the antenna gain factor, G_A. For a 3-sector cell, this gain is less than 3. If the implementation loss from the ideal gain is assumed to be 1 dB, $G_A = 2.4$.

Voice activity monitoring (VAM) increases RR utilization based on the activity factor in speech. In the idle periods of one speech connection, RRs can be reallocated to provide service to other connection requests. This reduces the average signal power of all users from a uniform population and ensures, based on the weak law of large numbers, that the interference is nearly average most of the time. This factor is denoted as the voice activity gain, G_v, which by measurement has been established as $G_v = 2.5$ for 40% voice activity in a period. VAM can be modeled either by the indicator of a gain condition, $1(G_{\pi,ij,t} \geq 1/act\%)$, or by the indicator of the random event of on-off voice activity, $1(\kappa_{ij,1,t} > 0)$, where, by convention, service type $s = 1$ denotes live voice and $\kappa_{ij,1,t}$ the voice activity on the link between node i and node j at time t.

In general, the indicator $u_{ij,s,t}$ is the product of factors that include $1(\mathrm{BER}_s \leq 10^{-n})$ for service type s; this factor can itself be factored into a product of indicators of the acceptable range of values of the RR parameters that together comprise the QoS for type s. For example, $1(BER_t \geq 10^{-n}) = 1(P_{i,t} \geq p_i) \cdot 1(G_{\text{coding},ij,t} \geq 10^{y/10}) \cdot 1(G_{A,j,t} \geq g) \cdot 1(P_{\text{avg. path loss},ij,t} \leq \pi)$ $1(I_{\text{co-ch. interf.},ij,t} \leq \varsigma)$. For fading links, the condition $P_{\text{avg. path loss},ij,t} \leq \pi$ can be replaced with the random event of the number

of replicas received at node j of a signal transmitted from node i, according to a discrete-event distribution, and the relative path loss on each replica that follow a Rayleigh or Rician distribution. The joint distribution of the number of reflected paths and the amplitude of the reflections is required to determine the expected value of the indicator of the multipath event.

Bandwidth: transport channels, data rates and power. In the 3G standard of Wideband Code-division Multiple Access (W-CDMA), options are available to integrate multi-rate services: (1) trade off processing gain for increased data rate in the same spread bandwidth and (2) pair up basic transport data channels until the required rate is obtained. The phrase "basic channel" refers to the CBR transmission with the highest processing gain. The RRM in the RNC fully controls the choice of appropriate coding scheme, interleaving parameters, and rate-matching parameters.

The MAC protocol controls the multimedia data stream delivered to the physical layer over the transport channels. If an MS wants to transmit data of different services, e.g., a real-time service and packet data service, it is assigned two sets of transport formats, one for real-time service and one for packet data service. As for a single service, the MS may use any transport format assigned for real-time services, whereas it may only use the transport formats for the data service. The MS is assigned a specific output power/rate threshold. The aggregate output power/rate will never exceed the threshold. Thus, the transport formats for data service fluctuate adaptively to transport formats used for the speech service.[16]

Delay. Packet flow control mechanisms can assign high priorities to mission-critical traffic or traffic sensitive to delay variation, ensuring that it will be advanced in the queue for transmission before all other traffic types. Prioritization at a single-processor node can be modeled by re-ordering the service times $\left(\tau_{i,n}^{s}, s=1,\ldots,S\right)$ at node i, conditioned on the number of packets, $Q_{i,s,t}$, of each type s currently at the node at time t and the last service completion time $\tau_{i,n^*}^{s^*}$ before t. The custom queueing feature of some packet switch equipment allows reserving specific quantities of bandwidth for each type s to ensure specific streams a minimum quantity of bandwidth. The MVPP model adapts service rates $\left(\sigma_{i,s,t}, s=1,\ldots,S\right)$, up to a maximum allowable total rate, i.e., $\sum_{s=1}^{S}\sigma_{i,s,t} \le \Lambda_{i,t}$, conditioned on the number of packets of each type and the last service completion time, to reduce any backlog of high-priority packets at the node.

Sources of delay and delay variation are the number of handoffs that occur over the duration of a call due to user mobility, handoff failures, and queueing latency when data packets are retransmitted in response to unrecoverable block errors. The queueing delay and its variation at a node can be bounded to not exceed maximum delay and variance targets, Δ_s and $\sigma_{\Delta,s}^2$, respectively, as part of the QoS for service type s. The average allowable delay for type s is denoted $\overline{\Delta}_s$. Other sources of delay, such as radio propagation, are assumed relatively small or are allocated to mechanisms in the wired network.

Little's formula states that the average number of customers in a queueing system in steady-state is equal to the arrival rate of customers to the system, times the average time spent in the system. The result makes no specific assumption regarding the arrival distribution or the service time distribution; nor does it depend upon the number of servers in the system or upon the particular queueing discipline within the system. In terms of the MVPP model of the queue of packets for service type s, the random arrival rate of packets of type s and the delay limits, the delay condition can be approximated instantaneously at time t or by a time average over an observation interval $[0,T)$,

$$1\left(Q_{i,s,t} < \Delta_s\left[u_{0i,s,t}\alpha_{s,t} + \sum_{j\in\{\text{location area of node } i\}} u_{ji,s,t}\sigma_{j,s,t}\right]\right), \tag{9}$$

$$1\left(\int_0^T Q_{i,s,\tau}d\tau < \overline{\Delta}_s\left[\int_0^T u_{0i,s,v}\alpha_{s,v}dv + \sum_{j\in\{\text{location area of node } i\}}\int_0^T u_{ji,s,w}\sigma_{j,s,w}dw\right]\right). \tag{10}$$

The delay variation condition can be expressed in terms of instantaneous variance of queueing delay at node i at time t as

$$1\left(\left(Q_{i,s,t} - E[Q_{i,s,t}]\right)^2 < \sigma_{\Delta,s}^2\left[\left(u_{0i,s,t}\alpha_{s,t} + \sum_{j\in\{\text{location area of node } i\}} u_{ji,s,t}\sigma_{j,s,t} - E\left[u_{0i,s,t}\alpha_{s,t} + \sum_{j\in\{\text{location area of node } i\}} u_{ji,s,t}\sigma_{j,s,t}\right]\right)^2\right]\right) \tag{11}$$

where expectation is taken with respect to the probability measure $\wp$.

3.6. Performance Metrics

Call Dropping and Blocking. Delays, call interruptions and losses at the network level are due to handoff decisions in the presence of new and existing connection requests for the limited RRs at the nodes. Moreover, in micro- and picocells of an HCN, frequent handoffs greatly increase the signaling load on network control processors to negotiate the same QoS on the new links along the path. If the queueing of non-real-time traffic fails to maintain QoS due to buffer overflows and excessive delays, one or more active services in the call may have to be dropped to release more system resources to higher-priority services or to support new calls. In an HCN a handoff attempt may also be dropped or blocked from another cell or cell layer if there are no feasible RR allocations to service the call in the new location. Calls arriving to the network are blocked from entry for the same reasons. Mechanisms of dropping and blocking due to resource constraints are represented in the model by the functions $u_{ij,s,\cdot}$ that randomly modulate the rates of the packet flows. These functions contain terms of the form $1(\text{Available}\, RR_{j,t} \geq \rho_s)$, where $RR_{j,t}$ is the radio resource at node j at time t required at level ρ_s to maintain type s QoS.

Metrics of system performance, either local or global, can be represented by the model as the expected number of blocked calls, handoffs, or dropped calls, either for a given service type or all service types. These quantities are merely the expectations of the counting processes for the corresponding network events over the observation period, summed over the indices of the nodes and service types of interest. Based on the semimartingale representations of these underlying MVPPs and on the assumption that the rates of arrivals and service completions are non-explosive, the Fubini Theorem can be applied to represent these expectations, with respect to the σ-finite measure $\wp$, as sums, over the service types and nodes of interest, of the integrals of the expected (F_t)-predictable rates of the corresponding packet flows over $[0,T]$. For example, the random number of blocked calls of service type s to node j over $[0,T]$ is represented as the $(F_t, \wp)$-semimartingale

$$N_{\text{blocked},j,s,[0,T]} = \left[N_{\text{blocked},j,s,[0,T]} - \int_0^T (1-u_{0j,s,v}) 1(Q_{j,s,v-} \geq q_{j,s}) \alpha_{s,v} dv \right] + \int_0^T (1-u_{0j,s,v}) 1(Q_{j,s,v-} \geq q_{j,s}) \alpha_{s,v} dv, \qquad (12)$$

with expected value

$$E[N_{\text{blocked},j,s,[0,T]}] = E\left[\int_0^T (1-u_{0j,s,v}) 1(Q_{j,s,v-} \geq q_{j,s}) \alpha_{s,v} dv \right] = \int_0^T (P(Q_{j,s,v-} \geq q_{j,s}) - P(u_{0j,s,v}(Q_{j,s,v-} \geq q_{j,s}))) \overline{\alpha}_{s,v} dv \qquad (13)$$

where the second term in the integrand on the right-hand side of (13) is a joint probability, $\overline{\alpha}_{s,t}$ is the $\wp$-mean of the arrival rate of type s at time t, and $q_{j,s}$ is buffer size at node j for service type s. Expression (13) can be generalized to provide representations of the average number of handoffs or dropped calls for any or all service types, as the expectations of the indicator functions for those events, summed over the appropriate indices of interest, then integrated over the interval $[0,T]$. Instantaneous mean rates of handoffs or dropped calls are then the integrands of the average number of the corresponding network events, at some time $t \in [0,T]$.

4. PARTIALLY OBSERVED POINT PROCESSES

The dynamic nature of WINs makes it difficult to provide QoS due to the need to update routing state information, as represented by the time histories of the sample paths of $\boldsymbol{U}_t(\omega) = (u_{ij,s,t}(\omega); i,j = 0,1,\ldots,N+M_{\max}; s = 1,\ldots,S)$ and of the global network state, $\boldsymbol{Q}_t(\omega)$. Knowledge of network events on which CHO procedures are based is inherently incomplete, both temporally and spatially. In general, MS i is at most aware of the network events within the AS_i cells comprising its active set, since the time of the last handoff. The BSs may share similar call processing event history with the other C_j cells within the location area defined by the control range of the corresponding BSC j or MSC j. Imprecise knowledge of uncontrolled or unobserved phenomena, such as multipath fades and radio interference from outside the network, further limits the information available to controllers at all nodes and the ability of network designers to estimate accurately the PDFs corresponding to these random phenomena.

One approach to this problem is to group network links into two types: stationary and transient. Stationary links are those between stationary nodes or slowly moving nodes, such as BSs and office or residential users, likely to exist for a long time. Transient links are those between mobile nodes moving very fast. A link <i,j> can be typed as a stationary or transient link, provided both nodes are stationary or at least one of the nodes is mobile, respectively. To reduce the probability of link outages, stationary links are always preferred when choosing a QoS path. When an existing path is severed, the corresponding traffic flow is rerouted to another feasible path. During the period after the old path is broken and before the new path is set up, best-effort routing is used to redirect the traffic flow. For this reason, wireless networks normally can

only provide soft QoS, which means the required QoS is not guaranteed for some transient time periods, when the routing path is severed or the network is partitioned due to the motion of some network nodes.

The MVPP approach encompasses the situation of partial observations. Recall that the MVPPs in the system are progressively measurable with respect to the family of increasing, right-continuous σ-algebras $(F_t, t \in [0,T])$, with $F_t \subseteq F_v \subseteq F_T$ for $t \le v, t, v \in [0,T]$ and $F_t = \sigma\{\omega : (\boldsymbol{U}_v(\omega), \boldsymbol{Q}_v(\omega)), 0 \le v \le t\}$, generated by the augmented global network state. Partial or incomplete information is modeled by families of increasing, right-continuous σ-subalgebras, $(G_t, t \in [0,T])$, with $G_t \subseteq F_t \subseteq F_T$ for $t \in [0,T]$ and $G_t = \sigma\{\omega : \boldsymbol{X}_v(\omega), 0 \le v \le t\}$, an internal history generated by some process $\boldsymbol{X}_t$ constructed from observable events at the nodes to time t. The cumulative partial observation σ-subalgebras, or, equivalently, the state process generating them, can be partitioned by nodes forming a link, by cell layer and/or by service type, into a union of local observations available at the mobile and fixed nodes, i.e, $G_t = \bigcup_{i=0}^{M} \bigcup_{j=0}^{N} \bigcup_{l=1}^{3} \bigcup_{s=1}^{S} G_{ij,l,s,t}$, where the local observations at time $t \in [0,T]$ may not be disjoint, that is, $G_{ij,l,s,t} \cap G_{i^*j^*,l^*,s^*,t} \neq \varnothing$, $(i, j, l, s) \neq (i^*, j^*, l^*, s^*)$, for nodes within a common active set or common location area. Note that each local σ-subalgebra of partial observations, $G_{ij,l,s,t} = \sigma\{\omega : \boldsymbol{X}_{ij,l,s,v}(\omega), 0 \le v \le t\}$, can be the internal history of some process defined in terms of local events. In notation, $G_t = \bigcup_{i=0}^{M} \bigcup_{j=0}^{N} \bigcup_{l=1}^{3} \bigcup_{s=1}^{S} \sigma\{\omega : \boldsymbol{X}_{ij,l,s,v}(\omega), 0 \le v \le t\}$ This construction leads to a local filtration of the completely observed network state and of the integrated conditional rates corresponding to the predictable components of the state's underlying point processes. The theory of the filtration problems for partially observed, MVPPs can be found in the cited works of Brémaud and others.[1,2] Let $\hat{E}[f_t | G_t]$ or $\hat{f}_t$ denote the (G_t)-predictable version of the conditional expectation of the right-continuous, (F_t)-progressively measurable random process f_t with respect to the history of observations to time t, i.e., $E[f_t | G_t]$. Then, it can be shown that the partially observed components of the network state lead to a recursive form for the G_t-filter of $Q_{i,s,t}$ that resembles a discrete-space version of a Kalman filter, based on an application of Theorem T1, Chapter IV.[1] The structure of the filter is based on the decomposition of the network state component $Q_{i,s,t}$ into its constituent MVPPs and their corresponding integrated conditional rates shown in (6), (7) and (8).

To make use of the theorem, the semimartingale representations of state are recast in terms of new state variables, $Z_{i,s,t}(n) = 1(Q_{i,s,t} = n)$ for $n \in \mathbf{Z}_+$ a positive integer and $\hat{Z}_{i,s,t}(n) = E\left[Z_{i,s,t}(n) \middle| G_t^X \right]$, where X is one or more of the observed processes, discrete- or continuous-space, that generate the internal history of partial observations, (G_t). An explicit form of the G_t-filter is comprised of the G_t-innovations processes

$$\hat{M}^{A}_{i,s,t} = N^{A}_{i,s,t} - \int_0^t \hat{\alpha}^{X}_{i,s,v} dv \text{ and } \hat{M}^{D}_{i,s,t} = N^{D}_{i,s,t} - \int_0^t E\left[\sigma_{i,s,v}\left(\sum_{j \neq i} u_{ij,s,v} \right) 1(Q_{i,s,v-} > 0) \middle| G_v^X \right] dv \tag{14}$$

for exogenous arrivals to and departures from, respectively, the queue of packets of type s at node i to time t, and the corresponding innovations gains, $K^{A}_{i,s,t}(n)$ and $K^{D}_{i,s,t}(n)$ to yield the equation, for $n \in \mathbf{Z}_+$,

$$\begin{aligned} \hat{Z}^{X}_{i,s,t}(n) = P[Q_{i,s,0} = n] &+ \int_0^t \hat{f}^{X}_{i,s,v}(n) dv + \int_0^t K^{A}_{i,s,v}(n) \left(dN^{A}_{i,s,v} - \hat{\alpha}^{X}_{i,s,v} dv \right) \\ &+ \int_0^t K^{D}_{i,s,v}(n) \left(dN^{D}_{i,s,v} - \hat{E}\left[\sigma_{i,s,v}\left(\sum_{j \neq i} u_{ij,s,v} \right) 1(Q_{i,s,v-} > 0) \middle| G_v^X \right] \right) dv, \end{aligned} \tag{15}$$

where

$$f_{i,s,t}(n) = \left(Z_{i,s,t}(n-1) 1(n > 0) - Z_{i,s,t}(n) \right) \alpha_{i,s,t} + \left(Z_{i,s,t}(n+1) - Z_{i,s,t}(n) 1(n. > 0) \sigma_{i,s,t}\left(\sum_{j \neq i} u_{ij,s,t} \right) \right) \tag{16}$$

and the random rates in the second and third terms on the right-hand side of (15) are predictable versions of the (G_t^X)-rates of $N_{i,s,t}^A$ and $N_{i,s,t}^D$, respectively. In order to achieve a more recursive structure for the filter, equations for the innovation gains in (15) must be determined and solved in terms of the MVPPs and other processes that generate the state and partial observations. Applying result R6 in [1] to the filtration of $Q_{i,s,t}$, the innovation gains in (15) for service type s at mobile node i at time t, given that the last observed state of the queueing subnetwork of type s, $\boldsymbol{Q}_{s,t-}(\boldsymbol{n}) = (n_1, n_2, \ldots, n_i, \ldots, n_j, \ldots)$, can be computed and depends on the point processes that generate the partial observations,

$$K_{i,s,t}^A(\boldsymbol{n}) = -\hat{Z}_{i,x,t-}^A(n_i) + \frac{u_{ii,s,t-}\hat{\sigma}_{i,x,t-}^A 1(n_i>0)\hat{Z}_{i,x,t-}^A(n_i) + \left(u_{0i,s,t-}\hat{\alpha}_{s,t-} + \sum_{j\in\{\text{BSs in the active set of MS } i\}, i\neq j} u_{ji.s.t-}\hat{\sigma}_{j,x,t-}^A 1(n_j>0)\hat{Z}_{j,x,t-}^A(n_j)\right)1(n_i>0)\hat{Z}_{i,x,t-}^A(n_i-1)}{\left(u_{0i,s,t-}\hat{\alpha}_{s,t-} + \sum_{j\in\{\text{BSs in the active set of MS } i\}, i\neq j} u_{ji.s.t-}\hat{\sigma}_{j,x,t-}^A 1(n_j>0)\hat{Z}_{j,x,t-}^A(n_j)\right) + u_{ii,s,t-}\hat{\sigma}_{i,x,t-}^A\left(1-\hat{Z}_{i,x,t-}^A(0)\right)} \tag{17}$$

$$K_{i,s,t}^D(\boldsymbol{n}) = -\hat{Z}_{i,s,t-}^D(n_i) + \frac{u_{ik.s.t-}\hat{\sigma}_{i,x,t-}^D 1(n_i>0)\hat{Z}_{i,s,t-}^D(n_i) + \left(\sum_{k\in\{\text{BSs in the active set of MS } i\}, k\neq i} u_{ik.s.t-}\hat{\sigma}_{i,x,t-}^D\right)\hat{Z}_{i,s,t-}^D(n_i+1)}{\left(\sum_{k\in\{\text{BSs in the active set of MS } i\}} u_{ik.s.t-}\hat{\sigma}_{i,x,t-}^D\right)\left(1-\hat{Z}_{i,s,t-}^D(0)\right)} \tag{18}$$

Note that the exact form of the set of filtration equations (14)-(15) may include terms corresponding to either one or both of the innovations gains (17) and (18), provided that either of the point processes $N_{i,s,t}^A$ or $N_{i,s,t}^D$ or both are generators of the partial observations algebra (G_t^X), respectively. Although somewhat tedious in form, versions of this filtration can be applied to recursively update estimates of specific WIN system parameters, based on the evolution of observed network events within connected neighborhoods of mobile and fixed nodes. Indeed, many QoS-based routing algorithms recently proposed are based on either local states or imprecise observations of more comprehensive network states.

5. CONCLUSIONS

MVPP models for the packet flows of the integrated services among nodes of a wireless multimedia network have been proposed to represent the random, real-time behavior of call processing events. The analytical models lead, using the methods of the theory of semimartingale decompositions of point processes, to representations of transient packet flows. In this manner, the flow models form the mathematical basis on which to construct real-time algorithms for the decentralized control of call admission, handoff, routing, and congestion, based on perfect or incomplete observations of the history of network events available to the controllers at the nodes. The general model is shown to be extendible to a wide range of statistical properties ascribed to the arrivals, service processing, handoffs, and routing of various types of integrated services in recently published studies of wireless multimedia telecommunications.

Further examination remains of the limitations of the proposed real-time models to represent accurately all IP services, as well as the exploration of the ways in which the models can be embedded in the structure of the layers of wireless IP protocols. A future tradeoff analysis between network delays and the span of simultaneous call event observations available to network controllers is necessary to determine the performance limits of practical implementations of the multimedia control algorithms. The next goal of this research is the development of efficient real-time, adaptive control algorithms of multimedia streams based on specific cases of the MVPP models.

ACKNOWLEDGMENTS

The author wishes to thank FIT for its support and to commend the special mobile groups (SMGs) of 3rd Generation Partnership Project (3GPP) and the technical teams of the TIA 45.5 committees as part of the 3rd Generation Partnership Project 2 (3GPP2) for incorporating features for the adaptive QoS provisioning to enable packet-based multimedia services.

REFERENCES

1. P. Brémaud, *Point Processes and Queues: Martingale Dynamics*, Springer-Verlag, New York, 1981.
2. J. Walrand and P. Varaiya, "Flows in queueing networks: a martingale approach," *Math. Oper. Resrch.*, vol. 6, pp. 387-404, 1981.
3. W. S. Hortos Jr., *Partially Observable Point Processes and the Control of Packet Radio Networks*, doctoral dissertation, University of Michigan, UMI Dissertation Information Services, Ann Arbor, May 1990.
4. R. Boel, P. Varaiya and Wong, "Martingales on jump processes, I: representation results," *SIAM J. on Control*, vol. 15, no. 5, pp. 999-1021, Aug. 1975.
5. V. S. Frost and B. Melamed, "Traffic modeling for telecommunication networks," *IEEE Comm. Mag.,* vol. 32, no.3, pp. 70-81, Mar. 1994.
6. W. Willinger, M. S. Taqqu, W. E. Leland, and D. V. Wilson, "Self-similarity in high-speed packet traffic: analysis and modeling of Ethernet traffic measurements," *Statist. Sci.*, vol.10, no. 1, pp.67-85, 1994.
7. A. S. Acampora and M. Naghshineh, "Control and quality-of-service provisioning in high-speed microcellular networks," *IEEE Personal Comm. Mag.*, pp. 36-43, 1994.
8. T. V. J. G. Babu, T. Le-Ngoc, and J. F. Hayes, "Performance of a priority-based dynamic capacity allocation scheme WATM systems," *Proc. IEEE GLOBECOM '98*, vol. 4, pp. 2234-2238, Nov.1998.
9. W.E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Networking*, vol. 2, no.1, pp.1-15, Feb.1994.
10. M. Parulekar and A. Makowski, "Tail probabilities for a multiplexer with self-similar traffic," *Proc. IEEE INFOCOM '96*, vol. 3, pp. 1452-1459, Mar. 1996.
11. A. T. Andersen and B. F. Nielsen, "An application of superpositions of two-state Markovian sources to the modelling of self-similar behaviour," *Proc.IEEE INFOCOM '97*, vol. 1, pp. 196-204, Apr. 1997.
12. N. Likhanov, B. Tsybakov, and N. D. Georganas, "Analysis of an ATM buffer with self-similar ("fractal") input traffic," *Proc. IEEE INFOCOM '95*, vol. 3, pp. 985-992, Apr.1995.
13. W. M. Lam and G. W. Wornell, "Multiscale representation and estimation of fractal point processes," *IEEE Trans. Sig. Process.*, vol. 43, no. 11, pp. 2606-2617, Nov. 1995.
14. M. A. Krishnam, A. Venkatachalam and J. M. Capone, "Self-similar point process through a fractal construction," submitted to *Proc. IEEE INFOCOM 2000*, Tel Aviv, Israel, 8 pages.
15. M. Grossglauser and J.-C. Bolot, "On the relevance of long-range dependence in network traffic," *IEEE/ACM Trans, Networking*, vol. 7, no.5, pp. 629-640, Oct. 1999.
16. ETSI/SMG2, "The ETSI UMTS Terrestrial Radio Access (UTRA) ITU-R RTT candidate submission," submitted to ITU-R TG 8/1 as an IMT-2000 proposal, Jun.1998.