Optimization of real-time protocols for wireless packet-switched multimedia networks based on partially observed multivariate point processes

William S. Hortos

Optimization of real-time protocols for wireless packet-switched, 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

Next-generation wireless networks have been designed to transport integrated multimedia services based on a cellular extension of a packet-switched architecture using variants of the Internet protocol (IP). Each call, arriving to or active within the network, carries demand for one or more services in parallel, where each service type has 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. MS roaming among the nodes of the WIN is controlled by handoff procedures between base stations (BSs), or BS controllers (BSCs), and the MSs. Both handoff and CAC procedures are typically embedded in the media access control (MAC) and logical link control (LLC) layers of the WIN protocol stack. Performance metrics, such as, probabilities of call blocking and handoff, control procedures within the WIN protocol. Earlier published results by the author present a method for the performance analysis of wireless multimedia networks based on multivariate point-process (MVPP) models and their semi-martingale representations. The models describe the finite-horizon, transient behavior of the packet flows in integrated multimedia traffic. This paper extends the earlier work to focus on the optimization of real-time control procedures embedded in the WIN protocol stack. The point processes corresponding to the packet flow of each service type and to collateral random network events are decomposed into right-continuous, pure jump processes and predictable, integrated random rate processes. Control of the set of point processes describing WIN behavior is implemented through the construction of an absolutely continuous change of a reference probability measure on network events to a controlled measure. With respect to the constructed controlled measure, the integrated random rates in the semi-martingale representations of the network point processes explicitly depend on the parameters of the protocol mechanisms for packet access, routing, switching, handoff, power control, and other resource allocations. The MVPP control approach via constructed probability measure also supports predictive models of packet flows that incorporate measurement-based estimates of the probability distributions for voice, video, data, and other Internet traffic as well as radio path distortions. Optimization of the performance metrics, represented in terms of the MVPPs and the control parameters of candidate WIN protocols for multimedia services, are discussed for both partial and complete observations of network events. The models are further used to develop optimal, recursive stochastic filters of network state, based on partial or incomplete observations of packet flow dynamics.

Keywords: Wireless IP networks; real-time protocols; integrated multimedia services; resource allocation; quality-of-service (QoS) routing; multivariate random point processes; adaptive estimation; stochastic optimal control; martingale representations; controlled probability measures

1. INTRODUCTION

Adaptive techniques have been proposed to enhance mobility and portability performance in current and next-generation wireless multimedia networks. Quality of service (QoS) provisioning in all classes of network services, including voice, video, data 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) to parallel and possibly supplant the existing circuit-switched networks has been initiated by telecommunications standards bodies and infrastructure equipment manufacturers. An increasing level of multimedia traffic will be transported in packets 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. Consequently, 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 call events for real-time services. In the past, Markov and Bayesian methods have been employed to formulate and evaluate control algorithms that adapt to changing network conditions in real time, and to enable decentralized management procedures to handle network resources. These management procedures ensure predefined levels of QoS to different traffic classes of multimedia services in complex hierarchical cellular networks (HCNs).

This paper extends previous development by this author of analytical models of the behavior of packet flows in wireless multimedia network, based on the theory of semi-martingale representations of multivariate point processes (MVPPs) with randomly modulated rates.[1] In that work, the models achieve a comprehensive representation of the controllable and observable, real-time call processing events. The MVPPs are used to represent the transient packet flows of information 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.

The extension of the models in this paper is based on an application of Girsanov's theorem to control the MVPP rates through an absolutely continuous change of probability measure from a reference measure on network events. The controlled rates represent the mechanisms of packet entry, routing and switching in the network, as 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. The counting processes on packet events as well as inter-event times can follow general non-stationary probability distribution functions (PDFs) for call arrivals, cell occupancy, call holding, integrated service loading, and service completion. 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, handoff and blocking, depend on non-stationary path loss distributions due to fading, reflections and user mobility, traffic loading, as well as on the observable network history. The set of these random variables, with the addition of technical mathematical assumptions, constitute the admissible class of controls which influence network models through the change of probability measure on network events.

The extent of observed network history available 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. Observations of network events can be partial based on incomplete state information or incomplete in time, either deterministic instants or at random instants according to a known PDF.

The MVPP models generalize Poisson point processes, exponentially distributed call processing, and other renewal processes commonly assumed to formulate queueing structures for the evaluation of the asymptotic performance of control schemes for packet networks at or near equilibrium. The actual WIN dynamics, 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, $[0,T], T<\infty$. The MVPP approach also 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 proposed MVPP models of real-time network events are not only mathematically tractable and extendible, but can 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.

Over two decades ago, Brémaud [2], Walrand and Varaiya [3] and others established the theoretical foundation for general semi-martingale 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 semi-martingale equation in continuous time or between random stopping times.[2] In cases of incomplete observations of network behavior, the effects of a real-time adaptive control algorithm can be 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 [4] and that of Boel, Varaiya and Wong [5] in the control of point processes with incomplete information.

The optimal control problem is then formulated based on the MVPP models of WIN behavior. Performance metrics, such as, average waiting time or 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, necessary to achieve QoS requirements, modulate and constrain the stochastic conditional rates in the semi-martingale respresentations of the MVPPs. Under special cases of network observability, stochastic dynamic programming conditions, similar in form to Hamilton-Jacobi-Bellman condtions, can be formulated as sufficient conditions to characterize the optimal control policies for the optimization problem. Methods are suggested on how to extend the MVPP approach to Markov models and to cases of incomplete information.

2. FEATURES OF WIRELESS MULTIMEDIA NETWORKS

The WIN operations that handle the flow of multimedia packets during the input, support, and completion of a call are first introduced. 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, $\mathrm{F}^a(t), \tau_n^a < t \leq \tau_{n+1}^a$, 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., $\mathrm{F}_n^a(t), \tau_n^a < t \leq \tau_{n+1}^a$, 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 at most *S* service types. Each type requires a distinct 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. Without the exponential assumption on the inter-service times, superposition of the streams of service completions will not, in general, form a renewal process.[6]

Modeling assumptions of the MVPP approach must allow the self-similar behavior [7] of World Wide Web (WWW) traffic. Indeed, the only assumptions required in the general analytical development are those that support the semi-martingale 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 $(\Omega; F_t; \wp)$ or controlled probability space $(\Omega; F_t; \wp^U)$.

2.1. Integrated service classes

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 (jitter) noticeably degrades the QoS of 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 supported by available bit rate (ABR), 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 delay, e.g., many web pages include real-time video and audio clips.

While some QoS measures are expresssed in terms of the interrelated parameters of receive signal strength (RSS), signal-to-interference ratio (SIR), bit-error ratio (BER), and frame-error ratio (FER), others distinguish service classes by their tolerance to fixed delay and delay variations. All classes of service types are accommodated in the model.[8] Both CBR and VBR services are assumed in each class. 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 include 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 generate must be delivered in a certain temporal sequence. 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 availability.

Class 2 services include non-real-time, delay-sensitive, connection-oriented services with limited delay requirements such as MPEG-2 video, 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 a VBR 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 such as WWW downloads. 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 (ABR) and other resources, after the QoS of active services from the other two classes have been met first.

The 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 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 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 picocell. Handoffs and new call admissions depend on three factors: availability of RRs in the active set of cells in the targeted layer to meet the QoS needs of the call, the match between the layer and MS mobility, and interference induced in the designated cell by acceptance of the connection request.

2.3. Radio resources and transport channels

Key features in 3G radio transmission technology enable control of the transport of multiple parallel services with different QoS requirements on one wireless connection. A flexible, dual-mode packet access scheme allows transmission either on common channels (CCHs) or 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 a single long packet or 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 the method and characteristics of data transfer 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 regulates target SIR, 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. Control of 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 portion of mobility support involves resource allocation to maintain the QoS of the established connection(s). If sufficient RRs are not allocated or available, QoS may not be met. This may lead to loss of the connection, i.e., handoff failure and subsequent call dropping, if a reduced QoS level cannot be negotiated. It is common practice to give higher priority to handoff requests than to new call requests and 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 measure of the pilot or beacon channels in the downlink to the MS from the BSs. During the 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*). The "best" BS in the active set is the one with highest value of the signal quality measure, while all other BSs of the set are within the handoff margin, 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*, allows the MS to be connected to more than one sector within the same cell. In *hard handoff*, a connection on the current frequency may be severed as it moves to a new frequency during handoff.

3. ELEMENTS OF CONTROLLED NETWORK MODELS

This section describes specific representations of a wide range of the traffic, link distortions, control and other operational elements of multimedia WINs in terms of the real-time MVPP models.

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. 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 $b_{s,n} = 0$ indicates 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 counting processes, every random process commonly used in telecommunications can be constructed.[6]

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 [2, 4], 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 $\sigma_{i,s,t\wedge\tau_{i,n+1}^{s}} = -\frac{d\,\mathrm{F}_{i,s,t\wedge\tau_{i,n+1}^{s}}^{d} / dt}{\left(1-\mathrm{F}_{i,s,t\wedge\tau_{i,n+1}^{s}}^{d}\right)}$, 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 can 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 $\left(\tau_{s,n}^{a}\right)$.

Voice and other CBR applications. Voice traffic and interactive video inject 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. 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 can be 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, in the worst case, dropped.

VBR 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.[9] For example, an MPEG-2 video encoder generates a bit stream which is modeled at the video frame level. MPEG-2 frames can be of type intra (*I*), predictive (*P*), or bi-directional (*B*). Here only the *I* and *P* frame types are considered, with the *I* frames being generated at scene changes. The *I* frame bit rates contribute to the largest amplitudes, while the *P* frames transmit the differential information in successive frames and result in a distribution of moderately valued frame rates. VBR video source model considered is represented by an I state discrete-time Markov chain, with a transition probability matrix $\mathbf{P}_V$ and a rate vector $\mathbf{R}_V = [r_1, r_2, \ldots, r_I]$. The rate r_I represents the number of bits generated per video frame when the process is in state *i*. The last state *I* represents the intraframe state and the remaining states corresponding to the *P* frames. The diagonal elements of $\mathbf{P}_V$ are dominant except that p_{II}=0. The latter transition results from an immediate transition from an *I* frame to a *P* frame. The diagonally dominant structure signifies that VBR video is characterized by strong short-term correlations.

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.

ABR applications. Data applications, such as file transfers or multimedia mail, function with a wide range of available bandwidth. Packets for such connection-less services can be buffered in queues at the nodes. Packet-switched data networks can adequately support ABR services with best-effort QoS guarantees. Arrival and service completions for ABR applications are best modeled with PDFs with parameters that can be adjusted to the available RRs of idle signal processors, transport channels, etc., such as those for marked renewal processes. Recent studies [7, 10] reveal that packet loss and delay behavior is very different in simulations using actual aggregrate self-similar traffic data with LRD rather than traditional MMPP models.

3.2. Self-similar traffic

Self-similar (or fractal) stochastic processes have been proposed as accurate models of certain categories of traffic in high-bandwidth communications networks (e.g., LAN traffic, ABR traffic, WAN traffic). Studies of LAN traffic [10] and WAN traffic [7] question the commonly assumed models of Poisson distributions and renewal processes. Were traffic to follow a Poisson or Markov arrival process, it would have a characteristic burst length that would be smoothed by averaging over a sufficiently long time scale. However, measurements of real traffic indicate significant variance (burstiness) on a wide range of time scales. Such traffic is described statistically using the concept of self-similarity. Self-similarity is a 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. The degree of self-similarity of a series can be expressed with 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 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 $\mathrm{f}(t) = \alpha k^{\theta} t^{-\theta-1}, \theta, k > 0, t \geq k,$ and its cumulative distribution function (CDF) is $\mathrm{F}(t) = P[\tau \leq t] = 1 - (k/t)^{-\theta}$. The parameter k represents the smallest possible value of the random variable. Pareto distributions have been used for WWW file transfers in the analyses by ETSI study groups of proposed 3G data traffic. Pareto distributions have a number of properties that are qualitatively different from those more commonly used, e.g., 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. As θ decreases, an arbitrarily large portion of the probability mass may be present in the tail of the distribution.

Approximate distributions for self-similar point processes. Applying 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 [11], the superposition of two-state Markov sources [12], the mixture of exponentials to fit the heavy-tail distributions, the superposition of N "on-off" processes with sub-exponential "on" periods [13], deterministic chaotic-maps, and self-similar (fractal) point processes.[14] In each of these models, it is 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 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.[15] The new model decomposes the self-similar process in a way that is tractable for characterization and control of packet traffic. 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}. The probability density function of inter-arrival times for the corresponding point process is

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

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. The potential problem of LRD can be avoided by observing WIN 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.[16] 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).[16] Consequently, *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. *Truncated* forms of standard heavy-tailed CDFs are thus sufficient for the development of accurate models.

One such example for the inter-arrival time distribution is a truncated Pareto $\mathrm{F}_{\mathrm{T}_C}(t)$ is a truncated Pareto PDF defined by

$$\mathrm{F}_{\mathrm{T}_C}(t) = \begin{cases} \left(\dfrac{t+k}{k}\right)^{-\theta} & \text{, if } 0 < t < \mathrm{T}_C < \infty \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where $1 < \theta < 2$ and the parameter T_C is referred to as the *cutoff lag*. Truncated versions of (3) can also be defined. The cutoff lag eliminates correlation in the input process beyond a lag equal to T_C.

3.3. Class of admissible control policies: admission, handoff, switching and routing

The extent of control of the WIN packet flows is limited to CAC, handoff, switching and routing among network nodes. Handoff of a call from BS *i* to BS *j* is modeled by a random vector $\boldsymbol{u}_{ij,t} = (u_{ij,1,t}, u_{ij,2,t}, \ldots, u_{ij,S,t})$. The *s*'th component of $\boldsymbol{u}_{ij,t}$ is the indicator of the effect of the handoff event 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. Call admission control (CAC) is modeled by vector $\boldsymbol{u}_{0j,t} = (u_{0j,1,t}, u_{0j,2,t}, \ldots, u_{0j,S,t})$, *i*=0, that regulates new connections of calls to BS *j* incoming from gateways of the wired packet-switched infrastructure. The components of $\boldsymbol{u}_{ij,t}$ are (F_t)-predictable (more strongly, left-continuous) indicator functions of random events and Borel measurable. As such, expectations of the $\boldsymbol{u}_{ij,t}$ with respect to $\wp$ and Ω, are the probabilities of call access, blocking or handoffs that influence the flow of services through the WIN.

The $(N + M_{\max} + 1) \times (N + M_{\max} + 1) \times S$ time-varying control array,

$$\boldsymbol{U}_t = (u_{ij,s,t}; i,j = 0, \boldsymbol{1}, \ldots, N + M_{\max}; s = 1, \ldots, S),\ t \in [0, T]$$

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. It is also assumed that controls are closed under concatenation in time, that is, *U*=[*U*1,*U*2,*t*], $t \in [0, T]$, is also a control if U_1 and U_2 are controls. Controls are also assumed to satisfy a stochastic causality property. The set of control arrays U_t, $t \in [0, T]$, defined by the above conditions, are referred to the *class of admissible control policies,* $\mathcal{U}$..

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, can also modeled by the variables, $u_{ijk,s,t}$, adding a third index to designate the HCN layer.

3.4. Network state

A candidate state process of the WIN requires sufficient dimensionality 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 semi-martingale representation in terms of a discrete-valued jump, right-continuous, zero-mean, $(F_t, \wp)$-local ($(F_t, \wp^U)$-local) martingale and an integrated conditional random rate process with respect to the family of σ-algebras $\left(F_t, t \in [0,T]; F_t \subseteq F\right)^2$, generated by evolution of observed network events to time t, i.e.,

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

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

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

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, the observed network history to time t on which the conditional rates are estimated, and the admissible control policy $U \in \mathcal{U}$.. Indicator functions in (5) and (6) show that uninterrupted packet processing cannot occur when no packets of type s are in the node. For the special 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 $\tau^s_{i,n}$ and $\tau^s_{i,n+1}$, i.e., the event $\{\tau^s_{i,n} \le t \le \tau^s_{i,n+1}\}$, for each service type s.

At the fixed nodes of the WIN, multiple parallel processors typically handle at most L simultaneous calls, each of which may carry up to S active services. Therefore, at most L×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

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

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 control terms $u_{ij,s,t-}$ in (7) and (8) 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 $\left(F_t, t \in [0,T]; F_t \subseteq F\right)$ 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. Thus, it is natural to partition each admissible control array, $U=[U_C,U_{NC}]$, into two components, corresponding to the controlled and uncontrolled network events that influence connectivity.

3.5. QoS and radio resource constraints

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 $\left(u_{ij,s,t}\right)$. 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*.

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. Voice activity monitoring (VAM) can be modeled either by the indicator of a gain condition, $1\left(G_{\pi,ij,t} \geq 1/act\%\right)$, where G_v is the voice activity gain, or by the indicator of the random event of on-off voice activity, $1\left(\kappa_{ij,1,t} > 0\right)$, 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 entry $u_{ij,s,t}$ is the product of factors that include $1\left(\mathrm{BER}_s \leq 10^{-n}\right)$ for service type *s*. This factor can, in turn, be factored into a product of indicators of the range of RR parameter values that together comprise QoS for type *s*. For example, $1\left(BER_t \geq 10^{-n}\right) = 1\left(P_{i,t} \geq p_i\right) \cdot 1\left(G_{\text{coding},ij,t} \geq 10^{y/10}\right) \cdot 1\left(G_{A,j,t} \geq g\right) \cdot 1\left(P_{\text{avg. path loss},ij,t} \leq \pi\right) \;\; 1\left(I_{\text{co-ch.interf.},ij,t} \leq \varsigma\right)$. 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. Mechanisms of dropping and blocking due to resource constraints are represented in the model by the $u_{ij,s,\cdot}$, which contain terms of the general form $1\left(\text{Available}\, RR_{j,t} \geq \rho_s\right)$, where $RR_{j,t}$ is the radio resource at node *j* at time *t* required at level ρ_s to maintain QoS for type *s*.

4. CONSTRUCTION OF CONTROLLED PROBABILITY MEASURES

A family of probability measures $\wp^U$ on the network events Ω is constructed from a reference measure and the class of admissible control policies $\mathcal{U}$ defined above. The role of the Radon-Nikodym derivatives or likelihood ratios is fundamental to the construction. An absolutely continuous change of measure is given in terms of the local description of the MVPPs that define the network, that is, $dN_{ijs,t}\, d\wp \rightarrow dN_{ijs,t}\, d\wp^U$, that is, the change of conditional rates $p_{ojs}\alpha_t \rightarrow u_{0js,t}\alpha_{s,t}$ and $p_{ijs}\sigma_{is,t} 1(Q_{is,t-} > 0) \rightarrow u_{ijs,t}\sigma_{is,t} 1(Q_{is,t-} > 0)$ for $U \in \mathcal{U}$.

The following theorem, repeated without proof from [4], is a variation of the result by Doleans-Dade in [17] applied to the MVPPs constructed for the WIN.

Theorem 1. Let $\left(\breve{N}^A_{0js,t}, \breve{N}^D_{ijs,t}\right)$, for $\mathrm{i} = 0,1,\ldots,N+M_{\max}$; $j = 1,\ldots,N+M_{\max}$; $s = 1,\ldots,S$ be the counting processes defined for packet arrivals and service completions, adapted to the network history $\left(F_t, t \in [0,T]; F_t \subseteq F\right)$, and let

$p_{ojs}\alpha_t$, $p_{ijs}\sigma_{js,t}1(Q_{j,t-} > 0)$ for $\mathrm{i} = 0,1,\ldots,N + M_{\max}$; $j = 1,\ldots,N + M_{\max}$; $s = 1,\ldots,S$ be the $(\wp,F_t)$-predictable conditional rates of $\breve{N}^A_{0js,t}, \breve{N}^D_{ijs,t}$, respectively, where the constants p_{ijs} satisfy $0 < p_{ijs} \le 1$ and $\sum_{\mathrm{j=1}}^{\mathrm{M_{max}}} p_{ijs} \le \mathrm{M}_{\max}$ for $i = 0,1,\ldots,\mathrm{M}_{\max}$; $s = 1,\ldots,S$. Let $u_{ijs,t} / p_{ijs}$ for $\mathrm{i} = 0,1,\ldots,N + M_{\max}$; $j = 1,\ldots,N + M_{\max}$; $s = 1,\ldots,S$ for $(\wp,F_t)$-predictable processes, such that for all $t\in [0,T]$ and all indices i, j, s:

$$\int_0^t \left(u_{0js,t} / p_{0js}\right) p_{0js}\alpha_{js,v}\,dv < +\infty, i = 0 \tag{9}$$

$$\int_0^t \left(u_{ijs,t} / p_{ijs}\right) p_{ijs}\sigma_{js,v}\,dv < +\infty, i \ge 1 \tag{10}$$

Define the processes L^U_t by

$$L^U_t = \prod_i \prod_j \prod_s L^U_{ijs,t}, \tag{11}$$

where

$$L^U_{0js,t} = \left[\prod_{n\ge1} \int_0^t \left(u_{0js,\tau_{0js,n}} / p_{0js}\right) 1\left(\tau_{0js,n} \le t\right) \alpha_{s,v}\,dv\right] \cdot \exp\left(\int_0^t \left(p_{0js} - u_{ojs,v}\right) \alpha_{s,v}\,dv\right), \mathrm{i} = 0, \tag{12}$$

$$L^U_{ijs,t} = \left[\prod_{n\ge1} \int_0^t \left(u_{ijs,\tau_{ijs,n}} / p_{ijs}\right) 1\left(\tau_{ijs,n} \le t\right) \sigma_{is,v}\,dv\right] \cdot \exp\left(\int_0^t \left(p_{ijs} - u_{ijs,v}\right) 1\left(Q_{is,v-} > 0\right) \sigma i_{s,v}\,dv\right), \mathrm{i} \ge 1, \tag{13}$$

where the sequences $\{\tau_{ijs,n}, n\ge1\}$ refers to the sequence of F_t-stopping times associated with the jump transitions in the MVPPs $\left(\breve{N}^A_{0js,t}, \breve{N}^D_{ijs,t}\right)$. Then L^U_t is a nonnegative, local $(\wp,F_t)$-martingale and a $(\wp,F_t)$-supermartingale. Furthermore, for each i, j, s, the factor process $L^U_{ijs,t}$ is a nonnegative, local $(\wp,F_t)$-martingale and a $(\wp,F_t)$-supermartingale.

The next result, stated without the proof found in [4], formalizes the effect of the change of measure on the local description of the MVPPs underlying the network state models.

Theorem 2. Let the conditions of Theorem 1 hold. Furthermore, suppose that $E\left[L^U_t\right] = 1$ for each admissible control policy $U\in\mathcal{U}$.. Define the probability measure $\wp^U$ by $\frac{d\wp^U}{d\wp} = L^U_t$. Then each $\breve{N}^A_{0js,t}$ has the conditional $(\wp^U,F_t)$-rate $u_{0js,t}\alpha_{s,t}, i = 0$; each $\breve{N}^D_{ijs,t}$, for $i = 0,1,\ldots,\mathrm{M}_{\max}$; $j = 1,\ldots,\mathrm{M}_{\max}$; $s = 1,\ldots,S$ over the observation interval $[0,T]$.

5. THE OPTIMIZATION PROBLEM FOR THE CONTROLLED NETWORK

The optimal control problem is formulated in terms of performance metrics on WIN performance for admissible $U\in\mathcal{U}$., leading to optimization of cost functionals. Optimality conditions for the "best" policies $U\in\mathcal{U}$, given the observation σ-algebras on the network events, take the form of generalized backward recursive relations.

5.1. Performance metrics

Throughput and Capacity. The primary performance metric is network throughput, the time average of the number of packets for each service type delivered from the source to the designated terminal per unit of time. Other similar measures of interest for the WIN include the probability distribution of the number of transmissions from each network node, the time average of packet delay, the fraction of channel capacity used for successful transmission, the probability of successful packet transmission.

From the counting processes defined in Section 3, the number of packets of service type s, $s = 1,\ldots,S$, is the given as the sum of the individual components in the queueing state $Q_{,t}$, that is, $Q_{\mathrm{total},s,t} = \sum_{i=1}^{\mathrm{N+M_{max}}} Q_{i,s,t}$. Under the assumptions stated in Sections 3 and 4, $Q_{\mathrm{total},\,s}$ is a right-continuous, $(\wp,F_t)$-supermartingale and a non-explosive MVPP with a conditional random rate that is the sum over i of the rates of the $Q_{i,s,t}$ given in (5).

The *link throughput of service type s* over the subinterval $(v,t] \subset [0,T]$ for any $v<t$ from node i to node j is given by $\tilde{N}_{ijs,w} = N^{C}_{ijs,w} - N^{NC}_{ijs,w}$, that is, the difference between controlled and uncontrolled or lost packets of type s transported between the nodes over the interval. Similarly, the *node i throughput of type s* over $(v,t] \subset [0,T]$ is denoted as $\mathrm{L}_{is,t} - \mathrm{L}_{is,v}$, where L_{is} is defined for $i = 1,\ldots,N + M_{\max}; s = 1,\ldots,S$ and $v \in [0,T]$ by $\mathrm{L}_{is,v} = \sum_{j=1}^{\mathrm{N+M_{max}}} \tilde{N}_{ijs,v}$. Lastly, the *system throughput of service type s* over $(v,t]$ is given by $\mathrm{L}_{s,t} - \mathrm{L}_{s,v}$, where L_s is defined for all $v \in [0,T]$ by $L_{s,v} = \sum_{i=1}^{\mathrm{N+M_{max}}} \tilde{N}_{i0,v}$, where it is assumed that once a packet of service type s is successfully reaches its target node, it cannot be returned during the call.

The average throughput rates corresponding to the stochastic throughput processes defined above are formed by taking the expectation with respect to the probability measure $\wp$, or the controlled probability measure $\wp^{U}$, divided by the length of the subinterval $t - v$. Note that the expressions for the throughput metrics are merely linear combinations of the MVPPs $\left(\tilde{N}^{A}_{0js,t}, \tilde{N}^{D}_{ijs,t}\right)$ and so share with them the same ($\wp$,F_t)-semi-martingale structure described in Sections 3 and 4.

Define $U_C([0,t])$ as the set of values of all admissible control arrays U_C with entries that satisfy the conditions in Theorem 1 over $[0.t]$. The *mean link capacity from node i to node j* at time $t \in [0,T]$ is defined as $K_{ij,t} = \max_{U_C([0,t])} E^{U}\left[\tilde{N}_{ijs,t}\right]$, while the *mean link capacity at node i* at time t is defined as $K_{i,t} = \max_{U_C([0,t])} E^{U}\left[\mathrm{L}_{is,t}\right]$, where $E^{U}[\cdot]$ represents integration with respect to $\wp^{U}$ so that the $N_{ijs,\cdot}$ admit ($\wp^{U}$, F_t)-rates that depend on the values $u_{ijs,t}(\omega), \omega \in \Omega$. in $U_C([0,t])$.

Call Dropping and Blocking. Metrics of the expected number of blocked calls, handoffs, or dropped calls, either for a given service type or all service types are again the expectations of the counting processes for the corresponding network events over the observation period, summed over the indices of interest. Based on the semi-martingale representations of the underlying MVPPs and an assumption that rates of arrivals and service completions are non-explosive, the Fubini Theorem is 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 $(F_t, \wp)$-semi-martingale

$$N_{\text{blocked},j,s,[0,T]} = \left[N_{\text{blocked},\,j,s,[0,T]} - \int_0^T (1 - u_{0j,s,v}) 1(Q_{j,s,v-} \geq q_{j,s}) \alpha_{s,v} dv\right] + \int_0^T (1 - u_{0j,s,v}) 1(Q_{j,s,v-} \geq q_{j,s}) \alpha_{s,v} dv, \tag{14}$$

with expected value

$$E\left[N_{\text{blocked},j,s,[0,T]}\right] = E\left[\int_0^T (1 - u_{0j,s,v}) 1(Q_{j,s,v-} \geq q_{j,s}) \alpha_{s,v} dv\right] = \int_0^T \left(P(Q_{j,s,v-} \geq q_{j,s}) - P\left(u_{0j,s,v}(Q_{j,s,v-} \geq q_{j,s})\right)\right) \overline{\alpha}_{s,v} dv \tag{15}$$

where the second term in the integrand on the right-hand side of (15) 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 (15) can be generalized to represent the average number of handoffs or dropped calls for any or all service types, as the expectations of the indicators 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 integrands of the average number of the corresponding network events, at some time $t \in [0,T]$.

Prioritization at a single-processor node can be modeled by re-ordering the service times $\left(\tau^{s}_{i,n}, s = 1,\ldots,S\right)$ 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 $\tau^{s*}_{i,n*}$ 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 allows adaptation of the service rates $(\sigma_{i,s,t}, s = 1,\ldots,S)$, up to a maximum allowable total rate, i.e., $\sum_{s=1}^{S} \sigma_{i,s,t} \leq \Lambda_{i,t}$, 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.

Delay. 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 $\sigma^2_{\Delta,s}$, respectively, as part of the QoS for service type *s*. The average allowable delay for type s is denoted $\bar{\Delta}_s$. Other sources of delay 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 assumptions regarding the arrival distribution or service distribution; nor does it depend upon the number of servers in the system or upon the 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 $[0,T)$,

$$1\left(Q_{i,s,t} < \Delta_s\left[u_{0i,s,t}\alpha_{s,t} + \sum_{j\in\{\text{location area of node } i\}} u_{ji,s,t}\sigma_{j,s,t}\right]\right), \tag{16}$$

$$1\left(\int_0^T Q_{i,s,\tau}d\tau < \bar{\Delta}_s\left[\int_0^T u_{0i,s,v}\alpha_{s,v}dv + \sum_{j\in\{\text{location area of node } i\}} \int_0^T u_{ji,s,w}\sigma_{j,s,w}dw\right]\right). \tag{17}$$

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

$$1\left(\left(Q_{i,s,t} - E\left[Q_{i,s,t}\right]\right)^2 < \sigma^2_{\Delta,s}\left[\left(u_{0i,s,t}\alpha_{s,t} + \sum_{j\in\{\text{location area of node } i\}} u_{ji,s,t}\sigma_{j,s,t} - E\left[u_{0i,s,t}\alpha_{s,t} + \sum_{j\in\{\text{location area of node } i\}} u_{ji,s,t}\sigma_{j,s,t}\right]\right)^2\right]\right) \tag{18}$$

where expectation can be taken with respect to the probability measure $\wp$ or the controlled probability measure $\wp^U$.

5.2. Optimal control criterion

Each performance metric described in the preceding can be the expressed as a general cost functional to be optimized over admissible control policies $U=[U_C,U_{NC}]\in\mathcal{U}$.. To each admissible control array *U*, there is a unique cost functional of form:

$$C(U) = E^U\left[\int_0^T c(t,U_t)dt + f_T\right] \tag{19}$$

The terms in (19) are assumed to obey the following conditions. For each $U\in\mathcal{U}$., the instantaneous cost $(c(t,U_t), t\in[0,T])$ is a composite process $c(t,U_t(\omega),\omega) = (c\cdot U)(t,\omega)$, where c is (F_t)-adapted for each value of $U_t(\omega)\in \boldsymbol{P}$, the space of values of the fixed routing arrays. The function *c* is Lebesgue measurable with respect to *t* and continuous in the sample path values $U_t(\omega)$ for all $\omega\in\Omega$; while *c* itself has left-continuous sample paths with finite right-hand limits at each discontinuity for each $U_t(\omega)$ for all $\omega\in\Omega$. The process *c* is thus progressively measurable with respect to the family of σ-algebras, $(F_t, t\in[0,T]; F_t\subseteq F)$, with left-continuous sample paths for each $U\in\mathcal{U}$.. The terminal cost f_T, a nonnegative, *F*-measurable and $\wp^U$-integrable function for each $U\in\mathcal{U}$, represents the cost incurred at or after the end of the network operation at *t=T*.

The optimal control problem for the WIN network is the determination of control policies is $U^*\in\mathcal{U}$. over the observation interval $[0,T]$ that satisfies $C(U^*) = \inf_{U\in U} C(U)$, where the minimum instead of the infimum can be taken, assuming the $U\in\mathcal{U}$.. take values in a compact set and the instantaneous and terminal costs are a..s. $\wp^U$.bounded for $U\in\mathcal{U}$.. A control policy *U**, if it exists, that satisfies the criterion is called an *optimal control policy*.

5.3. Recursive optimality conditions: complete network observations

Recursive optimality conditions that characterize optimal policies for the controlled real-time MVPP models of the network have been developed.[4] For a more tractable exposition, the results are presented without proof for the case of complete observations of network history. The local structure of the optimality conditions resemble the Hamilton-Jacobi-Bellman dynamic programming conditions. With complete observations, the *conditional cost function* $\phi(U_1,U_2,t)$ for the admissible control policy concatenated at time t from U_1 and U_2 in $\mathcal{U}$., obeys $\phi(U_1,U_2,t)=E^{[U_1,U_2,t]}\left[\int_t^T c(v,U_{2,v})dv+f_T \middle| F_t\right]=E^{U_2}\left[\int_t^T c(v,U_{2,v})dv+f_T \middle| F_t\right]=\phi(U_2,U_2,t)$, based on the causality conditions imposed on admissible control policies $U\in\mathcal{U}$.. Hence, the *optimal cost-to-go function* $W_t^U=\inf\limits_{\hat{U}\in \mathrm{U}}\phi(U,\hat{U},t)=\inf\limits_{\hat{U}\in \mathrm{U}}\phi(\hat{U},\hat{U},t)$, that is, $W_t^U=W_t$ does not depend on the control policy U.

Suppose the network state is an array-valued process, $(X_t, t\in[0,T])$, formed from the network MVPPs. The next theorem introduces functions w_n, generalizing the optimal cost-to-go functions used in dynamic programming conditions of optimality. The following result summarizes a protracted analytical development in [4] of local optimality conditions for the optimal control of general packet-switched radio networks, for cases of partial and complete observations of the state process.

Theorem 3. Suppose the control policies have complete observations of the state $(X_t, t\in[0,T])$and, for every $U\in\mathcal{U}$., the state has a local description in terms of the conditional ($\wp^U$, F_t)-rates. Then $U=U^*$ is optimal in $\mathcal{U}$. if and only if there exist functions, $w_n(t,t_0,x_0,\ldots,t_n,x_n)$, measurable in their arguments and absolutely continuous in t, such that, for $K=N+M_{\max}$, τ_n the n-th stopping or transition time of $\boldsymbol{X}$, and e_{jks} is the $K+1\times K+1\times S$ array with all zero entries except 1 in the i, j, s position,

$$\frac{\partial w_n(t,\tau_0,\boldsymbol{X}_0,\ldots,\tau_n,\boldsymbol{X}_n)}{\partial t}+\min_{U\in \mathrm{U}}\Bigg\{\sum_s\sum_{l=1}^{K}\left[\alpha_{s,t}(u_{0ls,t})\cdot(w_{n+1}(t,\tau_0,\boldsymbol{X}_0,\ldots,\tau_n,\boldsymbol{X}_n+e_{0ls})-w_n(t,\tau_0,\boldsymbol{X}_0,\ldots,\tau_n,\boldsymbol{X}_n))\right]+$$

$$\sum_s\sum_{j=1}^{K}\sigma_{js,t}1(Q_{js,t-}>0)\left[\sum_{k=1}^{K}(u_{jks,t})\cdot(w_{n+1}(t,\tau_0,\boldsymbol{X}_0,\ldots,\tau_n,\boldsymbol{X}_n+e_{jks})-w_n(t,\tau_0,\boldsymbol{X}_0,\ldots,\tau_n,\boldsymbol{X}_n))\right]+$$

$$\sum_s\sum_{m=1}^{K}\sigma_{ms,t}1(Q_{ms,t-}>0)\cdot(u_{m0s,t})\cdot(w_{n+1}(t,\tau_0,\boldsymbol{X}_0,\ldots,\tau_n,\boldsymbol{X}_n+e_{m0s})-w_n(t,\tau_0,\boldsymbol{X}_0,\ldots,\tau_n,\boldsymbol{X}_n))+c(t,U_t)\Bigg\}=0 \tag{20}$$

for $\tau_n\le t<\tau_{n+1,}$ and

$$w_n(t,\tau_0,\boldsymbol{X}_0,\ldots,\tau_n,\boldsymbol{X}_n)=f_T \text{ for } \tau_{\mathrm{n}}\le T<\tau_{n+1}\,. \tag{21}$$

The minimum in (20) is attained at optimal control policies $U_t^*(\omega)$ a.s. $\wp^{U*}$. Furthermore, for the optimal cost-to-go process at time $t\in[0,T]$ takes the form

$$W_t=\sum_{n=0}^{\infty}1(\tau_n\le t<\tau_{n+1})\cdot w_n(t,\tau_0,\boldsymbol{X}_0,\ldots,\tau_n,\boldsymbol{X}_n). \tag{22}$$

5.4. Optimality conditions with Markov assumptions

The optimality conditions in (20)-(22) are greatly simplified when the control policies U depend only on the last observation of the network state to time t, i.e., $U_t(\omega)=U(t,\boldsymbol{X}_{t-}(\omega))$ for each $\omega\in\Gamma$, $\Gamma\in F_t$ and the assumptions for the MVPPs underlying the network state are such that $\boldsymbol{X}$ is a Markov process. This is the special case of Markov control of a Markov process. Although, as explained earlier, these assumptions do not accurately represent the dynamics of the WIN packet flows and interdependence of events occurring at the nodes, they do allow conditions (20)-(22) to be expressed in terms of the *optimal cost-to-go function* $W_t=V_t=V_t(\boldsymbol{X}_{t-})$to realize dynamic programming conditions where the infinitesimal generator for V_t depends on the conditional ($\wp^{U*}$, F_t)-rates for MVPPs given in (6)-(8).

6. PARTIALLY OBSERVED, CONTROLLED 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. 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.

The control approach for the MVPPs extends to the case of partial observations. The MVPPs 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 $t \leq v, t, v \in [0,T]$ and $F_t = \sigma\{\omega : (U_v(\omega), \boldsymbol{Q}_v(\omega)), 0 \leq v \leq t\}$, generated by the augmented network state. Partial or incomplete state 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 $G_t = \sigma\{\omega : \boldsymbol{X}_v(\omega), 0 \leq v \leq t\}$, an internal history generated by a 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 local observations at time $t \in [0,T]$ may not be disjoint, that is, $G_{ij,l,s,t} \cap G_{i^* j^*,l^*,s^*,,t} \neq \emptyset$, $(i, j, l, s) \neq (i^*, j^*, l^*, s^*)$, for nodes within a common active set or common location area. Note that each local σ-subalgebra of partial observations, $G_{ij,l,s,t} = \sigma\{\omega : \boldsymbol{X}_{ij,l,s,v}(\omega), 0 \leq v \leq t\}$, can be the internal history of some process defined in terms of local events. In notation, $G_t = \bigcup_{i=0}^{M}\bigcup_{j=0}^{N}\bigcup_{l=1}^{3}\bigcup_{s=1}^{S} \sigma\{\omega : \boldsymbol{X}_{ij,l,s,v}(\omega), 0 \leq v \leq t\}$ 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 work of Brémaud and others.[2, 3] 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, 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.[2] 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).

Alternatively, partial temporal information is modeled by sequences of increasing, right-continuous (F_t)– stopping times (ξ_n), termed the observation instants. The observation instants can be random or deterministic. The partial temporal observation subalgebras $(G_t, t \in [0,T])$, are defined as the stopped σ-algebras $G_t = F_{t \wedge \xi_n}$ for $t \in [0,T]$ and $\xi_{n-1} \leq t < \xi_n$. G_t is thus the history generated by the network state process observe only at the instants ξ_n.

To apply the theorem to models of controlled network processes, the semi-martingale representations of state are recast in terms of new state variables and the admissible control policies, $Z_{i,s,t}(n) = 1(Q_{i,s,t} = n)$ 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 processes, discrete- or continuous-space, that generate the internal history of partial state observations, (G_t). An explicit form of the G_t-filter consists of the G_t-innovations processes

$$\hat{M}^A_{i,s,t} = N^A_{i,s,t} - \int_0^t \hat{\alpha}^X_{i,s,v}\,dv \text{ and } \hat{M}^D_{i,s,t} = N^D_{i,s,t} - \int_0^t E\left[\sigma_{i,s,v}\left(\sum_{j \neq i} u_{ij,s,v}\right) 1(Q_{i,s,v-} > 0) \middle| G_v^X\right] dv \tag{23}$$

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 $n \in \mathbf{Z}_+$,

$$\begin{aligned}\hat{Z}^X_{i,s,t}(n) = P[Q_{i,s,0} = n] + \int_0^t \hat{f}^X_{i,s,v}(n)dv + \int_0^t K^A_{i,s,v}(n)\left(dN^A_{i,s,v} - \hat{\alpha}^X_{i,s,v}\,dv\right) \\ + \int_0^t K^D_{i,s,v}(n)\left(dN^D_{i,s,v} - \hat{E}\left[\sigma_{i,s,v}\left(\sum_{j \neq i} u_{ij,s,v}\right) 1(Q_{i,s,v-} > 0) \middle| G_v^X\right]\right) dv,\end{aligned} \tag{24}$$

where

$$f_{i,s,t}(n) = \left(Z_{i,s,t}(n-1)1(n>0) - Z_{i,s,t}(n)\right)\alpha_{i,s,t} + \left(Z_{i,s,t}(n+1) - Z_{i,s,t}(n)1(n.>0)\sigma_{i,s,t}\left(\sum_{j\neq i} u_{ij,s,t} \right)\right). \tag{25}$$

The random rates in the second and third terms on the right-hand side of (24) are predictable versions of the $\left(G_t^X\right)$-rates of $N^A_{i,s,t}$ and $N^D_{i,s,t}$, respectively. In order to achieve a more recursive structure for the filter of the controlled process, equations for the innovation gains in (25) must be solved in terms of the MVPPs, control policies, and other processes that generate the state and partial observations. Applying result R6 in [2] to the filtration of $Q_{i,s,t}$, the innovation gains in (24) 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 based on the point processes that generate the partial observations,

$$K^A_{i,s,t}(\boldsymbol{n}) = -\hat{Z}^A_{i,x,t-}(n_i) + \frac{u_{ii,s,t-}\hat{\sigma}^A_{i,x,t-}1(n_i>0)\hat{Z}^A_{i,x,t-}(n_i) + \left(u_{0i,s,t-}\hat{\alpha}_{s,t-} + \sum_{j\in\{\text{BSs in the active set of MS}\,i\},\, i\neq j} u_{ji.s.t-}\hat{\sigma}^A_{j,x,t-}1(n_j>0)\hat{Z}^A_{j,x,t-}(n_j) \right) 1(n_i>0)\hat{Z}^A_{i,x,t-}(n_i-1)}{\left(u_{0i,s,t-}\hat{\alpha}_{s,t-} + \sum_{j\in\{\text{BSs in the active set of MS}\,i\},\, i\neq j} u_{ji.s.t-}\hat{\sigma}^A_{j,x,t-}1(n_j>0)\hat{Z}^A_{j,x,t-}(n_j) \right) + u_{ii,s,t-}\hat{\sigma}^A_{i,x,t-}\left(1-\hat{Z}^A_{i,x,t-}(0)\right)} \tag{26}$$

$$K^D_{i,s,t}(\boldsymbol{n}) = -\hat{Z}^D_{i,s,t-}(n_i) + \frac{u_{ik.s.t-}\hat{\sigma}^D_{i,x,t-}1(n_i>0)\hat{Z}^D_{i,s,t-}(n_i) + \left(\sum_{k\in\{\text{BSs in the active set of MS}\,i\},\, k\neq i} u_{ik.s.t-}\hat{\sigma}^D_{i,x,t-} \right)\hat{Z}^D_{i,s,t-}(n_i+1)}{\left(\sum_{k\in\{\text{BSs in the active set of MS}\,i\}} u_{ik.s.t-}\hat{\sigma}^D_{i,x,t-} \right)\left(1-\hat{Z}^D_{i,s,t-}(0)\right)} \tag{27}$$

The exact form of the set of filtration equations (23)-(24) may include terms corresponding to either one or both of the innovations gains (26) and (27), provided that either of the point processes $N^A_{i,s,t}$ or $N^D_{i,s,t}$ or both are generators of the partial observations algebra $\left(G_t^X\right)$, respectively. While this filtration reveals dependence on entries in the admissible control arrays and can be applied to recursively update estimates of selected WIN system processes, based on the evolution of observed network events, the controls cannot generally be "separated" from and solved in terms of the estimates of the network state. Thus, optimal control policies cannot be determined explicitly in terms of the network state estimates, derived from the filtration equations, as in Kalman filter applications to the optimization of linear-quadratic control systems. Restricting control policies in $U\in\mathcal{U}$ and terms in the cost functionals to depend only on the processes generating the internal history of partial observations to force separability of the optimal controls and state estimates may lead to "optimal" controls that cause significantly suboptimal performance in actual network implementations.

7. CONCLUSIONS

Control of the MVPP models for the packet flows of the integrated services among nodes of a wireless multimedia network has been formulated through an absolutely continuous change of probability measure on the random, real-time call processing events. The MVPP models are shown to encompass a wide range of statistical properties ascribed to the arrivals, service processing, handoffs, and routing of various types of integrated services in studies of wireless multimedia systems. The probability measures are constructed from a reference measure on network events via likelihood ratios that depend explicitly on the elements in admissible control arrays that influence call admission, handoff, switching and routing. In turn, the conditional random rates in the semi-martingale decompositions, with respect to family of network observations and the new controlled measures, of the point processes also depend on the admissible control entries. Based on the control approach by change of measure, the optimization problem is constructed to determine admissible control policies that minimize several network performance metrics. For the case of complete observations of network events, backward recursive optimality conditions have been given that characterize the admissible control policies that optimize the performance criterion. The impact of Markov controls and Markov network state on the optimality conditions is discussed. The filtration equations of the partially observed, controlled network processes have been developed. The issues created for the optimization problem based on the controlled MVPP models, when only partial state information is available, are discussed.

Further study is required to establish conditions on the network structure, performance criteria and controls that allow separation of the optimal control policies and the estimates of the network processes, given partial temporal and state observations of network events. In the case of partial temporal observations, examination should be made of the "best" sequence of observation instants to minimize the number of "lost" events, similar to the application of importance sampling techniques in real-time system simulations. The results of this research should reduce the extent of observations required to support the optimal policies in RLC, MAC and higher layers of wireless IP protocols. Another goal of future research is the development of efficient real-time control algorithms of multimedia streams based on specific 3G operating scenarios.

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 routing to enable IP multimedia services.

REFERENCES

1. W. S. Hortos, "Real-time performance analysis of wireless multimedia networks based on partially observed, multivariate point processes," *Digital Wireless Commun.II*, *Proc. SPIE*, 4045-05, Orlando, FL, April 2000.
2. P. Brémaud, *Point Processes and Queues*: *Martingale Dynamics*, Springer-Verlag, New York, 1981.
3. J. Walrand and P. Varaiya, "Flows in queueing networks: a martingale approach," *Math. Oper. Resrch.*, vol. 6, pp. 387-404, 1981.
4. 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.
5. 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.
6. V. S. Frost and B. Melamed, "Traffic modeling for telecommunication networks," *IEEE Comm. Mag.,* vol. 32, no.3, pp. 70-81, Mar. 1994.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. 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.
15. 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.
16. 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.
17. C. Doléans-Dade and P.A. Meyer, "Intégrales stochastique par rapport aux martingales locales," *Séminaire Probabilités, IV, Lecture Notes in Mathematics*, vol. 124, Springer-Verlag, Berlin, pp. 77-107, 1970.