EURASIP Journal on
Wireless Communications and Networking
a SpringerOpen Journal

RESEARCH Open Access

Call admission control for integrated multimedia service in heterogeneous mobile hotspots

Wei Song[1*], Peijian Ju[1] and Yu Cheng[2]

Abstract

Mobile hotspots are a promising trend to offer ubiquitous multimedia services even in public transit vehicles such as buses, trains, and airplanes. However, it is very challenging due to high mobility, fast channel fading, and stringent multimedia quality-of-service (QoS) constraints. Effective admission control is necessary to limit the admitted traffic so that accepted users are provided QoS guarantee. In this paper, we develop a comprehensive analytical framework for the performance of interactive data service and conversational video service in mobile hotspots with heterogeneous wireless technologies. We jointly consider the contention-based wireless local area network (WLAN) at the link layer, the highly varying wireless wide area network (WWAN) due to vehicle mobility and multipath fading, adaptive modulation and coding for the WWAN link at the physical layer, and batch packet arrivals of video traffic at the application layer. Based on the analytical approach, the maximum numbers of users are derived for QoS assurance. Simulation results verified the validity of the analysis. Numerical results demonstrated the effectiveness of the analytical approach for admission control and the effects of network parameters such as the traffic buffer size and the transmission distance.

Keywords: Mobile hotspots, Video QoS, Markov-modulated fluid source, Adaptive modulation, Coding

1 Introduction

The rapid advance and breakthrough of wireless technologies are well supporting the establishment of pervasive wireless infrastructure. Wireless local area networks (WLAN) are now widely deployed in hotspot areas, e.g., offices, airports, cafés, and hotels. The mainstream handheld wireless devices such as smartphones, laptops, and tablets are equipped with built-in WLAN interfaces. Internet access can be enabled by integrating the WLAN with wireless wide area networks (WWAN), such as the long-term evolution (LTE) cellular networks and IEEE 802.16 wireless metropolitan area networks, also known as WiMAX for worldwide interoperability for microwave access. The indoor hotspots usually serve slow-moving or static users in a residential or business environment. A natural evolution is to extend the hotspot service to moving vehicles such as buses, trains, and airplanes. Such a vehicular network is the so-called *mobile hotspot*, which consists of a group of end users that move as a whole in a public transit.

Although there is no unified network architecture for mobile hotspots, Figure 1 illustrates a typical architecture based on existing wireless technologies [1,2]. A key network entity is the access point (AP) mounted in the vehicle to form a WLAN among a collection of wireless user terminals. A dual-mode AP further acts as a mobile router and communicates with a base station (BS) through roof-top antenna(s). The BS is in turn connected with a fixed or wireless backbone, and the Internet via wireline links. In the industry, there are already successful deployments of mobile hotspots. Aircell (http://www.aircell.com) installs *Gogo Inflight System* for onboard Internet access. Gogo is a ground-to-air system using cellular technology evolution-data optimized (EV-DO) and IEEE 802.11 WLAN. The connection

*Correspondence: wsong@unb.ca
[1]Faculty of Computer Science, University of New Brunswick, Fredericton, Canada
Full list of author information is available at the end of the article

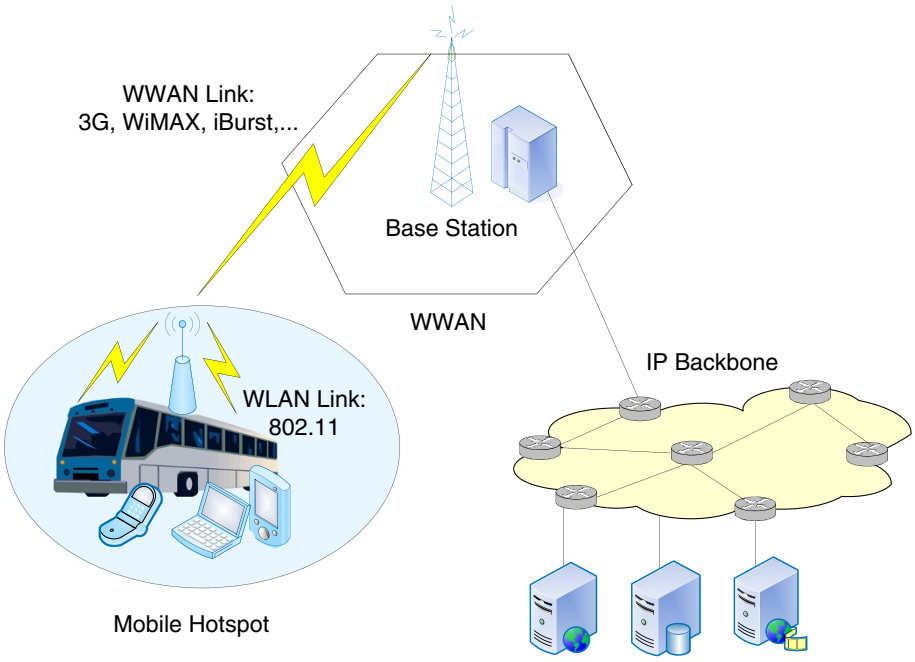

Figure 1 Mobile hotspots with two-hop relay over heterogeneous wireless links.

speed was approximately 500 to 600 kbit/s for downloads and 300 kbit/s for uploads [3]. Red Ball Internet (http://www.redballinternet.com) is another solution for high-rate data connectivity in mobile hotspots. It employs IEEE 802.20 broadband wireless access with high-capacity spatial division multiple access (HC-SDMA).

To offer multimedia services in mobile hotspots, many challenging issues need to be addressed due to high mobility and fast channel fading. A link adaptation scheme is developed in [4] for the WiMAX-based downlink of the mobile hotspot between the WiMAX BS and the AP mounted in the vehicle. Both the BS and AP are equipped with the multiple input and multiple output (MIMO) antenna system. The proposed adaptation algorithm selects an appropriate burst profile for the MIMO transmission mode and the modulation and coding settings to maximize downlink throughput. Another link-layer solution for mobile hotspots is discussed in [5]. It is based on a novel architecture called *information raining*, in which a number of repeaters are placed along the railway track and multiple antennas are installed on the roof of coaches. Downlink packets are decomposed into smaller fragments and relayed to the vehicle via multiple adjacent repeaters. A heuristic algorithm matches repeaters with antennas to optimize the throughput.

The network-layer routing and mobility issues are addressed in [6] for mobile hotspots. The authors evaluate two solutions, namely, the network mobility basic support protocol (NEMO) [7] and the session initiation protocol (SIP) [8]. As an application-layer solution, the SIP-based approach is observed to feature easy deployment but suffer from long handoff latency. The transport-layer throughput of TCP-friendly rate control is analyzed in [2] for mobile hotspots. The results reveal the effects of link bandwidth, retransmission limit, buffer size, vehicle velocity, and the number of users. A cross-layer approach is proposed in [9] for video transmission in a metro passenger information system based on IEEE 802.11p, also known as wireless access in vehicular environments (WAVE). Focusing on the train-ground link, the authors jointly consider video coding parameters and handoff decisions to minimize video distortion.

We notice that many previous studies [4,5,9] focus on the wireless section between the AP of the mobile hotspot and the BS. As shown in Figure 1, the two-hop relay can not only facilitate easy implementation but also benefit end users. Pack et al. address both the WLAN and WWAN links with heterogeneous characteristics in [1,2]. A Rayleigh fading channel is considered for the dedicated WWAN link. The error process at the packet level is characterized by a two-state Markov model. The WLAN downlink and uplink within the mobile hotspot are modeled as a discrete-time $M/M/1/K$ queue and a discrete-time $M/M/1$ queue, respectively. Although the performance analysis targets at multimedia streaming flows, the traffic characteristics under study are relatively simplified.

The statistics collected from Allot's worldwide mobile operator customers show that video streaming, file sharing, and Web browsing are three types of major applications for mobile data bandwidth usage [10]. Voice over IP (VoIP) and Instant Messaging (IM) have gained additional share and continue to be the fastest-growing application type. To provide high-quality multimedia services in the mobile hotspot, it is essential to limit the number of admitted users. Hence, we first develop a comprehensive analytical framework to evaluate the performance of integrated multimedia services in mobile hotspots. Our analysis focuses on both video and data services and takes into account the unique traffic characteristics of different services, such as on-off flow dynamics and burst arrival nature of video traffic. A closed queueing network with multiple classes is used to model multi-service flows and contention-based random access of the WLAN link. On the other hand, we assume that the WWAN link is subject to Nakagami fading and modeled by a finite-state Markov chain (FSMC) when adaptive modulation and coding (AMC) is used to address channel variation. Then we apply the analytical framework to determine the number of videos and data users that can be admitted to the mobile hotspot with acceptable quality of service (QoS). The effective numerical evaluation enables adaptive admission control with varying user traffic.

The remainder of this paper is organized as follows. In Section 2 and Section 3, we introduce the network model and traffic model for this study, respectively. To enable an effective admission control, an analytical framework is developed in Section 4. Numerical results are discussed in Section 5, followed by Conclusions in Section 6.

2 Channel models for two-hop heterogeneous mobile hotspots

In this work, we consider a mobile hotspot shown in Figure 1. The two-hop relay structure not only facilitates easy implementation but also takes advantage of the complementary strengths of the WLAN and WWAN technologies. On one hand, since end users within the moving vehicle are relatively stationary to the AP, the lightweight-designed WLAN is sufficient to provide a reliable high-rate transmission. On the other hand, the ubiquitous and large-area coverage of WWAN provides enhanced mobility support. Different from ordinary mobile stations, the AP can afford more complex design such as the multi-antenna system [4] because of the larger physical dimensions and less power constraints. As compared to the direct connection between end users and the BS, the two-hop relay can relieve the energy consumption for end users and share the powerful transmission facility of the AP [1]. Moreover, the multi-user traffic from a mobile hotspot can be multiplexed at the AP before relayed toward the BS. The multiplexing gain can be exploited to reduce resource occupancy.

2.1 Overview of mobile hotspot modeling

There are many previous studies that analyze the performance of mobile hotspots through simulations or measurements. These experimental results reveal the fundamental insights on the impact of various system configurations. Nonetheless, numerical evaluation is necessary to facilitate adaptation with dynamic traffic and network conditions. An effective analytical framework is proposed in [1,2] to evaluate the throughput and packet loss of mobile hotspots. In this paper, we further extend the performance analysis to multimedia services with unique traffic characteristics and apply more generic channel models for the two-hop heterogeneous links of mobile hotspots. Before introducing the details, we would briefly discuss the modeling rationale and present an outline of the combination of these models for performance analysis.

In Section 2.2, the WLAN channel is modeled with saturated data traffic and non-saturated video traffic [11]. The access delays of video and data packets are then captured by state-dependent geometric distributions. On the other hand, the WWAN link is assumed to experience Nakagami fading [12] and use AMC to address channel variation. In Section 2.3, a FSMC for the WWAN link is introduced for completeness [13]. In Section 3, we present the flow-level and packet-level models of multimedia traffic. As defined in Section 3.1, users are assumed to alternate between 'on' and 'off' phases. During the on phase, data users are saturated with packets to transmit. The packet-level video traffic is non-saturated and features a burst arrival structure as shown in Section 3.2. Furthermore, the multi-user traffic from the mobile hotspot can be aggregated at the AP before it is relayed toward the BS. The aggregate traffic is modeled by a Markov-modulated fluid [14] in Section 3.3.

Based on the system model of mobile hotspots in Section 3, we analytically evaluate the flow-level performance in terms of data response time and packet-level delay and loss probability in Section 4. In Section 4.1, the steady-state probabilities of multimedia flows and data response time are analyzed with a multi-class closed queueing network, which combines the on-off flow model and the WLAN channel model. The state-dependent packet service time of the WLAN channel are obtained to extend a discrete-time $D/G/1$ queueing system in Section 4.2. The batch arrivals of video packets are incorporated into the delay analysis. Finally, in Section 4.3, the performance of aggregate traffic over the WWAN channel is evaluated by the fluid approach, which combines the Markov-modulated fluid traffic model and the FSMC channel model.

2.2 Channel model of WLAN with contention-based random access

Taking advantage of the group mobility feature and extended capability of the vehicle-mounted AP, end users in the mobile hotspot can greatly benefit from the two-hop relay. The wireless link between the AP and user terminals is usually based on IEEE 802.11 WLAN. Here, we consider the distributed coordination function (DCF) for contention-based random access. Given that there are n_v active video flows and n_d active data flows, we use the analytical model in [11] to derive the average service time of video and data packets, denoted by $1/\phi_v(n_v, n_d)$ and $1/\phi_d(n_v, n_d)$, respectively. Here, the packet service time is defined as the time duration from the moment that a packet becomes the head of a node's sending queue to the instant that it finishes transmission. As data files are usually pre-stored at application servers, we assume saturated traffic for data service. In contrast, the packets from a video flow are non-saturated and arrive at a mean rate of λ_v.

When a video flow transmits in a slot, a collision happens if any other video or data flow transmits in the same slot. The collision probability p_v is given by

$$p_v = 1 - \left[1 - \frac{\lambda_v}{\phi_v}\tau_v\right]^{n_v - 1} (1 - \tau_d)^{n_d}, \tag{1}$$

where τ_v and τ_d are the transmission probabilities of a video flow and a data flow in a slot, respectively, which are given by Equations 6 and 7 in [11]. They depend on the initial backoff window, retransmission limit, and maximum backoff stage. Similarly, the collision probability of a data flow sending in a slot is

$$p_d = 1 - \left[1 - \frac{\lambda_v}{\phi_v}\tau_v\right]^{n_v} (1 - \tau_d)^{n_d - 1}. \tag{2}$$

In solving the four equations for p_v, p_d, τ_v, and τ_d, we obtain these four unknown variables. Let T_{sv} and T_{sd} denote the time durations of a successful packet transmission for a video flow and a data flow, respectively. The corresponding collision durations are denoted by T_{cv} and T_{cd}. Then the average packet service time, including the transmission time, backoff time, and collision time, is obtained as

$$\frac{1}{\phi_v(n_v, n_d)} = \left[(n_v - 1)\frac{\lambda_v}{\phi_v} + 1\right] T_{sv} + n_d T_{sd} + \overline{W}_v(p_v) + \frac{1}{n_c}\left[\left((n_v - 1)\frac{\lambda_v}{\phi_v} + 1\right)\overline{T}_{cv} + n_d \overline{T}_{cd}\right] \tag{3}$$

$$\frac{1}{\phi_d(n_v, n_d)} = n_v \frac{\lambda_v}{\phi_v} T_{sv} + n_d T_{sd} + \overline{W}_d(p_d) + \frac{1}{n_c}\left[n_v \frac{\lambda_v}{\phi_v}\overline{T}_{cv} + n_d \overline{T}_{cd}\right], \tag{4}$$

where n_c is the average number of video/data flows involved in a collision, $\overline{W}_v(p_v)$ is the average backoff time of a video flow as a function of p_v, $\overline{W}_d(p_d)$ is the average backoff time of a data flow as a function of p_d, $\overline{T}_{cv}$, and $\overline{T}_{cd}$ are the average collision time of a packet from a video flow and a data flow, respectively. Here, $\overline{T}_{cv}$ and $\overline{T}_{cd}$ are functions of T_{cv} and p_v, and T_{cd} and p_d, respectively. For detailed derivation of (3) and (4), interested readers can refer to Equations 8 and 9 of [11].

In [15,16], the authors demonstrate the near-memoryless behavior of the packet service time, which can be accurately approximated by an exponential distribution or a geometric distribution. To enable a discrete-time analysis, we use a small time unit τ_p to discretize the time scale. Then, we have the following probability mass function (PMF) of a geometric distribution to model the video packet service time $S_v(n_v, n_d)$ with n_v active video users and n_d active data users

$$\begin{aligned} &P\left[S_v(n_v, n_d) = k\right] = \left[1 - \nu_v(n_v, n_d)\right]^{k-1} \nu_v(n_v, n_d) \\ &\nu_v(n_v, n_d) = \phi_v(n_v, n_d)\tau_p, \quad k = 1, \ldots. \end{aligned} \tag{5}$$

Similarly, the data packet service time $S_d(n_v, n_d)$ is approximated by another geometric distribution with a mean $1/\nu_d(n_v, n_d) = [\phi_d(n_v, n_d)\tau_p]^{-1}$.

2.3 FSMC model for WWAN channel with Nakagami fading and AMC

The WWAN link between the AP and the BS is highly varying and fast fading due to the high velocity of the vehicle. According to the channel fading analysis based on empirical measurements in [12], we assume that the WWAN link is a non-line-of-sight multipath fading channel, which exhibits Nakagami flat fading with additive white Gaussian noise (AWGN) [17]. The received signal at time slot k is denoted by $y(k) = f(k) \cdot x(k) + n(k)$, where x is the transmitted signal, n is an AWGN noise, and f is the random fading amplitude. The marginal distribution of f follows a Nakagami distribution with a probability density function (PDF) given by [18]

$$p_f(\alpha) = \frac{2\alpha^{2m-1}}{\Gamma(m)}\left(\frac{m}{\Omega}\right)^m \exp\left(-\frac{m}{\Omega}\alpha^2\right), \quad \alpha > 0,\ \Omega > 0,\ m \geq 0.5 \tag{6}$$

where $\Gamma(\cdot)$ is the Gamma function, α is the fading amplitude, $\Omega = E[f^2]$ is the second moment or local mean power, and m is the parameter indicating fading severity, given by

$$m = \frac{E^2[f^2]}{\mathrm{Var}[f^2]} = \frac{\Omega^2}{E[(f^2 - \Omega)^2]}. \tag{7}$$

With $m = 1$, Equation 6 becomes the Rayleigh distribution. When $0.5 \leq m < 1$, the channel fading is more severe than Rayleigh fading. While $m > 1$, the fading is less severe than Rayleigh fading. When $m \to \infty$, the distribution becomes an impulse, which means there is no

fading at all. The received signal-to-noise ratio (SNR) over the Nakagami fading channel follows a gamma distribution with a PDF given by

$$p_s(\gamma) = \frac{\gamma^{m-1}}{\Gamma(m)} \left(\frac{m}{\overline{\gamma}}\right)^m \exp\left(-\frac{m}{\overline{\gamma}}\gamma\right), \qquad \gamma \geq 0, \quad (8)$$

where m is the shape parameter, $\overline{\gamma}/m$ is the scale parameter, and $\overline{\gamma}$ is the average received SNR which depends on the large-scale fading such as path loss. Considering a log-distance model (Chapter 3 in [19]), we have the average SNR (in dB) at a transmitter-receiver distance d as follows:

$$\overline{\gamma}(d) = \overline{\gamma}(d_0) + 10\kappa \log_{10}\left(\frac{d}{d_0}\right), \quad (9)$$

where $\overline{\gamma}(d_0)$ is the average SNR at a reference distance d_0, and κ is the path loss exponent.

To enhance transmission performance, AMC has been widely used to match transmission modes with time-varying channel conditions [18]. Given N transmission modes, the SNR range is partitioned into $(N+1)$ non-overlapping consecutive intervals, with the boundaries denoted by γ_n, $n = 0, 1, ...N+1$, where $\gamma_0 = 0$ and $\gamma_{N+1} = +\infty$. The transmission mode n is chosen when $\gamma_n \leq \gamma < \gamma_{n+1}$. As an example, Table 1 shows the profiles of five transmission modes, where β_n is the modulation and coding rate in terms of bits/symbol for mode n. To avoid deep fades, when $\gamma_0 \leq \gamma < \gamma_1$, mode 0 is assumed to have a transmission rate $\beta_0 = 0$.

Assume that the AMC is applied frame by frame over a constant duration τ_f and channel state transitions only happen between adjacent states. Then, the WWAN mobile channel is characterized by an FSMC model, which is defined by an $(N+1) \times (N+1)$ state transition probability matrix $\boldsymbol{P_c}$ [20]. The infinitesimal generating matrix of the corresponding continuous-time Markov process is given by $\boldsymbol{B_c} = (1/\tau_f)(\boldsymbol{P_c} - \boldsymbol{I_{N+1}})$, where $\boldsymbol{I_{N+1}}$ is an identity matrix. It is worth mentioning that motion results in Doppler frequency shift in the received signal, while Doppler shift in turn affects the level crossing rates for the FSMC model, which is a measure of the rapidity of the fading.

3 Flow-level and packet-level traffic models for multimedia services

3.1 On-off flow-level traffic dynamics

According to delay-sensitivity, popular mobile applications can be broadly categorized into four classes, namely, conversational, interactive, streaming, and background [21]. The background class is the most delay-tolerant and of the best-effort service nature. As the most delay-sensitive class, the conversational class is characterized by two-way conversational communication pattern and subject to strict delay bound of tens or hundreds of milliseconds. Voice telephony and video telephony are typical services of the conversational class. The interactive class comprises services of a request-response pattern, such as Web browsing, E-mail server access, and voice messaging. It features a variable call duration, which depends on the file size and available bandwidth. As a main performance criterion, the call duration is also known as the response time [22,23] to measure the service responsiveness, e.g., how fast a Web page is successfully downloaded and appears after it has been requested. Although the response time should be bounded to maintain fluent interactions, the delay requirement is far less stringent than that of the conversational class. Approximately, a transfer delay of 2 to 4 s is acceptable to most interactive services. The streaming class is meant for services having the content played back at the receiver during delivery. Streaming services are primarily unidirectional, such as video surveillance, movie clip streaming, and audio streaming. As the time relation between information entities of a streaming flow must be preserved, the transfer delay variation needs to be bounded. A playout buffer can be introduced to counter against traffic burstiness and absorb delay jitter resulting from network bandwidth variation.

As seen, there are particularly two major types of video services, i.e., conversational video (e.g., video telephony) and video streaming [24]. Conversational video is characterized by stringent end-to-end latency constraint and two-way traffic with a bursty pattern due to the use of live video encoder. In contrast, video streaming usually only involves with one-way downlink traffic. Video streams can be pre-stored at application servers and allow for a pre-rolling delay (normally less than 10 s) before the start of video playback. As such, video streaming applications are

Table 1 Modulation and coding settings of $(N+1)$ transmission modes

Mode	0	1	2	3	4	5
Modulation	Null	BPSK	QPSK	8-QAM	16-QAM	32-QAM
β_n (bits/symbol)	0	1	2	3	4	5
a_n	1	134.9972	129.6907	112.9110	87.9240	73.3388
g_n	0	0.9777	0.4046	0.1459	0.0478	0.0149
φ_n (dB)	0	9.7000	13.5000	18.0000	23.2000	28.5000

more concerned with playback smoothness, and the delay constraint is relaxed in some sense.

Our study focuses on interactive data service and conversational video service. Given a finite user population in the vehicle mobile hotspot, N_v and N_d denote the maximum numbers of video flows and data flows admitted in the hotspot, respectively. Suppose each user alternates between on and off phases [25,26] as illustrated in Figure 2. After the completion of a video or data flow, the user takes a random idle time before starting the next service request. The idle time is assumed to follow an exponential distribution with means of $1/\lambda_1$ and $1/\lambda_2$ for video and data users, respectively. The response time of data flows, i.e., the data flow duration, depends on the data file size and available bandwidth. In contrast, the video flow duration is independent of the transmission rate and subject to a strict constraint so as to preserve the intrinsic time relation of the flow. The video flow duration is usually in the order of minutes [27] and assumed to be exponentially distributed with a mean $1/\mu_1$.

3.2 Packet-level video traffic with batch arrivals

Although the response time of data flows needs to be bounded to ensure smooth interactions, data service can accept elastic bandwidth. Packet arrivals in the active period are assumed to be saturated. Suppose there is a fixed size L_d for data packets. The number of packets in a data flow (denoted by X_d) follows a geometric distribution with a mean of $1/\eta_d$.

On the other hand, video flows have a strict rate requirement to maintain the time relation of video packets. This is because video traffic is usually encoded and compressed over frames captured at fixed intervals. To remove temporal redundancy, intracoded (I) frames are interleaved with predicted (P) frames and bidirectionally coded (B) frames. As observed in [28], a hypothesized and independent distribution such as gamma, Erlang, and Weibull distribution can capture the statistics of video frame size, given that the fixed inter-arrival time and batch structure of video bursts are preserved. Assume an exponential distribution for the video flow duration and a fixed inter-arrival time of video bursts, which is denoted by $\tau_b = H\tau_p$ (H time units). The number of bursts in a video flow (denoted by X_v) then follows a geometric distribution. Thus, we have the mean active period of video flows $1/\mu_1 = \overline{X}_v\tau_b$. Given a fixed video packet size L_v, each video traffic burst is segmented into a batch of packets that arrive simultaneously. The number of packets in a batch is modeled by a negative binomial distribution (NB), which is a discrete analog of Gamma distribution. The PMF of the batch size (denoted by A) is given by

$$\mathrm{P}[A=k] = \binom{k+r-2}{k-1}(1-p)^r p^{k-1}, \quad r>0, \quad 0<p<1, \quad k=1,\ldots, \tag{10}$$

where the binomial coefficient

$$\binom{k+r-2}{k-1} = \frac{(k+r-2)(k+r-3)\ldots(r)}{(k-1)!}. \tag{11}$$

The parameters p and r can be obtained by fitting the mean and variance of the batch size:

$$\overline{A} = 1 + \frac{rp}{1-p}, \qquad \sigma_A^2 = \frac{rp}{(1-p)^2}. \tag{12}$$

To interpret the physical meaning of p and r, we can view the number of packets in a batch as the outcome of a sequence of independent Bernoulli trials. Given p as the probability of 'success' in each trial, the number of successes to observe r 'failures' follows an NB distribution in (10).

3.3 Markov-modulated fluid model for aggregate traffic

As shown in Figure 1, traffic flows from the mobile hotspot can be first aggregated at the AP before relayed toward the BS. The multiplexed traffic exhibits high correlation, due to on-off flow dynamics and long-range dependency of video traffic. As it is complex and expensive to apply service differentiation at the AP, we consider the traffic from the mobile hotspot as an aggregate flow and characterize it with a Markov-modulated fluid model in Figure 3. A histogram-based technique [29] can be used to measure the varying traffic rate of an aggregate flow over a fixed interval τ_a. According to predefined rate boundaries, the aggregate flow is classified into M states. The fluid rate (in packets/s) for state i is defined by the average rate of state-i traffic segments, denoted by R_i, $i = 0, 1, \ldots, M-1$. The transition probability from state i to state j is estimated from the normalized relative frequency of transitions, that is, $p_{i,j} = f_{i,j}/f_i$, where $f_{i,j}$ is the total number of transitions from state i to j and f_i is the total number of transitions out of state i. The resulting $M \times M$ matrix of state transition probabilities (denoted by $\boldsymbol{P}_\mathbf{a}$) can be translated into a corresponding infinitesimal generating matrix in the continuous-time domain, given by $\boldsymbol{B}_\mathbf{a} = (1/\tau_a)(\boldsymbol{P}_\mathbf{a} - \boldsymbol{I}_\mathbf{M})$, where $\boldsymbol{I}_\mathbf{M}$ is an identity matrix.

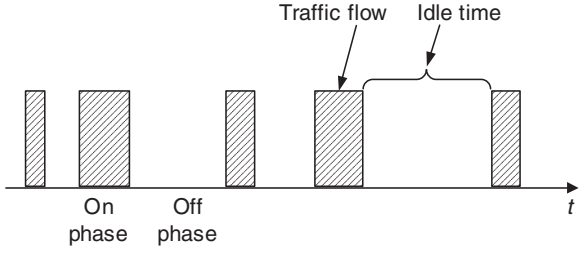

Figure 2 On-off flow dynamics of traffic flows.

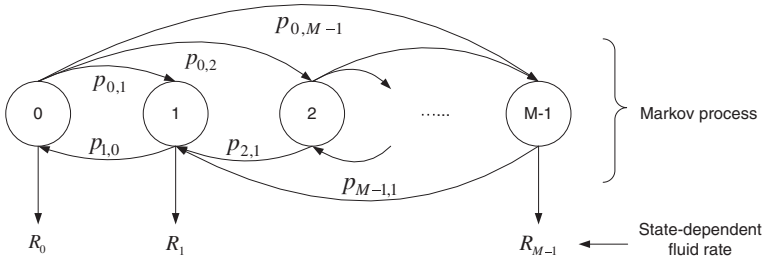

Figure 3 Markov-modulated fluid model for a multiplexed traffic flow.

4 Delay and loss analyses for multimedia services in mobile hotspots

4.1 Flow-level analysis for WLAN link with contention access

Suppose that N_v video users and N_d data users at maximum are admitted into the mobile hotspot. As illustrated in Figure 2, video and date users alternate between on and off phases. For the flow-level dynamics, it is well accepted to assume that the idle time before requesting a new flow is exponentially distributed. The video flow duration is independent of the transmission rate and modelled by an exponential distribution with a mean in the order of minutes. On the other hand, the data flow duration is variable depending on the number of packets in a flow (X_d) and packet service time (S_d). When there are n_v ($n_v \le N_v$) active video users and n_d ($n_d \le N_d$) active data users, based on the analytical model in Section 2.2, we derive the average service time of video and data packets $1/\nu_v(n_v, n_d)$ and $1/\nu_d(n_v, n_d)$ (in time unit), respectively. The total time to transmit all the packets of a data flow in such a state is described by a probability generating function (PGF) as follows[a]:

$$\begin{aligned} D_d(z) &= \sum_{k=1}^{\infty} \mathrm{P}[D_d = k]\, z^k = \sum_{n=1}^{\infty} \mathrm{P}[X_d = n] \sum_{k=1}^{\infty} \mathrm{P}[S_{d,1} + \ldots \\ &\quad + S_{d,n} = k | X_d = n]\, z^k \\ &= \sum_{n=1}^{\infty} \mathrm{P}[X_d = n] \cdot [S_d(z)]^n = X_d(S_d(z)) \\ &= \frac{\nu_d \eta_d z}{1 - (1 - \nu_d \eta_d) z}, \end{aligned} \tag{13}$$

where X_d and S_d follow geometric distributions with PGF given by

$$X_d(z) = \frac{\eta_d z}{1 - (1 - \eta_d) z}, \qquad S_d(z) = \frac{\nu_d z}{1 - (1 - \nu_d) z}. \tag{14}$$

As seen in (13), the expected data call duration follows a geometric distribution of a mean $1/(\nu_d \eta_d)$ when there are n_v active video users and n_d active data users. Nonetheless, the data call duration is state dependent with varying active users. Therefore, we need to average the call duration with flow-level dynamics. To measure service responsiveness, we refer to the overall data call duration as data response time and evaluate its first and second moments in the following.

According to the DCF access mechanism, packets from active flows take turns to be served, which is similar to the processor sharing (PS) discipline of a queueing system or a time-sharing computer system. A fair share of the total serving capacity is dedicated to each active flow, that is, the WLAN channel operates like a symmetric queueing system [30]. Hence, we model the dynamics of video and data flows in the mobile hotspot by a closed queueing network in Figure 4, where two customer classes of finite populations N_v and N_d are served in a PS manner. The queueing state is defined as the numbers of active video and data users, i.e., $\boldsymbol{n} = (n_1, n_2)$, where $0 \le n_1 \le N_v$ and $0 \le n_2 \le N_d$. Using the analytical approach in [31], we have the state transition rate matrix Q to describe the queueing dynamics of an underlying birth-and-death process, whose elements are given by

$$Q(\boldsymbol{n} - \boldsymbol{e_j}, \boldsymbol{n}) = \lambda_j \cdot (N_j - n_j + 1), \;\; j = 1, 2 \tag{15}$$

$$Q(\boldsymbol{n} + \boldsymbol{e_j}, \boldsymbol{n}) = \mu_j(\boldsymbol{n} + \boldsymbol{e_j}) \cdot (n_j + 1), \;\; j = 1, 2$$

$$Q(\boldsymbol{n}, \boldsymbol{n}) = -\sum_{j=1}^{2} \Big[\lambda_j \cdot (N_j - n_j) + (1 - \delta_{\boldsymbol{n},\boldsymbol{0}}) \cdot \mu_j(\boldsymbol{n}) \cdot n_j \Big],$$

where $N_1 = N_v$, $N_2 = N_d$, $\mu_1(\boldsymbol{n}) = \mu_1$, $\mu_2(\boldsymbol{n}) = \nu_d(\boldsymbol{n})\eta_d$, $\boldsymbol{e_j}$ is a two-dimensional unit vector, whose only non-zero element is the jth element equal to 1, and δ is the Kronecker delta symbol. Here, the service rate $\mu_2(\boldsymbol{n})$ depends on the average number of packets in a data flow (η_d^{-1}) and the average packet service time ($\nu_d(\boldsymbol{n})^{-1}$) obtained from the WLAN model. Then, stationary state probabilities $\boldsymbol{\pi}_{vd}$ are obtained by solving the following linear equation system:

$$\boldsymbol{\pi}_{vd} \boldsymbol{Q} = \boldsymbol{0}, \qquad \boldsymbol{\pi}_{vd} \boldsymbol{1} = 1 \tag{16}$$

where $\boldsymbol{0}$ and $\boldsymbol{1}$ are column vectors of all zeros and all ones, respectively.

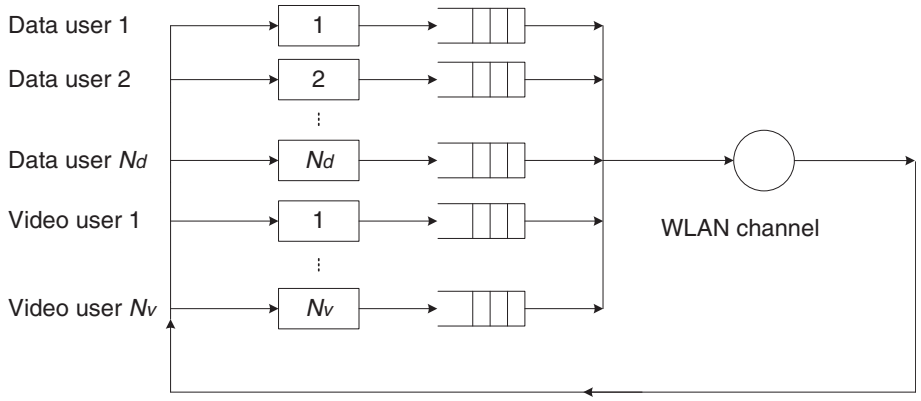

Figure 4 Closed queueing network modeling at flow level for multi-service users in WLAN mobile hotspots.

To further evaluate the response time of data flows, we need to consider a modified birth-and-death process, in which a tagged data flow sees $\boldsymbol{k}$ active flows in service at its time of entry. Here, $\boldsymbol{k} = (k_1, k_2)$, where $0 \le k_1 \le K_1$, $0 \le k_2 \le K_2$, $K_1 = N_\mathrm{v}$, and $K_2 = N_\mathrm{d} - 1$ (with one less than in the original process). The equilibrium distribution of the modified process is obtained similarly and denoted by $\tilde{\boldsymbol{\pi}}_\mathrm{vd}$. Using the approach in [31], we obtain the first and second moments of data response time as

$$\overline{\Gamma}_\mathrm{d} = \mathrm{E}[\,\Gamma_\mathrm{d}] = -\tilde{\boldsymbol{\pi}}_\mathrm{vd}\boldsymbol{M}^{-1}\mathbf{1}, \qquad \mathrm{E}[\,\Gamma_\mathrm{d}^2] = -2\tilde{\boldsymbol{\pi}}_\mathrm{vd}\boldsymbol{M}^{-2}\mathbf{1}, \tag{17}$$

where

$$M(\boldsymbol{k} - \boldsymbol{e_j}, \boldsymbol{k}) = \lambda_j \cdot (K_j - k_j + 1), \;\; j = 1, 2 \tag{18}$$

$$M(\boldsymbol{k} + \boldsymbol{e_j}, \boldsymbol{k}) = \mu_j(\boldsymbol{k} + \boldsymbol{e_j} + \boldsymbol{e_J}) \cdot (k_j + 1), \;\; j = 1, 2$$

$$M(\boldsymbol{k}, \boldsymbol{k}) = -\sum_{j=1}^{2} \Big[\lambda_j \cdot (K_j - k_j) + \mu_j(\boldsymbol{k} + \boldsymbol{e_J}) \cdot (k_j + \delta_{j,J})\Big],$$

and $J = 2$ indicates that the above derivation is for the class of data flows.

4.2 Packet-level analysis for video traffic over WLAN link

For interactive data service with saturated traffic, the packet access delay is evaluated in Section 2.2, while the response time of data flows is analyzed with the closed queueing network in Section 4.1. On the other hand, conversational video service is more sensitive to packet delay because of strict real-time requirement. In this section, we further investigate the video packet delay, taking into account batch arrival structure and varying serving capacity. Due to channel variations and flow-level dynamics, the available packet service rate is time-varying and state dependent. It is generally intractable to analytically evaluate the performance with such complex traffic and channel models. In [22,32] the performance bounds are derived for different regimes. Motivated by such findings, we can approximate the performance with lower and upper bounds, assuming that the traffic is served with a constant *average* capacity in the fluid regime or the capacity available in a state in the quasi-stationary regime. Specifically, given the state-dependent service time $S_\mathrm{v}(\boldsymbol{n})$ for video packets, we have the average serving capacity

$$\overline{S}_\mathrm{v} = \sum_{\boldsymbol{n}} \boldsymbol{\pi}_\mathrm{vd}(\boldsymbol{n}) S_\mathrm{v}(\boldsymbol{n}). \tag{19}$$

Then, by assuming a constant service time $\overline{S}_\mathrm{v}$, a lower bound for the video packet delay (denoted by T_v) can be obtained as

$$\overline{T}_\mathrm{v,l} = \mathrm{E}\big[T_\mathrm{v} | \overline{S}_\mathrm{v}\big]. \tag{20}$$

On the other hand, an upper delay bound is a probabilistic average of the conditional delay with the serving capacity available in each state, i.e.,

$$\overline{T}_\mathrm{v,u} = \sum_{\boldsymbol{n}} \boldsymbol{\pi}_\mathrm{vd}(\boldsymbol{n}) \mathrm{E}\big[T_\mathrm{v} | S_\mathrm{v}(\boldsymbol{n})\big]. \tag{21}$$

To evaluate the packet delay with a certain serving capacity, $\overline{S}_\mathrm{v}$ or $S_\mathrm{v}(\boldsymbol{n})$, we use a $D/G/1$ queueing model since video packet batches arrive over fixed burst intervals. The delay experienced by a tagged packet in a video batch (T_v) consists of three independent components: (1) the waiting time of the first packet of that batch to be served, denoted by W_b; (2) the waiting time due to the transmission of the packets of that batch queued before the tagged packet, denoted by W_q; and (3) the transmission time of the tagged packet S_v, which is modeled by a geometric distribution.

Firstly, W_b is the waiting time of the first packet generated in a video batch. To evaluate W_b with a queueing system, each video batch can be regarded as a single customer whose service time is the total transmission time of all packets in a batch. An analytical approach is introduced in [33] for the waiting time of a $D/G/1$ queue, where the

inter-arrival time is deterministic and the service time follows a general distribution. As defined in Sections 2.2 and 3.2, the packet transmission time follows a geometric distribution in (5), and the batch size is modeled by a negative binomial distribution in (10). Then the PGF of the total service time of a video batch is given by

$$\begin{aligned} G(z) &= \sum_{k=1}^{\infty} \mathrm{P}[\,G=k]\, z^k = \sum_{n=1}^{\infty} \mathrm{P}[\,A=n] \sum_{k=1}^{\infty} \mathrm{P}[\,S_{\mathrm{v},1} + \ldots \\ &\quad + S_{\mathrm{v,n}} = k | A=n]\, z^k \\ &= \sum_{n=1}^{\infty} \mathrm{P}[\,A=n] \cdot [\,S_{\mathrm{v}}(z)]^n = A(S_{\mathrm{v}}(z)) \end{aligned} \tag{22}$$

where $A(\cdot)$ and $S_{\mathrm{v}}(\cdot)$ are respectively the PGF of the NB distribution and geometric distribution that are as follows:

$$A(z) = \frac{z(1-p)^r}{(1-zp)^r}, \qquad S_{\mathrm{v}}(z) = \frac{\nu_{\mathrm{v}} z}{1-(1-\nu_{\mathrm{v}})z}. \tag{23}$$

Then using the approach in [33], we obtain the PGF of the waiting time W_{b} as

$$W_{\mathrm{b}}(z) = \frac{\Phi \cdot (z-1) \prod_{k=1}^{H-1} (z - z_k)}{z^H - G(z)} \tag{24}$$

where H is the burst interval in time units, z_1, ..., z_{H-1} are the unique roots of $z^H - G(z) = 0$ within the unit circle $|z| < 1$, and Φ is a normalization constant. When H is large, we can apply the Muller method [34] to numerically find the roots z_k. The normalization constant Φ is calculated by

$$\Phi = \lim_{z \to 1^-} \frac{z^H - G(z)}{(z-1)\prod_{k=1}^{H-1}(z-z_k)}. \tag{25}$$

Based on (24), the mean and variance of W_{b} are respectively given by

$$\overline{W}_{\mathrm{b}} = -\frac{H(H-1) - G''(1^-)}{2[\,H - G'(1^-)]} + \sum_{k=1}^{H-1} \frac{1}{1-z_k} \tag{26}$$

$$\begin{aligned} \sigma^2_{W_{\mathrm{b}}} &= -\sum_{k=1}^{H-1} \frac{z_k}{(1-z_k)^2} - \frac{H(H-1)(H-2) - G'''(1^-)}{3[\,H - G'(1^-)]} \\ &\quad - \frac{H(H-1) - G''(1^-)}{2[\,H - G'(1^-)]} + \left[\frac{H(H-1) - G''(1^-)}{2[\,H - G'(1^-)]} \right]^2. \end{aligned} \tag{27}$$

Next, W_{q} is the waiting time of a tagged packet to transmit all the other packets that are generated in the same video batch of the tagged packet but queued before it. Clearly, W_{q} depends on the batch size A and the position of the tagged packet. According to the analysis in [35], the probability that an arbitrary tagged packet falls within a batch of a size k is given by $k \cdot \mathrm{P}[\,A=k]\ / \overline{A}$. If n packets from the same video batch as the tagged packet are queued prior to it, the batch size must be no less than n. Hence, we obtain the PGF of the number of packets queued before the tagged one as

$$Y(z) = \sum_{n=1}^{\infty} z^n \sum_{k=n+1}^{\infty} \frac{k \cdot \mathrm{P}[\,A=k]}{\overline{A}} \cdot \frac{1}{k} = \frac{1 - A(z)}{\overline{A} \cdot (1-z)}. \tag{28}$$

The packet transmission time S_{v} is assumed to be geometrically distributed with its PGF given in (23). The PGF of the waiting time W_{q} is then

$$W_{\mathrm{q}}(z) = Y(S_{\mathrm{v}}(z)) = \frac{1 - A(S_{\mathrm{v}}(z))}{\overline{A} \cdot (1 - S_{\mathrm{v}}(z))}. \tag{29}$$

Hence, the kth factorial moment of W_{q} is obtained from (29) as follows

$$\mathrm{E}\big[(W_{\mathrm{q}}-1)\ldots(W_{\mathrm{q}}-k+1)\big] = \lim_{z \to 1^-} \frac{d^k W_{\mathrm{q}}(z)}{dz^k}. \tag{30}$$

The mean and variance of W_{q} are then

$$\overline{W}_{\mathrm{q}} = \lim_{z \to 1^-} \frac{dW_{\mathrm{q}}(z)}{dz}, \quad \sigma^2_{W_{\mathrm{q}}} = \lim_{z \to 1^-} \frac{d^2 W_{\mathrm{q}}(z)}{dz^2} + \overline{W}_{\mathrm{q}} - (\overline{W}_{\mathrm{q}})^2. \tag{31}$$

Since the overall packet delay consists of three independent components, the corresponding PGF of the packet delay is obtained from (23), (24), and (29) as

$$T_{\mathrm{v}}(z) = W_{\mathrm{b}}(z) \cdot W_{\mathrm{q}}(z) \cdot S_{\mathrm{v}}(z). \tag{32}$$

The mean and variance of packet delay can be evaluated accordingly by

$$\overline{T}_{\mathrm{v}} = \overline{W}_{\mathrm{b}} + \overline{W}_{\mathrm{q}} + \frac{1}{\nu_{\mathrm{v}}}, \qquad \sigma^2_{T_{\mathrm{v}}} = \sigma^2_{W_{\mathrm{b}}} + \sigma^2_{W_{\mathrm{q}}} + \frac{1-\nu_{\mathrm{v}}}{\nu^2_{\mathrm{v}}}. \tag{33}$$

4.3 Multiplexed traffic over highly varying WWAN link

Given two heterogeneous links to support mobile hotspots, there are various causes for packet loss. Inside the mobile hotspot, we focus on packet loss due to collisions over the WLAN link [2], in view of the small scale of a vehicle and the relatively stationary AP and user terminals. The collision probabilities p_{v} and p_{d} for video and data packets are derived in (1) and (2). Given a maximum retransmission limit L_r, we have the packet loss probability due to collisions

$$P_{\mathrm{cv}} = p_{\mathrm{v}}^{L_r+1}, \qquad P_{\mathrm{cd}} = p_{\mathrm{d}}^{L_r+1}. \tag{34}$$

Moreover, the highly varying WWAN link may introduce packet loss due to Nakagami channel fading. Consider AMC based on SNR ranges γ_n for N transmission modes.

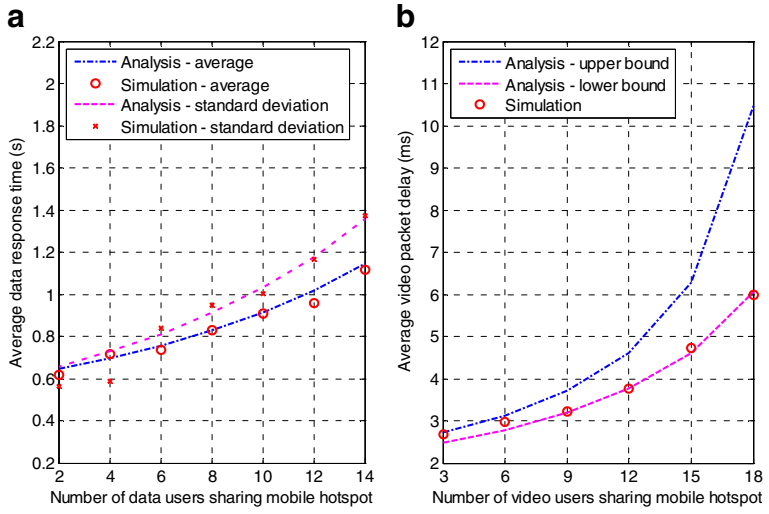

Figure 5 Comparison of simulation and analytical results in terms of data flow response time and video packet delay. (**a**) Data flow response time with $N_v = 5$. (**b**) Video packet delay with $N_d = 7$.

The probability of choosing mode n is given by

$$q_n = \int_{\gamma_n}^{\gamma_{n+1}} p_s(\gamma)\, d\gamma = \frac{\Upsilon(m, \frac{m\gamma_n}{\overline{\gamma}}) - \Upsilon(m, \frac{m\gamma_{n+1}}{\overline{\gamma}})}{\Gamma(m)}, \quad n = 0, 1, ..., N, \tag{35}$$

where $p_s(\gamma)$, given in (8), is the PDF of the received SNR over the Nakagami fading channel, and $\Upsilon(m, x) = \int_x^\infty t^{m-1} e^{-t}\, dt$ is the upper incomplete gamma function. The packet error rate in the presence of AWGN noise can be approximated by a piecewise exponential function [36]:

$$\varepsilon_n(\gamma) \approx \begin{cases} 1, & \text{if} 0 < \gamma < \varphi_n \\ a_n e^{-g_n \gamma}, & \text{if} \gamma \geq \varphi_n. \end{cases} \tag{36}$$

Normally, $\gamma_n > \varphi_n$, then the packet error rate of mode n is given by [37]

$$\begin{aligned} \zeta_n &= \frac{1}{q_n} \int_{\gamma_n}^{\gamma_{n+1}} a_n e^{-g_n \gamma} p_s(\gamma)\, d\gamma \\ &= \frac{1}{q_n} \frac{a_n}{(b_n)^m} \left(\frac{m}{\overline{\gamma}}\right)^m \frac{\Upsilon(m, b_n\gamma_n) - \Upsilon(m, b_n\gamma_{n+1})}{\Gamma(m)}, \end{aligned} \tag{37}$$

where $b_n = (m/\overline{\gamma}) + g_n$. Obviously, $\zeta_0 = 1$ for mode 0 in deep fades. Therefore, we obtain the packet loss probability due to channel fading as

$$P_f = \frac{\sum_{n=0}^{N} \beta_n \zeta_n q_n}{\sum_{n=0}^{N} \beta_n q_n}, \tag{38}$$

where β_n is the modulation and coding rate in bits/symbol. According to the search algorithm in [37], the SNR ranges γ_n $(n = 1, ..., N)$ are determined so that $\zeta_n = P_0$, which naturally leads to $P_f = P_0$.

Further, the aggregate traffic from the mobile hotspot is multiplexed as an M-state Markov-modulated fluid flow at the AP before relayed over the WWAN link toward the BS. To bound the packet delay for real-time traffic, a buffer limit B_t (packets) can be applied to the transmission queue. Then, buffer overflow also leads to data loss. Based on the level crossing rates of N transmission modes, the WWAN link with AMC and Nakagami fading is characterized by an $(N+1)$-state FSMC model [20]. Therefore, the statistics of the queue length can be analyzed with a fluid approach [38]. We defined a coupled traffic and link state $k = (N+1)(i-1) + j$, which indicates that the aggregate flow is in state i and the WWAN link is in state j, where $0 \leq i \leq M-1$, $0 \leq j \leq N$, $0 \leq k \leq U-1$, and $U = (N+1)M$.

Let $F_k(x)$ denote the joint probability that the system is in state k and the queue length is no greater than x. Then, the equilibrium queue length distribution satisfies [39]

$$\frac{d\boldsymbol{F}(x)}{dx}\boldsymbol{D} = \boldsymbol{F}(x)\boldsymbol{B} \tag{39}$$

$$\boldsymbol{D} = \boldsymbol{R} \oplus [-\boldsymbol{C}], \quad \boldsymbol{B} = \boldsymbol{B_a} \oplus \boldsymbol{B_c} = \boldsymbol{B_a} \otimes \boldsymbol{I_{N+1}} + \boldsymbol{I_M} \otimes \boldsymbol{B_c}, \tag{40}$$

where $\boldsymbol{F} = [F_0(x), F_1(x), ..., F_{U-1}(x)]$; $\boldsymbol{R} = \text{diag}\,[R_0, R_1, ..., R_{M-1}]$ is a diagonal matrix with the diagonal elements being the multiplexed traffic rates (packets/s) at M states; $\boldsymbol{C} = \text{diag}[C_0, C_1, ..., C_N]$ is also a diagonal matrix giving the effective data rates at $N+1$ transmission modes; $\boldsymbol{B_a}$ and $\boldsymbol{B_c}$ are the generating matrices of the aggregate fluid and the WWAN link, respectively; and $\oplus$ and $\otimes$ are Kronecker sum and Kronecker product, respectively.

Table 2 System parameters for numerical analysis

Symbol	Definition	Value
$1/\lambda_1$	Average idle time of video users (s)	50
$1/\lambda_2$	Average idle time of data users (s)	12
$1/\mu_1$	Average video flow duration (s)	82.5
τ_b	Video burst interval (s)	0.033
$\overline{X}_v$	Average number of bursts in a video flow	2,500
$\overline{X}_d$	Average number of packets in a data flow	400
p	Parameter of NB distribution for batch size A	0.6946
r	Parameter of NB distribution for batch size A	0.1200
L_v	Video packet size (byte)	750
L_d	Data packet size (byte)	480
L_a	WWAN link packet size (byte)	250
M	Number of traffic states of aggregate flow	4
N	Number of valid transmission modes over WWAN link	5
P_0	Target packet error probability due to WWAN channel fading	0.001
T_f	AMC adaptation period (s)	0.001
f_c	Carrier frequency of WWAN link (MHz)	900
m	Fading severity parameter of WWAN link	1.1∼2.2
v	Velocity of mobile hotspot (m/s)	5.6
κ	Path loss exponent of WWAN link	2.75
d_0	Reference distance between vehicle and BS (m)	400
$\overline{\gamma}(d_0)$	Average received SNR at reference distance (dB)	20

The solution to the above differential equation system takes the following form:

$$\boldsymbol{F}(x) = \sum_{k=0}^{U-1} a_j \boldsymbol{\Phi_k} e^{u_k x}, \tag{41}$$

where u_k and $\boldsymbol{\Phi_k}$ are the eigenvalues and *row* eigenvectors of $\boldsymbol{BD}^{-1}$, respectively. It is straightforward to have $u_0 = 0$, $\boldsymbol{\Phi_0} = \boldsymbol{\psi}$, and $\boldsymbol{\psi B} = \boldsymbol{0}$. The a_ks are constants to be determined by invoking boundary conditions [38]. That is, for an arbitrary coupled system state k,

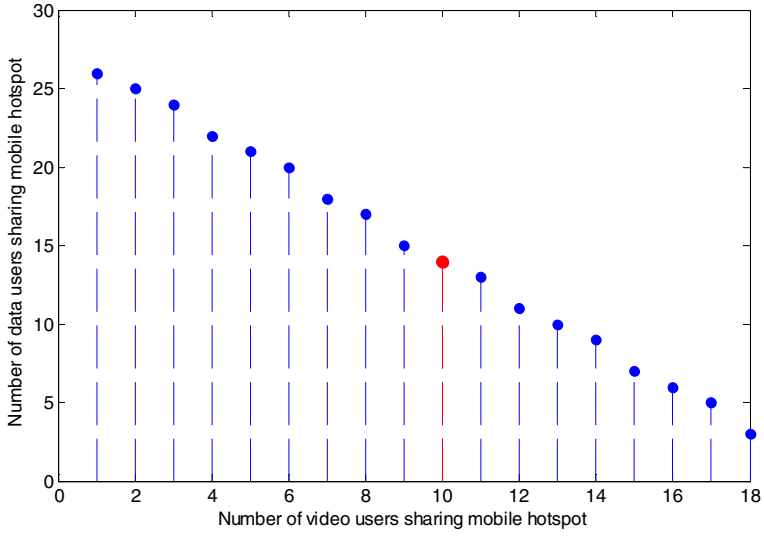

Figure 6 Maximum numbers of video and data users (N_v, N_d) admissible in the mobile hotspot.

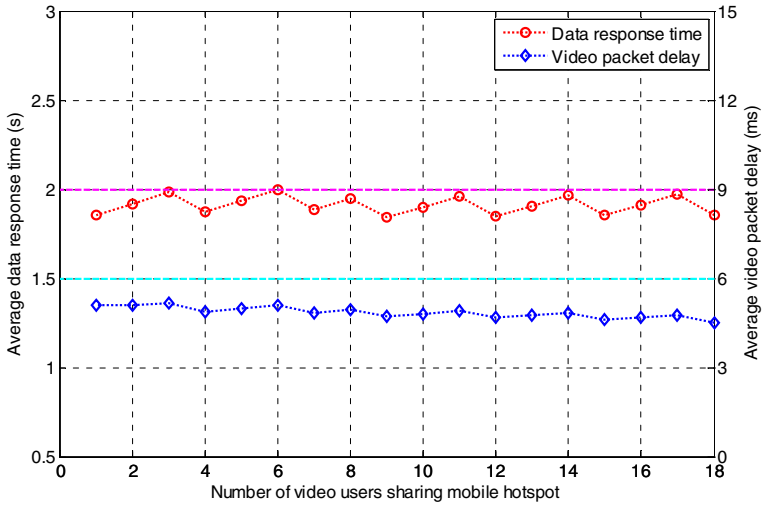

Figure 7 Mean data response time and video packet delay for cases at the boundary of the admission region.

$$k = (N+1)(i-1)+j,$$

$$\begin{aligned} F_k(0^+) &= 0, \quad \text{if } R_i > C_j \\ F_k(B_t^-) &= \psi(k), \text{ if } R_i \le C_j. \end{aligned} \tag{42}$$

The general queue length distribution in equilibrium is then $F(x) = \sum_{k=0}^{U-1} F_k(x)$. Thus, the probability of packet loss due to buffer overflow is given by

$$P_o = 1 - \frac{\tilde{\lambda}_a}{\lambda_a}, \quad \lambda_a = \sum_{i=0}^{M-1}\sum_{j=0}^{N} R_i \psi((N+1)(i-1)+j) \tag{43}$$

$$\tilde{\lambda}_a = \lambda_a - \sum_{i=0}^{M-1}\sum_{j=0}^{N} (R_i - C_j)\left[\psi((N+1)(i-1)+j) - F(B_t^-)\right]. \tag{44}$$

As packet loss is attributed to access collision, channel fading, and buffer overflow, the overall packet loss probability is obtained from (34), (38), and (43) as

$$P_l = 1 - (1-P_c)(1-P_f)(1-P_o). \tag{45}$$

The packet delay over the WWAN link is obtained by the Little's law as

$$\overline{T}_a = \frac{1}{\tilde{\lambda}_a}\int_0^{B_t} \left[1 - F(x)\right] dx. \tag{46}$$

5 Numerical results and discussions

Although data and video users alternate between active and idle states, the number of admitted users should be restricted with admission control so that admitted users are provided QoS assurance. Based on the analytical framework in Section 4, we can determine the maximum numbers of videos and data users admissible to the mobile hotspot so that the data response time and video packet delay are bounded. In this section, we first present numerical results to validate the accuracy of the analysis and demonstrate its application to admission control.

Table 2 gives the system parameters for numerical analysis. In particular, the WLAN link settings follow the specification of IEEE 802.11. The WWAN link adopts the AMC algorithm in [37]. The data traffic model refers to the evaluation standards for the universal mobile telecommunications system (UMTS) [25,26]. The video traffic parameters are based on H.264/AVC video traces of *Tokyo Olympics* and *NBC News* [40]. These video sequences have a CIF (352 × 288) resolution, a fixed rate of 30 frames/s, a GoP size of 16, 7 B frames between two I/P key pictures, and a quantization step-size indexed at 38.

5.1 Validation of analytical framework

Using the approaches in Section 4, we evaluate the data flow response time and video packet delay. As our numerical analysis involves complex calculations, we use the Symbolic Math Toolbox of MATLAB 7.10.0 (R2010a) (MathWorks Inc., Natick, MA, USA) to solve nonlinear equation systems, one-sided limit, derivatives, and definite integrals. To verify the accuracy of the analysis, we develop an event-driven simulator with C++ for a mobile hotspot using the channel models presented in Section 2. Multimedia traffic is generated according to the traffic models in Section 3.

Figure 5a compares the simulation results with the analytical results in terms of data response time when there are $N_v = 5$ video users and varying data users. As seen, the analytical results match well the simulation results. Figure 5b shows the video packet delay with $N_d = 7$ data users and varying video users. It can be seen that the analytical results right fall between the upper bound and the lower bound.

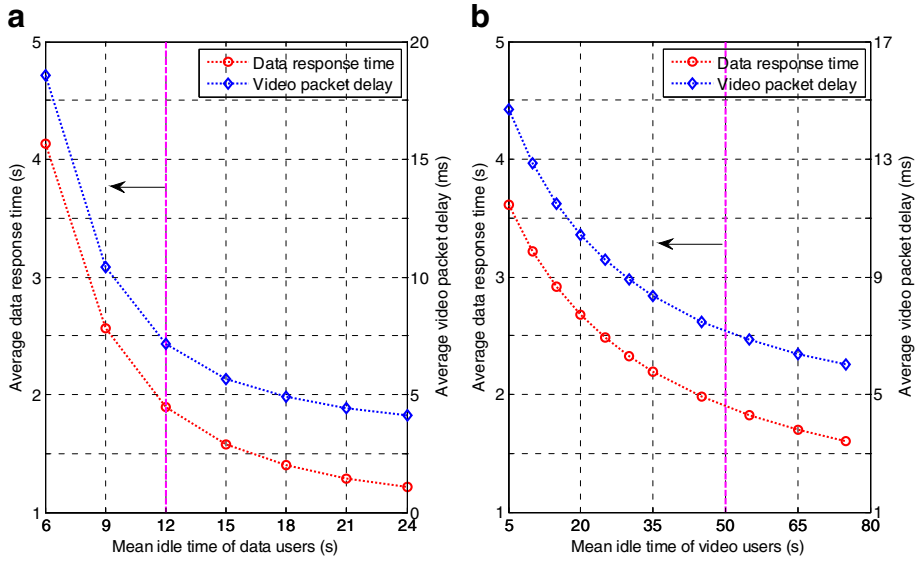

Figure 8 Performance outage with user traffic variation when $N_v = 10$ and $N_d = 14$. (**a**) With varying mean idle time of data users. (**b**) With varying mean idle time of video users.

5.2 Application to admission control for mobile hotspots

Figure 6 shows the pairs of (N_v, N_d) as the admission region with the system parameters in Table 2. For example, as shown in the red circle, when $N_v = 10$, the maximum number of data users allowed in the mobile hotspot is $N_d = 14$. For the points at the boundary of the admission region, we present the corresponding video and data performance in Figure 7. It can be seen that the data response time and video packet delay are bounded if the numbers of video and data users are within the admission region.

A useful aspect of the analytical framework is to effectively adapt the admission control with traffic and channel variations. The admission region in Figure 6 is obtained with an average idle time of 50 s and 12 s for video and data users, respectively. Figure 8 shows the performance variation with $N_v = 10$ and $N_d = 14$ when user activities are dynamically changing. When users become more active,

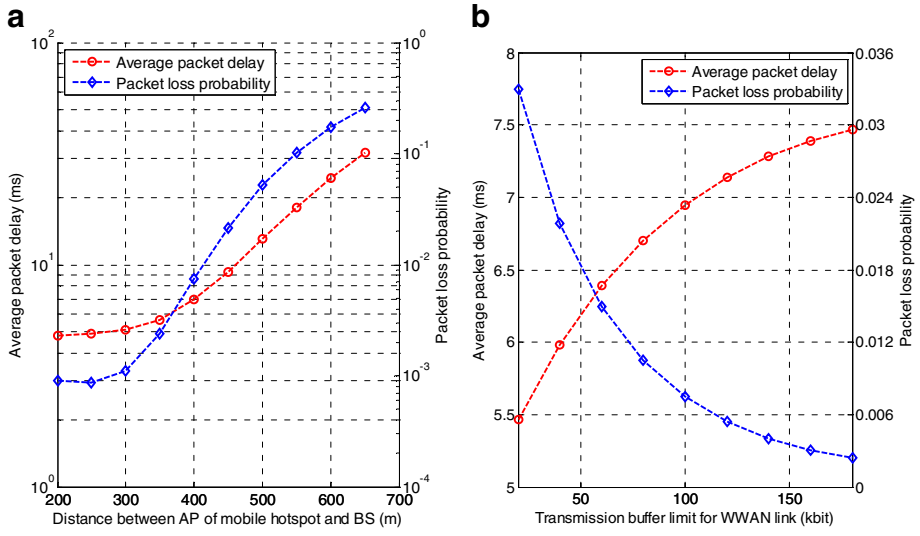

Figure 9 Performance variation with AP-BS distance and transmission buffer limit. (**a**) Average packet delay vs. AP-BS distance. (**b**) Average packet delay vs. transmission buffer limit of WWAN link.

i.e., with a shorter idle time, the admission region must be updated timely to avoid performance outage. An effective and accurate analytical framework is important to enable adaptive admission control.

The performance of the mobile hotspot also depends on the distance between the AP and the BS and the transmission buffer limit [2]. As discussed in Section 4.3, the average received SNR of the WWAN link affects the fading loss and achievable data rate. The distance between the AP and the BS is an essential factor to determine the average SNR ($\overline{\gamma}$) as well as the fading severity parameter (m). As found in [41,42], the fading severity parameter is inversely proportional to the transmitter-receiver separation distance. An empirical relationship is also developed in [43]. Figure 9a clearly demonstrates the effect of the distance. As seen when the vehicle moves away from the BS and the distance is greater than 400 m, the packet loss probability violates the upper bound of 1%. The mobile hotspot can be handed over to a closer or stronger BS in the vicinity. In view of the trade-off between the packet delay and loss probability illustrated in Figure 9b, we can also enlarge the transmission buffer size to mitigate data loss if the increasing packet delay is still acceptable.

6 Conclusions

In this paper, we study admission control for heterogeneous mobile hotspots to offer ubiquitous multimedia services in a vehicular environment. To derive the maximum numbers of admissible users, we develop a comprehensive analytical framework to evaluate the achievable performance. A Nakagami fading channel with AMC is considered for the WWAN link, whereas the contention-based WLAN link is characterized by a state-dependent birth-and-death process. Taking into account on-off user dynamics at the flow level, we model the time-varying packet service time in the WLAN with state-dependent geometric distributions. Moreover, the proposed analytical approach addresses the unique video traffic feature of batch packet arrivals. The delay and loss performance within the mobile hotspot is effectively analyzed by means of a closed queueing network in a processor-sharing manner and a discrete-time $D/G/1$ queueing system with batch arrivals. Further, we apply the well-known fluid approach to evaluate the performance of Markov-modulated aggregate traffic from the mobile hotspot over a highly varying WWAN channel modeled by a finite-state Markov chain. As seen in the simulation and analytical results, the multimedia QoS requirement is satisfied when we limit the number of users admitted in the mobile hotspots using the proposed approach. In the future work, it would be interesting to use the proposed analysis to develop an efficient handoff algorithm adaptive to traffic and channel variations.

Endnote

[a]The parameters n_v and n_d are skipped for presentation clarity.

Competing interests

The authors declare that they have no competing interests.

About the Authors

Wei Song (M'09) received her Ph.D. degree in electrical and computer engineering from the University of Waterloo, Canada, in 2007. Since 2008, she has worked as a postdoctoral research fellow at the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley. In July 2009, she joined the Faculty of Computer Science, University of New Brunswick, Fredericton, Canada, as an Assistant Professor. She received a Best Paper Award from IEEE WCNC 2007, a Top 10% Award from IEEE MMSP 2009 and a Best Student Paper Award from IEEE CCNC 2013. Her current research interests include the heterogeneous interworking of wireless networks, cooperative wireless networking, mobile hotspots, cross-layer design for multimedia QoS provisioning. She co-chaired the Wireless Access Track of IEEE VTC Fall 2010, the General Symposium of IWCMC 2011, and the Wireless Communications Symposium of IEEE GLOBECOM 2011. She is also an Associate Editor of *IEEE Transactions on Vehicular Technology*, and an Editor of Wiley's journal of *Wireless Communications and Mobile Computing*.

Peijian Ju (S'12) received a B.S. degree in electronics and information engineering and an M.S. degree in electronic science and technology from Huazhong University of Science and Technology, Wuhan, China, in 2009 and 2011, respectively. Currently, he is working toward a Ph.D. degree with the Faculty of Computer Science, University of New Brunswick, Fredericton, Canada. His research interests include multiple access control and cross-layer design for cooperative wireless networks.

Yu Cheng (S'01-M'04-SM'09) received the B.E. and M.E. degrees in Electrical Engineering from Tsinghua University, Beijing, China, in 1995 and 1998, respectively, and the Ph.D. degree in Electrical and Computer Engineering from the University of Waterloo, Waterloo, Ontario, Canada, in 2003. From September 2004 to July 2006, he was a postdoctoral research fellow in the Department of Electrical and Computer Engineering, University of Toronto, Ontario, Canada. Since August 2006, he has been with the Department of Electrical and Computer Engineering, Illinois Institute of Technology (IIT), Chicago, Illinois, USA, and now as an Associate Professor. His research interests include next-generation Internet architectures and management, wireless network performance analysis, network security, and wireless/wireline interworking. He received a Postdoctoral Fellowship Award from the Natural Sciences and Engineering Research Council of Canada (NSERC) in 2004, and a Best Paper Award from the conferences QShine 2007 and ICC 2011. He received the National Science Foundation (NSF) CAREER Award in 2011 and IIT Sigma Xi Research Award in the junior faculty division in 2013. He served as a Co-Chair for the Wireless Networking Symposium of IEEE ICC 2009, a Co-Chair for the Communications QoS, Reliability, and Modeling Symposium of IEEE GLOBECOM 2011, a Co-Chair for the Signal Processing for Communications Symposium of IEEE ICC 2012, a Co-Chair for the Ad Hoc and Sensor Networking Symposium of IEEE GLOBECOM 2013, and a Technical Program Committee (TPC) Co-Chair for WASA 2011. He is an Associated Editor for *IEEE Transactions on Vehicular Technology* and New Books & Multimedia Column Editor for *IEEE Network*.

Author details

[1]Faculty of Computer Science, University of New Brunswick, Fredericton, Canada. [2]Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL, USA.

Acknowledgements

This research was supported by research grants from Natural Sciences and Engineering Research Council (NSERC) of Canada and New Brunswick Innovation Foundation (NBIF).

Received: 30 December 2012 Accepted: 15 May 2013
Published: 29 May 2013

References

1. S Pack, H Rutagemwa, XS Shen, JW Mark, L Cai, Performance analysis of mobile hotspots with heterogeneous wireless links. IEEE Trans Wireless Commun. **6**(10), 3717–3722 (2007)
2. S Pack, XS Shen, JW Mark, L Cai, Throughput analysis of TCP friendly rate control in mobile hotspots. IEEE Trans. Wireless Commun. **7**, 193–203 (2008)
3. WS Mossberg, Internet-a-Gogo: airlines to offer in-flight access. WSJ (2008). http://online.wsj.com/article/SB121382851874286403.html. Accessed 5 August 2012
4. MM Hasan, JW Mark, X Shen, A link adaptation scheme for the downlink of mobile hotspot. Wireless Commun. Mobile Comput. **12**(16), 1458–1470 (2011)
5. DH Ho, S Valaee, Information raining and optimal link-layer design for mobile hotspots. IEEE Trans. Mobile Comput. **4**(3), 271–284 (2005)
6. S Pack, XS Shen, JW Mark, J Pan, Mobility management in mobile hotspots with heterogeneous multihop wireless links. IEEE Commun. Mag. **45**(9), 106–112 (2007)
7. V Devarapalli, R Wakikawa, A Petrescu, P Thubert, Network mobility (NEMO) basic support protocol. IETF RFC 3963 2005 http://tools.ietf.org/html/rfc3963. Accessed 5 August 2012
8. C Huang, C Lee, J Zheng, A novel SIP-based route optimization for network mobility. IEEE J. Select. Areas Commun. **24**(9), 1682–1691 (2006)
9. L Zhu, FR Yu, B Ning, T Tang, Cross-layer design for video transmissions in metro passenger information systems. IEEE Trans. Veh. Technol. **60**(3), 1171–1181 (2011)
10. Allot Communications, Allot mobile trends report H2. (2011-2012)
11. W Song, H Jiang, W Zhuang, Performance analysis of the WLAN-first scheme in cellular/WLAN interworking. IEEE Trans. Wireless Commun. **6**(5), 1932–1952 (2007)
12. J Yin, G Holland, T ElBatt, F Bai, H Krishnan, in *Proceedings of the First International Conference on Communications and Networking in China (Chinacom'06)*. DSRC channel fading analysis from empirical measurement (Beijing, 25–27 October 2006)
13. Q Liu, S Zhou, GB Giannakis, Cross-layer scheduling with prescribed QoS guarantees in adaptive wireless networks. IEEE J. Select. Areas Commun. **23**(5), 1056–1066 (2005)
14. D Anick, D Mitra, MM Sondhi, Stochastic theory of a data handling system with multiple sources. The Bell, Syst. Tech. J. **61**(8), 1871–1894 (1982)
15. H Bobarshad, M van der Schaar, MR Shikh-Bahaei, A low-complexity analytical modeling for cross-layer adaptive error protection in video over WLAN. IEEE Trans. Multimedia. **12**(5), 427–438 (2010)
16. A Abdrabou, W Zhuang, Service time approximation in IEEE 802.11 single-hop *ad hoc* networks. IEEE Trans. Wireless Commun. **7**, 305–313 (2008)
17. N Nazirah, AZA Izzati, N Fisal, SK Seyd Yusof, SHS Ariffin, M Abbas, in *Proceedings of the 9th WSEAS International Conference on System Science and Simulation in Engineering (ICOSSSE)*. Cross-layer routing approach in high speed mobile wireless networks, (2010), pp. 238–243
18. MS Alouini, AJ Goldsmith, Adaptive modulation over Nakagami fading channels. Wireless Personal Commun. **13**, 119–143 (2000)
19. V Garg, *Wireless Communications and, Networking*. (Morgan Kaufmann, San Francisco, 2007)
20. J Razavilar, KJR Liu, SI Marcus, Jointly optimized bit-rate/delay control policy for wireless packet networks with fading channels. IEEE Trans. Commun. **50**(3), 484–494 (2002)
21. 3GPP, Services and service capabilities. 3GPP TS 22.105 V10.0.0 (2011) http://www.3gpp.org/ftp/Specs/html-info/22105.htm. Accessed 5 August 2012
22. F Delcoigne, A Proutière, G Régnié, Modeling integration of streaming and data traffic. Perform. Eval. **55**(3-4), 185–209 (2004)
23. J Cao, M Andersson, C Nyberg, M Kihl, in *Proceedings of the 10th International Conference on Telecommunications (ICT)*, vol. 2, Tahiti, Papeete, French Polynesia. Web server performance modeling using an M/G/1/K*PS queue (IEEE Piscataway, 2003), pp. 1501–1506
24. S Wenger, H.264/AVC over IP. IEEE Trans. Circuits Syst. Video Technol. **13**(7), 645–656 (2003)
25. 3GPP, Selection procedures for the choice of radio transmission technologies of the UMTS. 3GPP TS 30.03 V3.2.0 (1998)
26. 3GPP, IP transport in UTRAN. 3GPP TR 25.933 V5.4.0 (2004). http://www.3gpp.org/ftp/Specs/html-info/3003U.htm Accessed 5 August 2012
27. J Catone, Any way you slice it: YouTube dominates online video (2008). http://www.sitepoint.com/. Accessed 5 August 2012
28. DMB Masi, MJ Fischer, DA Garbin, Video frame size distribution analysis. Telecommunications Rev. **19**, 74–86 (2008)
29. P Skelly, M Schwartz, S Dixit, A histogram-based model for video traffic behavior in an ATM multiplexer. IEEE/ACM Trans. Netw. **1**(4), 446–459 (1993)
30. FP Kelly, *Reversibility and Stochastic Networks*. (Wiley, New York, 1979)
31. D Mitra, JA Morrison, Asymptotic expansions of moments of the waiting time in closed and open processor-sharing systems with multiple job classes. Adv. Appl. Prob. **15**(4), 813–839 (1983)
32. T Bonald, A Proutière, On performance bounds for the integration of elastic and adaptive streaming flows. ACM SIGMETRICS Performance Eval. Rev. **32**, 235–245 (2004)
33. LD Servi, D/G/1 queues with vacations. Operations Res. **34**(4), 619–629 (1986)
34. SD Conte, C de Boor, *Elementary Numerical Analysis: an Algorithmic Approach*. (McGraw-Hill, New York, 1972)
35. PJ Burke, Delays in single-server queues with batch input. Operations Res. **23**(4), 830–833 (1975)
36. Q Liu, S Zhou, GB Giannakis, Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links. IEEE Trans. Wireless Commun. **2**(5), 1746–1775 (2004)
37. Q Liu, S Zhou, GB Giannakis, Queuing with adaptive modulation and coding over wireless links: Cross-layer analysis and design. IEEE Trans. Wireless Commun. **4**(3), 1142–1153 (2005)
38. M Schwartz, *BroadBand Integrated Networks*. (Prentice Hall, Upper Saddle River, 1996)
39. F Wan, L Cai, E Shihab, A Gulliver, Admission region of triple-play services in wireless home networks. Comput. Commun. **33**(7), 852–859 (2010)
40. P Seeling, M Reisslein, B Kulapala, Network performance evaluation with frame size and quality traces of single-layer and two-layer video: a tutorial. IEEE Commun. Surv. Tut. **6**(3), 58–78 (2004)
41. L Cheng, BE Henty, DD Stancil, F Bai, P Mudalige, Mobile vehicle-to-vehicle narrow-band channel measurement and characterization of the 5.9 GHz dedicated short range communication (DSRC) frequency band. IEEE J. Select. Areas Commun. **25**(8), 1501–1516 (2007)
42. MI Azam, AUH Sheikh, in *Proceedings of the IEEE 13th International Multitopic Conference (INMIC)*. Error performance over frequency-selective variable Nakagami fading channel with RAKE reception (IEEE Islamabad, 14–15 December 2009)
43. G Dib, Vehicle-to-vehicle channel simulation in a network simulator. Master's thesis, Carnegie Mellon University 2009

doi:10.1186/1687-1499-2013-142
Cite this article as: Song *et al.*: Call admission control for integrated multimedia service in heterogeneous mobile hotspots. *EURASIP Journal on Wireless Communications and Networking* 2013 **2013**:142.