Cross-Layer Multipath Transmission of Elastic and Streaming Traffic over Heterogeneous Wireless Networks and Its Performance Analysis*

Wei Song

Faculty of Computer Science
University of New Brunswick
Fredericton, NB, Canada
wsong@unb.ca

Abstract. Next-generation wireless networks are expected to be heterogeneous by integrating multiple broadband access technologies. Popular wireless devices become equipped with various network interfaces. Multihoming support can be enabled to allow for multiple simultaneous associations with heterogeneous networks. Taking advantage of multihoming capability, we investigate multipath transmission of elastic and streaming traffic over heterogeneous wireless networks. A new flow splitting and multipath transmission scheme is proposed by exploiting cross-layer information of bulk data and video streams. The heavy-tailedness of elastic flow size is mitigated by balancing the traffic load, while the large bandwidth requirement of streaming flows is satisfied by aggregating fractional bandwidth available in multiple networks. Based on the leaky bucket algorithm and a nearly decomposable Markov process, the flow-level performance is evaluated in terms of flow blocking probabilities and data loss probability. Also, we analyze the fine-granular packet-level performance with respect to transfer delay using a batch arrival queueing model. Numerical results demonstrate the performance gain of the multipath transmission scheme.

Key words: Heterogeneous wireless networking, multipath streaming, multihoming support, elastic and streaming services.

1 Introduction

Nowadays, a variety of broadband access options are offered by the proliferating wireless networks such as the third-generation (3G) cellular networks, IEEE 802.11 wireless local area networks (WLAN), and IEEE 802.15 wireless personal area networks (WPAN). Aiming at different application environments, these

* This research was supported by a Discover Grant from Natural Sciences and Engineering Research Council (NSERC) of Canada and a Start-up Grant from New Brunswick Innovation Foundation (NBIF).

QSHINE 2010, November 17-19, Houston, United States

DOI 10.4108/eai.17-11-2010.2245427

wireless networks will coexist and provide an integrated heterogeneous access to mobile users. To exploit the complementary strengths of heterogeneous networks, multi-radio and smart wireless devices will be the mainstream for future wireless networks. To coordinate the heterogeneous wireless access and multi-radio devices, network selection is one of the major issues that are researched intensively in the literature. Depending on a decision algorithm, an incoming traffic flow is automatically assigned to a best available network. An ongoing traffic flow can also be dynamically migrated between different networks via vertical handoff by monitoring available bandwidth, channel status, and topology change. Many centralized and distributed network selection algorithms are proposed to provide the *always-the-best* connectivity and enhance quality-of-service (QoS) [9,14,16]. Basically, access selection aims to sharing the heterogeneous network resources along the time scale at the flow level.

Taking one step further, we can exploit the multihoming and multi-streaming support and apply flow splitting and data stripping for multipath transmission over heterogeneous wireless links. Multihoming enables a wireless device to maintain multiple simultaneous associations with more than one attachment point. Multi-streaming allows data to be partitioned into multiple streams and delivered independently to the application at the receiver. Multi-streaming can prevent head-of-line blocking problem that occurs in the regular transport control protocol (TCP). The stream control transmission protocol (SCTP) is one of the well-known transport-layer specifications that offer multihoming and multi-streaming capabilities. The original SCTP is designed to improve throughput and reliability by exploiting multiple paths. It is further extended to support host mobility and even interworking functionality [7]. If multihoming and multi-streaming capabilities are enabled for multi-radio devices, a traffic flow can be split into multiple streams and delivered simultaneously over multiple network interfaces. As such, the access selection problem is addressed from a different perspective.

In this paper, we study bandwidth sharing for integrated heterogeneous wireless networks by means of multipath transmission and flow splitting. Taking advantage of multihoming and multi-streaming capabilities of multi-radio devices, a novel cross-layer multipath transmission scheme is proposed to make use of any fractional bandwidth available in the integrated networks. Traffic flows are broadly classified into two primary categories: elastic and streaming [2]. Here, we consider two representative services, i.e., bulk data transfer and video streaming. The proposed scheme exploits the application-layer knowledge such as data file size and coding and compression structure to enable simultaneous data stripping across multiple networks. Video frames arriving in a burst are dispatched to heterogeneous links of different capacities based on frame types and frame grouping.

Also, we evaluate the flow-level performance of the multipath transmission scheme in terms of flow blocking probabilities and transfer delay. The analysis is based on a straightforward two-dimensional Markov chain, which is nearly decomposable in the quasi-stationary regime. At the packet level, a batch ar-

rival queueing process [6] is employed to analyze the statistics of packet transfer delay. As demonstrated in the numerical results, multipath transmission significantly outperforms a randomized access selection with respect to bandwidth aggregation.

The remainder of this paper is organized as follows. In Section 2, we introduce the network model and traffic model for this study. A novel multipath transmission is proposed and analyzed in Section 3 for elastic and streaming traffic, exploiting application-layer information and multihoming capability of multi-radio devices. Numerical results are presented in Section 4, followed by conclusions in Section 5.

2 System Model

In this study, we consider a heterogeneous wireless infrastructure integrating multiple access options, as shown in Fig. 1. Multi-radio devices are considered to be multihoming capable with an extended SCTP protocol. We focus on a scenario that three network interfaces can be activated simultaneously for power efficiency and interference mitigation. The study can be extended to relax this limitation. Moreover, we assume that a middleware deployed at both the application server and user device deals with the splitting and merging of traffic flows across available networks. The middleware employs both the application-layer and network-layer information in load splitting. Specifically, the available bandwidth over each associated network can be estimated by exploiting the multiple interfaces of user devices.

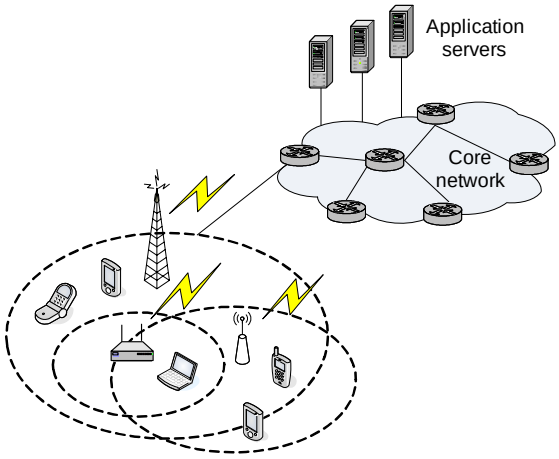

Fig. 1. System model of integrated heterogeneous wireless networks.

Nowadays, the wireless network capacity is boosted significantly with advanced techniques such as adaptive modulation and coding, power control, and multiple input and multiple output (MIMO). A variety of new services become

proliferating, such as video streaming, Web browsing, file transfer, and conversational video. These applications are broadly classified into two categories, i.e., elastic and streaming services. Elastic services such as file transfer and Web browsing can accept varying rates depending on available bandwidth. The main performance metric is the transfer delay or equivalently the throughput of traffic flows. On the other hand, streaming services require to preserve an intrinsic rate. For instance, video streaming plays back the video content at the receiver during the delivery. A stringent rate requirement needs to be satisfied to prevent data overflow and depletion at the playout buffer. In this study, we focus on flow splitting and multipath transmission for two representative elastic and streaming services: bulk data transfer and video streaming.

2.1 Elastic Traffic

An elastic flow is characterized by the size of bulk data to be transferred. It is widely observed that the size of Web documents and data files is heavy-tailed and presents high variability. Here, we model the elastic flow size L_e with a Weibull distribution as in [10], whose probability density function (PDF) is given by

$$f_e(x) = \frac{\alpha_e}{\beta_e}\left(\frac{x}{\beta_e}\right)^{\alpha_e - 1} e^{-(x/\beta_e)^{\alpha_e}}$$
$$0 < \alpha_e \leq 1, \qquad \beta_e > 0, \qquad x > 0 \tag{1}$$

where α_e is the shape parameter and β_e is the scale parameter. The exponential distribution is a special case of the Weibull distribution with $\alpha_e = 1$, while the Weibull distribution is heavy-tailed when $0 < \alpha_e < 1$. The smaller the value of α_e, the heavier the tail of a Weibull distribution.

2.2 Streaming Traffic

The essential traffic characteristics of streaming flows are the flow duration and variable rate. In this study, we take video streaming flows as an example. It is known that video traffic is inherently long-range dependent and highly correlated due to compression coding. In the third-generation (3G) cellular networks, H.264 Advanced Video Coding (AVC) is recommended for high-quality video [1]. To remove temporal redundancy, intracoded (I) frames are interleaved with predicted (P) frames and bidirectionally coded (B) frames. I frames are compressed versions of raw frames independent of other frames, whereas P frames only refer preceding I/P frames and B frames can refer both preceding and succeeding frames. A sequence of video frames from a given I frame up to the next I frame comprise a group of pictures (GoP). Because P and B frames are encoded with reference to preceding and/or succeeding I/P frames, the transmission traffic follows the batch arrivals shown in Fig. 2. Here, the GoP follows a structure of size 16 such as "$I_0 P_4 B_1 B_2 B_3 P_8 B_5 B_6 B_7$ $P_{12} B_9 B_{10} B_{11} I_{16} B_{13} B_{14} B_{15} \ldots$" In contrast, video frames are decoded and displayed at the receiver in a reorganized order.

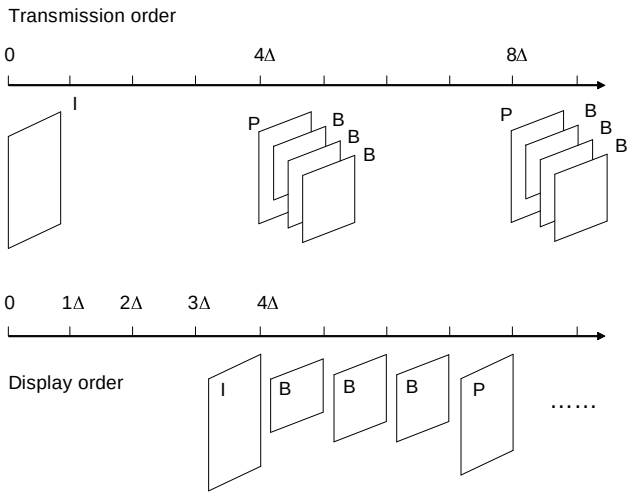

Fig. 2. Transmission and display orders of video frames.

In the literature, there has been extensive work modeling the varying rate and frame size of video traffic [5,11]. To capture both frame size variation and autocorrelation, we extend the Markov-modulated Gamma-based model (MMG) proposed in [11] for performance analysis. For a video stream consisting of a GOP sequence, video clips are grouped into a small number of shot classes depending on the GOP size. The size of I, P, and B frames in a class is modeled by an axis-shifted Gamma distribution. In the original MMG model, the GOP size boundaries for classification are geometrically separated. As observed in [8], the size of video frames based on H.264 exhibits heavy-tailed property. That is, extremely large frames exist with a non-negligible probability. To discern differences for large-size video clips in classification, we propose to use the following sigmoid function to determine the class boundaries:

$$x_i = \frac{1}{1 + e^{-\alpha_s \cdot (i - \beta_s)}}, \qquad i = 1, 2, ..., K+1 \tag{2}$$

where K is the number of video classes. As this sigmoid function takes values within $(0, 1)$, we map the frames size in the range of $\left[X_{min}, X_{max}\right]$ such that

$$x_1 = \frac{X_{min}}{\theta_s \cdot X_{max}}, \qquad x_{K+1} = \frac{1}{\theta_s} \tag{3}$$

where θ_s $(0 < \theta_s < 1)$ is a scale factor.

Take a video trace coded with single layer H.264/AVC as an example [13]. We choose the video sequence of *Tokyo Olympics* with a common intermediate format (CIF) resolution (352×288), a fixed frame rate at 30 frames/s, a GoP size of 16 with 3 B frames between I/P key pictures, and a quantization step size indexed at 24. Fig. 3 shows the size boundaries to classify GOPs according to a geometric function or a sigmoid function, respectively. As seen, the S-shaped size boundaries can also differentiate differences when the GOP size is very large.

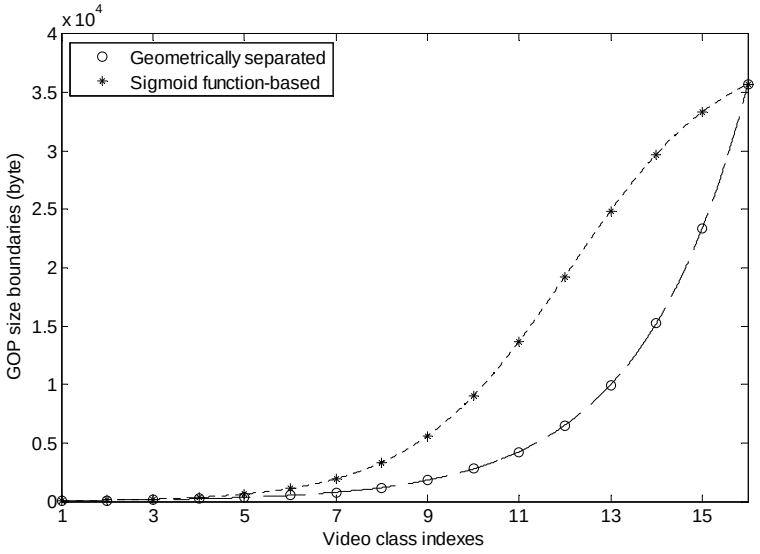

Fig. 3. Size boundaries to classify video clips.

According to the size boundaries, video clips are classified into K shot classes and the transition probability p_{ij} from class i to class j can be estimated from normalized relative frequency of transitions:

$$p_{ij} = f_{ij}/f_i \tag{4}$$

where f_{ij} 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 matrix of transition probabilities, denoted by P, can be translated into a corresponding infinitesimal generating matrix in a continuous-time domain, denoted by M, as follows

$$M = g(P - I) \tag{5}$$

where g is the rate of GOPs and I is the identity matrix.

In the original MMG model [11], the size of each type of frames in a class is modeled with an axis-shifted Gamma distribution. To render tractable analysis, we decouple the flow-level and packet-level traffic models. For each video class (state), the video traffic is considered as a fluid flow of rate γ_i $(i = 1, 2, ..., K)$. At the finer packet-level, video frames generated in a burst are fragmented into packets for transmission. We use a batch arrival process to model the packet-level traffic.

3 Multipath Transmission Exploiting Multihoming Capability

As shown in Fig. 1, a multihomed wireless device is assumed to capable of accessing three different networks simultaneously. In this section, we will discuss

how to split an elastic or streaming flow across multiple access networks. A novel cross-layer multipath transmission scheme is proposed to aggregate available bandwidth and enhance QoS provisioning.

3.1 Flow Splitting for Elastic Traffic

For an elastic flow defined in (1), due to the heavy-tailedness and high variability, it is difficult to determine how much bandwidth is required exactly to satisfy a transfer delay bound. To mitigate the variability and bandwidth requirement, we can split an elastic flow across three available access depending on a probability q_i. The purpose is to decompose a heavy-tailed flow into the superposition of a number of light-tailed substreams. A recursive algorithm is proposed in [3] to fit heavy-tailed distributions with a finite mixture of exponentials, i.e., a hyper-exponential distribution. Here, we approximate the heavy-tailed Weibull distribution in (1) with a three-stage hyper-exponential distribution, in which each exponential component corresponds to the size of a substream flow over an available access. That is,

$$\hat{f}_e(x) = \sum_{i=1}^{3} q_i e^{-\varphi_i x}$$
$$\sum_{i=1}^{3} q_i = 1, \quad 0 < \varphi_i < 1, \quad x > 0. \tag{6}$$

The parameters q_i and φ_i can be derived by fitting the first three moments. For the Weibull distribution in (1), the k^{th} raw moment is given by

$$m_k = \mathrm{E}\left[L_e^k\right] = \alpha_e^k \Gamma\left(1 + \frac{k}{\beta_e}\right). \tag{7}$$

The approximation is sufficient for engineering applications.

3.2 Data Stripping for Video Streaming

According to video codec, a number of B frames are generated between two key I/P frames depending on the target encoding bit rate. As shown in Fig. 2, each video traffic burst consists of an I or P frame and a number of B frames. For example, a video trace with a GOP size of 16 can have 0 to 15 B frames in a traffic burst. Suppose three networks are available with data rates at R_1, R_2, and R_3, respectively, where $R_1 \geq R_2 \geq R_3$. Since I/P frames are usually of a large size and important reference to decode other related frames, we consider assigning the single I/P frame in each burst toward the network with the highest data rate R_1. Next, we sort the B frames of a burst in an ascending order and denote the i^{th} B frame size by B_i, where $i = 1, ..., S$ and S is the number B frames in a burst. Then, the first k of these B frames are distributed to the second network such that

$$\frac{\sum_{i=1}^{k} B_i}{\sum_{i=1}^{S} B_i} \approxeq \frac{R_2}{R_2 + R_3}. \tag{8}$$

The remaining B frames go to the third network.

Exploiting cross-layer information such as frame type and size and channel rate, we split a video streaming flow into three substreams. Each substream requires a smaller bandwidth from the associated network. In other words, the fractional resources available in each network are aggregated to support a bandwidth-demanding streaming flow. As discussed in Section 2.2, a Markov-modulated Gamma-based model can be used to characterize the video substream. Here, we use the leaky bucket algorithm in [12] to derive the effective bandwidth of video streams to bound data loss probability. The leaky bucket algorithm emulates data transmission over a channel via introducing a virtual token pool which has a finite size B_T and generates tokens at a rate r. A data buffer of size B_D can also be involved to mitigate data loss due to traffic variations. A token is required from the token pool to transmit a data unit and data loss occurs if the data buffer is full and the token pool is empty. Since data loss probability $P_{L,s}$ depends on token generation rate r, the effective bandwidth of a given video substream can be derived to bound $P_{L,s} \leq \varepsilon$.

3.3 Flow-Level Performance

For each available network, we consider a restricted access mechanism as in [16] to share the bandwidth between elastic and streaming substreams. For presentation brevity, we take one network as an example and omit subscripts for the network in the following. According to the restricted access mechanism, for a network with a capacity C, streaming flows are offered a preemptive priority over elastic flows and only occupies up to a minimum amount of bandwidth R ($R < C$). Based on the leaky bucket analysis, an effective bandwidth of b is derived for streaming substreams to bound data loss probability. Thus, the network can admit N_s streaming flows at most, where $N_s = \lfloor \frac{R}{b} \rfloor$. As elastic flows can adapt to varying bandwidth, all the bandwidth unused by streaming traffic is shared equally by active elastic flows. Totally N_e elastic flows are admissible to bound flow transfer delay.

For the flow-level performance analysis, it is assumed that streaming and elastic substreams arrive as Poisson processes with a mean rate λ_s and λ_e, respectively. The duration of streaming flows is considered to be exponentially distributed with a mean $1/\mu_s$. A heavy-tailed elastic flow is split into substreams of an exponentially distributed size, as discussed in Section 3.1. Hence, the flow-level performance can be evaluated with a two-dimensional Markov chain, which is nearly decomposable under a quasi-stationary assumption [2]. That is, the number of elastic flows is assumed to evolve rapidly with respect to the streaming traffic and attain a stationary regime. This is due to the fact that elastic and streaming flows evolve at different time scales. The time required to transfer a data file such as a Web page should be bounded within seconds, whereas the mean duration of video streaming flows is usually in the order of minutes. In the

quasi-stationary regime, the two-dimensional Markov chain is decomposed into an $M/M/K/K$ queue for streaming flows and an $M/M/1$ processor sharing (PS) queue for elastic flows. Thus, we have the streaming and elastic flow blocking probabilities ($P_{B,s}$ and $P_{B,e}$, respectively) and flow transfer delay (T_e) as follows

$$P_{B,s} = \psi_0 \frac{\rho_s^{N_s}}{N_s!} \tag{9}$$

$$P_{B,e} = \sum_{i=0}^{N_s} \psi_i \frac{\big[1-\rho_e(i)\big]\rho_e(i)^{N_e}}{1-\rho_e(i)^{N_e+1}} \tag{10}$$

$$T_e = \sum_{i=0}^{N_s} \psi_i \frac{\rho_e(i)^{N_e+1}\big[N_e\rho_e(i) - N_e - 1\big] + \rho_e(i)}{\lambda_e\big[1-\rho_e(i)^{N_e}\big]\big[1-\rho_e(i)\big]} \tag{11}$$

$$\rho_s = \frac{\lambda_s}{\mu_s}, \quad \rho_e(i) = \frac{q\lambda_e}{\varphi(C - i\cdot b)}, \quad i = 0, 1, ..., N_s \tag{12}$$

where ψ_i denotes the steady-state probability of i ongoing streaming flows, given by

$$\psi_i = \frac{\rho_s^i}{i!}\left[\sum_{i=0}^{N_s}\frac{\rho_s^i}{i!}\right]^{-1}. \tag{13}$$

The traffic parameters q, φ, and b are derived as in Section 2.1 and 2.2. Considering flow splitting across multiple available networks, we have the overall flow blocking probabilities and flow transfer delay

$$P_{B,s} = 1 - \prod_{i=1}^{3}\Big[1 - P_{B,s}^{(i)}\Big] \tag{14}$$

$$P_{B,e} = 1 - \prod_{i=1}^{3}\Big[1 - P_{B,e}^{(i)}\Big] \tag{15}$$

$$T_e = \max\big[T_e^{(i)}\big]. \tag{16}$$

3.4 Packet-Level Performance

In Section 3.3, we analyze the flow-level performance of the proposed scheme and the discreteness of data units are neglected. For elastic services such as file transfer, user QoS experience is more concerned with flow-level dynamics. In contrast, streaming services are also sensitive to finer packet-level performance due to real-time playback during delivery. In this section, we further evaluate the packet-level performance of video streaming traffic.

As seen in Fig. 2, video frames are generated in burst according to coding and compression algorithms. For each traffic burst, a random number of video frames are disseminated toward each network channel. We assume that these application data are segmented into fixed-size packets for transmission. The transmission time of the packets, denoted by τ, is used to discretize the time scale for analysis

purpose. Hence, the packet transmission process can be modeled by a $D^{[A]}/D/1$ queueing system, which has a constant batch interarrival time, general batch-size distribution, and deterministic service rate. As the size of video frames can be modelled with a Gamma distribution [8,11], we characterize the batch size with a negative binomial distribution, which is a discrete analog of Gamma distribution. The probability mass function (PMF) of the batch size (A) is then given by

$$f_s(k) = \mathrm{P}[A = k] = \binom{k+r-1}{r-1}(1-\eta)^r \eta^k \tag{17}$$
$$r > 0, \quad 0 < \eta < 1, \quad k = 0, 1, ...$$

where the binomial coefficient

$$\binom{k+r-1}{r-1} = \frac{(k+r-1)(k+r-2)...(r)}{k!}. \tag{18}$$

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

$$\mathrm{E}[A] = r\frac{\eta}{1-\eta}, \qquad \mathrm{Var}[A] = r\frac{\eta}{(1-\eta)^2}. \tag{19}$$

Following the probability generating function (PGF) technique in [6], we can evaluate the statistics of packet transfer delay through a $D^{[A]}/D/1$ queueing system. Let Q_i denote the number of backlog packets in the transmission buffer at the end of time slot i. The evolution of buffer occupancy is thus given by

$$Q_i = \Big[Q_{i-1} + A_i - 1\Big]^+. \tag{20}$$

When the system converges to an equilibrium ($i \to \infty$), the queue occupancy Q reaches a steady state defined by a PMF function $q_s(k)$. The following generating function [6] can be obtained by multiplying (20) by z^k and summing over k

$$Q(z) = \sum_{k=0}^{\infty} q_s(k) z^k = \frac{(1-\zeta)(z-1)}{z - A(z)} \tag{21}$$

where ζ is the batch (traffic burst) arrival rate per time unit and $A(z)$ is the PGF of the batch arrival size A, given by

$$A(z) = \sum_{k=0}^{\infty} f_s(k) z^k = \left[\frac{\eta}{1-(1-\eta)z}\right]^r. \tag{22}$$

Hence, the k^{th} factorial moment of queue occupancy (Q) can be obtained from (21) as follows

$$\Omega_k = \mathrm{E}\Big[Q(Q-1)...(Q-k+1)\Big] = \lim_{z \to 1^-} \frac{d^k Q(z)}{dz^k}. \tag{23}$$

According to the generalized Little's formula [4], we have the k^{th} raw moment of the queueing delay

$$W_k = \frac{\Omega_k}{\hat{\zeta}^k}. \tag{24}$$

In particular, the mean and variance of packet transfer delay (in seconds) is obtained as

$$T_s = \tau \left(\frac{Q^{'}(1^{-})}{\hat{\zeta}} + 1 \right) \tag{25}$$

$$\sigma_{T_s}^2 = \tau^2 \left[\frac{Q^{''}(1^{-}) + Q^{'}(1^{-}) - \left(Q^{'}(1^{-})\right)^2}{\hat{\zeta}^2} \right]. \tag{26}$$

4 Numerical Results

In this section, we present numerical results to evaluate the performance of our proposed multipath transmission scheme. Table 1 gives the system parameters for numerical analysis. Assuming the average elastic flow size is $\mathrm{E}[L_e]$ = 1280 KB, we take the shape and scale parameter in (1) as $\alpha_e = 0.7$ and $\beta_e = 1011.2$. Fitting the first three moments, we can obtain the parameters of the

Table 1. System parameters for numerical analysis.

Symbol	Value	Definition
$\mathrm{E}[L_e]$	1280	Mean elastic flow size (KB)
α_e	0.7	Shape parameter of elastic flow size L_e
β_e	1011.2	Scale parameter of elastic flow size L_e
α_s	0.5856	Sigmoid function parameter for segmentation
β_s	12.0683	Sigmoid function parameter for segmentation
θ_s	1.1	Scale factor to segment video streaming flow
S	7	Number of B frames between two I/P frames
K	6	Number of video classes for traffic modelling
ε	0.01	Upper bound for data loss probability
λ_s	$0.001 \sim 0.02$	Mean arrival rate of streaming flows (/s)
μ_s^{-1}	30	Mean duration of streaming flows (min)
λ_e	$0.8 \sim 0.1$	Mean arrival rate of elastic flows (/s)
C_1	40	Available bandwidth of network 1 (Mbit/s)
C_2	20	Available bandwidth of network 2 (Mbit/s)
C_3	12	Available bandwidth of network 3 (Mbit/s)

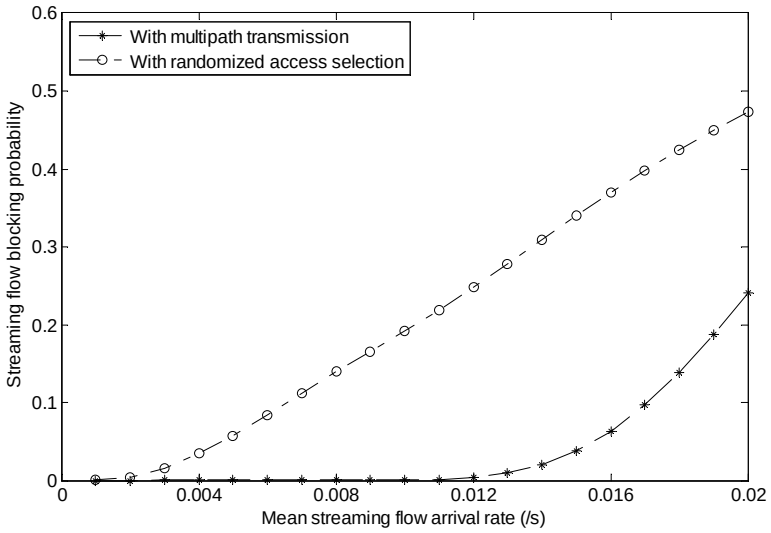

Fig. 4. Flow blocking probability of streaming traffic ($P_{B,s}$).

hyper-exponential distribution for the superposition of substreams: $q_1 = 0.1186$, $q_2 = 0.4314$, $q_3 = 0.4500$, $\varphi_1 = 2.7799 \times 10^{-4}$, $\varphi_2 = 0.0010$, and $\varphi_3 = 0.0011$ (kbit^{-1}).

For video streaming flows, we consider H.264/AVC video sequences of *Tokyo Olympics* from the video trace library of Arizona State University [13]. These video sequences have a CIF resolution, a fixed frame rate at 30 frames/s, a GoP size of 16, and 7 B frames between two I/P key pictures. The quantization level varies with the step size and a higher quantization index (between 0 and 51) results in a lower encoding bit rate. The proposed data stripping algorithm in Section 3.2 exploits the coding and compression structure to enable multipath transmission. To bound data loss probability at the flow level, the effective bandwidth can be derived with the leaky bucket algorithm [12]. For instance, at the quantization level 42, the effective bandwidth requirements of the three video substreams to upper bound data loss probability by 0.01 are $b_1 = 658.9$ kbit/s, $b_2 = 413.5$ kbit/s, and $b_3 = 215.3$ kbit/s, respectively.

4.1 Flow-Level Performance

Fig. 4 and Fig. 5 illustrate the flow-level performance of the proposed multipath transmission scheme evaluated by (9 - 14) and that of a randomized access selection algorithm [15]. For the randomized selection algorithm, an incoming flow requests admission to an available network with a probability ϑ_i. These selection probabilities ϑ_i can be determined to maximize the acceptable traffic load to the overall integrated system. As seen in Fig. 4, multipath transmission significantly reduces the blocking probability for streaming flows. Although an ideal access selection is expected to choose a most desirable network for an

incoming traffic flow, it is possible that none individual network satisfies the overall bandwidth requirement. Multipath transmission can effectively aggregate available bandwidth across multiple networks to serve split substreams requiring smaller bandwidth and achieves a much lower blocking probability.

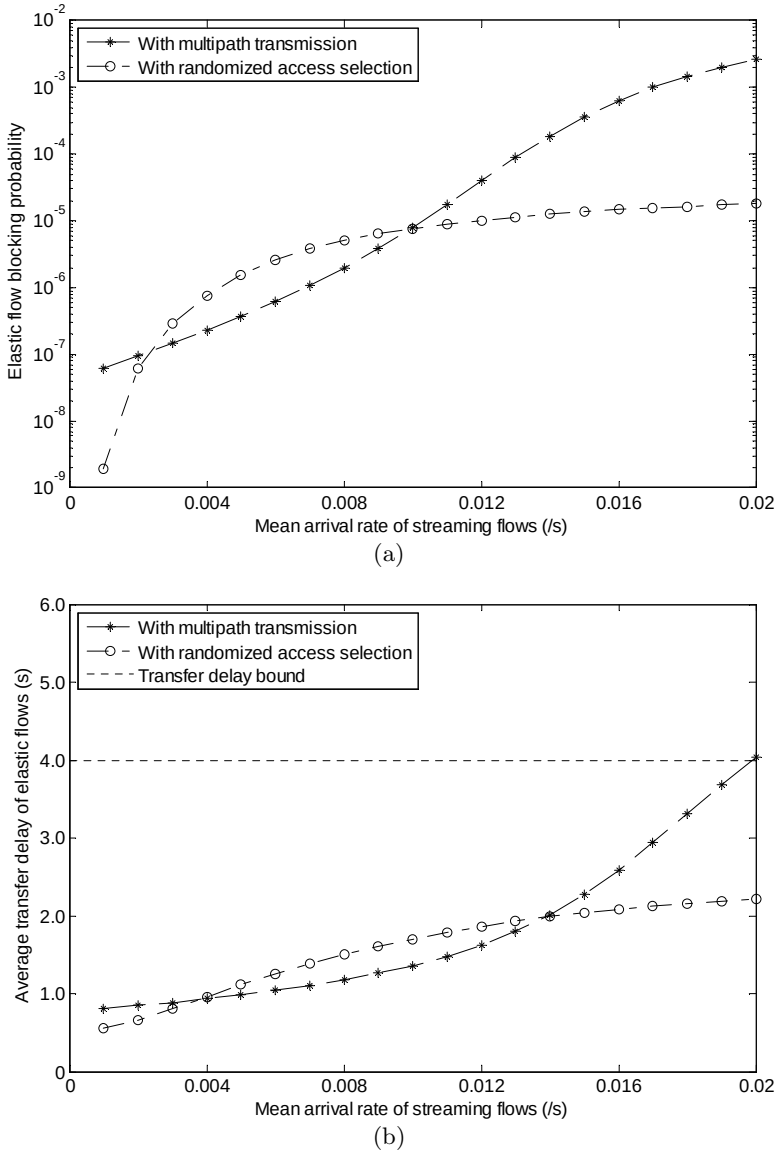

Fig. 5. Flow-level performance of elastic traffic. (a) Flow blocking probability ($P_{B,e}$). (b) Flow transfer delay (T_e).

As shown in Fig. 5, the randomized selection provides a better performance for elastic flows in some cases. With the randomized selection, streaming flows are more likely to be declined from the system. Consequently, more bandwidth unused by streaming traffic can be shared by elastic flows. As a result, a lower blocking probability and transfer delay is achievable for elastic flows at the expense of a significantly higher blocking probability for streaming traffic. Nevertheless, the performance of elastic flows with multipath transmission is well acceptable.

4.2 Statistics of Packet Transfer Delay

As shown in Fig. 2, video frames are generated in burst and segmented into packets for transmission. To ensure smooth playback at the receiver, the burst of application data need to be delivered at a speed higher than the burst arrival rate. Considering the video traces at 30 frames/s, we have the delay bound at 266.7 ms, given 7 B frames between two key I/P frames. Fig. 6 shows the average transfer delay of video streaming bursts. The statistics of packet transfer delay can be evaluated with the the batch arrival queueing process and PGF technique in Section 3.4. It can be seen in Fig. 6 that multipath transmission effectively reduces the average transfer delay of video frames. A primary reason for the performance gain is that the traffic burstiness is balanced and mitigated with flow splitting.

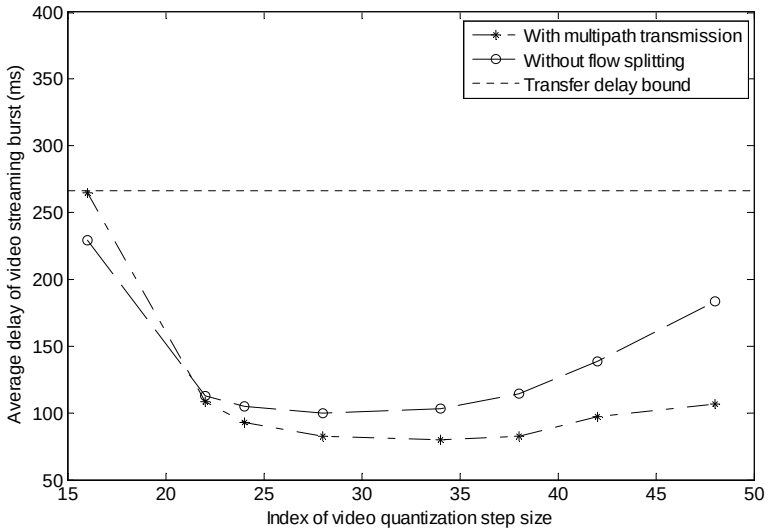

Fig. 6. Average transfer delay of video streaming bursts at different quantization levels.

5 Conclusions and Future Work

In this paper, we propose a cross-layer multipath transmission scheme for elastic and streaming flows. The proposed scheme exploits application-layer and network-layer information to enable flow splitting and data stripping across multiple heterogeneous networks. A heavy-tailed elastic flow is thereby decomposed into multiple substreams of an exponentially distributed size. A much smaller effective bandwidth is required for multiple associated networks to support video substreams. Moreover, we evaluate the flow-level and packet-level performance in terms of flow blocking probabilities and transfer delays. Our analytical approach takes into account the heavy-tailed size of elastic flows and burstiness of video streaming flows. The $D^{[A]}/D/1$ queuing process with batch arrivals is considered to analyze packet transfer delay of video streaming traffic, which is sensitive to this finer packet-level performance due to real-time playback.

Numerical results are presented to demonstrate the performance of the proposed multipath transmission scheme. A significant performance gain is particularly observed for bandwidth-demanding video streaming traffic. The QoS enhancement is attributed to the effective aggregation of available bandwidth across multiple networks. In our future work, we will investigate how to adapt multipath transmission with network variations. Especially, in high mobility conditions, link capacities may fluctuate rapidly due to severe fading or switch of network attachment points. Hence, the flow splitting needs to be dynamically adapted and cope with network variations.

References

1. 3GPP. Transparent end-to-end packet-switched streaming service (PSS); protocols and codecs. 3GPP TS 26.234 V9.3.0, June 2010.
2. F. Delcoigne, A. Proutire, and G. Rgni. Modeling integration of streaming and data traffic. *Perform. Eval.*, 55(3-4):185–209, Feb. 2004.
3. A. Feldmann and W. Whitt. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Perform. Eval.*, 31(3-4):245–279, Jan. 1998.
4. D. Gross and C. M. Harris. *Fundamentals of Queueing Theory*. Wiley, 1974.
5. D. P. Heyman. The GBAR source model for VBR videoconferences. *IEEE/ACM Trans. Networking*, 5(4):554–560, Aug. 1997.
6. A. Y.-M. Lin and J. A. Silvester. On the performance of an ATM switch with multichannel transmission groups. *IEEE Trans. Commun.*, 41(5):760–770, 1993.
7. L. Ma, F. Yu, V. C. M. Leung, and T. Randhawa. A new method to support UMTS/WLAN vertical handover using SCTP. *IEEE Wireless Commun. Mag.*, 11(4):44–51, Aug. 2004.
8. D. M. B. Masi, M. J. Fischer, and D. A. Garbin. Video frame size distribution analysis. *The Telecommunications Review*, 19:74–86, Sept. 2008.
9. D. Niyato and E. Hossain. Dynamics of network selection in heterogeneous wireless networks: An evolutionary game approach. *IEEE Trans. Veh. Technol.*, 58(4):2008–2017, May 2009.

10. K. M. Rezaul and A. Pakšstas. Web traffic analysis based on EDF statistics. In *Proc. 7th Annual PostGraduate Symposium on the Convergence of Telecommunications, Networking and Broadcasting (PGNet)*, June 2006.
11. U. K. Sarkar, S. Ramakrishnan, and D. Sarkar. Markov-modulated Gamma-based framework. *IEEE/ACM Trans. Networking*, 11(4):638–649, Aug. 2003.
12. M. Schwartz. *BroadBand Integrated Networks*. Prentice Hall, 1996.
13. P. Seeling, M. Reisslein, and B. Kulapala. Network performance evaluation with frame size and quality traces of single-layer and two-layer video: a tutorial. *IEEE Communications Surveys & Tutorials*, 6(3):58–78, Third Quarter 2004.
14. P. Si, H. Ji, and F. R. Yu. Optimal network selection in heterogeneous wireless multimedia networks. *ACM/Springer Wireless Networks*, 16(5):1277–1288, Aug. 2009.
15. W. Song, Y. Cheng, and W. Zhuang. Improving voice and data services in cellular/WLAN integrated network by admission control. *IEEE Trans. Wireless Commun.*, 6(11):4025–4037, Nov. 2007.
16. W. Song, H. Jiang, and W. Zhuang. Performance analysis of the WLAN-first scheme in cellular/WLAN interworking. *IEEE Trans. Wireless Commun.*, 6(5):1932–1952, May 2007.