Cost Efficient Media Streaming Algorithms for Rate-dependent Pricing Strategies in Heterogeneous Wireless Networks

Ahmed H. Zahran and Cormac J. Sreenan
Department of Computer Science
University College Cork, Ireland
[a.zahran, cjs]@cs.ucc.ie

Abstract

The future of wireless networking is envisioned as an integrated system of wireless radio access technologies with heterogeneous features. This heterogeneity combined with the characteristic diversity of provided services create promising opportunities for improving the utility of both operators and users. In this paper, we investigate the opportunity to reduce the average cost of streaming sessions by benefiting from the system embedded heterogeneity and streaming application buffering capability. First, we analyze the optimal streaming strategy for a theoretical infinite session. Based on this analysis, we propose pseudo-optimal and greedy-optimal adaptive media streaming algorithms for heterogeneous wireless networks. The performance of these algorithms is compared to a naive greedy streaming approach using NS2 simulations. The results show that the greedy-optimal algorithm reduces the average session cost down to 73.9% *of the average cost incurred on using greedy algorithm. This cost saving is realized at an insignificant increase in signaling load and session blocking probability. Hence, we strongly recommend the developed greedy-optimal algorithm for media streaming in next-generation heterogeneous wireless networks.*

1. Introduction

The mobile industry has witnessed a dramatic technological evolution over the past fifteen years. Currently, 3G technologies, such as UMTS and CDMA2000, enable universal digital packet-switched voice and data services. In parallel, disruptive access technologies, such as HiperMan, WiMAX, WiFi, and HiperLan, support cheaper and higher data rate service within their limited coverage. Hence, the integration of these technologies is foreseen as a possible approach to provide high data rate services to users in densely populated areas, where short-range technologies are becoming increasingly pervasive [1]. Additionally, this integration is strongly propelled by the interest in introducing new services such as gaming, conferencing, and media streaming. This combination of applications and access technologies boost the system characteristic diversity creating a real heterogeneous system.

The embedded system heterogeneity creates promising opportunities for both network operators and system users to improve their utility. For example, the users can benefit from the cost diversity in different networks to reduce their bills. For In [2], Liang et al. employ proactive document prefetching in cheap networks using a probabilistic framework. This excessive interest in cheaper networks combined with the interest in supporting QoS requirements of real-time applications in these networks represent the main drives for adopting more advanced pricing schemes. In [3], Badia et. al. propose a *rate-dependent* pricing scheme for WLANs to support voice over IP QoS requirements. In this work, we investigate real-time media streaming in heterogeneous settings considering rate-dependent pricing schemes.

Streaming applications have a special buffering capability by which the application can save future parts of the stream to be played-out later. Combining this buffering capability with the cost and resource variation of wireless access technologies in integrated systems, one can clearly anticipate a possible opportunity for cost savings by buffering the stream in cheaper resource-rich intermittent technologies, e.g. WLANs in a 3G-WLAN integrated system. In this paper, we launch a new initiative toward investigating the optimal cost streaming strategy in heterogeneous systems. We propose new heuristic rate control algorithms for streaming applications in heterogeneous networks based on the optimal streaming policy of a theoretical infinite session. The proposed algorithms, especially the greedy-optimal algorithm, have shown noticeable savings in the average session cost in comparison to a natural naive greedy streaming strategy. Using NS2 simulations, we show that this cost

This research was made possible thanks to the Science Foundation of Ireland.

978-0-7695-3333-9 /08 $25.00 © 2008 IEEE
DOI 10.1109/NGMAST.2008.32

savings are realized at the expense of insignificant signaling and session blocking probabilities.

The rest of this paper is organized as follows. Section 2 presents important background and related work. The system model, problem statement, and the optimal streaming strategy guidelines are developed in Section 3. We then present the proposed media streaming algorithms for heterogeneous wireless networks in Section 4 followed by their performance evaluation in Section 5. Finally we conclude and present future research work directions in Section 6.

2. Background and Related Work

Audio and video streaming popularity is significantly growing in both wired and wireless domains. Many techniques at different system levels have been proposed in the literature [4] to improve the user streaming experience. Initial playout latency has been proposed in order to avoid playout interruption due to transmission errors. Adaptive media playout is also proposed to decrease initial latency and compete transmission delay jitter [5]. Layered video compression is another technique that enables the user to have different levels of stream quality of service (QoS). Last but not least, rate control enables changing the streaming rate according to the available system resources. Note that rate control can be source-based, receiver-based, or hybrid [4].

Few papers have addressed media streaming in heterogeneous systems. Many of these papers [6, 7] focus on improving the session quality through adjusting the streaming rate according to the available network resources. In [8], using an experimental testbed, the authors demonstrate the possibility of performing seamless policy-triggered vertical handoff (VHO) using Mobile IPv4 while running video sessions. These papers mainly focus on benefiting from bandwidth variations during VHOs ignoring other aspects of system characteristic diversity. On contrary, we additionally consider cost changes that naturally accompany VHOs.

In this work, we investigate possible average session cost reduction approaches as a novel research issue in heterogeneous wireless systems. The cost reduction is feasible by controlling the streaming speed to buffer data in the cheaper intermittent network and minimize the utilization of the expensive one. Practically, stream rate control can be performed using real-time streaming protocol (RTSP) [9]. Additionally, proactive VHO algorithms, e.g. [10], combined with the media independent handover framework IEEE 802.21 [11] are strongly recommended for efficient stream control framework. Using this framework, the application can track the received signal strength of the current base station to predict the network transition instant. Hence, most of the stream control signaling is transmitted over the cheap network to reduce signaling cost.

3. Problem Statement and Solution approach

In this section, we present the theoretical basis for our proposed algorithms. We first present the system model followed by the original problem formulation. We then present an optimization framework for a theoretical infinite session whose solution is used to establish the main guidelines for our proposed heuristic algorithms.

3.1. System Model

In our model, we assume a two-tier integrated wireless system composed of networks N_x for $x \in \{u, i\}$, where u and i correspond to the technologies that provide universal and intermittent coverage respectively. Each network has a *non-decreasing rate-dependent* cost profile [3], denoted as $\chi_x(r_x)$ where r_x represents the data service rate in network N_x. Typically, r_x is non-negative and upper-bounded by a maximum service rate of r_{xmax} in network N_x, i.e. $r_x \in [0, r_{xmax}]$. Note that the design of the network cost profiles in heterogeneous integrated systems is a non-trivial task because the cost plays a role in inter-network handoff decision.

Figure 1 shows a typical scenario for a streaming session in a two-tier heterogeneous system. Generally, we assume that the session duration follows a generic heavy tailed distribution [12]. Clearly, during the session lifetime, the user encounters different events including session start, technological transitions, and session end as it traverses dual and single coverage zones at t_i's. The durations spent by the user in different coverage combinations τ_i are assumed to have generic phase-type (PH) distributions. Note that $\overline{\tau_i}$ represents the residual time distribution of τ_i. To this end, it is worth noting that the parameters of these distributions can be estimated using the user mobility information that are collected on the run using similar framework to that proposed in [13] for mobility modeling in heterogeneous systems. At the aforementioned instants, the streaming application buffer status, denoted as x_k, is governed by the following differential equation

$$x_{k+1} = x_k + (r_x - r_o)(t_{k+1} - t_k)\,, \tag{1}$$

where r_o represents the average play-out data rate. That is to say that the buffer status at any instant equals the buffered information at the previous instant in addition to the difference between the downloaded and consumed data.

3.2. Original Problem

Based on this model, our goal is to find the optimal streaming policy in each network that minimizes the average session cost J_{av}. In this context, the average session

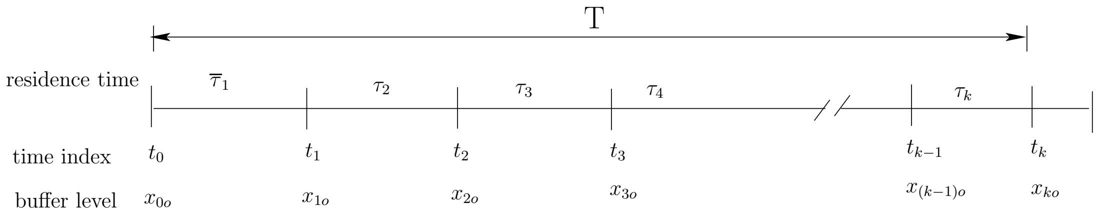

Figure 1. Streaming Session

cost is expressed as

$$J_{av} = E\left[\frac{J}{T}\right] = E\left[\frac{\sum_{i=0}^{n} J_i}{\sum_{i=0}^{n} \tau_i}\right] , \qquad (2)$$

where J represents the total session cost and J_i represents the session cost during the time interval $[t_{i+1},\, t_i)$, i.e. τ_i. To this end, it is worth noting that there are practical bounds on the streaming rate in each stage. These bounds are an upper bound, r_{imax} determined by the network as the maximum average download rate per user and a lower bound r_{imin} determined to satisfy the play-out quality by maintaining the buffer level above a pre-specified initial play-out latency, ρ.

Our objective is to determine the optimal streaming policy that minimizes (2) such that the aforementioned constraints are satisfied. To this end, the presented problem fits sequential decision stochastic programming, whose solution is based on computationally expensive scenario generation techniques. Clearly, mobile devices with their limited processing capacities are not capable of handling these computations within the required VHO delay. Hence, in the following analysis, we simplify the original problem assuming infinite session duration, i.e. $T \to \infty$ or $n \to \infty$, seeking intuitions to develop a heuristic solution for the optimal streaming policy. This approach is mainly propelled by the interest in reducing the average cost of long sessions for which cost saving benefits are viable.

3.3. Theoretical Session

Under the infinite session duration assumption, the process turns to be a delayed Markov regenerative process [14] in which transition instants to a specific network N_i represent regenerative epochs at which the process satisfies Markov property. In [14, Theorem 7.6-7], it is shown that under bounded cost assumption, there exist a stationary policy π_s^* that minimizes the average cost. Additionally, it is also shown [14, Theorem 7.5] that the average session cost given a specific initial state x for any stationary policy π_s is expressed as

$$J_{av} = \frac{E_{\pi_S}[J_\tau | x]}{E_{\pi_s}[\tau | x]} , \qquad (3)$$

where J_τ represents the total cost incurred during a regenerative cycle duration τ. That is to say that the optimal average session cost estimated over the entire session duration under a specific stationary policy equals the optimal average session cost estimated over one regenerative cycle using the same policy. Hence, our objective is to determine a stationary policy π_s^* that minimizes eq. (3).

In a two-tier integrated system, each cycle composes of two stages spent in unique and dual coverage zones. Hence, the denominator of eq. (3) equals the expected value of the sum of the unique and dual coverage zone residence times. Clearly, this value only depends on the user mobility parameters. Hence, the expected regenerative cycle duration represents the expected value of the sum of two PH distributions. According to properties of PH-type distributions [15], this sum is also a well-defined PH distribution. The math details of this operation is omitted here due to page limits. However, interested readers are referred to [15]. To sum up, minimizing the average session cost implies minimizing the numerator of eq. (3).

Generally, the numerator of (3) can be expressed as

$$\begin{aligned} E_{\pi_s}[J_\tau | x] &= E[J_\tau | x] \\ &= \mu_d \chi_d(r_d) + \mu_u \chi_u(r_u) \\ &= \sum_{i=d,u} \mu_i \chi_i(r_i) , \end{aligned}$$

where μ_i represents the mean zone residence time in N_i. Generally, this problem fits the constrained dynamic programming framework [16]. However, this type of problems is usually associated with overwhelming computations that may not fit the real-time decision requirement in the considered problem. Nevertheless, we benefit from the problem characteristics, specifically the non-decreasing nature of the cost functions $\chi_i(r_i)$, by solving the unconstrained version of the problem and then tune the solution of the unconstrained problem version to fit the original problem constraints. This tuning is simply done by setting the estimated streaming rate to the nearest feasible value if it falls outside the acceptable solution domain. Consequently, the problem degenerates to the basic dynamic programming problem [16]

$$\min_{r_i} \sum_{i=d,u} \mu_i \chi_i(r_i) .$$

Following the dynamic programming solution framework, the problem is solved sequentially backward considering one stage each step. Due to the monotonic nature of the cost profile, the minimization of the last stage implies streaming at the minimum possible rate, i.e. stop streaming that consequently results in zero cost. However, stopping streaming may lead the buffered content to drop below ρ and consequently contradicts with the QoS constraint. Hence, the rate choice at this stage is mainly determined by the lower bound constraint. That is to say

$$x_{du} + (r_u - r_o)\mu_u \geq \rho \, ,$$

where x_{du} represents the expected buffer level at the network transition. Hence, the practical minimum streaming policy for the last stage is

$$r_u = r_o + \frac{\rho - x_{du}}{\mu_u} \, . \tag{4}$$

Proceeding to the next stage, we need to estimate r_d that

$$\min_{r_d} \mu_d \chi_d(r_d) + \mu_s \chi_s(r_o + \frac{\rho - x_{du}}{\mu_u}) \, . \tag{5}$$

For purpose of illustration, we assume that $\chi_d(r_d) = b_d r_d$ and $\chi_u(r_u) = a_u r_u^2 + b_u r_u$ in which a_u, b_u, and b_d can be tuned to fit different cost profiles. Substituting in (5), differentiating w.r.t r_d and equating with zero, the optimal streaming policy in the dual coverage zone can be expressed as

$$r_d = \frac{1}{\mu_d}\left(\frac{b_u - b_d}{2a_u} + (\rho - x_0) + r_o(\mu_d + \mu_u)\right) \, . \tag{6}$$

Consequently, the streaming rate in the unique coverage stage can be expressed as

$$r_u = \frac{b_d - b_u}{2a_u} \, .$$

To this end, it worth noting that the selected rate should also satisfy the non-negativity nature of the streaming rates. For example, if $b_d < b_u$, the streaming rates degenerate to $r_u = 0$ and $r_d = \frac{1}{\mu_d}\left((\rho - x_0) + r_o(\mu_d + \mu_u)\right)$. Clearly, (4) and (6) define the optimal streaming strategy for an infinite session. This strategy has two main components

- In the dual coverage, the developed policy suggests buffering sufficient data to minimize the usage of the expensive resources during this cycle.

- In the unique coverage, the application should stream at a rate that satisfy the QoS condition by the end of the expected residence time.

4. Streaming Algorithms

In this section, we present possible streaming strategies in next-generation heterogeneous wireless networks. First, we introduce a greedy media streaming (GMS) algorithm as a *natural* behaviour for the users in the heterogeneous settings. We then propose the pseudo-optimal media streaming (POMS) and greedy-optimal media streaming (GOMS) algorithms following the aforementioned optimal streaming guidelines. The main difference between both algorithms is that GOMS applies the same optimal framework in a greedy fashion.

4.1. GMS Algorithm

The main idea of GMS algorithm is to take an advantage of the cheap network by streaming at the maximum possible rate whenever this network is visited. In the expensive network, the application stops the streaming process provided that the buffered data is more than ρ. If ρ is crossed, GMS starts to stream at the nominal rate; i.e. r_o. Note that if the application is started in the expensive network, it streams at the stream average rate; otherwise, it streams at the cheap network maximum service rate.

4.2. POMS Algorithm

POMS algorithm follows the guidelines presented in Section 3.3. Hence, the application sets the streaming rate on switching to the expensive network, r_u^{po}, to

$$r_u^{po} = \min\left(r_{umax}, \, \max\left(0, r_o + \frac{\rho - x_{du}}{\mu_u}\right)\right) \, . \tag{7}$$

In (7), the max and min functions are introduced to satisfy the non-negative rate and maximum streaming rate conditions. If the buffered data drops below ρ, the application requests streaming at the stream average rate; i.e. r_o. This situation occurs when the reduced streaming rate estimated by (7) and the buffered data are insufficient to satisfy the application consumption during its residence in the expensive network.

On moving to the cheaper intermittent network, the application sets the streaming rate to the value that enables buffering the expected data consumption in the expensive network. Hence, the application sets the streaming rate to

$$r_d^{po} = \min\left(r_{dmax}, \, \max\left(0, \, \frac{(\rho - x_{ud}) + r_o(\mu_d + \mu_u)}{\mu_d}\right)\right) \, , \tag{8}$$

where x_{ud} represents the buffer status when the user moves from the unique to dual coverage. Similar to the unique coverage case, the application requests rate adjustment to

Table 1. Simulation Parameters

Param.	Value	Param.	Value	Param.	Value
r_{dmax}	2MBps	r_{umax}	50 KBps	r_o	25 KBps
a_u	3.2e-9	b_u	0	b_d	70e.6
μ_T	22min	θ_T	1.11	ρ	6 sec

the average rate if the streaming rate drops below a super buffering threshold, denoted as $\overline{\rho}$. It is worth noting that $\overline{\rho}$ is chosen to ensure that the buffered stream is at least the expected application consumption in the expensive network. Hence, $\overline{\rho}$ can be expressed as

$$\overline{\rho} = \rho + \mu_u \, .$$

This rate adjustment rule is introduced to avoid the buffer depletion below the designated buffered data amount determined by the optimal streaming guidelines.

The aforementioned rules define POMS behaviour on network switching and buffer depletion. In case of session start, POMS sets the initial streaming rate to r_o if the session starts in the expensive network. On the other hand, if the session starts in the cheaper intermittent access technology, the application streams according to (8) after replacing μ_d with $\overline{\mu}_d$, which represents the mean of the dual coverage residual time. To this end, it is worth pointing out that designated streaming rates under this policy implicitly adapts to the user mobility variation through the mean values of the unique and dual coverage zone residence times.

4.3. GOMS Algorithm

GOMS algorithm applies the same streaming rules of POMS in a greedy fashion. That is to say on moving into the cheaper intermittent network, GOMS sets the streaming rate to the network maximum service rate until it crosses $\overline{\rho}$ beyond which the application streams at the average streaming rate r_o. The key idea behind the greedy-optimal algorithm is to enforce the optimal streaming policy guidelines. Note that POMS may not have enough time to buffer the amount suggested by the optimal policy guidelines especially with the well-known high variability in mobility pattern statistics.

5. Numerical Results

In order to asses the performance of the proposed algorithms, we simulate a 3G-WLAN integrated system using NS-2 [17]. Table 1 shows the default values of the system parameters. We adopt zone residence time model [13, 18] for mobility simulations. The parameters of the proposed cost profiles provide cost incentive for the users to persist on using the 3G technology for low data rate applications.

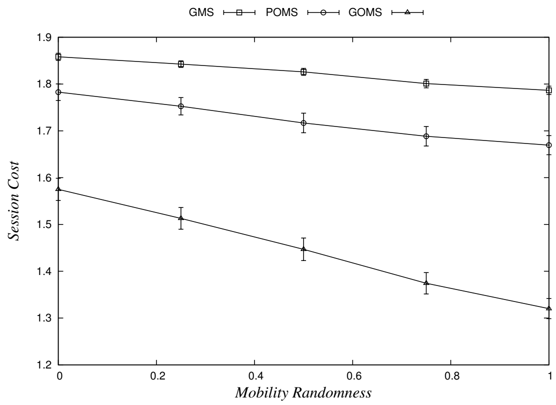

Figure 2. Session cost versus mobility Randomness

The chosen maximum network service rates represent possible rate allocation strategy of CDMA2000 and 802.11g. The session has a hyper-exponential distribution with mean μ_T and coefficient of variation θ_T. In the following figures, we compare the performance of the GMS, POMS, and GOMS algorithms for different mobility patterns that scans the complete mobility spectrum from random walkers to fluid-flow travellers. In the figures, 0 and 1 on the abscissa correspond to random-walk and fluid flow respectively. Additionally, each point represents the mean of 1000 sessions with its corresponding 95% confidence interval.

Figure 2 plots the average session cost versus different mobility patterns. Clearly, the optimized streaming algorithms reduce the average session cost for the complete mobility spectrum. More importantly, the optimal policy enforcement in GOMS significantly reduces the average session cost. Hence, we highly recommend the optimized algorithms to the natural naive GMS policy. The decreasing trend in the session cost as the pattern changes from random-walk to fluid flow mobility is a natural consequence to the corresponding drop in the variability of the residence times and the increase in the frequency of dual coverage visits as the mobility becomes less random, i.e. toward fluid-flow mobility.

Figure 3 plots the executed VHO versus the mobility patterns for the presented algorithms. In this context, executed VHO corresponds to a situation in which an active application requests resource allocation during session activity when it moves to a new network. Note that the application may not request any resource allocation as it moves to the expensive network when it has buffered data. Clearly, the figure shows a natural increasing trend in the VHO rate as the user mobility pattern changes from random walk to

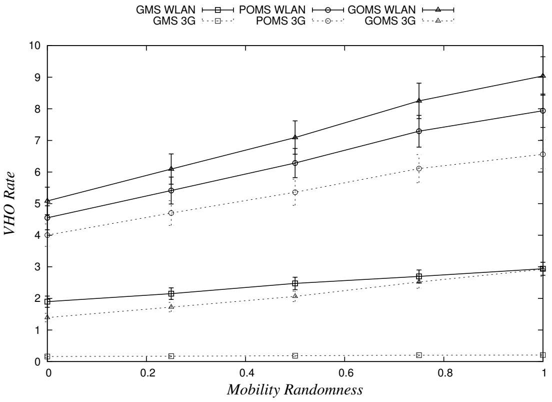

Figure 3. Session VHO versus mobility Randomness

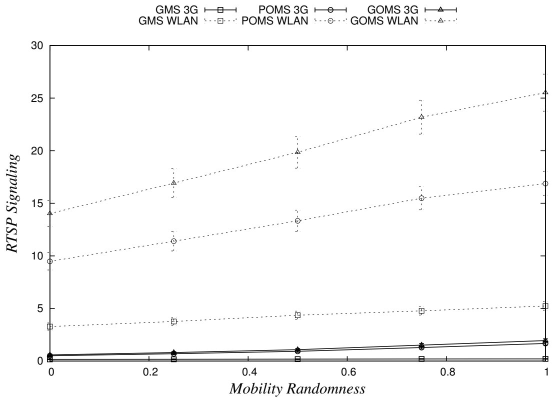

Figure 4. Session RTSP signaling versus mobility Randomness

fluid flow due to the frequency increase of zone transitions toward fluid-flow. Additionally, the figure shows that the greedy algorithm results in the lowest VHO rate in both networks due to the same reasons mentioned for GMS reduced HHO rate. The figure also shows a noticeable difference between the VHO rate in the 3G network for GOMS in comparison to POMS. This difference is interpreted by the success of the greedy behaviour in buffering sufficient data and consequently, the application rarely performs VHOs to the 3G network. In contrast, POMS in many situations performs this type of VHOs at a reduced rate to satisfy the second component of optimal policy. Finally, we would like to point out that the fewer VHOs in WLAN produced by POMS in comparison to GOMS has two possible reasons. The first is the persistence of GOMS on using WLAN in each visit at least in the nominal rate. Note that POMS may not persist on WLAN usage if it has enough buffered data, which may have been accumulated from a previous prolonged WLAN visit. The second reason is the higher forced termination probability of POMS as will be shown later. It is worth noting that similar results are obtained for horizontal handoff (HHO) signaling load. However, they are omitted due to page limitation.

Figure 4 plots the RTSP signaling in 3G and WLAN networks for the presented algorithms versus mobility randomness. For all the presented algorithms, the noticeable gap between the RTSP signaling in both networks is due to adopting a proactive VHO strategy. Hence, most of the RTSP messages are transmitted in the cheaper network just after moving into the dual zone or proactively before leaving the WLAN. Clearly, GMS results in the fewest RTSP signaling load for the same aforementioned reasons. Additionally, the higher RTSP signaling of GOMS in WLAN is due to the algorithmic multiple rate adjustments, i.e. reducing the streaming rate to the average rate after buffering the recommended amount. However, it is worth noting that this increase has insignificant cost as it is transmitted over the cheaper network.

Figure 5 plots the forced termination probabilities due to both VHOs and HHOs for different algorithms versus different mobility patterns. As a natural consequence for the fewest executed handoffs performed by GMS, it also has the lowest forced termination probabilities. On contrary, POMS results in the largest blocking probability as a natural consequence for performing more blockable handoffs, i.e. executed HHOs and VHOs to the cellular network. For similar reasons, GOMS represents a compromised approach for the two other approaches.

6. Conclusion and Future Work

The service integration of wireless access technologies combined with the huge characteristic diversity of different applications create a new heterogeneous networking paradigm that opens the door to novel research opportunities. In this paper, we investigate the optimal cost streaming strategy for media streaming in this networking paradigm. Additionally, we propose adaptive rate media streaming algorithms for heterogeneous systems. The proposed greedy optimal algorithm shows significant cost reductions to the natural naive greedy streaming approach at the expense of slight increase in signaling load and session blocking probabilities. Hence, we foresee the greedy-optimal algorithm as an initial promising heuristic solution for media streaming in next-generation heterogeneous wireless networks. As future work, we are interested in addressing several questions

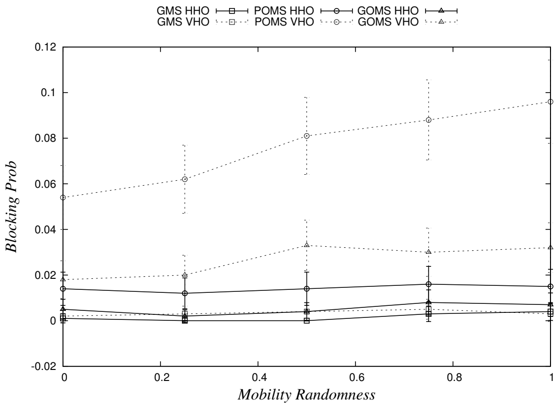

Figure 5. Session blocking probabilities versus mobility Randomness

including how does it perform under different network settings and how far it performs in comparison to the optimal streaming mechanism for heterogeneous wireless networks.

References

[1] M. M. Buddhikot, G. Chandranmenon, S. Han, Y. W. Lee, and S. M. L. Salgarelli, "Integration of 802.11 and Third Generation Wireless Data Networks," in *Proc. of IEEE INFOCOM*, San Francisco, US, Apr. 2003, pp. 503 – 512.

[2] B. Liang, S. Drew, and D. Wang, "Performance of multiuser network-aware prefetching in heterogeneous wireless systems," *to appear in ACM-Springer Wireless Networks*.

[3] L. Badia, S. Merlin, A. Zanella, and M. Zorzi, "Pricing VoWLAN services through a micro-economic framework," *IEEE Wireless Commun.*, vol. 13, no. 1, pp. 6–13, Feb. 2006.

[4] D. Wu, Y. Hou, W. Zhu, Y.-Q. Zhang, and J. Peha, "Streaming video over the Internet: approaches and directions," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 11, no. 3, pp. 282–300, Mar 2001.

[5] E. Steinbach, N. Farber, and B. Girod, "Adaptive playout for low latency video streaming," in *Proc. International Conference on Image Processing*, vol. 1, 2001, pp. 962–965.

[6] L.-J. Chen, T. G. Y. Sun, M. Sanadidi, and M. Gerla, "Adaptive video streaming in vertical handoff: a case study," in *Proc of The First Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services*, 2004, pp. 111 – 112.

[7] J.-W. Ding, C.-T. Lin, and K.-H. Huang, "ARS: an adaptive reception scheme for handheld devices supporting mobile video streaming services," in *Proc. International Conference on Consumer Electronics (ICCE '06)*, 2006, pp. 141– 142.

[8] A. Schorr, A. Kassler, and G. Petrovic, "Adaptive media streaming in heterogeneous wireless networks," in *Proc. of IEEE 6th Workshop on Multimedia Signal Processing*, 2004, pp. 506 – 509.

[9] H. Schulzrinne, A. Rao, and R. Lanphier, "Real Time Streaming Protocol," RFC 2326, April 1998.

[10] A. Zahran, B. Liang, and A. Saleh, "Signal Threshold Adaptation for Vertical Handoff in Heterogeneous Wireless Networks," *ACM/Spring Mobile Networks and Applications (MONET)*, vol. 11, no. 4, pp. 625 – 640, Aug 2006.

[11] IEEE, "Local and Metropolitan Area Networks: Media Independent Handover Services," Draft Standard, Feb. 2007.

[12] M. Li, M. Claypool, R. Kinicki, and J. Nichols, "Characteristics of streaming media stored on the Web," *ACM Trans. Inter. Tech.*, vol. 5, no. 4, pp. 601–626, 2005.

[13] A. H. Zahran and B. Liang, "A Generic Framework for Mobility Modeling and Performance Analysis in Next-Generation Heterogeneous Wireless Networks," *IEEE Commun. Mag.*, vol. 45, no. 9, pp. 92–99, Sep. 2007.

[14] S. M. Ross, *Applied Probability Models with Optimization Applications*. San Francisco: Holden-Day, 1970.

[15] G. Latouche and V. Ramaswami, *Introduction to Matrix analytic Methods in Stochastic Modeling*, ser. ASA-SIAM series on Statistics and Applied Probability. SIAM, 1999.

[16] D. P. Bertsekas, *Dynamic Programming and Optimal Control*. Athena Scientific, 2007.

[17] "Network Simulator." [Online]. Available: http://www.isi.edu/nsnam/ns/

[18] A. H.Zahran, "Modeling and Design of Next-Generation Heterogeneous Wireless Networks," Ph.D. dissertation, University of Toronto, 2007.