Threshold-Based Media Streaming Optimization for Heterogeneous Wireless Networks

Ahmed H. Zahran and Cormac J. Sreenan

Abstract—The integration of different wireless access technologies combined with the huge characteristic diversity of supported services in next-generation wireless systems creates a real heterogeneous network. In this paper, we propose a generic practical framework that optimizes media streaming in heterogeneous systems by taking advantage of cost and resource characteristic diversity of the integrated access technologies and the buffering capability of streaming applications. The proposed optimization framework represents a means to compromise the tradeoff between different performance metrics including streaming monetary cost, signaling load, and session quality. Additionally, it accommodates different design challenges including mobility randomness, limited processing capacity, and handoff delay requirements. The simulation results provide important insights on the design of pricing profiles in integrated systems. Additionally, the results show that significant cost savings can be realized using the newly proposed streaming management algorithms and optimization framework.

Index Terms—Media streaming algorithm, heterogeneous wireless network, cost optimization, service pricing.

1 Introduction

The future of wireless networking is envisioned as a converged system of different radio access technologies that have heterogeneous characteristics [1], [2], [3]. In this networking paradigm, roaming across heterogeneous networks, commonly known as vertical handoff (VHO) [4], represents a unique process that brings both design challenges and system benefits. Improving network resource utilization and enhancing user perceived Quality-of-Service (QoS) are among the foreseen benefits while keeping such transitions seamless is the main design challenge. This challenge requires careful handling for the VHO process at different layers [5], especially with the increasing penetration of revenue-generating resource-demanding services such as media streaming [6].

The integration of third-generation (3G) cellular systems and wireless local area networks (WLAN) is a typical example for next-generation heterogeneous wireless networks (NgHWN). Typically, 3G systems provide expensive universal coverage with limited resources while WLANs represent cheap resourceful alternatives within their limited coverage, commonly denoted as hotspots. This characteristic diversity creates new room for improving the user streaming experience in NgHWN by taking advantage of the possible rate control of streamed media. Generally, rate control is used to improve the user streaming experience as the network conditions change [7]. For example, the streaming application may control its buffered media playout rate to conceal occasional delays and loss variations [8], [9]. Additionally, the application can instruct the streaming server to change the streaming rate, e.g., by changing the stream encoding, in response to drastic changes in the available network resources.

In NgHWNs, changing the streaming rate as the user roams across heterogeneous networks is a principal requirement to maintain seamless VHOs due to the huge difference in the utilized bandwidth during the VHO process and the limited buffering space in 3G systems. Several studies [10], [11], [12] experimentally prove the feasibility of seamless streaming in different systems, such as 802.11b-CDMA2000 [10], 802.11b-UMTS [12], and GPRS-WLAN [11] using different adaptive cross-layer designs. This success is rooted to employing proactive VHO policies, tracking the signal deterioration in the hotspots, and proactively adjusting the streaming rate before leaving hotspots. It is worth mentioning that the media independent handover framework (IEEE 802.21) [13], supports signal strength tracking and application notification through its event and information services, respectively.

The coexistence of different access technologies drives another research avenue based on the existence of multipath diversity in NgHWN. In [14], Pan et al. employ this diversity by streaming the media in different networks to realize smooth handoff. In [15], Jurca and Frossard study rate and path allocation so that the quality at the receiver is maximized. In [16], Yang and Chen optimize the streaming quality by selecting the cell and technology used for multicasting different video stream layers to different users. Although all the aforementioned studies provide important insights on media streaming in heterogeneous systems, none of them has considered the monetary service cost in their analysis.

Access and service pricing plays a significant role in different operational aspects of any communication system, especially resource management and revenue maximization

- *The authors are with the Department of Computer Science, University College Cork, Cork, Ireland. E-mail: {a.zahran, cjs}@cs.ucc.ie.*

Manuscript received 23 Feb. 2009; revised 1 Oct. 2009; accepted 12 Oct. 2009; published online 11 Jan. 2010.
For information on obtaining reprints of this article, please send e-mail to: tmc@computer.org, and reference IEEECS Log Number TMC-2009-02-0068.
Digital Object Identifier no. 10.1109/TMC.2010.17.

[17]. This role is magnified in NgHWN because service cost is considered a principal factor in the VHO decision [18], [19]. Typically, cost heterogeneity is an inherent characteristic of NgHWN. This cost heterogeneity is considered a core motivation for several studies [20], [21]. In [20], Liang et al. employ proactive document prefetching before exiting cheaper networks using a probabilistic framework. In [21], Chen et al. model the cost of data delivery as linear programming assuming the availability of network topology and mobility information through global positioning system support. This assumption is applicable to specific scenarios such as predetermined paths in transportation systems but may not be applicable for more general scenarios.

In [22], [23], we developed new stream management algorithms that minimize the average monetary session cost of streaming applications in two-tier wireless systems under user random mobility patterns. Under this setting, the solution of the stream management problem becomes computationally expensive. Hence, it contradicts with the limited processing capacity of mobile devices and VHO delay requirements. In [22], we propose stream management heuristics based on the analysis of a theoretical infinite streaming session. The proposed heuristic results in noticeable reduction in the session cost in comparison to greedy media streaming (GMS), in which the mobile terminal (MT) downloads at the maximum possible rate in the cheaper network and avoids using the expensive network. In [23], we develop a pseudo-optimal GMS (PGMS) as an optimized version of GMS. The developed algorithm results in further cost reduction in comparison to previously proposed heuristics. These studies show a tradeoff between the streaming monetary cost, the signaling load, and session QoS.

In this paper, we propose controlled streaming to threshold and stop (CSTS) and controlled streaming to threshold and recourse (CSTR) as new streaming management policies for two-tier heterogeneous wireless systems. The proposed strategies are designed to enhance the application adaptability to different cost profiles and mobility behavior. Additionally, we present a generic analytical framework to optimize media streaming for any threshold-based streaming policies in heterogeneous systems. This framework accommodates the aforementioned practical limitations and naturally adapts to user mobility patterns. Furthermore, we present a means to tune the algorithm performance to adapt to different cost components such as monetary and signaling costs. The presented results show that CSTR and CSTS realize significant cost reductions in comparison to previously proposed algorithms.

The rest of this paper is organized as follows: Section 2 presents the system model and problem formulation. The optimization approach and proposed streaming strategies are presented in Section 3 followed by the analysis of this framework in Section 4. In Section 5, we present the simulation results and then conclude in Section 6.

2 System Model and Problem Formulation

In this section, we first present the system model and core assumptions. We then formulate the media streaming problem in heterogeneous systems pointing out associated challenges.

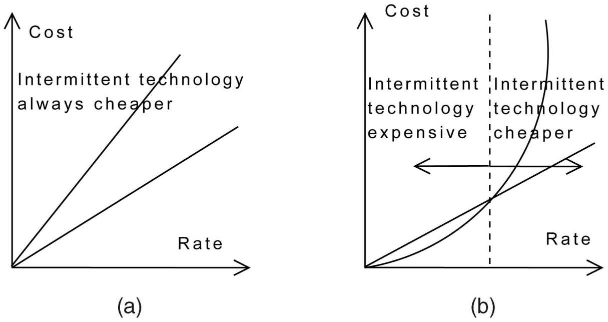

Fig. 1. Cost profiles in a two-tier system. (a) Cost Profile A and (b) Cost Profile B.

2.1 System Model

In our model, we consider a two-tier integrated wireless system composed of networks N_n for $n \in \{u, i\}$, where u and i correspond to the technologies that provide universal and intermittent coverage, respectively. For example, the 3G cellular network and WLAN, respectively, represent the technologies with the universal and intermittent coverage in a 3G-WLAN integrated system. Note that for the rest of this paper, we will consider 3G-WLAN for ease of presentation. However, the presented material can be generalized to any two-tier integrated systems. We assume that each network has a continuous nondecreasing rate-dependent cost profile, denoted as $\chi_n(r_n)$, where r_n represents the average[1] data service rate in network N_n. The assumption of monotonic service cost strategies is motivated by the interest in QoS-enabled integrated systems [26], [27], [28]. These studies propose rate dependent pricing strategies for several purposes including optimizing network performance, controlling user behavior, and reducing network congestion. Note that the commonly adopted usage-based pricing strategies are equivalent to linear rate dependent strategies for continuous usage of the resource as is the case for media streaming. To illustrate, if $\chi_n(r) = ar$, the monetary service cost for a specific period t equals $(ar)t = a(rt)$ in which (rt) represents the size of downloaded data and a represents the cost per data unit.

We typically assume that the streaming monetary cost of considered media in WLANs is cheaper than the corresponding cost in the cellular network. The reverse case is trivial as the user will always maintain its association with the universal network. Fig. 1 shows two possible pricing schemes in two-tier systems. Fig. 1a shows a case in which the service cost of the 3G cellular network is always more expensive than that of WLANs for all service rates. For similar cost profiles, handing off to the WLAN can always reduces the streaming cost. Fig. 1b shows a more interesting case in which the operators of the cellular system choose to provide a cheaper service for a specific low data rate range, which is upperbounded by the intersection point of the cost-rate curves as shown in Fig. 1b. Noting that the service cost is expected to be a major factor in VHO decision, cheaper price represents a major incentive for users not to roam to WLANs

[1] Practically, fixed service data rate may not be realizable. However, by using admission control and scheduling mechanisms that represent intrinsic parts of different standards [24], [25], the user can be guaranteed an average service rate.

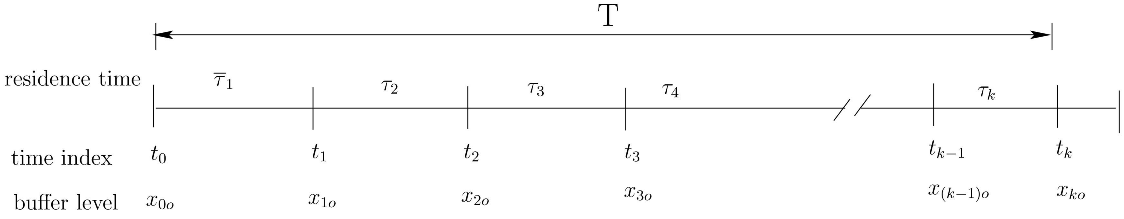

Fig. 2. Streaming session.

as long as the application requirements are below the intersection point. The latter strategy may be propelled by the interest of the cellular networks in maintaining the high revenue associated with applications such as voice telephony and audio streaming. If the application bandwidth requirement is above the intersection rate, the intermittent technology becomes more favorable and cost optimization of streaming applications is viable.

Fig. 2 shows a typical scenario for a streaming session in a two-tier heterogeneous system. Generally, we assume that the session duration, denoted as T, follows a generic heavy tailed distribution [29]. Clearly, after the session start at t_o, the mobile terminal encounters subsequent transitions between unique and dual technological zones at specific instants t_j until the session ends at t_k. The durations spent by the user in different technological zones, denoted as τ_n, are assumed to have generic probability and cumulative density functions, denoted as $f_n(\tau_n)$ and $F_n(\tau_n)$, respectively. Note that $\overline{\tau_n}$ represents the residual time distribution of τ_n.

The dynamics of the streaming application buffer level, $x(t)$, is governed by the following differential equation:

$$x'(t) = r_n(t) - r_o, \tag{1}$$

where r_n represents the server streaming rate during τ_n and r_o represents the average playout data rate. Typically, r_n is non-negative and upper-bounded by a maximum service rate of r_{nmax} in network N_n, i.e., $r_n \in [0,\ r_{nmax}]$. Practically, rate control messages are communicated to the streaming server using a stream management protocol such as Real-Time Streaming Protocol (RTSP) [30]. In this work, we assume the streamed media is precoded with a fixed data rate, i.e., the model assumes that the user chooses a specific media quality for the entire session. Additionally, we assume proactive stream management in which the streaming rate is adjusted before leaving the cheaper network to reduce the signaling cost and ensure seamless VHOs. It is worth noting that we also assume that the application proactively tunes the streaming rate to maintain the buffer level above the initial playout latency level when the rate of change of the buffer level is negative. This assumption ensures that the buffer always has sufficient data for smooth playout of the media. Clearly, this buffer model ignores buffer fluctuations that may evolve due to playout rate variability and occasional channel degradation. Generally, these fluctuations are of insignificant impact on the considered rate optimization problem due to the large time scale difference between these fluctuations and the zone residence times. Hence, the average value of streaming and playout rates are considered good approximate values.

To this end, it is worth pointing out that the presented problem ignores possible horizontal handoff (HHO) when the user roams between different cells in the overlay 3G cellular network. Generally, HHOs should not affect the streaming policy design as the application is not expected to change its streaming rate as it moves from one cell to another within the same technology. However, HHOs may affect the streaming quality if the next-cell does not have enough resources. The latter impact can be ignored for efficient streaming strategies that significantly reduce the session HHO rate as will be shown later.

2.2 Streaming Optimization Problem

The objective of the media streaming optimization problem is to establish the optimal streaming policy that minimizes a generic streaming cost function, which can be expressed as

$$J_{av} = \int_0^T \chi(x(t))dt. \tag{2}$$

The optimal cost can be realized by tuning the server streaming rate on each transition based on the buffer status and the cost profile of different operators. Considering the stochastic nature of user mobility, the presented problem falls under the umbrella of multistage stochastic sequential decision problems [31]. These types of problems take one of two forms: 1) multistage stochastic programs or 2) stochastic dynamic programs. The solution of these programs is computationally demanding because it usually involves generating and averaging a tremendous number of sample trajectories. Hence, this approach is not suitable for online estimation of streaming parameters due to the limited processing capacity of handheld devices and the real-time requirement of VHO decisions. Consequently, we focus on developing streaming policies that accommodate these limitations as detailed in the Section 3.

3 Optimized Streaming Algorithms

In this section, we develop novel practical cost optimization strategies in heterogeneous systems. We first present a generic optimization framework based on which we propose three streaming policies for heterogeneous networks.

3.1 Optimization Framework

The main objective of our optimization framework is to develop a practical means for minimizing the cost of streaming sessions in NgHWNs. Typically, monetary cost saving is mainly attained by buffering the media in the cheaper intermittent network and minimizing the utilization of the expensive one. Hence, we introduce the notion of *optimization cycle* that is determined by any subsequent visits to the cheaper and expensive network in series. Fig. 3

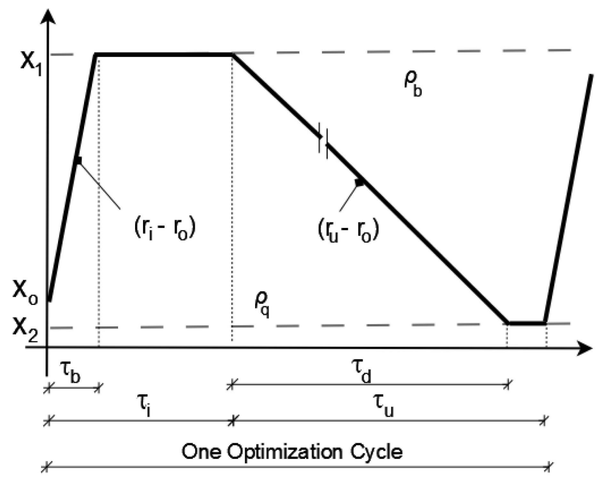

Fig. 3. Optimization cycle.

illustrates a typical optimization cycle showing the buffer level variations in its two time stages, which correspond to the durations spent by the user under the coverage of intermittent and universal networks. In this figure, ρ_q represents the buffer level that corresponds to the Initial playout latency. Generally, playout latency is employed to ensure the presence of sufficient media in the application buffer to start media playout. This delay is intended to avoid successive playout interruption due to buffer depletion caused by channel condition degradation. Consequently, the streaming strategies are designed to maintain the buffer level above this threshold. Additionally, ρ_b represents a buffering threshold whose value is optimized in our framework. Furthermore, x_o, x_1, and x_2 represent the buffer level at the start of the optimization cycle, on exiting the cheap network, and at the end of the optimization cycle, respectively.

In the first stage of the optimization cycle, the application streams the media at a rate that is larger than the nominal stream rate, i.e., $r_i > r_0$. Hence, the buffer level increases at a rate of $(r_i - r_o)$ until the buffer level reaches the buffering threshold. At this instant, the application readjusts the streaming rate to the nominal stream rate, i.e., r_o, in order to maintain the buffer level until the mobile terminal exits the hotspot. In the second stage of the optimization cycle, the application relies on the buffered data by setting the streaming rate r_u such that $0 \leq r_u < r_o$. Hence, the buffer is depleted at a rate of $(r_o - r_u)$. As the buffer level approaches the QoS threshold, ρ_q, in the expensive network, the application proactively adjusts the streaming rate to the nominal stream rate. Hence, the streaming policy will not result in any playout interruptions unless the extra bandwidth allocation request is denied by the 3G network.

To this end, the design of the streaming strategy under this framework involves the choice of a set of parameters including the buffering threshold ρ_b and the streaming rates r_i and r_u. Each strategy assigns a subset of these parameters as static values and optimizes the design values of the parameters in the complementary subset. The developed optimization framework is based on the foundations of stochastic optimization [32] that involves two steps

1. developing a deterministic version of the problem and
2. solving the problem using one of the well-established optimization tools.

Clearly, the streaming strategy parameters can only be optimized when the user starts the session in a WLAN or when it is handed off to a WLAN during the session. Note that if the session starts in the cheaper network, the streaming policy parameters will be optimized using the residual residence time distribution of the cheaper network, i.e., $\overline{\tau}_i$. If the streaming session starts in the cellular network, the application will adjust the server streaming rate to the nominal rate, i.e., r_o to minimize the session cost and satisfy QoS constraint. Finally, as the user continues to roam across the two networks, the design parameters are estimated online according to different operating conditions such as the mobility parameters, buffer level, and service cost.

3.2 Streaming Strategies

In this section, we first introduce the greedy media streaming (GMS) algorithm as a natural behavior for the users in NgHWN. We then propose the Greedy streaming to threshold and stop (GSTS),[2] controlled streaming to threshold and stop (CSTS), and controlled streaming to threshold and recourse (CSTR) algorithms. The latter three algorithms are based on the optimization cycle idea and are optimized using the framework presented in Section 4.

3.2.1 Greedy Media Streaming (GMS)

GMS is a typical strategy to reduce the session cost in heterogeneous system by taking 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 until the buffered data is depleted to ρ_q. At this instance, GMS readjusts the streaming rate to the nominal rate; i.e., r_o.

3.2.2 Greedy Streaming to Threshold and Stop (GSTS)

GSTS uses static values for the streaming rates r_i and r_u and only optimizes the buffering threshold value. In the cheaper network, the application sets the streaming rate to the maximum available bandwidth, i.e., $r_i = r_{imax}$. In the expensive network, GSTS pauses the streaming process, i.e., $r_u = 0$. The buffering threshold value is optimized to minimize the session cost using the cost optimization framework that is explained in the following section. The main intuition behind introducing the buffering threshold is to reduce the streaming monetary cost in the cheaper network after securing a sufficient amount of the streamed media in the application buffer. However, this strategy leads to increasing the network signaling load as the application will require more visits to finish downloading the whole media file.

3.2.3 Controlled Streaming to Threshold and Stop (CSTS)

In the cheaper network, CSTS optimizes the values of both the streaming rate and the buffering threshold. In the expensive network, CSTS pauses streaming, i.e., $r_u = 0$. Note that reducing the streaming rate in the cheaper intermittent

[2] Note that PGMS [23] is renamed GSTS for consistency.

network typically reduces the session cost as the streamed media is buffered at a lower cost rate. On the other hand, buffering the media at reduced streaming rates may lead to a failure in securing the target amount before exiting the cheaper network. This failure increases the probability of buffer depletion, which is followed by resource allocation request in the expensive network to avoid streaming interruption. Such a request involves excess signaling and may degrade the streaming QoS if the bandwidth allocation request is denied in the 3G cellular systems. The optimal choice of the streaming policy parameters should compromise the tradeoff between monetary cost, signaling load, and session QoS.

3.2.4 Controlled Streaming to Threshold and Recourse (CSTR)

CSTR represents the most flexible streaming strategy as it optimizes all the optimization cycle parameters including r_i, ρ_b, and r_u. Note that streaming on moving to the expensive network enables the application to use the expensive network at a lower cost rate before the buffer depletion. The high variability of the user mobility magnifies the importance of streaming in the expensive network because the user in many cases may not have time to secure enough media in the application buffer. In CSTR, the application considers a recourse estimate for r_u on moving to the expensive network because the exact value of the residence time in the cheap network and consequently, the buffer level on moving to the expensive network, become known at the end of the first stage of the optimization cycle. Hence, the recourse value of r_u represents a more accurate estimate for the value determined in the first half of the optimization cycle when all the mobility information is unknown. In the recourse phase, the streaming rate is determined such that the expected streaming cost of the second part of the optimization cycle in minimized, as detailed in the following section.

4 Cost Analysis and Optimization

In this section, we present a generic optimization framework that is applicable to GSTS, CSTS, and CSTR. We focus on the analysis of CSTR as it represents the most general streaming strategy. The analysis of CSTS and GSTS can be derived as special cases of CSTR as will be shown later. First, we present the cost analysis of an optimization cycle. We then present our cost optimization framework.

4.1 Optimization Cycle Cost Analysis

Typically, each optimization cycle has two stages that correspond to the duration spent in the cheaper and expensive networks. Our cost analysis follows the same sequence by presenting the cost of each stage in series.

4.1.1 Cost Analysis in the Cheaper Network

As the user starts a new optimization cycle, by moving from cellular network to a WLAN or starting a new session in the WLAN, the application should determine the values of the streaming strategy parameters, i.e., r_i, ρ_b, r_u. Typically, r_i is lower bounded by the nominal streaming rate and upper bounded by the technology maximum streaming rate, i.e., $r_o \leq r_i \leq r_{imax}$. Once, this value is communicated to the streaming server, the buffered media builds up at a rate of $(r_i - r_o)$. Let τ_b represent the time required for the buffered data level to reach the optimized buffering threshold. Then, τ_b can be expressed as $\tau_b = \frac{\rho_b - x_o}{r_i - r_o}$. Assuming that there is enough media to be streamed, we have two possible scenarios:

- The user moves out of the cheap network before hitting the buffering threshold, i.e., $\tau_i \leq \tau_b$. Hence, the user maintains a constant streaming rate during its residence in the cheaper network.
- The buffered media increases to the buffering threshold, i.e., $\tau_i > \tau_b$. Hence, at τ_b, the application adjusts the streaming rate to maintain the buffer level at the buffering threshold.

Note that in both cases, the application proactively adjusts the streaming rate to be used in the expensive network before exiting the cheaper network. To this end, the buffer level at the beginning of the second stage can be expressed as

$$x_1 = \begin{cases} x_o + (r_i - r_o)\tau_i, & \tau_i \leq \tau_b \\ \rho_b, & \tau_i > \tau_b. \end{cases}$$

Generally, we define three cost components including monetary cost, signaling cost, and QoS cost. In the cheaper intermittent network, these costs are denoted C_{im}, C_{is}, and C_{iq}, respectively. These costs can be expressed as

$$C_{im} = \begin{cases} \chi_i(r_i)\tau_i, & \tau_i \leq \tau_b \\ \chi_i(r_i)\tau_b + \chi_i(r_o)(\tau_i - \tau_b), & \tau_i > \tau_b, \end{cases}$$

$$C_{is} = \begin{cases} 2\chi_{is}, & \tau_i \leq \tau_b \\ 3\chi_{is}, & \tau_i > \tau_b, \end{cases}$$

$$C_{iq} = 0,$$

where χ_{is} represents the signaling cost for stream rate adjustment in the cheaper intermittent network. This cost includes resource allocation (deallocation) communication with the access network and RTSP message communication with the streaming server. Note that the signaling cost only considers the signaling resulting from the stream management and ignores the signaling incurred for other core functions such as location update [33] as the latter cost is incurred independent of the streaming strategy. Typically, the MT adjusts the streaming rate at least twice on entering and exiting WLANs. An additional rate adjustment is performed on hitting the buffering threshold ρ_b. The QoS cost is set to zero because we assume that WLANs have sufficient resources and will not deny the user resource allocation request. Additionally, the streaming rate in the cheaper network is always higher than the application playout rate. Hence, the user continuously maintains the application buffer above the QoS threshold.

4.1.2 Cost Analysis in the Expensive Network

On moving to the expensive network, the application determines its streaming rate, which is typically $0 \leq r_u \leq r_o$. Hence, the buffer depletes at a rate of $(r_0 - r_u)$ creating two possible cases

- the user enters a dual coverage zone before the buffer depletion to the QoS threshold ρ_q. Note that the buffer depletion duration may be expressed as $\tau_d = \frac{x_1 - \rho_q}{r_o - r_u}$ or $\overline{\tau_d} = \frac{\rho_b - \rho_q}{r_o - r_u}$ for $\tau_i \leq \tau_b$ or $\tau_i > \tau_b$, respectively.

- the buffer depletes while the user is still in the expensive network. In this case, the application proactively readjusts the streaming rate to r_o.

Consequently, the monetary, signaling, and QoS cost, respectively, donated as C_{um}, C_{us}, and C_{uq}, in the expensive network can be expressed as

$$C_{um} = \begin{cases} \chi_u(r_u)\tau_u, & \tau_u \le \tau_d \\ [\chi_u(r_u)\tau_d + \chi_u(r_o)(\tau_u - \tau_d)], & \tau_u > \tau_d, \end{cases}$$

$$C_{us} = \begin{cases} \chi'_{us}, & \tau_u \le \tau_d \\ \chi_{us} + \chi'_{us}, & \tau_u > \tau_d, \end{cases}$$

$$C_{uq} = \begin{cases} \chi_{uq}(r_u), & \tau_u \le \tau_d \\ \chi_{uq}(r_o - r_u), & \tau_u > \tau_d, \end{cases}$$

where χ'_{us} and χ_{us} represent the signaling cost in the expensive network excluding and including RTSP signaling and $\chi_{qu}(r_u)$ represents the QoS cost in the expensive network. Note that the RTSP signaling of the first rate adjustment is proactively sent in WLANs. Additionally, the QoS cost in the cellular network is considered to represent the possible resource allocation denial due to insufficient resources. This denial may lead to sluggish streaming if the application buffer is completely depleted. The QoS cost is assumed to be rate dependent as it is intuitively expected that the larger the requested the bandwidth is, the higher the probability of denying the resource allocation request.

4.2 Cost Optimization

In this section, we first derive an expression for the expected cost of the optimization cycle based on the presented cost analysis. We then formulate the cost optimization problem and present our solution approach.

4.2.1 *The Expected Cost of an Optimization Cycle*

The expected monetary cost, C_m, for an optimization cycle can be expressed as

$$\begin{aligned} C_m = &\int_0^{\tau_b} \chi_i(r_i)t_1 f_i(t_1)dt_1 \\ &+ \int_{\tau_b}^{\infty} [\chi_i(r_i)\tau_b + \chi_i(r_o)(t_1 - \tau_b)] f_i(t_1)dt_1 \\ &+ \int_0^{\tau_b}\int_0^{\tau_d} \chi_u(r_u)t_2 f_u(t_2)dt_2 f_i(t_1)dt_1 \\ &+ \int_0^{\tau_b}\int_{\tau_d}^{\infty} [\chi_u(r_u)\tau_d + \chi_u(r_o)(t_2 - \tau_d)] \\ &\quad f_u(t_2)dt_2 f_i(t_1)dt_1 \\ &+ \int_{\tau_b}^{\infty}\int_0^{\overline{\tau_d}} \chi_u(r_u)t_2 f_u(t_2)dt_2 f_i(t_1)dt_1 \\ &+ \int_{\tau_b}^{\infty}\int_{\overline{\tau_d}}^{\infty} [\chi_u(r_u)\overline{\tau_d} + \chi_u(r_o)(t_2 - \overline{\tau_d})] \\ &\quad f_u(t_2)dt_2 f_i(t_1)dt_1. \end{aligned}$$

After some algebraic manipulation, integrating by parts and simplifying, C_m can be expressed as

$$\begin{aligned} C_m = &\frac{\chi_i(r_o)}{\mu_i} + \frac{\chi_u(r_o)}{\mu_u} + (\chi_i(r_i) - \chi_i(r_o))I_1 \\ &+ (\chi_u(r_u) - \chi_u(r_o))[I_2 + I_3], \end{aligned} \tag{3}$$

where

$$\begin{aligned} I_1 &= \int_0^{\tau_b} (1 - F_i(t_1))dt_1, \\ I_2 &= \int_0^{\tau_b}\int_0^{\tau_d} (1 - F_u(t_2))dt_2\ f_i(t_1)dt_1, \quad \text{and} \\ I_3 &= \int_{\tau_b}^{\infty}\int_0^{\overline{\tau_d}} (1 - F_u(t_2))dt_2\ f_i(t_1)dt_1. \end{aligned}$$

Note that the first two terms in (3) correspond to the monetary cost for a mobile terminal streaming at the nominal stream rate and staying exactly the average residence time in each network. The third term represents the excess cost for streaming at a higher rate in the cheaper network while the last term in (3) represents the cost savings due to streaming at a lower rate in the cellular network.

Similarly, the expected signaling cost, C_s, can be expressed as

$$\begin{aligned} C_s = &2\chi_{is} + \chi_{is}[1 - F_i(\tau_b)] + \chi'_{us} \\ &+ \chi_{us}\{[1 - F_u(\overline{\tau_d})][1 - F_i(\tau_b)] + I_4\}, \end{aligned} \tag{4}$$

where $I_4 = \int_0^{\tau_b}[1 - F_u(\tau_d)]f_i(t_1)dt_1$. Typically, the first term in Section 4.2.1 is unavoidable resource management signaling on entering and exiting the cheaper network. In contrast, the second term is an additional cost due to rate adjustment after hitting the buffering threshold. The third term represents the cost associated with resource allocation as the user moves to the expensive network. The last term represents the signaling cost in the expensive network including the penalty that may be incurred if the buffer is depleted. Last but not least, the expected QoS cost, C_q, can be expressed as

$$\begin{aligned} C_q = &\chi_q(r_u) + \chi_q(r_o - r_u) \\ &\{[1 - F_u(\overline{\tau_d})][1 - F_i(\tau_b)] + I_4\}. \end{aligned} \tag{5}$$

Note that the first term in (5) corresponds to the possibility of resource allocation denial on performing VHO to the expensive network while the second term corresponds to the penalty for choosing a low data rate, which speeds the buffer depletion.

4.2.2 *Optimization Problem Formulation*

We formulate the optimization problem as a weighted sum of different cost components as follows:

$$\begin{aligned} \min \quad & C_{obj} = \sum_{l \in \{m,\, s,\, q\}} k_l C_l \\ \textit{such that} \quad & \rho_b \ge \rho_q \\ & r_o \le r_i \le r_{imax} \\ & 0 \le r_u \le r_o, \end{aligned} \tag{6}$$

where k_l represents the cost coefficient of C_l. These cost coefficients are used to tune the algorithm performance as will be shown later. To this end, the presented problem represents a deterministic program that can be solved for the streaming algorithm design parameters. Generally, the resultant program is nonlinear in the design parameter due to the nonlinearities in residence distributions and cost profiles. Hence, this problem is classified as a nonlinear optimization program with boundary constraints. The solution of nonlinear programs is generally realized using different numerical techniques. Practically, the adopted optimization technique should be simple, easy to implement, and applicable to any model without restrictions. All these features are satisfied by *direct search* methods [34]. More importantly, these methods are known for their speed and robustness when the application target is one or two digits of accuracy, which matches the requirements of the presented program.

Clearly, the objective function involves the solution of integrals I_1-I_4. Although each of these integrals can be calculated numerically, we develop a framework to analytically estimate these integrals to speed the solution of (6). In this framework, we employ phase-type (PH) distributions [35] to represent the residence times of different zones. PH distributions are chosen due to their flexibility and analytical tractability.

We assume that the mobile terminal records the durations spent in different dual and unique coverage zones. We fit the durations spent in a specific zone based on its coefficient of variation, denoted as θ_x[3] to hyper-exponential, exponential, and hypoexponential distributions for $\theta_x > 1$, $\theta_x = 1$, and $\theta_x < 1$, respectively. These distributions are specifically chosen due to their simple structure and computationally inexpensive fitting procedures [36]. To illustrate the simplicity of the fitting procedure, we present the case of hyperexponential distributions. The probability distribution function of hyperexponential random variables is expressed as $f_X(x) = \sum_{i=1}^{k} q_i\lambda_i e^{-\lambda_i x}$ in which k represents the number of exponential stages, λ_i represent the exponential parameter of stage i, and q_i represents the probability of stage i. Given the sample mean (μ) and coefficient of variation θ_x of a specific residence time, λ_1 and λ_2 can be estimated as

$$\lambda_1 = \mu \bigg/ \left[1 - \sqrt{\frac{q_2}{q_1}\frac{\theta_x^2 - 1}{2}}\right]$$

$$\lambda_2 = \mu \bigg/ \left[1 + \sqrt{\frac{q_1}{q_2}\frac{\theta_x^2 - 1}{2}}\right],$$

where the parameters q_1 and q_2 can be assigned any values that satisfy the restrictions $q_1, q_2 \geq 0,\ q_1 + q_2 = 1$, and $\lambda_1,\ \lambda_2 > 0$. Closed forms for the integrals I_1-I_4 can be obtained for using integration by parts. These closed forms are calculated for different combinations of residence time distributions but are not shown here for brevity.

4.2.3 Recourse Phase

The solution of (6) provides all the design parameters including ρ_b, r_i, and r_u. However, these parameters are estimated with all the involved stochastic parameters unknown. Clearly, at the end of the first stage of the optimization cycle, the buffer level, x_1, is known. Consequently, a recourse estimate of r_u can be considered based on the uncovered information. Similar to the previous analysis, the updated r_u can be estimated by minimizing a weighted sum of different expected costs for the second stage given the buffer level at the beginning of this stage. Similar to the previous analysis, the streaming rate in the expensive network can be obtained by solving the following program:

$$\min k_m\left\{\frac{\chi_u(r_o)}{\mu_u} + (\chi_u(r_u) - \chi_u(r_o))\int_0^{\tau_d}(1 - F_u(t_2))dt_2\right\} + k_s\chi'_{us} + k_q\chi_q(r_u) + [k_s\chi_{us} + k_q\chi_q(r_o - r_u)][1 - F_u(\tau_d)] \tag{7}$$

$$such\ that\ \ 0 \leq r_u \leq r_o,$$

where $\tau_d = \frac{x_1 - \rho_q}{r_o - r_u}$. Equation (7) also represents a deterministic program that can be solved using standard optimization tools.

To this end, the presented analytical framework is applicable to all the proposed streaming strategies. Yet further modifications are considered depending of the streaming strategy. For example, if greedy streaming is considered in the cheaper network, then $r_i = r_{imax}$. Similarly, if a stream pause is considered in the expensive network as in GSTS and CSTS, the streaming rate in the expensive network is set to zero, i.e., $r_u = 0$. Hence, the associated costs in the expensive networks will be expressed as follows:

$$C_{um} = \begin{cases} 0, & \tau_u \leq \tau_d \\ [\chi_u(r_o)(\tau_u - \tau_d)], & \tau_u > \tau_d, \end{cases}$$

$$C_{us} = \begin{cases} 0, & \tau_u \leq \tau_d \\ \chi_{us}, & \tau_u > \tau_d, \end{cases}$$

$$C_{uq} = \begin{cases} 0, & \tau_u \leq \tau_d \\ \chi_{uq}(r_o), & \tau_u > \tau_d. \end{cases}$$

5 Simulation and Numerical Results

The performance of the presented algorithms under different operating scenarios is investigated using NS-2 [37] simulations. We have simulated an integrated cellular-WLAN heterogeneous system, with square cells for simplicity of illustration. Each overlay cellular cell is composed of N square subdivisions, where WLANs are randomly located with probability P_{wo}. When an MT is handed off to another cell, it experiences a new random WLAN topology. For mobility simulation, we adopt a two-dimensional Gauss-Markov movement model [38]. In this model, the MT velocity is assumed to be correlated in time and is modeled by a Gauss-Markov process. In its discrete version, at time n, the MT velocity in each dimension, v_n, is given by

$$v_n = \alpha_v v_{n-1} + (1 - \alpha_v)\mu_v + \sqrt{1 - \alpha_v^2}\, x_{n-1}, \tag{8}$$

where α_v, $0 \leq \alpha_v \leq 1$, represents a velocity memory factor, μ_v is the asymptotic mean of v_n, and x_n is an independent and stationary Gaussian process with zero mean and standard deviation σ_v, where σ_v is the asymptotic standard deviation of v_n. The Gauss-Markov model can be easily tuned to represent a wide range of user mobility patterns between the two extreme cases of random-walk and constant velocity fluid-flow as α_v changes from 0 to 1. Additionally, it has been shown that the Gauss-Markov model successfully emulates real mobility patterns due to

[3] $\theta_x = \frac{\sigma_x}{\mu_x}$, where σ_x and μ_x represent the standard deviation and the mean of the corresponding measurements, respectively.

TABLE 1
Simulation Parameters

Param.	Value	Param.	Value	Param.	Value
a_i	3.2e-7	b_i	4e-3	r_{imax}	2Mbps
a_u	3.2e-8	b_u	4e-4	r_{umax}	50 KBps
μ_T	22min	θ_T	1.11	r_o	25 KBps
ρ	6 sec	P_v	0.02	P_h	0.03
α_v	0.75	σ_v	2.5	μ_v	0
P_{wo}	0.3	N	100		

its capturing temporal dependency, which is an important mobility characteristic.

We assume that the network pricing profiles of the cellular and WLAN technologies are expressed as $\chi_u(r) = a_u r^2 + b_u r$ and $\chi_i(r) = a_i r^2 + b_i r$, respectively. Note that these cost profiles are used for illustration purposes only. The design of optimal cost profiles in heterogeneous integrated systems is a nontrivial design issue and is beyond the scope of this work. We assume linear blocking probabilities in 3G systems. That is to say that the probability that a specific handoff in 3G system is blocked varies from 0 to P_v and P_h for vertical and horizontal handoffs, respectively, as the requested rate changes from 0 to r_{umax}. Typically, vertical blocking probability is expected to be smaller than horizontal blocking probability because the impact of VHOs is more critical due to their higher frequency.

The session duration is modeled using a hyperexponential distribution with mean μ_T and coefficient of variation θ_T to capture the high variability of session durations. In the following sections, we compare the performance of the GMS, GSTS, CSTS, and CSTR algorithms for different pricing strategies and mobility patterns. For the optimization problem solution, we use the parallel direct search method from the OPT++ optimization package [39]. Table 1 shows the default values of the simulation parameters of a two-tier system. The chosen maximum network service rates represent possible per user rate allocation strategies of CDMA2000 and 802.11g. In the shown figures, each point represents the performance metric for the mean of 500 sessions with its corresponding 95 percent confidence interval.

5.1 Pricing Strategies

In this section, we study the impact of different cost profiles including linear and nonlinear pricing strategies. Additionally, in order to focus on monetary cost, we reset several simulation parameters including signaling cost coefficient, QoS cost coefficient, and system blocking probabilities, i.e., $k_s = k_q = P_v = P_h = 0$.

5.1.1 Linear Pricing Strategies

Fig. 4 plots the average cost rate for different algorithms versus cost coefficient ratio, i.e., $\frac{b_u}{b_i}$. In this figure, $a_u = a_i = 0$ and $\chi_u(r_{umax}) = 100$. The figure also shows the ideal minimum cost, which is estimated assuming full knowledge of the user residence times in different technologies. Typically, as the cost coefficient ratio increases, WLANs cost profile becomes cheaper and more appealing to the user. The figure shows that using GMS not only incurs the lowest cost but also leads to near-optimal cost. The figure also shows that GSTS comes next to GMS followed by CSTS and CSTR. This result conforms with the previous result as it shows that using a greedy strategy under linear cost strategies incurs lower cost and minimizes usage of the expensive network. These results show that linear pricing schemes, which are equivalent to current usage-based schemes, are not effective pricing strategies to minimize the heterogeneous user resource-aggressive behavior in WLANs. The same result can be intuitively generalized for concave pricing profiles. Hence, in the rest of our study, we consider pricing profiles with convex shapes.

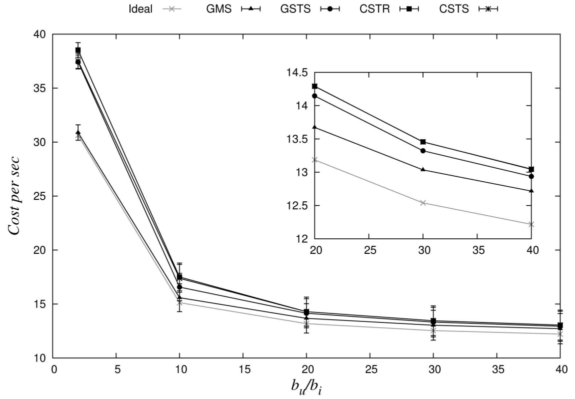

Fig. 4. Session cost versus cost coefficient ratio for linear pricing profiles.

5.1.2 Nonlinear Pricing Strategies

Fig. 5 plots the average session cost rate versus cost coefficient ratio for nonlinear pricing profiles. In this figure, the cost profile of the expensive network uses the coefficients shown in Table 1. The figure shows that controlled streaming policies realize the lowest cost followed by GSTS and GMS. For example, with one order of magnitude difference in the service cost of both networks, the monetary cost incurred using GMS is 1.5 times that of GSTS. Additionally, the figure shows that as the WLAN becomes cheaper, the incurred cost of these algorithms converges, specifically as the cost ratio increases beyond 30 as shown in the subplot. In contrast to the linear pricing case, the figure shows that the controlled streaming strategies result in a noticeable suboptimal cost margin to the ideal cost. However, the largest cost gap between the controlled streaming strategies cost and the

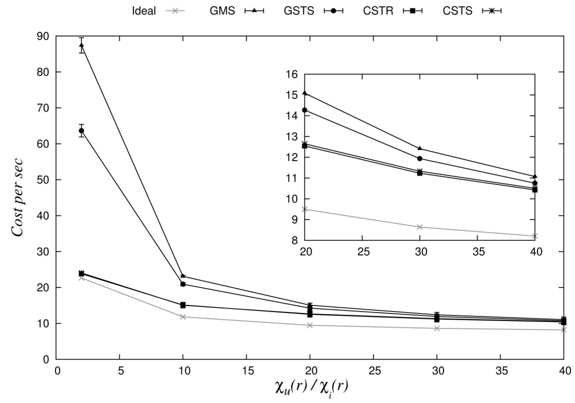

Fig. 5. Session cost versus cost ratio for nonlinear pricing profiles.

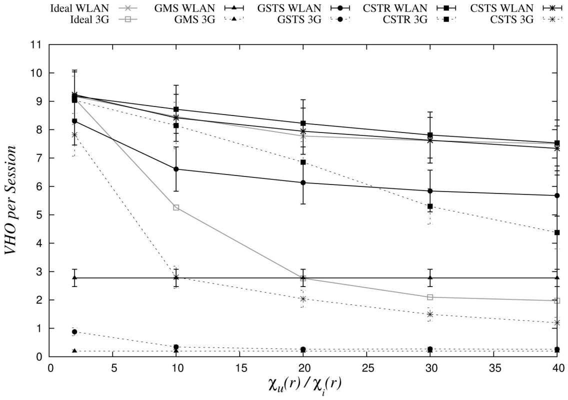

Fig. 6. Session VHO versus cost ratio for nonlinear pricing profiles.

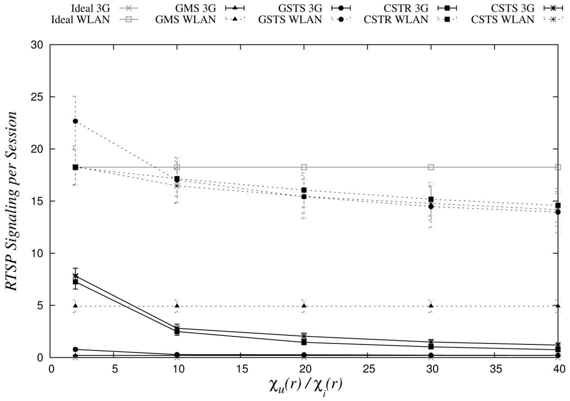

Fig. 7. Session RTSP signaling versus cost ratio for nonlinear pricing profiles.

ideal case is less than 27 percent. This cost difference is because the ideal cost is obtained with complete information of the residence times in different zones. Hence, it optimizes the streaming rates for the entire session in comparison to an optimization cycle as is the case for the presented algorithms. The ideal solution only represents a benchmark that is practically infeasible.

Fig. 6 plots the average executed-VHO signaling of different algorithms versus cost coefficient ratio for nonlinear pricing profiles. In this context, an executed VHO corresponds to a resource allocation request after network switching. Note that when the application pauses streaming on moving to the expensive network policy, it will not request resource allocation on moving to the expensive network as long as it has sufficient media in its buffer. Clearly, the figure shows a noticeable difference in the VHO signaling load in the integrated network. This difference is due to minimizing the usage of the expensive network, especially for the algorithms that pause the streaming in the expensive network. Additionally, the figure shows that GMS produces the least signaling followed by GSTS, CSTS, and CSTR.

Typically, GMS has the least signaling load in both networks as it aggressively maximizes the buffered media in the cheaper network. This behavior consequently minimizes the probability of the buffer depletion in the cellular network. Additionally, it enables the application to complete the stream download in fewer WLAN visits in comparison to threshold based strategies. The greedy behavior in WLANs also enables GSTS to successfully buffer the media to the target threshold, and consequently, minimizes its dependence on the cellular network as shown in Fig. 6. Additionally, the greedy behavior slightly reduces the number of WLANs visits of GSTS in comparison to other threshold-based media streaming strategies. Last but not least, the recourse phase in CSTR results in a noticeable increase in the cellular VHO signaling load that reaches 3.8 times that of CSTS. This increase is interpreted by the conservative rate choice of controlled streaming strategies in cheaper networks and the high variability of user mobility. Both factors do not allow the application in many cases to secure enough media in the application buffer. Hence, the application is forced to perform VHOs on moving to the expensive cellular network.

Fig. 7 plots RTSP signaling load for different algorithms versus the cost coefficient ratio for nonlinear pricing profiles. The noticeable difference between the RTSP signaling in WLANs and cellular 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. Hence, RTSP signaling in the 3G system corresponds to the buffer depletion to the QoS threshold ρ_q. The figure shows that GMS produces the lowest RTSP signaling in both networks for the same reasons behind the low VHO signaling of GMS. The figure also shows that GSTS produces higher signaling load than both CSTS and CSTR although the latter algorithms perform more WLAN VHOs as shown in Fig. 6. This result is due to the greedy behavior of GSTS that enables the application to secure the required media and consequently perform rate readjustment in the WLAN. The figure also shows that CSTS and CSTR send more RTSP messages in the cellular network. These messages reflect the frequency of the application buffer depletion in the cellular network. The figure also shows that the recourse part in the CSTR streaming policy slightly reduces the RTSP signaling load in the cellular system as it prolongs the time required to deplete the buffer to ρ_q.

Fig. 8 plots the average executed horizontal handoff (HHO) signaling load for different algorithms versus the cost coefficient ratio for nonlinear pricing profiles. In this context, an executed HHO corresponds to a cell transition during which the user is located in a unique coverage zone while the application is streaming. Note that the user will not request resource allocation if it is crossing cells through a dual coverage zone or if it is relying on the buffered data. Hence, in the latter two cases, the application will not initiate any HHO procedure. Clearly, the figure shows that GMS produces the least signaling load followed by GSTS, CSTS, and then CSTR. The figure shows also that introducing a recourse phase in CSTR leads to significant increase in the HHO signaling load. This increase reaches three times the signaling load of CSTS and 10 times of that of GMS.

To sum up, we find that the policy component of the proposed streaming strategies have different impacts on different performance metrics. The greedy behavior in the cheaper network successfully enables the application to buffer the media and significantly reduces the signaling load as shown for GMS and GSTS. However, it increases the session cost, especially for relatively expensive WLANs.

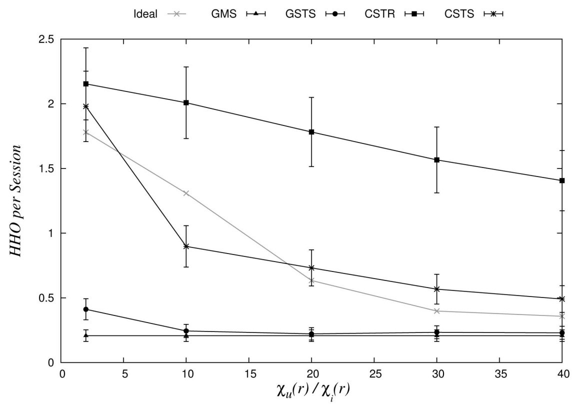

Fig. 8. Session HHO versus cost ratio for nonlinear pricing profiles.

Stopping the stream on moving to the expensive network represents an effective approach to reduce the signaling load, especially when combined with greedy streaming that ensures buffering the targeted amount. The controlled streaming proves its ability to significantly reduce the session monetary cost, irrespective of the adopted policy in the expensive network.

5.2 Algorithm Cost Weights and Performance Tuning

In this section, we investigate the impact of the algorithm cost weights, i.e., k_m, k_s, and k_q, on different performance metrics. The shown results are produced assuming the default cost parameters shown in Table 1. Additionally, we set $k_s = k_q = k$ and vary k while keeping $k_m = 1$ and we assume zero cost for signaling in WLANs, i.e., $\chi_{is} = 0$.

Fig. 9 shows the average session cost rate versus the signaling and QoS cost weights for different algorithms. The figure shows that CSTR has the lowest monetary cost followed by GSTS, GSTS, and them GMS. Additionally, the figure shows that the monetary cost of GSTS, CSTR, and CSTS can be tuned by varying the weights of different costs. The figure shows also that the lowest monetary cost for CSTS is attained with $k_q = k_s = 0$, i.e., when only the monetary cost is considered in the optimization framework. In contrast, the lowest monetary cost for CSTR and CSTS is realized with $k_m = k_s = k_q = 1$, i.e., all costs are equally considered in the objective function.

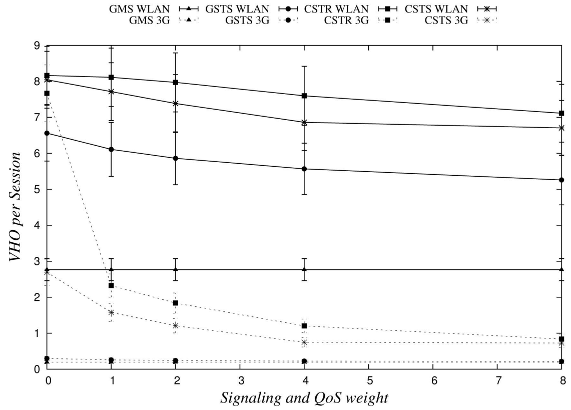

Fig. 10. VHO rate versus algorithm cost weight.

Fig. 10 plots the average session VHO signaling load for different algorithms versus their corresponding cost weights. Clearly, the figure shows that the variation of cost weights slightly reduces the VHO signaling in WLANs as it triggers the streaming strategy to increase the buffering threshold. In contrast, the figure shows that changing the cost coefficient has significant impact on VHO signaling of controlled streaming strategy in the cellular networks. For example, the 3G VHO signaling load of CSTR drops to approximately 30 percent as the cost weight k changes from 0 to 1. This drop is interpreted by securing enough data in the buffer before exiting the cheaper network by increasing the streaming rate. Furthermore, the figure also shows that CSTR always produces the largest signaling load in 3G networks for all cost weights due to the recourse strategy in the cellular network.

Fig. 11 plots the average session RTSP signaling load for different algorithms versus the algorithm signaling and QoS costs. Clearly, the figure shows that the increase in cost weight does not lead to significant changes in the RTSP signaling load in WLAN. In contrast, the increase of k from 0 to 1 leads to a significant drop in cellular RTSP signaling load to 40 percent. This drop is due to the decrease in VHOs to the cellular systems as shown in Fig. 10 and the more conservative estimate of the streaming rate in the cellular network on performing a VHO to it. Fig. 12 plots the average session HHO signaling load for different algorithms versus

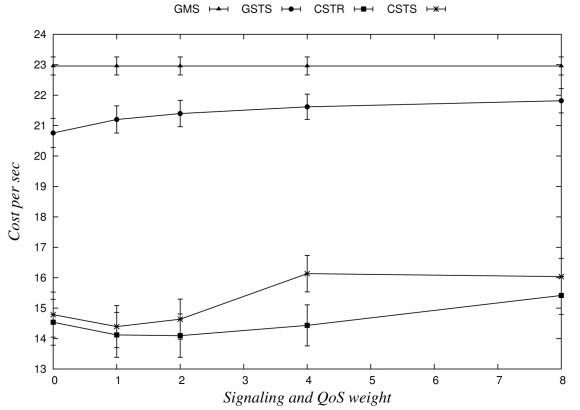

Fig. 9. Session cost rate versus algorithm cost weight.

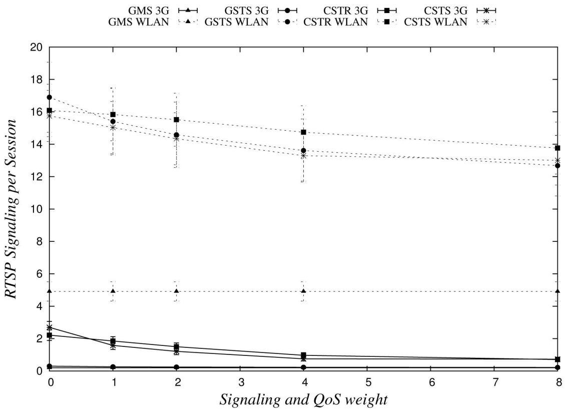

Fig. 11. RTSP signaling versus algorithm cost weight.

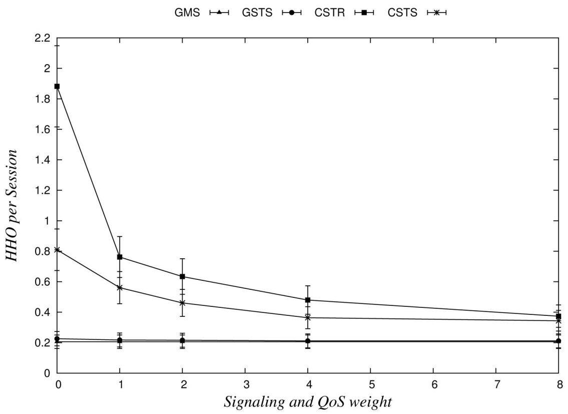

Fig. 12. Session HHO signaling versus algorithm cost weight.

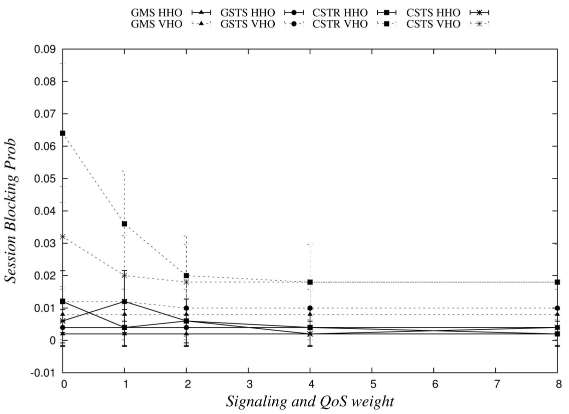

Fig. 13. Session blocking probability versus algorithm cost weight.

the algorithm signaling and QoS cost weight. Clearly, the figure shows significant drop in controlled streaming HHO signaling load. Intuitively, this drop is a natural consequence for decreasing the dependence on the cellular network due to its expensive signaling load. The drop in the frequency of both VHO and HHO leads to noticeable drop in the session blocking probability[4] as shown in Fig. 13

Clearly, Figs. 9, 10, 11, 12, and 13 show that the performance of the presented algorithms are significantly affected by the cost coefficient parameters. Additionally, these figures indicate the user interest in reducing his monetary cost and the operator interest in reducing the signaling load do not intersect. For example, the minimum cost for CSTS occurs at $k = 1$ while the lowest signaling load occurs as $k \rightarrow \infty$. Hence, the choice of the cost coefficients of the presented algorithms can be optimized to compromise the tradeoff between user and operator perspectives. In order to achieve this target, we define a total cost, C_{tot}, that is defined as

$$\begin{aligned} C_{tot} = \overline{C}_m + S_{VHO} + S_{HHO} + S_{RTSP} \\ + 100 * (B_{HHO} + B_{VHO}), \end{aligned}$$

where $\overline{C}_m$ represents the session cost rate, S_{VHO}, S_{HHO}, S_{RTSP}, respectively, represent the signaling load in 3G cellular system for VHO, HHO, and RTSP, and B_{HHO} and B_{VHO}, respectively, represent the HHO and VHO blocking probabilities. Fig. 14 plots the total cost for different algorithms versus QoS and signaling cost coefficients for $k_m = 1$. Clearly, the Figure shows that the total cost for CSTS and CSTR can be minimized by setting $k = 2$ and $k = 4$, respectively. Considering the fact that the session monetary cost rate for CSTR at $k = 4$ is less that the corresponding cost of CSTS at $k = 2$, we find that using CSTR with $k = 4$ represents a better stream management policy in this scenario.

6 Conclusion

The characteristic heterogeneity of different access technologies in next-generation integrated wireless systems brings both benefits and challenges to the system design. In this

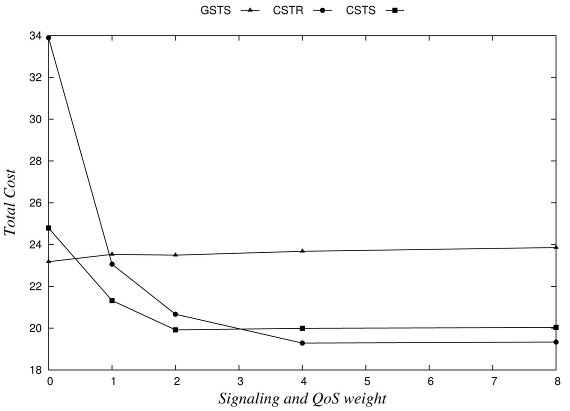

Fig. 14. Total cost versus QoS and signaling costs.

paper, media streaming optimization in two-tier systems is investigated. Novel media streaming strategies that understand the inherent system heterogeneity are developed. Additionally, the design parameters of these strategies are optimized using an analytical framework that accommodates the limited processing power of mobile devices and the real-time delay requirements of vertical handoffs. The performance of the developed strategies is simulated under different operating scenarios including different cost profiles of network operators and user mobility patterns. The results show that current usage-based cost profiles will not prevent user aggressive behavior especially in cheaper networks. Additionally, the results show that the developed streaming strategies realize noticeable monetary session cost reductions under advanced cost profiles. As future work, we are interested in extending the optimization framework to consider cost optimization with QoS improvement using scalable video.

Acknowledgments

The authors wish to acknowledge the support of Science Foundation Ireland (SFI).

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," *Proc. IEEE INFOCOM*, pp. 503-512, Apr. 2003.

[4] Typically, this probability does not represent a complete session dropping but corresponds to the chances of having a sluggish streaming due to insufficient resources.

[2] ETSI, "Universal Mobile Telecommunications System (UMTS); 3GPP System to Wireles Local Area Network (WLAN) Interworking; System Description," Technical Report ETSI TS 123 234 version 7.7.0 Release 7, Aug. 2008.

[3] 3GPP2, "CDMA2000 WLAN Interworking," 3GPP2 S.R0087-A, Feb. 2006.

[4] M. Stemm and R.H. Katz, "Vertical Handoffs in Wireless Overlay Networks," *ACM Mobile Networks and Applications,* vol. 3, no. 4, pp. 335-350, 1998.

[5] I. Akyildiz, Y. Altunbasak, F. Fekri, and R. Sivakumar, "AdaptNet: An Adaptive Protocol Suite for the Next-Generation Wireless Internet," *IEEE Comm. Magazine,* vol. 42, no. 3, pp. 128-136, Mar. 2004.

[6] S. Ohmori, Y. Yamao, and N. Nakajima, "The Future Generations of Mobile Communications Based on Broadband Access Technologies," *IEEE Comm. Magazine,* vol. 38, no. 12, pp. 134-142, Dec. 2000.

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

[8] E. Steinbach, N. Farber, and B. Girod, "Adaptive Playout For Low Latency Video Streaming," *Proc. Int'l Conf. Image Processing,* vol. 1, pp. 962-965, 2001.

[9] C. Sreenan, J.-C. Chen, P. Agrawal, and B. Narendran, "Delay Reduction Techniques for Playout Buffering," *IEEE Trans. Multimedia,* vol. 2, no. 2, pp. 88-100, June 2000.

[10] L.-J. Chen, T.G.Y. Sun, M. Sanadidi, and M. Gerla, "Adaptive Video Streaming in Vertical Handoff: A Case Study," *Proc. First Ann. Int'l Conf. Mobile and Ubiquitous Systems: Networking and Services,* pp. 111-112, 2004.

[11] A. Kassler, A. Schorr, L. Chen, C. Niedermeier, C. Meyer, M. Helbing, and M. Talanda, "Multimedia Communication in Policy Based Heterogeneous Wireless Networks," *Proc. IEEE 59th Vehicular Technology Conf.,* 2004.

[12] M. Bernaschi, F. Cacace, R. Clementelli, and L. Vollero, "Adaptive Streaming on Heterogeneous Networks," *Proc. First ACM Workshop Wireless Multimedia Networking and Performance Modeling (WMuNeP '05),* pp. 16-23, 2005.

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

[14] Y. Pan, M. Lee, J.B. Kim, and T. Suda, "An End-to-End Multipath Smooth Handoff Scheme for Stream Media," *IEEE J. Selected Areas in Comm,* vol. 22, no. 4, pp. 653-663, Mar. 2004.

[15] D. Jurca and P. Frossard, "Media-Specific Rate Allocation in Heterogeneous Wireless Networks," *Proc. Int'l Packet Video Workshop,* 2006.

[16] D.-N. Yang and M.-S. Chen, "Bandwidth Efficient Video Multicasting in Multiradio Multicellular Wireless Networks," *IEEE Trans. Mobile Computing,* vol. 7, no. 2, pp. 275-288, Feb. 2008.

[17] D. Niyato and E. Hossain, "Competitive Pricing in Heterogeneous Wireless Access Networks: Issues and Approaches," *IEEE Network,* vol. 22, no. 6, pp. 4-11, Nov./Dec. 2008.

[18] J. McNair and F. Zhu, "Vertical Handoffs in Fourth-Generation Multinetwork Environments," *IEEE Wireless Comm.,* vol. 11, no. 3, pp. 8-15, June 2004.

[19] 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.

[20] B. Liang, S. Drew, and D. Wang, "Performance of Multiuser Network-Aware Prefetching in Heterogeneous Wireless Systems," *ACM-Springer Wireless Networks,* vol. 15, no. 1, pp. 99-110, Jan. 2009.

[21] H. Chen, H. Wu, S. Kumar, and N.-F. Tzeng, "Minimum-Cost Data Delivery in Heterogeneous Wireless Networks," *IEEE Trans. Vehicular Technology,* vol. 56, no. 6, pp. 3511-3523, Nov. 2007.

[22] A.H. Zahran and C.J. Sreenan, "Cost Efficient Media Streaming Algorithms for Rate-Dependent Pricing Strategies in Heterogeneous Wireless Networks," *Proc. First IEEE Int'l Workshop Future Multimedia Networking,* pp. 485-491, Sept. 2008.

[23] A.H. Zahran and C.J. Sreenan, "PGMS: Pseudo Optimal Greedy Media Streaming Algorithm for Heterogeneous Wireless Networks," *Proc. Fourth IEEE Conf. Wireless and Mobile Computing, Networking, and Comm.,* pp. 12-15, Oct. 2008.

[24] D. Gao, J. Cai, and K.N. Ngan, "Admission Control in IEEE 802.11e Wireless LANs," *IEEE Network,* vol. 19, no. 4, pp. 6-13, July/Aug. 2005.

[25] C. Cicconetti, L. Lenzini, E. Mingozzi, and C. Eklund, "Quality of Service Support in IEEE 802.16 Networks," *IEEE Network,* vol. 20, no. 2, pp. 50-55, Mar./Apr. 2006.

[26] L. Badia, S. Merlin, A. Zanella, and M. Zorzi, "Pricing VoWLAN Services through a Micro-economic Framework," *IEEE Wireless Comm.,* vol. 13, no. 1, pp. 6-13, Feb. 2006.

[27] J. Lau and B. Liang, "Optimal Pricing for Selfish Users and Prefetching in Heterogeneous Wireless Networks," *Proc. IEEE Int'l Conf. Comm. (ICC '07),* June 2007.

[28] V. Gazis, N. Alonistioti, and L. Merakos, "Toward a Generic 'Always Best Connected' Capability in Integrated WLAN/UMTS Cellular Mobile Networks (and Beyond)," *IEEE Wireless Comm.,* vol. 12, no. 3, pp. 20-29, June 2005.

[29] M. Li, M. Claypool, R. Kinicki, and J. Nichols, "Characteristics of Streaming Media Stored on the Web," *ACM Trans. Internet Technology,* vol. 5, no. 4, pp. 601-626, 2005.

[30] H. Schulzrinne, A. Rao, and R. Lanphier, "Real Time Streaming Protocol," IETF RFC 2326, Apr. 1998.

[31] M.S. Casey and S. Sen, "The Scenario Generation Algorithm for Multistage Stochastic Linear Programming," *Math. Operations Research,* vol. 30, no. 3, pp. 615-631, 2005.

[32] S.S. Rao, *Engineering Optimization: Theory and Practice,* third ed. Wiley, 1996.

[33] I. Akyildiz, J. McNair, J. Ho, H. Uzunalioglu, and W. Wang, "Mobility Management in Next-Generation Wireless Systems," *Proc. IEEE,* vol. 87, no. 8, pp. 1347-1384, Aug. 1999.

[34] R.M. Lewis, V. Torczon, and M.W. Trosset, "Direct Search Methods: Then and Now," *J. Computational and Applied Math.,* vol. 124, nos. 1/2, pp. 191-207, 2000.

[35] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling.* SIAM, 1999.

[36] G. Bolch, S. Greiner, H. de Meer, and K.S. Trivedi, *Queuing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications,* second ed. Wiley, Aug. 1998.

[37] NS2, "Network Simulator," http://www.isi.edu/nsnam/ns, 2010.

[38] B. Liang and Z.J. Haas, "Predictive Distance-Based Mobility Management for Multi-Dimensional PCS Networks," *IEEE ACM Network,* vol. 11, no. 5, pp. 718-732, Oct. 2003.

[39] J.C. Meza, R.A. Oliva, P.D. Hough, and P.J. Williams, "OPT++: An Object-Oriented Toolkit for Nonlinear Optimization," *ACM Trans. Math. Software,* vol. 33, no. 2, p. 12, June 2007.

Ahmed H. Zahran received the BSc and MSc degrees in electrical engineering from the Electronics and Electrical Communication Department at Faculty of Engineering, Cairo University, in 2000 and 2002, respectively, and the PhD degree at the Department of Electrical and Computer Engineering, University of Toronto, in 2007. He is a postdoctoral researcher with the MISL group at the Computer Science Department, University College Cork, Ireland. His research interests span different topics in wireless mobile networking such as network architecture, mobility and resource management, and modeling and performance evaluation. He won the best paper award at the IFIP Networking 2005 conference.

Cormac J. Sreenan received the PhD degree in computer science from Cambridge University. He is a professor of computer science at University College Cork (UCC) in Ireland. Prior to joining UCC in 1999, he was on the research staff at AT&T Labs—Research, Florham Park, New Jersey, and at Bell Labs, Murray Hill, New Jersey. At UCC, he directs the Mobile & Internet Systems Laboratory (MISL), which is a group of more than 10 research staff and students with research activities in multimedia and wireless networking and systems. He is currently on the editorial boards of the *ACM/Springer Multimedia Systems Journal* and the *Elsevier Computer Networks Journal*. In the past, he has served as a guest editor for *Communications of the ACM*, the *IEEE Journal on Selected Areas in Communications*, and the *IEEE Wireless Communications Magazine*. He is a fellow of the British Computer Society.