Multi-Service Load Sharing for Resource Management in the Cellular/WLAN Integrated Network

Wei Song, *Student Member, IEEE*, and Weihua Zhuang, *Fellow, IEEE*

***Abstract*—With the interworking between a cellular network and wireless local area networks (WLANs), an essential aspect of resource management is taking advantage of the overlay network structure to efficiently share the multi-service traffic load between the interworked systems. In this study, we propose a new load sharing scheme for voice and elastic data services in a cellular/WLAN integrated network. Admission control and dynamic vertical handoff are applied to pool the free bandwidths of the two systems to effectively serve elastic data traffic and improve the multiplexing gain. To further combat the cell bandwidth limitation, data calls in the cell are served under an efficient service discipline, referred to as *shortest remaining processing time* (SRPT) [1]. The SRPT can well exploit the heavy-tailedness of data call size to improve the resource utilization. An accurate analytical model is developed to determine an appropriate size threshold so that data calls are properly distributed to the integrated cell and WLAN, taking into account the load conditions and traffic characteristics. It is observed from extensive simulation and numerical analysis that the new scheme significantly improves the overall system performance.**

***Index Terms*—Cellular/WLAN interworking, resource management, quality of service, load sharing, vertical handoff, admission control.**

I. Introduction

As two most popular wireless networks, the cellular network and wireless local area network (WLAN) are complementary in terms of mobility support, quality of service (QoS) provisioning, deployment strategy, etc. With the cellular/WLAN interworking, the complementary strengths of the two networks can be combined to provide QoS enhancement for multiple services. The multi-service traffic load should be appropriately shared across the interworked systems so as to efficiently utilize the overall resources. Especially, in the overlay area with both cellular and WLAN access, the load sharing is essentially important as the heterogeneous mobility and QoS support of the interworked systems can significantly affect the service provisioning and overall resource utilization.

There have been some research works on the load sharing for cellular/WLAN interworking via admission control, which properly assigns incoming calls to a target system in the overlay network. In [2], optimal and adaptive strategies are proposed only for data users based on user mobility and traffic characteristics. An optimal joint session admission control scheme based on a semi-Markov decision process (SMDP) is proposed in [3] for multimedia traffic. The overall network revenue is maximized under QoS constraints. Nonetheless, much research attention is paid to the vertical handoff calls involved with user mobility at WLAN boundary crossing. It is known that most WLANs are deployed in indoor environments like cafés, offices, and hotels. Users within these areas are most static or only maintain a pedestrian-level mobility. To efficiently utilize the resources in the interworked systems, it is necessary to introduce the dynamic load transfer for low-mobility users staying within the overlay area. In [4], dynamic session transfer is studied for hierarchical integrated networks as an analogy to task migration in distributed operating systems. We also investigate the load sharing problem in [5] and consider both call assignment via admission control and load transfer via dynamic vertical handoff. The complementary QoS provisioning capabilities of the cellular network and WLANs are effectively exploited by multiple services. However, there are still not many analytical works that consider the dynamic vertical handoff within the overlay area, which is triggered by network states instead of user mobility. As the dynamics of both interworked systems are involved, the load sharing problem becomes very complex for a multi-service scenario.

In this paper, we propose a new load sharing scheme for voice and elastic data services in the cellular/WLAN integrated network. First, it uses admission control and dynamic vertical handoff to distribute real-time voice calls preferably to the cell. With ubiquitous cellular coverage and fine QoS provisioning, the voice traffic can be efficiently supported. The free bandwidths unused by voice in the two systems are then combined to effectively serve elastic data traffic for a large multiplexing gain. To further combat the cell bandwidth limitation, we consider an efficient service discipline, referred to as *shortest remaining processing time* (SRPT) [1], for data calls in the cell. The SRPT can well exploit the heavy-tailed property of data call size to make a good trade-off between user perceived QoS and grade of service (GoS) (e.g., call blocking probability). Data calls are assigned to the integrated cell and WLAN, respectively, based on a call size threshold. To appropriately determine the size threshold, we develop an analytical model to evaluate the system performance accu-

Manuscript received May 1, 2007; revised September 21, 2007; accepted January 16, 2008. The editors coordinating the review of this paper and approving it for publication were J. C. Hou and V. K. Bhargava.

W. Zhuang is with the Department of Electrical and Computer Engineering, University of Waterloo, Canada (e-mail: wzhuang@bbcr.uwaterloo.ca).

W. Song is with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, US (e-mail: wsong@eecs.berkeley.edu).

Digital Object Identifier 10.1109/TWC.2009.070455

rately and effectively. The model takes into account the heavy-tailedness of data traffic and also the dynamic vertical handoff triggered by network states.

The remainder of this paper is organized as follows. In Section II, we describe the multi-service traffic model and the system capacity model for the WLAN and cellular cell. In Section III, a new load sharing scheme is proposed and the system performance is evaluated analytically. Based on the analytical model, we further discuss the impact of data size threshold and develop a simple search algorithm to determine the size threshold. Numerical results are presented and analyzed in Section IV. Section V concludes this research.

II. System Model

A. Traffic Model for Voice and Elastic Data Services

With cellular/WLAN interworking, the complementary network strengths can be effectively combined to improve multi-service provisioning. In this study, we consider both voice service and elastic data services, which are typical services of the conversational class and interactive class defined for the universal mobile telecommunication system (UMTS) [6], respectively. The conversational class is meant for real-time services characterized by a two-way conversational communication pattern. For voice service, there is a stringent delay requirement. The interactive class includes non-real time services such as Web browsing and file transfer, which are tolerant of elastic bandwidth. If the download of a Web page or data file is viewed as a data call, the data call duration (i.e., the time to complete the file transfer) is dependent on the file size and occupied bandwidth. Also, the data transfer delay is referred to as *response time* to emphasize the interactive nature. The mean response time should be bounded to ensure fluent interaction. The delay bound is far less stringent than that of conversational services. For Web browsing, a transfer delay of 2 - 4 seconds per page is the proposed bound and a desirable target is 0.5 seconds.

As incoming calls are invoked independently by a large number of users, we assume that voice and data call arrivals are independent Poisson processes with mean rates denoted by λ_v and λ_d, respectively. The voice call duration, which is of an order of minutes, is assumed to be exponentially distributed with mean $(\mu_v)^{-1}$. For elastic data services such as Web browsing and file transfer, it is observed that the packet-level traffic presents asymptotic self-similarity and high variability over a wide range of time scales [7]. This is mainly attributed to the heavy-tailed document size, which is a very important call-level traffic characteristic affecting the QoS metrics of interest. To explore the impact of the heavy-tailed data call size (denoted by L_d) on performance, L_d is modeled by a Weibull distribution [8], whose probability density function (PDF) is given by

$$f_{L_d}(x) = \frac{\alpha_d}{\beta_d}\left(\frac{x}{\beta_d}\right)^{\alpha_d - 1} e^{-(x/\beta_d)^{\alpha_d}} \qquad (1)$$
$$0 < \alpha_d \le 1, \quad \beta_d > 0, \quad x > 0$$

where α_d is the shape parameter and β_d is the scale parameter. The PDF of the Weibull distribution is denoted by $W_b(x, \alpha_d, \beta_d)$ for simplicity. The mean of L_d is given by $\mathrm{E}[L_d] \triangleq \overline{L}_d = \beta_d\,\Gamma(1 + \frac{1}{\alpha_d})$, where $\Gamma(\cdot)$ is the Gamma function. The exponential distribution is actually a special case of the Weibull distribution with $\alpha_d = 1$, while the Weibull distribution is heavy-tailed if $0 < \alpha_d < 1$. The smaller the α_d value, the heavier the tail that occurs in a given Weibull distribution. To assess the degree of heavy-tailedness, *Weibull factor* is introduced in [9], which is defined as

$$W_{L_d} = x\frac{d}{dx}\Big[\ln(-\ln(1 - F_{L_d}(x)))\Big] \qquad (2)$$

where $F_{L_d}(\cdot)$ is the cumulative distribution function (CDF) of L_d. For a Weibull distribution, the Weibull factor actually equals the shape parameter α_d.

B. System Capacity of WLAN and Cellular Cell

It is well known that the complementary strengths of the cellular networks and WLANs have motivated their interworking. As WLANs operate at license-exempt frequency bands, a large bandwidth is available to support a high date rate, e.g., up to 11 Mbit/s in IEEE 802.11b. However, WLANs are usually deployed in disjoint hotspot areas and can only provide local coverage. In contrast, the cellular networks have well entrenched infrastructure providing ubiquitous coverage, but relatively low data rates are supported with current widely deployed third-generation (3G) networks. For example, the UMTS system (Release 1999) can provide a data rate up to 2 Mbit/s for low-mobility applications (up to 10 km/hr) [10]. There are also some enhancement technologies such as the high speed packet access (HSPA), which can promote the downlink packet rate of UMTS access network up to 14 Mbit/s. However, these broadband wireless technologies are still not widely applied to the cellular networks in operation. Also, the deployment of microcells or picocells in hotspots is not so cost-effective as WLAN deployment. Hence, we focus on the interworking of WLANs and 3G cellular networks with a much smaller cell capacity.

To maximize the interworking effectiveness, it is imperative to take into account the complementary QoS provisioning strengths of the two networks. For the WLAN, the contention-based access determines its limitation in service differentiation and hard QoS guarantee, e.g., to real-time services. As all services with different traffic characteristics compete together to access the WLAN channel, this complete sharing (CS) manner penalizes services with a larger bandwidth requirement and privileges services requiring only a smaller bandwidth and those with aggressive traffic [11]. Hence, the elastic data traffic can be efficiently supported by the WLAN, as the large bandwidth and flexible access are constructive to increasing the multiplexing gain.

At the call level, a flow capturing the WLAN channel cannot hold the channel to complete its transmission, different from the first come, first served (FCFS) discipline. Instead, packets from ongoing calls take turns to be served according to the contention among them, similar to the sharing of CPU power by jobs in time-sharing computer systems. This service discipline is referred to as *processor sharing* (PS). The service queue with PS discipline exhibits unique characteristics, which should be considered in the resource management and may significantly affect the utilization. Based on the queueing

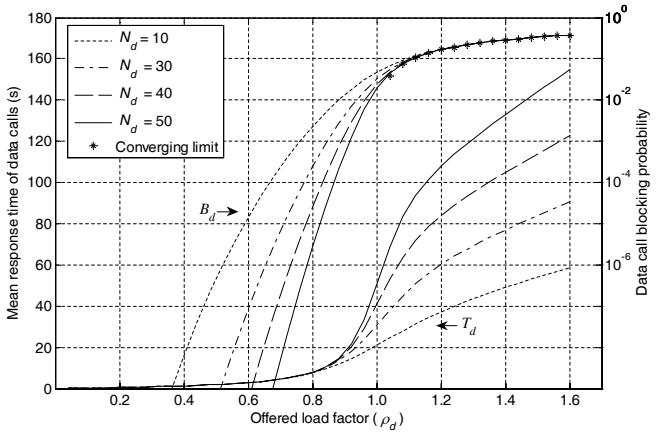


Fig. 1. Data call performance in terms of mean response time (T_d) and data call blocking probability (B_d) under PS service discipline versus offered load factor (ρ_d) and number of admitted data calls (N_d).

analysis for $M/G/1/K-PS$ queues, Fig. 1 is obtained to illustrate the dependence of performance on the offered traffic load factor (ρ_d) and number of admitted data calls (N_d). It can be seen that the mean response time T_d increases relatively slowly with ρ_d, when the system is underloaded with $\rho_d \ll 1$. For a moderately large value of N_d, the data call blocking probability B_d is very small and T_d is almost independent of N_d. However, when overload occurs with $\rho_d \geq 1$, T_d increases fast and almost linearly with ρ_d and N_d, while B_d converges fast to the limit $\frac{\rho_d - 1}{\rho_d}$ with a moderately large value of N_d [12]. Hence, admitting more calls is not effective to reduce the blocking probability in overload but may significantly degrade the perceived performance. It is important to ensure that the system operates in a normal load condition, so that the blocking probability is bounded and a sufficiently high throughput is maintained for admitted calls [12].

In [13], we have analyzed the WLAN capacity (i.e., the achievable throughput of the WLAN channel) for integrated voice and data services. The data throughput is observed to vary with the numbers of voice and data calls accommodated in the WLAN. When there is no voice call in service, the maximum achievable throughput is around 5.4 Mbit/s over a 11 Mbit/s physical channel. That is, the spectrum utilization at the medium access control (MAC) layer is around 50%. The data throughput is reduced by around 112 kbit/s to admit a new voice call, although the voice codec generates a packet stream only at a constant rate of 8 kbit/s in the example. As real-time traffic uses small payloads in packetization to meet the delay bound, the large protocol overhead indeed reduces the efficiency. Also, the simplified physical layer of WLAN to reduce implementation cost further exacerbates its weak support for real-time services [14].

To enhance QoS provisioning of WLANs, many mechanisms have been proposed [15]. For example, admission control can be applied at the access point to restrict the bandwidth occupancy of each service and enable certain QoS protection. It is observed in [16] that there exists an optimal operating point for the WLAN in the unsaturated case, beyond which the packet delay increases dramatically and the throughput drops quickly. When the packet service rate is larger than the arrival rate (network stability constraint) and the collision probability is small enough (e.g., less than 0.1), the service queue of a flow is almost empty and the packet delay is sufficiently small (say, less than 30 ms) to meet the requirement of real-time voice service. Based on the analytical model in [13], we can derive the WLAN admission region in terms of the maximum numbers of voice and data calls that can be simultaneously accommodated in the WLAN, denoted by a feasible set of vectors (n_v^w, n_d^w). Accordingly, the mean data packet service rate $\xi_d^w(n_v^w, n_d^w)$ (in bit/s) is obtained for each vector (n_v^w, n_d^w) in the admission region. Then, equipped with an admission control module, the access point of the WLAN decides whether to accept or reject an incoming call based on the numbers of ongoing calls and the admission region.

On the other hand, in the cellular network, reservation-based resource allocation is enabled with the centralized infrastructure. Real-time services can be supported efficiently and provided fine QoS guarantee. For example, voice calls with strict delay bound can be provided preemptive priority over data traffic, while data calls share the remaining bandwidth unused by voice traffic. As such, the QoS of voice calls is not degraded even when the system is overloaded with data traffic. Consider a cellular system based on code division multiple access (CDMA). Suppose that voice traffic is delivered with dedicated channels (DCH), while data traffic can be transported over the downlink shared channels (DSCH). Based on a cell load factor [17], the capacity of the more congested downlink can be modeled similar to [18]. The maximum numbers of simultaneously admitted voice and data users, denoted by (n_v^c, n_d^c), are limited to bound the interference level and satisfy user QoS requirements for the ratio of bit energy to noise and interference power spectral density $\left(\frac{E_b}{N_0}\right)$.

Based on the above system model, we investigate in this study how to properly share the multi-service traffic load across the interworked systems so as to maximize the overall resource utilization. In the following section, we propose a new load sharing scheme with QoS-awareness. The complementary strengths of the two networks in QoS provisioning are well exploited with admission control and dynamic vertical handoff. The characteristics of data call size are also taken into account in the load sharing.

III. Load Sharing between Integrated Cell and WLAN

A. Proposed Load Sharing Scheme with QoS-Awareness

As discussed in Section II-B, it is very inefficient to support real-time services in the WLAN due to excessive control overhead. In contrast, the cellular network has a strength in real-time service provisioning. The large cell size and ubiquitous cellular coverage can reduce handoff frequency and in turn the impact of handoff latency on delay-sensitive real-time traffic. Thus, in our load sharing scheme, an incoming voice call is preferably distributed to the cell, and overflows to the WLAN only if there is not sufficient free bandwidth for a voice call in the cell. The advantage of preferably assigning voice calls to the cell has also been observed in previous works such as [19]. On the other hand, we also consider dynamic transfer of ongoing voice calls in the WLAN to the cell via vertical handoff whenever the cell has free

bandwidth to accommodate more voice calls. The dynamic call transfer can be implemented with the policy-based framework discussed in [5]. When network under-utilization is detected after call completion or outgoing handoff, vertical handoff can be triggered and performed with the coordination of the WLAN access point and cellular radio network controller. However, the signaling overhead is inevitable with the dynamic call transfer. As mentioned in Section II-A, the data call duration is bounded within several seconds to guarantee responsiveness. To improve the control efficiency, we only consider the dynamic transfer of voice calls between the cell and the WLAN, as voice calls are relatively long-lived with an average duration in an order of minutes. By this means, voice calls are more concentrated in the cell and provisioned fine QoS guarantee. The bandwidths unused by voice traffic in the two systems can then be combined to effectively serve data calls. The rationale behind the idea can be understood by viewing the integrated cell and WLAN as two coupled queueing systems with service rates C_1 and C_2, respectively. By exploiting the cellular/WLAN interworking and vertical handoff, the integrated network performance within the overlay area can approach that of one queue with a larger service rate $(C_1 + C_2)$, which maximizes the multiplexing gain [20].

For elastic data calls, the response time depends on the bandwidth sharing manner and also the fluctuation of ongoing flow numbers [21]. As discussed in Section II-B, data traffic in the WLAN is served in a PS manner with the contention-based access. The mean response time varies with the offered load as shown in Fig. 1. On the other hand, because voice calls are preferably distributed to the cell via admission control and vertical handoff, the average bandwidth available to data traffic is relatively small when the voice traffic load is high. With the centralized control of base stations, it is necessary and feasible to serve data calls in the cell with a more efficient bandwidth sharing policy. In this study, we consider the shortest remaining processing time (SRPT) discipline [1], which is optimal in terms of minimizing the mean response time. Under the SRPT, only one call with the least remaining data to transmit is scheduled first and receives service at an instant. Given an incoming data call with a size smaller than the remaining data size of the call in service, the ongoing call is preempted and waits in the queue, while the new call is served subsequently. In contrast, under the PS, each ongoing call shares an equal quantum of service. As such, smaller-size data calls under the SRPT will not be stuck in the system for such a long duration as when the bandwidth is shared with data calls of a larger size.

It is known that the SRPT can significantly outperform the PS when the call size is heavy-tailed and the load is high. It may be suspected that the improvement of SRPT over PS comes at the expense of a longer response time for calls with a larger data size. Thus, the SRPT is often thought to be unfair as it favors short calls and penalizes long calls. An argument for this claim is the Kleinrock conservation law [22], which holds for service disciplines not making use of the size but is not necessarily true for size-based disciplines such as the SRPT. It is proved in [23] that, for any load condition and any continuous heavy-tailed size distribution with finite mean and variance, at least 99% of the data calls have a smaller response time under the SRPT than under the PS. These 99% of calls actually do significantly better, and the unfairness of SRPT diminishes with the heavy-tailed property. In addition, the control overhead of SRPT such as for preemption is also not higher than that of PS [23]. In practical systems, the PS may be implemented in a round robin manner and each call is preempted after receiving one quantum of service. In contrast, the preemption of SRPT only occurs when a new call of a smaller size arrives and there is less preemption overhead. Although the SRPT scheduling may involve higher implementation complexity and cost than the simple FCFS, significant performance improvement can be achieved. There is always the trade-off between complexity and performance. As the SRPT is applied at the call level instead of the packet level, the implementation complexity and cost should be affordable.

In this study, we consider some specific elastic data applications such as Web browsing and file transfer. They usually preserve a request-response pattern and are primarily unidirectional from application servers to user terminals. The Web documents or data files are pre-stored in the Web server or file server. It is possible to know the data call size *a priori* from session signaling. For example, a session description protocol (SDP) offer/answer mechanism has been proposed as an Internet draft for file transfer [24]. By introducing a set of new SDP attributes, it is possible to deliver some meta information of the file (such as content type and size) before the actual transfer. On the other hand, cross-layer design has become very popular and essential in the wireless domain to address the unique challenges such as the scarce radio resources and highly error-prone transmission conditions. The information exchange across different protocol layers can further improve the system performance. Hence, in our load sharing scheme, we exploit the meta information of data calls that can be passed to the network layer. In particular, a data call is distributed to the cell if the call size is not greater than a threshold Φ_d and the cell bandwidth available to data traffic is at least R_d^c. Otherwise, that data call is assigned to the WLAN. By properly determining the data size threshold (to be discussed in Section III-C), we can improve the resource utilization without degrading the user QoS experience.

B. Steady-State Probabilities of Interworked System

In the above load sharing scheme, we take into account the traffic characteristics of different services and the complementary QoS provisioning capabilities of the two networks. By taking advantage of the interworking and vertical handoff, the free bandwidths of the interworked systems are combined to maximize the multiplexing gain. Some previous works such as [5] have shown the performance improvement by simulation. In this section, we analytically evaluate the QoS metrics such as voice/data call blocking probabilities and mean response time of data calls, based on which we can appropriately determine the data size threshold.

As discussed in Section II-B, data calls in the WLAN share the available bandwidth in a PS manner. Under the PS, the mean response time is insensitive to the call size distribution if the overall service capacity is fixed. Nonetheless, due to

the random access in the WLAN, the bandwidth available to data traffic actually fluctuates not only with voice call arrivals/departures but also with the contention status. The insensitivity is generally lost in case of a varying capacity [25]. For data calls with a heavy-tailed size and high variability, the call-level performance such as mean response time even improves over the case with an exponentially distributed data call size. However, with admission control in place, the insensitivity can be retained for a high load condition, where proper resource allocation and load control are critical to prevent QoS violation. In a light load case, the call blocking probability is usually sufficiently low and all admitted calls are provided satisfactory QoS. Hence, we assume that the QoS of data traffic in the WLAN is insensitive to the heavy-tailed call size distribution. The insensitivity assumption is validated by the numerical results given in Section IV-A. Although conservative control is possible for a light load condition due to QoS underestimation, the control effectiveness is not affected very adversely.

Since data calls are assigned to the integrated cell and WLAN based on the bandwidth occupancy state of the cell and data call size, the data call arrivals to the cell and WLAN are still Poisson processes with mean rates denoted by λ_d^c and λ_d^w, respectively. Then, with the insensitivity assumption for data service in the WLAN, we can model the interworked systems with a three-dimensional Markov chain, in which the state (i, j, k) denotes the numbers of voice and data calls in the WLAN (i and j, respectively) and the number of voice calls in the cell (k). The steady-state probability is denoted by $\pi(i, j, k)$. Based on the bandwidth occupancy of voice traffic in the cell and the data call size distribution given in (1), the mean data call arrival rate to the cell can be derived as

$$\lambda_d^c = \lambda_d \cdot \delta_d^c \cdot \chi_d^c \tag{3}$$

$$\delta_d^c = \int_0^{\Phi_d} f_{L_d}(x)\ dx, \quad \chi_d^c = \sum_{(i,j)} \sum_{k:\ C_d^c(k)\ \geq\ R_d^c} \pi(i, j, k)$$

where δ_d^c is the fraction of data calls with a size not greater than Φ_d, $(1 - \chi_d^c)$ is the probability that such a data call is blocked by the cell due to congestion, and $C_d^c(k)$ is the maximum cell capacity available to data traffic when there are k voice calls in progress. Similarly, the mean data call arrival rate to the WLAN can be obtained as

$$\lambda_d^w = \lambda_d \Big[\delta_d^c \cdot (1 - \chi_d^c) + (1 - \delta_d^c)\Big] = \lambda_d \cdot \big(1 - \delta_d^c \cdot \chi_d^c\big). \tag{4}$$

As discussed in Section II-B, the WLAN capacity varies with the accommodated traffic load due to variable contention overhead. Here, the analytical model in [13] is adopted to capture the throughput degradation when more voice calls are admitted to the WLAN. We can derive the maximum numbers of voice and data calls that can be simultaneously carried by the WLAN (n_v^w, n_d^w). Accordingly, the mean data packet service rate $\xi_d^w(n_v^w, n_d^w)$ is also obtained for each vector (n_v^w, n_d^w) in the admission region. Based on the analytical model, we can derive the state transition rates of the aforementioned three-dimensional Markov chain, given at the top of next page, where N_v^c and N_v^w are the maximum numbers of voice calls admitted in the cell and the WLAN, respectively, $N_d^w(i)$ is the maximum number of data calls allowed in the WLAN with i voice calls in progress[1], $\xi_d^w(i, j)$ is the mean service rate provided to each data call when there are i voice calls and j data calls in the WLAN, and g_d^w is the mean size of data calls flowing to the WLAN. Note that the transition rate from state (i, j, k) to state $(i - 1, j, k)$ consists of two components. One is due to the completion of the i voice calls in the WLAN with a mean rate of $i \cdot \mu_v$, and the other is due to the completion of the k voice calls in the cell with a mean rate $k \cdot \mu_v$. When one of the k voice calls in the cell completes and makes room for a new voice call, one of the i voice calls in the WLAN can be handed over to the cell. According to our load sharing scheme and the overall data call size distribution given in (1), g_d^w can be derived as

$$g_d^w = \frac{(1 - \chi_d^c) \displaystyle\int_0^{\Phi_d} x f_{L_d}(x)\ dx + \int_{\Phi_d}^{\infty} x f_{L_d}(x)\ dx}{\delta_d^c \cdot (1 - \chi_d^c) + (1 - \delta_d^c)}. \tag{6}$$

The first term in the numerator of (6) corresponds to data calls of a size not greater than Φ_d, which are blocked by the cell due to congestion with a probability $(1 - \chi_d^c)$ and overflow to the WLAN. The second term in the numerator accounts for the data calls that have a size larger than Φ_d and are assigned to the WLAN to request admission. The denominator is a normalization constant for the size distribution of data calls flowing to the WLAN.

Due to the interdependence between i and k as shown in the state transition rates of (5), the size of the state space does not explode with the third dimension of the Markov chain (k), i.e., the number of voice calls in the cell. The steady-state probabilities $\pi(i, j, k)$ can be obtained by solving a very sparse linear system of balance equations. Then, the voice call blocking probability B_v is given by

$$B_v = \sum_{\substack{(i,j):\ i\ \leq\ N_v^w \\ j\ >\ N_d^w(i+1)}} \pi(i, j, N_v^c). \tag{7}$$

That is, an incoming voice call is blocked if there are N_v^c voice calls in the cell and not sufficient spare capacity is available for one more voice call, and if the WLAN is also congested with i voice calls and j data calls, which means that, with the j data calls already in progress, the admission of one more voice call in the WLAN will result in delay violation to the admitted i voice calls.

As illustrated in Fig. 1, when overload occurs, the mean response time under the PS increases dramatically with the offered load and the number of admissible calls (N_d), while the call blocking probability converges and cannot be reduced by increasing N_d. In contrast, in an underload case, the call blocking probability is sufficiently small with a reasonably large value of N_d and the mean response time is almost independent of N_d. Similar phenomenon is observed for the SRPT discipline. Hence, the QoS of data calls can be assured by maintaining an underload condition for data traffic in the cell. This can be achieved by properly determining the data

[1] N_v^c, N_v^w, and $N_d^w(i)$ are obtained from the admission regions of the cell and the WLAN, i.e., the feasible sets of vectors (n_v^c, n_d^c) and (n_v^w, n_d^w), respectively. Here, $N_v^c = \max(n_v^c)$, $N_v^w = \max(n_v^w)$, and $N_d^w(i) = \max(n_d^w)$, given $n_v^w = i$.

$$\begin{array}{ll}
(i,j,k)\rightarrow(i,j,k+1): & \lambda_v, \quad \text{if } i\le N_v^w,\ j\le N_d^w(i),\ k\le N_v^c-1\\
(i,j,k)\rightarrow(i,j,k-1): & k\cdot\mu_v, \quad \text{if } i=0,\ j\le N_d^w(i),\ 1\le k\le N_v^c\\
(i,j,k)\rightarrow(i+1,j,k): & \lambda_v, \quad \text{if } i\le N_v^w-1,\ j\le N_d^w(i+1),\ k=N_v^c\\
(i,j,k)\rightarrow(i-1,j,k): & (i+k)\cdot\mu_v, \quad \text{if } 1\le i\le N_v^w,\ j\le N_d^w(i),\ k=N_v^c\\
(i,j,k)\rightarrow(i,j+1,k): & \lambda_d^w, \quad \text{if } i\le N_v^w,\ 0\le j\le N_d^w(i)-1,\ k\le N_v^c\\
(i,j,k)\rightarrow(i,j-1,k): & j\cdot\xi_d^w(i,j)/g_d^w, \quad \text{if } i\le N_v^w,\ 1\le j\le N_d^w(i),\ k\le N_v^c
\end{array} \tag{5}$$

size threshold Φ_d. Then, the data call blocking probability B_d can be obtained as

$$B_d=\Big[\delta_d^c\cdot(1-\chi_d^c)+(1-\delta_d^c)\Big]B_d^w=(1-\delta_d^c\cdot\chi_d^c)\cdot B_d^w \tag{8}$$

where B_d^w is the data call blocking probability of the WLAN and is given by

$$B_d^w=\sum_{\substack{(i,j):\ i\le N_v^w\\ j+1>N_d^w(i)}}\ \sum_{k=0}^{N_v^c}\pi(i,j,k). \tag{9}$$

That is, the admission of a new data call should not degrade the WLAN capacity so much that the bandwidth requirement of ongoing voice calls cannot be satisfied. From the Little's law, the mean response time of data calls served in the WLAN can be obtained as

$$T_d^w=\frac{1}{\lambda_d^w\cdot(1-B_d^w)}\sum_{\substack{(i,j):\ i\le N_v^w\\ j\le N_d^w(i)}}\ \sum_{k=0}^{N_v^c}j\cdot\pi(i,j,k). \tag{10}$$

On the other hand, the mean response time of data calls admitted to the cell can be obtained from the $M/G/1-SRPT$ queue. This is because data call arrivals to the cell is still a Poisson process with a mean rate λ_d^c given in (3). The data call blocking probability is negligibly small if an underload condition is guaranteed by the threshold Φ_d. The average cell bandwidth allocated to data calls is

$$\overline{C}_d^c=\sum_{\substack{(i,j):\ i\le N_v^w\\ j\le N_d^w(i)}}\ \sum_{k=0}^{N_v^c}C_d^c(k)\cdot\pi(i,j,k). \tag{11}$$

Then, based on the formulas in [1], the mean response time is approximated by

$$T_d^c=\int_0^{\Phi_d}\frac{1}{\delta_d^c}f_{L_d}(x)\ \Gamma_d^c(x)\ dx \tag{12}$$

where $\frac{1}{\delta_d^c}f_{L_d}(x)\ (0<x\le\Phi_d)$ is the PDF of the size of data calls in the cell, and $\Gamma_d^c(x)$ is the conditional response time for a data call of size x, given by

$$\Gamma_d^c(x)=\int_0^y\frac{dt}{1-\rho_d^c(t)}+\frac{\lambda_d^c\Big[\int_0^y t^2 g_{L_d}(t)\ dt+y^2(1-G_{L_d}(y))\Big]}{2\big[1-\rho_d^c(y)\big]^2} \tag{13}$$

$$\rho_d^c(y)=\lambda_d^c\int_0^y t\cdot g_{L_d}(t)\ dt,\quad y=\frac{x}{\overline{C}_d^c} \tag{14}$$

$$g_{L_d}(t)=\frac{1}{\delta_d^c}W_b(t,\alpha_d,\beta_d/\overline{C}_d^c),\quad 0<t\le\Phi_d/\overline{C}_d^c. \tag{15}$$

Here, $g_{L_d}(\cdot)$ denotes the PDF of a bounded Weibull distribution and $G_{L_d}(\cdot)$ the corresponding CDF. In contrast to the data call size distribution $W_b(x,\alpha_d,\beta_d)$ given in (1), the scale parameter β_d is proportionally modified with $\overline{C}_d^c$ to switch the unit from data call size to service time.

For comparison purpose, when data calls in the cell are served under the PS discipline, the mean response time can be approximated by [26]

$$T_d^c=\frac{(\overline{\rho}_d^c)^{N_d^c+1}(N_d^c\,\overline{\rho}_d^c-N_d^c-1)+\overline{\rho}_d^c}{\lambda_d^c\cdot\big[1-(\overline{\rho}_d^c)^{N_d^c}\big](1-\overline{\rho}_d^c)},\quad \overline{\rho}_d^c=\rho_d^c(\Phi_d/\overline{C}_d^c) \tag{16}$$

where $\overline{\rho}_d^c$ is the average load factor of data traffic in the cell, which can be obtained from (14), and N_d^c is the maximum number of data calls allowed in the cell. Considering the sharing of data traffic load based on call size, the overall mean response time of data calls can be evaluated by

$$T_d=\frac{\delta_d^c\,\chi_d^c\cdot T_d^c+\big[\delta_d^c\cdot(1-\chi_d^c)+(1-\delta_d^c)\big](1-B_d^w)\cdot T_d^w}{\delta_d^c\,\chi_d^c+\big[\delta_d^c\cdot(1-\chi_d^c)+(1-\delta_d^c)\big](1-B_d^w)}. \tag{17}$$

C. Determination of Data Size Threshold

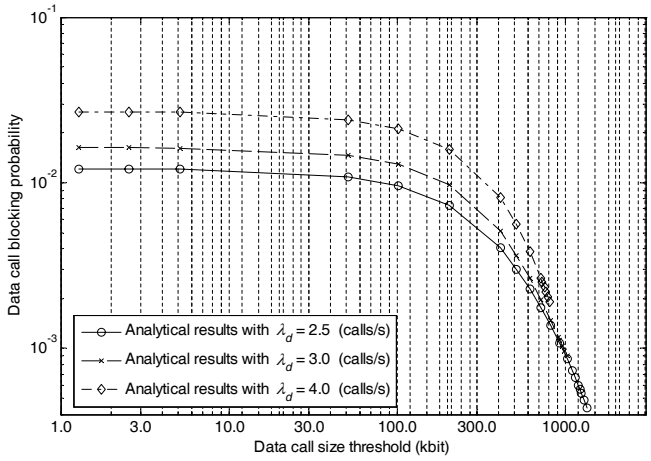
The proposed load sharing scheme aims at efficiently sharing the multi-service traffic load between the integrated cell and WLAN. Since voice calls are preferably distributed to the cell for high efficiency and fine QoS, data traffic should be properly balanced between the two systems correspondingly. Based on the observations in Section II-B, there are some important principles to follow in determining the data size threshold Φ_d.

First, an underload condition should be ensured for data traffic in the cell. That is, the data load factor in the worst case, denoted by $\hat{\rho}_d^c$, is less than 1:

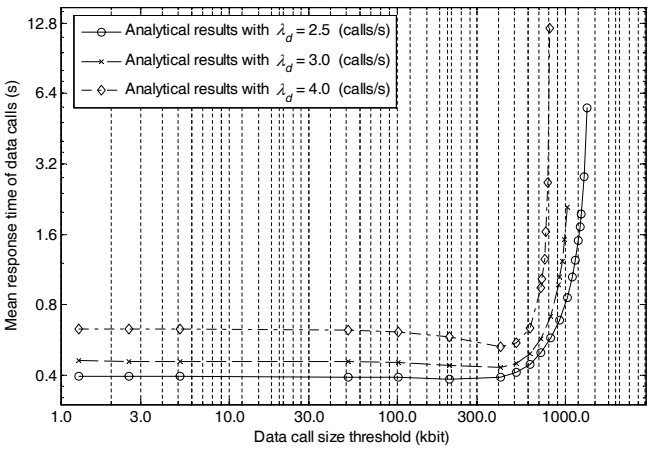
$$\hat{\rho}_d^c=\lambda_d^c\int_0^{\Phi_d/R_d^c}t\cdot\frac{1}{\delta_d^c}W_b(t,\alpha_d,\beta_d/R_d^c)\ dt<1 \tag{18}$$

where R_d^c is the minimum cell bandwidth available to data traffic, and $\frac{1}{\delta_d^c}W_b(t,\alpha_d,\beta_d/R_d^c)$, $0<t\le\Phi_d/R_d^c$, denotes the PDF of a bounded Weibull distribution with shape parameter α_d and scale parameter β_d/R_d^c. Moreover, data calls with a smaller size usually expect a shorter response time than those with a larger size. As data calls in the cell have a smaller size than most of those in the WLAN, our second principle is to guarantee that $T_d^c\le T_d^w$. The mean response time T_d^w and T_d^c are given by (10) and (12), respectively. Last, a good trade-off should be maintained between user-perceived QoS such as mean data response time and GoS in terms of call blocking probabilities. An appropriate threshold Φ_d^* can be determined to satisfy the following condition:

$$B_d(\Phi_d)<B_d(\Phi_d^*)\Rightarrow T_d(\Phi_d)>T_d(\Phi_d^*),\quad \forall\ \Phi_d\ne\Phi_d^*. \tag{19}$$



(a)



(b)

Fig. 2. Voice and data call performance versus data size threshold (Φ_d) with an exponentially distributed data call size ($W_{L_d} = 1.0$) and different load conditions of $\lambda_d = 2.5,\ 3.0$, and 4.0 (calls/s), respectively. (a) Data call blocking probability (B_d). (b) Mean data response time (T_d).

That is, the size threshold Φ_d should be chosen so that the mean response time T_d is minimized without increasing the data call blocking probability B_d. As such, the resource utilization is improved without degrading the QoS performance.

To evaluate the impact of data size threshold (Φ_d) on performance, we carry out some numerical analysis with the following system parameters: the mean voice call arrival rate $\lambda_v = 0.45$ (calls/s), average voice call duration $(\mu_v)^{-1} = 140$ (s), and average data call size $\overline{L}_d = 64$ (kbyte). Moreover, the parameters for the WLAN and the cell are the same as those used in [13] and [18], respectively.

Fig. 2 shows the impact of the data size threshold (Φ_d) on data call blocking probabilities (B_d) and mean data response time (T_d) in different load conditions (λ_d). It is observed that B_d and T_d only slightly decrease with Φ_d when Φ_d is relatively small. After a certain threshold such as $\Phi_d = 102.4$ (kbit), B_d begins to decrease faster with Φ_d. When Φ_d is sufficiently large, e.g., $\Phi_d \geq 640.0$ (kbit), T_d even increases exponentially with Φ_d. The phenomena observed in Fig. 2 can be explained as follows. First, the explosive increase of T_d with a large value of Φ_d is due to congestion in the cell. As seen from (3), more data traffic load is assigned to the cell when Φ_d is larger. Due to a small cell bandwidth and high occupancy by voice traffic, the data call performance is degraded substantially if the cell is overloaded. On the other hand, when Φ_d is relatively small, the decrease of T_d with Φ_d is attributed to the fact that the cell bandwidth unused by voice traffic can be efficiently utilized by small-size data calls under the SRPT. When Φ_d is sufficiently small to meet the underload condition, the larger the value of Φ_d, the more the data calls of a small size that can be assigned to the cell. Under the SRPT, the small-size data calls in the cell will not stay in the system for such a long duration as in the case where the bandwidth is shared among data calls of a large size in a PS manner.

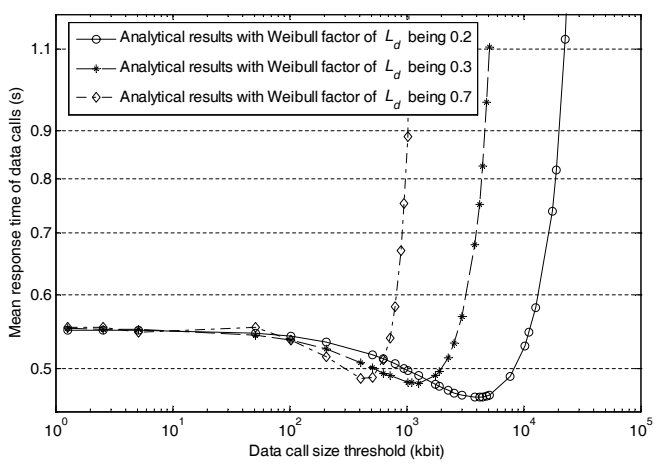


Fig. 3. Mean data response time (T_d) versus data size threshold (Φ_d) with mean data call arrival rate $\lambda_d = 3.6$ (calls/s) and different heavy-tailedness for data call size, i.e., $W_{L_d} = 0.2,\ 0.3$, and 0.7, respectively.

To further demonstrate the impact of the data size threshold with various heavy-tailedness degrees of data call size, we vary the shape parameter α_d in (1) and select the scale parameter β_d accordingly to keep the same mean value $\overline{L}_d$. Let the Weibull factor $W_{L_d} = \alpha_d$ denote the degree of heavy-tailedness. The smaller the value of W_{L_d}, the heavier the tail of the distribution of data call size. In terms of data call blocking probability, the impact of the size threshold Φ_d is similar to that of the exponential case shown in Fig. 2(a). With a smaller W_{L_d}, B_d decreases with Φ_d more slowly. Due to space limitation, the results are not shown here. Fig. 3 shows the dependence of the mean data response time T_d on the size threshold Φ_d with different heavy-tailedness of data call size. We can see that T_d first slowly decreases with Φ_d until a sufficiently large Φ_d leads to an explosive increase of T_d due to system overload. In contrast to Fig. 2(b) with an exponentially distributed data call size, the reduction of T_d with Φ_d is more evident in the heavy-tailed case. For a smaller W_{L_d} (say, 0.2), T_d decreases more slowly and can achieve an even smaller lower bound. This is due to the *"mice-elephants"* property of heavy-tailed distributions. A smaller W_{L_d} (i.e., a higher level of heavy-tailedness) implies that there is a larger fraction of even shorter data calls and that less data calls have a much larger size. Given the same size threshold Φ_d, more data calls can then be efficiently served under the SRPT in the cell. As a result, a smaller T_d is achievable with an appropriate size threshold.

TABLE I
SEARCH ALGORITHM FOR DATA SIZE THRESHOLD

1: Derive the WLAN capacity region in terms of vectors (n_v^w, n_d^w) to meet the stability constraints.
2: Derive the cell capacity region in terms of vectors (n_v^c, n_d^c) to satisfy the $\frac{E_b}{N_0}$ requirements.
3: Set the search range for the size threshold as $[\Phi_{d,min}, \Phi_{d,max}]$.
// Search for optimal Φ_d that minimizes T_d by Brent's method [27].
4: **for** $i = 1, ..., N_{iter}$ **do** // Try N_{iter} rounds of iterations at maximum.
// The constraints that $\hat{\rho}_d^c < \varepsilon$ and $T_d^c \leq T_d^w$ are incorporated by setting the evaluation of T_d to be infinitely large if these constraints are violated.
5: **if** A parabolic interpolation is acceptable **then**
6: Construct trial parabolic fits.
7: **else**
8: Resort to golden section search.
9: **end if**
10: **if** The desired precision is reached **then**
11: Exit the iteration loop.
12: **end if**
13: **end for**
14: Output the data size threshold that minimizes T_d and satisfies the preceding constraints, denoted by Φ_d^*.
15: Adapt the data size threshold in a range of $[\Phi_d^* \cdot (1-\tau),\ \Phi_d^* \cdot (1+\tau)]$ so as to minimize B_v and B_d while ensuring a T_d below the corresponding upper bound.

Taking into account the observations in Fig. 2 and Fig. 3, we propose a simple search algorithm, as given in Table I, to determine the data size threshold. Following the principles discussed at the beginning of this section, we apply the Brent's method [27] to find the optimal Φ_d^* that minimizes the mean data response time T_d. The constraints that $\hat{\rho}_d^c < 1$ and $T_d^c \leq T_d^w$ are incorporated in the Brent's method by setting the evaluation of T_d to be infinitely large if these constraints are violated. As a superlinear search method, the Brent's method can efficiently locate the minimum. In each iteration, the QoS metrics are evaluated only once with a given trial size threshold. The analytical model given in Section III-B can be employed to effectively evaluate the QoS metrics such as B_v, B_d, and T_d. Hence, the size threshold can be determined with an affordable running overhead and adapted to traffic load variations. Moreover, it is observed in Fig. 2(b) and Fig. 3 that T_d may be sensitive to Φ_d in the neighborhood of Φ_d^*. Therefore, the underload condition given in (18) is applied conservatively to guarantee system stability. As shown in Table I, the bound for the data load factor $\hat{\rho}_d^c$ is set to be ε, which is less than 1 and around 0.9. Based on Φ_d^*, the size threshold can further vary in a range of $[\Phi_d^* \cdot (1-\tau),\ \Phi_d^* \cdot (1+\tau)]$ so as to minimize B_v and B_d while ensuring a T_d below the corresponding upper bound.

IV. NUMERICAL RESULTS AND DISCUSSION

In this section, we first validate the analytical model given in Section III-B, and then compare the performance of the new load sharing scheme with the randomized load sharing scheme proposed in [18] and a service-differentiated scheme. For the randomized load sharing scheme, incoming voice and data calls are distributed to the WLAN with a probability θ_v^w and θ_d^w, respectively, and to the cell with a probability θ_v^c $(= 1 - \theta_v^w)$ and θ_d^c $(= 1 - \theta_d^w)$, respectively.

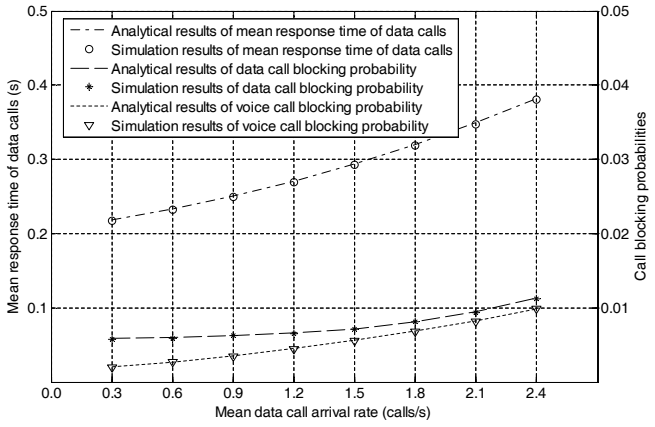


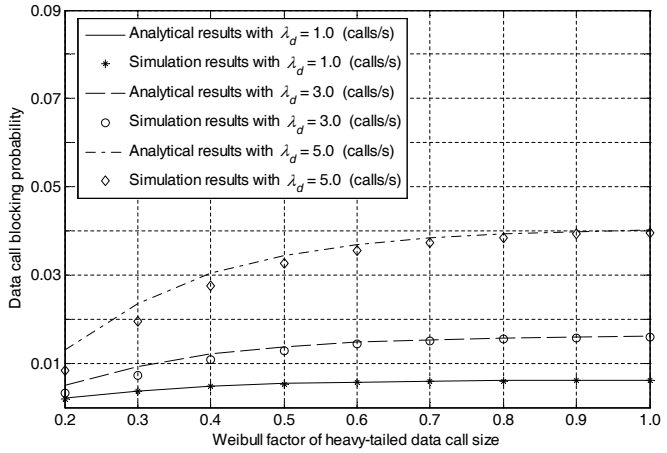
Fig. 4. Analytical and simulation results of voice and data call blocking probabilities (B_v and B_d, respectively) and mean data response time (T_d) versus mean data call arrival rate (λ_d) with an exponentially distributed data call size ($W_{L_d} = 1.0$).

The admission parameters θ_v^w and θ_d^w (or θ_v^c and θ_d^c) are determined to achieve the best performance [18]. For the service-differentiated scheme, voice calls are preferably admitted to the cell, while data calls are first distributed to the WLAN. A call rejected by its preferred network overflows to the overlay cell or WLAN to request admission. Further, dynamic call transfer is not considered in this scheme. The same system parameters are used as the preceding numerical analysis on the impact of data size threshold.

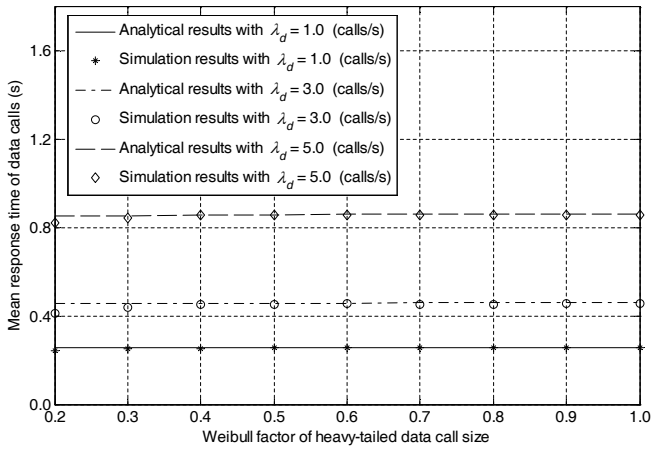
A. Accuracy Validation of Analytical Model

In Section III-B, we develop an analytical model for QoS evaluation. Based on the observations for $M/G/1/K-PS$ queueing systems, the performance of data calls in the WLAN are assumed to be insensitive to the data call size distribution with the contention-based access. The insensitivity assumption needs to be verified because the WLAN capacity is not fixed as in the analysis for $M/G/1/K-PS$ queues. To assess the validity of this assumption, we develop a discrete event-driven simulator with C/C++ language. Consistent with the system model given in Section II, a cellular cell and a contention-based WLAN are simulated to serve voice and data calls. More than 10^7 call arrivals and departures are generated in each simulation round to collect statistics on voice/data call blocking probabilities and mean response time of data calls.

Fig. 4 shows the analytical and simulation results of data call blocking probability (B_d) and mean data response time (T_d) when the data call size is exponentially distributed, i.e., the Weibull factor $W_{L_d} = 1$. A close match can be observed for different load conditions (λ_d). Fig. 5 further illustrates the cases with a heavy-tailed data call size, i.e., $0 < W_{L_d} < 1$. Similarly, the analytical results agree with the simulation results, except that the data call blocking probability is slightly overestimated when $W_{L_d} \leq 0.3$. This is due to the increase of heavy-tailedness with a small W_{L_d}. In our analytical model given in Section III-B, we assume that the performance of data calls in the WLAN is insensitive to the data call size distribution under the PS service discipline. Due to the varying WLAN capacity, the insensitivity is impaired and the call-level QoS may improve when a greater variability is induced with

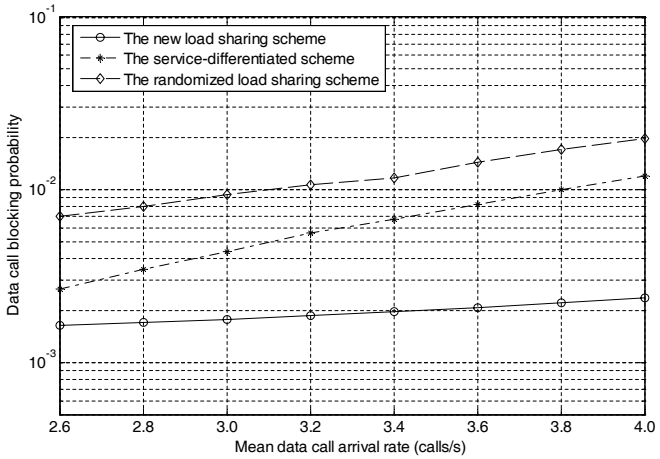


(a)

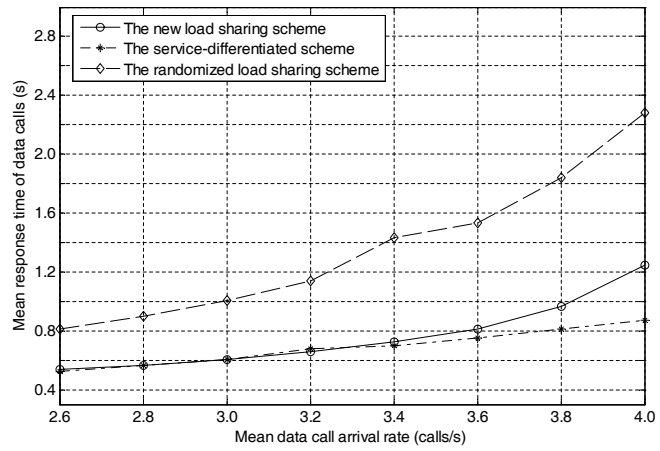


(b)

Fig. 5. Analytical and simulation results of voice and data call performance versus Weibull factor (W_{L_d}) with a heavy-tailed data call size and $\lambda_d = 1.0$, 3.0, and 5.0 (calls/s), respectively. (a) Data call blocking probability (B_d). (b) Mean data response time (T_d).



(a)



(b)

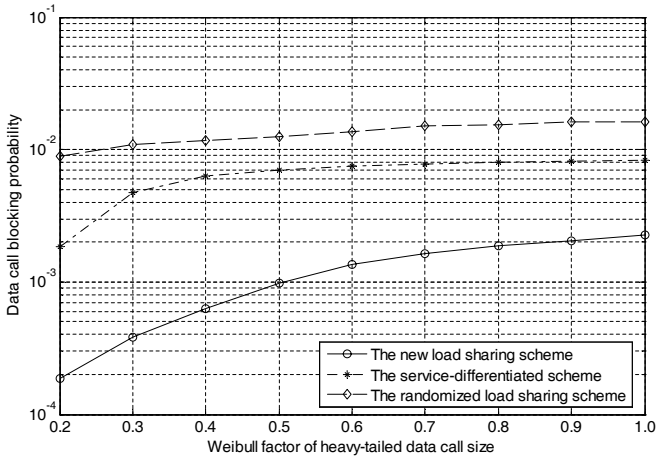
Fig. 6. Performance of different load sharing schemes versus mean data call arrival rate (λ_d) with an exponentially distributed data call size ($W_{L_d} = 1.0$). (a) Data call blocking probability (B_d). (b) Mean data response time (T_d).

the heavy-tailed call size [25]. Nonetheless, the insensitivity is expected to retain when the call blocking probabilities are sufficiently small. For example, as seen in Fig. 5, with a relatively light traffic load (say, $\lambda_d = 1.0$ calls/s) and a smaller data call blocking probability, the gap between the analytical results and simulation results is much smaller when $W_{L_d} \leq 0.3$. As the system is usually designed to ensure call blocking probabilities in the order of 10^{-3} - 10^{-2}, the analytical model in Section III-B is valid for the following performance analysis.

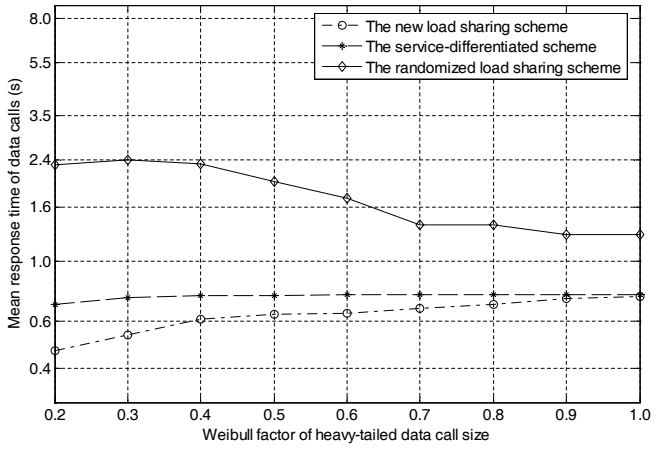
B. Performance Improvement with Proposed Load Sharing Scheme

Fig. 6 shows the performance of the three schemes in terms of data call blocking probability (B_d) and mean data response time (T_d). Significant performance improvement is observed with the new scheme. For example, in the case of $\lambda_d = 3.6$ (calls/s), B_d of the new scheme is 85.6% smaller than that of the randomized scheme, while T_d is 46.8% lower. A performance gain of 74.8% is achieved by the new scheme with respect to the service-differentiated scheme for B_d, although T_d of the two schemes is very close. In some cases, T_d of the service-differentiated scheme is even slightly lower than that of the new scheme. However, this low mean data response time of the service-differentiated scheme is achieved at the expense of much higher call blocking probabilities B_v and B_d. The new load sharing scheme still outperforms the other two schemes.

Fig. 7 shows the performance of the three schemes with different Weibull factors W_{L_d}, i.e., different heavy-tailedness degrees of the data call size. It can be seen that an even larger performance gain is achievable with the new scheme for B_d and T_d when W_{L_d} is smaller, i.e., the data call size is distributed with a heavier tail. For example, when $W_{L_d} = 0.2$, B_d of the new scheme is more than 95% smaller than those of the other two schemes, while the reduction is around 87.7% when $W_{L_d} = 0.8$. Similarly, when W_{L_d} decreases from 0.8 to 0.2, the reduction of T_d with respect to the randomized scheme increases from 49.6% to 79.7%. In comparison with the service-differentiated scheme, the new scheme reduces T_d by 7.7% when $W_{L_d} = 0.8$ and by 32.8% when $W_{L_d} = 0.2$. In addition, the reduction of T_d with W_{L_d} is due to the much higher call blocking probabilities, which restrict the total admissible traffic load to share the bandwidth.



(a)



(b)

Fig. 7. Performance of different load sharing schemes versus Weibull factor (W_{L_d}) with a heavy-tailed data call size and mean data call arrival rate $\lambda_d = 3.6$ (calls/s). (a) Data call blocking probability (B_d). (b) Mean data response time (T_d).

The significant performance gain observed in Fig. 6 and Fig. 7 lies in the fact that the new load sharing scheme not only takes advantage of the complementary QoS of the interworked systems in load distribution, but also exploits vertical handoff in dynamic call transfer to maximize the multiplexing gain. Moreover, the data size threshold can be appropriately determined with the approach given in Section III-C, which effectively takes into account the load conditions and heavy-tailedness of data call size. Nonetheless, the new scheme requires that the data call size be known *a priori* via session signaling. The signaling and control overhead for dynamic vertical handoff may increase the implementation complexity.

C. Overload Protection via SRPT Scheduling

As discussed in Section III-A, data calls in the cell are served under the SRPT, which can be enabled by the centralized resource allocation and benefit the system with the best performance achievable. The advantage of SRPT is particularly more evident in system overload when it is more challenging for the cell of small bandwidth to provide QoS guarantee. Fig. 8 compares the mean data response time (T_d)

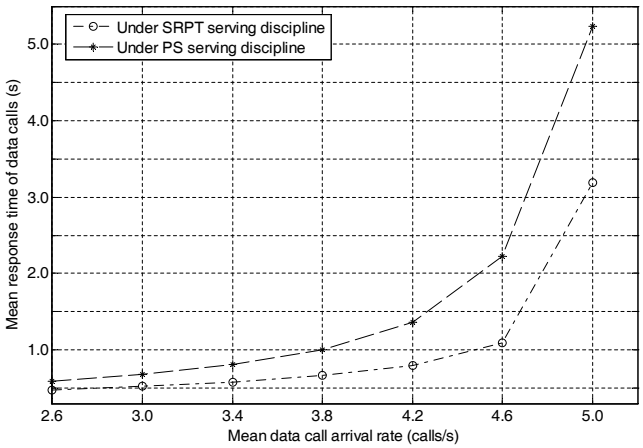


Fig. 8. Mean response time (T_d) of data calls under SRPT or PS service discipline applied in the cell.

when the SRPT or PS are applied respectively to serve data traffic in the cell. It can be seen that, when system overload occurs, T_d under the SRPT is much lower than that under the PS. At the same time, both service disciplines exhibit very close voice and data call blocking probabilities. It is known that there exists a trade-off between T_d and call blocking probabilities. That is, when more calls are admitted and share a given bandwidth, T_d increases although call blocking probabilities decrease. Hence, the observation of a significantly reduced T_d and similar call blocking probabilities implies a higher resource utilization under the SRPT.

V. Conclusions and Further Work

In this paper, we have investigated the load sharing problem for the cellular/WLAN integrated network so that the interworking is exploited to enhance multi-service provisioning. A new load sharing scheme has been proposed to effectively support voice and elastic data traffic in the integrated network. While voice calls are preferably distributed to the cell via admission control and dynamic vertical handoff, the radio resources unused by voice in the two systems are pooled to effectively serve elastic data traffic. To further overcome the cell bandwidth limitation, the efficient service discipline SRPT is applied for data calls in the cell, and only data calls with a size not greater than a threshold are admitted to the cell. The size threshold can be determined with the proposed analytical model, taking into account the load conditions and heavy-tailedness of data call size. It is observed that the new scheme significantly outperforms the randomized load sharing scheme and a service-differentiated scheme.

The cellular networks are evolving toward broadband wireless access, while many enhancement features can be introduced to WLANs with state-of-art techniques such as efficient packet scheduling. At the same time, more bandwidth-demanding services such as multimedia streaming become popular in the wireless domain. Considering the ever-increasing service demands, the load sharing is still a challenging issue, even with increased network capacity and enhanced QoS provisioning capability. In this study, we have observed that the system performance can be significantly improved by exploiting the data traffic characteristics such

as the heavy-tailed data call size in the load sharing. The study can be further extended to bandwidth-demanding multimedia services such as video streaming for the interworking of augmented cellular networks and WLANs, where unique characteristics presented by the scalable video traffic such as rate-adaptiveness should be taken into account.

REFERENCES

[1] L. E. Schrage and L. W. Miller, "The queue M/G/1 with the shortest remaining processing time discipline," *Operations Research*, vol. 14, no. 4, pp. 670–684, Jul./Aug. 1966.

[2] T. E. Klein and S.-J. Han, "Assignment strategies for mobile data users in hierarchical overlay networks: performance of optimal and adaptive strategies," *IEEE J. Select. Areas Commun.*, vol. 22, no. 5, pp. 849–861, June 2004.

[3] F. Yu and V. Krishnamurthy, "Optimal joint session admission control in integrated WLAN and CDMA cellular network," *IEEE Trans. Mobile Comput.*, vol. 6, no. 1, pp. 126–139, Jan. 2007.

[4] S. Lincke-Salecke, "Load shared integrated networks," in *Proc. 5th European Personal Mobile Commun. Conf. (EPMCC)*, Apr. 2003, pp. 225–229.

[5] W. Song, W. Zhuang, and Y. Cheng, "Load balancing for cellular/WLAN integrated networks," *IEEE Network*, vol. 21, no. 1, pp. 27–33, Jan./Feb. 2007.

[6] 3GPP, "Services and service capabilities," 3GPP TS 22.105 V8.4.0, June 2007.

[7] M. E. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic: evidence and possible causes," *IEEE/ACM Trans. Networking*, vol. 5, no. 6, pp. 835–846, Dec. 1997.

[8] K. M. Rezaul and A. Pakštas, "Web traffic analysis based on EDF statistics," in *Proc. 7th Annual PostGraduate Symposium on the Convergence of Telecommunications, Networking and Broadcasting (PGNet)*, June 2006.

[9] R. Daris and L. Torelli, "Some indices for heavy-tailed distributions," in *Proc. 31st Int'l ASTIN Colloquium*, June 2000, pp. 45–54.

[10] J. F. Huber, D. Weiler, and H. Brand, "UMTS, the mobile multimedia vision for IMT-2000: A focus on standardization," *IEEE Commun. Mag.*, vol. 38, no. 9, pp. 129–136, Sept. 2000.

[11] P. Tran-Gia and F. Hübner, "An analysis of trunk reservation and grade of service balancing mechanisms in multiservice broadband networks," in *Proc. IFIP Workshop TC6*, 1993, pp. 83–97.

[12] S. B. Fredj, S. Oueslati-Boulahia, and J. W. Roberts, "Measurement-based admission control for elastic traffic," in *Proc. 17th Int'l. Teletraffic Congress*, Dec. 2001, pp. 161–172.

[13] W. Song, H. Jiang, and W. Zhuang, "Performance analysis of the WLAN-first scheme in cellular/WLAN interworking," *IEEE Trans. Wireless Commun.*, vol. 6, no. 5, pp. 1932–1952, May 2007.

[14] A. K. Salkintzis, G. Dimitriadis, D. Skyrianoglou, N. Passas, and N. Pavlidou, "Seamless continuity of real-time video across UMTS and WLAN networks: challenges and performance evaluation," *IEEE Wireless Commun. Mag.*, vol. 12, no. 3, pp. 8–18, June 2005.

[15] H. Zhu, L. Ming, I. Chlamtac, and B. Prabhakaran, "A survey of quality of service in IEEE 802.11 networks," *IEEE Wireless Commun. Mag.*, vol. 11, no. 4, pp. 6–14, Aug. 2004.

[16] H. Zhai, X. Chen, and Y. Fang, "How well can the IEEE 802.11 wireless LAN support quality of service?" *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, pp. 3084–3094, Nov. 2005.

[17] J. Pérez-Romero, O. Sallent, R. Agusti, and M. A. Diaz-Guerra, *Radio Resource Management Strategies in UMTS*. New York: Wiley, 2005.

[18] W. Song, Y. Cheng, and W. Zhuang, "Improving voice and data services in cellular/WLAN integrated network by admission control," *IEEE Trans. Wireless Commun.*, vol. 6, no. 11, pp. 4025-4037, Nov. 2007.

[19] E. Vanem, S. Svaet, and F. Paint, "Effects of multiple access alternatives in heterogeneous wireless networks," in *Proc. IEEE WCNC*, vol. 3, Mar. 2003, pp. 1696–1700.

[20] R. W. Wolff, *Stochastic Modeling and the Theory of Queues*. Prentice Hall, 1989.

[21] S. B. Fredj, T. Bonald, A. Proutière, G. Régnié, and J. W. Roberts, "Statistical bandwidth sharing: a study of congestion at flow level," in *Proc. ACM SIGCOMM*, Aug. 2001, pp. 111–122.

[22] L. Kleinrock, *Queueing Systems, Volume 1: Theory*. New York: John Wiley and Sons, 1975.

[23] N. Bansal and M. Harchol-Balter, "Analysis of SRPT scheduling: investigating unfairness," *ACM SIGMETRICS Performance Evaluation Review*, vol. 29, no. 1, pp. 279–290, June 2001.

[24] M. Garcia-Martin, M. Isomaki, G. Camarillo, and S. Loreto, "A session description protocol (SDP) offer/answer mechanism to enable file transfer," Internet draft, June 2007.

[25] R. Litjens and R. J. Boucherie, "Elastic calls in an integrated services network: the greater the call size variability the better the QoS," *Perform. Eval.*, vol. 52, no. 4, pp. 193–220, May 2003.

[26] F. Delcoigne, A. Proutière, and G. Régnié, "Modeling integration of streaming and data traffic," *Perform. Eval.*, vol. 55, no. 3-4, pp. 185–209, Feb. 2004.

[27] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C, The Art of Scientific Computing (2nd ed.)* Cambridge: Cambridge University Press, 1999.

Wei Song (S'07) received the B.S. degree in electrical engineering from Hebei University, China, in 1998, the M.S. degree in computer science from Beijing University of Posts and Telecommunications, China, in 2001, and the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Canada, in 2007. She is now a postdoctoral research fellow. She received the Best Student Paper Award from IEEE WCNC 2007 and postdoctoral fellowship from NSERC of Canada in 2008. Her current research interests include the interworking of cellular networks and wireless local area networks, and multimedia service provisioning.

Weihua Zhuang (M'93-SM'01-F'08) received the Ph.D. degree in electrical engineering from the University of New Brunswick, Canada. Since October 1993, she has been with the Department of Electrical and Computer Engineering, University of Waterloo, Canada, where she is a Professor. Dr. Zhuang is a co-author of the textbook *Wireless Communications and Networking* (Prentice Hall, 2003). Her current research interests include multimedia wireless communications, wireless networks, and radio positioning. She received the Premiers Research Excellence Award in 2001 from the Ontario Government for demonstrated excellence of scientific and academic contributions, the Outstanding Performance Award in 2005 and 2006, and 2008 from the University of Waterloo, the Best Paper Awards from IEEE WCNC 2007, IEEE ICC 2007, and Qshine 2007 and 2008. Dr. Zhuang is a Fellow of the IEEE and serves as the Editor-in-Chief of IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and an Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, EURASIP JOURNAL ON WIRELESS COMMUNICATIONS AND NETWORKING, and INTERNATIONAL JOURNAL OF SENSOR NETWORKS.