Resource Management for QoS Support in Cellular/WLAN Interworking

Wei Song, Hai Jiang, Weihua Zhuang, and Xuemin (Sherman) Shen, University of Waterloo

Abstract

To provide mobile users with seamless Internet access anywhere and anytime, there is a strong demand for interworking mechanisms between cellular networks and wireless local area networks in the next-generation all-IP wireless networks. In this article we focus on resource management and call admission control for QoS support in cellular/WLAN interworking. In specific, a DiffServ interworking architecture with loose coupling is presented. Resource allocation in the interworking environment is investigated, taking into account the network characteristics, vertical handoff, user mobility, and service types. An effective call admission control strategy with service differentiation is proposed for QoS provisioning and efficient resource utilization. Numerical results demonstrate the effectiveness of the proposed call admission control scheme.

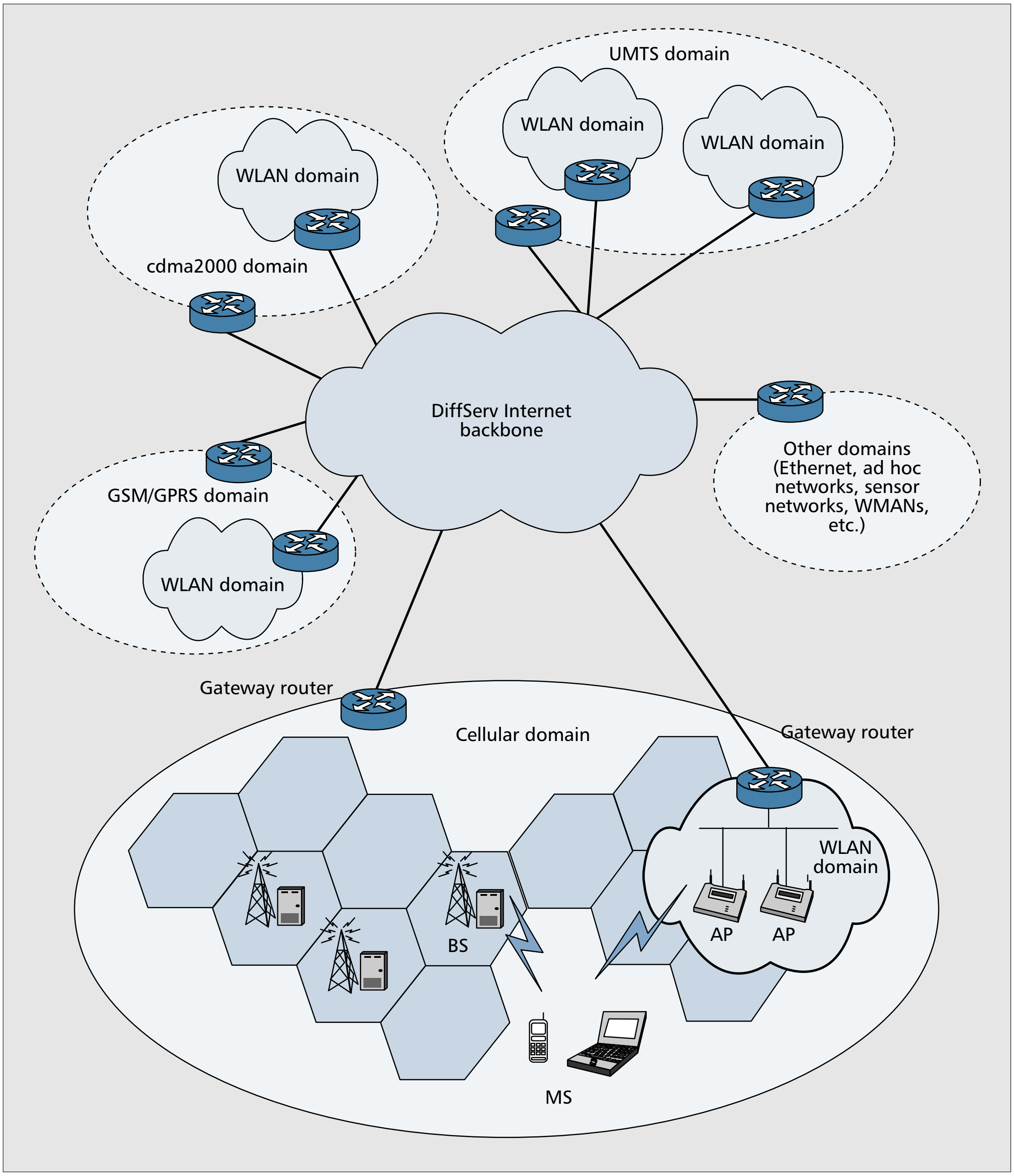
The past decade has witnessed the fast evolution and successful deployment of a number of wireless access networks. The two most promising ones are cellular networks and wireless local area networks (WLANs). Originally aimed at high-quality circuit-switched voice service with wide area coverage, cellular networks have been well deployed around the world. The second-generation (2G) cellular networks (e.g., Global System for Mobile Communications [GSM] and IS-95) were a revolution from analog to digital technology. 2.5G cellular networks such as General Packet Radio Service (GPRS) provide packet-switched low-rate (approximately up to 100 kb/s) data services. With the ever-increasing demand for high-rate multimedia services, after many indoor and outdoor tests, commercial third-generation (3G) cellular networks have been deployed in many countries, and are expected to provide multimedia services with a maximum bit rate of 2 Mb/s. On the other hand, WLANs have shown their potential to provide higher-rate data services at lower cost over local area coverage. Working in the license-exempt 2.4 GHz industrial, scientific, and medical (ISM) frequency band, the IEEE 802.11b WLAN offers a data rate up to 11 Mb/s, while the IEEE 802.11a WLAN and European Telecommunications Standard Institute (ETSI) HIPERLAN/2 can support a data rate up to 54 Mb/s in the 5 GHz frequency band. Compared to cellular networks, WLANs typically cover a smaller geographic area as a wireless extension to wired Ethernet. As a result, WLANs are feasible candidates for high-rate data service provisioning at hotspot areas with low user mobility.

Driven by the service *anywhere* and *anytime* concept, it is well accepted that fourth-generation (4G) wireless networks will be heterogeneous, integrating different networks to provide seamless Internet access for mobile users with multimode access capability. The well deployed cellular networks and WLANs will both be included along with other wireless access networks such as ad hoc networks, sensor networks, and wireless metropolitan area networks (WMANs). One major challenge in cellular/WLAN interworking is how to take advantage of the wide coverage and almost universal roaming support of cellular networks and the high data rates of WLANs. Many issues should be carefully addressed to achieve seamless interworking, such as mobility management, resource allocation, call admission control (CAC), security, and billing. This article focuses on resource allocation and CAC in the integration of cellular networks and WLANs. First, a differentiated services (DiffServ) interworking architecture with loose coupling is presented. Second, we discuss the resource allocation issue in cellular/WLAN interworking. Third, we propose an effective call admission scheme for voice and data services in integrated cellular and WLAN networks. Finally, the performance of the proposed scheme is evaluated to show its strengths and impact on the system.

Interworking Architecture

Many interworking architectures have been proposed to integrate cellular networks and WLANs for 4G wireless communications, mainly aimed at augmenting cellular networks with high-rate data services by WLANs in hotspots. Based on the interdependence between the two access networks, the interworking architectures can be classified into two categories.

Tight coupling architecture: The WLAN is connected to the cellular core network, and appears to the cellular core network as one cellular radio access network. For example, the integration point of WLANs to a GPRS/Universal Mobile Telecommunications System (UMTS) core network can be the serving GPRS support node (SGSN) [1] or gateway GPRS support node (GGSN) [2]. A user roaming across the two domains is based on the mobility management protocols of the cellular networks, thus enhancing the interdomain mobility management capability. The main disadvantages of the tight coupling approach are that:



■ Figure 1. *Loosely coupled cellular/WLAN interworking architecture in a DiffServ platform.*

- An interface in the cellular core network exposed to WLANs is required, which is a challenge as the two domains are likely to be developed and deployed independently by different operators.
- A large volume of WLAN traffic will go through the cellular core network, possibly making the latter a network bottleneck.
- WLANs need to have a protocol stack compatible with that of cellular networks. The induced complexity and cost may hamper the deployment of a tight coupling architecture [2].

Loose coupling architecture: The gateway directly connects the WLANs to the Internet backbone, and there is no direct link between the WLANs and the cellular core network [2]. The main advantage of the loose coupling approach is the independent deployment of the two domains. However, as the two domains are separated, the mobility signaling may traverse a relatively long path, thus inducing relatively high handoff latency.

On the other hand, to interconnect heterogeneous IP-based wireless access networks with the Internet backbone in 4G networks, it is well recognized that an all-IP DiffServ platform [3] is the most promising architecture to provision broadband seamless global access for the following reasons [4]:

- Based on a limited number of service classes, DiffServ is a scalable mechanism as no per-flow processing is needed in core networks.
- The DiffServ platform adopts a domain-based architecture, where each domain can freely and independently choose its own system mechanisms as long as its service level agreements (SLAs) with neighboring domains are satisfied [3]. Such a domain-based architecture allows the flexibility and convenience to deploy each domain independently, and to develop, modify, or exchange the techniques in a domain without a significant impact on the overall system.

- The newly emerged IEEE 802.11e draft for WLANs aims at provisioning quality of service (QoS) in a relative sense, which is similar to and can be mapped smoothly to the relative QoS in DiffServ.
- A fast handoff procedure is required for seamless roaming in and among wireless access networks. The popular solution for fast handoff is to use Mobile IP for interdomain (macro-) mobility and to use micromobility protocols for intradomain mobility. Micromobility protocols can be seamlessly incorporated in the domain-based DiffServ platform.

Loosely coupled interworking of cellular networks and WLANs matches well with the emerging evolution toward an all-IP 4G infrastructure, and can be naturally implemented in a domain-based DiffServ platform. Figure 1 illustrates the integration architecture of local coverage WLANs with different wide coverage cellular networks such as UMTS, code-division multiple access (cdma2000), and GSM/GPRS. The base station (BS) or access point (AP) provides the mobile station (MS) with Internet access. All the wireless domains are interconnected through the DiffServ Internet backbone to provide end-to-end Internet services to an MS.

Resource Allocation

For traditional cellular networks, resource allocation plays an essential role in effectively provisioning QoS guarantee to each MS during its traffic lifetime and efficiently utilizing scarce radio resources, and has been extensively studied. When WLANs are integrated with cellular networks, it is much more challenging to achieve QoS provisioning and efficient resource utilization, due to the heterogeneous nature of the networks and the limited QoS support in WLANs. To develop an effective resource allocation solution for a cellular/WLAN interworking scenario, various factors should be considered, weighed, and balanced.

Heterogeneous Network Characteristics

Resource allocation solutions differ in cellular networks and WLANs due to the heterogeneity in the physical, medium access, and link control layers. In cellular networks, based on a centralized architecture, the BS has the ability to provide QoS guarantee to MSs via properly scheduling their access to the wireless channel, taking advantage of the information available in the BS and collected from MSs. Furthermore, the schedulers located in different BSs can also coordinate with each other to improve overall system performance. On the other hand, in the current WLAN standard IEEE 802.11, two channel access functions are defined: the mandatory distributed coordination function (DCF) and optional point coordination function (PCF) in a centrally controlled manner. The most popularly commercialized DCF is based on carrier sense multiple access with collision avoidance (CSMA/CA) and binary exponential backoff, which cannot guarantee user QoS requirements. The IEEE 802.11e draft enhances the QoS provisioning capability by a differentiation mechanism. However, only relative QoS is provided. As a result, the fine QoS provisioning in cellular networks and the relatively weak QoS support capability of WLANs need to be taken into account for resource allocation in cellular/WLAN interworking.

Hierarchical Network Architecture

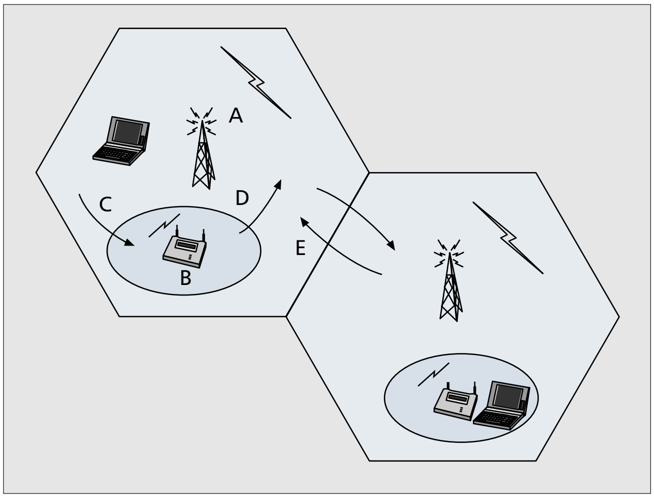
As shown in Fig. 1, cellular networks provide ubiquitous connectivity through wide area coverage (the upper tier), and WLANs (the lower tier) are deployed disjointly in hotspot local areas where the traffic intensity is usually much higher than that in other areas. With this two-tier overlay structure, both cellular access and WLAN access are available to MSs within WLAN-covered areas. As a result, the handoff from a cellular network to its overlaid WLANs is optional, and mainly happens to enhance QoS, lower cost, and balance traffic load. The handoff decision has a significant impact on the resource utilization efficiency and QoS guarantee. Due to network heterogeneity, the handoff between the cellular network and WLANs is referred to as *vertical handoff*, in contrast with *horizontal handoff* between BSs or APs within a homogeneous wireless network. The vertical handoff decision algorithm is an important aspect in cellular/WLAN interworking and is quite different from that in horizontal handoff [5].

Dynamic User Mobility

An MS with a higher moving speed may experience more handoffs during its call lifetime. The handoff procedure may cause extra delay, packet losses, and even connection interruption. The situation is even worse with vertical handoff as, in the loose coupling model, the handoff signaling is very likely to travel through a relatively long path compared to that in a horizontal handoff. Thus, it is desired to assign MSs with high mobility to a network with large coverage (i.e., the cellular network). This principle has been extensively studied in hierarchical cellular networks where microcells are overlaid with macrocells [6]. Furthermore, WLANs are generally deployed for indoor environments like cafés, offices, and airports. Users within these areas are mostly static or only maintain pedestrian-level mobility. The low user mobility level results in a heavy-tailed residence time of a user staying within a WLAN-covered area [7]. Consequently, within the coverage of a large cell, a uniform mobility model for the cell will not be applicable. The user mobility model varies even within the coverage of a single cell. As user residence time in a cell or WLAN is closely related to channel occupancy time, this new characteristic results in more complex analysis of resource allocation.

Multiple Traffic (Service) Types

Different traffic types usually require different QoS deliveries. Real-time services such as voice and video are sensitive to delay, while the main concern for delay-tolerant data service is throughput. Benefiting from a centralized architecture, cellular networks can serve real-time traffic effectively. In contrast, due to distributed control and the backoff mechanism in channel access, it is rather challenging for WLANs to meet the strict delay requirements of real-time services. However, WLANs are more efficient than cellular networks in serving bursty data traffic. Moreover, data service often experiences traffic load asymmetry in the uplink and downlink (e.g., the downlink is characterized by a larger traffic volume). The asymmetry level is likely to change with time. The mainstream cellular networks have adopted frequency-division duplexing (FDD), which is not flexible enough to handle such traffic load asymmetry. System capacity is limited by the uplink or downlink with a larger load, resulting in a waste of the precious radio bandwidth in the other link. Different from cellular networks, WLANs can be viewed as operating in a virtual time-division duplexing (TDD) mode, which can effectively handle the load asymmetry of data service. Also, the high service rate of WLANs can be fully utilized by data traffic. As multiservice support is a basic requirement for future wireless networks, service type becomes an important factor in resource allocation for cellular/WLAN interworking. At the same time, multiservice support offers opportunities to fully utilize the overall resources of the cellular network and WLANs. This is actually an important motivation to consider the interworking of the two networks.



■ Figure 2. *New and handoff call arrivals in cellular/WLAN interworking.*

Admission Control with Service Differentiation

With the unique characteristics presented by the integrated cellular and WLAN network, resource allocation becomes a very complex issue with many new challenges. As an important aspect, admission control helps to provide target QoS for multiple services and to utilize the overall resources in an efficient manner.

As discussed in the previous section, various new factors should be taken into account to develop an effective admission scheme for cellular/WLAN interworking. With the two-tier overlay structure, for an incoming service request into a WLAN-covered area, a decision needs to be made on whether to admit the incoming call to the WLAN or to the overlaying cell. Due to the heterogeneous underlying technologies, the choice can have a significant impact on resource utilization efficiency and QoS satisfaction. If the total resources of the two tiers are allocated to users by jointly considering factors such as available capacity, traffic characteristics, user mobility, and QoS support capability, higher utilization and better QoS assurance can be achieved.

Moreover, considering user mobility, handoff traffic should be differentiated from new traffic in terms of call admission. In particular, the admission of vertical handoff calls needs to be controlled appropriately. When a user moves from an area with only cellular coverage to an overlaid WLAN area, the ongoing call of the user can be handed over to the WLAN for balancing load, lowering cost, and so on. If there is no spare capacity in the WLAN to accommodate the handoff call, the call can remain in the cellular network. On the other hand, for a new call originating within a WLAN-covered area, either the covering cell or the WLAN is first selected for admission according to the service type, user mobility, and current network status. If the service request is rejected by the preferred network, it can overflow to the other network for admission, based on whether or not the benefit is larger than the cost.

The above problems are rarely addressed in the literature. Much research work [5] has been focused on the vertical handoff decision process for cellular/WLAN interworking, which aims at minimizing unnecessary handoffs and the impact of the ping-pong effect by properly considering the metrics collected from the cellular network and WLANs. However, little research considers the admission decision from the perspective of resource utilization by taking into account the unique user traffic and mobility characteristics in the heterogeneous environment. There is some related work on the similar problems in two-tier hierarchical cellular networks, in which small-sized microcells are overlaid with large macrocells. However, many proposed schemes [8] focus on allocating the resources of the two tiers depending on the user mobility level, and usually only one service type (i.e., voice) is considered. Also, in many schemes the problems are simplified under the assumption that continuous coverage is provided by both tiers, and there is no vertical handoff. In the following, we study resource sharing between voice and data services, and the related call admission issue in cellular/WLAN interworking.

Resource Sharing between Voice and Data Services

In this research we consider two most popular services: voice and interactive data service such as Web browsing and file transfer. Generally, voice traffic requires real-time transmission and is tolerant to a certain level of packet loss. The transmission rate of an active voice call is stable, thus having low elasticity in bandwidth requirements. In contrast, data traffic is normally delay-insensitive and tolerant to transmission rate variations, but requires reliable transmission. If download of a Web document or transfer of a data file is viewed as a packet data call, the data transfer time (i.e., the data call duration) usually depends on the file size and bandwidth allocated. Given the interactive nature of the data applications considered here, the mean transfer time for data calls should be bounded below a threshold. In addition, constraints on new call blocking probability and handoff call dropping probability need to be met.

To properly share the total bandwidth between voice and data services in each network, the restricted access mechanism [9] is used. Voice traffic is offered preemptive priority over data traffic, and occupies up to a certain amount of bandwidth to meet its strict QoS requirements. The remaining bandwidth is dedicated to data traffic. Moreover, to achieve higher resource utilization by taking traffic dynamics into account, all bandwidth unused by current voice traffic is shared equally by ongoing data flows. Then the number of data calls admitted in a cell or WLAN should be restricted, so that the constraint for mean data transfer time can be satisfied. This restricted access mechanism is to strike a good balance between high utilization and fine QoS guarantee.

Call Admission for Voice and Data Traffic

With the different traffic characteristics and service provisioning capability of each network, admission strategies should vary with service type. WLANs are usually deployed in disjoint hotspot areas with small coverage instead of continuous wide coverage (which can be provided by cellular networks). Consequently, to minimize the impact of latency and processing overhead induced by frequent vertical handoff, voice calls are admitted with a preference to the cellular network. Another reason for selecting the cellular network for voice calls is the fine QoS provisioning provided by the cellular network and required by voice traffic to meet its strict delay requirement. On the other hand, data traffic has better rate adaptation capability. With larger bandwidth provided by a WLAN, transmission of the packets belonging to a data call will finish sooner, consuming the allocated resources for less time. As a result, it is likely that a data call will end within the WLAN coverage and not need to hand over to the cellular network when the MS moves out of WLAN coverage. For voice calls, however, the channel occupancy time is not

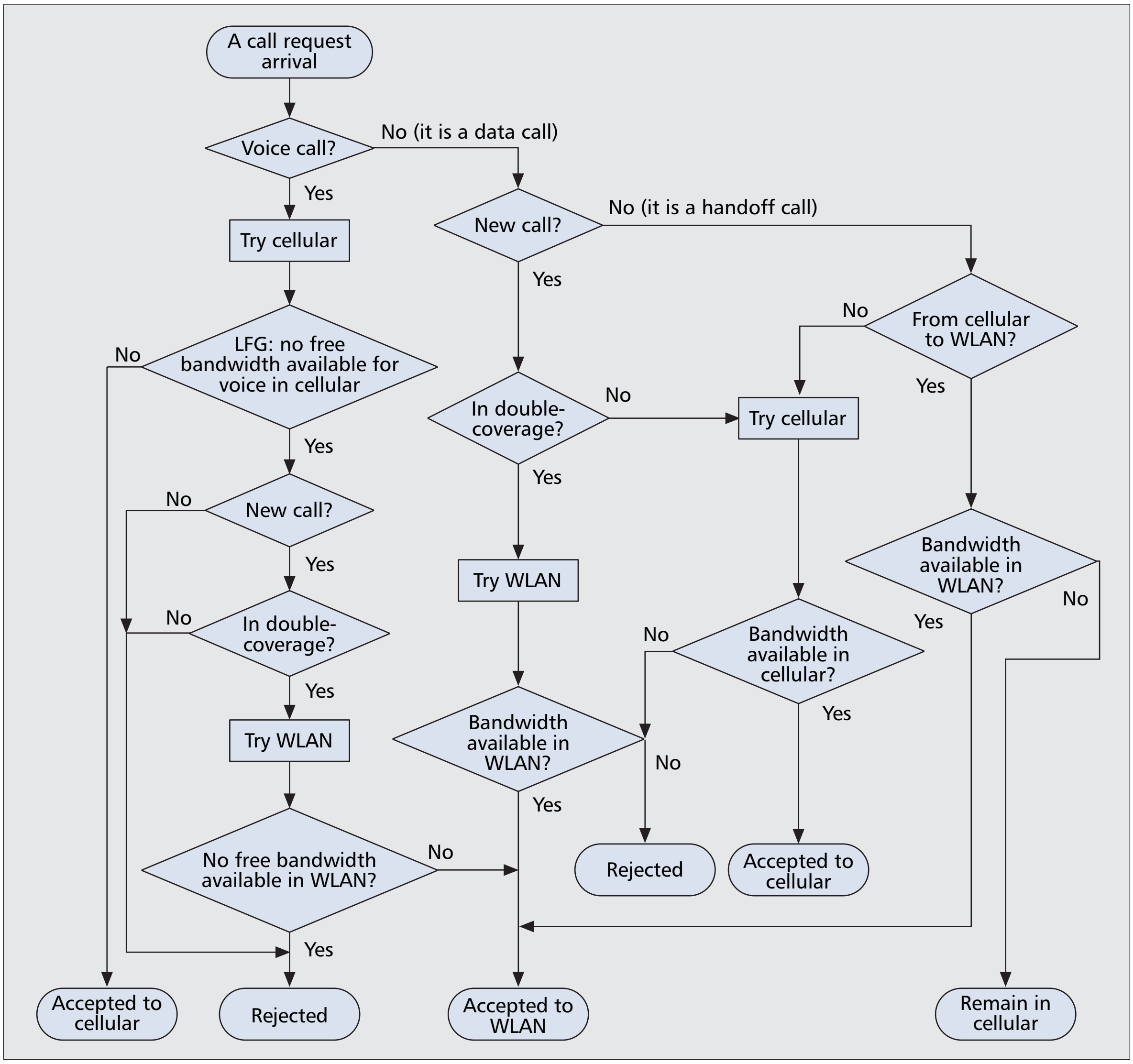


Figure 3. *Call admission procedure for cellular/WLAN interworking.*

reduced even when larger bandwidth is available. Only limited resources of the WLAN can be utilized by voice traffic, with a large overhead from frequent vertical handoff due to the small coverage of WLANs.

The details of the admission scheme are elaborated as follows. In Fig. 2 we illustrate the traffic arrivals in a simple scenario of cellular/WLAN interworking. First consider new call arrivals to the system:

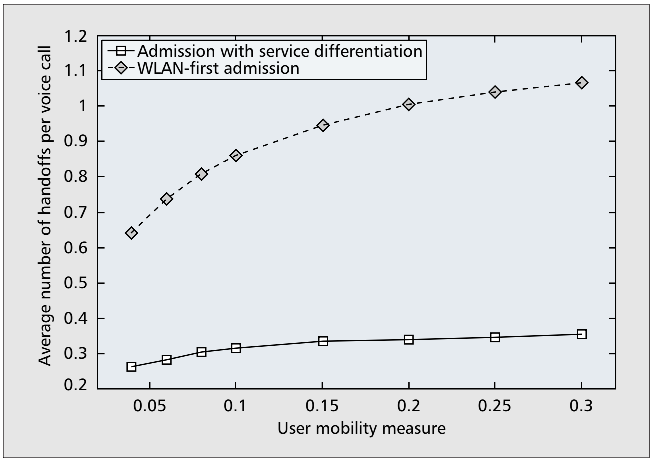
• In the area with only cellular coverage (referred to as *cellular-only area* in the following), area A in Fig. 2, both new voice and data calls have to request admission to the covering cell.

• There are two choices for the new voice calls in the area with WLAN coverage (area B in Fig. 2), which is referred to as *double-coverage area* in the following. A new voice call will first try to get admission to the cell. The request is rejected if there is not enough bandwidth to accommodate the new voice call even after all existing data calls in the cell are degraded to have the minimum bandwidth. Then the new voice call blocked by the cell overflows to the WLAN to request admission. Only when both the covering cell and WLAN reject the voice service request will the call leave the system.

• The first admission choice for new data calls in the double-coverage area is the WLAN. If the admission request of a data call is rejected by the WLAN due to lack of sufficient bandwidth, it will not try the cell, as the overflow of data traffic to the cellular part brings little benefit to data calls, but may severely affect the capability of the cellular network in carrying voice traffic.

• To provide priority to voice (handoff and new) calls in the cellular-only area, a limited fractional guard channel (LFG) policy [10] is used. A certain amount of cell bandwidth is reserved for prioritized voice traffic, including vertical handoff calls from the WLAN to its overlaying cell, horizontal handoff calls between neighboring cells, and new calls originating within the cellular-only area. The LFG policy is proven to be optimal in terms of minimizing the number of voice channels required subject to hard constraints on the new voice call blocking and handoff voice call dropping probabilities.

On the other hand, for an ongoing voice or data call moving out of the WLAN (arrow D in Fig. 2) or from one cell to another within the cellular network (arrow E in Fig. 2), a handoff should proceed; otherwise, the call will be dropped. The case is different for ongoing calls from a cell to its overlaid WLANs (arrow C in Fig. 2). If the mean transfer time for data calls is not violated when the incoming data call is admitted to the WLAN, the data call is handed over from the cell to the WLAN; otherwise, it will remain in the cell. In contrast, for an ongoing voice call, no handoff to the WLAN will proceed so that the voice call remains in the cellular network in



■ Figure 4. *Average number of handoffs per voice call (in comparison with WLAN-first scheme) vs. user mobility in double-coverage area.*

order to avoid the impact of handoff latency. This is also because real-time voice traffic cannot fully take advantage of the large bandwidth of WLANs to reduce the call duration and accordingly the channel occupancy time. The call admission procedure is summarized in Fig. 3.

Performance Evaluation

In our work we assume a traditional layered protocol stack. Using the services provided by the lower layer, each protocol layer deals with a specific task and provides transparent service to the upper layer. Such an architecture allows the flexibility to modify or change the techniques in a protocol layer without significant impact on the overall system design. Hence, the medium access control (MAC) and physical (PHY) layers can be viewed as transparent to the upper layers if their impact is represented by the total capacity provided to the network layer and the effective bandwidth requirement of each call. Our scheme can be applied to different MAC and PHY layer mechanisms, as long as the total capacity and effective bandwidth requirements of different traffic types are set accordingly.

We consider the case where one WLAN is located in each cell, and the overall system is assumed to be in statistical equilibrium, in which each cell is statistically the same as any other cell [11]. Due to the low user mobility pattern within hotspot areas, the residence time of an MS in a double-coverage area conforms to heavy-tailed distributions, while the cell residence time is exponentially distributed. The heavy-tailed distributions of residence times make performance analysis very complex. To address the problem, an effective approach of fitting heavy-tailed distributions with hyperexponential distributions is used.

Both voice and data call arrivals are assumed to be Poisson. The voice traffic model is relatively simple as the voice call duration is exponential, and a constant effective bandwidth is required to guarantee the QoS. On the contrary, data calls are more elastic to bandwidth variations, and the data call duration (i.e., data transfer time) depends on the data file size, allocated bandwidth to the data call, and system steady state distribution. It is difficult to obtain precise statistics for the transfer time of data calls even when the data file size is assumed to be exponentially distributed [11]. Measurements have demonstrated that the data file size follows heavy-tailed distributions such as lognormal and Pareto distribution. In our call admission scheme, upper and lower bounds are used to evaluate the mean data transfer time [12]. As all the bandwidth unused by current voice traffic is shared equally by ongoing data calls, the data service provisioning of the cell or the WLAN can be modeled as an M/G/1/K-Processor Sharing (PS) queue. For the M/G/1/K-PS queue, if the bandwidth available to data service is constant, the mean data transfer time is insensitive to the specific file size distribution. However, because the bandwidth allocated to data service actually varies with voice traffic dynamics and user mobility, the insensitivity does not hold anymore. As a result, upper and lower bounds of the mean data transfer time [12] are used for approximation.

By properly setting the effective bandwidths of voice calls and data calls, packet-level QoS such as packet delay and packet loss can be guaranteed, as long as the allocated bandwidth to a voice or data call is no less than the corresponding effective bandwidth requirement. In the following we focus on call-level QoS in terms of call blocking/dropping probabilities, mean data transfer time, and number of handoffs per voice call. Based on the proposed admission strategy, the admission region of a cell or WLAN for voice and data traffic can be derived with the voice call blocking/dropping probabilities and mean data transfer time bounded. Please refer to [12] for more details of the derivation of admission region. Given the system parameters listed in Table 1, we first obtain the numbers of voice channels required in a cell and a WLAN to meet the new call blocking probability and handoff call dropping probability constraints. Then, based on the upper bound for the mean data transfer time, the maximum numbers of data calls that can be accommodated in a cell and a WLAN are obtained. Due to the coupling between the cellular network and WLANs in terms of resource allocation, intuitively an increase of the admission region of one network may result in a reduction of that in the other. As the WLAN has larger bandwidth, it is more likely for the cellular part to become the service bottleneck. Therefore, we select the configuration of admission region which maximizes the numbers of voice and data calls in the cellular network. This is reasonable since one main purpose of cellular/WLAN interworking is to complement the cellular network by effectively utilizing the additional capacity provided by WLANs.

We compare our scheme with a common sense admission scheme referred to as *WLAN-first* scheme. In the WLAN-first scheme, both new voice calls and new data calls select the WLAN as their service preference in the double-coverage area to get the benefit of possible larger bandwidth and less cost [5]. If rejected, they then try the cellular network for

Parameter	Value
Cell capacity (Mb/s)	2
WLAN capacity (Mb/s)	5
Ratio of WLAN coverage area to that of the cellular cell	0.1
Average voice call duration time (s)	180
Average data file size (Kbytes)	64, 80
Mean data transfer time upper bound (s)	3
New voice call blocking probability upper bound	0.01
Handoff voice call dropping probability upper bound	0.001

■ Table 1. *Parameters used in performance evaluation.*

admission. Whenever WLAN coverage is available, the ongoing voice calls and data calls are handed over to the WLAN if there is enough spare bandwidth to accommodate the handoff calls while maintaining the QoS of existing calls.

Figure 4 shows the performance of our admission scheme and the WLAN-first scheme in terms of handoff rate. User mobility within the double-coverage area can be characterized by the WLAN residence time, which depends on factors such as user moving speed and the size of WLAN coverage. In general, a short residence time indicates a high mobility level. Here, we use the ratio of average voice call duration to mean WLAN residence time to indicate the user mobility level in the double-coverage area, and refer to it as *user mobility measure*. As seen in Fig. 4, the performance improvement for voice calls in our scheme is obvious. The average handoff number of a voice call is significantly reduced, which means much less latency and fewer packet losses due to handoffs. The improvement increases with higher user mobility. When mobility is higher, the WLAN-first scheme leads to much more frequent vertical handoffs and degrades the performance of voice calls. On the other hand, the call blocking/dropping probability of data service in our scheme is slightly higher than that of the WLAN-first scheme. This is because more voice calls are admitted to the cellular network in our scheme, resulting in less bandwidth available for data traffic in the cellular network. This trade-off is acceptable for better QoS provisioning to high-priority voice traffic.

Conclusions and Further Work

In this article we have investigated resource allocation issues in cellular/WLAN interworking for 4G wireless communications. Loose coupling in an all-IP DiffServ architecture has been considered for effective interworking. A new admission strategy for integrated voice and data services has been proposed, which prefers the cellular network for voice service and WLANs for data service, according to the characteristics of the cellular network and WLANs, the distinct features of voice and data traffic and their QoS requirements, and user mobility patterns. The admission strategy can effectively improve the performance of high-priority voice service and, at the same time, fully utilize the large bandwidth of WLANs for data service.

There are still many open issues that need further investigation. It is necessary to consider the emerging WLANs enhanced with better QoS provisioning capability. Resource allocation in the integrated cellular and WLAN network for video service is another important issue, since video is also a typical real-time service with strict delay/jitter requirements. In addition, with layered coding, video traffic is more adaptive to bandwidth variation than traditional voice service. As a result, multimedia services (including video) pose more technical challenges for QoS support in the heterogeneous cellular/WLAN networking environment.

Acknowledgments

This research has been supported by a grant from the Bell University Laboratories (BUL) at the University of Waterloo, Canada.

References

[1] A. Salkintzis, C. Fors, and R. Pazhyannur, "WLAN-GPRS Integration for Next-Generation Mobile Data Networks," *IEEE Wireless Commun.*, vol. 9, no. 5, Oct. 2002, pp. 112–24.

[2] M. Buddhikot *et al.*, "Integration of 802.11 and Third-Generation Wireless Data Networks," *Proc. IEEE INFOCOM*, vol. 1, Apr. 2003, pp. 503–12.

[3] B. Moon and H. Aghvami, "DiffServ Extension for QoS Provisioning in IP Mobility Environments," *IEEE Wireless Commun.*, vol. 10, no. 5, Oct. 2003, pp. 38–44.

[4] Y. Cheng *et al.*, "Efficient Resource Allocation for China's 3G/4G Wireless Networks," *IEEE Commun. Mag.*, vol. 43, no. 1, Jan. 2005, pp. 76–83.

[5] Q. Zhang *et al.*, "Efficient Mobility Management for Vertical Handoff between WWAN and WLAN," *IEEE Commun. Mag.*, vol. 41, no. 11, Nov. 2003, pp. 102–08.

[6] T. Klein and S.-J. Han, "Assignment Strategies for Mobile Data Users in Hierarchical Overlay Networks: Performance of Optimal and Adaptive Strategies," *IEEE JSAC*, vol. 22, no. 5, June 2004, pp. 849–61.

[7] S. Thajchayapong and J. M. Peha, "Mobility Patterns in Microcellular Wireless Networks," *Proc. IEEE WCNC*, vol. 3, Mar. 2003, pp. 1963–68.

[8] W. Shan, P. Fan, and Y. Pan, "Performance Evaluation of a Hierarchical Cellular System with Mobile Velocity-Based Bidirectional Call-Overflow Scheme," *IEEE Trans. Parallel and Distrib.Sys.*, vol. 14, no. 1, Jan. 2003, pp. 72–83.

[9] M. Naghshineh and A. S. Acampora, "QoS Provisioning in Microcellular Networks Supporting Multiple Classes of Traffic," *Wireless Networks*, vol. 2, no. 3, Aug. 1996, pp. 195–203.

[10] R. Ramjee, D. Towsley, and R. Nagarajan, "On Optimal Call Admission Control in Cellular Networks," *Wireless Networks*, vol. 3, no. 1, Mar. 1997, pp. 29–41.

[11] C. Leong *et al.*, "Call Admission Control for Integrated On/Off Voice and Best-Effort Data Services in Mobile Cellular Communications," *IEEE Trans. Commun.*, vol. 52, no. 5, May 2004, pp. 778–90.

[12] W. Song and W. Zhuang, "QoS Provisioning via Admission Control in Cellular/Wireless LAN Interworking," *Proc. 2nd Int'l. Conf. Broadband Networks*, Oct. 2005.

Biographies

WEI SONG (wsong@bbcr.uwaterloo.ca) received a B.S. degree in electrical engineering from Hebei University, Baoding, China, in 1998 and an M.S. degree in computer science from Beijing University of Posts and Telecommunications, China, in 2001. She is currently working toward a Ph.D. degree at the Department of Electrical and Computer Engineering, University of Waterloo, Canada. Her current research interests include resource allocation and QoS provisioning for the integrated cellular networks and wireless local area networks (LANs).

HAI JIANG [S'04] (hjiang@bbcr.uwaterloo.ca) received a B.S. degree in 1995 and an M.S. degree in 1998, both in electrical engineering, from Peking University, Beijing, China. He is currently working toward a Ph.D. degree at the Department of Electrical and Computer Engineering, University of Waterloo. His current research interests include QoS provisioning and resource management for multimedia communications in all-IP wireless networks.

WEIHUA ZHUANG [M'93, SM'01] (wzhuang@bbcr.uwaterloo.ca) received B.Sc. and M.Sc. degrees from Dalian Maritime University, China, and a Ph.D. degree from the University of New Brunswick, Canada, all in electrical engineering. Since October 1993 she has been with the Department of Electrical and Computer Engineering, University of Waterloo, where she is a full professor. She 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 Outstanding Performance Award in 2005 from the University of Waterloo and the Premier's Research Excellence Award (PREA) in 2001 from the Ontario Government for demonstrated excellence of scientific and academic contributions. She is an Associate Editor of *IEEE Transactions on Wireless Communications, IEEE Transactions on Vehicular Technology*, and *EURASIP Journal on Wireless Communications and Networking.*

XUEMIN (SHERMAN) SHEN [M'97, SM'02] (xshen@bbcr.uwaterloo.ca) received a B.Sc. (1982) degree from Dalian Maritime University, and M.Sc. (1987) and Ph.D. (1990) degrees from Rutgers University, New Jersey, all in electrical engineering. Currently, he is with the Department of Electrical and Computer Engineering, University of Waterloo, where he is a professor and the associate chair for graduate studies. His research focuses on mobility and resource management in interconnected wireless/wireline networks, UWB wireless communications systems, wireless security, and ad hoc and sensor networks. He is a co-author of two books, and has published more than 200 papers and book chapters in wireless communications and networks, control, and filtering. He was Technical Co-Chair for IEEE GLOBECOM '03, ISPAN '04, QShine '05, IEEE Broadnets '05, and WirelessCom '05, and is Special Track Chair of the 2005 IFIP Networking Conference. He serves as Associate Editor for *IEEE Transactions on Wireless Communications; IEEE Transactions on Vehicular Technology; Computer Networks; Dynamics of Continuous, Discrete and Impulsive — Series B: Applications and Algorithms; Wireless Communications and Mobile Computing* (Wiley); and *International Journal of Computer and Applications.* He has also served as Guest Editor for *IEEE JSAC, IEEE Wireless Communications*, and *IEEE Communications Magazine.* He received the Premier's Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada, for demonstrated excellence of scientific and academic contributions, and the Distinguished Performance Award in 2002 from the Faculty of Engineering, University of Waterloo, for outstanding contributions in teaching, scholarship, and service.