Toward Efficient Service-Level QoS Provisioning in Large-Scale 802.11-Based Networks

Tarik Taleb, Tohoku University
Abdelhamid Nafaa and Liam Murphy, University College Dublin
Kazuo Hashimoto, Nei Kato, and Yoshiaki Nemoto, Tohoku University

Abstract

Along with recent advances in mobile networking and portable computing technologies, there is a trend in the telecommunications industry toward the development of efficient ubiquitous systems that can provide a set of bandwidth-intensive and real-time services to multiple users while supporting their full mobility. Large-scale deployment of 802.11-based technologies will play an integral part in the construction of such ubiquitous wireless mobile systems. A challenging task in the development of such networks is efficient provisioning of QoS-enabled services for mobile users. In this context, we propose a scheme that constantly monitors the overall network performance to perform admission control and traffic conditioning at the 802.11-based access points and mobile terminals. The focus is on service-level fairness, where different flows from the same traffic class can still receive the same QoS level even if they have different bit rates. Furthermore, given the mobility of users, the success of any resource allocation and admission control model depends on the continuity of QoS guarantees across different WLANs. This article proposes a dynamic service level agreement negotiation protocol that allows mobile terminals to perform handoffs between different WLANs while maintaining the agreed level of service. End users also can change their service levels in response to changes in network conditions.

A transformation of network services is underway in the telecommunications world. Indeed, triple play has become the de facto services bundle for both telecommunications operators and service providers. The IPTV platform also is emerging as a framework encompassing all services (voice, TV, and Internet) in an interactive and integrated way. Another trend is the extension of the range of services through WiFi-enabled set-top boxes (STBs), enabling end users to access their triple-play services from laptops with full roaming capabilities. A likely scenario is one in which a service provider enables its subscribers to access its services across wireless LANs (WLANs) operated (or owned) by the service provider. Another likely scenario is when service providers establish service level agreements (SLAs) with third-party WLAN operators to extend the range of their services and ultimately, increase their average revenue per user (ARPU).

A serious challenge to this *service mobility* concept is the difficulty of providing efficient and uninterrupted guarantees of quality of service (QoS), as the services are essentially QoS-constrained and delivered to a large number of users roaming over differently loaded wireless networks. In 802.11-based networks, guaranteeing the same QoS metrics (e.g., loss rate, mean delay, mean jitter) for all flows belonging to the same traffic class (TC) is quite a challenging task. Despite this, maintaining a sustained application-level perceived QoS, regardless of the bit rate of each individual TC flow, is an imperative in most existing and forthcoming 802.11 networks (hotspots) [1, 2]. To achieve sustained QoS guarantees, we investigate the trade-off between achieving firm QoS performance and maximizing network utilization, in order to design an effective admission control (AC) mechanism.

As will be shown in this article, QoS enforcement can be addressed by medium access control (MAC)-level adaptation; however, service/user mobility support, that is, the way in which a service provider (SP) signals transiting service characteristics to visited WLANs through SLA agreements is a crucial issue. Current practice in the multimedia services distribution industry is based on SLA agreements, where both the SP and an end user agree on the terms of service provision in terms of QoS metrics. The SLA is composed of a legal part that states the terms of the service provisioning contract between a user and an SP and a technical part, the service level specification (SLS), that describes parameters such as: service characterization traffic conditioning policies, traffic class, expected performance of QoS metrics, and so on [3]. These SLA/SLS agreements should be available at all visited WLANs to:

- Enable the service after a user-to-service authentication process.
- Provide seamless QoS continuity by enforcing the appropriate QoS policies as specified in the SLS.

0890-8044/07/$20.00 © 2007 IEEE

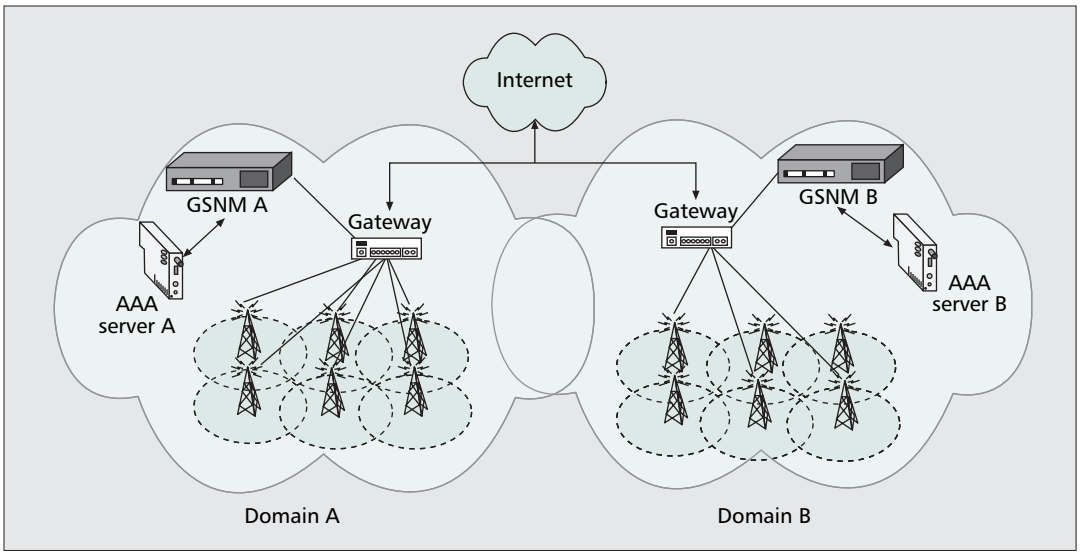


Figure 1. *Architecture of the proposed dynamic service negotiation protocol with its key components.*

Clearly, the scalability of the protocols used to dynamically signal SLA/SLS formats between the SP (owner of the transiting service) and the visited networks (WLAN operators) is critical.

In short, dynamic SLA negotiation is essential for efficient QoS provisioning in wireless mobile environments. A critical component is the mechanism to deliver QoS in a network comprised of many WLANs. We review existing service-level negotiation protocols and investigate a scalable way to dynamically signal SLSs across WLANs visited by mobile users. Particular attention is paid to the overhead entailed by the per-session SLS negotiation procedure. Another important component addressed here is how QoS metrics performance thresholds conveyed in an SLA/SLS are effectively enforced at each WLAN. We emphasize QoS provisioning at the MAC layer, since this layer is crucial for resource allocation in contention-based mediums such as 802.11-based networks. Efficient resource control and allocation in WLANs is also important to provide the network operator (NO) with better means to leverage and broker its underlying resources to SPs.

The article is organized as follows. We present a brief description of the envisioned network architecture and the functions of its key components. We briefly discuss the current service level negotiation mechanisms and propose an approach that best suits the features of wireless mobile networks. We discuss issues related to the QoS provisioning in 802.11-based networks. The article also proposes a QoS enforcement strategy based on QoS metric-driven, MAC-level adaptation and AC. Finally, we draw some important conclusions.

Architecture Description

This section outlines the key components of the envisioned network architecture. At the highest level, the SP plays a central role in the digital multimedia services distribution chain. An SP contracts an SLA with its subscriber to define the terms and conditions of the service delivery process. The SP is also responsible for negotiating resources with NOs to match the QoS requirements of provisioned services; for scalability reasons, this resource negotiation/commitment is usually performed at an aggregate level. Obviously, although the SP is in charge of service provisioning and activation, a tight cooperation with the NO is required to:

- Better leverage the underlying network resources.
- Be more responsive in accommodating new situations such as handoffs, network load changes, and so on.

In this regard, the NO is expected to develop effective models and interfaces to better control its underlying resources (e.g., using an AC mechanism) and ultimately, better broker these resources to third-party SPs.

The components of our targeted network architecture are depicted in Fig. 1, which portrays the coverage area of a number of WLANs forming different domains administrated by different NOs. Each domain is administered by a global service negotiation manager (GSNM) and an authentication authorization accounting (AAA) server. The latter is used to verify whether mobile users are authorized to access requested services; whereas, the former carries out the whole service negotiation procedure. Upon receiving a service initiation/renegotiation request from a mobile user, the GSNM uses information about outstanding requests and resource availability to accept or reject requests; again, this requires interaction with a distributed resource allocation mechanism deployed at each WLAN. If the request cannot be satisfied without degrading the QoS of already existing users, the GSNM may deny the request and send an immediate negative acknowledgment. Alternatively, a list of available (downgraded) service levels can be sent with a negative acknowledgment, to renegotiate a service level with the subscriber.

Generally speaking, QoS provisioning consists of two major operations: resource allocation and service-level negotiation/management. The mechanisms by which GSNMs deny requests or allocate resources for a user rely on QoS metric-driven MAC-level adaptation and resource allocation. QoS metrics performance bounds, as specified in the original SLS, are used to assess whether the current WLAN can support a new service request. Here, an SLS negotiation protocol is used to enable a-priori QoS commitment negotiation between the SP and the NO. The SLS negotiation protocol runs in parallel with other out-of-bound multimedia session signaling/control protocols, such as real-time streaming protocol (RTSP) and Session Initiation Protocol (SIP), which also are required.

Dynamic QoS Negotiation for Next-Generation Wireless Mobile Networks

The SLS-based QoS negotiation and commitment procedure (usually called QoS peering) proposed in our approach is slightly different from the one that is usually established between an SP and an NO [1]. The main difference is in the time-scale of this QoS peering process. Although resources are usually committed for longer terms (measured in months) and at aggregate

levels in large-scale backbone networks, the resources in our case (large-scale WLANs) are negotiated and committed on a per-session basis for the lifetime of the multimedia session or for the roaming time. We still assume that resources are committed for longer terms in the backbone network.

Current Dynamic SLS Negotiation Approaches

Several recent publications have examined new ways to enable dynamic service level negotiation between clients and SPs. A thorough discussion of the advantages and pitfalls of each can be found in [2]. Traditional examples include the Service Negotiation Protocol (SrNP) and the Common Open Policy service — service level specification (COPS-SLS) [3]. Originally designed for wired networks, a major issue with these protocols is that they require periodic signaling among network entities and terminals. When applied to wireless mobile networks, this signaling wastes a significant portion of the network bandwidth, resulting in poor scalability for large-scale deployment scenarios. To cope with this issue, Dynamic Service Negotiation Protocol (DSNP) [4] takes into account the constraints of the wireless environment on bandwidth and power and reduces the frequency of signaling messages. The basic concept behind DSNP operation is an immediate dissemination of the QoS profile of each user negotiating for a given service not only to the access point (AP) of the sub-domain where the user is currently located, but also to all the APs of adjacent sub-domains. In this way, when this user enters an adjacent sub-domain, it is immediately served at the same QoS level with no requirement for additional signaling from the user. DSNP is tailored to negotiate SLS over IP networks that ensure proper QoS metrics performance negotiation for end users and to improve the control of network operators over their underlying network resources. On the other hand, conventional streaming control protocols, such as RTSP and SIP are used to initiate and control multimedia sessions and operate throughout the lifetime of a multimedia session.

Although DSNP is a lightweight protocol and is seen as the most suitable scheme for dynamic mobile environments, its application to mobile networks gives rise to a number of issues. For example, whenever a user subscribes for a service level, the GSNM should advise its current AP and the neighboring APs of the new service requirements of this subscriber. This implies that APs should maintain a significantly large state table of user profiles. In addition, given that in DSNP there is no mechanism to inform APs when to erase profiles of departing users, APs must permanently maintain profile information on all users. This raises doubts about the scalability of DSNP when applied to highly dynamic mobile networks.

Proposed Dynamic Service Negotiation Scheme

The proposed dynamic QoS negotiation scheme supports service initiation, service negotiation, and service renegotiation upon handoff. WLANs are assumed to be equipped with resource allocation interfaces that allow the SLS negotiation protocol to check for resource availability in user-visited WLANs.

Service Initiation or Renegotiation — At initial connection set up, a subscriber searches for an AP through which it can communicate and then sends a service initiation request to the detected AP. The AP then forwards the message to its corresponding GSNM. Upon consulting its corresponding AAA server and the WLAN concerned, the GSNM makes the decision on whether to admit or reject the request. If the requested service level can be guaranteed, the GSNM informs the subscriber of the successful registration of its required SLS. This acknowledgment is sent in a negotiation response message via the same AP. If the requested SLS cannot be provided, a negative acknowledgment is sent to the subscriber with a list of available service levels. The subscriber then notifies the GSNM of its desired SLS via an SLS negotiation request message. A successful negotiation of SLS is acknowledged to the subscriber via a negotiation response message. After the service initiation procedure is finished, the subscriber starts exchanging actual data traffic.

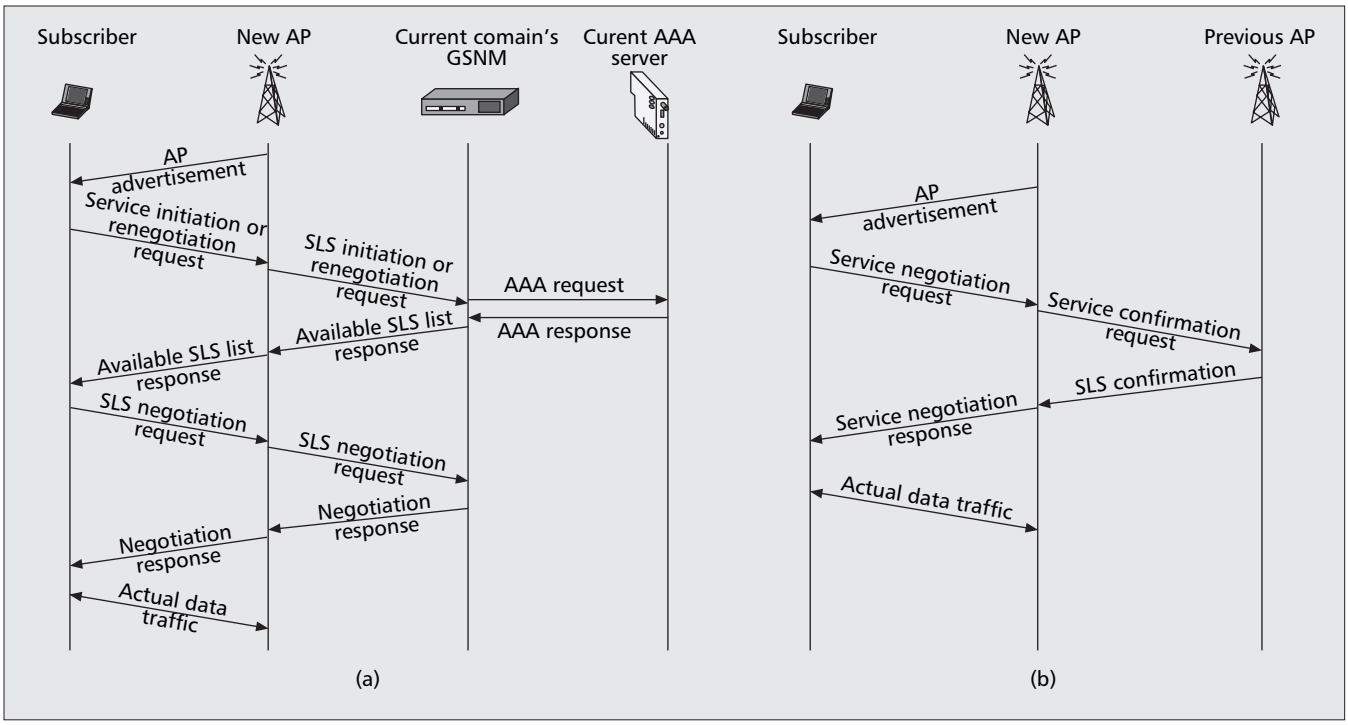
While using the service, a subscriber may wish to renegotiate a different service level using the same procedure. Service renegotiation also can be triggered by the GSNM in a proactive way. Indeed, in the case of service performance degradation, the GSNM may offer privileges to subscribers that agree to downgrade their current service levels. Note that the flexibility of a subscriber in terms of QoS may be explicitly mentioned in the SLS format. Similarly, if sufficient network resources become available, the GSNM can encourage subscribers to switch to a higher service level to receive better QoS. The major operations of these service initiation and renegotiation mechanisms are depicted in Fig. 2a.

Service Negotiation Upon Handoff — When a mobile node performs a handoff to a new AP within the same domain (i.e., intra-domain handoff), the mobile terminal issues a service negotiation request to the new AP. This AP verifies whether the requested service should be guaranteed by consulting the previous AP via an SLS confirmation request message. After the requested SLS is confirmed, the AP informs the mobile terminal of the negotiation results via a service negotiation response. The previous AP then erases the profile of the mobile user from its database, as this mechanism informs APs when to erase profiles of departing users. APs are thus no longer forced to permanently store user profiles, which improves the overall scalability of the system by allowing NOs to further broker their underlying resources to different SPs. Figure 2b portrays the major operations of service renegotiation upon an intra-domain handoff. Although consulting previously used APs about the SLSs of newly joining users adds a two-hop delay to the service renegotiation latency, this delay can be minimized by adding an encrypted key to SLSs. Indeed, upon a handoff, a user can send its SLS to the new AP with a key. The new AP uses this key to authenticate the user and the requested SLS. This is similar to the concept of the distributed system proposed in [3]. In this manner, if resources are available, the mobile user can immediately start accessing the requested service with no requirement of confirmation from the previous AP.

To further minimize the service negotiation delay, mobile users can initiate the service negotiation procedure as soon as they enter the overlapping area between the coverage areas of the two APs involved.

Performance Evaluation

Having described the details of the proposed dynamic service negotiation scheme, we now demonstrate via simulation how the scheme addresses all the aforementioned issues of DSNP. We conducted a simple simulation using the NS2 network simulator. We considered a network domain consisting of five APs, managed by the same GSNM and the same AAA server. The delay between the wireless gateway and APs was set to 1.5 ms. The number of mobile nodes roaming in the coverage areas of the five APs varied between five and 100. All mobile nodes were assumed to already have initiated their services. Our focus was on service negotiation upon intra-domain handoff, as this is the most frequent handoff performed by mobile terminals. The mobility of each mobile terminal was randomly set. For comparison purposes, we used the DSNP [2] scheme and the encrypted SLS scheme [3] (for further details see [5]).



■ Figure 2. *Major operations of the proposed dynamic service negotiation mechanism: a) service initiation or renegotiation; b) service negotiation upon intradomain handoff.*

Although the QoS negotiation delay depends highly on the underlying network topology, all simulation results indicate that our proposed scheme exhibits the highest negotiation delay. Indeed, compared with DSNP, our proposed scheme incurred an average additional negotiation delay of 6 ms for 5 MSs and 6.6 ms for 100 MSs, respectively. This result was expected, since the proposed scheme must confirm with the previous AP before granting a mobile terminal its requested service level. An alternative solution in wireless environments would be to make use of the mobility patterns of users to predict the next AP to which users will attach after handoff. Then, users could anticipate their service negotiation with the next point of attachment while they are still connected to their former AP.

Although the relatively longer negotiation delay is a drawback of the proposed scheme, its major advantages are its reduced signaling overhead and its shorter SLS storage table, as can be deduced from Fig. 3a and Fig. 3b. These figures show that DSNP requires a high number of signaling messages and causes replication of the users' profiles at all APs. This is inefficient as the storage capacity in APs will be limited and is unrealistic with respect to current commercial APs. In contrast, our proposed scheme tends to be more scalable as it reduces the number of signaling messages and does not require significant storage tables. The reduced signaling overhead of the proposed scheme is attributable to the fact that it locally negotiates service requirements of users, without involving GSNMs upon a handoff occurrence; whereas, its shorter SLS storage table is due to its ability to inform APs when to erase the profiles of departing users.

A network operator is more concerned about the scalability of SLS signaling/negotiation generated on a per-session/per-user basis, although storage is important, too. The negotiation delay is certainly a critical issue; however, this may be addressed using a network operator-assisted handover technique, where the SLS negotiation procedure is initiated together with the handover signaling handshake.

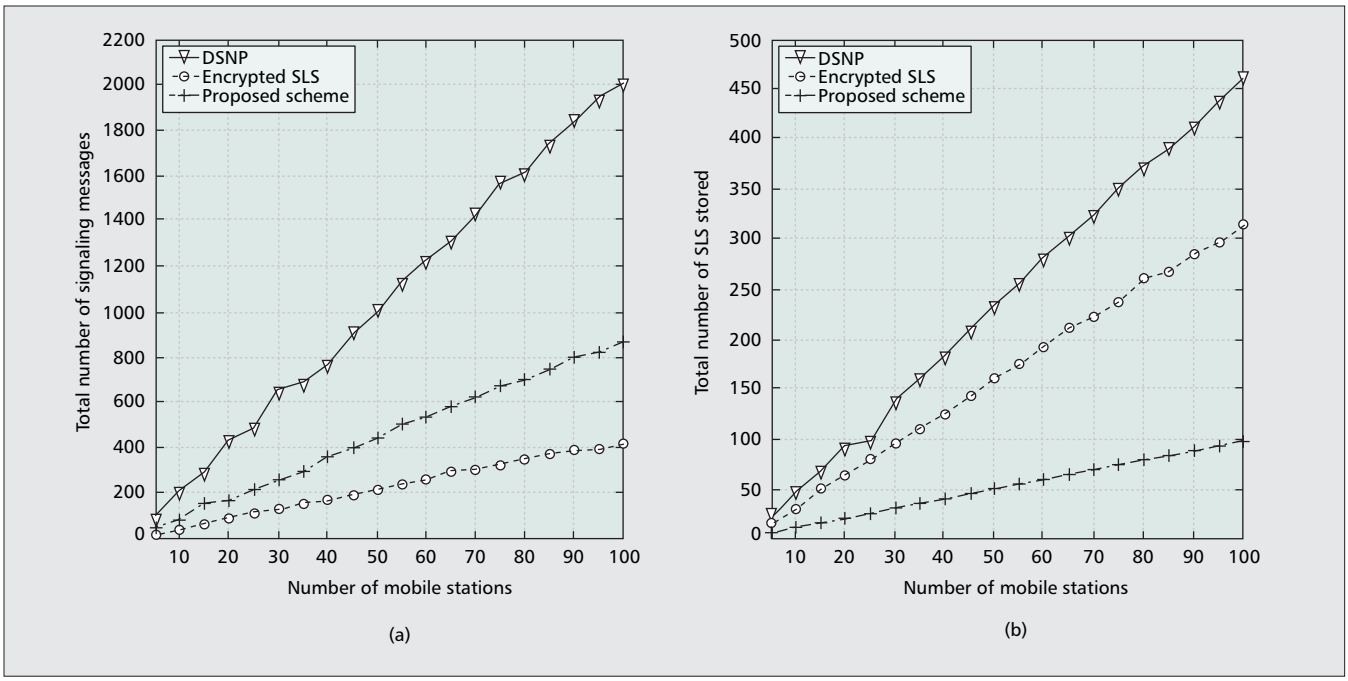
Sustained QoS Provisioning in 802.11-based Networks

In contention-based networks (e.g., 802.11), IP-level service differentiation is not sufficient to efficiently manage the shared resources, and additional QoS provisioning at the MAC layer is essential in this respect. For example, in IP-layer differentiated services (DiffServ), all IP traffic classes would be aggregated in a single MAC queue and consequently would receive the same service level at a MAC layer. Real-time flows would experience a loss of priority during transmission in WLANs. In addition, service differentiation at the IP layer basically would provide scheduling between different traffic classes by determining the service class to serve at each time instant. But this service differentiation has only local significance; in certain circumstances, a high-priority flow from a given station (A) may obtain the same network access opportunities as a best-effort flow from another station (B). This is particularly prevalent if station A is scheduling several traffic classes, while station B is currently handling only best-effort traffic.

The MAC layer is essential for resource allocation in 802.11-based networks. Resource allocation in 802.11-based networks is actually determined by the way different traffic classes handle contention situations; that is, the way different flows determine the appropriate spacing between consecutive transmission attempts in different network conditions [6].

Service Differentiation in 802.11-Based Networks

WLANs should enable a variable number of users with heterogeneous QoS requirements to share a common radio channel. By extending the MAC protocol of a WLAN to provide several traffic classes (TCs), the NO could fragment its QoS offer into several levels of guarantees, so as to accommodate any network configuration in terms of per-class network load. In this context, the network resources can be fully utilized by flows from a single TC, or the load can be distributed among the supported TCs according to the instantaneous per-class offered load; which ultimately gives more flexibility to an NO



■ Figure 3. *Performance evaluation in terms of signaling overhead and scalability: a) service negotiation signaling; b) data stored along the network.*

in brokering its underlying resources. Depending on the nature of content coding and the targeted applications (video conferencing, IP telephony, media streaming, etc.), the multimedia streams may be mapped into different TCs characterized by different guaranteed QoS metric thresholds.

The IEEE 802.11e task group has focused on finding better QoS mechanisms to support multimedia streams. The 802.11e standard includes enhanced distributed-channel access (EDCA), where each node maintains a backoff instance for each traffic class. During initialization, EDCA assigns static QoS parameters for each TC. Based on these parameters, the MAC protocol provides different service levels to each TC.

In the final 802.11e standard amendment [7], the contention window, CW[i]min/ CW[i]max values, which determine the backoff ranges (0, CW[i]max) of high-priority TCs has been narrowed in previous versions of 802.11e drafts and is still unchanged (e.g., TC3 has a CW range of [0, 15], and TC2 has a range of [0, 31]). This favors service differentiation in relaxed network conditions, while severely limiting the achievable throughput for a few backlogged high priority flows due to an excessive intra-TC collision rate when the number of flows increases [6]. The class-specific MAC parameters of EDCA (CW[i]min and CW[i]max) do not suit many network configurations and are only effective for a reduced number of multimedia flows. Independently of the network offered load, certain traffic classes (TCs) may experience degraded performance, while other classes experience low contention. This situation considerably reduces the flexibility of network operators in maximizing the profitability of their underlying resources.

Resource Allocation in 802.11-Based Networks

In a WLAN, it is crucial to restrict the volume of traffic in order to maintain the service quality of currently served traffic. If there are no limitations on the volume of traffic that is introduced to the network, performance degradation results because of higher backoff times. Effective resource allocation in IEEE 802.11 is difficult to achieve due to the intrinsic nature of its carrier sense multiple-access with collision-avoidance (CSMA/CA) scheme. Unlike point-coordinated wireless networks where bandwidth provision can be managed using only bandwidth availability information, resource allocation in 802.11 networks requires additional parameters and more advanced models. For the same overall offered load, the network may exhibit different performance (i.e., availability levels) depending on the number of competing flows and their respective bit rates. For instance, the network contention level (collisions) experienced by ten 100-kb/s active flows would be different from that experienced by two 500-kb/s active flows.

Even though some adaptive CW schemes focus on coordination between MAC parameters of different flows belonging to the same TC [6], they still provide *access opportunities fairness* rather than *service-level fairness*. In fact, if traffic were balanced between nodes, achieving fairness between flows from the same traffic class would require the MAC parameters on different nodes to be harmonized. However, traffic load is typically unbalanced in real WLAN deployments, with a large variation in TC volume from one node to the next.

Based on local network measurements, Zai et al. [8] suggest controlling the arrival rate at each station to satisfy a given threshold in the network, such as minimum throughput or maximum delay/jitter/loss rate. Their analytical model is designed to control acceptance of network flows based on a new metric (channel busyness ratio) and a rate control algorithm — call admission and rate control (CARC). In addition to its inapplicability to 802.11e-like protocols, where several traffic classes may simultaneously operate in the network, CARC tries to find the optimal network utilization (to maximize the throughput) but barely considers delay fluctuations.

Distributed admission control (DAC) and two-level protection and guarantee mechanisms [9] are combined to efficiently address the previously mentioned issues. DAC is a measurement-based mechanism that was considered by the 802.11e working group. However, the major problem with DAC-based approaches is that the overall network bandwidth is statically allocated to different TCs, so that each TC receives a fixed share of bandwidth that cannot be exceeded. Thus, streams from a given TC may be rejected while some bandwidth is underutilized in other TCs. This may severely affect the flexibility of the network operator as it is quite difficult to forecast the per-TC traffic volume in realistic WLAN deployments.

One characteristic of contention-based networks is the existence of a critical trade-off between the achieved network throughput and the delay guarantees [10]. It is not possible to fully utilize the network capacity while still satisfying strict delay requirements of different service classes. Generally speaking, increasing the throughput of a flow beyond a certain extent means increasing the enqueueing delays and thus, probably violating delay constraints. A candidate AC mechanism should enable all possible per-class load distributions as long as the QoS constraints are not violated.

QoS Metrics-Driven MAC-Level Adaptation and Admission Control

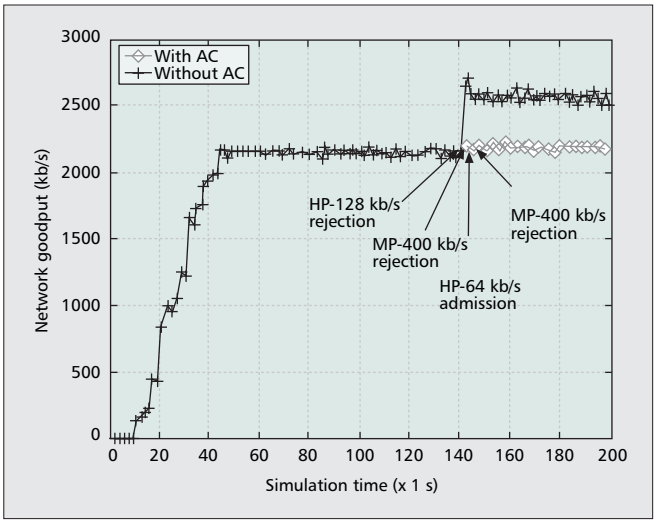
Although WLANs may be augmented with prioritization mechanisms, they still require an accurate model of network/service monitoring to effectively perform AC and traffic conditioning at the network access. In [6], a simple empirical adaptation model was proposed that reacts to increased loss rate and accordingly adjusts the CW size to keep the loss rate of each traffic class within a predefined maximum threshold. However, results in [8] clearly showed that delays are, to a certain extent, uncorrelated with loss rate. Although the loss rate may be controlled, the total delay is an independent process that depends on the network load and the mean access delay.

Instead of adapting the CW size based only on collision events or network conditions, we propose to react to the achieved QoS metrics performance that provides service-level fairness regardless of flow bit rates [11]. Our model monitors several important parameters such as the contention level (collision ratio) per traffic class, overall network load, MAC queue length, and so on. Our basic idea is to dynamically adjust the CW size of each individual TC flow based on an accurate assessment of the overall MAC-level enqueueing delay that depends on:

- The MAC queue length (number of packets).
- The packet service time (PST), that is, the elapsed time interval between when a packet arrives at the front of the queue and when it is received by the receiver.

The PST considers the backoff interval (including freezing periods), transmission delay (with eventual retransmissions), and associated physical overhead. The CW size is readjusted on a regular basis to keep the MAC-level enqueueing delay within a predefined per-TC maximum delay threshold. In our study, we arbitrarily fixed the maximum delay to 0.5 second for high priority (HP) flows and 0.8 second for medium priority (MP) flows.

We developed a delay estimation mechanism based on the instantaneous queue length (packet inter-arrival rate) assessment. Therefore, the achievable throughput with the potential degradations (mean loss rate) may be predictable as well. Using the packet arrival rate for a given TC flow that is a-priori known from earlier contracted SLSs, it is possible to capture the queue dynamics based on instantaneous network activities (e.g., using the M/M/1/K model). The objective is to predict the impact of the new stream's admission on overall network performance. We assess the consequences of increasing the packet arrival rate (offered load) of a given TC/station before actually admitting any new flow. The packet arrival rate augmentation corresponds to the offered load of the new stream requesting to be admitted. By extending our delay prediction model using an M/M/1/K model, it is possible to integrate the loss rate constraint with the maximum tolerated delay constraint. Therefore, upon a new stream's arrival, we can a-priori assess the impact of increasing the packet arrival rate under delay and loss constraints and ultimately predict the degradation of all active flows in the network. This fully distributed model would give a network operator (or SP) the means to accept or reject a new stream.



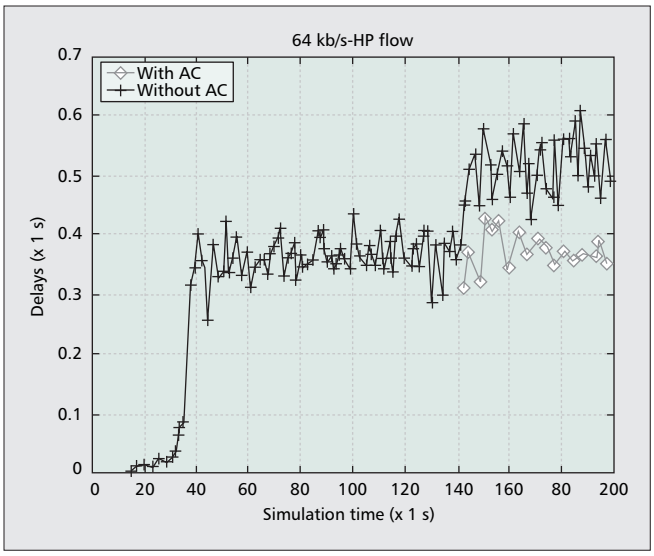
■ Figure 4. *Admission control: overall network utilization.*

We performed an experiment where we progressively increased the overall network load by continuously backlogging new flows; we used an 802.11b network with 11 Mb/s of network capacity and 16 wireless terminals. Each of these terminals can generate up to two flows, representing two uniquely prioritized traffic classes: high priority (HP) and medium priority (MP). From t = 10 s, a new flow from a different priority (HP, MP) is started until network saturation at t = 140 s. The objective of such network dimensioning is to assess if our adaptation model can react to frequent changes in relative (per-class) network load while still meeting the QoS requirement. Note that we use two different bit rates for each TC to evaluate the ability of our MAC-level adaptation mechanism to achieve service-level fairness among flows from the same TC regardless of the offered load of each single flow, that is, intra-TC QoS enforcement.

In Fig. 4 and Fig. 5, we evaluate the performance of our AC scheme by comparing it to a scheme that uses only delay-based MAC-level adaptation. We refer to these two operation modes as with-AC and without-AC, respectively. The same network dimensioning is used as described previously.

As can be seen in Fig. 4, when the network is sufficiently relaxed (before t = 140 s), there is sufficient bandwidth available, and both the with-AC and the without-AC schemes, achieve similar throughputs, carrying the offered load. However, under stressed conditions, the scheme without AC gains a significant advantage over our scheme (with AC). The goodput gain reaches about 20 percent when the load is around 2.4 Mb/s (between t = 140 s and t = 200 s). At this point, the AC mechanism rejects three entering flows: a 128-kb/s-HP flow at t = 140 s, a 400-kb/s-MP flow at t = 141 s, and finally, a 400-kb/s-MP flow at t = 143 s; while at the same time, a 64-kb/s-HP flow was accepted at t = 142 s based on QoS degradation predictions.

However, the bandwidth gain in the without-AC scheme comes with a serious degradation of the QoS of all active multimedia flows, as revealed by delay measurements of the single HP flows. As can be seen from Fig. 5, HP flows experience high delays from t = 140 s, when the first flow is accepted in the network. Their performance is further degraded with the admission of other flows. Here, all flows try to further reduce the sizes of their respective CWs to cope with increasing PST, but result in an increase in the contention level and yield a counterproductive situation.



■ Figure 5. *End-to-end delays for 64 kb/s HP flows.*

Conclusions

In this article, we present a QoS architecture for real-time service provisioning in large-scale 802.11-based networks. This architecture aims to ensure QoS continuity for mobile users as they roam across different WLANs. We address the issue of end user mobility by proposing a protocol for dynamic service-level agreement negotiation that dynamically signals QoS characteristics of each individual end user's service during handoffs. Our proposed protocol considerably reduces the signaling overhead and improves system scalability, which allows network operators to support more SLAs and thus better leverage their underlying resources. We also address the issue of resource allocation and admission control within a single WLAN by focusing on providing service-level fairness among flows belonging to the same traffic class. By considering key network performance metrics and application-level QoS requirements, we derive an accurate delay estimation model to adjust the contention window of network flows, and propose an effective admission control model to protect already-active flows.

As briefly discussed, latency associated with SLA negotiation during handoffs is an important issue to manage when provisioning real-time services. This delay could be considerably reduced by anticipating end user handoffs. Therefore, our future work will focus on the integration of SLA negotiation with a context-aware network-assisted handoff management strategy.

References

[1] EU-IST Integrated Project, "End-to-End QoS through Integrated Management of Content, Networks and Terminals," IST-2003-507637/ENTHRONE, 6th EU Framework Program, 2003-2007.
[2] P. Trimintzios *et al.*, "Service-Driven Traffic Engineering for Intradomain Quality of Service Management," *IEEE Network*, vol. 17, no. 3, May 2003.
[3] T. Nguyen *et al.*, "COPS-SLS: A Service Level Negotiation Protocol for the Internet," *IEEE Commun. Mag.*, May 2002, pp. 158–65.
[4] J.C. Chen *et al.*, "Design and Implementation of Dynamic Service Negotiation Protocol (DSNP)," *Elsevier J. Comp. Commun.*, June 2006.
[5] T. Taleb *et al.*, "A Dynamic Service Level Negotiation Mechanism for QoS Provisioning in NGEO Satellite Networks," *Proc. IEEE ICC*, Glasgow, Scotland, June 2007.
[6] A. Nafaa *et al.*, "Sliding Contention Window (SCW): Towards Backoff Range-based Service Differentiation over IEEE 802.11 Wireless LAN Networks," *IEEE Network*, vol. 19, issue 4, July/Aug. 2005.
[7] IEEE 802.11e, "Wireless LAN Medium Access Control (MAC) Enhancements for Quality of Service (QoS)," July 2005.
[8] H. Zhai *et al.*, "How Well Can the IEEE 802.11 Wireless LAN Support Quality of Service?," *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, Nov. 2005, pp. 3084–94.
[9] Y. Xiao *et al.*, "Protection and Guarantee for Voice and Video Traffic in IEEE 802.11e wireless LANs," *Proc. IEEE INFOCOM '04*, vol. 3, Hong Kong, Mar. 2004, pp. 2152–62.
[10] M. Barry *et al.*, "Distributed Control Algorithms for Service Differentiation in Wireless Packet Networks," *Proc. IEEE INFOCOM '01*, Anchorage, AK, Apr. 22–26, 2001.
[11] A. Nafaa, "Provisioning of Multimedia Services in IEEE 802.11 Networks: Facts and Challenges," *IEEE Wireless Commun.*, Sept. 2006.

Biographies

TARIK TALEB (taleb@aiet.ecei.tohoku.ac.jp) received his B.E. degree with distinction in information engineering, and his M.E. and Ph.D. degrees in computer science from Tohoku University in 2001, 2003, and 2005, respectively. He is currently working as an assistant professor at Tohoku University, Japan. From October 2005 to March 2006, he worked as a research fellow with the Intelligent Cosmos Research Institute, Sendai, Japan. His research interests are in the field of wireless networking, intervehicular communications, satellite and space communications, congestion control protocols, network management, handoff and mobility management, and network security. His recent research has also focused on on-demand media transmission in multicast environments. He is on the editorial board of *IEEE Wireless Communications*. He also serves as secretary of the Satellite and Space Communications Technical Committee of ComSoc (2006–present). He is a recipient of the 2007 Funai Foundation Science Promotion Award, the 2006 IEEE Computer Society Japan Chapter Young Author Award, the Niwa Yasujirou Memorial Award (February 2005), and the Young Researcher's Encouragement Award from the Japan chapter of the IEEE Vehicular Technology Society (VTS) (October 2003).

ABDELHAMID NAFAA obtained his Master's and Ph.D. degrees in 2001 and 2005, respectively, from the University of Versailles-SQY, France, where he was involved in several national and European projects: NMS, IST-ENTHRONE1, IST-ATHENA, and IST-IMOSAN. He is a Marie Curie Research Fellow under the EU-FP6 EIF Marie Curie action that seeks broader synergy in the European research space. He was granted the Marie Curie award to undertake research work at University College Dublin (UCD) in the area of multimedia services distribution over carrier-grade networks. Before joining UCD, he was an assistant professor at the University of Versailles-SQY and acted as a technology consultant for U.S. and European-based companies in the area of reliable multimedia communication over WiFi technology and IMS-based multicasting in DVB-S2 satellite networks, respectively. Recently, he was involved in a government-funded project (i.e., VidAs) that aims to develop a P2P-based VOD services distribution for network operators. He is a co-author of over 25 technical journal or international conference papers on multimedia communications.

LIAM MURPHY [M] received a B.E. in electrical engineering from University College Dublin (UCD) in 1985, and an M.Sc. and Ph.D. in electrical engineering and computer science from the University of California, Berkeley in 1988 and 1992 respectively. He is currently a senior lecturer in computer science at UCD, where he is director of the Performance Engineering Laboratory (http://www.perf-englab.com). He has published over 100 refereed journal and conference papers on various topics, including multimedia transmissions, dynamic and adaptive resource allocation algorithms, and software development. His current research projects involve mobile and wireless systems, computer network convergence issues, and Web services performance issues. He is a director of crovan (http://www.crovan.com), a UCD/Dublin City University (DCU) campus company spun out of Enterprise Ireland funded research.

KAZUO HASHIMOTO received his M.S. degree in computer science from Brown University and his Ph.D. degree in information science from Tohoku University, Sendai, Japan, in 1986 and 2001, respectively. Currently, he is a professor in the Graduate School of Information Sciences, Tohoku University. From 2001 to 2005, he managed KDDI Labs USA as president and CEO. During this period he directed all R&D activities in collaboration with major research institutions and universities in the United States. His current research interests are in the field of network security, network management, data mining, and multimedia information retrieval.

NEI KATO [SM] received his M.S. and Ph.D. degrees from the Graduate School of Information Sciences, Tohoku University, in 1988 and 1991, respectively. Since 1991 he has been working for Tohoku University and is currently a full professor at the Graduate School of Information Sciences. His research interests are computer networking, wireless mobile communications, image processing, and neural networks. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan. He has served on several technical program and organizing committees of international conferences. Since 2006 he has served as a technical editor of *IEEE Wireless Communications*.

YOSHIAKI NEMOTO [SM] received his B.E., M.E., and Ph.D. degrees from Tohoku University in 1968, 1970, and 1973, respectively. He is a full professor in the Graduate School of Information Sciences and served as director of the Information Synergy Center, Tohoku University. His research interests are microwave networks, communication systems, computer network systems, image processing, and handwritten character recognition. He is a recipient of the 2005 Distinguished Contributions to Satellite Communications award from IEEE ComSoc and a co-recipient of the 1982 Microwave Prize from the IEEE MTT Society. He is a member of IEICE and a fellow member of the Information Processing Society of Japan.