An Integrated Load Balancing Scheme for Future Wireless Networks

Eng Hwee Ong, Jamil Y. Khan
School of Electrical Engineering and Computer Science
University of Newcastle, Australia, NSW 2308
Email: enghwee.ong@studentmail.newcastle.edu.au, jamil.khan@newcastle.edu.au

***Abstract*—With the emerging IEEE 802.11n standard, the WLAN is poised as a promising ubiquitous networking technology to support multimedia applications where providing QoS becomes imperative. However, the 802.11 WLAN is not designed to support delay sensitive traffic. This problem is magnified during a handover and typically results in excessive handover latency and packet loss. In addition, a 802.11 WLAN handover process is predominantly based on the physical layer detection without QoS considerations. This often causes overloading of access points and consequently all its associated connections would suffer from high delay. The former can be resolved by reducing handover latency to achieve seamless handover and the latter can be mitigated by employing link layer detection in the 802.11 WLAN handover process and having an appropriate admission control scheme. Although the IEEE 802.11e standard supports prioritized QoS, it cannot guarantee strict QoS required by real-time services under heavy load. In this paper, we proposed an integrated load balancing scheme incorporating (i) QoS-based fast handover to support seamless handover by eliminating both detection and scanning phases from the 802.11 WLAN handover process; and (ii) soft admission control to protect QoS of existing connections when resources are low. This synergy allows us to perform QoS-related handover opportunistically and guarantees service QoS during *and* after handover respectively. Simulations showed that our proposed integrated load balancing scheme is capable of providing seamless handover and QoS provisioning for real-time VoIP services in terms of bounded delay and packet loss when considering multimedia traffic. Particularly, our proposed scheme exhibits both throughput and QoS fairness which jointly optimize overall system utilization.**

I. Introduction

Seamless mobility and multimedia communications are the driving forces of future wireless networks. The IEEE 802.11 WLAN permits connectivity of $11Mbps$ using the 802.11b and high speed data rate of $54Mbps$ using the 802.11a/g necessary to support multimedia services at low cost. The advent of 802.11n will further increase connection speed up to $600Mbps$. The increasing popularity of multimedia applications such as VoIP, video streaming and data have made their unification over WLAN compelling since they can now leverage on pervasive 802.11 networks of high bandwidth for user mobility. However, supporting real-time VoIP and video services over WLAN poses numerous challenges such as admission control and QoS provisioning.

Zhai *et al.* [1] found that WLAN attains maximum throughput and low delay when operating in unsaturated mode due to low collision probability, suggesting that admission control is a suitable strategy for real-time traffic due to its low bandwidth but strict delay requirements. One of the main challenges in QoS provisioning for WLAN is to support real-time VoIP connection with seamless handover since dynamic network conditions may result in unacceptably high packet delay and consequently packet loss. VoIP requires one-way end-to-end delay of less than 150ms [2] but can tolerate some packet loss rate of up to 2% [3]. This implies that the total handover latency and packet loss should not exceed these bounds in order to sustain an undisruptive VoIP call of acceptable QoS. We focus on minimizing Layer 2 handover latency which composes of detection delay, scanning delay, authentication delay and reassociation delay, where detection and scanning delays are the dominating cost [4], [5].

In this paper, we proposed an integrated load balancing scheme featuring (i) QoS-based fast handover to support seamless handover; and (ii) soft admission control to protect QoS of existing connections, both by estimating critical network QoS parameters [6]. The basic idea is to protect QoS of real-time services from network overloading by redistributing network load opportunistically through QoS-related handovers. To the best of our knowledge, there is no prior research on load balancing scheme that guarantees service QoS and jointly optimizes system utilization by considering QoS-based fast handover in conjunction with soft admission control. The remainder of the paper is organized as follows. Section II discusses related work. Section III describes our proposed integrated load balancing scheme and architecture. Section IV illustrates the simulation model. Section V presents the simulation results and Section VI concludes the paper.

II. Related Work

Various load metrics such as number of active connections, gross load, packet loss and throughput have been proposed in literature and it is one of the key elements in any load balancing scheme. For circuit-switched cellular networks such as GSM, load balancing is traditionally based on number of active calls per cell as its load metric since the load contributed by each user is the same. However, Bianchi *et al.* [7] showed that load balancing in packet-switched wireless networks such as WLAN can be improved by using additional "packet level" load metrics such as gross load which considers number of stations together with retransmission probability and packet loss. Bazzi *et al.* [8] developed a measurement-based call admission control to protect QoS of existing connections by

978-1-4244-3062-8/08/$25.00 ©2008 IEEE

denying incoming calls when resources are low. However, the parameters of their call admission control require tuning for different traffic mixes, hence not adaptive to dynamic network conditions.

Balachandran *et al.* [9] presented an adaptive load balancing solution where a centralized admission control server contains load information of all access points. However, this approach requires additional central server which increases network signaling overheads, creates bottleneck and prone to single point of failure. Velayos *et al.* [10] proposed a decentralized load balancing scheme using throughput per access point as their load metric. However, the major pitfall of this scheme is that station will experience service outage during a handover since station first disassociate from an old access point and can only reassociate with an underloaded access point after some searching time has elapsed.

Our contributions differ from related works in three significant ways (i) we guarantee service QoS during handover by enabling seamless handover with QoS-based fast handover *and* guarantee service QoS after handover by operating network in unsaturated mode with a soft admission control; (ii) we exploit estimated critical network QoS parameters as criterion to select the best target network for handover and as load metric for soft admission control, both of which are adaptable to varying network conditions; and (iii) our distributed handover architecture provides network-assisted discovery compatible with IEEE 802.21 media independent handover infrastructure [11], thereby supports single transceiver stations, horizontal and vertical handovers. In addition, the terminal-oriented decision mechanism supports always best connected services.

III. PROPOSED INTEGRATED LOAD BALANCING SCHEME AND ARCHITECTURE

The concept of our proposed integrated load balancing scheme leverages on estimation of critical network QoS parameters, specifically, packet delay in this work. We perform the bootstrap approximation in first stage to estimate the short-term stationary dynamic QoS parameters in an access point. We further account for the effects of non-stationary components in the second stage by performing the sequential Bayesian estimation with cumulative sum (CUSUM) monitoring in stations while listening to beacon broadcasts containing QoS parameter estimates. We refer readers to [6] for a more detailed description.

Accordingly, station would select the best access point according to their delay estimates which enabled us to obviate both detection and scanning phases of the 802.11 handover process, leading to significant Layer 2 handover latency reduction. The delay estimates are then augmented as load metric to devise a measurement-based soft admission control which is simple yet effective as it considers dynamic network conditions prevalent in broadband WLANs. The key idea is to ensure that the delay threshold of an access point is not violated when accepting new connections, which effectively protects QoS of existing connections by maintaining WLAN in an unsaturated mode. Soft admission control is important when considering multimedia traffic, since traditional hard admission control that applies pre-determined network capacity directly as admission threshold is ineffective against such dynamic traffic.

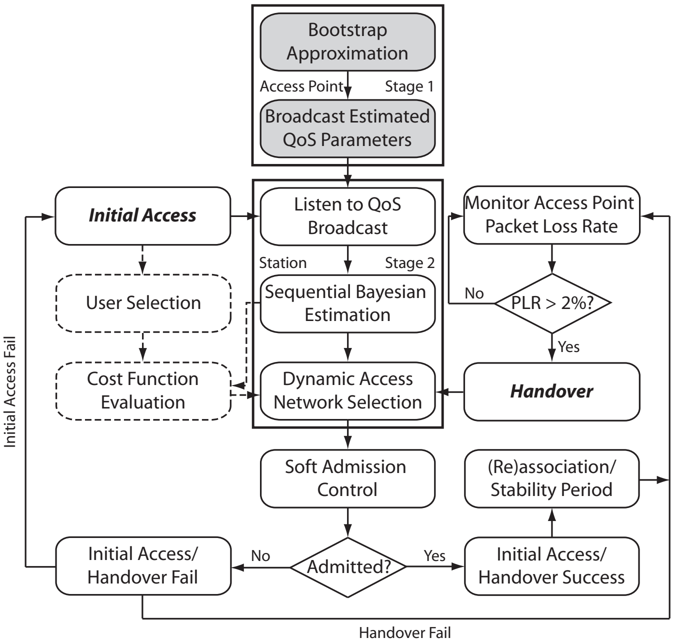
Bootstrap Approximation
Access Point
Stage 1
Broadcast Estimated QoS Parameters
Initial Access
Listen to QoS Broadcast
Monitor Access Point Packet Loss Rate
Station
Stage 2
User Selection
Sequential Bayesian Estimation
No
PLR > 2%?
Yes
Initial Access Fail
Cost Function Evaluation
Dynamic Access Network Selection
Handover
Soft Admission Control
(Re)association/ Stability Period
Initial Access/ Handover Fail
No
Admitted?
Yes
Initial Access/ Handover Success
Handover Fail

Fig. 1. Distributed terminal-oriented, network-assisted handover architecture.

We further advocate a distributed terminal-oriented, network assisted handover architecture suitable for future wireless networks as depicted in Fig. 1. The shaded blocks refer to network entities while the unshaded blocks refer to terminal entities. Always best connected services which consider both network conditions and user preferences during a network selection can also be supported by completing the blocks in dashed lines. However, these are outside the scope of this work. Our proposed handover architecture can be triggered by two events, viz. *initial access* to network where station would choose the best network according to their service QoS (packet delay) requirement and *handover* when the network QoS (packet loss rate) exceeds 2% for the case of VoIP services. Soft admission control located in each station would arbitrate the prevailing traffic load and QoS in terms of delay between a source and target access points. Upon admission, station would perform (re)association with the selected target access point during (handover) initial access. Otherwise, the station would continue to monitor the packet loss rate of its associated access point when handover fails or listen to QoS broadcast should initial access fails. A stability period of two beacon intervals is enforced before the other stations can make the next handover attempt to prevent the ‘ping pong’ effect.

The fundamental of our proposed handover architecture is *network-assisted discovery* such that source access point broadcasts measurement report of neighboring access points together with its own, compatible with the IEEE 802.21 media independent handover framework. As illustrated in Fig. 2,

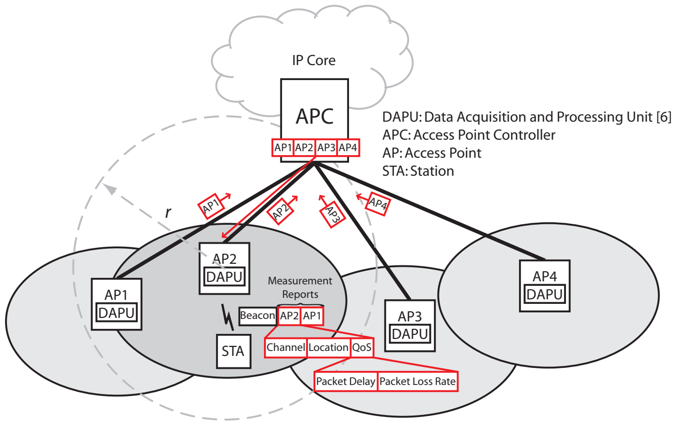

Fig. 2. Location-based broadcast in proposed handover architecture.

the estimated QoS parameters of an access point together with their channel number, location would be encapsulated in a packet as measurement report and transmitted to the access point controller (APC) periodically. The APC would then collect these measurement reports from every access point in their subnet. The consolidated measurement reports of the source and neighboring access points would finally be disseminated from APC using a *location-based* broadcast. This is motivated by the fact that it is not viable for stations to handover to a distant access point, making it unnecessary for station to monitor their network conditions. Since the locations of access points are usually fixed, it is feasible for an APC to maintain a location map information locally. Each APC can then exploit the location of source access point as center of circle with radius r to filter out any access points that lie outside the circumference, thereby disseminating only measurement reports of three nearest neighboring access points for broadcast by source access point.

The advantage of our proposed handover architecture is twofold. First, it supports prevailing single transceiver station without any hardware modifications by requiring source access point to broadcast information of neighboring access points in addition to its own. Second, scanning procedures for handover decisions can be eliminated since station listening to the broadcast would be able to get information of prospective neighboring access points. Consequently, our total Layer 2 handover latency as illustrated in Fig. 3 is significantly reduced as both detection and scanning delay is obviated. We exploited the fact that VoIP connections can tolerate some packet loss rate of 2% and utilize this as link layer detection to trigger handover. Since the best target access point is available from dynamic access network selection algorithm at the same instance, we do not incur any detection delay. Accordingly, it consists of 2-way handshake processing delay of typically $1ms$ required by the soft admission control, average channel switch time of $12ms$, authentication delay of less than $1ms$ [12] and average reassociation delay of $15.37ms$. We note that reassociation delay can be further reduced to $1.69ms$ by applying neighbor graph technique [13]. Therefore, our total Layer 2 handover latency is approximately $16ms$ to $30ms$ as opposed to existing total handover latency of more than $1000ms$ [5] when including link layer detection delay. Although physical layer detection is widely deployed to exclude link layer detection delay, the total handover latency can still be as high as $420ms$ [4].

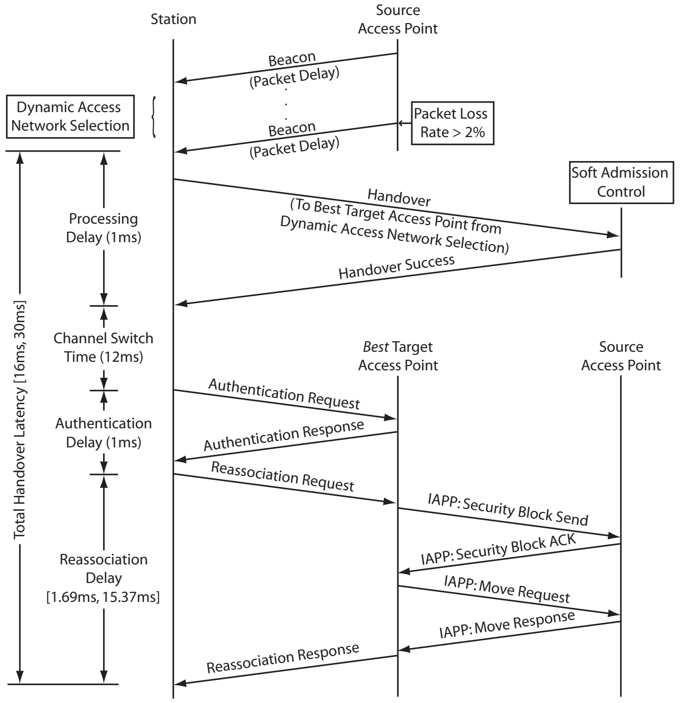

Fig. 3. Seamless handover process of integrated load balancing scheme.

IV. Simulation Model

Without loss of generality, we used the wireline-to-wireless network topology as shown in Fig. 4 in order to focus on delay within each BSS. We assumed codec delay of $40ms$, packetization delay of $20ms$ at both sender and receiver and backbone network delay of $30ms$. Therefore, the wireless network delay in both uplink and downlink should be less than $60ms$ in order to meet the one-way end-to-end delay requirement of VoIP packets. We simulated a typical hotspot scenario with one 802.11b and one 802.11g access point operating with maximum data rate of $11Mbps$ and $54Mbps$ respectively. We assumed that at least one legacy station has associated with the 802.11g access point. However, the legacy station do not transmit any traffic therefore all the system resources are available for 802.11g stations. We subject our simulation to multimedia traffic source as summarized in Table. I. Voice station generates VoIP stream using G.711 codec with silence suppression. Video station generates traffic according to MPEG-4 trace (Jurassic Park) [14] at 25 frames/sec and data station generates best effort FTP traffic.

The simulation models were developed using OPNET™ Modeler® 14.0 with Wireless Module. We further assumed no hidden terminals and excluded RTS/CTS mechanism from our simulation. We also incorporated MAC service data unit

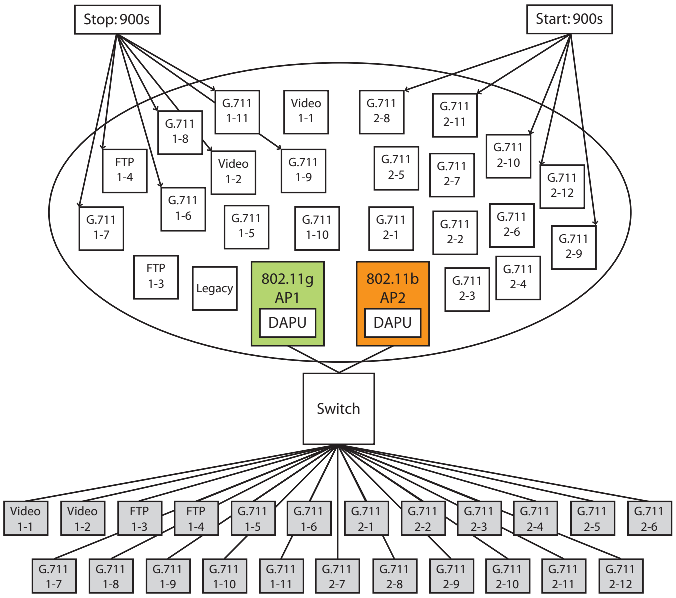

Fig. 4. Simulation model with mixed 802.11b/g access points.

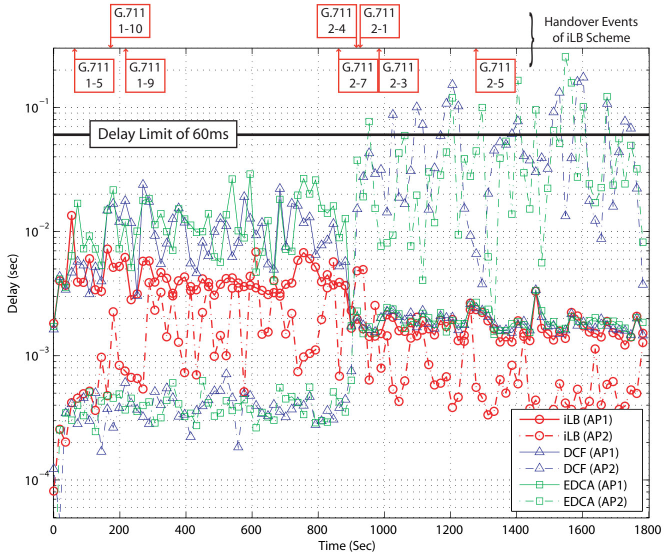

Fig. 5. Average downlink delay.

TABLE I
TRAFFIC GENERATION PARAMETERS.

Traffic Type	Packet Size (Bytes)	Inter-arrival (ms)	Avg. Data Rate (kbps)
Voice-G.711	80	10	64
Data-FTP (UL)	750	100	60
Data-FTP (DL)	3750	50	600
Video-High Quality	MPEG-4 trace	40	770

(MSDU) lifetime limit mechanism to discard MSDUs from the transmitter queue if they exceed the MSDU lifetime before successful transmission. The MSDU lifetime for voice and video packets are chosen as $50ms$ and $100ms$ respectively. Data packets have MSDU lifetime of $1s$. All stations in our simulations are roaming capable to support handover events. Mobility model is not considered since we are interested in QoS-related handover rather than radio-related handover.

In our simulation, we initially introduced an unbalanced load of two FTP, two video, seven G.711 stations in BSS 1 and seven G.711 stations in BSS 2. At time $900s$, one FTP, one video and five G.711 connections from BSS 1 were stopped, while five G7.11 connections from BSS 2 were started. These discrete events induce imbalance traffic load during our simulation for evaluating the responsiveness of our integrated load balancing scheme under such dynamic network conditions. We note that no perturbations are injected after $900s$ in order to observe the steady state performance. Finally, we investigate the performance of our proposed integrated load balancing scheme from two critical aspects. First, we examine its QoS performance in terms delay and packet loss of access points which reflects the capability of WLAN to support VoIP services. Second, we quantify the effect of load balancing on the overall system utilization by adopting the balance index introduced in [15] to reflect the used capacity and QoS condition in each access point. Suppose x_i is the total throughput or delay of access point i, then the balance index can be defined as,

$$\mathrm{B}\left(x\right) = \left(\sum_i x_i\right)^2 \Big/ n\left(\sum_i x_i^2\right), \qquad (1)$$

where n is the number of access points over which the load will be redistributed. The balance index is a continuous function which is independent of scale. It is bounded between 0 and 1 such that it has a value of 1 when all access points have the exactly the same throughput or delay and a value of $1/n$ when access points are extremely unbalance, which is 0 in the limit as $n \to \infty$.

V. Simulation Results

The simulation results presented include QoS performance of our integrated load balancing (iLB) scheme evaluated in terms of delay and packet loss. We then compare it with the 802.11b/g distributed coordination function (DCF) and the 802.11e enhanced distributed channel access (EDCA) which represent the cases without load balancing. Each VoIP connection has duplex traffic which eventually results in higher downlink load, leading to the classical bottleneck at access point for infrastructure-based WLAN [16]. Therefore, we excluded the uplink results due to space limit and focus on the average downlink delay and packet loss rate associated with each access point as shown in Fig. 5 and Fig. 6 since they are the limiting factors.

In our simulation, access point 1 with multimedia traffic is mildly overloaded while access point 2 with voice only traffic is highly overloaded for both DCF and EDCA. The overloading is predominantly due to physical layer detection of the existing 802.11 WLAN handover process which lacks QoS considerations. As a result, no handover is triggered since all

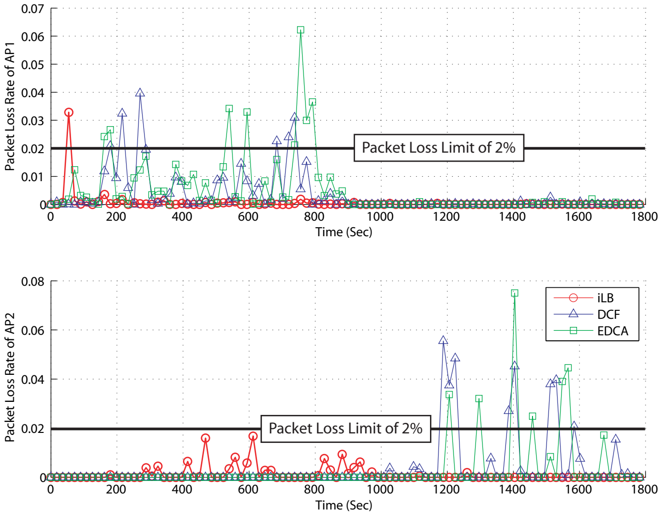

Fig. 6. Average downlink packet loss rate.

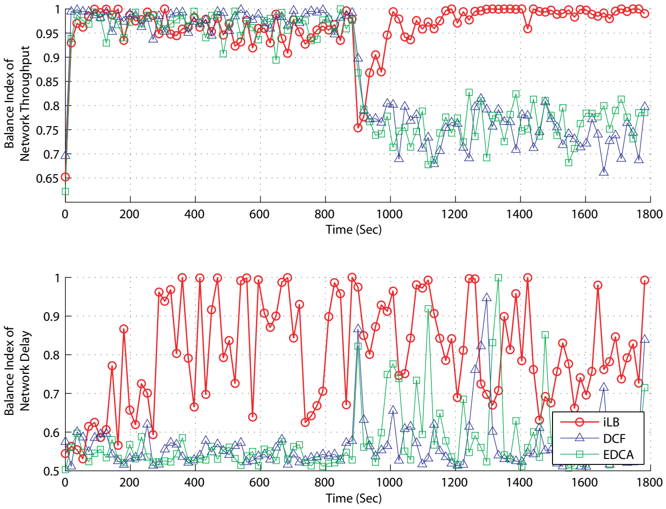

Fig. 7. Balance index of network throughput and delay.

stations are within good coverage region of their access points. On the contrary, QoS-related handovers are observed with iLB since it supports link layer detection which triggers a handover when packet loss rate of associated access point exceeds 2%. This together with soft admission control effectively mitigate overloading of both access points. We note that a handover will only be triggered on the conditions that (i) packet loss rate of source access point is more than 2% and; (ii) there exists a target access point which can better meet the delay requirement of VoIP services. Finally, the handover attempt can only be completed if the target access point can still accept connections when subjected to soft admission control. As such, there will be no additional loss associated with a particular handover when sucessfully triggered and its service QoS shall be guaranteed after handover since WLAN will operate in unsaturated mode to protect QoS of existing voice connections.

It is evident that both DCF and EDCA are unable to support the strict QoS requirements of real-time VoIP services, where the delay incurred by WLAN should be less than $60ms$ and the packet loss rate should be less than 2%. Accordingly, DCF and EDCA have an average downlink delay of up to $170ms$ and $250ms$ in access point 2 respectively. In addition, DCF and EDCA have an average downlink packet loss rate of up to 4% and 6% respectively in both access points. These observations are due to buffer overflow phenomenon in both access points which is operating beyond its maximum capacity and hence experienced excessive delay and consequently packet loss. Although the QoS prioritization mechanism of EDCA achieves the best uplink performance in both access points, it has the worst downlink performance in terms of average delay and packet loss rate when subjected to heavy load. Particularly, we can see that DCF performs *better* than EDCA which suggests that the smaller contention window sizes in EDCA cause increased collisions which have a strong negative impact on downlink performances. We believed that iLB could effectively mitigate this problem, particularly when EDCA is utilized for voice traffic of same priority which reduces to classical DCF scenario.

With the introduction of iLB scheme, we achieved an average downlink delay of less than $14ms$ together with an average downlink packet loss rate of less than 2% in both access points throughout the simulation. The initial packet loss rate of 3.2% is a result of our link layer detection that triggers QoS-related handovers. Clearly, the access points which used to be the bottleneck are now able to support real-time VoIP connections in presence of multimedia traffic with bounded average delay and packet loss rate. We note that iLB also exhibits throughput and QoS fairness which jointly improve overall system utilization in contrast to DCF and EDCA as shown in Fig. 7. The balance index of network throughput for DCF and EDCA without load balancing is 0.86 which improves to 0.96 with iLB. Similarly, the balance indexes of network delay for DCF and EDCA without load balancing are 0.56 and 0.58 respectively which improve to 0.81 with iLB. We attained optimal load balancing since our estimated packet delay metric directly optimizes the expected packet delay, making it adaptive to dynamic network conditions. This augmented our soft admission control in allowing us to support multimedia traffic, which is not possible with the traditional hard admission control technique. We remark that our iLB scheme provides a normalized approach to effectuate load balancing irrespective of network data rates as shown in our simulation comprising of a mixture of 802.11b and 802.11g access points. Moreover, our proposed handover architecture supports access network heterogenity through the notion of broadcasting. Therefore, we argue that our iLB scheme can be fully extended to support vertical handovers in future heterogeneous wireless networks.

We have demonstrated the importance of integrated load balancing scheme in future wireless networks. However, any derived benefit comes at a cost to both network and terminal which we would briefly discuss. For network, there will be no additional signaling overhead as measurement reports are appended to beacons, which are periodically broadcasted by an access point to announce its existence. There would be some storage and communication overheads for updating these measurement reports. However, these would not impose heavy loads since location-based broadcast is restricted to only three nearest neighboring access points. For terminal, we consider computational complexity which would manifest as power consumption. Although our proposed scheme requires additional computations to perform network selection, we expect it to be minimal since our algorithm has linear time complexity of $O(n)$. Moreover, the exclusion of scanning phases in our QoS-based fast handover scheme helps to offset this incremental computational cost.

VI. Conclusion and Future Work

We presented an integrated load balancing (iLB) scheme that leverages on QoS-based fast handover to provide seamless handover and soft admission control to protect QoS of existing connections when resources are low. By means of estimating critical QoS parameters, we are able to eliminate both detection and scanning phases from the 802.11 WLAN handover process and devise a robust soft admission control to support multimedia traffic, otherwise not conceivable with hard limiting approaches. We showed by induction that our iLB scheme is able to support seamless handover with total Layer 2 handover latency of $16ms$ to $30ms$. We further demonstrated by simulation that a bounded average downlink delay of less than $14ms$ and a bounded average downlink packet loss rate of less than 2% is achievable. These satisfy the stringent QoS requirements of real-time VoIP connections in presence of multimedia traffic. In summary, iLB offers four main benefits, viz. (i) QoS guarantee during handover with fast handover; (ii) QoS guarantee after handover with soft admission control; (iii) exhibits both throughput and QoS fairness which jointly improve overall system utlization; and (iv) normalized load balancing solution irrespectively of access network heterogenity.

In this work, we have achieved significant QoS enhancements over DCF by employing iLB. We also showed that IEEE 802.11e standard could not guarantee the QoS requirement of real-time VoIP services without an appropriate load balancing mechanism. For future work, we would extend our iLB scheme to the 802.11e standard and sought further performance gains by optimizing EDCA parameter set adaptively using the notion of our QoS parameters estimation technique. We would also investigate the performance of iLB in heterogeneous networking environments.

References

[1] H. Zhai, X. Chen, and Y. Fang. How well can the IEEE 802.11 wireless LAN support quality of service? *IEEE Transactions on Wireless Communications*, 4(6):3084–3094, November 2005.

[2] ITU-TG.114. One-way transmission time. 2003.

[3] C. Shim, L. Xie, B. Zhang, and C. Sloane. How delay and packet loss impact voice quality in voip. Technical report, Qovia, Inc., December 2003.

[4] A. Mishra, M. Shin, and W. Arbaugh. An empirical analysis of the ieee 802.11 mac layer handoff process. *SIGCOMM Comput. Commun. Rev.*, 33(2):93–102, 2003.

[5] H. Velayos and G. Karlsson. Techniques to reduce the IEEE 802.11b handoff time. In *Communications, 2004 IEEE International Conference on*, volume 7, pages 3844–3848, June 2004.

[6] E. H. Ong and J. Y. Khan. Dynamic access network selection with QoS parameters estimation: A step closer to ABC. In *Vehicular Technology Conference, 2008. VTC Spring 2008. IEEE*, pages 2671–2676, Marina Bay, Singapore, May 2008.

[7] G. Bianchi and I. Tinnirello. Improving load balancing mechanisms in wireless packet networks. In *Communications, 2002. ICC 2002. IEEE International Conference on*, volume 2, pages 891–895, 2002.

[8] A. Bazzi, M. Diolaiti, and G. Pasolini. Measurement based call admission control strategies in infrastructured IEEE 802.11 WLANs. In *Personal, Indoor and Mobile Radio Communications, 2005. PIMRC 2005. IEEE 16th International Symposium on*, volume 3, pages 2093–2098, September 2005.

[9] A. Balachandran, P. Bahl, and G. M. Voelker. Hot-spot congestion relief in public-area wireless networks. In *Mobile Computing Systems and Applications, 2002. Proceedings Fourth IEEE Workshop on*, pages 70–80, 2002.

[10] H. Velayos, V. Aleo, and G. Karlsson. Load balancing in overlapping wireless LAN cells. In *Communications, 2004 IEEE International Conference on*, volume 7, pages 3833–3836, June 2004.

[11] IEEE-P802.21/D10.0. Draft ieee standard for local and metropolitan area networks: Media independent handover services. April 2008.

[12] I. Ramani and S. Savage. Syncscan: practical fast handoff for 802.11 infrastructure networks. In *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, volume 1, pages 675–684, March 2005.

[13] A. Mishra, M. Shin, and W. A. Arbaush. Context caching using neighbor graphs for fast handoffs in a wireless network. In *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, volume 1, March 2004.

[14] F. H. P. Fitzek and M. Reisslein. MPEG-4 and h.263 video traces for network performance evaluation. *IEEE Network*, 15(6):40–54, November/December 2001.

[15] D. Chiu and R. Jain. Analysis of the increase and decrease algorithms for congestion avoidance in computer networks. *Comput. Netw. ISDN Syst.*, 17(1):1–14, 1989.

[16] S. Shin and H. Schulzrinne. Balancing uplink and downlink delay of voip traffic in wlans using adaptive priority control (apc). In *QShine '06: Proceedings of the 3rd international conference on Quality of service in heterogeneous wired/wireless networks*, page 41, New York, NY, USA, 2006. ACM.