(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
26 June 2008 (26.06.2008)

PCT

(10) International Publication Number
WO 2008/076945 A2

(51) International Patent Classification:
G08C 15/00 (2006.01)

(21) International Application Number:
PCT/US2007/087677

(22) International Filing Date:
14 December 2007 (14.12.2007)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/875,269	14 December 2006 (14.12.2006)	US
11/956,141	13 December 2007 (13.12.2007)	US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier application:
US 11/956,269 (CON)
Filed on 13 December 2007 (13.12.2007)

(71) Applicant *(for all designated States except US)*: **NTT DOCOMO, INC.** [JP/JP]; Sanno Park Tower, 11-1, Nagata-cho, 2-chome, Chiyoda-Ku, Tokyo, 100-6150 (JP).

(72) Inventors; and

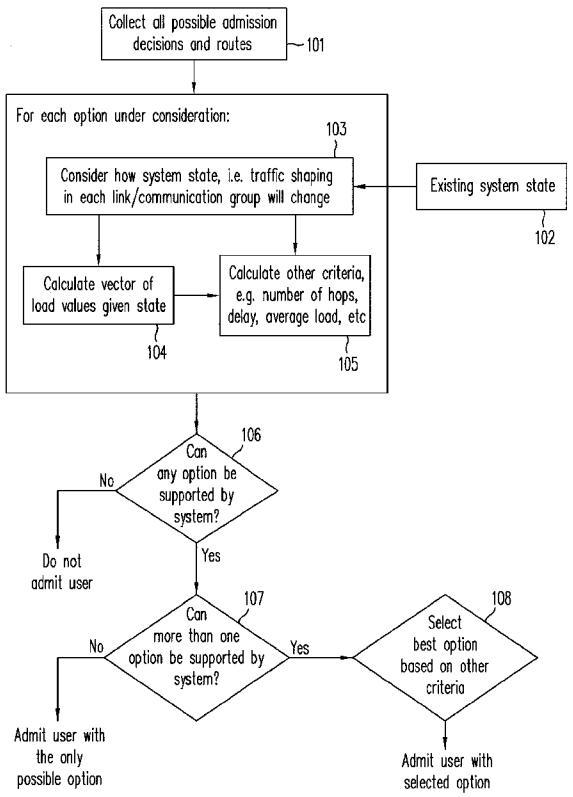
(75) Inventors/Applicants *(for US only)*: **RAMPRASHAD, Sean, A.** [GB/US]; 960 Terrace Drive, Los Gatos, California 94024 (US). **LI, Danjue** [CN/US]; 371 Elan Village, Unit 320, San Jose, California 95134 (US). **KOZAT, Ulas C.** [TR/US]; 3612 Flora Vista Avenue, #349, Santa Clara, California 95051 (US). **PEPIN, Christine** [CA/US]; 712 Hans Avenue, Mountain View, California 94040 (US).

(74) Agent: **KWOK, Edward C.**; MacPherson Kwok Chen & Heid LLP, 2033 Gateway Place, Suite 400, San Jose, California 95110 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY,

[Continued on next page]

(54) Title: METHOD AND APPARATUS FOR MANAGING ADMISSION AND ROUTING IN MULTI-HOP 802.11 NETWORKS TAKING INTO CONSIDERATION TRAFFIC SHAPING AT INTERMEDIATE HOPS



(57) Abstract: A method and a system control admission for voice transmission over a multi-hop network. The method takes multiple parameters into account in order to make a decision and ensures that the system remains stable. The parameters may include options (e.g., aggregation level, bursting level, and transmission rate) and constraints (e.g., the number of users, access points, or sensing range of each user). The method computes the load of the network given each set of options and constraints and compares it against the packetization interval of the voice codec to check whether or not the system is stable. An algorithm of the method finds the best trade-offs between overhead reduction (e.g., due to contention and to packet headers) and a solution robust to channel errors, if links are noisy. If several stable solutions corresponding to different options exist, additional criteria (e.g., the least number of hops, the highest transmission rate) may be used to determine the final "best" solution.

WO 2008/076945 A2

TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

Method and Apparatus for Managing Admission and Routing in Multi-hop 802.11 Networks Taking into Consideration Traffic Shaping at Intermediate Hops

CROSS REFERENCE TO RELATED APPLICATIONS

The present application relates to and claims priority of (a) U.S. provisional patent application no. 60/875,269, filed on December 14, 2006; and (b) U.S. patent application no. 11/956,141, filed December 13, 2007, both of which are incorporated herein by reference. For the US designation, the present application is a continuation of the aforementioned U.S. patent application no. 11/956,141.

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention relates generally to data traffic over a multi-hop network. More specifically, the present invention relates to managing admission and routing of data streams transmitted over a multi-hop 802.11 based network (or a similar network) that provides traffic-shaping and rate-adaptation services at each hop.

2. Discussion of the Related Art

A mechanism for both controlling admission and making routing decisions is a desirable tool for efficient network management. In a 802.11 network, such a mechanism determines for each new user which access point to associate with and which route, among the multiple potential routes, traffic and packets associated with the user should follow to reach their respective destinations. Such a mechanism may be driven, at least in part, by resource availability (e.g., bandwidth and the number of access points) and the users' Quality-of-Service (QoS) requirements (e.g., transmission rate, throughput, delay, and signal-to-noise ratio). A new user may be denied admission when the required resources are not available, or when user QoS requirements (for both the new user and existing users) would not be satisfied if the new user is admitted. For an interactive voice communication application (e.g., "voice over IP" or "VoIP"), the resource and QoS requirements may include transmission throughput, delay and packet loss. Thus, a desirable admission control policy ensures that the network can support a given number of users without violating the QoS constraints, or the limits of the system. In a multi-hop network, the admission control policy should also

prevent bottlenecks, and therefore serves to limit the number of users that can be supported.

A number of admission control algorithms have been proposed for a single-hop wireless network. For example, the article "Channel quality dependent scheduling for flexible wireless resource control," by Z. Jiang, L. Chang, and N. K. Shankaranarayanan, published in the Proceedings of IEEE Globecom 2001, discloses a method that optimizes only throughput. The article "Distributed multi-hop scheduling and medium access with delay and throughput constraints in mind," by V. Kanodia, C. Li, A. Sabharwal, B. Sadeghi, and E. Knightly, published in Proceedings of ACM MobiCom 2001, discloses a method that optimizes both the throughput and the delay. For a 802.11-based multi-hop network, the article "Admission control for multihop wireless backhaul networks with QoS support" ("LNPWZ06"), by S. Lee, G.. Narlikar, M. Pal, G. Wilfong, and L. Zhang, published in the Proceedings of IEEE WCNC 2006, Las Vegas NV, April 2006, discloses a method that meets each user's QoS requirements in delays and connection rates.

Other admission control algorithms take into consideration other options, such as aggregation of voice packets at access-points in the scenario of a multi-hop network. Aggregation of voice packets is disclosed, for example, in the article "A Joint Traffic Shaping and Routing Approach to Improve the Performance of 802.11 Mesh Networks" ("WiOpt06"), by C. Pepin, U. Kozat, and S. A. Ramprashad, published in the Proceedings of WiOpt 2006, April 3-7, 2006 and in the U.S. patent application "Method for Improving Capacity in Multi-Hop Wireless Mesh Networks" ("NPA06"), serial no. 11/531,384, by S. Ramprashad, C. Pepin, and U. Kozat.

Once a user is admitted into the network, a routing scheme of the system determines proper routing of the user's data packets to their destinations. Routing in a multi-hop wireless network has been extensively discussed in the literature. In earlier wireless routing protocols (e.g., AODV, DSR, DSDP, and TORA), routing is handled independently from the lower layers, with path discovery being performed in a best-effort fashion without regard for system performance or QoS. Such routing algorithms are intended mainly for Mobile Ad Hoc networks (MANETs), where finding a connected path has priority (see, e.g., http://www.ietf.org/html.charters/manet-charter.html).

It is important to note that such methods do not directly consider important issues at the lower layers. Specifically, if a network is such that there are high overheads in handling and transmitting packets at lower communication layers, as is the case of 802.11 networks , a routing protocol that does not take such overheads and lower layers into consideration can lead to poor system performance.

Taking QoS into consideration in path discovery is discussed, for example, in the

article "A high-throughput path metric for multi-hop wireless routing," D. S. J. De Couto, D. Aguayo, J. Bicket, and R. Morris,in Proceedings of ACM Mobicom 2003, San Diego CA, September 2003. That article discusses a method that minimizes the expected total number of transmissions and retransmissions required to successfully deliver a packet. A new metric is devised which incorporates the effects of link loss ratios, asymmetry in the loss ratios in the two directions of each link, and interference among the successive links of a path. The metric is implemented in the DSDV and DSR routing protocols and is shown to provide better performance than a minimum hop-count metric, particularly for paths with two or more hops.

The prior art includes complex approaches that jointly route and schedule packets to satisfy user QoS requirements. Such approaches can be ineffective due to the complexity of the problem. For example, solving the joint scheduling and routing problem to satisfy required connection rates in a multi-hop wireless network is a known NP-complete problem. See, e.g., the LNPWZ06 article discussed above. The prior art approaches are also ineffective because they focus on only one or two parameters at a time (e.g., throughput and delay) for each user independently. Therefore, in general, the resulting solutions cannot easily scale with the increase in users or access points. In fact, as new users join the network, such approaches may even create bottlenecks.

Beyond such general approaches one may consider more carefully the applications being admitted into the system and by doing so be able to improve performance and simplify the optimization problem. For example, for voice applications (e.g., VoIP applications), or media applications in general that generate small packet sized data persistently, many benefits may be achieved using traffic-shaping at intermediate hops as shown in "WiOpt06" and "NPA06". In such a scenario one can directly consider important issues at lower layers such as the Medium Access Control (MAC) and Physical (PHY) layers. Specifically, by carefully selecting lower-layer mechanism such as aggregation and bursting levels at each hop one can reduce the inherent inefficiencies under the 802.11 protocol for voice or similar traffic. With this the number of VoIP calls that may be supported by such a system may increase substantially. Admission and routing control in such a scenario should take into account such lower layer mechanisms.

SUMMARY

According to one embodiment of the present invention, a method and apparatus control admission of data streams over a multi-hop 802.11 network. The admission control method specifically takes into consideration the joint effects of multiple mechanisms, such as traffic-shaping (e.g., aggregation and bursting) and routing or forwarding packets through multiple wireless links or hops. Because an admission decision impacts both the aggregation and bursting capabilities of the system, and thus the behavior of MAC and PHY layers, these

MAC/PHY layer factors are considered directly in admission control. Additional considerations may include those for existing mechanisms already in use in the system, such as adapting the physical transmission rate at each hop to channel conditions (e.g., signal-to-noise ratio).

According to one embodiment of the present invention, the method views the nodes (i.e., access points, gateway access points, terminals) and system as a collection of communication groups. A communication group is a set of flows originating from 802.11 nodes that are contending for a common wireless resource. See for example the illustrations in Figures 2 and 3 where flows (links) are labeled with a number "n" indicating their group membership. One can also consider such a group to be the interfaces (a MAC and an associated PHY) provided on such nodes that contend among themselves for a common wireless resource. The mechanism admits a user into the network (i.e. select for the user both a suitable access point (AP) and selects a route to the gateway access point (GAP)) in a manner such that the resulting changes in the states of different communication groups in the system achieve both a feasible (i.e., stable) system and desirable performance attributes (e.g., maximizing the number of users that can be supported by the system). Here, the "state" of a communication group includes the implicit or directed states of lower layer mechanisms, such as the aggregation or bursting levels each node in a group will use to transmit traffic.

According to one embodiment of the present invention, when a new user enters the system, an access point is selected for association by the new user . This will be the access point with which the user's terminal communicates into the multi-hop network. Furthermore, a route from that access-point to the GAP is assigned. In some instances, the route may be implicitly assigned upon admission because of a simple network topology. In other instances the system may consider a number of possible routes based on the admission decision. In other situations, it can be that the user terminal is connected to an access-point which is itself a GAP, and thus no route is necessary. Correctly managed under the present invention, admission (and associated parameter settings) leads to system parameter values that achieve a desirable performance level (e.g. maximized efficiency).

Further, according to one aspect of the present invention, a method of the present invention not only takes into account admission and routes as a decision independent of lower layers, but directly takes into account the lower-layer traffic-shaping mechanisms used for transferring packets between access-points. It does do by either effectuating an implicit change in traffic-shaping mechanism used by each node in the system, or by directly changing the traffic-shaping mechanism used by each node in the system. An implicitly directed change would include, for example, eliciting a known deterministic (or average) reaction of the access point to a new flow being routed through it. The admission and route

decision defines which access points receive additional traffic, and thus the reaction can be predicted. For example, the system may have access points that implement a rule which aggregates, when possible, one incoming packet from each flow into a common next hop packet, sends the packets in burst transmissions, or performs both, when the flows have the same next-hop access point destination. A directed change specifies how each node should handle incoming flows to aggregate and burst in common transmissions.

In both cases (i.e., implicit and direct changes), the admission and routing decision implies a change in state variables in each communication group. The state variables associated with a communication group may include, for example, an aggregation level (e.g., the number of packets that may be multiplexed together in a larger packet), a bursting level (i.e., the number of packets that can be sent in a transmission burst) and a physical-layer (PHY) transmission rate that are used by the nodes in accessing the common physical medium under contention.

One advantage of the present invention is that desirable traffic-shaping parameters for each communication group are determined at the same time the best admission and routing strategy are determined for each new user.

The present invention does so jointly in providing an admission control method for traffic, e.g. voice transmission, over a multi-hop 802.11 network. The method takes multiple parameters into account in order to make a decision and ensures that the system remains stable. The parameters may include options (e.g., aggregation level, bursting level, and transmission rate) and constraints (e.g., the number of users, access points, or sensing range of each user). The method computes the load of each of the communications groups in the network given each set of options and constraints, as determined by a potential admission and routing decision, and compares it against the wireless resources available to each communication group. Note, a load is calculated for each wireless resource being used by the system, i.e. there is a load calculated for each communication group. In the case of voice, it is convenient to consider a measure of load in terms of the channel occupancy time relative to the packetization interval of the voice codec, i.e. the amount of wireless resources (e.g., time) consumed in each communication group per packetization interval of the voice codecs. This is a convenient measure and time interval in VoIP, as a voice coder produces one packet per packetization interval in each direction of a call. In more general cases, one can consider any interval and the average time the wireless medium of a given communication group is busy during that interval. Even more generally, one can simply consider the long-term average time the wireless medium of a communication group is busy.

For a given option (admission, routing, and traffic shaping option) one then checks the load required to handle the number and type (size, bursting, etc) of packets that would be

transferred in each communication group against the available wireless resources. By doing so one can check whether or not the system operates in a stable fashion for the given option. Furthermore, among options that can operate in a stable fashion, the method has a measure of load on the system for each communication group. Among these options for stable operation, one can then select an option based on other criteria, e.g., the least number of hops, the highest transmission rate, the lowest load over all communication groups, or the lowest average load.

In investigating options, one can also consider the effects of packet losses, channel impairments, physical layer rates, and adaptation. Such effects are all accounted for in the load calculation. Therefore an option may additionally include, beyond routing, admission, bursting and aggregation, an implicit or direct control of the physical layer data rate used by each transmission in a communication group.

In comparison, prior art admission control methods rely on other considerations and parameters, and often only to decide whether or not a user is admissible. Connection rate and delay are the two most common parameters. Such techniques do not fundamentally consider a system which is additionally optimized by traffic-shaping. The idea of traffic-shaping, as defined and managed in 802.11 multi-hop networks, is unique. Stability and efficiency may be achieved jointly with admission and traffic-shaping.

The present invention is better understood upon consideration of the detailed description below and the accompanying drawings. The present invention is applicable, for example, to a voice-over-IP (VoIP) application on a multi-hop 802.11 network.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 illustrates a method that takes admission, routing and traffic-shaping into consideration in the process of admitting a new user, in accordance with one embodiment of the present invention.

Figure 2 shows system 200 to which the present invention is applicable.

Figure 3 shows alternative system 300 to which the present invention is applicable.

Figure 4 shows an example of using both aggregation and bursting at a transmission opportunity to improve efficiency.

Figure 5 shows a system with two GAPs which is not tree-rooted and for which the principles also apply).

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

The present invention provides, in an 802.11-based or similar network, a method to admit and route users. The method routes the user's traffic through multiple-hop wireless links, determines and provides (implicitly or explicitly) traffic-shaping options at one or more of the hops. In certain situations, the admissibility condition is also a stability condition that each communication group has to satisfy for a given set of parameters or options (e.g., aggregation level, bursting level, and connection rates). The solution determines an acceptable set of parameters (more than one solution may exist) and the route and traffic profile across the network, for each user and communication group, without overloading or violating Quality of Service (QoS) constraints imposed for existing and new users.

According to one embodiment of the present invention, a central or "control" entity collects data regarding the state of the network and computes the loads for each communication group as would occur under different potential options. This control entity can exist anywhere in the network; in fact, it may be collocated with or implemented in the GAP. The control entity may collect information from and may take advantage of entities within the network that include automated mechanisms for adaptation (e.g. some communication groups may be able to adapt certain parameters autonomously as mentioned above, based the links they support). One example of such an automated mechanism is the relay access point (RAP) which makes local decisions on aggregation and bursting levels in each of its outbound link, based on corresponding inbound traffic parameters, such as the number and types of inbound users, the number and types of inbound aggregated or burst flows from other RAPs, and the SNRs on different links.

Unlike the prior art, a method under the present invention takes into account multiple parameters (e.g., aggregation level, bursting level, and connection rates) for selectively admitting users. In the detailed description below, "system load" may be determined from calculating the relative channel occupancy time in each communication group. System load may be used as a metric for stability. A stable system results when the communication groups can support the system load. For voice traffic, for example, if the time interval between voice packet generation for each user is x milliseconds, and the time required to transmit all packets generated over the x milliseconds interval is y milliseconds, for all traffic associated with users having links passing through a given communication group, y being greater than x, the system is unstable. The stability condition ensures that the voice packets of all voice calls are serviced in a timely manner, without a long queuing time and without a large packet loss rate (e.g. less than 1%, overall). One can also consider a more aggressive condition for stability, such as the y being less than x by a known gap, i.e. a link budget.

According to one embodiment of the present invention, the stability condition is deemed only a limiting condition. That is, there are many possible admission and routing

options that can support a user and enable the system to perform in a stable fashion, providing the required quality of service to all users. A "system load" vector may be provided to adapt the traffic-shaping options under a given admission or routing selection. The system load vector may also be used to indicate unstable communication groups.

One approach to admission that takes into account multiple parameters is disclosed in the article "Analyzing and Managing Traffic Shaping in the Transmission of Voice over Multi-Hop 802.11 Networks" ("LKRP06"), by D. Li, U. C. Kozat, S. A. Ramprashad, C. Pepin, *unpublished DoCoMo Internal Document*, Nov 2006. This article provides a mathematical description of the load calculation that can be used effectively.

The article under review "An Analysis of Joint Aggregation, Bursting and Rate Adaptation Mechanisms for Increasing VoIP Capacity in Multi-hop 802.11 Networks", S. A. Ramprashad, D. Li, U. C. Kozat, and C. Pepin, under review by IEEE Transactions on Wireless Communications, submitted Feb 2007, revised Oct 2007. ("RLKP07") describes the concept in greater detail. In particular this article provides detailed mathematical descriptions of another method of load calculation which accounts for the transmission of packets in greater detail. To provide an in-depth discussion for these load calculation techniques, the LKRP06 and RLKP07 articles are attached herewith as Appendices A and B, and are hereby incorporated by reference in their entireties.

Figure 2 shows system 200 to which the present invention is applicable. As shown in Figure 2, system 200 includes a number of access points, including relay access points (RAP1 to RAP3) and a gateway access point (GAP), thus creating a multi-hop network. (Technically, the network of Figure 2 is a mixed 2-hop and 3-hop tree-rooted network. Figure 5 shows a system with two GAPs which is not tree-rooted and for which the principles also apply). The GAP provides connectivity to another network (e.g., a wired network). For example, the GAP may be a high-speed wired connection to the Internet. System 200 provides wireless multi-hop paths from mobile terminals to the GAP. Such a system may include multiple GAPs as in Figure 5. In Figure 2, the multi-hop network consists of multiple communication groups, with each communication group and its links being labeled by the same reference number (e.g., communication group 1 includes four links associated with the GAP). A communication group may be defined by the radio channel or the radio interfaces shared by the links. The radio channel or interfaces are a common wireless resource that is shared by the associated originating or destination mobile terminals or access points. One example of a wireless resource is an 802.11 channel, or a 802.11 hybrid scheme, such as HCCA. The wireless resource may also consist of different time intervals (e.g., a communication group may be refined to be the distributed contention time-interval or a centralized controlled (Point Coordination Function –PCF mode) interval).

The network supports a number of mobile clients, providing each mobile client a route (i.e., a set of one or more connected links) to at least one GAP. Clients may connect to any one of the four access points RAP1-RAP3 or the GAP. The packets exchanged between the GAP and the mobile client may traverse only one-hop to the GAP (e.g. in Figure 2 clients that connect directly to the GAP over links labeled "1"), or multiple hops to the GAP. (Although the basic network topology illustrated in this detailed description is a tree-structure, the method of the present invention is applicable to more general network topologies.) As shown in Figure 2, access points can participate in several communication groups. One implementation provides an access point with different wireless network interface cards that allow them to receive and transmit simultaneously in each group. Alternatively, a single interface card is used with appropriate time-sharing between its use for multiple communication groups.

One may generalize the concept of communication groups to simply links that use the same channel or strongly interfere with each other. For example, Figure 3 illustrates another allocation of communication groups, in which a relay access point RAP3 uses the same time and channel as the GAP. Such an arrangement results in a system with less performance than the one shown in Figure 2. However, such an arrangement may be user or preferable when the number of independent available 802.11 channels or time slots in the system is inadequate.

Each communication group may be characterized by the following state variables: (a) the aggregation level used in each of the outbound flows to other terminals, RAPs or GAPs; (b) the bursting level used in each of the outbound flows to other terminals, RAPs or GAPs; (c) the aggregation level used in each of the inbound flows from other terminals, RAPs or GAPs; (d) the bursting level used in each of the inbound flows from other terminals, RAPs or GAPs; (e) the number of mobile terminals directly supported by the APs in the group; (f) the packet generation rate (e.g., in packets/second) and the data rate (e.g., in mega or kilo bits/second) in each flow and from each user; (g) the physical layer data rate of each link or flow; (h) the channel state seen by each flow; and (i) the medium access control (MAC) parameter settings or each MAC mechanism in the group. While this list of parameters is not exhaustive, these parameters may be used to estimate a "load" on each communication group, according to one embodiment of the present invention, as discussed below. Many of these parameters are traffic-shaping parameters. Other state variables are possible, and include, for example, the present buffer levels at each transmitting terminal or MAC mechanism.

While communication groups are assumed not to interfere with each other in the wireless medium (for the most part), communication groups affect each other through the nature of their respective flows. For example, in Figure 2, the state of one communication

group "4" affects the state of communication group "1". This is because the packets sent to (from) mobiles in group 4 do eventually terminate (originate) at the GAP, and therefore do use the link RAP3 ←→GAP in group "1". Furthermore, the traffic-shaping options of a first communication group on one RAP or GAP can affect the traffic-shaping options of other communication groups or links involve in forwarding the traffic of the first communication group. For example, the traffic-shaping settings and state of RAP1 applied on its group "2" link to RAP2 affects the QoS in communication group "2". The selected traffic-shaping options may even have an effect on the QoS seen in group "1" since the flow from RAP1→RAP2 eventually flows to the GAP over a group "1" link. Further, in a communication group, traffic-shaping inbound traffic (i.e., the number and the types of packets) can affect traffic-shaping on outbound packets. After all, inbound traffic is collected together for outbound transmissions.

In this detailed description, one focus application is VoIP traffic. Stability with respect to the packetization interval of the voice codec may be used as a performance parameter for an VoIP application. For simplicity, the following model assumes that the network support only VoIP applications and considers stability with respect to the packetization interval. If all users use the same coder and packetization interval, the method of the present invention described below computes a channel occupancy time (i.e., the "load") of each communication group as the expected (or maximum) time to transfer on the links in the communication group one packet per user that uses this group. Typically, a user's route to the GAP may include multiple links and thus involves multiple communication groups.

At each access point (i.e., at the GAP and at each of the RAPs), packets are collected for transmission over the wireless links along the assigned routes. The collected packets to a common destination (over a common link) may be processed as a group in various ways to allow efficient use of the wireless resources. Specifically, an access point may choose between (i) several options of "aggregating" multiple voice or transmission packets into larger single transmission packets and (ii) using a single transmission opportunity (e.g., an opportunity granted after contention in the Carrier Sense Multiple Access-CSMA scheme of 802.11) to "burst" multiple transmission packets over the wireless medium. Using both the common transmission packet payloads and the transmission opportunities efficiently reduces overheads associated with data transmission over the 802.11 wireless interface. As discussed in the WiOPT06, LKRP06 and RLKP07 articles, overheads can consume more than 2/3 of the total available wireless resource, in some instances, leaving 1/3 or less of the wireless resource for actual data transmission. By placing more data in each packet, or by transmitting more packets at each transmission opportunity, the relative overhead per bit of payload information transmitted over an 802.11 link is reduced.

Given the number and the types of flows supported by a communication group, a combination of system parameters (e.g., "aggregation level", "bursting level", "physical layer rate, and "signal-to-noise ratio on the link") affect the resource utilization efficiency of the wireless resource each communication group has over the wireless medium. For a given communication group, such parameters are given values that result in the best efficiency, or a desired level of efficiency or performance (e.g., delay). Figure 4 shows an example of using both aggregation and bursting at a transmission opportunity to improve efficiency. As shown in Figure 4, at a transmission opportunity, a burst of two transmission packets (i.e., packets 401 and 402) and acknowledgement of packet 403, separated from each other by a short interframe spacing (SIFS). Transmission packets 401 and 402 are each formed by aggregating three voice data packets. Here the aggregation level on flow "i" is indicated by "A(i)", and the bursting level by "B(i)" as in LKRP06 and RLKP07. More generally, each packet in a burst transmission may use a different aggregation level. Here, in the mathematical framework below and in the Appendices, we use"A(i,j)" to refer to the aggregation level of the j-th packet in a burst of flow "i", where "j" goes from 1 to B(i). All "A(i,j)" and "B(j)" values can also be statistical quantities.

According to one embodiment of the present invention, when a new user enters the system, an access point is selected for the new user to associate and a route to the GAP is assigned. In some instances, the route may be implicitly assigned upon admission due to the simple topologies. Admission (and associated parameter settings) is correctly managed such that the resulting system parameter values achieve desirable performance (e.g. maximized efficiency). Further, a method of the present invention not only takes into account admission and routes, but also defines (implicitly or explicitly) a traffic-shaping mechanism (e.g., aggregation and bursting level) for transferring packets collected at one access point to the next access point.

Figure 1 illustrates a method that takes admission, routing and traffic-shaping into consideration in the process of admitting a new user, in accordance with one embodiment of the present invention. At step 101, as each new user (or a group of users) joins the network, the method first collects information regarding possible admission decisions and routes that are to be, or can be, considered for admitting the user into the system. At step 102, the method collects the existing state of the network (e.g., the aggregation levels, the bursting levels, the SNRs, and the number and types of users in each communication group). The admission control algorithm includes primarily the following steps: (i) decompose or note the map of the multi-hop network into communication groups; (ii) compute the load in each communication group for different possible route and traffic-shaping options (steps 103-105); (iii) check the stability condition for each communication group for every possible scenario (step 106); (iv) consider further load attributes if stability is achieved for more than one

scenario (step 107); (iv) admit or reject each user to the network, based on the best solution.

In Figure 1, the search for stable scenarios is conducted by *theoretically* computing the load vector (vector consisting of load values for each communication group) for each admission and routing scenario. According to one embodiment of the present invention, for each aggregation level, bursting level, transmission rate, and number of users or access points in a communication group, a channel occupancy time is computed for that communication group. Each scenario takes into account the constraints of the multi-hop network. The constraints may include the numbers of access points, users, and current connections, and the sensing range of each user. Setting of proper constraints generally limits the search space. In one embodiment where the users may have several access points in their sensing range and the network allows re-association of the users (without violating the stability condition), then scenarios where users who are already admitted into the network are re-associated to different access points may be investigated.

For each possible admission option, the state of the system may change as result of a new user's admission. In fact, multiple changes may result from each admission option, and multiple admission options exist. For example, the number of users and the types of users in each communication group may change, and a change in one communication group typically leads to a change in another communication group. Such changes follow directly from the fact that the traffic for a new user is carried by the assigned route or routes to and from one of the GAPs, the possible routes based on the admission options considered or selected.

Further, and of significance, a method of the present invention determines how each change in a communication group (e.g., whether directed centrally or as a result of self-adaptation) changes traffic-shaping settings (e.g., aggregation, bursting levels and PHY rate on its links). Traffic shaping has a strong influence on system performance, and admission decisions affect traffic-shaping. The method may have options to re-associate existing users (i.e., changing the APs with which the user are associated and their routes to the GAP) as part of the admission consideration and the change in system state.

For each potential admission option being considered, and the associated joint route and traffic-shaping changes implicit in selecting that option, a new "load" may be computed for each communication group to provide a new network state. These potential "load" measurements can then be used for making admission control decisions and to select between routing options and traffic-shaping options, where many options exist for a given admission option. The computed "load" indicates a relative (fractional) or weighted channel occupancy time for each communication group. For the entire network, the "system load" for a given option may be expressed as a vector of load values, with each value in the vector corresponding to the load for a communication group.

The new "system load" for a given admission, routing and traffic-shaping option may in fact contain load values that show instability in one or more of the communication groups (i.e., the communication groups cannot support the admission option). For example, the load on one or more communication groups exceeds a pre-determined limit for acceptable operation (e.g., a load which exceeds the wireless medium's ability to deliver an acceptable QoS in delay or packet loss). Alternatively, a number of potential admission and routing options may allow all communication groups to operate acceptably. In that situation, the final admission and routing decision among these possibilities may be made based on other criteria, such as the least number of hops, the highest signal-to-noise ratio, minimization of the maximum load over all groups, and load balancing, given the current constraints of the network.

As discussed above, loads are calculated for each communication group. In calculating such loads it is useful to note that there are different types of load components. Some components are incurred per packet transmission, such as 802.11 PHY and MAC headers, while others are incurred per transmission opportunity, such as contention overheads. Still other load components are invariant to how many packets are transmitted, or how many transmission opportunities are used, such as the actual voice (or media) data being transmitted. The former two types of loads are intimately connected with traffic-shaping, since the loads can be amortized, using the techniques of aggregation and bursting, over multiple voice-packets.

Many main system state parameters may be used to calculate a system load value that can then be used in making admission, routing and traffic-shaping decisions. Details of two such calculations can be found in LKRP06 and RLKP07 (attached as Appendices), with LKRP06 providing a simple method, and RLKP07 providing a more detailed method. For example, each inbound or outbound flow "*i*" in a communication group, one can consider traffic-shaping variables (a) aggregation level *A(i)* or *A(i,j)*, (b) bursting level *B(i)*, and (c) physical layer data rate *PHY(i)*. Many possible combinations of such parameters may be selected. A particular combination may refer to a particular admission and routing option.

Aggregation level *A(i)* refers to the number of packets (e.g., voice packets) that are aggregated into a single 802.11-packet in flow *i*. Bursting level *B(i)* refers to the number of 802.11-packets that are transmitted per transmission opportunity in flow *i*. If different aggregation levels are used for different packets in a burst one can use the notation "*A(i,j)*" as described above to refer to the aggregation level of the "j-th" packet in a burst. *PHY(i)* refers to physical layer data rate that is used to transmit the *i*-th flow. Under the 802.11 standard, *PHY(i)* takes values of 1, 2, 11, ..., 54 mega-bits-sec (Mbs). To illustrate the use of both aggregation and bursting, Figure 4 shows a flow in which *A(i)* and *B(i)* have values 3

and 2 respectively (here $A(i,1)=A(i,2)=3$). Figure 4 also assumes the aggregation includes exactly one voice packet from each user in the communication group. In practice, $A(i)$ is more realistically the effective aggregation level (i.e., some 802.11-packets may contain two or more voice data packets from a single user; but each user provides on the average one voice packet per 802.11-packet).

In this detailed discussion, aggregate level $A(i)$ and bursting level $B(i)$ are assumed the same at every transmission opportunity. Although other values of aggregation level $A(i)$ and bursting value $B(i)$ are possible. $B(i)$ need not be the same at every transmission opportunity. $B(i)$ may be, for example, an average value over many bursts.

Each admission and routing decision includes an assumption regarding a level of data packet traffic that results on each link in each communication group per packetization interval. For example, for VoIP traffic, a 2-way constant bit-rate (or constant packet rate) traffic may be assumed. If one voice packet is generated per call per direction within each frame-interval ("voice-packet-interval"), the number of voice packets transported within that communication group is given by 2 × the number of users using links in the communication group. It may also be 4 × the number of users using links in the communication group, at a RAP, if inbound and outbound likes use the same group, i.e. where the voice packet appears in both inbound and outbound links in the same wireless interface using the same channel. Of course, the assumption single voice packet per voice-packet-interval per direction is invalid for a variable bit-rate service, or where a single user may use multiple routes. Compensation for the incorrect assumption can be provided in the calculation as needed and described in RLKP07.

In all cases, these models suggest that a load to be borne by the communication group may be associated with each admission and routing decision based on the number of packets (per packetization interval) supported by flows in the communication group. This number defines the possible aggregation and bursting parameter options "*A(i)*" (or "*A(i,j)*") and "*B(i)*" that may be considered for flow "i" in the group. Such parameters may also be implicit on the operation of the nodes in response to the new traffic it sees, as discussed above. An option may consider a joint selection of admission, routing and set of parameter choices for all flows in the communication group. A load for the group can then be calculated. In the following, the parameters $A(i)$, $B(i)$ and $PHY(i)$ are used to illustrate calculating a channel occupancy required to support that load. This calculation is done for all groups affected by the admission and routing possibility (option) being tested. LKRP06 and RLKP07 provide in greater detail the calculation described below.

To do the calculation other application- or scenario-specific parameters may be used. These parameters may include: (i) the number of bits $V(i)$ per voice packet; (ii) packet error

rate *P(i)*; and (iii) a statistical description of the probability of successful transmisssions *p(i, j,m)*. *V(i)* refers to the number of bits per voice packet within the *i*-th flow. As with *A(i)* and *B(i)*, *V(i)* may be provided the same fixed value for each user to simplify calculation of the load, although the analysis can be extended to cases where different users generates different bits per voice packet. *V(i)* may be used to model not only speech data generated by the speech or audio encoder, but also other overhead quantities such as (a) Internet Protocol (IP) header overheads; (b) Real Time Transport (RTP) overheads; User Datagram Protocol (UDP) header overheads, and (d) any other overhead (on average or amortized over users) within the payload of each 802.11 packet. Packet error rate *P(i)* refers the packet error rate of each transmitted 802.11 packet in the *i*-th flow, which is a function of *V(i)*, *PHY(i)*, and *A(i)*. See, e.g., the "LKRP06" and "RLKP07" article mentioned above. To transmit a burst of packets may take multiple transmission attempts because of collisions and errors on the wireless medium, which may be accounted for in the calculation. Transmission probability *p(i,j,m)* refers to the probability that it takes "*m*" transmit opportunities to transfer the "*j*" packet in a burst for flow i.

In one embodiment of the present invention, the load (i.e., "channel occupancy time") for each packetization interval consists of three additive parameters: (a) successful transmission time T_s; (b) unsuccessful transmission time T_f; and (c) the time when nodes are not transmitting but are in contention for the wireless medium, known as the back-off time T_b. Successful transmission time T_s refers to the time spent successfully transmitting packets. Unsuccessful transmission time T_f refers to the time wasted due to transmission failures. Back-off time T_b refers to the time spent in back-off mode by each MAC in the communication group.

Successful transmission time T_s includes 3 components: (a) the time spent transmitting voice (or another application payload) bits on each link, (b) the time spent transmitting an acknowledgement (ACK) packet, and (c) additional overhead, such as interframe (IF) spacing. The time spent transmitting an application payload (e.g., voice) is a parameter invariant to the number of successful voice-packets or successful transmission opportunities used to transmit the data. Assuming *V(i)*, *A(i)*, and *B(i)* are fixed and the same for each user on link *i*, $V(i) \times A(i) \times B(i)$ bits are transmitted during each packet-interval (e.g., one voice-packet per user) over the link. The time required to transmit these bits, assuming the same physical layer data rate *PHY(i)* for each user, is simply $V(i) \times A(i) \times B(i)/PHY(i)$ in microseconds, if *PHY(i)* is expressed in mega-bits/second.

The time spent transmitting an ACK packet acknowledging is spent following successful transmissions. In theory, one or more ACK packets may be sent per transmission opportunity to cover all packets burst transmitted during that transmission opportunity. For

example, if there are no hidden terminals (i.e. all nodes in a group can sense transmissions from all other nodes) and if there are no other channel impairments resulting in bit or symbol losses, it is shown in LKRP06 and RLKP07 that if the first packet in a burst transmission is successful, then all other packets in the burst are successful. Therefore, for a set of *B(i)* burst packets on ACK would be used to acknowledge the burst. Assuming that the *B(i)* burst of 802.11-packets are all part of the same transmission opportunity, one ACK packet is provided for transmission of $V(i) \times A(i) \times B(i)$ bits. Traffic- shaping via both aggregation (*A(i)*) and burst processing *B(i)* amortizes the overhead over multiple voice packets.

Alternatively, one ACK packet may be sent for each 802.11-packet transmitted in a burst. There are two cases in which this can happen. The first case is when the system choses to do so (this practice may not conform to the 802.11 standard). The second case is when there are channel impairments such as hidden terminals or bit or symbol losses due to noise on the wireless channel. In the second case individual packets in a burst can be lost, terminating the burst transmission and resulting in retransmissions of the lost packets as described in RLKP07. This is where the parameter "*p(i,j,m)*" helps to describe the average number of ACK packets that are transmitted. For greater detail, see, the RLKP07 article. Also see the 802.11e/D13.0 standard[1], and the 802.11a standard[2] for greater details on the ACK mechanism. Note, in all cases the aggregation mechanism helps to amortize ACK overheads from multiple speech packets by creating a single ACK for an aggregated transmission. Bursting can further amortize ACKs over multiple aggregated packets, depending on the ACK mechanism and channel impairments.

Additional overheads, such as inter-frame spacings (IFS) (e.g. Short IFS (SIFS), the Preamble (PLCP), PHY header, and MAC header, are seen per 802.11- transmitted packet, while others such as some of the Distributed IFS (DIFS)) spacing occur once per transmission opportunity. As with ACKs, these overheads are amortized over many voice packets by the process of aggregation and bursting leading to increased efficiency by these processes. The exact values of these times are defined in the standard. See, e.g., the articles LKRP06 and RLKP07 , the 802.11e/D13.0 standard, and the 802.11a standard.

---

[1] IEEE P802.11e/D13.0, Part II Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: Amendment Medium Access Control (MAC) Quality of Service (QoS) Enhancements, IEEE, Jan 2005

[2] IEEE Computer Society, "Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: Higher-speed Physical Layer (PHY) extension in the 5 GHz band}," IEEE Std. 802.11a-1999 R2003.

Time wasted due to transmission failures (i.e., time T_f) can result from collisions in the transmission medium, or errors in transmitted information due to channel impairments. Failures can happen at each transmission opportunity and for each packet transferred within a burst of transmission. Depending on the channel impairments, aggregation level $A(i)$ can affect the packet loss rate. T_f depends on the number (or expected number) of transmission failures per transmitted burst (i.e., on transmission probability $p(i,j,m)$). See, e.g., the articles LKRP06 and RLKP07 for a description of how to calculate time T_f, and the 802.11e/D13.0 standard and the 802.11a standard, for the overheads relevant to each failure.

In calculating time T_f per transmission attempt, the main time not included is time for ACK packets that are missing from 802.11-packets, or which are not correctly received. Since transmissions lost due to collisions and failures are often a statistical process, time T_f is often given as an expected value, and not an deterministic value as in time T_s. See, for example, the LKPR06 and RLKP07 articles. Other ways for calculating a value representative of time T_f includes the distribution of the T_f value, or the probability that time T_f exceeds a predetermined time value.

Time T_b spent in back-off mode by each MAC in the communication group represents the time wasted while the channel is idle, or when MAC mechanisms in the system are in a random back-off state, contending for access to the channel. Time T_b is a central part of the Distributed Coordination Function (DCF) mode under the 802.11 standard, which is based on a Carrier Sense Multiple Access (CSMA) scheme. See, e.g., the descriptions in the 802.11e/D13.0 standard and the 802.11a standard.

The total system load is then estimated by taking into account all links and users, and adding all the individual contributions to the values T_s, T_f and T_b. Once the totals are known over all traffic within the communication group, the load for the communication group is given by the sum:

$$load = T_s + T_f + T_b$$

This time-based load may be converted to a relative value. For example, in an VoIP application, when the voice-packet interval is D_s, and the load is the time spent transmitting all voice packets generated by users, or transmitted in flows during that interval, a "relative load" may be used, given by:

$$relative\ load = (T_s + T_f + T_b)/D_s$$

For stability, the relative load for all communication groups may be either *strictly* less than, or at times sufficiently less than or equal, to 1. The time load and the relative load values given above for each communication group form a vector of load values for a

admission option. If stability is achieved with more than one option, the relative load of each communication group for each option can be used to compare options. For example, the maximum load or the average load estimated across communication groups for each option can be used to compare options. Comparison of options can be refined using other metrics, such as the number of hops of each flow or delays.

The above detailed description is provided to illustrate specific embodiments of the present invention and is not intended to be limiting. Numerous modifications and variations within the scope of the present invention are possible. The present invention is set forth in the accompanying claims.

Appendix A

Analyzing and Managing Traffic-Shaping in the Transmission of Voice over Multi-Hop 802.11 Networks

Danjue Li Ulaş C. Kozat Sean A. Ramprashad Christine Pépin
DoCoMo USA Labs
3240 Hillview Avenue, Palo Alto, CA, 94304
dli@ucdavis.edu, {kozat,ramprashad,pepin}@docomolabs-usa.com

Abstract

Due to low cost and easiness of deployment, multi-hop 802.11 networks have become an attractive solution for providing *last-mile* broadband (wireless) access in urban environments. To support real-time streaming applications such as Voice-over-IP (VoIP) over such networks, one of the main challenges is how to address the inefficiencies (i.e., overheads) of the Distributed Coordination Function (DCF) used in the MAC layer to access the wireless medium and the associated PHY overheads. In this paper, we build upon our previous work on aggregation-based traffic shaping to design a more comprehensive scheme that can perform network-aware joint aggregation and bursting as well as PHY-rate adaptation in presence of channel errors. Through extensive simulations, we demonstrate that more than 3-fold call-capacity improvements can be gained over conventional transport methods. Meanwhile, we also develop a theoretical framework to predict the voice capacity of a system which implements our proposed scheme, and provide certain guidelines on how to properly adapt traffic shaping settings and PHY-rates to optimize voice capacity under various wireless channel conditions. Finally, this framework reveals itself as a useful tool in planning and managing a multi-hop network where the *load* of the network (i.e., the channel occupancy time of its users) is the metric used in a centralized admission control strategy.

I. Introduction

We consider the problem of transmitting voice traffic efficiently over a multi-hop 802.11 network. Such networks present an attractive solution for providing *last-mile* broadband (wireless) access in many urban environments. This is true given the low cost of 802.11 end-points, the availability of existing

Danjue Li is an intern at DoCoMo USA Labs working towards her Ph.D. at University of California Davis. The other authors are employees of DoCoMo USA Labs

power sources (lamp posts, etc.), the fact that 802.11 is a widely used standard-based solution, and the possibility of using only a few gateways to provide (wired) connectivity outside of the net. However, such networks do suffer from similar inefficiency problems as single-hop systems due to inherent overheads in accessing and transmitting on the wireless medium [1] using the Distributed Coordination Function (DCF). In multi-hop scenarios, such inefficiencies can propagate with flows across the network resulting in bottlenecks and even larger losses in efficiency relative to single-hop scenarios.

Specifically, while the simplicity of the 802.11 MAC and PHY designs are clearly attractive and successful in flexibly accommodating a range of traffic types and applications, the inefficiencies they have can be severe for important applications such as Voice-over-IP (VoIP). Fundamentally, the use of small payload sizes, constant bit-rate traffic and stringent delay requirements are not well matched to the underlying 802.11 access and transmission mechanisms.

In our previous paper [1], we proposed to use aggregation-based traffic shaping along with routing to improve the efficiency of multi-hop voice transmission, and demonstrated its efficacy via ns2 [2] simulations. The basic premise is that as traffic moves toward a (wired) gateway the intermediate (wireless) "relay" nodes can be intelligent by modifying the traffic characteristics in terms of timing, overheads, etc., to make subsequent links (MAC/PHY's) in the path more efficient.

Based on the same premise, in this paper we extend the analysis by jointly considering aggregation and bursting strategies, and propose a theoretical framework which can be used to analyze and manage such networks. Specifically, we bound voice capacity for different aggregation, bursting and/or PHY-layer data rate settings, looking in details at the contention ("α") and header ("β") overheads defined in [1]. The theory also enables us to consider more broadly "unbalanced" [1] scenarios, i.e., scenarios where different relays use different settings and/or support an unequal number of clients. This helps us to address many more scenarios, and even $n > 2$ hop scenarios, which can not be easily characterized through ns2 simulations [1].

Another contribution we make in this work is addressing the issue of managing the system in the presence of channel impairments. Such an issue is particularly important when transmitting large packets, such as those generated by relays that aggregate many smaller packets (flows) into common larger packets (flows) [1]. Using our proposed theoretical framework, we look at how other options, such as bursting smaller (possibly aggregated) packets, perform relative to aggregation-only in such error cases. Even more importantly, we consider how one should adapt the underlying PHY transmission rate jointly with traffic shaping and routing.

The rest of the paper is organized as follows. In Section II, we discuss the inefficiency problem of

the current DCF scheme and provide a survey on related works. Then we present our proposed *joint Traffic shaping, rate Adaptation and Routing (TAR)* mechanism for improving the efficiency of 802.11 multi-hop networks in Section III. In Section IV, we provide a theoretical framework to analyze the voice capacity of a multi-hop network without channel impairments, and then in Section V, we use this theory to explore adaptive TAR configurations as a function of channel impairments. Here the relative benefits of both aggregation and bursting, i.e., relative decrease in overheads "α" and "β", can clearly be seen. In Section VI we compare the theory with ns simulations, and look at how and why they differ. Given the results in Section V and Section VI, we discuss in Section VII how to manage such networks in a practical setting, where wireless channel quality and voice traffic patterns change constantly or are uncertain. Issues of network planning and admission control are all jointly discussed in this section. Finally, in Section VIII, we summarize the work and discuss some open problems.

II. Background and Related Work

In 802.11-based VoIP, multiple users with small payload sizes compete for the shared wireless medium. This results in an inefficient use of the resources when the DCF access scheme is used at the Medium Access Control (MAC) layer. In this section, we examine the reasons of such inefficiencies and give a brief survey on how this issue has been addressed in the literature.

A. Inefficiencies of the DCF Scheme

The inefficiencies of the DCF scheme can be represented by two factors "α" and "β". The α factor refers to the impact of the overheads incurred by the idle periods used for contention resolution and the wasted time slots due to packet collisions [1]. These types of overheads are inherent to the distributed nature of the DCF access mechanism. We denote α, $0 \leq \alpha \leq 1$, as the ratio of such overheads to the raw bandwidth of the wireless medium. If the bandwidth is W, then $(1-\alpha)W$ is the effective bandwidth that can be used for successful transmissions.

Out of the remaining bandwidth, there are additional overheads due to PHY/MAC headers, frame separations and control messages generated during a transmission. Let β, $0 \leq \beta \leq 1$, denote the ratio of those overheads over the remaining bandwidth $(1-\alpha)W$. Then $(1-\beta)(1-\alpha)W$ is the effective bandwidth that is used for actual data transmissions, as illustrated in Figure 1. Experimental results [3] have shown that when using short packets, such as those generated by the ITU-T Rec. G.711 [4] with 10 *msec* duration, the joint effect of α and β can lower the effective bandwidth utilization to about 10% of the total resource.

It is worth noting that the β factor has one unique quality that will be important to consider later. Unlike the α factor that changes monotonically with many parameters, the β factor is affected by the maximum transfer unit (MTU) of the MAC layer. When a packet exceeds the MTU size its payload is fragmented into multiple packets. All fragments are sent as consecutive transmissions in a single transmission opportunity. Therefore at each fragmentation boundary additional β-related overheads are added and the β factor has discontinuous jumps. See illustration in Figure 1 [1].

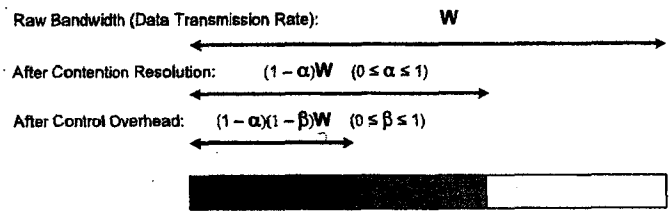


Fig. 1. Inefficiency of bandwidth utilization for DCF

B. Related Work

To address the aforementioned bandwidth inefficiency problems, 802.11e [5] extends the original 802.11 standard to provide additional mechanisms for bursting multiple packets in the same transmission opportunity to reduce the α factor. In 802.11n [6] [7], in addition to bursting, different types of aggregation are used at the MAC layer as well as the PHY layer to further reduce both α and β factors. In a similar spirit, Wang *et al.* [8] proposed to aggregate packets from multiple users at an access point (AP) and use a multicast transmission to better utilize transmission opportunities in the downlink direction. Header compression and silent suppression have also been investigated in the past for voice traffic to improve system capacity [3]. Although they focus on reducing the amount of data to be transmitted instead of improving the efficiency of the DCF mechanism, silence suppression in particular does have the side effect of affecting α overheads by reducing contention.

All of the above works consider one-hop WLAN as the underlying networking environment. In the setting of 802.11 multi-hop networks, solutions have been focusing on distributing the load over different wireless channels or routes [9] [10] [11]. Unlike those works, we explicitly utilize routing to leverage the capability of traffic shaping via aggregation and/or bursting in subsequent hops of a multi-hop network. This in fact balances the advantages of concentrating traffic on fewer routes and distributing the load over multiple routes [1].

We extend this concept further in this paper, and focus on analytically modeling voice capacity for different aggregation and bursting settings, as well as considering the issue of PHY data rate settings. This latter issue is of particular importance when considering the effects of (re)transmission efforts. Although there are some previous studies investigating a voice capacity model for IEEE802.11 WLANs [12] [13] [14], most of these works assume identical user data rates and error-free channels. Although Medepalli et. al [15] took channel errors and different user data rates into consideration, and quantified their impact on system capacity in a single hop scenario, they did not consider the joint relationship among traffic shaping, PHY-rate adaptation and routing in a multi-hop setting. To the best of our knowledge, our work is the first attempt to quantify the joint impact of different traffic shaping strategies and PHY-rate adaptations on voice capacity, and provide a systematic way to use the analytical results in the design and management of multi-hop 802.11 networks.

III. TAR: Enhancing Voice Capacity of Multi-Hop 802.11 Networks via Joint Traffic Shaping, Rate Adaptation and Routing

A. The Method

To improve the efficiency of the DCF scheme, we propose to implement a *joint **Traffic shaping, rate Adaptation and Routing*** (TAR) mechanism at intermediate relay access points (RAPs) and the gateway access point (GAP), in a network as shown in Figure 2. Specifically, each access point aggregates (i.e., multiplexes) arriving flows and transmits these aggregated packets as bursts within the same transmission opportunity, at a transmission rate adapted to the noise in the wireless channel. Adequate routing allows this traffic shaping to take place. By performing routing and traffic shaping decisions jointly one can improve the efficiency of 802.11 multi-hop networks.

Different from our previous work [1] which relies on performing only aggregation/de-aggregation or bursting, we here consider aggregation and bursting together. Furthermore, we adaptively select the aggregation and/or bursting level together with the transmission (PHY) rates based on the traffic pattern and wireless medium state. The rational behind such a joint consideration is that aggregation and bursting impact the system in different ways, and each has its own advantages and disadvantages.

Aggregation can efficiently suppress both α and β factors by reducing the number of contending flows and sharing the MAC/PHY headers, frame separations and control overheads. Bursting focuses on suppressing α overheads only and is thus less effective than aggregation. As shown in a previous paper [1] using a larger aggregation level can potentially increase voice capacity. However, this is only true when the impact of channel impairments, such as bit error, is negligible. If packet losses due

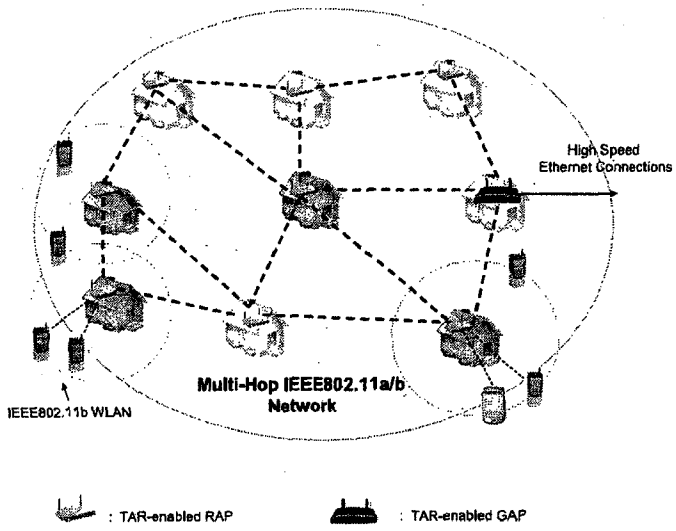


Fig. 2. Supporting VoIP over Multi-hop IEEE802.11a/b Networks

to bit/symbol/packet-errors are non-negligible, aggregation might lose its advantage over bursting. For example, for independent and identically distributed (i.i.d.) bit/symbol-errors (and a given error rate) packets generated at a large aggregation level suffer from higher packet loss rates than shorter packets. When the packet loss rate becomes large enough, the overhead created by retransmitting long erroneous packets can exceed that saved by reducing α and β factors per each transmission. In this case, bursting smaller (possibly aggregated) packets may be a better choice. Therefore, by combining aggregation and bursting together one can try to balance the two strategies, achieving better system performance. Furthermore, parameters such as the bit/symbol/packet loss rate depend on the underlying PHY rate. When traffic shaping loses its advantages because of high packet loss rates, one may try to balance this loss with the PHY rate adaptation options available in 802.11.

B. System Model and Assumptions

Consider a tree-rooted two-hop 802.11 network model as in Figure 3. The model encompasses the following components: (1) wireless VoIP clients that act as both the source and sink of VoIP flows; (2) Wireless TAR-enabled RAPs that can forward and process many VoIP flows; (3) A TAR-enabled GAP through which VoIP sessions are established with hosts outside the network. Although we focus on two-hop networks for the sake of clarity, our analysis can easily be extended to more than two hops. We assume that all voice traffic is generated at a data rate of 64 *kbps*, according to the ITU-T Rec. G.

SLOT δ	SIFS	DIFS	T_{PHY}	MAC	FCS	IP	UDP	RTP	ACK	Min. Rate	MTU
$20\mu s$	$10\mu s$	$50\mu s$	$192\mu s$ (L) $96\mu s$ (S)	30Byte	4Byte	8Byte	20Byte	12Byte	14Byte	1Mbps	2346Byte

TABLE I

IMPORTANT CONSTANTS FOR IEEE802.11X(X=A;B)

711 [4], and is packetized into equally sized packets at fixed intervals.

As illustrated in Figure 3, multiple communication groups exist: between a RAP and the associated VoIP clients, and between the GAP and the RAPs. Each RAP/GAP is equipped with multiple wireless interface cards, one for each communication group that the node is participating in. Each WNIC can operate simultaneously in parallel, i.e. receiving on one interface and transmitting on the other one at the same time. To avoid interference, we require each communication group operate orthogonally to each other.

In order to develop a theoretical framework for a TAR-enabled VoIP system, we make the following assumptions:

- each voice call includes one uplink flow (i.e. flow running from a RAP to the GAP) and one downlink flow (i.e. flow running from the GAP to a RAP), and these two flows start close in time. Uplink flows and downlink flows are shaped independently.
- all the stations connected to the same RAP are within carrier sensing range of each other so that whenever a node is transmitting, all other nodes can detect its transmission, i.e. there are no hidden terminals. Basic access without RTS/CTS assistance is considered in the analysis, however we do provide discussion on how RTS/CTS affects voice capacity later in Section VII.
- for clients and access points, their attempts to access the channels are mutually independent. RAPs will take advantage of the idle time required by the GAP to fulfill all their backoff requirements to minimize the medium idle time of the system.

IV. VOICE CAPACITY WITHOUT CHANNEL IMPAIRMENTS

In this section, we present a theoretical framework to estimate channel occupancy created by different aggregation and bursting options. This will enable us to explore the relative balance of different options, including their effective α and β factors. Extending the results in [1], it will also help us to estimate voice capacity of a TAR-enabled wireless network where the impact of channel impairments is negligible.

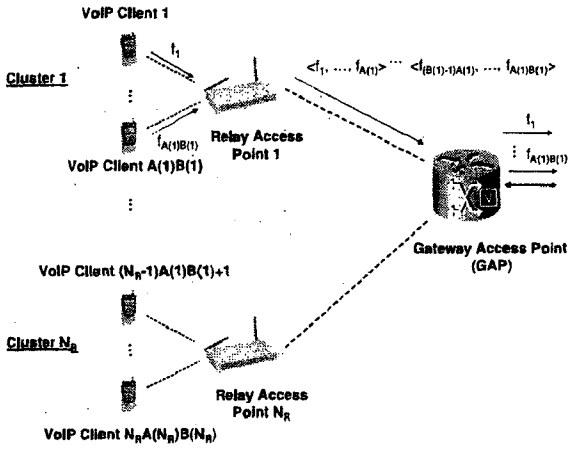


Fig. 3. System Model for A Two-hop IEEE802.11x-based Wireless Network: $< f_1, \ldots f_{A(i)} > \ldots < f_{(B(i)-1)A(i)+1}, \ldots f_{(B(i)A(i))} >$ refers to a traffic shaping profile which enables node i to first aggregate $A(i)$ flows together and then burst out $B(i)$ of the aggregated flows. Joint traffic shaping is applied to both uplink and downlink traffic

A. Channel Occupancy Time

Table I summarizes the important constants for an IEEE802.11b system. Although different 802.11 standards differ from each other in terms of transmission overheads, without losing generality, hereinafter, we consider IEEE802.11b in our analysis. Let D_s be the packetization period, i.e. the time period for generating voice packets, of the voice codec.

The main idea underlying the analysis is to count the number of packet exchanges that can be held in D_s by calculating average wireless resources, i.e. long term average of channel occupancy time T required to support one packet exchange per user. This approach has been originally proposed by Hole et. al [14] to predict voice capacity of WLANs. We extend such analysis by explicitly including consideration for traffic shaping strategies, channel impairments and different PHY transmission rates. For a TAR-enabled system, the "exchanges" mentioned above are a pair of bursts of aggregated voice packets as shown in Figure 4, one burst for uplink traffic and one burst for downlink traffic.

When the impact of channel errors is negligible, we are only concerned with the impact of packet collisions on voice capacity. In this case, T mainly includes the time of successfully transmitting the burst of aggregated packets, T_s, the time wasted due to collision-caused transmission failures, T_f, and

associated backoff time T_b, i.e.

$$T = E(T_s + T_f + T_b) = E(T_s) + E(T_f) + E(T_b) \quad (1)$$

For a given node i, $i \in \{1, ..., N_R\}$, T is a function of its TAR parameters, i.e. $T = T(R(i), A(i), B(i))$, where $R(i)$, $A(i)$, and $B(i)$ denote PHY transmission rate, aggregation level, and bursting level, respectively. RAPs and GAP can be configured using same TAR parameters in a *balanced case* or different ones in an *unbalanced case.*

In (1), T_s consists of the transmission time of voice packets T_v, the transmission time of ACK message T_{ack}, and any necessary inter-frame spacings (IFS). Since packets can be dropped due to collisions, the long term average of T_s depends on the probability of having a successful transmission. Let m denote the number of undergone collisions/retransmissions before a success as m, and m is a geometric random variable with probability mass function: $p(m) = (1 - P_C)P_C^m$, where P_C is the collision probability. Note that when channel errors are not negligible, we need jointly consider channel errors and collision loss together to estimate $p(m)$. By assuming the maximum number of retransmissions is M_{max}, we can estimate the probability having a successful transmission within M_{max} retransmission attempts as $P(I = 0) = \sum_{m=0}^{M_{max}} p(m)$, where $I = 0$ when a successful transmission happens within M_{max} retries and $I = 1$, otherwise. Based on Figure 4(c), T_s could be expressed as:

$$\begin{aligned} E(T_s) &= E(T_s|I=0)P(I=0) + E(T_s|I=1)P(I=1) \\ &= 2(DIFS+B(i)(T_v+2SIFS+T_{ack})-SIFS)\,P(I=0) \\ &= 2(DIFS+B(i)(T_v+2SIFS+T_{ack})-SIFS)\sum_{m=0}^{M_{max}^x} p(m) \end{aligned} \quad (2)$$

where $T_v = T_{header} + T_{data}$, and T_{header}, T_{data}, and T_{ack} are defined by:

$$T_{header} = T_{PHY} + \frac{(MAC + FCS) \cdot 8}{R(i)}, \quad (3)$$

$$T_{data} = \frac{A(i) \cdot (D_s R_s + (RTP + UDP + IP) \cdot 8)}{R(i)}, \quad (4)$$

$$T_{ack} = T_{PHY} + \frac{8 \cdot ACK}{R_p} \quad (5)$$

In (5), R_p is the minimum data rate of IEEE802.11b. Note that in above equations, we can set $B(i) = 1$ to obtain the analysis for aggregation-only scenario (as shown in Figure 4(a)), and $A(i) = 1$ for bursting-only scenario (as shown in Figure 4(b)), respectively.

Since there is no hidden terminal in the system, collision can only happen to the first aggregated packet in a burst, as illustrated by Figure 5(a). Depending on how many collisions the first packet undergoes

before being successfully transmitted, the wasted channel time, T_f, can be estimated as:

$$\begin{aligned} E(T_f) &= E(T_f|I=0)P(I=0) + E(T_f|I=1)P(I=1) \\ &= \frac{\sum_{m=0}^{M_{max}} p(m) m \tau_f}{\sum_{m=0}^{M_{max}} p(m)} \sum_{m=0}^{M_{max}} p(m) + M_{max} + 1)\tau_f (1 - \sum_{m=0}^{M_{max}^x} p(m)) \\ &= \left(\sum_{m=0}^{M_{max}} p(m) m + (M_{max}+1) \sum_{m=M_{max}+1}^{\infty} p(m) \right) \tau_f \end{aligned} \quad (6)$$

where τ_f is time wasted when one transmission attempt fails, i.e., $\tau_f = 2(DIFS + T_v + T_{ATO})$. Note that 2 here comes from counting in both uplink and downlink traffic. Here, T_{ATO} denotes ACKTimeout and we set it to be a conservative value of $SIFS$ plus ACK transmission time at the minimum transmission rate plus a slot time δ, i.e. $T_{ATO} = T_{ack} + SIFS + \delta$ [16]. To obtain P_C, we can either use a bi-dimensional Markov chain model [17] or approximate it using $1/(CWmin + 1)$ [15]. Considering that voice calls normally start randomly in time, generating traffic only every D_s, collisions mainly happen between GAP and RAPs with probability $1/(CWmin + 1)$. Figure 6 provides some sensitivity check on P_C approximation. From it, we can see that voice capacity is not very sensitive to small changes of P_C and quite flat around $1/(CWmin+1)$. Given that P_C measured through ns simulations to be around 10% and voice capacity to be 40, 40, 42 for those cases in Figure 6, respectively, having an approximation to be $1/(CWmin + 1)$ is therefore accurate enough for our analysis.

Let U_r be the length of r^{th} backoff in unit of time slots, and then the cumulative time spent on the backoff process can be expressed as: $T_b = \sum_{r=0}^{m} U_r$. To estimate $E(T_b)$, let us denote the maximum collision window size at r^{th} transmission attempt as W_r. Following the binary exponential backoff procedure of DCF, we have $W_r = \min\{CWmax, 2^r(CWmin + 1) - 1\}, r = 0, 1, ..., M_{max}$. Thus,

$$\begin{aligned} E(T_b) &= E(\sum_{r=0}^{m} U_r|I=0)P(I=0) + E(\sum_{r=0}^{m} U_r|I=1)P(I=1) \\ &= \sum_{m=0}^{M_{max}} p(m) \sum_{r=0}^{m} \frac{W_r}{2}\delta + \sum_{m=M_{max}+1}^{\infty} p(m) \sum_{r=0}^{M_{max}} \frac{W_r}{2}\delta \end{aligned} \quad (7)$$

By substituting $E(T_s)$, $E(T_f)$ and $E(T_b)$ in (1) using (2), (6), and (7), we can obtain the long term average of channel occupancy time $T(R(i), A(i), B(i))$.

Assume that there is at least one bi-directional call existing between a RAP and the GAP, then for a network with N_R RAPs, the total channel occupancy time will be:

$$T_{WLAN} = \sum_{i=1}^{N_R} T(R(i), A(i), B(i)) \quad (8)$$

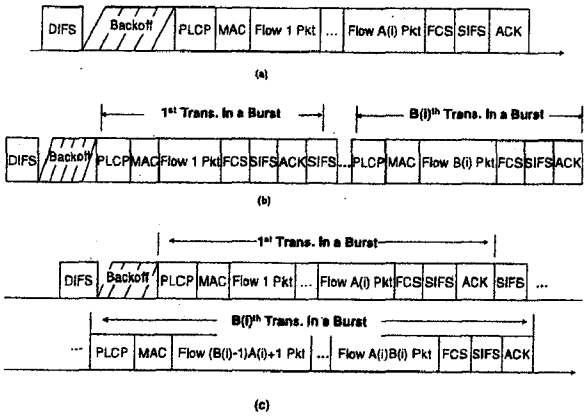


Fig. 4. Basic access of 802.11 DCF, (a) a successful transmission when APs can only perform aggregation, and aggregation level = A(i);(b) a successful transmission when APs can only perform bursting, and bursting level = B(i); (c) a successful transmission when APs are TAR-enabled using aggregation level = A(i), bursting level = B(i);

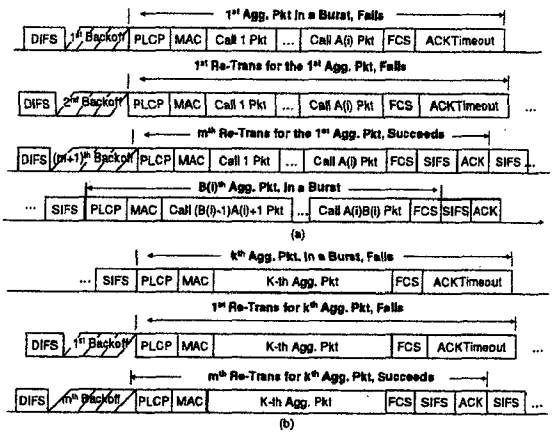


Fig. 5. Wasted time due to transmission failure: (a) when a packet exchange fails due to collision, failure can only happen to the first transmission in a burst;(b) when the packet exchanged fails due to channel errors, failure can happen any time during the bursting procedure and here shows k-th aggregated packet fails its first transmission during the bursting process and it is successfully transmitted after m retransmissions;

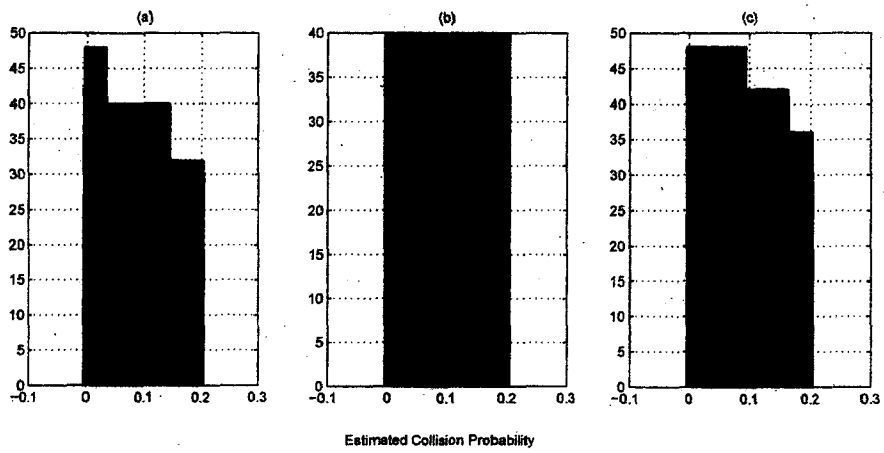


Fig. 6. Sensitivity Analysis on the Estimation Accuracy of Collision Probability: CWmin =1, (a) $N_R = 4, SNR = 16dB$ (b) $N_R = 5, SNR = 16dB$ (c)$N_R = 6, SNR = 16dB$

B. Voice Capacity Analysis

To support the offered load T_{WLAN} while maintaining a stable system, T_{WLAN} has to be less than assigned wireless resource D_s, i.e. $T_{WLAN} \leq D_s$. When this system stability criteria is satisfied, the number of voice calls that the system supports is: $\boldsymbol{C} = \sum_{i=1}^{N_R} A(i)B(i)$. Mathematically, this can be formulated as an optimization problem, i.e.:

$$\max_{A(i),B(i)\in Z^+} \sum_{i=1}^{N_R} A(i)B(i) \tag{9}$$

$$\text{S.T.} \quad A(i) \leq N_{max} = \min\{N_0, A_{max}\}, \forall i : 1 \leq i \leq N_R \tag{10}$$

$$A(i)B(i) \leq N_{max}, \forall i : 1 \leq i \leq N_R \tag{11}$$

$$E(T_s(R(i), A(i), B(i)) \leq TXOP(i), \forall i : 1 \leq i \leq N_R \tag{12}$$

$$\sum_{i=1}^{N_R} T(R(i), A(i), B(i)) \leq D_s \tag{13}$$

where N_0 is the maximum number of calls a single hop 802.11 network can support, and A_{max} is the maximum aggregation level without causing fragmentation. From the discussion in Section II, we have known that when the size of an aggregated packet reaches MTU limit, fragmentation takes place to increase transmission reliability, and all resulting fragments are transmitted sequentially using a bursting mechanism. As the result, the benefit of performing aggregation is compromised by extra overheads incurred by fragmentation. To avoid such a circumstance, we set the maximum non-fragmented aggregation level as A_{max}. Given MTU, D_s and R_s, A_{max} can be computed using $\left\lfloor \frac{8\,MTU}{D_s R_s + 8(IP+UDP+RTP)} \right\rfloor$.

To solve the optimization problem, we can either use a greedy algorithm to search the complete solution space or decouple it into a two-stage searching, i.e. obtaining a solution for the balanced case first and then searching around the obtained solution to choose the best TAR parameters for the general case. Let $\mathcal{A}$ and $\mathcal{B}$ denote the size of candidate sets for aggregation level and bursting level, respectively. Compared to a complete search, the two-stage searching can reduce the complexity from $O(\mathcal{A}^{N_R}\mathcal{B}^{N_R})$ to $O(\mathcal{AB})$.

C. Reduced Overheads

Based on the above analysis, we can also obtain the average α and β overheads for supporting voice calls over a TAR-enabled system in a channel-error negligible case as:

$$\alpha = \sum_{i=1}^{N_R} \frac{E(T_f) + E(T_b)}{T_{WLAN}} \tag{14}$$

$$\beta = \sum_{i=1}^{N_R} \frac{E(T_s) - 2B(i)T_v \sum_{m=0}^{M_{max}} p(m)}{N_R \cdot E(T_s)} \tag{15}$$

As the comparison, we also obtain the α and β overheads for a system which supports the same amount of voice calls without implementing aggregation and bursting strategies:

$$\alpha_0 = \frac{E(T_f') + E(T_b)}{E(T_f') + E(T_b) + E(T_s')} \tag{16}$$

$$\beta_0 = \frac{2(DIFS+T_{header}+SIFS+T_{ack})}{E(T_s')} \sum_{m=0}^{M_{max}} p(m) \tag{17}$$

where T_f' and T_s' are updated values for a TAR-disabled scenario, respectively, by setting $A(i)$ and $B(i)$ in (2) and (4) to be 1.

V. Voice Capacity with Channel Impairments

We now extend the above analysis to model the voice capacity when channel impairments such as bit errors are not negligible. To that end, we discuss how to model error performance of wireless channel first and then proceed to present details on voice capacity modeling.

A. Error Performance of IEEE802.11b PHY Modes

IEEE802.11b provides four different PHY modes at 2.4GHz with each corresponding to a different modulation scheme (Table II). At low data rate, i.e. 1Mbps and 2Mbps, a differentially encoded M-PSK (M=2, 4) is used to modulate the information while Complementary Code Keying(CCK) is adopted as the basis for the high rate physical layer extension to deliver data rates of 5Mbps and 11Mbps. To analyze the error performance of each PHY mode, we assume an additive white Gaussian noise (AWGN) channel in

Standard	Mode	Date Rate	Code Rate	Modulation
802.11b	1	1 MB	1	DBPSK
	2	2 Mbps	1	DQPSK
	3	5.5 Mbps	1	CCK
	4	11 Mbps	1	CCK

TABLE II

DIFFERENT PHY MODES OF IEEE802.11B

following discussions. We can establish the similar error performance analysis with other more realistic channel models as well, although the computations are much more complicated.

Symbol error rate (SER) for coherent detection of a differentially encoded M-PSK (M=2, 4) [18] is:

$$P_s(M) = \begin{cases} erfc(\sqrt{\frac{E_s}{N_0}})\left(1 - \frac{1}{2}\ erfc(\sqrt{\frac{E_s}{N_0}})\right), & \text{M=2} \\ 2\ erfc(\sqrt{\frac{E_s}{2N_0}}) - 2\ erfc^2(\sqrt{\frac{E_s}{2N_0}}) + erfc^3(\sqrt{\frac{E_s}{2N_0}}) - \frac{1}{4}\ erfc(\sqrt{\frac{E_s}{2N_0}}), & \text{M=4} \end{cases} \quad (18)$$

where $\frac{E_s}{N_0}$ is signal-to-noise ratio(SNR) of received signal, and $erfc(x)$ is the error function, i.e.:

$$erfc(x) = \frac{2}{\sqrt{\pi}} \int^{\infty} e^{-t^2} dt, x \in \Re \quad (19)$$

To compute the SER for higher rate, we decouple the modulation process into two sequential processes, CCK modulations for unrotated complex code selection and DQPSK based phase rotation, i.e.

$$P_s(M) = \begin{cases} P_{CCK(2)} + (1 - P_{CCK(2)}) \cdot P_s(4) & \text{M=5.5,} \\ P_{CCK(6)} + (1 - P_{CCK(6)}) \cdot P_s(4) & \text{M=11,} \end{cases} \quad (20)$$

where $P_{CCK(k)}, k = 2, 6$ is defined by:

$$P_{CCK(k)} \leq \sum_{i=0}^{2^k-1} \sum_{j=0, j\neq i}^{2^k-1} Q\left(\sqrt{\frac{2E_b}{N_0} r_c D_{ij}}\right) P(S_i), \quad (21)$$

where $Q(x) = \frac{1}{2} erfc\left(\frac{x}{\sqrt{2}}\right)$, D_{ij} is the *Hamming Distance* between a codeword i and a codeword j, and r_c is the code rate, which is 1 for modulation schemes used in IEEE802.11b. $P(S_i)$ is the probability that codeword S_i is transmitted and its long term average can be approximated by $P(S_i) = \frac{1}{2^k}$.

Based on above analysis, we can obtain the SER curves of IEEE802.11b as shown in Figure 7.

B. Channel Occupancy Time

When the impact of channel errors such as bit errors is not negligible, packets can be lost due to both collisions and channel errors. Different from collisions, channel error caused packet losses can happen to both data packets and ACK packets anytime during the transmission (as shown in Figure 5). Since

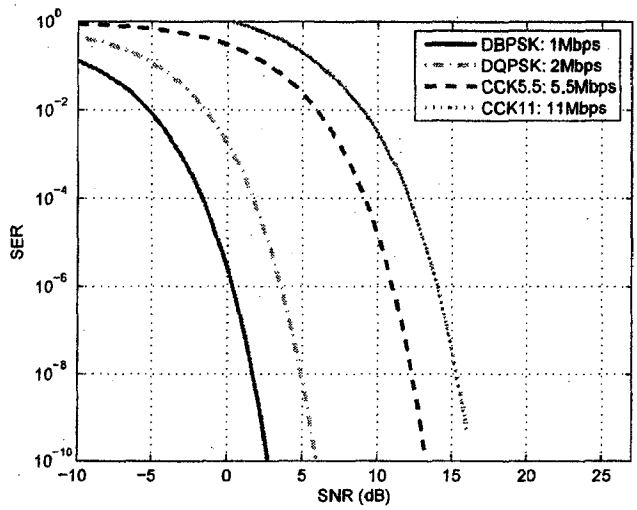


Fig. 7. SER Performance of different modulation schemes used in IEEE802.11b

ACK packet is small (14bytes), the probability of ACK packet error is quite low and can be neglected. Thus, by taking the union of the two events, channel errors and collisions, we can obtain the PLR as:

$$P = P_C + (1 - P_C) P_E = P_C + (1 - P_C)(1 - (1 - P_S)^L) \tag{22}$$

where P_S is channel SER defined by (18) or (20), depending on its modulation scheme, P_C is the collision probability defined in Section IV, and L is the packet length in byte including its MAC header, i.e. $L = MAC + A(i)(RTP + UDP + IP + \frac{D_s R_s}{8})$.

Let $T^e(R(i), A(i), B(i))$ be the long term average time of channel occupancy time of node i when packet transmissions can suffer from both collision and channel errors. $R(i)$, $A(i)$ and $B(i)$ are similarly defined as in Section IV, except that $R(i)$ is no longer fixed to be 11Mbps. Instead, it is a variable determined by measured channel quality. With those priori, T^e can be computed as:

$$T^e(R(i), A(i), B(i)) = E(T_s^e) + E(T_f^e) + E(T_b^e), \tag{23}$$

where $E(T_s^e)$ is the time used to successfully transmit the burst of aggregated packets when considering both collisions and channel errors. $E(T_f^e)$ is the wasted time slots due to collisions and channel errors, and $E(T_b^e)$ is the time spent on performing backoff. Different from the case where channel impairment is negligible, here channel errors can cause a retransmission to any packet in a burst, not just the first one, and retransmission limit M_{max} is applied to each of them as well. Since collision can only happen to the first packet in the burst when the system has no hidden terminal, depending on their locations, packets in the same burst suffer from different PLRs, which result in different probability of retransmissions. For

the first packet, the probability to have m retransmissions is defined by $p'(m) = (1-P)P^m, m \geq 0$, where P is defined by (22). For the rest, we can estimate their retransmission probability using:

$$p''(m) = \begin{cases} (1-P)P_E P^{m-1} & m \geq 1 \\ 1 - P_E & m = 0 \end{cases} \quad (24)$$

Accordingly, we can compute $E(T_f^s)$ and $E(T_f^e)$ as the following, respectively:

$$E(T_s^e) = 2(DIFS + T_v + SIFS + T_{ack}) \sum_{m=0}^{M_{max}} p'(m) + (B(i)-1) \cdot 2(2SIFS + T_v + T_{ack}) \sum_{m=0}^{M_{max}} p''(m) \quad (25)$$

$$E(T_f^e) = \sum_{m=0}^{M_{max}} p'(m) m\tau_f + \sum_{m=M_{max}+1}^{\infty} p'(m)(M_{max}+1)\tau_f + (B(i)-1)\left(\sum_{m=0}^{M_{max}} p''(m) m\tau_f + \sum_{m=M_{max}+1}^{\infty} p''(m)(\tau_f' + M_{max}\tau_f) \right) \quad (26)$$

where τ_f' is the time wasted if the first transmission attempt of $k^{th}(k \geq 2)$ packet in the burst fails, $\tau_f' = 2(SIFS + T_v + T_{ATO})$. $E(T_b^e)$ can be computed based on the backoff procedure when collision and channel errors are both present. Figure 5, shows the transmission of an aggregated packet k in the burst. As shown by Figure 5, when the sender node detects a loss, it chooses a new backoff value from the doubled maximum contention window (unless it already reaches $CWmax$), and starts retransmitting from the last unacknowledged packet. Retransmission continues until the packet is acknowledged or it reaches M_{max}. Then the system resets its maximum contention window size to be $CWmin$ and proceeds to transmit the rest in that burst. Following such a backoff procedure, we can express $E(T_b^e)$ as:

$$E(T_b^e) = \sum_{m=0}^{M_{max}} p'(m) \sum_{r=0}^{m} \frac{W_r}{2}\delta + \sum_{m=M_{max}}^{\infty} p'(m) \sum_{r=0}^{M_{max}} \frac{W_r}{2}\delta + (B(i)-1)\left(\sum_{m=1}^{M_{max}} p''(m) \sum_{r=1}^{m} \frac{W_r}{2} + \sum_{m=M_{max}}^{\infty} p''(m) \sum_{r=1}^{M_{max}} \frac{W_r}{2} \right)\delta \quad (27)$$

where W_r is the maximum contention window size used for r^{th} backoff as defined in Section IV-A, and P is PLR defined by (22). Substituting $E(T_s^e)$, $E(T_f^e)$ and $E(T_b^e)$ into (23), we can obtain the channel occupancy time T^e for supporting bi-directional voice calls between RAP i and the GAP.

Given T^e, we can compute the total channel occupancy time for supporting N_R TAR-enabled RAPs with configurations $(R(i), A(i), B(i)), i = 1, ...N_R$ as:

$$T_{WLAN}^e = \sum_{i=1}^{N_R} T^e(R(i), A(i), B(i)) \quad (28)$$

C. Voice Capacity Analysis

From the discussion in previous section, we know that the problem of finding voice capacity of a multi-hop 802.11b network can be mathematically formulated as an optimization problem. Following the same spirit, when channel errors are not negligible, we can estimate the voice capacity by solving the following optimization problem:

$$\max_{R(i)\in R^+, A(i), B(i)\in Z^+} \sum_{i=1}^{N_R} A(i)B(i) \tag{29}$$

$$\text{S.T.} \quad A(i) \le N_{max} = \min\{15, A_{max}\}, \forall i : 1 \le i \le N_R \tag{30}$$

$$A(i)B(i) \le N_{max}, \forall i : 1 \le i \le N_R \tag{31}$$

$$E(T_s^e(R(i), A(i), B(i))) \le TXOP(i), \forall i : 1 \le i \le N_R \tag{32}$$

$$\sum_{i=1}^{N_R} T^e(R(i), A(i), B(i)) \le D_s \tag{33}$$

$$R(i) \in \{10^6, 2\cdot 10^6, 5.5\cdot 10^6, 11\cdot 10^6\}, \forall i : 1 \le i \le N_R \tag{34}$$

using two-stage searching approach proposed in Section IV-A. In this optimization problem, (33) is the system stability constraint when both channel error and collisions cause non-negligible channel losses.

D. α and β Overheads

Similarly, we can also evaluate the α and β overheads in this case by substituting $E(T_s)$, $E(T_f)$ and $E(T_b)$ in (14) and (15) using $E(T_s^e)$, $E(T_f^e)$ and $E(T_b^e)$, i.e.:

$$\alpha^e = \sum_{i=1}^{N_R} \frac{E(T_f^e) + E(T_b^e)}{T_{WLAN}} \tag{35}$$

$$\beta^e = \sum_{i=1}^{N_R} \frac{E(T_s^e) - 2T_v\left(\sum_{m=0}^{M_{max}} p'(m) + (B(i)-1)\sum_{m=0}^{M_{max}} p''(m)\right)}{N_R \cdot E(T_s^e)} \tag{36}$$

VI. Comparing Theory to Simulation

A. Simulation setup

We simulate a multi-hop wireless mesh network as shown by Figure 3, where N_R intermediate RAPs and a GAP perform traffic shaping for both uplink and downlink traffic. Aggregated packets will be de-aggregated at RAPs and the GAP for downlink and uplink traffic, respectively, before being forwarded to next hop. Voice traffic for both upflows and downflows are generated at a constant bit rate $64kbps$ according to ITU-T Rec. G.711 standard and be packed into a packet every $30ms$ [1]. This together gives us equally-sized voice packets, each consisting of 240 bytes of voice payload. To avoid high collisions caused by synchronized traffic generation, we randomize the time when a node starts to generate voice

[1]This is a good compromise between the resulting packetization delay and load on the network due to CSMA and transmission access overheads

traffic. Since voice traffic is a delay-sensitive traffic and can only tolerate certain level of packet losses, to support voice calls with satisfying quality, we require that for each pair of uplink/downlink flows, its PLR has to be no more than 1% and 99% percentile of the round-trip time cumulative distribution function (CDF) cannot exceed $250ms$. However, from our experiments [1], we find out that when the system satisfies the PLR constraint, it also has quite low transmission delays, while if the PLR constraint is violated, the round trip delay starts increasing in an unbounded fashion. This is due to that fact that when the PLR constraint is violated, the system is most likely overloaded and comes into an unstable status, where queue might be built up, leading to unbounded delays. Therefore, we can simply use the PLR to examine the feasibility of a traffic shaping setting. If all flows under a given traffic shaping setting satisfy the PLR requirement, such setting is feasible and otherwise, it is infeasible.

To implement joint traffic shaping strategies, we extend current ns-2.26 implementation with an implementation of 802.11e EDCF [19] by adding a traffic shaping buffer on each node. This buffer is used to accumulate enough packets before performing aggregation and/or bursting. We choose buffering deadline to be equivalent to $30ms$, i.e. the packetization interval of voice codec. It means that every $30ms$, node i aggregates a single packet from each of $A(i)$ users into a super packet and then bursts out $B(i)$ of those super packets after contending for the channel. This buffering delay depends upon the relative transmission/arrival times of each packet in the first hop. If we assume that voice traffic has a relatively low delay jitter over the first hop, we can be confident that with a traffic shaping buffering delay T_{buf} equivalent to the packetization interval, $A(i) \cdot B(i)$ packets are present at each traffic shaping instance. Other parameters are set according to Table I except that we choose the minimum contention window size CW_{min} to be 7 time slots and the maximum window size CW_{max} to be 15 time slots.

To determine voice capacity of a network through ns simulations, we use the following criteria:

$$C_s = \max_k \{k \in Z^+ | 1 \leq i \leq 2k, PLR_i \leq 1\%\} \tag{37}$$

where PLR_i is the PLR of the i^{th} flow. In other words, obtaining voice capacity via ns simulations is equivalent to finding out the maximum number of users supported by a feasible traffic shaping setting of a stable system.

B. Performance evaluation

To evaluate our proposed model, we present our results from three different aspects. First, we would like to examine how much we can gain from performing traffic shaping, how channel quality can affect voice capacity, and how different traffic shaping schemes affect the aforementioned α overhead and β

overhead. When presenting this part of result, we assume that the system uses balanced traffic shaping setting, i.e., each RAP/GAP is using same $A(i)$ and $B(i)$. Then, we will show a comparison between using balanced traffic shaping setting and using unbalanced traffic shaping setting. In the end, we will look into voice capacity improvement brought by performing rate adaptation.

N_R	Simulation		Analysis	
	$(A(i), B(i))$	Capacity	$(A(i), B(i))$	Capacity
1	(8,2), (2,8)	16	(8,2),(2,8)	16
2	(5,3),(5,3)	30	(8,2),(2,8)	32
3	(5,3)	45	(8,2)	48
4	(5,2)	40	(5,2)	40
5	(4,2)	40	(4,2)	40
6	(7,1)	42	(7,1)	42
7	(6,1)	42	(6,1)	42
8	(4,1)	32	(5,1)	40
9	(4,1)	36	(4,1)	36
10	(3,1)	30	(3,1)	30
11	(2,1)	22	(2,1)	22
12	(1,1)	12	(1,1)	12

TABLE III

VOICE CAPACITY WHEN JOINT TRAFFIC SHAPING IS SUPPORTED: NEGLIGIBLE CHANNEL IMPAIRMENT

1) The impact of traffic shaping: Table III, IV, and V compare voice capacity predicted by using our proposed model to voice capacity obtained through ns simulations when three different traffic shaping schemes are implemented: a) *joint aggregation and bursting*; b) *aggregation only*; and c) *bursting only*. In these three tables, both model-predicted voice capacity and simulation-obtained voice capacity are evaluated for different N_R value, i.e. $N_R \in \{1, 2, ...15\}$, and all nodes are operating at $11Mbps$. Table III, IV, and V also give out corresponding $A(i)$ and $B(i)$ settings for achieving those voice capacities. All simulation results hereinafter are average results from multiple simulation runs. From those results, we can see that our proposed model can accurately predict voice capacity in almost all the cases. Although there are a few cases where the predicted voice capacity is higher than the achievable voice capacity in ns simulations, the difference is small and it is often caused by slightly different $A(i)$ and $B(i)$ values. For example, in Table III, when $N_R = 3$, the achievable voice capacity in ns simulations is 45 and it is achieved by setting aggregation level $A(i)$ to be 5 and bursting level $B(i)$ to be 3. This

N_R	Simulation		Analysis	
	$A(i)$	Capacity	$A(i)$	Capacity
1	5	15	5	15
2	8	16	8	16
3	8	24	8	24
4	8	32	8	32
5	8	40	8	40
6	7	42	7	42
7	6	42	6	42
8	4	32	5	40
9	4	36	4	36
10	3	30	3	30
11	2	22	2	22
12	1	12	1	12

TABLE IV

VOICE CAPACITY WHEN ONLY AGGREGATION SCHEME IS SUPPORTED: NEGLIGIBLE CHANNEL IMPAIRMENT

N_R	Simulation		Analysis	
	$B(i)$	Capacity	$B(i)$	Capacity
1	15	15	15	15
2	8	16	8	16
3	6	18	6	18
4	4	16	4	16
5	3	15	3	15
6	3	18	3	18
7	2	14	2	14
8	2	16	2	16
9	2	18	2	18
10	1	10	1	10
11	1	11	1	11
12	1	12	1	12

TABLE V

VOICE CAPACITY WHEN ONLY BURSTING SCHEME IS SUPPORTED: NEGLIGIBLE CHANNEL IMPAIRMENT

configuration is feasible according to our model as well. However, our model provides other feasible traffic shaping settings, $(A(i), B(i)) = (8,2) or (2,8)$, which yields a higher voice capacity, i.e. 48 calls. This voice capacity is not achievable in ns simulations. The mismatch here basically shows that our model is providing a tight upper bound on voice capacity, which means that the system cannot support voice calls more than the capacity predicted by the model. Otherwise, the system becomes instable, suffering from high PLR and long round-trip delay.

To better elaborate the impact of traffic shaping, Figure 8 plots the voice capacity for different traffic shaping schemes. In addition to the above observations, from Figure 8, we can also see that our proposed joint traffic shaping scheme outperforms the two schemes which consider aggregation and bursting exclusively. When N_R is large enough, the joint traffic shaping scheme falls back to the *aggregation-only* scheme. If N_R reaches certain point, i.e. $N_R = 10$, the three schemes all use the same traffic shaping setting, $A(i) = 1, B(i) = 1$, yielding the same voice capacity.

When channel bit error is not negligible, we re-evaluate the impact of traffic shaping. Figure 9 shows the model-predicted performance versus simulation-obtained performance when channel SER is $5 \cdot 10^{-4}$. Again, we can see that when taking channel errors into account, our proposed model can still be very

accurate in predicting voice capacity of the system. Compared to Figure 8 where we have negligible channel bit error, Figure 9 shows much lower voice capacity for schemes where aggregation is used to shape the traffic. This happens because compared to original voice packets, aggregated packets are much longer and as a result more vulnerable to channel bit errors. Frequent retransmissions triggered by packet losses increase the α overheads, causing voice capacity degradation. However, such performance degradation does not appear to the *bursting-only* scheme. Since bursting keeps the original packet size, it is less affected by channel conditions as shown by Figure 9. Such observations provide us certain guidelines in choosing proper traffic shaping setting: when channel quality is good, higher aggregation level can be used to improve voice capacity; while when channel quality deteriorates, switching to a lower aggregation level might be a better choice.

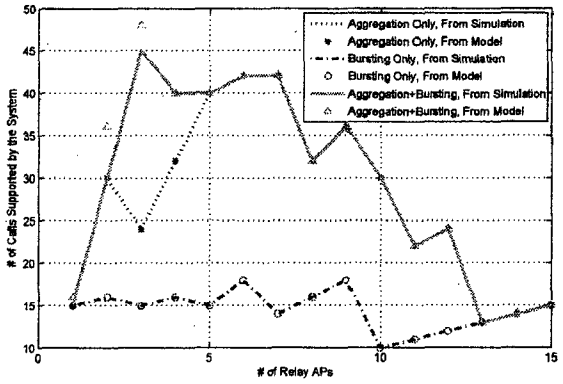


Fig. 8. Voice capacity when we have negligible bit errors

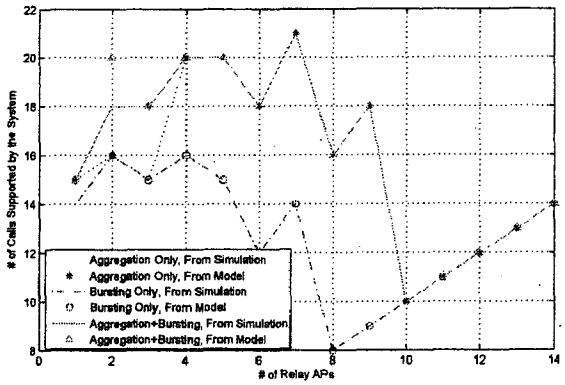


Fig. 9. Voice capacity when we have $SER = 5 \cdot 10^{-4}$

Figure 10 shows voice capacity for a wide range of signal-to-noise-ratio (SNR), where N_R is fixed to be 4 and different transmission rates are considered. We present both model-predicted results and simulation-obtained results in Figure 10, where dots denote voice capacity predicted by our proposed model while lines denote voice capacity obtained via ns simulations. Figure 11 illustrates the traffic shaping settings deduced from the model for different SNR values. From Figure 10, we can see that when nodes are operating on high data rates, i.e. $5.5Mbps$ and $11Mbps$, if SNR value drops to be below certain threshold, voice capacity deteriorates very fast. This is due to that the system tries to reduce the aggregation level to maintain a better error resilience, as shown in Figure 11. Note that for a given PHY rate, aggregation level does not always decreases monotonically. Instead, at certain points, higher aggregation level is used for a lower SNR value. Such fluctuations happen because at those points, there exist more than one traffic shaping settings which can yield the same voice capacity, and the

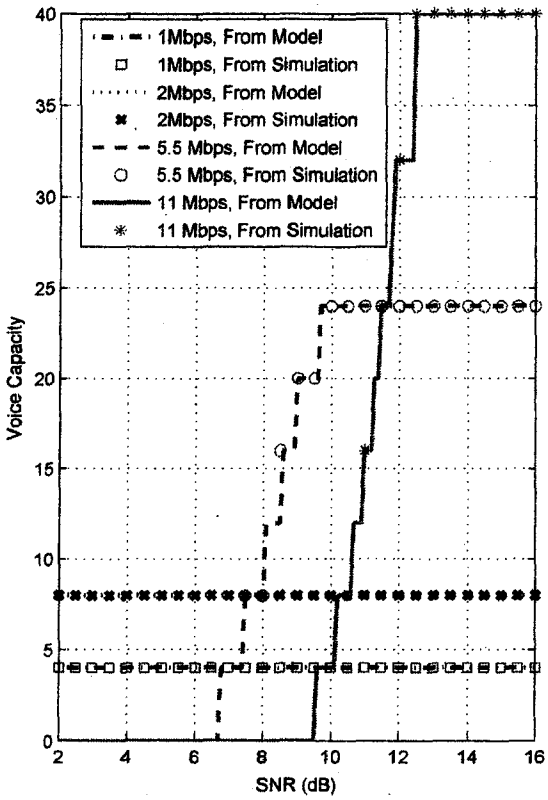


Fig. 10. Voice capacity under different channel conditions: $N_R = 4$

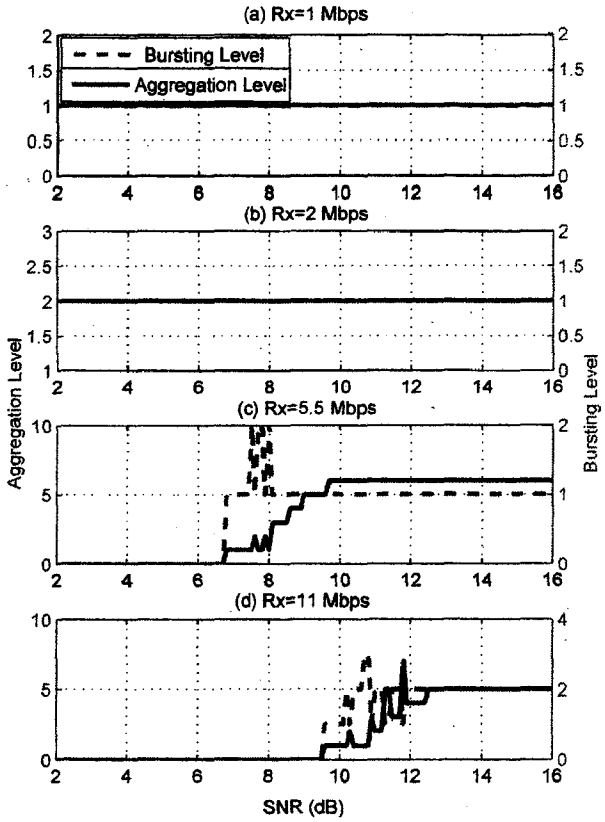


Fig. 11. Traffic shaping settings: $N_R = 4$

system chooses the one which consumes the least amount of channel time. Depending on the tradeoff between reducing β overhead by sharing the PHY/MAC header among more packets and increasing α overhead by more frequently retransmitting larger packets, fluctuation of aggregation level can appear. The α and β overheads associated with different schemes are shown in Figure 12 for $R = 11Mbps$ and Figure 12 for $R = 5.5Mbps$. From these two figures, we can see that as channel quality deteriorates, both overheads tend to increase regardless of small glitches appearing in this big tendency. Different from aggregation-based scheme which emphasizes more on reducing β overhead, bursting-based scheme mainly works on reducing α overhead and is much less efficient when dealing with β overhead. By implementing aggregation and bursting jointly, the system reaches an intermediate position for both α overhead reduction and β overhead reduction. One interesting observation from these two figures is that

using bursting can always result in lower α and β overheads. However, when aggregation is taken into consideration, only β overhead is always efficiently reduced by using traffic shaping. Depending on channel quality and aggregation level, α overhead in traffic shaping cases might even be higher than the case where we do not consider traffic shaping at all.

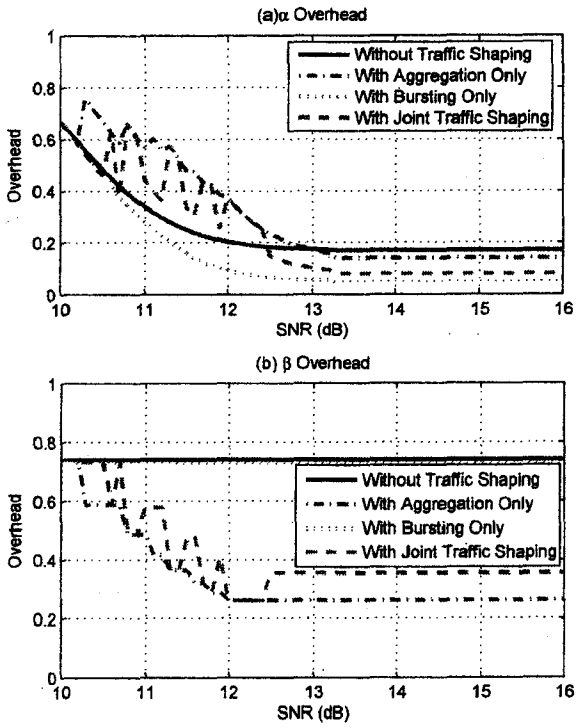


Fig. 12. Overhead measurement: $R = 11Mbps$

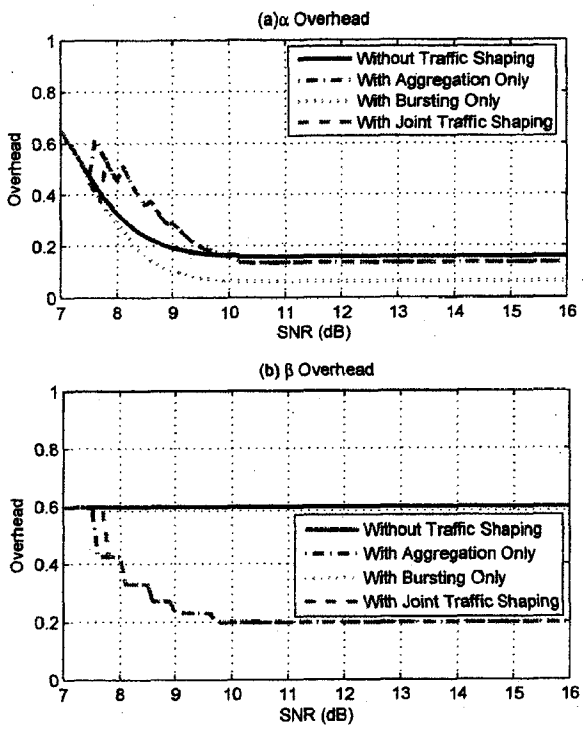


Fig. 13. Overhead measurement: $R = 5.5Mbps$

2) Balanced traffic shaping vs. unbalanced traffic shaping: In all of the above results, we use balanced traffic shaping settings for RAPs and the GAP, i.e. $A(i) = A, B(i) = B, R(i) = R$, and voice capacity is the maximum number of voice calls supported by a stable system under such settings. However, from model-based analysis and ns simulations, we both observe that when the system is reaching its capacity under balanced traffic shaping settings, it does not use up all its wireless resource. In some cases, the residue of wireless resource might be enough to support one or even more voice calls. One way to fully utilize the wireless resource is to use unbalanced traffic shaping settings for RAPs and the GAP. In other words, some RAPs and the GAP might use different aggregation level and/or bursting level. To show the benefit of performing unbalanced traffic shaping, we first examine voice capacity for a TAR-enabled

system with a fixed number of RAPs. Here, we choose N_R to be 4 again and exhaustively search for the optimal $(A(i), B(i))$ setting for given SNR and $R(i)$. Since our previous results have shown that our model can very accurately predict voice capacity when using balanced traffic shaping setting, hereinafter, we mainly present results for balanced cases from our proposed model. As illustrated by Figure 14, using unbalanced traffic shaping does further improve the voice capacity of a TAR-enabled system. Depending on PHY layer transmission rate and channel quality, the capacity gain brought by unbalanced traffic shaping setting varies. When APs all operate on $11Mbps$ and SNR value of received signals is high, the system can *squeeze* in 5 more voice calls. However, when the system operates on $5.5Mbps$ or below, the capacity gain by using unbalanced traffic shaping setting is quite trivial. Figure 14 is for a fixed

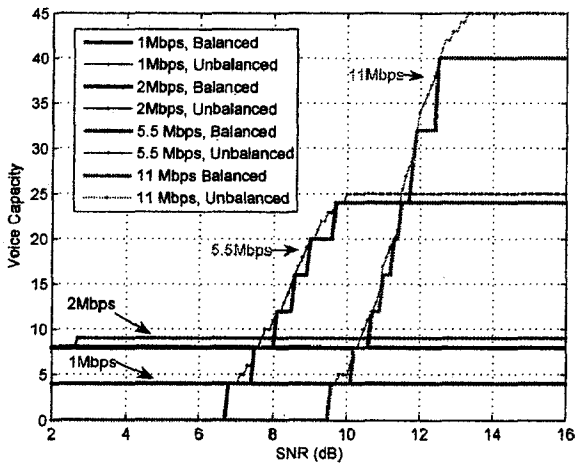


Fig. 14. Balanced traffic shaping Vs. unbalanced traffic shaping: $N_R = 4$

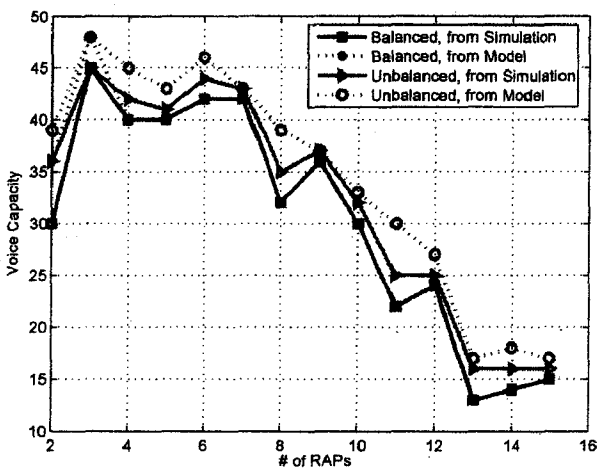


Fig. 15. Comparison of voice capacity improvement achieved by using unbalanced traffic shaping setting: $R = 11Mbps, SER = 0.0$

number of RAPs. Figure 15 shows the impact of using unbalanced traffic shaping when the system has various number of RAPs. Due to space limit, we only show the results for $R = 11Mbps, SER = 0.0$. From Figure 15, we can see that our proposed model can still provide a good prediction on voice capacity when RAPs and the GAP are using unbalanced traffic shaping setting, although it is not as accurate as when we consider balanced traffic shaping setting. Compared to model-predicted results, simulation-based results have shown more trivial performance improvement brought by using unbalanced traffic shaping settings. Meanwhile, from our previous analysis in Section IV, we have shown that the

computational complexity of finding voice capacity for a balancedly configured TAR-enabled system is much lower than the complexity of an unbalancedly configured system. Therefore, we choose balanced traffic shaping settings to configure TAR-enabled system in our later discussion on network management.

3) Rate adaptation: For a TAR-enabled system, when wireless channel quality deteriorates, the system chooses to switch to a modulation scheme which can provide better error resilience. Among four modulation schemes existing in 802.11b, BPSK has highest error resilience but lowest bit rate, while CCK11 has highest bit rate but is the least resilient to channel errors. Therefore, switching modulation schemes based on wireless channel quality is actually requiring nodes in the system to be able to adapt their transmission rates. Previous literature has demonstrated the efficacy of rate adaptation on improving wireless system throughput. In the rest of this session, we examine the impact of rate adaption on voice capacity of a TAR-enabled system, especially its joint force with traffic shaping. Since in the wireless mesh network we consider in this work, all RAPs within in one hop range are likely to get the same wireless channel quality, they tend to adapt their rate in a similar way, i.e. using the same $R(i)$ before and after the rate adaptation. Figure 16 shows the impact of rate adaptation evaluated via our proposed model for a fixed number of RAPs, i.e. $N_R = 4$.

In Figure 16, (a), (b) and (c) are results for a system without implementing any traffic shaping strategy, a TAR-enabled system with balanced traffic shaping settings, and a system with unbalanced traffic shaping settings, respectively. Solid curves in this figure are voice capacity without performing rate adaptation and blue dotted curves are capacity after adapting PHY rate. In Figure 16 (a), we can see that system capacity decreases very fast when operating at $11Mbps$ in SNR range $9.1dB - 11.9dB$ or $5.5Mbps$ in SNR range $6.0dB - 9.0dB$. We observe the similar patterns in Figure 16(b)(c). However, by performing rate adaptation, the system becomes more resilient to the deterioration of channel quality, as illustrated by blue dotted lines in all three plots in Figure 16.

From Figure 16, we can also see that by implementing traffic shaping, those points when voice capacity becomes decreasing are shifted to a righter position in SNR axis. For example, for a system operating at $11Mbps$, its voice capacity begins to decrease from $SNR = 11.9dB$ when no traffic shaping strategy is implemented. However, if balanced traffic shaping strategy is implemented, the decreasing starts earlier, i.e. $SNR = 12.4dB$. When considering unbalanced traffic shaping, this capacity decreasing point appears further right in the SNR axis, i.e. $SNR = 13.1dB$. Such a tendency can be explained by the decreased error resilience of a scheme using aggregation to send longer packets during transmissions.

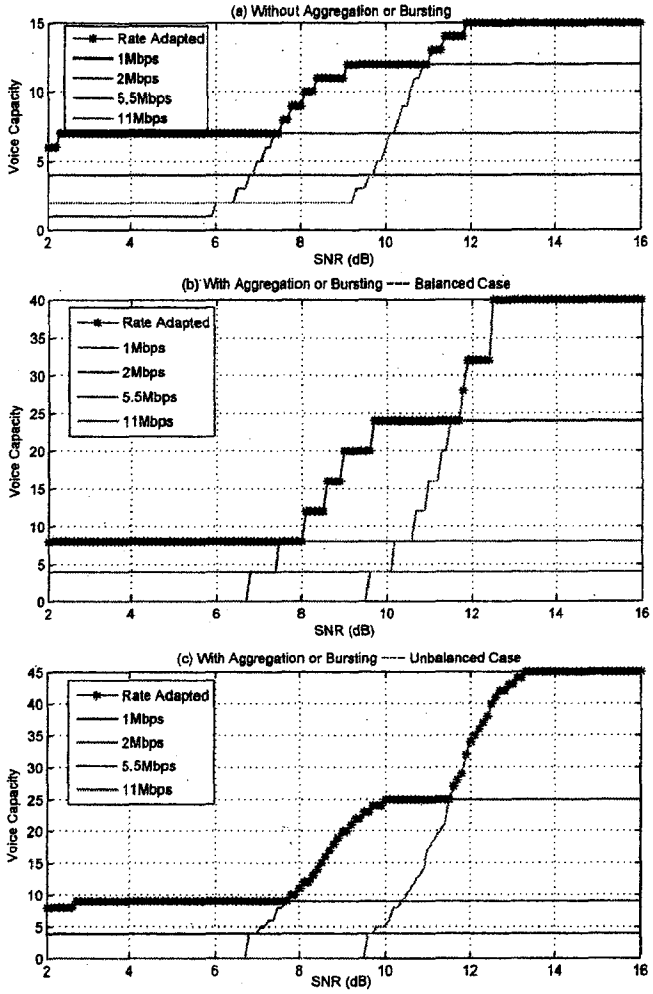


Fig. 16. The impact of rate adaptation evaluated via the proposed model: $N_R = 4$

VII. Rate Adaptation and Network Management

Our previous analysis and presented results so far reveal some important observations in the planning and design of two-hop wireless mesh networks. In this section, we discuss how to use our proposed model to manage such networks in a practical setting, where wireless channel quality and voice traffic patterns change constantly or are uncertain.

A. Implementation issues of a TAR-enabled system

Figure 17 shows the generic architecture of a TAR-enabled node. As illustrated by Figure 17, by measuring signal-to-noise ratio(SNR) of received wireless signals, each node can estimate its wireless

channel quality. SNR measurements will then be collected by TAR control unit. Depending on whether the system implements centralized control mechanism or distributed one, SNR measurements will be processed differently. If a centralized control scheme is implemented, TAR control unit of each RAP forwards its collected information to the GAP and retrieve updates on TAR configuration(i.e. aggregation level, bursting level, PHY transmission rate and routing decision) from the GAP. If a distributed control scheme is implemented, each RAP processes SNR information locally to obtain TAR updates by looking up the prebuilt table to determine the best traffic shaping settings and PHY mode for the next transmission attempt. Meanwhile, *call manager* updates its serving list using received routing related information.

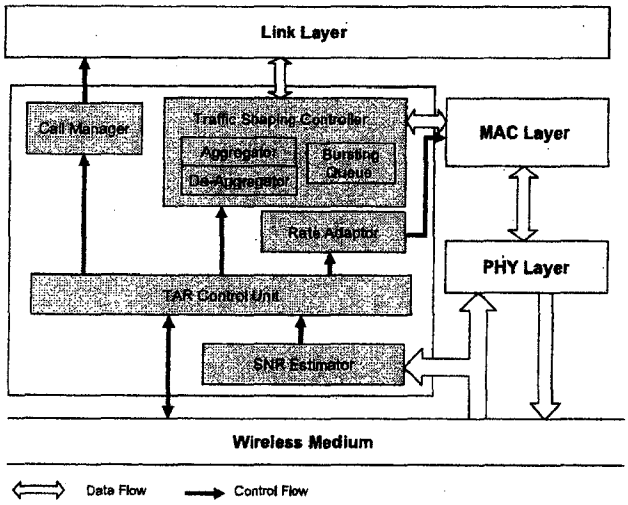


Fig. 17. Generic architecture of A TAR-enabled Node

B. Network planning

One of the key conclusions we can draw from previous results is that when given an abundant number of relay nodes, we should not blindly apply a naive load balancing approach to equally share all wireless users among the relays. System voice capacity curves have suggested that we should first identify an optimal number of relay nodes and then enforce load balancing among the identified number of relay nodes. We give an example here to elaborate this point in more details. Suppose we have a network with 10 relay nodes and there are 40 VoIP users to support. Brute-force load balancing would assign 4 users per relay. But according to Table III, the system can at most support 30 voice calls if requiring 10 relay nodes are all used, i.e. $[N_R, (A(i), B(i))] = [10, (3, 1)]$. Obviously, putting 4 calls on each of these

10 users is not a feasible scenario. However, our results show that there are other scenarios which can support 40 VoIP users, $[6,(7,1)],[7,(6,1)],[5,(4,2)],[4,(5,2)],[3,(5,3)]$. Instead of using all 10 nodes, each of those feasible scenarios only chooses a subset of 10 relay nodes, and relay selection naturally involves additional criteria such as geographical proximity, channel quality, and low buffering latency.

Meanwhile, capacity curves as in Figure 8 along with traffic shaping setting curves as in Figure 11 reveal a structure for mesh network layout. For instance, if a capacity of 40 users is desired, a floor plan with 4 relay nodes each serving 10 users using traffic shaping setting $(A(i),B(i)) = (5,2)$ or a plan with 5 nodes each serving 8 users using traffic shaping setting $(A(i),B(i)) = (4,2)$ can both attain the desired user capacity when wireless channel quality is good. Given candidate layouts, we can make the final decision by considering additional factors such as the availability of non-interfering channels.

Since in a practical setting, wireless channel quality changes constantly. When channel quality deteriorates very fast, we need to find a way to increase error resilience. From Figure 11 and Figure 16, we can see that the system always tries to lower its aggregation level first when detecting a channel quality degradation. If channel quality still deteriorates, then the system chooses to decrease its PHY rate and increase aggregation level accordingly. This process repeats as channel quality keeps decreasing. Such a network-aware PHY-rate and traffic shaping setting adapting process will result in different voice capacity while channel quality varies. As mentioned previously, different desired voice capacity affects the choice of relay nodes as well. Therefore, when planning a mesh network layout, we should also consider wireless channel quality and make the system able to quickly adapt to network status.

C. Network management

Managing wireless mesh networks includes two parts: how to admit new users into the system (i.e. which AP to associate each newly joined user to) and how each AP operates given the users they support.

To find out how to admit new users into the system, we need first to decide the system status when the new user sends out the join-in request. System status is decided by many variables, i.e., transmission rates of each RAP/GAP, traffic shaping settings used by the system, number of users existing on each AP and etc.

(see provisional patent for more details)

VIII. Conclusion

In this paper, we address the bandwidth inefficiency problem during transmitting voice traffic over multi-hop 802.11 wireless networks. Instead of using aggregation and bursting exclusively, we leverage

both to design a joint traffic shaping scheme which can effectively reduce α overhead as well as β overhead. We propose to use relays and routing decisions in such a way that when they are combined with proper traffic shaping mechanisms, substantial increase in the system capacity can be attained in terms of the number of VoIP users that can be supported. In addition, when traffic shaping loses its advantages due to high bit/symbol/packet-losses, we propose to use PHY-rate adaptation options in IEEE802.11 to balance the capacity degradation. The choice of traffic shaping settings and PHY-rate depend on network status. To analyze the voice capacity of a system using our proposed strategies, we develop a theoretical framework, which also provides an easy way to evaluate the feasibility of a given network floor plan. Our simulation results demonstrate that our framework can accurately predict the voice capacity of a system with given number of relay nodes, and compared to conventional transport methods, the proposed joint traffic shaping can boost up the capacity significantly in all cases. For some combination of relay and traffic shaping setting, the voice capacity can be even twice more. Meanwhile, our simulation results also show that by adapting PHY-rate and traffic shaping to network status, the system can provide better error resilience and maintain more stable voice capacity.

Although we focus on wireless network topologies with a depth of two hops and a single GAP, most of the insights gained over this network topology can be carried over to more general network. However, there exist more interesting and complicated network planning and network management problems, which we will address in our future work. The other issue we have not considered is in this paper is that there might co-exist different types of background traffic. Unlike VoIP, other services such as video streaming can generate more variable and bursty traffic. How to design a traffic shaping strategy which can accommodate different types of services is something that we would like to address in the future as well.

REFERENCES

[1] C. Pepin, U. C. Kozat, and S. Ramprashad, "A Joint Traffic Shaping and Routing Approach to Improve the Performance of 802.11 Mesh Network," in *IEEE WIOP*, April 2006.

[2] "Network simulator," http://www.isi.edu/nsnam/ns/.

[3] Y. Gwon, James Kempf, Raghu Dendukuri, and Ravi Jain, "Experimental Results on IP-layer Enhancement to VoIPv6 over IEEE 802.11b Wireless LAN," in *IEEE WiNMee*, April 2005.

[4] ITU-T, "Pulse Code Modulation (PCM) of Voice Frequencies," ITU-T Recommendation G.711,1988 (Blue Book), 1993.

[5] IEEE Computer Society, "Part 11: Wireless Medium Access Control (MAC) and Physical Layer (PHY) specifications: Medium Access Control (MAC) Enhancements for Quality of Service (QoS)," 2005, IEEE Std. 802.11eD13.0.

[6] "TGn Sync Proposal for 802.11n," `http://www.tgnsync.org/techdocs`.

[7] "WWiSE proposal for 802.11n," `http://www.wwise.org/technicalproposal.htm`.

[8] W. Wang, S. C. Liew, and V. O. K. Li, "Solutions to Performance Problems in VoIP over a 802.11 Wireless LAN," *IEEE Transactions on Vehicular Technology*, vol. 54, no. 1, January 2005.

[9] J. Zhu and S. Roy, "A 802.11 Based Dual-Channel Reservation MAC Protocol for In-Building Multihop Networks," *Springer Mobile Networks and Applications*, vol. 10, no. 5, 2005.

[10] A. Raniwala and T. Chiueh, "Architecture and Algorithms for an IEEE 802.11-Based Multi-Channel Wireless Mesh Network," in *IEEE INFOCOM*, Miami, FL, March 2005.

[11] R. Draves, J. Padhye, and B. Zill, "Routing in Multi-Radio, Multi-Hop Wireless Mesh Networks," in *ACM Mobicom*, Philadelphia, PA, Sept. 26-Oct. 1 2004.

[12] S. Garg and M. Kappes, "Can I add a VoIP call?," in *IEEE ICC*, Anochorage, Alaska, June 2003.

[13] K. Medepalli, P. Gopalakrishnan, D. Famolari, and T. Kodama, "Voice capacity of IEEE 802.11b, 802.11a and 802.11g Wireless LANs," in *IEEE GLOBECOM*, Dallas, Texas, Nov. 2004.

[14] D. Hole and F. Tobagi, "Capacity of an IEEE802.11b WLAN Supporting VoIP," in *IEEE ICC*, Paris, France, June 2004.

[15] K. Medepalli, P. Gopalakrishnan, D. Famolari, and T. Kodama, "Voice Capacity of IEEE802.11b and 802.11a Wireless LANs in the Presence of Channel Errors and Different User Data Rates," in *IEEE VTC*, Los Angeles, CA, Fall 2004.

[16] IEEE Computer Society, "Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: Higher-speed Physical Layer (PHY) extension in the 2.4 GHz band," IEEE Std. 802.11b-1999 (R2003).

[17] G. Bianchi, "Performance analysis of the ieee 802.11 distributed coordination function," *IEEE JSAC*, vol. 18, 2000.

[18] M. K. Simon, S. M. Hinedi, and W. C. Lindsey, *Digital Communication Techniques: Signal Design and Detection*, Prentice-Hall, Inc., 1995.

[19] "NS2 802.11e package," `http://www.tkn.tu-berlin.de/research/802.11e\_ns2`.

Appendix B

An Analysis of Joint Aggregation, Bursting and Rate Adaptation Mechanisms for Increasing VoIP Capacity in Multi-Hop 802.11 Networks

Sean A. Ramprashad Danjue Li Ulaş C. Kozat Christine Pépin

DoCoMo USA Labs, Palo Alto, CA, 94304

danjli@cisco.com, {ramprashad,kozat,pepin}@docomolabs-usa.com

Abstract

Due to the low cost and ease of deployment, single-hop and multi-hop 802.11 networks have become attractive solutions for providing *last-mile* broadband (wireless) access in urban environments. However, a critical issue in using such networks to support applications such as Voice over IP is the widely known overheads in the Medium Access Control (MAC) layer and Physical (PHY) layer for each transmission. The effect of these overheads can be more severe in multi-hop deployments, and can limit the number of VoIP calls per Gateway Access Point (GAP) in a multi-hop system to be no greater than that of a single-hop single-GAP system.

In this paper, we build upon our previous work that showed that a multi-hop network, with one GAP using a single channel, can in fact support more users than a single-hop network. To do so functions such as aggregation have to be intelligently used at relays in the network. We build on this prior work giving a theoretical framework that considers the joint tradeoff of aggregation, bursting, and PHY rate adaptation given a set of admission and routing decisions. We compare call-capacity estimates from the theory to those derived from simulations, and illustrate through examples the advantages of considering routing/admission decisions jointly with bursting and aggregation.

I. Introduction

We consider the problem of transmitting voice traffic efficiently over a multi-hop 802.11 network. Such networks are attractive solutions for providing *last-mile* broadband (wireless) access in many urban and corporate environments. This is true given the low cost of 802.11 end-points, the availability of power sources (lamp posts, etc.), the fact that 802.11 is a widely used standard, and the possibility of needing only a few gateways for wired connectivity outside of the net. However, such networks do suffer from efficiency problems similar to those in single-hop systems due to inherent overheads in the Medium Access Control (MAC) and Physical (PHY) layers when accessing and transmitting on the wireless medium [1]. In multi-hop scenarios, such inefficiencies can propagate and concentrate along flows across the network. Furthermore, these can be severe for important applications such as Voice-over-IP (VoIP) that have strict delay

requirements and which generate persistent two-way traffic with small payload sizes. Such traffic can create bottlenecks as it concentrates towards gateways.

In [1] a joint mechanism of admission/routing and traffic shaping was proposed to improve the efficiency of multi-hop VoIP transmission. The benefits of such an approach was demonstrated via ns-2 simulations where it was shown that more than a 3-fold increase in VoIP call-capacity is possible for 2-hop systems that use the joint mechanism relative to 2-hop systems that simply forward packets. The basic premise is that as traffic moves toward a (wired) gateway one can use the intermediate (wireless) "relay" nodes intelligently to modify the traffic characteristics in terms of timing, overheads, aggregation, bursting, etc., to make the use of subsequent links more efficient. In particular the approach notes, as exploited in 802.11n and observed in other studies [2] [3] [4] [5], that by aggregating payloads destined to a single (next-hop) destination into common transmitted packets one can amortize a significant portion of the MAC/PHY overheads. The result is increased efficiency and VoIP call-capacity. In fact, by pro-actively delaying packets above the MAC layer to facilitate greater aggregation in relays one can increase capacity without increasing the end-to-end system delay across multiple hops [1, Fig. 4]. Similar observations are made in related testbed study in [6] that considers opportunistic use of aggregation and other network topologies under different constraints.

Methods to increase efficiency by using aggregation and bursting [4] [5], and even multicasting (in downlinks) [8], have been proposed for single-hop scenarios. However for multi-hop scenarios an important point made in [1] is that the ability and extent to which aggregation and bursting are used often depends on routing and admission decisions, e.g. routing can be used to concentrate traffic into relays. In particular, as shown in [1], forcing a user to use more hops to the GAP (e.g. use a longer route through a relay than connect directly to the GAP) can in fact increase the number of calls the system (GAP) can support. In fact we show later (Section VII) that one can consider routing/admission purposefully with traffic-shaping to achieve a number of different objectives including, but not limited to, maximizing voice call capacity. Furthermore options beyond those explored in [1] such as PHY rate adaptation are of interest in scenarios that operate in the presence of transmission errors. One therefore has many options and decisions to consider in both designing and managing such jointly optimized multi-hop systems.

However, understanding the effect that a given joint parameter-settings and decisions can have on the system performance is not simple. In fact, analyzing the joint effect of all possible

aggregation, bursting, PHY settings, and admission/routing operations via simulations/testbeds is often not practical given the large number of possible combinations that can be considered, even for a moderate sized network. This poses the problem of how to understand, design and operate these systems. This is in part why [1] limits investigation to a few "balanced" 2-hop scenarios. Similarly other simulation/testbed based studies, as in [6], are pragmatically limited to a few scenarios (in [6] a 6-node linear topology and a 15-node example) without explicit joint consideration of all parameters. In fact, though both [1] (directly) and [6] (implicitly) point to the benefits of joint optimization, comparison of these studies is not possible since the scenarios/networks considered are different.

In this paper we want to explore and understand the impact of these joint optimizations in more detail in a more general setting. The framework describes such a setting considering the operation of the Distributed Coordination Function (DCF). Though not addressed, networks using a centralized Point Coordination Function (PCF), which save additionally on overheads such as those due to contention, would also benefit from similar joint methods that amortize many inherent MAC/PHY overheads. Using DCF as first step, the goal of the paper is to quantify the underlying advantage/effect of a given joint selection. We limit ourselves to scenarios whereby the multi-hop network can be broken down into non-interfering communication groups with no hidden terminals. By doing so we focus mainly on the inherent advantages/disadvantages of joint MAC/PHY-parameter/admission/routing settings, ignoring for simplicity the additional degradations that may occur from interference and/or hidden nodes in practical deployments. Such issues can be addressed in the analysis by introducing additional loss terms. The impact of some of these issues can be mitigated in part by using Request-to-Send/Clear-to-Send (RTS/CTS) messages as in the analysis in [5], though our analysis will not explicitly model such a process/option. There are also related problems of path discovery, channel/transmission-error estimation, delay estimation, etc., in the practical management of such a system. Some of these are considered in [6]. Such issues are, however, beyond the scope of this paper.

The paper provides a theoretical framework that includes many of the main parameters of interest, e.g. aggregation, bursting, PHY rate, and symbol error-rate, on each link in the multi-hop network. The admission and routing choices implicitly define which links are present and what aggregation and bursting settings are possible. We are primarily interested in general tradeoffs, and thus the framework is similar in spirit to the capacity-based analyses outlined in [9] [10] [11].

However, we do compare the theory with results obtained via ns-2 simulations for some basic scenarios of interest. Under the assumption of non-interfering groups, for a given setting the load on each wireless resource in the multi-hop network is calculated. This allows one to explore conveniently a large number of possible parameter combinations and their feasibility, i.e. their ability to support the present VoIP flows. The framework therefore provides a tool to help understand, manage and plan such networks, allowing one to identify potential bottlenecks and heavily loaded areas in the network. We also give illustrative examples whereby classic methods that try to distribute load, as in [12] [13] [14], may not support as many voice calls as methods that jointly select admission/route and MAC/PHY parameters.

The rest of the paper is organized as follows. In Section II, we briefly discuss the inefficiencies of the current 802.11 DCF mechanism, leading into our proposed *Joint Traffic shaping, Admission, Routing and Rate-adaptation (JTARR)* approach. In Section III we provide a theoretical framework to analyze the loads in a multi-hop network for the case of no channel impairments, describing in Section IV how to use these loads to give an upperbound on VoIP capacity. In Section V we extend this theory for cases with channel impairments. In Section VI we compare the theory with ns-2 simulations, looking in detail at joint tradeoffs of parameters. In Section VII we discuss briefly how to manage such networks in a practical setting, showing how the framework can be used to address a number of different objectives, including maximizing VoIP capacity. Finally, in Section VIII, we summarize the work and discuss some open problems.

II. JTARR: Joint Traffic Shaping, Admission, Routing and Rate Adaptation

To begin, consider a DCF managed single-hop scenario using an 11 Mb/s PHY rate, a 1 Mb/s basic rate, a single channel, and a single access point (AP) that by default serves as the GAP. Mobile users connect to this AP, and each uses 2-way constant bit-rate 64 kb/s flows (consistent with ITU-T Rec. G.711 [15]) to support a VoIP call. If the payload of each 802.11 packet represents 30 msec of speech data with an additional 40 byte IP/RTP/UDP overhead, an 802.11b system may only be able to support 15 VoIP calls [16, Table 4]. Here the payloads occupy about 20% of the total time on the wireless medium. If the packet interval is reduced to 10 msec, the system may only be able to support 6 calls with payloads occupying now only about 10% of the time on the medium [16, Table 4]. Furthermore note that if additional APs connect to the single GAP using the same channels, simply forwarding VoIP packets from other users/APs, the

entire system can support no more calls than the corresponding single-hop system. In fact, if all APs/users use the same channel as the GAP the system throughput often degrades significantly with the increasing number of hops, e.g. as shown in [6, Fig. 1].

Inefficiencies are particularly apparent in 802.11-based VoIP deployments since VoIP uses small payload sizes. Given the inherent DCF MAC/PHY layer overheads per transmitted packet, the relative inefficiencies increase with decreased payload size. Here it is worth noting that for single-hop systems increasing the packet duration may not be feasible in VoIP since this incurs a delay penalty. However, multi-hop systems allow for additional functions across multiple packets. To improve the efficiency we propose to implement for multi-hop networks a ***Joint Traffic shaping, Admission, Routing and Rate adaptation*** (JTARR) mechanism at intermediate relay access points (RAPs) and GAPs. For this paper we will consider networks as illustrated in Figures 1 and 2, where access points are allowed to aggregate and/or burst packets. In such networks by correctly admitting users, routing traffic, and also properly delaying the release of packets to the MAC layer, it is possible for APs to use (or evoke as in bursting) such mechanisms to increase the efficiency of the 802.11 MAC/PHY layers [1].

Aggregation is the process by which payloads from different IP packets are included into one 802.11 transmitted packet. This can reduce the number of contending transmission packets, and amortizes MAC/PHY-headers and control overheads over multiple voice packets. Bursting is the process by which multiple 802.11 packets are transmitted in a single transmission opportunity. This focuses on suppressing contention/idle-slot overheads and is thus less effective than aggregation. However, if packet losses due to bit/symbol/packet-errors are non-negligible, then aggregation may have disadvantages since it relies on using longer packets. Furthermore, since the bit/symbol/packet loss rate depends on the underlying PHY rate, one may try to jointly adapt aggregation and bursting strategies with 802.11 PHY rate options. Our goal is to present a framework quantifying the general tradeoffs of potential joint settings.

A. General Model, Assumptions and Examples

As a first step to understanding tradeoffs, we consider in this paper simplified models of potential JTARR-enabled VoIP systems. Specifically, the models are such that the analysis can be broken down into an analysis per communication group. A communication group is defined as a set of *flows* that contend with each other for the same wireless resource. Such a group is defined

by a given wireless channel in a given geographic location. A single flow represents a number of VoIP links with the same origin and destination (in a given communication group) that are bundled together through aggregation and/or bursting to use common transmission opportunities. Such links can originate/terminate at clients or access points.

For simplicity, we assume that the channels of different communication groups operate orthogonally to each other and that all nodes originating/terminating flows within the group can sense each other, i.e. there are no hidden terminals. This orthogonality can be possible if different groups use different channels and/or if different groups use the same channel but are sufficiently separated geographically, as in the illustrations in Figures 1 and 2. For practical deployments, where the number of channels or geographic separation may be restricted, additional problems such as interference between groups and hidden terminals do arise. To keep the analysis necessarily simple and focused on the underlying tradeoffs, we will not explore these issues directly. However, when the effect of symbol errors is considered in Section V there is the potential to account in part for such effects by including additional error- vents.

In the model access points (APs or RAPs or GAPs) are equipped with multiple wireless interfaces. Interfaces can operate in parallel, i.e., receive and/or transmit on interfaces simultaneously. Take Figure 1(a). In this example we consider using 4 channels and two gateways where access points may have up to two interfaces. Here the channel that a flow/link uses is synonymous with the term communication group. RAP1, RAP2 and GAP2 each have one interface on Channel 1 that is used for connections between these access points. RAP1 has an additional interface on Channel 4 for connection to GAP1. Each access point uses *only* its "Primary Channel" to communicate with users in their service area. Given the constraint of 4 channels, GAP1 uses Channel 4 for this purpose and RAP1 uses Channel 1 for this purpose. RAP2 and GAP2 use additional channels, channel 2 and 3 respectively, to serve users directly. Additional channels could be used to create more interfaces, or different primary channels. Alternatively, additional channels can expand the coverage area as shown in Figure 1(b) where groups re-use the same channels but are geographically separated.

Users have multiple options in such systems. For example in Figure 1(a) "user1" can potentially be associated (admitted) with RAP1, GAP1 or GAP2. Similarly "user2" and "user3" also have multiple admission possibilities. In such a system, one could consider admission decisions (and also routing decisions) independently of other parameters. Here it would make sense to try to

admit a given user first to GAP1 or GAP2, if possible, then to RAP1 or RAP2, if possible. However, as will be shown later this is not the best *joint* decision that can be made.

B. A 2-Hop Tree-Rooted Example

In the simulations in Section VI we will focus on a 2-hop tree-rooted network with a single GAP as illustrated by Figure 2. Group 1 in this example is also helpful in following the framework developed in the next sections. The 2-hop network in Figure 2 has the following components: (1) K communication groups, each associated with the primary channel of an access point; (2) wireless VoIP clients that act as both the ultimate sources (in the uplink) and sinks (in the downlink) of VoIP flows; (3) wireless JTARR-enabled RAPs that can forward and process many VoIP flows; and (4) a JTARR-enabled GAP through which VoIP sessions are established with hosts outside the network. In this example, as in Figure 1(a), RAPs have two interfaces and the GAP has one interface. If an additional non-interfering channel is available the GAP could use it to serve its clients directly thus creating an additional $(K+1)^{th}$ group.

C. Defining Links and Flows

In our approach we look at the load placed on the wireless channel of each communication group. This load defines the fraction of time the channel is "in use", i.e. is being held by nodes in the group. To calculate the loads in a multi-hop system we need to define how the VoIP packets traverse the system. In Figure 2, each voice call includes one uplink direction and one downlink direction, and packets travel over one or two hops. For other topologies, one can more generally define the routes of each voice-packet between a client and a terminating GAP. Given the packets sent through a RAP, or to a client or GAP, there are different bursting, aggregation and rate options that one may consider for each wireless link in each group. Using the load estimates, we can determine the feasible set of options for each group. Hereinafter we shall use the term "channel occupancy time" and "load" interchangeably.

Let a communication group consist of a number of flows contending for the same wireless resource. These flows are generated by N_n nodes, $node(1), \ldots, node(N_n)$, which can be mobile clients, RAPs, and/or GAP(s). For example in Group 1 of Figure 2 $node(1), \ldots, node(K-1)$ are RAPs, $node(K)$ is the GAP, and $node(K+1), \ldots, node(N_n)$ are clients that connect directly to the GAP. In other groups the situation may be different. We use the index $i = 1, \ldots, N_F$ to index the N_F flows being supported by this group. In Figure 2 Group 1 all packets from

clients directly served by each RAP are combined into one uplink flow, and all packets destined for clients served by a given RAP are combined in the GAP into one downlink flow. With this there are a total of $N_F = 2N_n - 2$ flows in Group 1. Let S_k be the set of flow indices of flows originating at node(k)

$$S_k = \{i : \text{Flow } i \text{ is transmitted by node}(k)\} \quad (1)$$

We assume each flow has its own JTARR settings. The bursting level $B(i)$ denotes for flow i how many aggregated packets are burst in a single transmission opportunity. Within a burst, let $A(i,j)$, $j = 1,\ldots,B(i)$, denote the number of voice packets included in the j-th transmitted packet of a burst. The $A(i,j)$ value is chosen so that the length of any aggregated packet does not exceed the maximum transfer unit (MTU) limit of an 802.11 packet. The maximum value we denote as A_{max}. Furthermore let $R(i)$ and $H(i)$ denote the PHY layer data rate and PHY layer basic rate used for flow i.

We consider that all voice calls generate constant bit rate traffic at a source rate of R_s bits per second for each direction. Voice traffic is packed for transport into packets representing D_s seconds of voice each. The payload of packets at the MAC layer includes a 40 byte IP/UDP/RTP header. That is, each voice packet at the MAC layer includes $(R_sD_s + 8 \cdot 40)$ bits of data in the payload. A flow is therefore supporting (on average) the transmission of at least one voice packet per D_s seconds for each (voice-call)-direction it is supporting. Table I further summarizes the important constants we will use in equations and the analysis. With these assumptions and definitions we can completely define the traffic in our system and proceed with a load analysis.

III. Analyzing Loads per Group without Channel Impairments

In this section, we present a theoretical framework to estimate the channel occupancy time for a given group. We begin by focusing on the case where channel impairments are negligible.

A. Channel Occupancy Analysis

Given the traffic supported by a group "$\mathcal{G}$" we look at a measure, "$T_{\mathcal{G}}$", of the expected load generated by the flows in the group in a D_s interval of time. We begin by assuming that there are no channel errors and that transmitted packets are lost only due to collisions. In this case the load comprises three components: 1) T_s, the time for successfully transmitting bursts of aggregated packets; 2) T_f, the time wasted due to collision-caused transmission failures; and

3) T_b, the associated time spent in random backoff when contending for the channel. These times are often statistical. We therefore look at long term average (expected) values, i.e. Load $= E[T_s + T_f + T_b] = E[T_s] + E[T_f] + E[T_b]$. We will calculate a lowerbound T_G on this load.

B. Probability of Successful and Failed Transmissions

A key set of statistics used in subsequent sections is the probability of successful and failed transmissions. All quantities $E[T_s]$, $E[T_f]$ and $E[T_b]$ depend on these statistics. We now consider the case of packets being lost due to collisions only. Since there are no hidden terminals in the system, a collision occurs only when two or more terminals start transmissions simultaneously. Therefore in a collision event the first packet in a burst is lost terminating the transmission burst. In our system we limit the number of retransmissions to M_{max}, where each packet has M_{max} such attempts *independent* of other packets. Therefore, for *each packet* there are $M_{max} + 2$ cases one needs to consider: $M_{max} + 1$ successful cases corresponding to $0,\ldots,M_{max}$ failed transmissions followed by a good transmission, and the case of all $M_{max} + 1$ possible transmissions failing.

Define $p(m, i, j)$ as the probability of having exactly m *failed* transmissions for the j^{th} packet in flow i. Define $P_{succ}(i, j) = 1-p(M_{max}+1, i, j)$. This defines the probability the j^{th} packet was transmitted successfully within $M_{max} + 1$ attempts. Clearly once the first packet is transmitted successfully, then all other packets $j > 2$ follow successfully in the burst with probability one since we consider only losses due to collisions. However, when the first (or any) packet fails to be transmitted after $M_{max} + 1$ attempts it is dropped. Define $P_{fail}(i, j)=1 - P_{succ}(i, j)$. If a packet is dropped the next packet re-starts the contention/bursting process as shown in Figure 3.

An exact calculation of $p(m, i, j)$ values rests on many assumptions [9] [11], e.g. the probability of colliding on the first transmission attempt and the additional effect due to the growth of backoff window limits. Here we make a simplifying assumption. For the purpose of average load and voice capacity calculations, especially in heavily loaded scenarios, it suffices to assume that collisions happen independently of each other with an approximate steady-state probability P_C as in [9] [17]. Such heavily loaded scenarios are the most important in understanding voice capacity and in managing the system. Under these assumptions we approximate $p(m, i, j)$ by

$$p(m,i,j) = \begin{cases} P_{fail}(i,j-1)(1-P_C) + P_{succ}(i,j-1) & m = 0 \\ (1-p(0,i,j))\,(1-P_C)P_C^{m-1} & 0 < m \leq M_{max} \\ (1-p(0,i,j))\,P_C^{m-1} & m = M_{max}+1 \end{cases} \quad (2)$$

One can calculate values recursively starting with $j = 1$ and $m = 0$ using $P_{fail}(i, 0) = 1$. As shown in Figure 3,

$$P_{first}(i, j) = P_{fail}(i, j - 1) \qquad P_{burst}(i, j) = P_{succ}(i, j - 1) \tag{3}$$

are the probabilities of packet j being the first packet in a burst transmission, and packet j being a second, or third, etc. in a burst transmission, respectively.

In our paper we will also need to assume a value P_C. If we assume that collisions happen mainly between the GAP and the RAPs or clients (and less between RAPs or between clients), one value to consider is $P_C = 1/(CWmin + 1)$ [11]. Another is suggested in [5, (1)]. Using the equations outlined in Sections III-C to V, Figure 4 shows the sensitivity of capacity to the P_C value. We can see that the voice capacity prediction, as defined in Section IV, is not that sensitive to small changes of P_C and is quite flat around $1/(CWmin + 1)$. Therefore we use $P_C = 1/(CWmin + 1)$ for our calculations though in general one can use any value of P_C or any form of $p(m, i, j)$ in Sections III-C to V. Finally, it should also be noted that $P_{fail}(i, j)$ has significance in analyzing VoIP traffic. VoIP applications can tolerate values $P_{fail} > 0$ depending on the voice codec and packet loss concealment functions used. Typical acceptable loss limits $LOSS_{max}$ range from 1% to 5%.

C. T_s: *Occupancy of successful transmissions*

For VoIP we look at load values (channel occupancy times) *within* a D_s interval of time. In a D_s interval each flow in a group, on average, services one voice-packet from each client using that flow. In our system the flow also has to transmit the group of such packets as one set of burst/aggregated packets on average within no more than D_s seconds of time. Let $T_s(i)$ denote the time spent on successful transmissions of voice packets from flow i. Here we include headers, the transmission time of ACK packets T_{ack}, and any necessary inter-frame spacings (IFS). If all packets in the burst are sent the time is simply

$$\bar{T}_s(i) = DIFS - SIFS + \sum_{j=1}^{B(i)} (T_v(i, j) + 2SIFS + T_{ack}) \tag{4}$$

where,

$$T_v(i, j) = T_{header}(i) + T_{data}(i, j) \qquad T_{PHY}(i) = PHY/H(i) \tag{5}$$

$$T_{header}(i) = T_{PHY}(i) + \frac{(MAC+FCS)\cdot 8}{R(i)} \qquad T_{ack}(i) = T_{PHY}(i) + \frac{8\cdot ACK}{R(i)} \tag{6}$$

$$T_{data}(i, j) = \frac{A(i,j)}{R(i)} (D_s R_s + (IP + UDP + RTP) \cdot 8) \tag{7}$$

The parameter values used above for 802.11b can be found in Table I. If the $j-1^{th}$ packet fails, then the j^{th} packet becomes the first in a number of subsequent burst attempts. Here define

$$P_{succ2}(i,j) = P_{first}(i,j)(1-P_C) + \sum_{m=1}^{M_{max}} p(m,i,j) \quad (8)$$

This defines the probability the j^{th} packet was transmitted successfully as the first in a burst. Note, $P_{succ}(i,1) = P_{succ2}(i,1)$ since $P_{first}(i,1) = P_{fail}(i,0) = 1$. A packet that is successfully transmitted as the first in the burst has a DIFS period before transmission, and a packet successfully transmitted as the 2^{nd}. 3^{rd}, etc. in the burst has a SIFS period. $E[T_s(i)]$ follows as:

$$\begin{aligned} E[T_s(i)] &= \sum_{j=1}^{B(i)} P_{succ2}(i,j)\left(DIFS + T_v(i,j) + SIFS + T_{ack}\right) \\ &\quad + \sum_{j=2}^{B(i)} P_{burst}(i,j)\left(2SIFS + T_v(i,j) + T_{ack}\right) \end{aligned} \quad (9)$$

Finally, we sum over flows

$$E[T_s] = \sum_{i=1}^{N_F} E[T_s(i)] \quad (10)$$

At this point note that while we have focused on the case of constant bit-rate (CBR) VoIP traffic, all using the same packet interval D_s and source rate R_s, the model could include other traffic types with other statistics. Here the value $T_{data}(i,j)$ would be a statistical quantity, reflecting a range of possible values depending on what sized/type packets, at a given time, are aggregated and burst together in the flow. Individual values have to conform to constraints such as the MTU size, and in (10) the expected value $E[T_{data}(i,j)]$ is ultimately the quantity of interest. For simplicity in the exposition we will focus on the CBR case with constant D_s and R_s.

D. T_f: Occupancy of unsuccessful transmissions

We now look at the time wasted due to collisions *within* each D_s interval of time. A collision event affects only the first packet in a burst. When packet j in flow i is the first in a burst and a transmission attempt fails, the time wasted by this flow is given by $\tau_f(i,j)$ where:

$$\tau_f(i,j) = DIFS + T_v(i,j) + T_{ATO} \quad (11)$$

Here T_{ATO} denotes the ACK timeout, i.e., the time a sender has to wait before declaring its packet as lost. We set it to be equal to $T_{ATO} = T_{PHY} + SIFS + \delta$ [18]; δ is defined in Table I. Using $p(m,i,j)$, the expected number of such failed *first* transmissions is given by:

$$F_1(i,j) = \sum_{m=1}^{M_{max}+1} mp(m,i,j) \quad (12)$$

Note, under the collision-only assumption $j > 1$ packets never fail if they are transmitted as the second, third, etc. in a burst. This is implicitly assumed by (12). With (12),

$$E[T_f(i)] = \sum_{j=1}^{B(i)} F_1(i,j)\tau_f(i,j) \quad (13)$$

In calculating $E[T_f]$ however one can not simply sum over all i as in (10) since collisions involve overlaps of two or more flows. However, if we consider networks where one node, e.g. the "root" node, serves many flows, a lowerbound to $E[T_f]$ accounting for many independent events is

$$E[T_f] \geq \max_{k=1,..,N_n} \sum_{i \in \mathcal{S}_k} E[T_f(i)] = E[\hat{T}_f] \quad (14)$$

Here flows originating from a given node obviously do not collide with each other, and therefore one can sum across such flows. We can expect that the maximum value in (14) is often given by the node with the heaviest load, e.g. the root node. If the number of flows for such a load is a large fraction of the total number of flows then such a bound may in fact be tight.

E. T_b: Time due to random backoff

On the first attempt to obtain a transmission opportunity, a transmitting node may see one of two cases. In the first case, on getting a packet, the node senses the channel is free for a period DIFS. In this case no backoff is performed and the packet is transmitted. In the second case the channel was not free for a period of DIFS. Here the node proceeds with contending for the channel using a random backoff. Assume the transmitting node has failed to transmit the first packet in a burst after m attempts, and let U_r be the length of the r^{th} backoff used in units of time slots. The cumulative time spent on backoffs for flow i after m such attempts is:

$$\sum_{r=1}^{m} U_r(i)\delta \quad (15)$$

Denote the maximum collision window size at the r^{th} transmission attempt as W_r, i.e. given a backoff is performed $U_r \sim$ Uniform$[0, W_r]$. Following the backoff procedure, we have $W_r = \min\{CWmax, 2^{(r-1)}\ (CWmin + 1)\ -1\}$, $r = 1, ..., M_{max} + 1$. To calculate $E[U_r]$ we need to consider the possibility that the node saw the channel free when it first sensed the channel, and so could have transmitted without a backoff, i.e. $U_1 = 0$. We ignore this possibility, assuming it happens with small probability under heavy loading. We therefore assume $E[U_r] = W_r/2\ \forall r$.

To calculate $E[T_b(i)]$ we need to calculate on average how many backoff processes occur for flow i in a D_s interval of time. Here we need to consider all the possible cases where the

j^{th} packet, $j = 1, \ldots, B(i)$, becomes the first in a burst. Note, when a sender node detects a lost packet, it re-contends for the channel with a new backoff value and on obtaining a TXOP starts retransmitting from the last unacknowledged packet. Retransmission continues until either the packet is acknowledged or the M_{max} retransmission limit for that packet is reached. Packets exceeding this limit are dropped. After a drop the system resets its maximum contention window size to $CWmin$ and proceeds to transmitting the remaining packets in that burst. For simplicity we define $W_{M_{max}+2} = 0$, and with this $E[T_b(i)]$ is given by:

$$E[T_b(i)] = \left(\sum_{j=1}^{B(i)} P_{first}(i,j)\right) \frac{W_1}{2}\delta + \left(\sum_{j=1}^{B(i)} \sum_{m=1}^{M_{max}+1} p(m,i,j) \sum_{r=2}^{m+1} \frac{W_r}{2}\right) \delta \quad (16)$$

The first term accounts for the probability that a backoff is triggered by packet j. The second term counts for any r, $r > 1$, backoffs. As in (14) we bound $E[T_b]$.

$$E[T_b] \geq \max_{k=1,\ldots,N_n} \sum_{i \in S_k} E[T_b(i)] = E[\hat{T}_b] \quad (17)$$

The assumption underlying (17) is that the backoff mechanism is applied only to one flow at a time within S_k. Finally we define the load lowerbound

$$T_G = E[T_s] + E[\hat{T}_f] + E[\hat{T}_b] \quad (18)$$

IV. Voice Capacity Analysis

A. Relationship to the Hole-Tobagi Bound

One can at this point consider using (10) and (17) to lowerbound the load for a system without collisions. This lowerbound load can then be used to upperbound the voice capacity. In a single-hop system without any aggregation and bursting options, i.e. $B(i) = 1 = A(1,1)$ $\forall i$, the equations can define the load for supporting Q voice calls in a system. Here we have Q clients and one GAP, i.e. $Q+1$ nodes and $2Q$ flows. The GAP supports Q downlink flows. Therefore the lowerbound of the load for Q calls, L_Q, is

$$L_Q = \sum_{i=1}^{2Q} E[T_s(i)] + \max_{k=1,\ldots,Q+1} \sum_{i \in S_k} E[T_b(i)] = \left(\sum_{i=1}^{2Q} E[T_s(i)]\right) + \sum_{\text{GAP flows}} E[T_b(i)] \quad (19)$$

One can then use this to find $Q_{max} = \max_Q$ such that $L_Q \leq D_s$. Under the assumption of no collisions $p(0,i,j) = P_{succ}(0,i,1) = 1$ $\forall j$, $E[T_b(i)] = \bar{T}_s(i)$, and $E[T_b(i)] = CW_{min}\delta/2$. With $B(i) = A(i,j) = 1$ in $\bar{T}_s(i)$ (4), it follows that Q_{max} is the same as the upperbound derived by

Hole and Tobagi in [10]. In their result, the authors consider the case of no collisions where only the GAP backs off. The results are the same because the $\max_k$ operation in (19) is achieved by the node which is the GAP, i.e. the result is that only backoffs of the GAP are counted.

B. Illustrating General Case with a 2-hop Network

The group-wise analysis allows us to compute an upperbound on the voice capacity (based on a lowerbound on load) for many systems by extending the approach in Section IV-A. Indeed, in order to support calls in a system $T_{\mathcal{G}}$ has to be less than the packetization interval D_s for all groups $\mathcal{G}$. For illustration let us consider the 2-hop scenario as described in Section II-B with $N_R = K - 1$ RAPs, each with a different primary channel, and a single GAP using a $(N_R + 1)^{th}$ channel. RAPs connect to the GAP using this $(N_R + 1)^{th}$ channel. There are N_R RAP defined groups, each consisting of flows between the given RAP and its respective clients. We will assume that these groups do not use bursting and aggregation, and so each has a limit $N_{max} = Q_{max}$ on the number of clients it can support as described in Section IV-A.

Under the restriction that this Q_{max} limit is not violated for first-hop (RAP) groups, the voice capacity is given by the maximum number of calls that can be supported by flows in the 2^{nd}-hop, i.e. flows GAP↔RAPs and GAP↔clients using the $N_R + 1^{th}$ channel (group). Let us assume that each RAP is supporting one uplink RAP→GAP flow. These N_R flows are given indices $i = 1, \ldots, N_R$. Additionally we assume that there are N_c clients directly connected to the GAP. This gives N_c uplink client→GAP flows with indices $i = N_R+1, \ldots, N_u$ where $N_u = N_R+N_c$. We will assume $B(i) = A(i, 1) = 1$ for $i = N_R + 1, \ldots, N_u$. The number of voice-packets in D_s seconds supported by these uplink flows is given by

$$\text{\# packets in } D_s \text{ secs.} \;=\; \sum_{i=1}^{N_R} \sum_{j=1}^{B(i)} A(i,j) + \sum_{i=N_R+1}^{N_u} 1 \qquad (20)$$

Assuming one voice-packet per call, this is also the number of voice calls supported by uplinks.

For each uplink flow assume that the GAP supports a downlink flow transferring the same number of packets per second. Indicating such flows by $i = N_u + 1, \ldots, 2N_u$, the constraint is

$$\sum_{j=1}^{B(i)} A(i,j) \;=\; \sum_{j=1}^{B(i+N_u)} A(i+N_u, j),\ 1 \le i \le N_u \qquad (21)$$

The voice capacity is the maximum value of (20) such that the system load does not exceed D_s.

$$\text{Capacity Upperbound} \;=\; \max_{B(i),A(i,j)\in Z^+} \sum_{i=1}^{N_F} \sum_{j=1}^{B(i)} A(i,j), \text{ assuming (21) holds} \qquad (22)$$

$$\text{given} \quad \sum_{j=1}^{B(i)} A(i,j) \leq N_{max} \ \forall i, \ A(i,j) \leq A_{max} \ \forall i,j \tag{23}$$

$$T_{\mathcal{G}} \leq D_s \text{ for all groups, and } P_{fail} \leq LOSS_{max} \tag{24}$$

$T_{\mathcal{G}}$ often reduces to a form where the GAP is the node achieving the maximum in (14) and (17).

Later we will consider for simplicity "balanced" scenarios where for RAP↔GAP flows $B(i) = \bar{B} \ \forall i$ and $A(i,j) = \bar{A} \ \forall i,j$. For these scenarios the capacity can be obtained by an exhaustive search on $\bar{B}$ and $\bar{A}$. In more general cases, where RAPs can use different parameters and/or support varied numbers of clients the search space can be large and a capacity analysis can be less meaningful. Often in looking at bounds for such "*unbalanced*" cases we search locally around optimal solutions for balanced cases. However it is straightforward during network operation to use loads to test the feasibility of a particular set of parameters in supporting clients, i.e. to test $T_{\mathcal{G}} \leq D_s \forall \mathcal{G}$ where parameters are possibly implicit on an admission and/or routing decision. Note $T_{\mathcal{G}}$ is actually a lowerbound on load. In a practical system one should allow for a budget $\delta > 0$ and test $T_{\mathcal{G}} \leq D_s - \delta$, as discussed in Sections VI-A and VI-B.

V. Load and Voice Capacity with Channel Impairments

We now extend the analysis for the case when channel impairments are not negligible. Here packets may be lost due to both collisions and channel errors. Unlike losses due to collisions, losses due to channel errors can happen to both data packets and ACK packets anytime during the transmission. We focus on channel errors due to additive noise, ignoring losses due to interference from other wireless nodes. Assume that a byte represents a symbol and that symbol errors occur in an independently identically distributed (i.i.d) fashion. We also assume that $H(i) << R(i)$ in our analysis and ignore errors in symbols transmitted with the basic rate $H(i)$. If P_S is the symbol error rate (SER), then the packet error rate $P_E(i,j) = 1 - (1 - P_S)^{L(i,j)}$. Here $L(i,j)$ consists of all symbols transmitted at the data rate $R(i)$, i.e. $L(i,j) = MAC + A(i,j)(RTP + UDP + IP + \frac{D_s R_s}{8})$ for aggregated packets. Since an ACK packet is small (14 bytes), we will assume that the probability of an erroneous ACK packet due to additive noise can be neglected.

We consider the long-term average of the channel occupancy for group $\mathcal{G}$ when transmissions are subject to both collision and channel errors. As in $T_{\mathcal{G}}$ we consider a lowerbound $T_{\mathcal{G}}^e$

$$T_{\mathcal{G}}^e = E[T_s^e] + E[\hat{T}_f^e] + E[\hat{T}_b^e], \tag{25}$$

Parallel to the definitions used with T_G, $E[T_s^e]$ is the time used to successfully transmit aggregated packets when considering both collisions and channel errors, $E[T_f^e]$ is the wasted time slots due to collisions and channel errors, and $E[T_b^e]$ is the time spent on random backoff. Values $E[\hat{T}_f^e]$ and $E[\hat{T}_b^e]$ are lowerbounds for $E[T_f^e]$ and $E[T_b^e]$ respectively.

A collision can only happen to the first packet in the burst. Symbol errors can affect any of the packets in a burst. Thus only the first aggregated packet transmitted in a burst sees both impairments. By taking the union of the two events (channel errors and collisions) we assume a steady state packet loss rate for the first packet in a burst transmission given by

$$P(i,j) = P_C + (1 - P_C)P_E(i,j) \tag{26}$$

For other aggregated packets there are two cases to consider. If such a packet is being transmitted in a burst following a previous successful transmission it can only be in error through symbol errors, i.e. the probability of a loss is given by $P_E(i,j)$. If on this attempt this packet is in error, then on retransmission this packet now becomes the first packet in retransmissions. It also becomes the first packet in a burst if the previous packet was never transmitted successfully. In these last two cases the packet sees the loss probability given by (26).

Under collisions and symbol errors let $p'(m,i,j)$ be the probability that the j^{th} packet in flow i had m *failed* transmissions. We define $P'_{fail}(i,j) = p'(M_{max}+1,i,j)$ and $P'_{succ}(i,j) = 1 - P'_{fail}(i,j)$ as in Section III-B. We assume that each packet in a burst has M_{max} retransmission opportunities independent of other packets. It follows:

$$p'(m,i,j) = \begin{cases} P'_{fail}(i,j-1)(1-P(i,j)) + P'_{succ}(i,j-1)(1-P_E(i,j)) & m = 0 \\ (1-p'(0,i,j))\,(1-P(i,j))P^{m-1}(i,j) & 0 < m \leq M_{max} \\ (1-p'(0,i,j))\,P^{m-1}(i,j) & m = M_{max}+1 \end{cases} \tag{27}$$

We can calculate values recursively starting with $m = 0$ and $j = 1$ using $P'_{fail}(i,0) = 1$.

Define $P'_{first}(i,j) = P'_{fail}(i,j-1)$ and $P'_{burst}(i,j) = P'_{succ}(i,j-1)$, and let $P'_{succ2}(i,j) = P'_{first}(i,j)(1-P(i,j)) + \sum_{m=1}^{M_{max}} p'(m,i,j)$ as in Section III-C.

$$\begin{aligned} E[T_s^e(i)] \;=\; & \textstyle\sum_{j=1}^{B(i)} P'_{succ2}(i,j)\,(DIFS + T_v(i,j) + SIFS + T_{ack}) \\ & + \textstyle\sum_{j=2}^{B(i)} P'_{burst}(i,j)(1-P_E(i,j))\,(2SIFS + T_v(i,j) + T_{ack}) \end{aligned} \tag{28}$$

Define $F'_1(i,j) = P'_{first}(i,j)P(i,j) + \sum_{m=1}^{M_{max}+1}(m-1)p'(m,i,j)$ which accounts for failed

transmissions of the j^{th} packet as the first in a burst. Then

$$E[T_f^e(i)] = \sum_{j=1}^{B(i)} \left(F'_1(i,j)\tau_f(i,j) + P'_{burst}(i,j)P_E(i,j)\tau'_f(i,j) \right) \quad (29)$$

Here $\tau'_f(i,j)$ is the time wasted when a packet fails but is not first in a burst, i.e., $\tau'_f(i,j) = (SIFS + T_v(i,j) + T_{ATO})$ and $\tau_f(i,j)$ is defined in (11). We can lowerbound $E[T_f^e]$ as

$$E[T_f^e] \geq \max_{k=1,\ldots,N_n} \sum_{i \in \mathcal{S}_k} E[T_f^e(i)] = E[\hat{T}_f^e] \quad (30)$$

As in (16), assume $W_{M_{max}+2} = 0$. We can express $E[T_b^e(i)]$ as:

$$E[T_b^e(i)] = \left(\sum_{j=1}^{B(i)} P'_{first}(i,j) \right) \frac{W_1}{2}\delta + \left(\sum_{j=1}^{B(i)} \sum_{m=1}^{M_{max}+1} p'(m,i,j) \sum_{r=2}^{m+1} \frac{W_r}{2} \right) \delta \quad (31)$$

As in (30) we bound $E[T_b^e]$ as follows:

$$E[T_b^e] \geq \max_{k=1,\ldots,N_n} \sum_{i \in \mathcal{S}_k} E[T_b^e(i)] = E[\hat{T}_b^e] \quad (32)$$

Substituting $E[T_s^e]$, $E[\hat{T}_f^e]$ and $E[\hat{T}_b^e]$ into (25), we obtain $T_{\mathcal{G}}^e$. Section IV describes the analysis of voice capacity for a 2-hop network. One can do the same analysis for the case $P_E > 0$ by simply substituting $T_{\mathcal{G}}^e$ for $T_{\mathcal{G}}$ in the analysis.

Though not considered in our analysis, interference between wireless nodes and from hidden nodes can create additional losses. Some of these losses, including losses of ACK packets which may now be non-negligable, can be roughly accounted for by increasing P_C in (26). The losses of ACK packets may also create situations whereby packets are delivered yet nodes keep retransmitting. Interference in itself affects the assumed channel signal-to-noise ratio, and thus P_E. Both P_C and P_E may now be both functions of the group and/or individual nodes in question at a particular point in time. Some of these losses can be mitigated by using the RTS/CTS mechanism in 802.11, which for example would reduce P_C and also the time wasted in each collision. Here additional signalling overheads would have to be included in the analysis.

VI. Comparing Theory and Simulations

To demonstrate the usefulness of the theoretical framework we provide analysis and comparisons to simulations for the two-hop wireless network described in Section IV. In the system traffic shaping is performed for the uplink and downlink traffic directions between the GAP and

RAPs. Aggregated packets are de-aggregated at the RAPs before being forwarded to clients. The GAP similarly de-aggregates before forwarding to clients outside the 802.11 network.

For simulations we use ns-2.26 extended with an 802.11e EDCF package [19]. We added a traffic shaping function and buffer at each RAP and GAP. The buffers are used to accumulate packets before aggregating and bursting functions. Both theory and simulations use the parameter values in Table I. In the general case, a RAP, say node(k^*), takes one packet from each of its $C(k^*)$ clients, delays, aggregates and then forwards these aggregated packets (on average once every D_s msec) using one or *more* burst transmissions. A client is a member of only one flow. Let $\mathcal{U}_{k^*}$ be the index set of all uplink RAP→GAP flows in $\mathcal{S}_{k^*}$ serviced by node(k^*). Then since that RAP serves $C(k^*)$ clients, parameters satisfy $\sum_{i\in\mathcal{U}_{k^*}} \sum_{j=1}^{B(i)} A(i,j) = C(k^*)$. Similarly in the downlink direction the GAP accepts one packet from each client every D_s msec and groups those destined for a common RAP into one or more flows. For simplicity we force each downlink flow to have the same aggregation/burst parameters used by the corresponding uplink flow. The examples use $H(i)$=1 Mb/s and the limits $A_{max} = 8$ and $M_{max} = 4$. For rate $R(i)$=11 Mb/s the constraint $\sum_{i\in\mathcal{U}_{k^*}} \sum_{j=1}^{B(i)} A(i,j) = C(k^*) \leq 15$ represents the first-hop limit. VoIP packets on all calls are generated using CBR coding with $R_s = 64$ kb/s and $D_s = 30$ msec.

To avoid collisions in simulations caused by overly synchronized traffic, we randomize the start time of calls at clients as in [16]. In considering simulated voice capacity we tolerate a low level of packet loss. From prior observations in [1], a 2-hop system that satisfies a low packet loss constraint also has low transmission delays, e.g. a round trip including packetization delays between clients and the GAP of below 150 msec. Therefore, in determining capacity via simulations we deem that a scenario (the VoIP call count, associated traffic shaping, RAPs, etc.) is supportable if and only if the packet loss on all flows is below 1% in the simulation. The simulated VoIP capacity is the maximum supportable call count under that constraint. In fact, the simulated capacity values given later are average values; each supportable scenario tested represents multiple simulation runs with different (random) client start-times.

A. Capacity Analysis: Balanced without Channel Impairments

We first consider the case without channel impairments comparing the voice capacities predicted by our model to those obtained through simulations. Tables II, III, and IV summarize cases for which joint aggregation and bursting, aggregation-only (i.e. $B(i) = 1\ \forall i$), and bursting-only

(i.e. $A(i,j) = 1\ \forall i,j$) are used, respectively. Also note that we limit ourselves for simplicity to presenting only "balanced" scenarios. Here we consider that clients connect only to RAPs ($N_c = 0$, $N_u = N_R$), each RAP serves the same number of clients, all flows use the same $B(i) = \bar{B}\ \forall i$, and the bursts use the same aggregation level $\bar{A} = A(i,j) \forall i,j$ for all transmitted packets. Both the theoretical prediction in (22) and simulations use this constraint. The tables show the $\bar{A}$ and $\bar{B}$ achieving the maximum number of calls for each (balanced) scenario. If different RAPs support different numbers of clients, using different traffic-shaping parameters, one could find cases supporting a larger number of clients. However, the search space considering all cases is too large to simulate and present.

For joint-options in Table II, that allow $\bar{A}, \bar{B} \geq 1$, a RAP uses only one uplink RAP→GAP flow, and the GAP only one downlink GAP→RAP flow per RAP, to service calls. For Tables III and IV multiple flows are often used to support uplinks and downlinks, i.e. $|\ \mathcal{U}_k\ | > 1$. Using more flows is less efficient as expected (and shown). Except in limited cases, e.g. $\bar{B} = 1$ for the case $N_R = 1$ in Table III, the tables show that by using aggregation and bursting at relays the system can often support many more clients than if relays simply forwarded packets (i.e. $\bar{A} = \bar{B} = 1$), or if all clients had connected directly to the GAP. Indeed in this latter case the system can only support 15 calls. One can also see that aggregation is more efficient than bursting in these error-free scenarios. We can also see from these tables that the simulation results and theory agree very well. Note, the upperbound provided by the theory does not discount the scenarios gained by simulations, i.e. the theory overestimates the feasible set of scenarios. In all cases overestimates are on the order of 1 or 2 extra calls per RAP. Such observations can be used to guide the headroom "δ" mentioned in Section IV-B which is necessary in adjusting lowerbound load (upperbound capacity) values for use in practical systems. Some differences (in part) are due to the constraints of considering balanced only cases.

We have not considered "unbalanced" cases. Here one can expect, as in the case of $N_R = 11$ in Table III, that if some nodes use a slightly higher aggregation level than others that both simulations and theory would allow more calls to be supported. Also note that under error-free operations and for a given bursting level $\bar{B}$ that two scenarios, one with $A(i,j) > 0\ i = 1,\ldots,\bar{B}$, the other with $\hat{A}(i,j) > 0\ i = 1,\ldots,\bar{B}$, have the same load if $\sum_{j=1}^{\bar{B}} A(i,j) = \sum_{j=1}^{\bar{B}} \hat{A}(i,j)$. For example, the balanced case $N_R = 4$ with simulated $(\bar{A}, \bar{B}) = (5,2)$ has the same load as the case $A(i,1) = A_{max} = 8$ and $A(i,2) = 2$. Both cases are therefore feasible.

B. Aggregation and Bursting Tradeoffs with Symbol Errors

With channel impairments, the impact of traffic shaping needs to be re-evaluated. Figure 5 shows the model-predicted performance versus the simulation-obtained performance with an i.i.d. SER rate of $5 \cdot 10^{-4}$. We can see that with high-levels of channel errors, our proposed model can provide a well-matched upperbound on capacity relative to the simulated capacity. The differences seen, as before, are often on the order of 1 or 2 extra calls per RAP. Compared to error-free cases in Tables II to IV, the SER=$5 \cdot 10^{-4}$ case in Figure 5 has a lower voice capacity, in particular for error-free schemes that benefited from high levels of aggregation. Indeed, longer aggregated packets have higher packet loss rates. In contrast, the relative degradation in performance for bursting-only schemes is less.

C. PHY Layer Rate Adaptation

The observations in the previous section suggest that in the presence of errors using lower aggregation levels and higher bursting levels may be preferable. We now examine how useful such adaptations are when additionally allowing for adaptation in the PHY rate, $R(i)$. Figure 6 shows the voice capacity for a wide range of signal-to-noise-ratios (SNRs) for the case $N_R = 4$ using $R(i) = 2$, 5.5 and 11 Mb/s. Note, for 2.0 and 5.5 Mb/s cases the limits $\sum_{i \in \mathcal{U}_{k^*}} \sum_{j=1}^{B(i)} A(i,j) = C(k^*) \leq 2$ and ≤ 7 represent the respective first-hop limits used. In all cases $H(i) = 1$ Mb/s.

The figure shows "balanced-cases" only. A few points are labeled with the corresponding SER and the selected "balanced-case" parameters $(\bar{A}, \bar{B})$ that maximize voice capacity[1]. What the figure shows is that below a certain SNR the aggregation levels and voice capacities drop dramatically within a 2 to 3 dB SNR range. Beyond this range switching to a lower PHY rate (and a larger $\bar{A}$) is preferable. Given the higher first-hop limit (15 calls) in the 11 Mb/s case, bursting is useful to allow RAPs to service up to 12 clients per RAP without increasing the aggregation level beyond 6. In fact, at high SNR it is possible to allow some RAPs to admit more than 12 users with $B(i) = 2$ using aggregation levels no greater than 8. This is not shown since only the "balanced-cases" are illustrated. For the 2.0 and 5.5. Mb/s balanced-cases bursting is less useful given the lower first-hop limits. Still, in unbalanced cases, where some RAPs may serve less than 7 users, other RAPs at 5.5 Mb/s can benefit from using $B(i) > 1$ with $A(i,j) < 7$

[1]At low SER "balanced-case" values predict the feasibility of other "unbalanced-case" values as mentioned in Section VI-A.

to service up to 7 users with as low an $A(i,j)$ as possible. The setting depends on the SER experienced in the RAP↔GAP link and the load contributed by other RAPs. The feasibility of a given setting is simply a test on corresponding loads over all communication groups.

How such observations are used in practice does however depend on the rate adaptation mechanism. Many systems have existing PHY rate adaptation mechanisms. If these mechanisms maintain a low SER, e.g. a SER$< 10^{-5}$ switching before the sharp drops in voice capacity, the rate adaptation mechanism operates independently and the system could simply set $B(i)$ and $A(i,j)$ according to the minimum SER rate. If there is access to the SER, e.g. though channel strength measurements, etc., and to the adaptation mechanism then there is a 2 to 3 dB SNR region in which capacity can be increased using finer adaptations of $B(i)$, $A(i,j)$ and $R(i)$.

VII. Network Management

Our previous analysis and results so far reveal some important observations in planning and managing two-hop wireless mesh networks. One of the key conclusions we can draw from the results is that given a system with a number of relay nodes, one should not blindly apply a naive load balancing approach to equally share all wireless users among the relays. For instance, suppose we have a network with 10 relay nodes and there are 40 VoIP users to support. Brute-force load balancing would try to assign 4 users per relay. However according to Table II (balanced case), a balanced system can support only 30 voice calls if 10 relay nodes are used. The setting $[N_R, (\bar{A}, \bar{B})] = [10, (3,1)]$ yields the largest capacity with all relays having the same number of users, and putting $\bar{A} = 4$ calls on each of these 10 RAPs is not a feasible scenario. However, our results show that there are other scenarios that can support 40 VoIP users, i.e., $[N_R, (\bar{A}, \bar{B})] = [6, (7,1)], [7, (6,1)], [5, (4,2)], [4, (5,2)], [3, (5,3)]$. Thus a better approach rests on identifying a good selection of N_R before enforcing load balancing.

Another point to note, related to the above example, is that by pro-actively using RAPs one can increase the number of calls a system can support. Indeed, in Figure 2 one can consider a scenario where communication groups cover geographic areas overlapping with the GAP, but use different frequency channels. RAPs therefore act as collectors of many short inefficient transmissions thus reducing the load of transmissions to GAP through traffic shaping.

The use of system load calculation to manage the system depends on the objective. If one wants to minimize the total system load, i.e. min $\sum_{\mathcal{G}} T_{\mathcal{G}}$, then a strategy that limits the number

of hops by admitting users first to GAPs when possible, and then first-hop RAPs, etc., makes sense. For example in Figure 1(a), user1, user2 and user3 would be admitted to either G1 or G2. If one wants to minimize the maximum load on a GAP, then admitting some users to relays first helps. Here if no other users are present in Figure 1(a) then admitting users2&3 to RAP1, using G2↔RAP1 to support an aggregated flow for users2&3, and admitting user1 to G1 could be a good choice. If one wants to minimize $max_{\mathcal{G}} T_{\mathcal{G}}$, then admitting and routing users1&2&3 depends on the existing users and flows in the system. In all cases, the feasibility of a choice of admission/routing for a user can be made by considering all possible aggregation and bursting choices for that admission/routing. This would be followed by testing combinations to find one that satisfies $T_{\mathcal{G}} \leq (D_s - \delta)\ \forall \mathcal{G}$, and that meets additional objectives as stated above.

VIII. Conclusion

In this paper we look at the problem of transmitting VoIP traffic efficiently over multi-hop 802.11 wireless networks. Specifically we look at advantages that may be gained by using aggregation and bursting at relays, and by considering such traffic shaping strategies jointly with PHY rate adaptation, admission and routing. We show that indeed aggregation and bursting at RAPs can substantially increase the number of VoIP clients a network can support. Also depending on the overall goal in managing the network, e.g. maximize the number of clients supportable by the network, minimize the load on the GAP, minimize the sum load across the network, etc., joint consideration of traffic-shaping and admission/routing can yield benefits.

However, one challenge in doing such joint optimizations is the large number of potential parameter and decision combinations that may be considered, even for a moderately sized network. It is often impractical to exhaustively consider all such combinations through simulations. We provide a theoretical framework that can be used to analyze the load on each wireless resource in the network as a function of aggregation, bursting, and PHY rate selections. This framework allows us to explore the potential benefits and VoIP capacity limits. It also provides a tool that can be used to plan and manage the operation of such networks.

Although our simulations focused on balanced scenarios with two-hop topologies, many of the insights gained and methods used can be applied to more general networks and unbalanced cases. One issue we have not considered directly is the coexistence of VoIP with different types of background traffic, though consideration of such traffic can be accounted for as mentioned

in Section III-C. Another extension for future work is direct inclusion of RTS/CTS mechanisms that can be used to reduce losses from collisions, and be used to simplify modeling when there are hidden nodes. Common to all such extensions is the underlying observation that joint optimization of traffic shaping and PHY rate adaptation, and further consideration of the effect of admission and routing, can yield significant benefits in multi-hop 802.11 systems.

REFERENCES

[1] C. Pépin, U. C. Kozat, and S. A. Ramprashad, "A Joint Traffic Shaping and Routing Approach to Improve the Performance of 802.11 Mesh Networks," in *4th Int. Symp. on Modeling and Optimization of Mobile, Ad Hov, and Wireless Networks (WiOPT)*, Boston, MA, April 2006.

[2] "TGn Sync Proposal for 802.11n," `http://www.tgnsync.org/techdocs`.

[3] "WWiSE proposal for 802.11n," `http://www.wwise.org/technicalproposal.htm`.

[4] S. Lawrence, A. Biswas, and A. A. Sahib, "A Comparative Analysis of VoIP Support for HT Transmission Mechanisms in WLAN," in *IEEE Int. Conf. on Dist. Computing Systems Workshops*, June 2007.

[5] C. Liu and A. P. Stephens, "An Analytic Model for Infrastructure WLAN Capacity with Bidirectional Frame Aggregation," in *Wireless Comm. and Net. Conf. (WCNC)*, March 2005, pp. 113–119.

[6] D. Niculescu, S. Ganguly, K. Kim, and R. Izmailov, "Performance of VoIP in a 802.11 wireless mesh network," in *Int. Conf. on Wireless Communications and Mobile Computing*, Vancover, BC, July 2006, pp. 869 – 874.

[7] IEEE Computer Society, "Part 11: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Medium Access Control (MAC) Enhancements for Quality of Service (QoS)," 2005, IEEE Std. 802.11eD13.0.

[8] W. Wang, S. C. Liew, and V. O. K. Li, "Solutions to Performance Problems in VoIP over a 802.11 Wireless LAN," *IEEE Transactions on Vehicular Technology*, vol. 54, no. 1, January 2005.

[9] K. Medepalli, P. Gopalakrishnan, D. Famolari, and T. Kodama, "Voice capacity of IEEE 802.11b, 802.11a and 802.11g Wireless LANs," in *IEEE GLOBECOM*, Dallas, Texas, Nov. 2004.

[10] D. Hole and F. Tobagi, "Capacity of an IEEE802.11b WLAN Supporting VoIP," in *IEEE ICC*, Paris, France, June 2004.

[11] K. Medepalli, P. Gopalakrishnan, D. Famolari, and T. Kodama, "Voice Capacity of IEEE802.11b and 802.11a Wireless LANs in the Presence of Channel Errors and Different User Data Rates," in *IEEE VTC*, Los Angeles, CA, Fall 2004.

[12] J. Zhu and S. Roy, "A 802.11 Based Dual-Channel Reservation MAC Protocol for In-Building Multihop Networks," *Springer Mobile Networks and Applications*, vol. 10, no. 5, 2005.

[13] A. Raniwala and T. Chiueh, "Architecture and Algorithms for an IEEE 802.11-Based Multi-Channel Wireless Mesh Network," in *IEEE INFOCOM*, Miami, FL, March 2005.

[14] R. Draves, J. Padhye, and B. Zill, "Routing in Multi-Radio, Multi-Hop Wireless Mesh Networks," in *ACM Mobicom*, Philadelphia, PA, Sept. 26-Oct. 1 2004.

[15] ITU-T, "Pulse Code Modulation (PCM) of Voice Frequencies," ITU-T Recommendation G.711,1988 (Blue Book), 1993.

[16] S. A. Ramprashad and C. Pépin, "A study of silence suppression and real speech patterns and their impact on VoIP capacity in 802.11 networks," in *IEEE Int. Conf. on Multimedia and Expo*, Beijing, China, July 2007, pp. 939–942.

[17] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE JSAC*, vol. 18, 2000.

[18] IEEE Computer Society, "Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: Higher-speed Physical Layer (PHY) extension in the 2.4 GHz band," IEEE Std. 802.11b-1999 (R2003).

[19] "NS2 802.11e package," `http://www.tkn.tu-berlin.de/research/802.11e\_ns2`.

SLOT δ	SIFS	DIFS	*PHY*	MAC	FCS	IP	UDP	RTP	ACK	Min. Rate	MTU
$20\mu s$	$10\mu s$	$50\mu s$	192 bits (L)	30 Bytes	4 Bytes	8 Bytes	20 Bytes	12 Bytes	14 Bytes	1 Mb/s	2346 Byte

TABLE I

IMPORTANT CONSTANTS FOR IEEE802.11B

N_R	Simulation		Analysis	
	$(\bar{A}, \bar{B})$	Capacity	$(\bar{A}, \bar{B})$	Capacity
1	(8,2) or (2,8)	16	(2,9)	18
2	(5,3) or (3,5)	30	(6,3) or (3,6)	36
3	(5,3)	45	(8,2)	48
4	(5,2)	40	(6,2)	48
5	(4,2) or (8,1)	40	(4,2) or (8,1)	40
6	(7,1)	42	(8,1)	48
7	(6,1)	42	(6,1)	42
8	(4,1)	32	(5,1)	40
9	(4,1)	36	(4,1)	36
10	(3,1)	30	(3,1)	30
11	(2,1)	22	(3,1)	33

TABLE II

VOICE CAPACITY WITH BOTH AGGREGATION AND BURSTING (NEGLIGIBLE CHANNEL IMPAIRMENTS)

CLAIMS

We claim:

1. A method for admitting a user into a multi-hop network, comprising:

dividing the multi-hop network into a plurality of communication groups each corresponding to a shared resource;

for each communication group, evaluating the effect of admitting the user into the communication group, taking into account quality of service considerations; and

admitting or rejecting the new user based on the evaluations of the communication groups.

2. A system having a mechanism for admitting a user into a multi-hop network, comprising:

a plurality of communication groups forming the multi-hop network each corresponding to a shared resource; and

a control entity having (a) means for evaluating, for each communication group, the effect of admitting the user into the communication group, taking into account quality of service considerations, and (b) means for admitting or rejecting the new user based on the evaluations of the communication groups.

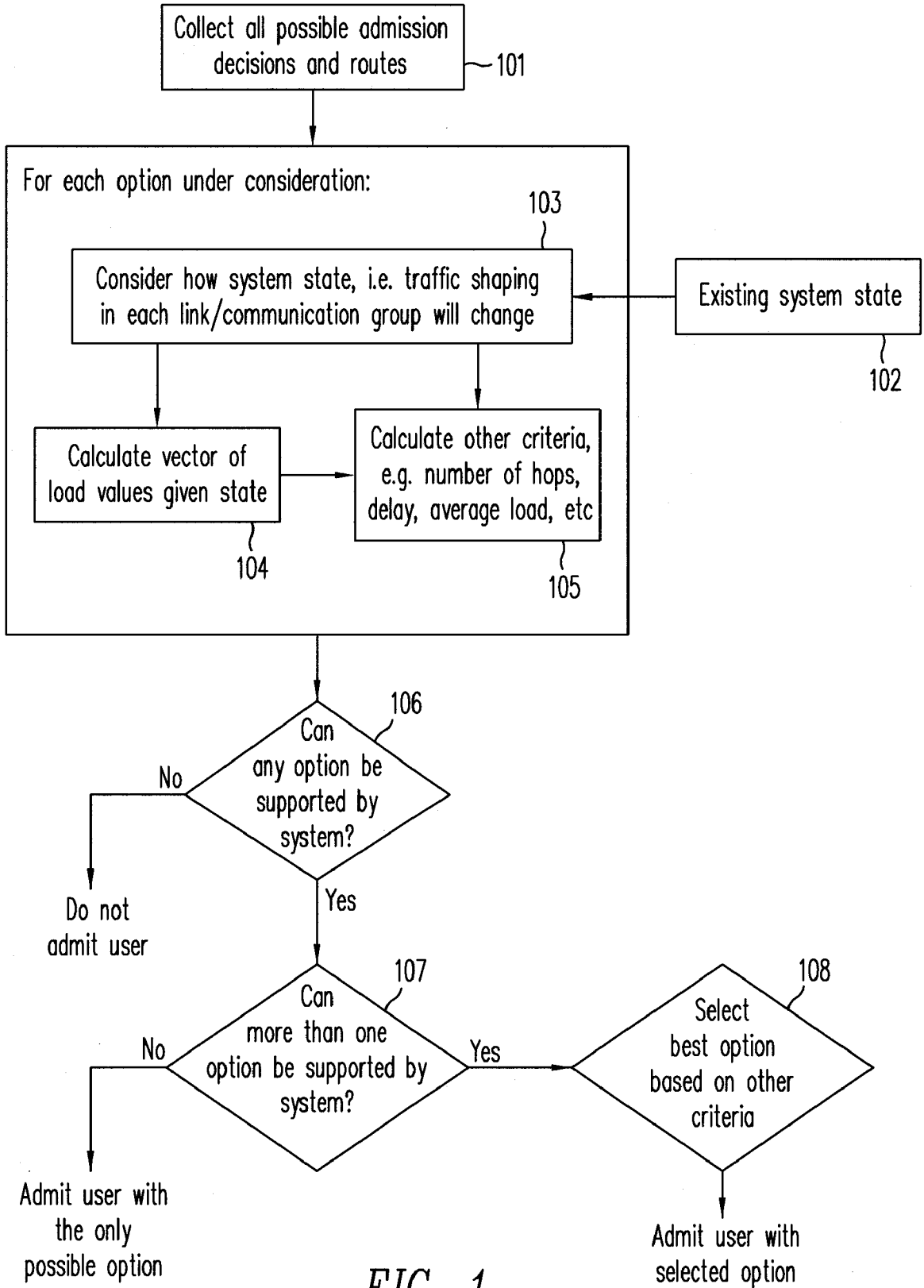

Collect all possible admission decisions and routes
101
For each option under consideration:
103
Consider how system state, i.e. traffic shaping in each link/communication group will change
Existing system state
102
Calculate vector of load values given state
104
Calculate other criteria, e.g. number of hops, delay, average load, etc
105
106
Can any option be supported by system?
No
Do not admit user
Yes
107
Can more than one option be supported by system?
No
Admit user with the only possible option
Yes
108
Select best option based on other criteria
Admit user with selected option

FIG. 1

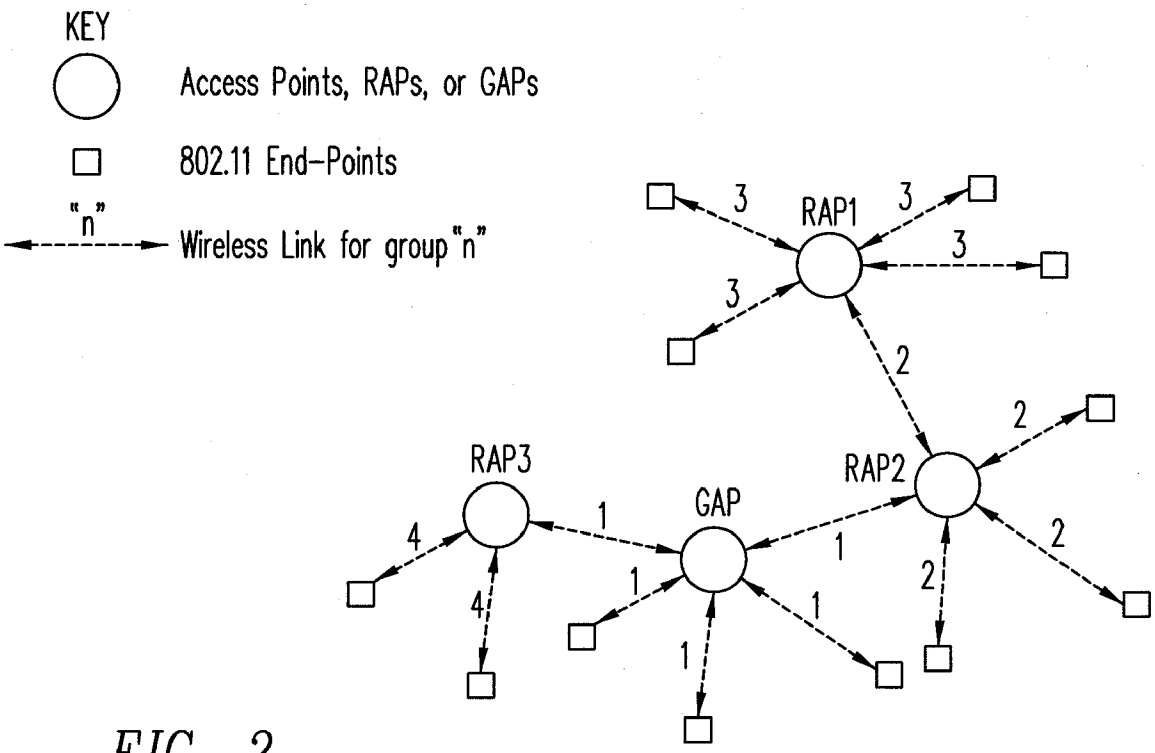

KEY
Access Points, RAPs, or GAPs
802.11 End-Points
"n"
Wireless Link for group "n"
RAP1
RAP2
RAP3
GAP

FIG. 2

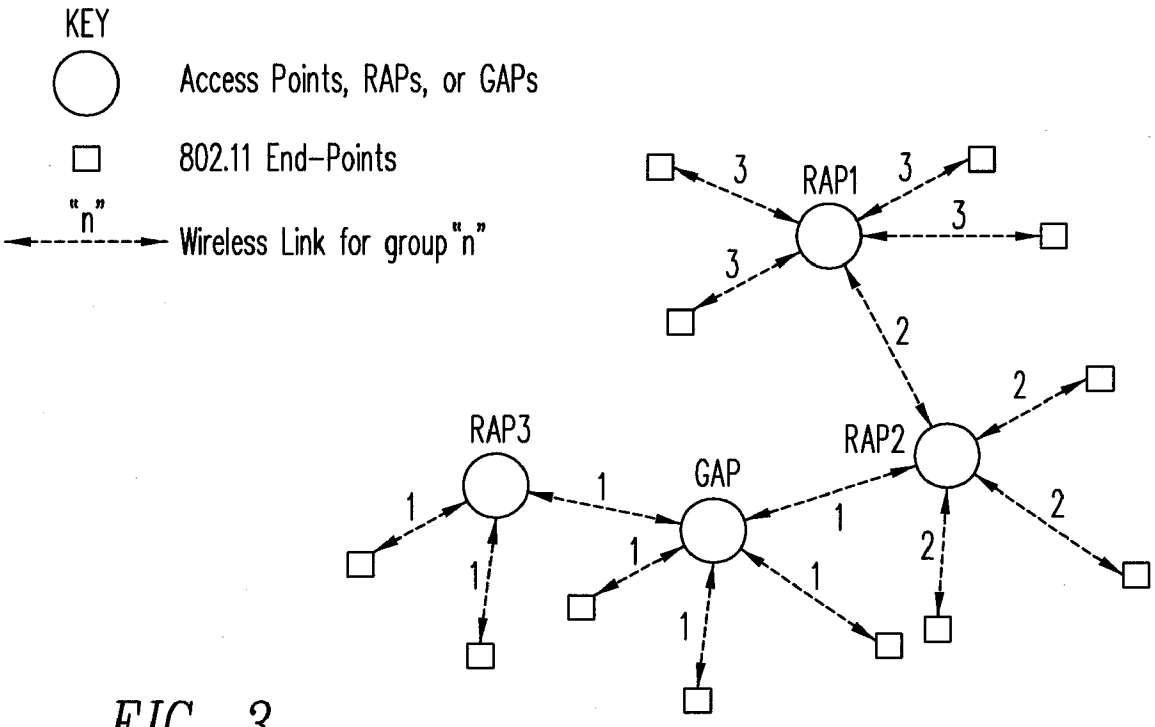

KEY
Access Points, RAPs, or GAPs
802.11 End-Points
"n"
Wireless Link for group "n"
RAP1
RAP2
RAP3
GAP

FIG. 3

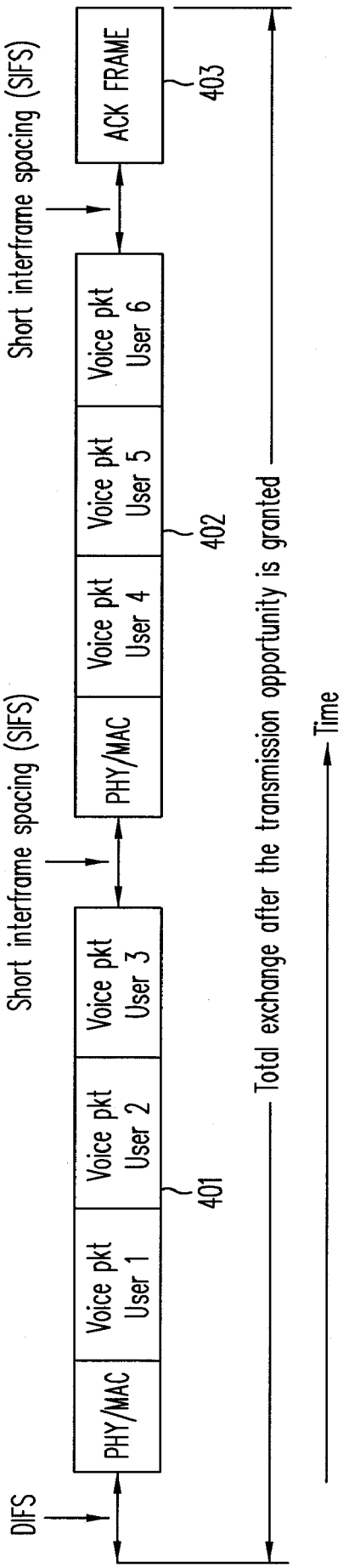

DIFS
PHY/MAC
Voice pkt User 1
Voice pkt User 2
Voice pkt User 3
401
Short interframe spacing (SIFS)
PHY/MAC
Voice pkt User 4
Voice pkt User 5
Voice pkt User 6
402
Short interframe spacing (SIFS)
ACK FRAME
403
Total exchange after the transmission opportunity is granted
Time

FIG. 4

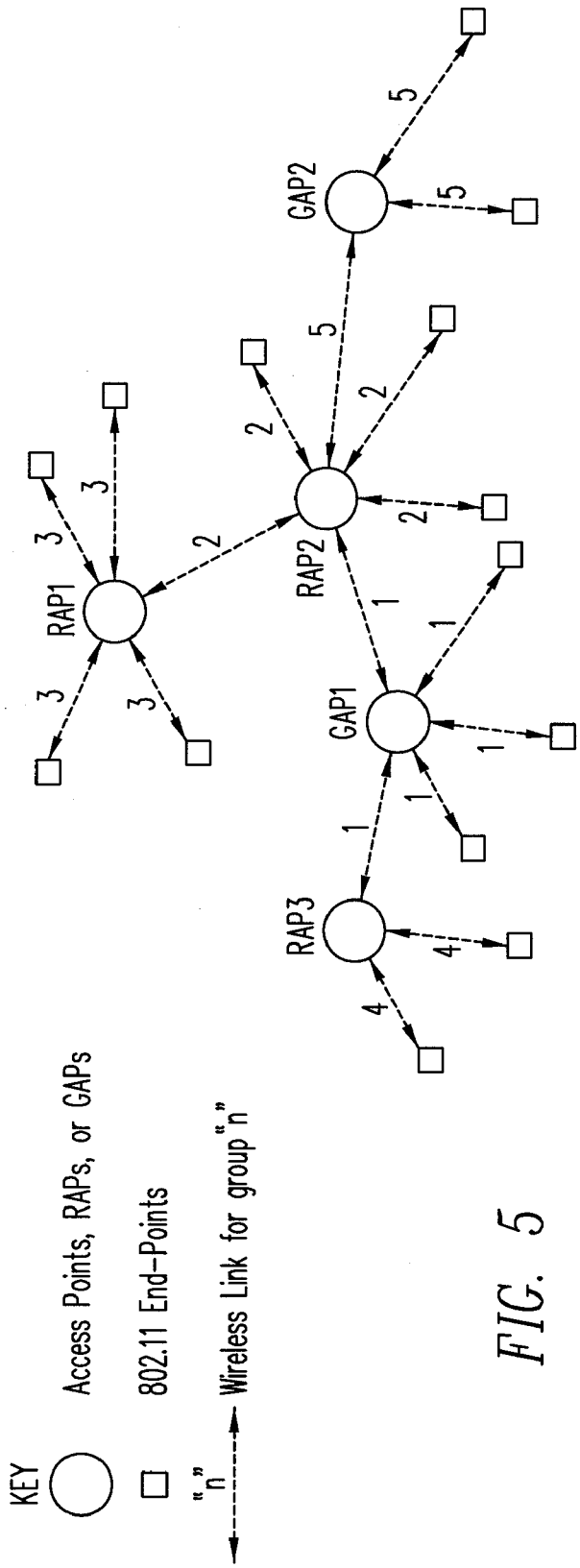

RAP1
RAP2
RAP3
GAP1
GAP2
KEY
Access Points, RAPs, or GAPs
802.11 End-Points
"n"
Wireless Link for group "n"

FIG. 5