

The Effect of First-Hop Wireless Bandwidth Allocation on End-to-End Network Performance

Lili Qiu, Paramvir Bahl, Atul Adya
Microsoft Research
One Microsoft Way, Redmond, WA 98052
{liliq, bahl, adya}@microsoft.com

ABSTRACT

With the increasing popularity of handheld devices and wireless local area networks (LANs), real-time applications such as Internet telephony are poised to become ubiquitous. While there has been a substantial amount of research on quality of service problems in the Internet, most end-to-end bandwidth allocation approaches, such as RSVP, have had limited success due to scalability and deployment issues. Starting with the observation that reserving bandwidth in the Internet backbone requires substantial infrastructure support, but reserving bandwidth in the first hop does not, we only focus on the first-hop reservation. We evaluate several first-hop allocation schemes and determine their effectiveness in improving end-to-end performance. Since utilization of the reserved first-hop bandwidth depends on the remaining Internet path throughput, we characterize this throughput using traces collected from a popular Web site. Our analysis shows that different clients experience widely different throughputs, and that a significant portion of the clients receive very low throughput (e.g. less than 20 Kbps). We then evaluate several bandwidth allocation schemes for various congestion scenarios. Our results show that the scheme which takes into account of both the application data rate and available Internet path bandwidth yields the best performance. Moreover, the scheme performs even better if it adapts to the changing path properties. We discuss how path bandwidth can be measured without active probing, how frequently it needs to be measured, and how this measurement is incorporated into the first-hop bandwidth allocation algorithm.

Categories and Subject Descriptors

C.2.1 [Computer-Communication Networks]: Network Architecture and Design—*Wireless communication*; C.2.5 [Computer-Communication Networks]: Local and Wide-Area Networks—*Internet*

General Terms

Design, Measurement, Performance

Keywords

Bandwidth allocation, wireless QoS, measurement, simulation.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

NOSSDAV'02, May 12-14, 2002, Miami, Florida, USA.
Copyright 2002 ACM 1-58113-512-2/02/0005 ...\$5.00.

1. INTRODUCTION

The ubiquitous availability of the Internet combined with the low cost of using it for long-distance voice and video communications has made IP telephony [8, 7] an attractive alternative to traditional landline and cell-phone based communications. Today, many commercial instant messaging products give users the ability to make voice calls to other users anywhere in the world. Until recently, end devices that allowed IP telephony were fairly cumbersome, e.g., laptops and desktop computers equipped with speakers and microphones. With the advent and growing popularity of IEEE 802.11-based wireless LANs [17] and personal device assistants (PDA) with embedded speakers, a microphone, and sometimes even a miniature camera, IP telephony has received a much needed boost, becoming a serious competitor to cellular and wireless telephony.

However, all is not rosy: wireless LAN based IP-telephony has some problems including bandwidth management in the wireless hop. While the backbone networks in the Internet have become faster, wireless technologies generally used in the first-hop are still slow and can get congested in a public environment (e.g., a mall or a conference) [2]. Consequently, for a wireless LAN based IP-telephony product to be successful, it is important that the limited first-hop bandwidth be managed efficiently. Stated differently, the local network must guarantee a certain level of service quality (i.e., sufficient bandwidth) to the individual user while accommodating a large number of users in the network.

In this paper, we focus on the problem of how to manage bandwidth reservation in the first wireless hop when the remaining path the application traverses uses the Internet best effort service. With many enterprises moving towards wireless LAN connectivity and upgrading their backbone networks to 100 Mbps and higher bandwidths, we feel it is important to focus on the wireless first-hop. Specifically, we are interested in understanding if there is an efficient first-hop bandwidth reservation scheme, and the extent to which the reservation in the first hop improves end-to-end performance.

To answer these questions, we analyze the throughput of Internet paths using traces collected at a popular Web site. Our analysis shows that different clients experience widely different throughput, and a significant portion of clients receive very low throughput (e.g. less than 20 Kbps). This observation suggests that wireless bandwidth allocation needs to take Internet path properties into account.

Next we study several first-hop bandwidth allocation schemes. Using simulations we evaluate their performance under various congestion scenarios. Our results show that the scheme that considers both application data rate and the Internet bandwidth yields the best performance. Moreover, the performance of this scheme can be improved further by making it adaptive, i.e., the allocation varies ac-

Date	Time	# packets	# clients
20 Dec. 2000	6:53 PM - 9:01 PM	100.0 million	134,475
24 Jan. 2001	10:08 AM - 11:21 AM	20.38 million	53,811

Table 1: Summary of the two traces analyzed in this paper.

ording to the changing Internet bandwidth. We also discuss how a passive Internet bandwidth measurement mechanism can be incorporated into the allocation algorithm.

The rest of the paper is organized as follows. In Section 2, we analyze throughput of Internet paths using packet level traces collected at the *microsoft.com* Web site. In Section 3 we describe several bandwidth management techniques. In Section 4, we evaluate the performance of these bandwidth management schemes under different congestion scenarios. We discuss implementation issues in Section 5, and survey previous work in Section 6. We conclude in Section 7.

2. MOTIVATION: THROUGHPUT OF INTERNET PATHS

In this section, we motivate the need for performing bandwidth allocation on the wireless first-hop based on the available Internet throughput for the connection. We analyze the Internet throughput by using packet level traces collected at the *microsoft.com* Web site via the *tcpdump* tool. We captured the incoming and outgoing Web traffic, software download traffic, and streaming media traffic. Table 1 summarizes the two traces we analyze in this paper.

2.1 Throughput Distribution Across Clients

We begin by asking the question: how often is it the case that an application which traverses the Internet is rate-limited by its Internet path compared to its first wireless hop.

For each TCP connection in our traces, we compute the throughput for every 50 Kbytes sent. That is, we filter out very short connections, where throughput has not stabilized (e.g., the connection is in the TCP slow start phase [13]). Figure 1 shows the CDF of the average throughput experienced by all clients. Different clients experience widely different throughput, from 1 Kbps up to 10 Mbps. It is evident that the Web server was not the bottleneck of the end-to-end performance for most clients. Meanwhile it re-confirms the fact that end-to-end performance varies significantly for different hosts, as reported in [3]. In addition, 33.8% of the clients have throughput less than 20 Kbps, i.e., less than stereo quality audio or video encoding rate. This suggests that the Internet path can often become a bottleneck for real-time applications. For example, if an IP-based wireless phone call is made from a mall to an international site or to a remote home user on a dialup-line, the local available wireless bandwidth may be high but the end-to-end Internet path bandwidth would be very low. This leads us to conclude that it would be useful to consider congestion level of the Internet path when making the bandwidth allocation decision in the first hop. When the Internet path is congested, allocating a lower bandwidth at the first hop would admit more connections without decreasing quality of the existing connections.

2.2 Temporal Stability of Throughput

We now analyze the temporal stability of throughput. First, we study how the throughput varies. For the entire period of a trace, we compute the ratio between the maximum and minimum throughput, and the ratio between the maximum and mean throughput seen by each host. Figure 2 (a) and (b) plot the cumulative distribution of the corresponding ratios, respectively. As the figure shows, when considering the variation between the maximum and mini-

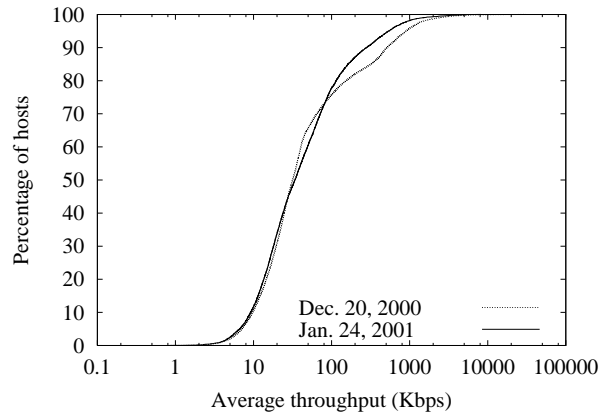
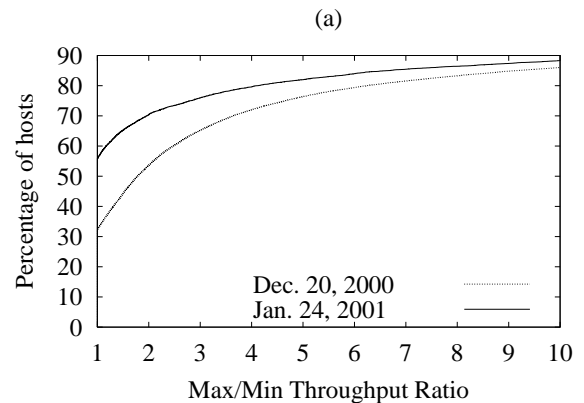
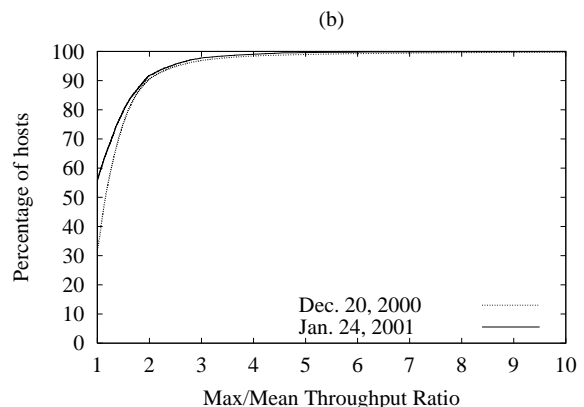


Figure 1: Cumulative distribution of average throughput to different hosts.



(a) CDF of the ratio between the maximum and minimum throughput.



(b) CDF of the ratio between the maximum and mean throughput.

Figure 2: Variation of Internet throughput seen by different hosts.

imum throughput, around 70% of the hosts have throughput variation within a factor of 2, and around 80% of the hosts have throughput variation within a factor of 4 during the trace period; when considering the variation between the maximum and mean throughput, over 90% of the hosts have throughput variation within a factor of 2. These results are consistent with the previous studies, which reported similar degree of stability [3, 21]. On the other hand, throughput for some hosts vary by up to three orders of magnitude.

Next, we examine how long a client's throughput remains stable. We use the notion of operational throughput constancy introduced in [21]. We say that the throughput remains operationally constant in a region when the ratio between the maximum and minimum observed values is less than a factor of ρ . Figure 3 shows the distribution of the size of the maximum steady regions when ρ varies from 1.2 to 20 for the two traces. From Figure 3, we observe that the bandwidth remains steady on the time scale of minutes.

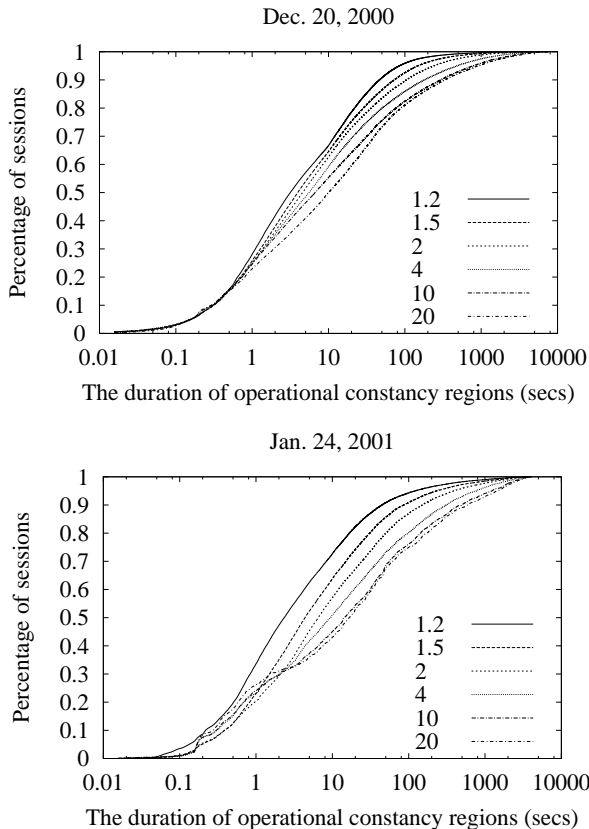


Figure 3: Cumulative distribution of operational constancy regions when ρ varies from 1.2 to 20.

Note that since our throughput estimation is based on short Web transfers, the stability region we are able to detect is limited by the duration of a client's browser session. In other words, we tend to under-estimate the duration for which throughput remains stable. For example, suppose a client's bandwidth to the Web site is stable for several hours, but the client connects to the Web site for 3 minutes only, then the maximum stationarity period that can be determined using the Web trace is 3 minutes.

Applying the heuristic described in [1], we find that 90% of the sessions last less than 1000 seconds (i.e. around 16 minutes), where a session refers to a sequence of requests initiated by a user without prolonged pauses in-between. As a result, our analysis may be conservatively biased towards shorter stability periods. However,

longer stability periods do not have a negative impact on the performance of the bandwidth reservation schemes described in Section 3. In any case, we can conclude that there exists several minutes of throughput stability. This observation suggests that we can use past throughput information to guide future adaptation.

To summarize, our analysis shows that throughput varies widely across clients on the Web site. Moreover, a significant portion of clients experience low throughput. This occurs most likely due to the slow dialup lines and international links, and implies that it is useful to incorporate the Internet bandwidth information when making bandwidth allocation decision at the first wireless hop. Furthermore, as the previous studies showed, we also observe that throughput remains stable on the time scale of minutes. This suggests that we can use the past throughput to predict the future and allocate bandwidth accordingly.

3. BANDWIDTH ALLOCATION

We now explore bandwidth allocation techniques that try to utilize the available bandwidth as judiciously as possible. Our motivation is that it is not efficient to reserve more wireless bandwidth than what an application can use. An application may end up using less wireless bandwidth than it specifies either because it generates data at a slower rate or because the bottleneck is at other links (e.g., Internet path has lower available bandwidth). To address this issue, we propose to *passively* monitor the throughput of applications. For those applications that have significant reserved bandwidth left unused, we adjust the allocated bandwidth according to their usage.

With this in mind, we consider the following reservation schemes:

1. No reservation: best effort.
2. *R0*: reserve at the rate that the source specifies. For CBR (constant bit-rate) traffic, it is the actual data rate, and for VBR (variable bit-rate) traffic, it is the average data rate, assuming the average data rate is available.
3. *R1*: reserve at $\min(s, f * I)$, where s is the rate specified by the source, I is the throughput of the Internet path, and f is a tolerance factor that takes into account of the error in prediction of Internet path's bandwidth. Note that *R0* is a special case of *R1* in which f is ∞ . Unless otherwise specified, we use $f = 1$.
4. *R2*: same as *R1*, except this scheme periodically re-adjusts its allocation. We denote *period* as the time interval in which allocation is re-adjusted. Unless otherwise specified, we use $f = 1$, and *period* = 60 seconds.

4. PERFORMANCE EVALUATION

In this section, we use extensive simulations in the *ns-2* network simulator [16] to evaluate the performance of the different bandwidth management schemes for the first wireless hop. We first discuss our simulation methodology, and then present results for various simulation scenarios.

4.1 Simulation Setup

For our simulation we use the network topology shown in Figure 4. An Internet path is modeled as a single link whose bandwidth is equal to the throughput of the path. The senders employ TFRC, a TCP-friendly congestion control scheme for real-time applications; the receivers periodically report their perceived round-trip time and loss rate back to the senders [10]. This information is used for rate adaptation. In the reservation-based schemes, the bandwidth of the first hop link between the sender and the wireless access point (AP) is equal to the reservation rate honored by the AP. In the best effort

scheme, the bandwidth is set to 100 Mbps, since in this case the throughput is only limited by the bandwidth of the physical link between the source and access point, as well as the Internet path, but not by the reservation rate (since there is no reservation in the best effort case).

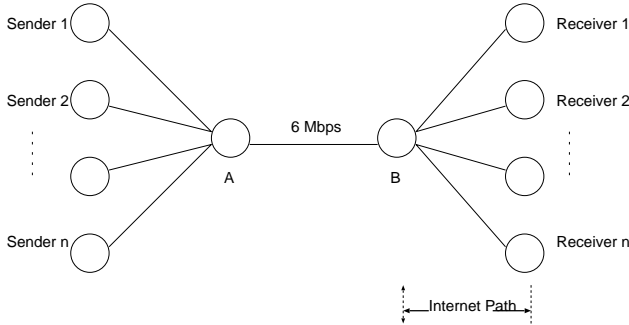


Figure 4: Simulation topology.

4.2 Scenario 1: Congestion at the first wireless hop

In this scenario we examine the case where congestion occurs only at the first wireless hop, and the Internet paths have sufficient bandwidth. In our simulation, we set the sender's desired audio packet rate to 48 Kbps. Figure 5 shows that initially when the number of connections is small (fewer than 125), the average throughput per connection is around 48 Kbps for all the schemes. As the number of connections increases beyond the capacity of the first hop link, the performance of connections begins to degrade in the best effort scheme, whereas all the reservation schemes are able to sustain close to 48 Kbps throughput for the admitted connections. Of course, this performance comes at the expense of denying additional connections. So we conclude that a reservation-based scheme can effectively provide QoS when the congestion occurs in the portion of the path over which it has control.

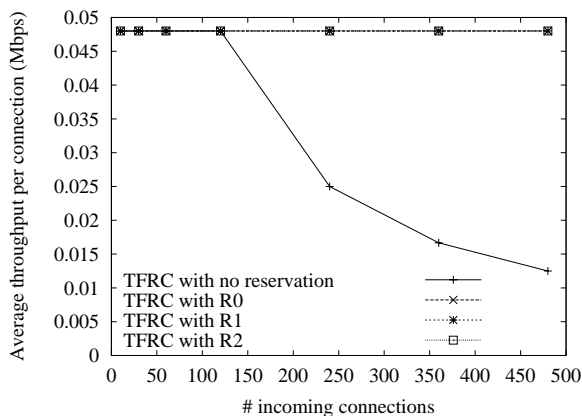
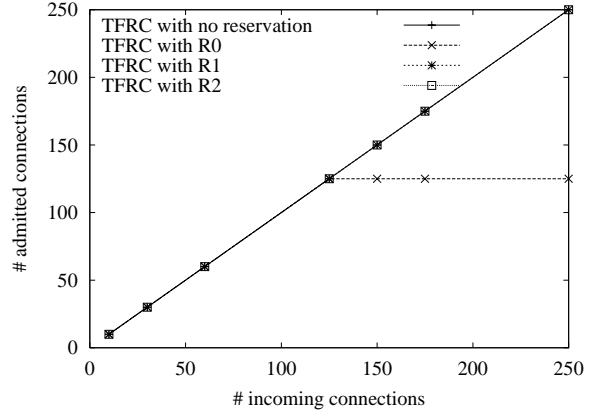


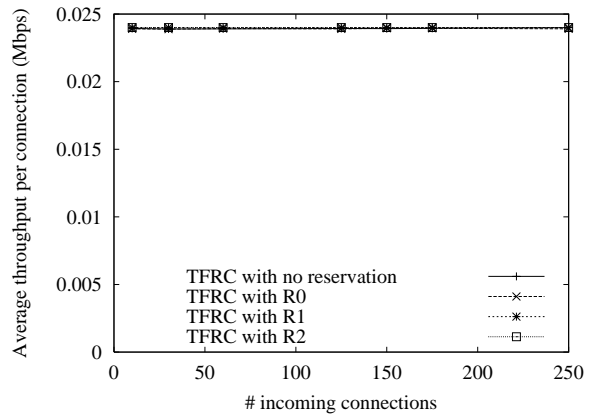
Figure 5: Scenario 1: The desired sending rate of all sources is 48 Kbps, and all Internet paths have 1 Mbps.

4.3 Scenario 2: Congestion at the Internet path

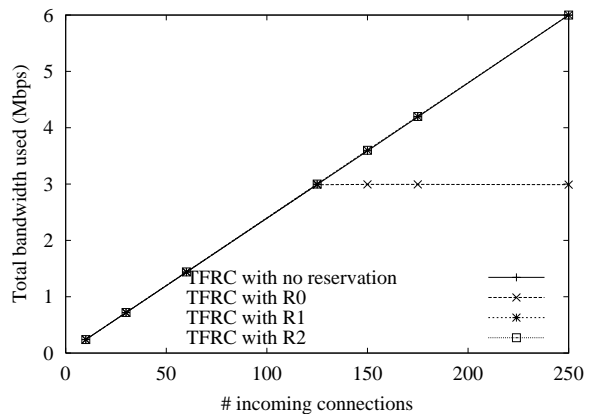
We now study the other extreme, where the congestion occurs on the Internet path, over which the first-hop reservation has no control. As before, the desired sending rate is 48 Kbps. We set the bandwidth of the Internet path to 24 Kbps. As Figure 6 shows,



(a) The number of admitted connections.



(b) The average bandwidth of admitted connections.



(c) The utilization of the wireless link.

Figure 6: Scenario 2: The desired sending rate of all sources is 48 Kbps, and all Internet paths have only 24 Kbps bandwidth.

the quality of the admitted connections is similar across all the schemes: all receiving around 24 Kbps, the bandwidth of the Internet paths. On the other hand, the number of admitted connections are the same for all the schemes, except for reservation scheme R0. This is because R0 allocates 48 Kbps for each connection, even though all the connections use less than that due to the congestion at the Internet. Thus, R0 is sub-optimal since it results in unnecessarily denying connections. This is also evident from Figure 6(c), where connections are denied even though the link is only half utilized (3 Mbps out of 6 Mbps are used). Based on the results, we can see that when congestion occurs in the portion of the paths over which the reservation scheme has no control, reservation cannot help improve QoS. A naive reservation scheme, such as R0, unnecessarily denies connections without improving the performance of the admitted connections. However, the reservation schemes that take into account of Internet congestion information, such as R1 and R2, yield similar performance as the best-effort scheme, all out-performing R0.

4.4 Scenario 3: Congestion at both the Internet and the first wireless hop

Let's turn to the final scenario, where some connections are rate limited by the Internet paths, and others are rate limited by the first wireless hop.

4.4.1 Fixed Internet Bandwidth

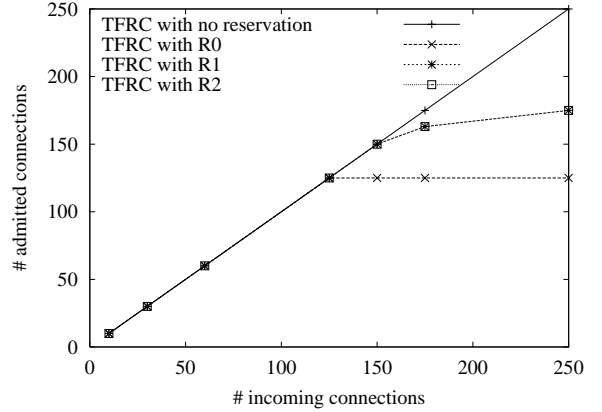
In our first experiment, all sources send 48 Kbps audio packets as before, and the bandwidth of an Internet path is randomly assigned to be either 24 Kbps or 96 Kbps. Figure 7 shows the results. As we can see, initially when the wireless link is lightly loaded, all the schemes perform similarly. As the number of connections increases, R0 starts to deny connections, and some of them are denied even though the link is not fully utilized as shown in Figure 7(c). As the number of connections increases further and fully loads the wireless link, R1 and R2 also start to deny connections. In contrast, the best effort scheme does not deny any connections at the expense of degrading the performance of existing connections.

In our second experiment, the sources send one of the following audio packets: 16 Kbps, 24 Kbps, 32 Kbps, 48 Kbps. Figure 8 summarizes the results. We use normalized quality, along with the number of admitted connections and utilization of links as our performance metrics. We define the *normalized quality* as the ratio between the actual throughput of the connection versus its desired throughput. The results are similar as before: the schemes R1 and R2 can both admit more connections than R0, while maintaining similar quality. In comparison, the performance of best effort traffic degrades significantly as the wireless link becomes congested.

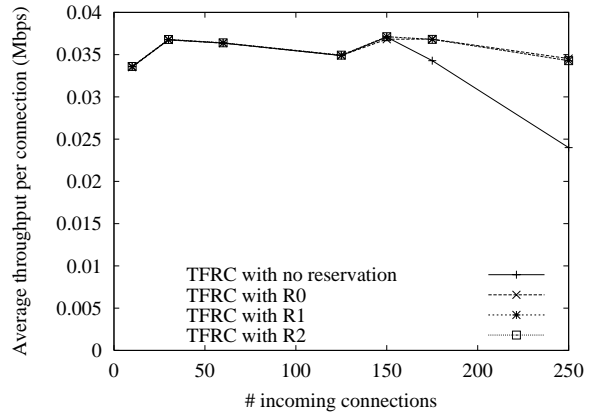
4.4.2 Varying Internet Bandwidth

So far we have fixed the bandwidth of the Internet paths, and also assumed that the senders have precise knowledge of the Internet throughput. In the following experiments, we use the throughput in the real Internet trace for our simulation. In particular, we randomly pick hosts from the Dec. 2000 throughput trace, and assign their perceived throughput to the bandwidth of the Internet paths in the simulation topology. The bandwidth of the links vary according to the trace. The schemes R1 and R2 estimate the initial throughput for the Internet path as the average of the first twenty throughput samples for the host. The desired sending rate of a source is either one of the above five CBR rates, or the rate of the video traces we collected, where Table 2 shows statistics of the video traces.

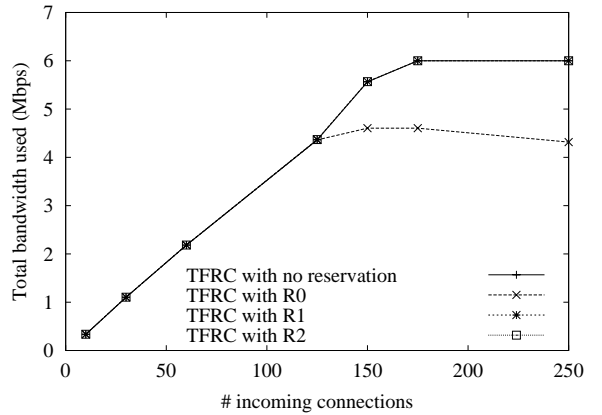
In our first experiment, connections arrive and depart according to a Poisson distribution. The average connection duration is 8



(a) The number of admitted connections.



(b) The average bandwidth of admitted connections.

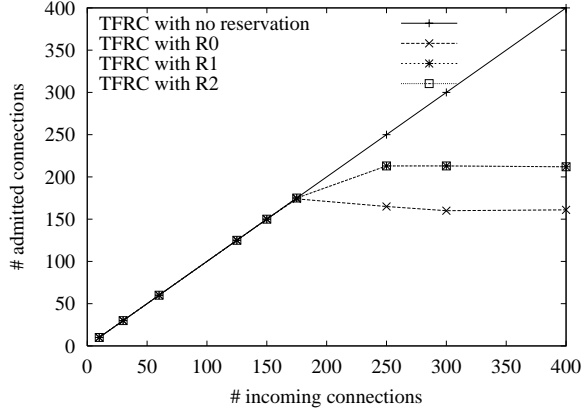


(c) The utilization of the wireless link.

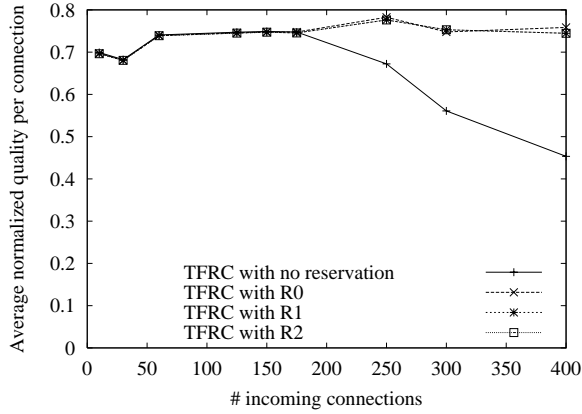
Figure 7: Scenario 3: The desired sending rate of all sources is 48 Kbps, and Internet paths have either 24 Kbps or 96 Kbps bandwidth.

Trace	mean rate (Kbps)	max rate (Kbps)	min rate (Kbps)
1	37.50	125.52	5.66
2	16.85	97.84	0.22
3	46.95	136.01	19.80
4	37.52	100.02	4.77
5	37.52	88.51	14.23

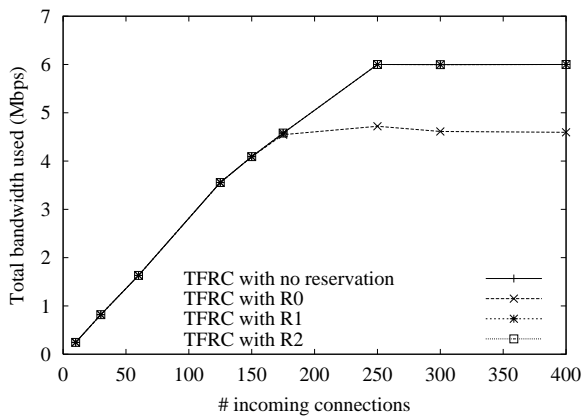
Table 2: Statistics of the video traces used in the simulations.



(a) The number of admitted connections.



(b) The normalized quality of admitted connections.



(c) The utilization of the wireless link.

Figure 8: Scenario 3: The desired sending rate of a source is one of the followings: 16 Kbps, 24 Kbps, 32 Kbps, 48 Kbps, and 64 Kbps, and Internet paths have either 24 Kbps or 96 Kbps bandwidth.

minutes, corresponding to the average duration of a phone call [12], which is one of our target applications. Figure 9(a) shows the number of connections admitted versus time, and Figure 9(b) shows CDF of the normalized quality of admitted connections.

We make the following observations. First, reservation schemes provide much better quality than the best effort scheme. For example, the median normalized quality is 0.46 in the best effort case, and is 0.88 when R0 is used.

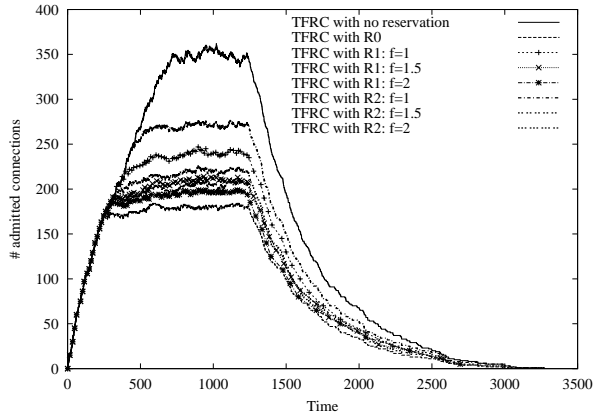
Second, both R1 and R2 can achieve performance similar to that of R0, while admitting more connections.

Third, by choosing different tolerance factor used in R1 and R2, we can trade off the number of admitted connections for better quality. For example, when tolerance factor is 1, more connections are admitted, but the performance is noticeably worse than the case when the tolerance factor is 2. To account for error in measurement and prediction of the Internet bandwidth, a tolerance factor greater than 1 should be used. This also makes it possible to increase allocation at the wireless hop when congestion in the Internet path gets alleviated. In contrast, when the tolerance factor is 1, the bandwidth allocation can only decrease since the connection cannot send more than what has been initially allocated at the wireless hop. In our simulations, we find that a tolerance factor in the range (1, 2] provides a good tradeoff between the number of admitted connections and their quality.

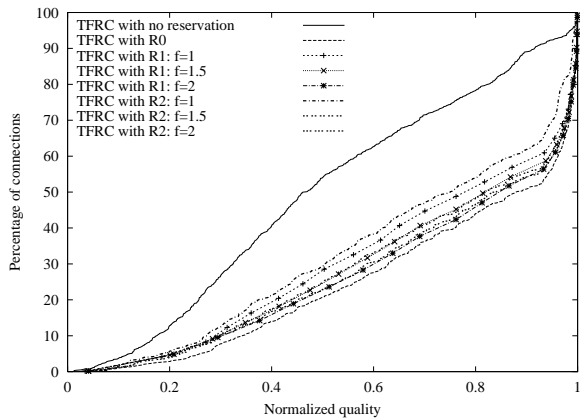
Finally, the non-adaptive scheme performs slightly worse than the adaptive scheme on average, except when the tolerance factor equals to 1. (When the tolerance factor is equal to 1, the adaptive scheme can only reduce its allocation during the session, and possibly decrease the quality of the connection.) Simulations using other adaptation periods ranging from 5 seconds to 10 minutes yield comparable performance. This suggests that in many cases the Internet throughput does not fluctuate so quickly that necessitates fine-grained adaptation, and adaptation on the time scale of minutes is sufficient.

On the other hand, for those hosts that do experience large fluctuations in throughput, the non-adaptive schemes perform poorly. In comparison, the adaptive scheme helps to improve throughput of such hosts by over 40%. (Note that when congestion on an Internet path is alleviated in the middle of a session, it is possible that the wireless link is already fully loaded. In that case, we do not increase wireless bandwidth allocation for that connection.) In addition, the adaptive schemes eliminate the need to pro-actively measure the network throughput. Instead we only need to passively monitor existing connections and adapt bandwidth allocation according to the throughput they have seen recently. Thus, we not only avoid the probing overhead, but also reduce the setup latency that is incurred in R1, since R1 measures the throughput before allocating bandwidth whenever it lacks recent information about the bandwidth to the destination host.

We conduct a second experiment in which the duration of all connections is fixed at 1 hour. Figure 10(a) shows the number of admitted connections versus the number of incoming connections, and Figure 10(b) and (c) show the CDF of the normalized quality of admitted connections when the number of incoming connections are 300 and 400, respectively. The results are qualitatively the same as above. Notice that even when the connection duration is increased to one hour, most connections achieve similar performance regardless of whether adaptation of bandwidth allocation occurs during the session or not. On the other hand, adaptation helps increase some connections' normalized quality by up to 73%.



(a) The number of admitted connections versus time.



(b) CDF of the normalized quality of admitted connections.

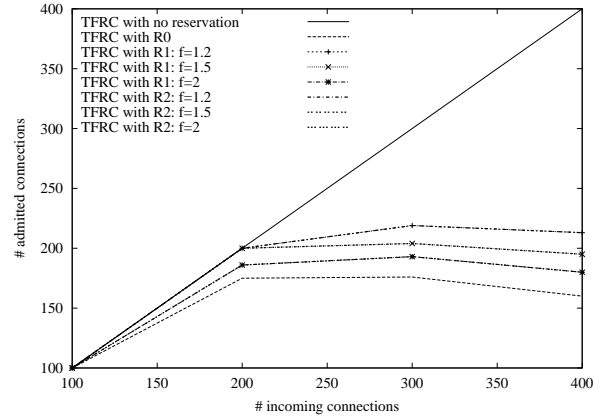
Figure 9: Scenario 3: The desired sending rate of a source is any of the above five audio packet rates or the trace-based video traffic rate; and the bandwidth of the Internet paths are assigned according to the Dec. 20, 2000 packet trace. Connection duration is exponentially distributed with average of 8 minutes.

5. IMPLEMENTATION ISSUES

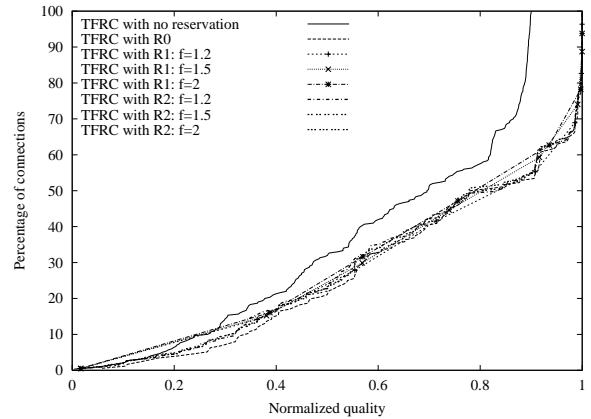
We now briefly discuss how one would implement a first-hop bandwidth allocation scheme. The bandwidth allocation scheme is essentially made up of two components: an *Access Server* and an *Allocation Server*. The Allocation server monitors the Internet throughput periodically, and re-adjusts the reservation accordingly; and the Access Server polices users as appropriate.

Figure 11 illustrates the three different ways in which a network designer might consider implementing the schemes described in this paper. Option 1 is to implement the Allocation Server in the network, and implement the Access Server in the client; option 2 is to implement both servers in the wireless AP; and option 3 is to implement them on a computer acting as a gateway between the wireless subnet and the rest of the network. Options 2 and 3 are similar in flavor. In both cases, the Allocation Server accepts client requests and reserves wireless bandwidth for the first hop.

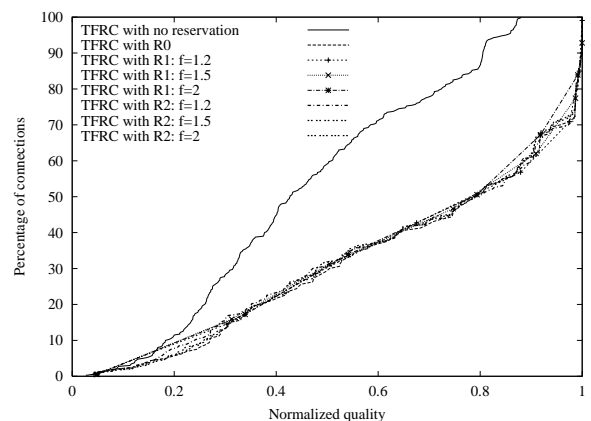
In option 1, the monitoring and policing mechanism are built into the client. A piece of software residing inside the client QoS scheduler monitors the end-to-end bandwidth, and re-adjusts its reservation with the network accordingly. As a result, option 1 requires that the clients sharing an access point behave in a cooperative fash-



(a) # admitted connections vs. # incoming connections.



(b) CDF of admitted connections' normalized quality ($Conn = 300$).



(c) CDF of admitted connections' normalized quality ($Conn = 400$).

Figure 10: Scenario 3: Same as Figure 9 except that the duration of all connections is fixed to be 1 hour, where $Conn$ is the number of incoming connections.

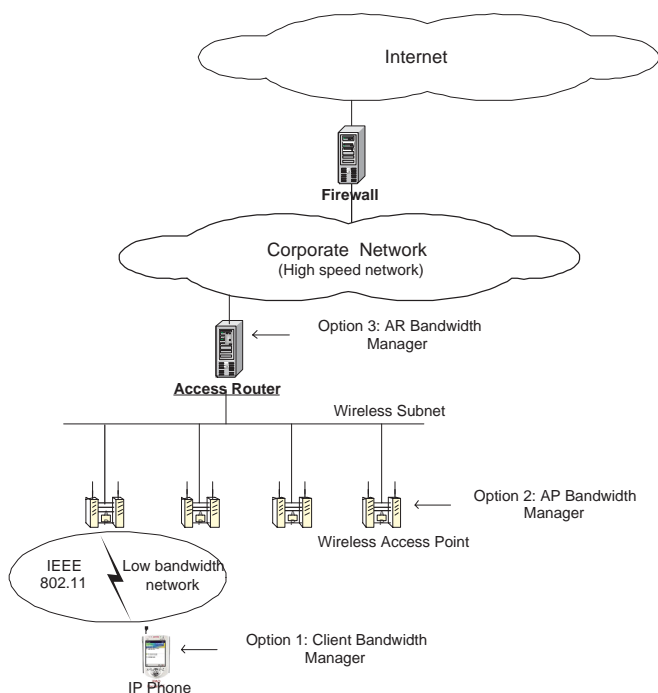


Figure 11: Different options for placing the Bandwidth Manager.

ion (i.e., they are not malicious). In contrast, options 2 and 3 do not require clients cooperation. On the other hand, option 1 has the advantage that it is easier to upgrade client software (e.g., by a software patch download) than to upgrade the firmware of wireless access points.

6. RELATED WORK

There are several studies on characterizing Internet path properties. For example, Balakrishnan *et al.* [3] used the 1996 Olympic Games Web traces to analyze the spatial and temporal stability of TCP throughput. The authors reported that throughput to a client tends to remain stable (i.e., within a factor of 2) for many tens of minutes. Zhang *et al.* [21] studied the stability of Internet path properties using traces collected from the NIMI infrastructure. As in [3], they observed that throughput remained stable on the time scales of minutes. Our trace analysis complements the existing studies, and reports similar degree of temporal stability in throughput.

In the past decade, there has been significant research work on admission control, and many protocols have been proposed. For example, Zhang *et al.* [20] proposed RSVP, a signaling protocol to reserve resources at all the routers along the path. While RSVP provides guaranteed quality of service (i.e., integrated service), it has significant scalability problems, and has not been widely deployed in today's Internet. Instead of providing such a hard performance guarantee, several measurement-based admission control algorithms [11, 14, 6, 9] have been proposed to provide a soft guarantee. Allowing occasional performance degradation enables more efficient utilization of network resources while still providing acceptable service.

There have been several recent proposals on endpoint admission control that use the differentiated service at the backbone and add control algorithms at the endpoints to provide real-time services. Breslau *et al.* [5] gave an extensive evaluation of these schemes.

Like endpoint admission control schemes, we also perform admission control using passive measurement at or close to the sender (i.e., at the wireless access point). However, our work differs in that we consider the case in which backbone only provides best-effort service, not even the services provided by DiffServ (as a result, we provide weaker QoS guarantees). While the reservation schemes we study cannot provide hard QoS guarantees (since part of the network path has no QoS support), we show that their ability in providing better end-to-end performance and their efficiency in utilizing the link can be enhanced by incorporating the Internet congestion information.

Most wireless QoS work has focused on the MAC layer. For example, IEEE 802.11e adds QoS support to the existing 802.11b and 802.11a standards. Tandagopal *et al.* [15] investigated techniques to achieve MAC layer fairness in shared channel wireless networks. Viadya *et al.* [18] proposed a distributed algorithm for fair scheduling in a wireless LAN. In [4], the authors extended the 802.11 Distributed Coordination Function (DCF) to provide service differentiation for delay sensitive and best-effort traffic.

Subnet Bandwidth Manager [19] is a signaling protocol for RSVP-based admission control over IEEE 802-style networks. It can potentially incorporate the bandwidth allocation techniques discussed in this paper.

7. CONCLUSION

With the lack of infrastructure support for both Integrated Services and Differential Services in today's Internet, providing applications with a better quality of service is an interesting and challenging problem. In this paper, we studied several bandwidth allocation techniques for managing bandwidth in the first wireless hop. We have shown how these schemes perform for various traffic conditions. Our results show that schemes that incorporate Internet path characteristics perform the best. Furthermore, the scheme that adapts to changing path bandwidth yields even better performance. Such an adaptive scheme also has a desirable property that it measures Internet bandwidth without active probing. With the growing popularity of hand-held devices and wireless LANs, providing good network serve guarantees for real-time traffic such as IP-based audio/video telephony will become important, and first-hop reservation schemes are a step towards that goal.

8. ACKNOWLEDGEMENTS

We would like to thank Venkata N. Padmanabhan, Scott Hogan, Rob Emanuel, Chris Darling, and Al Lee for helping us collect the packet traces from the microsoft.com Web site.

9. REFERENCES

- [1] A. Adya, P. Bahl, and L. Qiu. Analyzing Browse Patterns of Mobile Clients. In *Proceedings of SIGCOMM Internet Measurement Workshop 2001*, November 2001.
- [2] A. Balachandran, G. Voelker, P. Bahl, and V. Rangan. Characterizing User Behavior and Network Performance in a Public Wireless LAN. In *To appear in Proc. of ACM SIGMETRICS*, June 2002.
- [3] H. Balakrishnan, S. Seshan, M. Stemm, and R. H. Katz. Analyzing Stability in Wide-Area Network Performance. In *Proceedings of ACM SIGMETRICS '97*, 1997.
- [4] M. Barry, A. T. Campbell, and A. Veres. Distributed Control Algorithms for Service Differentiation in Wireless Packet Networks. In *In Proc. of IEEE INFOCOM*, April 2001.
- [5] L. Breslau, E. W. Knightly, S. Shenker, I. Stoica, and H. Zhang. Endpoint Admission Control: Architectural Issues

- and Performance. In *Proceedings of ACM SIGCOMM '2000*, August 28 - September 1 2000.
- [6] C. Casetti, J. Kurose, and D. Towsley. An Adaptive Algorithm for Measurement-based Admission Control in Integrated Services Packet Networks. In *Workshop for High-Speed Networks*, October 1996.
- [7] J. Davidson, J. Peters, B. Gracely, and J. Peters. *Voice over IP Fundamentals*. Cisco Press, 2000.
- [8] B. Douskalis. Prentice Hall, 1999.
- [9] S. Floyd. Comments on Measurement-based Admissions Control for Controlled-Load Services. *submitted to Computer Communications Review*, July 1996.
- [10] S. Floyd, M. Handley, and J. Padhye. Equation-Based Congestion Control for Unicast Applications. In *Proceedings of ACM SIGCOMM '2000*, August 28 - September 1 2000.
- [11] R. J. Gibbens, F. P. Kelly, and P. B. Key. A Decision-Theoretic Approach to Call Admission Control in ATM Networks. *IEEE Journal of Selected Areas in Communication*, pages 1101 – 1114, August 1995.
- [12] E. Heitfield and A. Levy. Parametric, Semi-parametric and Non-Parametric Models of Telecommunications Demand: An Investigation of Residential Calling Patterns. In *Information Economics and Policy*, 2001.
- [13] V. Jacobson. Congestion Avoidance and Control. In *Proceedings of ACM SIGCOMM*, 1988.
- [14] S. Jamin, P. Danzig, S. Shenker, and L. Zhang. A Measurement-based Admission Control Algorithm for Integrated Services Packet Networks. *IEEE/ACM Transactions on Networking*, December 1996.
- [15] T. Nandagopal, T. Kim, X. Gao, and V. Bharghavan. Achieving MAC layer fairness in wireless packet networks. In *In Proc. of ACM MOBICOM*, August 2000.
- [16] NS-2 - Network Simulator. <http://www.isi.edu/nsnam/ns/>.
- [17] L. M. S. C. of the IEEE Computer Society. Wireless LAN medium access control (MAC) and physical layer (PHY) specifications. *IEEE Standard 802.11, 1999 Edition*, 1999.
- [18] N. Vaidya, P. Bahl, and S. Gupta. Distributed Fair Scheduling in a Wireless LAN. In *In Proc. of ACM MOBICOM*, August 2000.
- [19] R. Yavatkar, D. Hoffman, Y. Bernet, F. Baker, and M. Speer. SBM (Subnet Bandwidth Manager): A Protocol for RSVP-based Admission Control over IEEE 802-style networks. *RFC 2814*, May 2000.
- [20] L. Zhang, S. Deering, D. Estrin, S. Shenker, and D. Zappala. RSVP: A New Resource ReSerVation Protocol. In *IEEE Network Magazine*, September 1993.
- [21] Y. Zhang, N. Duffield, V. Paxson, and S. Shenker. On the Constancy of Internet Path Properties. In *Proceedings of ACM SIGCOMM Internet Measurement Workshop*, November 2001.