

Towards Provisioning Diffserv Intra-Nets

Ulrich Fiedler, Polly Huang, Bernhard Plattner

Compute Engineering and Networks Laboratory
Swiss Federal Institute of Technology
ETH-Zentrum, Gloriastrasse 35
CH-8092 Zurich, Switzerland
{fiedler, huang, plattner}@tik.ee.ethz.ch

Abstract. The question of our study is how to provision a diffserv (differentiated service) intra-net serving three classes of traffic, i.e., voice, real-time data (e.g. stock quotes), and best-effort data. Each class of traffic requires a different level of QoS (Quality of Service) guarantee. For VoIP the primary QoS requirements are delay and loss; for real-time data response-time. Given a network configuration and anticipated workload of a business intra-net, we use ns-2 simulations to determine the minimum capacity requirements that dominate total cost of the intra-net. To ensure that it is worthwhile converging different traffic classes or deploying diffserv, we cautiously examine capacity requirements in three sets of experiments: three traffic classes in i) three dedicated networks, ii) one network without diffserv support, and iii) one network with diffserv support. We find that for the business intra-net of our study, integration without diffserv may need considerable over-provisioning depending on the fraction of real-time data in the network. In addition, we observe significant capacity savings in the diffserv case; thus conclude that deploying diffserv is advantageous. The relations we find give rise to, as far as we know, the first rule of thumb on provisioning a diffserv network for increasing real-time data.

1 Introduction

Integration of services for voice, real-time data, and best-effort data on IP intra-nets is of special interest, as it has the potential to save network capacity that dominates costs [1]. Moreover, Bajaj et al. [2] suggested that when network utilization is high, having multiple service levels for real-time traffic may reduce the capacity requirement. While Bajaj et al. examined voice and low quality video in a diffserv network, we are interested in voice and emerging transaction-oriented applications. Such applications are web-based and include support for response-time-critical quotes, and transactions on securities, stocks, and bonds. We expect to derive guidelines to provision an IP intra-net with multiple levels of services for such applications.

Recently Kim et al. evaluated provisioning for voice over IP (VoIP) in a diffserv network [3]. They found a roughly linear relationship between the number of VoIP connections and the capacity requirement. The multiplexing gain was in agreement with the telecommunication experience. Although they investigated the effect of competing web traffic on VoIP in diffserv networks, they did not measure performance for any type of traffic other than VoIP.

It is not clear yet how to provision for web-like data traffic. This provisioning problem is of special interest because in the future a portion of this data traffic may be real-time. This work addresses the problem of provisioning web traffic with response-time requirement on diffserv networks that carry, in addition to real-time data, VoIP and other best effort data. Our goal is to measure with simulations the benefit of integration and differentiation of services. Therefore we try to answer, which of the following three cases requires lower capacity.

1. Keeping voice, real-time data, and best-effort data on dedicated IP networks
2. Integrating voice, real-time data, and best-effort data in a conventional IP network without diffserv support
3. Integrating voice, real-time data, and best-effort data in a network with diffserv support for the three classes of traffic

To determine the minimal capacity requirement, we measure quality of VoIP and real-time data. For VoIP, similar to Kim et al. [3], we measure end-to-end delay and loss. For real-time data traffic, we measure response-times of web down-loads.

We observe that case two needs significant over-provisioning compared to case one when the fraction of real-time data is small to medium. If all data traffic is real-time, the capacity requirements for case one and two are not very different. For case three, we find a lower capacity requirement than that of case one. This capacity requirement shows an approximately linear relationship to the fraction of real-time data; more real-time data require more capacity. Compared to dedicated networks, the percentage of capacity savings for integration with diffserv remains approximately constant for a small to medium fraction of real-time data. These results give a guideline on how to provision diffserv networks for web traffic with response-time requirements.

The rest of this report is organized as follows: Section 2 describes the simulation environment, topology, source model and explains measurement metrics for QoS requirements to assess capacity provisioning. In Section 3 we explore capacity provisioning for voice, real-time and best-effort data services on dedicated networks. In Section 4 we present results on provisioning a conventional IP network without diffserv support for these services. In Section 5 we investigate provisioning and capacity savings with diffserv support. In Section 6 we discuss our results, limitations, and further work.

2 Setup

This section introduces the simulation environment, topology, source model, and measurement metrics for QoS requirements to assess capacity provisioning.

2.1 Simulation Environment

We use the ns-2 [4] as our simulation environment. We have augmented ns-2 with diffserv additions by Murphy [5], and scripts that explicitly model the interactions of HTTP/1.1 [6]. We have customized routines for collecting performance statistics.

The diffserv additions model diffserv functionality. Diffserv [7–9] provides different levels of service by aggregating flows with similar QoS requirements. At the network edges, packets are classified and marked with *code-points*. Inside the network, packets are forwarded solely depending on their code-points. We consider three levels of service based on different forwarding mechanisms as specified by the IETF.

- *expedited forwarding (EF)*

This forwarding mechanism implements a virtual wire [10]. It is intended for delay sensitive applications such as interactive voice or video. Wroclawski and Charny [11] propose to implement guaranteed service[12] based on EF. We use it to forward voice traffic.

- *assured forwarding (AF)*

This forwarding mechanism is intended for data that should not be delayed by network congestion [13]. Wroclawski and Charny [11] propose to implement controlled-load service[14] based on AF. We use AF to differentiate real-time data traffic from best-effort data traffic.

- *best effort (BE)*

This forwarding mechanism implements best-effort service. We use it to forward best-effort data traffic.

The diffserv package implements three queues to buffer the packets of these three service classes. The queues are serviced with a deficit round-robin scheduler. We do no conditioning at ingress and configure a single level of drop precedences for AF, i.e. we mark all AF traffic as AF11.

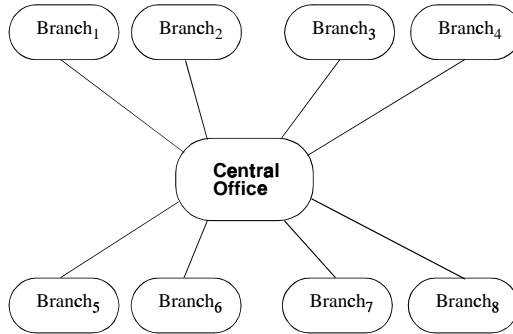


Fig. 1. Intra-Net Topology

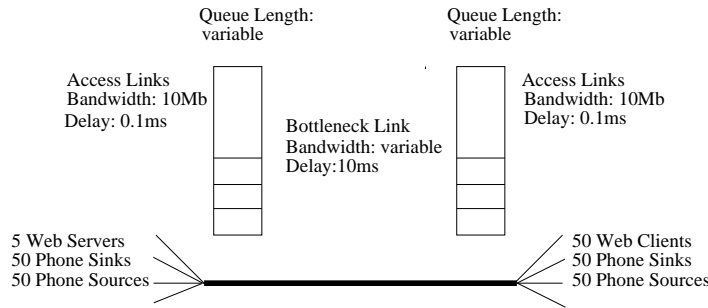


Fig. 2. Simulation Topology

2.2 Topology

For our simulation we choose a simple dumbbell topology as depicted in Figure 2. This topology may e.g. represent the bottleneck link in an intra-net, on which capacity is scarce (see Figure 1). It may also be viewed in a more general sense, since many have argued that there is always a single bottleneck link on any network path [2], which justifies our choice to start experiments with such a network topology. We distribute sources as depicted in Figure 2. We position 50 phones and 50 web clients at the right side of the bottleneck and 50 phones and 5 web servers at left side. Access links have 10Mb in capacity and 0.1ms in delay. The bottleneck link has a propagation delay of 10ms to model the worst case for a connection going through all of Europe. The variables in this simulation are the capacity of the bottleneck link and its queue size. If not explicitly mentioned, queue size is set to 20KB.

2.3 Source Model

We model voice traffic as VoIP. A telephone is a source/sink pair that produces CBR traffic at call times. We assume no compression or silence suppression; the source thus produces a 64Kb stream. As packetization for VoIP is not standardized, we assume that each packet contains 10ms of speech. We think that this is a realistic tradeoff between packetization delay and overhead due to IP and UDP headers. The net CBR rate of a phone is thus 86.4Kb.

To model call duration and inter-call time we use parameters as depicted in Table 1. In residential environments, call durations are simply exponentially distributed with a mean of 3 minutes [15]. However, from our partners we know that a bi-modal distribution for call durations, representing long and short calls, better suits the situation of a busy hour in a large bank. Both long

Table 1. User/Session Related Parameters to Generate Voice Traffic

Parameter	Distribution	Average
Call duration	Bimodal	long call: 20% Short Call: 80%
Long Call	Exponential	8 min
Short Call	Exponential	3 min
Inter-Call time	Exponential	15 min

and short calls, as well as the inter-call time, were represented by exponential distributions. This model offers a mean load of 800Kb on the bottleneck link given that 50 phone pairs interact.

We model data traffic as web traffic. We explicitly model the traffic generated by request/reply interactions of HTTP/1.1 between web server and clients. This modeling of HTTP/1.1 interactions includes features like persistent connections and pipelining. We keep TCP connections at servers open for further down-loads and eliminate the need for a client to wait for the server's response before sending further requests. We model a requested web page as an index page plus a number of embedded objects. To generate the corresponding traffic, we need the distributions for the following entities:

1. size of requested index objects
2. number of embedded objects
3. size of embedded objects
4. server selection
5. think time between two successive down-loads.

Table 2. User/Session Related Parameters to Generate Web Traffic

Parameter	Distribution	Average	Shape
Size of Index Objects	Pareto	8000 B	1.2
Size of Embedded Objects	Pareto	4000 B	1.1
Number of Embedded Objects.	Pareto	20	1.5
Think Time	Pareto	30 sec	2.5

We use parameters as depicted in Table 2. Distributions of object sizes and number of embedded objects were determined with web crawling in a bank's intra-net [16]. Measured distributions basically agree with those of Barford et al. [17,18]. Our model does not address the matching problem, i.e. which request goes to which server. Instead we randomly chose the server for each object. Server latencies were not modeled, as this exceeds the scope of this work. This model produces a mean load of 850Kb on the bottleneck link in direction from servers to clients, given that five web servers interact with 50 web clients on a dedicated network with 2.4Mb provisioned at the bottleneck link. In experiments with integrated networks, we measure slightly more data traffic than voice traffic on the bottleneck link given the number of sources and topology in Figure 2.

2.4 QoS Requirements

To determine the capacity requirement for a given network, we need to define QoS requirements for each type of traffic. We chose these requirements as realistically as possible. For this reason we

were in contact with a large Swiss bank. Thus, the QoS requirements for voice and real-time data are derived from the practice.

For VoIP transmission, we measure end-to-end delay and packet loss. We define three requirements: one for the end-to-end delay and the other two for packet loss. End-to-end packet delay for VoIP consists of coding and packetization delay as well as network delay. Coding and packetization delays can be up to 30ms. Steinmetz and Nahrstedt [19] state that one-way lip-to-ear delay should not exceed 100–200ms based on testing with human experts. In conventional telephony networks, users are even accustomed to much lower delays. For these reasons we define 50ms as an upper limit for end-to-end network delay. The impact of loss to VoIP depends on the speech coder. We assume the ITU-T G.711 coder [20], which is widely deployed in conventional telephones in Europe. To quantify the impact of packet loss to packetized speech coded with this coder, we group successive losses of VoIP packets into *outages*. The notion of outages comes from Paxson who investigated on end-to-end Internet packet dynamics [21]. Such outages are usually perceived as a crackling sound when played out at the receiver's side. We think that number and duration of outages may characterize the impact of loss reasonably well for this coder in situations of low or moderate loss. Since there is no standard on how to assess human perception of VoIP transmission with this coder, we conducted trials in our lab, in which we mapped pre-recorded speech on outage patterns of our simulations. From these trials, we define the requirements on outages for VoIP as follows: The number of outages must not exceed 5 per minute. Outage duration must not exceed 50ms. This approach is similar to the one used by Kim et al. [3].

For real-time data we measure the fraction of response-times of web page down-loads that are below five seconds. We define this fraction as the *five seconds response time quantile*. We define the *response-time* as the time that has elapsed between sending out the first TCP *syn* packet and receiving the last TCP packet containing relevant data. To control QoS for this type of real-time data traffic, banks measure the percentage of data that is received in five seconds. Banks require this quantile to exceed some limit in the very high nineties. After discussion with partners, we set this limit to 99%.

3 Dedicated Networks

We investigate the provisioning for dedicated networks for VoIP and web traffic to be able to assess the benefit in terms of capacity savings when integrating these networks. We determine provisioning by assessing QoS for each type of traffic.

3.1 Voice

For VoIP traffic we measure the capacity needed to accommodate the traffic generated by 50 phones. We measure end-to-end delay, and number and duration of outages at increasing link capacity.

Figure 3 (top) depicts maximum end-to-end delay of VoIP packets. We find that the maximum delay decreases from 145ms at 1.2Mb to 11ms at 1.9Mb. The QoS requirement for end-to-end packet delay, to be less than 50ms, is reached at 1.6Mb. For capacities larger than 1.8Mb, the end-to-end delay equals the physical propagation delay. This means that queues in the system are empty at this capacity. Figure 3 (bottom) demonstrates the number of outages for VoIP on a dedicated network decreases from 25 at 1.2Mb to zero at 1.8Mb. For larger capacities the number of outages remains zero. The QoS requirement for the number of outages, to be less than five per minute, is met at 1.5Mb. At 1.8Mb outages disappear as queues become empty. Figure 4 depicts the maximum outage durations. Maximum outage duration sharply drops at 1.4Mb and becomes zero at 1.8Mb. The QoS requirement for the maximum outage duration, to be less than 50ms, is reached at 1.5Mb. Reviewing all QoS requirements at a time, we find that the end-to-end delay requirement is met at 1.6Mb, whereas the requirements on the number and duration of outages are both met at 1.5Mb.

We summarize our findings on provisioning dedicated network for VoIP:

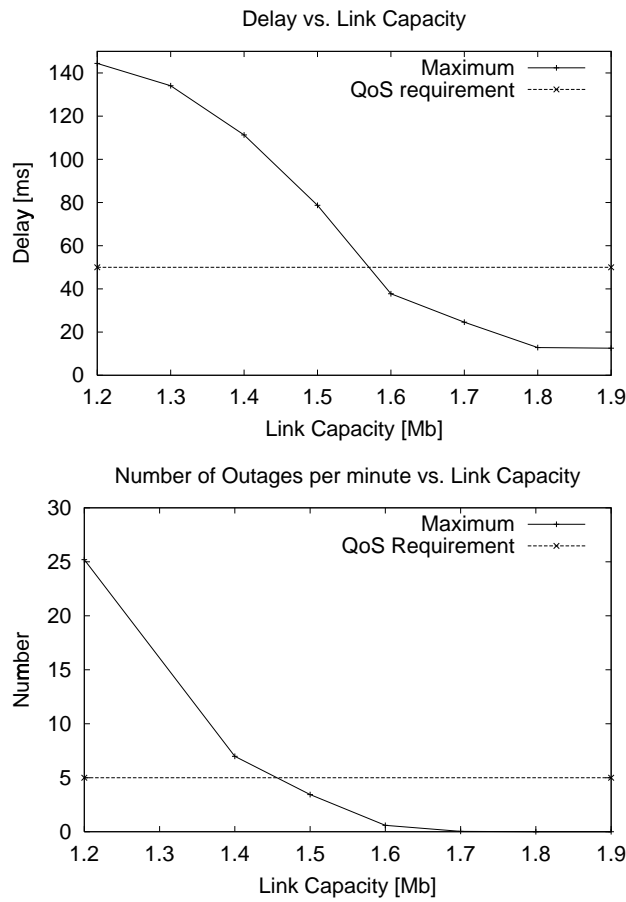


Fig. 3. Packet Delay and Outage Frequency for VoIP on a Dedicated Network

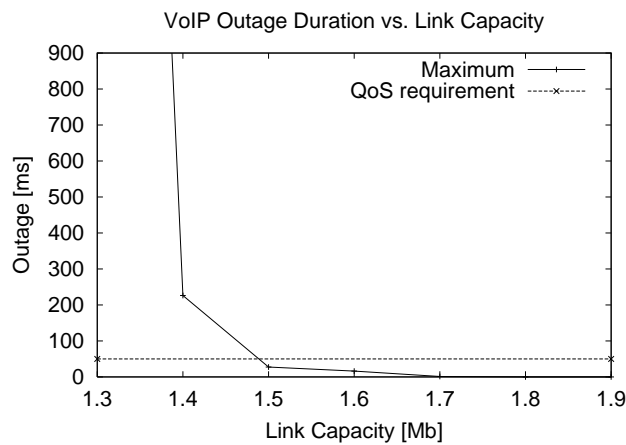


Fig. 4. Outage Duration for VoIP on a Dedicated Network

- Outages and queuing delay totally disappear at 1.8Mb.
- The stringent end-to-end delay requirement of 50ms is the dominant factor that limits provisioning for VoIP in a dedicated network to 1.6Mb.

3.2 Real-Time Data

In this section, we measure the capacity requirement for real-time data traffic on a dedicated network. We model real-time data as web traffic with a response-time requirement.

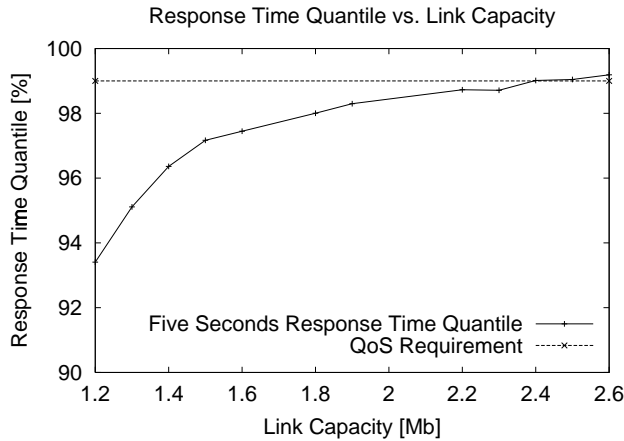


Fig. 5. Five Seconds Response Time Quantile for Web Traffic on a Dedicated Network

We start with all data traffic in our experiment being real-time, i.e. all 50 web clients requesting web pages with a response-time requirement. Figure 5 depicts the fraction of web pages that can be down-loaded within five seconds at increasing capacities. This fraction grows from 93.5% at 1.2Mb to 99.3% at 2.6Mb. The curve is concave from above, which means that additional capacity has a stronger impact on response-times at lower capacities and a smaller impact at higher capacities. The QoS requirement for real-time data, that 99% of the web pages can be down-loaded in less than five seconds, is met at 2.4Mb.

These response-time quantiles are measured at a queue size of 20KB at both ends of the bottleneck link. Since queue size 20KB is already larger than the bandwidth delay product, which is required to make TCP perform at its maximum, varying queue size has not much impact on response-times.

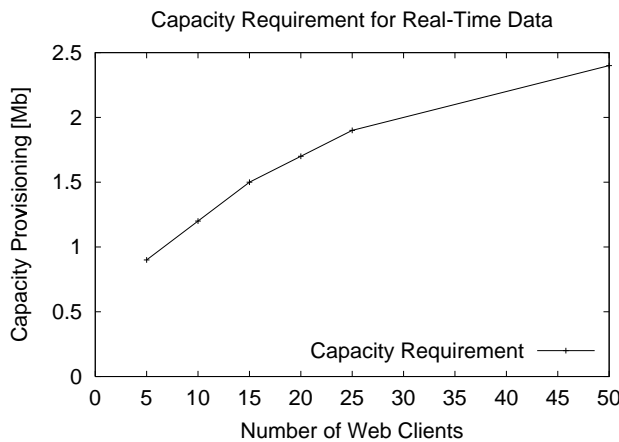


Fig. 6. Multiplexing for Real-Time Data Traffic

As we want to experiment with various fractions of data traffic having a real-time requirement, we investigate the capacity requirement of an increasing number of web clients requesting pages with a response-time requirement. Figure 6 shows that the capacity requirement grows less and less with increasing number of clients. It grows from 0.9Mb for 5 web clients to 2.4Mb for 50 web clients.

We summarize our findings on provisioning dedicated networks for real-time data:

- The capacity requirement for 50 web clients requesting web pages with response-time requirement is met at 2.4Mb.

3.3 Three Dedicated Networks

We determine the capacity requirement for a set of dedicated networks as follows. A dedicated network for VoIP has a capacity requirement of 1.6Mb. To accommodate the real-time data traffic on a dedicated network, we need between zero and 2.4Mb depending on the number of clients requesting web pages with response-time requirement. The remaining best-effort traffic on a dedicated network needs marginal capacity as it has no QoS requirements to meet. We assume that this capacity for best-effort traffic on a best-effort network is zero, which is in favor of the dedicated networks case. Summing up these requirements, we conclude that dedicated networks need between 1.6Mb and 4.0Mb provisioning depending on the fraction of real-time traffic.

4 Integrated Networks

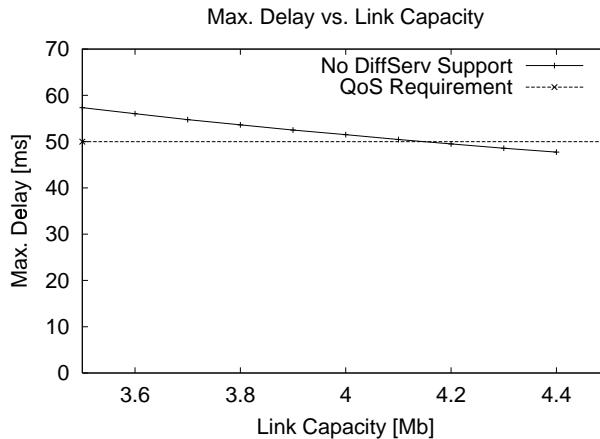


Fig. 7. Maximal end-to-end Delay without Diffserv Support

In this section, we investigate capacity requirements for an integrated network without diffserv support. A network without diffserv support cannot differentiate between best-effort and real-time data. Therefore, we have to over-provision the network such that VoIP meets its QoS requirements and all web traffic, best-effort and real-time, meet the response-time requirement.

For VoIP packets we measure a maximum end-to-end delay of 58ms at 3.5Mb that linearly decreases with increasing capacity to 45ms at 4.4Mb (Figure 7). In contrast to the dedicated networks case, these values remain considerably high. To measure the impact of loss for VoIP, we depict number and duration of outages at increasing capacity. Figure 8 (top) shows that outages start at 5.5 per minute for 3.5Mb and slowly decrease to 2.8 per minute at 4.5Mb. The QoS requirement of a maximum of five outages per minute is met at 3.6Mb. Outage durations (see

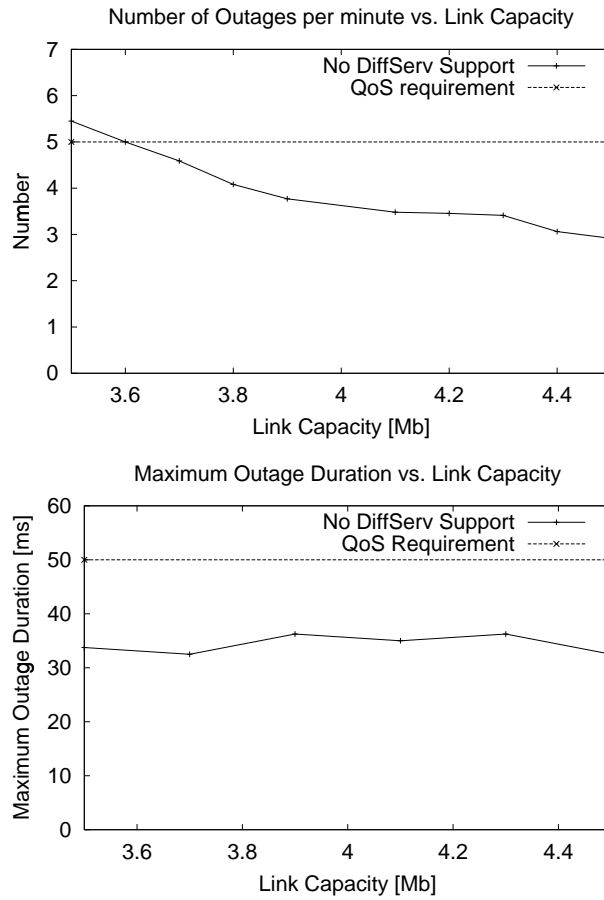


Fig. 8. Number of Outages and Outage Duration without DiffServ Support

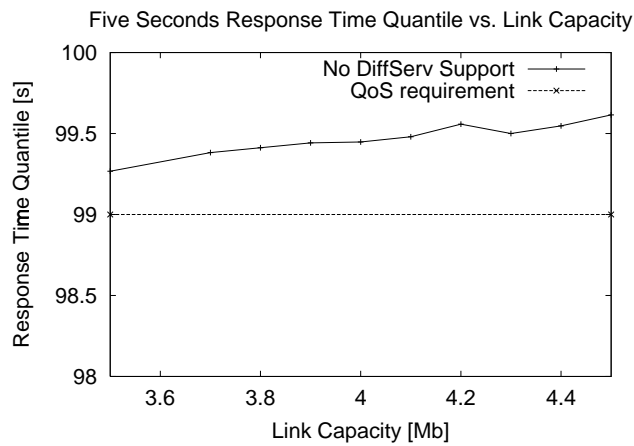


Fig. 9. Five Seconds Response Time Quantile without DiffServ Support

Figure 8, bottom) show no clear trend. They fluctuate between 30ms and 40ms in the studied interval between 3.5Mb and 4.5Mb. This is below the QoS requirement of 50ms. Other than in the dedicated networks case, there is no capacity at which VoIP delay sharply drops and outages

disappear. We presume that this effect comes from the bursty nature of competing web traffic, which cannot be alleviated in the studied range of capacities.

For data traffic, we monitor the response-time of all web pages, as a network without diffserv support cannot differentiate between real-time and best-effort data. Figure 9 depicts the fraction of all down-loads below five seconds. This fraction increases from 99.2% for 3.6Mb to 99.6% for 4.5Mb. The QoS requirement, that 99% of the web pages can be down-loaded in five seconds, is fulfilled for all studied capacities. We presume that the reason for this excellent performance for real-time data is the unlimited use of the total capacity at the bottleneck link.

To determine the capacity requirement for a network without diffserv support, we review QoS requirements for VoIP and data traffic. VoIP's delay is the dominant requirement that sets the capacity requirement to 4.2Mb.

We summarize our results for provisioning an integrated network without diffserv support:

- An integrated network without diffserv support cannot differentiate services. Therefore we have to over-provision such that all data traffic meets the QoS requirement for real-time data.
- The stringent requirement on VoIP's delay is the dominant factor setting the capacity requirement for an integrated network without diffserv support to 4.2Mb.

5 DiffServ Support

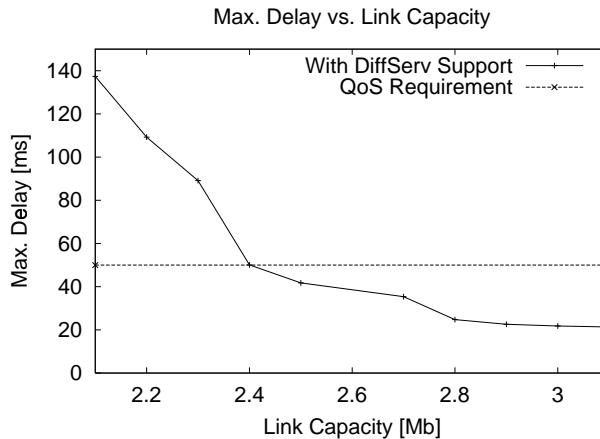


Fig. 10. Maximal end-to-end Delay with DiffServ Support

In this section we investigate provisioning an integrated network for voice, real-time, and best-effort data with diffserv support. To differentiate three levels of service, we mark the traffic as follows:

1. Voice traffic as expedited forwarding (EF).
2. Real-time data traffic as assured forwarding (AF).
3. The remaining traffic as best effort (BE).

After some first trials, in which we have observed larger capacity requirements for a diffserv network than for a network without diffserv support, we realized a performance problem caused by the deficit round-robin (DRR) scheduler in Murphy's diffserv package. The DRR scheduler implements a queue for each level of service and assigns a fixed number of byte-credits to each queue on a per-round basis. These byte-credits can then be used to forward packets of back-logged

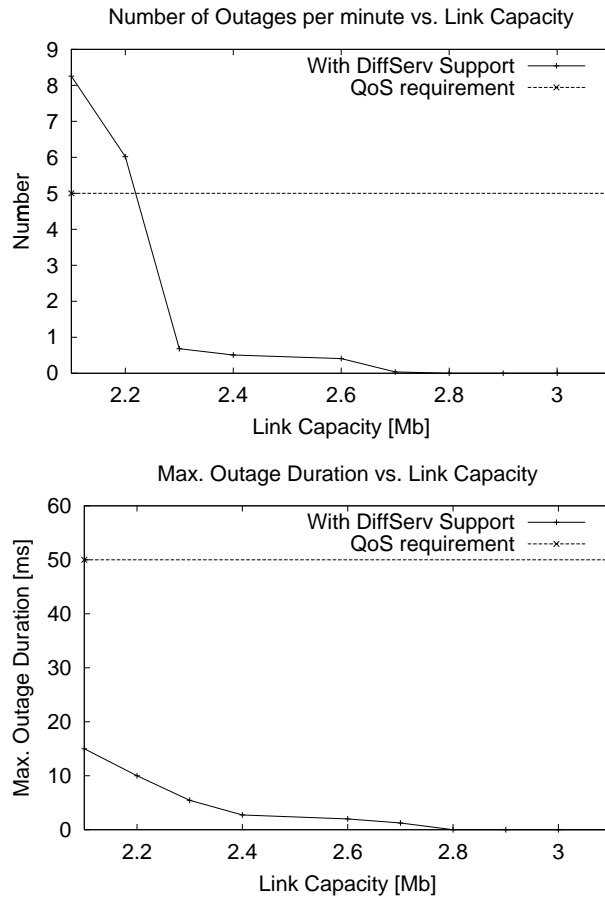


Fig. 11. Number of Outages and Outage Duration with DiffServ Support

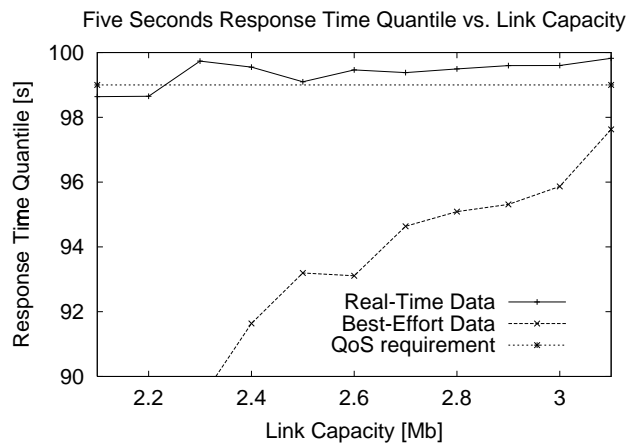


Fig. 12. Five Seconds Response Time Quantiles with DiffServ Support

queues. If a queue is not back-logged, the byte-credits can be unlimitedly accumulated to be used when packets arrive on this queue, which then causes high queuing delays for packets in other queues. We have found in our simulations that bursty real-time traffic has led the AF queue

to accumulate enough byte-credits to cause large delay and outages for VoIP traffic in EF. The performance problem is now also reported on the web page of the diffserv package[5]. We thus revised the scheduler and modified it into a weighted round-robin (WRR) based on packet counts. This WRR scheduler has the advantage that we can give tight delay bounds for every queue. With this modification, we were able to measure significantly lower capacity requirements than in the integration without diffserv case.

Before varying the fraction of data traffic that is real-time, we have thoroughly studied a setup in which 20% of the web clients have requested real-time data and 80% of the web clients have requested best-effort data. We have found that the capacity requirement is minimized when service rates for the queues in the WRR scheduler are configured such that QoS for voice and real-time data is simultaneously at the performance limit. We show figures that depict QoS metrics vs. capacity for such a configuration. For VoIP, Figure 10 shows that end-to-end delay lowers from 140ms at 2.1Mb to 20ms at 3.1Mb. The QoS requirement of 50ms is fulfilled at 2.4Mb. Figure 11 (top) shows that the number of outages for VoIP drops from eight per minute at 2.1Mb to zero at 2.8Mb and remains zero for capacities larger than 2.8Mb. Figure 11 (bottom) shows that the maximal outage duration for VoIP drops from 15ms at 2.1Mb to zero at 2.8Mb and remains zero for capacities larger than 2.8Mb. The QoS requirement for outage durations, not to exceed 50ms, is met for all capacities in the interval studied. We see that all QoS metrics for VoIP in an integrated network with diffserv support evolve similar to the corresponding QoS metrics in a dedicated network. Delay, and the number and duration of outages sharply drop at some sufficient capacity around 2.4Mb. The delay requirement determines the provisioning from the VoIP side.

For real-time data, the fraction of web pages that can be down-loaded in less than five seconds increases from 98.7% at 2.1Mb to 99.7% at 3.1Mb (Figure 12). The 99% QoS requirement, determining provisioning from the real-time data side, is met at 2.3Mb. For best-effort data, the fraction of pages with down-loads times of less than five seconds increases from around 80% at 2.1Mb (not shown) to 97.6% at 3.1Mb. This is a clear service differentiation between real-time traffic and best-effort traffic.

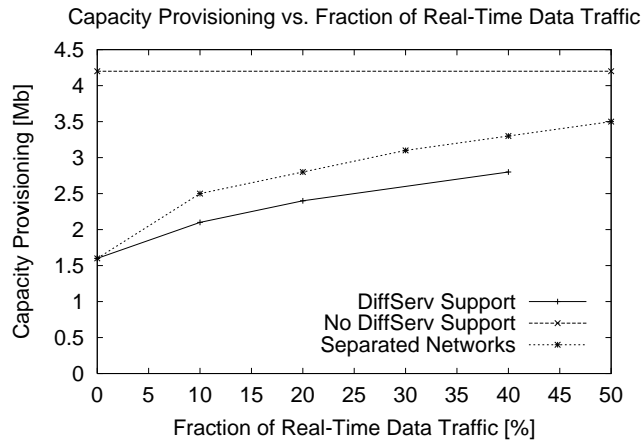


Fig. 13. Provisioning Requirement versus Fraction of Real-Time Data Traffic

Next, we vary the fraction of data traffic that is real-time to investigate the relationship between provisioning requirements and the fraction of real-time data traffic. We find an approximately linear correlation for small to medium fractions of data traffic being real-time; more real-time traffic needs more capacity (Figure 13). Compared to the dedicated networks case, the savings in capacity are approximately 15% for the WRR scheduler with our source model that produces slightly more data traffic than voice traffic. First measurements with a weighted fair queuing scheduler seem to

show slightly higher savings. In addition, we see that a network without diffserv support needs up to 60% over-provisioning.

We additionally experimented with conditioning the real-time data traffic with a token bucket and found that configuration of rate and depth had no significant effect on provisioning.

6 Conclusion

In this study, we investigate QoS provisioning of intra-nets serving three classes of traffic, i.e., voice, real-time data and best-effort data. We find that integration on a network without diffserv support needs significant over-provisioning if a small to medium fraction of data traffic is real-time. If all data traffic is real-time, provisioning for dedicated networks and integrated networks with or without diffserv support is not much different. For all other traffic compositions we find lower capacity requirements for networks with diffserv support than for dedicated networks. Compared to dedicated networks, the percentage of savings for integration with diffserv remains constant for a small to medium fraction of real-time data. Our results provide a rule of thumb towards capacity planning for real-time data.

For the networks with diffserv support, we find that choice and implementation of the scheduling algorithm had significant impact on capacity requirements. We are aware that the WRR scheduler, which we used in Section 5 does not lead to the optimal performance. However, we can give tight delay bounds for this scheduler, which we think is crucial as delay seems to be the limiting factor. To clarify this issue, we are currently experimenting with a weighted fair queuing scheduler[22].

In addition, we plan to further vary traffic compositions, particularly to study variations of offered load for voice and data, and investigate more complex topologies. To generalize our results for the Internet, we plan to investigate capacity provisioning on randomly generated topologies with Zipf's law connectivity [23]. Finally, we would like to stress that, in addition to the results reported in this paper, our simulation package is publicly available. It will enable the research community, network service providers, banks and other large enterprises to derive provisioning guidelines for their networks.

7 Acknowledgements

We would like to thank many people for helpful discussions particularly Marcel Dasen, Thomas Erlebach, George Fankhauser, Felix Gaertner, Matthias Gries, Albert Kuendig, Sean Murphy, and Burkhard Stiller.

References

1. A. M. Odlyzko, "The internet and other networks: Utilization rates and their implications," *Information Economics and Policy*, vol. 12, pp. 341–365, 2000.
2. S. Bajaj et. al., "Is service priority useful in networks?," in *Proceedings of the ACM Sigmetrics '98*, Madison, Wisconsin USA, June 1998.
3. G. Kim et. al., "Qos provisioning for voip in bandwidth broker architecture: A simulation approach," in *Proceedings of the communication networks and distributed systems modeling and simulation conference (CNDS)'01*, Phoenix, Arizona, USA, Jan. 2001.
4. L. Breslau et. al., "Advances in network simulations," *IEEE Computer*, May 2000.
5. S. Murphy, "Diffserv package for ns-2," available at <http://www.teltec.dcu.ie/~murphys/ns-work/diffserv/index.html>.
6. R. Fielding et. al., "Hypertext transfer protocol — http/1.1," RFC 2616, Internet Request For Comments, June 1999.
7. S. Blake et. al., "An architecture for differentiated services," RFC 2475, Internet Request For Comments, Dec. 1998.
8. J.Wroclaski D. Clark, "An approach to service allocation in the internet," IETF Draft, July 1997.

9. L. Zhang K. Nichols, V. Jacobson, "A two-bit differentiated services architecture for the internet," IETF Draft, Apr. 1999.
10. V. Jacobson et. al., "An Expedited Forwarding PHB," RFC 2598, Internet Request For Comments, June 1999.
11. A. Charny J. Wroclawski, "Integrated service mappings for differentiated services networks," IETF Draft, Feb. 2001.
12. S. Shenker et al., "Specification of guaranteed quality of service," RFC 2212, Internet Request For Comments, Sept. 1997.
13. J. Heinanen et. al., "Assured Forwarding PHB Group," RFC 2597, Internet Request For Comments, June 1999.
14. J. Wroclawski, "Specification of the controlled-load network element service," RFC 2211, Internet Request For Comments, Sept. 1997.
15. Siemens, *Telephone traffic theory tables and charts*, Siemens Aktiengesellschaft, Berlin - Muenchen, Germany, 1981, 3rd edition.
16. M. Dasen, "Up-to-date information in web accessible information resources," Ph.d. thesis – tik-schriftenreihe nr. 42, ETH Zurich, June 2001.
17. P. Barford et. al., "Changes in web client access patterns: Characteristics and caching implications," *World Wide Web, Special Issue on Characterization and Performance Evaluation*, vol. 2, no. 2, pp. 15–28, 1999.
18. P. Barford and M. E. Crovella, "Generating representative web workloads for network and server performance evaluation," in *Proc. of Performance'98/ACM SIGMETRICS'98*, 1997.
19. R. Steinmetz and K. Nahrstedt, *Multimedia: Computing, Communications & Applications*, Prentice-Hall, 1995.
20. International Telecommunication Union Telecom Standardization Sector, "Itu-t g.711 - pulse code modulation *pcm* of voice frequencies," available at <http://www.itu.int>, Nov. 1988.
21. V. Paxson, "End-to-end internet packet dynamics," *IEEE/ACM Transactions on Networking*, vol. 7, no. 3, pp. 277–292, June 1999.
22. C. Partridge, "Weighted fair queueing," in *Gigabit Networking*. 1994, p. 276, Addison-Wesley Publishing.
23. Faloutsos et. al., "On power-law relationships of the internet topology," in *Proceedings of ACM SIGCOMM*, Stockholm, Sweden, Aug. 1999.