A Low Cost Approximation for a Controlled Load Service in a bridged/switched LAN Environment

Peter Kim

pk@hplb.hpl.hp.com
Hewlett-Packard Laboratories
Filton Rd., Bristol BS34 8QZ, U.K.

Abstract

This paper reports the design and the implementation results of a Controlled Load service in a bridged/switched Local Area Network (LAN). In contrast to other approaches, the service was built based on traffic aggregation and simple two level Static Priority scheduling within LAN switches. All complex control functions such as the per-flow packet classification, policing and traffic reshaping are only performed within end-nodes (hosts or routers) of the LAN. Admission Control is used to restrict the access to the new service. Measurement results received in a bridged test network exibit low average end-to-end delay characteristics and no packet loss despite: (1) the simplicity of the traffic control implemented in our LAN switches, (2) a worst-case network load, and (3) the use of long range dependent data sources within the experiments.

1 Introduction

Packet switching data networks such as the Internet are currently migrating towards multi-service networks [1], [2]. This is driven by the use of applications with a variety of performance constraints and the widening commercial use of the Internet. In contrast to the traditional Best-Effort service deployed today, the new services will offer quality of service and include diverse service commitments such as probabilistic or deterministic performance guarantees.

One approach for a multi-service Internet that was proposed by the Internet Engineering Task Force (IETF) is the Integrated Service Packet Network. A core component of it is the service model which defines the end-to-end behaviour of the network. So far, the Guaranteed- [3] and the Controlled Load [4] service have been put forward as proposed Internet standards.

End-to-end service guarantees can however only be provided when the mechanisms which enable these guarantees are supported at all intermediate links along the data path. This includes LAN technology which is typically located at both ends of the data path, where large bridged networks often interconnect hundreds of users. The IETF Integrated Services over Specific Link Layers (ISSLL) working group was chartered with the purpose of exploring the mechanisms required to support the new services over various link layer technologies. Reference [5] describes the framework for providing this functionality in shared and switched IEEE 802 type LANs. Our work was carried out in this context.

This paper presents a low cost approach to provide Controlled Load quality of service within bridged/switched LANs. We focused on Controlled Load because the low assurance level of this service and the design freedom given by the ISSLL framework enable a variety of tradeoffs in the design of the traffic control mechanisms within the network. Our design aimed at a cost effective solution in which LAN switches only use a simple Static Priority scheduler with two priority levels and classify data packets only based on the User Priority proposed in IEEE P802.1p/Q [6], [7]. Both resulted from the strong cost constraints in the LAN market in which bandwidth overprovisioning is a competitive (and the typical) alternative to resource allocation. Our approach was further motivated by the fact that many next generation LAN switches will only support priority scheduling and P802.1p/Q classification. More advanced (e.g. per-flow-) control mechanisms will probably not be widely available in the near future and thus delay the deployment of algorithms relying on these mechanisms indefinitely.

The remainder of this paper is organized as follows. In Section 2, we discuss the service capabilities of the Best-Effort service. This is based on experimental results received in our testbed. Section 3 outlines the specification of the Controlled Load service. Section 4 discusses design decisions and describes the packet scheduling process that was used to enforce the required service quality. Section 5 provides the admission control conditions. This is followed by an evaluation of the new service in Section 6. Related work is studied in Section 7. A summary and our conclusions can be found in Section 8.

2 Performance Characteristics under Network Overload

The part of the end-to-end delay that can be controlled by admission control is the queuing delay. For a single LAN switch, this delay depends on: (1) the burstiness of the arriving data traffic, (2) the arrival- and the service data rate, and (3) the buffer capacity of the switch. To investigate the performance of a switched LAN that uses the traditional First-In-First-Out (FIFO) service discipline within switches, we measured: (1) the end-to-end delay and the packet loss rate versus the network load, and (2) the packet loss rate in dependence of the buffer space. Before discussing the results, we however first outline our measurement methodology, which was also applied to achieve the measurement results reported in Section 6.

2.1 Measurement Methodology

To describe this, we consider the test network illustrated in Figure 1. It consists of two LAN switches and ten HP-UX 9.05 UNIX workstations. All delay measurements were taken by a workstation which we called the Measurement Client. It had two LAN adapter cards. One of them (1a) was exclusively used for sending data, the second one (1b) was used for receiving. All data packets generated by the Measurement Client were addressed to a pre-defined multicast group which was joined with the receiver interface 1b. By using the same workstation for sending and receiving test packets, we could use the same clock for determining the start and finish time of each measurement and thus avoided timing discrepancies that would have occurred if we had used two separate workstations. The time was measured using PA-RISC register CR 16 [8] which provides a 10-ns tick on a 100 MHz HP C100 workstation.

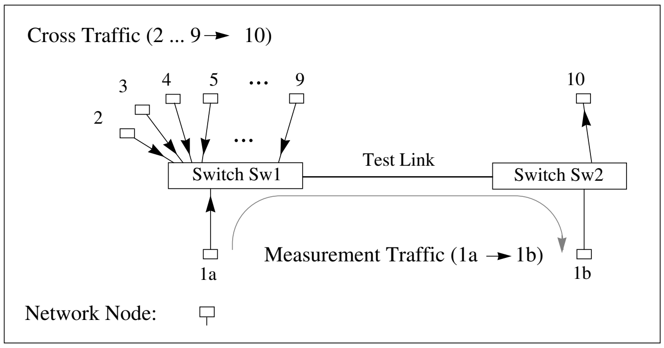

Figure 1. Experimental Setup for Measuring the Packet Delay and Loss Characteristics.

All time-stamps are taken within the device driver of the LAN adapter card in the UNIX kernel. The measured delay is the link layer end-to-end delay. It includes: (1) the time for transferring the data packet from kernel memory to the adapter card, (2) the queuing delay within the test network, (3) the time for transferring the packet back to the kernel memory, and (3) the relevant operating system overhead. The accuracy of the overall approach is in the order of 40 μs . A detailed analysis of it can be found in [9].

The measurement setup further included eight other workstations which imposed cross traffic. We called these Traffic Clients. The Test Link in Figure 1 was a 100 Mbit/s half-duplex IEEE 802.12 link. All data packets were forwarded in one direction, from switch Sw1 to Sw2. This avoided any medium access contention and led to network conditions that were basically identical to those in a full-duplex switched network. Filter entries in the switches ensured that the measurement- and the cross traffic only interacted at the entrance to the test link.

To measure the packet loss rate in the network, we used two different approaches. The first was based on the MIB counter IfOutDiscards [10]. It counts the data packets dropped at a particular switch port and thus enabled us to determine the total packet loss rate at the entrance to the test link. The Measurement Client was used to determine the loss rate for a single test flow. This was based on sequence numbers carried in the test packets.

To generate realistic traffic pattern within the network, we used a traffic trace driven approach. For this, we first recorded 2 hour test traces for several multimedia applications using a LAN Traffic Monitor. In the experiments, the traces were then passed to a UNIX kernel based Traffic Generator which generated an almost identical data stream to the original trace monitored on the network. To generate *N* homogeneous data sources from the same Traffic Client, we multiplexed *N* copies of the original trace into a single trace file, where each of the *N* copies had a different, randomly chosen start offset into the original trace. On reaching the end of the trace a source wrapped around to the beginning. Random start offsets were further applied for all Traffic Clients to avoid synchronisation effects.

Trace Name	Encoding Scheme	Total Number of Packets	Average Data Rate (Mbit/s)	Hurst Parameter
OVision	MPEG-1 Video	844438	1.286	0.60
MMC1	JPEG Video	2078674	2.973	0.47
MMC2	JPEG Video	1722701	2.611	0.84

Table 1. Basic Characteristics of the Application Traces.

Table 1 shows the basic characteristics of the application traces. We recorded one MPEG-1 encoded video trace from the OptiVision Communication System [11] and two JPEG encoded traces from the conferencing system MMC [12]. A description of the exact configurations used during the recording is given in [9]. Reference [9] further contains a statistical analysis of the traces and a comparison to a 2 hour JPEG encoded Star-Wars trace extensively studied in [13]. A Variance Time Plot analysis led to the results for the Hurst parameter *H*. *H* is often used to quantify the degree of long range dependence. The result for MMC2 suggest that this trace has strong dependencies. The estimates received for the two other traces, OVision and MMC1 indicate burstiness occuring over much shorter time intervals.

Source Name	Ave. Data Rate (Mbit/s)	Model Parameters					
		Peak Rate (Pkts/s)	I (ms)	N (Pkts)	Peak/ Ave. Ratio	β ON-Period	β OFF-Period
POO1	0.321	64	325	20	2	1.2	1.1
POO3	0.262	256	360	10	10	1.9	1.1

Table 2. Source Model Parameters for the Pareto Sources.

For comparison, we also used two “artificial” traffic sources with infinite variance distributions for generating test traces. Following [14], this was based on an ON/OFF source model with Pareto distributed ON and OFF times. The authors of [14] showed that the superposition of many sources whose ON and OFF periods exhibit infinitive variance produces network traffic with long range dependencies. Table 2 shows the model parameters we used in our experiments. These follow the suggestions made in [14] for LAN traffic. They are further identical to those in [15]. We however used fixed sized packets of 1280 byte length. A comparison to our application traces in Table 1 showed that traces derived from these source models maintain their peak to average ratio over significantly longer time scales.

2.2 Delay and Loss Characteristics

Using the setup illustrated in Figure 1, we first performed four different measurements. These were based on traces generated from the application traces MMC2 and OVision, and the traffic sources models POO1 and POO3. In each measurement we only used homogeneous flows produced from the same application trace or the same source model. The Measurement Client always injected a single flow into the test network and measured the delay and loss rate for the corresponding data packets. The cross traffic varied from zero to a total load of about 90 Mbit/s. The required trace files were pre-computed. All data packets used the Best-Effort service. The measurement interval for a single measurement point was 30 minutes.

Figure 2, Figure 3 and Figure 4 show the results. For all test sources, we can more or less observe a certain threshold in the load-delay and load-loss curves: the results are low as long as the network load stays below the threshold. As soon as the network utilization however exceeds the threshold, delay and loss increase significantly faster. Note that even though IEEE 802.12 is the standard for a shared 100 Mbit/s LAN, the actual maximum data throughput across a single link is variable and only about 93 Mbit/s when 1280 byte data packets are used for the data transmission. This can be explained with the overhead imposed by the Demand Priority medium access protocol and is extensively analysed in [16], [17] in the context of the Guaranteed service.

The maximum delay is determined by the burstiness of the traffic and the network load. All results are limited by a bound of about 23 ms which corresponds to 256 kbytes of output buffer space used in switch Sw1. As soon as the maximum delay reaches this bound, the output FIFO queue is full and packet loss occurs as can be observed in Figure 2 and Figure 4.

For all four test sources, the results for the average load stay almost constant over a load range of over 60 Mbit/s. Even for low loss rates smaller than 0.1%, the average delay remains in the order of a few milliseconds, typically below 10 ms. From this, two simple conclusions can be drawn: (1) if the network administrator can ensure that the network is always operating below the load-threshold, then resource reservation is probably not required unless an application has guaranteed service constraints. This is well known. (2) Since the average delay does not significantly increase with the network load, there seem to be little gain in supporting *several* higher priority levels to differentiate service classes with a different *average* delay within high speed LAN switches. Even when several classes were implemented, these would provide an average delay which would be hard to distinguish for existing real-time applications which typically have an end-to-end delay budget of 150 ms [18], [19].

It remains to remark that we received similar load-delay and load-loss curves in experiments in which the Test Link in Figure 1 consisted of: (1) a full-duplex Ethernet (100BaseT) link, (2) a shared half-duplex 802.12 link, and (3) a shared multihub 802.12 network. In the latter two cases, cross traffic was injected from both switches Sw1 and Sw2. This led to longer tails in the delay distribution. The results for the average delay however did not differ significantly in all experiments.

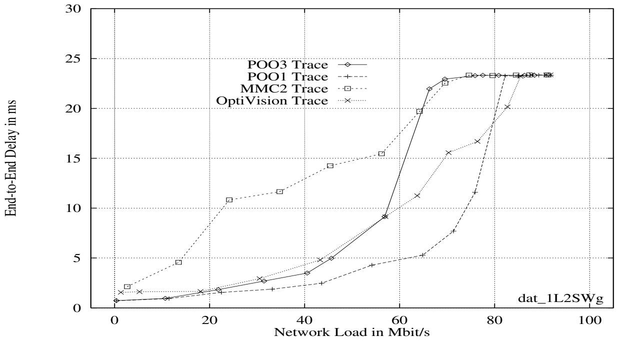

Figure 2. Maximum End-to-End Packet Delay for different Flow Types in Dependence of the Network Load.

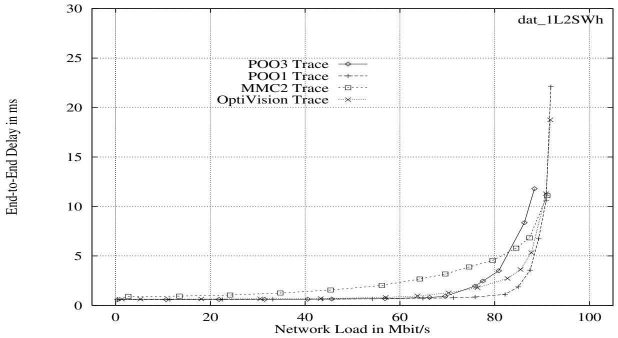

Figure 3. Average End-to-End Packet Delay for different Flow Types in Dependence of the Network Load.

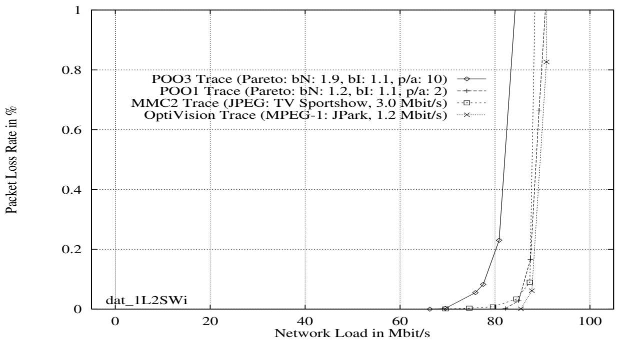

Figure 4. Packet Loss Rate for different Flow Types in Dependence of the Network Load.

2.3 Impact of the Buffer Space in LAN Switches

Given the strong impact of the buffer capacity in LAN switches on the loss characteristics and the desire to reduce costs, we measured the packet loss rate in dependence of the buffer space. This used the same setup as applied to investigate the delay characteristics. We performed six experiments based on MMC2 and POO3 test sources because these exhibited the worst behaviour in the experiments reported in the previous section. In the first test, we loaded the network with 24 MMC2 video flows: 23 were generated by the Traffic Clients and 1 was generated by the

Measurement Client. Switch Sw1 in Figure 1 had an output buffer of just 16 kbytes for each of its ports. We then measured the packet loss rate of the aggregated traffic at the output port from switch Sw1 to Sw2. Afterwards, we increased the buffer size in Sw1 and repeated the test using the same setup but a larger buffer space. After measuring the loss characteristics for the entire buffer space range, we performed the same set of experiments with 28 and with 32 MMC2 flows on the network.

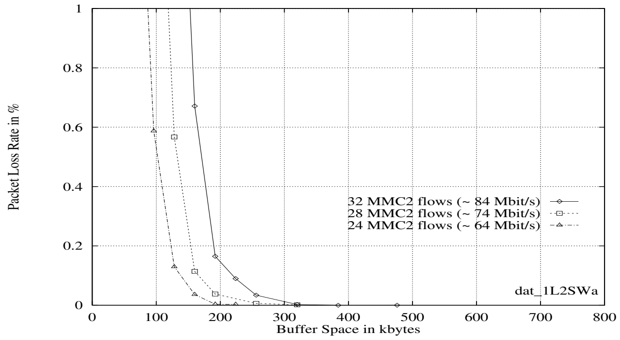

Figure 5. The Packet Loss Rate for different Sets of MMC2 Flows in Dependence of the Buffer Space in the Switch.

Figure 5 shows the results for the total packet loss rate in switch Sw1. For all three test sets, a minimum buffer space of far less than 200 kbytes prevents packet loss rates larger than 1%. As expected, the slope of the loss-curves becomes flatter such that significantly more memory is required to completely eliminate the packet loss in Sw1.

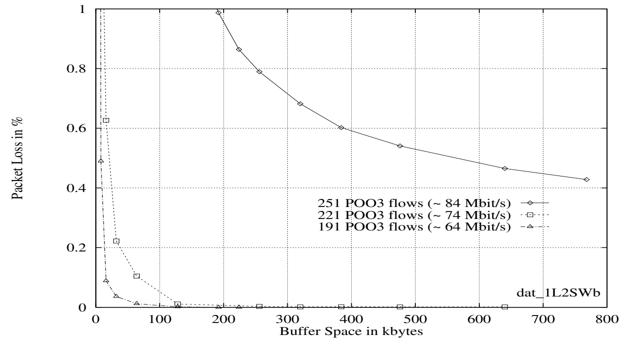

Figure 6. The Packet Loss Rate for different Sets of POO3 Flows in Dependence of the Buffer Space in the Switch.

Figure 6 contains the results for the POO3 test sources measured using the same setup. Each test on average generated the same network load as the corresponding MMC2 test (≈ 64, ≈ 74, ≈ 84 Mbit/s). The results however differ significantly from the ones in Figure 5. Much more buffer space is required to eliminate packet loss in the switch. This is not surprising considering the infinite variance of the Pareto distribution. For a load of about 84 Mbit/s (the upper curve in Figure 6) and a buffer space of 768 kbytes, the packet loss rate is still 0.428%. The slope of the curve only decays slowly. The loss rate however significantly decreases when the load falls below a certain utilization which occurs in Figure 6 between 84 and 74 Mbit/s. This is caused by the limited peak rate of the POO3 Pareto source. Unlike the results in Figure 5, we can observe an extremely long tail in all loss-curves in Figure 6. However, to achieve rates of under 1%, only buffer space of far less than 100 kbytes is required.

The optimum amount of buffer space to be used in LAN switches is hard to determine. In this section, we could observe that increasing the buffer space decreased or even eliminated the loss in the test switch, provided traffic bursts were temporary and moderate. This required large buffer sizes because the loss rate and the buffer space are not linearly correlated. The results have also shown that more buffer space does not always help. In case the traffic characteristics exceeded a certain load and burstiness threshold then even a large amount of buffer space could only insignificantly reduce the loss rate. To provide quality of service under these conditions requires traffic control such as for example resource reservation with admission control.

3 Characteristics of the Controlled Load Service

The Controlled Load service [4] attempts to approximate the service that an application would receive from the Best-Effort service under unloaded network conditions. It was motivated by the fact that existing applications typically work well across lightly loaded networks, but become unusable when the network is overloaded. In contrast to the Guaranteed service, Controlled Load does not give absolute guarantees for any service parameters such as the delay or the packet loss rate. The service specification is intentionally minimal which will allow a wide range of implementation approaches.

Admitted flows may assume: (1) a very low packet loss rate *close to the packet error rate* of the transmission medium, and (2) a low average delay. More precisely, the average *queuing delay* should not be significantly larger than the flows *burst time*. If a flow's traffic is characterized using a Token Bucket Filter[1] then the burst time is given by: δ / r, where δ denotes the token bucket depth (or the burst size), and *r* the token generation rate. The difference to the Best-Effort service is that the above conditions are guaranteed even when the network is congested.

4 Packet Scheduling Process

During our experiments in Section 2.2, we found that the network maintains an almost constant average delay over a long load range[2]. Existing delay sensitive applications will thus see little difference between an empty (~0 Mbit/s) and a moderately loaded (~60 Mbit/s) network segment in the data path. This behaviour can be exploited to provide Controlled Load service.

---

1 The Token Bucket Filter is sometimes also called Leaky Bucket or (δ, r) Regulator. For an analysis of the scheme, see [20].

2 For example, the average delay for the *OptiVision* application in Figure 3 (playing the MPEG encoded adventure movie *Jurassic Park)* only increases by 0.631 ms while the network load increases from ~1.3 Mbit/s (1 *OVision* flow) to ~70 Mbit/s (54 *OVision* flows).

In contrast, packet loss must be watched carefully. We observed that it may occur long before the application may be able to detect a change in the average delay. Furthermore, the Controlled Load service definition specifies a target loss rate close to the packet error rate of the transmission medium. This is extremely low in LAN's. The IEEE 802.12 standard [21] for example specifies a minimum bit error rate of less than 1 bit error in 10^8 bits for a UTP cabling.

Due to these constraints, we focus on controlling the packet loss rather than the average delay and attempt to provide a loss free packet delivery service as, we believe, is expected from a Controlled Load service in a LAN environment. Based on the observations in Section 2.2, we do not attempt to derive a bound for the average delay for each admitted flow (and compare the result to the flow's burst time), since we can expect the average delay to be sufficiently low, provided there is no packet loss in the network. Considering the small load-delay variations observed in Figure 3, it also seems questionable whether a calculus will be able to provide accurate upper bounds for the average delay that are useful in the admission control.

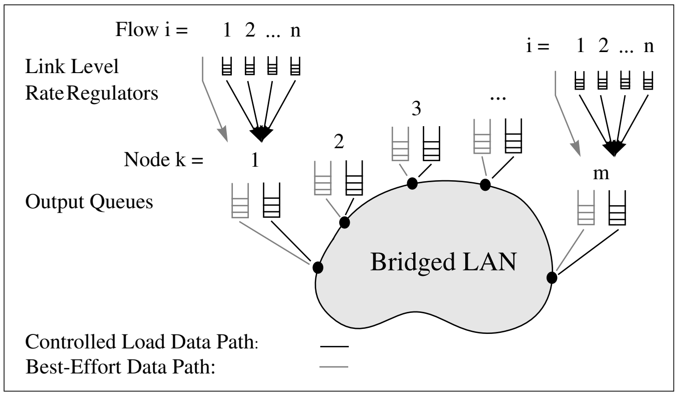

Figure 7. Traffic Reshaping Points for the Controlled Load Service.

Our admission control conditions were derived based on the assumptions that the switched LAN: (1) supports a priority mechanism with at least two priority levels, and (2) has a deterministic link access. The use of full-duplex switched links in the network typically simplify the analysis since they eliminate the need to consider the contention between different network nodes sharing the same physical medium. In our model, LAN switches may however also be interconnected by shared medium links, provided the link technology supports a prioritized medium access. In our test network for example, we used shared half-duplex and multihub IEEE 802.12 technology which supports two priority levels (high- and normal) for the medium access.

Best effort data packets are forwarded using normal priority. The access to this service is not restricted. The Controlled Load service uses the high priority medium access mechanism. This is subject to admission control. As illustrated in Figure 7, all high priority traffic is policed and reshaped on a per-flow basis at the entrance (hosts or routers) of the LAN. The rate regulators required in these nodes are implementations of the Token Bucket filter ((δ, r) Regulator).

In LAN switches, all Controlled Load traffic is aggregated into the high priority queue of the Static Priority scheduler. When forwarded across a shared link, the 802.12 high priority medium access isolates the aggregated Controlled Load flows from the Best-Effort traffic. Since LAN switches do not differentiate individual data flows, these may become more and more bursty as they travel across several links. This is due to the interaction with other Controlled Load traffic.

The admission control ensures that on each link, the Controlled Load service, on average, never consumes more resources than a pre-defined allocation limit. Fortunately, data packets in LANs are delivered based on a single data distribution tree, enforced by the standard 802.1 Spanning Tree protocol [22]. Feedback effects[1] as observed in wide area networks can not occur. It can further be assumed that the number of bridges between the source and the destination in the LAN is limited. On average, we expect this to be of the order of two to five bridges.

It remains to remark that the packet processing described in this section differs substantially from the Integrated Services architecture standardized for the network layer - which requires core mechanisms such as packet classification, policing or reshaping to be implemented on a per-flow basis within routers. Our model however still complies to the IETF ISSLL framework for providing Integrated Services across 802 type LANs.

5 Admission Control

In contrast to our packet scheduling model which could be applied in a variety of LANs e.g. full-duplex switched 100BaseT networks, there is no standard mechanism for performing admission control for existing technologies such as 802.3 Ethernet, 802.5 Token Ring, or 802.12 Demand Priority. This is because each of these technologies has a different medium access mechanism and therefore schedules data packets according to its own policy. Another factor to be considered is the network topology which can be shared, half-duplex- or full-duplex switched. The admission control conditions used to enforce Controlled Load service will thus typically be technology and topology specific and must be defined separately for each LAN technology.

This section provides the admission control conditions for 802.12 LANs. We use a parameter based approach. All results are derived using the Token Bucket (δ, r) traffic characterisation. In [20], it is shown that the Token Bucket enforces the amount of data which can leave the regulator in any time interval Δt. A data flow i conforms to its (δ^i, r^i) characterization if in any existing time interval Δt no more than $b^i(\Delta t)$ bits leave the token bucket, where $b^i(\Delta t) \leq \delta^i + r^i \Delta t$ is the Traffic Constraint Function of flow i.

Link layer network resources are reserved on a link-by-link basis, where a link may be a shared multihub 802.12 segment or a half-duplex- or full-duplex link. The admission control consists of a Bandwidth- and a Buffer Space Test. The Bandwidth Test proves that sufficient spare bandwidth is available such that Stability is maintained when the new flow is admitted. More pre-

[1] See for example [23] for a discussion of feedback effects.

cisely: assuming a link with N flows admitted, where each flow i obeys its Traffic Constraint Function $b^i(\Delta t)$ at the entrance of the bridged LAN, then Stability is given when:

$$\lim_{\Delta t \to \infty} \sum_{i \in N} b^i(\Delta t) - C_s \cdot \Delta t = -\infty \tag{5.1}$$

holds, where C_s denotes the link capacity available for serving data. C_s is computed in [9] (see Equation 7.21 in Chapter 7). Intuitively, the network is stable when: (1) the burst size δ^i is bounded for each flow i, and (2) the sum of the average data rates of all admitted flows is smaller than the available link capacity C_s.

As we could observe in [17], the data throughput in 802.12 networks depends significantly on the hub cascading level and the packet sizes used for the data transmission. To consider these dependencies, the Bandwidth Test is based on a Time Frame concept which enabled us to bind the Demand Priority medium access overhead. For each Controlled Load flow, network bandwidth is allocated according to the average data rate r specified in the traffic characterisation (δ, r).

The Buffer Space Test checks that there is sufficient buffer space available such that none of the flow's data packets is dropped due to a queue overflow in the network. For this we first derived an approximation of the Traffic Constraint Function of a single flow after it traversed a single link. The result enabled us to determine the buffer space. Before this is discussed, we however first describe the Bandwidth Test. Furthermore, in the remaining of this paper, we use the term *real-time* flow to denote a data flow using the Controlled Load service.

5.2 The Bandwidth Test

Theorem 1: *Consider an 802.12 link with m nodes, where each node k has n real-time flows which are already admitted. Assume a time frame of TF, a link speed of* C_l *and that the packet count for flow i on node k over the time interval TF is* $pcnt_k^i$. *Further let* D_{pp} *and* D_{it} *be the topology specific worst-case per-packet overhead and normal priority service interrupt time, respectively. Furthermore, assume that each real-time flow i on each node k has a bounded burst size* δ_k^i *and obeys its traffic characterisation* (δ_k^i, r_k^i) *at the entrance of the bridged network. A new Controlled Load flow* ν *with the traffic characterisation* (δ^ν, r^ν) *and a packet count* $pcnt^\nu$ *can be admitted such that Stability is maintained if:*

$$r^\nu < \frac{TF - D_{it} - \frac{1}{C_l}\sum_{k=1}^{m}\sum_{i=1}^{n} r_k^i \cdot TF - \sum_{k=1}^{m}\sum_{i=1}^{n} pcnt_k^i \cdot D_{pp} - pcnt^\nu \cdot D_{pp}}{TF/C_l} \tag{5.2}$$

The formal proof can be found in [9] (see Chapter 7.2). The theorem is derived from Theorem 1 in [16] which accurately computes the available link bandwidth in 802.12 networks, but allocates resources based on the peak-data rate. The most significant difference is that the time frame *TF* in [16] is always also an upper bound on the delay for all Guaranteed service flows. This does not hold for Equation 5.2 due to the average data rate allocation. The parameters D_{pp} and D_{it} reflect the overhead of the Demand Priority medium access. Results for them can be found in [17]. Parameter C_l denotes the physical link speed. In contrast to the available data throughput C_s used in Equation 5.1, C_l is constant and equal to 100 Mbit/s. The packet count $pcnt_k^i$ represents the number of packet overheads which real-time flow i on network node k can consume within the time frame *TF*. It is equivalent to the maximum number of Controlled Load data packets that flow i is allowed to send within *TF*.

Theorem 1 applies to the most general case: a link including a shared multihub segment with k nodes, each of which forwarding n Controlled Load flows into the network. Links with other topologies such as single hub segments, half-duplex (k = 2) or full-duplex (k = 1) links can be viewed as special cases. The Bandwidth Tests for these topologies only differ in the use of the topology specific results for the per-packet overhead D_{pp} and the normal priority service interrupt time D_{it}.

5.3 The Output Traffic Constraint Function

To derive an approximation of the Traffic Constraint Function of real-time flow i after it traversed a (potentially shared) link, we used the analysis technique proposed by Cruz in [20]. For this, we first defined $b_{in}^i(t)$ and $b_{out}^i(t)$ as the Input- and Output Traffic Constraint Function of flow i, respectively. Function $b_{in}^i(t)$ describes the traffic pattern that enters the high priority output queue at the LAN switch selected for analysis, whereas $b_{out}^i(t)$ describes the flow's traffic at the exit of the corresponding link. It represents the traffic that is then passed into the output queue of the next LAN switch in the data path. This assumes output buffered LAN switches which we also used in our test network. The result of the analysis can then recursively be applied to determine $b_{out}^i(t)$ on each link along the data path of flow i.

We further defined a minimum service rate $R^i_{min\_Nk}$ for each real-time flow i on the link. It is the minimum data throughput which a flow is guaranteed to receive from the network. $R^i_{min\_Nk}$ is enforced by the round-robin service policy embedded into the Demand Priority medium access protocol. For example, on a half-duplex 802.12 link connecting two LAN switches each of which forwarding a single real-time flow into the link, we have: $R^i_{min\_Nk} \approx C_s/2$, provided the packet size distribution of both flows is similar.

Theorem 2: *Consider an 802.12 link with m network nodes and assume that each real-time flow i obeys its Input Traffic Constraint Function* $b_{in}^i(t)$. *Then let* D_{it}, P_{max} *and* $R^\nu_{min\_Nk}$ *be the Normal Priority Service Interrupt Time, the Maximum Link Packet Size and the Minimum Service Rate of flow* ν *selected for the analysis, respectively. Furthermore let* Δ *and H denote two* *time variables, where: $\Delta \geq 0$ and $H \geq 0$. If Theorem 1 applies (Stability) and the network has a total capacity of at least C_s available for serving data, then the output traffic of flow ν at the exit of the segment is bounded by:*

$$b^{\nu}_{out}(t) \leq \max_{\Delta \geq 0}\left[b^{\nu}_{in}(t + \Delta + D_{it}) - R^{\nu}{}_{min\_Nk} \cdot \left(\Delta - (m-1) \cdot \frac{P_{max}}{C_s} - H\right)\right] \quad (5.3)$$

The proof of Theorem 2 and the derivation of the parameters Δ and H is quite lengthy and was thus omitted in this paper. Both can however be found in [9]. Theorem 2 states that the traffic pattern of flow ν injected by node k is most distorted when the maximum amount of data defined by $b^{\nu}_{in}(t)$ is instantaneously passed into the high priority output queue, but is afterwards only served with the minimum service rate $R^{\nu}{}_{min\_Nk}$. This implicitly assumes the link to be temporarily busy with serving data from: (1) real-time flows $i \neq \nu$ which enter the link at the same node k, and (2) real-time flows injected by other nodes j with $j \neq k$ on the same shared link.

Note here that a Controlled Load service flow may temporarily be served with a rate that is significantly smaller than its allocated bandwidth r. To illustrate this, assume for example a half-duplex switched link with 2 network nodes ($k = 2$), each of which passes a single flow into the shared link. Let: $r^1 + r^2 < C_s$ but $r^1 + r^2 \approx C_s$, $r^1 = 3 \cdot r^2$ and $\delta^1 \gg 0$, $\delta^2 \gg 0$, where C_s denotes the available data rate and (δ^1, r^1), (δ^2, r^2) are the traffic characterisations of the two flows, respectively. Furthermore, assume that resources have been reserved and that both flows use the same fixed packet size for the data transmission. In this case, we find that although less bandwidth is reserved for the flow on node $k = 2$, this flow may nevertheless temporarily consume half of the network capacity due to the average data rate allocation and the round-robin service policy.

It can easily be shown that the longest interval for this effect is given by: $\Delta t \leq \delta^2/((C_s/2) - r^2)$, provided the flow on $k = 2$ obeys its traffic characterisation and does not send more data into the link. During the time interval Δt, node $k = 1$ is however only served with a rate of: $C_s/2 < r^1$, which causes the data in the output queue on this node to grow. The data backlog on node $k = 1$ is only reduced after a maximum of δ^2 data was cleared on node $k = 2$. Similar observations can be made for multihub 802.12 networks.

5.4 The Buffer Space Test

An upper bound on the buffer space required for flow ν follows directly from flow ν's Output Traffic Constraint Function $b^{\nu}_{out}(t)$ for the case that: $t = 0$. More formally, we have:

Theorem 3: *Consider an 802.12 link with m nodes and assume that each real-time flow i obeys its Input Traffic Constraint Function: $b^{i}_{in}(t)$. If Theorem 1 applies (Stability) and the data traffic of the new flow ν at the exit of the link is bounded by the corresponding Output Traffic Constraint Function $b^{\nu}_{out}(t)$ specified by Theorem 2, then the buffer space sS^{ν} required for flow ν at the entrance to the network segment is bounded by:*

$$sS^{\nu} \leq b^{\nu}_{out}(0) \quad (5.4)$$

This follows from the definitions made for the computation of the Output Traffic Constraint Function $b^{\nu}_{out}(t)$ and is formally proven in [9].

6 Controlled Load Service Evaluation

To evaluate the new service, we measured the end-to-end delay and the packet loss rate in several test networks with different topologies. All experiments were based on the methodology described in Section 2.1. We used the application traces: *MMC1*, *MMC2*, *OVision*, and the source model traces: *POO1* and *POO3*, whose characteristics were also discussed in Section 2.1. The corresponding Token Bucket (δ, r) parameters used for these traces are shown in Table 3 and Table 4.

For each trace, we carried out six different measurements which differ in respect to the allocated resources and the location where data flows entered the test network. The latter aimed at an estimation of the impact of the flow distribution and of the flow location on the delay characteristics. For all test traces, we selected the Token Bucket parameters such that a large number of flows could be admitted. This was based on initial experiments which showed that a low bandwidth utilization led to low packet delays despite of the larger total burst size available for admitted flows. Worst-case delays in the network were typically achieved with burst sizes in the order of a few kbytes, because this also allowed a large number of flows to be admitted.

Column 7 in both tables lists the maximum length of the rate regulator queue at the source node. Whenever this limit was exceeded, arriving data packets were dropped. In all experiments using the traces 1 - 4, this however never occurred. Packet loss within the rate regulator was only observed in the POO3 tests which we thus marked with an asterisk (*). The loss can be explained with the infinite variance of the Pareto source which occasionally generated hundreds of data packets within a single ON interval. Since this does not reflect the behaviour of any application known to us, especially when we consider the average packet generation rate of 10 and the average data rate of 0.262 Mbit/s, we believe that cutting the extreme tail of the Pareto distribution actually led to more realistic results.

Column 8 in Table 3 shows the Packet Count (*pcnt)* which was used by the admission control for each application flow. The listed results were measured with the Time Window algorithm proposed in [9]. This algorithm was however not used for Pareto sources. For these, the admission control computed the Packet Count based on the fixed packet size of 1280 bytes. This removed the overhead typically introduced by the Time Window algorithm and enabled a resource allocation up to the capacity limit of the LAN.

In the following, we discuss the measurement results received in two selected test networks which we called the *1HDL*- and the *4HDL Test Network*. The 1HDL Test Network included two

Test Number	Trace	Encoding Scheme	Average Data Rate (Mbit/s)	Per-Flow Resources allocated at the Link Layer			
				Data Rate r (Mbit/s)	Burst Size δ (kbytes)	Max.Reg.Queue Length (Pkts)	*pcnt* (Pkts), (*TF* = 20ms)
1a	MMC2	JPEG Video	2.611	3.4	10.5	153	20
1b	MMC2	JPEG Video	2.611	3.4	10.5	153	20
1c	MMC2	JPEG Video	2.611	3.4	10.5	153	20
1d	MMC2	JPEG Video	2.611	3.4	25.4	144	29
1e	MMC2	JPEG Video	2.611	3.4	25.4	144	29
1f	MMC2	JPEG Video	2.611	3.4	25.4	144	29
2a	MMC1	JPEG Video	2.973	3.1	10.0	40	19
2b	MMC1	JPEG Video	2.973	3.1	10.0	40	19
2c	MMC1	JPEG Video	2.973	3.1	10.0	40	19
2d	MMC1	JPEG Video	2.973	3.1	23.0	31	26
2e	MMC1	JPEG Video	2.973	3.1	23.0	31	26
2f	MMC1	JPEG Video	2.973	3.1	23.0	31	26
3a	OVision	MPEG-1 Video	1.286	1.8	6.0	137	11
3b	OVision	MPEG-1 Video	1.286	1.8	6.0	137	11
3c	OVision	MPEG-1 Video	1.286	1.8	6.0	137	11
3d	OVision	MPEG-1 Video	1.286	1.8	15.8	129	20
3e	OVision	MPEG-1 Video	1.286	1.8	15.8	129	20
3f	OVision	MPEG-1 Video	1.286	1.8	15.8	129	20

Table 3. Source and Token Bucket Parameters for the Tests using the Application Traces *MMC1*, *MMC2* and *OVision*.

Test Number	Source Model	Average Data Rate (Mbit/s)	Peak /Average Rate Ratio	Per-Flow Resources allocated at the Link-Layer		
				Data Rate r (Mbit/s)	Burst Size δ (kbytes)	Max.Reg. Queue Length (Pkts)
4a	POO1	0.321	2	0.66	1.25	1
4b	POO1	0.321	2	0.66	1.25	1
4c	POO1	0.321	2	0.66	1.25	1
4d	POO1	0.321	2	0.60	5.0	1
4e	POO1	0.321	2	0.60	5.0	1
4f	POO1	0.321	2	0.60	5.0	1
5a	POO3	0.262	10	0.44	1.25	430 (*)
5b	POO3	0.262	10	0.44	1.25	430 (*)
5c	POO3	0.262	10	0.44	1.25	430 (*)
5d	POO3	0.262	10	0.44	2.50	430 (*)
5e	POO3	0.262	10	0.44	2.50	430 (*)
5f	POO3	0.262	10	0.44	2.50	430 (*)

Table 4. Source and Token Bucket Parameters for the Tests using the Pareto Sources.

LAN switches and is basically identically to the network illustrated in Figure 1. The 4HDL Test Network consisted of five LAN switches and four test links. All switches were interconnected by half-duplex switched 802.12 links. Each setup included real-time data flows traversing the test links in both directions. This typically allowed the admission of larger burst sizes, which then led to a stronger tail in the delay distribution than e.g. observed for full-duplex links. The results reported in this paper thus reflect the worst-case observations made during our experiments.

6.1 Delay and Loss Characteristics in the 1HDL Test Network

The setup of the 1HDL Test Network only differed by three additionally Traffic Clients from the setup shown in Figure 1. These were connected to switch Sw2, bore the network node numbers 10, 11 and 12, and generated cross traffic which entered the Test Link at switch Sw2. High priority cross traffic was sent by the nodes 2 - 9 (connected to Sw1) and the nodes 10 and 11 (connected to switch Sw2). Node 12 generated Best-Effort data traffic equivalent to more than 80 Mbit/s (which was not rate regulated) such that the Test Link between the switches Sw1 and Sw2 was always overloaded. This further used 1500 byte data packets to achieve the worst-case impact in respect to the normal priority service interrupt time.

In all experiments, we measured the end-to-end delay of a single flow sent by the Measurement Client. This did not include the delay in the rate regulator since we were interested in the characteristics of the queuing delay in the network. The measured delay was thus basically introduced in the high priority output queue of switch Sw1. Additionally, we measured: (1) the packet loss rate of the total traffic that entered the test link through Sw1, and (2) the average Controlled Load data rate. For the latter, we set appropriate filter entries in switch Sw2 such that a copy of all high priority data packets was also forwarded to a particular switch port (not shown). We then periodically read the MIB counter ifOutOctet [10] of this port, which counts the total number of bytes forwarded.

The details of the flow distribution in the test network and the measurement results are shown in Table 5 and Table 6. In all experiments, only homogeneous flows were admitted[1]. The

Test Number	Trace	Number of Flows admitted	Topology Information			Measured Parameters						
			Number of Flows sent from: Sw1 : Sw2	Number of Flows sent from Nodes: 2 ... 9	Number of Flows sent from Nodes: 10, 11	High Priority Data Rate (Mbit/s)	Pkt. Loss Rate (%)	Ave. Delay (ms)	90.0 % (ms)	99.0 % (ms)	Max. Delay (ms)	Ave. Packet Size (Bytes)
1a	MMC2	22	22 : 0	(7), 2	0, 0	59.120	0	1.407	2.865	6.325	13.275	1383
1b	MMC2	22	17 : 5	(2), 2	2, 3	58.775	0	1.322	2.605	5.735	12.025	1383
1c	MMC2	22	11 : 11	(3), 1	5, 6	59.127	0	1.018	1.615	3.745	10.515	1383
1d	MMC2	20	10 : 4	(2), 1	2, 2	37.972	0	1.173	1.775	4.395	10.485	1382
1e	MMC2	20	10 : 7	(2), 1	3, 4	45.656	0	1.477	2.895	6.745	15.525	1383
1f	MMC2	20	10 : 10	(2), 1	5, 5	53.086	0	1.690	3.525	8.025	19.365	1383
2a	MMC1	24	24 : 0	(2), 3	0, 0	71.775	0	1.743	3.835	6.855	13.545	1356
2b	MMC1	24	18 : 6	(3), 2	3, 3	71.820	0	1.643	3.665	7.235	16.615	1356
2c	MMC1	24	12 : 12	(4), 1	6, 6	71.901	0	1.160	2.125	4.535	11.745	1357
2d	MMC1	15	11 : 4	(3), 1	2, 2	44.933	0	1.219	2.155	5.265	14.425	1356
2e	MMC1	19	11 : 8	(3), 1	4, 4	56.825	0	1.517	3.175	7.155	15.655	1356
2f	MMC1	22	11 : 11	(3), 1	5, 6	65.917	0	1.947	4.435	9.335	18.985	1356
3a	OVision	42	42 : 0	(6), 5	0, 0	54.025	0	0.797	0.985	1.895	6.395	1332
3b	OVision	42	32 : 10	(3), 4	5, 5	54.125	0	0.786	0.975	1.885	6.785	1332
3c	OVision	42	21 : 21	(6), 2	10, 11	53.937	0	0.752	0.915	1.465	6.595	1332
3d	OVision	24	20 : 4	(5), 2	2, 2	30.750	0	0.721	0.825	1.015	5.655	1333
3e	OVision	30	20 : 10	(5), 2	5, 5	38.556	0	0.731	0.855	1.105	5.965	1333
3f	OVision	36	20 : 16	(5), 2	8, 8	46.741	0	0.750	0.885	1.385	7.375	1332

Table 5. Measured Packet Delay and Loss Rate for the Application Traces in the 1HDL Test Network.

Test Number	Source	Number of Flows admitted	Topology Information			Measured Parameters					
			Number of Flows sent from: Sw1 : Sw2	Number of Flows sent from Nodes: 2 ... 9	Number of Flows sent from Nodes: 10, 11	High Priority Data Rate (Mbit/s)	Pkt. Loss Rate (%)	Ave. Delay (ms)	90.0 % (ms)	99.0 % (ms)	Max. Delay (ms)
4a	POO1	132	132 : 0	(19), 16	0, 0	47.206	0	0.703	0.825	1.035	2.125
4b	POO1	132	102 : 30	(17), 12	15, 15	47.246	0	0.702	0.835	1.075	2.395
4c	POO1	132	66 : 66	(9), 8	33, 33	46.084	0	0.684	0.795	1.025	2.415
4d	POO1	60	50 : 10	(7), 6	5, 5	20.244	0	0.660	0.715	0.835	1.325
4e	POO1	80	50 : 30	(7), 6	15, 15	27.121	0	0.663	0.725	0.875	1.455
4f	POO1	100	50 : 50	(7), 6	25, 25	36.164	0	0.667	0.745	0.925	1.635
5a	POO3	202	202 : 0	(26), 25	0, 0	67.203	0	1.059	1.045	12.045	20.405
5b	POO3	204	154 : 50	(20), 19	25, 25	68.527	0	1.078	1.095	11.795	26.915
5c	POO3	204	104 : 100	(19), 12	50, 50	69.754	0	0.839	1.005	2.975	18.625
5d	POO3	141	101 : 40	(16), 12	20, 20	45.541	0	0.705	0.805	1.055	12.445
5e	POO3	171	101 : 70	(16), 12	35, 35	58.819	0	0.733	0.855	1.285	13.685
5f	POO3	201	101 : 100	(16), 12	50, 50	69.283	0	0.840	1.005	3.345	25.475

Table 6. Measured Packet Delay and Loss Rate for the Pareto Sources in the 1HDL Test Network.

maximum for this is provided in Column 3. Each result represents the allocation limit for the resources (δ, r) specified in Table 3 and Table 4. More specifically: all results for the tests: $a-c$ in Column 3 correspond to the bandwidth limit for the setup (additionally flows were rejected by the Bandwidth Test), whereas the results for the tests: d - f reflect the maximum buffer space in Sw1 (additionally flows were rejected by the Buffer Space Test). Column 4 shows the ratio of flows sent through switch Sw1 (*downstream*) into the Test Link versus the number of flows sent by switch Sw2 (*upstream*).

The columns 5 and 6 provide detailed information about how many flows entered the test network at each node. The first number (in brackets) in Column 5 specifies the number of flows that entered the network at node 2, whereas the second defines the number sent from each of the other nodes (3 - 9) connected to switch Sw1. This differentiation was required since the total number of flows could typically not be evenly distributed amongst all nodes. Column 6 shows the number of flows sent by node 11 and node 12. The setup for Test 1a thus for example included 7 MMC2 flows from node 2, and two from each of the nodes: 3 - 9. If we additionally consider the single flow sent by the Measurement Client, we get: $1 + 7 + 7 \cdot 2 = 22$ for the number of JPEG video flows on the Test Link in this experiment. In Test 1f, we had 2 flows from node 2, 1 flow from each of the

[1] The following parameters were used for the admission control: (1) a time frame of $TF = 20$ ms, (2) a per-packet overhead of: $D_{pp\_HD} = 8.555\ \mu s$ and a normal priority service interrupt time of: $D_{it\_HD} = 252.67\ \mu s$ corresponding to 100 m UTP cabling, (3) a minimum service rate of: $R_{MIN\_N1} = C_s / 2$ for each of the two LAN switches ($m = 2$), and (4) a buffer space of 256 kbytes for each high priority output queue in switch Sw1 and switch Sw2.

nodes: 3 - 9, and 5 flows from node 10 and node 11. This resulted in 20 MMC2 flows in the network. For each test, the measurement interval was again 30 minutes.

The measurement results are shown in the columns 7 - 12. In all the experiments reported in Table 5 and Table 6, we did not observe a single packet loss for high priority traffic. The loss measurements covered all real-time flows that entered the test link at switch Sw1. Since their total number was always at least as high as the total number of real-time flows from node 11 and node 12, it is very likely that we also had no packet losses at the high priority output queue of switch Sw2. The largest bandwidth reservations were made in the tests 5b and 5c, in which we allocated: $204 \cdot 0.44$ Mbit/s $=$ 89.76 Mbit/s for Controlled Load traffic on the link. We further find that all results for the average delay are in the order of 1 ms. This is sufficient for a Controlled Load service.

The highest average delays were measured in the MMC1 and MMC2 tests. Even though the MMC1 trace is less bursty than the MMC2 trace, we measured a lower average delay for the latter. This is because more resources were allocated for each individual MMC2 flow (as shown in Table 3) which led to a lower high priority data rate on the test network and thus to a lower average packet delay. Furthermore, the results for both video traces MMC1 and MMC2 are higher than the results achieved with the Pareto sources. This behaviour could however already be observed in Figure 5 and Figure 6 in Section 2.3.

The lowest average delays (~0.7 ms) were received with the Pareto source model POO1. This is not surprising since in these experiments, we also reserved resources at peak data rate (tests a - c) or very close to the peak data rate (tests d - f) of each flow.

The largest maximum delays were measured in the POO3 tests. A long tail in the distribution of the results can be identified. This is similar to the characteristics found for this source in Figure 6. Note that result for test 5f (25.475 ms) is higher than the maximum delay observed in Figure 2 for a switch with 256 kbytes buffer space. This can be explained by the different experimental setup used: in Section 2.2, all traffic entered the test link at the same switch, whereas in Test 5f the half-duplex switched link was loaded from both switches. In the worst case, when the output queues of Sw1 and Sw2 are full and both queues receive the same service (equivalent to $C_s/2$), then the maximum delay could be as high as twice ($\sim 2 \cdot 23 = 46$ ms) the result observed in Figure 2. We may thus see long maximum delays in bridged networks, provided this consists of bridges interconnected by half-duplex switched links, and the traffic is bursty over long time scales.

In contrast, the results received in the MMC1 and MMC2 measurements are typically significantly lower despite of the higher average delays measured for them. To illustrate the differences between the application- and the source model traces we plotted the delay distribution function of the MMC1 tests: 2d - 2f and the POO3 tests 5d - 5f in Figure 8. Figure 9 and Figure 10 show the corresponding distribution density for Test 2f and Test 5f. Similar graphs can be plotted for the other results, but are omitted here. We selected the tests: 2f and 5f for illustration because they provided the largest results for the maximum delays in the 4HDL Test Network discussed in the following.

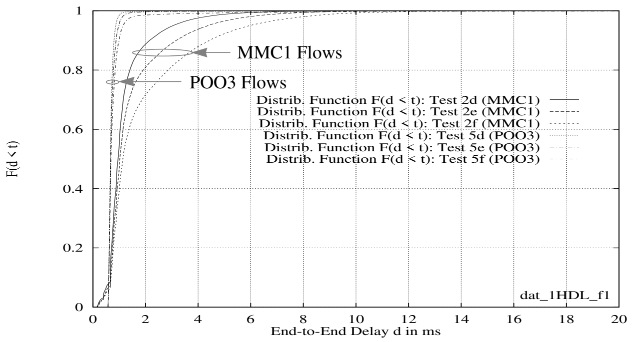

Figure 8. Delay Distribution Function for the Tests 2d - 2f (MMC1) and 5d - 5f (POO3) in the 1HDLTest Network.

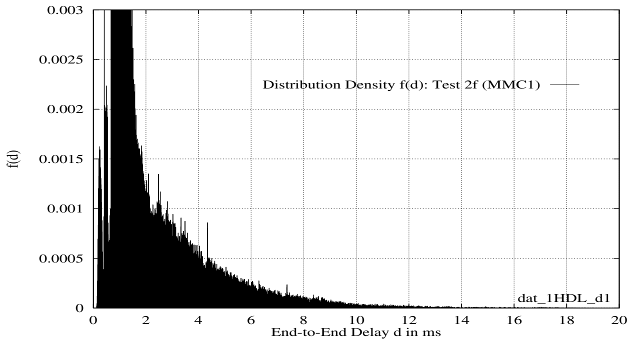

Figure 9. Delay Distribution Density corresponding to Test 2f (MMC1, 1HDL Topology) in Figure 8.

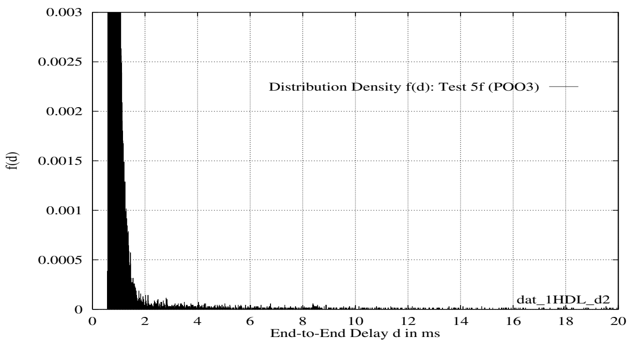

Figure 10. Delay Distribution Density corresponding to Test 5f (POO3, 1HDL Topology) in Figure 8.

6.2 Delay and Loss Characteristics in the 4HDL Test Network

Figure 11 illustrates the measurement setup in the 4HDL Test Network. The upper part of the picture shows the network topology, the lower part the path of the data flows during the experiments. Controlled Load flows entered the test network at the nodes: 2 - 13 (high priority cross traffic) and at the Measurement Client (measurement traffic). The latter was connected to switch Sw1 (sending interface - 1a) and switch Sw5 (receiving interface - 1b), respectively.

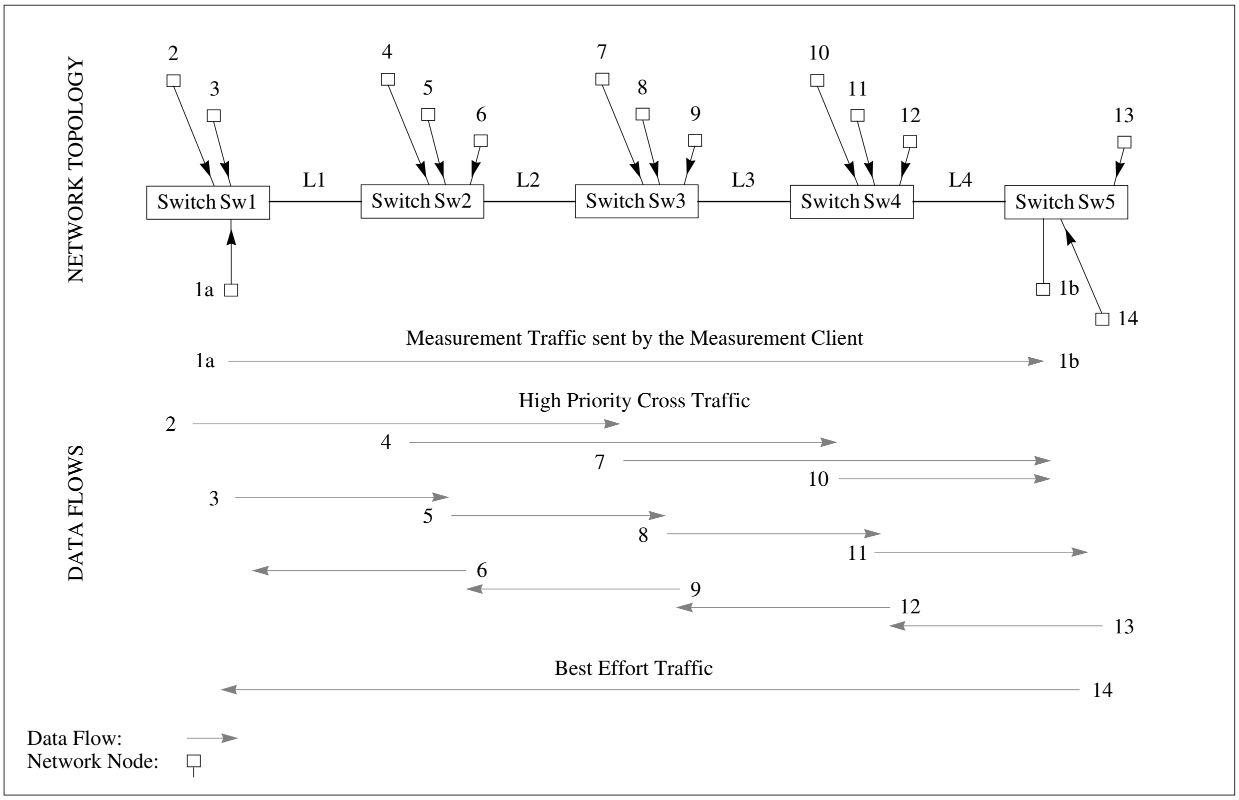

Figure 11. Measurement Setup in the 4HDL Test Network.

All flows from the nodes: 2, 4, 7 and 10 traversed two half-duplex links along the data path of the measurement traffic as illustrated in Figure 11. In contrast, the flows sent from the nodes: 3, 5, 8 and 11 only travelled across a single link downstream with the measurement traffic. Upstream Controlled Load flows were generated at the nodes: 6, 9, 12 and 13 and also only forwarded across a single half-duplex switched link. The particular data path of each flow was enforced by installing appropriate filter entries in the LAN switches. Best Effort traffic was only sent by node 14. It traversed the entire upstream data path from switch Sw5 to switch Sw1 and had the same characteristics as in the previous experiments.

In this topology, we measured the packet loss rate at the output queue of the switches: Sw1 (to link L1), Sw2 (to L2), Sw3 (to L3) and Sw4 (to L4). In switch Sw1 for example, this included the real-time flows from the nodes: 2, 3 and the measurement traffic; in switch Sw4, this detected loss of data packets from the nodes: 7, 10, 11 and from the Measurement Client. The average high priority data rate was recorded for link L4. The Measurement Client measured the delay for all data packets of a single flow which traversed the entire test network.

Since the test network had a half-duplex switched topology, the admission control could use the same topology specific parameters (D_{pp_HD}, D_{it_HD}, R_{MIN_N1}, $m = 2$) as used in the single link (1HDL) test network. In contrast to the link-by-link admission policy in which the buffer space requirements grow along the data path, we configured the admission control such that it did not increase the buffer space allocation for flows traversing several switches. Instead, the admission control treated all Controlled Load flows as if these had entered the test network at the corresponding local switch, neglecting any traffic distortions along the data path already traversed by these flows. This led to the same bandwidth- and buffer space reservations on all half-duplex links in the bridged test network.

Such an allocation strategy assumes that the statistically distributed use of buffer space in switches can compensate for the traffic distortions introduced in the network. We believe that in a bridged LAN environment, this is likely to be the case whenever resources are conservatively allocated at the edge of the network. This is supported by the facts that the number of LAN switches within the data path is typically small, and feedback effects can not occur in LANs. Furthermore, in real networks we do not expect reservations to be made up to the capacity limit of the network to prevent the starvation of the Best-Effort service. More realistic (Controlled Load) Utilization Factors f of $f \le 0.9$ however are likely to enforce sufficient spare bandwidth such that packet loss for real-time flows is eliminated or at least significantly reduced even in large topologies. In our experiments we nevertheless allocated resources up to the (theoretical) link capacity ($f = 1.0$) to test the worst case.

Table 7 and Table 8 contain the results measured in the 4HDL Test Network. Figure 12, Figure 13 and Figure 14 show the end-to-end delay distribution function and the distribution density for selected MMC1 and POO3 tests. These represent the equivalent graphs to Figure 8, Figure 9 and Figure 10 shown for the 1HDL Test Network in the previous section.

Test Number	Trace	Number of Flows admitted (L1, L2, L3, L4)	Topology Information					Measured Parameters					
			Number of Flows sent: *downstream : upstream*	Number of Flows sent from Nodes:			Cross-Switch Reser-vation (Mbit/s)	High Priority Data Rate on L4 (Mbit/s)	Pkt. Loss Rate (%)	Ave. Delay (ms)	90.0 % (ms)	99.0 % (ms)	Max. Delay (ms)
				2, 4, 7, 10	3, 5, 8, 11	6, 9, 12, 13							
1a	MMC2	22	22 : 0	7	(14),7	0	27.20	58.357	0	3.541	6.975	12.195	23.635
1b	MMC2	22	17 : 5	8	(8), 0	5	30.60	58.623	0	3.746	7.525	12.845	26.385
1c	MMC2	22	11 : 11	5	(5), 0	11	20.40	59.016	0	2.670	5.105	9.545	21.735
1d	MMC2	20	10 : 4	4	(5), 1	4	17.00	37.272	0	2.588	4.645	8.675	20.525
1e	MMC2	20	10 : 7	4	(5), 1	7	17.00	45.972	0	3.595	6.925	12.395	25.445
1f	MMC2	20	10 : 10	4	(5), 1	10	17.00	54.065	0	4.249	8.395	14.705	30.535
2a	MMC1	24	24 : 0	11	(12),1	0	37.20	71.789	0	5.448	9.965	15.215	25.065
2b	MMC1	24	18 : 6	6	(11),5	6	21.70	71.791	0	6.655	12.535	19.445	33.555
2c	MMC1	24	12 : 12	5	(6), 1	12	18.60	71.875	0	5.660	11.015	17.094	31.885
2d	MMC1	15	11 : 4	5	(5),0	4	18.60	44.887	0	3.158	6.125	10.695	19.815
2e	MMC1	19	11 : 8	5	(5),0	8	18.60	56.936	0	4.988	9.995	16.795	33.165
2f	MMC1	22	11 : 11	5	(5),0	11	18.60	66.104	0	6.393	12.655	20.145	36.625
3a	OVision	42	42 : 0	20	(21),1	0	37.80	53.973	0	1.561	1.965	3.655	12.195
3b	OVision	42	32 : 10	11	(20),9	10	21.60	54.320	0	1.679	2.195	4.285	14.135
3c	OVision	42	21 : 21	10	(10,0	21	19.80	53.519	0	1.668	2.135	3.865	13.595
3d	OVision	24	20 : 4	8	(11),3	4	16.20	30.245	0	1.361	1.575	1.835	8.085
3e	OVision	30	20 : 10	8	(11),3	10	16.20	38.143	0	1.442	1.715	2.225	8.025
3f	OVision	36	20 : 16	8	(11),3	16	16.20	46.225	0	1.571	1.935	3.085	12.855

Table 7. Measured Packet Delay and Loss Rate for the Application Traces in the 4HDL Test Network.

Test Number	Source	Number of Flows admitted (L1, L2, L3, L4)	Topology Information					Measured Parameters					
			Number of Flows sent: *downstream : upstream*	Number of Flows sent from Nodes:			Cross-Switch Reser-vation (Mbit/s)	High Priority Data Rate on L4 (Mbit/s)	Pkt. Loss Rate (%)	Ave. Delay (ms)	90.0 % (ms)	99.0 % (ms)	Max. Delay (ms)
				2, 4, 7,10	3, 5, 8, 11	6, 9, 12,13							
4a	POO1	132	132 : 0	65	(66), 1	0	43.56	47.535	0	1.363	1.625	1.955	7.955
4b	POO1	132	102 : 30	50	(52), 1	30	33.66	47.105	0	1.383	1.655	2.015	7.815
4c	POO1	132	66 : 66	32	(33), 1	66	21.78	48.278	0	1.421	1.705	2.085	7.735
4d	POO1	60	50 : 10	20	(29), 9	10	12.60	20.536	0	1.217	1.355	1.535	7.995
4e	POO1	80	50 : 30	20	(29), 9	20	12.60	26.641	0	1.272	1.455	1.685	7.445
4f	POO1	100	50 : 50	20	(29), 9	50	12.60	36.593	0	1.326	1.545	1.815	7.665
5a	POO3	202	202 : 0	90	(111),21	0	44.44	66.786	0	5.048	16.415	38.375	65.755
5b	POO3	204	154 : 50	50	(103),53	50	22.44	67.738	0	7.174	23.855	48.485	88.405
5c	POO3	204	104 : 100	50	(53), 3	100	22.44	69.872	0	5.482	16.755	48.555	78.315
5d	POO3	141	101 : 40	50	(50), 0	40	22.44	46.478	0	2.214	1.855	24.835	52.875
5e	POO3	171	101 : 70	50	(50), 0	70	22.44	57.192	0	3.289	5.025	36.705	63.985
5f	POO3	201	101 : 100	50	(50), 0	100	22.44	69.536	0	6.129	19.325	54.125	88.015

Table 8. Measured Packet Delay and Loss Rate for the Pareto Sources in the 4HDL Test Network.

The organization of the results is similar as discussed for the 1HDL Test Network. There are only a few minor differences which we clarify in the following. Column 3 shows the number of flows admitted on *each* of the four half-duplex switched links. The columns 4 - 8 again contain informations about the flow distribution. This uses the same notation as described in the previous section. In Test 1a (MMC2) for example, 7 flows entered the network at each of the nodes: 2, 4, 7 and 10. The setup additionally included 14 flows sent by node 3 and 7 flows generated at each of the nodes: 5, 8 and 11. The nodes: 6, 9, 12, 13 did not pass any flows into the test network in this experiment. Considering the single flow sent by the Measurement Client, and the path information for the bridged test network in Figure 11, we have: 1 + 14 + 7 = 22 flows for link L1, and: 1 + 7 + 7 + 7 = 22 MMC2 flows for the links: L2, L3 and L4. Switch Sw2 then only forwarded the 7 flows from node 2 and the single flow from the Measurement Client onto link L2. This resulted in a Cross Switch Reservation of: $8 \cdot 3.4\ \text{Mbit/s} = 27.2\ \text{Mbit/s}$ from link L1 to L2. The same result is received for all other switches, which is why we listed it in Column 8 in Table 7 and Table 8.

Test 1f included: (1) 4 flows from each of the nodes: 2, 4, 7 and 10, (2) 5 flows from node 3 and 1 flow sent by each of the nodes: 5, 8, 11, and (3) 10 flows generated at each of the nodes: 6, 9, 12 and 13. This led to 20 MMC2 video flows on each of the half-duplex switched links. For the Cross Switch Reservation, we receive: $(1 + 4) \cdot 3.4\ \text{Mbit/s} = 17.0\ \text{Mbit/s}$ in this case.

The columns 9 - 14 show the measured parameters. We omitted the results for the average packet size since these were basically identical to those in Table 5. As in the previous tests, the measurement interval for each test was 30 minutes.

We can first observe that all results for the packet loss rate are zero. In all tests in the 4HDL Test Network, we did not detect the loss of a single Controlled Load data packet in any of the five LAN switches. This shows that the admission control was sufficiently conservative to prevent buffer overflow within the switches despite: (1) a worst case allocation (f = 1.0) on each link in the test network, and (2) the use of data sources with long range burst characteristics. Looking at the results in Column 11, we find that in some tests, the average delay increased significantly (MMC2, MMC1, POO3) in comparison to the single link (1HDL) topology, whereas this did not occur to the same extent with other sources (OVision, POO1). The largest growth of the average delay can be observed in the POO3 tests (214 - 630%) even though the absolute results (2.2 - 7.2 ms) are still in the order of the values received for MMC1 (3.2 - 6.7 ms). For the video sources MMC1 and MMC2 we achieved growth rates of: 159 - 388% and 120 - 183%, respectively. In contrast to this, the results in the POO1 experiments only increased by: 84 - 108%.

Even though the average delays received in the MMC1 and POO3 tests are similar, their distributions nevertheless differ significantly. This can be seen in Figure 12. The results for MMC1 are only distributed over a short time range. We measured maximum delays of just: 19.8 - 36.6 ms. This corresponds to the behaviour observed for this trace in the 1HDL Test Network.

The results obtained for the POO3 tests again exhibit a strong tail in their distribution with maximum delays of: 52.9 - 88.4 ms. Even for the 99.0 percentile we still have values between: 24.8 - 54.1 ms. Considering the distribution in Figure 14, one can expect even larger maxima when the experiments are repeated with longer measurement intervals than 30 minutes. All recorded results are however still significantly smaller than the theoretical maximum (~184 ms) of the queuing delay in the 4HDL Test Network.

The maximum delays obtained with each of the other four traces are much smaller than the results achieved with the POO3 sources. The lowest values were again measured in the POO1 tests (7.4 - 7.9 ms) due to the peak rate allocation strategy applied for this source. In general, we found that the results for each source type exhibited the same fundamental characteristics (e.g. short or long range distribution) in all test topologies used in our experiments. This can for example be observed in the graphs shown for MMC1 Test 2f (Figure 9, Figure 13) and POO3 Test 5f (Figure 10, Figure 14).

We further found it hard to quantify the impact of the network topology and of the flow distribution on the delay characteristics. A conclusion other than that this impact is difficult to predict, can not be drawn from our results. Some of the factors that influence the delay and loss characteristics are: (1) the resources allocated for each Controlled Load flow, leading to a certain burstiness of the total traffic and thus to a certain average high priority data rate on the link, (2) the amount of cross traffic that traverses several switches, (3) the ratio of the flows forwarded upstream and downstream in respect to the measurement traffic, (4) the number of network nodes with Traffic Clients connected to each switch, (5) speed mismatches between input and output links (not tested), and (6) the traffic characteristics and the data path of the best effort traffic.

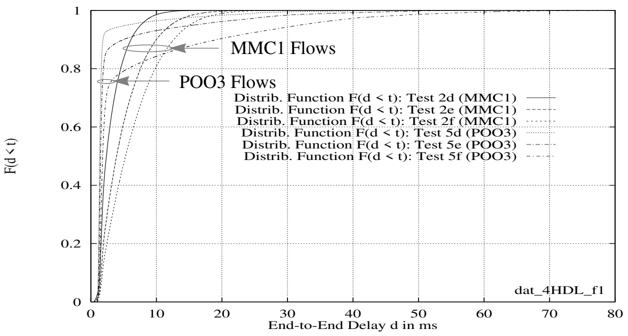

Figure 12. Delay Distribution Function for the Tests 2d - 2f (MMC1) and 5d - 5f (POO3) in the 4HDLTest Network.

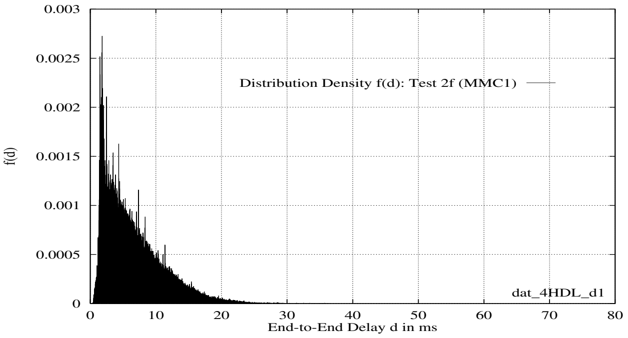

Figure 13. Delay Distribution Density corresponding to Test 2f (MMC1, 4HDL Topology) in Figure 12.

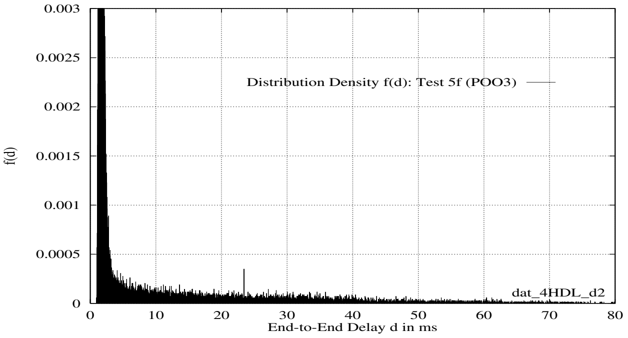

Figure 14. Delay Distribution Density corresponding to Test 5f (POO3, 4HDL Topology) in Figure 12.

During our experiments we however found that although these dependencies changed the results, they never led to a significant increase of the average delay, persistent packet loss in the network, or consistently higher maximum delays than discussed in this section. We thus believe that a Controlled Load service can be provided by using admission control, but without any per-flow traffic control mechanisms in LAN switches (e.g. just simple Static Priority scheduling and P802.1p/Q classification) - independent whether these switches are interconnected by shared 802.12 multihub segments or half-duplex- or full-duplex switched links.

6.3 Resource Utilization Issues

Another evaluation criteria beside the quality of the new service is the resource utilization that can be achieved with a particular admission control algorithm. Table 9 shows the minimum- and maximum results observed for each source type during the experiments in the 4HDL Test Network.

Since the packet size used for the data transmission has a strong impact on the data throughput, we used packet size specific measurement results for the computation of the resource utilization. The packet sizes listed in Column 2 for MMC2, MMC1 and OVision are average results and were taken from Column 13 in Table 5. The Pareto sources used a fixed packet size of 1280 byte. Column 3 in Table 9 shows the worst-case data throughput measured on a half-duplex switched link for the packet sizes in Column 2. The columns 4 and 5 contain the minimum and maximum results for the measured Controlled Load data rate. These results come from Table 7 and Table 8.

Trace / Source Model	Packet Size	Measured Parameters			Min. Resource Utilization (%)	Max. Resource Utilization (%)
		Max. Throughput (Mbit/s)	Min. Data Rate (Mbit/s)	Max. Data Rate (Mbit/s)		
MMC2	1383	93.35	37.27	59.02	39.92	63.22
MMC1	1356	93.26	44.89	71.87	48.13	77.06
OVision	1332	93.15	30.24	54.32	32.46	58.31
POO1	1280	92.91	20.54	48.28	22.11	51.96
POO3	1280	92.91	46.48	69.87	50.03	75.20

Table 9. Minimum and Maximum Resource Utilization for the Measurement Results in Table 7 and Table 8.

Table 9 indicates that a large resource utilization can be achieved with our admission control. Most of the results are however optimistic since we typically allocated resources close to the average data rate of real-time flows. More realistic are the results for the Pareto POO1 source (22.11 - 51.96%) because in these tests resources were reserved at peak data rate. Even with a loose allocation policy caused by the often occuring difficulties in selecting appropriate Token Bucket parameters, we believe that a sufficiently large number of Controlled Load flows can be supported to make the deployment of our approach worthwhile.

7 Related Work

We are not aware of any approach to provide Controlled Load quality of service that was specifically designed for IEEE 802 type LANs (or 802.12 networks). Much research has however been performed in the context of wide area networks.

The simplest approach is probably the Simple Sum algorithm. This was used in [15] to provide Controlled Load quality of service. The underlying service discipline was Weighted Fair Queuing (WFQ) which isolated real-time traffic on a per-flow basis in each router in the network. In contrast to this, our simple Static Priority scheduler cannot control the Controlled Load flows which may thus become burstier as they traverse the switched LAN. This may lead to a much stronger tail of the end-to-end packet delay distribution which we can also observe when we compare our results with the simulation results in [15].

Many approaches for admission control have been proposed based on the concept of the *Effective Bandwidth*. These could be used to provide Controlled Load service. The concept includes the computation of the bandwidth requirement $C(\varepsilon)$ (the Effective Bandwidth) for a class of flows such that their stationary data arrival rate exceeds $C(\varepsilon)$ with a probability of not more than ε. More formally [24]: $Prob(C(\varepsilon) < R_A) \leq \varepsilon$, where R_A denotes the aggregated data rate and ε the overflow probability.

One approach to compute the Effective Bandwidth is to choose a statistical source model for the data arrival process at a switch and to select appropriate values for the model parameters. Afterwards the effective bandwidth is derived based on ε and the model. This approach was used in: [24], [25], [26], [27], [28], [29], [30]. Many of these schemes are measurement-based since they also use input parameters which are estimated using on-line measurements within switches or routers. An alternative approach is based on the theory of Large Deviation. Instead of choosing a statistical source model, the authors of [31], [32], [33] estimate the large deviation rate function, which is directly related to the Effective Bandwidth. This uses load measurements within the (ATM) switches over pre-defined time blocks.

In spite of the rich research on statistical service guarantees, we used a parameter based approach to provide Controlled Load service guarantees in 802.12 networks. This was because: (1) most of the above admission control schemes depend on a full-duplex switched network and can thus not easily be reused for admission control in 802.12 LANs. (2) We focused on low costs which basically excluded any per-flow traffic control or complex measurement based approaches within LAN switches. (3) Furthermore, we believe that accurate probabilistic end-to-end service guarantees will be hard to derive in bridged LANs. The task is particulary hard in 802.12 networks because: (1) the medium access can be shared, (2) the data throughput is not constant, and (3) the use of Static Priority scheduling in switches which enables interactions between different Controlled Load flows.

8 Conclusions

In this paper, we have shown how Controlled Load quality of service could be enforced in a bridged/switched LAN. We first defined the packet scheduling process and then provided admission control conditions. The latter were specific for IEEE 802.12 networks. The second part of the paper included a performance evaluation of the new LAN service.

The Controlled Load service was built based on simple Static Priority scheduling in LAN switches, and traffic policing and reshaping mechanisms deployed only in hosts and routers at the entrance to the bridged network. We believe that this approach can be used independently of the LAN technology and thus for a variety of high speed LANs, including shared, half-duplex and full-duplex switched link topologies, provided that the underlying link technologies have a deterministic medium access and the service access for each link is restricted by admission control. Unlike the packet scheduling process, the admission control conditions are link technology specific. Our conditions consider the Demand Priority protocol overhead to accurately provide the required service quality.

To obtain most realistic evaluation results, we built the Controlled Load service in an 802.12 test network. Our measurement results showed that the Controlled Load service quality can be enforced under worst-case load and traffic conditions. Packet loss in the network will be extremely rare, in particular with realistic Utilization Factors of: $f < 1$. We also showed that the maximum delays in bridged networks can potentially be very large when the data sources are bursty over long time scales. The measured results for the average delay however always remained in the order of a few milliseconds which is sufficient to support existing time sensitive, but adaptive applications.

The main advantages of the approach presented in this paper are its simplicity and applicability. Both also represent the main differentiator to many solutions discussed in the previous section. Our scheme only depends on static priority scheduling and 802.1p/Q classification within the network. This will ensure low implementation costs. Since both mechanisms will be supported by many next generation switch products, the Controlled Load service could immediately be deployed, provided network nodes are able to regulate real-time flows. Several operating system vendors have already announced the support for rate regulation in future product releases. Furthermore, only those nodes which use the Controlled Load service (or other services based on the high priority medium access) would have to be updated.

References

[1] D. Clark, S. Shenker, L. Zhang, *Supporting Real-Time Applications in an Integrated Services Packet Network: Architecture and Mechanism*, in Proc. of ACM SIGCOMM'92, pp. 14 - 26, August 1992.

[2] P. P. White, J. Crowcroft, *The Integrated Services in the Internet: State of the Art*, in Proceedings of the IEEE, Vol. 85, No. 12, pp. 1934 - 1946, December 1997.

[3] S. Shenker, C. Partridge, R. Guerin, *Specification of the Guaranteed Quality of Service*, IETF, RFC 2212, September 1997.

[4] J. Wroclawski, *Specification of the Controlled-Load Network Element Service*, IETF, RFC 2211, September 1997.

[5] A. Ghanwani, J. W. Pace, V. Srinivasan, A. Smith, M. Seaman, *A Framework for Providing Integrated Services Over Shared and Switched IEEE 802 LAN Technologies*, Internet Draft draft-ietf-issll-is802-framework-04.txt, March 1998.

[6] ISO/IEC, IEEE P802.1p/D8 - *Supplement to MAC Bridges: Traffic Expediting and Dynamic Multicast Filtering*, September 1997.

[7] ISO/IEC, IEEE P802.1Q/D8 - *Draft Standard for Virtual Bridged Local Area Networks*, September 1997.

[8] Hewlett-Packard, *PA-RISC 1.1 Architecture and Instruction Set*, Reference Manual, Part No: 09740-90039, September 1992.

[9] P. Kim, *Resource Reservation in Shared and Switched Demand Priority Local Area Networks*, PhD Dissertation, University of London, September 1998, available from UCL at: http://www.cs.ucl.ac.uk/staff/J.Crowcroft/paststudents.html.

[10] K. McCloghrie, M. Rose, *Management Information Base for Network Management of TCP/IP-based Internets: MIB-II*, IETF, RFC 1213, March 1991.

[11] OptiVision Inc., *OptiVision Live MPEG Communication System,* User's Guide, Version 1.2 f, September 1996.

[12] N. Leymann, *Eine Videokomponente für das Videokonferenzsystem Multimedia Collaboration*, Diploma Thesis, in German, Technical University of Berlin, August 1996.

[13] M. Garret, W. Willinger, *Analysis, Modelling and Generation of Self-Similar VBR Video Traffic*, in Proc. of ACM SIGCOMM'94, pp. 269 - 279, September 1994.

[14] W. Willinger, M. S. Taqqu, R. Sherman, D. V. Wilson, *Self-Similarity Through High Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level*, in Proc. of ACM SIGCOMM'95, pp.100 - 113, August 1995.

[15] S. Jamin, S. Shenker, P. Danzig, *Comparison of Measurement-based Admission Control Algorithms for Controlled Load Service*, in Proc. of IEEE INFOCOM'97, pp. 973 - 980, April 1997.

[16] P. Kim, *Deterministic Service Guarantees in 802.12 Networks, Part I: the Single Hub Case*, in IEEE Transactions on Networking, Vol. 6, No. 5, pp. 645 - 658, October 1998.

[17] P. Kim, *Deterministic Service Guarantees in 802.12 Networks, Part II: the Cascaded Network Case*, in Proc. of IEEE INFOCOM'98, pp. 1376 - 1383, March 1998.

[18] International Telecommunication Unit (ITU-T), General Characteristics of International Telephone Connections and International Telephone Circuits, *Recommendation G.114 - One-Way Transmission Time*, February 1996.

[19] L.C. Wolf, C. Griwodz, R. Steinmetz, *Multimedia Communication*, in Proceedings of the IEEE, Vol. 85, No. 12, pp. 1925 - 1932, December 1997.

[20] R. Cruz, *A Calculus for Network Delay Part I: Network Elements in Isolation*, in IEEE Transactions on Information Theory, Vol. 37, No. 1, pp. 114 - 131, January 1991.

[21] ISO/IEC, *ANSI/IEEE Standard 802.12 - Demand Priority Access Method, Physical Layer and Repeater Specification for 100 Mbit/s Operation*, November 1995.

[22] ISO/IEC, *ANSI/IEEE Standard 802.1D - Media access control (MAC) bridges*, July 1993.

[23] H. Zhang, *Service Disciplines for Guaranteed Performance Service in Packet-Switching Networks*, in Proceedings of the IEEE, Vol. 83, No. 10, pp. 1373 - 1396, October 1995.

[24] R. Guerin, H. Ahmadi, M. Naghshineh, *Equivalent Capacity and Its Application to Bandwidth Allocation in High-Speed Networks*, in IEEE JSAC, Vol. 9, No. 7, pp. 968 - 981, Sept. 1991.

[25] G. Kesidis, J. Walrand, C. Chang, *Effective Bandwidth for Multiclass Markov Fluids and Other ATM Sources*, in IEEE Transactions on Networking, Vol. 1, No. 4, pp. 424 - 428, August 1993.

[26] S. Abe, T. Soumiya, *A Traffic Control Method for Service Quality Assurance in an ATM Network*, in IEEE JSAC, Vol. 12, No. 2, pp. 322 - 331, February 1994.

[27] R. Gibbens, F. Kelly, P. Key, *A Decision-Theoretic Approach to Call Admission Control in ATM Networks*, in IEEE JSAC, Vol. 13, No. 6, pp. 1101 - 1113, August 1995.

[28] S. Floyd, *Comments on Measurement-based Admission Control for Controlled Load Services*, draft submitted to ACM Computer Communication Review, (also available at: http://www-nrg.ee.lbl.gov/floyd/), July 1996.

[29] R. Gibbens, F. Kelly, *Measurement-based connection admission control*, in 15th International Teletraffic Congress, June 1997.

[30] Z. Dziong, M. Juda, L. Mason, *A Framework for Bandwidth Management in ATM Networks - Aggregate Equivalent Bandwidth Estimation Approach*, in IEEE JSAC, Vol. 5, No. 1, pp. 134 - 147, February 1997.

[31] N. G. Duffield, J. T. Lewis, N. O'Connell, R. Russell, F. Toomey, *Entropy of ATM Traffic Streams: A Tool for Estimating QoS Parameters*, in IEEE JSAC, Vol. 13, No. 6, pp. 981 - 990, August 1995.

[32] S. Crosby, I. Leslie, J. Lewis, N. O'Connell, R. Russell, F. Toomey, *Bypassing Modelling: an Investigation of Entropy as a Traffic Descriptor in the Fairisle ATM network*, in Proc. of 12th U.K. Teletraffic Symposium (UKTS 95), February 1995.

[33] R. Vesilo, V. Solo, *Techniques for Adaptive Estimation of Effective Bandwidth in ATM Networks*, in IEEE Globecom, pp. 1344 - 1348, November 1997.