A Pragmatic Approach for Service Provisioning Based on a Small Set of Per-Hop Behaviors

Volker Sander  
Central Institute for Applied Mathematics,  
Forschungszentrum Jülich GmbH, 52425 Jülich, Germany  
Email: sander@fz-juelich.de

Markus Fidler  
Department of Computer Science, Aachen University  
Ahornstr. 55, 52074 Aachen, Germany  
Email: fidler@informatik.rwth-aachen.de

Abstract—In this paper we describe the implementation of a network providing advanced services such as a Premium service that aims at providing low loss, low delay, and low delay jitter and an Olympic service that allows for a service differentiation in terms of delay within three additional classes. Our implementation of this network is based on the Differentiated Services Architecture, which is the most recent approach of the Internet Engineering Task Force towards Quality of Service. Access to service classes is controlled by a Bandwidth Broker, which can perform Traffic Engineering by means of Multiprotocol Label Switching. The Premium service is implemented as Expedited Forwarding and the Olympic service as a group of Assured Forwarding Per-Hop-Behavior. We present a thorough evaluation of the proposed services implemented by the careful assignment of micro-flows to a small set of Per-Hop Behaviors.

I. INTRODUCTION

Emerging Grid applications consist of complex mixes of flows, with a variety of data rates and widely differing latency requirements. Applications with these characteristics arise in areas as remote visualization, analysis of scientific databases, and teleimmersion. Table I lists the various flows of a future teleimmersion application together with their networking demands [6]. Characteristics such as these place substantial demands on networks which cannot be fulfilled by today’s Best-Effort (BE) Internet.

<table>
<thead>
<tr>
<th>Flow Type</th>
<th>Latency</th>
<th>Bandwidth</th>
<th>Reliability</th>
<th>Dynamic QoS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>&lt; 50 ms</td>
<td>64 Kb/s</td>
<td>Yes</td>
<td>Low</td>
</tr>
<tr>
<td>text</td>
<td>&lt; 100 ms</td>
<td>64 Kb/s</td>
<td>Yes</td>
<td>Low</td>
</tr>
<tr>
<td>Audio</td>
<td>&lt; 30 ms</td>
<td>128 Kb/s</td>
<td>No</td>
<td>Medium</td>
</tr>
<tr>
<td>Video</td>
<td>&lt; 100 ms</td>
<td>5 Mb/s</td>
<td>No</td>
<td>Medium</td>
</tr>
<tr>
<td>Tracking</td>
<td>&lt; 10 ms</td>
<td>128 Kb/s</td>
<td>No</td>
<td>Medium</td>
</tr>
<tr>
<td>Database</td>
<td>&lt; 100 ms</td>
<td>&gt; 1 Gb/s</td>
<td>Yes</td>
<td>High</td>
</tr>
<tr>
<td>Simulation</td>
<td>&lt; 50 ms</td>
<td>&gt; 1 Gb/s</td>
<td>Mixed</td>
<td>High</td>
</tr>
<tr>
<td>Haptic</td>
<td>&lt; 10 ms</td>
<td>&gt; 1 Mb/s</td>
<td>Mixed</td>
<td>High</td>
</tr>
<tr>
<td>Rendering</td>
<td>&lt; 30 ms</td>
<td>&gt; 1 Gb/s</td>
<td>No</td>
<td>Medium</td>
</tr>
</tbody>
</table>

TABLE I  
FLOWS AND REQUIREMENTS OF TELEIMMERSION APPLICATIONS

The Differentiated Services (DS) [2] framework defines an architecture for implementing scalable service differentiation in the existing Internet by an aggregation of flows to a small number of different traffic classes. DS can be complemented by Traffic Engineering (TE), which is of special interest, if not only a relative differentiation between services shall be implemented, but, if absolute service guarantees need to be given. We apply Multi-Protocol Label Switching (MPLS) [13] for this purpose. MPLS is based on a functional decomposition of the network layer into a control component and a forwarding component. This distinction gives a number of options for the implementation of the control component. In [9], [15] we presented the General-purpose Architecture for Reservation and Allocation (GARA), which implements an advance reservation framework using heterogeneous resource ensembles. GARA includes a DS reservation manager, which can be used as a Bandwidth Broker (BB) for an automated DS network management and, which allows to apply DS TE by means of MPLS.

In this paper, we describe an implementation of a Premium service based on the Expedited Forwarding (EF) Per-Hop-Behavior (PHB) [5] and an Olympic service based on the Assured Forwarding (AF) PHB group [10]. We perform a careful evaluation of the services by applying both Transport Control Protocol (TCP) and User Datagram Protocol (UDP) flows, and address the question about the impact on the achievable service when these heterogeneous applications share a single PHB. The remainder of the paper is organized as follows: Sections II describes the DS Architecture in detail. In section III we show results obtained from measurements in the implemented network. Section IV concludes the paper.

II. DIFFERENTIATED SERVICES ARCHITECTURE

The DS architecture [2] addresses the scalability problems of Integrated Services by defining the behavior of aggregates. Packets are identified by simple markings that indicate according to which aggregate behavior they should be treated. In the core of the network, routers need not determine to which flow a packet belongs, only which aggregate behavior should be used. Edge routers mark packets and indicate whether they are within profile or if they are out of profile, in which case they might even be discarded by a dropper at the edge router. A particular marking on a packet indicates a PHB that has to be applied for forwarding of the packet. Currently, the EF PHB [5] and the AF PHB group [10] are specified. Though the architecture allows for the definition of additional PHBs, by setting the 6-bit Differentiated Services Code Point (DSCP), end-to-end guarantees require the support of a certain PHB in all pertaining domains thus making an end-to-end deployment with only a few PHBs more feasible.

This work was supported in part by the Path Allocation in Backbone networks (PAB) project funded by the German Research Network (DFN) and the Federal Ministry of Education and Research (BMBF).
The EF PHB is intended for building a service that offers low loss, low delay, and low delay jitter, namely a Premium service. The specification of the EF PHB was recently redefined to allow for a more exact and quantifiable definition. Besides the Premium service a so called Olympic service [10] is proposed by the IETF to be based on the AF PHB group by extending it by means of a class based over-provisioning. Three of the four currently defined classes of the AF PHB group are used for such an Olympic service. The service differentiation between the three classes Gold, Silver, and Bronze is proposed to be performed by the means of admission control, i.e. assigning only a light load to the Gold class, a medium load to the Silver class and a high load to the Bronze class.

The Differentiated Services architecture pushes complexity to the edges of the network: packets are marked to belong to an aggregate behavior either by applications or by edge routers. If edge routers mark packets, which is the more general solution, they may choose to do so on a per-flow basis or on any other criteria. In this scenario, of course, the question arises which packets will get marked. This is especially the case when the environment is dynamic, i.e. when varying flows should be able to use the available services. Here, a particular resource manager called a Bandwidth Broker (BB) comes into place. A BB is a middleware service which controls and facilitates the dynamic access to network services of a particular administrative domain. BBs are also viewed as the Policy Decision Point (PDP) of the controlled domain. The concept of a BB is typically associated with the Differentiated Services architecture. In this context, the task of a BB is to control the configuration of the edge routers of a single DS domain. By performing a careful admission control BBs are a fundamental building block for the provision of network services on top of DS aggregates. The GARA research prototype is a BB which addresses these issues. This paper presents a thorough evaluation of the achievable service classes provided by the admission control of a BB in a DS environment which does not rely on the assumption of a broad variance of aggregates. Instead, heterogenous aggregates consisting of elastic and non-elastic flows are explicitly addressed.

III. EXPERIMENTAL STUDIES

We report on experiments designed to examine a DS implementation based on commodity products. In the following subsections we first give the experimental configuration and then address the implementation and evaluation of the different traffic classes. We show problems observed when using the BE service and address these with similar measurements using the Olympic or the Premium service.

A. Experimental Configuration

Our experimental configuration comprises a laboratory testbed at the Research Centre Jülich, donated by Cisco Systems. The testbed allows controlled experimentation with basic DS mechanisms. Four Cisco Systems 7200 series routers were used for all experiments. These are either connected by OC3 ATM connections, by Fast Ethernet, or by Gigabit Ethernet connections. End-system computers are connected to routers by switched Fast Ethernet connections. Hence, the minimum MTU size of our testbed is that of the end-systems: 1500 B. To create a point of congestion, we configured an ATM Permanent Virtual Circuit (PVC) between an ingress and an interior router to 60 Mb/s.

We performed several experiments demonstrating the performance of high-end TCP applications [15] like a Guaranteed Rate (GR) file transfer with deadline and typical UDP applications like video streaming [8] or videoconferencing. The following tools have been used for traffic generation:

- gen_send/gen_recv – BE UDP traffic generator. This traffic generator was applied to generate BE UDP traffic with a mean rate of 50 Mb/s and different burst characteristics. These UDP flows do not aim to model any specific application, but we assume that the applied burst characteristics reflect effects that occur in today’s and in the future Internet. TCP streams are initially bursty, UDP based real-time applications are emerging, which create bursts, for example by intra-coded frames in a video sequence. Further on burst sizes increase in the network, due to aggregation and multiplexing [4].

- rude/crude – Delay-sensitive UDP traffic generator. This traffic generator allows to measure the one-way delay and delay jitter. In our experiments we used real-time traffic patterns from script files, which we created from publicly available video traces [8]. We applied IP fragmentation for the transmission of frames that exceed the MTU, which we consider as being allowed here, since we configured the DS classes to prevent from dropping fragments. The sequence, which we applied for the experimental results shown in this paper, is a television news sequence produced by the ARD. The sequence is MPEG-4 encoded with a minimum frame size of 123 B, a maximum frame size of 17.055 KB, a mean rate of 0.722 Mb/s and a peak rate of 3.411 Mb/s. The Hurst parameter is about 0.5 and decays with an increasing aggregation level. Figure 1 illustrates a part of the traffic profile of the sequence.

- ttcp – TCP stream generator. We used the widely known TCP benchmark ttcp to generate TCP load. In the experiments reported on in this paper we selected an end-system which was not capable of generating a rate of more than 1.8 MB/s and if not stated otherwise we applied a socket buffer corresponding to a maximum window size of about 15 MTU.

B. Implementation and Evaluation of the Best Effort Service

Applying the plain BE service to the video test application used throughout this paper we generate the baseline for our evaluation. Our configuration allocates the remaining capacity of the ATM bottleneck link, which is not used by any other class, to the BE class. In the following experiments no other class than BE is used, resulting in an assignment of 60 Mb/s of the bottleneck ATM link to the BE class. The tx-ring-limit parameter on the ATM interface card that specifies the queue size, which is assigned to the applied ATM PVC, was set to 16 particles each of 512 B allowing to store up to four MTU on the ATM interface. This value is by far smaller than the default value, but it has to be applied to allow for an efficient QoS
implementation [7]. The BE layer 3 queue was configured to hold at most 256 packets. We consider this queue size, which is a trade off between delay and loss rate, as being feasible for BE TCP traffic, which is rather sensitive to packet drops than to queuing delay in a range of a few tens of milliseconds.

In figure 2 the delay measured when transmitting the news sequence in the BE class is shown. Congestion is generated by applying an UDP stream with two bursts, each of ten seconds duration. As can be seen from figure 2, the delay is bounded to about 42 ms, showing some minor effects on the measurements due to tail-drop in the router. The delay corresponds to an effective data rate on the ATM interface of about 48 Mb/s after subtracting the ATM induced overhead. While this delay is acceptable for streaming video applications, it can be critical for real-time video applications like video conferencing.

For a service differentiation in terms of delay, the Olympic service [10] proposed by the IETF is realised by admission control and a class based over-provisioning. We carried out experiments with the transmission of the news sequence in each of the Olympic classes, with the classes configured according to Table II. Within each of the Olympic classes a differentiation of the drop probability for differently marked excess traffic can be performed by applying Multiple Random Early Detection (M-RED). Nevertheless, we consider excess traffic in an over-provisioned class as harmful for the BE class. Therefore we mark the conforming traffic and drop excess traffic in the over-provisioned classes. The layer 3 queue size of each of the three Olympic classes was configured to 128 packets in the WFQ environment. Consequently, the ingress meter and marker are based on a token bucket with a confirmed information rate of 2.4 Mbit/s for all Olympic classes, which leads to the over-provisioning factors given in Table II. A confirmed burst size of 32 MTU is used at the ingress. This value is intentionally smaller than the queue size that is applied in the core, to avoid packet drops in the Olympic classes within the network, to avoid a high utilization of the queuing space, and thus to reduce queuing delays. Besides it has to be noted that the WFQ queue size is configured in packets, which can be smaller than the MTU, whereas the confirmed burst size that is used by the meter and marker is configured in bytes.

Figure 3 shows the measured delay for the news sequence in the Bronze Class and the impacts of congestion in the BE class on the Bronze class. Compared to the transmission of the sequence within the BE class, which is shown in Figure 2, the delay is reduced significantly. Furthermore, packet drops did not occur in the Bronze class. Thereby AF based services can be applied as GR service without packet loss for conforming traffic. The delay and delay jitter differentiation, which can be achieved in addition by the Olympic service, is shown in Figure 4 and 5 for the Silver and the Gold class respectively, compared to the Bronze class in Figure 3.

Additionally, we present experiments with TCP in the Bronze class and demonstrate how TCP can be configured in a GR environment to achieve the desired throughput. We show that, if the pertaining class is configured properly, packet drops do not occur, which prevents from halving the TCP congestion window. The data rate instead corresponds to the capacity allocated for the flow. To avoid effects on the RTT by upstream congestion, the acknowledgements are also transmitted in the Bronze class. The maximum window size is in our experiments controlled by setting the socket buffer size. The resulting RTT can be computed according to the bandwidth-delay product: $W = R \cdot RTT$, with $W$ denoting the maximum window size and $R$ denoting the configured GR capacity. The RTT adjusts to the available or configured capacity and to the configured maximum window size.

### Table II

<table>
<thead>
<tr>
<th>Class</th>
<th>Percent</th>
<th>Gross Capacity</th>
<th>Net Capacity</th>
<th>Over-Provision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bronze</td>
<td>5 %</td>
<td>3 Mbit/s</td>
<td>2.4 Mbit/s</td>
<td>≥ 1 x</td>
</tr>
<tr>
<td>Silver</td>
<td>10 %</td>
<td>6 Mbit/s</td>
<td>4.8 Mbit/s</td>
<td>≥ 2 x</td>
</tr>
<tr>
<td>Gold</td>
<td>15 %</td>
<td>9 Mbit/s</td>
<td>7.2 Mbit/s</td>
<td>≥ 3 x</td>
</tr>
</tbody>
</table>

**C. Implementation and Evaluation of an Olympic Service**

A Weighted Fair Queuing (WFQ) environment is used for the implementation of the Olympic service based on three AF PHB classes. Within these classes GARA is capable of managing the allocated resources and the relative load in order to allow...
For these experiments we configured the Bronze class to 25% of the bottleneck link capacity, corresponding to a net data rate of about 1.6 MB/s. Figure 7 shows the RTT for a configured socket buffer of 15 MTU. Congestion in the BE class starts after 10 s and leads to an increase in the RTT, which corresponds to the queuing delay added by queuing the data of a complete TCP window. Figure 8 shows the corresponding throughput. At the beginning the application limits the data rate to about 1.8 MB/s and after the BE downstream congestion started, the limitation is given by the configured capacity for the Bronze class at about 1.6 MB/s and from the TCP point of view leads to a limitation of the sending rate by the offered window. The same effect on the throughput can be observed, if the maximum window is increased by configuring a socket buffer of 32 MTU. Figure 9 and figure 10 show the resulting RTT and throughput for this configuration. Again the RTT by increased queuing delay is adjusted to the available capacity and the window size, being about twice as high as in the previous experiment.

From these TCP experiments it can be seen that during periods of BE congestion WFQ acts as an aggregate traffic shaper with a rate corresponding to the configured WFQ weight. The achieved TCP throughput is independent of the TCP window size, as shown in Figure 8 and 10.

D. Implementation and Evaluation of a Premium Service

In a first experiment the Premium service was implemented based on EF using Priority Queuing (PQ). The ingress router was configured to apply a meter and marker with a confirmed information rate of 4.8 Mb/s and a burst size of 32 MTU. Excess traffic is dropped. The parameters that were applied at the ingress router were reflected by the core configuration. The PQ scheduler was bound to 10% of the bottleneck link capacity, corresponding to about 4.8 Mb/s. Bursts of up to 48 KB are permitted in the core. Figure 6 shows the results of a transmission of the news sequence. A reduction of the transmission delay and delay jitter especially for big video frames, which lead to packet bursts, becomes obvious for PQ compared to the WFQ settings in Table II. Here the tx-ring-limit parameter, which is used to configure the outgoing non-preemptive interface queuing capacity, is of major importance [7].

The following series of experiments were applied to an implementation of the EF PHB with the goal to analyze the behavior of a heterogeneous aggregate. It uses WFQ to emulate strict PQ by provisioning 99% of the available capacity to the EF aggregate. This is to ensure that the queue of the EF aggregate caused by possible bursts is minimized to reduce the queuing delay. Note that the BB prototype GARA performs a careful admission control and is thus preventing the starvation of the BE traffic. The particular challenge is caused by applying elastic and non-elastic traffic in a single EF aggregate. The setup consisted of following three flows, which passed an ATM bottleneck link of 100 Mb/s capacity:

- The first flow entering the testbed was a delay-sensitive Premium UDP flow. It ran from the beginning to the end of the measurement. GARA acted as a BB to associate the flow to the EF PHB. The related UDP traffic generator was configured to achieve a rate of 40 Mbps by constantly
submitting 1 KB packets every 0.2 milliseconds. The receiver continuously reported the delay calculated from the time-stamps in the packets.

- The second flow in the experiment was a GR TCP flow which was roughly active in the time interval between 20 and 60 seconds. Emulating a distributed supercomputing application, we created a bursty TCP stream which was injecting data in chunks of 256 KB into the network, using the EF PHB. Every 8th message contained two chunks, i.e. 512 KB. The average rate of the flow was 16 Mb/s. Using GARA, we claimed a slightly higher guaranteed bandwidth reservation, allowing bursts of up to one full chunk.

- The third flow started during the experiment was a BE UDP flow which was roughly active in the time interval between 40 and 80 seconds. Our main intention was to create a heavy congestion by submitting 750 byte packets at a frequency of 10000 Hz, to achieve a rate of 60 Mb/s. To demonstrate the impact of a single Premium flow under congestion, the competing UDP flow was still active after the TCP flow ends. The BE flow thus consumed a significant amount of the available capacity.

Figure 11 shows that the selected single-aggregate implementation is not appropriate for providing delay-sensitive services when bursty TCP flows use the same aggregate in parallel. If a burst introduced by the TCP flow exceeds the available output link capacity, packets get queued on the IP-layer queue. Because packets of the Premium UDP flow are also queued, the delay variation increases significantly. We can easily revalidate the result illustrated in Figure 11 by some simple calculations. The Premium service is used by a UDP application which is transmitting data at a rate of 40 Mb/s. This application shares the aggregate with a TCP flow which is injecting bursts of 256 KB at the link speed of its Fast Ethernet interface, i.e. at a rate of 100 Mb/s. Hence, the 256 KB of data enter the EF aggregate within 20 milliseconds. The total amount of data entering the EF aggregate in this time interval is thus 356 KB. Assuming an ATM overhead of 20 %, the EF aggregate is served at a rate of 80 Mb/s. We thus know that 200 KB of EF data leave the router during this 20 ms interval. Consequently, at the end of the interval there exists an EF queue of 156 KB that leads to an upper delay boundary of 15 ms.

In order to inject a traffic profile which is conforming to the Service Level Agreement with the peered downstream domain, the egress router of a DS domain might be enforced to shape out the traffic of a whole aggregate. When this is applied to the scenario illustrated above, the impact caused by the bursts of the TCP stream might be amplified by the queuing introduced by traffic shaping. Figure 12 illustrates the impact of traffic shaping when it is performed for an aggregate. Traffic shaping can be viewed as an additional constraint which limits the EF capacity of the output link by shaping the rate to the given traffic profile. Hence, EF packets get queued whenever TCP bursts cause the shaper to become active. In the illustrated scenario, the shaper became active whenever the TCP application produced a burst of two data chunks every 2 s.

As traffic shaping over an aggregate has a negative impact on the delay variation of a Premium flow, the EF implementation proposed here uses a flow-based service differentiation on the
output interface of the egress router. In detail, GARA applied an additional router internal packet marking mechanism, called “qos-group” which facilitates an efficient packet classification. The router configuration propagated by GARA extended the basic DSCP marking by also assigning the “qos-group” for the related flow. This additional classification was then used to update the configuration of the output interface of the ingress router to shape out the related GR flow. This demonstrates that real application can actually use DS, especially if access to services is automated by a BB such as GARA. The pragmatic approach of limiting the assumptions made about the underlying PHBs addresses potential deployment constraints and facilitates the negotiation of traffic trunk encoding between peered domains.

Our future work will focus on larger scenarios with several possible bottleneck links. These require complex TE mechanisms and an advanced resource management, which we aim at addressing with GARA and MPLS.

IV. CONCLUSIONS AND FUTURE WORK

We have presented a quantitative evaluation of a DS implementation providing a Premium service and an Olympic service automatically configured by GARA. Our evaluation addressed the QoS demand of heterogeneous types of flows sharing a single PHB. The experiments presented used commodity hardware. This demonstrates that real application can actually use DS, especially if access to services is automated by a BB such as GARA. The pragmatic approach of limiting the assumptions made about the underlying PHBs addresses potential deployment constraints and facilitates the negotiation of traffic trunk encoding between peered domains.

Our future work will focus on larger scenarios with several possible bottleneck links. These require complex TE mechanisms and an advanced resource management, which we aim at addressing with GARA and MPLS.

REFERENCES