Performance of TCP/IP with MEDF Scheduling

Ruediger Martin, Michael Menth, and Vu Phan-Gia

Department of Distributed Systems, Institute of Computer Science
University of Würzburg, Am Hubland, D-97074 Würzburg, Germany
{martin|menth|phan}@informatik.uni-wuerzburg.de

Abstract To achieve Quality of Service (QoS) in Next Generation Networks (NGNs), the Differentiated Services architecture implements appropriate Per Hop Behavior (PHB) for service differentiation. Common recommendations to enforce appropriate PHB include Weighted Round Robin (WRR), Deficit Round-Robin (DRR) and similar algorithms. They assign a fixed bandwidth share to Transport Service Classes (TSCs) of different priority. This is a viable approach if the ratio of high priority traffic TSC_{high} over low priority traffic TSC_{low} is known in advance. If TSC_{high} holds more and TSC_{low} less users than expected, the QoS for TSC_{high} can be worse than for TSC_{low}. As shown in preceding work, the Modified Earliest Deadline First (MEDF) algorithm heals this problem on the packet level. Therefore, we investigate its impact in congested TCP/IP networks by simulations and show its attractiveness as a powerful service differentiation mechanism.

1 Introduction

Current research for multi-service Next Generation Networks (NGNs) focuses amongst others on the provision of Quality of Service (QoS) for different service classes. The Differentiated Services architecture [1], [2] achieves QoS by implementing appropriate Per Hop Behavior (PHB) for different Transport Service Classes (TSCs). Flows of different TSCs compete for the resources buffer space and forwarding speed in the routers. Mechanisms that assign those resources divide buffer space among different TSCs (buffer management) and control the order in which packets are dequeued and forwarded (scheduling). Therefore, those mechanisms can be characterized along two dimensions: space and time.

Common examples and recommendations [3] [4] to enforce appropriate PHB are algorithms like Weighted Round Robin (WRR), Class Based Queueing (CBQ) [5], and Deficit Round-Robin (DRR) [6]. The common goal is to assign a fair share of network resources to different TSCs. The share is set in advance and fixed independently of the actual traffic mix. This behavior is desirable in many situations. In a network where a ratio q of high priority TSC (TSC_{high}) traffic over low priority TSC (TSC_{low}) traffic is expected, e.g. due to network admission control, the algorithms can be used to assign this share. The low priority TSC (TSC_{low}) uses the remaining bandwidth where a fraction of $1-q$ is guaranteed. However, if resources are scarce and buffers always contain packets of both classes, these algorithms enforce the share q regardless of the current traffic mix. Particularly, if the TSC_{high} traffic exceeds the limit set by the control parameters, it suffers from QoS degradation.

The authors of [7] introduced the priority algorithm Modified Earliest Deadline First (MEDF). They showed that MEDF prefers TSC_{high} over TSC_{low} on the packet level equally regardless of the traffic mix ratios. This is a clear advantage of MEDF compared to the previously mentioned algorithms that assign a fixed share for the whole TSC_{high} aggregate.

In this paper we focus on the impact of MEDF in TCP/IP networks. For saturated TCP sources conventional algorithms are problematic because of the fixed share of bandwidth assigned to each traffic class regardless of the current number of flows. We bring the MEDF algorithm into play to achieve a relative traffic-mix-independent per-flow-prioritization among TSCs. But still, this behavior is easily configurable by per-class relative delay factors.

The algorithms that work on the IP packet level impact the performance of adaptive TCP flows by packet loss and delay (round trip time). Packet loss is influenced by space priority mechanisms, delay by time priority mechanisms. In this work we combine MEDF with space priority mechanisms like Full Buffer Sharing (FBS) and Buffer Sharing with Space Priority (BSSP) and contrast it to time priority mechanisms like First In First Out (FIFO) and Static Priority (SP).

This work is structured as follows. In Section 2 we present the algorithms under study in detail. Section 3 discusses the simulation environment, the respective parameters used for our performance evaluation study, and presents the results obtained from our simulations. Sections 4 and 5 finally conclude this work with a short summary and outlook on future research.

2 Space and Time Priority Mechanisms

Network congestion arises where different flows compete for resources at routers in the network. To avoid this problem at least for a certain subset of high priority flows, flows of higher priority should receive preferential service as opposed to low priority flows. Basically, if packet arrivals exceed the router forwarding speed temporarily or permanently, congestion arises and buffers fill up. This leads to longer network delays and high packet loss rates, to degraded Quality of Service. Buffer sizes and forwarding speed are fixed parameters for given networks. To assign these scarce resources, we can limit the space available to the respective flows (buffer management) or we can dequeue the packets depending on their priority (scheduling). Thus, mechanisms to achieve service differentiation can be divided along two dimensions: space and time. Combinations of both are also possible.

2.1 Space Priority Mechanisms

We use two kinds of space priority mechanisms for our performance evaluation: Full Buffer Sharing and Shared Buffers with Space Priority. In [10] we compare a third space priority mechanism Random Early Detection gateways [11]. RED was originally designed to detect incipient congestion by measuring the average queue length. Several improvements have been suggested for instance in [12] and [13] to achieve fairness in the presence of non-adaptive connections and to introduce TSC priorities. We omit this section for lack of space here and refer to our technical report [10].

In the following sections, we denote the router buffer by B and packets by P. The function $S(B)$ refers to the maximum buffer size and $F(B)$ to the current fill level of the buffer. The function $enqueueTail(P, B)$ enqueues the packet P into the buffer B. The function $drop(P)$ drops the packet P if the algorithms cannot accept the packet.

Full Buffer Sharing (FBS). The FBS strategy allows all flows to share the same buffer irrespective of their priority. If not mentioned differently, we use this mechanism as default in our simulations.

Buffer Sharing with Space Priority (BSSP). The BSSP queueing strategy (cf. Alg. 1) is threshold based and allows packets to occupy buffer space available for their TSC and for all TSCs of lower priority. Let TSC_i, $i \in \{0, \ldots, n-1\}$ be TSCs of different priority, 0 being the highest priority. TSC_i can at most demand space BS_i^{max} in the buffer, where $BS_i^{max} \geq BS_{i+1}^{max}$ and BS_0^{max} is set to the actual buffer size. The concept is illustrated in Fig. 1 for three TSCs and fully described in Alg. 1 with $F(B, TSC_i)$ denoting the space in the buffer B that is currently filled by TSC_i. There is a guaranteed amount of buffer space for the highest priority class only, lower priority classes possibly find their share taken by classes of higher priority. This concept resembles the Russian dolls bandwidth constraints model (RDM) suggested by the IETF traffic engineering working group (TEWG) in [14].

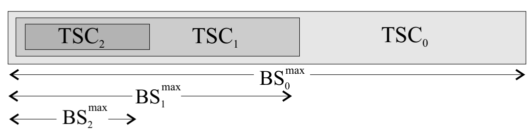

Figure 1. Buffer Sharing with Space Priority for $i = 3$ TSCs

Require: Packet P, Buffer B, max TSC Buffer Size BS_i^{max} for $i = 0 \ldots n-1$
{ max Buffer Size $S(B) = BS_0^{max}$ }
i = TSC(P)
if $\sum_{j=i}^{j=(n-1)} F(P, TSC_j) \leq BS_i^{max}$ **then**
$enqueTail(P, B)$
else {space limit exceeded for TSC i}
$drop(P)$
end if

Algorithm 1: Buffer Sharing with Space Priority ENQUEUE

2.2 Time Priority Mechanisms

Once packets arrive at the queue and the space priority mechanism assigns available buffer space, i.e., it decides whether the packet is accepted or dropped, the time priority mechanism decides which packet to dequeue next. This decision on the packet level influences the delay and therefore the TCP sending rate via RTT. We contrast two time priority mechanisms to Modified Earliest Deadline First (MEDF).

First in First Out (FIFO). FIFO leaves the prioritization to the enqueueing option and is used as the performance baseline to compare with. Packets proceed in the order they arrive and are accepted by the space priority mechanism.

Static Priority (SP). The Static Priority concept chooses TSC_{high} packets in FIFO order as long as packets of that class are in the buffer. TSC_{low} packets wait in the router queue until low priority packets only are available. Then they are also dequeued in a FIFO manner until new TSC_{high} packets arrive.

Modified Earliest Deadline First (MEDF). In the context of the UMTS Terrestrial Radio Access Network, the authors of [7] introduced a modified version of the Earliest Deadline First (EDF) algorithm called Modified Earliest Deadline First (MEDF). It supports n only different TSCs, but in contrast to EDF it is easier to implement. Packets are stored in n TSC specific queues in FIFO manner. They are stamped with a modified deadline that is their arrival time plus an offset $M_i, 0 \leq i < n$, which is characteristic for each TSC. The MEDF scheduler selects the packet for transmission that has the earliest due date among the packets in the front positions of all queues. For only two TSCs, this is the choice between two packets and sorting according to ascending deadlines is not required. The difference $|M_i - M_j|$ between two TSCs i and j is a relative delay advantage that influences the behavior of the scheduler. We are interested in the performance of this scheduling algorithm in the presence of adaptive traffic, here TCP.

For our simulations we use two TSCs whose queues are implemented as shared buffers such that the space priority mechanisms are applicable. With two TSCs we set the MEDF parameters to $M_{high} = 0$ and $M_{low} = x, x \in \{0s, 0.1s, 0.5s, 1.0s, 1.5s\}$. Thus, TSC_{high} obtains no additional delay. The deadline for TSC_{low} packets is increased by the M_{low} parameter.

3 MEDF Performance Evaluation

In this section we describe the general goals and approach of our performance evaluation study and present the results. We used the network simulator (NS) version 2 [15] to run the experiments deploying the RENO TCP implementation [16]. Standard simulation methods as replicate-delete were applied to obtain statistically reliable results of the non-ergodic random processes. In the following sections we only give average values as the simulated time was chosen to yield very narrow confidence intervals. Our goal is the measurement of the prioritization of TSC_{high} traffic. For that purpose, we define the normalized bandwidth ratio. Let n_{high} be the number of TSC_{high} flows, n_{low} the number of TSC_{low} flows. The functions $B(TSC_{high})$ and $B(TSC_{low})$ denote the bandwidth used by all TSC_{high} and TSC_{low} flows, respectively. The normalized bandwidth ratio $\overline{B}_{ratio}(TSC_{high}, TSC_{low})$ is the amount of bandwidth used by TSC_{high} per flow divided by the amount of bandwidth used by TSC_{low} per flow:

$$\overline{B}_{ratio}(TSC_{high}, TSC_{low}) = \frac{\frac{B(TSC_{high})}{n_{high}}}{\frac{B(TSC_{low})}{n_{low}}}$$

A mechanism with traffic-mix-independent per-flow-prioritization among TSCs exhibits the same normalized bandwidth ratio regardless of the traffic mix. The number of saturated TCP sources is the same for both TSCs in the following if not mentioned otherwise.

3.1 MEDF Characteristics

To isolate the general behavior more easily and to eliminate unpredictable side effects, we start with single link simulations and extend it to multiple links.

MEDF Single Link Scenario

Simulation environment. We use the classical dumbell topology for our single link simulation environment. A number of TSC_{high} TCP traffic sources and a number of TSC_{low} TCP traffic sources connect to Router A. Router A uses a space and a time priority mechanism described above and sends the packets over a single link to router B. Router B has sufficient capacity to serve the link and its single task is to distribute the arriving packets to the corresponding destinations.

We choose the number of simultaneous active TCP connections n as $n_{min} \cdot 2^i, i \in \{0, \ldots, 8\}$, n_{min} being the minimum number of TCP connections to get a theoretical load of 100% on the link. Otherwise there is no overload, space and time priorities do not have effect, and the flow control is not active. Here $n_{min} = 2$. The packet size $S(P)$ is a common standard value of 500 Bytes including headers. Regarding the link parameters, with the link bandwidth being $C_l = 1.28 Mbit/s$, we set the link propagation delay D_{prop} to 46.875 ms so that the theoretical round trip time RTT sums up to $RTT = 2 \cdot (n_{links} \cdot D_{prop} + (n_{links}+1) \cdot D_{TX}) = 2 \cdot (1 \cdot 46.875ms + 2 \cdot 3.125ms) = 100ms$, where $D_{TX} = \frac{S(P)}{C_l}$ is the transmission delay to send a packet and n_{links} the number of links between routers A and B.

The default value for the buffer size S_{Buffer} is 160 packets so that a router is able to store packets for 0.5 seconds transmission. We use the parameters mentioned here as default parameters and write down the respective values in the following text only if they are set differently. Other parameters like algorithm specific settings are subject to the analysis and we indicate their values appropriately.

Simulation. Figure 2 shows the normalized bandwidth ratio $\overline{B}_{ratio}(TSC_{high}, TSC_{low})$ for traffic mixes $n_{high} : n_{low}$ of $1:3$,$1:1$, and $3:1$ with MEDF parameter $M_{low} = 0.5s$, i.e., one buffer size. The link bandwidth is the x-axis parameter. The value $n_{min} = 2$ is omitted here and in the following figures as there is virtually no priority for the minimum number of users. The link capacity is fully shared between the single user of each class, thus, they both reach the maximum rate. This behavior – as expected – is sound for lack of competition on the link.

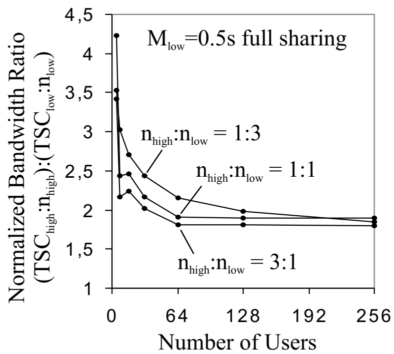

Figure 2. MEDF prioritization independent of the traffic mix

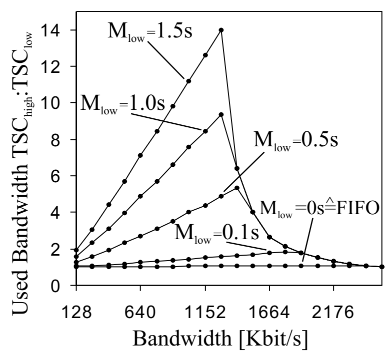

Figure 3. MEDF prioritization for two TSCs

The figure shows the traffic-mix-independent per-flow-prioritization property of MEDF. The small differences for low congestion are due to the slight influence of the buffer space. Few high priority flows can occupy relatively more buffer space per flow in contrast to many high priority flows under low congestion. However, the difference is negligibly small and the normalized bandwidth ratio converges very quickly. Opposed to that, conventional algorithms like WRR are insensitive to the traffic mix and therefore the normalized bandwidth ratio would decrease severely with the ratio of TSC_{high} traffic over TSC_{low}. We further emphasize that this property is achieved by a single parameter per class and originates from the relative delay advantage controlled by MEDF.

The traffic-mix-independent per-flow-prioritization property of MEDF was already shown in [7] on the packet level. Due to this property on the TCP flow level as well, we use the same number of saturated TCP sources for both TCSs in the following.

For the minimum number of users $n_{min} = 2$, there is virtually no prioritization for lack of competition. Prioritization of TSC_{high} traffic reaches its maximum at $n = 4$ users (2 users per TSC) and degrades with a rising number of users. As we cannot simulate any value between two and four users – one and two users per TSC – we vary the bandwidth while keeping the number of users fixed at a value of 4 to derive the basal characteristics of the algorithm by having a more continuous range in Fig. 3. This demonstrates the behavior at various levels of slight congestion in real networks.

At a bandwidth of 1.280 Mbit/s this experiment corresponds to a simulation with default values and 4 users, at a bandwidth of 2.560 Mbit/s it is equivalent to 2 users. Higher offset values M_{low} lead directly to a higher prioritization of TSC_{high} packets. The throughput ratio rises with the bandwidth which is inversely proportional to the number of users. Low bandwidth (same holds for many users) limits the rate that connections for TSC_{high} can achieve dramatically. Besides, the actually measured round trip time increases and shortens the maximum obtainable rate. Thus, TSC_{low} connections are able to grasp a higher relative share of the bandwidth. The bandwidth ratio rises until it reaches a maximum. Here, slowly sufficient capacity becomes available for both TSCs and low priority packets can use more of the additional bandwidth. At 2.560 Mbit/s there is virtually no competition for bandwidth anymore.

Another important aspect that can be seen here to understand MEDF is the interaction of M_{low} and the round trip time. The round trip time rises both with increasing traffic on the link and with decreasing bandwidth. Low bandwidth results in a longer transmission delay. The delay advantage becomes smaller relative to the round trip time and the prioritization decreases.

The MEDF parameter M_{low} can be used to adjust the priority ratio for the anticipated level of competition for network resources. If sufficient resources are available, the MEDF algorithm does not influence normal network operation. For very scarce resources – here large numbers of users and low bandwidth, respectively – the network is under heavy overload and anticipatory action like admission control to block some of the connections must be taken to prevent such situations. Otherwise, only a very small portion of the overall bandwidth remains for each TSC_{high} flow anyway — no matter whether they receive preferential service or not. For low and medium overload, MEDF shows a very clear and easy adjustable behavior.

MEDF Multi Link We now extend our single link experiment to multiple links to assess the influence of MEDF on TSC priority if applied multiple times.

Simulation environment. Figure 4 shows the simulation topology for the multi link experiment in the case of two links. If we simply add additional links and routers, the first router receives the packets from the TCP sources in an unordered way and applies the priority algorithm. Thus, the packets arrive at the router serving the next link one by one and the priority algorithm has no additional effect. To overcome this problem, we introduce cross traffic. Additional TCP sources connect to the interior routers and generate traffic that crosses the way of the measured traffic.

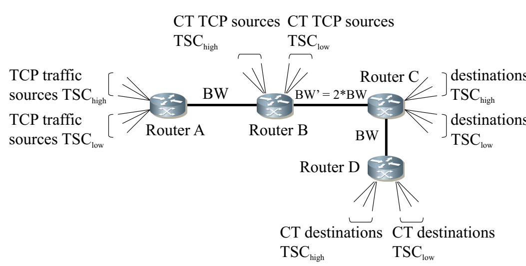

Figure 4. Multi link simulation topology

It is important to send the cross traffic over the same number of links to account for comparable round trip times for the measured traffic and the cross traffic. Furthermore, the round trip time for both the single link and the multi link experiment should be the same. Otherwise, significant parameters that depend on the round trip time such as the maximum rate that can be achieved by a TCP connection are different and the experiments are not comparable. Therefore, we calculate the new link propagation delay $D_{prop} = \frac{46.875-(n_{link}-1)\cdot D_{TX}}{n_{link}} ms$.

The TCP connections need the same bandwidth per flow on all links. If the bandwidth differs from link to link, the link with the lowest capacity becomes the bottleneck and dominates the observable effects. However, doubling the bandwidth of the links with cross traffic solves this problem.

Simulation. Figure 5 shows the effect of MEDF over multiple links. We used the standard parameters with $M_{low} = 1s$, i.e., twice the buffer size, and the default Full Buffer Sharing mechanism as buffer management.

In general, the degree of prioritization of TSC_{high} increases with the number of links on the path, hence, with the number of applications of MEDF scheduling instances. However, when the competition for network resources is low, the increase in priority is much more obvious. The reason behind this is similar to the situation for the single link experiment. The bandwidth theoretically available to a single connection is higher, hence, the actually measured round trip time is lower. Therefore, few TSC_{high} connections achieve higher rates in contrast to the situation when the network is highly overloaded. Rising competition for network resources makes the conditions for TSC_{high} more disadvantageous. TSC_{low} now obtains a larger share of the bandwidth. The priority does not increase linearly if additional links are added. The overall bandwidth ratio can be controlled by setting the MEDF parameter appropriately.

MEDF and Space Priority We now consider the MEDF characteristics with the usage of space priority mechanisms. Figure 6 shows the influence of the buffer sharing option. FIFO with FBS leads to an even division of available bandwidth between both TSCs as no packet preferences exist. FIFO with BSSP spreads the bandwidth equally as long as there is enough buffer space available (n $\leq$ 2). Then it reaches its maximum when router buffers fill completely and slightly flattens under heavy traffic load.

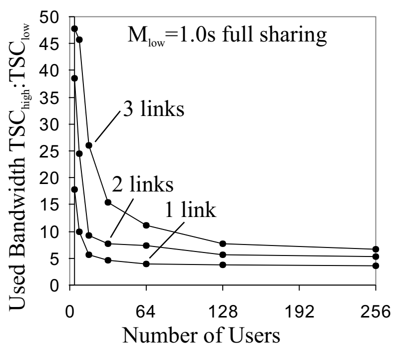

Figure 5. MEDF prioritization in a multi link topology

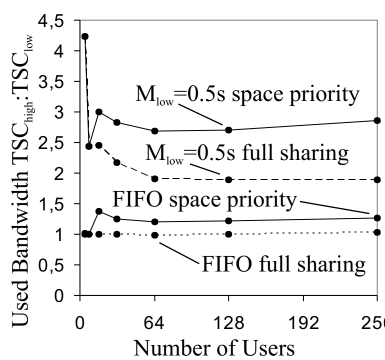

Figure 6. MEDF and the impact of space priority

MEDF with parameter $M_{low} = 0.5s$ and FBS clearly outperforms both FIFO experiments and exhibits the behavior characterized in the preceding sections. If we add

BSSP, we observe a superposition of the MEDF curve and the curve for FIFO with BSSP. For few users we clearly identify the typical MEDF characteristics, for more users the router buffers fill completely and the space priority comes into play. Thus, space priority prohibits the typical decrease of the bandwidth ratio.

3.2 MEDF in Comparison to other Priority Mechanisms

We used FIFO as the comparison baseline in the previous experiments. FIFO does not prioritize the traffic in time and therefore is one extreme of the spectrum of time priority mechanisms. Another extreme is Static Priority (SP).

Static Priority (SP) Under network congestion, the time priority mechanism Static Priority leads to starvation of TSC_{low} regardless of the buffer management in use. There are always TSC_{high} packets waiting in the router queues. SP dequeues those packets and even though the TSC_{low} packets occupy most of the buffer space, their chance to leave the buffer is very low and, thus, the TCP timers for those connections expire. Accordingly, the TCP source tries to re-establish the connection but will suffer from starvation again. As a consequence, SP is completely inadequate for severely congested networks. In contrast to MEDF it does not consider a maximum delay for low priority traffic to prevent this effect.

For a comparison to RED, a pure space priority algorithm, we refer to our technical report [10].

4 Conclusion

In this work we examined the impact of the pure time priority (packet scheduling) mechanism Modified Earliest Deadline First (MEDF) in congested TCP/IP networks. Conventional algorithms like Weighted Round Robin (WRR), Deficit Round Robin (DRR) or Class Based Queuing (CBQ) assign fixed bandwidth shares among Transport Service Classes (TSCs) of different priorities. This is problematic with a varying number of users per TSC and saturated TCP sources. If the TSC of high priority (TSC_{high}) holds more users than expected and the TSC of low priority (TSC_{low}) holds fewer users, then the Quality of Service (QoS) for TSC_{high} can be worse then for TSC_{low}. MEDF, however, achieves a relative traffic-mix-independent per-flow-prioritization among all TSCs.

In contrast to MEDF, Static Priority (SP) leads to starvation of low priority traffic while First In First Out (FIFO) effects no prioritization at all. MEDF achieves the desired priority ratio of the high priority TSC over the low priority TSC in realistic overload situations by its adjustable parameter M_{low} which reflects a relative delay advantage.

Full Buffer Sharing (FBS) was the default buffer management scheme in our experiments. To estimate the influence of the buffer management algorithms, we combined MEDF with Buffer Sharing with Space Priority (BSSP). The results showed an increased prioritization of TSC_{high}.

In conclusion, MEDF has powerful service differentiation capabilities and our performance study revealed that it is an attractive mechanism to achieve service differentiation for TCP flows in congested networks. MEDF is especially interesting since it does not require per-class bandwidth configuration which might be problematic in the presence of unknown traffic mix.

5 Outlook

The practical adaptation of the relative delay parameter M_{low}, especially its dependence on the propagation delay, is an interesting field of further research.

Acknowledgment

The authors would like to thank Prof. Tran-Gia for the stimulating environment which was a prerequisite for that work.

References

1. Blake, S., Black, D.L., Carlson, M.A., Davies, E., Wang, Z., Weiss, W.: RFC2475: An Architecture for Differentiated Services. ftp://ftp.rfc-editor.org/in-notes/rfc2475.txt (1998)
2. Grossman, D.: RFC3260: New Terminology and Clarifications for Diffser. ftp://ftp.rfc-editor.org/in-notes/rfc3260.txt (2002)
3. Davie, B., et al.: An Expedited Forwarding PHB (Per-Hop Behavior). ftp://ftp.rfc-editor.org/in-notes/rfc3246.txt (2002)
4. Charny, A., et al.: Supplemental Information for the New Definition of the EF PHB (Expedited Forwarding Per-Hop Behavior). ftp://ftp.rfc-editor.org/in-notes/rfc3247.txt (2002)
5. Floyd, S., Jacobson, V.: Link-sharing and Resource Management Models for Packet Networks. IEEE/ACM Transactions on Networking **3** (1995)
6. Shreedhar, M., Varghese, G.: Efficient Fair Queueing Using Deficit Round-Robin. IEEE/ACM Transactions on Networking **4** (1996)
7. Menth, M., Schmid, M., Heiß, H., Reim, T.: MEDF - A Simple Scheduling Algorithm for Two Real-Time Transport Service Classes with Application in the UTRAN. IEEE INFOCOM 2003 (2003)
8. Ramabhadran, S., Pasquale, J.: Stratified round Robin: a low complexity packet scheduler with bandwidth fairness and bounded delay. Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications Karlsruhe, Germany (2003) 239–250
9. Stiliadis, D., Varma, A.: Efficient fair queueing algorithms for packet-switched networks. IEEE/ACM Transactions on Networking (TON) **6** (1998) 175–185
10. Martin, R., Menth, M.: Performance of TCP/IP with MEDF Scheduling. Technical Report 323, University of Würzburg (2004)
11. Floyd, S., Jacobson, V.: Random Early Detection Gateways for Congestion Avoidance. IEEE/ACM Transactions on Networking **1** (1993) 397–413
12. Anjum, F., Tassiulas, L.: Balanced-RED: An Algorithm to Achieve Fairness in the Internet. IEEE INFOCOM 1999 (1999)
13. Bodin, U., Schelén, O., Pink, S.: Load-tolerant Differentiation with Active Queue Management. SIGCOMM Computer Communication Review (2000)
14. Le Faucheur, F.: Russian Dolls Bandwidth Constraints Model for Diff-Serv-aware MPLS Traffic Engineering. Internet Draft TEWG (2003)
15. Fall, K., Varadhan, K.: The ns Manual. http://www.isi.edu/nsnam/ns/doc/ns_doc.pdf (2003)
16. Stevens, W.R.: TCP/IP Illustrated, Volume 1. Addison-Wesley Longman (1994)