Best-Effort Low-Delay Service

Jörg Diederich
Institute of Operating Systems and Computer Networks
Technical University of Braunschweig
Germany
Email: dieder@ibr.cs.tu-bs.de

Mark Doll, Martina Zitterbart
Institute of Telematics
University of Karlsruhe (T.H.)
Germany
Email: {doll|zit}@tm.uka.de

Abstract

The Differentiated Services (DiffServ) approach is intended to provide Quality of Service (QoS) in IP-based networks. This is a very important issue not only in wire-line fixed networks but also in IP-based wireless mobile networks. Currently, services within the DiffServ approach including Premium Service, Assured Service and Best-Effort Service are intended for fixed networks. We have proposed the Mobile Differentiated Services QoS model [5] to enhance the above services so they are better suited for wireless mobile networks. One part of this QoS model is the so-called Best-Effort Low-Delay (BELD) Service which has the same low-delay and low-jitter characteristic as Premium Service, but a certain probability for packet loss. This service is especially suited for delay-sensitive and loss-tolerant applications and can take favor of unused Premium Service resources to increase the network utilization. This paper presents details on the BELD Service including initial simulation results showing the feasibility of BELD service.

1. Introduction

Current third generation mobile networks such as UMTS networks are based on IP to provide data communication services. However, means to provide QoS are currently not included: Instead, a separate circuit-switched network is maintained for applications which require a minimum bandwidth to work properly (e.g., mobile telephony). In future releases of UMTS networks (Release 4/5), it is planned to join both, the packet-switched and the circuit-switched network, to a so-called *All-IP network* where even data from QoS-sensitive applications such as mobile telephony are forwarded over an IP-based network infrastructure. However, QoS mechanisms have to be included in such an All-IP network to ensure the QoS requirements of QoS-sensitive applications, for example, using the Differentiated Services (DiffServ) [12] approach. One problem is that the legacy service model of DiffServ does not take the special characteristics of wireless mobile networks into account, e.g., the occurance of handoffs or the special characteristics of wireless links. Therefore, we have developed the Mobile Differentiated Services QoS model (*MoDiQ*) [5]. *MoDiQ* especially deals with the problem of handoff resource shortage leading to an interruption of a communication session after a handoff. For this purpose, a simple handoff prioritization scheme is proposed which reserves some resources for handoffs exclusively to avoid handoff resource shortages. Such a prioritization scheme is based on the trade-off between the probability of a handoff resource shortage and the probability that a new session request is blocked. Decreasing the number of handoff resource shortages always means to increase the probability of a blocked new session at the same time. This way, the utilization of resources decreases because resources have to be reserved for handoffs. However, handoff resources are not always fully utilized because they cannot be reserved in general with a very high accuracy. This is because mobility patterns of typical mobile terminals (e.g., in cars) are not predictable to 100%.

BELD Service has two simultaneous objectives: It increases the utilization of Premium Service resources and provides a separate service for delay-sensitive, but loss-tolerant applications. For such applications, no tailored service is available in the current DiffServ service model. Instead, they must make use of Premium Service, the only low-delay service in the legacy DiffServ service model. However, Premium Service is expected to be the most expensive service since Premium Service traffic is treated with a very high priority within the network. Therefore, using this service may create unnecessary costs for applications which require a low delay in packet forwarding but can compensate packet losses up to a certain rate.

This paper is organized as follows. Section 2 introduces the DiffServ service model and the *MoDiQ* service model. Section 3 describes details on BELD Service and its components. In Section 4, the feasibility of BELD Service is shown by simulations. Finally, Section 5 summarizes this article.

2. Background

2.1. The Legacy Differentiated Service Model

As shown in Figure 1, the legacy DiffServ service model [15] consists of three services: Premium Service, Olympic Service (also called Assured Service), and Best-Effort Service. The Best-Effort Service is targeted at elastic applications, Olympic Service at delay-insensitive applications with minimal bandwidth demands, and Premium Service [15] at delay-sensitive real-time applications. Since this paper is focused on delay-sensitive applications, Premium Service is described in more detail in the following section.

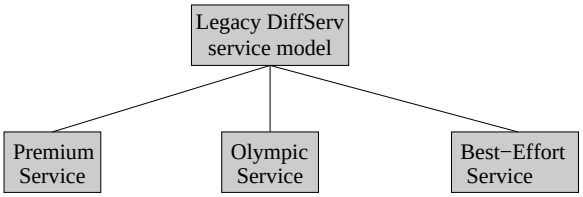


Figure 1. The legacy DiffServ service model

2.2. Premium Service

Premium Service [15] is characterized as a low-delay, low-jitter, low-loss end-to-end service. It is specified by the QoS parameter 'peak rate' which is expected to be available as soon as the service user wants to use it. Hence, Premium Service should be indistinguishable from a physical line with the same fixed data rate (the negotiated peak rate). The following requirements have to be met:

- A DiffServ node should forward Premium Service packets with a higher priority than packets from services without low-delay characteristics. This is the *high-priority forwarding requirement* of Premium Service which ensures that the Premium Service aggregate on an output interface receives its configured rate almost independent of the amount of other traffic.

- The number of low-delay packets in the domain, or within the network respectively, must be limited so that queues of low-delay packets do not occur. For this purpose, it is necessary to coordinate node configuration, which determines the capacity on a link reserved for Premium Service, with admission control and traffic conditioning, which control the amount of Premium Service traffic inserted into a domain. This constitutes the *traffic limitation requirement* of Premium Service. It ensures for all nodes in the domain that the total incoming Premium Service traffic destined to each single output interface is smaller than the configured rate of that output interface.

The first requirement is related to a single node and, therefore, formulated in a DiffServ-specific per-hop behavior (PHB). The second requirement constitutes the Premium Service 'rules' and is related to a DiffServ domain or a network as a whole [11, 15]. Both are discussed in the following paragraphs.

2.2.1 The Expedited Forwarding Per-Hop Behavior

The Expedited Forwarding PHB proposal [3] defines a forwarding behavior with a bound on the delay and the delay jitter that a single EF packet may experience. The high-priority forwarding requirement is supposed to be implemented using a simple priority scheduling discipline or any other prioritization scheduling discipline. Furthermore, Premium Service traffic should on average experience no or only small queues. Two main situations are considered where queuing of these packets is necessary even if the high-priority forwarding requirement is fulfilled.

- *Busy-interface effect*: A Premium Service packet arrives when the transmission of a large packet (up to the MTU) from a different service (e.g., Best-Effort Service) has just started. In this case, the Premium Service packet has to be queued because preemption of packets in transmission is not supported by most link layer technologies. Finally, the Premium Service packet can be sent immediately after the transmission of the Best-Effort Service packet.

- *Simultaneous fan-in effect*: Two or more Premium Service packets arrive simultaneously on different input interfaces having the same destination output interface. In this case, only one Premium Service packet can be sent immediately, the remaining packets have to be queued and are sent subsequently.

The EF PHB considers both effects separately and provides two delay bounds: The first one relates to queuing delays caused by effects on the output interface (for example, the busy-interface effect; further effects are described in the EF PHB definition [3] but are not relevant in this paper). The second one relates to effects with regard to the input interfaces (e.g., the simultaneous fan-in effect):

- *Aggregate Delay Bound*: Each EF packet of the EF aggregate on the outgoing interface is guaranteed not to leave the node later than the *error term* E_a after its ideal departure time (i.e., the departure time without

the delay caused by the busy-interface effect). This error term E_a is to be specified by the vendor of an EF-compliant node and characterizes the performance of the node. If the busy-interface effect is the only reason for delaying Premium Service packets, E_a is defined as $E_a = \frac{MTU}{C}$ with C being the line rate of the outgoing link. For example, E_a is $8\,ms$ for an MTU of 1500 bytes for Best-Effort Service packets and a line rate of $C = 1.5\,\text{Mbit}/s$.

- *Individual Delay Bound*: The delay of each EF packet of an EF flow, entering the EF-compliant node from a single input interface, is bound by the *error term* E_p. If this error term depends only on the number N of incoming interfaces, it is defined as $E_p = \frac{MTU \cdot (N-1)}{C}$ since only a single Premium Service packet can be sent immediately in case of the simultaneous fan-in effect. The remaining $N - 1$ packets experience a queuing delay. Continuing the above example, E_p is $16\,ms$ if the EF-compliant node has three incoming interfaces.

2.2.2 Rule: Traffic Limitation Requirement

While the above described PHBs ensure the high-priority forwarding requirement, the rules implement the traffic limitation requirement of Premium Service. Three QoS components, viz. node configuration, traffic conditioning, and admission control, are involved in limiting the Premium Service traffic to ensure a low-delay.

2.2.3 Node Configuration

Node configuration is necessary for Premium Service to appropriately configure the EF PHB on each node. Constraints on the configurable parameters (the EF rate and the EF buffers) are described below.

EF Rate It has been shown that the EF rate on a link must not exceed 50% of the link capacity to ensure a bound on the jitter [11]. This is because of the busy-interface effect caused by the presence of Best-Effort Service packets.

EF Buffers The two effects 'busy-interface' and 'simultaneous fan-in' determine the minimal buffer size necessary to accommodate a worst-case of Premium Service packet arrivals at a node, which is described as follows.

Figure 2 depicts a worst-case situation with regard to the maximum queuing delay where both, the busy-interface effect and the simultaneous fan-in effect occur. It shows the state of a single interior node with three incoming links and a single outgoing link. Each interior node implements a simple priority scheduling but only the high-priority queue is shown for simplicity reasons. The figure depict a series of three parts to elaborate the behavior over time with the uppermost part being the first in time and the bottommost part being the last.

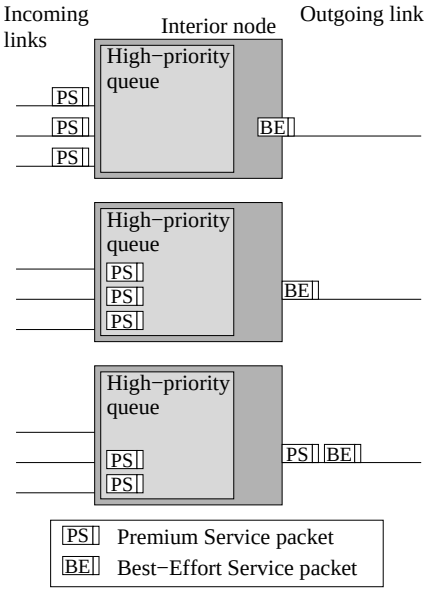


Figure 2. Simultaneous packet arrival

Depending on the internal processing speed of the DiffServ node, the size of the high-priority EF buffer on the interior node must, therefore, be at least equal to $MTU \cdot N$ where N is the number of incoming interfaces (the *in-degree of a router*).

It is ensured that the high-priority EF buffers have been emptied before the next burst of Premium Service packets may arrive if the incoming aggregates are perfectly shaped and if the above mentioned traffic limitation requirement of Premium Service is ensured.

The maximum delay a Premium Service packet experiences in a single DiffServ node is the sum of the maximum delays determined by the error terms E_a and E_p. In case both are determined only by the busy-interface effect and the simultaneous fan-in effect, the maximum per-packet delay calculates to $E_a + E_p = \frac{MTU}{C} + \frac{MTU \cdot (N-1)}{C} = \frac{MTU \cdot N}{C}$. It is $24\,ms$ for an MTU of 1500 bytes, a line rate of $1.5\,\text{Mbit}/s$ and three incoming interfaces. However, larger EF buffers may be required, even if the EF aggregates entering the domain are perfectly shaped. This is caused by the possible aggregation of bursts within a domain if a per-flow traffic shaping is not applied on each interior node.

2.2.4 Traffic Conditioning

Traffic conditioning, the second component to implement the rules for Premium Service, is needed for two purposes:

1. To limit the amount of traffic inserted into a domain (on the boundary nodes).

2. To limit the effect of burst aggregation in the domain (on boundary nodes and possibly on interior nodes).

The primary motivation of the first is to avoid an overload of EF resources in the domain-global sense so that no more Premium Service traffic enters the domain than the domain is able to carry. Therefore, Premium Service packets which exceed the negotiated rate, are policed (i.e., dropped).

The second purpose relates to a per-node scale, i.e., its task is to avoid an overload of buffers resulting from burst aggregation. In this case, traffic shaping is necessary on the boundary node to enforce compliance of the incoming EF aggregate with a maximal burstiness which may be negotiated between two domains. On interior nodes, traffic conditioning may not be required if the buffer sizes accommodate the burst aggregation effects. This may be true, for example, for small domains with few interior nodes only. However, traffic shaping on the output link of interior nodes may be reasonable in large domains [8] (but the tradeoff is a potentially higher delay because of the delay in shaping).

2.2.5 Admission Control

While traffic conditioning ensures that already admitted Premium Service data flows adhere to their negotiated rate, a peak-rate-based admission control ensures that the sum of the negotiated Premium Service resources is equal to or smaller than the capabilities of the domain. Therefore, it is necessary to reserve resources in the network for a particular session to ensure the low-delay characteristic.

2.3. The Mobile Differentiated Service Model

The *MoDiQ* service model [5] extends the legacy DiffServ service model from three to six service classes as depicted in Figure 3. In contrast to the legacy DiffServ service

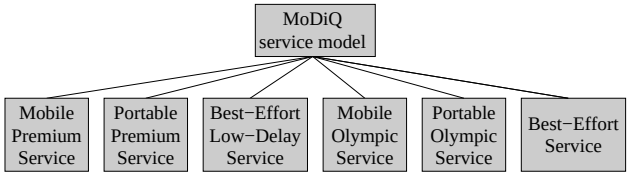


Figure 3. The MoDiQ service model

model, the *MoDiQ* service model proposal provides assurances on the handoff success probability for both legacy services, viz. Premium Service and Olympic Service. Furthermore, separate services without such an assurance are available for portable terminals to increase the efficiency of resource utilization.

Therefore, the legacy Premium Service class is divided into two parts: *Mobile Premium Service* provides low-delay packet delivery with support for assurances on the handoff success probability whereas the *Portable Premium Service* is a low-delay service with no such support. Analogously, the Olympic Service class is split into a *Mobile Olympic Service* and a *Portable Olympic Service*.

For delay-sensitive and loss-tolerant applications, there is currently no service available in the legacy DiffServ service model. Thus, the *MoDiQ* service model contains a third low-delay service called *Best-Effort Low-Delay (BELD) Service* which is described in detail in this paper.

3. Best-Effort Low-Delay Service

BELD Service is defined as having a low-delay, low-jitter characteristics similar to Premium Service. The major difference is that BELD Service packets have a higher probability of being dropped. It is built from a new PHB called *Expedited Forwarding with Dropping* (EFD) [6] which has been proposed to the IETF for discussion. This PHB can be used to construct BELD Service together with an appropriate traffic conditioning and an optional admission control component. These components are described in detail in the following sections.

3.1. Expedited Forwarding with Dropping PHB

The EFD PHB is an enhancement to the Expedited Forwarding (EF) PHB [3]. It achieves a low-delay low-jitter characteristic similar to the two Premium Services by utilizing EF resources, i.e., resources exclusively available for Premium Service traffic, which are currently not in use. Within a single DiffServ node, these EF resources consist of the EF bandwidth and the EF buffer per outgoing link.

Bandwidth For an EFD aggregate, no bandwidth is exclusively available (i.e., reserved by node configuration) at an EFD-compliant DiffServ node. Instead, EFD packets utilize unused EF bandwidth following the main purpose of BELD Service to increase the utilization of the EF resources. Hence, there is no node configuration necessary for the EFD PHB with regard to bandwidth.

Buffers Similar to the resource 'bandwidth', EFD packets utilize EF buffers only, additional buffers must not be added. Otherwise, additional queuing delays could occur.

As described previously, queuing of EF packet can become necessary because of the busy-interface effect or the simultaneous fan-in effect. Both situations affect the queuing delay a single EF packet experiences and both cannot be

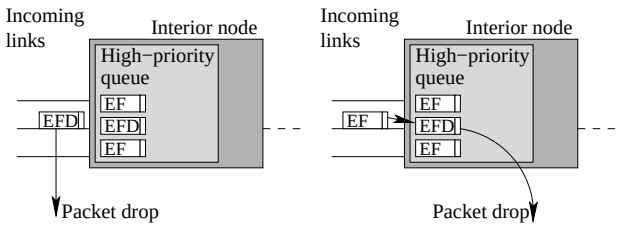
avoided. Thus, both have to be considered for worst-case delay or delay jitter bound calculations in scenarios with only Premium Service traffic. For the characterization of an EF-compliant node [3], that means that both effects are already included in the error terms E_a and E_p because these bounds are said to hold 'independent of the amount of non-EF traffic' offered to the EF node.

When introducing EFD traffic, the probabilities for both effects increase because the utilization of EF bandwidth increases. Therefore, EFD packets will affect the average queuing delay and the average jitter of EF packets. However, as the worst-case probabilities for both situations have to be considered already in case of no EFD traffic, the maximum bound on the delay for a single EF node is not affected by the introduction of EFD traffic.

Node configuration may be used to limit the amount of EF buffers to be used by EFD packets. For example, if an EFD-compliant node has a high in-degree, the EF buffers must be rather large to accommodate the simultaneous fan-in effect. In such a case, it may be desirable to limit the amount of EF buffers which EFD packets are allowed to use.

3.1.1 Function Overview

The main function of the EFD PHB is to ensure that EF packets can utilize EF resources independent of the amount of EFD traffic in the network. For this purpose, two situations have to be considered which are both based on the 'busy-interface effect' (cf., Fig. 4):



(a) EFD packet arriving at a node with filled EF buffers

(b) EF packet arriving at a node with filled EF buffers

Figure 4. EFD PHB: Basic behavior

1. An EFD packet arrives and there are not sufficient resources left in the EF buffer to store the EFD packet. In this case, the EFD packet has to be dropped (cf., Fig. 4 (a)).

2. An EF packet arrives and there is not sufficient space in the EF buffers. If the buffer is filled with EF packets, this is the rare case of an EF packet to be dropped (which is independent from whether this node implements the EFD PHB or not). If there are EFD packets within the EF buffer, one or several EFD packets have to be dropped to release buffer resources for the arriving EF packet (cf., Fig. 4 (b)).

3.1.2 Applicability

In general, the EFD PHB can be implemented on every node within a DiffServ domain, which also implements the EF PHB. At a first glance, it is not necessary on interior nodes but only at network boundaries in the following situation:

The basic EF PHB requirement is that the sum of the incoming EF traffic, destined to a certain outgoing interface, must be smaller or equal to the configured EF capacity on that outgoing interface. If the boundary nodes limit the EF traffic so that this requirement is fulfilled in all interior nodes, packet loss should not occur. Thus, if boundary nodes ensure that also the sum of the EF and the EFD traffic fulfills this requirement, packet drop of EFD packets will not occur in the domain either. As explained above, dropping EFD packets is the main task of the EFD PHB. It is, therefore, not necessary to implement the EFD PHB on those nodes where no EFD packets are to be dropped. Thus, the EFD PHB appears to be only necessary on nodes which perform traffic conditioning, such as the boundary nodes.

The EFD PHB is, however, necessary on interior nodes, though, since already a concrete implementation of the EF PHB cannot comply with the above mentioned basic EF requirement in all situations. For this reason, the EF PHB definition has introduced the error terms E_a and E_p to characterize to what extent an EF-compliant node may violate the EF PHB requirements. As an example, EF packets may rarely have to be dropped even within a domain if the sum of the EF traffic for a single outgoing interface exceeds the configured EF rate temporarily. This may, therefore, occur in case of the EFD PHB as well: It may become necessary to drop EFD packets preferentially in the rare case where the sum of the EF traffic and the EFD traffic exceeds the configured EF rate on the outgoing link temporarily.

For shared-media access networks, such as Ethernet or IEEE-802.11-based Wireless LANs, a prioritization mechanism is necessary to implement service differentiation (e.g., a Wireless LAN based on the enhanced distributed coordination function [9]). This can be activated from the IP layer by an appropriate translation of the QoS parameters and is already necessary, for example, to implement an EF-based service over a shared-media access network. To provide EFD-like QoS mechanisms for such a QoS-enhanced media access, this prioritization mechanism must be extended to support BELD Service. This enhancement can be simple, e.g., using the next lower priority compared to EF traffic.

3.2 Traffic Conditioning

For BELD Service, traffic conditioning has two purposes:

1. It is necessary at domain boundaries to avoid that EF/EFD traffic can preempt other traffic without limitations. In contrast to networks with only EF traffic, where only misbehaving traffic sources enforce a traffic conditioning, admission control cannot avoid such a situation in networks with EFD traffic. This is because admission control is not mandatory to implement a best-effort-like service such as BELD Service.

2. It protects EF resources from overload.

The latter means to ensure that the sum of EF traffic and EFD traffic arriving at a DiffServ domain is in conformance to the negotiated EF rate. This function can be performed using a token bucket. To keep EF queues within the network small, strict traffic shaping is necessary at the domain boundary for EF traffic. A token bucket performing traffic conditioning requires the usage of a small queue for traffic shaping if the output flow should not be bursty. This is especially important on boundary nodes, where several EF aggregates are merged.

When introducing additional EFD packets into the network, the sum of the EF/EFD packets might exceed the configured EF rate. This leads to extensive queuing and packet dropping of EFD packets at the token bucket queue. In this case, the token bucket must implement the EFD PHB to avoid dropping of EF packets. However, EF packets experience a higher queuing delay in this overload case because an arriving EF packet might see an empty bucket.

As an example, Figure 5 depicts two adjacent domains where domain A is allowed to send EF traffic with a rate of $40\,\mathrm{kbit}/s$. To limit the inserted EF/EFD traffic, domain B in-

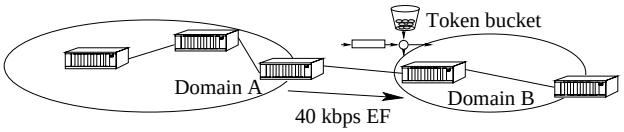


Figure 5. Example: Traffic conditioning

stalls a token bucket at the domain boundary with the MTU as maximum burst size to shape the aggregated traffic flow. Furthermore, the token generation rate is $40\,\mathrm{kbit}/s$ and the token bucket queue size is a single MTU-sized packet. Owing to the introduction of BELD Service traffic, the terminals in domain A send more than $40\,\mathrm{kbit}/s$ temporarily. In this case, an EF packet in the token bucket queue must wait up to $\frac{MTU}{40kbit/s}$ at the token bucket for the generation of sufficient tokens. This *traffic shaping delay* is $20\,ms$ for an example MTU of 100 bytes which is a reasonable packet size for a low-delay class intended for interactive applications. The traffic shaping delay decreases when the configured token bucket rate increases.

The traffic shaping delay, resulting from an empty bucket on an EF packet arrival, might be caused by an EFD packet which was sent immediately before the EF packet arrived. In such a case, EFD packets 'steal' tokens which belong to the EF resources. This cannot be avoided if the EFD packets share the token bucket with the EF packets. Thus, the introduction of EFD traffic might increase the maximum delay of EF packets. This would not occur in absence of EFD packets if all sender of EF packets conform to their negotiated rate. Nevertheless, the new maximum delay is still bounded which depends on the buffer size of the traffic shaper.

3.3 Admission Control

Admission control is normally not necessary for a 'best-effort' service such as BELD Service. In this case, packet loss of BELD Service packets can occur for an extended period of time if EF packets alone highly utilize the EF resources or in case of a high amount of BELD Service traffic.

Nevertheless, admission control may still be performed for two reasons:

1. To limit packet loss in BELD Service.

2. To avoid additional delays of EF packets at the network boundary because of traffic conditioning.

Packet loss in BELD Service should be limited if adaptive applications having a low-delay requirement can only deal with a limited packet loss rate such as, for example, mobile telephony applications based on the AMR codec. Furthermore, EF packets experience a delay in the queue of a traffic-shaping token bucket in case the sum of the EF and the EFD traffic exceeds the configured token generation rate. To avoid this delay, admission control could limit the amount of BELD Service traffic inserted into the network. The challenge of admission control for BELD Service is to find a good compromise between the utilization of Premium Service resources and the packet loss rate of BELD Service and to estimate the amount of resources available for BELD Service in the future.

3.4 Related Work

In adding a drop precedence scheme to Premium Service, both, Mobile Premium Service and BELD Service appear similar to the real-time class of the SIMA model [13]. However, *MoDiQ* does not aim at dynamically adjustable packet loss rates but at the provisioning of two different low-delay services: Mobile Premium Service with a low

packet loss rate and BELD Service with a higher packet loss rate. Additionally, SIMA is a qualitative service with neither admission control nor guarantees on the available bandwidth. Mobile Premium Service gives such a guarantee and even for BELD Service packet loss can be limited by means of admission control.

BELD Service is similar to the service models for the Alternative Best-Effort Service [10] or the Equivalent DiffServ proposals [7] in that all provide a low-delay Best-Effort Service. However, Alternative Best-Effort Service and Equivalent DiffServ are targeted at networks without support for quantitative services whereas BELD Service is a complement to Mobile Premium Service, which gives quantitative assurances using admission control means.

4. Feasibility of BELD Service

This section elaborates on the feasibility of BELD Service using the network simulator ns-2. It shows the influence of introducing BELD Service on the service characteristics of Mobile Premium Service. The following QoS parameters have to be considered:

- Resource utilization: This is the primary performance metric to measure the gain of introducing BELD Service. It should be higher than for scenarios with only Mobile Premium Service traffic.
- Packet loss: Should not occur for Mobile Premium Service but for BELD Service in case of high loads.
- Average packet delay and jitter: Should increase owing to a higher utilization of EF buffers.
- Maximum packet delay and jitter: Should only increase if traffic shaping becomes necessary.

The simulation model is as shown in Figure 6. Traffic is sent from the mobile terminals via the base station to the communicating partner. Nine base stations provide connectivity for a 3x3 mobility cell topology. The distance between two base stations is 700 m horizontally and vertically which is a typical distance for mobile networks in a densely populated city area. The cell size is 800 m so the coverage areas of two neighboring base stations overlap up to 100 m to enable soft handoffs without interruptions of connectivity. The handoff control algorithm is based on a hysteresis [14] which can avoid subsequent handoffs between two base stations within a short period of time (the so-called *flip-flop effect*). The wireless network is based on the IEEE 802.11 standard to simulate realistic effects such as collisions on the air interface. Each of the nine base stations is connected to a wired router via a 1 Mbit/s link, which constitutes the bottleneck in this scenario. In the following simulations, there is a simple priority scheduler combined with a policing and

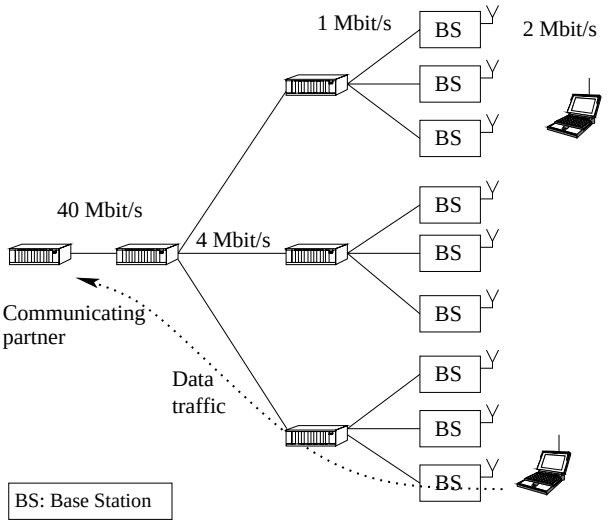


Figure 6. The simulation model

shaping token bucket on each base station for the link towards the communicating partner. The priority scheduler ensures that Mobile Premium Service and BELD Service packets are forwarded preferentially. A token bucket is used for traffic conditioning to ensure both, that the sum of the Mobile Premium Service and BELD Service traffic does not exceed the configured rate and that the stream is burst-free. The configured rate of the token bucket is chosen according to the configured capacity of each base station (50 Mobile Premium Service flows with 2 kbit/s). To accommodate the 'simultaneous fan-in effect', the EF buffers on the base station must be sized to carry at maximum 50 EF packets. Since this would lead to very high delays if BELD Service packets are allowed to utilize the whole EF buffer, BELD Service packets are restricted to use only one-fifth of the EF buffer corresponding to ten MTU-sized packets. The remaining links in the backbone are over-dimensioned, so no resource shortages occur there. The simulated application is mobile telephony with 2 kbit/s CBR traffic. The mobile terminals move according to a random mobility pattern, which is such that the center cell has a higher resource utilization than the remaining cells.

The number of mobile terminals injected into a simulation is determined by the *offered load* [2]. Intuitively, the offered load is defined such that at an offered load of 0%, no new sessions arrive. At an offered load of 100%, the new session arrival rate is such that no new session request has to be blocked and all resources of the cell are busy under the following assumptions:

- All terminals are static/portable.
- All sessions start simultaneously.

- The session duration is constant for all sessions.

In reality, new session requests will be blocked earlier than at an offered load of 100% because the above assumptions do not hold in reality. In the following simulations, the offered load relates to a basic level of Mobile Premium Service traffic only. A variable amount of BELD Service traffic is inserted into the scenario in addition to this basic offered load. To achieve a fair comparison between scenarios with/without BELD Service traffic, all simulation sets are run twice: In the first run, all additional sessions inserted into the simulations request Mobile Premium Service. In the second run, the same mobility pattern is reused but the additional sessions request BELD Service this time. This way, a direct comparison between identical scenarios with/without BELD Service becomes possible. BELD Service sessions are always admitted, there is no admission control for BELD Service. This way, the influence of BELD Service onto the delay of Mobile Premium Service traffic should be evaluated. As discussed previously, an additional delay of Mobile Premium Service packets can occur as a result of a delay in traffic shaping: If the amount of BELD Service and Mobile Premium Service traffic is larger than the configured rate of the token bucket on the boundary node, traffic shaping and policing becomes necessary to protect interior nodes in the domain from congestion.

4.1 Resource Utilization

The primary objective of BELD Service is to increase the utilization of Mobile Premium Service resources. Figure 7 depicts the difference in the accepted sessions for an offered load varying between 50% and 150% and an additionally introduced amount of Mobile Premium Service (MPS in the figure) and BELD Service traffic between 0% and 40%. All simulation results have been validated by running the simulations several times to gain average values, the errors bars depict the maximum and minimum values. The simulation

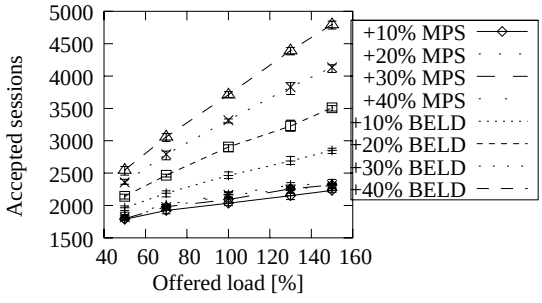


Figure 7. Accepted sessions in BELD Service

results of the first simulation run with only Mobile Premium Service sessions (the four bottommost graphs) are almost identical independent whether 10%, 20%, 30% or 40% additional Mobile Premium Service sessions have been added. This is because admission control of Mobile Premium Service denies additional new session requests when the sum of the negotiated Mobile Premium Service peak rates exceeds the configured Mobile Premium Service link resources. The small increase in the number of accepted sessions even at high offered loads is, among other reasons, caused by the exponential distribution of the session duration: Since an exponential distribution of the sessions duration leads to many short sessions, some sessions can still be accepted even at a high offered load when short sessions terminate (cf., App. A.2.4 in [4]).

In contrast, the number of accepted sessions of the second simulation set with additional BELD Service sessions is increasing for both, an increasing offered load and an increasing percentage of additional BELD Service sessions. This is as expected since there is no admission control for these additional BELD Service sessions.

4.2 Packet Loss

Figure 7 depicts the percentage of packet loss for BELD Service packets (Mobile Premium Service packets do not experience any drops which is, thus, not shown in Fig. 8).

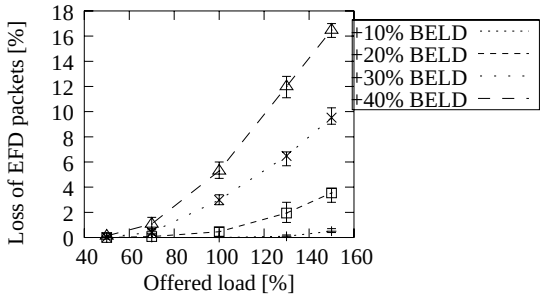
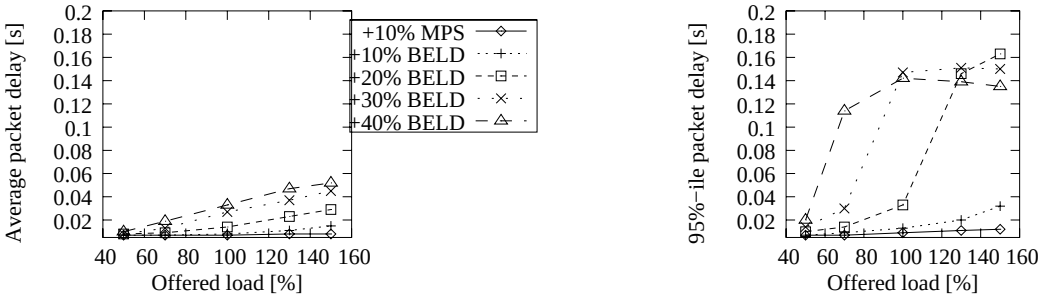


Figure 8. Packet loss in BELD Service

Almost no packet loss occurs for BELD Service packets when 10% additional BELD Service sessions are inserted into the simulation. This is because BELD Service packets can utilize unused Premium Service capacity which is reserved for handoff purposes but not in use. For 20% additional BELD Service sessions, packet loss occurs at an offered load of 100% and higher and for 30% and 40% additional BELD Service sessions above an offered load of 70%. These numbers are important for the following discussion on the packet delay: Drops of BELD Service packets are a sign that the sum of the data rates of the Mobile Premium Service and BELD Service flows exceeds the configured EF rate of the domain. Thus, the boundary nodes drop BELD



(a) Average delay

(b) 95%-ile delay

Figure 9. Influence of BELD Service on Mobile Premium Service packet delay

Service packets which increases the delay of Mobile Premium Service packets owing to a delay in traffic shaping.

4.3 Delay

To separate the effects of the media access schemes on the wireless link, the delay is measured between the base station and the communicating partner only. Figure 9 depicts the average delay and the 95%-ile of the delay for Mobile Premium Service sessions. The delay of BELD Service packets is not shown here because it is in general the same or slightly lower than the delay of Mobile Premium Service packets because of the limitations on the EF buffer usage.

For comparison, both figures depict the delay from the first simulation run with 10% additional Mobile Premium Service sessions inserted into the network. The average delay is constant at a low value of $8\,ms$. The upper 95%-ile bound on the delay remains also almost constant over the viewed offered load range, but at about $10\,ms$. However, the results are different for the second simulation run with additional BELD Service sessions:

- If no packet loss occurs for BELD Service packets, the delay and the 95%-ile delay are the same as for Mobile Premium Service packets. This is true, for example, for the '+10% BELD Service' case up to an offered load of 130% and for the '+20% BELD Service' case up to an offered load of 70%.

- If there are losses for BELD Service packets, the delay increases significantly, For example, in the '+30% BELD Service' case, packet loss occurs at an offered load of 70%. In this case, the 95%-ile of the delay increases to above $20\,ms$ for an offered load of 70% and then rapidly higher for high offered loads. The average delay starts increasing from an offered load of 100% where the amount of packet loss is becoming significant (i.e., larger than 1%).

- The rate of change for the 95%-ile delay decreases for the scenarios with many additional BELD Service sessions (e.g., '+40% BELD Service') at high offered loads. This is because of the high amount of packet drops which cuts the tail of the probability distribution for the packet delay.

Furthermore, the maximum delay increases to up to $375\,ms$ because of the large EF buffer (50 packets). However, the theoretical bound, as gained from the following Equation, is not exceeded as expected.

$$\frac{\textit{Buffer size}}{\textit{EF rate}} = \frac{800\,\textit{bit} \cdot 50\,\textit{entries}}{100.000\,\textit{bit/s}} = 400\,\textit{ms} \qquad (1)$$

In different scenarios with a lower number of potentially simultaneously arriving Mobile Premium Service packets, the maximum delay will be significantly lower.

Thus, the introduction of too much BELD Service traffic affects the average delay and the 95%-ile delay of packets from Mobile Premium Service sessions. It depends on the type of the application and on the network whether this influences an application significantly or not. If the application can tolerate a delay of up to $100\,ms$, the introduction of additional 10% BELD Service sessions is applicable in this scenario: The average Mobile Premium Service delay remains below $15\,ms$, the 95%-ile below $40\,ms$ and the packet loss rate of BELD Service remains below 1%. Therefore, BELD Service can be used in this case, for example, for a low-cost mobile telephony service using the AMR codec [1] which produces a comprehensible speech service even under a packet loss rate of 4%. The shaping delay depends heavily on the token bucket rate, i.e., the aggregated amount of Mobile Premium Service traffic arriving at the domain boundary. Thus, the decision if admission control

is necessary must be taken for each network individually and cannot be generalized.

4.4 Jitter

The jitter is very low (less than $4\,ms$) even for high offered loads. Similar to the result for the packet delay, the average jitter increases if the amount of inserted BELD Service traffic leads to an additional traffic shaping delay. The 95%-ile of the jitter also increases if a traffic shaping delay occurs but it remains below $15\,ms$ in this scenario. This does not change even if an additional $400\,\text{kbit}/s$ background traffic is inserted at the base station. Hence, the introduction of BELD Service should not influence interactive applications such as mobile telephony which require a low jitter.

5. Summary and Future Work

The Best-Effort Low-Delay Service is one part of the *MoDiQ* proposal to enhance the legacy DiffServ service model to be used in wireless mobile networks. It is a complement to the Premium Services in that is has the same low-delay characteristics, but gives no assurances on the packet loss rate. It has two main objectives, viz. to increase the utilization of network resources and to provide an especially tailored service for interactive real-time applications which require a low delay but can deal with a certain amount of packet loss. BELD Service is applicable for both, wired and wireless networks. However, it is especially suited for the latter since some Premium Service resources are unused there, for example, in presence of a handoff resource reservation scheme, which reserves some resources in a cell to avoid a resource shortage after a handoff.

The simulations have shown that BELD Service can in principle increase the network utilization at a moderate packet loss rate. In the simulated scenario, an additional 10% of BELD Service sessions does not affect the Mobile Premium Service QoS parameters such as the delay or delay jitter. However, these simulation are preliminary only as they show the feasibility of BELD Service in one particular scenario only. They do not show how the additional number of BELD Service sessions, to be introduced without affecting the delay of Mobile Premium Service sessions, can be gained without experimentation. Furthermore, the increase in network utilization comes at the cost of a higher average delay and a possibly higher maximum delay caused by traffic shaping. Whether this is a problem or not for a real-time application using Premium Service, depends on a concrete, more realistic network scenario, which has to be evaluated in future work.

It is expected that more BELD Service sessions can be added in a scenario where not only the unused handoff resources can be utilized by BELD Service sessions but also unused resources from non-constant-bit-rate traffic (e.g., due to silence suppression in voice applications).

References

[1] AMR speech Codec: General description (Release 4). Technical report, 3GPP: Technical Specification Group Services and System Aspects, Apr. 2000.

[2] S. Choi and K. Shin. A comparative study of bandwidth reservation and admission control schemes in QoS-sensitive cellular networks. *ACM Wireless Networks*, 6(4):289–305, 2000.

[3] B. Davie, A. Charny, J. Bennet, K. Benson, J.-Y. Le Boudec, W. Courtney, S. Davari, V. Firoiu, C. Kalmanek, K. Ramakrishnam, and D. Stiliadis. An Expedited Forwarding PHB. RFC (Proposed Standard) 3246, IETF, Mar. 2002.

[4] J. Diederich. *Simple and Scalable Quality of Service for Wireless Mobile Networks*. Shaker Verlag, Aachen, Germany, July 2003. Doctoral thesis, University of Karlsruhe.

[5] J. Diederich, L. Wolf, and M. Zitterbart. A Mobile Differentiated Services QoS Model. In *Proc. of the 3rd IEEE Workshop on Applications and Services in Wireless Networks (ASWN)*, Berne, Switzerland, July 2003. Accepted for publication.

[6] J. Diederich and M. Zitterbart. An Expedited Forwarding with Dropping PHB. Internet Draft, IETF, Oct. 1999. Work in progress.

[7] B. Gaidioz and P. Primet. The Equivalent Differentiated Services Model. Research Report 2002-09, RESAM, INRIA, France, Feb. 2002.

[8] L. Georgiadis, R. Guérin, V. Peris, and K. Sivarajan. Efficient Network QoS Provisioning Based on per Node Traffic Shaping. *IEEE/ACM Transactions on Networking*, 4(4):482–501, Aug. 1996.

[9] D. Gu and J. Zhang. QoS Enhancement in IEEE802.11 Wireless Local Area Networks. *IEEE Communications Magazine*, 41(6):120–124, June 2003.

[10] P. Hurley, J.-Y. Le Boudec, P. Thiran, and M. Kara. ABE: Providing a Low-Delay Service within Best-Effort. *IEEE Network*, 15(3):60–69, May 2001.

[11] V. Jacobson, K. Nichols, and K. Poduri. The 'Virtual Wire' Per-Domain Behavior. Internet Draft, IETF, July 2000. Work in progress.

[12] K. Nichols, S. Blake, F. Baker, and D. Black. Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers. RFC (Proposed Standard) 2474, IETF, Dec. 1998.

[13] J. Ruutu and K. Kilkki. Simple Integrated Media Access – a Comprehensive Service for Future Internet. In *Proc. of the IFIP Conference on Performance of Information and Communications Systems (PICS)*, Lund, Sweden, May 1998.

[14] N. Tripathi, J. Reed, and H. VanLandingham. Handoff in Cellular Systems. *IEEE Personal Communications Magazine*, 5(6):26–37, Dec. 1998.

[15] L. Zhang, V. Jacobson, and K. Nichols. A Two-bit Differentiated Services Architecture for the Internet. RFC (Informational) 2638, IETF, July 1999.