Group Communication in Differentiated Services Networks

Roland Bless and Klaus Wehrle
Institute of Telematics
Universität Karlsruhe (TH)
D-76128 Karlsruhe, Germany
{bless,wehrle}@telematik.informatik.uni-karlsruhe.de

Abstract

The Differentiated Services (DiffServ) approach will also bring benefits for multicast applications which need quality of service support. For instance, a highly reliable multicast service can be provided based on the proposed Expedited Forwarding behavior [8]. Such a service may also be used advantageously in a global computing cluster infrastructure, e.g., for distribution of synchronization messages. However, DiffServ multicast services have not been addressed in a very detailed manner yet.

This paper illustrates some of the problems which will arise when IP Multicast is used in DiffServ networks without taking special precautions into account for providing it. Those problems mainly lead to situations in which other service users are affected adversely. In order to retain the benefits of the DiffServ approach, a quite simple and scalable solution for those problems is needed, not resulting in additional complexity or costs in a DiffServ domain. The proposed architecture in this paper requires only an additional entry for the DiffServ Codepoint in multicast routing tables and some support by management mechanisms. The discussion of the related problems and presentation of the solution is illustrated and confirmed by some measurements performed with a Linux implementation of DiffServ and an adapted Linux Multicast Router.

1 Introduction

Internet services offering a better quality than the current deployed best-effort service are urgently required. In addition to group communication, many advanced applications (such as global cluster computing) need certain assurances from the network layer, e.g., a maximum delay, a minimum packet loss rate or guaranteed transmission rate. Global cluster computing scenarios will benefit from such high quality dissemination services for their message passing infrastructure [10]. Currently used IP mechanisms are not able to offer such guarantees.

The IETF attempted to meet these trends in defining the Integrated Services (IntServ) architecture, which provided quality based services even for group communication scenarios in the Internet. However, the IntServ Architecture shows some inherent scalability problems if applied Internet-wide, especially within backbone areas. Because service differentiation in the Internet was and is still required, the Differentiated Services (DiffServ) Architecture [1, 3] was developed to overcome these scaling problems. Scalability is achieved by avoiding complexity and maintaining per-flow state information in core routers and pushing unavoidable complexity to the network edges. Therefore, individual flows belonging to the same service are aggregated, thereby eliminating the need for complex classification or managing state information per flow in interior routers.

On the other hand, reduced complexity in routers makes it more complex to provide such better services together with IP Multicast. Basic DiffServ mechanisms (which are briefly described in section 2) can be likewise applied to IP Multicast forwarding, whereas providing better services with quality of service (QoS) based on a combination of DiffServ and multicast is not straightforward. Some problems which emerge from this fact are described in section 3. An architecture for solving for those problems is suggested in section 4. Simplicity of the solution was a major objective in order to not defeat the so far gained advantages of DiffServ. In section 5, the previously described problems and their solution are demonstrated by measurement results of an implementation.

2 Differentiated Services and IP Multicast

In the DiffServ Architecture services can be constructed from *per-hop forwarding behaviors (PHB)* and some related *traffic conditioning* actions (e.g., metering, marking, shaping or dropping) which are applied to packets along their

path. The forwarding behavior that a packet experiences is identified by a so-called *codepoint* in the IP packet header. Each codepoint (DSCP) is a specific value conveyed by the *Differentiated Services Field (DS-field)* that replaces a part of the common *Type of Service* (ToS) field in IPv4 packets and the class field in IPv6 packets [1]. Different PHBs may use distinct queueing mechanisms in order to achieve the intended differential forwarding treatment of packets.

A packet is usually classified and marked to receive a particular forwarding behavior in the first DiffServ-capable node along its path (the so-called 'First-Hop Router', cf. Fig. 1). Classification at this early stage may still be at the granularity of an end-to-end 'microflow' by considering multiple fields of a packet, e.g., source/destination addresses, ports and protocol ID number. Such a classification rule for packets is part of a *Traffic Conditioning Specification (TCS)* which also comprises a corresponding traffic profile, i.e., a description of a traffic stream's properties. Thus, packets are marked according to their corresponding traffic profile that is selected by the classifier.

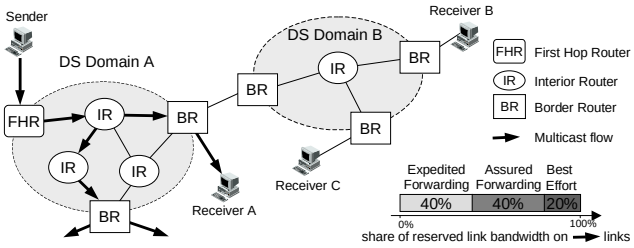


Figure 1. Example of two DiffServ Domains using IP Multicast with reserved bandwidth

Packets on the same link in a particular direction carrying the same codepoint are denoted as *Behavior Aggregate (BA)*. After initial setting of the codepoint, subsequent nodes on the path typically operate only on those aggregates. Therefore, 'interior routers' (cf. Fig. 1) only have to classify packets by their specific codepoint and treat them with the corresponding forwarding mechanism. Consequently, this leads to higher scalability by avoiding complexity and per-flow state information especially in 'core' routers located in inner network regions.

Besides first-hop routers there are also 'border routers' interconnecting a DiffServ (DS) domain with other domains (the DiffServ Architecture subsumes both types in the notion of *'boundary nodes'*). Their task is to police and condition traffic that is leaving a DS domain or flowing into it. Thus, they are often acting as ingress and egress border router. Classification will be simpler compared to first-hop routers, because traffic has to be differentiated only in dependence on the behavior aggregate and incoming or outgoing link. Therefore, they are also highly scalable, because they do not have to keep and maintain per flow-states and reservation information. They will use profiles for different (BA,link) pairs instead.

It is important to notice that the first-hop router mainly determines the service that an incoming packet will experience, whereas all subsequent routers typically cannot change codepoints for a specific microflow, but only for all packets of an aggregate. Changing a codepoint for packets of a particular flow in the interior network requires usually per-flow classification, thus leading to the same scalability problems which the IntServ approach possesses.

2.1 Services and Per-Hop Behaviors

Based on the *Expedited Forwarding PHB (EF)* [8] a guaranteed bandwidth, low delay and low loss service can be provided, showing the same characteristics as a 'virtual leased line'. This is achieved by keeping EF queues very small or almost empty, which in turn can only be accomplished by guaranteeing that the maximum arrival rate of an aggregate is less than that aggregate's minimum departure rate. Thus, admission control, policing and shaping are ordinarily needed with respect to the guaranteed rate as a required configuration parameter. Usually, all packets waiting in the EF queue for transmission are served before all packets waiting in queues of other PHBs.

In contrast, a service based on the *Assured Forwarding* PHB Group (AF) [2] may permit a statistical guaranteed rate only. It would allow senders to use additional available capacity while providing a minimum base rate. Packets exceeding the negotiated rate are marked for subsequent treatment with higher drop precedence. Consequently, bursts of packets belonging to an AF flow may transit successfully a DS domain if enough capacity is available.

Because packets carry the DSCP which determines their forwarding treatment, the basic DiffServ mechanisms work also for multicast ('mc-') packets. Replicated packets may get the same DSCP as the original incoming mc-packet, and, consequently show the same forwarding behavior. Applying the EF PHB to a multicast context leads to a service with very interesting properties. The very low packet loss characteristic makes it suitable as a basis for a highly (but not absolute) reliable multicast service. Packet loss cannot be fully precluded, because of aggregation effects that may lead to packet dropping. Nevertheless, in reality packet losses should occur so infrequently that many applications can tolerate these losses, or, if this is not the case, that at least very simple retransmission schemes can be applied.

2.2 Management of Differentiated Services

Admission control and resource reservation are at least for EF-based services required. Furthermore, installation

and updating of traffic profiles in boundary nodes are necessary. Additionally, offering *services on demand* requires some kind of signaling and automatic admission control procedures. Therefore, the concept of *Bandwidth Brokers* was already suggested by Van Jacobson at a very early stage of DiffServ development (cf. RFC 2638). In this concept, the *Bandwidth Broker (BB)* is a dedicated node in each DS domain, keeping track of the amount of available and reserved bandwidth for services within the domain, and, processing admission control requests from customers or BBs of adjacent domains. Moreover, it installs or alters traffic profiles in boundary routers and automatically negotiates parameters of service level specifications bilaterally with adjacent domains.

Protocols for signaling a reservation request to a DS Domain are required. For accomplishing end-system signaling to DiffServ domains RSVP may be used with new DS specific reservation objects. RSVP is mainly designed for use in multicast scenarios and is already supported by many operating systems. However, when applying RSVP to a DiffServ network some problems will arise which are briefly described in the next section.

3 Problems of Providing IP Multicast in DS Domains

As mentioned before, while the basic DiffServ mechanisms work also with IP Multicast, supplying DiffServ multicast services is not straightforward. Although potential problems and the complexity of providing multicast with Differentiated Services are considered in a separate section of [3], both aspects have to be discussed in greater detail. The simplicity of the DiffServ Architecture and its router models is necessary to reach high scalability, but it causes also fundamental problems in conjunction with the provision of IP Multicast in DS domains.

For subsequent considerations we assume, unless stated otherwise, at least a unidirectional point-to-multipoint communication scenario in which the sender transmits packets with an assigned 'better' PHB than the traditional default PHB, resulting in a service of better quality compared to the default best-effort service. In order to accomplish this, a traffic profile corresponding to the traffic conditioning specification has to be installed in the sender's first-hop router. Furthermore, it must be assured that the corresponding resources are available on the path from the sender to all receivers, possibly requiring adaptation of traffic profiles at involved domain boundaries. But providing suitable resources is difficult, because receivers can dynamically join and leave an mc-group anytime, thereby leading to a dynamic resource consumption. If this fact is not considered, it will lead to the problem described in the next section.

3.1 Neglected Reservation Subtree Problem

Typically, resources for some DiffServ services must be reserved before they will be actually used. But in an mc-scenario, group membership is often highly dynamic, therefore limiting the use of a sender-initiated resource reservation in advance. Unfortunately, dynamic addition of new members of the mc-group using Differentiated Services can adversely affect existing other traffic, if resources were not explicitly reserved before use.

IP Multicast packet replication usually takes place when the packet is handled by the routing process. Thus, a DiffServ capable node would also copy the content of the DS field [1] into the IP packet header of every replicate. Consequently, replicated packets get exactly the same DS codepoint as the original packet, and, therefore experience the same forwarding treatment as the incoming packets of this mc-group. Normally, the replicating node cannot test whether a corresponding reservation exists for a particular flow of replicated packets on an output link (resp. its corresponding interface), because a flow-specific traffic profile is usually not available in boundary and interior nodes (except in first-hop nodes).

When a new receiver joins an IP mc-group, the corresponding multicast routing protocol (e.g., DVMRP, PIM-DM or PIM-SM) accomplishes that the multicast tree is expanded by a new subtree which connects the new receiver to the already existing mc-tree. As a result of tree expansion and missing per-flow classification mechanisms (cf. section 2), the new receiver will implicitly use the 'value-added service' of better quality.

If the additional amount of resources which are consumed by the new part of the mc-tree are not taken into account by the domain management (cf. section 2.2), the currently provided QoS level of other receivers (with correct reservations) will be adversely affected or violated. This negative effect on existing traffic contracts by a neglected reservation – in the following designated as *Neglected Reservation Subtree Problem (NRS Problem)* – must be avoided under any circumstances.

One can distinguish two distinct major cases of the NRS Problem. In order to compare their different effects a simple example of a share of bandwidth is illustrated in Fig. 1. Three types of services (respectively their corresponding behavior aggregates) share the bandwidth of the considered output link: Expedited Forwarding, Assured Forwarding and the traditional Best-Effort service. In this example we assume, in accordance to the implementation in KIDS [4], that routers perform simple priority queueing, where EF has the highest and Best-Effort the lowest assigned priority. When Weighted Fair Queueing (WFQ) would be used, the described effects would also occur, only with minor differences.

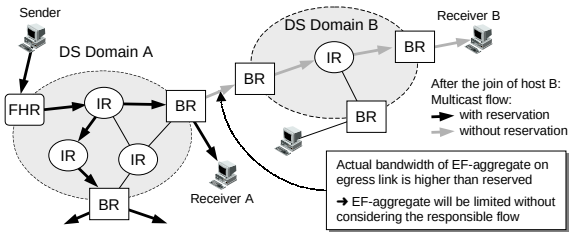
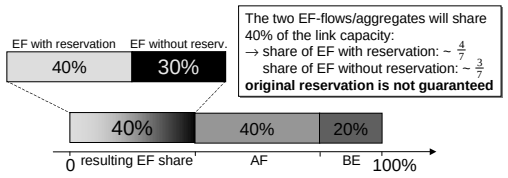
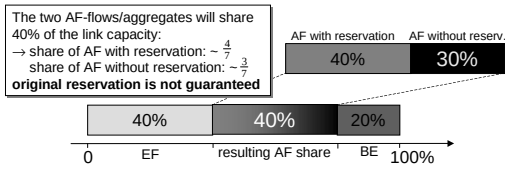


Figure 2. NRS Problem case 1, after join of B

The NRS Problem occurs in two different cases:
Case 1 — If the branching point of the new subtree and the previous mc-tree is an (egress) border router (cf. Fig. 2), the additional mc-flow increases the amount of used resources for the corresponding aggregate and will be greater than the originally reserved amount. Consequently, the policing component in the egress border router discards packets until the traffic aggregate is conforming to the traffic contract. But during discarding packets the router cannot identify the responsible flow (because of missing flow classification functionality at this level), and, thus randomly discards packets, whether they belong to a correctly reserved flow or not. As a result, there will be no longer any service guarantee for the reserved flows.



(a) Expedited Forwarding



(b) Assured Forwarding

Figure 3. Resulting share of bandwidth between Border Routers with a neglected reservation

Fig. 3 shows the resulting share of bandwidth in cases when (a) Expedited Forwarding and (b) Assured Forwarding are used for the flow causing the NRS Problem. Assuming that the additional traffic would use another 30% of link bandwidth, Fig. 3 (a) illustrates that the resulting EF aggregate (70% of the outgoing link bandwidth) is throttled down to its originally reserved 40%. In this case, the amount of dropped EF bandwidth is equal to the amount of excess bandwidth. The marked parts in Fig. 3 indicate that the complete EF aggregate is affected by packet losses. The other services, e.g., AF or Best-Effort, are not disadvantaged. Fig. 3 (b) shows the same situation for AF. The only difference is that AF is now affected by discards and remaining services will get their guarantees.

In either case, packet losses are restricted to the misbehaving service class by the traffic meter and policing mechanisms in border routers. Moreover, the latter problem (case 1) occurs only in egress border routers, because they are normally responsible, that not more traffic leaves the DS domain, than the following ingress border router will accept. Therefore, those violations of service level agreements will be already detected and processed in egress border routers.

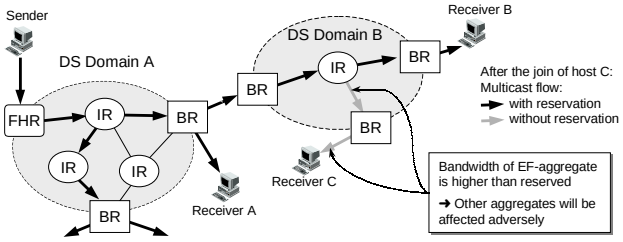
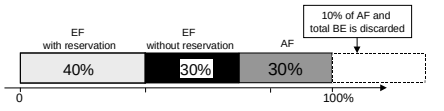


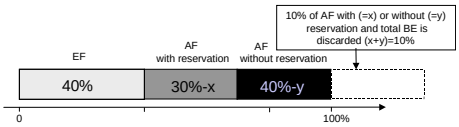
Figure 4. NRS Problem case 2, after join of host C

Case 2 — The NRS Problem can also occur, if the branching point between the previous mc-tree and the new subtree is located in an interior router (as shown in Fig. 4). Because the router is usually not equipped with metering or policing functions it will not recognize any excess amount of traffic and will forward the new mc-flow. If the latter belongs to a higher priority service, such as EF, bandwidth of the aggregate is higher than the aggregate's reservation and it will steal bandwidth from lower priority services. The additional amount of EF without a corresponding reservation is forwarded together with the aggregate that has a reservation. This results in no packets losses for EF as long as the resulting aggregate is not higher than the output link bandwidth. Because of its higher priority, EF gets as much bandwidth as needed and as is available (strictly speaking, it is implementation dependent whether interior routers have something like a maximum configured service rate).

As a result, there is no restriction for EF, but as Fig. 5 (a) shows, other services will be extremely disadvantaged by this use of non-reserved resources. Their bandwidth is stolen by the new additional flow. In this case, the additional 30% EF traffic preempts resources from the AF traffic, which in turn preempts resources from the best-effort traffic, resulting in 10% packet losses for the AF aggregate and complete loss of best-effort traffic. The example



(a) Expedited Forwarding



(b) Assured Forwarding

Figure 5. Resulting share of bandwidth between Interior Routers with a neglected reservation

in Fig. 5 (b) shows that this can also happen with lower priority services like AF. When a reservation for flows of a lower priority service is neglected, other services (with even lower priority) can be reduced in their quality (in this case the best-effort service). As shown in the example, the service's aggregate causing the problem can itself be affected by packet losses (10% of the AF aggregate is discarded). Besides the described problems of case 2, case 1 will occur in the next border router which performs traffic metering and policing for flows of the service aggregate.

It must be noted, that directly applying RSVP to DiffServ would also result in an NRS Problem, because a receiver has to join the IP mc-group before sending a resource reservation request (RESV message) in order to receive the sender's PATH messages at first. Thus, the join for receiving PATH messages may cause an NRS Problem if this situation is not handled in a special way (like the solution in 4.1).

3.2 Dynamics of Arbitrary Sender Change

Because Differentiated Services are unidirectional by definition, we also consider the point-to-multipoint communication being of unidirectional nature. However, in traditional IP Multicast any node can send packets spontaneously and asynchronously to an mc-group, respectively to its multicast group address (therefore, IP Multicast offers a multipoint-to-multipoint service). For some applications it is an important feature that should also be available in case a specific service other than best-effort is used within the group.

Differentiated Services possess conceptually a simplex character. Therefore, for every mc-tree implied by a sender resources must be reserved separately if *simultaneous sending* should be possible with a better service. This is even true if shared multicast delivery trees are used (e.g., with PIM-SM or Core Based Trees). Unless single-source or extended multicast schemes (such as EXPRESS [7] or access control of Simple Multicast [9]) are used, there is no possibility to ensure in the IP layer that only one sender is transmitting at the same time. Otherwise, the NRS problem will occur again.

4 An Architecture for Providing Multicast in Differentiated Services Networks

The problems described in the previous sections are mainly caused by the simplicity of the DiffServ Architecture. Solutions have to be developed that do not introduce an additional amount of complexity that diminishes the scalability of this approach. In this paper, an architecture is suggested to provide simple solutions for the previously described problems.

4.1 Solution for the NRS Problem

Usage of resources which where not reserved before must be precluded. Basically, there are two possible solutions: first, existing mc-routing protocols have to be adapted, so that a join is only processed if resource availability was ensured before. Second, mc-routing protocols are not changed, but a join does not cause uncontrolled consumption of resources. In order to avoid fundamental changes in existing mc-routing protocols, the second solution is proposed.

In our example (which is depicted in Fig. 6), we want to consider the case when the join of a new receiver to a DS mc-group requires grafting of a whole new subtree to an already existing multicast delivery tree (step A). At first, the connecting node which joins both trees converts the codepoint (and therefore the Per-Hop Behavior) to a codepoint of a PHB which is similar to the default PHB in order to provide a best-effort-like service (no guarantees) for the new subtree (step B).

However, the re-marked packets should be separated from existing ordinary default PHB traffic in order to avoid unfairness being introduced. This unfairness would result from the fact that a high amount of re-marked packets is brought into the outgoing default BA at once. If the rate at which re-marked packets are inserted into the outgoing default aggregate is not reduced, those re-marked packets will probably cause discarding of other flow's packets in this BA if resources are scarce. Therefore, re-marked packets from this mc-group should be discarded more aggressively than other packets in this outgoing aggregate in order to limit the resources used by them. This could be accomplished by using a *Limited Effort (LE)* PHB (and a related DSCP) for those packets [5, 6]. Traffic within the Limited Effort BA should get a minimum service rate, but it is also limited in relation to the default (best-effort) BA. Thus, best-

effort traffic preempts Limited Effort traffic up to its minimum configured service rate, whereby Limited Effort traffic may utilize unused BE resources. Merely dropping packets more aggressively at the re-marking node is not sufficient, because there may be enough resources in the outgoing BA to transmit every re-marked packet and not requiring discarding any other packets within the same BA. Nonetheless, in the next node resources may be short for this particular BA. Therefore, those 'excess' packets should be identifiable at this next node.

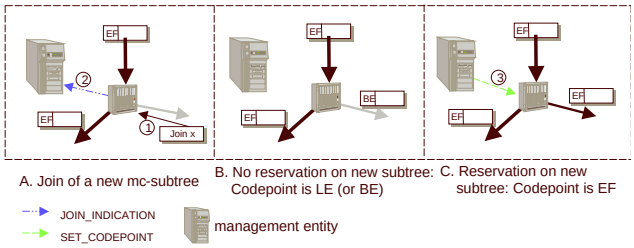


Figure 6. Sequence of the proposed solution

Consequently, a join of new receivers without prior reservation would only use specific resources in a controlled way and not cause an NRS Problem. A decoupling of joining an mc-group and the resource reservation process is achieved in this way. The better service will be only provided if a reservation request was processed by the management (e.g., Bandwidth Brokers) before. In case of a successful admission test, the re-marking node will be instructed by the management entity (*DS Domain Manager – DSDM*) to stop re-marking and to use the original codepoint again ③. Because reservation requests might also be initiated by the sender, an incoming JOIN request of a new receiver subtree should be also forwarded by a boundary node to the management node (indicated by the JOIN_INDICATION message in step A ② of Fig. 6), so that the re-marking node can be instructed (via the SET_CODEPOINT message in step C ③) to immediately use the same DSCP for replicated packets belonging to this group as for incoming packets (Fig. 6). Furthermore, besides indicating to use the original DSCP or not, the SET_CODEPOINT message may also be used to change the DSCP values in the mc-routing table. For this purpose, it carries the new DSCP value which should then be set in replicated packets.

The proposed solution does not require any additional classification of mc-groups within an aggregate. Because every mc-packet has to be processed by the multicast forwarding process, which looks up next hops in the mc-routing tables, the addition of an extra byte for containing the DSCP per output link (resp. child virtual interface) in each mc-routing table entry results in nearly no additional cost. Packets will be replicated according to the mc-routing process, so this is also the right place for setting appropriate DSCP values of replicated packets. Their DSCP values are not copied from the incoming original packet by default, but from the additional field in the mc-routing table entry of the corresponding output link. Initially, it contains the DSCP of the LE PHB. When a packet arrives with the default PHB, no re-marking is usually necessary, and, the outgoing replicates should also get the same codepoint in order to retain the behavior of todays common mc-groups using the default PHB. Otherwise, the specified DSCP from the corresponding mc-routing table entry is used in the replicated outgoing packets instead.

Moreover, the management entity (DSDM) must have detailed knowledge of the current multicast tree topologies, in order to make admission control decisions and to determine involved branching points. The ease of getting such information depends on the used routing protocols. In case of MOSPF, it would be easy due to its link-state nature, whereas for other routing protocols, such as PIM-DM/SM or DVMRP, changes of the tree structure have to be tracked (in case of PIM, routing topology information is known from unicast routing). As a possible solution, boundary nodes may forward mc-routing messages to the DSDM in order to inform it about a join or prune request of a subtree. To keep the complexity of interior nodes low, this task should be preferably handled by boundary routers. Additionally, a mechanism must be supplied for instructing a (branching) node to change its marking behavior and the DSCP value in the related mc-routing table entry (something like the SET_CODEPOINT message). This mechanism may be also incorporated into an existing mc-routing protocol as an extension. Alternatively, following a policy-based approach, it is possible to use the COPS protocol (cf. RFC 2748) for communicating change of mc-policies (an equivalent to exchange of SET_CODEPOINT and JOIN_INDICATION messages).

In summary, only those receivers will obtain a better service within a DiffServ mc-group, which actually reserved the according resources previously in the new subtree with assistance of the management. Otherwise, they get a quality that is similar to best-effort (or even less, however, the proposed LE PHB is not strictly required). This feature solves also the RSVP problem of having to join the mc-group in order to receive PATH messages first. If an already established mc-tree does not deliver a higher quality service before the DSDM receives a reservation request, the request is simply forwarded to the DSDM of the next upstream domain.

4.2 Solution for Arbitrary Sender Change

Every participant would have to initiate an explicit reservation if a receiver wants to make sure that it is possible to send with a value-added service to the group, regardless whether other senders already use the same service class si-

multaneously. This would require a separate reservation for each sender rooted mc-tree, which could be accomplished by the proposed architecture. Nevertheless, single-source multicast schemes with access control would be a great help in order to ensure that only one sender can transmit data at once. First-hop routers should therefore always classify multicast packets in dependence of the sender's address and mc-group address.

5 Proof of the Neglected Reservation Subtree Problem

In the following sections, it is showed that the NRS Problem actually exists and occurs in reality. Hence, we investigated the problem and its solution using the Linux-based implementation KIDS, which is described in an early version more detailed in [4]. Furthermore, we implemented the proposed solution for the NRS Problem using a standard 2.3 Linux kernel and modified the multicast routing table as well as the multicast routing behavior.

5.1 Implementation of the proposed solution

As described in section 4.1, the proposed solution for avoiding the NRS Problem is just adding one byte to the routing table entries in each Multicast router. In the Linux OS the multicast routing table is implemented by the *Multicast Forwarding Cache (MFC)*. The MFC is a hash table consisting of an `mfc_cache` entry for each combination of the following three parameters: sender's IP address, multicast group address and incoming interface. The routing information in an `mfc_cache` entry is kept in an array of TTLs for each virtual interface. When the TTL is zero, a packet matching to this `mfc_cache` entry will not be forwarded on this virtual interface. Otherwise, if the TTL is less than the packet's TTL, the latter will be forwarded on the interface with a decreased TTL.

In order to set an appropriate codepoint if bandwidth is allocated on an outgoing link, we added a second array of bytes for specifying the codepoint, that should be used on a virtual interface. The first six bits of the byte contain the DSCP that should be used and the seventh bit indicates, whether the original codepoint in the packets has to be changed to the specified one (=0) or has to be left unchanged (=1). The default entry of the codepoint byte is zero, so initially all packets will be re-marked to Best Effort.

Furthermore, we modified the multicast forwarding code for considering this information while replicating multicast packets. To change an `mfc_cache` entry we implemented a daemon for exchanging the control information (e.g. `JOIN_INDICATION`- and `SET_CODEPOINT`-messages) with a DSDM. Currently, the daemon uses a proprietary protocol, but it is planned to migrate to the COPS protocol.

5.2 Test Environment and Execution

In order to proof NRS Problem case 1, as described in the above section, a testbed as shown in Fig. 7 was built. It is a reduced version of the network shown in Fig. 4 and consists of two DS-capable routers, a first-hop router and an egress border router. The absence of interior routers does not have any effects to proof the described problem. The

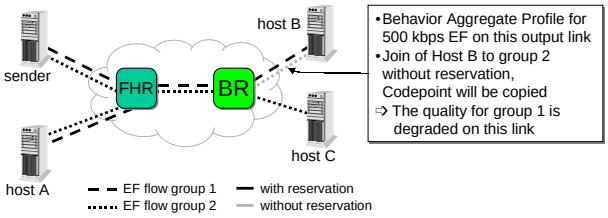
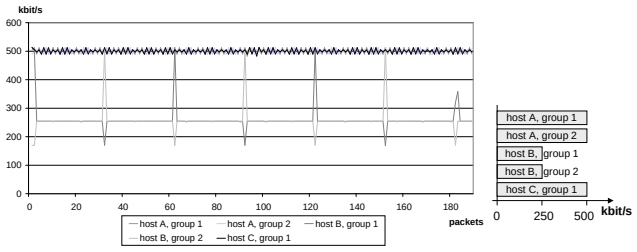


Figure 7. Evaluation of NRS Problem in Fig. 4

testbed comprises two Personal Computers used as DiffServ routers (Pentium III at 450 Mhz, 128 MB Ram, 3 network cards Intel eepro100), as well as one sender and three receiver systems (also PCs). KIDS has been installed on the routers and an *mrouted* was used to perform multicast routing. The network was completely built of separate 10BaseT Ethernet segments in full-duplex mode. There are no performance bottlenecks in the presented measurements on part of the software router, because even a PC at 200 Mhz has no problem to handle up to 10 Mbps DS traffic on each link [4].

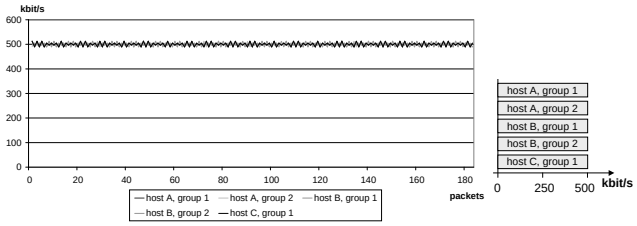
The sender generates two shaped UDP traffic flows of 500 kbps (packets of 1000 byte constant size) each and sends them to multicast group 1 (233.1.1.1) and 2 (233.2.2.2). In both measurements receiver A has a reservation along the path to the sender for each flow, receiver B has reserved for flow 1 and C for flow 2. Therefore, two static profiles are installed in the first-hop router with 500 kbps EF and a token bucket size of 10000 byte for each flow. In the egress border router one profile has been installed for the output link to host B and one related for the output link to host C. Each of them permits up to 500 kbps EF, but only the EF aggregate carried on the outgoing link is considered.

In measurement 1 hosts join to the groups as shown in Fig. 2. Those joins are using a reservation for the group towards the sender. Only the join of host B to group 2 has no admitted reservation. As described in section 3.1 this will cause the NRS Problem (case 1). Metering and policing mechanisms in the egress border router throttle down the EF aggregate to the reserved 500 kbps, whether individual flows have reserved or not. Fig. 8 shows the obtained



(a) Interarrival time (b) Average throughput

Figure 8. Results of measurement 1



(a) Interarrival time (b) Average throughput

Figure 9. Results of measurement 2

results. Hosts A and C received their flows without any interference. But host B received data from group 1 only with half of the reserved bandwidth, so one half of the packets has been discarded. Fig. 8 also shows that receiver B got the total amount of bandwidth for group 1 and 2, that is exactly the reserved 500 kbps. Flow 2 uses EF without actually having reserved any bandwidth and additionally violated the guarantee of group 1 on that link.

For measurement 2 the previously presented solution (cf. section 4) has been installed in the border router. Now it checks during duplication of packets, whether the codepoint has to be changed to (Limited) Best-Effort or whether it can be just copied. In this measurement it changed the codepoint for group 2 on the link to B to best-effort. Results of this measurement are presented in Fig. 9. Each host gets its flows with the reserved bandwidth and without any packet loss. Packets from group 2 are re-marked in the border router so that they are subsequently treated as best-effort traffic. In this case, they get the same bandwidth as the EF flow (500 kbps), because there is not enough other traffic on the link present (uncongested link), and thus, there is no need to discard packets. In case of congestion its bandwidth share will be decreased.

The above measurements confirm that the NRS Problem is to be taken seriously and that the presented solution will solve it.

6 Summary and Future Work

Global cluster computing infrastructures may profit from a quality of service oriented point-to-multipoint message passing service, which could be provided by using Differentiated Services mechanisms with IP Multicast. But aspects of providing DiffServ within multicast groups were not addressed in very much detail so far. In this paper, two fundamental multicast provisioning problems were identified: resource usage conflicts due to neglected reservations as well as support of different senders within a group. The proposed scalable and efficient solution uses an additional entry in the multicast routing table within DS nodes. It holds a DS field (merely one byte) in order to control the setting of DSCP values in replicated packets. To effectively control this, additional support from a separate resource management has to be provided, too. Moreover, the solution is widely decoupled from any particular multicast routing protocol.

In the future, an open solution for implementing the additional management functionality (e.g., a means for setting the DSCP value in the multicast routing table) has to be developed. Furthermore, we are currently exploring new DiffServ forwarding behaviors which support the quick and reliable forwarding of bursty traffic, that is typical for messages used in cluster computing scenarios.

References

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

[2] F. Baker, J. Heinanen, W. Weiss, and J. Wroclawski. Assured Forwarding PHB Group. RFC2597, June 1999.

[3] S. Blake et al. An Architecture for Differentiated Services. RFC 2475, Dec. 1998.

[4] R. Bless and K. Wehrle. Evaluation of Differentiated Services using an Implementation under Linux. In *Proceedings of IWQoS'99, London, June 1999*. IEEE Press, 1999.

[5] R. Bless and K. Wehrle. A Limited Effort Per-Hop Behavior. Internet-Draft – draft-bless-diffserv-le-phb-00.txt, Feb. 2001. Work in progress.

[6] B. Carpenter and K. Nichols. A Bulk Handling Per-Domain Behavior for Differentiated Services. Internet-Draft – draft-ietf-diffserv-bh-pdb-02.txt, Jan. 2001. Work in progress.

[7] H. W. Holbrook and D. R. Cheriton. IP Multicast Channels: EXPRESS Support for Large-scale Single-source Applications. In *Computer Communication Review*, volume 29. ACM SIGCOMM, Oct. 1999.

[8] V. Jacobson, K. Nichols, and K. Poduri. An Expedited Forwarding PHB. RFC 2598, June 1999.

[9] R. Perlman et al. Simple Multicast: A Design for Simple, Low-Overhead Multicast. Internet-Draft – draft-perlman-simple-multicast-03.txt, Oct. 1999. Work in progress.

[10] A. Roy et al. MPICH-GQ: Quality-of-Service for Message Passing Programs. In *Proceedings of the IEEE/ACM SC2000 Conference*. IEEE Press, Nov. 2000.