Performance of Hierarchical Aggregation in Differentiated Services Networks°

Revised version, February 2004

Susana Sargento and Rui Valadas

University of Aveiro/Institute of Telecommunications, 3810 Aveiro, Portugal,
susana@ua.pt, rv@det.ua.pt

Abstract - We address the use of hierarchical aggregation in DiffServ networks. We propose two analytical models to study the tradeoffs between signaling load and resource utilization. In the case of the signaling load, we introduce a novel performance metric that captures, simultaneously, the state information stored and the rate of signaling messages processed at routers. In the first analytical model, based on multidimensional birth-death processes, the offered load is detailed at the flow level, which allows accurate assessment of the signaling load. The second analytical model accommodates time-varying offered loads, which allows studying the tradeoffs between the time-scale of the aggregate demand and the time-scale of signaling. Our results, which also include analysis using measured traces, show that hierarchical aggregation can introduce very high signaling gains with a small penalty in terms of resource utilization, allowing significant savings in terms of network cost.

1 Introduction

In the IntServ architecture [1] resources are reserved for individual flows, i.e., on a per-flow basis, using the RSVP protocol. This implies that every time a new flow requests admission in the network, there must be signaling messages exchanged between the various network elements (hosts and routers) in the flow's path; moreover a state for each flow needs to be maintained at all routers along the flow's path. Both these factors contribute to the lack of scalability attributed to the IntServ architecture.

The reservation of resources for aggregates of flows (instead of individual flows) has been proposed in the context of DiffServ architecture [2,3,4,5,6], as a means of reducing significantly the signaling load and the state information stored at routers, while still providing the same QoS for real time flows. To support aggregation, an extension to RSVP that allows RSVP signaling messages to be hidden inside an aggregate, was recently defined in [3]. This actually leads to a new architecture that reuses most features of the DiffServ architecture, but includes some enhancements in order to match the strong service model of IntServ. This architecture was already considered for use in an IP-based access network with QoS support [11].

° Submission to the Telecommunication Systems Journal

In the simplest case, all edge routers reserve bandwidth end-to-end, i.e., between ingress and egress routers of a network domain; this reservation can be updated in bulks much larger than the individual flow's bandwidth. Whenever a flow requests admission at an ingress router, the router checks if there is enough bandwidth to accept the flow on the (end-to-end) aggregate leading to the egress router. If resources are available, the flow will be accepted, without any need for signaling the core routers. Otherwise, the core routers will be signaled in an attempt to increase the aggregate's bandwidth. If this attempt succeeds, the flow will be admitted; otherwise, it will be rejected. Thus, with aggregation, signaling messages are only exchanged when the aggregate's bandwidth needs to be updated. The efficiency of aggregation depends heavily on the matching between the aggregate reservation and the aggregate demand. If the bulk size is too large, the signaling load will be minimal but either reserved resources will be under-utilized or there will be unnecessarily blocked flows. Otherwise, with a too small bulk size the signaling load may approach that of per-flow signaling.

Within a large network domain, the need to set-up a lot of end-to-end aggregates can lead to poor resource utilization. One way to alleviate this problem is to partition the domain in areas and to have end-to-end aggregates between the area border routers. The resource utilization can be increased since an area aggregate can now be shared by flows coming from different domain edge routers. However, the signaling load also increases, since the area border routers need to be signaled on a per-flow basis. In fact, every time a new flow arrives at a domain edge router, there is the need to check if there are sufficient resources in every aggregate that the flow traverses within the domain. We recall that the facility for partitioning a domain into areas is already included in several routing protocols, e.g. OSPF and ISIS, and also in MPLS, again motivated by scalability reasons.

An extension of the RSVP protocol [3] has been developed to handle the signaling in networks using flow aggregation (with either end-to-end or area aggregates). This much facilitates the interworking with IntServ networks (where end-to-end reservations are performed via RSVP). The (basic) RSVP protocol is used between the edge routers of an aggregation region to perform admission control and to upgrade the bandwidth of the aggregates. Whenever there is the need to install, remove or update the bandwidth of an aggregate, the ingress router sends a RSVP Path message towards the egress router, where it declares the aggregate bandwidth required for reservation; the egress router responds with a RSVP Resv message. These messages will be processed by all core routers, which can accept or deny the request based on the available resources. Note that the set-up of a new aggregate or the attempt to increase an existing aggregate's bandwidth is triggered by the arrival, at the ingress router, of a RSVP Path message sent by an end-user requesting a flow's admission. The (extended) RSVP protocol handles the transport of end-user RSVP messages inside aggregation regions. If, upon the arrival of an end-user RSVP Path message at the ingress router, there is sufficient bandwidth to accept the flow, the router changes the IP protocol field of the Path message from RSVP to the newly defined RSVP-E2E-IGNORE protocol type. The RSVP Path message will then be transparently sent inside the aggregation region without any processing in the core routers. When the egress router of the aggregation region

receives this message, restores the IP protocol field back to RSVP and forwards the message towards the receiver. The corresponding RSVP Reservation message will be hidden from the core routers in the same way.

The architecture described above requires the support of RSVP signaling and state maintenance in all network elements, as in the IntServ architecture. However, the signaling load decreases significantly because the cores routers are only signaled when there is an attempt to update the aggregate's bandwidth; only the edge routers need to process signaling on a per-flow basis. Moreover, the memory requirements of core routers are much lower since state maintenance is performed on an aggregate basis (and not on a per-flow basis as in IntServ). Therefore, the proposed solution, despite requiring some enhancements to the (basic) DiffServ architecture, is able to support the same QoS for real time flows achieved by the IntServ architecture. Moreover, it constitutes a framework that allows the definition of different trade-offs between signaling load and resource utilization.

In this paper we analyze the tradeoffs between signaling load and resource utilization in a network domain that can be partitioned in areas. For this purpose we develop two different analytical models. In the first model, called per-flow load model, the offered load is detailed at the flow level. It assumes flow arrivals according to a Poisson process and exponentially distributed flow durations, such that a multi-dimensional birth-death process can describe the number of flows in a domain. Note that, although a Poisson model may not be appropriate for packet level traffic, it is widely used for flow level traffic, given that flows are usually generated by a large number of independent users. The second analytical model, called aggregate load model, accommodates an offered load whose average aggregate bandwidth is time-varying. In particular, we will assume a sinusoidal variation. In this model, the offered load is not detailed at the flow level. In addition, we assess the impact of a measured data via discrete-event simulation.

This paper is organized as follows. Section 2 presents the system model. Sections 3 and 4 present the two analytical models: the per-flow load model and the aggregate load model. Section 5 discusses the results obtained with the analytical models, and the ones obtained with discrete-event simulations with a measured trace. Finally, section 6 concludes the paper.

Our contribution is the following. First we develop a per-flow load model that allows detailed characterization of the signaling load and the resource utilization in hierarchical network domains using aggregation. In particular, we introduce a signaling metric that is well adapted to the case of flow aggregates. Then, we develop an aggregate load model that permits studying the tradeoffs between the time-scale of the aggregate demand and the time-scale of signaling. Third, we carry out numerical studies that clearly show the advantages of structuring network domains into areas.

2 System Model

Consider a network domain partitioned in areas (Fig. 1). We refer to the routers in the edge of the domain as Domain Border Routers (DBRs), and the routers in the edge of each area as Area Border Routers (ABRs). DBRs also play the role of ABRs.

Sessions of packet flows are offered between DBRs and can be aggregated in pipes of reserved bandwidth called aggregates. We consider two cases, where the bandwidth is reserved end-to-end between DBRs (called end-to-end aggregates) or reserved end-to-end between ABRs (called area aggregates). In the first case, sessions traverse a single aggregate (between DBRs) whereas, in the second case, they traverse a concatenation of area aggregates. Each aggregate (end-to-end or area) will have ingress and egress routers and will (possibly) traverse several other routers, called internal routers. We assume that the aggregate's bandwidth can be adjusted over time through appropriate signaling. The ingress router of each aggregate will process signaling messages on a per-flow basis, whereas the internal routers will only process signaling messages when the aggregate's bandwidth is to be updated. Aggregates traverse one or more areas; in each area, they travel through an ABR pair (an ingress and an egress ABR). Let $\mathcal{J} = \{1, 2, ..., J\}$ be the set of ABR pairs in the domain. We consider that each ABR pair has a bottleneck capacity C_j, which corresponds to the lowest capacity among the links that belong to the route between the two ABRs. While the bandwidth of an aggregate can vary over time, it is limited by the bottleneck capacities of the ABR pairs it traverses. Let $\mathcal{H} = \{1, 2, ..., H\}$ be the set of aggregates. Aggregates are defined by (i) their bandwidth $r_h(t)$, (ii) their origin and destination ABRs, (iii) their route $\mathcal{R}_h \subseteq \mathcal{J}$ described by the ABR pairs they traverse and (iv) the number of internal routers m_h (i.e. not including ingress and egress routers) traversed by the aggregate.

As mentioned before, the traffic is offered between ingress and egress DBRs. Let $\mathcal{K} = \{1, 2, ..., K\}$ be the set of sessions. We consider two offered traffic models. In the first one, called per-flow load model, a session k is characterized by (i) ingress and egress DBRs, (ii) route $\mathcal{H}_k \subseteq \mathcal{H}$ described in terms of the aggregates it traverses, (iii) the bandwidth of each packet flow b_k, (iv) and the traffic intensity $\rho_k = \lambda_k/\mu_k$. Specifically, to allow Markov modeling, we assume that packet flows arrive according to a Poisson process with rate λ_k and have exponentially distributed durations with mean $1/\mu_k$. We also consider that the bandwidth of an aggregate can be adjusted in steps of a bulk bandwidth, which we denote by q_h. The second model, called aggregate load model, considers a time-varying aggregated offered load. Session k is simply characterized by (i) ingress and egress DBRs, (ii) route $\mathcal{A}_k \subseteq \mathcal{J}$ described in terms of the ABR pairs it traverses and (iii) bandwidth $r_k(t)$. In this case, a session is meant to represent an aggregate of flows. We consider that the bandwidth of an aggregate is adjusted at the beginning of fixed time intervals, matching exactly the requirement for the interval (i.e. the maximum $r_k(t)$ over the interval).

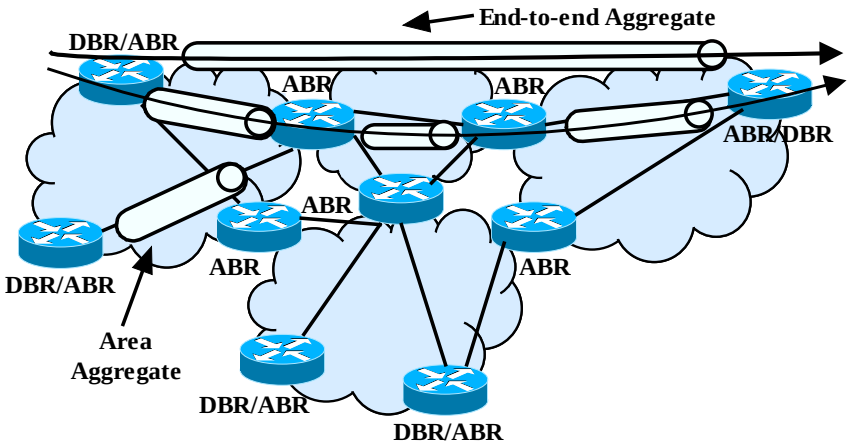

Fig. 1 System model.

We further define $\mathcal{K}_h \subseteq \mathcal{K}$ as the set of sessions that traverse aggregate h, and $\mathcal{H}_j = \{1, 2, ..., H_j\} \subseteq \mathcal{H}$ as the set of aggregates that traverse ABR pair j. Note also that different aggregates can be used for different service classes.

3 Per-Flow Load Model

In this section, we develop a continuous-time Markov process (more specifically, a multi-dimensional birth-death process) that characterizes the system state under the assumption of flow arrivals according to a Poisson process and exponentially distributed flow durations. The system state is characterized by vector $n = (n_1, n_2, ..., n_K)$, where n_k represents the number of flows of session k in the system. New flows requesting admission in the domain can be accepted if there is enough bandwidth in each of the aggregates they traverse; they can also be accepted if the bandwidth in all aggregates that do not observe previous condition can be increased to accommodate the flows.

We consider as a metric for assessing the signaling overhead and the amount of state information, the (total) rate of signaling messages processed by all routers in a domain. The signaling messages correspond to attempts of updating the reservation state at a router. In particular, a signaling message may attempt installing (or uninstalling) a flow or aggregate, or may attempt increasing (or decreasing) the bandwidth of an aggregate. With this metric we capture not only the number of reservations at routers but also the frequency of their updates, which is an important factor in terms of router cost. For example, the metric considered in [4] was the average number of flows in a domain. This only captures the amount of state information, which is clearly insufficient, especially when dealing with flow aggregates.

In a domain without aggregates (i.e., using only per-flow reservations), upon a flow arrival all routers in the flow's path will process a signaling message. In a domain with aggregates, signaling messages will always be

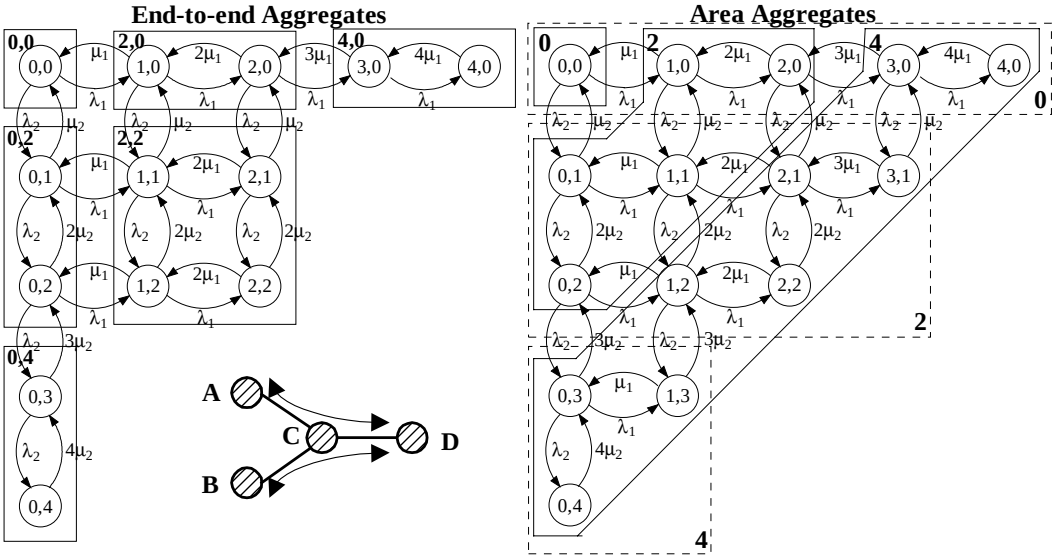

Fig. 2 Domain with 3 areas and respective state diagram with end-to-end and area aggregates.

processed by the ingress router of each aggregate, but the internal routers of the aggregates only process signaling messages if there is an attempt of updating the bandwidth of the aggregate. Note that, in the case of a session traversing multiple aggregates, as in the case of a domain with areas, a flow arrival may provoke attempts of bandwidth updates in more than one aggregate.

Consider the simple example of Fig. 2, that corresponds to a domain partitioned in 3 areas. There are two sessions, one offered between DBRs A and D and the other between DBRs B and D. Fig. 2 represents the Markov chains for the cases of end-to-end aggregates and area aggregates. In the first case, there are 2 end-to-end aggregates; in the second one, there are 3 area aggregates. We consider that the flow's bandwidth of both sessions is $b_1 = b_2 = 1$ unit, the bulk size is $q_h = 2$ units in all aggregates and the bottleneck capacity of all ABR pairs is $C_j = 4$ units.

In Fig. 2, states are grouped according to the bandwidth of the corresponding aggregates; each group is enclosed in a polygon and its respective bandwidth is indicated in one of the polygon's corners. In the case of area aggregates we show polygons for areas CD (fixed line) and BC (dotted line). For example, in the case of end-to-end aggregates, in state (2,1) both aggregates have 2 units of reserved bandwidth; the first aggregate from A to D is utilized at 100% and the second one, from B to D, at 50%. States (3,1) and (1,3) are only allowed in area aggregates, since there is a single aggregate in area CD shared by both sessions, whose bandwidth can grow up to the limit of 4 units; this illustrates the higher utilizations that can be achieved with this type of aggregation. Signaling messages attempting to update an aggregate's bandwidth are driven by transitions between states belonging to different polygons. Take the example of the state (1,1) and area aggregates. Transitions to either state (2,1) or state (1,2), drive signaling messages in the aggregate of area CD, but not in

the aggregates of other areas. However, the transition from state (0,2) to state (0,3) drive signaling messages both in area CD and in area AC (or BC).

In the general case, the state space of the Markov chain, is defined by:

$$\mathcal{S} = \left\{ n \in \mathcal{I}^K : \sum_{h \in \mathcal{H}_j} \left\lceil \sum_{k \in \mathcal{K}_h} n_k b_k \right\rceil_h \le C_j,\ j = 1,..., J \right\} \tag{1}$$

where $\mathcal{I}$ is the set of non-negative integers and $\lceil x \rceil_h$ is the lowest multiple of q_h higher than x. The inner sum in the state space definition corresponds to the bandwidth reserved for each aggregate, which is always a multiple of the bulk bandwidth q_h. The outer sum corresponds to the overall bandwidth, reserved for all aggregates, in ABR pair $j \in \mathcal{J}$.

From the state space the limiting state probabilities can be easily calculated through standard techniques. We denote by π_n the limiting probability of state n. The reserved resource utilization, i.e., the percentage of the reserved bandwidth that is utilized by the admitted traffic in all ABR pairs is given by

$$U = \frac{1}{J} \sum_{n \in \mathcal{S}} \left[\sum_{j \in \mathcal{J}} \frac{\sum_{h \in \mathcal{H}_j} \left(\sum_{k \in \mathcal{K}_h} n_k b_k \right)}{\sum_{h \in \mathcal{H}_j} \left\lceil \sum_{k \in \mathcal{K}_h} n_k b_k \right\rceil_h} \right] \pi_n \tag{2}$$

To model the signaling load, let n_k^+ be a state (possibly not belonging to $\mathcal{S}$) reached from $n \in \mathcal{S}$ through increasing the number of session's k flows by one unit, i.e., $n_k^+ = (n_1, ..., n_k+1, ..., n_K)$, and let $\left(n, n_k^+\right)$ represent an (upward) transition (possibly not allowed within the state space) from state n to state n_k^+. There are two types of upwards transitions: allowed and forbidden within the state space. The sets of allowed transitions, $\mathcal{A}$, and forbidden transitions, $\mathcal{F}$, are defined by

$$\mathcal{A} = \left\{ \left(n, n_k^+\right) : n, n_k^+ \in \mathcal{S} \right\} \tag{3}$$

$$\mathcal{F} = \left\{ \left(n, n_k^+\right) : n \in \mathcal{S}, n_k^+ \notin \mathcal{S} \right\} \tag{4}$$

In order to describe an aggregate's bandwidth adjustment we introduce the indicator function $I_h^{\left(n, n_k^+\right)}$ that equals 1 whenever there is an allowed or forbidden transition $\left(n, n_k^+\right)$ driven by an arrival of a session's k flow that can no longer be accommodated in the bandwidth currently reserved for the aggregate h it traverses, i.e.,

$$I_h^{(n,n_k^+)} = \begin{cases} 1, & \sum_{k \in \mathcal{K}_h} n_k b_k + b_k > \left\lceil \sum_{k \in \mathcal{K}_h} n_k b_k \right\rceil_h \\ 0, & otherwise \end{cases} \quad (5)$$

If $I_h^{(n,n_k^+)} = 1$ for $(n, n_k^+) \in \mathcal{A}$, i.e., for an allowed transition, there will be a successful reservation update; otherwise, if the transition is forbidden, the reservation update fails. In both cases, signaling messages will be processed at all routers of the aggregate. The number of routers of an aggregate *h* that suffer an attempt of reservation update upon an (n, n_k^+) transition is $1+m_h I_h^{(n,n_k^+)}$ (we do not consider the egress router of the aggregate). A transition in the opposite direction provokes reservation updates in the same number of routers.

We define γ_A^B as the rate of signaling messages, where *A* indicates whether only the internal routers are considered (*I*) or all domain routers (*D*) and *B* indicates whether we are considering all reservation attempts (*T*) or only successful reservation attempts (*S*). Using these definitions:

$$\gamma_I^T = \sum_{(n,n_k^+) \in \mathcal{A}} \left(\lambda_k \pi_n + (n_k + 1)\mu_{k+1} \pi_{n_k^+}\right) \sum_{h \in \mathcal{H}_k} m_h I_h^{(n,n_k^+)} + \sum_{(n,n_k^+) \in \mathcal{F}} (\lambda_k \pi_n) \sum_{h \in \mathcal{H}_k} m_h I_h^{(n,n_k^+)} \quad (6)$$

$$\gamma_D^T = \sum_{(n,n_k^+) \in \mathcal{A}} \left(\lambda_k \pi_n + (n_k + 1)\mu_{k+1} \pi_{n_k^+}\right) \sum_{h \in \mathcal{H}_k} \left(m_h I_h^{(n,n_k^+)} + 1\right) + \sum_{(n,n_k^+) \in \mathcal{F}} (\lambda_k \pi_n) \sum_{h \in \mathcal{H}_k} \left(m_h I_h^{(n,n_k^+)} + 1\right) \quad (7)$$

where the first term in each equation represents the rate of successful signaling messages, i.e., γ_I^S and γ_D^S, respectively. Note that, whenever $I_h^{(n,n_k^+)} = 0$, there are signaling messages only in the ingress router of the aggregate; when $I_h^{(n,n_k^+)} = 1$, all routers in the aggregate's path process signaling messages. Our signaling studies take as a reference the rate of signaling messages in per-flow reservations (IntServ), which we denote by $\gamma_{A,ref}^B$. In this case, each time a new flow arrives, all routers in the flow's path process signaling messages. Thus:

$$\gamma_{I,ref}^T = \sum_{(n,n_k^+) \in \mathcal{A}} \left(\lambda_k \pi_n + (n_k + 1)\mu_{k+1} \pi_{n_k^+}\right) \sum_{h \in \mathcal{H}_k} m_h + \sum_{(n,n_k^+) \in \mathcal{F}} (\lambda_k \pi_n) \sum_{h \in \mathcal{H}_k} m_h \quad (8)$$

$$\gamma_{D,ref}^T = \sum_{(n,n_k^+) \in \mathcal{A}} \left(\lambda_k \pi_n + (n_k + 1)\mu_{k+1} \pi_{n_k^+}\right) \sum_{h \in \mathcal{H}_k} (m_h + 1) + \sum_{(n,n_k^+) \in \mathcal{F}} (\lambda_k \pi_n) \sum_{h \in \mathcal{H}_k} (m_h + 1) \quad (9)$$

where again the first term in each equation represents the rate of successful signaling messages, i.e., $\gamma_{I,ref}^S$ and $\gamma_{D,ref}^S$, respectively. Note that $\sum_{h \in \mathcal{H}_k} (m_h + 1)$ represents the total number of routers in flow's *k* path. Note

also that the signaling in per-flow reservations is a particular case of the signaling with aggregation, when there is one end-to-end aggregate per session and a bulk size equal to the session's flow bandwidth. Finally, we define a signaling gain as the ratio between the signaling rate with per-flow reservations and the one with aggregation, i.e., $G_A^B = \gamma_{A,ref}^B / \gamma_A^B$.

4 Aggregate Load Model

In this section we present a model that considers a time-varying offered load, which is an extension to multiple areas of the model described in [6]. In particular, we consider that the aggregate traffic of session k is characterized by a sinusoid with random phase

$$r_k(t) = d_k + e_k \cos\left(\frac{2\pi}{T}t + \theta_k\right) \tag{10}$$

where d_k is the mean bandwidth of the aggregate, e_k is the amplitude of the sinusoid, T is the sinusoid period, and θ_k is the random phase uniformely distributed in $[0,2\pi]$. This model is motivated by the behavior of a large number of observed traces of traffic aggregates that exhibit a near-deterministic periodic long-term trend. Consider, for example, the trace of Fig. 3, which corresponds to traffic observed at the ingress router of Qbone "PSC" [7]. The period T corresponds to 24 hours.

As mentioned before, we also assume that (i) the reservation of aggregate h is updated at the beginning of fixed time intervals of duration τ, and that (ii) the amount of bandwidth to be reserved in the beginning of time interval matches exactly the requirement for that time interval. The desired bandwidth reservation, at time interval x_h = 1, 2, ..., T/τ, in aggregate h, corresponds to the maximum offered bandwidth calculated over this interval, i.e.,

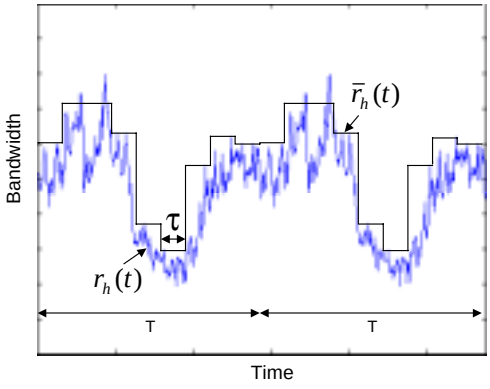

Fig. 3 Traffic observed at the ingress router of Qbone "PSC".

$$r_{h,x_h} = \max_{(x_h-1)\tau \le t < x_h\tau} \sum_{k\in\mathcal{K}_h} r_k(t) \tag{11}$$

Note that at the beginning of time interval x_h the bandwidth of the aggregate (i.e. the bandwidth effectively reserved) may or may not be adjusted to r_{h,x_h}, depending on the bandwidth available at the ABR pairs traversed by aggregate *h*. This is illustrated in Fig. 3 where $r_h(t)$ and $\bar{r}_h(t)$ represent the offered load and the bandwidth effectively reserved to the aggregate, respectively. To avoid trivialities we assume that T/τ is an integer.

We define the probability of overload of session *k*, P_k, as the fraction of bandwidth of session *k* that cannot be admitted. To calculate this metric we consider a reduced load approximation [9], where the traffic offered to an ABR pair is reduced according to the overload suffered by the sessions using that ABR pair in the other ABR pairs traversed by those sessions. Denoting the probability of overload at ABR pair *j* as L_j, the reduced load offered to ABR pair *j* at time interval x_h is

$$\hat{r}_j(t) = \sum_{h\in\mathcal{H}_j} \sum_{k\in\mathcal{K}_h} r_k(t) \prod_{l\in\mathcal{A}_k-\{j\}} (1-L_l) \tag{12}$$

and the reduced desired bandwidth reservation is

$$\hat{r}_{j,x_h} = \sum_{h\in\mathcal{H}_j} \hat{r}_{h,x_h} = \sum_{h\in\mathcal{H}_j} \max_{(x_h-1)\tau \le t < x_h\tau} \left(\sum_{k\in\mathcal{K}_h} r_k(t) \prod_{l\in\mathcal{A}_k-\{j\}} (1-L_l) \right) \tag{13}$$

The overload probability at ABR pair *j* is the ratio of the overload bandwidth (bandwidth that cannot be reserved, calculated over the set of aggregates of ABR pair *j*) and the desired bandwidth reservation. This probability is approximated by

$$L_j \approx \frac{\left(\frac{\tau}{T}\right)^{H_j} \left[\sum_{x_1=1}^{T/\tau} \cdots \sum_{x_{H_j}=1}^{T/\tau} \left(\sum_{h\in\mathcal{H}_j} \hat{r}_{h,x_h} - C_j \right)^+ \right]}{\frac{\tau}{T} \left(\sum_{h\in\mathcal{H}_j} \sum_{x_h=1}^{T/\tau} \hat{r}_{h,x_h} \right)}, \; j = 1,2,\ldots,J \tag{14}$$

The numerator corresponds to the mean overload bandwidth in ABR pair *j*. The summations in the numerator perform all possible combinations of relative phases between aggregates and $(\tau/T)^{H_j}$ is the probability of each combination. The denominator represents the mean desired bandwidth reservation, calculated over the set of aggregates in the ABR pair *j*. The detailed derivation of this result is presented in [8]. The set of *J* non-linear equations with *J* unknowns in (11) can be solved using the method of repeated substitutions [9]. The probability of overload of session *k* is then given by

$$P_k \approx 1 - \prod_{j \in \mathcal{A}_k} \left(1 - L_j\right) \tag{15}$$

The reserved resource utilization is defined as the ratio of average offered load to average reserved bandwidth:

$$U \approx \frac{1}{J} \sum_{j \in \mathcal{J}} \frac{\left(1 - L_j\right) E\left(\sum_{h \in \mathcal{H}_j} \sum_{k \in \mathcal{K}_h} r_k(t) \prod_{l \in \mathcal{A}_k - \{j\}} \left(1 - L_l\right) \right)}{\left(\frac{\tau}{T}\right)^{H_j} \left[\sum_{x_1=1}^{T/\tau} \cdots \sum_{x_{H_j}=1}^{T/\tau} \min\left(\hat{r}_{j,x_h}, C_j\right) \right]} \tag{16}$$

where $E(x)$ represents the expected value of x.

5 Numerical Investigations

In this section we present numerical examples and simulations to study the tradeoffs between signaling load and utilization. First, we consider a domain with the Dumbbell topology depicted in Fig. 4, with two peripheral areas in each side of the domain and a central area. Later, we will consider both a core network and an access network. In the experiments that consider the domain depicted in Fig. 4, there are 4 sessions, ACDE, ACDF, BCDF and BCDE, all traversing the central area, denoted by r_{xy}, where x is the origin and y is the destination. The number of routers inside each area is 4. Thus each session travels through 15 routers (not including the domain egress router). Except the case of real aggregate simulations, the bandwidth of each ABR pair is 32 Mb/sec in all areas. We compare two types of network domains: (i) domains with only end-to-end aggregates and (ii) domains with only area aggregates.

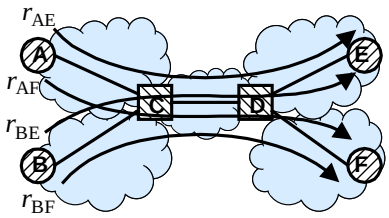

Fig. 4 Dumbbell topology.

In the figures presented bellow, we denote end-to-end aggregation by “End-to-end” and area aggregation by “Area”. In area aggregation we consider domains with one service class, denoted by “1 class”, and two service classes, denoted by “2 classes”. In the case of two service classes, we consider that sessions r_{AE} and r_{AF} belong to one class and sessions r_{BE} and r_{BF} belong to another, leading to one aggregate in the left peripheral areas

and two aggregates in all other areas. We will consider two cases regarding the bulk size or τ/T, the normalized time interval: the same bulk size (or τ/T) for all aggregates in the domain, called fixed bulk size (or fixed τ/T), and bulk size (or τ/T) proportional to each aggregates' offered load, denoted by "Prop. bulk" (or "Prop. τ/T"). In the later case, the *xx*-axis will represent the bulk size (or τ/T) of peripheral areas. We will consider as the metric for assessing the signaling load, the signaling gain of end-to-end and area aggregation over per-flow signaling.

5.1 Per-Flow Load Model

In these studies we assume that sessions are characterized by b_k = 1 Mb/sec, λ_k = 8 sec^{-1} and $1/\mu_k$ = 1 sec. Thus the session's average offered bandwidth is 8 Mb/sec.

Fig. 5 (a) shows the signaling gains with both end-to-end aggregation and area aggregation, considering all routers in the domain. We present signaling gains considering (i) all reservation attempts (solid lines, denoted by "All") and (ii) only successful reservations (dashed lines, denoted by "Successful"). When the bulk size equals the flows' bandwidth the signaling rate is the same as in per-flow signaling (so a unitary gain is obtained). Results show that the gains over per-flow signaling increase with the bulk size. For a bulk size of 8 Mb/sec the gains considering only successful reservations are approximately 15 with end-to-end aggregation and 4 with area aggregation. These gains will approach 15 and 5, respectively, which are obtained in the limit, as the bulk size increases, when no signaling takes place in internal routers. Note that the number of routers that process signaling messages on a per-flow basis (i.e., every time a flow arrives or departs), is 1 with end-to-end aggregation (the ingress DBR), 3 with area aggregation (the ingress ABRs of each area) and 15 without aggregation, which explains the referred gains. The gains in end-to-end aggregation almost reach the limit when the bulk size is 8 Mb/sec. Given that in the central area there are 4 aggregates, each with an average offered load of 8 Mb/sec, and that the bandwidth of the ABR pair is 32 Mb/sec, the system trend is to have the bandwidth of all aggregates adjusted to the bulk size at all times. This leads to very few bandwidth update attempts and, therefore, very few signaling messages in internal routers. The gains obtained with the three cases of area aggregation are very similar, the one in the case of one service class and proportional bulks being slightly larger (in this case, the bulk size in the central area is twice the bulk size in the peripheral areas).

Consider now the difference between the signaling gains of (i) all reservation attempts and (ii) successful reservations. For a 8 Mb/sec bulk size, with end-to-end aggregation, the signaling gain decreases from 15 to 8.5, reflecting the significant number of reservation requests that cannot be established, i.e., a relatively high blocking probability. With area aggregation, the signaling gains are almost the same, showing that almost all reservation attempts turn into successful reservations.

Fig. 5 (b) considers the signaling gains achieved by internal routers, i.e., not including DBRs, in the case of end-to-end aggregation, and not including DBRs and ABRs, in the case of area aggregation. This metric is very important, as it represents the gains that can be achieved by the routers that are supposed to have the lowest cost. As seen before, the gains corresponding to all domain routers are biased by the number of routers that need to perform per-flow signaling and router costs are most certainly not a linear function of the signaling load. We first notice that the signaling gains of internal routers are higher than the domain ones. Considering only successful reservations and fixed bulk sizes, we observe that the signaling gains in area aggregation with one and two service classes are similar (recall that with one service class there is only one aggregate in the central area, and with two service classes there are two). There are two opposite effects. First, with one service class the aggregates are always shared by more than one session; the overall traffic inside an aggregate becomes smoother, which contributes to reducing the signaling rate. On the other hand, the resource sharing in area aggregation with one service class increases the number of admitted flows, which contributes to increas-

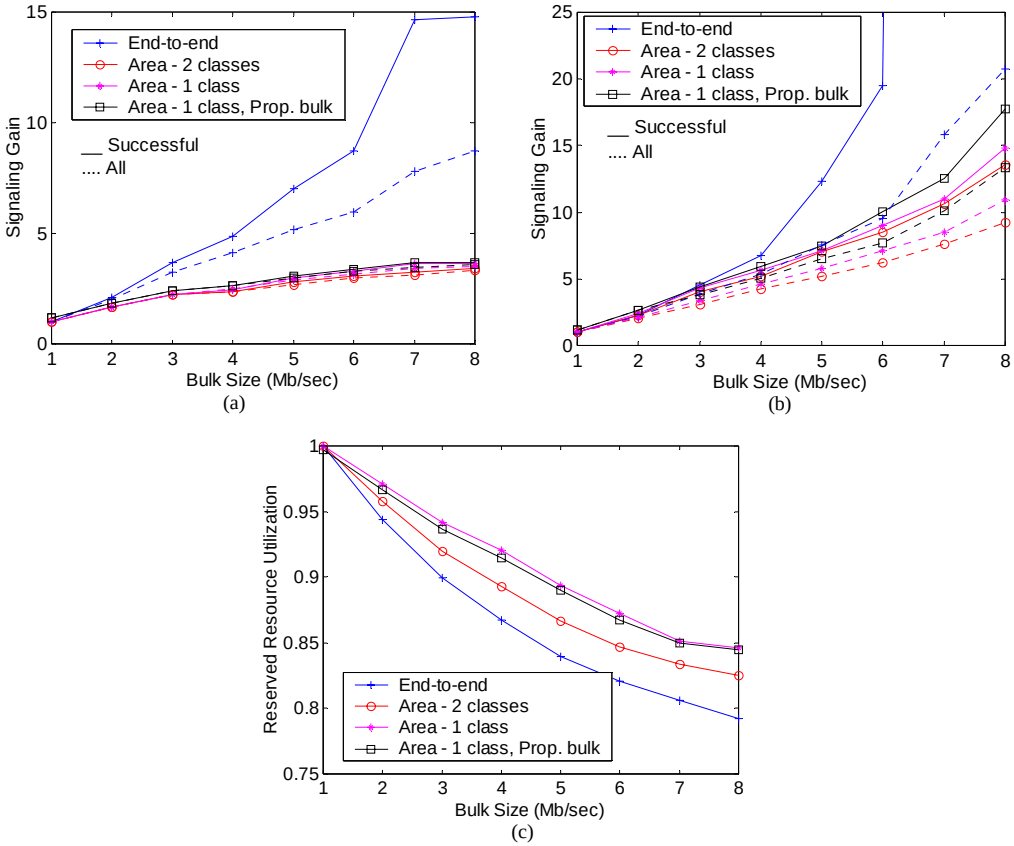

Fig. 5 Signaling gains of (a) all routers in the domain, (b) internal routers and (c) reserved resource utilization (per-flow model).

ing the signaling rate. These two opposite effects balance out, which explains the similarities in terms of gains. Considering proportional bulk sizes, we notice that the signaling gains increase. This is due to the larger bulk size of the central area, which provokes a reduction in its signaling load. The signaling gains in end-to-end aggregation are higher for bulk sizes larger than approximately 4 Mb/sec. This is again explained by the system trend, to have the bandwidth of all aggregates adjusted to the bulk size at all times, for large bulk size values.

Fig. 5 (c) depicts the reserved resource utilization. The resource utilization increases in area aggregation because the resource sharing is larger, compared with end-to-end aggregation. In area aggregation with one service class, the four sessions share the same aggregate in the central area. Consider the cases of area aggregation with one service class, with fixed and proportional bulk sizes. The utilization achieved with proportional bulk size is slightly smaller, but quite close to the one obtained with fixed bulk size. This reflects the good tradeoff between signaling and utilization that can be achieved in this case. As an example, to achieve 84% of resource utilization, the bulk size must be lower than 5 Mb/sec in end-to-end aggregation, 6 Mb/sec in area aggregation with two service classes, and 8 Mb/sec in area aggregation with one service class and with proportional and fixed bulk sizes.

5.2 Aggregate Load Model

In these experiments the mean bandwidth of the cosine wave is d_k = 5.3 Mb/sec and the amplitude is also e_k = 5.3 Mb/sec, so as to reproduce an overload situation.

Fig. 6 depicts the reserved resource utilization and the overload probability in the domain. The results concerning the case of area aggregation with one service class and a proportional τ/T are presented only for $\tau/T \leq 1/2$. Note that the τ/T values of the central area are twice the ones of the peripheral area, and the *xx*-axis is representing those of peripheral areas.

As τ/T increases, the frequency of reservation updates decreases, leading to lower resource utilization and higher overload probabilities.

For small τ/T, the reserved resource utilization and overload probabilities obtained with all types of aggregation are approximately the same. We recall that with end-to-end aggregation there are four aggregates in the central area (one aggregate per session) and with area aggregation and one service class there is only one aggregate (whose resources are shared by all four sessions). With end-to-end aggregation the four sessions will only share resources when their aggregates do so, while with area aggregation the four sessions will always share resources irrespective of τ/T. For small τ/T, the aggregates are adjusted frequently (in relation to the time-scale of the offered load), and significant resource sharing takes place in both types of aggregation. This is done at the cost of a high signaling rate. As τ/T increases, and the signaling rate decreases, the reservations

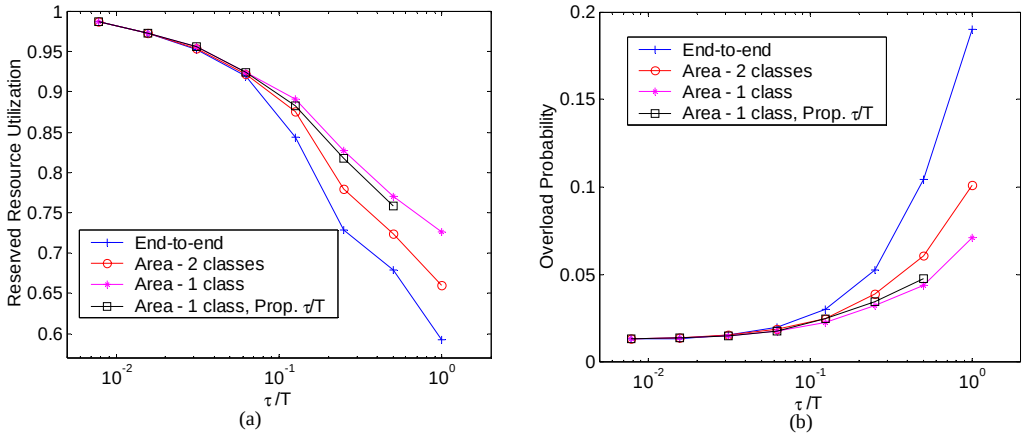

Fig. 6 (a) Overload probability and (b) reserved resource utilization (aggregate model).

are made for longer time intervals, leading to a decrease in the resource utilization (and higher overload probabilities). This affects more a system with end-to-end aggregation because, as mentioned before, in area aggregation the four sessions still share resources. As an example, to reach an utilization larger than 75% and an overload probability smaller than 4%, $\tau/T \leq 1/2$ in area aggregation with one service class and $\tau/T \leq 1/6$ in end-to-end aggregation. The case of area aggregation with one service class and a proportional τ/T achieves a resource utilization slightly smaller (and a slightly larger overload probability), when comparing with the case of a fixed τ/T. This is due to the decrease in the frequency of the reservations updates in the central area. However, the differences between the utilization and overload results achieved in these cases are very small, reflecting that the increase in the τ/T of the central area has little impact in the resource utilization and overload probability. The increase in the number of service classes increases the number of aggregates required in each area. Therefore, the case of two service classes is an intermediate case between end-to-end and area aggregation with one service class.

Although this model does not detail the offered load at the flow level, it is still possible to derive a (rough) approximation for the signaling gains, by noting that a unitary reserved resource utilization is achieved when using per-flow signaling. From Fig. 6 (a), it can be seen that a unitary utilization is approximately obtained when τ/T=1/256. Thus, as an example, the signaling gains will be 32 when τ/T=1/8, and 128 when τ/T=1/2.

In general, the results obtained with this model confirm the ones obtained with the per-flow load model. Although this model accommodates a time-varying offered load, it does not allow the determination of the exact signaling load. In the next section, we will present a simulation study, based on measured aggregates, that determines the signaling gains for time-varying offered loads.

5.3 Simulations with Measured Aggregate

We consider now a traffic trace measured at NLANR on December 1, 1999 [10] and evaluate the system performance via discrete event simulation. This trace is characterized by a very large variance and noise. The information available includes the arrival time, the duration, and the number of bytes of each flow. The total number of flows is 64087. The average flows' bandwidth is 19.6 Kb/sec, but approximately 80% of the flows have a bandwidth bellow the mean. The total average bandwidth is 1.43 Mb/sec and the variance is 0.144 Mb/sec. The simulation results correspond to averages taken over a total of 20 runs; in all runs, the phase (time instant of beginning) of each aggregate was chosen randomly. This section considers, in subsection 5.3.1, the domain depicted in Fig. 4, and in subsections 5.3.2 and 5.3.3, two larger domains representing access and core networks, respectively.

5.3.1 Dumbbell Topology

In order to study a scenario with overload, we decreased the bandwidth of all areas of the Dumbbell network in section 5.1 to 10 Mb/sec.

Fig. 7 (a) shows the signaling gains with both end-to-end and area aggregation, considering (i) all reservation attempts and (ii) only successful reservations. The results confirm the ones obtained with the per-flow load model.

Fig. 7 (b) shows the signaling gains of internal routers. In all types of aggregation, the signaling gains increase sharply with the bulk size reaching much higher values than in the per-flow load model case. With a 1.25 Mb/sec bulk size and only successful reservations, the signaling gains reach 900 in end-to-end aggregation, 1100 in area aggregation with two service classes, 1600 in area aggregation with one service class and fixed bulk size, and 1800 with one service class and proportional bulk sizes. This is essentially due to the larger ratio between the bulk size and the flow's bandwidth. We also notice that the signaling gains are always larger with area aggregation than with end-to-end aggregation. This is explained by the higher burstiness of the traffic and the larger resource sharing that is possible with area aggregates. The overall traffic in the aggregate becomes smoother and the number of signaling attempts decreases. Note also that we are still far from the limiting situation of the per-flow load model case, since the maximum bulk size value considered in the experiments still allows sufficient granularity in the bandwidth adjustment process. This, in fact, represents a more realistic scenario. When taking into account non-successful reservation attempts, it can be seen that the gains of area aggregation over end-to-end aggregation are effectively higher. This can be explained by the lower blocking with area aggregation, which reflects in a lower number of unsuccessful signaling attempts. Area aggregation with two service classes is again an intermediate case between the other two.

We also notice from Fig. 7 (b) that signaling gains of 124 in area aggregation with one service class and fixed bulk sizes, and 160 in area aggregation with proportional bulk sizes, are obtained when the ratio between the bulk size and the mean flows' bandwidth is 8 (i.e. with a bulk size of 156.8 Kb/s). For the same ratio in the per-flow load experiments (i.e. a bulk size of 8 Mb/sec), the gains were 15 and 18, respectively. This difference can be explained by the asymmetry of the distribution of the flows' bandwidth in the real aggregate, which is 80% bellow the mean (recall that in the per-flow load model the flow's bandwidth is fixed). Thus, in this case, the bandwidth update attempts are provoked by a much smaller number of flows (mostly the flows with larger bandwidth), leading to a decrease in the signaling rate.

The reserved resource utilization is depicted in Fig. 7 (c). Results show again that area aggregation achieves higher resource utilization, slightly decreasing with proportional bulk sizes. As an example, to achieve an utilization larger than 90%, the bulk size should be lower than 800 Kb/sec in area aggregation with one service class (with fixed and proportional bulks), 400 Kb/sec in area aggregation with two service classes, and 300

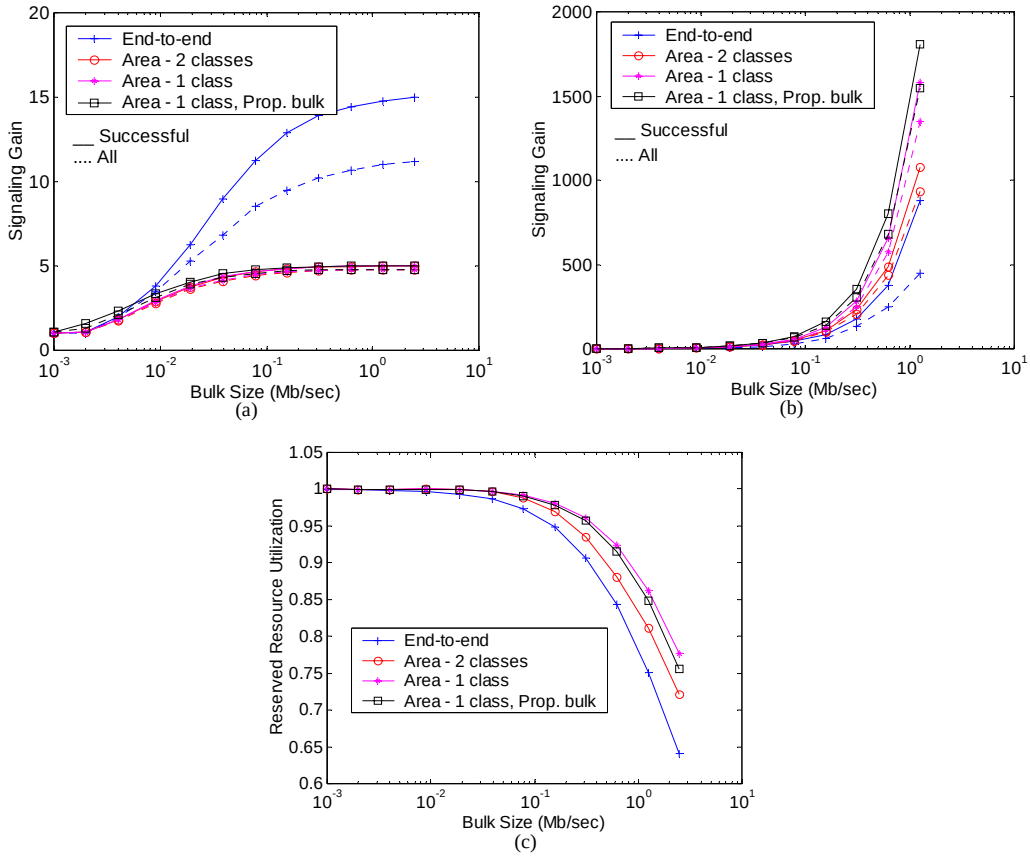

Fig. 7 Signaling gains of (a) all routers in the domain, (b) internal routers and (c) reserved resource utilization (measured aggregate).

Kb/sec with end-to-end aggregation. For these bulk sizes, the signaling gain in internal routers with end-to-end aggregation is 150, with area aggregation, one service class and fixed bulk size is 1250, and with area aggregation, one service class and proportional bulk sizes is 1500.

The results of these studies show that area aggregation, compared with end-to-end aggregation, can achieve larger signaling gains and larger utilizations, even for relatively small networks. They also show that area aggregation with proportional bulk sizes can raise the signaling gains with very little impact on the utilization. In the next two sections, we will study the performance of aggregation in larger network domains.

5.3.2 Access Network

We consider an access network having a tree topology, as depicted in Fig. 8. The domain contains 8 areas, and each area, except the right area, contains two ABR pairs. There are a total of 8 sessions in the domain, each from a different left DBR to the right DBR. We assume that all sessions belong to the same service class. The number of routers inside each area is 4. Thus, each session traverses 25 routers (not including the egress DBR). Since all sessions have different origin/destination pairs and the domain supports a single service class, there is an aggregate per session in end-to-end aggregation, and one aggregate in each ABR pair in area aggregation. The bandwidth of each ABR pair is presented in the figure in Mb/sec. In area aggregation, we consider the cases of fixed and proportional bulk sizes. In the latter case, the bulk size in the central area is twice the bulk size in the 6 left areas and the bulk size in the right area is 4 times larger than the one in the 6 left areas.

The signaling gains in internal routers are presented in Fig. 9 (a). These gains are very large, higher than

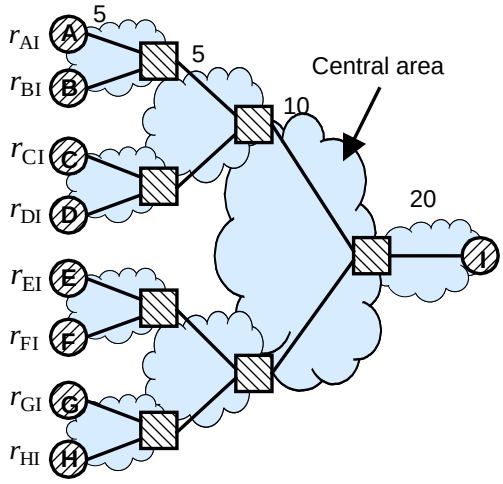

Fig. 8 Access network.

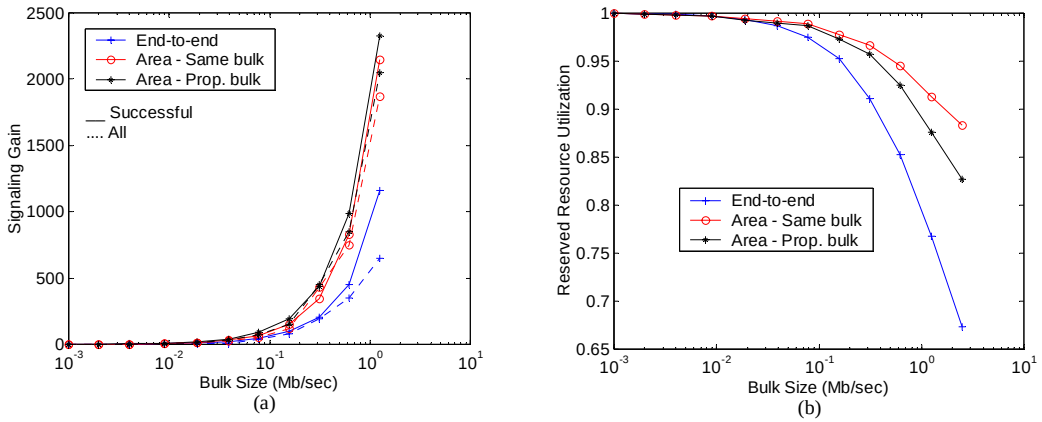

Fig. 9 (a) Signaling gains in internal routers and (b) reserved resource utilization (access domain).

2000 in both cases of area aggregation. There is also an increase in the signaling gains when proportional bulks are considered. The gains with area aggregation achieved in the case of all reservation attempts are approximately 4 times larger than the ones with end-to-end aggregation. Fig. 9 (b) presents the reserved resource utilization, which is always larger than 88% with area aggregation and fixed bulk sizes. Note that, in area aggregation, the right area has 8 sessions all sharing the same aggregate. The utilization in area aggregation with proportional bulks is slightly smaller than the one obtained with fixed bulk size, but much higher than the one obtained with end-to-end aggregation. As an example, to achieve an utilization always larger than 92%, the bulk size can grow up to 1.25 Mb/sec in area aggregation with a fixed bulk size, and only to 312 Kb/sec in end-to-end aggregation.

5.3.3 Core Network

The core network is based on a Dumbbell topology, as depicted in Fig. 10. The domain contains 5 areas, with 2 ABR pairs in each peripheral area and one ABR pair in the central one. The number of routers inside each area is 4. There are 16 sessions traversing the domain: each left DBR has a session to each right DBR. We

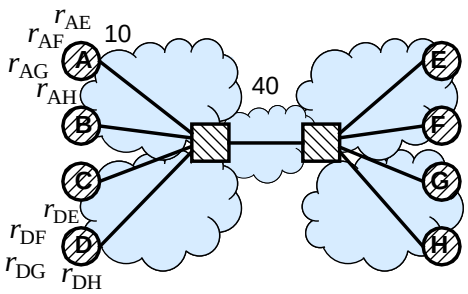

Fig. 10 Core domain - Dumbbell topology.

assume that all sessions belong to the same service class. As in the case of the access network, there is an aggregate per session in end-to-end aggregation, and an aggregate per ABR in area aggregation. Note that all 16 sessions traverse the central area. The bandwidth of each ABR pair is represented in the figure in Mb/sec. In area aggregation, we also consider the cases of fixed and proportional bulk sizes.

The signaling gains in internal routers presented in Fig. 11 (a) are, once more, much larger in area aggregation than in end-to-end aggregation, and even larger when proportional bulks sizes are considered. In terms of resource utilization, Fig. 11 (b), we observe that there is a large difference between the utilization obtained with end-to-end and area aggregation. The one of area aggregation is always larger than 90%. Moreover, a proportional bulk size does not degrade the utilization. The large resource sharing in area aggregation enables to achieve an utilization larger than 95% with bulk sizes of 1.25 Mb/sec, while in end-to-end aggregation the bulk may only increase until 150 Kb/sec.

The results obtained with the core and access networks show that the overall performance gains of area aggregation over end-to-end aggregation are likely to increase with the size of the network and can achieve very large values.

6 Conclusions

We analyzed the tradeoffs between signaling and resource utilization in DiffServ networks partitioned in areas (hierarchical domains) using flow aggregates that can be dynamically adjusted. These tradeoffs are studied using two analytical models. In the first model, based on multidimensional birth-death processes, the offered load is detailed at the flow level, which allows accurate assessment of the signaling load. The second model accommodates time-varying offered loads, which allows studying the tradeoffs between the time-scale of the

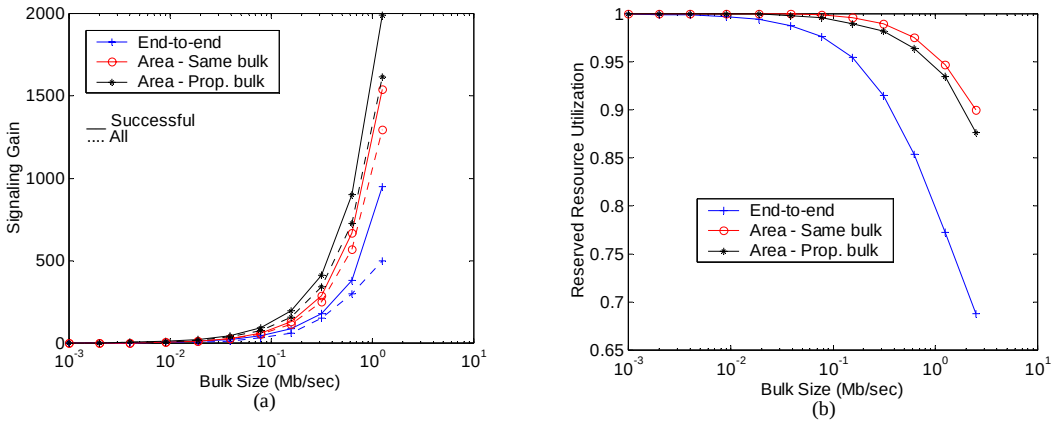

Fig. 11 (a) Signaling gains in internal routers and (b) reserved resource utilization (core domain).

aggregate demand and the time-scale of signaling. We also analyze the effect of a measured flow aggregate, through discrete-event simulation, on network domains with a relatively large size, representing both access and core networks. Our results show that structuring a network domain in areas achieves high performance gains, which can contribute to reduce significantly the cost of core routers.

7 References

1. R. Braden et al., “Integrated Services in the Internet Architecture: An Overview”, IETF RFC 1633, June 1994.
2. D. Awduche et al., “RSVP-TE: Extensions to RSVP for LSP Tunnels”, IETF RFC 3209, December 2001.
3. F. Baker el al., “Aggregation of RSVP for IPv4 and IPv6 Reservations”, IETF RFC 3175, September 2001.
4. P. Pan et al., “BGRP: A Tree-based Aggregation Protocol for Inter-Domain Reservations”, Journal of Comm. and Networks, 2(2), pp. 157-167, June 2000.
5. O. Schelén and S. Pink, “Aggregation Resource Reservations over Multiple Routing Domains”, In Proceedings of IWQoS’98, Napa, CA, May 1998.
6. H. Fu and E. Knightly, “Aggregation and Scalable QoS: A Performance Study”, In *Proc. of IWQoS '01*, June 2001.
7. Internet Page of Qbone, http://tombstone.oar.net/sitemap.html.
8. S. Sargento and R. Valadas, “Aggregation Performance in Hierarchical Domains”, University of Aveiro Technical Report, May 2002.
9. K. Ross, “Multiservice Loss Models for Broadband Telecommunication Networks”, *Springer*, 1995.
10. Internet Page of NLANR, http://moat.nlanr.net/Traces/ Kiwitraces/auck2.html.
11. S. Sargento et al., “IP-based Access Networks for Broadband Multimedia Services”, IEEE Communications Magazine, Special Issue on Broadband Access, February 2003