CONNECTION-ORIENTED FLOW AGGREGATION WITH A RSVP OVER ATM EXAMPLE

Martin Lorang

Abstract

This paper deals with the IP flow to ATM VC mapping of the RSVP over ATM implementation of the ACTS project DIANA. As part of an edge router at the ingress from an IP to an ATM network, this implementation provides signalling translation from RSVP to ATM, handles the admission of flows to ATM VCs and controls an IP over ATM queuing discipline. It is thus an example of a traffic descriptor and QoS parameter based resource reservation that strictly guarantees end-to-end QoS. In order to address scalability issues, a concept of massive aggregation of flows to VCs and a dynamic bandwidth management is applied to reduce the control overhead induced by signalling and the maintenance of per-flow state information.

1 Introduction

Today, the Internet offers a flexible, but simple connectionless best-effort service and for sure this is inadequate for QoS sensitive applications. For this reason, the IETF has been working on extensions to the current Internet protocol suite in order to enable service guarantees (Integrated Services) or at least differentiation (Differentiated Services).

In Integrated Services, RSVP allows applications to request QoS for individual flows. However, the wide scale deployment of RSVP must be approached with care because the processing of (periodically refreshed) signalling messages, packet classification based on almost all header fields, sophisticated multicast support and the handling of per-flow state is challenging in backbone routers.

Differentiated Services improve scalability by classifying and marking packets only once at the ingress to a DS capable network. The goal is to obtain a particular forwarding behaviour on DS routers, resulting in relative QoS. Absolute guarantees with such an approach can only be given with signalling alike mechanisms, e.g. by overlaying a net of bandwidth brokers that control resource reservations on routers in their domain.

Therefore, this paper investigates a scalable connection-oriented approach directly building upon signalling based approaches, namely Integrated Services and ATM. In order reduce the control overhead induced by signalling and per-flow state, flow aggregation and a dynamic bandwidth management is applied.

Using a prototype developed by the ACTS project DIANA, aggregation and resource calculation algorithms are discussed and evaluated. Those functions are embedded in a process architecture that is briefly explained in section 2. Section 3 and 4 deal with an effective bandwidth based resource calculation for Controlled Load flow aggregates and its integration with a heuristic dynamic bandwidth management algorithm, respectively. Since the flow admission, resource calculation and scheduling framework must not only meet the expected QoS requirements but also - and especially - reduce flow state and signalling overhead, the simulative and experimental activities presented in section 5 focus on the performance of the dynamic control of flow aggregates, as carried out by the combination of the resource calculation and dynamic bandwidth management algorithm. While RSVP over ATM as such is a promising option to quickly introduce guaranteed QoS to broadband networks, the paper concludes with alternative concepts towards the introduction of connection-oriented flow aggregation to pure IP networks and extensions towards multi-level aggregation.

References that introduce to Differentiated Services, Integrated Services and RSVP over ATM, material that supports the statements above as well as further implementation details can be found in [1] and [2].

2 Overview of the Process Architecture, Traffic Control and Resource Allocation Functions

DIANA´s RSVP over ATM traffic control architecture is implemented in a router at the border between an IP

Martin Lorang, University of Stuttgart, IND, lorang@ind.uni-stuttgart.de

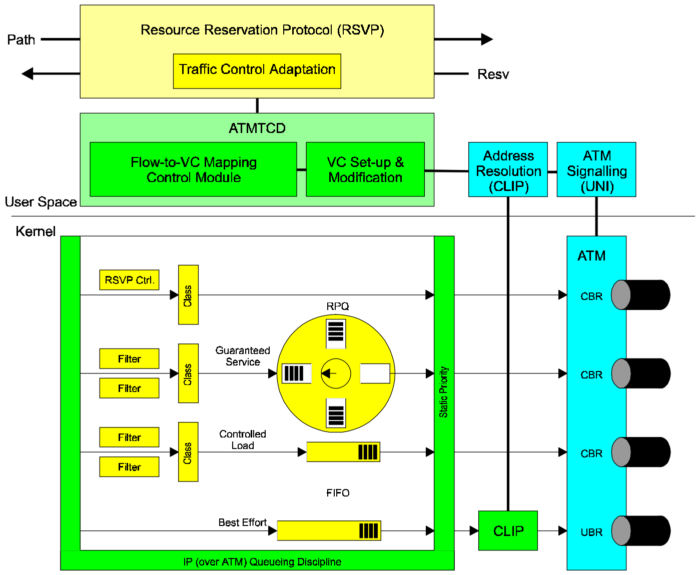

Fig. 1: DIANA´s RSVP over ATM traffic control architecture

network and an ATM network, at a point where a sufficiently large number of IP flows can be aggregated into a single VC towards a common egress of the ATM network.

The RSVP demon process controlling this architecture starts interacting with ATM processes when traffic control related RSVP messages arrive. For this purpose, RSVP has been enhanced with an ATM specific Link Layer Dependent Adaptation Layer (LLDAL). DIANA´s LLDAL implementation is separated into a traffic control adaptation to RSVP and a demon process ATMTCD that controls most of the interworking tasks (address resolution, VC set-up) and traffic management functions (flow-to-VC mapping, dynamic bandwidth management, queue management & schedulers). This process creates, modifies or deletes entries in a flow-to-VC mapping table, which contain information about a single VC and all the flows mapped to that VC. Moreover, an aggregation filter associated with each entry ensures that only flow requests that match the entry's aggregation criterion are mapped to the VC.

Whenever ATMTCD cannot assign an incoming RSVP reservation request to one of the existing entries, it creates a new one, starts the ATM address resolution procedure, and, if this succeeds, orders a VC. When the ATM connection is established, ATMTCD sends a confirmation back to the RSVP demon which then forwards the reservation message. Conversely, if a matching entry already exists, the flow is added and resources are re-calculated. In order to reduce the control overhead induced by signalling for flows with a possibly low bandwidth demand and short lifetime, the aforementioned dynamic bandwidth management scheme over-allocates bandwidth and re-negotiates a VC only if the resource calculation for the flows mapped to that record yields a bandwidth that crosses a re-negotiation threshold.

The RSVP over ATM process ensemble introduced in Fig. 1 controls an IP over ATM priority queuing discipline that - as further discussed in [2] - includes a RPQ scheduler [3] ensuring isolation of Guaranteed Service flows with different delay requirements if they are aggregated to the same VC.

For Controlled Load service flows that do not include delay bound information, separate FIFO queues can be used. In addition, a default queue for best-effort traffic is foreseen making use of the Classical IP over ATM (CLIP) set-up procedures. Unlike the other classes, the default best-effort class does not classify packets. Otherwise, ATMTCD is in charge of installing the IP packet filters that allow to map packets to the classes that provide the forwarding behaviour specified through RSVP signalling. As an additional feature, the queuing discipline assigns dedicated RSVP over ATM control CBR-VCs to protect RSVP signalling messages from being dropped in the ATM network when congestion occurs.

3 Resource Calculation for the Controlled Load Service Class

The resource calculation function considered here is an application of a heuristic extension of the concept of effective bandwidth in a bufferless system based on Chernoff´s large deviation approximation [4] to a system with both buffer and bandwidth resources [5]. Both approaches yield expressions for the effective bandwidth that can be derived from a more general definition in [7] applied to stochastic processes with stationary independent increments [6]. The heuristic extension is additionally restricted to leaky bucket constraint sources. However, this special case exactly matches the requirements of the Controlled Load service class.

The authors of [5] claim that leaky bucket constraint, but non-deterministic sources in a system with buffer behave as if they passed individual, lossless virtual buffer/trunk systems, each of which consumes buffer and bandwidth in proportion to the system capacity *B* and *C*, and then arrived with a modified, leaky bucket parameters dependent rate at the bufferless server. As a result of this heuristic assumption, the modified (effective) bandwidth *e* for lossless performance has been derived as

$$e=\frac{P}{1+BC^{-1}\cdot B_T^{-1}\cdot(P-r)} \quad \text{if} \quad r\le\frac{B_T}{BC^{-1}} \tag{1}$$

or $e=r$ if the above condition does not hold, where *P*, *r*, B_T are the peak cell rate and the leaky bucket parameters mean token rate and token buffer size, respectively. *B*/*C* is the maximum delay in the system. As a consequence, the dynamic determination of *C* implies a variable value of *B.* With the moment generating function $M_j(s)$ of sources of class *j,* Chernoff´s approximation gives an upper bound of the actual loss as

$$\log P_{loss}\le-(sC-\sum_{j=1}^{J}K_j\cdot\log M_j(s)) \quad , \tag{2}$$

which holds for all $s\ge0$ thus being the most accurate if the left hand side is minimal, i. e.

$$\sum_{j=1}^{J}K_j\cdot\frac{M_j'(s)}{M_j(s)}=C \quad , \tag{3}$$

the sum of the effective bandwidths of all traffic sources in the system. K_j denotes the number of sources of class *j*. Assuming deterministic, periodic on-off sources with the modified peak rate e_j and activity $w_j=r_j/e_j$ in uniformly distributed random phase relation to each other, the authors of [5] obtain

$$M_j(s)=1-w_j+w_j\cdot\exp(s\cdot e_{0,j}) \tag{4}$$

If a loss bound *L* is introduced, the Bahadur-Rao refinement of the bound taken into account [8] and $C=C(s)$ interpreted as the variable resource demand of the flow aggregate not being subject to any limitation, (2) and (3) can be combined and completed to

$$s\cdot\sum_{j=1}^{J}K_j\cdot\frac{M_j'(s)}{M_j(s)}-\sum_{j=1}^{J}K_j\cdot\log M_j(s)+\frac{1}{2}\cdot\log2\pi\cdot s^2\sum_{j=1}^{J}K_j\cdot\left(\frac{M_j''(s)}{M_j(s)}-\left(\frac{M_j'(s)}{M_j(s)}\right)^2\right)\ge\log\frac{1}{L} \tag{5}$$

In this inequation, the solution *s* can be regarded as the statistical multiplexing working point that changes and has to be re-computed whenever a new source is added to or removed from the aggregate. The more sources build this aggregate the better is the multiplexing gain although the various source types contribute differently. Since $M'(s)^2\le M''(s)M(s)$ [6], expression $M'(s)/M(s)$, the effective bandwidth of every source, is strictly increasing in *s* while the multiplexing gain decreases. On the other hand, if a source is added to the aggregate, a term $s\cdot M'(s)\cdot M(s)^{-1}-\log M(s)$ is added to the Chernoff approximation in (5), which is equal to zero for $s=0$ and strictly increasing thereafter, and a non-negative $2\pi\cdot s^2(M''(s)\cdot M(s)^{-1}-(M'(s)\cdot M(s))^2)$ to the Bahadur-Rao refinement, which can be easily verified using again the relation from [6]. Hence, a new source increases the continuous lhs of (5) allowing for a smaller solution *s* and a lower effective bandwidth for all sources. Note that negative contributions of the Rao-Bahadur term should be ignored because they impair the conservative Chernoff bound otherwise. The complexity of (5) can be reduced slightly if the Rao-Bahadur refinement is simply ignored, however, the actual issue remains: The lhs of (5) is only a precise lower bound of $-\log P_{loss}$ if the number of sources is high. If (4) replaces $M(s)$ in (5), the Rao-Bahadur refinement is dropped and $s\rightarrow\infty$, (5) converges to

$$\sum_{j=1}^{J}K_j\cdot\log\frac{1}{w_j}\ge\log\frac{1}{L} \tag{6}$$

If only a few sources build the aggregate, this condition may not be met. At first glance, this does not seem to be a problem since, as pointed out in [5], allocating the sum of the deterministic effective bandwidths e_j of all sources ensures lossless multiplexing. However, the transition from deterministic to statistical multiplexing when new sources enter can only be made successfully if a sufficiently precise initial value for *s* is provided. It turns out that the Newton algorithm deployed to solve (5) needs a significant number of iterations to converge, even if heuristic estimates and corrections are made. Once the transition is made, (5) usually has to be evaluated only once, i. e. the first iteration returns a solution that is precise enough, as also reported in [6].

4 Heuristic Dynamic Bandwidth Management Algorithm

The goal of a dynamic bandwidth management algorithm is to over-allocate bandwidth to an extent that is justified by the savings in signalling processing overhead. The algorithm introduced below is admittedly a simplistic approach that neither makes any assumptions about stochastic models of incoming reservation requests nor tries to estimate the stochastic characteristics of those. It serves only a single purpose, namely demonstrating the potential of connection-oriented flow aggregation, and is not supposed to perform particularly well.

The algorithm consists of two levels: Level A measures the mean capacity increase *M* with the arrival of a reservation request *R*. However, extreme and rare deviations from the mean are taken into account only up to a maximum because over-allocating bandwidth for very large but seldom occurring types of reservation requests is inefficient. The calculation procedure below resembles TCP´s retransmission timeout calculation [10],

$$
\begin{aligned}
&Err = max(M - R, Err)\\
&M = M + \alpha Err\\
&D = D + \alpha(|Err| - D)\\
&G = M + D
\end{aligned}
$$

and is carried out whenever a reservation request arrives. As opposed to TCP, the dynamic bandwidth management algorithm uses the same weight α for both the calculation of the mean *M* and the mean deviation *D*. Increasing the expression for *D* as in TCP would overrate peaks, a strategy that is not appropriate for a dynamic bandwidth management, as mentioned above. Another difference is the calculation of α

$$\alpha = \exp\left(\frac{\ln(c_\alpha)}{K_{OA} N_\alpha}\right)$$

which dynamically adapts α dependent on the current K_{OA} the measured G is multiplied with when determining the amount of bandwidth over-allocation. If $c_\alpha \le 1$, α increases if K_{OA} decreases and the algorithm gets more responsive to deviations from the estimated mean *M*. $K_{OA} N_\alpha$ is the number of updates needed until the contribution of an reservation request to the mean *M* or mean deviation *D* has reduced to c_α times the initial contribution. While G is re-computed whenever a reservation request reaches the system, K_{OA} is only updated occasionally, in regular intervals of length T , and in the current version of the algorithm even only in unit steps in order to ensure that each change affects a significant share of arriving reservation requests. For this very reason, both the estimated mean *M* and the mean deviation *D* are included in *G* above. The update procedure of the over-allocation factor K_{OA} , level B of the algorithm, works as follows:

- Measure the wasted data volume and the number of re-negotiations at every level A update.
- If K_{OA} was modified last time, check whether the costs have been reduced since then. If yes, consolidate the update, otherwise go back to the previous, obviously better value. Increase the thresholds D_L and D_{XXL} if a modification does not succeed in saving costs and decrease them else. Return.
- Identify the dominant absolute cost and counter-act by incrementing or decrementing K_{OA} . Per definition, cost A dominates cost B if $C_A > D_{XXL} \cdot C_B$ or $C_A > D_M \cdot C_B$ and $\Delta C_A > D_M \Delta C_B$, the cost savings or increase since the last update. Remember the change to make the cost evaluation above during the next level B update
- Reset the measurements of wasted data volume and the number of renegotiations used above.

5 Evaluation of the Dynamic Resource Allocation Functions

Since the measurements only involve signalling and no data traffic, a simple experimental platform consisting of two Linux PCs connected back-to-back via ATM is sufficient. The PCs run the edge router´s RSVP over ATM software simultaneously with Tcl controlled signalling message generators and receivers doing nothing but initiating and releasing RSVP sessions. A GN Nettest Interwatch traffic analyser helps to monitor the measurements and to verify that the set-up works properly.

The signalling message generators modelled a set of relatively heterogeneous Markovian on-off ($\lambda = \mu$) sources with two dominant source types, denoted as type 1 and 6 in the table below. With those sources the behaviour of the dynamic resource allocation function can be demonstrated, however, future experimentation will have to involve correlated and autocorrelated sources.

Fig. 2 shows simulation results for a system that only runs level A of the dynamic bandwidth algorithm. Fixed over-allocation factors are pre-configured regardless of the costs this setting may imply and so the renegotiation ratio drops quickly as the bandwidth over-allocation grows. Fig. 3 depicts results for the algorithm including level B. In principle, the results do not differ, but it is interesting to see how expensive re-negotiation has to be before it is worth over-allocating. The simulations in both figures do not implement the effective bandwidth approach discussed in section 3 but take the deterministic effective bandwidth *e*. As opposed to the simulation, the edge router implementation integrates the resource allocation function and the dynamic bandwidth management. With the source scenario under consideration, the aggregate is practically always at working points that allow to profit from the smoothing effects of the statistical multiplexing and more flows can benefit from over-allocation.

Type	1	2	3	4	5	6	7	8
Number	100	32	16	12	10	25	10	5
e [b/s]	10000	30000	45000	60000	75000	100000	200000	500000
w [b/s]	0.5	0.5	0.5	0.5	0.5	0.5	0.25	0.1
r [b/s]	5000	15000	22500	30000	37500	50000	50000	50000
P [b/s]	12500	37500	56250	75000	93750	125000	250000	625000
B_T [b]	3000	9000	13500	18000	22500	30000	80000	230000
$\lambda = \mu[1/s]$	0.004	0.002	0.002	0.002	0.002	0.002	0.002	0.002

The remaining parameters needed to reproduce the results are summarised in the table below

L	BC^{-1}	c_α	N_α	I	D_M	D_{XXL}	T
0.0001	0.1 s	0.1	10	0.05	1.2, 1.4, ..., 2.0	2.4, 2.8, ..., 4.0	120 s

6 Conclusions

As stated earlier, RSVP over ATM is a traffic descriptor and QoS parameter based connection-oriented architecture that strictly guarantees QoS end-to-end. Since ATM is not particularly useful for connection-less traffic, RSVP over ATM should be complemented by an IP switching backbone network architecture that is optimised for those streams. The resulting network would thus comprise Integrated Services routers in the access area and combined IP/ATM switches in the backbone. In this case, the edge routers would then

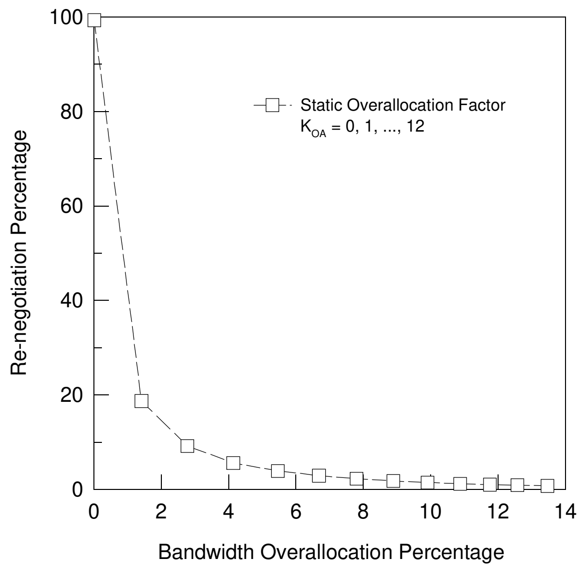

Fig. 2:Simulated re-negotation ratio for static over-allocation factors

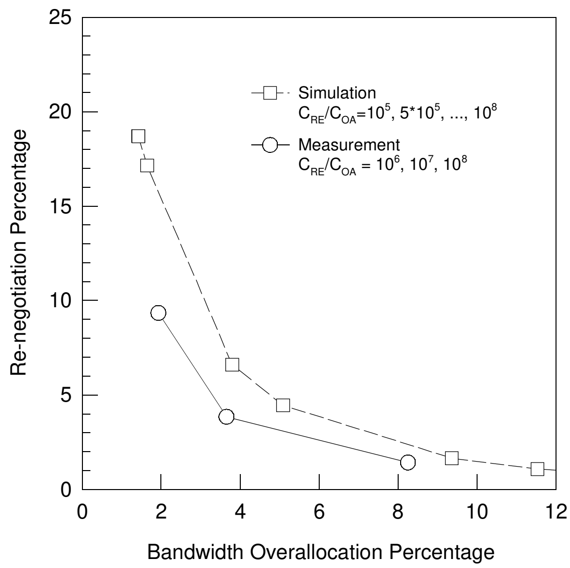

Fig. 3: Measured and simulated re-negotiation ratio for cost dependently controlled over-allocation factors

separate connection-oriented from connection-less traffic. In this way, the requirements of both connection-oriented and connection-less traffic can be satisfied, and if the connection-oriented network is improved in the direction of flow aggregation to reduce state information and signalling overhead in the backbone, both in a scalable way - as demonstrated above.

Extensions to RSVP supporting aggregation in a pure IP based network have been proposed in [11], [12] and recently in [13]. While the first two discuss both TOS based or tunnel based aggregation - the latter being relatively similar to the RSVP over ATM approach - and some relatively simple RSVP protocol extensions, the more recent contribution explicitly proposes a DS field classification within the aggregation region and considers multi-level aggregation, dynamic bandwidth management issues, deaggregation node determination and routing issues. In fact, multi-level aggregation and deaggregation requires extensive reservation and other protocol support. In RSVP over ATM, which can be regarded as a simple special case, a switched VC through an ATM network is equivalent to a layer 2 link from the IP´s routing protocol point of view. IP routing information is propagated as usual, and the end-point of the aggregate can be easily identified by the ATM address resolution mechanisms. In the case of multi-level aggregation, a generalised approach may replace ATM´s address resolution requests and replace by deaggregation point identification request and response messages, respectively, which include a stack in which both aggregation and deaggregation points register themselves unless they are already logically bypassed by aggregation tunnels [1]. Since those mechanism only map aggregation to deaggregation points but rely on routing protocol mechanisms to identify which destinations can be reached via an aggregate between them, they work without routing protocol changes.

For Differentiated Services networks, so called bandwidth brokers have been considered that manage resource information and that may inititiate end-to-end reservation requests along a chain of bandwidth brokers managing separate DS domains [14]. Such mechanims may vary the grade of connection-orientation from a purely connection-less to a fully reservation capable, connection-oriented network, using similar mechanisms as the ones investigated in this paper.

References

[1] M. Meuli (ed.), *Evaluation and Demonstration of Various Approaches towards QoS in IP Networks*, ACTS project DIANA, Deliverable 5, February 2000.

[2] E. Aarstad, L. Burgstahler and M. Lorang, „Description of the Flow-to-VC Mapping Control Module in DIANA´s RSVP over ATM Architecture”, *in QoS Summit´99*, Paris, France, November 16-19, 1999.

[3] Liebeherr, D. E. Wrege and D. Ferrari, „Exact Admission Control for Networks with a Bounded Delay Service“, in *IEEE/ACM Transactions on Networking*, Vol. 4, No. 6, pp. 885-901, December 1996.

[4] J. Y. Hui, „Resource Allocation for Broadband Networks“, in *IEEE Journal on Selected Areas in Communications*, Vol. 6, No. 9, pp. 1598-1608, December1988.

[5] A. Elwalid, D. Mitra and R. H. Wentworth, „A New Approach for Allocating Buffers and Bandwidth to Heterogeneous, Regulated Traffic in an ATM Node”, in *IEEE Journal on Selected Areas in Communications*, Vol. 13, No. 6, pp. 1115-1127, August 1995.

[6] F. Kelly, „Notes on Effective Bandwidth”, in *Stochastic Networks: Theory and Applications (ed. F. Kelly, S. Zachary and I. Ziedins)*, Oxford University Press, 1996, pp. 141-168.

[7] C.-S. Chang, „Stability, Queue Length, and Delay of Deterministic and Stochastic Queueing Networks”, in *IEEE Transactions on Automatic Control*, Vol. 39, No. 5, pp. 913-931, May 1994.

[8] J. A. Bucklew, *Large Deviation Techniques in Decision, Simulation, and Estimation*, Wiley, 1990.

[9] P. Billingsley, *Probability and Measure - Second Edition*, Wiley, New York, 1986.

[10] W. R. Stevens, *TCP/IP Illustrated, Volume 1 - The Protocols*, Addison Wesley, 1994

[11] S. Berson and S. Vincent, *Aggregation of Internet Integrated Services State.* Work in Progress, IETF Integrated Services Working Group, <draft-berson-classy-approach-01.ps >, November 1997.

[12] R. Guerin, S. Blake and S. Herzog, *Aggregating RSVP-based QoS Requests.* Work in Progress, IETF Integrated Services Working Group, <draft-guerin-aggreg-rsvp-00.txt>, November 1997.

[13] F. Baker, C. Iturralde, F. Le Faucheur and B. Davie, *Aggregation of RSVP for IPv4 and IPv6 Reservation,* Work in Progress, IETF Integrated Services over Specific Link Layers Working Group, <draft-ietf-issll-rsvp-aggr-02.txt>, March 2000.

[14] B. Teitelbaum, S. Hares, L. Dunn, R. Neilson, V. Narayan and F. Reichmeyer, “Internet2 QBone: Building a Testbed for Differentiated Services”, in *IEEE Network*, Vol. 13, No. 5, pp. 8-16, September/October 1999.