Admission Control for End-to-End Multimedia Content Delivery with Quality of Service Guarantees

S.Georgoulas, G.Pavlou
University College London, UK
{S.Georgoulas,G.Pavlou}@ee.ucl.ac.uk

E.Borcoci
Universitatea Politehnica Bucuresti (UPB), Romania, eugen.borcoci@elcom.pub.ro

K.H.Ho
University of Surrey, UK
K.Ho@surrey.ac.uk

E.Vraka
ERICSSON, UK
effimia.vraka@ericsson.com

Abstract

End-to-End Quality of Service guaranteed delivery for multimedia content is a challenging issue, especially in multi-domain environments and heterogeneous network infrastructures. The approach proposed by the ENTHRONE project to solve the End-to-End Quality of Service problem in a scalable manner is to establish, and activate when needed, long-term Quality of Service enabled aggregate pipes over multi-domain environments for the subsequent transport of individual flows from multimedia content providers to multimedia content consumers. Based on this approach, this paper proposes admission control schemes both at the granularity of individual flows and aggregate demands. Through simulations we show the proper joint operation of the schemes, their ability to provide Quality of Service, resource utilization gains and to minimize service rejection probabilities.

1. Introduction

End-to-End (E2E) Quality of Service (QoS) guaranteed content delivery is still an open and challenging issue, especially in multi-domain environments and heterogeneous network infrastructures. Various business/service models have been proposed to tackle this issue, involving entities such as *Service Providers (SPs), Content Providers (CPs), Network Providers (NPs), Access Network Providers (ANPs)* and *Content Consumers (CCs)*.

The method proposed by the ENTHRONE project [1] to solve the E2E QoS problem in a scalable manner is to establish long-term QoS-enabled aggregate pipes over multi-domain environments for the subsequent transport of individual flows from CPs to CCs. That is, based on forecasted data, provider Service Level Agreements and Specifications (pSLAs/pSLSes) between interconnected SPs/NPs in this E2E chain from the CPs until the ANPs are established. After a sequence of pSLAs has been established in this E2E chain, an SP is able to offer services to individual CCs and a customer SLA between the SP and each interested CC is established (cSLA/cSLS). The key goal is to assure for each individual cSLS flow the desired E2E QoS guarantees while optimizing at the same time the utilization of the reserved resources. In order to do so, dynamic decisions based on run-time information about the status of the resources are needed in the form of admission control schemes, both at the granularity of individual flows (cSLSes) and at the granularity of aggregate demands (pSLSes).

In this paper we focus on developing and evaluating through simulation the effectiveness of admission control schemes in order to provide QoS guarantees for multimedia content delivery using the ENTHRONE architecture as a reference point. The rest of this paper is organized as follows. In Section 2 we briefly describe the ENTHRONE architecture and discuss the placement of the admission control schemes. In Section 3 we elaborate on the functionality of the cSLS and pSLS admission control schemes and in Section 4 we evaluate their joint performance. Finally in Section 5 we conclude, summarizing our findings and giving some directions for future work.

2. Admission Control Placement in the ENTHRONE Architecture

A simplified version of the ENTHRONE architecture that focuses on cSLS/pSLS related issues in the E2E chain is shown in Figure 1.

978-0-7695-3130-4/08 $25.00 © 2008 IEEE
DOI 10.1109/WIAMIS.2008.40

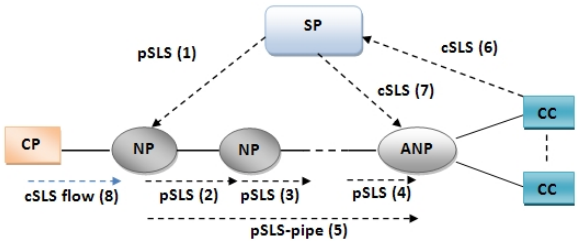

Figure 1. Simplified ENTHRONE architecture

The numbers in brackets indicate the sequence of actions performed before the delivery of any content. Prior to accepting any cSLA, the SP establishes a set of QoS enabled paths (aggregate pipes), based on contracts (pSLAs) agreed between the SP and the NP adjacent to the CP and between neighboring NPs downstream (actions 1-4). This way, a pSLS pipe until the ANP is constructed (action 5), with segments of it traversing individual NPs. After this pSLS pipe is constructed, the SP can accept cSLA requests from CCs taking also into account the resources in the ANP (actions 6-7) and CCs can request content delivery (action 8). ENTHRONE makes a clear distinction between the subscription phase (logical construction) of cSLSes/pSLSes and the actual activation/invocation of them, which may happen at a later time. This is an assumption we also adopt and the admission control schemes presented in this paper deal with the invocation phase. When a pSLS invocation request is made by the SP to the first NP in the E2E chain, all the NPs in the E2E chain must perform admission control to check whether this request can be accommodated. If this process is successful then the pSLS invocation request is accepted. In a similar manner, when a cSLS invocation request is made by CCs to the SP, the SP must perform admission control taking into account the pSLS pipe currently invoked resources and also request from the ANP to check the availability of resources in the 'last-mile' connections to the CCs.

The above discussion reveals that pSLS admission control functions must be implemented in all NPs along the E2E chain taking into account the available resources at each NP and at the interconnecting inter-domain links for the specific QoS class (that is, intra- and inter-domain traffic trunks (TTs)) whereas cSLS admission control functions must be implemented by the SP itself taking into account the currently invoked resources of the pSLS pipe and also by the ANP taking into account the available resources of the 'last-mile'.

In this paper we address pSLS admission control performed by NPs and cSLS admission control performed by the SP, but not by the ANP. For the time being we will assume that the 'last-mile' bandwidth is sufficient, therefore it does not affect the overall cSLS admission control decision.

3. Admission Control Schemes

3.1. cSLS Admission Control Scheme

As cSLS admission control scheme performed by the SP we adopt the measurement-based admission control (MBAC) scheme presented in [2]. This MBAC scheme employs the Gaussian effective bandwidth approach. Using the peak rate of the cSLS flow requesting admission, measurements of the current aggregate traffic load (mean rate, variance) at the entrance of the pSLS pipe and the target packet loss rate (PLR) that the SP wants to enforce, a bandwidth value $C_{required}$ can be derived, which is needed in order for the existing cSLS flows and the new one requesting admission to be served with the actually incurred PLR maintained in values lower than the target PLR. This target PLR value, enforced by the SP, can be derived based: a) on the total E2E PLR that the content to be delivered can sustain (and still be of the quality defined in the cSLA), b) on the PLR values that the NPs in the E2E chain have agreed to enforce for the subscribed pSLSes, and c) on the PLR that the ANP will guarantee at the access network part. This bandwidth value is compared with the currently invoked bandwidth value of the pSLS pipe $C_{pSLS-pipe-invoked}$. If $C_{required} \leq C_{pSLS-pipe-invoked}$ then the new cSLS flow is admitted.

This cSLS admission control procedure is triggered upon each cSLS invocation request. It is worth noting that the use of the peak rate of the cSLS flow requesting admission in the admission control decision does not mean that this cSLS flow -if admitted- will be allocated resources equal to its peak rate. For future admission control decisions of other cSLS flows, its real traffic contribution will be reflected in the aggregate traffic measurements

3.2. pSLS Admission Control Scheme

pSLS admission control is independent of the cSLS invocation process and is intended for aggregate traffic, i.e. the whole pSLSes aggregate traffic the particular TT carries. Our current pSLS admission control scheme is based on some simple metrics.

Each NP multiplexes within its TTs, distinct pSLS segments for pSLS pipes from upstream SPs/NPs. Instead of allocating fixed resources to each pSLS segment within the NP's TTs, our scheme assumes that there is a minimum value of bandwidth guaranteed for each pSLS segment $C_{pSLS-segment-\min}$ and a

maximum one $C_{pSLS-segment-\max}$, which can be reached through a pSLS update procedure triggered by upstream SPs/NPs. Initially, the minimum bandwidth is allocated to all pSLS segments. There is an amount of spare resources for each TT, managed by the NP, which are used to increase the capacity of the pSLS segments when needed. In order for our scheme to be cost-effective, we assume that the spare resources are *not* adequate for all the pSLS segments to reach their maximum capacity concurrently, assuming that when one pSLS segment is heavily loaded, there may be others that are under-loaded and can release resources.

The criterion for deciding when a pSLS segment should request release of resources or increase of its capacity is chosen to be very simple: the utilization of it. An upper threshold in utilization is set, so that when crossed, pSLS admission control is triggered by the upstream SP/NP to examine if some additional resources from the pool of TT spare resources can be allocated to this pSLS segment. Similarly, crossing a lower threshold in utilization can also trigger the pSLS admission control mechanism, this time to release some resources. In all cases, the invoked pSLS segment capacity should be maintained between its minimum and maximum value and we also define a step bandwidth value *S* by which the pSLS segment capacity can be increased or decreased at every pSLS update epoch. In order not to have the case where instantaneous crossings of the utilization thresholds lead to triggering of the pSLS admission control process, we define a period *T* over which the upstream SPs/NPs estimate their pSLS segment utilization and compare it with the upper and lower threshold.

Once the request for increase/decrease of the bandwidth allocation for a pSLS segment succeeds, the request for a similar increase/decrease is propagated to downstream NPs and if the admission control decision E2E is positive, then the bandwidth of the whole pSLS pipe is increased/decreased accordingly.

4. Performance Evaluation

To evaluate the joint operation of the cSLS/pSLS admission control schemes we use the network simulator *ns-2* [3] and the topology of Figure 2.

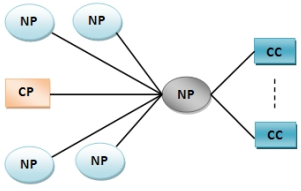

Figure 2. Simulation topology

In this topology, four upstream NPs and a CP are attached to the NP in question, which performs pSLS admission control. The CP generates streams upon CC requests, which are multiplexed with the traffic of the upstream NPs in the same TT of the NP. For the sake of simplicity and since we have assumed that the 'last-mile' bandwidth of the ANP is sufficient and does not affect the overall cSLS admission control decision, the CCs are connected at the edge of the NP. The SP, even though not shown in Figure 2, manages and directs the whole procedure of requests from CCs and delivery to CCs, performing cSLS admission control and triggering, together with the upstream NPs, the pSLS admission control procedure of the NP.

We assume that the cSLS requests from the CCs are for MPEG-4 encoded video clips. In our simulations we use three different levels of video quality in order to 'simulate' user preference as well as different terminal capabilities. To do so, three different trace files from [4] are used. The high quality trace has peak rate 3.1Mbps and average rate 0.58Mbps, the medium quality trace has peak rate 1.5Mbps and average rate 0.18Mbps and the low quality trace has peak rate 1.5Mbps and average rate 0.11Mbps. CCs send requests for one of the three quality video streams and we model the arrival processes of the cSLS requests as Poisson arrival processes. We also assume that the duration of the video clips is exponentially distributed with an average of 300sec (5mins) and that the target PLR the SP aims to enforce is 10^{-3}. This PLR value is low enough to allow for additional losses that downstream NPs as well as the ANP in a realistic E2E chain will incur (10^{-2} is roughly a representative upper limit for allowed E2E PLR for multimedia content [5]).

The upstream NPs inject already aggregated normally distributed Variable Bit Rate (VBR) traffic to the NP's TT. In order to simulate a variety of load conditions, the average rate of the upstream NPs' traffic varies between 20Mbps and 90Mbps and the arrival rates for the cSLS requests vary between 350 requests/hour and 1500 requests/hour and in a way in time so that we have: a) the situation where all of them (upstream NPs/SP) generate sufficient traffic demand to trigger the pSLS admission control of the NP, b) the situation where all of them underutilize their current bandwidth allocations within the TT and request release of resources, and c) intermediate situations.

Regarding the bandwidth allocations within the TT, all four upstream NPs and also the SP are allocated $C_{pSLS-segment-\min}$ equal to 50Mbps, which, as aforementioned, is hardly reserved. Also, the maximum bandwidth $C_{pSLS-segment-\max}$ that the

upstream NPs and the SP can request is 100Mbps and the bandwidth of the TT is set to 375Mbps, meaning that it is not possible for all pSLS segments within the TT to simultaneously reach the maximum allowed capacity. We also assume that the step *S* by which the pSLS segment capacity can be increased/decreased each time is 5Mbps, the minimum time period *T* that each upstream NP and the SP can trigger pSLS admission control is 1min and the utilization thresholds for triggering pSLS admission control are 0.25 (underutilization) and 0.50 (overutilization).

To demonstrate the benefits of deploying the joint cSLS/pSLS scheme we also simulate the scenario where no pSLS admission control takes place and that for cSLS admission control the SP implements the simplest possible peak rate admission control; that is the SP only adds the peak rate of the new cSLS request to the cumulative peak rate of the already invoked cSLSes and compares it with $C_{pSLS-pipe-invoked}$. In this scenario all pSLS segments are, from the beginning, allocated fixed bandwidth values equal to 75Mbps so that their sum is equal to the TT capacity.

The joint cSLS/pSLS scheme satisfies the PLR target (achieves a value of 0.93×10^{-3}), whereas the cSLS peak rate scheme achieves zero PLR, since it does not account for any statistical multiplexing. Figures 3 and 4 show the utilization of the SP's pSLS segment, the utilization of the NP's TT, and the cSLS rejection probabilities for the three quality levels.

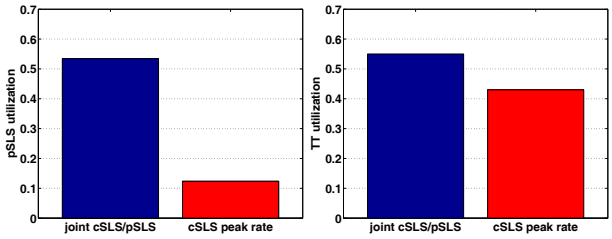

Figure 3. SP pSLS utilization (left) and NP TT utilization (right)

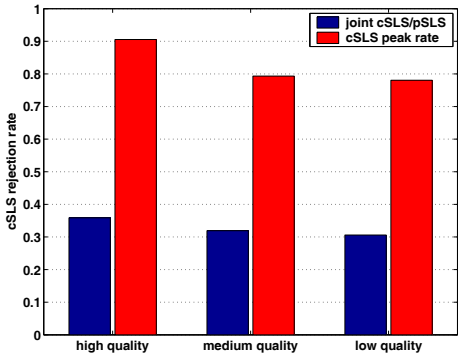

Figure 4. cSLS rejection rate

As it can be seen, the joint cSLS/pSLS scheme achieves significantly higher utilization compared to the cSLS peak rate scheme with respect to SP's pSLS segment utilization since a) the cSLS scheme in the former case allows for statistical multiplexing gains and also b) the pSLS scheme in the former case allows the dynamic allocation/release of resources. With respect to the NP's TT utilization, the joint cSLS/pSLS scheme achieves again higher utilization due to the higher upstream NPs/SP traffic contributions in the NP's TT. It is worth noting that the performance gap between the joint cSLS/pSLS scheme and the cSLS peak rate scheme would have been even higher if we hadn't set a minimum hardly reserved bandwidth value for each pSLS segment. Regarding cSLS rejection rates, the joint cSLS/pSLS scheme achieves much lower rejection rates compared to cSLS peak rate scheme for all three quality levels.

5. Conclusions and Future Work

In this paper we presented a joint cSLS/pSLS admission control scheme, using the ENTHRONE architecture as a reference point. We showed by means of simulations that the introduction of even simple admission control functions at per-flow and per-aggregate level are capable of providing significant resource utilization gains and reducing significantly service rejections while maintaining QoS within acceptable levels. In the future we will attempt to provide more complete and realistic solutions by investigating and incorporating cSLS admission control at the ANPs, considering more complex pSLS admission control schemes (e.g. see [6]) as well as more NPs in the E2E chain and more types of traffic for the cSLS requests

Acknowledgement. This work was undertaken in the context of the IST ENTHRONE phase 2 project.

6. References

[1] www.ist-enthrone.org.
[2] S. Georgoulas et al "Heterogeneous Real-time Traffic Admission Control in Differentiated Services Domains", IEEE GLOBECOM 2005.
[3] K. Fall et al "The *ns* manual" (www.isi.edu/nsnam/ns/ns_doc.pdf).
[4] http://www.tkn.tu-berlin.de/research/trace/ltvt.html.
[5] T. Chaded "IP QoS Parameters", TF-NGN, November 2000.
[6] E. Borcoci et al "Service Invocation Admission Control algorithm for Multi-domain IP Environments", Elsevier Computer Networks, November 2007.