Traffic shaping and dimensioning of an external overload controller in service architectures

Andersson, Jens; Nyberg, Christian; Kihl, Maria

Published in:
Proceedings. 2006 31st IEEE Conference on Local Computer Networks

DOI:
10.1109/LCN.2006.322165

2006

TRAFFIC SHAPING AND DIMENSIONING OF AN EXTERNAL OVERLOAD CONTROLLER IN SERVICE ARCHITECTURES

Jens Andersson, Christian Nyberg and Maria Kihl
Lund Institute of Technology, Sweden
{jens.andersson, christian.nyberg, maria.kihl}@telecom.lth.se

Abstract

This paper investigates the dimensioning of a server used for external overload control in a service architecture. Great savings can be obtained by an operator if the dimensioning analysis is performed correctly. As one of the main parts of this paper it is shown that Poissonian arrivals is a good assumption for some services in service architectures. Methods that can be used for dimensioning are presented and examples are provided.

1. Introduction

In the new generation of service architectures for telecommunication networks, access to telecom network capabilities, is provided from the Internet via a gateway architecture [4]. In Figure 1 an example of such an architecture is presented. By opening the telecommunication networks it is foreseen that the pace of development of new services will increase. Typically a Service Provider (SP) provides customers residing in the Internet with an application triggered by an HTTP request. The SP is connected to an Application Server (AS), which can be seen as a gateway to the telecommunication network. The AS translates the requests from the SP into telecom specific requests according to, for example, the Parlay standard, [4]. Each SP signs a Service Level Agreement (SLA) with the owner of the AS. The role of the SLA is to clarify which level of service the parties can expect from each other, for example, availability and maximum delay. To avoid violation of the SLA an Overload Control (OC) device should be used. The AS is equipped with an external OC, see [3]. The OC can choose either to reject or forward an incoming request. Either a request is rejected or forwarded the amount of processing is approximately the same. It is important that the OC device does not get overloaded. Therefore, a traffic restriction of how an SP can send its requests is agreed in the SLA. This restriction is equivalent to a Token Bucket (TB), see [6].

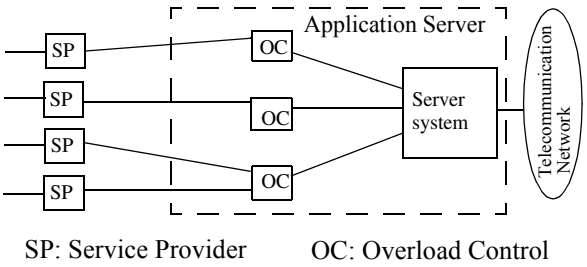

Figure 1. Example of a telecom web service gateway architecture

In this paper we investigate and propose methods for dimensioning an external OC device. To be able to perform our dimensioning analysis the arrival process to a service architecture is needed, see Section 2. Section 3 presents and compares two methods for analysing the departures from a TB. The dimensioning is based on the result of these analysis.

2. The arrival process

It is common that Poissonian arrivals are assumed to a service architecture. However, the assumption is usually not motivated with any measurements from a real system. We have been provided with a log from an operator of an Intelligent Network (IN) service architecture. The log was recorded over a time interval of thirty minutes, and contains request for a certain not media-stimulated application. The interarrival times between two consecutive arrivals have been calculated and these times have been fitted to different known distributions by using the distribution fitting tool in Matlab, see Figure 2. The exponential distribution is the smoother line which is not sampled and as seen it fits very well. This strengthens the assumption that poissonian arrivals to the service architecture is correct.

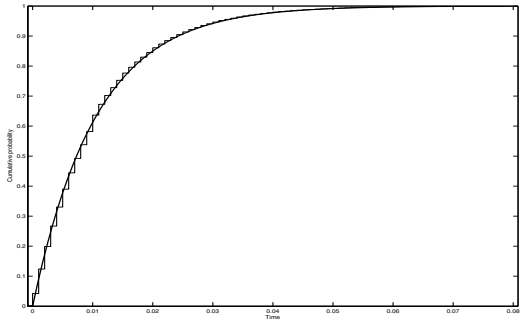

Figure 2. The time between arrivals from a real traffic log is fitted with the exponential distribution

3. SLAs and dimensioning analysis

We compare two methods to perform dimensioning analysis, namely; Queuing theory and Network calculus, [1], analysis. The latter of these is simpler to perform but also the more pessimistic method. By extending the results obtained in Sidi et al. [2], some results on the arrivals to the OC device based on queuing theory are obtained. For calculations and thorough descriptions, see [6]. The distribution function for the departures from a TB when assuming three different values for the arrival rate λ, (5, 10, 20) per second to the TB and token rate, *D*, set to 10/s is shown in Figure 3. In the plotted configuration the bucket size, *M*, equals 10 and the request queue size equals 4. Based on these results and with some further calculations the mean number of departures can be found, see Figure 4.

If the Network Calculus bound is considered the function describing the number of departures during interval *t* takes the shape of *M+t/D*.

The choice of time unit, arrival process and which measurement method that is used will all have impact on the dimensioning process. According to our contact with operators and suppliers, an example of a formulation of an SLA is: *If the requests sent are conformant to a TB with a burst size of M and a token rate of D tokens per time unit, the mean time until a request has been served is less than* t_0 *seconds.*
For evaluation of dimensioning methods for some other formulations, see [6].

The OC device can be modelled as an G/D/1 queue as it has deterministic service times. Queuing theory then gives the mean time in system, see [5]. When the worst case analysis is considered an estimation of the mean sojourn time in the OC device can be calculated as described in [6]. To meet a requirement of a mean time in the OC less than 0.1 the capacity should be set to a relative measure of 24 according to worst case analysis. Note that this is a worst case bound and independent of the arrival process. If poissonian arrivals are assumed

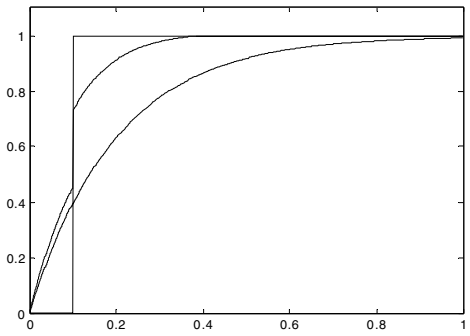

Figure 3. The distribution function plotted for λ=5, 10, 20 corresponding to the curves from the lower to the upper.

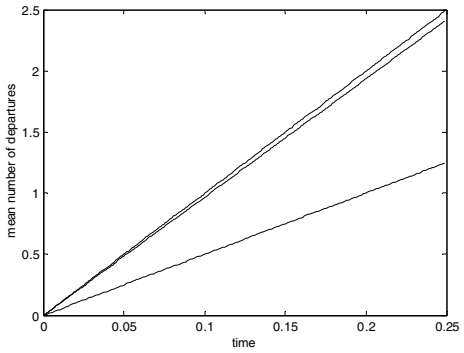

Figure 4. Mean number of departures calculated from the mean interdeparture time for λ=5, 10, 20

the mean time in system can be bounded by queuing theory. For λ set to 5, 10 or 20 this correspond to a capacity respectively set to 14, 17 and 18 for the constraint to be fulfilled. Compared to the worst case analysis this is a reduction of about 42%, 30% and 25% respectively.

4. Results and discussions

For correct dimensioning of an OC device, information about the arrivals is needed. We have analysed real logs of arriving requests to an IN service architecture and shown that they form a Poisson process.

Queuing theory and Network calculus were used to derive bounds of the required capacity at the OC device. Queuing theory results in tighter bounds but requires that the assumption about the arrival process is correct. The bounds derived with queuing theory are statistical bounds and can not be used when the SLA contains hard time constraints, which cannot be violated. Network calculus can give such upper bounds. By examples it is shown how using correct dimensioning analysis may result in great savings by the operators.

REFERENCES

[1] J-Y. Boudec and P. Thiran, *NETWORK CALCULUS A Theory of Deterministic Queuing Systems for the Internet*, LNCS 2050, Springer Verlag May 10, 2004

[2] M. Sidi, W-Z Liu, I. Cidon and I. Gopal, "Congestion Control Through Input Rate Regulation", IEEE Transactions on Communications, vol 41 nbr 3, March 1993

[3] M. J. Whitehead and P. M. Williams, "Adaptive network overload controls", BT Technology Journal, Vol 20 No 3, July 2002

[4] Parlay consortium, www.parlay.org

[5] L. Kleinrock, *Queueing systems, Volume I: Theory*, John Wiley & Sons, 1975

[6] J. Andersson, C Nyberg, M. Kihl, "Dimensioning of an external overload controller", Technical report, Lund University, Department of Communication System, 2006