Managing end-to-end resource reservations

Extended Abstract

Luis Almeida[1,2]
[1]IT - University of Porto
Porto, Portugal
lda@fe.up.pt

Moris Behnam[2]
[2]IDT - Mälardalen University
Västerås, Sweden
moris.behnam@mdh.se

Paulo Pedreiras[3]
[3]IT - University of Aveiro
Aveiro, Portugal
pbrp@ua.pt

Categories and Subject Descriptors

C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design—*Network communications, Distributed networks*

Keywords

Computer networks, Resource reservations, Scalability, Adaptability

1. CONTEXT

Currently, there is a strong push for real-time applications distributed over large geographical areas, fuelled by interactive multimedia, remote interactions, cloud-based time-sensitive services and many cyber-physical systems [3]. These applications rely on many different resources, from the end nodes where they execute, to the networks over which they communicate. However, supporting global compositionality in this realm, i.e., guaranteeing applications performance a priori independently of load variations, is a significant challenge. It calls upon resource reservations across the system that comply to performance metrics agreed at the time of application deployment, what is normally called Service Level Agreement (SLA). This requires well defined application resource requirements, from the computing end nodes, possibly including multiple internal resources, to the network.

In spite of existing technical solutions for enforcing reservations, from shaping in networks to virtualization in processors and reservations in static distributed systems [2], such solutions typically fall short on at least two aspects, scalability and adaptability. The former is needed to cope with large numbers of reservations to achieve the desired segregation and isolation among many applications, while the latter is needed to mitigate the effects of overprovisioning that micro-reservations can have in the efficiency of the whole system. These two requirements apply particularly to the network, given its central position in supporting the applications referred above and, in general, emerging paradigms such as the Internet-of-Things. Nevertheless, adaptability naturally extends from the network to the ends nodes, for the sake of efficiency, too.

On the other hand, the scalable solutions that we have today, e.g., deployed in current Internet, provide at best class-based Quality-of-Service (QoS) support [1]. This grants protection against interference from traffic of lower QoS delay tolerant classes but not within the same class, which is aggravated when the number of similar applications grows, e.g. VoIP calls.

Finally, supporting scalability and adaptability requires agile resource management. If the former points to a distributed management architecture, the latter is better achieved with a centralized one. This apparent conflict is typically dealt with using clustering. Local reservations are managed at the cluster level while end-to-end reservations require a distributed cluster-level protocol. An example of such approach is the Stream Reservation Protocol (SRP) in Audio-Video Bridges, but it is still class-oriented.

2. OPEN PROBLEMS

Therefore, given the considerations above, we herein formulate what we consider to be open problems to achieve the desired compositionality of distributed real-time applications in large open systems, in a resource efficient way. Given a distributed real-time application that will execute in N end nodes connected to a large network, how to:

- Formulate its resource requirements and interfaces?
- Express adaptivity in such requirements/ interfaces?
- Support scalable and adaptive network reservations?
- Analyze the requirements feasibility?
- Carry out global admission control and enforce the needed reservations?
- Track and distribute slack?

3. ACKNOWLEDGMENTS

With support from the Portuguese Gov. through FCT grants CodeStream (PTDC/EEI-TEL/3006/2012) and ServCPS (PTDC/EEAAUT/122362/2010).

4. REFERENCES

[1] G. Bertrand, S. Lahoud, M. Molnar, and G. Texier. Qos routing and management in backbone networks. In *Intell. QoS Tech. and Network Manag.t: Models for Enhancing Comm.*, pages 138–159. IGI Global, 2010.

[2] N. Serreli, G. Lipari, and E. Bini. Deadline assignment for component-based analysis of real-time transactions. In *CRTS 2009 Proceedings*, December 2009.

[3] F. Xia, L. Ma, J. Dong, and Y. Sun. Network qos management in cyber-physical systems. In *ICESS 2008 Proceedings*, pages 302–307. IEEE, July 2008.