On overloaded queues

Alain Jean-Marie

INRIA and LIRMM-University of Montpellier II,
161 Rue Ada, F-34292 Montpellier Cedex
ajm@lirmm.fr

Consider a queue with infinite storage capacity. When the rate of the input traffic into this queue is larger than the maximal service rate of the server, the queue is in overload. This situation is undesirable since it causes the queue to grow indefinitely, thereby making the quality of service delivered to clients worsen as time passes. Accordingly, classical queuing theory (and practice) is mostly concerned with stable, or underloaded, queues.

But queues under overload may have interesting properties. This depends quite a lot on the service discipline. For instance, in [1, 2], it has been found that the output rate of a Processor Sharing queue in overload, measured in customers per unit time, is not a constant and actually depends on the input rate and the Laplace transform of the service time distribution. Initially obtained as a curiosity in queuing theory, this result has found applications in the analysis of the Internet traffic, thanks to the combination of facts: the network is quite often in overload, and it tends to serve long-lived information flows in a Processor-Sharing fashion [3]. Other applications for queues in overload have been for web applications [4].

This talk will review properties of the output process and the waiting times in queues under overload, for various service disciplines. For the Processor Sharing queue, we will show the effect of overload on the discrimination of customers through their service time, the "elephants *vs* mice" problem. Other issues such as the finiteness of the expected conditional response time will be discussed [5].

References

[1] S. Yashkov, "On a heavy traffic limit theorem for the M/G/1 processor sharing queue", *Commun. Statist. - Stochastic Models 9*, **3**, 467–471, 1993.

[2] Alain Jean-Marie and Philippe Robert, "On the Transient Behavior of the Processor-Sharing Queue", *QUESTA*, **17**, 129–136, 1994.

[3] Thomas Bonald and Jim Roberts, "Congestion at flow level and the impact of user behaviour", *Computer Networks*, **42**, 521–536, 2003.

[4] Nikhil Bansal and Mor Harchol-Balter, "Analysis of SRPT Scheduling: Investigating Unfairness", *Proc. ACM Sigmetrics 2001*, Cambridge, Massachusetts, USA, 2001.

[5] Edward G. Coffman, Jr, Richard R. Muntz and H. Trotter, "Waiting Time Distributions for Processor-Sharing Systems", *JACM*, **17**, 1, pp. 123–130, january 1970.