TOP (2014) 22:458–459
DOI 10.1007/s11750-014-0329-9

DISCUSSION

Comments on: Queueing models for the analysis of communication systems

Harry G. Perros

Published online: 10 May 2014

Over the last 30 years, Dr. Bruneel's team has produced a continuous stream of high-quality papers, which focus on the use of generating functions in the analysis of complex queueing systems that arise in communication networks.

A crucial assumption in the use of generating functions is that time is slotted. That is, the queueing system under study has to be analyzed in discrete time. The implication of this is that in the paper, the authors have to assume that all packets are of fixed size; each packet takes one slot to arrive and one slot to get serviced. This assumption arises naturally in the case of ATM networks, which were introduced towards the end of the 1980s. In an ATM network, a packet is of fixed size and it is equal to 53 bytes. Consequently, it makes sense to depict ATM networks using discrete-time queueing theory. However, the predominant technology at this time is the IP network, where packets are of variable length, typically bounded by an upper bound of 1,500 bytes. Consequently, the results presented in the paper are not directly applicable, since a queueing system that represents a buffer in a router is a continuous-time system. Having said that, there are cases where we can safely assume fixed-size IP packets. For instance, this would occur if we download a large file. The file will be typically split into a number of fixed-size IP packets equal to the maximum allowed, with the exception of the last packet (which of course can be safely assumed that it is also of the maximum packet length). This situation arises in HTTP adaptive streaming. This is a new scheme for downloading video from a server, whereby the client controls the rate at which the server transmits the video stream based on its understanding of the network congestion, as evidenced by the rate of arrivals at the destination and

This comment refers to the invited paper available at doi:10.1007/s11750-014-0330-3.

H. G. Perros (✉)
Raleigh, USA
e-mail: hp@csc.ncsu.edu

buffer thresholds. In this case, a video is split into fixed intervals of 5 s (other interval durations are also possible). Within each interval, the video content is transmitted in IP packets with the maximum allowed length. Another example occurs in optical burst switching, where depending on the algorithm for constructing bursts, we may well end up with fixed length optical packets. However, in general, I would venture to say that this is a restricting assumption.

The authors introduce a well-constructed arrival process consisting of messages that are split up in fixed-size packets, and the introduced correlation makes it a very powerful and versatile process. The ensuing analysis of obtaining the pgf of the delay in the low class is well-constructed and at times intimidating!

The analysis is focused on a single queue. Of course, communication networks consist of a multiplicity of queues and I was wondering how can this work be extended to a tandem network of queues. How can one describe the departure process from one queue?

Expression 26 is quite elegant, but eventually it would have to be solved so that one can obtain numerical results. How easy is to solve it numerically and what are the possible instabilities in the numerical calculations?

I was pleased to see that the authors have obtained an expression for the tail probabilities. Too much emphasis is placed on the use of averages when defining SLAs, but percentiles are a more realistic performance metric as it provides a statistical upper bound on the metric. A given percentile of the delay can be obtained by adding up the individual probabilities for each m. However, is this an effective way? Is it possible that one can construct an expression for computing percentiles directly?

Overall, I enjoyed reading the paper, as I have done so many times in the past!!