Delay Analysis of a Place Reservation Queue

Bart Feyaerts

Supervisor(s): Sabine Wittevrongel

I. Introduction

Modern communication networks have to support an increasingly diverse range of applications, each bearing their own particular set of Quality of Service (QoS) requirements. Real-time applications (e.g. multimedia streaming, telephony, gaming, ...) are characterized by hard delay requirements: the mean delay and the delay jitter of the data packets should typically be kept minimal. For non-real-time applications (e.g. www, ftp, e-mail, ...) on the other hand, delay requirements are typically less stringent, but packet loss should be (very) low. In short, the network should differentiate between the various types of traffic, in order to meet up with the traffic types' QoS requirements. Over the last decades, many solutions have been proposed, varying in complexity and effect. A common method of supporting differentiated QoS, is the use of priority scheduling in the network nodes, opposed to standard First-In First-Out (FIFO) scheduling. In what follows, we will divide all network traffic in two separate classes: class 1 will cover delay-sensitive traffic, class 2 consists of delay-tolerant traffic.

In FIFO, all packets are treated equally, regardless of their requirements. The other extreme is posed by the Absolute Priority (AP) scheduling mechanism, where different traffic types are mapped to different priority levels, and transmission priority is always given to packets of the highest priority level. In this particular case, we will consider class-1 traffic to be high-priority and class-2 traffic to be low-priority. When the server of an AP system becomes available, the first class-1 packet, present in the queue, if any, is served first. Class-2 packets can then only enter the server in the absence of class-1 packets. AP has been studied extensively for various cases. It has been shown that AP does indeed decrease high-priority packet delay at the expense of low-priority traffic. This may lead to excessive low-priority packet delays or so-called starvation.

To counter the problem of packet starvation, a new priority scheduling mechanism was proposed in [1], called the Reservation Discipline (RD). With RD, a dummy packet, the reserved place R, is inserted in the queue, to act as a placeholder for future class-1 packets. As a class-1 packet arrives at the system, it is inserted at R's position and R gets reinserted at the end of the queue. Class-2 packets are simply inserted at the end of the queue, just as they would have been under FIFO scheduling. If multiple packets arrive at the system during the same time slot, first the class-1 packets are inserted one by one and then the class-2 packets are appended in arrival order. This way, class-1 packets are favoured twice: not only are they inserted before the class-2 packets that arrived during the same slot; the first class-1 packet of each slot is inserted at R's position, allowing the packet to jump over some class-2 packets. Since any class-2 packet can be jumped over only once, there is a limit to the disadvantage sustained. Research has been done before for service times of exactly one slot, see [2]. We have studied RD in case of geometric and general service times, see [3] and [4] respectively.

II. Methodology

Although we would eventually like to find values for delay probabilities, we try to avoid the direct use of probabilities. Fact is that

calculations that involve probabilities easily become complicated. As an example, we introduce two independent random variables X and Y, of which we want to determine $\mathrm{Prob}[X+Y=n]$. Even the direct calculation of this simple probability requires the use of a convolution. A PGF (probability generating function) approach keeps the calculations concise. The PGF of a random variable R (in what follows shortened to $R(z)$) is defined as $E\left[z^R\right]$, where $E[.]$ denotes the expected value of the quantity between brackets. Thus, the PGF of $X+Y$ can easily be obtained as $E\left[z^X\right]E\left[z^Y\right]$, which only involves a product but no convolution.

PGFs possess some useful properties. One of these is the moment generating property, which allows us to recursively calculate the consecutive moments of a random variable, using

$$\left.\frac{\mathrm{d}^n R(z)}{\mathrm{d}z^n}\right|_{z=1} = E\left[\prod_{i=0}^{n-1}(X-i)\right].$$

In practice this implies that the mean $E[R]$ of a random variable R is given by $R'(1)$, the first derivative of the PGF at $z=1$. The variance of that random variable R can then be calculated as $R''(1)+R'(1)-R'(1)^2$. Moreover, if we are interested in the probabilities itself, we can apply various inversion techniques for PGFs.

The first step in the analysis was to create an appropriate mathematical model to fit the system. The model involves a discrete-time single-server queueing system and two packet streams of which the numbers of arrivals from slot to slot are independent and identically distributed (*iid*). The transmission times of the packets are *iid* from packet to packet and have a general distribution. Based on this model, we composed a vector of random variables that allows us to adequately monitor the system state and to derive the desired performance measures. These variables were:

- the number of slots until service completion of the packet in the server (if any);
- the number of information packets in the system;
- the position of R in the queue.

This so-called *system state vector* was found to be Markovian, meaning that the distribution of the system state in a random slot can be calculated solely from the system state in the preceding slot. The relation between the system states in any two consecutive slots is given by the system equations. We then established a relationship between the joint PGFs of the system states in two consecutive slots by means of the system equations. From this relation, we then derived the joint PGF of the system state during a random slot in the steady state.

By means of the system state, expressions could also be found for the delays of a random class-1 and class-2 packet arriving during a slot in equilibrium. We then calculated the PGFs of the delays, allowing us to easily construct expressions for the mean values and the variances of the delays of both packet classes.

III. Conclusions

Numerical tests and simulations have proven that RD effectively solves the problem of starvation. There is, however, an important payoff: the class-1 delay performance decreases significantly, in relation to the class-1 delay performance under AP.

References

[1] W. Burakowski and H. Tarasiuk, On new strategy for prioritising the selected flow in queuing system, *Proceedings of the COST 257 11th Management Committee Meeting* (Barcelona, January 2000), COST-257 TD(00)03

[2] S. De Vuyst, S. Wittevrongel and H. Bruneel, Place reservation: delay analysis of a novel scheduling mechanism, *Computers & Operations Research*, vol. 35, no. 8, pp. 2447–2462 (2008)

[3] B. Feyaerts, S. De Vuyst, S. Wittevrongel and H. Bruneel, Analysis of a discrete-time priority queue with place reservations and geometric service times, *Proceedings of the Sixth Conference on Design, Analysis, and Simulation of Distributed Systems, DASD 2008* (Edinburgh, June 2008)

[4] B. Feyaerts and S. Wittevrongel, Performance analysis of a priority queue with place reservation and general transmission times, *Proceedings of the Fifth European Performance Engineering Workshop, EPEW 2008* (Mallorca, September 2008)