Real-time vs. Data Traffic: a DiffServ Performance Analysis

Thomas Demoor

Supervisor(s): Joris Walraevens, Dieter Fiems, Herwig Bruneel

I. Motivation

The huge difference between Quality of Service (QoS) demands for real-time traffic flows and data traffic flows is often neglected in packet-based telecommunication networks, such as the Internet. Real-time traffic, such as multimedia streams, can often endure some packet loss but requires low delays and/or low delay jitter. Data traffic benefits from low packet loss, hence avoiding retransmissions, but has less stringent delay characteristics.

In the nodes (routers, etc.) of the network, packets typically have to wait before being transmitted to the next node and a queue is present in order to preserve waiting packets. In most networks, all traffic is handled analogously. Packets are transmitted in order of arrival with acceptable QoS because the output bandwidth is underused. However, when bandwidth is scarce (accces networks, wireless links, future Internet) we are forced to handle packets smarter in order to achieve acceptable QoS. Differentiated Services [1] (DiffServ) is one of these smarter approaches. It classifies packets according to their QoS requirements. By basing the order in which packets are transmitted on class-dependent priority rules, QoS can be optimized over all traffic classes.

This article studies a system with a single server serving two queues, one per priority class, and an Absolute Priority scheduling algorithm . This is the most drastic scheduling method, as low-priority packets (class 2) are only served if there are no high-priority packets (class 1) in the system, but it is practical and easy to implement. In DiffServ terminology one would say that class-1 traffic has Expedited Forwarding Per-Hop Behaviour (PHB) and class-2 traffic has Default PHB.

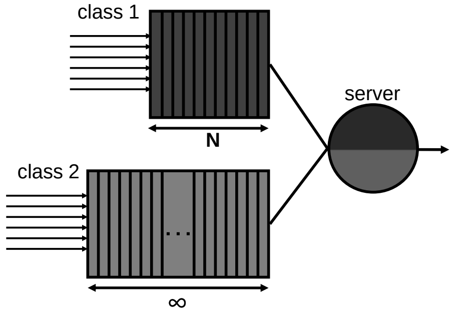

Figure 1. Formal system layout.

Although practical queues are of course finite, analytic studies of queueing systems generally assume infinite queue capacity facilitating the mathematical analysis of the system. In general, this assumption does not considerably alter the results as the studied queues have a sufficiently large capacity in practice. However, class-1 packets are delay-sensitive. It is hence useless to queue too many of these packets concurrently. Furthermore, because they have priority, class-1 packets would also monopolize the server causing starvation of class-2 traffic. Therefore, class-1 queue capacity should be as small as possible while still meeting the required packet loss constraints for this

T. Demoor is with the Department of Telecommunications and Information Processing, Ghent University (UGent), Gent, Belgium. E-mail: thdemoor@telin.ugent.be .

traffic and hence finite (exact) class-1 queue capacity is to be preferred above the infinite capacity approximation. On the other hand, the loss-sensitivity of class-2 packets results in a class-2 queue capacity as large as practically feasible, justifying the assumption of an infinite class-2 queue capacity. Hence, we limit the capacity of the class-1 queue to N packets and assume that the class-2 queue has infinite capacity. The system is depicted in figure 1.

II. Methodology

Time is divided into fixed-length slots corresponding to the transmission time of a packet. The server can handle one packet simultaneously and service takes one slot (deterministic). Class-1 packets are served with absolute priority over class-2 packets. Within a class the queueing discipline is First In First Out (FIFO). Packets can only enter the server at the beginning of a slot, even if arriving in an empty system. Assume that for both classes the number of arrivals in consecutive slots form a sequence of independent and identically distributed random variables. Let $a_{i,k}$ be the number of class-i ($i = 1, 2$) packet arrivals during slot k. The arrivals of both classes are characterized by the joint probability mass function

$$a(m, n) = \text{Prob}[a_{1,k} = m, a_{2,k} = n] \,. \quad (1)$$

This allows the arrivals of both classes to be correlated within a slot. As the class-2 queue has infinite capacity, we assume that all class-2 arrivals are accepted into the system. On the other hand, the class-1 queue can contain up to N packets. The queue management algorithm (QMA) determines if a class-1 arrival is accepted into the system or rejected.

We investigate the evolution of the system. The performance measures of interest are the class-i ($i = 1, 2$) system content u_i, i.e. the amount of class-i packets contained by the system at the beginning of a slot, the class-i delay d_i, i.e. the number of slots a random class-i packet resides in the system, and the class-1 packet loss ratio. The evolution of the system from slot to slot is illustrated in figure 2.

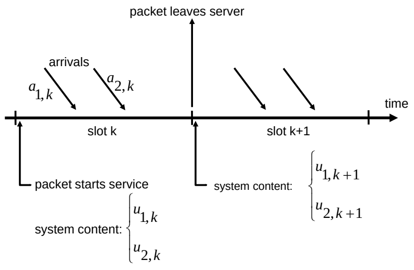

Figure 2. Evolution in time from slot k to slot k+1

An analytical solution technique yields probability generating functions (z-transforms) enabling us to determine all moments (mean, variance, etc.) of the packet delay and system content for both classes as well as the class-1 packet loss. This technique involves inversion of matrices of size N, the class-1 queue capacity, and hence the computational effort increases with increasing N. The improvement over a model where both queues have infinite capacity [2] is investigated.

III. Conclusions

Previous models assumed that the system had infinite capacity for all traffic classes but in a DiffServ router the capacity for real-time packets is typically small to prevent real-time traffic from monopolizing the system. The presented model takes the exact queue capacity into account increasing the accuracy of the results and allowing determination of high-priority packet loss and its influence on the performance of the system.

References

[1] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang and W. Weiss *An Architecture for Differentiated Services*, IETF RFC 2475, 1998.

[2] J. Walraevens, B. Steyaert and H. Bruneel, *Performance analysis of a single-server ATM queue with a priority scheduling* Computers & Operations Research, vol. 30, no. 12, 2003, pp. 1807-1829.