

Approximating Low Latency Queueing Buffer Latency

Martin J. Fischer, Ph.D.

Denise M. Bevilacqua Masi, Ph.D.

John F. Shortle, Ph.D.

Low Latency Queueing (LLQ) is an Internet Protocol (IP) router discipline that is being used to ensure that performance-sensitive high priority traffic, such as voice and video, receive their high level of performance, while allowing less performance-sensitive traffic, such as email or best-effort IP, to receive some portion of the bandwidth. In this paper, we develop a simple analytic approximation for the buffer latency (expected buffer delay) for each traffic class using the LLQ system. The approximation is validated via a simulation model.

Introduction

Industry is moving away from circuit-switched technology to Internet Protocol (IP) technology for telecommunications applications, including voice traffic. For example, the National Communications System manages several emergency telecommunications programs for federal government users. They are investigating the evolution of these networks toward IP capability, particularly in regard to ensuring continuity of priority traffic during emergencies. When IP networks become overloaded, packets get dropped and other quality of service (QoS) measures, such as packet latency and jitter, are significantly degraded. At some point, the QoS for voice and video packets becomes poor enough that their communication is lost. As Voice over IP (VoIP) and video become more prevalent in these networks, these issues become increasingly important—particularly since there is no dedicated communication path for a voice or video call.

There are several different mechanisms available that attempt to provide QoS in an IP network. [1, 2] Packet traffic can simply be handled on a First Come First Served (FCFS) basis. With a high enough bandwidth and under normal traffic conditions, this can be sufficient. [3, 4, 5] In priority queueing (PQ), higher priority real-time traffic (e.g., VoIP traffic) is transmitted before lower priority traffic (e.g., data traffic), with separate buffers for each class of traffic. Weighted Fair Queueing (WFQ) allocates the bandwidth fairly to network data traffic using weights to determine the amount of bandwidth allowed for different flows of traffic. Class-Based Weighted Fair Queueing (CBWFQ) extends weighted fair queueing to multiple user-defined traffic classes, rather than individual flows of traffic. Under CBWFQ, a certain portion of the bandwidth is set aside for each of the classes; when one class of traffic is not utilizing its bandwidth, the other class is allowed to overflow and use the bandwidth.

Low Latency Queueing (LLQ) combines PQ and CBWFQ and is being used frequently on the Internet with these multiple classes. The term *Low Latency Queueing* or *LLQ* is widely accepted and used within the IP community. It is apparent that there are many options available; modeling is essential to determine which QoS mechanism is most appropriate in advance, rather than using a trial and error approach on a real network. Under the assumption of Poisson packet arrivals, analytic queueing results are available for FCFS and PQ, but not for CBWFQ or LLQ. Reference [6] presents a summary of available results. In practice, there would likely be several classes of data traffic in CBWFQ. LLQ combines CBWFQ with PQ, and typically voice traffic is assigned to the higher priority queue (the Expedited Forwarding [EF] traffic class) than the data classes which use CBWFQ. In the future, there may be greater than one class of EF traffic. Requests to the Internet Assigned Numbers Authority (IANA) for additional EF classes, which could be used for disaster response VoIP or video traffic, have been submitted to the Internet Engineering Task Force (IETF). [7, 8] The CBWFQ traffic is normally referred to as Assured Forwarding (AF) and Best Effort (BE) traffic.

Figure 1 presents the configuration of LLQ. LLQ is composed of PQ and CBWFQ modules. The traffic that is critical and needs to be served is typically assigned to the PQ module. Within that module, service is rendered on a priority basis by the assigned priority of each traffic class. Service is given to the highest traffic class that is present, and there is no service interruption. Under CBWFQ, the bandwidth is shared in accordance to the weights assigned to the CBWFQ traffic classes. These weights are designed to ensure the traffic class gets a certain portion of the available bandwidth. The CBWFQ traffic is never selected for service when there is PQ traffic waiting to be served.

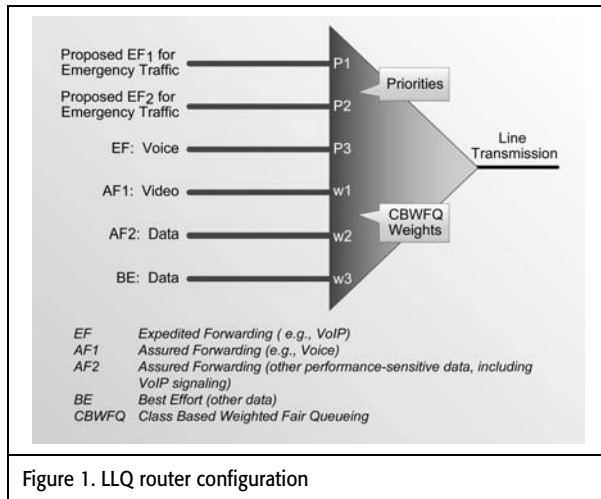


Figure 1. LLQ router configuration

LLQ has the ability to allow critical traffic (voice and video) to receive priority service (PQ module), while sharing the residual bandwidth among the CBWFQ traffic and assuring there is bandwidth that is made available to all traffic. The structure of LLQ allows surges in one traffic class to not dominate the use of the bandwidth. This is immediate in the CBWFQ module, and is controlled by the buffer sizes and by Call Admission Control in the PQ module.

Usually there are five or six traffic classes that utilize an LLQ router. These are assigned to one of the classes using the system. The traffic assigned to the PQ module is referred to as EF traffic. The traffic assigned to the CBWFQ module is referred to as AF and BE traffic. Each of the EF, AF, and BE traffic streams has its own buffers. There is no preemption of service in the LLQ system. When a packet completes service, the router looks at the PQ traffic first, and if there are packets waiting in the EF1 buffer, the first one is selected for transition. If no EF1 packets are waiting, it looks at the EF2 buffer and so on. If there are no PQ packets waiting for service, the router turns to the CBWFQ module. Within each PQ class, service is rendered on an FCFS basis.

Service in the CBWFQ module is different than in the PQ module. In that module, packets are selected for service based on the class weights. There are several versions of the CBWFQ rule [9]—a rule associated with Golestani [10] and a random selection rule we define. In this paper we consider the random rule. For the numerical experiments considered in this paper, the random rule is an upper bound on the Golestani rule as identified in reference 11. The goal of any CBWFQ rule is to ensure each AF or BE traffic class receives a share of the free bandwidth in proportion to its weight, if the traffic demand requires it.

Under the LLQ system, traffic, such as voice and video that require real time service, is typically assigned to the PQ

module. The traffic that does not require such a high QoS is assigned to the CBWFQ module. The usual placement of LLQ routers is at the edges of an IP network on lower speed lines. [4, 5]

In this paper we develop an analytically simple approximation for the expected buffer delay for each traffic class. It is based on the standard non-preemptive priority queueing (NPPQ) model. That model, and its use to date in modeling the LLQ system, is discussed in the following section. We then discuss the approximation and numerical comparisons with simulation results, followed by some concluding remarks.

The non-preemptive priority queueing model

We conducted a simulation analysis of the LLQ system and observed, among other things, that the NPPQ model plays a very important role in the modeling of the LLQ system. [11] Table 1 is a summary of the observations we drew from that analysis.

Table 1. LLQ simulation observations
Observations
PQ traffic performance is invariant under the random and Golestani CBWFQ rules
PQ traffic analytics are given by NPPQ model
Random CBWFQ is an upper bound on Golestani CBWFQ
Random CBWFQ average buffer latencies equal standard M/G/1 results
CBWFQ module cannot be decoupled from PQ module
CBWFQ performance at boundaries, monotonic behavior, and points of equality

The second observation states that the NPPQ model plays a very important role in developing analytic models for the LLQ system. First, the NPPQ model can be used to give the expected buffer delay for the PQ traffic and it does not matter what rule under which the CBWFQ module is operating. At the boundaries of the CBWFQ (weights equal to 0 or 1), the LLQ system behaves as an NPPQ system. We extend that result in the next section by showing that the complete LLQ system can be approximated by a NPPQ model.

The LLQ system which includes emergency traffic streams is composed of five or six classes of traffic with three classes being served by the PQ module and the remaining two or three classes by the CBWFQ module. We assume that the arrival process of each class of traffic is Poisson with rate λ_i and that the first two moments of the service time for class i are ES_i and ES_i^2 . The service distribution for each class can be general. Let

the class load be $\rho_i = \lambda_i ES_i$ and the total load be $\rho = \sum_{i=1}^J \rho_i$,

where J is the number of classes. We are looking at the LLQ system in steady state and assume $\rho < 1$. The normal independence assumptions among the classes and within each class are also assumed. We also assume each class has an infinite buffer.

The standard NPPQ system with J classes is one where the priority order is class 1, then class 2, then class 3, and so on. There is no service preemption and the server serves all waiting class 1 packets before serving class 2 and so on. Within a class, service is on an FCFS basis. If $W_q(i)$ is the expected buffer delay for the class i packet, then we have

$$W_q(i) = \frac{\sum_{j=1}^J \lambda_j ES_j^2}{2(1-\sigma_{i-1})(1-\sigma_i)}, \text{ where } \sigma_i = \sum_{j=1}^i \rho_j \text{ with } \sigma_0 = 0. \quad (1)$$

This result was first given in the 1950s. [12, 13, 14] Equation (1) will play a central role in what follows.

Returning to the LLQ system, we assume there are six classes ($J = 6$) and the first three are PQ traffic, then for $i = 1, 2, 3$ using the results presented in reference [11], we have

$$W_q(i) = \frac{\sum_{j=1}^6 \lambda_j ES_j^2}{2(1-\sigma_{i-1})(1-\sigma_i)}, \text{ where } \sigma_i = \sum_{j=1}^i \rho_j \text{ with } \sigma_0 = 0. \quad (2)$$

The numerator in $W_q(i)$ is summed over all six classes including the CBWFQ traffic, and accounts for the fact that there is no service interruption of packets from any class. For the PQ traffic, we need to only have the sum of the first three loads be less than one. Results exist for the case where one of the PQ packet classes can have loads greater than one. [12] We are not considering that case here. There are more analytic results available for the NPPQ system, but here we are dealing with buffer latency.

The approximate LLQ model

We need to develop an approximation for the expected packet delay for the CBWFQ traffic. Let $\alpha_i, i = 4, 5, 6$ be the probability that the class i packets are chosen for service if there are no PQ packets waiting and there are packets from each of the CBWFQ classes waiting. They are tied to the class i weights, $w_i, i = 4, 5, 6$, as shown earlier in Figure 1. The probabilities, $\alpha_i, i = 4, 5, 6$, are given by

$$\alpha_i = \frac{w_i / ES_i}{\sum_{j=4}^6 w_j / ES_j}. \quad (3)$$

Equation (3) reflects the fact that one cannot use the weights w_i directly because those weights are the percent of time the packet would occupy the server (i.e., the percent of bandwidth for the packet class). Dividing by the expected service time and normalizing gives us the desired probabilities that the class i packets are chosen.

As motivation for the approximation, consider the situation where there are only two CBWFQ classes, say 4 and 5. In this case, we have $\alpha_5 = 1 - \alpha_4$; so if $\alpha_4 = 0$, then $\alpha_5 = 1$. In this case, the system behaves as an NPPQ system with five classes where the priority order is 1, 2, 3, 5, 4. Equation (2) can be extended to account for the fact that class 5 has priority over class 4 with a slight modification in the definition of σ_j .

The central idea of this paper is to approximate an LLQ system as a NPPQ system where arriving CBWFQ packets are redistributed into priority classes *at packet arrival times*. Specifically, for the 6-class LLQ system considered in this paper (with 3 PQ classes and 3 CBWFQ classes), the approximating NPPQ system has 6 classes labeled 1, 2, 3, 4*, 5*, and 6*. The PQ packets are handled in the same manner as in the original LLQ system. The CBWFQ packets, on the other hand, are assigned to one of the priority classes 4*, 5*, and 6* at the arrival times of the packets. Aside from the process of assigning packets to priority classes, the system behaves like a standard NPPQ system.

To motivate the probabilities for redistributing the CBWFQ packets, consider the following scenario. Suppose that three packets—a class 4 packet, a class 5 packet, and a class 6 packet—arrive simultaneously to an empty system. Under the LLQ system, the probability that the packets are served in the order i then j then k ($i \neq j \neq k$) is

$$P(i, j, k) = \alpha_i \frac{\alpha_j}{(\alpha_j + \alpha_k)}. \quad (4)$$

To achieve the same ordering of packets under the approximate NPPQ system, we assign the three packets to the priority classes 4*, 5*, and 6* according to these probabilities (that is, $P(i, j, k)$ is the probability that i is assigned to 4*, j is assigned to 5*, and k is assigned to 6*).

More generally, for the approximating NPPQ system, a class 4 packet is assigned to class 4* with probability $P(4, 5, 6) + P(4, 6, 5)$, to class 5* with probability $P(5, 4, 6) + P(6, 4, 5)$, and to class 6* with probability $P(5, 6, 4) + P(6, 5, 4)$. Analogous assignments are made for class 5 and class 6 packets.

Let ρ_i^* be the load for class i^* in the approximate system. Then, we have

$$\begin{aligned}\rho_4^* &= \rho_4[P(4,5,6) + P(4,6,5)] + \rho_5[P(5,4,6) + P(5,6,4)] + \\ &\quad \rho_6[P(6,4,5) + P(6,5,4)] \\ \rho_5^* &= \rho_4[P(5,4,6) + P(6,4,5)] + \rho_5[P(4,5,6) + P(6,5,4)] + \\ &\quad \rho_6[P(4,6,5) + P(5,6,4)] \\ \rho_6^* &= \rho_4[P(5,6,4) + P(6,5,4)] + \rho_5[P(4,6,5) + P(6,4,5)] + \\ &\quad \rho_6[P(4,5,6) + P(5,4,6)]\end{aligned}\quad (5)$$

For $i = 4^*, 5^*, 6^*$, let $Wq^*(i)$ be the expected buffer delay in the approximate NPPQ system, then

$$Wq^*(i) = \frac{\sum_{l=1}^6 \lambda_l ES_l^2}{2(1 - \sigma_{i-1})(1 - \sigma_i)}, \text{ where } \sigma_i = \sum_{j=1}^3 \rho_j + \sum_{j=4}^i \rho_j^* \text{ with } \sigma_3 = \sum_{j=1}^3 \rho_j \text{ and } i = 4^*, 5^*, 6^*. \quad (6)$$

We can now approximate the classes 4, 5, and 6 expected buffer delay via

$$\begin{aligned}Wq(4) &= Wq^*(4)(P(4,5,6) + P(4,6,5)) + Wq^*(5)(P(5,4,6) + \\ &\quad P(6,4,5)) + Wq^*(6)(P(5,6,4) + P(6,5,4)) \\ Wq(5) &= Wq^*(4)(P(5,4,6) + P(5,6,4)) + Wq^*(5)(P(4,5,6) + \\ &\quad P(6,5,4)) + Wq^*(6)(P(4,6,5) + P(6,4,5)) \\ Wq(6) &= Wq^*(4)(P(6,4,5) + P(6,5,4)) + Wq^*(5)(P(4,6,5) + \\ &\quad P(5,6,4)) + Wq^*(6)(P(4,5,6) + P(5,4,6)).\end{aligned}\quad (7)$$

For the PQ traffic classes, we use the $Wq(i)$ given by Equation (2).

Numerical examples

In this section we compare the approximation presented above with the results of simulation models we have developed. [9, 15] We look at two examples—one with five classes and the other with six. In both cases, there are three PQ traffic classes. From the discussion above, the NPPQ model given by Equation (2) is exact, so comparisons will only be made to the performance of the CBWFQ classes of traffic.

We assume that the packet arrival process for each traffic class is a Poisson process independent of the other classes. We allow the packet-size distribution of each class to be general; in the example considered here, they are shown in Table 2.

Classes 1 and 4 have deterministic packet size distributions, classes 2 and 5 exponential, and classes 3 and 6 D_3 . This distribution is one where the packets come in one of three sizes each with a certain probability. The specifics of D_3 are presented in Table 3.

The D_3 probability distribution represents the types of sizes one sees on the Internet. [16] We assume each class has an infinite buffer and the line speed is a T1 rate (equal to 1,536 kbps). Two total loads for the five class examples were looked at—

0.73 and 0.83; recall we are only considering cases where the total load is less than one. Table 4 presents the loads for each of the five classes. The packet arrival rate is given in terms of packets per second.

Class	Service Distribution	Mean – kb	Second Moment – kb ²
1 and 4	D	1.60	2.56
2 and 5	M	6.78	91.83
3 and 6	D_3	5.56	61.25

Size (kb)	Probability
0.32	0.5
6	0.1
12	0.4

Class	λ_i	ρ_i	λ_i	ρ_i
1	45.0	0.05	50	0.05
2	58.8	0.26	65	0.29
3	30.0	0.11	35	0.13
4	50.0	0.05	55	0.06
5	58.8	0.26	70	0.31
Total		0.73		0.83

The buffer delays are given in ms and for both cases, the approximation is very good. There are several points to be made from Figures 2 and 3. Earlier (in Table 1) we presented a result that stated the load weighted average of the class waiting times equals the waiting time of an arbitrary customer in a non-priority M/G/1 system. This result is called Kleinrock's Law of Conservation [14]. Both the simulation and approximate models validate this law.

The second result is the value of $Wq(4)$ and $Wq(5)$ at which they are equal which can be found from the Kleinrock's Law of Conservation. For the approximation, the value of w_4 at which this occurs is

$$w_4 = \frac{ES_4}{ES_4 + ES_5}. \quad (8)$$

For these two cases, the value of w_4 at which this occurs is $w_4 = 0.19 = 1.6/(1.6+6.78)$.

We saw the same crossover happen with the simulation, but not at exactly the same w_4 . The reason is because the simulation is the actual random CBWFQ rule and it only randomly selects the next CBWFQ packets to be served when there are class 4 and 5 packets waiting and no packets from the PQ traffic. Thus, the loads play a role. Our approximation randomly distributes the class 4 and 5 packets to the 4* and 5* priority class independent of the traffic loads.

For the six class examples, we again consider two loads (0.73 and 0.86). The individual loads are presented in Table 5.

Figures 4 and 5 present the comparison for the six class examples. In these examples, w_4 was held fixed at 0.2 and w_5 was varied (we note that this would imply $w_6 = 1 - w_4 - w_5 = 0.8 - w_5$ for the example).

Class	λ_i	ρ_i	λ_i	ρ_i
1	40	0.04	45	0.05
2	50	0.22	58.8	0.26
3	25	0.09	30	0.11
4	45	0.05	50	0.05
5	50	0.22	58.8	0.26
6	30	0.11	35	0.13
Total		0.73		0.86

The results for the six class example were the same as the five class example; the approximation is very good. Also the behavior with respect to Kleinrock's Law of Conservation holds.

A very interesting observation is when a class weight is zero, its buffer expected delay is given by the NPPQ model with that class being the lowest priority. For the five class example, when one of the CBWFQ weights is zero, then the NPPQ model can be used to generate buffer delays for all classes.

The weights at which all buffer expected delays are equal is

$$w_i = \frac{ES_i}{ES_4 + ES_5 + ES_6}. \quad (9)$$

Concluding remarks

We have shown that performance of the PQ traffic in the LLQ is given by Equation (2). We then approximated the CBWFQ module of the LLQ system with a variation on the standard NPPQ model and made comparisons with a simulator. We approximate this dynamic changing of the CBWFQ discipline with one that is static and CBWFQ packets are distributed to priority classes 4*, 5*, and 6* upon arrival. Then a standard NPPQ model is used and the results decoupled back to the original class via Equation (7). We feel the approximation is quite good and certainly noteworthy because of the simplicity of the analytics associated with the approximation.

The approximation is based on several assumptions (Poisson arrivals, infinite buffers, and total load less than one), and results are given only for packet buffer latencies. In order to expand the results to finite buffers, loads greater than one, other packet arrival processes, as well as generating results for packet loss and buffer delay quantiles, more analysis is required to better understand the NPPQ under more general assumptions. We have shown here and in reference [11] that the NPPQ model plays a very important role in modeling of the LLQ discipline.

Acknowledgements

This work was partially funded by the National Communications Systems under Contract Number NBCH-D-02-0039 and Noblis' Center for Network-Based Systems, which is a research collaboration of Noblis and George Mason University. ■

Notes and references

1. Semeria, C., "Supporting Differentiated Service Class: Queue Scheduling Disciplines," *Juniper White Paper*, 2001; http://www.juniper.net/solutions/literature/white_papers/200020.pdf.
2. Cisco Systems, *IOS Quality of Service Solutions Configuration Guide*, Release 12.2 (Congestion Management Overview chapter); http://www.cisco.com/en/US/products/sw/iosswrel/ps1835/products_configuration_guide_chapter09186a00800b75a9.html.
3. Davie, B. "Is QoS Necessary? Quality of Service Mechanisms vs. Bandwidth Provisioning," Fall 2002 Seminar/Public Lecture Series, Stanford University School of Engineering, U.S. Asia Technology Management Center, 2002.
4. Fischer, M. J. and D. M. B. Masi, "A Quantitative Analysis of the Voice and Data Quality of Service Problem," *The Telecommunications Review*, vol. 18, Noblis, Falls Church, VA, 2007.
5. Fischer, M. J., D. M. B. Masi, and P. V. McGregor, "Making an Efficient Integrated Services Enterprise Network," *ITPro*, September/October 2007.
6. Masi, D. M. B. and M. J. Fischer, "Voice over Internet Protocol (VoIP) Performance Models—A Comprehensive Approach," International Conference on Telecommunication Systems—Modeling and Analysis (ICTSMA), Dallas, TX, November 17–20, 2005.

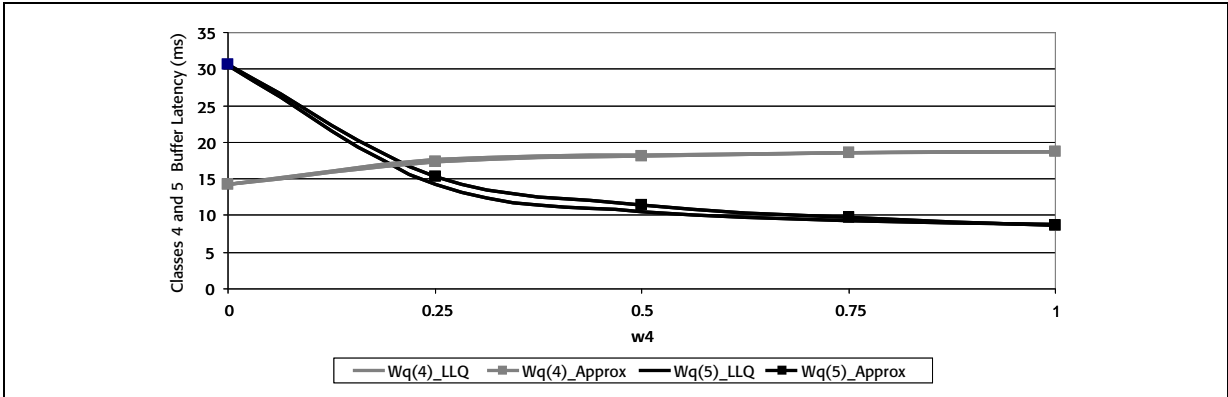


Figure 2. The five class comparison of approximation and simulation for $\rho = 0.73$

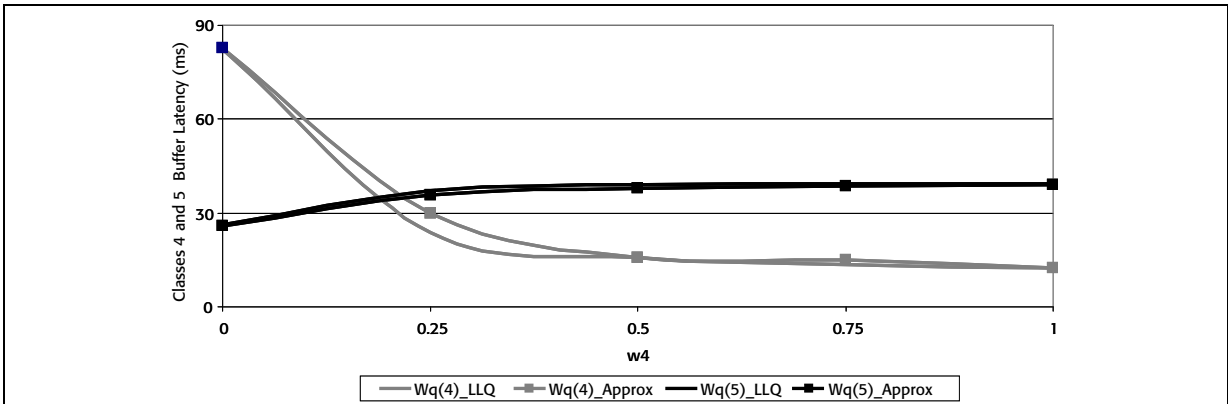


Figure 3. The five class comparison of approximation and simulation for $\rho = 0.83$

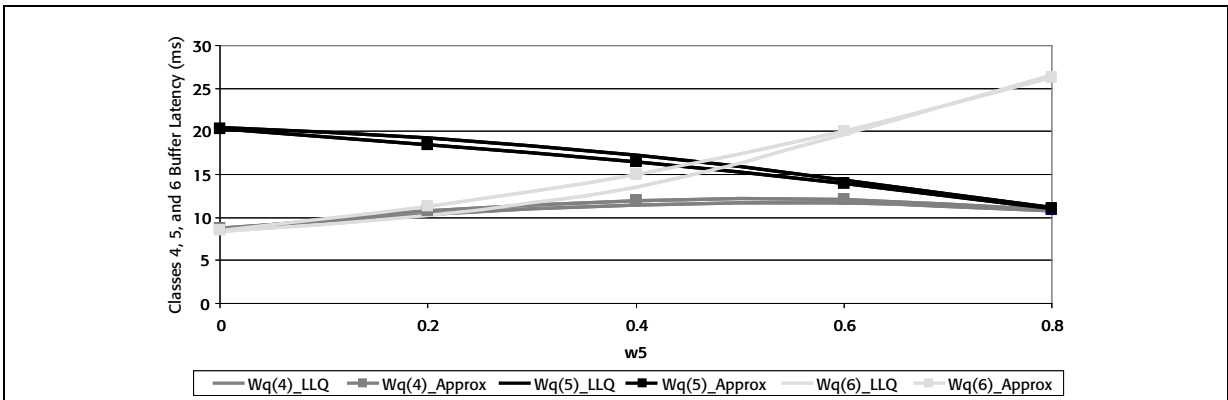


Figure 4. The six class comparison of approximation and simulation for $\rho = 0.73$

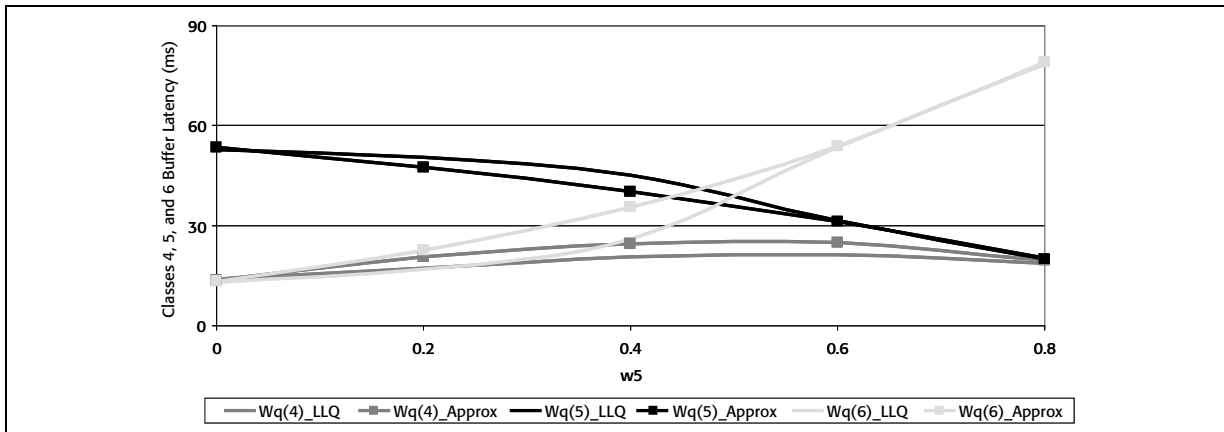


Figure 5. The six class comparison of approximation and simulation for $\rho = 0.85$

7. Baker, F., J. Polk, and M. Dolly, "An EF DSCP for Capacity-Admitted Traffic," draft-ietf-tsvwg-admitted-realtime-dscp-00, December 2006; <http://tools.ietf.org/id/draft-ietf-tsvwg-admitted-realtime-dscp-00.txt>.
8. Baker, F., J. Polk, and M. Dolly, "DSCPs for Capacity-Admitted Traffic," draft-ietf-tsvwg-admitted-realtime-dscp-01, March 2007; <http://tools.ietf.org/id/draft-ietf-tsvwg-admitted-realtime-dscp-01.txt>.
9. Masi, D. M. B., M. J. Fischer, and D. A. Garbin, "Modeling the Performance of Low Latency Queueing for Emergency Telecommunications," Winter Simulation Conference, Washington, DC, December 2007.
10. Golestani, S. J., "A Self-Clocked Fair Queueing Scheme for Broadband Applications," *Proceedings of the IEEE INFOCOM*, 1994.
11. Fischer, M. J., D. A. Garbin, and D. M. B. Masi, "A Discussion of Low Latency Queueing Performance Modeling," *Networking and Electronic Commerce Research Conference*, Garda, Italy, October 2007.
12. Cohen, J. W., "The Single Server Queue," North-Holland Publishing Company, New York, 1969.
13. Gross, D. and C. M. Harris, "Fundamentals of Queueing Theory," Third Edition, John Wiley, New York, NY, 1998.
14. Kleinrock, L., "Queueing Systems: Volume 2," Wiley, 1976.
15. Masi, D. M. B., M. J. Fischer, and D. A. Garbin, "Modeling the Performance of Class-Based Weighted Fair Queueing with OPNET and Custom Simulators," OPNETWORK Conference, Washington, DC, August 2007.
16. Thompson, K., G. J. Miller, and R. Wilder, "Wide Area Internet Traffic Patterns and Characteristics," *IEEE Network*, November/December 1997.

About the authors



Martin J. Fischer is a senior fellow at Noblis where his experience includes network design and performance analysis in telecommunications. He has published more than 50 articles in refereed journals. He received his doctorate degree in operations research from Southern Methodist University. Contact him at mfischer@noblis.org.



Denise M. Bevilacqua Masi is a senior principal engineer at Noblis where her experience and research interests include queueing theory and simulation applied to telecommunications networks. She received her doctorate degree in information technology and engineering at George Mason University. Contact her at dmasi@noblis.org.



John F. Shortle is an associate professor of systems engineering and operations research at George Mason University (GMU). He is a member of the Center for Air Transportation Systems Research at GMU and a member of the Center for Network-Based Systems, a collaborative initiative between Noblis and GMU. His experience includes developing stochastic, queueing, and simulation models to optimize networks and operations. His research interests include simulation and queueing applications in telecommunications and air transportation. He received his doctorate degree in operations research from UC Berkeley. Contact him at jshortle@gmu.edu.