Manuscript submitted to
AIMS' Journals
Volume **X**, Number **0X**, XX **200X**

Website: http://AIMsciences.org

pp. **X–XX**

INFLUENCE OF REAL-TIME QUEUE CAPACITY ON SYSTEM CONTENTS IN DIFFSERV'S EXPEDITED FORWARDING PER-HOP-BEHAVIOR

Thomas Demoor

Department of Telecommunications and Information Processing, Ghent University
St-Pietersnieuwstraat 41, 9000 Gent, BELGIUM

Joris Walraevens, Dieter Fiems, Stijn De Vuyst and Herwig Bruneel

Department of Telecommunications and Information Processing, Ghent University
St-Pietersnieuwstraat 41, 9000 Gent, BELGIUM

Abstract. This paper studies a single-server non-preemptive priority queue with two traffic classes in order to model Expedited Forwarding Per-Hop Behavior in the Differentiated Services (DiffServ) architecture. Generally, queueing models assume infinite queue capacity but in a DiffServ router the capacity for high priority traffic is typically small to prevent this traffic from monopolizing the output link and hence causing starvation of low-priority traffic. The presented model takes the exact (finite) high-priority queue capacity into account. Analytical formulas for the system content of each class are determined as well as the high-priority packet loss ratio. For each class, service of a packet takes a (different) general independent distribution. The issues this causes are resolved by using spectral decomposition. Numerical examples indicate the considerable impact of the finite capacity on system performance.

1. Introduction.

The rapid development of modern telecommunication networks has resulted in a wide variety of performance demands for various types of traffic. Evidently, allowing all traffic to meet their Quality of Service (QoS) requirements is of paramount importance. One of the more popular attempts to supply improved QoS is Differentiated Services (DiffServ) [1],[2], a computer networking architecture in Internet Protocol (IP) networks that distributes packets in various traffic classes. It provides QoS differentiation by basing the order in which packets are transmitted on class-dependent priority rules. In DiffServ each packet is forwarded according to its Per-Hop Behavior (PHB). Obviously, implementation of DiffServ is particularly interesting in networks that struggle to provide acceptable QoS because bandwidth is limited and/or variable.

2000 *Mathematics Subject Classification.* Primary: 60J10, 60K25; Secondary: 68M12.

Key words and phrases. Discrete-time priority queue, performance analysis, DiffServ, spectral decomposition.

This paper was presented at the QTNA2009 conference which was held in Singapore during July 29-August 1, 2009. An earlier and brief version of this paper was published in the Conference Proceedings. The reviewing process of the paper was handled by Wuyi Yue, Yutaka Takahashi and Hideaki Takagi as Guest Editors.

The second and third authors are Postdoctoral Fellows with the Research Foundation Flanders (FWO-Vlaanderen), Belgium.

A rather coarse, but very practical, classification distributes packets in two traffic classes. Real-time traffic, such as streaming video, requires low delays but can endure a small amount of packet loss. On the other hand, data traffic, such as file transfer, benefits from low packet loss but has less stringent delay characteristics.

This paper considers a two-class priority queueing system representing a DiffServ implementation where real-time traffic (Expedited Forwarding PHB) has strict priority scheduling over data traffic (Default PHB). Although this scheduling algorithm is quite drastic, as data packets are only served if the system is void of real-time packets, it minimizes the delay of the real-time packets. Furthermore, upon arrival of a real-time packet, its delay is known, in contrast to queues with more intricate scheduling algorithms where, f.i. the delay could be influenced by future arrivals. This provides an upper bound for the number of real-time packets the system should be able to contain, given the allowed maximum delay for this type of packets.

As real-time packets receive absolute priority, they can occupy the server (almost) permanently, denying data traffic of any service, if no admission control is performed. Therefore, the amount of real-time traffic allowed into the system should be regulated. Moreover, queueing a very large amount of real-time packets is useless anyway as they require small delays. These two observations emphasize the importance of limiting the capacity for real-time packets, evidently, without neglecting packet loss constraints. On the other hand, data packets require a very low amount of loss to achieve their QoS requirements. Therefore, the system capacity for data packets should be as large as practically feasible. Hence, we can assume that the capacity for data packets is sufficiently large to be approximated by infinity but that the capacity for real-time packets should be modelled exactly.

In the literature, various priority queueing systems (a.o. [8],[5]) have been discussed with infinite queue capacity, as this facilitates mathematical analysis of the system. In contrast, the queueing model studied in this paper considers finite capacity for real-time packets and infinite capacity for data packets, as explained in the former paragraph. This paper is an expansion of [3] where service of a packet was assumed to be deterministically equal to a single slot for data traffic and it is an extended version of [4]. As packet sizes of real-time and data packets are typically dissimilar, the current contribution expands the differentiation amongst packet sizes of both classes as we allow service times to take a (different) general distribution for each class. When a real-time packet arrives during service of a data packet (thus in a system void of real-time packets), it does not interrupt the service of the data packet. This is called non-preemptive priority and it causes the performance of real-time traffic to be dependent on data traffic, which was not the case in our previous contributions. The complexity introduced by the use of general service times is resolved by applying the spectral decomposition theorem (e.g. [6]).

The presented model is related to [10] where both queues are presumed to have infinite capacity. Finite queue capacity is considered in [9] as well, albeit by a different methodology, but only packet loss is investigated under the restriction of single-slot service times. Assessing the impact of the finite real-time queue capacity on the system contents is the main purpose of the current contribution.

The remainder of this paper is organized as follows: first the model under consideration will be thoroughly described. In section 3, several performance measures for our system are determined analytically. Next, the unfinished work is obtained. Afterwards, the results are investigated in some (numerical) examples. The paper is concluded in section 6.

2. **Model.** This paper studies a discrete-time single-server two-class non-preemptive priority queueing system where class-1 (real-time) packets receive strict priority over class-2 (data) packets. Packets are handled in a First-In-First-Out (FIFO) manner within a class. We limit the capacity of the class-1 queue to N packets such that real-time packets that arrive at a full queue are dropped by the system. The system can hence contain up to $N+1$ class-1 packets simultaneously, N in the queue and 1 in the server. In contrast, the class-2 queue has infinite capacity. Time is divided into fixed-length slots and a packet can only enter the server at slot boundaries, even if arriving in an empty system.

Let s_i $(i = 1, 2)$ denote a generic random service time of a class-i packet . These independent variables have corresponding pgfs $S_i(z)$ and mean values μ_i $(i = 1, 2)$. Observing the system at the beginning of a slot happens after the departure (if any) at the slot boundary but before arrivals in that slot.

We assume that, for both classes, the numbers of arrivals in consecutive slots form a sequence of independent and identically distributed (i.i.d.) random variables. We define $a_{i,k}$ as 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 (pmf)

$$a(m, n) = \Pr[a_{1,k} = m, a_{2,k} = n] \tag{1}$$

which allows us to take into account dependence between both classes. The corresponding probability generating function (pgf) is denoted by

$$A(z_1, z_2) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} a(i, j) z_1^i z_2^j \ . \tag{2}$$

The partial pgf of the number of class-2 arrivals in a slot with i $(0 \le i \le N)$ and i or more class-1 arrivals are respectively denoted by $A_i(z)$ and $A_i^*(z)$. We establish

$$A_i(z) = \mathrm{E}[z^{a_{2,k}}\ 1\{a_{1,k} = i\}] = \sum_{j=0}^{\infty} a(i, j) z^j \ , \ A_i^*(z) = \sum_{j=i}^{\infty} A_j(z) \ . \tag{3}$$

The indicator function $1\{.\}$ evaluates to 1 if its argument is true and to 0 if it is false. The mean number of class-1 and class-2 arrivals per slot are respectively expressed as

$$\bar{a}_1 = \sum_{i=1}^{\infty} i A_i(1) \ , \ \bar{a}_2 = \frac{d}{dz} A_0^*(z)\bigg|_{z=1} = A_0^{*\prime}(1) \ . \tag{4}$$

Hence, the arrival loads per class are $\rho_1 = \bar{a}_1\mu_1$ and $\rho_2 = \bar{a}_2\mu_2$. The mean number of total arrivals and total arrival load are respectively denoted by $\bar{a}_T = \bar{a}_1 + \bar{a}_2$ and $\rho_T = \rho_1 + \rho_2$.

3. **Analysis.** First, we review the spectral decomposition theorem for non-diagonalisable matrices as it will be used frequently in the remainder of this paper. In section 3.2, the equations for the system contents of both classes at the beginning of so-called start-slots are established. The next subsection addresses the characterization of arrivals during a class-i service. This enables determination of the system contents at start-slots in subsection 3.4. Finally, the system contents at the beginning of random slots are derived from those at start-slots.

3.1. **Spectral decomposition of non-diagonalisable matrices.** Let us consider a square $m \times m$ matrix $\mathbf{A}$ and a scalar function f. The spectral decomposition theorem allows us to express the image of $\mathbf{A}$ under f by evaluating f (and its derivatives) in the eigenvalues of $\mathbf{A}$, see e.g. [6].

In this paper, the function f is typically a power series $f(z) = \sum_{n=0}^{\infty} f_n z^n$ and the matrix $\mathbf{A}$ is non-diagonalisable. Such a matrix $\mathbf{A}$ cannot be reduced to a completely diagonal form by a similarity transform. However, any square matrix can be reduced to a form that is almost diagonal, called the Jordan normal form $\mathbf{J}$. Based on this reduction, it is possible to prove that the matrix $f(\mathbf{A})$ can be uniquely defined as

$$f(\mathbf{A}) = \sum_{j=1}^{s} \sum_{i=0}^{k_j-1} \frac{1}{i!} f^{(i)}(\lambda_j) \, (\mathbf{A} - \lambda_j \mathbf{I})^i \, \mathbf{G}_j \ , \tag{5}$$

see formula (7.9.9) in [6]. In this expression, $\{\lambda_1, \ldots, \lambda_s\}$ $(s \leq m)$ are the s distinct eigenvalues of $\mathbf{A}$, k_j denotes the index of eigenvalue λ_j and $f^{(i)}$ is the ith derivative of f. Obviously, it is required that the function f and its derivatives exist in the eigenvalues, i.e.

$$\lambda_j \in \mathrm{dom}\ f^{(i)} \,, \qquad j = 1, \ldots, s \,, i = 0, \ldots, k_j - 1 \,. \tag{6}$$

The matrices $\mathbf{G}_j$ are called the constituents or spectral projectors of $\mathbf{A}$ belonging to the eigenvalue λ_j and have the following properties:

- $\mathbf{G}_j$ is idempotent, i.e. $\mathbf{G}_j^2 = \mathbf{G}_j$.
- $\mathbf{G}_1 + \mathbf{G}_2 + \ldots + \mathbf{G}_s = \mathbf{I}$, with $\mathbf{I}$ the $m \times m$ identity matrix.
- $\mathbf{G}_j \mathbf{G}_{j'} = \mathbf{0}$ whenever $j \neq j'$ $(1 \leq j, j' \leq s)$.

In general, the matrices $\mathbf{G}_j$ need to be calculated from the transformation matrix $\mathbf{P}$, for which $\mathbf{J} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$. Specifically, if $\mathbf{P}$ is partitioned conformably as

$$\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1} = \begin{bmatrix} \mathbf{P}_1 & \mathbf{P}_2 & \cdots & \mathbf{P}_s \end{bmatrix} \begin{bmatrix} \mathbf{J}_1 & & & \\ & \mathbf{J}_2 & & \\ & & \ddots & \\ & & & \mathbf{J}_s \end{bmatrix} \begin{bmatrix} \mathbf{Q}_1 \\ \mathbf{Q}_2 \\ \vdots \\ \mathbf{Q}_s \end{bmatrix} , \tag{7}$$

with $\mathbf{J}_j$ the Jordan segment corresponding with eigenvalue λ_j, then the projectors $\mathbf{G}_j$ are

$$\mathbf{G}_j = \mathbf{P}_j \mathbf{Q}_j \qquad (j = 1, \ldots, s) \ . \tag{8}$$

We also note that the columns of $\mathbf{P}_j$ span the space of the right eigenvectors of $\mathbf{A}$ corresponding to λ_j while the rows of $\mathbf{Q}_j$ span the space of its left eigenvectors.

This spectral decomposition theorem provides us with a very powerful tool from a computational point of view. Instead of having to evaluate the matrix power series $\sum_{n=0}^{\infty} f_n \mathbf{A}^n$, we only need to evaluate the function f and its derivatives for scalar arguments and compute a finite number of matrix multiplications. The downside is that the eigenvalues of $\mathbf{A}$ have to be calculated, as well as the matrices $\mathbf{G}_j$. But once this is done, $f(\mathbf{A})$ can easily be calculated for any function f satisfying (6). In subsection 3.3, it will become clear that in our case the downsides are virtually non-existent as the eigenvalues and spectral projectors are surprisingly easy to obtain.

3.2. **Relating consecutive start-slots.** A start-slot is a slot where service of a packet can start. In Figure 1, the evolution of the system is exemplified for an ad-hoc case. Hereby, we hope to clarify the concept of start-slots and give some insights into the system studied in this paper. On the left, the queueing system is depicted

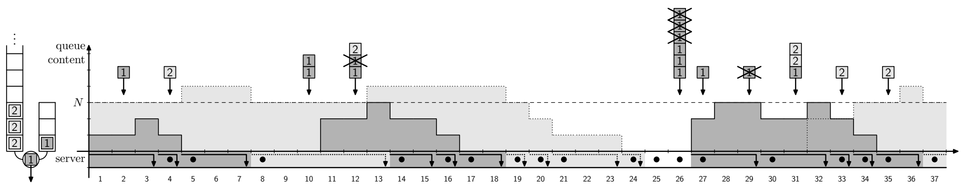

FIGURE 1. Evolution of the finite/infinite non-preemptive queueing system with $N = 3$ over 37 slots. The contents of both the high-priority queue (1-packets: dark grey) and low-priority queue (2-packets: light grey) are shown, as well as the content of the server. An arriving high-priority packet that can not be accommodated is lost. Start-slots are indicated with a '$\bullet$'.

for $N = 3$. Class-1/class-2 packets are dark-/lightgrey. Each class has a dedicated queue, both queues are served by the same server. The evolution of the system contents, influenced by arrivals and completed services, is depicted during 37 slots. The following events are particularly interesting. In slot 10, class-1 packets arrive in a system void of class-1 packets. Although class-1 packets have priority, class-1 service can only start after the class-2 service in progress is completed. Therefore, class-1 performance is dependent on class-2 traffic. Slot 12 exemplifies that the class-1 queue can only hold N packets, while slot 26 demonstrates that packets can not enter the server upon arrival, but only at slot boundaries.

Note that a slot where the system is empty at the beginning of the slot is a start-slot as well (slots 25 and 26 in the example). Evidently, the time epoch between the beginning of two consecutive start-slots consists of s_1, s_2 or a single slot(s), depending on the type of packet in the server (if any). Therefore, study of the evolution of the system during a service time is of paramount importance to the analysis. Let $e_{i,k}^j$ represent the number of class-i arrivals during a class-j service that starts in slot k. We have

$$e_{i,k}^j = \sum_{m=0}^{s_j - 1} a_{i,k+m} \ . \tag{9}$$

Notice that the $e_{i,k}^j$ are i.i.d. (for different k) as the $a_{i,k}$ are i.i.d. and independent of s_j.

Let $n_{i,l}$ denote the class-i system contents at the beginning of start-slot l, constituted by the class-i queue contents (thus excluding the server) and by the packet in the server if this packet is of class i. The set $\{(n_{1,l}, n_{2,l}), l \geq 1\}$ forms a Markov chain. Assume that start-slot l corresponds with slot k. Relating start-slots l and $l + 1$ establishes the following set of system equations:

$$\begin{aligned} n_{1,l+1} &= \begin{cases} \min(N, a_{1,k}), & \text{if } n_{1,l} = 0, n_{2,l} = 0 \\ \min(N, e_{1,k}^2), & \text{if } n_{1,l} = 0, n_{2,l} > 0 \\ \min(N, n_{1,l} - 1 + e_{1,k}^1), & \text{if } n_{1,l} > 0 \end{cases} , \\ n_{2,l+1} &= \begin{cases} a_{2,k}, & \text{if } n_{1,l} = 0, n_{2,l} = 0 \\ n_{2,l} - 1 + e_{2,k}^2, & \text{if } n_{1,l} = 0, n_{2,l} > 0 \\ n_{2,l} + e_{2,k}^1, & \text{if } n_{1,l} > 0 \end{cases} . \end{aligned} \tag{10}$$

The system equations can be explained as follows: if the system is empty, start-slot $l+1$ is the next slot $(k+1)$ and thus only the arrivals during slot k contribute to the system contents. If $n_{1,l}=0, n_{2,l}>0$, a class-2 packet starts service at the beginning of start-slot l and it leaves the system immediately before start-slot $l+1$. For each class, admitted arrivals during this class-2 service contribute to the system contents at the beginning of start-slot $l+1$. On the other hand, if $n_{1,l}>0$, a class-1 packet starts service at the beginning of start-slot l and it leaves the system immediately before start-slot $l+1$. For each class, admitted arrivals during this class-1 service contribute to the system contents at the beginning of start-slot $l+1$. Note that the class-1 system contents at the beginning of start-slots cannot exceed N, the class-1 queue capacity.

3.3. **Arrivals during a service.** In this subsection, the number of arrivals during a class-j $(j=1,2)$ service is characterized. The partial pgfs of the number of class-2 arrivals during a class-j service, during which i $(0\leq i\leq N)$ and i or more class-1 packets arrive are respectively denoted by $E_i^j(z)$ and $E_i^{j*}(z)$. We have

$$E_i^j(z)=\mathrm{E}[z^{e_{2,k}^j}\ 1\{e_{1,k}^j=i\}]\ ,\ E_i^{j*}(z)=\sum_{m=i}^{\infty}E_m^j(z)\ . \tag{11}$$

Obtaining these partial pgfs is, in general, a tedious task. We have

$$E_i^j(z)=\frac{1}{i!}\frac{d^i}{dx^i}S_j\big(A(x,z)\big)\Big|_{x=0}\ ,\ E_i^{j*}(z)=S_j\big(A(1,z)\big)-\sum_{k=0}^{i-1}E_k^j(z)\ . \tag{12}$$

However, from a computational point of view, this is infeasible for general $S_j(z)$ and $A(z_1,z_2)$. The most common approach would be to invert this two-dimensional transform using the (fast) Fourier-series method. When solving the system under consideration however, following alternative method is interesting, especially from a numerical point of view as a lot of the computational effort is reused in the further analysis of the system.

In this subsection, these pgfs are obtained using the spectral decomposition theorem. Packets only leave the system at the end of a service. Therefore, in each slot during a service, the queue contents evolve according to the $(N+1)\times(N+1)$ matrix

$$\mathbf{Y}(z)=\begin{bmatrix} A_0(z) & A_1(z) & \cdots & A_{N-1}(z) & A_N^*(z)\\ 0 & A_0(z) & \cdots & A_{N-2}(z) & A_{N-1}^*(z)\\ \vdots & \ddots & \ddots & \vdots & \vdots\\ \vdots & & \ddots & A_0(z) & \vdots\\ 0 & \cdots & \cdots & 0 & A_0^*(z)\end{bmatrix}\ . \tag{13}$$

More precisely, for $1\leq i,j\leq N+1$, given that the class-1 queue content is $i-1$ during the previous slot, $\mathbf{Y}(1)_{ij}$ is the probability that it is $j-1$ in the current slot (this is the probability that $j-i$ class-1 packets are effectively allowed into the system), while $\mathbf{Y}(z)_{ij}$ is the partial pgf of the packets added to the class-2 queue.

The partial pgfs $E_i^j(z)$ and $E_i^{j*}(z)$ are found as elements of the matrices $\mathbf{E}_j(z)=S_j\big(\mathbf{Y}(z)\big),\ j=1,2$. Using spectral decomposition, these matrices are readily obtained because of the special eigenstructure of $\mathbf{Y}(z)$. As this matrix has a triangular form, the eigenvalues simply are its diagonal elements. There are two distinct eigenvalues: $\lambda_1=A_0^*(z)$, with index 1, and $\lambda_2=A_0(z)$, with index N. The corresponding spectral projectors are easily shown to be independent of z and are given by

$$\mathbf{G}_1 = \begin{bmatrix}\mathbf{0} & \cdots & \mathbf{0} & \mathbf{e}\end{bmatrix} \ , \ \mathbf{G}_2 = \begin{bmatrix}\mathbf{I} & -\mathbf{e} \\ \mathbf{0}^T & 0\end{bmatrix} \ . \tag{14}$$

Here $\mathbf{I}$ denotes the identity matrix of appropriate size, $\mathbf{x}^T$ is the transpose of vector $\mathbf{x}$ and $\mathbf{e}$ and $\mathbf{0}$ indicate the column vectors of appropriate size with all elements equal to 1 and 0 respectively.

Hence, spectral decomposition (5) yields

$$\begin{aligned}\mathbf{E}_j(z) = S_j\big(\mathbf{Y}(z)\big) &= \begin{bmatrix} E_0^j(z) & E_1^j(z) & \cdots & E_{N-1}^j(z) & E_N^{j*}(z) \\ 0 & E_0^j(z) & \cdots & E_{N-2}^j(z) & E_{N-1}^{j*}(z) \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & & \ddots & E_0^j(z) & \vdots \\ 0 & \cdots & \cdots & 0 & E_0^{j*}(z)\end{bmatrix} \\ &= S_j\big(A_0^*(z)\big)\mathbf{G}_1 + \sum_{k=0}^{N-1} \frac{S_j^{(k)}\big(A_0(z)\big)}{k!}\big(\mathbf{Y}(z) - A_0(z)\mathbf{I}\big)^k \mathbf{G}_2 \ .\end{aligned} \tag{15}$$

Note that $\mathbf{E}_1(z)$ and $\mathbf{E}_2(z)$ share all factors except (the derivatives of) the functions $S_1(z)$ and $S_2(z)$. Especially note that the (computationally expensive) powers of $\big(\mathbf{Y}(z) - A_0(z)\mathbf{I}\big)$ are shared.

3.4. **System contents at the beginning of start-slots in steady state.** The partial pgf of the class-2 system contents at the beginning of start-slot l that has class-1 system contents equal to i is denoted by

$$N_{i,l}(z) = \mathrm{E}[z^{n_{2,l}} \ 1\{n_{1,l} = i\}] \ . \tag{16}$$

Let us organize these pgfs into a row vector of $N+1$ elements

$$\mathbf{n}_l(z) = \big[N_{i,l}(z)\big]_{i=0..N} \ , \tag{17}$$

which corresponds with the system contents at the lth start-slot. Furthermore, define the $(N+1)\times(N+1)$ matrices

$$\mathbf{H}_1 = \begin{bmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0\end{bmatrix} , \ \mathbf{H}_2 = \mathbf{I} - \mathbf{H}_1, \ \mathbf{D} = \begin{bmatrix}\mathbf{0}^T & 0 \\ \mathbf{I} & \mathbf{0}\end{bmatrix} \ , \tag{18}$$

and the row vector of $N+1$ elements

$$\mathbf{h}_1 = \begin{bmatrix}1 & 0 & \cdots & 0\end{bmatrix} \ . \tag{19}$$

By conditioning on the state of the server at start-slot l, a relation between $\mathbf{n}_l(z)$ and $\mathbf{n}_{l+1}(z)$ is derived from the system equations (10). We have

$$\mathbf{n}_{l+1}(z) = \mathbf{n}_l(0)\mathbf{H}_1\mathbf{Y}(z) + \big(\mathbf{n}_l(z) - \mathbf{n}_l(0)\big)\mathbf{H}_1\frac{1}{z}\mathbf{E}_2(z) + \mathbf{n}_l(z)\mathbf{H}_2\mathbf{D}\mathbf{E}_1(z) \ . \tag{20}$$

This can be explained as follows. The first term corresponds with an empty server. Therefore, $n_{2,l} = 0, n_{1,l} = 0$ and start slot $l+1$ is the next slot thus we take into account the arrivals in a single slot (start-slot l). The second term represents the evolution of the system when a class-2 service starts at start-slot l. This yields that $n_{2,l} > 0, n_{1,l} = 0$, that by start-slot $l+1$ the class-2 packet in service will have left the system and that we need to consider arrivals during a class-2 service. The final term corresponds with a class-1 packet starting service at start-slot l. Then $n_{1,l} > 0$ the class-1 packet in service will have left the system by start-slot $l+1$ and packets arriving during this class-1 service need to be accounted for.

Assume that the system has reached steady state and define following steady-state values

$$\mathbf{n}(z) = \lim_{l\to\infty} \mathbf{n}_l(z) = \lim_{l\to\infty} \mathbf{n}_{l+1}(z) = \big[N_i(z)\big]_{i=0..N} \ . \tag{21}$$

Taking the limit of (20) for $l \to \infty$ induces

$$\mathbf{n}(z)\big(z\mathbf{I} - \mathbf{H}_1\mathbf{E}_2(z) - z\mathbf{D}\mathbf{E}_1(z)\big) = \mathbf{n}(0)\mathbf{H}_1\big(z\mathbf{Y}(z) - \mathbf{E}_2(z)\big) \ . \tag{22}$$

Furthermore, note that

$$\mathbf{n}(0)\mathbf{H}_1 = N_0(0)\mathbf{h}_1 \ . \tag{23}$$

Hence, (22) becomes

$$\mathbf{n}(z)\big(z\mathbf{I} - \mathbf{H}_1\mathbf{E}_2(z) - z\mathbf{D}\mathbf{E}_1(z)\big) = N_0(0)\mathbf{h}_1\big(z\mathbf{Y}(z) - \mathbf{E}_2(z)\big) \ . \tag{24}$$

The constant $N_0(0)$ is the only unknown. It is found in two steps. First, evaluation of (24) in $z = 1$ produces

$$\mathbf{n}(1)\big(\mathbf{I} - \mathbf{H}_1\mathbf{E}_2(1) - \mathbf{D}\mathbf{E}_1(1)\big) = N_0(0)\mathbf{h}_1\big(\mathbf{Y}(1) - \mathbf{E}_2(1)\big) \ . \tag{25}$$

As the matrices $\mathbf{E}_j(1)$, $j = 1, 2$ are right-stochastic by construction, each row of matrix $[\mathbf{I} - \mathbf{H}_1\mathbf{E}_2(1) - \mathbf{D}\mathbf{E}_1(1)]$ sums to 0 and it hence has rank N and is not invertible. We thus require an additional relation in order to obtain the vector $\mathbf{n}(1)$. The normalization condition provides $\mathbf{n}(1)\mathbf{e} = 1$. Combining this with (25) yields

$$\mathbf{n}(1) = N_0(0)\left[\mathbf{h}_1\big(\mathbf{Y}(1) - \mathbf{E}_2(1)\big)\middle\| \frac{1}{N_0(0)}\right]\left[\mathbf{I} - \mathbf{H}_1\mathbf{E}_2(1) - \mathbf{D}\mathbf{E}_1(1)\middle\|\mathbf{e}\right]^{-1} \ . \tag{26}$$

By $[\mathbf{A}\|\mathbf{b}]$ we denote the matrix $\mathbf{A}$ with the last column replaced by the column vector $\mathbf{b}$ and by $[\mathbf{a}\|b]$ the vector $\mathbf{a}$ with the last element replaced by b. Second, derivation of (24) with respect to z yields

$$\begin{aligned}&\mathbf{n}(z)\big(\mathbf{I} - \mathbf{H}_1\mathbf{E}_2'(z) - \mathbf{D}\mathbf{E}_1(z) - z\mathbf{D}\mathbf{E}_1'(z)\big)\\ &+ \mathbf{n}'(z)\big(z\mathbf{I} - \mathbf{H}_1\mathbf{E}_2(z) - z\mathbf{D}\mathbf{E}_1(z)\big) = N_0(0)\mathbf{h}_1\big(\mathbf{Y}(z) + z\mathbf{Y}'(z) - \mathbf{E}_2'(z)\big) \ .\end{aligned} \tag{27}$$

Observe that $\mathbf{E}_j(1)$ $(j = 1, 2)$ and $\mathbf{Y}(1)$ are right-stochastic matrices by construction. Therefore,

$$\big(\mathbf{I} - \mathbf{H}_1\mathbf{E}_2(1) - \mathbf{D}\mathbf{E}_1(1)\big)\mathbf{e} = \mathbf{0} \ , \mathbf{h}_1\mathbf{Y}(1)\mathbf{e} = 1 \ . \tag{28}$$

Keeping these identities in mind, evaluation of (27) in $z = 1$ and multiplication of both sides of the resulting equation by $\mathbf{e}$ yields

$$N_0(0) = \frac{\mathbf{n}(1)\big(\mathbf{I} - \mathbf{H}_1\mathbf{E}_2'(1) - \mathbf{D}\mathbf{E}_1(1) - \mathbf{D}\mathbf{E}_1'(1)\big)\mathbf{e}}{1 + \mathbf{h}_1\big(\mathbf{Y}'(1) - \mathbf{E}_2'(1)\big)\mathbf{e}} \ . \tag{29}$$

Substitution of (26) for $\mathbf{n}(1)$ provides $N_0(0)$.

Now that $N_0(0)$ has been obtained, (26) provides $\mathbf{n}(1)$, the probability mass function (pmf) of the class-1 system contents at the beginning of a start-slot in steady state and from (24), the pgf of the class-2 system contents at the beginning of a start-slot in steady state are found as $\mathbf{n}(z)\mathbf{e}$. From this pgf, all moments can be determined by multiple differentiation of (24) with respect to z and evaluation in $z = 1$.

3.5. **System contents at the beginning of random slots in steady state.** The class-i system contents at the beginning of slot k are denoted by $u_{i,k}$. Note that $0 \leq u_{1,k} \leq N+1$ as the class-1 queue can hold up to N packets and the server can hold a single packet. Define the vector $\mathbf{u}(z)$, of $N+2$ elements, containing the partial pgfs of the system contents (of both classes) at the beginning of a random slot in steady state. We have

$$\mathbf{u}(z) = \big[U_i(z)\big]_{i=0..N+1} = \lim_{k\to\infty} \big[\mathrm{E}[z^{u_{2,k}}\ 1\{u_{1,k}=i\}]\big]_{i=0..N+1} \ . \tag{30}$$

The vector $\mathbf{u}(z)$ is obtained by conditioning on the state of the server. We have

$$\begin{aligned}\mathbf{u}(z) = &\lim_{k\to\infty} \big[\mathrm{E}[z^{u_{2,k}}\ 1\{u_{1,k}=i,\ \text{no service}\}]\big]_{i=0..N+1} \\ &+ \lim_{k\to\infty} \big[\mathrm{E}[z^{u_{2,k}}\ 1\{u_{1,k}=i,\ \text{class-2 service}\}]\big]_{i=0..N+1} \\ &+ \lim_{k\to\infty} \big[\mathrm{E}[z^{u_{2,k}}\ 1\{u_{1,k}=i,\ \text{class-1 service}\}]\big]_{i=0..N+1} \ .\end{aligned} \tag{31}$$

If a class-j ($j=1,2$) packet is in service during a random slot, the time epoch between the beginning of that slot and the beginning of the preceding start-slot is called the elapsed service time s_j^- of that packet. Keeping in mind that a random service slot is part of a larger service time with higher probability, the pgf of the elapsed service time of the class-j packet in service in a random slot has been determined in [7]. We have

$$S_j^-(z) = \frac{S_j(z)-1}{\mu_j(z-1)} \ . \tag{32}$$

Let $f_{i,k}^j$ represent the number of class-i arrivals during the elapsed class-j service time up to slot k. We have

$$f_{i,k}^j = \sum_{m=1}^{s_j^-} a_{i,k-m} \ . \tag{33}$$

The matrix containing the corresponding partial pgfs is again found using the spectral decomposition theorem (5). Recall that, in each slot during a service, the queue contents evolve according to the matrix $\mathbf{Y}(z)$, defined in (13). Therefore, the number of arrivals during an elapsed class-j service time are given by

$$\begin{aligned}\mathbf{F}_j(z) &= S_j^-\big(\mathbf{Y}(z)\big) \\ &= S_j^-\big(A_0^*(z)\big)\mathbf{G}_1 + \sum_{k=0}^{N-1} \frac{S_j^{-(k)}\big(A_0(z)\big)}{k!}\big(\mathbf{Y}(z)-A_0(z)\mathbf{I}\big)^k\mathbf{G}_2 \ .\end{aligned} \tag{34}$$

Note that almost all factors were previously obtained in (15), which is very interesting from a computational point of view.

The state of the server in a random slot k equals the state of the server in the preceding start-slot l, as in an empty system these slots coincide and in a non-empty system a class-j ($j=1,2$) packet enters the server in start-slot l and remains there until the following start-slot. Therefore, we can express the system contents at a random slot as the system contents at the preceding start-slot augmented with the arrivals during the elapsed service time. This observation enables following calculations. First,

$$\begin{aligned}&\mathrm{E}[z^{u_{2,k}}\ 1\{u_{1,k}=i\} \mid \text{no service}] \\ &= \mathrm{E}[z^0\ 1\{u_{1,k}=i\} \mid u_{1,k}=u_{2,k}=0] = 1\{i=0\} \ .\end{aligned} \tag{35}$$

Second,

$$\begin{aligned}&\mathrm{E}[z^{u_{2,k}}\ 1\{u_{1,k}=i\}\mid \text{class-2 service}]\\&=\mathrm{E}[z^{n_{2,l}+f^2_{2,k}}\ 1\{\min(N,f^2_{1,k})=i\}\mid \text{class-2 service}]\\&=\mathrm{E}[z^{n_{2,l}}\mid n_{1,l}=0,n_{2,l}>0]\ \mathrm{E}[z^{f^2_{2,k}}\ 1\{\min(N,f^2_{1,k})=i\}]\ .\end{aligned} \tag{36}$$

This equals 0 for $i \geq N+1$ as the class-1 queue can only contain up to N packets and it is hence impossible that more than N class-1 packets are admitted during an elapsed service time. Finally,

$$\begin{aligned}&\mathrm{E}[z^{u_{2,k}}\ 1\{u_{1,k}=i\}\mid \text{class-1 service}]\\&=\mathrm{E}[z^{n_{2,l}+f^1_{2,k}}\ 1\{\min(N,n_{1,l}+f^1_{1,k})=i\}\mid \text{class-1 service}]\\&=\sum_{j=1}^{i}\mathrm{E}[z^{n_{2,l}}\ 1\{n_{1,l}=j\}\mid n_{1,l}>0]\ \mathrm{E}[z^{f^1_{2,k}}\ 1\{\min(N-j,f^1_{1,k})=i-j\}]\ .\end{aligned} \tag{37}$$

Note that this equals 0 for $i=0$ as a class-1 service can only be in progress when there are class-1 packets in the system.

On average, the time epoch between the beginning of start-slots l and $l+1$ consists of a single slot if the system is empty ($\Pr[n_{1,l}=n_{2,l}=0]$), of μ_2 slots if a class-2 packet is served ($\Pr[n_{1,l}=0,n_{2,l}>0]$) or of μ_1 slots if a class-1 packet is served ($\Pr[n_{1,l}>0]$). Therefore, γ, the steady-state probability that a random slot is a start-slot, is defined as

$$\begin{aligned}\gamma&=\lim_{k\to\infty}\Pr[\text{slot k is a start-slot}]\\&=\frac{1}{N_0(0)+\big(N_0(1)-N_0(0)\big)\mu_2+\big(1-N_0(1)\big)\mu_1}\ .\end{aligned} \tag{38}$$

Recall that if the server is idle during a slot, that slot is a start-slot. Therefore, the probability that the system is empty at the beginning of a random slot is given by

$$\begin{aligned}U_0(0)&=\lim_{k\to\infty}\Pr[u_{1,k}=u_{2,k}=0]\\&=\lim_{k,l\to\infty}\Pr[n_{1,l}=n_{2,l}=0,\text{slot k is a start-slot}]=\gamma N_0(0)\ .\end{aligned} \tag{39}$$

In steady state, the system is in stochastic equilibrium. Therefore, on average, the amount of packets effectively accepted by the system equals the amount of packets served by the system. This yields that the effective total load is found as $\rho_T^e = 1-U_0(0)$. The effective class-1 load and mean number of effective class-1 arrivals are therefore expressed as $\rho_1^e=\rho_T^e-\rho_2$ and $\bar{a}_1^e=\rho_1^e/\mu_1$ respectively. In a random slot, the server is empty, serving a class-1 packet or serving a class-2 packet. The probabilities of these events are

$$\begin{aligned}\Pr[\text{no service}]=&U_0(0)\ ,\\\Pr[\text{class-1 service}]=&\big(1-U_0(0)\big)\frac{\rho_1^e}{\rho_T^e}=\rho_1^e\ ,\\\Pr[\text{class-2 service}]=&\big(1-U_0(0)\big)\frac{\rho_2}{\rho_T^e}=\rho_2\ .\end{aligned} \tag{40}$$

Substituting equations (35)-(37) and (40) in (31) produces

$$\begin{aligned}\mathbf{u}(z) = U_0(0)\begin{bmatrix}1 & 0 & \cdots & 0\end{bmatrix} + \frac{\rho_2}{N_0(1)-N_0(0)}\big(\mathbf{n}(z)-\mathbf{n}(0)\big)\mathbf{H}_1\mathbf{F}_2(z)\begin{bmatrix}\mathbf{I} & \mathbf{0}\end{bmatrix} \\ + \frac{\rho_1^e}{1-N_0(1)}\mathbf{n}(z)\mathbf{D}\mathbf{F}_1(z)\begin{bmatrix}\mathbf{0} & \mathbf{I}\end{bmatrix} \ .\end{aligned} \tag{41}$$

The class-1 and class-2 system contents are respectively characterized by $\mathbf{u}(1)$ and $\mathbf{u}(z)\mathbf{e}$.

Using Little's law, the average delay can be found from the average system contents. Furthermore, the class-1 packet loss ratio, this is the fraction of class-1 packets that is rejected by the system, is easily obtained. It is given by

$$PLR_1 = \frac{\bar{a}_1 - \bar{a}_1^e}{\bar{a}_1} \ . \tag{42}$$

4. **Unfinished Work.** The total unfinished work at the beginning of slot k, denoted by $w_{T,k}$, is defined as the number of slots it takes to serve all packets in the system at the beginning of slot k, when no new packets arrive from slot k on. Furthermore, the unfinished work of class-j ($j = 1, 2$) at the beginning of slot k, denoted by $w_{j,k}$, is defined as the number of slots of the total unfinished work that are effectively spent on serving class-j packets.

As service times can have an infinite support, we cannot explicitly track the class-1 unfinished work, as we did for the class-1 system contents. Instead, it is also tracked by a pgf. Therefore, define the bivariate pgf of the unfinished work as

$$W(z_1, z_2) = \lim_{k\to\infty} \mathrm{E}[z_1^{w_{1,k}} z_2^{w_{2,k}}] \ . \tag{43}$$

Each class-j packet in the system at slot k contributes a class-j service time to the class-j unfinished work, except for the packet in service, only the remaining service time of that packet, denoted by s_j^+, should be accounted for. The elapsed and remaining service time are obviously correlated. In [7], p. 31, the bivariate pgf of the elapsed and remaining class-j service time is obtained as

$$S_j^*(x, y) = \mathrm{E}[x^{s_j^-} y^{s_j^+}] = y\frac{S_j(x) - S_j(y)}{\mu_j(x-y)} \ . \tag{44}$$

As each class-1 packet in the queue adds a class-1 service to the unfinished work of class-1, define the column vector

$$\mathbf{s}_1(z) = \left[S_1(z)^i\right]_{i=0..N}^T \ . \tag{45}$$

Analogously to the system contents, $W(z_1, z_2)$ is found by conditioning on the state of the server and by relating slot k to the preceding start-slot l. We have

$$\begin{aligned}W(z_1, z_2) = U_0(0) + \frac{\rho_2(\mathbf{n}(S_2(z_2)) - \mathbf{n}(0))}{S_2(z_2)(N_0(1) - N_0(0))}\mathbf{H}_1 S_2^*(\mathbf{Y}(S_2(z_2)), z_2)\mathbf{s}_1(z_1) \\ + \frac{\rho_1^e \mathbf{n}(S_2(z_2))}{1 - N_0(1)}\mathbf{D} S_1^*(\mathbf{Y}(S_2(z_2)), z_1)\mathbf{s}_1(z_1) \ .\end{aligned} \tag{46}$$

The spectral decomposition theorem enables calculation of this expression, as $S_j^*(x, y)$ can be seen as a function in a single argument x by keeping y constant at z_1. Hence we can evaluate it for $x = \mathbf{Y}(S_2(z_2))$ using spectral decomposition.

5. **Numerical Examples.** We aim to give some qualitative insight into the main factors governing system performance rather than performing a case study of a simplified model of a DiffServ router. Therefore, the choice of distributions (and their parameters) might seem ad hoc but any i.i.d. distributions may be chosen for the numerical results as long as the system is stable for these inputs. We study an output-queueing switch with L inlets and L outlets and two types of traffic as in [10]. On each inlet of the switch a batch arrives according to a Bernoulli process with parameter ν_T. A batch contains b (fixed) packets of class 1 with probability ν_1/ν_T or b packets of class 2 with probability ν_2/ν_T (with $\nu_1+\nu_2=\nu_T$). Incoming packets are routed uniformly to the outlets where they arrive at a queueing system as described in this paper. Therefore, all outlets can be considered identical and analysis of one of them is sufficient. The arrival process at the queueing system can consequently be described by the pmf

$$a(bn,bm)=\frac{L!\left(\frac{\nu_1}{L}\right)^n\left(\frac{\nu_2}{L}\right)^m\left(1-\frac{\nu_T}{L}\right)^{L-n-m}}{n!m!(L-n-m)!}\ , \tag{47}$$

for n and m integers with $n+m\leq L$ and by $a(p,q)=0$, for all other values of p and q. Obviously the number of arrivals of class-1 and class-2 are negatively correlated as there can be no more than $Lb-i$ class-2 arrivals in a slot with i class-1 arrivals. For increasing values of L, the correlation increases and the numbers of arrivals of both types become uncorrelated for L going to infinity. We now study a 4×4 output-queueing switch and assume the batch size $b=5$. The remainder of this section can be divided into two parts. First, our system will be compared with the system described in [10], to which we shall refer as 'the infinite system', because it is equivalent to our model with $N=\infty$. Evidently, packet loss is not accounted for in the infinite system and we shall study the impact hereof under varying traffic conditions. In the second part, the influence of the general service times on system performance will be examined.

In the first place, consider

$$\nu_1=\nu_2=0.02\ ,\ S_1(z)=1/4z^2+1/2z^3+1/4z^4\ ,\ S_2(z)=1/2z+1/2z^2\ . \tag{48}$$

Note that the arrival load is constant, we have $\rho_T=0.9$. The longer class-1 service times reflect the fact that real-time packets are often larger than data packets.

Figure 2 depicts the packet loss ratio versus the class-1 queue capacity N. Obviously the packet loss decreases with increasing N. The region between 10^{-2} and 10^{-3}, which we have marked in grey, is particularly interesting. Most real-time applications tolerate this amount of packet loss, but the infinite system does not capture packet loss and we will show that this causes it to be inaccurate for other performance measures in this region as well. A packet loss ratio over 10^{-2} causes the QoS delivered to real-time applications to be unacceptable and is hence impractical in a DiffServ setting. Systems with a very small packet loss ratio ($<<10^{-3}$) are accurately modelled by the infinite system.

The mean and the standard deviation of the system contents at the beginning of random slots of both classes are plotted versus N in Figure 3. The effect of the priority scheduling is clear as the class-2 system contents exceed those of class-1, despite that, on average, the system receives the same amount of packets of each class and that class-1 packets generally have longer service times. The values increase for increasing N and clearly converge to the values corresponding with the infinite system, represented by the horizontal dotted lines. This validates that, for

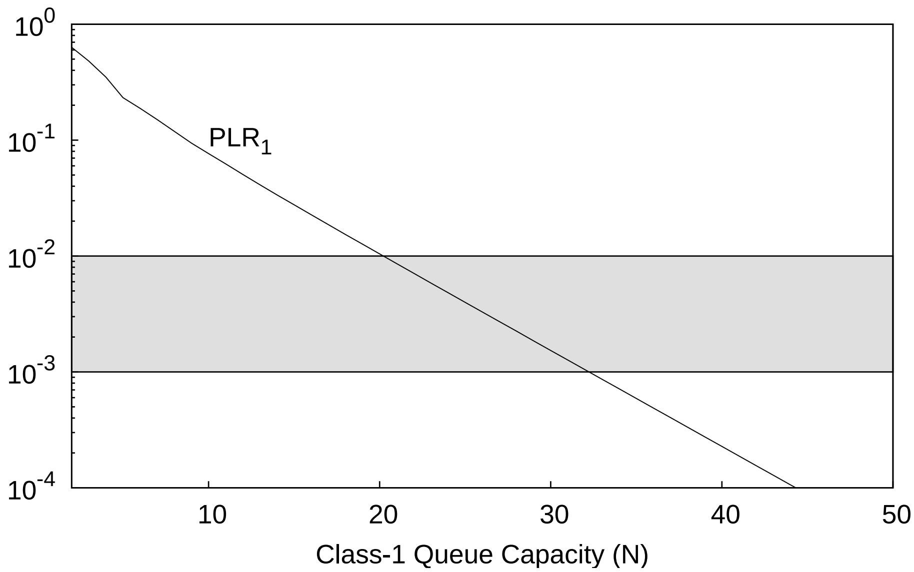

FIGURE 2. Packet loss ratio versus class-1 queue capacity.

N going to infinity, the system considered in this paper tends to the infinite system, as the number of dropped class-1 packets tends to zero. However, in our region of interest (again marked in grey) the infinite system considerably overestimates the mean value and standard deviation of the system contents of both classes. For instance, at $N = 21$, the smallest value for N where $PLR_1 < 10^{-2}$, the mean and standard deviation of the class-2 system contents are overestimated by 14% and 17% respectively.

the total arrival load approaches 100%. Next, for Figure 4 we fix $N = 20$ and vary $\nu_1 = \nu_2$ to consequently vary the total arrival load ρ_T between 0% and 100% while plotting the mean system contents of both classes. Instead of showing the packet loss on a separate figure, we have immediately marked the region of interest, where $10^{-3} < PLR_1 < 10^{-2}$, in grey. Again the effect of the priority scheduling is apparent. Especially note the class-2 starvation as the curves for the system under consideration and the infinite system seem to be close together, due to the steep slope, but for $\lambda_T \cong 0.875$, where the PLR_1 approaches the 1% boundary, the error introduced by using the infinite model over our model amounts to 8 and 13% for class-1 and class-2 respectively.

We now focus on the effect of the service times. Consider

$$N = 20,\ \nu_1 = 0.06,\ S_1(z) = z^2\ ,\ \nu_2 = 0.06/i,\ S_2(z) = z^i\ . \tag{49}$$

Note that $\mu_2 = i$. By increasing i, the average class-2 service time increases, while the average number of class-2 arrivals decreases. Therefore, the arrival load ρ_T remains constant at 0.9. Figure 5 depicts the average system contents at the beginning of random slots of both classes for average class-2 service time μ_2 from 1 to 40. Evidently, the decrease of $E[u_2]$ is caused by the decrease of μ_2 and hence λ_2, i.e., on average less packets arrive. This figure exemplifies the effect of class-2 traffic on class-1 traffic, as increasing values of μ_2 yield an increased probability that a class-1 packet arrives during the service time of a class-2 packet in a system

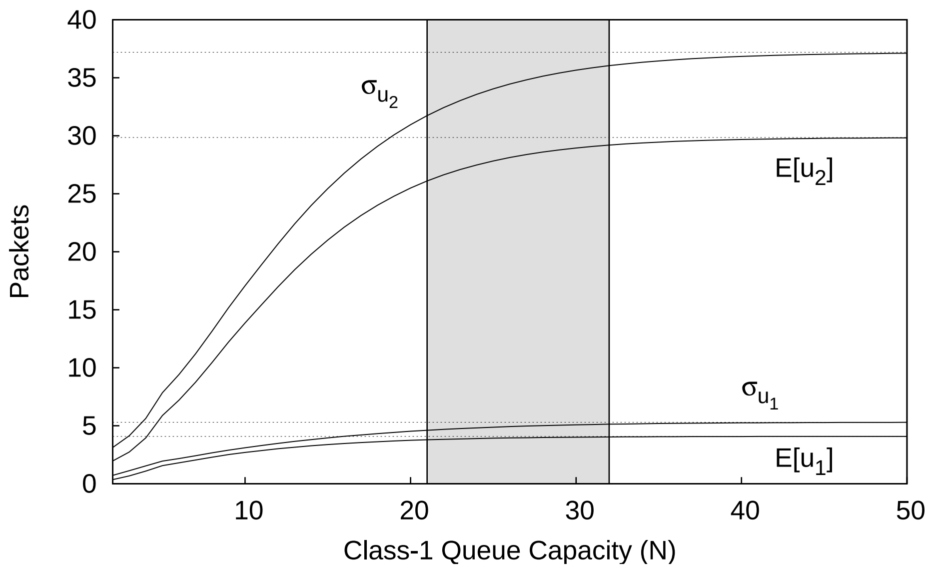

FIGURE 3. System contents versus class-1 queue capacity.

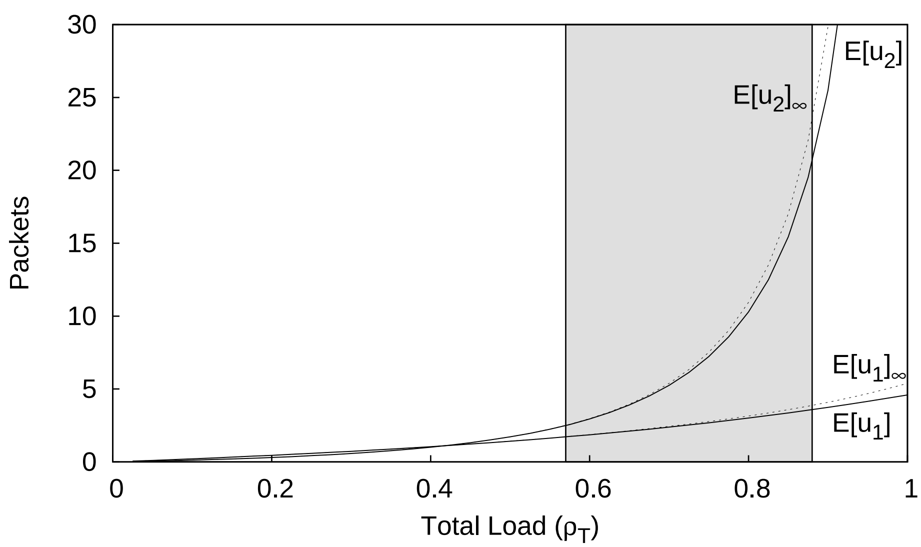

FIGURE 4. Average system contents vs arrival load.

void of class-1 packets and has to wait until the end of this service before it can enter the server.

Finally, consider

$$N = 20 \ , \ \nu_1 = \nu_2 = 0.15 \ , \ S_2(z) = \frac{z}{2-z} \ . \qquad (50)$$

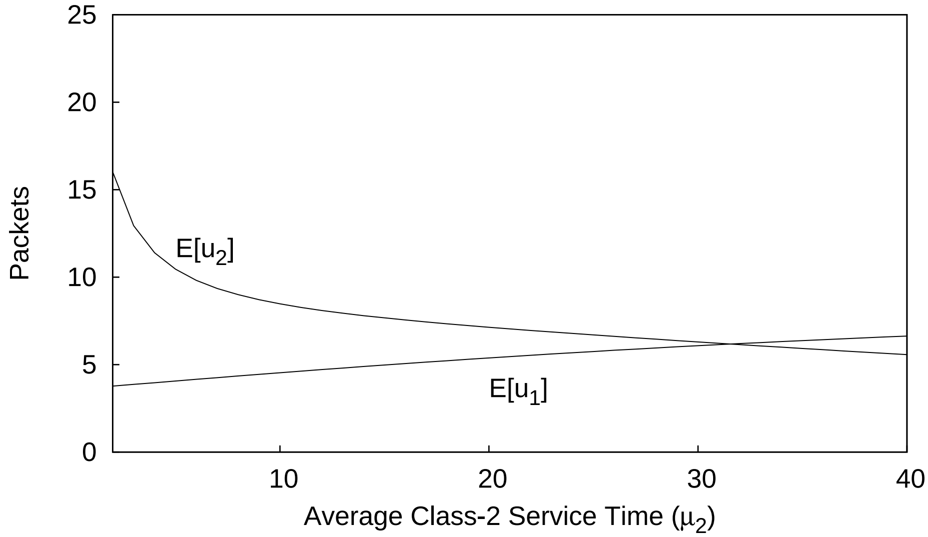

FIGURE 5. Average system contents versus average class-2 service time.

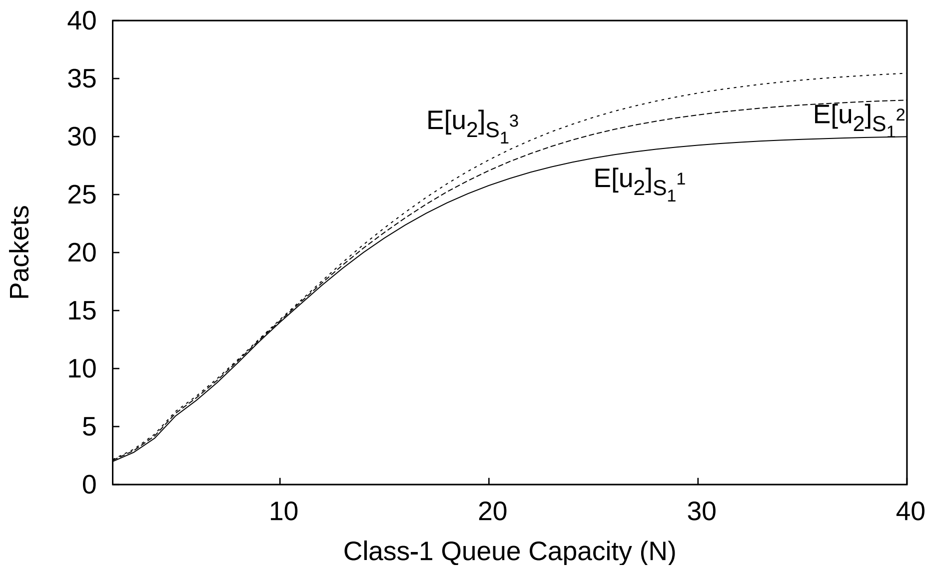

FIGURE 6. Average system contents versus class-1 queue capacity N.

In Figure 6 we have plotted the mean class-2 system contents versus the class-1 queue capacity for three distributions for the class-1 service times with the same mean $\mu_1 = 4$ slots, yielding $\rho_T = 0.9$, but different amounts of variance. In order of increasing variance, we have

$$S_1^1(z) = z^4 \ , \ S_1^2(z) = \frac{z}{4-3z} \ , \ S_1^3(z) = 0.7z + 0.3z^{11} \ . \tag{51}$$

Note that we have used the (shifted geometric) distribution for $S_2(z)$ and $S_1^2(z)$ to exemplify that service time distributions with unbounded support can be handled. The only requirement is that the first N derivatives of the pgf of the service time can be calculated as they are needed in (15) and (34). As expected, larger values of N correspond with higher class-2 system contents and higher variance of class-1 service times induces higher class-2 system contents. However, for smaller values of N, the effect is less explicit as class-1 packet loss is high in this case, hence dampening the effect of class-1 (service times) on class-2 performance.

6. Conclusions.

A two-class non-preemptive priority queue with finite capacity for high-priority packets has been studied in order to model a DiffServ router with Expedited Forwarding Per-Hop Behavior for high-priority traffic. This enables determination of high-priority packet loss. In a DiffServ router, the capacity for high-priority packets is often small to prevent this traffic from monopolizing the system and should thus not be approximated by infinity. Analytical formulas for system content of both traffic classes were determined making extensive use of the spectral decomposition theorem to cope with the difficulties that arise when considering general service times for both classes. Several numerical examples indicate the impact of small but practically feasible amounts of real-time packet loss on system performance, which is considerably different from what was predicted in existing models.

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] B. Carpenter and K. Nichols, *Differentiated services in the internet*, Proceedings of the IEEE, **90** (2002), 1479–1494.

[3] T. Demoor, J. Walraevens, D. Fiems, S. De Vuyst, and H. Bruneel, *Mixed finite-/infinite-capacity priority queue with interclass correlation and general class-1 service times (eds. K. Al Begain, D. Fiems and G. Horvath)*, "Lecture Notes in Computer Science: ASMTA'09", Springer-Verlag, Berlin, 5513 (2009), 264–278.

[4] T. Demoor, J. Walraevens, D. Fiems, S. De Vuyst, and H. Bruneel, *Modelling queue sizes in an Expedited Forwarding DiffServ router with service differentiation*, "Proceedings of QTNA'09", Singapore, (2009), paper 12.

[5] M. Mehmet Ali, and X. Song, *A performance analysis of a discrete-time priority queueing system with correlated arrivals*, Performance Evaluation, **57** (2004), 307-339.

[6] C. Meyer, Matrix Analysis and Applied Linear Algebra. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2000.

[7] H. Takagi, Queueing Analysis, Volume 3: Discrete-Time Systems. North-Holland Publishing Co., Amsterdam, 1993.

[8] T. Takine, B. Sengupta, and T. Hasegawa, *An analysis of a discrete-time queue for broadband ISDN with priorities among traffic classes*, IEEE Transactions on Communications, **42** (1994), 1837–1845.

[9] J. Van Velthoven, B. Van Houdt, and C. Blondia, *The impact of buffer finiteness on the loss rate in a priority queueing system*, "Lecture Notes in Computer Science: ASMTA'06" (eds. A. Horvath and M. Telek), Springer-Verlag, Berlin, 4045 (2006), 211–225.

[10] J. Walraevens, B. Steyaert, and H. Bruneel, *Performance analysis of the system contents in a discrete-time non-preemptive priority queue with general service times*, Belgian Journal of Operations Research, Statistics and Computer Science (JORBEL), **40** (2000), 91–103.

Received xxxx 20xx; revised xxxx 20xx.

E-mail address: thdemoor@telin.ugent.be
E-mail address: jw@telin.ugent.be
E-mail address: df@telin.ugent.be
E-mail address: sdv@telin.ugent.be
E-mail address: hb@telin.ugent.be