Uncovering the evolution from finite to infinite high-priority capacity in a priority queue

Joris Walraevens, Thomas Demoor, Dieter Fiems and Herwig Bruneel
Department of Telecommunications and Information Processing
Email: {jw,thdemoor,df,hb}@telin.UGent.be

***Abstract*—Infinite capacity queues are often used as approximation for their finite real-world counterparts as they are mathematically tractable. It is generally known that tail probabilities of low-priority system content in a two-class priority queue with infinite capacity for customers of both priority classes can be non-exponential, even if the interarrival time and service time distributions are exponentially decaying. In contrast, when the capacity for the high-priority customers is finite, tail probabilities of low-priority system content are always exponentially decaying. Therefore, using the results for one as an (accurate) approximation for the other is not obvious. From an analytical point of view, the non-exponentiality in the infinite case is caused by the arisal of an implicitly defined function, a root of the kernel, in the probability generating function for the low-priority system content. However, up till now, it has been unclear how this non-exponentiality suddenly emerges when taking the limit from to the finite to the infinite case. Our main contribution is that, under the restriction of a maximum of two arrivals per slot, a recurrence relation in the high-priority capacity is constructed resulting in an explicit expression for the corresponding generating function for the finite case. Amazingly, this expression contains all roots of the kernel in the infinite case. Taking the limit of this expression leads to the well-known behavior for the infinite case as the root inside the complex unit circle dominates the other roots uncovering the evolution from the finite to the infinite case. Furthermore, we investigate under which circumstances the standard tail characterizations are inaccurate.**

I. Introduction

Networks endure an ever-increasing tension as numerous applications, each requiring different Quality of Service (QoS) standards, concurrently utilize the network infrastructure. Priority queues are ideal tools to provide differentiated service. Therefore, they are widely implemented and have been studied extensively. It has been observed that modelling the high-priority queue capacity as finite or infinite leads to different low-priority tail behavior, which is closely related to packet loss, but how this shift in behavior arises remains unclear.

Abate and Whitt [1] were the first to prove that tails in an infinitely-sized priority queue are not necessarily exponential, even if the distributions of interarrival and service times are exponentially decaying. They heavily rely on singularity analysis of the Laplace transform of the low-priority waiting time in the complex plane and characterize three types of tails of the waiting time $w(t)$ of low-priority customers in a two-class $M/G/1$ priority queue, namely (i) $\sim \alpha t^{-3/2} e^{-\eta t}$, (ii) $\sim \alpha e^{-\eta t}$ and (iii) $\sim \alpha t^{-1/2} e^{-\eta t}$, with α and η constants depending on the arrival and service-time distributions. Depending on the parameters of the arrival and service processes, one of these three types of tail behavior appears.

Type (i) is encountered when the 'priority effect' dominates (large low-priority waiting time due to blocking by high-priority customers). In this case, the tail of the low-priority waiting time is related to the tail of the busy period of the high-priority queue, and the Laplace transform of the latter is expressed as an implicitly defined function. This implicit function, a solution of the so-called kernel equation, is the origin of the non-exponentiality of the tail, as it has a branch cut in the complex plane, rather than simple poles (the latter lead to exponentially decaying tails). This type is observed when the load of the low-priority class is low. On the other hand, exponential tails (ii) are encountered when the 'queueing effect' dominates (large low-priority waiting time due to many low-priority customers blocking each other) and is observed when the load of the low-priority class is high. Type (iii) is the boundary between the other cases (both of the aforementioned effects are equally strong). Furthermore, using large deviations principles, these results were verified [2]. See [3], [4] for divisions of the parameter space according to the three types and for the occurrence of the three types of tails for other random variables (f.i., the low-priority system content).

In contrast, when the capacity for the high-priority customers is limited, tail probabilities of the low-priority system content are *always* exponentially decaying (i.e., for all possible values of the involved parameters). Here, all singularities of the transform are (simple) poles, leading to purely exponential tails [5]. This is also apparent using matrix-analytic techniques [6], [7], where the terms "levels" and "phases" are used for the two dimensions of the queueing system. The high-priority capacity corresponds to the number of phases but treating an infinite amount of phases remains an open problem. Furthermore, it has been shown [8], [9] that truncation can lead to erroneous results concerning tail behavior. Recent research in matrix-analytic techniques has therefore focused on trying to cope with an infinite number of phases. Primary attention has been paid to obtaining the boundary condition for exponentiality, i.e. finding conditions under which the tails are exponential, for several subclasses of random walks [10], [11]. Consequently, the methods from literature can either handle infinite or finite capacity, but the evolution/limit from finite to infinite capacity is still not fully discovered (although for the QBD subcase some recent results give some hope [12], [13]).

The current contribution clarifies the evolution from the

finite to the infinite case[1] in the case of a discrete-time priority queue with a maximum of two high-priority arrivals in a slot. This restriction leads to a simpler model but allows us to outline the analysis method clearly. In the infinite case, the kernel of the functional equation of the bivariate probability generating function (pgf) of the system contents of both classes plays a capital role, as mentioned above. For the finite case, we extend previous work [5], where the matrix pgf of the low-priority system content was found. This previous work (the finite and infinite case) is summarized in section II. Next, we construct an important recurrence relation in the high-priority queue capacity in section III. A crucial relation between the characteristic polynomial of this recurrence relation in the finite case and the kernel in the infinite case is identified. We discover that, in the finite case, *all* roots of the kernel influence system behavior but *cancel* out each others branch cuts. In the limit to the infinite case, we show that the root inside the unit circle *dominates* the other roots, and the branch cut of this root is no longer canceled. Section IV provides some applications of the obtained explicit expression for $U_N(z)$.

II. Model and previous results

Consider a discrete-time absolute priority queueing system with class-1 customers having service priority over class-2 customers. Slots correspond to the (deterministic) service times. The class-1 queue capacity is limited to $N (\geq 0)$ customers, while there is no limit on the number of class-2 customers in the system. Arrivals occur i.i.d. from slot to slot. Let $a_i\,(i = 1, 2)$ denote the number of arrivals of class i in a random slot. The bivariate probability mass function (pmf) of the arrivals in a random slot is given by

$$a(m, n) = \text{Pr}[a_1 = m, a_2 = n]\,. \tag{1}$$

Let us denote partial probability generating functions (pgfs) of the number of class-2 arrivals in a slot given that the number of class-1 arrivals in that slot equals (exceeds) i by $A_i(z)$ ($A_i^*(z)$ resp.). This yields

$$A_i(z) = \text{E}[z^{a_2}\ 1\{a_1 = i\}] = \sum_{j=0}^{\infty} a(i, j)z^j\,, \tag{2}$$

$$A_i^*(z) = \sum_{l=i}^{\infty} A_l(z)\,, \tag{3}$$

with $1\{\cdot\}$ the indicator function. These partial pgfs characterize the numbers of arrivals in a slot (and thus the arrival process) completely. For instance, the normalization condition amounts to $\sum_{i=0}^{\infty} A_i(1) = A_0^*(1) = 1$; the mean numbers of class-1 and class-2 arrivals are given by

$$\lambda_1 = \sum_{i=0}^{\infty} iA_i(1) = \sum_{i=1}^{\infty} A_i^*(1), \tag{4}$$

$$\lambda_2 = \sum_{i=0}^{\infty} A_i'(1) = A_0^{*'}(1). \tag{5}$$

[1]We call the priority queue with finite (infinite resp.) high-priority capacity the finite (infinite) case from now on

We restrict the maximum number of class-1 arrivals per slot to two. Thus, $\forall z:\ A_i(z) = A_i^*(z) = 0,\ i > 2$. And, to exclude the degenerate case where there is no queueing for class-1, let $A_2(1) > 0$. We focus on the steady-state class-2 system content in the finite and infinite case. In [5], it is shown that the pgf of the class-2 system content in a system with class-1 capacity equal to N (the finite case) is given by

$$U_N(z) = (z - 1)p_{0,N}\big[1\ 0\ \cdots\ 0\big]\boldsymbol{X}_N(z)\boldsymbol{P}_N(z)^{-1}\mathbf{1}\,, \tag{6}$$

with $p_{0,N}$ the probability that the system is empty at the beginning of a slot (void of class-1 and class-2 customers), the $(N + 1) \times (N + 1)$ transition matrix $\boldsymbol{X}_N(z)$ given by

$$\boldsymbol{X}_N(z) = \begin{bmatrix} A_0(z) & A_1(z) & A_2(z) & 0 & \cdots & 0 & 0 \\ zA_0(z) & zA_1(z) & zA_2(z) & 0 & \cdots & 0 & 0 \\ 0 & zA_0(z) & zA_1(z) & zA_2(z) & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & & \cdots & & zA_0(z) & zA_1(z) & zA_2^*(z) \\ 0 & & \cdots & & 0 & zA_0(z) & zA_1^*(z) \end{bmatrix}. \tag{7}$$

for $N \geq 1$ and $\boldsymbol{X}_0(z) = [A_0^*(z)]$, with $\boldsymbol{P}_N(z) = z\boldsymbol{I} - \boldsymbol{X}_N(z)$ and with $\mathbf{1}$ an appropriately-sized column vector of 1's. Here and throughout the remainder, we assume that $N \geq 2$ (unless stated otherwise). The cases $N = 0$ and $N = 1$ are easily handled directly from (6) and we will include these in the results when appropriate. The determinant of $\boldsymbol{P}_N(z)$ plays a crucial role, as this determinant is the denominator of $U_N(z)$ through $\boldsymbol{P}_N(z)^{-1}$. Its roots are potentially poles of $U_N(z)$ and the decay of the geometric term(s) in the pmf/tail of the class-2 system content is characterized by these poles.

The infinite case is studied in [4]. Here, the pgf of the class-2 system content is given by

$$U_\infty(z) = p_{0,\infty}\frac{A_0^*(z)(z - 1)(Y_1(z) - 1)}{(z - Y_1(z))(A_0^*(z) - 1)}, \tag{8}$$

with $p_{0,\infty} = 1 - \lambda_1 - \lambda_2$. As $Y_1(z)$ is the unique root of the kernel with $|x| < 1$ when $|z| < 1$, the kernel plays a crucial role. The kernel is given by

$$F(x, z) = \sum_{i=0}^{\infty} A_i(z)x^i - x \tag{9}$$

$$= A_2(z)x^2 + (A_1(z) - 1)x + A_0(z). \tag{10}$$

As the kernel turns out to be quadratic in x due to the restriction on the class-1 arrivals, it has two roots given by

$$Y_1(z) = \frac{1 - A_1(z) - \sqrt{(1 - A_1(z))^2 - 4A_0(z)A_2(z)}}{2A_2(z)} \tag{11}$$

and

$$Y_2(z) = \frac{1 - A_1(z) + \sqrt{(1 - A_1(z))^2 - 4A_0(z)A_2(z)}}{2A_2(z)} \tag{12}$$

The square-root in the expression of $Y_1(z)$ causes the non-exponential tail probabilities in $U_\infty(z)$, as it gives rise to branch cuts and branch points (points where the expression under the square root equals 0). This paper will unveil that $Y_1(z)$ also appears in the expression for $U_N(z)$, but that its square-root is in fact canceled by the square-root of $Y_2(z)$.

III. Calculation of $U_N(z)$

In this section, an explicit expression for $U_N(z)$ is established. In the process, we uncover a crucial relation between the characteristic polynomial of a recurrence relation for the determinant in the finite case and the kernel in the infinite case. Furthermore, the expression for $U_N(z)$ also correctly converges to $U_\infty(z)$ when taking the limit for N. First, some manipulations on the matrices in (6) will be performed. Note that we start count of rows and columns at 0.

Lemma 1. *The function $U_N(z)$ can be written as*

$$U_N(z) = (z-1)p_{0,N} \frac{\left(\sum_{i=0}^{2} A_i(z) \sum_{j=0}^{N} \boldsymbol{adj}(\boldsymbol{P}_N(z))_{ij}\right)}{z^N D_N(z)}, \; N \geq 2, \quad (13)$$

with $\boldsymbol{adj}(\boldsymbol{P}_N(z))$ the adjugate matrix of $\boldsymbol{P}_N(z)$ and $D_N(z)$ the determinant of

$$\boldsymbol{Q}_N(z) = \begin{bmatrix} z-A_0(z) & -A_1(z) & -A_2(z) & \cdots & 0 & 0 \\ -A_0(z) & 1-A_1(z) & -A_2(z) & \cdots & 0 & 0 \\ 0 & -A_0(z) & 1-A_1(z) & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & & \cdots & -A_0(z) & 1-A_1(z) & -A_2^*(z) \\ 0 & & \cdots & 0 & -A_0(z) & 1-A_1^*(z) \end{bmatrix}. \quad (14)$$

Proof: Multiplication of vector $\begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}$ and matrix $\boldsymbol{X}_N(z)$ in (6) results in $\begin{bmatrix} A_0(z) & A_1(z) & A_2(z) & 0 & \cdots & 0 \end{bmatrix}$. Multiplication of the adjugate matrix of $\boldsymbol{P}_N(z)$ with $\mathbf{1}$ leads to the column vector $[\sum_{j=0}^{N} \boldsymbol{adj}(\boldsymbol{P}_N(z))_{ij}]$. Furthermore, all elements of $\boldsymbol{P}_N(z)$ but the ones in the first row have a factor z. Therefore, it is easily seen that $\det(\boldsymbol{P}_N(z)) = z^N D_N(z)$. ∎

Let us commence by calculating $D_N(z)$. To that end, a linear homogeneous recurrence relation for $\{D_N(z)\}_{N=0}^{\infty}$ is constructed, which turns out to be crucial. This recurrence relation is then solved by means of generating functions.

Theorem 1. *The determinant $D_N(z)$ is a solution of the recurrence relation*

$$D_N(z) = (1-A_1(z))D_{N-1}(z) - A_0(z)A_2(z)D_{N-2}(z), \quad N \geq 2 \quad (15)$$

with seed functions

$$D_0(z) = z-1, \quad (16)$$

$$D_1(z) = z(1-A_1^*(z)) - A_0(z). \quad (17)$$

Proof: We first subtract the second row of $Q_N(z)$ from its first row, which does not affect its determinant $D_N(z)$. Laplace expansion along the last row (and then last column) of this matrix leads to

$$D_N(z) = (1-A_1^*(z))E_{N-1}(z) - A_0(z)A_2^*(z)E_{N-2}(z), \quad N \geq 2, \quad (18)$$

with $E_N(z)$ the determinant of the $(N+1)\times(N+1)$ matrix

$$\begin{bmatrix} z & -1 & 0 & 0 & \cdots & 0 & 0 \\ -A_0(z) & 1-A_1(z) & -A_2(z) & 0 & \cdots & 0 & 0 \\ 0 & -A_0(z) & 1-A_1(z) & -A_2(z) & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & & \cdots & & -A_0(z) & 1-A_1(z) & -A_2(z) \\ 0 & & \cdots & & 0 & -A_0(z) & 1-A_1(z) \end{bmatrix}. \quad (19)$$

Notice that the matrices giving rise to $E_N(z)$ and $D_N(z)$ only differ in the last column. Again performing Laplace expansion of (19) along the last row (and then last column), it is clear that $E_N(z)$ fulfills a recurrence relation:

$$E_N(z) = (1-A_1(z))E_{N-1}(z) - A_0(z)A_2(z)E_{N-2}(z), \quad N \geq 2, \quad (20)$$

with seed functions

$$E_1(z) = (1-A_1(z))E_0(z) - A_0(z), \quad (21)$$

$$E_0(z) = z. \quad (22)$$

Eliminating $E_{N-2}(z)$ from expressions (18) and (20) leads to

$$D_N(z) = E_N(z) - A_2(z)E_{N-1}(z), \; N \geq 1. \quad (23)$$

Note that expression (17) can also be obtained from this expression.

Since $D_N(z)$ is a linear combination of $E_N(z)$ and $E_{N-1}(z)$ for $N \geq 1$, $D_N(z)$ fulfils the same recurrence equation as $E_N(z)$, i.e.,

$$D_N(z) = (1-A_1(z))D_{N-1}(z) - A_0(z)A_2(z)D_{N-2}(z), \quad N \geq 3. \quad (24)$$

In order for this recurrence relation to be valid for $N = 2$, $D_0(z)$ should be chosen as in (16), which completes the proof. ∎

Remark 1. *Note that $D_0(z)$ and $D_1(z)$ are chosen such that the recurrence relation is valid for all $N \geq 2$. Therefore, $D_0(z)$ has no real meaning as determinant.*

Remark 2. *Theorem 1 states that $\{D_N(z)\}_{N=0}^{\infty}$ is a linear homogeneous recurrence relation of order 2. The order is due to the maximum number of class-1 arrivals in a slot.*

Solving the linear homogeneous recurrence relation in theorem 1 can easily be achieved, for instance by means of generating functions.

Lemma 2. *The generating function $D(x,z)$ of $\{D_N(z)\}_{N=0}^{\infty}$, defined as*

$$D(x,z) = \sum_{N=0}^{\infty} D_N(z)x^N, \quad (25)$$

is given by

$$D(x,z) = A_0(z)\frac{(1-A_0^*(z)-(z-1)A_2(z))x + (z-1)}{F(A_0(z)x, z)}, \quad (26)$$

with $F(x,z)$ the kernel in the infinite case (expression (10)).

Proof: Multiplying all terms in (15) by x^N and summing over all valid N leads to an expression for $D(x,z)$ as a function of $D_0(z)$ and $D_1(z)$. Inserting these seed functions leads to (26). ∎

Remark 3. *The denominator of the generating function is directly related to the characteristic polynomial of the underlying recurrence relation and the roots of this polynomial lead*

to geometric terms in the final expression of $D_N(z)$. Both surprisingly and crucially, the characteristic polynomial is related to the kernel $F(x,z)$ in the infinite case (see expression (26)), and the root $Y_1(z)$ of the kernel will thus appear in the final expression of $D_N(z)$ in the finite case. This turns out to be the crucial link between the finite and the infinite cases.

Theorem 2. *The determinant $D_N(z)$ is given by*

$$D_N(z) = \frac{(1-A_0^*(z))}{A_2(z)(Y_2(z)-Y_1(z))}\left[\frac{z-Y_1(z)}{1-Y_1(z)}\left(\frac{A_0(z)}{Y_1(z)}\right)^N - \frac{z-Y_2(z)}{1-Y_2(z)}\left(\frac{A_0(z)}{Y_2(z)}\right)^N\right], \quad N \geq 0\,. \quad (27)$$

Here $Y_1(z)$ and $Y_2(z)$ are as defined in (11) and (12).

Proof: The function $D_N(z)$ is calculated by writing expression (26) as power series in x. This can be done by partial fraction expansion of the rational expression. As the denominator equals $F(xA_0(z),z)$, its roots in x are $Y_1(z)/A_0(z)$ and $Y_2(z)/A_0(z)$. The partial fraction expansion then equals

$$D(x,z) = \frac{(z-Y_1(z))(A_2(z)Y_1(z)-A_0(z))Y_2(z)}{(Y_1(z)-Y_2(z))A_0(z)\left(1-x\frac{A_0(z)}{Y_1(z)}\right)} + \frac{(z-Y_2(z))(A_2(z)Y_2(z)-A_0(z))Y_1(z)}{(Y_2(z)-Y_1(z))A_0(z)\left(1-x\frac{A_0(z)}{Y_2(z)}\right)}. \quad (28)$$

Writing both terms on the right as geometric series (cf. (25)), identifying the coefficients of x^N on both sides, and using that $Y_1(z)$ and $Y_2(z)$ are roots of kernel (10) leads to (27). ∎

We still need to calculate the numerator of (13). However, as the used methodology is completely analogous to the one used for the denominator: write the terms of the numerator as functions of several determinants, which are very similar to $D_N(z)$. For all these determinants, suitable linear recurrence relations are then constructed and solved. Therfore, we have omitted the details here and immediately state the resulting lemma.

Lemma 3. *The numerator of expression (13) for $U_N(z)$ is equal to*

$$\frac{(z-1)p_{0,N}A_0^*(z)z^N\left[\left(\frac{A_0(z)}{Y_1(z)}\right)^N - \left(\frac{A_0(z)}{Y_2(z)}\right)^N\right]}{A_2(z)(Y_2(z)-Y_1(z))}. \quad (29)$$

Theorem 3. *The pgf $U_N(z)$ of the class-2 system content is given by*

$$U_N(z) = \left(\frac{1-\lambda_1}{1-(A_2(1)/A_0(1))^N}-\lambda_2\right)\frac{(z-1)A_0^*(z)}{(1-A_0^*(z))} \frac{\left(Y_2(z)^N - Y_1(z)^N\right)}{\left(\frac{z-Y_1(z)}{1-Y_1(z)}Y_2(z)^N - \frac{z-Y_2(z)}{1-Y_2(z)}Y_1(z)^N\right)}. \quad (30)$$

Proof: The numerator of $U_N(z)$ is given by (29), while the denominator is given by $z^N D_N(z)$; $D_N(z)$ is given by equation (27). Finally, $p_{0,N}$ is calculated by using the normalization condition $U_N(1)=1$, leading to

$$p_{0,N} = \frac{1-\lambda_1}{1-(A_2(1)/A_0(1))^N} - \lambda_2. \quad (31)$$

∎

Corollary 1. *The correct limiting behavior from the finite to the infinite case is established as $\lim_{N\to\infty} U_N(z) = U_\infty(z)$.*

Proof: For z inside the complex unit circle, it is easily proved that $|Y_1(z)| < 1 < |Y_2(z)|$ (through f.i. Rouché's theorem and the implicit function theorem). Therefore, when taking the limit of (30) for $N \to \infty$, the terms in $Y_2(z)^N$ dominate those in $Y_1(z)^N$, both in numerator and denominator. Furthermore, for a stable system, $\lambda_1 < 1$, which results in $A_2(1) < A_0(1)$. Therefore $\lim_{N\to\infty}(A_2(1)/A_0(1))^N = 0$. These two observations immediately lead to (8). ∎

Corollary 2. *If the $A_i(z)$ ($i = 0,1,2$) are meromorphic, $U_N(z)$ is meromorphic and thus cannot have branch points.*

Proof: If the $A_i(z)$ ($i=0,1,2$) are meromorphic so are the seed values of all recurrence relations used in this paper. As the recurrence relations consist of basic operations and the meromorphic functions form a field with respect to the usual pointwise operations followed by redefinition at the removable singularities, evidently $D_N(z)$, $F_N(z)$, $H_N(z)$, $K_N(z)$, $L_N(z)$, $M_N(z)$ and finally $U_N(z)$ are all meromorphic. ∎

Remark 4. *This corollary asserts that $U_N(z)$ (expression (30)) cannot contain branch points and thus the square root in $Y_1(z)$ is canceled by the square root in $Y_2(z)$ for each finite N. This is highly comparable to the expression of the N-th Fibonacci number,*

$$\frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^N - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^N, \quad (32)$$

that 'seems' to contain $\sqrt{5}$ while the Fibonacci numbers are obviously integer as the square-roots in both terms cancel out.

IV. Applications

Although the main point of this paper is obtaining an "explicit" expression for $U_N(z)$, achieved in (30), and studying the limit behavior to $U_\infty(z)$, this explicit expression can also be used for other purposes. By taking the derivative of (30) and evaluating in $z=1$, the average of u_{2_N}, the class-2 system content in a priority queue with class-1 capacity N is

$$\mathrm{E}[u_{2_N}] = \lambda_2 + \frac{1}{Y_2(1)^N(1-\lambda_T)+\lambda_2)}\left[Y_2(1)^N\lambda_{12} + \frac{Y_2(1)^N\lambda_{11}\lambda_2}{2(1-\lambda_1)} + \frac{(Y_2(1)^N-1)\lambda_{22}}{2} + \frac{\lambda_2}{Y_2(1)-1} + \frac{[(1-\lambda_1)Y_2'(1)-\lambda_2Y_2(1)]Y_2(1)^{N-1}N}{(Y_2(1)^N-1)}\right]. \quad (33)$$

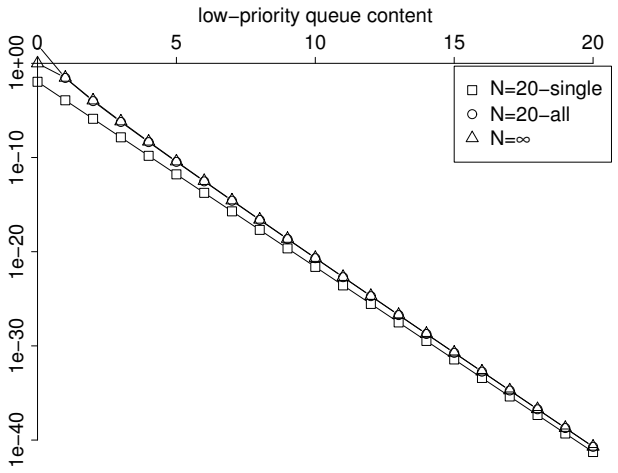

Fig. 1. Low-priority system content for $N = 20$ single pole and all poles and $N = \infty$ for $\lambda_1 = 0.4,\ \lambda_2 = 0.01$

Furthermore, this allows us to study how u_{2_N} approaches u_{2_∞}, the class-2 system content in a priority queue with infinite capacity. For arbitrarily large N , we have

$$\mathrm{E}[u_{2_N}] \sim \mathrm{E}[u_{2_\infty}] + \frac{[(1-\lambda_1)Y_2'(1) - \lambda_2 Y_2(1)]N}{(1-\lambda_T)Y_2(1)^{N+1}}. \quad (34)$$

Evidently, $\mathrm{E}[u_{2_N}] < \mathrm{E}[u_{2_\infty}]$ as $Y_2(1) > 1,\ Y_2'(1) < 0$. Also, as $Y_2(1) > 1$, the linear evolution in N in the numer is dampened by the appearance of $Y_2(1)^N$ in the denominator so the "convergence rate" slows down as N increases.

Also, the explicit expression allows more efficient calculation of the the tail behavior. The tail of u_{2_N} is obtained by investigating the behavior of its pgf around its poles. First, the poles need to be calculated numerically, as an analytic description of the poles remains to be found. Next, the residue needs to be computed in these poles. Both these operations are more efficient using (30) instead of (6), especially for large N. In figure 1, we plot the tail of $u_{2_{20}}$ using only the dominant pole (squares), all poles (circles) and the tail of u_{2_∞} (triangles) for the 2×2 switch arrival process [4] with $\lambda_1 = 0.4,\ \lambda_2 = 0.01$. The single (dominant) pole approximation performs badly (difference $\sim 10^2$) as the poles lie close together. When all poles are taken into account, the results lie very close to the ones obtained in the infinite case.

V. Conclusions and outlook

In this paper, first and foremost, an explicit expression for the low-priority system content in a priority queue with high-priority system capacity equal to N and unbounded low-priority capacity (the finite case) is established uncovering the evolution to the priority queue with unbounded capacity for both classes (the infinite case). In the infinite case, the root of the kernel that lies inside the complex unit circle plays a crucial role. In the current contribution, a recurrence relation in the high-priority capacity is constructed for the finite case. We discover a crucial relation between the characteristic polynomial of this recurrence relation and the kernel and we show that all roots of the kernel pop up in the resulting explicit expression for the system content. However, the resulting tail behavior is exponential as these roots cancel out each others branch cuts. Additionally, taking the limit of this expression leads to the correct behavior for the infinite case as the root inside the complex unit circle dominates the other roots allowing for non-exponential behavior.

For practical applications, the explicit expression for the finite case is an additional tool, as it can be easily inverted analytically or numerically. Furthermore, although the dominant pole governs the behavior of the true tail, in the region of practical interest, one might need to take multiple poles into account. The used method should be applicable for a broader class of queues. We expect that at least the priority/tandem models from [2], where fluid flow and large deviations principles were used, falls within this class. Furthermore, we hope to relax the restriction on the high-priority arrivals to any number. This will of course affect the degree of the kernel and the order of the recurrence relation.

Acknowledgment

The first and third authors are Postdoctoral Fellows with the Research Foundation, Flanders (F.W.O.-Vlaanderen), Belgium.

References

[1] J. Abate and W. Whitt, "Asymptotics for M/G/1 low-priority waiting-time tail probabilities," *Queueing Systems*, vol. 25, no. 1-4, pp. 173–233, 1997.

[2] I. Adan, M. Mandjes, W. Scheinhardt, and E. Tzenova, "On a generic class of two-node queueing systems," *Queueing Systems*, vol. 61, no. 1, pp. 37–63, 2009.

[3] K. Laevens and H. Bruneel, "Discrete-time multiserver queues with priorities," *Performance Evaluation*, vol. 33, no. 4, pp. 249–275, 1998.

[4] 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, pp. 1807–1829, 2003.

[5] T. Demoor, J. Walraevens, D. Fiems, S. De Vuyst, and H. Bruneel, "Influence of real-time queue capacity on system contents in diffserv's expedited forwarding per-hop-behavior," *Journal of Industrial and Management Optimization*, vol. 6, no. 3, pp. 587–602, 2010.

[6] 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*, vol. 4054, pp. 211–225, 2006.

[7] K. Al-Begain, A. Dudin, A. Kazimirsky, and S. Yerima, "Investigation of the M(2)/G(2)/1/infinity, N queue with restricted admission of priority customers and its application to HSDPA mobile systems," *Computer Networks*, vol. 53, no. 8, pp. 1186–1201, 2009.

[8] D. Kroese, W. Scheinhardt, and P. Taylor, "Spectral properties of the tandem Jackson network, seen as a quasi-birth-and-death process," *Annals of Applied Probability*, vol. 14, no. 4, pp. 2057–2089, 2004.

[9] Y. Sakuma and M. Miyazawa, "On the effect of finite buffer truncation in a two-node Jackson network," *Journal of Applied Probability*, vol. 42, no. 1, pp. 199–222, 2005.

[10] M. Miyazawa and Y. Zhao, "The stationary tail asymptotics in the GI/G/1-type queue with countably many background states," *Advances in Applied Probability*, vol. 36, no. 4, pp. 1231–1251, 2004.

[11] M. Miyazawa, "A Markov renewal approach to M/G/1 type queues with countably many background states," *Queueing Systems*, vol. 46, no. 1-2, pp. 177–196, 2004.

[12] ——, "Tail Decay Rates in Double QBD Processes and Related Reflected Random Walks," *Mathematics of Operations Research*, vol. 34, no. 3, pp. 547–575, 2009.

[13] F. Guillemin and J. van Leeuwaarden, "Rare event asymptotics for a random walk in the quarter plane," *Queueing Systems*, vol. 67, no. 1, pp. 1–32, 2011.