Joint Queue-Length Distribution for the Non-Preemptive Multi-Server Multi-Level Markovian Priority Queue

Josef Zuk \(^1\) and David Kirszenblat \(^2\)

\(^1\)Defence Science and Technology Group (DSTG)
\(^2\)Affiliation not available

December 7, 2023

Abstract

Explicit results are obtained using simple and exact methods for the joint queue-length distribution of the M/M/c queue with an arbitrary number of non-preemptive priority levels. This work is the first to provide explicit results for the joint probability generating function and joint probability mass function for a general number of priority levels. A fixed-point iteration is developed for the stationary balance equations, which enables direct computation of the joint queue-length distribution. A multi-variate probability generating function is also derived, from which the joint probability mass function can be computed by means of a multi-dimensional fast Fourier transform method.
Joint Queue-Length Distribution for the Non-Preemptive Multi-Server Multi-Level Markovian Priority Queue

Josef Zuk
Defence Science and Technology Group, Melbourne, Australia, josef.zuk@defence.gov.au

David Kirszenblat
Defence Science and Technology Group, Melbourne, Australia, david.kirszenblat@defence.gov.au

Explicit results are obtained using simple and exact methods for the joint queue-length distribution of the $M/M/c$ queue with an arbitrary number of non-preemptive priority levels. This work is the first to provide explicit results for the joint probability generating function and joint probability mass function for a general number of priority levels. A fixed-point iteration is developed for the stationary balance equations, which enables direct computation of the joint queue-length distribution. A multi-variate probability generating function is also derived, from which the joint probability mass function can be computed by means of a multi-dimensional fast Fourier transform method.

Key words: queueing theory; non-preemptive priority; queue length distribution
MSC2000 subject classification: Primary: 90B22; secondary: 60K25, 60J74
OR/MS subject classification: Primary: Queues: Priority; secondary: Queues: Markovian

1. Introduction

This work is concerned with the development of practical algorithms for computing the joint queue-length distribution for the non-preemptive Markovian priority queue with a general number of priority levels. In the most recent edition of their textbook, Shortle et al. [11] remark that ‘the determination of stationary probabilities in a non-preemptive Markovian system is an exceedingly difficult matter, well near impossible when the number of priorities exceeds two’. Elmelegy [6] has also commented that ‘given the immense literature studying non-preemptive priority queuing systems, it is hard to find a simple and exact method that calculates the performance measures of non-preemptive priority systems with more than two priority levels’. The present discussion serves to fill this knowledge gap.

Previous work on the non-preemptive priority $M/M/c$ queue has been reviewed recently in [14]. The vast majority of effort concerning the joint distribution has focused on the two-level problem. The single source of previous work addressing the joint queue-length distribution for more than two priority levels comprises the papers of Wignall [12] and Wignall and Enns [13]. These consider single-server systems involving multiple queues with distinct arrival rates ranked by priority level, and with feedback – requiring deterministic or probabilistic transitions between queues before system exit is achieved. The present problem can, in principle, be constructed as a special case of this scheme. In [12, 13], equations that must be solved recursively are presented for the multivariate probability generating function (PGF); but no general solution is provided, and manual solution becomes increasingly cumbersome as the number of priority levels grows beyond a small number. No actual probability mass functions (PMFs) are computed. By contrast, we present an explicit closed-form expression for the joint PGF given any number of priority levels, that lends itself to practical numerical evaluation of the joint PMF. In [12], explicit results are confined to the probability that the system is empty, and the probability that a given queue is being served at a random observation. In [13], the distribution of the maximum queue length during a busy period
in the presence of probabilistic feedback is also given. The approach that we have adopted in the present paper may be viewed as an extension of the method employed by Cohen [3] for the two-level case. With it, we are able, in the words of Neuts [10], ‘to obtain actual results, that is numbers and insight from numbers’.

Priority-level numbers significantly greater than two are encountered in numerous real-world applications, such as health care [5]. In a hospital emergency department (ED), arrivals are prioritized according to patient acuity level [9]. Most hospitals operate with at least five acuity levels. In a forthcoming paper, the results obtained here will be applied to the ambulance ramping problem [1], in which arrivals to the ED by ambulance or as walk-ins are categorized into three priority levels, corresponding to high, intermediate or low patient acuity. The model is further complicated by the fact that there are two arrival classes (ambulance and walk-in) each with their own arrival rate and each of which contain patients of multiple priority levels. Thus, there is an entanglement between arrival classes and priority levels.

The number of servers is denoted by $c$ and the number of priority levels by $K$. Each priority level is associated with a Poisson arrival rate $\lambda_\kappa$, $\kappa = 1, 2, \ldots, K$, leading to a total arrival rate $\lambda = \sum_{\kappa=1}^{K} \lambda_\kappa$. A common service rate $\mu$, associated with an exponential distribution, is assumed for all priority levels. Thus, the total traffic intensity is given by $r = \lambda/(c\mu)$. As we are interested in the state-state queue-length distribution, possible values of $r$ will be limited to the ergodic region $r < 1$. The level traffic intensity for priority $\kappa = 1, 2, \ldots, K$ is defined as $r_\kappa = \lambda_\kappa/(c\mu)$, so that $r = \sum_{\kappa=1}^{K} r_\kappa$. It is also convenient to introduce priority-level fractions $0 \leq \nu_\kappa \leq 1$, summing to unity, such that $r_\kappa = r\nu_\kappa$. Alternatively, the $\nu_\kappa \geq 0$ may be chosen without constraint, provided we set $r_\kappa = r\nu_\kappa/\parallel\nu\parallel_1$. This facilitates exploration of different distributions among the priority levels for a given constant total load on the system. To test the numerical performance of the algorithms developed here, we shall fix $r$ and compute distributions for a random sample of vectors $\nu = (\nu_1, \nu_2, \ldots, \nu_K)$ chosen independently from the unit probability simplex in $K$-dimensions $\Delta_K \equiv \{ \nu \in \mathbb{R}^K : \parallel \nu \parallel_1 = 1, \nu \geq 0 \}$. We also define the partial (per server) traffic intensity as $\rho \equiv \lambda/c\mu$ so that $r = c\rho$, in line with the notation of [8].

The rest of the paper is organized as follows: In Section 2, we set up the stationary balance equations for the model. It is shown that, for any number of priority levels, they can be stated in compact form comprising just a single equation. This leads to a nearest-neighbour relationship among the elements of the joint PMF, that is subsequently used as a diagnostic test of the computational procedures developed later. It is also shown that the balance equation can be solved directly for the joint PMF by means of a fixed-point iteration (FPI). While the FPI does not constitute an efficient computational algorithm, it provides an important benchmark for verifying the correctness of the vastly more efficient methodology discussed in the remainder of the paper. In Section 3, the general balance equation is used to derive an explicit closed-form expression for the multi-variate PGF of the joint distribution. It is shown that the correct marginals follow from the joint PGF. Section 4 describes a multi-dimensional fast Fourier transform-based (FFT) method that computes the joint PMF from the multi-variate PGF. It is equally applicable to the marginal distributions. Measures of performance for various diagnostic tests are presented in Section 5, the results of which are discussed there. Conclusions follow in Section 6. Various technical details appear in the Appendices.

2. Stationary Balance Equations For ease of illustration, we shall begin by considering the concrete example of a system with three priority levels: high, intermediate, low. Let the single-subscripted quantity $p_k$ denote the probability that the queue is empty and that $k$ of the $c$ servers are busy. Let the triple-subscripted quantity $p_{\ell mn}$ denote the probability that there are $\ell$ high-priority clients, $m$ intermediate-priority clients and $n$ low-priority client in the queue, and that all
servers are busy. Since, apart from the identification \( p_c = p_{000} \), the probabilities \( p_k \) and \( p_{\ell mn} \) are exhaustive and mutually exclusive, we must have that

\[
\sum_{k=0}^{c-1} p_k + \sum_{\ell,m,n=0}^{\infty} p_{\ell mn} = 1. \tag{1}
\]

Figure 1 displays the state transition diagram for the two-level case, which is easily drawn and captures all the qualitative features of the general problem. With the definitions above, the stationary balance equations for the three-level system are given by

\[
\begin{align*}
(a) \quad \lambda p_k &= (k+1)\mu p_{k+1} \\
(b) \quad (\lambda + c\mu)p_{000} &= \lambda p_{c-1} + c\mu(p_{010} + p_{011} + p_{001}) \\
(c) \quad (\lambda + c\mu)p_{00n} &= \lambda_3 p_{00n-1} + c\mu(p_{01n} + p_{01n+1} + p_{00n+1}) \\
(d) \quad (\lambda + c\mu)p_{0mn} &= \lambda_2 p_{0m-1n} + \lambda_3 p_{0mn-1} + c\mu(p_{1mn} + p_{0m+1n}) \\
(e) \quad (\lambda + c\mu)p_{\ell mn} &= \lambda_1 p_{\ell-1mn} + \lambda_2 p_{\ell m-1n} + \lambda_3 p_{\ell mn-1} + c\mu p_{\ell+1mn}.
\end{align*}
\]

We adopt the convention that negative valued subscripts correspond to zero probability. It follows immediately from (a) that

\[
p_k = \frac{1}{k!} \left( \frac{\lambda}{\mu} \right)^k p_0,
\]

for \( k = 0, 1, \ldots, c \). Therefore,

\[
p_{000} \equiv p_c = \frac{1}{c!} \left( \frac{\lambda}{\mu} \right)^c p_0. \tag{4}
\]

With the observation that \( \lambda p_{c-1} = c\mu p_c = c\mu p_{000} \), we can recast the collection of balance equations above for the \( p_{\ell mn} \) into the single equation

\[
(\lambda + c\mu)p_{\ell mn} = \lambda_1 p_{\ell-1mn} + \lambda_2 p_{\ell m-1n} + \lambda_3 p_{\ell mn-1} + c\mu(p_{\ell+1mn} + \delta_0 p_{\ell m+1n} + \delta_0 p_{\ell+1mn} + \delta_0 p_{\ell mn+1} + \delta_0 p_{\ell mn}). \tag{5}
\]
One should note that (5) is homogeneous in $p_{\ell mn}$, which means that its solution is determined only up to an arbitrary scale factor. It is convenient to set $\tilde{p}_{\ell mn} = p_{\ell mn}/p_{000}$ so that $\tilde{p}_{000} = 1$. The normalization condition (1) then implies that

$$\frac{1}{p_{000}} = \sum_{k=0}^{c-1} \frac{c!}{k!} \left( \frac{\mu}{\lambda} \right)^{-k} + \sum_{\ell, m, n = 0}^\infty \tilde{p}_{\ell mn}.$$  

Thus, we may solve (5) for $\tilde{p}_{\ell mn}$ by setting $\tilde{p}_{000} = 1$, and subsequently determine $p_{000}$ from (6), so that the probabilities can be recovered via $p_{\ell mn} = p_{000} \cdot \tilde{p}_{\ell mn}$.

Next, let us write

$$\Phi(\ell) \equiv \sum_{m,n} p_{\ell mn},$$

and observe that we also have

$$\sum_{m,n=0}^\infty p_{\ell m-1n} = \sum_{m,n=0}^\infty p_{\ell mn-1} = \Phi(\ell).$$

Moreover,

$$\sum_{m,n=0}^\infty \delta_{0\ell} p_{\ell m+1n} = \delta_{0\ell} \left[ \Phi(0) - \sum_{n=0}^\infty p_{00n} \right],$$

$$\sum_{m,n=0}^\infty \delta_{0\ell} \delta_{0m} p_{\ell mn+1} = \delta_{0\ell} \sum_{n=0}^\infty p_{00n} - p_{000},$$

and so we see that these three contributions sum to $\Phi(0)$. Accordingly, (5) implies that

$$(\lambda_1 + c\mu) \Phi(\ell) = \lambda_1 \Phi(\ell - 1) + c\mu \left[ \Phi(\ell + 1) + \delta_{0\ell} \Phi(0) \right],$$

for $\ell = 0, 1, \ldots$. If we let $\Delta \Phi(\ell) \equiv \Phi(\ell) - \Phi(\ell - 1)$, then $\Delta \Phi(\ell + 1) = r_1 \Delta \Phi(\ell)$ for $\ell = 1, 2, \ldots$, subject to $\Delta \Phi(1) = -(1 - r_1) \Delta \Phi(0)$ and where $r_1 = \lambda_1/(c\mu)$. This is solved by $\Delta \Phi(\ell) = \varphi(\ell) \cdot \Delta \Phi(0)$, with $\varphi(\ell) \equiv -(1 - r_1) r_{\ell-1}$. It is equivalent to the first-order recurrence equation

$$\Phi(\ell) = \Phi(\ell - 1) + \varphi(\ell) \cdot \Phi(0),$$

for $\ell = 1, 2, \ldots$ which, in turn, is solved by

$$\Phi(\ell) = \Phi(0) \cdot \left[ 1 + \sum_{k=0}^{\ell} \varphi(k) \right] = \Phi(0) r_{\ell}^r,$$

for $\ell = 0, 1, \ldots$. The quantity $\Phi(\ell)$ represents the probability that there are $\ell$ high-priority clients in the queue and that all servers are busy. For $\ell \geq 1$, it simply gives the probability of $\ell$ high-priority clients in the queue without qualification because having one or more high-priority clients in the queue implies that all servers must be busy. Let $\Psi(\ell)$ denote the probability of $\ell$ high-priority clients in the queue. Then, for $\ell \geq 1$, $\Psi(\ell) = \Phi(\ell) = \Phi(0) r_{\ell}^r$ and

$$1 = \sum_{\ell=0}^\infty \Psi(\ell) = \Psi(0) + \sum_{\ell=1}^\infty \Phi(\ell),$$
from which it follows that \( \Psi(0) = 1 - [r_1/(1 - r_1)] \Phi(0) \). Thus, we may write

\[
\Psi(\ell) = [1 - \Phi(0)/(1 - r_1)] \delta_{0\ell} + \Phi(0)r_1^\ell. \tag{14}
\]

The normalization condition (1) implies that

\[
\sum_{\ell=0}^{\infty} \Phi(\ell) = 1 - P_{NW}, \quad P_{NW} \equiv \sum_{k=0}^{c-1} p_k, \tag{15}
\]

where \( P_{NW} \) represents the no-wait probability. We may equate this with the summation over \( \Phi(\ell) \) as given by (12) to obtain \( \Phi(0) = (1 - r_1)(1 - P_{NW}) \). Hence, (14) can be re-expressed in the more direct form

\[
\Psi(\ell) = P_{NW}\delta_{0\ell} + (1 - P_{NW})(1 - r_1)r_1^\ell. \tag{16}
\]

This result applies equally to the general problem with an arbitrary number of priority levels. The only undetermined component in the foregoing discussion is the value of \( p_0 \), which represents the probability that the system is empty, and is given by \([8]\)

\[
\frac{1}{p_0} = \sum_{k=0}^{c} \frac{\rho^k}{k!} + \frac{\rho^{c+1}}{c!(c - \rho)}, \tag{17}
\]

where \( \rho \equiv \lambda/\mu \). Knowledge of \( p_0 \) allows us to obtain \( p_{000} = p_c \). Another way of characterizing it is as follows: Let the random variable \( N_{sys} \) represent the number of clients in the system. It is well-known that, for some constant \( A \), \( \Pr(N_{sys} = n) = Ar^n \) for all \( n \geq c \). By construction, \( p_{000} = \Pr(N_{sys} = c) \). We also have \( P_{NW} = \Pr(N_{sys} \leq c - 1) \). Hence,

\[
P_{NW} = \sum_{n=0}^{c-1} \Pr(N_{sys} = n) = 1 - \sum_{n=c}^{\infty} \Pr(N_{sys} = n)
= 1 - \frac{A r^c}{1 - r} = 1 - \frac{1}{1 - r} \Pr(N_{sys} = c), \tag{18}
\]

which leads to the result \( p_{000} = (1 - r)(1 - P_{NW}) \).

The stationary balance equations as given in (5) are easily extended to the general problem of an arbitrary number \( K \) of priority levels. In order to establish a compact expression, we introduce the lattice vectors \( n \equiv [n_1, n_2, \ldots, n_K] \in \mathbb{Z}^K \), and let \( e_\kappa, \kappa = 1, 2, \ldots, K \) denote the standard unit Cartesian coordinate basis vectors in \( \mathbb{Z}^K \). Then the stationary balance equations for \( K \) priority levels can be expressed as

\[
(1 + r)p_n = \prod_{j=1}^{K} \delta_{0n_j} p_n + \sum_{\kappa=1}^{K} r_\kappa p_{n-e_\kappa} + \prod_{j=1}^{K} \delta_{0n_j} p_{n+e_\kappa}, \tag{19}
\]

where we recall the convention that \( p_n \equiv 0 \) if \( n_\kappa < 0 \) for any \( \kappa = 1, 2, \ldots, K \). If we introduce the boundary set\(^1\)

\[
B \equiv \{n \in \mathbb{Z}^K : n_\kappa \leq 0 \text{ for some } \kappa = 1, 2, \ldots, K\}, \tag{20}
\]

then, for all \( n \not\in B \), we have the interior stationary balance equations

\[
p_n = \frac{1}{1 + r} \left[ p_{n+e_1} + \sum_{\kappa=1}^{K} r_\kappa p_{n-e_\kappa} \right]. \tag{21}
\]

\(^1\)Strictly speaking, this is the boundary and beyond or, equivalently, the set of non-interior points.
Thus every probability in the interior region is a positive weighted sum of its lower nearest neighbours plus its upper highest priority neighbour. While it is not a numerically stable proposition to attempt to first solve for $p_n$ on the boundary set $B$ and then use (21) to propagate the solution into the interior region, the relationship (21) does provide a robust diagnostic test of where a candidate solution for $p_n$ behaves as it should. We shall call this the nearest-neighbour test. The two-dimensional instance of this test was used extensively for the two-level problem in [14]. The wait-conditional joint queue-length PMF is given by $P(n) = (1-r)p_n$. The full unconditional joint PMF is then expressed as

$$P_{\text{full}}(n) = P_{\text{NW}}\delta(n) + (1 - P_{\text{NW}})P(n)$$

$$= P_{\text{NW}}\delta(n) + (1 - P_{\text{NW}})(1 - r)p_n,$$

where $\delta(n) \equiv \prod_{\kappa=1}^{K} \delta_{0n_{\kappa}}$, and recalling that $\bar{p}_0 = 1$.

The system (19) can be solved directly using an FPI. However, this necessitates truncation of the problem to a maximum queue size for each priority level. Consider the space $A^K_\infty$ of countably infinite matrices in $K$-dimensions, whose elements are non-negative and indexed by the $K$-dimensional vector $n = [n_1, n_2, \ldots, n_K]$, with $n_\kappa = 0, 1, \ldots$, for each $\kappa = 1, 2, \ldots, K$. In other words, $n \in \mathbb{N}_0^K$ — the space of $K$-tuples of non-negative integers. We define the matrix mapping $M: A^K_\infty \to A^K_\infty$ by

$$A' = M(A), \quad A' = \frac{1}{1+r} \left[ \prod_{\kappa=1}^{K} \delta_{0\kappa} \cdot A_n + \sum_{\kappa=1}^{K} \left( r_{\kappa} A_{n-e_{\kappa}} + \prod_{j=1}^{\kappa-1} \delta_{0j} \cdot A_{n+e_{\kappa}} \right) \right].$$

We observe that the mapping $M$ is sum preserving:

$$A' = M(A) \Rightarrow \sum_{n \in \mathbb{N}_0^K} A_n = \sum_{n \in \mathbb{N}_0^K} A_n,$$

assuming that the sum is finite. This is equivalent to $\|A'\|_1 = \|A\|_1$ for the $L_1$ matrix norm, and follows from (10), which trivially extends to the general multi-level case.

We shall consider the collection of probabilities $p_n$ as defining a matrix $P \in A^K_\infty$. The $L_1$ and $L_\infty$ norms are given, respectively, by

$$\|P\|_1 \equiv \sum_n |P_n|, \quad \|P\|_\infty \equiv \max_n |P_n|.$$
more efficient FFT method developed in the following sections. Not only will it confirm that the joint queue-length distribution has been computed correctly, there will be an implied confirmation of the multi-variate PGF on which the FFT method rests. Results of this comparison are illustrated in Figure (7) and discussed later on in the results section. It is difficult to use Monte Carlo (MC) simulation to check a multi-variate distribution in its entirety due to size requirements and lack of appropriate hypothesis tests. We have, however, performed the more straightforward task of checking marginal distributions against discrete-event MC simulation. The FPI method does have the singular advantage that one may expect it to be robust to the incorporation into the model of a variety of complications (e.g. unequal service rates, or more complex queue disciplines), whereas the closed-form PGF may not survive such modifications.

3. Probability Generating Function Generalizing the approach of Cohen [3], let us introduce the collection of multi-variate functions of $K - 1$ continuous variables

$$G_{\ell}(u) \equiv P_0 \cdot \sum_{m_1, \ldots, m_{K-1} = 0}^{\infty} \tilde{p}_{\ell m_1 \ldots m_{K-1}} u_{m_1} \cdots u_{m_{K-1}},$$

for $\ell = 0, 1, \ldots$, where $u \equiv [u_1, \ldots, u_{K-1}]$. With the identification $P_0 \equiv 1 - r$, we have

$$G_0(0) = 1 - r, \quad \sum_{\ell = 0}^{\infty} G_{\ell}(1) = 1,$$

and $G_{\ell}(u)$ represents the PGF for the wait-conditional joint queue-length PMF, such that

$$P(n) = \prod_{\kappa = 1}^{K-1} \frac{1}{n_{\kappa+1}} \frac{\partial^{n_{\kappa+1}}}{\partial u_{\kappa}^{n_{\kappa+1}}} G_{n_1}(u_1, \ldots, u_{K-1}) \bigg|_{u=0}. \quad (28)$$
Then, the PGF for the full unconditional PMF is given by
\[
G_{\text{full, } \ell}(u) = P_{NW} \delta_{0\ell} + (1 - P_{NW}) G_{\ell}(u).
\] (29)

Summing over the stationary balance equations yields the linear recurrence relations
\[
G_{\ell+1} + [\alpha - (1 + r)] G_{\ell} + r_1 G_{\ell-1} = 0,
\] (30)
for \( \ell \geq 1 \), where
\[
\alpha \equiv r \cdot u = \sum_{\kappa=1}^{K-1} r_{\kappa} u_{\kappa+1}.
\] (31)

The characteristic equation reads
\[
\lambda^2 + [\alpha - (1 + r)] \lambda + r_1 = 0,
\] (32)
and is solved by \( \lambda = \lambda_{\pm} \) with
\[
\lambda_{\pm} = \frac{1}{2} \left[ 1 + r - \alpha \pm \sqrt{(1 + r - \alpha)^2 - 4r_1} \right].
\] (33)

When \( u = 1 \equiv [1, \ldots, 1] \), in which case \( \alpha = r - r_1 \), we require, from the foregoing discussion of the high-priority marginal, that \( G_\ell \propto r_1^\ell \), and we have \( \lambda_- = r_1 \) whereas \( \lambda_+ = 1 \). It follows immediately that
\[
G_{\ell}(u) = G_0(u) \lambda_-(u),
\] (34)
for \( \ell = 0, 1, \ldots \). It remains to solve the problem for \( \ell = 0 \).

For this purpose, let us write \( n \equiv (\ell, m) \) with \( m = (m_1, \ldots, m_{K-1}) \in \mathbb{N}^{K-1} \) and \( q_{m}^\ell \equiv p_{\ell m} \). Then, for \( \ell = 0 \), and on setting
\[
Q_m \equiv \sum_{j=1}^{K-1} q_{m+e_j}^0 \cdot \prod_{i=1}^{j-1} \delta_{0m_i},
\] (35)
the birth-death equations (19) may be cast as
\[
(1 + r)q_m^0 = \prod_{j=1}^{K-1} \delta_{0m_j} \cdot q_m^0 + \sum_{j=1}^{K-1} r_{j+1} q_{m-e_j}^0 + q_{m+e_j}^1 + Q_m.
\] (36)

Summing over the multi-indices \( m \) with the corresponding powers of \( u \), we see that the stationary balance equations (19) require that
\[
(1 + r)G_0(u) = G_0(0) + \alpha G_0(u) + G_1(u) + \sum_{m \in \mathbb{N}^{K-1}} Q_m \cdot \prod_{j=1}^{K-1} u_j^{m_j}.
\] (37)

Noting that \( G_1(u) = \lambda_- G_0(u) \), and using the identity \( \lambda_+ + \lambda_- = 1 + r - \alpha \), we arrive at
\[
\lambda_+ G_0(u) = G_0(0) + \sum_{m \in \mathbb{N}^{K-1}} Q_m \cdot \prod_{j=1}^{K-1} u_j^{m_j},
\] (38)
which may be further manipulated to yield

\[
\lambda_+ G_0(u_1, \ldots, u_{K-1}) = \sum_{k=1}^{K-1} \frac{1}{u_k} \left[ G_0(0, \ldots, 0, u_k, \ldots, u_{K-1}) - G_0(0, \ldots, 0, u_{k+1}, \ldots, u_{K-1}) \right] + G_0(0, \ldots, 0)
\]

\[
= \frac{1}{u_1} G_0(u_1, \ldots, u_{K-1}) + \sum_{k=2}^{K-1} \left( \frac{1}{u_k} - \frac{1}{u_{k-1}} \right) G_0(0, \ldots, 0, u_k, \ldots, u_{K-1}) + \left( 1 - \frac{1}{u_{K-1}} \right) G_0(0, \ldots, 0).
\]

(39)

On formally setting \( u_K \equiv 1 \), this may be simplified as

\[
G_0(u_1, \ldots, u_{K-1}) = \frac{1}{1/u_1 - \lambda_+(u_1, \ldots, u_{K-1})} \sum_{k=2}^{K} \left( \frac{1}{u_{k-1}} - \frac{1}{u_k} \right) G_0(0, \ldots, 0, u_k, \ldots, u_{K-1}).
\]

(40)

It should be noted that \( G_0'(u_1, \ldots, u_{K-1}) \equiv G_0(0, \ldots, 0, u_k, \ldots, u_{K-1}) \) is the result for an appropriately aggregated \((K-k+1)\)-dimensional problem, obtained by treating the \( k \) highest priority levels as a single high priority level with level traffic intensity \( r_{hi} = \sum_{k=1}^{K} r_k \). One may also note that \( G_0(0, \ldots, 0) = P_0 \).

One immediate consequence of (40) is that the marginal distribution of the aggregation of the top \( p \) priority levels is geometric, for all \( p = 1, 2, \ldots, K \). From (34), the PGF for this distribution is

\[
G^{(p)}_{agg}(u) \equiv \sum_{\ell=0}^{\infty} u^{\ell} G_{\ell}(u^{(p)}) = \frac{G_0(u^{(p)})}{1 - u \lambda_-(u^{(p)})},
\]

(41)

where \( u^{(p)} \equiv (u_1, \ldots, u_{K-p}) \). Only a single term survives the summation in (40) in the evaluation of \( G_0(u^{(p)}) \), to produce

\[
G_{agg}^{(p)}(u) = \frac{1-u}{1-u \lambda_+(u^{(p)})[1-u \lambda_-(u^{(p)})]} G_0(0^{p-1}, 1_{K-p}).
\]

(42)

Now, we have

\[
\alpha^{(p)} \equiv r \cdot u^{(p)} = (\sigma_p - \sigma_1)u + r - \sigma_p.
\]

(43)

where \( \sigma_p \equiv \sum_{k=1}^{p} r_k \). It follows that we can write

\[
\lambda_\pm(u^{(p)}) = \frac{1}{2} \left[ 1 + r_{agg} - r_{lo} u \pm \sqrt{(1 + r_{agg} - r_{lo} u)^2 - 4 r_{hi}} \right],
\]

(44)

with

\[
\begin{align*}
& r_{lo} = \sigma_p - \sigma_1, \quad r_{hi} = \sigma_1, \quad r_{agg} = r_{lo} + r_{hi} = \sigma_p.
\end{align*}
\]

(45)

Hence,

\[
\frac{1-u}{(1-u \lambda_+)(1-u \lambda_-)} = \frac{1}{1-\sigma_p u},
\]

(46)

and the desired result follows, consistent with intuitive expectations.

The solution of the system (40) is derived in Appendix A. Here, we explain the overall strategy: We begin by noting that the recurrence relation (30) is similar to equation (1.2.10) of Cohen [3], but with his analogue of \( r_2 u_1 \) replaced by our \( \alpha \) defined in (31). The analogue of (1.2.12) still holds, but the analogue of (1.2.11) acquires additional complicating terms of the form \( G_0(0, \ldots, 0, u, v, w, \ldots) \). However, these can be dealt with by means of a simple aggregation trick. This is because the leading zeros mean that we looking at probabilities of various numbers of low priority entities in
the queue and no entities with the highest \( n \) priorities. But this is the same as no entities of a single highest priority level aggregated from the highest \( n \) priorities. Thus, we have to solve a lower-dimensional problem. The relevant observation can be expressed, for example in a 4-level problem, as \( G_0(0, v, w) = G'_0(v, w) \), \( G_0(0, 0, w) = G'_0(w) \), and of course \( G_0(0, 0, 0) = G'_0() = P_0 \). The PGFs with lesser numbers of arguments refer to appropriately aggregated lower-dimensional problems. Therefore, one has to start with Cohen’s two-dimensional problem and successively work upwards to the full \( K \)-dimensional problem. The recurrence relations take on a simpler, more transparent structure, that is amenable to explicit solution, when one changes notation by writing the arguments of the PGFs backwards. Accordingly, it is convenient to present the result for \( G_0() \) in terms of new variables \((z_1, z_2, \ldots, z_{K-1})\) where ascending indices are associated with ascending priority level. Thus \( z_k = u_{K-k} \), in which case

\[
G_0(0, \ldots, 0, u_k, \ldots, u_{K-1}) = G_0(0, \ldots, 0, z_{K-k}, \ldots, z_2, z_1) \tag{47}
\]

or, equivalently, \( G'_0(u_k, \ldots, u_{K-1}) = G'_0(z_{K-k}, \ldots, z_2, z_1) \). The result is then given for a new version of \( G_0() \) re-parameterized according to \( G_0(z_1, \ldots, z_{K-1}) \equiv G'_0(z_{K-1}, \ldots, z_2, z_1) \), which is just \( G'_0() \) with its arguments flipped.

A self-contained summary of Appendix A is as follows: To construct the PGF for the problem of \( K \) non-preemptive priority levels, we consider the joint PMF \( p_{\ell m_1 \ldots m_{K-1}} \), where \( \ell \) enumerates the number in the queue of the highest priority level, and write the PGF as

\[
G_\ell(z_1, \ldots, z_{K-1}) \equiv \sum_{m_1 \ldots m_{K-1} = 0}^\infty p_{\ell m_1 \ldots m_{K-1}} z_1^{m_1} \ldots z_{K-1}^{m_{K-1}}, \tag{48}
\]

so that \( z_1 \) is associated with the lowest priority level and \( z_{K-1} \) with the next-to-highest priority level. We focus our attention on the \( \ell = 0 \) component \( G_0(z_1, \ldots, z_{K-1}) \), and write \( P_0 \equiv G_0(0, \ldots, 0) \). Then, we have

\[
G_0(z_1, \ldots, z_{K-1}) = P_0 \prod_{\kappa=1}^{K-1} \frac{1 - z_k \zeta_+ (z_1, \ldots, z_{K-1})}{1 - z_k \zeta_+ (z_1, \ldots, z_{K})}. \tag{49}
\]

The \( \zeta_+(z_1, \ldots, z_{K}) \) solve the quadratic equation

\[
\zeta^2 + [\beta(z_1, \ldots, z_{K}) - (1 + r)] \zeta + \sigma_{K-k} = 0, \tag{50}
\]

for \( \kappa = 1, 2, \ldots, K - 1 \). We have defined \( \sigma_{K-k} \equiv \sum_{k=1}^{K-\kappa} r_k \), so that \( \sigma_{K-K} = \sum_{k=1}^{K-k} r_k \) and \( \sigma_K = r \). Also,

\[
\beta(z_1, \ldots, z_{K}) \equiv \sum_{k=1}^{K-1} z_k r_{K+1-k}, \tag{51}
\]

for \( \kappa = 1, 2, \ldots, K - 1 \). We adopt the conventions that \( z_0 \equiv 1, \beta() \equiv 0 \). The solutions \( \zeta_\pm \) are given by

\[
\zeta_\pm (z_1, \ldots, z_{K}) = \frac{1}{2} \left[ 1 + r - \beta(z_1, \ldots, z_{K}) \pm \sqrt{(1 + r - \beta(z_1, \ldots, z_{K}))^2 - 4 \sigma_{K-k}} \right]. \tag{52}
\]

It follows that \( \zeta_+() = 1, \zeta_-() = r \).

The explicit expression for the two-level problem \((K = 2)\) is

\[
G_0(z_1) = P_0 \frac{1 - z_1}{1 - z_1 \zeta_+(z_1)} = P_0 \frac{1 - z_1}{1 - \frac{z_1}{2} \left[ 1 + r - r_2 z_1 + \sqrt{(1 + r - r_2 z_1)^2 - 4 r_1} \right]}, \tag{53}
\]
which is in agreement with Cohen [3], noting that \( P_0 = 1 - r \) if \( G_0(.) \) is to represent the wait-
conditional PGF. The explicit expression for the three-level problem \( (K = 3) \) is

\[
G_0(z_1, z_2) = P_0 \cdot \frac{1 - z_1 - z_2 \zeta_+ (z_1)}{1 - z_1 \zeta_+ (z_1)} \cdot \frac{1 - z_2 \zeta_+ (z_1, z_2)}{1 - z_1, z_2} = P_0 \cdot \frac{1 - z_1}{2} \left[ 1 + r - r_3 z_1 + \sqrt{(1 + r - r_3 z_1)^2 - 4(r_1 + r_2)} \right] \\
\times \frac{1 - z_2}{2} \left[ 1 + r - r_3 z_1 + \sqrt{(1 + r - r_3 z_1)^2 - 4(r_1 + r_2)} \right] \frac{1 - z_1}{2} \left[ 1 + r - r_3 z_1 - r_2 z_2 + \sqrt{(1 + r - r_3 z_1 - r_2 z_2)^2 - 4r_1} \right].
\]

(54)

The result for the full wait-conditional PGF is given by

\[
G_\ell(z_1, \ldots, z_{K - 1}) = G_0(z_1, \ldots, z_{K - 1}) \cdot \zeta_\ell (z_1, \ldots, z_{K - 1}),
\]

(55)

for \( \ell = 0, 1, \ldots \).

### 3.1. Marginal Probabilities

We first consider the low-priority marginal PMF for the two-
level \( (K = 2) \) problem. The low-priority marginal PGF is a function of a single continuous complex
variable \( z \), and is obtained by summing (55) over the discrete high-priority queue lengths, to yield

\[
G_{l0}(z) = \sum_{\ell=0}^{\infty} G_\ell(z) = \frac{(1 - r)(1 - z)}{[1 - z\zeta_+(z)][1 - \zeta_-(z)]},
\]

(56)

with

\[
\zeta_\pm(z) = \frac{1}{2} \left[ 1 + r - r_2 z \pm \sqrt{(1 + r - r_2 z)^2 - 4r_1} \right],
\]

(57)

which agrees with Cohen’s result [3]. By using the quadratic identities

\[
\zeta_+(z) + \zeta_-(z) = 1 + r - r_2 z , \quad \zeta_+(z) \cdot \zeta_-(z) = r_1,
\]

(58)

which also imply that

\[
[1 - \zeta_+(z)] \cdot [1 - \zeta_-(z)] = r_2(z - 1),
\]

(59)

one is able to express \( G_{l0}(z) \) in various equivalent and more convenient forms:

\[
G_{l0}(z) = \frac{1 - r \cdot r - \zeta_-(z)}{1 - r z} = \frac{1 - r}{\zeta_+(z) - r}.
\]

(60)

Since \( \zeta_+(1/r) = r, \zeta_-(1/r) = r_1/r \), the first form shows that \( G_{l0}(z) \) has a pole at \( z = 1/r > 1 \). We
also have \( \zeta_+(1) = 1, \zeta_-(1) = r_1 \), but it should be clear that the apparent pole at \( z = 1 \) implied by
(56) is spurious due to a cancellation with the numerator.

The \( p \)-th marginal for the general \( K \)-level problem, with \( p = 1, 2, \ldots, K - 1 \), has its PGF given by

\[
G_{\text{mrg}}^{(p)}(z_p) \equiv \sum_{\ell=0}^{\infty} G_{\ell}(1, \ldots, z_p, \ldots, 1) = \frac{G_0(1, \ldots, z_p, \ldots, 1)}{1 - \zeta_-(1, \ldots, z_p, \ldots, 1)},
\]

(61)

and pertains to the \( p \)-th lowest priority level or, equivalently, the \((K+1-p)\)-th highest priority level. Each \( G_{\text{mrg}}^{(p)}(z_p) \) has the same functional form as \( G_{l0}(z) \) for the two-level problem. Let us relabel (57) as

\[
\zeta_\pm(z; r_{hi}, r_{lo}) = \frac{1}{2} \left[ 1 + r_{\text{sum}} - r_{lo} z \pm \sqrt{(1 + r_{\text{sum}} - r_{lo} z)^2 - 4r_{hi}} \right],
\]

(62)
by setting \( r_{hi} = r_1, r_{lo} = r_2, r_{sum} = r_{hi} + r_{lo} \), so that we also have
\[
G_{lo}(z; r_{hi}, r_{lo}) = \frac{1 - r_{sum}}{\zeta_{+}(z; r_{hi}, r_{lo}) - r_{sum}}.
\] (63)

Then \( G_{mrg}(z_p) = G_{lo}(z; r_{hi}, r_{lo}) \) on making the identifications
\[
r_{lo} = r_{K+1-p}, \quad r_{hi} = \sum_{k=1}^{K-p} r_k.
\] (64)

The derivation of this result is relegated to the second appendix. Its interpretation is as follows: All priority levels above the \( p \)-th lowest level are aggregated into a single high level whose level traffic intensity is the sum of the aggregated values. The low level is identified with the \( p \)-th lowest level and assigned its level traffic intensity, namely \( r_{K+1-p} \). Finally, all priority levels below the \( p \)-th are discarded, so the overall traffic intensity for the equivalent two-level problem becomes \( r_{sum} = r_{lo} + r_{hi} \).

### 3.2. Exclusively-Low Probabilities
To consider the case where there are assumed to be only the lowest-priority clients in the queue, we must set \( z_2 = z_3 = \cdots = z_{K-1} = 0 \), to obtain the PGF for the exclusively-low case
\[
G_{xlo}(z_1) = G_0(z_1, 0, \ldots, 0),
\] (65)

which is given by (49) and (55) as
\[
G_{xlo}(z_1) = P_0 \cdot \frac{1 - z_1}{1 - z_1 \zeta_{+}(z_1, 0, \ldots, 0)}.
\] (66)

On setting \( r_{lo} = r_K, r_{hi} = \sigma_{K-1} = r - r_{lo} \), we can write
\[
\beta(z_1, 0) = r_{lo} z_1, \quad \zeta_{\pm}(z_1, 0) = \frac{1}{2} \left[ 1 + r - r_{lo} z_1 \pm \sqrt{(1 + r - r_{lo} z_1)^2 - 4 r_{hi}} \right].
\] (67)

Thus, we see that \( \zeta_{\pm}(z_1, 0) \) is identical with its counterpart for the two-level (\( K = 2 \)) problem, so that we can write \( \zeta_{\pm}(z_1, 0) = \zeta_{\pm}(z_1) \). Using the identity
\[
\frac{1 - z_1}{1 - z_1 \zeta_{+}(z_1)} = \frac{1 - \zeta_{-}(z_1)}{\zeta_{+}(z_1) - r},
\] (68)

we arrive at the result
\[
G_{xlo}(z_1) = P_0 \cdot \frac{1 - \zeta_{-}(z_1)}{\zeta_{+}(z_1) - r} = P_0 \left[ 1 + r_{lo} \cdot \frac{z_1}{\zeta_{+}(z_1) - r} \right].
\] (69)

This is equivalent to
\[
G_{xlo}(z_1) = 1 - r + r_{lo} z_1 G_{lo}(z_1).
\] (70)

It follows that
\[
P_{xlo}(n) = (1 - r) \delta_{n0} + (1 - \delta_{n0}) r_{lo} P_{lo}(n - 1),
\] (71)

where \( P_{lo}(n) \) is the wait-conditional marginal PMF for the lowest-priority level. One should note that this relationship is independent of the number of priority levels \( K \).
3.3. Exclusively-High Probabilities  To consider the case where there are assumed to be only the highest-priority clients in the queue, we must set $Z_1 = Z_2 = \cdots = Z_{K-1} = 0$, to obtain the PMF for the exclusively-high case

$$P_{xhi}(\ell) = G_{\ell}(0,0,\ldots,0) = P_0 \zeta_-(0),$$

as given by (49) and (55). On setting $r_{hi} = \sigma_1 = r_1$, $r_{lo} = r - r_{hi}$, we can write

$$\beta(0) = 0, \quad \zeta_+(0) = \frac{1}{2} \left[ 1 + r \pm \sqrt{(1+r)^2 - 4r_{hi}} \right].$$

Therefore

$$P_{xhi}(\ell) = (1-r) \left[ \frac{1+r-\sqrt{(1+r)^2 - 4r_{hi}}}{2} \right]^\ell,$$

for $\ell = 0, 1, \ldots$. One should note that this result is independent of the number of priority levels $K$.

4. FFT Mixture Method  Let us begin by considering the numerical evaluation of the marginal PMFs. Without loss of generality, we may restrict our attention to the low-priority PMF for the two-level problem. We have

$$P_{lo}(n) = \frac{1}{n!} \left. \frac{d^n g(z)}{dz^n} \right|_{z=0}, \quad g(z) = \frac{1-r}{\zeta_+(z) - r},$$

where $g(z) = G_{lo}(z)$ as given by (60). Invoking Cauchy’s integral theorem, we can write this as

$$P_{lo}(n) = \int_C \frac{dz}{2\pi i} \frac{g(z)}{z^{n+1}},$$
for \( n = 0, 1, \ldots, \) where the integration contour \( C \) encircles the origin anti-clockwise and remains within the radius of convergence of the Taylor expansion of \( g(z) \), which is given by \( \eta_c = 1/r > 1 \). If we choose \( C \) to be the circle of radius \( \eta < \eta_c \) centred about the origin, and write \( z = \eta e^{-i\theta} \), for \( 0 \leq \theta < 2\pi \), then

\[
P_{\eta}(n) = \frac{1}{\eta^n} \int_0^{2\pi} \frac{d\theta}{2\pi} e^{in\theta} g(\eta e^{-i\theta}).
\]  

(77)

Approximation of the integral by an \( N \)-interval trapezoidal rule on the grid \( \theta_k = 2\pi k/N, k = 0, 1, \ldots, N \), yields the finite sum

\[
P_{\eta}(n) \simeq \frac{1}{\eta^n N} \sum_{k=0}^{N-1} e^{2\pi i nk/N} g(\eta^{-2\pi i k/N}),
\]  

(78)

which is proportional to the inverse discrete Fourier transform of the sequence \( h(k) = g(\eta^{-2\pi i k/N}) \), \( k = 0, 1, \ldots, N - 1 \). When \( N \) is chosen to be a power of 2, it can be implemented as an inverse fast Fourier transform (IFFT), in which case we write \( P_{\eta}(n) \simeq \eta^{-n} \text{IFFT}[h](n) \). Extraction of queue-length probabilities from a PGF by means of a discrete Fourier transform has been previously considered by Daigle [4]. We shall adopt an alternative general method due to Fornberg [7] that is better suited to the multi-variate problem.

The RHS of (77) is independent of the value of the contour radius \( \eta \). This is no longer true in (78) as a consequence of the approximation. The dependence on the integration radius in using an FFT to numerically compute derivatives of analytic functions has been studied by Bornemann [2]. He found that there exists an optimal radius that minimizes the error, and that this optimal value is close to the radius of convergence.

The dependence of the approximated solution on the contour radius can be used to one’s advantage, as originally pointed out by Fornberg [7]. Consider the numerical differentiation of the analytic function \( g(z) \). According to the FFT method, we must evaluate

\[
H(n; \eta) \equiv \eta^{-n} \text{IFFT}_k[h(k; \eta)](n),
\]

\[
h(k, \eta) \equiv g(\eta^{-2\pi i k/N}),
\]  

(79)

for some contour radius. We may equally evaluate the weighted mixture

\[
H_M(n) \equiv \sum_{m=1}^{M} f_m H(n; \eta_m),
\]  

(80)

for some collection of contour radii \( \eta_m \) and mixture coefficients \( f_m \) that sum to unity \( \sum_{m=1}^{M} f_m = 1 \). As discussed in [7], the first \( M \) aliasing terms can be cancelled using (in principle) any \( M \) distinct contour radii \( \eta_m < \eta_c \) provided ones makes an appropriate choice of the coefficients \( f_m \). To achieve this, the mixture coefficients \( f_m \) must solve the matrix equation

\[
\begin{bmatrix}
1 & 1 & \cdots & 1 \\
\eta_1^N & \eta_2^N & \cdots & \eta_M^N \\
\eta_1^{2N} & \eta_2^{2N} & \cdots & \eta_M^{2N} \\
\vdots & \vdots & \ddots & \vdots \\
\eta_1^{(M-1)N} & \eta_2^{(M-1)N} & \cdots & \eta_M^{(M-1)N}
\end{bmatrix}
\begin{bmatrix}
f_1 \\
f_2 \\
f_3 \\
\vdots \\
f_M
\end{bmatrix}
= \begin{bmatrix}
1 \\
0 \\
0 \\
\vdots \\
0
\end{bmatrix}.
\]  

(81)

The solution for the mixture coefficients \( f_m \) is given by

\[
\frac{1}{f_m} = \prod_{\ell=1}^{M} \left( 1 - \frac{\eta_{m\ell}}{\eta_{m\ell}} \right),
\]  

(82)
for $m = 1, 2, \ldots, M$. Therefore, the residual errors after the FFT mixture are given by

$$\varepsilon_M(n) = c_{MN}(n) \sum_{m=1}^{M} f_m \eta_m^{MN} = (-1)^{M-1} c_{MN}(n) \prod_{m=1}^{M} \eta_m^n,$$

(83)

for some constants $c_{MN}(n)$. The result for the summation over $m$ follows directly from (82). One may note that, when applied to (75), $c_{MN}(n) = P_0(n + MN)$.

4.1. Marginal Distribution Without loss of generality, we consider the marginal distribution for the low-priority level in the two-level ($K = 2$) problem. For any fixed $r$, worst case performance of the FFT method is observed to occur as $r_{hi} \to 0^+$. Thus, we shall set $r_{lo} = r$, $r_{hi} = 0$. In this case, the wait-conditional PGF is trivially given by

$$g(z) = \frac{1 - r}{1 - rz} = \sum_{n=0}^{\infty} p_n z^n,$$

(84)

with $p_n = (1 - r)^n r^n$, for $k = 0, 1, \ldots$. To approximate the $p_n$ via an $N$-point FFT, we apply the trapezoidal rule on the grid $z_k = \eta e^{-2\pi ik/N}$ for $k = 0, 1, \ldots, N - 1$, to obtain

$$p_n = \frac{1}{N} \sum_{k=0}^{N-1} g(z_k) e^{2\pi i k n / N}.$$  

(85)

This leads us to consider the ratio

$$\frac{p_n}{(1 - r)^n} = \frac{1}{N (r\eta)^n} \sum_{k=0}^{N-1} e^{2\pi i k n / N}$$

(86)

that will approach unity as $N \to \infty$. With $\xi \equiv r\eta < 1$, we introduce the ratio function

$$R_N(\xi, n) = \frac{1}{N \xi^n} \sum_{k=0}^{N-1} e^{2\pi i k n / N}$$

(87)

Given (86), we aim to achieve $R_N(\xi, n) \approx 1$ for all $n = 0, 1, \ldots, N - 1$. We may also observe that

$$R_N(\xi, n) = \frac{1}{\xi^n} I_N(\xi, n) = \frac{1}{\xi^n} \text{IFFT}_k \left[ (1 - \xi e^{-2\pi i k / N})^{-1} \right] (n).$$

(88)

The ratio function can be evaluated by explicitly summing the series to give

$$R_N(\xi, n) = \frac{1}{1 - \xi^n},$$

(89)

independent of $n$. Thus, $\delta R_N(\xi) \equiv R_N(\xi, n) - 1 \approx \xi^n$ for $\xi^n \ll 1$. If the desired discretization relative error is set to $10^{-\alpha}$, then $\xi$ is determined by $\xi^n = 10^{-\alpha}$.

We shall be directly computing the quantity $I_N(\xi, n)$ in (88) via a dedicated FFT routine. Since $R_N(\xi, n) \approx 1$, we have that $I_N(\xi, n) \approx \xi^n$. This cannot be too small for relevant values of $n$ due to machine arithmetic limits. In double-precision arithmetic, we must ensure that $I_N(\xi, n) \gg 10^{-16}$ for the largest $n = 0, 1, \ldots, N_{\text{max}} < N$ of interest, $N$ being the FFT size, which must be greater than $N_{\text{max}}$ in order to prevent aliasing. If we take the absolute FFT error (i.e. round-off error) to be $\varepsilon_{\text{fft}} \approx 10^{-15}$, then the relative FFT error is $10^{-15} / \xi^{N_{\text{max}}} \sim 10^{-15+n_{\text{max}}/N}$. Best numerical
performance is obtained by equating the two competing and opposing sources of error, i.e. the relative discretization error with the relative FFT error. This yields the relationship

$$\alpha = \frac{15}{1 + N_{\text{max}}/N}$$

(90)

for the overall relative error exponent. It implies that, for twelve decimal places of accuracy ($\alpha = 12$), we require that $N \approx 4N_{\text{max}}$.

Let us now consider the FFT mixture scheme. In order to specify the $M$ contour radii, we introduce a spread parameter $s$, whose value we typically take to be $s \simeq 0.05$, and spread factors $\zeta_m = (1 - s \cdot (m - 1)/(M - 1))$, for $m = 1, 2, \ldots, M$. Then, we set $\xi_m = \zeta_m \xi$, which gives all of the radii in terms of the largest one $\xi_1 = \xi$, as yet undetermined. The smallest is given by $\xi_M = \zeta_M \xi = (1 - s)\xi$.

It is useful to observe that, in the present context, (83) reads

$$\epsilon_M(n)/[(1 - r)r^n] = (-1)^{M-1} \prod_{m=1}^{M} \xi_m^{n},$$

(91)

the RHS being independent of $n$.

We revisit the foregoing error analysis adapted to the mixture method, and begin by estimating the FFT error. If the mixture ratio function as given by the trapezoidal rule were the actual object of evaluation by means of an FFT, then the exact result would be

$$R_{\text{mix}}^{\text{fft}}(\xi, n) = \sum_{m=1}^{M} \frac{f_m(\xi)}{1 - \xi_m^{n}}$$

(92)

while the approximate result rendered by the FFT is

$$R_{\text{fft}}^{\text{fft}}(\xi, n) = \sum_{m=1}^{M} \frac{f_m(\xi)}{1 - \xi_m^{n}} \left(1 \pm \epsilon_{\text{fft}} \frac{\xi_m}{\xi_m^{n}}\right)$$

(93)

where $\epsilon_{\text{fft}}$ denotes the absolute error in the FFT computation due to machine arithmetic precision. Thus, the error in the ratio function due to FFT precision is

$$\Delta R_{\text{fft}}^{\text{fft}}(\xi, n) \equiv \left|R_{\text{fft}}^{\text{fft}}(\xi, n) - R_{\text{mix}}^{\text{fft}}(\xi, n)\right|$$

(94)

$$\leq \epsilon_{\text{fft}} \sum_{m=1}^{M} \frac{|f_m(\xi)|}{1 - \xi_m^{n}} \frac{1}{\xi_m} \sim \epsilon_{\text{fft}} \sum_{m=1}^{M} \frac{|f_m(\xi)|}{1 - \xi_m^{n}} \frac{1}{\xi_m^{n}}$$

where the final step represents a worst-case bound. Given that $\xi_m^{n} \ll 1$, we have

$$\sum_{m=1}^{M} \frac{|f_m(\xi)|}{1 - \xi_m^{n}} \frac{1}{\xi_m} \simeq \sum_{m=1}^{M} \frac{|f_m(\xi)|}{\xi_m^{n}} \simeq \frac{|f_M(\xi)|}{\xi_M^{n}}.$$

(95)

So, we see that the FFT error is dominated by the smallest contour radius $\xi_M$. It follows that it is estimated by

$$\Delta R_{\text{fft}}^{\text{fft}}(\xi, n) \sim \epsilon_{\text{fft}} |f_M(\xi)|/\xi_M^{n} \sim \epsilon_{\text{fft}}/\xi_M^{n},$$

(96)
since $|f_M(\xi)| \simeq 1$ for a sufficiently large spread $s$, and we take $\varepsilon_{\text{fft}} = 10^{-15}$.

To estimate the discretization error, we consider the mixture ratio function

$$R_{\text{mix}}^{\text{Nfft}}(\xi, n) = \sum_{m=1}^{M} \frac{f_m(\xi)}{1 - \xi_m^{\text{Nfft}}} = \sum_{m=1}^{M} f_m(\xi) + \sum_{m=1}^{M} \xi_m^{\text{Nfft}} f_m(\xi) + \cdots = 1 + (-1)^{M-1} (\xi_1 \xi_2 \cdots \xi_M)^{\text{Nfft}} + \cdots .$$

(97)

Thus, the discretization error due to application of the trapezoidal rule is

$$\Delta R_{\text{tpz}}^{\text{Nfft}}(\xi, n) \equiv R_{\text{mix}}^{\text{Nfft}}(\xi, n) - 1 \simeq (g\xi)^{MN_{\text{fft}}},$$

(98)

where we have introduced the geometric mean

$$g \equiv \left( \prod_{m=1}^{M} \xi_m \right)^{1/M} .$$

(99)

We proceed to equate the trapezoidal-rule error of (98) with the FFT error of (96) at the largest desired queue length $n = N_{\text{max}}$, and set them to a common error level $10^{-\alpha}$, i.e.

$$(g\xi)^{MN_{\text{fft}}} = 10^{-15}/\xi_{M_{\text{max}}}^{\text{Nmax}} = 10^{-\alpha} .$$

(100)

Then, after taking $\alpha = 12$, $M = 4$, eliminating $\xi$ from the equations yields the relationship $N_{\text{fft}} = N_{\text{max}}/(1 - N_{\text{max}}\chi)$ with $\chi \equiv \log_{10}((1 - s)/g)$. For $s = 0.05$, $M = 4$, we obtain $\chi = 0.0037$. So, it suffices to choose $N_{\text{fft}} \simeq N_{\text{max}}$. Consequently, the largest contour radius is determined as

$$\xi = g^{-1} \cdot 10^{-12/(MN_{\text{fft}})},$$

(101)

with $N_{\text{fft}}$ chosen to be the smallest power of two exceeding $N_{\text{max}}$.

As a sanity check, we apply the FFT-mixture method as described here to the computation of the low-priority marginal PMF for the two-level problem of the case $r_{hi} = 0$, and plot in Figure 2 various errors for the ratio function as a function of queue length. The maximum queue length considered was that sufficient to attain a tail probability $P_{\text{min}} = 10^{-12}$. The target error level was set to $10^{-12}$ (i.e. $\alpha = 12$). The exact (pre-discretization) ratio function is everywhere unity in the present case. The overall error (blue curve) represents the difference in the computed ratio function from unity. The FFT error (red curve) represents the difference in the computed ratio function from the exact value of the discretized ratio function. The discretization error (orange curve) represents the difference in the exact discretized ratio function from unity. The results are consistent with expectations.

4.2. Joint Distribution The expression (49) for $G_0(z)$ contains spurious algebraic singularities that cancel out between denominator and numerator, as can be seen by observing that $\zeta(1) = 1$. This has implications for numerical evaluation, and limits the choice of integration contour radii in the application of Cauchy’s theorem to $|z_{\epsilon}| < 1$ whereas, in principle, one could extend this to at least $|z_{\epsilon}| < 1/r$. Fortunately, the algebraic singularities can be eliminated by re-structuring the integrand into an equivalent form according to the following argument: Let us write (49) as

$$G_0(z) = P_0 \cdot \prod_{\kappa=1}^{K-1} \frac{P^+_\kappa(z)}{Q^+_\kappa(z)} ,$$

(102)
where
\[ \mathcal{P}_k^\pm(z) = 1 - z_n \zeta_{k-1}^\pm(z), \quad \mathcal{Q}_k^\pm(z) = 1 - z_n \zeta_k^\pm(z), \] (103)

with \( \zeta_k^\pm(z) \equiv \zeta_{\pm}(z_1, \ldots, z_k), \) for \( k = 0, 1, \ldots, K - 1. \) In particular, \( \zeta_0^+() = 1, \quad \zeta_0^-() = r. \) Then, as derived in Appendix C, we have the result
\[ \frac{\mathcal{P}_k^+(z)}{\mathcal{Q}_k^+(z)} = \frac{\mathcal{Q}_k^-(z)}{\mathcal{P}_k^-(z)}, \] (104)

which yields the equivalent form for the PGF
\[ G_\ell(z_1, \ldots, z_{K-1}) = P_0 \prod_{k=1}^{K-1} \frac{1 - z_n \zeta_-(z_1, \ldots, z_k)}{1 - z_n \zeta_-(z_1, \ldots, z_{K-1})} \zeta^-_k(z_1, \ldots, z_{K-1}). \] (105)

This result can also be established directly via an argument based on the fact that the marginal queue-length distribution for the aggregation of the highest priority levels is a geometric distribution, for all \( p = 1, 2, \ldots, K. \) The details are presented in Appendix D. It is also interesting to note that the change of integration variables \( z \rightarrow w \) such that \( w_k(z_1, \ldots, z_k) = \zeta_-(z_1, \ldots, z_k), \) for \( k = 1, 2, \ldots, K - 1, \) leads to the completely meromorphic form
\[ G_\ell(z_1, \ldots, z_{K-1}) \cdot \prod_{k=1}^{K-1} dz_k = P_0 W_\ell K - 1 \cdot \prod_{k=1}^{K-1} \frac{1 - z_n w_k}{1 - z_n w_{K-1}} \cdot \prod_{k=1}^{K-1} \sigma K - k w_k^2 dw_k, \] (106)

where
\[ z_k = \frac{w_k}{w_{K-1}^2 + \sigma_{K+1-k} - \sigma_{k} w_k - w_{K-1} w_k}, \]
\[ = \frac{1}{r_{K+1-k}} \left( \frac{w_{K-1}^2 w_k}{w_{K-1}^2 - w_k} \right), \] (107)

and we adopt the convention that \( z_0 \equiv r = \sum_{k=1}^{K} r_k. \) The second product is the Jacobian of the variable transformation. We have not yet found a practical use for this representation.

Application of Cauchy’s integral theorem in each of \( K - 1 \) dimensions yields the expression for the joint PMF
\[ P(\ell, n_{K-1}, \ldots, n_1) = \prod_{k=1}^{K-1} \int_{C_k} \frac{dz_k}{2\pi i z_k^{n_k+1}} \cdot G_\ell(z_1, \ldots, z_{K-1}), \] (108)

where the anti-clockwise closed contours around the origin \( C_k \) can be taken to be circles of radius \( \eta_k \ll 1/r. \) For simplicity, we adopt a common radius \( \eta_k = \eta \) in each dimension.

We introduce the multi-dimensional inverse FFT
\[ \mathcal{F}_\ell(n; \eta) = \eta^{-N(K-1)} \cdot \text{IFT}k [G_\ell(\eta e^{-2\pi i k/N}))(n), \] (109)

where, for simplicity, the FFT size \( N \) is taken to be the same for every dimension, and \( n = [n_1, n_2, \ldots, n_{K-1}] \). According to the mixture method,
\[ P(\ell, n_{K-1}, \ldots, n_1) \simeq \sum_{m=1}^{M} f_m \mathcal{F}_\ell(n; \eta_m), \] (110)

with the contour radii \( \eta_m \) and the coefficients \( f_m \) chosen as for the marginal distributions discussed in the preceding section.
Using the mixture, we execute $M$ runs of an FFT of size $N$ in each of the $K-1$ dimensions. Without the mixture technique, comparable accuracy would require a single FFT execution of size $MN$. Therefore the no-mixture/mixture timing ratio for $K$ priority levels is

$$\frac{(MN)^{K-1} \log_2 (MN)^{K-1}}{M \cdot N^{K-1} \log_2 (N^{K-1})} = M^{K-2} \left[ 1 + \frac{\log_2 M}{\log_2 N} \right] \approx M^{K-2}.$$  

We see that there is no benefit in implementing the mixture method for the two-level ($K = 2$) problem, but it is increasingly advantageous as the number of priority levels rises. Testing of numerous cases confirms this performance differential.

5. Numerical Tests

5.1. Aggregation Test  As discussed for the two-level problem in [14], the aggregated queue-length distribution describes the total number of entities in the queue, regardless of priority level. This is equivalent to the queue-length distribution of the basic M/M/$c$ queueing model with traffic intensity $r = \sum_{k=1}^{K} r_k$, which is known to be a simple geometric distribution. Hence, the exact aggregate PMF is given by

$$P_{\text{agg}}^{(\text{ex})}(k) = (1 - r) r^k,$$

for $k = 0, 1, 2, \ldots$. Consequently, for $k > 0$,

$$[\Delta \ln P_{\text{agg}}](k) \equiv \ln(P_{\text{agg}}(k)) - \ln(P_{\text{agg}}(k-1)) = \ln r,$$

independent of $k$.

One diagnostic test of the computational methodology is to check how well the aggregate PMF constructed from the computed joint PMF reproduces the exact result. This test is more convenient than similarly testing against marginals as only finite summations are required. The aggregate PMF is obtained from the joint PMF as

$$P_{\text{agg}}(k) = \sum_{n \in \mathbb{N}_0} P(n),$$

for $k = 0, 1, 2, \ldots$. We then consider the measure of performance (MOP)

$$\Xi_{\text{agg}} \equiv -\max_{k \geq 1} \left\{ \log_{10} (|\Delta \ln P_{\text{agg}}(k) - \Delta \ln P_{\text{agg}}^{(\text{ex})}|) \right\},$$

where the maximum is taken over all values $0 \leq k \leq n_{\text{lim}}$ such that $P_{\text{agg}}^{(\text{ex})}(k) > P_{\text{agg}} > 0$, for some threshold level $P_{\text{min}}$ since one cannot expect the numerical methods to maintain performance down to arbitrarily small tail probabilities. Since we are working in double-precision arithmetic, all MOPs of this kind are capped at a maximum allowed value of 16. The interpretation of $\Xi_{\text{agg}}$ (and similarly for all of the subsequent MOPs) is that it indicates the number of decimal places of numerical agreement in the worst case.

5.2. Nearest-Neighbour Test  A direct consequence of the identity (21) is that the joint PMF at any given interior point $n$ is a positively weighted sum of the joint PMF values at $K + 1$ of its $2K$ nearest neighbours. We denote by $P_{\text{nn}}(n)$ the PMF evaluated at the point $n$ by means of (21). Then, we consider the MOP

$$\Xi_{\text{nn}} \equiv -\max_{n \in \mathbb{N}^K} \left\{ \log_{10} (|\ln(P(n)) - \ln(P_{\text{nn}}(n))|) \right\},$$

where the maximum is taken over all values $n \in [1, n_{\text{lim}}]^K$ such that $P(n) > P_{\text{min}} > 0$. 
5.3. Xhi-Test  We denote by $P_{\text{xhi}}^{(\text{ex})}(\ell)$ the theoretical exact expression of the exclusively-high PMF as given by (74), which we compare with the result obtained directly from the joint PMF as

\[ P_{\text{xhi}}(\ell) = P(\ell,0,\ldots,0). \]  

If we define, for all $\ell > 0$,

\[ \Delta \ln P_{\text{xhi}}^{(\text{ex})}(\ell) \equiv \ln(P_{\text{xhi}}(\ell)) - \ln(P_{\text{xhi}}(\ell - 1)), \]  

then $\Delta \ln P_{\text{xhi}}^{(\text{ex})} = \ln[1 + r - \sqrt{(1 + r)^2 - 4r_{\text{hi}}}/2]$, independent of $\ell$. The MOP for the xhi-test is then taken to be

\[ \Xi_{\text{xhi}} \equiv -\max_{\ell \geq 1} \left\{ \log_{10} \left( |\Delta \ln P_{\text{xhi}}^{(\text{ex})}(\ell)| - \Delta \ln P_{\text{xhi}}^{(\text{ex})} \right) \right\}, \]  

where the maximum is taken over all values $0 \leq \ell \leq n_{\text{lim}}$ such that $P_{\text{xhi}}^{(\text{ex})}(\ell) > P_{\text{min}} > 0$. When applied to the FFT method, the xhi-test simply checks how accurately the numerical FFT mixture algorithm reproduces the Cauchy residue theorem for a simple pole at the origin.

5.4. Xlo-Test  The exclusively-low PMF, as obtained directly from the computed joint PMF, is given by

\[ P_{\text{xlo}}(n) = P(0,\ldots,0,n). \]  

It is theoretically related to the lowest-priority marginal PMF according to (71). This allows one to compare the computed $P_{\text{xlo}}(n)$ against a known accurate evaluation of the marginal $P_{\text{lo}}(n)$. Such a marginal PMF is provided by the quadratic recurrence method developed in [14] for the two-level problem, recalling that the two-level problem is sufficient for the calculation of all marginals of the multi-level problem. The relevant MOP is taken to be

\[ \Xi_{\text{xlo}} \equiv -\max_{n > 0} \left\{ \log_{10} \left( |\ln(P_{\text{xlo}}(n)) - \ln(r_{\text{lo}}P_{\text{lo}}(n - 1))| \right) \right\}, \]  

where the maximum is taken over all values $0 < n \leq n_{\text{lim}}$ such that $P_{\text{lo}}(n) > P_{\text{min}} > 0$.

5.5. FPI Test  In the FPI test, we compute the joint PMF using both the FPI and FFT methods, and check their agreement. The MOP is simply taken to be

\[ \Xi_{\text{fpi}} \equiv -\max_{n \in [1,n_{\text{lim}}]^K} \left\{ \log_{10} \left( |\ln(P_{\text{fpi}}(n)) - \ln(P_{\text{fpi}}(n))| \right) \right\}, \]  

where the maximum is taken over all values $n \in [1,n_{\text{lim}}]^K$ such that $P_{\text{fpi}}(n) > P_{\text{min}} > 0$.

5.6. Results  Figures 3–6 present the results of the numerical tests described above. MOP values prior to worst-case minimization, relevant to the FFT mixture computations, are displayed on the vertical axes against the relevant queue lengths. Each test pertains to $K = 4$ priority levels repeated with 30 randomized level traffic intensities, all corresponding to total traffic intensity $r = 0.9$, as shown by the blue lines and dots. The red curve picks out an individual run. The cyan curve delineates the overall worst-case behaviour. The PMF array is generated up to a maximum queue length of $N_{\text{max}} = 100$ in each priority dimension. We see that $\Xi_{\alpha} > 9.5$ for $\alpha = \text{agg}, \text{xhi}, \text{xlo}, \text{nn}$.

Figure 7 compares the joint PMF arrays as computed via the FFT and FPI methods, for $K = 3$ priority levels with 30 randomized level traffic intensities summing to $r = 0.75$. All elements whose probabilities exceeded a threshold tail level of $P_{\text{min}} > 10^{-10}$ were included. The worst case for each lowest-priority queue length is plotted. We see that $\Xi_{\text{fpi}} \gtrsim 2$. Decreasing performance the further one progresses into the tail is a reflection of the truncation error in the FPI method.
Figure 3.

Note. Aggregation test for \( K = 4 \) priority levels with 30 randomized level traffic intensities corresponding to total traffic intensity \( r = 0.9 \). The number of decimal places of agreement with the exact result is plotted on the vertical axis as a function of aggregated queue length. The maximum aggregated queue length of 100 includes data points with PMF above \( P_{\text{min}} = 2.4 \times 10^{-6} \).

Figure 4.

Note. XLo-test for \( K = 4 \) priority levels with 30 randomized level traffic intensities corresponding to total traffic intensity \( r = 0.9 \). The number of decimal places of agreement is plotted on the vertical axis as a function of lowest-priority queue length. The maximum queue length of 60 includes data points with PMF above \( P_{\text{min}} = 1.0 \times 10^{-6} \).
**Figure 5.**

*Note.* Xhi-test for $K = 4$ priority levels with 30 randomized level traffic intensities corresponding to total traffic intensity $r = 0.9$. The number of decimal places of agreement is plotted on the vertical axis as a function of highest-priority queue length. The maximum queue length of 100 includes data points with PMF above $P_{\text{min}} = 1.0 \times 10^{-20}$.

**Figure 6.**

*Note.* Nearest-neighbour test for $K = 4$ priority levels with 30 randomized level traffic intensities corresponding to total traffic intensity $r = 0.9$. The number of decimal places of agreement is plotted on the vertical axis as a function of lowest-priority queue length. All joint-PMF data points above $P_{\text{min}} = 1.0 \times 10^{-10}$ that occur within a maximum queue length of 100 in each priority dimension were considered.
Finally, in Figure 8, we plot the results of the combined set of tests for \( K = 7 \) priority levels with a randomly generated set of level traffic intensities that sum to a total traffic intensity of \( r = 0.9 \). The PMF array is generated up to a maximum queue length of \( N_{\text{max}} = 15 \) in each priority dimension. For the nearest-neighbour test, array elements with probabilities above \( P_{\text{min}} = 10^{-6} \) are considered, and the results are presented as the worst case for each lowest-priority queue length. The other curves are self-explanatory. It can be observed that the algorithm holds up well in this large example, where the FFT size is \( 0.25 \) GB, and the size of the PMF array is \( 2 \) GB.

The results obtained confirm (i) numerical stability, (ii) internal consistency, (iii) consistency between methods, and (iv) agreement with theoretical exact values.

6. Conclusions Simple methods for accurate computation of the joint queue-length distribution for a non-preemptive multi-level priority queue have been developed. An explicit multi-variate PGF for the joint queue-length distribution has been derived in closed form for the first time. With the aid of an FFT mixture method, it is used to numerically compute the joint queue-length PGF and the queue-length marginals. A direct method based on a fixed-point iteration of the stationary balance equations has also been developed. Comparable explicit and exact results are not available elsewhere in the existing literature.

Appendix A: Multi-Variate PGF In order to solve (40) for

\[
G_0(u_k, \ldots, u_{K-1}) = G_0(0, \ldots, 0, u_k, \ldots, u_{K-1}) ,
\]

we consider a general problem for a family of functions \( G_0(.) \) that are differentiated from one another according to the number of arguments that they take. The system to be solved has the
FFT Multi-Level Combo Test
(r = 0.9, K = 7, P_{\text{min}} = 1.0e-06)

Figure 8.

Note. All FFT tests for K = 7 priority levels with a random set of level traffic intensities corresponding to total traffic intensity r = 0.9. The number of decimal places of agreement is plotted on the vertical axis as a function of the relevant queue length. Joint-PMF data points were computed to a maximum queue length of 15 in each priority dimension. For the nearest-neighbour test, all all points above P_{\text{min}} = 1.0 \times 10^{-6} were considered.

following structure: For each n = 1, 2, \ldots, K - 1,

\[ G_0(x_1, \ldots, x_n) = \frac{1}{1/x_1 - \lambda_+(x_1, \ldots, x_n)} \sum_{k=2}^{n+1} \left( \frac{1}{x_{k-1}} - \frac{1}{x_k} \right) G_0(x_k, \ldots, x_n), \]  

(124)

with \( x_{n+1} \equiv 1 \), and \( G_0() = P_0 \). The functions \( \lambda_\pm(x_1, \ldots, x_n) \) are solutions of the quadratic equation

\[ \zeta^2 - [1 + r - \alpha(x_1, \ldots, x_n)]\zeta + \sigma_{K-n} = 0, \]  

(125)

with

\[ \alpha(x_1, \ldots, x_n) = \sum_{k=1}^{n} x_k r_{K-n+k}. \]  

(126)

We observe that \( G_0'(u_k, \ldots, u_{K-1}) = G_0(x_1, \ldots, x_n) \), being the application with \( n = K - k \) arguments, in which case

\[ \alpha(u_k, \ldots, u_{K-1}) = \sum_{\ell=1}^{K-k} u_{\ell+k-1} r_{\ell+k} = \sum_{\ell=k}^{K-1} u_\ell r_{\ell+1}, \]  

(127)

which involves only the \( K - k \) lowest priorities with level traffic intensities \( r_{k+1}, \ldots, r_K \), while the constant term of the quadratic \( \sigma_k = \sum_{\ell=1}^{k} r_\ell \) aggregates the disjoint highest \( k \) priorities with level traffic intensities \( r_1, \ldots, r_k \). This is consistent with the interpretation of \( G_0'(u_k, \ldots, u_{K-1}) \) as solving the \((K+1-k)\)-level sub-problem with the \( k \) highest priorities aggregated into a single level while keeping the total traffic \( r \) intensity constant.

Next, we introduce

\[ J(x_n, \ldots, x_1) \equiv G_0(x_1, \ldots, x_n), \]
\[ \zeta_\pm(x_n, \ldots, x_1) \equiv \lambda_\pm(x_1, \ldots, x_n), \]  

(128)
and set $z_k \equiv x_{n+1-k}$ for $k = 1, 2, \ldots, n$, so that $z_1 = x_n$, $z_n = x_1$, and the convention $x_{n+1} = 1$ translates to $z_0 = 1$. Then (124) becomes

$$J(z_1, \ldots, z_n) = \frac{1}{1/z_n - \zeta^+(z_1, \ldots, z_n)} \sum_{k=2}^{n+1} \left( \frac{1}{z_{n+2-k}} - \frac{1}{z_{n+1-k}} \right) J(z_1, \ldots, z_{n+1-k})$$

$$= \frac{1}{1/z_n - \zeta^+(z_1, \ldots, z_n)} \sum_{\ell=0}^{n-1} \left( \frac{1}{z_{\ell+1}} - \frac{1}{z_\ell} \right) J(z_1, \ldots, z_\ell),$$

(129)

under the change of summation variable $\ell = n + 1 - k$, and where $J() = G_0(0, \ldots, 0) = P_0$. Let us now write $J_\ell(z) \equiv J(z_1, \ldots, z_\ell)$ and $J_0 \equiv J()$, so that the subscript $\ell$ indicates the number of elements (i.e., dimension) of the vector argument $z$. Then, we have

$$J_n = \frac{1}{1/z_n - \zeta^+(z_1, \ldots, z_n)} \sum_{\ell=0}^{n-1} \left( \frac{1}{z_{\ell+1}} - \frac{1}{z_\ell} \right) J_\ell,$$

(130)

for $n = 1, 2, \ldots, K - 1$. Finally, let $J'_n \equiv (1/z_n - \zeta^+(z_1, \ldots, z_n)) J_n$ so that we obtain the recurrence

$$J'_n = \sum_{\ell=0}^{n-1} W_\ell J'_\ell,$$

(131)

with

$$W_\ell = \left( \frac{1}{z_{\ell+1}} - \frac{1}{z_\ell} \right) \frac{1}{1/z_n - \zeta^+(z_1, \ldots, z_\ell)},$$

(132)

for $\ell \geq 1$, and $W_0 = 1/z_1 - 1$.

For $n \geq 1$, we can write

$$J'_{n+1} = W_n J'_n + \sum_{\ell=0}^{n-1} W_\ell J'_\ell = (1 + W_n) J'_n,$$

(133)

which is solved by

$$J'_n = \prod_{\ell=1}^{n-1} (1 + W_\ell) J'_1, \quad J'_1 = W_0 J'_0,$$

(134)

where we have set $J'_0 = P_0$. After some algebraic manipulation, this leads to the explicit representation

$$J_n(z) = P_0 \prod_{\ell=1}^{n} \frac{1 - z_\ell \zeta^+(z_1, \ldots, z_{\ell-1})}{1 - z_\ell \zeta^+(z_1, \ldots, z_\ell)},$$

(135)

where the identity $\zeta^+(\cdot) = 1$ has been invoked. The quantity that is ultimately of interest for the $K$-level problem is $J_{K-1}(z) = G_0(z) = G_0(z_1, \ldots, z_{K-1})$, where the arguments of $G_0(.)$ are now reversed.

To complete the specification of the solution, we recall that

$$\alpha(u_m, \ldots, u_{K-1}) = \sum_{k=m}^{K-1} u_k r_{k+1},$$

(136)

and set $z_m = u_{K-m}$ for $m = 1, 2, \ldots, K - 1$, in order to define the function $\beta(z)$ via

$$\beta(u_{K-1}, \ldots, u_m) \equiv \alpha(u_m, \ldots, u_{K-1}),$$

(137)
so that
\[
\beta(z_1, \ldots, z_{K-m}) = \sum_{k=m}^{K-1} z_{K-k} r_{k+1}.
\] (138)

Equivalently, since \( m \) is arbitrary,
\[
\beta(z_1, \ldots, z_n) = \sum_{\ell=1}^{n} z_\ell r_{K+1-\ell}.
\] (139)

The functions \( \zeta_\pm(z_1, \ldots, z_n) \) solve the quadratic equation
\[
\zeta^2 - [1 + r - \beta(z_1, \ldots, z_n)]\zeta + \sigma_{K-n} = 0.
\] (140)

The two branches are given by
\[
\zeta_\pm(z_1, \ldots, z_n) = \frac{1}{2} \left[ 1 + r - \beta(z_1, \ldots, z_n) \pm \sqrt{(1 + r - \beta(z_1, \ldots, z_n))^2 - 4\sigma_{K-n}} \right],
\] (141)
for \( n = 1, 2, \ldots, K-1 \). Noting that \( \sigma_K = \sum_{k=1}^{K} r_k = r \), and with the convention \( \beta() \equiv 0 \), we recover the relationships \( \zeta_+(1) = 1, \zeta_-(1) = r \) from extension to the case \( n = 0 \).

**Appendix B: Extraction of Marginals** In this appendix, we show that all marginals for the general \( K \)-level priority problem can be inferred from knowledge of the structure of the PGF for the two-level problem. It is convenient to introduce the notation
\[
\xi_k^\pm(z) \equiv \zeta_\pm(z_1, \ldots, z_k),
\] (142)
where the integer subscript indicates the number of elements (i.e. the dimension) of the vector argument. Then, from (61) and (105), it follows that
\[
\sum_{\ell=0}^{\infty} G_\ell(z) = \sum_{k=0}^{K-1} \frac{1 - z_k \zeta_k^-} {1 - z_{k+1} \zeta_k^-}(z),
\] (143)
where we have set \( z_K \equiv 1 \), and noted that
\[
P_0 = 1 - r = 1 - \zeta_-(1) = 1 - \zeta_0^-(z).
\] (144)

To extract the \( p \)-th marginal, we must consider the vector argument \( z = z^{(p)} \) where we set \( z_k = 1 \) for all \( k \neq p \). Thus \( z^{(p)} = 1 + (z_p - 1)e_p \). We may observe that the \( k \)-th term of the product on the RHS of (143) is equal to unity, expect when \( k = p - 1 \) or \( k = p \). This leads to a representation of the PGF for the \( p \)-th marginal as
\[
G^{(p)}_{\text{merg}}(z_p) = \sum_{\ell=0}^{\infty} G_\ell(z^{(p)}) = \frac{1 - z_p \zeta_p^-} {1 - z_p \zeta_p^-}(z^{(p)}) \frac{1 - z_{p-1} \zeta_{p-1}^-} {1 - z_{p-1} \zeta_{p-1}^-}(z^{(p)})
\] (145)
for \( p = 1, 2, \ldots, K-1 \). Since the identity \( \beta_k(1) = r - \sigma_{K-k} \) implies that \( \zeta_k^+(1) = 1, \zeta_k^-(1) = \sigma_{K-k} \), we have
\[
\zeta_{p-1}^+(z^{(p)}) = \zeta_{p-1}^-(1) = \sigma_{K+1-p}.
\] (146)

Next, we observe that
\[
\beta_p(z^{(p)}) = \beta_p(1, \ldots, z_p, \ldots, 1) = \sum_{k=1}^{p-1} r_{K+1-k} + r_{K+1-p} z_p.
\] (147)
Setting

\[ r_{lo} = r_{K+1-p}, \quad r_{hi} = \sum_{k=1}^{K-p} r_k = \sigma_{K-p}, \quad r_{sum} \equiv r_{lo} + r_{hi} = \sigma_{K+1-p}, \]

we obtain

\[ r - \beta_p(z^p) = r_{sum} - r_{lo} z_p. \]

Therefore,

\[ \zeta_p^\pm(z^p) = \frac{1}{2} \left[ 1 + r - \beta_p(z^p) \pm \sqrt{(1 + r - \beta_p(z^p))^2 - 4\sigma_{K-p}} \right] \]

\[ = \frac{1}{2} \left[ 1 + r_{sum} - r_{lo} z_p \pm \sqrt{(1 + r_{sum} - r_{lo} z_p)^2 - 4r_{hi}} \right] \]

\[ = \zeta_p^\pm(z_p), \]

where the final identification with \( \zeta_p^\pm(z) \) for the two-level problem is associated with the mappings \( r_1 \leftarrow r_{hi}, r_2 \leftarrow r_{lo}, r \leftarrow r_{sum} \). Consequently,

\[ G^p_{mrg}(z_p) = \frac{1 - r_{sum}}{1 - r_{sum} z_p} \cdot \frac{1 - z_p \zeta_-^-(z_p)}{1 - \zeta_-^-(z_p)}, \]

for \( p = 1, 2, \ldots, K - 1 \). Hence, on comparing with (105) for the case \( K = 2 \), we see that the desired result for the marginal PGFs is established. Alternatively, by multiplying numerator and denominator by \( 1 - \zeta_+(z_p) \) and using the identities given in (58), one recovers the first form in (60).

**Appendix C: PGF Ratio** Considering the quantities \( P^\pm_\kappa(z), \Omega^\pm_\kappa(z) \) as defined in (103), we shall show that

\[ P^+_\kappa(z)P^-_\kappa(z) = \Omega^+_{\kappa}(z)\Omega^-_{\kappa}(z), \]

which implies that one can eliminate the removable singularities in (102) by making the substitution

\[ P^\pm_\kappa(z)/\Omega^\pm_\kappa(z) \rightarrow \Omega^\mp_\kappa(z)/P^\pm_\kappa(z). \]

In order to first evaluate \( P^\pm_\kappaP^-_\kappa \), we let \( a \equiv 4\sigma_{K+1-\kappa} \) and \( b \equiv 1 + r - \beta_{\kappa-1} \). Then it is easy to see that

\[ P^+_\kappaP^-_\kappa = 1 - b z_\kappa + \frac{a}{4} z_\kappa^2 \]

\[ = 1 - (1 + r - \beta_{\kappa-1})z_\kappa + \sigma_{K+1-\kappa} z_\kappa^2. \]

Similarly, if we now let \( a' \equiv 4\sigma_{K-\kappa} \) and \( b' \equiv 1 + r - \beta_{\kappa}, \)

\[ \Omega^+_{\kappa}\Omega^-_{\kappa} = 1 - b' z_\kappa + \frac{a'}{4} z_\kappa^2 \]

\[ = 1 - (1 + r - \beta_{\kappa})z_\kappa + \sigma_{K-\kappa} z_\kappa^2 \]

\[ = 1 - (1 + r - \beta_{\kappa-1} - r_{K+1-\kappa} z_\kappa)z_\kappa + \sigma_{K-\kappa} z_\kappa^2 \]

\[ = 1 - (1 + r - \beta_{\kappa-1})z_\kappa + (\sigma_{K-\kappa} + r_{K+1-\kappa}) z_\kappa^2 \]

\[ = 1 - (1 + r - \beta_{\kappa-1})z_\kappa + \sigma_{K+1-\kappa} z_\kappa^2. \]

The desired result follows from the equality of the right-hand sides of the foregoing pair of equations.
Appendix D: Alternative Derivation of the Joint PGF Recalling from (55) that \( G_r(z_1, \ldots, z_{K-1}) = G_0(z_1, \ldots, z_{K-1})z_\ell^\ell \), we obtain for the full \( K \)-dimensional joint PGF,

\[
G(z_1, \ldots, z_K) \equiv \sum_{\ell=0}^\infty z_\ell^\ell G_r(z_1, \ldots, z_{K-1}) = \frac{G(z_1, \ldots, z_{K-1}, 0)}{1 - z_K \zeta_{K-1}(z)}, \tag{156}
\]

where \( \zeta_{K-1}(z) \equiv \zeta_{(z_1, \ldots, z_{K-1})} \) such that the subscript denotes the number of arguments. We introduce the notation \( G(z_1, \ldots, z_p | z) \equiv G(z_1, \ldots, z_p, z, \ldots, z) \) for any \( z \in \mathbb{C} \) and \( p = 0, 1, \ldots, K \). For \( p = 0 \), we have \( G(\cdot | z) = G(\cdot, \ldots, z) \), while for \( p = K \), we have \( G(z_1, \ldots, z_K | z) = G(z_1, \ldots, z_K) \), independent of \( z \).

Since the marginal distribution of the aggregation of the top \( p \) priority levels is also geometric, as argued in the discussion below (41), we have that

\[
G(z_1, \ldots, z_{K-p} | z) = \frac{G(z_1, \ldots, z_{K-p}, 0)}{1 - z_{K-p} \zeta_{K-p}(z)}, \tag{157}
\]

for arbitrary \( z \) and \( p = 1, 2, \ldots, K \), where the rates \( \zeta_{K-p}(z) \) are to be determined. When \( p = 1 \), we recover (156). Setting \( z = z_{K-p+1} \) in (157) yields

\[
G(z_1, \ldots, z_{K-p-1} | z) = \frac{G(z_1, \ldots, z_{K-p+1}, 0)}{1 - z_{K-p+1} \zeta_{K-p+1}(z)}, \tag{158}
\]

for \( p = 1, \ldots, K \). Next we write (157) as

\[
G(z_1, \ldots, z_{K-p+1} | z) = \frac{G(z_1, \ldots, z_{K-p+1}, 0)}{1 - z_{K-p+1} \zeta_{K-p+1}(z)}, \tag{159}
\]

and choose \( z_{K-p+1} = z \) so that

\[
G(z_1, \ldots, z_{K-p} | z) = \frac{G(z_1, \ldots, z_{K-p+1}, 0)}{1 - z_{K-p+1} \zeta_{K-p+1}(z_1, \ldots, z_{K-p}, z)}. \tag{160}
\]

Now we set \( z = z_{K-p+1} \) to obtain

\[
G(z_1, \ldots, z_{K-p} | z_{K-p+1}) = \frac{G(z_1, \ldots, z_{K-p+1}, 0)}{1 - z_{K-p+1} \zeta_{K-p+1}(z)}. \tag{161}
\]

We observe that (158) and (161) express equations for the same quantity. Thus, on equating the RHS of each equation, we arrive at the recursion

\[
G(z_1, \ldots, z_k | 0) = \frac{1 - z_k \zeta_k(z)}{1 - z_k \zeta_{k-1}(z)} G(z_1, \ldots, z_{k-1} | 0), \tag{162}
\]

for \( k = 1, 2, \ldots, K-1 \), which leads to the result

\[
G_0(z_1, \ldots, z_{K-1}) = G(z_1, \ldots, z_{K-1} | 0) = P_0 \prod_{k=1}^{K-1} \frac{1 - z_k \zeta_k(z)}{1 - z_k \zeta_{k-1}(z)}, \tag{163}
\]

where we have used the fact that \( P_0 = G(\cdot | 0) = G(0, \ldots, 0) \).

In order to determine the rates \( \zeta_k(z) \), we multiply the stationary balance equation (19) by \( z_1^{n_1} \ldots z_{K-k}^{n_{K-k+1}} z_k^{n_k + \cdots + n_1} \), and sum over all indices \( n_1, \ldots, n_K \) to obtain

\[
\left[ \beta_k(z) - r + \frac{\sigma_{K-k}}{z \zeta_k(z)} \right] G(z_1, \ldots, z_k | z) = G(z_1, \ldots, z_k | 0), \tag{164}
\]
where
\[
\beta_k(z) = \sum_{\kappa=1}^{k} z_{\kappa} r_{K-k+1}, \quad \sigma_k = \sum_{\kappa=1}^{k} r_{\kappa},
\]
and we have used the relation
\[
\frac{d}{dz} G(z_1, \ldots, z_k \mid z) \bigg|_{z=0} = \zeta_k(z) G(z_1, \ldots, z_k \mid 0),
\]
that follows from (157). We also use (157), expressed as
\[
G(z_1, \ldots, z_k \mid 0) = (1 - z \zeta_k(z)) G(z_1, \ldots, z_k \mid z),
\]
to obtain
\[
\left[ z^2 \zeta_k^2 - (1 + r - \beta_k) z \zeta_k + \sigma_{K-k} \right] G(z_1, \ldots, z_k \mid z) = 0,
\]
provided \( z \neq 0 \). Setting \( z = 1 \) yields
\[
\zeta_k^2 - (1 + r - \beta_k) \zeta_k + \sigma_{K-k} = 0,
\]
which is solved by (52). Therefore, the result (105) is established.

**Acknowledgments** The authors gratefully acknowledge useful discussions with Dr. Stephen Bocquet.

**References**


