QUEUEING SYSTEMS WITH RESTRICTED WORKLOAD:
AN EXPLICIT AND RECURSIVE EXPRESSION

Nico M. van Dijk

Research memorandum 1987-50
Dec. '87

VRIJE UNIVERSITEIT
FACULTEIT DER ECONOMISCHE WETENSCHAPPEN
EN ECONOMETRIE
AMSTERDAM
Queueing systems with restricted workload:
An explicit and recursive expression

Nico M. van Dijk
Free University, Amsterdam, The Netherlands

Abstract

Queueing systems are studied with Poisson arrivals, general service requirements and a capacity restriction on the total residual service requirement (workload). A closed form expression is obtained for the steady state workload distribution. A recursive representation of this expression as well as the loss probability is provided.

Queueing * workload * recursive representation.
1. Introduction

Ever since Erlang's loss formula from the twenties, queueing systems with lost arrivals have been widely studied and applied to fields such as telecommunication, computer performance evaluation and manufacturing. Generally, however, in these studies the acceptance or rejection of arriving jobs is based upon only the configuration of the jobs already present such as their total number, class specification or particular positions within the queue. However, in various practical situations the acceptance or rejection of an arriving job may naturally depend on the total residual workload or service requirement from the jobs present and the arriving job. Particularly the case where the total residual workload is not allowed to exceed some threshold seems most realistic.

For instance, an upper limit on the workload can be imposed so as to guarantee that the system (machine, server) will not work too long without pauses. Or a finite workload constraint may correspond to a restricted temporarily storage for work to be worked off (e.g. loads to be unloaded or data to be read). In contrast with related finite flow or dam models, in service systems it can be practical that a job requirement is either fully accepted or fully rejected and not just the excess amount. For instance, a truck has to be unloaded completely or a program is to be read from begin to end.

Closed form expressions for systems with restricted workload have not been reported. Laplace Stieltjes transforms are given in Cohen [2], [4] when only the excess amount is rejected while for the case of full rejection, as in this paper, exact results are given only for the exponential and deterministic case in Cohen [4] and Gavish and Schweitzer [7]. Bounds for the general service case are developed in Geish and Kobayashi [8] and Ross [14]. Asymptotic results when the capacity restriction tends to infinity can be found in Cohen [2], [4], De Kok and Tijms [5], Iglehart [10], Van Ommeren [16], Van Ommeren and De Kok [17] and Wyner [18]. Approximations are proposed in De Kok and Tijms [5] and Tijms [15].
In particular it turned out that systems with workload depending blocking are highly sensitive to distributional forms of service requirements. Moreover, even in the exponential case the steady state distribution does not exhibit a standard type scaled geometric form (such as for GI/M/1 queues).

This paper is concerned with single server non-exponential queueing systems with a restricted workload capacity. The main result is a closed form expression of state dependent geometric form for the steady state workload distribution. Moreover, a recursive representation of this distribution as well as the loss probability is derived. The proof of the expression is of interest in itself (cf. remark 2.2).

The organization is as follows. Section 2 first derives the steady state distribution of a detailed workload description under the assumption of a last-come first-served discipline. Next from this distribution the steady state distribution of the total workload is directly concluded also for the first-come first-served case. Section 3 shows that this distribution as well as performance measures such as the loss probability can be recursively determined.

2. Model and result

Consider a single-server system with Poisson arrivals and general service requirements (as in an M/G/1-queue) but with a restricted capacity \( K \) on the total workload (the total amount of residual service requirements from the customers present). An arriving customer is rejected if its service request yields an excess; that is if its required service length exceeds \( K-w \), where \( w \) is the workload (total residual service requirement) from the customers already present. The interarrival and service times are assumed to be independent with arrival parameter \( \lambda \) and continuous density function \( q(.) \) of the service requirement with mean \( r \). Assume \( \lambda r < 1 \).
LCFS-case. First let us assume a last-come first-served preemptive discipline and denote by \((w_1, \ldots, w_n)\) the state in which \(n\) customers are present of which the \(j\)-th customer, in order of arrival, has a residual service requirement \(w_j\) \((j=1, \ldots, n)\). Only the \(n\)-th customer is thus being served. For \(n>0\), let \(\pi(w_1, \ldots, w_n)\) be the corresponding stationary density while for \(n=0\) we write \(p(0)\) for the stationary probability of the empty state. Furthermore, for all \(k \geq 0\) and \(w_1, \ldots, w_k\) and \(w\) let

\[
V(w|w_1, \ldots, w_k) = \frac{1}{r} \int_w^{K(w_t + \ldots + w_k)} q(s) ds.
\]

Lemma 2.1 Under the assumption of a last-come first-served preemptive discipline, we have for all \(n>0\)

\[
\pi(w_1, \ldots, w_n) = p(0)[\lambda r]^n \prod_{k=1}^n V(w_k|w_1, \ldots, w_{k-1}).
\]

Proof. Note that the system under consideration with the given state description is a Markov process which changes only by jumps (the arrivals) according to a bounded jump intensity and by deterministic drifts (up to the service completion of the last entered job). Therefore, one easily verifies that the corresponding semigroup of transition probabilities satisfies the weak continuity condition (2.18) in Dynkin [6], p. 54. This condition guarantees that the semigroup is stochastically continuous. As a result, by theorem 1.9 and theorem 2.3 of Dynkin [6], the stationary distribution \(\pi(\cdot)\) is uniquely determined by the stationary forward or backward infinitesimal Kolmogorov equation provided it is contained in the domain \(D_A\) of the weak infinitesimal operator, that is the set of functions \(f\) for which

\[
Af(x) = \lim_{h \to 0} h^{-1} \int P_h(x;dy)[f(y)-f(x)]
\]

is well-defined for any \(x\) while the right-hand side remains uniformly bounded in all \(x\) for \(h\) sufficiently small, where \(P_h(x;\cdot)\) represents the transition probability measure over time \(h\) from out of state \(x\). As in examples 2.15 and 2.18 of Dynkin [6], for the present system one easily verifies that \(D \subset D_A\), where \(D\) is the set of all real-valued functions on the state space \(S=\{(w_1, \ldots, w_n); w_i>0, i=1, \ldots, n, n \geq 0\}\) which are bounded.
and have a bounded and continuous derivative with respect to $w_n$. Now, first observe that $\pi(.)$ as defined by (2.2) is contained in $\mathcal{D}$ by assumption of $\lambda r < 1$ and continuity of $q(.)$. (Note that expression (2.1) can be seen as a truncated excess probability of a renewal process and is thus bounded by 1.) As a result, it suffices to verify the forward infinitesimal Kolmogorov equations for stationarity. To this end, similarly to Gavish and Schweitzer [7] and Keilson [11], by conditioning upon a time point $t-\Delta t$ while considering a time point $t$ and using the continuity with respect to the last entered component, the stationary forward Kolmogorov equations require that for any $\Delta t$ and with $o(\Delta t)/\Delta t \to 0$ as $\Delta t \to 0$, for $n>0$:

\begin{equation}
(2.3) \quad \pi(w_1, \ldots, w_n) = \\
\pi(w_1, \ldots, w_{n-1})\lambda\Delta t q(w_n) +
\pi(w_1, \ldots, w_n+\Delta t)[1-\lambda\Delta t] +
\Delta t \int_0^\Delta t \pi(w_1, \ldots, w_{n-1}, r) dr [1-\lambda\Delta t] +
\frac{K}{\Delta t} - \int_0^\Delta t (w_1 + \ldots + w_n) q(s) ds \frac{o(\Delta t)}{\Delta t}
\end{equation}

with the boundary condition for $n=0$:

\begin{equation}
(2.4) \quad p(0) = \\
p(0)[1-\lambda\Delta t] +
\frac{K}{\Delta t} p(0)\Delta t [1 - \int_0^{\Delta t} q(s) ds] + \int_0^{\Delta t} \pi(r)dr [1-\lambda\Delta t] + o(\Delta t).
\end{equation}

By dividing by $\Delta t$, letting $\Delta t \to 0$ and using the continuity for the last entered component again, the stationary infinitesimal forward Kolmogorov equation becomes:
with the boundary condition

\[(2.6) \quad \pi(0^+) = \lambda p(0) \left[ \int_0^K q(s) \, ds \right],\]

where \(0^+\) indicates that for that component the right-hand limit at 0 is to be read. By substituting expression (2.2) we directly conclude that the first term within braces in the right-hand side of (2.5) is itself equal to 0. Again from expression (2.2) we find

\[(2.7) \quad \pi(w_1, \ldots, w_n, 0^+) = \lambda \pi(w_1, \ldots, w_n) \left[ \int_0^{K - (w_1 + \ldots + w_n)} q(s) \, ds \right],\]

so that by substitution also the second term within braces in the right-hand side is itself equal to 0. We have thus proven (2.5). As expression (2.2) also guarantees (2.7), the proof is completed.

**FCFS-case.** Now let us assume a first-come first-served discipline and denote by state \(w\) the total residual service requirement (workload) from the customers present. Let \(F(w)\) be the corresponding stationary distribution function. As the total workload process, however, does not depend on which customers are actually being served, we have

\[(2.8) \quad F(w) = \sum_{n=1}^{\infty} \int_{w_1 + \ldots + w_n \leq w} \pi(w_1, \ldots, w_n) \, dw_1 \ldots dw_n + p(0)\]

where \(\pi(w_1, \ldots, w_n)\) is given by (2.2). The steady state workload distribution as well as other performance measures of interest such as the loss probability of a customer are hereby in principle determined.
Remark 2.2. Note that the proof of lemma 2.1 or rather (2.5), is actually established by showing partial balance of the first and second term of the right-hand side of (2.5). In contrast, in Gavisch and Schweitzer [7] a similar equation has been derived but has been solved in only the special exponential case. These balances in fact can be seen as continuous analogues (and in fact came up as such) of well-known partial or local balance principles in discrete state queueing descriptions for proving product forms. This observation is of interest in itself. More remarkably, while partial balance principles are generally related to service form independent results (cf. Barbour [1], Cohen [3], Hordijk and Van Dijk [9], Kelly [12], Schassberger [13]), here we have obtained an explicit service form dépendant result.

3. Recursive representation

This section is concerned with a recursive representation of the workload distribution (2.8) as well as the loss probability of an arriving customer. To this end, we define for all \( k \geq 1 \) and \( w_1 + \ldots + w_k \leq t < K \):

\[
U_k(w_1, \ldots, w_k-1) = \int_{w_k}^{t-(w_1 + \ldots + w_k-1)} q(s)ds,
\]

where for \( k = 1 \) we can substitute \( w_1 = \ldots = w_{k-1} = 0 \) and write \( U_1(w_1) \), and for all \( n > 1 \) and \( t < K \):

\[
\Phi_k(n) = \int_{w_1 + \ldots + w_k \leq t} \prod_{k=1}^{n} U_k(w_1, \ldots, w_k-1)dw_1 \ldots dw_n,
\]

while \( \Phi_0(0) = 1 \). Then \( F(w) \) from (2.8), and \( p(n) \) the steady state probability of \( n \) customers as according to (2.2) (that is under the last-come first-served assumption) can be written as

\[
F(w) = p(0) \sum_{n=0}^{\infty} \lambda^n \Phi_n(n) \quad (w \leq K),
\]

\[
p(n) = p(0) \lambda^n \Phi_0(n) \quad (n \geq 0).
\]
From expression (3.1), however, we conclude that

\[(3.5) \quad U_t(w_k|w_1, w_2, \ldots, w_{k-1}) = U_{t-w_1}(w_k|w_2, \ldots, w_{k-1}).\]

By conditioning (3.2) to the first component \(w_1\) and substituting (3.5), we derive

\[(3.6) \quad \phi_n(n) = \int U_t(w_1)[ \int \Pi U_t(w_k|w_2, \ldots, w_{k-1})dw_2 \ldots dw_n]dw_1 = \int U_t(w_1)\phi_{t-w_1}(n-1)dw_1.\]

As a result, by (3.3), (3.4) and the recursion (3.6), the normalizing constant and the distribution \(F(w)\) can be computed by

\[p(0) = \left[\sum_{n=0}^{\infty} \lambda^n \phi_n(n)\right]^{-1},\]

\[F(w) = p(0) \sum_{n=0}^{\infty} \lambda^n \phi_n(n) \quad \text{for} \quad w \leq K,\]

where

\[\phi_n(k) = \int [\int q(s)ds] \phi_{t-u}(k-1)du \quad (t \leq K, k \geq 1),\]

\[\phi_t(0) - 1 \quad (t \leq K),\]

As a particular performance measure of interest, let us also consider the loss probability \(B\) of an arriving customer. Again noting that this is not dependent upon the actual last-come or first-come first-served discipline, in either case we can use expression (2.2) and calculate \(B\) by

\[(3.8) \quad B = p(0) \int q(u)[1 + \sum_{n=1}^{\infty} \pi(w_1, \ldots, w_n)dw_1 \ldots dw_n]du.\]

To this end, define for all \(n \geq 1\) and \(s \leq t \leq K:\)
Then, similarly to (3.6) we obtain from (3.1):

\[ \psi_{s,t}(n) = \int_0^t U_t(u) \psi_{s,t-u}(n-1) du \]

with \( \psi_{s,t}(0)=1 \), and by (3.8) and (3.9):

\[ B = p(0) \int q(u)[1 + \sum_{n=1}^{K} \lambda^n \psi_{a,k}(n)] du \]

Hence the loss probability \( B \) is recursively determined by (3.7) and (3.10). Similar recursions can be given for other performance measures.

References


