A SINGLE SERVER QUEUE IN A HARD-REAL-TIME ENVIRONMENT

K S TRIVEDI ET AL. MAY 85 CS-1985-12 AFOSR-TR-85-0747

UNCLASSIFIED AFOSR-84-8132
CS-1985-12

A Single Server Queue in a Hard-Real-Time Environment

Francois Baccelli
INRIA
France

Kishor S. Trivedi
Duke University
We consider a single server first in first out queue in which each arriving task has to be completed within a certain period of time (it's deadline). More precisely, each arriving task has its own deadline - a non-negative real number - and as soon as the response time of one task exceeds its deadline, the whole system is considered to have failed. (In that sense the deadline is hard). The main practical motivation for analysing such queues comes from the need to evaluate mathematically the reliability of computer systems working with real time constraints (space or aircraft systems for instance). We shall therefore be mainly concerned with the analytical characterization of the transient behavior of such a queue in order
to determine the probability of meeting all hard deadlines during a finite period of time (the "mission
time"). The probabilistic methods for analysing such systems are suggested by earlier work on impatience
in telecommunication systems [1], [2].
A Single Server Queue in a Hard-Real-Time Environment

Francois Baccelli**
INRIA
Rocquencourt
France

and

Kishor S. Trivedi
Department of Computer Science
Duke University
Durham, NC 27706 USA

ABSTRACT

We consider a single server first in first out queue in which each arriving task has to be completed within a certain period of time (its deadline). More precisely, each arriving task has its own deadline - a non-negative real number - and as soon as the response time of one task exceeds its deadline, the whole system is considered to have failed. (In that sense the deadline is hard). The main practical motivation for analyzing such queues comes from the need to evaluate mathematically the reliability of computer systems working with real time constraints (space or aircraft systems for instance). We shall therefore be mainly concerned with the analytical characterization of the transient behavior of such a queue in order to determine the probability of meeting all hard deadlines during a finite period of time (the "mission time"). The probabilistic methods for analyzing such systems are suggested by earlier work on impatience in telecommunication systems [1], [2].

Additional keywords: functional equations,

*This work was supported in part by the National Science Foundation under grant number USNSF MCS 83-0200, by the Army Research Office under grant DAAGH-84-0046, and by the Air Force Office of Scientific Research under grant number AFOSR-84-0132.

**This work was begun when Dr. Baccelli was a visiting scientist in the Department of Computer Science at Duke University.
1. Introduction

In this paper, we are concerned with the analysis of a system operating in a hard-real-time environment. In such a system, each job (or task) must be completed within a specified period of time after a request for its execution arrives. If any job fails to complete within its deadline, the entire system is considered to have failed. For a description of such systems and their analysis in a deterministic environment see [9).

We consider here a single server queueing system in which job arrival stream is Poisson and the job service requirement is generally distributed. Each job has a deadline associated with it so that if the response time of a job exceeds its deadline, we will assume that the system has failed. The system can also fail due to a breakdown experienced by the server. The completion time (or the actual service time) of a job once scheduled will be allowed to depend, in general, on the job sequence number. In this way, graceful degradation of the server can also be taken into account. The aim of this paper is to derive an expression for the average number of jobs completed before system failure.

M/G/1 queueing system with server breakdown and repair has been analyzed by Gaver [4] while and M/M/n queueing system with server breakdown and repair was analyzed by Mitrany and Avi-Itzhak [11]. Baccelli and Trivedi [3] analyzed and M/G/2 standby redundant system with breakdown and repair. These studies carried out steady state analysis. Approximate transient analysis of such a queueing system has been carried out by Meyer [10] assuming no repair while Kulkarni, Nicola, Trivedi and Smith [8] have extended this analysis to allow for (possibly imperfect) repairs. The latter effort [7,8], also allows for deadline constraints to be imposed but only in the case that no resource contention is permitted. Analysis in the present paper is exact and allows for deadline constraints, queueing and server breakdown.

After defining the basic model in the next section, we then derive the functional equations for the queueing system in section 3. These equations are specialized to the case of an M/G/1 queue with exponentially distributed deadlines and independent geometric server failure process in section 4.
2. The Basic Model

Although various computational assumptions will be introduced later on, we first state the queueing problem in its generality.

Let \( \{\sigma_n, n \geq 1\} \) and \( \{\delta_n, n \geq 1\} \) represent the respective completion time and deadline length of the \( n \)-th job. \( \sigma_n \) represents the actual time job \( n \) spends in the server once scheduled. This may take into account a possible degradation of the server's capacity. Therefore the statistics of \( \sigma_n \) will not be stationary in general (for instance the time to serve a unit-customer may increase with the age of the system) and \( \sigma_n \) is allowed to be infinite (for instance when the server has failed before the completion of the job). Let \( r_n, n \geq 1 \) represent the time between the arrivals of customers \( n-1 \) and \( n \). \( r_n \) will be assumed a.s. finite.

We shall represent the successive response times by a sequence of defective random variables \( R_n, n \geq 0 \):

- \( R_0 \) is a given a.s. finite random initial condition,
- \( R_{n+1} \) represents the response time of the \( n+1 \)-st customer provided its completion time is finite and provided the \( n \)-th customer experienced a finite response time and met its deadline. Otherwise, \( R_{n+1} \) is infinite.

Mathematically, the evolution of \( R_n \) is described by equations (1) and (2):

\[
\begin{align*}
\{R_{n+1} < \infty\} &= \left\{ \{R_n < \infty\} \cap \{\sigma_{n+1} < \infty\} \cap \{R_n < \delta_n\} \right\} \\
&= \bigcap_{k=1}^{n} \left( \{\sigma_k < \infty\} \cap \{R_k < \delta_k\} \right) \cap \{\sigma_{n+1} < \infty\}, \quad n \geq 1, \\
\{R_1 < \infty\} &= \{\sigma_1 < \infty\}.
\end{align*}
\]

and in the event \( \{R_{n+1} < \infty\} \):

\[
R_{n+1} = [R_n - r_{n+1}]^+ + \sigma_{n+1}, \quad n \geq 0
\]

where \( [a]^+ \) denotes \( \max(a,0) \). Hence, on \( \{R_{n+1} < \infty\} \), \( R_{n+1} \) is fully determined from (2) and the knowledge of an initial value of \( R_0 \). In the complementary event, \( \{R_{n+1} = \infty\} \), \( R_{n+1} \) is determined by
We introduce now computational assumptions to be used in the sequel of the paper:

- The sequences \( \{ \tau_n, n \geq 1 \}, \{ \sigma_n, n \geq 1 \} \) and \( \{ \delta_n, n \geq 1 \} \) will be assumed independent.

- The arrival stream will be assumed Poisson of intensity \( \lambda \).

- The hard deadlines will be assumed to be i.i.d. exponentially distributed random variables with parameter \( \delta \).

- The \( \sigma_n \)'s will be assumed to be independent but not necessarily identically distributed random variables. We shall denote as \( \gamma_n(s) \) the quantity \( E[e^{-s\sigma_n}1_{s < \infty}], \Re(s) \geq 0 \).

3. The Functional Equation

Let \( s \in \mathbb{C}, \Re(s) \geq 0 \). From equations (1) and (2) we obtain:

\[
e^{-sR_s+1}1_{(R_s+1) < \infty} = e^{-s(R_s-\tau_s+\delta)} e^{-\sigma_s+1}1_{(R_s < \infty)}1_{(R_s < \delta)}1_{(\sigma_s+1 < \infty)}.
\]

Let

\[
\phi_n(s) = E[e^{-sR_s}1_{(R_s < \infty)}], \quad n \geq 0.
\]

Taking the expectation on both sides of (3) and using independence and distributional assumptions, we get

\[
\phi_{n+1}(s) = \gamma_{n+1}(s)E[e^{-s(R_s-\tau_s)+\delta}1_{(R_s < \delta)}1_{(R_s < \infty)}]
\]

\[
= \gamma_{n+1}(s)E[e^{-s(R_s-\tau_s)+\delta} e^{-R_s}1_{(R_s < \infty)}]
\]

\[
= \gamma_{n+1}(s)\{E[e^{-(\delta+s)R_s}1_{(R_s < \infty)}] + E[e^{-sR_s} \int_0^R e^{-s(\delta-t)} e^{-\lambda t} dt 1_{(R_s < \infty)}]\}.
\]

So that \( \phi_n(s) \) satisfies the recurrence equation

\[
\phi_{n+1}(s) = \frac{\gamma_{n+1}(s)}{\lambda-s}[\lambda \phi_n(s + \delta) - s \phi_n(s + \delta)], \quad n \geq 0.
\]

For \( s \in \mathbb{C}, |s| < 1 \), let \( G(s,x) \) and \( F(s,x) \) be defined by the power series
\[ G(s,z) = \sum_{n \geq 1} \gamma_n(s)z^n, \]
\[ F(s,z) = \sum_{n \geq 0} \phi_n(s)z^n. \]

Using the recurrence equation (5), we get
\[ F(s,z) - \phi_0(s) = \frac{1}{\lambda - \delta} \sum_{n \geq 0} z^{n+1} \gamma_{n+1}(s) (\lambda \phi_n(s + \delta) - s \phi_n(\lambda + \delta)). \] (6)

Then using Hadamard's theorem [5] we conclude that \( F(s,z) \) satisfies the functional integral equation (see the Appendix):
\[ F(s,z) = \phi_0(s) + \frac{1}{\lambda - \delta} \left[ \frac{\lambda}{2i \pi} \int_{C} F(s+\delta,u)G(s+\delta,\frac{z}{u})du \right. \]
\[ - \frac{s}{2i \pi} \int_{C} F(\lambda+\delta,u)G(\lambda+\delta,\frac{z}{u})du \right], \]
where the contour integral is taken on any circle with center at 0 and of radius \( R \) where \( |z| < R < 1 \).

Before considering special cases, let us note how to get features of practical interest from the knowledge of the solution of (7). Let \( N^* \) be the stopping time defined by:
\[ N^* = \inf\{n \geq 1 | R_n \geq \delta_n \}. \] (8)

The very definition of \( \phi_n(s) \) and (1) entail:
\[ \phi_n(0) = P[R_n < \infty] = P[\bigcap_{k=1}^{n-1} (R_k < \delta_k) \cap (\sigma_n < \infty)] \]
\[ = \alpha_n \cdot P[N^* \geq n], \quad n \geq 1, \] (9)

where \( \alpha_n = P(\sigma_n < \infty) \).

So that the mathematical expectation of \( N^* \), the number of customers served before (and including) the first hard failure is given by (in case \( \alpha_n = \alpha \) for all \( n \geq 1 \):
\[ E[N^*] = \frac{1}{\alpha} \times \lim_{z \to 1} (F(0,z) - \phi_0(0)). \] \hspace{1cm} (10)

More generally, denoting by \( \Gamma \) a circle of center 0 and radius \( R < 1 \), one gets from (8) and (9)

\[ P[N^* \geq n] = \frac{1}{\alpha 2i \pi} \int \frac{F(0,z)}{z^{n+1}} \, dz. \] \hspace{1cm} (11)

4. A Special Case: Systems with geometrically distributed catastrophic failures.

Consider a system in which the server is subject to catastrophic non-repairable failures (i.e., when the failure occurs, the server stops functioning forever). Assuming that the number \( N \) of customers that this system can process before the first failure (and disregarding hard deadline constraints) is geometrically distributed with parameter \( \alpha \) so that

\[ P[N = n] = (1-\alpha)\alpha^n. \] \hspace{1cm} (12)

Such a situation will occur in case the server can experience a catastrophic failure with rate \( \lambda_f \) so that the probability of a successful job completion \( \alpha = \int_0^\infty e^{-\lambda_f t} \, dF_S(t) = G(\lambda_f), \)

where we assumed that the service requirements of customers are i.i.d. with common distribution functions \( F_S \) and Laplace Stieltjes transform \( G(s) \). In this case, we get the following representation for \( \phi_n(s) \):

\[ \phi_1(s) = \alpha G(s) \] \hspace{1cm} (13)

and from (7) we get:

\[ F(s,x) = \phi_0(s) + az - \frac{G(s)}{\lambda - s} \left[ \lambda F(s + \delta, x) - s F(\lambda + \delta, x) \right]. \] \hspace{1cm} (14)

Let

\[
\begin{align*}
\alpha(s,z) &= \frac{\lambda az G(s)}{\lambda - s} \\
\beta(s,z) &= \phi_0(s) - \frac{s az G(s)}{\lambda - s} F(\lambda + \delta, x)
\end{align*}
\] \hspace{1cm} (15)

Equation (14) can be rewritten as
\[ F(s, z) = a(s, z)F(s + \delta, z) + b(s, z), \]  

so that for any finite integer \( n \) and for \( \text{Re}(s) > \lambda \)

\[ F(s, z) = b(s, z) + \sum_{k=1}^{n} \left( \prod_{i=0}^{k-1} a(s + i \delta, z) \right) b(s + k \delta, z) \]

\[ + \left( \prod_{i=0}^{n} a(s + i \delta) \right) F(s + (n + 1) \delta, z). \]

Let us show that \( R_n(s, z) \) the last term in the right hand side of (17), converges to zero as \( n \) goes to infinity. For large \( n \),

\[ |a(s + n \delta, z)| \sim \frac{\lambda \alpha}{\delta} \frac{1}{n} |G(s + n \delta)| \]

\[ \leq K(z) \frac{1}{n}, \]

so that the convergence of this remainder is faster than \( 1/n ! \):

\[ |R_n(s, z)| \sim f_n(s, z) \leq K(z) \frac{1}{n !}. \]  

Hence in the limit, (17) yields the expansion:

\[ F(s, z) = \sum_{k=0}^{\infty} \left( \prod_{i=0}^{k-1} a(s + i \delta, z) \right) b(s + k \delta, z), \]

where

\[ \prod_{i=0}^{k} a(s + i \delta, z) = 1. \]

Using our definitions (equations (15)) we get:

\[ F(s, z) = \sum_{k \geq 0} (\lambda \alpha z)^k \left( \prod_{i=0}^{k-1} \frac{G(s + i \delta)}{\lambda - s - i \delta} \right) \phi_0(s + k \delta) \]

\[ - F(\lambda + \delta, z) \alpha z \left( \sum_{k \geq 0} (\lambda \alpha z)^k \left( \prod_{i=0}^{k} \frac{G(s + i \delta)}{\lambda - s - i \delta} \right) (s + k \delta) \right). \]

We remain with the problem of determining the unknown function \( F(\lambda + \delta, z) \). For this, let us multiply both sides of (15) by \((\lambda - s)\) and take the limit as \( s \) approaches \( \lambda \). Since \( F(s, z) \) has to be analytic in \( s \) at \( \lambda \), the L.H.S. has to vanish implying then the relation:
\[ F(\lambda + \delta, x) = \frac{\sum_{k \geq 0} \frac{(xz)^k}{k!} \prod_{i=1}^b G(\lambda + i \delta) \cdot \phi_0(\lambda + (k+1)\delta)}{\sum_{k \geq 0} \frac{(xz)^k}{k!} \prod_{i=1}^b G(\lambda + i \delta)(\lambda + k \delta)} , \]

where

\[ z = -\frac{\lambda \alpha}{\delta} . \]

One should notice first that in each term in the R.H.S. of (19), there are possibly singularities other than \( s = \lambda \) situated inside the right half plane: \( s = \lambda - i \delta \) for those \( i \geq 1 \) such that \( \lambda - i \delta \geq 0 \) - if any - is such a singularity. Actually, we can check directly that removing any of these possible singularities (as we did for \( s = \lambda \)) provides the same expression for the unknown \( F(\lambda + \delta, x) \) so that the function

\[ F(s, z) = \phi_0(s) + \frac{\lambda \alpha zG(s)}{(\lambda - s)} \cdot \left\{ \sum_{k \geq 0} \frac{(xz)^k}{k!} \nu_k(s) \phi_0(s + (k+1)\delta) \right\} \]

\[ - \sum_{k \geq 0} \frac{(xz)^k}{k!} \nu_k(s)(s + k \delta) \cdot \sum_{k \geq 0} \frac{(xz)^k}{k!} \nu_k(\lambda)(\lambda + k \delta) \]

where

\[ \nu_k(s) = \prod_{i=1}^b G(s + i \delta) \frac{i \delta}{i \delta + s - \lambda} \]

is an analytic function of \((s, z)\) for \( \text{Re}(s) \geq 0 \) and \(|z| < 1\).

Accordingly, equation (11) provides the following expression for the expected number of customers served before the first failure:

\[ E[N^*] = \sum_{k \geq 0} \frac{z^k}{k!} \nu_k(0) \phi_0(k \delta + \delta) \]
5. Numerical Results

Next we give some numerical results based on equation (22). In figure 1, we have plotted $E[N^*]$ as a function of $1/\delta$ with $\lambda_f = 0.0001$ for the case of an exponential service time distribution with mean $0.5$ and the arrival rate $\lambda = 1$. We also compare the numbers obtained by equation (22) with those obtained by simulation. In figure 2, we keep $\lambda_f = 0.0001$ and vary the service time distribution, which is assumed to be gamma distributed with mean $0.5$. We vary the shape parameter $\alpha_0$ of the gamma distribution over the values $0.5, 1$ and $5.0$. We have assumed in the numerical example that $\phi_0(s) = 1$.

6. Conclusion

We have studied an $M/G/1$ queueing system with server breakdown and hard deadlines on job response time. Thus, a transient analysis of the system is performed in order to determine the average number of jobs completed before system failure. Extensions of the model in the direction of a more general "server" with multiple processors, subject to failure/repair type of degradation in the sense of [7,8] is needed.

Acknowledgement

We would like to thank Gianfranco Ciardo for his programming assistance with this paper.
Appendix

Let $G(z)$ and $F(z)$ be two analytic functions in the domain $|z| < 1$ with expansions

$$G(z) = \sum_{n \geq 0} \gamma_n z^n,$$
$$F(z) = \sum_{n \geq 0} \phi_n z^n,$$

Let $\Gamma$ be a circular contour of center $0$ and radius $R$ where $|z| < R < 1$ for a given $z$ so that $|z| < 1$.

For $u$ in the ring shaped domain $|z| < |u| < R$, $F(u) \cdot G(\frac{z}{u})$ is analytical in $u$ and has the Laurent expansion

$$F(u) \cdot G\left(\frac{z}{u}\right) = \sum_{n \geq 0} \sum_{k \geq 0} \phi_n \gamma_k z^k u^{n-k}$$

Hence, the coefficient of $u^{-j}, j \in N$ in this expansion is given by the contour integral

$$\sum_{n \geq 0} \phi_n \gamma_{n+j} z^{n+j} = \frac{1}{2i\pi} \int_{\Gamma} \frac{F(u)G\left(\frac{z}{u}\right)}{u^{1-j}} \, du.$$
References


Figure 1. Average number of jobs completed before system failure.
Figure 2. The effect of varying the service time distribution.
END

FILMED

11-85

DTIC