Analysis of SITA queues with many servers and space-time geometry

Eitan Bachmat  
Assaf Natanzon

1. INTRODUCTION AND PRELIMINARIES

SITA queues were introduced in [5] as a means for reducing job size variance at individual hosts in a server farm. It turns out that SITA queues are mathematically very interesting. For example, they satisfy a duality that is typical of automorphic forms in number theory. This leads to useful queueing theoretic insights, [1] and is also related to interesting number theoretic questions, [8], [6].

In this paper we will consider other aspects of SITA queues which turn out to be related to two dimensional Lorentzian geometry. In particular we will be interested in the behavior of SITA queues as $h \to \infty$. The tail of the waiting time function was studied in detail in [9]. We will concentrate on the average waiting time $E(W)$, rather than the tail since it leads to some interesting analogy and insights.

A SITA queue consists of $h$ hosts, numbered $1, \ldots, h$ and a set of cutoffs $0 = s_0 < s_1 < s_2 < \ldots < s_{h-1} < s_h = \infty$. Assuming that all hosts are identical, and given a job of size $x$, we assign it to host $i$ such that $s_{i-1} \leq x < s_i$.

More generally, if the hosts are not identical, we assume that each job is assigned some size $x$ and we proceed as before. We will assume that the hosts are coupled which means that the time it takes a job of size $x$ to process at host $j$ is a function $t_j(x)$. If the hosts are identical we take $t_j$ to be the identity function. We say that host strengths are linearly coupled if there exist constants $c_j$, $j > 1$ such that $t_j(x)/t_j(x) = c_j$ for all $x$. In this case, we again take $t_j$ to be the identity (job sizes are measured by the time they take on host 1).

2. CHOOSING GOOD CUTOFFS

One of the major problems in SITA queues is finding the optimal cutoffs, given all the other parameters, i.e., $X$, $t_j$, $\lambda$ and the target function $E(W)$. Several heuristics such as SITA-E and SITA-V, [5] have been suggested, however, the ratio between their performance and that of the optimal choice is not bounded, [2]. We present a choice of cutoffs which always offers reasonably good performance in the stochastic sense.

Given a SITA system, let $s_1^{\text{qbal}}, \ldots, s_h^{\text{qbal}}$ be the cutoffs which balance average queue length at the hosts. In some cases it may not be possible to find cutoffs which completely balance (equalize) average queue length. In those cases, we let $s_i^{\text{qbal}}$ be those cutoffs which minimize the maximal average queue length among all hosts. We have the following simple and basic result.

**Theorem 1.** Consider a SITA queue which satisfies the assumptions of Little’s law. For any, $s_1, \ldots, s_{h-1}$ we have

$$hE(W)(f, h, \rho, s_1, \ldots, s_{h-1}) \geq E(W)(f, h, \rho, s_1^{\text{qbal}}, \ldots, s_{h-1}^{\text{qbal}})$$

**Proof:** Given any set of cutoffs $s_1, \ldots, s_{h-1}$, let $p_i$ denote the portion of jobs which arrive at host $i$ of the SITA system, namely, the probability that $s_{i-1} \leq x < s_i$. We have

$$E(W) = \sum p_i E(W_i)$$

where $E(W_i)$ is the average waiting time at the $i$th host. For the same set of cutoffs we may also define $M(s_1, \ldots, s_{h-1}) = \max p_i E(W_i)$. Obviously, we have for any set of cutoffs $M(s_1, \ldots, s_{h-1}) \leq E(W)(s_1, \ldots, s_{h-1}) \leq hM(s_1, \ldots, s_{h-1})$. We note that the arrival rate to host $i$ is $\lambda p_i$, where $\lambda$ is the system arrival rate. Let $E(Q_i)$ be the average queue length at host $i$. By Little’s law, we have that

$$E(Q_i) = \lambda p_i E(W_i)$$

by definition $s_1^{\text{qbal}}, \ldots, s_h^{\text{qbal}}$ minimizes the function

$$\max_i E(Q_i)$$

which by (2) is the same as minimizing $M$.

We have

$$E(W)(s_1, \ldots, s_{h-1}) \geq M(s_1, \ldots, s_{h-1}) \geq M(s_1^{\text{qbal}}, \ldots, s_h^{\text{qbal}}) \geq E(W)(s_1^{\text{qbal}}, \ldots, s_h^{\text{qbal}})/h$$

Multiplying both ends of the equation by $h$ we get the desired result. q.e.d.

3. THE MANY SERVER CASE

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 2001 ACM X-XXXXX-XX-X/XX/XX ...$5.00.
The result above shows that cutoffs which balance queue length are never far from optimal when minimizing average waiting time, though, the average queue lengths for the optimal cutoffs can be highly unbalanced. This occurs for bounded Pareto distributions \( P_{\alpha,n} \) with \( \alpha_n \to 2 \) from below and \( I_n \to \infty \) at a fast enough pace. The same sequence shows that the constant \( h \) is in general, the best possible. However, as we will show, for a fixed bounded distribution the queue length balancing policy becomes optimal.

From now on, we will assume Poisson arrivals, that the job sizes are drawn from an i.i.d. generic distribution \( X \), and (at first) a fixed rate \( \lambda \) of arrivals regardless of the number of hosts. Our target function will be the average normalized waiting time \( E(N) = E(W)/E(X) \). We could have used \( E(W) \) just as well, however, is invariant under change of time units and also the duality of \( \text{SITA} \) queues, hence it is preferable.

We consider the case where \( X \) is some fixed distribution of job sizes in the range \([1, p]\) with a smooth strictly positive density function \( f \), a fixed arrival rate and the number of hosts \( h \) tending to infinity (becoming large). This asymptotic regime, was first considered in [4] under the name, providing a formula for choosing the cutoffs. We say that optimal cutoffs can be highly unbalanced. This occurs for \( \lambda \) and \( \rho/h \) tending to infinity (becoming large). This asymptotic performance of a \( \text{SITA} \) queue, with \( h \) identical hosts, a fixed arrival rate \( \lambda h \) and \( \rho \) and a job size distribution given by an infinitely differentiable density function \( f \) on the interval \([1, p]\). There exists a unique, differentiable, strictly increasing function \( g \) which is an \( h \)-asymptotically optimal cutoff formula.

The function \( g \) is a solution to the differential equation

\[
   f(g(x))g(x)g'(x) = c
\]

where \( c > 0 \) is some constant, and \( g \) satisfies the boundary conditions \( g(a) = 1 \) and \( g(b) = p \). The equation says that in the \( h \)-asymptotic regime the asymptotically optimal cutoffs, balance the queue length in each server and also balance load. The same conclusions hold for the case in which the arrival rate is \( \lambda = \lambda h \), i.e., grows linearly with the number of hosts.

We have

\[
   E(N) \approx \frac{\rho}{2h(L_f(1)) L_f(2)} \int_{1}^{\infty} f(x) g(x) dx,
\]

where \( \rho = h \lambda E(X) \).

**Proof** Each host is responsible for job sizes in a small range given by equation (3). Let \( dx_i = (b - a)/h \). Let \( ds_i = s_i - s_{i-1} \). We have

\[
   ds_i = s_i - s_{i-1} = g(a + (b - a)i/h) - g(a + (b - a)(i - 1)/h)
   \approx g'(x_i)(b - a)/h = g'(x_i)dx_i
\]

where \( f(h) \sim g(h) \) denotes the assertion that \( f(h)/g(h) \to 1 \) as \( h \to \infty \). Let \( k \) be a function on the positive reals and \( I \) an interval in the positive reals. Consider the incomplete Mellin transform of \( f \) w.r.t. \( I \), given by \( L_{f,I}(s) \). Let \( I = [s_{i-1}, s_i] \) denote the interval of job sizes assigned to host \( i \) if \( I = (0, \infty) \), we obtain the Mellin transform \( L_f(s) \).

Since \( f \) is assumed to be bounded from above, we see that the utilization \( \rho_i \) on each host tends to zero and hence \( 1 - \rho_i \to 1 \). From the Pollaczek-Khinchine formula we verify that the contribution of host \( i \) to the computation of \( E(N) \) is given by

\[
   E(N_i) = \frac{\rho}{2h(L_f(1)) L_f(2)} \int_{1}^{\infty} f(x_i) g(x_i) dx_i
\]

Similarly

\[
   E(N) = \sum_i E(N_i)
\]

Summing over \( i \) and recalling that \( dx_i = \frac{b - a}{h} \) we have the Riemann sum expression

\[
   E(N) \approx \frac{\rho}{2h(L_f(1)) L_f(2)} \int_{1}^{\infty} f(x_i) g(x_i) dx_i
\]

Taking the limit as \( h \) goes to infinity we see that

\[
   E(N) \approx \frac{\rho(b - a)}{2h(L_f(1)) L_f(2)} \int_{1}^{\infty} f(x_i) g(x_i) dx_i
\]

We need to minimize the integral expression. We note that the functional of interest has the form

\[
   F(x, g, g') = \tilde{L}(g)g'^2
\]

which is a special case of the more general form

\[
   F(x, g, g') = \tilde{L}(x, g)g'^2
\]

This bares a strong resemblance to functionals of the form

\[
   F(x, g, g') = \tilde{L}(x, g)g'^2
\]

which is the proper time of a causal curve in a two dimensional space-time with a Lorentzian metric written in light coordinates. As such it has been studied extensively, see for example [7]. It has also been studied in the context of maximal increasing subsequences among non uniformly distributed points, [3]. The maximizers of the proper time functional are the geodesics. The main difference between the proper time functional and the normalized waiting time functional is that the function \( \sqrt{z} \) which is applied to \( g' \) in the first case is strictly convex, while the function \( z^2 \) which is
applied to \( g' \) in the second case is strictly concave. This difference in particular is responsible for the fact that a generic functional to the identical server case. We get after the change of coordinates the previous results about uniqueness of the solution and having a conserved quantity (A Beltrami equation). We note that the same holds in Lorentzian geometry, the equation \( A(x)B(g)\sqrt{g'f} \) is the geodesic equation of flat Minkowski space, but in non-standard coordinates where one applies monotone changes of variables to the \( x \) and \( y \) coordinates separately. Essentially we are still in the Minkowski space case.

We may consider, more generally, the case of (non-linearly) coupled servers. To consider the large \( h \) limit, we again assume that the functions \( t_{c,h}(t) \), lead in the limit to a continuous function \( T(x,g) \). The resulting functional will have the form

\[
L(x,g)(g')^2
\]

As we noted previously. This is similar to the proper time functional in Lorentzian geometry. However, this time, the Lorentzian metric is not flat. It is in general a curved space-time, which no coordinate change will convert to Minkowski space. In analogy with Lorentzian geometry, we do not expect in this case to always have a unique solution and examples can be constructed where the solution is not unique. The solution need not be unique since the Euler-Lagrange equation is not a Beltrami type equation anymore and we do not have a balanced quantity. Going from the linearly coupled model to the coupled model is analogous to introducing curvature in space-time!

4. REFERENCES


