Supplement for "Allocation of Service Time in a Multi-Server System"

Muhammad El-Taha
Department of Mathematics and Statistics
University of Southern Maine
96 Falmouth Street
Portland, ME 04104-9300
Email:eltaha@usm.maine.edu
and
Bacel Maddah
Grado Department of Industrial and Systems Engineering
Virginia Tech
Blacksburg, VA 24061-0118
Email:bmaddah@vt.edu

September 9, 2005

Service Time Distributions

We present a brief summary of four distributions characterized by high variability, namely, Two Phase Hyperexponential, Erlang-Hyperexponential, Weibull, and truncated Pareto distributions.

Two Phase Hyperexponential (H_2).

We consider a Two Phase Hyperexponential which density function is given by $g(t) = 2p^2\mu e^{-2p\mu t} + 2(1-p)^2\mu e^{-2(1-p)\mu t}, (0 \leq p \leq \frac{1}{2})$. Therefore

$$
\begin{aligned}
E[S] &= 1/\mu, \;\; E[S^2] = 1/(2p(1-p)\mu^2), \\
P(S > T) &= pe^{-2p\mu T} + (1-p)e^{-2(1-p)\mu T}, \\
E[S_1] &= \frac{1}{\mu} - \frac{1}{2\mu}(e^{-2p\mu T} + e^{-2(1-p)\mu T}), \\
E[S_1^2] &= \frac{1}{2p(1-p)\mu^2} - \left(\frac{1}{2p\mu^2} + \frac{T}{\mu}\right) e^{-2p\mu T} - \left(\frac{1}{2(1-p)\mu^2} + \frac{T}{\mu}\right) e^{-2(1-p)\mu T}.
\end{aligned}
$$

This distribution is interpreted as a mixture of two exponential probability distributions.

Erlang-Hyperexponential (EH_2).

The density function of this distribution, a mixture of an exponential and an Erlang with two exponential phases, is given by

$g(t) = 4p^3\mu^2 te^{-2p\mu t} + 2(1-p)^2\mu e^{-2(1-p)\mu t}, (0 \leq p \leq 1)$. Therefore

$$
\begin{aligned}
E[S] &= 1.5/\mu, \;\; E[S^2] = (3-2p)/(2p(1-p)\mu^2), \\
P(S > T) &= 2p^2\mu T e^{-2p\mu T} + pe^{-2p\mu T} + (1-p)e^{-2(1-p)\mu T}, \\
E[S_1] &= \frac{1.5}{\mu} - \left(\frac{1}{\mu} + pT\right) e^{-2p\mu T} - \frac{1}{2\mu} e^{-2(1-p)\mu T}, \\
E[S_1^2] &= \frac{3-2p}{2p(1-p)\mu^2} - \left(\frac{3}{2p\mu^2} + \frac{3T}{\mu} + 2pT^2\right) e^{-2p\mu T} - \left(\frac{1}{2(1-p)\mu^2} + \frac{T}{\mu}\right) e^{-2(1-p)\mu T}.
\end{aligned}
$$

Weibull.

We consider the familiar Weibull distribution with shape factor α and scale factor β. The density function is given by $g(t) = \alpha\beta^{-\alpha}t^{\alpha-1}e^{-(t/\beta)^\alpha}$. Therefore

$$
\begin{aligned}
E[S] &= (\beta/\alpha)\Gamma(1/\alpha), \text{where } \Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt, \\
E[S^2] &= (2\beta^2/\alpha)\Gamma(2/\alpha), \;\; P(S > T) = e^{-(T/\beta)^\alpha}, \\
E[S_1] &= \int_0^T e^{-(t/\beta)^\alpha}dt, \;\; E[S_1^2] = \int_0^T \alpha\beta^{-\alpha}t^{\alpha+1}e^{-(t/\beta)^\alpha}dt + T^2 e^{-(T/\beta)^\alpha}.
\end{aligned}
$$

The above two integrals have no closed form solutions. In our analysis we evaluate these integrals numerically.

Truncated Pareto.

We consider a truncated Pareto distribution with a density function given by

$g(t) = \frac{\alpha l^\alpha}{1-(l/u)^\alpha}t^{-\alpha-1}, l \leq t \leq u$, where u and l are the largest and smallest job sizes and $0 < \alpha < 2$. In all the numerical examples in this paper, we fix u at 10^{10} and we select l in such a way as to keep the mean of the distribution equal to 3000. Therefore, for $l \leq T \leq u$,

$$
E[S] = 3000, \;\; E[S^2] = \frac{\alpha l^\alpha (l^{2-\alpha} - u^{2-\alpha})}{(\alpha-2)[(1-(l/u)^\alpha]},
$$

$$
\begin{aligned}
P(S>T) &= \frac{(l/T)^\alpha - (l/u)^\alpha}{1-(l/u)^\alpha}, \\
E[S_1] &= \frac{\alpha l^\alpha (l^{1-\alpha} - T^{1-\alpha})}{(\alpha-1)[(1-(l/u)^\alpha]} + TP(S>T), \\
E[S_1^2] &= \frac{\alpha l^\alpha (l^{2-\alpha} - T^{2-\alpha})}{(\alpha-2)[(1-(l/u)^\alpha]} + T^2 P(S>T).
\end{aligned}
$$

Harchol-Balter [10] indicates that such a job size distribution is found in many real world computer systems and cites many empirical studies that support this fact.

Proof of Analytic Results

Proof of Lemma 3.1. Conditioning on $\{S > T\}$, equation (7) implies that

$$E[S] = E[S_1] + P(S > T)E[S_2]\ . \tag{S.1}$$

By replacing $\lambda P(S > T)E[S_2]$ by its value from (S.1) the stability conditions are reduced to

$$\lambda E[S] - r < \lambda E[S_1] < m\ . \tag{S.2}$$

Notice that, for $T = 0$, $\lambda E[S_1] = 0$, i.e. all work is diverted to Station 2. On the other hand if $T = t_u$, where t_u is the smallest value of t such that $P(S > t) = 0$, then $\lambda E[S_1] = \lambda E[S]$, i.e. all work is processed through Station 1.[1] Moreover, $\lambda E[S_1]$ is increasing (and continuous) in T for $T \in (0, t_u)$ as shown in Section 2.1.1. Therefore, since $m > 0$ and $\lambda E[S] - r < m$ (from the overall stability condition), there always exists a T such that (S.2) holds. ∎

Proof of Lemma 3.2. The proof follows from (S.2) and the fact that $E[S_1]$ is increasing in T for $T \in (0, t_u)$. ∎

Proof of Corollary 3.3. The proof follows from Lemma 3.2 and the fact that $E[S_1]$ is increasing in T for $T \in (0, t_u)$. Note that $\overline{T} = \max_{m=1,\ldots c-1} \overline{T}_m = \overline{T}_{c-1}$ and $\underline{T} = \min_{m=1,\ldots c-1} \underline{T}_m = \underline{T}_1$. ∎

[1]Note that $t_u < \infty$ for bounded distributions only. For unbounded distributions, $t_u = \infty$, implying that all work cannot be processed by Station 1 except for an infinite T.

Proof of Lemma 3.4. Note that if $\rho \to 1$, (i.e. $\lambda E[S] \to c$), (S.2) implies that $E[S_1] \to (m/c)E[S]$, or equivalently $\rho_1 \to 1$, and $T^*_m \to T^b_m$. (Recall that $E[S_1]$ is continuous and increasing in T for $T \in (0, t_u)$.) Finally, (S.1) implies that $\lambda P(S > T^b_m)E[S_2] = \lambda E[S] - \lambda E[S_1] \to c - m = r$, or equivalently $\rho_2 \to 1$. ∎

Proof. of Theorem 3.7. Consider a balanced series system with $m = c - 1$, that is $r = 1$. Since $E[S^2]$ approaches infinity for the three distributions above, then, similar to the proof of Theorem 3.6, it suffices to show that $C^2_S > C^2_{S_2}$. To prove (i), using results from the first section of this supplement, we obtain

$$E[S_2] = \frac{e^{-2p\mu T^b_m} + e^{-2(1-p)\mu T^b_m}}{2\mu(pe^{-2p\mu T^b_m} + (1-p)e^{-2(1-p)\mu T^b_m})},$$

$$E[S_2^2] = \frac{\frac{1}{2p\mu^2}e^{-2p\mu T^b_m} + \frac{1}{2(1-p)\mu^2}e^{-2(1-p)\mu T^b_m}}{pe^{-2p\mu T^b_m} + (1-p)e^{-2(1-p)\mu T^b_m}}.$$

As $p \to 0$, $E[S_2] \to (1 + e^{-2\mu T^b_m})/(2\mu e^{-2\mu T^b_m})$, $E[S_2^2] \to 1/(2p\mu^2 e^{-2\mu T^b_m})$, and $C^2_S \to 1/(2p)$. Therefore, with $C^2_{S_2} = E[S_2^2]/(E[S_2])^2 - 1$,

$$\lim_{p\to 0} \frac{C^2_S}{C^2_{S_2}} = \frac{(1 + e^{-2\mu T^b_m})^2}{4e^{-2\mu T^b_m}} > 1,$$

which completes the proof of (i).

To prove (ii), it can be shown that when $S \sim EH_2$,

$$E[S_2] = \frac{\left(\frac{1}{\mu} + pT^b_m\right)e^{-2p\mu T^b_m} + \frac{1}{2\mu}e^{-2(1-p)\mu T^b_m}}{2p^2\mu T^b_m e^{-2p\mu T^b_m} + pe^{-2p\mu T^b_m} + (1-p)e^{-2(1-p)\mu T^b_m}},$$

$$E[S_2^2] = \frac{\left(\frac{3}{2p\mu^2} + \frac{T}{\mu}\right)e^{-2p\mu T^b_m} + \frac{1}{2(1-p)\mu^2}e^{-2(1-p)\mu^2 T^b_m}}{2p^2\mu T^b_m e^{-2p\mu T^b_m} + pe^{-2p\mu T^b_m} + (1-p)e^{-2(1-p)\mu T^b_m}}$$

As $p \to 0$, $E[S_2] \to (1/\mu + (1/2\mu)e^{-2\mu T^b_m})/(e^{-2\mu T^b_m})$, $E[S_2^2] \to 3/(2p\mu^2 e^{-2\mu T^b_m})$, and $C^2_S \to 2/(3p)$. Therefore,

$$\lim_{p\to 0} \frac{C^2_S}{C^2_{S_2}} = \frac{4}{9}\frac{(1 + \frac{1}{2}e^{-2\mu T^b_m})^2}{e^{-2\mu T^b_m}} = \frac{4}{9} + \frac{1}{9}\left(4e^{2\mu T^b_m} + e^{-2\mu T^b_m}\right) > 1,$$

where the last inequality follows since it can be easily shown that $4e^x + e^{-x} > 5$ for all $x > 0$. This completes the proof of (ii).

To prove (iii), it can be shown that when $S \sim Wbl$,

$$
\begin{aligned}
E[S_2] &= \frac{\frac{\beta}{\alpha}\Gamma\left(\frac{1}{\alpha}\right) - \int_0^{T_m^b} e^{(t/\beta)^\alpha} dt}{e^{-(T_m^b/\beta)^\alpha}} , \\
E[S_2^2] &= \frac{\frac{2\beta^2}{\alpha}\Gamma\left(\frac{2}{\alpha}\right) - \int_0^{T_m^b} \alpha\beta^{-\alpha} t^{\alpha+1} e^{-(t/\beta)^\alpha} dt - {T_m^b}^2 e^{-(T_m^b/\beta)^\alpha}}{e^{-(T_m^b/\beta)^\alpha}} - 2T_m^b E[S_2].
\end{aligned}
$$

As $\alpha \to 0$, $E[S_2] \to (\beta e/\alpha)\Gamma(1/\alpha)$ (since $\Gamma(\infty) = \infty$), $E[S_2^2] \to (2\beta^2 e/\alpha)\Gamma(2/\alpha) - 2T_m^b(\beta e/\alpha)\Gamma(1/\alpha)$, and $C_S^2 \to 2\alpha\Gamma(2/\alpha)/(\Gamma(1/\alpha))^2$. Therefore,

$$
\lim_{\alpha \to 0} \frac{C_S^2}{C_{S_2}^2} = e > 1,
$$

which completes the proof of (iii). ∎

Proof of Lemma 3.8. The proof of (i) follows by noting that $\rho_1 = \lambda E[S_1]/m$ and $E[S_1]$ is non-decreasing in T (see Section 2.1.1). The proof of (ii) also follows from the fact that $E[S_1]$ is non-decreasing in T by noting that $\rho_2 = \lambda(E[S] - E[S_1])/(c - m)$. To prove (iii), consider the derivative of $E[S_1^2]$ in (3) with respect to T

$$
\frac{dE[S_1^2]}{dT} = 2TP(S > T) .
$$

Noting that $C_{S_1}^2 = E[S_1^2]/(E[S_1])^2 - 1$, then

$$
\frac{dC_{S_1}^2}{dT} = \frac{2P(S > T)(TE[S_1] - E[S_1^2])}{(E[S_1])^3} \geq 0 ,
$$

where the last inequality follows from the fact that $S_1 \leq T$. ∎

Proof of Lemma 3.9. The proof is achieved by deriving a closed form expression for $C_{S_2}^2$. When

$S \sim H_2$, the expressions for $E[S_2]$ and $E[S_2^2]$ in the proof of Theorem 3.7 imply that

$$C_{S_2}^2 = \frac{2e^{-2\mu T}(-4 + \frac{1}{p(1-p)})}{(e^{-2p\mu T} + e^{-2(1-p)\mu T})^2} + 1 \; .$$

Therefore,

$$\frac{dC_{S_2}^2}{dT} = \frac{-4\mu e^{-2\mu T}(1-2p)(-4 + \frac{1}{p(1-p)})(e^{-2p\mu T} - e^{-2(1-p)\mu T})}{(e^{-2p\mu T} + e^{-2(1-p)\mu T})^3} \leq 0 \; ,$$

for $0 \leq p \leq \frac{1}{2}$. ∎

Proof of Lemma 3.10. The proof follows directly from the above two lemmas. When $T = 0$,we have $C_{S_2}^2 = C_S^2$. Since, when $S \sim H_2$ and $p < \frac{1}{2}$, $C_{S_2}^2$ is strictly decreasing in T, we obtain $C_{S_2}^2 < C_S^2$. When $T = \infty$, equations (2) and (3) imply that $C_{S_1}^2 = C_S^2$. Since $C_{S_1}^2$ is strictly increasing in T, we conclude that $C_{S_1}^2 < C_S^2$. ∎