SAND REPORT

SAND2002-0942
Unlimited Release
Printed May 2002

Maximum Utilization of Parallel Computers

Vitus Leung, John DeLaurentis

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865)576-8401
Facsimile: (865)576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.doe.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800)553-6847
Facsimile: (703)605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/ordering.htm

SAND2002-0942
Unlimited Release
Printed May 2002

Maximum Utilization of Parallel Computers

Vitus Leung
Optimization & Uncertainty Estimation Department

John DeLaurentis
Computational Mathematics & Algorithms Department

Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-1110

Abstract

Two GAO reports have questioned the utilization of computing resources in the Department of Energy. While Jones and Nitzberg have observed that utilization peaks at 60-80% for a variety of architectures and allocation policies. This investigation examines the theoretical and observed average maximum efficiencies of massively parallel computers, and shows that the observed efficiency at Sandia is very nearly optimal.

Here, the average maximum efficiency is defined as the expected utilization or efficiency when the queue is nonempty. We have developed a model that allows us to compare the observed efficiency with the average maximum efficiency, and allows us to forecast the expected maximum efficiency given the average size of the smallest task, S, waiting in the queue. The model predictions are in excellent agreement with the measured efficiencies obtained from the Sandia data.

The average number of idle processors may be estimated by analyzing the embedded renewal process. We let $Y(t)$ denote the number of idle processors, and we set S equal to the random variable representing the size of the smallest task in the queue. The stopping time T^* denotes the time when exactly S processors become available; the process $Y(t)$ is reset to zero when the level S is attained. Also, we let N denote the number of processors, $\gamma = E[S/N]$, and we set $\rho(\gamma)$ equal to the efficiency when $E[S/N] = \gamma$; that is, $\rho(\gamma)$ is the maximum efficiency when the smallest waiting task requires, on average, γN nodes. We show that $\rho(\gamma) = \lambda^{-1}E[S/N]/E[T^*]$, where λ^{-1} is the mean of the exponentially distributed completion times.

Acknowledgements

We would like to thank Sue Kelly of Sandia National Laboratories' Scalable Systems Integration Department for providing us with data. We thank Robert Carr of the Optimization & Uncertainty Estimation Department for helpful discussions.

Contents

List of Figures

1 Introduction

Two GAO reports (see [2] and [3]) have questioned the utilization of computing resources in the Department of Energy. While Jones and Nitzberg (see [5]) have observed that utilization peaks at 60-80% for a variety of architectures and allocation policies. This investigation examines the theoretical and observed average maximum efficiencies of massively parallel computers, and shows that the observed efficiency at Sandia is very nearly optimal.

Here, the average maximum efficiency is defined as the expected utilization or efficiency when the queue is nonempty. We have developed a model that allows us to compare the observed efficiency with the average maximum efficiency, and allows us to forecast the expected maximum efficiency given the average size of the smallest task, S, waiting in the queue. The dependence on the smallest task reflects the scheduling policy of not blocking the smallest task in the queue; this is the policy at Sandia. The model predictions are in excellent agreement with the measured efficiencies obtained from the Sandia data.

The average number of idle processors may be estimated by analyzing the embedded renewal process. We let $Y(t)$ denote the number of idle processors, and we set S equal to the random variable representing the size of the smallest task in the queue. The stopping time T^* denotes the time when exactly S processors become available; the process $Y(t)$ is reset to zero when the level S is attained, see Fig. 1. Also, we let N denote the number of processors, $\gamma = E[S/N]$, and we set $\rho(\gamma)$ equal to the efficiency when $E[S/N] = \gamma$; that is, $\rho(\gamma)$ is the maximum efficiency when the smallest waiting task requires, on average, γN nodes.

In terms of the renewal process, we define the average maximum efficiency, $\rho(\gamma)$, $0 < \gamma < 1$, by

$$\rho(\gamma) \equiv \lim_{M\to\infty} \left(1 - \overline{Y}_M/N\right), \tag{1.1}$$

where we assume that the average, for $t_j = j\Delta t$, $\Delta t > 0$,

$$\overline{Y}_M \equiv \frac{1}{M}\sum_{j=1}^{M} Y(t_j), \tag{1.2}$$

converges in mean square (the limit in expression (1.1) is convergence in mean square). Assuming that the completion times (the time required for a processor to complete a task) are independent, exponentially distributed random variables, $S \geq 1$, and that S is constant between renewal times , it can be shown that $\overline{Y}_M$ converges. Also, under the same assumptions on the completion times and S, it is possible to derive a theoretical expression for $\rho(\gamma)$, namely,

$$\rho(\gamma) \equiv \lim_{M\to\infty} \left(1 - \overline{Y}_M/N\right) = 1 - \frac{E\left[\int_0^{T^*} Y(t)dt\right]}{N\, E[T^*]} = \lambda^{-1}\frac{E[S/N]}{E[T^*]}, \tag{1.3}$$

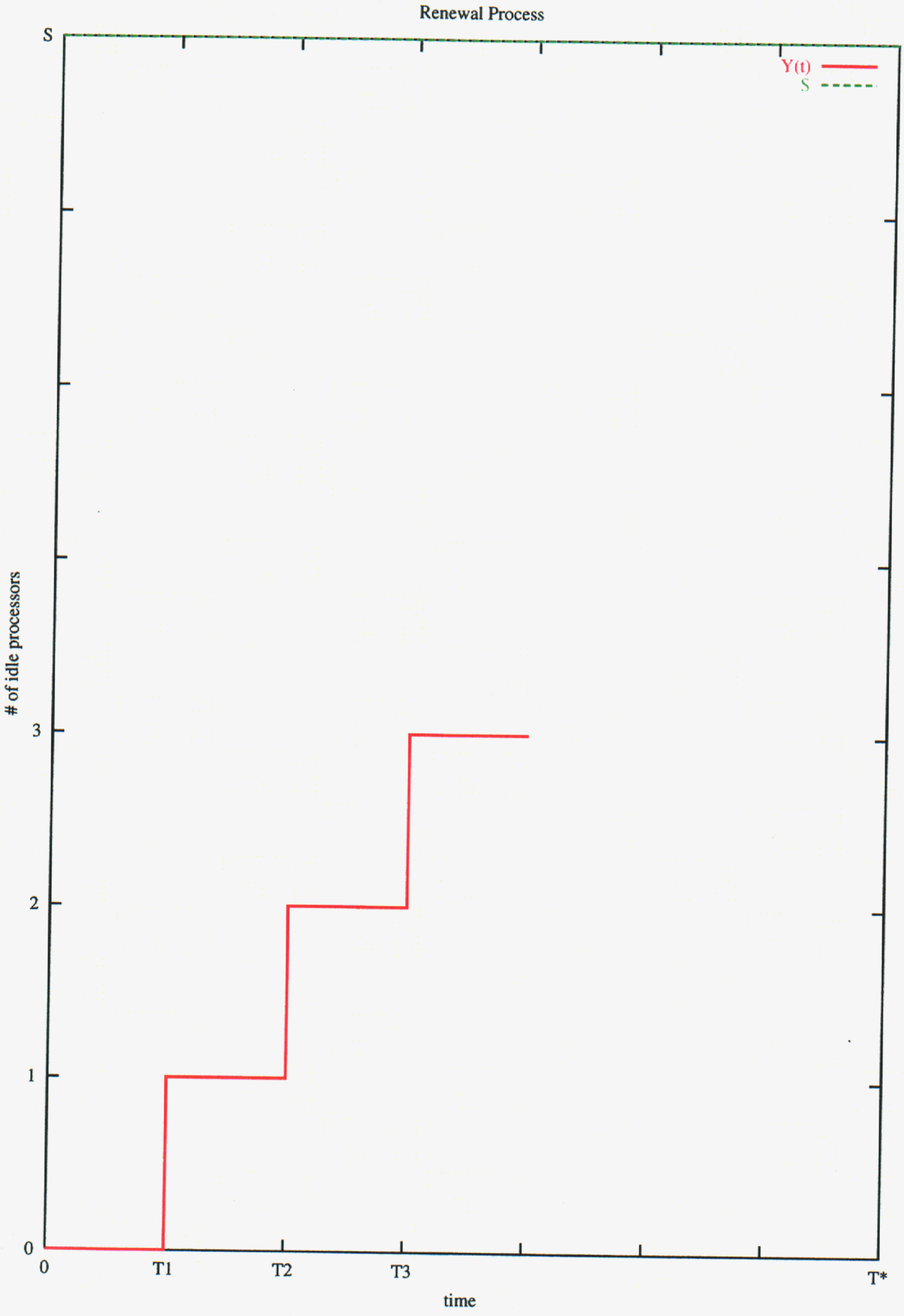

Figure 1: Renewal Process

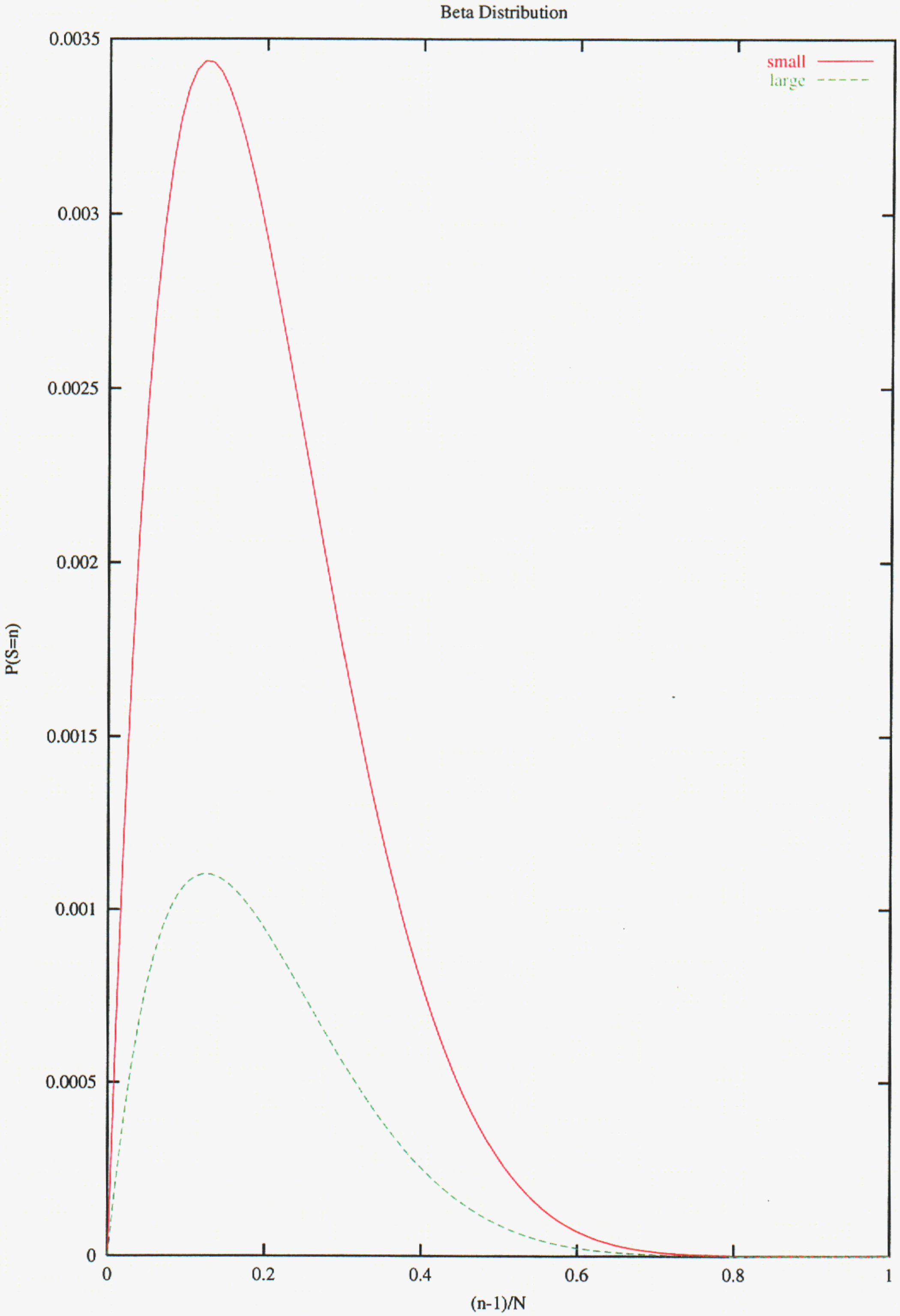

Figure 2: Beta Law ($\alpha = 2, \beta = 8$)

where λ^{-1} is the mean of the exponential random variable. Expression (1.3) allows us to make a comparison between theoretical and experimental efficiencies.

We have observed, from empirical studies, that the distribution of S approximately obeys a Beta law (see Fig. 2),

$$P(S = n) \approx \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(\frac{n-1}{N}\right)^{\alpha-1} \left(1 - \frac{n-1}{N}\right)^{\beta-1} \frac{1}{N}, \tag{1.4}$$

where $n = 1, \ldots, N$; $\alpha, \beta \geq 1$, $\Gamma(z)$ denotes the gamma function and $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ is the approximate normalization constant. The mean of S/N is given by $\gamma = E[S/N] \approx \alpha/(\alpha+\beta)$; that is, the average number of nodes required for the smallest task is approximately $\left(\frac{\alpha}{\alpha+\beta}\right) N$. At Sandia, there are two possible values for N depending on whether the machine is in large or small configuration.

If α is an integer and β is "large" (but not too large since we also require that $\frac{\Gamma(\alpha+\beta)}{N\Gamma(\alpha)\Gamma\beta} \rightarrow 0$, as $N \rightarrow \infty$) we have (see Fig. 3)

$$\rho(\gamma) \approx -\frac{\gamma}{\ln(1-\gamma)} = \frac{1}{1 + \gamma/2 + \gamma^2/3 + \cdots}, \tag{1.5}$$

In particular, if $E[S/N] \approx .2$ the model predicts a maximum efficiency of 90%; $\rho(.2) \approx -.2/\ln(1-.2) \approx .9$, which is in excellent agreement with the measured maximum utilization of approximately 90%. Here, the measured maximum efficiency was obtained by computing the efficiency over periods in which the queue was nonempty.

The ratio of the average maximum efficiency and the observed utilization may be used to provide a relative efficiency. Using the relative efficiency as a figure of merit, we found that the Sandia computers performed better than expected. The Sandia data may be summarized as follows:

- Measured maximum efficiency is 90% over 73% of snapshots
- Measured low efficiency is 49% over 27% of snapshots
- Observed efficiency is approximately 79% over 100% of snapshots
- Relative efficiency $\equiv .79/.9 \approx .88$ or approximately 88%

The key point is that the ratio of the observed efficiency, 79%, to the maximum efficiency, 90%, or the relative efficiency is approximately 88%, which is nearly optimal. Additionally, the curve for $\rho(\gamma)$ (see Fig. 3) may be used to predict the maximum utilization for other values of the parameter γ, provided estimates of the expected value $E[S/N]$ are available for the time period of interest. We further broke out the smallest job waiting in the Sandia queues into fourteen ranges and collected the corresponding efficiencies. Fig. 4 shows the top seven ranges. Note that not all the ranges are populated in both the small and large configurations. The difference between the predicted and observed values could be due to the sparsity of the observed values.

The Sandia data also revealed some trends for average daily usage, see Fig. 5. The highest average daily usage occurs on Fridays, 94%. The average daily usage decreases steadily over the weekend with 83% on Saturdays and 75% on Sundays

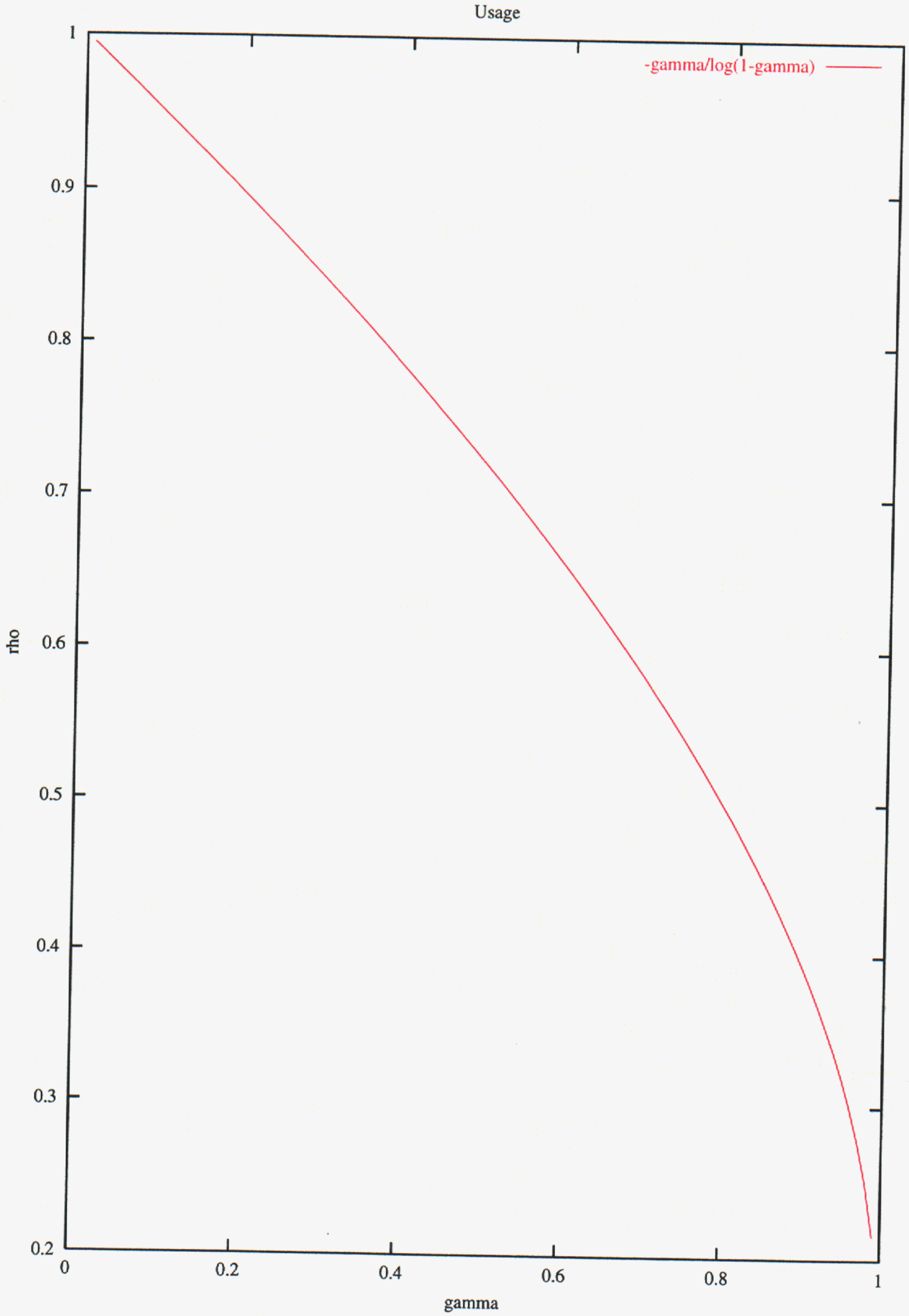

Figure 3: Maximum Efficiency

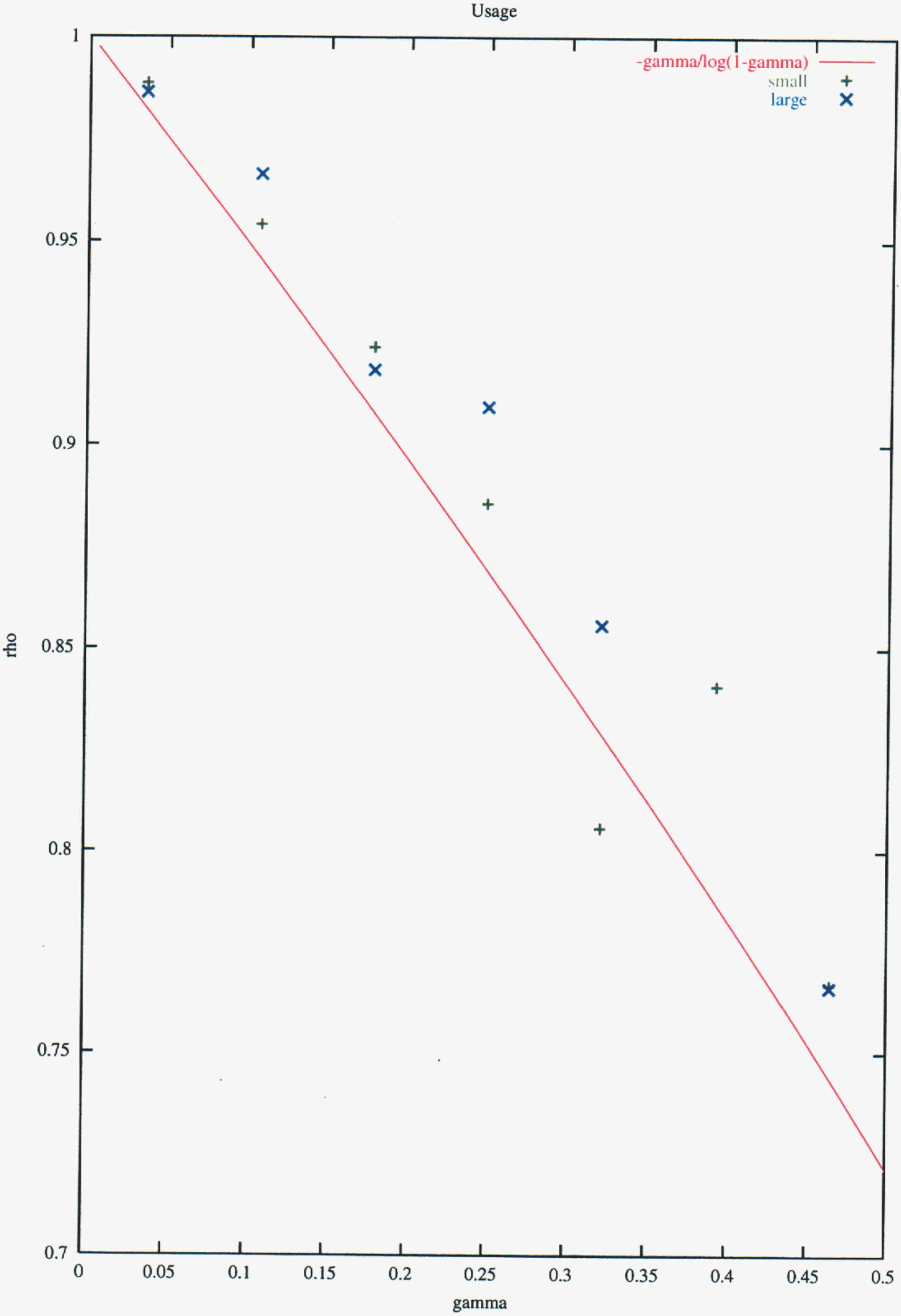

Figure 4: Maximum Efficiencies

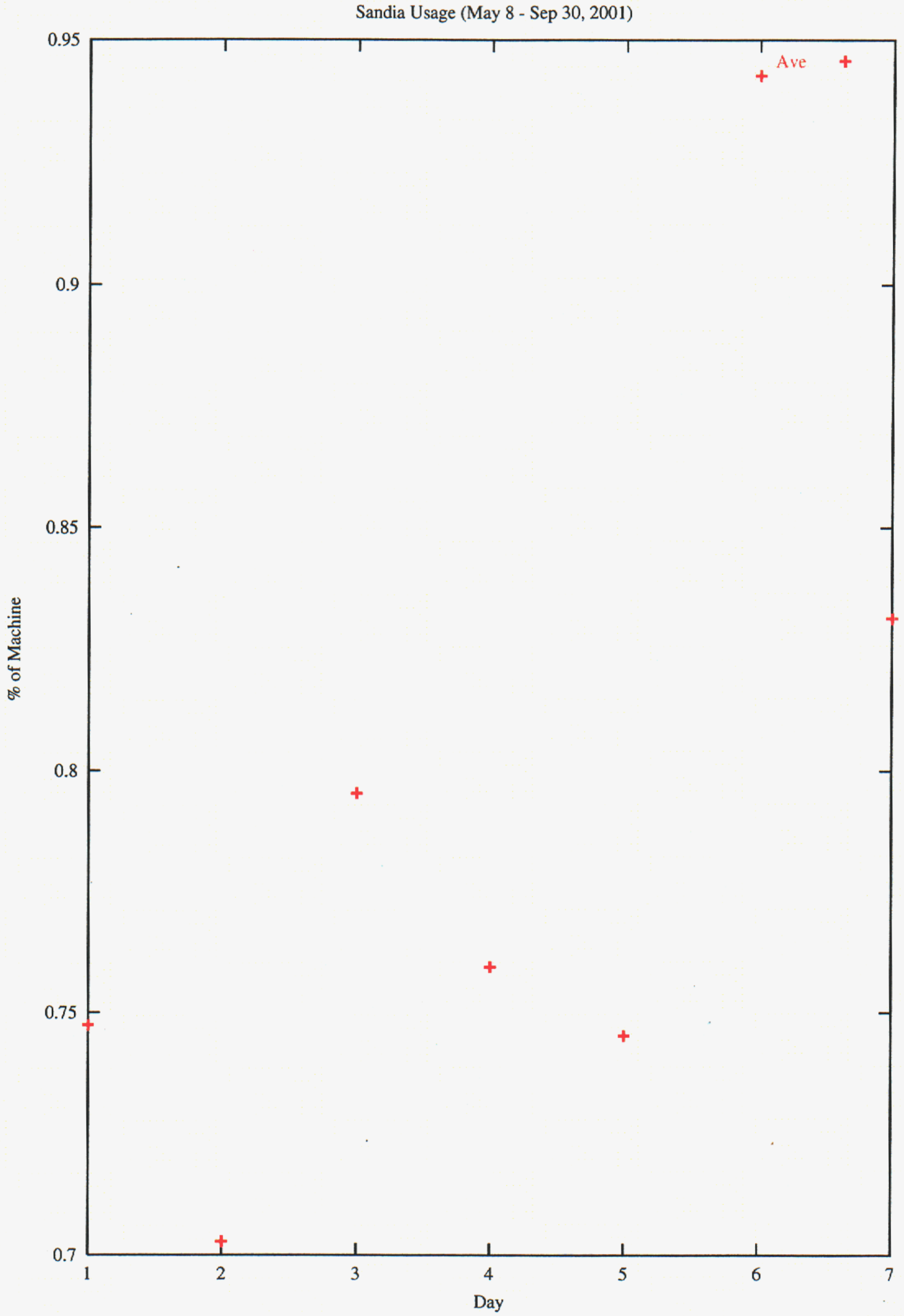

Figure 5: Average Daily Usage

before reaching the lowest average daily usage on Mondays, 70%. A similar trend occurs during the work week with 80% on Tuesdays, 76% on Wednesdays, and 75% on Thursdays. These trends could be the result of the Sandia queues being loaded with production jobs before the weekend, but the job queues being exhausted before the work week begins and the Sandia queues being loaded with development jobs shortly after the work week begins, but the development work diminishing over the work week going into the next weekend cycle.

In the second section, we define the embedded renewal process more precisely, and derive relation (1.3). The third section presents an asymptotic analysis for the distribution of the smallest task size, and, in turn, this expansion provides an estimate for $\rho(\gamma)$.

2 Expected Maximum Efficiency

In this section we present an outline of the proof that mean square limit of $\overline{Y}_M$ exists and equals $\lambda^{-1}E[S/N]/E[T^*]$ so that $\rho(\gamma) = \lambda^{-1}E[S/N]/E[T^*]$. The key idea is that the number of idle processors may be represented by an embedded renewal process. In turn, we use the renewal theorem and the fact the process is approximately covariance stationary to derive convergence.

We begin by defining the renewal process more precisely. We let S_1 denote the size of the smallest task in the queue for the first renewal period, and we let $T_1^* = T_{S_1}$ be the time for enough processors to become available to accommodate the task of size S_1. For the second renewal period we set S_2 equal to the smallest task in the queue, we let T_{S_2} be the time for S_2 processors to become available, and we define $T_2^* = T_{S_1} + T_{S_2}$. The i^th renewal period is defined similarly, if S_i denotes the size of the smallest task in the queue during the i^th period, $T_i^* = T_{S_1} + T_{S_2} + ... + T_{S_i}$, where T_{S_i} is the time required for S_i processors to become available. We let $Y(t)$ denote the number of idle processors at time t, $Y(t) = Y(t - T_i^*)$ if $t > T_i^*$. Here, we assume that the number of idle processors is immediately reset to zero when enough processors become available for the smallest task in the queue, and the process begins again with another randomly chosen minimum task size. Also, we assume that the size of the smallest task remains constant throughout the period. The renewal process may be viewed as a hitting process that begins anew whenever the number of idle processors reach a random level S.

Next, to analyze the stopping times, we assume that the time, τ, required for a processor to complete a task is exponentially distributed with mean λ^{-1} ($P(\tau > t) = \exp(-\lambda t)$), and is independent of the other processors' completion times. If k processors are idle, the time, τ_k, for another processor to become available is given by the minimum over $N - k$ independent, exponentially distributed random variables. For a given renewal period, let us denote by; τ_0 the time for the first processor to become idle, $\tau_0 + \tau_1$ the time for the second to become available, $T_k \equiv \tau_0 + \tau_1 + ... + \tau_{k-1}$ the time for the k^{th} processor to become idle, and for, say, the first renewal period,

we note $T_{S_1} = \tau_0 + ... + \tau_{S_1-1}$, see Fig. 1.
The following is a partial list of symbols:

N = number of processors,
τ_k = duration of time that exactly k processors are idle,
$T_k = \tau_o + \cdots + \tau_{k-1}$ = time that the k^{th} processor becomes idle,
S_i = size of the smallest task in the queue during renewal period i,
$T_1^* = T_{S_1} = \sum_{k=0}^{S-1} \tau_k$ = time for exactly S_1 processors to become available,
Y(t) $= \sum_{k=0}^{S-1} I_{[T_k,T_{k+1})}(t)$ = number of processors idle at time t if $t < T_{S_1}$,

where $I_{[T_k,T_{k+1})}(t)$ denotes the indicator function on the set $[T_k, T_{k+1})$.

The distribution of τ_k is given by the minimum of $N-k$ independent, identically distributed random variables, τ_{k_j}, that is,

$$P(\tau_k > t) = P\left(\min_{1\leq j\leq N-k} \tau_{k_j} > t\right) = P(\tau_{k_j} > t)^{N-k} = e^{-\lambda(N-k)t}. \tag{2.1}$$

Here, the $N-k$ random variables represent the $N-k$ processors that are working on a task, so that, exactly k processors are idle. In the last step we used the assumption that τ_{k_j} is exponentially distributed with parameter λ.

Let us establish a renewal equation for

$$A(t) = E[Y(t)]. \tag{2.2}$$

We condition on the time of the first renewal, $T_1^* = s$, and consider two cases; $t < s$, so that $T_1^* > t$, and $t \geq s$ so that $Y(t) = Y(t - T_1^*)$. We have

$$E[Y(t)|T_1^* = s] = \begin{cases} E[Y(t)|T_1^* = s], & t < s, \\ A(t-s), & t \geq s. \end{cases} \tag{2.3}$$

Invoking the law of total probabilities, we obtain

$$\begin{aligned} A(t) &= E[Y(t)] \\ &= \int_0^\infty E[Y(t)|T_1^* = s]P(T_1^* \in ds) \\ &= \int_0^t E[Y(t-s)|T_1^* = s]P(T_1^* \in ds) + \int_t^\infty E[Y(t)|T_1^* = s]P(T_1^* \in ds) \\ &= a(t) + \int_0^t A(t-s)P(T_1^* \in ds) \end{aligned} \tag{2.4}$$

where $a(t) = \int_t^\infty E[Y(t)|T_1^* = s]P(T_1^* \in ds)$, and $\int_B P(T_1^* \in ds) \equiv P(T_1^* \in B)$ denotes the probability measure for T_1^*.

We have shown that $A(t)$ satisfies the renewal equation

$$A(t) = a(t) + \int_0^t A(t-s)P(T_1^* \in ds). \tag{2.5}$$

By the renewal theorem (see Karlin and Taylor [6]) we obtain

$$\lim_{t\to\infty} A(t) = \frac{1}{E[T_1^*]} \int_0^\infty a(t)dt, \tag{2.6}$$

where $a(t) = \int_t^\infty E[Y(t)|T_1^* = s]P(T_1^* \in ds)$.

Interchanging the order of integration we arrive at the expression

$$\begin{aligned} \int_0^\infty a(t)dt &= \int_0^\infty \int_t^\infty E[Y(t)|T_1^* = s]P(T_1^* \in ds)dt \\ &= \int_0^\infty \int_0^s E[Y(t)|T_1^* = s]dtP(T_1^* \in ds) = E\left[\int_0^{T_1^*} Y(t)dt\right] \end{aligned} \tag{2.7}$$

The change in order of integration is justified by the estimate

$$\int_t^\infty E[Y(t)|T_1^* = s]P(T_1^* \in ds) = 0(e^{-\lambda t})$$

obtained from expression (2.1). We have, from expressions (2.6) and (2.7),

$$\lim_{t\to\infty} A(t) = \frac{E\left[\int_0^{T^*} Y(t)dt\right]}{E[T^*]}, \tag{2.8}$$

where $T^* \equiv T_1^*$. It follows that

$$\frac{1}{M}\sum_{j=1}^{M} E[Y(t_j)] = \frac{1}{M}\sum_{j=1}^{M} A(t_j) \to \mu \equiv \frac{E\left[\int_0^{T^*} Y(t)dt\right]}{E[T^*]}, \tag{2.9}$$

as $M \to \infty$. Now assuming

$$\frac{1}{M^2}\sum_{1\leq j,k\leq M} E[(Y(t_j) - \mu)(Y(t_k) - \mu)] \to 0, \tag{2.10}$$

as $M \to \infty$, it follows that (see Karlin and Taylor [6])

$$E\left[(\bar{Y}_n - \mu)^2\right] \to 0, \tag{2.11}$$

as $n \to \infty$, where $\bar{Y}_M \equiv \frac{1}{M}\sum_{j=1}^M Y(t_j)$. Expression (2.10) follows from the observation that for $|t - s|$ sufficiently large, $Y(t)$ and $Y(s)$ are "essentially" independent, see Karlin and Taylor [6].

We have arrived at the desired result,

$$\bar{Y}_M \equiv \frac{1}{M}\sum_{j=1}^{M} Y(t_j) \to \frac{E\left[\int_0^{T^*} Y(s)ds\right]}{E[T^*]}, \tag{2.12}$$

as $M \to \infty$; convergence is in mean square. We have shown the first half of expression (1.3), namely, that $\bar{Y}_M$ converges, and that it converges to the ratio $E\left[\int_0^{T^*} Y(s)ds\right] \Big/ E[T^*]$. It remains to show that this ratio can be expressed in terms of the random variable $S \equiv S_1$.

The expected values $E\left[\int_0^{T^*} Y(t)dt\right]$ and $E[T^*]$ (recall that $T^* \equiv T_1^*$) may be expressed in terms of the exponential random variables τ_k (see expression 2.1); that is,

$$\begin{aligned}
E[T^*] &= E\left[\sum_{k=0}^{S-1} \tau_k\right] \\
&= \sum_{m=1}^{N} \left(\sum_{k=0}^{m-1} E[\tau_k]\right) P(S=m) \\
&= \sum_{m=1}^{N} \left(\sum_{k=0}^{m-1} \tfrac{1}{\lambda(N-k)}\right) P(S=m) \\
&= \lambda^{-1} \sum_{k=0}^{N-1} \tfrac{1}{N-k} P(S>k),
\end{aligned} \tag{2.13}$$

and

$$\begin{aligned}
E\left[\int_0^{T^*} Y(t)dt\right] &= E\left[\int_0^{T_S} Y(t)dt\right] \\
&= \sum_{m=1}^{N} E\left[\sum_{m=0}^{S-1} [k\tau_k | S=m]\right] P(S=m) \\
&= \sum_{m=1}^{N} E\left[\sum_{k=0}^{m-1} \tfrac{k}{\lambda(N-k)}\right] P(S=m) \\
&= \tfrac{1}{\lambda} \sum_{k=0}^{N-1} \tfrac{k}{N-k} P(S>k).
\end{aligned}$$

It follows that

$$\begin{aligned}
E\left[\int_0^{T^*} Y(t)dt\right] &= \tfrac{1}{\lambda} \sum_{k=0}^{N-1} \tfrac{N-(N-k)}{N-k} P(S>k) \\
&= \tfrac{N}{\lambda} \sum_{k=0}^{N-1} \tfrac{1}{N-k} P(S>k) - \tfrac{1}{\lambda} \sum_{k=0}^{N-1} P(S>k) \\
&= N\,E[T^*] - \lambda^{-1} E[S].
\end{aligned} \tag{2.14}$$

Here, we used the identity $\sum_{k=0}^{N-1} P(S>k) = E[S]$. Using expressions (2.12) and (2.14), we have arrived at the desired conclusion,

$$1 - \frac{1}{N}\bar{Y}_M \to 1 - \frac{E\left[\int_0^{T^*} Y(s)ds\right]}{N\,E[T^*]} = 1 - \frac{N\,E[T^*] - \lambda^{-1}E[S]}{N\,E[T^*]} = \frac{\lambda^{-1}E[S]}{N\,E[T^*]}, \tag{2.15}$$

as $M \to \infty$; convergence is in mean square. This completes the proof of relation (1.3).

This relation allows the experimentalist to compare the statistic $\bar{Y}_M$, obtained from observations, with the theoretical prediction, provided the distribution of the smallest task size, S, is known.

3 Distribution of the Smallest Task

We turn, now, to an analysis of the distribution of the smallest task size, S, and we use an asymptotic expansion for the distribution to obtain an analytical expression for the ratio $\lambda^{-1}E[S/N]\,E[T^*] = \rho(\lambda)$. As we noted in the introduction, our empirical studies indicate that S may be approximately represented by a beta distribution; that is,

$$P(S=n) = \frac{c}{N}\left(\frac{n-1}{N}\right)^{\alpha-1}\left(1-\frac{n-1}{N}\right)^{\beta-1}, n = 1, ..., N, \tag{3.1}$$

where c is chosen so that $\sum_{n=1}^{N} P(S=n) = 1$, $\alpha > 0$, and $\beta \geq 1$ (we note that $P(S=1) = 0$). To relate the beta distribution to the efficiency, $\rho(\gamma)$, for a given $\gamma \equiv E[S/N]$, the parameters α and β must be chosen so that $E[S/N] = \gamma + O(1/N)$, for the specified γ, $0 < \gamma < 1$. For this we use the beta integral

$$\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}dy, \tag{3.2}$$

which holds true for $\alpha > 0$ and $\beta > 0$. This integral may also be used to provide an estimate of c; that is, assuming $\alpha > 0$ and $\beta \geq 1$, we have

$$\begin{aligned} 1 &= \sum_{n=1}^{N} P(S=n) = c\sum_{n=1}^{N-1}\left(\tfrac{n}{N}\right)^{\alpha-1}\left(1-\tfrac{n}{N}\right)^{\beta-1}\tfrac{1}{N} \\ &= c\int_0^{1-\frac{1}{N}} y^{\alpha-1}(1-y)^{\beta-1}dy + O\left(N^{-1}\right) \\ &= c\int_0^1 y^{\alpha-1}(1-y)^{\beta-1}dy + O\left(N^{-1}\right) = c\tfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} + O\left(N^{-1}\right). \end{aligned} \tag{3.3}$$

It follows that $c = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} + O(N^{-1})$.

The expected value of S/N is approximately given by (assuming that $\beta \geq 1$),

$$\begin{aligned} E[S/N] &= c\sum_{n=1}^{N-1}\left(\tfrac{n}{N}\right)^{\alpha}\left(1-\tfrac{n}{N}\right)^{\beta-1}\tfrac{1}{N} + \tfrac{1}{N} \\ &= c\int_0^{1-\frac{1}{N}} y^{\alpha}(1-y)^{\beta-1}dy + O\left(N^{-1}\right) \\ &= c\int_0^1 y^{\alpha}(1-y)^{\beta-1}dy + O\left(N^{-1}\right) \\ &= c\tfrac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} + O\left(N^{-1}\right) = \tfrac{\alpha}{\alpha+\beta} + O\left(N^{-1}\right). \end{aligned} \tag{3.4}$$

In the last step we used the fact that $c = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} + O(N^{-1})$, and the identity $\Gamma(z+1) = z\Gamma(z)$. In order that $E[S/N]$ satisfy the relation $E[S/N] = \gamma + O(1/N) \approx \gamma$ we set

$$\frac{\alpha}{\alpha+\beta} = \gamma. \tag{3.5}$$

We note that the parameters α and β are not uniquely determined by this identity.

To evaluate $E[T^*]$ we need an expression for the incomplete beta integral, namely,

$$\sum_{v=0}^{k}\binom{r}{v}p^v(1-p)^{r-v} = (r-k)\binom{r}{k}\int_p^1 t^k(1-t)^{r-k-1}dt, \tag{3.6}$$

where, if r is a real number, v a positive integer, we set $\binom{r}{v} = \frac{r(r-1)...(r-v+1)}{v(v-1)...(1)}$, $\binom{r}{0} \equiv 1$, and we assume k a positive integer, $r-k > 0$, $0 < p \leq 1$. Identity (3.6) may be verified by differentiating both sides with respect to p, see Feller [1] or Hogg and Craig [4]. (By allowing $p \downarrow 0$ this identity may be used to derive relation (3.2) for integer α and real β; in this case, $\alpha = k+1$ and $\beta = r-k$.) Using identity (3.6), we have, for $\alpha = k+1$ and $\beta = r-k$,

$$\begin{aligned} P(S > m) &= c \sum_{j=m}^{N-1} \left(\tfrac{j}{N}\right)^{\alpha-1} \left(1 - \tfrac{j}{N}\right)^{\beta-1} \tfrac{1}{N} \\ &= c \textstyle\int_{m/N}^{1} y^k (1-y)^{r-k-1} dy + O\left(N^{-1}\right) \\ &= \tfrac{c}{(r-k)\binom{r}{k}} \sum_{v=0}^{k} \binom{r}{v} p_m^v (1-p_m)^{r-v} + O\left(N^{-1}\right), \end{aligned} \tag{3.7}$$

where $p_m = m/N$. For $E[T^*]$, by applying the estimate (3.7) and expression (2.13), we obtain,

$$\begin{aligned} E[T^*] &= \lambda^{-1} \sum_{m=1}^{N-1} \tfrac{1}{N-m} P(S > m) \\ &= \tfrac{\lambda^{-1} c}{(r-k)\binom{r}{k}} \sum_{m=1}^{N-1} \tfrac{1}{N} \tfrac{1}{(1-p_m)} \sum_{v=0}^{k} \binom{r}{v} p_m^v (1-p_m)^{r-v} + O\left(N^{-1}\right) \\ &= \tfrac{\lambda^{-1} c}{(r-k)\binom{r}{k}} \sum_{v=0}^{k} \binom{r}{v} \sum_{m=1}^{N-1} p_m^v (1-p_m)^{r-v-1} \tfrac{1}{N} + O\left(N^{-1}\right). \end{aligned} \tag{3.8}$$

The inner sum in expression (3.8) may rewritten as

$$\begin{aligned} \sum_{m=1}^{N-1} \left(\tfrac{m}{N}\right)^v \left(1 - \tfrac{m}{N}\right)^{r-v-1} \tfrac{1}{N} &= \textstyle\int_0^1 y^v (1-y)^{r-v-1} dy + O\left(N^{-1}\right) \\ &= \tfrac{\Gamma(v+1)\Gamma(r-v)}{\Gamma(r+1)} + O\left(N^{-1}\right) \\ &= \tfrac{v!}{r(r-1)\ldots(r-v)} + O\left(N^{-1}\right) \\ &= \tfrac{1}{(r-v)\binom{r}{v}} + O\left(N^{-1}\right). \end{aligned} \tag{3.9}$$

Combining expressions (3.8), (3.9), and using the relation $c = \frac{\Gamma(r+1)}{\Gamma(k+1)\Gamma(r-k)} + O\left(N^{-1}\right)$, since $\alpha = k+1$ and $\beta = r-k$, we obtain

$$\begin{aligned} E[T^*] &= \tfrac{\lambda c}{(r-k)\binom{r}{m}} \sum_{v=0}^{k} \binom{r}{v} \tfrac{1}{(r-v)\binom{r}{v}} + O\left(N^{-1}\right) \\ &= \tfrac{\lambda^{-1}}{(r-k)\binom{r}{k}} \tfrac{\Gamma(r+1)}{\Gamma(k+1)\Gamma(r-k)} \sum_{v=0}^{k} \tfrac{1}{r-v} + O\left(N^{-1}\right) \\ &= \lambda^{-1} \sum_{v=0}^{k} \tfrac{1}{r-v} + O\left(N^{-1}\right). \end{aligned} \tag{3.10}$$

Using relation (1.3), expression (3.4), and expression (3.10), we obtain for the efficiency $\rho(\gamma)$,

$$\rho(\gamma) = \frac{\lambda^{-1} E[S/N]}{E[T^*]} = \frac{\gamma}{\sum_{v=0}^{k} \frac{1}{r-v}} + O\left(\frac{1}{N}\right). \tag{3.11}$$

Rewriting expression (3.5) using the substitutions $\alpha = k+1$ and $\beta = r-k$, together with expression (3.11), we obtain the relations

$$\rho(\gamma) \approx \frac{\gamma}{\sum_{v=0}^{k} \frac{1}{r-v}}, \tag{3.12a}$$

$$\gamma = \frac{k+1}{r+1}, \tag{3.12b}$$

where k is an integer, r is a real number, and $r - k \geq 1$.

We may approximate the denominator in equation (3.12a) using the expansion for harmonic sums ($\sum_{k=1}^{n} \frac{1}{k} = \ln n + \gamma_0 + \frac{1}{2n} + O\left(\frac{1}{n^2}\right)$, where γ_0 is Euler's constant), provided $r - k$ is sufficiently large, and making the substitution $\gamma = \frac{k+1}{r+1}$,

$$\begin{aligned} \sum_{v=0}^{k} \frac{1}{r-v} &= \ln(r) - \ln(r-k-1) + O\left(\tfrac{1}{r-k}\right) \\ &= -\ln\left(1 - \tfrac{k+1}{r}\right) + O\left(\tfrac{1}{r-k}\right) = -\ln(1-\gamma) + O\left(\tfrac{1}{r-k}\right), \end{aligned} \tag{3.13}$$

where $r - k >> 1$.

It follows that

$$\rho(\gamma) = -\frac{\gamma}{\ln(1-\gamma)} + O\left(\frac{1}{r-k}\right) = \left(1 + \frac{\gamma}{2} + \frac{\gamma^2}{3} + ...\right)^{-1} + O\left(\frac{1}{r-k}\right). \tag{3.14}$$

In particular for $\gamma = .2$, we have that $\rho \approx .9$, which is in close agreement with the measured efficiency, see Fig. 3.

We remark that the parameters r and k can not grow too rapidly since we must have $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{1}{N} = \frac{\Gamma(r+1)}{\Gamma(k+1)\Gamma(r-k)} \frac{1}{N} \to 0$, as $N \to \infty$, where $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ is the normalization constant for the beta distribution.

4 Summary

This work was motivated, in part, by an attempt to explain the observation that utilization peaks in the 60-80% range for a variety of parallel architectures and allocation procedures. Our study confirms the intuition which holds that the efficiency of a parallel computer is effectively limited by the fact that the machine must spend a significant fraction of the time "draining" to accommodate a new task; more importantly, we have quantified the relationship between efficiency and the size of the smallest task in the queue. We have shown that the average maximum efficiency decreases, approximately, as the $-\gamma/\ln(1-\gamma)$ for increasing γ, where the average size of the smallest task equals γN.

The functional relationship between the smallest task size and average maximum efficiency may provide some useful insights into the observed efficiencies. For γ "small", the function $-\gamma/\ln(1-\gamma)$ is approximately linear, so the inefficiency associated with modest sized tasks is not significant. On the other hand, as γ increases, the function $-\gamma/\ln(1-\gamma)$ decreases more rapidly than linear, so that the incurred inefficiency becomes more significant. This may partially explain the observed uniform grouping of efficiencies in the 60-80% range.

The key idea in the analysis is the observation that the underlying stochastic process is a renewal process. The renewal theorem and the asymptotic covariance stationarity are used to prove convergence and to derive an explicit form for the limit. The assumption that the smallest task size, approximately, obeys a Beta law allows us to derive an analytic expression for the expected maximum efficiency.

References

[1] William Feller, *An Introduction to Probability Theory and Its Applications*, John Wiley & Sons, 1968.

[2] *Information Technology: Department of Energy Does Not Effectively manage Its Supercomputers*, GAO/RCED-98-208.

[3] *Supporting Congressional Oversight: Framework for Considering Budgetary Implications of Selected GAO Work*, GAO-01-447.

[4] Robert V. Hogg and Allen T. Craig, *Introduction to Mathematical Statistics*, Macmillan, 1978.

[5] James Patton Jones and Bill Nitzberg, *Scheduling for Parallel Computing: A Historical Perspective of Achievable Utilization*, NASA Ames Research Center, 1998.

[6] Samuel Karlin and Howard Taylor, *A First Course in Stochastic Processes*, Academic Press, 1975.

UNLIMITED RELEASE
INITIAL DISTRIBUTION:

1	MS	0310	R. W. Leland, 9220
1		0312	W. J. Camp, 9200
1		0318	P. Yarrington, 9230
1		0807	V. G. Kuhns, 9338
1		0807	J. P. Noe, 9338
1		0847	S. A. Mitchell, 9211
1		1109	S. M. Kelly, 9224
1		1109	G. F. Quinlan, 9224
1		1110	R. D. Carr, 9211
10		1110	J. M. DeLaurentis, 9214
1		1110	D. W. Doerfler, 9224
1		1110	W. E. Hart, 9211
10		1110	V. J. Leung, 9211
1		1110	D. E. Womble, 9214
1		1111	B. A. Hendrickson, 9226
1		1227	M. J. Hannah, 5001
1		9018	Central Technical Files, 8945-1
2		0899	Technical Library, 9616
1		0612	Review & Approval Desk, 9612 For DOE/OSTI

Intentionally Left Blank