# Qualitative Behavior of the EQS Parallel Processor Allocation Policy

Rajesh K. Mansharamani
Mary K. Vernon

Technical Report #1192

November 1993

# Qualitative Behavior of the EQS Parallel Processor Allocation Policy *

Rajesh K. Mansharamani and Mary K. Vernon
*mansha@cs.wisc.edu* *vernon@cs.wisc.edu*

Computer Sciences Department
University of Wisconsin
1210 West Dayton Street
Madison, WI 53706.

December 9, 1993

Several studies of multiprogrammed parallel systems have observed that dynamic equiallocation policies have high performance for a variety of specific parallel workloads. However, only very incomplete information is available about which workload parameters are key determinants of policy performance and how the mean response times of equiallocation policies behave as a function of key workload parameters. This paper addresses these issues for an idealization of the Spatial EQuiallocation policy (EQS) and a workload model that characterizes the essential features of parallel applications with respect to scheduling discipline performance. Important features of the workload model include general distribution for available job parallelism, controlled correlation between available parallelism and total job processing requirement, general distribution of processing requirement per class of jobs in the correlation model, and general nondecreasing deterministic job execution rates (i.e., speedups) that represent synchronization and communication overheads as well as load imbalance for parallel programs.

The performance of EQS is analyzed using sample path analysis to derive bounds and using highly efficient and extensively validated interpolation approximations to derive estimates for mean response time ($\overline{R}_{EQS}$). The bounds show that under exponential job processing requirements (demands) and any concave nondecreasing job execution rate function for all jobs $\overline{R}_{EQS}$ is minimum when all jobs are fully parallel and is maximum when all jobs are fully sequential. The upper bound is also shown to hold under very general workload conditions. The approximation is used to obtain the demand and parallelism parameters that are key determinants of EQS performance and to study the behavior of $\overline{R}_{EQS}$ as a function of changes in the workload. Mean response time is shown to decrease with stochastic increase in available parallelism, decrease in variability of parallelism, and increase in correlation. Under certain potentially realistic assumptions, the mean response time is also shown to be fairly insensitive to parallel program overheads.

*This research was partially supported by the National Science Foundation under grants CCR-9024144 and CDA-9024618.

# 1 Introduction

Dynamic equiallocation policies are a class of parallel processor scheduling policies that attempt to allocate processing power equally to all jobs, subject to the constraint that no job is allocated more processors than its available parallelism. Several studies [30, 11, 10, 18, 7, 27, 19, 20, 16] of multiprogrammed parallel systems have observed that dynamic equiallocation policies have high performance for a variety of workloads, where the performance metric is mean response time. However, the performance characteristics of equiallocation policies are not yet thoroughly understood. Typical questions that remain unanswered are:

- What are key workload parameters that affect scheduling policy performance? In particular, what measures of job processing requirement, job parallelism, and correlation between the two are key *determinants* of policy performance?
- What is the qualitative behavior of equiallocation policies as a function of the key workload parameters? For example, how does the policy performance respond to changes in workload parallelism or to changes in the correlation between processing requirement and parallelism?

In previous work, the principal barriers to addressing these questions have been restrictive workload assumptions and numerical solution techniques that do not readily yield insight into key parameters and policy behavior. In particular, previous simulation studies necessarily make specific assumptions about distributions of cumulative processing demand and parallelism, and previous analytic studies of equiallocation policies [10, 27] either assume that task service times are exponentially distributed, or assume that job service time is exponentially distributed. These assumptions are made for analytic tractability yet the solution techniques involve explicit enumeration of the state space, which yields no direct insight and grows exponentially in the number of processors.

In this paper recent sample path analyses and interpolation approximation techniques [1, 16] are extended to analyze an idealization of the Spatial EQuiallocation policy (EQS) (defined in Section 2) for a workload model that we believe is broadly applicable and uses only a few parameters to characterize the essential properties of parallel applications with respect to scheduling discipline performance. Significant features of the workload model include general distribution of *available* job parallelism[1], controlled correlation between available parallelism and total job processing requirement (i.e., demand), general distribution of demand per class of jobs in the correlation model, and general deterministic nondecreasing job execution rates (i.e. speedup curves) which represent synchronization and communication overheads as well as load imbalance in

[1]The available parallelism, $N$, of a job is the number of processors the system scheduler believes the job can productively use.

parallel programs. The key extensions to the workload model in this paper as compared with [16] are (1) the model of correlation and (2) several equations that constrain the system parameter space. The constraints on the parameter space aid in evaluating the qualitative behavior of EQS and in identifying stress tests for validating the interpolation approximations.

The performance of EQS is analyzed using sample path analysis to derive bounds and interpolation approximations to derive estimates for mean response time ($\overline{R}_{EQS}$). The bounds show that under exponential job processing demands and any concave nondecreasing job execution rate function for all jobs, $\overline{R}_{EQS}$ is minimum when all jobs are fully parallel and is maximum when all jobs are fully sequential. Further proofs show that the upper bound holds under more general workload conditions that include general interarrival times, general demands, general available parallelism, and general nondecreasing execution rates, with arbitrary dependencies among these workload variables.

The central interpolation approximation is derived by showing that the EQS system under constant available parallelism reduces to a symmetric queue [8] and then interpolating among the mean response time estimates at these extreme points to obtain an approximation for $\overline{R}_{EQS}$ for general workloads. The mean response time approximation is extensively validated against simulation and shown to be very accurate across the workload parameter space. The approximation yields insight into the key determinants of EQS performance and the behavior of $\overline{R}_{EQS}$ as a function of these parameters, and is highly efficient to evaluate – systems with hundreds of processors can easily be analyzed.

The main results derived in this paper, under the assumptions that jobs can dynamically adapt to their processor allocation and have a common nondecreasing execution rate function, are as follows:

- The key determinants of $\overline{R}_{EQS}$ are job arrival rate, the mean total processing requirement, $\overline{D}$, and the mean job service time on an otherwise empty system, $\overline{S}$. More specifically, $\overline{R}_{EQS}$ does not depend on demand parameters other than $\overline{D}$, such as coefficient of variation, and $\overline{S}$ contains all information about parallelism, correlation, and execution rate parameters that are needed to determine $\overline{R}_{EQS}$.
- $\overline{R}_{EQS}$ increases linearly in each of $\overline{D}$ and $\overline{S}$ (given that offered load remains fixed).
- The performance of EQS improves with (stochastic) increase in available parallelism. In particular, the performance of EQS is optimal when all jobs are fully parallel and is pessimal when all jobs are fully sequential.
- For workloads with a concave job execution rate and no correlation between total processing requirement and available parallelism, $\overline{R}_{EQS}$ decreases when the variability of available parallelism in the workload decreases. More specifically, for a fixed mean available parallelism, $\overline{R}_{EQS}$ is minimum when the coefficient of variation, $C_N$, of available parallelism is minimum and is maximum when $C_N$ is maximum.

- An increase in correlation between mean demand and available parallelism improves the performance of EQS (given a fixed overall mean demand).
- $\overline{R}_{EQS}$ remains bounded over the same range of job arrival rates for sublinear execution rates as for linear execution rates, and is relatively insensitive to parallel program overheads if the workload is not fully parallel and the execution rate is nearly linear for small processor allocations.

Although these results are derived assuming a single execution rate function for all jobs, it appears that most of the results are likely to hold more generally as long as job execution rate on $j \leq N$ processors is either uncorrelated or positively correlated with available parallelism $N$, as clarified in Section 7.

The remainder of this paper is organized as follows. The workload model, system assumptions, and constraints on model parameters are presented in Section 2, which also contains a summary of the notation used throughout the paper. Bounds on mean response time for EQS are presented in Section 3. and mean response time approximations for EQS are derived in Section 4. The qualitative behavior of EQS is studied in Section 5 for uncorrelated workloads and in Section 6 for correlated workloads. Finally Section 7 contains the conclusions of this work.

# 2 System Model

We consider an open system model with $P$ identical processors and a central job queue as shown in Figure 1. The centralized queueing model is a conceptual model; actual implementations of the scheduling policy may in general allow for distributed queue access. We assume zero job scheduling and preemption overhead, since this is an idealized system model aimed at understanding qualitative performance characteristics of the EQS scheduling policy. In practical implementations preemption overhead will exist but preemption frequency should be limited so as to guarantee that overhead is a small fraction of the application processing time.

Below we define the EQS scheduling policy (Section 2.1), the basic workload model as it was defined in [16] (Section 2.2), and extensions to that workload model to represent correlation between total job processing requirement and available parallelism (Section 2.3). Constraints that exist among workload parameters are discussed in Section 2.4, and the notation used throughout the remainder of the paper is given in Section 2.5.

## 2.1 The EQS Scheduling Policy

Dynamic equiallocation (EQ) policies allocate an equal fraction of processing power to each job in the system unless a job has smaller available parallelism than the equiallocation value, in which case each such

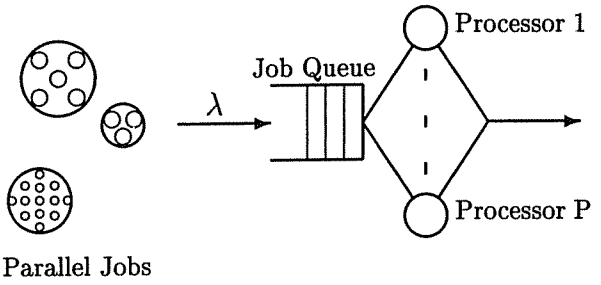



Figure 1: Open System Model

job is allocated processing power equal to its available parallelism, and the equiallocation value is recursively recomputed for the remaining jobs. For example, if there are five jobs in a 100-processor system and the available parallelism per job is (50, 25, 100, 10, 10), then the allocation of processing power is (27.5, 25, 27.5, 10, 10). The available parallelism of a job is defined to be the number of processors the system scheduler believes the job can make productive use of. Reallocation of power can occur on job arrivals, job departures, and changes in a job's available parallelism.

The Spatial EQuiallocation policy (EQS) is an EQ policy in which processing power is allocated spatially for integral allocation and temporally for fractional allocation. For example, if a job is to receive an allocation of 27.5 units of processing power, then it is allocated 27 processors and it receives an additional 0.5 units of processing power by time sharing an additional processor (i.e., the job alternately executes on 27 and 28 processors). Ignoring variations in implementation details, the EQS policy was first defined in [30].

In this paper we analyze the EQS policy and also comment on the applicability of the results for temporal and hybrid spatial/temporal equiallocation policies. In the context of the workload model defined in the next section, all EQ policies have the same performance when job execution rates are linear.

## 2.2 Basic Workload Model

The goal is to have a simple workload model that is broadly applicable, characterizes the essential features of parallel workloads with respect to scheduling disciplines, contains a small number of parameters, and is easy to analyze. To achieve broad applicability, few restrictions are made on the distribution of important system parameters, such as available job parallelism and total processing demand. To keep the parameter

space simple and to facilitate ease of analysis, a simple characterization of job execution rates and correlation between demand and parallelism is assumed.

Jobs arrive to the system according to a Poisson process with rate $\lambda$ as shown in Figure 1. All jobs are assumed to be statistically identical. Each job is characterized by the following random variables.

(1) Total service demand (execution time on one processor) $D$,

(2) Available parallelism $N \in \{1, 2, \ldots, P\}$,

(3) Execution rate function (ERF) $E : [0, P] \rightarrow [0, N]$, which is nondecreasing and has the following properties:

$$E(x) \begin{cases} \leq x, & 0 \leq x \leq N, \\ = E(N), & N < x \leq P. \end{cases}$$

The system operates as follows. Upon arrival each job joins the central job queue. At each time, $t \geq 0$, the $P$ processors are allocated to jobs present in the queue according to the EQS processor allocation policy. If $a(t)$ processors (possibly fractional) are allocated to a job at time $t$, then its demand is satisfied at rate $E(a(t))$. In other words, $E(k)$ is the *speedup* of the job if the job is allocated $k$ processors throughout its execution, and if the allocation can vary $E(x)$ is also assumed to be the *instantaneous rate* at which the job executes whenever it is allocated $x$ processors. The job leaves the system upon completion of its total demand, $D$. The available parallelism, $N$, of a job is the number of processors the system scheduler believes the job can productively use. The workload model assumes that $N$ is an upper bound on the actual number of processors, $m$, the job can productively use (i.e., by definition, $E(x) = E(N)$ for $N < x \leq P$, and if $m$ is less than $N$ then $E(j) = E(m), m < j \leq N$.)

The following is assumed about $N$ and $E$.

- $N$ has a general (bounded) distribution with mean $\overline{N}$, coefficient of variation $C_N$, and probability mass function $\underline{p} = (p_1, \ldots, p_P)$, where $p_k = \Pr[N = k]$, $k = 1, \ldots, P$.

- $E$ is derived from a *deterministic* nondecreasing function $\gamma$, such that $\gamma(x) = x$ for $0 \leq x \leq 1$, and $\gamma(x) \leq x$ for $1 < x \leq P$.

  For a job with available parallelism $N$, $E(N) = \gamma(N)$. When fewer than $N$ processors are allocated to the job, the execution rate $E$ depends on more detailed characteristics of the applications. Most results in this paper are derived assuming that the work for a job can be dynamically redistributed across the

number of processors allocated to it such that it executes as if it had available parallelism equal to the processor allocation, i.e., $E(j) = \gamma(j)$, for $1 \leq j < N$. This could be appropriate for applications based on the work queue model, or in some cases where the processes of a job are timeshared on the allocated processors. In cases where the allocated processing power, $x$, is nonintegral a linear interpolation between $\gamma(\lfloor x \rfloor)$ and $\gamma(\lceil x \rceil)$ is used to compute $E(x)$.

Note that other assumptions about job execution rate on fewer than $N$ processors are possible. For example, one might assume that the parallelism overhead is about the same on fewer processors as on $N$ processors, i.e., $E(j) = \frac{j}{N}\gamma(N)$ for $1 \leq j < N$, which could represent a system with jobs that have fixed parallelism in which overhead is primarily due to message passing software and processing load is balanced across the processors, e.g., through judicious cyclic rotation of processes. As another example, if communication overheads are fixed for a given available parallelism but the load is only balanced when $j$ evenly divides $N$, then $E(j) = \frac{1}{\lceil N/j \rceil}\gamma(N)$, for $1 \leq j < N$.

As shown by the above examples, for given assumptions about the application characteristics, the function $\gamma$ *determines* the ERF $E$. Thus $\gamma$ will be called the execution rate determinant (ERD) of the workload in the remainder of this paper. The ERD $\gamma$ is said to be *linear* if $\gamma(x) = x$, for all $0 \leq x \leq P$.

In this paper the performance properties of the EQS policy are for the most part studied under the assumption that jobs can dynamically adapt to the processor allocation such that $E(j) = \gamma(j)$, $j < N$, as described above. Some of the results can be expected to hold for other assumptions about $E(j), j < N$. We comment on this further as the results are developed and in Section 7.

The service time of a job on $N$ processors is denoted by the random variable $S = D/\gamma(N)$, with mean denoted by $\overline{S}$.

The workload model defined above contains three simplifications each of which represents a trade-off between analytic tractability and the simplicity of the parameter space on the one hand, and generality of the model on the other hand. The first is the assumption of constant available parallelism per job, the second is the assumption of a fixed execution rate, $E(k)$, whenever the job is allocated $k$ processors, and the third is the assumption of a single function $\gamma$ that determines the execution rate for all jobs. The first assumption is realistic for certain systems and/or workloads, for example, if the job is based on a work queue model and can continuously adapt to any given number of processors up to a maximum value of $N$ throughout (most of) its lifetime, or if the system scheduler assumes the job's parallelism is fixed (as in the CM-5). Similarly, the second assumption is realistic for certain cases of dynamic scheduling (i.e., when execution rates are nearly linear and/or when parallelism overheads including load imbalance are relatively evenly distributed throughout the execution of the program, on any number of processors). Furthermore,

since the purpose of the model is to analyze *scheduling policy* behavior and performance, as opposed to obtaining precise mean response times for the applications, assumptions that approximately represent key workload characteristics while keeping the model tractable and the parameter space simple, are acceptable even when they don't precisely describe the behavior of individual applications. For example, if jobs have varying available parallelism, one can view the model with constant available parallelism as capturing the contention that occurs between phases of different jobs, where a phase is a portion of the job in which available parallelism is constant. Similarly, although jobs actually have differing degrees of sublinearity, one can view the model as representing how policy generally performs as execution rates are more or less sublinear. Extensions that would further increase the applicability of the model yet preserve its tractability and parameter simplicity would be desirable, but appear to be quite difficult to obtain.

## 2.3 Correlation Model

The workload model defined so far is very similar to the workload model in [16]. In this section the model is extended to allow correlation between $D$ and $N$.

It is unknown whether or how job processing demand is correlated with parallelism in real workloads. The most general way to model correlation is to specify an arbitrary joint distribution of $D$ and $N$, but this approach can complicate both the analysis and exploration of the design space. A simpler model that still permits a wide range of correlation, can be obtained by assuming that for a job with available parallelism $N$, its *mean demand* is either independent of $N$ with probability $q$ or is linearly correlated with $N$ with probability $1-q$. Varying $q$ from 0 to 1 thus allows us to control the workload correlation in the model. Below the parameters of the correlation model are defined more precisely.

The demand of a job with available parallelism $N$ is drawn from a general distribution, $\mathcal{F}_D^u$, with mean $A$ and coefficient of variation $C_v$, and then with probability $1-q$ it is scaled by the factor $\frac{cN}{A}$, where $A$, $C_v$, and $c$ are constants independent of $N$. Thus the mean demand of the job with available parallelism $N$ is given by

$$\Delta_N = \begin{cases} A, & \text{with probability } q, \\ cN, & \text{with probability } 1-q. \end{cases}$$

Note that even for the scaled demand cases the coefficient of variation of processing requirement is equal to $C_v$.

Let $r$ denote the *correlation coefficient* of $\Delta_N$ and $N$. That is,

$$r \equiv \frac{E[\Delta_N N] - E[\Delta_N]E[N]}{\sigma_{\Delta_N}\sigma_N}, \quad \sigma_{\Delta_N}, \sigma_N \neq 0 \tag{1}$$

Define $r$ to be 0 when $\sigma_{\Delta_N} = 0$ or $\sigma_N = 0$. The following lemma shows how $A$ and $c$ are related to $\overline{D}$ (i.e., the mean demand of the workload across all jobs) and $\overline{N}$, and how $q$ is related to $r$. This lemma shows that the workload correlation is specified by the single parameter $r$.

**Lemma 2.1** *For the correlation model given by (1),*

$$A = \overline{D}, \quad c = \overline{D}/\overline{N}, \quad \textit{and} \quad q = 1 - r^2.$$

**Proof.** By definition of $\Delta_N$,

$$\Delta_N = \begin{cases} A, & \text{with probability } q, \\ cN, & \text{with probability } 1-q. \end{cases}$$

Thus, $E[\Delta_N] = \overline{D} = qA + (1-q)c\overline{N}$, for all $0 \leq q \leq 1$. Setting $q = 1$ yields $A = \overline{D}$, and setting $q = 0$ yields $c = \overline{D}/\overline{N}$.

To prove that $q = 1 - r^2$, note first that either $\sigma_{\Delta_N} = 0$ or $\sigma_N = 0$ implies that $\Delta_N = \overline{D}$ with probability 1. Thus $q = 1 - r^2$ for these cases. For $\sigma_{\Delta_N} > 0$ and $\sigma_N > 0$, we evaluate the RHS of equation (1). First note that

$$E[\Delta_N N] = qA\overline{N} + (1-q)cE[N^2].$$

Using this and $E[\Delta_N] = \overline{D}$ and further simplifying we obtain,

$$E[\Delta_N N] - E[\Delta_N]\overline{N} = (1-q)\overline{D}\,\overline{N}\,C_N^2. \tag{2}$$

Also,

$$\begin{aligned} E[\Delta_N^2] &= qA^2 + (1-q)c^2E[N^2], \\ \sigma_{\Delta_N}^2 &= E[\Delta_N^2] - E[\Delta_N]^2 = (1-q)\overline{D}^2 C_N^2. \end{aligned}$$

Substituting $\sigma_{\Delta_N} = \sqrt{1-q}\,\overline{D}\,C_N$ and the RHS of (2) in (1), yields

$$r = \frac{(1-q)\overline{D}\,\overline{N}\,C_N^2}{\sqrt{1-q}\,\overline{D}\,C_N\,\sigma_N} = \sqrt{1-q},$$

which results in $q = 1 - r^2$ as required. ∎

A consequence of this lemma is that $r = 0$ implies that $q = 1$ and thus that $D$ and $N$ are independent.

## 2.4 Parameter Constraints

The workload model defined above is not only general but is also easy to parameterize. Important generalizations in the workload model include the general distribution of available parallelism, general distribution of job demand for jobs with no correlation, general nondecreasing ERD, and controlled correlation between demand and parallelism. Varying workload parameters, such as $C_D$ and $r$, allows us to explore the design space more thoroughly than in the past. Nearly all previous performance studies of parallel processor scheduling policies have assumed specific distributions for demand and/or parallelism. Furthermore, the authors are not aware of any study that has allowed controlled correlation between demand and parallelism. (Some previous studies have considered specific extremes of our correlation model such as $r = 0$ and $r = 1$, cf. [12, 11, 32]. In i.i.d. task service time models there is implicitly a high correlation between demand and parallelism and there is no opportunity to vary demand and parallelism parameters independently.)

Workload parameters of immediate interest to us are mean and coefficient of variation in demand, i.e., $\overline{D}$ and $C_D$, mean and coefficient of variation of available parallelism, i.e., $\overline{N}$ and $C_N$, correlation coefficient $r$, execution rate determinant $\gamma$, and mean service time $\overline{S}$. These parameters must satisfy certain relationships which constrain the system design space. The parameters $\overline{D}$, $\overline{N}$, $\gamma$, and $r$ can vary freely within their feasible ranges (e.g., $0 \leq \overline{D} < \infty$, $1 \leq \overline{N} \leq P$, or $0 \leq r \leq 1$), which is why they are the free parameters of the model. Below, the constraints on the other parameters of interest, i.e., $C_D$, $C_N$, and $\overline{S}$ are identified. The constraints delineate the model parameter space which will be useful in evaluating the qualitative behavior of EQS as well as for identifying stress tests for validating mean response time approximations.

The overall coefficient of variation, $C_D$, in demand (after unconditioning on $N$) can vary freely between 0 and $\infty$ only when $D$ and $N$ are independent, i.e., $r = 0$. For $r > 0$, it can be verified using Lemma 2.1 that $C_D$ depends on $C_v$, $r$, and $C_N$ as follows:

$$C_D^2 = (1 + C_v^2)(1 + r^2 C_N^2) - 1. \tag{3}$$

Since $N$ is bounded above by $P$, it follows that $C_N$ cannot be unbounded. For a given $\overline{N}$, the following

constraints on $C_N$ are derived in Appendix A,

$$0 \leq C_N \leq \sqrt{\frac{\overline{N}(P+1)-P}{\overline{N}^2} - 1} \ . \tag{4}$$

The lower bound is attained when $N$ is constant and integer-valued for all jobs, i.e., $N = k$, where $k \in \{1, \ldots, P\}$. The upper bound is attained when $N$ has a two-point p.m.f. with nonzero mass only at 1 and P.

Appendix A also derives the following results that constrain $\overline{S}$. When $r = 0$ and $\gamma$ is concave[2], $\overline{S}$ is minimum when $C_N$ is minimum and $\overline{S}$ is maximum when $C_N$ is maximum. When $r = 1$, $\gamma$ is concave, and $N/\gamma(N)$ is concave, $\overline{S}$ is maximum when $C_N$ is minimum and $\overline{S}$ is minimum when $C_N$ is maximum. (The result for $r = 1$ holds for i.i.d. exponential task service times [21].) For concave $\gamma$ and $N/\gamma(N)$, $\overline{S}$ decreases with workload correlation $r$. (Note that $N/\gamma(N)$ is concave for the concave ERDs considered in the experiments in this paper.)

## 2.5 Notation

Table 1 summarizes the notation for the model parameters and variables. Under the implicit assumption of Poisson arrivals the following notation will be used to characterize specific workloads.

$$(\lambda,\ \mathcal{F}_N,\ \mathcal{F}_D^u,\ r,\ \gamma,\ E(j)),$$

$\lambda$ = job arrival rate

$\mathcal{F}_N$ = distribution of $N$, e.g., $N = P$, Uniform(1,P)

$\mathcal{F}_D^u$ = distribution of demand for jobs with mean demand independent of parallelism e.g., exp($\mu$)

$r$ = correlation coefficient

$\gamma$ = execution rate determinant. By default $\gamma$ is a general nondecreasing ERD. The notation $\gamma \in \mathcal{E}^c$, specifies that $\gamma$ belongs to the class of concave and nondecreasing ERDs, $\mathcal{E}^c$. To specify the linear ERD, the notation $\gamma^l$ is used.

$E(j)$ = job execution rate on $j < N$ processors , e.g., $E(j) = \gamma(j)$ in the case of jobs that can dynamically and efficiently redistribute their work.

[2] A function $f : (a, b) \to I\!R$ is concave if $f(\alpha x + (1-\alpha)y) \geq \alpha f(x) + (1-\alpha)f(y)$, for all $x, y \in (a, b)$ and $\alpha \in (0, 1)$. Conversely, $f$ is convex if $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$ [24]. Informally, a function is concave if the line joining two points on the function lies on or below the function values between the two points, and is convex if the line lies on or above the function values.

A general distribution of demand or available parallelism, general ERD, or arbitrary value of $r$ between 0 and 1, will be indicated simply by leaving the notation as $\mathcal{F}_D^u$, $\mathcal{F}_N$, $\gamma$, or $r$, respectively.

Experimental results in this paper will make use of the following bounded-geometric distribution for available job parallelism (similar to the distribution in [11, 9]):

**Definition 2.1** *$N$ has a bounded-geometric distribution with parameters $P_{max}$ and $p$ if*

$$N = \begin{cases} P, & \textit{with probability } P_{max}, \\ \min(X, P), & \textit{with probability } 1 - P_{max}, \end{cases} \qquad \textit{where } X = \textit{Geometric(p).}$$

In some experiments, three specific bounded-geometric distributions for N will be examined. These distributions are given in Table 2. Discussion of these workloads is contained in [16]. Another distribution for $N$ that will be used is the following two-point pmf:

**Definition 2.2** *$N$ has a $K_2(a, b, \alpha)$ distribution if*

$$N = \begin{cases} a, & \textit{with probability } \alpha, \\ b, & \textit{with probability } 1 - \alpha. \end{cases} \qquad 0 \leq \alpha \leq 1$$

Note that the $K_2(1, P, \alpha)$ distribution is the bounded-geometric distribution with $P_{max} = 1 - \alpha$ and $p = 1$.

The following two types of ERDs are used in the experiments:

- $\gamma(k) = k^c$, for $k = 1, 2, \ldots$, $0 \leq c \leq 1$,
- $\gamma(k) = (1 + \beta)k/(k + \beta)$, for $k = 1, 2, \ldots$, $0 \leq \beta \leq \infty$.

Both ERDs are concave and nondecreasing as shown in Figure 2. The second ERD is derived from a type of execution signature in [5].

# 3 Mean Response Time Bounds for the EQS Policy

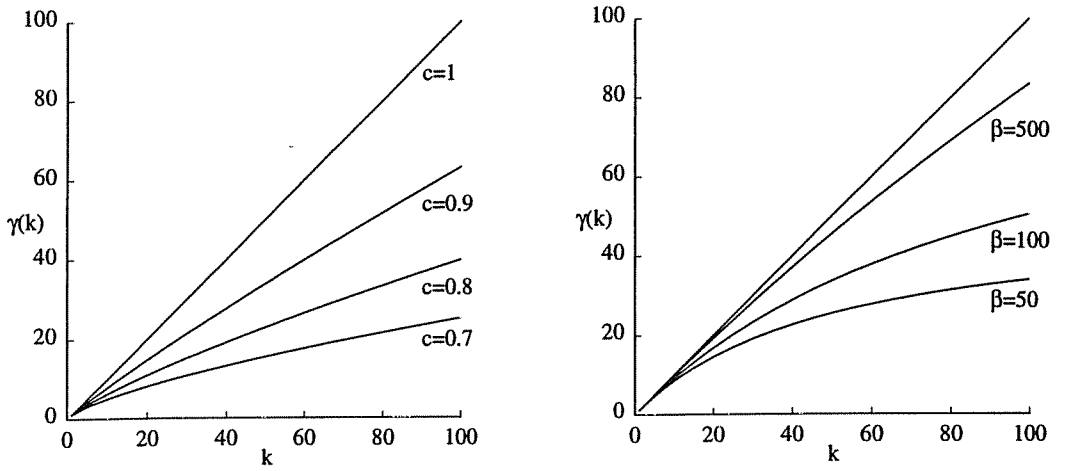
This section first derives lower and upper bounds on $\overline{R}_{EQS}$ for the workload ($\lambda$, $\mathcal{F}_N$, $\exp(1/\overline{D})$, $r = 0$, $\gamma \in \mathcal{E}^c, E(j) = \gamma(j)$). These bounds show that the mean response time is minimum when all jobs are fully parallel (i.e., $N = P$) and is maximum when all jobs are fully sequential (i.e., $N = 1$), all else being equal. Note that these bounds are derived assuming $N$ and $D$ are independent, $D$ is exponential, the workload

Table 1: Notation

| | |
|---|---|
| $P$ | Number of processors in the system |
| $\lambda$ | Arrival rate of jobs |
| $D$ | Total job demand |
| $\mathcal{F}_D^u$ | Distribution of demand for "uncorrelated" jobs |
| $\overline{D}$ | Overall mean job demand |
| $C_D$ | Overall coefficient of variation of demand |
| $\rho$ | Offered load $\lambda\overline{D}/P$ |
| $N$ | Available job parallelism |
| $\mathcal{F}_N$ | Distribution of available parallelism |
| $p_k$ | Probability$[N = k]$, $k = 1, \ldots, P$ |
| $\underline{p}$ | $(p_1, p_2, \ldots, p_P)$ |
| $\overline{N}$ | Average available parallelism |
| $C_N$ | Coefficient of variation of available parallelism |
| $r$ | Correlation between $N$ and $D$ (as defined in (1)) |
| $\gamma$ | Execution rate function (ERD) of the workload |
| $\mathcal{E}^c$ | Class of concave and nondecreasing ERDs |
| $\gamma^l$ | Linear execution rate function |
| $\overline{S}$ | Mean job service time |
| $S_n$ | Normalized mean service time $\overline{S}/\overline{D}$ |
| $\overline{R}_{EQS}$ | Mean response time of EQS |

Table 2: Three Bounded-Geometric Distributions for $N$

| Symbol | Parallelism | $P_{max}$ | $p$ | P=20 | | P=100 | |
|---|---|---|---|---|---|---|---|
| | | | | $\overline{N}$ | $C_N$ | $\overline{N}$ | $C_N$ |
| H | High | 0.9 | 1.0 | 18.10 | 0.31 | 90.10 | 0.33 |
| M | Moderate | 0.1 | 1/(0.4P) | 8.70 | 0.77 | 43.14 | 0.80 |
| L | Low | 0.1 | 0.9 | 3.00 | 1.89 | 11.00 | 2.70 |




(a) $\gamma(k) = k^c$ (b) $\gamma(k) = (1+\beta)k/(k+\beta)$

Figure 2: Two types of ERDs

ERD is concave and nondecreasing, and each job can dynamically redistribute its work across its processor allocation. The assumption of exponential job demand is probably not a serious limitation in this case, since the approximate analysis in this paper as well as the simulation experiments reported in this and previous papers indicate that $\overline{R}_{EQS}$ depends only on mean demand and not on distribution of demand.[3] We also show that the upper bound for $\overline{R}_{EQS}$ holds under more general workload assumptions, which include general job arrival times, job demands, available parallelisms, and execution rates, with arbitrary dependencies among these workload variables.

The lower and upper bounds in this section are generalizations of the bounds in [1] for the EQS policy, and are obtained as corollaries of more general bounds, which show that the performance of EQS improves with "increase" in available parallelism. For example, for workloads with the same available parallelism for all jobs $\overline{R}_{EQS}$ decreases as available parallelism increases. In [1] it was shown that the mean response time of any processor conserving policy[4] under exponential job demands and *linear* job execution rates is minimum when $N = P$ and maximum when $N = 1$. Note that the generalizations below are only with respect to the EQS policy and do not hold for all processor conserving policies.

[3]The bounds also hold for the generalized exponential distribution since it can be shown analytically that the mean response time of EQ is the same under exponential and generalized exponential demands [17].

[4]A processor conserving policy does not allocate more processors to a job than the job can productively make use of, and it does not leave a processor idle if any job can make use of that processor.

## 3.1 Lower and Upper Bounds: $\mathcal{F}_D^u = \exp,\ r = 0,\ \gamma \in \mathcal{E}^c$

It will be shown that under the workload $(\lambda, \mathcal{F}_N, \exp(1/\overline{D}), r = 0, \gamma \in \mathcal{E}^c, E(j) = \gamma(j))$, the performance of EQS is optimal when all jobs are fully parallel and is pessimal when all jobs are fully sequential. The bounds follow as an immediate consequence of the following theorem.

**Theorem 3.1** *If $\ell$ and $m$ are constants such that $\ell \leq m$, then under the workload assumptions* $(\lambda, \cdot, \exp(1/\overline{D}), r = 0, \gamma \in \mathcal{E}^c, E(j) = \gamma(j))$,

$$\overline{R}_{EQS}(m \leq N \leq P) \ \leq \ \overline{R}_{EQS}(1 \leq N \leq \ell).$$

**Proof.** See Appendix B. ∎

The intuition for Theorem 3.1 is that whenever the number of jobs in each system is equal, the total job completion rate in the system with higher available parallelism is greater than or equal to the job completion rate in the other system.

Setting $\ell = m = P$ in Theorem 3.1 yields the following lower bound on $\overline{R}_{EQS}$:

**Corollary 3.1** *Under the workload assumptions* $(\lambda, \cdot, \exp(1/\overline{D}), r = 0, \gamma \in \mathcal{E}^c, E(j) = \gamma(j))$,

$$\overline{R}_{EQS}(\mathcal{F}_N) \ \geq \ \overline{R}_{EQS}(N = P).$$

In [1] a corresponding bound was given for all processor conserving policies assuming exponential demands and the linear ERF. As in [1], a tighter lower bound can be obtained when $N \not\equiv P$ by using the fact that $\overline{R}_{EQS} \geq \overline{S}$. This yields the following bound, which henceforth will be referred to as the $N = P$ lower bound:

$$\overline{R}_{EQS}(\mathcal{F}_N) \ \geq \ \max\left\{\overline{S},\ \overline{R}_{EQS}(N = P)\right\}, \quad \text{under } (\lambda, \cdot, \exp(1/\overline{D}), r = 0, \gamma \in \mathcal{E}^c, E(j) = \gamma(j)). \tag{5}$$

Setting $\ell = m = 1$ in Theorem 3.1 yields the following bound on $\overline{R}_{EQS}$, which henceforth will be referred to as the $N = 1$ upper bound:

**Corollary 3.2** *Under the workload assumptions* $(\lambda, \cdot, \exp(1/\overline{D}), r = 0, \gamma \in \mathcal{E}^c, E(j) = \gamma(j))$,

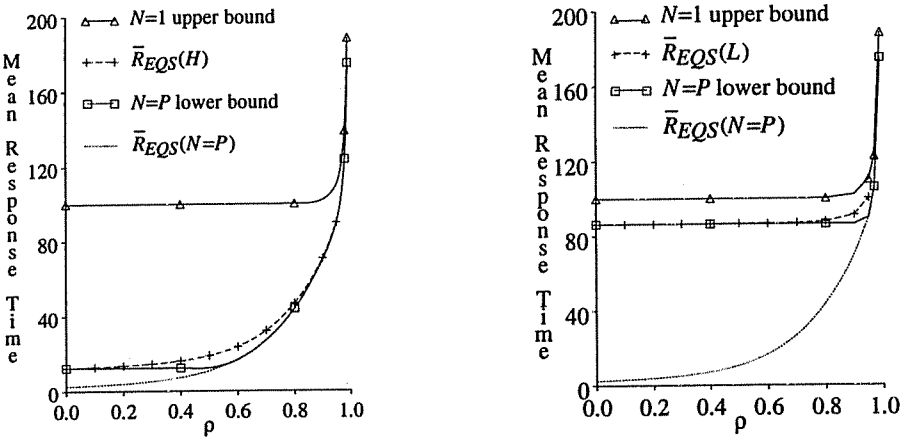
$$\overline{R}_{EQS}(\mathcal{F}_N) \ \leq \ \overline{R}_{EQS}(N = 1). \tag{6}$$

For the linear ERF, the bounds in (5) and (6) can be shown to reduce to the following [1]:

$$\max\left(\overline{D}E[1/N], \frac{1}{1-\rho}\frac{\overline{D}}{P}\right) \ \leq \ \overline{R}_{EQS}(\lambda, \mathcal{F}_N, \exp(1/\overline{D}), r = 0, \gamma^l, E(j) = \gamma(j)) \ \leq \overline{R}_{M/M/P}.$$

## 3.2 Experimental Evaluation of the $N = P$ and $N = 1$ Bounds

For the workload $(\lambda, \cdot, \exp(1/\overline{D}), r = 0, \gamma \in \mathcal{E}^c, E(j) = \gamma(j))$ and for many distributions of $N$, the mean system response time will lie closer to the $N = P$ lower bound than the $N = 1$ upper bound, primarily because the lower bound is the maximum of the mean service time and the mean reponse time when $N = P$. For example, for a given distribution of $N$, $N \not\equiv 1$, the $N = P$ bound is exact when $\rho \to 0$ but this is not true of the $N = 1$ bound. This point is further illustrated by comparing simulation estimates of $\overline{R}_{EQS}$ against the bounds for a 100 processor system, the H and L distributions of $N$ given in Table 2, exponential job demand $D$ with mean $\overline{D} = P = 100$, and ERD $\gamma(k) = k^{0.8}$, $k = 1, 2, \ldots, P$.[5] As seen from Figure 3(a) and (b) the $N = 1$ upper bound is rather loose for workloads with high average available parallelism, but is much tighter when average available parallelism is low. Conversely, $\overline{R}_{EQS}(N = P)$ is tighter for the H workload, but looser for the L workload. Taking the maximum of $\overline{S}$ and $\overline{R}_{EQS}(N = P)$, i.e., the $N = P$ bound, results in a tight bound for both high and low average available parallelism.




(a) $H$ workload (b) $L$ workload

Figure 3: Tightness of $N = 1$ and $N = P$ bounds for $\overline{R}_{EQS}$: $D = \exp$, $r = 0$

ERD $\gamma(k) = k^{0.8}$
$\overline{D} = P = 100$

[5] All simulation experiments in this paper have 95% confidence intervals with less than 10% half-widths, and in almost all cases the half-widths are less than 5of the estimate. The confidence intervals were generated using the regenerative method whenever feasible and otherwise the method of batch means.

## 3.3 Upper Bound under General Workloads

The $N = 1$ upper bound will now be derived under more general workload assumptions, i.e., general arrivals, general available parallelism, general demands, and general nondecreasing execution rates, with arbitrary dependencies among these workload variables. The upper bound follows as a direct consequence of the following theorem.

**Theorem 3.2** *Let $\Gamma_I$ be a system with the EQS policy and primitive workload variables $\{(A_i, D_i, N_i \geq k, E_i), i = 1, 2, \ldots\}$, where $A_i$ is job $i$'s arrival time, $D_i$ its total demand, $N_i$ its available parallelism, and $E_i$ its execution rate function. Let these primitive variables have arbitrary marginals (given that $N_i \geq k$, and the other variables make sense, e.g., $D_i \geq 0$) with arbitrary dependencies among them. Let $\Gamma_{II}$ be a system with the EQS policy and the same workload as $\Gamma_I$ except that $N_i = k$ for all $i = 1, 2, \ldots$. Then*

$$\overline{R}_{\Gamma_I} \leq \overline{R}_{\Gamma_{II}}, \quad k = 1, 2, \ldots, P.$$

**Proof.** See Appendix B. ∎

The intuition for Theorem 3.2 is that system $\Gamma_I$ allocates at least as much processing power to each unfinished job as $\Gamma_{II}$ does.

Setting $k = 1$ in Theorem 3.2 yields the following result.

**Corollary 3.3** *Let $\Gamma_I$ be a system with the EQS policy and primitive workload variables $\{(A_i, D_i, N_i, E_i), i = 1, 2, \ldots\}$. Let these primitive variables have arbitrary marginals with arbitrary dependencies among them. Let $\Gamma_{II}$ be a system with the EQS policy and the same workload as $\Gamma_I$ except that $N_i = 1$ for all $i = 1, 2, \ldots$. Then*

$$\overline{R}_{\Gamma_I} \leq \overline{R}_{\Gamma_{II}}, \quad k = 1, 2, \ldots, P.$$

*More specifically,*

$$\overline{R}_{EQS}(\mathcal{F}_N) \quad \leq \quad \overline{R}_{EQS}(N = 1), \quad \text{under } (\lambda, \cdot, \mathcal{F}_D^u, r, \gamma, E(j)).$$

Considering only constant values of $N$ in Theorem 3.2 yields the following corollary.

**Corollary 3.4** *Consider a system with the EQS policy with general $\{(A_i, D_i, E_i), i = 1, 2, \ldots\}$ (i.e., these primitive variables have arbitrary marginals with arbitrary dependencies among them). Then*

$$\overline{R}_{EQS}(N = P) \leq \ldots \leq \overline{R}_{EQS}(N = k) \leq \overline{R}_{EQS}(N = k - 1) \leq \ldots \leq \overline{R}_{EQS}(N = 1), \quad k = P, \ldots, 2.$$

*where $N = k$ denotes $N_i = k$, for all $i = 1, 2, \ldots$.*

This corollary shows that for workloads with constant available parallelism the performance of EQS improves as available parallelism increases. This result is generalized in Section 5.2.

# 4 Mean Response Time Estimates for EQS

In this section mean response time estimates for EQS are derived under the workload assumptions ($\lambda$, $\mathcal{F}_N$, $\mathcal{F}_D^u$, $r$, $\gamma$, $E(j) = \gamma(j)$). In Section 4.1 the EQS system under constant available parallelism is reduced to a known queueing system to obtain an exact expression for $\overline{R}_{EQS}$. In Section 4.2 an approximate estimate of $\overline{R}_{EQS}$ is derived for general distribution of available parallelism and arbitrary correlation between job processing requirement and parallelism. The approximation for $\overline{R}_{EQS}$ is shown to be an interpolation among the exact results for constant values of available parallelism, which yields additional insight into the behavior of the policy.

Before proceeding, the known queueing system that will be used in deriving both the exact and approximate mean response times is the *symmetric queue*, which is defined as follows [8].

**Definition 4.1** *A queue is a* symmetric *queue if it operates in the following manner:*

*(i) The service requirement of a job is a random variable whose distribution may depend upon the class of the job.*

*(ii) A total service effort is supplied at the rate $\phi(j)$, where $j$ is the total number of jobs in the queue.*

*(iii) A proportion $\alpha(l, j)$ of this effort is directed to the job in position $l \in \{1, 2, \ldots, j\}$; when this job leaves the queue, jobs in positions $l+1, l+2, \ldots, j$ move to positions $l, l+1, \ldots, j-1$, respectively.*

*(iv) When a job arrives at the queue it moves into position $l \in \{1, \ldots, j+1\}$ with probability $\alpha(l, j+1)$; jobs previously in positions $l, l+1, \ldots, j$ move to positions $l+1, l+2, \ldots, j+1$, respectively, where $j$ is the total number of jobs in the queue as seen by the arrival.*

*Note that $\phi(j) > 0$ if $j > 0$, and $\sum_{l=1}^{j} \alpha(l, j) = 1$.*

## 4.1 Reductions under Constant Available Parallelism

For a system with constant available parallelism across all jobs, i.e., $N = k$, where $1 \leq k \leq P$, the EQS system reduces to a symmetric queue, which leads to an exact solution for $\overline{R}_{EQS}$ for any general distribution

for demand and any general non-decreasing ERD $\gamma$, as given by the following theorem.

**Theorem 4.1** *For the system* $(EQS,\ \lambda,\ N=k,\ \mathcal{F}_D^u,\ r=0,\ \gamma,\ E(j)=\gamma(j))$, $1 \le k \le P$,

$$\overline{R}_{EQS}(N=k,r=0) = \frac{b}{\lambda}\left\{\sum_{i=1}^{P} \frac{(P\rho)^i}{(i-1)!\,E(k)^{\min(i,m)} \prod_{j=m+1}^{i} E(P/j)} + \frac{(P\rho)^P}{P!\,E(k)^m \prod_{j=m+1}^{P} E(P/j)} \frac{\rho}{1-\rho}\left(\frac{1}{1-\rho}+P\right)\right\}, \tag{7}$$

*where* $m = \lfloor P/k \rfloor$, $\rho = \lambda\overline{D}/P$, *and*

$$b = \left[1 + \sum_{i=1}^{P} \frac{(P\rho)^i}{i!\,E(k)^{\min(i,m)} \prod_{j=m+1}^{i} E(P/j)} + \frac{(P\rho)^P}{P!\,E(k)^m \prod_{j=m+1}^{P} E(P/j)} \frac{\rho}{1-\rho}\right]^{-1}.$$

**Proof.** Let $\Gamma_k$ denote the system $(EQS,\ \lambda, N=k, \mathcal{F}_D^u, r=0, \gamma, E(j)=\gamma(j))$, $1 \le k \le P$. We first show that $\Gamma_k$ is a symmetric queue, and then derive the mean response time of $\Gamma_k$ for a general distribution of job demand.

System $\Gamma_k$ satisfies conditions (i) through (iv) for a symmetric queue in Definition 4.1. The total service effort supplied when there are $j$ jobs in $\Gamma_k$ is $\phi(j) = j \cdot \min(E(k), E(P/j))$ since $E$ is non-decreasing. (Note that for $j \ge P$, $\phi(j) = P$, because $E(x) = \gamma(x) = x$ for $0 \le x \le 1$.) From the definition of the EQ policy, $\alpha(l,j) = 1/j$, $l = 1,\ldots,j$, since each job in $\Gamma_k$ gets an equal fraction of processing power. Note that this does not hold if available parallelism is not constant across all jobs.

The mean response time of a job in system $\Gamma_k$ can be derived from Theorems 3.8 and 3.10 of [8], which give the following steady state probability of $i$ jobs in the queue for the stationary symmetric queue with arbitrary distribution of job service time:

$$\pi_i = \frac{ba^i}{\prod_{l=1}^{i}\phi(l)}, \quad i = 0,1,2,\ldots \tag{8}$$

where

$$a = \lambda\overline{D}, \quad \text{and} \quad b = \left[\sum_{i=0}^{\infty} \frac{a^i}{\prod_{l=1}^{i}\phi(l)}\right]^{-1}.$$

Substituting $\phi(l) = l\min(E(k), E(P/l))$ into equation (8) and Little's Result

$$\overline{R}_{EQS}(N=k) = \frac{\sum_{i=1}^{\infty} i\pi_i}{\lambda},$$

yields the mean response time for $\Gamma_k$ as given in (7), where $m = \lfloor P/k \rfloor$. That is, $m$ is the maximum number of jobs that can execute simultaneously without contention for processors. The derivation of (7) uses the fact that if there are $P$ or more jobs in the system then the total service effort of the symmetric queue is $P$. ∎

**Remark:** The symmetric queue reduction and equation (8) actually hold for any nondecreasing ERF $E$, and the mean response time formula (7) holds for any nondecreasing $E$ such that $E(x) = x$ for $0 \leq x \leq 1$ (i.e., (7) and (8) hold for more general $E$ than $E(j) = \gamma(j)$).

An important observation from equation (7) is that $\overline{R}_{EQS}(N = k)$ *depends only on the mean job demand and not on higher moments of job demand.* This property is a generalization of the corresponding property for Processor Sharing (PS) systems. Note that when $P = 1$ or when $N = 1$ the EQS policy is identical to the PS policy, and thus $\overline{R}_{EQS}(N = 1)$ equals $\overline{R}_{M/G/P\ PS} = \overline{R}_{M/M/P}$. For the case of the linear ERF it was shown in [16] that when $P \bmod k = 0$, $\overline{R}_{EQS}(N = k) = \overline{R}_{M/G/c\ PS} = \overline{R}_{M/M/c}$, where $c = P/k$. The same does not hold for nonlinear ERFs, however, if $k > 1$.

## 4.2 Approximation for $\overline{R}_{EQS}$: general $N$ and $r$

Given equation (7) for $\overline{R}_{EQS}(N = k, r = 0)$, the following interpolation on the pmf of $N$ might be used to approximate the mean response time for general distributions of $N$ (cf. [16]):

$$\overline{R}_{EQS}(\mathcal{F}_N, r = 0) \approx \hat{R}^p_{EQ} = \sum_{k=1}^{P} p_k \overline{R}_{EQS}(N = k, r = 0), \quad \text{under } (\lambda, \cdot, \mathcal{F}^u_D, r = 0, \gamma, E(j) = \gamma(j)). \tag{9}$$

However, it is not immediately obvious whether or how one might use the equations for constant available parallelism to obtain mean response time estimates for correlated workloads. We thus take an alternate approach to compute the mean response time for workloads with arbitrary correlation $0 \leq r \leq 1$, general distributions of demand and available parallelism, and general nondecreasing ERD $\gamma$. The result of this alternate approach will provide a justification for the above interpolation, an interpolation on the estimates in equation (7) for $\overline{R}_{EQS}(r = 1)$, and an interpolation on the parameter $r$ between the approximations for $r = 0$ and $r = 1$.

The approximate mean response time for the EQS policy for the workload $(\lambda, \mathcal{F}_N, \mathcal{F}^u_D, r, \gamma, E(j) = \gamma(j))$ is derived by (1) classifying jobs according to their available parallelism, (2) computing the mean response time for each class of jobs by approximating the *average interference* from other classes of jobs, and (3)

computing the overall mean response time as a weighted sum of the approximate mean response times per class. The particular approximate representation of average interference by other job classes yields a system for each class that reduces to a symmetric queue, from which the class mean response time is computed.

Let a job with available parallelism $k$ belong to class $C_k$, for $k = 1, \ldots, P$. Let $\overline{R}_{EQS,C_k}$ denote the mean response time of class $C_k$ under the workload $(\lambda, \mathcal{F}_N, \mathcal{F}_D^u, r, \gamma, E(j) = \gamma(j))$. Clearly,

$$\overline{R}_{EQS} = \sum_{k=1}^{P} p_k \overline{R}_{EQS,C_k}. \tag{10}$$

For each class $C_k$, the approximate processor contention from classes other than $C_k$ is modeled by assuming each such class has available parallelism $k$, but retains its total service demands as before. More precisely, let $\overline{R}_{EQS,C_k}$ be approximately equal to the mean response time of class $C_k$ in a system $\Gamma_k$ which is like the original system except that a class $C_j$ job in $\Gamma_k$ has demand $D_j$ and available parallelism $k$, where $\overline{D}_j = q\overline{D} + (1-q)cj$, $q = 1 - r^2$ and $c = \overline{D}/\overline{N}$, as per the correlation model in Section 2.3. The instantaneous load, including available parallelism and execution rate, of class $C_j$ jobs is not accurately modeled by assuming that class $C_j$ jobs have parallelism $k$. However, the offered load by class $C_j$ jobs is accurately modeled since the arrival rate and distribution of processing requirement of the class are as in the actual system. Since efficiency is underestimated when parallelism is overestimated and vice versa, the interference experienced by $C_k$ may be somewhat underestimated for lower $k$ and somewhat overestimated for higher $k$. Note that these errors will tend to cancel each other in the calculation of overall mean response time, $\overline{R}_{EQS}$.

The approximation for $\overline{R}_{EQS,C_k}$ is derived by solving for the mean response time of class $k$ in system $\Gamma_k$. Note that in system $\Gamma_k$ there are $P$ job classes, $C_1, \ldots, C_P$, where $C_j$ has available parallelism $k$ and demand $D_j$. Since all jobs have the same available parallelism and since the definition of a symmetric queue permits multiple job classes with different service demand distributions (see Definition 4.1), the system again reduces to a symmetric queue. As before, the total service effort with $j$ jobs in the queue is $\phi(j) = j \cdot \min(E(k), E(P/j))$, $j \geq 0$, and the fraction of effort for job $i$ is $\alpha(i,j) = 1/j$, for $i = 1, \ldots, j$. Furthermore, equation (8) holds also for the case of multiple classes with different distribution of demand (see Theorem 3.8 and 3.10 of [8]), and thus $\overline{R}_{\Gamma_k} = \overline{R}_{EQS}(N = k, r = 0)$.

The mean response time of class $k$ in $\Gamma_k$, $\overline{R}_{\Gamma_k,C_k}$, is obtained from part (ii) of Theorem 3.10 of Kelly [8]. Using the notation in this paper, this theorem can be stated as follows.

Given there are $Q$ customers in the symmetric queue, the classes of the customers are independent and the probability the customer in a given position is of class $C_k$ is $\frac{\lambda_k \overline{D}_k}{\lambda \overline{D}}$, where $\lambda_k$ is the arrival rate of class $C_k$ and $\overline{D}_k$ is the mean demand of class $C_k$.

Thus given $Q$ jobs in system $\Gamma_k$, the of number, $Q_k$, of jobs of class $C_k$ is binomially distributed with parameters $Q$ and $u_k$ where $u_k := \lambda_k \overline{D}_k/(\lambda \overline{D}) = p_k \overline{D}_k/\overline{D}$. Therefore $\overline{Q}_k = \overline{Q} u_k$ and Little's law yields the mean response time of class $C_k$ in $\Gamma_k$ as

$$\overline{R}_{\Gamma_k, C_k} = \frac{\overline{Q}_k}{\lambda_k} = \frac{\overline{Q} u_k}{\lambda p_k} = \frac{\overline{D}_k}{\overline{D}} \overline{R}_{\Gamma_k}.$$

Since $\overline{R}_{\Gamma_k} = \overline{R}_{EQS}(N = k, r = 0)$,

$$\overline{R}_{EQS, C_k} \approx \overline{R}_{\Gamma_k, C_k} = \frac{\overline{D}_k}{\overline{D}} \overline{R}_{EQS}(N = k, r = 0).$$

Substituting the above in (10), yields under the workload assumptions $(\lambda, \cdot, \mathcal{F}_D^u, \cdot, \gamma, E(j) = \gamma(j))$ that

$$\overline{R}_{EQS}(\mathcal{F}_N, r) \approx \sum_{k=1}^{P} p_k' \overline{R}_{EQS}(N = k, r = 0), \quad p_k' = p_k \frac{\overline{D}_k}{\overline{D}} = p_k \left(1 - r^2 + r^2 \frac{k}{\overline{N}}\right), \tag{11}$$

where the expression for $p_k'$ was derived as per the correlation model described in Section 2.3.

Further insight can be obtained from equation (11) by making the following observations. When $r = 0$, $p_k' = p_k$, for $k = 1, 2, \ldots, P$, and approximation (11) reduces to the interpolation approximation in (9). On the other hand when $r = 1$ it follows from (11) that under the assumptions $(\lambda, \cdot, \mathcal{F}_D^u, \cdot, \gamma, E(j) = \gamma(j))$,

$$\overline{R}_{EQS}(\mathcal{F}_N, r = 1) \approx \sum_{k=1}^{P} p_k \frac{k}{\overline{N}} \overline{R}_{EQS}(N = k, r = 0).$$

Finally, for $r$ between 0 and 1 and $(\lambda, \cdot, \mathcal{F}_D^u, \cdot, \gamma, E(j) = \gamma(j))$,

$$\begin{aligned}
\overline{R}_{EQS}(\mathcal{F}_N, r) &\approx \sum_{k=1}^{P} p_k \left\{1 - r^2 + r^2 \frac{k}{\overline{N}}\right\} \overline{R}_{EQS}(N = k, r = 0) \\
&= (1 - r^2) \sum_{k=1}^{P} p_k \overline{R}_{EQS}(N = k, r = 0) + r^2 \sum_{k=1}^{P} p_k \frac{k}{\overline{N}} \overline{R}_{EQS}(N = k, r = 0) \\
&\approx (1 - r^2) \overline{R}_{EQS}(\mathcal{F}_N, r = 0) + r^2 \overline{R}_{EQS}(\mathcal{F}_N, r = 1),
\end{aligned} \tag{12}$$

which can be interpreted as an interpolation on $r$.

## 4.3 Validations

We validated approximation (11) against simulation estimates of $\overline{R}_{EQS}$ for several distributions of $D$ and $N$, linear as well as sublinear $\gamma$, and uncorrelated as well as correlated workloads. As noted in Section 3 almost all simulation estimates have 95% confidence intervals with less than 5% half-widths (less than 10% in all cases), and whenever possible the regenerative method was used to obtain the confidence intervals. The batch means method was used if obtaining the regenerative cycles was too time consuming (e.g., for workloads with low average available parallelism).

The parameter values used in the validation experiments are as follows:

(i) $P$: 20,100

(ii) $\mathcal{F}_D^u$: deterministic, two-stage Erlang, exponential, two-stage hyperexponential, Gamma.

The approximations for $\overline{R}_{EQS}$ suggest that mean response time is insensitive to the distribution of demand provided the mean demand is fixed. Within the limits of statistical error, the simulation results also show this to be the case. For all validation experiments the mean job demand, $\overline{D}$, is equal to $P$. Thus offered load $\rho \equiv \lambda\overline{D}/P = \lambda$.

(iii) $\rho$: 0.1 to 0.9

(iv) $\mathcal{F}_N$: bounded-geometric, uniform

Table 3 lists the parameter settings for all distributions of $N$ considered in the validations. The parameter settings for the bounded geometric distributions are arranged in groups of three, and within each group in order of decreasing $\overline{N}$. It can be shown that for a fixed value of $\overline{N}$, the bounded-geometric distribution with lowest $C_N$ has $P_{max} = 0.0$ and the bounded-geometric distribution with highest $C_N$ has $p = 1$ [15]. Thus, the first group of three are low $C_N$ workloads, the last group are high $C_N$ workloads, and the middle group are workloads with intermediate $C_N$.

Note that for a fixed $\overline{N}$ the $K_2(1, P, \frac{P-\overline{N}}{P-1})$ distribution, or equivalently the bounded-geometric distribution with $p = 1$ and $P_{max} = \frac{\overline{N}-1}{P-1}$, distribution has highest $C_N$ over all distributions of $N$, and the constant $N$ distribution has lowest $C_N$ for integer-valued $\overline{N}$. For constant N, approximation (11) is exact, which is why this case is not included in the validation experiments.

| Distribution | Parameter Settings | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Bounded- | $P_{max}$ | 0.0 | 0.0 | 0.0 | 0.1 | 0.1 | 0.1 | 0.9 | 0.5 | 0.1 |
| Geometric | $p$ | 0.005 | 1/(0.5P) | 1/(0.1P) | 0.01 | 1/(0.4P) | 0.9 | 1 | 1 | 1 |
| Uniform | (P/2,P), (1,P), (1,P/2) | | | | | | | | | |

Table 3: Validation Workloads for $N$: P=20,100

(v) $\gamma$: $\gamma(k) = k$, $\gamma(k) = k^c$ $0 < c < 1$, $\gamma(k) = (1+\beta)k/(k+\beta)$ $0 < \beta < \infty$, $k = 1, 2, \ldots, P$

In the absence of extensive data for real workloads the models are validated against three types of ERDs. The first is simply the linear ERD. The second is a simple algebraic choice of a concave sublinear ERD, whereas the third is derived from a type of *execution signature* given in [5]. For the ERD $\gamma(k) = k^c$, the validations include c = 0.7, 0.8, and 0.9, which are plotted for $P = 100$ in Figure 2(a). At $c = 0.7$, $\gamma(20) = 8.14$ and $\gamma(100) = 25.12$ which are quite low compared to their linear counterparts of 20 and 100, respectively. The value of $c = 0.7$ therefore tests the accuracy of the models for highly sublinear ERDs. For the ERD $\gamma(k) = (1+\beta)k/(k+\beta)$ the following values of $\beta$ are used in the validations: $\beta = 20, 50, 100$ for $P = 20$, and $\beta = 50, 100, 500$ for $P = 100$. The smaller values of $\beta$ are used as stress tests whereas the larger values are used to evaluate the accuracy of the models when the ERD is close to linear, but not exactly linear. Figure 2(b) plots these ERDs for $P = 100$.

(vi) $r$: 0, 0.5, 1.

The total number of data points in our validations was 2561[6]. Figure 4 summarizes the validations by means of a histogram of relative error. There was no appreciable difference in the histograms for $r = 0$, $r = 0.5$, and $r = 1$ and as a result the histograms are not presented separately for these cases. Note that approximation (11) is extremely accurate since all data points in Figure 4 are with 15% of simulation estimates. The largest errors (10 – 15%) were observed for correlated workloads with low to moderate $\overline{N}$ (i..e, $0.1P \leq \overline{N} \leq 0.5P$), high $C_N$, and moderate to high execution rate sublinearity.

We also ran a few experiments for a nonconcave ERD, specifically, $\gamma(k) = P/\lceil P/k \rceil$, $k = 1, \ldots, P$, which is a step function with perfect speedup when $k$ evenly divides $P$. We found approximation (11) to have a similar level of accuracy for this ERD as well. The results derived from approximation (11) in Sections 5 and 6 hold within the accuracy of the model, which is expected to be high for concave ERDs and is likely to be high for nonconcave ERDs as well.

[6]Many of the simulations were run on the Condor distributed system [2].

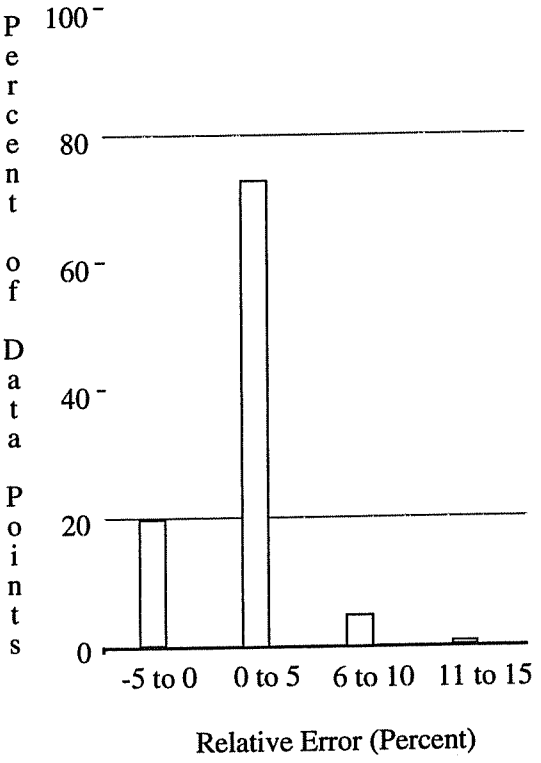



Figure 4: Summary of Validations: 2561 data points

r=0, 0.5, 1,
P=20,100

# 5 Behavior of $\overline{R}_{EQS}$ with respect to Key Parameters: $r = 0$

For the sake of simplicity we focus on uncorrelated workloads in this section and then generalize the results for correlated workloads in Section 6. The goal is to identify which workload parameters are key determinants of policy performance and to determine the functional dependence of EQS on key parameters. Section 5.1 points out that it is quite straightforward to determine the key parameter of job demand and discusses how $\overline{R}_{EQS}$ varies with the key demand parameter. In Section 5.2 the behavior of $\overline{R}_{EQS}$ is examined as a function of several different parameters of available parallelism and the key parallelism parameter is identified. Section 5.3 presents insights into the behavior of EQS as a function of sublinearity in the workload ERD and Section 5.4 presents a summary of the results for uncorrelated workloads.

## 5.1 $\overline{R}_{EQS}$ as a function of job demand

Approximation (9) is a weighted sum of the mean response times of EQS under constant available parallelism. We noted in Section 4.1 that the mean response time of EQS under constant available parallelism depends only on the first moment of demand and not higher moments. Thus the weighted sum of $\overline{R}_{EQS}(N = k, r = 0)$ in approximation (11) depends only on $\overline{D}$ and not on higher moments of demand. This means that $\overline{R}_{EQS}$ is independent of $C_D$ for all distributions of job demand and all distributions of job parallelism. Simulation studies have also verified this result for specific demand distributions [11, 9].

The dependence of $\overline{R}_{EQS}$ on $\overline{D}$ can be readily obtained from approximation (11). For a given $\rho$, $\overline{R}_{EQS}$ is directly proportional to $\overline{D}$. This is because $\overline{R}_{EQS}(N = k, r = 0)$ given by equation (7) is directly proportional to $\overline{D}$ for a given $\rho$ (because $\lambda = \rho P/\overline{D}$). These results were also shown for the linear ERF in [16].

## 5.2 $\overline{R}_{EQS}$ as a function of job parallelism

To understand the behavior of $\overline{R}_{EQS}$ as a function of available parallelism, $N$, we need to know which parameters of $N$ are principal determinants of $\overline{R}_{EQS}$. Natural candidates are $\overline{N}$ and $C_N$. Another measure of $N$ that could be a key determinant when $r = 0$ is $E[1/\gamma(N)]$ since the mean service time is $\overline{D}E[1/\gamma(N)]$.

A possible approach to determining if a given parameter of $N$, say $E[f(N)]$, uniquely determines $\overline{R}_{EQS}$ is to test whether $\overline{R}_{EQS}$ remains unchanged across *all* distributions of $N$ that have a given $E[f(N)]$, for each possible value of $E[f(N)]$. In other words, if $\min\{\overline{R}_{EQS}(\mathcal{F}_N) : E[f(N)] = x\} = \max\{\overline{R}_{EQS}(\mathcal{F}_N) :$

| | |
|---|---|
| **minimize** $\sum_{k=1}^{P} R_k\, p_k$ | **maximize** $\sum_{k=1}^{P} R_k\, p_k$ |
| subject to: | subject to: |
| (i) $\underline{p} \geq 0$ | (i) $\underline{p} \geq 0$ |
| (ii) $\sum_{k=1}^{P} p_k = 1$ | (ii) $\sum_{k=1}^{P} p_k = 1$ |
| (iii) $\sum_{k=1}^{P} f(k)\, p_k = E[f(N)] = a$ | (iii) $\sum_{k=1}^{P} f(k)\, p_k = E[f(N)] = a$ |

Figure 5: Linear Programs for Min and Max of $\overline{R}_{EQS}$

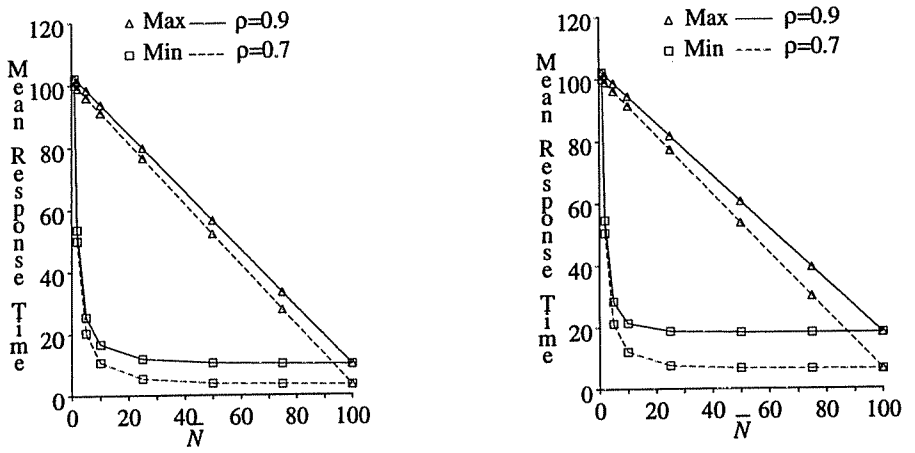
$E[f(N)] = x\}$ for all feasible $x$, then $E[f(N)]$ is a parameter that uniquely determines $\overline{R}_{EQS}$. To use this approach we must obtain the minimum and maximum of $\overline{R}_{EQS}(\mathcal{F}_N, r = 0)$ over all distributions of $N$ for each value of $E[f(N)]$. A key observation about approximation (9) is that $\overline{R}_{EQS}(N = k, r = 0)$ does not depend on the pmf, $\underline{p}$ (see equation (7)). Thus, for given fixed values for $\lambda$, $\overline{D}$, and $\gamma$, $\hat{R}^{\underline{p}}_{EQ}$ in approximation (9) can be viewed as a linear combination of the $p_k$'s and we can use linear programming [4] to obtain the minimum and maximum mean response times. The generic form of the linear program is given in Figure 5, where $R_k$ denotes $\overline{R}_{EQS}(N = k)$.

Below, the linear programs of Figure 5 are used to determine whether $\overline{N}$, $C_N$, or $E[1/\gamma(N)]$ uniquely determine $\overline{R}_{EQS}$.

### 5.2.1 $\overline{R}_{EQS}$ versus $\overline{N}$

Setting $f(N) = N$ in Figure 5 we obtain linear programs that minimize and maximize the estimator $\hat{R}^{\underline{p}}_{EQ}$ for a given $\overline{N}$, $\lambda$, $\overline{D}$, and $\gamma$ over all possible pmfs $\underline{p}$ such that the expected value of $N$ is $\overline{N}$. For $P = 100$ and specific values of $\lambda$, $\overline{D}$, and $\gamma$, the linear programs were solved for $\overline{N} = 1, 2, 5, 10, 25, 50, 75$, and 100 using the Simplex Method of linear programming [4]. Figures 6(a) and (b) plot the envelopes obtained from the minimum and maximum values of $\hat{R}^{\underline{p}}_{EQ}$ versus $\overline{N}$ for $\overline{D} = P$, two different ERDs, and two different values of $\rho = \lambda\overline{D}/P = \lambda$. The minimum value of $\hat{R}^{\underline{p}}_{EQ}$ for a given $\overline{N}$ was obtained for a distribution of $N$ with low $C_N$ (typically $K_2(\lfloor\overline{N}\rfloor, \lceil\overline{N}\rceil, \lceil\overline{N}\rceil - \overline{N})$). The maximum value was obtained for the $K_2(1, P, \frac{P-\overline{N}}{P-1})$ distribution of $N$.

Figure 6 clearly shows that for uncorrelated workloads, $\overline{N}$ alone does not adequately capture the influence




(a) Linear ERF

(b) $\gamma(k) = 101k/(k+100)$

Figure 6: Envelopes of $\overline{R}_{EQS}$ versus $\overline{N}$

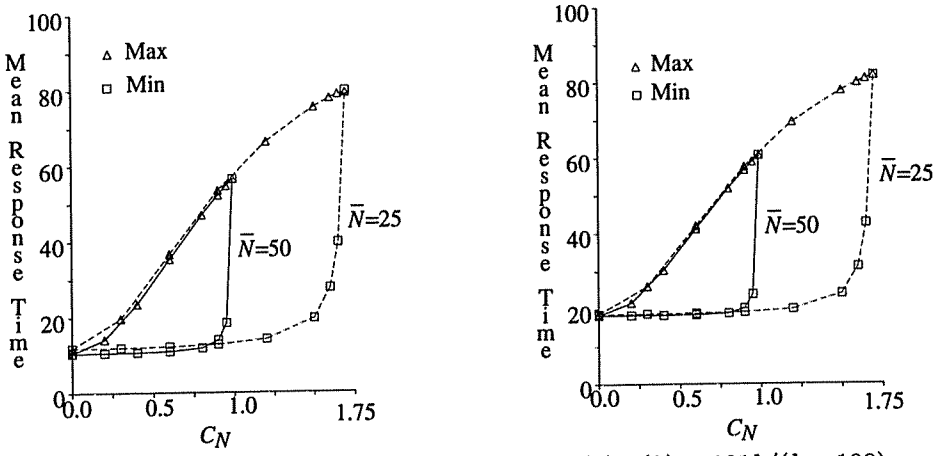
$\overline{D} = P = 100$

of $\mathcal{F}_N$ on the behavior of $\overline{R}_{EQS}$. For example, at $\rho = 0.9$ and $\overline{N} = 25$ in Figure 6(a), $\hat{R}^p_{EQ}$ ranges from a minimum of 11.88 when $N = 25$, to a maximum of 79.83 when $N$ has the $K_2(1, 100, \frac{75}{99})$ distribution.

Although $\overline{N}$ does not in general uniquely determine $\overline{R}_{EQS}$, the envelopes in Figure 6 provide useful bounds on $\overline{R}_{EQS}$ and lead to two useful observations. First, for each of the given parameter settings and across all distributions of $N$, $\hat{R}^p_{EQ}$ is maximum when $N = 1$ and minimum when $N = P$. This is consistent with the bounds for $\overline{R}_{EQS}$ that were derived in Section 3, where the upper bound was derived for general demands and the lower bound was derived for exponential demands. For the envelopes in Figure 6 job demand has a general distribution. Second, the plots for the maximum value of $\hat{R}^p_{EQ}$ versus $\overline{N}$ in Figures 6 reveal an interesting property of the $K_2(1, P, \frac{P-\overline{N}}{P-1})$ distribution of $N$ – namely, that the response time for this distribution decreases linearly as the mean available parallelism increases (i.e. as the fraction of fully parallel jobs increases). This observation is only for a specific distribution of $N$; results below show that the result also holds for other distributions of $N$.

### 5.2.2 $\overline{R}_{EQS}$ versus $C_N$

We next examine whether $C_N$ and $\overline{N}$ together uniquely determine $\overline{R}_{EQS}$ for a given $\lambda$, $\overline{D}$, and $\gamma$. Figures 7(a) and (b) plot envelopes of $\hat{R}^p_{EQ}$ versus $C_N$ for two values of $\overline{N}$ and two different ERDs, for systems with




(a) Linear ERF

(b) $\gamma(k) = 101k/(k+100)$

Figure 7: Envelopes of $\overline{R}_{EQS}$ versus $C_N$

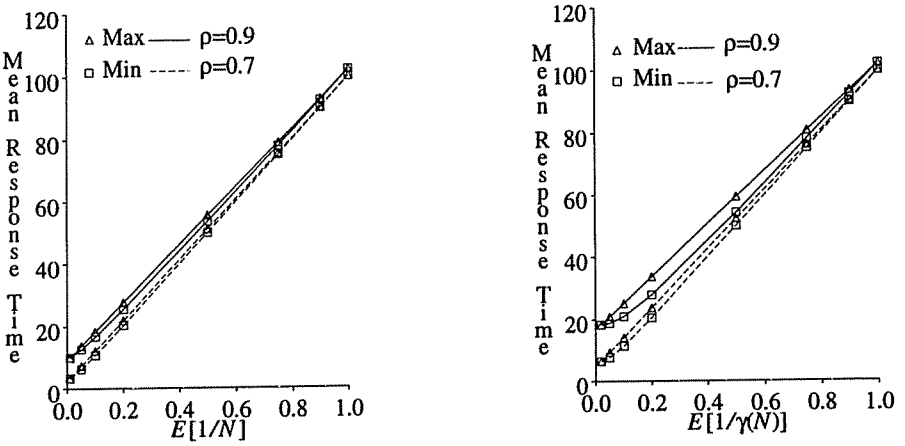
$\overline{D} = P = 100$,
$\rho = 0.9$

$P = 100$, $\overline{D} = P$, and $\rho = 0.9$. As before, the envelopes were obtained using linear programming. Note that for each value of $\overline{N}$, the range of $C_N$ is constrained as specified in (4). As was the case in Figure 6 the envelopes of $\hat{R}^p_{EQ}$ versus $C_N$ are very similar for both types of ERDs. The envelopes also have similar shape and orientation for both values of $\overline{N}$ and for different values of $\rho$ (not shown). However, unlike the envelopes for $\overline{N}$ there is no particular pattern to the distributions of $N$ that yield the minimum or maximum value of $\hat{R}^p_{EQ}$ at different values of $C_N$.

The plots in Figure 7 show that $C_N$ and $\overline{N}$ together are not sufficient to determine the behavior of $\overline{R}_{EQS}$ as a function of workload parallelism. However, the envelopes show that, for the parameter values examined, $\hat{R}^p_{EQ}$ is minimum when $C_N$ is minimum and is maximum when $C_N$ is maximum, and that the range of possible mean response times is low for low $C_N$.

### 5.2.3 $\overline{R}_{EQS}$ versus $E[1/\gamma(N)]$

The linear programs in Figure 5 with $f(N) = 1/\gamma(N)$ are used next to obtain envelopes of $\hat{R}^p_{EQ}$ versus $E[1/\gamma(N)]$ for given values of $\lambda$ and $\overline{D}$, and a given function $\gamma$. Note that $E[1/\gamma(N)]$ can vary from $1/\gamma(P)$ (when $N = P$) to 1 (when $N = 1$). Figure 8(a) and (b) plots the envelopes for two different ERDs and two different values for $\rho$, given that $P = 100$ and $\overline{D} = P$.




(a) Linear ERF

(b) $\gamma(k) = 101k/(k+100)$

Figure 8: Envelopes of $\overline{R}_{EQS}$ vs $E[1/\gamma(N)]$

$\overline{D} = P = 100$

For the linear ERD in Figure 8(a) we observe that there is very little spread between the minimum and maximum values of $\overline{R}_{EQS}$ for a fixed value of $E[1/N]$. For sublinear ERDs as in Figure 8(b) the spread is somewhat larger but is still quite small. These results provide evidence that $E[1/\gamma(N)]$ almost uniquely determines $\overline{R}_{EQS}(r=0)$ and is thus the key parameter of available parallelism for uncorrelated workloads.

If in general $\overline{R}_{EQS}$ increases (nearly) linearly as a function of $E[1/\gamma(N)] = \overline{S}/\overline{D}$ holds in general for workloads $(\lambda, \mathcal{F}_N, \mathcal{F}_D^u, r=0,\ \gamma,\ E(j)=\gamma(j))$ then the following simple interpolation approximation for $\overline{R}_{EQS}$ should be accurate for these workloads:

$$\overline{R}_{EQS}(\mathcal{F}_N, r=0) \approx \left(\frac{E\left[\frac{1}{\gamma(N)}\right] - \frac{1}{\gamma(P)}}{1 - \frac{1}{\gamma(P)}}\right) \overline{R}_{EQS}(N=1, r=0) + \left(\frac{1 - E\left[\frac{1}{\gamma(N)}\right]}{1 - \frac{1}{\gamma(P)}}\right) \overline{R}_{EQS}(N=P, r=0). \quad (13)$$

We validated the above approximation against simulation estimates of $\overline{R}_{EQS}$, for the concave ERDs shown in Figure 2 as well as the linear ERD. For all experiments we found the interpolation on $E[1/\gamma(N)]$ to be extremely accurate. (More than 95% of the validations had relative errors between -5% and 15%, and the maximum relative error was about 30%.) Thus, within the accuracy of the above interpolation approximation, $E[1/\gamma(N)]$ is a key determinant of EQS mean response time.

The qualitative behavior of $\overline{R}_{EQS}$ versus the key parameter $E[1/\gamma(N)] = \overline{S}/\overline{D} \equiv S_n$ yields the following insights into the performance of EQS as a function of $N$ when $r = 0$.

(1) Since $\overline{R}_{EQS}$ increases nearly linearly in $E[1/\gamma(N)]$, a workload with a lower value of $E[1/\gamma(N)]$ has a smaller mean response time.

(2) Since $1/\gamma(P) \leq E[1/\gamma(N)] \leq 1/\gamma(1)$ it follows that $\overline{R}_{EQS}(N = P) \leq \overline{R}_{EQS}(\mathcal{F}_N) \leq \overline{R}_{EQS}(N = 1)$. That is, $N = P$ *is optimal and* $N = 1$ *is pessimal* for the workload ($\lambda$, $\mathcal{F}_N$, $\mathcal{F}_D$, $r = 0$, $\gamma$), which is generalization of the mean response time bounds of Section 3.

(3) We next consider distributions of $N$ between the two extremes of $N = 1$ and $N = P$. In particular, consider two distributions $\mathcal{F}_{N_1}$ and $\mathcal{F}_{N_2}$ such that $N_2 \leq_{st} N_1$ (i.e., $P[N_2 \leq n] \geq P[N_1 \leq n], 1 \leq n \leq P$). Under this condition it is shown in [25] that $E[f(N_1)] \leq E[f(N_2)]$ for any nonincreasing function $f$. Setting $f = 1/\gamma$ it follows that $E[1/\gamma(N_1)] \leq E[1/\gamma(N_2)]$ and thus $\overline{R}_{EQS}(\mathcal{F}_{N_1}) \leq \overline{R}_{EQS}(\mathcal{F}_{N_2})$. Thus, *a stochastic increase in available parallelism leads to a decrease in mean response time* for the EQS policy. Hence the EQS policy does not discourage and may encourage users to increase program parallelism (as long as the ERD is nondecreasing).

(4) A stochastic increase in parallelism also increases the mean parallelism. What if the mean parallelism is held constant but the variability in parallelism changes? More precisely, consider $\overline{N}_1 = \overline{N}_2$ and $N_1 \leq_v N_2$, which means that $E[f(N_1)] \leq E[f(N_2)]$ for all convex functions $f$ [25]. If $\gamma$ is concave then $1/\gamma$ is convex and it follows that $\overline{R}_{EQS}(\mathcal{F}_{N_1}) \leq \overline{R}_{EQS}(\mathcal{F}_{N_2})$ if $N_1 \leq_v N_2$. Thus the mean response time of EQS decreases with a decrease in variability of $N$ if $\gamma$ is concave and $\overline{N}$ remains fixed. Note that for the bounded distributions considered in this paper the highest variability in $N$ for a fixed $\overline{N}$ is when $N$ has a $K_2(1, P, .)$ distribution and the least variability in $N$ is when $N$ is constant. Also recall (from 4) that for a given $\overline{N}$, the $K_2(1, P, \cdot)$ has the highest $C_N$ and the constant distribution has the lowest $C_N$. Thus for a given $\overline{N}$, $\overline{R}_{EQS}$ is maximum when $C_N$ is highest and is minimum when $C_N$ is lowest. This results generalizes the corresponding results for specific workloads in Figures 6 and 7.

Note that results (1) and (2) above contrast with studies of fork-join queueing systems that have shown parallelism to be harmful for other scheduling disciplines, particularly at high loads [14, 26, 3].

## 5.3 $\overline{R}_{EQS}$ as function of ERD Sublinearity

Intuition suggests that system performance should improve with a decrease in synchronization and communication overheads. This is also shown analytically, since an increase in $\gamma$ decreases $E[1/\gamma(N)]$ which in turn decreases $\overline{R}_{EQS}$. This section addresses the following further questions about the behavior of $\overline{R}_{EQS}$ as a function of ERD sublinearity:

- How stable is the system as a function of ERD sublinearity?
- Precisely how does $\overline{R}_{EQS}$ behave as the ERD sublinearity increases for given functional forms of $\gamma$?
- How does the behavior of $\overline{R}_{EQS}$ change with the functional form of $\gamma$?

(i) **System stability versus degree of sublinearity**

Under the assumption of negligible preemption and scheduling overhead, and the fairly weak assumption that $E(x) = x$ for $0 \leq x \leq c$ where $c$ is a constant greater than zero, the stability condition for a system with the EQS scheduling policy for any ERD $\gamma$ is the same as for the linear ERD, that is $\lambda < P/\overline{D}$ or $\rho < 1$. This stability property of the EQS policy is not shared by several other processor scheduling policies for parallel systems. For example, consider the FCFS policy with a workload having $N = P$ and ERD $\gamma$. This system behaves like an $M/G/1$ system with mean service time $\overline{x} = \overline{D}/\gamma(P)$ and thus the stability condition is $\lambda < \gamma(P)/\overline{D}$. That is, the upper bound on arrival rate for stable operation of the FCFS system depends on $\gamma(P)$ and degrades as the sublinearity of $\gamma$ increases. If $\gamma(P) = P/2$ then the upper bound on $\lambda$ is half that of the EQS system.

(ii) **Sensitivity of $\overline{R}_{EQS}$ to ERD sublinearity and type**

The sensitivity of $\overline{R}_{EQS}$ to the degree of ERD sublinearity is examined for the following two specific ERD functions. In each function the degree of sublinearity is controlled by a single parameter.

(a) $\gamma(k) = k^c$, $k = 1, 2, \ldots, N$, $0 \leq c \leq 1$. When $c = 0$ we obtain the constant ERD $\gamma(k) = 1$, and when $c = 1$ we obtain the linear ERD $\gamma(k) = k$. Thus we control the degree of sublinearity by varying $c$ from 0 to 1. This ERD is plotted in Figure 2(a) for different values of $c$.

(b) $\gamma(k) = \dfrac{(1+\beta)k}{k+\beta}$, $k = 1, 2, \ldots, N$, $0 \leq \beta < \infty$. When $\beta = 0$, we obtain $\gamma(k) = 1$, and when $\beta = \infty$ we obtain the linear ERD. Thus we control the degree of sublinearity by varying $\beta$ from 0 to $\infty$. This ERD is plotted in Figure 2(b) for several values of $\beta$.

Figure 9 plots $\overline{R}_{EQS}$, estimated from approximation (9), versus ERD sublinearity, $\gamma(P)/P$, for each of the above ERDs, the H and L workloads of Table 2, and two different values of $\rho$. For each curve, $P = 100$ and $\overline{D} = P$. For both ERD types we observe that sublinearity has a fairly small impact on overall mean response time for the L workload, since a significant fraction of the jobs are sequential and the service time for sequential jobs is independent of ERD sublinearity. On the other hand, for the H workload the ERD sublinearity has a significant impact on mean response time. Furthermore, the precise behavior of $\overline{R}_{EQS}$ as a function of ERD sublinearity differs for the two different types of ERDs, and the difference increases as $\rho$ increases.

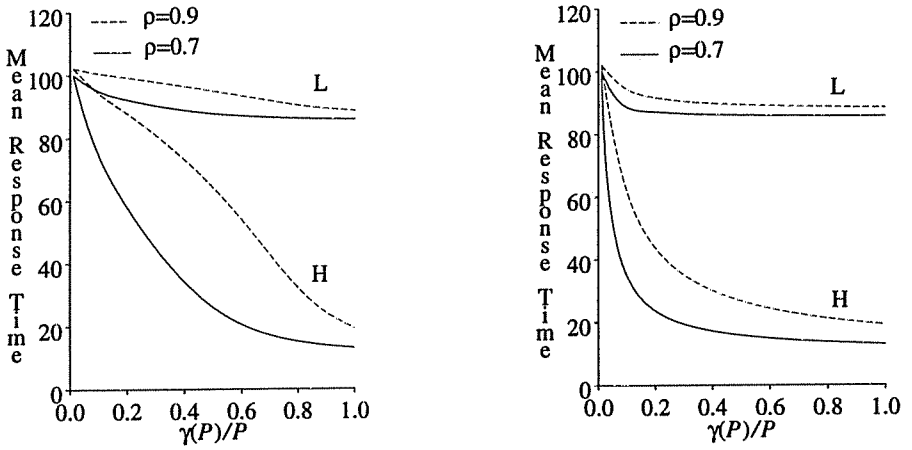
For the ERD $\gamma(k) = (1+\beta)k/(k+\beta)$ the mean response time of EQS decreases dramatically when $\gamma(P)$ increases from 1 to $0.5P$, and then decreases only very gradually as $\gamma(P)$ varies from $0.5P$ to $P$.[7] For the ERD $\gamma(k) = k^c$, as $\rho$ increases the mean response time decreases more gradually for $\gamma(P)$ in the range of 1 to $0.5P$. $\overline{R}_{EQS}$ behaves differently (at moderate to high load) under these two ERD types because of the different behavior of these ERDs when processor allocation is low, say in the region of 0-0.20P (see Figure 2). At higher load, jobs are allocated fewer processors, and for any fixed average allocation of processors $k < P$, say k=10, the curves (in Figure 2) that correspond to particular increases in $\gamma(P)$ more rapidly approach rate $k$ for the ERD controlled by $\beta$ than for the ERD controlled by $c$. (Note that for the $H$ workload and $\rho > 0.7$, the mean number of jobs in the system as obtained from the interpolation on $\underline{p}$ is greater than 10 under all ERDs.)

One conclusion of this sensitivity study is that, as intuition might suggest, the EQS policy provides better performance to workloads that have the initial part of their ERDs (say the first 10-20%) close to linear. Another conclusion is that if the ERD has this property, $\overline{R}_{EQS}$ is relatively insensitive to ERD sublinearity in the range of $\gamma(P) > 0.5P$, particularly if the workload is not fully parallel and $\rho$ is less than 0.9.

## 5.4 Summary of insights for $r = 0$

In this section the following properties of the EQS policy for uncorrelated workloads with $E = \gamma$ were derived from the interpolation approximation in (9).

[7] Note that the degree of insensitivity of $\overline{R}_{EQS}$ to ERD sublinearity when $\gamma(P) > 0.5P$ is partially due to the fact that the H workload contains a fraction of sequential jobs, whose service times dominate in the overall mean service time estimate. For a fully parallel workload, the decrease in $\overline{R}_{EQS}$ as $\gamma(P)$ increases is still gradual for $\gamma(P) > 0.5$, but has somewhat more negative slope than the H workload.




(a) $\gamma(k) = k^c$ (b) $\gamma(k) = (1+\beta)k/(k+\beta)$

Figure 9: $\overline{R}_{EQS}$ vs ERD sublinearity

$$\overline{D} = P = 100$$

(i) $\overline{D}$ *and* $S_n \equiv E[1/\gamma(N)]$ *as well as* $\lambda$ *and* $P$*, are the key determinants of* $\overline{R}_{EQS}$.

(ii) $\overline{R}_{EQS}$ *increases linearly with* $\overline{D}$ *for a given value of* $\rho$ *and is insensitive to higher moments of* $D$ *(e.g.,* $C_D$*).*

(iii) *Parallelism Considered Useful:* $\overline{R}_{EQS}$ decreases with a stochastic increase in available parallelism. In particular, $N = P$ is optimal and $N = 1$ is pessimal for the EQS policy.

(iv) For a concave ERD $\gamma$ $\overline{R}_{EQS}$ *decreases with a decrease in the variability of available parallelism.*

(v) *Graceful degradation with ERD sublinearity:* In the absence of preemption and scheduling overhead, the stability condition for the EQS system is the same for sublinear ERDs as for the linear ERD (i.e., $\lambda < P/\overline{D}$). Furthermore, if the workload ERD is close to linear when processor allocation is $0 - 20\%$ of P, $\overline{R}_{EQS}$ is relatively insensitive to ERD sublinearity at higher processor allocations, given that the applications have at least 50-60% efficiency on P processors.

# 6 Behavior of $\overline{R}_{EQS}$ for Correlated Workloads

In Section 5 the behavior of EQS was studied using the following approximation for uncorrelated workloads:

$$\overline{R}_{EQS}(\mathcal{F}_N, r=0) \approx \sum_{k=1}^{P} p_k \overline{R}_{EQS}(N=k, r=0), \quad \text{under } (\lambda, \cdot, \mathcal{F}_D^u, \cdot, \gamma, E(j)=\gamma(j)). \tag{9}$$

To study the behavior of $\overline{R}_{EQS}$ for correlated workloads, we make a key observation about the following approximation for $\overline{R}_{EQS}$ under general workload conditions, which was derived in Section4.2:

$$\overline{R}_{EQS}(\mathcal{F}_N, r) \approx \sum_{k=1}^{P} p'_k \overline{R}_{EQS}(N = k, r = 0), \quad p'_k = p_k \frac{\overline{D}_k}{\overline{D}} = p_k \left(1 - r^2 + r^2 \frac{k}{\overline{N}}\right). \tag{11}$$

Comparing (11) with (9) shows that $\overline{R}_{EQS}(\mathcal{F}_N, r)$ is obtained by replacing $p_k$ in (9) by $p'_k$. We note that $p'_k \geq 0$ and $\sum_{k=1}^{P} p'_k = 1$. Hence $\underline{p}' = (p'_1, \ldots, p'_P)$ is a pmf for a random variable $N' \in \{1, \ldots, P\}$. This implies that if we use the random variable $N'$ instead of $N$ in approximation (9) we will obtain an estimate for $\overline{R}_{EQS}(\mathcal{F}_N, r)$. Thus we can view the behavior of $\overline{R}_{EQS}$ under a correlated workload as equivalent to the behavior of $\overline{R}_{EQS}$ under an uncorrelated workload with a different distribution of available parallelism. Formally,

$$\overline{R}_{EQS}(\mathcal{F}_N, r) \approx \overline{R}_{EQS}(\mathcal{F}_{N'}, r = 0), \quad \text{under } (\lambda, \cdot, \mathcal{F}_D^u, \cdot, \gamma, E(j) = \gamma(j)). \tag{14}$$

## 6.1 $\overline{R}_{EQS}$ as a function of job demand and parallelism

When $r > 0$, as in the case of $r = 0$, $\overline{D}$ is the only determinant of $\overline{R}_{EQS}$ with respect to job demand. This is true because approximation (11) is a weighted sum of the mean response times of EQS under constant available parallelism, which depend only on the first moment of demand and not on higher moments. (Note that the weights $p'_k$ do not depend on $\overline{D}$ since the ratio of $\overline{D}_k/\overline{D}$ is independent of $\overline{D}$, for $k = 1, 2, \ldots, P$.)

Regarding the key determinant of $\overline{R}_{EQS}$ with respect to the distribution of available parallelism for correlated workloads, in Section 5 we showed that $E[1/\gamma(N)]$ is the key determinant of $\overline{R}_{EQS}(\mathcal{F}_N, r = 0)$ (given that $\lambda$, $\overline{D}$, and $\gamma$ are fixed). This, together with approximation (14), implies that $E[1/\gamma(N')]$ is the key determinant for $\overline{R}_{EQS}(r)$. Simplifying $E[1/\gamma(N')]$ we get

$$\begin{aligned} E\left[\frac{1}{\gamma(N')}\right] &= \sum_{k=1}^{P} p'_k \frac{1}{\gamma(k)} \\ &= \sum_{k=1}^{P} p_k \frac{\overline{D}_k}{\overline{D}} \frac{1}{\gamma(k)} \\ &= \frac{1}{\overline{D}} \overline{S} \equiv S_n. \end{aligned}$$

Thus $S_n$ *is the key parameter for job parallelism, workload correlation, and job execution rate determinant.* Moreover, the result that $\overline{R}_{EQS}$ increases (nearly) linearly as a function of $S_n$, as per Figure 8, holds for

correlated workloads since it holds for all distributions of $N$ in uncorrelated workloads. We can show that under nondecreasing $\gamma$, $S_n(r)$ is minimum when $N = P$ and maximum when $N = 1$, and a stochastic increase in $N$ causes $S_n$ to decrease. Thus for correlated workloads the $N = P$ workload has optimal performance and the $N = 1$ workload has pessimal performance. Unlike the case of $r = 0$ for a fixed $\overline{N}$, $S_n$ does not necessarily decrease with decrease in variability of $N$. For example, when $r = 1$ it follows from Theorem A.2 in Appendix A that for concave $\gamma$ and concave $N/\gamma(N)$, $S_n$ is minimum when variability in $N$ is maximum and is maximum when variability in $N$ is minimum. Thus, since property (v) in Section 5.4 is expected to hold generally for uncorrelated workloads, all of the properties of $\overline{R}_{EQS}$ summarized in that section apply to workloads with $r > 0$, except property (iv).

## 6.2 $\overline{R}_{EQS}$ as a function of $r$

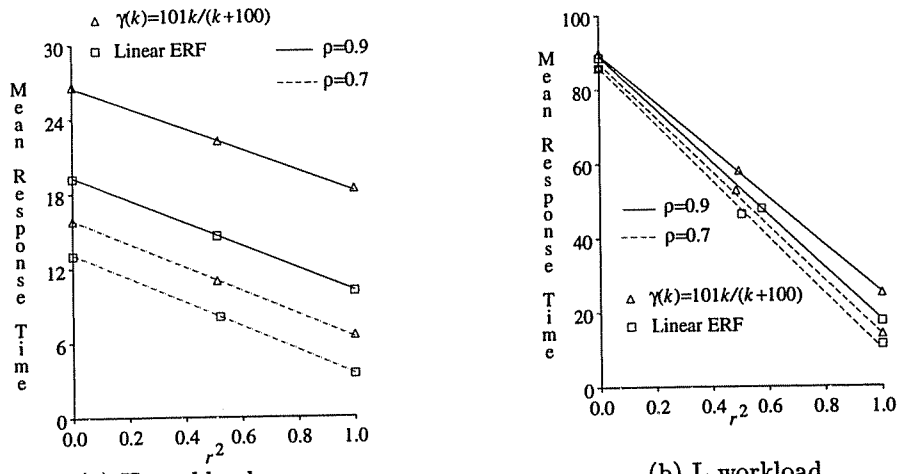
We now study the behavior of $\overline{R}_{EQS}$ when workload correlation increases. Recall from (14) that the behavior of EQS under correlated workloads and a distribution of available parallelism $\mathcal{F}_N$ is the same as the behavior of EQS under no correlation and a distribution of available parallelism $\mathcal{F}_{N'}$. The pmf of $N'$, $\underline{p}'$, is related to the pmf of $N$, $\underline{p}$, as follows:

$$p'_k = p_k \frac{\overline{D}_k}{\overline{D}} = p_k \left(1 - r^2 + r^2 \frac{k}{\overline{N}}\right).$$

Thus, $p'_k < p_k$ for $k < \overline{N}$ and $p'_k > p_k$ for $k > \overline{N}$. As a result the random variable $N'$ has stochastically higher available parallelism than $N$ (i.e., $N' \geq_{st} N$). The intuitive reason for the increase in effective available parallelism is that under correlated workloads, jobs that have small demands and exit the system quickly have on average smaller parallelism and leave behind jobs that have larger available parallelisms on average.

As seen in Section 6.1 a stochastic increase in parallelism causes $\overline{R}_{EQS}(r = 0)$ to decrease and hence $\overline{R}_{EQS}(r)$ decreases with correlation, under the given model of workload correlation and given that $\overline{D}$ remains unchanged. The intuition for this result is that as $r$ increases, larger demand jobs have larger available parallelisms, and this causes them to complete faster than if they had lower parallelisms as in uncorrelated workloads. (Consider for example the case where a sequential job in an uncorrelated workload runs on one processor but the remaining processors are idle.)

Concerning the quantitative behavior of $\overline{R}_{EQS}$ as a function of $r$, Figure 10 depicts $\overline{R}_{EQS}$ (as obtained




(a) H workload

(b) L workload

Figure 10: $\hat{R}^{Sn}_{EQS}$ versus $r$

$\overline{D} = P = 100$

from approximation (11) versus $r^2$ for the H and L workloads, two types of ERDs (one linear and one sublinear), and two values of $\rho$. The quadratic dependence of $\overline{R}_{EQS}$ on $r$ (under the given workload model), which is shown in the interpolation on $r$ (12), is evident in the figure. $\overline{R}_{EQS}$ decreases more sharply for the L workload than for the H workload, but for both workloads there is a significant decrease in mean response time as $r$ increases. The results show that EQS is a high performance policy under highly correlated workloads. Note that again that this property is not shared by all scheduling disciplines.

# 7 Conclusions

In this paper the performance of the idealized Spatial EQuiallocation policy, EQS, was analyzed under a workload model that includes general distribution of total job processing requirement (demand), general distribution of available parallelism, a general nondecreasing execution rate determinant (ERD) for all jobs, and controlled correlation between mean demand and available parallelism. First, sample path analysis was used to derive mean response time bounds for EQS. These bounds show that for exponential demands that are uncorrelated with available parallelism and the same concave ERD for all jobs EQS has optimal performance when all jobs are fully parallel, and under general demand, available parallelism, execution rate, and correlation, EQS has pessimal performance when all jobs are fully sequential. Second, approximate analysis was used to solve for $\overline{R}_{EQS}$ under the general assumptions of the workload model. This analysis

was made possible by reducing the EQS system under the assumption of constant available parallelism to a known queueing system, namely Kelly's symmetric queue. Extensive validations against simulation show that the mean response time approximation is extremely accurate; all 2561 estimates validated were within 15% of simulation estimates, and about 95% of the cases had under 5% relative error.

The approximate analysis yielded the insights that, within the accuracy of the model, (1) mean total processing requirement $\overline{D}$, and normalized mean service time, $S_n := \overline{S}/\overline{D}$, are the key determinants of $\overline{R}_{EQS}$; $\overline{R}_{EQS}$ increases linearly in each of these determinants, (2) $\overline{R}_{EQS}$ decreases with a stochastic increase in available parallelism, in particular, it is optimal when all jobs are fully parallel and pessimal when all jobs are fully sequential, (3) for uncorrelated workloads and a concave workload ERD $\overline{R}_{EQS}$ decreases with decrease in variability of available parallelism, (4) $\overline{R}_{EQS}$ decreases with increase in workload correlation (for fixed $\overline{D}$), and (5) in the absence of preemption and scheduling overhead the EQS system has the same stability condition for sublinear ERDs as it does for the linear ERD, and its mean response time is relatively insensitive to parallel program overheads if the workload is not fully parallel and the ERD is nearly linear for small processor allocations.

Although the above results were derived assuming that all jobs have the same execution rate function, $\gamma$, careful thought reveals that the results are likely to hold more generally as long as job execution rate on $j \leq N$ processors is uncorrelated with available parallelism $N$. The results except for the third are also likely to hold when the execution rate on $j \leq N$ processors is positively correlated with available parallelism. Thus, the key properties of the system that lead to the nice performance behavior are (a) equiallocation of processing power, (b) jobs can dynamically redistribute their work among their allocated processors, and (c) jobs with available parallelism $n > j$ generally execute at least as efficiently on $j$ processors as jobs with available parallelism $j$. The results explain why several previous studies of systems that satisfy assumptions (a) and (b) have observed high performance. If property (c) does not hold for a given equiallocation system (e.g., an EQ policy under workloads where $\dot{E}(j) = (j/N)\gamma(N)$) then some of the results should continue to hold and other results do not hold. For example, insensitivity of mean response time to coefficient of variation in demand, $C_D$, should continue to hold, but the result that mean response time decreases with "increase" in available parallelism will not necessarily hold, and system performance can be expected to be more sensitive to ERF sublinearity. Thus, for high-performance multiprogrammed parallel systems, the development of architectural and software support that allows jobs to dynamically and efficiently redistribute

their work across their processor allocation is highly desirable.

## Acknowledgements

The authors acknowledge Vikram Adve, Asit Dan and David Nicol for comments and discussions that improved the quality of this work.

## References

[1] R. Agrawal, R. Mansharamani, and M. Vernon. Response Time Bounds for Parallel Processor Allocation Policies. Technical Report # 1152, Computer Sciences Department, University of Wisconsin-Madison, June 1993.

[2] A. Bricker, M. Litzkow, and M. Livny. Condor Technical Summary. Technical Report TR 1069, Computer Sciences Department, University of Wisconsin, Madison, WI, January 1992.

[3] C. Chang, R. Nelson, and D. Yao. Optimal Task Scheduling on Distributed Parallel Processors. *Proceedings of Performance'93.*

[4] G. Dantzig. Linear Programming and Extensions. Princeton University Press, Princeton, 1963.

[5] L. Dowdy. On the Partitioning of Multiprocessor Systems. Technical Report, Vanderbilt University, Nashville, TN, July 1988.

[6] G. Grimmett, and D. Stirzaker. Probability and Random Processes. Oxford University Press, 1989.

[7] A. Gupta, A. Tucker, and L. Stevens. Making Effective Use of Shared Memory Multiprocessors: The Process Control Approach. Technical Report, Computer Sciences Department, Stanford University, Stanford, CA, July 1991.

[8] F. Kelly. Reversibility and Stochastic Networks. John Wiley & Sons, 1979.

[9] S. Leutenegger. Issues in Multiprogrammed Multiprocessor Sharing. Ph.D. Thesis, Technical Report TR 954, Computer Sciences Department, University of Wisconsin, Madison, WI, August 1990.

[10] S. Leutenegger, and R. Nelson. Analysis of Spatial and Temporal Scheduling Policies for Semi-Static and Dynamic Multiprocessor Environments. Research Report, IBM T.J. Watson Research Center, Yorktown Heights, August 1991.

[11] S. Leutenegger, and M. Vernon. The Performance of Multiprogrammed Multiprocessor Scheduling Policies. *Proceedings of the ACM SIGMETRICS Conference on Measurement & Modeling of Computer Systems*, May 1990, pp. 226-236.

[12] S. Majumdar, D. Eager, and R. Bunt. Scheduling in Multiprogrammed Parallel Systems. *Performance Evaluation Review 16*, 1 (May 1988), 104-113.

[13] S. Majumdar, D. Eager, and R. Bunt. Characterisation of programs for scheduling in multiprogrammed parallel systems. *Performance Evaluation*, Vol. 13, 1991, pp. 109-130.

[14] A. Makowski, and R. Nelson. Distributed Parallelism Considered Harmful. Technical Report RC 17449, IBM Research Division, 1991.

[15] R. Mansharamani. Efficient Analysis of Parallel Processor Scheduling Policies. Ph.D. Thesis, Computer Sciences Department, University of Wisconsin, Madison, WI, November 1993.

[16] R. Mansharamani, and M. Vernon. Approximate Analysis of Parallel Processor Allocation Policies. *In preparation.*

[17] R. Mansharamani, and M. Vernon. Benefits of Instant Job Service for Parallel Processor Scheduling. *In preparation.*

[18] C. McCann, R. Vaswani, and J. Zahorjan. A Dynamic Processor Allocation Policy for Multiprogrammed, Shared Memory Multiprocessors. *ACM Transactions on Computer Systems 11*, 2 (May 1993), 146–178.

[19] V. Naik, S. Setia, and M. Squillante. Scheduling of Large Scientific Applications on Distributed Memory Multiprocessor Systems. *Proceedings of the 6th SIAM Conference on Parallel Processing for Scientific Computation.* IBM Research Report RC 18621, T. J. Watson Research Center, Yorktown Heights, Jan. 1993.

[20] V.Naik, S. Setia, and M. Squillante. Performance Analysis of Job Scheduling Policies in Parallel Supercomputing Environments. To appear, *Proceedings of Supercomputing'93*, November 1993. IBM Research Report RC 19138, September 1993.

[21] R. Nelson. A Performance Evaluation of a General Parallel Processing Model. *Proceedings of the ACM SIGMETRICS Conference on Measurement & Modeling of Computer Systems*, May 1990, pp. 13-26 .

[22] R. Nelson, and D. Towsley. A Performance Evaluation of Several Priority Policies for Parallel Processing Systems. COINS Technical Report 91-32, Computer and Information Sciences, University of Massachusetts, Amherst, MA, May 1991. (To appear in JACM.)

[23] R. Nelson, D. Towsley, and A. Tantawi. Performance Analysis of Parallel Processing Systems. *IEEE Transactions on Software Engineering*, April 1988, pp. 532-540.

[24] A. Roberts, and D. Varberg. *Convex Functions.* Academic Press, New York, 1973.

[25] S. Ross. *Stochastic Processes.* New York, Wiley 1983.

[26] S. Setia, M. Squillante, and S. Tripathi. Analysis of Processor Allocation in Multiprogrammed Parallel Processing Systems. Technical Report CS-TR-2840, University of Maryland, College Park, MD, February 1992.

[27] S. Setia, and S. Tripathi. An Analysis of Several Processor Partitioning Policies for Parallel Computers. Technical Report CS-TR-2684, University of Maryland, College Park, MD, May 1991.

[28] S. Stidham. A last word on $L = \lambda W$. *Operations Research 22*, 2 (1974), 417-421.

[29] D. Towsley, C. Rommel, and J. Stankovic. Analysis of Fork-Join Program Response Times on Multiprocessors. *IEEE Transactions on Parallel and Distributed Systems*, July 1990, pp. 286-303.

[30] A. Tucker, and A. Gupta. Process Control and Scheduling Issues for Multiprogrammed Shared-Memory Multiprocessors. *Proceedings of the 12th ACM Symposium on Operating System Principles*, December 1989, pp. 159-166.

[31] J. Walrand. Introduction to Queueing Networks. Prentice Hall 1988.

[32] S. Zhou, and T. Brecht. Processor-pool-based Scheduling for Large-Scale NUMA Multiprocessors. *Performance Evaluation Review 19*, 1 (May 1991), 133-142.

# Appendix

## A Constraints on System Parameters

The constraints on system parameters developed below delineate the design space for evaluating the qualitative behavior of the EQS policy and are also used to identify stress tests for validating mean response time approximations. Below achievable lower and upper bounds on $C_N$ and $\overline{S}$ are derived as a function of the free parameters of the model, namely $\overline{N}$, $\overline{D}$, $r$, and $\gamma$.

### A.1 Constraints on $C_N$

In the workload model $N$ is bounded above by $P$ because a program cannot make use of more than $P$ processors. As a result, the coefficient of variation of $N$, $C_N$ is generally not independent of $\overline{N}$. For example, if $\overline{N} = 1$ or $\overline{N} = P$, $C_N = 0$. The following proposition provides bounds on $C_N$ in terms of $\overline{N}$.

**Proposition A.1** *For a given $\overline{N}$,*

$$0 \leq C_N \leq \sqrt{\frac{\overline{N}(P+1) - P}{\overline{N}^2} - 1}.$$

*The lower bound is achieved when $N$ is constant and integer-valued for all jobs, and the upper bound is achieved when $N$ has a $K_2(1, P, \frac{P-\overline{N}}{P-1})$ distribution.*

**Proof.** The lower bound is trivial. The derivation of the upper bound is as follows. Since $C_N = \sigma_N/\overline{N}$ an upper bound for $\sigma_N$ is needed. By definition,

$$\sigma_N^2 = E[N^2] - \overline{N}^2$$

$$E[N^2] = \sum_{i=1}^{P} p_k k^2 = \sum_{i=1}^{P} p_k f(k), \tag{15}$$

where $f(x) = x^2$. An upper bound on $E[N^2]$ can be derived by observing that $f$ is a convex function, that is,

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y), \qquad 0 \leq \alpha \leq 1.$$

Choosing $\alpha$ such that $\alpha \cdot 1 + (1-\alpha)P = k$, that is, $\alpha = (P-k)/(k-1)$, gives the following bound for $f(k)$,

$$f(k) = f(\alpha \cdot 1 + (1-\alpha)P) \leq \alpha f(1) + (1-\alpha) f(P) = \frac{P-k}{P-1} \cdot 1 + \frac{k-1}{P-1} P^2.$$

Using this bound in (15) it follows that,

$$\begin{aligned} E[N^2] &\leq \sum_{k=1}^{P} p_k \left( \frac{P-k}{P-1} \cdot 1 + \frac{k-1}{P-1} P^2 \right) \\ &= \frac{P-\overline{N}}{P-1} + \frac{\overline{N}-1}{P-1} P^2 \\ &= \overline{N}(P+1) - P. \end{aligned}$$

Hence,

$$C_N^2 = \frac{E[N^2] - \overline{N}^2}{\overline{N}^2} \leq \frac{\overline{N}(P+1) - P}{\overline{N}^2} - 1,$$

which yields the required upper bound. From the derivation it follows that this upper bound is attained when $N$ has nonzero mass only at 1 and P. ∎

## A.2 Constraints on $\overline{S}$

First consider $\overline{S}$ for general ERDs in terms of $\overline{D}$, $r$, $N$, and $\gamma$.

**Lemma A.1**

$$\overline{S} = (1-r^2)\,\overline{D}\,E\left[\frac{1}{\gamma(N)}\right] \;+\; r^2\,\frac{\overline{D}}{\overline{N}}\,E\left[\frac{N}{\gamma(N)}\right]. \tag{16}$$

**Proof.** Since $S = D/\gamma(N)$ the correlation model in Section 2.3 yields

$$\begin{aligned}\overline{S} = E\left[\frac{D}{\gamma(N)}\right] &= \sum_{k=1}^{P} p_k \left\{ q\,\frac{A}{\gamma(k)} \;+\; (1-q)\,\frac{c\,k}{\gamma(k)} \right\} \\ &= q\overline{D}E\left[\frac{1}{\gamma(N)}\right] \;+\; (1-q)\,\frac{\overline{D}}{\overline{N}}E\left[\frac{N}{\gamma(N)}\right],\end{aligned}$$

from which (16) follows. ∎

Equation (16) shows that

$$\overline{S} = (1-r^2)\,\overline{S}(r=0) \;+\; r^2\overline{S}(r=1).$$

For the case of $r = 0$ we have the following bounds on $\overline{S}$.

**Theorem A.1** *For a concave ERD $\gamma$, a given value of $\overline{N}$, and $r = 0$,*

$$\frac{\overline{D}}{\overline{N}} \le \frac{\overline{D}}{\gamma(\overline{N})} \le \overline{S} \le \overline{D}\left(\frac{P-\overline{N}}{P-1} + \frac{\overline{N}-1}{P-1}\cdot\frac{1}{\gamma(P)}\right).$$

*The lower bound is attained when $N$ is constant (i.e., $C_N = 0$) and the upper bound is attained when $N$ has a $K_2(1, P, \frac{P-\overline{N}}{P-1})$ distribution (maximum $C_N$).*

**Proof.** Since $\gamma$ is concave, $1/\gamma$ is convex [24] and the proof is similar to the proof of Proposition A.1. ∎

For the linear ERD the bounds on $\overline{S}$ for $r = 0$ reduce to the following form:

**Corollary A.1** *For the linear ERD, a given value of $\overline{N}$, and $r = 0$,*

$$\frac{\overline{D}}{\overline{N}} \le \overline{S} \le \overline{D}\left(1 - \frac{\overline{N}-1}{P}\right).$$

When $r = 1$ the reverse conditions for minimum and maximum $\overline{S}$ are obtained, as shown in the following theorem.

**Theorem A.2** *If $N/\gamma(N)$ is concave then for a given $\overline{N}$ and $r = 1$,*

$$\frac{\overline{D}}{\overline{N}}\left(\frac{P-\overline{N}}{P-1} + \frac{\overline{N}-1}{P-1}\cdot\frac{P}{\gamma(P)}\right) \le \overline{S} \le \frac{\overline{D}}{\gamma(\overline{N})}.$$

*The lower bound is attained when $N$ has a $K_2(1, P, \frac{P-\overline{N}}{P-1})$ distribution (maximum $C_N$) and the upper bound is attained when $N$ is constant (i.e., $C_N = 0$).*

**Proof.** See [15] ∎

From Theorems A.1 and A.2 we obtain the following corollary.

**Corollary A.2** *If $\gamma$ is a concave ERD such that $N/\gamma(N)$ is concave then*

$$\overline{S}(r=1) \;\le\; \overline{S}(r=0).$$

**Proof.** From Theorem A.2 $\overline{S}(r=1) \leq \overline{D}/\gamma(\overline{N})$ under the given conditions for $\gamma$, and from Theorem A.1 $\overline{D}/\gamma(\overline{N}) \leq \overline{S}(r=0)$. ∎

To summarize the above results, note that for a concave ERD $\gamma$, for uncorrelated workloads $\overline{S}$ is minimum when $C_N$ is minimum and $\overline{S}$ is maximum when $C_N$ is maximum. For a concave ERD $\gamma$ such that $N/\gamma(N)$ is concave, and for fully correlated workloads, $\overline{S}$ is maximum when $C_N$ is minimum and $\overline{S}$ is minimum when $C_N$ is maximum. For the latter conditions on $\gamma$, $\overline{S}$ decreases with workload correlation $r$. (The above derivations only proved $\overline{S}(r=1) \leq \overline{S}(r=0)$, but using (16) in addition to this bound shows that that $\overline{S}$ decreases with $r$.)

# B Proofs of Theorems 4.1 and 4.2

**Theorem 3.1** *If $\ell$ and $m$ are constants such that $\ell \leq m$, then under the workload assumptions* $(\lambda, \cdot, \exp(1/\overline{D}), r=0, \gamma \in \mathcal{E}^c, E(j)=\gamma(j))$,

$$\overline{R}_{EQS}(m \leq N \leq P) \quad \leq \quad \overline{R}_{EQS}(1 \leq N \leq \ell).$$

Let $\Gamma_I$ = $(EQS,\ \lambda,\ m \leq N \leq P,\ \exp(1/\overline{D}),\ r=0,\ \gamma \in \mathcal{E}^c,\ E(j)=\gamma(j))$, and
let $\Gamma_{II}$ = $(EQS,\ \lambda,\ 1 \leq N \leq \ell,\ \exp(1/\overline{D}),\ r=0,\ \gamma \in \mathcal{E}^c,\ E(j)=\gamma(j))$.
The following lemma is used in the proof of this theorem.

**Lemma B.1** *Suppose there are $K$ jobs in system $\Gamma_I$ such that the allocation of processing power to these jobs is $(a_1, a_2, \ldots, a_K)$, and suppose there are $M \geq K$ jobs in system $\Gamma_{II}$ such that the allocation of processing power to these jobs is $(b_1, b_2, \ldots, b_M)$. Then*

$$\sum_{i=1}^{K} \gamma(a_i) \geq \sum_{i=1}^{K} \gamma(b_i).$$

*(Note that the summation is from 1 to $K$ on both sides.)*

**Proof.** Since the ERD $\gamma$, which is the same for both systems, is *concave and nondecreasing*

$$\sum_{i=1}^{K} \gamma(b_i) \leq K\gamma\left(\frac{\sum_{i=1}^{K} b_i}{K}\right) \leq K\gamma\left(\frac{P}{K}\right) \tag{17}$$

Since $b_i \leq \ell$, $i = 1, 2, \ldots, K$, and $\gamma$ is nondecreasing

$$\sum_{i=1}^{K} \gamma(b_i) \leq K\gamma(\ell).$$

The above inequality together with (17) yields

$$\sum_{i=1}^{K} \gamma(b_i) \leq K \min(\gamma(\ell), \gamma(P/K)) \leq K \min(\gamma(m), \gamma(P/K)), \tag{18}$$

where the last inequality follows because $\ell \leq m$ and $\gamma$ is nondecreasing.
To see that

$$K \min(\gamma(m), \gamma(P/K)) \leq \sum_{i=1}^{K} \gamma(a_i), \tag{19}$$

consider the following two cases:

(i) $m \geq P/K$:
If $m \geq P/K$ then $a_i = P/K$, $i = 1, 2, \ldots, K$ (since $P/K$ is the equiallocation number and the available parallelism of each job in $\Gamma_I$ is at least $m$). Hence

$$K \min(\gamma(m), \gamma(P/K)) = K\gamma(P/K) = \sum_{i=1}^{K} \gamma(a_i).$$

(ii) $m < P/K$:
Since $m < P/K$ each job in $\Gamma_I$ is allocated at least $m$ processors. That is, $a_i \geq m$, $i = 1, 2, \ldots, K$. Hence

$$K \min(\gamma(m), \gamma(P/K)) = K\gamma(m) \leq \sum_{i=1}^{K} \gamma(a_i).$$

This proves inequality (19). The lemma follows from inequalities (18) and (19). ∎

**Proof of Theorem 3.1.** This theorem is proved by sample path analysis, making use of the following observations:

(i) If a job is allocated processing power $x$ then the residual life time of the job is exponentially distributed with rate $\gamma(x)\mu$.

(ii) If there are $k > 0$ jobs in system $\Gamma_i$, $i = I, II$, at time $t$ with the $j^{th}$ job having a processor allocation of $x_j$, $1 \leq j \leq k$, then the time to the next departure from $\Gamma_i$ is exponentially distributed with rate $\sum_{j=1}^{k} \gamma(x_j)\mu$.

Let $Q_i(t)$ be the number of jobs in system $\Gamma_i$ at time $t$, $i = I, II$. Let $\alpha_k^I(t) = \gamma(a_k)/P$, $k = 1, 2, \ldots, Q_I(t)$, where $a_k$ is the processor allocation to the $k^{th}$ job in $\Gamma_I$ at time $t$. Similarly let $\alpha_k^{II}(t) = \gamma(b_k)/P$, $k = 1, 2, \ldots, Q_{II}(t)$, where $b_k$ is the processor allocation to the $k^{th}$ job in $\Gamma_{II}$ at time $t$. Thus the $k^{th}$ job in $\Gamma_i$ departs with rate $\alpha_k^i(t)P\mu$, $i \in \{I, II\}$.

*Coupling of Sample Paths in $\Gamma_I$ and $\Gamma_{II}$*

Fix the arrival times of jobs to be the same in $\Gamma_I$ and $\Gamma_{II}$. Fix sequences of integers $\{N_i^I\}_{i=1}^{\infty}$ and $\{N_i^{II}\}_{i=1}^{\infty}$ for available job parallelisms in $\Gamma_I$ and $\Gamma_{II}$ respectively, where $m \leq N_i^I \leq P$ and $1 \leq N_i^{II} \leq \ell$, $i = 1, 2, \ldots$. Consider that *potential job completions* [31] occur in each of $\Gamma_I$ and $\Gamma_{II}$ at jumps of a Poisson process with rate $P\mu$. Fix the same potential completion instants $\{T_i\}_{i=1}^{\infty}$ in both $\Gamma_I$ and $\Gamma_{II}$. To generate *actual* job completion times in $\Gamma_I$ and $\Gamma_{II}$ let $\{U_i\}_{i=1}^{\infty}$ be i.i.d. Uniform[0,1) random variables. At the $r^{th}$ potential completion instant $T_r$, the $k^{th}$ job in $\Gamma_i$ departs if

$$U_r \in \left[ \sum_{j=1}^{k-1} \alpha_j^i(T_r^-), \sum_{j=1}^{k} \alpha_j^i(T_r^-) \right), \quad k = 1, 2, \ldots, Q_i(t), \quad i \in \{I, II\}. \tag{20}$$

This ensures that the probability that the $k^{th}$ job departs from $\Gamma_i$ is $\alpha_k^i(T_r^-)$.

*Sample Path Analysis*

Using the above coupling of sample paths we show by an induction over time that for every sample path, for all $t \geq 0$

$$Q_I(t) \leq Q_{II}(t). \tag{21}$$

We carry out the induction only over arrival instants and potential completion instants since no jobs depart in between these event times. Let $\{t_i\}_{i=0}^{\infty}$ be the sequence of arrival and potential completion times arranged in increasing order. Let both $\Gamma_I$ and $\Gamma_{II}$ start out with zero jobs each. Then clearly (21) is satisfied at $t = t_0$. Assume that (21) is true for all $t \leq t_j$. Since no jobs arrive or depart in $(t_j, t_{j+1})$ (21) is also true for all $t_j < t < t_{j+1}$. We now prove that (21) is true at $t = t_{j+1}$. Consider all possible events at time $t_{j+1}$.

1. Job Arrival:
   By the induction hypothesis it follows that

$$Q_I(t_{j+1}) = Q_I(t_j) + 1 \leq Q_{II}(t_j) + 1 = Q_{II}(t_{j+1}).$$

2. Potential Completion:

   (a) No departure from each of $\Gamma_I$ and $\Gamma_{II}$:

$$Q_I(t_{j+1}) = Q_I(t_j) \leq Q_{II}(t_j) = Q_{II}(t_{j+1}).$$

   (b) Departure from $\Gamma_I$ but not from $\Gamma_{II}$:

$$Q_I(t_{j+1}) = Q_I(t_j) - 1 \leq Q_{II}(t_j) - 1 = Q_{II}(t_{j+1}) - 1 < Q_{II}(t_{j+1}).$$

   (c) Departure from each of $\Gamma_I$ and $\Gamma_{II}$:

$$Q_I(t_{j+1}) = Q_I(t_j) - 1 \leq Q_{II}(t_j) - 1 = Q_{II}(t_{j+1}).$$

   (d) Departure from $\Gamma_{II}$ but not from $\Gamma_I$:
   This implies that

$$U_r \in \left[0, \sum_{i=1}^{Q_{II}(t_j)} \alpha_i^{II}(t_{j+1}^-)\right), \quad \text{and} \quad U_r \in \left[\sum_{i=1}^{Q_I(t_j)} \alpha_i^I(t_{j+1}^-),\ 1\right), \tag{22}$$

   where $t_{j+1} = T_r$, the $r^{th}$ potential completion instant, $1 \leq r \leq j+1$. Since these two intervals overlap, we have

$$\sum_{i=1}^{Q_I(t_j)} \alpha_i^I(t_{j+1}^-) < \sum_{i=1}^{Q_{II}(t_j)} \alpha_i^{II}(t_{j+1}^-). \tag{23}$$

   Since $Q_I(t_j) \leq Q_{II}(t_j)$ (induction hypothesis) we have from Lemma B.1 that

$$\sum_{i=1}^{Q_I(t_j)} \alpha_i^{II}(t_{j+1}^-) = \frac{1}{P} \sum_{i=1}^{Q_I(t_j)} \gamma(b_i) \leq \frac{1}{P} \sum_{i=1}^{Q_I(t_j)} \gamma(a_i) = \sum_{i=1}^{Q_I(t_j)} \alpha_i^I(t_{j+1}^-). \tag{24}$$

   (23) and (24) together imply

$$\sum_{i=1}^{Q_I(t_j)} \alpha_i^{II}(t_{j+1}^-) \leq \sum_{i=1}^{Q_I(t_j)} \alpha_i^I(t_{j+1}^-) < \sum_{i=1}^{Q_{II}(t_j)} \alpha_i^{II}(t_{j+1}^-),$$

   which shows that $Q_I(t_j) < Q_{II}(t_j)$. Hence

$$Q_I(t_{j+1}) = Q_I(t_j) \leq Q_{II}(t_j) - 1 = Q_{II}(t_{j+1}).$$

This completes the proof by induction. Thus, we have shown for every sample path that $Q_I(t) \leq Q_{II}(t)$, $\forall t \geq 0$. Hence for every sample path

$$\overline{Q}_I = \lim_{t\to\infty} \frac{1}{t} \int_0^t Q_I(s)ds \leq \lim_{t\to\infty} \frac{1}{t} \int_0^t Q_{II}(s)ds = \overline{Q}_{II},$$

from which it follows by Little's Law [28] that $\overline{R}_{\Gamma_I} \leq \overline{R}_{\Gamma_{II}}$ for every sample path. Now uncondition on arrival times, available parallelisms, and potential completion times. ∎

**Remark:** Note that the above proof does not require the assumption of Poisson arrivals. The arrival process can be any GI process.

**Theorem 3.2** *Let $\Gamma_I$ be a system with the EQS policy and primitive workload variables $\{(A_i, D_i, N_i \geq k, E_i), i = 1, 2, \ldots\}$, where $A_i$ is job $i$'s arrival time, $D_i$ its total demand, $N_i$ its available parallelism, and $E_i$ its execution rate function. Let these primitive variables have arbitrary marginals (given that $N_i \geq k$, and the other variables make sense, e.g., $D_i \geq 0$) with arbitrary dependencies among them. Let $\Gamma_{II}$ be a system with the EQS policy and the same workload as $\Gamma_I$ except that $N_i = k$ for all $i = 1, 2, \ldots$. Then*

$$\overline{R}_{\Gamma_I} \leq \overline{R}_{\Gamma_{II}}, \quad k = 1, 2, \ldots, P.$$

**Proof.** Let $\mathcal{Q}_I(t)$ be the set of jobs in system $\Gamma_I$ at time $t$, and likewise, let $\mathcal{Q}_{II}(t)$ be the set of jobs in system $\Gamma_{II}$ at time $t$. We prove this theorem by suitably coupling sample paths for $\Gamma_I$ and $\Gamma_{II}$, and showing that for every sample path $\mathcal{Q}_I(t) \subseteq \mathcal{Q}_{II}(t)$, for all $t \geq 0$, from which it will follow that $\overline{R}_{\Gamma_I} \leq \overline{R}_{\Gamma_{II}}$.

*Coupling of Sample Paths in $\Gamma_I$ and $\Gamma_{II}$*

Fix $\{A_i, D_i\}_{i=1}^{\infty}$ as the same for both $\Gamma_I$ and $\Gamma_{II}$. For system $\Gamma_I$ choose a sequence $\{N_i^I\}_{i=1}^{\infty}$ such that $N_i^I \geq k$, $i = 1, 2, \ldots$. For system $\Gamma_{II}$ fix $N_i^{II} = k$ for all $i = 1, 2, \ldots$. Pick a sequence of execution rate functions $\{E_i^I\}_{i=1}^{\infty}$ for $\Gamma_I$ where $E_i^I$ is nondecreasing, $i = 1, 2, \ldots$. Fix the execution rate function for job $i$ in system $\Gamma_{II}$ as $E_i^{II}(x) = E_i^I(x)$, for $0 \leq x \leq k$, and $E_i^{II}(x) = E_i^I(k)$, $x \geq k$.

*Sample Path Analysis*

Under the above coupling of sample paths we show by induction over time that for every pair of coupled sample paths, for all $t \geq 0$,

$$\mathcal{Q}_I(t) \subseteq \mathcal{Q}_{II}(t). \tag{25}$$

Let $a_i^I(t)$ and $a_i^{II}(t)$ be the allocations of processing power to job $i$ in system $\Gamma_I$ and $\Gamma_{II}$, respectively, at time $t$. Note that $a_i^I(t) = 0$ if $i \notin \mathcal{Q}_I(t)$, and $a_i^{II}(t) = 0$ if $i \notin \mathcal{Q}^{II}(t)$. From (25) it follows that

$$a_i^I(t) \geq a_i^{II}(t), \quad i \in \mathcal{Q}^I(t), \tag{26}$$

because

$$\begin{aligned} a_i^{II}(t) &= \min(k, P/|\mathcal{Q}_{II}(t)|) \\ &\leq \min\left(N_i^I, P/|\mathcal{Q}_{II}(t)|\right), \qquad \text{since } N_i^I \geq k \\ &\leq \min\left(N_i^I, P/|\mathcal{Q}_I(t)|\right), \qquad \text{since } |\mathcal{Q}_I(t)| \leq |\mathcal{Q}_{II}(t)| \\ &\leq a_i^I(t). \end{aligned}$$

The last inequality holds because if job $i$ gets $N_i^I$ processors in $\Gamma_I$ then $a_i^I(t) = N_i^I$ and if job $i$ gets less than $N_i^I$ processors then it gets at least as many as the equiallocation number $P/|\mathcal{Q}_I(t)|$, by definition of the EQS policy.

We carry out the induction over arrival and departure times in $\Gamma_I$ and $\Gamma_{II}$. Let $\{t_i\}_{i=0}^{\infty}$ be the sequence of arrival and departure times in $\Gamma_I$ and $\Gamma_{II}$ arranged in increasing order. Let both $\Gamma_I$ and $\Gamma_{II}$ start out with zero jobs each at $t = 0$. Then clearly (25) is satisfied at $t = t_0$. Assume that (25) is true for all $t \leq t_j$. Since no jobs arrive or depart in $(t_j, t_{j+1})$ it follows that (25) is true for all $t < t_{j+1}$. We now prove that (25) is true at $t = t_{j+1}$. Consider all possible events at time $t_{j+1}$.

1. Arrival of job k:
   By the induction hypothesis it follows that

$$\mathcal{Q}_I(t_{j+1}) = \mathcal{Q}_I(t_j) \cup \{k\} \subseteq \mathcal{Q}_{II}(t_j) \cup \{k\} = \mathcal{Q}_{II}(t_{j+1}).$$

2. Departure from $\Gamma_I$ only:

$$\mathcal{Q}_I(t_{j+1}) \subset \mathcal{Q}_I(t_j) \subseteq \mathcal{Q}_{II}(t_j) = \mathcal{Q}_{II}(t_{j+1}).$$

3. Departure from both $\Gamma_I$ and $\Gamma_{II}$:
Suppose job $\ell$ departs from $\Gamma_I$ and job $m$ departs from $\Gamma_{II}$. Then we have the following cases depending on how $\ell$ is related to $m$:

   (a) $\ell = m$:

$$\mathcal{Q}_I(t_{j+1}) = \mathcal{Q}_I(t_j) - \{\ell\} \subseteq \mathcal{Q}_{II}(t_j) - \{m\} = \mathcal{Q}_{II}(t_{j+1}).$$

   (b) $\ell \neq m$:
   Depending on whether or not $m$ and $\ell$ are present in $\mathcal{Q}_I(t_j)$ and $\mathcal{Q}_{II}(t_j)$, respectively, we have the following cases:

      (i) $m \in \mathcal{Q}_I(t_j)$:
      This is impossible. The reason is that $m \in \mathcal{Q}_I(t_j) \Rightarrow m \in \mathcal{Q}_I(t_{j+1})$ because $\ell \neq m$. Since $\mathcal{Q}_I(s) \subseteq \mathcal{Q}_{II}(s)$, for all $0 \leq s < t_{j+1}$, it follows from (26) that $a_m^I(s) \geq a_m^{II}(s)$ for $A_m \leq s < t_{j+1}$. Since job $m$ has not departed from $\Gamma_I$ by time $t_{j+1}$, we have

$$D_m > \int_0^{t_{j+1}} E_m^I(a_m^I(s))ds \geq \int_0^{t_{j+1}} E_m^{II}(a_m^{II}(s))ds.$$

      Hence job $m$ has not departed from $\Gamma_{II}$ by time $t_{j+1}$, which is a contradiction.

      (ii) $m \notin \mathcal{Q}_I(t_j)$, $\ell \in \mathcal{Q}_{II}(t_j)$:
      Since $m \notin \mathcal{Q}_I(t_j)$, we have by the induction hypothesis that $\mathcal{Q}_I(t_j) \subset \mathcal{Q}_{II}(t_j)$ and since $\ell$ departs $\Gamma_I$ but not $\Gamma_{II}$ at time $t_{j+1}$ it follows that

$$\mathcal{Q}_I(t_{j+1}) \subset \mathcal{Q}_{II}(t_{j+1}).$$

      (iii) $m \notin \mathcal{Q}_I(t_j)$, $\ell \notin \mathcal{Q}_{II}(t_j)$:
      Similar to case (i), it is impossible that $\ell \in \mathcal{Q}_I(t_j)$ and $\ell \notin \mathcal{Q}_{II}(t_j)$.

4. Departure from $\Gamma_{II}$ only:
   Suppose job $m$ departs from $\Gamma_{II}$. Then either $m \in \mathcal{Q}_I(t_j)$ or $m \notin \mathcal{Q}_I(t_j)$. The former case is impossible as in case 3(b)(i). In the latter case $\mathcal{Q}_I(t_j) \subset \mathcal{Q}_{II}(t_j)$ and $\mathcal{Q}_I(t_{j+1}) \subseteq \mathcal{Q}_{II}(t_{j+1})$.

This completes the proof by induction which shows that for every sample path $\mathcal{Q}_I(t) \subseteq \mathcal{Q}_{II}(t)$. Therefore, job $i$ departs from $\Gamma_I$ at least as early as it does from from $\Gamma_{II}$, from which it follows that the response time of job $i$ in $\Gamma_I$ is less than or equal to its response time in $\Gamma_{II}$ for every sample path, $i = 1, 2, \ldots$. Therefore $\overline{R}_{\Gamma_I} \leq \overline{R}_{\Gamma_{II}}$ for every sample path. Unconditioning on $\{(A_i, D_i, N_i^I, E_i^I), i = 1, 2, \ldots\}$ yields the final result. ∎