Adv. Appl. Prob. **41,** 546–575 (2009)
Printed in Northern Ireland

JOB REPLICATION ON MULTISERVER SYSTEMS

YUSIK KIM,* **
RHONDA RIGHTER * AND
RONALD WOLFF,* *University of California, Berkeley*

Abstract

Parallel processing is a way to use resources efficiently by processing several jobs simultaneously on different servers. In a well-controlled environment where the status of the servers and the jobs are well known, everything is nearly deterministic and replicating jobs on different servers is obviously a waste of resources. However, in a poorly controlled environment where the servers are unreliable and/or their capacity is highly variable, it is desirable to design a system that is robust in the sense that it is not affected by the poorly performing servers. By replicating jobs and assigning them to several different servers simultaneously, we not only achieve robustness but we can also make the system more efficient under certain conditions so that the jobs are processed at a faster rate overall. In this paper we consider the option of replicating jobs and study how the performance of different 'degrees' of replication, ranging from no replication to full replication, affects the performance of a system of parallel servers.

Keywords: Stochastic scheduling; grid computing; job replication

2000 Mathematics Subject Classification: Primary 68M20
Secondary 90B36; 90B22

1. Introduction

When scheduling jobs over multiple resources, it is natural to take advantage of the parallel structure, where different jobs are served simultaneously on different servers. Replicating the same job on different servers is usually considered a waste of resources; in many cases, such as for manufacturing systems, it is not even feasible. However, for most computational jobs, replicating a job is not only feasible but also practically effortless. We investigate situations where using the option to replicate can be beneficial.

Recently, a new paradigm for distributed computing, called grid computing [3], has emerged and is becoming popular. One distinctive feature of grid computing is the unusual characteristics of the resources. The resource pool is a massive group of autonomous and unreliable servers connected by the Internet, where each server is owned and/or controlled by an independent entity. Examples are normal home users with PCs, high performance gaming consoles with idle CPU cycles, and in a corporate environment, where computers assigned to employees are left on most of the time. Currently, deployed grid applications include SETI@home [5] for the search of intelligent life forms in outer space and Folding@home [6] for studying the structure of protein molecules. In both cases, the job is a computation to be performed on an assigned batch of data.

Received 11 December 2008; revision received 9 February 2009.
* Postal address: Department of Industrial Engineering and Operations Research, 4141 Etcheverry Hall, Berkeley, CA 94720, USA.
** Email address: ykim@newton.berkeley.edu

The use of replication is not new. In the reliability literature, the life of a device is extended by having redundant components (replicates) for specific functions [7]. Also, in the grid computing literature, job replication has been used to control reliability (or fault tolerance) by estimating the number of replicas to guarantee a certain probability of successful job completion [8]. In contrast, the primary goal of replication in this paper is to increase efficiency, that is, to either minimize the amount of time to complete a set of jobs or to maximize the service capacity of a group of servers.

We define the service time of a job submitted to a server as the time between submission and completion of the job on that server. In grid computing, the service time depends on the load at the server, i.e. how the server is being used by the local owner at that time. That is, it includes local queueing delays and downtimes, and, hence, can be highly variable. By submitting replicates of the same job to multiple servers and using the first to complete, we are protected from the case where the service time of one of them is unusually long. But replicating jobs is done at the expense of sacrificing the opportunity to process more jobs simultaneously. Thus, there is a trade-off between completing a job more quickly (by replicating) and serving different jobs in parallel (by not replicating). Essentially, we are comparing the performance of processing short jobs sequentially versus long jobs in parallel. Hence, we introduce the notion of the 'degree' of replication and seek to optimize it.

To better illustrate this idea, suppose that there are 10 parallel servers. If we let each job simultaneously occupy two servers (i.e. replicate twice), we would be able to process five different jobs at a time, whereas if we let each job occupy five servers, we would only be able to process two different jobs at a time. The latter would be said to have a higher degree of replication. The goal of this paper is to characterize the performance of a system in terms of the degree of replication.

We consider two performance objectives. One is to minimize the makespan for a fixed number of jobs to be processed, and the other is to maximize the effective, or long-run, service rate when arrivals occur according to a Poisson process. We also briefly consider minimizing the expected response time. For all of our objectives, we assume that the degree of replication must be chosen at the outset and that it is held fixed thereafter. That is, we consider a design problem rather than a control problem.

Unless stated otherwise, we assume that service times are independent and identically distributed (i.i.d.) both spatially and temporally. That is, the service times of jobs across servers, replicated or not, are i.i.d., and the service times of subsequent jobs assigned to the same server are i.i.d. as well. The spatial independence assumption is reasonable since the service time duration of a job at a server depends primarily on the server (how busy it is, for example) rather than the job. As for the assumption of distribution homogeneity across servers, we assume that we group servers of similar service capacity, and we consider the scheduling problem for each such group separately. Such grouping eases implementation as well as analysis. The assumption that successive service times on a given server have the same distribution is reasonable assuming that future changes in server speed and availability are unpredictable, and given that we are tackling a design problem, so cannot change our decision in the light of new information about the current speed of a server. For the model with job arrivals (Section 4), it turns out that our objective to minimize the effective service rate in a system is not sensitive to any temporal dependency structure of service times. But, for the model with a fixed number of jobs in which our objective is to minimize makespan (Section 3), the results will be different when successive service times are not independent. For simplicity of analysis, we begin by assuming temporal independence of the service times and provide

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

some discussion in the sequel of how positive temporal dependence changes the results we obtain, assuming independence.

There has been work done on replication for cases where the optimal amount of replication is an extreme value, i.e. minimal or maximal replication. Borst *et al.* [1] considered a multiserver discrete-time queueing system with batch arrivals with random size. The service time of each job has a geometric distribution. It was shown that a policy that distributes the jobs over the servers as 'evenly as possible', i.e. minimal replication, minimizes both the number of jobs in the system jointly across time as well as the mean response time of the jobs.

Koole and Righter [4] proved that, for a multiserver queueing system, if the service times have a *new worse than used* (NWU) distribution (defined in Section 2) then a maximum replication policy, i.e. processing jobs sequentially using all servers for each job, stochastically maximizes the number of completed jobs jointly across time for an arbitrary arrival process of jobs. They obtained an analogous result for the optimality of minimal replication for two servers and new better than used (NBU) service times.

After some preliminary definitions in Section 2, we consider minimization of the expected makespan, defined as the expected time required to complete all jobs, for a finite number of jobs in Section 3.

Dobber [2] approximated the mean and standard deviations of order statistics necessary for computing the expected makespan in terms of the mean and standard deviations of the service distribution, and used the approximations to compare the performance of various degrees of replication. Our approach is more analytic than Dobber's. We give a condition that guarantees monotonicity of the expected makespan in the degree of replication and characterize the makespan random variable for particular distributions. For Bernoulli service time distributions, we show that as long as the mean is at most $\frac{2}{3}$, maximal replication will be preferred to minimal replication. We describe the effect of service time variability by showing that the performance of maximal replication relative to minimal replication gets better with more variable service times. Bounds and asymptotic results are also presented.

In Section 4 we assume a general arrival stream of jobs and approach the problem from a queueing perspective. We investigate how the degree of replication affects the effective service rate and, hence, the overall system load, ρ. We prove that if maximal replication is optimal for the system load objective, it is also optimal for the makespan objective. We provide monotonicity conditions for the system load in the degree of replication, and show the optimality of minimal and maximal replications for NBU and, respectively, NWU service times. A closed-form expression for the system load is found for the Bernoulli, shifted Bernoulli, and Pareto distributions. We show that, for Bernoulli(p) service times, $p = 1/\mathrm{e}$ is an upper threshold for maximal replication to be optimal.

Our general conclusion is that the degree of replication should increase with the variability of the distribution.

2. Preliminaries

Here we define terms and relations that we will use frequently. For more detail and proofs, see [9, pp. 3–112]. In what follows, X and Y are nonnegative random variables with finite expectations.

2.1. Stochastic order relations

Definition 1. (*Stochastic order.*) The random variable X is said to be greater than Y in the stochastic order, written as $X \geq_{\mathrm{st}} Y$, if $F_X(x) \leq F_Y(x)$ for all x.

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

An alternative definition is $X \geq_{\mathrm{st}} Y$ if and only if $\mathrm{E}[h(X)] \geq \mathrm{E}[h(Y)]$ for all increasing h.

Definition 2. (*Increasing and decreasing convex orders.*) The random variable X is said to be greater than Y in the increasing convex order, written as $X \geq_{\mathrm{icx}} Y$, or the decreasing convex order, written as $X \geq_{\mathrm{dcx}} Y$, if $\mathrm{E}[h(X)] \geq \mathrm{E}[h(Y)]$ for all increasing or, respectively, decreasing convex h.

We also have the tail characterization of the increasing convex order:

$$X \geq_{\mathrm{icx}} Y \quad \Longleftrightarrow \quad \int_x^\infty \bar{F}_X(u)\,\mathrm{d}u \geq \int_x^\infty \bar{F}_Y(u)\,\mathrm{d}u \quad \text{for all } x.$$

Definition 3. (*Increasing and decreasing concave orders.*) The random variable X is said to be greater than Y in the increasing concave order, written as $X \geq_{\mathrm{icv}} Y$, or the decreasing concave order, written as $X \geq_{\mathrm{dcv}} Y$, if $\mathrm{E}[h(X)] \geq \mathrm{E}[h(Y)]$ for all increasing or, respectively, decreasing concave h.

Observe that the increasing concave order can be defined through the decreasing convex order and vice versa:

$$\begin{aligned} X \geq_{\mathrm{icv}} Y \quad &\Longleftrightarrow \quad X \leq_{\mathrm{dcx}} Y, \\ X \geq_{\mathrm{dcv}} Y \quad &\Longleftrightarrow \quad X \leq_{\mathrm{icv}} Y. \end{aligned}$$

Definition 4. (*Convex order.*) The random variable X is said to be greater than Y in the convex order, written as $X \geq_{\mathrm{cx}} Y$, if $\mathrm{E}[h(X)] \geq \mathrm{E}[h(Y)]$ for all convex h.

Alternative definitions are as follows:

$$\begin{aligned} X \geq_{\mathrm{cx}} Y \quad &\Longleftrightarrow \quad \mathrm{E}[h(X)] \geq \mathrm{E}[h(Y)] \quad \text{for all convex } h \\ &\Longleftrightarrow \quad X \geq_{\mathrm{icx}} Y \quad \text{and} \quad X \geq_{\mathrm{dcx}} Y \\ &\Longleftrightarrow \quad X \geq_{\mathrm{icx}} Y \quad \text{and} \quad \mathrm{E}[X] = \mathrm{E}[Y]. \end{aligned}$$

From the last equivalence, $X \geq_{\mathrm{cx}} Y$ implies that $\mathrm{E}[X] = \mathrm{E}[Y]$. By taking $h(x) = x^2$, this also implies that $\mathrm{var}(X) \geq \mathrm{var}(Y)$. Convex ordering may be used as a measure of relative variability.

2.2. New better and new worse than used distributions

We define classes of distributions where there is an order relation between a nonnegative random variable and an aged version of itself.

Definition 5. (*NBU and NWU.*) The random variable X is called NBU if $X \geq_{\mathrm{st}} (X - t \mid X > t)$ for all $t \geq 0$. If the opposite inequality holds then the random variable X is called NWU.

The terminology 'larger is better' originated in the reliability literature, where a longer lifetime is considered to be better.

NWU is equivalent to $\mathrm{P}(X > x + t) \geq \mathrm{P}(X > x)\,\mathrm{P}(X > t)$, which is equivalent to $H(x + t) \leq H(x) + H(t)$ for all x and t, i.e. subadditive, and NBU is equivalent to $\mathrm{P}(X > x + t) \leq \mathrm{P}(X > x)\,\mathrm{P}(X > t)$, which is equivalent to $H(x + t) \geq H(x) + H(t)$ for all x and t, i.e. superadditive. Here H is the cumulative hazard rate function, $H(x) = -\log \bar{F}(x)$.

Definition 6. (*New better and new worse than used in expectations.*) The random variable X is called new better than used in expectation (NBUE) if $\mathrm{E}[X - t \mid X > t] \leq \mathrm{E}[X]$. If the opposite inequality holds then the random variable X is called new worse than used in expectation (NWUE).

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

A class of distributions depending on the variability relative to the exponential distribution can be defined.

Definition 7. (*Harmonic new better and harmonic new worse than used in expectations.*) The random variable X is called harmonic new better than used in expectation (HNBUE) if $X \leq_{\text{cx}} \text{exponential}(\mu)$, where $\text{E}[X] = 1/\mu$. If the opposite inequality holds then the random variable X is called harmonic new worse than used in expectation (HNWUE).

We have

$$\text{NWU} \Rightarrow \text{NWUE} \Rightarrow \text{HNWUE},$$
$$\text{NBU} \Rightarrow \text{NBUE} \Rightarrow \text{HNBUE}.$$

Since the coefficient of variation of an exponential random variable (RV) is 1, and convex ordering implies ordering in the variance, NWU, NWUE, and HNWUE distributions have coefficients of variation larger than or equal to 1.

2.3. Dependence measures and orders

We define measures of positive dependence and how two n-dimensional random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be ordered by the strength of positive dependence among their components. For more details and proofs, see [10, pp. 387–404].

Definition 8. (*Weak positive association (WPA).*) If the RVs $X_1, \ldots, X_n$ satisfy

$$\text{cov}(h_1(X_{i_1}, X_{i_2}, \ldots, X_{i_k}), h_2(X_{j_1}, X_{j_2}, \ldots, X_{j_{n-k}})) \geq 0$$

for all choices of disjoint subsets $\{i_1, \ldots, i_k\}$ and $\{j_1, \ldots, j_{n-k}\}$ of $\{1, 2, \ldots, n\}$, and all increasing functions h_1 and h_2 for which the above covariance is defined, then $X_1, X_2, \ldots, X_n$ are said to be weakly positively associated.

WPA characterizes an existence of positive dependence, but does not measure its strength. To measure the relative strength of positive dependence between two random vectors, we introduce the *supermodular order*. A function $\phi\colon \mathbb{R}^n \to \mathbb{R}$ is *supermodular* if, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, it satisfies

$$\phi(\boldsymbol{x}) + \phi(\boldsymbol{y}) \leq \phi(\boldsymbol{x} \wedge \boldsymbol{y}) + \phi(\boldsymbol{x} \vee \boldsymbol{y}),$$

where the operators '$\wedge$' and '$\vee$' denote coordinatewise minimum and maximum, respectively.

Definition 9. (*Supermodular order.*) The random vector $\boldsymbol{X}$ is said to be greater than $\boldsymbol{Y}$ in the supermodular order, written as $\boldsymbol{X} \geq_{\text{sm}} \boldsymbol{Y}$, if

$$\text{E}[\phi(\boldsymbol{X})] \geq \text{E}[\phi(\boldsymbol{Y})]$$

for all supermodular functions $\phi\colon \mathbb{R}^n \to \mathbb{R}$.

It can be shown that if $\boldsymbol{X}^{\text{I}}$ is a vector of independent RVs having the same marginals as $\boldsymbol{X}$, and $\boldsymbol{X}$ is weakly positively associated, then $\boldsymbol{X} \geq_{\text{sm}} \boldsymbol{X}^{\text{I}}$ (see Theorem 9.A.23 of [10]).

Definition 10. (*Positive upper and positive lower orthant dependence orders.*) The random vector $\boldsymbol{X}$ is said to be greater than $\boldsymbol{Y}$ in the positive upper orthant order, written as $\boldsymbol{X} \geq_{\text{uo}} \boldsymbol{Y}$, or the positive lower orthant order, written as $\boldsymbol{X} \geq_{\text{lo}} \boldsymbol{Y}$, if $\text{P}(\boldsymbol{X} > \boldsymbol{x}) \geq \text{P}(\boldsymbol{Y} > \boldsymbol{x})$ or, respectively, $\text{P}(\boldsymbol{X} \leq \boldsymbol{x}) \geq \text{P}(\boldsymbol{Y} \leq \boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^n$.

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

Definition 11. (*Positive quadrant dependence (PQD) order.*) The random vector $\boldsymbol{X}$ is said to be greater than $\boldsymbol{Y}$ in the PQD order, written as $\boldsymbol{X} \geq_{\mathrm{PQD}} \boldsymbol{Y}$, if $\boldsymbol{X} \geq_{\mathrm{uo}} \boldsymbol{Y}$ and $\boldsymbol{X} \geq_{\mathrm{lo}} \boldsymbol{Y}$.

Note that when $\boldsymbol{X}$ and $\boldsymbol{Y}$ are bivariate, upper and lower orthant dependence orders are equivalent, and, furthermore, if $\boldsymbol{X} \geq_{\mathrm{PQD}} \boldsymbol{Y}$ then

$$\mathrm{cov}(\boldsymbol{X}) \geq \mathrm{cov}(\boldsymbol{Y})$$

when $\boldsymbol{X}$ and $\boldsymbol{Y}$ have the same marginals.

Let $1_{\{.\}}$ denote the indicator function. Since the functions $\phi_{\boldsymbol{x}} = 1_{\{\boldsymbol{y}:\ \boldsymbol{y} > \boldsymbol{x}\}}$ and $\psi_{\boldsymbol{x}} = 1_{\{\boldsymbol{y}:\ \boldsymbol{y} \leq \boldsymbol{x}\}}$ are supermodular for each fixed $\boldsymbol{x}$, it is immediate that

$$\boldsymbol{X} \geq_{\mathrm{sm}} \boldsymbol{Y} \quad \Longrightarrow \quad \boldsymbol{X} \geq_{\mathrm{PQD}} \boldsymbol{Y}.$$

3. Analysis of makespan under finite workload

In our first model we have a finite number of jobs to process using a predetermined number of parallel servers, and we are interested in the makespan, i.e. the amount of time spent to complete all jobs. For simplicity, we assume that there are n processors and n jobs. We consider l-replication policies that replicate each job l times, where l is a factor of n. So, for the case $l = 1$, i.e. the minimal replication policy, we run n different jobs on n different processors in parallel. For the case $l = n$, i.e. the maximal replication policy, all jobs are processed sequentially, where each job at the time of execution occupies all n processors with its replicas. For the case $1 < l < n$, we partition the processors and jobs into m groups of l processors, where $lm = n$ and each group consists of l jobs waiting to be processed by the group of processors. See Figure 1. Other assumptions we make are that there is no communication delay for transmitting problem data across the system, there is no overhead induced when starting a new job, and there is no penalty for interrupting a running job. These assumptions are all reasonable in a grid computing environment where the users of donated computers download the software once, and thereafter just input/output data is transmitted.

Let us denote the service time of the kth replication of the ith job of group j as X_{ijk}, $i = 1, \ldots, l$, $j = 1, \ldots, m$, $k = 1, \ldots, l$. Because of our setup, we can also think of index k as representing the kth server in group j. Assume that X_{ijk} is nonnegative and i.i.d. for all i, j, and k. We may express the makespan for l-replications as

$$M_{(l,n)} := \max_{j \in \{1,\ldots,m\}} \sum_{i=1}^{l} \min_{k \in \{1,\ldots,l\}} X_{ijk}. \tag{1}$$

We can see that finding a closed form for the distribution of the makespan is generally intractable since it is an l-fold convolution of a function raised to the mth power. Consequently, computing the expected value analytically is very difficult. However, for the purpose of motivating the idea of replication, we estimated $\mathrm{E}[M_{(l,n)}]$ through simulated values of log-normally distributed service times X for different coefficients of variation, $c_v = \sqrt{\mathrm{var}(X)}/\,\mathrm{E}[X]$. See Table 1. Recall that an RV X has a log-normal distribution with parameters μ and σ if $\log_{\mathrm{e}} X \sim N(\mu, \sigma^2)$. Note that the mean and variance of X is $\exp\{\mu + \sigma^2/2\}$ and $(\exp\{\sigma^2\} - 1)\exp\{2\mu + \sigma^2\}$, respectively. Observe that when c_v is reasonably large, replication drastically reduces the expected makespan.

For the extreme cases of minimal and maximal replications, either the convolution or the exponent vanishes and this difficulty can be avoided. In particular, the maximal replication

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

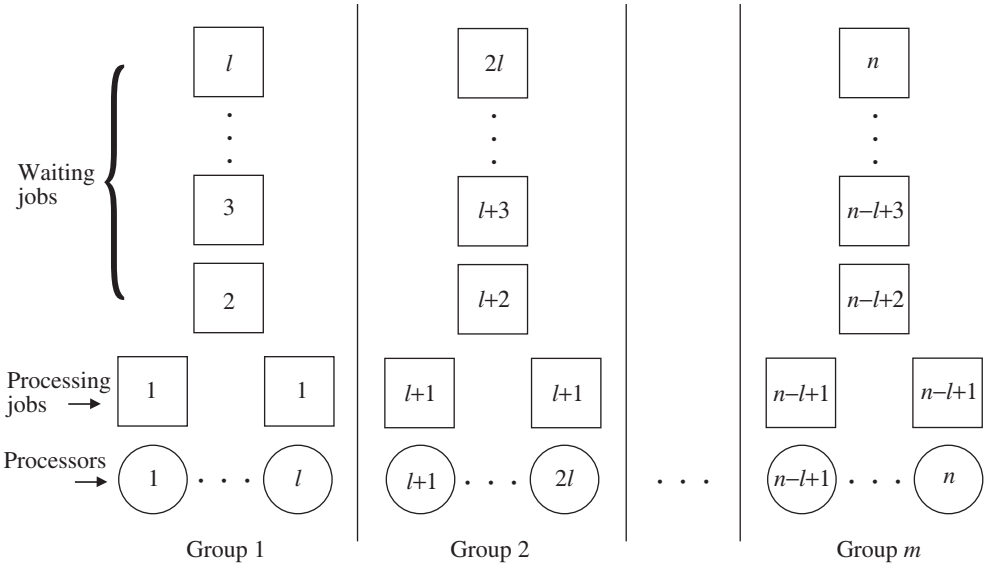


FIGURE 1: The beginning epoch of an l-replication policy where n processors are partitioned into m groups of l each. Each group has l jobs assigned, where each job is replicated l times and processed concurrently.

TABLE 1: Expected makespan for 100 jobs with log-normal service time distribution with different coefficients of variation on 100 servers. The minimum in each column is marked in bold.

l	c_v			
	0.1003	0.7951	1.3108	11.2010
1	**1.289 144**	6.194 503	14.029 006	477.300 393
2	2.148 741	3.637 286	5.118 202	27.839 179
4	3.861 831	**3.384 693**	3.369 276	4.987 980
5	4.706 410	3.524 322	**3.223 273**	3.329 165
10	8.827 199	4.429 550	3.266 562	1.282 913
20	16.837 766	6.297 849	3.974 073	0.768 918
25	20.782 238	7.178 278	4.339 773	0.691 326
50	40.108 005	11.133 747	6.012 986	0.592 256
100	77.877 122	18.012 447	8.850 298	**0.583 138**

policy ($l = n$ and $m = 1$) has makespan

$$M_{(n,n)} = \sum_{i=1}^{n} \min_{k\in\{1,\ldots,n\}} X_{i1k}$$

and the minimal replication ($l = 1$ and $m = n$) has makespan

$$M_{(1,n)} = \max_{j\in\{1,\ldots,n\}} X_{1j1}.$$

Note that when evaluating $\mathrm{E}[M_{(1,n)}]$ or $\mathrm{E}[M_{(n,n)}]$, we do not need the temporal independence assumption since, for maximal replication, we take the expectation over a sum.

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

For service time distributions where maximal replication is optimal regardless of the number of available servers, n, we have the following result.

Theorem 1. *If $M_{(r,r)} \leq_{\text{st}} M_{(k,r)}$ for all r and all factors k of r, then $M_{(1/2,n)} \geq_{\text{st}} M_{(l,n)}$ when l is even and a factor of n.*

Proof. Fix the degree of replication l and observe that *within* each group of l servers, maximal replication takes place. So the time it takes for a group to finish processing its l jobs is $M_{(l,l)}$. For $2 \leq k \leq l$, let $M^i_{(k,l)}$, $i = 1, \ldots, n/l$, be i.i.d. versions of $M_{(k,l)}$, i.e. $M^i_{(k,l)} =_{\text{st}} M_{(k,n)}$. Then

$$M_{(l,n)} =_{\text{st}} \max_{1 \leq i \leq n/l} M^i_{(l,l)}.$$

Splitting each group into two subgroups,

$$M_{(1/2l,n)} =_{\text{st}} \max_{1 \leq i \leq n/l} M^i_{(1/2,l)} \geq_{\text{st}} \max_{1 \leq i \leq n/l} M^i_{(l,l)} =_{\text{st}} M_{(l,n)}.$$

When n and l are powers of 2, we can apply Theorem 1 repetitively to get a monotonicity property.

Corollary 1. *If $M_{(n,n)} \leq_{\text{st}} M_{(l,n)}$ for all l and n such that $2 \leq l \leq n$, where l and n are powers of 2, then $M_{(l,n)}$ stochastically decreases in l for all n.*

The optimality of maximal replication for NWU service times [4] does not depend on the number of servers, n, and by applying Theorem 1 we have the following result.

Corollary 2. *If $2 \leq l \leq n$, where l and n are powers of 2, and if service times are NWU, $M_{(l,n)}$ stochastically decreases in l.*

Since NWU distributions are highly variable (e.g. the coefficient of variation is greater than or equal to 1 and larger than the exponential distribution in the convex ordering sense), this is consistent with the general observation in the following section that more variability favors more replication.

3.1. The impact of service time variability on the expected makespan

Recall that the motivation behind considering job replication was that we expected replication to have an advantage over nonreplication when the service times are highly variable. It turns out that this is indeed true under certain conditions. Here we establish an ordering between expected makespans under two convex ordered service time distributions using the same policy. But we were only able to show the result for minimal replication and maximal replication policies. We use convex ordering as our measure of relative variability. We begin with the following lemma.

Lemma 1. *The following implications hold, where M and M' are the makespans with generic service times X and X', respectively.*

1. $X' \geq_{\text{icx}} X \Rightarrow M'_{(1,n)} \geq_{\text{icx}} M_{(1,n)} \Rightarrow \text{E}[M'_{(1,n)}] \geq \text{E}[M_{(1,n)}]$.
2. $X' \geq_{\text{icv}} X \Rightarrow M'_{(n,n)} \geq_{\text{icv}} M_{(n,n)} \Rightarrow \text{E}[M'_{(n,n)}] \geq \text{E}[M_{(n,n)}]$.

Proof. For part 1, we have

$$M_{(1,n)} = \max_{k \in \{1,\ldots,n\}} X_k.$$

Letting $f(\boldsymbol{x}) = \max(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^n$, we can see that $f(\boldsymbol{x})$ is convex in $\mathbb{R}^n$ and increasing convex componentwise. Let g be an arbitrary, univariate, increasing convex function. Then it follows

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

that the composition $g \circ f : \mathbb{R}^n \to \mathbb{R}$ is also increasing convex componentwise. By changing one component at a time,

$$\begin{aligned} \mathrm{E}[g \circ f(X_1, X_2, \ldots, X_n)] &\le \mathrm{E}[g \circ f(X_1', X_2, \ldots, X_n)] \\ &\le \mathrm{E}[g \circ f(X_1', X_2', \ldots, X_n)] \\ &\le \cdots \\ &\le \mathrm{E}[g \circ f(X_1', X_2', \ldots, X_n')], \end{aligned}$$

and it follows that $M_{(1,n)} = f(X_1, \ldots, X_n) \le_{\mathrm{icx}} f(X_1', \ldots, X_n') = M'_{(1,n)}$.

For part 2,

$$M_{(n,n)} = \sum_{j=1}^{n} \min_{i \in \{1,\ldots,n\}} X_{ij}.$$

Let $f(\boldsymbol{x}) = \sum_{j=1}^{n} \min(\boldsymbol{x}_j) : \mathbb{R}^{n^2} \to \mathbb{R}$, where $\boldsymbol{x}_j \in \mathbb{R}^n$, $j = 1, 2, \ldots, n$. Since $\min(\boldsymbol{x}_j)$ is increasing concave componentwise, so is $f(\boldsymbol{x})$. For any increasing concave function g, $g \circ f$ is also increasing concave componentwise. So, by using the same reasoning as for part 1,

$$\mathrm{E}[g \circ f(\boldsymbol{X})] \le \mathrm{E}[g \circ f(\boldsymbol{X}')],$$

and we conclude that $M_{(n,n)} = f(\boldsymbol{X}) \le_{\mathrm{icv}} f(\boldsymbol{X}') = M'_{(n,n)}$.

Theorem 2. *If* $X' \ge_{\mathrm{cx}} X$ *then* $\mathrm{E}[M'_{(1,n)}] \ge \mathrm{E}[M_{(1,n)}]$ *and* $\mathrm{E}[M'_{(n,n)}] \le \mathrm{E}[M_{(n,n)}]$.

Proof. The relation $X' \ge_{\mathrm{cx}} X$ is equivalent to $X' \ge_{\mathrm{icx}} X$ and $X' \ge_{\mathrm{dcx}} X$. From Lemma 1 we have $\mathrm{E}[M'_{(1,n)}] \ge \mathrm{E}[M_{(1,n)}]$. Also, $X' \ge_{\mathrm{dcx}} X$ is equivalent to $X' \le_{\mathrm{icv}} X$, so $\mathrm{E}[M'_{(n,n)}] \le \mathrm{E}[M_{(n,n)}]$ follows from Lemma 1.

We can conclude that minimal replication performs better for less variable service time distributions and that maximal replication performs better for more variable service time distributions. Theorem 2 also implies that if the maximal replication policy is better than the minimal replication policy for service time X, then the same is true for service times larger than X in the convex sense.

By the optimality of maximal replication for exponential service time distributions [4], it is possible to show that maximal replication is better than minimal replication for service times that are more variable than the exponential.

Corollary 3. *If X is HNWUE (see Definition 7) then maximal replication yields a smaller makespan than minimal replication.*

3.2. Geometric service time distributions

Suppose that the X_{ijk} are i.i.d. geometric(p). Then,

$$\min_{1 \le k \le l} X_{ijk} \sim \text{geometric}(1 - (1-p)^l)$$

and

$$Y_l := \sum_{i=1}^{l} \min_{1 \le k \le l} X_{ijk} \sim \text{negative binomial}(l, 1 - (1-p)^l),$$

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

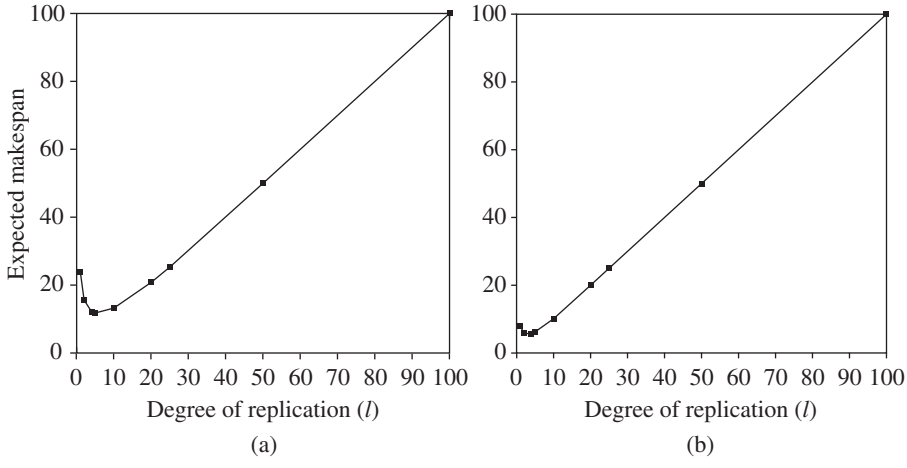


FIGURE 2: Expected makespans under geometric(p) service times for $n = 100$ and (a) $p = 0.2$, (b) $p = 0.5$.

which has the distribution function

$$F_{Y_l}(x) = \sum_{j=l}^{x} \binom{x}{j} (1 - (1-p)^l)^j (1-p)^{l(x-j)}.$$

Finally, the distribution function of $M_{(l,n)}$ can be found:

$$\mathrm{P}(M_{(l,n)} \le x) = [F_{Y_l}(x)]^{n/l},$$

and its mean is

$$\mathrm{E}[M_{(l,n)}] = \sum_{x=1}^{\infty} [1 - \{F_{Y_l}(x)\}^{n/l}].$$

It can be seen that, for geometric(0.2) service times, using five-replication reduces the expected makespan of 100 jobs by half compared to scheduling the jobs in parallel. See Figure 2.

This result may seem like a contradiction to the optimality of maximal replication for NWU service times since the geometric distribution is memoryless. But, if we extend the support from the nonnegative integers to the nonnegative reals, the geometric distribution is not NWU. It is required in [4] that the support be the real line for the optimality result to hold. It may also seem to contradict the result of Borst *et al.* [1], but they permitted the degree of replication to change over time.

3.3. Bernoulli service times

Consider the case when the service time distribution has probability mass accumulated on two points. Such two-point distributions are interesting because they are easy to analyze and yet they capture the effect of individual grid resources being in two states (e.g. busy/idle). They also provide insight into the analysis of having general bimodal service time distributions. We also use some of the theory we develop here for general bounded distributions in Subsection 3.1.

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

We first consider the Bernoulli distribution taking values 0 or 1 and later extend to the case where the distribution is shifted away from 0.

Suppose that the service time distribution of a job is Bernoulli with success probability p:

$$X = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1 - p. \end{cases}$$

The expected makespan for maximal replication is

$$\mathrm{E}\left[\sum_{i=1}^{n} \min_{k\in\{1,\dots,n\}} X_{i1k}\right] = np^n$$

and for minimal replication it is

$$\mathrm{E}\left[\max_{j\in\{1,\dots,n\}} X_{1j1}\right] = 1 - (1-p)^n.$$

We observe that, for sufficiently large n, maximal replication will perform better than minimal replication. This result relies heavily on the fact that X_{ijk} has a nonzero probability of taking the value 0, so, for a large number of replications, the minimum of the service times will be 0 with high probability.

Let us now find an expression for the expected makespan for l-replication. Since

$$\min_{k\in\{1,\dots,l\}} X_{ijk} = \begin{cases} 1 & \text{with probability } p^l, \\ 0 & \text{with probability } 1 - p^l, \end{cases}$$

we can see that $\sum_{i=1}^{l} \min_{k\in\{1,\dots,l\}} X_{ijk} := Y_j$ has a binomial distribution, $B(l, p^l)$. The expected makespan is

$$\begin{aligned} \mathrm{E}[M_{(l,n)}] &= \mathrm{E}\left[\max_{j\in\{1,\dots,m\}} Y_j\right] \\ &= \sum_{y=1}^{l} \mathrm{P}\left(\max_{j\in\{1,\dots,m\}} Y_j \geq y\right) \\ &= \sum_{y=1}^{l} \left\{1 - \mathrm{P}\left(\max_{j\in\{1,\dots,m\}} Y_j \leq y-1\right)\right\} \\ &= \sum_{y=1}^{l} \{1 - [\mathrm{P}(Y_1 \leq y-1)]^m\} \\ &= l - \sum_{y=0}^{l-1} \{F(y)\}^m, \end{aligned}$$

where $F(\cdot)$ is the distribution function of $B(l, p^l)$. It is easy to calculate this value for various n, l, and p (see Figure 3). The shapes of the graphs for different combinations of (n, p) were all either monotonically increasing, monotonically decreasing, or increasing and then decreasing in l, which we conjecture to be true in general.

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

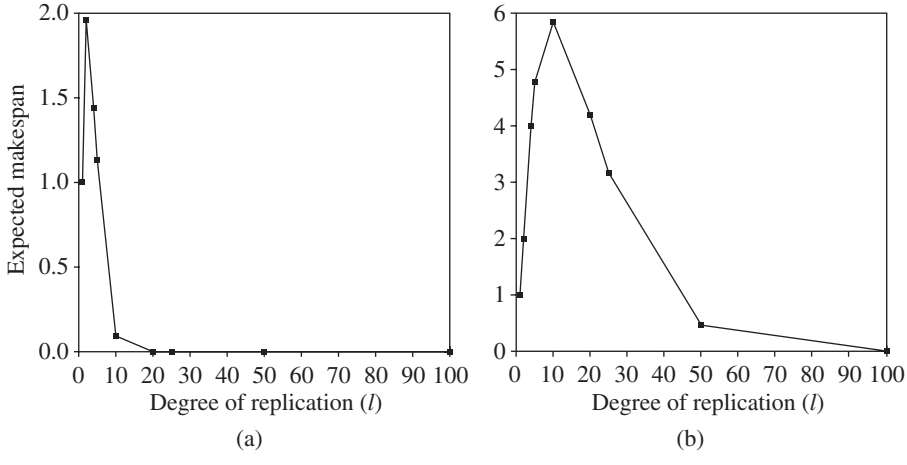


FIGURE 3: Expected makespans under Bernoulli(p) service times for $n = 100$ and (a) $p = 0.5$, (b) $p = 0.9$.

Conjecture 1. *For Bernoulli service times,* $\mathrm{E}[M_{(l,n)}] \geq \min\{\mathrm{E}[M_{(1,n)}], \mathrm{E}[M_{(n,n)}]\}$, *i.e. either maximal replication or minimal replication is optimal.*

As p approaches 1, it is evident that the maximal replication makespan will approach n, which is larger than 1, the upper bound of the minimal replication makespan. By the results of Figure 3, there is reason to believe that there is a certain upper threshold for p such that maximal replication is better than minimal replication for all n.

Lemma 2. *When service times are i.i.d. Bernoulli(p) on a system with n jobs and n servers, there exists a* c_n, $0 \leq c_n \leq 1$, *such that the expected makespan for maximal replication is no larger than that of minimal replication if and only if* $0 \leq p \leq c_n$.

Proof. The expected makespan is np^n for maximal replication and $1-(1-p)^n$ for minimal replication. For $n = 1$, minimal and maximal replications are equivalent for all p. We only need to verify the $n \geq 2$ case. For fixed $n \geq 2$, let us consider the function $f_n(p) = np^n + (1-p)^n - 1$ on the interval $[0, 1]$. Note that $f_n(0) = 0$ and $f_n(1) = n - 1 > 0$, so we can restrict our attention to the form of the function $f_n(p)$ on the open interval $(0, 1)$. Observe the second derivative of $f_n(p)$:

$$f_n''(p) = n^2(n-1)p^{n-2} + n(n-1)(1-p)^{n-2} > 0;$$

we see that $f_n(p)$ is convex on $(0, 1)$ for $n \geq 2$. Also, $p = 1/(n^{1/(n-1)} + 1)$ is always a root of the equation $f_n'(p) = 0$, and we conclude that $f_n(p)$ has a minimum on $(0, 1)$, and since $f_n(0) = 0$, the minimum value must be strictly smaller than 0. The convexity of $f_n(p)$, $f_n(0) = 0$, and the existence of a point $0 < a < 1$, where $f_n(a) < 0$, imply that the function has a single root, c_n, in the interval $(0, 1)$, so it can be partitioned as $(0, c_n] \cup (c_n, 1)$, where $f_n(p) \leq 0$ on $(0, c_n]$ and $f_n(p) > 0$ on $(c_n, 0)$.

Theorem 3. *For* $n \geq 2$, *the sequence* $\{c_n\}$ *described in Lemma 2 is increasing in n.*

Proof. Observe that the second term of $f_n(p) = np^n + (1-p)^n - 1$ decreases in n and consider the first term np^n. Letting n vary continuously while holding p fixed, np^n is decreasing

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

on $n \geq 1/\log(1/p)$ or, equivalently, $p \leq \mathrm{e}^{-1/n}$. Defining $d_n = \mathrm{e}^{-1/n}$, $f_n(p)$ is decreasing in n for all $p \leq d_n$. Note that d_n increases in n. If $c_n \leq d_n$ for $n \geq 2$ then we can conclude that c_n is increasing in n. To see if this condition holds, we exploit the fact that c_n is an upcrossing point of $f_n(p)$ and evaluate the sign of $f_n(d_n)$. Now $c_n \leq d_n$ is equivalent to $f_n(d_n) \geq 0$, where

$$f_n(d_n) = \frac{n}{\mathrm{e}} - 1 + (1 - \mathrm{e}^{-1/n})^n.$$

Since the third term is always positive, it can be seen that at least for $n \geq 3$, $f_n(d_n)$ is nonnegative. It remains to verify that $c_2 \leq c_3$, where $c_2 = \frac{2}{3}$. Now, $c_2 \leq c_3$ is equivalent to $f_3(c_2) \leq 0$, which is indeed true since $f_3(\frac{2}{3}) = -\frac{2}{27}$.

Corollary 4. *When service times are i.i.d.* Bernoulli(p) *with n jobs and n servers, if* $p \leq \frac{2}{3}$ *then the expected makespan of maximal replication is no larger than that of minimal replication for all* $n \geq 1$.

Proof. Since $c_2 = \frac{2}{3}$ and c_n is increasing for $n \geq 2$, $f_n(c_2) \leq 0$ for all $n \geq 2$. Also, since $f_1(c_2) = 0$, $f_n(c_2) \leq 0$ for all $n \geq 1$. Consequently, for all $p \leq c_2$ and all $n \geq 1$, $f_n(p) \leq 0$.

Combining Corollary 4 with our earlier conjecture, we believe that maximal replication is optimal whenever $p \leq \frac{2}{3}$ for Bernoulli(p) service times.

Note that the coefficient of variation of Bernoulli(p) is $\sqrt{(1-p)/p}$, so the smaller p is the more variable X is in this sense, and again we have a result consistent with our general observation that more variability favors more replication.

For service time distributions that have a bounded support, we can use Corollary 2 and the results for two-point distributions to compare the performance of minimal and maximal replications. Without loss of generality, we may consider RVs bounded by 0 from below since adding a constant to both of the RVs under comparison does not affect the convex order relation.

Lemma 3. *Let X be an RV with support* $[0, b]$ *and mean* μ*, and let Y be the corresponding two-point distribution on the points* $\{0, b\}$ *with the same mean. Then* $X \geq_{\mathrm{cx}} Y$.

Proof. We use the tail characterization of convex ordering, i.e. $X \geq_{\mathrm{cx}} Y$ if and only if $\int_a^\infty \bar{F}_X(x)\,\mathrm{d}x \geq \int_a^\infty \bar{F}_Y(x)\,\mathrm{d}x$ for all a. Suppose that $\mathrm{P}(Y = b) = p$, so that $\mathrm{E}[X] = \mathrm{E}[Y] = pb$. Now consider the function

$$h(y) = \int_y^b [\bar{F}_X(x) - \bar{F}_Y(x)]\,\mathrm{d}x = \int_y^b [\bar{F}_X(x) - p]\,\mathrm{d}x,$$
$$h'(y) = -\bar{F}_X(y) + p.$$

Observe that, owing to the monotonicity of F_X, the sign of $h'(y)$ changes from negative to positive as y increases from 0 to b. Also, since $h(0) = h(b) = 0$, we conclude that $h(y) \leq 0$ on $[0, b]$. By the tail characterization of increasing convex ordering, $X \geq_{\mathrm{icx}} Y$. Since $\mathrm{E}[X] = \mathrm{E}[Y]$, we have $X \geq_{\mathrm{cx}} Y$.

This lemma together with Theorem 2 provides a link between the performance of a bounded service time and a two-point distributed service time.

Corollary 5. *Let X be a bounded random variable denoting the service time, and let* X' *be the two-point distribution with the same mean having mass on the two boundary points of X. If minimal replication yields a smaller makespan than maximal replication for* X'*, then minimal replication yields a smaller makespan than maximal replication for X as well.*

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

By using the result of Lemma 2, it is a simple task to compare the performance of minimal replication to the performance of maximal replication, provided that c_n can be computed.

Corollary 6. *Let X be a random service time with upper bound b and mean μ. If $\mu/b > c_n$ then minimal replication yields a smaller makespan than maximal replication, where c_n is as defined in Lemma 2.*

Proof. Since b is an upper bound and 0 is a (trivial) lower bound for X, we may consider a two-point distribution Y on $\{0, b\}$, where $\mathrm{P}(Y = b) = \mu/b$. Then $\mathrm{E}[Y] = \mu$, and if $\mu/b > c_n$, minimal replication is better than maximal replication for service times Y, and, consequently, the same holds for X.

3.4. Shifted Bernoulli service times

Let us now consider the more realistic case where the lower bound of the two-point distribution is δ $(0 < \delta < 1)$ away from 0. Now we may have an optimal l that is strictly between 1 and n. Let the service time be X, where

$$X = \begin{cases} 1 & \text{with probability } p, \\ \delta & \text{with probability } 1 - p, \end{cases}$$

and

$$Z = \frac{X - \delta}{1 - \delta}$$

is now a Bernoulli(p) RV. Expressing X in terms of Z and substituting it into (1), we obtain an expression for the l-replication expected makespan for service time X:

$$\begin{aligned} \mathrm{E}[M_{(l,n)}] &= \mathrm{E}\left[\max_{j\in\{1,\dots,m\}} \sum_{i=1}^{l} \min_{k\in\{1,\dots,l\}} X_{ijk}\right] \\ &= \mathrm{E}\left[\max_{j\in\{1,\dots,m\}} \sum_{i=1}^{l} \min_{k\in\{1,\dots,l\}} \{(1-\delta)Z_{ijk} + \delta\}\right] \\ &= (1-\delta)\,\mathrm{E}\left[\max_{j\in\{1,\dots,m\}} \sum_{i=1}^{l} \min_{k\in\{1,\dots,l\}} Z_{ijk}\right] + \delta l. \end{aligned}$$

Note that the expectation part of the last expression is the makespan for the l-replication with Bernoulli service times taking values 0 or 1. We may use the result derived for the Bernoulli service times to plot this expression for various l, m, n, and p (see Figure 4).

It is interesting to see in these examples that there exist cases where the optimal degree of replication occurs at a nonextreme l value.

When we have a strictly positive lower bound on the service time, processing jobs in sequence creates an overhead which can be avoided by using minimal replication. So we can expect minimal replication to be favorable over maximal replication in this case.

Theorem 4. *For the shifted Bernoulli service time distribution, the expected makespan of minimal replication is smaller or equal to that of maximal replication for all $0 \le p \le 1$ if*

$$\delta \ge 1 - \frac{1 - 1/n}{1 - \hat{p}^{n-1}},$$

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

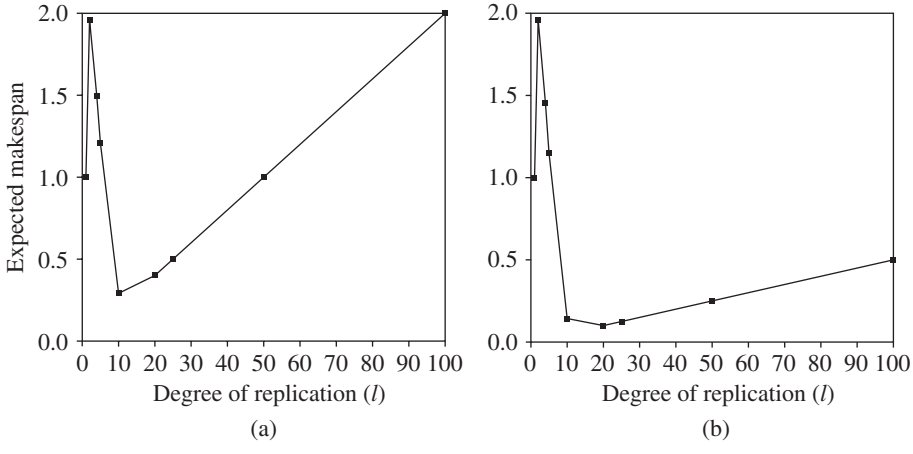


FIGURE 4: Expected makespan for $n = 100$ and (a) $(p, \delta) = (0.5, 0.02)$, (b) $(p, \delta) = (0.5, 0.005)$. Nonextreme solutions where 10 or 20 replications of each job is optimal.

where

$$\hat{p} = \frac{1}{1 + n^{1/(n-1)}}$$

when $n \geq 2$.

Proof. The expected makespan of maximal and minimal replications is $np^n + n\delta(1 - p^n)$ and $1 - (1 - \delta)(1 - p)^n$, respectively. For $n = 1$, these share the same value. Henceforth, we assume that $n \geq 2$. For fixed n, let $f_n(p) = \mathrm{E}[M_{(n,n)} - M_{(1,n)}]$, where

$$\begin{aligned} f_n(p) &= (1 - \delta)\{(1 - p)^n - n(1 - p^n)\} + n - 1, \\ f_n'(p) &= (1 - \delta)\{-n(1 - p)^{n-1} + n^2 p^{n-1}\}, \\ f_n''(p) &= (1 - \delta)\{n(n - 1)(1 - p)^{n-2} + n^2(n - 1)p^{n-2}\}. \end{aligned}$$

Observe that $f_n''(p) > 0$ since $0 < \delta < 1$ and $n \geq 2$. So $f_n(p)$ is convex in p. The positive real root of the equation $f_n'(p) = 0$ is

$$\hat{p} = \frac{1}{1 + n^{1/(n-1)}},$$

which lies strictly between 0 and 1. Evaluating the minimum value of $f_n(p)$,

$$f_n(\hat{p}) = n(1 - \delta)(\hat{p}^{n-1} - 1) + n - 1.$$

A condition equivalent to $f_n(\hat{p}) > 0$ is

$$\delta \geq 1 - \frac{1 - 1/n}{1 - \hat{p}^{n-1}}.$$

In the above proof, note that $f_n(0) = \delta(n - 1) > 0$, $f_n(1) = n - 1 > 0$, and $f_n(p)$ is convex. So minimal replication is better than maximal replication near extreme values of p, which is also when the variance of the service time is small. Also, near $p = \hat{p}$, depending on n and δ, either maximal replication is better than minimal replication or at least the performance gap between them is minimized.

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

3.5. General discrete service time distributions

For the general discrete distribution defined on a finite number of points, we show how to compute $\mathrm{E}[M_{(l,n)}]$ exactly. Let X have the discrete distribution

$$X = \begin{cases} a_1 & \text{with probability } p_1, \\ \vdots & \\ a_k & \text{with probability } p_k, \end{cases}$$

let $\boldsymbol{a}^\top = (a_1, \ldots, a_k)$, where $(a_1 \le a_2 \le \cdots \le a_k)$, and let $\boldsymbol{p}^\top = (p_1, \ldots, p_k)$, where $\boldsymbol{p} > \boldsymbol{0}$ and $\boldsymbol{a} \ge \boldsymbol{0}$. Then X can be represented as $X = \boldsymbol{a}^\top \boldsymbol{Y}$, where $\boldsymbol{Y}$ is distributed multinomial$(1, \boldsymbol{p})$. Let $\{X_i; i = 1, \ldots, l\}$ be i.i.d. samples of X. We have

$$\mathrm{P}(\min\{X_1, \ldots, X_l\} \ge a_i) = \left(\sum_{j=i}^{k} p_j\right)^l,$$

$$\mathrm{P}(\min\{X_1, \ldots, X_l\} = a_i) = \left(\sum_{j=i}^{k} p_j\right)^l - \left(\sum_{j=i+1}^{k} p_j\right)^l = \tilde{p}_i,$$

so $\min\{X_1, \ldots, X_l\} \stackrel{\mathrm{D}}{=} \boldsymbol{a}^\top \tilde{\boldsymbol{Y}}$, where $\tilde{\boldsymbol{Y}}$ is distributed multinomial$(1, \tilde{\boldsymbol{p}})$. Here '$\stackrel{\mathrm{D}}{=}$' denotes equality in distribution. We have

$$\begin{aligned} \sum_{i=1}^{l} \min\{X_1^i, \ldots, X_l^i\} &\stackrel{\mathrm{D}}{=} \sum_{i=1}^{l} \boldsymbol{a}^\top \tilde{\boldsymbol{Y}}_i \\ &\stackrel{\mathrm{D}}{=} \boldsymbol{a}^\top \sum_{i=1}^{l} \tilde{\boldsymbol{y}}_i \\ &\stackrel{\mathrm{D}}{=} \boldsymbol{a}^\top \boldsymbol{Z}, \end{aligned}$$

where $\boldsymbol{Z}$ is multinomial$(l, \tilde{\boldsymbol{p}})$. It remains to find the distribution function of the RV $\boldsymbol{a}^\top \boldsymbol{Z}$:

$$\mathrm{P}(\boldsymbol{a}^\top \boldsymbol{Z} \le x) = \sum_{\substack{\boldsymbol{1}^\top z = l \\ z \ge 0}} f(z) 1_{\{\boldsymbol{a}^\top z \le x\}}(z),$$

where $f(z)$ is the probability mass function of multinomial$(l, \tilde{\boldsymbol{p}})$. Note that the number of summands is of $O(l^{k-1})$. So when k is fixed, the time complexity of computing the distribution is polynomial in l. Let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_m$ be i.i.d. samples with the same distribution as $\boldsymbol{Z}$. Finally, the distribution function of $\{\max \boldsymbol{a}^\top \boldsymbol{Z}_1, \ldots, \boldsymbol{a}^\top \boldsymbol{Z}_m\}$ is $\{\mathrm{P}(\boldsymbol{a}^\top \boldsymbol{Z} \le x)\}^m$ and the expectation can be computed by summing the tail probability. So,

$$\mathrm{E}[M_{(l,n)}] = \sum_{x=0}^{la_k} [1 - \{\mathrm{P}(\boldsymbol{a}^\top \boldsymbol{Z} \le x)\}^m].$$

Example 1. The expected makespan of a system with eight servers and service time distribution

$$\boldsymbol{a}^\top = [1, 2, 3, 4], \qquad \boldsymbol{p}^\top = \left[\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{8}\right],$$

is $\mathrm{E}[M_{(1,8)}] = 3.552\,372$, $\mathrm{E}[M_{(2,8)}] = 3.627\,524$, $\mathrm{E}[M_{(4,8)}] = 4.480\,234$, and $\mathrm{E}[M_{(8,8)}] = 8.031\,373$.

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

We can use an appropriate discrete distribution to obtain an approximation for general bounded continuous distributions. A difficulty is that the precision of the approximation depends on k, the number of points in the discrete RV's support, which impacts on the complexity exponentially. But, since any bounded distribution can be stochastically bounded by an appropriate bounded discrete distribution, and a fairly good bound can be obtained with a relatively small k, this method can be useful for worst-case analysis.

3.6. Bounds for the l-replication makespan

For most service time distributions, finding the expected makespan for the general l-replication policy is analytically intractable. However, in some cases, by using its lower and upper bounds, we can determine whether the optimal number of replications is a nonextreme replication policy (i.e. $2 \le l \le n-1$) by comparing it to the expected makespan of the analytically tractable cases of $l = 1$ and $l = n$. If, for example, there is some $2 \le l \le n-1$ that has an upper bound smaller than both the makespan of minimal and maximal replications, can conclude that the optimal number of replications is somewhere in the middle. We provide a lower bound and an upper bound of $M_{(l,n)}$, the l-replication makespan.

Theorem 5. *For $M_{(l,n)}$, the following inequality holds:*

$$l\,\mathrm{E}[X_{(1,l)}] \le \mathrm{E}[M_{(l,n)}] \le l\,\mathrm{E}[X_{(n-l+1,n)}],$$

where $X_{(i,j)}$ is the ith order statistic (in increasing order) from a sample of size j.

Proof. By applying Jensen's inequality, we obtain a lower bound for the l-replication makespan:

$$\mathrm{E}\left[\max_{j\in\{1,\dots,m\}}\sum_{i=1}^{l}\min_{k\in\{1,\dots,l\}}X_{ijk}\right] \ge \max_{j\in\{1,\dots,m\}}\mathrm{E}\left[\sum_{i=1}^{l}\min_{k\in\{1,\dots,l\}}X_{ijk}\right] = l\,\mathrm{E}[X_{(1,l)}].$$

For an upper bound, we have

$$\begin{aligned}\mathrm{E}\left[\max_{j\in\{1,\dots,m\}}\sum_{i=1}^{l}\min_{k\in\{1,\dots,l\}}X_{ijk}\right] &\le \mathrm{E}\left[\sum_{i=1}^{l}\max_{j\in\{1,\dots,m\}}\min_{k\in\{1,\dots,l\}}X_{ijk}\right]\\ &\le l\,\mathrm{E}[X_{(n-l+1,n)}].\end{aligned}$$

The first inequality holds because, for any i.i.d. sequence $Y_{ij},\ i = 1,\dots,m,\ j = 1,\dots,l$,

$$\max_{i\in\{1,\dots,m\}}\sum_{j=1}^{l}Y_{ij} = \sum_{j=1}^{l}Y_{i^*,j} \le_{\mathrm{st}} \sum_{j=1}^{l}\max_{i\in\{1,\dots,m\}}Y_{ij},$$

where i^* is the maximizing index.

For the second inequality, consider an i.i.d. sample of size n laid out as a double array having l rows and m columns, where $lm = n$. Then

$$\max_{j\in\{1,\dots,m\}}\min_{k\in\{1,\dots,l\}}X_{ijk} \le_{\mathrm{st}} X_{(n-l+1)}$$

can be verified by the following procedure. Take the minimum value within each column to obtain m numbers. Since the largest of the m numbers cannot be larger than the other $l-1$ numbers in its column, the largest it can be is the lth largest of the n. The inequality follows by combining this with a sample path argument for the stochastic order.

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

When the support of the service time RV does not contain 0, we have the following result.

Lemma 4. *Let the service time distribution X be a strictly positive RV such that $X > c > 0$. Then*

$$\mathrm{E}[M_{(l,n)}] \geq lc.$$

Proof. Let X_{ijk} be i.i.d. RVs having the same distribution as X. Since $X > c$,

$$\min_{k\in\{1,2,\ldots,l\}} X_{ijk} > c \quad \text{for each } i, j,$$

$$\sum_{j=1}^{l} \min_{k\in\{1,2,\ldots,l\}} X_{ijk} > lc \quad \text{for each } i,$$

$$M_{(l,n)} = \max_{i\in\{1,2,\ldots,m\}} \sum_{j=1}^{l} \min_{k\in\{1,2,\ldots,l\}} X_{ijk} > lc.$$

We can see here that, for sufficiently large n, more replication will result in a higher makespan when service times are bounded away from 0.

3.7. Asymptotic properties

Since grid computing involves a large set of servers, we study the asymptotic properties of the makespan RV as $n \to \infty$. The makespan RV for maximal replication,

$$\sum_{j=1}^{n} \min\{X_{1,j}, \ldots, X_{n,j}\},$$

is of particular interest since, for large n, it is a sum of many small RVs and we can expect it to have some limiting properties.

Theorem 6. *Let $\{Y_{i,j}; i, j = 1, \ldots, n\}$ be i.i.d. exponential(λ) RVs, and let $\{X_{i,j}; i, j = 1, \ldots, n\}$ be an i.i.d. random sample of size n^2 from a distribution stochastically smaller than exponential(λ), i.e. $X_{i,j} \leq_{\mathrm{st}} Y_{i,j}$. Then,*

$$\lim_{n\to\infty} \mathrm{P}\left(\sum_{j=1}^{n} \min\{X_{1,j}, \ldots, X_{n,j}\} \leq \frac{1}{\lambda}\right) = 1.$$

Proof. Since, for each (i, j), $X_{i,j} \leq_{\mathrm{st}} Y_{i,j}$,

$$\mathrm{P}(X_{i,j} > a) \leq \mathrm{P}(Y_{i,j} > a) \quad \text{for } i, j = 1, \ldots, n,$$

$$\prod_{i=1}^{n} \mathrm{P}(X_{i,j} > a) \leq \prod_{i=1}^{n} \mathrm{P}(Y_{i,j} > a),$$

and, consequently,

$$\min\{X_{1,j}, \ldots, X_{n,j}\} \leq_{\mathrm{st}} \min\{Y_{1,j}, \ldots, Y_{n,j}\} \quad \text{for each } j.$$

Note that the right-hand side has distribution exponential($n\lambda$). Summing each side with respect to j, we have

$$\sum_{j=1}^{n} \min\{X_{1,j}, \ldots, X_{n,j}\} \leq_{\mathrm{st}} \sum_{j=1}^{n} \min\{Y_{1,j}, \ldots, Y_{n,j}\}.$$

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

Let us represent the distribution functions of the left-hand side RV as F_n and the right-hand side RV as G_n. Then an equivalent expression for the above stochastic order is $F_n(x) \geq G_n(x)$. Note that G_n is the distribution function of Erlang$(n, n\lambda)$, which converges to the constant $1/\lambda$ in distribution as $n \to \infty$. Since $F_n(x) \geq G_n(x)$ holds pointwise for all n and the (pointwise) convergence of $G_n(x)$ is guaranteed,

$$\liminf_{n\to\infty} F_n(x) \geq G_\infty(x)$$

for each x in the domain. This implies that, for $x \geq 1/\lambda$, $\lim_n F_n(x) = 1$, from which we can conclude that

$$\lim_{n\to\infty} \mathrm{P}\left(\sum_{j=1}^{n} \min\{X_{1,j}, \ldots, X_{n,j}\} \leq \frac{1}{\lambda}\right) = 1.$$

Note that in the above theorem we do not make any assumptions on the convergence of $\sum_{j=1}^{n} \min\{X_{1,j}, \ldots, X_{n,j}\}$. What this theorem tells us is that if the service time has an exponential RV as a stochastic upper bound, the probability of the maximal replication makespan being less than some constant approaches 1 as n goes to ∞.

Recall that, for NWU service times, a new service time is stochastically smaller than the remaining service time after some service has been completed. Also, from [4], if the service time distribution is NWU, maximal replication is optimal, i.e. the expected makespan is minimized when $l = n$. We also have the following result.

Theorem 7. *Let* $\{X_n; n = 1, 2, \ldots\}$ *be i.i.d. NWU RVs. Then*

$$\lim_{n\to\infty} \mathrm{P}(M_{(n,n)} \leq \mathrm{E}[X_1]) = 1.$$

Proof. Let H be the cumulative hazard rate function of X. Recall that X is NWU if and only if H is subadditive, i.e.

$$H(nx) \leq nH(x) \quad \Longleftrightarrow \quad \bar{F}(nx) \geq \bar{F}^n(x) \quad \Longleftrightarrow \quad 1 - \bar{F}^n(x) \geq F(nx).$$

Observe that the left-hand side of the last inequality is the distribution function of $\min\{X_1, \ldots, X_n\}$ and that the right-hand side is the distribution function of X/n. So we have $\min\{X_1, \ldots, X_n\} \leq_{\mathrm{st}} X/n$ and, consequently, $\sum_{j=1}^{n} \min\{X_{1,j}, \ldots, X_{n,j}\} \leq_{\mathrm{st}} \sum_{j=1}^{n} X_j/n$. The right-hand side of the inequality converges to $\mathrm{E}[X]$ with probability 1 by the small law of large numbers and the result follows.

A more refined result for the asymptotic distribution of the maximal replication makespan can be found using the central limit theorem for triangular arrays of RVs, i.e. if $\{Z_{ni}; n = 1, 2, \ldots; i = 1, \ldots, n\}$ is a triangular array satisfying $\mathrm{E}[Z_{ni}] = 0$, $\mathrm{var}(Z_{ni}) = \sigma_n^2$, and $\lim_n \sigma_n^2 \to \sigma^2 > 0$, then $n^{-1/2} \sum_{i=1}^{n} Z_{ni} \xrightarrow{\mathrm{D}} N(0, \sigma^2)$, where '$\xrightarrow{\mathrm{D}}$' denotes convergence in distribution. Let the X_i^j, $i, j = 1, \ldots, n$, be i.i.d. RVs representing the service time and consider the RV

$$X_{nj} := \frac{\min\{X_1^j, \ldots, X_n^j\}}{b_n}$$

for $j = 1, \ldots, n$ and some given sequence $\{b_n\}$. Then $\{X_{nj}; n = 1, 2, \ldots; j = 1, \ldots, n\}$ forms a triangular array. For the central limit theorem to hold, it is sufficient to have

$$\mathrm{var}(X_{nj}) = \frac{\mathrm{var}(\min\{X_1^j, \ldots, X_n^j\})}{b_n^2} \to \sigma^2,$$

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

where $\sigma^2 > 0$ is a constant. Then

$$\sum_{j=1}^{n} X_{nj} - n\,\mathrm{E}[X_{nj}] \to N(0, \sigma^2),$$

or if we let $\alpha_n := \mathrm{E}[\min\{X_1^j, \ldots, X_n^j\}]$,

$$\frac{1}{b_n}\sum_{j=1}^{n} \min\{X_1^j, \ldots, X_n^j\} - \frac{n}{b_n}\alpha_n \to N(0, \sigma^2).$$

So, for large n, the maximal replication makespan is approximately distributed as $N(n\alpha_n, b_n^2\sigma^2)$. Note that the sequence $\{b_n\}$ is determined by how fast the variance of $\min\{X_1^j, \ldots, X_n^j\}$ approaches 0 (provided that X_i^j is bounded from below).

For example, when $X_{i,j}$ is uniform(0,1)-distributed,

$$\mathrm{E}[\min\{X_1^j, \ldots, X_n^j\}] = \frac{1}{n+1},$$

$$\mathrm{var}(\min\{X_1^j, \ldots, X_n^j\}) = \frac{2}{(n+2)(n+1)} - \left(\frac{1}{n+1}\right)^2 = \frac{n}{(n+2)(n+1)^2},$$

so $\alpha_n = 1/(n+1)$, $b_n = 1/n$, and $\sigma^2 = 1$. Note that, for uniform service times, minimal replication is preferred over maximal replication because, while there is no difference in the expected makespan, the variance of maximal replication is larger than that of minimal replication by a factor of n.

3.8. The impact of intertemporal dependence of service times

Now, we consider the case where there is positive dependence among successive service times for each server. A scenario where successive service times are positively dependent is when service times depend on how 'busy' the grid server is or when service times for our jobs can be modeled as waiting times in the underlying queue of work at the local server.

We compare the l-replication makespan for two cases where one ($\boldsymbol{X}$) has a stronger temporal positive dependence among successive service times on each server than the other ($\boldsymbol{Y}$). For the two cases, let $\boldsymbol{X}_k^j$ and $\boldsymbol{Y}_k^j$ denote l-dimensional random vectors of service times on server k of group j, where $k = 1, \ldots, l$ and $j = 1, \ldots, m$, so that $\boldsymbol{X}_k^j = (X_{1jk}, X_{2jk}, \ldots, X_{ljk})^\top$ and $\boldsymbol{Y}_k^j = (Y_{1jk}, Y_{2jk}, \ldots, Y_{ljk})^\top$. For any (fixed) group j, let us assume that $\boldsymbol{X}_k^j \geq_{\mathrm{sm}} \boldsymbol{Y}_k^j$ for all servers $k = 1, \ldots, l$ and that the $\boldsymbol{X}_k^j$ as well as the $\boldsymbol{Y}_k^j$ are independent across k so that servers are independent of each other but jobs on the same server have positively dependent service times. See Figure 5 for an illustration. Then we are comparing

$$M_{(l,n)} = \max_{1\leq j\leq m} \sum_{i=1}^{l} \min_{1\leq k\leq l} X_{ijk} \tag{2}$$

with

$$M'_{(l,n)} = \max_{1\leq j\leq m} \sum_{i=1}^{l} \min_{1\leq k\leq l} Y_{ijk}. \tag{3}$$

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

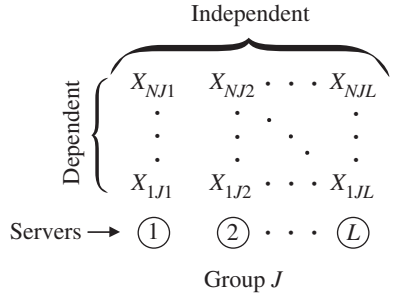


FIGURE 5: Here X_{iJk} is the service time of a replica of job i of (fixed) group J assigned to server k. There is positive dependence in the temporal (vertical) direction but independence across servers.

First we show that temporal dependence of service times is preserved after taking the minimum across servers, i.e. if we hold $j = J$ fixed,

$$\left(\min_{1\le k\le m} X_{1Jk}, \dots, \min_{1\le k\le m} X_{lJk}\right) \ge_{\text{sm}} \left(\min_{1\le k\le m} Y_{1Jk}, \dots, \min_{1\le k\le m} Y_{lJk}\right)$$

holds when $\boldsymbol{X}_k^J \ge_{\text{sm}} \boldsymbol{Y}_k^J$. To establish this relation, we need Corollary 9.A.10 of [10], which we restate for convenience.

Theorem 8. ([10, Corollary 9.A.10].) *Let $\boldsymbol{X} = (X_1, \dots, X_n)^\top$ and $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top$ be two random vectors such that $\boldsymbol{X} \ge_{\text{sm}} \boldsymbol{Y}$, and let $\boldsymbol{Z}$ be an m-dimensional random vector which is independent of $\boldsymbol{X}$ and $\boldsymbol{Y}$. Then*

$$(h_1(X_1, \boldsymbol{Z}), \dots, h_n(X_n, \boldsymbol{Z})) \ge_{\text{sm}} (h_1(Y_1, \boldsymbol{Z}), \dots, h_n(Y_n, \boldsymbol{Z}))$$

whenever $h_i(x, \boldsymbol{z})$, $i = 1, \dots, n$, are all increasing or all decreasing in x for every $\boldsymbol{z}$.

We use Theorem 8 to show the following generalization of Theorem 9.A.12 of [10].

Theorem 9. *Let $\boldsymbol{X}_k$ and $\boldsymbol{Y}_k$ for $k = 1, \dots, L$ be N-dimensional random vectors such that $\boldsymbol{X}_k$ and $\boldsymbol{Y}_k$ are independent over k. Let $f_i : \mathbb{R}^L \to \mathbb{R}$, $i = 1, \dots, N$, be a continuous increasing function. If $\boldsymbol{X}_k \ge_{\text{sm}} \boldsymbol{Y}_k$ for all k then*

$$\begin{aligned}&(f_1(X_{11}, \dots, X_{1L}), \dots, f_N(X_{N1}, \dots, X_{NL}))\\ &\quad \ge_{\text{sm}} (f_1(Y_{11}, \dots, Y_{1L}), \dots, f_N(Y_{N1}, \dots, Y_{NL})),\end{aligned}$$

where $\boldsymbol{X}_k = (X_{1k}, X_{2k}, \dots, X_{Nk})^\top$, $k = 1, \dots, L$.

Proof. Let $\boldsymbol{Z}^{(0)} = (\boldsymbol{X}_2, \boldsymbol{X}_3, \dots, \boldsymbol{X}_L)$. Since $\boldsymbol{X}_1, \boldsymbol{Y}_1$, and $\boldsymbol{Z}^{(0)}$ are independent and $\boldsymbol{X}_1 \ge_{\text{sm}} \boldsymbol{Y}_1$, by Theorem 8 we have

$$(f_1(X_{11}, \boldsymbol{Z}^{(0)}), \dots, f_N(X_{N1}, \boldsymbol{Z}^{(0)})) \ge_{\text{sm}} (f_1(Y_{11}, \boldsymbol{Z}^{(0)}), \dots, f_N(Y_{N1}, \boldsymbol{Z}^{(0)})).$$

By letting f_i depend only on the first argument and row i of $\boldsymbol{Z}^{(0)}$, we have

$$\begin{aligned}&(f_1(X_{11}, X_{12}, \dots, X_{1L}), \dots, f_N(X_{N1}, X_{N2}, \dots, X_{NL}))\\ &\quad \ge_{\text{sm}} (f_1(Y_{11}, X_{12}, \dots, X_{1L}), \dots, f_N(Y_{N1}, X_{N2}, \dots, X_{NL})).\end{aligned}$$

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

Now, letting $\boldsymbol{X}_2$ take the role of $\boldsymbol{X}_1$ and letting $\boldsymbol{Z}^{(1)} = (\boldsymbol{Y}_1, \boldsymbol{X}_3, \ldots, \boldsymbol{X}_L)$, and applying Theorem 8 once more, we have

$$\begin{aligned}&(f_1(Y_{11}, X_{12}, \ldots, X_{1L}), \ldots, f_N(Y_{N1}, X_{N2}, \ldots, X_{NL}))\\ &\quad \geq_{\mathrm{sm}} (f_1(Y_{11}, Y_{12}, X_{13}, \ldots, X_{1L}), \ldots, f_N(Y_{N1}, Y_{N2}, Y_{N3}, \ldots, X_{NL})).\end{aligned}$$

Repeating this procedure, we arrive at

$$\begin{aligned}&(f_1(X_{11}, \ldots, X_{1L}), \ldots, f_N(X_{N1}, \ldots, X_{NL}))\\ &\quad \geq_{\mathrm{sm}} (f_1(Y_{11}, \ldots, Y_{1L}), \ldots, f_N(Y_{N1}, \ldots, Y_{NL})).\end{aligned}$$

Taking $f_i(\boldsymbol{x}) = \min_{1\leq k\leq l}(x_k)$ for all i, we have

$$\left(\min_{1\leq k\leq l} X_{1jk}, \ldots, \min_{1\leq k\leq l} X_{ljk}\right) \leq_{\mathrm{sm}} \left(\min_{1\leq k\leq l} Y_{1jk}, \ldots, \min_{1\leq k\leq l} Y_{ljk}\right) \tag{4}$$

when j is fixed. From (2) and (3), we can see that the next step in comparing $M_{(l,n)}$ and $M'_{(l,n)}$ is to sum the terms in (4). Now, since $\phi(\boldsymbol{x}) = \sum_i x_i$ is a linear function with positive coefficients, it is an increasing supermodular function and we can apply Theorem 9.A.16 of [10] to relate the supermodular ordering of individual terms to the increasing convex ordering of the sums of those terms.

Theorem 10. ([10, Theorem 9.A.16].) *If $\boldsymbol{X} \leq_{\mathrm{sm}} \boldsymbol{Y}$ then $\phi(\boldsymbol{X}) \leq_{\mathrm{icx}} \phi(\boldsymbol{Y})$ for any increasing supermodular function $\phi\colon \mathbb{R}^n \to \mathbb{R}$.*

So, from (4),

$$\sum_{i=1}^{l} \min_{1\leq k\leq l} X_{ijk} \leq_{\mathrm{icx}} \sum_{i=1}^{l} \min_{1\leq k\leq l} Y_{ijk}.$$

Finally, since $\max\{\boldsymbol{x}\}$ is increasing convex, we have, from (2) and (3),

$$\mathrm{E}[M_{(l,n)}] \leq \mathrm{E}[M'_{(l,n)}].$$

Recall that if $\boldsymbol{X}$ is WPA, and $\boldsymbol{X}^{\mathrm{I}}$ has the same marginals as $\boldsymbol{X}$ but are independent, $\boldsymbol{X} \geq_{\mathrm{sm}} \boldsymbol{X}^{\mathrm{I}}$ (see Section 2). Using this relation between WPA and the supermodular order, we obtain the following result which can be used to compare the expected makespan of the positive temporal dependence case with the corresponding makespan assuming temporal independence.

Theorem 11. *Let $\boldsymbol{X}_k^j$, $k = 1, \ldots, l$ and $j = 1, \ldots, m$, be weakly positively associated, and let $\boldsymbol{Y}_k^j$ be a vector of independent RVs that have the same marginals as $\boldsymbol{X}_k^j$ for all j and k. Then,*

$$\mathrm{E}\left[\max_{1\leq j\leq m} \sum_{i=1}^{l} \min_{1\leq k\leq l} X_{ijk}\right] \geq \mathrm{E}\left[\max_{1\leq j\leq m} \sum_{i=1}^{l} \min_{1\leq k\leq l} Y_{ijk}\right].$$

We have identified many cases in which minimal replication or maximal replication is optimal when service times are temporally independent. The following corollary extends these results to the case in which there is positive dependence for jobs on the same server.

Corollary 7. *Under the same assumptions as in Theorem 11, if minimal replication or maximal replication is optimal for service times $\boldsymbol{Y}_k^j$, $k = 1, \ldots, l$ and $j = 1, \ldots, m$, then they are also optimal for $\boldsymbol{X}_k^j$.*

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

Proof. The expected makespan for minimal replication,

$$\mathrm{E}\Big[\max_{1\le j\le n} X_{1j1}\Big],$$

and maximal replication,

$$\mathrm{E}\Big[\sum_{i=1}^{n} \min_{1\le k\le n} X_{i1k}\Big] = n\,\mathrm{E}\Big[\min_{1\le k\le n} X_{11k}\Big],$$

do not depend on the index i and, therefore, they do not depend on any dependence across i. By Theorem 11, the expected makespan when there is WPA of service times within a server is always longer than when they are independent. So if minimal or maximal replication is optimal for the independent case, it must also be optimal for the WPA case.

So, for example, if our conjecture that either minimal or maximal replication is optimal for independent Bernoulli service times is true, it is also true when service times are WPA.

4. Performance analysis with an arrival stream of jobs

In the previous section we made the assumption that there are a finite number of jobs waiting to be processed. Now we allow jobs to arrive according to some arrival process and we select the number of servers among the n that will be grouped together so that each group is responsible for processing a single unique job at a time. See Figure 6. If we decide on a group of size l, the amount of time required to complete a particular job will be the minimum among the service times of its l replications, which we define as the *effective service time* of a job. Let Y denote the effective service time. Then

$$Y = \min_{k\in\{1,2,\ldots,l\}} X_k,$$

where the X_k are i.i.d. service times at individual servers. We investigate the behavior of the system as a function of l, the degree of replication.

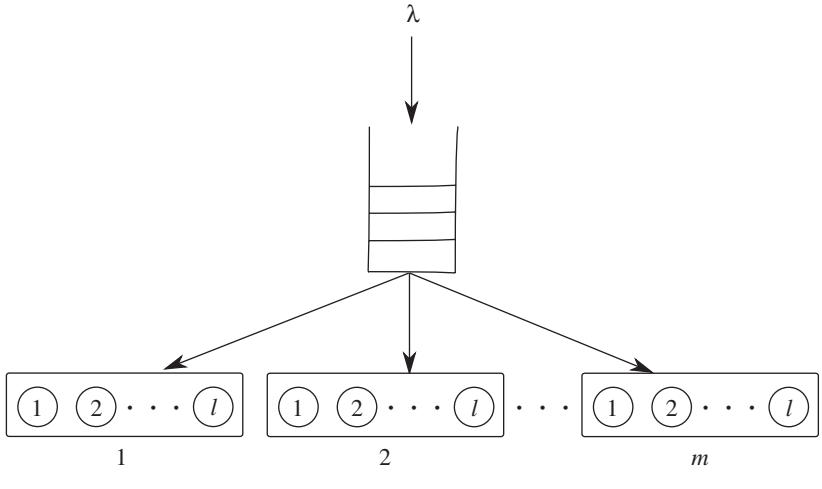


FIGURE 6: A G/G/m queue under l-replication. A group of l processors form a 'team' to simultaneously process replicas of a single job.

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

With l as the group size, $m = n/l$ is the number of groups, where we assume that l divides n for simplicity of discussion. Then the queueing system under this job assignment policy is effectively a G/G/m queue with generic effective service time Y.

Let $\mu(l) = 1/\,\mathrm{E}[\min_{k\in\{1,2,\ldots,l\}} X_k]$ be the effective service rate of a single group as a function of l.

The service capacity of a system of parallel servers is usually represented as the sum of the capacities of the individual parallel servers. In our case,

$$\mu(l)\frac{n}{l} = m\mu(l)$$

is the service rate of the overall system. When there is a stream of jobs arriving at the system at a certain rate, say λ, the overall congestion of the system can be represented as the *system load*

$$\rho(l) = \frac{\lambda}{m\mu(l)} = \frac{\lambda}{n(\mu(l)/l)}.$$

The minimization of $\rho(l)$ or, equivalently, the maximization of service capacity $n\mu(l)/l$ will be our main objective. We consider this objective because it impacts on many performance measures of a queueing system. Most notably, as ρ approaches 1, the delay probability approaches 1 and the expected time a job spends in the system approaches ∞. In this sense we can maximize the stability region of the system by minimizing ρ.

Since $\rho(l)$ is proportional to $l/\mu(l)$, we will often argue the optimality of the system load through $l/\mu(l)$.

Note that this objective is not sensitive to any temporal dependence of service times within a server.

In the following subsections we study $\rho(l)$ for our queueing model. In Subsection 4.7 we will consider the more difficult but more appropriate objective of minimizing the mean response time in steady state.

4.1. The system load (ρ) as a function of l

We may express the system load as

$$\rho(l) = \frac{\lambda}{m\mu(l)} = \frac{\lambda l}{n\mu(l)} = \frac{\lambda}{n}\int_0^\infty l\{\bar{F}(x)\}^l\,\mathrm{d}x.$$

Recall that NWU is equivalent to having a subadditive cumulative hazard rate, which implies that $\bar{F}(lx) \geq \bar{F}^l(x)$. Under this condition, we have

$$\int_0^\infty l\{\bar{F}(x)\}^l\,\mathrm{d}x \leq \int_0^\infty l\bar{F}(lx)\,\mathrm{d}x = \mathrm{E}[X],$$

where X is the RV denoting the service time with cumulative distribution function F. Thus, we have the following upper bound of the system load when F is NWU:

$$\rho(l) \leq \frac{\lambda}{n}\mathrm{E}[X].$$

Similarly, when F is NBU, we have the following lower bound:

$$\rho(l) \geq \frac{\lambda}{n}\mathrm{E}[X].$$

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

Note that $\lambda \, \mathrm{E}[X]/n$ is equal to $\rho(1)$, the system load, for minimal replication. So we conclude that, for NBU service times, minimal replication is optimal and, for NWU service times, minimal replication is the worst.

We now show that the optimality of maximal replication for minimizing $\rho(l)$ implies its optimality for our makespan objective, minimizing $\mathrm{E}[M_{(l,n)}]$.

Theorem 12. *If maximal replication is optimal for the queueing system with objective $\rho(l)$, it is also optimal for the makespan objective $\mathrm{E}[M_{(l,n)}]$ when there are n jobs to do.*

Proof. The following relations hold for all $1 \le l \le n$:

$$M_{(l,n)} = \max_{1 \le i \le m} \sum_{j=1}^{l} \min_{1 \le k \le l} X_{ijk} \ge_{\mathrm{st}} \sum_{j=1}^{l} \min_{1 \le k \le l} X_{1jk},$$

$$\mathrm{E}[M_{(l,n)}] \ge l \, \mathrm{E}\Big[\min_{1 \le k \le l} X_{11k}\Big] = \frac{l}{\mu(l)}.$$

Since we assume that $l/\mu(l)$ is minimized at $l = n$,

$$\mathrm{E}[M_{(n,n)}] = \frac{n}{\mu(n)} \le \frac{l}{\mu(l)} \le \mathrm{E}[M_{(l,n)}] \quad \text{for all } l.$$

By the fact that

$$\frac{n}{\mu(n)} = \mathrm{E}[M_{(n,n)}]$$

we conclude that $\mathrm{E}[M_{(n,n)}] \le \mathrm{E}[M_{(l,n)}]$ for all l.

Note that the inequality

$$\mathrm{E}[M_{(l,n)}] \ge \frac{l}{\mu(l)}$$

is what provides the relation between the system load and the expected makespan. Establishing a similar optimality result for the converse of the above theorem would involve imposing a lower bound on $\mathrm{E}[M_{(l,n)}]$, which leads to a dead end since we would not be able to compare the lower bound with $l/\mu(l)$.

4.2. Impact of the service time variability on the system load

Lemma 5. *If $X' \ge_{\mathrm{cx}} X$ then $\mathrm{E}[X'_{(1,n)}] \le \mathrm{E}[X_{(1,n)}]$, where $X'_{(1,n)}$ and $X_{(1,n)}$ are the first-order statistics from i.i.d. samples of size n distributed as X' and X, respectively.*

Proof. The relation $X' \ge_{\mathrm{cx}} X$ is equivalent to $X' \le_{\mathrm{cv}} X$. Since the function $\min\{\boldsymbol{x}\}$, $\boldsymbol{x} \in \mathbb{R}^n$, is concave in $\boldsymbol{x}$, we have $\mathrm{E}[\min\{\boldsymbol{X}'\}] \le \mathrm{E}[\min\{\boldsymbol{X}\}]$.

Since $\mu(n) = 1/\mathrm{E}[X_{(1,n)}]$ is the system service rate when maximal replication is used, we can see that the system service rate is ordered according to the convex ordering of service time distributions. This together with the fact that $\mathrm{E}[X'] = \mathrm{E}[X]$ when $X' \ge_{\mathrm{cx}} X$ implies the following corollary.

Corollary 8. *If maximal replication yields a smaller system load than minimal replication for service time X, the same is true for $X' \ge_{\mathrm{cx}} X$.*

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

4.3. Conditions for monotonic $\rho(l)$

Let us first consider the simple case of comparing the value of $\mu(2)/2$ to $\mu(1)(=\mu)$. We would like to find conditions where the service rate is at least doubled by using two servers to do a job. We want

$$\mathrm{E}\left[\frac{X}{2}\right] \geq \mathrm{E}[\min\{X_1, X_2\}],$$

where X, X_1, and X_2 are i.i.d. A sufficient condition for this is

$$\frac{X}{2} \geq_{\mathrm{st}} \min\{X_1, X_2\},$$

which is equivalent to

$$\bar{F}(2x) \geq \bar{F}^2(x)$$

for all x in the support of F.

Lemma 6. *For $n = 1, 2, \ldots$, if $X \geq_{\mathrm{st}} 2\min\{X_1, X_2\}$ then*

$$2^{n-1}\min\{X_1, \ldots, X_{2^{n-1}}\} \geq_{\mathrm{st}} 2^n \min\{X_1, \ldots, X_{2^n}\}$$

and if $X \leq_{\mathrm{st}} 2\min\{X_1, X_2\}$ then

$$2^{n-1}\min\{X_1, \ldots, X_{2^{n-1}}\} \leq_{\mathrm{st}} 2^n \min\{X_1, \ldots, X_{2^{n-1}}\},$$

where $X, X_1, X_2, \ldots$ are i.i.d.

Proof. The proof is by induction. We are given that the initial condition holds. Assume that $2^{n-1}\min\{X_1, \ldots, X_{2^{n-1}}\} \geq_{\mathrm{st}} 2^n \min\{X_1, \ldots, X_{2^n}\}$ holds. Then,

$$\begin{aligned}\min\left\{\frac{X_1}{2}, \ldots, \frac{X_{2^n}}{2}\right\} &\geq_{\mathrm{st}} \min\{\min\{X_{11}, X_{12}\}, \ldots, \min\{X_{2^n,1}, X_{2^n,2}\}\} \\ &=_{\mathrm{st}} \min\{X_1, \ldots, X_{2^{n+1}}\}, \\ 2^n \min\{X_1, \ldots, X_{2^n}\} &\geq_{\mathrm{st}} 2^{n+1}\min\{X_1, \ldots, X_{2^{n+1}}\}.\end{aligned}$$

What the lemma implies is that if $X \geq_{\mathrm{st}} 2\min\{X_1, X_2\}$ or $X \leq_{\mathrm{st}} 2\min\{X_1, X_2\}$ then $l/\mu(l)$ and, thus, $\rho(l)$ is monotonically decreasing or, respectively, increasing in l when n and l are powers of 2.

Corollary 9. *If the service time distribution is NWU or NBU then maximal replication or, respectively, minimal replication yields the smallest system load, ρ.*

It can be easily seen that the NWU condition implies that $X \geq_{\mathrm{st}} 2\min\{X_1, X_2\}$. Recall that an RV is NWU if $\mathrm{P}(X > x) \leq \mathrm{P}(X - t > x \mid X > t)$ for all $t, x \geq 0$, and by taking $t = x$ we have $X \geq_{\mathrm{st}} 2\min\{X_1, X_2\}$. Also, $X \geq_{\mathrm{st}} 2\min\{X_1, X_2\}$ is equivalent to $H(2x) \leq 2H(x)$ for all x, where H is the cumulative hazard rate function. However, $X \geq_{\mathrm{st}} 2\min\{X_1, X_2\}$ does not imply NWU; the following serves as a counterexample. Define

$$H(x) = \begin{cases}\sqrt{x}, & 0 \leq x \leq 2^0, \\ 2^n G\left(\dfrac{x}{2^n}\right), & 2^n \leq x \leq 2^{n+1} \text{ for } n = 0, 1, \ldots,\end{cases}$$

where $G(x) = (x-1)^2 + 1$. Note that, for this H, $\frac{1}{2}H(x) \leq H(x/2)$ holds everywhere. But $H(1) = 1$, $H(3) < 3$, and $H(4) = 4$, so $H(1) + H(3) < H(4)$ and H is not subadditive everywhere. Thus, $\frac{1}{2}H(x) \leq H(x/2)$ is a necessary but not a sufficient condition for X to have an NWU distribution.

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

4.4. Bernoulli service times

For a Bernoulli(p) service time X that takes the value 0 or 1, there is no stochastic order relation between $X/2$ and $\min\{X_1, X_2\}$, and we cannot use Lemma 6 to claim monotonicity of $\rho(l)$. However, in this case, it is straightforward to find an expression for $\mu(l)$:

$$\mathrm{E}\Big[\min_{1\le i\le l} X_i\Big] = p^l, \qquad \mu(l) = p^{-l}.$$

Since $\mu(l)$ is an increasing exponential function, it can be seen that $l/\mu(l)$ increases until $l = -1/\log p$ and decreases thereafter. See Figure 7(a). This implies that either minimal replication or maximal replication, depending on n and p, yields the lowest system load. That is, in contrast to the earlier conjecture for the makespan for a finite number of jobs, the unimodal property is verified for the system load objective.

Lemma 7. *For service times distributed Bernoulli(p), if $p \le n^{-1/(n-1)}$ then maximal replication yields a lower system load than minimal replication for a system with $n(>1)$ servers, and, therefore, maximal replication is optimal. Otherwise, minimal replication is optimal.*

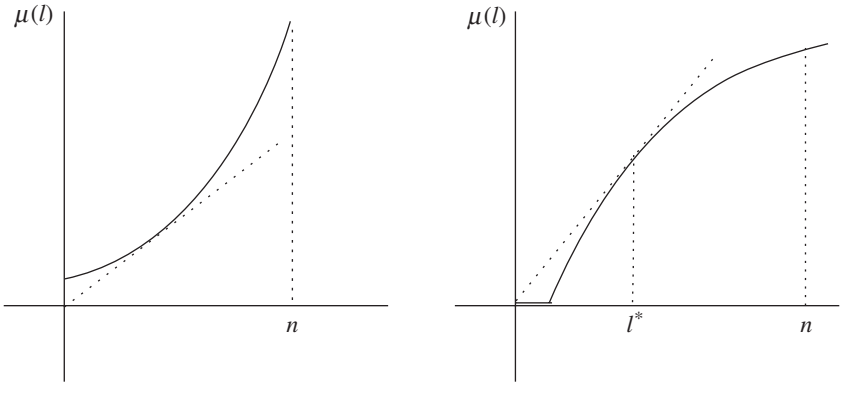
Proof. The system loads for maximal and minimal replications are respectively

$$\rho(n) = \frac{\lambda}{\mu(n)} = \lambda p^n \quad \text{and} \quad \rho(1) = \frac{\lambda}{n\mu(1)} = \lambda\frac{p}{n}.$$

Taking the ratio of maximal replication to minimal replication, we have np^{n-1}. We compare this ratio to 1. A condition for maximal replication being better than minimal replication is

$$np^{n-1} \le 1 \quad \Longleftrightarrow \quad p \le n^{-1/(n-1)}.$$

Corollary 10. *For service times distributed Bernoulli(p), if $p \le 1/\mathrm{e}$ then maximal replication minimizes the system load for all $n > 1$.*



(a) Two-point distributed service time (b) Pareto-distributed service time

FIGURE 7: The form of $\mu(l)$ for different service time distributions. The dotted tangent line indicates where $l/\mu(l)(\propto \rho(l))$ is largest or smallest.

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

Proof. Using the inequality

$$\frac{\log n}{n-1} < 1 \quad \text{for } n > 1,$$

the result is immediate.

Corollary 11. *For service times distributed* Bernoulli(p), *if* $p \le n^{-1/(n-1)}$ *then maximal replication is optimal for the makespan objective,* $\mathrm{E}[M_{(l,n)}]$.

Proof. This follows immediately from the results of Theorem 12 and Lemma 7.

We can also infer from Corollaries 10 and 11 that, when $p \le 1/\mathrm{e}$, maximal replication minimizes the expected makespan for all n. These conclusions are stronger than that of Lemma 2 and Corollary 4 since we now have *optimality* of maximal replication rather than a comparison to minimal replication.

4.5. General two-point distributed service times

Since a general two-point distribution is essentially a scaled and translated Bernoulli distribution, the stochastic order properties are equivalent to those of Bernoulli RVs. Therefore, the system load will not in general be monotone in l. Let us consider an RV $Y = a + bX$, where $a, b > 0$ and $X \sim$ Bernoulli(p). For this case,

$$\frac{1}{\mu(l)} = a + bp^l.$$

By taking the derivative of $l/\mu(l)$,

$$\frac{\mathrm{d}}{\mathrm{d}l}\frac{l}{\mu(l)} = a + bp^l(1 + l\log p),$$

we can see that, for l near 0, the sign is positive and, for a large enough l, the sign is negative, meaning that $l/\mu(l)$ starts increasing and beyond a certain threshold, it decreases. This is when we assume that l can take values ranging from 0 to ∞. But the valid ls are integers in the range $1 \le l \le n$. In this case we can see that, depending on the values of a, b, and p, it is possible for $l/\mu(l)$ to be strictly increasing.

4.6. Pareto service time distributions

What initially motivated us to consider replication was the high variability of service times in grid computing. So heavy-tailed distributions such as the Pareto distribution are perfect candidates for analysis. Recall that the Pareto distribution with location parameter $c > 0$ and shape parameter $k > 0$ has the cumulative distribution function

$$F(x) = 1 - \left(\frac{c}{x}\right)^k, \qquad x \ge c.$$

Depending on the parameter k, the mean and variance can be either finite or infinite:

$$\begin{aligned}
&\mu < \infty \quad \text{and} \quad \sigma^2 < \infty \quad \text{when } k > 2,\\
&\mu < \infty \quad \text{and} \quad \sigma^2 = \infty \quad \text{when } 1 < k \le 2,\\
&\mu = \infty \quad \text{and} \quad \sigma^2 = \infty \quad \text{when } 0 < k \le 1.
\end{aligned}$$

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

An interesting property of the Pareto distribution is that the minimum of two Pareto RVs is also a Pareto RV. So if the service time follows Pareto(c, k) then $X_{(1,l)}$ follows Pareto(c, lk). Thus, even if the original service time distribution has an infinite mean and/or infinite variance, by increasing l (i.e. replicating l times), it is possible to reduce the mean and variance to a finite number. Since the distribution of $X_{(1,l)}$ is explicitly known, $\rho(l)$ can be found easily:

$$\rho(l) = \frac{\lambda}{n}\frac{l}{\mu(l)}, \qquad \frac{1}{\mu(l)} = \mathrm{E}[X_{(1,l)}] = \begin{cases} \dfrac{lkc}{lk-1}, & l > 1/k, \\ \infty, & 0 < l \le 1/k, \end{cases}$$

and $l/\mu(l)$ achieves its minimum value at $l^* = 2/k$. See Figure 7(b) for the graph of the function $\mu(l)$. Hence, if $n \le 2/k$, maximal replication is optimal. Note that although the Pareto distribution is highly variable, it is not NWU (when maximal replication is optimal) because its support is bounded away from 0.

4.7. The mean response time

Although the server utilization is one dimension of the performance of a system which, interestingly, we can control using different degrees of replication, a more important performance measure is the *mean response time*, defined as the average time a job spends in the system. However, this measure is difficult to obtain analytically. When jobs arrive according to a renewal process, a very simple approximation of the mean response is

$$\begin{aligned} \mathrm{E}[W(\mathrm{GI/G/m})] + \mathrm{E}\Big[\min_{1\le i\le l} Y_i\Big] &\simeq \frac{c_a^2 + c_s^2(l)}{2}\,\mathrm{E}[W(\mathrm{M/M/m})] + \mathrm{E}\Big[\min_{1\le i\le l} Y_i\Big] \\ &= \frac{c_a^2 + c_s^2(l)}{2}\,\frac{1}{m\mu(l) - \lambda}\,\mathrm{P}(W(\mathrm{M/M/m}) > 0) + \mathrm{E}\Big[\min_{1\le i\le l} Y_i\Big] \\ &= \frac{c_a^2 + c_s^2(l)}{2}\,\frac{1}{\lambda}\,\frac{\rho(l)}{1-\rho(l)}\,\mathrm{P}(W(\mathrm{M/M/m}) > 0) + \mathrm{E}\Big[\min_{1\le i\le l} Y_i\Big], \end{aligned} \tag{5}$$

where $W(\cdot)$ denotes the time spent waiting in the queue and $Y_i \sim G$ is the service time. As can be seen here, the term $c_s^2(l)$ depends on l through the first and second moments of the first-order statistic from an i.i.d. sample of size l which is generally difficult to characterize. Also, the delay probability depends on l through factorials and the number of terms in a summation.

Observe that, for exponential distributions, $\bar{F}(2x) = \bar{F}^2(x)$, and, hence, there is no difference in the system load $\rho(l)$ for different values of l. Thus, in the exponential case, the problem reduces to the one fast server versus many slow servers case where the one fast server is known to have a smaller response time and we can conclude that maximal replication is optimal for minimizing the mean response time. Indeed, from [4], maximal replication minimizes the mean response time for NWU service times.

For Pareto(c, k) service times, recall that the system load is minimized at $l^* = 2/k$. But the effective service time of a group of l servers is distributed Pareto(c, lk), and the variance is infinite when $lk = 2$. So we can expect the choice of l that minimizes the system load $\rho(l)$ to yield a very long response time. We provide an example for comparing the system load to the response time approximated by (5) for Pareto service times. See Table 2. Observe that, when we do not replicate enough, the large variability negatively affects the performance, and as we

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press

TABLE 2: Approximate response times and system loads for Pareto(3,0.5) service times with $n = 100$ and Poisson arrivals with rate $\lambda = 2$ for different degrees of replication, l, where $\rho(l) < 1$. The squared coefficient of variation of the effective service time of a group of l servers is denoted by $c_s^2(l)$.

l	Mean response time	System load $\rho(l)$	$c_s^2(l)$
1	∞	∞	∞
2	∞	∞	∞
4	∞	0.000 727	∞
5	0.001 679	0.500 000	0.800 000
10	0.245 289	0.750 000	0.066 667
20	∞	1.333 333	0.012 500
25	∞	1.630 435	0.007 619
50	∞	3.125 000	0.001 739
100	∞	6.122 449	0.000 417

replicate too much, the effective service time does not decrease fast enough to compensate for the additional overhead of processing more jobs in sequence.

References

[1] BORST, S., BOXMA, O., GROOTE, J. F. AND MAUW, S. (2003). Task allocation in a multiserver system. *J. Sched.* **6,** 423–436.

[2] DOBBER, M. (2006). Robust applications in time-shared distributed systems. Doctoral Thesis, Vrije Universiteit Amsterdam.

[3] FOSTER, I., KESSELMAN, C. AND TUECKE, S. (2001). The anatomy of the grid: enabling scalable virtual organizations. *Internat. J. High Performance Comput. Appl.* **15,** 200–222.

[4] KOOLE, G. AND RIGHTER, R. (2008). Resource allocation in grid computing. *J. Sched.* **11,** 163–173.

[5] KORPELA, E. *et al.* (2001). SETI@home-massively distributed computing for SETI. *Comput. Sci. Eng*. **3,** 78–83.

[6] LARSON, S. M., SNOW, C. D., SHIRTS, M. AND PANDE, V. S. (2009). Folding@home and genome@home: using distributed computing to tackle previously intractable problems in computational biology. Preprint. Available at http://arxiv.org/abs/0901.0866.

[7] LEISTMAN, A. L. AND CAMPBELL, R. H. (1986). A fault-tolerant scheduling problem. *IEEE Trans. Soft. Eng.* **12,** 1088–1089.

[8] LITKE, A., SKOUTAS, D., TSERPES, K. AND VARVARIGOU, T. (2007). Efficient task replication and management for adaptive fault tolerance in mobile grid environments. *Future Generation Computer Systems* **23,** 163–178.

[9] SHAKED, M. AND SHANTHIKUMAR, J. G. (1994). *Stochastic Orders and Their Applications*. Academic Press, Boston, MA.

[10] SHAKED, M. AND SHANTHIKUMAR, J. G. (2007). *Stochastic Orders*. Springer, New York.

https://doi.org/10.1239/aap/1246886623 Published online by Cambridge University Press