Characterizing The Load of A Single-Server System

Jun Liu
Computer Science Department, University of North Dakota
Grand Forks, ND 58202
jliu@cs.und.edu

March 11, 2006

Abstract

The load of a computer is an important metric and is useful in many applications, such as balancing load across heterogeneous computers. In this paper, a new metric for characterizing the load of a uniprocessor computer is studied. This metric is defined as a ratio of the average waiting time experienced by computing jobs to the average idle time of the processor. The definition of the new load metric is based on two important observations: 1) the average idle time between serving consecutive computing tasks shrinks as the load becomes heavier; 2) the average waiting time experienced by computing jobs increases as the load becomes heavier.

The new load metric aims to being able to compare different load status of a single-server system—the typical abstraction of a uniprocessor computer. The validity of comparing two different load status using the new load metric is subject to the satisfaction of certain conditions. These conditions are that relations of stochastic ordering need to be satisfied between the two processes of physical observations under different load status, *e.g.* the process of idle time and the processes of waiting time. When these conditions are not fully satisfied, the new load metric only serves as the first-moment approximation to the status of operations of a single-server system.

Numerical analysis on the effectiveness of the new load metric has been conducted by simulating operations of a single-server system under different types of job arrivals, the new load metric has been demonstrated to have the ability of characterizing the status of operations of the system, and the limitations on its ability have also been demonstrated. The ability of the new load metric in characterizing different load status is also compared to the ability of the traditional load metric—the average utilization, by comparing their effectiveness on serving as job assignment criteria for balancing load across multiple uni-processor systems with different service capabilities. The preliminary results suggest that the new load metric serves balancing the load more evenly across multiple systems than the average utilization.

Jun Liu
Computer Science Department
University of North Dakota
Grand Forks, ND 58202
jliu@cs.und.edu

1 Introduction

The computational load of a computer is an important metric for characterizing the operational status of the computer. The computational load of a computer within a time period is typically characterized by the average utilization of the processor (CPU) during this time period, *i.e.* the ratio of the amount of time spent in processing computing tasks (jobs) within this time period to the duration of the period. The load of a computer closely relates to other observable performance metrics. For instance, when the load of a computer is high, the average waiting time of computing tasks drastically increases [9]. In general, a heavily loaded computer is vulnerable to sudden external changes because there is no sufficient amount of available resources in ready to sudden changes. The computational load of a computer has vast applications in admission control and scheduling of tasks, especially in environments consisting of heterogeneous computers. For example, the computational load can be used in balancing load across computers of non-uniform service capabilities in order to prevent individual computers from being more heavily loaded than others. When computers are with different capabilities, the same amount of task assignment may result in different load. Therefore, the load status is useful for assigning computing tasks to computers with respect to their load status. Whether the average utilization of the processor is the appropriate metric characterizing the load of a computer? A uniprocessor computer is typically modeled as a single-server system: a work-conserving processor and a buffer of unbounded space. Computing tasks are first queued in the buffer if the processor is busy, and the processor handles tasks at a first-come-first-serve order. Hence, it is important to know what feature of the operation of a single-server system can be characterized by the average utilization of the processor. From the point of view of the processor, its idle time reflects the load of the system. From the point of view of tasks waiting to be processed, their waiting time reflects the load of the system. For example, Table 1 illustrates the average idle time of the processor and the average waiting time of tasks measured at different values of average utilization. In cases (1), (2), and (3), even though values of average utilization are similar, but the operation of a single-server system in different cases should be different because either the average idle time or the average waiting time is different. In cases (3), (4), and (5), even though values of average utilization are very different, but the operation of a single-server system shares some common feature in different cases because either the average idle time or the average waiting time is similar. Hence, comparison of different load status of a single-server system using average utilization is difficult because the average utilization lacks the ability of revealing the status of operations of a single-server system reflecting concerns from both the processor and the tasks.

	Average Utilization %	Average CPU Idle Time (in seconds)	Average Waiting Time (in seconds)
(1)	76.199%	0.14383	0.72390
(2)	75.48%	0.12323	0.49921
(3)	73.401%	0.02732	0.10270
(4)	51.07%	0.20316	0.10340
(5)	37.048%	0.12812	0.02207

Table 1: The average idle time of the processor and the average waiting time of tasks measured at different values of average utilization of a single-server system.

In this paper, we propose a new metric which is more appropriate for characterizing the load status

of a uniprocessor computer than the average utilization. This new metric is derived from both the idle time of the processor of a single-server system and the waiting time of tasks in the system. Both the idle time and waiting time are important physical measures reflecting the operation of a single-server system driven by task arrivals, and they are directly related to the load status of a single-server system in the way that: the mass of the probability distribution of idle time becomes more concentrated in a narrow region close to zero as the load of the system becomes heavier; the mass of the probability distribution of waiting time of tasks generally spreads out across a wider region as the load of the system becomes heavier. Correspondingly, the average idle time and the average waiting time generally shrinks and increases, respectively, as the load of the system becomes heavier. This new load metric is defined as the ratio of the average waiting time to the average idle time. The value of this ratio generally becomes larger as a single-server system becomes more heavily loaded. This new load metric is also purposed to aid comparison of different load status. The validity of comparison of different load status using this new load metric is subject to the satisfaction of certain conditions.

The effectiveness of the new load metric on characterizing the load status of the system needs to be examined. Stochastic expressions are commonly used for expressing the typical behaviors of the operation of a single-server system. Since the new load metric is still defined on the average values of physical observations, and many stochastic features of an entire probability distribution of the values of a physical observation can not be captured by the average value of this physical observation. Thus, a method of representing the status of operation of a single-server system is needed for examining the effectiveness of the new load metric, which should be able to reveal more stochastic features of the operation of the system. The entropy derived from the probability distribution of a physical observation can do for this purpose, because the entropy derived from a probabilistic distribution reflect the difference between the distribution itself and a uniform distribution. In general, the more the mass of a probability distribution is concentrated in a narrow region, the smaller the value of the entropy derived from the probabilistic distribution is [1, 4]. Thus, the entropy expression of the operation of a single-server system can be defined as the ratio of the entropy of waiting time to the entropy of idle time, and the value of this ratio generally becomes larger as the load status of a single-server system becomes more heavier.

When a physical measure is abstracted into a random variable X, the set of possible values of X is denoted as $\mathcal{X}$. The probability distribution of X can be expressed with a probability mass function $p_X(x)$ where $x \in \mathcal{X}$. The entropy of X can be used to characterize the entire distribution of X and is defined as [3]

$$\begin{aligned} H(X) &= E_{p_X}\Big[\log \frac{1}{p_X(x)}\Big] \\ &= \sum_{x\in\mathcal{X}} p_X(x) \log \frac{1}{p_X(x)}. \end{aligned} \tag{1}$$

The entropy of a random variable can be interpreted as the length (number of bits) of the shortest description of the probability distribution of this random variable [3].

Our main results on characterizing the load status of a single-server system are as follows. First, the load status of a single-server system can be approximated by a new load metric which is the ratio of the average waiting time of tasks to the average idle time of the processor. The effectiveness of the new load metric has been demonstrated by comparing the values of the new load metric to the entropy expression of the status of operation of a single-server system under different types of task arrivals. Second, the necessary conditions for the validity of comparison of different load status using the new load metric is demonstrated. Third, a Monte-Carlo method is proposed to examine the

satisfaction of the necessary conditions for valid comparison of different load status using the new load metric. Fourth, load-balancing mechanisms can be developed based on the new load metric, and the new load metric can be further potentially used in constructing cluster-oriented load metrics for balancing load across clusters.

The rest of this paper is organized as follows. The previous work related to our work is described in Section 2. The new load metric and its efficiency of representing the load status are described in Section 3. A demonstration of balancing load adopting the new load metric as a criterion for assigning jobs to receive actual service in a system is illustrated in Section 4. The new load metric is discussed in Section 5, and our work is summarized in Section 6.

2 Related Work

The load information of a computer has vast applications in admission control or scheduling of computing tasks, especially in environments consisting of heterogeneous computers.

Fei *et al.* [6] studied the method of selecting the best server from a cluster based on the criteria of shortest response time. In their work, the response time of serving requests is the time duration between the moment sending a request to a server and the moment receiving the complete response to the request. The response time of serving a request consists of the time the request waits in the buffer of a server. As discussed in our work, the waiting time alone does not act well as a sign of the load status of a server. Moreover, selection of servers using dynamic response time tends to result in unbalanced loads across a cluster. Godfrey *et al.* [7] studied the technique of balancing load across heterogeneous servers organized by a structure of Distributed Hash Tables (DHTs). In their work, the full service capacity of a server is used for distributing requests to servers. Assignment of tasks based on full computing capacities of servers may only work well in situations when the specifications of tasks are very similar; otherwise, this load-balancing technique does not see to work well in general. The load metric proposed in our work can be adopted in their work for their load-balancing technique to function well under general task specifications.

Stochastic representation of features of a system has been widely used in performance analysis of systems, and the entropy is among the methods of representing the stochastic features. Balaji *et al.* [15] has made use of the entropy method to characterize the changes of stochastic features of task arrivals to a single-server system when the task arrivals go through the system. It has been shown that the stochastic features of a number of types of task arrivals can be enriched by going through the system under different types of service policies.

Majorization and stochastic ordering are theoretical tools for comparing performance of systems under different conditions. When a system is viewed as a function preserving a particular relation, the output from the system is comparable if the corresponding input to the systems is assumed to satisfy the particular relation. The methods of majorization and stochastic ordering serve for establishing relations between different input and between the corresponding output.

The methods of majorization and stochastic ordering have been presented in a number of good sources: Hardy *et al.* [8], Marshall *et al.* [13], Bhatia *et al.* [1], etc. A key technique for examining the satisfaction of majorization or stochastic ordering between two physical measures is the Schur-convexity [13]. When two physical measures satisfy a relation of majorization or stochastic ordering, the functional results of the two physical measures have the same relation if the function is Schur-convex. Specifically, Morales *et al.* [14] showed that the respective entropy of two physical measures satisfy an inequality if a relation of stochastic ordering is satisfied between the two measures. Ebrahimi *et al.* [4] explored the properties of entropy of random variables. In their work, a very useful inequality relation has been shown between entropy of two random variables satisfying

a relation of stochastic ordering. Their work builds the foundation of our work on characterizing the status of operation of a single-server system.
The methods of majorization and stochastic ordering have also been applied in analysis of the performance of systems. Chang [2] discussed the application of stochastic ordering in the theoretical study of balancing load and scheduling in multi-server systems. Coffman *et al.* [5] also made use of the technique of stochastic ordering in comparing the fairness of job assignment in balancing load across processors in parallel systems. Koole *et al.* [10] studied the comparison of queue length distribution of queueing systems fed with on-off sources of different stochastic specifications.

3 Characterizing the load of a single server system with infinite buffer space

A work-conserving single-server system is commonly used for modeling a uniprocessor computer. A work-conserving single-server system consists of a buffer of unlimited space and a processor of constant processing rate. Computing tasks are processed in a first-come-first-serve order, and they wait in the buffer if the processor is busy. For a series of tasks $\{1, 2, \cdots, i, \cdots\} (i \in \mathbb{N})$ arriving at the single-server system, their arrival time and service time requested are denoted as $\{a_n, n \in \mathbb{N}\}$ and $\{S_n, n \in \mathbb{N}\}$, respectively. The departure time of tasks is denoted as $\{d_n, n \in \mathbb{N}\}$. Furthermore, the inter-arrival time between consecutive tasks $n-1$ and n is derived as $A_n = a_n - a_{n-1}$, and the inter-departure time between tasks $n-1$ and n is denoted as $D_n = d_n - d_{n-1}$. The process of inter-arrival time and inter-departure time are denoted as $\{A_n, n \in \mathbb{N}\}$ and $\{D_n, n \in \mathbb{N}\}$, respectively. The average inter-arrival time and the average service time requested are denoted as $E[A]$ and $E[S]$, respectively. ($E[A] < \infty$ and $E[S] < \infty$) The load of a single-server system is traditionally expressed as the average utilization ρ of this system, and $\rho = E[\text{fraction of time the processor is busy}]$.
Both the waiting time of tasks and the idle time of the processor are important metrics for characterizing the operation of a single-server system. The waiting time of task n, denoted as W_n, is the amount of time the task waits in the buffer before it is being processed. Following from standard queueing analysis [12, 9], W_n can be expressed into:

$$W_n = d_n - a_n - S_n = (W_{n-1} + S_{n-1} - A_n)^+ \tag{2}$$

where $(x)^+ = \max\{0, x\}$. The idle duration of the processor between processing tasks $n-1$ and n, denoted as I_n, is the interval starting at the moment when task $n-1$ finishes its processing and ending at the moment when task n begins its processing. I_n can be expressed into

$$I_n = (a_n - d_{n-1})^+ = (A_n - S_{n-1} - W_{n-1})^+ \tag{3}$$

as it is illustrated in Figure 1. The process of waiting time and the process of idle time are denoted as $\{W_n, n \in \mathbb{N}\}$ and $\{I_n, n \in \mathbb{N}\}$, respectively. The average waiting time and the average idle time are denoted as $E[W]$ and $E[I]$, respectively. The traditional load metric ρ can be expressed into $1 - \frac{E[I]}{E[A]}$.

3.1 A New Metric for Characterizing Load Status

We claim that the ratio $\frac{E[W]}{E[I]}$ with respect to a sequence of tasks is more appropriate than the average utilization for characterizing the load of a single-server system driven by this sequence of tasks. The new load metric is still a first-moment approximation to the actual status of the operation of

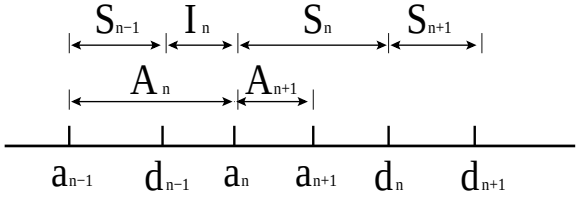


Figure 1: Illustration of The Idle Time.

a single-server system, and it can not fully characterize the stochastic features of the two physical observations, idle time and waiting time, that directly related to the operation of the system. In order to demonstrate the effectiveness of the new load metric, it is necessary to find a way to characterize more of the stochastic features of the two important observations. Hence, the entropy method is under consideration for expressing the probability distribution of a physical observation. When it is to characterize the stochastic features of a sequence of observations $\vec{X} = (X_1, \cdots, X_n)$, the average entropy (also called entropy rate) is used to express the entire vector $\vec{X}$, denoted as $H_{ER}(\vec{X})$, which represents the average number of bits used in expressing an individual random variable in $\vec{X}$. The average entropy of $\vec{X}$ is defined based on the joint entropy of $\vec{X}$, denoted as $H(X_1, \cdots, X_n)$, and $H_{ER}(\vec{X})$ is expressed into

$$H_{ER}(\vec{X}) \quad = \quad \frac{H(X_1, \cdots, X_n)}{n}. \tag{4}$$

The joint entropy of $(X_1, \cdots, X_n)$ is expressed into

$$\begin{aligned} & H(X_1, \cdots, X_n) = \\ & - \sum_{(x_1, \cdots, x_n)} p(x_1, \cdots, x_n) \cdot \log p(x_1, \cdots, x_n) \end{aligned} \tag{5}$$

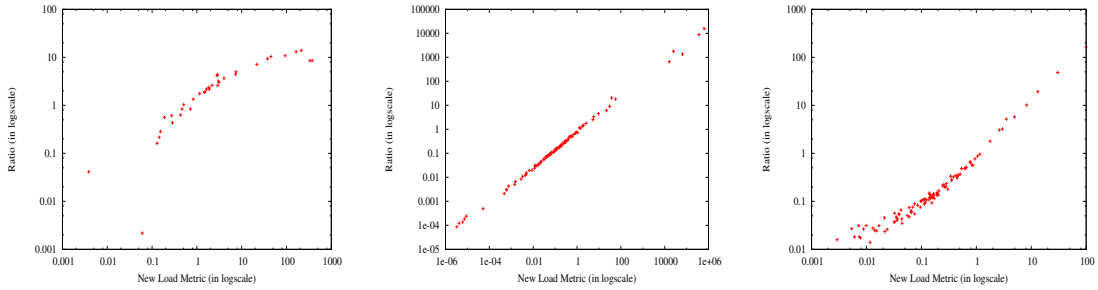
where $p(x_1, \cdots, x_n)$ is the joint probability mass distribution function (pmf) of $(X_1, \cdots, X_n)$, and $(x_1, \cdots, x_n) \in \mathbb{R}^n$ are the arguments of the joint pmf.
When $\vec{I} = (I_2, \cdots, I_n)$ and $\vec{W} = (W_2, \cdots, W_n)$ are used for characterizing the operation of a single-server system driven by a sequence of n task arrivals, the ratio $\frac{H_{ER}(\vec{W})}{H_{ER}(\vec{I})}$ is defined as a reference index of the status of the operation of a single-server system.

3.2 The Effectiveness of the New Load Metric

The effectiveness of the new load metric on characterizing the load status of a system can be evaluated by examining the relation between the new load metric $\frac{E[W]}{E[I]}$ and the reference index $\frac{H_{ER}(\vec{W})}{H_{ER}(\vec{I})}$ under various types of task arrivals.

3.2.1 Poisson Task Arrivals

When both the service time of tasks and the inter-arrival time between consecutive tasks follows exponential distributions, the process of task arrivals forms a Poisson process. The probability mass function (pmf) of an exponential distribution is denoted as $p(x) = \lambda \cdot e^{-\lambda \cdot x}$ $(\lambda > 0)$ with a mean of $\frac{1}{\lambda}$. When the average inter-arrival time and average service time of tasks are denoted as $E[A]$ and $E[S]$, respectively, the corresponding average utilization of a single-server system is $\frac{E[S]}{E[A]}$. Moreover, it is usually assumed that $E[S] \leq E[A]$ for the stable operation of the system. Setting



(a) Poisson Task Arrivals

(b) Task Arrivals with Highly Variable Service Time

(c) Task Arrivals with Dependent Service Time

Figure 2: The effectiveness of the new load metric in characterizing the operation of a single-server system under different patterns of task arrivals, as compared to the effectiveness of the traditional load metric—the average utilization.

different mean values for the distributions of inter-arrival time and service time results in different values of utilization. When the mean inter-arrival time is set to be 1.0 and the mean service time ranges between $[0.1, 1.0]$, the relation between the new load metric $\frac{E[W]}{E[I]}$ and the reference index $\frac{H_{ER}(\vec{W})}{H_{ER}(\vec{I})}$ is shown in Figure 2 (a). The ratio $\frac{H_{ER}(\vec{W})}{H_{ER}(\vec{I})}$ increases with larger values of the new load metric, which corresponds to higher load.

3.2.2 Task Arrivals with Highly Variable Service Time

In order to generate task arrival patterns of highly variable service time requested by tasks than the service time in Poisson arrivals, the service time of tasks is made to follow Weibull distributions [16], and the inter-arrival time between consecutive tasks still follows exponential distributions. The probability mass function of a Weibull distribution is denoted as

$$p(x) = \frac{b}{a}\left(\frac{x}{a}\right)^{(b-1)} e^{-(\frac{x}{a})^b} \quad (a > 0, b > 0)$$

with a being the scale parameter and b being the shape parameter. The Weibull distribution is a versatile distribution that can take on the characteristics of other types of distributions, based on the value of the shape parameter b which controls the decay rate of the tail of a Weibull distribution. For example, a Weibull distribution of $b = 1$ mimics an exponential distribution; a Weibull distribution of $0 < b < 1$ has a longer tail than the distribution of an exponential distribution; a Weibull distribution of $b > 1$ has a shorter tail than the distribution of an exponential distribution. A long-tailed distribution has a larger variance than a short-tailed distribution. For generating task arrival patterns of different variances, a Weibull distribution with $b = 1.0$ is used for the distribution of inter-arrival time of tasks and Weibull distributions with $0 < b \leq 2.0$ are used for the distribution of service time of tasks. Under task arrivals of highly variable service time, the behaviors of the entropy rate of the idle time and the entropy rate of the waiting time are shown in Figure 2 (b). It can be seen that the new load metric can still perform a good job of distinguishing different status of the operation of a single-server system.

3.2.3 Task Arrivals with Dependent Service Time

Task arrivals with dependent service time can be introduced when rigid autocorrelation structures are used for generating the service time requested by tasks. A slow-delay autocorrelation structure on the amount of aggregated service time formed at increasing time scales forms a long-range dependence structure on task arrivals, *i.e.* the variance of the amount of aggregated service time requested by tasks do not decay quickly as the length of the time scale increases [11]. A single-server system fed with task arrivals with dependent service time is usually with a highly dynamic load compared to the load when the system is fed with task arrivals with independent service time. The process of fractional Gaussian noise (fGn) is an approach to model a physical observation with a long-range dependence structure. A fGn process is a Gaussian process with its autocorrelation function represented as (in [11])

$$\rho(k) \to k^{-\beta} \quad (k \to \infty)$$

where k is the lag of the autocorrelation and β measures the degree of the dependence structure exhibited in a physical process. Specifically, the fact $0 < \beta < 1$ indicates the existence of a long-range dependence. Under task arrivals with their service time being modeled as fGn series and their inter-arrival time following exponential distributions, the behaviors of the entropy rate of the idle time and the entropy rate of the waiting time are shown in Figure 2 (c). The dispersed region at the lower-left corner of Figure 2 (c) also demonstrates the limited ability of the new load metric in characterizing the operation of a single-server system when task arrivals are with dependence, because the dispersed region states that different status of the operation of the system results in the same or very close values of the new load metric.

3.3 Discussion

Seen from Figure 2, even though the new load metric has the ability of distinguishing different status of the operation of a single-server system, the ability of the new load metric is still limited. The new load metric can not well distinguish different load status when task arrivals are with dependence, and this is especially true when the system is lightly loaded (ref. Figure 2 (c)). The possible cause to this fact is that the first-moment approximation to idle time can not well capture the rich stochastic features of the distribution of idle time when the distribution spans a wide range. Furthermore, the new load metric is not appropriate for charactering the load under extremely regulated task arrivals, *i.e.* both the service time of tasks and the inter-arrival time between consecutive tasks are constant. When the inter-arrival time and the service time of tasks are denoted as A and S, respectively, the corresponding average utilization of a single-server system is $\frac{S}{A}$. The value of the new load metric is 0 when $A \geq S$. In this case, the average utilization is the only appropriate load metric. The relation of entropy rate of idle time and the average utilization under regulated task arrivals is shown in Figure 3, and the values of entropy rate of waiting time of tasks are 0 at all values of average utilization. The entropy rate of idle time decreases as the average utilization increases. This fact is coincident with the shrinking idle time as utilization becoming high. If the entropy rate of the idle time is denoted as H, then 2^H denotes the size of the support set of idle time. As the average utilization becomes high, the size of the support set of idle time shrinks, and thus, the entropy rate of idle time decreases.

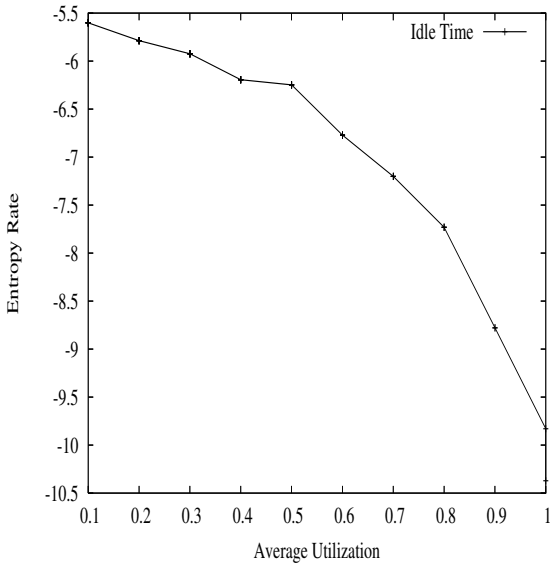


Figure 3: Relation of entropy rate of the idle time and the average utilization under regulated task arrivals.

4 Balancing Load across multiple single-server systems using the new load metric

In order to demonstrate the new load metric's ability of distinguishing different load status, the new load metric is used for balancing load across multiple single-server systems with unequal service capabilities. A cluster of 5 single-server systems with unequal service abilities is used in this demonstration. Arrivals of computing tasks are from a stream mixed from multiple arrival patterns. The inter-arrival time follows a distribution mixed of 5 different exponential distributions, and the service time of jobs follows a distribution mixed from exponential, Weibull, and fGn distributions. The 5 hosts are set purposely with different initial load status by putting various-length backlogs of pending jobs in different hosts before a load-balancing mechanism starts to function. A load-balancing mechanism aims to balancing the load metric values measured among the 5 hosts, and it adopts a simple strategy of assigning a newly arrived job to a host with the current lightest load, *i.e.*with highest availability. Two different load metrics are used in this demonstration: the new load metric and the average utilization. A load metric is updated when a job finishes processing in a host in order to reflect the dynamic changes of load in a host. The effectiveness of a load-balancing mechanism adopting a particular load metric is evaluated by observing the fairness on average idle time across different host, as well as the fairness on average waiting time across the hosts. The effectiveness of the two load-balancing mechanisms adopting the new load metric and the utilization, respectively, are illustrated in Figure 4.

For the load-balancing mechanism using the new load metric, the availability of a host is defined as $e^{(\bar{I}-\bar{W})}$ where $\bar{I}$ and $\bar{W}$ are the average idle time and the average waiting time, respectively. This expression avoids the computational problems when either $\bar{I}$ or $\bar{W}$ is zero. The load-balancing mechanism assigns a newly arrived job to the host with the current highest value of availability. A highly available host has its value of $e^{(\bar{I}-\bar{W})}$ prominently different from others (ref. Figure 4 (a)-(3)). Aided by the metric of availability, differences among average idle time and among average waiting time, respectively, are made small over time (ref. Figure 4 (a)-(1) and (a)-(2)).

For the load-balancing mechanism using the average utilization, the most available host is the one

with the current smallest value of average utilization. The load-balancing mechanism assigns a newly arrived job to the host with the current smallest value of average utilization. Since the average utilization only focuses on the business of the processor itself, a host with a high average utilization will not be able to receive new jobs even if its backlog of pending jobs is not long. Therefore, a large amount of new jobs can be cumulated at those hosts with a lower average utilization over time. The load in different hosts can oscillate largely, and correspondingly, it is not good to reduce the differences among average idle time and among waiting time (ref. Figure 4 column (b)).
This demonstration only shows the potentials of applying the new load metric in balancing load across multiple hosts with different service capabilities. More in-depth work needs to be performed on the subject of balancing load. Moreover, the metric of availability defined based on the new load metric, *i.e.*$e^{(\bar{I}-\bar{W})}$, also offers an opportunity to characterize the overall availability of a cluster of hosts. If the operation of each host is assumed independent from other hosts, then the overall availability of hosts in a cluster can be expressed as a sum of individual availability metric. Aided by the cluster-oriented availability metric, inter-cluster load-balancing can be studied.

5 Discussion

Two metrics are considered for characterizing the load status of a single-server system: the ratio $\frac{E[W]}{E[I]}$ and the ratio $\frac{H_{ER}(W)}{H_{ER}(I)}$. As we know, the entropy expressions of physical observations to the operation of a single-server system reveal more stochastic features of the probabilistic distributions of the physical observations, *i.e.* the entropy expression of a physical observation (*e.g.* the idle time) reveals the degree of uncertainty of the probabilistic distribution of the physical observation. In contrast, the average value of a physical observation does not reveal many stochastic features of the probability distribution of the observation. In this circumstance, it is necessary to note that the ratio $\frac{E[W]}{E[I]}$ is still a good load metric for two reasons: 1) the ratio $\frac{E[W]}{E[I]}$ is easier to evaluate than the ratio $\frac{H_{ER}(W)}{H_{ER}(I)}$; 2) in many cases, the ratio $\frac{E[W]}{E[I]}$ has a similar ability as the ratio $\frac{H_{ER}(W)}{H_{ER}(I)}$ on distinguishing two different load status. The fact of being simple to be evaluated makes the ratio $\frac{E[W]}{E[I]}$ a good candidate load metric in realistic applications.

6 Conclusion

We proposed a new load metric for approximating the status of operation of a single-server system, in order to remedy the incomplete representation of the status of operation when using the traditional load metric—the average utilization. This new load metric is expressed as the ratio of the average waiting time of tasks to the average idle time of the processor in a single-server system. Both the idle time and the waiting time reflect different aspects of the load of a single-server system. Since this new load metric is still a first-moment approximation to the status of operation of a single-server system, entropy representations of the idle time and of the waiting time are used to examine the effectiveness of the new load metric on representing the load status of the system. The ability of the new load metric in characterizing the load status of operations of a single-server system has been demonstrated under different job arrival patterns.
This new load metric aims at gaining the ability of comparing different load status. The validity of comparison of different load status using this new load metric is subject to the satisfaction of certain conditions. These conditions are that relations of stochastic ordering need to be satisfied between two processes of the idle time and between the two processes of the waiting time, respectively, derived under two different load status. If these conditions can not be satisfied, the difference on

load status suggested by the new load metric can only be treated as a first-moment approximation to the actual difference between the status of operations of the two systems.
The difference on the abilities between the new load metric and the traditional load metric—average utilization—is also compared. The comparison is conducted through comparing their effectiveness on serving as criteria for job assignments, in order to balance load across multiple single-server systems in a cluster. Preliminary results showed that the load-balancing mechanism adopting the new load metric could provide a better fairness among average idle time and among average waiting time among the multiple systems, as compared to a load-balancing system adopting the average utilization.

References

[1] Rajendra Bhatia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer, New York, 1996.

[2] Cheng-Shang Chang. A new ordering for stochastic majorization: theory and applications. *Advances in Applied Probability*, 24:604–634, 1992.

[3] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, 1991.

[4] Nader Ebrahimi, Ehsan S. Soofi, and Hassan Zahedi. Information properties of order statistics and spacings. *IEEE Transactions on Information Theory*, 50(1):177–183, January 2004.

[5] Jr. Edward G. Coffman, John Bruno, and Peter Downey. Scheduling independent tasks to minimize the makespan on identical machines. *Probability in the Engineering and Informational Sciences*, 9(3):447–456, 1995.

[6] Zongming Fei, Samrat Bhattacharjee, Ellen W. Zegura, and Mostafa H. Ammar. A novel server selection technique for improving the response time of a replicated service. In *Proceedings of INFOCOM'98*, pages 783–791, 1998.

[7] Brighten Godfrey and Ion Stoica. Heterogeneity and load balance in distributed hash tables. In *Proceedings of INFOCOM'05*, 2004.

[8] G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, Cambridge, UK, second edition, 1952.

[9] Leonard Kleinrock. *Queueing Systems*, volume I. Theory. John Wiley & Sons, 1975.

[10] Ger Koole, Zhen Liu, and Don Towsley. Comparing queueing systems with heterogeneous on-off sources. Technical Report WS-533, Vrije Universiteit Amsterdam, 1999.

[11] Will E. Leland, Murad S. Taqq, Walter Willinger, and Daniel V. Wilson. On the self-similar nature of Ethernet traffic. In Deepinder P. Sidhu, editor, *ACM SIGCOMM*, pages 183–193, San Francisco, California, 1993.

[12] David V. Lindley. The theory of queues with a single server. *Proceedings of Cambridge Philosophy Society*, 48:277–289, 1952.

[13] Albert W. Marshall and Ingram Olkin. *Inequalities: Theory of Majorization and Its Applications*, volume 143 of *Mathematics in Science and Engineering*. Academic Press, 1979.

[14] Domingo Morales, Leandro Pardo, and Igor Vajda. Uncertainty of discrete stochastic systems: General theory and statistical inference. *IEEE Transactions on Systems, Man, and Cybernetics, Part A. Systems and Humans*, 26(6):681–697, November 1996.

[15] Balaji Prabhakar and Robert Gallager. Entropy and the timing capacity of discrete queues. *IEEE Transactions on Information Theory*, 49(2):357–370, 2003.

[16] Wallodi Weibull. A statistical theory of the strength of material. In *Proceedings of Royal Swedish Institute for Engineering Research*, volume 151, Stockholm, 1939.

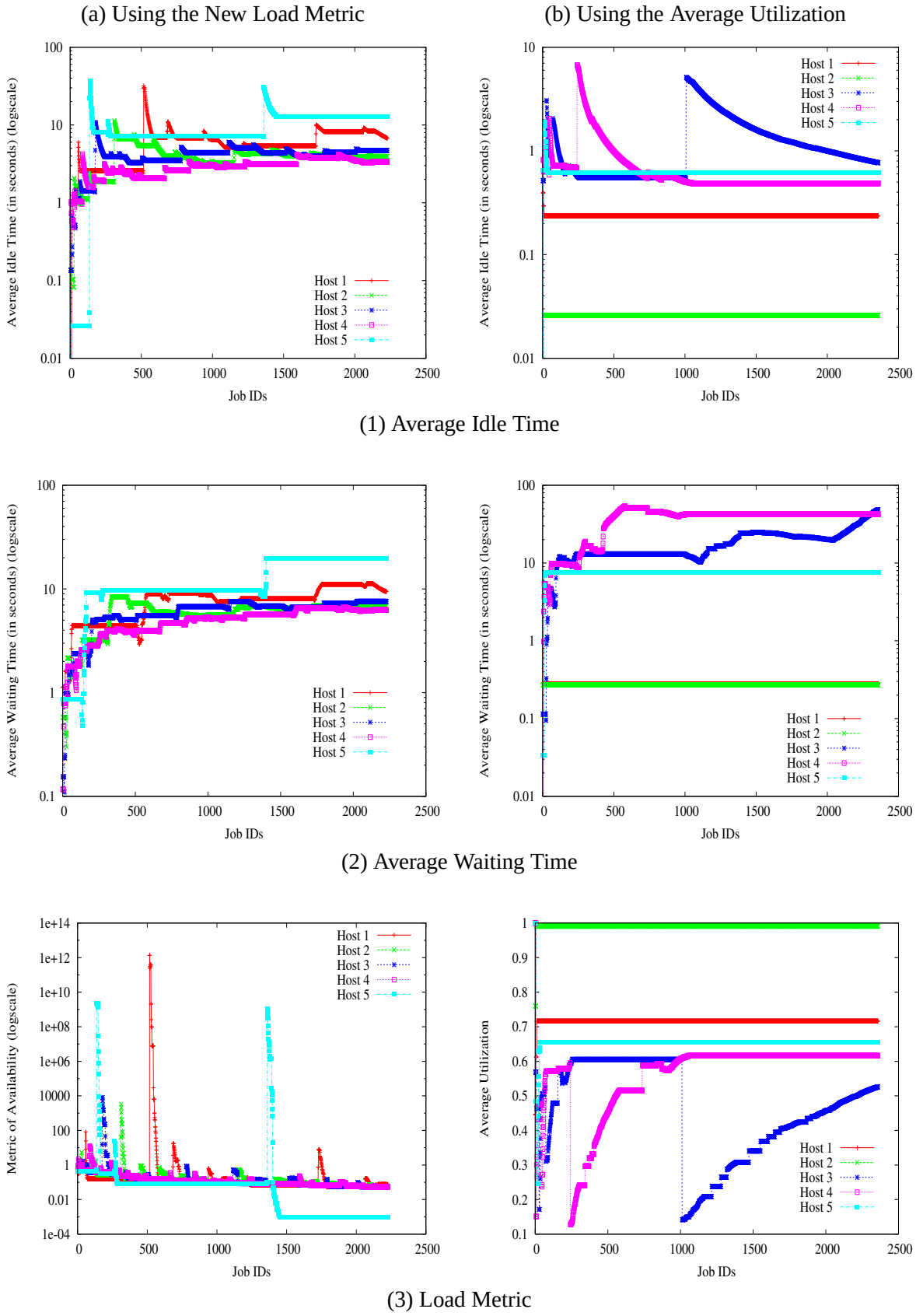


Figure 4: The history of changes on average idle time, average waiting time when different metrics are used for balancing load. There are 5 computing hosts with unequal service abilities.