

December 25, 2008

RT0826

Computer Science 22 pages

# Research Report

Differentiating the performance of systems more reliably

Takayuki Osogami

IBM Research, Tokyo Research Laboratory

IBM Japan, Ltd.

1623-14 Shimotsuruma, Yamato

Kanagawa 242-8502, Japan

## Limited Distribution Notice

This report has been submitted for publication outside of IBM and will be probably copyrighted if accepted. It has been issued as a Research Report for early dissemination of its contents. In view of the expected transfer of copyright to an outside publisher, its distribution outside of IBM prior to publication should be limited to peer communications and specific requests. After outside publication, requests should be filled only by reprints or copies of the article legally obtained (for example, by payment of royalties).



# Differentiating the performance of systems more reliably

Takayuki Osogami

December 25, 2008

## Abstract

The system of measuring the performance of a Web system using a workload generator can be modeled as a closed interactive system. In such a system, the throughput and the mean response time are related by the response time law. However, we find that a measured throughput and a corresponding measured mean response time can have significantly different accuracy. As a result, one metric may be more reliable than the other to identify the better of two given configurations of a Web system, which is an important problem that appears frequently in practice. Using simulation, we derive rules of thumb that characterize when throughput is more reliable than mean response time. Also, we explain these rules of thumb analytically. Specifically, we refine the response time law using the central limit theorem and formally define the asymptotic reliability of an estimator of a metric. Using these analytical frameworks, we provide insights into when and why one metric is more reliable than the other.

## 1 Introduction

The configuration of a computer system needs to be adjusted so that we can enjoy its best possible performance at a given environment. For example, a Web system has many configuration parameters whose particular values can have a significant impact on the performance. A common practice is to measure the performance for different configurations of a Web system and choose the configuration that achieved the best measured performance. Popular metrics of the performance include mean response time and throughput: the mean response time is the average time between when a request is issued and when the request is completed, and the throughput is the average number of requests processed in a unit time. These metrics can be measured by using a workload generator that feeds requests to a Web system and recording the times when the requests are completed.

The question that we address in this paper is which of the two metrics, mean response time or throughput, we should base our judgment on in determining the better of two given configurations. One might argue that both metrics are important and should be balanced on the ground that throughput is a metric from a system's perspective and mean response time is a metric from a users' perspective. Alternatively, one might argue that mean response time is a better metric, since

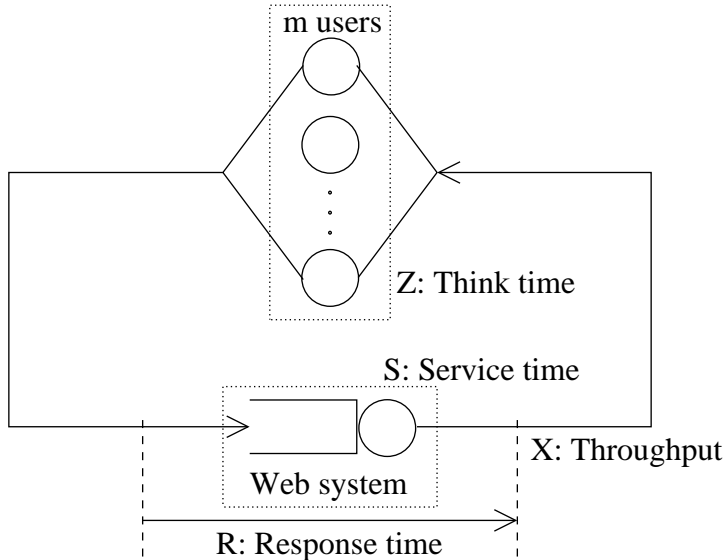


Figure 1: A closed interactive system model of a Web system.

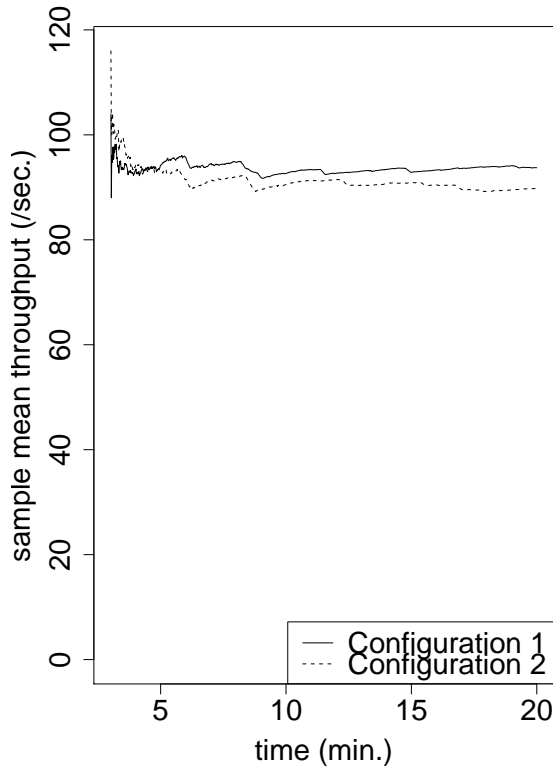
we use more information to calculate mean response time than to calculate throughput. Notice that, to calculate the throughput, we only need to record the number of completed requests and the duration of measurement. To calculate the mean response time, we also need to record when each request is issued and completed. Another argument might be that the two metrics are equivalent, since one metric can be calculated from the other using the response time law, which we will explain shortly.

Observe that a Web system under measurement can be modeled as a closed interactive system as shown in Figure 1. A workload generator emulates a fixed number of users who send requests to the Web system. After a user issues a request, the user waits for the result of the request and then thinks (pauses) for a random period before issuing the next request. In such a closed interactive system, there is a relation, known as the response time law (see p. 46 from [9]) or the interactive response time law (see Section 33.5 from [6]), between the mean response time,  $r$ , and the throughput,  $x$ :

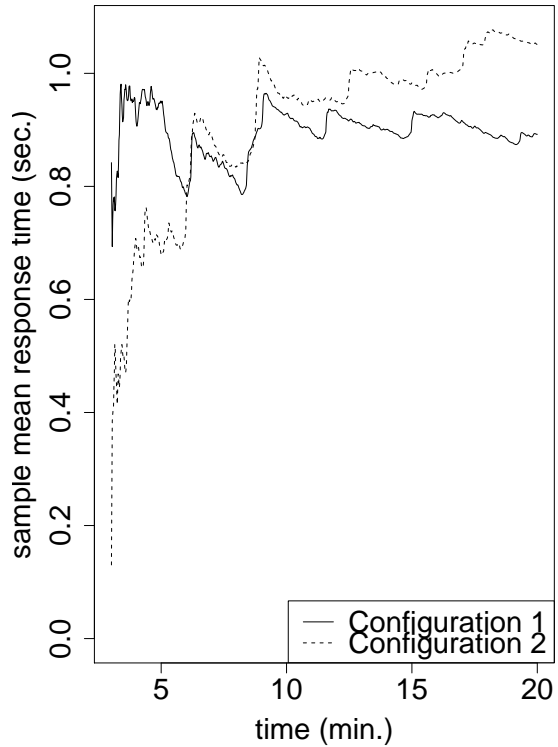
$$r = \frac{m}{x} - z,$$

where  $z$  is the mean think time, and  $m$  is the number of users. Note that  $z$  and  $m$  are known, since they are specifiable parameters of the workload generator. Hence, if  $x$  is measured, then  $r$  can be calculated, and vice versa.

None of the above three arguments do not precisely answer our question, as is illustrated by Figure 2. Figure 2(a) shows the measured throughputs of a Web system for two particular configurations as a function of the measurement time. Figure 2(b) shows the corresponding measured mean response times. Figure 2 is drawn using the experimental results in [12], and the settings of the experiments are detailed in [12] and also repeated briefly in Section 2 of this paper. Observe that the measured throughputs appear to be more stable over time than the corresponding mean



(a) Throughput



(b) Mean response time

Figure 2: The measured throughput (a) and the measured mean response time (b) of a Web system with two particular configurations studied in [12]. The two configurations differ in three parameters, (`MaxClients`, `HeapMax`, `PreparedStatementCacheSize`). The parameters of Configuration 1 are set as (400,192,20) and those of Configuration 2 are set as (300,128,30). `ThreadPoolMax`, which is also varied in [12], is set as 40 for both configurations.

response times. Also observe that the relative difference between the two measured throughputs tends to be smaller than that between the corresponding mean response times at a given time. In this example, if throughput was measured for more than five minutes, Configuration 1 would be identified as the better, which appears to be the correct decision. However, if mean response time was measured for eight minutes, either of the two configurations might be identified as the better.

In this paper, we study the throughput and the mean response time measured or simulated for a closed interactive system. In particular, we characterize which of the two metrics is more reliable to precisely identify the better of two given configurations and under what conditions. Specifically, we will find that mean response time tends to become more reliable than throughput when given two configurations have closer performances, when the load is lower, when the service time has lower variability, or when the think time has higher variability. Also, throughput tends to be more reliable than mean response time when the load is intermediate. When the load is

high, the two metrics tend to have similar reliability. These rules of thumbs derived based on simulation-based studies constitute the first contribution of this paper.

The second contribution of this paper is an analytical study of the asymptotic reliability of the two metrics. We analyze the asymptotic variabilities of a measured mean response time and a measured throughput when measurement time approaches infinity. Then we relate the asymptotic variabilities of the two metrics, which can be seen as a central limit theorem (CLT) version of the response time law. The CLT version of the response time law is then used to explain the findings with simulations and measurements. Our explanation provides insights into when and why one metric is more reliable than the other.

The rest of the paper is organized as follows. In Section 2, we start with a statistical analysis of the measured mean response time and the measured throughput from a Web system studied in [12] and investigate the observation from Figure 2. In Section 3, we study the reliability of the two metrics with simulations for a wide range of settings and derive rules of thumb. In Section 4, we state the CLT version of the response time law, which will be used in Section 5 to study the asymptotic reliability of the two metrics. In Section 6, we review the related work in the literature.

## 2 Statistical analysis of measured performances

In this section, we study the traces from the measurement of a Web system in [12]. In Section 2.1, we formally describe how mean response time and throughput are measured, and study how well the response time law relates the two measured metrics. In Section 2.2, we study the variabilities of the two measured metrics. In the setting under consideration, we will find that the measured throughput of a configuration tends to have smaller variability than the corresponding mean response time. However, we will also find that the difference between the throughputs of two configurations tends to be smaller than that between the corresponding mean response times. In Section 2.3, we discuss which of the two metrics is more reliable to identify the better of two given configurations. Our conclusion will be that, in the particular settings under consideration, throughput can be significantly more reliable than mean response time particularly when the measurement time is short.

Before discussing our findings, we briefly state the settings of the experiments, whose details are found in [12]. The Web system under consideration runs Trade3, a standard benchmark application for online stock brokerage. A machine is dedicated to generate workload with Web Performance Tools V1.9 (WPT). Specifically, WPT emulates  $m = 500$  users, where each user issues a random request such as *buy*, *sell*, or other standard trading operations. After issuing a request, the user waits for the results of the request returned by the Web system. After receiving the results, the user thinks for a random period, uniformly distributed between a second and eight seconds (i.e.,  $z = 4.5$  seconds), before issuing a new request. When we start measuring the performance of a given configuration, 500 users log in at random times so that all of the users are expected to log in within two minutes. The first three minutes is then considered to be the warmup period and ignored. Therefore, for each configuration, the completion time and the response time of each request are stored for 17 minutes after the warmup period.

## 2.1 Estimators

Let  $t$  be the time since the warmup period ends (and measurement starts) for a given configuration. Let  $L(t)$  be the number of requests completed from time 0 to  $t$ . Let  $R_\ell$  be the response time of the  $\ell$ -th completed requests for  $\ell = 1, 2, \dots$  after time 0. Then, at time  $t$  such that  $L(t) > 0$ , the time-average throughput and the sample-average response time are respectively given by

$$\bar{X}(t) = \frac{L(t)}{t} \text{ and } \bar{R}(t) = \frac{\sum_{\ell=1}^{L(t)} R_\ell}{L(t)}. \quad (1)$$

Observe that  $\bar{X}(t)$  and  $\bar{R}(t)$ , respectively, are natural estimators of the throughput,  $x$ , and the mean response time,  $r$ , at time  $t$ . Indeed, Figure 2 shows  $\bar{X}(t)$  and  $\bar{R}(t)$  for two particular configurations. If the weak law of large numbers holds for these natural estimators, we have  $\bar{X}(t) \Rightarrow x$  and  $\bar{R}(t) \Rightarrow r$  as  $t \rightarrow \infty$ , where  $\Rightarrow$  denotes weak convergence.

Each dot in Figure 3(a) shows the estimated throughput,  $\tilde{x} = \bar{X}(T)$ , and the estimated mean response time,  $\tilde{r} = \bar{R}(T)$ , after  $T = 17$  minutes of measurement for one of the 225 configurations studied in [12]. The solid curve in Figure 3(a) shows the response time law. Observe that the response time law well characterizes the relation between  $\tilde{r}$  and  $\tilde{x}$  but not exactly. If the response time law held exactly, the higher throughput a configuration has, the shorter mean response time the configuration has. Note, however, that there are pairs of configurations such that one configuration has higher throughput and yet has longer mean response time than the other. Therefore, depending on whether we judge based on throughput or mean response time, we can make different conclusion as to which configuration has better performance.

Observe, in Figure 3(a), that the relative difference between the mean response times of given two configurations tends to be larger than that between the corresponding throughputs. Specifically,  $\tilde{r}$  ranges between 0.78 and 4.26, differing by a factor of 5.5. By contrast,  $\tilde{x}$  ranges between 57.5 and 93.7, differing by a factor of 1.63. These observations might appear to suggest that mean response time would be more reliable than throughput, but this turns out to be largely false as we will see below.

## 2.2 Variance parameters of the estimators

Next, we examine the accuracy of  $\bar{X}(t)$  and  $\bar{R}(t)$  by estimating their variance parameters,  $V_X$  and  $V_R$ , respectively. Specifically, let  $\Delta_L(t) \equiv L(t) - L(t-1)$  be the measured throughput in the interval  $(t-1, t]$  for  $t = 1, 2, \dots, T$ , where  $T = 1,020$  seconds. Also, let  $K(t) \equiv \frac{t}{L(t)} \sum_{\ell=1}^{L(t)} R_\ell$  be the sample-average response time at  $t$  multiplied by  $t$ , and we define  $\Delta_K(t) \equiv K(t) - K(t-1)$  for  $t = 1, 2, \dots, T$ . We estimate  $V_X$  and  $V_R$ , respectively, with  $\tilde{V}_X = \frac{1}{T} \sum_{\ell=1}^T (\Delta_L(\ell) - \tilde{x})^2$  and  $\tilde{V}_R = \frac{1}{T} \sum_{\ell=1}^T (\Delta_K(\ell) - \tilde{r})^2$ , where recall that  $\tilde{x} = \bar{X}(T)$  and  $\tilde{r} = \bar{R}(T)$ .

A rationale for estimating with  $\tilde{V}_X$  and  $\tilde{V}_R$  is provided by assuming that  $L(t)$  and  $K(t)$  are Brownian motions. In Appendix A.1, we show that  $\tilde{V}_X$  is the maximum likelihood estimator of  $V_X$  provided that  $L(t)$  is the Brownian motions with drift  $x$  and variance parameter  $V_X$  (i.e., the variance is  $V_X t$  at time  $t$ ), and that  $\tilde{V}_R$  is the maximum likelihood estimator of  $V_R$  when  $K(t)$  satisfies analogous conditions.

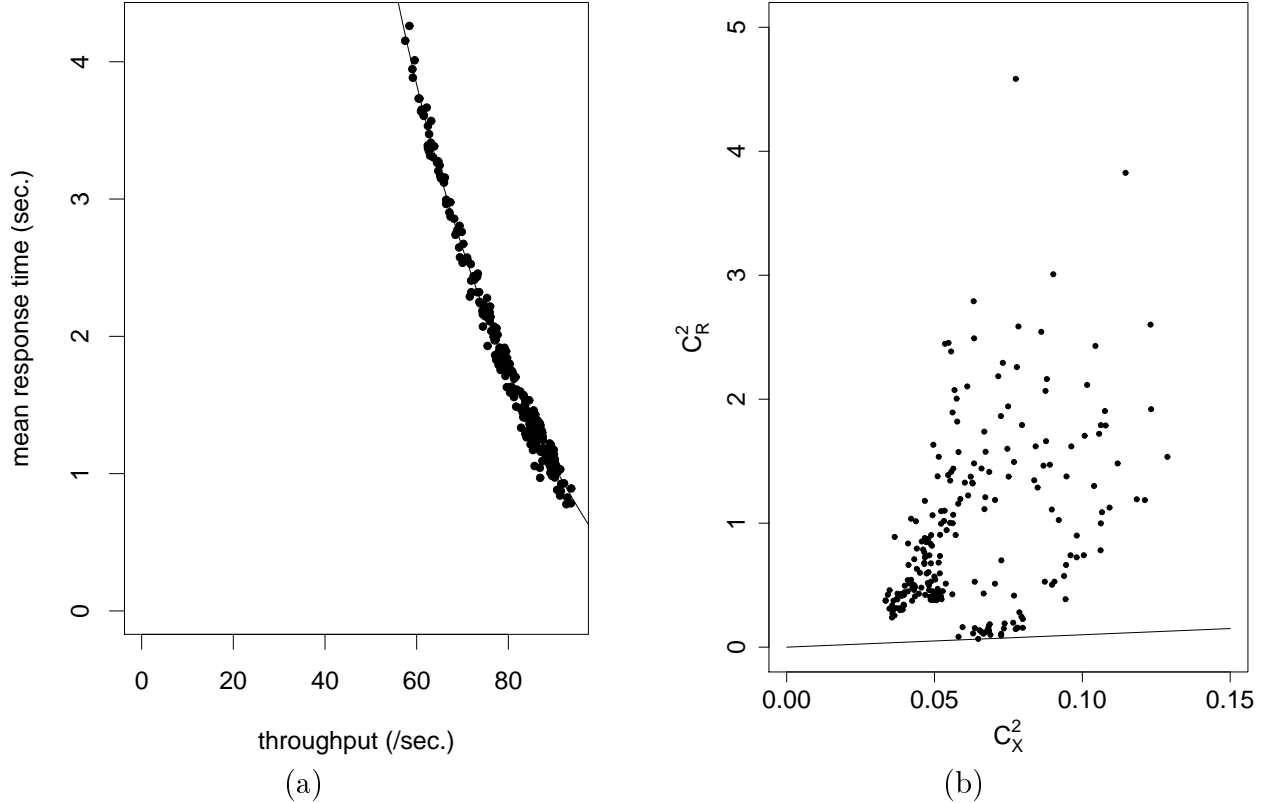


Figure 3: For each of the 225 configurations, we plot (a)  $(\tilde{x}, \tilde{r})$  and (b)  $(\tilde{C}_X^2, \tilde{C}_R^2)$ . The curve in (a) shows the response time law, and the line in (b) shows  $C_X^2 = C_R^2$ .

We normalize  $\tilde{V}_X$  and  $\tilde{V}_R$  in such a way that  $\tilde{C}_X^2 \equiv \tilde{V}_X/\tilde{x}^2$  and  $\tilde{C}_R^2 \equiv \tilde{V}_R/\tilde{r}^2$  to isolate the impact of the mean values from the variances. In Figure 3(b), each dot shows  $(\tilde{C}_X^2, \tilde{C}_R^2)$  for one of the 225 configurations. The solid line shows  $\tilde{C}_R^2 = \tilde{C}_X^2$ . Observe that all of the dots are above the solid line. This suggests that  $\bar{R}(t)$  has higher variability than  $\bar{X}(t)$ . In fact,  $\bar{R}(t)$  appears to have much higher variability than  $\bar{X}(t)$ , since  $\tilde{C}_R^2/\tilde{C}_X^2$  ranges between 1.03 and 59.2. Therefore, contrary to Figure 3(a), Figure 3(b) appears to suggest that throughput might be more reliable than mean response time.

### 2.3 Mean response time or throughput

The above arguments suggest that throughputs of two given configurations tend to have lower variability but be closer to each other than the corresponding mean response times. To investigate this tradeoff, we study the relative variability of the difference between the throughputs of two given configurations and the analogous quantity between the corresponding mean response times. For  $i = 1, 2$ , let  $x^{(i)}$  be the true throughput of Configuration  $i$  and  $\bar{X}^{(i)}(t)$  be its natural estimator at time  $t$ . Similarly, let  $r^{(i)}$  be the corresponding true mean response time and  $\bar{R}^{(i)}(t)$  be its natural estimator. The following coefficients of variation can be used to study the relative variability of

the difference between the two estimated values:

$$C_{\Delta X}(t) \equiv \frac{\sqrt{\mathbf{V}(\bar{X}^{(1)}(t) - \bar{X}^{(2)}(t))}}{|x^{(1)} - x^{(2)}|} \quad \text{and} \quad C_{\Delta R}(t) \equiv \frac{\sqrt{\mathbf{V}(\bar{R}^{(1)}(t) - \bar{R}^{(2)}(t))}}{|r^{(1)} - r^{(2)}|}, \quad (2)$$

where  $\mathbf{V}(\cdot)$  denotes the variance. Unfortunately, Figure 3 suggests that  $C_{\Delta X}(t)$  is hard to estimate accurately, since  $x^{(1)}$  and  $x^{(2)}$  can be very close to each other relative to the error in the measurement.

Therefore, in this section, we estimate and compare  $D_{\Delta X}(t) \equiv |x^{(1)} - x^{(2)}| C_{\Delta X}(t)$  and  $D_{\Delta R}(t) \equiv |x^{(1)} - x^{(2)}| C_{\Delta R}(t) = \frac{x^{(1)}x^{(2)}}{m} \sqrt{\mathbf{V}(\bar{R}^{(1)}(t) - \bar{R}^{(2)}(t))}$ , which follows from the response time law. Notice that  $C_{\Delta X}(t) < C_{\Delta R}(t)$  iff  $D_{\Delta X}(t) < D_{\Delta R}(t)$ . We estimate  $D_{\Delta X}(t)$  with  $\tilde{D}_{\Delta X}(t) \equiv \sqrt{\tilde{V}_X^{(1)}(t) + \tilde{V}_X^{(2)}(t)}$  and  $D_{\Delta R}(t)$  with  $\tilde{D}_{\Delta R}(t) \equiv \frac{\tilde{x}^{(1)}\tilde{x}^{(2)}}{m} \sqrt{\tilde{V}_R^{(1)}(t) + \tilde{V}_R^{(2)}(t)}$ . Here,  $\tilde{V}_X^{(i)}(t)$  and  $\tilde{V}_R^{(i)}(t)$ , respectively, denote the estimated variance parameters of  $\tilde{X}^{(i)}(\cdot)$  and  $\tilde{R}^{(i)}(\cdot)$  at time  $t$ . We calculate  $\tilde{V}_X^{(i)}(t)$  and  $\tilde{V}_R^{(i)}(t)$  analogously to  $\tilde{V}_X$  and  $\tilde{V}_R$  in Section 2.2 but using only the measurement by  $t$ . Also,  $\tilde{x}^{(i)}$  is calculated analogously to  $\tilde{x}$  in Section 2.1.

In Figure 4, the solid curve shows the average value of  $\tilde{D}_{\Delta X}(t)$  over all of the  $\binom{225}{2}$  pairs of the 225 configurations. Similarly, the dashed curve shows the average value of  $\tilde{D}_{\Delta R}(t)$ . Notice that these average values do not necessarily well represent how  $C_{\Delta X}(t)$  and  $C_{\Delta R}(t)$  compare with each other, since they are biased toward the pair of the configurations having large  $|x^{(1)} - x^{(2)}|$  (recall that  $C_{\Delta X}(t)$  and  $C_{\Delta R}(t)$  are multiplied with  $|x^{(1)} - x^{(2)}|$  to define  $D_{\Delta X}(t)$  and  $D_{\Delta R}(t)$ ). Nonetheless, we study these average values, since individual values of  $\tilde{D}_{\Delta X}(t)$  and  $\tilde{D}_{\Delta R}(t)$  are prone to the inaccuracy of measurement. In the following sections, we will be able to directly compare  $C_{\Delta X}(t)$  and  $C_{\Delta R}(t)$  using simulation and analysis.

Observe that the solid curve is below the dashed curve, suggesting that the relative variability of  $\bar{X}^{(1)}(t) - \bar{X}^{(2)}(t)$  is smaller than that of  $\bar{R}^{(1)}(t) - \bar{R}^{(2)}(t)$ . This suggests that, in the settings under consideration, throughput is more reliable than mean response time to identify the better of two given configurations. A primary conclusion of this section is that there appears to be a significant difference between the reliability of throughput and that of mean response time particularly when measurement time is short. This motivates us to further investigate the reliability of the estimators with simulation and analysis.

### 3 Simulation-based study of reliability

Since experimenting with a real Web system is time-consuming and its results depend on minute conditions, we now switch to simulation. We start by describing the conditions of the simulation in Section 3.1. In Section 3.2, we discuss the results of the simulations.

#### 3.1 Conditions of the simulation

We consider a closed interactive system with  $m$  users. In the simulation, each user behaves as follows. A user issues a request to the Web server, where the service time of the request is chosen

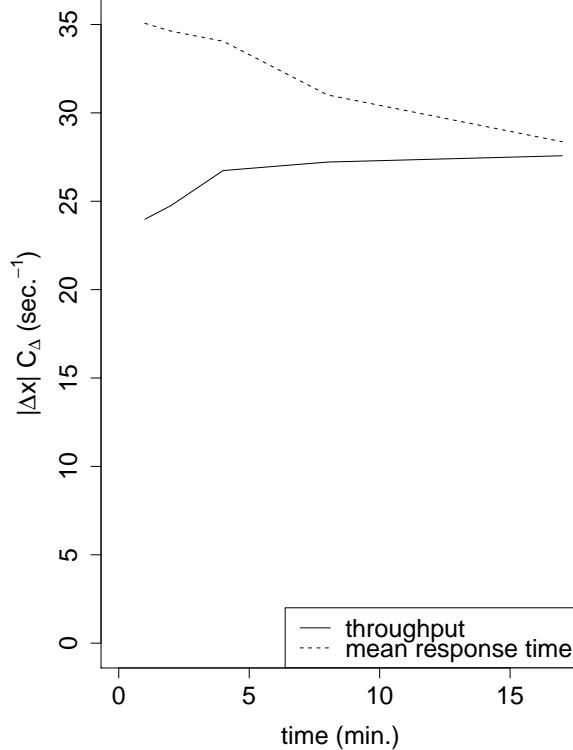


Figure 4: The average values of  $\tilde{D}_{\Delta X}(t)$  (solid line) and  $\tilde{D}_{\Delta R}(t)$  (dashed line) over the  $\binom{225}{2}$  pairs of configurations.

independently from a common service-time distribution with mean  $s$ . The request is processed by the Web server according to Processor Sharing, where each request receives equal share of the processing power. The result of the request is then returned to the user. After receiving the result, the user thinks before issuing the next request, where the think time is chosen independently from a common think-time distribution with mean  $z$ . We assume that the service times and the think times are independent.

To study the system at the steady state, we start the simulation at time  $t_0 = -10 m(z + s) < 0$ . Then each user would be expected to have issued ten requests by time 0 if the Web server were always busy. To increase the speed of convergence to the steady state, we choose the number of requests being processed at the Web server and the number of users thinking at  $t_0$  from the steady-state distribution [1]. Then we choose the remaining service-times from the service-time distribution and the remaining think-times from the think-time distribution. At time  $t$ , the throughput and the mean response time are estimated using the natural estimators defined with Equation (1). Recall that  $L(t)$  denotes the number of requests completed from time 0 to  $t$ , so that the completions during the warm-up period  $[t_0, 0]$  are ignored.

Throughout Section 3.2, we will study  $C_{\Delta R}(t)$  and  $C_{\Delta X}(t)$ , as defined with Equation (2), for a given pair of configurations. For  $i = 1, 2$ , we obtain  $x^{(i)}$  and  $r^{(i)}$  in Equation (2) from known analytical expressions [1]. The variances in Equation (2) are estimated with sample variances by

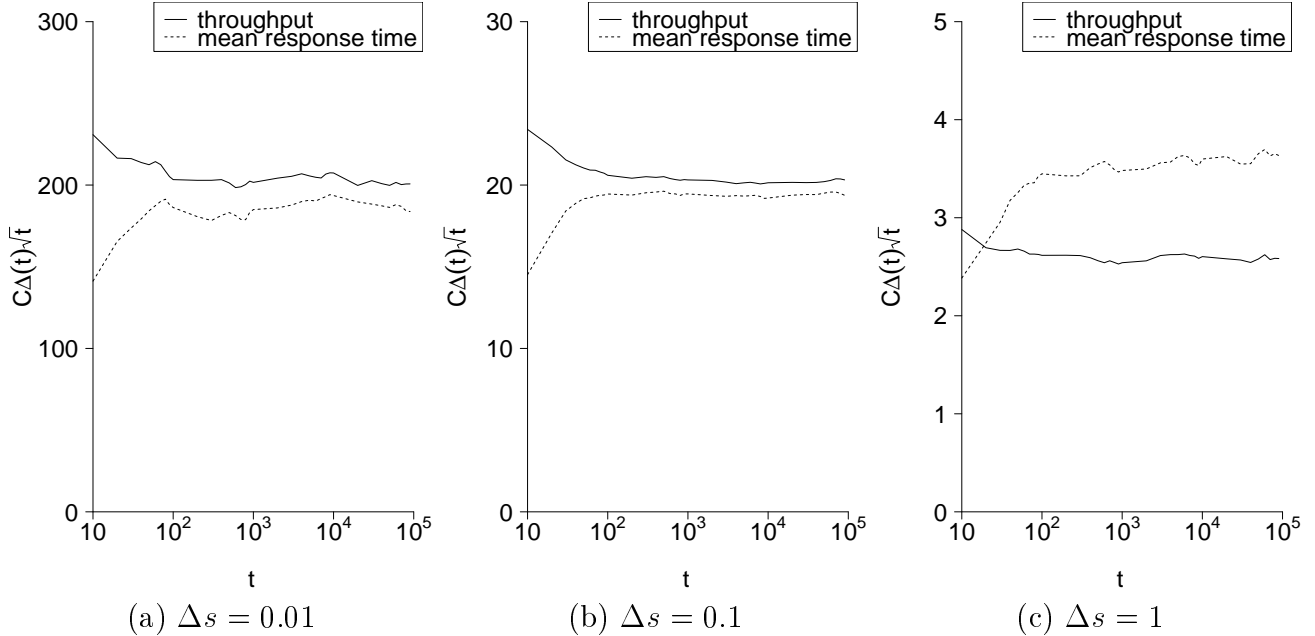


Figure 5: The impact of  $\Delta s$  on  $\sqrt{t}\tilde{C}_{\Delta X}(t)$  (solid lines) and  $\sqrt{t}\tilde{C}_{\Delta R}(t)$  (dashed lines). The service times have exponential distributions with varied means, the think time has an exponential distribution with  $z = 10$ , and  $m = 12$ . The two configurations differ in their mean service times: one is fixed at 1, and the other is  $1 + \Delta s$ , which is varied as indicated in each column.

repeating the simulation at least 1,000 times in each setting. For example, we estimate  $C_{\Delta X}(t)$  by  $\tilde{C}_{\Delta X}(t) = \sqrt{\tilde{V}_X^{(1)}(t) + \tilde{V}_X^{(2)}(t)}/|x^{(1)} - x^{(2)}|$ , where  $\tilde{V}_X^{(i)}(t)$  is the sample variance of  $\bar{X}^{(i)}(t)$  for  $i = 1, 2$  (note that  $\tilde{V}_X^{(i)}(t)$  in Section 2.3 is calculated differently from  $\tilde{V}_X^{(i)}(t)$  using only one sample). Analogously,  $\tilde{C}_{\Delta R}(t)$  denotes the estimated  $C_{\Delta R}(t)$ . In the figures in Section 3.2, we will show  $\sqrt{t}\tilde{C}_{\Delta R}(t)$  and  $\sqrt{t}\tilde{C}_{\Delta X}(t)$ , since  $\tilde{C}_{\Delta R}(t)$  and  $\tilde{C}_{\Delta X}(t)$  tend to decrease proportionally to  $1/\sqrt{t}$ .

### 3.2 Results of the simulations

Figure 5 shows  $\sqrt{t}\tilde{C}_{\Delta X}(t)$  with solid lines and  $\sqrt{t}\tilde{C}_{\Delta R}(t)$  with dashed lines. The two configurations differ only in the processing speeds of the Web system. Specifically, the mean service time of Configuration 1 is fixed as  $s^{(1)} = 1$ , and that of Configuration 2 is varied such that  $s^{(2)} = 1.01$  in Figure 5(a),  $s^{(2)} = 1.1$  in Figure 5(b), and  $s^{(2)} = 2$  in Figure 5(c). For all cases, the service times have exponential distributions, the number of users is  $m = 12$ , and the think time has an exponential distribution with mean  $z = 10$ . With these settings, we have  $r^{(1)} = 3.63$  and  $x^{(1)} = 0.880$ . For Configuration 2, we have  $r^{(2)} = 3.71$  and  $x^{(2)} = 0.875$  in Figure 5(a),  $r^{(2)} = 4.45$  and  $x^{(2)} = 0.831$  in Figure 5(b), and  $r^{(2)} = 14.1$  and  $x^{(2)} = 0.498$  in Figure 5(c).

Observe, in Figure 5, that  $\tilde{C}_{\Delta X}(t) > \tilde{C}_{\Delta R}(t)$  when the difference between the two configurations is small (Figure 5(a)), and that  $\tilde{C}_{\Delta X}(t) < \tilde{C}_{\Delta R}(t)$  for a sufficiently large  $t$  when the difference is

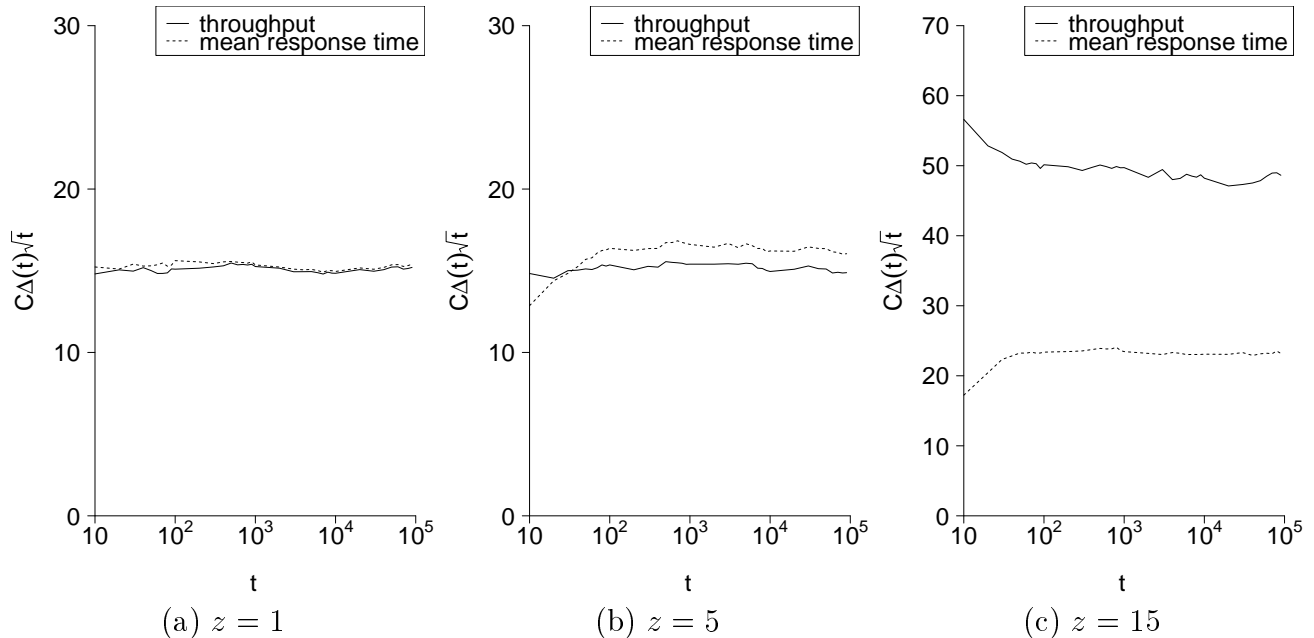


Figure 6: The impact of load on  $\sqrt{t} \tilde{C}_{\Delta X}(t)$  (solid lines) and  $\sqrt{t} \tilde{C}_{\Delta R}(t)$  (dashed lines). The settings are the same as in Figure 5(b) except that  $z$  is varied as indicated in each column.

large (Figure 5(c)). This observation can be explained by the convexity of the function,  $r = m/x - z$  (see Figure 3). When  $\Delta s \equiv |s^{(1)} - s^{(2)}|$  is small (Figure 5(a)),  $\Delta x \equiv |x^{(1)} - x^{(2)}|$  and  $\Delta r \equiv |r^{(1)} - r^{(2)}|$  are also small, which in turn suggests that  $|\Delta r / \Delta x| \approx m / (x^{(1)})^2$ . When  $\Delta s \equiv |s^{(1)} - s^{(2)}|$  is large (Figure 5(c)), the convexity suggests that  $|\Delta r / \Delta x| < m / (x^{(1)})^2$ . Hence, a rule of thumb is that mean response time tends to become relatively more reliable than throughput as the two configurations have closer performances.

Below we fix  $s^{(1)} = 1$  and  $s^{(2)} = 1.1$  to study the impact of changing other parameters on  $\tilde{C}_{\Delta X}(t)$  and  $\tilde{C}_{\Delta R}(t)$ . We first study the impact of load by varying  $z$ . Figure 6 shows  $\sqrt{t} \tilde{C}_{\Delta X}(t)$  with solid lines and  $\sqrt{t} \tilde{C}_{\Delta R}(t)$  with dashed lines, where the conditions of the simulations are the same as in Figure 5(b) except that  $z$  is varied as indicated in each column. Recall that  $z = 10$  in Figure 5(b). Figure 6(a) suggests that  $\tilde{C}_{\Delta X}(t) \approx \tilde{C}_{\Delta R}(t)$  when the load is high. When the load is intermediate (Figure 6(b)), we have  $\tilde{C}_{\Delta X}(t) < \tilde{C}_{\Delta R}(t)$  for a sufficiently large  $t$ . When the load is low (Figure 5(b) and Figure 6(c)), we have  $\tilde{C}_{\Delta R}(t) < \tilde{C}_{\Delta X}(t)$ .

The load is also affected by  $s$  and  $m$ . However, we find that the following rules of thumb hold regardless of the particular causes of high or low load: (i) when the load is low, mean response time tends to be more reliable than throughput; (ii) when the load is intermediate, throughput tends to be more reliable than mean response time; (iii) when the load is high, throughput and mean response time tend to have similar reliability. The rules of thumb will be explained analytically in Section 5.

Figure 7 illustrates the impact of the variability of the service times on  $\sqrt{t} \tilde{C}_{\Delta X}(t)$  and  $\sqrt{t} \tilde{C}_{\Delta R}(t)$ .

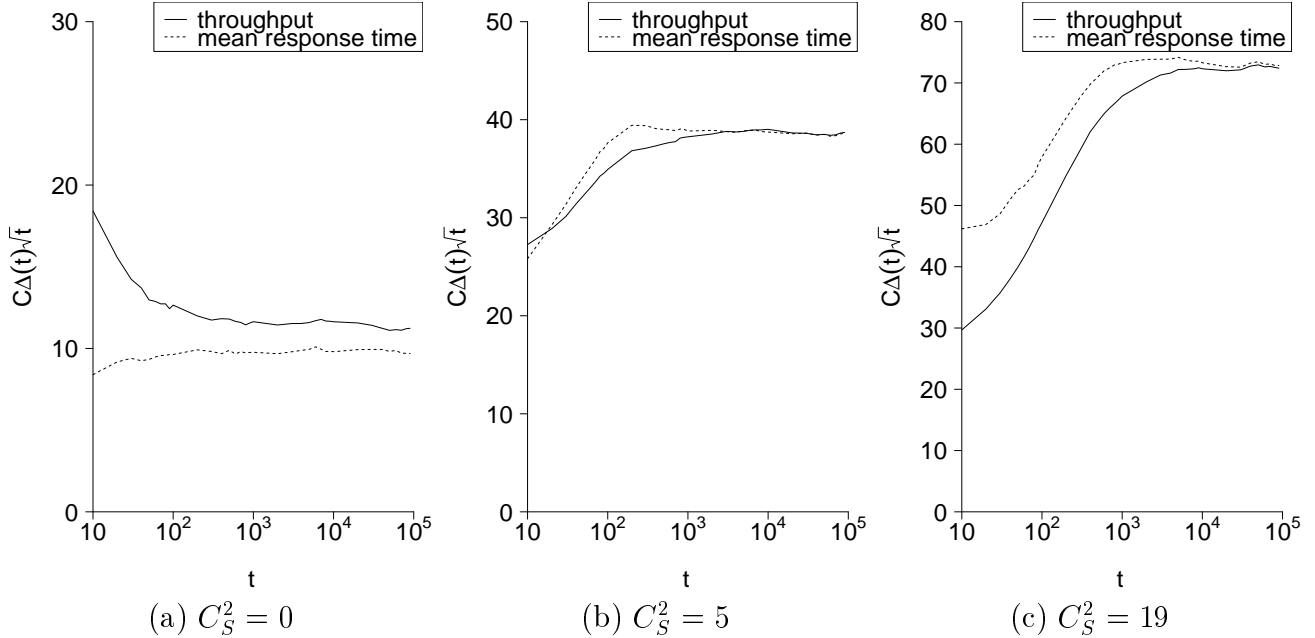


Figure 7: The impact of service time variability on  $\sqrt{t}\tilde{C}_{\Delta X}(t)$  (solid lines) and  $\sqrt{t}\tilde{C}_{\Delta R}(t)$  (dashed lines). The settings are the same as in Figure 5(b) except that the distributions of the service times are varied in such a way that the squared coefficient of variation,  $C_S^2$ , of  $S^{(1)}$  becomes as indicated in each column.

The conditions of the simulations are the same as in Figure 5(b) except that the distributions of the service times,  $S^{(i)}$  for Configuration  $i$  ( $i = 1, 2$ ), are varied in each column. Specifically, the service times are constants,  $S^{(1)} = 1$  and  $S^{(2)} = 1.1$ , in Figure 7(a). In Figures 7(b)-(c), the service times have Weibull distributions. Recall that the Weibull distribution with shape parameter  $a$  and scale parameter  $b$  has the cumulative distribution function,  $F(x) = 1 - \exp(-(x/b)^a)$ , for  $x \geq 0$ . In Figure 7(b),  $S^{(1)}$  has parameters,  $a = 1/2$  and  $b = 1/2$ , such that  $\mathbb{E}[S^{(1)}] = 1$  and  $\mathbb{V}(S^{(1)}) = 5$ ; also,  $S^{(2)}$  has parameters,  $a = 1/2$  and  $b = 1.1/2$ , such that  $\mathbb{E}[S^{(2)}] = 1.1$  and  $\mathbb{V}(S^{(2)}) = 5$ . In Figure 7(c),  $S^{(1)}$  has  $a = 1/3$  and  $b = 1/6$  such that  $\mathbb{E}[S^{(1)}] = 1$  and  $\mathbb{V}(S^{(1)}) = 19$ ; also,  $S^{(2)}$  has  $a = 1/3$  and  $b = 1.1/6$  such that  $\mathbb{E}[S^{(2)}] = 1.1$  and  $\mathbb{V}(S^{(2)}) = 19$ .

Comparing Figure 7 and Figure 5(b), we find that both  $\sqrt{t}\tilde{C}_{\Delta X}(t)$  and  $\sqrt{t}\tilde{C}_{\Delta R}(t)$  are larger when the service times are more variable. Observe that  $\sqrt{t}\tilde{C}_{\Delta R}(t)$  is more sensitive to the variability of the service time particularly when  $C_S^2 \equiv \mathbb{V}(S^{(1)})/\mathbb{E}[S^{(1)}]^2 \leq 1$  (Figure 7(a)). This makes intuitive sense, since the variability of the service time can directly affect the variability of the response time. Overall, a rule of thumb is that mean response time tends to become more reliable than throughput, when the service time is less variable.

Figure 8 illustrates the impact of the variability of the think time on  $\sqrt{t}\tilde{C}_{\Delta X}(t)$  and  $\sqrt{t}\tilde{C}_{\Delta R}(t)$ . The conditions of the simulations are the same as in Figure 5(b) except that distribution of the think time,  $Z$ , is varied in each column. In Figure 8(a), the think time is a constant,  $Z = 10$ .

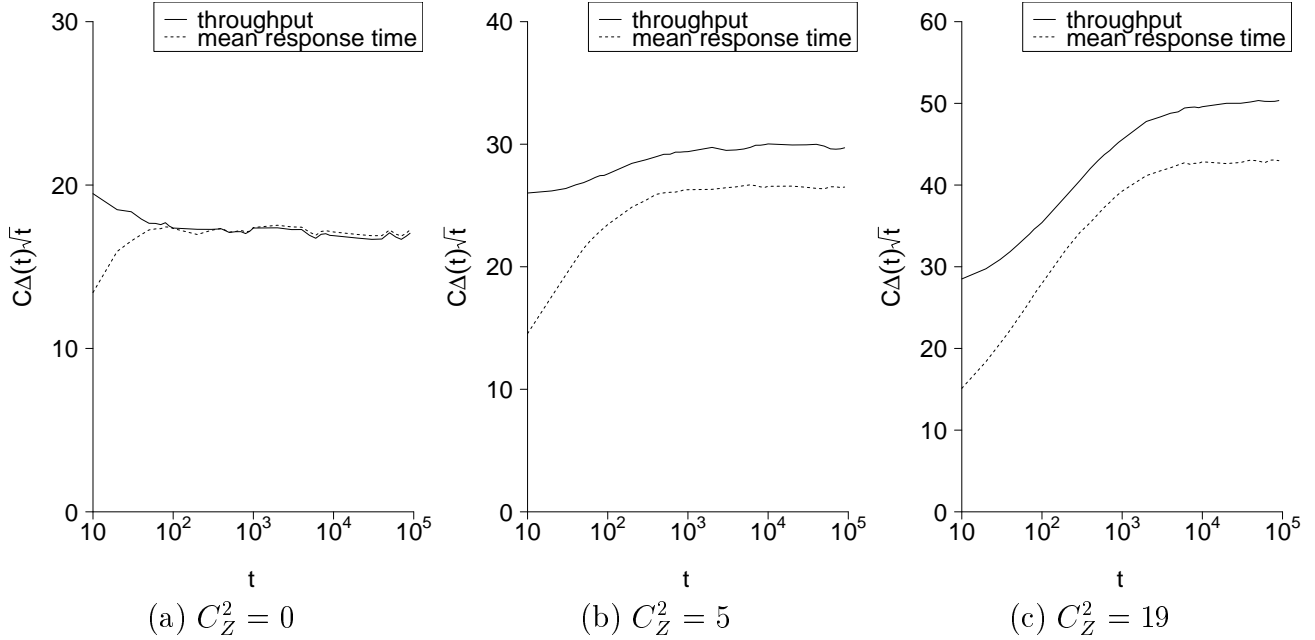


Figure 8: The impact of think time variability on  $\sqrt{t}\tilde{C}_{\Delta X}(t)$  (solid lines) and  $\sqrt{t}\tilde{C}_{\Delta R}(t)$  (dashed lines). The settings are the same as in Figure 5(b) except that the think time distribution is varied in such a way that the squared coefficient of variation becomes as indicated in each column.

In Figure 8(b),  $Z$  has a Weibull distribution with parameters,  $a = 1/2$  and  $b = 10/2$ , such that  $E[Z] = 10$  and  $V(Z) = 500$ . In Figure 8(c),  $Z$  has a Weibull distribution with  $a = 1/3$  and  $b = 10/6$  such that  $E[Z] = 10$  and  $V(Z) = 1,900$ .

Comparing Figure 8 and Figure 5(b), we find that both  $\sqrt{t}\tilde{C}_{\Delta X}(t)$  and  $\sqrt{t}\tilde{C}_{\Delta R}(t)$  are larger when the think time is more variable. Figure 8(c) suggests that  $\sqrt{t}\tilde{C}_{\Delta X}(t)$  is more sensitive to the variability of the think time particularly when  $C_Z^2 \equiv V(Z)/E[Z]^2 > 1$ . This sensitivity is in contrast to the service time variability but makes intuitive sense. Observe that the variability of the think time can directly affect the variability of the throughput but only indirectly affect the variability of the response time: i.e., the variability of the think time affects the variability of the number of requests being processed at the Web server, which in turn affects the variability of the response time. Overall, a rule of thumb is that mean response time tends to be more reliable than throughput when the think time is more variable.

A detail is that the simulation is repeated 10,000 times for Figures 7(b)-(c) and for Figures 8(b)-(c), while it is repeated 1,000 times for the rest of the figures in this section. This is because more repetition is needed to obtain stable results with more variable distributions.

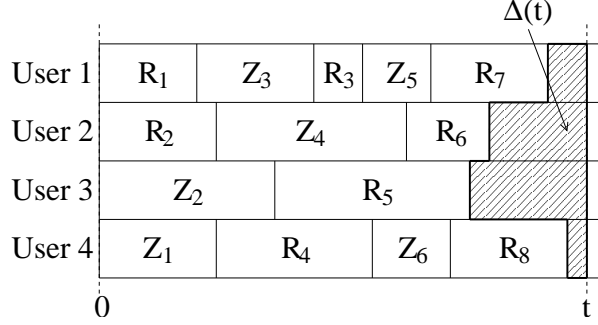


Figure 9: The shaded area illustrates the definition of  $\Delta(t)$  for a closed interactive system having  $m = 4$  users. Each user alternates between thinking (whose duration is  $Z_\ell$ ) and waiting (whose duration is  $R_\ell$ ).  $L(t) = 8$  response times and  $M(t) = 6$  think times are recorded by time  $t$ .

## 4 The response time law

To investigate the observations in the previous sections, we now refine the response time law with the central limit theorem (CLT). Although the response time law is frequently used, we cannot find its formal statement in the literature. In Section 4.1, we start by formally stating the response time law. In Section 4.2, we refine the response time law with the CLT.

### 4.1 A weak-convergence version

To state the response time law formally, let  $M(t)$  be the number of think times recorded up to time  $t$ . Let  $Z_\ell$  be the  $\ell$ -th recorded think time. Then the sample-average think times at time  $t$  is

$$\bar{Z}(t) = \frac{\sum_{\ell=1}^{M(t)} Z_\ell}{M(t)}. \quad (3)$$

Similar to  $\bar{X}(t)$  and  $\bar{R}(t)$  defined with Equation (1), note that  $\bar{Z}(t)$  is the natural estimator of the mean think time,  $z$ , at time  $t$ . If the weak law of large numbers holds for  $\bar{Z}(t)$ , then  $\bar{Z}(t) \Rightarrow z$  as  $t \rightarrow \infty$ . The response time law relates the limiting values of the natural estimators.

**Proposition 1 (The response time law (weak-convergence version))** *Let  $\Delta(t) \equiv mt - (L(t)\bar{R}(t) + M(t)\bar{Z}(t))$ , where  $\bar{X}(t)$ ,  $\bar{R}(t)$ , and  $\bar{Z}(t)$  are as defined with Equation (1) and Equation (3) for a closed interactive system with  $m$  users (see Figure 9 for an illustration of  $\Delta(t)$ ). Suppose that*

$$(\bar{X}(t), \bar{R}(t), \bar{Z}(t), \Delta(t)/t) \Rightarrow (x, r, z, 0) \quad (4)$$

as  $t \rightarrow \infty$ , where  $x$ ,  $r$ , and  $z$  are constants such that  $0 < x < \infty$  and  $0 \leq r, z < \infty$ . Then  $r = m/x - z$  holds among these constants.

**Proof:** By the definition of  $\Delta(t)$ , we have the following relation:

$$\frac{\Delta(t)}{t} + \frac{L(t)}{t} \bar{R}(t) + \frac{M(t)}{L(t)} \frac{L(t)}{t} \bar{Z}(t) = m. \quad (5)$$

Note that  $|L(t) - M(t)| \leq m$  by the definitions of  $L(t)$  and  $M(t)$ . Also, recall that  $L(t)/t = \bar{X}(t) \Rightarrow x$  as  $t \rightarrow \infty$ . Thus, it follows that  $M(t)/L(t) \Rightarrow 1$  as  $t \rightarrow \infty$ . Now, letting  $t \rightarrow \infty$  in Equation (5) and using Equation (4), we obtain  $x(r+z) = m$  by the continuous mapping theorem (see Section 3.4 from [16]). ■

## 4.2 A central limit theorem version

When appropriately normalized, the natural estimators are expected to converge in distribution to normal random variables under the conditions that the CLT holds. Namely, as  $t \rightarrow \infty$ , we expect to have  $\sqrt{t}(\bar{X}(t) - x) \Rightarrow N(0, V_X)$ ,  $\sqrt{t}(\bar{R}(t) - r) \Rightarrow N(0, V_R)$ , and  $\sqrt{t}(\bar{Z}(t) - z) \Rightarrow N(0, V_Z)$ , where  $N(\mu, V)$  denotes a normal random variable with mean  $\mu$  and variance  $V$ . Then, for a large  $t$ , it is suggested that  $\bar{X}(t) \approx N(x, V_X/t)$  and  $\bar{R}(t) \approx N(r, V_R/t)$ . Observe that  $V_X$  indicates how quickly  $\bar{X}(t)$  converges to  $x$ , and an analogous observation holds for  $V_R$ . To study how  $V_X$  and  $V_R$  compare to each other, we consider the following CLT version of the response time.

**Lemma 1 (A CLT version of the response time law)** *Suppose that a closed interactive system satisfies the assumptions made in Proposition 1 with Equation (4) replaced by a slightly stronger condition:*

$$\left( \bar{X}(t), \bar{R}(t), \bar{Z}(t), \Delta(t)/\sqrt{t} \right) \Rightarrow (x, r, z, 0). \quad (6)$$

*Under the condition that  $\sqrt{t}(\bar{Z}(t) - z) \Rightarrow N(0, V_Z)$  holds for a constant  $V_Z$ ,  $\sqrt{t}(\bar{X}(t) - x) \Rightarrow N(0, V_X)$  holds iff  $\sqrt{t}(\bar{R}(t) - r) \Rightarrow N(0, V_R)$  holds for constants,  $V_X$  and  $V_R$ . Also, let  $\mathbf{V}$  and  $\mathbf{W}$  be covariance matrices such that*

$$\mathbf{V} = \begin{pmatrix} V_R & V_{RZ} \\ V_{RZ} & V_Z \end{pmatrix} \quad \text{and} \quad \mathbf{W} = \begin{pmatrix} V_R & V_{RZ} & -\frac{m(V_R+V_{RZ})}{(r+z)^2} \\ V_{RZ} & V_Z & -\frac{m(V_Z+V_{RZ})}{(r+z)^2} \\ -\frac{m(V_R+V_{RZ})}{(r+z)^2} & -\frac{m(V_Z+V_{RZ})}{(r+z)^2} & \frac{m^2(V_R+V_Z+2V_{RZ})}{(r+z)^4} \end{pmatrix},$$

*and let  $\mathbf{0}$  denote a zero vector. Then, as  $t \rightarrow \infty$ , if it holds that*

$$\sqrt{t} \left( \bar{R}(t) - r, \bar{Z}(t) - z \right) \Rightarrow N(\mathbf{0}, \mathbf{V}), \quad (7)$$

*it also holds that*

$$\sqrt{t} \left( \bar{R}(t) - r, \bar{Z}(t) - z, \bar{X}(t) - x \right) \Rightarrow N(\mathbf{0}, \mathbf{W}). \quad (8)$$

Lemma 1 can be proved by following the framework developed by Glynn and Whitt in [3, 4, 5], where they prove a CLT version of Little's law (in Section 6, we will discuss the related work in more detail). For completeness, we provide our own self-contained proof of Lemma 1 in Appendix A.2.

## 5 Analytical study of reliability

The CLT version of the response time law allows us to compare the asymptotic reliability of throughput and that of mean response time. In Section 5.1, we introduce a formal framework to compare the asymptotic reliabilities of estimators. In Section 5.2, we discuss the results of the analysis.

### 5.1 Asymptotic reliability of an estimator

The reliability of an estimator can be evaluated using the asymptotic variance parameter. For example, for  $i = 1, 2$ , let  $x^{(i)}$  and  $\bar{X}^{(i)}(t)$ , respectively, be the true throughput of Configuration  $i$  and its natural estimator at time  $t$ . When the conditions in Lemma 1 are satisfied, we have  $\bar{X}^{(i)}(t) \approx N(x^{(i)}, V_X^{(i)}/t)$  for a large  $t$ . When the two configurations are measured independently,  $\bar{X}^{(1)}(t)$  and  $\bar{X}^{(2)}(t)$  are independent. Therefore, for a large  $t$ , we have  $\bar{X}^{(1)}(t) - \bar{X}^{(2)}(t) \approx N(x^{(1)} - x^{(2)}, (V_X^{(1)} + V_X^{(2)})/t)$ . This motivate us to use  $V_X^{(1)} + V_X^{(2)}$  as a metric for evaluating the asymptotic reliability of throughput.

To compare the asymptotic reliability of throughput and that of mean response time, we normalize the asymptotic variance parameters as follows:

**Definition 1** For  $i = 1, 2$ , let  $\bar{A}^{(i)}(t)$  be the estimator of a metric,  $A$ , for Configuration  $i$ . Similarly, let  $\bar{B}^{(i)}(t)$  be the corresponding estimator of a metric,  $B$ . For  $i = 1, 2$ , suppose that, as  $t \rightarrow \infty$ , we have  $\bar{A}^{(i)}(t) \Rightarrow a^{(i)}$ ,  $\bar{B}^{(i)}(t) \Rightarrow b^{(i)}$ ,  $\sqrt{t}(\bar{A}^{(i)}(t) - a^{(i)}) \Rightarrow N(0, V_A^{(i)})$ , and  $\sqrt{t}(\bar{B}^{(i)}(t) - b^{(i)}) \Rightarrow N(0, V_B^{(i)})$ , where  $a^{(i)}$ ,  $b^{(i)}$ ,  $V_A^{(i)}$ , and  $V_B^{(i)}$  are constants such that  $a^{(1)} \neq a^{(2)}$  and  $b^{(1)} \neq b^{(2)}$ . We say that  $A$  is **asymptotically more reliable than  $B$**  iff  $(V_A^{(1)} + V_A^{(2)})/(a^{(1)} - a^{(2)})^2 < (V_B^{(1)} + V_B^{(2)})/(b^{(1)} - b^{(2)})^2$ .

### 5.2 Analytical results

Suppose that the performances of two configurations of an interactive closed system are measured independently. Then the asymptotic reliability of throughput and that of mean response time are characterized by the following lemma:

**Lemma 2** Consider two configurations of an interactive closed system. Let  $x^{(i)}$  be the throughput and  $r^{(i)}$  be the mean response time of Configuration  $i$ . Also, let  $\bar{R}^{(i)}(t)$ ,  $\bar{Z}^{(i)}(t)$ , and  $\bar{X}^{(i)}(t)$ , respectively, be the natural estimators of mean response time, mean think time, and throughput of Configuration  $i$  at time  $t$ . We assume that  $(\bar{R}^{(1)}(t), \bar{Z}^{(1)}(t), \bar{X}^{(1)}(t))$  and  $(\bar{R}^{(2)}(t), \bar{Z}^{(2)}(t), \bar{X}^{(2)}(t))$  are independent. Also, we assume that the conditions in Lemma 1 hold for each configuration, so that  $\sqrt{t}(\bar{R}^{(i)}(t) - r^{(i)}, \bar{Z}^{(i)}(t) - z, \bar{X}^{(i)}(t) - x^{(i)}) \Rightarrow (\hat{R}^{(i)}, \hat{Z}^{(i)}, \hat{X}^{(i)})$  as  $t \rightarrow \infty$ , where  $\hat{R}^{(i)} = N(0, V_R^{(i)})$ ,  $\hat{Z}^{(i)} = N(0, V_Z)$ , and  $\hat{X}^{(i)} = N(0, V_X^{(i)})$ . Also, let  $V_{RZ}^{(i)}$  be the covariance between  $\hat{R}^{(i)}$  and  $\hat{Z}^{(i)}$ . Then throughput is asymptotically more reliable than mean response time iff

$$\left(\frac{x^{(1)}}{x^{(2)}}\right)^2 (V_R^{(1)} + V_Z + 2V_{RZ}^{(1)}) + \left(\frac{x^{(2)}}{x^{(1)}}\right)^2 (V_R^{(2)} + V_Z + 2V_{RZ}^{(2)}) < V_R^{(1)} + V_R^{(2)}. \quad (9)$$

**Proof:** Recall Definition 1, and let  $C_{\Delta X}^2 \equiv (V_X^{(1)} + V_X^{(2)})/(x^{(1)} - x^{(2)})^2$  and  $C_{\Delta R}^2 \equiv (V_R^{(1)} + V_R^{(2)})/(r^{(1)} - r^{(2)})^2$ . It suffices to show that  $C_{\Delta X}^2/C_{\Delta R}^2 < 1$  is equivalent to Inequality (9). Using the expression of  $\mathbf{W}$  in Lemma 1 and Equation (6), we can eliminate  $V_X^{(i)}$  in the expression of  $C_{\Delta X}^2$  and  $r^{(i)}$  in the expression of  $C_{\Delta R}^2$  for  $i = 1, 2$ . After simplification, we obtain that

$$\frac{C_{\Delta X}^2}{C_{\Delta R}^2} = \frac{\left(\frac{x^{(1)}}{x^{(2)}}\right)^2 (V_R^{(1)} + V_Z + 2V_{RZ}^{(1)}) + \left(\frac{x^{(2)}}{x^{(1)}}\right)^2 (V_R^{(2)} + V_Z + 2V_{RZ}^{(2)})}{V_R^{(1)} + V_R^{(2)}} \quad (10)$$

Substituting the above expression into  $C_{\Delta X}^2/C_{\Delta R}^2 < 1$  and multiplying the both hands with  $V_R^{(1)} + V_R^{(2)}$ , we obtain Inequality (9). ■

Although Lemma 2 provides exact conditions for throughput to be asymptotically more reliable than mean response time under the given assumptions, the conditions do not fully allow intuitive understanding. To gain further insights, we consider the limiting case where the performances of the two configurations are close to each other. Specifically, let  $x^{(1)} = x^{(2)}(1 + \varepsilon)$ , and assume that  $|\varepsilon| \ll 1$ . When the two configurations have very different performances, we can accurately identify the better of the two configuration with short measurement time regardless of whether we measure throughput or mean response time. Therefore, the case where  $|\varepsilon| \ll 1$  not only simplifies the expression in Lemma 2 but also is important in application.

**Theorem 1** *Let  $x^{(1)} = x^{(2)}(1 + \varepsilon)$  and assume  $|\varepsilon| \ll 1$ . Then, under the assumptions in Lemma 2, Inequality (9) can be expressed as*

$$V_Z + V_{RZ}^{(1)} + V_{RZ}^{(2)} + \left(V_R^{(1)} + 2V_{RZ}^{(1)} - V_R^{(2)} - 2V_{RZ}^{(2)}\right) \varepsilon + o(\varepsilon) < 0. \quad (11)$$

*In particular, as  $\varepsilon \rightarrow 0$ , Inequality (9) converges to the following expression:*

$$V_Z + V_{RZ}^{(1)} + V_{RZ}^{(2)} < 0. \quad (12)$$

**Proof:** Substitute  $x^{(1)} = x^{(2)}(1 + \varepsilon)$  into Equation (10) and simplify the equation with the relation,  $(1 + \varepsilon)^k = 1 + k\varepsilon + o(\varepsilon)$  for  $k = \pm 1$ . Then we obtain

$$\frac{C_{\Delta X}^2}{C_{\Delta R}^2} = 1 + 2\frac{V_Z + V_{RZ}^{(1)} + V_{RZ}^{(2)}}{V_R^{(1)} + V_R^{(2)}} + 2\frac{V_R^{(1)} + 2V_{RZ}^{(1)} - V_R^{(2)} - 2V_{RZ}^{(2)}}{V_R^{(1)} + V_R^{(2)}} \varepsilon + o(\varepsilon).$$

Substituting the above expression into  $C_{\Delta X}^2/C_{\Delta R}^2 < 1$  and multiplying the both hands with  $V_R^{(1)} + V_R^{(2)}$ , we obtain Inequality (11). Also, letting  $\varepsilon \rightarrow 0$  in Inequality (11), we obtain Inequality (12). ■

In Inequality (12), notice that  $V_{RZ}^{(i)}$  is a parameter that affects the covariance between  $\bar{R}^{(i)}(t)$  and  $\bar{Z}^{(i)}(t)$  for  $i = 1, 2$ . Specifically, under the assumptions of Lemma 1, the covariance can be approximated with  $V_{RZ}^{(i)}/t$  for a large  $t$ . We expect that the covariance is usually negative with the following reasoning: when the think times are shorter, the Web system receives more requests, which in turn makes response times longer. Therefore, Inequality (12) implies that throughput tends to be asymptotically more reliable than mean response time when the negative correlation between the think times and the mean response times are stronger or when the think time has lower variability.

To isolate the impact of the mean value on the variance and the impact the variance on the covariance, we use the following notations. For  $i = 1, 2$ , let  $\rho_{RZ}^{(i)} \equiv V_{RZ}^{(i)}/\sqrt{V_R^{(i)} V_Z}$ ,  $C_R^{(i)} \equiv \sqrt{V_R^{(i)}}/r^{(i)}$ , and  $C_Z \equiv \sqrt{V_Z}/z$ . Observe that  $C_R^{(i)}$  would be to the coefficient of variation as  $V_R^{(i)}$  was to the variance. Also,  $\rho_{RZ}^{(i)}$  would be to the correlation as  $V_{RZ}^{(i)}$  was to the covariance. Then Inequality (12) can be restated as follows:

$$\rho_{RZ}^{(1)} C_R^{(1)} r^{(1)} + \rho_{RZ}^{(2)} C_R^{(2)} r^{(2)} < -C_Z z. \quad (13)$$

We use Inequality (13) to gain further insights into the following rules of thumb that we have found in Section 3:

- i. mean response time tends to be more reliable than throughput when
  - a. two configurations have similar performances,
  - b. the load is low,
  - c. the service times have low variability, or
  - d. the think times have high variability;
- ii. throughput tends to be more reliable than mean response time when the load is intermediate;
- iii. throughput and mean response time tend to have similar reliability when the load is high.

Recall that, in Section 3, Rule i-a is explained by the convexity of the function,  $r = m/x - z$ . Thus, we consider the other rules.

Rules i-b, ii, and iii can be explained by the impact of the load on the correlation between the think times and the response times. When the load is very low, the Web system is idle most of the time regardless of how much the think times deviate from their mean value. Then most of the requests are processed individually without sharing the processing power. As a result, at the very low load, we have  $\rho_{RZ}^{(i)} \approx 0$ , which makes Inequality (13) unsatisfied. When the load is very high, the Web system keeps processing  $m$  or  $m - 1$  requests most of the time, and how the think times deviate from their mean value only has a small impact on the response times. It is when the load is intermediate that the think times and the response times have a relatively strong correlation, which tends to make Inequality (13) satisfied.

Rule i-c can be explained by the impact of the variability of the service time on the variability of the response time. Specifically, when the service time is less variable, the response time is less

variable, which in turn makes  $C_R^{(i)}$  smaller. Since  $\rho_{RZ}^{(i)}$  is usually negative as discussed above, a smaller  $C_R^{(i)}$  makes Inequality (13) less likely to be satisfied. Rule i-d can be explained similarly. When the think time is more variable,  $C_Z$  is larger, which in turn makes Inequality (13) less likely to be satisfied.

## 6 Related work

This paper is partially based on its preliminary version that appeared in [11]. However, in [11], we only study the *accuracy* of estimators for a single configuration and not the *reliability* of the estimators to identify the better of two given configurations. Also, Section 2 is entirely new. The experiments in Section 3 are new. Section 4 corresponds to Section 2 from [11]. However, the proof of Proposition 1 is new and the proof of Lemma 1 (in Appendix A.2) is considerably simplified. Section 5 is new, although some of the concepts are also used in [11].

The mathematical techniques needed to prove the CLT version of the response time law are introduced in [3, 4, 5], where a CLT version of *Little's law* is proved. Although the CLT version of the response time law can be proved based on what has been known in the literature, our interpretation of the CLT version of the response time law is entirely new. In particular, the problem of whether to measure mean response time or throughput in a closed interactive system is first studied formally in this paper. Our work is also related to *indirect estimation* techniques [8], which we briefly review below.

Indirect estimation is a technique that estimates a metric of interest by estimating another metric and calculating the metric of interest using a known relation between the two metrics. Indirect estimation is often asymptotically more efficient than directly estimating the metric of interest when the simulation time approaches infinity (*i.e.* an indirect estimator estimator often has a smaller coefficient of variation than a direct estimator). For example, Law shows that it is asymptotically more efficient to estimate the mean number of jobs,  $L$ , in an M/G/1 queue by estimating the mean delay,  $W$ , in the queue and calculating  $L$  via Little's law,  $L = \lambda W$ , than to estimate  $L$  directly [7]. The result in [7] is extended to a G/G/s queue by Carson and Law [2]. An important assumption in [2, 7] is that the arrival rate,  $\lambda$ , is known. Glynn and Whitt show that indirect estimation and direct estimation of  $L$  have the same asymptotic efficiency if  $\lambda$  is unknown and needs to be estimated [5].

Another example of indirect estimation is the use of a relation between the delay and the idle period in the GI/G/1 queue. The mean delay,  $W$ , in the GI/G/1 queue can be expressed using the mean equilibrium idle period,  $I_e$ , and a quantity that can be calculated from the service time and the interarrival time distributions (see p. 475 from [17]). Minh and Sorli show that, under heavy traffic, it is more efficient to estimate  $W$  in a GI/G/1 queue by estimating  $I_e$  and calculating  $W$  via the relationship between  $W$  and  $I_e$  than directly estimating  $W$  [10]. In [10],  $I_e$  is estimated by the first two sample moments of the idle period. Recently, Wang and Wolff propose a superior method for estimating  $I_e$  directly [15].

Yet another example of indirect estimation is the use of Little's law to estimate the blocking probability,  $p_b$ , in the G/G/s/0 model. Specifically, Little's law allows us to calculate  $p_b$  from the

mean number of busy servers, the arrival rate, and the mean service time. Srikant and Whitt show that it is more efficient to estimate  $p_b$  by estimating the mean number of busy servers and calculating the  $p_b$  via Little's law than directly estimating  $p_b$  as the ratio of losses to arrivals [13, 14].

## 7 Concluding remarks

The response time law could also be used for indirect estimation. For example, throughput may be estimated indirectly by estimating mean response time and calculating throughput using the response time law. However, indirect estimation via the response time law has not been investigated in the literature. The problem addressed in this paper could be restated in terms of indirect estimation. Then the problem would be whether direct estimation of throughput is more reliable than indirect estimation of throughput. Note, however, that we discuss the reliability of estimators *with respect to determining the better of two given configurations*. This is in contrast to the prior work, which studies the efficiency of estimators *with respect to determining the performance of a given configuration*. Our point of view is that, in the context of determining the better of two given configurations, it is logical to compare the estimator's reliability as defined in this paper rather than to compare the efficiency of direct and indirect estimation.

We emphasize that our study applies to a Web system whose performance is measured *with a workload generator* not to a Web system with users who are not under control. The use of a workload generator is a common practice to optimize the Web system before being used for critical purposes. The assumptions that  $m$  and  $z$  are known are justified with the use of a workload generator.

One might argue that Theorem 1 suggests that the rest of our work is not needed if we make the think time deterministic, since the theorem implies that throughput and mean response time have similar reliability when two given configurations have similar performances. However, we believe that we should set the think time stochastic in a workload generator. Even though the response time law implies that the variability of the think time has no impact on throughput and mean response time, the response time law requires assumptions that do not necessarily hold in practice. For example, cache might be effective when think time is short. Then stochastic think time implies that the cache is sometimes effective, which would be different with deterministic think time.

We also remark that, despite the assumptions that do not hold in practice, our simulation-based results and analytical results have practical relevance and implications. For example, our statistical analysis of the traces from the measurement of real Web systems suggests that throughput and mean response time can have significantly different reliabilities, which suggests the relevance of our simulation-based and analytical study. Our results also provide insights into the causes of the difference in the reliabilities.

## References

- [1] F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios. Open, closed, and mixed networks of queues with different classes of customers. *Journal of the ACM*, 22(2):248–260, 1975.
- [2] J. S. Carson and A. M. Law. Conservation equations and variance reduction in queueing simulations. *Operations Research*, 28(3):535–546, 1980.
- [3] P. W. Glynn and W. Whitt. A central-limit-theorem version of  $L = \lambda W$ . *Queueing Systems: Theory and Applications*, 1(2):191–215, 1986.
- [4] P. W. Glynn and W. Whitt. Ordinary CLT and WLLN versions of  $L = \lambda W$ . *Mathematics of Operations Research*, 13(4):674–692, 1988.
- [5] P. W. Glynn and W. Whitt. Indirect estimation via  $L = \lambda W$ . *Operations Research*, 37(1):82–103, 1989.
- [6] R. Jain. *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. John Wiley & Sons, New York, NY, 1991.
- [7] A. M. Law. Efficient estimators for simulated queueing systems. *Management Science*, 22(1):30–41, 1975.
- [8] A. M. Law and W. D. Kelton. *Simulation Modeling and Analysis*. McGraw-Hill, third edition, 2000.
- [9] E. D. Lazowska, J. Zahorjan, G. S. Graham, and K. C. Sevcik. *Quantitative System Performance: Computer System Analysis Using Queueing Network Models*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1984.
- [10] D. Minh and R. Sorli. Simulating the GI/G/1 queue in heavy traffic. *Operations Research*, 31(5):966–971, 1983.
- [11] T. Osogami. Relations in the central limit theorem version of the response time law. In *Proceedings of the Fourth International Conference on the Quantitative Evaluation of Systems*, pages 69–78, Edinburgh, Scotland, September 2007.
- [12] T. Osogami and S. Kato. Optimizing system configurations quickly by guessing at the performance. In *Proceedings of The SIGMETRICS 2007*, pages 145–156, San Diego, CA, June 2007.
- [13] R. Srikant and W. Whitt. Simulation run lengths to estimate blocking probabilities. *ACM Transactions on Modeling and Computer Simulation*, 6(1):7–52, 1996.
- [14] R. Srikant and W. Whitt. Variance reduction in simulations of loss models. *Operations Research*, 47(4):509–523, 1999.

- [15] C. Wang and R. W. Wolff. Efficient simulation of queues in heavy traffic. *ACM Transactions on Modeling and Computer Simulation*, 13(1):62–81, 2003.
- [16] W. Whitt. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer-Verlag, New York, NY, 2001.
- [17] R. W. Wolff. *Stochastic Modeling and the Theory of Queues*. Prentice Hall, Upper Saddle River, NJ, 1989.

## A Appendix

### A.1 Maximum likelihood estimator of the variance parameter

We argue that the estimators,  $\tilde{V}_X$  and  $\tilde{V}_R$ , used in Section 2.2 are the maximum likelihood estimators of  $V_X$  and  $V_R$ , respectively, under the conditions that  $L(t)$  and  $K(t)$  are Brownian motions. The assumption of a Brownian motion may be partially justified by the observation,  $\bar{X}(t) \approx N(x, V_X/t)$ , in Section 4.2. Since  $\bar{X}(t) = L(t)/t$ , we then have  $L(t) \approx N(xt, V_X t)$  for a large  $t$ .

Specifically, suppose that  $L(t)$  is a Brownian motion such that  $L(1)$  is a normal random variable with mean  $x$  and variance  $V_X$ . Then, given the observations,  $L(i)$  for  $i = 1, \dots, T$ , it can be shown that the maximum likelihood estimators of  $x$  and  $V_X$ , respectively, are  $\tilde{x}$  and  $\tilde{V}_X$  as defined in Section 2.1 and Section 2.2. Similarly, suppose that  $K(t)$  is a Brownian motion such that  $K(1)$  has mean  $r$  and variance  $V_R$ . Then, given the observations,  $K(i)$  for  $i = 1, \dots, T$ , the maximum likelihood estimators of  $r$  and  $V_R$ , respectively, are  $\tilde{r}$  and  $\tilde{V}_R$ .

### A.2 Proof of Lemma 1

We start by showing that Equation (7) implies Equation (8). Let  $\hat{R}$  and  $\hat{Z}$  be the random variables such that  $\sqrt{t}(\bar{R}(t) - r, \bar{Z}(t) - z) \Rightarrow (\hat{R}, \hat{Z})$  as  $t \rightarrow \infty$ . We will represent the limiting random variable of  $\sqrt{t}(\bar{X}(t) - x)$  in terms of  $\hat{R}$  and  $\hat{Z}$ .

After subtracting  $(rL(t) + zM(t))/t$  from the both hands of Equation (5), we multiply the both hands with  $\sqrt{t}$ . Then we obtain after simplification that

$$\begin{aligned} & \frac{\Delta(t)}{\sqrt{t}} + \frac{L(t)}{t} \sqrt{t}(\bar{R}(t) - r) + \frac{M(t)}{L(t)} \frac{L(t)}{t} \sqrt{t}(\bar{Z}(t) - z) \\ &= \left( m - xr - xz \frac{M(t)}{L(t)} \right) \sqrt{t} - \left( r + z \frac{M(t)}{L(t)} \right) \sqrt{t} \left( \frac{L(t)}{t} - x \right). \end{aligned} \quad (14)$$

By the response time law, we have  $m = x(r + z)$ , which can be substituted into the first term to obtain that

$$\left( m - xr - xz \frac{M(t)}{L(t)} \right) \sqrt{t} = xz \frac{(L(t) - M(t))/\sqrt{t}}{L(t)/t}.$$

Similar to the arguments in the proof of Proposition 1, we can show that  $(L(t) - M(t))/\sqrt{t} \Rightarrow 0$  and  $L(t)/t \rightarrow x$  as  $t \rightarrow \infty$ . Hence, the first term on the right-hand side weakly converges to 0 as  $t \rightarrow \infty$ . Also, recall from the proof of Proposition 1 that  $M(t)/L(t) \Rightarrow 1$  as  $t \rightarrow \infty$ .

Using the definition of  $\bar{X}(t) \equiv L(t)/t$ , we can solve Equation (14) for  $\sqrt{t}(\bar{X}(t) - x)$ . Now, we let  $t \rightarrow \infty$  and apply Equation (4) and Equation (6). Then, using the definitions of  $\hat{R}$  and  $\hat{Z}$ , we obtain  $\sqrt{t}(\bar{X}(t) - x) \Rightarrow -\frac{x}{r+z}(\hat{R} + \hat{Z})$  by the continuous mapping theorem. This implies Equation (8), since  $x = m/(r+z)$  by Proposition 1 and  $(\hat{R}, \hat{Z}) = N(\mathbf{0}, \mathbf{V})$ .

The above argument also shows that  $\sqrt{t}(\bar{X}(t) - x)$  converges to a normal random variable under the conditions that  $\sqrt{t}(\bar{R}(t) - r)$  and  $\sqrt{t}(\bar{Z}(t) - z)$  converge to normal random variables,  $\hat{R}$  and  $\hat{Z}$ , as  $t \rightarrow \infty$ . Hence, it only remains to show that  $\sqrt{t}(\bar{R}(t) - r)$  converges to a normal random variable under the conditions that  $\sqrt{t}(\bar{X}(t) - x)$  and  $\sqrt{t}(\bar{Z}(t) - z)$  converge to normal random variables as  $t \rightarrow \infty$ . We will represent the limiting random variable of  $\sqrt{t}(\bar{R}(t) - r)$  in terms of  $\hat{X}$  and  $\hat{Z}$ , assuming that  $\sqrt{t}(\bar{X}(t) - x, \bar{Z}(t) - z) \Rightarrow (\hat{X}, \hat{Z})$  as  $t \rightarrow \infty$ .

We can solve Equation (14) for  $\sqrt{t}(\bar{R}(t) - r)$ . Now, we let  $t \rightarrow \infty$  and apply Equation (4) and Equation (6). Then, using the definitions of  $\hat{X}$  and  $\hat{Z}$ , we obtain

$$\sqrt{t}(\bar{R}(t) - r) \Rightarrow -\frac{r+z}{x}\hat{X} - \hat{Z}(t) \tag{15}$$

by the continuous mapping theorem. Since  $\hat{X}$  and  $\hat{Z}$  are normal random variable, the right-hand side of Equation (15) is a normal random variable. This completes the proof of the lemma.