

The N-Burst/G/1 model with heavy-tailed service-times distribution *

Ronit Nossenson and Hagit Attiya
Dept. of Computer Science
Technion, Haifa, 32000, Israel
ronitt@cs.technion.ac.il, hagitt@cs.technion.ac.il

Abstract

This study introduces a new analytic queuing model, the N-Burst/G/1 model with heavy-tailed service-time distribution, which captures many of the issues that affect Web servers as observed by empirical studies. An asymptotic calculation of the model's waiting-time distribution is presented; this relies on calculating the waiting-time distribution in the M/G/1 model with heavy-tailed service-time distribution. Finally, using real data and simulation we verify the model's assumptions and demonstrate its accuracy.

1. Introduction

Empirical studies show that Web servers suffer from both self-similar document arrival process and heavy-tailed service-time distribution [7, 8, 14].

Until recently, models of Web servers consist of processes with either a realistic arrivals or a realistic service, but not both (for example, [12, 16]). Boxma and Cohen [4] consider a realistic heavy-tailed service-time distribution with a Poisson arrival process that is not self-similar. Respectively, other models [10, 17] consider an ON/OFF arrival process but assume Exponential service-time distribution, which is not heavy-tailed.

Recent simulations [15] show that in the typical situation, with medium or low utilization levels, processes with realistic arrivals or with a realistic service, but not with both, fail to estimate the *maximum number of clients in the system* and the *average response time*. Thus, an accurate queuing model for Web servers should consider both a realistic arrival process and a realistic service-time distribution.

It is difficult to obtain explicit performance parameters for a queue with general arrival process and general service-time distribution (G/G/1 queue) with heavy-tailed inter-

arrivals and/or service-time distributions [5]. Such a model is hard to analyze since it is not a Markovian process and it does not fit into any known solution of the G/G/1 queue which usually depends on moments that do not exist in the case of heavy-tailed distributions.

Two theoretical studies concerning such processes were proposed [5, 18]. Boxma and Cohen [5] obtain *heavy-traffic* results for the waiting time of a single server queue with heavy-tailed distributions. Xia and Liu [18] analyze the asymptotic tail distribution of stationary waiting times and stationary virtual waiting times in a single server queue with a long-range dependent arrival process and sub-exponential service times. Both studies conclude that the waiting time distribution is dominated either by the arrival distribution or by the service-time distribution, depending on which one has the heaviest tail. These models provide important breakthrough toward a realistic analytical model for Web servers. Yet, it is not clear how to apply these models when evaluating Web servers applications and/or policies. Moreover, no research have been done to study the accuracy of these models in the content of simulation and real data from a Web server system.

This paper introduces and evaluates a new stochastic queuing model for Web servers called the N-Burst/G/1 model with heavy-tailed service-time distribution. We calculate an asymptotic formula of the model's waiting-time distribution and evaluate its accuracy using real data and simulations. The model assumptions regarding the arrival and service processes distributions can be justified by a recent simulation study [15] as described in Section 3. Using the methods introduced in [13], it is simple to fit this model to real data.

The N-Burst/G/1 model combines the *N-Burst* arrival process with a class of heavy-tailed service-time distribution that was introduced in [4]. The N-Burst arrival process [17] is a superposition of N identical independent ON/OFF sources with matrix exponential truncated heavy-tailed ON time distribution and exponential OFF time distribution. A matrix exponential truncated heavy-tailed distribution [11] mimics a heavy-tailed distribution: its reliability function

* This research was supported by the *Israel Science Foundation* (grant number 105/01).

shows heavy-tailed behavior for a limited range, and drops off exponentially thereafter. Note that this arrival process is in fact a Markov Modulated Poisson Process (MMPP).

This paper calculates the model's waiting-time distribution, based on the waiting-time distribution of the M/G/1 model with identical service process. The tail exponent ν of the class of heavy-tailed distributions studied here is in the interval $[1..2]$. In [4], the calculation of the tail of the waiting-time distribution of the M/G/1 model is given together with a calculation of the waiting-time distribution in the case the tail exponent is equal to 1.5. Here we calculate asymptotic formula of this waiting-time distribution for a tail exponent that is a rational number $\nu \in [1, 2]$.

The N-Burst/G/1 model introduced herein is an extension of the N-Burst/ME/1 model [13] provided by replacing the Matrix-Exponential service-time distribution with a real heavy-tailed service-time distribution. Our simulations show that replacing this Matrix-Exponential distribution with real heavy-tailed distribution results in a more accurate model, which predicts the reality in a much more realistic (i.e., pessimistic) manner.

2. Definitions

In this section we describe some central concepts used in this paper.

Definition 2.1 A random variable X follows a heavy-tailed distribution (with tail index ν) if

$$P[X > t] \sim C \left(\frac{t}{\theta}\right)^{-\nu}, \text{ for } t \rightarrow \infty$$

where C and θ are positive parameters and $\nu > 1$.

Heavy-tailed distributions are characterized by extremely high variability, which increases sharply as ν decreases. Such a distribution has infinite variance; if $\nu \leq 1$, then it also has infinite mean.

Greiner et al. [11] introduce a family of hyper-exponential distributions called *matrix exponential truncated heavy-tailed* (in short, ME-THT), see Figure 1. They showed that these distributions with growing number of phases asymptotically approach heavy-tailed distributions.

Next we describe such distribution with T phases. By definition, the probabilities

$$p_i = \frac{[\theta^{i-1}(1-\theta)]}{(1-\theta^T)}, \quad i = 1, \dots, T$$

of entering phase i decay geometrically with the factor θ , $0 < \theta < 1$. The state holding times grow geometrically

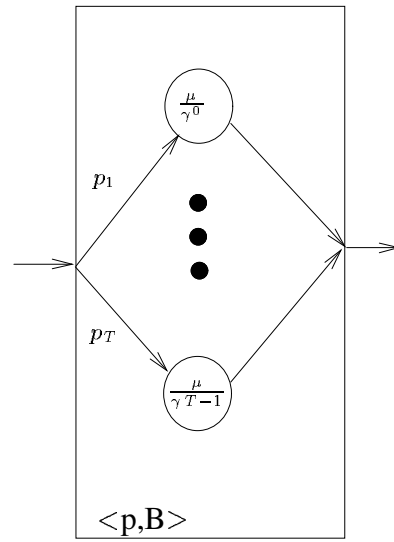


Figure 1. Phase diagram for the ME-THT distribution with T phases.

with the factor

$$\gamma = \frac{1}{\theta^{1/\alpha}}$$

This distribution has the following reliability function [11]:

$$R(x) = \frac{1-\theta}{1-\theta^T} \sum_{i=0}^{T-1} \theta^i \exp\left(-\frac{\mu}{\gamma^i} x\right)$$

In order to have heavy-tailed behavior with tail-exponent α and mean \bar{x} , we need to have [11]:

$$\mu = \frac{1-\theta}{1-\theta^T} \frac{1-(\theta\gamma)^T}{1-\theta\gamma} \frac{1}{\bar{x}}$$

The *heavy-tail range*, $Rng(R(x))$, is defined as the mean of the largest exponential phase. This largest phase is mainly responsible for the drop-off in the reliability function [11]:

$$Rng(R(x)) = \frac{\gamma^{T-1}}{\mu}$$

The variable θ can be chosen freely in the open interval $(0, 1)$. The larger the value of θ is, the more phases are necessary to obtain the same heavy-tail range.

3. N-Burst/G/1 Model with Heavy-Tailed Service Time Distribution

This section introduces the N-Burst/G/1 model with heavy-tailed service time distribution. Section 3.1 de-

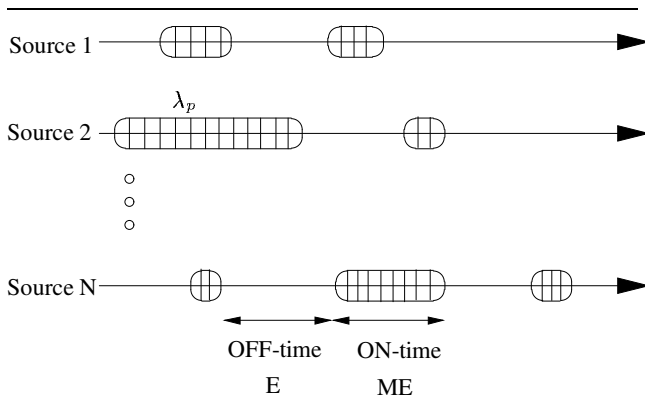


Figure 2. The N-Burst Arrival Process: traffic from N ON/OFF sources are multiplexed together.

scribes the N-Burst arrival process and Section 3.2 describes the class of service-time distributions.

3.1. Arrival Process

The N-Burst arrival process [17] is a superposition of traffic streams from N independent, identical sources of ON/OFF type (see Figure 2): each source emits requests at a Poisson-rate λ_p during its ON-time, and transmits nothing during its OFF-time. Let k be the mean rate of the individual source (the average for the ON- and OFF-times together), then the N sources collectively generate requests at mean rate $\lambda = Nk$.

The distribution of the duration of the ON periods is a matrix exponential truncated heavy-tailed (ME-THT), as described in Section 2. The choice of the actual ON time distribution has a great impact on the performance. The distribution of the OFF time duration is exponential.

These assumptions regarding the distributions of the ON periods, the OFF periods and the inter-arrivals periods (within the ON times) can be justified using a recent result [15]. Specifically, the behavior of the realistic arrival process in a Web server is effected mainly by the relation between the shape parameters of the distributions of the ON periods and OFF periods, and not from their absolute values [15]. Thus, it suffices to chose only one heavy-tailed distribution (and we chose the ON periods as ME-THT). In addition, the shape parameter of the inter-arrivals periods (within the ON times) has no effect on the behavior of the realistic arrival process [15]. Since this shape parameter is responsible for the heavy-tailed behavior, its lack of influence means that it suffices to use an exponential distribution for these inter-arrivals periods.

3.2. Service-Time Distribution

Our model uses the following class of service time distributions $B(t)$, $t \in [0, \infty)$, which was introduced in [4]:

$$B(t) = 1 - \frac{s^{2-\nu}}{\Gamma(2-\nu)} \delta \int_0^\infty e^{-s\theta} \frac{\theta}{(\theta+t)^\nu} d\theta \quad (1)$$

with $1 < \nu < 2$, $s > 0$, $0 < \delta \leq 1$, where s and δ are constants and $\Gamma(\cdot)$ is the Gamma function (Details regarding the Gamma function can be found in [2].)

The first moment of $B(t)$ is [4, Equation 2.11]:

$$\beta \triangleq \int_0^\infty t dB(t) = \frac{2-\nu}{\nu-1} \frac{\delta}{s} \quad (2)$$

The second moment of $B(t)$ is infinite. Note that $B(t)$ has the following asymptotic heavy-tail behavior [4, Equation 1.2]:

$$1 - B(t) = O(t^{-\nu}), t \rightarrow \infty$$

with $1 < \nu < 2$.

4. Waiting-Time Distribution

In this section we calculate the waiting time distribution of the N-Burst/G/1 model with heavy-tailed service-time distribution.

The indicator random variable $I(t)$ is 1 if and only if the number of sources that are simultaneously in their ON interval in the *entire* time interval $[0, t]$ does not change, otherwise, $I(t) = 0$. $P_I(t)$ denotes the probability $\Pr(I(t) = 1)$.

The indicator random variable $I(i, t)$ is 1 if and only if exactly i sources are simultaneously in their ON interval in the *entire* time interval $[0, t]$, otherwise, $I(i, t) = 0$. We denote $P_i(t) = \Pr(I(i, t) = 1)$. Note that $P_I(t) = \sum_{i=0}^N P_i(t)$.

The probability $P_i(t)$ is calculated as follows. Assume the ON time distribution has the reliability function $R(x) \sim \frac{1}{x^\alpha}$, as described in Section 2. The distribution of the time period with i independent bursts simultaneously in the ON intervals is heavy-tailed with exponent $i \cdot \alpha - (i - 1)$, with the reliability function, $R_i(x) \sim \frac{1}{x^{i \cdot \alpha - (i - 1)}}$ [17]. The probability $P_i(t)$ is equal to probability that none of these i simultaneously ON bursts finishes before time t , $R_i(t)$.

According to the Total Probability Theorem the probability for a client to wait for at least the time interval $[0, t]$, equal to the probability to wait for this time, given $I(t) = 1$ plus the probability to wait for this time, given $I(t) = 0$. That is,

$$W(t) = P_I(t) \cdot W(t|I(t) = 1) + (1 - P_I(t)) \cdot W(t|I(t) = 0) \quad (3)$$

The probability that at least one change occurs in the number of simultaneously ON bursts during the time interval $[0, t]$ is $O(P_I(t)^2)$. For example, consider the case where there are i_1 simultaneously ON bursts during the time interval $[0, t_1]$ and $i_1 + 1$ simultaneously ON bursts during the time interval $[t_1, t]$, that is, another source turns ON starting from point t_1 and on. Note that the probability for this scenario is $P_{i_1}(t_1) \cdot P_{i_1+1}(t - t_1)$. Thus, $W(t)$ is dominated by $P_I(t) \cdot W(t|I(t) = 1)$. Then,

$$\begin{aligned} W(t) &\simeq P_I(t) \cdot W(t|I(t) = 1) \\ &= \sum_{i=0}^N P_i(t) \cdot W(t|I(i, t) = 1) \end{aligned} \quad (4)$$

Since the arrival process is a Markov-modulated Poisson Process (MMPP) which is a special case of the Markovian arrival process, the Pollaczek-Khintchine formula holds. Thus, the calculation of the conditional waiting time $W(t|\dots)$ is equal to the waiting time in the M/G/1 model with the corresponding arrival rates.

The calculation of the arrival rate is as follows. Assume we have i sources that are simultaneously in their ON interval in the entire time interval $[0, t]$, and the other $(N - i)$ sources act as ON/OFF. Recall that each source emits requests at a Poisson-rate λ_p during its ON-time, and that k is the mean rate of the individual source (the average for the ON- and OFF-times together), as described in Section 3.1. Then the N sources collectively generate requests at mean rate $\lambda = i \cdot \lambda_p + (N - i) \cdot k$, where $i \cdot \lambda_p$ represents the rate generated by the i sources that are simultaneously in their ON interval in the entire time interval, and $(N - i) \cdot k$ represents the other $(N - i)$ ON/OFF sources.

To calculate the waiting time of our model it is left to calculate the waiting-time distribution of the M/G/1 queue with heavy-tailed service time distribution. This distribution was calculated in [4] for $\nu = 1.5$. Unfortunately, in the general case $\nu \in [1, 2]$, the calculation and the explicit expression for the waiting-time distribution was omitted from [4] and only the the distribution of its tail is specified. Thus, we next calculate this waiting-time distribution of the M/G/1 queue.

4.1. Waiting-Time Distribution of an M/G/1 Queue for a Rational Number $\nu \in [1, 2]$

In this section we study the waiting time distribution $W(t)$ of the M/G/1 queue with traffic load $a < 1$ and with service-time distribution $B(t)$ for the case $1 < \nu < 2$, as discussed in Section 3.2. The method used here to find this waiting time distribution is first to calculate its Laplace-Stieltjes transform (LST) and then calculate the waiting time distribution, using Doetsch's theorem [9].

The definition of the Laplace-Stieltjes transform of the service-time distribution $B(t)$ is as follows (for details regarding this transformation, see for example [1]).

$$\beta\{\rho\} \triangleq \int_0^\infty e^{-\rho t} dB(t)$$

with real $\rho \geq 0$. We denote the LST of $W(t)$ by $\omega\{\rho\}$.

Recall that β is the average of the service-time distribution $B(t)$ (Section 3.2). The Pollaczek-Khintchine formula states that for positive real ρ [6, Page 256],

$$\omega\{\rho\} = \frac{1 - a}{1 - a \frac{1 - \beta\{\rho\}}{\beta\rho}} \quad (5)$$

Let $0 < \mu = 2 - \nu < 1$. It was shown [4, Section 2, Equation 2.15]:

$$\frac{1 - \beta\{\rho\}}{\beta\rho} = \frac{1}{1 - \frac{\rho}{s}} + \frac{1}{\mu} \cdot \frac{\frac{\rho}{s}}{(1 - \frac{\rho}{s})^2} - \frac{1}{\mu} \cdot \frac{(\frac{\rho}{s})^{1-\mu}}{(1 - \frac{\rho}{s})^2}$$

with $s > 0$.

Assuming μ is rational, $\mu = \frac{M}{N}$, with $M < N$, $M, N \in \{1, 2, \dots\}$ and $\text{g.c.d.}(M, N) = 1$, we have:

$$\frac{1 - \beta\{\rho\}}{\beta\rho} = \frac{1}{1 - \frac{\rho}{s}} + \frac{N}{M} \cdot \frac{\frac{\rho}{s}}{(1 - \frac{\rho}{s})^2} - \frac{N}{M} \cdot \frac{(\frac{\rho}{s})^{\frac{N-M}{N}}}{(1 - \frac{\rho}{s})^2} \quad (6)$$

with $s > 0$. Define $y = (\frac{\rho}{s})^{\frac{1}{N}}$. Thus, $y^N = (\frac{\rho}{s})$. Equation 6 implies

$$\begin{aligned} \frac{1 - \beta\{\rho\}}{\beta\rho} &= \frac{1}{1 - y^N} + \frac{N}{M} \cdot \frac{y^N}{(1 - y^N)^2} - \frac{N}{M} \cdot \frac{y^{N-M}}{(1 - y^N)^2} \\ &= \frac{1 - y^N + \frac{N}{M}y^N - \frac{N}{M}y^{N+M}}{(1 - y^N)^2} \end{aligned} \quad (7)$$

Replacing $\frac{1 - \beta\{\rho\}}{\beta\rho}$ in Equation 5 with the result of Equation 7 yields

$$\begin{aligned} \omega\{\rho\} &= \frac{1 - a}{1 - a \left\{ \frac{(1 - y^N) + \frac{N}{M}y^N - \frac{N}{M}y^{N+M}}{(1 - y^N)^2} \right\}} \\ &= \frac{(1 - a) \cdot (1 - y^N)^2}{(1 - y^N)^2 - a \left[1 - y^N + \frac{N}{M}y^N - \frac{N}{M}y^{N+M} \right]} \end{aligned} \quad (8)$$

Define:

$$A(y) \triangleq (1 - y^N)^2 - a \cdot \left[1 - y^N + \frac{N}{M}y^N - \frac{N}{M}y^{N+M} \right]$$

Substituting $A(y)$ in Equation 8 yields

$$\omega\{\rho\} = (1 - a) \frac{(1 - y^N)^2}{A(y)} \quad (9)$$

Both $A(y)$ and $(1 - y^N)^2$ are polynomial of degree $2N$, and $y = 1$ is a zero of both of them. Thus, the function $\frac{(1-y^N)^2}{A(y)}$ has at most $2N - 1$ poles (singularity points). Denote the other zeros of $A(y)$ by $y_n, n = 1, 2, \dots, 2N - 1$, and assume first that the multiplicity of y_n is one for all $n = 1, \dots, 2N - 1$. We can write for Equation 9

$$\omega\{\rho\} = (1 - a) \sum_{n=1}^{2N-1} \frac{c_n}{y - y_n}$$

Each term in this sum can be written as an absolutely convergent power series in y . So we may put

$$\omega\{\rho\} = (1 - a) \sum_{k=0}^{\infty} w_k y^k \quad (10)$$

To determine the w_k , we will apply the Cauchy's integral theorem with the counterclockwise simple closed contour

$$C_\varepsilon \triangleq \{y : |y| < \varepsilon\}, 0 < \varepsilon < Y$$

With

$$|y| < Y \triangleq \min(|y_1|, |y_2|, \dots, |y_{2N-1}|)$$

Recall that the Cauchy's integral theorem states that if $f(z)$ is analytic in some simply connected region R , then

$$\int_C f(z) dz = 0$$

for any closed contour C completely contained in R . That is, the contour integral along any path not enclosing a pole is 0. Also recall that the Cauchy's integral formula states that

$$f(z_0) = \frac{1}{2\pi i} \int_C \frac{f(z) dz}{z - z_0}$$

where the integral is a contour integral along the contour C enclosing the point z_0 .

The residue of a function f around a point z_0 is defined by

$$Res_{z_0} f = \frac{1}{2\pi i} \int_C f(z) dz$$

where C is counterclockwise simple closed contour, small enough to avoid any other poles of f . Note that C_ε is by its definition indeed small enough to avoid any other poles, and hence

$$w_k = \frac{1}{2\pi i} \int_{C_\varepsilon} y^{-(k+1)} \frac{(1 - y^N)^2}{A(y)} dy \quad (11)$$

The integral in Equation 11 is regular in y , except at $y = 0$, and at the poles $y = y_n, n = 1, \dots, 2N - 1$. Thus,

$$w_k = - \sum_{n=1}^{2N-2} y_n^{-(k+1)} [(1 - y_n^N)^2] \cdot \lim_{y \rightarrow y_n} \frac{y - y_n}{A(y)} \quad (12)$$

If the multiplicity of y_n is not one for all $n = 1, \dots, 2N - 1$, that is, some poles with multiplicity larger than one, then the calculation of the residues at such poles requires the appropriate adaptation.

Denote $A^{(1)}(y) = \frac{d}{dy} A(y)$, so that

$$Y_n \triangleq \lim_{y \rightarrow y_n} \frac{y - y_n}{A(y)} = [A^{(1)}(y_n)]^{-1}$$

Also, define

$$Z_n \triangleq \frac{(1 - y_n^N)^2 Y_n}{y_n}$$

for $n = 1, \dots, 2N - 1$.

Substituting Z_n in Equation 12 yields

$$w_k = - \sum_{n=1}^{2N-2} y_n^{-k} \cdot Z_n \quad (13)$$

Using Equation 13 in Equation 10 implies

$$\begin{aligned} \omega\{\rho\} &= (a - 1) \sum_{k=0}^{\infty} \sum_{n=1}^{2N-2} y_n^{-k} \cdot Z_n \cdot y^k \\ &= (a - 1) \sum_{n=1}^{2N-2} Z_n \sum_{k=0}^{\infty} \left(\frac{y}{y_n}\right)^k \end{aligned}$$

Recall that $y = \frac{\rho}{s}$, so, we have

$$\omega\{\rho\} = (a - 1) \sum_{n=1}^{2N-2} Z_n \sum_{k=0}^{\infty} y_n^{-k} \left(\frac{\rho}{s}\right)^k$$

Applying Doetsch's theorem [9], for $s > 0$, and every finite $H \in \{0, 1, 2, \dots\}$

$$W(t) = (a - 1) \sum_{n=1}^{2N-2} Z_n \cdot \left[\sum_{k=0}^H \frac{1}{\Gamma(-\frac{k}{N})} y_n^{-k} (st)^{-\frac{(k+N)}{N}} + O((st)^{-\frac{(N+H+1)}{N}}) \right]$$

with $\Gamma(-m) = 0$, for $m = 0, 1, 2, \dots$, and $\Gamma(-m)\Gamma(1 + m) = -(\frac{\sin m\pi}{\pi})^{-1}$, m not an integer.

5. Simulations

We study the performance of the N-Burst/G/1 model using the *Client in the System* process. This process describes the number of clients in the system along the time axis, that is, the number of clients in the queue plus one (the client that is currently serviced). Other performance measurement parameters that we use are the *maximum number of clients* that are simultaneously in the system and the *average response time* experienced by the clients. These two parameters determine resources required by the system, such as buffer-size and computing power, and the system Quality of Service.

Section 5.1 describes a simulation series comparing the N-Burst/G/1 with its corresponding Markovian model, the N-Burst/ME/1 model [13]. To emphasize the differences between these models, we feed the same arrivals to both servers, the *G-server* with Pareto heavy-tailed service-times distribution and the *ME-server* with matrix exponential truncated heavy-tailed service-times distribution. Both servers share the same mean service-time. These simulations provide better understanding of the influence of the service-times distribution on the model's performance.

Section 5.2 describes a simulation analyzing the performance of these models comparing with the so-called *Realistic* process that considers both realistic arrivals and heavy-tailed service-times distribution. The comparison between the error factor of each model in estimating the maximum number of clients in the system and the average response time, demonstrates the improvement achieved by replacing the ME-server with the G-server.

5.1. Comparison Between the N-Burst/G/1 and the N-Burst/ME/1 models

Recall that to create heavy-tailed service behavior, the shape parameter of the service-times distribution should be smaller than 2.0 (see Section 2). Here we present simulation results with the following shape parameter values: 1.3 (Figure 3), 1.1 (Figure 4), 0.8 (Figure 5) and 0.6 (Figure 6), covering all different system behaviors observed by us. Specifically, we found out that there is no difference in the system behavior when the shape parameter is above 1.3. Also, the system is already overloaded when the shape parameter is smaller than 0.6.

Table 1 summarizes the comparison of the maximum number of clients in the system and the average response time in these models.

Following the characterization of [3, 8], with the parameter fitting method described in [13], the N-Burst arrival process has the following parameters. The number of clients

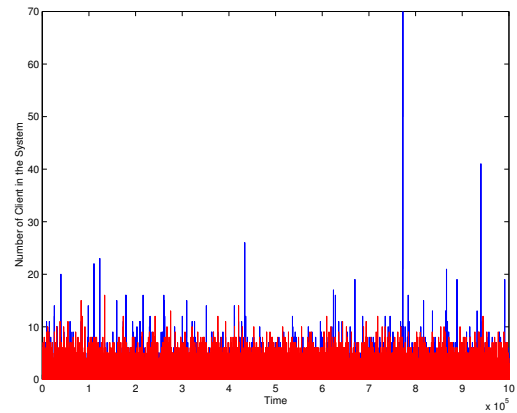


Figure 3. The client in the system processes of the N-Burst/G/1 (blue) and the N-Burst/ME/1 (red) models. The service-time shape parameter is 1.3

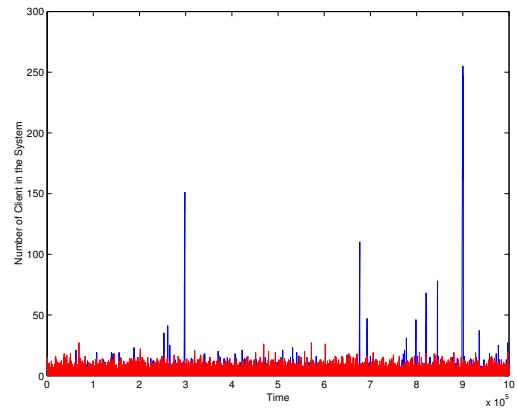


Figure 4. The client in the system processes of the N-Burst/G/1 (blue) and the N-Burst/ME/1 (red) models. The service-time shape parameter is 1.1

is 2; the ON-times shape parameter is 1.3; the ON-times θ is 0.5; the ON-times Number of phases is 5; the ON-times average is 1.0 the OFF-times average is 100.0; the inter-arrivals average is 0.5.

The immediate conclusion is that the N-Burst/G/1 model consistently represents a more bursty process than the N-Burst/ME/1 model. When the shape parameter of the service-times distribution is 1.3 or larger, both processes are rather smooth, with very small peaks (see Figure 3). This is a result of selecting the above mentioned arrival process parameters and in particular, choosing the On-times shape parameter as 1.3 too. (See [15] for a detailed discussion of the relation between the parameters of the arrival and the service processes.)

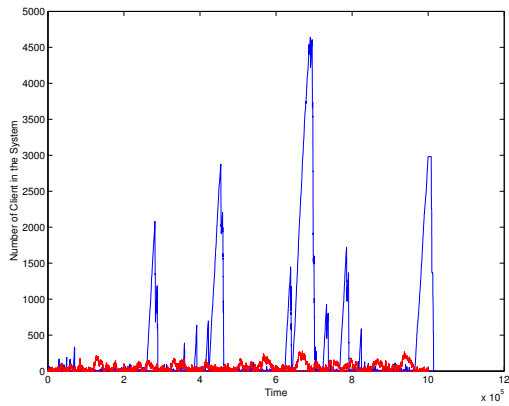


Figure 5. The client in the system processes of the N-Burst/G/1 (blue) and the N-Burst/ME/1 (red) models. The service-time shape parameter is 0.8

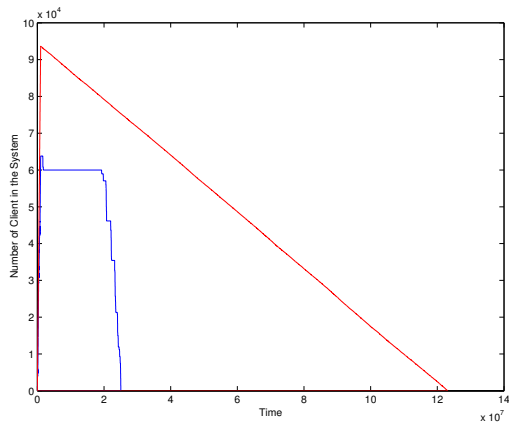


Figure 6. The client in the system processes of the N-Burst/G/1 (blue) and the N-Burst/ME/1 (red) models. The service-time shape parameter is 0.6

Decreasing the shape parameter of the service-times distribution below 1.3 and yet keeping the system not totally over-loaded (with service-times shape parameter larger than 0.6) highlights the difference between the models. As can be seen from Figures 4 and 5, the bursty nature of the N-Burst/G/1 model is emphasized by its graph line shape, which is much less smooth than the N-Burst/ME/1 graph's. In addition, as described in Table 1, the N-Burst/G/1 model predicts that the values of the maximum number of clients in the system and the average response time are larger in one order of magnitude than the corresponding values predicted by the N-Burst/ME/1 model.

Finally, when the system is over-loaded, although the N-Burst/ME/1 model predicts a more pessimistic scenario, it

Service shape parameter	Model	Maximum clients	Average response
1.3	N-Burst/G/1	70	0.729
	N-Burst/ME/1	16	0.545
1.1	N-Burst/G/1	255	10.306
	N-Burst/ME/1	27	1.867
0.8	N-Burst/G/1	4642	4053.430
	N-Burst/ME/1	278	580.036
0.6	N-Burst/G/1	63807	14273183.730
	N-Burst/ME/1	93659	61196254.758

Table 1. Comparison of the N-Burst/G/1 and the N-Burst/ME/1 models.

still represents a less bursty process. In both models, the clients accumulate rather fast. The difference between the models lays in the clients departures processes. As can be seen from Figure 6, the clients departures in the N-Burst/ME/1 graph's have a constant rate while the clients departures in the N-Burst/G/1 graph's have a variable rate (a 'stairs' graph). This phenomenon is a result of the extremely high variability of the service-times distribution in the N-Burst/G/1 model. On one hand, most of the samples in this distribution are very small, causing the steep decreases, on the other hand, once in a while a very large sample arrives and causes a long horizontal line.

5.2. Comparing the Realistic Process with the N-Burst/G/1 and N-Burst/ME/1 Models

In this section we describe a simulation comparing between three processes: the Realistic process, the N-Burst/G/1 and the N-Burst/ME/1. We create the Realistic process by feeding a Realistic arrival process into a G-server. The N-Burst/G/1 and the N-Burst/ME/1 processes are created by feeding the corresponding N-Burst arrival process into a G-server and a ME-server respectively.

We determine the parameters of the Realistic arrival and service processes using the characteristic of [3, 8] and the parameter fitting method described in [13] to construct the corresponding N-Burst arrival process and the ME-server. Specifically, the Realistic arrival and service processes have the following parameters. For the arrival process: the On-Times location parameter is 1.0 and the shape parameter is 1.3; the Off-Times location parameter is 50.0 and the shape parameter is 1.3; the inter-arrival times (within the On-times) location parameter is 0.1 and the shape parameter is 1.3. For the service-times distribution: the location parameter is 0.05 and the shape parameter is 0.8.

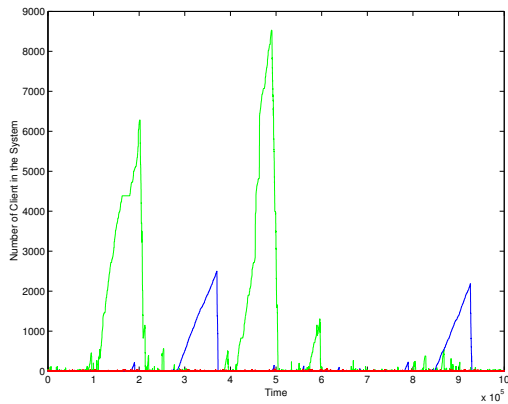


Figure 7. The client in the system processes of the R/R/1 (green), the N-Burst/G/1 (blue) and the N-Burst/ME/1 (red) models.

Model	Max clients	Error factor	Average response	Error factor
Realistic	8528	—	21875.618	—
N-Burst/G/1	2501	3.410	7532.427	2.904
N-Burst/ME/1	57	149.614	217.176	100.728

Table 2. Comparison of the Realistic process with the N-Burst/G/1 and the N-Burst/ME/1 models.

The simulation results appear in Figure 7. The comparison of the models maximum number of clients in the system and the average response time together with the calculation of their error factors appear in Table 2. The error factor is given by dividing the value of the real data result with the value of the model estimated result.

The simulations clearly indicate that the N-Burst/G/1 model is much more accurate than the N-Burst/ME/1 model. It has better estimations in several order of magnitude than the corresponding estimations predicted by the N-Burst/ME/1 model. Also, as can be seen from Figure 7, the N-Burst/G/1 model represents a more bursty process than the N-Burst/ME/1 model.

References

- [1] Mathworld web page. <http://mathworld.wolfram.com/>.
- [2] E. Artin. *The gamma function*. Holt, Rinehart and Winston, New York, 1964. translated from the German by M. Butler.
- [3] P. Barford and M. Crovella. Generating representative Web workloads for network and server performance evaluation. In *SIGMETRICS '98/ PERFORMANCE '98 Joint International Conference on Measurement and Modeling of Computer Systems*, pages 151–160, Madison, Wisconsin, United States, June 1998.
- [4] O. J. Boxma and J. W. Cohen. The M/G/1 queue with heavy-tailed service time distribution. *IEEE Journal on Selected Areas in Communications*, 16(5):749–763, June 1998.
- [5] O. J. Boxma and J. W. Cohen. Heavy-traffic analysis for the GI/G/1 queue with heavy-tailed distributions. *Queueing Systems*, 33(1):177–204, Jan. 1999.
- [6] J. W. Cohen. *The single server queue*. Elsevier science publishers B.V., Amsterdam, The Netherlands, 1982.
- [7] M. Crovella and A. Bestavros. Self-similarity in world wide web traffic evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846, Dec. 1999.
- [8] S. Deng. Empirical model of www document arrivals access link. In *Proceedings of the 1996 IEEE International Communications Conference (ICC'96)*, pages 17–23, Vancouver, Canada, June 1996.
- [9] G. Doetsch. *Handbuch der Laplace-Transformation*. Birkhauser Verlag, Basel, 1955.
- [10] A. Erramilli, O. Narayan, and W. Willinger. Experimental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Transactions on Networking*, 4(2):209–223, 1996.
- [11] M. Greiner, M. Jobmann, and L. Lipsky. The importance of power-tail distributions for modeling queueing systems. *Operations Research*, 47(2):313–326, Mar. 1999.
- [12] M. Harchol-Balter, B. Schroeder, N. Bansal, and M. Agrawal. Size-based scheduling to improve web performance. *ACM Transactions on Computer Systems (TOCS)*, 21(2):207–233, 2003.
- [13] R. Nossenson and H. Attiya. Evaluating web server performance with an extension of the N-burst model. Technical Report CS-2002-10, Department of Computer Science, Technion, 2002.
- [14] R. Nossenson and H. Attiya. The distribution of file transmission duration in the web. In *International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS)*, pages 647–654, Montreal, Canada, July 2003.
- [15] R. Nossenson and H. Attiya. Evaluating self-similar processes for modeling web servers. In *International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS)*, CA, USA, July 2004.
- [16] I. A. Rai, G. Urvoy-Keller, and E. W. Biersack. Analysis of las scheduling for job size distributions with high variance. In *Proceedings of the 2003 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, pages 218–228. ACM Press, 2003.
- [17] H. Schwefel and L. Lipsky. Impact of aggregated, self-similar on/off traffic on delay in stationary queueing models. In *Performance Evaluation 43*, pages 203–221, 2001.
- [18] C. H. Xia and Z. Liu. Queueing systems with long-range dependent input process and subexponential service times. In *Proceedings of the 2003 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, pages 25–36. ACM Press, 2003.