

Using Latency Quantiles to Engineer QoS Guarantees for Web Services

Ulrich Fiedler and Bernhard Plattner

Compute Engineering and Networks Laboratory
Swiss Federal Institute of Technology
ETH-Zentrum, Gloriastrasse 35
CH-8092 Zurich, Switzerland
{fiedler, plattner}@tik.ee.ethz.ch

Abstract. Simulations with web traffic usually generate input by sampling a heavy-tailed object size distribution. As a consequence these simulations remain in transient state over all periods of time, i.e. all statistics that depend on moments of this distribution, such as the average object size or the average user-perceived latency of downloads, do not converge within periods practically feasible for simulations. We therefore investigate whether the 95-th, 98-th, and 99-th latency percentiles, which do not depend on the extreme tail of the latency distribution, are more suitable statistics for the performance evaluation. We exploit that corresponding object size percentiles in samples from a heavy-tailed distribution converge to normal distributions during periods feasible for simulations. Conducting a simulation study with ns-2, we find a similar convergence for network latency percentiles. We explain this finding with probability theory and propose a method to reliably test for this convergence.

1 Introduction

Evaluating performance of web services for QoS purposes is a difficult problem due to the great variability of web traffic. An important characteristic of web traffic is that it usually shows bursts within a wide range of times scales [1]. This characteristic is called self-similarity and has been shown to be a consequence of a related observation, the heavy-tail in the size distribution of downloaded objects [2]. Modeling self-similarity in simulations with web traffic is important given that it has been shown that self-similarity has a significant negative impact on network performance [3] [4]. However, generating the input to self-similar web traffic in simulations by sampling a heavy-tailed object size distribution with infinite variance has severe implications on stability. Crovella and Lipsky [5] report that the convergence of the average object size of a sample to the average of the heavy-tailed object size distribution used to generate this sample requires simulation periods that are magnitudes too long to be practically feasible. Also, the distribution of output statistics that depend on all moments of the heavy-tailed object size distribution does not converge during practically feasible simulation periods. Therefore, the simulation remains in transient state for all practically

feasible simulation periods. A similar statement can be made, considering the fact that heavy-tails are always finite in any physical or simulation environment, as long as the tail is “sufficiently long”, e.g. several orders of magnitude beyond the average. Hence, to enable performance evaluation with simulation, there is a need to investigate meaningful output statistics that do not inherently have this dependency on moments of the object size distribution and thus can converge within feasible simulation periods.

In this paper, we take a end-user’s perspective in a client/server scenario for web services. We propose a performance analysis method for simulations with web traffic which is based on the latency quantiles of system components such as network, server/cache, client. The latencies of these components essentially sum up to the user-perceived latency of web downloads. We exploit (i) that latency quantiles are naturally suited to describe QoS and (ii) that quantiles of interest do not depend on the extreme tail of a distribution and hence not on the moments of the distribution. (i) can be explained with the fact that the p -th quantile of user-perceived latency of web downloads equating to t_0 seconds means that a p fraction of downloads is faster than t_0 . If p is represented by a percentage value, we call the p th quantile a *percentile*. Thus we argue that it is meaningful to characterize system performance for web traffic with high percentiles of the user-perceived latency such as the 95-th, 98-th, or 99-th percentile (see [6] for further details). A similar statement holds for the performance characterization of system components.

Therefore, in this paper, we explore the convergence of network latency quantiles. We exploit the fact that, if network utilization is not too high the relation between network latency and object size can be approximated as linear around latency quantiles of interest. In this case, a network latency quantile can converge when the corresponding object size quantile converges. From probability theory we know that the corresponding object size quantile converges to a normal distribution at rate $n^{-1/2}$ where n is the sample size. This convergence is fundamentally different from the convergence of the sample’s average object size to an α -stable distribution at rate $n^{1/\alpha-1}$ where α is the tail index of the heavy-tailed object size distribution. As a consequence, the amount of time required to converge object size quantiles of interest and thus latency quantiles is magnitudes smaller than the amount of time required to converge the average object size which is a minimum requirement to converge the whole system. This large difference in time to converge continues to hold under the assumption of realistic limits to the object size distribution inherent to common operating systems. Under higher utilization probability theory let us still expect convergence of latency quantiles to normal although the rate will be slower than $n^{-1/2}$. The sample size required to converge an object size quantile can then be viewed as an estimation of the initial phase of the convergence of the corresponding latency quantile.

Conducting a simulation study with ns-2, we find the expected convergence of the 95-th, 98-th, and 99-th network latency percentiles for both low and high utilization. With low utilization we mean utilizations that are reported average

to private networks (see [7]). With high utilization we mean utilizations that are known as an upper limit to what is acceptable during the busiest period (see [8] on provisioning procedures). We therefore propose a method that enables us to reliably test for convergence. In case of convergence the method additionally provides accurate estimates of the p -th network latency quantile which can be exploited to engineer QoS guarantees.

Finally, we argue that both, the estimation of the initial phase of the convergence, as well as the test method can also be applied to evaluate latency quantiles which are associated with system components other than the network.

Hence, the main research contributions of this paper are:

1. We give evidence that quantiles of user-perceived latencies are suitable statistics to evaluate performance of web services for QoS purposes.
2. We give evidence that latency quantiles, in contrast to other statistics such as the average latency, converge within an amount of time which is practically feasible for simulations. As a consequence, engineering QoS guarantees for latency quantiles of web services becomes feasible.
3. We provide lower bounds that estimate the initial phase required to converge latency quantiles.

The rest of this paper is structured as follows: In section 2 we review workload modeling with respect to convergence. In section 3 we determine sample sizes required to estimate object size quantiles in simulation. In section 4 we propose a method to test for convergence of network latency quantiles. In section 5 we apply this method to simulation results of a client/server scenario. Finally, we conclude in section 6.

2 Web Workload Modeling

In this section, we shortly review web workload modeling with respect to convergence.

For our analysis of convergence, we assume that the web traffic in the simulation is generated with a SURGE [9] type of model. We follow [10] and assume that the model accounts for probability distributions for the following user/session attributes:

- inter-session time between sessions from different users
- pages per session to quantify the number of web pages accessed within a session by the same user
- think time to quantify the time between completion of a download and initiation of the next request
- number of embedded objects per page
- inter-object time to quantify time between requests of embedded objects
- object size.

With respect to convergence, the probability distributions of interest are the object size distribution and the think time distribution since it is the heavy-tails

of these distributions that are the essential cause for the great variability and the self-similarity of web traffic [11] [4]. We say here that a distribution with cumulative density function (CDF) F is *heavy-tailed* with *tail index* α if

$$1 - F(x) \sim x^{-\alpha} \quad \text{for } n \rightarrow \infty \quad \text{with } \alpha \in (0, 2] \quad (1)$$

where $a(x) \sim b(x)$ means

$$\lim_{n \rightarrow \infty} \frac{a(x)}{b(x)} = 1$$

. We note that more general definitions are possible (see e.g. [12]). Effects from the heavy-tail in the object size distribution clearly dominate the effects from the heavy-tail in the think time distribution [4]. We therefore focus our analysis of simulation input on the object size distribution. We follow the approach of [10] and model the object size distribution with a *ParetoII* [13] distribution with CDF

$$F(x) = 1 - \frac{1}{(1 + \frac{x}{s})^\alpha} \quad x \in [0, \infty[\quad (2)$$

This ParetoII distribution has two free parameters: the average a , and the shape parameter α which equals to its tail index. $s = a * (\alpha - 1)$ is a dependent parameter.

3 Characterization of Simulation Input

In this section, we characterize object size quantiles in simulation input with respect to convergence to determine minimal simulation durations. We follow [5] and assume that for any dependent parameter in simulation output to converge, the corresponding parameter in simulation input has to converge. Thus, presumably for the 95-th, 98-th, or 99-th network latency quantile in simulation output to converge, the corresponding object size quantile in simulation input has to converge. Of course, much more input may be necessary to converge the 95-th, 98-th, or 99-th network latency percentiles. We employ quantile estimation techniques in statistics to estimate minimal sample sizes required to converge object size quantiles of interest from a ParetoII distribution.

3.1 Distribution of the Sample's Quantile

Under assumption of independent sampling, the expected value of the p -th sample's quantile is given by $x_p = F^{-1}(p)$. The probability density distribution of the p -th quantile of a random variable can then be derived as follows (see e.g. [14], section 3.7, p. 101).

Let $X_{(1)}, \dots, X_{(n)}$ be the ordered observations from a i.i.d random variable. Let $X_{(k)}$ be the p -th quantile where $k = np$ if np is an integer, and $k = \lfloor np + 1 \rfloor$ if np is not an integer. The event $x \leq X_{(k)} \leq x + dx$ occurs if $k - 1$ observations are less

than x , one observation is in the interval $[x, x + dx]$, and $n - k$ observations are greater than $x + dx$. The probability of any particular arrangement of this type is $F^{k-1}(x)f(x)[1 - F(x)]^{n-k}dx$. By the multinomial theorem, there are $n\binom{n-1}{k-1}$ such arrangements. Thus, the probability density distribution of the sample's p -th quantile is given by:

$$f_k(x) = n\binom{n-1}{k-1}(F(x))^{k-1}(1 - F(x))^{n-k}f(x) \quad (3)$$

The corresponding distribution $F_k(x)$ from which we can infer confidence intervals at a given sample size can be obtained by numerical integration. Moreover, it is of interest to denote that this distribution $F_k(x)$ converges to a normal distribution at rate $n^{-1/2}$. This follows from the following theorem:

Theorem 1 (Limit Theorem for Sample's Quantiles).

Let X_1, \dots, X_n be n independent observations on a random variable X with CDF F . Let $X_{(k)}$ be the p -th quantile where $k = np$ if np is an integer, and $k = \lfloor np + 1 \rfloor$ if np is not an integer. Let (i) $F(x)$ admit a continuous PDF $f(x)$ for all x . Further let (ii) the p -th quantile x_p of F be unique and $f(x_p) > 0$. Then the distribution of the sample's p -quantile $X_{(k)}$ converges to a normal distribution:

$$\sqrt{n}(X_{(k)} - x_p) \rightarrow \mathcal{N}(0, \sigma^2) \quad \text{for } n \rightarrow \infty \quad \text{with } \sigma = \frac{\sqrt{p(1-p)}}{f(x_p)}$$

For a proof of this limit theorem, which is the quantile's equivalent to the more commonly known central limit theorem (CLT) for the sample's average, refer to Rao [15], section 6f.2, p.423. The proof is essentially straightforward from Equation 3.

Both, Equation 3 and Theorem 1 can now be applied to evaluate the convergence of object size quantiles in simulation input. Theorem 1 applies for a heavy-tailed ParetoII distribution (see Equation 2 for CDF) since the ParetoII distribution fulfills the required regularity condition specified in the theorem. In detail, $f(x) = F'(x)$ of a ParetoII distribution is continuous for all $x \in [0, \infty)$ and all quantiles x_p of a ParetoII distribution are unique with $f(x_p) > 0$ since the CDF is strictly monotonous. This is although the ParetoII distribution does not fulfill the regularity condition of the central limit theorem (see [5]).

This leads to the following fundamental implications:

3.2 Implications

Object size quantiles obtained by sampling a heavy-tailed ParetoII distribution behave completely different than the average object size in the sample. The distribution of the p -th sample object size quantile converges to a normal distribution at rate $n^{-1/2}$ for sample size $n \rightarrow \infty$. Thus the p -th object size quantile can be estimated from relatively small samples since (i) the normal distribution is symmetric and has fast decaying exponential tails which lead to relatively small confidence intervals and (ii) the rate $n^{-1/2}$ is "fast". The distribution of

the average object size quantile converges to a α -stable distribution at a rate $n^{1/\alpha-1} < n^{-1/2}$ where $\alpha < 2$ is the tail index of the object size distribution (see [5]). Thus estimating the average object size requires extremely large samples since (i) the α -stable distribution is usually skewed and has itself heavy-tails, which, particularly for α close to 1, leads to very large confidence intervals and (ii) the rate $n^{1/\alpha-1}$ get extremely slow for $\alpha \rightarrow 1$.

To evaluate sample sizes required to estimate object size quantiles, we can now iterate the sample size n and determine the corresponding confidence intervals and expected values. We can perform this evaluation by numerical integration of the probability density given in Equation 3 (method 1) or employ Theorem 1 as a large sample approximation (method 2). Method 2 is computationally very cheap since approximations for the confidence interval immediately follow from evaluating the variance of the approximated normally distributed sample's quantiles:

$$s_n^2 = \frac{p(1-p)}{n * f^2(x_p)} \quad (4)$$

| Accuracy of the 99-th Percentile | Sample Size |
|----------------------------------|------------------|
| 1% | $2.8 \cdot 10^6$ |
| 2% | $7.2 \cdot 10^5$ |
| 3% | $3.2 \cdot 10^5$ |
| 5% | $1.2 \cdot 10^5$ |
| 10% | $3.1 \cdot 10^4$ |

Table 1. Sample Size Required to Estimate the 99-th Object Size Percentile

| Percentile (5% Acc.) | Sample Size |
|----------------------|------------------|
| 95-th | $2.6 \cdot 10^4$ |
| 98-th | $6.0 \cdot 10^4$ |
| 99-th | $1.2 \cdot 10^5$ |
| 99.9-th | $1.2 \cdot 10^6$ |
| 99.99-th | $1.1 \cdot 10^7$ |
| Average | 10^8 |

Table 2. Sample Size Required to Estimate other Object Size Percentiles

To perform a numerical evaluation we define the *accuracy* with which we can estimate a random variable such as the object size from a samples of size n as:

$$Accuracy = \max\left\{\left|\frac{L_n - E_n}{E_n}\right|, \left|\frac{U_n - E_n}{E_n}\right|\right\} \quad (5)$$

Here E_n is the expected value and L_n, U_n are the lower and upper bound of the confidence interval. We denote that evaluations in this paper are at confidence level 95%. To produce numerical values we assume that the average in the ParetoII object size distribution is 12KB and that the shape parameter $\alpha = 1.2$. These are the same values as in [10]. The exact values obtained with method 1 are listed in Table 1 and Table 2. The approximation to these values obtained with method 2 maximally differ by ± 0.2 in the mantissa.

We refer to Crovella and Lipsky [5] to compare these sample sizes to the sample sizes required to estimate the average object size from a sample. They analyze convergence to α -stable and roughly approximate the sample size required to estimate the average with a k digit accuracy as

$$n \geq c_2 * (10^{-k})^{-\frac{1}{1-1/\alpha}} \quad (6)$$

where $c_2 \approx 1$. This can be applied to estimate the average object size in the sample. Setting $10^{-k} = 0.05$ leads to the value of comparison which we added to Table 2.

This value of comparison shows that object size quantiles which are of interest to engineer guarantees on the response times of web downloads, converge at sample sizes which are magnitudes smaller than the samples sizes required to converge the average object size. This difference also holds at the presence of realistic bounds to the object size distribution inherent to common operating systems such as a 2.1GB or 4.2GB upper bound. A 2.1GB upper bound to the object size distribution leads to a required sample size of approximately 10^7 instead of 10^8 in Table 2 which can be calculated with the standard formulae based on the CLT.

Moreover, we can show that the difference between the sample size required to estimate quantiles and the sample size required to estimate the sample's average gets more pronounced when the tail index $\alpha \rightarrow 1$ (see Table 3). We denote that Table 3 lists estimates with 5% accuracy.

| Tail Index α | 99-th Percentile | Average (w/o bound) | Average (2.1GB bound) |
|---------------------|------------------|------------------------|--------------------------|
| 1.1 | $1.3 \cdot 10^5$ | 10^{14} | 10^{12} |
| 1.2 | $1.2 \cdot 10^5$ | 10^8 | 10^7 |
| 1.3 | $9.1 \cdot 10^4$ | 10^6 | 10^6 |

Table 3. Sample Size Required for Estimation of 99-th Quantile and Average

We summarize the findings of this section as follows: We have analyzed the convergence of object size quantiles in simulation input which is necessary that we can see convergence of latency quantiles in simulation output. We have applied quantile estimation techniques to derive this convergence which is to normal at a rate $n^{-1/2}$. This fundamentally differs from the convergence of the average object size to a α -stable distribution at a rate $n^{1/\alpha-1}$. As a consequence, the sample size required to converge the 95-th, 98-th, and 99-th object size quantile is several orders of magnitudes smaller than the sample size required to converge the average object size for α close to 1.

4 Characterization of Simulation Output

In this section, we analyze latency quantiles in simulation output with respect to convergence. We refer to probability theory to show that latency quantiles can be expected to converge to normal. We show how to reliably test for this convergence. We then conduct a simulation study to show that latency quantiles converge within periods which are practically feasible to simulations. As a consequence it becomes feasible to engineer QoS guarantees for network latency quantiles of web downloads.

Formally, convergence of latency quantiles cannot be treated in the same way as the convergence of object size quantiles. The initial assumption in Theorem 1, that the observations are independent, does not hold for the latency quantiles given that concurrent downloads can be from the same server or can share the bottleneck link on the network. Hence, the observed latencies are correlated. Literature on probability theory ([16] section 8.3) indicates that quantiles of correlated observations continue to converge to a normal distribution at a rate $n^{-1/2}$, where n is the sample size, if two conditions are fulfilled. First, a regularity condition on the distribution like (i) and (ii) in Theorem 1 is required. We think it is reasonable to assume such regularity for a latency distribution. In detail, this regularity means (i) to assume that the latency distribution F can for any latency be arbitrarily closely approximated with a differentiable function, (ii-a) that the latency associated with the quantile occurs in the simulation and (ii-b) that latencies very close to the quantile do also occur. Second, dependence of the observations must be sufficiently weak. Sufficiently weak dependence of observations means that autocorrelations of the observations decay so fast that that the convergence to normality at a $n^{-1/2}$ is not perturbed. This is e.g. the case when dependencies that depend on the lag only lead to autocorrelations which are summable over all lags (for details refer to [17]). We expect this to apply if network utilization is low. At higher utilization we expect that latency quantiles also converge to a normal distribution. However, this convergence is at a rate slower than $n^{-1/2}$ since the observations of latencies are known to be long-range dependent. The expectation can be justified with Theorem 8.2 in [17] since quantiles can be written as M-estimators.

4.1 Testing for Convergence to Normality

We propose to (i) produce and (ii) analyze normal plots for increasing sample sizes n to test when and whether latency quantiles become convergent. We (i) apply the frequently used normal plot¹ method (see e.g. Rice[14] p. 321–328) to a set of latency quantiles obtained from simulation runs with different seeds to the random number generator. This method only leads to qualitative results. We therefore (ii) enhance this method with a fully fledged statistical test to obtain quantitative results (see [14]) which can be successively monitored for increasing sample size n . We call this analyzing the normal plot. Moreover, we propose to (iii) monitor the rate of convergence. We call this consistency check.

The normal plot is produced as follows: Let $Y_{(k),j}$ be the latency quantile $Y_{(k)}$ estimated from a sample with index j which was obtained from a simulation run with a specific seed to the random number generator. Hence,

Simulation run 1 $\rightarrow Y_{1,1} \leq \dots \leq Y_{(k),1} \dots \leq Y_{max,1}$
 ...
 Simulation run $m \rightarrow Y_{1,m} \leq \dots \leq Y_{(k),m} \dots \leq Y_{max,m}$

To produce the normal plot, we arrange the estimated latency quantiles $Y_{(k),1} \dots Y_{(k),m}$ in ascending order:

$Y_{(k),1} \dots Y_{(k),m} \rightarrow Y_{(k),(1)} \dots Y_{(k),(m)}$.

Then we exploit that if this ordered set is consistent with normality, the expected value of $Y_{(k),(i)}$ is the $\frac{i}{m+1}$ quantile of a normal distribution with unknown parameters μ and σ :

$$E(Y_{(k),(i)}) = \mathcal{N}^{-1}(\mu, \sigma^2)\left(\frac{i}{m+1}\right) \quad (7)$$

Not knowing the parameters μ and σ of the normal distribution $\mathcal{N}(\mu, \sigma^2)$, we can exploit that any quantile of a normal distribution can be closely approximated with the corresponding quantile of the standard normal distribution $\mathcal{N}(0, 1)$. The approximation relation is (see [14]):

$$\mathcal{N}^{-1}(\mu, \sigma^2)\left(\frac{i}{m+1}\right) \approx \sigma * \mathcal{N}^{-1}(0, 1)\left(\frac{i}{m+1}\right) + \mu \quad (8)$$

Therefore a *normal plot* plots the $Y_{(k),(i)}$ against the $\frac{i}{m+1}$ quantile of the standard normal distribution. If the data in the set is close to normal distributed, the result of the plot is close to a straight line. Any deviation in the data from normality such as skewness or subexponential tails can be visually inspected. However, care needs to be taken in classifying a set as representing data which is consistent with a normal distribution. Due to the ordering process

¹ sometimes also called normal probability plot or Q-Q plot

$Y_{(k),1} \dots Y_{(k),m} \rightarrow Y_{(k),(1)} \dots Y_{(k),(m)}$ normal plots always tend to look somewhat linear.

To be reliable, we need to extend this qualitative test of visual inspection to a hypothesis test which produces quantitative results. Moreover, we want this test to provide accurate estimates of the parameters σ and μ of the normal distribution which can be exploited to engineer QoS guarantees for latencies. Therefore, we apply linear regression between the set and its expected values. The correlation coefficient of the linear regression, which quantifies the deviation from linearity, can now be exploited to test the hypotheses that the data in the ordered set is consistent with normality. For $i = 30$ data points [14] reports that, if the data is consistent with normality, 10% of plots have a correlation coefficient below 0.9707, 5%, have a coefficient below 0.9639, and 1% have a coefficient below 0.9490. Values for $i = 40$ are 0.9767, 0.9715, and 0.9597. These values for coefficients can thus be used in a hypothesis test as critical values at desired significance level. This hypothesis test should have sufficient power to distinguish a set consistent with a normal distribution from a set consistent with a heavy-tailed α -stable distribution given that the correlation is sensitive to outliers at the extremes of the ordered set. Moreover, the slope and intercept of the linear regression provide accurate estimates of the parameters μ and σ of the normal distribution.

To enhance the robustness of the test, we additionally monitor the rate of convergence, i.e. check the consistency. We take the estimate for the standard deviation from the linear regression and check whether this estimate is consistent with a $n^{-1/2}$ rate of convergence. We do this by plotting the estimated standard deviation times \sqrt{n} and check whether this is constant.

5 Results

We now apply this test method to investigate the convergence of network latency quantiles from a ns-2 (network simulator version 2) [18] simulation study of a client server scenario for web services. To perform this study we have extended ns-2 version ns-2.1b9a with our own implementation of the hyper text transfer protocol HTTP/1.1[19] on top of ns-2's FullTCP. The implementation is available via the ns-2 contributed code web page. The implementation explicitly models the HTTP interactions and includes HTTP/1.1 features like pipelining of requests for embedded pages and persistent connections between client and servers. To facilitate analysis, we limit user/session attributes in workload generation to the minimum set which is relevant to study convergence properties. This set includes the think time to quantify the time between completion of a download and initiation of the next request, the number of embedded objects per page, and object sizes. What we do not account for are session-related attributes which have distributions with fast converging tails. We also do not account for the effects of server and client latencies on the network. We also randomly chose the web server where a request goes to. We follow [10] with the choice of parameters for the think time, object size, and number of embedded objects per page distribu-

tions. We work with two models (see Table 4): The first model (“coarse model”) allows us to directly associate object sizes and download times to enable a in-depth analysis. The second set (“accurate model”) removes this simplification and models web workload at its full complexity. To keep results with both models comparable, we adjusted the think time in the first model such that the resulting average utilization of both workload model is equal. This means that we have kept the ratio between *average number of objects times average object size* and *average think time* constant.

| Workload Model | Object Size Distribution | Embedded Objects. Per Page Distribution. | Think Time Distribution |
|----------------|--|--|--------------------------------------|
| Coarse Model | ParetoII Average 12 KB Shape 1.2 | None | ParetoII Average 10s Shape 2.0 |
| Accurate Model | ParetoII Average 12 KB Shape 1.2 | ParetoII Average 3 Shape 1.5 | ParetoII Average 40s Shape 2.0 |

Table 4. Probability Distributions for Web Traffic Generation

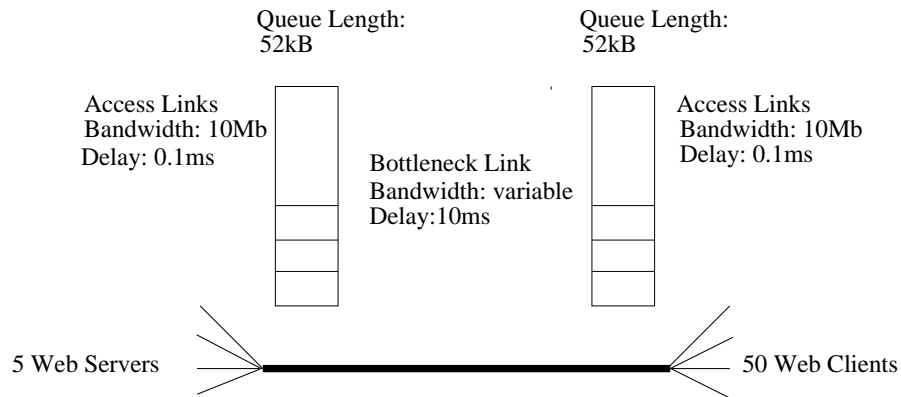


Fig. 1. Validation Topology

We then start with a simulating the dumbbell topology of Figure 1 which essentially models a bottleneck link. This bottleneck link can be viewed as an abstraction of a transoceanic link or a critical backbone link in private network. We assign a 10ms propagation delay to this bottleneck link which may also be viewed in a more general sense, since it has been argued that there is always a

single bottleneck link on any network path which is usually not fast moving[20]. Access links to the bottleneck have a capacity of 10 Mb/s and a propagation delay of 0.1ms. Clients and servers are attached to the access links. Queue sizes are set to 52KB.

5.1 Simulation Study

We start our investigation of convergence of network latency quantiles with the coarse model for web traffic generation (see Table 4 for parameters). We vary the capacity at the bottleneck link to obtain samples at different network utilizations. We define *link utilization* as the amount of traffic transported over the link per time unit in proportion to the link capacity. We consider three cases for the link capacity: 6400Kb/s, 2560Kb/s, and 640Kb/s. The 6400Kb/s case leads to a utilization of slightly more than 7% (see Table 5) which is roughly equivalent to what [7] reports as average in private networks. We then gradually increase utilization up to 64% which is known as an upper limit to what's acceptable during the busiest period (see [8] on provisioning procedures). We refer to the 6400Kb/s case as low utilization, to the 2560Kb/s case as medium utilization, to the 640KB/s case as high utilization.

| Capacity | Utilization | Loss Rate |
|----------|-------------|--------------|
| 640Kb/s | 64% | 0.8% |
| 2560Kb/s | 17% | $\leq 0.1\%$ |
| 6400Kb/s | 7.0% | $\leq 0.1\%$ |

Table 5. Bottleneck Link: Utilization and Loss Rate

Since our goal is to investigate convergence of latency quantiles we run very long simulations to generate 30, respectively 40, samples with different seeds to the random number generator. Each of these simulations terminate after the first 500,000 requested objects have been completely downloaded. This corresponds to approximately 28 hours of simulation time. Typically more than 500k objects have been completely downloaded during this period. At this sample size the 99-th object size percentile in simulation input has converged to 3% accuracy (see Table 1). The 98-th and 95-th quantile have converged even further. We have verified that the largest object size over all simulations runs at each utilization investigated is larger than 2.1GB, i.e. very close to the object size limit inherent to ns-2. We have also verified that neither the average object size nor the average network latency converge in our simulations.

At low utilization we find convergence which is consistent with a $n^{-1/2}$ rate for all latency quantiles investigated. Therefore, we investigate whether our findings about convergence of network latency percentiles under low utilization continue to hold when we model the full variability of the structure inherent to

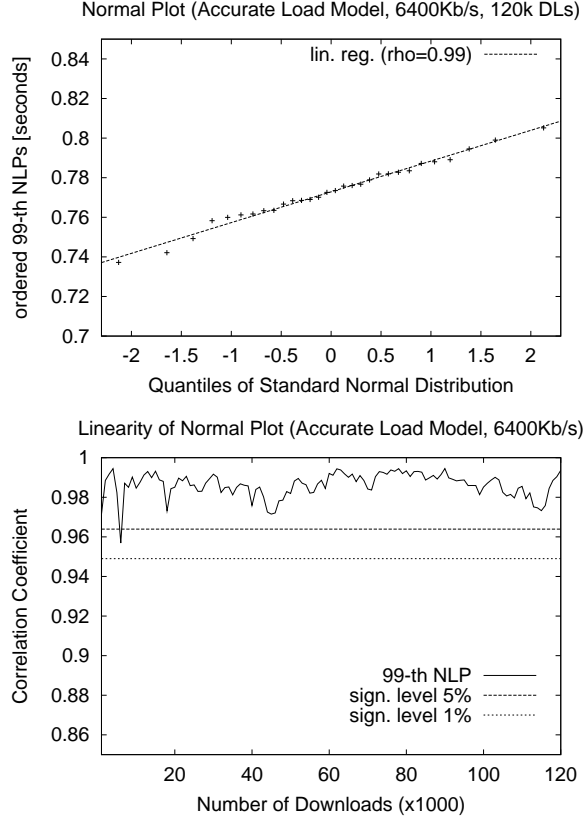


Fig. 2. Normal Plot and Linearity (99-th NLP, Accurate Model, Low Util.)

web pages. We thus repeat the simulations for low utilization with the accurate model for web traffic generation. We run simulation with 30 different seeds to obtain samples. Each of these simulations terminate after the first 120,000 requested web pages have been completely downloaded. We successively apply the convergence test and find that the 95-th, 98-th, and 99-th network latency percentile converge to normality at a $n^{-1/2}$ rate. Figure 2 depicts that the correlation coefficient from the normality test remains above all critical values after some initial phase. The same finding can be reported for the 95-th and 98-th network latency quantile. For all quantiles investigated, i.e. the 95-th, 98-th and 99-th percentile, the convergence is at a $n^{-1/2}$ rate (see Figure 3) given that deviations from constant are not larger than in the corresponding consistency check for object size quantiles in simulation input (not depicted).

Table 6 lists the sample sizes required to estimate the 99-th network latency quantile in simulation output at 5% accuracy. The listed sample sizes, which base on the accuracy definition given with equation 5, have been obtained as

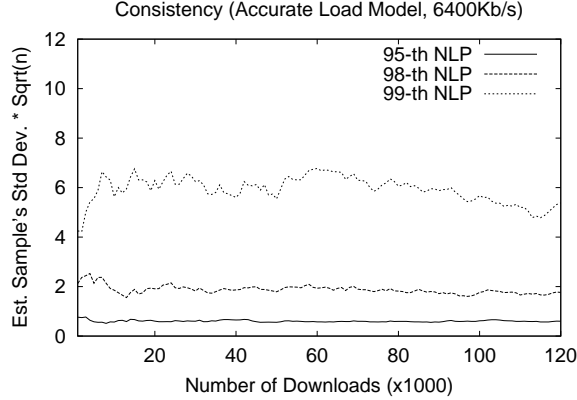


Fig. 3. Consistency of Convergence (Accurate Model, Low Util.)

| Accuracy | #Downloads |
|----------|------------------|
| 1% | $3.6 \cdot 10^6$ |
| 2% | $9.0 \cdot 10^5$ |
| 5% | $1.5 \cdot 10^5$ |
| 10% | $3.6 \cdot 10^4$ |

Table 6. Sample Size Required for Estimation of the 99-th Network Latency Percentile

follows: The expected latency quantile has been estimated from the normality plot at 120k downloads. The confidence interval radius has been approximated with $1.96 * s_n$ where s_n^2 is the quantile’s sample variance which has been evaluated with $s_n^2 = \sigma^2/n$. σ in turn has been estimated from the data of Figure 3. The values listed in Table 6 are slightly larger than in Table 1 which lists sample sizes required to converge object size quantiles in simulation input. A similar observation can be made for the 95-th and 98-th network latency quantile. Thus, under low utilization sample sizes obtained from evaluating the convergence of object size quantiles in simulation input turn out to be good approximations for sample sizes required to estimate latency quantiles from simulation output.

At medium utilization we do not find convergence for any of the latency quantiles investigated within samples sizes we have analyzed. We explain this finding with the fact that the latency distribution in the “estimated confidence interval” around the quantiles of interest does not exhibit sufficient regularity which is required for convergence. Presumably this comes from discontinuities in TCP’s reaction to minimal packet loss.

For the simulations with the coarse workload model we find that the 99-th and 98-th latency percentiles converge to a normal distribution at high utilization (see Figure 4 for results obtained with applying the test method described in section 4.1 to 40 simulation runs). The estimated confidence intervals for these

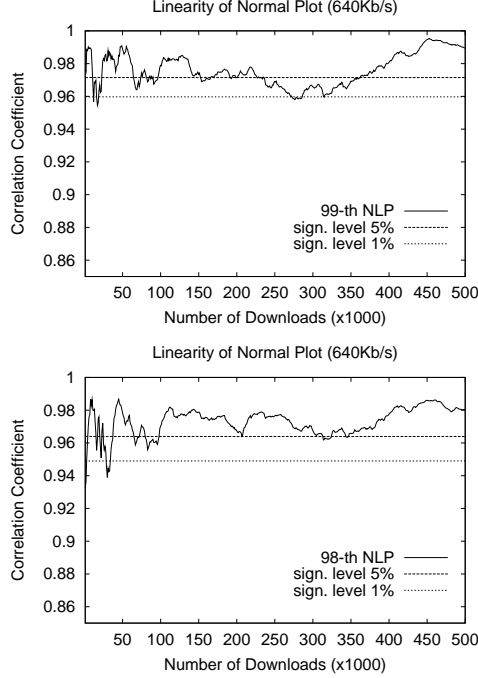


Fig. 4. Linearity of Normal Plot (Coarse Model, High Util.)

latency quantiles after 500,000 downloads are: 7.35 ± 0.92 seconds for the 99-th percentile and 5.08 ± 1.37 seconds for the 98-th percentile. The 95-th latency quantile does not converge within sample sizes that we have analyzed. However, the convergence of the 99-th and 98-th latency percentile is not consistent with a $n^{-1/2}$ rate (see Figure 5 for a log-log plot of sample variance vs. sample size). Such convergence at a $n^{-1/2}$ rate would result in a line parallel to the reference line entitled with Hurst parameter $H = 0.5$. The convergence is also not completely consistent with a slower rate $n^{-\beta}$ with $\beta < 1/2$ which is expected for a long-range dependent correlation structure among observations of latency quantiles. Such a correlation structure would result in a straight line with smaller slope (see e.g. the reference line for Hurst parameter $H = 0.9$ which is to be expected for the corresponding on/off process (see [11])). Moreover, the 99-th and the 98-th latency percentiles converge at different rates which is to be explained with the fact that the simulation has not yet reached stability. Nevertheless, we argue that it is possible to give guarantees for these latency quantiles based on estimating an upper bound of the confidence intervals to the variance of latency quantiles at sample size n . Such an estimation can be obtained by grouping simulations and evaluating the variance of latency quantiles at sample size n for each group. In our case this implies to perform e.g. 20 times 40 simulation runs

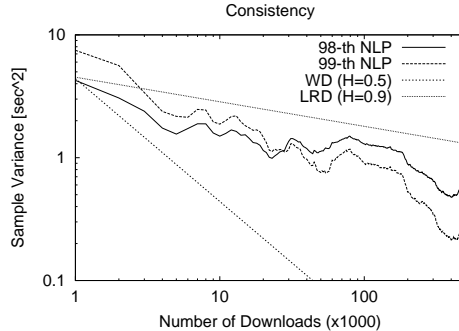


Fig. 5. Consistency of Convergence (Coarse Model, High Util.)

instead of 40 to estimate the confidence interval bounds. However, for practical applications some rough approximation from Figure 5 may already be sufficient.

We summarize the findings of this sections as follows: We have referred to probability theory to explain that the sample's p -th latency quantile can converge to normal at a $n^{-1/2}$ rate, where n is the sample size, if utilization is low. If utilization is high, the sample's p -th latency quantile can continue to converge to normal. However, the rate will be slower than $n^{-1/2}$. Hence, we have proposed a method which enables us to reliably test for such convergence. This method is based on (i) producing normal plots, (ii) analyzing normal plots by monitoring the correlation coefficient which quantifies the linearity of the plot, and (iii) checking the rate of convergence. In case of convergence this methods additionally provides accurate estimates of the p -th latency quantile. We have applied this method to the output of a simulation study with ns-2. We have observed that network latency quantiles in simulation output converge to a normal distribution at rate $n^{-1/2}$ if the utilization is low. We have also observed that network latency quantiles continue to converge to normal with a slower rate if utilization is high.

6 Summary and Conclusion

In this paper, we have investigated whether the 95-th, 98-th, or 99-th percentile of user-perceived latencies are suitable statistics to measure performance of web services and hence to engineer QoS guarantees for web services. We have exploited that (i) latency quantiles have a natural interpretation in evaluating QoS, (ii) quantiles do not depend on the extreme tail of the distribution and thus not on moments of the distribution, and (iii) user-perceived latency is a sum of the latencies of system components, which essentially are network, server/cache, and client. We have analyzed the convergence of simulation input to determine minimal simulation durations necessary to estimate the latency quantiles of interest from simulations. We have applied quantile estimation techniques to derive the

convergence of the p -th object size quantile which is to a normal distribution at a rate $n^{-1/2}$ where n is the sample size. This convergence is fundamentally different from the convergence of the average object size to a α -stable distribution at a rate $n^{1/\alpha-1}$. As a consequence, the sample size required to converge the 95-th, 98-th, and 99-th object size quantile is several orders of magnitudes smaller than the sample size required to converge the average object size. The large difference in amount of time to converge continues to hold under the assumption of realistic limits to the object size distribution inherent to common operating systems.

We have referred to probability theory to explain that latency quantiles under low utilization converge to a normal distribution at a $n^{-1/2}$ rate. We have proposed a method to reliably test for such convergence. We have validated the test method in a simulation study with ns-2. Moreover, we have found that network latency quantiles converge to normal distributions under high load which we also explain with probability theory.

As a consequence engineering QoS guarantees for web services based on latency quantiles becomes feasible since we argue that the proposed method can as well accurately estimate latency quantiles associated with server/cache and client. In order to further clarify this issue, we plan to explore the impact of network topology and document popularity on our results.

7 Acknowledgements

We would like to thank many people for helpful discussions particularly Martin Maechler, Polly Huang, and Samarjit Chakraborty.

References

1. W.E.Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson, "On the Self-Similar Nature of Ethernet Traffic (Extended Version)," *IEEE/ACM Transactions on Networking*, vol. 2, no. 1, pp. 1–15, Dec. 1994.
2. M. Crovella and A. Bestavros, "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 835–846, Dec. 1997.
3. A. Erramilli, O. Narayan, and W. Willinger, "Experimental Queueing Analysis with Long-Range Dependent Packet Traffic," *IEEE/ACM Transactions on Networking*, vol. 4, no. 2, pp. 209–223, Apr. 1996.
4. K. Park, G. T. Kim, and M. E. Crovella, "On the Relationship between File Sizes, Transport Protocols, and Self-Similar Network Traffic," in *Proceedings of the Fourth International Conference on Network Protocols (ICNP'96)*, Columbus, Ohio, USA, Oct. 1996, pp. 171–180.
5. M. Crovella and L. Lipsky, "Simulations with Heavy-Tailed Workloads," in *Self-Similar Network Traffic and Performance Evaluation*, K.Park and W. Willinger, Eds., chapter 3, pp. 89–100. Wiley-Interscience, NY, 2000.
6. U. Fiedler, P. Huang, and B.Plattner, "Towards Provisioning Diffserv Intra-Nets," in *Proceedings of IWQoS'01*, Karlsruhe, Germany, June 2001, pp. 27–43, Springer.

7. A. M. Odlyzko, "The Internet and other Networks: Utilization Rates and their Implications," *Information Economics and Policy*, vol. 12, pp. 341–365, 2000.
8. S. Ben Fredj et. al., "Statistical Bandwidth Sharing: A Study of Congestion at Flow Level," in *Proceedings of SIGCOMM'01*, San Diego, California, USA, Aug. 2001, pp. 111–122, ACM.
9. P. Barford and M. Crovella, "Generating Representative Web Workloads for Network and Server Performance Evaluation," in *Proceedings of Performance '98/ACM SIGMETRICS '98*, Madison, Wisconsin, USA, June 1998, pp. 151–160.
10. A. Feldmann et. al., "Dynamics of IP traffic: A Study of the Role of Variability and the Impact of Control," in *Proceedings of SIGCOMM'99*, Cambridge, Massachusetts, USA, Sept. 1999, ACM.
11. Walter Willinger, Murad S. Taqqu, Robert Sherman, and Daniel V. Wilson, "Self-Similarity through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level," *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 71–86, 1997.
12. C. Goldie and C. Kluppelberg, "Subexponential Distributions," in *A Practical Guide to Heavy Tails: Statistical Techniques for Analysing Heavy Tails*, R. Feldman R. Adler and M.S. Taqqu, Eds., pp. 435–460. Birkhauser, Basel (CH), 1997.
13. Norman L. Johnson, Samuel Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, vol. 1 of *Wiley Series in Probability and Mathematical Statistics*, Wiley, NY, 2 edition, 1994.
14. J. Rice, *Mathematical Statistics and Data Analysis, 2nd edition*, Duxbury Press, 1995.
15. C. Radhakrishna Rao, *Linear Statistical Inference and Its Applications*, Wiley, New York, 2 edition, 1973.
16. Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*, Wiley, NY, 1986.
17. Jan Beran, *Statistics for Long-Memory Processes*, Chapman & Hall, NY, 1994.
18. L. Breslau et. al., "Advances in Network Simulations," *IEEE Computer*, May 2000.
19. R. Fielding et. al., "Hypertext Transfer Protocol — HTTP/1.1," RFC 2616, Internet Request For Comments, June 1999.
20. S. Bajaj et. al., "Is Service Priority Useful in Networks?," in *Proceedings of the ACM Sigmetrics '98*, Madison, Wisconsin USA, June 1998.