ON THE SIMULATION OF QUEUES WITH PARETO SERVICE

PABLO JESÚS ARGIBAY-LOSADA
ANDRÉS SUÁREZ-GONZÁLEZ
CÁNDIDO LÓPEZ-GARCÍA
RAÚL FERNANDO RODRÍGUEZ-RUBIO
JOSÉ CARLOS LÓPEZ-ARDAO
DIEGO TEIJEIRO-RUIZ
ETSE de Telecomunicación, Universidade de Vigo, 36200 Vigo, Spain

Abstract: In M/G/n queues —with G a heavy-tailed distribution— the tail of G has low probability but a dramatic impact on the performance of the system. The analytical treatment of M/G/n queues is difficult, so many times we must use simulation to study them. But the simulation of systems using heavy-tailed distributions presents difficulties. We need efficient simulation methods to study those systems, and we can use M/G/1 systems as workbenches since they have some analytical results to check the simulation results with. In this paper we try to gain some insight into the nature of those difficulties, and propose, develop and analyze a method to speed up simulations of M/G/1 systems when G is heavy-tailed.

keywords: heavy tails, queue systems, steady state.

1. INTRODUCTION

M/G/n queues —where G is a heavy-tailed service time distribution— are used to model queue systems where a range of values of the service time, whose probability is very low, have a drastic impact on the overall performance of the system. The Pareto distribution is one of these heavy-tailed distributions and it has been proposed as the page size distribution in Web servers or as the file size distribution in FTP servers. The accurate analytical treatment of M/G/n systems is very difficult and in many cases it cannot be applied. Simulation is a possible method to study them. But simulations with heavy-tailed random variables present some additional difficulties, and care must be taken when extracting conclusions from the results of these simulations. It is necessary to have accurate and efficient simulation methods. Efficient because we need to generate big quantities of data for our simulation study to be accurate enough. And their accuracy can be checked by means of comparisons with known results from simpler systems with analytical solution. One of these simpler queue systems that can be studied analytically is the M/P/1 queue. M/P/1 systems can be used then as a workbench for more efficient simulation methods, able to deal with the heavy-tail problematic. The slow convergence of the simulations to the steady state may be an important problem of the simulation of M/P/1 queues.

Recent studies have shown the problems involved in simulating M/P/1 queues. The reason of these problems is the heavy-tailed condition of the Pareto: very high values of the demanded service time appear with very low —but not negligible— probabilities, in such a way that their effect in the waiting time distribution is drastic. The heavy-tailed condition decisively contributes to rise the mean queue waiting time. But problems relating to practical aspects of computer simulation like finite machine resolution and finite and low simulation time make the simulations underestimate the parameters of interest, typically the mean queue waiting time. Gross [Gross et al, 2002] studies the impact of finite resolution random number generation on the mean queue waiting time estimation. It is interesting to know how many of these problems can be avoided with better simulation techniques and computer resources, and how the power-tailed condition effectively limits our efforts to speed up the simulations.

In this paper we investigate this problem, try to get insight in the impact of the transient period in the mean value, and propose a method to try to start the simulation near the steady state. We compare the proposed method with the traditional start from empty system.

2. HEAVY-TAILED DISTRIBUTIONS

A random variable (RV) X, with cumulative distribution function (cdf) $\mathrm{F}(x)$, is said to be heavy-tailed if its complementary distribution function, $1-\mathrm{F}(x)$, has

Proceedings 17th European Simulation Multiconference
(c) SCS Europe BVBA, 2003 ISBN 3-936150-25-7

an hyperbolic decaying tail:

$$\exists \alpha > 0 \left| \lim_{k \to \infty} \frac{1 - \mathrm{F}\,(x)}{x^{-\alpha}} = \mathsf{c} \in (0, \infty) \right.$$

The Pareto cdf, clearly heavy-tailed, is given by $\mathrm{F}\,(x) = 1 - (\mathsf{m}/x)^{\alpha} \quad \forall x \geq \mathsf{m} > 0$, where m is called the scale parameter, and α is called the shape parameter. In [Gross et al, 2002] a Pareto distribution with $\mathsf{m} = 1$ is used in a M/P/1 queue to show the problems that appear when simulating such system when α is near 2. In this paper we also fix m to 1 to demonstrate the benefits of our method in the same scenario. The Pareto probability density function (pdf) is given by $\mathrm{f}\,(x) = \alpha \cdot \mathsf{m}^{\alpha}/x^{\alpha+1} \quad x > \mathsf{m} > 0$. The Pareto k^{th} order moment exists if and only if $\alpha > k$. Its mean value exists if and only if $\alpha > 1$ and is given by $\overline{X} = \alpha \cdot \mathsf{m}/(\alpha - 1)$. Its second order moment exists if and only if $\alpha > 2$ and is given by $\overline{X^2} = \alpha \cdot \mathsf{m}^2/(\alpha - 2)$

3. PARETO TAIL PROBLEMS

Recent research has shown that the estimation using computer simulation of the mean queue waiting time of a M/P/1 queue, $\overline{W}$, converges very slowly to its theoretical value when α approximates 2 [Gross et al, 2002]: simulation run-lengths as long as some million observations do not give estimations of $\overline{W}$ close to the exact theoretical value in these cases.

The **Pollaczek-Khinchin** formula states that the mean queue waiting time is directly proportional to the second order moment of the service time in a M/G/1 queue:

$$\overline{W} = \frac{\lambda \cdot \overline{S^2}}{2 \cdot (1 - \rho)}$$

where λ is the average arrival rate of customers, S the demanded service time random variable and ρ the utilization factor of the queue system —$\rho = \lambda \cdot \overline{S}$—.

If we have a limited resolution random number generator we will not be able to generate the extremely large values of S that appear occasionally in the actual system, so the measured $\overline{S^2}$ will tend to be low, and this will probably make the estimation of $\overline{W}$ low. Even if we have an infinite resolution random number generator, we can give a rough estimation of how many observations of customer queue waiting times we need before getting close to the real mean value. If we have a random number generator with finite resolution which is only able to produce numbers between 0 and K, we will loose in the simulation service times greater than K. But the appearance in our simulation of values beyond a certain number is not only a matter of resolution of the random number generator, but relates to the intrinsic probability of that value, or range of values.

If we have a range of values whose probability is p, the mean number of trials in order to get one value in that range is $\frac{1}{\mathsf{p}}$. The weight of the tail of a Pareto beyond a certain limit K, i.e. the probability of getting one value in the range (K, ∞), is given by $\mathsf{K}^{-\alpha}$, so the probability of getting all the values smaller than K in r trials is $(1 - \mathsf{K}^{-\alpha})^r$.

In the Pareto case, when α is near 2, the tail has a great influence in the value of its second order moment. For example, we select the utilization factor of the system $\rho = 0.5$. We choose a shape parameter $\alpha = 2.1$, so $\overline{S} = 1.909$ and $\overline{S^2} = 21$. If we generate a sample of 1 million observations, the probability of getting all the values smaller than K, i.e., the probability of having a sample indistinguishable of that from a truncated Pareto with truncation parameter K is $\mathsf{P} = (1 - \mathsf{K}^{-\alpha})^{10^6} \simeq \mathrm{e}^{-\frac{10^6}{\mathsf{K}^{\alpha}}}$. Considering the service time RV, S, whose pdf is a Pareto, and a service time RV S_{t}, whose pdf is a truncated Pareto from the former, with truncation value K, we have $\mathrm{f}_{S_{\mathrm{t}}}(x) = \frac{\mathrm{f}_S(x)}{1 - \Pr(x < \mathsf{K})} \quad x < \mathsf{K}$ with $\overline{S_{\mathrm{t}}} = \frac{\mathsf{K}^{\alpha} - \mathsf{K}}{\mathsf{K}^{\alpha} - 1} \cdot \overline{S}$ and $\overline{S_{\mathrm{t}}^2} = \frac{\mathsf{K}^{\alpha} - \mathsf{K}^2}{\mathsf{K}^{\alpha} - 1} \cdot \overline{S^2}$. If we now impose a probability of 99 percent about having obtained a truncated Pareto, the correspondent K is 6434. and the mean value of the associated truncated Pareto, $\overline{S_{\mathrm{t}}}$, is $0.999934 \cdot \overline{S}$. So intuitively the probability of getting a mean value of 0.999934 times the theoretical value —this ratio will represent the accuracy in the estimation— is 99 percent. This may be considered negligible —we are correctly estimating the mean value of the Pareto RV. But the second order moment of the truncated Pareto is $S_{\mathrm{t}} = 0.58 \cdot \overline{S^2}$, what means that with a probability of 99 percent we are underestimating the theoretical value of the second order moment, with the estimation being 0.58 times the theoretical value.

So we see that with a probability of 99 percent we will also underestimate $\overline{W}$ in a factor of at least 0.58, nearly half the theoretical, and the cause is we are generating too few service times to be able to reach the steady-state.

This means that the *a priori* high value of the run size of the simulation is in practice a very low one when the service time distribution is heavy-tailed. To have more accurate results we need a much larger number of samples. For example, if we impose an accuracy of 99 percent in the second order moment estimation — equivalent to the accuracy in the mean queue waiting time—, we obtain $\mathsf{K} = 10^{20}$, so with a probability of 99 percent we will need no less than 10^{40} samples. If we want an accuracy of 90 percent, with a confidence of 99 percent we will need no less than 10^{19} samples. Fig. 1 details this. In it we plot the tolerance —one minus the accuracy— versus the number of samples.

These examples show that although the M/P/1 process, with α near 2, is ergodic in theory, the run sizes of the simulations needed to check that ergodicity will

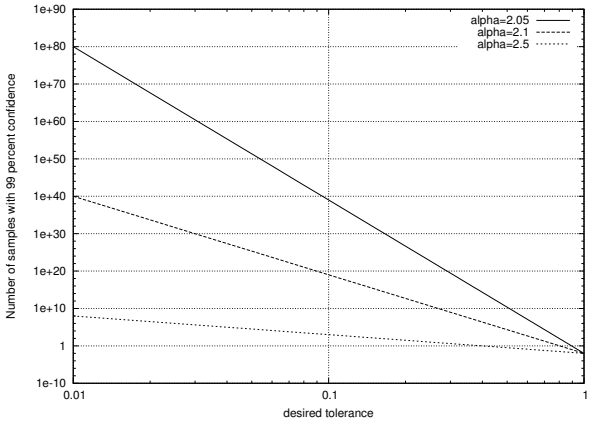

Fig. 1. Number of samples required for a given tolerance in the mean queue waiting time

probably be too high to consider the system ergodic in practice.

So we can see that estimating $\overline{W}$ when the service time is heavy-tailed and the shape parameter α is slightly greater than 2 —large variance— will probably be computationally very expensive if we start our simulations from an empty system. Thus, traditional simulation methods based on computating the samples of the involved RVs —the interarrival times of the customers and their service times—, will be too expensive due to the large amount of samples that must be generated before obtaining a representative set of samples of the involved processes.

4. CHOSEN INTERVAL LENGTH

We have developed a framework simulation model to achieve a greater accuracy in the simulations of M/P/1 queues with α slightly greater than 2. Its main idea is to try to initialize the simulation almost in steady-state. We can take advantage of our knowledge of the arrival process of the M/G/1 queue. When a user arrives at a M/G/1 system, it will possibly find some users in the queue and one in the resource. The queue waiting time of the arriving customer will be the residual life of the user in the resource, S_{r}, —i.e., the remaining time that user will stay in the system— plus the service times of all the customers in the queue before our customer arrived. The distribution of the residual life of the customer in the resource will depend on the distribution of the service time, as it happens to the whole service time demanded by this user, L. Its pdf is given by [Kleinrock, 1975],

$$\mathrm{f}_L(x) = \frac{x \cdot \mathrm{f}_S(x)}{\overline{S}}$$

where $\mathrm{f}_S(x)$ is the pdf of the demanded service time. From its definition, we can note that its mean, $\overline{L}$, will be $\overline{S^2}/\overline{S} = \overline{S} \cdot (1 + \mathrm{C}^2)$.

So if our customer arrives to the system while there is somebody in the resource, it will arrive randomly in an interval described by $\mathrm{f}_L(x)$, and, in average, it will have to wait $\overline{L}/2$ for the client in the resource to finish, plus some amount of time due to the users in the queue. If we denote M the number of clients in queue when the user in the resource entered it, and N the number of clients who arrived between the user in the resource began service and our user arrival, we can say that the queue waiting time for a user that has to wait is:

$$W = S_r + \left(\sum_{i=1}^{M} S_i + \sum_{j=1}^{N} S_j \right) \tag{1}$$

where the term between brackets represents the waiting time due to the customers in queue when our client arrived. So we can express the $\overline{W}$ in the system as

$$\overline{W} = \rho \left(\frac{\overline{L}}{2} + \overline{M} \cdot \overline{S} + \lambda \cdot \frac{\overline{L}}{2} \cdot \overline{S} \right) \tag{2}$$

where we have used the fact that the waiting time will be non-null with probability ρ, and that the number of arrivals between the service start of the user in the resource and our client arrival is a RV with mean $\lambda \cdot \overline{L}/2$. To see what is the distribution of M, we can consider the queue of our M/P/1 system as another M/G/1 system. Since the departing customers from a M/G/1 system see the same distribution of the number of users in the system as the one seen by a random observer, the departures from the queue —to enter the resource— see samples of the Q RV.

Finally, we can write Equation (2) as

$$\overline{W} = \rho \cdot \left(\frac{\overline{L}}{2} + \overline{Q} \cdot \overline{S} + \lambda \cdot \frac{\overline{L}}{2} \cdot \overline{S} \right)$$

To achieve a good estimation of $\overline{W}$, we can simulate then a system where a user finds another customer being served, and whose service time follows the distribution $\mathrm{f}_L(x)$, obtainable from $\mathrm{f}_S(x)$. The number of users who arrive between the time the user in the resource began being served and our client arrival will be a Poisson RV whose mean will be known. The queue length when the selected interval began is one sample of the queue length distribution. If we could calculate a good estimation of the queue length Q, the sample of W when $W > 0$ would be obtained from Eq. (1).

We can approximate the theoretical convergence ratio of the classical simulation method (that starting from an empty system and simulating the system along the continuous time axis) and our proposed method using the considerations on probabilities of appearance of high-value samples we used in Section 3:

4.1. **Classical Method**

Consider the service time RV, S, whose pdf is a Pareto, and a service time RV S_{t}, whose pdf is a truncated Pareto from the former, with truncation value K

$$\mathrm{f}_{S_{\mathrm{t}}}(x) = \frac{\mathrm{f}_S(x)}{1 - \Pr(x < \mathsf{K})} \qquad x <= \mathsf{K}$$

Its first and second moments are

$$\overline{S_{\mathrm{t}}} = \frac{\mathsf{K}^\alpha - \mathsf{K}}{\mathsf{K}^\alpha - 1} \cdot \overline{S} \tag{3}$$

$$\overline{S_{\mathrm{t}}^2} = \frac{\mathsf{K}^\alpha - \mathsf{K}^2}{\mathsf{K}^\alpha - 1} \cdot \overline{S^2} \tag{4}$$

and we see that $\overline{S_{\mathrm{t}}} < \overline{S}$ and $\overline{S_{\mathrm{t}}^2} < \overline{S^2}$.

The probability of getting one sample value of a Pareto less than K is $1 - \mathsf{K}^{-\alpha}$. If we generate a sample of size $\mathrm{N_t}$ of a Pareto distribution, the probability that this sample is indistinguishable of one of a truncated Pareto with truncation value K, i.e., the probability of all those samples are less than K is $(1 - \mathsf{K}^{-\alpha})^{\mathrm{N_t}}$, so with this probability we are getting a sample indistinguishable of one from a truncated Pareto whose truncation value will be K or less, so in this case an upper bound for the second moment is given by Eq.(4).

So if we want to calculate with a given confidence P an upper bound for the second moment with $\mathrm{N_t}$ samples of our untruncated Pareto process, we do the following:

$$\mathrm{P} = \left(1 - \frac{1}{\mathsf{K}^\alpha}\right)^{\mathrm{N_t}} \simeq \mathrm{e}^{-\frac{\mathrm{N_t}}{\mathsf{K}^\alpha}} \Rightarrow \mathsf{K} \simeq \left(\frac{\mathrm{N_t}}{-\ln \mathrm{P}}\right)^{\frac{1}{\alpha}}$$

So the estimated second moment with a confidence of P over $\mathrm{N_t}$ samples, $\widehat{S^2}[\mathrm{N_t}, \mathrm{P}]$, is

$$\widehat{S^2}[\mathrm{N_t}, \mathrm{P}] = \frac{\mathsf{K}^\alpha - \mathsf{K}^2}{\mathsf{K}^\alpha - 1} \cdot \overline{S^2} \simeq \frac{\frac{\mathrm{N_t}}{-\ln \mathrm{P}} - \left(\frac{\mathrm{N_t}}{-\ln \mathrm{P}}\right)^{\frac{2}{\alpha}}}{\frac{\mathrm{N_t}}{-\ln \mathrm{P}} - 1} \cdot \overline{S^2}$$

If we denote $\widehat{W}[\mathrm{N_t}, \mathrm{P}]$ the estimated $\overline{W}$ with probability P over $\mathrm{N_t}$ samples, and define the accuracy in the estimation of $\overline{W}$ as

$$A_1[\mathrm{N_t}, \mathrm{P}] = \frac{\widehat{W}[\mathrm{N_t}, \mathrm{P}]}{\overline{W}}$$

it results that the estimated acccuracy in $\overline{W}$ with a confidence of P over $\mathrm{N_t}$ samples, $A_1[\mathrm{N_t}, \mathrm{P}]$ is

$$A_1[\mathrm{N_t}, \mathrm{P}] = \frac{\widehat{W}[\mathrm{N_t}, \mathrm{P}]}{\overline{W}} = \frac{\widehat{S^2}[\mathrm{N_t}, \mathrm{P}]}{\overline{S^2}} = \frac{\frac{\mathrm{N_t}}{-\ln \mathrm{P}} - \left(\frac{\mathrm{N_t}}{-\ln \mathrm{P}}\right)^{\frac{2}{\alpha}}}{\frac{\mathrm{N_t}}{-\ln \mathrm{P}} - 1} \tag{5}$$

4.2. **Proposed Method**

We use the relationship

$$\overline{W} = \left(\frac{\overline{L}}{2} + \lambda \frac{\overline{L}}{2} \overline{S} + \overline{Q} \cdot \overline{S}\right) \cdot \rho \tag{6}$$

where L is the distribution of the chosen interval length. If S is a Pareto with shape parameter α, is easy to see that L will be a Pareto with shape parameter $\alpha_2 = \alpha - 1$. To calculate the estimation of $\overline{L}$ as function of the number of samples, n, we use the same method as above.

Considering the Pareto RV L, the estimation of the mean of the correspondent truncated Pareto, L_{t} with shape parameter $\alpha_2 = \alpha - 1$ and with truncation value K is given by Eq (3)

$$\overline{L_{\mathrm{t}}} = \frac{\mathsf{K}^{\alpha_2} - \mathsf{K}}{\mathsf{K}^{\alpha_2} - 1} \cdot \overline{L} = \frac{\mathsf{K}^{\alpha-1} - \mathsf{K}}{\mathsf{K}^{\alpha-1} - 1} \cdot \overline{L}$$

In Equation (6) there is a term that represents the average value of Q. We will estimate $\overline{Q}$ producing an initial number of busy periods, randomly choosing one point in time and calculating the Q when the selected customer in service entered the resource. So if we generate n samples of this initial simulation, we think we can reasonably suppose that our estimation of $\overline{Q}$ will tend to $\overline{Q}$ with the same speed like the one we estimate $\overline{L}$ with. That supposition has been backed with simulation results shown in Fig. 3, where we plot the empirical pdf of the estimated $\overline{W}$ with 100 simulation runs of the classical method and the proposed method, and it can be seen that the obtained mean values are close to those predicted by the analytic expression in Eq. (7). Using this supposition, and if we denote A_2 the accuracy in the estimation of $\overline{L}$, $\widehat{L}/\overline{L}$, we have

$$\overline{W} = \left(\frac{\overline{L}}{2} + \lambda \cdot \overline{S} \cdot \frac{\overline{L}}{2} + \overline{Q} \cdot \overline{S}\right) \cdot \rho$$

$$= \left(\frac{\widehat{L}}{2 \cdot A_2} + \lambda \cdot \overline{S} \cdot \frac{\widehat{L}}{2 \cdot A_2} + \frac{\widehat{Q} \cdot \overline{S}}{A_2}\right) \cdot \rho = \frac{\widehat{W}}{A_2}$$

$$A_2 = \frac{\widehat{W}}{\overline{W}} = \frac{\widehat{L}}{\overline{L}} = \frac{\mathsf{K}^{\alpha-1} - \mathsf{K}}{\mathsf{K}^{\alpha-1} - 1}$$

The confidence for $\mathrm{N_t}$ samples being from a truncated Pareto with shape parameter $\alpha - 1$ and truncation point K or less is

$$\mathrm{P} = \left(1 - \frac{1}{\mathsf{K}^{\alpha-1}}\right)^{\mathrm{N_t}} \simeq \mathrm{e}^{-\frac{\mathrm{N_t}}{\mathsf{K}^{\alpha-1}}} \Rightarrow \mathsf{K} = \left(\frac{\mathrm{N_t}}{-\ln \mathrm{P}}\right)^{\frac{1}{\alpha-1}}$$

So an upper bound for the accuracy of the estimated $\overline{W}$ will be, for a confidence P and $\mathrm{N_t}$ samples,

$$A_2[\mathrm{N_t},\mathrm{P}] = \frac{\frac{\mathrm{N_t}}{-\ln \mathrm{P}} - \left(\frac{\mathrm{N_t}}{-\ln \mathrm{P}}\right)^{\frac{1}{\alpha-1}}}{\frac{\mathrm{N_t}}{-\ln \mathrm{P}} - 1} \qquad (7)$$

Fig. 2 compares the theoretical results for the upper bounds of the convergence rates of both methods, the classical one, given by Eq. (5), and the proposed one, given by Eq. (7), for a probability of 99 percent. We see that our method does not underestimate the real mean value of the queue waiting time as much as the traditional method. There is still a big difference between both estimations and the real value due to the fact mentioned in section 2: the probabilities involved for high values of the service times are too small for those values to appear in short simulations; but the improvement in the estimation is appreciable. This method can serve as basis for more improvements using known facts from the underlying processes, and we are working in the improvement of the simulation algorithm.

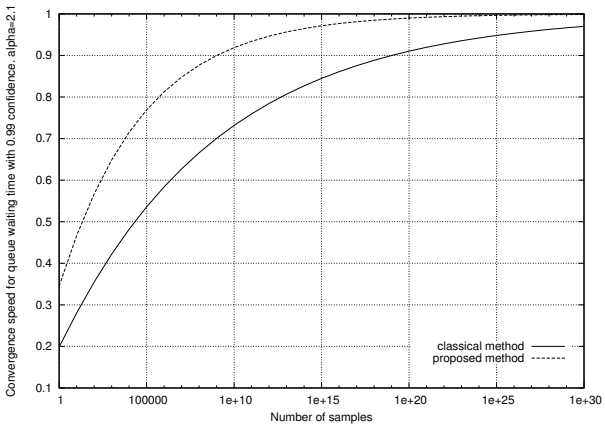

Fig. 2. Comparison between the 99 percent confidence upper bounds for the convergence rates of the classical and proposed method.

5. IMPLEMENTATION

To obtain samples of W with our method, we generate the value 0 with probability $1-\rho$, and with probability ρ a sample of the service time length found by a typical customer that has to wait. This is a Pareto with shape parameter $\alpha_2 = \alpha - 1$, where α is the shape parameter of the Pareto representing the service time. Next, we choose a random point in the generated interval which will represent the arrival instant of a typical client. The queue length in this moment will be the clients in queue when the selected interval began, plus the number of clients who arrived between the beginning of the interval and the arrival of our client. This last number is a Poisson RV with mean $\lambda \cdot U$, with $U = L - S_r$ the elapsed time since the beginning of the interval and our client arrival. We can directly generate samples of this RV. But the number of users in queue when the interval began follows the distribution of Q, which is unknown, so we will have to estimate it using a classical simulation. The waiting time of our client will be, then, the residual life of the interval plus the service times of the users in queue when it arrives.

6. PERFORMANCE

To evaluate the performance of simulations using this method, we note that it uses more samples of the random variables involved than the classical method. The classical method needs one interarrival time and one service time to produce one waiting time sample. Our method needs to generate one classical simulation to obtain estimates of Q, the queue length. To obtain one estimate of the waiting time, we need to estimate one queue length sample, Q_i, with a classical simulation; we need to generate one sample of a Pareto with shape parameter $\alpha - 1$; one Poisson to estimate how many customers arrive between the beginning of the chosen interval and our arrival, N, and $N + Q_i$ service times. Moreover, taking into account the fact that our estimates of Q will not be independent, because we are obtaining them from samples in one finite simulation, to reduce that dependence we can think of choosing a small proportion of estimates of Q from the total number of samples of our classical simulation. This makes the mean number of random values to generate one sample of the waiting time in our method bigger than that of the classical method. But that difference is not important enough to make the proposed method worst in performance than the classical.

If we have a M/P/1 with shape parameter α, the classical simulation will need 1 Poisson RV and 1 Pareto RV to obtain one sample of the waiting time. In our method, we need one sample of the queue length from a classical simulation. To reduce dependence between samples of it, we choose them sampling the classical simulation with a Poisson process with mean λ/n, with $n > 1$. We will need n samples of Q in the classical simulation to select one of them, Q_i, for computation, and that means n Poisson RVs and n Pareto RVs. We generate one more Pareto for the length of the selected interval, one Poisson for the number of arrivals in that interval prior to ours, N, and $Q_i + N$ Pareto RVs for the service times of all the arrivals. If we have a Pareto service time with $\alpha = 2.1$, and $\rho = 0.5$, using the **Pollaczek-Khinchin** formula we have $Q = 1.44$, and the average length of the chosen interval is 11. If we select in average one of every four samples of Q for computation, in the worst case, that in which we do not underestimate the theoretical $\overline{Q}$ —because in that case there are more service times to generate— we will need in average 4 Pareto + 4 Poisson + 1 Pareto + 1 Poisson + 1.44 Pareto + 1.44 Pareto = 7.88 Pareto RVs + 5 Poisson RVs. This implies that we need approximately 6.5 times more samples to obtain one sample of W than in the classical method. One fact that favours the

efficiency of our method is that it always produces samples of W with $W > 0$. The classical method generates interarrival and service times to produce the value $W = 0$ with probability $1 - \rho$. This is, we are wasting computer resources to generate one known value whose probability is known *a priori*. Our method only produces samples of W when $W > 0$, giving a mean value $W_{W>0}$. The final mean value of W will be $\rho \cdot W_{W>0}$. The lower is ρ, the better is our method in terms of efficiency compared with the classical method. So in the previous example, given that $\rho = 0.5$, we can consider our method to use $6.5/2 = 3.25$ times more samples than the classical one.

If we generate some simulations of the two methods, and represent the pdf of the estimated $\overline{W} = 5.5$, we obtain Figure 3, for which we run 100 simulations of 1 million samples of $\overline{W}$ in every method. It is clear that the proposed method has better accuracy. If we take into account that our method uses more samples and represent the pdf of the two methods but this time the classical method uses four times more samples to compensate the more samples used by our method, we obtain Figure 4, which uses 100 runs of 1 million values of the waiting time with the proposed method and 100 runs of 4 million values in the classical method. The difference between the accuracies in both methods is lower than that in Figure 3, but it is still appreciable that the proposed method works better, now with similar performance.

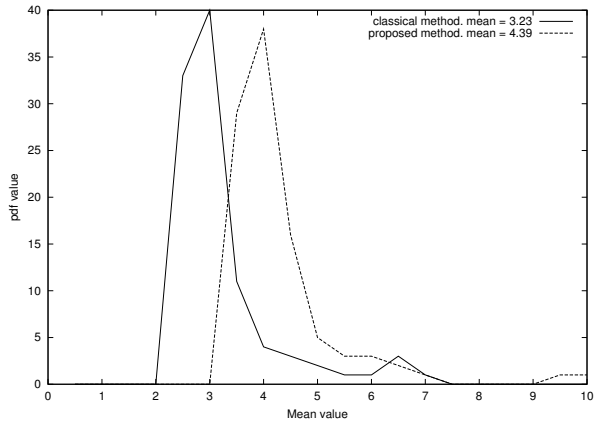

Fig. 3. Empirical pdf of $\overline{W}$ of the classical and the proposed methods with 100 runs of 1 million waiting times each.

7. CONCLUSION

The computer simulation of M/P/1 queues presents important difficulties due to the slow decaying tail of the Pareto distribution. This makes extremely high values, with great influence on the statistical figures of the system, appear with so low probabilities that if we want to simulate the physical underlying processes, generating demanded times and time arrivals, the cost in time will probably be prohibitive if we want accurate results. This forces to use all our knowledge of the statistics of the system inner processes, so the simulation can noticeably speed up.

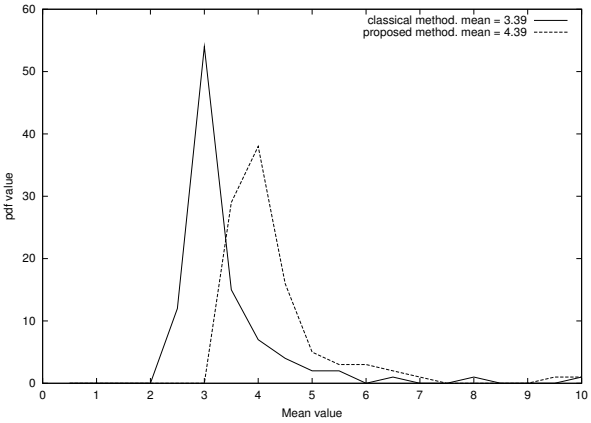

Fig. 4. Empirical pdf of $\overline{W}$ with 100 runs of 10^6 waiting times in the proposed method and $4 \cdot 10^6$ in the classical one.

REFERENCES

Gross, D., Shortle, J.F., Fischer, M.J. and Masi, D.M.B. 2002. "Difficulties in simulating queues with pareto service". In *Proceedings of the 2002 Winter Simulation Conference*, 2002.

Kleinrock, L. 1975. "Queueing systems". *Wiley & Sons*.

Takács, L. 1962. "Single server queue with poisson input simulation." In *Operations Research 10:388–397.*

Sigman, K. 1999. "A primer on heavy-tailed distributions." In *Queueing Systems. 33: 261-275.*

AUTHOR BIOGRAPHIES

PABLO JESÚS ARGIBAY-LOSADA is an assistant professor in the *Departamento de Enxeñería Telemática* at *Universidade de Vigo*. He received a telecommunication engineering degree from *Universidade de Vigo* in 2001. Nowadays, he is working toward his Ph.D. in the *Departamento de Enxeñería Telemática* at *Universidade de Vigo*. His e-mail address is <Pablo.Argibay@det.uvigo.es>.

ANDRÉS SUÁREZ-GONZÁLEZ is an associate professor in the *Departamento de Enxeñería Telemática* at *Universidade de Vigo*. He received a Ph.D. degree in telecommunication engineering from *Universidade de Vigo* in 2000. He is a member of ACM. His current research interests include simulation methodology and analysis of stochastic systems. His e-mail address is <asuarez@det.uvigo.es>.