Proceedings of the 2002 Winter Simulation Conference
E. Yücesan, C. H. Chen, J. L. Snowdon, and J. M. Charnes, eds.

DIFFICULTIES IN SIMULATING QUEUES WITH PARETO SERVICE

Donald Gross
John F. Shortle

Dept. of Systems Engineering & Operations Research
George Mason University
4400 University Drive, MS 4A6
Fairfax, VA 22030, U.S.A.

Martin J. Fischer
Denise M. B. Masi

Mitretek Systems
3150 Fairview Park Drive South
Falls Church, VA 22042, U.S.A.

ABSTRACT

M/G/1 queues, where G is a heavy-tailed distribution, have applications in Internet modeling and modeling for insurance claim risk. The Pareto distribution is a special heavy-tailed distribution called a power-tailed distribution, and has been found to serve as adequate models for many of these situations. However, to get the waiting time distribution, one must resort to numerical methods, e.g., simulation. Many difficulties arise in simulating queues with Pareto service and we investigate why this may be so. Even if we are willing to consider truncated Pareto service, there still can be problems in simulating if the truncation point (maximum service time possible) is too large.

1 INTRODUCTION

Queueing theory has long been employed to study congestion problems in a myriad of application areas. In some applications of queueing theory, the usual assumptions that made queueing analyses so productive (e.g., Poisson arrivals and exponential-type holding times) clearly do not hold. Two cases in point are in modeling traffic on the Internet and modeling financial claims on insurers. In these cases, very low probability of extremely high service values can arise (e.g., insurance claims as a result of the 9-11 terrorist attacks). Fowler (1999) details heavy-tailed distributions occurring in Internet traffic, at five of the seven OSI protocol levels: FTP transfers (application level) and session durations/size (session level) are among these. Data indicate that the Pareto distribution well describe these service times. Further, heavy-tailed distributions also play a significant role in portfolio and insurance models, where claim sizes can take on extremely large values (it can be shown that the probability of eventual ruin is the same as the stationary tail waiting probability for an M/G/1 queue, where the service times are Pareto random variables (Juneja et al. 1999)).

2 HEAVY-TAILED DISTRIBUTIONS AND THE PARETO

A cumulative distribution function, *F*(*x*), has a *power tail* if there exists positive constants *c* and *a* such that for $\overline{F}(x) = 1 - F(x)$

$$\lim_{x\to\infty} \frac{\overline{F}(x)}{x^{\alpha}} = c\,.$$

That is, the tail decays geometrically in the limit (as opposed to the more familiar exponential decay of the exponential, and gamma, for example). Power-tail distributions are a subset of a broader class of distributions whose tails decay more slowly than exponential, i.e.,

$$\lim_{x\to\infty} \left[e^{ax}\,\overline{F}(x)\right] = \infty\ .$$

This broader class is referred to as heavy-, fat- or long-tailed distributions, and include the lognormal and the Weibull (with shape parameter <1). These latter distributions have tails that decay more slowly than any exponential, but not as slowly as the Pareto, so that a power-tailed distribution is also a heavy-tailed distribution, but not necessarily the reverse.

The one-parameter (shape) version of the Pareto CDF is given by

$$F(x) = 1 - \frac{1}{(1+x)^{\alpha}} \qquad (x \ge 0)\,, \tag{1}$$

where α is the shape parameter. The corresponding density function is

$$f(x) = \frac{\alpha}{(1+x)^{\alpha+1}} \qquad (x \ge 0),$$

and it is straightforward to show that the Pareto is indeed a power-tailed distribution.

A major consequence of power-tailed behavior is the disappearance of moments. It is easy to see that for a Pareto to have its k^{th} moment, $E[X^k]$, we need $\alpha > k$. If $\alpha > 1$, then the mean, $E[X]$, is

$$E[X] = 1/(\alpha - 1)$$

and if $\alpha > 2$, it follows that

$$E[X^2] = \frac{2}{(\alpha-2)(\alpha-1)}.$$

Thus, no matter what the value of the parameter α, a Pareto random variable cannot have all its moments and hence does not have an analytic Laplace transform, which renders standard queueing analysis of M/G/1 impossible. Numerical methods which approximate the required Laplace transform (e.g., the Transform Approximation Method – TAM, Shortle et al. 2002) work well in many cases. However, TAM cannot be used for queueing networks; hence the requirement for simulation.

3 SIMULATION PERFORMANCE IN ESTIMATING MEAN QUEUE WAIT FOR M/P/1

To investigate how accurately we can simulate queues with Pareto service (the M/P/1 queue), we consider α values greater than 2. This allows comparison with theoretical results obtained from the Pollaczek-Khintchine (PK) formula (see Gross and Harris 1998, p. 212) as both mean and variance exist for the Pareto in this region and thus allows us to obtain results from the PK formula for the mean wait in queue, Wq. Figure 1 (all figures and tables appear at the end of the text) shows runs from ARENA simulations of varying run lengths for five Pareto service α values from 2.020202 to 3.5, yielding a range of coefficients of variation (CVs) from 1.53 to 10. We see that the closer α is to 2, the worse the percent error from the theoretical PK value for mean wait, Wq. For α greater than three, the simulation appears quite accurate, even for fairly small run lengths, but for α values in the low 2s, even run lengths as long as 20,000,000 transactions still produce sizable error. Crovella and Lipsky (1997) observed difficulties in estimating the mean of a heavy-tailed distribution with shape parameter α less than 1.7. However, in simulating Wq for an M/P/1, the variance is also needed; Sees (2001) showed that problems of simulating Wq for M/P/1 queues arise when α is less than 2.7.

Since for any finite simulation run length, there is always a maximum value of the random variables generated, we, in actuality, are simulating a truncated Pareto service distribution. It has also been argued that there is always a maximum file size or claim amount so, in reality, we are always dealing with truncated distributions. Therefore, we next turn our attention to the truncated Pareto distribution and consider M/PT/1 queues, where service times are truncated Pareto to gain insight as to what the problem may be in the poor results in simulating M/P/1 queues with α values near 2, and whether simulation does a better job for M/PT/1 than for M/P/1.

4 THE TRUNCATED PARETO DISTRIBUTION

Considering the untruncated Pareto CDF of equation (1), we see that $P\{X<T\} = F(T)=1-1/(1+T)^{\alpha}$, so that the truncated CDF becomes:

$$F_T(x) = \frac{1 - \dfrac{1}{(1+x)^{\alpha}}}{1 - \dfrac{1}{(1+T)^{\alpha}}}.$$

The first two moments of the truncated Pareto are:

$$ES = \frac{\alpha}{F(T)}\left[\frac{1}{\alpha(1+T)^{\alpha}} - \frac{1}{(\alpha-1)(1+T)^{(\alpha-1)}} + \frac{1}{\alpha(\alpha-1)}\right]$$

and

$$ES^2 = \frac{\alpha}{F(T)}\left[-\frac{1}{(\alpha-2)(1+T)^{(\alpha-2)}} + \frac{2}{(\alpha-1)(1+T)^{(\alpha-1)}}\right] + \frac{\alpha}{F(T)}\left[-\frac{1}{\alpha(1+T)^{\alpha}} + \frac{2}{\alpha(\alpha-1)(\alpha-2)}\right].$$

Figure 2 illustrates, for the α=2.083333 case how the truncated Pareto approaches the untruncated Pareto as the truncation point increases. Note the untruncated Pareto CV is 5 for this case and the first of the three graphs of the figure shows that it is not until the truncation point becomes greater than 10^{10}, does the CV get close to the untruncated value of 5 (note the log scale on the figure's x-axis). Further, the second and third graphs of the figure show the problem is really with the truncated variance slowly converging to the untruncated value. The mean converges rather quickly (at about 10^3, which is still about 3 orders of magnitude greater than the untruncated mean itself which is about .92). Figure 3 shows that the convergence slows as α gets near 2. This illustrates that even discarding a very, very small piece of the untruncated Pareto tail can have a significant effect in trying to simulate the M/P/1 queue. For example, for this case (α = 2.083333), $F(10^3)$=.99999944 so that the tail being discarded is on the order of 10^{-7}, but yields a CV of approximately 3, a 40%

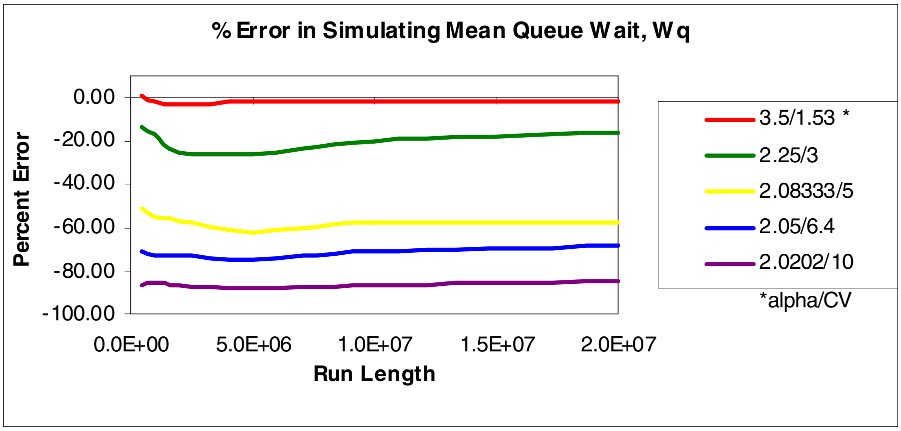


Figure 1: Percent Error in Simulating Mean Queue Wait

difference from the untruncated value of 5! Figure 4 compares the complementary CDF (tail values) of the Pareto, Lognormal, Weibull and Gamma all with the same mean and variance for a CV=6.403 (the Pareto α here is 2.05). We see that for the Pareto (the heaviest of the heavy-tailed distributions) to dominate the non-heavy-tailed Gamma, the value of the variate is about 200 times the mean, and for the Pareto to finally dominate the Lognormal, the variate is about 7000 times the mean.

We next look at how Wq for the truncated Pareto compares with that of the untruncated Pareto (M/PT/1 vs M/P/1) as the truncation point increases. Figure 5 shows that one needs very high truncation points before the truncated Wq nears the untruncated Wq, especially as α gets closer to 2.

5 ADDITIONAL SIMULATION RESULTS

Figure 6 shows Wq vs run size for M/P/1 cases where α=2.25, 2.083333 and 2.020202 (CV=3, 5 and 10). Again, as in Figure 1, simulation falls far short of estimating the PK values, even for run lengths of over 25 million transactions. But, if we take the maximum service time generated for each run, and compare simulation results with PK results for the M/PT/1, with the truncation point T equal to the maximum service time generated, the results track quite well. Table 1 shows the maximum service times generated for the 30 million run lengths, the tail probabilities (1-CDF) discarded beyond the truncation point and the resulting CVs in comparison to the untruncated CVs. Again we see the effect of the very small truncated tail on the CV actually attained.

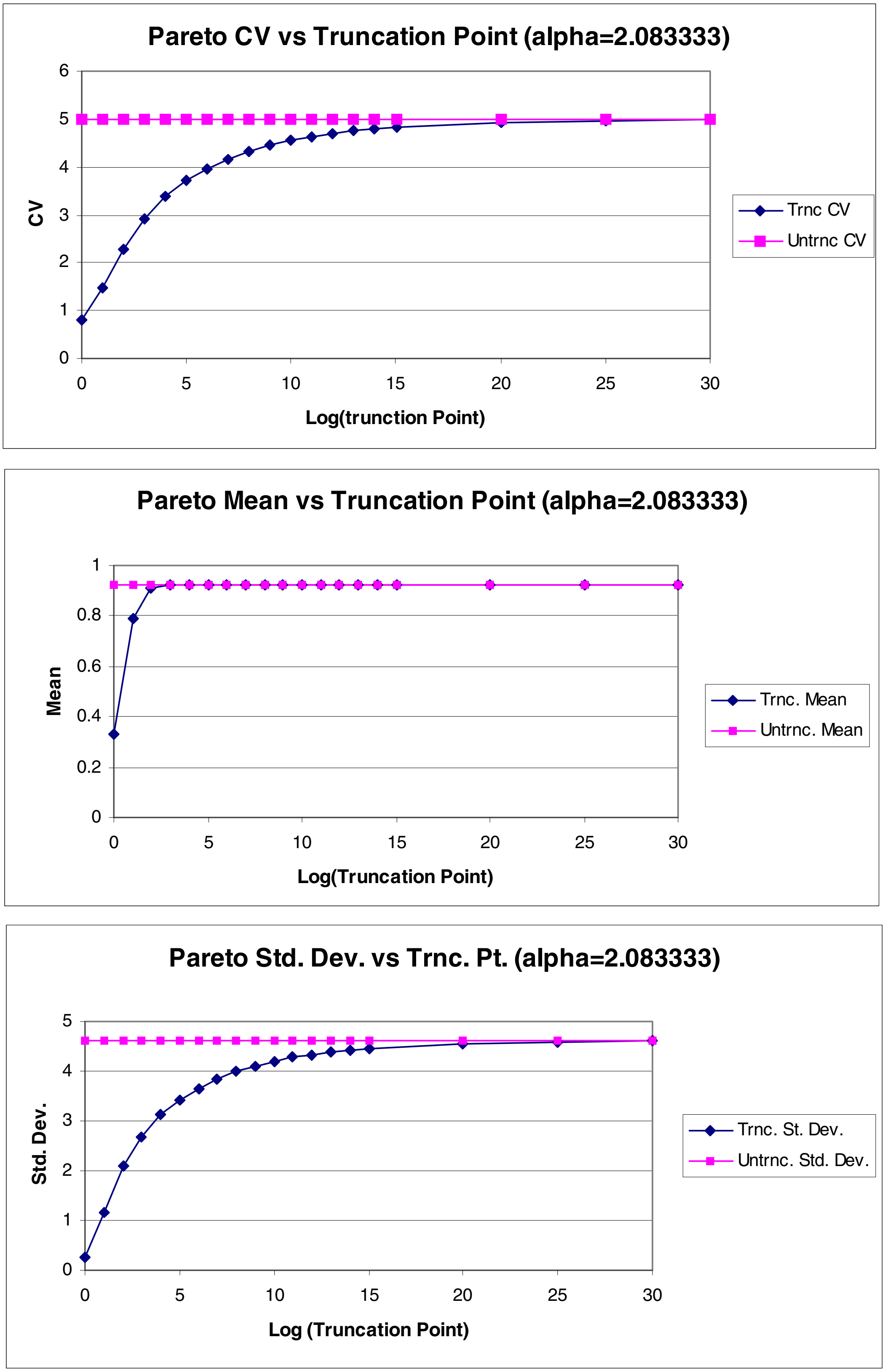
It is interesting to see how close we could theoretically come in simulating the M/P/1 by seeing what the maximum service time that could be potentially generated, assuming long enough run length. The number of significant digits of the ARENA random number generator (ARENA 5.0) is 12 or 13 digits. Table 2 shows the maximum service time possible (a random number drawn consisting of 12 or 13 .9s) and the maximum CV attainable compared to the untruncated CV. Table 3 shows that even if we could get 14 or 15 digit significance in generating 0-1 random numbers, we still fall far short of the untruncated CV. Again we see the extreme effect of discarding a miniscule tail for α values near 2.

6 CONCLUSIONS

The tail of the Pareto distribution significantly affects the performance of simulating mean queue wait in an M/P/1 queue. Very, very small tail probabilities can cause great errors in estimating mean queue wait, Wq, for Pareto shape parameter α ≤ 2.25 (CV ≥ 3). Since any simulation run is really a truncated M/PT/1, if we are really interested in Wq for the untruncated M/P/1, we can run into limitations imposed by the significant digits of the random number generator. Many argue that in reality, there is no such thing as an untruncated distribution since, for example, there is always a maximum value for service time (say a maximum file length or a maximum claim size). However, if the maximum is extremely high, there may still be problems in simulating distributions such as the Pareto, if one is interested in estimating mean performance measures.

ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation Grant DMII-0140232: Development of Procedures to Analyze Queueing Models with Heavy-Tailed Interarrival and Service Times.



Pareto CV vs Truncation Point (alpha=2.083333)

Pareto Mean vs Truncation Point (alpha=2.083333)

Pareto Std. Dev. vs Trnc. Pt. (alpha=2.083333)

Figure 2: Convergence of Truncated to the Untruncated Pareto for α=2.083333

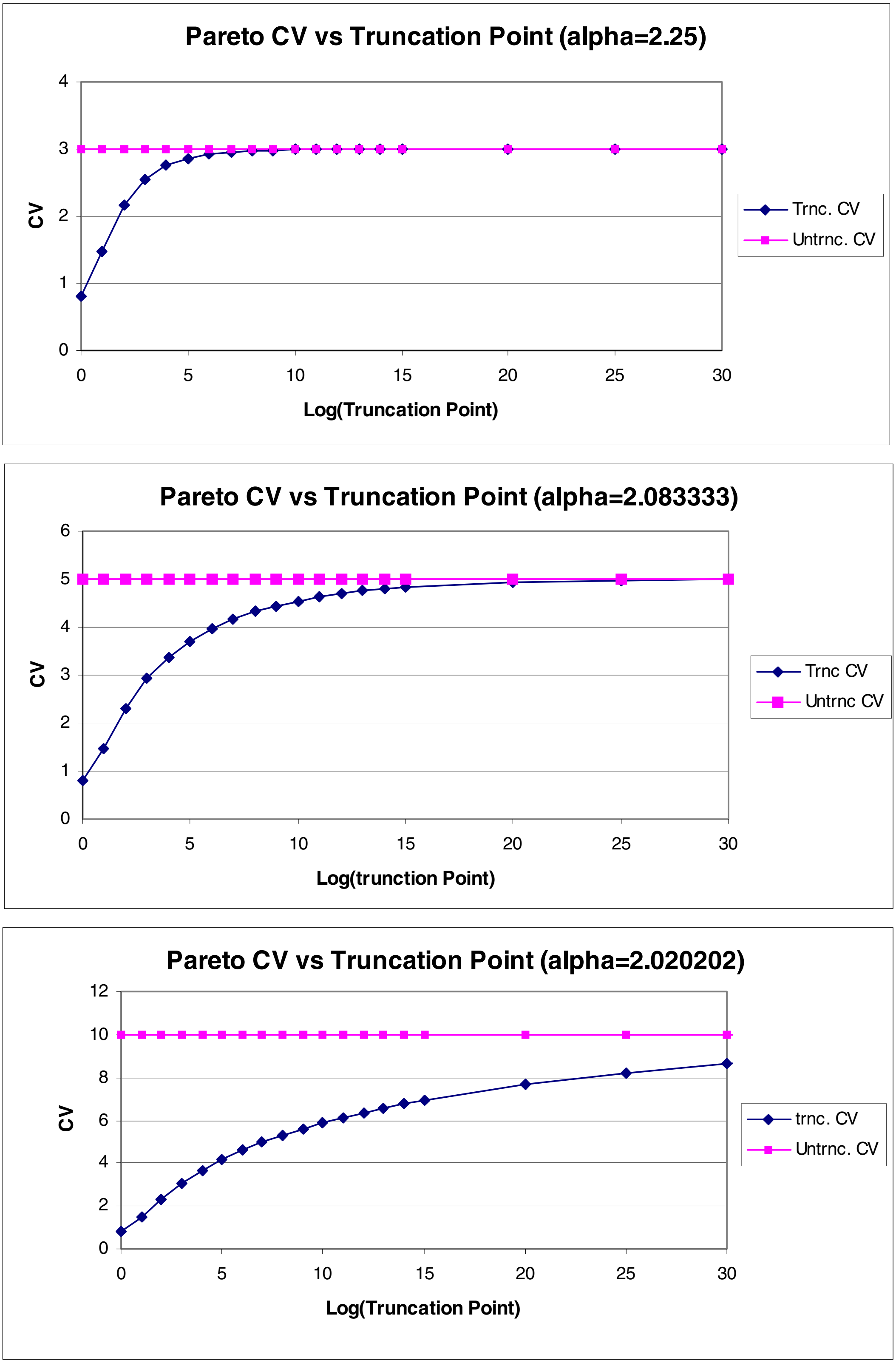


Figure 3: Convergence of Truncated to the Untruncated Pareto for Various α

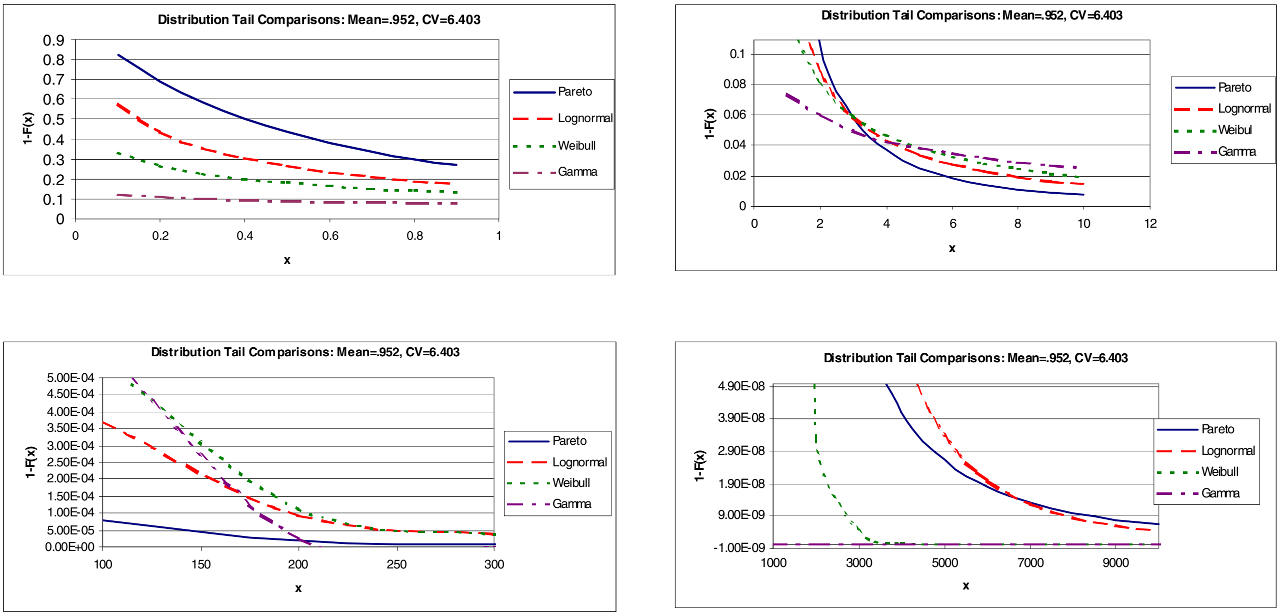


Figure 4: Heavy-Tailed Distributions and the Pareto

Table 1: Maximum Service Times Generated for 30M Run Lengths

alpha	Max ST-30M Run	F(max ST)	1 - F(max ST)	CV(untrnctd)	CV Attained
2.25	3358.0	0.999999988358003	1.16420E-08	3	2.6748
2.083333	6430.6	0.999999988358651	1.16413E-08	5	3.2978
2.05	7416.1	0.999999988358203	1.16418E-08	6.4	3.4720
2.020202	8458.1	0.999999988358421	1.16416E-08	10	3.6474

Table 2: Limits of the Random Number Generator

alpha	F(T):12 digit Arena RN	Max T (ST)	Max CV Attainable	F(T):13 digit Arena RN	Max T (ST)	Max CV Attainable	Untrunctd CV
2.25	0.99999999999900	215444.59	2.8892	0.99999999999990	599400.42	2.9146	3
2.083333	0.99999999999900	575446.27	3.9096	0.99999999999990	1737544.51	4.0174	5
2.05	0.99999999999900	713941.48	4.2380	0.99999999999990	2194812.14	4.3839	6.4
2.020202	0.99999999999900	870972.25	4.5876	0.99999999999990	2722281.74	4.7795	10

Table 3: Extending the Limits of the Random Number Generator

RN Digits	CV (α=2.25)	Untruncated (α=2.25)	CV (α=2.083333)	Untruncated (α=2.083333)	CV (α=2.05)	Untruncated (α=2.05)	CV (α=2.020202)	Untruncated (α=2.020202)
12	2.889	3	3.910	5	4.238	6.4	4.588	10
13	2.915	3	4.017	5	4.384	6.4	4.779	10
14	2.934	3	4.113	5	4.518	6.4	4.960	10
15	2.949	3	4.199	5	4.640	6.4	5.130	10

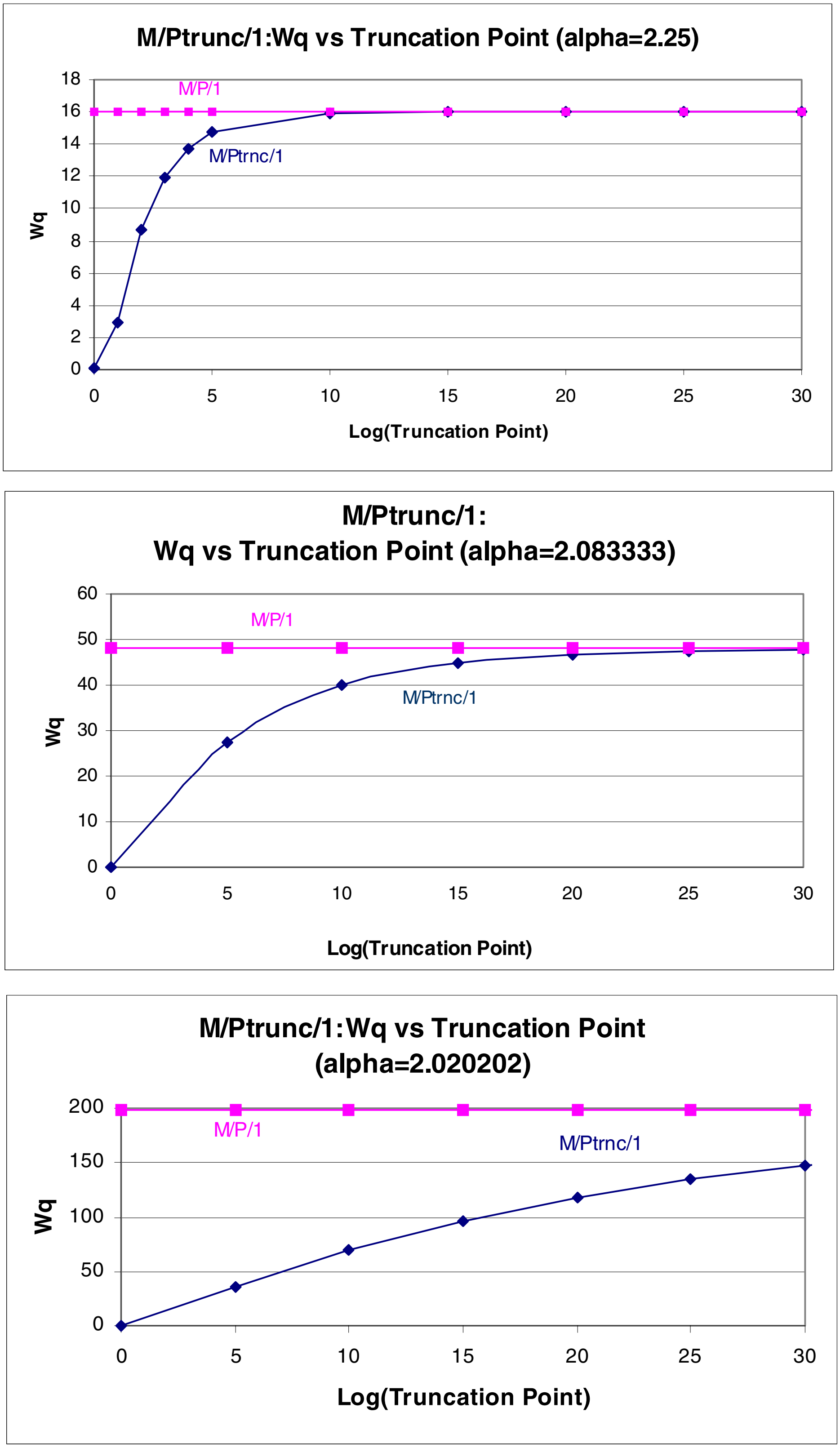


Figure 5: Wq vs. Truncation Point

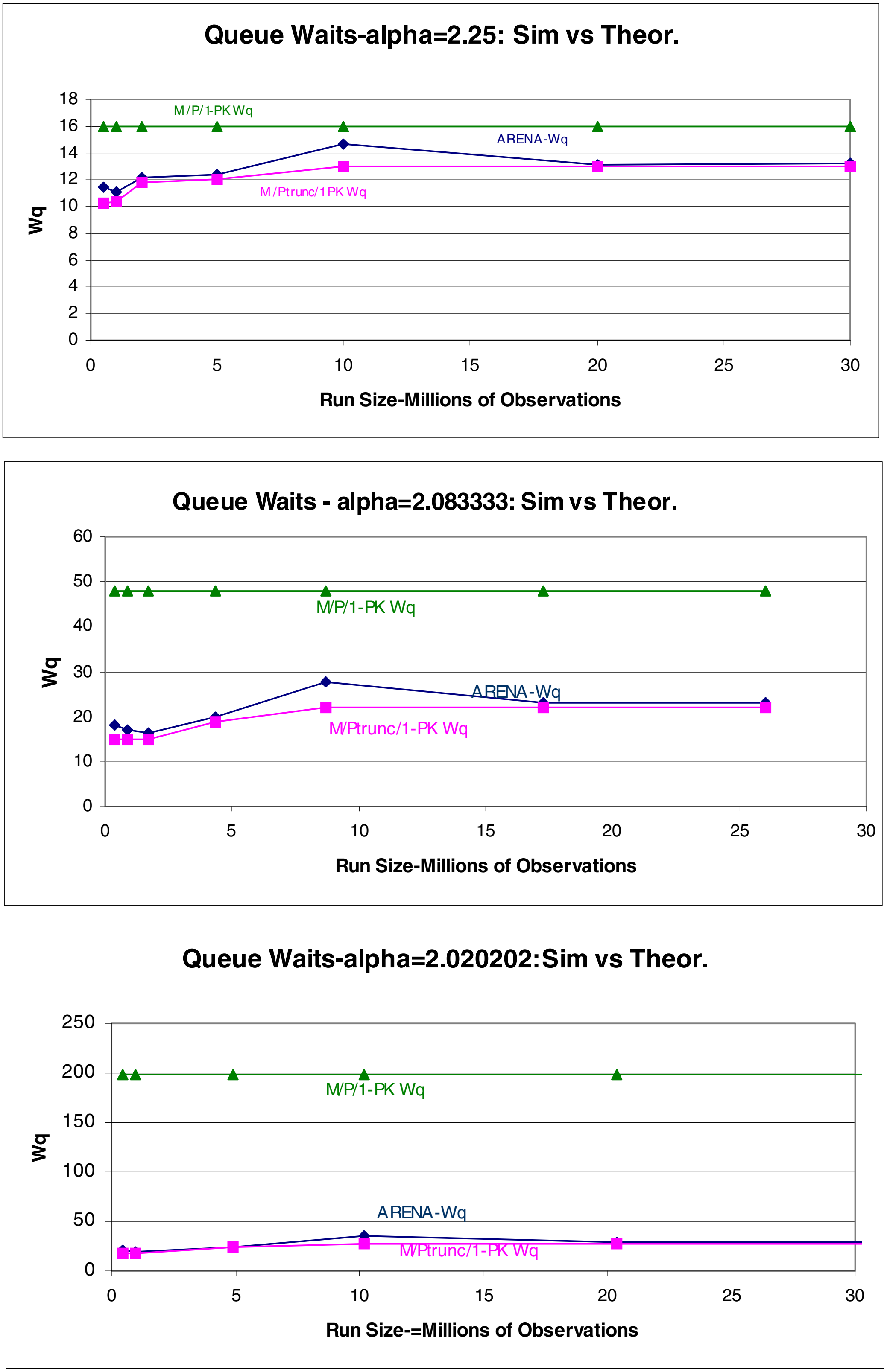


Figure 6: Simulated vs. Theoretical Wq

REFERENCES

Crovella, M. E. and L. Lipsky. 1997. Long-lasting Transient Conditions in Simulations with Heavy-Tailed Workloads. In *Proceedings of the 1997 Winter Simulation Conference*, ed. S. Andradottir, K. J. Healy, D. H. Withers, and B. L. Nelson, 1005-1012. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.Fowler, T. B. 1999. A Short Tutorial on Fractiles and Internet Traffic. *The Telecommunications Review* (McLean, VA) 10: 1-14.

Gross, D. and C. M. Harris. 1998. *Fundamentals of Queueing Theory*. 3rd ed. New York: John Wiley and Sons.

Juneja, S., P. Shahabuddin, and A. Chandra. 1999. Simulating heavy tailed processes using delayed hazard rate twisting. In *Proceedings of the 1999 Winter Simulation Conference*, ed. P. A. Farrington, H. B. Nemhard, D. T. Sturrock, G. W. Evans, 420-427. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Sees, J. C. 2001. Simulating M/Pareto/1 Queues. Ph.D. Dissertation, School of Information Technology and Engineering, George Mason University, Fairfax, Virginia.

Shortle, J.F., P.H. Brill, M.J. Fischer, D. Gross, and D.M.B. Masi. 2002. An Algorithm to Compute the Waiting Time Distribution for the M/G/1 Queue. Submitted to *INFORMS Journal on Computing*.

AUTHOR BIOGRAPHIES

DONALD GROSS is a Research Professor in the Department of Systems Engineering and Operations Research at George Mason University and Professor Emeritus of Operations Research, George Washington University. He is the co-author of the well-known book, *Fundamentals of Queueing Theory,* has numerous publications in the field of queueing theory, and is past president of INFORMS. He was Director, Operations Research and Production Systems at the National Science Foundation, 1988-1990; 1996. He has received the INFORMS Kimball Medal for Service to the Operations Research Profession. His email address is <dgross1@gmu.edu>.

JOHN F. SHORTLE was born in Santa Barbara, CA. He received a B.S. in mathematics from Harvey Mudd College in 1992. He received a Ph.D. and M.S. in operations research at UC Berkeley in 1996. He worked for three years at U S WEST Advanced Technologies developing stochastic, queueing, and simulation models to optimize networks and operations. In 2000, he won the INFORMS Daniel H. Wagner Prize for excellence in Operations Research Practice. He is currently an assistant professor of Systems Engineering at George Mason University. His research interests include simulation and queueing applications in telecommunications. His email address is <jshortle@gmu.edu>.

MARTIN J. FISCHER is a Senior Fellow in Mitretek's Center for Telecommunications and Advanced Technology. He has published over 30 articles in refereed journals. He received an M.S. degree and a Ph.D. degree in Operations Research from Southern Methodist University; and a B.A. degree and an M.S. degree in Mathematics from the University of New Hampshire and Florida State University, respectively. He has over 40 years experience in telecommunications, of which 30 years have been in the network design and performance analysis area. His email address is <mfischer@mitretek.org>.

DENISE M. B. MASI is a Principal Engineer in Mitretek's Center for Telecommunications and Advanced Technology. Her research interests include queueing theory and simulation applied to telecommunication networks. She received her Ph.D. degree in Information Technology and Engineering at George Mason University in 1998, an M.S. degree in Industrial Engineering from Purdue in 1989, and a B.S. in Industrial Engineering from Texas A&M University in 1988. Her email address is <dmasi@mitretek.org>.