Modelling Extreme Uncertainty: Queues with Pareto Inter-Arrival Times and Pareto Service Times

Raul Ramirez-Velarde[1[0000-0001-7186-1914]], Cristobal Pareja-Flores[2], Neil Hernandez-Gress[1[0000-0003-0966-5685]] and Laura Hervert-Escobar[1[0000-0003-2465-7106]]

[1] Tecnologico de Monterrey. Eugenio Garza Sada 2501 Sur, Col. Tecnológico, Monterrey, N. L., Mexico, 64849
[2] Departamento de Sistemas Informáticos y Computación, Facultad de Estudios Estadísticos, Universidad Complutense de Madrid,
28040 Madrid, Spain
rramirez@tec.mx

Abstract. When an operational parameter presents extremely high variability, uncertainty becomes extreme. Long-tail probability distributions can be used to model such uncertainty. We present a queuing system in which extreme uncertainty is modelled using long-tail probability distributions. There have been many queuing analyses for a single server queue fed by an M/G/traffic process, in which G is a Pareto distribution, that focus on certain limiting conditions. In this paper, we present a mathematical model to solve an infinite queuing system with one server where the inter-arrival time between jobs follows a Pareto probability distribution with shape parameter α and a scale parameter A. The system service time is also a Pareto probability distribution with shape parameter β and scale parameter B. We call this the P/P/1 queuing model.

Keywords: Extreme uncertainty, Pareto queues, long-tails.

1 Introduction

When an operational parameter presents extremely high variability, uncertainty becomes extreme. Long-tail probability distributions can be used to model such uncertainty. Many real-world systems exhibit extreme variability, where rare but significant events dominate system behaviour. Earthquakes, wildfires, and floods follow long-tailed distributions, with unpredictable inter-arrival times and durations. Similar patterns appear in network traffic, where bursts of data cause congestion, and in financial markets, where sudden crashes disrupt stability. Cybersecurity threats and human communication also display heavy-tailed activity, making prediction and management challenging.

This paper models a P/P/1 queuing system, where both inter-arrival and service times follow Pareto distributions, capturing extreme uncertainty. We provide mathematical formulations and simulations to analyse the system's behaviour, offering insights

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_16

relevant to telecommunications, disaster response, and computational workload management.

In this paper we will discuss a queuing system with the following characteristics:

- The inter-arrival time between jobs has a Pareto probability distribution with shape parameter α and a scale parameter A.
- The service time has a Pareto probability distribution with shape parameter β and scale parameter B.
- The queue is infinite.
- There is only one server.

We will call this the P/P/1 queuing system.

The probability distribution for random variable that represents the inter-arrivals time is defined by the Pareto I probability distribution with shape parameter α, and location parameter A:

$f(t) = \alpha \left(\frac{t}{A}\right)^{-\alpha-1}$, with $E[t] = \frac{\alpha A}{\alpha-1}$. The corresponding survival function is:

$$S(t) = 1 - F(t) = \left(\frac{t}{A}\right)^{-\alpha} \tag{1}$$

The probability distribution for the service time is also distributed as a Pareto I random variable with β as shape parameter and B as scale parameter.

$g(t) = \beta \left(\frac{t}{B}\right)^{-\beta-1}$, with $E[t] = \frac{\beta B}{\beta-1}$. The corresponding survival function is:

$$Z(t) = 1 - F(t) = \left(\frac{t}{B}\right)^{-\beta} \tag{2}$$

2 Previous Work

Many simulation studies have been undertaken to evaluate the performance of queueing systems with heavy-tail distributed inter-arrival times, execution times and transfer times. In this section, we review some of the most relevant work in this area, focusing on studies that have used the Pareto distribution to model heavy-tailed distributions.

Fischer and Cart [1] studied the properties and use of the Pareto distribution to model a M/Pareto/1 queue and a Pareto/M/1 queue. They showed that both systems can be used to model the transmission of information in a network, with the former being more suitable for switched networks and the latter being more suitable for packet transmission.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_16

To overcome the difficulties of simulating systems with heavy-tailed distributions, Argibay Losada et al. [2] proposed a method to speed up simulations. They used M/G/1 systems as workbenches and showed that their method could significantly reduce the simulation time. Gross et al. [3] investigated the difficulties of simulating queues with Pareto service. They considered truncated Pareto service and showed that it can lead to significant errors in the estimation of queue performance. Koh and Kim [4] studied the queue performance of Pareto/M/1/k using simulations. They investigated the queue behaviour with Pareto inter-arrival distribution and showed that the asymptotic and exact loss probabilities can be significantly different.

Fischer et al. [5] studied the one-parameter, two-parameter, and three-parameter Pareto distributions. They showed that the two-parameter Pareto distribution can result in lower congestion than the one-parameter Pareto distribution. Inmaculada et al. [6] derived estimators for the truncated Pareto distribution. They also investigated the distribution properties and illustrated its applicability in practice. Albrecher et al. [7] investigates parameter estimation for tempered Pareto-type distribution tempered with a general Weibull distribution in risk management and insurance, offering improved methods for parameter estimation.

Recent studies have further explored the impact of heavy-tailed distributions on queuing systems. Jiang et al. [8] quantified the efficiency of parallelism in systems prone to failures and exhibiting power law processing delays and channel availability. They characterized the performance of redundant and split parallelism schemes in terms of the power law exponent and delay distribution tail asymptotics.

Building on these results, we model both service times and arrival times as Pareto random variables, without truncation, and derive exact and asymptotic queuing behaviour models for a single server and investigate the resulting probability distributions.

3 Modelling Heavy Tails

A heavy-tailed distribution is a distribution, for which the tail is heavier than any exponential tail. In this distribution, the probability of observing a value far from the median is greater than it would be in the normal distribution. That is, the probability of extreme values is non-negligible.

More precisely, a distribution of a random variable X with function f, and cumulative function

$$F(x) = P[X \leq x],$$

where $P[X]$ is the corresponding probability density function, is said to be heavy right tailed if $\lim_{x\to\infty} e^{\varphi x} F(x) = \infty$, for any $\varphi > 0$. It tends to have infinite moments, such as infinite variance [9].

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_16

Note, that a moment is infinite, if the integral that defines the statistical moment converges too slowly to be integrated (divergent), therefore, the moment does not exist. Since, a heavy-tailed distribution is also a long-tailed distribution, it follows that

$\lim_{x\to\infty} P[X > x + t|X > x] = 1$, or, equivalently, $\lim_{x\to\infty} \frac{F(x+t)}{F(x)} = 1$, for any $t > 0$.

Intuitively, equation states that if x exceeds some large value, then it is equally likely that it will exceed an even larger value as well. For our problem domain, it means that if the system executing a task spends large amount of time, probably it will spend longer time to complete it [10].

3.1 Probability Distribution of Jobs on System

Let us start by investigating the probability that certain time lag, say the inter-arrival time (but it could also have been the execution time) will persist further into the future less than Δt units of time given the fact that we know it will persist more than τ time units. Assume $\Delta t \ll 1$, thus there cannot happen more than one event in Δt. Also, assume that the time is discrete, and the time slot is the smallest time interval considered in the system, for instance, seconds.

In other words, we wish to find out $P[t < \tau + \Delta t \,|t > \tau]$. For a Pareto r.v.,

$P[t < \tau + \Delta t\,] = 1 - \left(\frac{\tau+\Delta t}{A}\right)^{-\alpha}$. $P[t > \tau] = \left(\frac{\tau}{A}\right)^{-\alpha}$. Now, the probability

$P[t < t + \Delta t \; and \; t > \tau\,] = \left(\frac{\tau}{A}\right)^{-\alpha} - \left(\frac{\tau+\Delta t}{A}\right)^{-\alpha}$.

Therefore, $P[t < t + \Delta t \,|t > \tau] = \frac{\left(\frac{\tau}{A}\right)^{-\alpha} - \left(\frac{\tau+\Delta t}{A}\right)^{-\alpha}}{\left(\frac{\tau}{A}\right)^{-\alpha}} = 1 - \left(1 + \frac{\Delta t}{\tau}\right)^{-\alpha}$. We observe that the corresponding survival function $\psi_{arr}(\Delta t) = P[t > \tau + \Delta t \,|t > \tau]$ $= \left(1 + \frac{\Delta t}{\tau}\right)^{-\alpha}$ is a Lomax probability distribution, for which a power series approximation exists. Also, $\psi_{ser}(\Delta t) = \left(1 + \frac{\Delta t}{\tau}\right)^{-\beta}$ for the service time.

Since A is the minimum time for t, let us fix τ as A for inter-arrival times and B for service times, to address the general cases, as in any given moment an event will always have existed for at least A (or B) units of time before persisting into the future.

Recall that $e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots$ and that $(1 + x)^p = 1 + px + \frac{p(p-1)}{2}x^2 + \cdots$. In a Markovian queueing system with survival probability $F(t) = e^{-\lambda\Delta t} = 1 - \lambda\Delta t + \frac{(\lambda\Delta t)^2}{2} - \cdots$, we approximate

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_16

$P[No\ event] = P[\Delta t > \tau] \approx 1 - \lambda\Delta t$, and under the assumption that only one event can occur in the period Δt, $P[One\ event] = P[t < \tau] \approx \lambda\Delta t$ [11].

Similarly, with Pareto times:

$$P[No\ event] = P[\Delta t > \tau] = \left(1 + \frac{\Delta t}{A}\right)^{-\alpha} \approx 1 - \frac{\alpha\Delta t}{A} \tag{3}$$

and

$$P[One\ event] = P[\Delta t < \tau] = 1 - \left(1 + \frac{\Delta t}{A}\right)^{-\alpha} \approx \frac{\alpha\Delta t}{A} \tag{4}$$

Now we solve the state probabilities for the queueing system. Let $p_i(t)$, the probability that there are *i* jobs at time *t*. Then,

$$p_0(t + \Delta t) = p_0(t)\left(1 - \frac{\alpha\Delta t}{A}\right) + p_1(t)\left(\frac{\beta\Delta t}{B}\right)$$

This means that there are two ways in which there could be zero jobs on the system, either o job arrived (and no jobs left) or there was one job and then it finished and left. With some algebra:

$\frac{p_0(t+\Delta t) - p_0(t)}{\Delta t} = p_0(t)\frac{\alpha}{A} + p_1(t)\frac{\beta}{B}$, assuming stationary and steady state probabilities $\frac{dp_i(t)}{dt} = 0, then$

$$p_1(t) = \frac{\alpha B}{\beta A} p_0(t)$$

For one job we have:

$$p_1(t + \Delta t) = p_0(t)\frac{\alpha\Delta t}{A} + p_2(t)\frac{\beta\Delta t}{B} + p_1(t)\left(1 - \frac{\alpha\Delta t}{A}\right)\left(1 - \frac{\beta\Delta t}{B}\right)$$

Since Δt is so small, and there cannot happen more than one event in such small period, all powers of Δt vanish:

$$p_2(t) = \left(\frac{\alpha B}{\beta A}\right)^2 p_0(t)$$

And in general,

$$p_n(t) = \left(\frac{\alpha B}{\beta A}\right)^n p_0(t) \tag{5}$$

By calling $\rho = \frac{\alpha B}{\beta A}$ and we find that $p_n = \rho^n p_0$. Also,

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_16

$\sum_{n=0}^{\infty} p_n = \sum_{n=0}^{\infty} \rho^n p_0 = p_0 \left(\frac{1}{1-\rho}\right) = 1$, we find that $p_0 = 1 - \rho$.

Therefore, the number of jobs on a Pareto queuing system still follows a geometric probability distribution $p_n = \rho^n(1-\rho)$. Then.

$$E[n] = \sum_{n=0}^{\infty} n\rho^n(1-\rho) = \frac{\rho}{1-\rho}$$

A result supported by Whitt [12]. Nevertheless, for a heavy-tailed random variable we find that the probability of no event happening in a very small Δt interval is almost one, as events tend heavily to persist into the future, or $p_0 \approx 1$. Thus empirically, a closer approximation for the expected number of jobs on the system might be (called P/P/1 Series model):

$$E[n] = \frac{\rho}{(1-\rho)^2} \tag{6}$$

And the pdf is closer to $p_n = \rho^n$

Now let us explore the probability distribution of the residence time or system time, the time a job stays on the system from entering until it has been served and then exits.

3.2 Probability Distribution of System Time of M/M/1 Queue

Let us review first the Markovian case. Let us call $f_0(t)$ the probability distribution of the time a job stays on the system when there are no other jobs on the system, that is, the execution time. Assume all execution times are i.i.d. random variables with $f_0(t) = \mu e^{-\mu t}$. If there is one job on the system when a new job arrives, then the time the arriving job stays on the system is the execution time of both jobs, that is, the addition of two exponentially distributed random variables. Then, the probability distribution of the system time of one job when there is one other job already in the system is [11]:

$$f_1(t) = \int_0^t f_0(\alpha) f_0(t-\alpha) d\alpha = \int_0^t \mu e^{-\mu\alpha}\, \lambda e^{-\mu(t-\alpha)} d\alpha = \mu^2 t e^{-\mu t}$$

Following on that:

$$f_2(t) = \int_0^t f_1(\alpha) f_0(t-\alpha) d\alpha = \frac{\mu^3 t^2 e^{-\mu t}}{2}$$

And in general

$f_n(t) = \frac{\mu^{n+1} t^n e^{-\mu t}}{n!}$, $f_n(t)$ is then the probability distribution of the system time when there are n jobs on the system. Recall that $p_n$is the probability that there are n jobs on queue. Therefore, the probability distribution of the system time of any arriving job is (with $\rho = \frac{\lambda}{\mu}$),

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_16

$$f(t) = \sum_{n=0}^{\infty} p_n f_n(t)$$

$f(t) = \sum_{n=0}^{\infty} p_n f_n(t) = \sum_{n=0}^{\infty} \rho^n (1-\rho) \frac{\mu^{n+1} t^n e^{-\mu t}}{n!} = (\mu - \lambda) e^{-(\mu-\lambda)t}$, and

$$E[t] = \int_{t=0}^{\infty} t(\mu - \lambda) e^{-(\mu-\lambda)t} dt = \frac{1}{\mu - \lambda} \quad (7)$$

What are the probability distributions of the number of jobs in the system and the total system time when execution times and the inter arrival times are Pareto distributed random variables? Firstly, we will answer that question for self-similar variables. Secondly, for regular Pareto random variables.

3.3 Self-Similarity and Heavy-Tails

Now, we describe how heavy-tails can cause self-similarity (fractal behaviour), and long-range dependence. Let us first assume that we deal with polynomial decay of the tail of the probability distribution, and use the Pareto Distribution as an example of heavy tailed distribution [13]: $\boldsymbol{P[X > x] = x^{-\alpha} L(x)}$,

where $\alpha = 1/\gamma$ is an inverse of the extreme value index γ. L is a slow varying function at infinity, that is $\boldsymbol{\lim_{x \to \infty} \frac{L(\gamma x)}{L(x)} = 1}$, for $\boldsymbol{x > 0}$.

In non-self-similar data, the average of a series of samples tends to the population mean, as the number of samples increases [11]. That is:

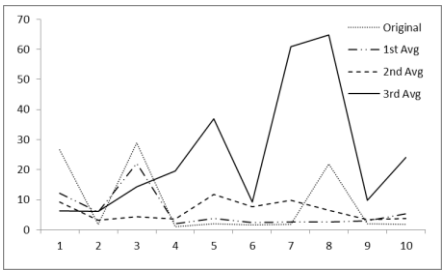

Fig. 1. Aggregation of self-similar data (Pareto distribution). Average does not smooth

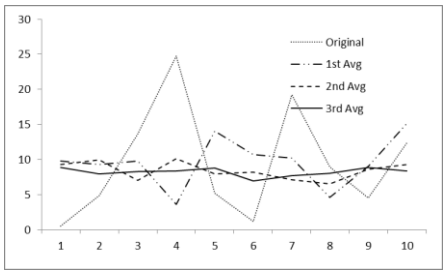

Fig. 2. Aggregation of non-self-similar data (exponential distribution). Average is smoothed

$P\left[\lim_{x \to \infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mu\right] = 1$, where $\mu = E[X_i]$, $Var\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] = \frac{\sigma^2}{n}$, and σ is the standard deviation of the population.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_16

The latter means that deviations of the sample mean with respect to the population mean decay proportionally to the size of the sample. Hence, as we aggregate data averaging larger collections, the averages become smoother, approaching the sample mean.

In [14], it is shown that in self-similar data something different happens: $Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \sigma^2 n^{-\alpha}$, where $\alpha < 1$. Hence, the sample mean converges to the population mean much slower. This implies that, with non-negligible probability, the execution time of a collection of jobs can be much larger or much smaller than the execution time computed using population mean. There is a non-negligible probability of having a exceedingly long runtime to solve the tasks, considerably longer than the population mean. It can create the self-similar profile shown in Figs. 1-2.

There are several, not equivalent, definitions of self-similarity. The standard one states that a continuous-time process $Y = \{Y(t), t \geq 0\}$ is self-similar (with self-similarity or Hurst parameter H), if it satisfies the condition [15]:

$$Y(t) \equiv a^{-H} Y(at), \forall \text{ t}, \forall \text{a} > 0, \text{ and } 0.5 < H < 1 \tag{8}$$

where the equality means that the expressions have equivalent probability distribution.

A process satisfying Eq. (8) can never really be stationary one as it requires that $Y(t) \equiv Y(at)$, (or rather the distribution of $\{Y(t+s) - Y(t)\}$ does not depend on t). As we show bellow, in our application, this does not hold, so we assume that $Y(t)$ has stationary increments. Let us $t = 1$ and a$= t$ in Eq. (1), thus,

$\text{Y(t)} \equiv t^H \text{Y(1)}, \forall \text{ t}, 0.5 < H < 1.$

Notice also, that eq. (8) in the context of time series analysis implies that [15] [16]:

$$z_n(t) = \sum_{i=1}^{n} X(i) \equiv n^H X(1) \tag{9}$$

Where the equality represents equality in probability distribution and $z_n(t)$ is the accumulation process for n jobs execution time.

We assume $z_1(t)$ to be the service time for any job, that is, $f_0(t) = f(z_1(t))$ is the probability distribution of the wait when there is no job been serviced when the new job arrives. That is that wait time is only the service time for the new job, and $f_n(t) = f(z_{n+1}(t))$. Then the probability distribution of the wait time is:

$$f(t) = \sum_{n=0}^{\infty} p_n f_n(t) \tag{10}$$

With

$$f_n(t) = f(z_{n+1}(t)) = (n+1)^H f(z_1(t)) \tag{11}$$

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_16

3.4 Using Fractional Differentiation to Find the Probability Distribution of Service Time

Where $f(z_n(t))$ is the probability distribution of the system time when there are n jobs on the system and $f(z_1(t))$ is the probability distribution of the system time when there is only one job on the system, or the execution time for one job (a presumably long-tailed pdf). Thus, the probability distribution for system time is:

$$f(t) = \sum_{n=0}^{\infty} p_n f_n(t) = \sum_{n=0}^{\infty} (n+1)^H f(z_1(t))\, \rho^n$$

$$f(t) = f(z_1(t)) \sum_{n=0}^{\infty} (n+1)^H \rho^n$$

Recall from fractional differentiation that:

$$\frac{d^\alpha x^k}{dx^\alpha} = \frac{k!}{(k-\alpha)!} x^{k-\alpha} \tag{12}$$

Now, we use Stirling's approximation:

$$\frac{k!}{(k-\alpha)!} = \frac{k^k e^{-k}\sqrt{2\Pi k}}{(k-\alpha)^{k-\alpha} e^{-(k-\alpha)}\sqrt{2\Pi(k-\alpha)}} = \frac{k^k}{(k-\alpha)^{k-\alpha}} e^{-\alpha} \sqrt{\frac{k}{k-\alpha}}$$

For $k \gg \alpha$,

$$\frac{k^k}{(k-\alpha)^{k-\alpha}} = \left(1 - \frac{\alpha}{k}\right)^{-k} (k-\alpha)^{\alpha} \sim e^{\alpha} k^{\alpha}$$

With,

$$\sqrt{\frac{k}{k-\alpha}} \sim 1$$

Therefore,

$$\frac{d^\alpha x^k}{dx^\alpha} = \frac{k!}{(k-\alpha)!} x^{k-\alpha} = e^{\alpha} k^{\alpha} e^{-\alpha} (1) x^{k-\alpha} = k^{\alpha} x^{k-\alpha} \tag{13}$$

Since $k \gg \alpha$ is not our case, we will test the accuracy of the previous approximation for $k = -1$ and $\alpha = 1$, since $\frac{d^1 x^{-1}}{dx^1} = -\frac{1}{x^2}$. That is:

$$\frac{d^\alpha x^k}{dx^\alpha} = \frac{dx^{-1}}{dx} = \frac{k!}{(k-\alpha)!} x^{k-\alpha} = \frac{(-1)!}{(-1-1)!} x^{-1-1} = \frac{(-1)!}{(-2)!} x^{-2}$$

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_16

$$= \frac{-1^{-1}}{(-2)^{-2}} e^{-1} \sqrt{\frac{-1}{-2}} (x^{-2}) = -\frac{4\sqrt{0.5}}{e} = -1.0405x^{-2} \sim -x^{-2}$$

Or a 4% error. Now with $\alpha = H$, and $k = -1$, and considering the chain rule for differentiation, we find:

$$\frac{d^H}{da}\left(\frac{1}{1-a}\right) = \frac{d^H}{da}(1-a)^{-1} = \frac{(-1)!}{(-1-H)!}(1-a)^{-1-H}\frac{d^H}{da}(-a)$$
$$= \frac{(-1)!}{(-1-H)!}(1-a)^{-1-H}\left(\frac{1!}{(1-H)!}H(-1)\right) \sim \frac{a^{1-H}}{(1-\alpha)^{1+H}}$$

Therefore if,

$$\sum_{n=0}^{\infty} a^n = \frac{1}{1-a}$$

Then

$$\frac{d^H}{da}\sum_{n=0}^{\infty} a^n = \frac{d^H}{da}\left(\frac{1}{1-a}\right)$$

thus,

$$\sum_{k=0}^{\infty} k^H a^{n-H} \sim \frac{a^{1-H}}{(1-\alpha)^{1+H}} \tag{14}$$

and,

$$\sum_{n=0}^{\infty} n^H \rho^n \sim \frac{\rho}{(1-\rho)^{1+H}} \tag{15}$$

Then, the probability distribution of the system time is:

$$f(t) = f(z_1(t)) \sum_{n=0}^{\infty} (n+1)^H \rho^n$$

With

$$\sum_{n=0}^{\infty} (n+1)^H \rho^n = \sum_{j=1}^{\infty} j^H \rho^{j-1} = \frac{1}{\rho}\sum_{j=0}^{\infty} j^H \rho^j$$

Then,

$$f(t) = f(z_1(t))\left(\frac{1}{\rho}\right)\left(\frac{\rho}{(1-\rho)^{1+H}}\right) =$$

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_16

$$f(t) = \frac{f(z_1(t))}{(1-\rho)^{1+H}}$$

Consequently,

$$E[t] = \frac{E[z_1(t)]}{(1-\rho)^{1+H}}$$

If $z_1(t)$ is has Pareto pdf then,

$$E[t] = \frac{\beta B}{(\beta-1)(1-\rho)^{1+H}} \tag{16}$$

Since Little's Law indicates that:

$$E[t] = E[n]E[arr\ time] = (E[n]+1)E[ser\ \ time] \tag{17}$$

Then it follows that $E[n] = \frac{1}{(1-\rho)^{1+H}} - 1$

Therefore (called P/P/1 Frac1 model),

$$E[n] = \frac{1}{(1-\rho)^{1+H}} - 1 = \frac{1-(1-\rho)^{1+H}}{(1-\rho)^{1+H}} \approx \frac{1-(1-(1+H)\rho)}{(1-\rho)^{1+H}} = \frac{(1+H)\rho}{(1-\rho)^{1+H}} \tag{18}$$

4 Modelling Sum of Pareto Random Variables

The central limit theorem (CLT) states that, under appropriate conditions, the distribution of a normalized version of the sample mean converges to a standard normal distribution. In the same manner, the addition of random variables with α-stable (long-tailed probability distribution), when normalized, approaches a well-defined stable limiting distribution which depends on α or β [17]. This will allow us to derive a quasi-asymptotic model for our queuing system.

The Generalized Central Limit Theorem states that the properly normalized sum $S_n = \sum_i^N z_i$ of many i.i.d. Pareto r.v.s may be approximated by a stable distribution:

$$\lim_{n\to\infty} P\left[\frac{S_n - b_n}{\frac{1}{n^{\beta}C_\beta}} < \varphi\right] = F_\beta(x) \tag{19}$$

In (14) $F_\beta(x)$ is a stable distribution with index β. The normalization coefficient is:

$$C_\beta = \left[\Gamma(1-\alpha)\cos\left(\frac{\pi\alpha}{2}\right)\right]^{1/\beta} \tag{20}$$

The shift coefficient is:

$$b_n = \frac{n\beta}{\beta-1} \tag{21}$$

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_16

4.1 Tail Asymptotic of Heavy-Tail Sums

According to Zaliapin et al [18] an approximation for the upper quantiles can be obtained by noticing that the tail $1-F_\beta(x)$ of a stable distribution has a simple asymptotic:

$$\lim_{x\to\infty} x^\beta(1-F_\beta) = C_\alpha^{-\beta}.$$

More plainly with r.v.s with Pareto distribution:

$$\lim_{n\to\infty} P[S_n > \varphi] = \theta = \left(\frac{\varphi - b_n}{\frac{1}{n^{\frac{1}{\beta}}}}\right)^{-\beta} = n(\varphi - b_n)^{-\beta} \tag{22}$$

And thus,

$$\mathrm{P}[S_n = t] = n\beta(t-b_n)^{-\beta-1} \tag{23}$$

Recall that $(1+x)^p \sim 1+px$. Also, $(a-b)^{-\beta} = a^{-\beta}\left(1+\frac{b}{a}\right)^{-\beta} \sim a^{-\beta}\left(1+\frac{\beta b}{a}\right)$.

Therefore,

$$P[S_n = t] = n\beta(t-b_n)^{-\beta-1} \sim n\beta t^{-\beta-1}\left(1+\frac{\frac{(\beta+1)n\beta}{(\beta-1)}}{t}\right) = n\beta t^{-\beta-1}\left(1+\frac{n\beta(\beta+1)}{t(\beta-1)}\right) \tag{24}$$

Therefore, the probability distribution for system time is:

$$f(t) = \sum_{n=0}^{\infty} p_n f_n(t) = \sum_{n=0}^{\infty} p_n\, P[S_{n+1} = t]$$

$$= \sum_{n=0}^{\infty} \rho^n(1-\rho)\,(n+1)\beta t^{-\beta-1}\left(1+\frac{(n+1)\beta(\beta+1)}{t(\beta-1)}\right)$$

$$= (1-\rho)\beta t^{-\beta-1}\left[\sum_{n=0}^{\infty}\rho^n(n+1) + \frac{\beta(\beta+1)}{t(\beta-1)}\sum_{n=0}^{\infty}(n+1)^2\rho^n\right]$$

$$= (1-\rho)\beta t^{-\beta-1}\left[\frac{1}{(1-\rho)^2} + \frac{\beta(\beta+1)}{t(\beta-1)}\left(\frac{\rho+1}{(1-\rho)^3}\right)\right]$$

$$f(t) = \beta t^{-\beta-1}\left[\frac{1}{(1-\rho)} + \frac{\beta(\beta+1)}{t(\beta-1)}\left(\frac{\rho+1}{(1-\rho)^2}\right)\right] \tag{25}$$

And by the Pareto scaling, if $y = Ax$, then $E[y] = AE[x]$, we have (P/P/1 Par Sum):

$$E[t] = B\left[\frac{\beta}{(1-\rho)(\beta-1)} + \frac{\beta(\beta+1)(1+\rho)}{(\beta-1)(1-\rho)^2}\right] \tag{26}$$

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_16

5 Results and Discussion

To validate the models, a discrete event simulation was carried out. The inter-arrival time of each job is Pareto I probability distribution with shape parameter α, and location parameter A, $f(t) = \alpha\left(\frac{t}{A}\right)^{-\alpha-1}$. For the simulation, $\alpha = 1.7$ and $A = 1.77059$, which makes the mean inter-arrival time of $E[arr\ time] = \bar{A} = 4.3$. The probability distribution for the service time is also distributed as a Pareto I random variable with β as shape parameter and B as scale parameter, $g(t) = \beta\left(\frac{t}{B}\right)^{-\beta-1}$. For the simulation, $\beta = 1.8$ and $B = 1.51111$, which makes the mean service time of $E[serv\ time] = \bar{S} = 3.4$. Each run consisted of one single batch of 1,000,000 arrivals, with the results shown in Table 1. Performance measures of Table 1 were estimated using observed value means allowing for transient period.

Simulation results yield a value of W=E[t]= 74.99923 with standard deviation of 46.5673, and value of L=E[n]= 17.75728 with standard deviation of 10.68285. Results for M/G/1 model are also shown.

	Sim	M/M/1	P/P/1 Series	P/P/1 Frac1	P/P/1 Par Sum	M/G/1
Eq. ρ		$\frac{\bar{S}}{\bar{A}}$	$\frac{\alpha B}{\beta A}$	$\frac{\alpha B}{\beta A}$	$\frac{\alpha B}{\beta A}$	$\frac{\bar{S}}{\bar{A}}$
Eq. E[t]		$\frac{1}{\mu-\lambda}$	$\frac{\rho}{(1-\rho)^2}\bar{A}$	$\left(\frac{(1+H)\rho}{(1-\rho)^{1+H}}\right)\bar{A}$	$B\left[\frac{\beta}{(1-\rho)(\beta-1)} + \frac{\beta(\beta+1)(1+\rho)}{(\beta-1)(1-\rho)^2}\right]$	$\frac{\rho^2+\lambda^2\sigma^2}{2(1-\rho)}+\rho$
Eq. E[n]		$\frac{\rho}{1-\rho}$	$\frac{\rho}{(1-\rho)^2}$	$\frac{(1+H)\rho}{(1-\rho)^{1+H}}$	$\frac{B}{\bar{A}}\left[\frac{\beta}{(1-\rho)(\beta-1)} + \frac{\beta(\beta+1)(1+\rho)}{(\beta-1)(1-\rho)^2}\right]$	$\frac{\frac{\rho^2}{\lambda}+\lambda\sigma^2}{2(1-\rho)}+\frac{1}{\mu}$
ρ	0.787431	0.806037	0.790697	0.790697	0.790698	0.790698
E[t]	74.99923	17.52916	77.612345	64.738760	405.389135	51.262932
E[n]	17.75728	4.155635	18.049383	15.055525	94.2765432	11.921612

Table 1. Comparing the average with derived models

6 Discussion

In Table 1 we can see that eq. (6), $\frac{\rho}{(1-\rho)^2}$, gives the best approximation to $E[n]$ whereas eq. (18), $\frac{\rho}{(1-\rho)^2}\bar{A}$, gives the best approximation to $E[t]$, meaning that model P/P/1 Series is the best approximation, closely followed by the P/P/1 Frac1 model.

It is also interesting to note that the performance of the well-known M/G/1 model is much better than the M/M/1 as simulation results give $L = E[n] = 17.76$, with

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_16

M/M/1 estimating $L = E[n] = 4.16$ and M/G/1 estimating $L = E[n] = 11.93$, a much better result.

Also, even though at first it appears that the P/P/1 Par Sum is way off, with a result of $L = E[n] = 94.28$, it is important to remember that it is an asymptotic model. In fact, the average maximum queue length observed, that is, maximum congestion, is $\bar{L}_{max} = 1{,}121.4$ jobs, with an all-simulations maximum of 3,438. These extremely large values result from rare, but high-impact events of extremely high service times. In these extreme cases, P/P/1 Par Sum model would yield better results.

7 Conclusion

The P/P/1 queueing system is a powerful tool for modelling a wide variety of real-world systems. In this paper, we have shown how the P/P/1 queue can be used to model systems with high variability in the inter-arrival and service times. Our results show that the P/P/1 Series model can accurately predict the mean number of jobs in the system and the mean residence time. This model assumes that the probability of no event happening in a very small Δt is almost one, as events tend heavily to persist into the future. This assumption is supported by our simulation results, which show that the P/P/1 Series model can closely approximate the mean number of jobs in the system and the mean residence time. We also find that the P/P/1 Frac1 model is also a close match with simulation results, although is fall short in estimating some parameters. Interestingly, the P/P/1 Par Sum model models better conditions in which there is congestion because the occurrence exceedingly large service time rare event.

We identify as current limitations of our work that the simulation assumes infinite queue capacity, which may not hold in real-world systems with finite resources. Also, the models rely on the assumption of steady-state conditions, which may not be valid during transient phases or under extreme variability.

Future research could explore extending this model to multi-server systems or incorporating additional real-world factors, such as varying service rates or priority queues. Also, our work will focus on the use of the P/P/1 Series model to model other real-world systems. And the use of different simulation techniques to improve the accuracy.

References

1. M. Fischer and H. Cart, "A Method for Analyzing Congestion in Pareto and Related Queues," Telecommunications Review, vol. 10, pp. 15-28, 1999.
2. P. Argibay Losada, A. Suarez Gonzalez, C. Lopez Garcia, R. Rodriguez Rubio and J. Lopez Ardao, "On the simulation of queues with Pareto Service," in Proc. 17th European Simulation Multiconference, Nottingham, United Kingdom, 2003.
3. D. Gross, M. Fischer, D. Masi and J. Shorte, "Difficulties in Simulating Queues with Pareto Service," in Proceedings of the Winter Simulation Conference, New Orleans, LA, USA, 2003.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_16

4. Y. Koh and K. Kim, "Evaluation of Steady-State Probability of Pareto/M/1/K Experiencing Tail-Raising Effect," Lecture Notes in Computer Science, vol. 2720, pp. 561-570, 2003.
5. M. Fischer, D. Bevilacqua Masi, D. Gross and J. Shorte, "One-Parameter Pareto, Two-Parameter Pareto, Three Parameter Pareto: Is There a Modelling Difference?," Telecommunications Review, vol. 16, pp. 79-92, 2005.
6. A. Inmaculada, M. Meerschaert and A. Panorska, "Parameter Estimation for the Truncated Pareto Distribution," Journal of the American Statistical Association, vol. 101, no. 473, pp. 270-277, 2006.
7. H. Albrecher, J. C. Araujo-Acuna and J. Beirlant, "Tempered Pareto-type modelling using Weibull distributions," ASTIN Bulletin: The Journal of the IAA, vol. 51, no. 2, pp. 509-538, 2021.
8. B. Jian, J. Tan, N. Shroff and D. Towsley, "Heavy Tails in Queueing Systems: Impact of Parallelism on Tail Performance," Journal of Applied Probability, vol. 50, no. 1, pp. 127-150, 2013.
9. S. Asmussen, Applied Probability and Queues, New York: Springer, 2003.
10. T. Rolski, H. Schmidli, V. Schmidli and J. Teugels, Stochastic processes for insurance and finance, Chichester: Wiley, 1999.
11. A. Leon-Garcia, Probability, Statistics, and Random Processes for Electrical Engineering, New York: Pearson, 2008.
12. W. Whitt, "Whitt, W. (2000). The impact of a heavy-tailed service-time distribution upon the M/GI/s waiting-time distribution," Queueing Systems, vol. 36, no. 1, pp. 71-87, 2000.
13. S. I. Resnick, "Heavy tail modeling and teletraffic data," Ann Statist, vol. 25, no. 5, pp. 1805-2272, 1997.
14. J. Beran, Statistics for Long-Memory Processes, New York: Taylor & Francis, 1994.
15. W. Willinger and V. Paxson, "Where mathematics meets the Internet," Notices of the American Mathematical Society, vol. 45, pp. 961-970, 1998.
16. R. V. Ramirez-Velarde and R. M. Rodriguez-Dagnino, "From commodity computers to high-performance environents: scalability analysis using self-similarity, large deviations and heavy-tails," Concurrency and Computation: Practice and Experience, vol. 22, no. 11, pp. 1494-1515, 2010.
17. R. V. Ramirez-Velarde and R. M. Rodriguez-Dagnino, "A gamma fractal noise source model for variable bit rate video servers," Computer Communications, vol. 27, no. 18, pp. 1786-1798, 2004.
18. I. V. Zaliapin, Y. Y. Kagan and F. P. Schoenberg, "Approximating the Distribution of Pareto Sums," Pure and Applied Geophysics, vol. 162, no. 6-7, pp. 1187-1228, 2005.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_16