Modelling Extreme Uncertainty: Estimating Maximum Queue Size of Systems with Pareto Inter-Arrival Times and Pareto Service Times

Raul Ramirez-Velarde[1[0000-0001-7186-1914]], Cristobal Pareja-Flores[2[0000-0001-7739-0236]], 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. We propose an approach to modelling maximum queue sizes for heavy-tail interarrivals and service times. We derive models for high percentiles of queue length based on the principle that for subexponential distributions, large deviations of cumulative workload are dominated by single extreme observation. This allows the distribution of aggregate workload to be approximated through the distribution of the maximum service time, leading to tractable models for extreme queue length quantiles. We derive parametric models that require fewer fitted parameters than extreme value methods, including generalized extreme value (GEV), generalized Pareto (POT), and power-law tail models. Event-driven Monte Carlo simulations of heavy-tailed single-server queue are used to evaluate and compare proposed models. We show that the model called Par Sum Exp gives best results.

Keywords: Heavy-tailed distributions, Monte Carlo simulation, Performance modelling

1 Introduction

Classical queueing theory, developed primarily in the early 20th century, relies heavily on exponential or light-tailed distributions for inter-arrival and service times. However, empirical measurements of modern communication networks, computer systems, cloud computing and web traffic have consistently revealed that actual traffic patterns exhibit fundamentally different statistical properties characterized by heavy-tailed or long-tailed distributions [1], [2] which introduces significant uncertainty in system design and parameter selection.

Queueing systems with heavy-tailed inter-arrival and service time distributions arise in diverse domains, including telecommunications, web traffic, finance, and healthcare.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_28

Unlike classical light-tailed models, they exhibit fundamentally different behaviour: extreme events occur with non-negligible probability, leading to large queue growth and making performance metrics difficult to estimate. Such dynamics are also observed in modern systems such as cloud and distributed computing (due to straggler effects), as well as in cybersecurity, transportation, and large-scale AI workloads, where rare events dominate system performance.

In [3], we presented the P/P/1 queueing 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. We showed that the model called P/P/1 Series was the one that best approximated the expected values of the queue size, server utilization and system wait time.

Simulation results showed that queue size in these systems has such high variance that the maximum queue size is usually much larger than the average queue size. In this paper we study the challenge of estimating high percentiles—the 90th, 95th, or 99th percentiles—of the maximum queue size over a given time horizon. These quantiles are critical for capacity planning, quality-of-service guarantees, and risk management. However, the slow polynomial decay of heavy-tailed distributions renders naive simulation inefficient and classical asymptotic approximations inaccurate for finite sample sizes.

Interestingly, simulations show that the maximum queue size is highly dependent on the standard deviation of the service time, confirming that large deviations on queue size are dominated by a single large jump rather than many moderate deviations as presented by [4], [5]. We show this result in Fig. 1.

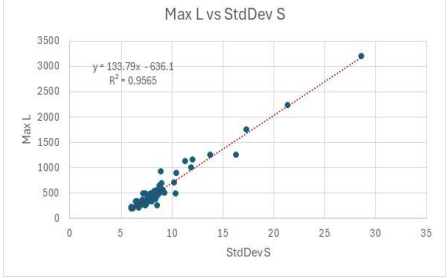

Fig. 1. Maximum queue size depends on service time standard deviation for Heavy-Tailed systems.

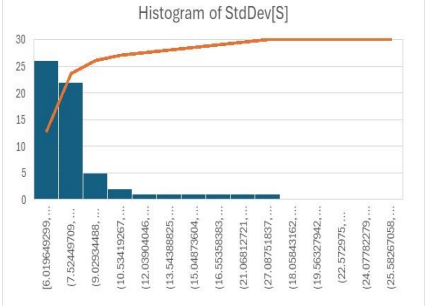

Fig. 2. Histogram of service time standard deviation

We also find that the standard deviation of the service time, that in our simulated system can potentially be infinite as result of the Pareto I random variables, is also a long-tailed random variable as shown in Fig. 2. Since the standard deviation shown in

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_28

Fig. 2 is empirically determined and highly variable it complicates modelling maximum queue size.

From an uncertainty quantification (UQ) perspective, this problem involves estimating extreme system responses under heavy-tailed uncertainty, where classical moment-based measures are insufficient and behaviour is driven by rare, high-impact events. Thus, estimating high quantiles of a queue becomes a tail-focused UQ problem involving both stochastic and model uncertainty.

In this paper we will discuss a queueing 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 queueing 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:

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

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

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.

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

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

In the rest of the paper, we develop the **Par Sum Exp** and **Pareto Max** models to approximate the high percentiles of the maximum queue size and compare them with known models for extreme value queue sizes.

2 Previous Work

Estimating high quantiles of maximum queue size in heavy-tailed systems presents several interrelated challenges. For example, [4] and [6] show that standard Monte Carlo simulation requires large sample sizes to observe sufficient tail events. For regularly

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_28

varying distributions with index α, the variance of tail probability estimators can be infinite when $\alpha \leq 2$. Whereas [7] show that asymptotic results provide theoretical guidance but their accuracy for finite buffer sizes or finite time horizons is often unclear. This paper also studies Extreme Value Theory (EVT) showing that for heavy-tailed distributions in the Fréchet domain of attraction, the maximum of n independent observations converge (after appropriate normalization) to a Fréchet distribution with shape parameter α. Also, the paper shows the generalized Pareto distribution (GPD) to approximate far-end-tail behaviour of workload in M/G/1 queues with Pareto service times is used. The GPD approximation, valid for exceedances over a high threshold, enables estimation of extreme quantiles through the peaks-over-threshold method.

Single-server queues with heavy-tailed service times are studied in [4] [6] [8] [9], providing insights into the "principle of one large jump" that governs heavy-tailed queue behaviour.

Asymptotic analysis has been carried out in [7] [6], showing that a key limitation of asymptotic methods is that they provide approximations valid only for very large buffer sizes. As Morozov et al. note, "the approximation is based on the asymptotic equivalence between the excess distribution over a high threshold and the generalized Pareto distribution" [7], requiring careful threshold selection in practice.

In [4], [5] and [10] apply large deviations theory demonstrating the principle of One Large Jump: For heavy-tailed random walks, large deviations are dominated by a single large jump rather than many moderate deviations. Blanchet et al. establishes uniform large deviations results for heavy-tailed single-server queues under heavy traffic, providing decay rates that are uniform over buffer levels

Fischer and Cart [11] 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.

Fischer et al. [12] 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. [13] derived estimators for the truncated Pareto distribution.

Recently, Jiang et al. [14] 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 [3] we derived the following models for total system time (Table 1):

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_28

Even though these models can approximate the mean of the queue size and sojourn time, they do not do well approximating the higher percentiles, which is the aim of this paper.

Table 1. Models derived in [3]

$f(t) = \frac{f(z_1(t))}{(1-\rho)^{1+H}}$ where $z_1(t)$ has Pareto pdf	$f(t) = \beta t^{-\beta-1}\left[\frac{1}{(1-\rho)} + \frac{\beta(\beta+1)}{t(\beta-1)}\left(\frac{\rho+1}{(1-\rho)^2}\right)\right]$
P/P/1 Frac1	P/P/1 Par Sum

3 Heavy-Tail Fits to Queue Length, Maximum Queue Length and High Queue Size Percentiles

3.1 Busy-Period Mechanism of Extreme Queue Length

Extreme queue lengths in heavy-tailed systems arise primarily through the behaviour of the busy period, defined as the interval during which the server remains continuously occupied. In Pareto service times, this implies that large deviations of workload are dominated by a single extreme service demand. This is called the principle of one large jump. An unusually long service time can therefore initiate a prolonged busy period during which arrivals accumulate and the queue grows substantially. Thus, the maximum queue length is largely determined by the largest busy period and its associated workload. We now proceed to characterize this busy period behaviour.

3.2 Pareto Distribution

A natural first approach to modelling extreme congestion in heavy-tailed queues is to fit parametric heavy-tailed distributions directly to the observed queue length L or to the maximum queue length L_{max}. The most common choice is the Pareto distribution, which was described before, motivated by the empirical observation of power-law decay in the upper tail.

When applied to L or L_{max}, the shape parameter α governs tail heaviness, while x_m sets the effective lower bound. This approach is attractive because of its simplicity and interpretability percentiles are available in closed form,

$$q_p = x_m(1-p)^{\left(-\frac{1}{\alpha}\right)}$$

However, global Pareto fits often overestimate extreme percentiles when fitted over the entire support. Therefore, we need to turn to statistical study of extreme values to search tools to approximate high percentiles in high-variance queueing systems.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_28

3.3 Extreme Value Theory for Maximum Queue Length

Extreme Value Theory (EVT) provides a principled framework for modelling the maximum queue size. Let $M_n = \max\{X_1, \ldots, X_n\}$ denote block maxima. Under broad conditions,

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) \to G(x),$$

where G is the Generalized Extreme Value (GEV) distribution [15],

$$G(x) = e^{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}}, 1 + \xi\left(\frac{x-\mu}{\sigma}\right) > 0.$$

The shape parameter ξ determines the tail class:

- ξ>0: Fréchet (heavy-tailed),
- ξ=0: Gumbel,
- ξ<0: Weibull.

In heavy-tailed queueing systems, the Fréchet class ($\xi > 0$) is most relevant. EVT block-maxima methods are asymptotically justified but can be data-inefficient when few maxima are available. In practice, GEV is fitted to block maxima.

3.4 Peaks-Over-Threshold (POT) and the Generalized Pareto Distribution

For estimating high percentiles (e.g., 90%, 95%, 99% and beyond), the Peaks-Over-Threshold (POT) approach is generally used. Rather than discarding data below block maxima, POT models all exceedance above a high threshold u [16].

Let $Y = X - u \mid X > u$, denote threshold exceedances. EVT shows that, for sufficiently high u,

$$P(Y \leq y) \approx 1 - \left(1 + \xi\frac{y}{\beta}\right)^{-\frac{1}{\xi}}$$

with density

$$f(y) = \left(\frac{1}{\beta}\right)\left(1 + \xi\frac{y}{\beta}\right)^{-1-\frac{1}{\xi}}, y \geq 0.$$

Here ξ is the tail index and β>0 a scale parameter. The unconditional tail is approximated by

$$P(X > x) \approx P(X > u)\left(1 + \xi\frac{x-u}{\beta}\right)^{-\frac{1}{\xi}},$$

yielding quantiles

$$q_p = u + \left(\frac{\beta}{\xi}\right)\left[\left(\frac{P(X>u)}{1-p}\right)^{\xi} - 1\right].$$

POT/GPD is particularly effective for estimating high percentiles of Lmax, since they focus exclusively on the tail, avoid contamination from the bulk, and provide

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_28

accurate estimates of high percentiles even with moderate sample sizes. Since in our simulation model we have $\alpha, \beta < 2$, which leads to volatile empirical standard deviations, given that POT/GPD is very sensitive to threshold selection, we chose the 80% percentile as threshold and use it to estimate 90% and 95% percentiles. The stability of the chosen threshold was verified using mean-excess and stability plots not shown in this paper for lack of space. GEV and GPD used in this paper are standard tools in UQ for modelling extremes, particularly in reliability analysis, risk assessment, and rare-event simulation.

3.5 Zipf Probability distribution

For discrete queue lengths, Zipf distributions can be used to extrapolate extreme percentiles when the tail of Lmax exhibits power-law scaling above a high threshold, indicating scale-free congestion dynamics [17].

For integer-valued $k \geq k_{\min}$:

$$P(X = k) = \frac{k^{-\alpha}}{\zeta(\alpha, k_{\min})}$$

where $\zeta(\cdot)$is the Hurwitz zeta function.

These well-established models are compared to the models we have derived: the Par Sum Exp and Pareto Max, which we now present.

3.6 Our Contribution: Models for Estimating High Percentiles in Queue length by Approximating the Sum by the Maximum of Pareto Random Variables

Now we derive two models called Par Sum Exp and Pareto Max. Par Sum Exp is derived by substituting the sum of Pareto variables in the Generalized Central Limit Theorem by the maximum observed value and then approximating using an exponential function. The model called Pareto Max is derived by converging $P[S_n < x]$ in a summatory.

Sum of Pareto Random Variables.

The addition of random variables with Pareto probability distribution, when suitably normalized, approaches a well-defined limiting distribution which depends on α or β [18].

The Generalized Central Limit Theorem (GCLT) states that the properly normalized sum $S_n = \sum_i^N z_i$ of many i.i.d. Pareto random variables with common distribution may be approximated by a stable distribution. Assuming β to be the tail index and $B = 1$ to simplify (we will carry out scaling later):

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_28

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

where,

$$b_n = \mathrm{n}\mu = \frac{n\beta}{\beta-1} \text{ and } a_n = n^{\frac{1}{\beta}}$$

Approximation to Exponential.

As observed in [19], in many cases of sequences of Pareto random variables, the largest observation M_n, has the same order of magnitude as the entire sum S_n, at least for the upper quantiles. Because Pareto distributions are subexponential, extreme realizations of the sum are asymptotically generated by a single large observation. Therefore, when modelling high quantiles of S_n (or queue extremes), the centring term b_n becomes negligible relative to extreme deviations, and the tail behaviour of S_n can be approximated by that of the maximum. Thus, for subexponentials in Eq. (1) we have $P[S_n > x] \approx P[M_n < x]$. We can now use the well-known distribution for the maximum:

$$P[S_n - b_n < x] = \prod_1^n P(X_i < x) = \left(1 - x^{-\beta}\right)^n \approx P[M_n < x]$$

Thus

$$\lim_{n\to\infty} P\left[\frac{S_n}{n^{\frac{1}{\beta}}} < x\right] = \lim_{n\to\infty} P\left[\frac{M_n}{n^{\frac{1}{\beta}}} < x\right] = \left(1 - \left(n^{\frac{1}{\beta}}x\right)^{-\beta}\right)^n = \left(1 - \frac{x^{-\beta}}{n}\right)^n = e^{-x^{-\beta}}$$

And the pdf is:

$$f(x) = \frac{d}{dx}\left(e^{-x^{-\beta}}\right) = \quad \beta x^{-\beta-1}e^{-x^{-\beta}}$$

$$f(t) = \sum_{n=0}^{\infty} p_n f_n(t) = \sum_{n=0}^{\infty} \rho^n(1-\rho)\alpha t^{-\beta-1}e^{-t^{-\beta}} = \beta t^{-\beta-1}e^{-t^{-\beta}}$$

Adding a normalization factor, if $C\int_1^\infty \beta\ x^{-\beta-1}e^{-x^{-\beta}}dx = 1$, then $C = \frac{1}{1-\frac{1}{e}} = \frac{e}{e-1}$

$$f(t) = \left(\frac{e}{e-1}\right)\beta t^{-\beta-1}e^{-t^{-\beta}}, t \geq 1, \beta > 1 \quad (2)$$

We call this the **Par Sum Exp** model. With scaling, the normalization factor $\frac{1}{1-e^{-1}}$ turns $\frac{1}{1-e^{-B}}$ which vanishes for large B. The final model is Eq. (3):

$$f(t) = B\beta t^{-\beta-1}e^{-\mathrm{B}t^{-\beta}}, t \geq A, \boldsymbol{\alpha} > 1 \quad (3)$$

$$F(T) = e^{-\mathrm{B}T^{-\beta}}, T > 0$$

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_28

Modelling Highest Observed Value.

If we compare the sum of Pareto random variables directly to its highest value,

$$P[S_n < x] = \left(1 - x^{-\beta}\right)^n \approx P[M_n < x]$$

Then, the pdf is,

$$P[S_n = x] \approx P[M_n = x] = n\beta x^{-\beta-1}\left(1 - x^{-\beta}\right)^{n-1}$$

And therefore:

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

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

$$f(t) = \frac{(1-\rho)\beta t^{-\beta-1}}{\left(1-\rho(1-t^{-\beta})\right)^2}, t > 1 \tag{4}$$

We call this the **Pareto Max** model. For this model

$$F[T] = \int_1^{\infty} \frac{(1-\rho)\beta t^{-\beta-1}}{\left(1-\rho+\rho t^{-\beta}\right)^2} dT = \frac{1-\rho}{1-\rho+\rho t^{-\beta}}, t > 0$$

After scaling we get the final model in Eq. (5):

$$f(t) = \frac{(1-\rho)\beta B^{\beta} t^{-\beta-1}}{\left(1-\rho\left(1-(t/B)^{-\beta}\right)\right)^2}, t > \mathrm{B}, \beta > 1 \tag{5}$$

$$F(T) = \frac{1-\rho}{1-\rho+(T/B)^{-\beta}}$$

4 Simulation Setup and Empirical Results for Maximum Queue Length

To determine which models allows us to determine best queue operational parameters under congestion, such as maximum queue capacity, a simulation was carried out with the following parameters. The inter-arrival time of each job is Pareto I probability distribution with shape parameter α, and location parameter A, $f(t) = \alpha A^{\alpha} t^{-\alpha-1}$. For the simulation, $\alpha = 1.7$ and $A = 1.77059$, which makes the mean inter-arrival time $E[arr\ time] = \bar{A} = 4.3$. 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, $g(t) = \beta B^{\beta} t^{-\beta-1}$. For the simulation, $\beta = 1.8$ and $B = 1.51111$, which makes the mean inter-arrival time $E[serv\ time] = \bar{S} = 3.4$.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_28

We carry out an event-driven simulation for a single-server queue with FCFS (First-Come First-Served) service discipline and infinite buffer. To generate the Pareto variables, we use the well-known procedure $X = x_m U^{-\frac{1}{\alpha}}$ with $U = Uniform(0,1)$. 300,000 events were generated in each simulation. As warm-up, the first 20% of events are discarded. 80 simulation replicas were created.

4.1 Key empirical results (means across 80 reps, 95% CI)

These are the average simulation results:

- U: 0.7893 [0.7862, 0.7925]
- E[S]: 3.4024 [3.3972,3.4077]
- Emp SD[S]: 10.2901 [9.1144,11.4659]
- W: 93.8440 [70.6348,117.0533]
- W_q: 90.4416 [67.2358,113.6474]
- L: 21.8206 [16.4057,27.2355]
- $L = \lambda W$: 21.8051 [16.3899,27.2204]
- L_q: 21.0313 [15.6174,26.4451]

One important observation is that for Pareto with shape $\alpha < 2$, the theoretical variance is infinite, so empirical standard deviations estimations can be volatile across replications. That explains the wide CIs for L, W, W_q.

The queue operates in a stable regime since $E[S] < E[A]$, giving traffic intensity $\rho \approx 0.79$, which ensures the existence of a stationary workload distribution for the single-server system. However, because the Pareto inter-arrival and service distributions have shape parameters $1 < \alpha, \beta < 2$, variance is infinite and convergence to stationarity can be slow, with large fluctuations caused by very long service times. To reduce transient effects, the simulation uses 300,000 events per replication with the first 20% discarded as warm-up. The results reflect a stable but highly-variance system which is characteristic of heavy-tailed queueing systems where extreme congestion events can dominate the entire system performance.

To create a baseline, we present Table 2 in which we compare mean simulation results to well-known models such as M/M/1, G/G/1 (Kingman) and M/G/1 (Pollaczek–Khinchine).

Table 2. Mean Simulation results vs well-known models

Model	ρ	W	Wq	L
Empirical	0.7893	93.8440	90.4416	21.8206
M/M/1	0.7893	16.1494	12.7470	3.7464
G/G/1 Kingman	0.7893	113.6830	110.2805	26.3728
M/G/1 P-K (empirical E[σ^2])	0.7893	83.2486	79.8461	19.3125

The Kingman and Pollaczek–Khinchine give reasonable approximations to the mean value for queue length but fall far too short for maximum queue length (results not shown). This is because, as shown in Fig. 3, the density of the Lmax is strongly right-skewed and heavy-tailed. The long right tail (values above ~2000–3000) visually

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_28

confirms the extreme-value behaviour. We now proceed to fit the models described in section 3.

4.2 Fitting High-Percentiles to Well-known Models

We fit the extreme value models presented in section 3: GEV block maxima, threshold GPD and Zipf. The results of the model fit for the tail of queue length, above 80th percentile, based on our simulations, are shown in Table 3, where we show the fitted parameters. And in Fig. 4, where we show the CCDF plots.

Table 3. Parameters of the maximum queue length $L_{\max}$. The empirical CCDF from simulation is compared with fitted extreme-value models: block-maxima GEV, threshold-based GPD (POT, u = 80th percentile), thresholded Pareto, and discrete Zipf. Differences in tail curvature highlight model-dependent extrapolation of extreme congestion.

Model	Threshold	Shape	Location	Scale
GPD (POT)	u = 80th percentile ≈ 1040	$\xi \approx 0.6043$	0 (fixed for exceedances)	$\beta \approx 508$
GEV (Block maxima)	Implicit (block maxima)	$\xi \approx 0.55$ (Fréchet)	$\mu \approx 419$	$\sigma \approx 241$
Zipf (discrete tail)	$k \geq k_{min} = ceil(u80) \approx 1041$	$\alpha \approx 2.98$	N/A	N/A

4.3 Approximating Queue Lmax Higher-Percentiles: GEV, POT/GPD and Zipf

We now evaluate the predictive accuracy of each model for high-percentile estimation. Simulation results yielded the following results for Lmax high percentiles: u90%=1,535, 90%CI=[1,096-2,422], u95%=2,412, 95%CI=[1,634-2,587]. With the fitted models, we proceed to compare the models against simulations for 90 and 95 percentiles. The results are shown in Table 4.

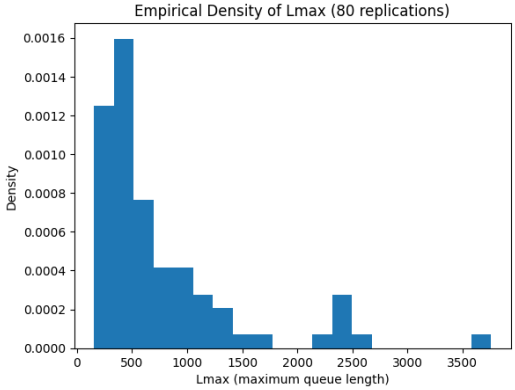

Fig. 3. Empirical density of the maximum queue length Lmax obtained from 80 independent simulation replications. The pronounced right skew and sparse extreme observations highlight the heavy-tailed nature of queue maxima.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_28

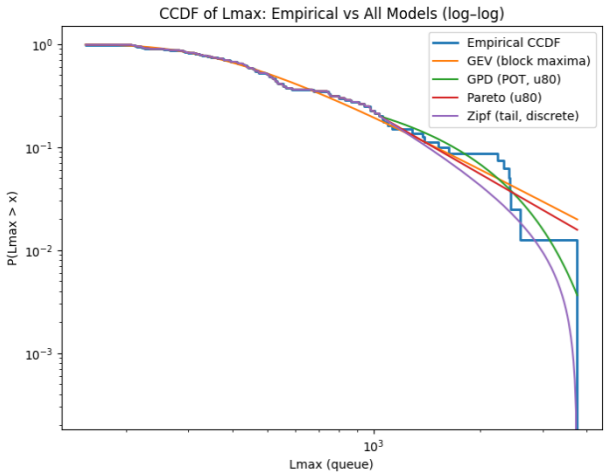

Fig. 4. Comparison of fitted extreme-value models for the maximum queue length Lmax. The GPD (POT) model is fitted above the 80th-percentile threshold, the GEV model is fitted to block maxima across replications, and Zipf model captures discrete power-law tail behaviour.

Table 4. Estimated values of ϕ such that P(Lmax < ϕ) equals 0.90 and 0.95, based on empirical simulation results and fitted extreme-value models.

Model	P(Lmax < ϕ) = 0.90	P(Lmax < ϕ) = 0.95
Empirical (simulation)	1535	2412
GPD (POT, u = 80th percentile)	1681	2238
GEV (block maxima)	**1494**	**2232**
Zipf (tail, k ≥ ceil(u80))	1477	2097

From Table 4 and Fig. 4 we find that even though all estimations fall within confidence intervals, the best fit is GEV (block-maxima). As Fréchet was chosen as best fit, we identify a heavy-tailed extreme-value regime (ξ>0). Whereas POT/GPD and Zipf are tail approximations, GEV models the entire extremal mechanism. POT overshoots at 90% percentiles and Zipf underestimates the far tail.

4.4 Approximating Queue Lmax Higher-Percentiles: Par Sum Exp and Pareto Max models

The Par Sum Exp and Pareto Max models derived in section 3.5 are models for total system time. We will find approximations for high percentiles of Lmax using Little's Law. Although Little's Law is exact for steady-state means, a common approximation for high-percentile queue lengths is to scale delay quantiles by the effective arrival rate, $q_p(L) \approx \lambda q_p(W)$. This heuristic scaling is widely used in heavy-tail queue approximations. It can be used as an approximation because in heavy-tailed systems, extreme queue lengths are typically generated by prolonged busy periods or unusually large

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_28

service requirements, during which the queue length evolves approximately as the cumulative number of arrivals over an extended sojourn period, yielding the sample-path scaling $L_{max} \approx \lambda T_{max}$ [16].

The results of the model fit for the tail of queue length based on simulation are shown in Table 5, where we show the fitted parameters and Fig. 3, where we show the CCDF plot.

Table 5. Fitted parameters for Par Sum Exp and Pareto Max for estimating the maximum queue length Lmax.

Model	B_{fit}	β_{fit}
Par Sum Exp	343,160	1.7
Pareto Max	962.121	1.7

Table 6. Estimated values of ϕ such that P(Lmax < ϕ) equals 0.90 and 0.95, based on empirical simulation results and fitted extreme-value models. Par Sum Exp model gives closest fit.

Model	P(Lmax < ϕ) = 0.90	P(Lmax < ϕ) = 0.95
Empirical (simulation)	1535	2412
Par Sum Exp	**1572**	**2401**
Pareto Max	1718	2693
GEV (block maxima)	1494	2232

As before, we compare Pareto Max and Par Sum Exp tail estimations to simulations for 90 and 95 percentiles and the GEV model which was the closest to simulation. In the case of Pareto Max model, the model was fit to exceedances above 80th percentile of Lmax which is 1040.2. The results are shown in Table 6.

Par Sum Exp model has two parameters. $\beta_{fit} = 1.7$ which is very close to parameters α and β from simulation parameters, and parameter B_{fit}. This represents an advantage over the GEV, POT/GPD and Zipf models for which all parameters need to be fit after simulation. Nevertheless, parameter B_{fit} cannot be easily obtained from simulation parameters as it's not directly linked to those but it's a much-scaled parameter linked to the tail behaviour of busy periods. This parameter must be fit at least once. Scaling laws for parameter B_{fit} are currently being researched.

5 Conclusions

In this paper we investigated the maximum queue length Lmax in an infinite memory single server queue with heavy-tailed interarrival and service times using Pareto probability distribution. We derived two models using the GCLT and used Monte Carlo simulations to obtain an empirical distribution to test the accuracy of the models. We also compared the derived models against Pareto tail, Generalized Pareto (POT), Generalized Extreme Value (block maxima), and Zipf distributions. We found that Lmax

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_28

exhibits heavy-tailed behaviour. Empirical diagnostics, including CCDF plots and extreme-value fitting, indicate that rare busy episodes dominate the upper tail.

We propose a reduced-parameter extreme-value modelling approach for UQ of rare congestion events, the Par Sum Exp model, which provides the most accurate approximation of high percentiles of Lmax. Specifically, the estimates obtained for the 90% and 95% percentiles were closer to empirical values than other models. This improvement was especially noticeable in the far tail, where classical EVT models tended either to overestimate or underestimate extreme quantiles. The accuracy of the Par Sum Exp model on high queue length percentiles appears to derive from its direct representation of the distribution of extreme system times rather than queue length. By mapping maximum system time to maximum queue length through a scaling relationship derived from Little's law, it captures the dominant mechanism generating extreme congestion, the occurrence of unusually long busy periods.

Future work will focus on scaling laws for deriving parameter A, which we know is determined by high quantiles of observed busy periods maxima. We will also investigate the sensitivity of the fitted parameters to different traffic intensities and tail indexes and extend the work to other queueing settings and network models.

References

[1] K. Park and W. Willinger, "Self-Similar Network Traffic: An Overview.," in *Self-Similar Network Traffic and Performance Evaluation* , Wiley, 2000, pp. 1-38.

[2] P. R. Jelenković, "Asymptotic analysis of queues with subexponential arrival processes," in *Self-Similar Network Traffic and Performance Evaluation*, Wiley, 2000, pp. 249-268.

[3] R. Ramirez-Velarde, C. Pareja-Flores, N. Hernandez-Gress and L. Hervert-Escobar, "Ramirez-Velarde, Raul, et al. "Modelling Extreme Uncertainty: Queues with Pareto Inter-arrival Times and Pareto Service Times," in *Lecture Notes in Computer Science*, vol. 15912, M. B. A. Z. Y. Paszynski, Ed., Singapore, Springer Nature Switzerland, 2025, pp. 222-235.

[4] S. Asmussen, K. Binswanger and B. Højgaard, "Rare Events Simulation for Heavy-Tailed Distributions," *Bernoulli,* p. 303–322, 2000.

[5] J. Blanchet and H. Lam, "Uniform large deviations for heavy-tailed queues under heavy traffic," *Bull. of the Mex. Math. Soc,* vol. 3, no. 19, pp. 183-199, 2013.

[6] N. K. Boots and P. Shahabuddin, "Simulating tail probabilities in GI/GI/1 queues and insurance risk processes with subexponential distributions," *ACM SIGMETRICS Performance Evaluation Review,* vol. 29, no. 3, pp. 38-39, 2001.

[7] E. V. Morozov, I. V. Peshkova and A. S. Rumyantsev, "Far-End-Tail Estimation of Queueing System Performance," *Journal of Mathematical Sciences,* vol. 248, no. 1, pp. 80-91, 2020.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_28

[8] P. Ramirez, R. Lillo and M. P. Wiper, "Bayesian analysis of a queueing system with a long-tailed arrival process," *Communications in Statistics—Simulation and Computation,* vol. 37, no. 4, pp. 697-712, 2008.

[9] C. M. Ramsay, "Exact waiting time and queue size distributions for equilibrium M/G/1 queues with Pareto service," *Queueing Systems,* vol. 57, no. 4, pp. 147-155, 2007.

[10] H. BERNHARD and B. DAS, "Heavy-tailed random walks, buffered queues," *Bernoulli,* vol. 26, no. 1, pp. 61-92, 2020.

[11] M. Fischer and H. Cart, "A Method for Analyzing Congestion in Pareto and Related Queues," *Telecommunications Review,* vol. 10, pp. 15-28, 1999.

[12] 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.

[13] 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.

[14] 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.

[15] S. Coles, An Introduction to Statistical Modeling of Extreme Values, London: Springer, 2001.

[16] P. Embrechts, C. Kluppelberg and T. Mikosch, Modelling Extremal Events for Insurance and Finance, London: Springer, 1997.

[17] M. E. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary physics,* vol. 46, no. 5, pp. 323-351, 2005.

[18] 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.

[19] W. Feller, An Introduction to Probability Theory and Its Applications, 2 ed., vol. 2, New York: Wiley, 1971.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_28