Prediction of Rare Latency Events

Leonard Scheerer, Max Helm*, Benedikt Jaeger*
*Chair of Network Architectures and Services
School of Computation, Information and Technology, Technical University of Munich, Germany
Email: leonard.scheerer@tum.de, helm@net.in.tum.de, jaeger@net.in.tum.de

***Abstract*—An increasing number of safety critical applications rely on networks and their latency bounds. Thus, providing estimates of worst case latencies is crucial for ensuring the quality-of-service requirements of such systems. This paper presents an approach to acquire these estimates using Extreme Value Theory, a statistical method that bases its predictions on models derived from real-world measurements. In particular, we analyze the latency values of a single flow of a virtualized network topology. Additionally, we compare the approach with alternative methods and present current applications of Extreme Value Theory in the networking area. We consider Extreme Value Theory a powerful tool for estimating the tail-end of network latency distributions. In many cases, it outperforms alternative approaches with similar goals.**

***Index Terms*—extreme value theory, latency measurement**

1. Introduction

Safety-related systems, such as those in medical, aerospace and security fields are known to be *time-critical*. That is to say, missing a deadline can have drastic consequences on the environment, on equipment or even human lives. More and more of these critical systems are now distributed and therefore rely on networks. One prominent example are vehicular networks that need to exchange real-time status updates between individual vehicles. For this reason, ultra-reliable and low-latency communication is a key service type in the next generation (6G) communication systems. Realizing networks with low end-to-end latency guarantees represents a major challenge facing 6G [1] [2].

This paper utilizes Extreme Value Theory (EVT) to model the extreme latency events of a single network flow. The raw latency data originate from an experiment conducted by Wiedner et al. The authors make use of real networking hardware to create a virtualized network topology on a single physical host. The detailed measurement setup is described in [3]. All code used to visualize and analyze the described data is available online at https://github.com/leonardscheerer/rare-latency-events.

The remainder of this paper is structured as follows: Section 2 introduces the theoretical background of Extreme Value Theory. In Section 3, various approaches to the prediction of extreme events are discussed. Section 4 applies EVT to real-world latency data. After presenting other applications of EVT in the networking area in Section 5, some concluding remarks are made in Section 6.

2. Background

This section introduces basic concepts and results in the field of Extreme Value Theory.

2.1. Extreme Value Theory

EVT is a branch of probability theory with the aim of describing the stochastic behavior of extremes, i. e., events on exceptionally large or small scales. The derived models are a solid theoretical basis for predicting the occurrence and extent of rare events and enable extrapolations to unobserved levels. This goal is unique amongst the statistical disciplines, as commonly, the objective is to model the ordinary, rather than the unordinary [4].

EVT had its first applications in the 1950's in the area of civil engineering, in which the frequency and magnitude of natural phenomena such as floods or earthquakes can be crucial information for the design of structural components of buildings [4]. In recent years, EVT has gained considerable traction in various other fields such as the social sciences, the medical profession, economics and even astronomy [5].

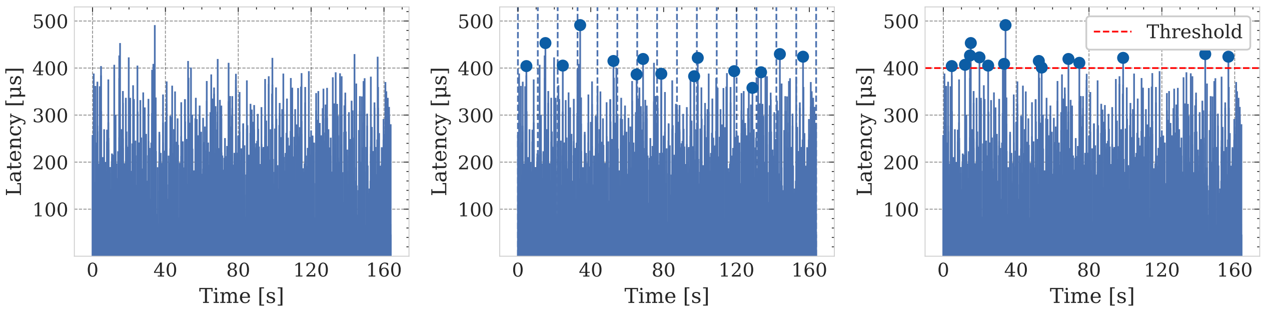
Conforming to the *extreme value paradigm*, the extrapolations used to predict extreme values are based on asymptotic arguments, i. e., on using mathematical limits as finite-level approximations. As a consequence, the results of EVT cannot be regarded as exact when applied to finite samples [4, Chapter 1].

Before trying to model the behavior of extreme events, it is first necessary to define what constitutes an extreme occurrence. There are two main ways to define such events, leading to two alternative methods of mathematical modeling [6]. Both approaches, termed the Block Maxima approach and the Peaks over Threshold approach, are used in practice and are briefly explained in the following two sections.

2.2. Block Maxima Approach

In the Block Maxima approach the observation period is divided into n non-overlapping blocks of equal size. The maximum of each block is deemed to be an extreme value. Figure 1a shows the raw latency data and Figure 1b highlights the extreme values when applying the Block Maxima approach with $n = 15$. When interested in exceptionally small events, extreme values are derived analogously with minima instead of maxima.

doi: 10.2313/NET-2023-11-1_08



(a) Raw data

(b) Block Maxima approach

(c) PoT approach

Figure 1: Latency data of a single network flow and their extreme values according to the Block Maxima and PoT approach.

The Block Maxima are subsequently modeled using the *Fisher-Tippett-Gnedenko* theorem. It states that under certain assumptions, mainly that these maxima are samples of independent and identically distributed random variables, the distribution of the maxima converge to one of three probability distributions: the Gumbel Distribution, the Fréchet Distribution or the Weibull Distribution. The distributions are also referred to as the extreme value type 1, type 2 and type 3 distributions, respectively [6].

All three distributions can be represented with a single distribution, the Generalized Extreme Value Distribution (GEV). The cumulative distribution function $F(x;\mu,\sigma,\xi)$ of the GEV is given by (1). It measures the probability that the random variable will take a value less than or equal to x.

$$F(x;\mu,\sigma,\xi) = \exp(-\max\{1+\xi\frac{x-\mu}{\sigma},0\}^{-\frac{1}{\xi}}) \qquad (1)$$

ξ is termed the extreme value index and maps to the aforementioned three distributions. To derive a robust model, ξ as well as the scale parameter σ and location parameter μ have to be fitted to the observed data using a suitable estimation method [5, Chapter 4].

2.3. Peaks over Threshold Approach

In the Peaks over Threshold (PoT) approach, we specify some threshold u. All values that exceed this threshold are considered extreme values. Figure 1a shows the raw latency data and Figure 1c highlights the extreme values when applying the PoT approach with $u = 400\,\mu\text{s}$. When interested in exceptionally small events, extreme values are derived analogously with threshold subceedances instead of threshold exceedances.

The obtained excesses, i. e., the amounts that the peaks exceed the threshold, follow the *Pickands-Balkema-De Haan* theorem. It states that with a sufficiently high threshold and under similar conditions to the Fisher-Tippett-Gnedenko theorem, the values of the excesses will converge to the Generalized Pareto Distribution (GPD). The cumulative distribution function $G(x;\sigma,\xi)$ of the GPD is given by

$$G(x;\sigma,\xi) = 1-\max\{1+\frac{\xi x}{\sigma},0\}^{-\frac{1}{\xi}} \qquad (2)$$

Similarly to (1), ξ determines the shape and σ the scaling of the distribution. Equation (2) does not contain a location parameter, as it is fixed to the previously chosen threshold [5, Chapter 4] [7].

3. Analysis

In this section we argue that dedicated methods are necessary to accurately predict extreme behavior. Afterwards we analyze selected modeling approaches for extremes, particularly in the context of latency events.

3.1. Traditional Methods

Traditional parametric statistical methods are ill-suited to model values at the very tail-end of a distribution. These statistical methods typically aim to be a good fit for a large proportion of the observed data, thus, accurately representing regions where most of the data fall. However, this comes at the price of a worse fit in the tails and therefore justifies the usage of dedicated approaches. Nevertheless, separate methods for modeling extreme values such as EVT are not needed – and possibly not suited – for estimating values that make up the top 10 %, 5 % or perhaps even 1 %. Rather, these methods focus on *extreme* (e. g. 0.1 %) outliers [8].

3.2. Modeling Approaches

Machine Learning. One possible approach to predicting rare latency events is machine learning. This method has emerged as a fast and reliable means to data-driven predictions. Wambura et al. [9] propose using a deep neural network for real-time stochastic extreme events prediction. The authors empirically confirm that their approach is fast and accurate. Their experimental results also suggest superior performance compared to well-known prediction methods. Nevertheless, application of deep learning to latency events is not straightforward and requires a large number of training samples due to a slow convergence in the training phase. Low learning efficiency can be combatted by the integration of knowledge of the environment such as estimated packet loss [10].

Network Calculus. Another possible approach to modeling and preventing high latencies are provable worst case upper latency bounds. This can be achieved via network calculus, a system theory for communication networks. The theoretical framework is built on the non-traditional min-plus and max-plus algebras [11]. Network Calculus and similar formal methods work on simplified assumptions and do not incorporate environmental events such as electromagnetic interferences. Additionally, the derived bounds are not tight [12].

Extreme Value Theory. The remainder of this paper focuses on applying EVT to the prediction of rare latency events. As discussed in Section 2, both the Block Maxima as well as the PoT approach are used in practice. The Block Maxima approach particularly lends itself to modeling data sets that already consist of block maxima, e. g. records of annual maximum sea-levels. In this case, the approach can incorporate all of the measurements and formulate accurate predictions. In practice, however, it is uncommon to have data of this form and following the Block Maxima approach may entail a wastage of information. Suppose, for example, that there are several recorded high events during one block. The block maximum takes precedence over all other events of a block. They are ignored as a consequence of this approach – even if they were noteworthy in the sense that they exceed the Block Maxima of other blocks [4]. This is not the case for the threshold exceedances in the PoT method. For this reason, the PoT approach is considered to utilize extreme observations more efficiently than the Block Maxima approach [13]. Note that both the maxima as well as the threshold excesses are assumed to be independent of each other in the respective theorems described in Section 2.2 and 2.3. Whereas this is often a reasonable assumption in the Block Maxima approach, as they are spaced out by construction, this cannot be said for the PoT approach. Thus, the PoT approach is often used in combination with special techniques such as declustering that aim to ensure that the data are independent [4, Chapter 5]. Based on the aforementioned benefits and drawbacks of the Block Maxima and PoT approaches, we deem the PoT technique to be more suitable for the characterization of the tail distribution of latencies.

4. Rare Latency Estimation

This section applies Extreme Value Theory to the real-world latency data introduced in Section 1 in order to predict rare latency events. We use the PoT approach described in Section 2.3 and assume that the raw data consist of a sequence of independent and identically distributed measurements.

4.1. Inference

To fit the generalized Pareto family to the observations, we first select a suitable threshold and subsequently estimate the characterizing scale parameter σ and shape parameter ξ.

Threshold Selection. Threshold selection is a crucial part of extreme value analysis following the Peaks over Threshold method. Too low a threshold is likely to lead to the Generalized Pareto Distribution not being a good fit for the threshold excesses, as a sufficiently high threshold is a requirement of the Pickands-Balkema-De Haan theorem. Too high a threshold results in very few exceedances – and thus less information – for the estimation of the model. One tool for the selection of an appropriate threshold is the mean residual life plot. Figure 2 shows the mean residual life plot of the latency data and its approximate 95 % confidence intervals based on the approximate normality of sample means.

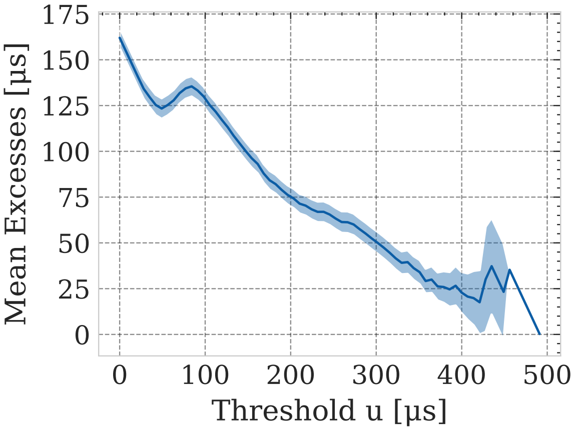


Figure 2: Mean residual life plot for the latency data of a single network flow.

The mean residual life plot depicts the average excess value over the given threshold for a set of different values of the threshold u. According to [4, Section 4.3.1], the mean residual life plot should be approximately linear in u above a threshold u_0 at which the GPD provides a valid approximation for the threshold excesses. In practice, the interpretation of the mean residual life plot often proves to be difficult as it involves a great deal of subjective judgement. Based on Figure 2, we decide to use a threshold of $u = 370\,\mu s$ because of the approximate linearity for from $u = 370\,\mu s$ to $u = 425\,\mu s$. This leads to 42 threshold exceedances, a proportion of about 3.32 %. It might be tempting to suggest a higher threshold such as $u = 425\,\mu s$ as there is some evidence for a linear relationship. However, this would result in only 4 exceedances, too few for a meaningful inference. Similarly, lower thresholds provide an excessive number of exceedances violating the asymptotic assumption of Extreme Value Theory.

Parameter Estimation. There are several fit methods to derive the parameters of (2). Using maximum likelihood estimation, we get

$$(\sigma, \xi) \approx (27.813, -0.064) \tag{3}$$

The 95 % confidence intervals for σ and ξ are [16.141, 39.485] and [−0.356, 0.227], respectively. We omit technical details and simply refer to [7, Chapter 3] and [4, Section 4.3.2].

4.2. Model Checking

Model checking consists of assessing the quality of a fitted generalized Pareto model based on plots and different metrics. In this paper, we focus on probability plots as a graphical technique to evaluate the quality of the parameter estimates in (3). In practice, however, various other means such as quantile plots, return level plots or density plots can also be useful to determine the goodness-of-fit of a model. The probability plot for the fitted model is shown in Figure 3.

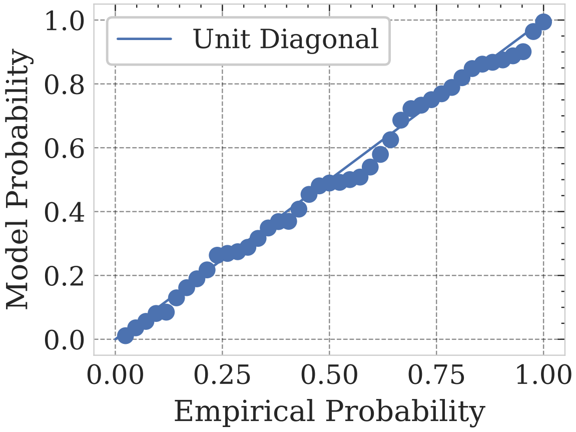


Figure 3: Probability plot for the latency data of a single network flow.

In general, probability plots are a tool for assessing the degree to which a data set follows a given distribution. In Figure 3, the data are plotted against the GPD with the model parameters of (3). The construction of probability plots ensures that the points should lie close to the unit diagonal if the data set follows the given distribution [4, Section 4.3.5]. No substantial departures from linearity can be seen, so that we deem our model to be suitable for extrapolation.

4.3. Extrapolation

After having estimated suitable parameter values, it is possible to utilize the derived model to predict extreme values. Usually, it is convenient to interpret extreme value models in terms of return periods and return levels. The former, the return period, corresponds to the average time between extreme events. The latter, the m-observation return level, describes the value that is expected to be exceeded exactly once in the next m observations [4, Section 4.3.3].

Using the model parameters of (3), the 10 000-observation return level is calculated to be approximately 505 µs, i. e., a latency value above 505 µs is expected to be witnessed only once every 10 000 observations. Assuming that the number of network packets per second stays constant at 7.7 packets/s, 10 000 observations correspond to about 22 min. Note that this is an extrapolation to an unobserved level. The data set only consists of 1264 observations sent over a period of 2.7 min with a maximum latency of about 491 µs.

The 95 % confidence interval is calculated to be approximately [429 µs, 580 µs] via the Delta Method. The confidence intervals often tend to be large, as uncertainty can be magnified in extrapolation.

5. Applications

This section is dedicated to presenting applications of Extreme Value Theory in the networking area.

CBA-EVT. Wang et al. [14] propose to use EVT in a medium access control (MAC) protocol designed for battery-powered wireless sensor networks (WSNs). WSNs are networks of spatially dispersed sensors that monitor physical conditions of the environment. Amongst other areas, they are used in earth sensing, e. g. for natural disaster prevention. It is paramount that MAC protocols for WSNs are energy-efficient to ensure that the sensors can serve their intended functions longer. CBA-EVT is such a MAC protocol that aims to be energy-efficient while also avoiding long latencies. It is named after the two theoretical methods that are the foundation of the protocol: Cost Benefit Analysis and Extreme Value Theory. For a given time slot, Extreme Value Theory in CBA-EVT is used to estimate the completion time of each node, i. e., the time after which no further packets will have to be received in this time slot. This can be used to enable the node to enter a low-power mode early during one time slot and thus saving energy without sacrificing latency.

Vehicular networks. Extreme Value Theory is also used to ensure stringent latency and reliability constraints in vehicular networks. Vehicle-to-vehicle safety applications are inherently time-critical, as individual vehicles rely on acquiring real-time status updates from each other. One commonly used metric is the age of information (AoI). It measures the time elapsed since the latest status update that reached its intended destination has been generated at its source. As argued by Abdel-Aziz et al. [2], minimizing the average AoI in vehicular networks cannot fulfill the unique requirements of ultra-reliable and low-latency vehicular communication. Instead, the authors use Extreme Value Theory to reduce the probability of outliers in the AoI distribution and show the achieved improvements of their approach with simulation results.

Wireless networks. Vehicular networks are a special case of wireless networks. As shown by Mouradian [12], Extreme Value Theory is particularly attractive for studying worst case delays in wireless networks. In contrast to wired networks, wireless networks are more susceptible to unpredictable behavior of the environment such as electromagnetic interference. These disturbances cannot be captured by formal methods like network calculus. As a result, statistical methods, especially Extreme Value Theory, are a valuable tool for the study of worst case delays in wireless networks.

6. Conclusion and Future Work

In this paper we discussed different prediction approaches for rare latency events. In particular, we looked at a general statistical method called Extreme Value Theory and the two main approaches therein: the Block Maxima approach and the Peaks over Threshold approach.

To investigate the Peaks over Threshold method in greater detail, we modeled the tail-end of the latency distribution of a single network flow. The latency data originated from a network experiment on a single physical host using real networking hardware. We consider this approach a powerful tool not only for estimating worst case latencies but also for many other applications in the networking area. However, the accuracy of the predictions is limited by the quality and amount of measurement data and by the assumptions about the data.

To reduce the complexity of the analysis, we assumed that the measurements are independent and identically distributed. The assumption of independence has already been relaxed by Helm et al. [15]. Future work can explore further relaxation of these assumptions and their effect on the accuracy and reliability of the predictions. The consequences of following the Block Maxima approach instead of the PoT method in the context of predicting high latencies also require further investigation.

References

[1] S. R. Pokhrel, J. Ding, J. Park, O.-S. Park, and J. Choi, "Towards enabling critical mMTC: A review of URLLC within mMTC," *IEEE Access*, vol. 8, pp. 131 796–131 813, 2020.

[2] M. K. Abdel-Aziz, C.-F. Liu, S. Samarakoon, M. Bennis, and W. Saad, "Ultra-reliable low-latency vehicular networks: Taming the age of information tail," in *2018 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2018, pp. 1–7.

[3] F. Wiedner, M. Helm, S. Gallenmüller, and G. Carle, "HVNet: Hardware-Assisted Virtual Networking on a Single Physical Host," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2022, pp. 1–6.

[4] S. Coles, J. Bawa, L. Trenner, and P. Dorazio, *An introduction to statistical modeling of extreme values*. Springer, 2001, vol. 208.

[5] J. Galambos, "Extreme value theory for applications," in *Extreme Value Theory and Applications: Proceedings of the Conference on Extreme Value Theory and Applications, Volume 1 Gaithersburg Maryland 1993*. Springer, 1994, pp. 1–14.

[6] M. I. Gomes and A. Guillou, "Extreme value theory and statistics of univariate extremes: a review," *International statistical review*, vol. 83, no. 2, pp. 263–292, 2015.

[7] L. Haan and A. Ferreira, *Extreme value theory: an introduction*. Springer, 2006, vol. 3.

[8] F. X. Diebold, T. Schuermann, and J. D. Stroughair, "Pitfalls and opportunities in the use of extreme value theory in risk management," *The Journal of Risk Finance*, vol. 1, no. 2, pp. 30–35, 2000.

[9] S. Wambura, H. Li, and A. Nigussie, "Fast memory-efficient extreme events prediction in complex time series," in *Proceedings of the 2020 3rd International Conference on Robot Systems and Applications*, 2020, pp. 60–69.

[10] C. She, C. Sun, Z. Gu, Y. Li, C. Yang, H. V. Poor, and B. Vucetic, "A tutorial on ultrareliable and low-latency communications in 6G: Integrating domain knowledge into deep learning," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 204–246, 2021.

[11] A. Van Bemten and W. Kellerer, "Network calculus: A comprehensive guide," 2016.

[12] A. Mouradian, "Extreme value theory for the study of probabilistic worst case delays in wireless networks," *Ad Hoc Networks*, vol. 48, pp. 1–15, 2016.

[13] A. Bücher and C. Zhou, "A horse race between the block maxima method and the peak–over–threshold approach," *Statistical Science*, vol. 36, no. 3, pp. 360–378, 2021.

[14] T.-L. Wang, J.-C. Kao, and S. A. Ciou, "CBA-EVT: A traffic-adaptive energy-efficient MAC protocol for wireless sensor networks," in *2014 Wireless Telecommunications Symposium*. IEEE, 2014, pp. 1–6.

[15] M. Helm, F. Wiedner, and G. Carle, "Flow-level Tail Latency Estimation and Verification based on Extreme Value Theory," in *2022 18th International Conference on Network and Service Management (CNSM)*. IEEE, 2022, pp. 359–363.