December 12, 2006

RT0696
Computer Science 11 pages

Research Report

Evaluating Availability under Heavy-tailed Repair Time

Sei kato and Takayuki Osogami

IBM Research, Tokyo Research Laboratory
IBM Japan, Ltd.
1623-14 Shimotsuruma, Yamato
Kanagawa 242-8502, Japan

Evaluating Availability under Heavy-tailed Repair Time

Sei Kato and Takayuki Osogami
IBM Research, Tokyo Research Laboratory
1623-14 Shimotsuruma, Yamato-shi, Kanagawa-ken, 242-8502 Japan
{seikato, osogami}@jp.ibm.com

Abstract

The time to recover from failures has a great impact on the availability of Information Technology (IT) systems. We find that the repair times have heavy-tailed power law distributions with scaling exponents close to one for two IT systems, an in-house system hosted by IBM and a high performance computing system at the Los Alamos National Laboratory. This means that the repair times of these systems have infinite variance and may also have infinite mean. As a result, a claasical metrics based on thie mean time to repair are not suitable for evaluating the availability of these systems. We propose a new metric, the T-year return value, for evaluating the reliability of IT systems. The T-year return value refers to the value that the mean repair time exceeds on average once every T years estimated based on the extreme value theory. We evaluate the T-year return values of the two IT systems and find that the T-year return value can well represent the system availability.

1 Introduction

Achieving high dependability is a major requirement in managing information technology (IT) systems. As IT systems play more roles in business and government activities, the importance of its dependability has been increased. To design and build a highly dependable IT system, it is of fundamental significance to measure and evaluate the availability of the IT system such as the frequency of failures and the time to recover from failures.

A significant amount of research has been devoted to measure and analyze the statistics of repair time of IT systems. The statistics of repair time was first studied by Long et al. [14]. They measured the time to repair (TTR) by polling Internet hosts periodically for three months, concluding that the repair time distribution is far from exponential. More intensive study on the statistics of repair time was carried out by Schroeder and Gibson [18]. They analyzed the statistical properties of repair time data, which were collected over nine years at Los Alamos National Laboratory (LANL) high performance computing (HPC) systems and concluded that the repair time is well modeled by a log-normal distribution. Also, based on the facts that the repair time distribution is non-exponential, many reliability models have been proposed with non-exponential distributions (in particlular, phase type distributions) [19, 2, 12].

While the prior work studies the distribution of repair times, there has bee no research that particularly studies long repair times, that is the tail of the repair time distribution. Note that long repair times can have significant impact on the availability of IT systems. We analyze the tail of the repair time distribution of two IT systems and find that the repair times have heavy-tailed power law distributions with scaling exponents close to one. This means that that the repair time has infinite variance and may have infinite mean, and hence the sample mean of the repair times does not converge or requires many samples to converge to the true mean repair time.

We propose a new metric that can provide an intuitive understanding of system availability even when the repair time have heavy-tailed power law distribution. The metric is referred as the T-year return value and is defined as the value that the repair time exceeds on average once every T years. We calculate the T-year return value, based on the extreme value theory, a theory developed for evaluating the maximum values of rare events.

The contributions of this paper are thus twofold. First, we study the statistical properties of the repair times of IT systems and find that the repair times have heavy-tailed power law distribution. The study of repair time provides us an insight as to how the system availability should be analyzed and evaluated. Second, we propose the T-year return value as a new metric for evaluating the system availability. We find that the T-year return value allows us to assess the system availability more effectively than classical metrics such as MTTR.

This paper is organized as follows. In Section 2, we analyze the statistical properties of repair time data of two IT systems and show that the repair times have heavy-tailed

of power law distributions. Section 3 gives a brief review of the extreme value theory. In Section 4, we analyze the T-year return value of the two IT systems and discuss the results of the analysis Section 5 is devoted to concluding remarks.

2 Repair Time Analysis

The prior work studies the statistical properties of the repair time distribution and shows that the cumulative distribution function (CDF) is "S"-shaped. In contrast, we focus on the statistics of the tail range since the statistics of rare events with long rapair times can have significant meaning in estimating the the mean time to repair. In this section, we discuss the statistical properties of repair time distribution in the tail range, showing that the repair time has a heavy-tailed power law distribution.

2.1 Repair Time Data and System Configuration

Analyzing the statistical properties of repair time requires large number of incident data. This is especially so when we study the tail of a distribution. Since incidents of IT systems do occur rarely, one needs to use the systematically collected incident data of large IT systems which is collected for a long period in production. To analyze the large amount of repair time data, we use the data of two large IT systems, whose incident data is stored in the incident management database in an organized way for years

2.1.1 An in-house system

One data used for analysis is the repair time data of an in-house system which is hosted by IBM. We use 332 incidents data which occurred in the system from April 1st, 2005 to February 27, 2006.

The data is extracted from an incident management database which stores records on every incident that occurred at all systems that include open systems and mission-critical systems. Each record contains incident description, time of occurrence, time of recovery, recovery process, business impact level and so forth. Note that the data stored in the database contains all incidents including those which do not affect the system. To analyze data more precisely, we extracted the incidents which did affect the system, because, in case of the incidents which do not have system impact, the restore time can be longer than expected.

The incident data are created as follows. When incidents are detected by a monitoring system, alerts are displayed on the monitoring console. Alternatively, operators are called in by users when the system is unavailable. Then the operator creates a new record and inputs the incident description and start time of the failure. Following that, the operator asks system engineers to repair the system or to seek the directions to repair. Operators input the time into the database every time when a remedy for repair is executed. And when finally the system is recovered, the incident end time and incident description is inputted to the database.

Since the data is created manually by operators, as pointed out by Schroeder and Gibson [18], the accuracy of data depends highly on the operator. To avoid using inaccurate data, we eliminated the incorrect data by checking the incident description for all incidents.

2.1.2 LANL HPC System

The other data we use for analysis is the one which has been collected on the LANL HPC system. LANL provides its computer operational data to support and enable computer science research [1]. In the following analysis, 3997 incidents are used which occurred in one of the LANL HPC system during May 6, 2002 to September 8, 2005.

LANL HPC system consists of 22 high-performance computing sub-systems, which is 18 SMP-based systems and four NUMA-based systems. Each sub-system varies in the number of nodes, the number of processors and the number of processors per node. More information on the system can be found in [1, 18]. The failure record contains the start time and end time of the failure, the system and node affected, and the root causes. Incident reporting system is much like to that of the in-house system.

Since each sub-system varies in production time, it is not reasonable to treat all repair time data of the whole system together. It is rational to analyze the data on only one sub-system because each incident of one sub-system can be assumed to be a realization with an identical distribution. Form this point of view, we use records on incidents which have been occurred on a sub-system whose system ID equals to 18, which corresponds to the system ID 7 in the reference [18]. This sub-system consists of 1024 nodes each of which has four processors and falls into four types according to the size of memory per node, 8, 16, 32 and 352 GB.

2.2 Analytical Results

2.2.1 Statistics summary

Statistics for both system are summarized in the Table 1. While 75% of the incidents are repaired within 201.5 minuites and 179.4 minutes for each system respectively, the repair time of the other 25% of the incidents vary very widely. The result shows that the median for the in-house system is slightly smaller than that of the LANL HPC system, whereas the 3rd quantile for the in-house system is larger. This implies that the repair time of the first 50%

Table 1. Summary statistics.

system	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
in-house system	0.0	16.00	49.0	504.8	201.5	35400.0
LANL HPC system	1.0	28.00	58.0	179.4	142.0	25370.0

of the in-house system is small, but long repair times are required to recover from the failures more often a the in-house system. incidents occur. There exists one incident that is repaired within one minute. This record gives that the minimum repair time of the in-house system equals to 0.0.

2.2.2 Analysis of time to repair

Figures 1 and 2 show the time series data of repair time X_i $(1 \leq i \leq n)$, the sample average series $\bar{X}_i = 1/i \sum_{j=1}^{i} X_j$ and the sample variance series $S_i^2 = 1/(i-1) \sum_{j=1}^{i} (X_j - \bar{X}_i)$ for the in-house system and the LANL HPC system. We see that the time series data shows bursts, which corresponds to rare events which have long repair time. The sample average data for the in-house system jumps at the same moment when the burst occurs and the saome average does not seems to converge. The same phenomena also can be seen in the sample average data for the LANL HPC system seems, but on the contrary, the sample average seems to converge to a certain value, whereas the sample variance does not seems to converge. The results that the sample average or sample variance fluctuate in time suggest that the distribution for the repair time is heavy-tailed and that the scaling exponent for the in-house system is smaller than 1, whereas that of LANL HPC system is in the range between 1 and 2.

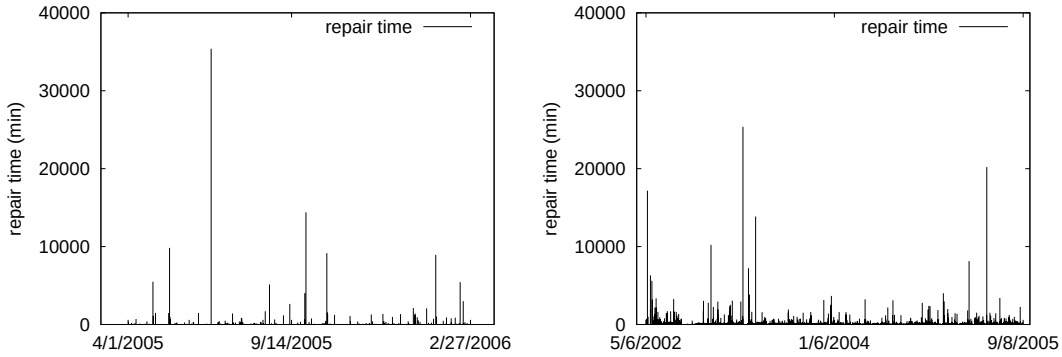
To see these property quantitatively, in Figure 3, we plot the complementary cumulative distribution function of repair time $F_c(x) = \Pr\{X > x\}$ in log-log plot together with the same in semi-log plot. As mentioned in the previous studies, the semi-log plot shows "S"-shaped, which means that the distribution is not exponential. With respect to the tail of the distribution, we find that the distribution of repair time for each system has a long tail. This means that the probability that the incident that requre very long repair time occurs does not decay exponentailly as $x \to \infty$. The power low of the repair time distribution is cleare in the case of LANL HPC system, where we see the scaling law in a wider range (Figure 3 (b)). The scaling exponent α such that,

$$F_c(x) \sim x^{-\alpha} \quad \text{as} \quad x \to \infty$$

is obtained to be 0.7 and 1.1 for each IT system respectively by the least square fit of data in power law region, which suggests that the both distribution is heavy-tailed. If $0 < \alpha \leq 1$, both mean and variance of the distribution is infinite whereas if $1 < \alpha \leq 2$, only variance is infinite. The scaling exponents obtained agree with the results in Figure 3, where the sample average does not converge in-house system case and only the variance does not converge in the LANL HPC case

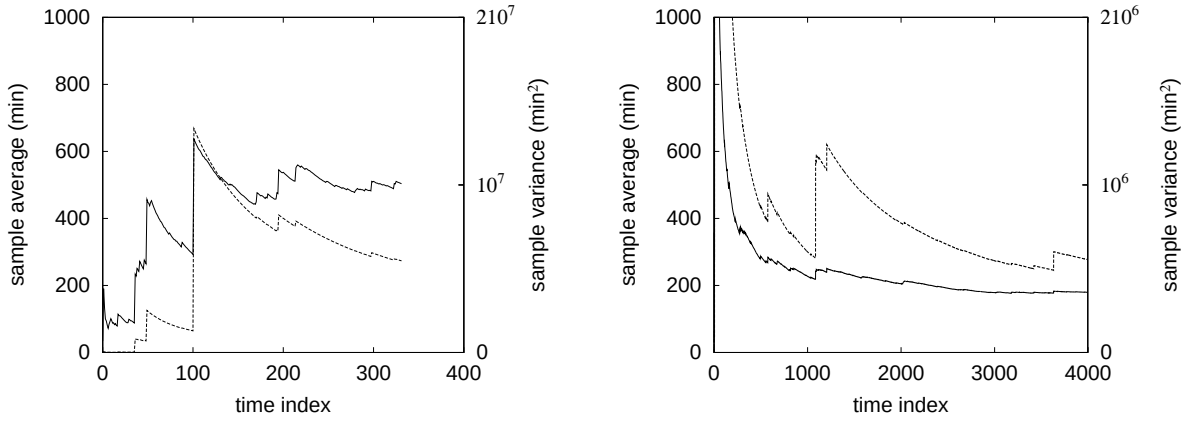
This result is interesting since if the repair time distribution is heavy-tailed with scaling exponent $\alpha \leq 1$, we cannot obtain the "real" mean time to repair. In this case, the sample average of repair time do not converge to the population mean value since the law of large number does not work because of the infinite moments of the distribution. This means that the sample average of repair time varies in time and thus do not represent an intrinsic value. In the case of $1 < \alpha \leq 2$, the sample variance do not converge to the population variance whereas the sample mean converge into the population mean. In this case, the sample mean does exist but its confidence interval is not obtained using the sample data. To avoid these difficulties, other representative values which represents a system availability should be offered. To meet these demands, in the following section, we propose a new metric, T-year return value, to represent the system availability which works well even when the repair time has a heavy-tailed distribution.

The arguments on the values of the scaling exponents may be appropriate to be mentioned here. To obtain the scaling exponent in the power law range more precisely, larger number of data is required. The log-log plot of the repair time distribution for the whole LANL HPC system and whole data of the in-house system data shows power law distribution over 3 orders of magnitude (Figure 4). The scaling exponent of this range is close to 1.0, which is slightly smaller than that of the previous analyzed sub-system. As mentioned previously, the data for the whole LANL HPC system contains the data for various systems and the production time. The repair time distribution for the whole in-house system data including those incidents data which affect the system shows the power law distribution with long scaling range (Figure 4). The scaling exponent of this region is also close to 1.0 We believe that the scaling exponent varies in the system but we have no clear explanation as to what determines the scaling exponent. The important findings here is that the repair time has a heavy-tailed distribution whose scaling exponent is less than 2.



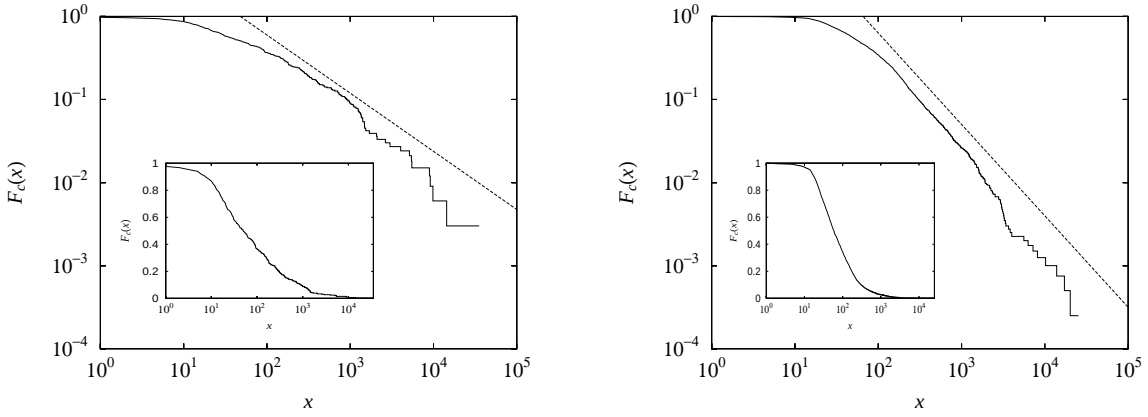
(a) In-house system (b) LANL HPC system

Figure 1. Time series plot for repair time for the in-house system (a) and for the LANL HPC system (b).



(a) In-house system (b) LANL HPC system

Figure 2. Sample average series (solid line) and sample variance series (dashed line) of repair time series for the in-house system (a) and for the LANL HPC system (b).



(a) In-house system

(b) LANL HPC system

Figure 3. Complementary cumulative distribution function of repair time for the in-house system (a) and the LANL HPC system (b). The dashed line for each shows $\propto x^{-0.7}$ in (a) and $\propto x^{-1.1}$ in (b) respectively. In the inset, we show the semi-log plot of the complementary cumulative distribution.

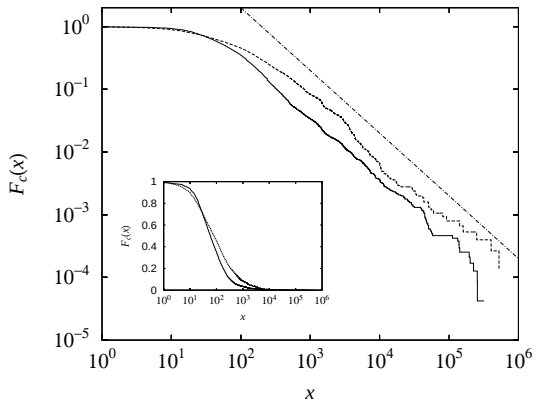


Figure 4. Log-log plot of the complementary cumulative distribution function for the in-house system (dashed) and for the LANL HPC system (solid). The slope denotes $\propto x^{-1.0}$. In the inset, semi-log plot of the CDF is shown.

3 Extreme Value Theory

In the previous section, we find the distribution of repair time is heavy-tailed. This leads us to apply the extreme value theory to the repair time data. In this section, we review the extreme value theory and introduce the T-year return value.

3.1 Model Formulation

The extreme value theory originally roots the study of limiting distribution of maximal values by Fisher and Tippett [9] and now forms a branch of statistics that studies the statistics of maximal or minimal value of rare events. After the success in applying the theory to engineering by Gumbel [11], the theory has been widely used in many fields such as hydrology [8, 17], meteorology [20], telecommunication engineering [21], actuarial science [15] and financial engineering [5, 13].

Suppose that there exits a series of block maximal $\{M_1, ..., M_n\}$ over a certain period of time. This is for example the time series data of monthly maximum rainfall data or the data of annual maximum sea levels. Then the extreme value theory tells us that the distribution of these maximal series approaches to a distribution family named Generalized Extreme Value distribution (GEV) given by Equation 2) in the limit of $n \to \infty$. Thus, the distribution of the block maximal M_i may be well approximated by a GEV for large number of n,

The main essence of the extreme value theory is contained in the following Fisher-Tippett theorem [9]. Let M_n

denote the maximal value for a sequence of n independent and identically distributed random variables $X_1, X_2, ..., X_n$ with a common probability distribution $F(x)$, i.e.,

$$M_n = \max\{X_1, ..., X_n\}.$$

The theorem says that if there exist sequences $\{a_n > 0\}$ and $\{b_n\}$ such that

$$\begin{aligned} \Pr\{(M_n - b_n)/a_n \leq x\} &= F^n(a_n x + b_n) \\ &\to G(x) \quad \text{as} \quad n \to \infty, \end{aligned} \tag{1}$$

where $G(x)$ is a non-degenerate distribution function, then G is a member of the Generalized Extreme Value Distribution (GEV) family given by

$$G_\xi(x) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}. \tag{2}$$

The GEV family are referred to as Fréchet distribution for $\xi > 0$, Gumbel distribution for $\xi < 0$, and Weibull distribution for $\xi = 0$. If Condition (1) is true, we say that F is in the maximum domain of attraction of G and we write as $F \in \text{MDA}(G)$. Using this definition, the theorem can be stated as follows: if $F \in \text{MDA}(G)$, then G is G_ξ for some parameter ξ. The domain of G_ξ is known to be large, and almost all well known distributions belong to MDA(G_ξ). In relation to our study, the MDA of the Fréchet distribution G_ξ for $\xi > 0$ should be mentioned here. It is shown by Gredenko [10] that $F_c(x) = x^{-1/\xi}L(x)$ if and only if $F \in \text{MDA}(G_\xi)$ for $\xi > 0$, where $L(x)$ is a slowly varying function which satisfies

$$\lim_{x\to\infty} L(ax)/L(x) = 1 \quad \text{for } a > 0.$$

This implies that the maximal distribution is the Fréchet distribution if and only if the distribution function is heavy-tailed power law distribution. Our particular interest is the MDA of the Fréchet distribution since, as seen in the previous section, the repair time distribution is heavy-tailed power law distribution. Thus, the maximal of repair times can be well modeled by the Fréchet distribution.

3.2 Threshold Models

The model described in the previous subsection uses the data of block maximal, but using only these values could be an inefficient approach. It turns out that we can use data which exceeds a certain threshold to study the distribution of the block maximal. Thus, the peak over threshold (POT) method, which focuses on the excesses above the threshold u has been developed [6, 7]. The classical extreme value theory mentioned previously suggests that the block maximal distribution is roughly approximated by the GEV, whereas this POT model suggests that a data series over a certain threshold is approximated by the Generalized Pareto distribution (GPD), which is given by Equation (4).

The conditional probability of the excess over the threshold u under the condition that X is larger than u is given as

$$\begin{aligned} F_u(y) &= \Pr\{X - u \leq y | X > u\} \\ &= \frac{F(y+u) - F(u)}{1 - F(u)}, \end{aligned} \tag{3}$$

for $0 \leq y < x_F - u$ where x_F is the right endpoint of the cumulative distribution function F defined by

$$x_F = \sup_{x\in\mathbb{R}}\{F(x) < 1\}.$$

The Pickands-Balkema-de Haan theorem [3, 16] shows that if and only if $F \in \text{MDA}(G_\xi)$, then there exists a measurable positive function $\sigma(u)$ such that

$$\lim_{u\to x_F} \sup_{0\leq x<x_F-u} \left|F_u(y) - G_{\xi,\sigma(u)}(y)\right| = 0,$$

where $G_{\xi,\sigma(u)}(y)$ is a Generalized Pareto Distribution (GPD),

$$G_{\xi,\sigma(u)}(y) = 1 - \left(1 + \frac{\xi y}{\sigma(u)}\right)^{-1/\xi} \tag{4}$$

and

$$\sigma(u) = \sigma + \xi\,(u - \mu)\,.$$

This theorem tells us that, for a sufficiently high threshold u, the distribution function of the excesses can be approximated by $G_{\xi,\sigma(u)}$ as

$$F_u(y) \approx G_{\xi,\sigma(u)}(y), \quad y > 0, \tag{5}$$

for some values of ξ and $\sigma(u)$ if and only if $F \in \text{MDA}(G_\xi)$. As discussed in Subsection 3.1, the distribution of maximal repair time can be modeled by the Fréchet distribution for sufficient number of samples, which leads that the repair time distribution belongs to the maximal attractor domain of the Fréchet distribution. This gives us a strong theoretical background that the distribution of the excesses for repair time can be well approximated by GPD. Thus, in the following analysis, we adopt the POT method and models the excesses of repair time by GPD.

3.3 Model Validation

To use the model for further inference, we need to check how well the model describes the data. The goodness of fit of the model to data can be evaluated by comparing the empirical distribution function and the model distribution function.

Let $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(m)}$ denote the series of excesses over the threshold u in ascending order. Then the empirical distribution function is given as,

$$\tilde{G}(y) = i/(m+1) \quad \text{for} \quad y_{(i)} \leq y < y_{(i+1)}.$$

The corresponding model distribution function for $\xi \neq 0$ is given by substituting the estimated parameters $\hat{\xi}$ and $\hat{\sigma}(u)$ to Equation (4) as,

$$\hat{G}(y) = 1 - \left(1 + \frac{\hat{\xi} y}{\hat{\sigma}(u)}\right)^{-1/\hat{\xi}}.$$

If the threshold model works well, the probability plot points

$$\left(\tilde{G}(y_{(i)}), \hat{G}(y_{(i)})\right), \quad i = 1, ..., m \tag{6}$$

should lie close to the straight line with gradient of 1 and y-intercept of 0. Thus this plot enables us to verfityvisually whether the data can be modeled by GPD.

Another commonly used method for diagnosis is to compare the quantiles of each distribution function. The probability plot given by Equation (6) describes how well the model probability distribution function predicts the probability. In contrast, the quantile plot can be used to evaluate the goodness of fit in the tail range.

To see the goodness of fit in the different scale, the quantile plot is used. By operating the inverse function to Equation (6), the corresponding quantile plot is obtained as,

$$\left(\hat{G}^{-1}(i/(m+1)), \tilde{G}^{-1}(i/(m+1))\right), \quad i = 1, ..., m,$$

where

$$\hat{G}^{-1}(y) = \frac{\hat{\sigma}}{\hat{\xi}}\left\{(1-y)^{-\hat{\xi}} - 1\right\}$$

and

$$\tilde{G}^{-1}(i/(m+1)) = y_{(i)}.$$

3.4 T-year return value

It is not meaningful to analyze the maximal value of the repair time distribution x_F, since in the previous section, we find that the scaling parameter ξ is positive, which means that the repair time distribution is not bounded above, i.e. $x_F = \infty$. Instead we adopt the T-year return value as a representative value, which represents the value that is exceeded on average once every T years. Note that T-year value does not depend on the degree of exceeds, but only depends on the number of times of excesses during T-years. In the context of the operational risk management, T-year return value is referred as the value at risk. Suppose that a GPD with parameter ξ and $\sigma(u)$ models the exceedances of a threshold u. Then from Equation (3), (4) and (5), the probability that a variable X is larger than x under the condition that $X > u$ is obtained as

$$\Pr\{X > x | X > u\} = \left[1 + \xi\left(\frac{x-u}{\sigma(u)}\right)\right]^{-1/\xi}.$$

If incidents occur λ times per year, then the T-year return value x_T is the solution of the following equation

$$\zeta_u \left[1 + \xi\left(\frac{x_T - u}{\sigma(u)}\right)\right]^{-1/\xi} = \frac{1}{\lambda T},$$

where $\zeta_u = \Pr\{X > x\}$. Consequently the T-year return value is given by

$$x_T = u + \frac{\sigma(u)}{\xi}\left[(\lambda T \zeta_u)^{\xi} - 1\right]. \tag{7}$$

Hence, substituting the estimated parameters into Equation (7), the estimated T-year return value $\hat{x}_T$ is obtained as

$$\hat{x}_T = u + \frac{\hat{\sigma}(u)}{\hat{\xi}}\left[(\hat{\lambda} T \hat{\zeta}_u)^{\hat{\xi}} - 1\right], \tag{8}$$

where $\hat{\zeta}_u$ is the portion of incidents exceeding u.

4 Analytical Results

We apply EVT to the following two data sets: repair time data of an in-house system which is hosted by IBM and data collected at the LANL HPC system. In this section, we see the goodness-of-fit of the GPD model and see the effectiveness of the T-year return value for the assessment of system availability.

4.1 Parameter Estimation

At this moment, there has not been established a universal way to decide the threshold value. In our study, we use the σ^* method [4] to choose an optimal threshold value. We chose threshold $u = 100$ for the in-house system case and $u = 200$ for the LANL HPC system case. The number of exceedances m is $m = 121$ and $m = 666$, respectively.

Maximum likelihood estimation is used to estimate the model parameters $\hat{\xi}$ and $\hat{\sigma}(u)$. We obtain the maximal likelihood estimators by maximizing the log likelihood function

$$\log L(X_1, ..., X_m | \xi, \sigma(u)) = -m \log \sigma(u) - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{m} \log\left(1 + \frac{\xi X_i}{\sigma(u)}\right),$$

solving the 2-dimensional linear equation for $\hat{\xi}$ and $\hat{\sigma}(u)$ numerically using the Newton-Raphson method. The initial value for ξ and $\sigma(u)$ are obtained by the rough estimation since the moment method does not work for positive scaling parameter.

4.1.1 In-house system

By using the 121 exceedances data over the threshold $u = 100$, maximizing likelihood estimates the parameter as $(\hat{\xi}, \hat{\sigma}) = (0.969, 266.3)$ with the corresponding maximized log-likelihood of -1341.2 . The inverse of the Fisher's information matrix gives the 95% parameter confidence interpal for $\hat{\xi}$ and $\hat{\sigma}$ as $[0.613, 1.32]$ and $[170.7, 361.8]$ for each. Recall that the scaling exponent α of power-law is given by the reciprocal of the scaling parameter ξ, i.e. $\alpha = \xi^{-1}$. Then the estimated scaling exponent is obtained as $[0.75, 1.63]$ with 95% confidence. The estimated scaling exponent α is a slightly above 1.0, but there is a possibility that the exponent is less than 1.0 since the confidence interval ranges over 1.0. Thus, in this case, we cannot distinguish whether that the scaling exponent is less than 1 or not.

4.1.2 LANL HPC system

The scaling and positioning parameters are estimated by maximizing the log likelihood function using the 666 exceedances repair time data of LANL HPC system. The parameters are estimated as $(\hat{\xi}, \hat{\sigma}) = (0.809, 170.6)$ with the corresponding maximized log-likelihood of -5331.6. The 95% parameter confidence intervals for $\hat{\xi}$ and $\hat{\sigma}$ are calculated as $[0.665, 0.953]$ and $[144.9, 196.4]$, which are smaller than those of the in-house system case. The corresponding confidence interval for scaling exponent α is $[1.05, 1.50]$. Contrary to the results of the in-house system, the scaling exponent is $1 < \alpha \leq$ with 95% confidence, which leads that the distribution is heavy-tailed with infinite variance, but finite mean.

4.2 Model Validation

The probability plot and the quantile plot are shown in Figure 5 and Figure 6. In Figure 5 (a) and (b), we see that the probability plot points lie diagonally, which means that the empirical model distribution is well described by the estimated model distribution. This is especially so in case of LANL HPC where the points are located very close to the diagonal line. On the contrary, there seems exist some quantile points which do not lie on the diagonal line, in both systems (Figure 6 (a) and (b)). These points correspond to the rare events whose repair time is above 10^3. As we can see in Figure (3), the power law distribution breaks in this tail region and this is the reason that these points do not lie on the diagonal line. On the other hand, in the region where $x < 10^3$, the model quantile agrees well with the empirical quantile (the inset of the Figure 6 (a) and (b)).

4.3 T-year return value

Since the formula (7) contains the estimated parameters, the variance of these parameters should be considered to estimate the confidence interval of the T-year return value $\hat{x}_T$. Suppose that the incident occurrence interval x follows the exponential distribution with rate λ. Then the maximal likelihood estimates and Fisher's information matrix gives the log-likelihood estimates $\hat{\lambda} = n/\sum_i^n X_i$ with variance $\hat{\lambda}^{-2}n^{-1}$. Similarly, if the number of exceedances over u follows the binomial distribution $\mathrm{Bi}(n, \zeta_u)$, then the log-likelihood estimates $\hat{\zeta}_u = m/n$ with variance $\hat{\zeta}_u(1-\hat{\zeta}_u)/n$. Then the variance-covariance matrix V for $(\hat{\lambda}, \hat{\zeta}_u, \hat{\xi}, \hat{\sigma}(u))$ is approximated as

$$V = \begin{pmatrix} \hat{\lambda}^{-2}n^{-1} & 0 & 0 & 0 \\ 0 & \hat{\zeta}_u(1-\hat{\zeta}_u)/n & 0 & 0 \\ 0 & 0 & v_{1,1} & v_{1,2} \\ 0 & 0 & v_{2,1} & v_{2,2} \end{pmatrix}$$

where $v_{i,j}$ denotes the (i, j) term of the variance-covariance matrix of ξ and $\sigma(u)$. Using the delta method, the variance for x_T is approximated as

$$\mathrm{Var}(x_T) \approx \nabla {x_T}^t V \nabla x_T,$$

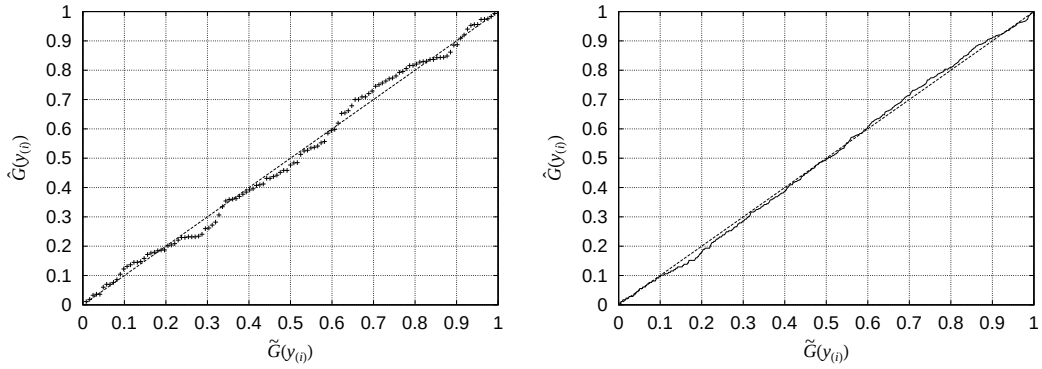
where

$$\nabla {x_T}^t = \left(\frac{\partial x_T}{\partial \lambda}, \frac{\partial x_T}{\partial \zeta_u}, \frac{\partial x_T}{\partial \xi}, \frac{\partial x_T}{\partial \sigma(u)} \right).$$

4.3.1 In-house system

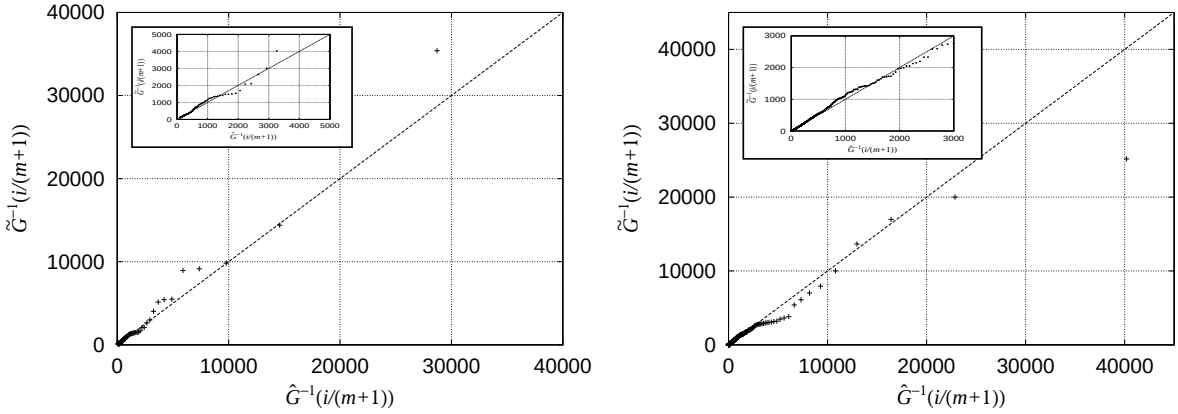
The arrival rate and the exceedance probability is calculated as $\hat{\lambda} = 332/(333/365) = 364.0$ and $\hat{\zeta_u} = 121/332 = 0.364$. Subsisting these estimators into Equation (8), we obtain 1-year return value as 31119.2 and the 1-month return value as 2640.8. In Figure 7 (a), we show the time series data together with the 1-year return value and 1-month return value. We see that the only 1 incident and 10 incidents exceeds the 1-year return value and 1-month return value, respectively during the observation period of April 1, 2005 and Feb. 27, 2006, which is 333 days. We find that these values are very close to the expected number of times when repair time exceeds each return value $333/365 = 0.912$ and $333/365 \times 12 = 10.9$. For reference, the 1-week return value is 560.3 which gives the realized exceedance as 46 whereas the expected value $333/365 \times 12 \times 4 = 43.8$. This also gives rather good agreement. Correspondences are summarized in the Table 2. The table shows that the estimated T-year return value predicts the return value with accuracy in every time scale.

The square root of variance of 1-year return value and 1-month return value are calculated as 19728.1 and 676.5,



(a) In-house sysetm (b) LANL HPC system

Figure 5. Probability plot for the in-house system (a) and the LANL HPC system (b). The dashed line corresponds to $\tilde{G}(y) = \hat{G}(y)$.

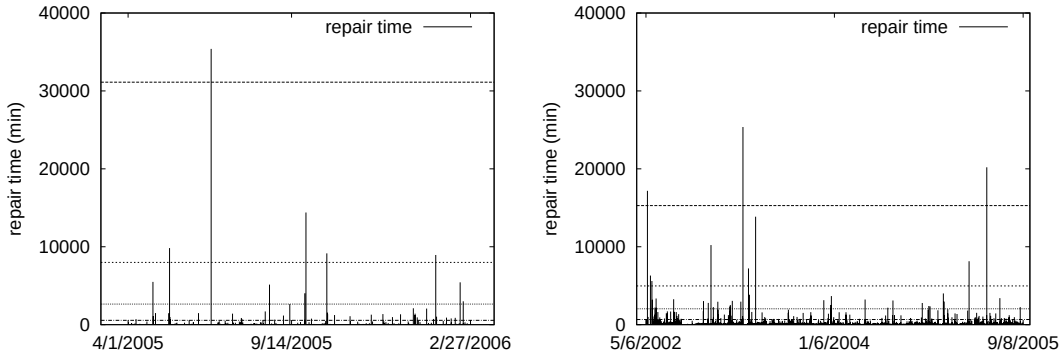


(a) In-house sysetm (b) LANL HPC system

Figure 6. Quantile plot for the in-house system and the LANL HPC system (b). The dashed line corresponds to $\tilde{G}^{-1}(y) = \hat{G}^{-1}(y)$. In the inset, the the same plot enlarged around the origin is drawn.

Table 2. The number of exceedances over the T-year return value.

system	1-week	1-month	1-quarter	1-year
in-house system (estimated)	43.8	10.9	3.65	0.921
in-house system (realized)	46	10	5	1
LANL HPC system (estimated)	160.7	40.2	13.4	3.35
LANL HPC system (realized)	172	38	9	3



(a) In-house sysetm (b) LANL HPC system

Figure 7. T-year return value for the in-house system (a) and for the LANL HPC system (b).The dashed line corresponds to the 1-year return value, the broken line corresponds to the quarterly return value, the dotted line corresponds to the 1-month return value, and the dash-dotted line corresponds to the 1-week return value. Note that the number of exceedance over the 1-year return value for the in-house system is 1, whereas that for the LANL HPC system is 3.

respectively. This leads the 95% confidence interval as $[-7547.8, 69876.3]$ and $[1314.8, 3966.7]$. The result gives rather large interval. This uncertainty is largely due to the uncertainty of the scaling parameter, which is caused by the short range of magnitude of the power law range.

4.3.2 LANL HPC system

The arrival rate λ and the exceedance probability ζ_u for the LANL HPC system are calculated as $\hat{\lambda} = 3997/(1222/365) = 1193.9$ and $\hat{\zeta_u} = 666/3997 = 0.167$. Then we obtain the 1-year return value as 15281.5 and 1-month return value as 2036.9. We see that the number of exceedances over the 1-year return value is 3 (Figure 7 (b)). The expected value during the observation period of May 6, 2002 - Sep. 8, 2005, that is 1222 days, is calculated as $1222/365 = 3.34$, which gives good agreement with the realized value 3. The number of exceedances over the the 1-month return is 38, whereas the expected value is 40.2, which also gives rather good agreement with the expected value. Again, also in the case of LANL HPC system, the table 2 shows that the T-year return value predicts the the return value with accuracy in every time scale, from 1-week to 1-year.

Agreement of the estimated T-year return value with the realized value in both system shows that by estimating the T-year return value, we can safely predict the worst case occurs during the next certain period of time in future while we cannot predict when occurs. The value gives rather good information on availability comparing to the classical value such as MTTR, since these classical values do not give us any information on the occurrence probability of incidents in future. Thus the T-year return value gives us the information on how the system is available or what degrees of incidents occur in a target IT system in a certain period in future.

The square roots of the variance for 1-year return value and for 1-month return value are calculated as 4062.7 and 219.0, leading the 95% confidence interval as $[7318.6, 23244.4]$ and $[1607.6, 2466.1]$. Comparing to interval of the in-house system, we see that the uncertainty is smaller.

5 Conclusion

We have found that repair times of two different IT systems have heavy-tailed power law distributions and the scaling exponents of the tails close to one. This implies that the repair time can have infinite variance and mean. In fact, we have seen that the sample mean and the sample valiance of the repair times of the two IT systems have large fluctuations and do not appear to converge in several years.

Since the mean time to repair (MTTR) is not a suitable metric for evaluating the availability of an IT system when the repair time can have infinite variance and mean, the T-year return value is proposed as a new metric for evaluating the system availability. We calculated the T-year return value for the two IT systems by modeling the repair time by the Generalized Pareto distribution. The 1-year return value and 1-month return value for each system are found to be very close to those of the expected values, which suggests that the T-year return value calculated as in this paper can well predict the value that the repair time exceeds on average once in T years.

At this moment we have no clear explanation as to why the repair time distribution has a heavy-tailed power law distribution and what determines the value of the scaling exponent. To answer these questions, far more studies in human systems of operators are required. More detailed analysis on the tail distribution is now under investigation and will be reported elsewhere.

The authors would like to thank to the LANL Computer Science Educational Institutes for publishing the repair time data. They are also grateful to T. Idé of TRL for his continuous encouragement.

References

[1] http://institutes.lanl.gov/data/.

[2] S. Asgapoor and J. Mathine. Reliability evaluation of distribution systems with non-exponential down times. *IEEE Trans. Power Sys.*, 12(2):579–584, 1997.

[3] A. Balkema and L. de Haan. Residual life time at great age. *Annals of Probability*, 2:792–804, 1974.

[4] S. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag, 2001.

[5] M. Cruz, R. Coleman, and G. Salkin. Modeling and measuring operational risk. *The Journal of Risk*, 1(1):63–72, 1998.

[6] A. Davison. *Statistical Extremes and Applications*, chapter Modelling excesses over high thresholds, with an application, pages 461–482. D. Reidel, 1984.

[7] A. Davison and R. Smith. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(3):393–442, 1990.

[8] L. de Haan. Fighting the arch-enemy with mathematics. *Statistica Neerlandica*, 44(2):45–68, 1990.

[9] R. A. Fisher and L. H. C. Tippett. Limiting forms of the frequency distribution of the largest and smallest member of a sample. *Proc. of the Cambridge Philosophical Society*, 24:180–190, 1928.

[10] B. Gredenko. Sur la distribution limite du terme maximum d'une s´erie al´eatoire'. *Annals of Mathematics*, 44:423–453, 1943.

[11] E. J. Gumbel. *Statistics of Extremes*. Columbia University Press, 1958.

[12] R. Kieckhafer, M. Azadmanesh, and Y. Hui. On the sensitivity of NMR unreliability to non-exponential repair distributions. In *Proc. of the 5th IEEE International Symposium on High-Assurance Systems Engineering (HASE 2000)*, pages 293–300, 2000.

[13] J. L. King. *Operational Risk: Measurement and Modeling*. John Wile & Sons, Ltd., 2001.

[14] D. Long, A. Muir, and R. Golding. A longitudinal survey of internet host reliability. In *Proc. of the 14th Symposium on Reliable Distributed Systems*, pages 2–9, 1995.

[15] A. McNeil. Estimating the tails of loss severity distributions using extreme value theory. *ASTIN Bulletin*, 27:117–137, 1997.

[16] J. Pickands. Statistical inference using extreme order statistics. *Annals of Statistics*, 3:119–131, 1975.

[17] R.W.Katz, M. B. Parlange, and P. Naveau. Statistics of extremes in hydrology. *Advances in Water Resources*, 25:1287–1304, 2002.

[18] B. Schroeder and G. Gibson. A large-scale study of failures in high-performance computing systems. In *Proc. of the International Conference on Dependable Systems and Networks (DSN 2006)*, pages 249– 258, 2006.

[19] B. Sericola. Interval-availability distributuion of 2-state systems with exponential failures and phase-type repairs. *IEEE Trans. on Reliability*, 43(2):335–343, 1994.

[20] R. L. Smith. Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statistical Science*, 4:367–393, 1989.

[21] M. Uchida. Traffi c data analysis based on extreme value theory and its applications. In *Proc. of the IEEE Global Telecommunications Conference (GLOBECOM 2004)*, pages 1418–1424, 2004.