

Evaluating Availability under Quasi-Heavy-tailed Repair Times

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 required to recover from failures has a great impact on the availability of Information Technology (IT) systems. We define a class of probability distributions named quasi-heavy-tailed distributions as those distributions whose time series graph of the sample mean shows intermittent jumps in a given period. We find that the distribution of repair time is quasi-heavy-tailed for three IT systems, an in-house system hosted by IBM, a high performance computing system at the Los Alamos National Laboratory, and a distributed memory computer at the National Energy Research Scientific Computing Center. This means that the mean time to repair estimated by observing incidents within a certain period could dramatically change if we observe incidents successively for another period. In other words, the estimated mean time to repair has large fluctuations over time. As a result, classical metrics based on the mean time to repair are not optimal for evaluating the availability of these systems. We propose to evaluate the availability of IT systems with the T -year return value, estimated based on extreme value theory. The T -year return value refers to the value that the repair time exceeds on average once every estimated T years. We find that the T -year return value is a sound metric of the availability of the three IT systems.

1 Introduction

High dependability is a major requirement for information technology (IT) systems. As IT systems play more roles in business and government activities, the importance of their dependability is increasing. To design and build a highly dependable IT system, it is of fundamental significance to measure and evaluate the availability metrics of the IT system, such as the frequency of failures and the average time to recover from failures.

A significant amount of research has been devoted to measuring and analyzing the statistics of repair times for IT

systems. The statistics of the repair times for an IT system was first studied systematically by Long et al. [17]. They measured the time to repair (TTR) by polling Internet hosts periodically for three months, and concluded that the repair time distribution is far from exponential. Schroeder and Gibson analyzed the statistical properties of repair time data collected over nine years at Los Alamos National Laboratory (LANL) for high performance computing (HPC) systems and concluded that the repair times are well modeled by a log-normal distribution [21]. Also, based on the facts that the repair time distribution is non-exponential, many availability models have been proposed with non-exponential distributions (in particular, with phase type distributions) [22, 4, 15].

While the prior work studied the distribution of repair times, there was no research that particularly studied long repair times, which is the tail of the repair time distribution. Note that a long repair time can have a significant impact on the availability of an IT system. We are analyzing the tail of the repair time distribution of three IT systems and are finding that the repair times have quasi-heavy-tailed distributions, which we define as distributions whose time series graph of the sample mean shows sudden jumps in some periods. As a result, the sample mean time to repair evaluated by observing incidents within a certain period could change dramatically if the observation were continued into another period. There exist robust statistical metrics such as the median, but these metrics do not do a good job at representing the availability of IT systems. Therefore new robust and intuitive metrics are needed for evaluating the availability of IT systems.

We propose to evaluate the availability of an IT system with the T -year return value, which can provide an intuitive understanding of system availability even when the repair time is quasi-heavy-tailed. The T -year return value is defined as the value that the repair time exceeds on average once every T years. The T -year return value would be helpful when we evaluate the failure risk of computer systems or to identify a sub-system whose failure risk is high. We calculate the T -year return value using 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 three IT systems and find that the repair times have quasi-heavy-tailed distributions. The study of repair times 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 system availability. We find that the T -year return value allows us to assess the system availability more robustly than classical metrics such as mean time to repair (MTTR).

This paper is organized as follows. In Section 2, we analyze the statistical properties of the repair time data of three IT systems and show that the repair times have quasi-heavy-tailed distributions. Section 3 gives a brief review of extreme value theory. In Section 4, we analyze the T -year return value of the three 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 this section, we focus on the statistics of the tail region, since the statistics of rare events with long repair times can be significant in estimating the mean time to repair. We will find that the distribution of repair times is quasi-heavy-tailed for these three IT systems. Hence, the mean time to repair has large fluctuations over time and is of limited suitability for evaluating the availability of IT systems.

2.1 Data and System Configuration

Analyzing the statistical properties of the repair times requires incident data for a long period, and particularly when we are studying the tail of the distribution. Since the incidents of IT systems occur rarely, we need to use the incident data for large IT systems that was systematically collected for a long period of production use. We use the data of three large IT systems, whose incident data was stored systematically in an incident management database for many years.

2.1.1 An in-house system

One data source used for analysis is the repair time data of an in-house system hosted by IBM, and on which enterprise applications are running. The data include 332 incidents data occurring from April 1, 2005 to February 27, 2006.

The data is extracted from an incident management database which stores records on every incident that occurred in all systems that include open systems and mission-

critical systems. Each record contains an incident description, the time of the occurrence, the time of the recovery, the recovery process, the business impact level, and so forth. To analyze the incident data more precisely, we extracted only those incidents that did affect the system, because, when the incidents do not affect the system, the repair time tend to be longer, since the operators do not pay much attention to those harmless incidents.

The record of an incident is created as follows. When an incident is detected by a monitoring system, an alert is displayed on the monitoring console. Alternatively, an operator is called by the users when the system is unavailable. Then the operator creates a new record and inputs the incident description and the start time of the failure. Following that, the operator asks system engineers to repair the system or to seek the directions for the repair. The operators input the time into the database every time when a recovery step is performed. When the system finally recovers, the incident end time and incident description are input into the database.

Since the data is created manually by operators, as pointed out by Schroeder and Gibson [21], the accuracy of data varies according to the operator. To avoid using inaccurate data, we eliminated the incorrect data such as improperly recorded data by checking the descriptions for all the incidents.

2.1.2 LANL HPC System

Another data source we used for analysis was collected on the LANL HPC system [1]. We use 3,997 incidents that occurred in one of the LANL HPC systems from May 6, 2002 to September 8, 2005.

The LANL HPC system consists of 22 high-performance computing sub-systems, which are 18 SMP-based sub-systems and four NUMA-based sub-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, 21]. The failure record contains the start time and end time of the failure, the system and node affected, and the root causes. The incident reporting system is much like that of the in-house system.

We analyzed the data on only one sub-system, because each incident of each sub-system can be taken as a realization from an identical distribution though the statistical properties differ among different sub-systems. In particular, we used the records on incidents that occurred on a sub-system whose system ID is 18, which corresponds to the system ID 7 in [21]. We chose this system for analysis, because this system has the largest number of observation samples. This sub-system consists of 1024 nodes, each of which has four processors and falls into one of four types according to the size of memory per node, 8, 16, 32, and

Table 1. Summary of statistics.

system	# incidents	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
in-house system	332	0.0	16.0	49.0	504.8	201.5	35400.0
LANL HPC system	3997	1.0	28.0	58.0	179.4	142.0	25370.0
NERSC Seaborg system	177	1.0	120.0	255.0	534.8	493.8	22230.0

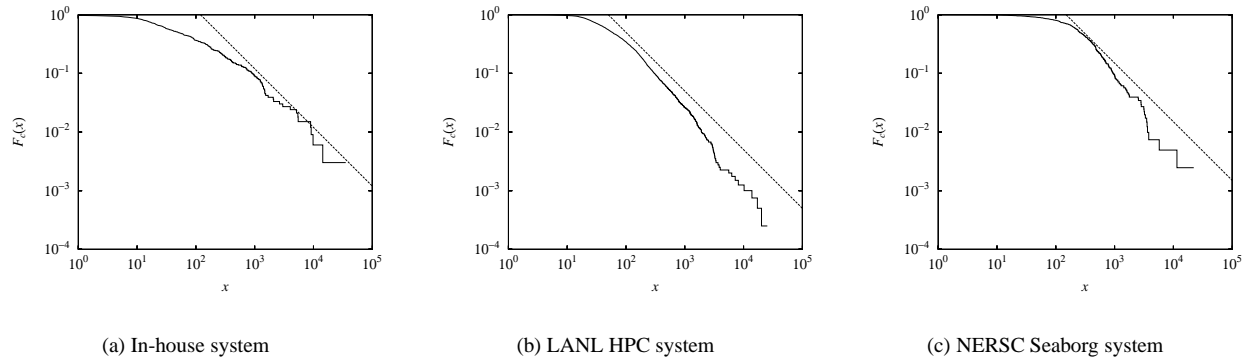


Figure 1. Log-log plot of the complementary cumulative distribution functions of the repair times (a) for the in-house system, (b) for the LANL HPC system, and (c) for the NERSC Seaborg system. The dashed line shows the power law distribution $F_c(x) \sim x^{-1}$.

352 GB.

2.1.3 NERSC Seaborg System

The third data source we used is the data collected on the distributed memory computer system named Seaborg of the National Energy Research Scientific Computing Center (NERSC) used for running scientific computing applications. We used data from 177 incidents that occurred in the system from July 2, 2001 to December 21, 2006.

The NERSC Seaborg system consists of 380 computing nodes with 16 processors per node, and each processor has a shared memory pool of between 16 and 64 GB. The computing nodes are connected with each other with a high-bandwidth, low-latency switching network. The specifications of the Seaborg system can be found in [2].

The system outage data for the NERSC Seaborg system was provided by the Petascale Data Storage Institute website at the NERSC [3]. The data is taken from the remedy problem tracking records that were created by the operators when an outage, scheduled or unscheduled, occurred in the system. This data contains the date and time when the failure occurred and when it was put back online, the outage cause (software or hardware), a category (scheduled or unscheduled), and failure descriptions.

2.2 Statistical Analysis Results

2.2.1 Summary of Statistics

Some statistics for the three IT systems are summarized in Table 1. While 75% of the incidents were repaired within 201.5 minutes, 142.0 minutes, and 493.8 minutes for each system, respectively, the repair times of the other 25% of the incidents vary widely. The median for the in-house system is slightly smaller than that of the LANL HPC system, whereas the third quartile for the in-house system is larger. Thus we expect that the tail distribution of the in-house system decays more slowly than the LANL HPC system. The value of the first quartile, median and the third quartile for the NERSC Seaborg system are larger than those for the in-house system and LANL HPC system. This means that a relatively small number of incidents are repaired in a short time for the NERSC Seaborg system.

There was one incident that was repaired within one minute, so that the minimum repair time of the in-house system equals 0.0. An example of an incident that requires short repair times is a case when a system reboots automatically. Another example is when false alerts are displayed. We classified the reboot as a real incident whereas false alert was not.

2.2.2 The shape of the distributions

In Figure 1, we plot the tail distributions of repair times, $F_c(x)$, which is the fractions of the repair times that exceed x , on a log-log scale. We find that $F_c(x)$ does not appear to decay exponentially as x becomes larger for the range of observed repair times. We see the tail distribution of the in-house system decays slowly compared with the other systems, as was pointed out in the previous section.

Note that, since only a finite number of repair times are observed, we cannot determine whether the distribution obeys the power law distribution or obeys other distributions such as the log-normal distribution. As will be discussed in the next subsection, in our study, it does not matter whether or not the distribution follows the power law distribution in assessing the availability of an IT system in a given period.

2.3 Quasi-heavy-tailed Distribution

A well-known statistical property of the heavy-tailed power law distribution,

$$F_c(x) \sim x^{-\alpha} \quad \text{as } x \rightarrow \infty, \quad (1)$$

is that the first q moments are infinite if $q - 1 < \alpha \leq q$. Let X_i be an independent and identically distributed (i.i.d.) sample having an infinite q -th moment (throughout, we assume non-negative distributions). Then the sample q -th moment $\langle X^q \rangle_j = 1/j \sum_{i=1}^j X_i^q$ will continue to show sudden jumps at a certain time of $j + 1$, whose magnitude $X_{j+1}^q/(j + 1)$ is comparable to $\langle X^q \rangle_j$. This jumping nature of the time series data of the sample q -th order moment is a representative property of the heavy-tailed power law distribution.

In practice, however, since the observation period for the outages of IT systems is limited, we cannot conclude whether or not the observed distribution follows the power law distribution. For example, we cannot or should not conclude the tail distributions of the repair times of the three IT systems follow the power law, decay by following the log-normal or exponential distribution, or is upper bounded by a maximal value. Also we cannot conclude that the time series data of the sample q -th moment of repair times continue showing sudden jumps and diverge to infinity.

Our question is what we can conclude with a limited number of observations. Specifically, can we conclude that the sample mean time to repair is likely to show large jumps in a given time scale? In other words, given a cumulative distribution function $F(x)$ for $x \leq a$ with $F(x)$ for $x > a$ unknown and the number N , of samples, can we determine whether or not the time series data of sample mean $\{\langle X^q \rangle_j : j = 1, \dots, N\}$ shows sudden jumps? Note that the distribution needs to be heavy-tailed for $\{\langle X^q \rangle_j : j = 1, \dots, N\}$ to

show jumps for **any** N , but does not have to be heavy-tailed for a **given** N .

To answer this question, we now define a new class of probability distributions named the quasi-heavy-tailed distribution (QHTD) with parameter (α, N) consisting of those distributions which meet the following two conditions:

1. The gradient of the cumulative distribution function in a log-log plot, $d \log F_c(x)/d \log x$ monotonically decreases,
2. Let h be the x -axis value at which the gradient of the cumulative distribution function in a log-log plot, $d \log F_c(x)/d \log x|_{x=h}$ becomes less than $-\alpha$. Then the value $F_c(h)$ is smaller than $1/N$.

The definition of QHTD relies neither on the tail distribution beyond the cutoff nor on the exact shape of the distribution below the cutoff. The definition is based only on the monotonic decrease of the gradient and the point where the gradient becomes less than $-\alpha$. We will argue that the sufficient conditions for a probability distribution function to show representative properties of the heavy-tailed power law distribution with parameter α in a given time scale with N samples is that the distribution belongs the QHTD.

Here is a simple explanation to show that these two conditions are sufficient for the time series graph to show jumps. Let us consider cumulative distribution functions $F_{\text{sol}}(x)$ and $F_{\text{dot}}(x)$ whose profiles are represented as the solid line and the dotted line in the Figure 2. Note that $F_{\text{sol}}(x) \geq F_{\text{dot}}(x)$ for $x \leq x_{1/N}$ and $F_{\text{sol}}(x) = F_{\text{dot}}(x)$ for $x > x_{1/N}$. For both F_{sol} and F_{dot} , a large value ($x > x_{1/N}$) is generated once in N samples in expectation. Since $F_{\text{sol}} \geq F_{\text{dot}}$ for $x \leq x_{1/N}$, the sample mean of the small value events ($x \leq x_{1/N}$) is smaller for the solid distribution. Therefore the large value is more likely to show a jump for the solid distribution. This suggests that if the dotted line distribution shows the representative property of heavy-tailed power law distribution, the solid line distribution also shows this property. Below, without loss of generality, we assume that $\delta = 1$ and $\alpha = 1$ so that $x_{1/N} = N$. The probability of realizing events whose repair time takes between a and b ($a < b$) is calculated as

$$P(a \leq X < b) = 1/a - 1/b. \quad (2)$$

Similarly, the probability of realizing events whose repair time is larger than N is calculated as $1/N$. Thus, a majority of events takes values of less than three, while those events whose repair time is larger than N occurs once in N events on average. When values of greater than N are ignored, the sample average of N events $\langle X \rangle_N$ is approximately $\log N$. Since a value greater than N is generated once and the sample average except the large value is $\log N$, the time series graph of the sample average shows a jump whose magnitude $X_j/N \geq 1$ is comparable with the sample average $\langle X \rangle_N$.

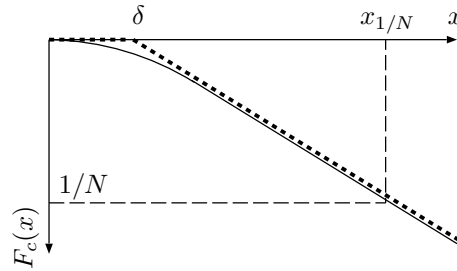


Figure 2. A schematic figure to explain sufficient conditions for a quasi-heavy-tailed distribution. The figure is plotted on a log-log scale. In the linear region, the complementary cumulative distribution function behaves as $F_c(x) \sim x^{-1}$.

Now let us see if the distributions for the three IT systems belong to the QHTD. From Figure 1, we see that the gradient of the complementary cumulative distribution function appears to decrease monotonically for all systems and that the gradient is ≥ -2 up to the number of samples observed. Thus the distributions for the three IT systems belongs to the QHTD for some $\alpha \leq 2$ and N . The treatment of the gradient of the complementary cumulative distribution function should be mentioned here. Note that the empirical distribution function is not only discontinuous but is also noisy. We can judge the first criteria after smoothing the slope of the complementary cumulative distribution functions.

2.4 Nonrobustness of the Sample Mean Time to Repair

In this subsection, we study the time series data of the sample means for the three IT systems and check whether or not the time series graph shows the representative properties of the heavy-tailed power law distribution.

Figure 3 shows the time series data of the repair time $\{X_i\}$ ($1 \leq i \leq n$), and Figure 4 shows the sample average series $\langle X \rangle_j = 1/j \sum_{i=1}^j X_i$ and the sample variance series $\langle X^2 \rangle_j = 1/(j-1) \sum_{i=1}^j (X_i - \langle X \rangle_j)^2$ for the three IT systems. We see that the time series data shows bursts, which correspond to rare events which have long repair times. The sample average data for the in-house system jumps at the same moment as when the burst occurs, and the sample averages does not seem to converge. The same phenomena also can be seen in the sample average and sample variance data for the LANL HPC system and the NERSC Seaborg system.

This means that the mean time to repair evaluated by observing incidents within a certain period could change dynamically if we observe incidents successively for some period. Thus classical metrics based on the mean time to repair are not suitable for evaluating the availability of these

systems.

3 Extreme Value Theory

In the previous section, we found the distribution of repair time is quasi-heavy-tailed, so that a robust metric is needed to represent the system availability. We propose to use the T -year return value, which is estimated based on extreme value theory, as a suitable metric for the system availability. In this section, we review extreme value theory and introduce the T -year return value.

3.1 Model Formulation

Extreme value theory originates with the study of the limiting distributions of maximal values by Fisher and Tippett [12] and now forms a branch of statistics that studies the statistics of maximal or minimal values of rare events. After the success in applying the theory to engineering by Gumbel [14], the theory has been widely used in many fields such as hydrology [10, 20], meteorology [23], telecommunications engineering [24], actuarial science [18], real-time systems [11], and financial engineering [7, 16].

The essence of extreme value theory is contained in the following Fisher-Tippett Theorem [12]. Let M_n denote the maximal value for a sequence of n i.i.d. random variables X_1, X_2, \dots, X_n with a common probability distribution $F(x)$. The theorem states that if there exist sequences $\{a_n > 0\}$ and $\{b_n\}$ such that

$$\Pr\{(M_n - b_n)/a_n \leq x\} = F^n(a_n x + b_n) \rightarrow G(x) \quad \text{as } n \rightarrow \infty, \quad (3)$$

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

$$G_\xi(x) = \exp \left\{ - \left[1 + \xi \left(\frac{x - \mu}{\sigma} \right) \right]^{-1/\xi} \right\}. \quad (4)$$

