Analysis and modeling of time-correlated failures in large-scale distributed systems

Citation for published version (APA):

Yigitbasi, M. N., Gallet, M., Kondo, D., Iosup, A., & Epema, D. H. J. (2010). Analysis and modeling of time-correlated failures in large-scale distributed systems. In *Proceedings of the 11th IEEE/ACM International Conference on Grid Computing (GRID 2010, Brussels, Belgium, October 25-28, 2011)* (pp. 65-72). Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/GRID.2010.5697961>

DOI:

[10.1109/GRID.2010.5697961](https://doi.org/10.1109/GRID.2010.5697961)

Document status and date:

Published: 01/01/2010

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Analysis and Modeling of Time-Correlated Failures in Large-Scale Distributed Systems

Nezih Yigitbasi

Delft University of Technology, the Netherlands
M.N.Yigitbasi@tudelft.nl

Matthieu Gallet

Ecole Normale Supérieure de Lyon, France
matthieu.gallet@ens-lyon.fr

Derrick Kondo

INRIA Grenoble, France
dkondo@imag.fr

Alexandru Iosup

Delft University of Technology, the Netherlands
A.Iosup@tudelft.nl

Dick Epema

Delft University of Technology, the Netherlands
D.H.J.Epema@ewi.tudelft.nl

Abstract—The analysis and modeling of the failures bound to occur in today's large-scale production systems is invaluable in providing the understanding needed to make these systems fault-tolerant yet efficient. Many previous studies have modeled failures without taking into account the time-varying behavior of failures, under the assumption that failures are identically, but independently distributed. However, the presence of time correlations between failures (such as peak periods with increased failure rate) refutes this assumption and can have a significant impact on the effectiveness of fault-tolerance mechanisms. For example, the performance of a proactive fault-tolerance mechanism is more effective if the failures are periodic or predictable; similarly, the performance of checkpointing, redundancy, and scheduling solutions depends on the frequency of failures. In this study we analyze and model the time-varying behavior of failures in large-scale distributed systems. Our study is based on nineteen failure traces obtained from (mostly) production large-scale distributed systems, including grids, P2P systems, DNS servers, web servers, and desktop grids. We first investigate the time correlation of failures, and find that many of the studied traces exhibit strong daily patterns and high autocorrelation. Then, we derive a model that focuses on the peak failure periods occurring in real large-scale distributed systems. Our model characterizes the duration of peaks, the peak inter-arrival time, the inter-arrival time of failures during the peaks, and the duration of failures during peaks; we determine for each the best-fitting probability distribution from a set of several candidate distributions, and present the parameters of the (best) fit. Last, we validate our model against the nineteen real failure traces, and find that the failures it characterizes are responsible on average for over 50% and up to 95% of the downtime of these systems.

Keywords—time-correlated failures, failure model, fault tolerance, real traces, trace-based analysis.

I. INTRODUCTION

Large-scale distributed systems have reached an unprecedented scale and complexity in recent years. At this scale failures inevitably occur—networks fail, disks crash, packets get lost, bits get flipped, software misbehaves, or systems crash due to misconfiguration and other human errors. Deadline-driven or mission-critical services are part of the typical workload for these infrastructures, which thus need to

be available and reliable despite the presence of failures. Researchers and system designers have already built numerous fault-tolerance mechanisms that have been proven to work under various assumptions about the occurrence and duration of failures. However, most previous work focuses on failure models that assume the failures to be non-correlated, but this may not be realistic for the failures occurring in large-scale distributed systems. For example, such systems may exhibit peak failure periods, during which the failure rate increases, affecting in turn the performance of fault-tolerance solutions. To investigate such time correlations, we perform in this work a detailed investigation of the time-varying behavior of failures using nineteen traces obtained from several large-scale distributed systems including grids, P2P systems, DNS servers, web servers, and desktop grids.

Recent studies report that in production systems, failure rates can be of over one thousand failures per year, and depending on the root cause of the corresponding problems, the mean time to repair can range from hours to days [1]. The increasing scale of deployed distributed systems causes the failure rates to increase, which in turn can have a significant impact on the performance and cost, such as degraded response times [2] and increased Total Cost of Operation (TCO) due to increased administration costs and human resource needs [3]. This situation also motivates the need for further research in failure characterization and modeling.

Previous studies [1], [4]–[8] focused on characterizing failures in several different distributed systems. However, most of these studies assume that failures occur independently or disregard the time correlation of failures, despite the practical importance of these correlations [9]–[11]. First of all, understanding if failures are time correlated has significant implications for proactive fault tolerance solutions. Second, understanding the time-varying behavior of failures and peaks observed in failure patterns is required for evaluating design decisions. For example, redundant submissions may all fail during a failure peak period,

Table I
SUMMARY OF NINETEEN DATA SETS IN THE FAILURE TRACE ARCHIVE.

System	Type	Nodes	Period	Year	Events
GRID ⁵⁰⁰⁰	Grid	1,288	1.5 years	2005-2006	588,463
COND-CAE ¹	Grid	686	35 days	2006	7899
COND-CS ¹	Grid	725	35 days	2006	4543
COND-GLOW ¹	Grid	715	33 days	2006	1001
TERAGRID	Grid	1001	8 months	2006-2007	1999
LRI	Desktop Grid	237	10 days	2005	1,792
DEUG	Desktop Grid	573	9 days	2005	33,060
NOTRE-DAME ²	Desktop Grid	700	6 months	2007	300,241
NOTRE-DAME ³	Desktop Grid	700	6 months	2007	268,202
MICROSOFT	Desktop Grid	51,663	35 days	1999	1,019,765
UCB	Desktop Grid	80	11 days	1994	21,505
PLANETLAB	P2P	200-400	1.5 year	2004-2005	49,164
OVERNET	P2P	3,000	2 weeks	2003	68,892
SKYPE	P2P	4,000	1 month	2005	56,353
WEBSITES	Web servers	129	8 months	2001-2002	95,557
LDNS	DNS servers	62,201	2 weeks	2004	384,991
SDSC	HPC Clusters	207	12 days	2003	6,882
LANL	HPC Clusters	4,750	9 years	1996-2005	43,325
PNNL	HPC Cluster	1,005	4 years	2003-2007	4,650

¹COND-* data sets denote the Condor data sets.

²This is the host availability version which is according to the multi-state availability model of Brent Rood.

³This is the CPU availability version.

regardless of the quality of the resubmission strategy. Third, understanding the temporal correlations and exploiting them for smart checkpointing and scheduling decisions provides new opportunities for enhancing conventional fault-tolerance mechanisms [2], [12]. For example, a simple scheduling policy could be to stop scheduling large parallel jobs during failure peaks. Finally, it is possible to devise adaptive fault-tolerance mechanisms that adjust the policies based on the information related to peaks. For example, an adaptive fault-tolerance mechanism can migrate the computation at the beginning of a predicted peak.

To understand the time-varying behavior of failures in large-scale distributed systems, we perform a detailed investigation using data sets from diverse large-scale distributed systems including more than 100K hosts and 1.2M failure events spanning over 15 years of system operation in total. Our main contribution is threefold:

- 1) We make four new failure traces publicly available through the Failure Trace Archive (Section II).
- 2) We present a detailed evaluation of the time correlation of failure events observed in traces taken from nineteen (production) distributed systems (Section III).
- 3) We propose a model for peaks observed in the failure rate process (Section IV).

II. METHOD

A. Failure Datasets

In this work we use and contribute to the data sets in the Failure Trace Archive (FTA) [13]. The FTA is an online public repository of availability traces taken from diverse parallel and distributed systems.

The FTA makes failure traces available online in a unified format, which records the occurrence time and duration of resource failures as an alternating time series of availability and unavailability intervals. Each availability or unavailability event in a trace records the start and the end of the event, and the resource that was affected by the event. Depending on the trace, the resource affected by the event can be either

a node of a distributed system such as a node in a grid, or a component of a node in a system such as CPU or memory.

Prior to the work leading to this article, the FTA made fifteen failure traces available in its standard format; as a result of our work, the FTA now makes available nineteen failure traces. Table I summarizes the characteristics of these nineteen traces, which we use throughout this work. The traces originate from systems of different types (multi-cluster grids, desktop grids, peer-to-peer systems, DNS and web servers) and sizes (from hundreds to tens of thousands of resources), which makes these traces ideal for a study among different distributed systems. Furthermore, many of the traces cover several months of system operation. A more detailed description of each trace is available on the FTA web site(<http://fta.inria.fr>).

B. Analysis

In our analysis, we use the autocorrelation function (ACF) to measure the degree of correlation of the failure time series data with itself at different time lags. The ACF takes on values between -1 (high negative correlation) and 1 (high positive correlation). In addition, the ACF reveals when the failures are random or periodic. For random data the correlation coefficients will be close to zero; and a periodic component in the ACF reveals that the failure data is periodic or at least it has a periodic component.

C. Modeling

In the modeling phase, we statistically model the peaks observed in the failure rate process, i.e., the number of failure events per time unit. Towards this end we use the Maximum Likelihood Estimation (MLE) method [14] for fitting the probability distributions to the empirical data as it delivers good accuracy for the large data samples specific to failure traces. After we determine the best fits for each candidate distribution for all data sets, we perform the goodness-of-fit tests to assess the quality of the fitting for each distribution, and to establish the best fit. As the goodness-of-fit tests, we use both the Kolmogorov-Smirnov (KS) and the Anderson-Darling (AD) tests, which essentially assess how close the cumulative distribution function (CDF) of the probability distribution is to the CDF of the empirical data. For each candidate distribution with the parameters found during the fitting process, we formulate the hypothesis that the empirical data are derived from it (the null-hypothesis of the goodness-of-fit test). Neither the KS or the AD tests can confirm the null-hypothesis, but both are useful in understanding the goodness-of-fit. For example, the KS-test provides a test statistic, D , which characterizes the maximal distance between the CDF of the empirical distribution of the input data and that of the fitted distribution; distributions with a lower D value across different failure traces are better. Similarly, the tests return p-values which are used to either reject the null-hypothesis

if the p-value is smaller than or equal to the significance level, or confirm that the observation is consistent with the null-hypothesis if the p-value is greater than the significance level. Consistent with the standard method for computing p-values [6], [13], we average 1,000 p-values, each of which is computed by selecting 30 samples randomly from the data set, to calculate the final p-value for the goodness-of-fit tests.

III. ANALYSIS OF AUTOCORRELATION

In this section we present the autocorrelations in failures using traces obtained from grids, desktop grids, P2P systems, web servers, DNS servers and HPC clusters, respectively. We consider the failure rate process, that is the number of failure events per time unit.

A. Failure Autocorrelations in the Traces

Our aim is to investigate whether the occurrence of failures is repetitive in our data sets. Toward this end, we compute the autocorrelation of the failure rate for different time lags including hours, weeks, and months. Figure 1 shows for several platforms the failure rate at different time granularities, and the corresponding autocorrelation functions. Overall, we find that many systems exhibit strong correlation from small to moderate time lags confirming that failures are indeed repetitive in many of the systems.

Many of the systems investigated in this work exhibit strong autocorrelation for hourly and weekly lags. Figures 1(a), 1(b), 1(e), 1(f), and 1(g) show the failure rates and autocorrelation functions for the GRID'5000, CONDOR (CAE), SKYPE, LDNS and LANL systems, respectively. The GRID'5000 data set is a one and a half year long trace collected from an academic research grid. Since this system is mostly used for experimental purposes, and is large-scale ($\sim 3K$ processors) the failure rate is quite high. In addition, since most of the jobs are submitted through the OAR resource manager, hence without direct user interaction, the daily pattern is not clearly observable. However, during the summer the failure rate decreases, which indicates a correlation between the system load and the failure rate. Finally, as the system size increases over the years, the failure rate does not increase significantly, which indicates system stability. The CONDOR (CAE) data set is a one month long trace collected from a desktop grid using the Condor cycle-stealing scheduler. As expected from a desktop grid, this trace exhibits daily peaks in the failure rate, and hence in the autocorrelation function. In contrast to other desktop grids, the failure rate is lower. The SKYPE data set is a one month long trace collected from a P2P system used by 4,000 clients. Clients may join or leave the system, and clients that are not online are considered as unavailable in this trace. Similar to desktop grids, there is high autocorrelation at small time lags, and the daily and weekly peaks are more pronounced. The LDNS data set is a two week long trace collected from DNS servers.

Figure 2. Daily and hourly failure rates for GRID'5000 and PLANETLAB platforms.

Unlike P2P systems and desktop grids, DNS servers do not exhibit strong autocorrelation for short time lags with periodic behavior. In addition, as the workload intensity increases during the peak hours of the day, we observe that the failure rate also increases. Finally, the LANL data set is a ten year long trace collected from production HPC clusters. The weekly failure rate is quite low compared to GRID'5000. We do not observe a clear yearly pattern as the failure rate increases during summer 2002, whereas the failure rate decreases during summer 2004. Since around 3,000 nodes were added to the LANL system between 2002 and 2003, the failure rate also increases correspondingly.

Last, a few systems exhibit weak autocorrelation in failure occurrence. Figure 1(c) and 1(d) show the failure rate and the corresponding autocorrelation function for the TERA-GRID and NOTRE-DAME systems. The TERAGRID data set is an eight month long trace collected from an HPC cluster that is part of a grid. We observe weak autocorrelation at all time lags, which implies that the failure rates observed over time are independent. In addition, there are no clear hourly or daily patterns, which gives evidence of an erratic occurrence of failures in this system. The NOTRE-DAME data set is a six month long trace collected from a desktop grid. The failure events in this data set consist of the availability/unavailability events of the hosts in this system. Similar to other desktop grids, we observe clear daily and weekly patterns. However, the autocorrelation is low when compared to other desktop grids.

B. Discussion

As we have shown in the previous section, many systems exhibit strong correlation from small to moderate time lags, which indicates probably a high degree of predictability. In contrast, a small number of systems (NOTRE-DAME, PNNL, and TERAGRID) exhibit weak autocorrelation; only for these systems, the failure rates observed over time are independent.

Figure 1. Failure rates at different time granularities for several platforms and the corresponding autocorrelation functions.

We have found that similar systems have similar time-varying behavior, e.g., desktop grids and P2P systems have daily and weekly periodic failure rates, and these systems exhibit strong temporal correlation at hourly time lags. Some systems (NOTRE-DAME and CONDOR (CAE)) have direct user interaction, which produces clear daily and weekly patterns in both system load and occurrence of failures—

the failure rate increases during work hours and days, and decreases during free days and holidays (the summer).

Finally, not all systems exhibit a correlation between work hours and days, and the failure rate. In the examples depicted in Figure 2, while GRID'5000 exhibits this correlation, PLANETLAB exhibit irregular/erratic hourly and daily failure behavior.

Figure 3. Parameters of the peak periods model. The numbers in the figure match the (numbered) model parameters in the text.

Our results are consistent with previous studies [7], [15]–[17] as in many traces we observe strong autocorrelation at small time lags, and that we observe correlation between the intensity of the workload and failure rates.

IV. MODELING THE PEAKS OF FAILURES

In this section we present a model for the peaks observed in the failure rate process in diverse large-scale distributed systems.

A. Peak Periods Model

Our model of peak failure periods comprises four parameters as shown in Figure 3: the peak duration, the time between peaks (inter-peak time), the inter-arrival time of failures during peaks, and the duration of failures during peaks:

- 1) **Peak Duration:** The duration of peaks observed in a data set.
- 2) **Time Between Peaks (inter-peak time):** The time from the end of a previous peak to the start of the next peak.
- 3) **Inter-arrival Time of Failures During Peaks:** The inter-arrival time of failure events that occur during peaks.
- 4) **Failure Duration During Peaks:** The duration of failure events that start during peaks. These failure events may last longer than a peak.

Our modeling process is based on analyzing the failure traces taken from real distributed systems in two steps which we describe in turn.

The first step is to identify for each trace the peaks of hourly failure rates. Since there is no rigorous mathematical definition of peaks in time-series, to identify the peaks we define a threshold value as $\mu + k\sigma$, where μ is the average and σ is the standard deviation of the failure rate, and k is a positive integer; a period with a failure rate above the threshold is a *peak period*. We adopt this threshold to achieve a good balance between capturing in the model

Table II
AVERAGE VALUES FOR THE MODEL PARAMETERS.

System	Avg. Peak Duration [s]	Avg. Failure IAT During Peaks [s]	Avg. Time Between Peaks [s]	Avg. Failure Duration During Peaks [s]
GRID'5000	5,047	13	55,101	20,984
CONDOR (CAE)	5,287	23	87,561	4,397
CONDOR (CS)	3,927	4	241,920	20,740
CONDOR (GLOW)	4,200	14	329,040	75,672
TERAGRID	3,680	35	526,500	368,903
LRI	4,080	78	58,371	31,931
DEUG	14,914	10	103,800	1,091
NOTRE-DAME	3,942	21	257,922	280,593
NOTRE-DAME (CPU)	7,520	33	47,075	22,091
MICROSOFT	7,200	0	75,315	90,116
UCB	21,272	23	77,040	332
PLANETLAB	4,810	264	47,124	166,913
OVERNET	3,600	1	14,400	382,225
SKYPE	4,254	11	112,971	26,402
WEBSITES	5,211	103	104,400	3476
LDNS	4,841	8	42,042	30,212
SDSC	4,984	26	84,900	6,114
LANL	4,122	653	94,968	21,193

extreme system behavior, and characterizing with our model an important part of the system failures (either number of failures or downtime caused to the system). A threshold excluding all but a few periods, for example defining peak periods as distributional outliers, may capture too few periods and explain only a small fraction of the system failures. A more inclusive threshold would lead to the inclusion of more failures, but the data may come from periods with very different characteristics, which is contrary to the goal of building a model for peak failure periods.

In the second step we extract the model parameters from the data sets using the peaks that we identified in the first step. Then we try to find a good fit, that is, a well-known probability distribution and the parameters that lead to the best fit between that distribution and the empirical data. When selecting the probability distributions, we consider the degrees of freedom (number of parameters) of that distribution. Although a distribution with more degrees of freedom may provide a better fit for the data, such a distribution can result in a complex model, and hence it may be difficult to analyze the model mathematically. In this study we use five probability distributions to fit to the empirical data: exponential, Weibull, Pareto, lognormal, and gamma. For the modeling process, we follow the methodology that we describe in Section II-C.

B. Results

After applying the modeling methodology that we describe in the previous section and Section II-C, in this section we present the peak model that we derived from diverse large scale distributed systems.

Table II shows the average values for all the model parameters for all platforms. The average peak duration varies across different systems, and even for the same type of systems. For example, UCB, MICROSOFT and DEUG are all desktop grids, but the average peak duration widely varies among these platforms. In contrast, for the SDSC, LANL, and PNNL platforms, which are HPC clusters, the average peak duration values are relatively close. The DEUG and UCB platforms have small number of long peak durations

Table III

EMPIRICAL DISTRIBUTION FOR THE PEAK DURATION PARAMETER. h DENOTES HOURS. VALUES ABOVE 10% ARE DEPICTED AS BOLD.

Platform / Peak Duration	1h	2h	3h	4h	5h	6h	$\geq 7h$
GRID'5000	80.56 %	13.53 %	3.38 %	1.33 %	0.24 %	0.12 %	0.84%
CONDOR (CAE)	93.75 %	3.13 %	—	—	—	—	3.12%
CONDOR (CS)	90.91 %	9.09 %	—	—	—	—	—
CONDOR (GLOW)	83.33 %	16.67 %	—	—	—	—	—
TERAGRID	97.78 %	2.22 %	—	—	—	—	—
LRI	86.67 %	13.33 %	—	—	—	—	—
DEUG	28.57 %	—	28.57 %	—	14.29 %	—	28.57 %
NOTRE-DAME	90.48 %	9.52 %	—	—	—	—	—
NOTRE-DAME (CPU)	56.83 %	17.78 %	9.84 %	3.49 %	5.40 %	3.49 %	3.17
MICROSOFT	35.90 %	33.33 %	25.64 %	5.13 %	—	—	—
UCB	9.09 %	9.09 %	—	—	—	9.09 %	72.73 %
PLANETLAB	80.17 %	13.36 %	3.71 %	1.27 %	0.53 %	0.53 %	0.43
OVERNET	100.00 %	—	—	—	—	—	—
SKYPE	90.91 %	4.55 %	—	4.55 %	—	—	—
WEBSITES	76.74 %	13.95 %	5.23 %	2.33 %	0.58 %	—	1.16
LDNS	75.86 %	13.79 %	10.34 %	—	—	—	—
SDSC	69.23 %	23.08 %	7.69 %	—	—	—	—
LANL	88.35 %	9.24 %	2.06 %	0.25 %	0.06 %	0.03 %	—
PNNL	85.99 %	10.35 %	1.75 %	0.96 %	0.64 %	0.16 %	0.16 %
Avg	74.79 %	11.3 %	5.16 %	1.01 %	1.14 %	0.7 %	5.85 %

resulting in higher average peak durations compared to the other platforms. Finally, as there are two peaks of zero length (single data point) in the OVERNET system, the average peak duration is zero.

The average inter-arrival time during peaks is rather low, as expected, as the failure rates are higher during peaks compared to off-peak periods. For the MICROSOFT platform, as all failures arrive as burst during peaks, average inter-arrival time during peaks is zero.

Similar to the average peak duration parameter, the average time between peaks parameter is also highly variable across different systems. For some systems like TERAGRID, this parameter is in the order of days, and for some systems like OVERNET it is in the order of hours.

Similarly, the duration of failures during peaks highly varies even across similar platforms. For example, the difference between the average failure duration during peaks between the UCB and the MICROSOFT platforms, which are both desktop grids, is huge because the machines in the UCB platform leave the system less often than the machines in the MICROSOFT platform. In addition, in some platforms like OVERNET and TERAGRID, the average failure durations during peaks is in the order of days showing the impact of space-correlated failures, that is multiple nodes failing nearly simultaneously.

Using the AD and KS tests we next determine the best fitting distributions for each model parameter and each system. Since we determine the hourly failure rates using fixed time windows of one hour, the peak duration and the inter-peak time are multiples of one hour. In addition, as the peak duration parameter is mostly in the range $[1h-5h]$, and for several systems this parameter is mostly $1h$ causing the empirical distribution to have a peak at $1h$, none of the distributions provide a good fit for the peak duration parameter. Therefore, for the peak duration model parameter, we only present an empirical histogram in Table III. We find that the peak duration for almost all platforms are less than $3h$.

Table IV shows the best fitting distributions for the model

Table IV

PEAK MODEL: THE PARAMETER VALUES FOR THE BEST FITTING DISTRIBUTIONS FOR ALL STUDIED SYSTEMS. E,W,LN, AND G STAND FOR EXPONENTIAL, WEIBULL, LOGNORMAL AND GAMMA DISTRIBUTIONS, RESPECTIVELY.

System	IAT During Peaks	Time Between Peaks	Failure Duration During Peaks
GRID'5000	— (see text)	LN(10.30,1.09)	— (see text)
CONDOR (CAE)	— (see text)	— (see text)	— (see text)
CONDOR (CS)	— (see text)	— (see text)	— (see text)
CONDOR (GLOW)	— (see text)	— (see text)	— (see text)
TERAGRID	— (see text)	LN(12.40,1.42)	LN(10.27,1.90)
LRI	LN(3.49,1.86)	LN(10.51,0.98)	— (see text)
DEUG	W(9.83,0.95)	— (see text)	LN(5.46,1.29)
NOTRE-DAME	— (see text)	W(247065.52,0.92)	LN(9.06,2.73)
NOTRE-DAME (CPU)	— (see text)	W(44139.20,0.89)	LN(7.19,1.35)
MICROSOFT	— (see text)	G(1.50,50065.81)	W(55594.48,0.61)
UCB	E(23.77)	— (see text)	LN(5.25,0.99)
PLANETLAB	— (see text)	LN(10.13,1.03)	LN(8.47,2.50)
OVERNET	— (see text)	— (see text)	— (see text)
SKYPE	— (see text)	W(123440.05,1.37)	— (see text)
WEBSITES	W(66.61,0.60)	LN(10.77,1.25)	— (see text)
LDNS	W(8.97,0.98)	LN(10.38,0.79)	LN(9.09,1.63)
SDSC	W(16.27,0.46)	E(84900)	LN(7.59,1.68)
LANL	G(1.35,797.42)	LN(10.63,1.16)	LN(8.26,1.53)
PNNL	— (see text)	E(160237.32)	— (see text)
Avg	W(193.91,0.83)	LN(10.89,1.08)	LN(8.09,1.59)

Table V

THE AVERAGE DURATION AND AVERAGE IAT OF FAILURES FOR THE ENTIRE TRACES AND FOR THE PEAKS.

System	Avg. Failure Duration [h] (Entire)	Avg. Failure Duration [h] (Peaks)	Avg. Failure IAT [s] (Entire)	Avg. Failure IAT [s] (Peaks)
GRID'5000	7.41	5.83	160	13
NOTRE-DAME (CPU)	4.25	6.14	119	33
MICROSOFT	16.49	25.03	6	0
PLANETLAB	49.61	46.36	1880	264
OVERNET	11.98	106.17	17	1
SKYPE	14.30	7.33	91	11
WEBSITES	1.17	0.97	380	103
LDNS	8.61	8.39	12	8
LANL	5.88	5.89	13874	653

parameters for all data sets investigated in this study. To generate synthetic yet realistic traces without using a single system as a reference, we create the *average system model* which has the average characteristics of all systems we investigate. We create the average system model as follows. First, we determine the *candidate distributions* for a model parameter with the distributions having the smallest D values for each system. Then, for each model parameter, we determine the best fitting distribution among the candidate distributions that has the lowest average D value over all data sets. After we determine the best fitting distribution for the average system model, each data set is fit independently to this distribution to find the set of best fit parameters. The parameters of the average system model shown in the "Avg." row represent the average of this set of parameters.

For the IAT during peak durations, several platforms do not have a best fitting distribution since for these platforms most of the failures during peaks occur as bursts hence having inter-arrival times of zero. Similarly, for the time between peaks parameter, some platforms (like all CONDOR platforms, DEUG, OVERNET and UCB platforms) do not have best fitting distributions since these platforms have inadequate number of samples to generate a meaningful model. For the failure duration between peaks parameter, some platforms do not have a best fitting distribution due to the nature of the data. For example, for all CONDOR platforms the failure duration is a multiple of a monitoring interval creating peaks in the empirical distribution at that

monitoring interval. As a result, none of the distributions we investigate provide a good fit.

In our model we find that the model parameters do not follow a heavy-tailed distribution since the p-values for the Pareto distribution are very low. For the IAT during peaks parameter, Weibull distribution provides a good fit for most of the platforms. For the time between peaks parameter, we find that the platforms can either be modeled by the lognormal distribution or the Weibull distribution. Similar to our previous model [13], which is derived from both peak and off-peak periods, for the failure duration during peaks parameter, we find that the lognormal distribution provides a good fit for most of the platforms. To conclude, for all the model parameters, we find that either the lognormal or the Weibull distributions provide a good fit for the average system model.

Similar to the average system models built for other systems [18], we cannot claim that our average system model represents the failure behavior of an actual system. However, the main strength of the average system model is that it represents a common basis for the traces from which it has been extracted. To generate failure traces for a specific system, individual best fitting distributions and their parameters shown in Table IV may be used instead of the average system.

Next, we compute the average failure duration/inter-arrival time over each data set and only during peaks (Table V). We compare only the data sets used both in this study and our previous study [13], where we modelled each data set individually without isolating peaks. We observe that the average failure duration per data set can be twice as long as the average duration during peaks. In addition, the average failure inter-arrival time per data set is on average nine times the average failure inter-arrival time during peaks. This implies that the distribution per data set is significantly different from the distribution for peaks, and that fault detection mechanisms must be significantly faster during peaks. Likewise, fault-tolerance mechanisms during peaks must have considerably lower overhead than during non-peak periods.

Finally, we investigate the fraction of downtime caused by failures that originate during peaks, and the fraction of the number of failures that originate during peaks (Table VI). In this table we present the results for only $k = 1$ for the GRID'5000 and NOTRE-DAME data sets and the average fraction over all systems; the results for several other values for k and for all systems can be found in [19]. We find that on average over 50% and up to 95% of the downtime of the systems we investigate are caused by the failures that originate during peaks.

V. RELATED WORK

Much work has been dedicated to characterizing and modeling system failures [1], [4]–[8]. While the correlation

Table VI
FRACTION OF DOWNTIME AND FRACTION OF NUMBER OF FAILURES
DUE TO FAILURES THAT ORIGINATE DURING PEAKS ($k = 1$).

System	Downtime %	Failures %
GRID'5000	49.19 %	62.56 %
NOTRE-DAME	53.40 %	69.09 %
Avg.	49.88 %	52.35 %

among failure events has received attention since the early 1990s [4], previous studies focus mostly on *space-correlated* failures, that is, on multiple nodes failing nearly simultaneously. Although the *time correlation* of failure events deserve a detailed investigation due to its practical importance [9]–[11], relatively little attention has been given to characterize the time correlation of failures in distributed systems. Our work is the first to investigate the time correlation between failure events across a broad spectrum of large-scale distributed systems. In addition, we also propose a model for peaks observed in the failure rate process derived from several distributed systems.

Previous failure studies [4]–[8] used few data sets or even data from a single system; their data also span relatively short periods of time. In contrast, we perform a detailed investigation using data sets from diverse large-scale distributed systems including more than 100K hosts and 1.2M failure events spanning over 15 years of system operation.

In our recent work [20], we proposed a model for space-correlated failures using fifteen FTA data sets, and we showed that space-correlated failures are dominant in most of the systems that we investigated. In this work, we extend our recent work with a detailed time-correlation analysis, and we propose a model for peak periods.

Closest to our work, Schroeder and Gibson [1] present an analysis using a large set of failure data obtained from a high performance computing site. However, this study lacks a time correlation analysis and focuses on well-known failure characteristics like MTTF and MTTR. Sahoo et al. [7] analyze one year long failure data obtained from a single cluster. Similar to the results of our analysis, they report that there is strong correlation with significant periodic behavior. Bhagwan et al. [21] present a characterization of the availability of the OVERNET P2P system. Their and other studies [22] show that the availability of P2P systems has diurnal patterns, but do not characterize the time correlations of failure events.

Traditional failure analysis studies [15], [16] report strong correlation between the intensity of the workload and failure rates. Our analysis brings further evidence supporting the existence of this correlation—we observe more failures during peak hours of the day and during work days in most of the (interactive) traces.

VI. CONCLUSION

In the era of cloud and peer-to-peer computing, large-scale distributed systems now consist of hundreds of thousands

of nodes. At this scale, providing high availability is a challenge, and overcoming it depends on the development of efficient fault-tolerance mechanisms. To develop new and improve existing fault-tolerance mechanisms, we need realistic models of the failures occurring in large-scale distributed systems. Traditional failure models in distributed systems were derived from small scale systems and often under the assumption of independence between failures. However, recent studies have shown evidence that there exist time patterns and other time-varying behavior in the occurrence of failures. Thus, we have investigated in this work the time-varying behavior of failures in large-scale distributed systems, and proposed and validated a model for time-correlated failures in such systems.

First, we have assessed the presence of time-correlated failures, using traces from nineteen (production) systems, including grids, P2P systems, DNS servers, web servers, and desktop grids. We found for most of the studied systems that, while the failure rates are highly variable, the failures still exhibit strong periodic behavior and time correlation.

Second, to characterize the periodic behavior of failures and the peaks in failures, we have proposed a peak model with four attributes: the peak duration, the failure inter-arrival time during peaks, the time between peaks, and the failure duration during peaks. We found that the peak failure periods explained by our model are responsible for on average for over 50% and up to 95% of the system downtime. We also found that the Weibull and the lognormal distributions provide good fits for the model attributes. We have provided best-fitting parameters for these distributions which will be useful to the community when designing and evaluating fault-tolerance mechanisms in large-scale distributed systems.

Last but not least, we have made four new traces available, three Condor traces and one TeraGrid trace, in the Failure Trace Archive, which we hope will encourage others to use the archive and also to contribute to it with new failure traces.

REFERENCES

- [1] B. Schroeder and G. A. Gibson, "A large-scale study of failures in high-performance computing systems," in *DSN '06*, 2006, pp. 249–258.
- [2] Y. Zhang, M. S. Squillante, and R. K. Sahoo, "Performance implications of failures in large-scale cluster scheduling," in *JSSPP*, 2004, pp. 233–252.
- [3] "The uc berkeley/stanford recovery-oriented computing (roc) project," <http://roc.cs.berkeley.edu/>.
- [4] D. Tang, R. Iyer, and S. Subramani, "Failure analysis and modeling of a vaxcluster system," jun 1990, pp. 244–251.
- [5] D. Oppenheimer, A. Ganapathi, and D. A. Patterson, "Why do internet services fail, and what can be done about it?" in *USITS'03*. USENIX Association, 2003, pp. 1–1.
- [6] D. Nurmi, J. Brevik, and R. Wolski, "Modeling machine availability in enterprise and wide-area distributed computing environments," in *Euro-Par*, 2005, pp. 432–441.
- [7] R. K. Sahoo, A. Sivasubramaniam, M. S. Squillante, and Y. Zhang, "Failure data analysis of a large-scale heterogeneous server environment," in *DSN '04*, 2004, p. 772.
- [8] J. Xu, Z. Kalbarczyk, and R. K. Iyer, "Networked windows nt system field failure data analysis," in *PRDC '99*, 1999, p. 178.
- [9] J. W. Mickens and B. D. Noble, "Exploiting availability prediction in distributed systems," in *NSDI'06*. USENIX Association, 2006, pp. 6–6.
- [10] K. Vaidyanathan, R. E. Harper, S. W. Hunter, and K. S. Trivedi, "Analysis and implementation of software rejuvenation in cluster systems," in *SIGMETRICS '01*. ACM, 2001, pp. 62–71.
- [11] F. Salfner, M. Lenk, and M. Malek, "A survey of online failure prediction methods," *ACM Comput. Surv.*, vol. 42, no. 3, pp. 1–42, 2010.
- [12] T. Z. Islam, S. Bagchi, and R. Eigenmann, "Falcon: a system for reliable checkpoint recovery in shared grid environments," in *SC '09*. ACM, 2009, pp. 1–12.
- [13] D. Kondo, B. Javadi, A. Iosup, and D. Epema, "The Failure Trace Archive: Enabling comparative analysis of failures in diverse distributed systems," in *CCGRID*, 2010, pp. 1–10.
- [14] J. Aldrich, "R. A. Fisher and the making of maximum likelihood 1912-1922," *Statistical Science*, vol. 12, no. 3, pp. 162–176, 1997.
- [15] X. Castillo and D. Siewiorek, "Workload, performance, and reliability of digital computing systems," jun 1981, pp. 84–89.
- [16] R. K. Iyer, D. J. Rossetti, and M. C. Hsueh, "Measurement and modeling of computer reliability as affected by system activity," *ACM Trans. Comput. Syst.*, vol. 4, no. 3, pp. 214–237, 1986.
- [17] B. Schroeder and G. A. Gibson, "Disk failures in the real world: what does an mttf of 1,000,000 hours mean to you?" in *FAST '07*. USENIX Association, 2007, p. 1.
- [18] U. Lublin and D. G. Feitelson, "The workload on parallel supercomputers: modeling the characteristics of rigid jobs," *J. Parallel Distrib. Comput.*, vol. 63, no. 11, pp. 1105–1122, 2003.
- [19] N. Yigitbasi, M. Gallet, D. Kondo, A. Iosup, and D. Epema, "Analysis and modeling of time-correlated failures in large-scale distributed systems," Delft University of Technology, Technical Report PDS-2010-004, May 2010. [Online]. Available: <http://pds.twi.tudelft.nl/reports/2010/PDS-2010-004.pdf>
- [20] M. Gallet, N. Yigitbasi, B. Javadi, D. Kondo, A. Iosup, and D. Epema, "A model for space-correlated failures in large-scale distributed systems," in *Euro-Par*, 2010, to appear. [Online]. Available: <http://www.st.ewi.tudelft.nl/~nezih/doc/space-correlated-europar10.pdf>
- [21] R. Bhagwan, S. Savage, and G. M. Voelker, "Understanding availability," in *IPTPS*, ser. Lecture Notes in Computer Science, M. F. Kaashoek and I. Stoica, Eds., vol. 2735. Springer, 2003, pp. 256–267.
- [22] J. Chu, K. Labonte, and B. N. Levine, "Availability and locality measurements of peer-to-peer file systems," in *Proceedings of ITCOM: Scalability and Traffic Control in IP Networks*, 2002.