A Model for Space-Correlated Failures in Large-Scale Distributed Systems

Matthieu Gallet[1,3], Nezih Yigitbasi[1,3], Bahman Javadi[2,3],
Derrick Kondo[2,3], Alexandru Iosup[1,3], and Dick Epema[1,3]

[1] Delft University of Technology, the Netherlands, [2] INRIA Grenoble, France
[3] The Failure Trace Archive. Web: `fta.inria.fr` Email: `contact@fta.inria.fr`

Abstract. Distributed systems such as grids, peer-to-peer systems, and
even Internet DNS servers have grown significantly in size and complexity
in the last decade. This rapid growth has allowed distributed systems to
serve a large and increasing number of users, but has also made resource
and system failures inevitable. Moreover, perhaps as a result of system
complexity, in distributed systems a single failure can trigger within a
short time span several more failures, forming a group of time-correlated
failures. To eliminate or alleviate the significant effects of failures on per-
formance and functionality, the techniques for dealing with failures re-
quire good failure models. However, not many such models are available,
and the available models are valid for few or even a single distributed sys-
tem. In contrast, in this work we propose a model that considers groups
of time-correlated failures and is valid for many types of distributed sys-
tems. Our model includes three components, the group size, the group
inter-arrival time, and the resource downtime caused by the group. To
validate this model, we use failure traces corresponding to fifteen dis-
tributed systems. We find that space-correlated failures are dominant in
terms of resource downtime in seven of the fifteen studied systems. For
each of these seven systems, we provide a set of model parameters that
can be used in research studies or for tuning distributed systems. Last,
as a result of our work six of the studied traces have been made available
through the Failure Trace Archive (`http://fta.inria.fr`).

1 Introduction

Millions of people rely daily on the availability of distributed systems such as
peer-to-peer file-sharing networks, grids, and the Internet. Since the scale and
complexity of contemporary distributed systems make the occurrence of failures
the rule rather than the exception, many fault tolerant resource management
techniques have been designed recently [1–3]. The deployment of these tech-
niques and the design of new ones depend on understanding the characteristics
of failures in real systems. While many failure models have been proposed for
various computer systems [3–6], few consider the occurrence of failure bursts. In
this work we present a new model that focuses on failure bursts, and validate it
with real failure traces coming from a diverse set of distributed systems.

The foundational work on the failures of computer systems [4, 7–9] has al-
ready revealed that computer system failures occur often in bursts, that is, the

occurrence of a failure of a system component can trigger within a short period a sequence of failures in other components of the system. It turned out that the fraction of bursty system failures is high in distributed systems; for example, in the VAXcluster 58% of all errors and occurred in bursts and involved multiple machines [4], and in both the VAXcluster and in Grid'5000 about 30% of all failures involve multiple machines [4, 10].

A bursty arrival breaks an important assumption made by numerous fault tolerant algorithms [1, 11, 12], that of independent and identical distribution of failures among the components of the system. However, few studies [4, 10, 13] investigate the bursty arrival of failures for distributed systems. Even for these studies, the findings are based on data corresponding to a single system–until the recent creation of online repositories such as the failure Failure Trace Archive [14] and the Computer Failure Data Repository [5], failure data for distributed systems were largely inaccessible to the researchers in this area.

The occurrence of failure bursts often makes the availability behavior of different system components to be correlated; thus, they are often referred to as component or *space-correlated failures*. The importance of space-correlated failures has been repeatedly noted: the availability of a distributed system may be overestimated by an order of magnitude when as few as 10% of the failures are correlated [4], and a halving of the work loss may be achieved when taking into account space-correlated failures [11].

This work addresses both scarcity problems, of the lack of traces, and of the lack of a model for space-correlated failures, with the following contributions:
 1. We make publicly and freely available through the Failure Trace Archive six new traces in standard format (Section 2);
 2. We propose a novel model for space-correlated failures based on moving windows (Sections 3);
 3. We propose and validate a fully automated method for identifying space-correlated failures (Sections 3 and 4, respectively). The validation uses failure traces taken from fifteen diverse distributed systems;
 4. We validate our model using real failure traces, and present for them the extracted model parameters (Section 5).

2   Background

In this section we present the terminology and the datasets used in this work.

2.1   Terminology

We follow throughout this work the basic concepts and definitions associated with system dependability as summarized by Avizienis et al. [15]. The basic threats to reliability are failures, errors, and faults occurring in the system. A *failure (unavailability event)* is an event in which the system fails to operate according to its specifications. A failure is observed as a deviation from the correct state of the system. An *error* is part of the system state that may lead to a failure. An *availability event* is the end of the recovery of the system from failure. As in our previous work [14], we define an *unavailability interval (downtime)* as a continuous period of a service outage due to a failure. Conversely, we define an *availability interval* as a contiguous period of service availability.
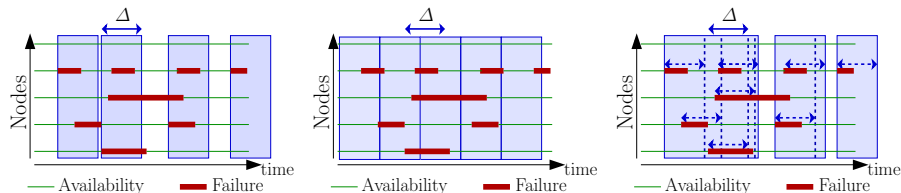
Fig. 1. Generative processes for space-correlated failures: (left) moving windows; (middle) time partitioning; (right) extending windows.

3.1 Space-Correlated Failures

We call *space-correlated failures* a groups of failures that occur within a short time interval; the seminal work of Siewiorek [7, 16], Iyer [4, 8], and Gray [9, 17] has shown that for tightly coupled systems space-correlated failures are likely to occur. Our investigation of space-correlated failures is hampered by the lack of information present in failure traces—none of the computer system failure traces we know records failures with sufficient detail to reconstruct groups of failures. We adopt instead a numeric approach that groups failures based on their start and finish timestamps. We identify three such approaches, moving windows, time partitioning, and extending windows, which we describe in turn.

Let $TS(\cdot)$ be the function that returns the time stamp of an event, either failure or repair. Let O be the sequence of failure events ordered according to increasing event time stamp, that is, $O = [E_i | TS(E_{i-1}) \leq TS(E_i), \forall i \geq 1]$.

Moving Windows We consider the following iterative process that, starting from O, generates the space-correlated failures with time parameter Δ. At each step in the process we select as the *group generator* F the first event from O unselected yet, and generate the space-correlated failure by further selecting from O all events E occurring within Δ time units from $TS(F)$, that is, $TS(E) \leq TS(F) + \Delta$. The process we employ ends when all the events in O have been selected. The maximum number of generated space-correlated failures is $|O|$, the number of events in O. The process uses a time window of size Δ, where the window "moves" to the next unselected event in O at each step. Thus, we call this process the generation of space-correlated failures through *moving windows*. Figure 1 (left) depicts the use of the moving windows for various values of Δ.

Time Partitioning This approach partitions time in windows of fixed size Δ, starting either from a hypothetical time 0 or from the first event in O. We call this process generation of space-correlated failures through *time partitioning*.

Extending Windows A group of failures in this approach is a maximal subsequence of events such that each two consecutive events are at most a time Δ apart, i.e., for each consecutive events E and F in O, $TS(F) \leq TS(E) + \Delta$. Thus, Δ is the size of the window that extends the horizon for each new event added to the group; thus, we call this second process generation of space-correlated failures through *extending windows*. We have already used this process to model the failures occurring in Grid'5000 [10].

The three generation processes, moving windows, time partitioning, and extending windows, can generate very different space-correlated failures from the
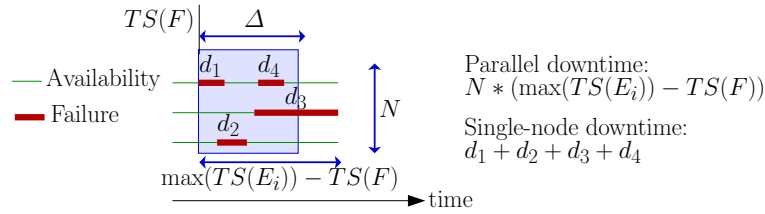
Fig. 2. Parallel and single-node job downtime for a hypothetical space-correlated failure.

same input set of events O (see Figure 1). The following two considerations motivate our selection of a single generation process from these three. First, time partitioning may introduce artificial time boundaries between failure events belonging to consecutive space-correlated failures, because each space-correlated failure starts at a multiple of Δ. Thus, the groups identified through time partitioning do not relate well to groups naturally occurring in the system, and may confuse the fault-tolerant mechanisms and algorithms based on them; the moving and extending windows do not suffer from this problem. Second, the extending windows process may generate infinitely-long space-correlated failures: as the extending window is considered between consecutive failures, a failure can occur long after its group generator (its first occurring failure). Thus, the groups generated through extending windows may reduce the efficiency of fault tolerance mechanisms that react to instantaneous bursts of failures. Thus, we select and use in the remainder of this work the generative processes for space-correlated failures through moving windows.

3.2 Model Components

We now build our model around the notion of space-correlated failures (groups) introduced in the previous section. The model comprises three components: the group inter-arrival time, the group size, and the group downtime. We describe each of these three components in turn.

Inter-Arrival Time This component characterizes the process governing the arrival of new space-correlated failures (including groups of size 1).

Size This component characterizes the number of failures present in each space-correlated failure.

Downtime This component characterizes the downtime caused by each space-correlated failure. When failures are considered independently instead of in groups, the downtime is simply the duration of the unavailability corresponding to each failure event. A group of failure may, however, affect users in ways that depend on the user application. We consider in this work two types of user applications: parallel jobs and single-node jobs. We define the *parallel job downtime* of a failure group as the product of the number of individual

nodes affected by the failure events within the group, and the time elapsed between the earliest failure event and the latest availability event corresponding to a failure within the group. We further define the *single-node job downtime* as the sum of the downtimes of each individual failure within the failure group. Figure 2 depicts these two downtime notions. The parallel job downtime gives an upper bound to the downtime caused by space-correlated failures for parallel jobs that would run on any of the nodes affected by failures. Similarly, the single-node job downtime characterizes the impact of a failure group on workloads dominated by single-node jobs, which is the case for many grid workloads [6].

3.3 Method for Modeling

Our method for modeling is based on analyzing in two steps failure traces taken from real distributed systems; we describe each step, in turn, in the following.

The first step is to analyze for each trace the presence of space-correlated failures comprising two or more failure events, for values of Δ between 1 second and 10 minutes. Tolerating such groups of failures is important for interactive and deadline-oriented system users.

The second step follows the traditional modeling steps for failures in computer systems [5,8]. We first characterize the properties of the empirical distributions using basic statistics such as the mean, the standard deviation, the min and the max, etc. This allows us to get a first glimpse of the type of probability distribution that could characterize the real data. We then 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 look at the degrees of freedom (number of parameters) of that distribution; while a distribution with more degrees of freedom may provide a better fit for the data, such a distribution can make the understanding of the model more difficult, can increase the difficulty of mathematical analysis based on the model, and may also lead to overfitting to the empirical datasets. Thus, we select five probability distributions to fit to the empirical data: exponential, Weibull, Pareto, lognormal, and gamma. The fitting of the probability distributions to the empirical datasets uses the Maximum Likelihood Estimation (MLE) method [18], which delivers good accuracy for the large data samples specific to failure traces.

After finding the best fits for each candidate distribution, goodness-of-fit tests are used to assess the quality of the fitting for each distribution, and to establish the best fit. We use for this purpose 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 of the KS and AD tests can confirm the null-hypothesis, but both are useful in understanding the goodness-of-fit. For example, the KS-test

Table 2. Selected failure group window size for each system.

Platform	Grid'5000	Websites	LDNS	LRI	Deug	SDSC	UCB
Window Size [s]	250	100	150	100	150	120	80

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.

4   Failure Group Window Size

An important assumption in this work is that space-correlated failures are present and significant in the failure traces of distributed systems. In this section we show that this is indeed the case. Section 3.1 the characteristics of the space-correlated failures are dependent on the window size Δ; we investigate this dependency in this section.

The importance of a failure model derives from the fraction of downtime caused by the failures whose characteristics it explains, from the total downtime of the system. For the model we have introduced in Section 3 we are interested in space-correlated failures of at least two failures. As explained in Section 3.1, the characteristics of the space-correlated failures depend on the window size Δ. Large values for Δ lead to more groups of at least two failures, but reduce the usefulness of the model for predictive fault tolerance. Conversely, small values for Δ lead to few groups of at least two failures, and effectively convert our model into the model for individual failures we have investigated elsewhere [14].

We assess the effect of Δ on the number of and downtime caused by space-correlated failures by varying Δ from one second to one hour; the most interesting values for Δ are below a few minutes, useful for proactive fault tolerance techniques. Figure 3 shows the results for each of the fifteen datasets (see Section 2.2). We distinguish in the figure the first seven systems, Grid'5000, SDSC, UCB, Deug, LDNS, LRI, and Websites, for which a significant fraction of the total system downtime is caused by space-correlated failures of size at least 2, when Δ is equal to a few minutes. For similar values of Δ, the space-correlated failures do not cause most of the system downtime for the remaining systems. We do not include in the distinguished systems Notre-Dame, Overnet and Microsoft and Skype, since the dependence of the depicted curves on Δ looks more like an artifact of the data, due to the regular probing of nodes.

The seven distinguished traces have similar dependency on Δ: as Δ increases slowly, the number of groups quickly decreases and the cumulative downtime quickly increases. Then, both slowly stabilize; this point, which occurs for values of Δ of a few minutes, is a good trade-off between small window size and large capture of failures into groups. We extract for each of the seven selected traces the best observed trade-off, and round it to the next multiple of 10 seconds; Table 2 summarizes our findings.
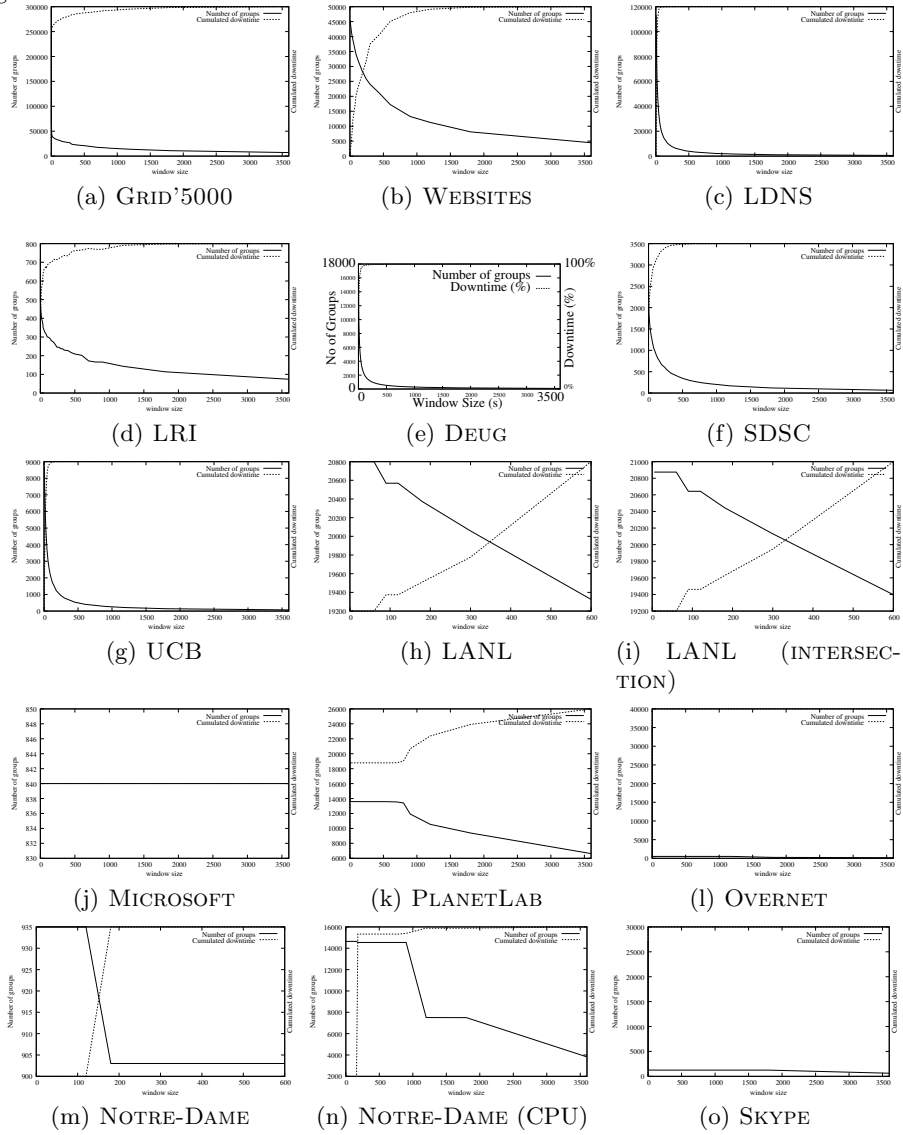
Fig. 3. Number of groups and cumulated downtime, for groups of at least 2 failures.

5 Analysis Results

In the previous section we have selected seven systems for which space-correlated failures are responsible for most of the system downtime. In this section, we present the results of fitting common distributions to the empirical distributions extracted from the failure traces of these seven traces selected. The space-correlated failures are generated using the moving windows method introduced in Section 3, and the values of Δ selected in Section 4.

The Failure Trace Archive already offers a toolbox (see [14, Section III.B] for details) for fitting common distributions to empirical data. We have adapted the tools already present in this toolbox for our model by extending the set of common distributions with the Pareto distribution, by adding a data preprocessing step that extracts groups of failures for a specific value of Δ, and by improving the output with automated graphing, tabulation, and summarization of results in text. These additions and are now publicly available as part of the FTA toolbox code repository.

5.1 Detailed Results

We have fitted to the empirical distributions five common distributions, exponential, Weibull, Pareto, lognormal, and gamma. We now present the results obtained for each model component, in turn.

Table 3. Failure Group Inter-Arrival Time: Best found parameters when fitting distributions to empirical data. Values in bold denote the best fit.

	Grid'5000	Websites	LDNS	LRI	Deug	SDSC	UCB
Exp	0.53	0.15	0.18	0.86	1.22	0.47	0.51
Weibull	0.44 0.79	0.16 1.21	0.12 0.74	0.46 0.63	0.23 0.47	0.13 0.57	0.07 0.48
Pareto	0.42 0.29	0.01 0.15	0.36 0.08	0.62 0.25	0.84 0.09	0.40 0.07	0.51 0.03
Logn	**-1.39 1.03**	**-2.17 0.76**	**-2.57 0.81**	**-1.46 1.28**	**-2.28 1.35**	**-2.63 0.86**	**-3.41 0.98**
Gamma	0.79 0.67	1.83 0.08	0.71 0.25	0.48 1.79	0.28 4.33	0.36 1.31	0.26 2.00

Failure Group Inter-Arrival Time To understand the failure group inter-arrival time, we consider for each failure group identified in the trace (including groups of size 1), the group generator (see Section 3.1). We then generate the empirical distribution from the time series corresponding to the inter-arrival time between consecutive group generators. Table 3 summarizes for each platform the parameters of the best fit obtained for each of the five common distribution we use in this work. The goodness-of-fit values for the AD and KS tests (see Section 3.3) are presented in Appendix A in Tables 9 and 10, respectively. These results reveal that the failure group inter-arrival time is not well characterized by a heavy-tail distribution as the p-values for the Pareto are low. Moreover, we identify two categories of platforms. The first category, represented by Grid'5000, LRI and Websites, is well-fitted by Log-Normal distributions. The second category, represented by Deug, LDNS, SDSC, UCB, is not well-fitted by any of the common distributions we tried; for these, the best-fits are either the lognormal or the gamma distributions.

Table 4. Failure Group Size: Best found parameters when fitting distributions to empirical data. Values in bold denote the best fit.

	Grid'5000	Websites	LDNS	LRI	Deug	SDSC	UCB
Exp	17.09	2.55	13.44	5.74	10.96	5.19	4.47
Weibull	12.82 0.71	2.87 1.60	15.12 2.29	5.76 1.01	12.12 1.39	4.94 0.93	5.05 2.52
Pareto	0.68 6.75	-0.06 2.68	-0.18 15.09	0.22 4.43	-0.03 11.26	0.22 3.70	-0.41 5.76
Logn	**1.88 1.25**	0.84 0.35	**2.52 0.41**	**1.32 0.77**	**2.15 0.70**	**1.19 0.70**	1.41 0.42
Gamma	0.64 26.78	**5.33 0.48**	6.23 2.16	1.30 4.40	2.22 4.94	1.23 4.24	**6.03 0.74**

Failure Group Size To understand the failure group size, we generate the empirical distribution of the sizes of each group identified in the trace (including

groups of size 1). Table 4 summarizes for each platform the parameters of the best fit obtained for each of the five common distribution we use in this work. The goodness-of-fit values for the AD and KS tests are presented in Appendix A in Tables 11 and 12, respectively. similarly to our findings for the failure group inter-arrival time, the results for the failure group size reveal heavy-tail distributions are not good fits. We find that the lognormal and gamma distributions are good fits for the empirical distributions.

Table 5. Failure Group Duration, $D_{\max}$: Best found parameters when fitting distributions to empirical data. Values in bold denote the best fit.

	GRID'5000	WEBSITES	LDNS	LRI	DEUG	SDSC	UCB
EXP	3.33e6	21225.18	2.48e6	2.46e5	1.18e5	67183.25	4071.25
WEIBULL	75972.13 0.28	10658.82 0.63	2.430e6 0.96	1.051e5 0.48	61989.86 0.54	35581.34 0.63	4131.60 1.03
PARETO	3.10 2686.08	0.73 5493.50	0.16 2.071e6	**1.71 24187.13**	1.53 15901.44	0.54 20627.60	0.09 3711.35
LOGN	**9.51 3.21**	**8.57 1.36**	**14.16 1.15**	10.41 2.45	**10.03 2.02**	**9.80 1.30**	**7.82 1.03**
GAMMA	0.14 2.362e6	0.46 46006.96	1.01 2.452e6	0.34 7.317e5	0.40 2.950e5	0.49 1.384e5	1.16 3509.88

Table 6. Failure Group Duration, D_{Σ}: Best found parameters when fitting distributions to empirical data. Values in bold denote the best fit.

	GRID'5000	WEBSITES	LDNS	LRI	DEUG	SDSC	UCB
EXP	4.40e5	10363.55	4.17e5	1.63e5	29979.27	30139.69	1500.92
WEIBULL	30951.59 0.33	6605.36 0.70	4.576e5 1.37	**80091.30 0.50**	13239.84 0.57	19008.04 0.69	1646.49 1.35
PARETO	2.54 2215.71	0.47 4258.00	-0.11 4.576e5	1.61 20672.26	0.91 5832.36	0.41 12570.49	-0.10 1645.39
LOGN	**8.89 2.71**	**8.20 1.13**	12.64 0.84	10.16 2.40	**8.67 1.62**	**9.25 1.16**	**7.01 0.81**
GAMMA	0.18 2.418e6	0.59 17462.56	**1.82 2.292e5**	0.36 4.484e5	0.40 74867.95	0.59 51497.86	1.82 825.92

Failure Group Duration The two last components of our model are the parallel- and single-node downtime of the space-correlated failures. To understand these two components, we generate for each the empirical data distribution using the durations of each group identified in the trace (including groups of size 1). The results of the fitting of the parallel downtime component are presented in Table 5, and the results of the fitting of the single-node downtime component are given in Table 6. Furthermore, the results of the AD and KS goodness of fit tests are shown in Tables 13 and 14 (parrallel downtime), and in Tables 15 and 16 (single-node downtime). Similarly to our previous findings in this section, we find that heavy-tail distributions such as Pareto do not fit well the empirical distributions. In contrast, the lognormal distribution is by far the best fit, with only two systems (LDNS and LRI) being better represented by the other distributions (the Gamma and Weibull distributions, respectively).

5.2 Results Summary

For all the component of our model and for all platforms, the most well-suited distribution is presented in Table 7. The main result is that Log-Normal distributions provide good results for almost all parts of our model. Thus, we can model most of node-level failures in the whole platform by groups of failures, each group being characterized by its size, its parallel downtime and its single-node downtime.

Table 7. Best fitting distribution for all model components, for all systems.

	Group size	Group IAT	$D_{\max}$	D_Σ
Grid'5000	Logn (1.88,1.25)	Logn (-1.39,1.03)	Logn (9.51,3.21)	Logn (8.89,2.71)
Websites	Gamma (0.84,0.35)	Logn (-2.17,0.76)	Logn (8.57,1.36)	Logn (8.20,1.13)
LDNS	Logn (2.52,0.41)	Logn (-2.57,0.81)	Logn (14.16,1.15)	Gamma (1.82,2.292e5)
LRI	Logn (1.32,0.77)	Logn (-1.46,1.28)	Weibull (1.051e5,0.48)	Weibull (80091.30,0.50)
Deug	Logn (2.15,0.70)	Logn (-2.28,1.35)	Logn (10.03,2.02)	Logn (8.67,1.62)
SDSC	Logn (1.10,0.70)	Logn (-2.63,0.86)	Logn (9.80,1.30)	Logn (9.25,1.16)
UCB	Gamma (6.03,0.74)	Logn (-3.41,0.98)	Logn (7.82,1.03)	Logn (7.01,0.81)

6 Related work

From the large body of research already dedicated to modeling the availability of parallel and distributed computer systems–see [3–6] and the references within–, relatively little attention has been given to space-correlated errors and failures [4, 10, 13], despite their reported importance [1, 3].

Table 8. Research on space-correlated availability in distributed systems.

Study	*System Type* System Name (Number of Systems/Total Size [nodes])	Data Source (Length)	Errors/ Failures	Gen. Process	Setup Type (Δ [min])
[4]	*SC* VAXcluster (1 sys./7)	Sys.logs (10 mo.)	Errors	time partitioning	manual (5 min.)
[13]	*NoW* Microsoft (1 sys./>50,000)	Msmts. (5 weeks)	Failures	instantaneous	manual (0 min.)
[10]	*Grid* Grid'5000 (15 cl./>2,500)	Sys.logs (1.5 years)	Failures	extending window	auto (0.5–60)
This study	*Various* Various (15 sys./>500,000)	Various (>6 mo. avg.)	Failures	moving window	auto (0.02–60)

Note: SC, NoW, Sys, Cl, Msmts, and Mo are acronyms for supercomputer, network of workstations, system, cluster, measurements, and months, respectively.

The main differences between this work and the previous work on space-correlated errors and failures is summarized in Table 8. Our study is the first to investigate the problem in the broad context of distributed systems, through the use of a large number of failure traces. Besides a broader scope, our study is the first to use a generation process based on a moving window, and to propose a method for the selection of the moving window size.

7 Conclusion and Future Work

It is highly desirable to understand and model the characteristics of failures in distributed systems, since today millions of users depend on their availability. Towards this end, in this study we have developed a model for space-correlated failures, that is, for failures that occur within a short time frame across distinct components of the system. For such groups of failures, our model considers three aspects, the group arrival process, the group size, and the downtime caused by the group of failures. We found that the best models for these three aspects are mainly based on the lognormal distribution.

We have validated this model using failure traces taken from diverse distributed systems. Since the input data available in these traces, and, to our knowledge, in any failure traces available to scientists, do not contain information about the space correlation of failures, we have developed a method based on moving windows for generating space-correlated failure groups from empirical data. Moreover, we have designed an automated way to determine the window size, which is the unique parameter of our method, and demonstrated its use on the same traces.

We found that for seven out of the fifteen traces investigated in this work, a majority of the system downtown is caused by space-correlated failures. Thus, these seven traces are better represented by our model than by traditional models, which assume that the failures of the individual components of the system are independently and identically distributed.

This study has also allowed us to contribute six new failure traces in standard format to the Failure Trace Archive, which we hope can encourage others to use the Archive and also to contribute to it with failure traces.

References

1. T. Heath, R. P. Martin, and T. D. Nguyen, "Improving cluster availability using workstation validation," in *SIGMETRICS*, 2002, pp. 217–227.
2. R. Bhagwan, K. Tati, Y. Cheng, S. Savage, and G. Voelker, "Total recall: System support for automated availability management," in *NSDI*, 2004, pp. 337–350.
3. R. Sahoo, A. Sivasubramaniam, M. Squillante, and Y. Zhang, "Failure data analysis of a large-scale heterogeneous server environment," in *DSN*, 2004, pp. 772–.
4. D. Tang and R. K. Iyer, "Dependability measurement and modeling of a multicomputer system." *IEEE Trans. Computers*, vol. 42, no. 1, pp. 62–75, 1993.
5. B. Schroeder and G. A. Gibson, "A large-scale study of failures in high-performance computing systems," in *DSN*, 2006, pp. 249–258.
6. A. Iosup, C. Dumitrescu, D. H. J. Epema, H. Li, and L. Wolters, "How are real grids used? the analysis of four grid traces and its implications," in *GRID*, 2006, pp. 262–269.
7. X. Castillo, S. R. McConnel, and D. P. Siewiorek, "Derivation and calibration of a transient error reliability model," *IEEE Trans. Computers*, vol. 31, no. 7, pp. 658–671, 1982.
8. R. K. Iyer, S. E. Butner, and E. J. McCluskey, "A statistical failure/load relationship: Results of a multicomputer study," *IEEE Trans. Computers*, vol. 31, no. 7, pp. 697–706, 1982.
9. J. Gray, "A Census of Tandem System Availability Between 1985 and 1990," in *IEEE Trans. on Reliability*, vol. 39, no. 4, October 1990, pp. 409–418.
10. A. Iosup, M. Jan, O. O. Sonmez, and D. H. J. Epema, "On the dynamic resource availability in grids," in *GRID*, 2007, pp. 26–33.
11. Y. Zhang, M. Squillante, A. Sivasubramaniam, and R. Sahoo, "Performance implications of failures in large-scale cluster scheduling," in *JSSPP*, 2004, pp. 233–252.
12. J. W. Mickens and B. D. Noble, "Exploiting availability prediction in distributed systems," in *NSDI*, 2006.
13. W. J. Bolosky, J. R. Douceur, D. Ely, and M. Theimer, "Feasibility of a serverless distributed file system deployed on an existing set of desktop PCs," in *SIGMETRICS*, 2000, pp. 34–43.
14. 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.
15. A. Avizienis, J.-C. Laprie, B. Randell, and C. E. Landwehr, "Basic concepts and taxonomy of dependable and secure computing," *IEEE Trans. Dependable Sec. Comput.*, vol. 1, no. 1, pp. 11–33, 2004.
16. T.-T. Y. Lin and D. P. Siewiorek, "Error log analysis: statistical modeling and heuristic trend analysis," in *IEEE Trans. on Reliability*, vol. 39, no. 4, October 1990, pp. 419–432.

17. J. Gray, "Why do computers stop and what can be done about it?" in *Symposium on Reliability in Distributed Software and Database Systems*, 1986, pp. 3–12.
18. J. Aldrich, "R. A. Fisher and the making of maximum likelihood 1912-1922," *Statistical Science*, vol. 12, no. 3, pp. 162–176, 1997.
19. D. Nurmi, J. Brevik, and R. Wolski, "Modeling machine availability in enterprise and wide-area distributed computing environments," in *Euro-Par*, 2005, pp. 432–.

A   Goodness-of-Fit Results

In this appendix we present additional tables to support the claims of good fit and the claims of best fit we made throughout Section 5. To this end, we present here two tables with the p-values of the AD and KS tests, respectively, for each model component. Consistent with the standard method for computing p-values [14, 19], each value we show is an average of 1,000 p-values, each of which is computed by selecting 30 samples randomly from the dataset.

Failure Group Inter-Arrival Time Tables 9 and 10 show the p-values for the AD and KS goodness of fit tests, respectively.

Failure Group Size Tables 11 and 12 show the p-values for the AD and KS goodness of fit tests, respectively.

Failure Group Duration, $D_{\max}$ Tables 13 and 14 show the p-values for the AD and KS goodness of fit tests, respectively.

Failure Group Duration, D_{Σ} Tables 15 and 16 show the p-values for the AD and KS goodness of fit tests, respectively.

Table 9. Failure Group Inter-Arrival Time: P-values for AD goodness of fit tests. Values in bold denote p-values above the test threshold for the significance level 0.5.

	Grid'5000	Websites	LDNS	LRI	Deug	SDSC	UCB
Exp	**0.185**	**0.329**	0.0175	0.0361	1.50e-5	1.25e-6	5.78e-7
Weibull	**0.221**	**0.491**	0.034	**0.287**	0.0242	0.0074	0.0047
Pareto	9.39e-5	1.09e-5	9.85e-7	6.56e-4	3.14e-6	2.57e-6	6.58e-8
Logn	**0.446**	**0.607**	**0.183**	**0.569**	**0.0801**	**0.156**	0.0153
Gamma	**0.2**	**0.55**	0.026	**0.147**	0.0072	0.0018	0.0012

Table 10. Failure Group Inter-Arrival Time: P-values for KS goodness of fit tests. Values in bold denote p-values above the test threshold for the significance level 0.5.

	Grid'5000	Websites	LDNS	LRI	Deug	SDSC	UCB
Exp	**0.0911**	**0.185**	8.27e-4	0.0035	3.10e-10	3.21e-11	1.61e-17
Weibull	**0.075**	**0.409**	5.27e-4	**0.153**	4.550-4	2.33e-4	4.38e-6
Pareto	3.04e-6	1.96e-7	3.16e-8	7.98e-5	1.01e-8	1.36e-6	5.21e-11
Logn	**0.333**	**0.466**	0.0496	**0.372**	0.0148	0.104	6.34e-4
Gamma	**0.0995**	**0.437**	0.0021	**0.0579**	3.49e-4	3.04e-5	1.67e-7

Table 11. Failure Group Size: P-values for AD goodness of fit tests. Values in bold denote p-values above the test threshold for the significance level 0.5**.**

	Grid'5000	Websites	LDNS	LRI	Deug	SDSC	UCB
Exp	**0.28**	**0.439**	**0.14**	**0.508**	**0.449**	**0.398**	**0.333**
Weibull	**0.502**	**0.498**	**0.597**	**0.516**	**0.639**	**0.408**	**0.734**
Pareto	**0.122**	**0.244**	4.86e-5	**0.079**	4.83e-4	**0.115**	0.0049
Logn	**0.575**	**0.684**	**0.711**	**0.647**	**0.708**	**0.622**	**0.765**
Gamma	**0.393**	**0.595**	**0.699**	**0.469**	**0.687**	**0.392**	**0.772**

Table 12. Failure Group Size: P-values for KS goodness of fit tests. Values in bold denote p-values above the test threshold for the significance level 0.5**.**

	Grid'5000	Websites	LDNS	LRI	Deug	SDSC	UCB
Exp	0.0232	1.49e-7	0.0011	0.0072	**0.119**	0.0026	3.41e-4
Weibull	0.0485	1.40e-6	**0.257**	0.008	**0.418**	8.32e-4	**0.206**
Pareto	1.19e-5	6.24e-14	9.14e-11	2.18e-11	2.09e-6	1.88e-8	4.74e-14
Logn	**0.144**	2.69e-4	**0.37**	**0.0651**	**0.421**	0.0267	**0.174**
Gamma	**0.0578**	2.16e-4	**0.376**	0.024	**0.44**	0.0073	**0.191**

Table 13. Failure Group Duration, $D_{\max}$**: P-values for AD goodness of fit tests. Values in bold denote p-values above the test threshold for the significance level** 0.5**.**

	Grid'5000	Websites	LDNS	LRI	Deug	SDSC	UCB
Exp	1.18e-6	0.0364	**0.563**	**0.0643**	0.047	0.0151	**0.565**
Weibull	**0.235**	**0.278**	**0.556**	**0.606**	**0.513**	**0.23**	**0.578**
Pareto	0.0118	5.99e-4	3.84e-5	**0.0593**	0.0061	2.74e-4	1.34e-5
Logn	**0.368**	**0.356**	**0.61**	**0.609**	**0.548**	**0.534**	**0.618**
Gamma	**0.0528**	**0.143**	**0.558**	**0.493**	**0.428**	**0.0962**	**0.571**

Table 14. Failure Group Duration, $D_{\max}$**: P-values for KS goodness of fit tests. Values in bold denote p-values above the test threshold for the significance level** 0.5**.**

	Grid'5000	Websites	LDNS	LRI	Deug	SDSC	UCB
Exp	5.41e-11	0.0083	0.004	0.0076	0.0085	0.0021	**0.465**
Weibull	**0.107**	**0.0733**	**0.303**	**0.292**	**0.348**	**0.111**	**0.477**
Pareto	0.0019	1.46e-5	0.0017	0.0035	0.0011	6.10e-5	5.27e-7
Logn	**0.223**	**0.223**	**0.263**	**0.253**	**0.347**	**0.383**	**0.481**
Gamma	0.023	**0.0709**	**0.251**	**0.236**	**0.301**	0.0291	**0.464**

Table 15. Failure Group Duration, D_{Σ}**: P-values for AD goodness of fit tests. Values in bold denote p-values above the test threshold for the significance level** 0.5**.**

	Grid'5000	Websites	LDNS	LRI	Deug	SDSC	UCB
Exp	4.85e-5	**0.0903**	**0.396**	**0.135**	0.0189	**0.0632**	**0.405**
Weibull	**0.26**	**0.222**	**0.592**	**0.627**	**0.455**	**0.243**	**0.59**
Pareto	0.0086	3.17e-4	1.23e-6	**0.0599**	0.0017	9.12e-5	5.02e-6
Logn	**0.393**	**0.342**	**0.602**	**0.583**	**0.576**	**0.535**	**0.626**
Gamma	**0.0775**	**0.158**	**0.606**	**0.547**	**0.23**	**0.139**	**0.602**

Table 16. Failure Group Duration, D_{Σ}**: P-values for KS goodness of fit tests. Values in bold denote p-values above the test threshold for the significance level** 0.5**.**

	Grid'5000	Websites	LDNS	LRI	Deug	SDSC	UCB
Exp	1.67e-7	0.0299	**0.233**	0.0209	0.004	0.0149	**0.276**
Weibull	**0.138**	0.0304	**0.48**	**0.325**	**0.37**	**0.136**	**0.485**
Pareto	0.0014	5.06e-6	1.92e-7	0.0033	1.78e-4	1.64e-5	1.13e-7
Logn	**0.24**	**0.183**	**0.471**	**0.235**	**0.414**	**0.373**	**0.494**
Gamma	0.0372	**0.0569**	**0.492**	**0.293**	**0.177**	**0.0527**	**0.473**