Quantifying Temporal and Spatial Correlation of Failure Events for Proactive Management

Song Fu and Cheng-Zhong Xu
Department of Electrical and Computer Engineering
Wayne State University, Detroit, MI 48202, USA
{song, czxu}@eng.wayne.edu

Abstract

Networked computing systems continue to grow in scale and in the complexity of their components and interactions. Component failures become norms instead of exceptions in these environments. Moreover, failure events exhibit strong correlations in time and space domain. In this paper, we develop a spherical covariance model with an adjustable timescale parameter to quantify the temporal correlation and a stochastic model to characterize spatial correlation. The models are further extended to take into account the information of application allocation to discover more correlations among failure instances. We cluster failure events based on their correlations and predict their future occurrences. Experimental results on a production coalition system, the Wayne State Grid, show the offline and online predictions by our predicting system can forecast 72.7% to 85.3% of the failure occurrences and capture failure correlations in cluster coalition environment.

1 Introduction

Networked computing systems continue to grow in scale and in the complexity of their components and interactions. In these systems, component failures become norms instead of exceptions. Failure occurrence as well as its impact on system performance and operation costs are becoming an increasingly important concern to system designers and administrators.

The growing complexity of hardware and software mandates autonomic management of failures in production systems. Failure prediction is a crucial technique for understanding emergent, system-wide phenomena and self-managing resource burdens. Based on the analysis of failure data in a system, a failure predictor can help determine possible occurrences of fatal events in the near future and help develop more effective failure tolerant solutions for improving system availability.

To achieve self-management of component failures in a system, we need an in-depth understanding of the cause of failures and their empirical and statistical properties. Past studies on component failures in production systems, such as IBM BlueGene/L supercomputer [12] and LANL high-performance computing clusters [17], revealed important patterns in failure distribution. Although the time-between-failure is highly non-linear, there exists the time-of-day and day-of-week patterns in long time spans [17, 15]. Temporal correlation aside, failure events, depending on their types, display strong spatial correlations: a small fraction of nodes may experience most of the failures in a coalition system [15] and multiple nodes may fail almost simultaneously [12]. These temporal and spatial correlation properties of failure events revealed by offline profiling provide important information for predicting the trend of failure dynamics.

Most of today's failure characterization approaches are heavily empirical, applying heuristics to explore temporal and spatial correlation of failures based on profiling. There lack formal models to quantify the temporal correlation among failures in different timescales. Most networked computing systems are hierarchical in structure and failures may occur in multiple scopes: node, cluster and system. There are no models to quantify the spatial correlation of failures for predicting their future distribution and locations in different scopes. It is known that there is dependency between the workload (its type and intensity) and the failure rate [17, 15]. However, there are few work on the impact of application allocation on failure correlation, either.

In this paper, we analyze and quantify both temporal and spatial correlations for proactive failure management in networked computing systems. We develop a covariance model with an adjustable timescale to quantify the temporal correlation and a stochastic model to describe spatial correlation. We utilize information of application allocation in a coalition system to discover more correlations among failure instances. We cluster failure events in a system based on their correlations and predict their future occurrences.

As a proof of concept, a prototype of failure predictor based on quantified failure correlation was developed and has

Table 1. Variables characterizing failure dynamics.

Variable	Description
fID	Failure identification number
$fLoct$	Location of a failure including compute node ID
$fType$	Classification of a failure based on its cause
$time$	Timestamp when a failure occurs
tbf	Time between successive failures
$fCount$	Number of failures in a subsystem for a time window
$usrUtil$	Percentage of CPU utilization that occurred while executing at the user level in a node
$sysUtil$	Percentage of CPU utilization that occurred while executing at the system level in a node
$frmUtil$	System frame utilization in a node
$pktCount$	Number of packets transmitted and received by a node for a time window
$ioCount$	Number of I/O requests to the physical disks of a node for a time window
$alloc$	Allocation information of nodes to application jobs
$sptCorr$	Spatial correlation among failures in a subsystem
$tmpCorr$	Temporal correlation among failures in a subsystem

been in operation since May 2006 on a production coalition environment: the Wayne State Grid [2]. The WSU Grid consists of three clusters located in three campus buildings of the Wayne State University and contains 40 high-performance compute servers in support of university-wide high-performance computing application programs. Offline and online failure predictions were performed with observed failures and on production traces from more than one and a half years of operations. The prediction results show our prediction system can forecast 85.3% of the failure occurrences cluster and system-wide, 72.7% in node-wide prediction, and capture the failure correlations in coalition clusters.

The rest of this paper is organized as follows: Section 2 presents the algorithms to explore temporal and spatial correlation among failure events. Section 3 describes the failure traces from the WSU Grid. The performance of offline and online prediction by our implemented prototype in the WSU Grid was evaluated in Section 4. Section 5 presents the related work and Section 6 summaries the paper.

2 Quantifying Temporal and Spatial Failure Correlation

In this section, we address two key issues in characterizing failure correlation: (a) What representation should we use to describe failure instances and the associated system performance variables? (b) How do we cluster failure signatures to identify temporal and spatial correlations among failure occurrences and to utilize the correlations for proactive failure management?

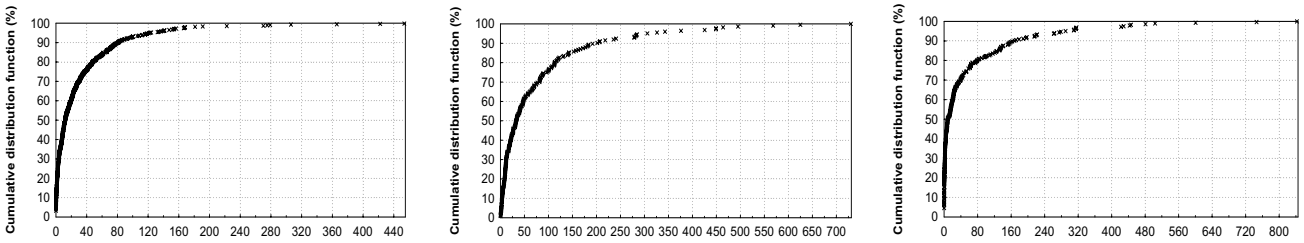
Without loss of generality, when we refer to a "failure" in the following discussion, we mean any anomaly caused by hardware defect, incorrect design, unstable environment or operator mistakes that makes services or compute nodes unavailable.

2.1 Failure Signatures

An important issue we address is that of extracting from a running system an indexable representation that distills the essential characteristic from a system state associated with a failure event. For this end, we define several performance variables of system characteristics in the face of failures. However, defining these variables is nontrivial. They should be able to present the difference between system states in normal execution and those in failures. They also need to capture the temporal and spatial correlations of failure events in multiple system components.

By investigating the structure of networked computing systems and the correlations of failure occurrences in a system, we define the performance variables used in our failure management framework, as listed in Table 1. They are raw data collected from the system event logs or derived from the raw data. The runtime states of a subsystem is characterized by its processor and memory utilization and the volume of communication and I/O operations. These performance metrics provide insightful information about the causes of failures. Variables, such as the number of failures in a time window, their types and intervals, are used to model the statistical characteristics of failure dynamics. Along with the nodal allocation information, these variables are utilized to establish the spatial and temporal correlations among failure events.

To quantify the correlation among failure events in a networked computing system, we need a representation that provides essential information about system performance status associated with a failure event. By clustering these representations, we are able to capture the failure dynamics and correlation. We will call such a representation a *failure signature*. Based on the performance variables defined in Table 1, a failure signature is constructed as a tuple (fID, $time$, $fLoct$, $fType$, $util$, $pktCount$, $ioCount$), where $util$ includes ($usrUtil$, $sysUtil$, $frmUtil$) of a compute node, and $pktCount$ and $ioCount$ measure the number of packets and I/O requests



(a) All Failures (b) Hardware Failures (c) Software Failures

Figure 1. Temporal distribution of hardware and software failure events in the Wayne State Grid from January 2, 2005 to January 6, 2006.

in the sampling period that immediately precedes the failure. With these failure signatures collected in a coalition system and the node allocation information *alloc*, we will analyze the failure distributions and correlations in both space and time domains.

2.2 Exploring Temporal and Spatial Correlation

The objective of applying clustering to a database of failure signatures is to find the natural grouping of these signatures that characterizes correlations among failure instances. The output of clustering is a set of groups, plus a characterization of each group. By inspecting the elements of failure signatures in each group, we can identify different regions of anomaly as well as a hint of the causing problems. In addition, the central signature of a group can be used as a syndrome of the failures, because it highlights the metrics that characterize a set of manifestations of the neighboring failure instances.

In order to render the description above operational, we specify distance metrics and clustering algorithms. We cluster failure signatures in two directions. One is to discover temporal locality among failure instances in the time domain. The other is to explore the causal dependency in the space domain. We design the clustering algorithms based on our observation from the Wayne State Grid.

2.2.1 Temporal Clustering

Studies in [17, 15, 12] found the skewness of failure distribution in time domain. Multiple failures may occur in a short time period. Liang [11] and Sahoo [14] used a fix time window to classify failure patterns for all types of failures. In reality, the time-between-failure (tbf) may follow various distributions for different types of failures. We profile time-between-failure in the WSU Grid from its failure traces. Figure 1 presents the cumulative distribution functions of tbf for the hardware, software and all failures. From the figure, we can see that tbf has a heavy tail distribution and its shape varies with the failure type. Software failures have a much more heavily tailed distribution in time than hardware failures. This discovery suggests that we should use an adjustable timescale to model and measure the temporal correlation among failures of different types.

By closely inspecting the system event logs and performance logs, we found that the temporal locality of failure events was mainly due to two causes:

- (T1) some faults [1] cause several failure instances occurred on multiple compute nodes in a short interval;
- (T2) a failure event may appear multiple times on a node before its root problem is solved.

To cluster failure signatures in the time domain, we define the distance between two failure events f_i and f_j as the elapsed time between them, denoted by $d_{i,j} = \| f_i - f_j \| = |t_{f_i} - t_{f_j}|$. We develop a spherical covariance model, based on recent advance of Bayesian statistics [4], to quantify the temporal failure correlations. The model characterizes the relations of failure instances in time space based on their distance between each other, even when they occur on different nodes. We assume the timers of compute nodes in a cluster are synchronized. The spherical covariance, $C_T(d)$, for temporal correlation is defined as:

$$C_T(d) = \begin{cases} 1 - \alpha\frac{d}{\theta} + \beta(\frac{d}{\theta})^3 & \text{if } 0 \le d \le \theta \\ 0 & \text{if } d > \theta \end{cases} \quad (2.1)$$

where θ is an adjustable timescale parameter for determining the temporal relevancy of two failure events, α and β are positive constants with $\alpha = 1 + \beta$. We use different values of θ to quantify temporal correlations of different types of failures. For example, with $\theta = 3$ hours for the WSU Grid, we can capture more than 40% software failures. For different types of failures, the value of θ varies. For other systems, θ can be determined by inspecting the cumulative distributions of the inter-failure time from their event logs. Two failures taken more than θ distance apart are considered as uncorrelated in time. $C_T(d)$ is nonnegative with limiting values of 1 at $d = 0$ and of 0 at $d = \infty$. After specifying the value of θ, we

[1] A fault is associated with incorrect state of a hardware or software component and it may cause a reduction in, or loss of, the capability of a component to perform a required function.

cluster the failure signatures of a compute node by comparing their $C_T(d)$ pair-wise. Failure signatures within a group is temporally correlated with high probability and likely to appear closely in time. The central signature is useful for investigating the root cause of the failure group and analyzing the distribution of inter-failure time among failure signatures in the same group. Algorithm 1 presents the pseudo-code of correlating failure signatures in time by a master node of a cluster. The master node first collects the failure signatures grouped by compute nodes in the cluster. It then inspects each pair of failure signature from different groups to calculates the temporal correlation. Although the algorithm needs to scan failure signatures of every compute node in a cluster, the total number of failure events occurred within a time window is quite limited.

Algorithm 1 ***Temporal clustering of failure signatures***

```
/* Temporal clustering on the master node of a subsystem */
MasterPredictor.TemplClustering() {
1:   collect signature groups from compute nodes in the cluster;
2:   θ = inter-failure time with p cumulative distribution;
3:   for any pair of nodes i, j in nodelist do
4:     T[i, j] =0;
5:     for any failure signature f on node i and g on j (t_f ≤ t_g) do
6:       d_{f,g} = |t_f − t_g|;
7:       if d_{f,g} ≤ θ then
8:         c_{f,g} = 1 − α * d_{f,g}/θ + β * (d_{f,g}/θ)^3;
9:         T[i,j] = T[i, j] + c_{f,g};
10:        if c_{f,g} ≥ C then
11:          Group_f = Group_f ∪ {g};
12:          remove g from the signature set;
13:        end if
14:      end if
15:    end for
16:  return T, Groups;
17:}
```

2.2.2 Spatial Clustering

In addition to temporal locality, failure events are correlated in the space domain. Multiple failures occur on different compute nodes and intersect with each other.

We analyzed the spatial distribution of application I/O failures in the WSU Grid and found some failure occurrences were clustered in space. By inspecting the system failure and performance traces, we found

- (S1) a failure may (nearly) simultaneously occur on multiple nodes in a cluster or across its border;
- (S2) a failure on a node may cause another failure happening on a different node.

The first case is common in parallel computing, where a single-program-multiple-data (SPMD) application runs on a set of nodes and a fatal software bug in the application will make multiple nodes come to failure. The second case happens among cooperative nodes. For example, a processor failure on one node may cause its running program to send wrong data to another node, which leads to an overflow and system dump. We refer to Case (S1) as *failure multiplication correlation* and to Case (S2) as *failure propagation correlation*. Note that in this paper we consider those propagation correlations that are caused by communication between failing nodes.

To find out what kind of nodes are likely to experience spatially correlated failures, we inspected the job scheduling information, the Portable Batch System (PBS) logs in WSU Grid. From these logs, we found that grid nodes had been allocated to grand applications in groups. It also indicates that faults triggered by parallel and distributed applications may cause multiple nodes to fail almost simultaneously. For example, 6 out of 10 nodes allocated to Job 538 failed on April 20, 2005. This job was to perform FFT in an image processing application. The bugs in the program caused the 6 nodes to experience endless loops. Reboot events were later found in their event logs. This indicates failures in coalition clusters are spatially correlated and the node allocation information is useful for predicting failure occurrences in the future.

We develop an aggregate stochastic model to cluster failure signatures in the space domain and use these groups for failure prediction. The model analyzes the probabilistic dependency among failure instances of different nodes, and combines the nodal failure statistics in a cluster into an aggregated state, which is further combined with failure states of other clusters into an aggregated system state. Modeling based on hierarchical decomposition and aggregation makes it possible to treat a large scale system by considering a reduced one with essentially the same features but with reduced complexity.

The analysis of failure multiplication correlation is based on the occurrence relations of failure events. Let set $F = \{f_1, f_2, \ldots, f_m\}$ denote all possible types of failures that may occur in a coalition clusters system, and $N = \{n_1, n_2, \ldots, n_r\}$ be the set of all compute nodes in the system. Random variables $\hat{n}_i$ and $\hat{f}_j$ are defined as: So, n_i and f_j indicate whether

$$\hat{n}_i = \begin{cases} 1 & \text{if node } n_i \text{ fails in a unit interval,} \\ 0 & \text{otherwise.} \end{cases} \quad \hat{f}_j = \begin{cases} 1 & \text{if failure } f_j \text{ occurs in a unit interval,} \\ 0 & \text{otherwise.} \end{cases}$$

a node fails or a failure happens. A unit interval is a small period of time when only one failure event can appear on a node. Time window is measured in unit intervals. Based on failure statistics, we measure the conditional probabilities $p(\hat{f}_j \mid \hat{n}_i)$, that is if node n_i fails, the probability that the failure is f_j, for $1 \le i \le r$ and $1 \le j \le m$.

Now, let's first consider the failure multiplication correlations among two nodes, say n_1 and n_2. The number of failures counted in a time window is $nodeFCount_i = \hat{n}_i \cdot w$. If we fix the window size in measurements, then the expected number of failures becomes,

$$E[nodeFCount_i] = w \cdot E[\hat{n}_i] = w \cdot p(\hat{n}_i) = w \sum_j p(\hat{n}_i \mid \hat{f}_j) \cdot p(\hat{f}_j).$$

We can further calculate the covariance of $nodeFCount_i$ of different compute nodes to analyze the correlations of these

variables. Assume the failure dynamics of the two nodes are monitored independently. Then according to the Bayesian theorem,

$$p(\hat{n}_1\hat{n}_2 \mid \hat{f}_j) = \frac{p(\hat{f}_j \mid \hat{n}_1)\cdot p(\hat{f}_j \mid \hat{n}_2)\cdot p(\hat{n}_1)\cdot p(\hat{n}_2)}{p(\hat{f}_j)^2}.$$

According to the inclusion-exclusion principle, the number of failure events after considering the failure multiplication correlation becomes

$$E[clusterFCount] = w \cdot \Big(\sum_i p(\hat{n}_i) - \sum_{i,k} \sum_j p(\hat{n}_i\hat{n}_k \mid \hat{f}_j)p(\hat{f}_j) + \cdots + \sum_j (-1)^r p(\hat{n}_1 \cdots \hat{n}_r \mid \hat{f}_j)p(\hat{f}_j)\Big). \quad (2.2)$$

By using these probabilities of failure distributions along with the temporal correlation among failure signatures, predictors in node, cluster and system wide calculate the number of failures that will occur in the prediction window with certain probability. Since $E[nodeFCount]$, $E[clusterFCount]$ and $E[sysFCount]$ are correlated, we use their corresponding prediction results to cross-verify each other. Then we refine the spatio-temporal correlations among the predicted failure instances by using the node allocation information in the system.

For failure propagation correlations, we define *propagation groups* to cluster failure signatures.

Definition 1 (Propagation relation) Let $\diamond$ be a relation on the set of failures F. It satisfies:

1. For any f_i and f_j in F, if f_i can cause f_j on another node, then f_i and f_j have relation $\diamond$, denoted as $f_i \diamond f_j$;
2. For any f_i, f_j and f_k in F, if $f_i \diamond f_j$ and $f_j \diamond f_k$, then $f_i \diamond f_k$. □

Relation $\diamond$ formulates the failure propagation dynamics between nodes. According to its definition, $\diamond$ is reflexive and transitive, and not symmetric. Thus, $(F, \diamond)$ is a partial order set. The propagation relation can be represented by Hasse diagrams. For partial order set $(F, \diamond)$, we calculate the transitive closure of relation $\diamond$ as $\mathcal{D}_\diamond$. The members of $\mathcal{D}_\diamond$ are groups of failure signatures that have possible propagation relations. Then we treat each member in $\mathcal{D}_\diamond$ as a unit and calculate its occurrence probability to compute nodes in N. After this transformation, the failure propagation correlation can be reformulated by the stochastic models that we use to analyze the failure multiplication correlation. Thus, we consider both of the spatial correlations in failure prediction.

In implementation, the predictor estimates the probabilities $p(\hat{f}_j \mid \hat{n}_i)$ and constructs the propagation relation $\diamond$ based on the failure statistics derived from event logs. For example, the predictor in node n_i counts the number of f_j occurrences and the total number of all failures in all past time windows and calculates the ratio between them as an estimate of $p(\hat{f}_j \mid \hat{n}_i)$. We set the length of an observation interval based on the measured mean-time-to-failures (MTTF) so that only one failure event can occur on a node in an interval. We also calculate the ratio of the number of intervals in which failures are observed on node n_i to the total number of intervals as the value of $p(\hat{n}_i)$. The master node of the system collects failure signatures from all compute nodes and estimate $p(\hat{f}_j)$ by (number of f_j instances)/(total number of failures) in the system. The master node also mines failure signatures to establish the propagation relations between failures. Failures f_j and f_k follow $f_j \diamond f_k$, if f_k occurs on a node, say A, after node A receives a message from another node B which suffers failure f_j. The probability of sending such a message equals to the inverse of the total number of messages sent by node B in the interval between occurrence time of failures f_j and f_k. In runtime, the predictor updates these probabilities and relation information using newly generated failure measures as system runs on. Then the correlations among failure signatures are analyzed. Node allocation information is utilized to refine failure correlations for prediction.

Algorithm 2 presents the pseudo-code of correlating failure signatures in space. Although the algorithm needs to scan failure events of every compute node in a cluster, the total number of failure events occurred within a time window is quite limited. The system wide predictor finds the failure correlations, utilizes cluster wide results and makes predictions in a similar way. The aggregate stochastic model reduces the state space of failure statistics and computational complexity, which facilitates online failure prediction in a coalition system.

Algorithm 2 ***Spatial correlating of failure signatures***

```
MasterPredictor.SpatioCluster() {
1:   for i = 1 upto nodelist.size do
2:       f_{j,i} = number of failure signature of type j on node i;
3:       n_i = f_{1,i} + f_{2,i} + ... + f_{m,i};
4:       p_{j,i} = f_{j,i}/n_i;
5:       p_{f_j} = (f_{j,1} + f_{j,2} + ... + f_{j,r})/(n_1 + n_2 + ... + n_r);
6:       p_{n_i} = number of intervals with failures on i / total number of intervals;
7:   end for
8:   for any pair of nodes i, j in nodelist do
9:       p_{n_{ij},f_k} = p_{k,i} * p_{k,j} * p_{n_i} * p_{n_j} / p^2_{f_k};
10:      S[i, j] = p_{n_{ij},f_1} * p_{f_1} + ... + p_{n_{ij},f_m} * p_{f_m} + p_i * p_{m_{ij}};
11:      if node i and j are allocated to the same job in
             nodeAlloc then
12:          S[i, j] = S[i, j] + 1;
13:      end if
14:  end for
15:  return S;
16:}
```

2.3 Proactive Failure Management

Occurrences of failures are quite dynamic in networked computing systems. The number of failure events varies with time. Numerically, its value is related to some performance metrics of the system, *e.g.* the resource utilization, the volume of communication and I/O operation. We model this

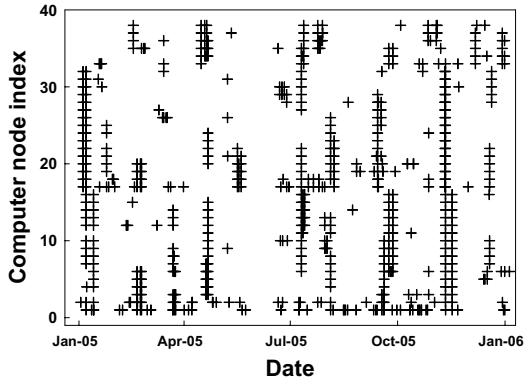


Figure 2. Failure events as a function of time on the WSU Grid.

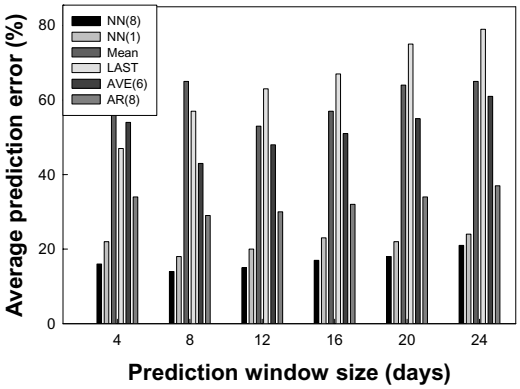


Figure 3. System-wide prediction accuracy using trace from the WSU Grid.

relationship by using function $\mathcal{F}$ in different scopes of a system, as

$$\begin{aligned} &\mathcal{F}_{node}(w, fCount_n, perf_n, tmpCorr_n) = 0 \\ &\mathcal{F}_{cluster}(w, fCount_c, perf_c, sptCorr_c, tmpCorr_c) = 0 \\ &\mathcal{F}_{system}(w, fCount_s, perf_s, sptCorr_s, tmpCorr_s) = 0 \end{aligned} \quad (2.3)$$

where w is the look-ahead window size of interest, *sptCorr* and *tmpCorr* are the spatial and temporal correlation among failure events in the corresponding scope. The performance state, $perf$, of a node can be represented by $(usrUtil, sysUtil, frmUtil, pktCount, ioCount)$. The $perf_c$ and $perf_s$ are composed of the mean and variance values of these performance variables in a cluster and the system, respectively. To find the essential performance variables for a failure instance, we analyze their probabilistic dependency among them in experiment evaluation sections.

In essence, predicting failures in a coalition system is to find approximate function $\mathcal{F}$. Failure events are highly nonlinear and it is difficult to find the relation between failure occurrences and performance states that fits various product systems. Instead of deriving function $\mathcal{F}$ directly, the predictor uses statistical learning approaches to perform failure prediction based on the current failure statistics and the resource utilization level. The prediction procedure can be expressed as follows,

$$x(w_{i+1}) = \mathcal{G}(x(w_i), x(w_{i-1}), \ldots, x(w_{i-k+1})) \quad (2.4)$$

where x denotes the measures of failure dynamics and $\mathcal{G}$ is the prediction function determined by a prediction mechanism with parameters' values in the k observation windows. The predictor's input layer is k consecutive measures in windows $w_i, w_{i-1}, \ldots, w_{i-k+1}$, obtained with the aid of a tapped delay line.

In this way, failure correlations spanning across multiple windows are kept for failure prediction. In essence, for a window size w, we can maintain the correlation information of a period of $k*w$ by using the order-k predictor, while being able to make failure predictions at a granularity of w at the same time. This scheme also increases the robustness of the failure predictor to noisy inputs because the noise effect of each measure fed to the predictor is suppressed by the multi-step looking back of the prediction mechanism.

3 Trace Collection and Characterization

In this study, we use event/failure trace files collected from the Wayne State Grid. First, we will briefly describe the architecture of the WSU Grid and the collected failure traces.

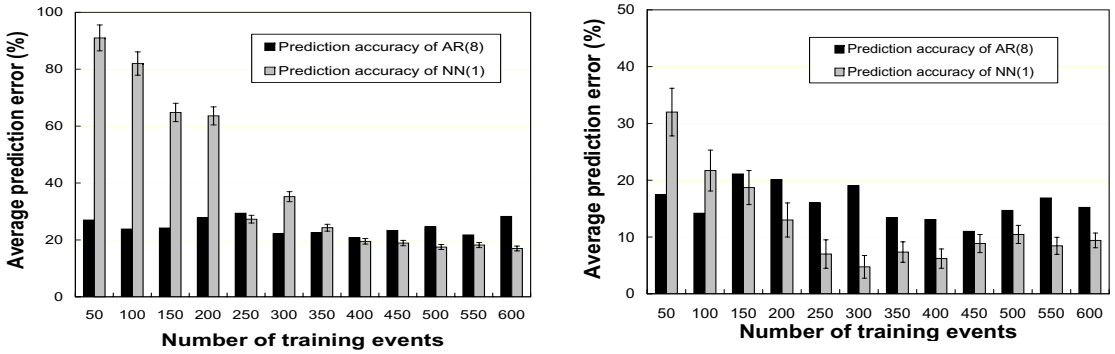
3.1 WSU Computational Grid

The WSU Grid consists of three Linux clusters, maintained by the Computing and Information Technology (CIT), the Institute for Scientific Computing (ISC), and the Department of Chemistry (CHM), in three separate buildings. It contains 40 high-performance compute servers dedicated to computational research. CIT and ISC clusters consist each of 16 nodes, and there are 8 nodes in the CHM cluster. Within each cluster, nodes are interconnected by gigabit Ethernet switches. Connections between clusters are through 100 Mbps fast Ethernet.

Typical applications running on the grid includes the molecular dynamics simulations, gene analysis, fluid dynamics simulation and more. These parallel applications ran on 8 to 30 nodes and some of them lasted for more than 10 days. The grid is also open to Wayne State students to execute their sequential and parallel programs.

3.2 Failure Traces

The event log files constitute a continuous 369-day trace between January 2, 2005 and January 6, 2006 from the three clusters. In the raw trace files, there are 35,547 event entries recorded and gathered in the system. The event logs have a total size of 596 MB and the PBS logs occupy 4.29 GB. To extract relevant event entries for our failure analysis, we first screened out events with lower severity levels, such as the informational and warning entries, because they did not af-



(a) Over-predictions (window size = 4 days).

(b) Under-predictions (window size = 4 days).

Figure 4. Cluster-wide failure prediction on the ISC cluster.

fect the availability of services and nodes. This step removes 97.5% entries from the raw logs. Then, we eliminated the duplicate adjacent entries from the result logs. This results in a 18.3% reduction in the number of entries of data, leaving only 726 failures, which include the hardware failures of compute nodes and network, the fatal failures of application programs and operating systems, operator's mis-operation and undetermined failures.

Among the remaining failure events, we observed four system maintenance operations conducted on June 3, 7, 9, and December 27, 2005 according to the administration logs. The first three ones were due to installing and debugging the PBS job scheduling system, and the last one was for system upgrade. After removing the events corresponding to these maintenance actions, we have the failure events distribution as shown in Figure 2. Time is represented on the X-axis and compute node index is on the Y-axis. A point is plotted each time a failure event occurs. The figure presents the burstiness of event occurrence and correlations of failures among compute nodes.

4 Experimental Results

As a proof of concept, we developed a prototype system for failure correlation quantification and management in the WSU Grid. We implemented several illustrating prediction algorithms by applying four time-series algorithms and a specific statical learning algorithm, the neural network, to learn and forecast failure dynamics. We used the gretl GNU time series library [1] and the Weka machine learning software [3] in the implementation.

The primary metric we used for evaluation is the relative prediction error, which quantifies the discrepancy between the actual and predicted values.

$$err = \frac{|\ PredictedValue - ActualValue\ |}{ActualValue} * 100\%,$$

where *err* is the absolute value of the relative prediction error, $PredictedValue$ is the number of failures predicted for the prediction window and $ActualValue$ is the number of failures observed for that window, where the window size is a configurable parameter. Mean error which we used for our computations is calculated by averaging all of the relative prediction errors. We conducted failure predictions in the Wayne State Grid and analyzed the accuracy of the system-wide, cluster-wide and node-wide predictions, by utilizing the temporal and spatial correlations among failure signatures quantified by approaches presented in Section 2.2.

4.1 Offline Prediction Performance

Figure 3 shows the performance of system-wide predictions. For illustration, we include four popular time-series algorithms: MEAN takes the average of previous measures as prediction; LAST uses the last measure; AVE(n) uses the average of last n measures; AR(m) is autoregressive; and one artificial neural network algorithm NN(n) using the last n measures to update the neural network and to predict failures. The predictor neural network has 4 input neurons to receive utilization and failure measures; 1 output neuron for prediction result; 3 hidden layers and 4 neurons in each hidden layer. The predictor uses these illustrating algorithms to forecast the number of failure instances that may occur in a prediction window. The first half of the failure trace is used for training, while the other half is used for prediction.From the figure, we can see AR performs the best among the time-series algorithms. But its prediction accuracy is still worse than the neural network-based predictors, because it does not characterize and adapt to the spatial correlations well in prediction. As described earlier, the high order predictor considers the interrelation of failure measures in consecutive windows and feed them as inputs to the network. The prediction error is reduced by 4.1% and the prediction accuracy reaches 85.3% when the prediction window size is 8 days.

On each cluster, we deploy a failure predictor on the master node. This predictor collects cluster-wide failure signatures and clusters them to explore failure correlations. The prediction error *err* only presents the absolute value of discrepancy between actual and predicted values. To provide more insight for performance analysis of our pre-

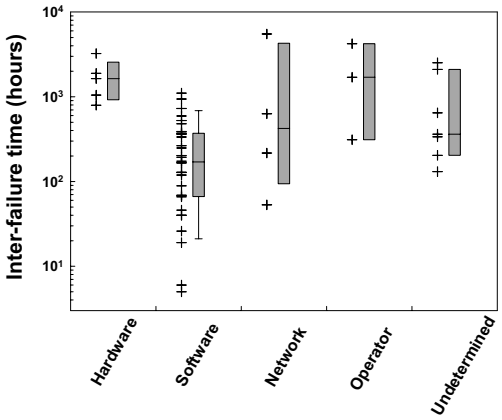


Figure 5. Node-wide failure signature clustering for Node 1 based on the causing problems.

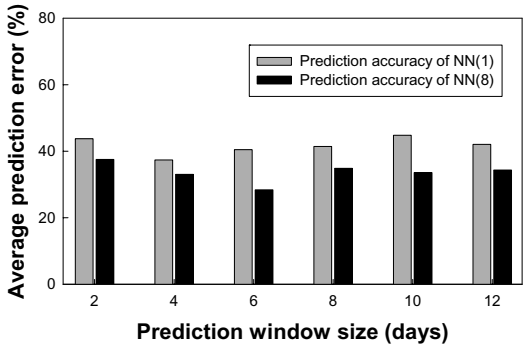


Figure 6. Node-wide failure prediction on Node 1 of the ISC cluster.

dictors, we distinguish two cases where (a) the number of predicted failures is less than the number of actually observed failures in a prediction window; (b) the former is greater than the latter. We refer to Case (a) as *under-prediction* and to Case (b) as *over-prediction*. The relative prediction error of a under-prediction is calculated by err = ($ActualValue$ - $PredictedValue$)/$ActualValue$ * 100%, and for over-prediction, it's the opposite. We conducted experiments to quantify the prediction error of under-predictions and over-predictions by using the NN failure predictor and the AR(8) on the three clusters. We measured the relative error of these two types of predictions. Figures 4(a) and 4(b) plot the average prediction error of over-predictions and under-predictions made by the two predictors on the ISC cluster. 95% confidence intervals are included in the figures. From these figures, we can see both predictors provide accurate results and the neural network based approach is a little better. Failure predictions on the CIT and CHM clusters display similar performance as that on the ISC cluster. Due to space limitation, we omit the experiment results on those two clusters in this paper.

To perform node-wide failure prediction, we cluster the failure signatures of a node according to the causes of the problems. Figure 5 plots the distributions of inter-failure time of 5-category failure events: hardware, software, network, operator, and undetermined failures, on Node 1 of the ISC cluster. A diagnosis memo in the administrative logs of WSU Grid record the causing problem of each failure event. The right box of each group marks the quartiles of the distribution and the horizontal line inside each box is the median. From the figure, we can see that the average interval between consecutive failures caused by software faults is the least one among the 5 categories, and 60.7% failure events on that node were caused by software faults.

We cluster the failure signatures of each category for Node 1. Temporal and spatial correlations are quantified based on the signature groups. Then the predictor forecasts the number of prospective failure instances in the next prediction window for each category. We calculate a weighted sum using the prediction result of each category and the prediction accuracy of that category in the previous predictions. This weighted sum is returned as a failure prediction. Figure 6 shows the prediction performance on Node 1 by using this weighted category approach. We notice that node-wide failure prediction is not as accurate as those in the cluster and system wide, due to the limited amount of failure events occurred on a compute node. We achieved up to 72.7% accuracy by using the NN(8) algorithm with the prediction window size equal to 6 days. Our failure prediction results are useful in coarse-grain job scheduling and node maintenance.

4.2 Online Prediction Performance

To evaluate the prediction performance in real system at runtime, we installed our predictors on compute nodes of each clusters and their master nodes in the WSU grid. By making on-line predictions, our failure predictors provide useful information for resource management, load distribution. In this experiment, first, we trained the predictors using failure event records between May, 2005 and April, 2006. Then, we evaluated the on-line prediction performance from May 12, 2006 to April 2, 2007. We record the failure predictions and compare them with the observed failure events later mined from the event logs.

Figure 7 depicts predicted and observed failure events from the entire computational grid during the online prediction based on the NN algorithm. Temporal and spatial correlations among failure occurrences were utilized for prediction. The node allocation information was used to refine correlations in predictions. From the figure, we can see the predictor can capture the trends of failure dynamics. An exception is noticed at time t = 32 when 21 failure events were observed while only 4 was predicted. Later, we checked the administration logs and found that on July 15, 2006, in the time interval of t = 32, a problem of a switch caused network breakdown in the CIT cluster and its compute nodes were un-

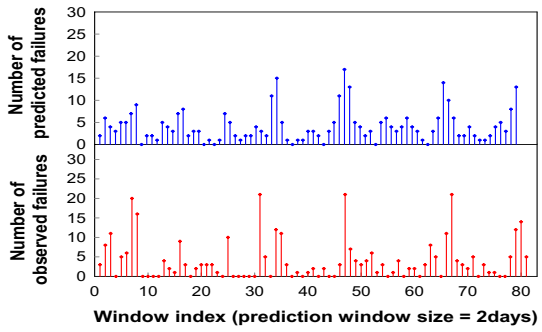


Figure 7. Online system-wide failure prediction in comparison with observed failure events.

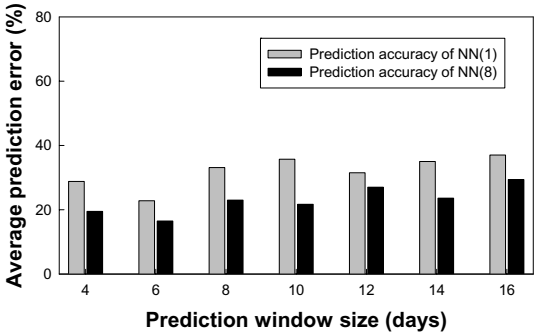


Figure 8. Online cluster-wide prediction on the ISC cluster.

available. To make a prediction, it took 2.17 seconds for the master node (a Pentium Xeon computer with 2.6 GHz processor and 2.5 GB of memory) to analyze the system wide failure events, find the failure correlations and make a prediction, after receiving the failure event data from the three clusters.

Figure 8 shows the performance of on-line predictions in the ISC cluster. We used predictors based on the NN(1) and NN(8) algorithms for prediction. According to these figures, the accuracy of online predictions is a little worse than that of offline predictions. This is due to the variance among limited number of prediction results. But still, the predictors provided useful information for autonomic management of the grid. Online predictions on the CIT and CHM clusters display similar performance as that on the ISC cluster. Due to space limitation, we omit the experiment results on those two clusters in this paper.

5 Related Work

To realize proactive management of failures, it is imperative to understand the characteristics of failure behaviors. Research in [17, 12, 15, 20] studied event traces collected from clusters and supercomputers. They found that failures are common in large-scale systems and their occurrences are quite dynamic, displaying uneven inter-arrival time. Sahoo et al. [15] found the correlation of failure rate with hour of the day and the distribution of failures across nodes. They reported that less than 4% of the nodes in a machine room experience almost 70% of the failures and found failure rates during the day to be four times higher than during the night. Similar result was observed by Schroeder and Gibson [17]. Iyer [10] and Castillo [5] reported a correlation between the workload intensity and the failure rate.

Xu et al. [19] performed a study of error logs collected from a heterogeneous distributed system consisting of 503 PC servers. They showed that failures on a machine tend to occur in bursts. They also observed a strong indication of error propagation across the network, which leads to the correlation between failures of different nodes. A recent study [8] collected failure data from three different clustered servers, and used Weibull distribution to model time-between-failure. Both these studies [19, 8] found that nodes which just failed are more likely to fail again in the near future. At the same time, it has also been found [18] that software related error conditions can accumulate over time, leading to system failing in the long run.

Salfner et al. [16] proposed a Similar Events Prediction method and demonstrated its effectiveness using field data of a telecommunication system. Their approach focused on time series-based failure prediction of single systems. In contrast, our approach emphasized on the impact of both temporal and spatial correlation of failure events in parallel computing systems. Hoffmann and Malek [9] proposed a function approximation method, called Universal Basis Functions, to model and forecast call availability in telecommunication systems. Their temporal model is elegant and could be applied to failure prediction of parallel computers. Challagulla et al. [6] applied machine learning techniques to predicting software defects. In our work, we apply statistical learning algorithms as illustration to predicting component failures in networked computer systems. The strong correlations among failure events result in better prediction precision by statistical learning. Besides, failure prediction aside, there are many works on failure detection. For example, Defago et al. [7] proposed formal accrual failure detectors that output a suspicion level instead of just a binary value. These techniques are complementary to our failure prediction approaches in constructing a comprehensive failure management infrastructure.

There were recent works utilizing temporal and/or spatial correlations of failures for failure prediction and proactive management. Sahoo et al. [14] inspected the eventset within a fixed time window before a target event for repeated patterns to predict the failure event of all types. Later, Liang et al. [11] profiled the time-between-failure of different failure types and applied a heuristic approach to detect failures by using a monitoring window of preset size corresponding to event type. Mickens and Noble [13] assumed the independency of failures among compute nodes and used the per-

node uptime data to predict whether a failure might occur on that node in the next time window of fixed size. In building classification rules, Sahoo et al. [14] took the ordering of events into consideration. They utilized a Bayesian network to analyze the causes of failures in a node individually. The spatial correlation among failure was considered by Liang et al. [11]. The authors analyzed the number of failures in every midplane of IBM BlueGene/L supercomputer. They found skewness in the distribution of network failures only, among the midplanes.

6 Conclusions

In this paper, we analyze and quantify the correlation among failure events in networked computing systems. Failure events are formally represented by failure signatures. By clustering signatures in the time and space domains, we explore the temporal and spatial correlations among failure occurrences. Node allocation information is utilized to refine the predicted correlations. Experimental results of offline and online prediction on a production coalition system present the feasibility of applying failure prediction to autonomic management for high-availability network computing.

The quantification and prediction of failure correlation by our approaches are useful for coarse-grain scheduling, resource management, and system maintenance. As an ongoing work, we are analyzing our proposed approaches rigorously and deriving their properties theoretically. In this work, the predictor forecasts the number of failure instances in a look-ahead window, based on the temporal and spatial correlation among failure events. However, it does not predict the time point when a failure will occur. As a future work, we will analyze the occurrence time of failures and design a mechanism to predict when a failure is going to happen.

Acknowledgments We would like to thank our chaperons, Roberto Baldoni, I-Ling Yen, Xavier Defago, and Miroslaw Malek, and the anonymous reviewers for their constructive comments and suggestions. We would also like to thank Philip Sokolowski and Michael thompson for their kind help in data collection from the Wayne State Grid. This research was supported in part by U.S. NSF grants CCF-0611750, DMS-0624849 and CNS-0702488.

References

[1] gretl: GNU Regression, Econometrics and Time-series Library. Available at: http://gretl.sourceforge.net/.

[2] Wayne State University, Grid computing. Available at: https://www.grid.wayne.edu/.

[3] Weka: The university of waikato. machine learning software in java. Available at: http://www.cs.waikato.ac.nz/ml/weka/.

[4] J. Berger, V. Oliveira, and B. Sansó. Objective Bayesian analysis of spatially correlated data. *Journal of American Statistical Association*, 96(456):1361–1374, 2001.

[5] X. CastiUo and D. P. Siewlorek. Workload, performance and reliability of digital computing systems. In *Proc. of Symp. on Fault-Tolerant Computing (FTCS)*, 1981.

[6] V. U. B. Challagulla, F. B. Bastani, I.-L. Yen, and R. A. Paul. Empirical assessment of machine learning based software defect prediction techniques. In *Proc. of Workshop on Object-Oriented Real-Time Dependable Systems*, 2005.

[7] X. Defago, P. Urban, N. Hayashibara, and T. Katayama. Definition and specification of accrual failure detectors. In *Proc. of IEEE Conf. on Dependable Systems and Networks (DSN)*, 2005.

[8] T. Heath, R. P. Martin, and T. D. Nguyen. Improving cluster availability using workstation validation. In *Proc. of ACM Conf. on Measurement and modeling of computer systems (SIGMETRICS)*, 2002.

[9] G. A. Hoffmann and M. Malek. Call availability prediction in a telecommunication system: A data driven empirical approach. In *Proc. of IEEE Symp. on Reliable Distributed Systems (SRDS)*, 2006.

[10] R. K. Iyer, D. Rossetti, and M. Hsueh. Measurement and modeling of computer reliability as affected by system activity. *ACM Trans. on Computer Systems*, 4(3):214–237, 1986.

[11] Y. Liang, Y. Zhang, A. Sivasubramaniam, M. Jette, and R. K. Sahoo. BlueGene/L failure analysis and prediction models. In *Proc. of IEEE Conf. on Dependable Systems and Networks (DSN)*, 2006.

[12] Y. Liang, Y. Zhang, A. Sivasubramaniam, R. Sahoo, J. Moreira, and M. Gupta. Filtering failure logs for a BlueGene/L prototype. In *Proc. of IEEE Conf. on Dependable Systems and Networks (DSN)*, 2005.

[13] J. Mickens and B. Noble. Exploiting availability prediction in distributed systems. In *Proc. of USENIX Symp. on Networked Systems Design and Implementation (NSDI)*, 2006.

[14] R. K. Sahoo, A. J. Oliner, I. Rish, and et al. Critical event prediction for proactive management in large-scale computer clusters. In *Proc. of ACM Conf. on Knowledge Discovery and Data Mining (SIGKDD)*, 2003.

[15] R. K. Sahoo, A. Sivasubramaniam, M. S. Squillante, and Y. Zhang. Failure data analysis of a large-scale heterogeneous server environment. In *Proc. of IEEE Conf. on Dependable Systems and Networks (DSN)*, 2004.

[16] F. Salfner, M. Schieschke, and M. Malek. Predicting failures of computer systems: A case study for a telecommunication system. In *Proc. of Workshop on Dependable Parallel, Distributed and Network-Centric Systems in conjunction with Intl. Parallel and Distributed Processing Symp.*, 2006.

[17] B. Schroeder and G. Gibson. A large-scale study of failures in hpc systems. In *Proc. of IEEE Conf. on Dependable Systems and Networks (DSN)*, 2006.

[18] K. Vaidyanathan, R. E. Harper, S. W. Hunter, and K. S. Trivedi. Analysis and implementation of software rejuvenation in cluster systems. In *Proc. of Conf. on Measurement and modeling of computer systems (SIGMETRICS)*, 2001.

[19] J. Xu, Z. Kalbarczyk, and R. K. Iyer. Networked Windows NT system field failure data analysis. In *Proc. of Pacific Rim Symp. on Dependable Computing (PRDC)*, 1999.

[20] P. Yalagandula, S. Nath, H. Yu, P. B. Gibbons, and S. Sesha. Beyond availability: Towards a deeper understanding of machine failure characteristics in large distributed systems. In *Proc. of USENIX WORLDS*, 2004.