Proactive Failure Management for Dependable Networked Computer Systems

Song Fu, Department of Computer Science and Engineering, University of North Texas

Cheng-Zhong Xu, Department of Electrical and Computer Engineering, Wayne State University

Motivation

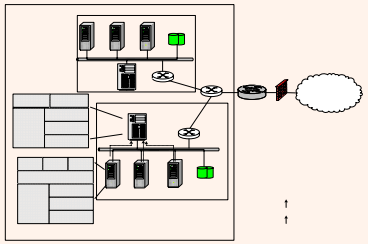
Networked computer systems continue to grow in scale and in the complexity of their components and interactions. In these systems, failures become norms instead of exceptions.

These require failure management tasks have significantly higher levels of automation.

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 aims to determine possible occurrences of fatal events in the near future and help develop more effective failure tolerant solutions for improving system availability.

Hierarchical Failure Prediction Framework

A networked computer system is hierarchical in structure and failures may occur in multiple scopes: node, cluster and system. Failures occurred at these system scopes are correlated in time and space.

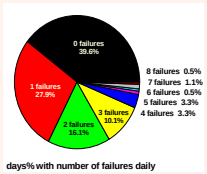


hPREFECTS: a hierarchical failure prediction framework for networked computer systems. The temporal and spatial correlation among critical events are quantified for failure prediction.

Failure & Performance Traces

From the Los Alamos National Laboratory HPC clusters: 22 clusters and 4,750 nodes; 23,739 failure records between 10/1995 and 9/2005.

Focus on one typical cluster #20: 256 nodes with1024 processors.

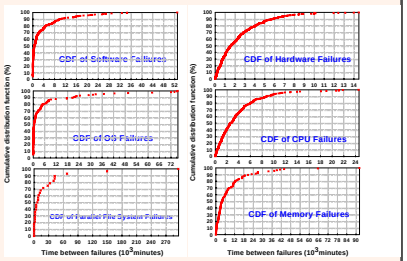


Daily distribution of failures in Cluster 20 between 9/1/03 and 8/31/04. 75.5% failures occurred in 30.1% days.

Temporal Correlation of Failures

Failures cluster in the time domain:

- Some faults cause several failure instances occurred in short intervals on multiple compute nodes
- A failure event may appear multiple times within a short period of time on a node before its root problem is solved.



Temporal distribution of major hardware and software failure events in LANL Cluster 20 from 9/1/2003 to 8/31/2005.

Time-between-failures (tbf) has a heavy tail distribution, which indicates failures cluster in the time domain, and its shape varies with the failure type. Software failures have a much more heavily tailed distribution in time than hardware failures.

Quantifying Temporal Correlation

Multi-scale spherical covariance model.

Distance between failure events *fi* & *fj*:

$$d_{i,j} = \| f_i - f_j \| = |t_{f_i} - t_{f_j}|$$

Spherical covariance to quantify the temporal correlations, based on Bayesian statistics

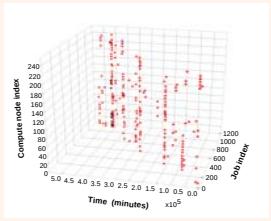
$$C_T(d) = \begin{cases} 1 - \alpha_\theta^d + \beta(\frac{d}{\theta})^3 & \text{if } 0 \le d \le \theta \\ 0 & \text{if } d > \theta \end{cases}$$

θ is an *adjustable* timescale parameter for determining the temporal relevancy of two failure events.

Spatial Correlation of Failures

Failures cluster in the space domain:

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



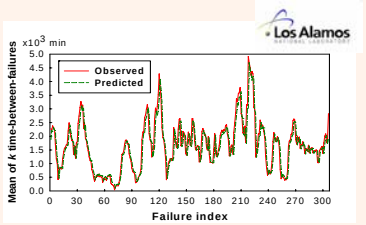
Spatial clustering of failures in Cluster 20 of the LANL HPC system.

Quantifying Spatial Correlation

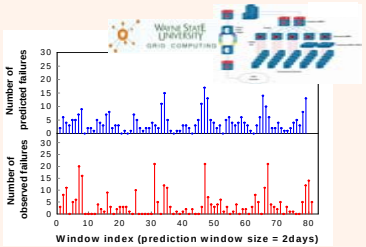
An *aggregate stochastic model* to cluster failure signatures in the space domain.

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.

Experimental Results



Offline failure prediction results using LANL trace with order-8 neural network predictor. Training samples are based on failure records from September 2003 to August 2004 and prediction is for September 2004 - August 2005. *Achieved 76.5% accuracy in offline prediction.*



Online failure prediction results in the Wayne State University Computational Grid from 5/12/2006 to 4/2/2007. *Achieved 70.7% accuracy in online prediction.* 2.17 seconds to make a system-wide prediction on master node.

Selected Publications

[1] S. Fu, Failure-Aware Resource Management for High-Availability Computing Clusters with Distributed Virtual Machines, JPDC 2010.

[2] S. Fu and C.-Z. Xu, Quantifying Event Correlations for Proactive Failure Management in Networked Computing Systems, JPDC 2010.

[3] S. Fu and C.-Z. Xu, Exploring Event Correlation for Failure Prediction in Coalitions of Clusters, ACM/IEEE SC 2007.

[4] S. Fu and C.-Z. Xu, Quantifying Temporal and Spatial Correlation of Failure Events for Proactive Management, IEEE SRDS 2007.

UNT UNIVERSITY OF NORTH TEXAS

WAYNE STATE UNIVERSITY