The Computer Failure Data Repository (CFDR)

Bianca Schroeder **Garth A. Gibson**
Carnegie Mellon University, {bianca,garth}@cs.cmu.edu

1 Motivation

Component failure in large-scale IT installations, such as cluster supercomputers or internet service providers, is becoming an ever larger problem as the number of processors, memory chips and disks in a single cluster approaches a million. Yet, virtually no data on failures in real systems is publicly available, forcing researchers working on system reliability to base their work on anecdotes and back of the envelope calculations, rather than empirical data.

This submission describes an effort currently underway at CMU to create a public *Computer Failure Data Repository (CFDR)*, sponsored by USENIX. The goal of the repository is to accelerate research on system reliability by filling the nearly empty collection of public data with detailed failure data from a variety of large production systems. Below we give a brief overview of the data sets we have collected so far, and discuss our ongoing efforts and the long-term goals of the CFDR.

2 Current data sources

The LANL data

The first data set that has been publicly released as part of the CFDR has been collected over the past 9 years at Los Alamos National Laboratory (LANL) and covers 22 high-performance computing systems, including a total of 4,750 machines and 24,101 processors. Those systems are mostly large clusters of SMP-based commodity hardware, but also include several large NUMA boxes. The data contains an entry for any failure that occurred during the 9-year time period and that resulted in a node outage. The data covers all aspects of system failures: software failures, hardware failures, failures due to operator error, network failures, and failures due to environmental problems (e.g. power outages). For each failure, the data includes start time and end time, the system and node affected, as well as categorized root cause information. To the best of our knowledge, this is the largest set of failure data studied in the literature to date, both in terms of the time-period it spans, and the number of systems and processors it covers, and the first to be publicly available to researchers (see [2] for raw data).

Node Type	#Systems	#Failures	#Nodes	#Procs.
2/4-way SMPs	18	12,607	4,672	15,101
128-256 proc. NUMA	4	8,486	78	9,000

Table 1. *The LANL data, collected 1995-2005.*

Storage failure data

Parts of our efforts have concentrated specifically on collecting storage related failure data. The reason is the potential severity of storage failures, which can not only cause temporary system unavailability, but in the worst case lead to permanent data loss. Moreover, disks have traditionally been viewed as perhaps the least reliable hardware component, due to the mechanical aspects of a disk.

We have been able to convince two high-performance computing (HPC) sites and one large internet service provider to provide hardware failure data from five different large-scale production clusters. The data sets vary in duration from 1 month to 5 years and cover a total of more than 70,000 hard drives from four different vendors. All disk drives included in the data were either SCSI or fibre-channel drives, commonly represented as the most reliable types of drives. Three of the data sets contain records for all types of hardware problems, not only storage related ones, and also contain information on the failure symptom and repair action.

Type of cluster	Duration	Total #Failures	Disk Count	Disk Type
HPC	08/01 - 05/06	1263	3406	10K RPM SCSI
HPC	01/04 - 07/06	14	520	10K RPM SCSI
Int. srv.	May 06	465	26,734	10K RPM SCSI
Int. srv.	09/04 - 04/06	667	39,039	15K RPM SCSI
Int. srv.	01/05 - 12/05	346	3734	10K RPM FC-AL

Table 2. *Overview of the hardware failure data sets.*

3 Work in progress & long-term goals

We are currently working toward three long-term goals.

Our first goal is to extend the number of data sets hosted by the CFDR to cover a large, diverse set of sites, as well as other types of data. Toward this end, we have established collaborations for data collection with another major HPC site, and two large commercial sites. We are also pursuing other types of data, including usage data (job logs and utilization measurements) and event logs, to facilitate the study of correlations between such data and system failures. For the LANL systems, we have recently added both usage data and event logs to the repository.

Second, we plan to study the existing data sets in more detail, with a focus on how the results can be used for better or new techniques for avoiding, coping and recovering from failures. For example, our initial analysis [1] of the LANL data shows that several common assumptions about failure processes (e.g. i.i.d. exponentially distributed time between failures) are not realistic in practice. One path for future work is to re-examine algorithms and techniques for fault-tolerant systems to understand where unrealistic assumptions result in poor design choices and for those cases explore new algorithms.

Third, we hope that our experiences from working with a variety of sites on collecting and analyzing failure data will lead to some *best practices* for failure data collection. Currently, data collection and analysis is complicated by the fact that there is no widely accepted format for anomaly data and there exist no guidelines on what data to collect and how. Providing such guidelines will make it easier for sites to collect data that is useful and comparable across sites.

Finally, we are seeking the assistance of all OSDI'06 attendees in making the CFDR a success by helping USENIX to identify other sources of failure data.

References

[1] Bianca Schroeder, Garth A. Gibson, "A large scale study of failures in high-performance-computing systems," In *International Conference on Dependable Systems and Networks (DSN'06)*.

[2] The raw data and additional information are available at: `www.lanl.gov/projects/computerscience/data`.