Workshop on Reliability Analysis of System Failure Data (RAF'07) MSR Cambridge, UK, March 2007.

The Computer Failure Data Repository (CFDR)

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

1 Motivation

System reliability is a major challenge in system design. Unreliable systems are not only major source of user frustration, they are also expensive. Avoiding downtime and the cost of actual downtime make up more than 40% of the total cost of ownership for modern IT systems. Unfortunately, with the large component count in today's large-scale systems, failures are quickly becoming the norm rather than the exception.

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 Collecting data

Obtaining failure data is extremely difficult due to the sensitive nature of this data. In our pursuit to create a public failure data repository, we have talked to more than a dozen companies and high-performance computing labs about contributing data. Our experiences in this process have led us to believe that it is highly unlikely to obtain data from *vendors* of IT equipment, due to their big fear of negative marketing and legal consequences. Instead, our approach has been to obtain data from large *end-users* of IT equipment, such as high-performance computing labs or internet services sites. We found that these sites are motivated to share data since they are facing a pressing need to provide reliability at scale and hope that researchers will be able to develop better solutions, if given real data to work with.

However, even obtaining data from end users is hard and sometimes impossible, since some vendors have NDAs in place with their customers that prevent them from sharing data about their systems. The most likely sites to provide data are therefore end-users that are big customers that have enough leverage with their vendors. Below we briefly describe the datasets we have been able to obtain so far.

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 [3] 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 three high-performance computing (HPC) sites and one large internet service provider to share hardware failure data from a number of large-scale production clusters. The data sets vary in duration from 1 month to 5 years and cover a total of more than 100,000 hard drives from at least four different vendors. The data include drives with SCSI and FC interfaces (commonly represented as the most reliable type of drives), as well as SATA interfaces. 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	3,406	10K RPM SCSI
HPC	01/04 - 07/06	14	520	10K RPM SCSI
HPC	12/05 - 08/06	360	14,208	SCSI & SATA
HPC	09/03 - 08/06	285	13,618	SATA
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	3,734	10K RPM FC-AL

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

3 Analyzing data

Our initial analysis of the collected failure data shows that many commonly held beliefs about failures are not realistic. Below we outline some sample results from our analysis of storage failures [2].

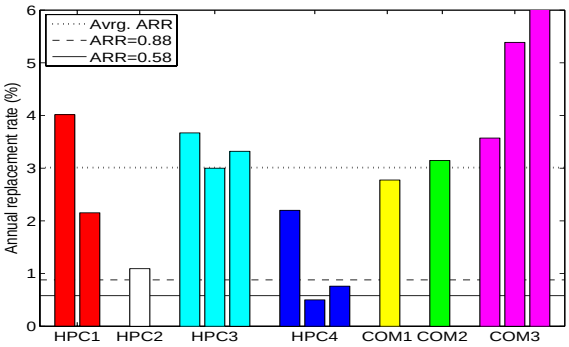

Figure 1. *Comparison of datasheet AFRs (solid and dashed line in the graph) and ARRs observed in the field.*

Manufacturers specify the reliability of a drive model in a drive's datasheet in terms of its *annualized failure rate (AFR)*, which is the percentage of disk drives in a population that fail in a test scaled to a per year estimation. We find that large-scale installation field usage appears to differ widely from nominal datasheet conditions. Figure 1 compares the annual replacement rates (ARR) for the different disk populations in our data (including only drives in their nominal lifetime of 5 years) to the datasheet AFRs (solid and dashed line). The field replacement rates were significantly higher (by a factor of 2-10) than we expected based on datasheet specifications. For drives outside their nominal lifetime (five to eight year old drives), field replacement rates were up to a factor of 30 higher than what the datasheet suggested.

Interestingly, the replacement rates of SATA disks (frequently described as lower quality) are not worse than the replacement rates of SCSI or FC disks (often believed to be the most reliable types of drives). In Figure 1, HPC4 (blue bars) consists of only SATA disks and doesn't exhibit higher ARRs than the other drive populations in the study. This may indicate that disk-independent factors, such as operating conditions, usage and environmental factors, affect replacement rates more than component specific factors.

We also find that changes in disk replacement rates as a function of drive age were more dramatic than often assumed, even during the early years of the lifecycle. Figure 2 shows the change in replacement rates as a function of drive age for one of the disk drive populations in our study. While replacement rates are often expected to be stable in year 2-5 of operation (before they start to increase due to wear-out in year 5-8), we observed a continuous increase in replacement rates, starting as early as in the second year of operation.

Finally, we analyzed the statistical properties of drive failures. A common assumption is that drive failures form a Poisson process, implying that the time between failures is exponentially distributed and that failures are independent. While many have suspected that the commonly made assumption of exponentially distributed time between failures is not realistic, previous studies have not found enough evidence to prove this assumption wrong with significant statistical confidence. Based on our analysis, we are able to reject the hypothesis of exponentially distributed time between disk replacements with high confidence. We identify as the key features that distinguish the empirical distributions from the exponential distribution, higher levels of variability and decreasing hazard rates. We find that the empirical distributions are fit well by a Weibull distributions with a shape parameter between 0.7 and 0.8.

We also find strong evidence for the existence of various types of correlations. For example, the empirical data exhibits significant levels of autocorrelation and long-range dependence.

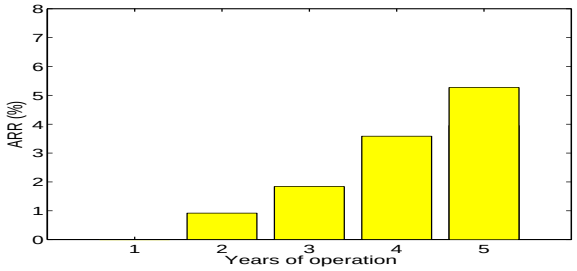

Figure 2. *ARR as a function of drive age in years.*

4 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 recent analysis of the LANL data [1] and the storage failure data [2] 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.

5 Acknowledgments

We would like to thank Gary Grider, Laura Davey and Jamez Nunez from the High Performance Computing Division at Los Alamos National Lab and Katie Vargo, J. Ray Scott and Robin Flaus from the Pittsburgh Supercomputing Center for collecting and providing us with data and helping us to interpret the data. We also thank the other people and organizations, who have provided us with data, but would like to remain unnamed. For discussions relating to the use of high end systems, we would like to thank Mark Seager and Dave Fox of the Lawrence Livermore National Lab.

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] Bianca Schroeder, Garth A. Gibson, "Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you?" In *5th Usenix Conference on File and Storage Technologies (FAST '07)*.

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