Fault-Tolerant Computer System Design
ECE 60872/CS 590

Continuous Distributions: Case Study

Saurabh Bagchi
ECE/CS
Purdue University

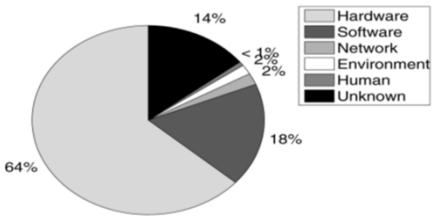
Case Study of Failures in an HPC Environment

- Paper: "A Large-Scale Study of Failures in High-Performance Computing Systems," Bianca Schroeder,Garth A Gibson, IEEE transactions on dependable and secure computing (TDSC), 2010, Vol.7(4), p.337-350.
- Data from a high-performance computing site, Los Alamos National Lab (LANL)
- Data from 22 high-performance production computing systems; 1996-2005
- Nonuniform-Memory-Access (NUMA) nodes, or two-way and four-way Symmetric-Multiprocessing (SMP) nodes.
- Total = 4,750 nodes, 24,101 processors.

Data Collection

- "Remedy database" created at LLNL in 1996
  - Failures are detected by an automated monitoring system that pages operations staff whenever a node is down
  - Operations staff create a failure record in the database; start time, system and node affected; then turn the node over to a system administrator for repair
  - Upon repair, the system administrator notifies the operations staff who then put the node back into the job mix; record end time of the failure
- Workload types: compute, graphics, or fe (front-end)
- Root causes: Human error; Environment, including power outages or A/C failures; Network failure; Software failure; and Hardware failure

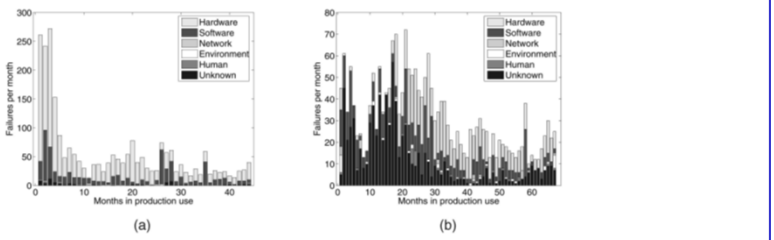
Root Cause Breakdown



- Hardware is the single largest component, with more than 50% of all failures
- Software is the second largest contributor, with around 20% of all failures
- The number of failures with unidentified root cause is significant: 14%

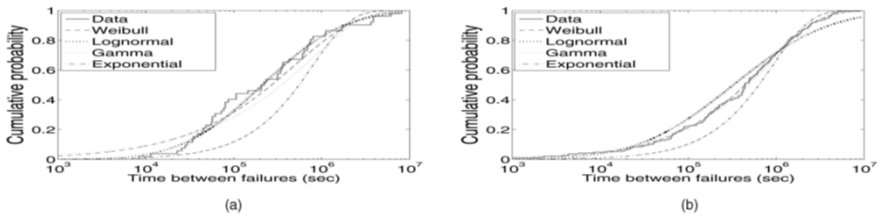
Failure Rate Analysis

- Look at failure rates over the entire lifetime of a system
- Failure rate as a function of system age follows one of two shapes



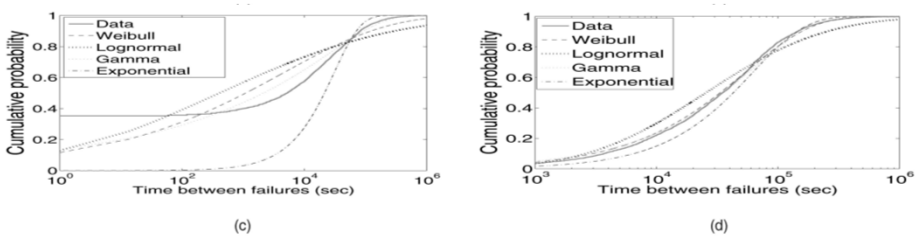
- (a) System 5 – Most common shape; failure rate drops during the early age of a system, as initial hardware and software bugs are detected and fixed and administrators gain experience
- (b) System 19 - Failure rate actually grows over a period of ~20 months, before it starts dropping.
  - Getting these systems into full production was a slow and painful process (?)

Time Between Failures (Node 22 – System 20)



- Early stage (Figure (a)): distribution between failures is well modeled by a Weibull or gamma distribution. Both distributions create an equally good visual fit and the same negative log-likelihood.
- The simpler exponential distribution is a poor fit
- Later stage (Figure (b)): the best fit is provided by the lognormal distribution, followed by the Weibull and the gamma distribution.
- The exponential distribution is an even poorer fit
- A numerical measure of goodness of fit is negative log likelihood

Time Between Failures (System 20)



- Early stage (Figure (c)): Not well captured by any of the standard distributions.
  - The reason is that an exceptionally large number (>30 percent) of interarrival times are zero, indicating a simultaneous failure of two or more nodes.
  - This indicates the existence of a tight correlation in the initial years of this cluster.
- Late stage (Figure (d)): Basic trend for system similar to the per node view