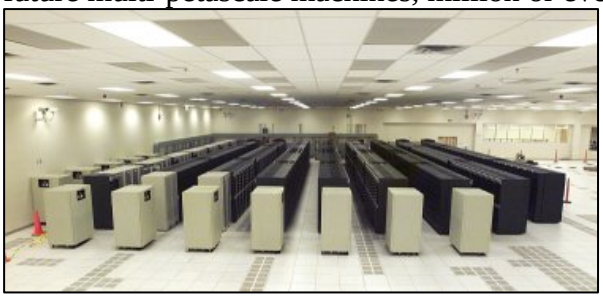
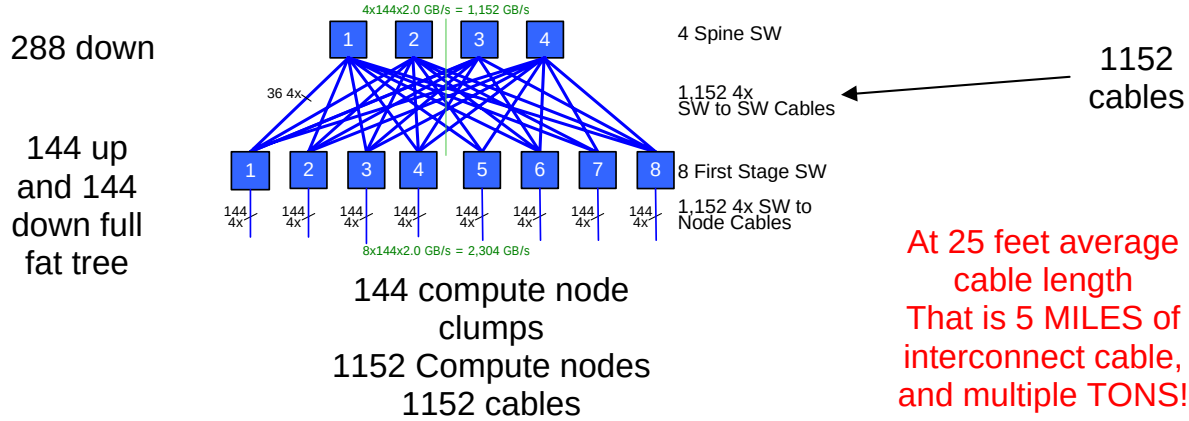
Statistical Analysis of Supercomputing Usage and Event Data (Toward Failure Prediction in a Large Supercomputing Environment)

Motivation

Large simulation codes running on large leadership class supercomputers are doing important science. Current leadership class machines have a few hundred thousand way parallelism. For future multi-petascale machines, million or even multi-million way parallelism is expected.



The above machine occupies about 1/3 acre and has 8000 processors, and 400 Terabytes if storage. Interconnect networks for existing machines consist of thousands of cables and many thousand parts.



The above diagram is a fat tree network of a modest supercomputer with 1152 nodes.
The latest supercomputer installations have over 10,000 disk drives. These environments are very complex. Typical applications that run on these systems involved one to several thousand processing units and often run for 6-18 months 24 hours per day. Failure is so common that applications stop computation and checkpoint their memory state to parallel file system disks ever few hours so they can restart after a processing unit failure. Mean Time To Interrupt (MTTI) for these jobs is typically a few hours. As supercomputers get bigger and bigger, it will become more important to understand failure at scale. Approaches for both avoiding failure and surviving through failure must be explored.

Goal

One way to learn more about failure in these environments is to study the operational data produced by these environments. LANL has kept interrupt data for every interrupt in this environment that has effected a running application for decades. This is valuable data, as it can be studied to look for ways we might predict or better deal with failure. A nice study by Garth

Gibson and Bianca Schroeder at Carnegie Melon University was done on the LANL interrupt data. This paper and other papers on LANL interrupt data are available at http://institute.lanl.gov/data/fdata/ . It is difficult to understand interrupt data without also looking at how these machines are used. For this reason LANL has kept usage data as well. Additionally, some machines at LANL have internal monitors for monitoring events that happen in that machines environment. LANL has kept this data as well. Additionally, it is important to understand what the machines look like and something about their environments to be able to work with these data. All of this data/information is available at http://institute.lanl.gov/data/fdata/ .

The goal of this work will be to characterize the usage and event data and look for correlations or relationships between the usage, event, environment, and interrupt data. The desire is to find out useful relationships that might be exploited to prevent, predict, or better deal with failure. The information should be provided in a co-authored LANL/Mines publishable paper or technical report form, suitable for submission to a relevant computer science conference. It is also expected that a presentation will be made at the end of this project to LANL describing the process and findings.

Student/Mentor/Machine Requirements

The students must have base knowledge in studies of raw data, simple statistics, correlation, regression, and time series analysis. Any statistical tool may be used, perhaps the R tool would be best given it is public. Most any sized modern personal computer/server could be used to do the analysis with, given these data add up to a few hundred megabytes in size and several million records. The statistics package chosen must be able to do the above statistical methods and also provide graphical output suitable for publication. The adviser should be able to organize the data analysis and paper writing activities to ensure a good measurement style academic publication will result. It is envisioned that the work can be split up into different characterization and analysis which would enable a multi-person team.

Communication with LANL

LANL primary and secondary contacts on this project will come to Golden at the beginning of the project and also at the end. Interim communication can be via email, teleconference, and we also have polycom and access grid video teleconferencing capabilities which work well for remote collaboration between small teams. At the initial face-to-face meeting, an explanation of the data, environment, and ideas as to how to proceed would be agreed to. LANL has a lecture on supercomputing I/O and we would be happy to give this lecture via video conference or in person during the initial visit. If Mines has polycom capability, there might be other supercomputing seminars that we might be able to arrange.

Preliminary Project Outline

- Read in and do sanity check of all relevant fields in the usage data, to get a feel for how clean the data is from an outlier/bad value point of view
- Read in and do sanity check of all relevant fields in the event data, to get a feel for how clean the data is from an outlier/bad value point of view
- Perform characterization of usage data (example few of the myriad of characterizations)
    - mostly over time

  - when are machines more heavily used, at night or during the week, etc.
  - has the workload changed over time
  - is it different for different sized jobs
  - is it different for different machines or different nodes of machines
  - was the machine heavily used early, mid or later in its lifetime
  - do users have profiles, does one user run most of the big jobs or a mix, does this change over time
  - are their common patterns across machines are their patterns unique to a particular machine

  I would expect this characterization would provide much information about how r people use supercomputers. This would require interaction with LANL as you might see a spike in usage associated with a milestone or other event that you wouldnt know about, with access to only the usage data. I would expect this portion of the study would generate lots of interesting time graphs and a lot of characterization discussion and text.
- Perform characterization on event data (example few of the myriad of characterizations)
  - including time series analysis
  - the event data will have less characterization, but similar things could be done looking at events over time
  - do events happen in clumps or spread out
  - do some events predict or highly correlate to other events
  - are their patterns of events, do the patterns change over time like during the life of a machine
  - do some events only happen during the infancy of a machine or late in its life
- look for simple correlating values between interrupt data and usage data (examples only)
  - this is of course looking for predictors of failure/interrupts in the usage data
  - we would expect highly used machines to fail more but that might not be true, it might only be certain kind of use or at certain times, etc.
- look for simple correlating values between interrupt data and event data (examples only)
  - this is of course looking for predictors of failure/interrupts in the event data
  - we would hope that events like high temp warnings etc. might predict future failure
  - this may not be simple, there may be many factors together that might predict interrupts or perhaps only certain kinds of interrupts
- look for combinations of event and usage data for correlation with failure (examples only)
  - this will be complex multivariate statistical analysis, one might even want to try to create predictor indexes that combine usage and event data etc.
- look for simple correlation between machine placement and interrupt data and event data
  - this is looking to see if where the machine is located in the room matters
  - this might be important for air flow issues in the room, vibration, etc.

Careful characterization and analysis of these data may provide incite into how these machines/environments are used, what real information machine monitors provide and how it relates to failure. The above functions are only examples. It is quite possible that in the process of characterization and analysis, a possible candidate relationship might pop up and we would want to dig deeper in that area.

Student Experience

This project will provide the students with:

- An introduction to LANL, what a National Lab does, and how we use supercomputers
- An introduction to supercomputing at extreme scale and an understanding of how supercomputing centers operate and how supercomputers are built, maintained and used
- A chance to work to produce analysis to assist in solving a world class computer science problem in high performance computing
- A chance to see how application of statistics and data analysis of operational data might benefit an organization
- A chance to collaborate with staff at a National Lab and be introduced to a small portion of a real problem that need to be solved
- A chance to be co-author on an academic paper that might end up being published in a relevant journal or conference proceeding. This will be helpful if you intend to go on to graduate school or if not, being published helps round out a resume. Co-authoring a relevant paper with a National Lab would look very good on a resume.