& Resource Usage

FRESCO: An Open Failure Data Repository for Dependability Research and Practice

Saurabh Bagchi, Carol Song (Purdue University)

Ravi Iyer, Zbigniew Kalbarczyk (University of Illinois at Urbana-Champaign)

Nathen DeBardeleben (Los Alamos)

Presentation available at: *engineering.purdue.edu/dcsl*

Roadmap

- Motivation
  - Why do we need an open data repository?
  - What are our plans
- Context
  - Large computing cluster at Universities
  - Demography of the cluster users
  - Challenge in supporting user needs
- Insights from analysis of Purdue's cluster
  - Cluster environment
  - The data set
  - Analysis and Results
  - Current status of the repository
  - The next steps

Motivation

- Dependability has become a necessary requisite property for computing systems that we rely on
- Dependable system design should be based on real failure modes of systems

BUT

- There does not exist any open failure data repository today for any recent computing infrastructure that is
  - large enough,
  - diverse enough, and
  - with enough information about the infrastructure and
  - the applications that run on them

PURDUE UNIVERSITY

So what do we do about it?

1. Stop bemoaning the lack of publicly available dependability dataset and start building one
2. But, who will want to share such data publicly?
3. Start with local IT organizations: Purdue and UIUC
4. Get NSF backing – some of the IT clusters have been set up with NSF funding
5. Collect an initial dataset for a small time window of
   1. Applications and libraries
   2. Resource usage – node and job level
   3. Health information – node and job level
6. Release the dataset after heuristic-based anonymization

PURDUE UNIVERSITY

National Science Foundation Context

- Planning grant from the National Science Foundation (NSF) in 06/14-06/15, $100K
  - Computational and Information Sciences Directorate (CISE)
  - Computing Research Infrastructure (CRI) Program
  - Deliverable: Data collection tools in place on Purdue's IT infrastructure; Requirements gathering
- Regular grant from NSF: 3 years, started 07/15, $1.2M
  - Involves UIUC with their Blue Waters cluster
  - Delivered: First release of the dataset with DOI
  - Deliverable:
    - Large set of diverse data, from the PI/co-PI's institutions plus others
    - Large set of users
    - Large set of analytical tools made available by the community

University Compute Cluster Context

- Computing clusters at university is not uncommon
- Users have a varying level of expertise
  - Writing own job scripts
  - Using scripts like a black box
- Varying user needs
  - High computation power
    - Analysis of large structures (Civil, Aerospace engineering)
  - High Lustre bandwidth for file operations
    - Working with multiple large databases/files (Genomics)
  - High Network Bandwidth
    - A parallel processing application

Goals of Data Analysis & Related Efforts

- Goals for data analysis of cluster resource usage and failures
  - Improve cluster availability
  - Minimize the maintenance cost
  - ``Customer satisfaction”
- Several synergistic efforts
  - XDMoD (U of Buffalo): NSF-funded project to audit utilization of the XSEDE cyberinfrastructure by providing a wide range of metrics on XSEDE resources, including resource utilization, resource performance, and impact on scholarship
  - XALT (U of Chicago, UT Austin): NSF-funded project to collect and understand job-level information about the libraries and executables that jobs use
  - [Past project] Computer Failure Data Repository (CFDR) (CMU, U of Toronto)

Purdue Cluster Details and Initial Dataset

Cluster Details

- Purdue's cluster is called Conte
  - 580 homogeneous nodes
  - Each node contains two 8 core Intel Xeon E5-2670 Sandy Bridge processors running at 2.6 GHz
  - Two Xeon Phi 5110P accelerator card, each with 60 cores
  - Memory: 64GB of DDR3, 1.6 GHz RAM
- 40Gbps FDR10 Infiniband interconnect along with IP
- Lustre file system, 2GB/s
- RHEL 6.6
- PBS based job scheduling using Torque
- Conte is a ``Community" cluster

Cluster Details

- Scheduling in Conte:
  - Each job requests for certain time duration, number of nodes and in some cases, amount of memory needed
  - When job exceeds the specified time limit, it is killed
  - Jobs are also killed by out-of-memory (OOM) killer scripts, if it exhausts available physical memory and swap space
- Node sharing:
  - By default only a single job is scheduled on a an entire node giving dedicated access to all the resources
  - However, user can enable sharing by using a configuration in the job submission scripts

Conte's Data Set

- Data set spans Oct '14 – Mar '15 (6 months)
  - ~500k jobs (489, 971 jobs)
  - 306 unique users
- Per job data
  - Accounting logs from PBS scheduler
    - Job owner details
    - Start/end time, resource requested/used
  - List of shared libraries used (using lsof)
- Node-wise performance data
  - Collected using Tacc stats
    - Lustre, Infiniband, Virtual memory, Memory and more…
- Syslog messages

System Usage Repository

- Objectives

URL: https://diagrid.org/resources/247

  - Enable systems research in dependability that relies on system usage and failure records from large-scale systems
  - Provide synchronized workload information, system usage information, some user information, hardware status information
  - Provide this repository for a diversity of workloads and diversity of computing systems
  - Provide common analytic tools to allow for dependability-related questions to be asked of the data

Questions for the Repository

URL: https://diagrid.org/resources/247

- What is the resource utilization (CPU, memory, network, storage) of jobs with a certain characteristic?
  - Characteristic could be the application domain, the libraries being used, parallel or serial, and if parallel, how many cores
  - We want to know this especially if resource utilization is anomalously high
- What is the profile of users submitting jobs to the cluster?
  - Are they asking for too much or too little resources in the submission script?
  - Are they making the appropriate use of the parallelism?
  - What kinds of problem tickets do they submit and how many rounds are needed to resolve these tickets?

Current status of the repository

- Workload traces from Conte
  - Accounting information (Torque logs)
  - TACC stats performance data
  - User documentation
- Privacy
  - Anonymize machine specific information
  - Anonymize user/group identifiers
- Library list is not shared
  - For privacy reasons

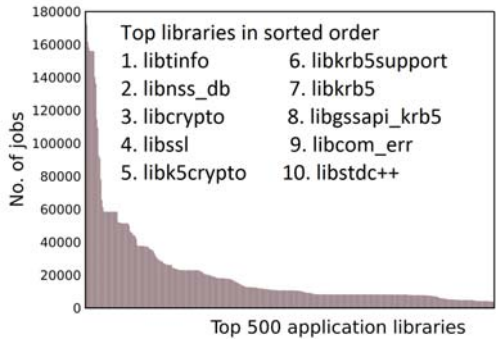
Insights from Data Analysis

- Few libraries (not pre-installed) are highly popular
- 45% jobs use less than 10% requested time
- 70% jobs use less than 50% requested memory
- 20% jobs showed memory thrashing
- Memory thrashing behavior of jobs the share the node and jobs that do not are exactly opposite
- Few jobs place very high demand for I/O and network resources

PURDUE UNIVERSITY

Hot Libraries

- Extract all dynamically linked libraries being used by the applications

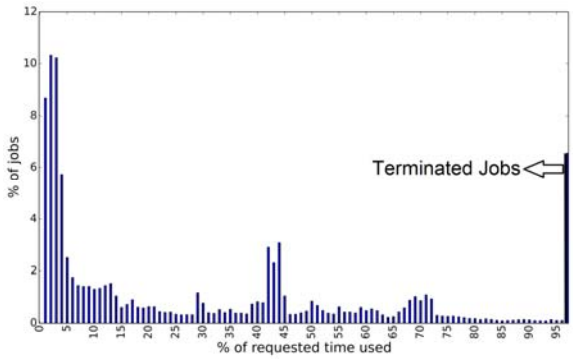


Sorted histogram of top 500 libraries as used by the jobs

1. Out of a total 3,629 unique libraries, some are used much more often
2. Each of the top 50 libraries is used by more than 80 % of the users

PURDUE UNIVERSITY

Resource Request Patterns



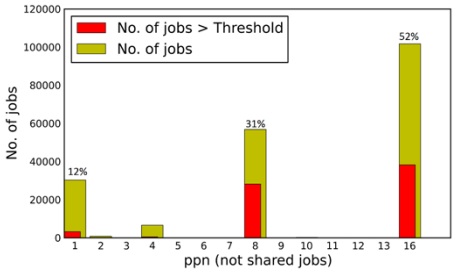
Percentage of the user requested time actually used by the jobs

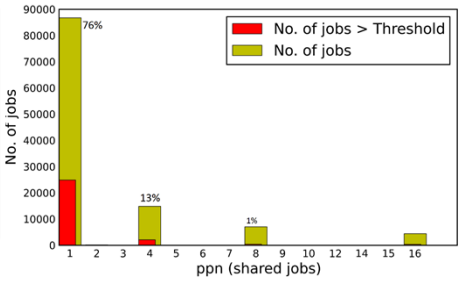
1. Almost 45% of jobs actually used less than 10% of requested time
2. But, scheduler during busy periods gives higher priority to shorter jobs

PURDUE UNIVERSITY

Shared v/s non-shared Jobs

- Jobs that share node have higher thrashing compared to jobs that do not share the node





PURDUE UNIVERSITY

Project Plans

PURDUE UNIVERSITY

Immediate next steps

- Continued data collection on Purdue clusters
- Analysis of Blue Waters logs
  - Analyze workloads on similar lines as Conte
  - Investigate the similarities and differences in the results
  - Identify the user behavior and workload characteristic for better cluster management
- Close the loop
  - Implement the remediation measures, e.g., move some jobs to a different cluster, increase the memory asked for
  - Check the effect of the remediation measures

PURDUE UNIVERSITY

A Wish List for Types of Data

- Accounting information
  - Job Owner ID, group ID
  - Start/End time
  - Resources requested/used
    - Memory, CPUs, Walltime ,etc
- Performance statistics
  - I/O usage (Lustre, Disk)
  - Network usage (IP, Infiniband)
  - Memory usage
  - Virtual memory statistics
- Library List per job
  - Shared objects used by the job

A Wish List for Types of Data

- User Tickets
  - Problem and resolution
- Failure resolution reports
  - Failure description
  - Root cause identification
  - Issue resolution

Conclusion

- It is important to analyze how resources are being utilized by users
  - Scheduler tuning, resource provisioning, and educating users
- It is important to look at workload information together with failure events
  - Workload affects the kinds of hardware-software failures that are triggered

Contributors

Purdue: Suhas Javagal, Subrata Mitra, Chris Thompson, Stephen Harrell, Chuck Schwarz

UIUC/NCSA: Saurabh Jha, Joseph Fullop, Jeremy Enos, Fei Deng, Jin Hao

DOE: Todd Gamblin, Ignacio Laguna, Dong Ahn

Thank you