High Performance Computing Resilience Testbed

Héctor J. Machín
University of Turabo
Research Alliance in Science and Math
Computer Science and Mathematics Division, Oak Ridge National Laboratory
Mentors: Stephen L. Scott, Thomas Naughton, Geoffroy Vallée, Anand Tikotekar

Motivation and Research Goal

- Probability of failures in HPC systems
  - Systems increase in size
  - Complex scientific computing applications
- Software and Hardware Failures
  - Increases application interruption
  - Decreases productive use of resources
  - Wastes allocation hours (costs)
- Checkpoints: fault tolerance in HPC systems
  - Application periodically stops useful work
  - Writes checkpoints to disk
- Transition to petascale computing †
  - Multicore size doubling each 30 months
  - System utilization drops to zero by 2013
  - 100% time only writing checkpoints

† B. Schroeder, G. Gibson – "Understanding Failures in Petascale Computers"

- Develop a resilience testbed
  - Fault injection techniques
  - Data gathering/monitoring
  - Analyzers for failure evaluation

Resources Used

Ganglia

- Scalable distributed monitoring system
  - gmond – metric collecting daemon
  - gmetad – storage daemon, round robin databases (RRD)
  - gmetric – tool to inject custom metrics

Ovis 1.1

- Real-time monitoring of clusters
  - RRD readers – read the RRD databases
  - RRD starter – start instances of RRD readers for all nodes
  - RRD killer – kills all RRD readers

Cbench

- Perl-based framework for HPC Systems
- Building, running and analyzing outputs of opensource codes

HPCC Parallel Benchmarks

- Test suite that examine the performance of HPC architectures
- Kernels with memory access patterns

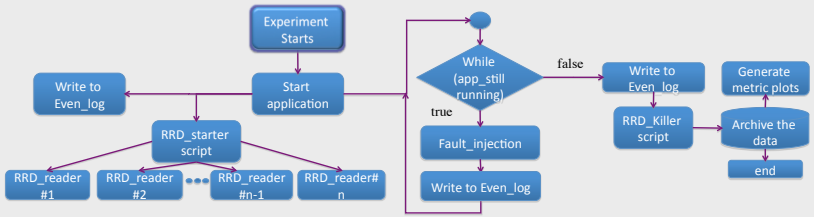

Fig 1. Research Methodology

System Name	# of Nodes	Procs Per Node	CPU Type	CPU Speed	Memory per Node
XTORC	64	1	Pentium IV	1695.7MHz	768MB

Table 1. System Specifications

Fig 2. XTORC Cluster System

Results and Analysis

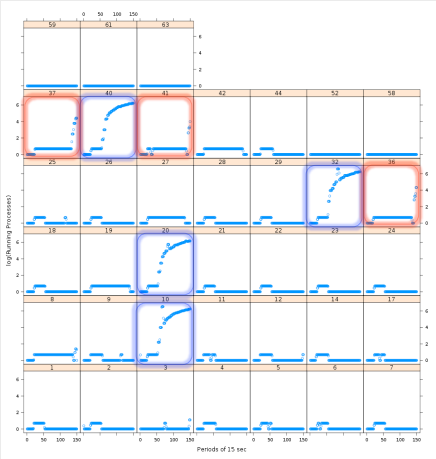

Fig 3. Running Processes vs. Periods 15 sec.

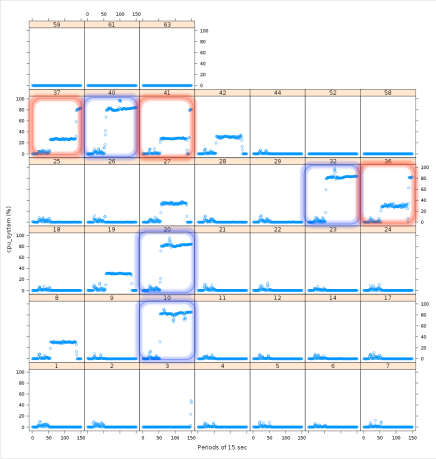

Fig 4. CPU System Percent vs. Periods 15 sec.

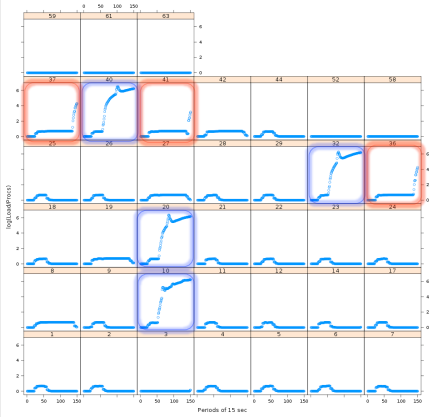

Fig 5. Load/Processor vs Periods of 15 sec.

Summary

- An emphasis on educational process
- Automation of several processes
  - Fault Injection
  - Data gathering/monitoring
  - Archive of data and events
  - Tools for failure visualization

Future Work

- Explore more fault generation techniques
- Explore the integrity of the faults injected
- Extended research on more complex HPC Systems

The Research Alliance in Math and Science program is sponsored by the Office of Advanced Scientific Computing Research, U.S. Department of Energy. The work was performed at the Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC under Contract No. De-AC05-00OR22725. This work has been authored by a contractor of the U.S. Government, accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

I would like to thank my mentors for the opportunity to work on this project. Also, I would like to thank Thomas Naughton for his unconditional and continued support. Finally, special thanks go to Debbie McCoy, who made provisions for this research experience. For more information visit my website: http://wiki.ornl.gov/sites/rams09/h_machin/Pages/default.aspx