University of Reading

School of Systems Engineering
MSc Dissertation Presentation

Title of presentation:

RAS Framework Engine Prototype

Name of Student: **Antonina Litvinova**

Supervisors: **Dr. Christian Engelmann**

Dr. George Bosilca

OAK RIDGE National Laboratory

THE UNIVERSITY of TENNESSEE UT

Director of the course:

Professor Dr. Vassil Alexandrov

22/09/2009

Contents

- Motivation
- Reliability, Availability and Serviceability (RAS) Framework for HPC systems
- Contribution
  - RAS Framework Engine Prototype
- Results
- Future work
- Acknowledgement, Paper and Questions

Motivation

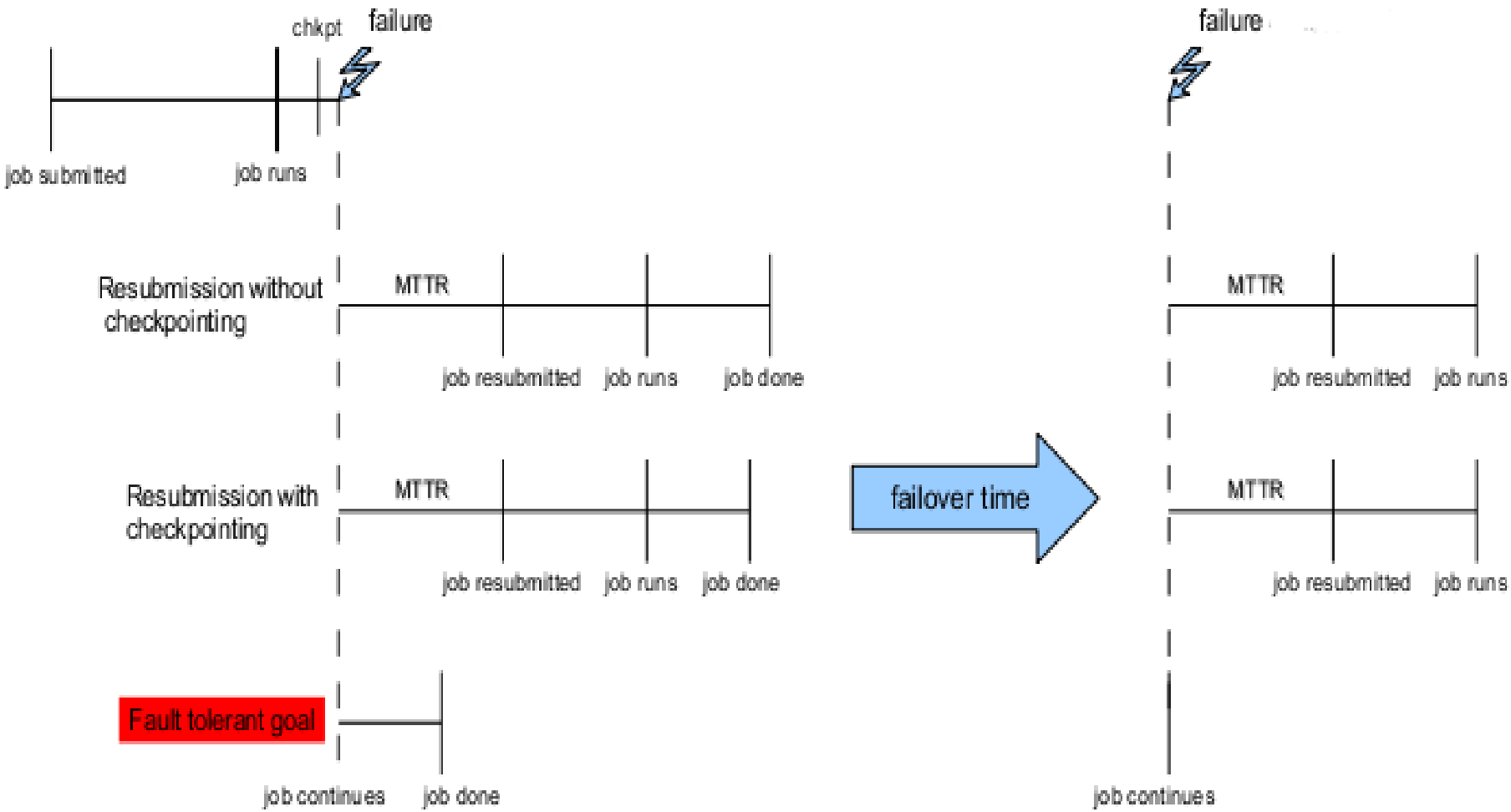
- High Performance Computing Systems reached Petaflops era
  - Roadrunner with 129,600 cores, 98 TB RAM (at Los Alamos National Laboratory, USA)
  - Jaguar with 150,152 cores, 300 TB RAM (at Oak Ridge National Laboratory, USA)
  - Kraken with 66, 000 cores (at Oak Ridge National Laboratory/University of Tennessee, USA)

- The trend is toward even larger-scale systems
- Significant increase in number of component and complexity
- Increase in failures
- Decrease in performance
- Reactive fault tolerance as checkpoint/restart is becoming less efficient
  - Experience failures
  - React to failures

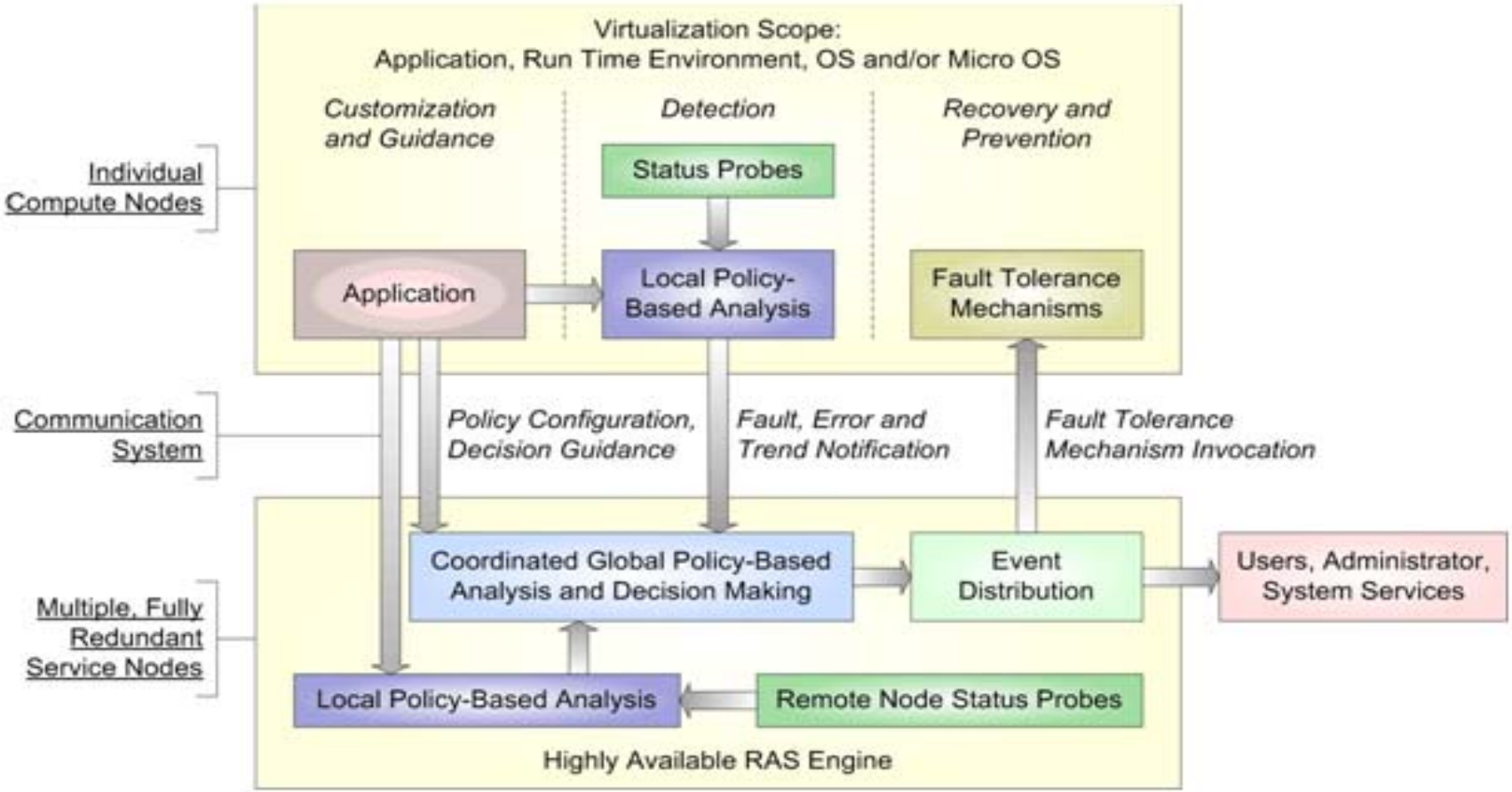
Proactive Fault Tolerance

- Proactive fault tolerance keeps parallel applications alive by
  - Avoid failures
  - Predict failures
  - Migration

Reactive vs. Proactive



RAS Framework



Reliability, Availability and Serviceability of a System

- Reliability
  - Hardware and software performance
  - Avoidance and robustness
  - Mean time between failures (MTBF)
- Serviceability
  - Component, device or system maintained and repaired

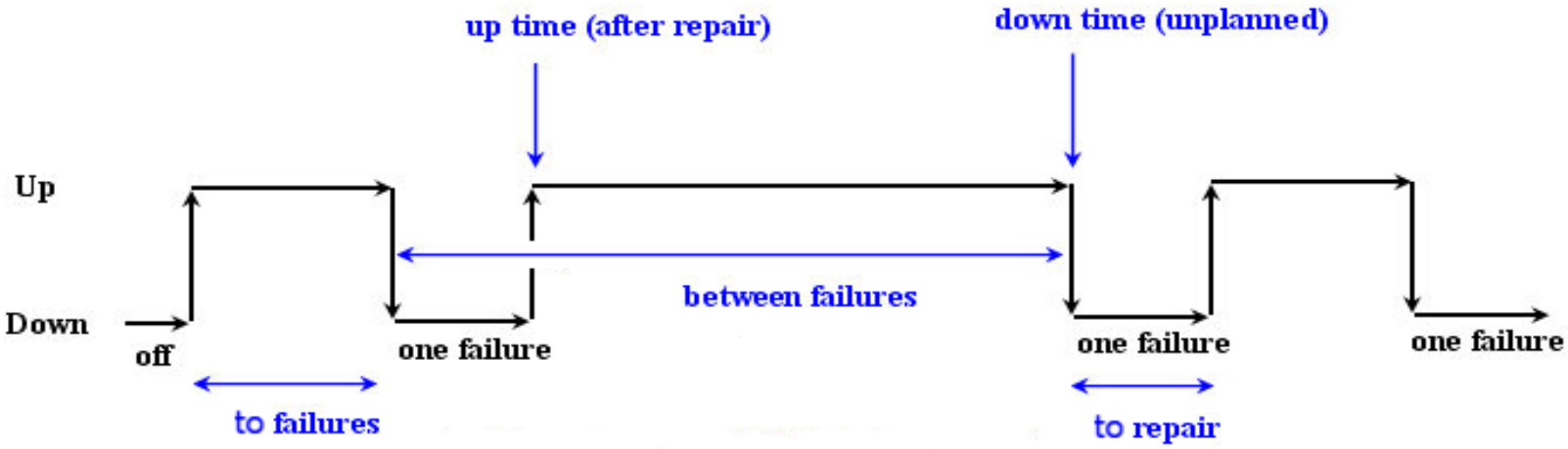
Reliability, Availability and Serviceability of a System

Mean time to failure (MTTF)

Mean time to repair (MTTR)

Mean time between failures (MTBF=MTTF+MTTR)

- Availability is MTTF/MTBF

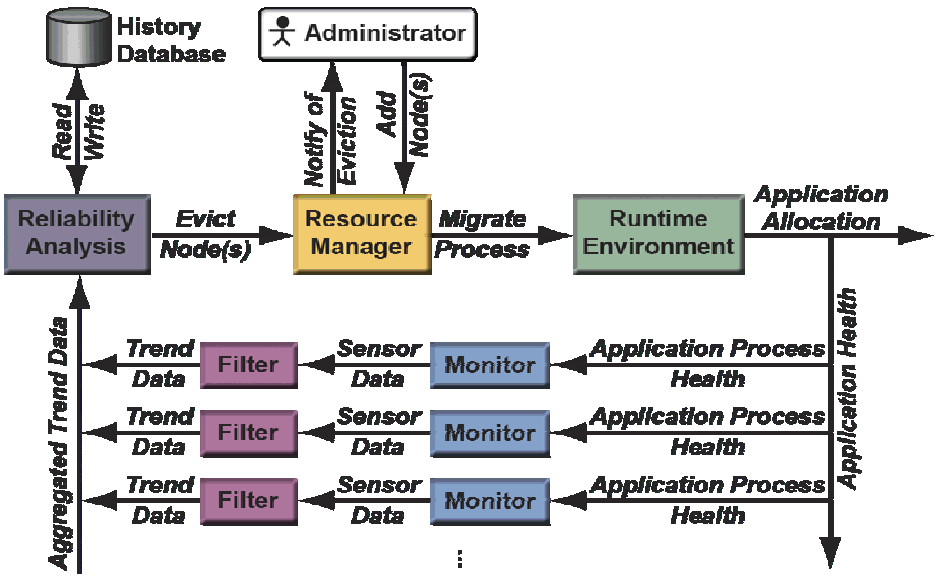


RAS Framework

- Reactive fault tolerance
- Proactive fault tolerance
- Reliability analysis
- Holistic fault tolerance
  - Reactive and Proactive

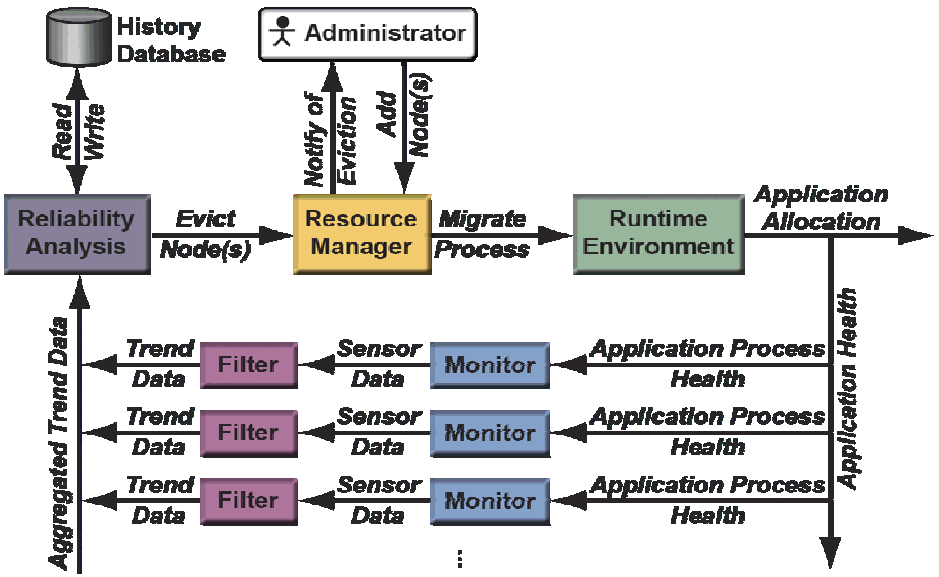
RAS Framework Engine Prototype

- Job and Resource Manager
- Monitoring System
- Event Logging
- Reliability Analysis
- Database
- Migration



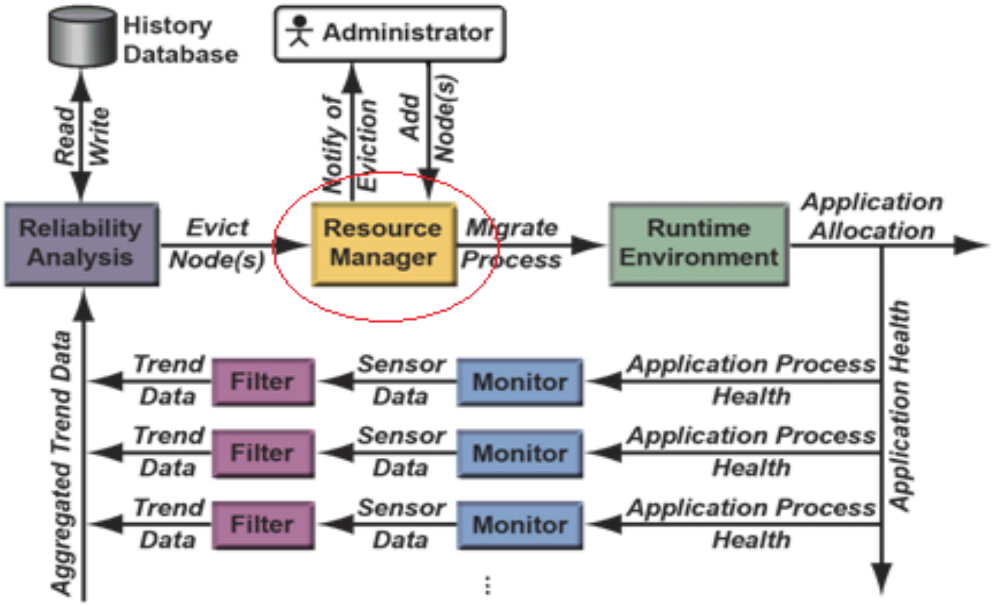
Contribution to RAS Framework Engine Prototype

- Database
- Daemons are interfaces between the Database and:
  - Monitoring System
  - Event Logging
  - Resource Manager



Job and Recourse Manager

- Accepts jobs
- Finds resources
- Submits jobs
- Provides the result

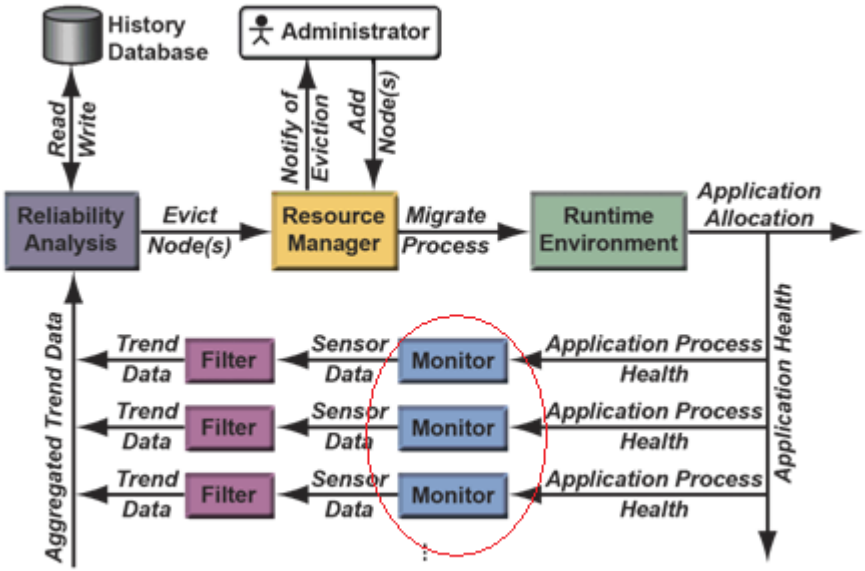


Must support migration

Collected data for Reliability Analysis

Monitoring System

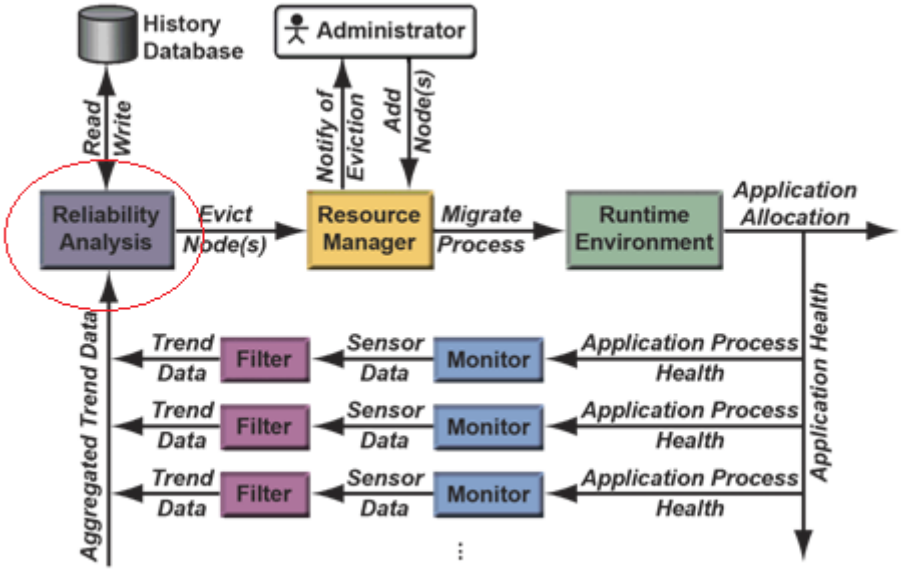
- Monitors resources
- Provides metrics values



- Collected data for Reliability Analysis

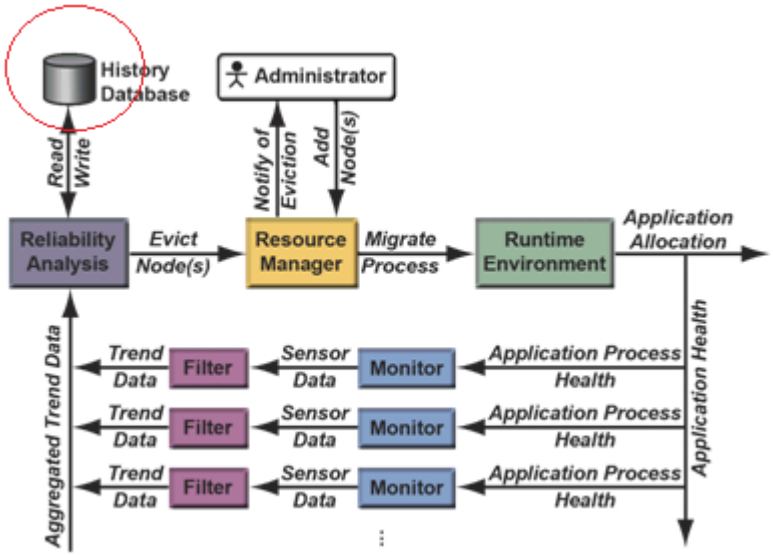
Reliability Analysis

- Analyses
- Makes predictions
- Trigger migration
- Stores data to the database



Database

- Has historical data and raw data
- Used by Reliability Analysis
- Archive data



Migration

- Job level
- Process level

- From a compute node to a node
- From a processor to a processor

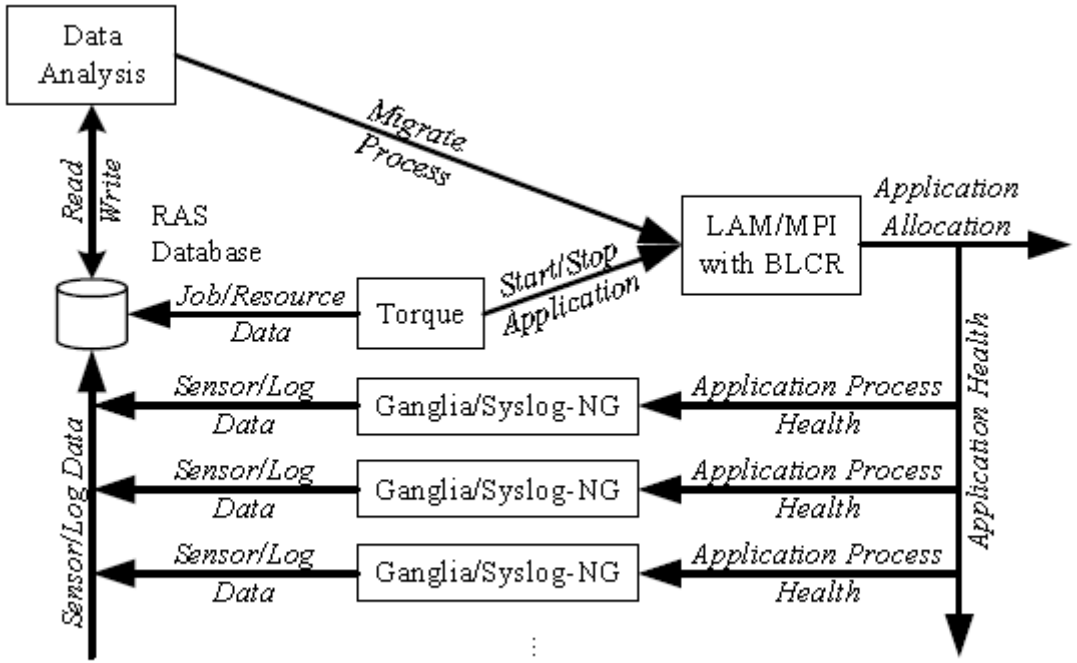
RAS Framework Engine Prototype

Uses:

- Torgue
- Ganglia (gmond)
- Syslog-ng
- MySQL

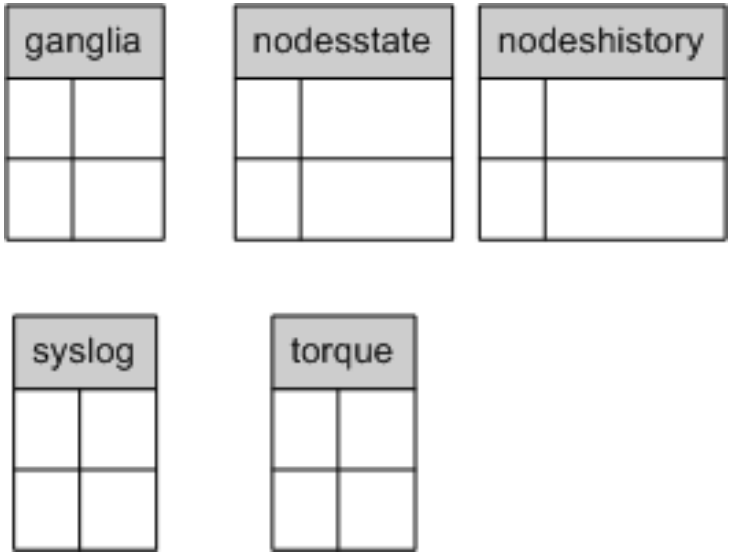
Consist of daemons:

- *Torquemysql*
- *Gangliamysqld*
- *Syslogmysqld*
- *Migrationd*
  and
  ras Database



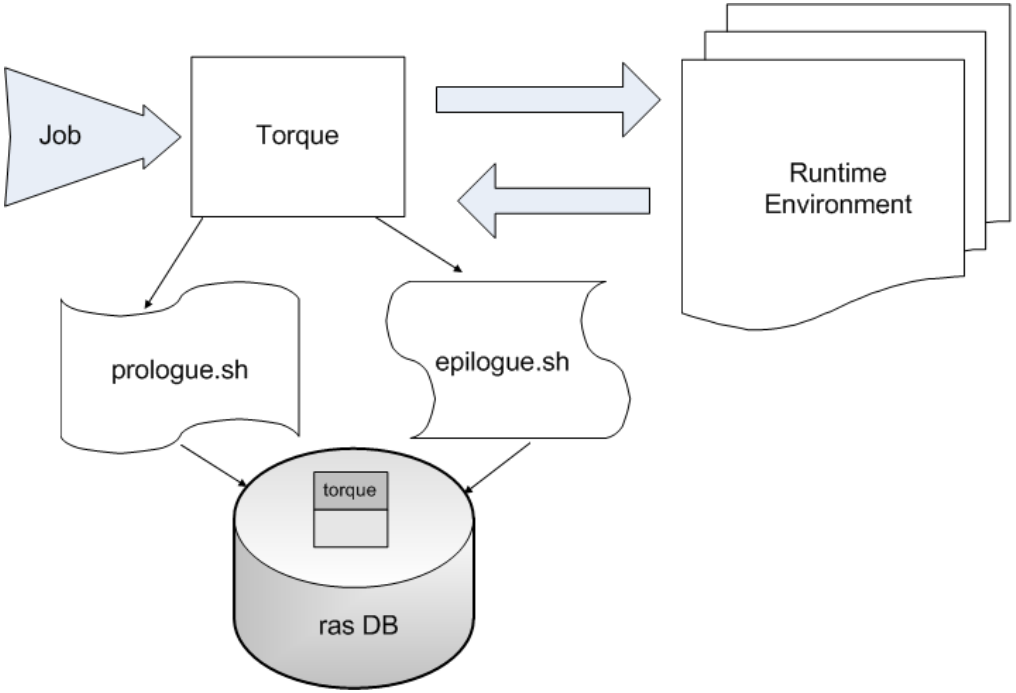
ras Database

- 5 tables
- MySQL
- No relations between tables
- Data types:
  - Int
  - Varchar
  - text



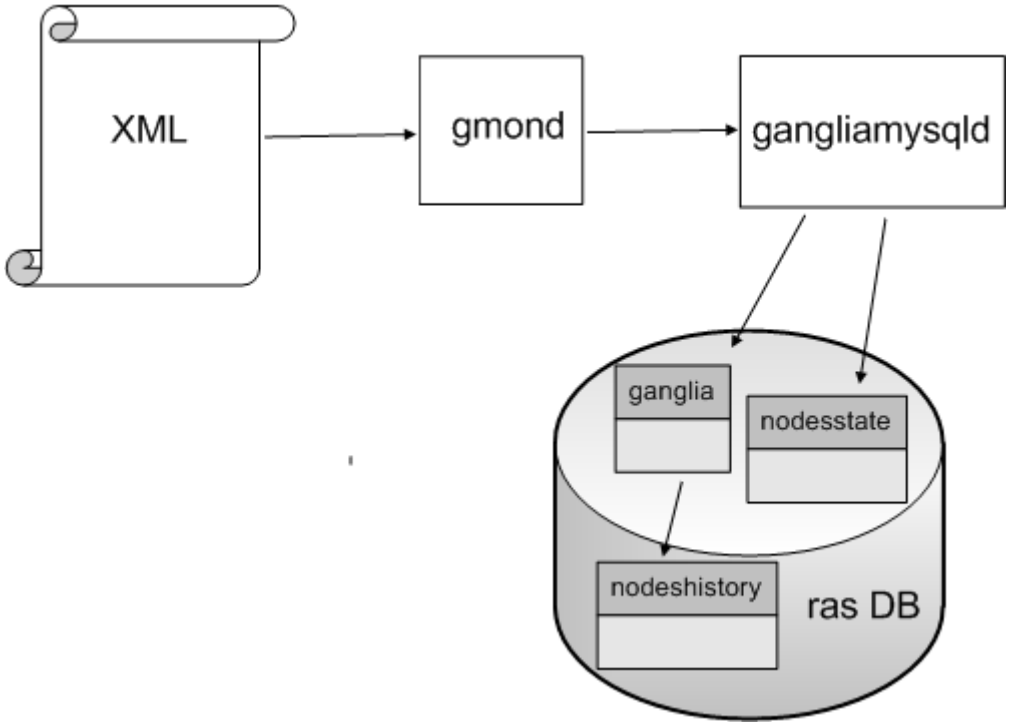
Torque/MySQL scripts

- Prologue script
- Epilogue script
- SQL statements



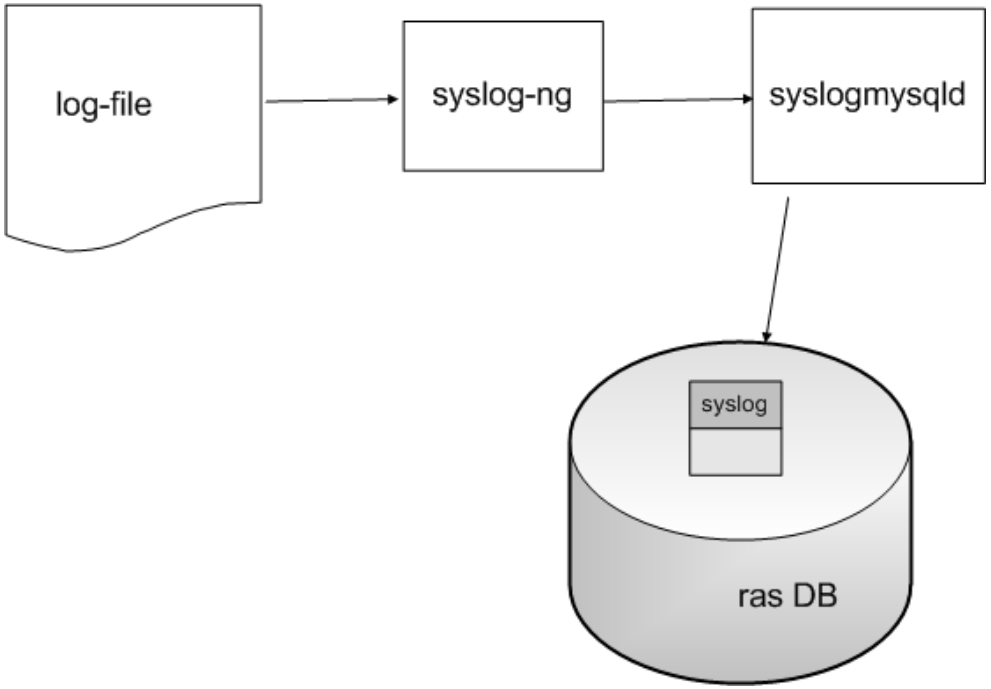
Ganglia/MySQL Daemon

- Exports in XML
- XSLTproc .xsl .xml -> .sql

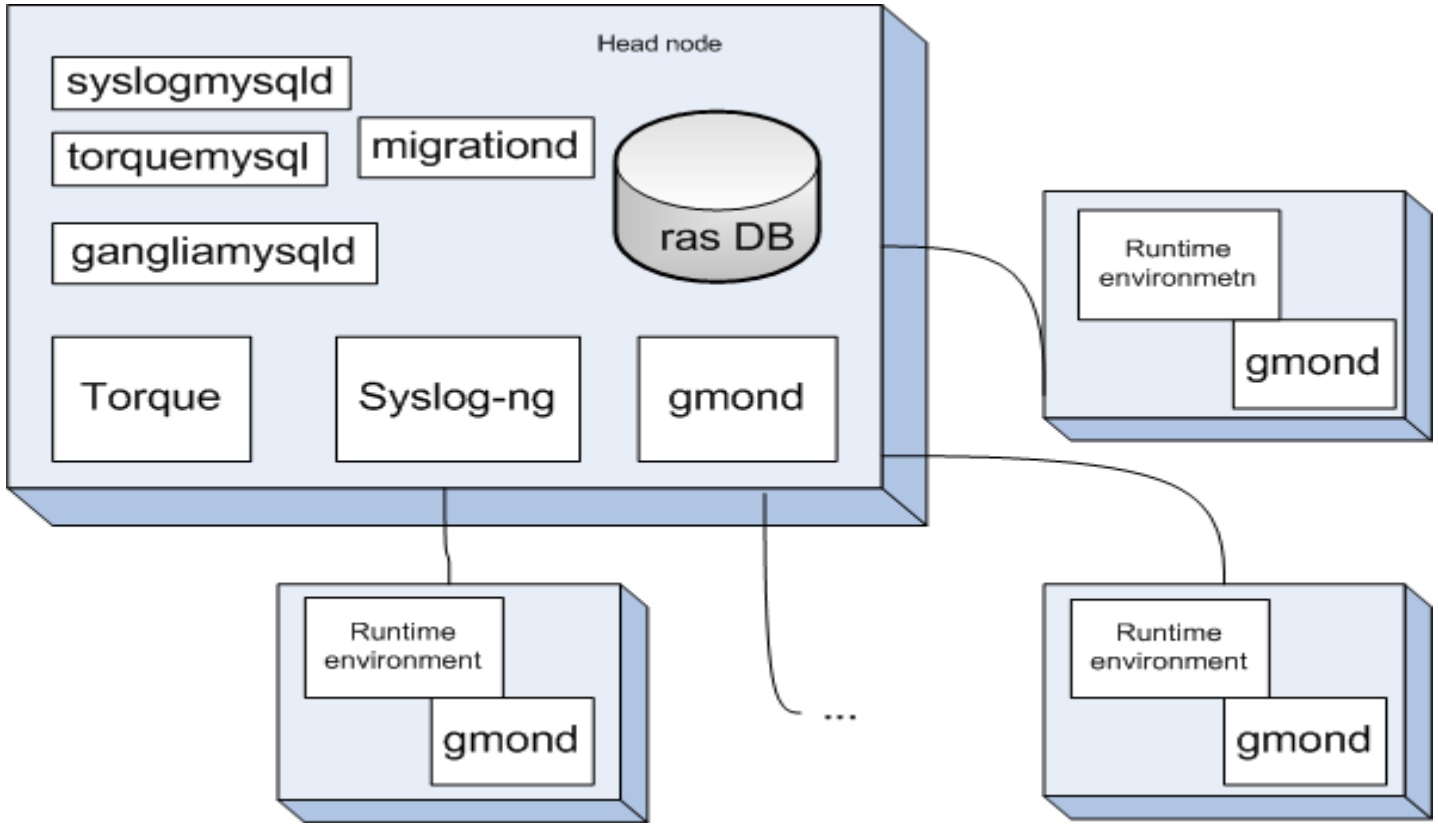


Syslog-ng/MySQL daemon

- Gets messages
- SQL statements



RAS Framework Engine Prototype integration on a system



Results

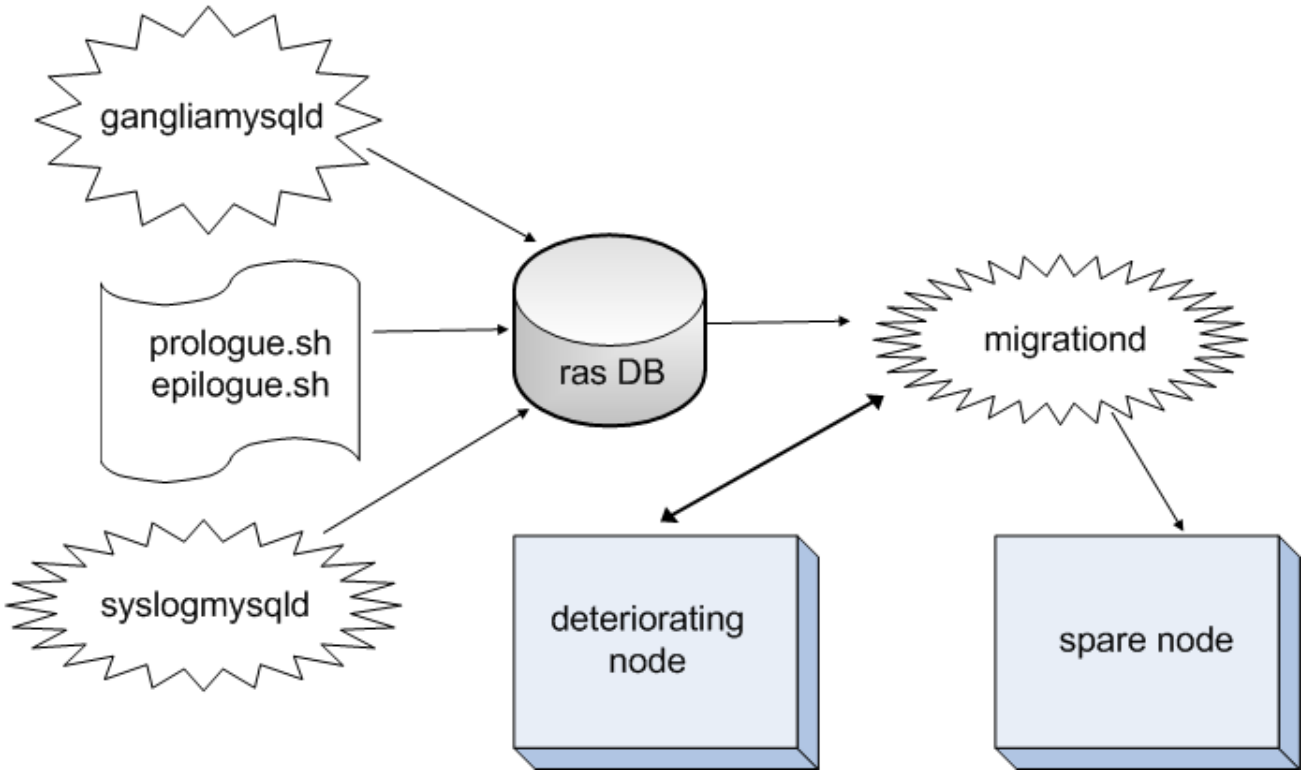
- Tested on a 48 node Linux cluster at ORNL
- Data stored from Ganglia, Torque, Syslog-ng

Limitations

- Database scalability
- Archiving data
- Time issues
- Gmond scalability

Future work: *Migration Daemon*

- Makes predictions
- Trigger migration



Paper

- A Proactive Fault Tolerance Framework for High Performance Computing

  28th IASTED International Conference on Parallel and Distributed Computing and Networks (PDCN), Innsbruck, Austria, February 16-18, 2010.

Acknowledgement

The project was 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.

References

- Stephen L. Scott, Christian Engelmann, Geoffroy R. Vallée, Thomas Naughton, Anand Tikotekar, George Ostrouchov, Chokchai (Box) Leangsuksun, Nichamon Naksinehaboon, Raja Nassar, Mihaela Paun, Frank Mueller, Chao Wang, Arun B. Nagarajan, and Jyothish Varma. A Tunable Holistic Resiliency Approach for High-Performance Computing Systems. Poster at the 14th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP) 2009, Raleigh, NC, USA, February 14-18, 2009.
- Christian Engelmann, Geoffroy R. Vallée, Thomas Naughton, and Stephen L. Scott. Proactive Fault Tolerance Using Preemptive Migration. In *Proceedings of the 17th Euromicro International Conference on Parallel, Distributed, and network-based Processing (PDP) 2009*, pages 252-257, Weimar, Germany, February 18-20, 2009. IEEE Computer Society, Los Alamitos, CA, USA. ISBN 978-0-7695-3544-9. ISSN 1066-6192. URL: http://www.csm.ornl.gov/~engelman/publications/engelmann09proactive.pdf
- High Availability for High-End scientific computing, Master's thesis, Kai Uhlemann 2006

Summary and Questions

- HPC systems, failures, Fault Tolerance
- The RAS Framework for HPC systems
- The RAS Framework Engine Prototype
- Results: data stored, has some limitations
- Future work: predictions
- Questions?