A Resilient Runtime Environment for HPC and Internet Core Router Systems

Tim Mattox, Ph.D. (timattox@cisco.com)

Technical Leader, Engineering

SPRTG Projects

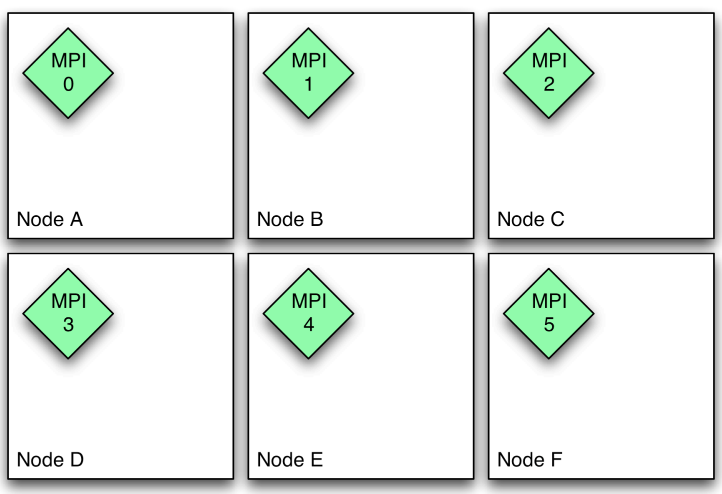
A Multiple Institution Project

- Cisco Team
  - Ralph Castain
  - **Timothy I. Mattox**
  - Robert M. Broberg
  - Jeffrey M. Squyres
- University Collaborators
  - Joshua Hursey, Indiana University
  - Chase Cotton, University of Delaware
  - Jonathan M. Smith, University of Pennsylvania
- Open MPI Project, http://www.open-mpi.org/

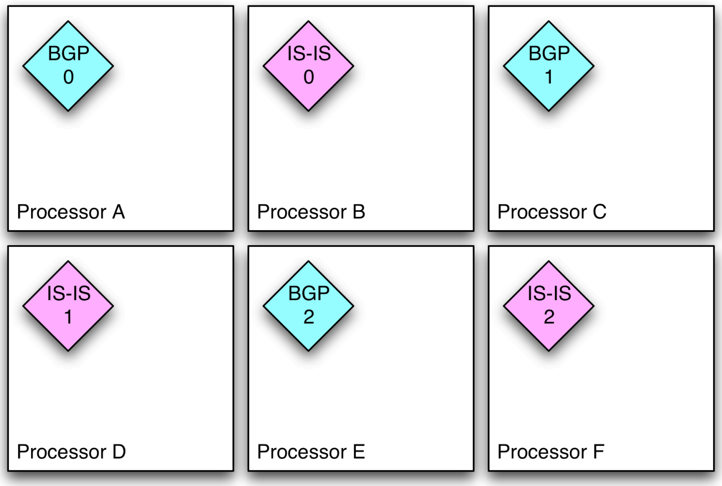
HPC and Internet Core Router Systems

- Highly parallel with various processor interconnects
- Trends that lower the whole system MTBF
  - Systems are growing in size and complexity
  - Increasing demands for new features
- Different fault tolerance needs
  - HPC Systems need long uptimes to effectively run large parallel applications
  - Internet Core Routers need non-stop operation to not disrupt services
    - IP Telephony
    - Video Conferencing

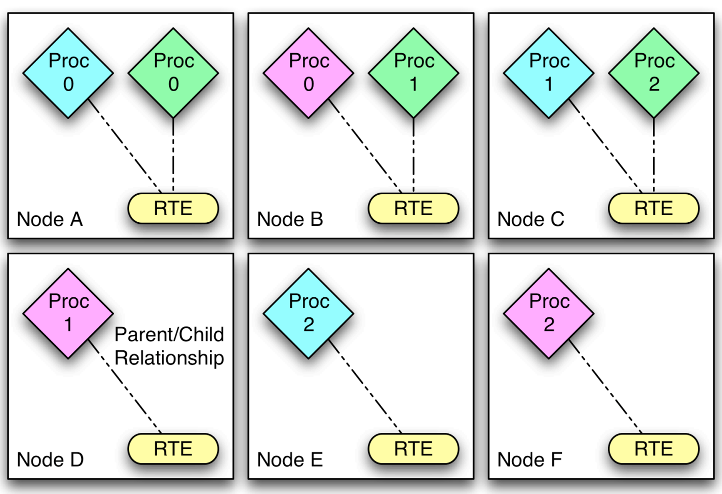
HPC System Architecture Slice



Internet Core Router Control Plane



Common Infrastructure



Open MPI's Runtime Environment (ORTE)

- Open Source (New BSD License)
  - 27 total Member, Partner, and Contributor organizations
- Modular Component Architecture (MCA)
  - Provides flexibility
  - Supports good software engineering practice

OPEN MPI

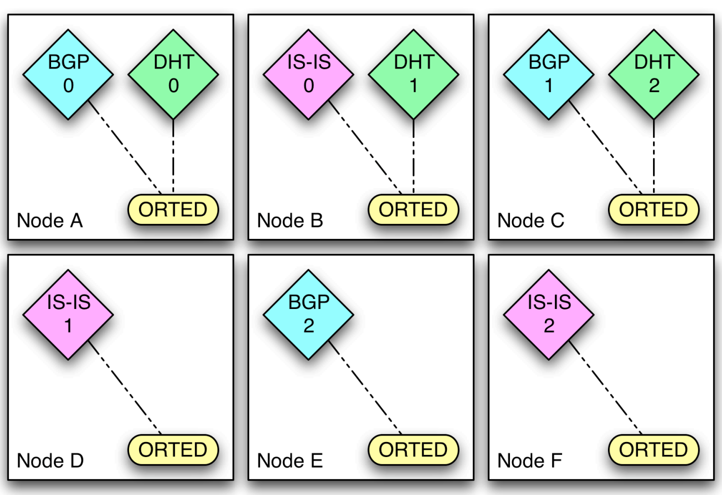
A Resilient Runtime Environment Needs

- Fault Detection
- Fault Recovery
- Fault Prediction
- Fault Group Model

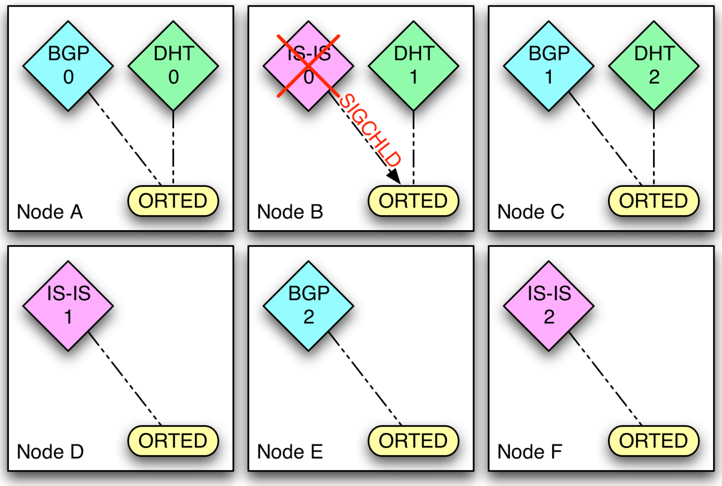
Our Additions/Enhancements to ORTE

- Sensor Framework
- Recovery Service (RecoS) Framework
- Resilient Mapper Component
- ClusterManager Routed Component

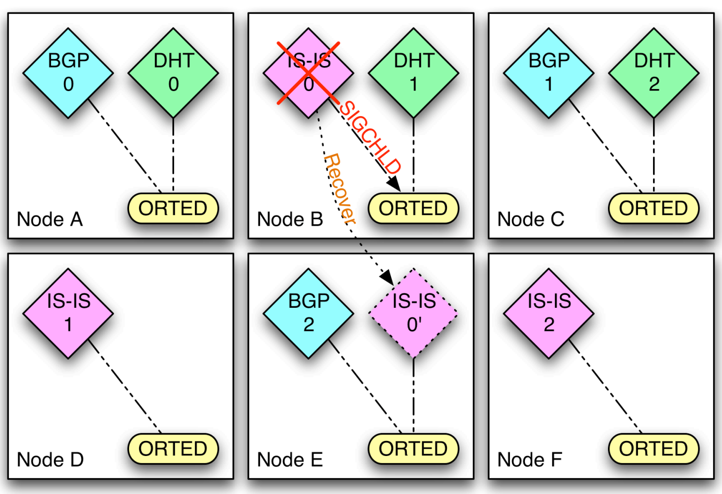
Example Fault Detection



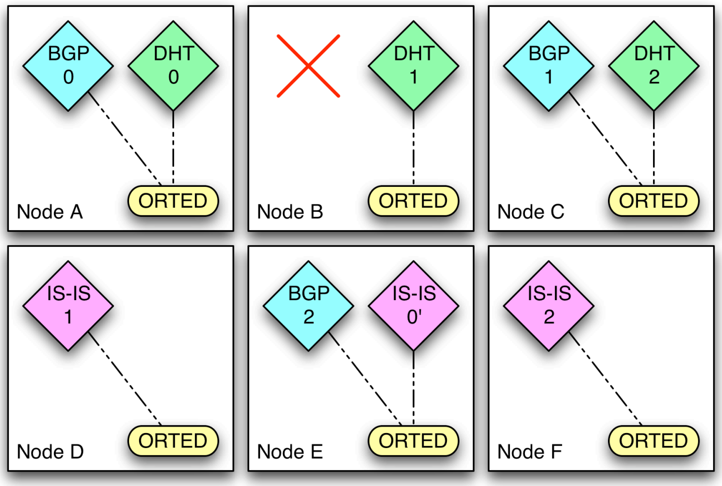
Example Fault Detection



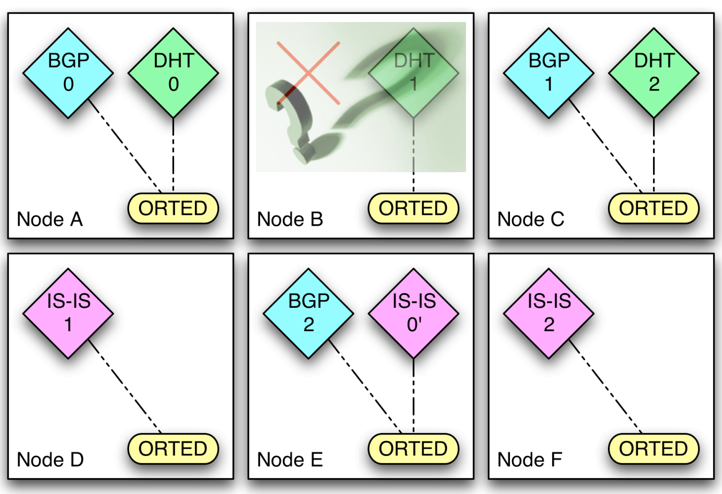
Example Fault Recovery



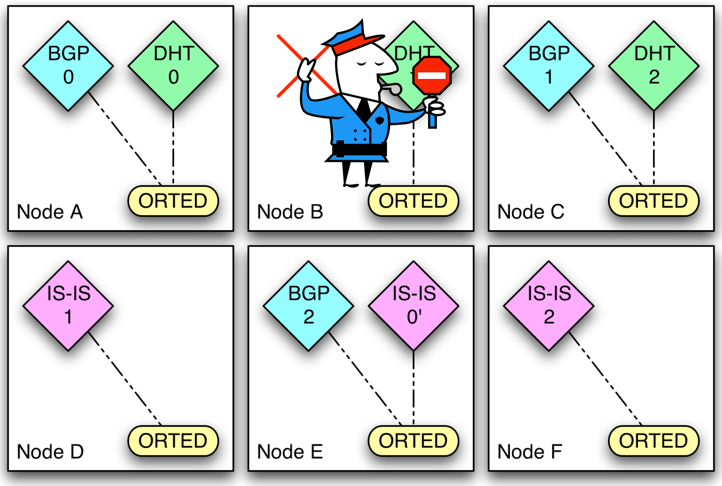
Example Fault Recovery



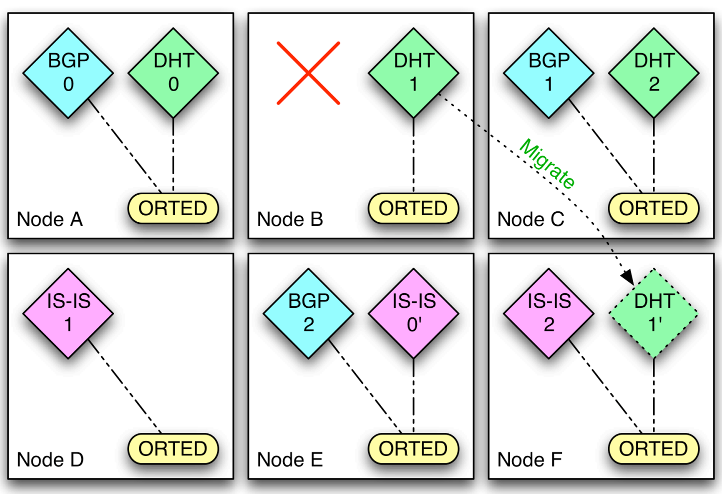
Example Fault Prediction



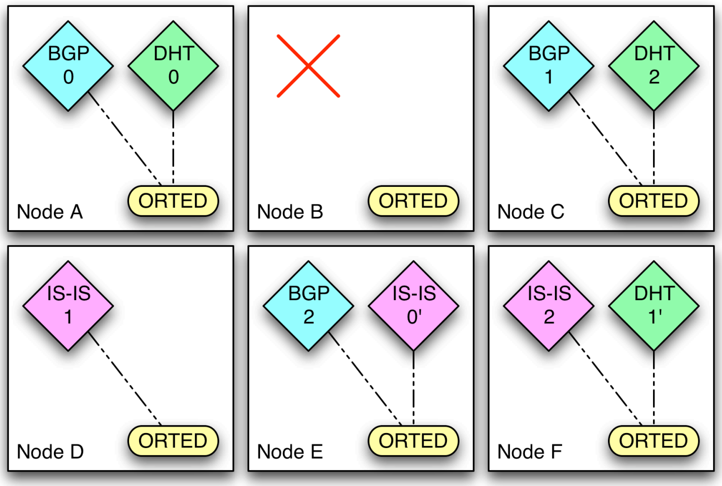
Example Fault Prediction



Example Fault Prediction



Example Fault Prediction



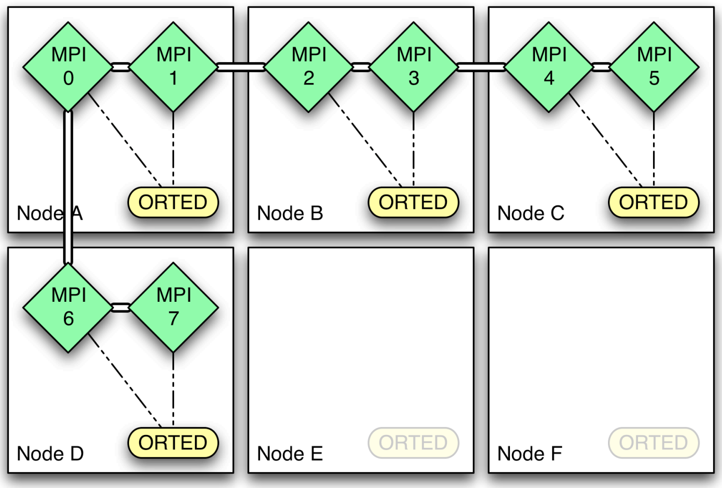
17

Preliminary Results

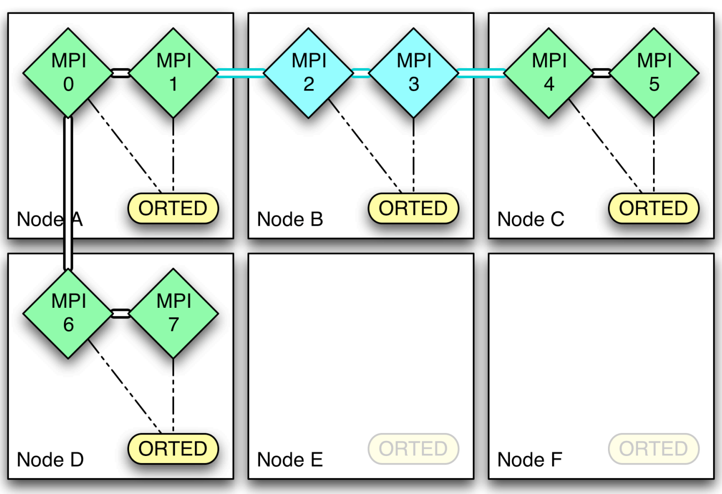
- Non-MPI process restart in ~6 milliseconds
  - Local shell script takes ~3 milliseconds to start a process
  - Remote shell script takes ~80 milliseconds via ssh
- MPI process migration vs. checkpoint/restart
  - 128 process LAMMPS metallic solid benchmark
  - 6 GB of state distributed on 32 nodes
  - Factor of five reduction in overhead migrating 4 processes vs. checkpoint/restart

18

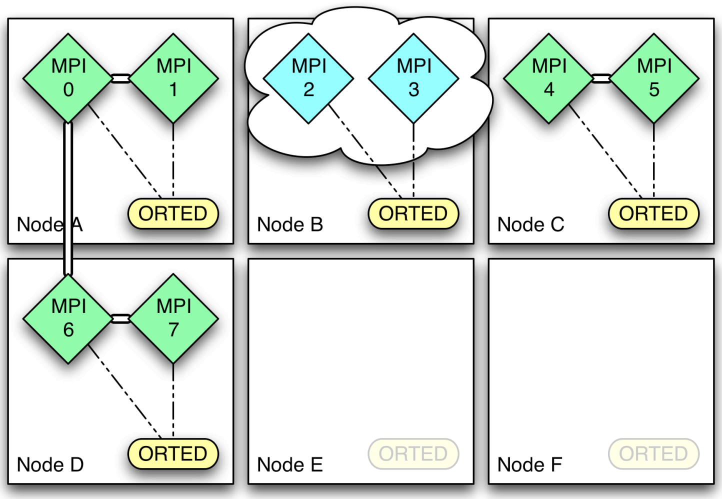
Example MPI Process Migration



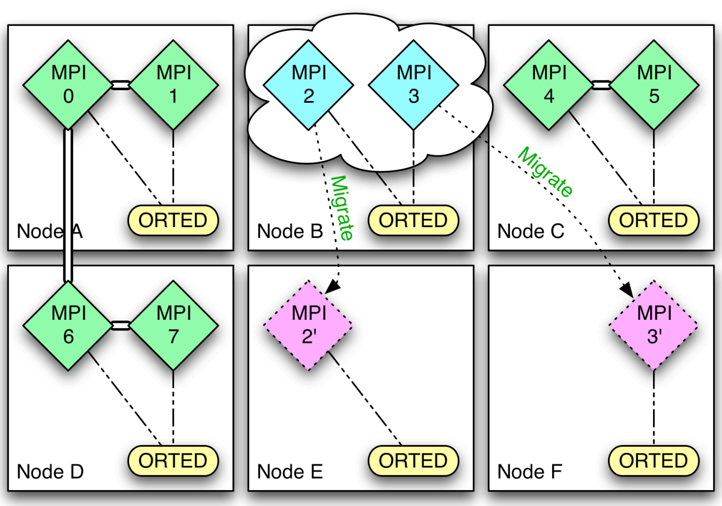
Example MPI Process Migration



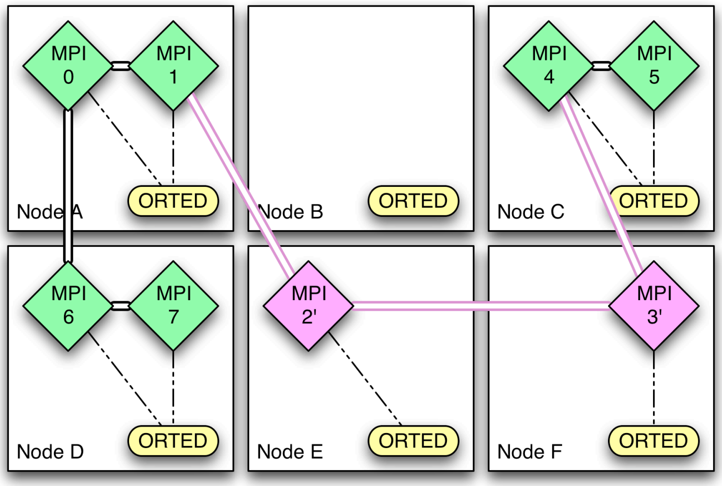
Example MPI Process Migration



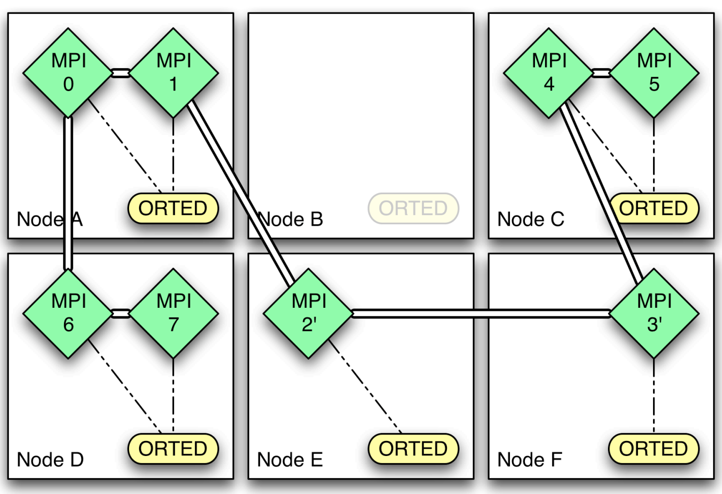
Example MPI Process Migration



Example MPI Process Migration



Example MPI Process Migration



Some Planned Future Extensions

- More sensor components
- More and better fault prediction algorithms
- More fault detection techniques
- Interface with more external fault notification systems

Conclusions

The overlap of goals for HPC and Internet Core Router System resiliency has resulted in a synergistic advancement in the Open MPI Runtime Environment software.

For more information:

See our poster (#47) in the Oregon Ballroom Lobby

Visit the Reliable Router Research (R3) website

http://r3.cis.upenn.edu/