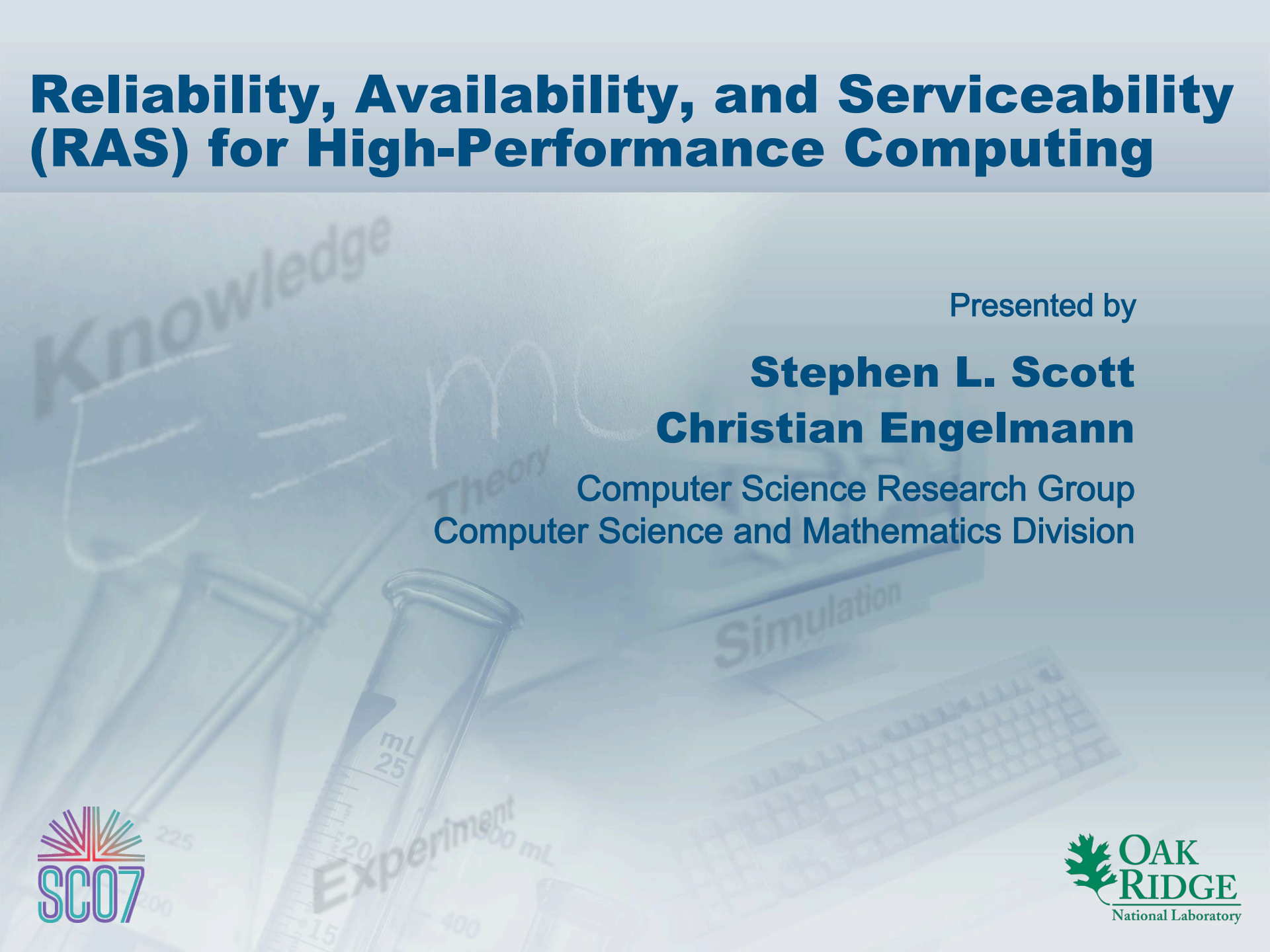
Reliability, Availability, and Serviceability (RAS) for High-Performance Computing

Presented by

Stephen L. Scott

Christian Engelmann

Computer Science Research Group
Computer Science and Mathematics Division



SC07

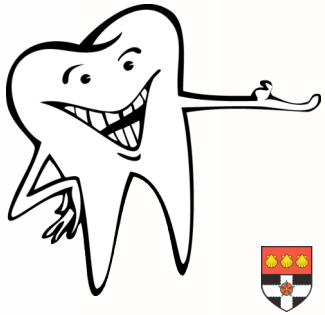
OAK RIDGE National Laboratory

Research and development goals

- Provide high-level RAS capabilities for current terascale and next-generation petascale high-performance computing (HPC) systems
- Eliminate many of the numerous single points of failure and control in today's HPC systems
  - Develop techniques to enable HPC systems to run computational jobs 24/7
  - Develop proof-of-concept prototypes and production-type RAS solutions

MOLAR: Adaptive runtime support for high-end computing operating and runtime systems

- Addresses the challenges for operating and runtime systems to run large applications efficiently on future ultrascale high-end computers
- Part of the *Forum to Address Scalable Technology for Runtime and Operating Systems (FAST-OS)*
- MOLAR is a collaborative research effort (www.fastos.org/molar)



OAK RIDGE National Laboratory

Office of Science U.S. DEPARTMENT OF ENERGY

THE OHIO STATE UNIVERSITY

NC STATE UNIVERSITY

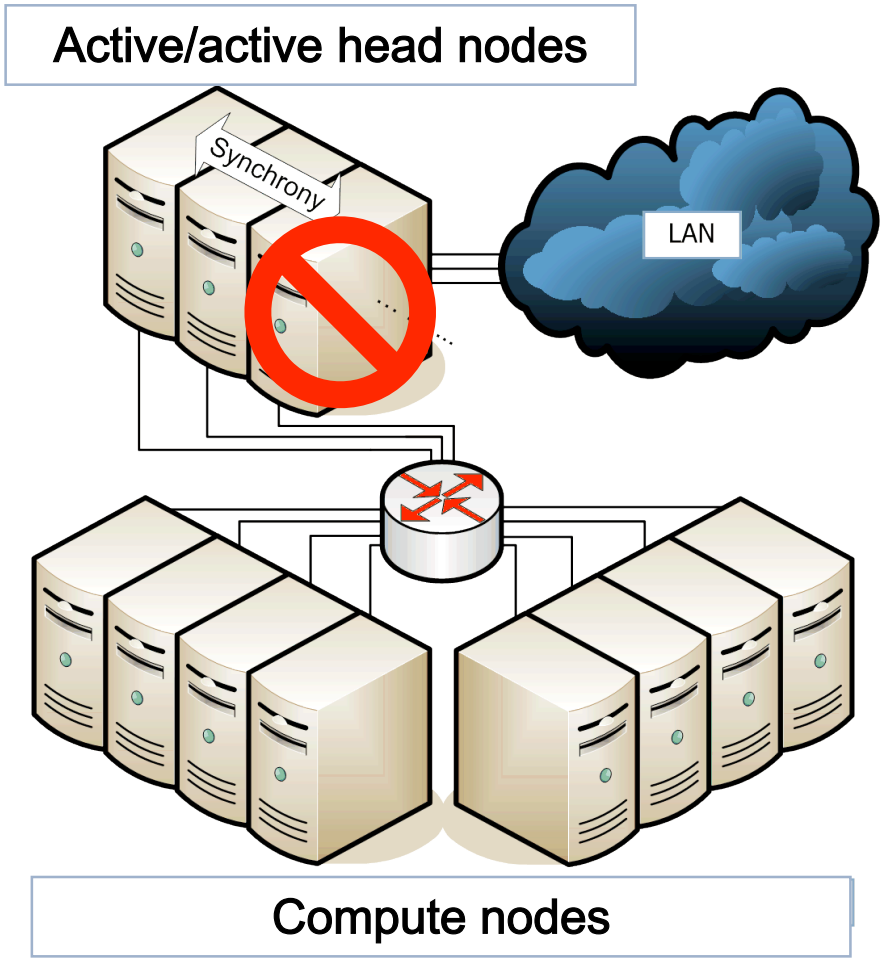
The University of Reading

CRAY

LOUISIANA TECH UNIVERSITY

OAK RIDGE National Laboratory

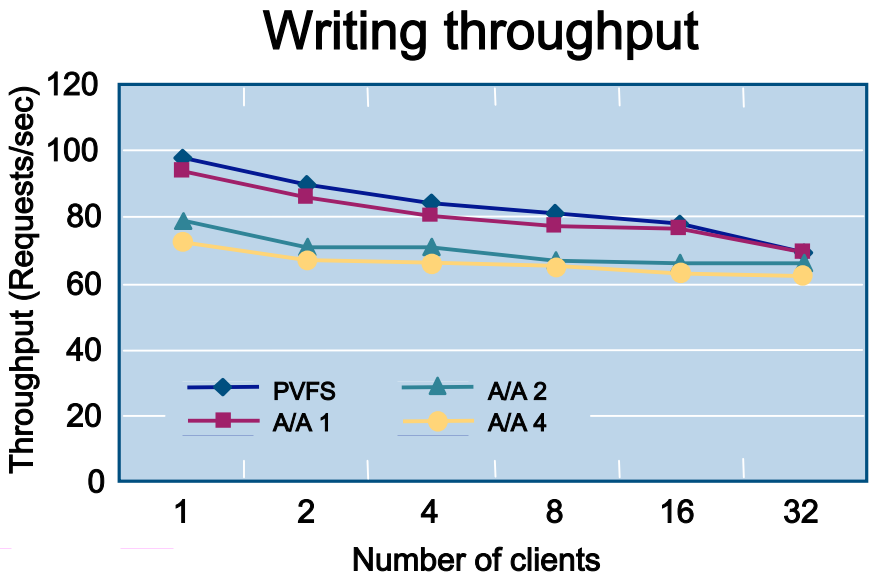
Symmetric active/active redundancy

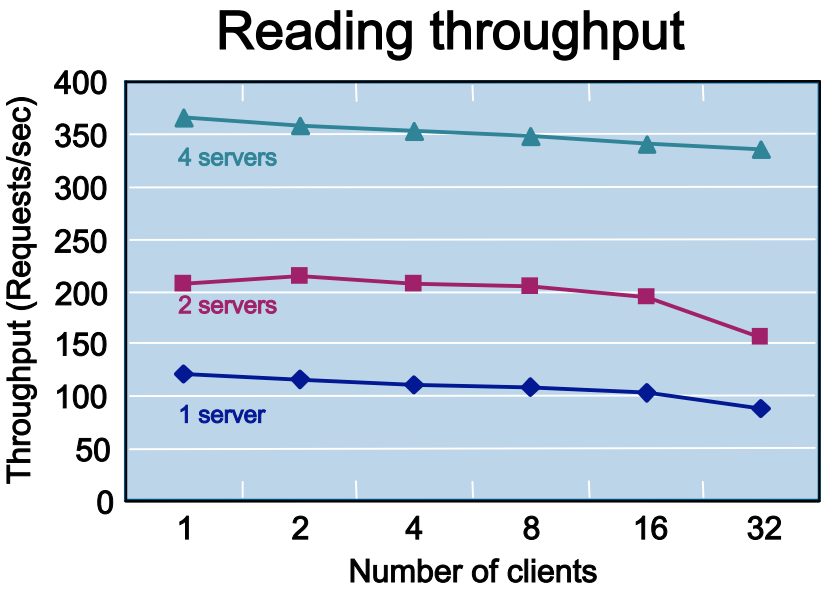


- Many active head nodes
- Workload distribution
- Symmetric replication between head nodes
- Continuous service
- Always up to date
- No fail-over necessary
- No restore-over necessary
- Virtual synchrony model
- Complex algorithms

➡ **Prototypes for Torque and Parallel Virtual File System metadata server**

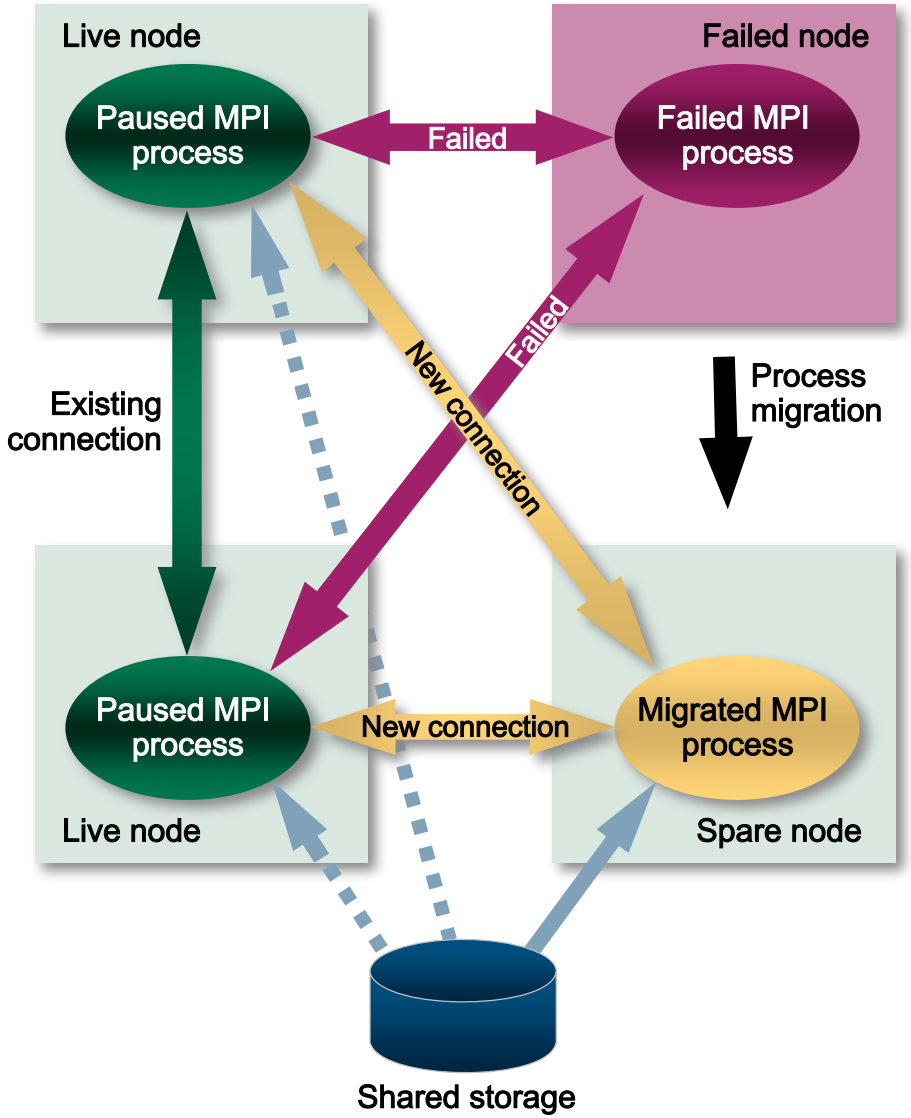
Symmetric active/active Parallel Virtual File System metadata server





Nodes	Availability	Est. annual downtime
1	98.58%	5d, 4h, 21m
2	99.97%	1h, 45m
3	99.9997%	1m, 30s

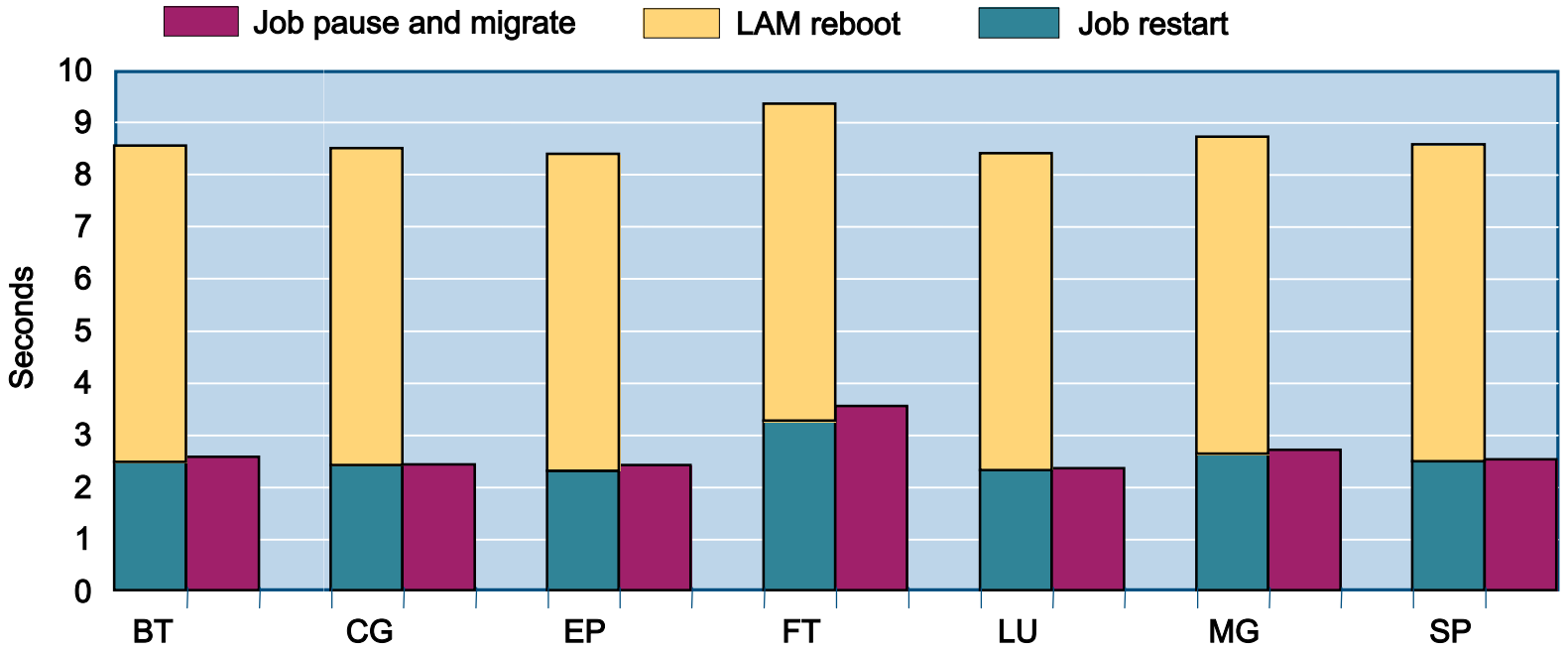
Reactive fault tolerance for HPC with LAM/MPI+BLCR job-pause mechanism



- Operational nodes: Pause
  - BLCR reuses existing processes
  - LAM/MPI reuses existing connections
  - Restore partial process state from checkpoint
- Failed nodes: Migrate
  - Restart process on new node from checkpoint
  - Reconnect with paused processes
- Scalable MPI membership management for low overhead

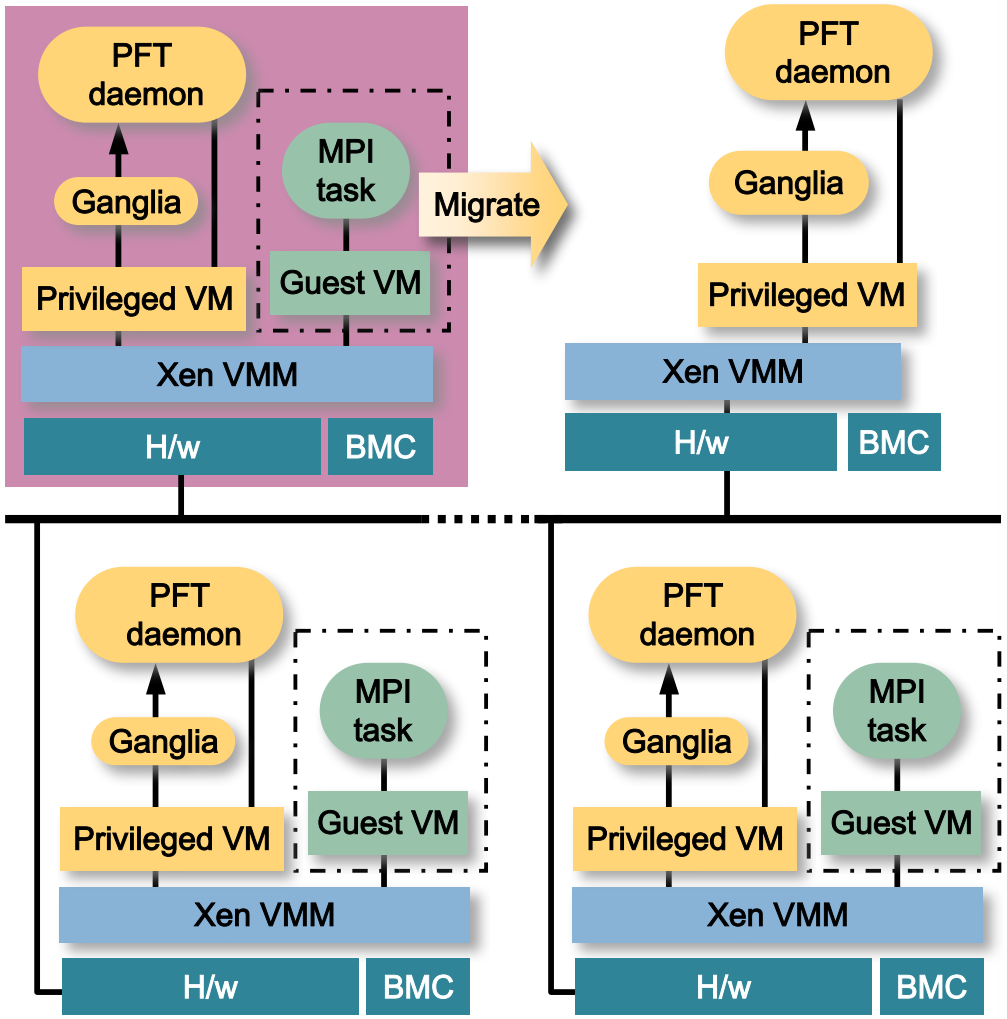
➡ **Efficient, transparent, and automatic failure recovery**

LAM/MPI+BLCR job pause performance



- 3.4% overhead over job restart, but
  - No LAM reboot overhead
  - Transparent continuation of execution
- No requeue penalty
- Less staging overhead

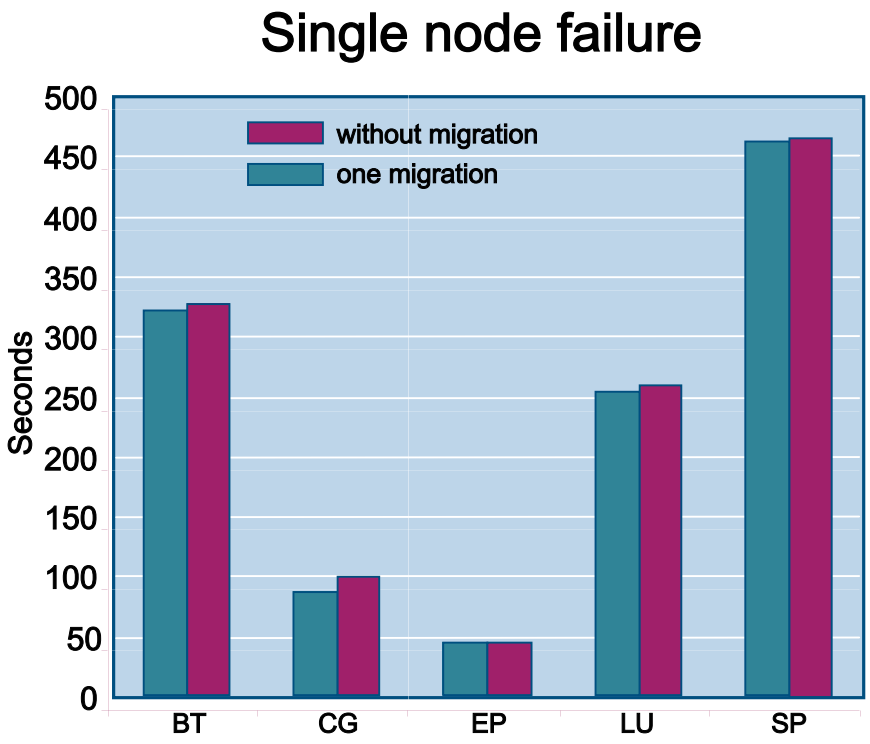
Proactive fault tolerance for HPC using Xen virtualization

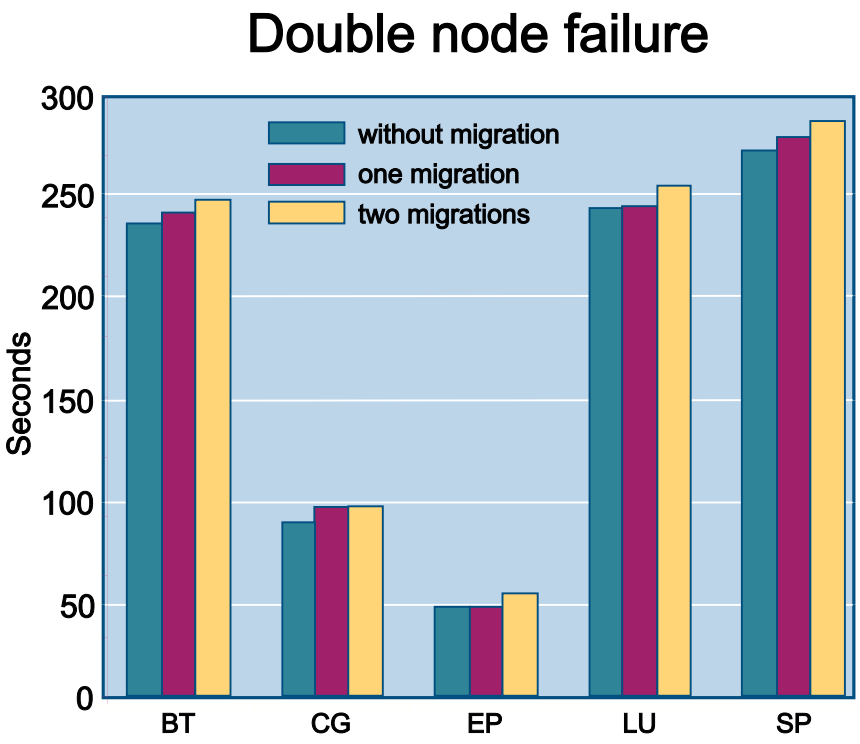


- Standby Xen host (spare node without guest VM)
- Deteriorating health
  - Migrate guest VM to spare node
- New host generates unsolicited ARP reply
  - Indicates that guest VM has moved
  - ARP tells peers to resend to new host

➡ **Novel fault-tolerance scheme that acts before a failure impacts a system**

VM migration performance impact

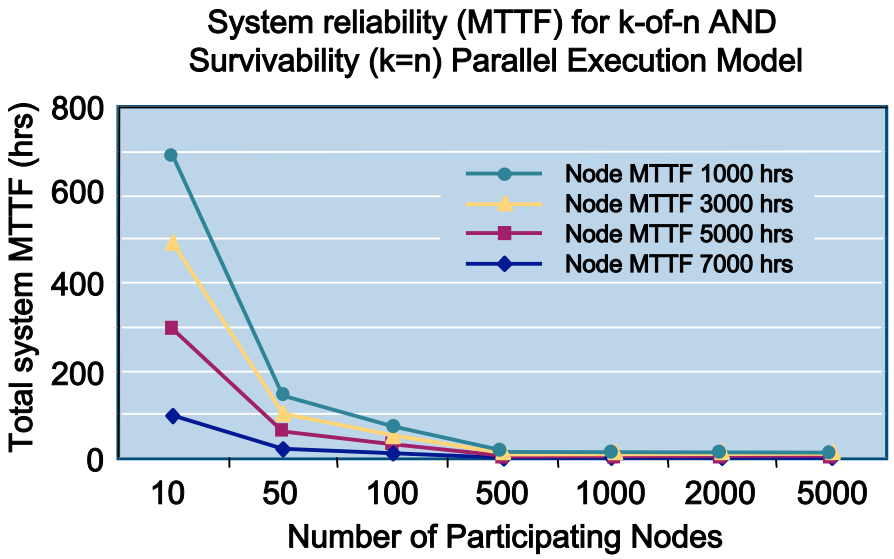


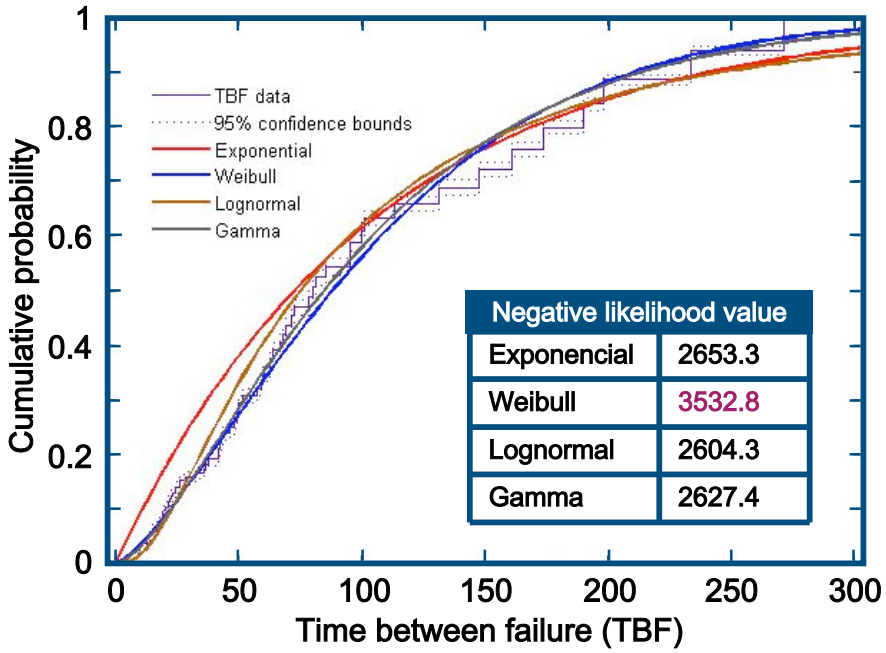


- Single node failure: 0.5–5% additional cost over total wall clock time
- Double node failure: 2–8% additional cost over total wall clock time

HPC reliability analysis and modeling

- Programming paradigm and system scale impact reliability
- Reliability analysis
- Estimate mean time to failure (MTTF)
- Obtain failure distribution: exponential, Weibull, gamma, etc.
- Feedback into fault-tolerance schemes for adaptation





Contacts regarding RAS research

Stephen L. Scott

Computer Science Research Group
Computer Science and Mathematics Division
(865) 574-3144
scottsl@ornl.ornl

Christian Engelmann

Computer Science Research Group
Computer Science and Mathematics Division
(865) 574-3132
engelmannc@ornl.ornl

OAK RIDGE National Laboratory