

Towards High Availability for High-Performance Computing System Services: Accomplishments and Limitations



Christian Engelmann^{1,2}, Stephen L. Scott¹,
Chokchai (Box) Leangsuksun³, Xubin (Ben) He⁴

¹ Oak Ridge National Laboratory, Oak Ridge, USA

² The University of Reading, Reading, UK

³ Louisiana Tech University, Ruston, USA

⁴ Tennessee Tech University, Cookeville, USA

Talk Outline

- Scientific high-end computing (HEC)
- Availability deficiencies of today's HEC systems
- Projects and accomplishments overviews
- High availability models for HPC system services
- Developed prototypes overview
- Existing limitations and most pressing issues

Scientific High-End Computing (HEC)

- Large-scale HPC systems.
 - Tens-to-hundreds of thousands of processors.
 - Current systems: IBM Blue Gene/L and Cray XT3
 - Next-generation systems: IBM Blue Gene/P and Cray XT4
- Computationally and data intensive applications.
 - 10 TFLOP – 1 PFLOP with 10 TB – 1 PB of data.
 - Climate change, nuclear astrophysics, fusion energy, materials sciences, biology, nanotechnology, ...
- Capability vs. capacity computing
 - Single jobs occupy large-scale high-performance computing systems for weeks and months at a time.

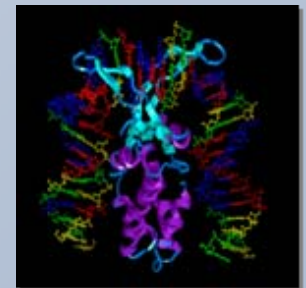
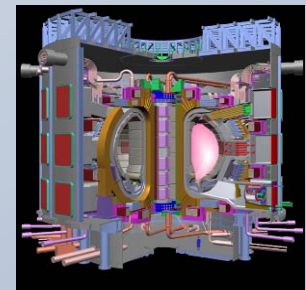
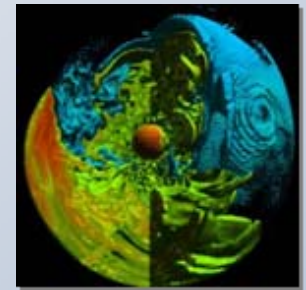
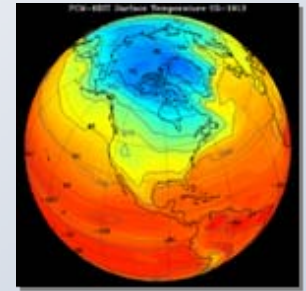
National Center for Computational Sciences

- **40,000 ft² (3700 m²) computer center:**
 - 36-in (~1m) raised floor, 18 ft (5.5 m) deck-to-deck
 - 12 MW of power with 4,800 t of redundant cooling
 - High-ceiling area for visualization lab:
 - 35 MPixel PowerWall, Access Grid, etc.
- **3 systems in the Top 500 List of Supercomputer Sites:**
 - **Jaguar:** 10. Cray XT3, MPP with 5212 Procs./10 TByte \Rightarrow 25 TFlop/s.
 - **Phoenix:** 17. Cray X1E, Vector with 1024 Procs./ 4 TByte \Rightarrow 18 TFlop/s.
 - **Cheetah:** 283. IBM Power 4, Cluster with 864 Procs./ 1 TByte \Rightarrow 4.5 TFlop/s.
 - **Ram:** SGI Altix, SSI with 256 Procs./ 2 TByte \Rightarrow 1.4 TFlop/s.

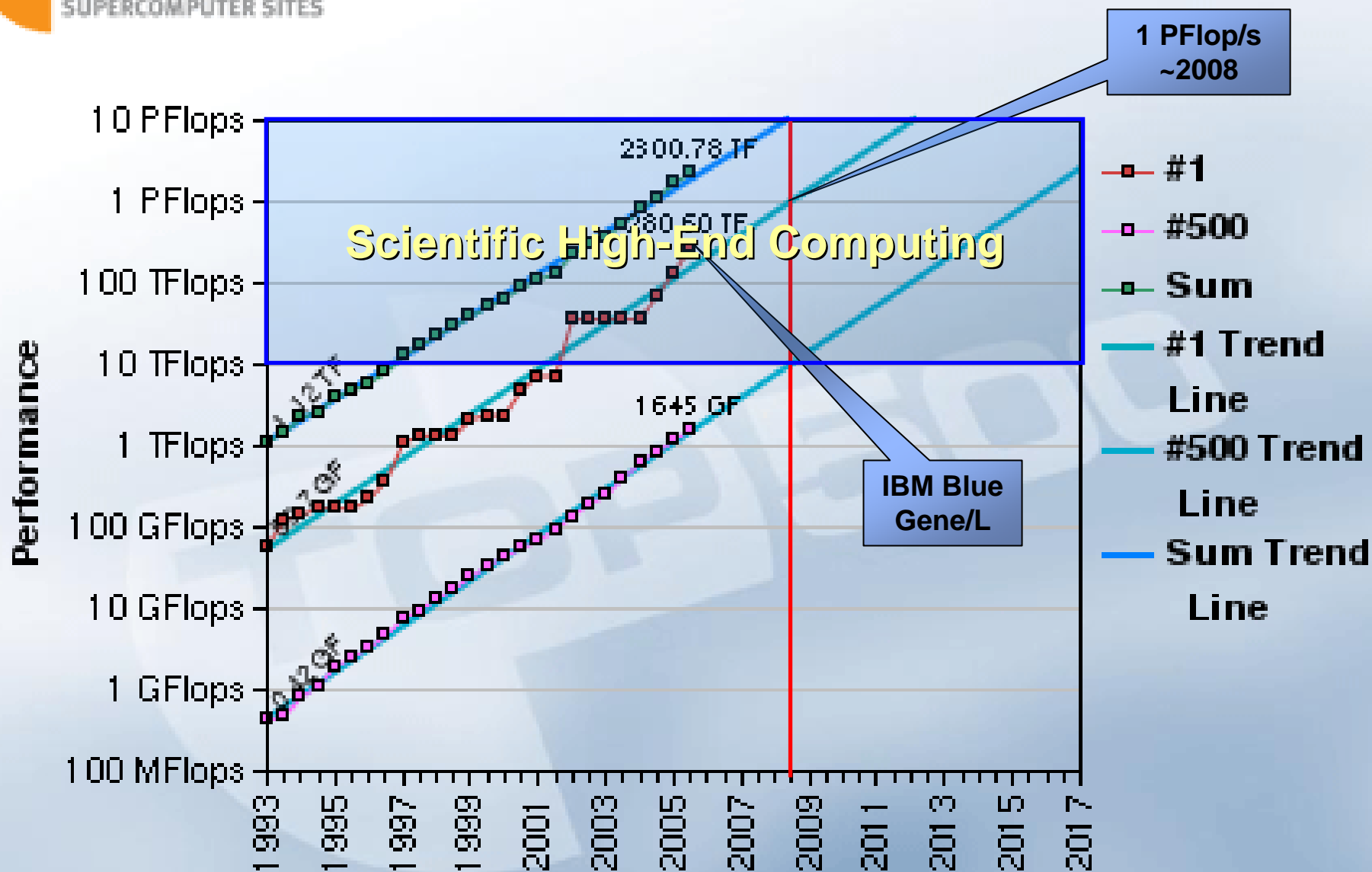


NCCS: At Forefront in Scientific Computing and Simulation

- Leading partnership in developing the National Leadership Computing Facility
 - Leadership-class scientific computing capability
 - 54 TFlop/s in 2006 (recent upgrade)
 - 100 TFlop/s in 2006 (commitment made)
 - 250 TFlop/s in 2007 (commitment made)
 - 1 PFlop/s in 2008 (proposed)
- Attacking key computational challenges
 - Climate change
 - Nuclear astrophysics
 - Fusion energy
 - Materials sciences
 - Biology



Projected Performance Development

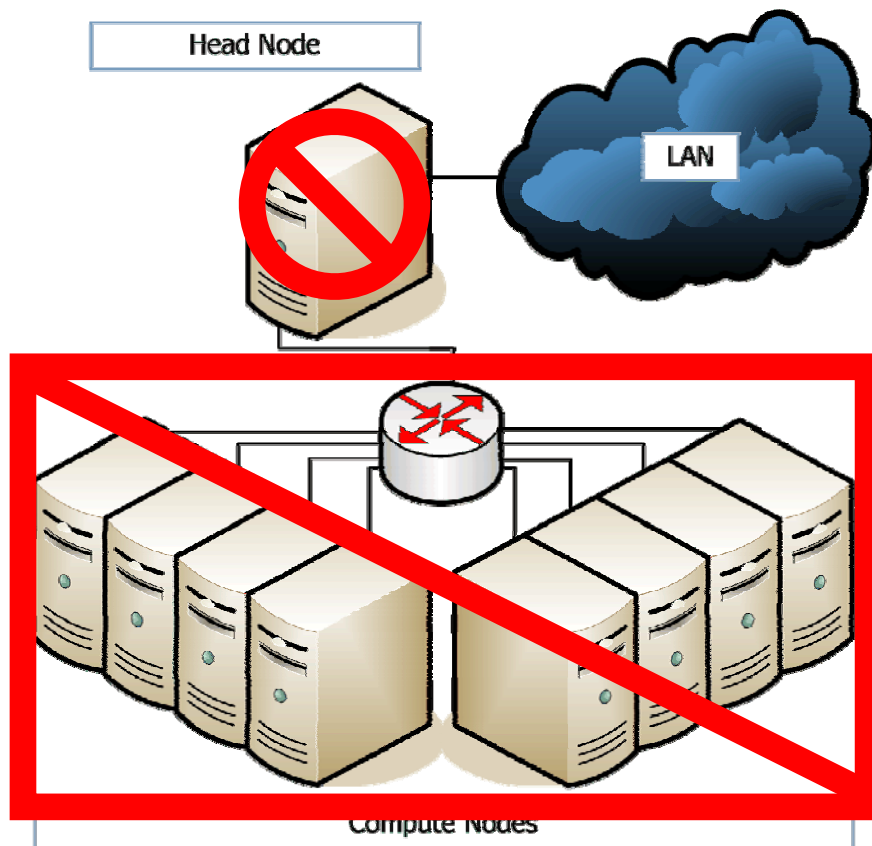


Availability Measured by the Nines

9's	Availability	Downtime/Year	Examples
1	90.0%	36 days, 12 hours	Personal Computers
2	99.0%	87 hours, 36 min	Entry Level Business
3	99.9%	8 hours, 45.6 min	ISPs, Mainstream Business
4	99.99%	52 min, 33.6 sec	Data Centers
5	99.999%	5 min, 15.4 sec	Banking, Medical
6	99.9999%	31.5 seconds	Military Defense

- Enterprise-class hardware + Stable Linux kernel = 5+
- Substandard hardware + Good high availability package = 2-3
- Today's supercomputers = 1-2
- My desktop = 1-2

Single Head/Service Node Problem



- Single point of failure.
- Compute nodes sit idle while head node is down.
- $A = \text{MTTF} / (\text{MTTF} + \text{MTTR})$
- MTTF depends on head node hardware/software quality.
- MTTR depends on the time it takes to repair/replace node.
- $\text{MTTR} = 0 \rightarrow A = 1.00$ (100%) **continuous availability.**

Projects Overview

- Initial **HA-OSCAR** research in active/standby technology for the batch job management system
- Ongoing **MOLAR** research in active/standby, asymmetric and symmetric active/active technology
- Recent **RAS LDRD** research in symmetric active/active technology
- ➔ 3-4 years of research and development in high availability for high-performance computing system services

Accomplishments Overview

- Investigated the overall background of HA technologies in the context of HPC
 - Detailed problem description
 - Conceptual models
 - Review of existing solutions
- Developed different replication strategies for providing high availability for HPC system services
 - Active/standby
 - Asymmetric active/active
 - Symmetric active/active
- Implemented several proof-of-concept prototypes

High Availability Models

- Active/Standby (Warm or Hot)
 - For one active component at least one redundant inactive (standby) component
 - Fail-over model with idle standby component(s)
 - Level of high-availability depends on replication strategy
- Active/Active (Asymmetric or Symmetric)
 - Multiple redundant active components
 - No wasted system resources
 - State change requests can be accepted and may be executed by every member of the component group

Active/Standby with Shared Storage

Active/Standby Head Nodes with Shared Storage



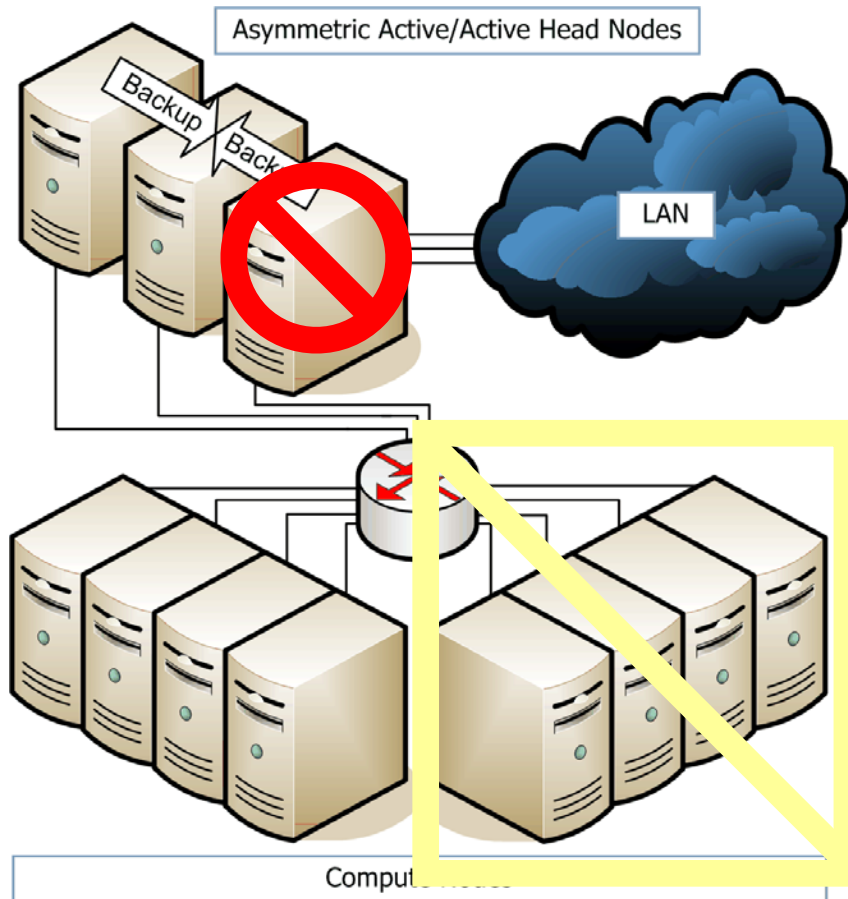
- Single active head node
 - Backup to shared storage
 - Simple checkpoint/restart
 - Fail-over to standby node
 - Possible corruption of backup state when failing during backup
 - Introduction of a new single point of failure
 - Correctness and availability are NOT guaranteed
- ➔ SLURM, meta data servers of PVFS and Lustre

Active/Standby Redundancy



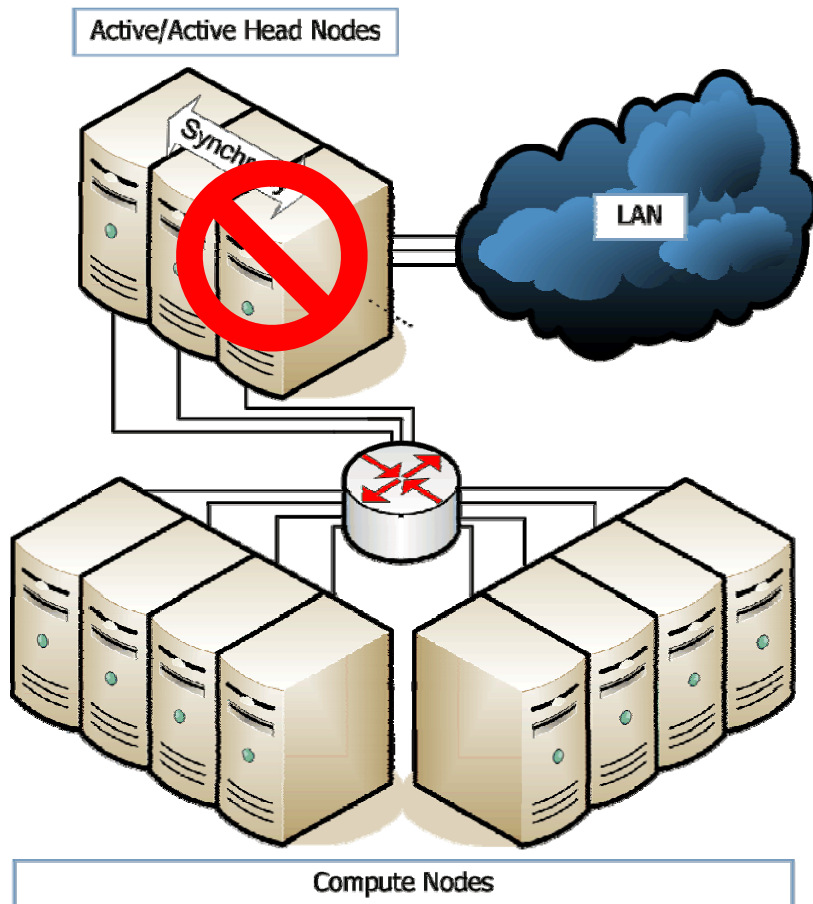
- Single active head node
- Backup to standby node
- Simple checkpoint/restart
- Fail-over to standby node
- Idle standby head node
- Rollback to backup
- Service interruption for fail-over and restore-over
- ➔ Torque on Cray XT
- ➔ HA-OSCAR prototype

Asymmetric Active/Active Redundancy



- Many active head nodes
 - Work load distribution
 - Optional fail-over to standby head node(s) ($n+1$ or $n+m$)
 - No coordination between active head nodes
 - Service interruption for fail-over and restore-over
 - Loss of state w/o standby
 - Limited use cases, such as high-throughput computing
- ➔ Prototype based on HA-OSCAR

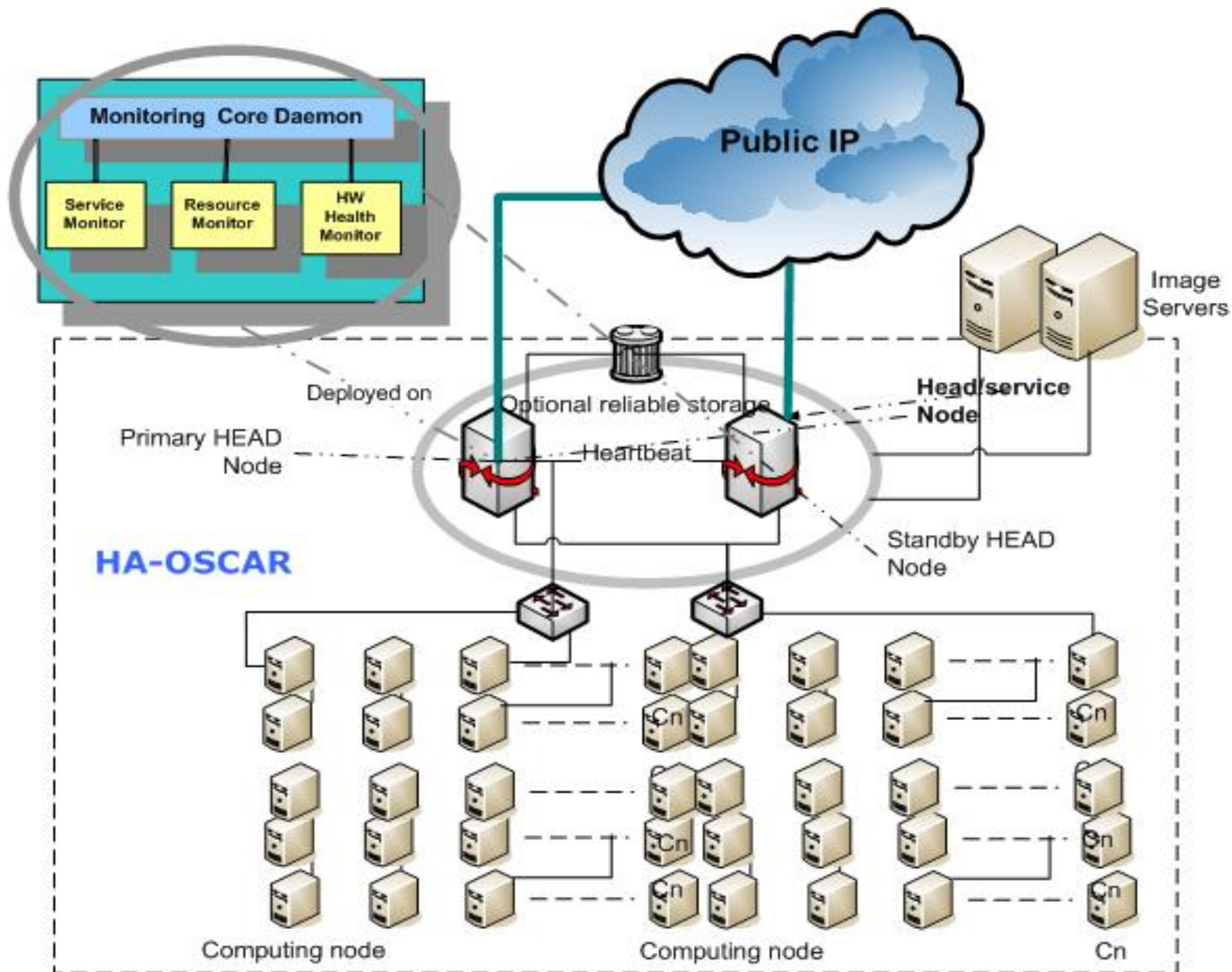
Symmetric Active/Active Redundancy

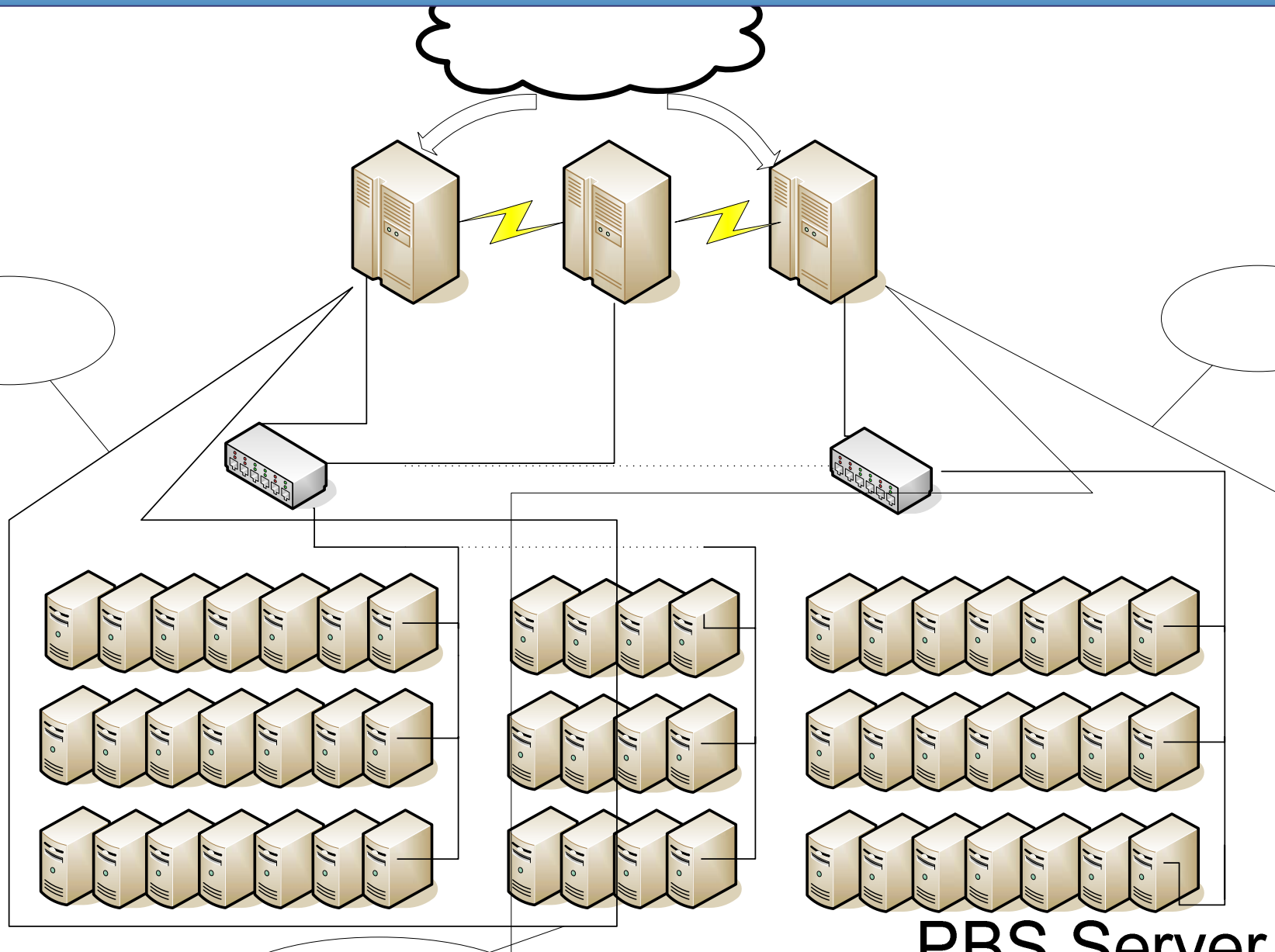


- Many active head nodes
- Work load distribution
- Symmetric replication between head nodes
- Continuous service
- Always up-to-date
- No fail-over necessary
- No restore-over necessary
- Virtual synchrony model
- **Complex algorithms**
- JOSHUA prototype for Torque

Developed Prototypes Overview

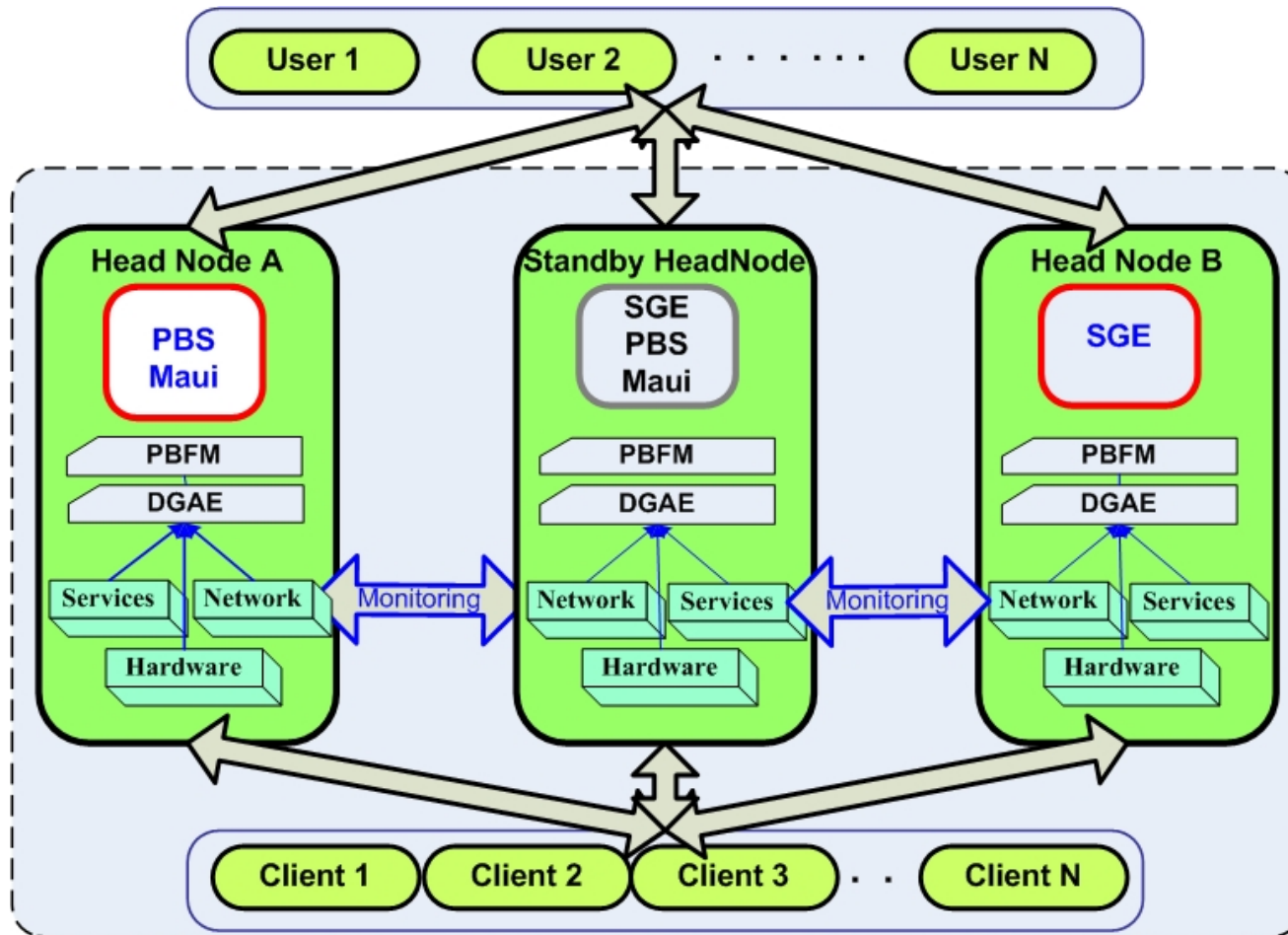
- Active/Standby HA-OSCAR
 - High availability for Open PBS/TORQUE
 - Integration with compute node checkpoint/restart
- Asymmetric active/active HA-OSCAR
 - High availability for Open PBS & SGE
 - High throughput computing solution
- Symmetric active/active JOSHUA
 - High availability for PBS TORQUE
 - Fully transparent replication



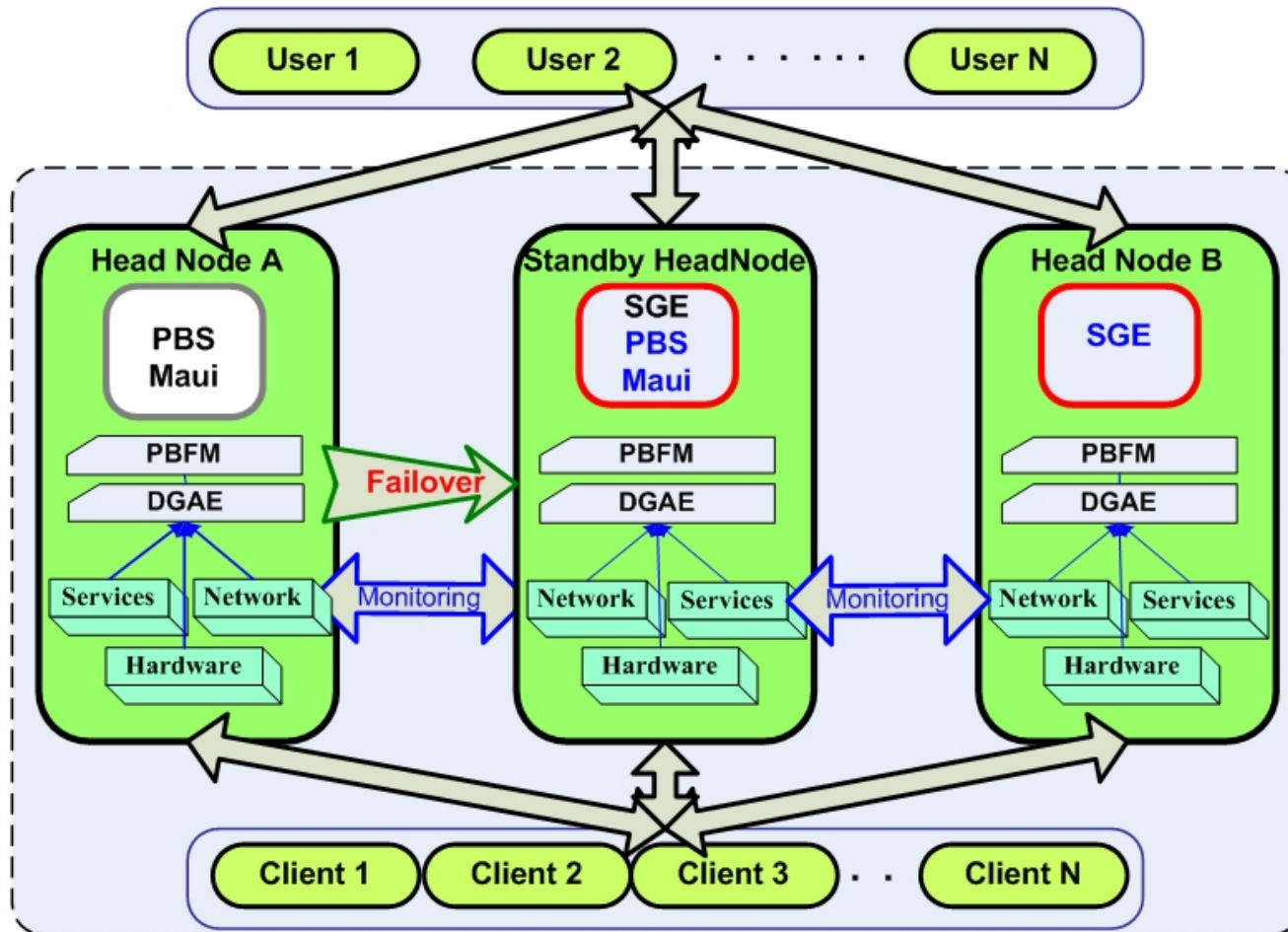


PBS Server
Submitting PBS jobs

Normal Active-Active Operation

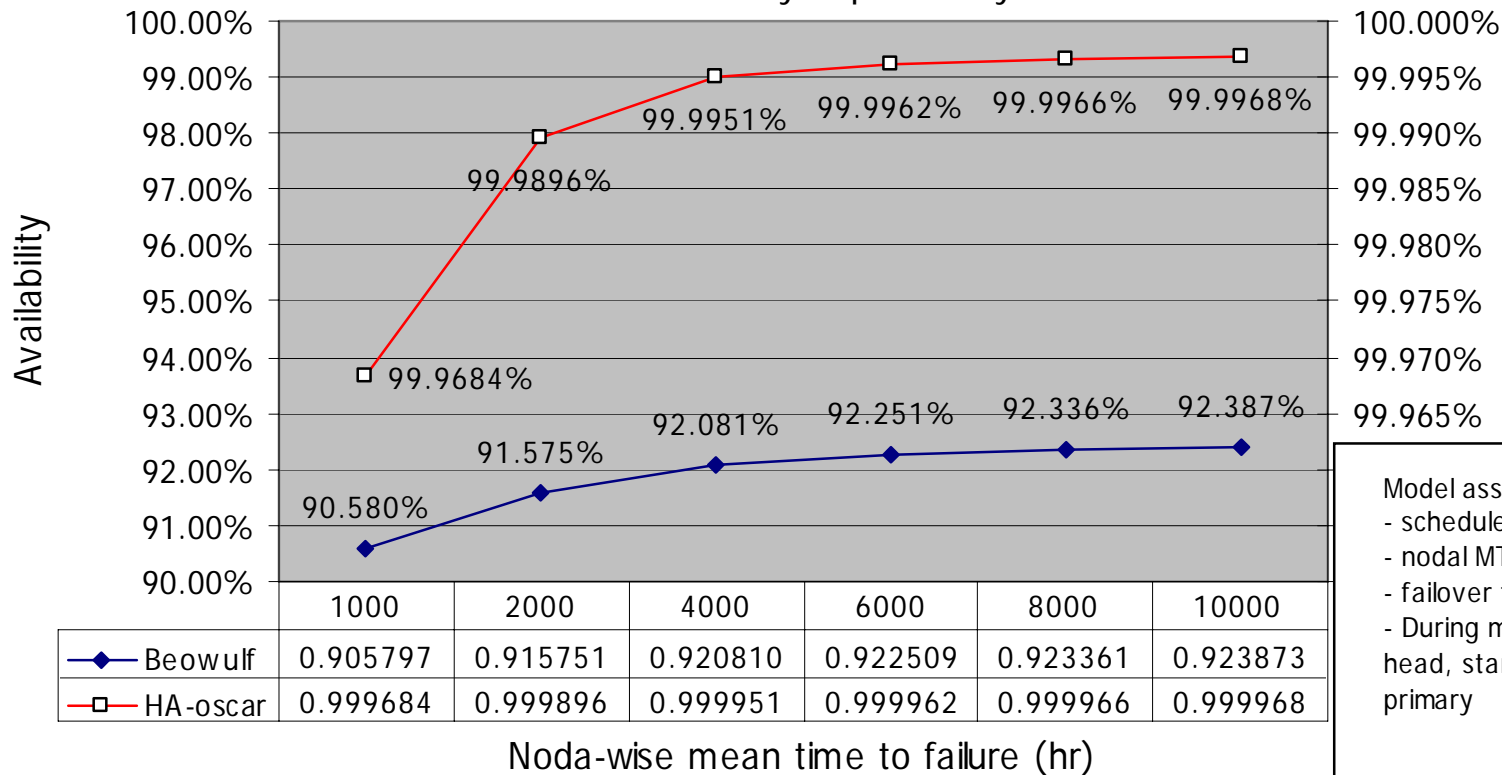


Failover Active-Active Operation



Asymmetric Active/Active Availability

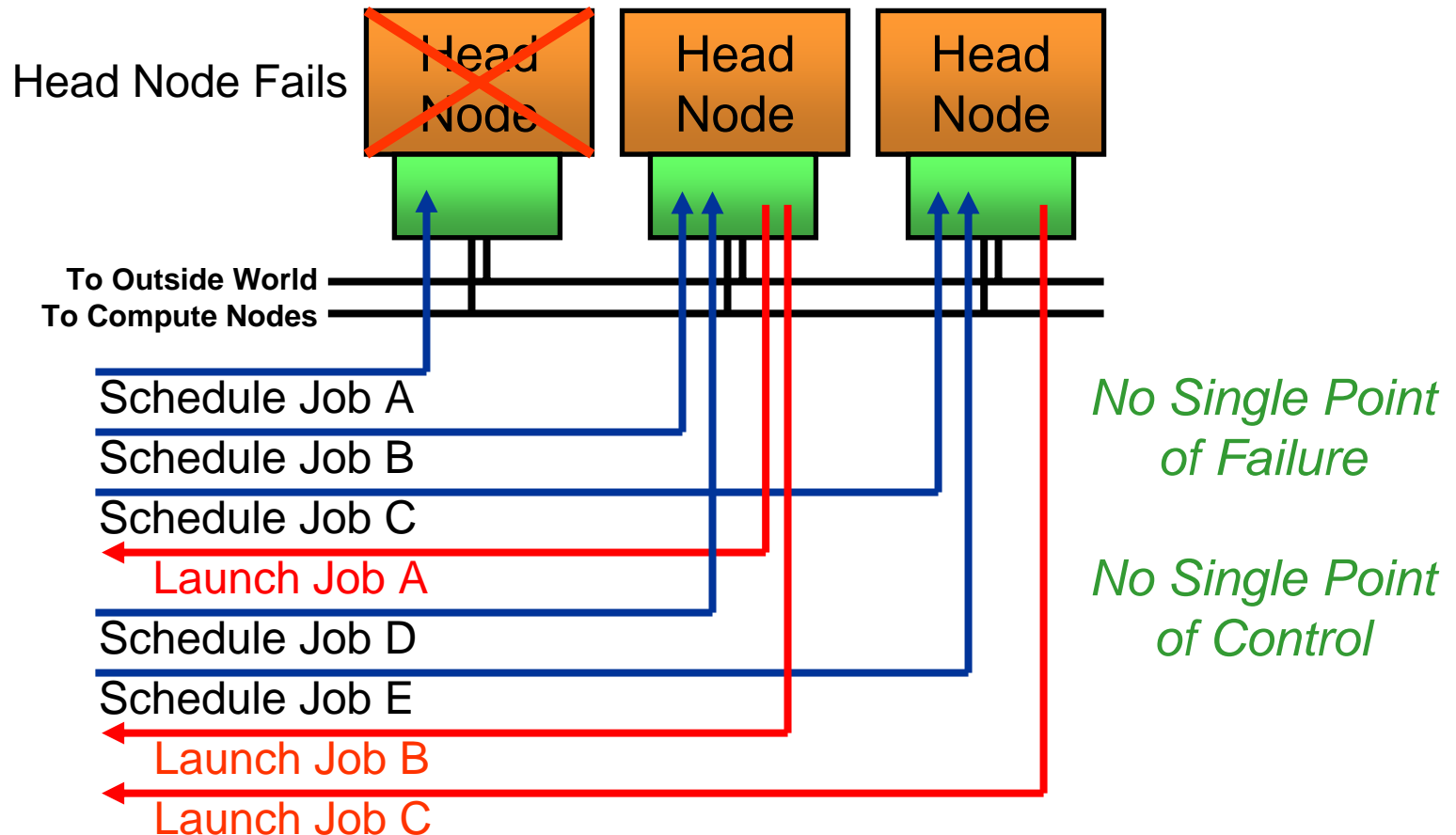
HA-OSCAR solution vs traditional Beowulf
Total Availability impacted by service nodes



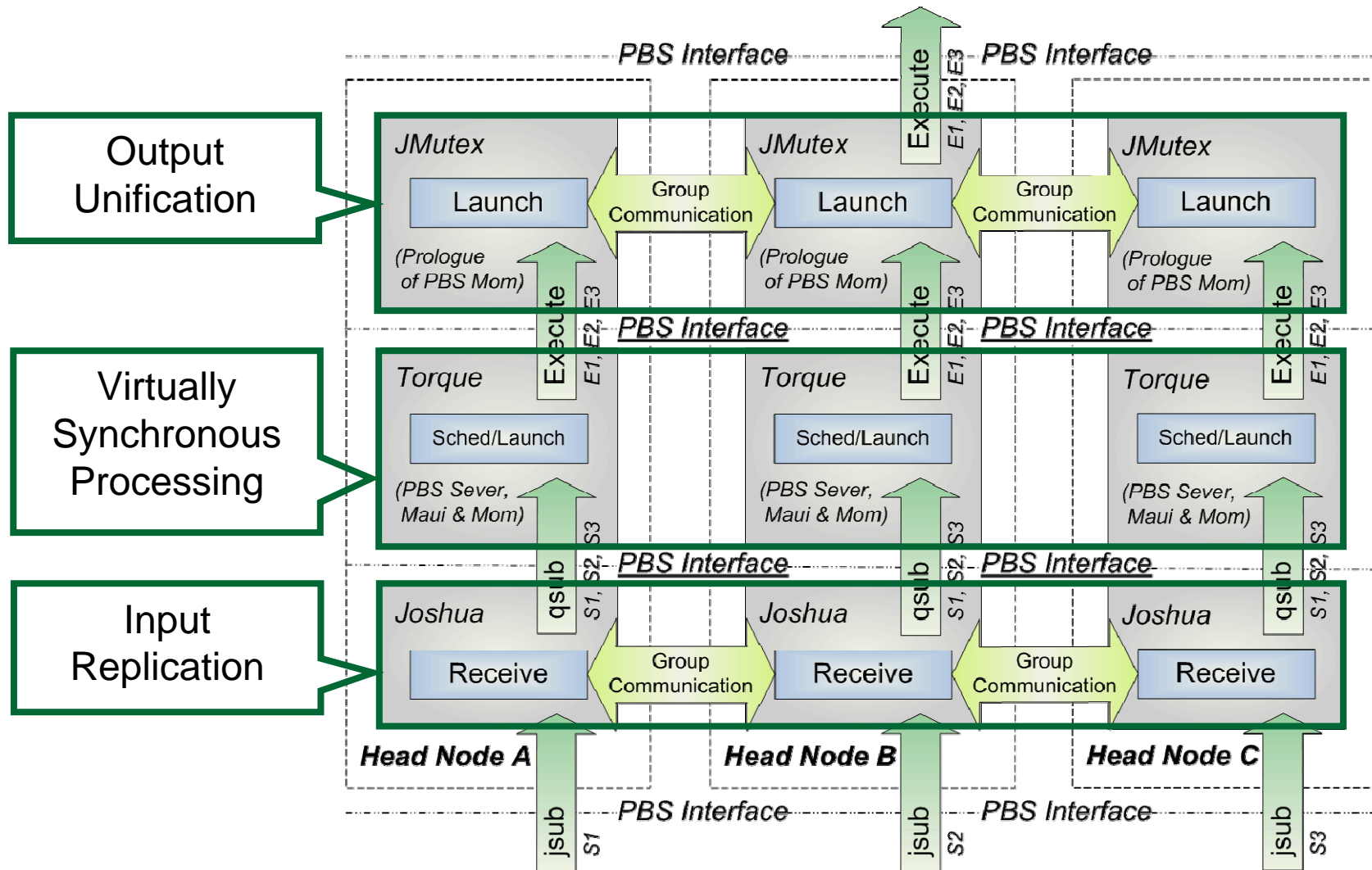
Model assumption:

- scheduled downtime=200 hrs
- nodal MTTR = 24 hrs
- failover time=10s
- During maintenance on the head, standby node acts as primary

JOSHUA: Symmetric Active/Active Replication for PBS Torque



Symmetric Active/Active Replication



Introduced Overhead

- Group communication system adds overhead for reliable and atomic multicast
- Latency increases with number of active nodes
- Throughput decreases with number of active nodes
- Overhead in acceptable range for this scenario
- Nodes: Pentium III 450MHz on 100MBit/s Ethernet

System	#	Latency	Overhead
TORQUE	1	98 ms	
JOSHUA/TORQUE	1	134 ms	36 ms / 37%
JOSHUA/TORQUE	2	265 ms	158 ms / 161%
JOSHUA/TORQUE	3	304 ms	206 ms / 210%
JOSHUA/TORQUE	4	349 ms	251 ms / 256%

Job Submission Latency Overhead

System	#	10 Jobs	50 Jobs	100 Jobs
TORQUE	1	0.93s	4.95s	10.18s
JOSHUA/TORQUE	1	1.32s	6.48s	14.08s
JOSHUA/TORQUE	2	2.68s	13.09s	26.37s
JOSHUA/TORQUE	3	2.93s	15.91s	30.03s
JOSHUA/TORQUE	4	3.62s	17.65s	33.32s

Job Submission Throughput Overhead

Symmetric Active/Active Availability

- $A_{\text{component}} = \text{MTTF} / (\text{MTTF} + \text{MTTR})$
- $A_{\text{system}} = 1 - (1 - A_{\text{component}})^n$
- $T_{\text{down}} = 8760 \text{ hours} * (1 - A)$
- Single node MTTF: 5000 hours
- Single node MTTR: 72 hours

Nodes	Availability	Est. Annual Downtime
1	98.58%	5d 4h 21m
2	99.97%	1h 45m
3	99.9997%	1m 30s
4	99.999995%	1s

Single-site redundancy for 7 nines does not mask catastrophic events.



Existing Limitations

- The active/standby and asymmetric active/active technology interrupts the service during fail-over
- Generic $n+1$ or $n+m$ asymmetric active/active configurations have not been developed yet
- The $2+1$ asymmetric active/active configuration uses two different service implementations
- The developed symmetric active/active technology has certain stability and performance issues
- All developed prototypes use a customized high availability environment
- Missing interaction with compute node fault tolerance mechanisms (except for HA-OSCAR for head node fail-over)

Most Pressing Issues

- For production-type deployment
 - Stability – guaranteed quality of service
 - Performance – low replication overhead
 - Interaction with compute node fault tolerance mechanisms – e.g. procedure for failing PBS mom
 - ➔ Testing, enhancements, and staged deployment
- For extending the developed technologies
 - Portability – ability to apply technology to different services
 - Ease-of-use – simplified service HA management (RAS)
 - ➔ Generic HA framework needed

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 ultra-scale 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

MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY



NC STATE UNIVERSITY



LOUISIANA TECH
UNIVERSITY



The University of Reading

CRAY THE SUPERCOMPUTER COMPANY

Towards High Availability for High-Performance Computing System Services: Accomplishments and Limitations



Christian Engelmann^{1,2}, Stephen L. Scott¹,
Chokchai (Box) Leangsuksun³, Xubin (Ben) He⁴

¹ Oak Ridge National Laboratory, Oak Ridge, USA

² The University of Reading, Reading, UK

³ Louisiana Tech University, Ruston, USA

⁴ Tennessee Tech University, Cookeville, USA