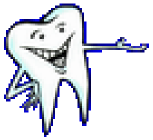
Symmetric Active/Active Metadata Service for Highly Available Cluster Storage Systems

Li Ou[1], Xubin (Ben) He[1], Xin Chen[1],
Christian Engelmann[2,3], Stephen L. Scott[2]

[1] Tennessee Tech University, Cookeville, USA
[2] Oak Ridge National Laboratory, Oak Ridge, USA
[3] The University of Reading, Reading, UK

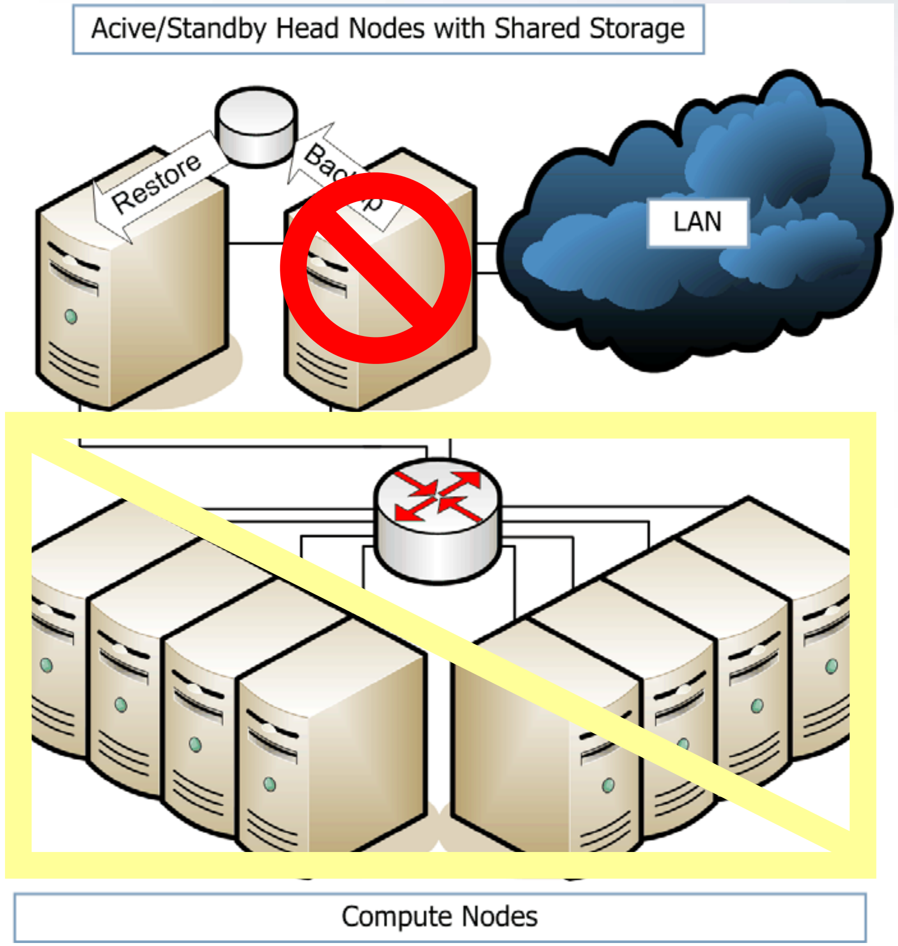


Background and Motivation

- Large-scale HPC systems with 10,000-210,000 processors
- 100-480 TFlop/s LINPACK performance and 10-100TB RAM.
- Current systems: IBM Blue Gene/P and Cray XT4 - Top500.org
- Next-generation: Peta-scale IBM Blue Gene/P and Cray ??
- Applications: Climate change, nuclear astrophysics, fusion energy, materials sciences, biology, nanotechnology, …
- Single application runs for days, weeks, and even months

- Reliability, availability, and serviceability is crucial for success
- Head and service nodes are single points of failure and control
- Efficient redundancy strategies are needed for high availability

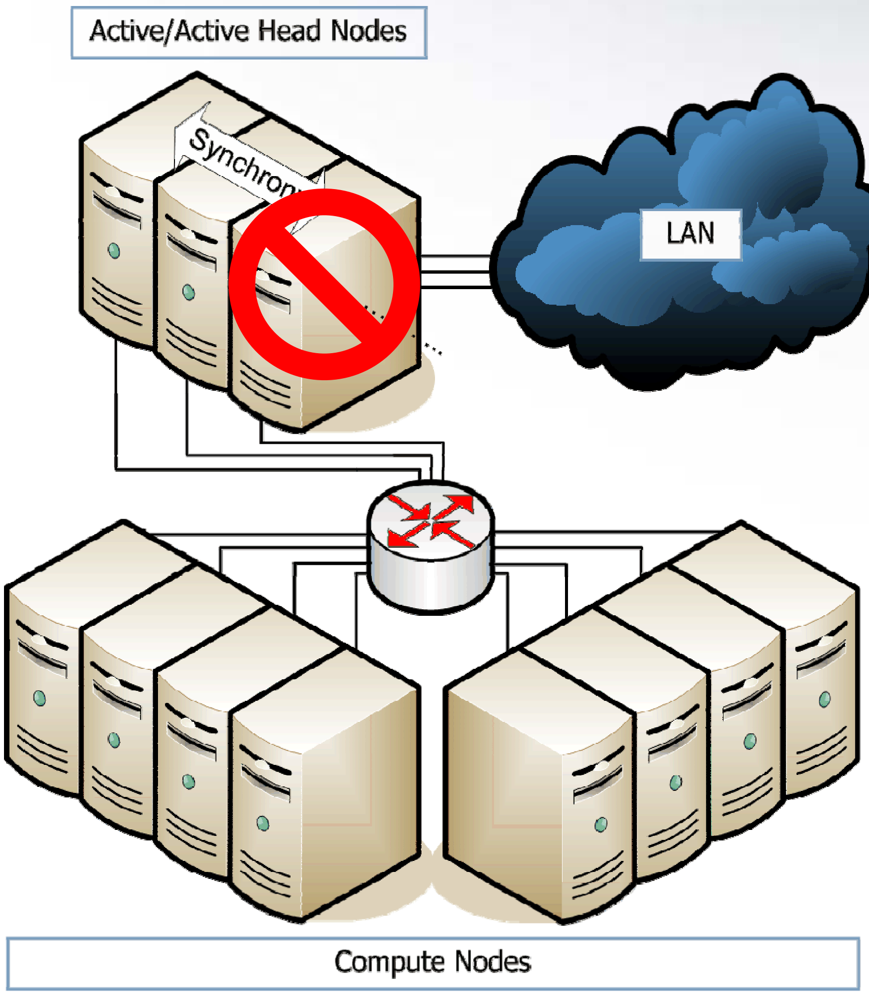
Active/Standby with Shared Storage



- Commonly employed technique
- 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 ALWAYS guaranteed

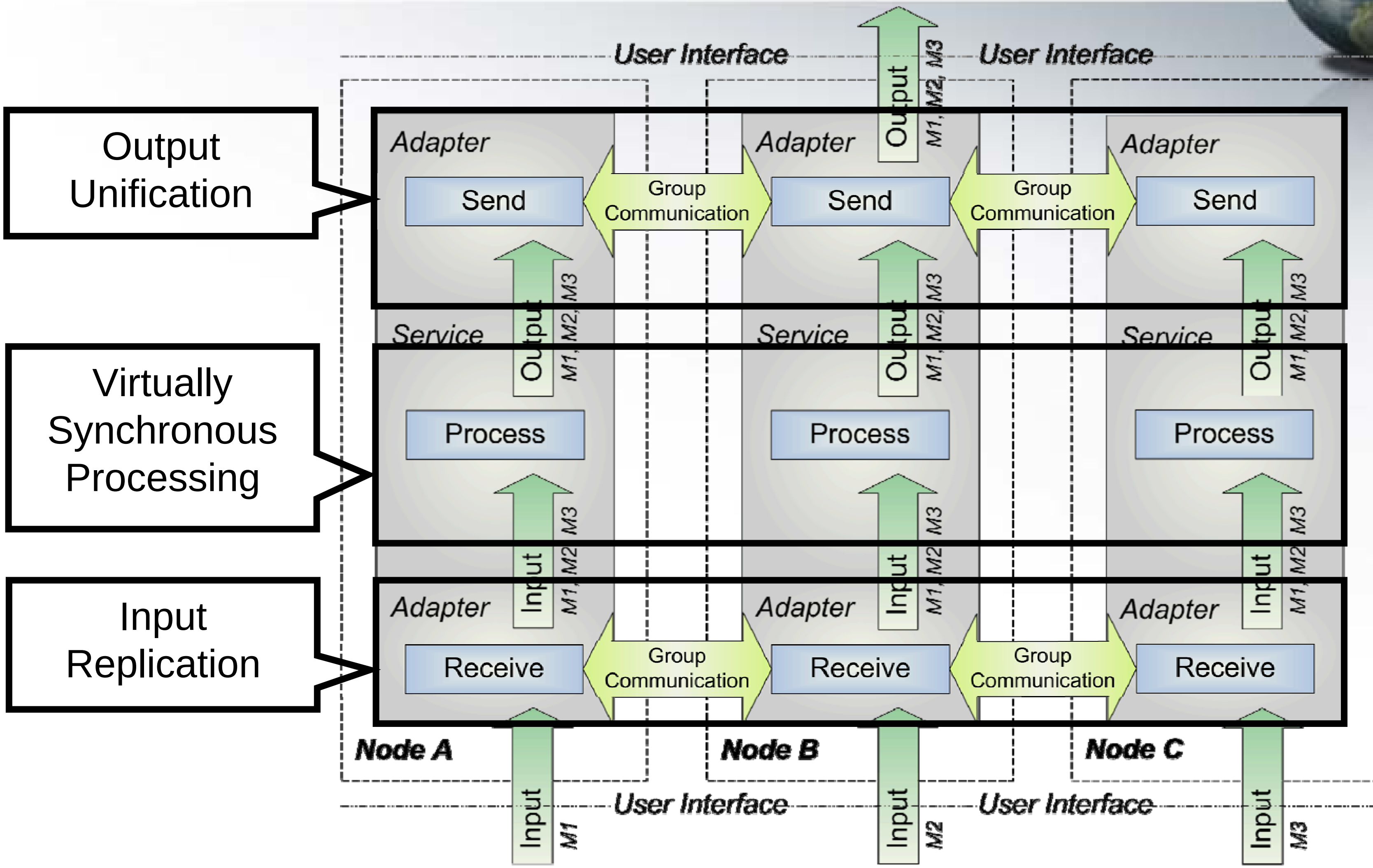
➔ Metadata servers (MDS) of Parallel Virtual File System (PVFS) and Lustre (not part of standard install)

Symmetric Active/Active Redundancy

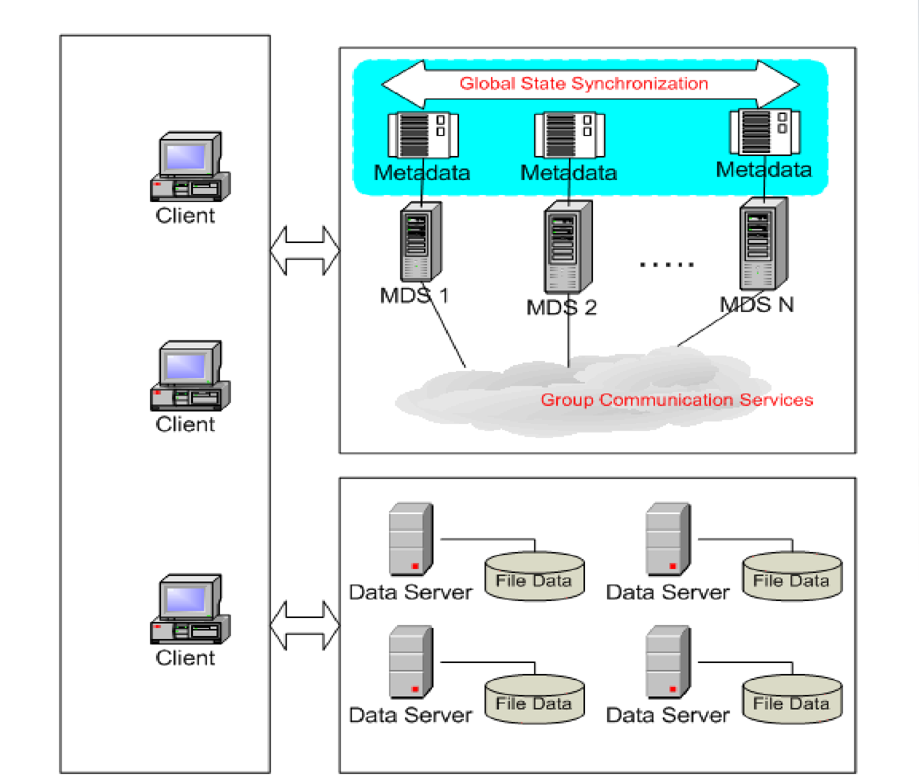


- Many active service nodes
- Work load distribution
- Symmetric replication between service nodes
- Continuous service
- Always up-to-date
- No fail-over, no restore-over
- State-machine replication
- Virtual synchrony model
- Complex algorithms (process group communication)
- JOSHUA prototype for Torque batch job queue and resource manager
- PVFS MDS (this paper)

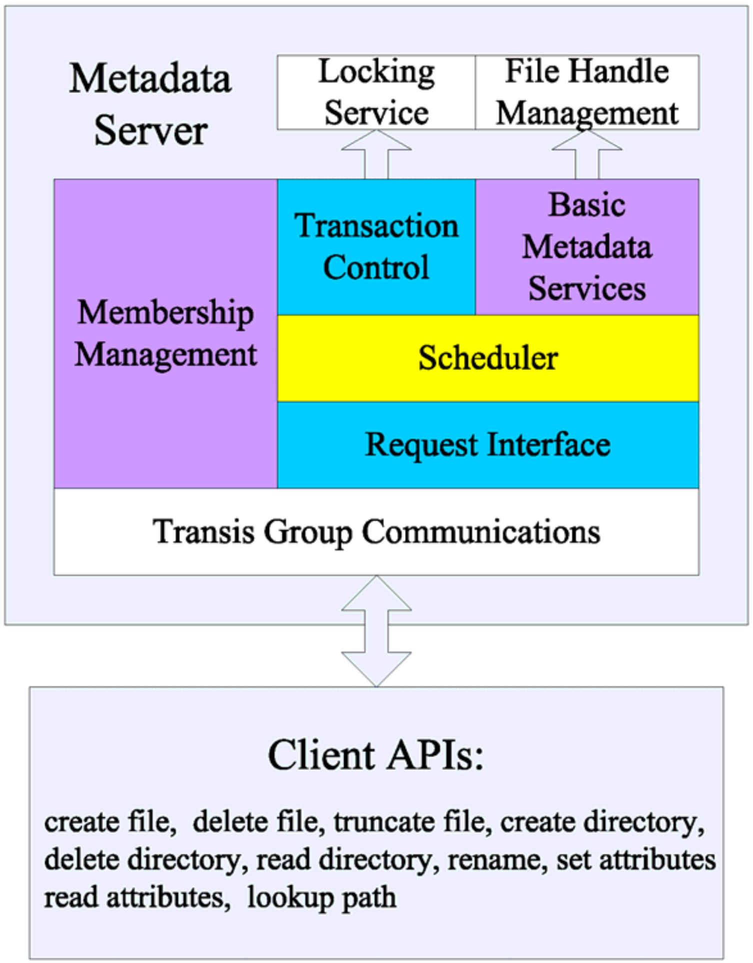
Internal Replication Method



Symmetric Active/Active Metadata Service Architecture



Symmetric Active/Active Metadata Service Design for the Parallel Virtual File System

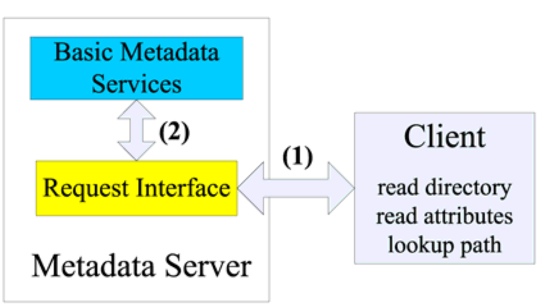


	read	update	write
read			X
update		X	X
write	X	X	X

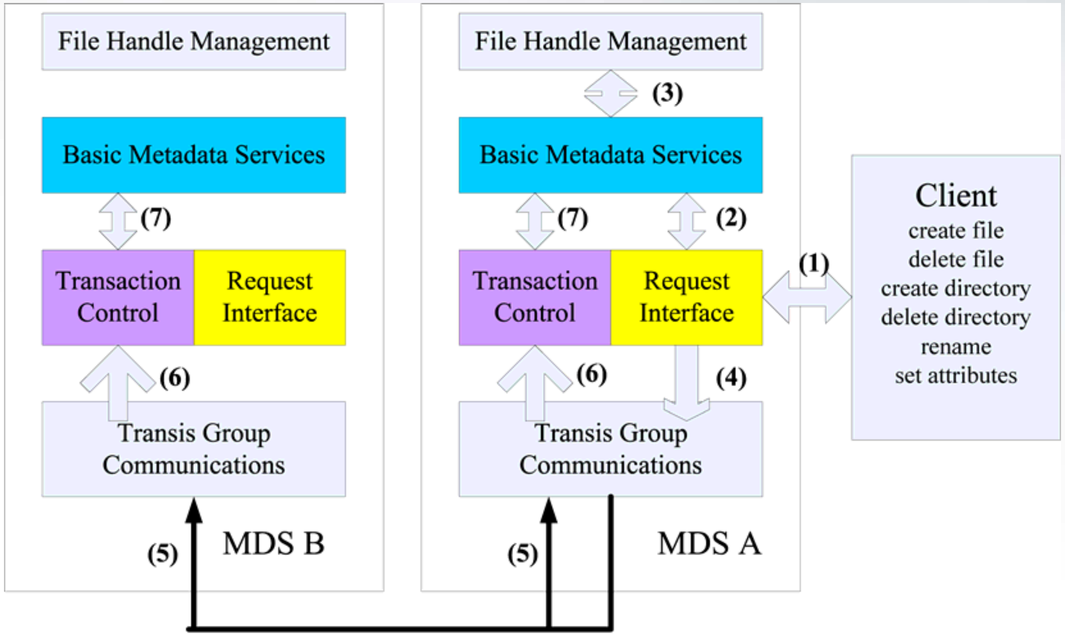
MDS Record Locking Conflicts Resolved by Transaction Control

Symmetric Active/Active PVFS Metadata Service Request Handling

MDS Read Request Handling



MDS Write Request Handling



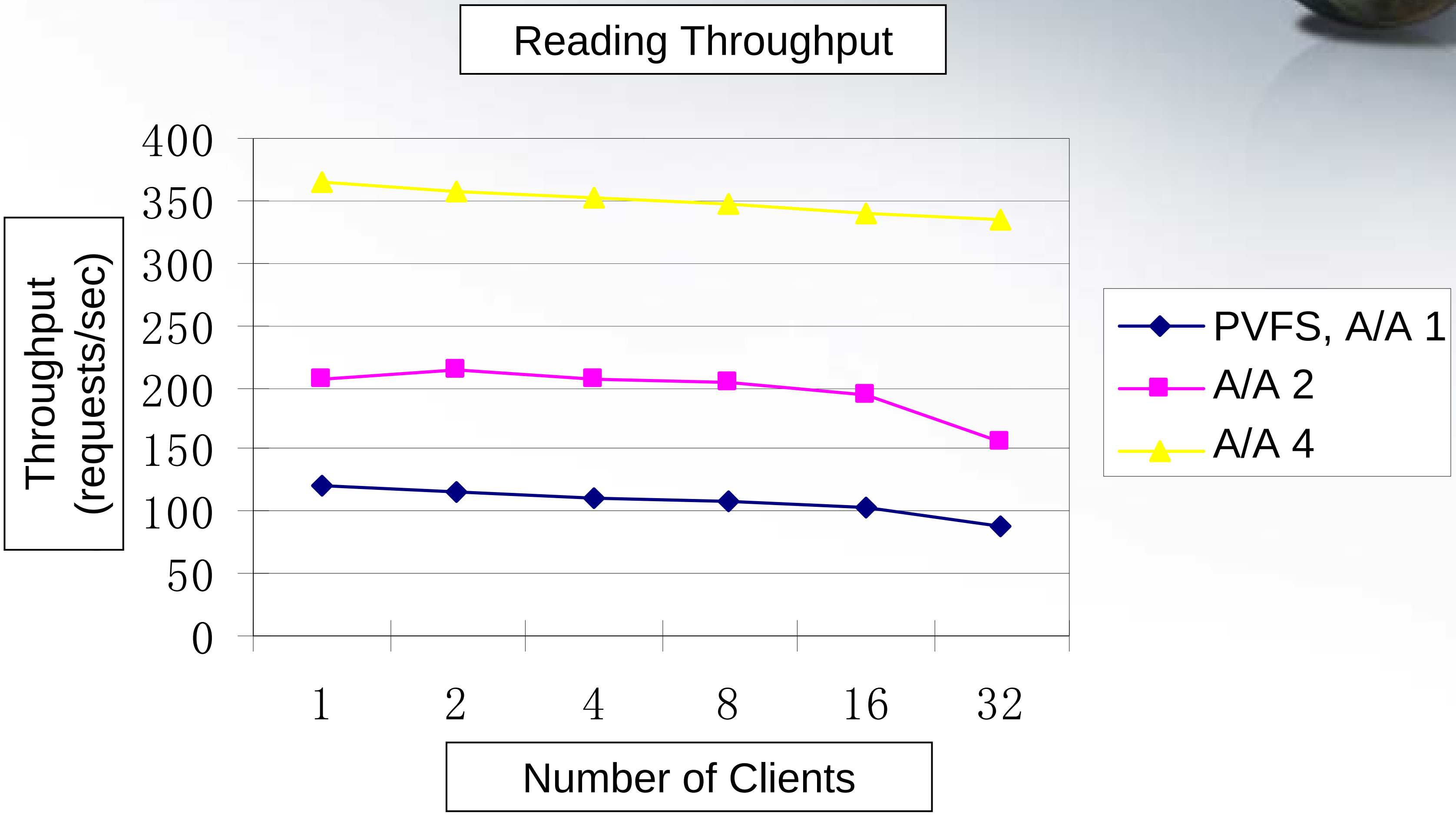
Experimental Setup

- Transis v1.03 group communication system with fast delivery protocol (see ICCCN 2007 paper)
- Parallel Virtual File System (PVFS) v2
- ORNL XTORC cluster
  - Dual Intel Pentium 2GHz nodes
  - 768MB memory and 40 GB hard disk space per node
  - 100MBit/s Fast Ethernet (full duplex)
  - Linux Fedora Core 5 operating system
- MPI-based benchmark using multiple clients to send concurrent MDS read and write requests

Symmetric Active/Active PVFS Metadata Service Performance and Overhead

100 Mbps Ethernet LAN Environment

Symmetric Active/Active PVFS Metadata Service Performance and Overhead



100 Mbps Ethernet LAN Environment

Symmetric Active/Active PVFS Metadata Service Availability Improvement

- $A_{component} = MTTF / (MTTF + MTTR)$
- $A_{system} = 1 - (1 - A_{component})\ n$
- $T_{down} = 8760\ hours * (1 – A)$
- 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.

Summary and Future Work

- Developed a symmetric active/active metadata service (MDS) for the Parallel Virtual File System (PVFS)
- Solution ensures no loss of service and no loss of state
- Minimal performance impact for MDS write requests
- Significant performance gain for MDS read requests
- Significant availability improvement for MDS
- Employed concepts are applicable to any networked file system that utilizes a MDS

- Ongoing work focuses on more complex scenarios in service-oriented architectures (SOAs), such as dependencies between multiple services

Symmetric Active/Active Metadata Service for Highly Available Cluster Storage Systems

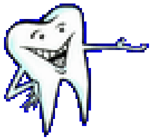
Li Ou[1], Xubin (Ben) He[1], Xin Chen[1],
Christian Engelmann[2,3], Stephen L. Scott[2]

[1] Tennessee Tech University, Cookeville, USA
[2] Oak Ridge National Laboratory, Oak Ridge, USA
[3] The University of Reading, Reading, UK

Tennessee Tech UNIVERSITY

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

The University of Reading



Office of Science
U.S. DEPARTMENT OF ENERGY