Scalability of InfiniBand-Connected LNET Routers

Team Light Coral

Computer System, Cluster, and Networking Summer Institute

Emily Baldwin — Wheaton College
Matthew Schauer — Georgia Institute of Technology
Jarrett Crews — New Mexico Institute of Mining and Technology

Susan Coulter — HPC-3
David Bonnie — HPC-3
Christopher Hoffman — HPC-3
Dane Gardner — Instructor

LA-UR-14-26018

Overview

Background

Objective

Cluster Set-Up

Benchmark Methods

Results

Obstacles

Future Work

Background

- Lustre File System
  - Servers
  - Network (LNET) router
  - Clients
- InfiniBand
  - FDR – 56 Gb/s
  - IP over IB
- IOR Performance Benchmarks

Objective

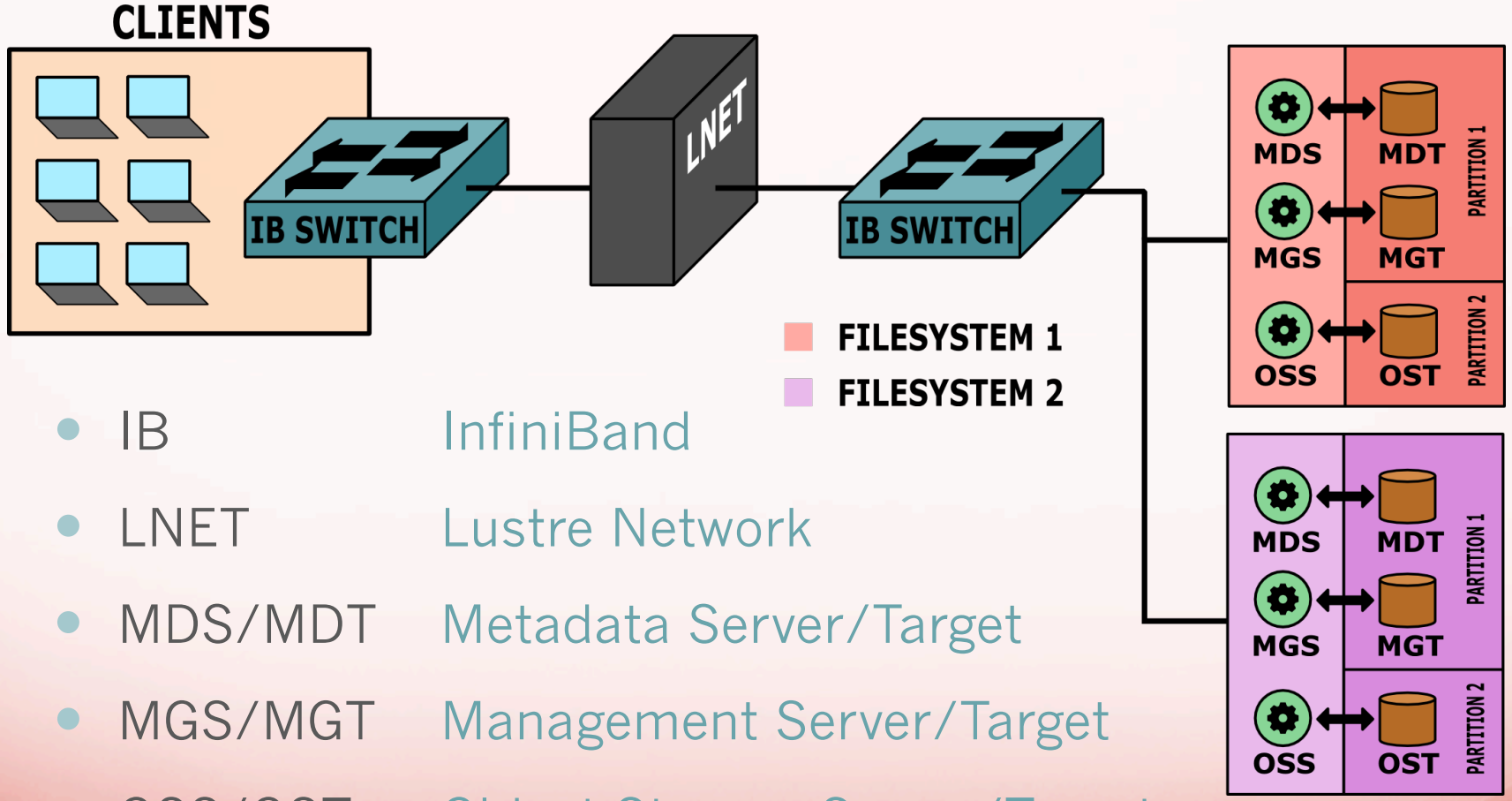
Is it possible to link multiple Lustre File Systems to a single LNET router?

If so, what is the read/write performance of multiple file systems from many client nodes?

Motivation

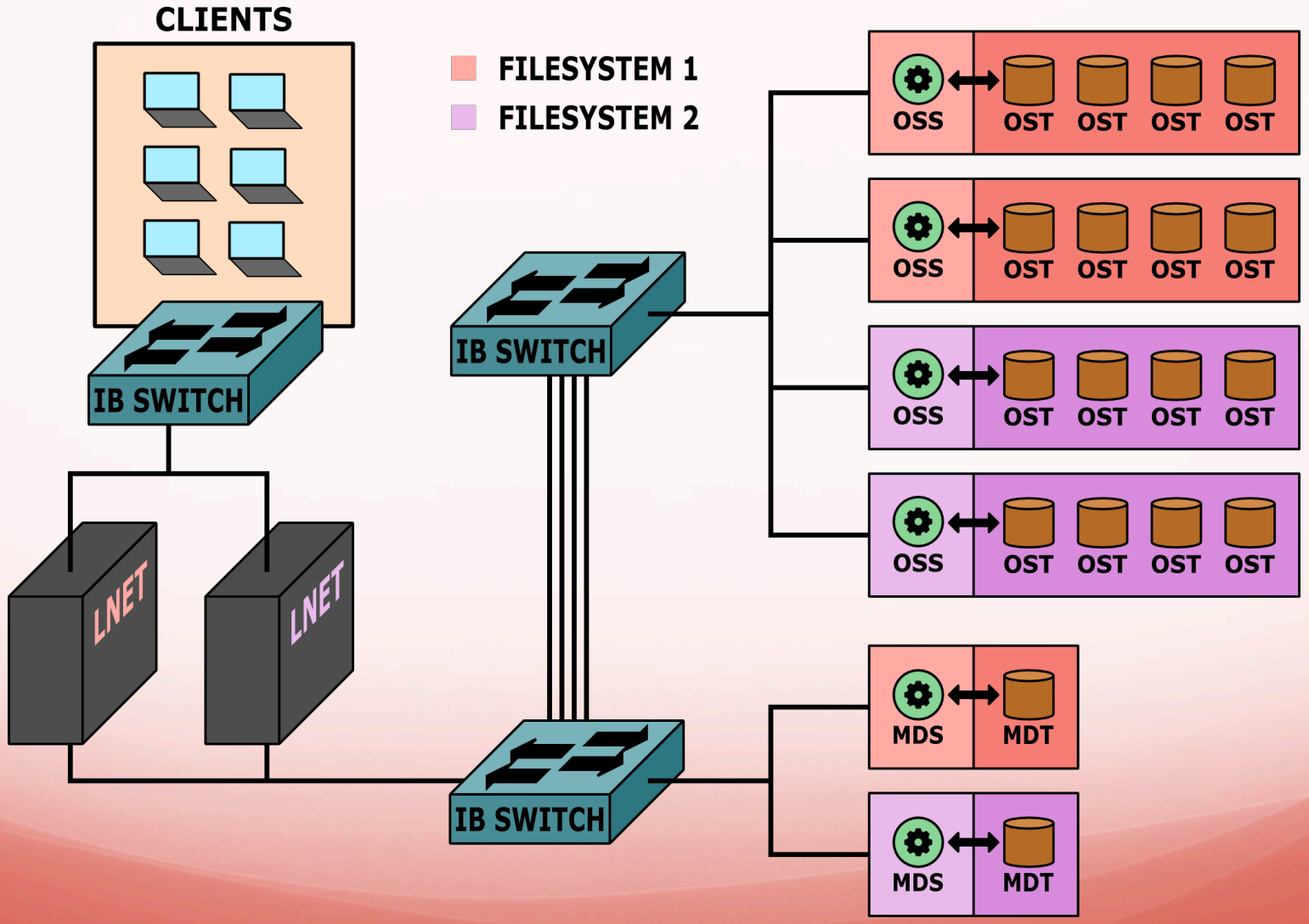
- LANL converting to Lustre from Panasas
- Never more than one file system per LNET router
  - Not cost-effective
  - Wasted router potential
- Multiple file systems per LNET router arrangement
  - No loss in performance?
  - No significant change in router utilization?
- Potential for easier transition from legacy machines

Cluster Set-Up



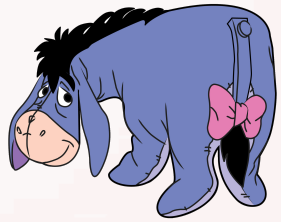
- IB InfiniBand
- LNET Lustre Network
- MDS/MDT Metadata Server/Target
- MGS/MGT Management Server/Target
- OSS/OST Object Storage Server/Target

Cluster Set-Up



Benchmark Methods

- IOR benchmarking tool
  - Writes/reads variable amounts of data
  - Parameters for file size, block size, files per node, etc.
  - Reports bandwidth statistics



- eeyore
  - Automates testing with IOR
  - Sequence a write, read, then simultaneous write/read
  - Script parameters include: file size, block size, nodes, and processes per node

Op	Mean	Max	Min	Stdev
w	535.73	544.19	528.31	5.64
r	410.05	416.34	405.79	3.88

Benchmark Methods

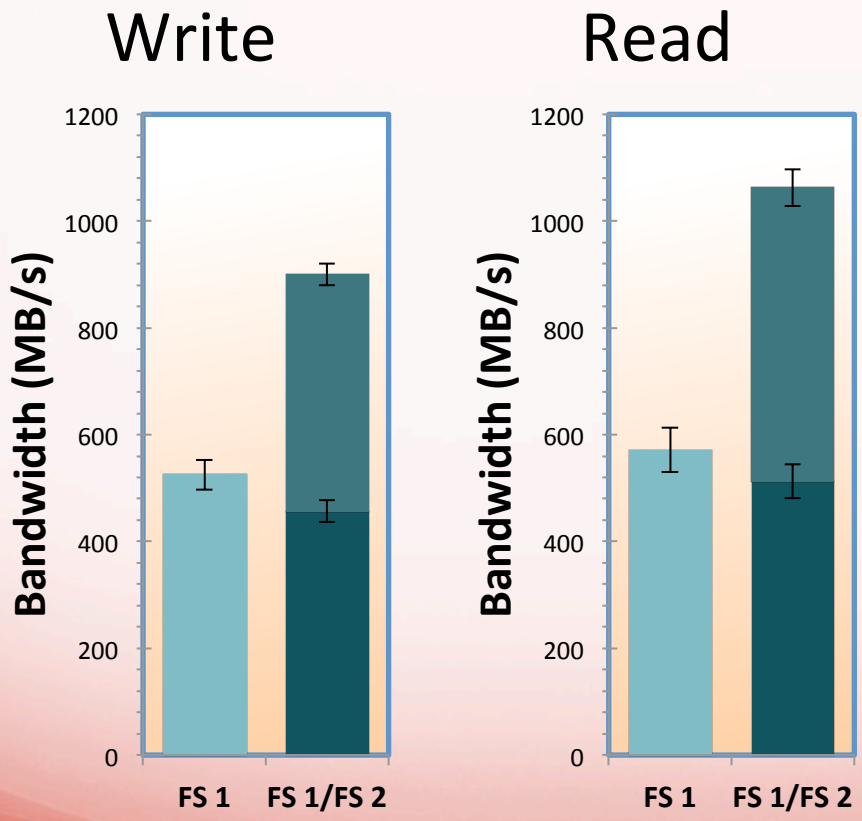
- Run each test *n* times, collect mean and standard deviation
- Test parameter combinations:

Number of Nodes	File Size/ Process	Block Size	Processes/Node	Total Transfer Size
6	32 GB	1 GB	1	192 GB
6	32 GB	512 MB	1	192 GB
6	32 GB	2 KB	1	192 GB
6	1 GB	1 GB	24	144 GB
6	1 GB	512 MB	24	144 GB
6	1 GB	2 KB	24	144 GB

Results

32 GB files, 512 MB block size

Write, then read

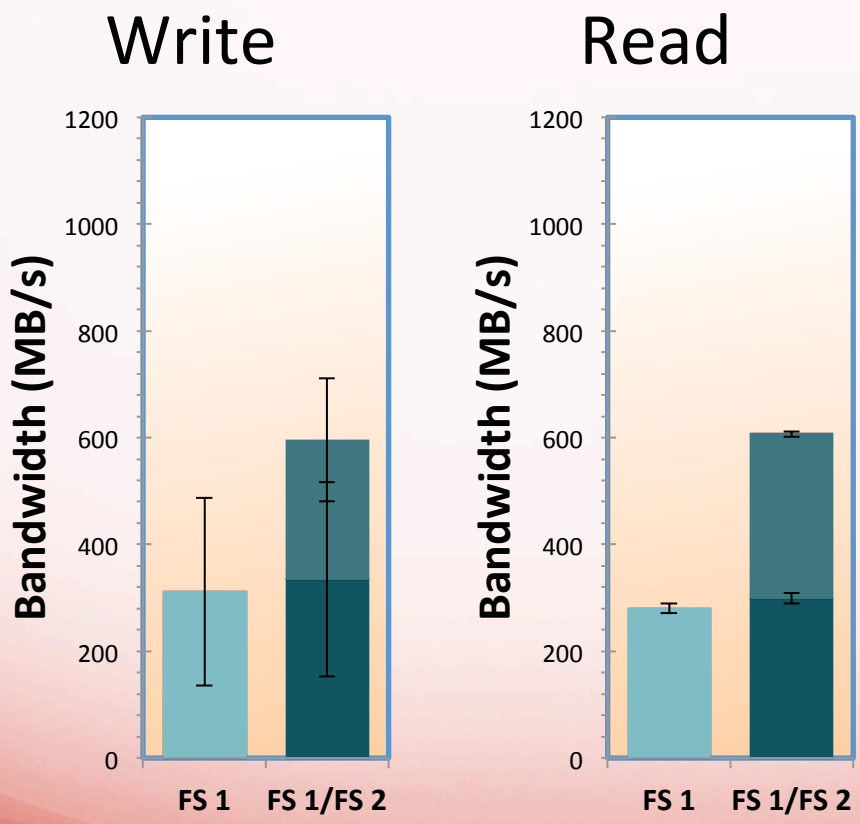


- 500 MB/s file system 1
  - Across 5 disks
- 400 MB/s file system 2
  - Bad DIMM
- Two file system bandwidth is sum of individual file system bandwidths
- Small standard deviation
  - Consistent results over many test runs

Results

32 GB files, 512 MB block size

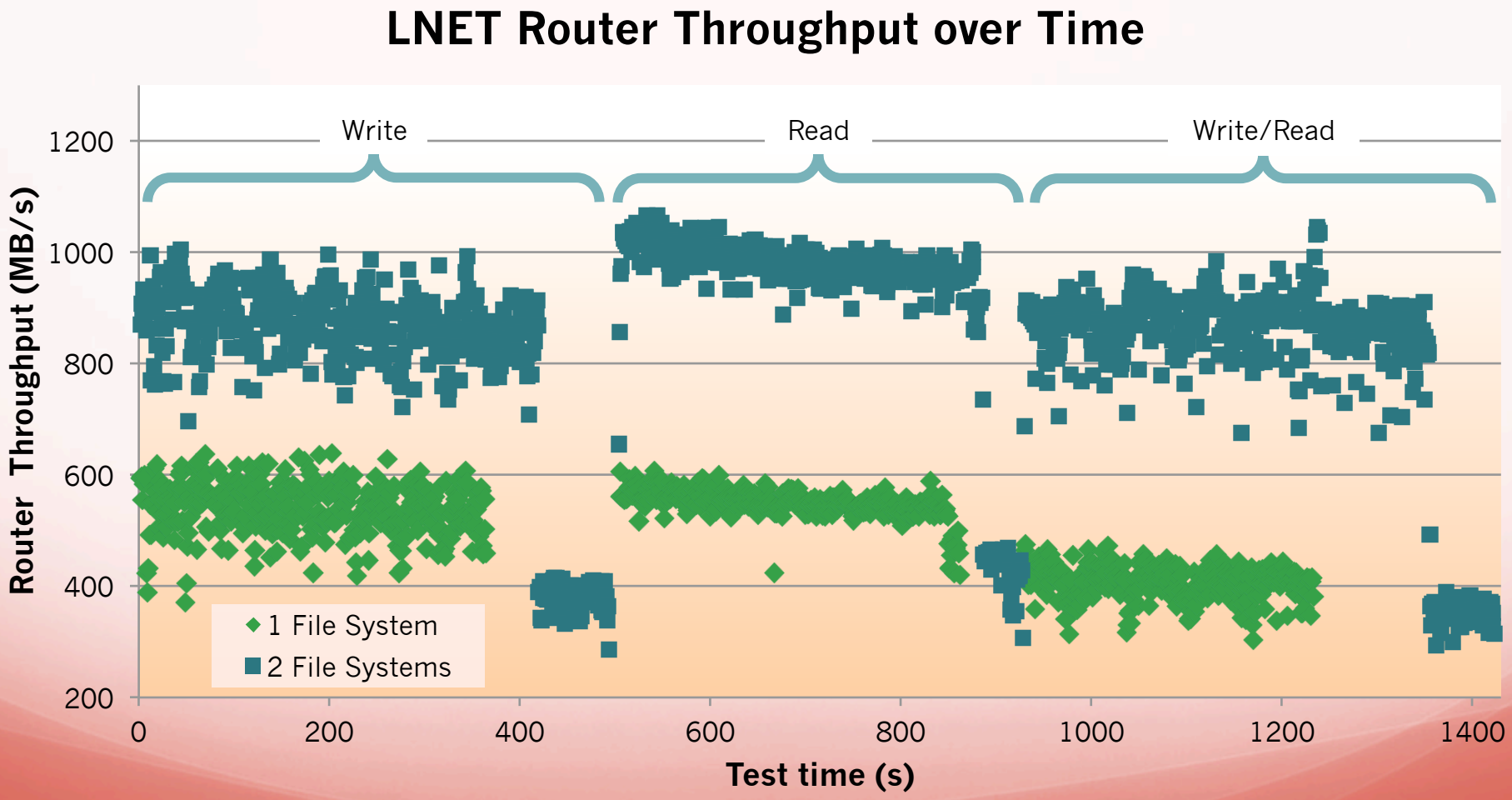
Simultaneous write and read



- Similar results to sequential write and read
- Large standard deviation on two file system writes

Results

32 GB files, 512 MB block size



Discussion

- LNET routers scale beyond a single file system
- Expected bottleneck does not exist in router
  - Negligible router CPU load
- Two file systems performed at expected capacity
- Scalability plausible
  - Bandwidth trend may not continue

Obstacles

- Lustre incompatibility with stock kernel
  - Server and client utilities
- 10% bandwidth loss
  - Removed LNET router
  - One file system performed slower than other
    - Discovered bad DIMM
  - Consistent results despite hardware issue

Future Work

- Scalability of LNET routers to more file systems
- More complex setups
  - Lustre file system components on different servers
  - Heterogeneous networks connected partially with InfiniBand and partially with Ethernet
  - Multiple Lustre networks with varying number of servers
  - Multiple routers connecting many Lustre networks

Los Alamos NATIONAL LABORATORY EST.1943

PRObE Parallel Reconfigurable Observational Environment

Thank You!

Emily Baldwin Wheaton College
baldwinemb@gmail.com
Matthew Schauer Georgia Institute of Technology
matthew.schauer.x@gmail.com
Jarrett Crews New Mexico Institute of Mining and Technology
jarrett.crews@gmail.com

Susan Coulter HPC-3
David Bonnie HPC-3
Christopher Hoffman HPC-3
Dane Gardner Instructor

New Mexico CONSORTIUM

ISTI INFORMATION SCIENCE & TECHNOLOGY INSTITUTE

NSF

Georgia Tech

NEW MEXICO TECH SCIENCE • ENGINEERING • RESEARCH • UNIVERSITY

Questions?

Background Lustre, IOR, InfiniBand

Objective >1 Lustre file systems, 1 LNET router

Cluster Set-Up Lustre file system, LNET router

Benchmark Methods bandwidth stats, parameters

Results nearly double bandwidth, scalability plausible

Obstacles Lustre kernel, 10% loss

Future Work more file systems, more complex setups