LA-UR-12-23586

GlusterFS

One Storage Server to Rule Them All

Team Members:

Matthew Broomfield (New Mexico Tech)

Eric Boyer (Michigan Tech)

Terrell Perrotti (South Carolina State University)

Mentors:

David Kennel (DCS-1)

Greg Lee (DCS-1)

Instructors:

Dane Gardner (NMC)

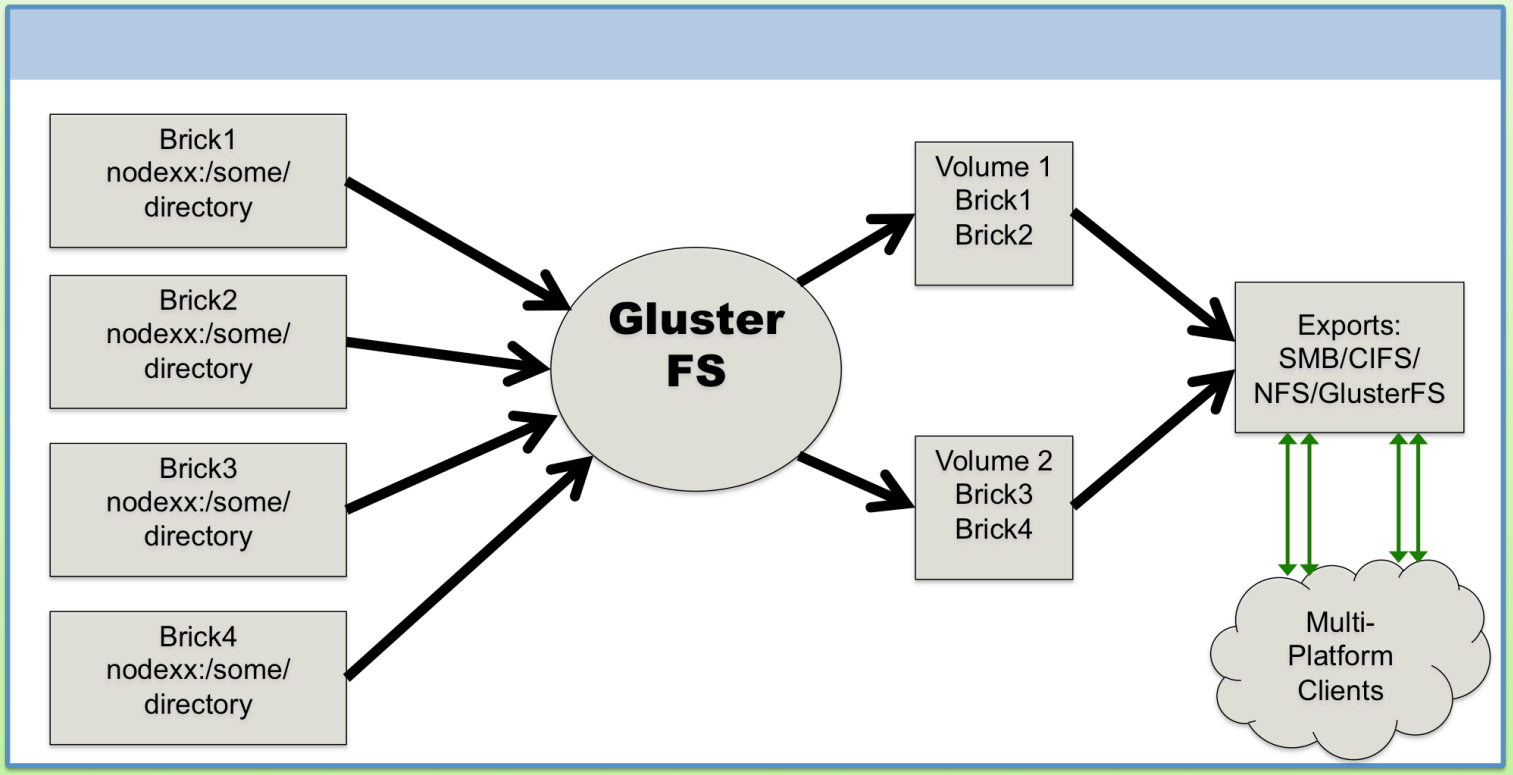
Andree Jacobson (NMC)

Outline

- Introduction to GlusterFS
- Services
- Administration
- Performance
- Conclusions
- Future Work

What is GlusterFS?

- GlusterFS is a Linux based distributed file system, designed to be highly scalable and serve many clients



LA-UR-12-23586

Why Use GlusterFS?

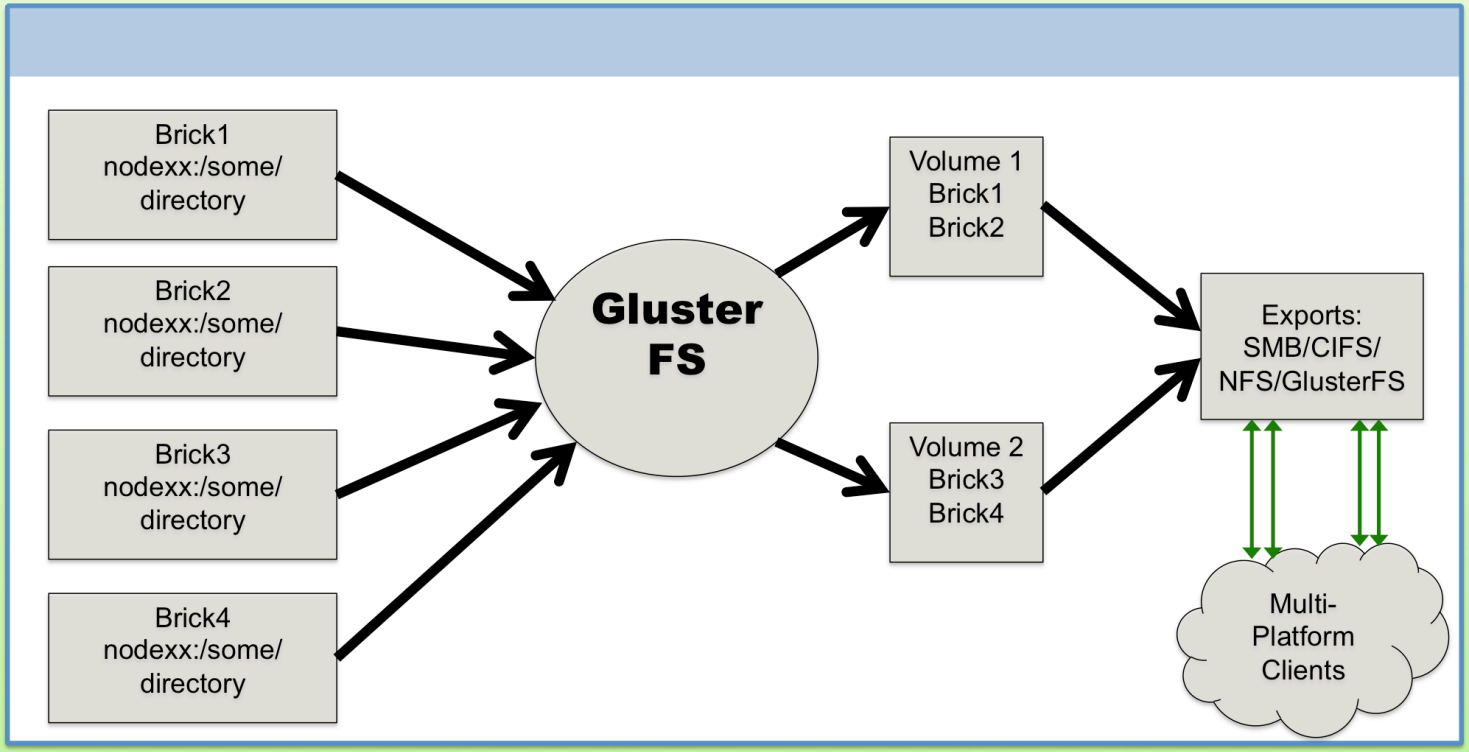
- No centralized metadata server
- Scalability
- Open Source
- Dynamic and live service modifications
- Can be used over Infiniband or Ethernet
- Can be tuned for speed and/or resilience
- Flexible administration

Where Is It Useful?

- Enterprise environments
  - Virtualization
- High Performance Computing (HPC)
- Works with Mac, Linux, and Windows clients

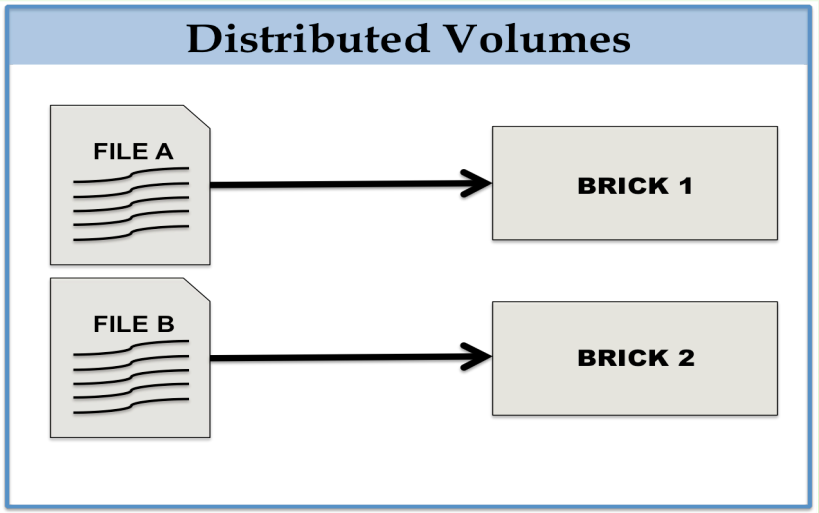
How Does It Work?

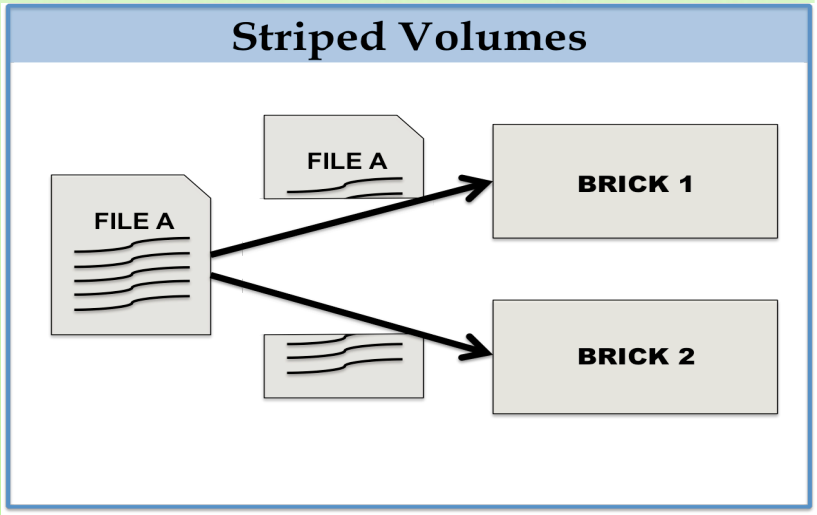
- Individual nodes export bricks (directories) to GlusterFS
- GlusterFS combines bricks into virtual volumes

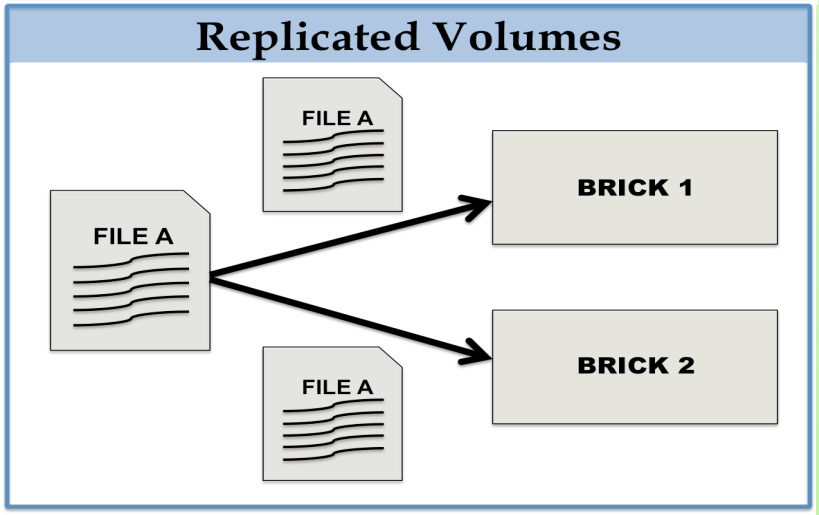


GlusterFS Volume Types

- Different GlusterFs volume types:





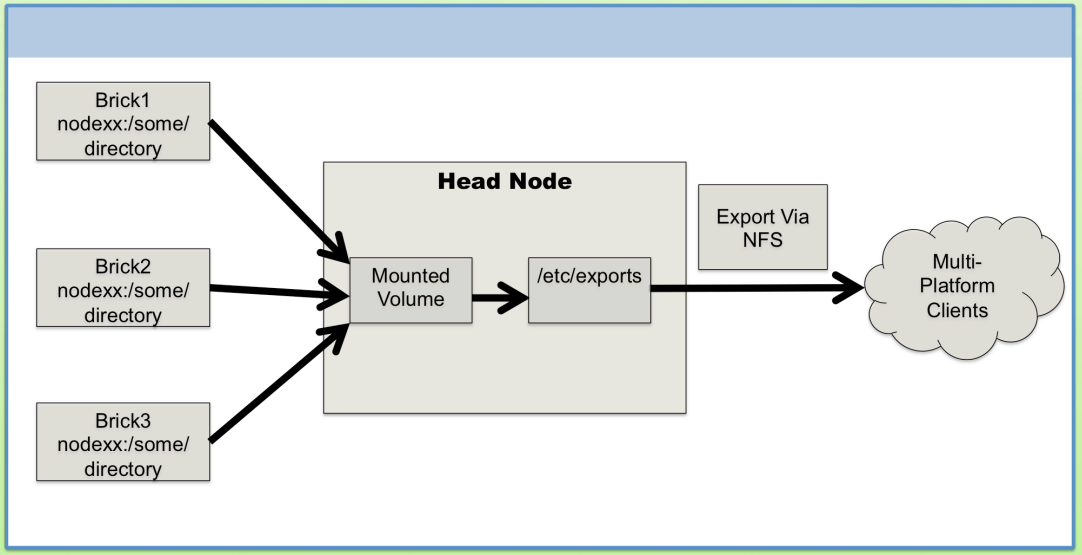


Data Control

- GlusterFS has a built in quota service and uses POSIX ACLs for user control
- POSIX ACLs
  - Can set individual users or group permissions
- Quotas
  - Given via directory
- Both apply to Mac and windows clients (NFS/ SAMBA)

Exporting

- GlusterFS volumes can be exported via NFSv3
  - POSIX ACLs are lost when exporting directly via NFS
  - Enable POSIX ACLs by mounting via GlusterFS and exporting via NFS
- SAMBA allows Windows users to modify NTFS permissions on files



Data Support

- Snapshots
  - We used the rsnapshot utility to enable snapshots
  - Use cron jobs to specify snapshot intervals and locations
- Auditing
  - The auditd utility can be used in conjunction with GlusterFS
  - Shows detailed file interactions
- GlusterFS built-in logging support
  - Performance
  - Diagnostics
  - Events (Warnings, Errors, General Information)

Administration

- GlusterFS has an intuitive CLI
  - Allows for quick volume tuning, shrinking, and expanding while the system is online and available
  - Easily integrated into current infrastructure

- Pitfalls
  - Latency induced when mounting and exporting
  - GlusterFS mounting/unmounting occasionally hung
  - Metadata is distributed, thus harder to remove

Total Cost of Operation

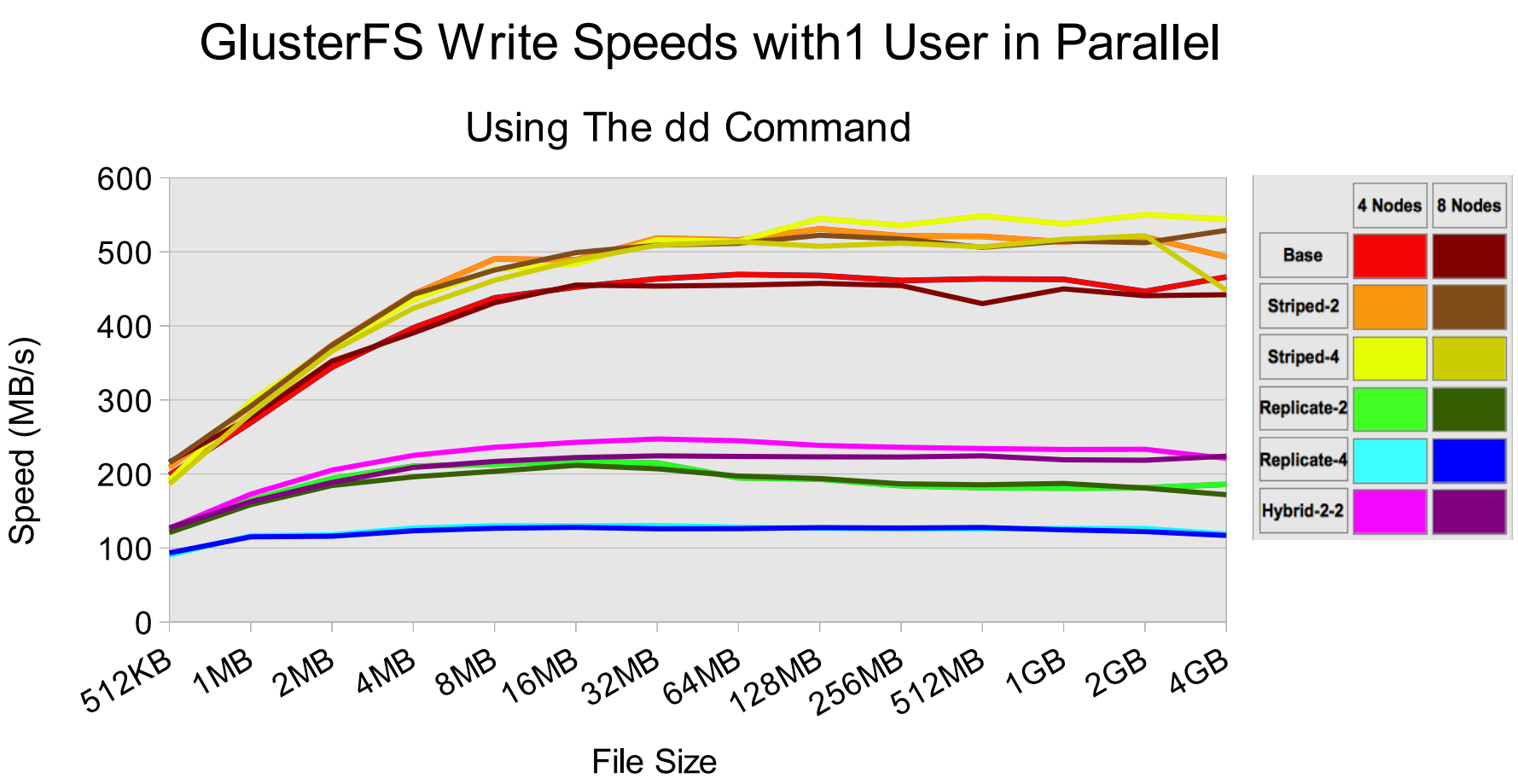
- Open Source
- Can use commodity hardware
- Can use 1 or 10Gbps Ethernet

GLUSTER

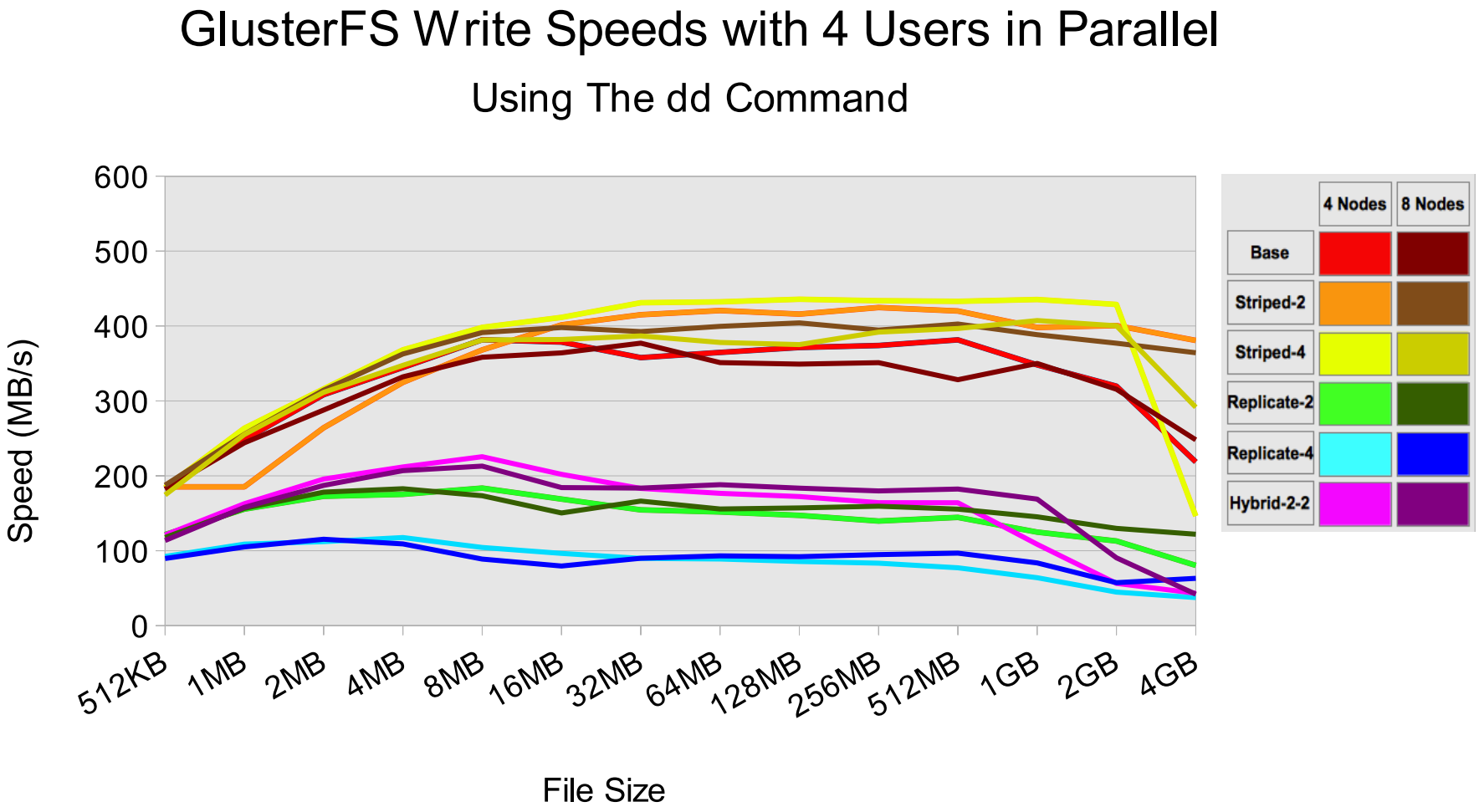
Performance Testing Key

- Base-N — Distributed across N nodes
- Striped-X-N — Striped across X nodes on an N node volume
- Replicated-X-N — Replicated across X nodes on an N node volume
- Hybrid-X-Y-N — Striped across X nodes, Distributed across Y nodes on an N node volume

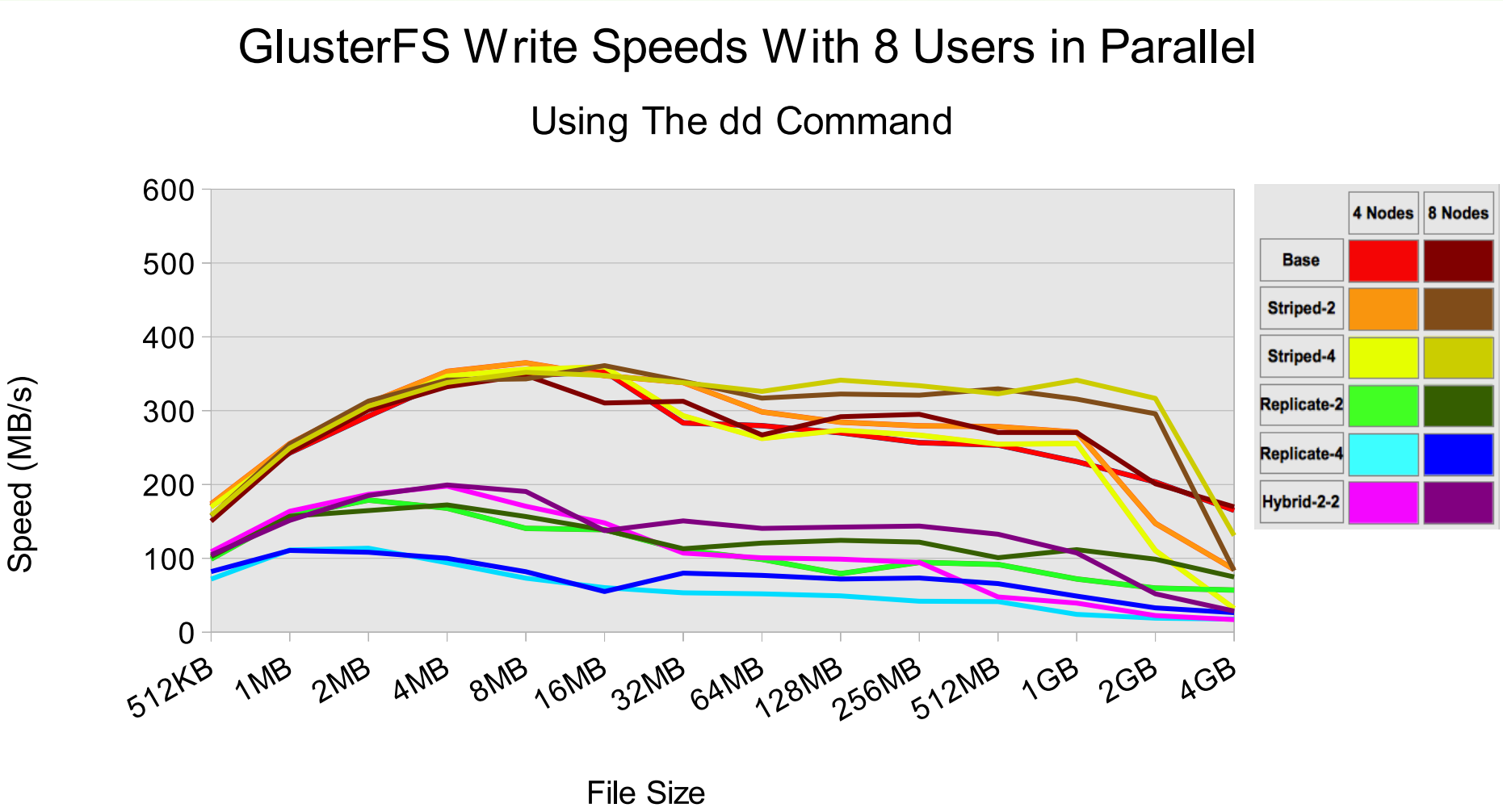
Write Performance



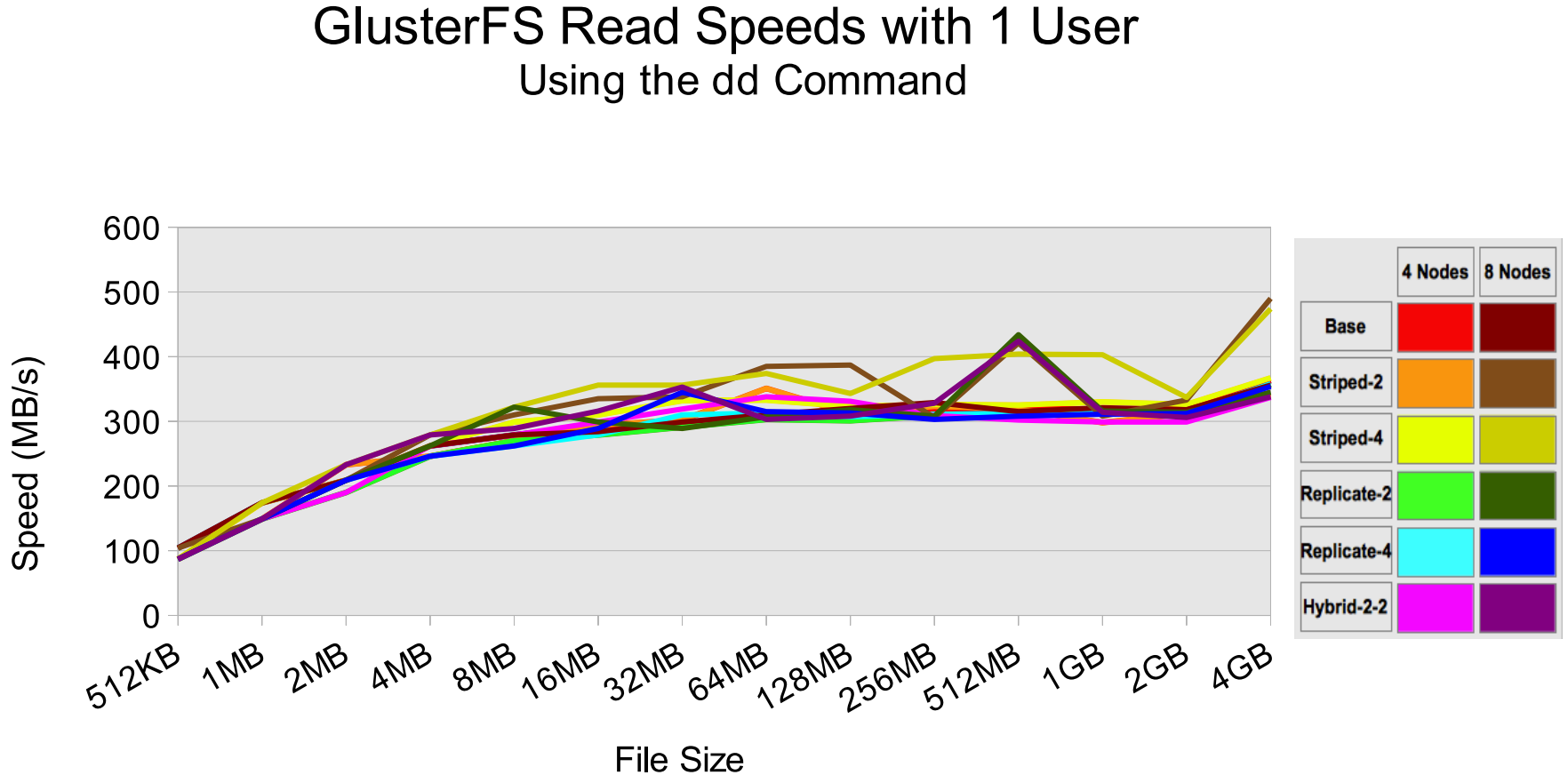
Write Performance



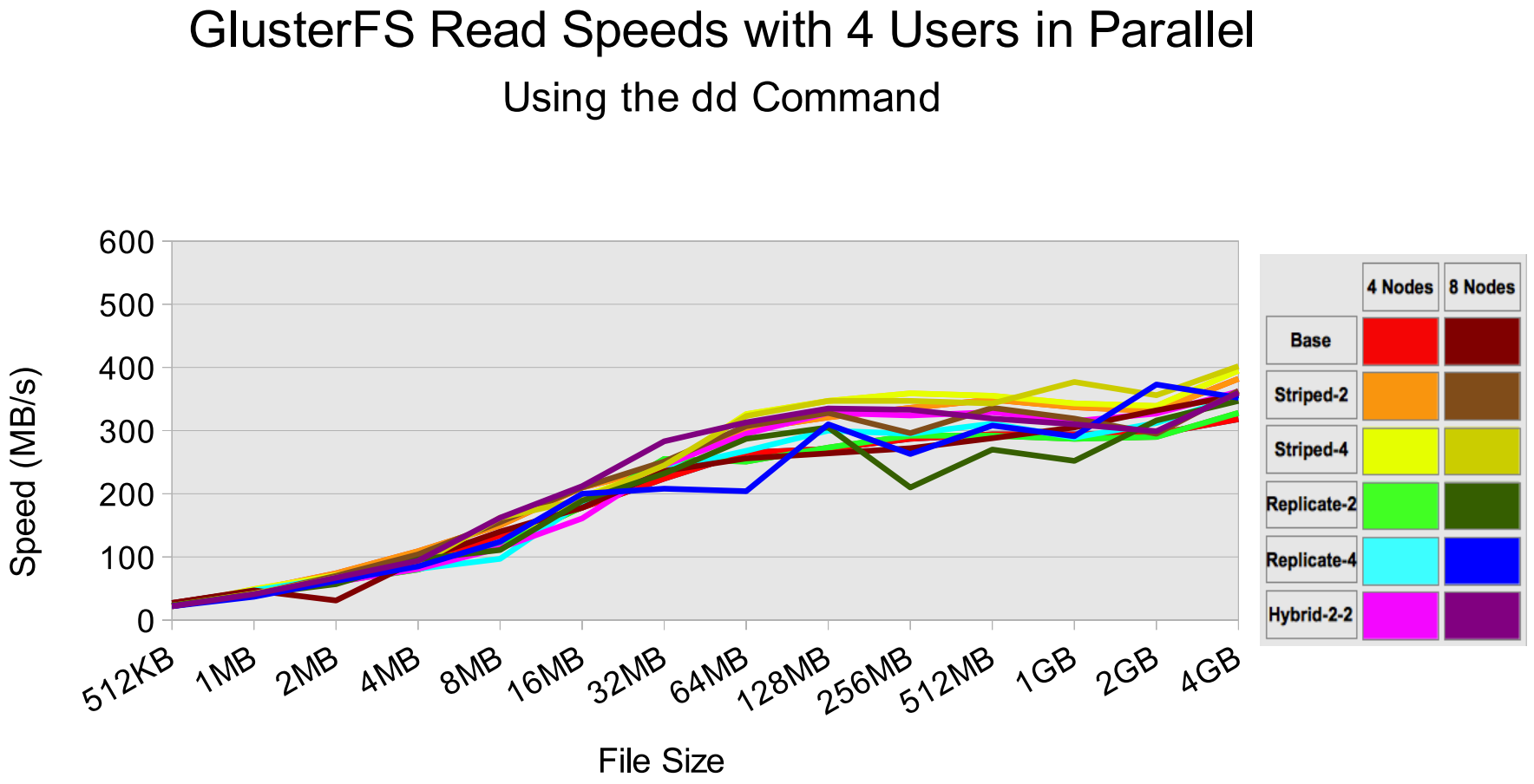
Write Performance



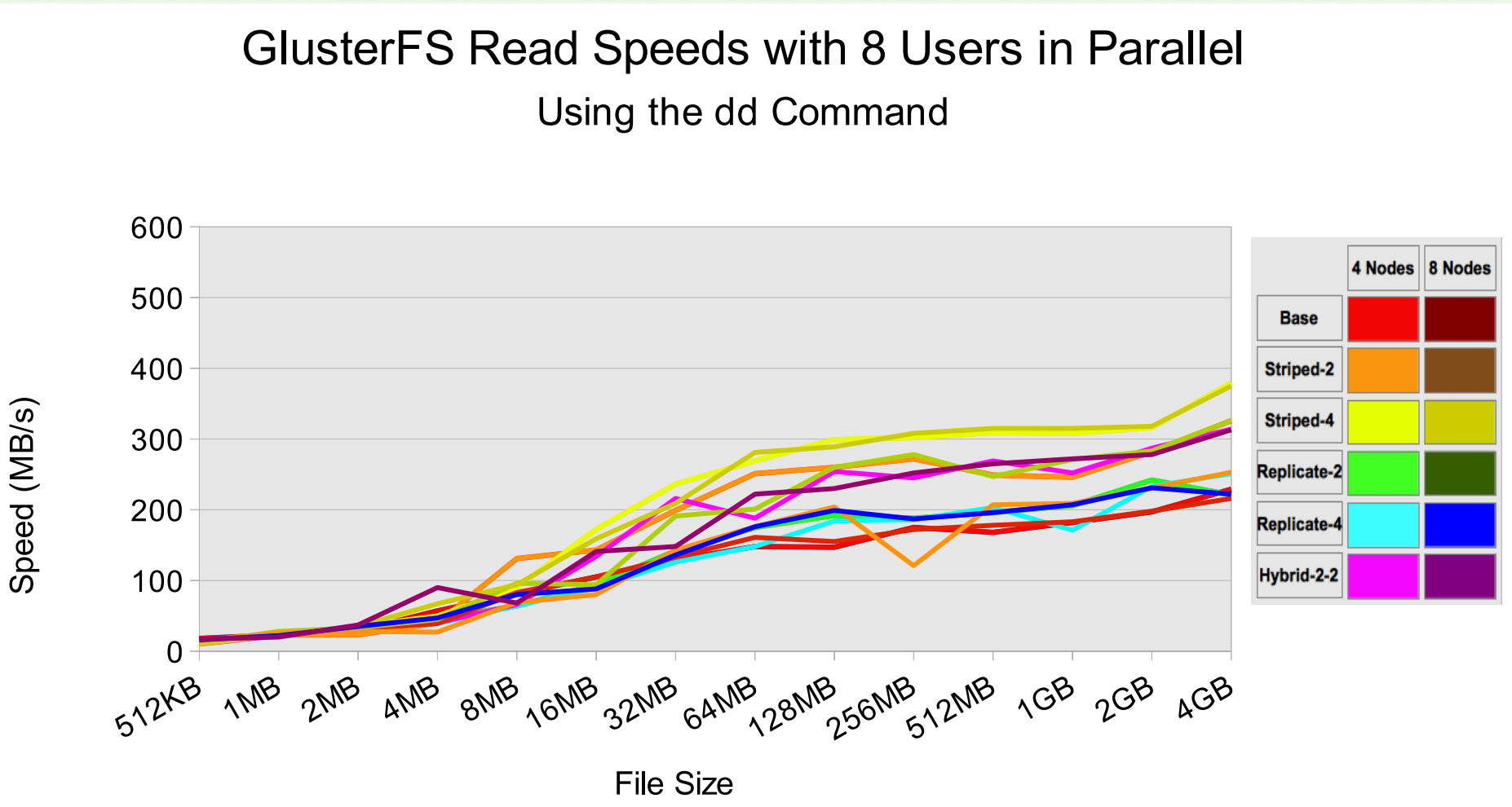
Read Performance



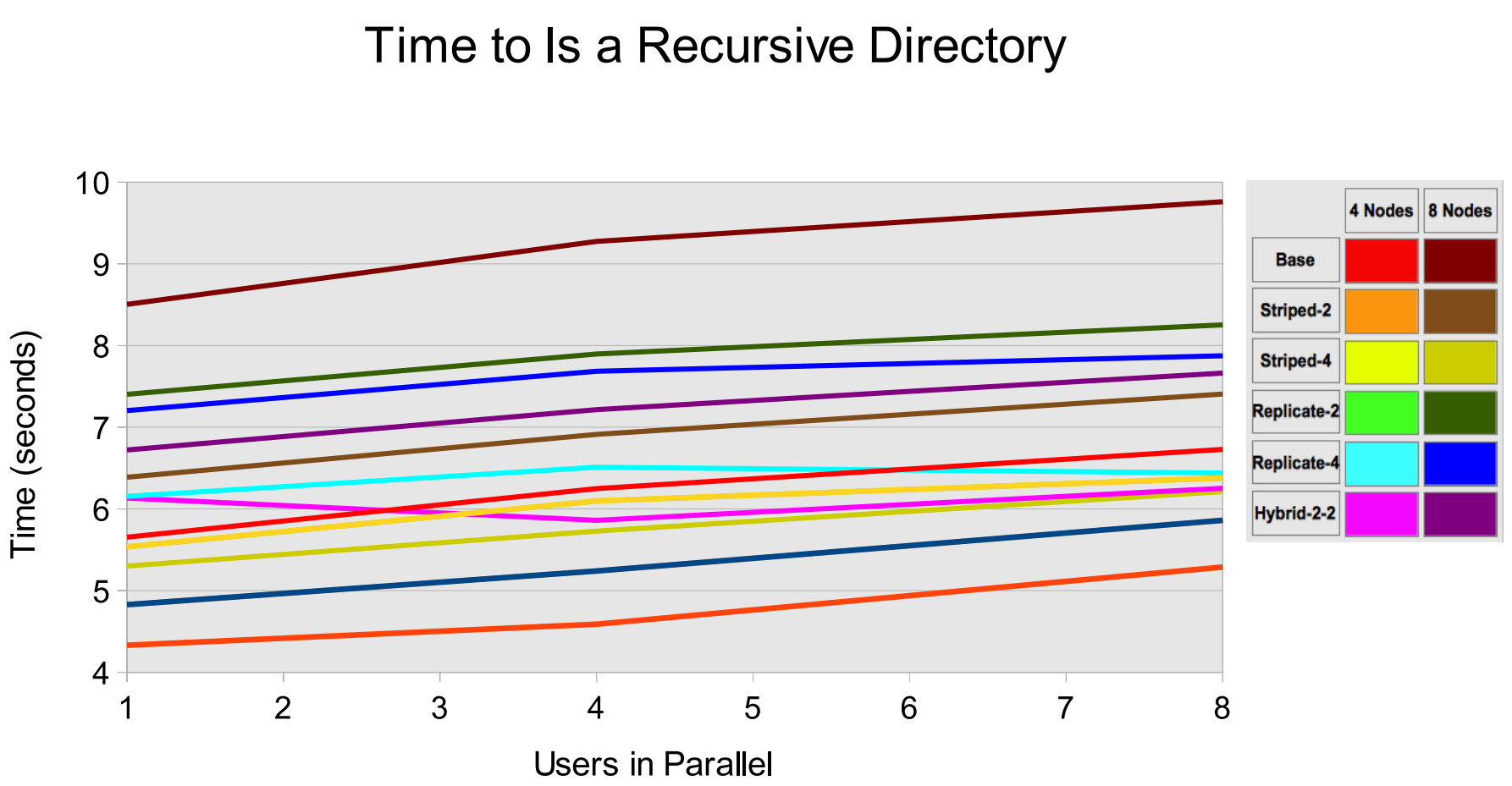
Read Performance



Read Performance

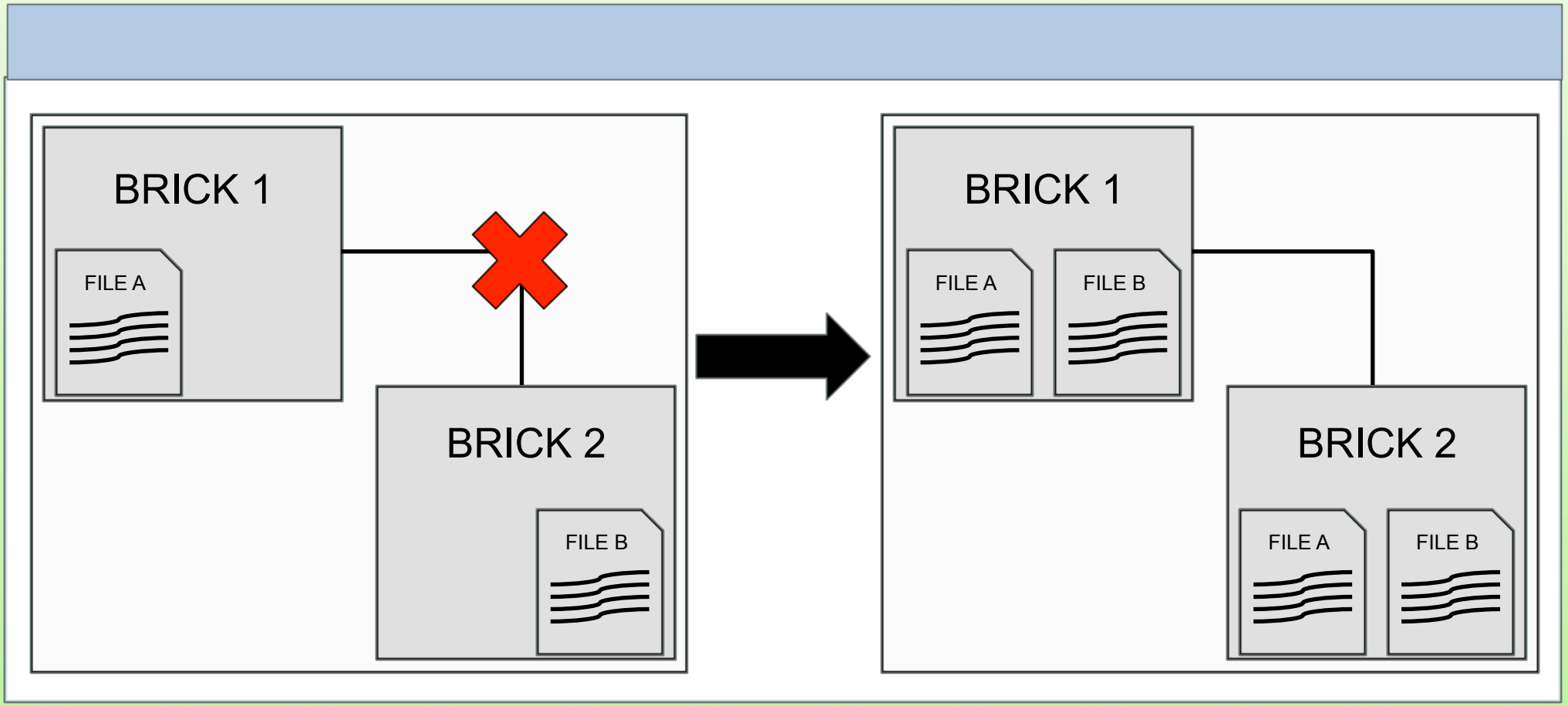


Ls Performance



Fault Tolerance

- Replicated volumes can self-heal



Conclusions

- GlusterFS proved to have widespread capabilities as a virtual file system
- Scalability is very dependent upon the underlying hardware
- Lack of built-in encryption and security paradigm
- Best suited in a general purpose computing environment

LA-UR-12-23586

Future Research

- GlusterFS over Infiniband
- Geo-replication
- Unified File and Object Storage
- Apache Hadoop
- Scalability for 1000's of nodes
- Using other filesystems on top of GlusterFS
- Testing different RAID types

Acknowledgments

- We would like to thank:
  - Los Alamos National Laboratory
  - New Mexico Consortium/PRObE
  - National Science Foundation
  - National Nuclear Security Administration
  - Gary Grider and the HPC division
  - Carol Hogsett and Josephine Olivas
  - Our mentors, David Kennel and Greg Lee
  - Our instructors, Dane Gardner and Andree Jacobson

LA-UR-12-23586

Questions??