A Comparative Experimental Study of Parallel File Systems for Large-Scale Data Processing

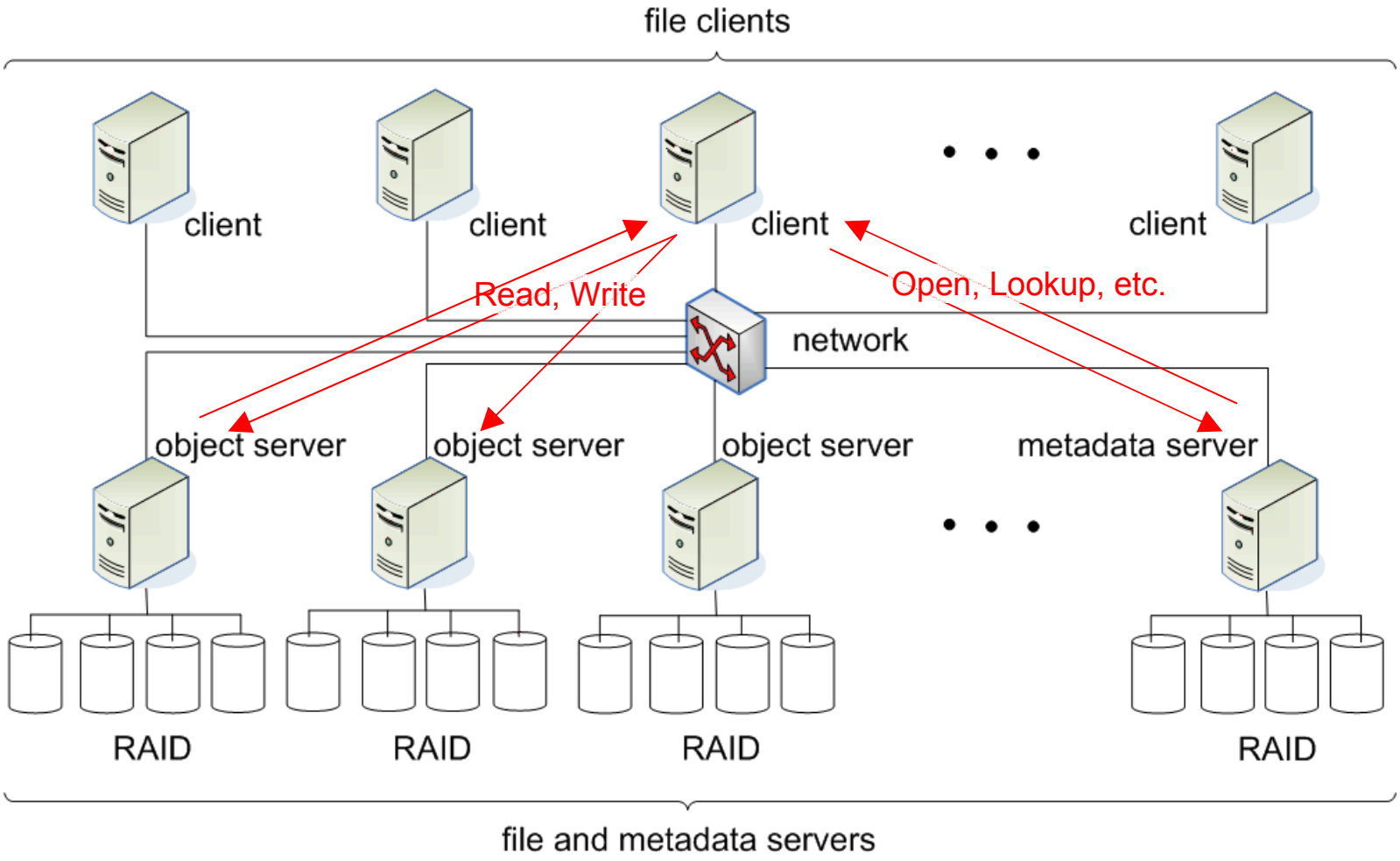
Z. Sebepou, K. Magoutis, M. Marazakis, A. Bilas

Institute of Computer Science (ICS)
Foundation for Research and Technology Hellas (FORTH)
Heraklion, Crete, Greece

Evolution

- Distributed file systems
  - NFS versions 2, 3, 4; AFS; *etc.*
- Shared-disk parallel file systems
  - Frangipani/Petal; GPFS; GFS; *etc.*
- Separating data/metadata paths to object storage
  - NASD; pNFS; Panassas; Lustre; PVFS; *etc.*

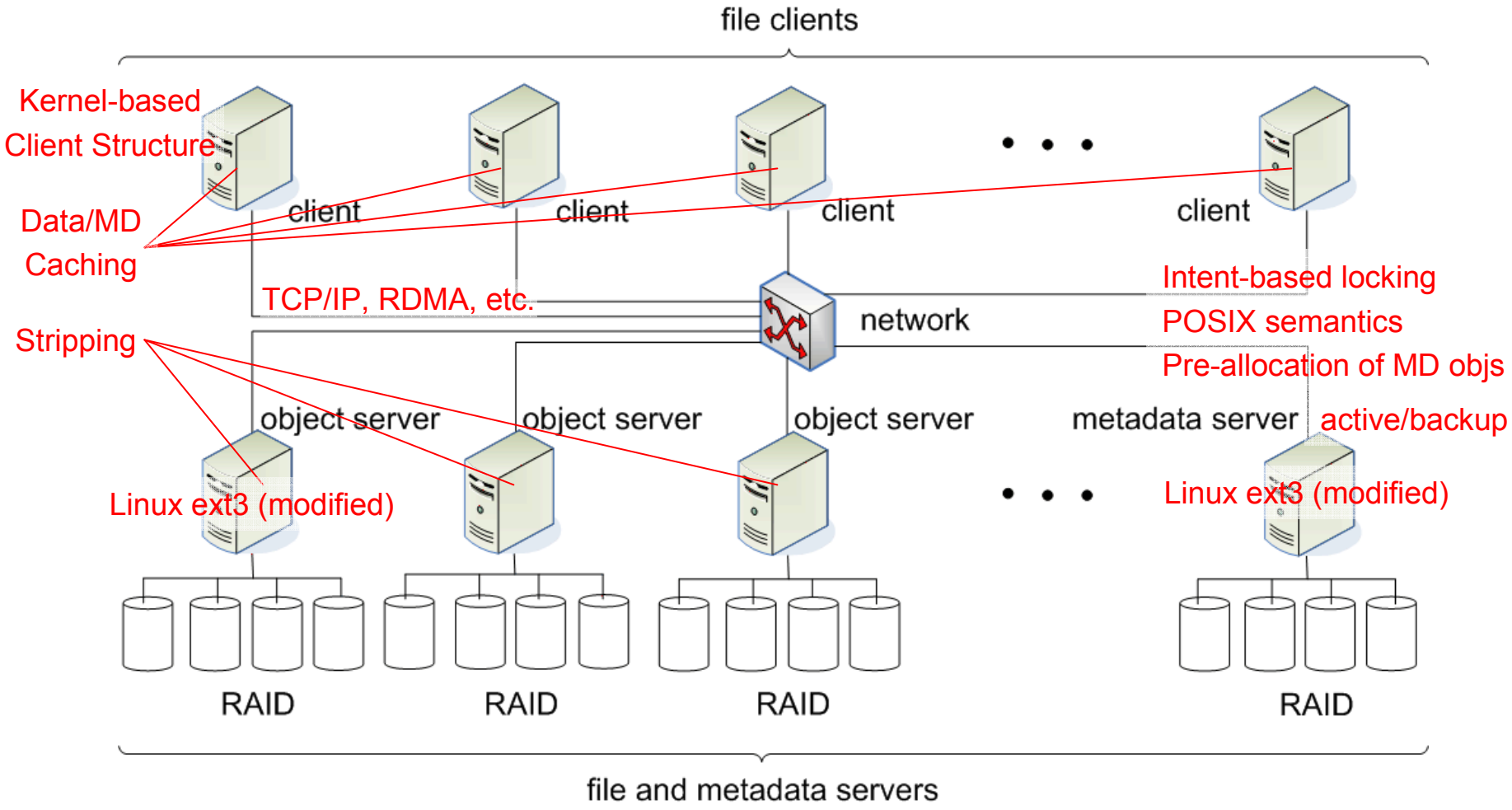
NASD-style Parallel File Systems



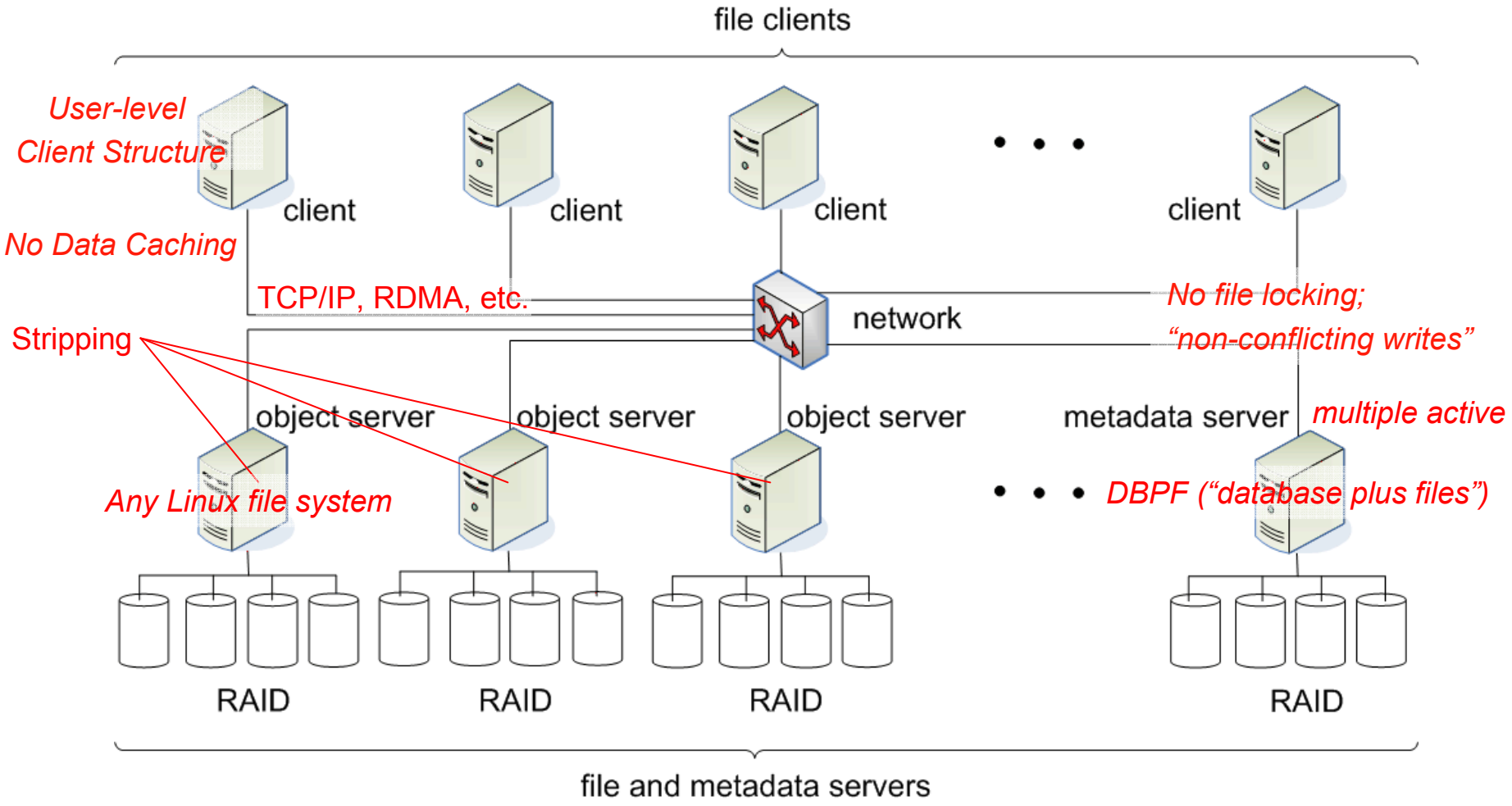
Lustre, PVFS

- Open-source systems, following the NASD paradigm
  - Lustre: Cluster File Systems, Inc.; acquired (2007) by Sun
  - PVFS: Clemson University, ANL, OSC
- Targeted for large-scale data processing
  - LLNL, ORNL, ANL, CERN
- Representative of different approaches to filesystem design
  - Client caching
  - Statelessness
  - Consistency and file access semantics
  - Portability

Lustre Architecture



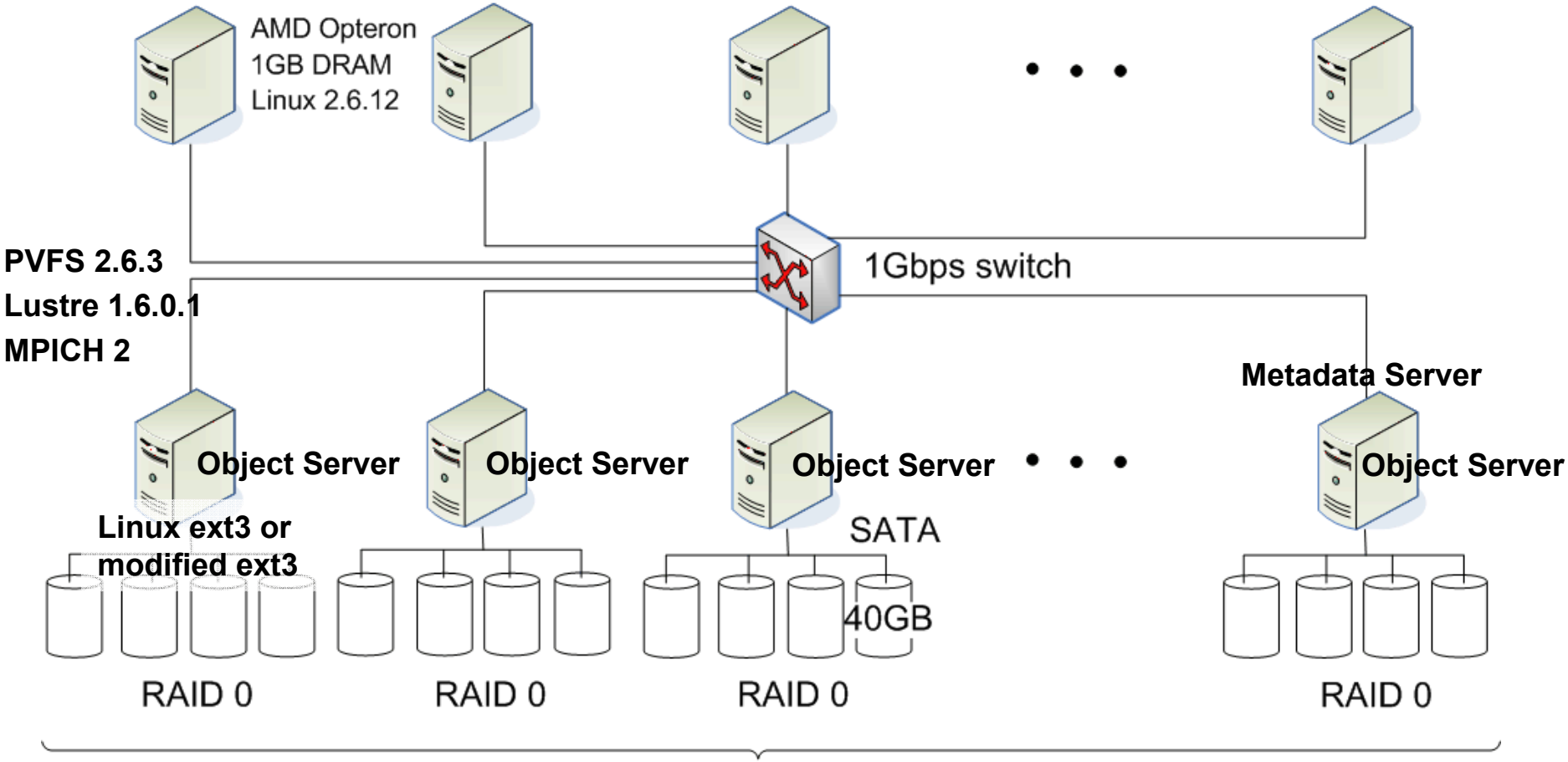
PVFS2 Architecture



Benchmarks

- Streaming; one client, many servers
  - *IOZone*
- Streaming: many clients, many servers
  - *Parallel I/O (MPI)*
- Metadata-intensive
  - *Cluster PostMark*
- Near-random I/O, optionally with data overlap
  - *Tile I/O (MPI)*
- User-perceived response time
  - `ls -lR` *on Linux kernel tree*

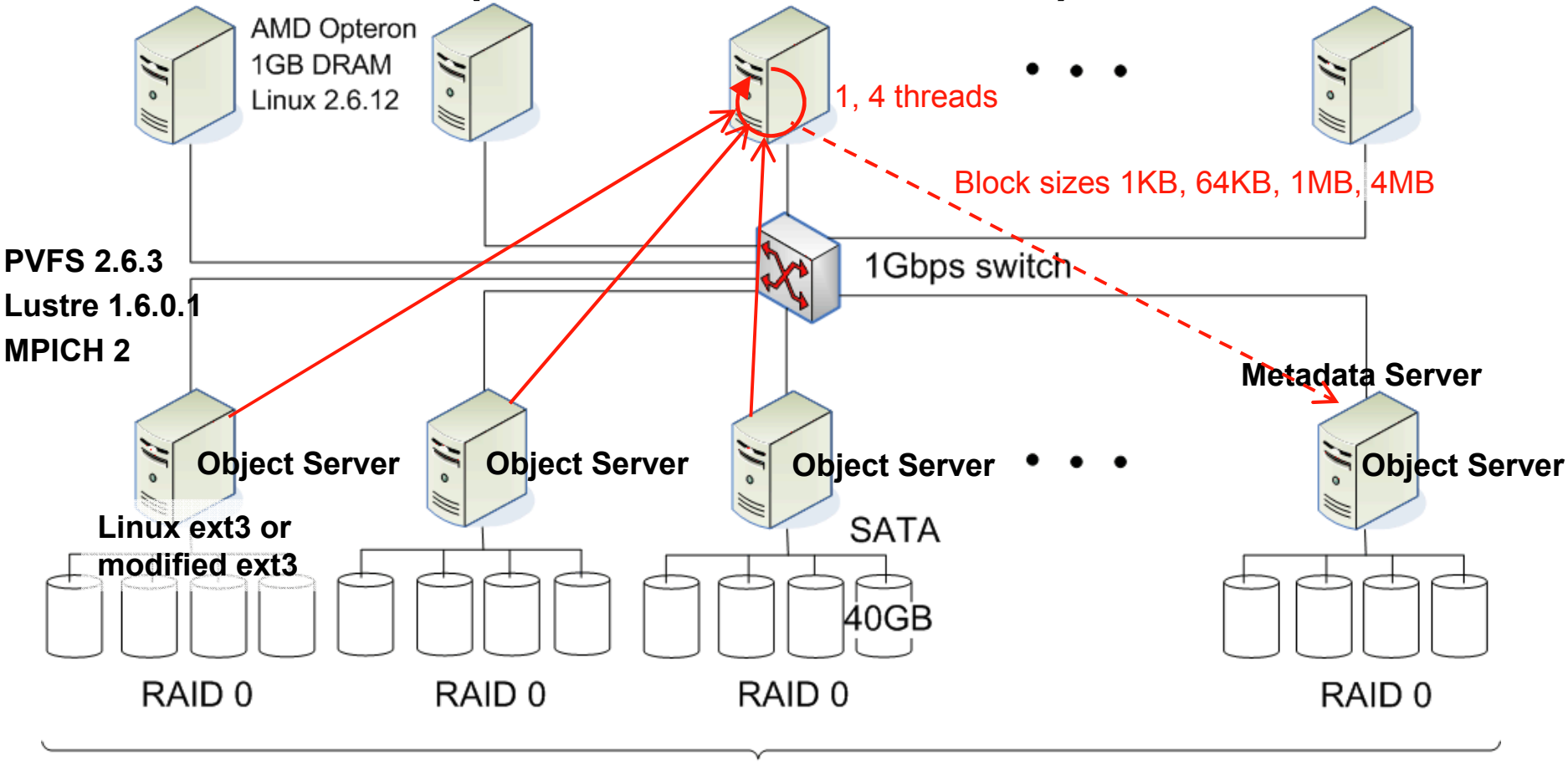
Experimental Testbed



PVFS2 stripe: 64KB (default)

Lustre stripe: 64KB, 256KB, 1MB (default)

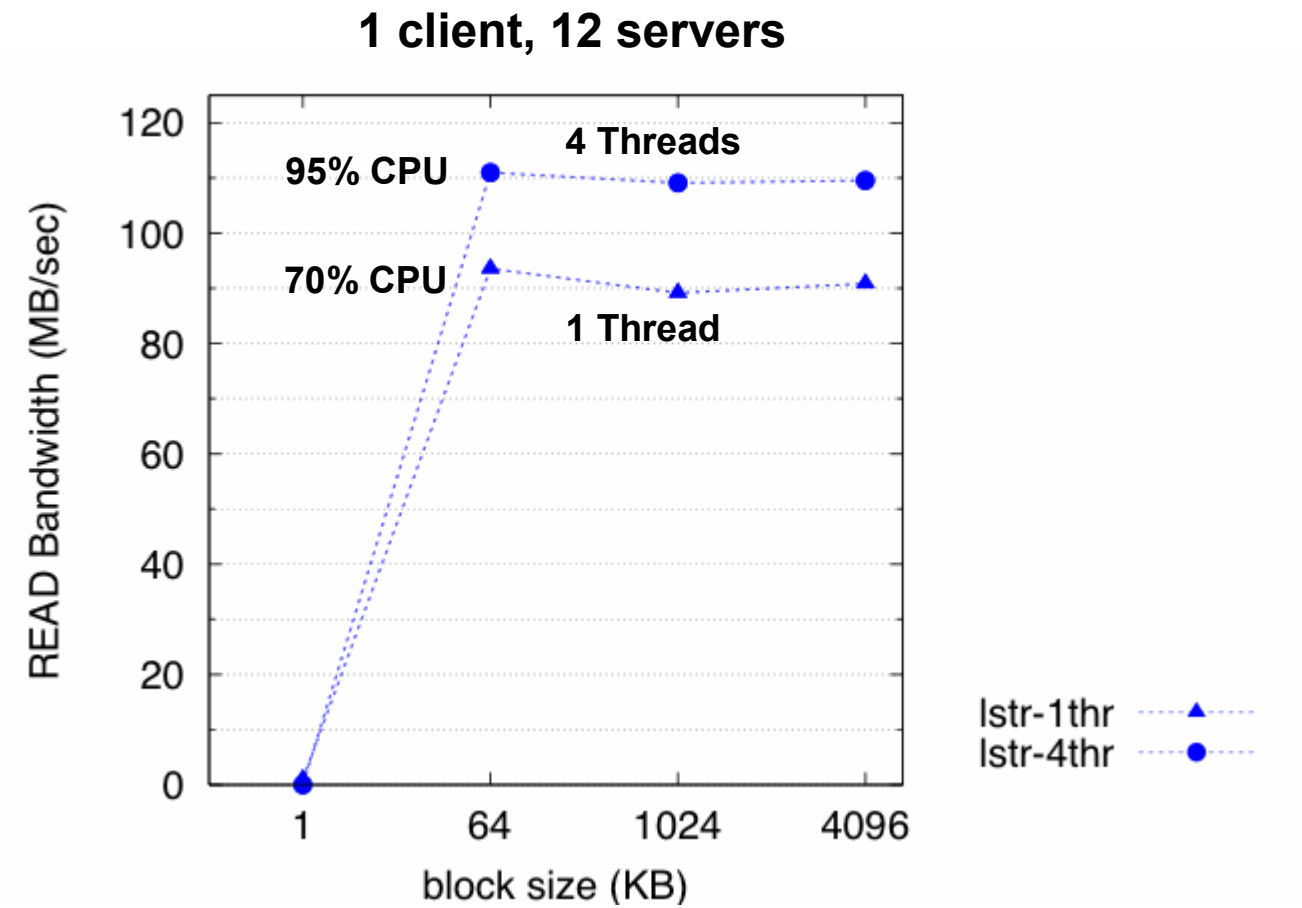
Streaming: One Client, Many Servers (IOZone Benchmark)



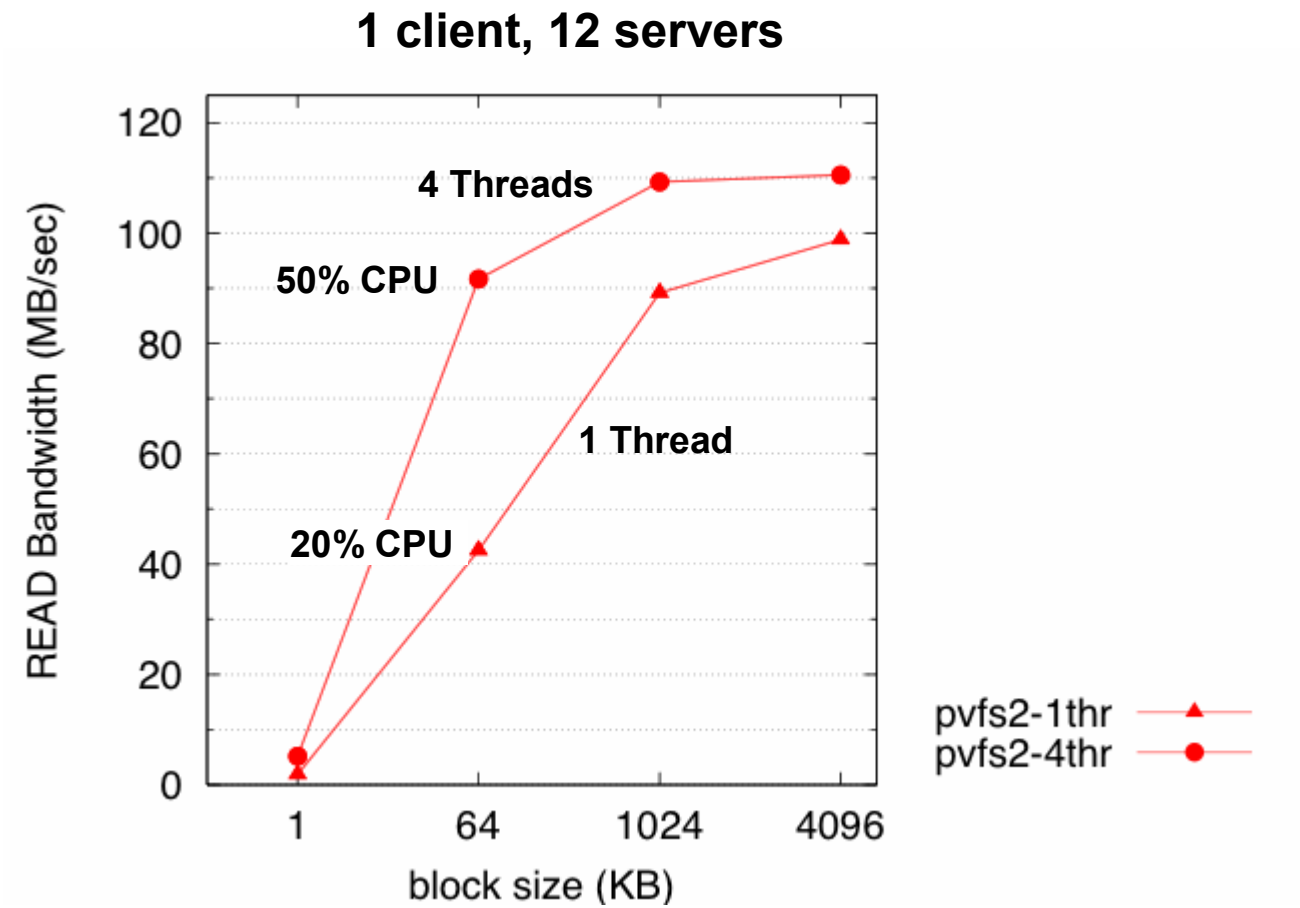
PVFS2 stripe: 64KB (default)

Lustre stripe: 64KB, 256KB, 1MB (default)

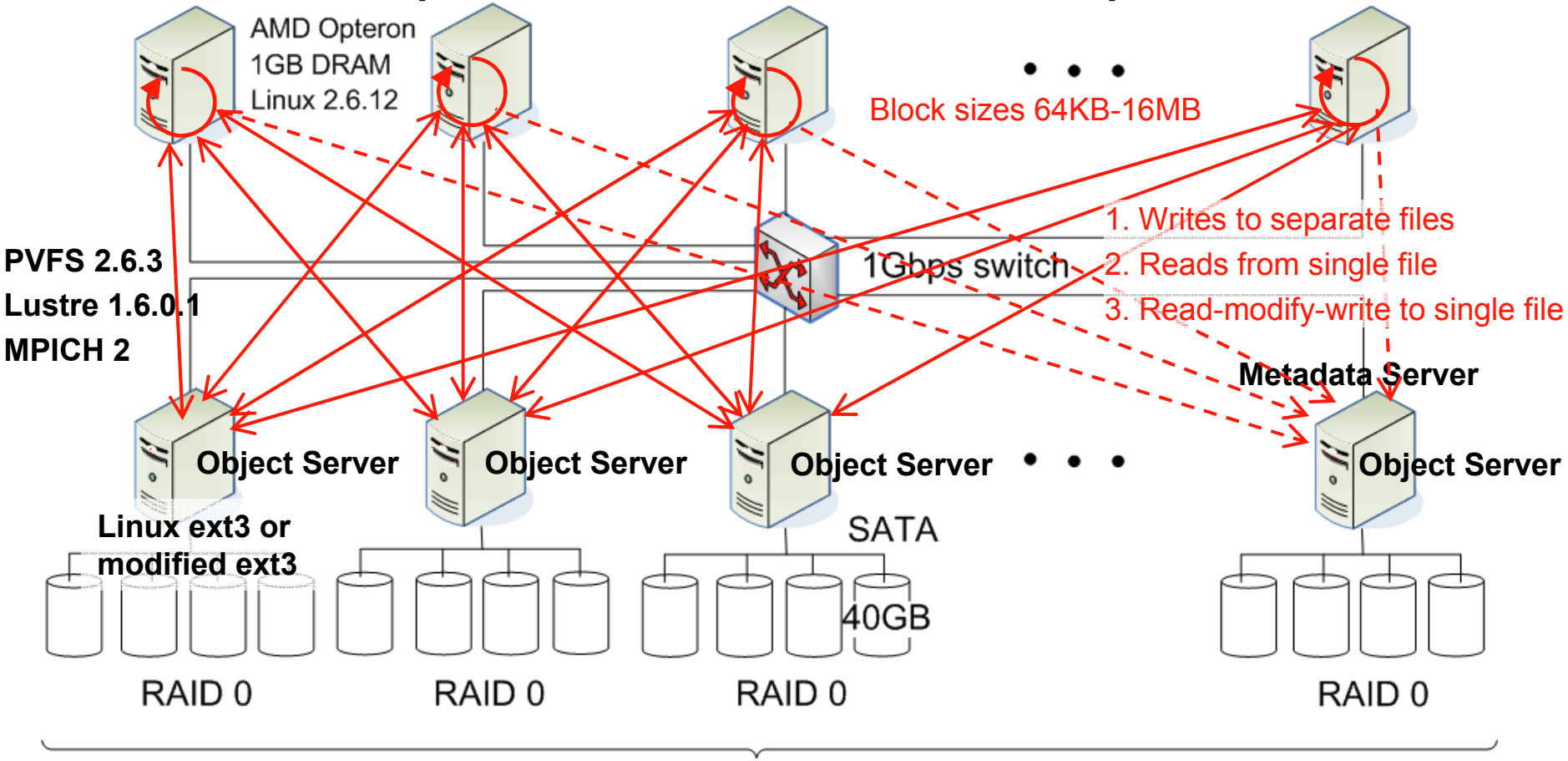
Streaming: One Client, Many Servers (IOZone Benchmark) – Lustre



Streaming: One Client, Many Servers (IOZone Benchmark) – PVFS



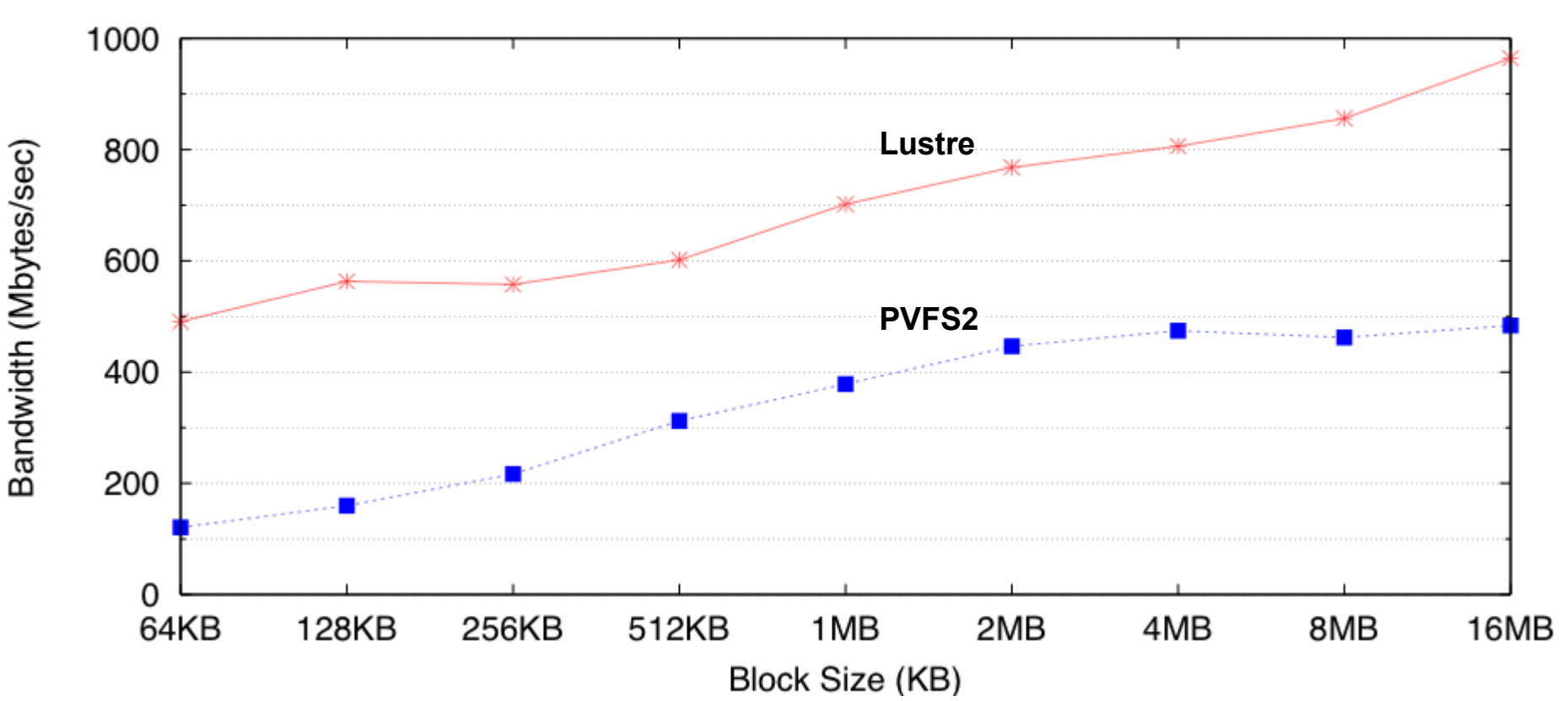
Streaming: Many Clients, Many Servers (Parallel I/O Benchmark)



PVFS2 stripe: 64KB (default)

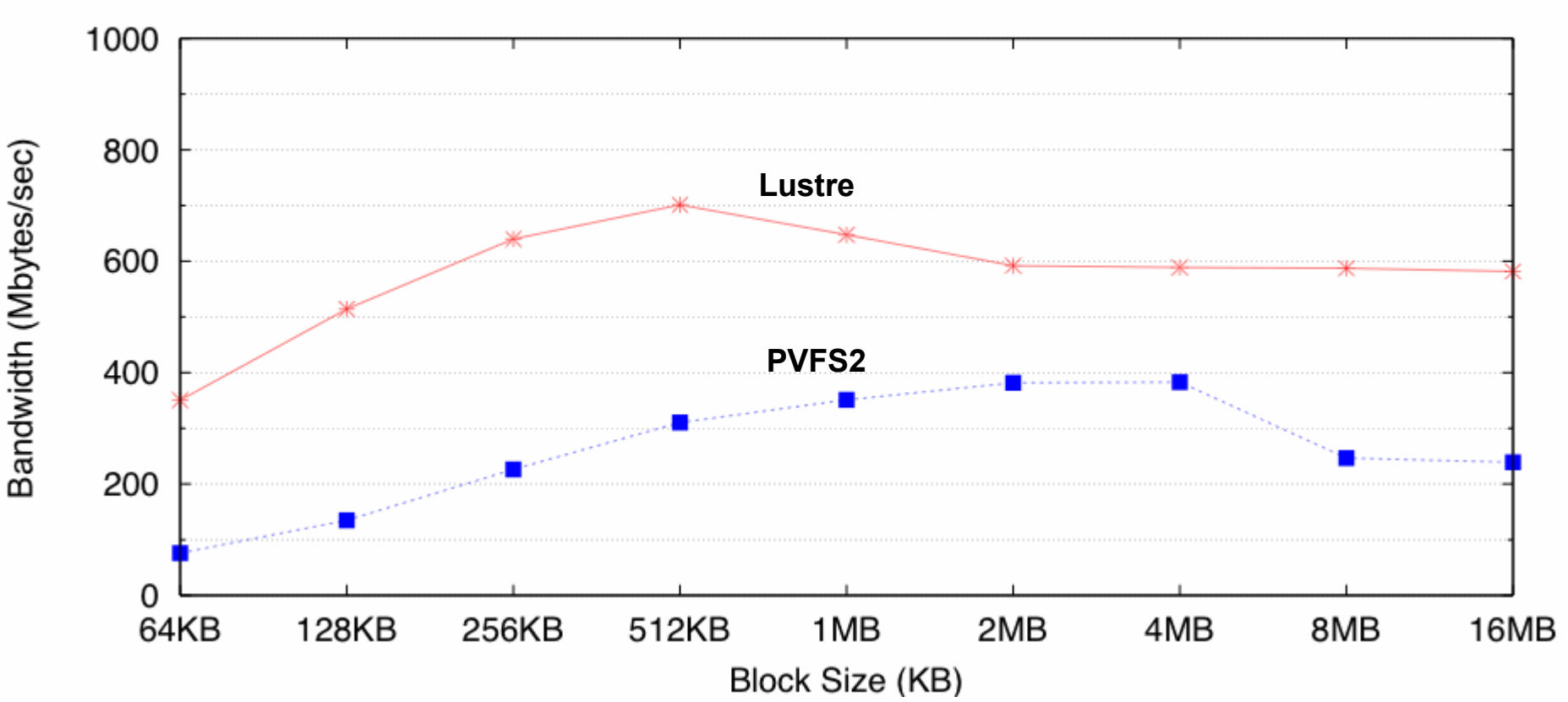
Lustre stripe: 64KB, 256KB, 1MB (default)

Streaming: Many Clients, Many Servers (Parallel I/O Benchmark)



Streaming: Many Clients, Many Servers (Parallel I/O Benchmark)

Reads from Single File; 12 clients, 12 servers



Cluster PostMark

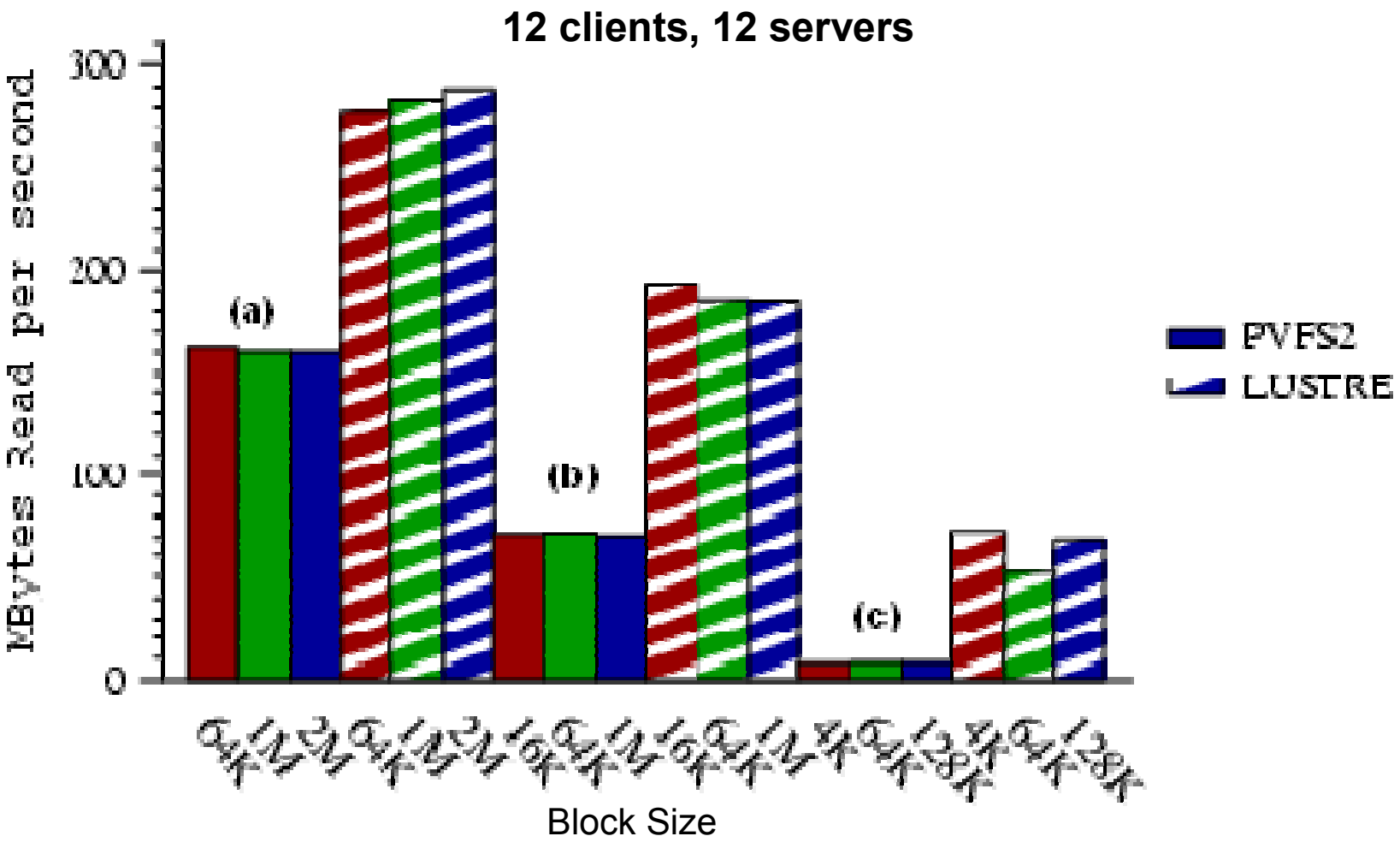
- Three configurations:

  (a) Few (100), large (10-100MB) files

  (b) Moderate number (800) of medium-size (1-10MB) files

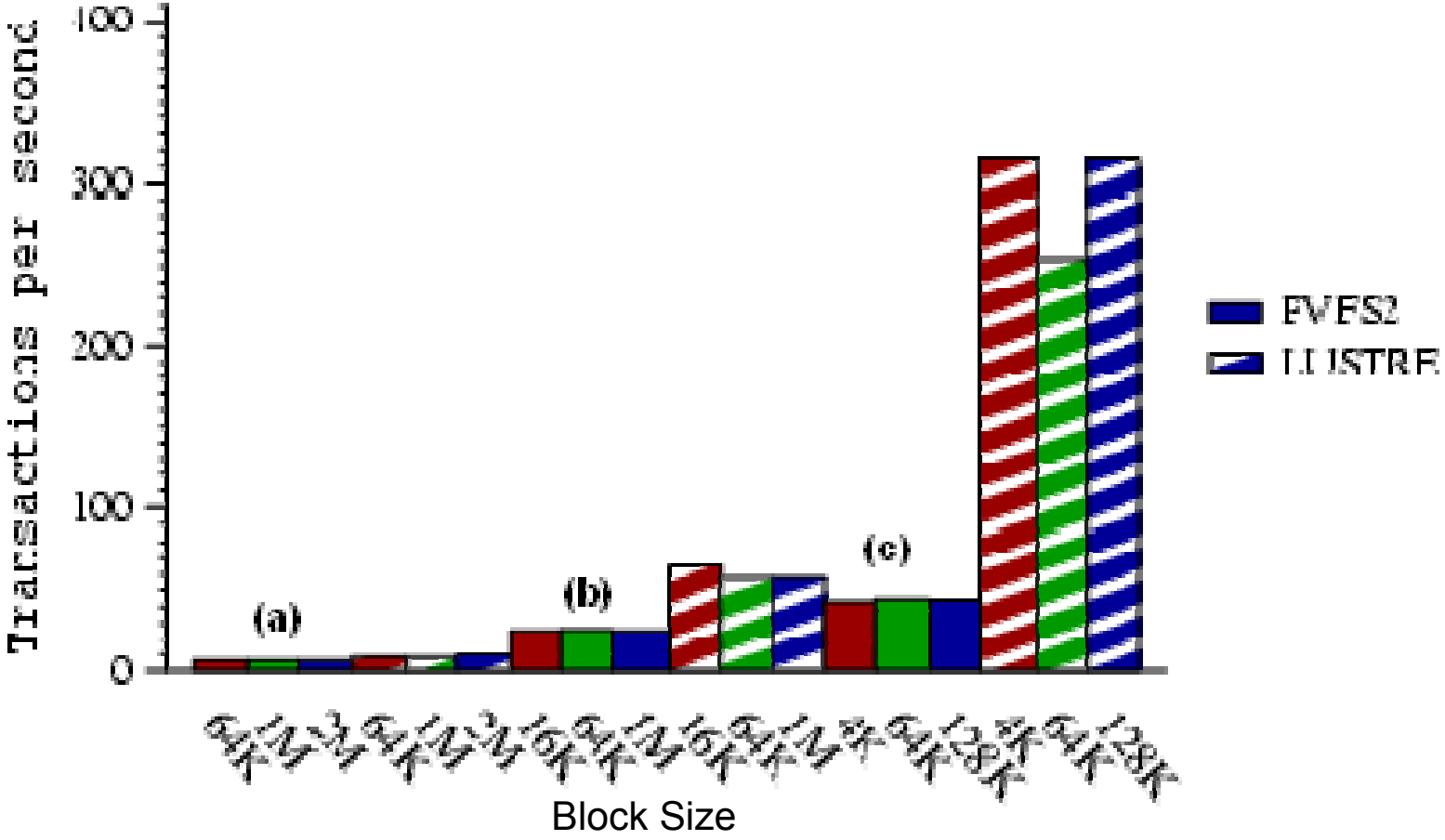
  (c) Many (8000), small (4-128KB) files

Cluster PostMark – Bandwidth



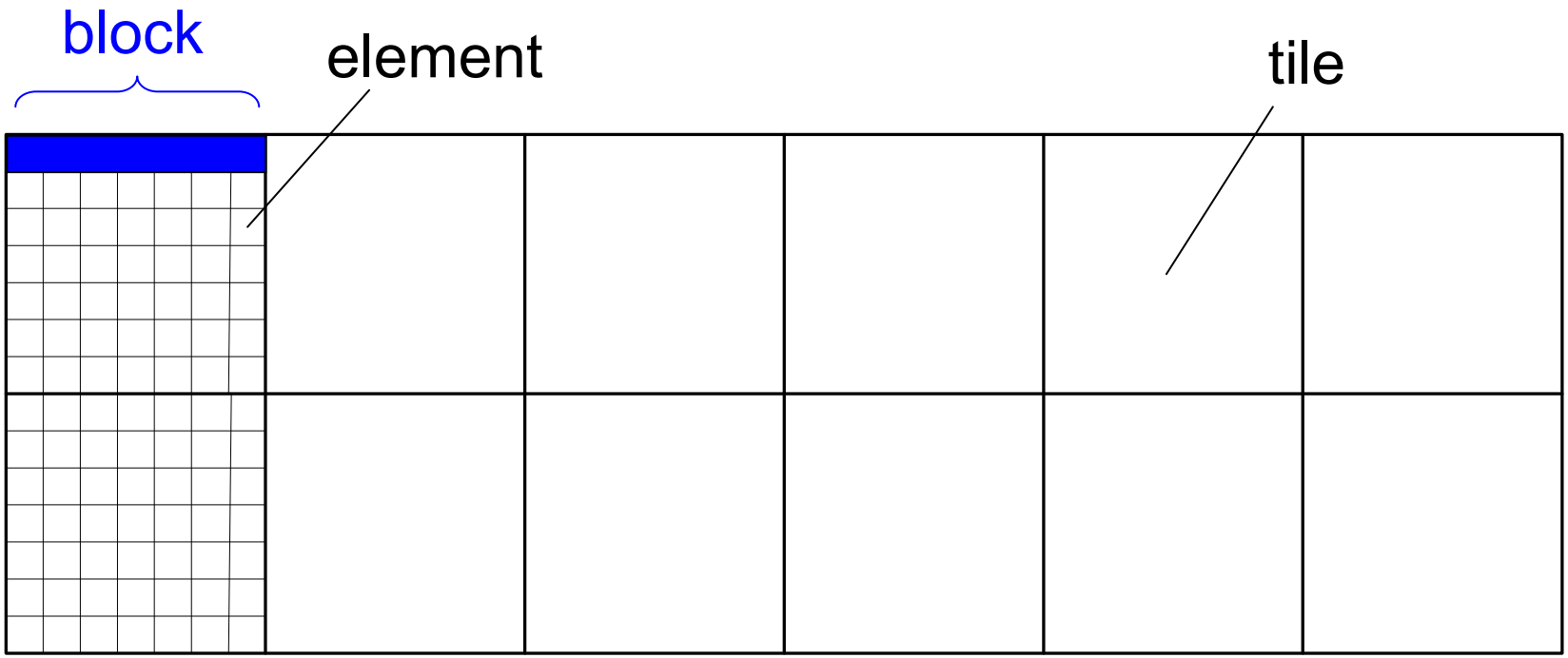
Cluster PostMark – Transactions

12 clients, 12 servers

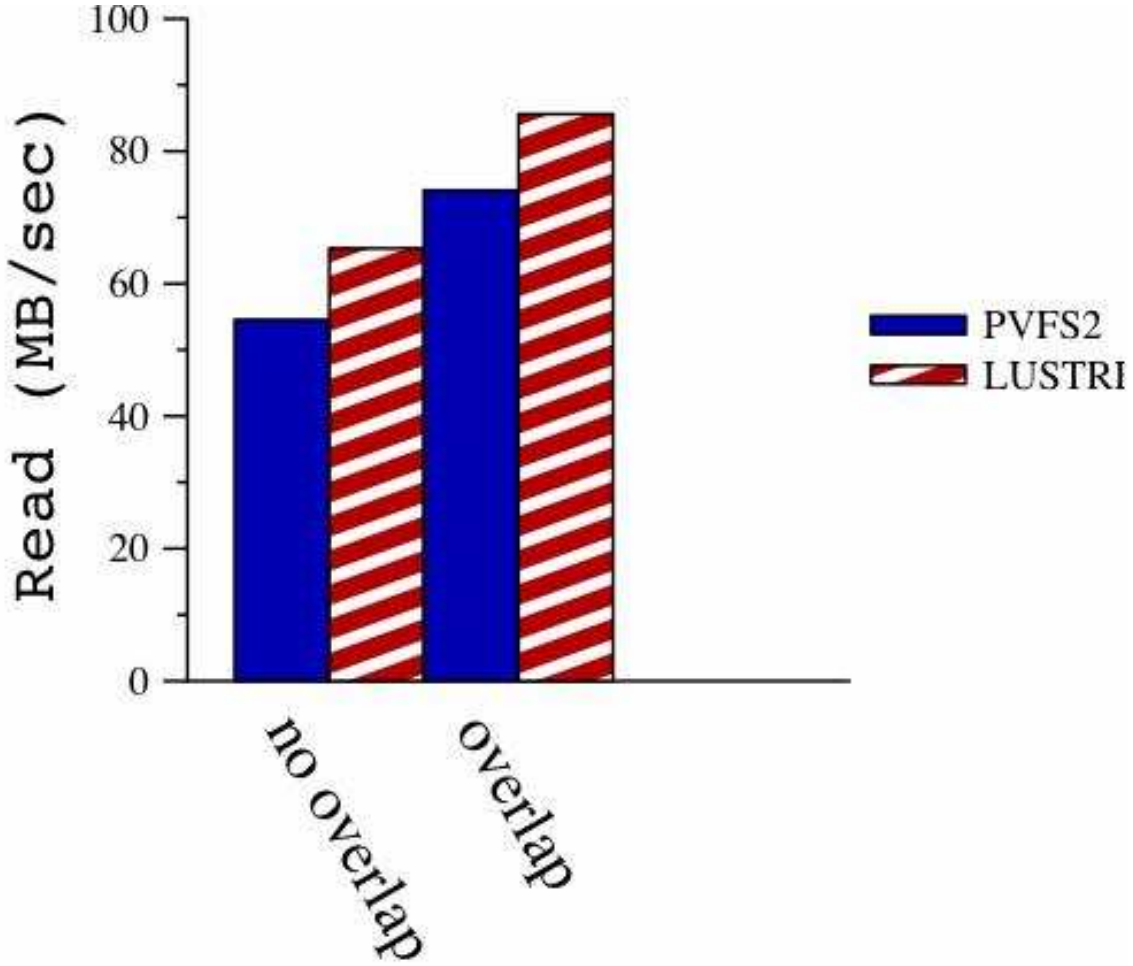


Near-random I/O, optionally with data overlap (Tile I/O)

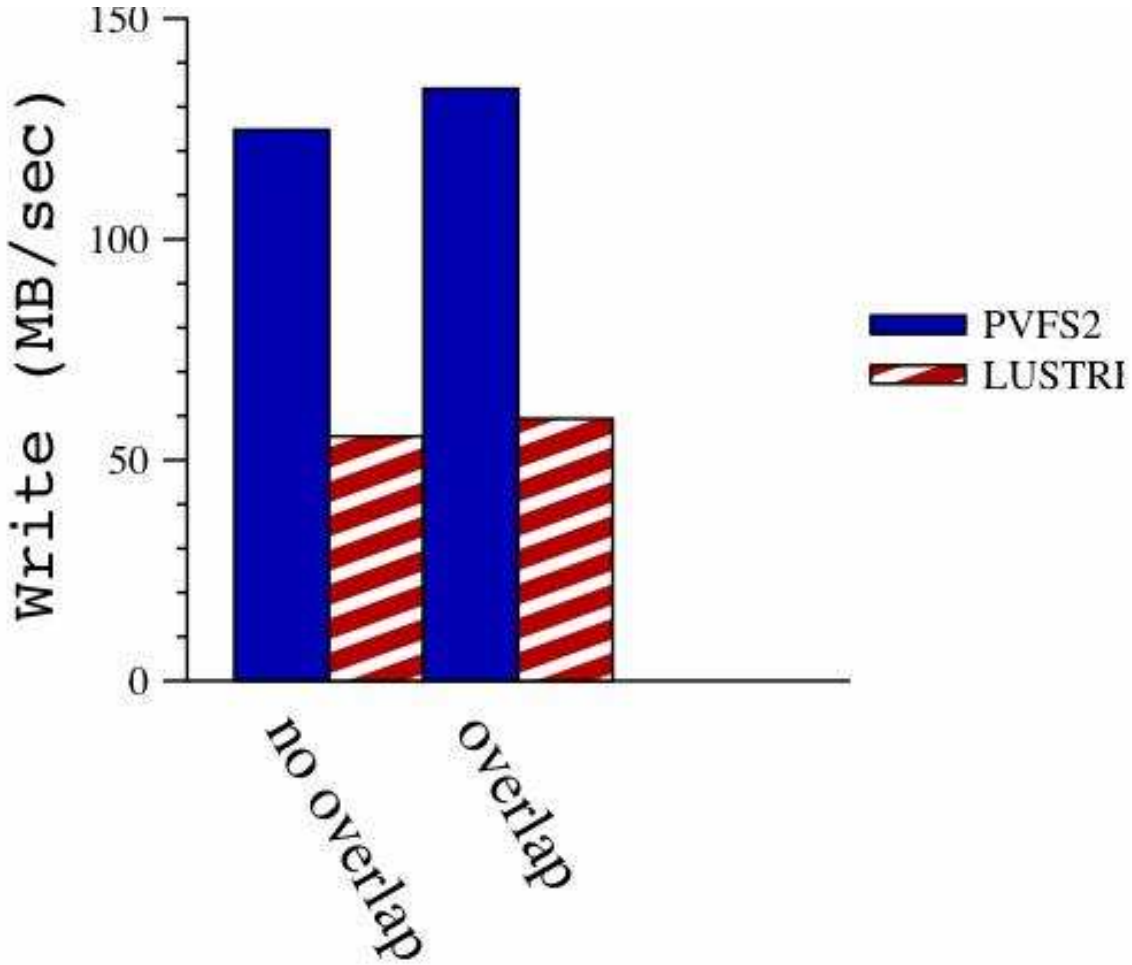
- two-dimensional logical structure overlaid on a single file
- each tile assigned to a separate client process



Tile I/O - Reads



Tile I/O - Writes



User-Perceived Response Time

`ls -lR` on Linux 2.6.12 kernel tree (~25,000 files)

File System	Response Time (sec)
Linux ext3 (local)	5.5
Lustre	58
PVFS2	80

Conclusions

- Scalable I/O bandwidth is achievable through parallel I/O paths to file servers
- Lustre's efficient metadata management is critical for metadata-intensive applications
- Lustre's consistency semantics are useful to some applications but cause unnecessary overhead to others that do not require them

Thank You!