
An Overview of Scalable Storage Virtualization and File System Research at FORTH-ICS

Kostas Magoutis

FORTH-ICS

magoutis@ics.forth.gr

TF-TERENA Meeting – Riga, September 12 2008



FORTH-ICS

- FORTH: Foundation for Research and Technology –
 - Largest research center in Greece
 - About 1000 employees
 - 25 year history
 - Located in Heraklion, Crete
- ICS: Institute of Computer Science
 - Largest of seven institutes under FORTH: 300+ personnel
 - Our Lab (CARV): computer architecture and computer systems, about 30 personnel
 - We are in the Computer Systems part of CARV
 - Expertise: communication protocols, storage systems, parallelism
 - Tools of trade: low-level systems software (kernel, controllers), simulation
 - Extensive prototyping infrastructure and expertise
- Unique environment
 - Research facilities & expertise
 - Transportation (Airport 20mins, 2nd in international traffic in Greece)
 - Supporting infrastructure for conferences, events, etc.



<http://www.ics.forth.gr/carv/scalable>



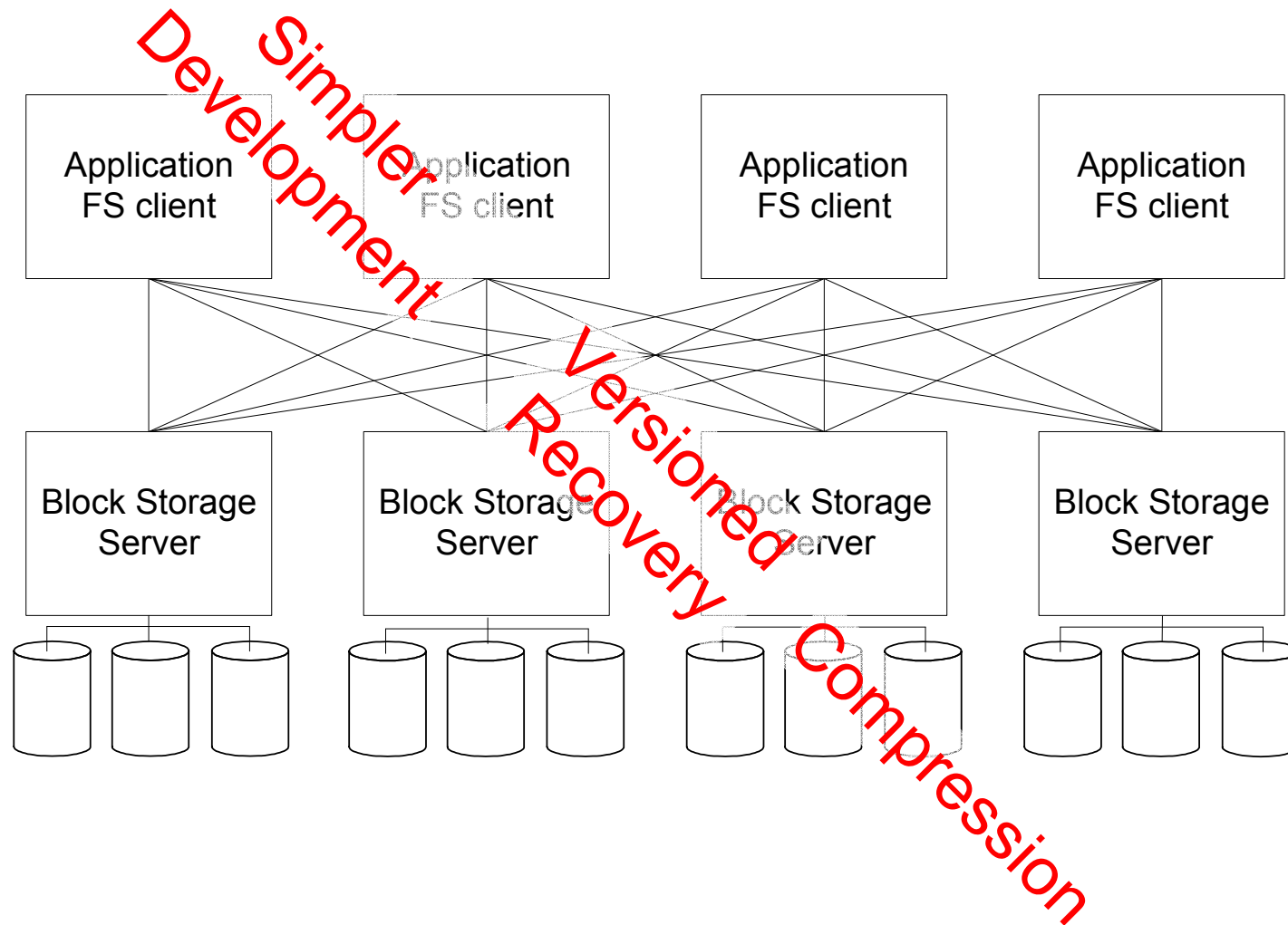
FORTH-ICS

CARV/scalable

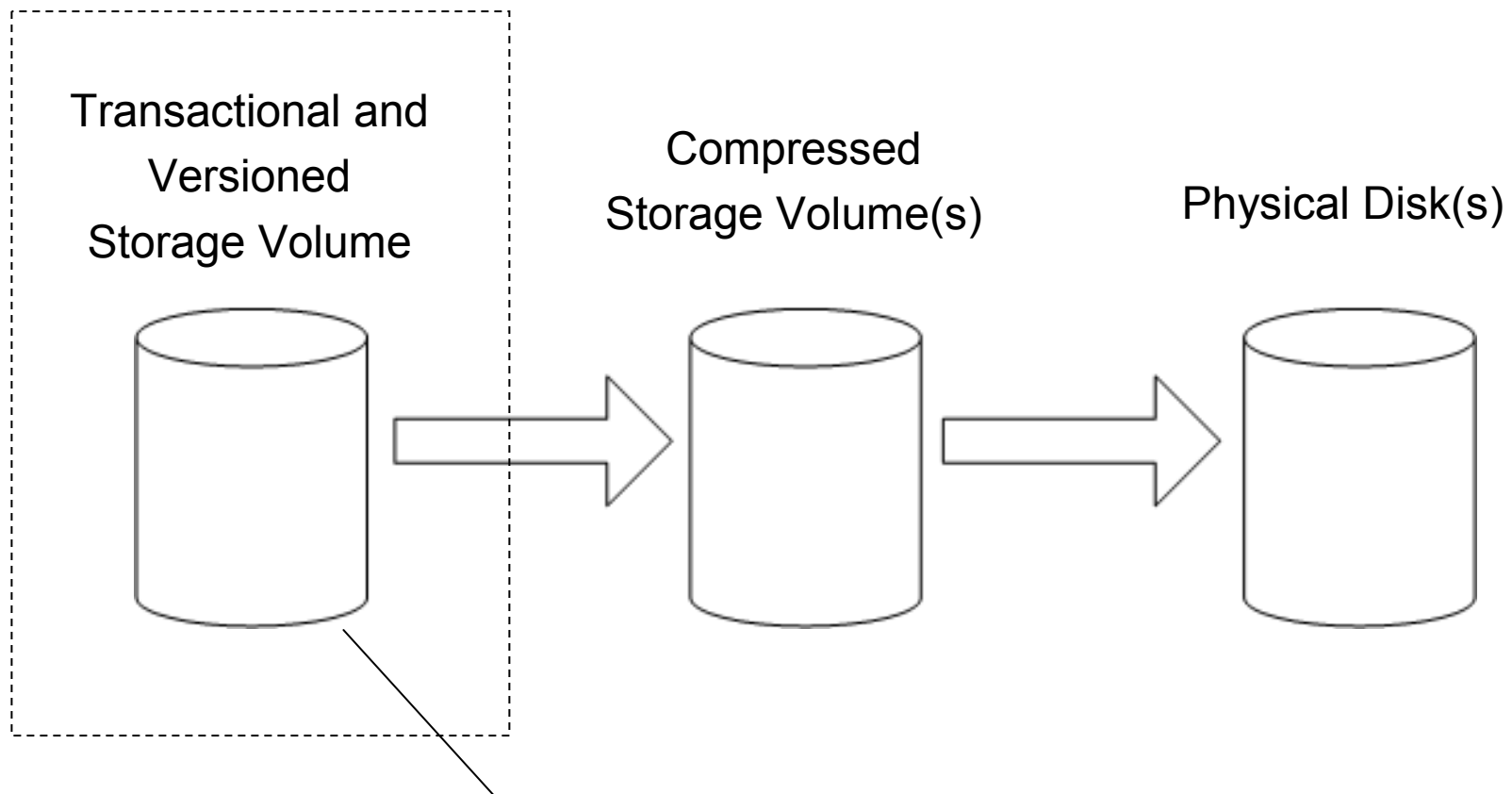
Scalable Storage Research at FORTH-ICS

- Block-level storage virtualization
 - Lightweight transactions
 - Versioning
 - Compression
- File systems
 - Comparative experimental study of parallel file systems
 - Efficient file transfer at 40Gbps / node

Scalable Block-level Storage Virtualization

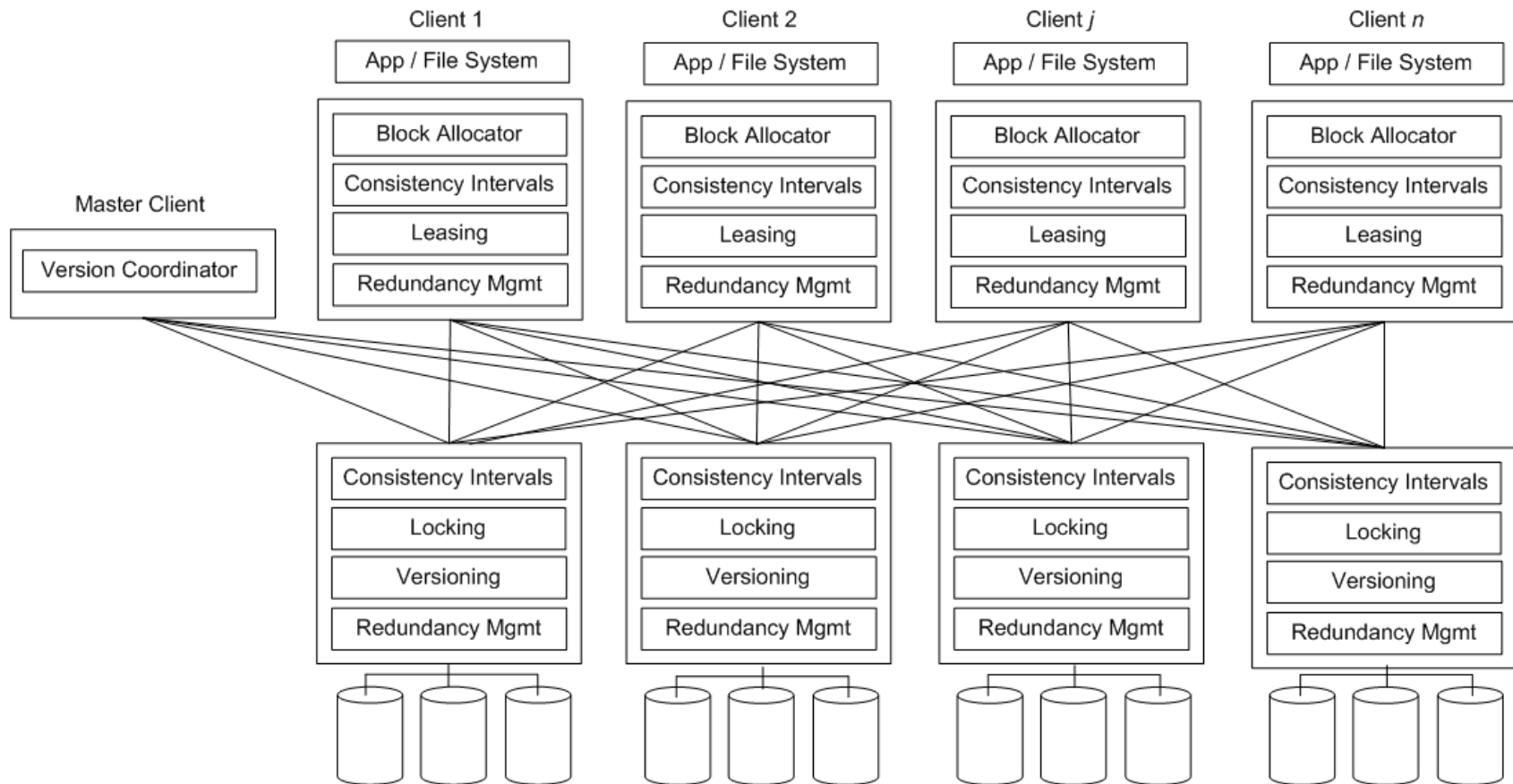


Storage Virtualization Mappings

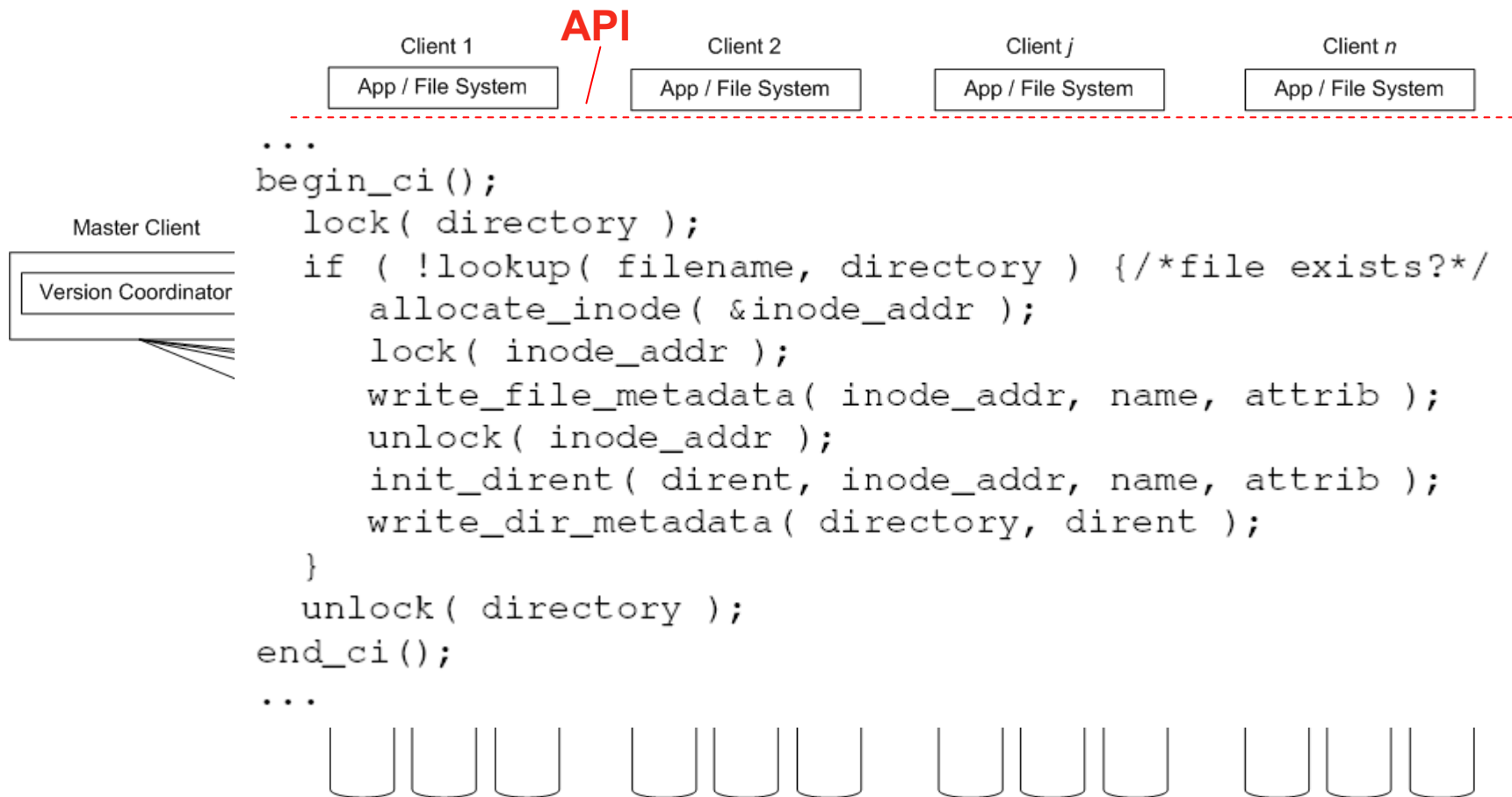


Recoverable Independent Block Devices (RIBD)

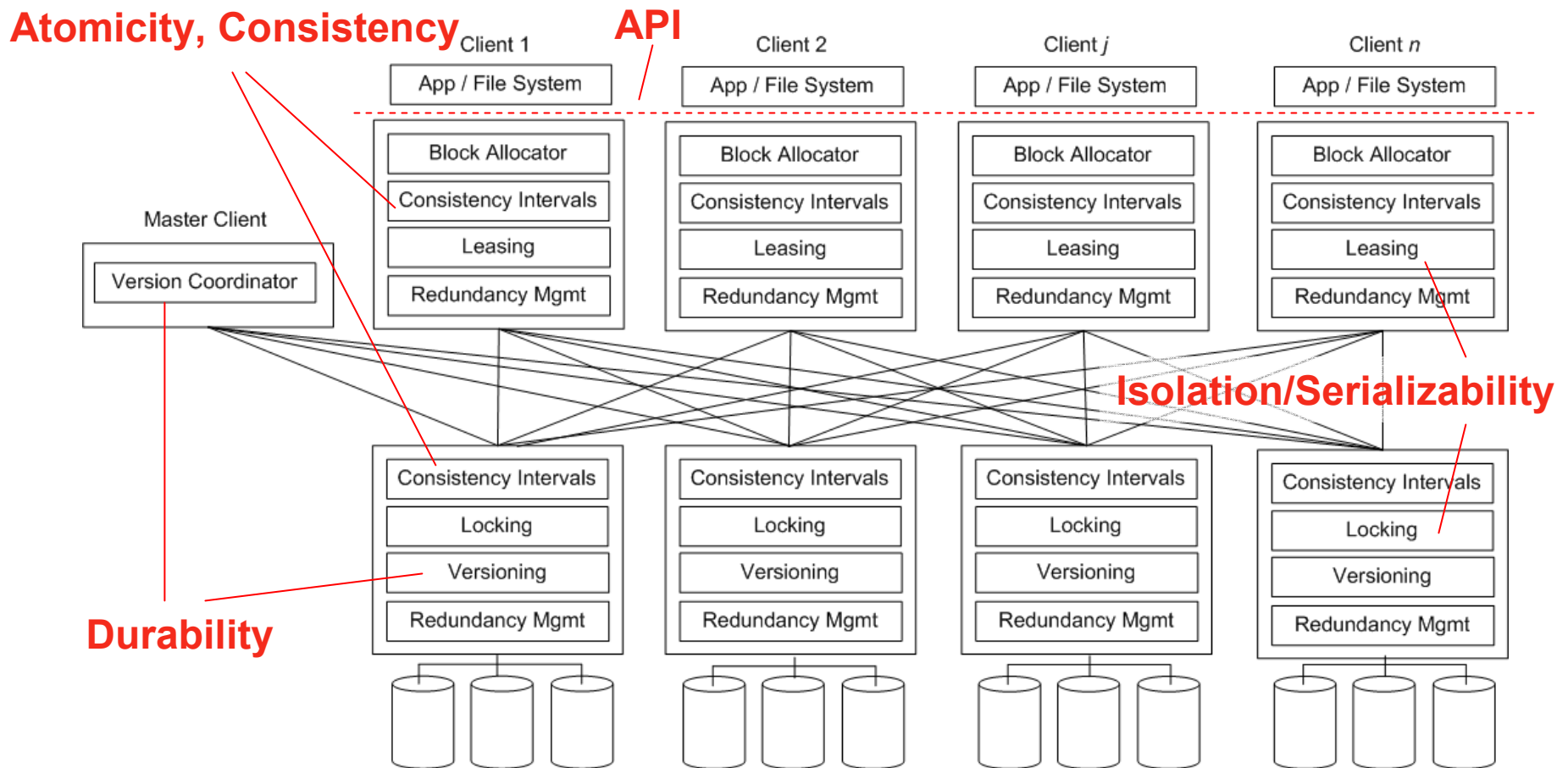
Recoverable Independent Block Devices (RIBD)



Recoverable Independent Block Devices (RIBD)



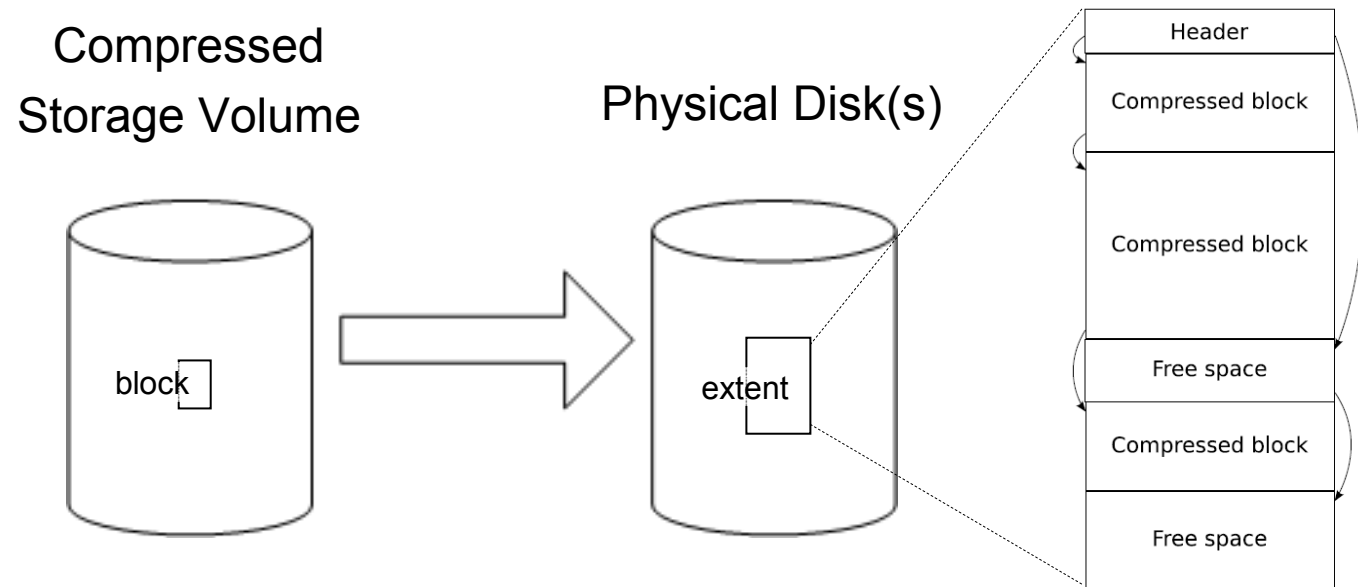
Recoverable Independent Block Devices (RIBD)



RIBD Evaluation

- Evaluated scalable file system (*OFS*) over RIBD
 - Simple stateless pass-through file system
- Compared to GFS, PVFS2 on range of benchmarks
 - IOZone, Clustered PostMark
- *OFS*/RIBD perform comparably to GFS, PVFS2 while providing stronger consistency guarantees

Block-level Compression Mapping



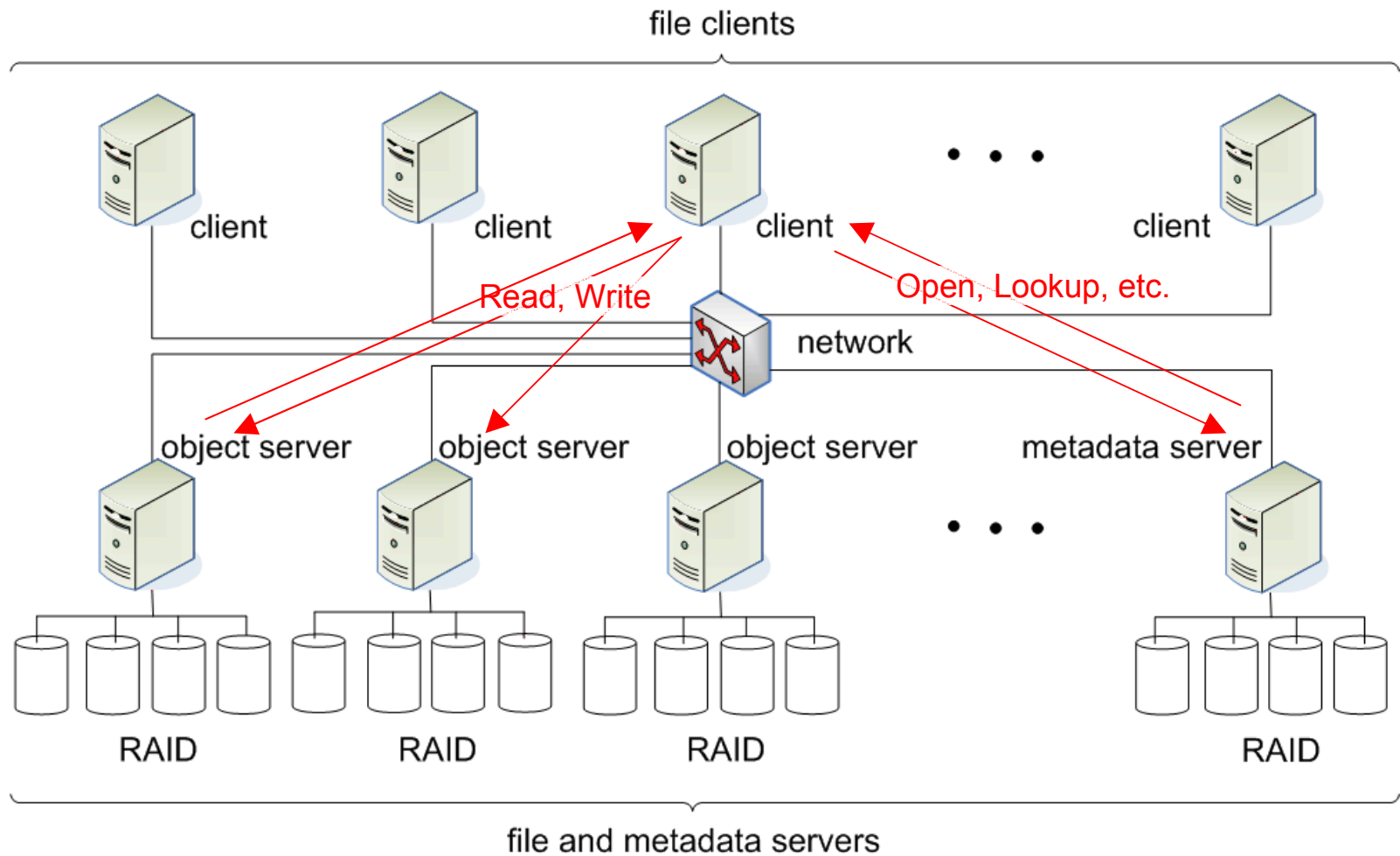
Design issues

- fixed-size blocks transformed into variable-size blocks by compression
- variable-size blocks packed into fixed-size extent on underlying physical disk
- data placement into extents and extent placement onto disks key for performance

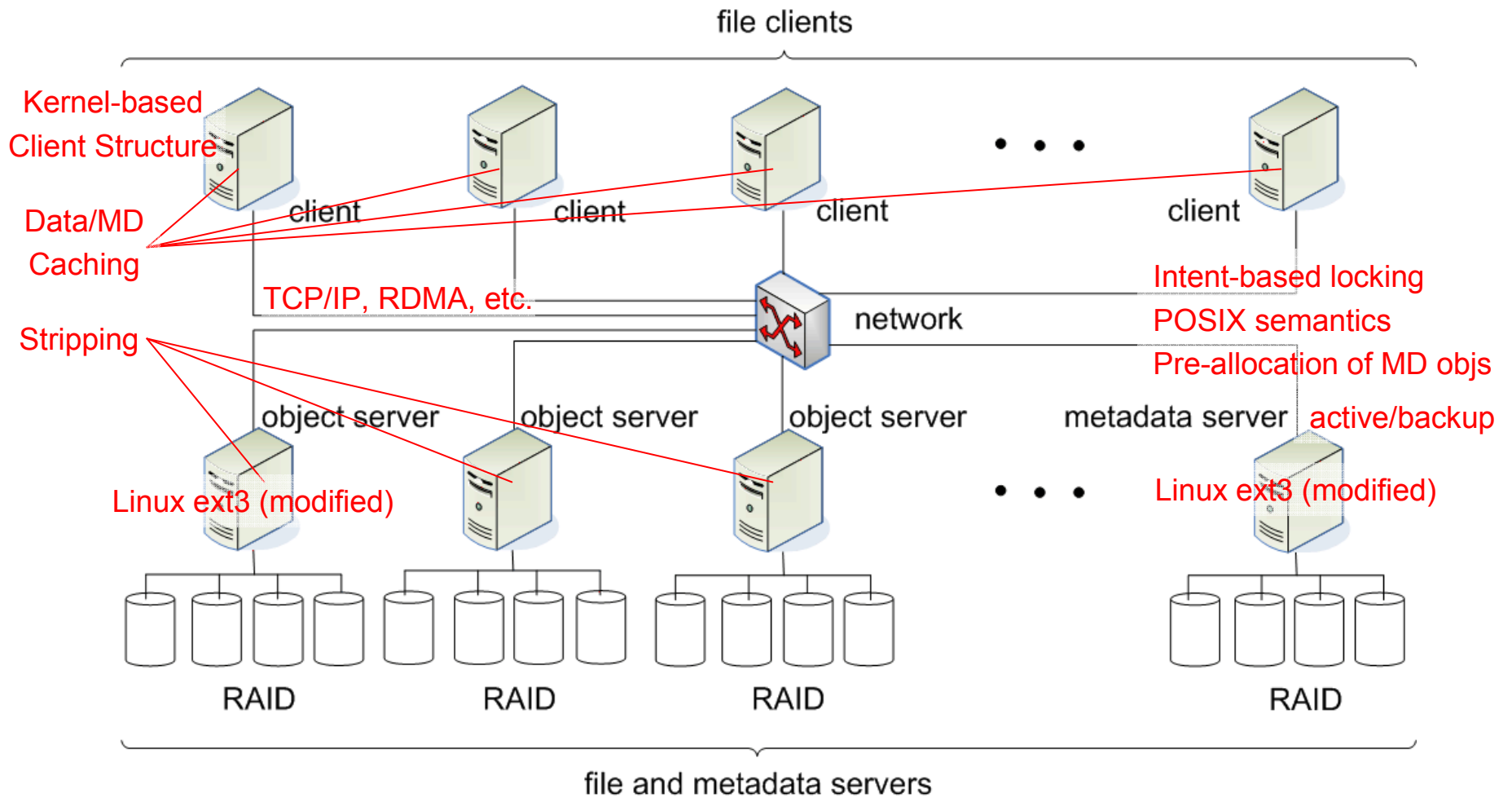
Comparative Experimental Study of PVFS, Lustre

- 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

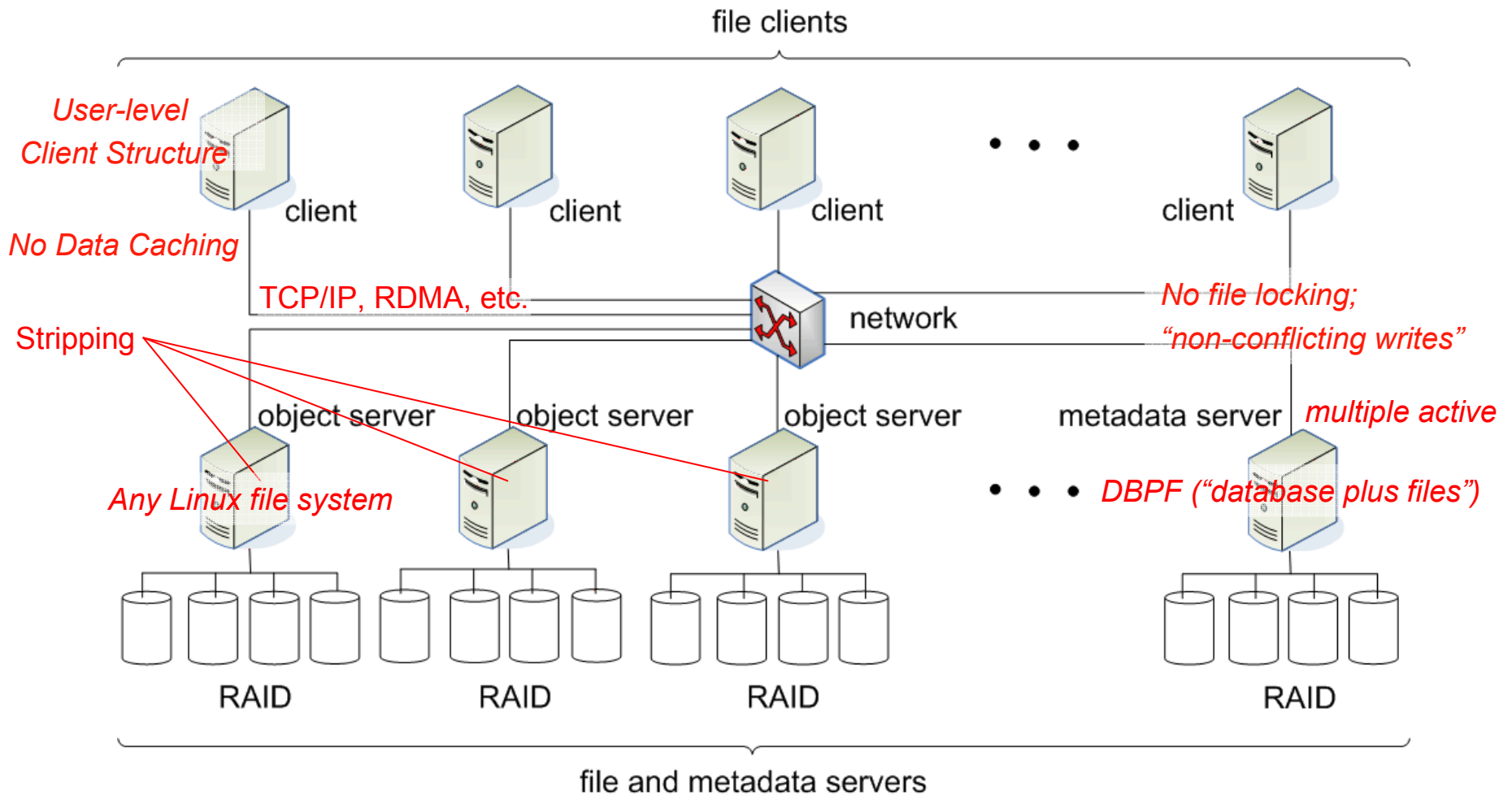
NASD-style Parallel File Systems



Lustre Architecture



PVFS2 Architecture



Summary of Results

- 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

Details can be found at USENIX Large-Scale Computing'08 paper



<http://www.ics.forth.gr/carv/scalable>