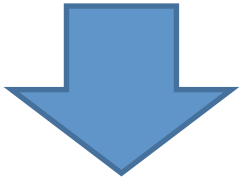
High Throughput, Low latency and Reliable Remote File Access

Hiroki Ohtsuji and Osamu Tatebe
University of Tsukuba, Japan
/ JST CREST

Motivation and Background

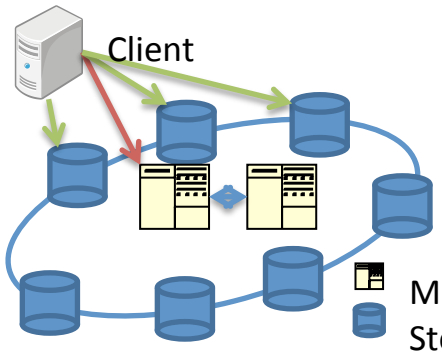
- Data-intensive computing is a one of the most important issue in many areas
- Storage systems for Exa-byte (10^{18})



Need a fast and reliable remote file access system

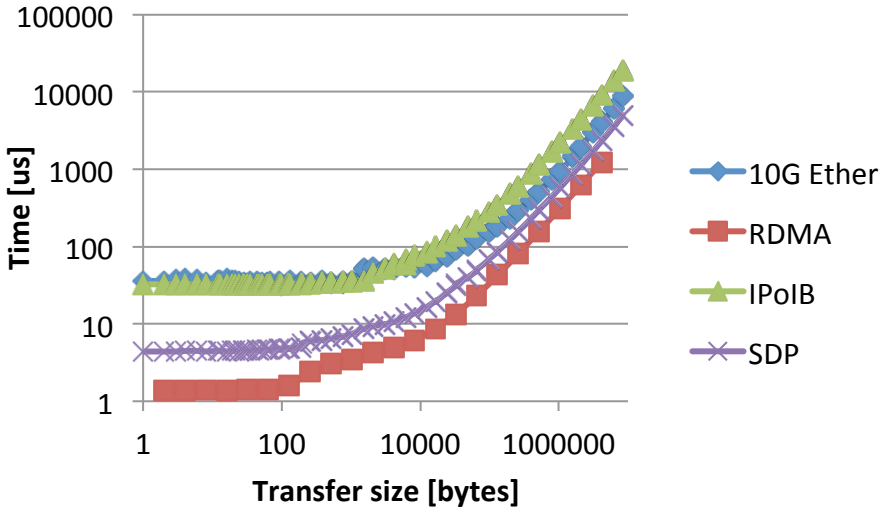
Motivation and Background(cont'd)

- Data sharing
  - Distributed file system
  - Clients access the data via Network
- Bottlenecks
  - Wide-area network
    - Long latency
  - Storage cluster
    - Overhead of network
- Fault tolerance
  - Suggestion: Congestion avoidance



Remote file access with RDMA

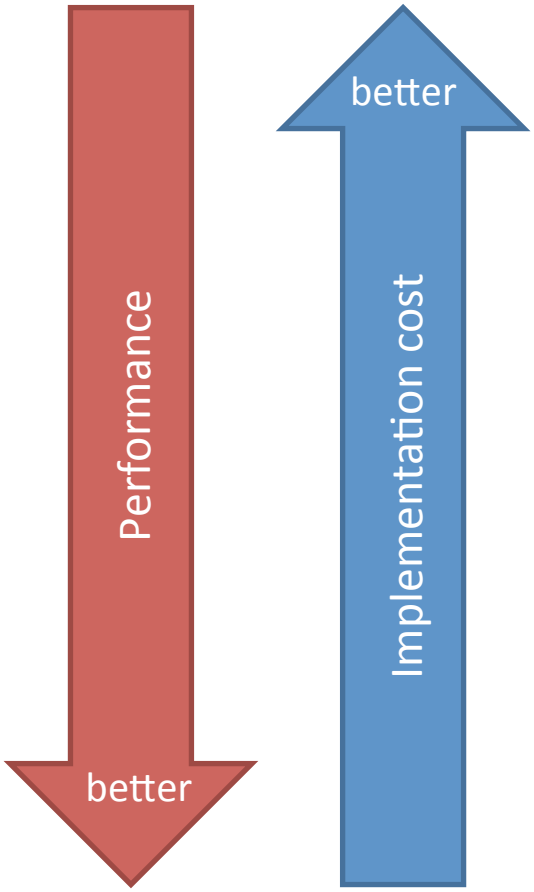
- Latency of Ethernet is at least 50 microseconds
  - Overhead of software
  - Protocol
  - Memory copy



- Flash memory based storage devices
  - 25µs latency (e.g. Fusion-io ioDrive), (HDD=5ms)
    - Network becomes a bottleneck of the system

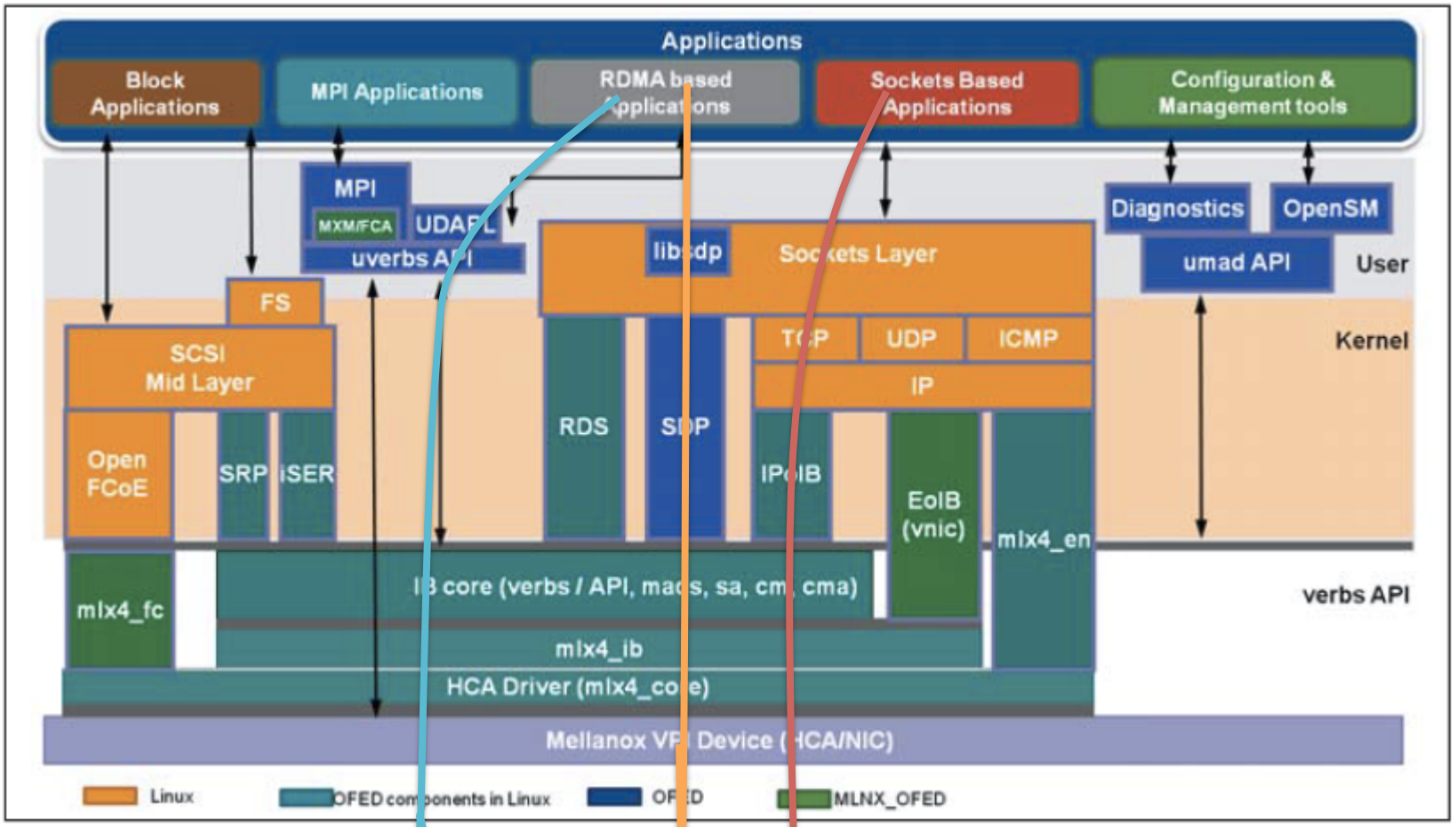
Usage of Infiniband

- IP over IB
  - Use the IP Protocol stack of operating systems
    - Pros
      - Can use as a network adapter
    - Cons
      - Inefficient
- SDP (Socket Direct Protocol)
  - Pros
    - Easy to use
      - Specify the LD_PRELOAD
    - Cons
      - Performance
- RDMA (Verbs API)
  - Pros
    - Low-latency
  - Cons
    - No compatibility with socket APIs



Structure of OFED

OFED : Drive and libraries for Infiniband



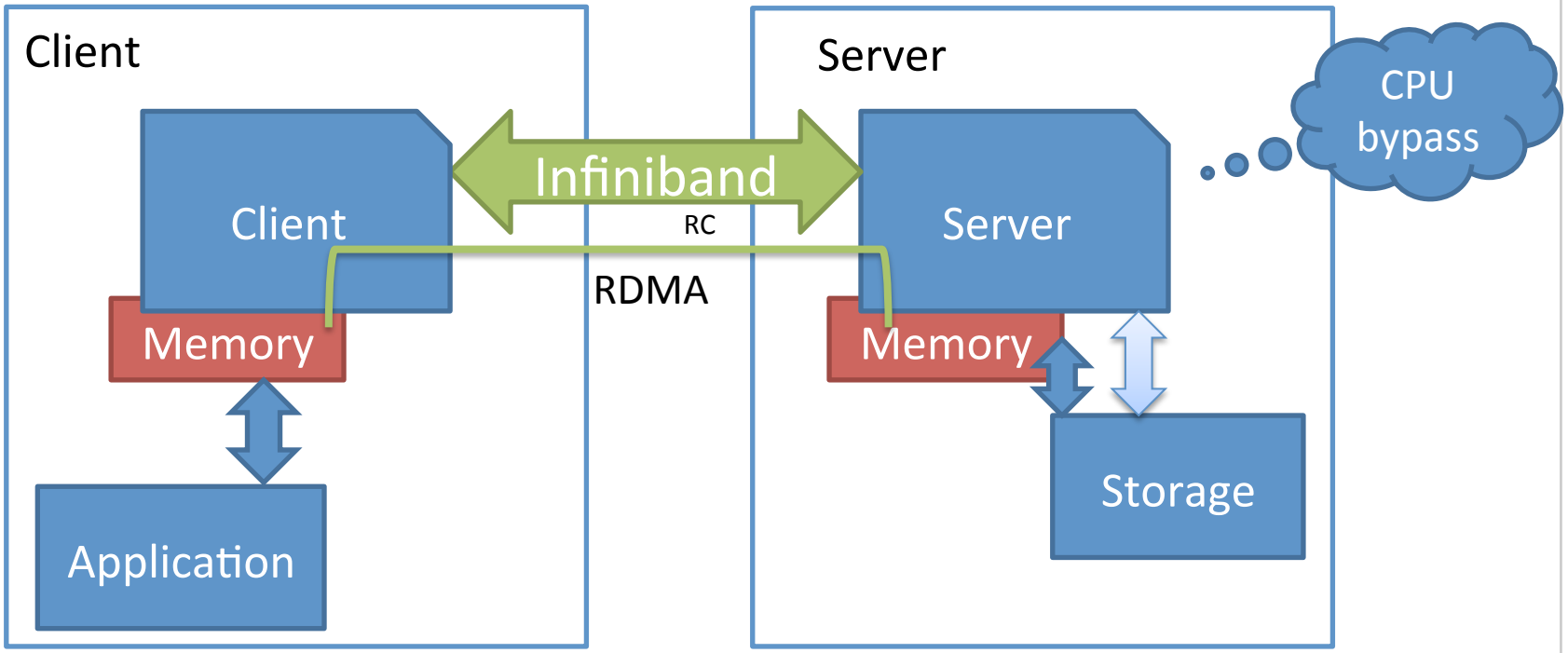
Verbs API SDP IPoIB

From ©Mellanox document

Remote file access with RDMA

Infiniband FDR (54.3Gbps)
Storage: Fusion-io ioDrive

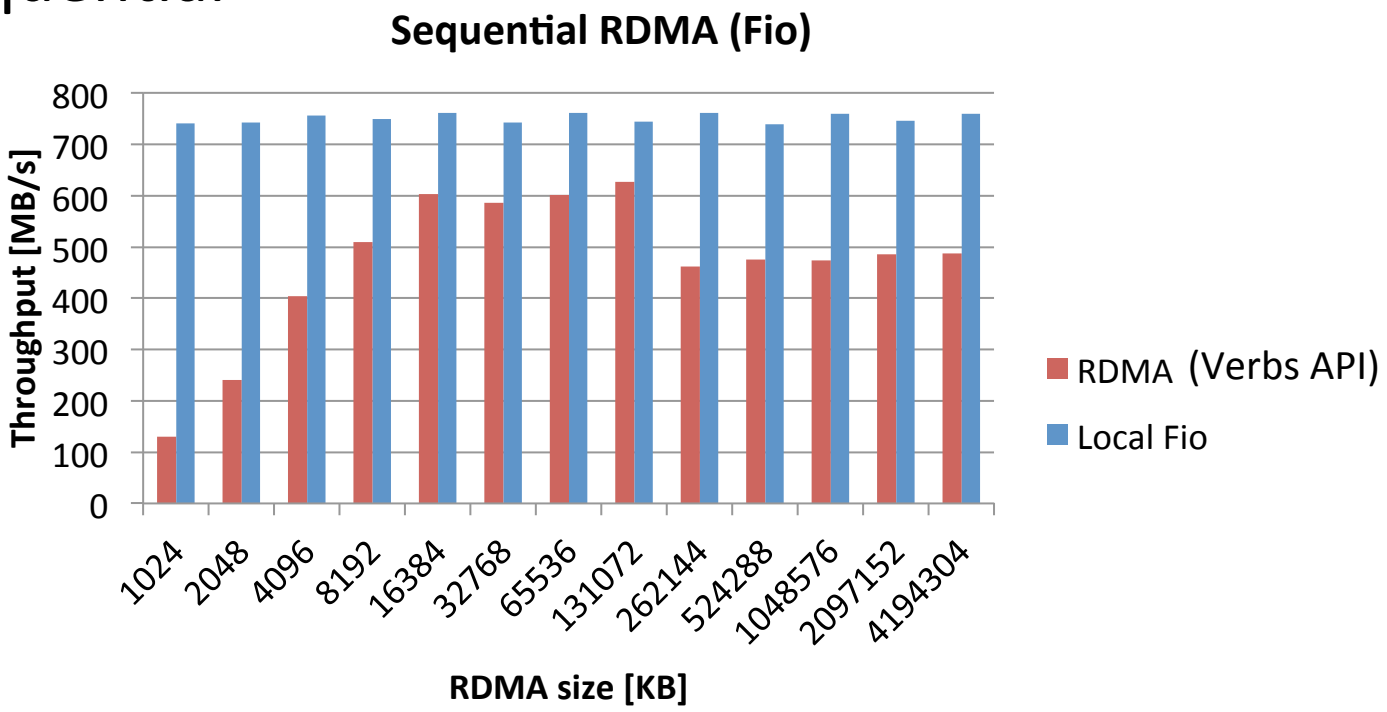
- Architecture



Low overhead remote file access with Verbs API

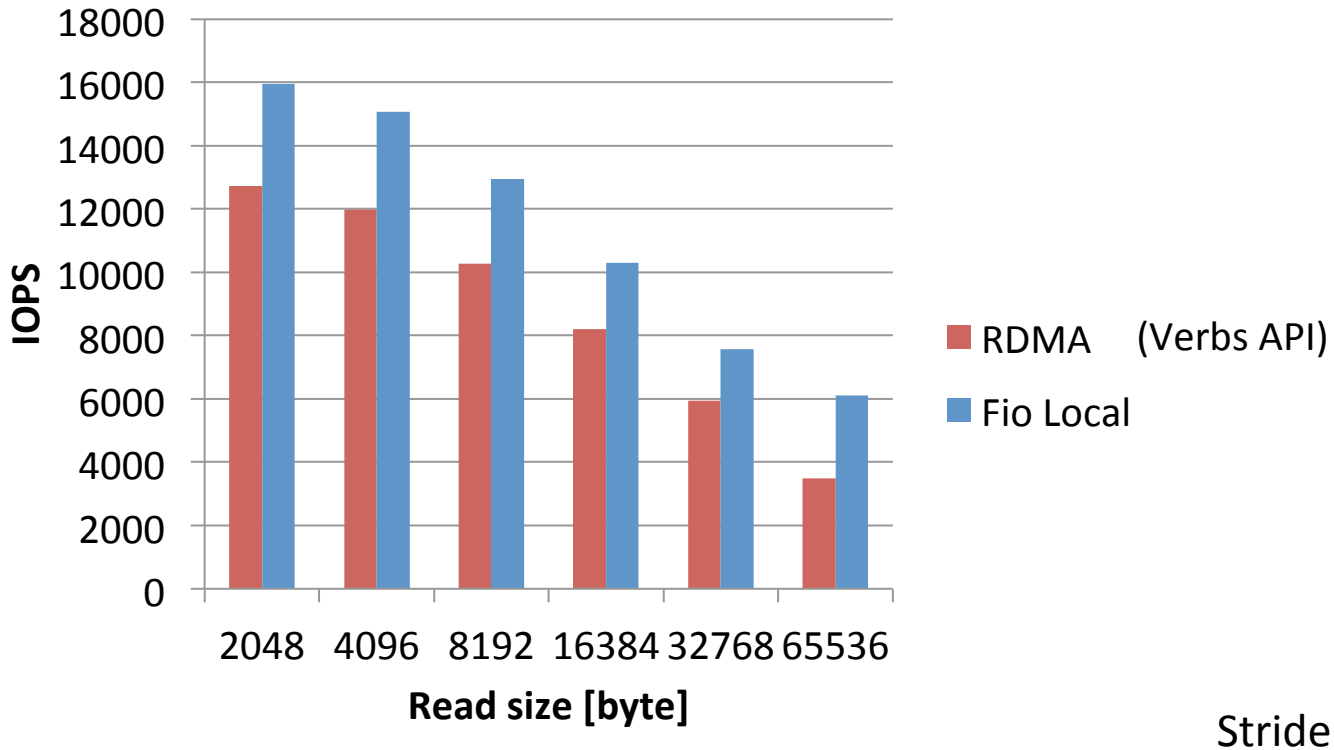
Preliminary Evaluation: Throughput

- A client accesses the file on the file server via Infiniband w/ Verbs API
  - Sequential



Preliminary Evaluation of IOPS

- Stride access from 2KB-64KB (seek 1MB)



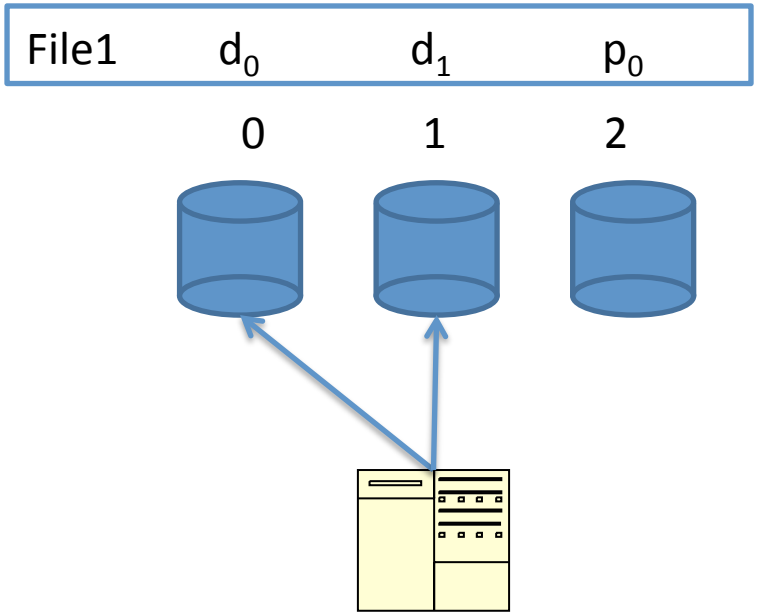
Stride = 1MB

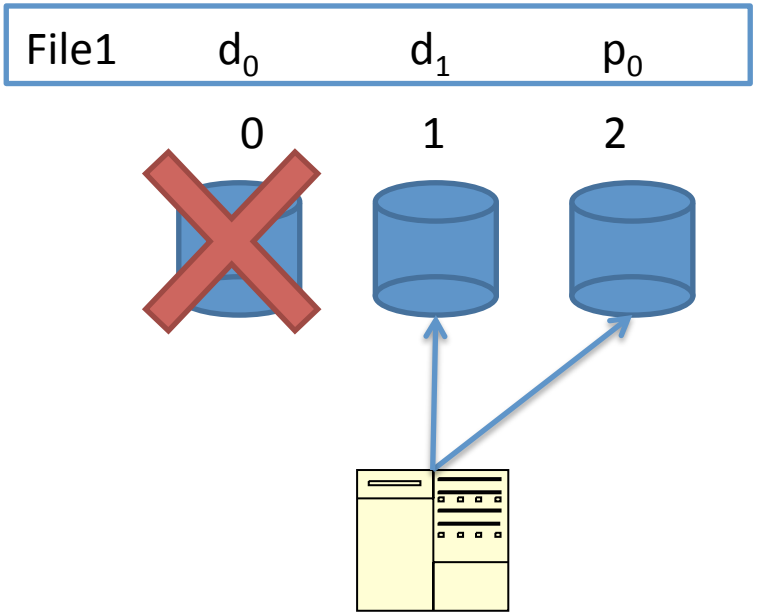
Congestion avoidance by using redundant data

- Concentration of access
  - There are hotspots (files) on the storage node
- Redundant data
  - Fault tolerance
  - Can be use to avoid congestion

Redundant data

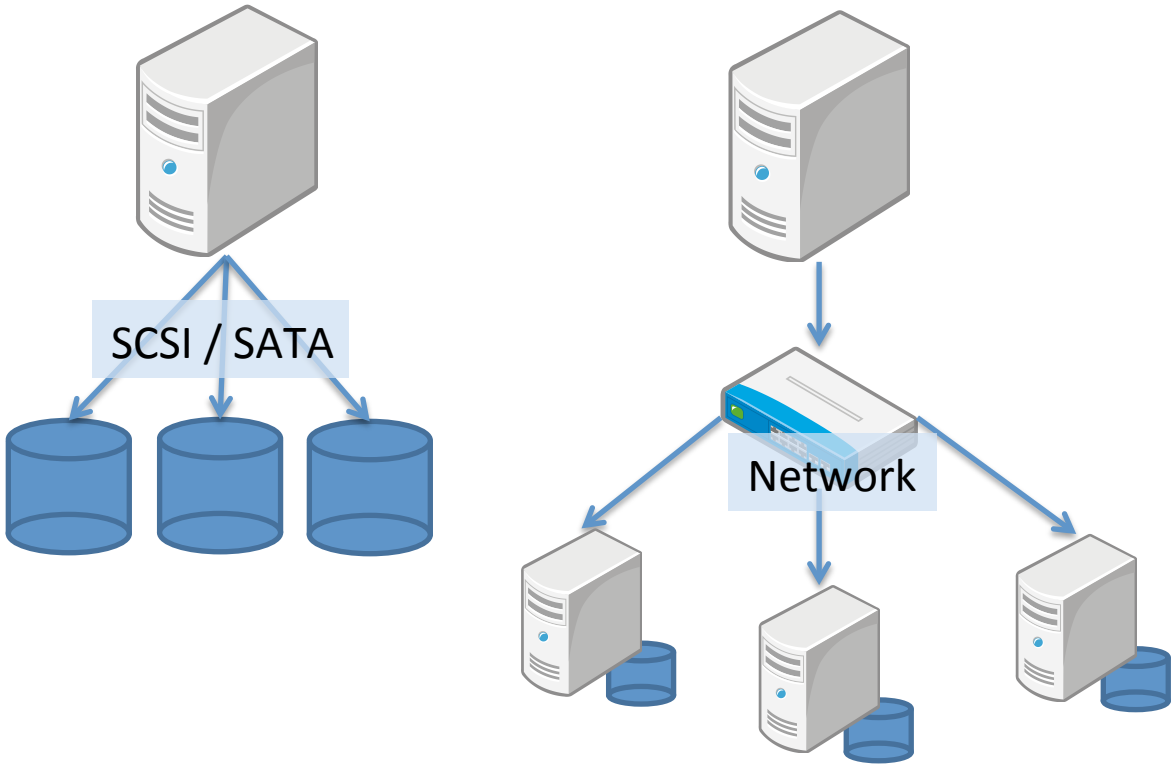
- Basic structure



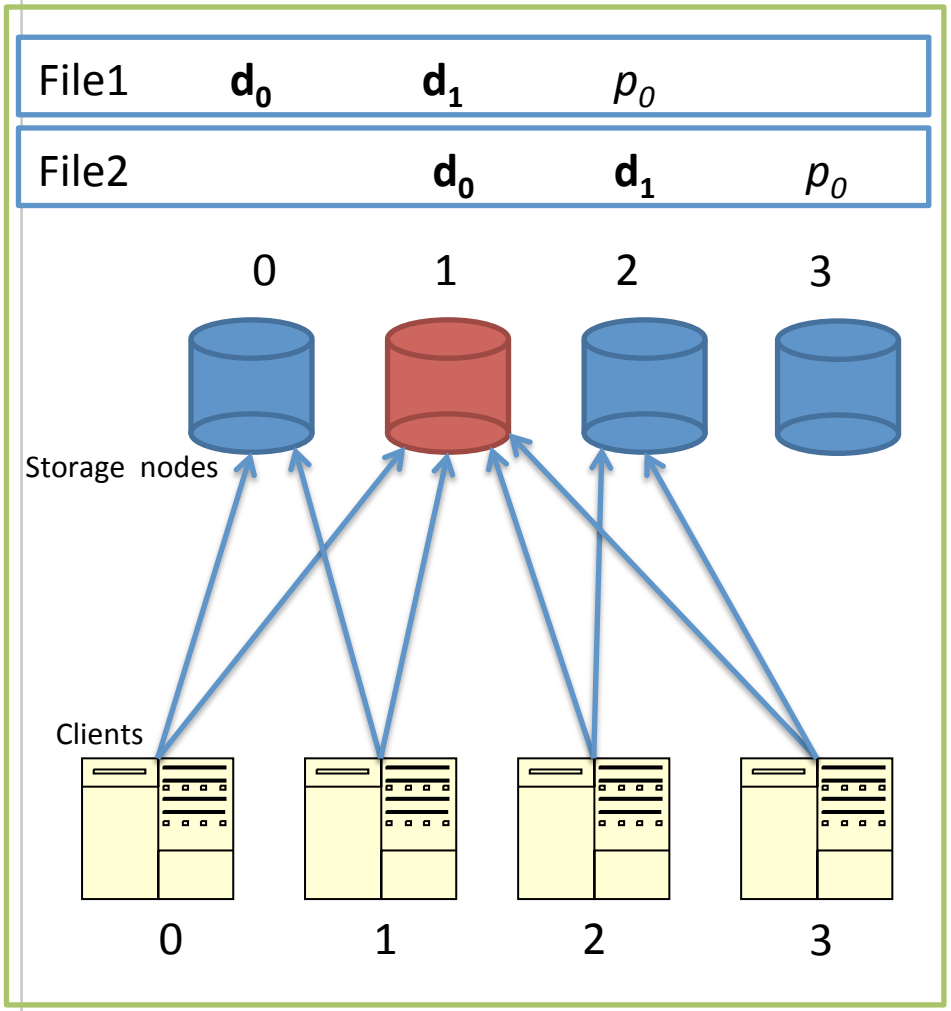


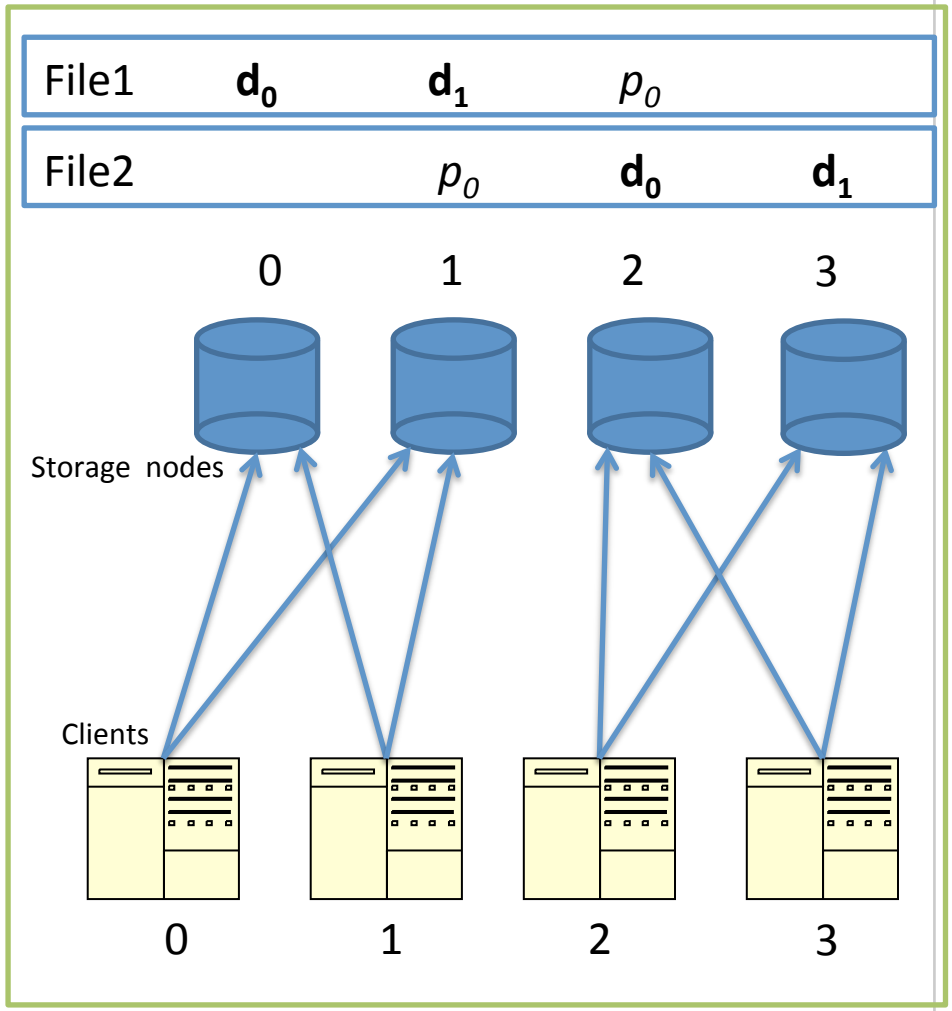
d_1 xor p_0 = d_0

- RAID: connected with SCSI / SATA
- →connected with network

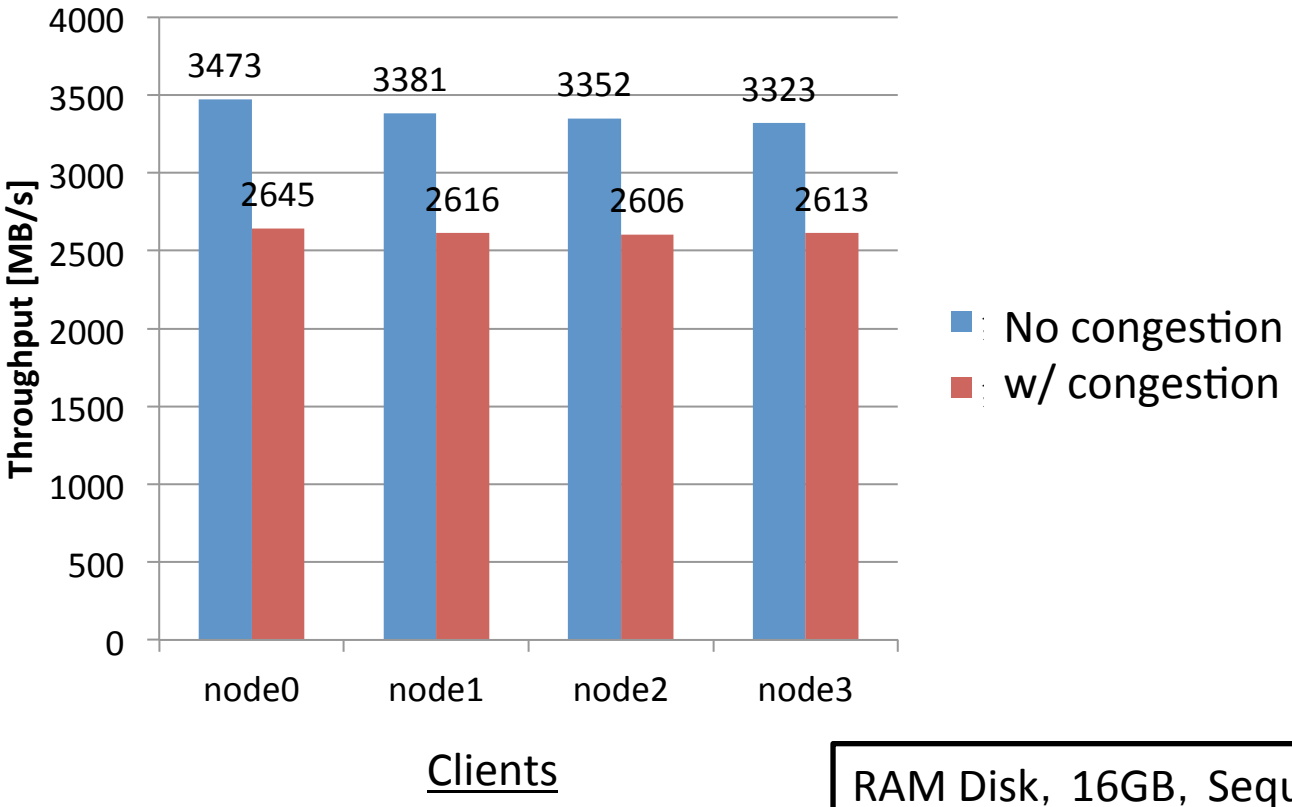


Performance deterioration



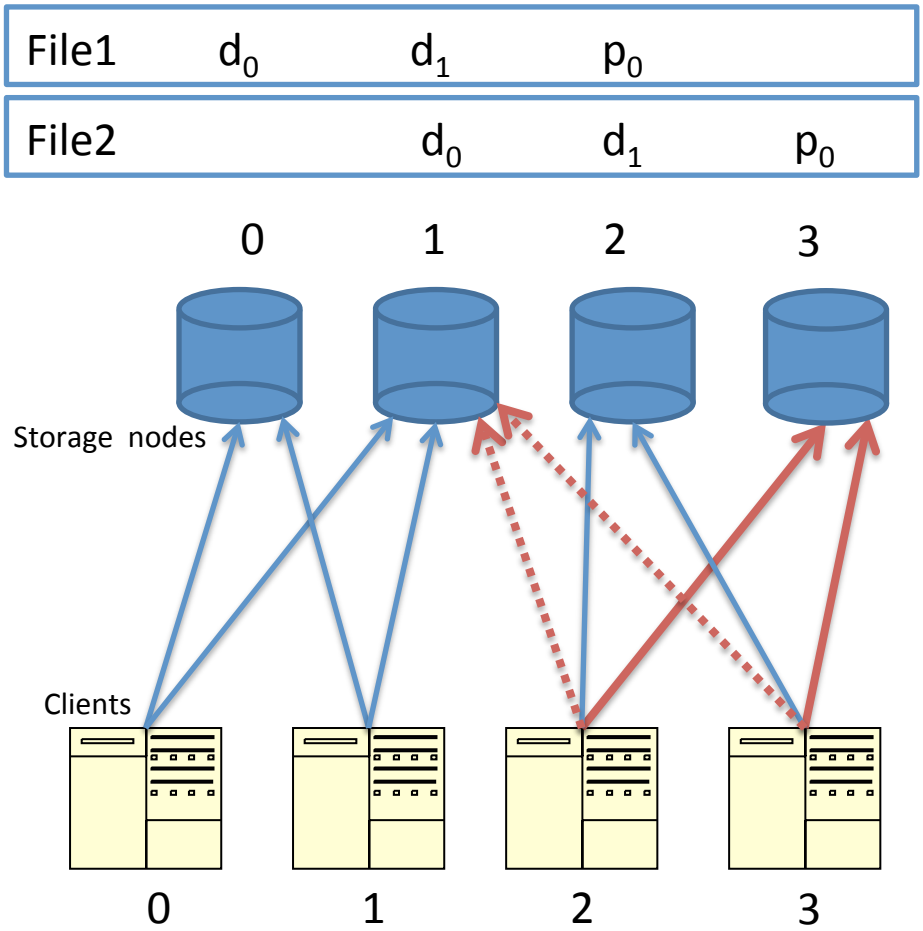


Performance deterioration(cont'd)



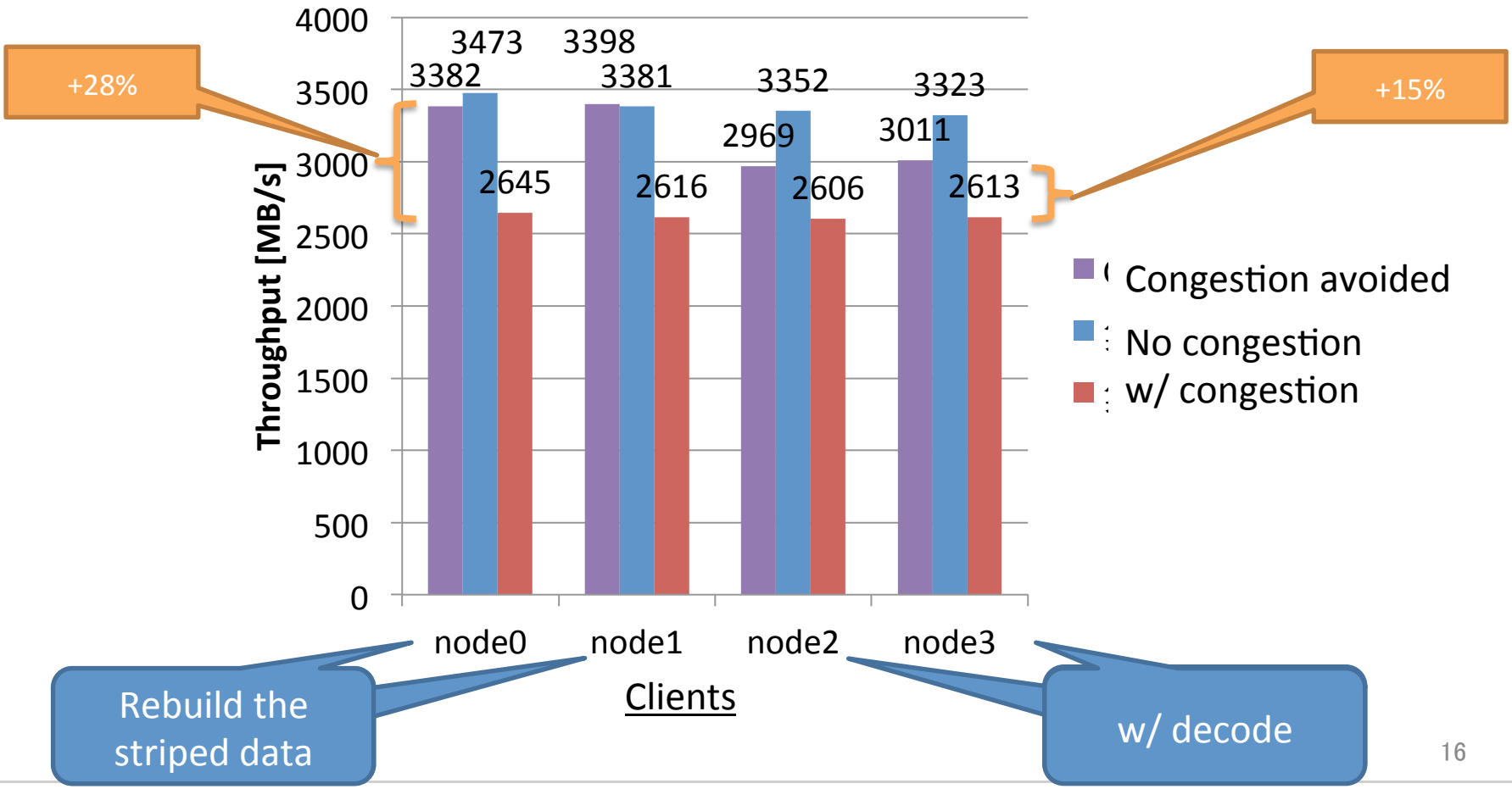
RAM Disk, 16GB, Sequential

Congestion avoidance



Performance evaluation

- Compare the cases



Related work

- Stephen C. Simms et al, Wide Area Filesystem Performance using Lustre on the TeraGrid, 2007.
- Wu, J., Wyckoff, P. and Panda, D.: PVFS over InfiniBand: Design and Performance Evaluation
- Erasure Coding in Windows Azure Storage, USENIX ATC '12
  - Shorten the latency by using redundant data
- HDFS RAID

Conclusion and Future work

- Remote file access with Infiniband RDMA
- Congestion avoidance
- Future work
  - How to detect the congestion
  - Writing of data(in progress)
    - w/o performance degradation