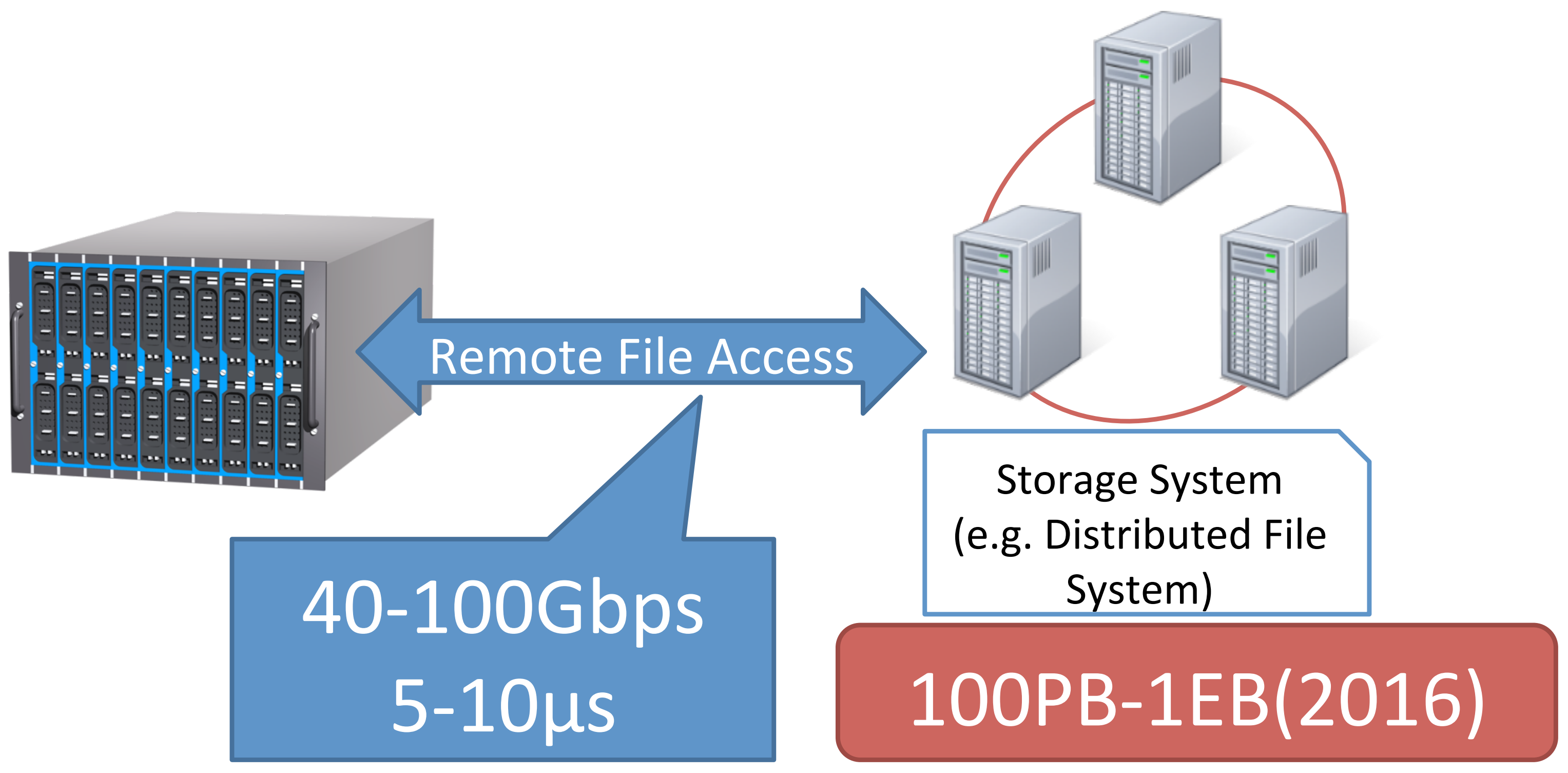
High Throughput, Low Latency and Reliable Remote File Access

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

Background



Exascale storage systems require the high bandwidth, low latency and reliable remote file access method.

・Latency and Bandwidth

Ethernet is a common network to connect storage nodes and client nodes. However, latency of Ethernet is at least a couple of hundreds microsecond. This is caused by the overhead of many hardware layers and the software stack. In order to accelerate the performance of applications, this should be eliminated and we need other sophisticated mechanisms to transfer the data. InfiniBand is one of the most suitable components to achieve this goal.

・Reliability

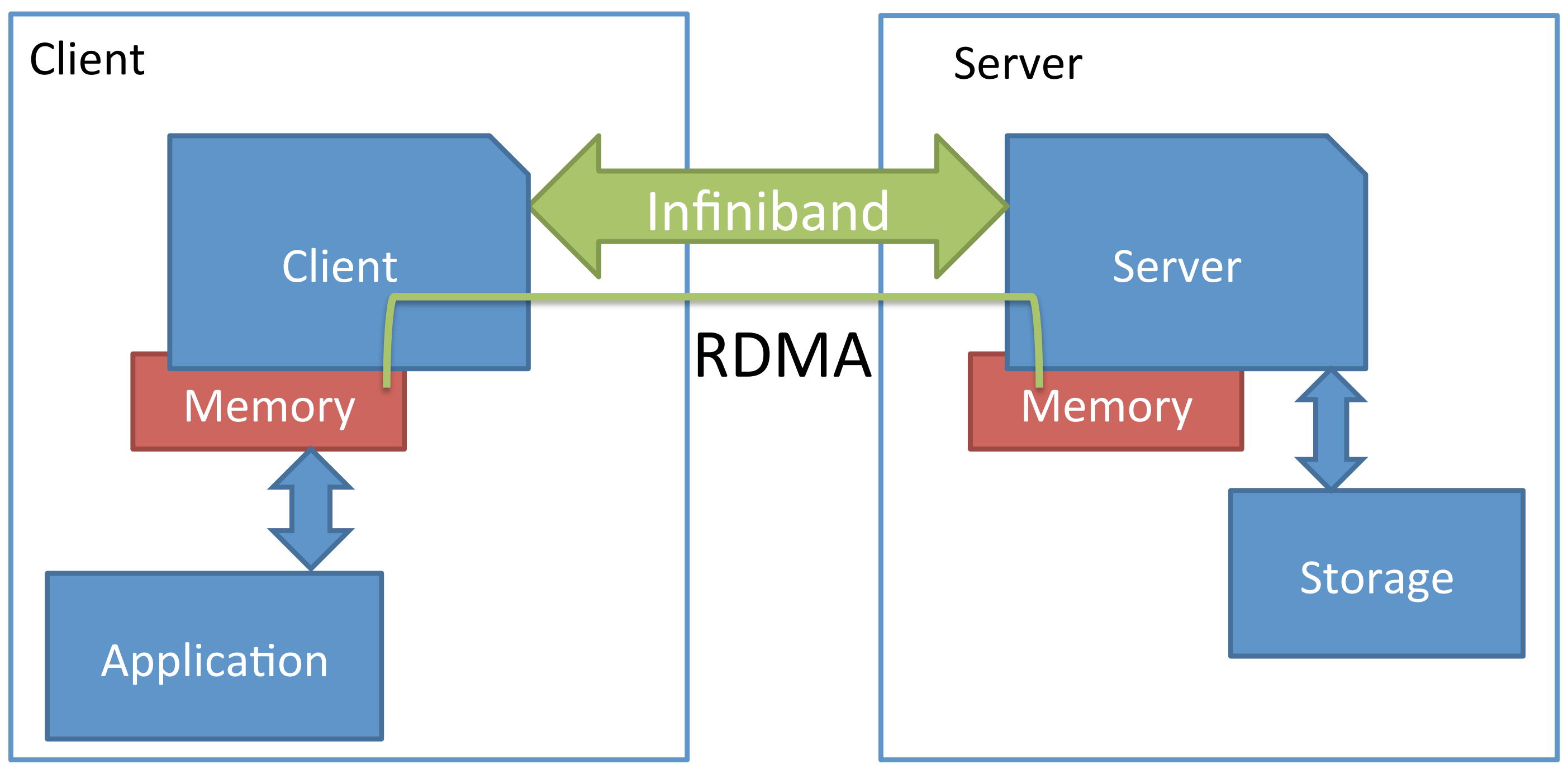
In terms of reliability, this poster describes how to securely store and access the data.

This poster mentions these three topics:

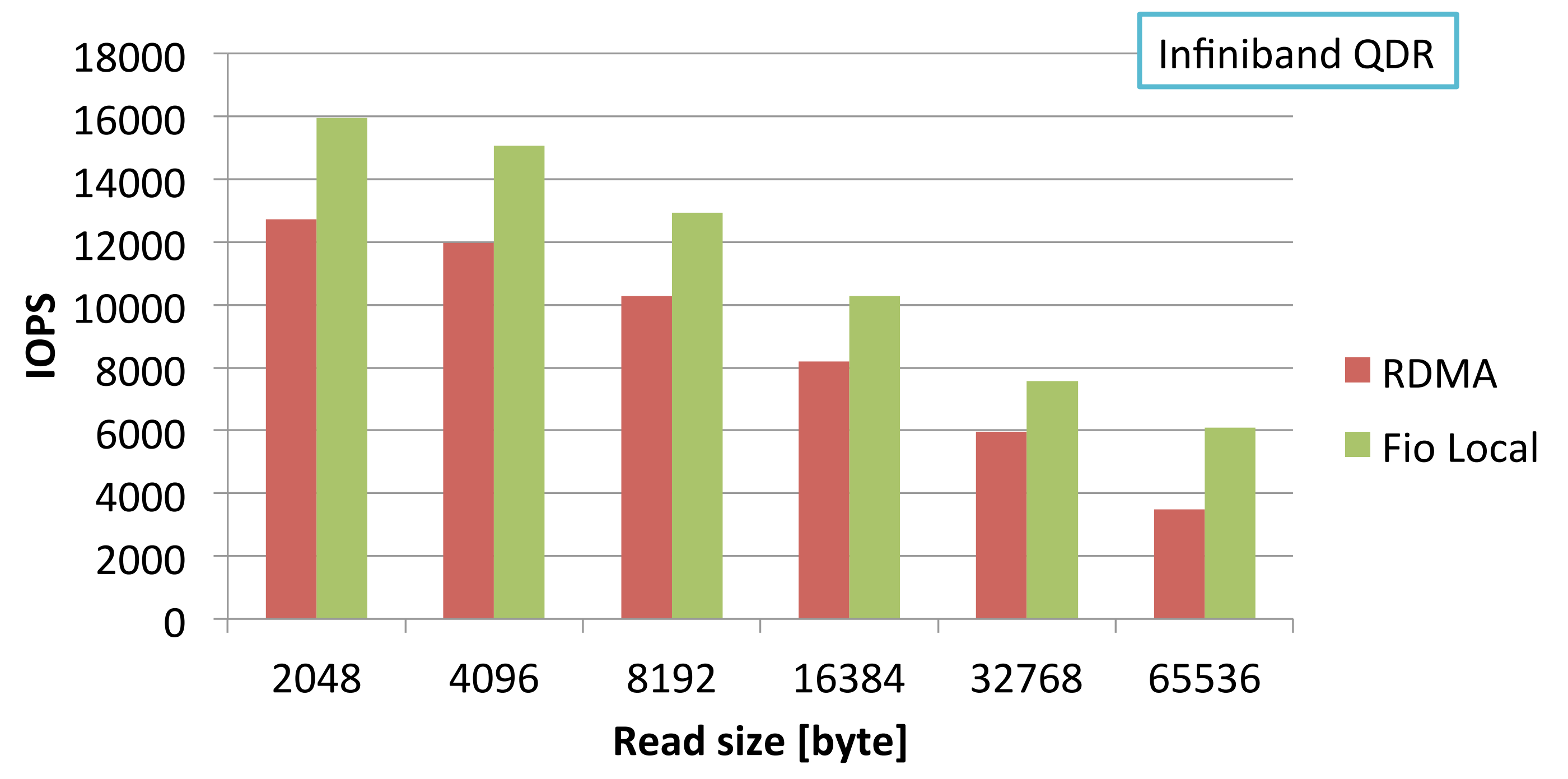
- Remote File Access with RDMA
  - Reduce the overhead of network communication with the CPU bypass architecture.
- Node-level redundancy
  - Replication is a famous method to prepare for failures. Redundant data can save amount of disk space, while replication takes space more than twice as large as size of the original data.
- Congestion avoidance
  - Redundant data provides more options to choose storage nodes. Some of combinations of storage nodes can avoid the network congestion.

Remote File Access with RDMA

Architecture

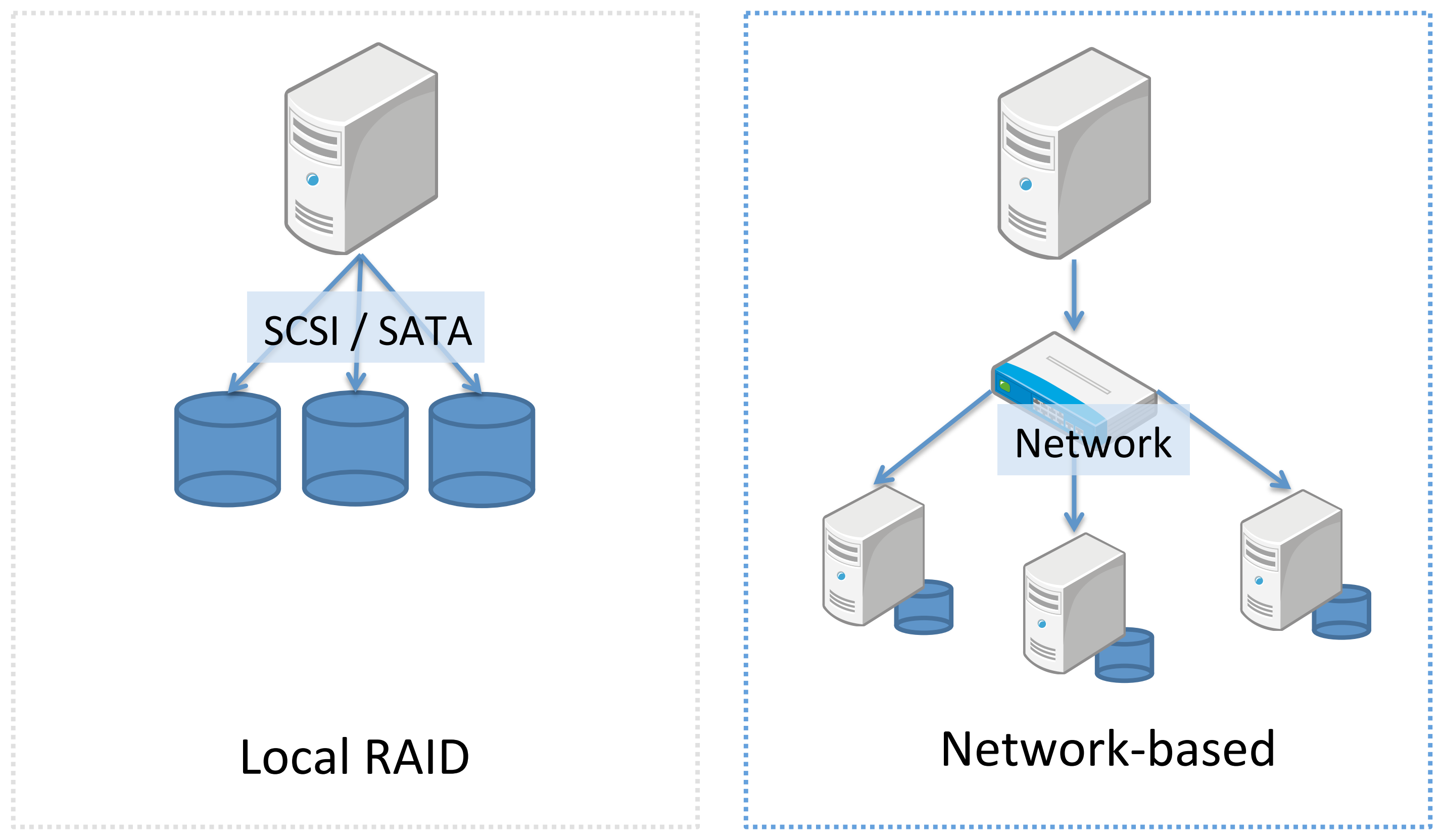


IOPS evaluation of stride access

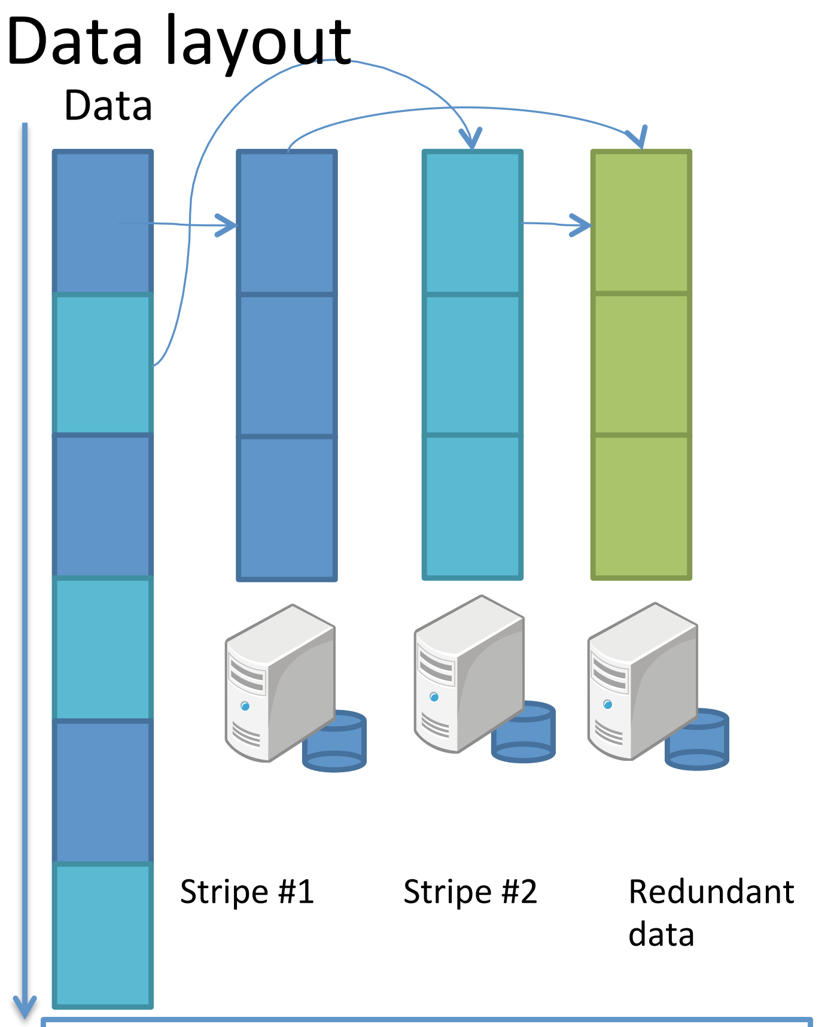


RDMA reduces the network latency and enhances the IOPS .

Node-level Redundancy

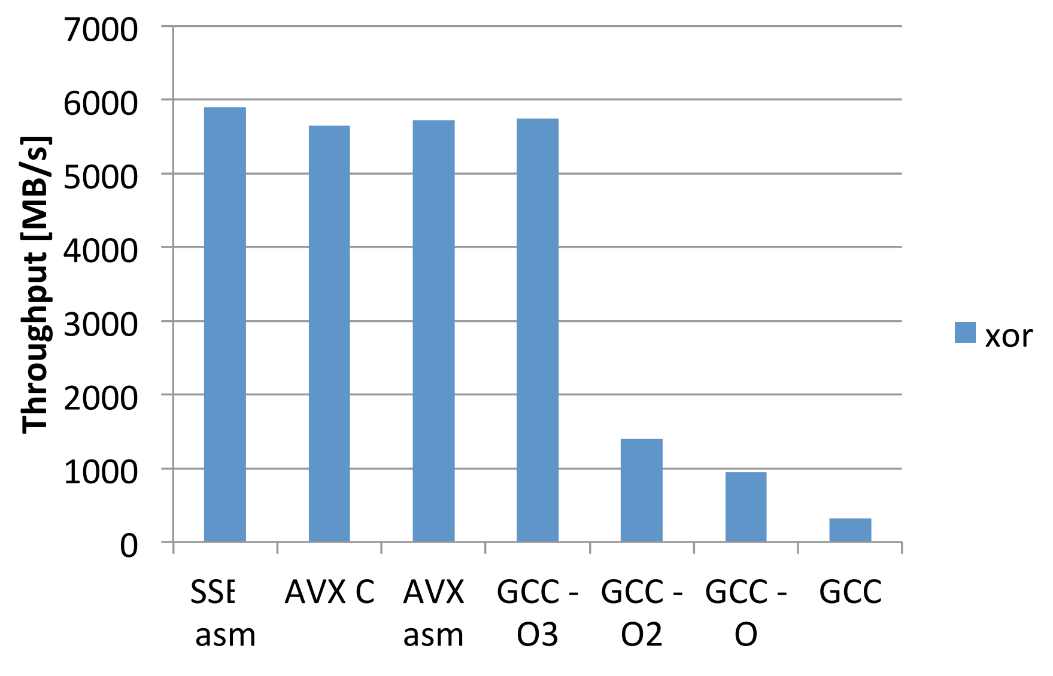


Data layout



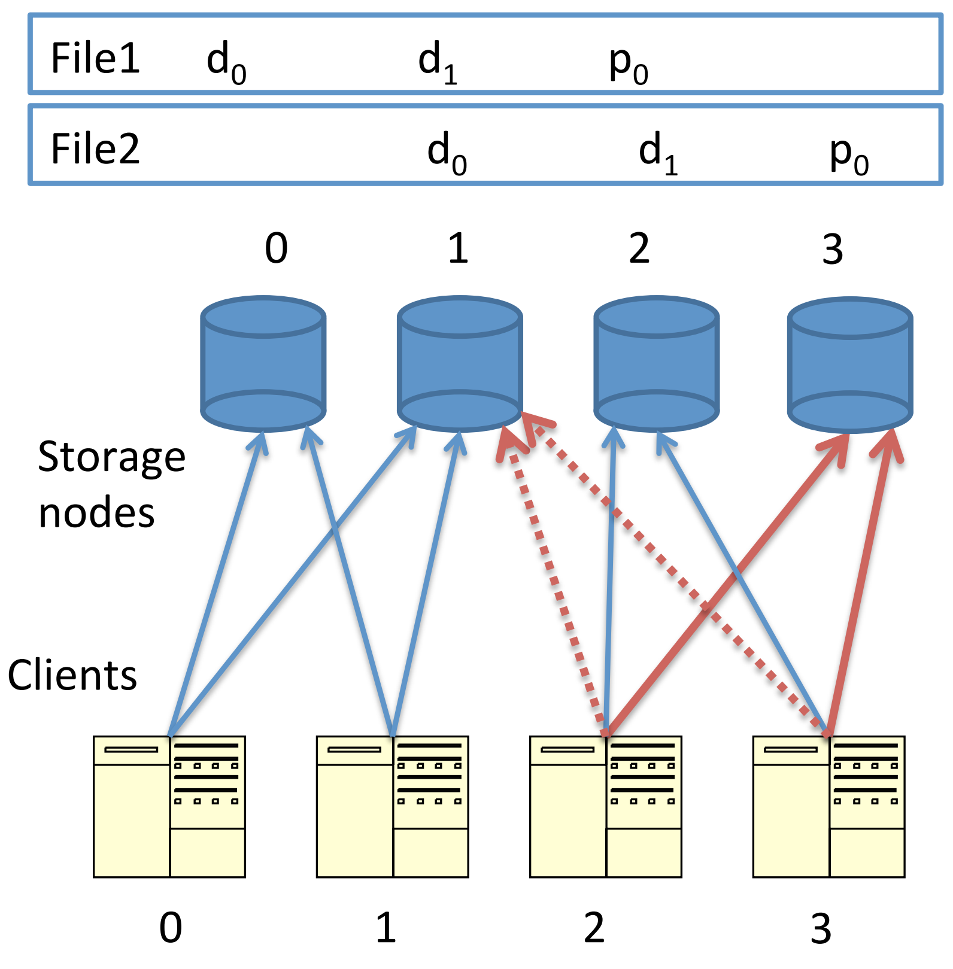
If stripe #1 is failed, the original data can be calculated from the data of stripe #2 and the redundant data.

XOR throughput (CPU)



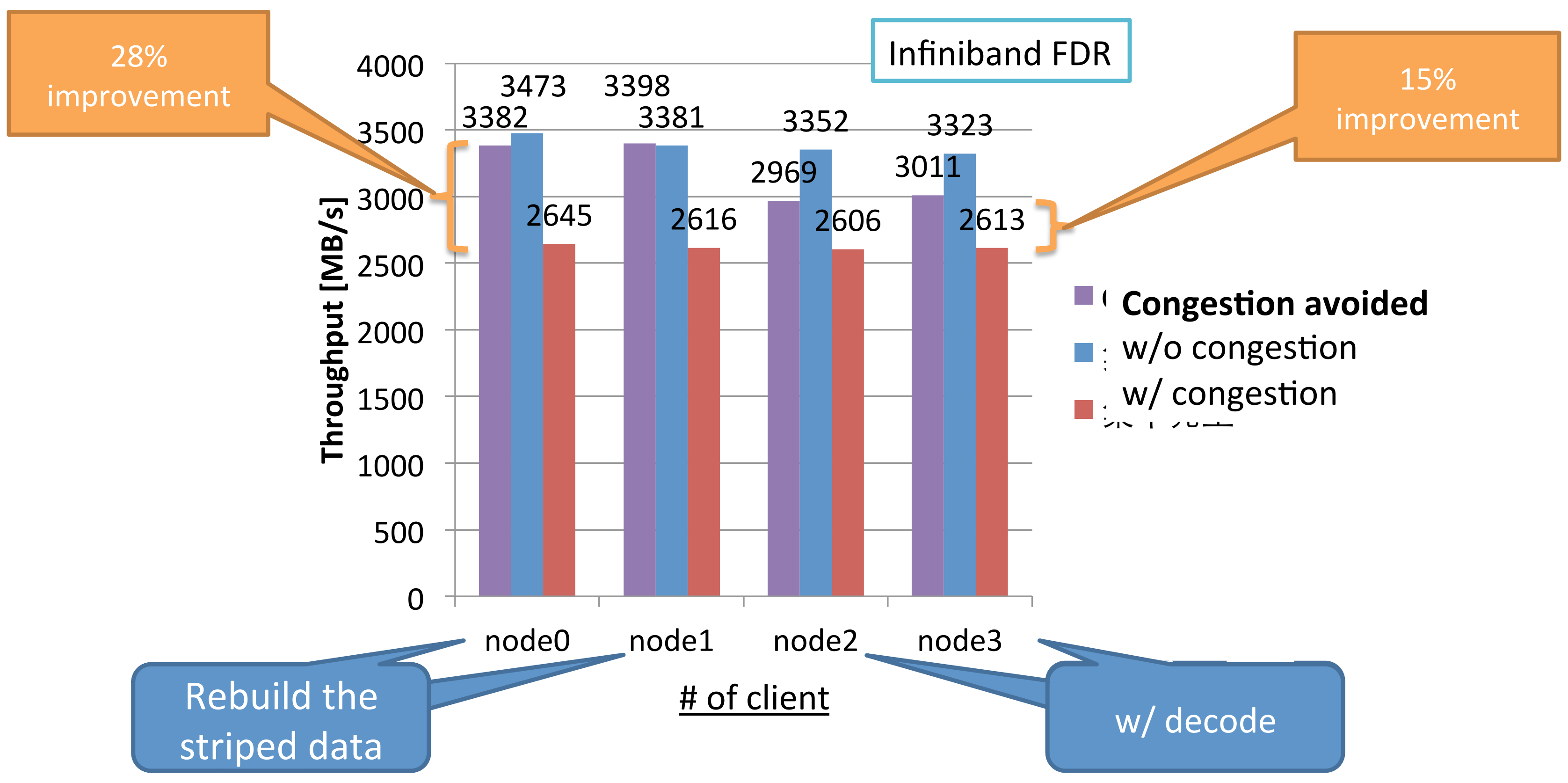
The XOR throughput ranks with throughput of InfiniBand FDR.

Congestion Avoidance



Network of storage node #1 will be congested if all clients access the files concurrently.
This congestion is avoided by the proposed method, which uses the redundant data to disperse the traffic.

Performance Evaluation


28% improvement
15% improvement
Infiniband FDR
Throughput [MB/s]
node0
node1
node2
node3
# of client
Congestion avoided
w/o congestion
w/ congestion
Rebuild the striped data
w/ decode