Pipelined Data Processing Architecture for Network Storage Systems

Hiroki Ohtsuji[1,2] and Osamu Tatebe[1]

[1]University of Tsukuba/JST CREST

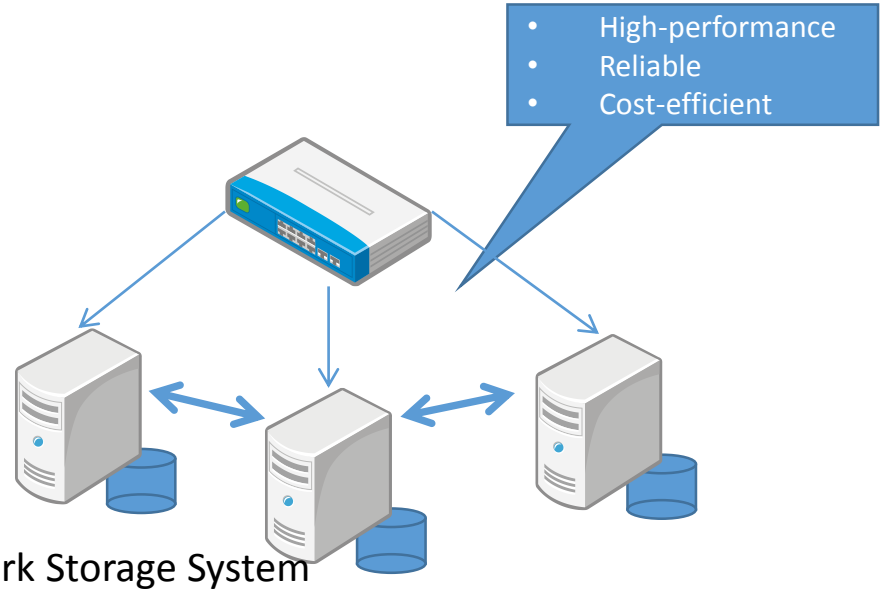
[2]JSPS Research Fellow

Abstract

- Exa-scale computing
  - 10^18 FLOPS, 10^18 Bytes
- Exa-scale systems require high-performance and reliable network storage systems
  - Scientific data
  - Big data

Our target:
Exa-scale storage system

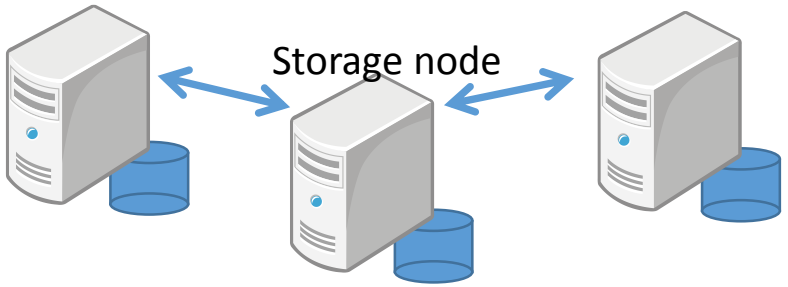
- Performance
- Reliability
  - Replication [1]
  - Erasure coding [2]



[1]CHERVENAK, A. L., FOSTER, I. T., KESSELMAN, C., SALISBURY, C., AND TUECKE, S. The data grid: Towards an architecture for the distributed management and analysis of large scientific datasets. JOURNAL OF NETWORK AND COMPUTER APPLICATIONS 23
(1999), 187–200.
[2]WEATHERSPOON, H., AND KUBIATOWICZ, J. D. Erasure coding vs. replication: A quantitative comparison. In In Proceedings of the First International Workshop on Peer-to-Peer Systems (IPTPS 2002 (2002).

Active-storage mechanism

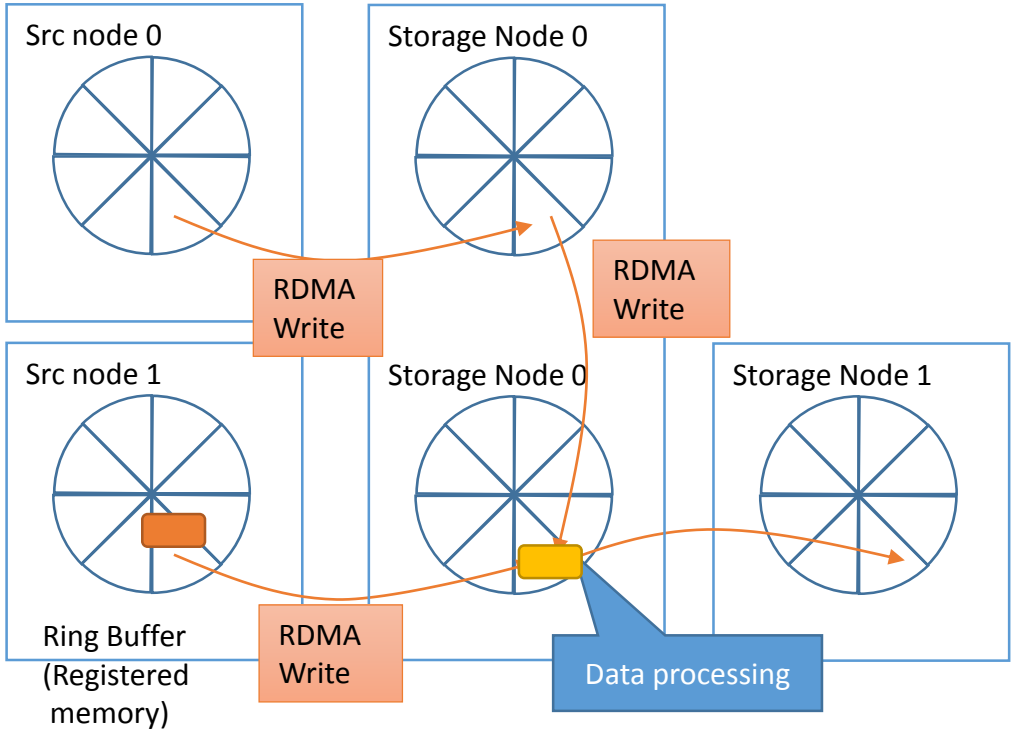
- W/o centralized controller
- Storage nodes themselves exchange data and process data blocks
  - Utilize network path(s) between storage nodes
- Building a data processing pipeline to implement erasure coding on the network storage system.



Data processing

Pipelined data processing

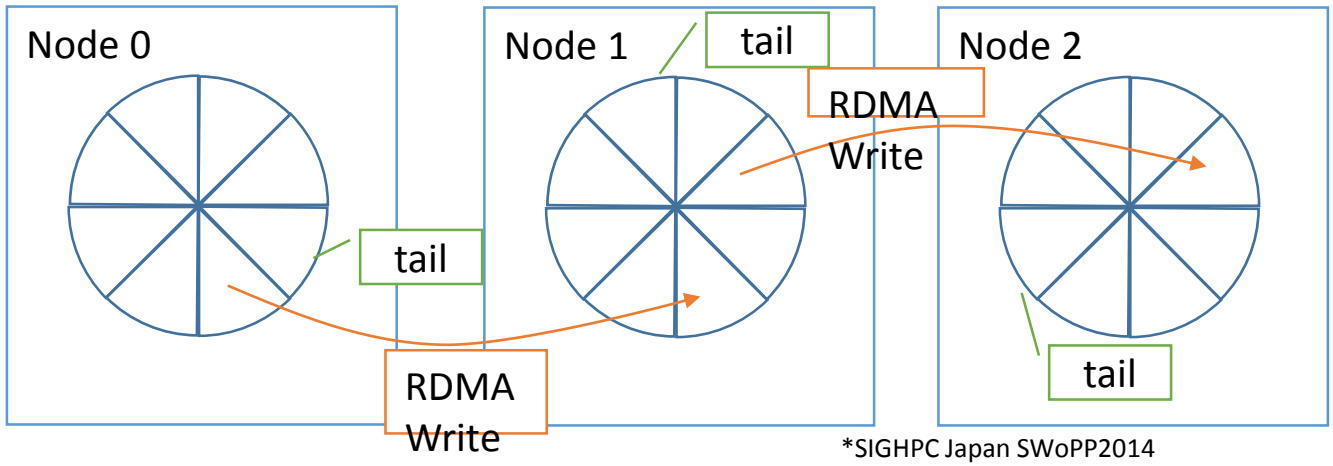
- Minimize the number of memory copies
- Utilize Remote Direct Memory Access (RDMA) of InfiniBand network



*SIGHPC Japan SWoPP2014

Zero-copy implementation

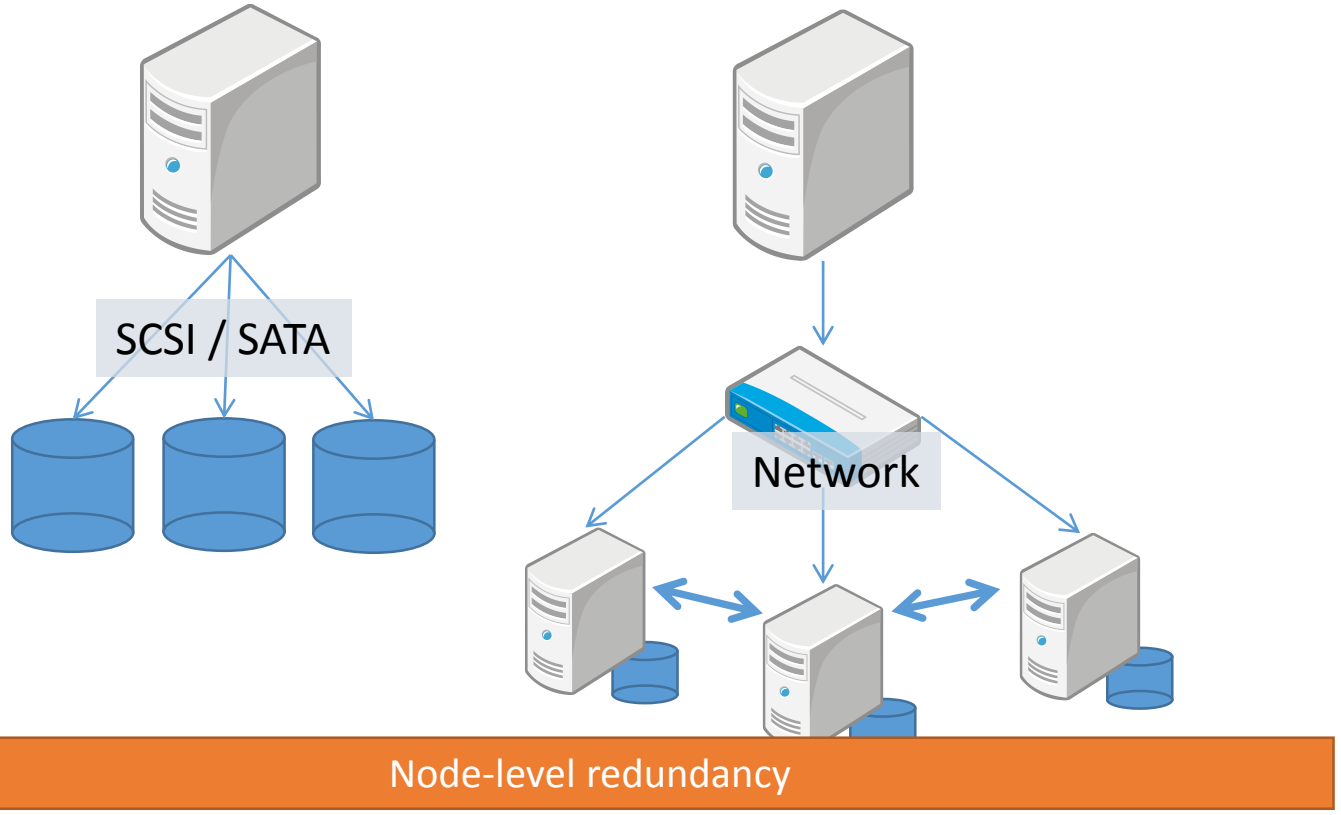
- `Share the buffer memory block`



*SIGHPC Japan SWoPP2014

Cluster-wide RAID

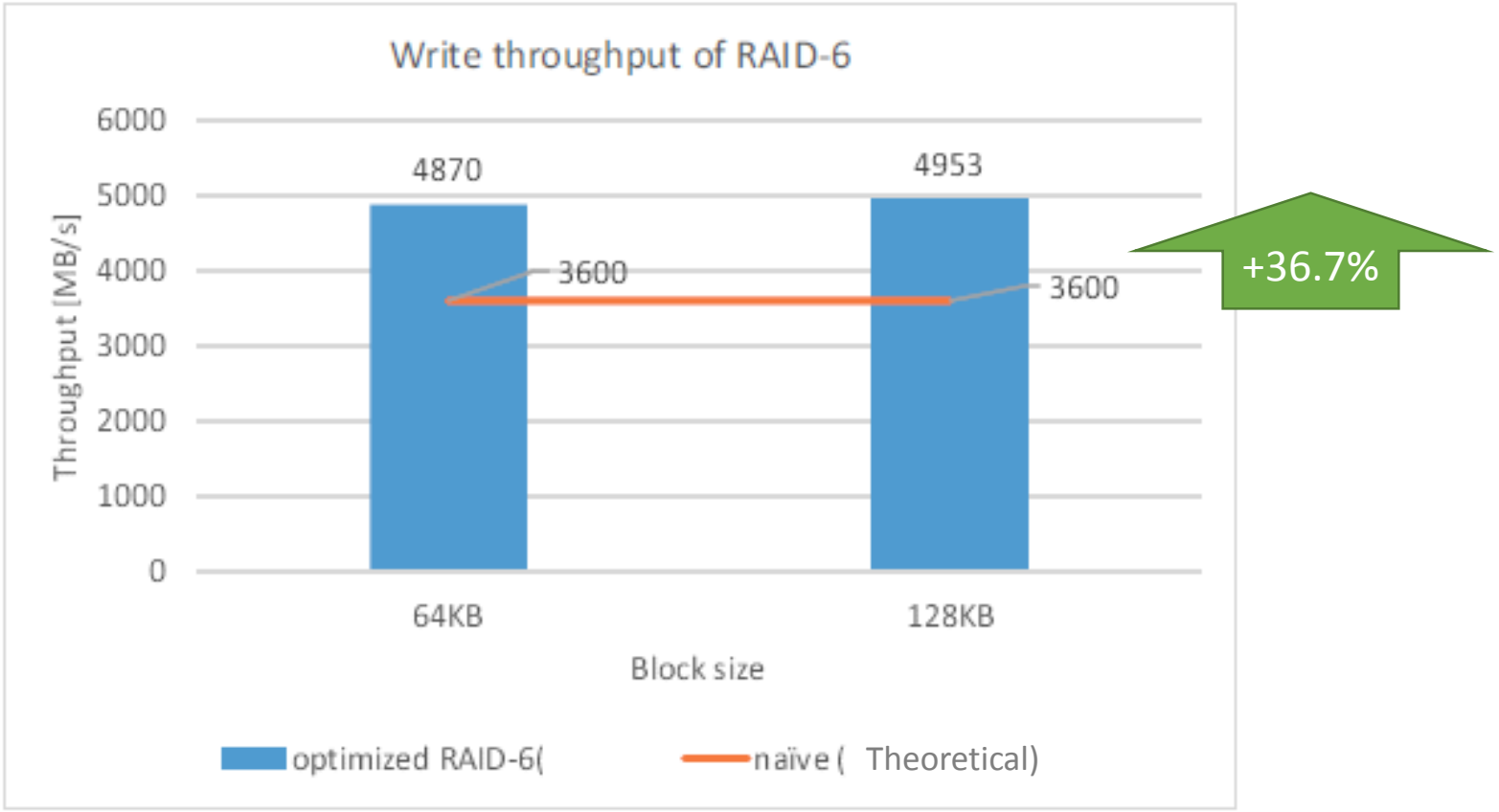
- Apply the active-storage mechanism to network RAID systems



*SIGHPC Japan SWoPP2014

Network RAID-6 w/ data pipeline

•



*SIGHPC Japan SWoPP2014

Conclusion

- Apply the active-storage mechanism to network storage system
  - Network-RAID
  - Erasure coding
- Performance improvement
  - +37% with Network-RAID-6