Breaking the Trade-off between Performance and Reliability of Network Storage System

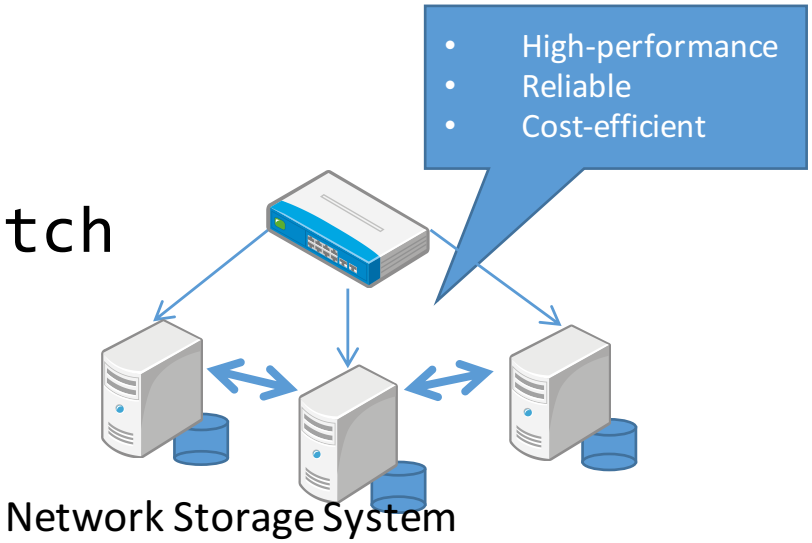
Hiroki Ohtsuji[1,2] and Osamu Tatebe[1]

[1]University of Tsukuba/JST CREST

[2]JSPS Research Fellow

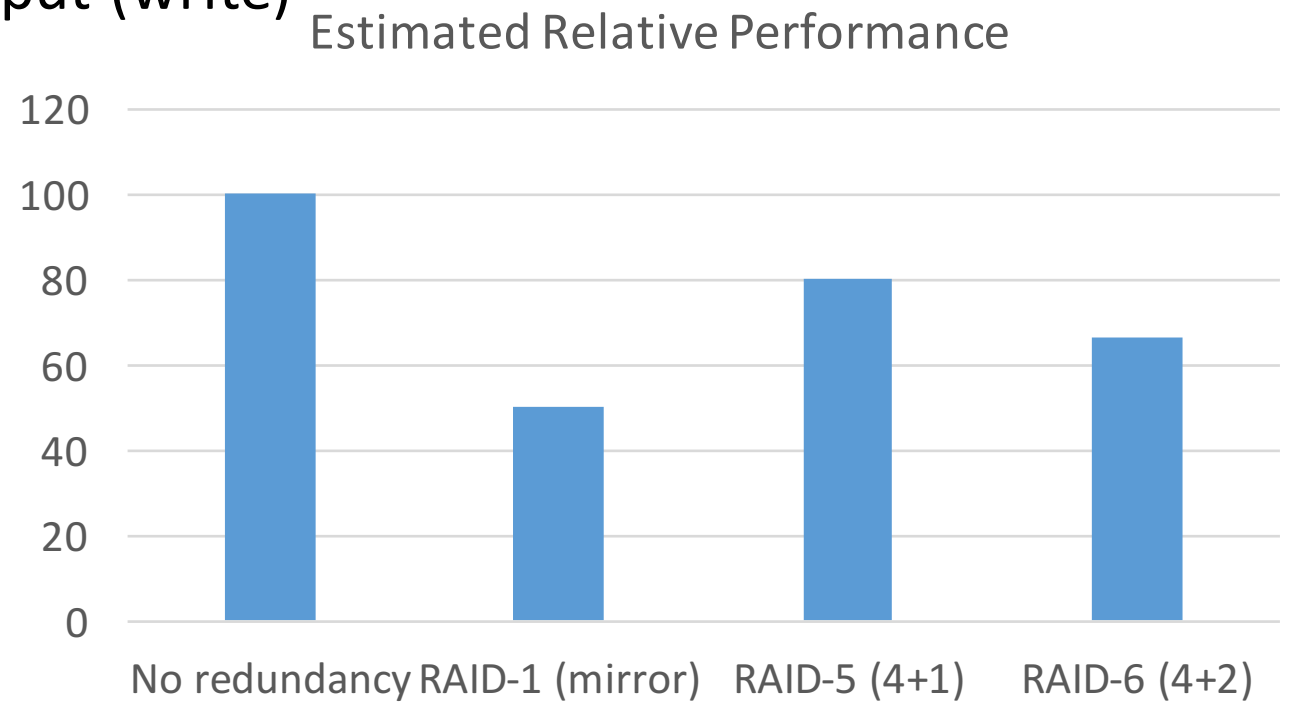
Background

- Requirements of exa-scale computing
  - High-performance and reliable storage systems
- Trade-off:
  - between performance and reliability of network storage system
- Optimization methods
  - Active-storage
  - Programmable network switch
- Adaptive strategy



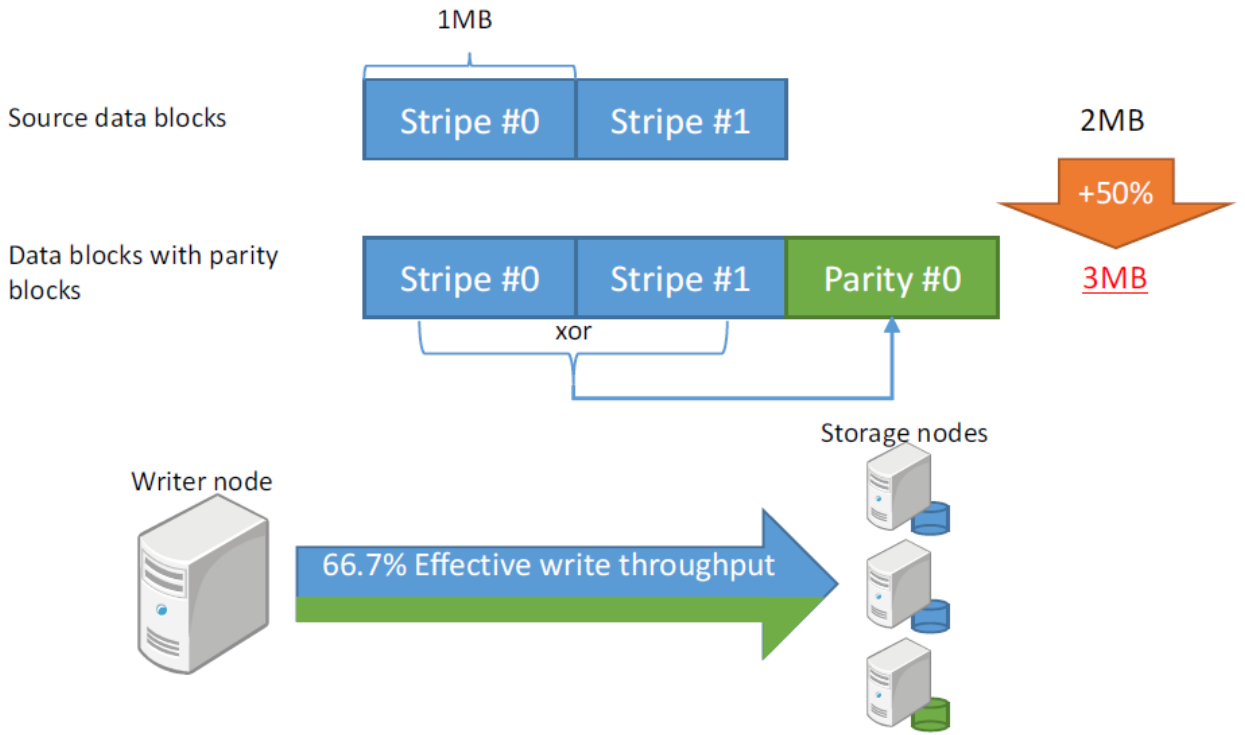
Trade-off: performance and reliability

- Performance
  - I/O throughput (write)
- Reliability
  - Data loss



P = (amount of original data / total amount of data with parity)

Performance degradation



(*2)

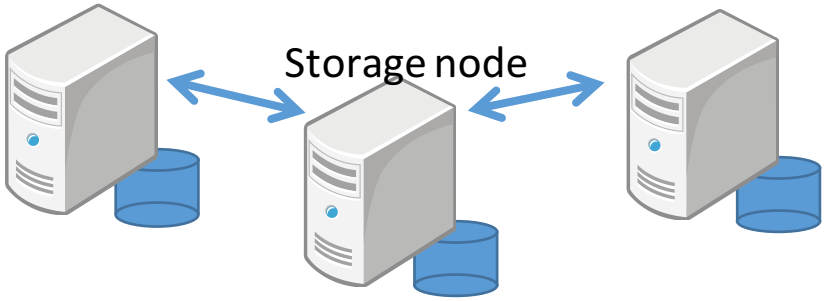
Total amount of traffic from a writer node is increased by parity block(s).
(33% degradation of write performance)

Our proposed optimization strategies

- Methods to optimize data write (to storage nodes)
  - Active-storage
    - Utilize computing capabilities of storage nodes
  - Programmable network switch
    - Utilize programmable functions of network switch

Active-storage mechanism (*1)

- Storage nodes exchange data blocks with each other to generate parity blocks

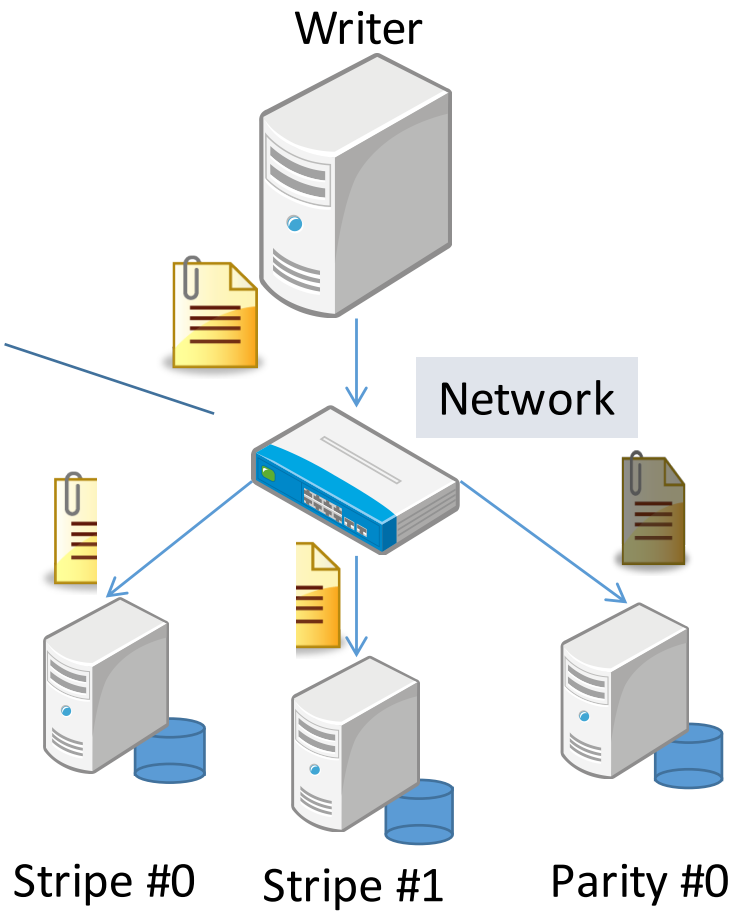


Off-loads the parity generation processes to storage nodes

*1 Hiroki Ohtsuji and Osamu Tatebe, “Active-Storage Mechanism for Cluster-wide RAID System”, International Conference on Data Science and Data Intensive Systems (IEEE DSDIS), pp.1-8, 2015 (to appear)

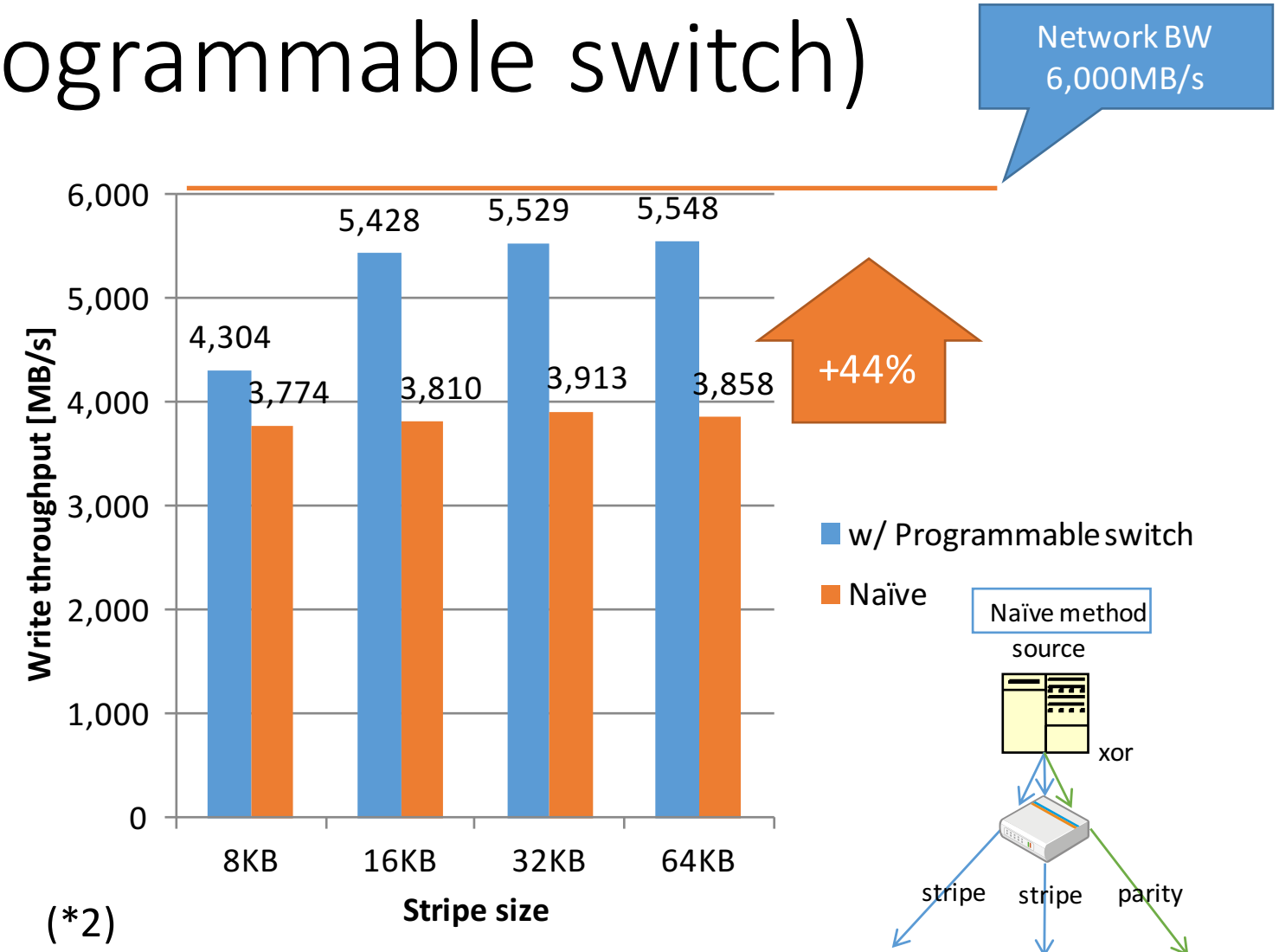
Programmable network switch (*2)

- Move the parity calculation process to network switch(s)
  - The writer sends data blocks to the programmable switch
  - The switch generates parity blocks
  - The amount of outbound traffic of the writer = the size of original data blocks
    - No traffic increase



*2 Hiroki Ohtsuji and Osamu Tatebe, “Network-based Data Processing Architecture for Reliable and High-performance Distributed Storage System”, 4th workshop on Big Data Management in Clouds (BigDataCloud), pp.1-12, 2015

Evaluation Result (w/ programmable switch)



(*2)

Adaptive strategy

- Hybrid architecture
  - Active-storage
    - Low scalability w/o additional network
    - Low-cost
  - Programmable network switch
    - Requires dedicated hardware
    - Good scalability

Conclusion and future work

- Conclusion
  - A hybrid architecture of optimized network storage system
    - Active-storage
    - Programmable network
  - Break the trade-off between performance and reliability

- Future work
  - Strategy to choose an optimal method

Acknowledgement

This works is supported by

- JST CREST “System Software for Post Petascale Data Intensive Science”,
- JST CREST “Extreme Big Data (EBD) Next Generation Big Data Infrastructure Technologies Towards Yottabyte/Year”
- JSPS KAKENHI Grant-in-Aid for JSPS Fellows “Network-oriented storage system for next generation supercomputing systems” FY2014-2015