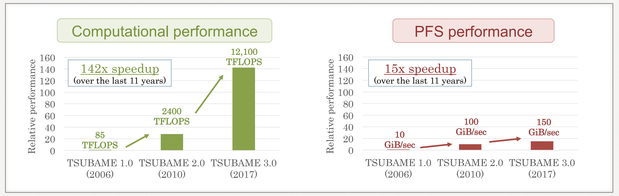
A Simulation-Based Analysis on the Configuration of Burst Buffers

Tianqi Xu†, Kento Sato‡ and Satoshi Matsuoka†
† Tokyo Institute of Technology / ‡ Lawrence Livermore National Laboratory

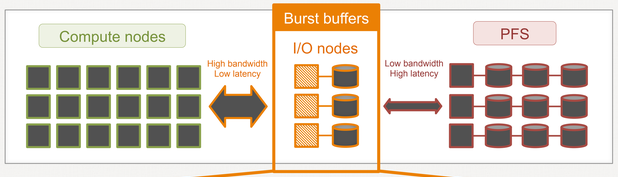
Background

Burst Buffer in HPC

- Computational performance has dramatically increased, but parallel file system (PFS) cannot catch up



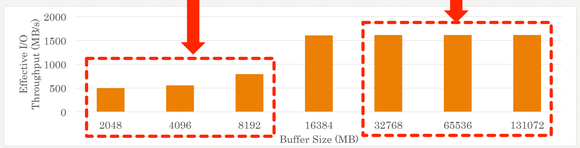
- Burst buffer (BB) systems are designed to alleviate the gap with higher performance but lower capacity



Why do burst buffer configurations matter ?

	Small-scale burst buffers	Large-scale burst buffers
Performance	Low bandwidth	High bandwidth
Capacity	Low capacity	High capacity
Cost	Low cost	High cost

Insufficient I/O throughput — Saturated effective I/O throughput

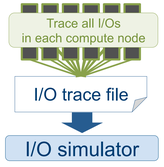


Supermovae

To find out the best trade-off between performance, capacity and cost, simulating I/O behaviors of applications in the two-level hierarchical storage is critical

Proposal

We simulate I/O behaviors of applications under different burst buffer configurations



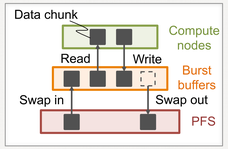
1. Trace all I/Os of parallel I/O applications in each compute node by using FUSE-based I/O tracer (MUSE). The I/O traces include read/write sizes and offsets with time stamps
2. Aggregate all I/O traces into a single file
3. Simulate I/O behaviors by using the I/O trace based on our "burst buffer model"

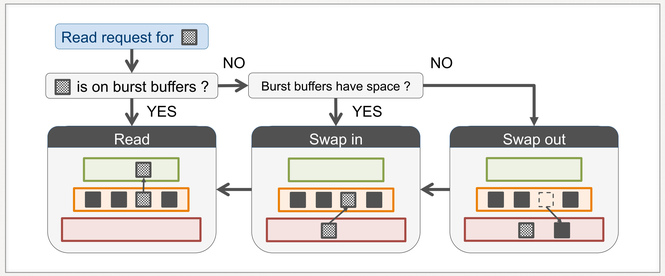
(MUSE: https://github.com/kento/MUSE)

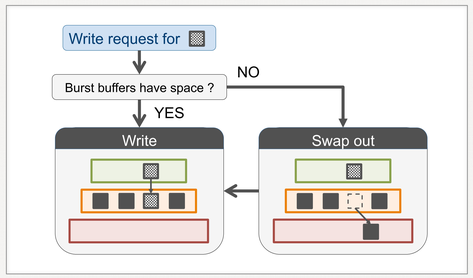
- Investigate the impact of different buffer sizes
- Estimate I/O performance of applications in different buffer sizes

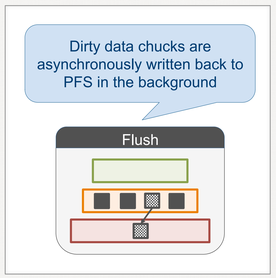
Burst buffer model

- Files are divided into "data chunks"
- Data chunks are moved between burst buffers and PFS (i.e., swap-in/out) based on a LRU algorithm
- Dirty data chunks are asynchronously flushed









Simulation Results

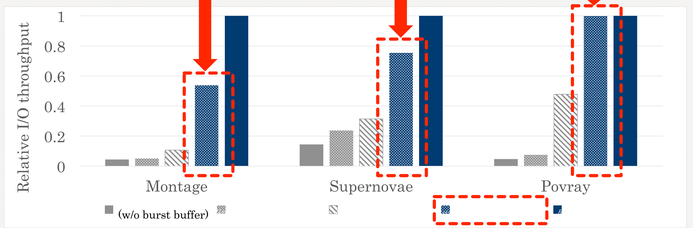
Simulating environment

System	Reedbush (@ University of Tokyo)
Burst buffer	DDN IME 14K x6 (Capacity: 209TB) **Latency: 20 us, Bandwidth 436.2 GB/s**
PFS	Lustre DDN SFA 14KE x3 (5.04 PB) **Latency 500 us, Bandwidth 145.2 GB/s**

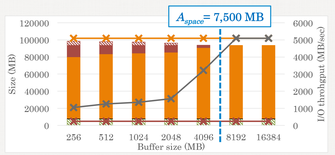
Simulating applications

Applications	Input Size (MB)	Access Space (A_{space}) (MB)	Total I/O (MB)
Montage	1,200	7,500	27,000
Supernovae	6,600	23,000	55,000
Povray	14	25	760

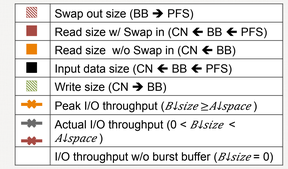
Applications can still achieve comparable I/O performance to its peak I/O performance even with half size of A_{space} of burst buffers



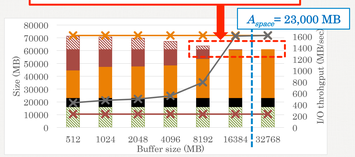
B_{isize}	Size of Burst buffers	A_{ispace}	Total I/O space that an applications access (i.e., Required storage space to run an application)



Montage: I/O operation break-down

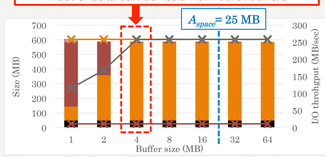


Less swap-out size contributes to the speedup



Supermovae: I/O operation break-down

Even with 4 MB of buffer size, most of data can be read from burst buffers



Povray: I/O operation break-down

Conclusion

- We investigate the performance under different buffer sizes
- Burst buffer configurations have huge impact to the performance
- We can still achieve the peak I/O performance with less than half of required storage space to run

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. (LLNL-POST-735955). This work was also supported by JST CREST Grant Number JPMJCR1303, Japan, and performed under the auspices of Real-World Big-Data Computation Open Innovation Laboratory, Japan.