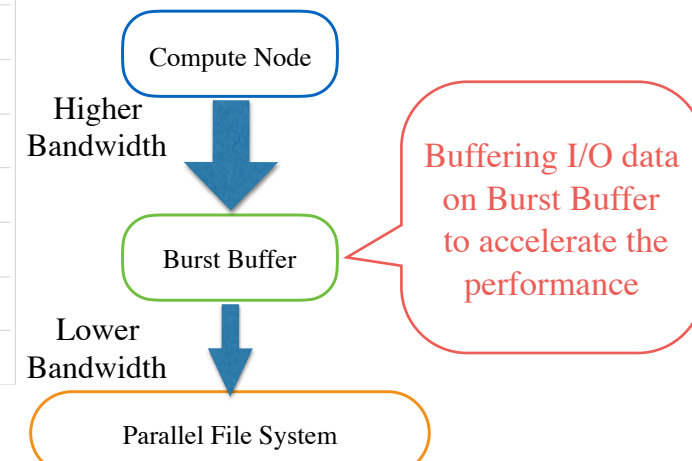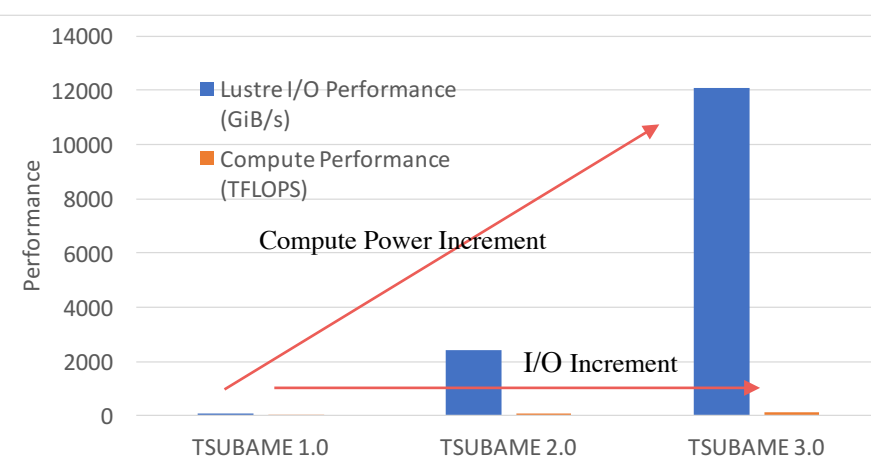# HuronFS : Hierarchical, User-level and On-demand Burst Buffer File System

Background

Static Burst Buffer in HPC

- Computational performance has dramatically increased, but parallel file systems (PFS) can not match the increment
- Burst buffer systems are designed to alleviate the gap:
  - Higher performance but lower capacity

Compute Node → Higher Bandwidth → Burst Buffer → Lower Bandwidth → Parallel File System

Buffering I/O data on Burst Buffer to accelerate the performance

Different increment between compute power & I/O
Compute : 142x speedup; I/O : 15x speedup

Problems with Current Static Burst Buffers

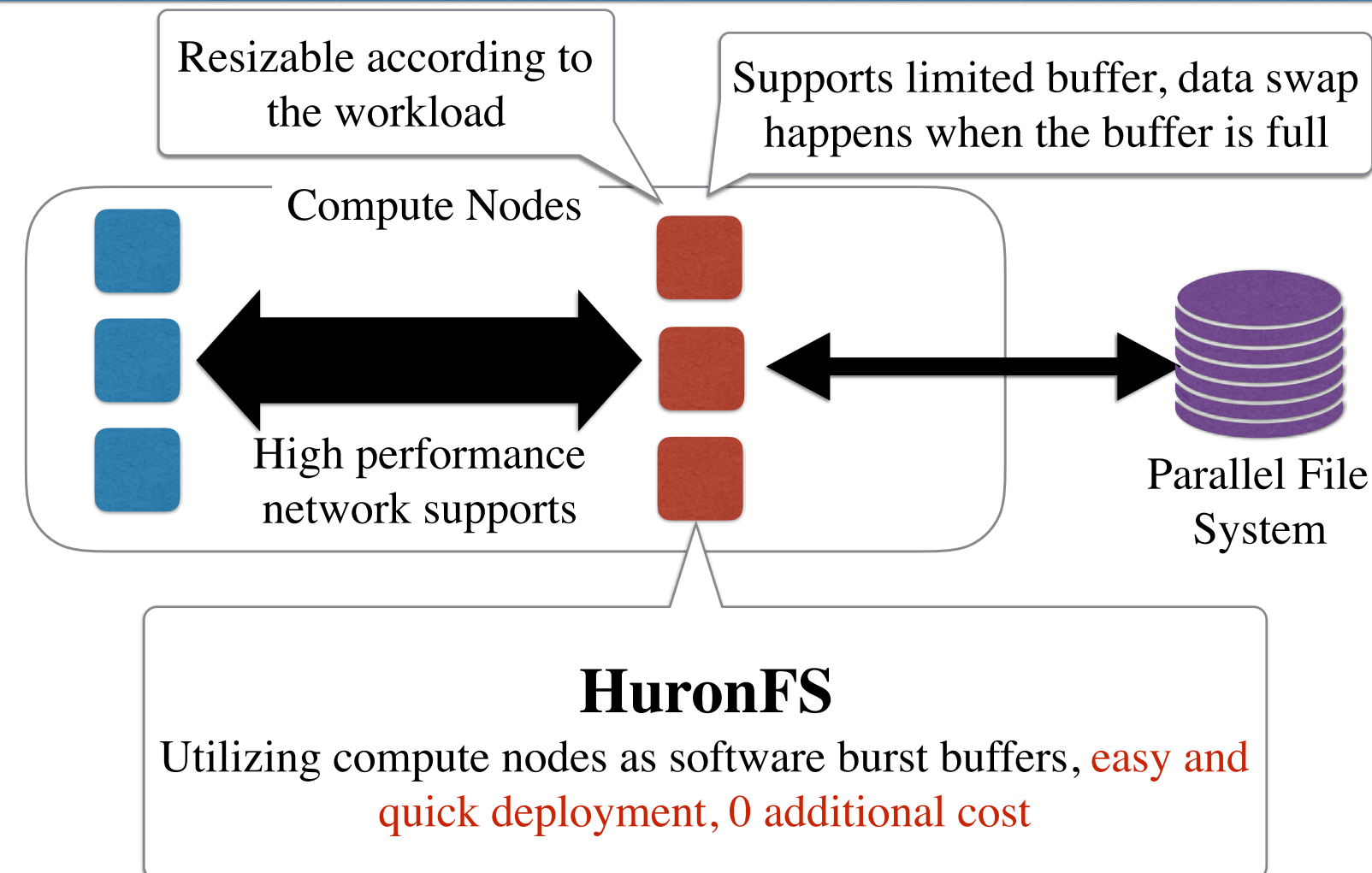| On deployment	On Submission
Requiring procurement and physical deployment.	Need to statically allocate before jobs.
The cost of having burst buffer system is high.	No extension during the jobs

- Moreover, current static burst buffer doesn't support data swap, which requires more buffer than total file size to start a job.
- Static burst buffer cannot adapt to I/O contention or workload.

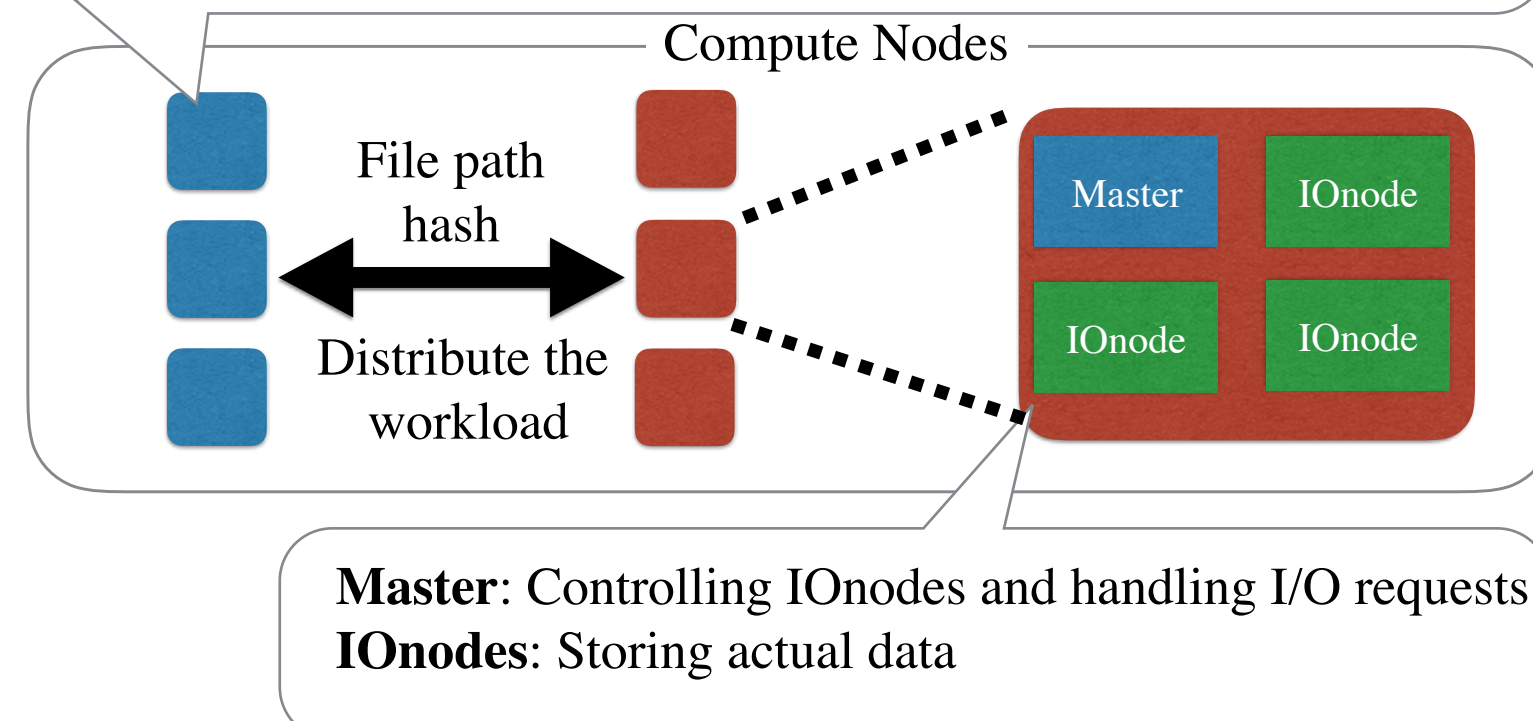We propose a software, user-level on demand burst buffer system for HPC (HuronFS) to solve these problems

Our Proposal : HuronFS

Overview of HuronFS

Resizable according to the workload

Supports limited buffer, data swap happens when the buffer is full

Compute Nodes ← High performance network supports → Parallel File System

HuronFS
Utilizing compute nodes as software burst buffers, easy and quick deployment, 0 additional cost

Overview of Architecture

- Built totally on user-level, so can be deployed by users
- **Implemented on top of FUSE, requires no code modification to use**

Compute Nodes

File path hash

Distribute the workload

Master | IOnode
IOnode | IOnode

Master: Controlling IOnodes and handling I/O requests
IOnodes: Storing actual data

Evaluation

System Name	A Supercomputer (44 nodes)
Memory	62 GB
Network	Mellanox FDR InfiniBand (54.54Gbps)
PFS	NFS (10Gbps Ethernet, 30TB)

# of Master	1
# of IOnode	1
IOnode Capacity	32GB
Client Buffer	100MB

Applications	Output File Size (MB)	Processes
Miranda IO	110 per node	4 per node (weak scaling)
BTIO	6400	64 processes (strong scaling)

Basic I/O
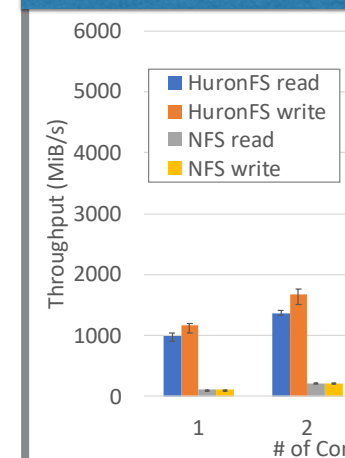
Sequential I/O

Random I/O

Benchmark	File Size
IOR	10MB per process

Benchmark	File Size	Transfer Size
IOR	1KB per process	8 Byte

Real Applications
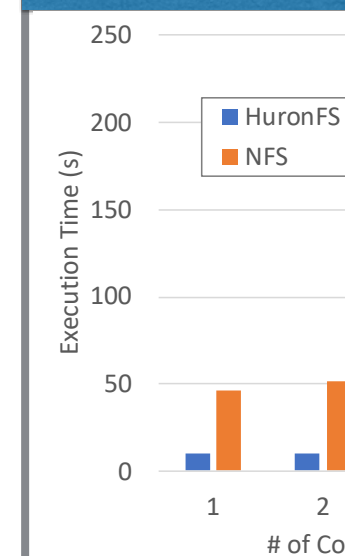
Miranda IO
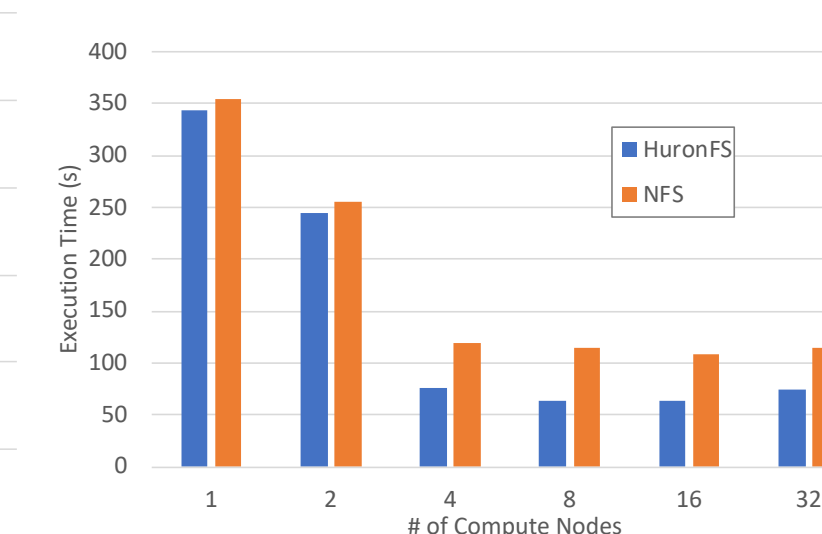
BTIO

Conclusion

- We propose a software, user-level on-demand burst buffer system for HPC
- We implemented our proposal and evaluated it on a real HPC system
- The results show that with our proposed system, applications can utilize the burst buffer to accelerate the I/O even without static burst buffer deployed.