BBAlloc: Towards Allocation based Management of Burst Buffer Systems

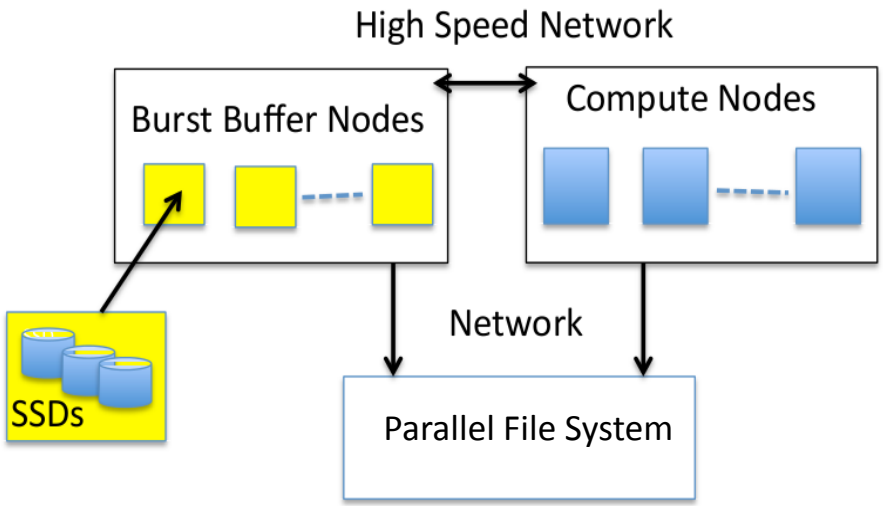
Sagar Thapaliya

PDSW 2015, Nov 16, 2015

Sagar Thapaliya, Purushotham Bangalore, Jay Lofstead, Kathryn Mohror, Adam Moody

BBAlloc for Burst Buffer (BB) Resource Management

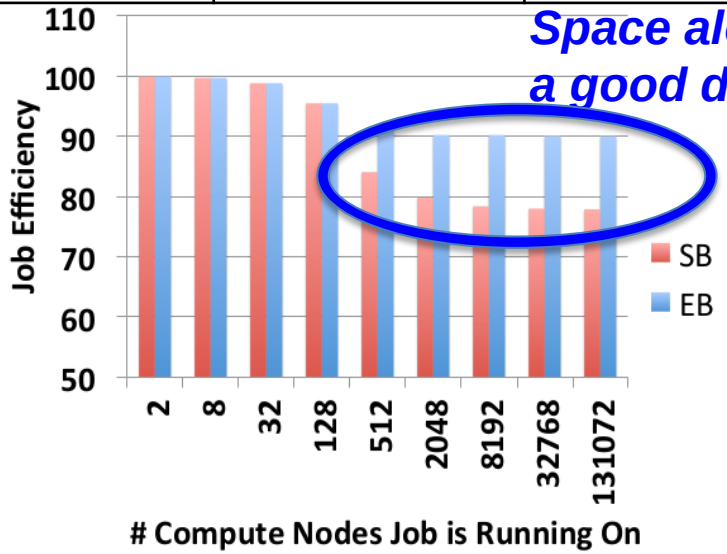
- BB Use Context
  - *Secondary storage*
  - Checkpointing, data staging
  - *Shared resource*
- Investigation
  - Management Issues
  - Right management approach

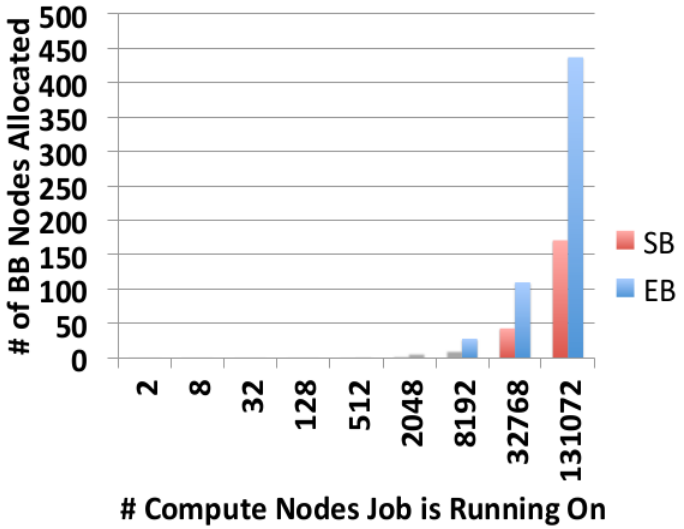


Allocation Criteria: Space v.s. Efficiency Requirement

- System configuration similar to Trinity cluster [1]
- E.g. HPC Job: write 256 MB per job process; I/O Interval: 1 hour
- Allocation (# BB nodes):
  - *Space based (SB)* : enough to store total data
  - *Efficiency based (EB)*: ComputeTime / WallClockTime * 100 → (90%)

# System BB Nodes	Compute Node Mem.	Compute Node Cores	# Compute Nodes	BB Node Space	BB Node Bandwidth
576	128 GB	32	19,000	6 TB	6 GB/s





[1] http://www.lanl.gov/projects/trinity/_assets/docs/trinity-overview-for-web.pdf

Space Sharing v.s. Time Sharing

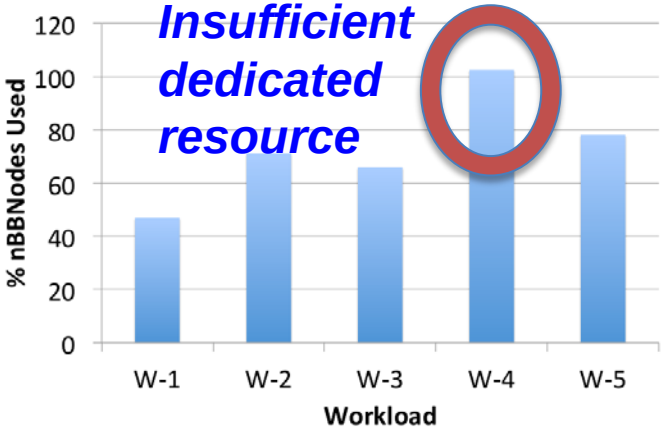
Space Sharing

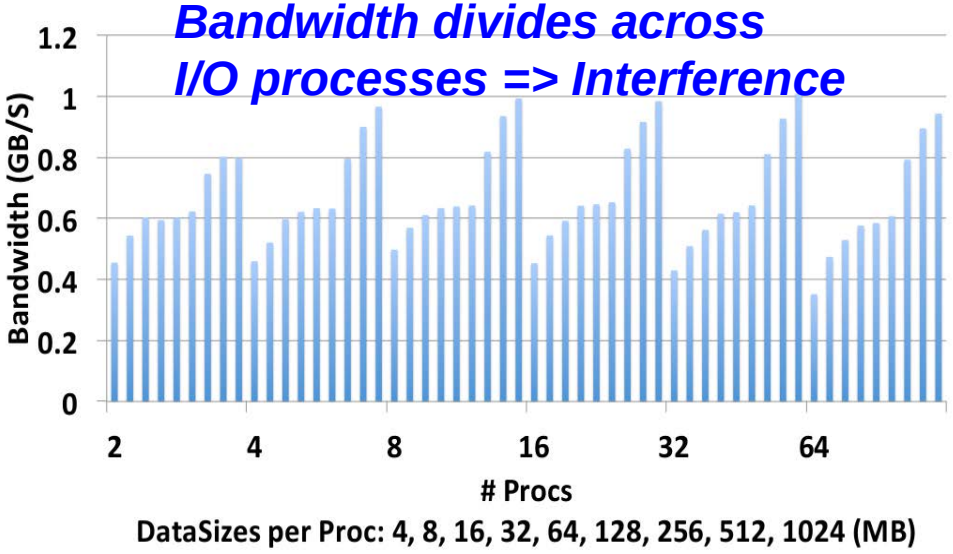
- Output size per CN: 33% of total memory
- 3 job sizes : Large, Medium, Small
  - # Compute Nodes: 4096, 1024, 10

Time Sharing

- Test bed on Catalyst Cluster [2]
- RDMA based write to BB node
- Intel 910 SSD as BB storage

W-1	W-2	W-3	W-4	W-5
1,5,20	1,10,20	2,5,20	1,10,200	1,5,200





[2] https://computing.llnl.gov/?set=resources&page=OCF_resources

Towards solution: BBAlloc

- Observation guided requirements
  - Address multiple aspects of jobs' resource requirements
  - Time sharing of BB to effectively support multiple jobs
  - Issues exist under time sharing
- BBAlloc
  - Framework to manage BB resources
  - Handle resource allocation (space, bandwidth)
  - Balance tradeoffs: job and whole system optimization

Thank You!

Sagar Thapaliya
sagar@uab.edu