OAK RIDGE National Laboratory

Summit Burst Buffer

Christopher Zimmer

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

Summit Storage Options

- Parallel File System
  - Spider-3 center-wide GPFS
  - 250 PB @ 2.5 TB/s
  - ~540 MB/s write performance per node when all nodes are writing

- Burst Buffer
  - 4,608 nodes with NVMe SSDs (Samsung PM1725a)
  - 7.3 PB Total
  - 9.67 TB/s aggregate write 27 TB/s aggregate read performance when using all nodes

What's a Burst Buffer?

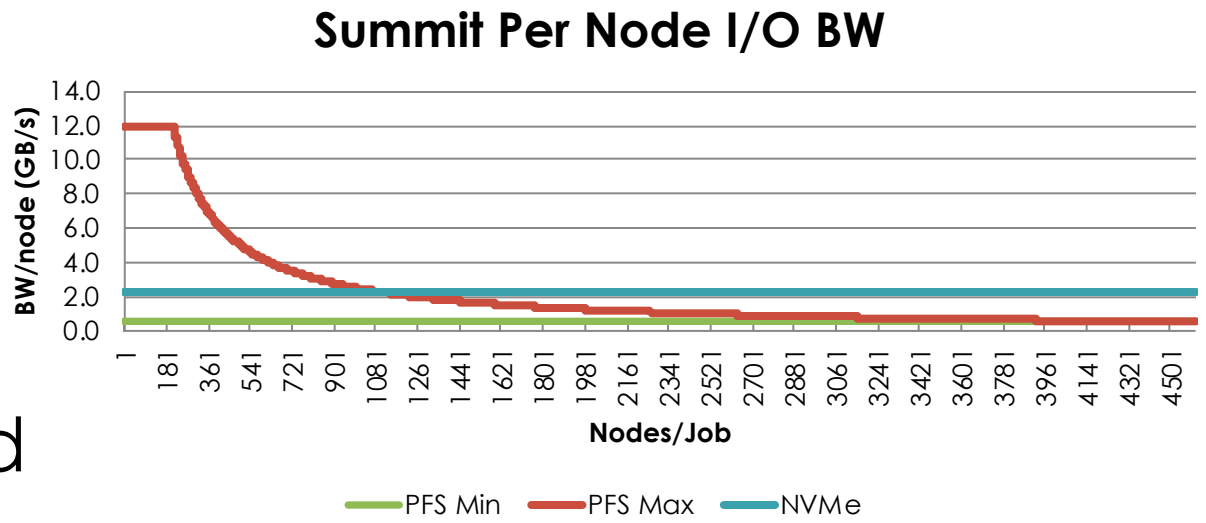
- Originally: A combination of software and hardware to accelerate phased periodic I/O
  - E.g. Applications checkpointing hourly
- Why it helps
  - The aggregate Summit NVMe's have ~4X more write bandwidth than the PFS and a larger factor more meta-data create performance.
  - Goal: shrinking a 5 minute hourly I/O phase for a 24 hour job to 2 minutes
    - Reduces I/O from 8% of application runtime to 3%
      - In early testing the meta-data performance improvement is even greater

Other NVMe Uses

- Machine Learning Training:
  - Each PM1725A offers 1million 4K reads per second
  - 1.6 TB for datasets
- Scratch space for temporary files
- Extended memory via mmap
  - Storage to reduce memory pressure for infrequently accessed data

When to use the Burst Buffers (Node Scale)

- Alpine GPFS Performance
  - Per node 12-14 GB/s (Without core isolation)
  - Aggregate 2.5 TB/s
    - Full system job will achieve 550 MB/s per node
- Node Local NVME
  - Samsung PM1725A
    - Write 2.1 GB/s
    - Read 5.5 GB/s
  - Scales linearly with Job Size
- Realistically benefit is realized
  - 150 Nodes



When to use continued:

- 24K Files to GPFS (4096 Nodes)
  - 24 TB of data written
  - 500 seconds spent creating and writing files
    - Most time spent in creation

- High IOP read workload (Full System)
  - 4k Random Reads (GPFS) ~100million
  - 4k Random Reads (NVMe) ~4.5 Billion
    - 1M per device

Using a Burst Buffer in a Job

- Interactively:
- bsub -Is -nnodes 1 -PSTF008 -W00:10 -alloc_flags "nvme" /bin/bash
  - jsrun -r1 df
    - /dev/mapper/bb-bb1 1452706772 33040 1452673732 1% /mnt/bb/cjzimmer
- Batch:
  - #!/bin/bash -l
  - #BSUB -P STF008
  - #BSUB -W 01:00
  - #BSUB -nnodes 1
  - #BSUB -alloc_flags "gpumps smt4 **nvme**"

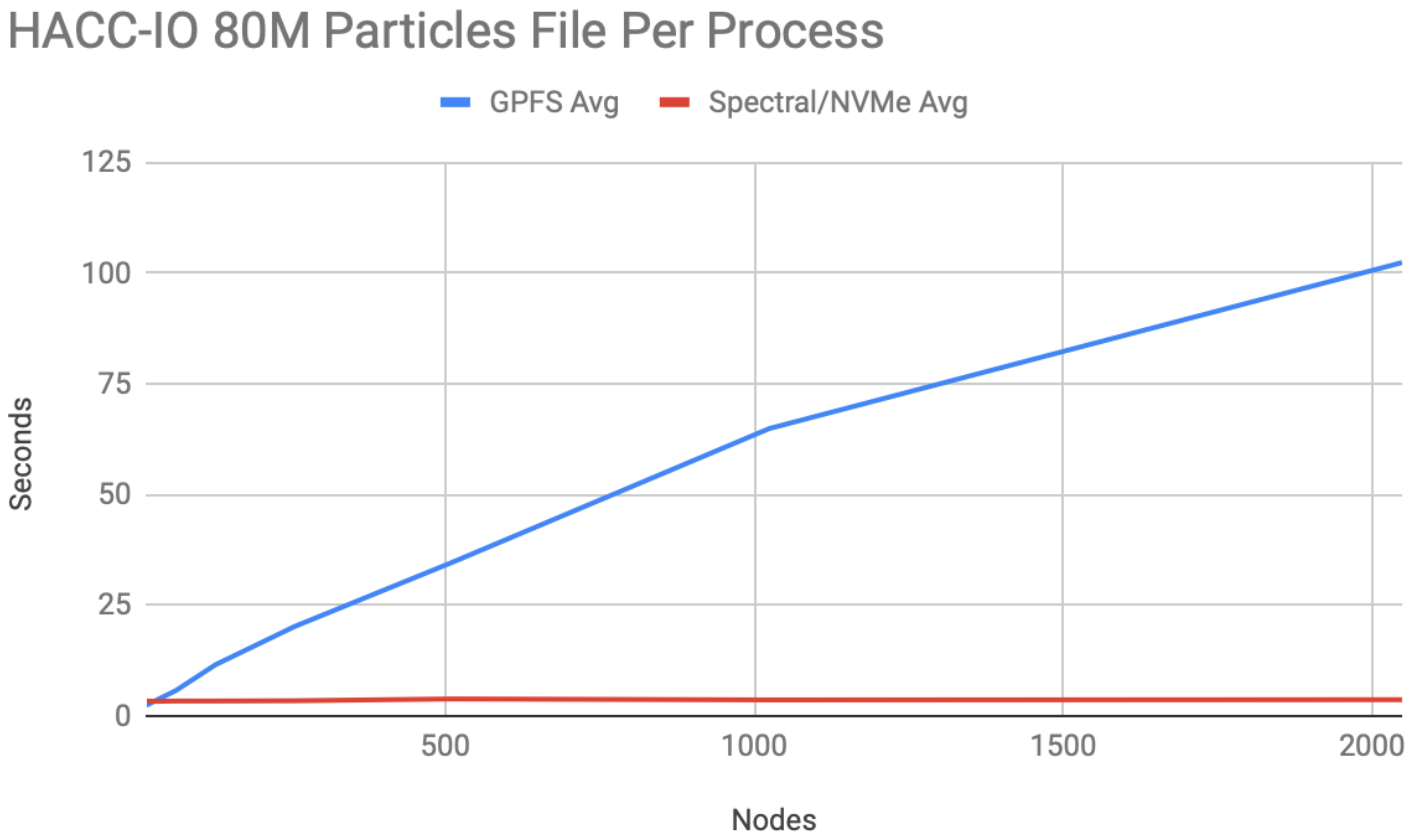
Other ways of using the burst buffers

- OLCF Provides Spectral:
  - Tranparent
  - No code changes
  - Automatically detects checkpoint files
    - Stages them to the burst buffer
    - Transfers them to the file-system upon close

- More in-depth presentation tomorrow

Performance

- Measured up to 2048 nodes



Thank you!

- Questions?