OAK RIDGE National Laboratory

Frontier NVMe

Christopher Zimmer

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

U.S. DEPARTMENT OF ENERGY

Frontier Storage Levels

- Three Levels of Storage in Frontier
  - Capacity Tier
    - Lustre
    - ~700PB Capacity 2-5 TB/s
  - Performance Tier
    - Lustre
    - ~10PB Capacity 10 TB/s
  - Node Local NVMe
    - XFS
    - ~32PB Capacity 37 TB/s Write 75 TB/s read

Open slide master to edit

Frontier Configuration

- Two Drives Per Node
  - LVM Stripe Configuration
  - 3.5 TB Capacity Per Node
  - ~4000MB/s Seq Write
  - ~9000MB/s Seq Read
- Allocating NVMes on Frontier/Crusher
  - Use "-C nvme" in sbatch/salloc
  - E.g. "salloc -C nvme -N1 -Astf008 -t00:30:00 /bin/bash"
  - Drives mount as "/mnt/bb/username" <- Your username
    - "/dev/mapper/nvme-bb 3561456576 3582172 3557874404 1% **/mnt/bb/cjzimmer**"

Open slide master to edit

NVME Use Cases (Machine Learning Training)

- High IOP Workloads
  - Machine Learning / AI Training
  - Workload iterates over data set reading and re-reading at high frequency
- Using the NVMe for this Workload
  - Traditionally – Shard the dataset into smaller chunks
  - Copy the chunks to the NVMe
  - Train on node-local Chunks

Open slide master to edit

NVME Use Cases (Checkpointing)

- Checkpointing on Frontier will be more important than previous OLCF machines
  - Scale + Technology Trends
- Checkpointing to PFS will work for most
- Certain applications may prefer to checkpoint to NVMe to reduce time in I/O.
  - Typical large volume file-per-process checkpoints

NVME Use Cases (Extending Memory)

- Node local NVMe devices can extend memory
  - See mmap MAP_FILE
- Purpose
  - Extend virtual address space to NVMe device
  - Much slower
  - Can aid in large in-memory database
  - Not frequently used

Open slide master to edit

Relevant Software

- High Velocity AI Cache (HVAC)
  - New software coming to Frontier to ease ML Training on NVMe
  - Eliminates the need to shard the dataset using the NVMes as a distributed cache
- UnifyFS
  - NVMe overlay file-system for checkpointing and restart.
- Spectral
  - Checkpointing software automatically abstracts NVMes. Write-only data no re-reading.

Relevant Software cont.

- Scalable Checkpoint Restart (SCR)
  - Handles hierarchical checkpointing and automatic restarting of failed applications

OAK RIDGE National Laboratory
Open slide master to edit

Thank you!

- Questions?

- Email : zimmercj@ornl.gov

OAK RIDGE National Laboratory

Open slide master to edit