Oak Ridge National Laboratory

An introduction to efficient I/O on Summit

Sarp Oral, PhD

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

U.S. DEPARTMENT OF ENERGY

Introduction

- Goal
  - Familiarize participants with Summit and its I/O subsystem
- Not by any means an exhaustive source of information
- Always check
  - https://olcf.ornl.gov
  - https://docs.olcf.ornl.gov/training/training_archive.html (training docs)
  - https://vimeo.com/olcf (Video channel for tutorials)
- If need more help
  - help@olcf.ornl.gov or 865-241-6536

ORNL Summit System Overview

System Performance

- Peak of 200 Petaflops (FP_{64}) for modeling & simulation
- Peak of 3.3 ExaOps (FP_{16}) for data analytics and artificial intelligence

The system includes

- 4,608 nodes
- Dual-rail Mellanox EDR InfiniBand network
- 250 PB IBM GPFS file system transferring data at 2.5 TB/s

Each node has

- 2 IBM POWER9 processors
- 6 NVIDIA Tesla V100 GPUs
- 608 GB of fast memory
  (96 GB HBM2 + 512 GB DDR4)
- 1.6 TB of non-volatile memory (Samsung PM1725A)

Primary Allocation Programs for Access to LCF

For more information, or to apply, please see the following links:

General Info on User Programs	https://www.olcf.ornl.gov/for-users/getting-started/#request-allocation
INCITE	http://www.doeleadershipcomputing.org/
ALCC	https://science.osti.gov/ascr/Facilities/Accessing-ASCR-Facilities/ALCC
DD	https://www.olcf.ornl.gov/for-users/documents-forms/olcf-directors-discretion-project-application/

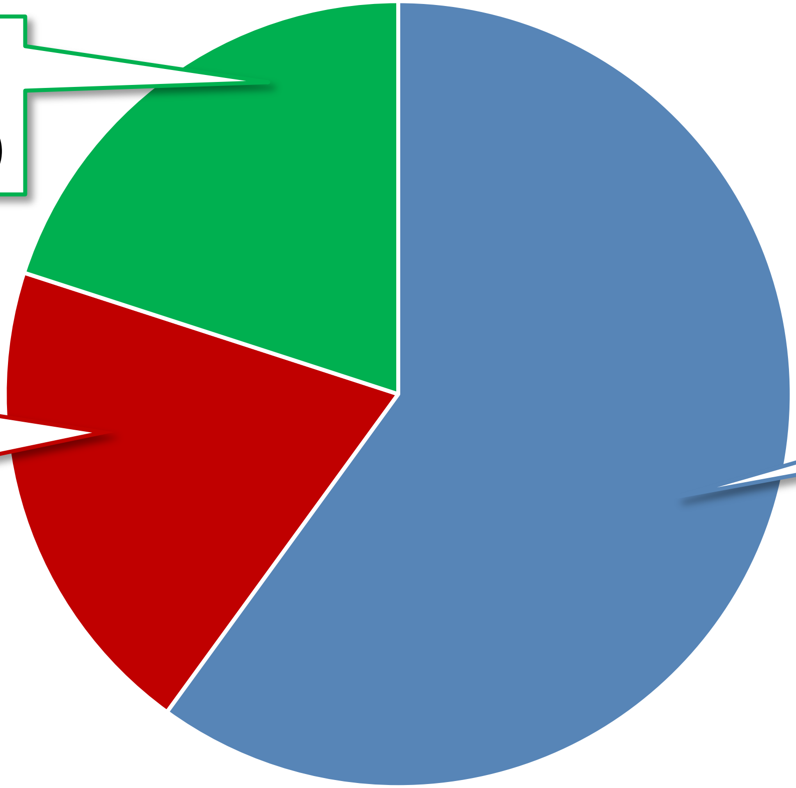
Primary Allocation Programs for Access to LCF

Current distribution of allocable hours

INCITE (Open Science)
- Large allocations (100,000s of node-hours)
- Solve most challenging problems in science and engineering
- Single simulations use 20%> of system
- Annual call for proposals

20% Director's Discretionary

(includes LCF strategic programs, ECP)



60% INCITE

Leadership-class computing

20% ALCC

ASCR Leadership Computing Challenge

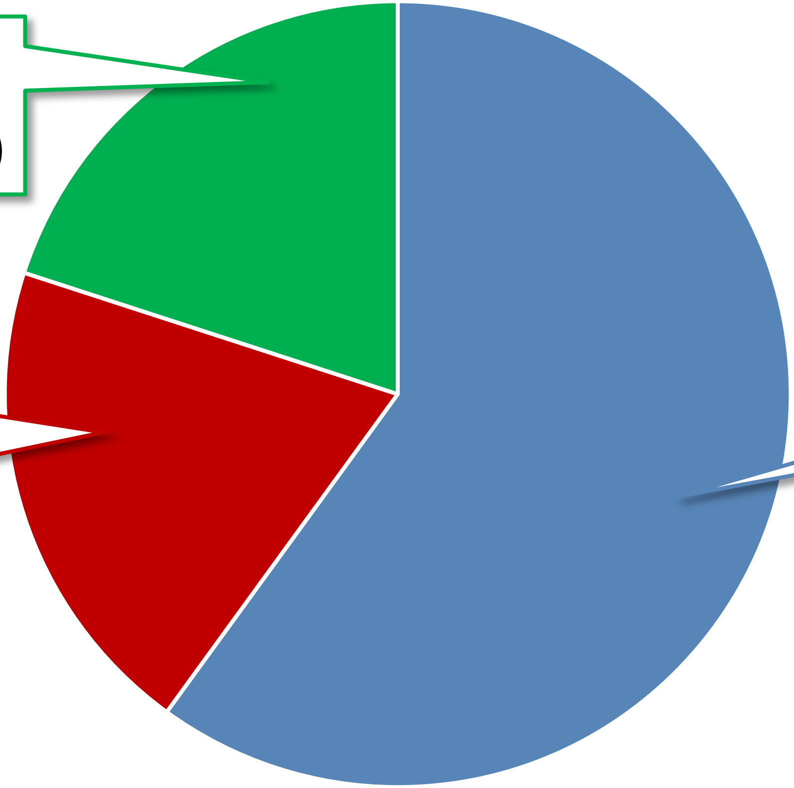
DOE/SC capability computing

Primary Allocation Programs for Access to LCF

Current distribution of allocable hours

ALCC (DOE Mission Science)
- Large allocations (100,000s of node-hours)
- Projects inline with mission of DOE
- Annual call for proposals

20% Director's Discretionary
(includes LCF strategic programs, ECP)



60% INCITE

Leadership-class computing

20% ALCC
ASCR Leadership Computing Challenge

DOE/SC capability computing

Open slide master to edit

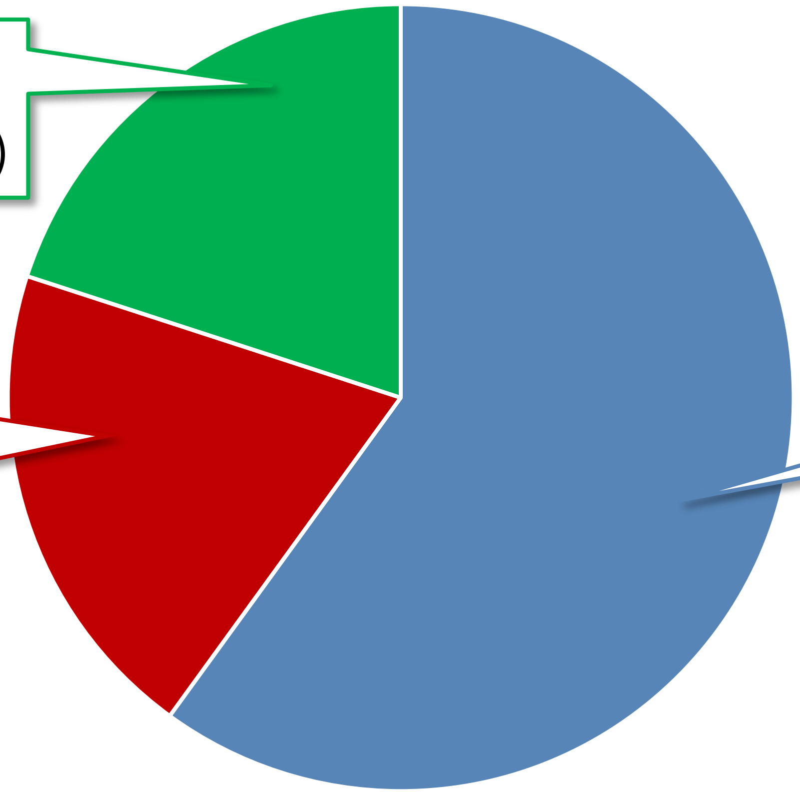
Primary Allocation Programs for Access to LCF

Current distribution of allocable hours

DD
- Smaller allocations (1000s -10,000s of node-hours)
- Intended as onramp for new projects / ECP
- Preparation for larger allocation programs
- Proposals accepted year round

20% Director's Discretionary

(includes LCF strategic programs, ECP)



60% INCITE

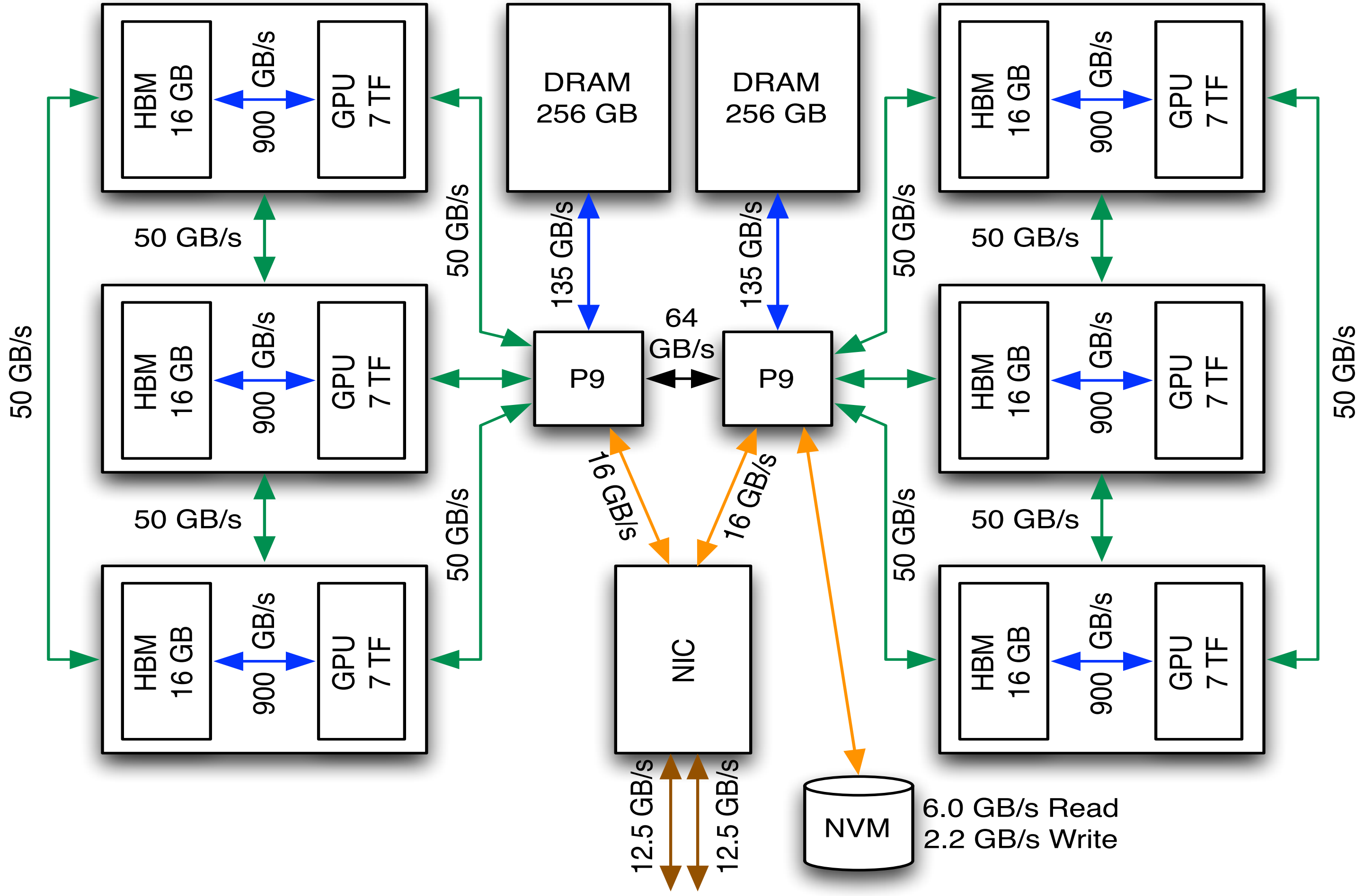
Leadership-class computing

20% ALCC

ASCR Leadership Computing Challenge

DOE/SC capability computing

Summit Node Schematic



TF	42 TF (6x7 TF)
HBM	96 GB (6x16 GB)
DRAM	512 GB (2x16x16 GB)
NET	25 GB/s (2x12.5 GB/s)
MMsg/s	83

- HBM/DRAM Bus (aggregate B/W)
- NVLINK
- X-Bus (SMP)
- PCIe Gen4
- EDR IB

HBM & DRAM speeds are aggregate (Read+Write).
All other speeds (X-Bus, NVLink, PCIe, IB) are bi-directional.

- Coherent memory across entire node
- NVLink v2 fully interconnects three GPUs and one CPU on each side of node
- PCIe Gen 4 connects NVM and NIC
- Single shared NIC with dual EDR ports

Comparison of Titan, Summit, and Frontier Systems

System Specs	Titan	Summit	Frontier
Peak	27 PF	200 PF	~1.5 EF
# cabinets	200	256	> 100
Node	1 AMD Opteron CPU 1 NVIDIA K20X Kepler GPU	2 IBM POWER9™ CPUs 6 NVIDIA Volta GPUs	1 HPC and AI Optimized AMD EPYC CPU 4 Purpose-Built AMD Radeon Instinct GPU
On-node interconnect	PCI Gen2 No coherence across the node	NVIDIA NVLINK Coherent memory across the node	AMD Infinity Fabric Coherent memory across the node
System Interconnect	Cray Gemini network 6.4 GB/s	Mellanox Dual-port EDR IB network 25 GB/s	Cray four-port Slingshot network 100 GB/s
Topology	3D Torus	Non-blocking Fat Tree	Dragonfly
Storage	32 PB, 1 TB/s, Lustre Filesystem	250 PB, 2.5 TB/s, IBM Spectrum Scale™ with GPFS™	2-4x performance and capacity of Summit's I/O subsystem.
Near-node NVM (storage)	No	Yes	Yes

Available I/O modules on Summit

- ADIOS
  - adios2/2.4.0 adios2/2.5.0 (D)
- HDF5
  - hdf5/1.8.18 hdf5/1.10.4 (D)
- NetCDF/PNetCDF
  - netcdf-cxx4/4.3.0 netcdf-fortran/4.4.4 netcdf/4.6.1 netcdf/4.6.2 (D) parallel-netcdf/1.8.1
- Darshan
  - darshan-runtime/3.1.7-hdf5pre110 darshan-runtime/3.1.7 (L,D) darshan-util/3.1.7 (D)
  - darshan-runtime/3.1.7-hdf5post110 darshan-util/3.1.6 darshan-util/3.2.1

Summit Storage Options

- Alpine Parallel File System (Spider-3)
  - Center-wide IBM SpectrumScale, single POSIX namespace
  - 250 PB usable formatted capacity
  - 2.5 TB/s sequential write; 2.2 TB/s random write
  - ~540 MB/s write performance per node when all nodes are writing
- Burst Buffer
  - 4,608 nodes with NVMe SSDs (Samsung PM1725a)
  - At scale (using all nodes)
    - 7.3 PB Total
    - 9.67 TB/s aggregate write
    - 27 TB/s aggregate read

Alpine Center-wide parallel file system

- Spider 3/Alpine
  - POSIX namespace, shared center-wide
  - Purged, 90-day window, not backed up
  - IBM SpectrumScale/GPFS
    - 77 ESS GL4, w/ O(30K) 10TB NL-SAS
    - IB EDR connected
  - 250 PB usable, formatted
    - ~90x of 2.8 PB DDR+HBM of Summit
  - 2.5 TB/s aggregate sequential write/read
  - 2.2 TB/s aggregate random write/read
  - 800K/s 32KB file transactions
    - create/open+write+close
  - ~30K 0B file create in a shared directory

- Each GL4
  - 2 P9 based NSD servers
  - 4 106 slot disk enclosures
  - 12 Gbps SAS connected (NSD – enclosure)
  - 422 disks in total organized in 2 distributed RAID sets
- Each NSD
  - 2 IB ConnectX-5 EDR ports connect to Summit
  - 2 IB ConnectX-5 EDR ports connect to the rest of OLCF

Summit burst buffer layer

- Each Summit compute node can write @ 12.5 GB/s to Alpine
  - Max out Alpine w/ 200 Summit compute nodes
- Each Summit node has a 1.6 TB Samsung PM1725a NVMe, <u>exclusive</u>
  - 6 GB/s read and 2.1 GB/s write I/O performance
  - 5 drive writes per day (DWPD)
  - Formatted as XFS (node-local file system)
  - Reformatted at the end of each job
- In aggregate Summit burst buffer layer
  - 7.4 PB @ 26.7 TB/s read and 9.7 TB/s aggregate write I/O performance
  - 4.6 billion IOPS in aggregate
  - 2.5 times the capacity of aggregate system DRAM and HBM

What is a Burst Buffer?

- An additional storage layer (hardware and software) to cater low-latency, high-bandwidth needs of applications, in a cost-effective manner
  - Parallel file systems provide high-bandwidth at the expense of latency
    - POSIX consistency semantics
- Burst buffer architectures
  - In-node (a.k.a *compute node local* or *node local*)
  - In-rack (not many examples yet)
  - In-system (a rough example might be DDN's IME)
- Summit has an in-node burst buffer layer

Why do we need a burst buffer?

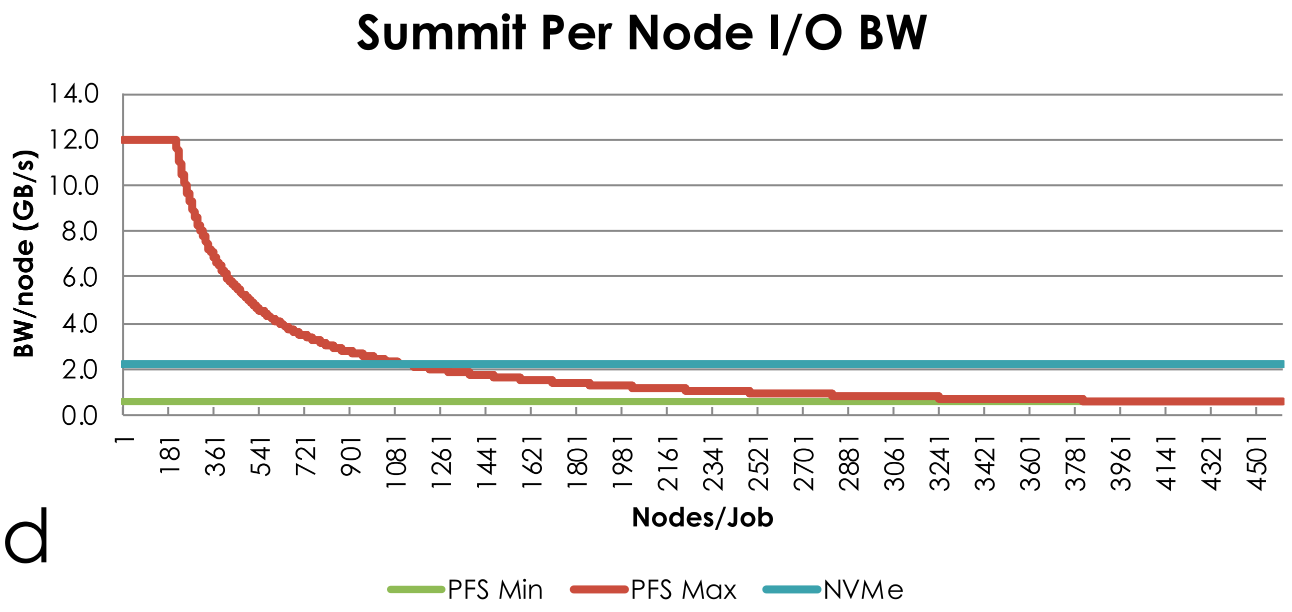
- Traditional modelling and simulation applications periodically write out their memory state
  - Rule of thumb, once every hour; X% of the memory
    - At OLCF our analysis show majority of applications write at most 15% of memory
    - Depends on the domain and application
    - Time series data dump can't be discarded
    - Checkpoint data can be reduced greatly (90%)
- ML/DL applications are even more demanding in terms of I/O
- We need a low-latency, cost-effective storage solution for ALL applications

How does a burst buffer help applications?

- On Summit NVMe’s in aggregate have ~4X more write bandwidth than Alpine (9.7 TB/s vs. 2.5 TB/s)
- The aggregate performance linearly increases with respect to increasing number of nodes
  - Low-latency and exclusive access with no shared resource contention resulting in very high small I/O and metadata performance

When to use a Burst Buffer (In-node architecture)?

- Alpine Performance
  - Per node 12-14 GB/s (Without core isolation)
  - Aggregate 2.5 TB/s
    - Full system scale 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



Pros and cons of burst buffer architectures

- In-node
  - Lowest access latency and bandwidth scales linearly
  - Most difficult to use
  - Cheapest solution, no need for extra hardware resources (e.g., servers, networking gear)
- As we move away from the node the access latency increases (in-rack or in-system) however, usage can become easier, while the cost increases

So, what is the problem then?

- How to aggregate and present in-node hardware storage devices at-scale in software as an effective I/O solution?
- Balance/optimize the performance (latency and bandwidth), capacity, ease of use, and cost

POSIX is <u>NOT</u> dead (and won't be for a long while)

- Almost all applications (at least at OLCF) are still asking for a POSIX shared namespace
  - easy to use,
  - doesn't require any application changes,
  - protects an application against its own harmful I/O patterns (e.g., overwrites) at the cost of heavy locking and synchronization

So, what is the problem then?

- Therefore as successful burst buffer software solution should
  - Resemble POSIX
    - relaxed perhaps, not in the strict consistency semantics sense, to keep the latency low
    - while providing a logical shared namespace abstraction on top of the physically distributed in-node storage devices

Acknowledgements

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725

Some of the contents are taken from:

“Summit Burst Buffer”, Christopher Zimmer, OLCF 2020 User Training
“Burst Buffer on Summit”, George S. Markomanolis, OLCF 2020 User Traning
“OLCF GPU Hackathon”, Tom Papatheodore, OLCF 2019 Training
“OLCF Overview for New Users”, Bill Renaud, OLCF 2020 Training