OAK RIDGE National Laboratory

A Step Towards the Final Frontier: Lessons Learned from Acceptance Testing of the First HPE/Cray EX 3000 System at ORNL

Verónica G. Melesse Vergara
Reuben Budiardja
Paul Peltz
Jeff Niles
Christopher Zimmer
Daniel Dietz
Christopher Fuson
Hong Liu
Paul Newman
James Simmons
Christopher Muzyn

Oak Ridge National Laboratory

Cray User Group 2021 (May 3, 2021)

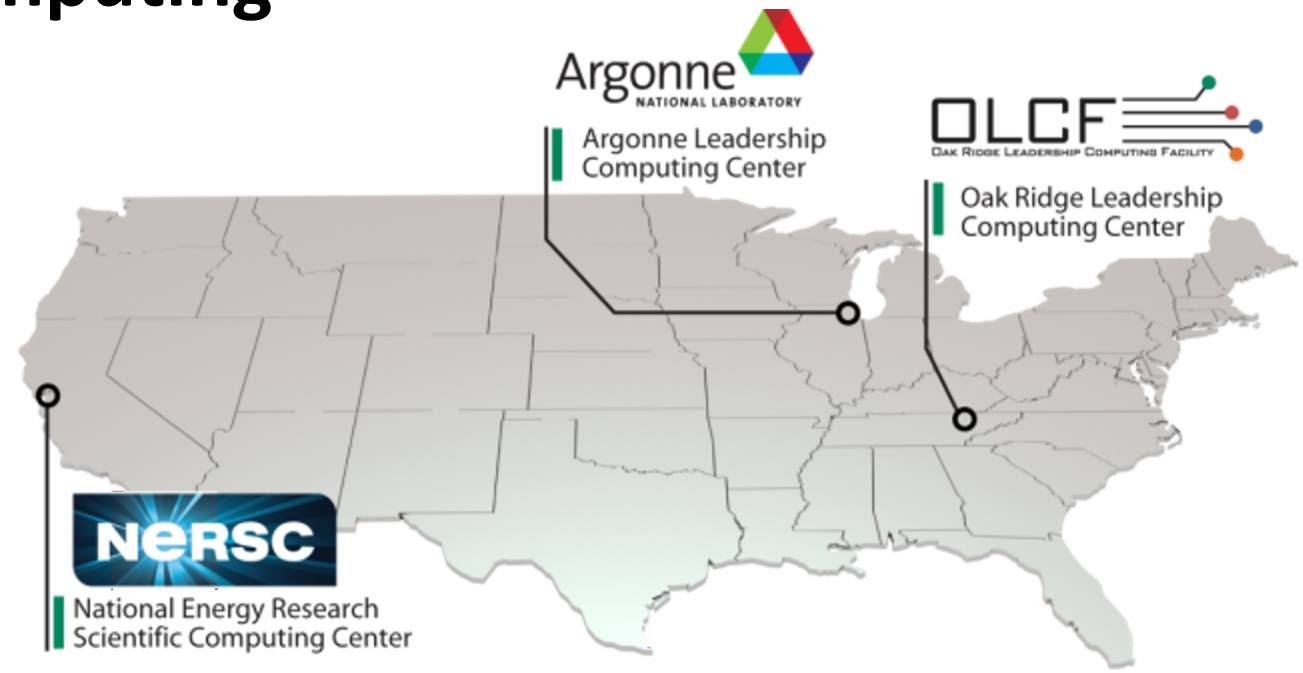
ORNL is managed by UT-Battelle, LLC for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

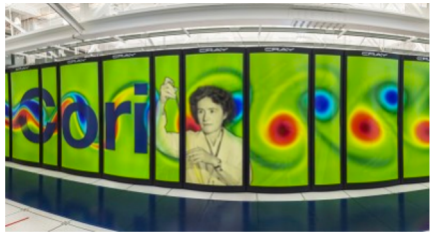
Outline

- The Oak Ridge Leadership Computing Facility (OLCF)
- Air Force Weather HPC11 system
- HPC11 Acceptance Testing
- HPC11 Compute Acceptance
- HPC11 Storage Acceptance
- Conclusions

The U.S. Department of Energy Office of Science and its role in computing



- DOE is leader in open High-Performance Computing
- Provide the world's most powerful computational tools for open science
- Access is free to researchers who publish
- Boost US competitiveness
- Attract the best and brightest researchers



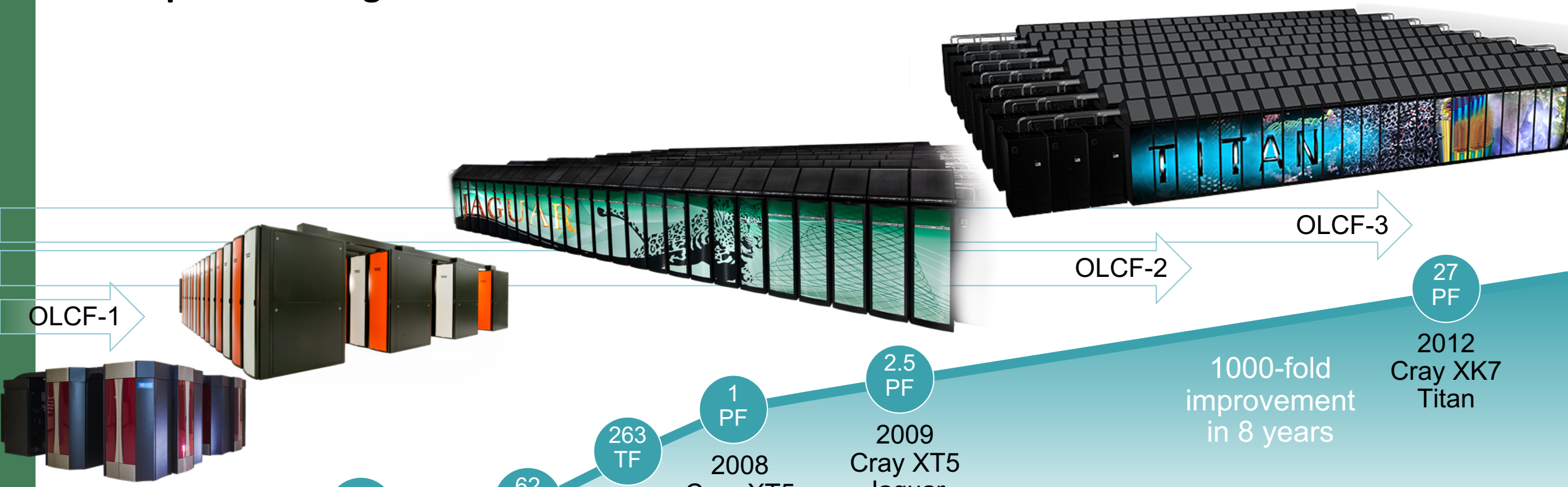
NERSC
Cori is 30 PF

ALCF
Theta is 12 PF

OLCF
Summit is 200 PF

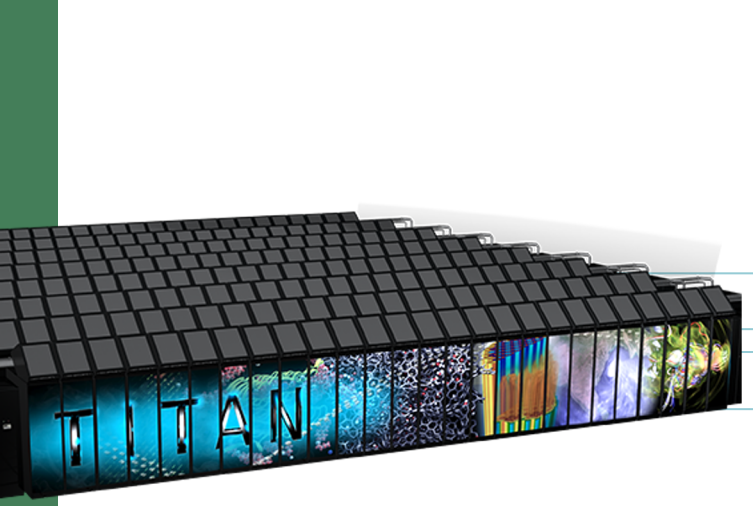
ORNL has delivered a series of leadership-class systems

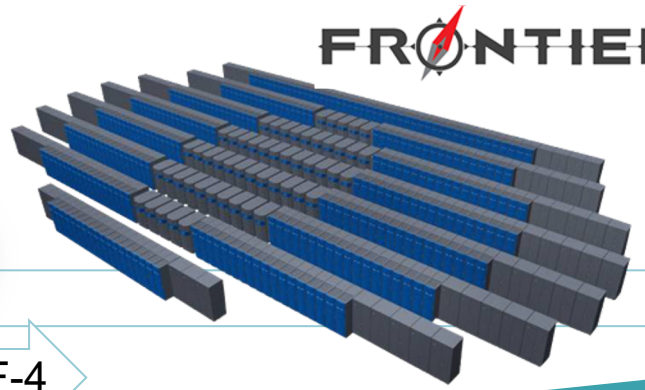
On scope • On budget • Within schedule



Titan, delivered world-class science research in support of our user community for over six years. It was decommissioned in August 2019.

We are building on this record of success to enable exascale in 2021





OLCF-4

OLCF-5

27 PF

2012
Cray XK7
Titan

200 PF

2018
IBM
Summit

500-fold improvement in 9 years

~1 EF

2021
Frontier

ORNL and Air Force Weather Strategic Partnership

- Interagency partnership between US Air Force and US Department of Energy's Oak Ridge National Laboratory
- Provide a high performance weather forecasting computer system
- System will primarily support work by the US Air Force Weather Wing
- First installation of the HPE Cray EX supercomputer in a federal facility.

Source: https://www.olcf.ornl.gov/2021/02/09/us-air-force-ornl-launch-next-generation-global-weather-forecasting-system/
Image credit: Jason Smith/ORNL, HPE Cray

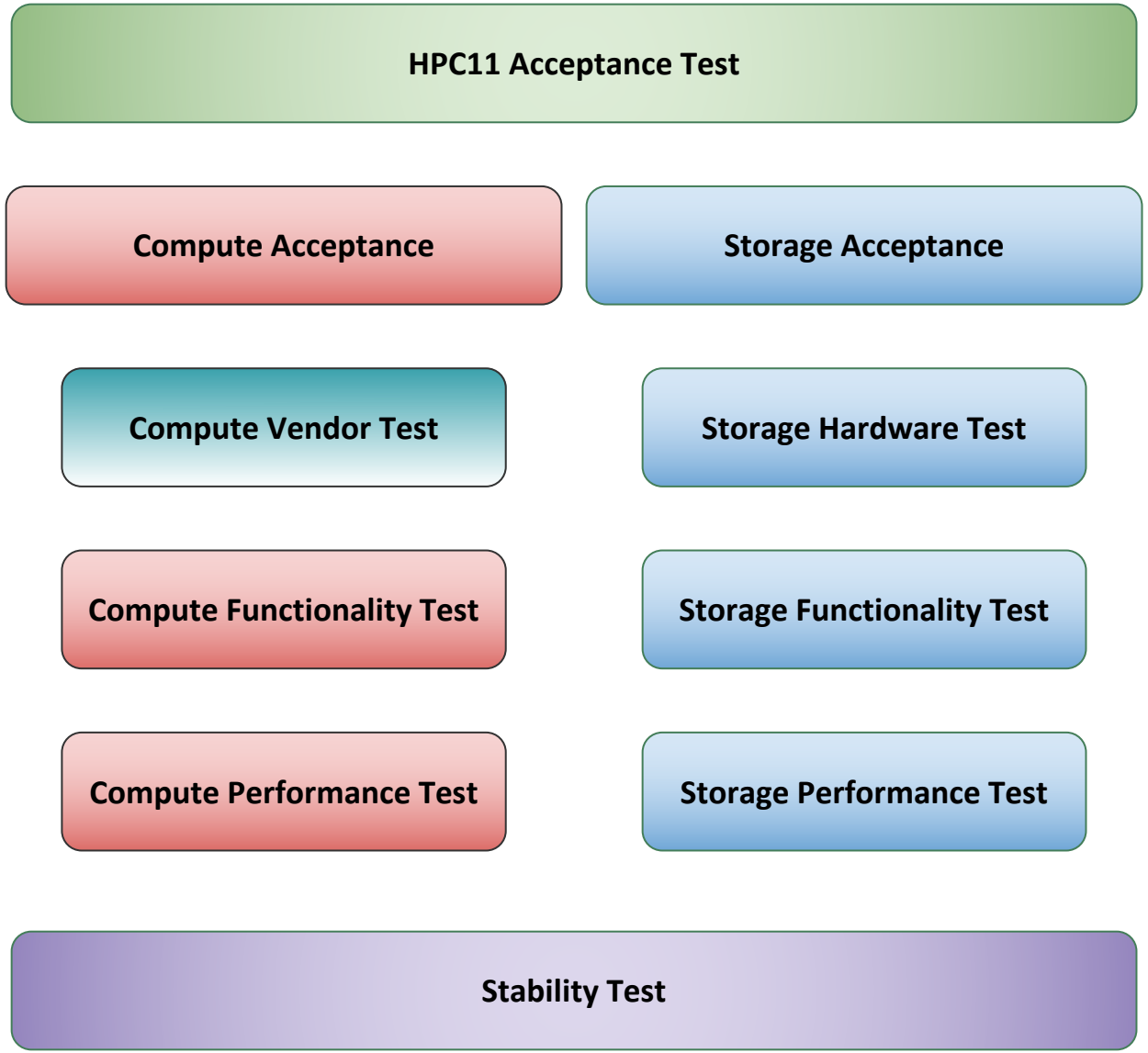
OAK RIDGE National Laboratory

HPC11 System Acceptance

HPC11 Miller and Fawbush

- Air Force Weather (AFW) HPC11 compute resource consists of two identical, independent compute systems:
  - Miller and Fawbush
  - Each with 800 compute nodes
    - Two 64-core AMD Rome CPU
    - 256 GB of memory
  - 100 Gbit Slingshot-10 interconnect
- Supported by two identical, independent file systems:
  - Storm and Cyclone
  - Lustre parallel file systems
- Software stack
  - HPCM for system software management
  - SLURM scheduler
  - Cray Programming Environment

HPC11 Acceptance Test



HPC11 Compute Acceptance Test: Vendor Test

- Executed by the vendor with results provided to ORNL
- Includes:
  - Hardware diagnostics
  - Contractual benchmarks:
    - UM 10.9
    - 4DVAR
  - High Performance LINPACK
  - STREAM

HPC11 Compute Acceptance Test: Functionality Test

- Ensures individual components of the hardware and software stack are working correctly
- Allows for verification of realistic workloads
- Precedes performance testing
- Includes:
  - System Administration: cold and warm boot of the full system, failure injection, telemetry data capture, among others
  - Network test: injection bandwidth per node, latency, global bandwidth
  - Scheduler and job launching tests: SLURM layout, job federation
  - Component tests: HPL, STREAM
  - Programming Environment tests: compilers, MPI, tools
  - Realistic workloads: math and I/O libraries

HPC11 Compute Acceptance Test: Performance Test

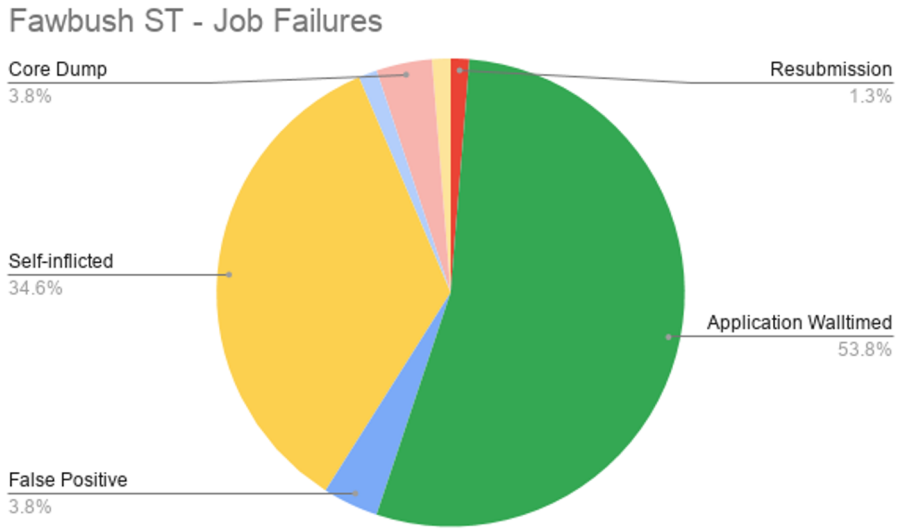
- Focuses on workloads specific to the individual program
- Replicate results submitted from VT:
  - UM 10.9
  - 4DVAR
- Execute OLCF applications in isolation to obtain reference values on a quiet system
  - LSMS: https://github.com/mstsuite/lsms
    - Locally Self-consistent Multiple Scattering
  - GenASiS: https://github.com/GenASiS/GenASiS_Basics
    - General Astrophysics Simulation System
  - minisweep: https://github.com/olcf/minisweep
    - Sn radiation transport miniapp for Denovo

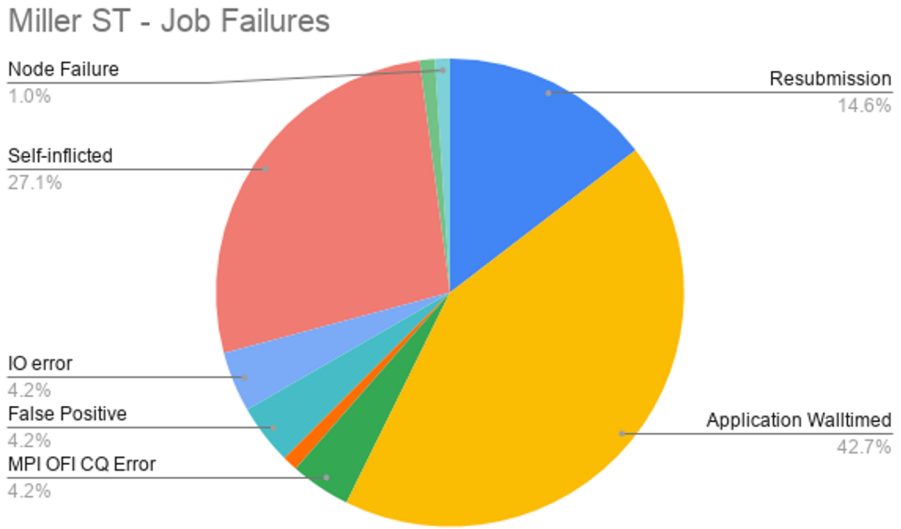
HPC11 Compute Acceptance Test: Stability Test

- Simulates a realistic workload on the system that combines:
  - Realistic continuous batch workload
  - Code development activities: compiling, job submission, data movement
- Stability test was managed by the OLCF Test Harness
  - https://github.com/olcf/olcf-test-harness
- Over 5,700 individual jobs were independently executed on each Fawbush and Miller
- Each system successfully completed a 14-day stability period:
  - 99.19% pass rate on Fawbush
  - 98.86% pass rate on Miller

HPC11 Compute Acceptance Test: Stability Test (cont'd)

- Stability test successfully demonstrated the reliability of the systems for the target workloads:
  - < 2.5% runtime variability for UM
  - < 4% runtime variability for 4DVAR
- All failures were classified and reported to HPE





HPC11 Compute Acceptance Test

- Compute acceptance identified several issues that were addressed before the system was accepted
- The OLCF Test Harness was able to capture several issues that could have impacted production workloads. A subset of those include:
  - UM test was able to detect a single CPU that was an early AMD test escape
    - A screen was conducted on all nodes and the defective part replaced
  - 4DVAR cases using a larger decomposition showed a higher rate of instability resulting in application walltimes
  - CCE 10 compiler bug reported for GenASiS and fixed in CCE 11
  - gdb4hpc unable to start in a multi-cluster SLURM environment
    - Bug is being investigated by HPE

HPC11 Storage Architecture

- Two identical, independent file systems, each with:
  - 1x DDN SFA14KX (10 enclosures), presented to 6x OSSs via SRP over direct-connect Infiniband
  - 1x DDN SFA200NV, presented to 2x MDSs via SRP over direct-connect Infiniband
  - 7.5PB usable capacity
  - ~110TB usable flash DoM capacity
  - Each server provides a 2x100G bonded (active-active) diverse ethernet HSN
  - Benchmarking (fio) showed performance of ~60GB/s write and ~65GB/s read at the block layer

HPC11 Storage Acceptance Test (cont'd)

- Several significant performance issues initially encountered:
  - Poor ksocklnd performance on our bonded ethernet interface (LU-14293)
    - iperf between two nodes was achieving 98Gbps and we were able to demonstrate 190Gbps with a 2->1 setup
    - lnet selftest could only hit ~20Gbps in a node to node test
    - ORNL backported a multiple-socket patch (LU-12815) to 2.12 which resolved this issue
  - Client hangs when using DoM with a fixed mdc lru_size (LU-14221)
    - ORNL typically set fixed lru_size to avoid the potential for a large memory footprint from dynamic
    - Clients would reliably hang with a reproducer that included large amounts of metadata operations
    - Resolved by backporting LU-11518
  - Multitude of grant related issues requiring a custom client, deviating from Cray client
  - Ongoing ZFS performance issues

HPC11 Storage Acceptance Test (cont'd)

- Performance then vs. now:
  - Initial hero benchmark numbers were ~35GB/s write and ~30GB/s read, single client performance ~2GB/s read/write
  - ~6GB/s read/write on clients
  - ~45GB/s write/read hero numbers

Conclusions

- Switching to HPCM for system software management, new to ORNL, was a learning experience but resulted in a successful deployment
- Executing identical systems in parallel introduces additional dependencies that must be carefully managed
  - Simultaneously accepting a file system pair complicated the process further
- Because the systems are independently managed, acceptance of Miller was able to proceed despite having to sustain a scheduled outage on Fawbush
- Using a workload that matches the intended use cases allows us to identify issues that directly impact the target user community
  - Augmenting the realistic workload with use cases from OLCF to increase test coverage of the programming environment

Acknowledgements

- HPE team:
  - Cathy Willis, Adam Sachitano, Jeff Becklehimer, Pete Johnsen, Eric Dolven, Kim Kafka
- US AFW team:
  - Frank Ruggiero, Bob Born, Dave Hladky
- ORNL team:
  - Dustin Leverman, Jim Rogers, Matt Ezell, Daniel Pelfrey, Joseph Voss, Brian Smith
- 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.

Thank you! Questions?