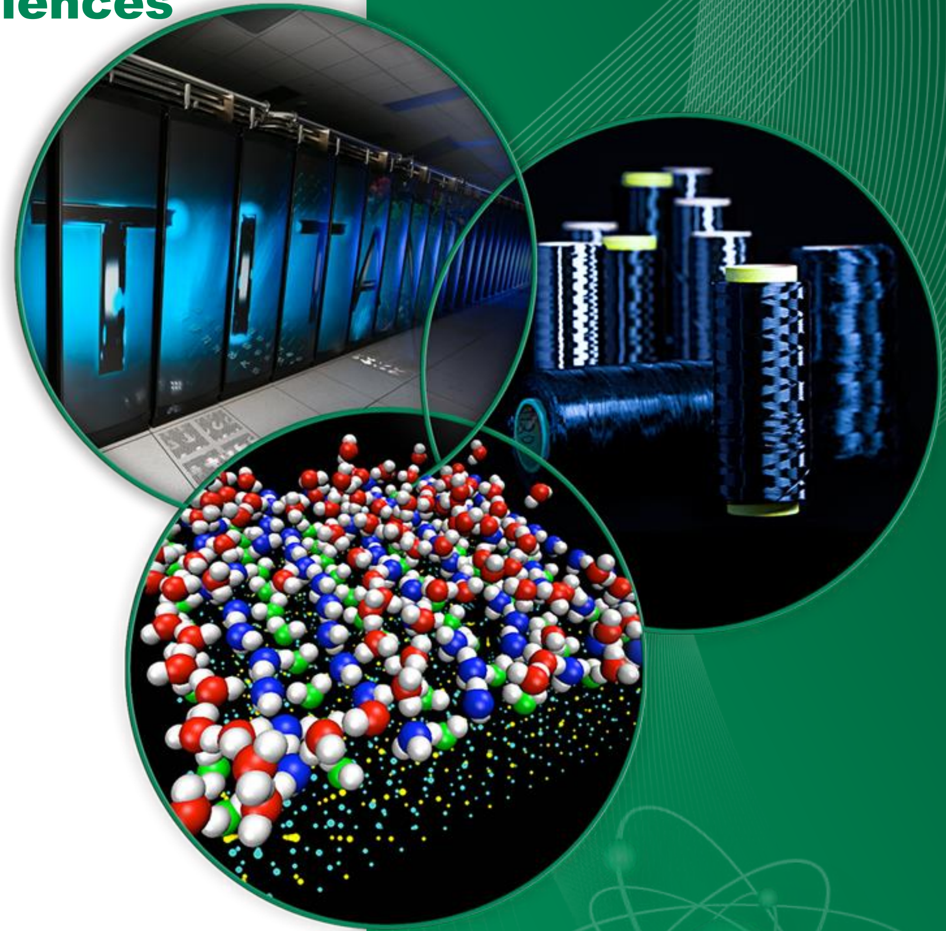
Oak Ridge National Laboratory
Computing and Computational Sciences



OFA Update by ORNL

Presented by:

Pavel Shamis (Pasha)

OFA Workshop
Mar 17, 2015

Acknowledgments

- Bernholdt David E.
- Hill Jason J.
- Leverman Dustin B.
- Curtis Philip B.

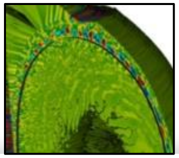
OLCF

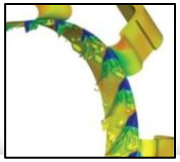
- The Oak Ridge Leadership Computing Facility (OLCF) was established at Oak Ridge National Laboratory in 2004 with the mission of accelerating scientific discovery and engineering progress by providing outstanding computing and data management resources to high-priority research and development projects.

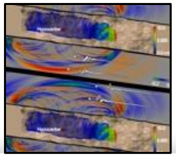
Jaguar: 2.3 PF

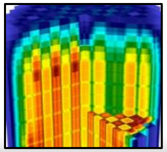
Titan: 27 PF

- ORNL's supercomputing program has grown from humble beginnings to deliver some of the most powerful systems in the world. On the way, it has helped researchers deliver practical breakthroughs and new scientific knowledge in climate, materials, nuclear science, and a wide range of other discipli









CORAL

- CORAL – Collaboration of ORNL, ANL, LLNL
- Objective – Procure 3 leadership computers to be sited at Argonne, Oak Ridge and Lawrence Livermore in 2017
  - Two of the contracts have been awarded with the Argonne contract in process
- Leadership Computers
  - RFP requests >100 PF, 2 GB/core main memory, local NVRAM, and science performance 5x-10x Titan or Sequoia

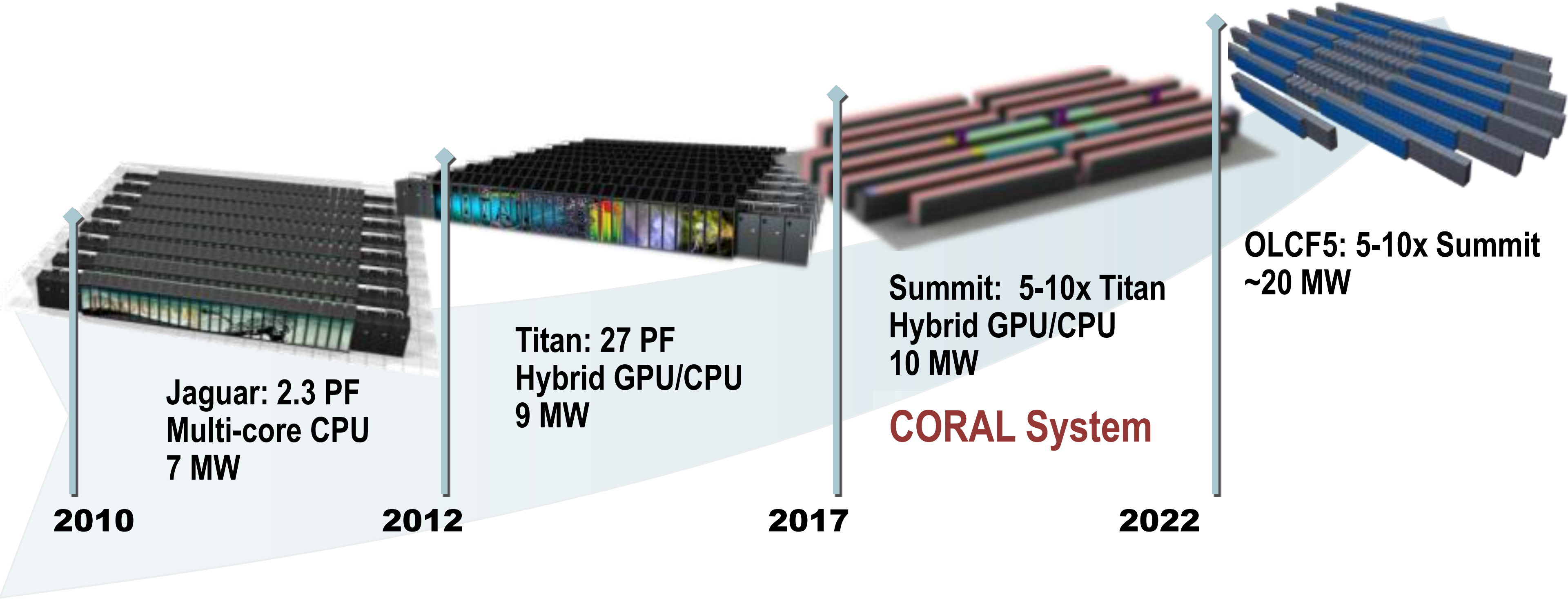
Oak Ridge National Laboratory

Lawrence Livermore National Laboratory

The Road to Exascale

Since clock-rate scaling ended in 2003, HPC performance has been achieved through increased parallelism. Jaguar scaled to 300,000 cores.

Titan and beyond deliver hierarchical parallelism with very powerful nodes. MPI plus thread level parallelism through OpenACC or OpenMP plus vectors



Jaguar: 2.3 PF
Multi-core CPU
7 MW

Titan: 27 PF
Hybrid GPU/CPU
9 MW

Summit: 5-10x Titan
Hybrid GPU/CPU
10 MW

CORAL System

OLCF5: 5-10x Summit
~20 MW

2010

2012

2017

2022

System Summary

Compute Node

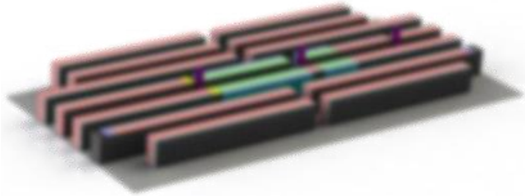
POWER® Architecture Processor
NVIDIA®Volta™
NVMe-compatible PCIe 800GB SSD
> 512 GB HBM + DDR4
Coherent Shared Memory

Compute Rack

Standard 19"
Warm water cooling

Compute System

Summit: 5x-10x Titan
10 MW

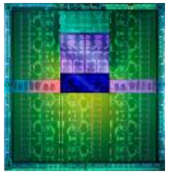


IBM POWER

- NVLink™

NVIDIA Volta

- HBM
- NVLink



Mellanox® Interconnect

Dual-rail EDR Infiniband®

OFA Software Stack

Summit VS Titan

Feature	Summit	Titan
Application Performance	5-10x Titan	Baseline
Number of Nodes	~3,400	18,688
Node performance	> 40 TF	1.4 TF
Memory per Node	>512 GB (HBM + DDR4)	38GB (GDDR5+DDR3)
NVRAM per Node	800 GB	0
Node Interconnect	NVLink (5-12x PCIe 3)	PCIe 2
System Interconnect (node injection bandwidth)	Dual Rail EDR-IB (23 GB/s)	Gemini (6.4 GB/s)
Interconnect Topology	Non-blocking Fat Tree	3D Torus
Processors	IBM POWER9 NVIDIA Volta™	AMD Opteron™ NVIDIA Kepler™
File System	120 PB, 1 TB/s, GFS™	32 PB, 1 TB/s, Lustre®
Peak power consumption	10 MW	9 MW

Present and Future Leadership Computers at OLCF, Buddy Bland

https://www.olcf.ornl.gov/wp-content/uploads/2014/12/OLCF-User-Group-Summit-12-3-2014.pdf

ORNL

- ORNL Joined OFA organization in December 2014
- InfiniBand Technology
  - Multiple InfiniBand installations
  - Substantial experience in management of InfiniBand networks
- OFA software stack users
  - ADIOS, Burst Buffers, CCI, Cheetah, Luster / Spider, Open MPI, Open SHMEM, STCI, UCCS/UCX, and more…
  - Substantial experience in research and development of HPC and RDMA software
- We participate in OFVWG and OFIWG efforts

Our Experience with OFA Software Stack

- There are multiple ways to get OFA software…
  - Mellanox OFED
  - OFA OFED
  - OFED packages within Linux distributions
- …and it is not easy choice

OFED

- Mellanox OFED
  - Up-to-date software stack
  - Network tools provide a *relatively* good level of information (device details, speeds, etc.)
  - Consistent CLI interface for *most* of the tools
  - All the above is true for Mellanox software/hardware only
- OFED
  - Community software stack (supports multiple technologies, vendors)
  - Very often it is behind Mellanox OFED in terms software features
  - Inconsistent CLI interfaces
  - Tools are way behind Mellanox OFED tools

OFED - Continued

- OFED packages within Linux distributions
  - Very easy to maintain, upgrade, install
    - “yum upgrade”
  - Security updates !
  - Based on OFED + Mellanox OFED (patches) ?
  - Packages are behind Mellanox OFED in terms software features

Wish list: Tools and Documentation

- We need better tools for management and monitoring IB networks
  - Currently we use ibdiagnet + scripting
  - It is relatively easy to fetch HCA’s information/statistic
  - It is difficult to extract information about switches
  - It is even more difficult to “connect the dots”
    - How do we identify “hot” spots and congestion in the network
    - Identification of bad cables/connectors
  - We want it free and open source ☺
- Documentation of best practices and user experience can be very helpful
  - Advance level: QoS tuning, balance loading with LMC bits, routing algorithms, etc..

Questions?

U.S. DEPARTMENT OF ENERGY | Office of Science

This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725