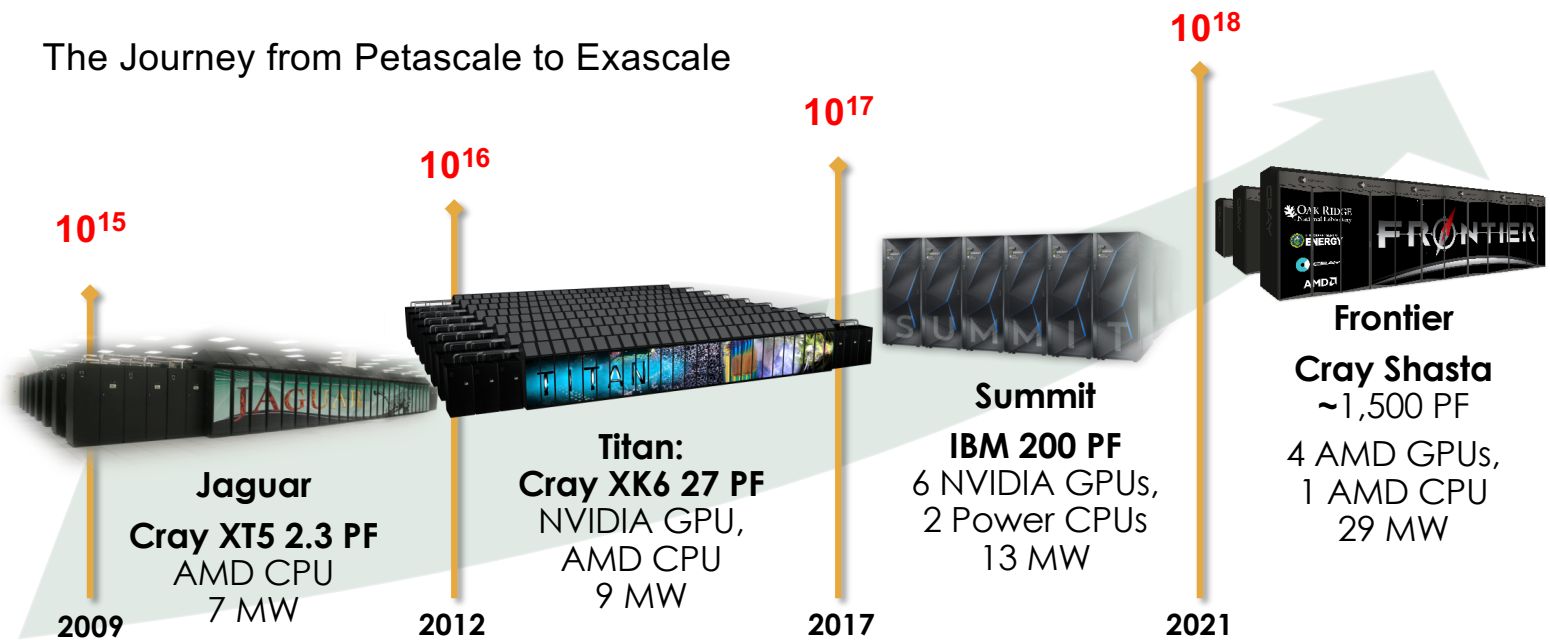
ORNL's Frontier Exascale Computer

Al Geist
Oak Ridge National Laboratory

Smoky Mountain Conference
August 27-30, 2019

Oak Ridge Leadership Computing Facility
Roadmap to Exascale

Mission: Providing world-class computational resources and specialized services for the most computationally intensive global challenges for researchers around the world.



Four Key Challenges to Reach Exascale

What is so special about Exascale vs. Petascale?
In 2009 there was serious concern that Exascale Systems may not be possible

Parallelism: Exascale computers will have billion-way parallelism (also termed concurrency). Are there more than a handful of applications that could utilize this?

Data Movement: Memory wall continues to grow higher - Moving data from the memory into the processors and out to storage is the main bottleneck to performance.

Reliability: Failures will happen faster than you can checkpoint a job. Exascale computers will need to dynamically adapt to a constant stream of transient and permanent failures of components.

Energy Consumption: Research papers in 2009 predicted that a 1 Exaflop system would consume between 150-500 MW. Vendors were given the ambitious goal of trying to get this down to 20 MW.

Exascale research efforts were started to address these challenges

After Several False Starts

Exascale False Starts: Who Remembers

- **Nexus / Plexus**
- **SPEC / ABLE**
- **Association Model**

We finally got traction with:

- **CORAL**
- **Exascale Computing Project**

Supercomputer Specialization vs ORNL Summit

- As supercomputers got larger and larger, we expected them to be more specialized and limited to just a small number of applications that can exploit their growing scale

- Summit's architecture with powerful, multiple-GPU nodes with huge memory per node seems to have stumbled into a design that has broad capability across:
  - Traditional HPC modeling and simulation
  - High performance data analytics
  - Artificial Intelligence

ORNL Pre-Exascale System -- Summit

System Performance

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

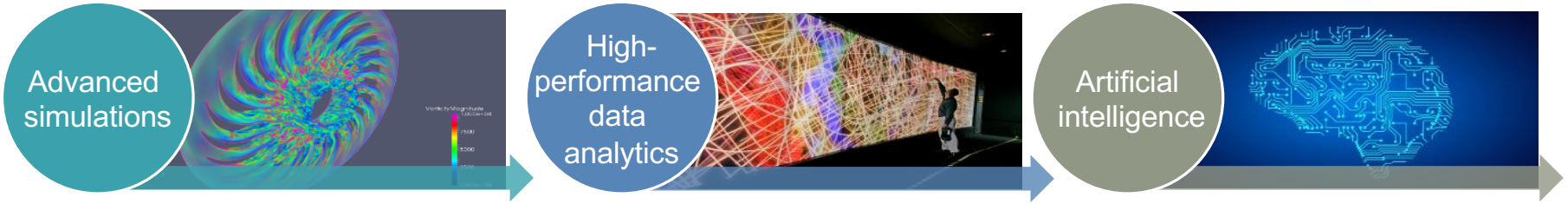
The system includes

- 4608 nodes
- Dual-rail Mellanox EDR InfiniBand network
- 250 PB IBM 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 NVM memory

Multi-GPU nodes Excel Across Simulation, Analytics, AI



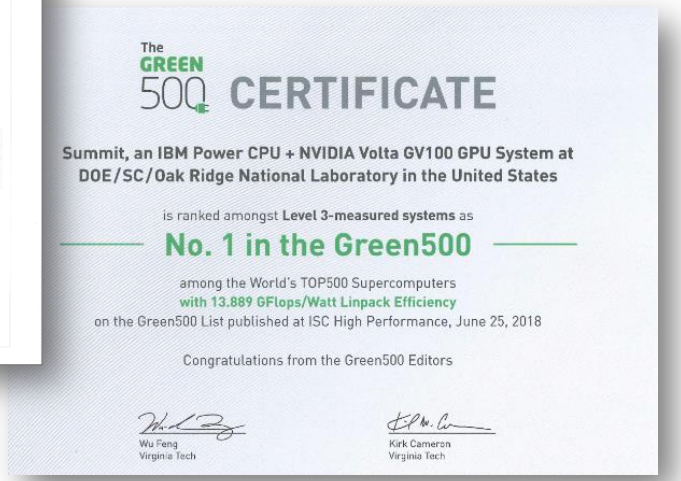
- Data analytics – CoMet bioinformatics application for comparative genomics. Used to find sets of genes that are related to a trait or disease in a population. Exploits cuBLAS and Volta tensor cores to solve this problem 5 orders of magnitude faster than previous state-of-art code.
  - **Has achieved 2.36 ExaOps** mixed precision (FP_{16}-FP_{32}) on Summit
- Deep Learning – global climate simulations use a half-precision version of the DeepLabv3+ neural network to learn to detect extreme weather patterns in the output
  - **Has achieved a sustained throughput of 1.0 ExaOps (FP_{16})** on Summit
- Nonlinear dynamic low-order unstructured finite-element solver accelerated using mixed precision (FP_{16} thru FP_{64}) and AI generated preconditioner. Answer in FP_{64}
  - **Has achieved 25.3 fold speedup** on Japan earthquake – city structures simulation
- **Half-dozen Early Science codes are reporting >25x speedup on Summit vs. Titan**

Multi-GPU Nodes Excel in Performance, Data, and Energy Efficiency

Summit achieved #1 on TOP500, #1 on HPCG, and #1 Green500

122 PF HPL
Shows DP performance

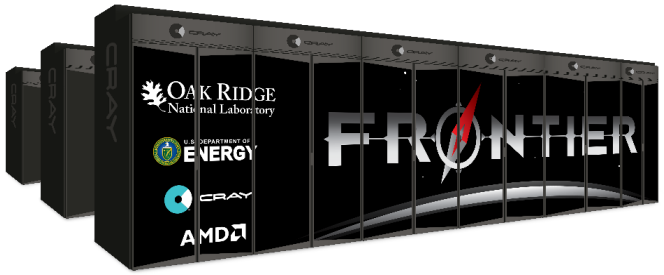
2.9 PF HPCG
Shows fast data movement



13.889 GF/W
Shows energy efficiency

Frontier Continues the Accelerated Node Design

begun with Titan and continued with Summit



Partnership between ORNL, Cray, and AMD

The Frontier system will be delivered in 2021

Peak Performance greater than 1.5 EF

Composed of more than 100 Cray Shasta cabinets

- Connected by Slingshot™ interconnect with adaptive routing, congestion control, and quality of service

Accelerated Node Architecture:

- One purpose-built AMD EPYC™ processor
- Four HPC and AI optimized Radeon Instinct™ GPU accelerators
- Fully connected with high speed AMD Infinity Fabric links
- Coherent memory across the node
- 100 GB/s node injection bandwidth
- On-node NVM storage

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 Kepler GPU	2 IBM POWER9™ CPUs 6 NVIDIA Volta GPUs	1 AMD EPYC CPU 4 AMD Radeon Instinct GPUs
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™	4x performance and 3x capacity of Summit's I/O subsystem.
On-node NVM	No	Yes	Yes
Power	9 MV	13 MV	29 MV

Moving Applications from Titan and Summit to Frontier

ORNL, Cray, and AMD are partnering to co-design and develop enhanced GPU programming tools.

- These new capabilities in the Cray Programming Environment and AMD's ROCm open compute platform will be integrated into the Cray Shasta software stack.

HIP (Heterogeneous-compute Interface for Portability) is an API developed by AMD that allows developers to write portable code to run on AMD or NVIDIA GPUs.

- The API is very similar to CUDA so transitioning existing codes from CUDA to HIP is fairly straightforward
- OLCF has HIP available on Summit so that users can begin using it prior to its availability on Frontier

In addition, Frontier will support many of the same compilers, programming models, and tools that have been available to OLCF users on both the Titan and Summit supercomputers

Solutions to the Four Exascale Challenges

How Frontier addresses the challenges

Parallelism: The GPUs hide between 1,000 and 10,000 way concurrency inside their pipelines so the users don't have to think about as much parallelism. Summit has shown the multi-GPU node design can do well in simulation, data, and learning.

Data Movement: Having High Bandwidth memory soldered onto the GPU increases BW an order of magnitude and GPUs are well suited for latency hiding.

Reliability: Having on-node NVM (Non-Volatile Memory) reduces checkpoint times from minutes to seconds. Cray adaptive network and system software aid in keeping system up despite component failures.

Energy Consumption: Frontier is projected to use less than 20 MW per 1 Exaflop – due in part to the 10 years of DOE investment in vendors for Exascale technologies. (FastForward, Design Forward, Pathforward)

Questions?

ORNL / Cray / AMD Partnership