OAK RIDGE National Laboratory

The Challenges of Exascale-Capable Facilities

Jim Rogers

Director, Computing and Facilities
National Center for Computational Sciences
Oak Ridge National Laboratory

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

U.S. DEPARTMENT OF ENERGY

The Era of ORNL's 'Modern' Computing Facilities

2004

- 20,000 ft2 (1,858 m2)
  - 2000 ft2 for largest system
- (2) 1.5MVA Transformers
- (3) 1,200-ton chillers
- 42°F (5.5 ° C) chilled water supply setpoint

Cray X1E - 18.5TF

- Traditional CRACs on perimeter with 55 °F (13 ° C)
- 250 lbs/ft2 raised floor on 3' pedestals.
- Mechanical, electrical, forced air plenum all below the raised floor
- PUE ~1.5 for the X1E

The Era of ORNL's 'Modern' Computing Facilities

2004

- 20,000 ft2
- 3MW Electrical
- 3600-tons chillers
- 42°F supply
- 1.5 PUE for X1E

Cray X1E - 18.5TF

2012

- 35,000 ft2 (3,251 m2)
  - 5,500 ft2 for largest system
  - 2,000 ft2 for file system
- 17MW Electrical Capacity
  - 9.6MW for largest system
- 6600-tons chillers
- 42°F (5.5 ° C) chilled water supply setpoint
- 1.29 PUE for Titan; 1.8 for air-cooled systems.

Cray XK7 - 27PF

The Era of ORNL's 'Modern' Computing Facilities

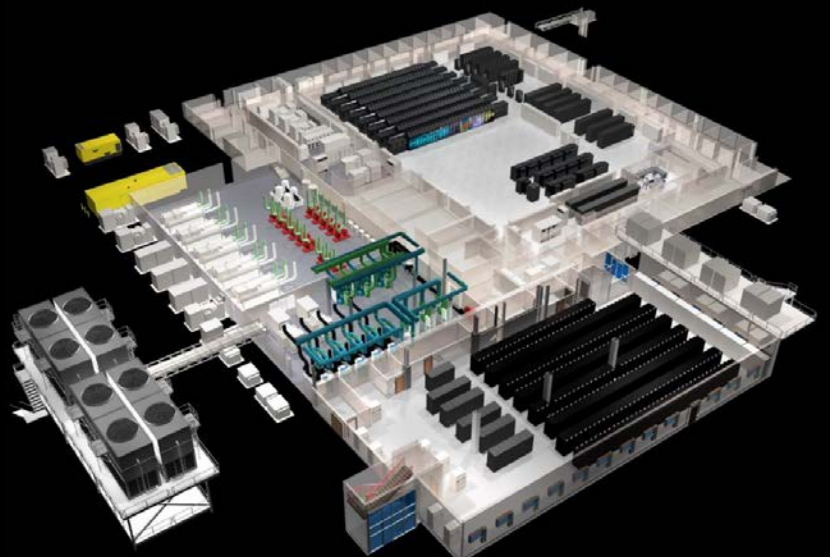
2004
- 20,000 ft2
- 3MW Electrical
- 3600-tons chillers
- 42°F supply
- 1.5 PUE for X1E

Cray X1E. 18.5TF

2012
- 35,000 ft2
- 17MW Electrical
- 6600-tons chillers
- 42°F (5.5 °C) supply
- 1.29 PUE for Titan

Cray XK7. 17.59PF

2018
- 35,000 ft2 + 11,000ft2
- 27MW Electrical Capacity
- 6600-tons chillers/7700-tons evaporative cooling
- Summit uses warm water (21°C)
- 1.05 PUE for Summit



OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

ORNL's Transition to Warmer Facility Supply Temperatures

Titan: Refrigerant-based per-rack cooling with direct rejection of heat to cold 5.5°C water

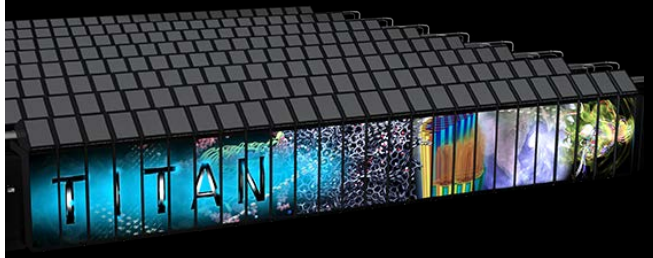
- ***Below dewpoint***
- ***100% use of chillers***

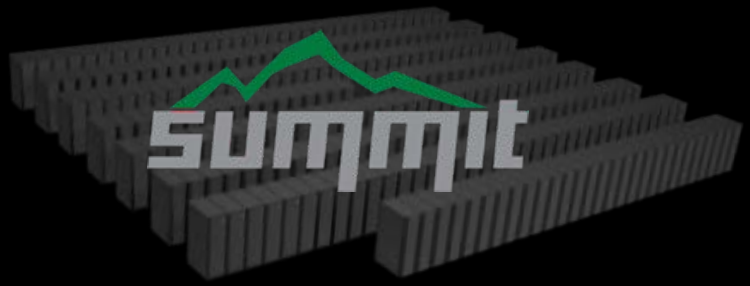
Summit: A combination of direct on-package cooling and RDHX with 21°C supply is > 95% room-neutral.

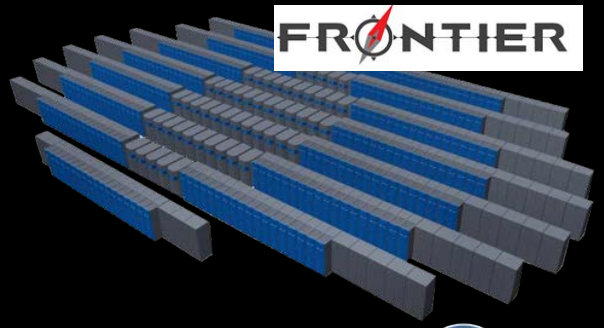
- ***Above dewpoint***
- ***Contribution by chillers ~20% of the hours of the year***

Frontier: Custom mechanical packaging is >95% room-neutral with a 30°C supply.

- ***~100% Evaporative Cooling, with supplemental HVAC for parasitic loads***







27 PF
2012
Cray XK7
Titan

200 PF
2018/2019
IBM
Summit

~1.5 EF
2021/2022
Frontier

OAK RIDGE National Laboratory

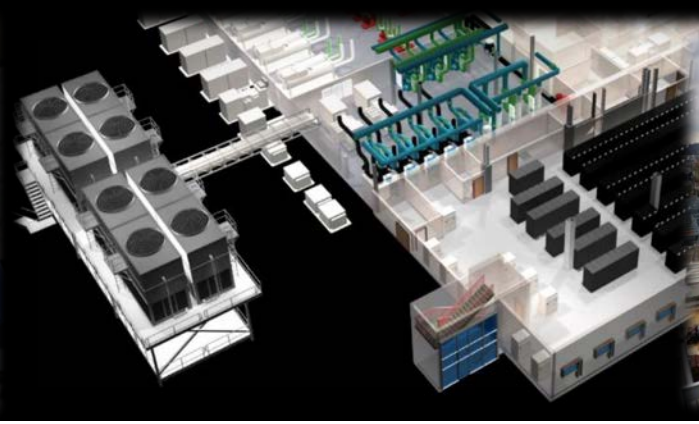
Annual PUE ~1.05

>100cabinets;
29MW design point
Annual PUE << 1.05

Facility Challenges for Exascale Systems

- Upgrades are very difficult and decisions can have huge impacts.
  - 2011 - ORNL used a rolling upgrade to move from Jaguar (Cray XE6) to Titan (XK7). Impacts to production capability, reliability for ~12 months
  - 2017 – 2018 - ORNL focused on a second/separate facility for Summit. 11,000 ft2 and 21°C Avoided interruptions to Titan/production, but becomes a constrained solution
  - 2019 – 2021 - Frontier reuses the physical space for Titan, but requires a third mechanical plant that can produce 16,000 tons of warm-water (to W3) cooling.

What is the appropriate strategy for your facility?

Difficult Decisions for Facility Providers

- What path is best for your facility?
  - Riken, with their transition from K to Fugaku, leverages existing infrastructure and footprint, but with an impact to production to allow removal/upfit/installation
  - Summit required a new facility and mechanical plant, eliminating impact to existing production, but with a constrained result
- Frontier leverages the old Titan space, but with considerable impact to other production users, and a massive facility undertaking.
  - O($100M) *facility* investment
  - 2+ year scope
  - Impact to 13.8/161kV infrastructure

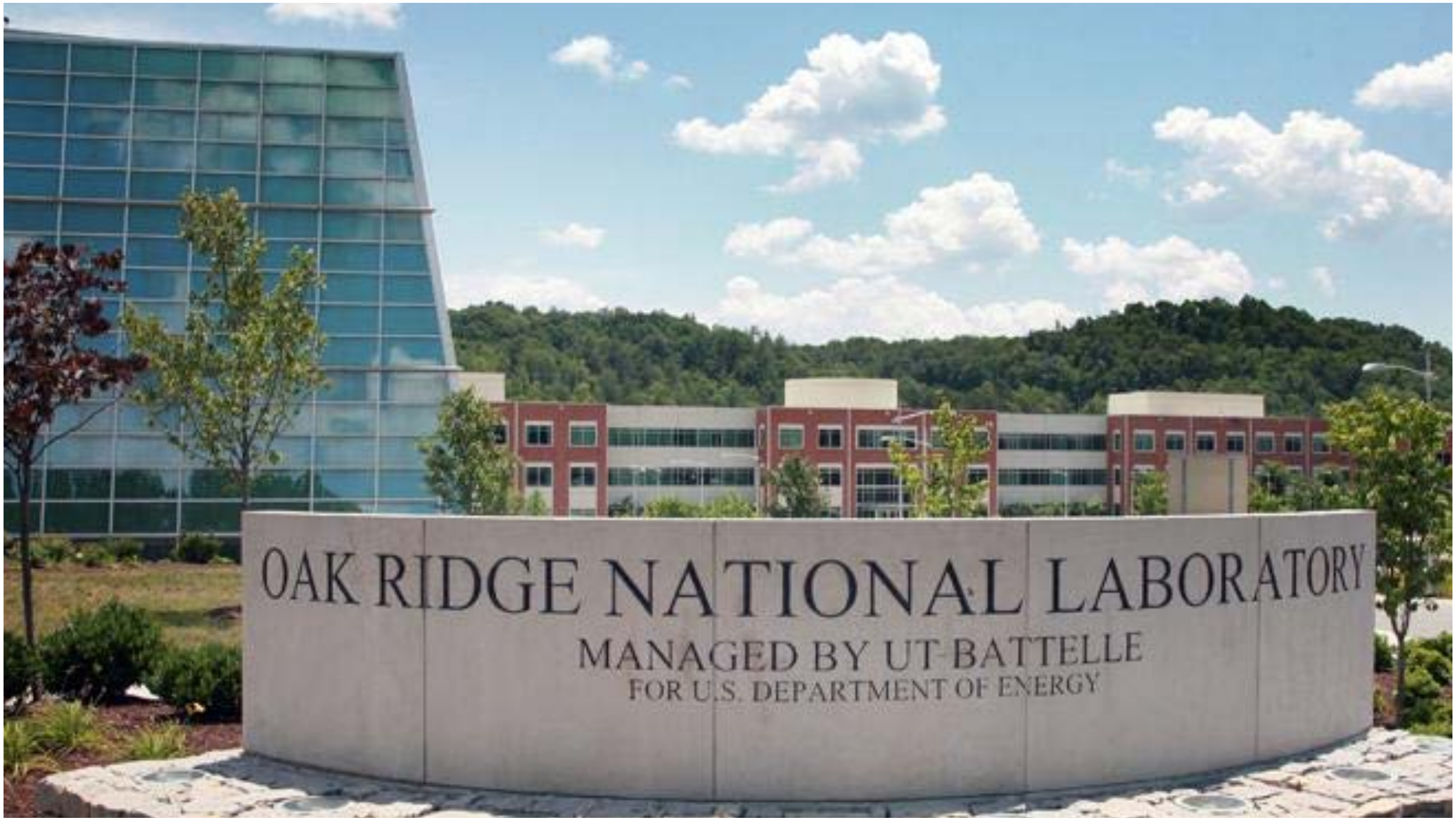
Initial Electrical Capacity of 40MW

Mechanical Capacity of up to 60MW

How do you avoid the tail wagging the dog?

Consider ORNL's Strategy Beyond Frontier/OLCF5

- Frontier enters production in 2021/2022, using
  - More than 50% of the space in the data center (> 100 cabinets + CDUs and other support infrastructure)
  - 29MW of power (design point) – 40MW provisioned to the premises
  - Commensurate mechanical/warm water cooling
  - Air-cooled file systems that still need 42F water (single-digit MW)
- How do you manage OLCF6 (~2025)
  - Insufficient space, unless it is an upgrade
  - + 20MW of electrical infrastructure is available via new 13.8kV and transformer upfit
  - Mechanical infrastructure is sufficient to 60MW
- And what about OLCF7? (~2029)
- And what about all the Strategic Partnerships?

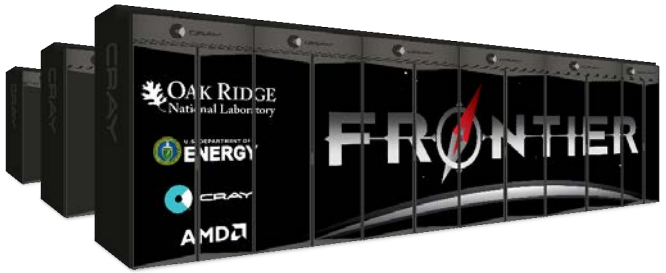

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE
FOR U.S. DEPARTMENT OF ENERGY

Supplemental

Frontier – Publicly Released Information, Nov 2019

Supplemental

Frontier Continues the Accelerated Node Design



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 injection bandwidth
- Near-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™	2-4x performance and capacity of Summit's I/O subsystem.
On-node NVM	No	Yes	Yes
Power	9 MW	13 MW	29 MW