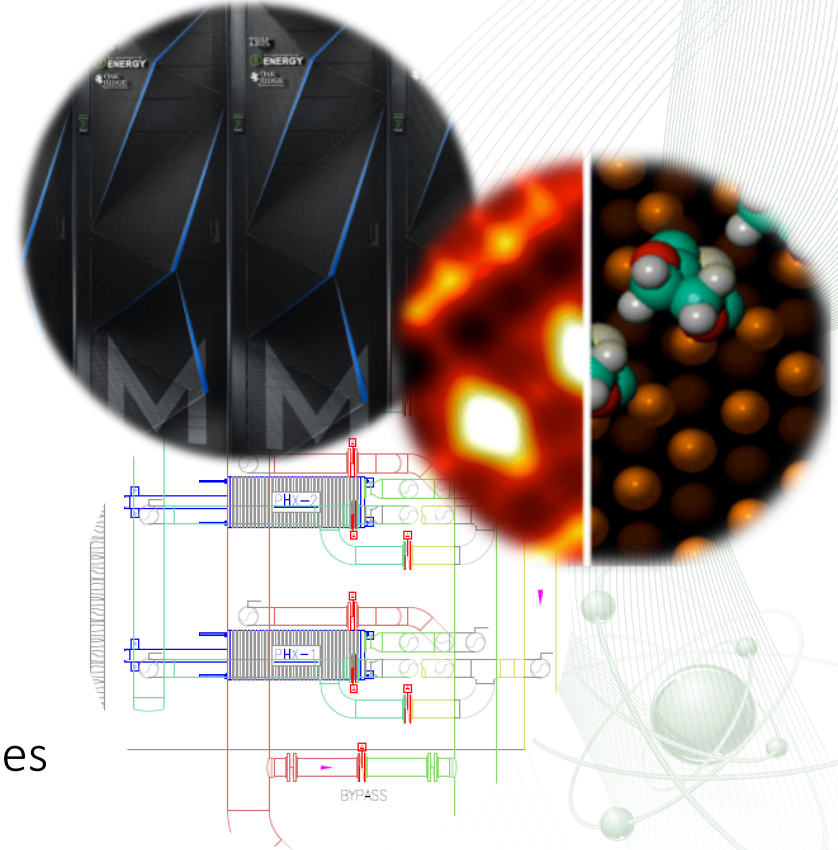
Designing an HPC Facility for Modern Systems



Jim Rogers

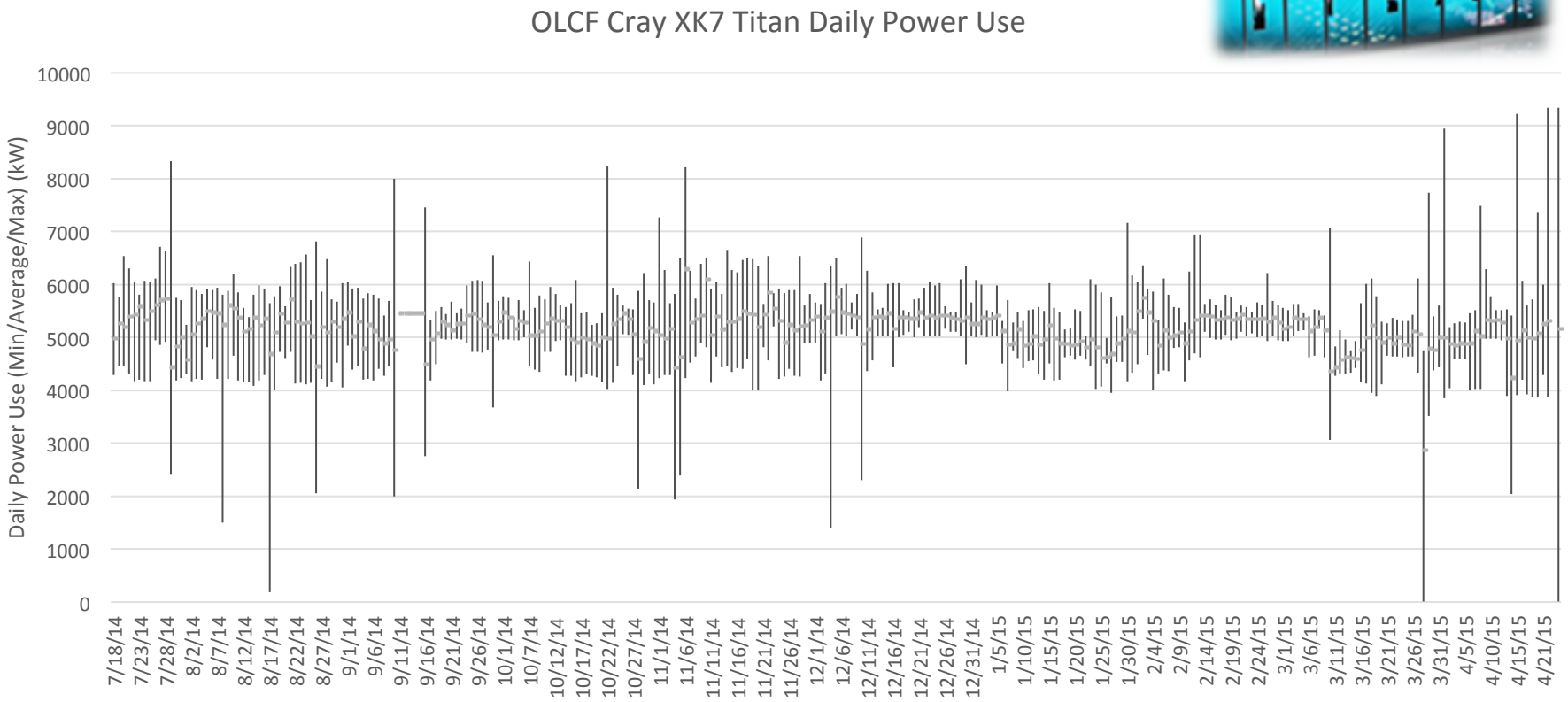
Director, Computing and Facilities
National Center for Computational Sciences
Oak Ridge National Laboratory

OAK RIDGE National Laboratory | OAK RIDGE LEADERSHIP COMPUTING FACILITY

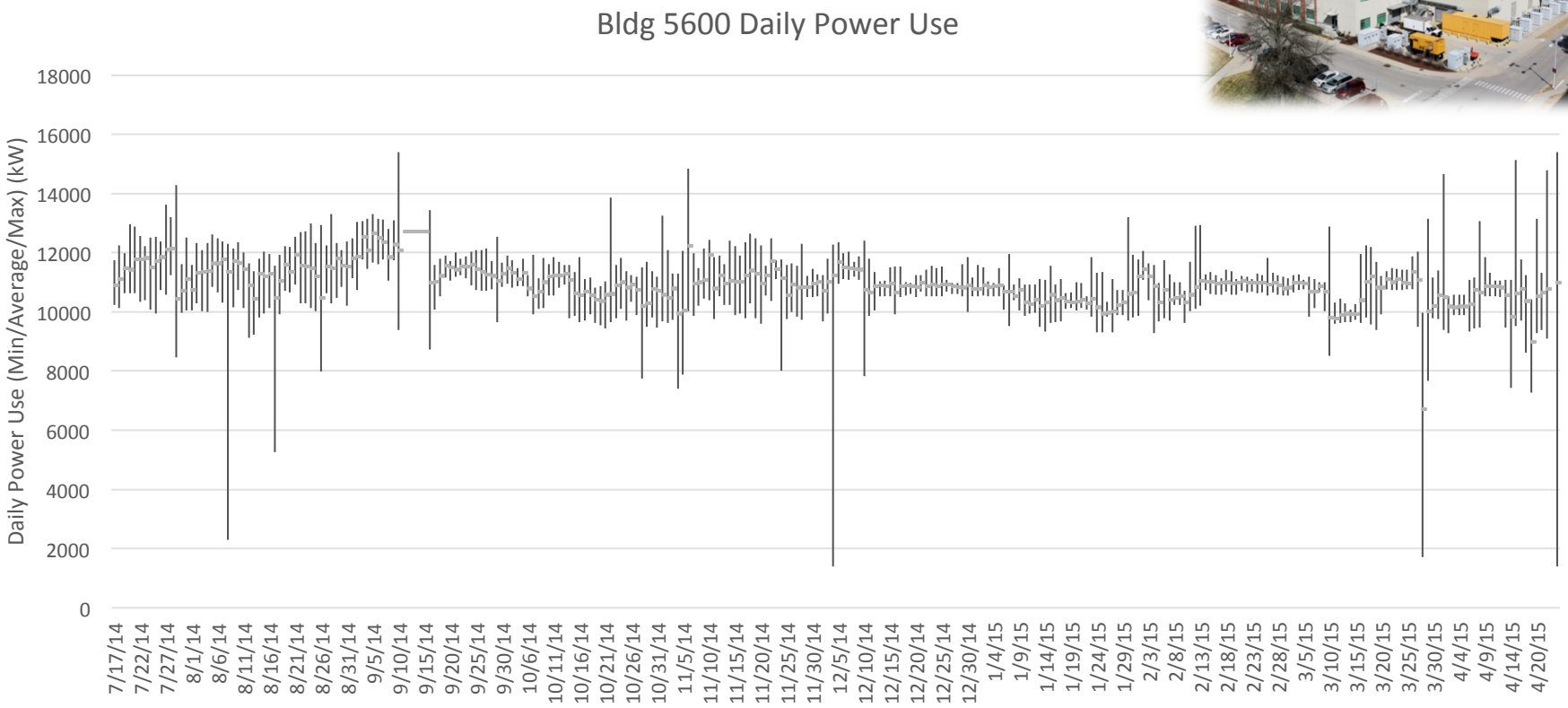
Designing an HPC Facility for Modern Systems

Abstract: In November 2014, Oak Ridge National Laboratory (ORNL) announced a subcontract with IBM for Summit, a 150PF 10MW supercomputer that presents significant challenges and opportunities. ORNL currently operates Titan, a 27PF Cray XK7 that has remained the fastest system in the United States for nearly three years. Summit will eventually supersede Titan, but in an operating environment that will dramatically shift. Gone are the traditional chillers, cold water temperatures and low flow rates. In their place, ORNL will provision a facility that can provide up to 20MW of electrical distribution capacity in a stepwise manner, 20℃ supply temperatures, and significant reductions in operating costs through evaporative cooling. This talk will describe the anticipated mechanical and electrical designs, the control systems, and the measurement systems not only for Summit, but also for the facility as a whole.

Current Single Largest System



Current Facility Power Use



(Current) Rule of Thumb for HPC Facility Planning

Ignore. >100%	• Manufacturer's Power Supply Aggregate Rating
Ignore. >100%	• TDP
Peak – 100%	• Diagnostics and Kernels designed to hammer CPUs, GPUs, Memory
75-85% of peak. Typically rolls down as I/O interferes	• HPL and other applications that can pipeline significant numbers of floating point operations
70%, with bursts to 85%	• Tuned applications/benchmarks targeted to your architecture
60%	• The majority of the applications at your facility/on this system
About 50%, indicating power management issues	• Applications that do not leverage architectural features, such as accelerators
30-35%. Can vary substantially	• Idle (OS-only)
~20% (varies widely)	• System Maintenance (Power-ON)
0%	• System Maintenance (Power-OFF)

Turbo Boost, Power Gating, Clock Frequency Manipulation, and other methods will definitely shift the nominal (60-70%) consumption upwards in near-term future systems.

2017 OLCF-4 Leadership System

Hybrid CPU/GPU architecture

Vendors: IBM, NVIDIA, and Mellanox

Baseline is 150PF (peak) and 5X Titan's Application Performance

Approximately 3,500 nodes, each with:

- Multiple IBM POWER9 CPUs and multiple NVIDIA Tesla® GPUs using the NVIDIA Volta™ architecture
- CPUs and GPUs completely connected with high speed NVLink™
- Large coherent memory: over 512 GB (HBM + DDR4)
- An additional 800 GB of NVRAM, which can be configured as either a burst buffer or as extended memory

Dual-rail Mellanox® EDR-IB full, non-blocking fat-tree interconnect

IBM Elastic Storage (GPFS™) - 1TB/s I/O and 120 PB disk capacity.

9.6 MW design point for the baseline system.

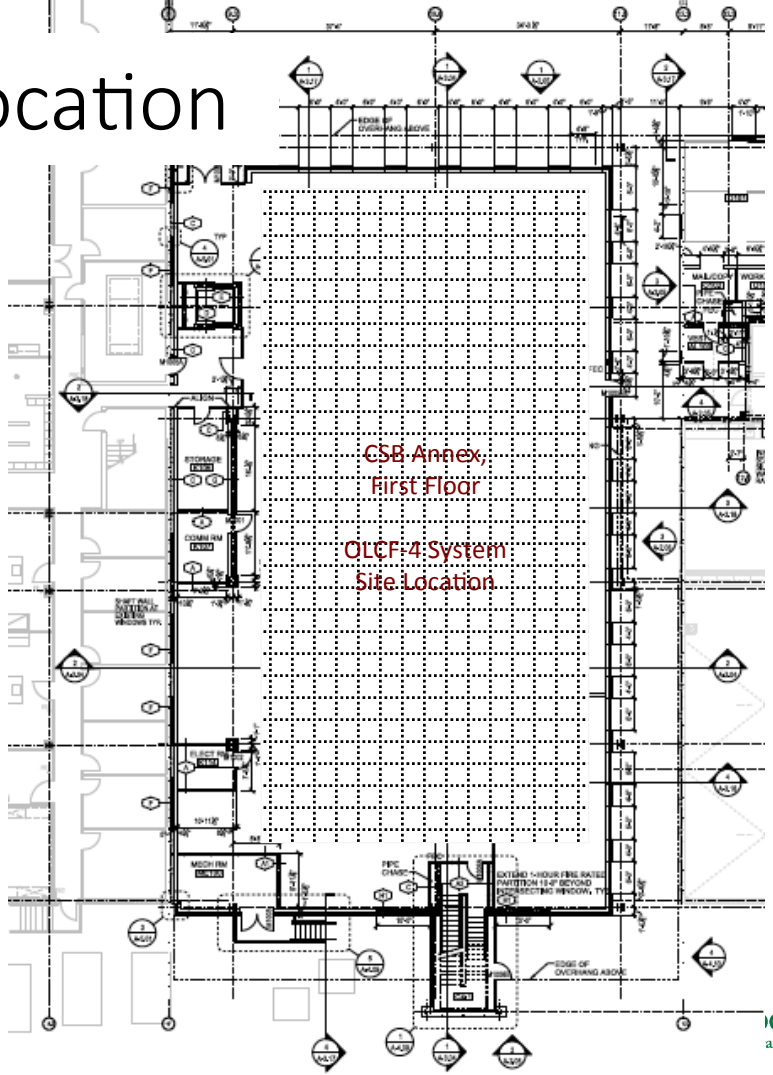
A single Statement of Work defines both the Summit and LLNL Sierra systems.
A Non-Recurring Engineering (NRE) subcontract provides support for required features that must be developed prior to system delivery.

DOE ASCR Computing Portfolio

System attributes	NERSC Now	OLCF Now	ALCF Now	NERSC Upgrade	OLCF Upgrade	ALCF Upgrade
Name Planned Installation	Edison	TITAN	MIRA	Cori 2016	Summit 2017-2018	Aurora 2018-2019
System peak (PF)	2.6	27	10	> 30	150	180
Peak Power (MW)	2	9	4.8	< 3.7	9.6	13
Node perf (TF)	0.460	1.452	0.204	> 3	> 40	> 15 times Mira
Node processors	Intel Ivy Bridge	AMD Opteron + NVIDIA Kepler	64-bit PowerPC A2	Intel Knights Landing (KNL)	IBM Power9 CPUs + NVIDIA Volta GPUs	Intel Knights Hill (KNH)
System size (nodes)	5,600 nodes	18,688 nodes	49,152	9,300 nodes; 1,900 nodes in data partition	~3,500 nodes	>50,000 nodes

OLCF-4 System Site Location

- ~11,000 ft^2 in CSBx.
  - Currently an unfinished space
- Slab on grade (no raised floor).
  - No issues with floor loading, delivery, freight elevators.
  - But, no convenience for interconnect and network cables.
- Delivery of water and electrical from above
- VESDA, life safety, lighting, and other systems compete with electrical and mechanical for space above.
- Customer (us) to provide up to 20MW electrical distribution capacity and commensurate mechanical cooling capability.
- Build-out schedule inside the computer room is separate from the build-out/delivery of new bulk mechanical and electrical distribution services to the facility



OAK RIDGE
LEADERSHIP
COMPUTING FACILITY

Facility Overview for 20 MW Power/Cooling

- 2015-2016- Bring power and cooling to the wall of the building (demarc)
  - Up to (7) new 3MVA transformers (~20MW)
  - Scales to meet system options, and future growth without disruption
  - Strict requirements for transformer selection
  - Minimum efficiency of 99.54% (extrapolated from DOE 2016 requirements)

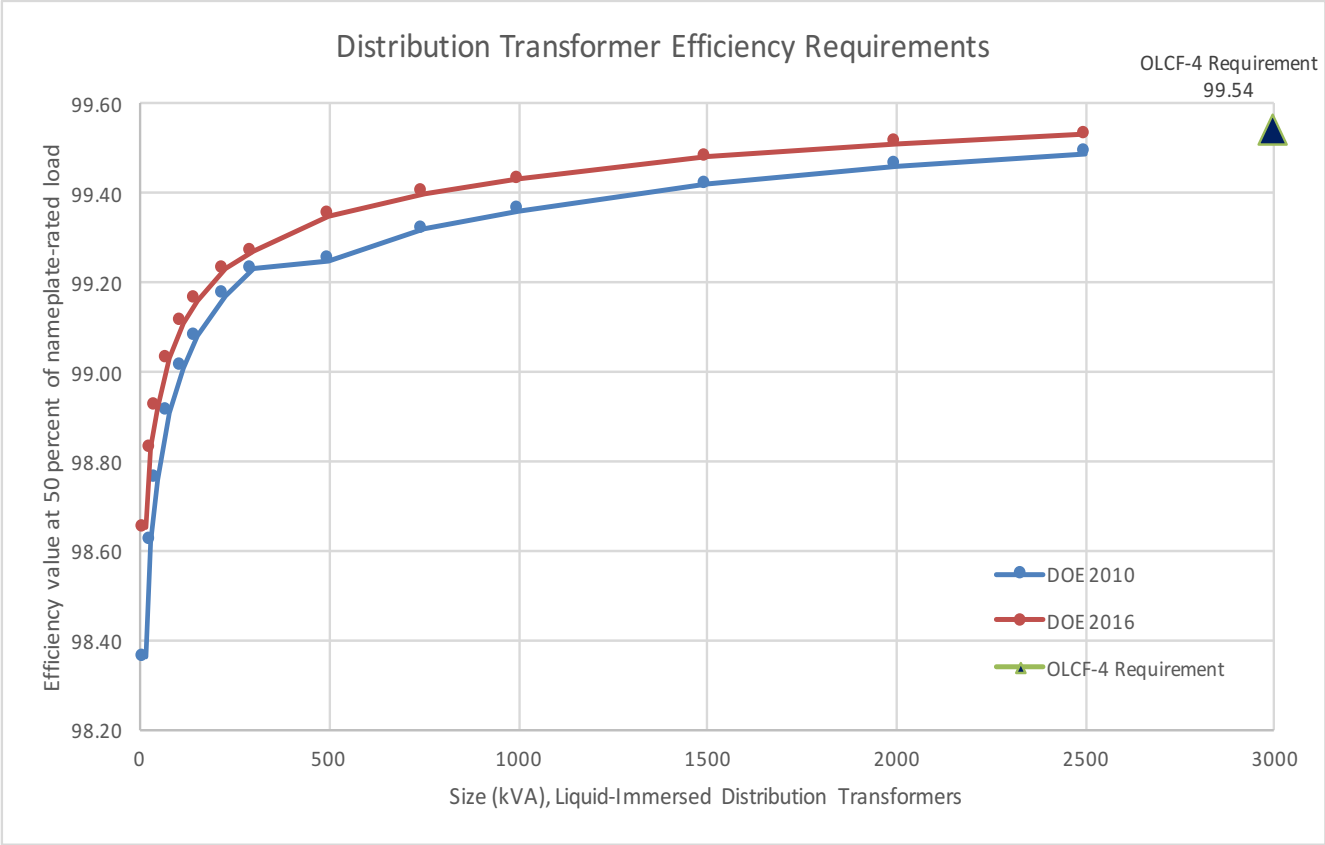
Transformer platforms provide room for (7) 3.5MVA Units

- 20 MW "warm water" cooling using heat exchangers, cooling towers, with cold-water trim from the existing CEP.

- 2016-2017- Prepare the room for the computers
  - Electrical distribution
  - Cooling Water Distribution (Aquatherm PP-R)
  - VESDA, sprinkler, life safety, lighting, cameras.

Cooling towers provide up to 5600 tons of capacity

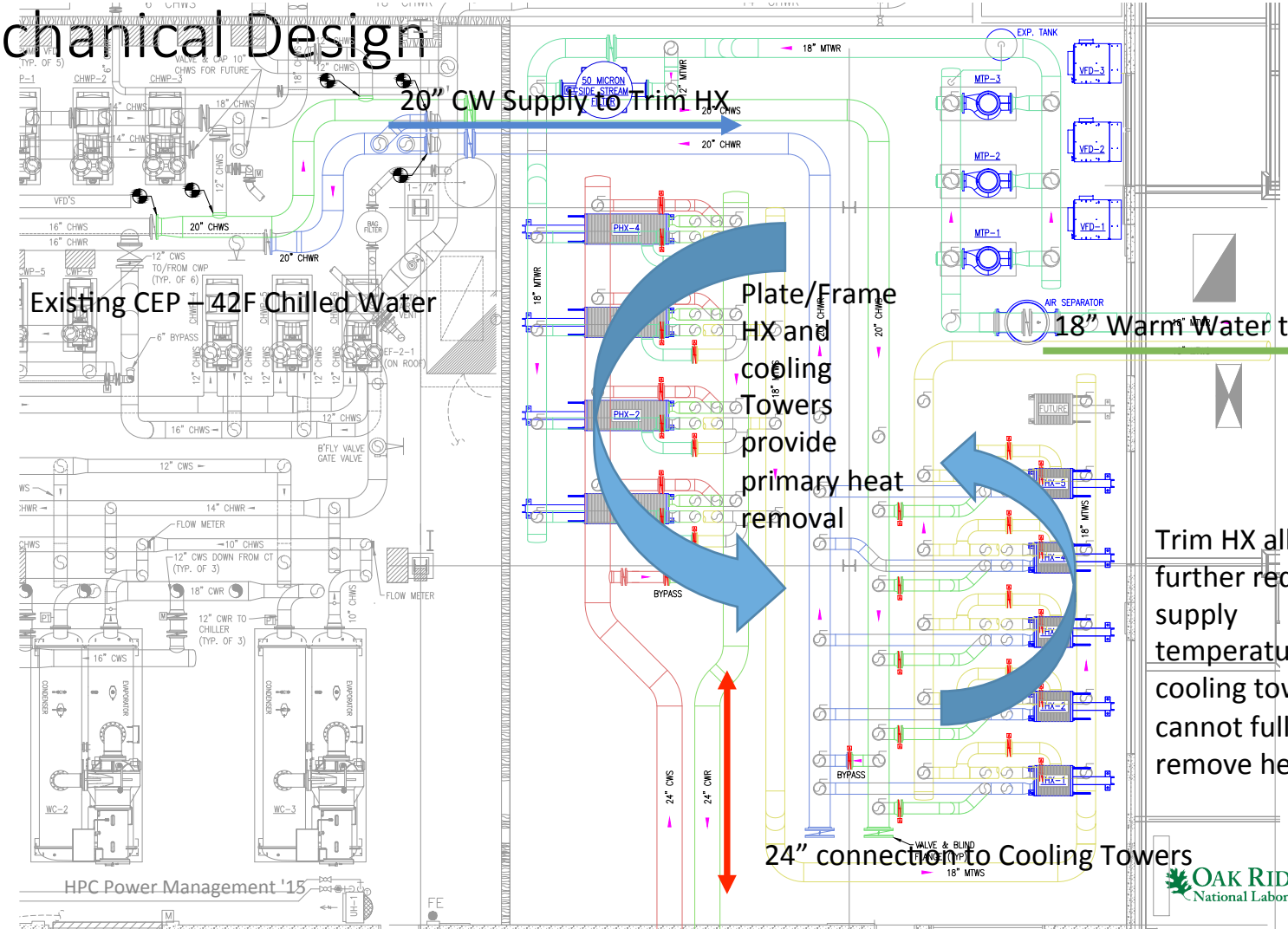
OLCF-4 Distribution Transformer Requirements



Beginning in 2016, new energy efficiency standards for distribution transformers will -

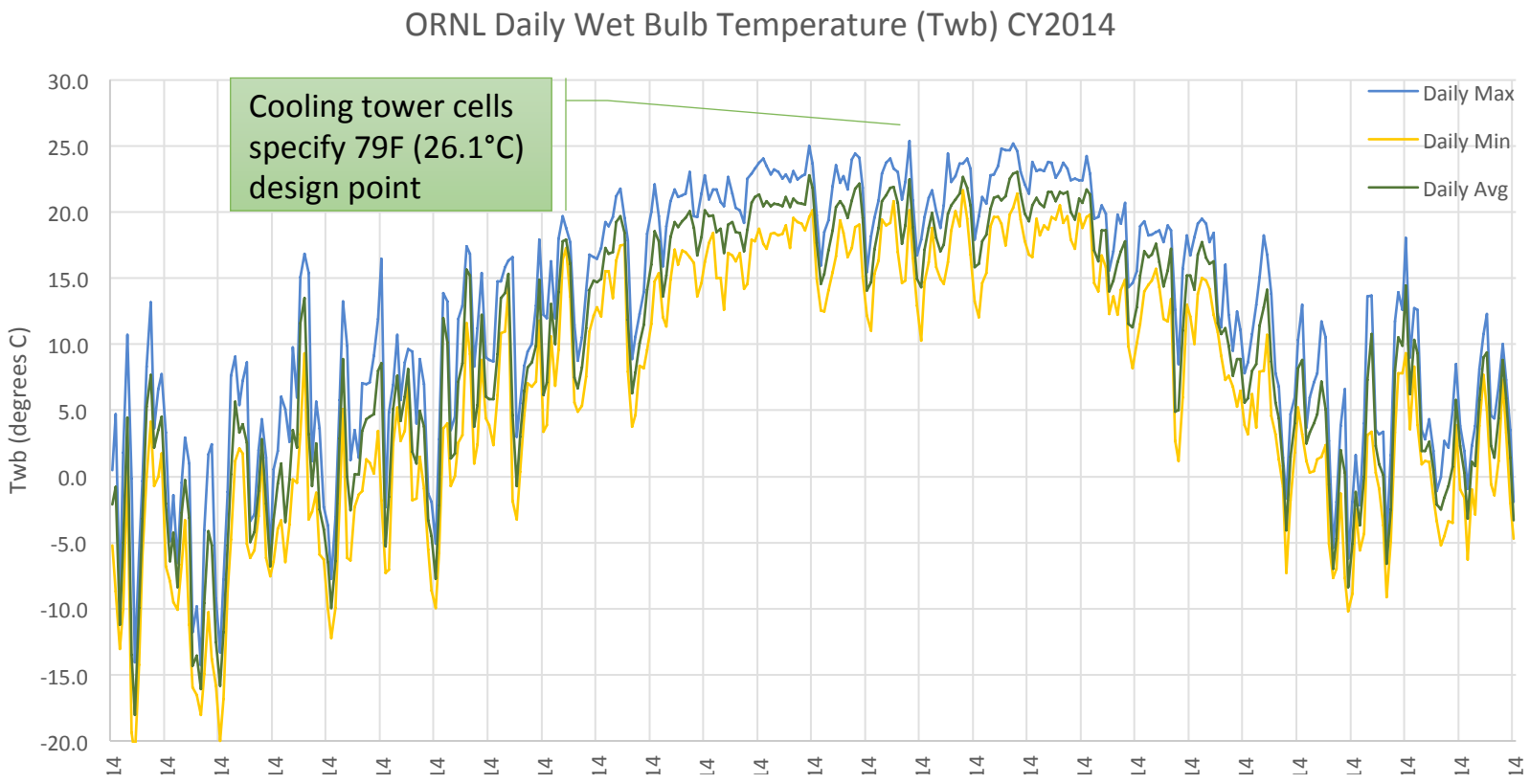
- Save 3.63 quadrillion BTUs ($1x10^6$ GW-h) of energy over the 30-year period 2016-2045.
- Decrease electrical losses by up to 8 percent for liquid-immersed transformers;
- Save up to $12.9 billion in total costs to consumers;
- Avoid 264.7 million metric tons of carbon dioxide emissions , the equivalent to the annual greenhouse gas emissions of about 51.75 million automobiles.

Mechanical Design



OAK RIDGE National Laboratory | OAK RIDGE LEADERSHIP COMPUTING FACILITY

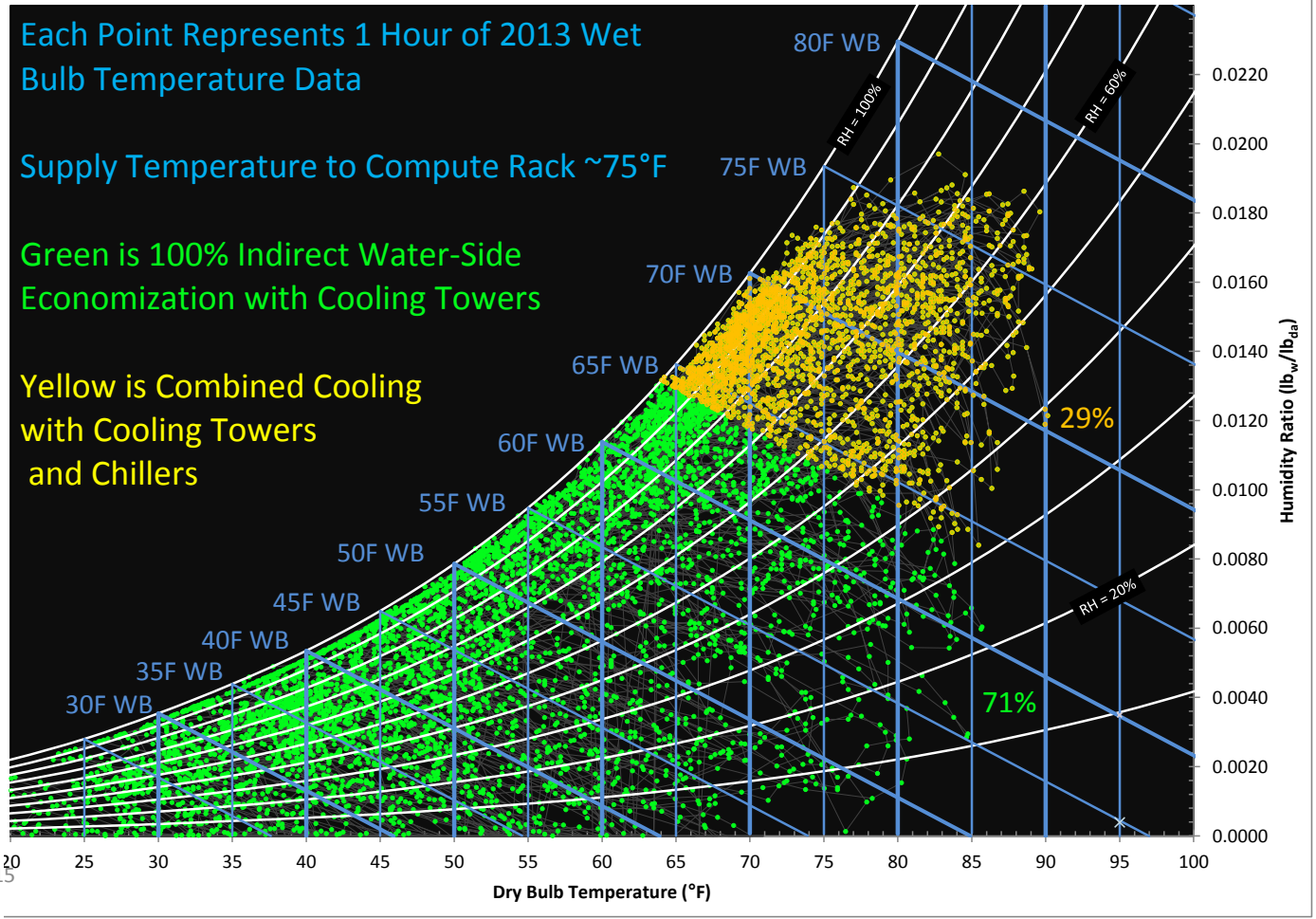
Environment Drives Evaporative Cooling Tower Selection



OAK RIDGE National Laboratory | OAK RIDGE LEADERSHIP COMPUTING FACILITY

Anticipated Water Side Economization for MWT Loop

- Assumes that ORNL can maintain operating conditions with a 75°F (23.9°C) supply.
- 71% of the hours in the year are fully 100% water-side economization.
- Contribution from chillers is substantially reduced->
  - Reduced operating cost
  - Better PUE (mostly pumping energy)

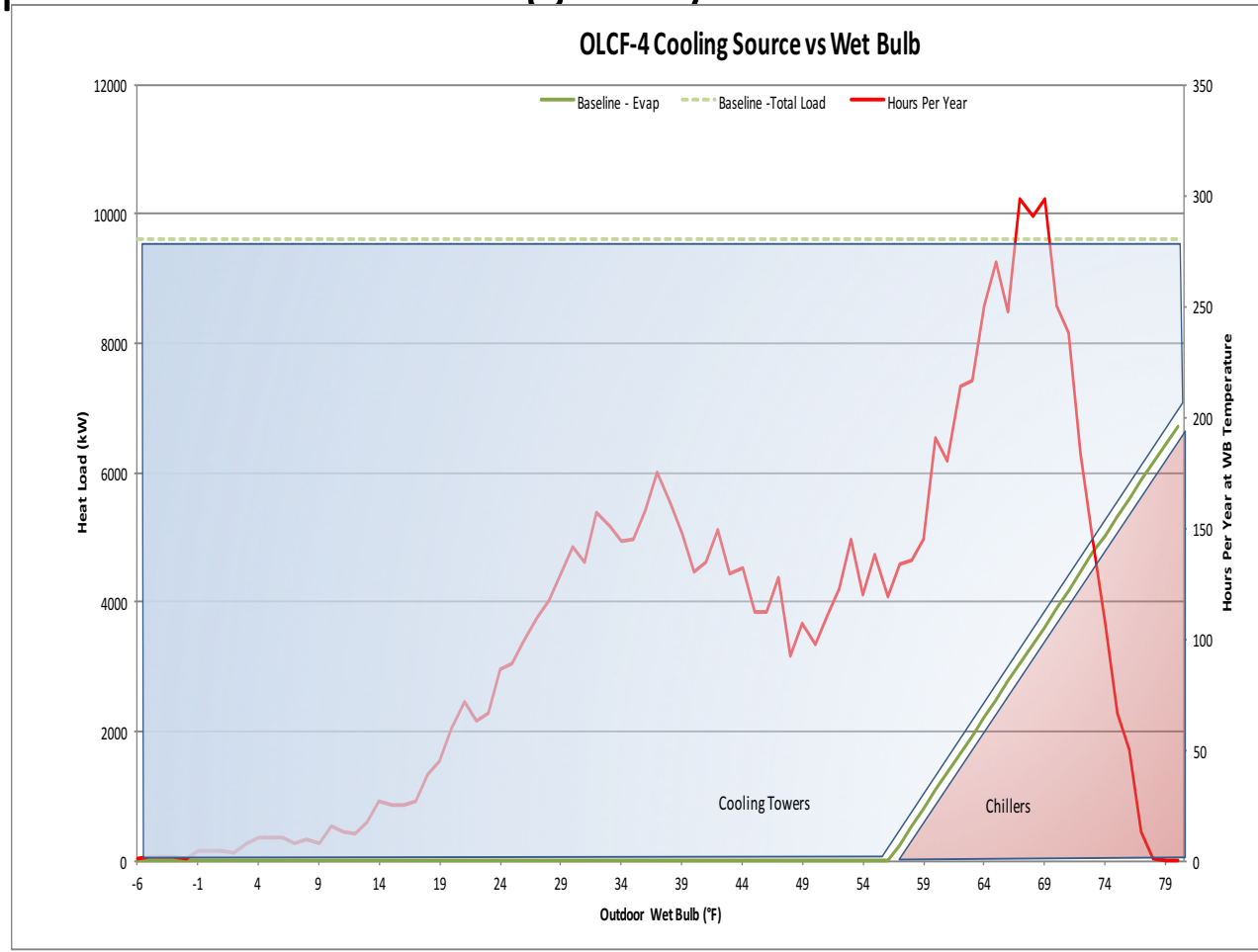


Predicting Evaporative Cooling Days

- Using a nominal design point:
  - 69°F/20.6°C supply temperature,
  - Cooling towers designed for 79°F (26°C) wet bulb,

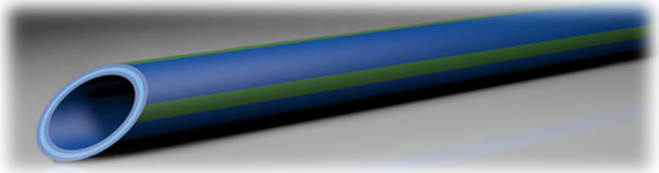
we anticipate significant 100% evaporative cooling days, with modest trim during worst case wet bulb conditions.

Raising the supply temperature (under evaluation) should further increase evaporative cooling envelope.



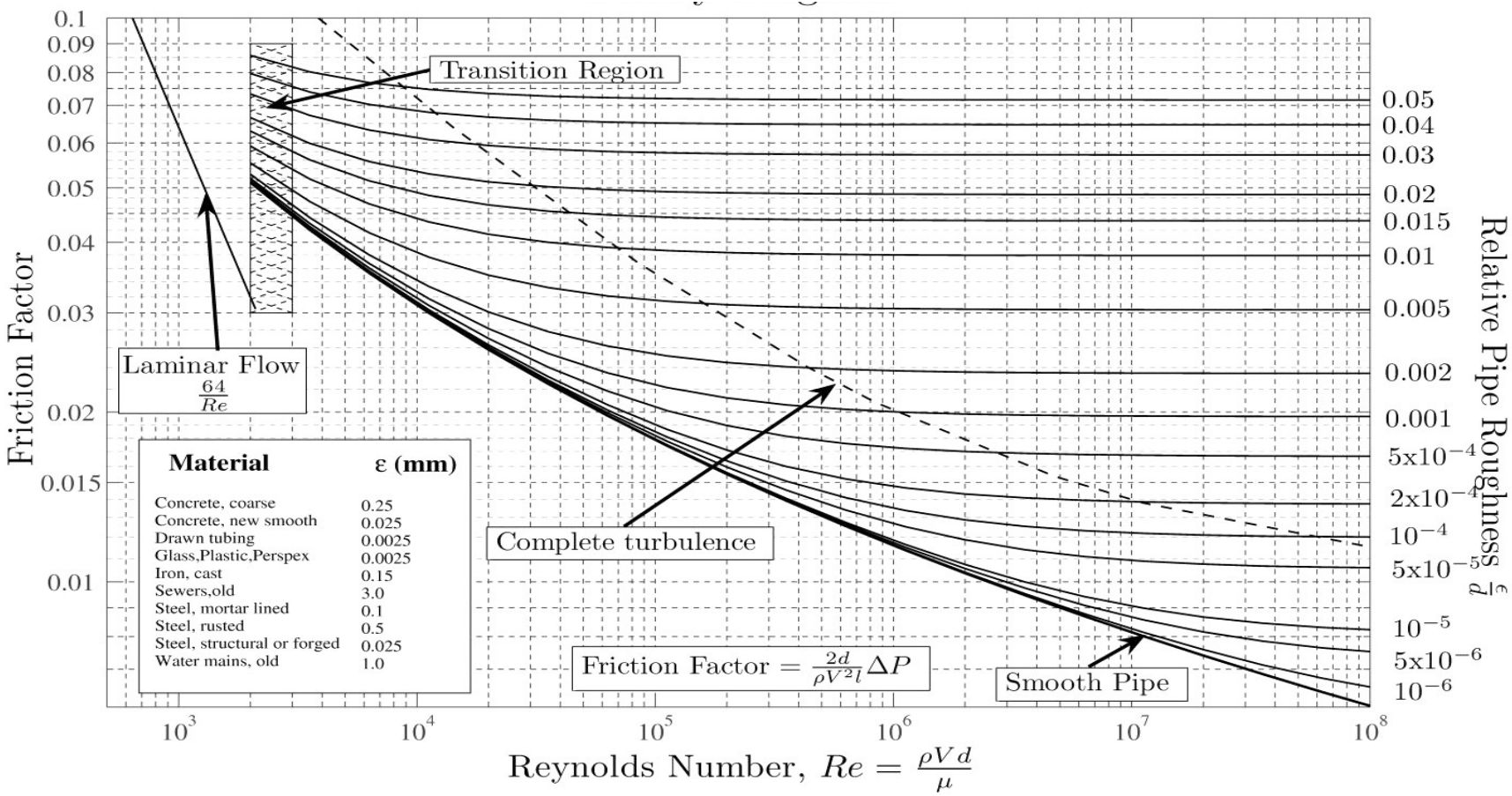
Reducing Energy Loss Through Material Selection

- Separate from the 20MW warm water cooling design of the new MEP
  - Secondary loop touch point in the MEP is the 18” Main to the computer room
- Constructed using Polypropylene-random (PP-R) thermoplastic
  - Sizes to 24”
  - Strong/Light. Impact Resistant
  - PP-R has a smooth interior surface, reducing turbulence, head pressure loss, and energy loss.
  - Reduced linear expansion vs. steel
  - Non-corroding, non-scaling (less chemical treatment)
  - Heat-fused fittings and components
  - Higher material cost than steel, but significantly reduced labor



- Summit: 18” Main (overhead) supplies branch headers; each branch header supplies a specific number of cabinets.
  - Design must accommodate the largest subcontract option (~7000 compute nodes)
  - Branch header design remains TBD. Potentially use a Tichelmann method to eliminate local flow control

Moody Diagram - friction factor, Reynolds number and relative roughness for fully developed flow in a circular pipe

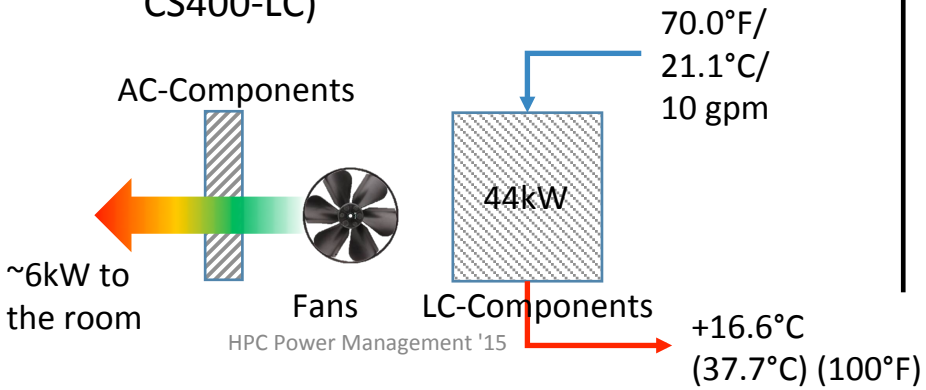


Material	ε (mm)
Concrete, coarse	0.25
Concrete, new smooth	0.025
Drawn tubing	0.0025
Glass,Plastic,Perspex	0.0025
Iron, cast	0.15
Sewers,old	3.0
Steel, mortar lined	0.1
Steel, rusted	0.5
Steel, structural or forged	0.025
Water mains, old	1.0

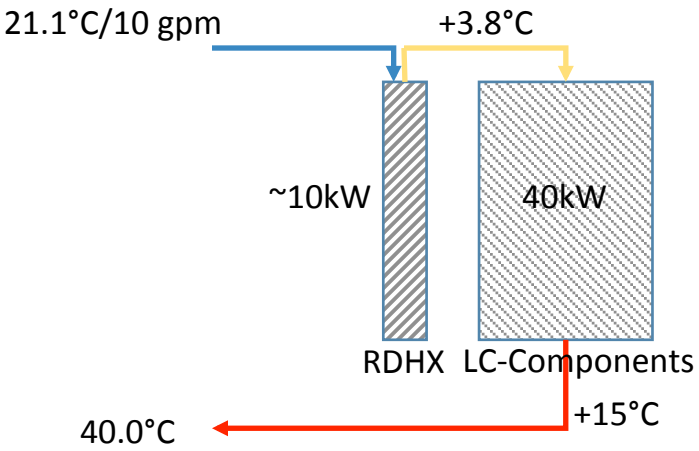
ACILITY

Different Approaches to Requirements for a 50kW Cabinet

- Parasitic Load to Air; Processor Load to Water
- Fans, power supplies, HCAs, and other residual heat is extracted with fans, and dumped to the room.
- CPU, GPU, and potentially DIMM loads are dumped to water.
  - Example- Asetek CDU (used in the Cray CS400-LC)

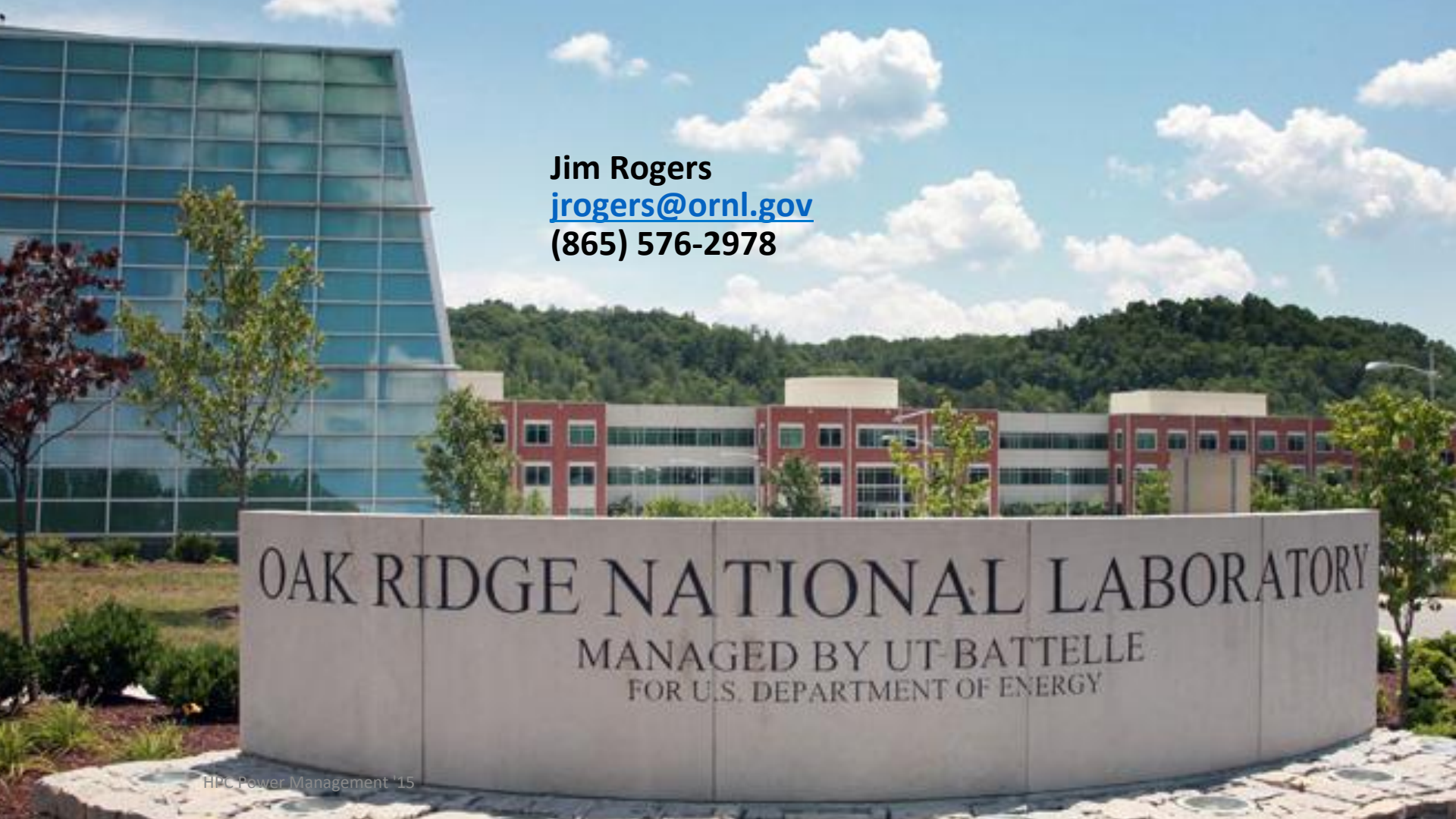


- Parasitic and Processor Loads to Water
- Summit - POWER9 and NVIDIA GPUS are liquid-cooled; Parasitic loads to RDHX.



All kW values are estimates for example purposes ONLY

OAK RIDGE National Laboratory | OAK RIDGE LEADERSHIP COMPUTING FACILITY



Jim Rogers
jrogers@ornl.gov
(865) 576-2978