A CORAL SYSTEM AND IMPLICATIONS FOR FUTURE HPC HARDWARE AND DATA CENTERS

Michael K Patterson
Senior Principal Engineer; Power, Packaging & Cooling
Intel, Technical Computing Systems Architecture and Pathfinding

Acknowledgement

Contributors, borrowed content, and data sources

Mark Seager, Intel

Ram Nagappan, Intel

Susan Coghlan, Argonne National Lab

Helmut Satzger, LRZ, Munich

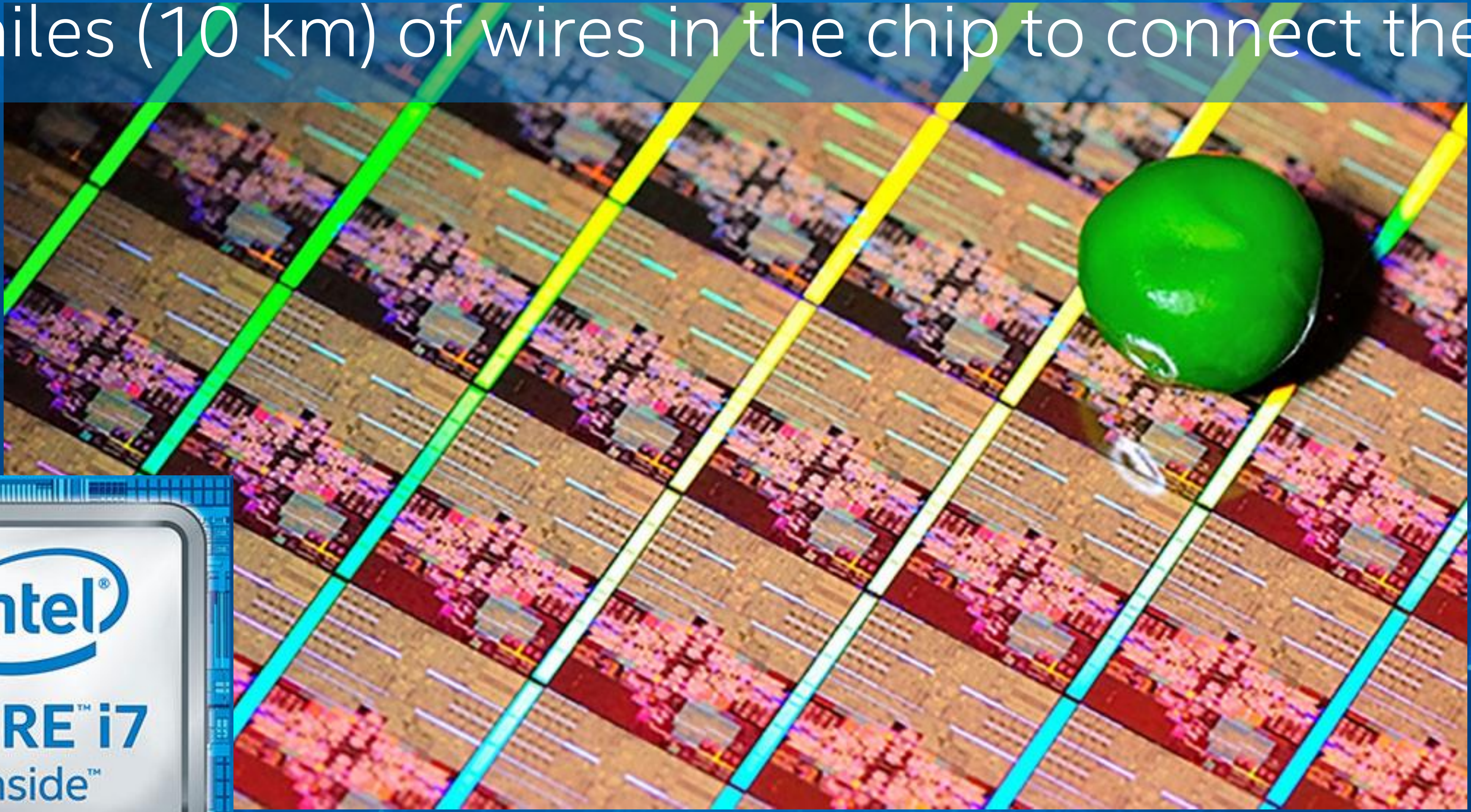
Jim Rogers, ORNL

Cray; Aurora system partners

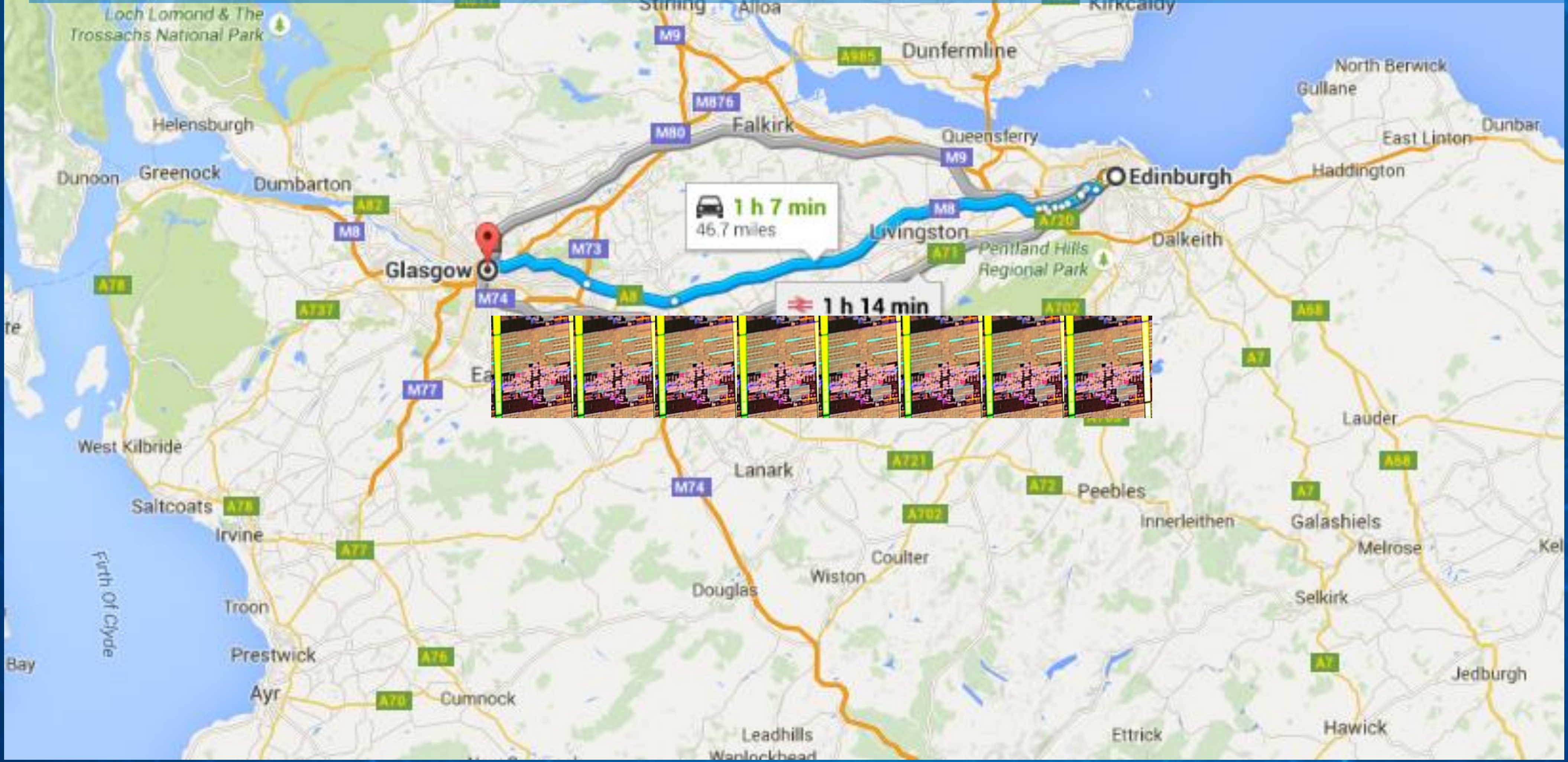
Appreciation

NCAR Team for continuing to invite us.

Intel Core™ i7 with over 1 billion transistors and over 6 miles (10 km) of wires in the chip to connect them



It would only take **eight** Core™ i7s to make a wire long enough to connect Glasgow to Edinburgh!



Intel Investments Paving the Way

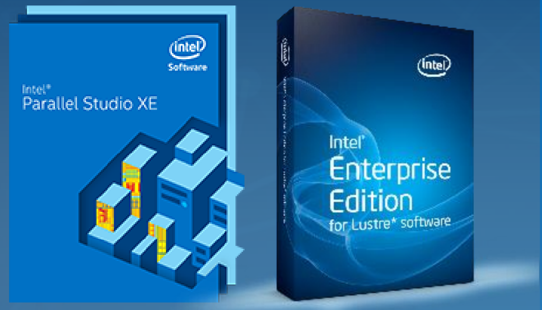
Holistic Approach to Cluster Solutions

CPU

Intel® Xeon® Processors

Intel® Xeon Phi™ Product Family

Software & Tools



Intel® Parallel Studio

Intel® Enterprise Edition for Lustre* software

Fabric

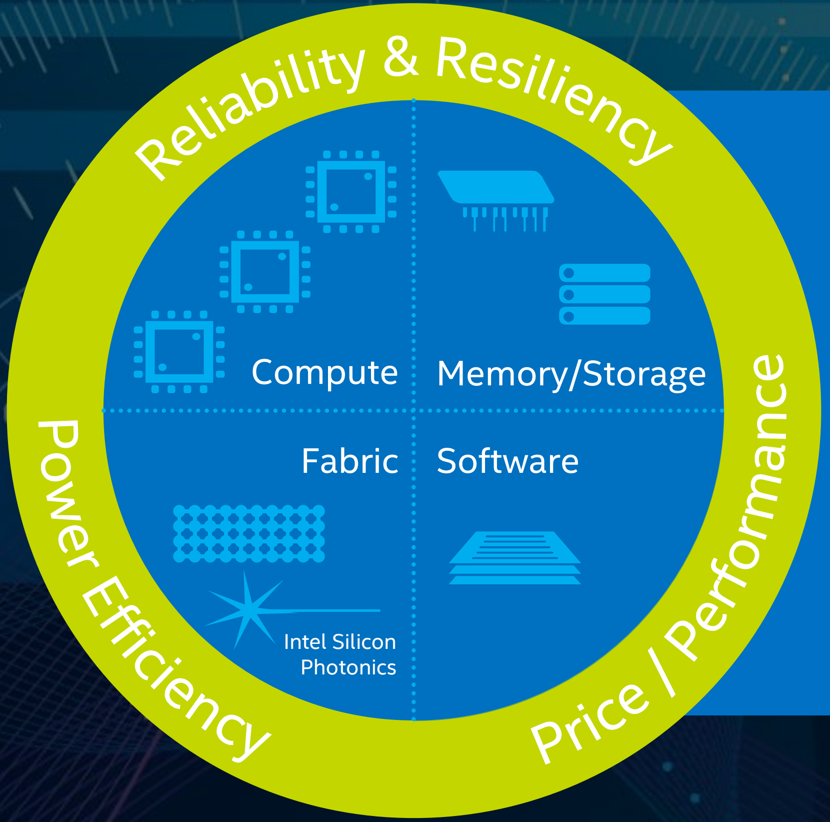
Intel® Omni-Path Architecture

Storage

Intel® Solid-State Drives (NVMe)

Intel's Scalable System Framework

A Configurable Design Path Customizable for a Wide Range of HPC & Big Data Workloads



- Small Clusters Through Supercomputers
- Compute and Data-Centric Computing
- Standards-Based Programmability
- On-Premise and Cloud-Based

Intel® Xeon® Processors
Intel® Xeon Phi™ Coprocessors
Intel® Xeon Phi™ Processors

Intel® True Scale Fabric
Intel® Omni-Path Architecture
Intel® Ethernet

Intel® SSDs
Intel® Lustre-based Solutions
Intel® Silicon Photonics Technology

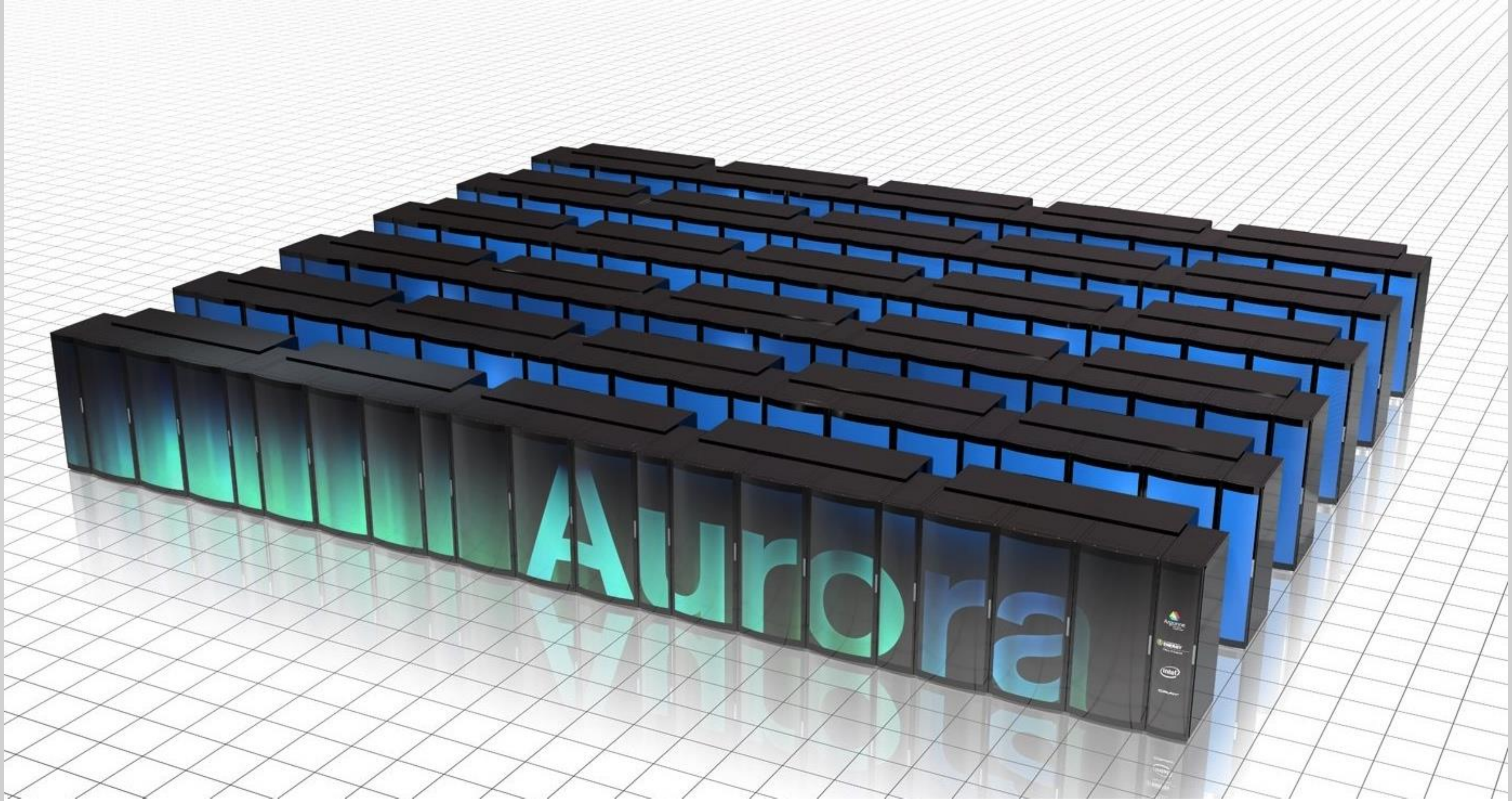
Intel® Software Tools
HPC Scalable Software Stack
Intel® Cluster Ready Program

CORAL

COLLABORATION OF OAK RIDGE, ARGONNE, AND LIVERMORE

- Acquire DOE 2018 – 2022 Leadership Computing Capability
- Three leadership class systems – one each at ALCF, LLNL, OLCF
  - With arch diversity between ALCF and OLCF
- ALCF: Intel (Prime) Cray (Integrator)
- OLCF: IBM (Prime)
- LLNL: IBM (Prime)

THE FUTURE



The Most Advanced Supercomputer Ever Built

An Intel-led collaboration with ANL and Cray to accelerate discovery & innovation

>180 PFLOPS
(option to increase up to 450 PF)

>50,000 nodes

13MW

2018 *delivery*

18X higher performance*

>6X more energy efficient*

Argonne NATIONAL LABORATORY

intel

Prime Contractor

CRAY

Subcontractor

Source: Argonne National Laboratory and Intel. *Versus ANL's current biggest system named MIRA (10PFs and 4.8MW)

Aurora | Built on a Powerful Foundation

Breakthrough technologies that deliver massive benefits

Compute

3rd Generation
Intel® Xeon Phi™

intel inside XEON PHI™

>17X performance†

FLOPS per node

>12X memory bandwidth†

>30PB/s aggregate in-package memory bandwidth

Integrated Intel® Omni-Path Fabric

Processor code name: Knights Hill

Interconnect

2nd Generation
Intel® Omni-Path

>20X faster†

>500 TB/s bi-section bandwidth

>2.5 PB/s aggregate node link bandwidth

File System

Intel® Lustre®*

>3X faster†

>1 TB/s file system throughput

>5X capacity†

>150TB file system capacity

Source: Argonne National Laboratory and Intel
*Other names and brands may be claimed as the property of others.

† Comparisons are versus Mira—Argonne National Laboratory's current largest HPC system, Mira. See Aurora Fact Sheet for details

Aurora Fact Sheet

System Feature	The Aurora Details	Comparison to Mira
Peak System Performance (FLOP/s)	180 - 450 PetaFLOP/s	10 PetaFLOP/s
Processor	Future Generation Intel® Xeon Phi™ Processor (Code name: Knights Hill)	PowerPC A2 1600 MHz processor
Number of Nodes	>50,000	49,152
Compute Platform	Intel system based on Cray Shasta next generation supercomputing platform	IBM Blue Gene/Q
Aggregate High Bandwidth On-Package Memory, local Memory and Persistent Memory	>7,000 Terabytes	768 Terabytes
Aggregate High Bandwidth On-Package Memory Bandwidth	>30 Petabytes/s	2.5 Petabytes/s
System Interconnect	2nd Generation Intel® Omni-Path Architecture with silicon photonics	IBM 5D torus interconnect with VCSEL photonics
Interconnect Aggregate Node Link Bandwidth	>2.5 Petabytes/s	2 Petabytes/s
Interconnect Bisection Bandwidth	>500 Terabytes/s	24 Terabytes/s
Interconnect Interface	Integrated	Integrated
Burst Buffer Storage	Intel® SSDs, using both 1st and 2nd Generation Intel® Omni-Path Architecture	None
File System	Intel® Lustre File System	IBM GPFS File System
File System Capacity	>150 Petabytes	26 Petabytes
File System Throughput	>1 Terabyte/s	300 Gigabyte/s
Intel Architecture (Intel® 64) Compatibility	Yes	No
Peak Power Consumption	13 Megawatts	4.8 Megawatts
FLOP/s Per Watt	>13 GigaFLOP/s per watt	>2 GigaFLOP/s per watt
Delivery Timeline	2018	2012
Facility Area for Compute Clusters	~3,000 sq. ft.	~1,536 sq. ft.

Argonne NATIONAL LABORATORY CRAY THE SUPERCOMPUTER COMPANY intel

For further information on Aurora, visit: intel.com/Aurora

Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

	Aurora
Processor	Xeon Phi™ Knights Hill
Nodes	>50,000
Performance	180 PF
Power	13 MW
Space	~3000 sq ft (~280 m²)
Cooling	Direct Liquid Cooling
Efficiency	>13 GF/w

All the details: **Aurora Fact Sheet** at **intel.com**

http://www.intel.com/content/www/us/en/high-performance-computing/aurora-fact-sheet.html?wapkw=aurora

Intel SSF enables Higher Performance & Density

A **formula** for more performance....

advancements in CPU architecture

✚ advancements in process technology

✚ integrated in-package memory

✚ integrated fabrics with higher speeds

✚ switch and CPU packaging under one roof

✚ all tied together with silicon photonics

= much higher performance & density

So what have we learned over the last three years?

Todays focus is on Power, Packaging, and Cooling (PPC)

- Power
  - 480Vac
  - >100 kW / cabinet
- Packaging
  - High density computing – significant computing in a small package
  - Weight becomes a key design parameter
- Cooling
  - Liquid cooling; for a number of reasons
  - Cooler is better, to a point

Power

Trends....

- Power now 480 Vac 3ph (400 Vac in Europe)
- >100 kW / cabinet
- In-cabinet 380 Vdc for optimized delivery
- Power management and power monitoring allows optimized performance and efficiency

Power Delivery Challenges in the horizon

Variable Power Cap

- Several reasons
  - Peak Shedding
  - Reduction in renewable energy

Power rate of change

- Ex: Hourly or Fifteen minute average in platform power should not exceed by X MW.

Controlled Power Ramp up/down – economic or technical issues

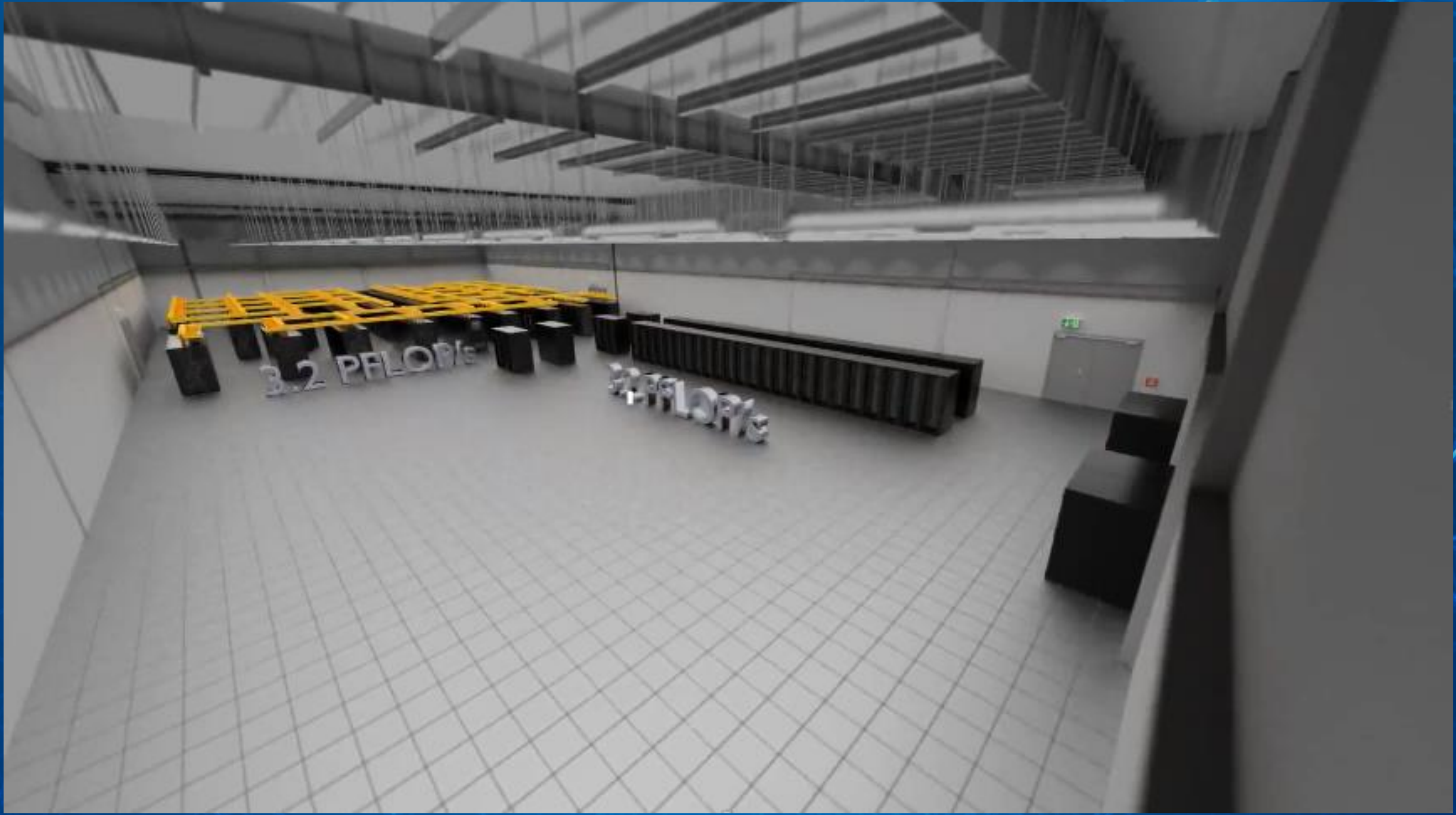
- Challenge to do this at a reasonable cost and with energy efficient mechanisms

Packaging

Rack and cluster weight and density

- Packaging
  - High density computing – network topology optimization and high node count per rack make for dense cabinets
- Rack weight density
  - Design limit: Floor tiles at 500 lbs/sf ~ 2500 kg/m2
- White space vs utility space
  - Compute density increasing, infrastructure support equipment is not
- What's the trend for machine room area?

I must need a huge data center for PetaScale and ExaScale computing – Right?


3.2 PFLOP/s

Quick Survey

In what year, do you expect to see "rack level" performance exceed 1 PF?

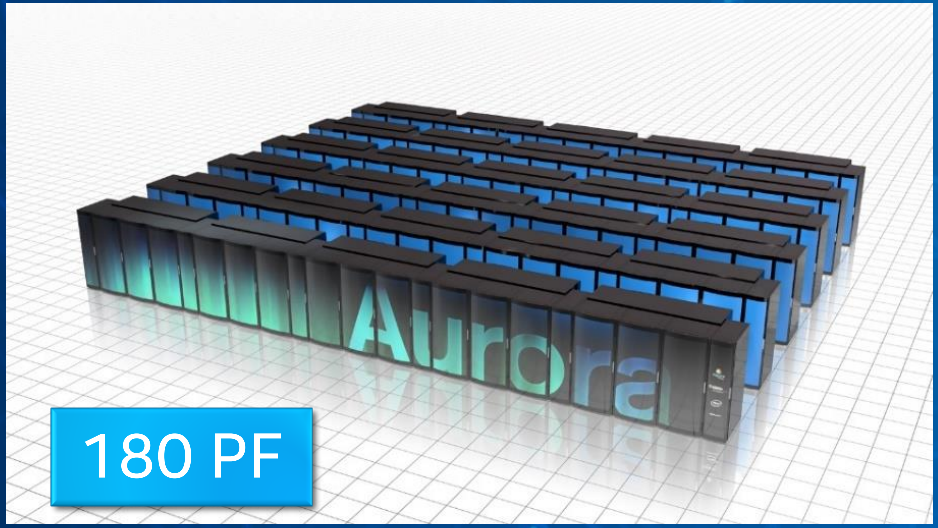
a) 2016

b) 2018

c) 2020

d) 2022

Performance density continues to increase



Source: lrz.de

System Feature	LRZ Phase 1	LRZ Phase 2	Mira	Titan	Summit	Aurora
Year	2012	2015	2012	2012	2017	2018
Perf Pflop/s	3.2	3.2	10	27	150	180
# of Nodes	9216	3096 Haswell	49,152	18,688	3500	>50,000 KNH
Power	2.3 MW	1.1 MW	4.8 MW	9 MW	9.6 MW	13 MW
Cluster Area (m^2) est.	546	182	143	400	418	279
Cluster Area (ft^2) est.	5875	1960	1540	4300	4500	3000
TF/m^2 Est.	6	18	70	67.5	359	645
TF/ft^2 est.	0.5	1.6	6.5	6.3	33.4	60

Do I need a huge data center?

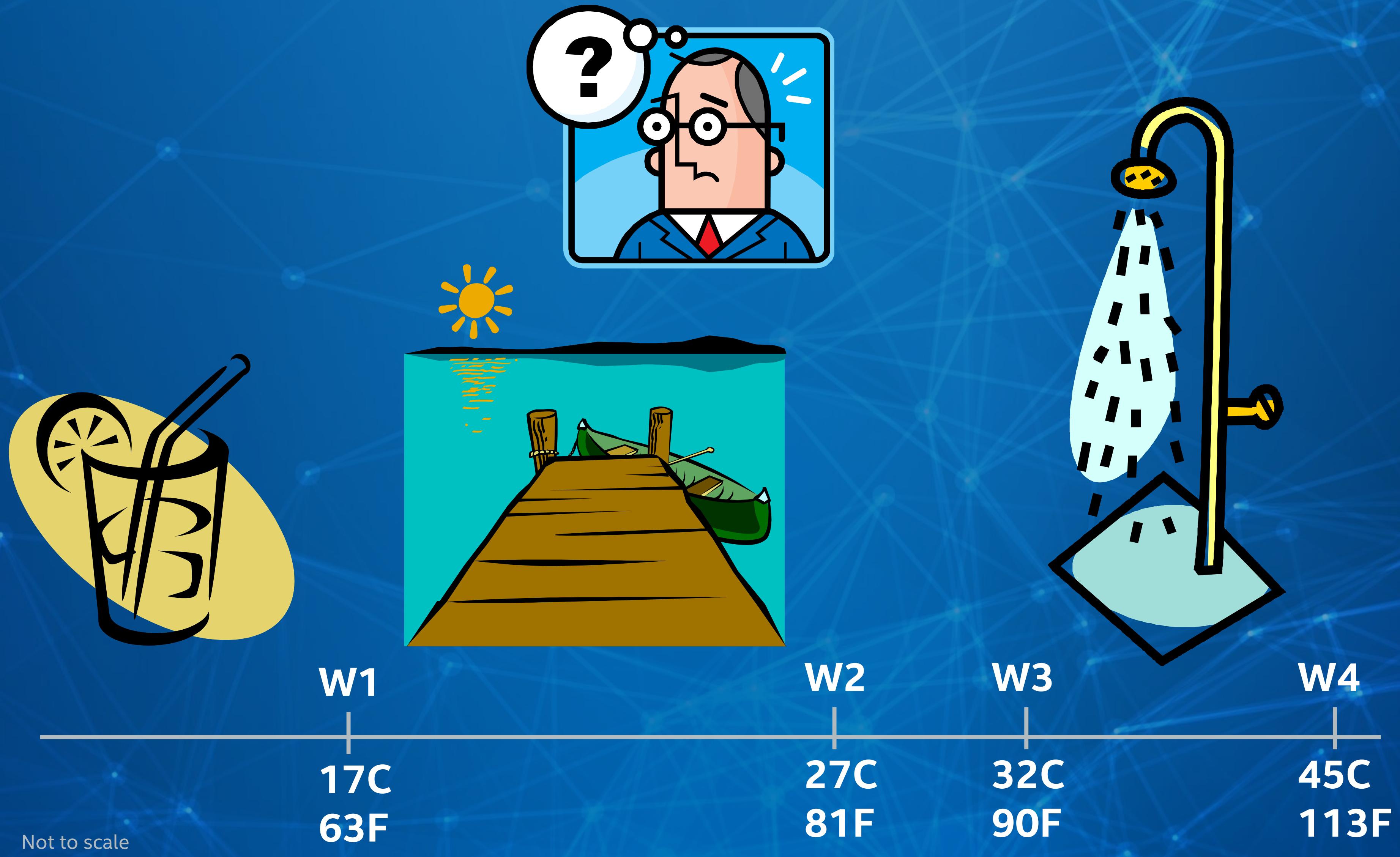
- Facility area for Compute Cluster does not have to be huge. Significant compute density in small packages
  - At Aurora density, the 3.2 LRZ PF step could fit in 5 m^2
- Don't forget:
  - If Storage is going to be large then you will need additional floor space.
  - If you are going to be using Xeon instead of Xeon Phi then you may need additional floor space
  - Utility and infrastructure space continues to grow

Cooling

Why liquid?

- Power per node continues to rise
- Rack density limits airflow path
- Increased thermal performance of liquid (vs air) allows more free-cooling
  - Thermal resistance from chip to liquid in a cold plate is smaller than chip to air over a heat sink
- Warmer or cooler?   “Warm-water cooling” has a good ring to it!

What does Warm-Water Cooling really mean?



intel

Just say no….

Warm Water Cooling

Instead, define either specific temperatures or functional….

Define the temperature at the facilities and IT water loop interface

W2, W3, W4

ASHRAE values help system vendor design system, guarantee performance

Define how the water temperature is made

Chiller

Cooling Tower

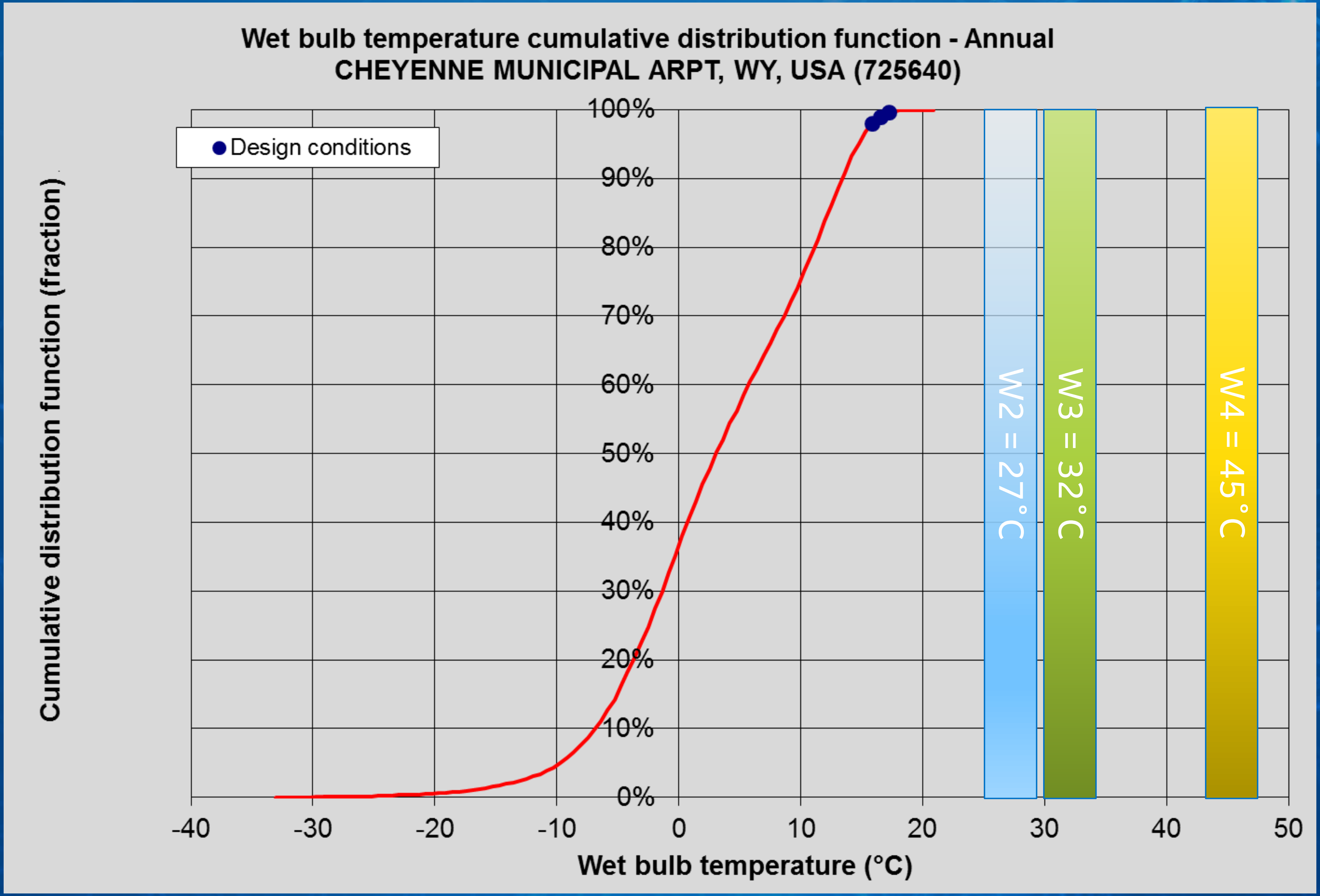
Dry Cooler

A proposal….

- As a starting point, use the coolest water you can make without a chiller
- Always be above the dewpoint (to prevent condensation in the machine)
- Cooler temperatures promote:
  - Lower leakage
  - More turbo frequencies
  - Higher stability
  - More time to recover in an upset condition
  - Better reliability
  - Reduced flow rates

Note - May consume more water, not applicable if after heat recovery

Why use "warm" water, when "cool" water costs the same?



Summary

Planning for Exascale needs to happen now; 180 PF in 2018

Designing for the future:

It's going to be Large!

kg/rack, kW/rack, perf/rack, power ramps and peak, pipe sizes, m2/m2

It may get Smaller!

Cluster footprint

High packaging density, high power, liquid cooling all enable best performance, efficiency, and TCO

Aurora

It's one more landmark.

It's the next one we have to reach.

But the journey does not stop there.

Argonne NATIONAL LABORATORY

intel

CRAY

Thanks for your attention

Questions?

michael.k.patterson@intel.com

Legal Disclaimer

Today's presentations contain forward-looking statements. All statements made that are not historical facts are subject to a number of risks and uncertainties, and actual results may differ materially.

NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL® PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT. INTEL PRODUCTS ARE NOT INTENDED FOR USE IN MEDICAL, LIFE SAVING, OR LIFE SUSTAINING APPLICATIONS.

Intel does not control or audit the design or implementation of third party benchmarks or Web sites referenced in this document. Intel encourages all of its customers to visit the referenced Web sites or others where similar performance benchmarks are reported and confirm whether the referenced benchmarks are accurate and reflect performance of systems available for purchase.

Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families. See www.intel.com/products/processor_number for details.

Intel, processors, chipsets, and desktop boards may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.