TRENDS IN HPC AND DATA CENTER POWER, PACKAGING, AND COOLING

Michael K Patterson, PhD, PE, DCEP

Power, Packaging, and Cooling
Intel, Technical Computing Systems Architecture and Pathfinding

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.

Intel, Intel Xeon, Intel Core microarchitecture, and the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

*Other names and brands may be claimed as the property of others

Copyright © 2011, Intel Corporation. All rights reserved.

Overview

Data Center Choices

IT & HPC Drivers

Metrics

Power

Cooling

Density

Resources

Data Center Choices

Air cooling vs liquid cooling?

Hot-aisle vs Cold-aisle?

Raised floor vs concrete?

Bricks and mortar vs containers?

New building vs existing?

UPS as part of HPC?

Rack density?

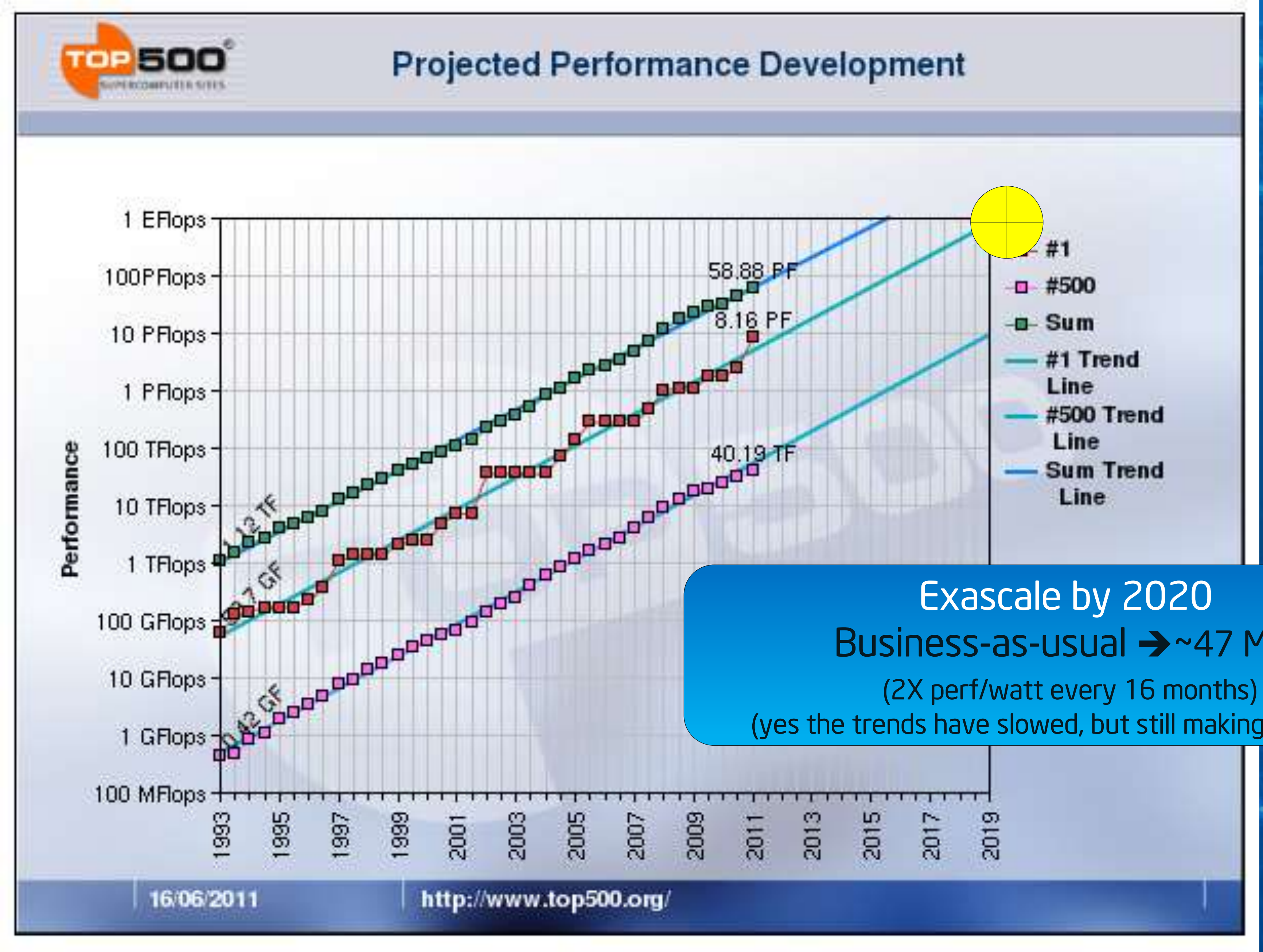
Feed from above or below?

1st cost or TCO?

Reliability level – Tier 1 To Tier IV?

HPC data centers; many hurdles

- Power and performance challenges exist to get to Exascale
- *Preparing for Exascale?* Aren't we a little early?
- The key facts…
    - Data Center life cycle – 10-15 years
    - HPC cluster life cycle – 3-5 years
- Leads to interesting results…



Exascale by 2020

Business-as-usual ➔~47 MW

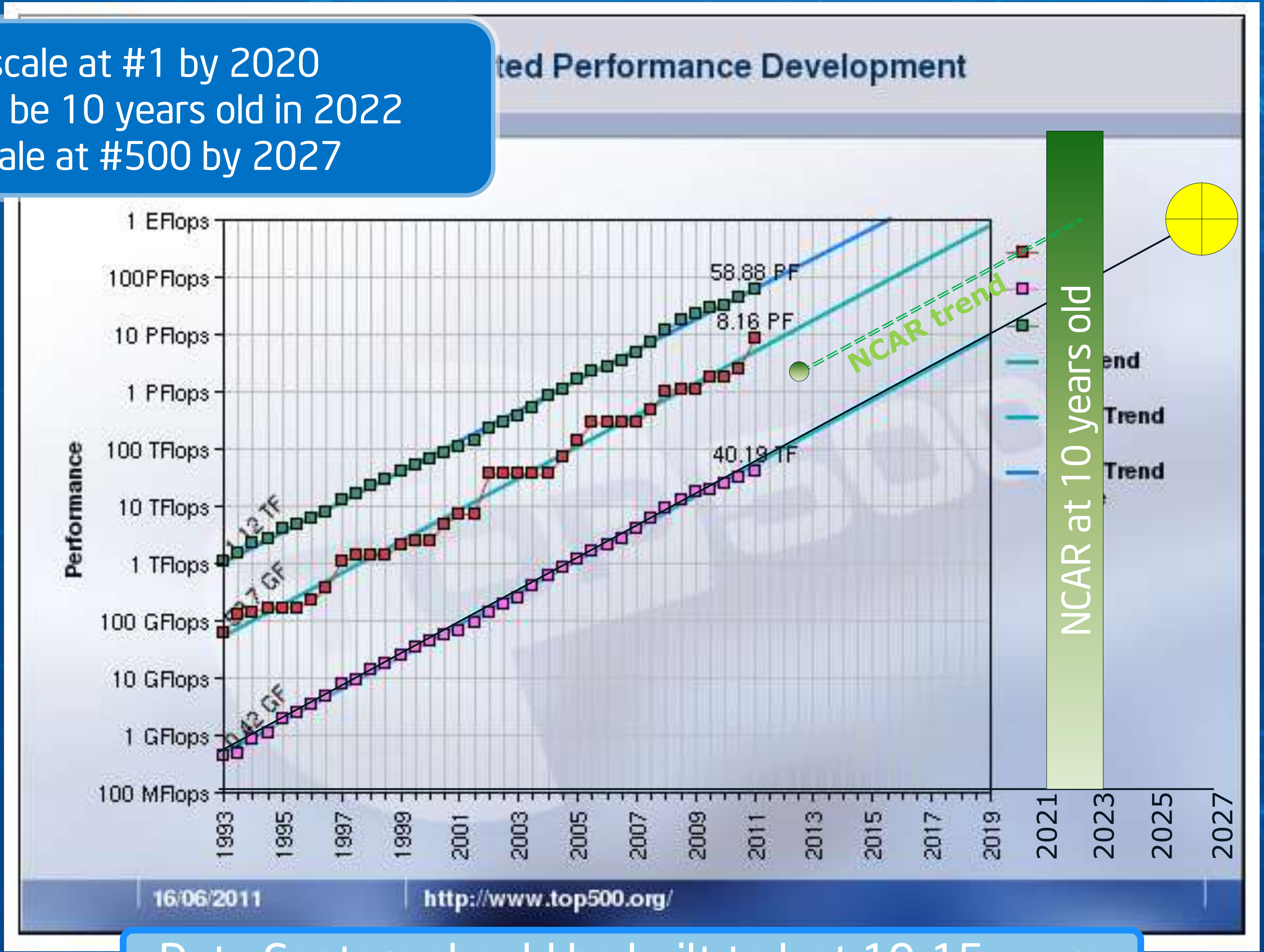
(2X perf/watt every 16 months)

(yes the trends have slowed, but still making progress)

NCAR - Home to an Exaflop SuperComputer

NCAR Yellowstone - New supercomputing center in Wyoming

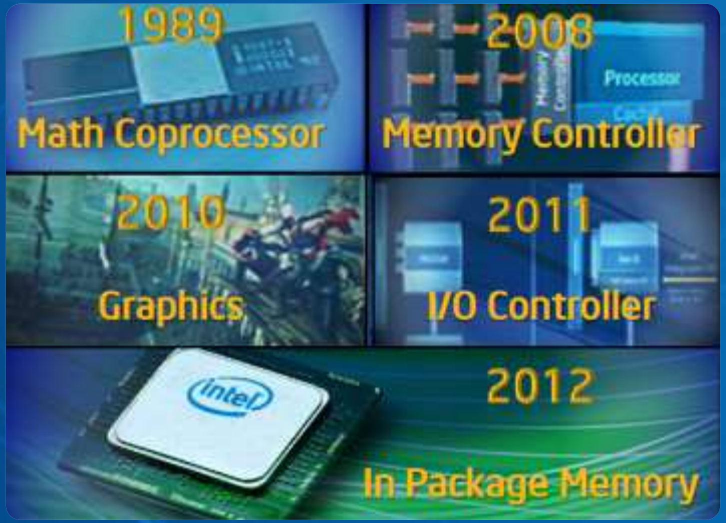
Exascale at #1 by 2020
NCAR will be 10 years old in 2022
Exascale at #500 by 2027



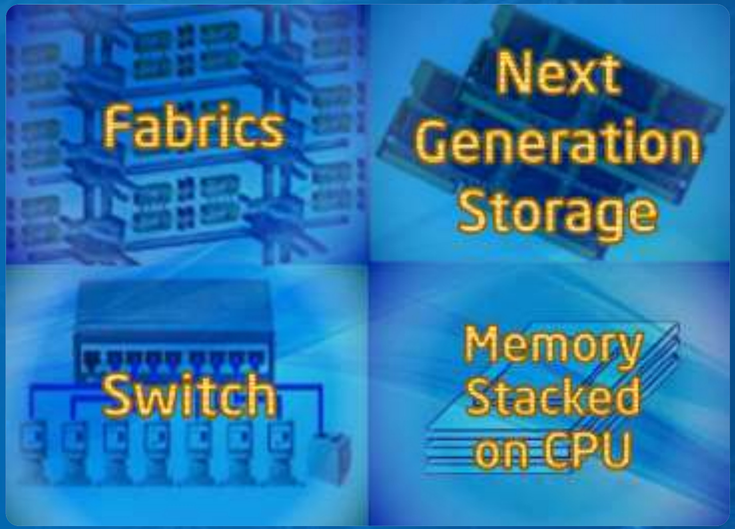
Data Centers should be built to last 10-15 years

Integration

Enabled by leading edge process technologies



Integrated Today



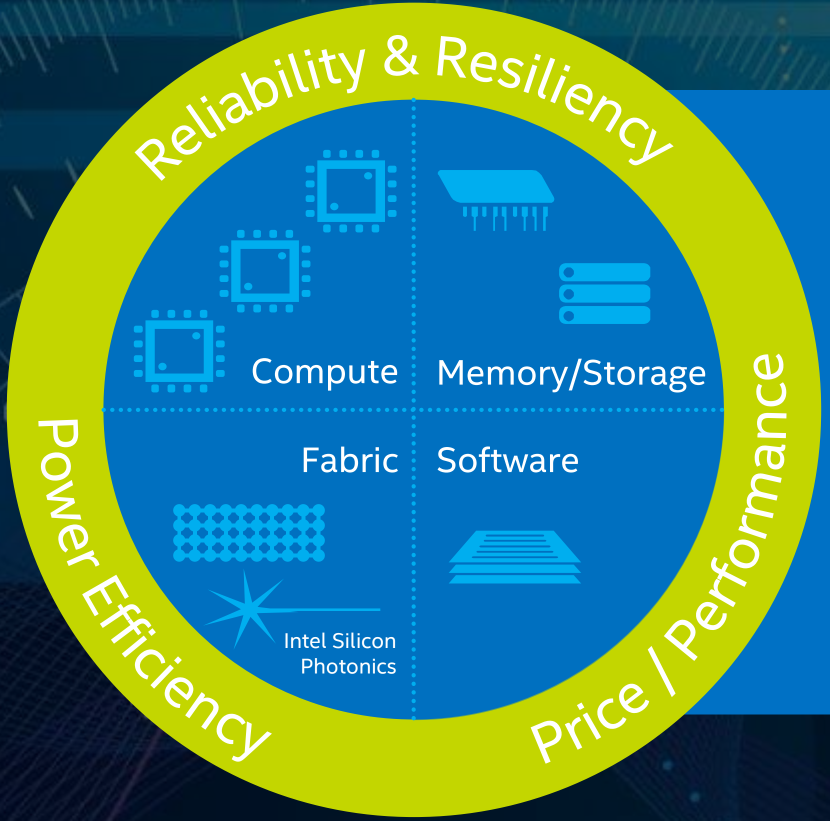
Possible Tomorrow**

System level benefits in cost, power, density, scalability & performance

**Future options are forecasts and subject to change without notice.

Intel's Scalable System Framework

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



Reliability & Resiliency

Compute

Memory/Storage

Fabric

Software

Intel Silicon Photonics

Power Efficiency

Price / Performance

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

Intel SSF enables Higher Performance & Density

A **formula** for more performance….

advancements in CPU architecture

✚ advancements in process technology

✚ integrated in-package memory

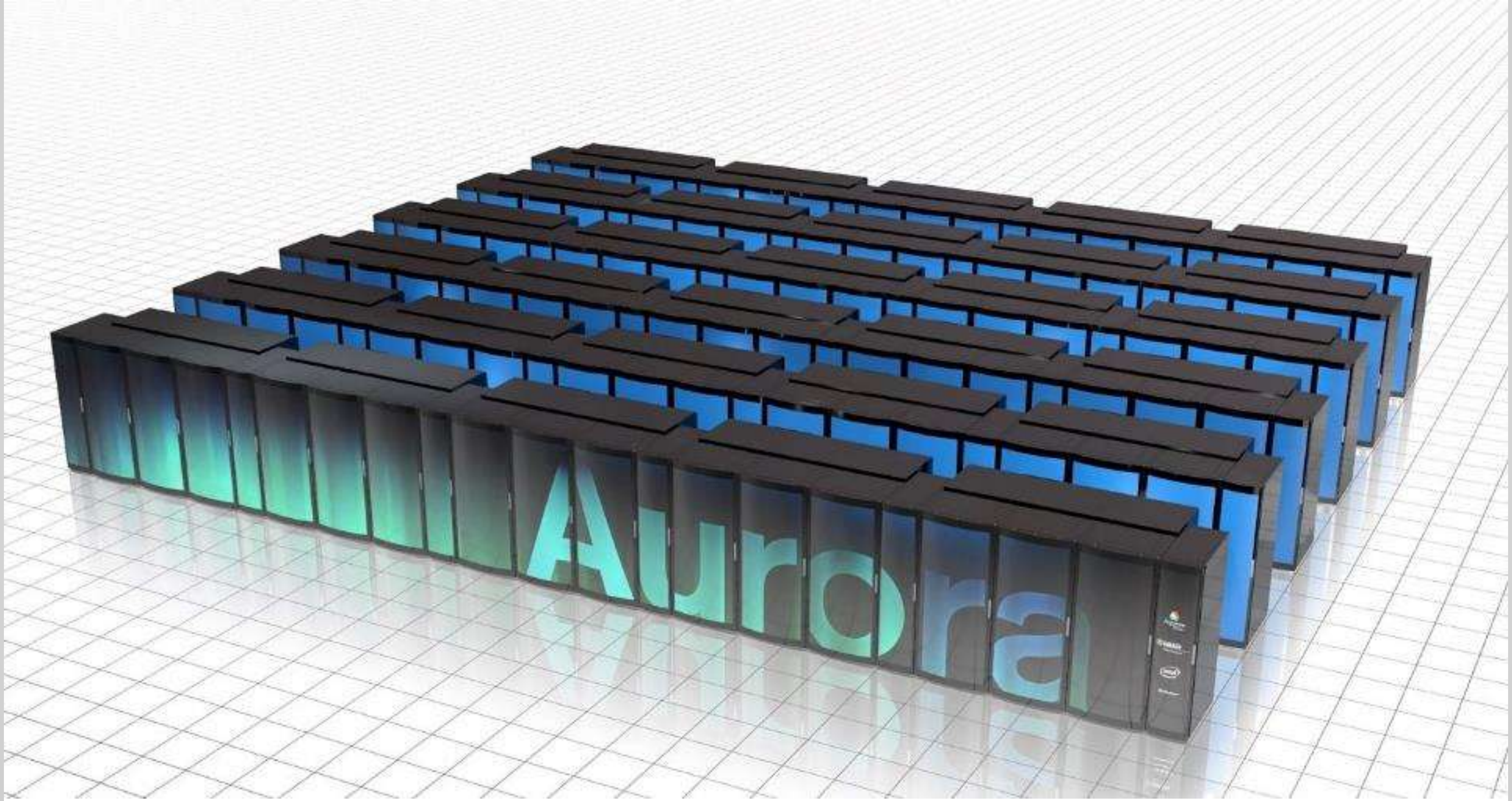
✚ integrated fabrics with higher speeds

✚ switch and CPU packaging under one roof

✚ all tied together with silicon photonics

= much higher performance & density

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†

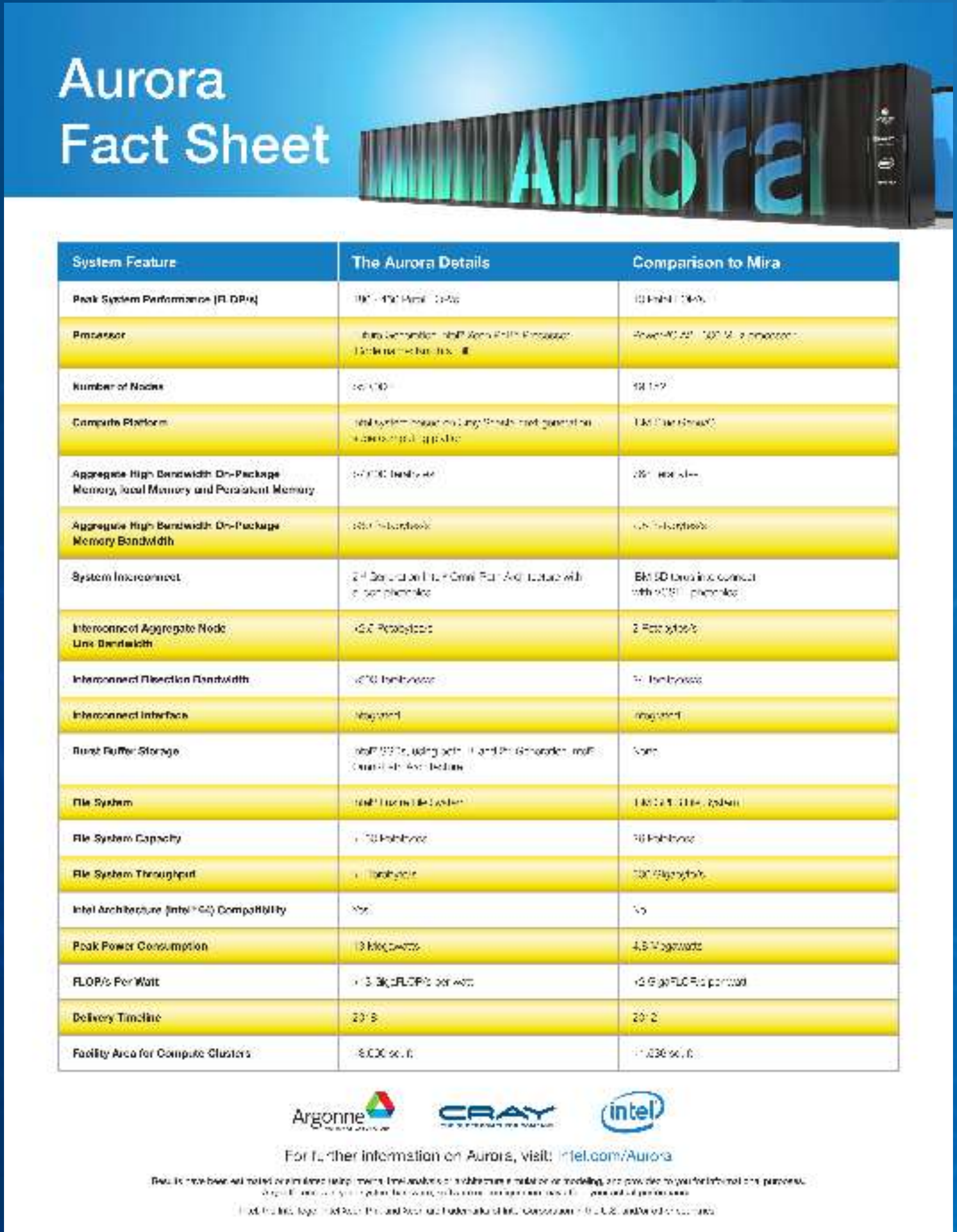
>1 TB/s file system throughput

>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 (FLOPs)	[illegible]	[illegible]
Processor	[illegible]	[illegible]
Number of Nodes	[illegible]	[illegible]
Compute Platform	[illegible]	[illegible]
Aggregate High Bandwidth On-Package Memory, local Memory and Persistent Memory	[illegible]	[illegible]
Aggregate High Bandwidth On-Package Memory Bandwidth	[illegible]	[illegible]
System Interconnect	[illegible]	IBM 5D torus interconnect with VCSEL photonics
Interconnect Aggregate Node Link Bandwidth	[illegible]	2 Petabytes/s
Interconnect Bisection Bandwidth	[illegible]	[illegible]
Interconnect Interface	Integrated	Integrated
Burst Buffer Storage	[illegible]	None
File System	[illegible]	[illegible]
File System Capacity	[illegible]	[illegible]
File System Throughput	[illegible]	[illegible]
Intel Architecture (Intel® 64) Compatibility	Yes	No
Peak Power Consumption	13 Megawatts	4.8 Megawatts
FLOP/s Per Watt	[illegible]	[illegible]
Delivery Timeline	2018	2012
Facility Area for Compute Clusters	[illegible]	[illegible]

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

	Aurora
Processor	Xeon Phi™ Knights Hill
Nodes	>50,000
Performance	180 PF Peak
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

How did we do this?

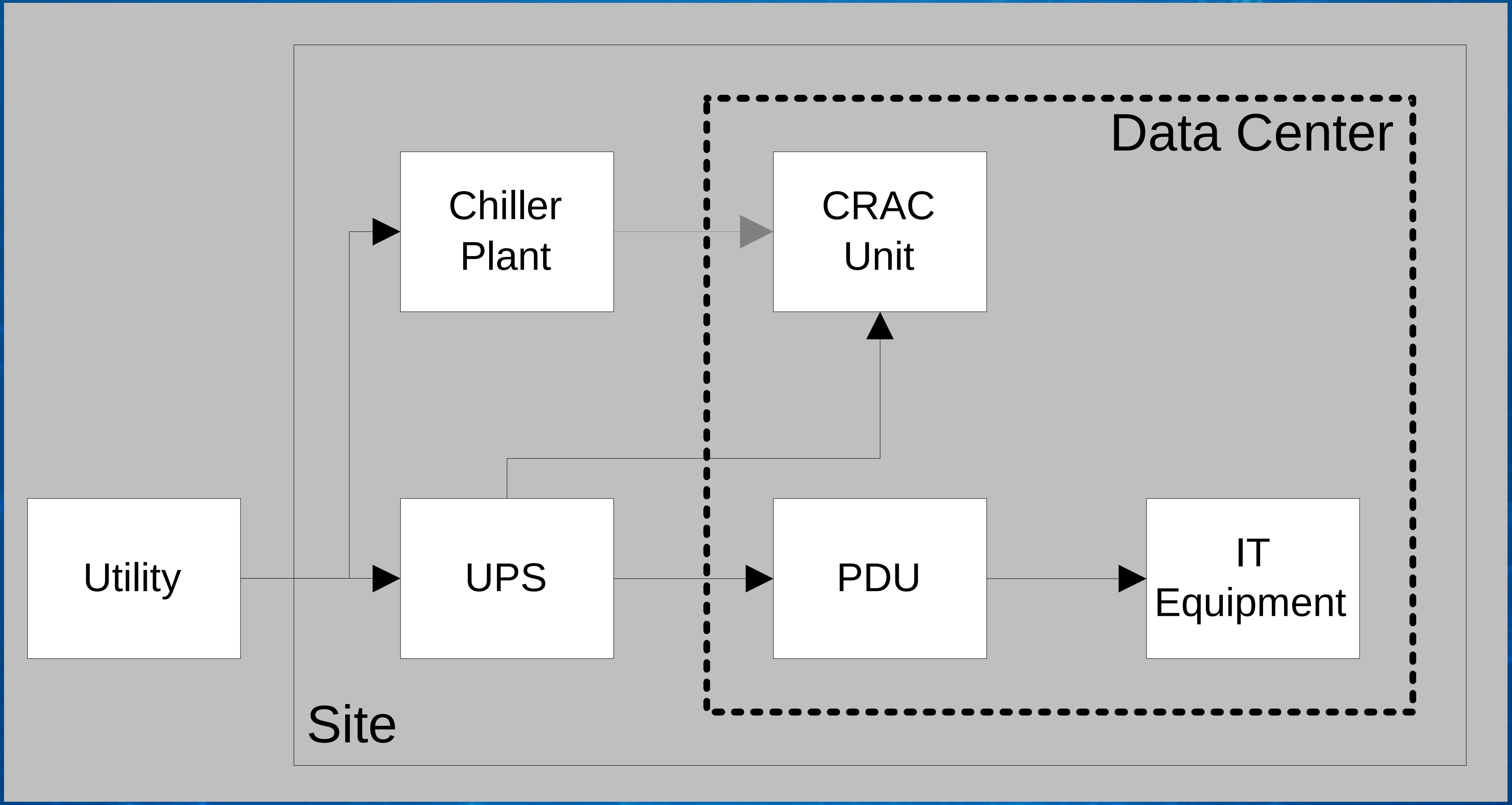
- In package memory
  - Closer to CPU, stacked memory
- Fabric Integration
  - Package connectivity
- Advanced Switches with higher radix and higher speeds
  - Closer integration of compute and switch
- Silicon Photonics
  - Low cost, outstanding performance but thermal challenges do exist
- All this drives changes

So what have we learned over the last three years?

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

- Metrics
  - How do we measure and compare?
- Power
  - 400Vac, 3ph, >100 kW / cabinet for the very high end
- Packaging
  - High density computing – significant computing in a small package
  - Weight becomes a key design parameter
- Cooling
  - Liquid cooling; good for some. Cooler is better, to a point
  - Aurora ~100% liquid cooled
  - Air cooling still very core to HPC

The Data center



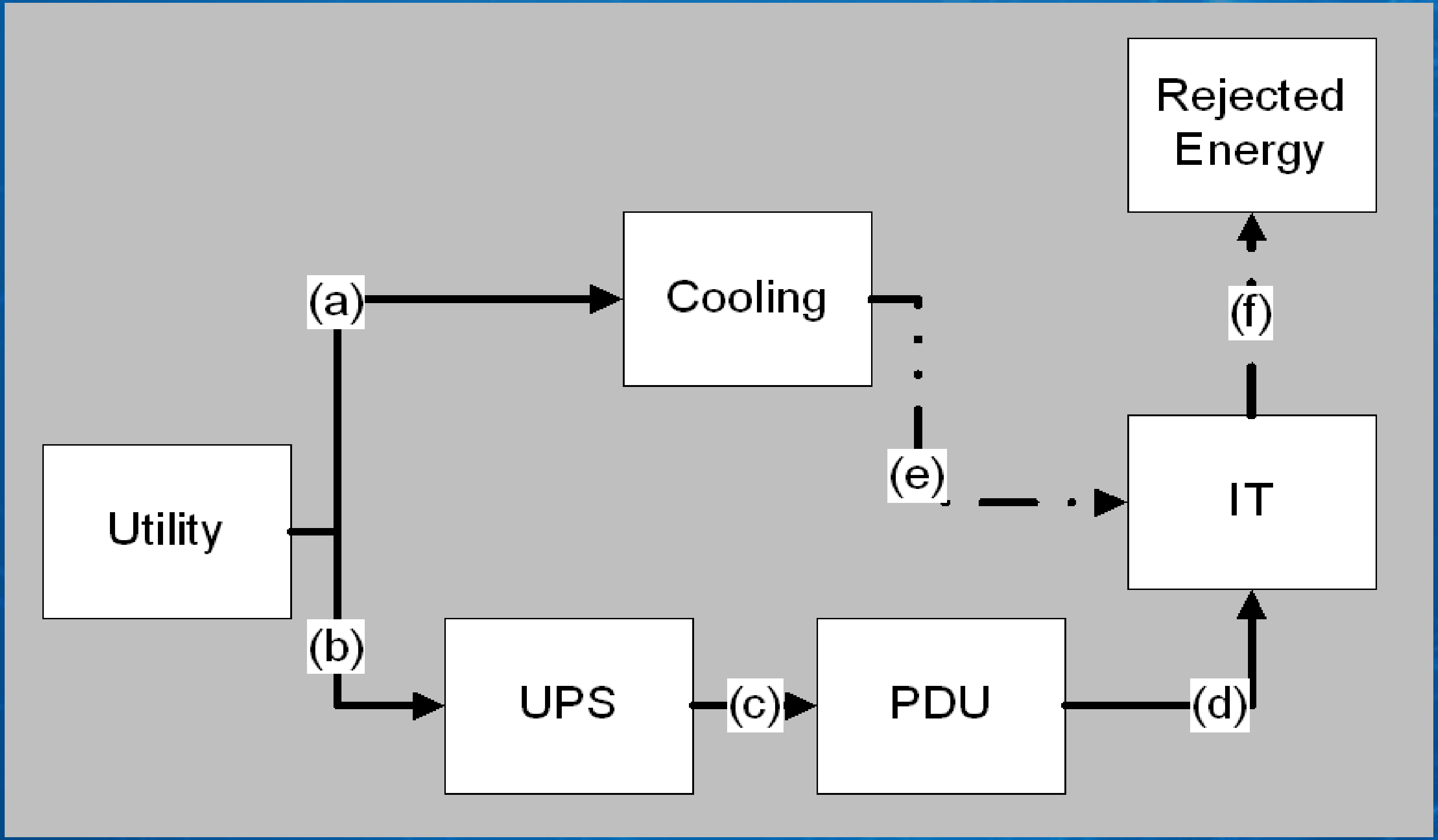
PUE

$$PUE = \frac{Total\ Data\ Center\ Annual\ Energy}{Total\ IT\ Annual\ Energy}$$

- Introduced in 2006 by Malone and Belady
- Developed and agreed to by EU Code of Conduct, DOE, EPA, Green Grid, ASHRAE, etc…
- Has led Energy Efficiency drive in Data Centers
  - PUE Average in 2007 ~ 2.5
  - Best in Class 2016:

NREL= 1.06, LRZ= 1.15, NCAR~1.2,
ORNL= 1.25, TU Dresden < 1.3

PUE – simple and effective



$$PUE = \frac{Total\ Energy}{IT\ Energy} = \frac{Cooling + PowerDistribution + Misc + IT}{IT} = \frac{a+b}{d}$$

PUEs: Reported and Calculated

	PUE
Global bank's best data center (of more than 100)	2.25
EPA Energy Star Average	1.91
Intel average	>1.80
Intel Jones Farm, Hillsboro	1.41
ORNL	1.25
T-Systems & Intel DC2020 Test Lab, Munich	1.24
Google	1.16
Leibniz Supercomputing Centre (LRZ)	1.15
Containers	1.1-1.6
National Center for Atmospheric Research (NCAR)	1.10
Yahoo, Lockport	1.08
Facebook, Prineville	1.07
National Renewable Energy Laboratory (NREL)	1.06

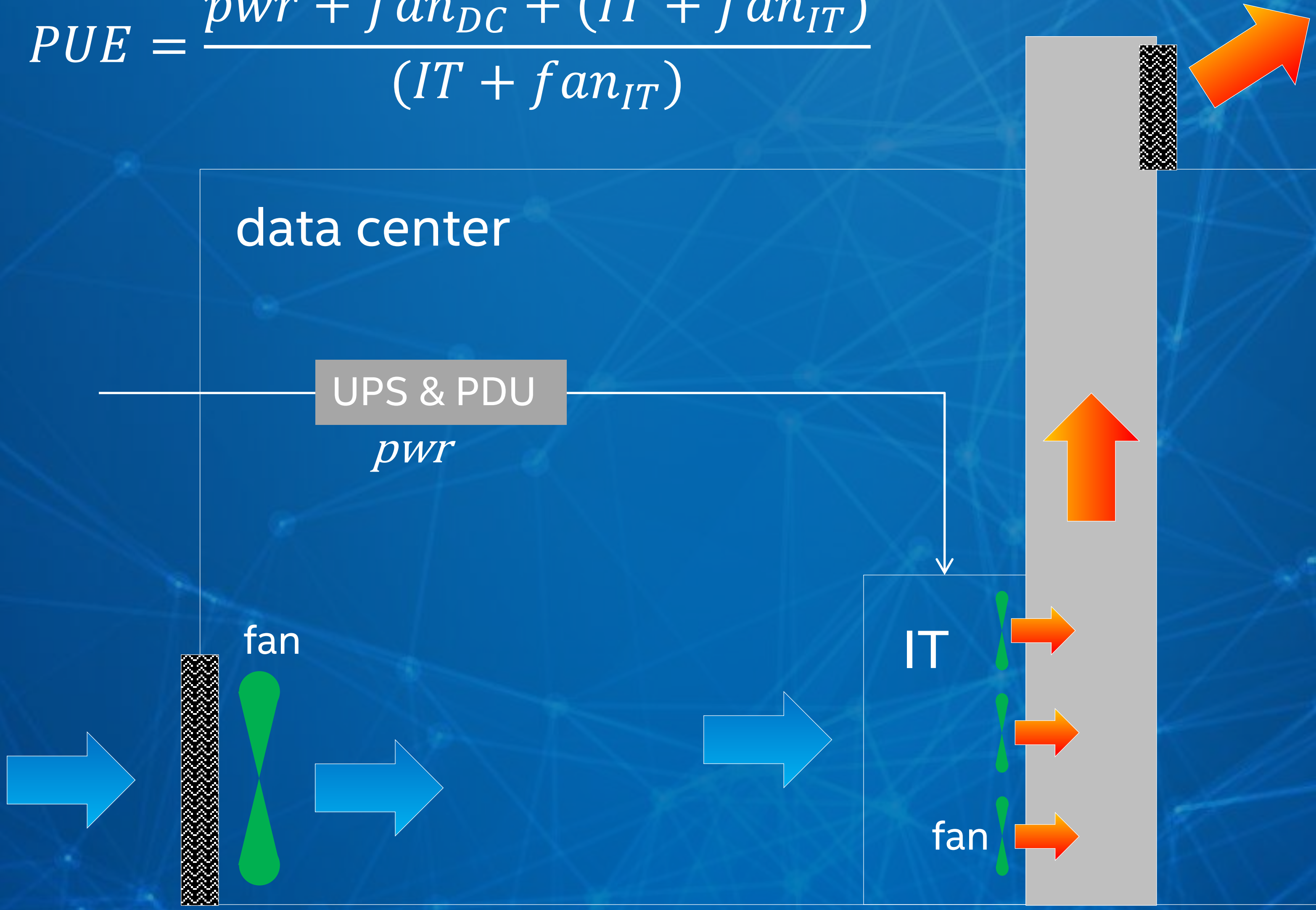
PUEs: Reported and Calculated

Liquid cooling is required for density, but not necessarily for efficiency

Global bank's best data center (of more than 100)	[illegible]	
EPA Energy Star Average	[illegible]	
Intel average	>1.80	
Intel Jones Farm, Hillsboro	1.41	A-FC
ORNL	1.25	LC
T-Systems & Intel DC2020 Test Lab, Munich	1.24	A-FC
Google	1.16	A-FC
Leibniz Supercomputing Centre (LRZ)	1.15	LC
Containers	1.1-1.6	
National Center for Atmospheric Research (NCAR)	1.10	LC
Yahoo, Lockport	1.08	A-FC
Facebook, Prineville	1.07	A-FC
National Renewable Energy Laboratory (NREL)	1.06	LC

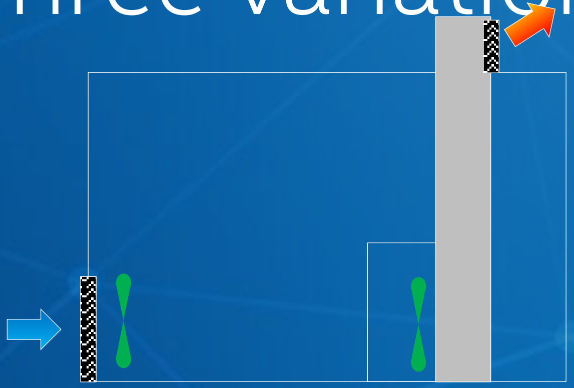
but PUE isn't perfect, consider.....

$$PUE = \frac{pwr + fan_{DC} + (IT + fan_{IT})}{(IT + fan_{IT})}$$



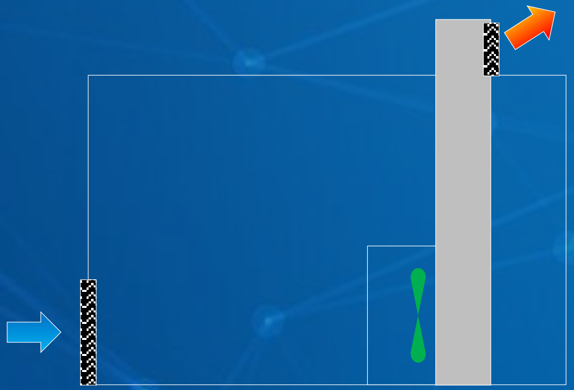
Three variations…

a)
both
fans



$$PUE_a = \frac{pwr + fan_{DC} + (IT + fan_{IT})}{(IT + fan_{IT})}$$

b)
IT
fans
only



$$PUE_b = \frac{pwr + (IT + fan_{IT})}{(IT + fan_{IT})}$$

c)
bldg
fan
only

$$PUE_c = \frac{pwr + fan_{DC} + IT}{IT}$$

$PUE_b < PUE_a < PUE_c$ but is (b) best?
We don't know….

Can we define a “server-PUE”? Maybe ITUE?

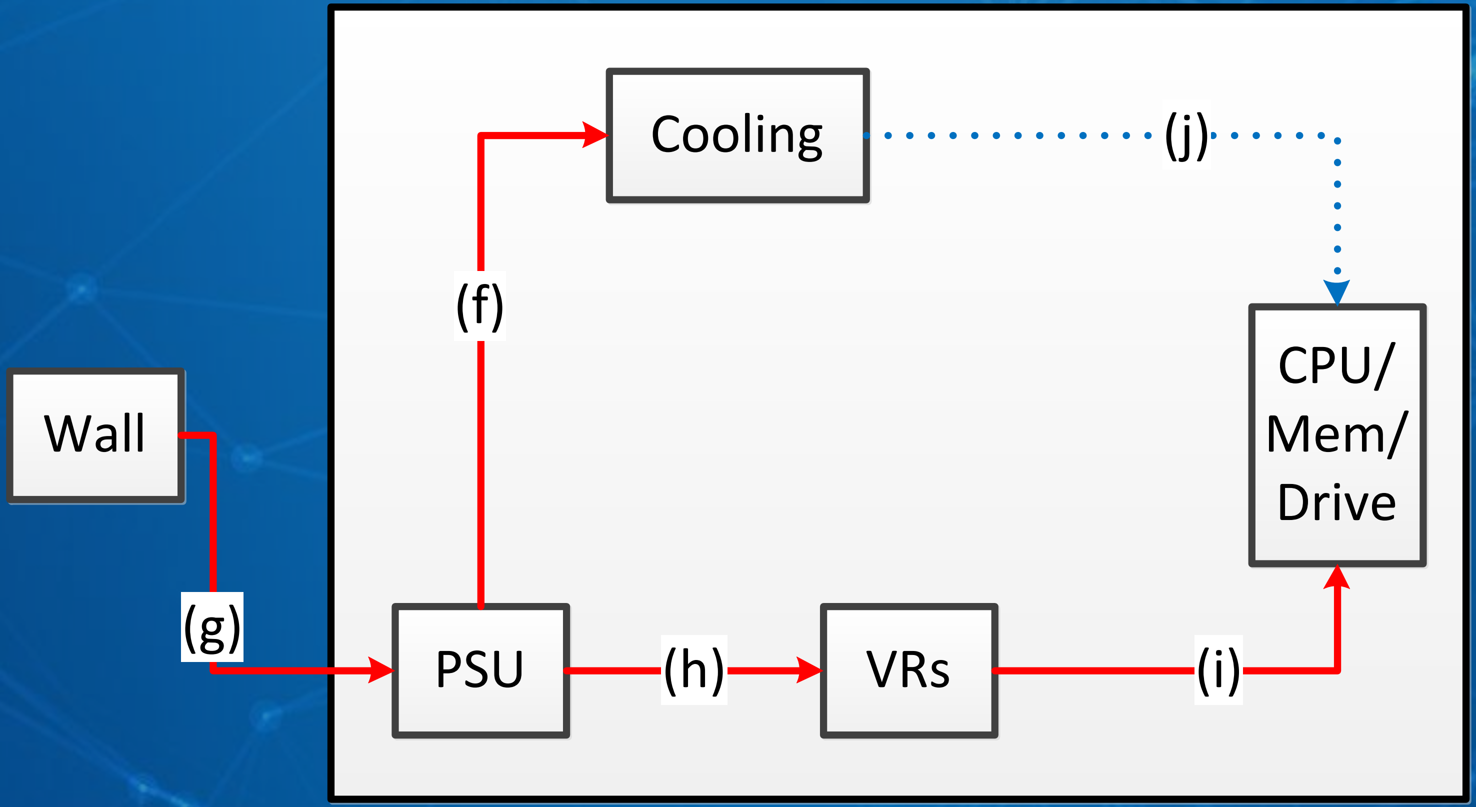
$$PUE = \frac{Total\ Energy}{IT\ Energy} = \frac{Pwr + Cooling + Misc + IT}{IT} = \frac{Infrastructure\ \ Burden + IT}{IT}$$

	Data Center	**Server**
Power dist losses	UPS, line losses, PDUs	PSU, VRs, board losses
Cooling losses	Chiller, CRAC, Pumps, Fans	Fans, Pumps
Misc losses	Security, Lighting, Building Control	Indicators, Platform Control
IT	Servers, Storage, Network	Processor, Memory, Disk

$$ITUE = \frac{Infrastructure\ \ Burden + Compute}{Compute} = \frac{Pwr + Cooling + Misc + Compute}{Compute}$$

$$ITUE = \frac{Total\ Energy\ into\ the\ IT\ Equipment}{Total\ Energy\ into\ the\ Compute\ Components}$$

ITUE



$$ITUE = \frac{total\ energy\ into\ the\ IT\ equipment}{total\ energy\ into\ the\ compute\ components} = \frac{g}{i}$$

The next step…

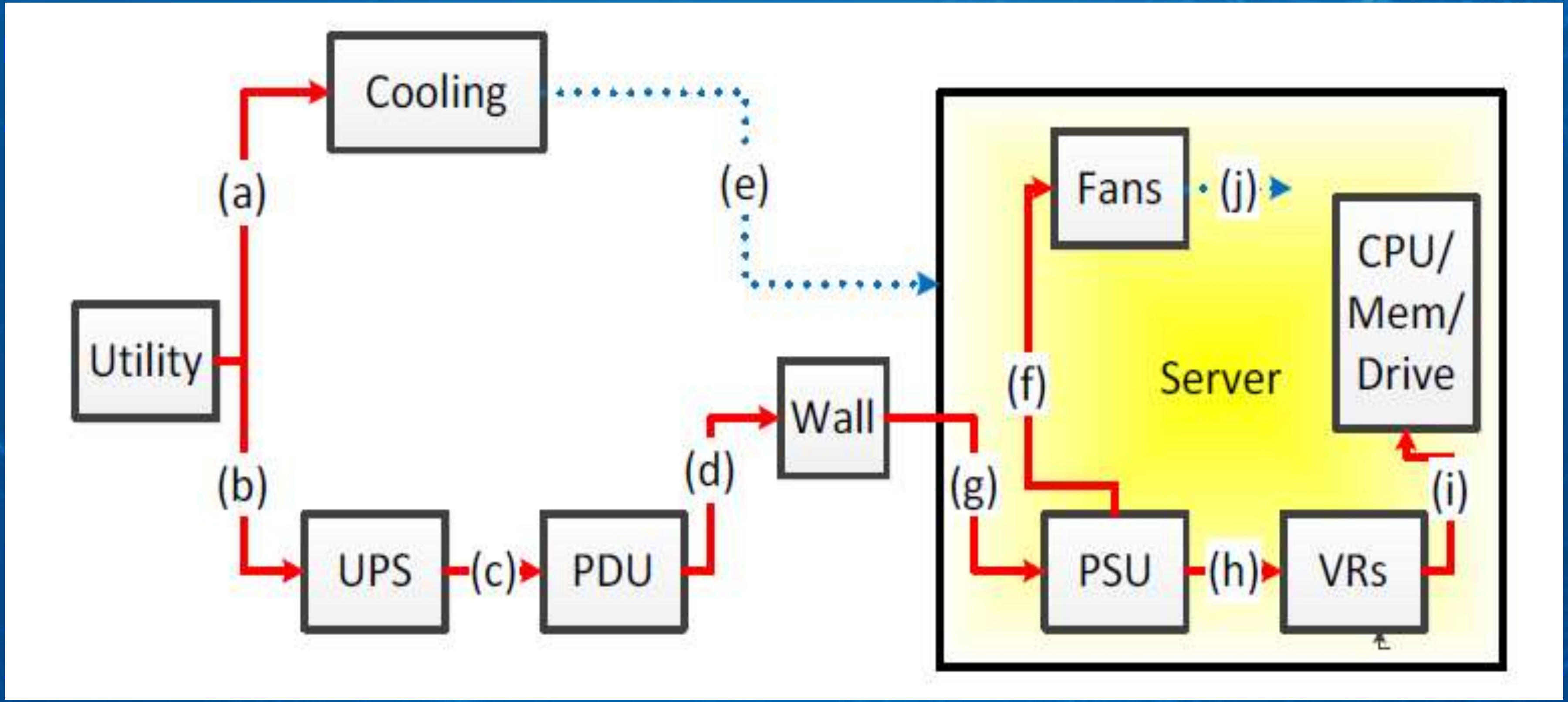
PUE and ITUE are both:

- dimensionless ratios
- Represent the burden or “tax” of infrastructure
- “1” is ideal, values larger than 1 are worse
- Values less than 1 are not allowed

- So why not:

$$TUE = PUE\ x\ ITUE$$

TUE



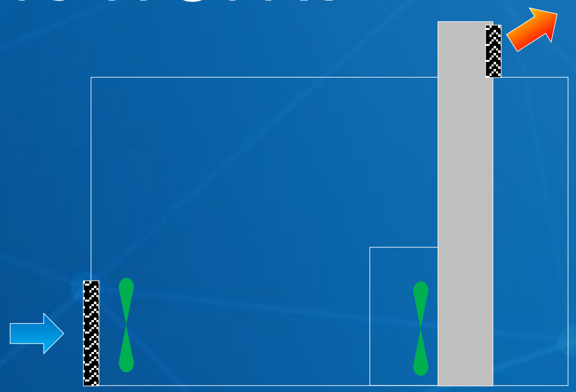
$$PUE = \frac{Total\ Energy}{IT\ Energy} = \frac{a+b}{d}$$

$$ITUE = \frac{Total\ Energy}{Compute\ Energy} = \frac{g}{i}$$

$$\boxed{\boldsymbol{TUE = ITUE \times PUE = \frac{a+b}{i}}}$$

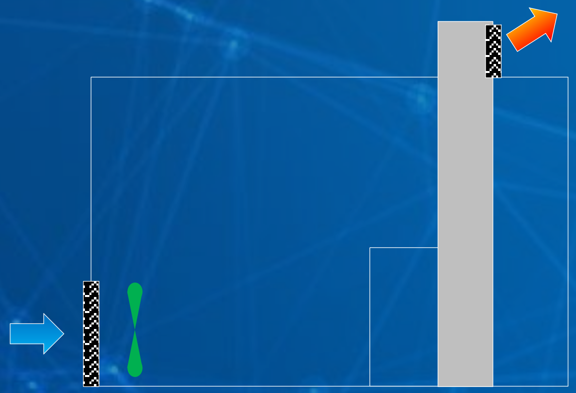
Does it work?

a) both fans



b) IT fans only

c) bldg fan only



$$TUE_a = \frac{pwr + fan_{DC} + fan_{IT} + compute}{compute}$$

$$TUE_b = \frac{pwr + fan_{IT} + compute}{compute}$$

$$TUE_c = \frac{pwr + fan_{DC} + compute}{compute}$$

The lowest TUE yields the lowest energy use. *Yes, it works!*

ITUE / TUE

TUE, a new energy-efficiency metric applied at ORNL's Jaguar

Michael K Patterson, Stephen W Poole, Chung-Hsing Hsu, Don Maxwell, William Tschudi, Henry Coles, David J Martinez, Natalie Bates

1 Introduction

- Paper available
  - email me or from ISC 13
  - Best Paper Award at ISC

- *Use the metric!*
  - Ask for projected ITUE in future procurements
    - Good cluster to cluster efficiency comparison
  - Begin to develop monitoring strategy

- Be aware of limits
  - Does not include workload / output
  - Difficult to use on older machines
  - Don't ask for everything; likely to expensive

Power

Trends in the very high end....

- 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

- More typical HPC
  - 400 Vac 3ph, 230Vac 1ph
  - 48 Vdc in the rack can reduce weight, cost, and size

- HVDC (380 Vdc) is an option; primary reasons why are 1st cost and renewables

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

Power

Europe primary power

- 400 Vac 3ph
- High density racks could use 3ph, 1ph if not high power; better options for PSUs (PUE/ITUE)
- Likely that most storage and system racks would do well on 230 Vac 1ph
- Consider rating schemes for equipment (PSU): Platinum, Gold, Silver, etc… (ITUE)
- Board power: same direction, higher cost components very often have a good ROI (ITUE)

UPS (Uninterruptable power supply)

- Generally HPC would rather spend money on compute than UPS, generally European power quality is good enough without
- Please don’t use UPS for power quality reasons. ☹ Also they waste 2-5%. (PUE)
- Do use UPS and redundant feeds for fabric and storage and service nodes

Power Management

- Tools available, must be built for the site needs

Packaging

Rack and cluster weight and density

- Strong correlation between weight and power
  - Some studies have shown kg/kW is ~constant across rack size
  - Goal is to reduce this ratio
- Packaging
  - High density computing – network topology optimization and high node count per rack (lots of power) make for dense cabinets
- Rack weight density
  - Design limit: Floor tiles at 500 lbs/sf ~ 2500 kg/m2 for high end. ☹ …more than many DCs
- 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?

Video Credit to Helmut Satzger, LRZ, Munich, thanks for sharing!


6.4 PFLOP/s
00:38.87

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 fits 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 & infrastructure space continues to grow

Rack density (kW/rack & kg/m2) have a wide range of choices, but the local data center may restrict these

Cooling

Why liquid? Why Air?

- Power per node continues to rise
- Rack density limits airflow path
- But air-cooling can cost less
- 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

Cooling

Air Cooling

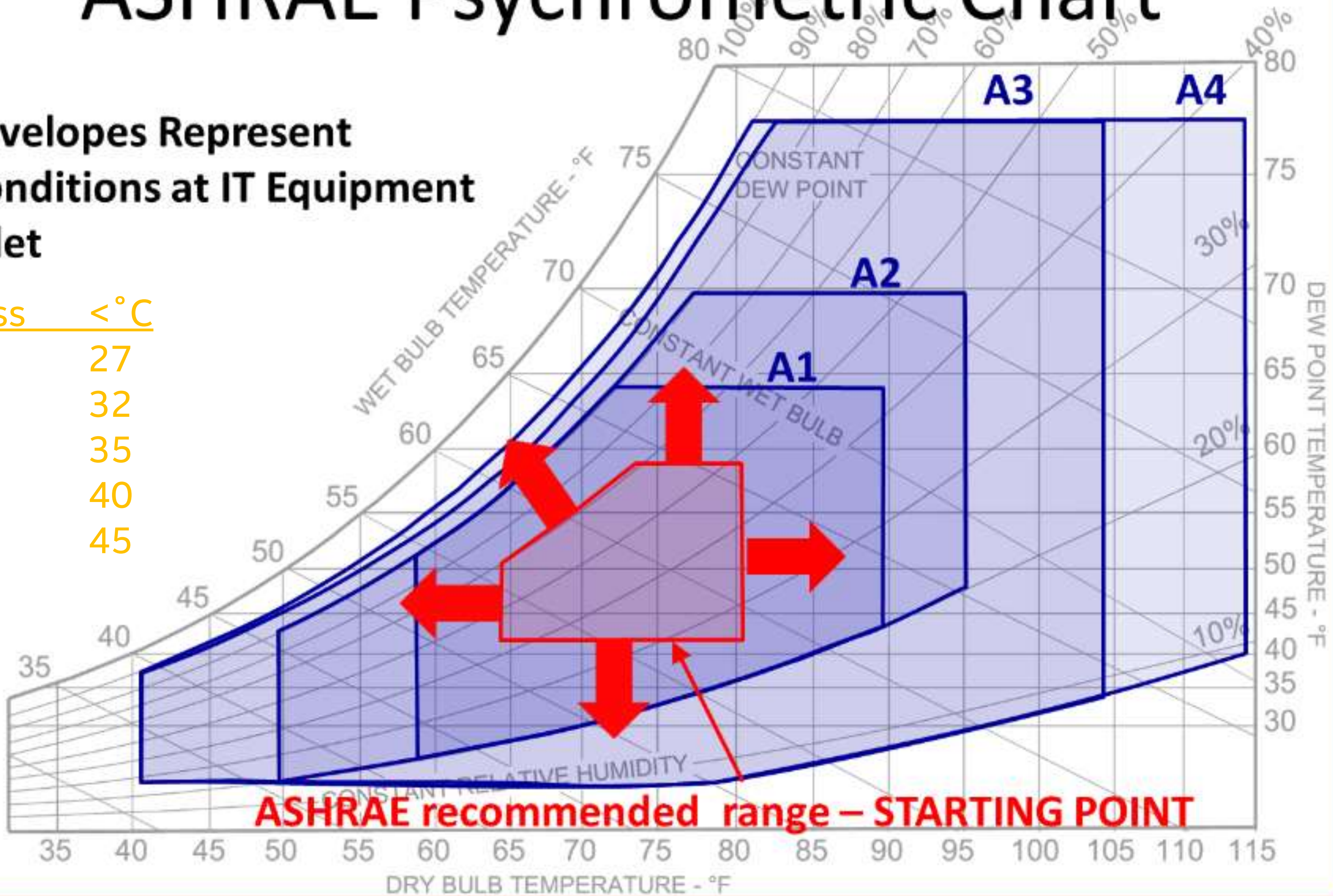
- ASHRAE A1 thru A4
- Containment a “must” for good PUE and low TCO
- Hot Aisle and Cold Aisle an operational choice, not an efficiency choice
- Free –air cooling should always be checked for applicability
  - Corrosion a real issue depending on air-quality
- Air-Cooling limits rack density, but good performance density can still be had

- If you do air-cooling in a new data center; the VERY FIRST MOST IMPORTANT consideration is to design the building around the airflow path. Get that done, then bring in the building architects. ☺

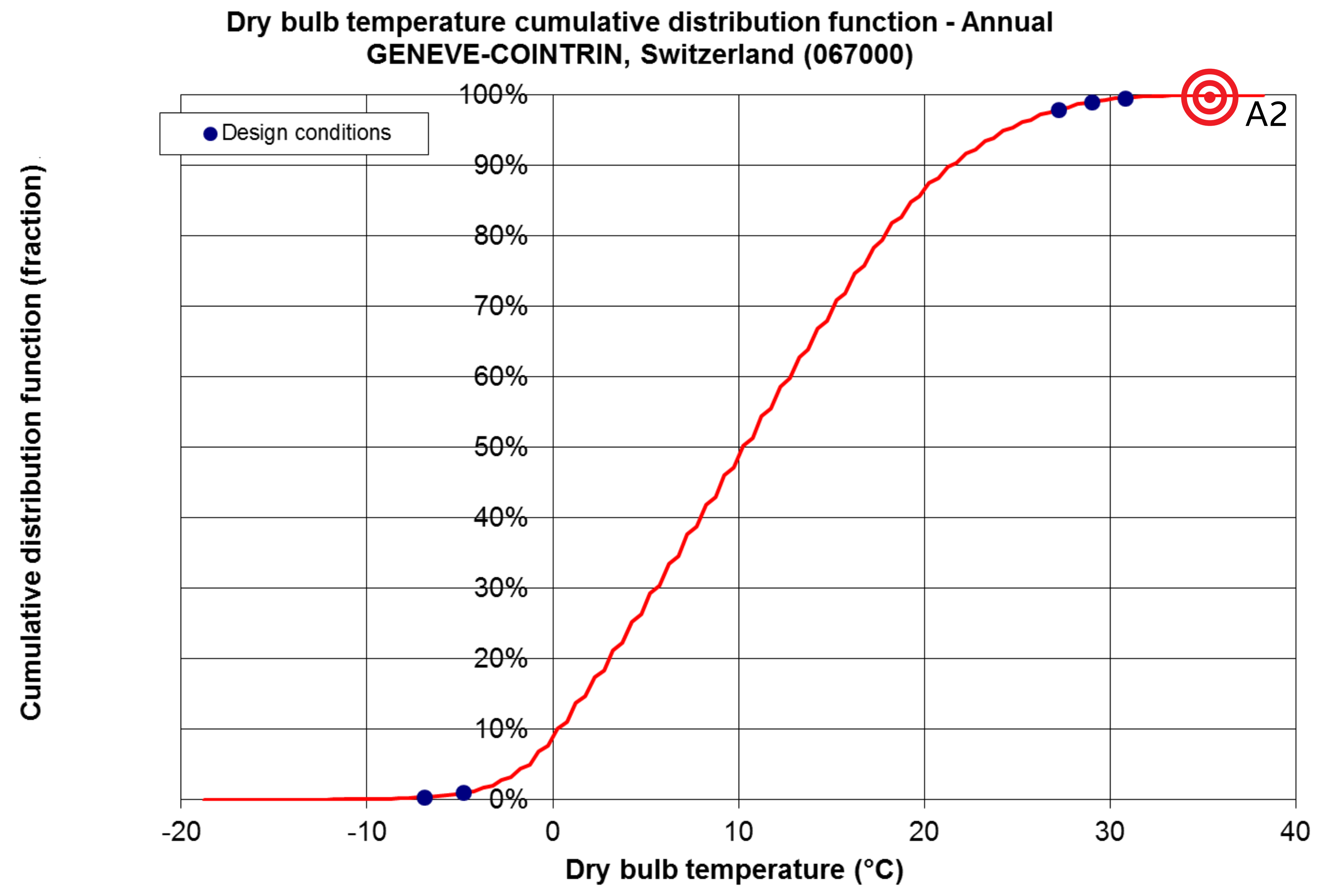
ASHRAE Psychrometric Chart

***Envelopes Represent Conditions at IT Equipment Inlet**

Class	<°C
R	27
A1	32
A2	35
A3	40
A4	45

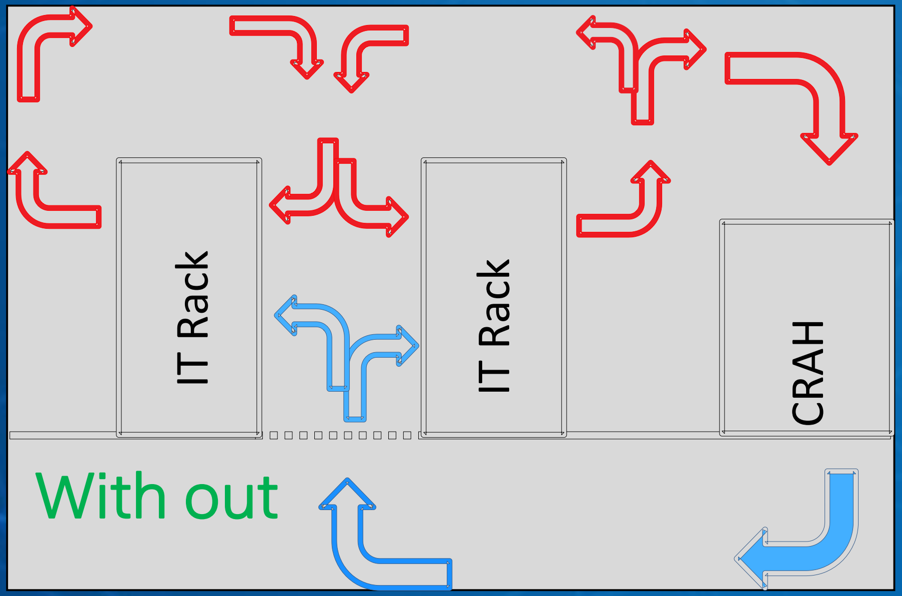


ASHRAE recommended range – STARTING POINT



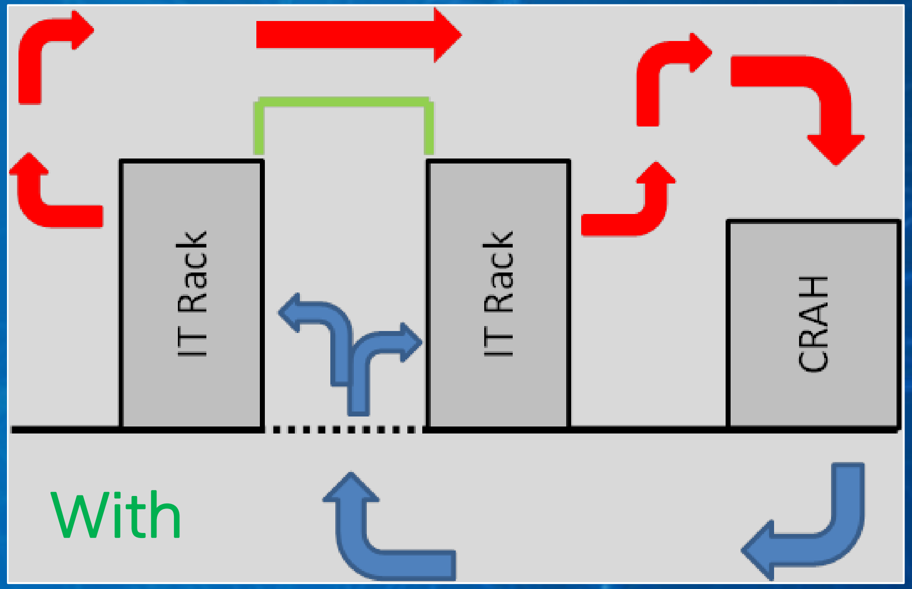
Dry bulb temperature cumulative distribution function - Annual
GENEVE-COINTRIN, Switzerland (067000)

Why containment?



Recirculation
Excess Air
Poor IT temperature

No mixing
Less airflow
Higher ∆T



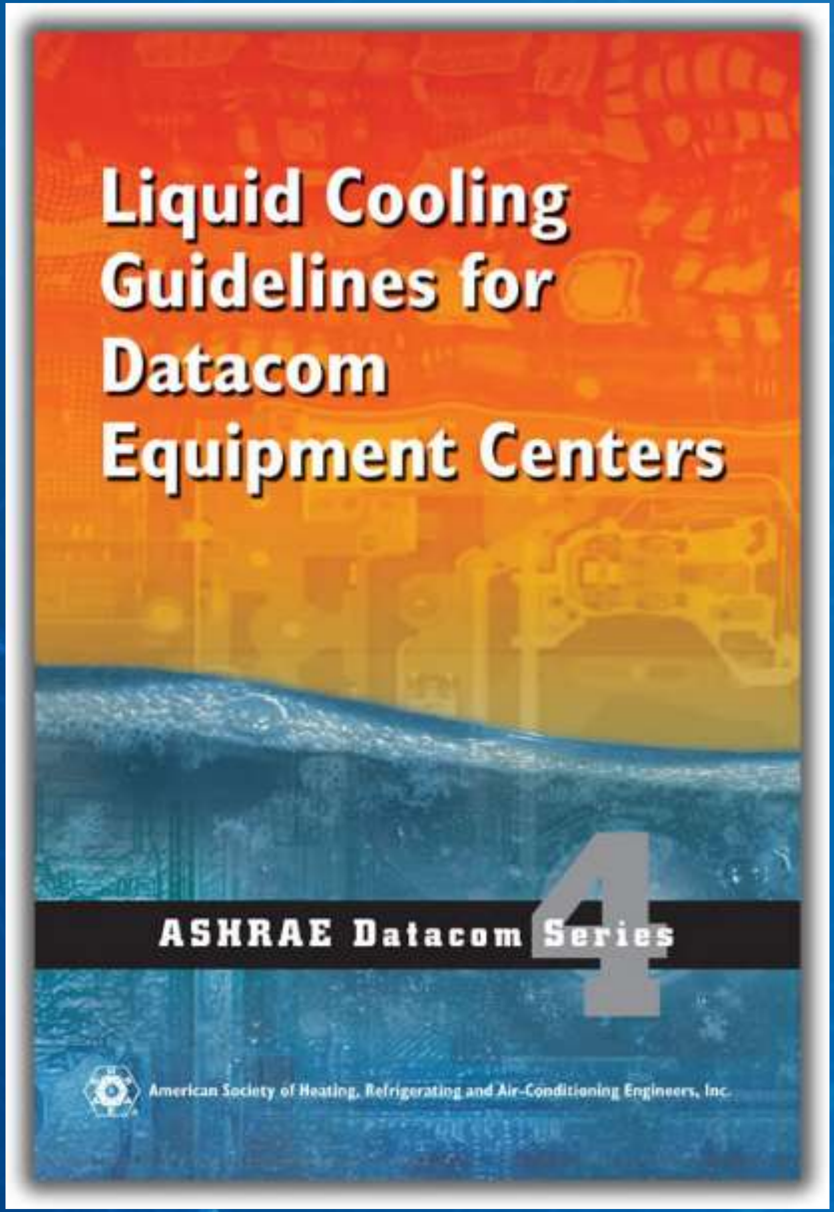
Airflow management is the number one tool in your toolbox for data center improvement – it can solve more problems than any other tool!

CERN cold aisle containment

Cooling

Liquid Cooling

- ASHRAE W1 thru W4
- Many different varieties, they have VERY different performance, cost, and efficiency results
- Water quality is an important issue
  - Consult ASHRAE guide; monitor!
- Immersion cooling is not on our roadmap, we still keep current but issues exist
  - Oil immersion
  - Two-phase immersion
- Liquid cooling can offer better performance, better reliability
  - High density systems (Aurora, etc, are fully liquid cooled)



Liquid Cooling Guidelines for Datacom Equipment Centers

4

ASHRAE Datacom Series

American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc.

All ASHRAE work has been incorporated into the 2nd Edition.

Tip: 2nd Edition now available for purchase in the ASHRAE bookstore.

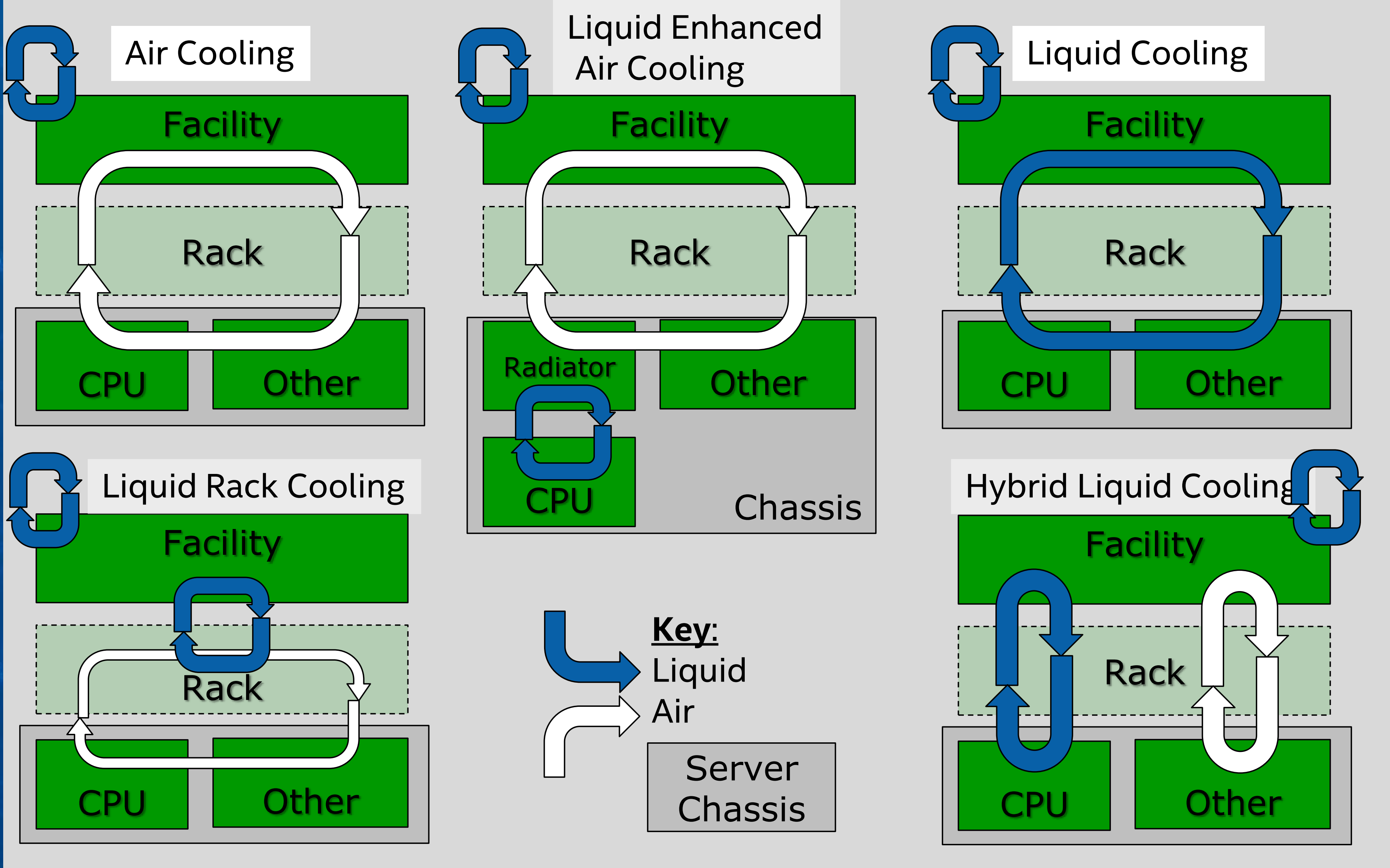
More important tip:
Chap 5 covers Facility water (FWS)
Chap 6 covers IT loop water (TCS)
These are very different! Specify the right water.

2011 ASHRAE Liquid-Cooled Thermal Guidelines

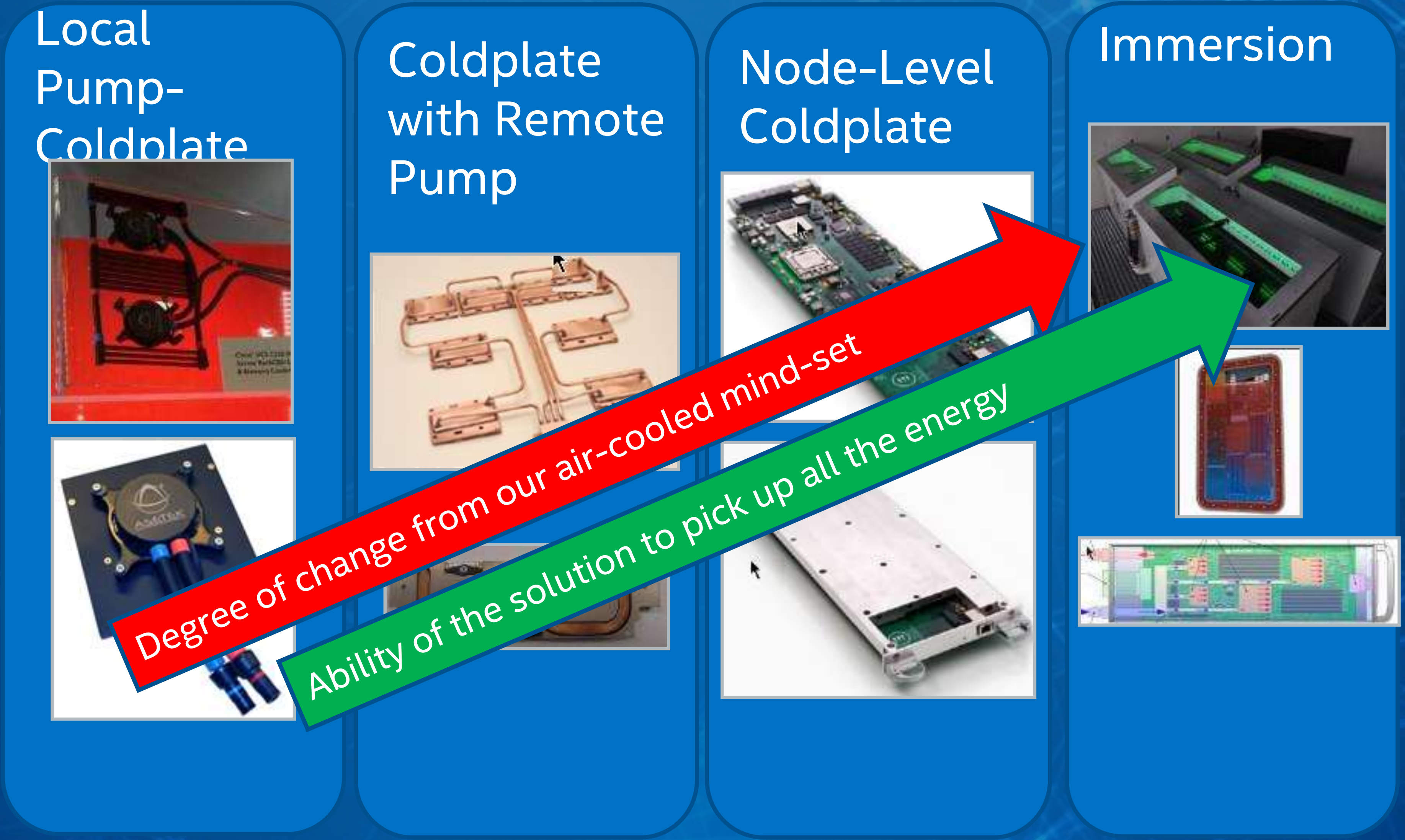
Classes	Typical Infrastructure Design		Facility Supply Water Temp (C)	IT Equipment Availability
	Main Cooling Equipment	Supplemental Cooling Equipment		
W1	Chiller/Cooling Tower	Water-side Economizer Chiller	2 – 17	Now available
W2			2 – 27	
W3	Cooling Tower	Chiller	2 – 32	Becoming available, dependent on future demand
W4	Water-side Economizer (with drycooler or cooling tower)	Nothing	2 – 45	
W5	Building Heating System	Cooling Tower	> 45	Not for HPC

Required Cooling Infrastructure: Balance of Silicon/Datacenter

System Definitions – all different, all about how close the liquid gets to the components



Liquid Cooling Technologies



Local Pump-Coldplate

Coldplate with Remote Pump

Node-Level Coldplate

Immersion

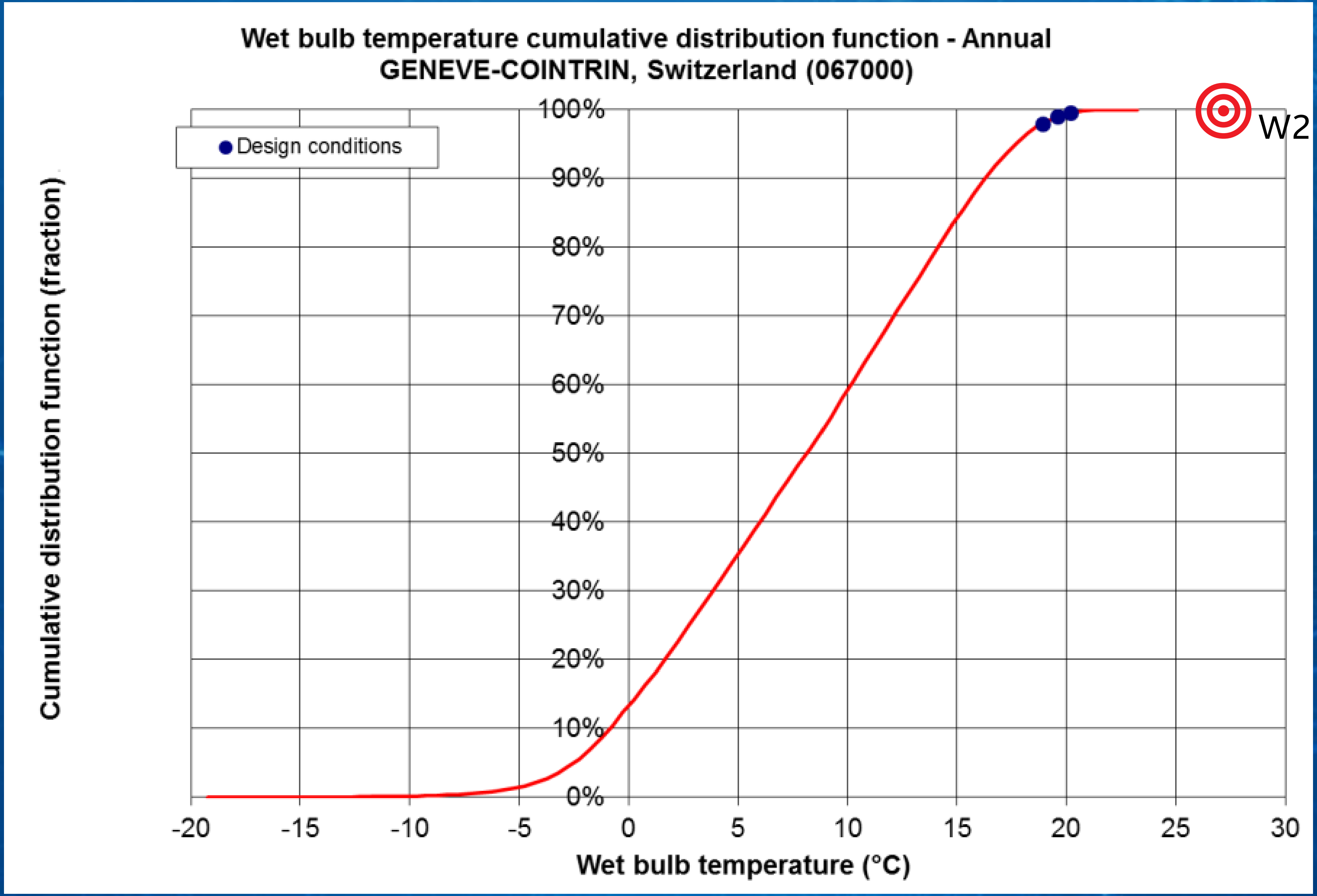
Degree of change from our air-cooled mind-set

Ability of the solution to pick up all the energy

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



Wet bulb temperature cumulative distribution function - Annual
GENEVE-COINTRIN, Switzerland (067000)

W2

ASHRAE

TC 9.9 Committee
http://tc99.ashraetcs.org/

Books
https://www.ashrae.org/resources--publications/bookstore/datacom-series

EE HPC WG

http://eehpcwg.llnl.gov/

Hot for Warm Water Cooling

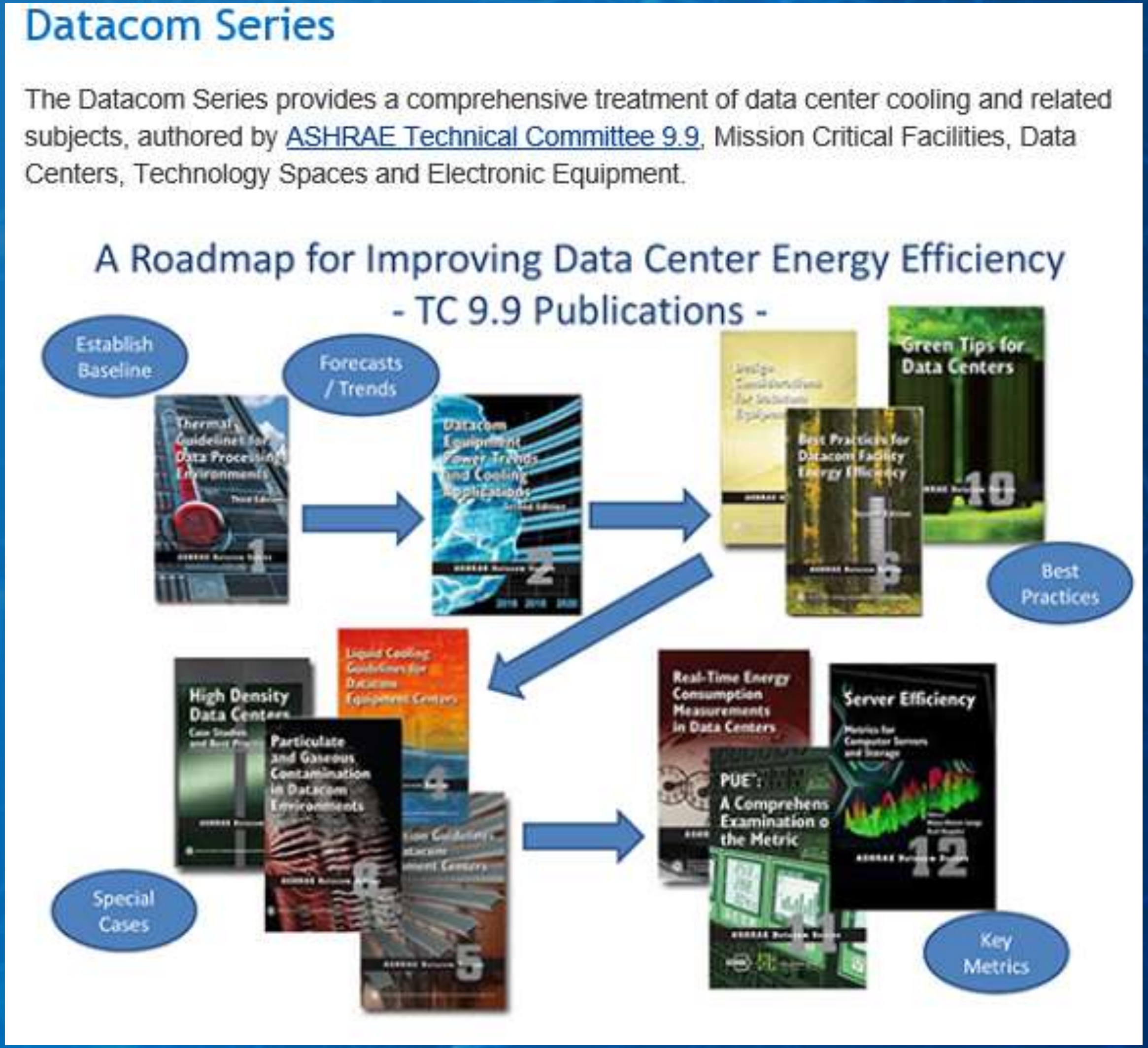
http://eetd.lbl.gov/sites/all/files/publications/lbnl-5128e.pdf

The Green Grid

http://www.thegreengrid.org/

http://www.thegreengrid.org/en/Global/Content/Tools/NAmericanFreeCoolingTool

EU Code of Conduct for Data Centres
http://iet.jrc.ec.europa.eu/energyefficiency/ict-codes-conduct/data-centres-energy-efficiency



Datacom Series

The Datacom Series provides a comprehensive treatment of data center cooling and related subjects, authored by ASHRAE Technical Committee 9.9, Mission Critical Facilities, Data Centers, Technology Spaces and Electronic Equipment.

A Roadmap for Improving Data Center Energy Efficiency

- TC 9.9 Publications -

Summary

- Data Center Design is straightforward, but can be daunting if not fully understood, unfortunately still very site-dependent
- Resources are available!
- Modular build out is best; plan for the end state, provision just for today
- PPC for HPC
  - Power delivery at higher voltages with less redundancy than Enterprise
  - Density has value, but Packaging can challenge most data centers
  - Air and Liquid Cooling have their place, choose the right one for performance and value

Thanks for your attention

Questions?

michael.k.patterson@intel.com