Preparing for Exascale; Energy Efficiency from the Silicon to the Data Center

Dr. Michael Patterson
Principal Engineer
Exascale System Architecture and Pathfinding

Legal Information

Intel in HPC: In 2009

Processors

Software

Intel® Enterprise Edition for Lustre* software

Intel in HPC: In 2013+

Processors

Fabrics

Intel® Truescale

Intel® Ethernet Products

Next Generation Interconnects

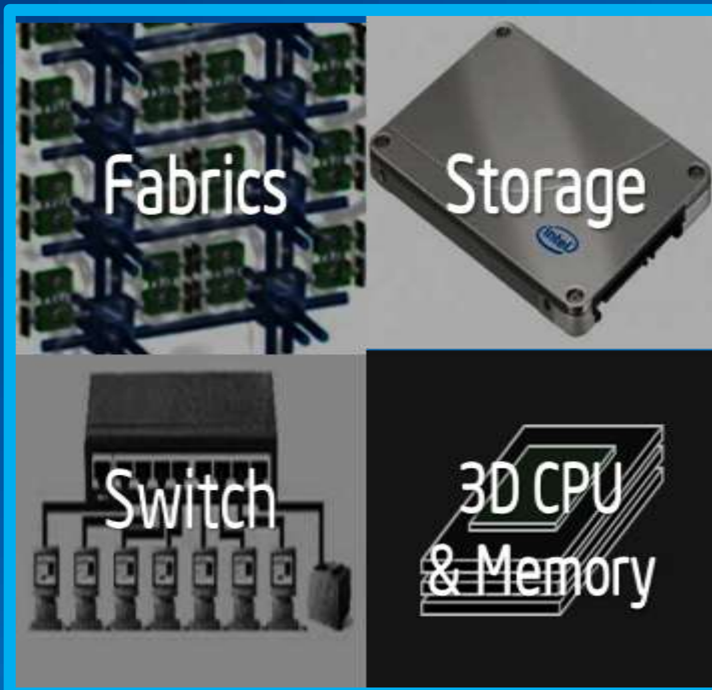
Coprocessor

Software

Intel® Enterprise Edition for Lustre* software

Integration Is The Key

Unprecedented Innovations Only Enabled by the Leading Edge Process Technology

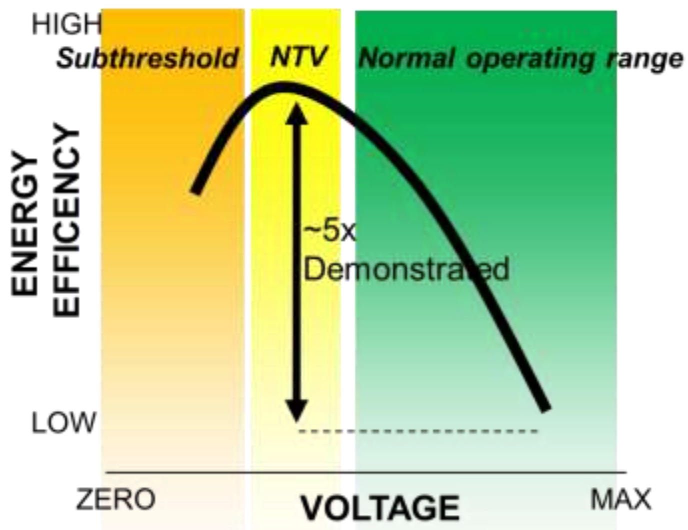


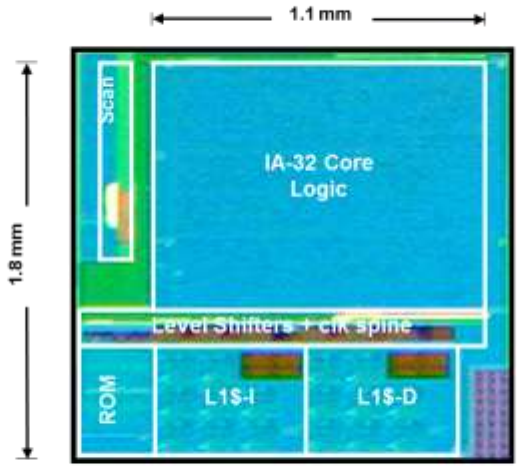
Integrated Today

The Possibilities For Tomorrow

Addressing the Power Challenge

Near Threshold Voltage Operation Demonstrated





Ultra-low Power	Energy Efficient	High Performance
280 mV	0.45 V	1.2 V
3 MHz	60 MHz	915 MHz
2 mW	10 mW	737 mW
1500 Mips/W	5830 Mips/W	1240 Mips/W

Technology	32nm High-K Metal Gate
Interconnect	1 Poly, 9 Metal (Cu)
Transistors	Core:6M
Core Area	$2mm^2$
Package	951 Pins FCBGA11

Wide Dynamic Range

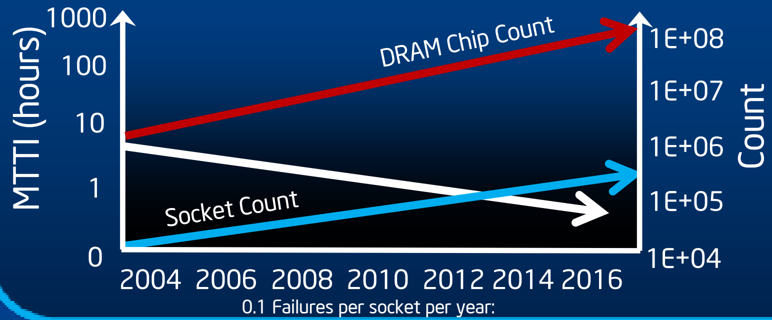
S. Jain, S. Khare et. al. "A 280mV-to-1.2V Wide-Operating-Range IA-32 Processor in 32nm CMOS," ISSCC 2012

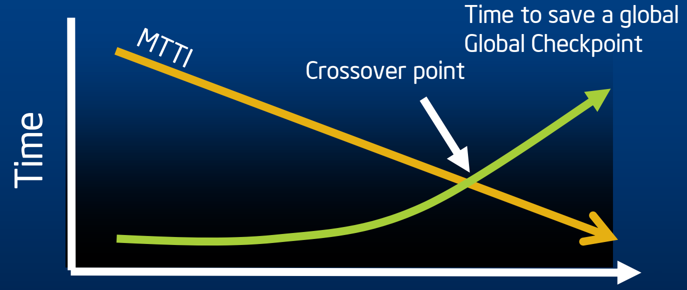
The voltage knob is the biggest knob we have, but it needs to be used intelligently

Source: Exascale Computing Study: Technology Challenges in achieving Exascale Systems (2008)

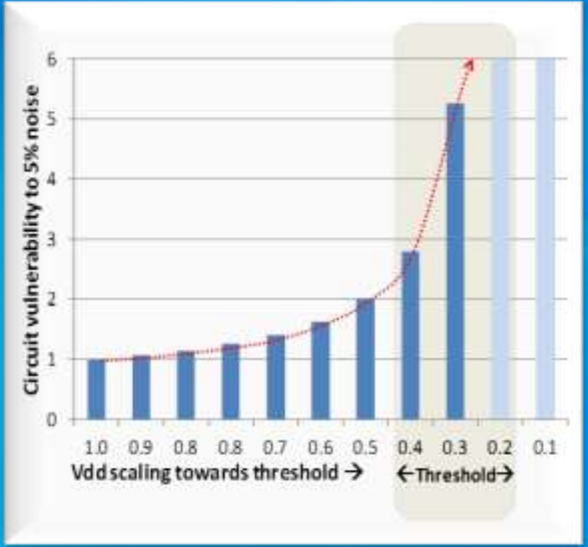
Transparent Hardware and System Software Recovery in the Face of Failures is Essential

MTTI Measured in Minutes

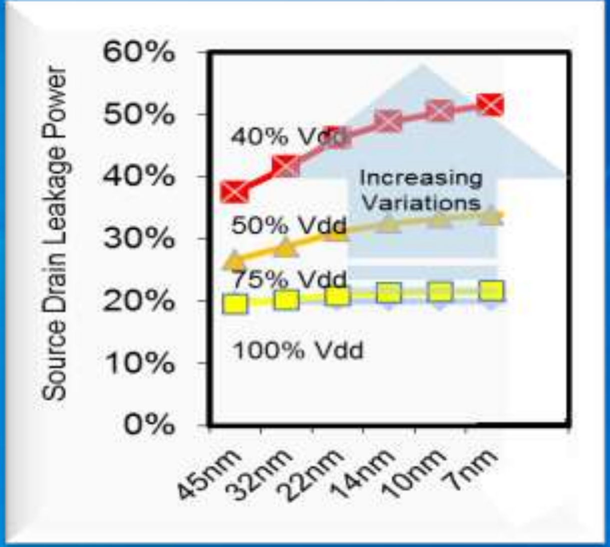




Vulnerability to Noise



Transistor Variations



Fabric Innovation Has to Accelerate to Balance Demands for Bandwidth, Latency, Resiliency, and Scalability

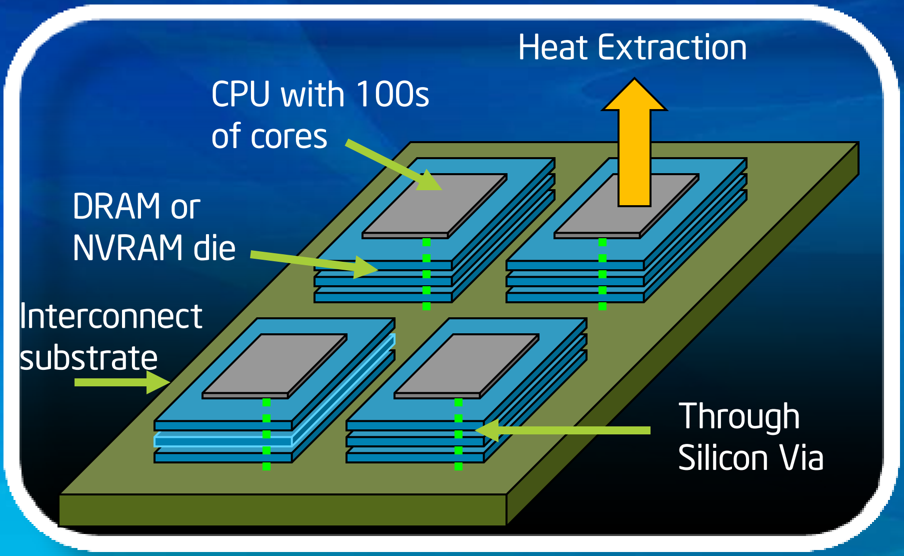
Challenge	HPC Clusters
Bandwidth Today: *10GB-20GB*	100s of GBs per second
Latency Today: *1000s of ns*	10s of nanoseconds across the fabric
Energy Efficiency Today: *12-25 pJ/bit*	3-5 pico Joules per bit per Link
Application	New Workloads in HPC, Big Data, Analytics, & Search
Scalability Today: *1000's of nodes*	10K-100Ks nodes in a Datacenter
Mgmt, Security, QoS	Cluster Fabric Mgmt

Intel's
Comprehensive Connectivity and Fabric Portfolio

- Fabrics are Becoming the Next Bottleneck to Unrelenting Need for Performance & HPC Workloads and Data in Cloud
- **Our Goal**: Innovate at the System, Node, and Fabric

intel

3D Integration of Compute, IO, And Memory Is the Only Solution For Energy Efficient BW



- Thin Logic and DRAM die
- Through silicon vias
- Energy efficient, high speed IO to buffer
- Detailed interface signals created on the logic die

Intel Exascale Labs — Europe

Strong Commitment To Advance Computing Leading Edge:

Intel collaborating with HPC community & European researchers
4 labs in Europe - Exascale computing is the central topic

ExaScale Computing Research Lab, Paris

Performance and scalability of Exascale applications

Tools for performance characterization

ExaCluster Lab, Jülich

Exascale cluster scalability and reliability

ExaScience Lab, Leuven

Comms avoiding algorithms

Architectural simulation

Scalable kernels and RT

Intel and BSC Exascale Lab, Barcelona

Scalable RTS and tools

New algorithms

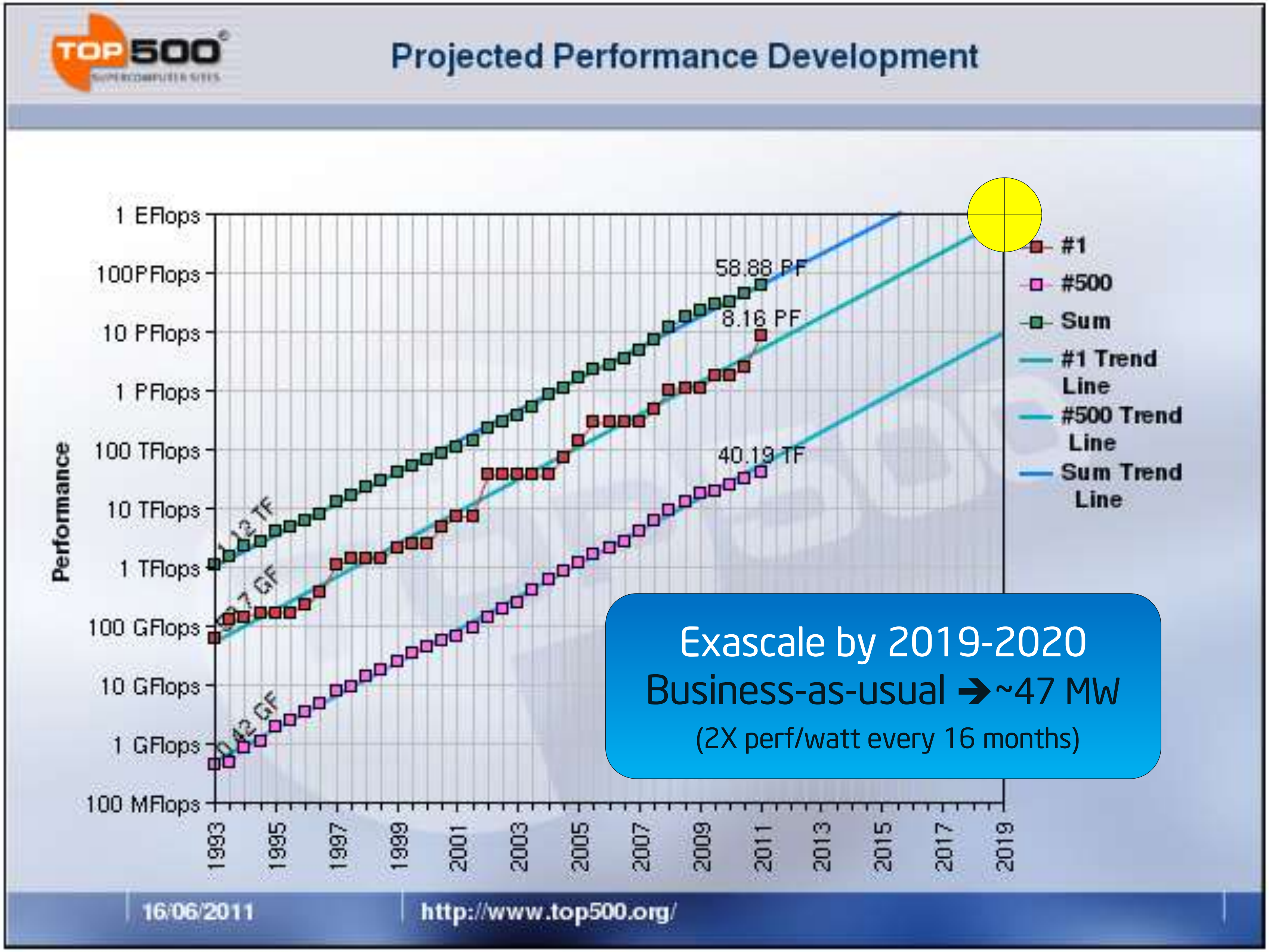
www.exascale-labs.eu

Signed Collaboration agreement

Daresbury Science & Innovation Campus

…but back to the topic at hand

- 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 2019-2020

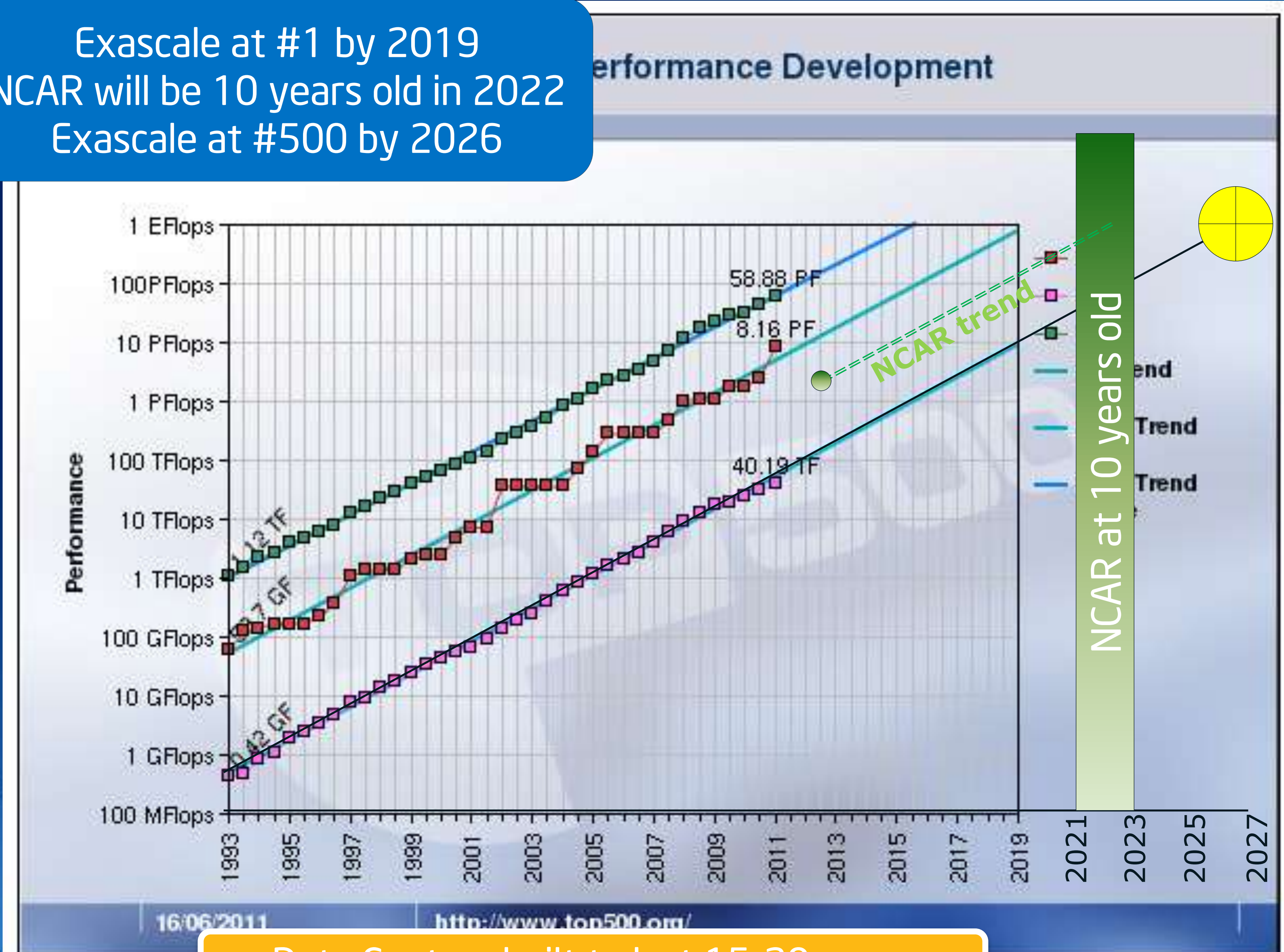
Business-as-usual → ~47 MW

(2X perf/watt every 16 months)

NCAR - Home to an Exaflop SuperComputer

NCAR Yellowstone
New supercomputing center in Wyoming

Exascale at #1 by 2019
NCAR will be 10 years old in 2022
Exascale at #500 by 2026



Data Centers built to last 15-20 years

The four essential elements

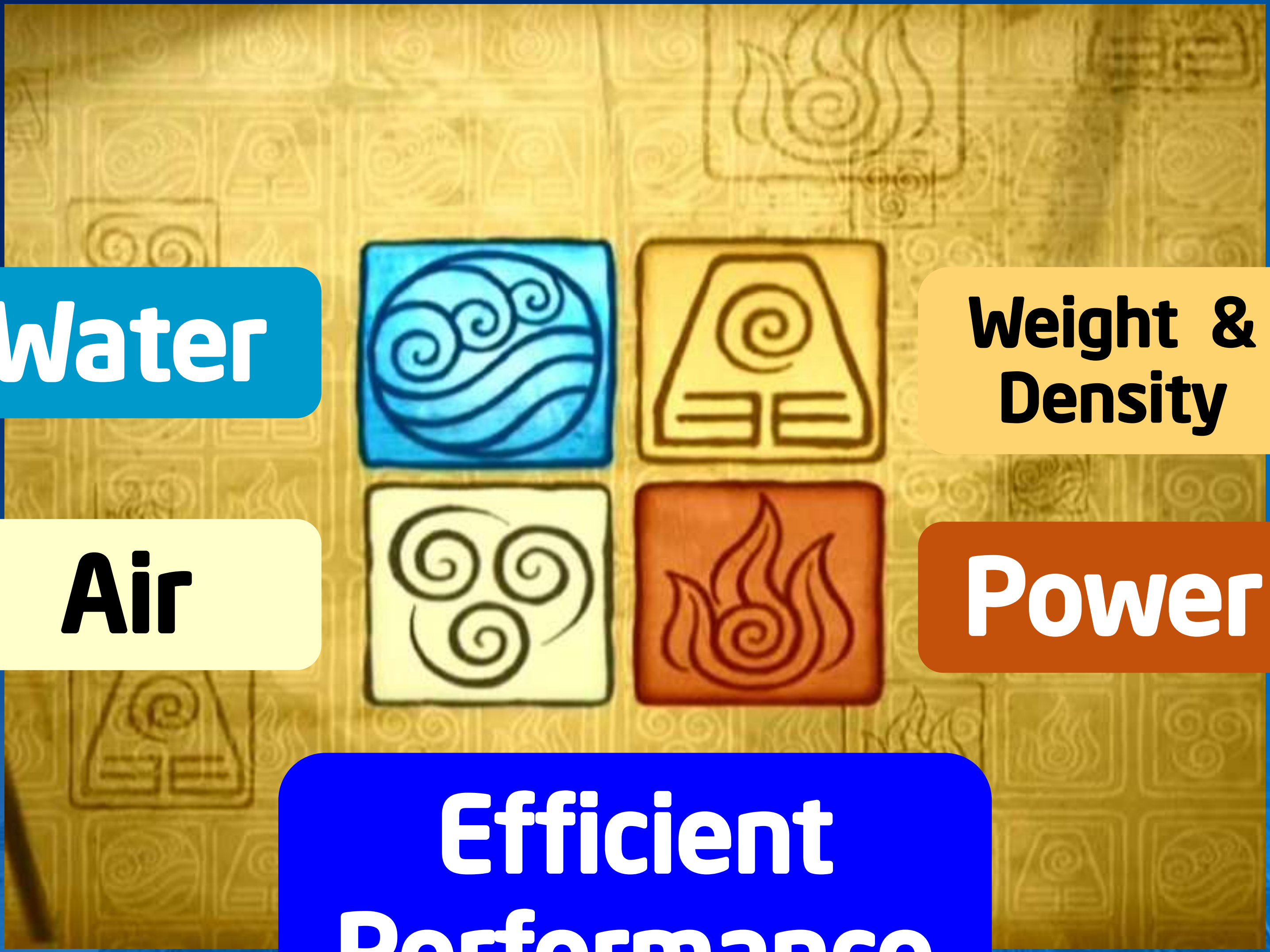
Water

Earth

Air

Fire

The four essential HPC elements



Water

Weight & Density

Air

Power

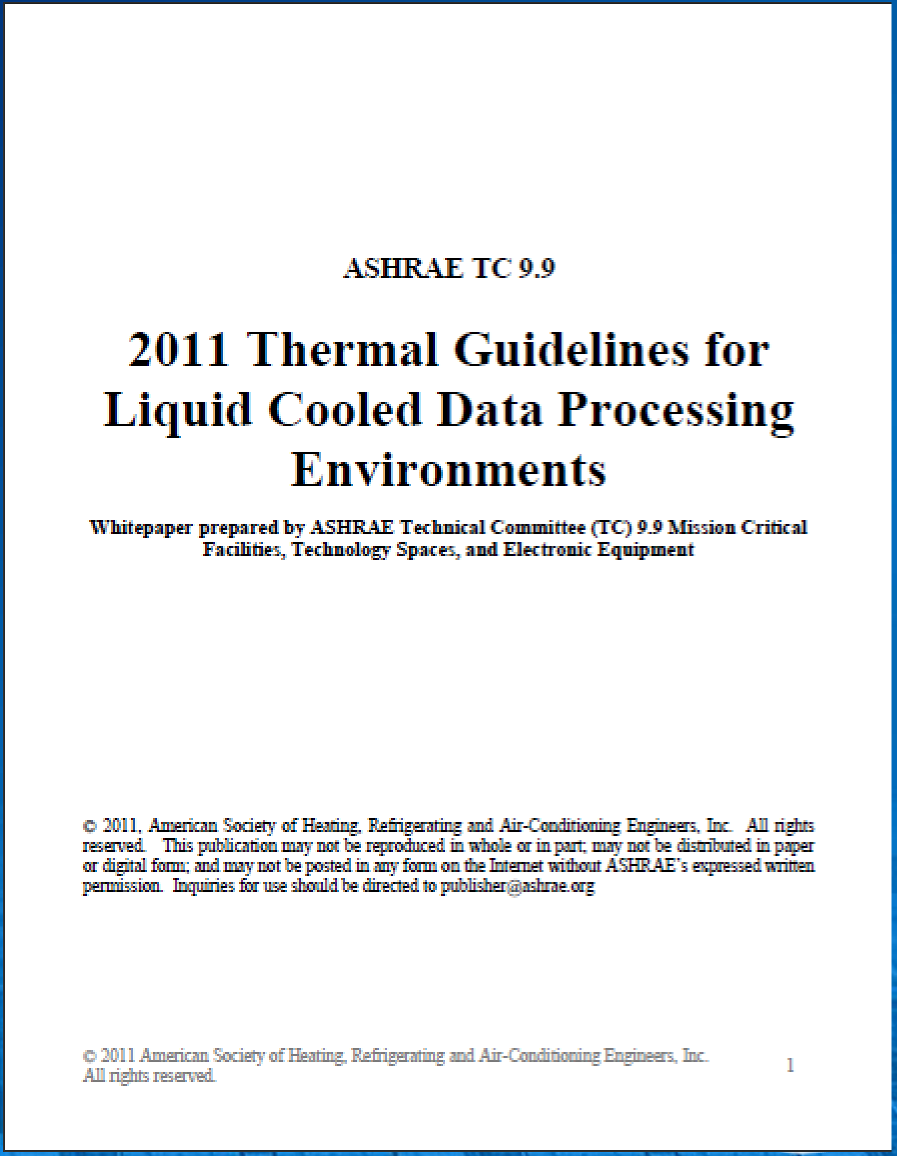
Efficient Performance

The four essential HPC elements

Water

ASHRAE Liquid Cooling Guidelines

- ASHRAE team worked to provide better guidance for liquid cooled systems
- Bull, Cray, Dell, HP, IBM, Intel, SGI, and others all participated
- Download at:
- http://tc99.ashraetcs.org or:
- email – michael.k.patterson@intel.com


ASHRAE TC 9.9

2011 Thermal Guidelines for Liquid Cooled Data Processing Environments

Whitepaper prepared by ASHRAE Technical Committee (TC) 9.9 Mission Critical Facilities, Technology Spaces, and Electronic Equipment

© 2011, American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc. All rights reserved. This publication may not be reproduced in whole or in part; may not be distributed in paper or digital form; and may not be posted in any form on the Internet without ASHRAE's expressed written permission. Inquiries for use should be directed to publisher@ashrae.org

© 2011 American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc. All rights reserved. 1

New Classes!

Table 1: 2011 ASHRAE Liquid Cooled Guidelines((I-P version in Appendix A)

	Typical Infrastructure Design		
Liquid Cooling Classes	**Main Cooling Equipment**	**Supplemental Cooling Equipment**	**Facility Supply Water Temp(C)**
W1(see Figure 3a)	Chiller/Cooling Tower	Water-side Economizer (w drycooler or cooling tower)	2 - 17
W2(see Figure 3a)			2 - 27
W3(see Figure 3a)	Cooling Tower	Chiller	2 - 32
W4(see Figure 3b)	Water-side Economizer (w drycooler or cooling tower)	N/A	2 - 45
W5(see Figure 3c) See Operational Characteristics	Building Heating System	Cooling Tower	>45

Water Quality Problems

- Corrosion – chemical attack on materials of construction (e.g. chloride corrosion in stainless steel)
- Scaling – chemical formation of deposits in cooling systems (e.g. hardness scaling)
- Fouling – particulate or physical blocking of channels or coating of surfaces (e.g. construction debris or dirt/dust blocking μchannels)
- Microbial – Biological activity in water systems (e.g. can lead to fouling or corrosion)

Things to know….

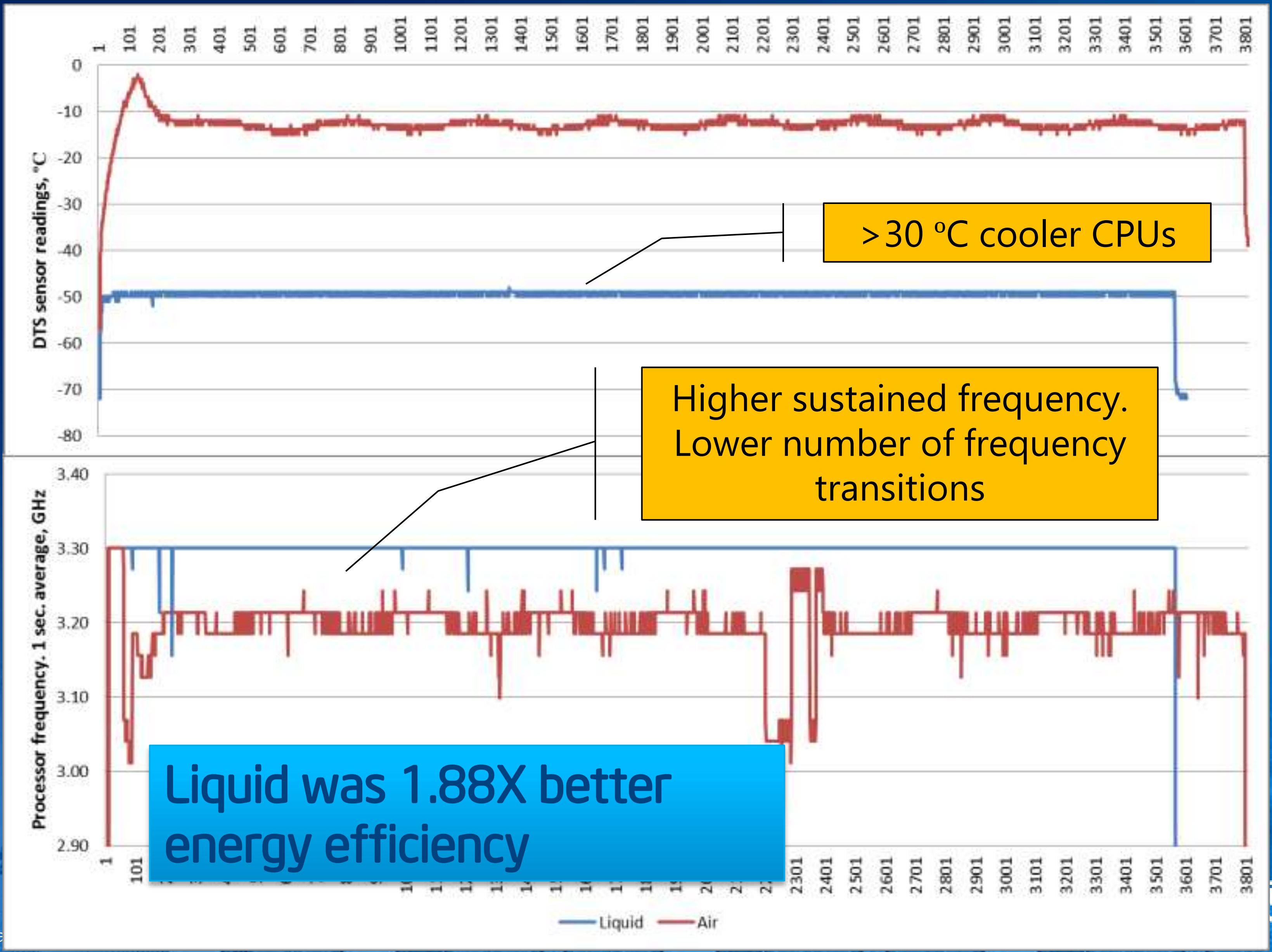
- Every water system is an on-going chemistry and biology experiment
- Closed loop systems need water quality monitoring and maintenance
- In water systems; there is no such thing as "zero" – there is always some bacteria, minerals, dissolved solids, just at trace levels

Ignoring water quality and water treatment guarantees failure

Case-study

- Compare direct-liquid cooled system vs. air-cooled
- Use as much as possible the same components
- Working with RSC, who uses EPSD Jefferson Pass board in their liquid-cooled HPC system design called Tornado
- The same board, CPUs, memory DIMMs, etc. used for benchmarks in liquid and air-cooled environment
- Cooling subsystem and power delivery are only two things which differ
- Consider meaningful HPC applications
- Likely not all apps will yield the same improvement: focus on those which are frequency bound and not power limited, Turbo to allow performance impact from the higher Turbo upsides

Observations



Liquid was 1.88X better energy efficiency

The four essential HPC elements

Air

Proceedings of the ASME 2013 International Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Microsystems
InterPACK2013
July 16-18, 2013, Burlingame, CA, USA

InterPACK2013-73163

A FIELD INVESTIGATION INTO THE LIMITS OF HIGH-DENSITY AIR-COOLING

Michael K Patterson
Intel Architecture Group
Intel Corporation
Dupont, Washington

Randall Martin
Clemson University
Clemson, South Carolina

J. Barr von Oehsen
Clemson University
Clemson, South Carolina

Jim Pepin
Clemson University
Clemson, South Carolina

Yogendra Joshi
G.W. Woodruff School of
Mechanical Engineering
Georgia Institute of Technology
Atlanta, Georgia

Vaibhav K Arghode
G.W. Woodruff School of
Mechanical Engineering
Georgia Institute of Technology
Atlanta, Georgia

Robin Steinbrecher
Intel Architecture Group
Intel Corporation
Dupont, Washington

Jeff King
Intel Architecture Group
Intel Corporation
Hillsboro, Oregon

ABSTRACT
In this paper we report on a field investigation into airflow management challenges in high density data centers. This field investigation has also served to validate laboratory investigations into high density air cooling issues. In data centers with significant power consumption, and consequently high cooling loads per rack, high volumes of room airflow are required to meet server cooling airflow requirements. These volumes of air can be difficult to deliver in raised floor hot aisle / cold aisle layouts. The velocity of the airflow is such that it creates a negative pressure near the bottom of the rack. This negative pressure entrains air from under and behind the rack, causing recirculation and warmer air being provided to the servers at the base of the rack. This can cause operational problems and server performance impacts. This phenomenon has been demonstrated in previous papers reporting on test data using particle imaging velocimetry (PIV) techniques. The current work validates those studies by looking at airflow, infrared thermography, and actual IT performance while the under rack recirculation flows are occurring. Additionally, we demonstrate significant improvement by employing rigorous airflow management practices. We also discuss the limitations of current CFD modeling, the majority of which does not have sufficient grid-wise resolution to capture the problem. Further we discuss typical operational conditions that have suppressed the problem (or perhaps the awareness of) to date. Finally, the paper recommends best practices to mitigate the problem in high density data centers.

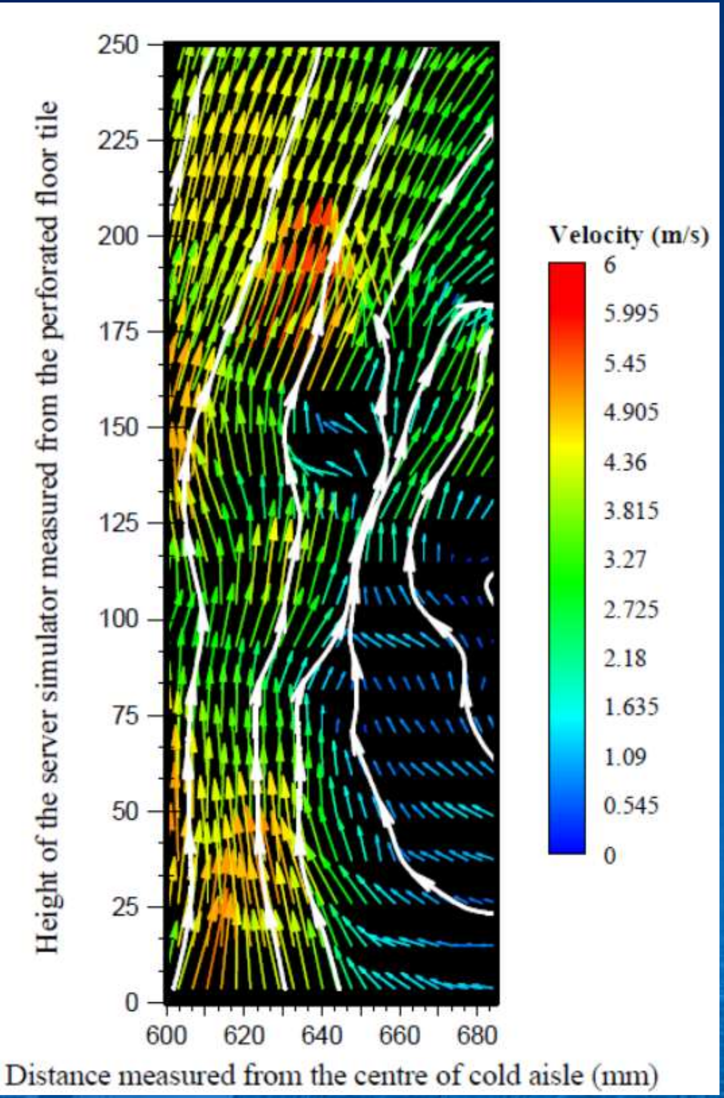
INTRODUCTION
Data Center design has become as big a challenge as the design of the IT systems that they support. Today's IT manufacturers have made the development and deployment of

ASME Paper from this summer's InterPACK

e-mail me for a copy

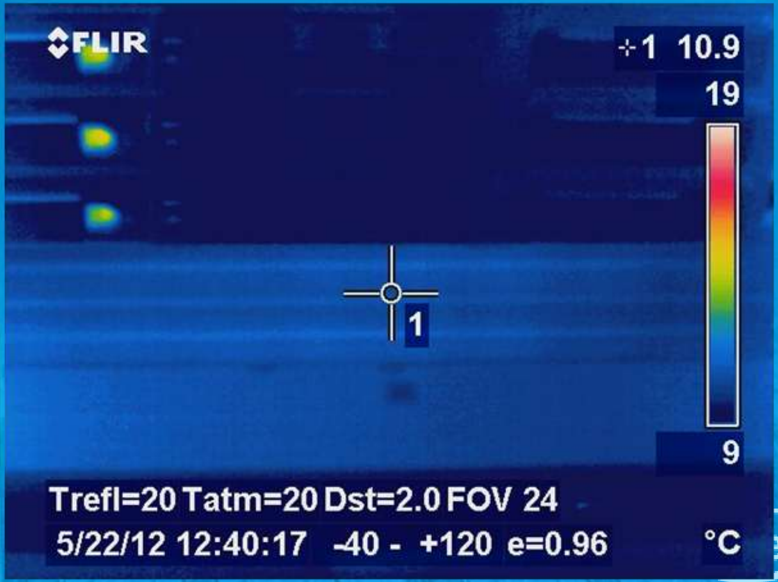
Documents the air-cooling issues we have found with high density airflow challenges on raised floors

Particle Imaging Velocimetry



- High velocity thru raised floor tiles causes a venturi effect
- Creates a negative flow at the server inlet
  - The server will get the airflow needed…but maybe not at the right temperature
- Experienced this with as low as 1000 cfm (470 l/sec)

Before and after...



For medium flowrates, airflow management best practices solved the problem

- containment
- blanking plates

High flowrates may still have negative flows

There will always be air-cooling in HPC, but the high end is trending towards liquid

Short term

- Colder data centers during peak use
- Extra cold aisle tiles
- Adding localized cooling
- Containment
  - Cold aisle
  - Hot aisle
  - Chimneys
- Additional supply ducting

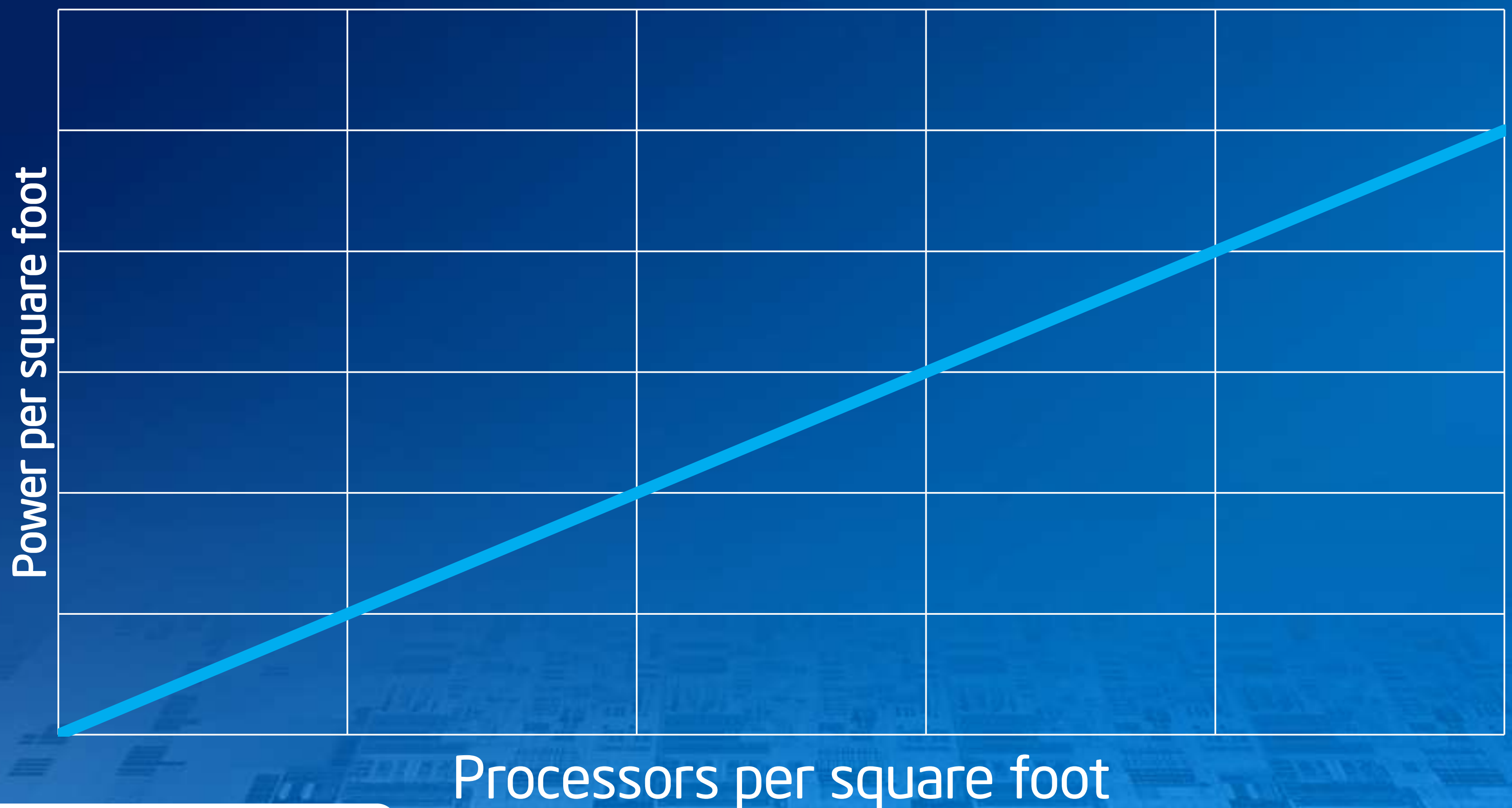
Long Term

- Recognize that HPC is moving beyond standard data center practices
- Trending from 30 kW to 60-200 kW racks
- Engage the suppliers and data center designers in advanced thermal management solutions
- Higher levels of vertical integration provided with the cluster procurement

The four essential HPC elements

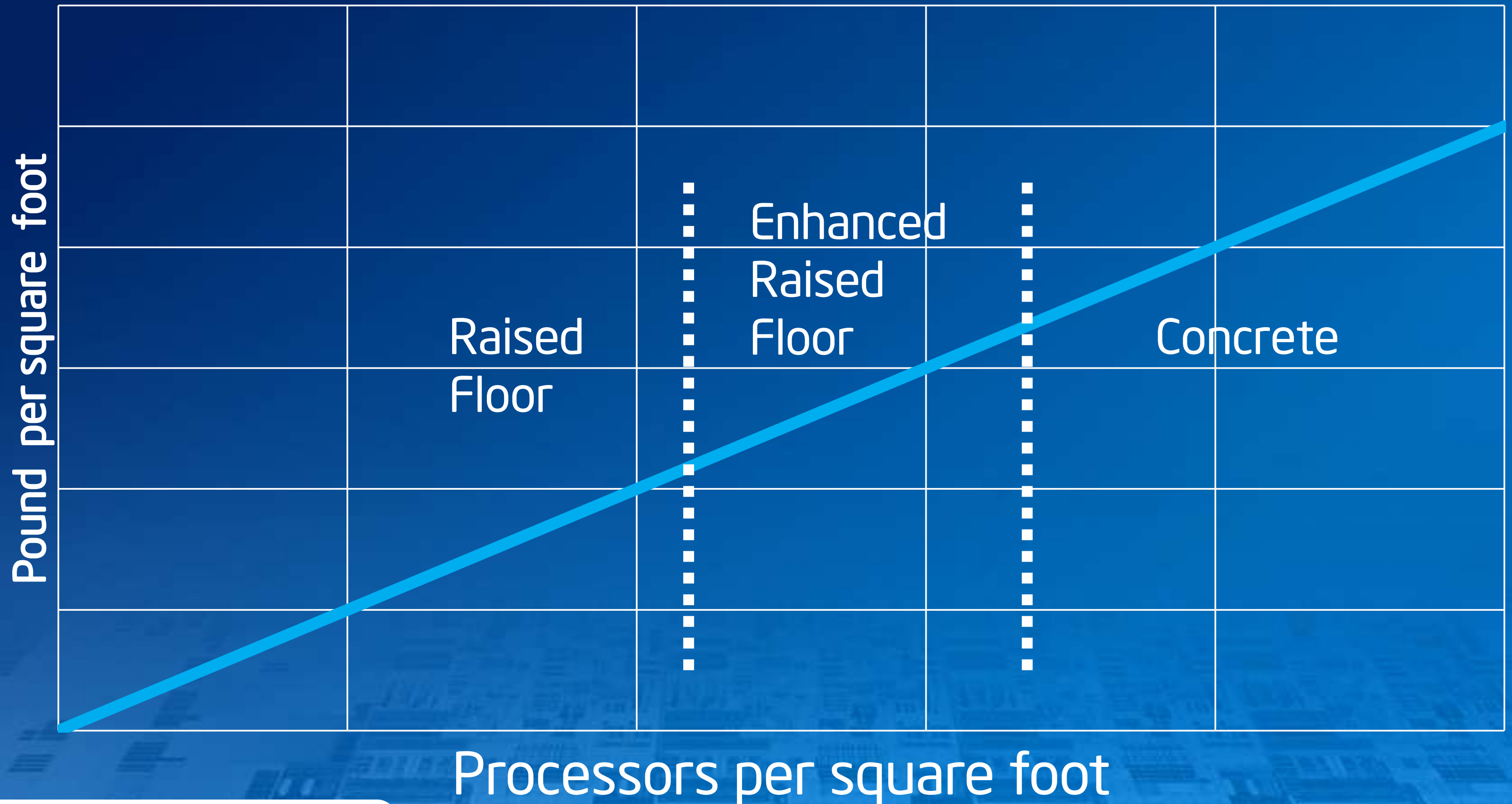
Weight & Density

Density Trending – Power



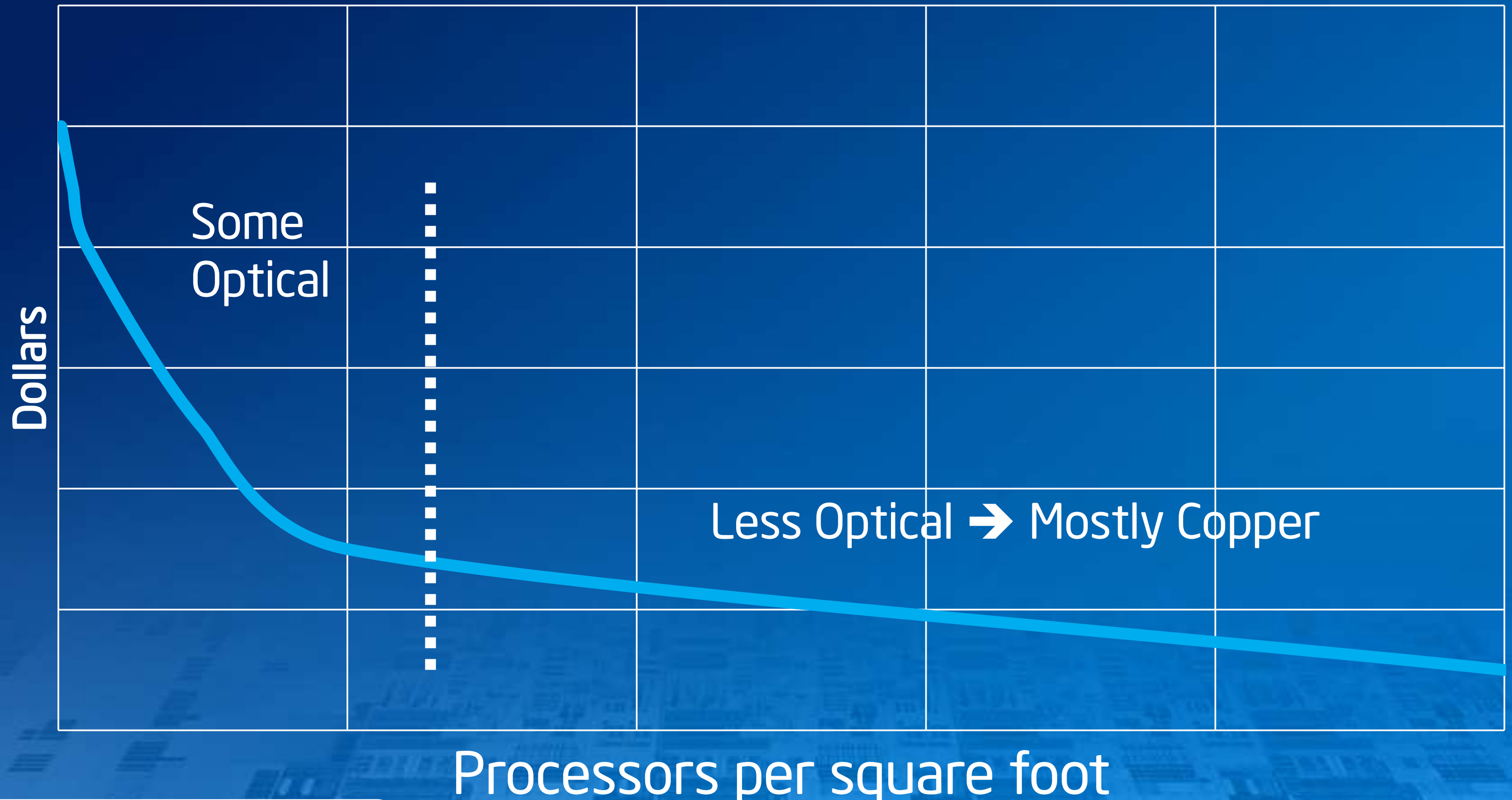
Not to scale, for demonstration purposes

Density Trending – Weight



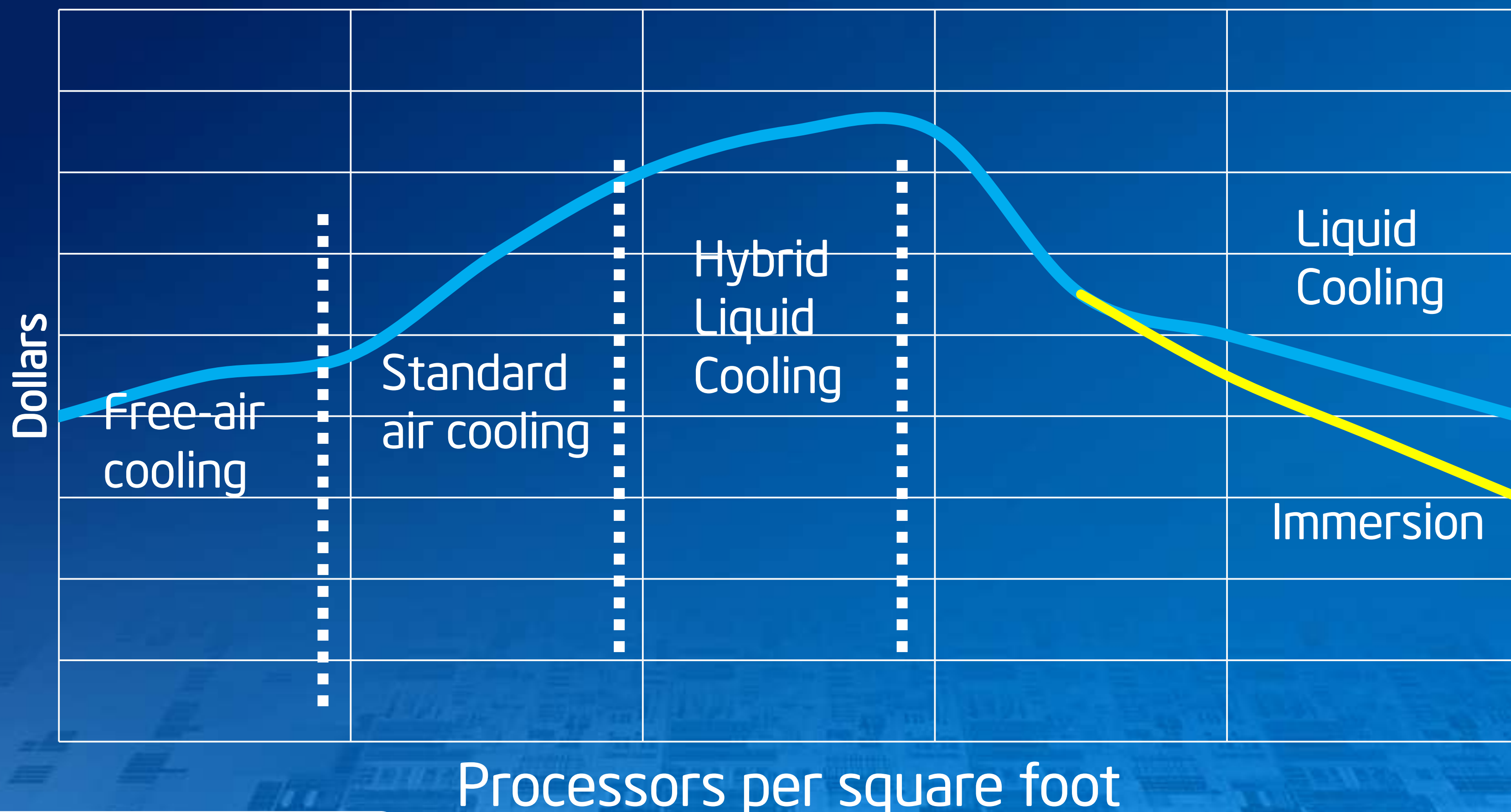
Not to scale, for demonstration purposes

Density Trending – Fabric Cost



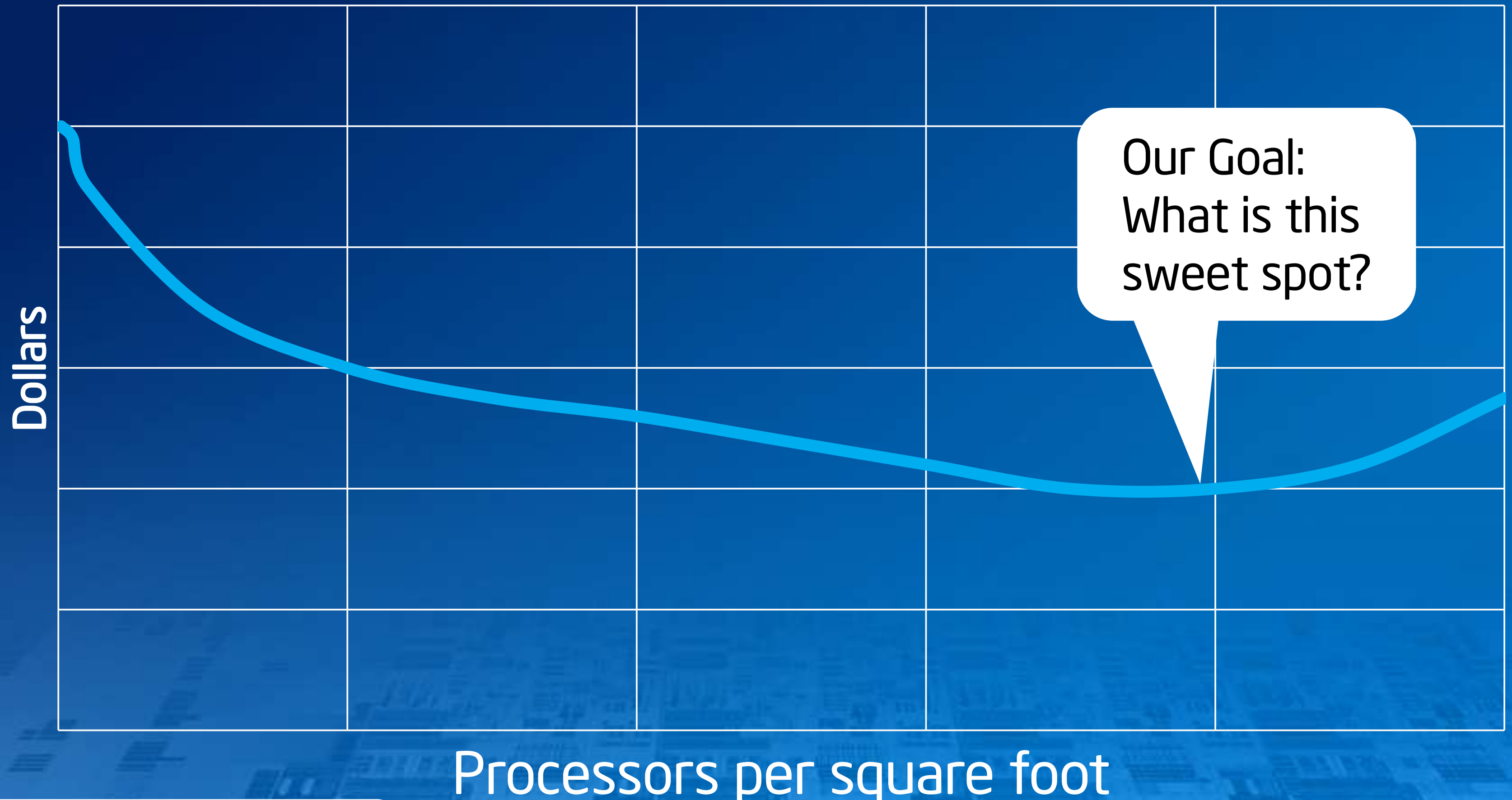
Not to scale, for demonstration purposes

Density Trending – Cooling Cost



Not to scale, for demonstration purposes

Density Trending – TCO: OpEx+CapEx



Not to scale, for demonstration purposes

The four essential HPC elements

Power

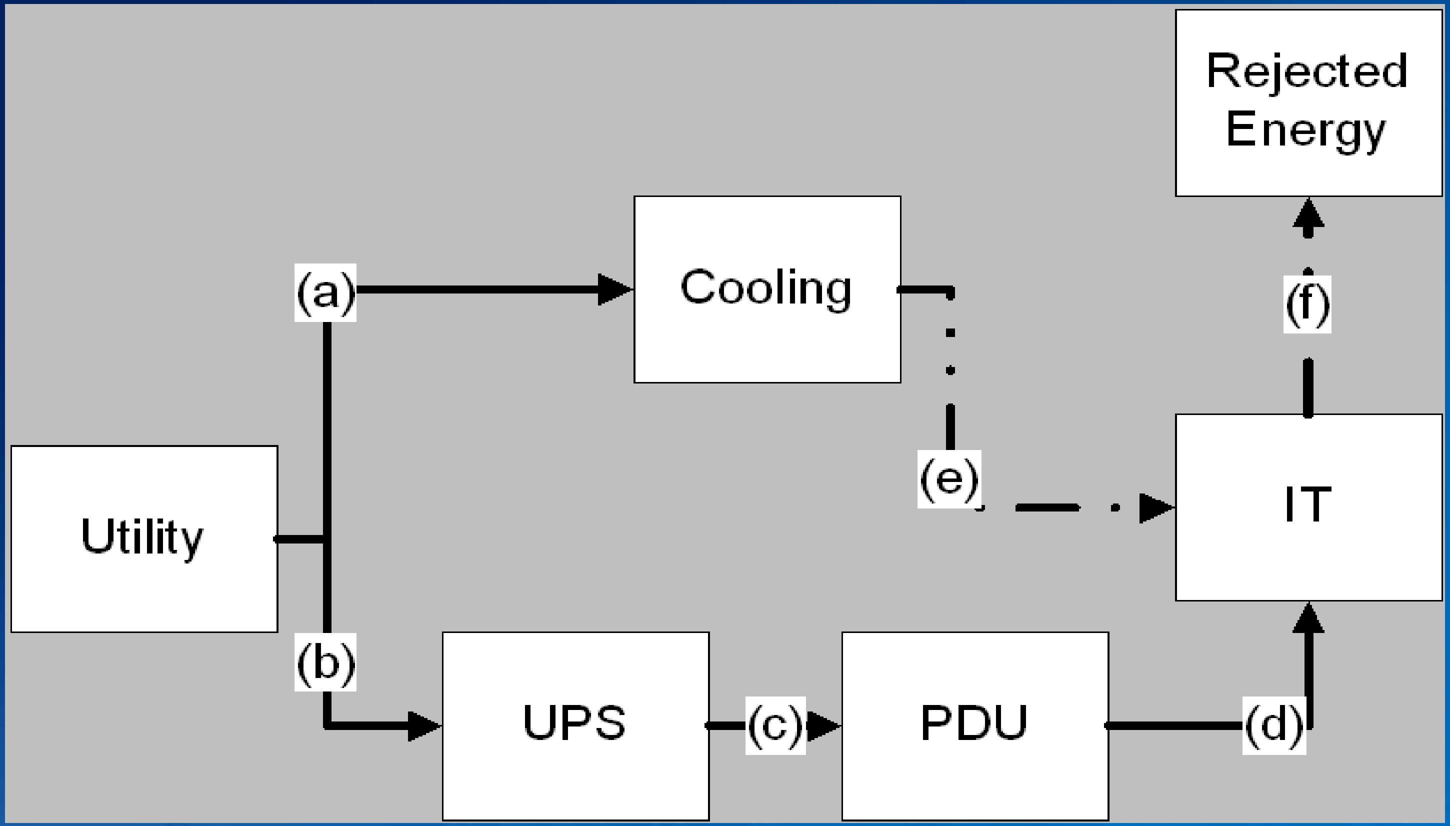
Utility Interaction

- As large systems trend towards 10 to 30 MW, planning and coordination with the utility must increase
- We are seeing Δpower/time requirements
  - Exploring specifics of these
  - May add features in software and resource manager to change start-up and ramp-down times

EEHPC WG

- WG brings together experts in both infrastructure and HPC computing systems
- Leadership role in providing analysis and recommendations in topics of importance to HPC community
  - Warm water cooling recommendations
  - PUE -> TUE
  - Data Center Energy Management Dashboard
  - Water cooling commissioning guidelines
  - Power measurement methodology – please use it for this springs Top 500 and Green 500 submissions

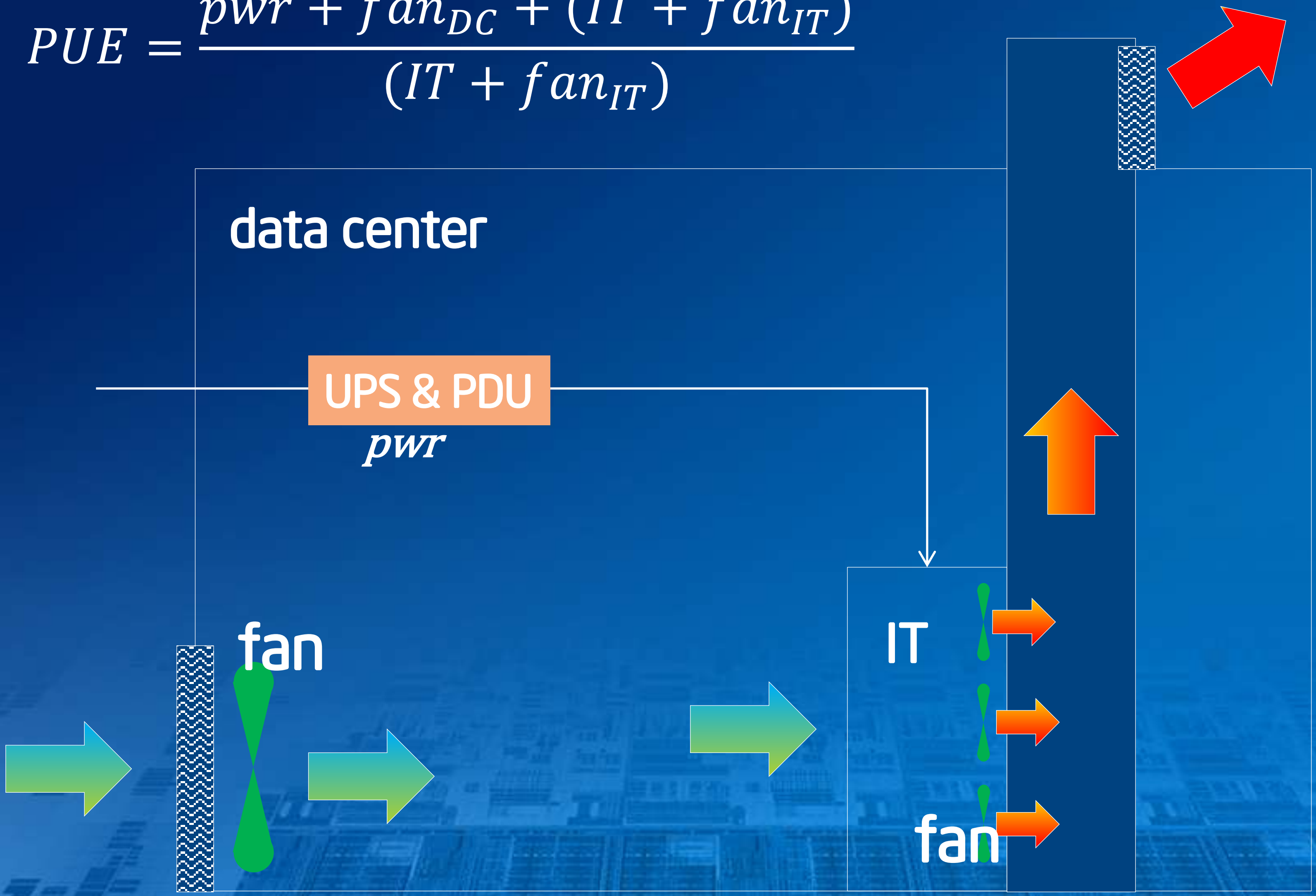
PUE Definition



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

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

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

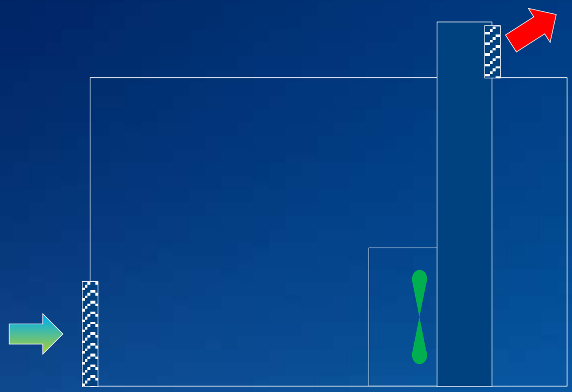


Three variations…

a) both fans

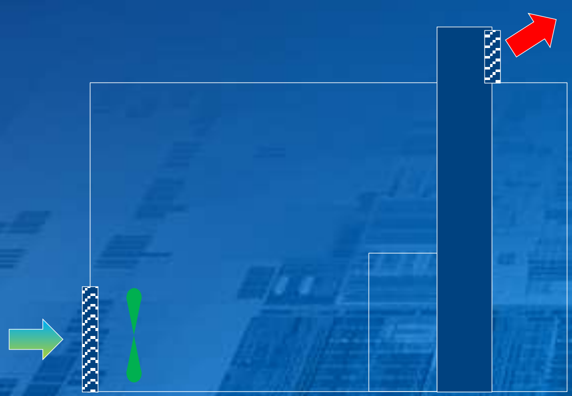
$$PUE_a = \frac{pwr + fan_{DC} + (IT + fan_{IT})}{(IT + fan_{IT})}$$

b) IT fans only



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

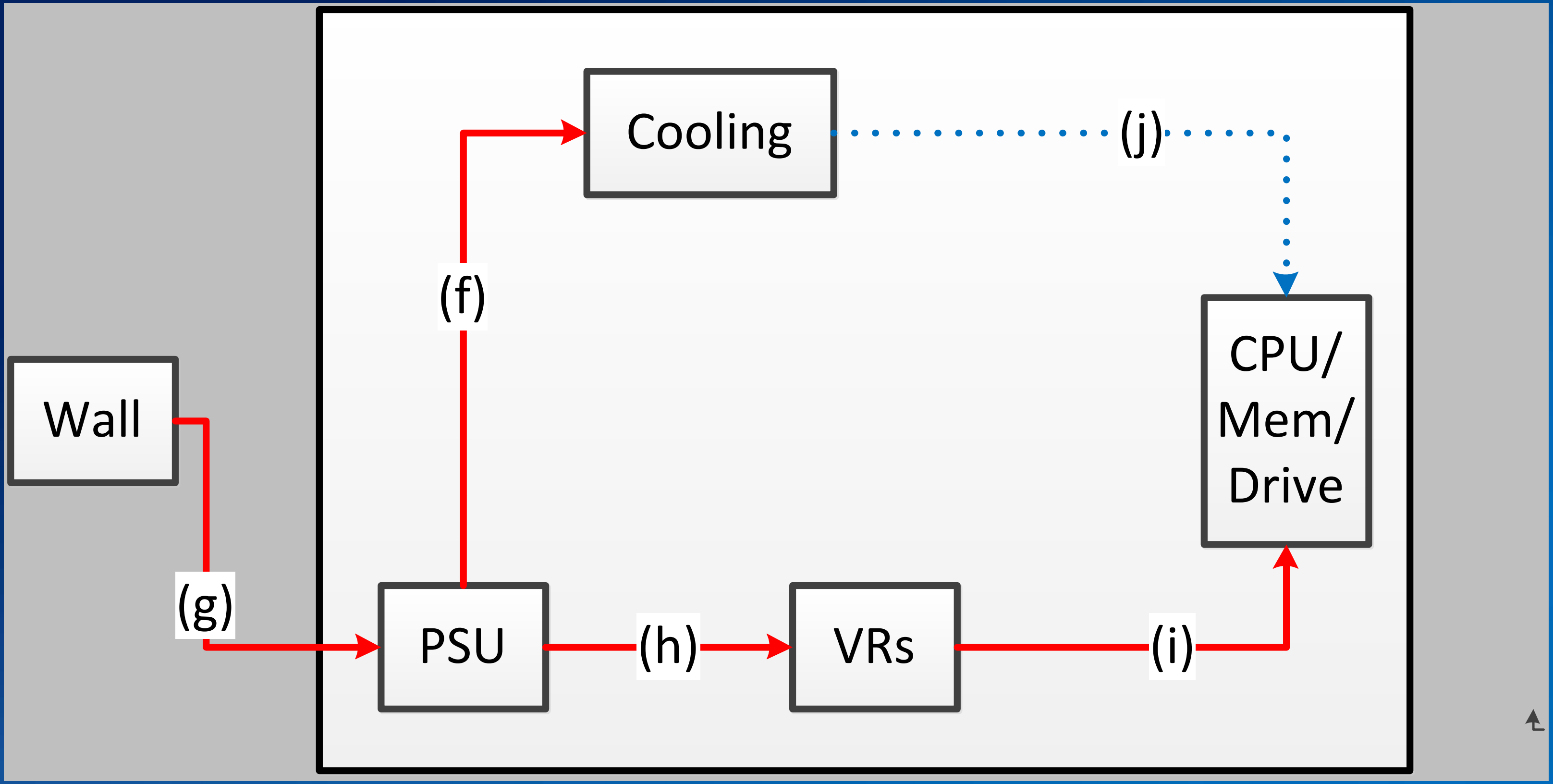
c) bldg fan only



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

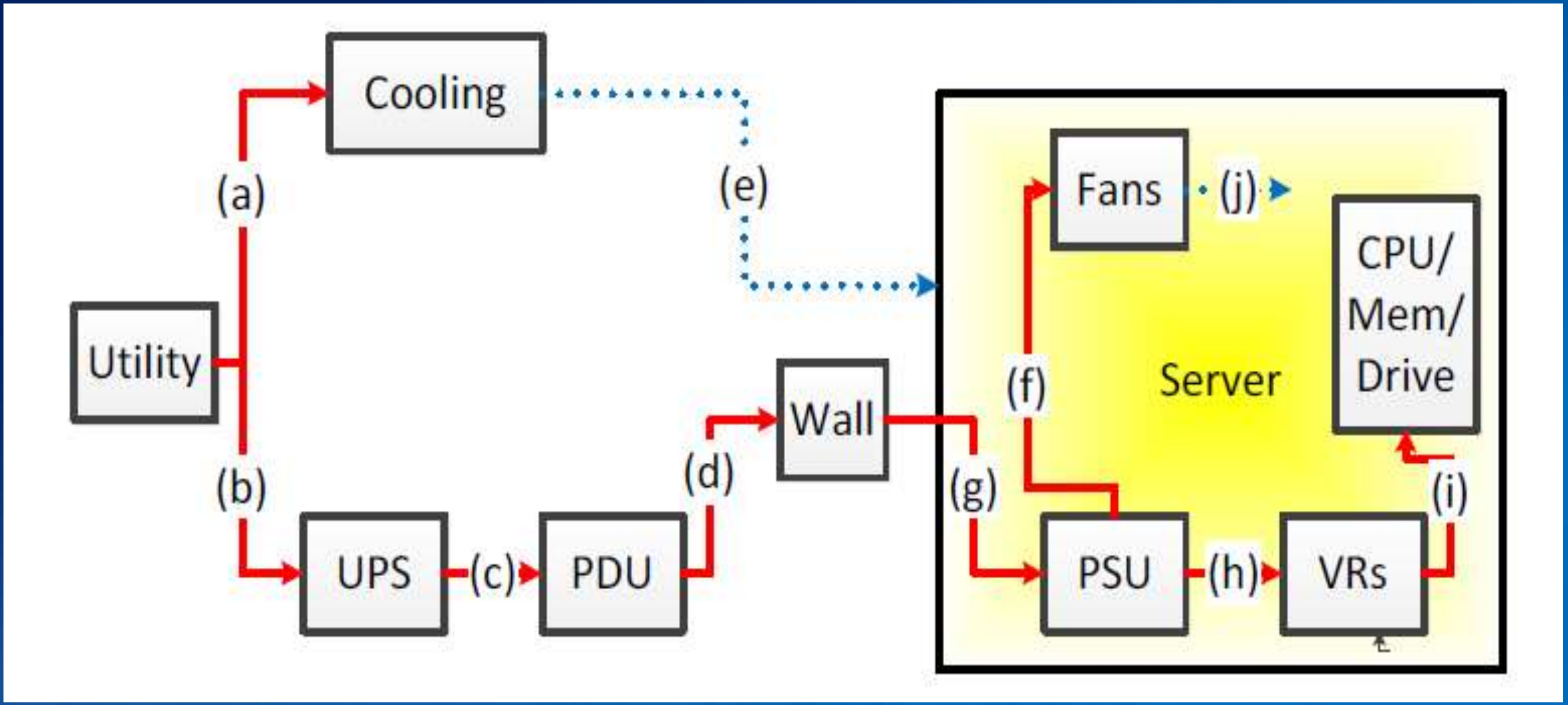
PUE_b < PUE_a < PUE_c but is (b) best?
We don't know….

ITUE



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

TUE



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

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

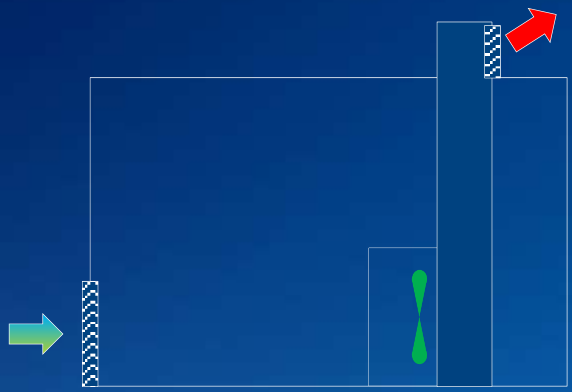
$$TUE = ITUE \times PUE = \frac{a+b}{i}$$

Does it work?

a) both fans

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

b) IT fans only



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

c) bldg fan only

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

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

TUE….

- ISC 2013, Leipzig
  - *TUE, a new energy-efficiency metric applied at ORNL's Jaguar*
  - http://eetd.lbl.gov/sites/all/files/isc13_tuepaper.pdf
- SC 2013, Denver
  - BoF: Total Power Usage Effectiveness: A New Take on PUE
  - Number of sites volunteered to evaluate and report

ExaScale Software energy and power challenges

- An ExaScale system in 2020 => ~millions of cores – a major challenge in terms of power consumption and data communication amongst all the cores
- Compute, network, memory, storage and datacenter infrastructure share limited total power envelope
  - Changing focus from just reducing power to sharing power efficiently to maximize science/kW
  - Goal is to get the most Science done within a fixed power budget
- Need smart balance of power between components for different applications
- Solution: interaction of hardware, resource-manager, and software at run-time to optimize for power or energy
  - establishing "knobs" and protocols for such interaction

The four essential HPC elements

Water

Weight & Density

Air

Power

Summary

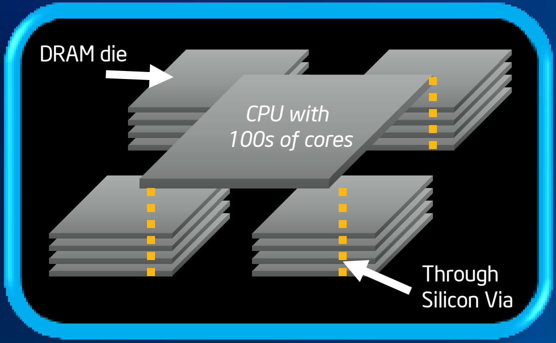
- Water
  - Cooling medium of the future for high density and high performance
  - Significant risk if water quality is not maintained
- Air
  - Air cooling will remain an important sector
  - Best practices required to push into high density
- Density
  - Pay attention to weight
  - Question raised floor use in new buildings
- Power
  - High density requires high power
  - ΔkW/min may be coming
  - EEHPC has new power measurement method
  - SW and Resource Mgr part of the solution

References and Resources

- Water
  - ASHRAE TC 9.9 – Water cooling book
  - http://tc99.ashraetcs.org/
- Air
  - ASHRAE TC 9.9 – range of books and papers
  - DC2020 - http://www.datacenter2020.com/
  - EU DC CoC - http://iet.jrc.ec.europa.eu/energyefficiency/sites/energyefficiency/files/best_practices_v3_0_8_2_final_release_dec_2011.pdf
- Density
  - ASHRAE TC 9.9 – Structural and Vibration book
- Power
  - EEHPC WG – http://eehpcwg.lbl.gov/
  - The Green Grid - http://www.thegreengrid.org/

The Path (Stair Steps) to Exascale….

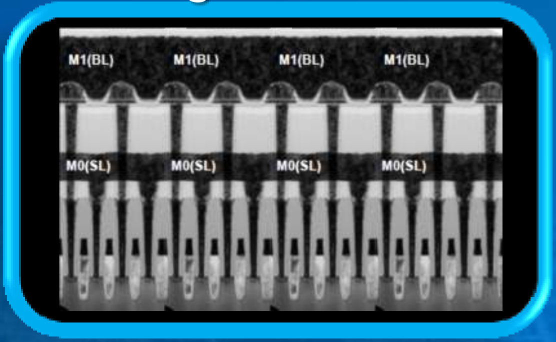
3D Integration



Next Generation Fabric

Software & Programmability

New Memory
Technologies & Hierarchies



….Will Be Provided by Well Optimized Technologies & Architectures
Co-Designed & Working WellTogether at the System Level

Thank You.   Questions?

michael.k.patterson@intel.com

intel®