Paradigm Shifts in HPC

Frank Baetke

Cetraro HPC Workshop – July 7th, 2014

1

Apollo (Warm-water cooled)

Think different.

Why use air, a commonly used “insulator”, as the default heat removal mechanism?

$$h_{water} = 50\text{-}100 \times h_{air}$$

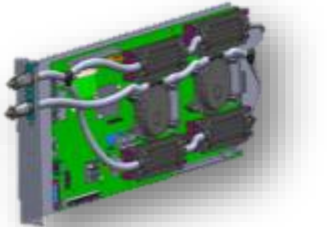
$$h = \frac{Q}{A * \Delta T}$$

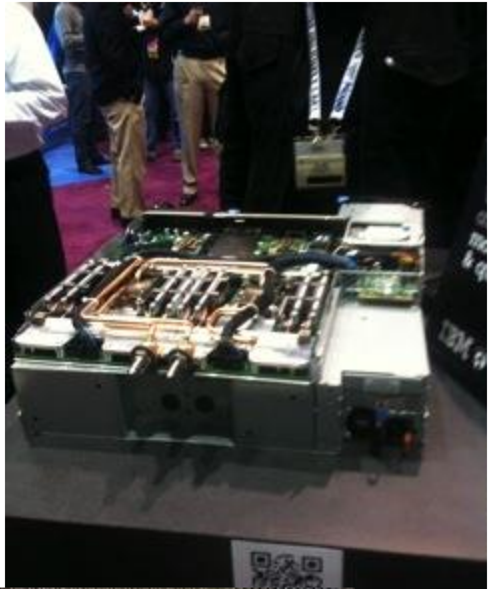
h: heat transfer coefficient
Q: heat input (W)
A: heat transfer surface area (m^2)
ΔT: Delta-T (K)

=

Liquid Cooling today

Components, cold-plates, immersion…





Hardcore Computer's Liquid Blade™ Server

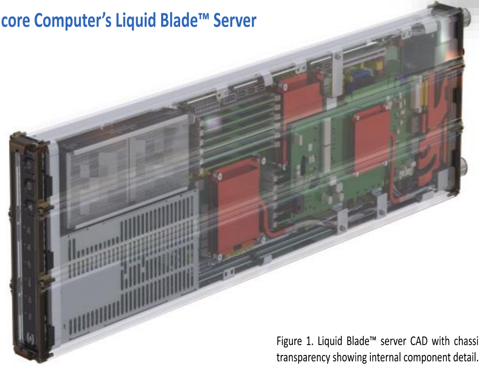
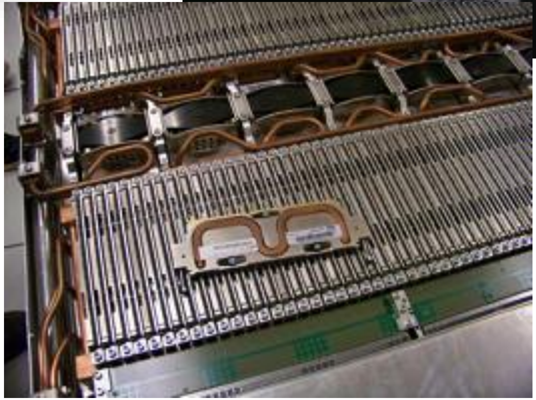
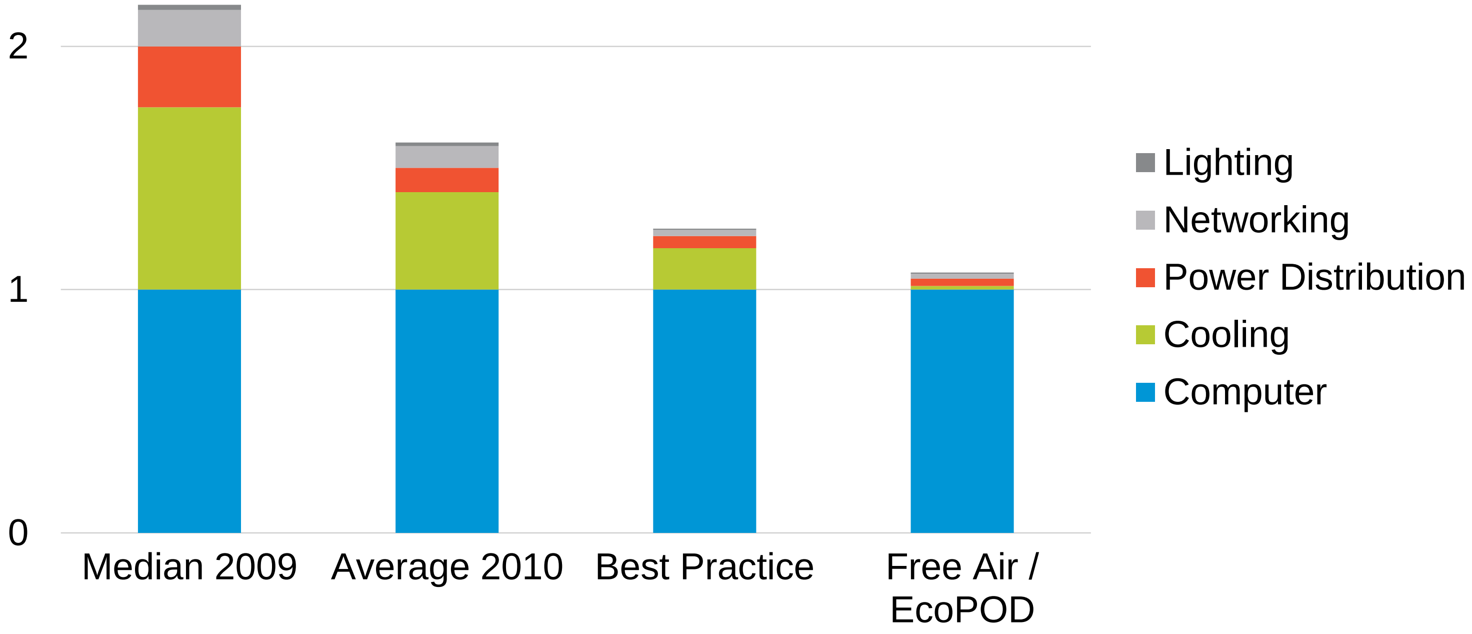


Figure 1. Liquid Blade™ server CAD with chassis transparency showing internal component detail.



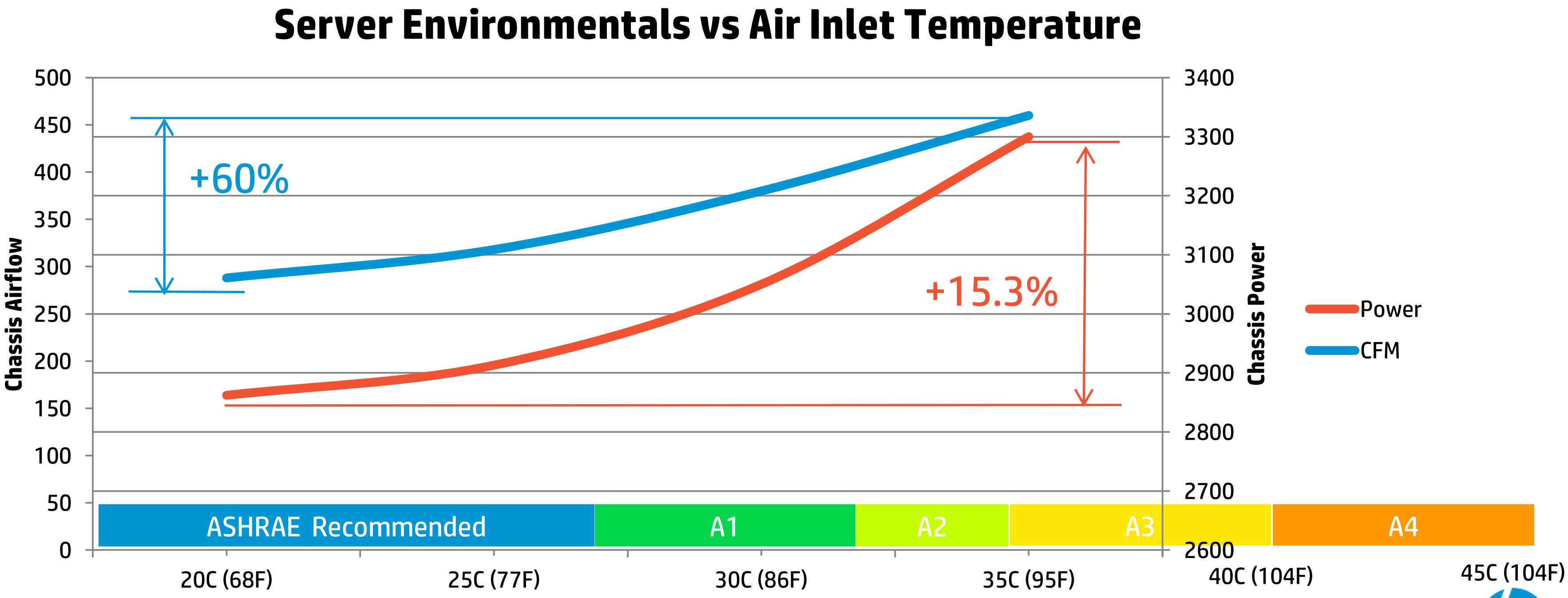
PUE

The “holy grail” of corporate IT?



PUE: “the untold story”

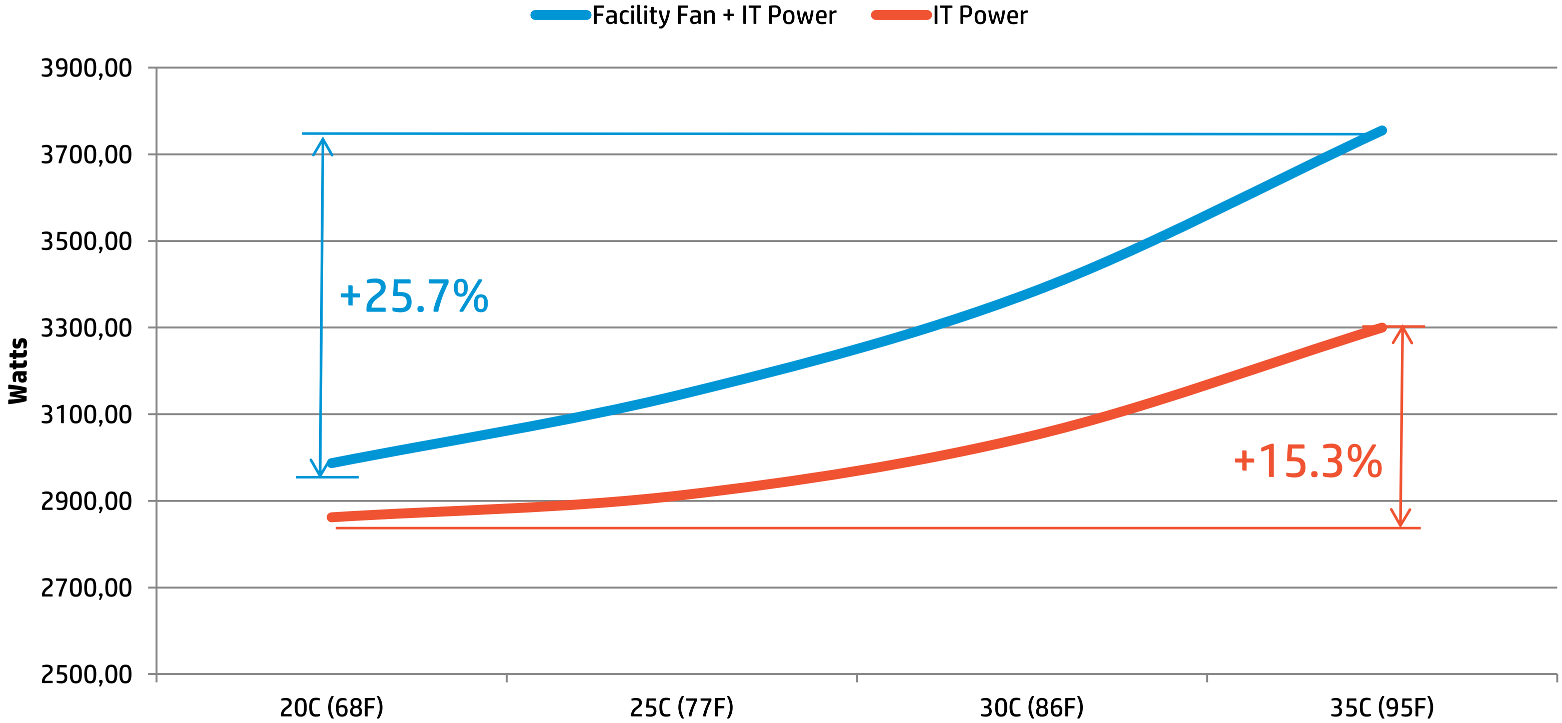
When you CAN’T afford Free Air Cooling



8x SL230 w/ E5-2670, 16x8GB, 4xSFF, Linpack

Free Air Cooling’s “dirty little secret”

Fan power α (fan speed)^3

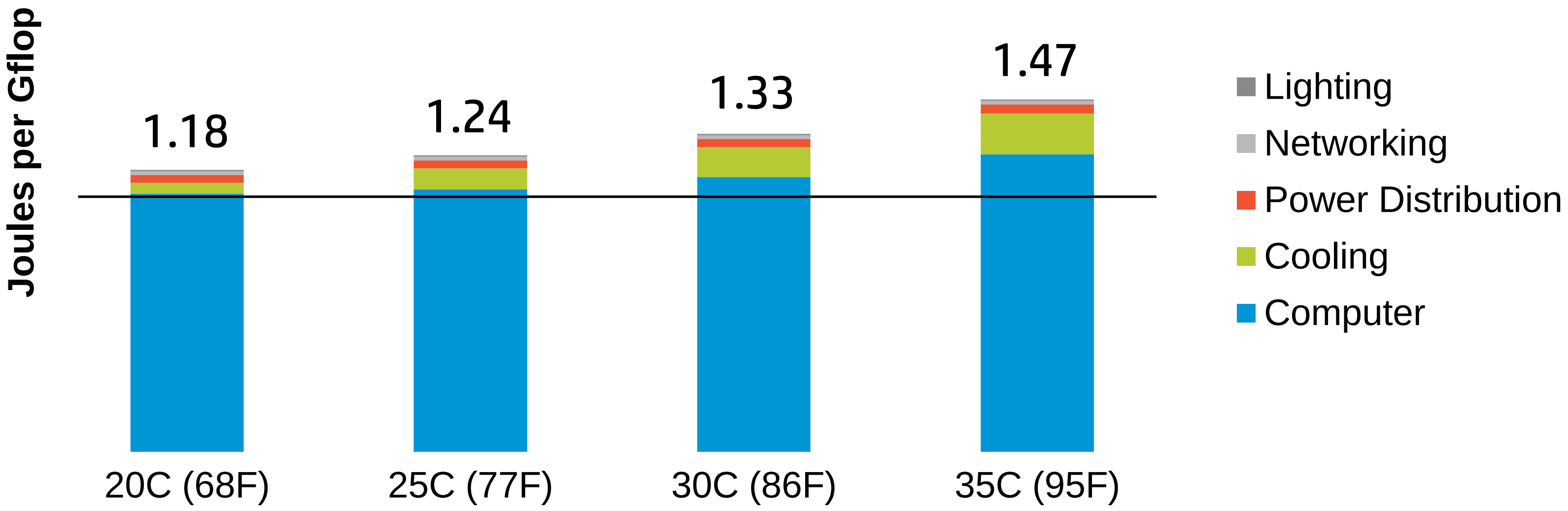


Baseline: 8% facility fan power @ 25C

What a ratio like PUE does not show

Looking at Energy per Compute Operation

≈ energy of 20C server inlet in a DC with PUE of 1.37



Apollo Rack: Liquid Cooling made Comfortable

“Datacenter in a rack”

- **Cooling**
  - Liquid: CPUs, GPUs, DIMMs
  - Air to Liquid heat-exchanger: remaining components

- **Power**
  - Up to 80kW (4x 30A 3ph 380-480VAC)
  - Cooling capacity: up to 100kW

- **Supporting infrastructure**
  - Integrated Fabrics: InfiniBand, Ethernet, Management
  - Pooled power, Battery backup unit…
  - Taking IPMI to new levels

The New HP Apollo 8000 System

Advancing the science of supercomputing

Scientific Computing

- Research computing
- Climate modeling
- Protein analysis

Manufacturing

- Product modeling
- Simulations
- Material analysis

Leading teraflops per rack for accelerated results

- **Up to 150 teraflops/rack with compute trays**
- > **250 teraflops/rack with accelerator trays**

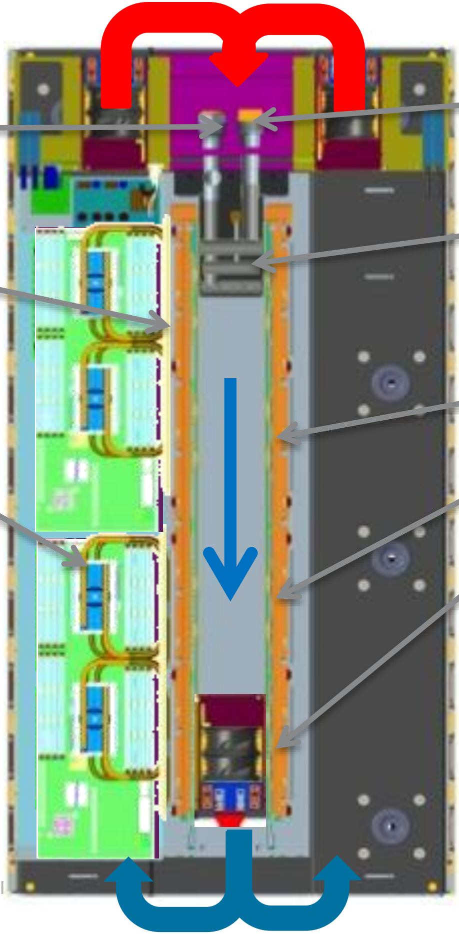
Efficient liquid cooling without the risk

- **Dry-disconnect** servers, intelligent Cooling Distribution Unit (iCDU) monitoring and isolation
- **Management** to enable facility monitoring, environmental controls and power management

Redefining data center energy recycling

- Save up to **3,800 tons** of CO2/year (790 cars)
- **Recycle water** to heat facility

Apollo Rack - Hybrid Cooling Concept



Back

Cool Water in

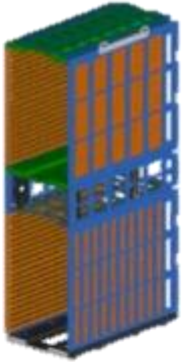
Warm Water Out

Air to Water
Heat Exchanger

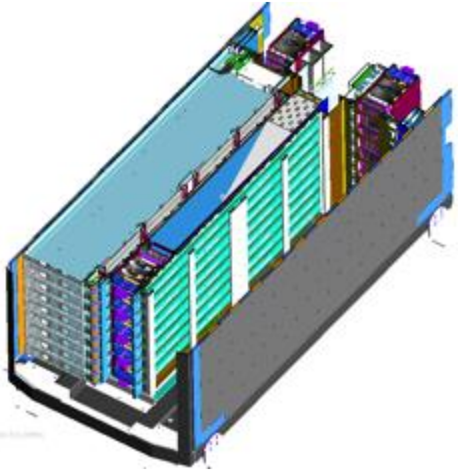
Dry Surface
Contact Area

Heat Pipes

Thermal Bus Bar
(Four Regions per side)



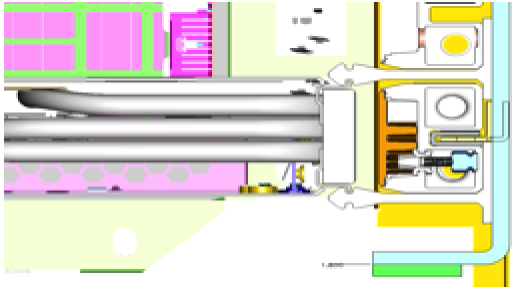
Modules



Front

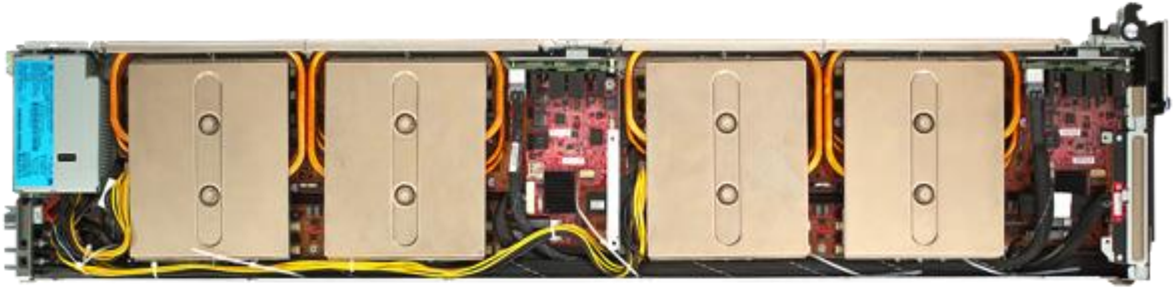
Cooling Technology

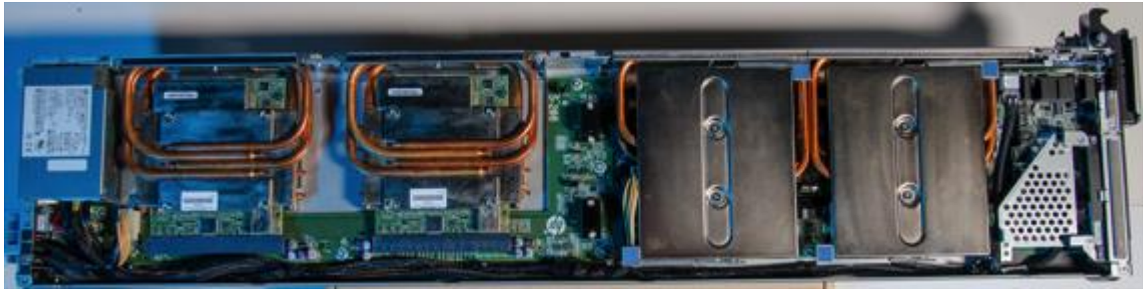
“dry-disconnect”



Trays

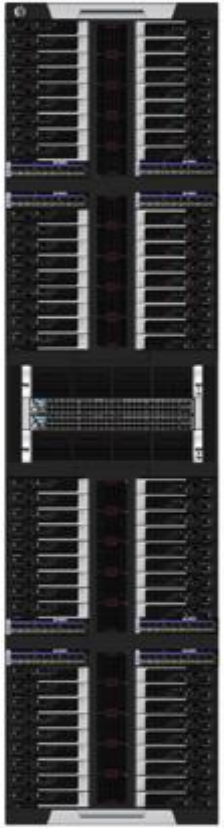
2x 2 CPUs, 2CPUs + 2 Phis



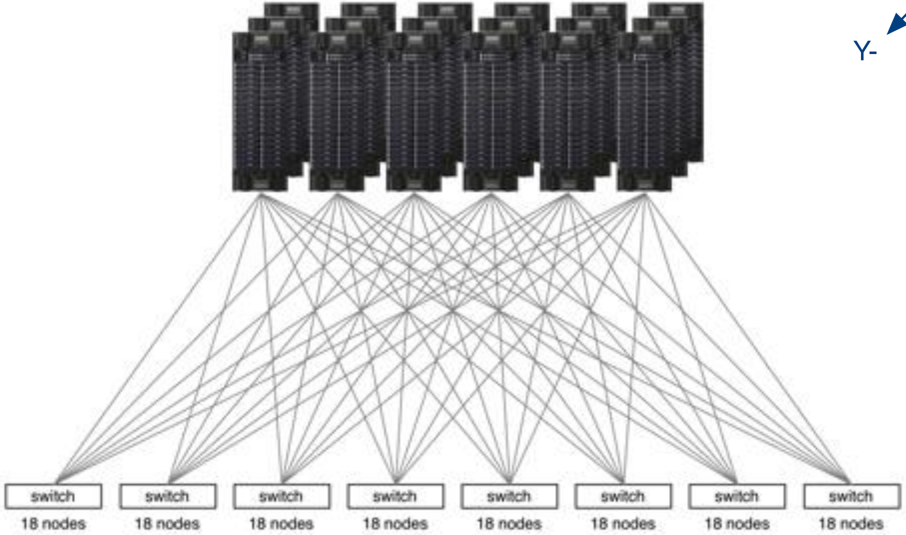


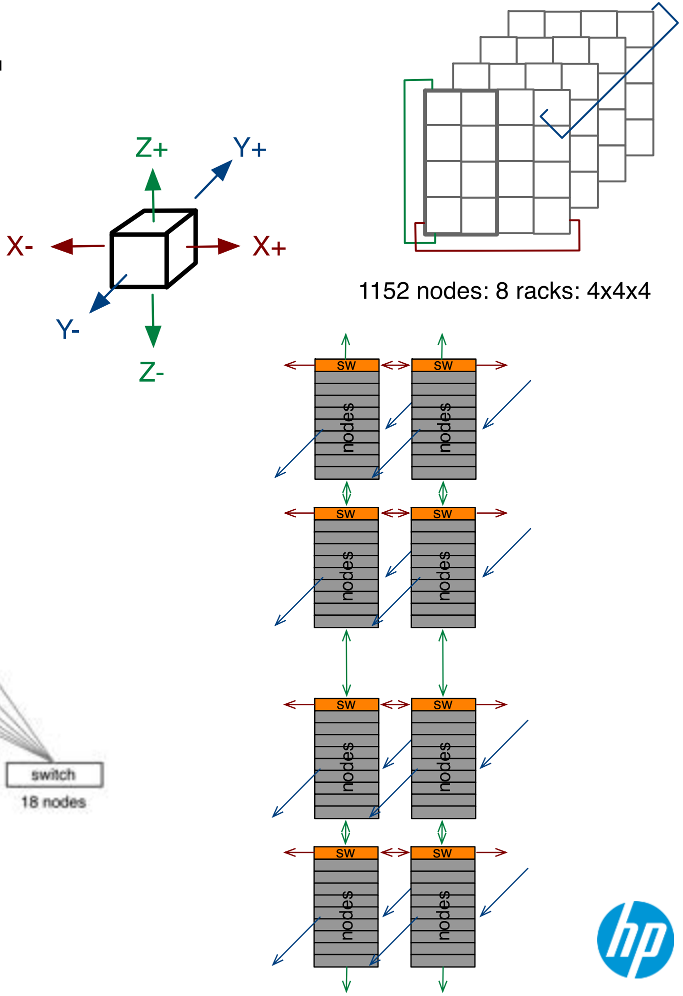
Fabrics

Infiniband: Fat-Tree, 3D Torus…



Config #1, 5 hops maxed out
1:1 bissection
5 hops, 18x 648p core switches
11664 nodes, 81 FlexRacks
Leaf switch integration
11664 QSFP cables





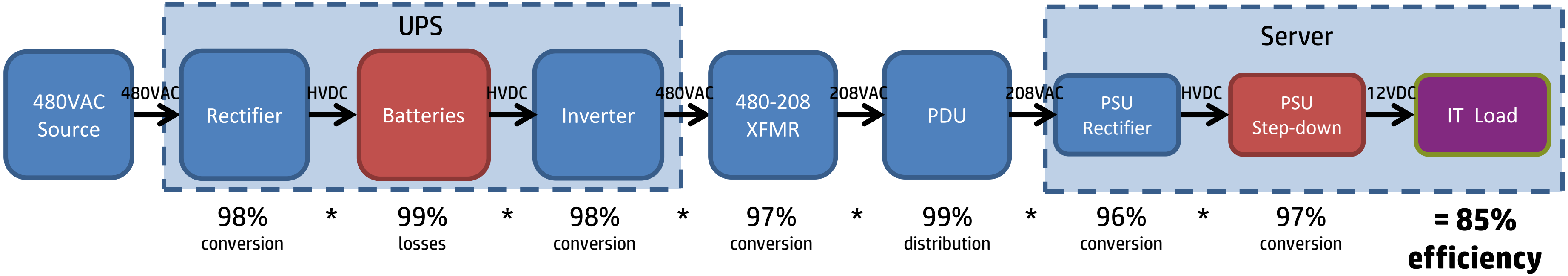
1152 nodes: 8 racks: 4x4x4

Power and Monitoring

Power Distribution Efficiency

High-Voltage AC to the Rack: Limiting conversion steps

Typical



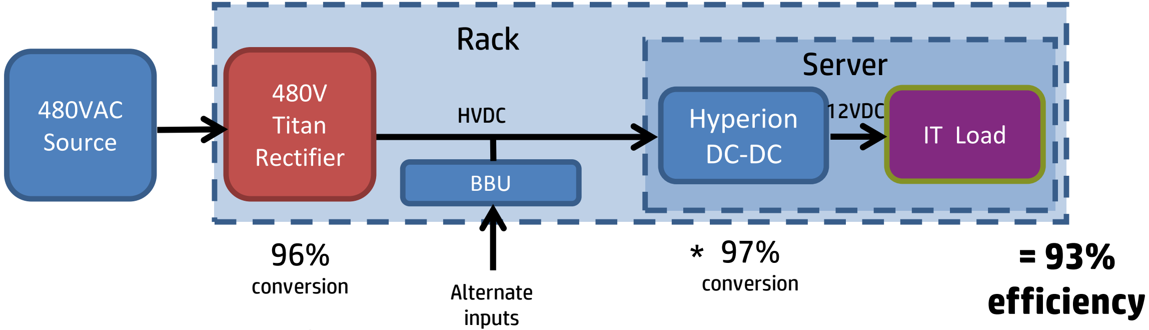
UPS

Server

480VAC Source → 480VAC → Rectifier → HVDC → Batteries → HVDC → Inverter → 480VAC → 480-208 XFMR → 208VAC → PDU → 208VAC → PSU Rectifier → HVDC → PSU Step-down → 12VDC → IT Load

98% conversion * 99% losses * 98% conversion * 97% conversion * 99% distribution * 96% conversion * 97% conversion **= 85% efficiency**

Apollo



Rack

Server

480VAC Source → 480V Titan Rectifier → HVDC → Hyperion DC-DC → 12VDC → IT Load

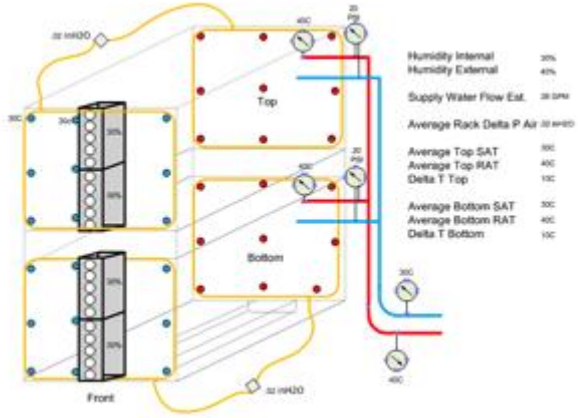
BBU

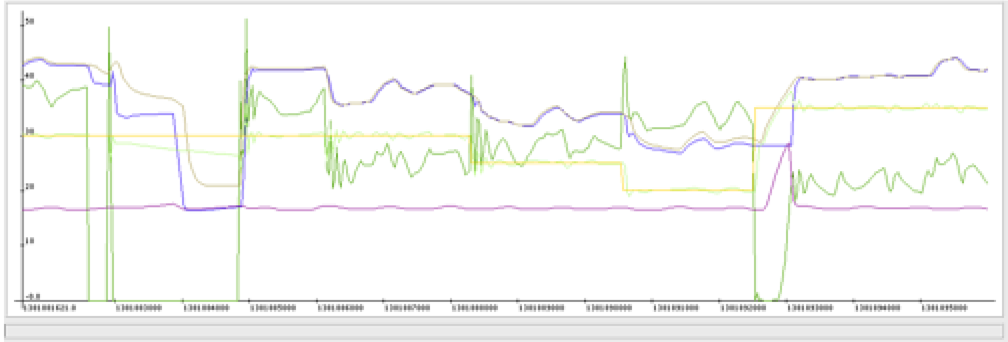
Alternate inputs

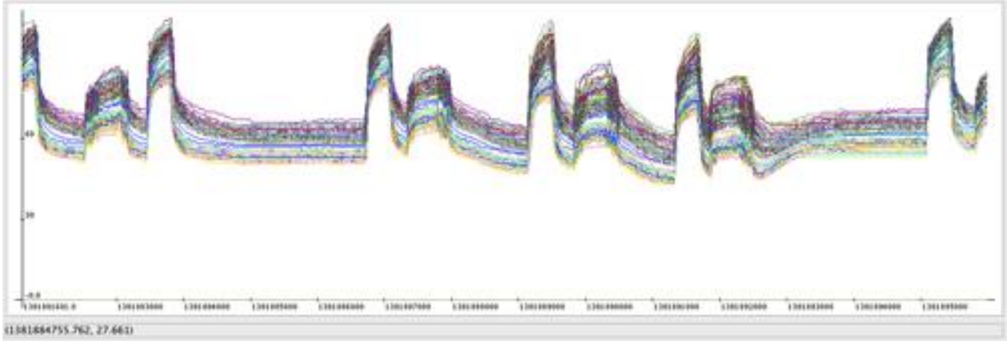
96% conversion * 97% conversion **= 93% efficiency**

Sensors and monitoring

You can’t optimize what you can’t measure



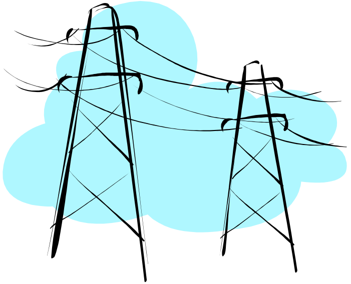


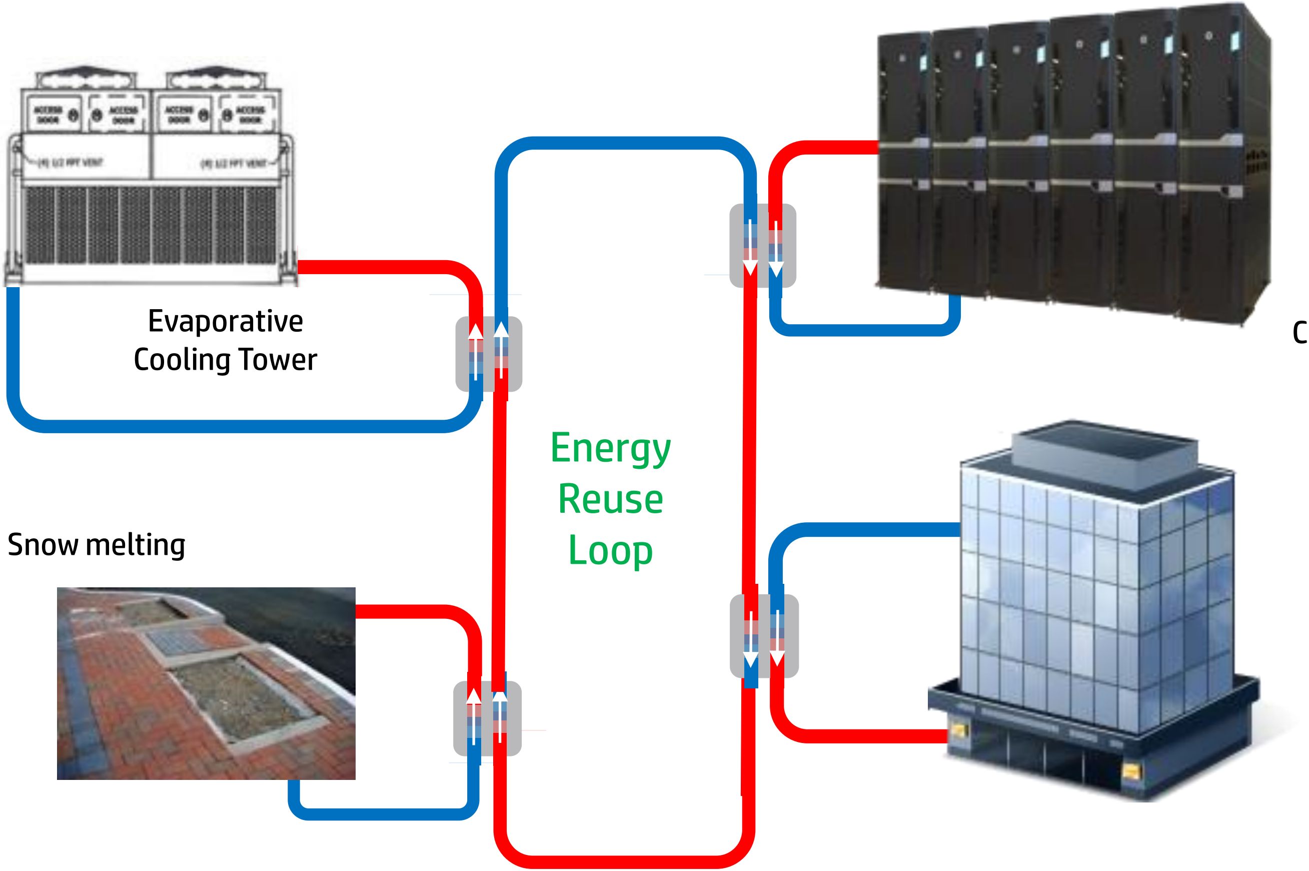


(1381884755.762, 27.661)

Systems

Chiller less Data Centers





Evaporative Cooling Tower

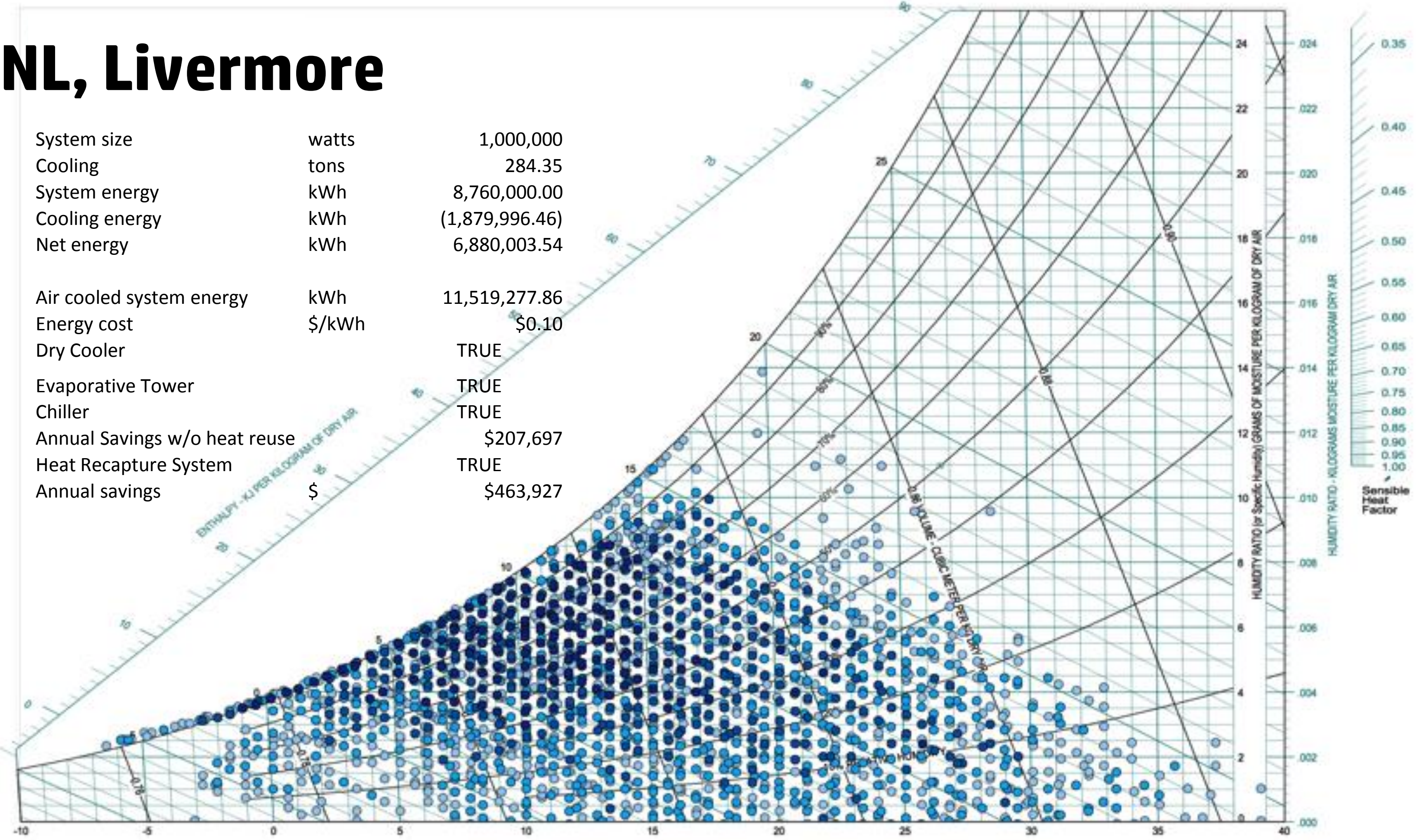
Energy Reuse Loop

Snow melting

Warm Water Cooled Servers

Building Heating

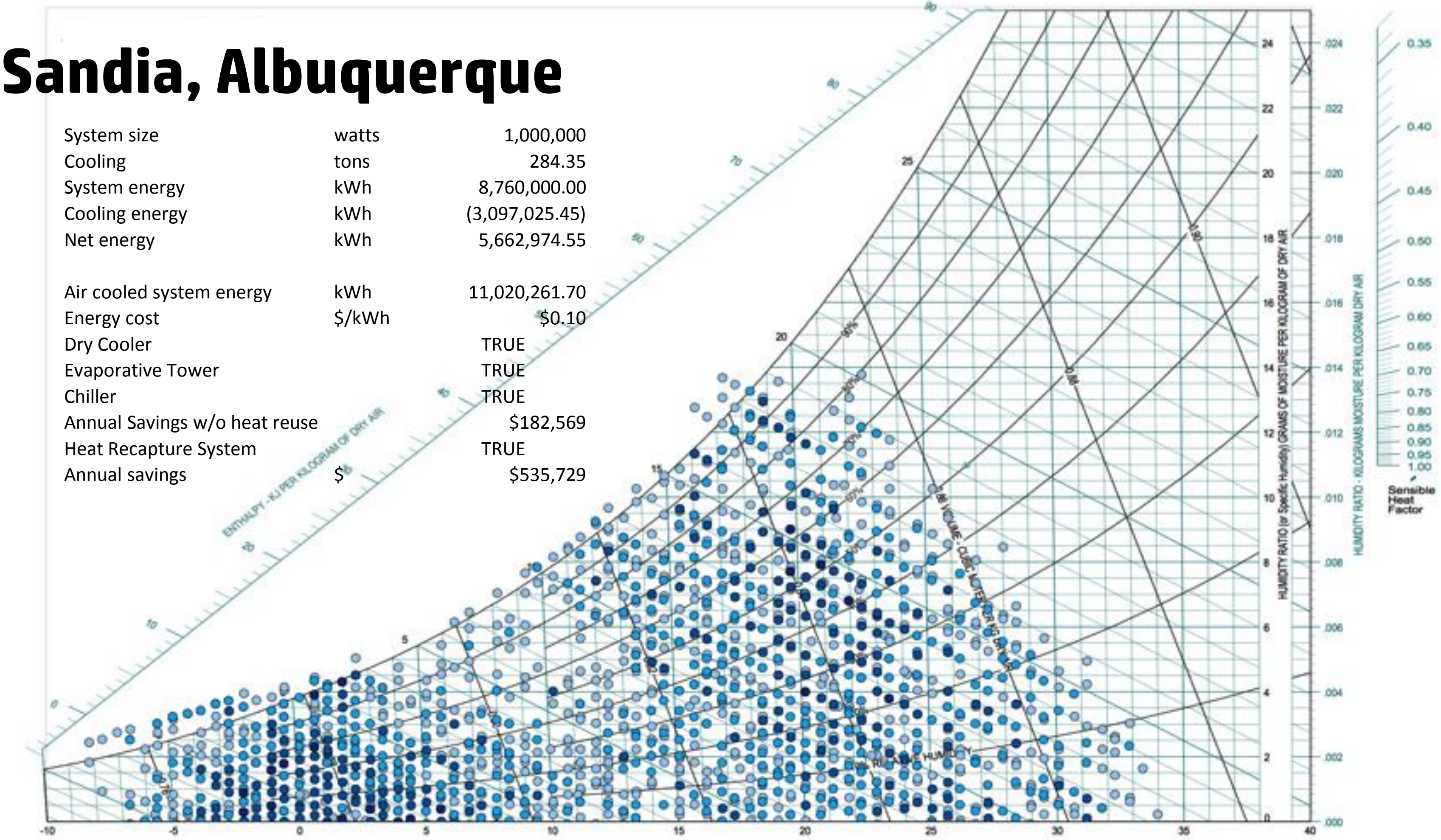
LLNL, Livermore



System size	watts	1,000,000
Cooling	tons	284.35
System energy	kWh	8,760,000.00
Cooling energy	kWh	(1,879,996.46)
Net energy	kWh	6,880,003.54
Air cooled system energy	kWh	11,519,277.86
Energy cost	$/kWh	$0.10
Dry Cooler		TRUE
Evaporative Tower		TRUE
Chiller		TRUE
Annual Savings w/o heat reuse		$207,697
Heat Recapture System		TRUE
Annual savings	$	$463,927

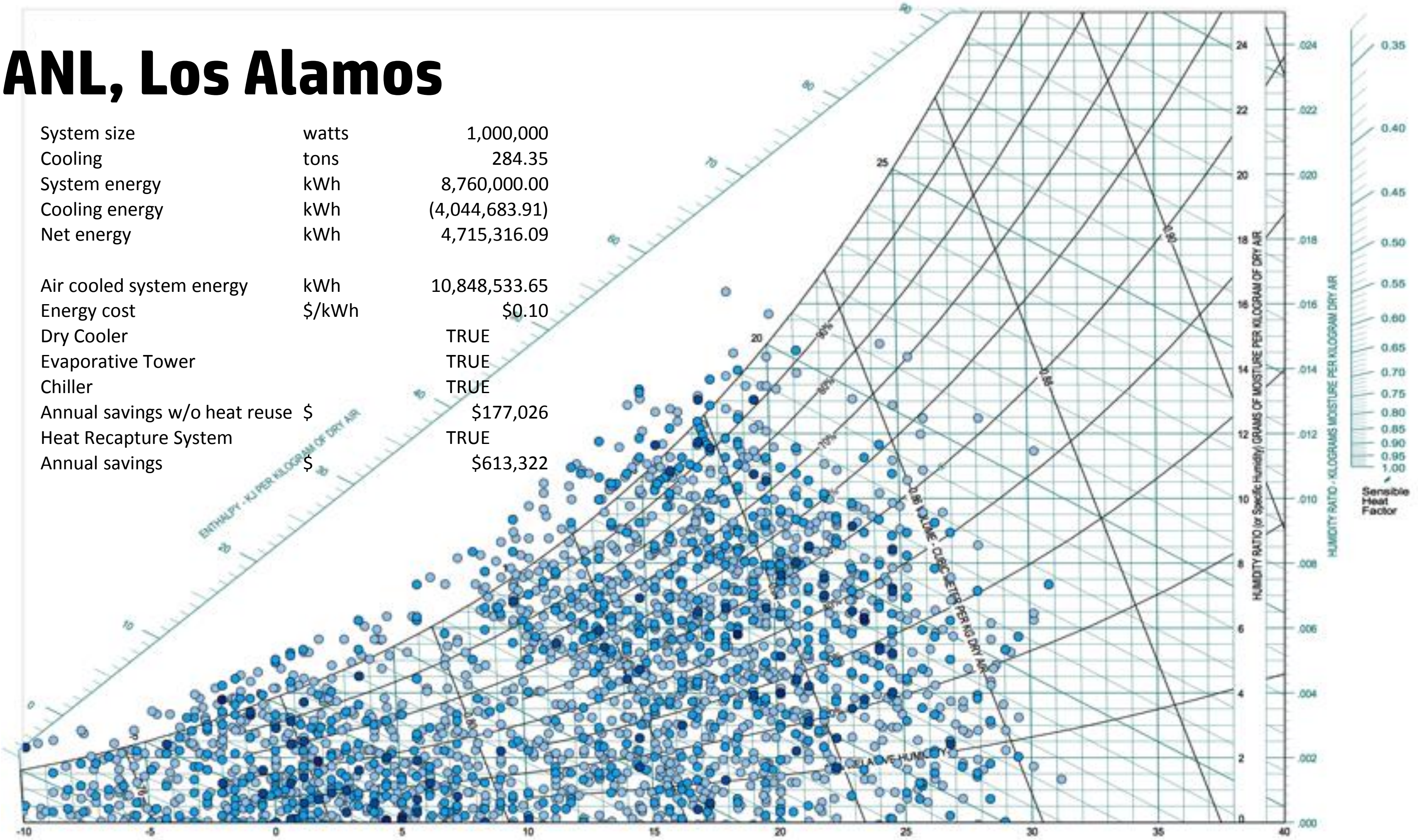
Sandia, Albuquerque

System size	watts	1,000,000
Cooling	tons	284.35
System energy	kWh	8,760,000.00
Cooling energy	kWh	(3,097,025.45)
Net energy	kWh	5,662,974.55
Air cooled system energy	kWh	11,020,261.70
Energy cost	$/kWh	$0.10
Dry Cooler		TRUE
Evaporative Tower		TRUE
Chiller		TRUE
Annual Savings w/o heat reuse		$182,569
Heat Recapture System		TRUE
Annual savings	$	$535,729



LANL, Los Alamos

System size	watts	1,000,000
Cooling	tons	284.35
System energy	kWh	8,760,000.00
Cooling energy	kWh	(4,044,683.91)
Net energy	kWh	4,715,316.09
Air cooled system energy	kWh	10,848,533.65
Energy cost	$/kWh	$0.10
Dry Cooler		TRUE
Evaporative Tower		TRUE
Chiller		TRUE
Annual savings w/o heat reuse	$	$177,026
Heat Recapture System		TRUE
Annual savings	$	$613,322



University of Tromsø in Norway

Forget cooling! Use the server room to heat the campus

International research hub focuses on global environmental issues, up close

- Increasing research demands, # of advanced calculations
- Energy consumption/sq. meter went up dramatically, 2 megawatts with plans for more
- Building new 400 sq. meter data center
- Expect to reduce 80% of energy costs for computer operation, saving 1.5M krone in operating budget/year

> “. . . the idea is to reduce electricity costs by sharing them with the rest of the university or other stakeholders heating.”
> -Svenn A. Hanssen , Head of IT department at the University of Tromsø

World’s largest supercomputer dedicated to advancing renewable energy research

- **$1 million in annual energy savings** and cost avoidance through efficiency improvements
- Petascale (one million billion calculations/ second)
- **6-fold increase** in modeling and simulation capabilities
- Average PUE of **1.06 or better**
- **Source of heat** for ESIF’s 185,000 square feet of office and lab spaces, as well as the walkways
- 1MW of data center power in under 1,000 sq. ft., **very energy-dense** configuration

2

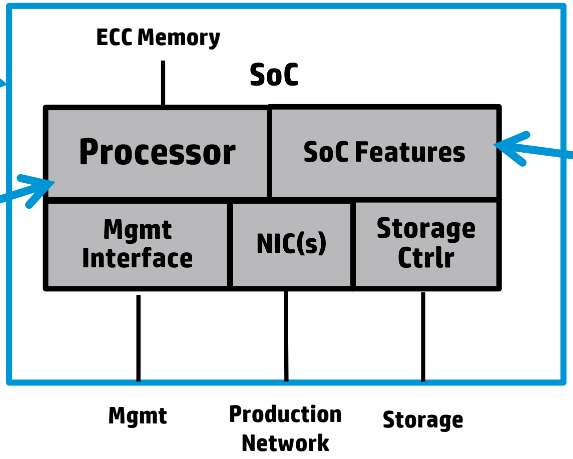
Server SoC & Application Specific Compute

Server SoCs Bringing Disruptive Value

TCO Reduction

Integration:
80% reduction in motherboard and chipset costs

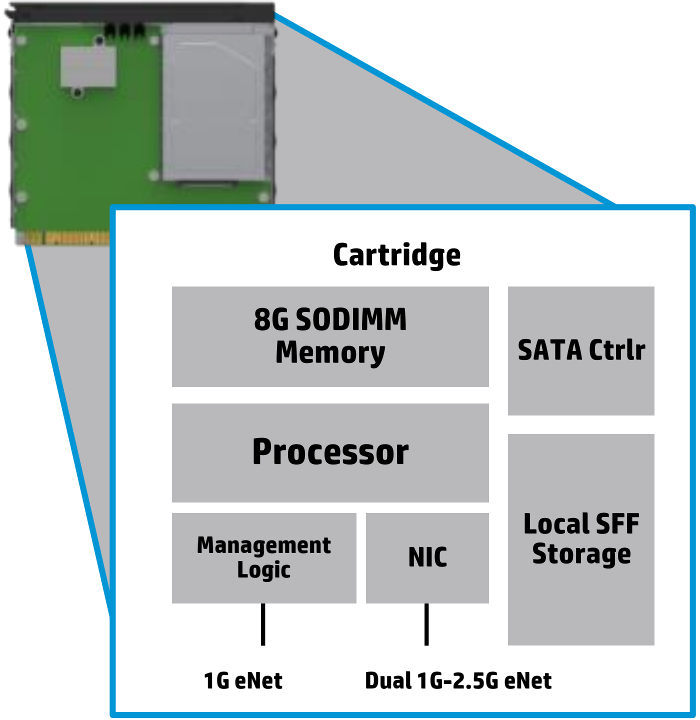
CPU design:
8 core CPU has enough compute capacity to host 95th percentile of virtual machine instances.



Value Creation

Accelerators:
Graphics Engine
Video Transcoding
H.265 4K
DSP
VoIP, Imaging
Network Processors
Packet processing,
FPGA
Pattern matching,
Math

Flexibility in cartridge design



Cartridge

8G SODIMM Memory

SATA Ctrlr

Processor

Management Logic

NIC

Local SFF Storage

1G eNet

Dual 1G-2.5G eNet

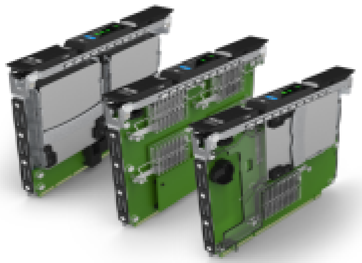
- **Complete server** on every cartridge
- **45 servers** per chassis
- **450 servers** per rack
- Example: dedicated server for hosting

“Moonshot” up close (front view)

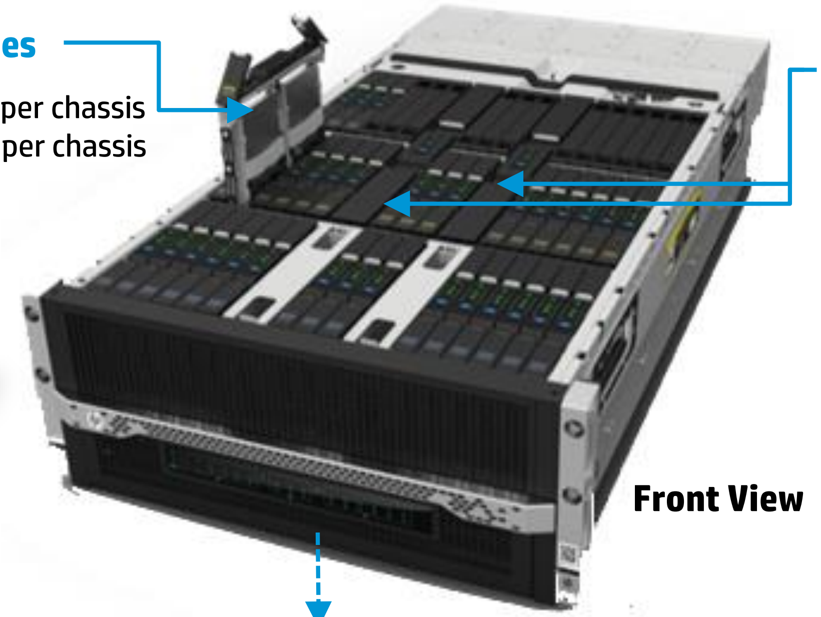
Delivering on the promise of extreme low energy computing

Top-loaded, hot plug cartridges

- Quad-Node cartridge =180 nodes per chassis
- Single-Node cartridge = 45 nodes per chassis

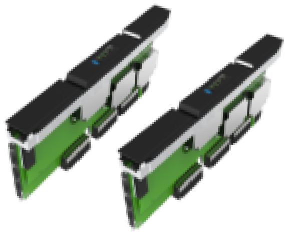


Compute, Storage, or
Both, x86 and ARM



Integrated A & B Switches

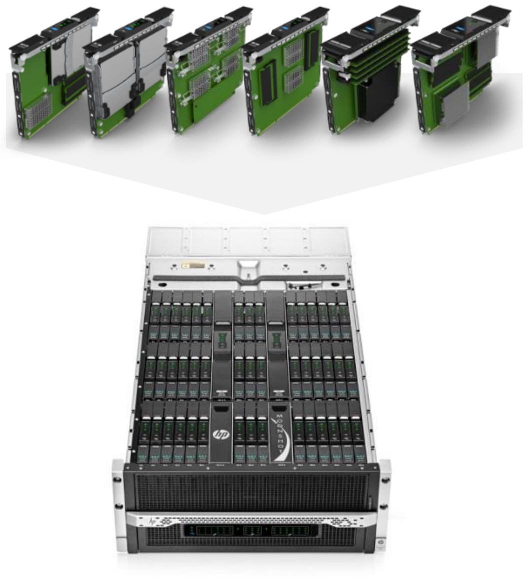
- 180x10G downlinks
- 6 x10G Stackable Uplinks



Front View

SL-APM and iPDU rack-level management

Special purpose cores



77% less costly**

89% less energy*

80% less space*

97% less complex*

HP Moonshot is the first step

* Based on HP internal analysis of HP Moonshot with ProLiant Moonshot Server Cartridges.

** Based on HP internal estimates of total cost to operate HP Moonshot with ProLiant Moonshot Server Cartridges as compared to traditional servers.

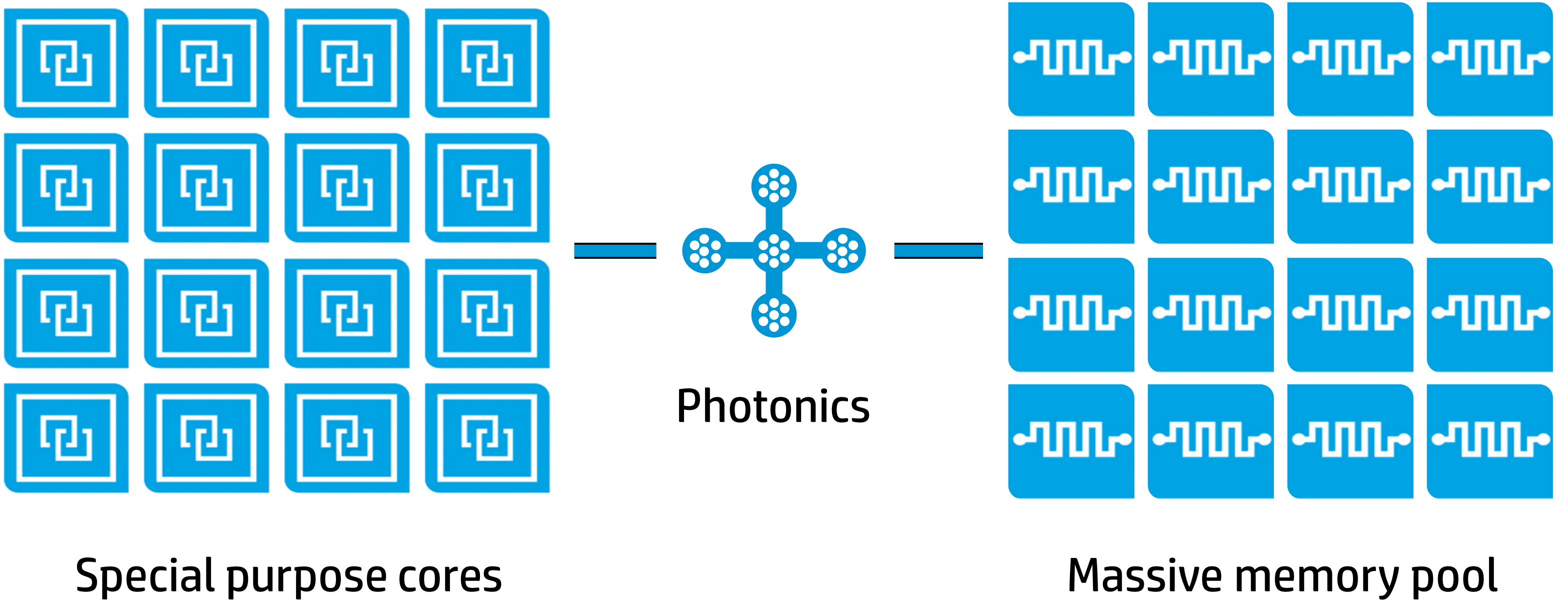
BitCoin - An Application Specific Compute Example

Orders of Magnitude Improvement in Short Timeframe (YMMV)

	X86 2011	GPU 2012	FPGA 2013	ASIC 2014
Million Hashes Per Second	7.5 1X	198 26X	800 105X	146,000 19,500X
Million Hashes per Joule	0.10 1X	1.3 13X	17.5 178X	913 9,300X

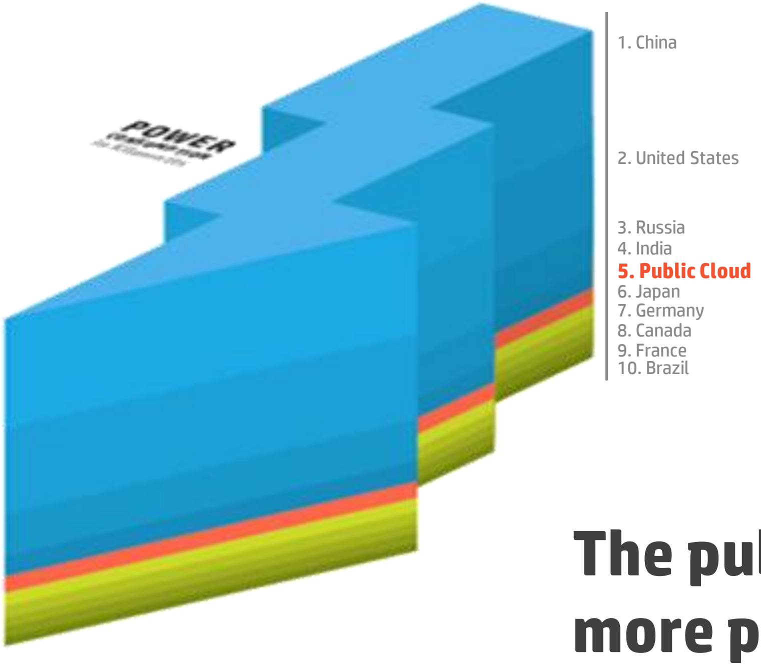
Source: Various Sampling of BitCoin Mining Hardware Performance
https://en.bitcoin.it/wiki/Mining_hardware_comparison

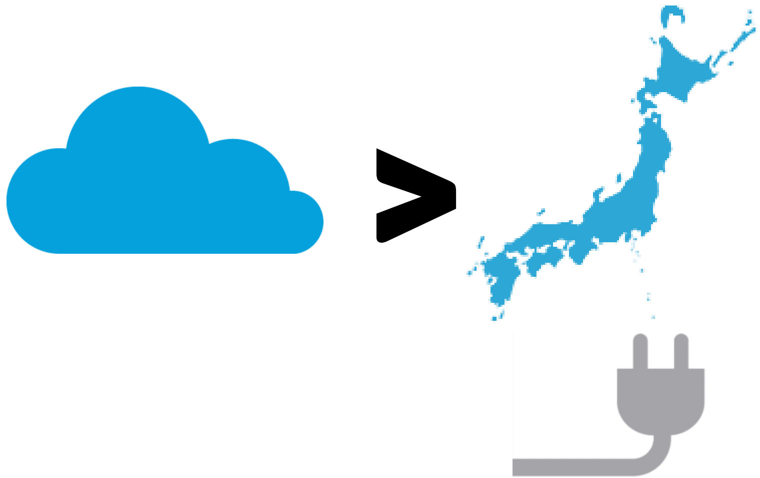
3 The Machine



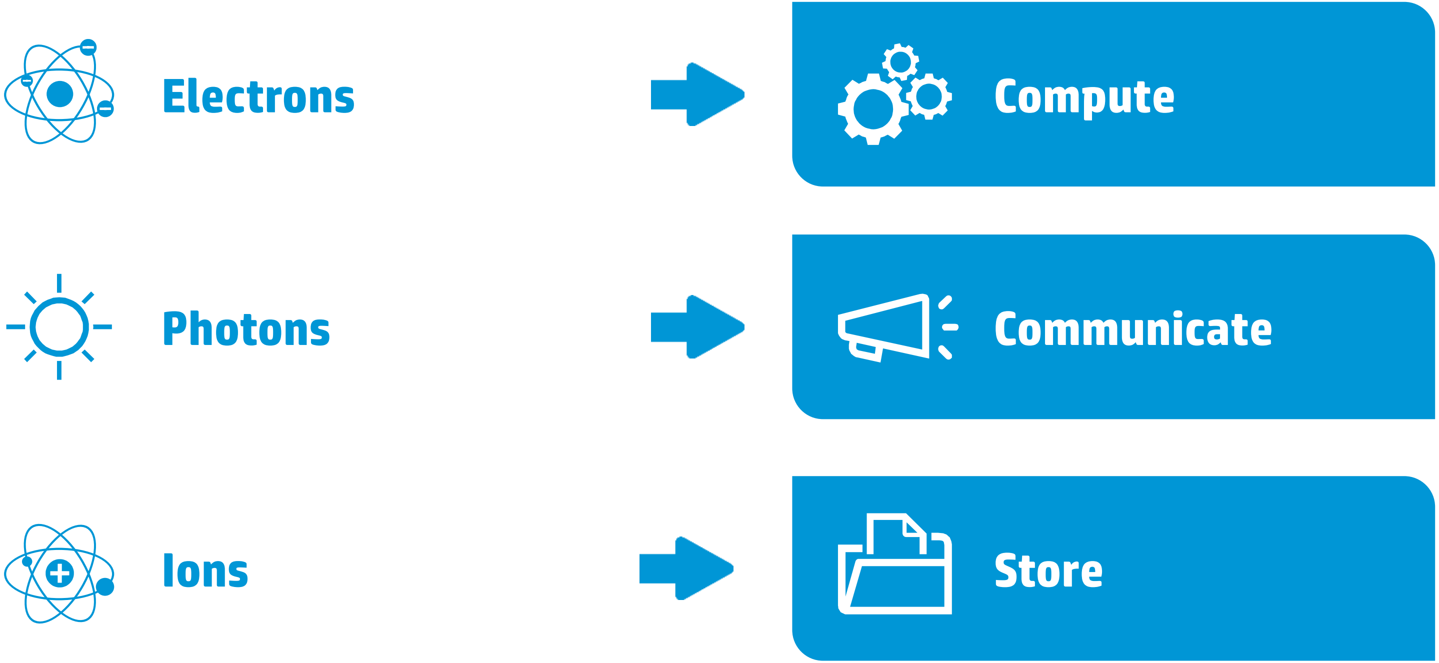
The Machine

Exascale Challenge: 1,000X Compute @ 2X Energy





The public cloud uses more power than Japan

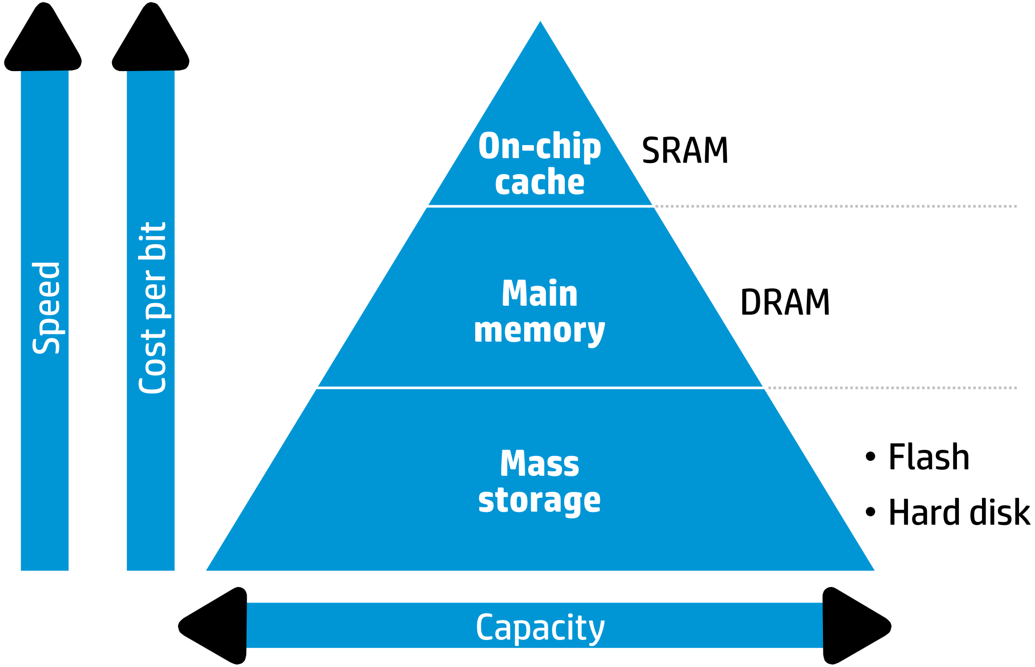


Electrons → Compute

Photons → Communicate

Ions → Store

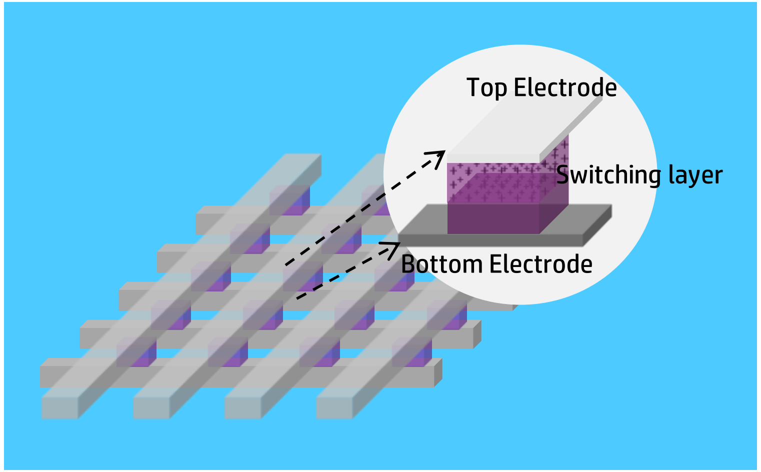
Universal Memory

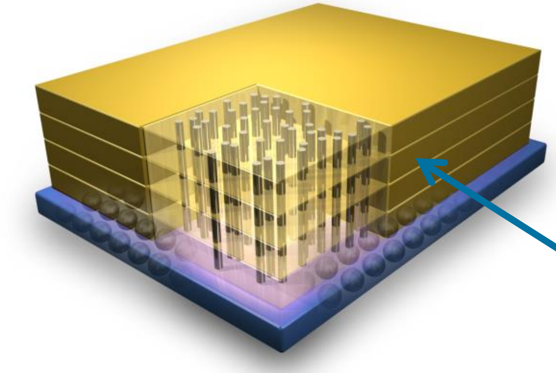


Universal memory obsoletes this hierarchy

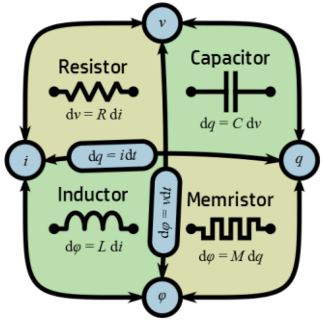
Universal Memory

HP Memristor





Through-silicon-via technology for hard drive like densities



Memristor Attributes

- DRAM-like performance
- Extremely dense
- Stackable in-die
- Very low-power
- Thermals good for die stacking
- Silicon fab friendly

Photonics and Fabrics

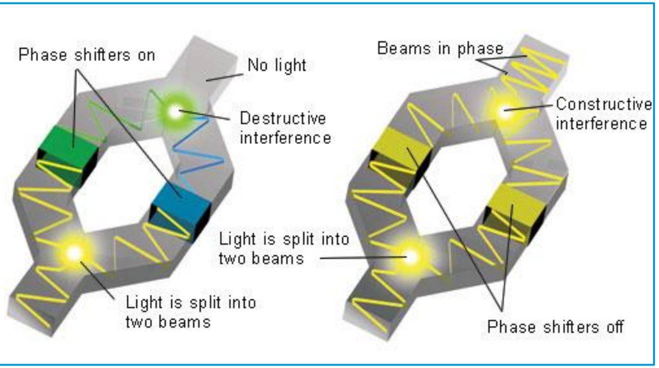
Photonics destroys distance

Photonics

This is about power consumption and application efficiency

3

Modulator

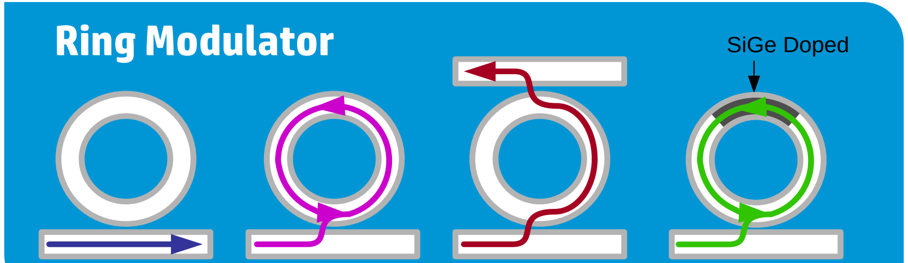


Phase shifters on
No light
Destructive interference
Light is split into two beams
Light is split into two beams
Beams in phase
Constructive interference
Phase shifters off

Industry investments in photonics

- Semiconductor lasers
- Light routing channels on silicon
- Light modulation by electrical signal
- Light path switching by electrical signal

Ring Modulator



Why photonics?

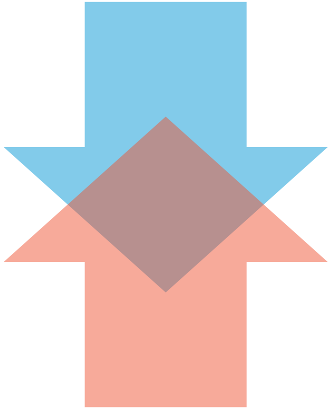
- High-bandwidth at extremely low power
- Distance matters little
- Compute subsystems can be redistributed for maximum space & thermal efficiency

No one fabric to rule them all

Fabrics optimized for node count and workload

	Processing Nodes
SoC/Blade/Cartridge	2 – 20
Zone/Chassis	20 – 200
Rack	200 – 2,000
Row	2,000 – 20,000
Datacenter(s)	20,000 – 200,000

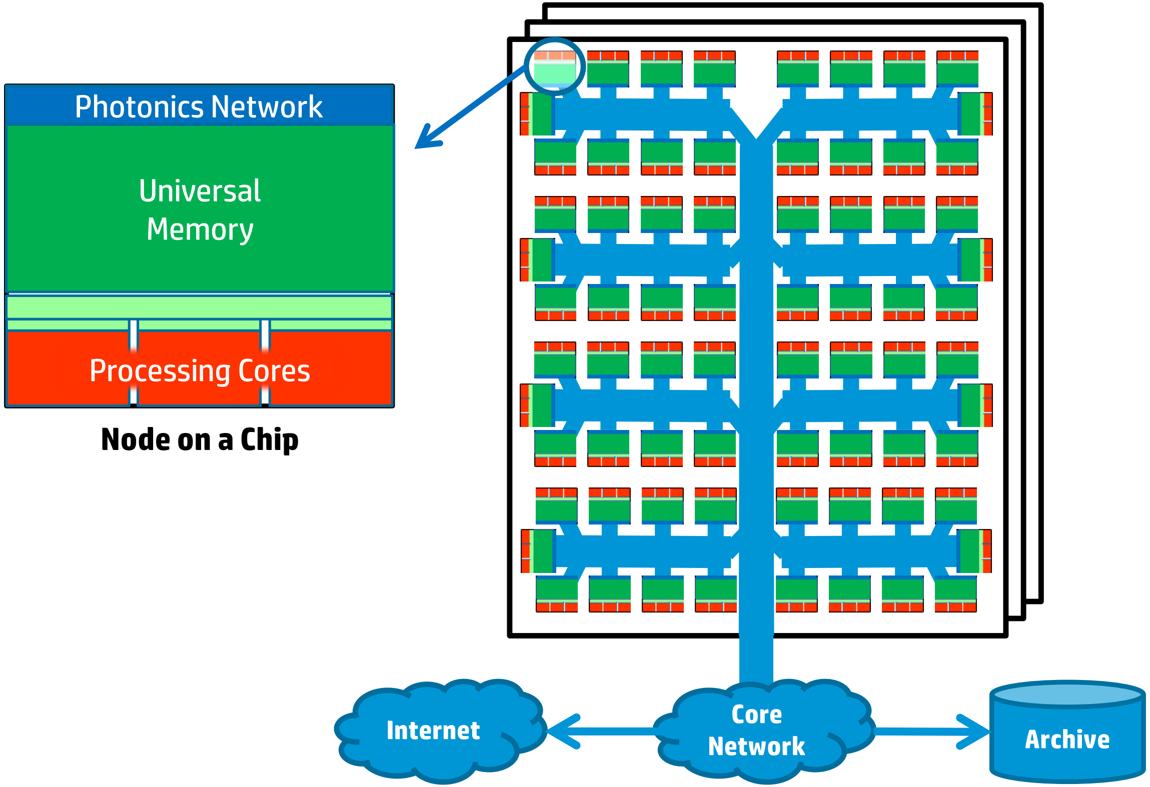
Memory Semantics

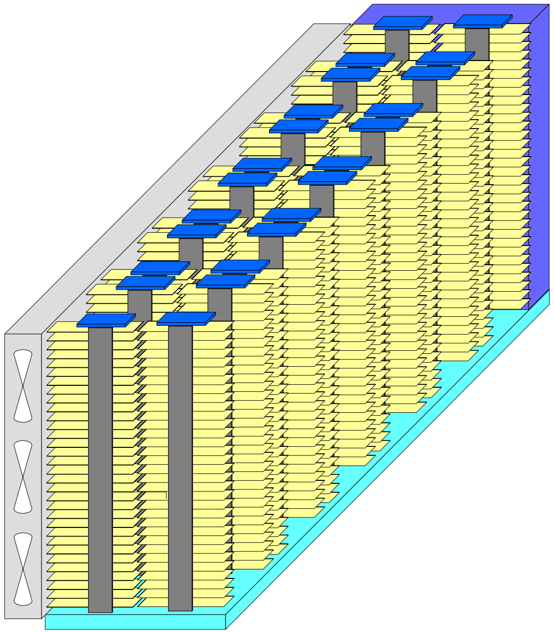


Socket Semantics
Ethernet (802.x)

Technologies Working Together

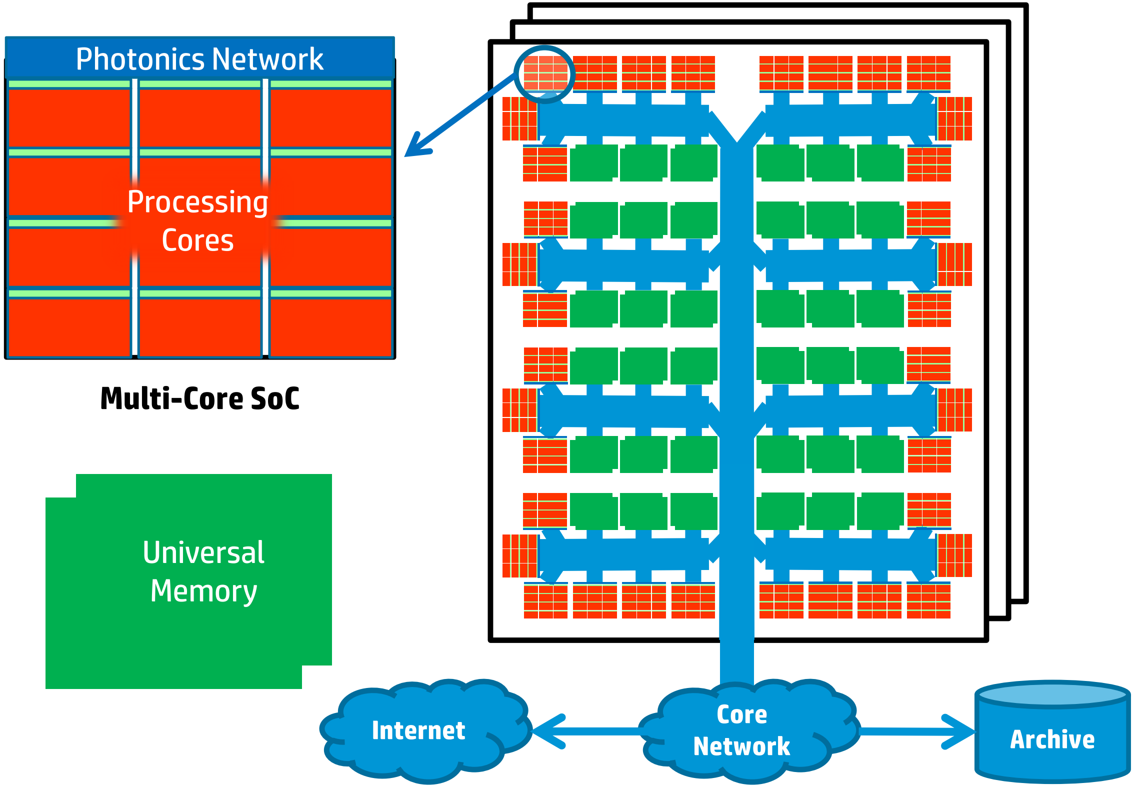
Example 1: Massive Shared-Nothing Compute Farm





Technologies Working Together

Example 2: Data-Centric HPC System



Photonics Network

Processing Cores

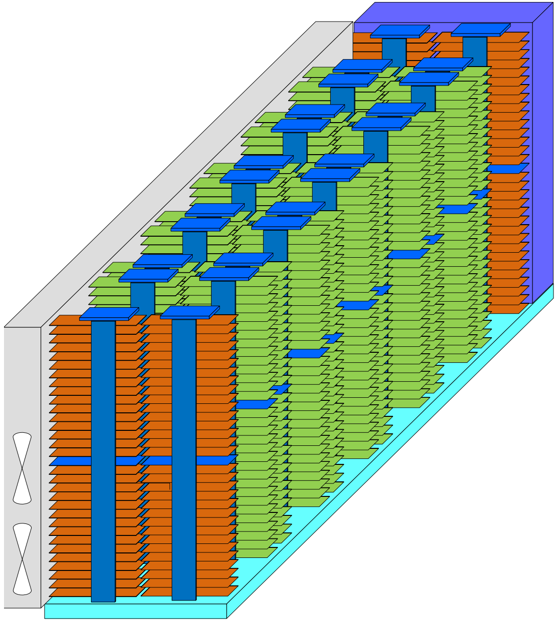
Multi-Core SoC

Universal Memory

Internet

Core Network

Archive



Thank you

Thank you