Standards-based Peta-scale Systems: Trends, Implementations and Solutions.

Dr. Frank Baetke
HP ISS/SCI Global Technology Programs

HPC Cetraro, June 21st, 2010

HP = HPC Leadership for a Changing World

Accelerating innovation through a converged infrastructure for the HPC data center

Performance

– Leadership performance, capacity and density, first in performance/watt

Efficiency

– Increased datacenter capacity and efficiency, accelerating time to deployment

Agility

– Simplified provisioning and management of scalable pools of standardized HPC resources

Confidence

– Industry quality leadership, successful customers, ongoing investments

Purpose-built HPC Servers

PURPOSE DRIVEN SCALE-OUT PRODUCT LINES

Density optimized for the data center

Shared infrastructure for accelerated service delivery

Extreme scale out datacenters with lean management

	DL	BL	SL
Design center	Rack	Blade enclosure in rack	Rack
Design focus	Versatility & value	Integrated & optimized, maximum redundancy	Cost & features optimized for extreme scale out
Application	General purpose	General purpose / private cloud / scale out	Web 2.0 / cloud / scale out
Management	Essential and advanced management HP Insight Dynamics	Advanced management-accelerated service delivery & change in minutes	Home grown management Basic management via IPMI/DCMI

The Most Successful Architecture Ever to Enter the TOP500

The Most Successful Architecture Ever to Enter the TOP500 – the BL-Series (c-Class)

New Performance/Density for HPC: HP ProLiant BL2x220c G6

BL2x220c G6	
Processor	Two 80W or 60W dual- or quad-core Intel Xeon 5500 Series processors per server node*
Memory	Registered or Unbuffered DDR3 6 DIMM Sockets per server 96GB max per server
Internal Storage	1 Non-Hot Plug SFF SATA HDD per server
Networking	2 integrated 1GbE Ethernet ports per server
Mezzanine Slots	1 PCIe Gen2 x8 mezzanine expansion slot per server
Additional Features	Internal USB 2.0 connector Optional internal SD Card slot (consumes the USB slot)
Management	ProLiant Onboard Administrator powered by iLO2
Density	32 server nodes in 10U enclosure

* 95W processors available through the SCI Private-plus process

SL-Series: HP PROLIANT SL6000

Ideal environments

#1 perf/watt

SPECpower_ssj2008
3106*

HP ProLiant SL160z G6	HP ProLiant SL165z G7	HP ProLiant SL170z G6	HP ProLiant SL2x170z G6
Maximum expansion with 18 DIMM slots and up to 2 PCIe slots	Maximum expansion with 12-core AMD processors and 24 DIMM slots	Maximum storage capacity with up to 6 LFF SATA or SAS hard drives	Maximum compute density with two servers per tray (1U)
Ideal Application	**Ideal Application**	**Ideal Application**	**Ideal Application**
HPC database tier Web memory-cache	HPC database tier Web memory-cache	Web Search Web database	HPC compute intensive Web front end

* Based on April 2010 published benchmarks. 12/11/07 SPEC announces the release of SPECpower_ssj2008, the first industry-standard SPEC benchmark that evaluates the power and performance characteristics of volume server class computers. The competitive benchmark results stated herein reflect results published on www.spec.org. See http://www.spec.org/power_ssj2008/results/power_ssj2008.html

Purpose-built HPC Storage

Complementary Scalable Storage Solutions for High Performance Computing

NEW!

X9000 Network Storage System

- Scalable performance and capacity
  - Scalable aggregate bandwidth
  - Scalable metadata, ideal for small files
- Shared datacenter multipurpose storage
  - Linux and Windows clients
  - NFS & CIFS support
  - Ideal for applications in media, FSI, bioinformatics, web/cloud

DDN Storage with Lustre

- Scalable performance and capacity
  - Scalable single-file bandwidth, with multiple writers
  - demanding bandwidth requirements
- Tightly coupled to HPC Linux clusters
- Ideal for parallel applications in traditional HPC

Purpose-built HPC Fabrics

Voltaire IB 4X QDR 36P RAF Managed Switch

What's New:

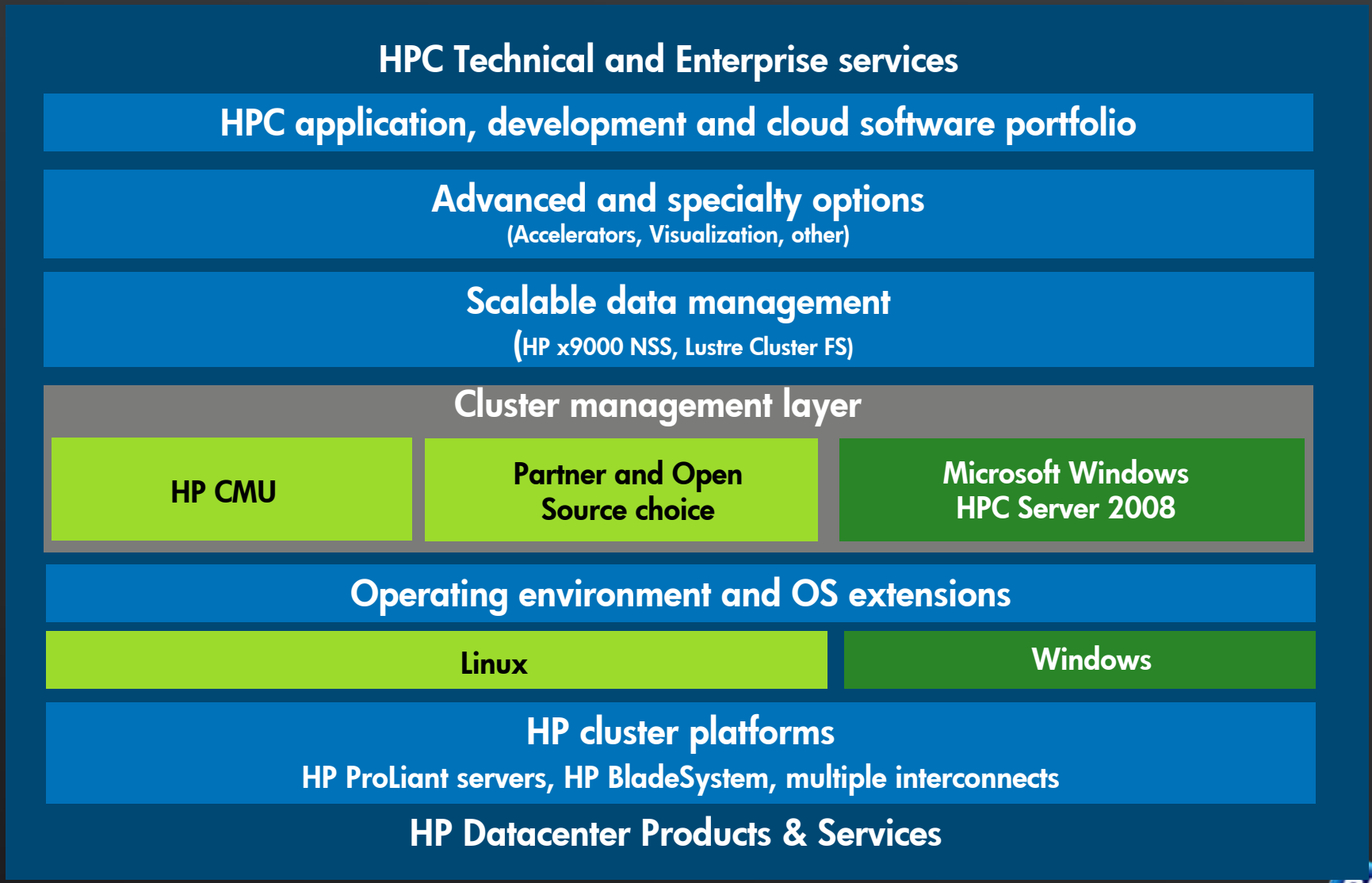
- Voltaire IB 4X QDR 36-port RAF managed switch with a reversed airflow fan unit enabling rear-to-front cooling
- Designed to support front cabling in clusters based on ProLiant SL6000/6500 servers
- 19" rack mountable chassis, 1U height, configurable with redundant Power Supplies, and Fan Units
- Aggregate data throughput: 2.88 Tb/s (QDR), 1440 Gb/s (DDR) or 720Gb/s (SDR)
- Port-to-port Latency: less than 100 nanoseconds latency
- On-board SM for fabrics up to 648 nodes

HP + 3Com – Leadership from Edge to Data Center Core

Data Center	Virtual Connect 	Edge 	End of Row 	DC Core 
Core	Core Routing 	Aggregation 	Campus Core 	Enterprise Core 
Edge	SMB 	Wireless 	Campus 	Edge Routing 	Enterprise 
Security	Edge Firewall 	Data Center 	Threat Management 	Intrusion Prevention 
Management	Network Node Manager Operations Center	Ops. Automation & Orchestration Integrated Management Center		

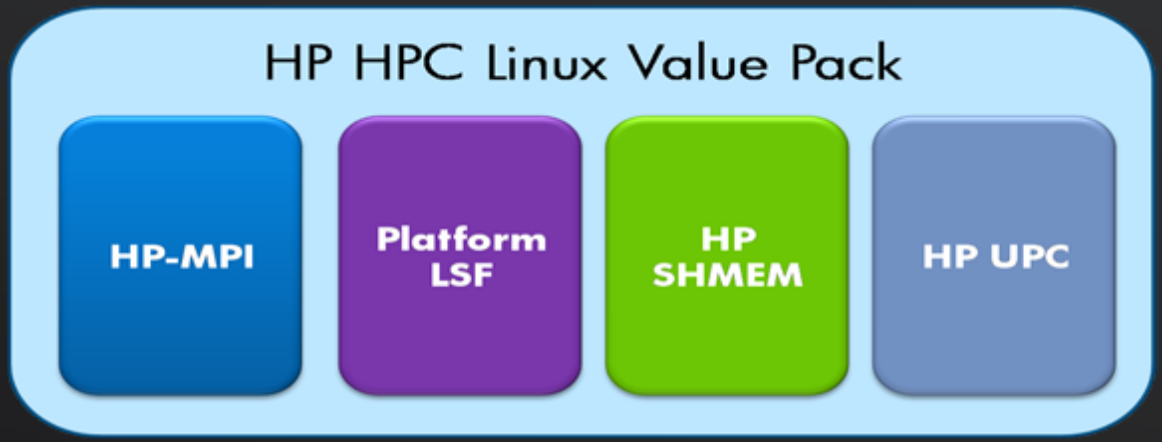
HPC Software Infrastructure

Unified Cluster Portfolio



A la Carte cluster options for HP Clusters

- Operating systems: RHEL, SLES, or customer-supported community distributions; Microsoft Windows HPC Server 2008
- Cluster Management: HP CMU, or third party, via SLMS or customer installed (e.g., ROCKS, Platform Cluster Manager)
- MPI: HP-MPI, or third party/open source; Windows MPI
- Workload manager: Platform LSF (via SLMS now), Altair PBS Pro (HP SKU), Adaptive Computing Moab (via SLMS)



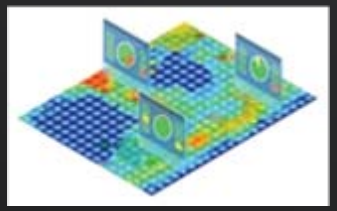
Datacenters – Power and Cooling

Trends at the Data Center: Significantly Improved Efficiencies

NEW!

HP Data Center Smart Grid

- Environmental Edge V5.1
  - 4D view of DC energy
  - Real-time PUE/DCiE metrics
  - Increase DC capacity by up to 25%
- 20’ Performance Optimized Datacenter*
  - Deploy modular datacenters quickly and efficiently, with a lower cost of entry

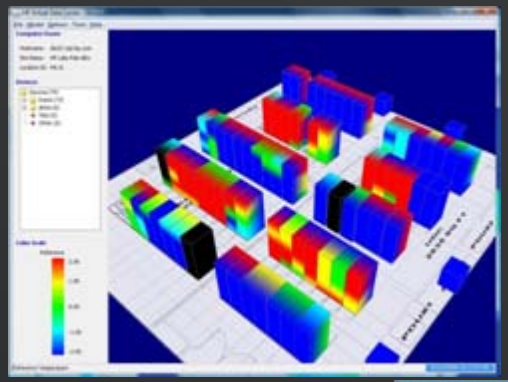


ADAPTIVE AND SCALABLE SOFTWARE FOR HPC Datacenters

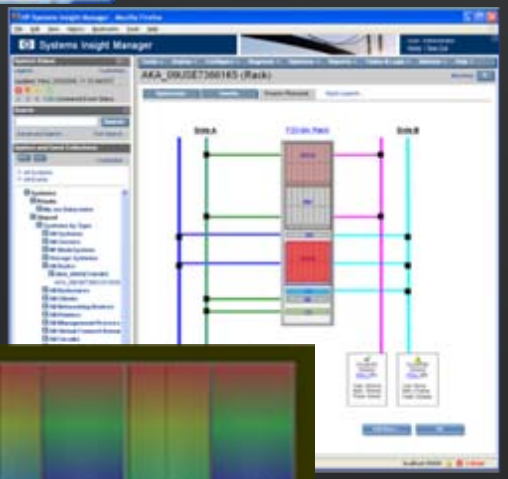
Edge Futures

- Integration with Insight Control
- One pane of glass power and cooling visualziation
- True 3D visualization
- Macro Data Center view
- Micro rack level view

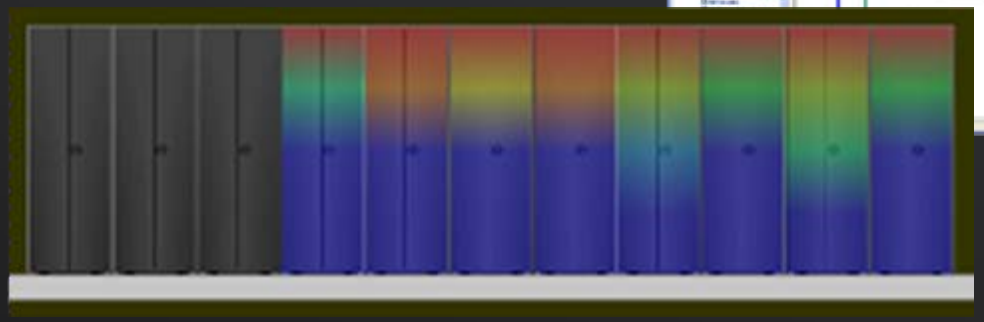
Edge 3D Visualization



IPM Rack View



Edge 2D Front View



Trend?
Are Next Generation Data Centers Ugly?

Advantages: Fast Deployment and Time to Operation: Efficient to Build and Rebuild

- Container backed into truck bay on mfg floor
- Racks assembled and then put into containers
- Truck pulls out with fully-configured container to the customer site…

HP POD products and concepts

Currently shipping - 2010 ***2011***

- 22 50U racks 40ft
- 600kW power capacity
- Designed for high density deployments – max 34kW per rack
- Flexible for redundant or non-redundant deployments

- 10 50U racks 20ft
- Modular design for better supply chain efficiency
- Flexibility to customize

- Rugged exterior
- EMI shielding
- Designed for portability

How Some Data Centers Will Look Like ….

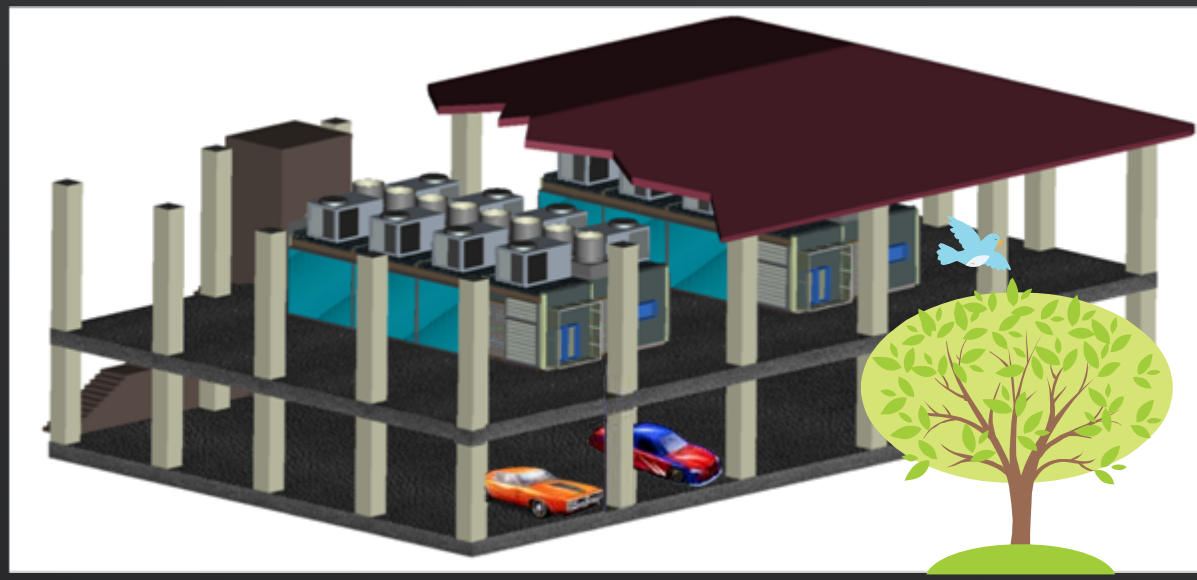
- Is arguably better along many dimensions
  - Scalable
  - Economic at small and large scales
  - Benefits of being co-located with other infrastructure

Illustration Courtesy of Microsoft; used here as a generic illustration of how future facilities might be built.

Holistic Data Center Portfolio: The Greenest 3 PFlop/s on the Planet I

Concept Details:

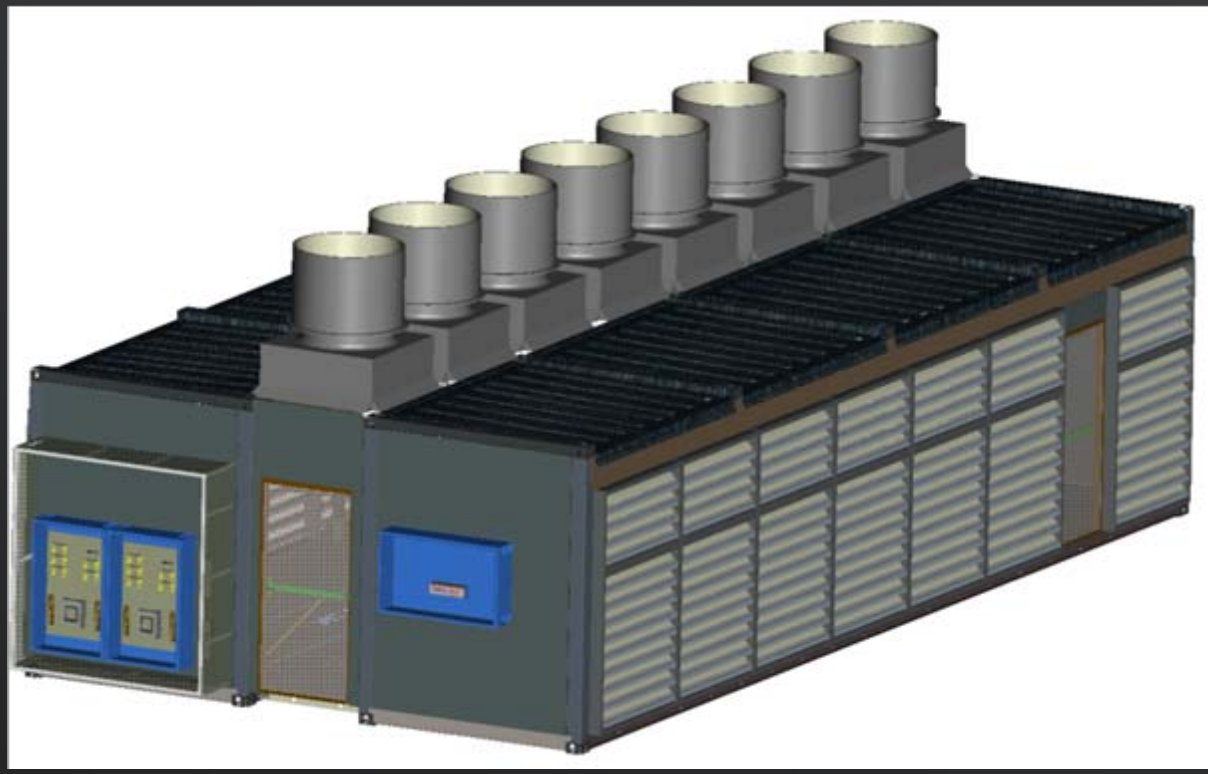
- Two HP Free Air PODs on platform.
- Minimum Area required is about 18 by 24 meters
- Supports 84 50U Racks
- GPU Assisted Processing
- Can distribute up to 2MW of power
- PUE of less than 1.1



Holistic Data Center Portfolio: The Greenest 3 PFlop/s on the Planet II

Concept Details:

- Up to 48" clearance in the front and back of every rack
- 42 50U 2800 lb racks
- Up to 8 225A busses for 32kW to every rack
- Make up fans, controlled dampers, evaporative coolers and filters for best year-round PUE



Trends in Efficiency

	5 yrs. ago	2010	2015
PUE	2, 3, Higher	1.1 Great	?
UPS Efficiency (Part of PUE)	94%	98%+	?
Power Supply Efficiency	75%	94%+	?
Fan Power per 2s Node	60+ W	2-10 W (< 1%) (some think 0)	?

Peta-scale Implementation Example: TITECH Tsubame 2.0

TSUBAME 2.0 Overview

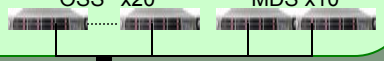
- Compute nodes: 2.4PFlops (CPU+GPU)
  - **New SL-node** >>1408)thin nodes, each with 2 Westmere-EP and 3 NVIDIA M2050
    - 1347 with 54GB and SSD 60GB, 41 with 96GB and SSD 120GB
    - Suse Linux Enterprise Server or Windows HPC Server
  - **DL580 G7** Medium (24) and Fat (10) nodes, with 2 NVIDIA S1070
    - Medium: 128GB plus SSD 120GB x4
    - Fat: 256BG plus SSD 120GB x4
- QDR InfiniBand, full bisection, non-blocking
  - Spine: **Voltaire Grid Director 4700** 12 x 324port
  - Edge: **Voltaire Grid Director 4036** 179 x 36 port and **4036E** 6 x 34port/10GbE 2 port
- Storage: 5.93PB
  - Lustre file system 5.93PB: **DDN SFA 10000** (10/rack, 5 racks) and DL360 G6 (30)
  - Home file system: 1.2PB: **DDN SFA 10000** (10/rack, 1 racks), BlueArc Mercury 100 (2) and DL360 G6 (30)
- Press release (Japanese):
  - http://www.gsic.titech.ac.jp/sites/default/files/pdf/TSUBAME/press.pdf

TSUBAME 2.0 System Overview

Storage system **Total 7.13PB (Lustre+ home)**

Lustre file system (DDN SFA10K) 5.93PB

MDS,OSS
HP DL360 G6 30nodes
Storage
DDN SFA10000 x5
(10 enclosure x5)
Lustre 5File System
OSS: 20 OST: 5.9PB
MDS: 10 MDT: 30TB



Home directory region 1.2PB

Storage Server
HP DL380 G6 4nodes
BlueArc Mercury 100 x2
Storage
DDN SFA10000 x1
10 enclosure x1

Existing system

Tapa System

SupreTitenet

SupreSinet3

Interconnect: Full bi-section non-blocking

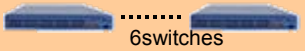
Core Switch

Voltaire Grid Director 4700 12switches
IB QDR: 324port

Edge Switch

Voltaire Grid Director 4036
179switches IB QDR : 36 port

Edge Switch(10GbE port)

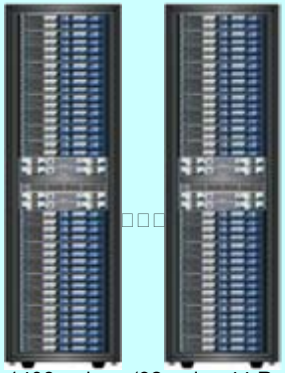


Voltaire Grid Director 4036E 6 switches
IB QDR:34port
10GbE: 2port

Management nodes

Compute nodes **2.4PFlops(CPU+GPU)**

"THIN" nodes



1408 SL nodes
CPU Intel Westmere-EP 2.93GHz
Turbo boost 3.196GHz 12Core/node
Mem: 54GB (1347 nodes)
96GB (41 nodes)
GPU NVIDIA M2050 515GFlops,3GPU/node
SSD 60GB x 2 120GB (54GB nodes)
120GB x 2 240GB (96GB nodes)
OS: Suse Linux Enterprise Server
Windows HPC Server

CPU Total: 215.99TFLOPS(Turbo boost 3.196GHz)
CPU+GPU: 2391.35TFlops
Memory Total 80.55TB
SSD Total 173.88TB

"Med" nodes

DL580 G7 24nodes
CPU Intel Nehalem-EX 2.0GHz
32Core/node
Mem:137GB(=128GiB)
SSD 120GB x 4 480GB
OS: Suse Linux Enterprise Server

CPU Total: 6.14TFLOPS

"Fat" nodes

DL580 G7 10nodes
CPU Intel Nehalem-EX 2.0GHz
32Core/node
Mem:274GB(=256GiB) 8nodes
549GB(=512GiB) 2nodes
SSD 120GB x 4 480GB
OS: Suse Linux Enterprise Server

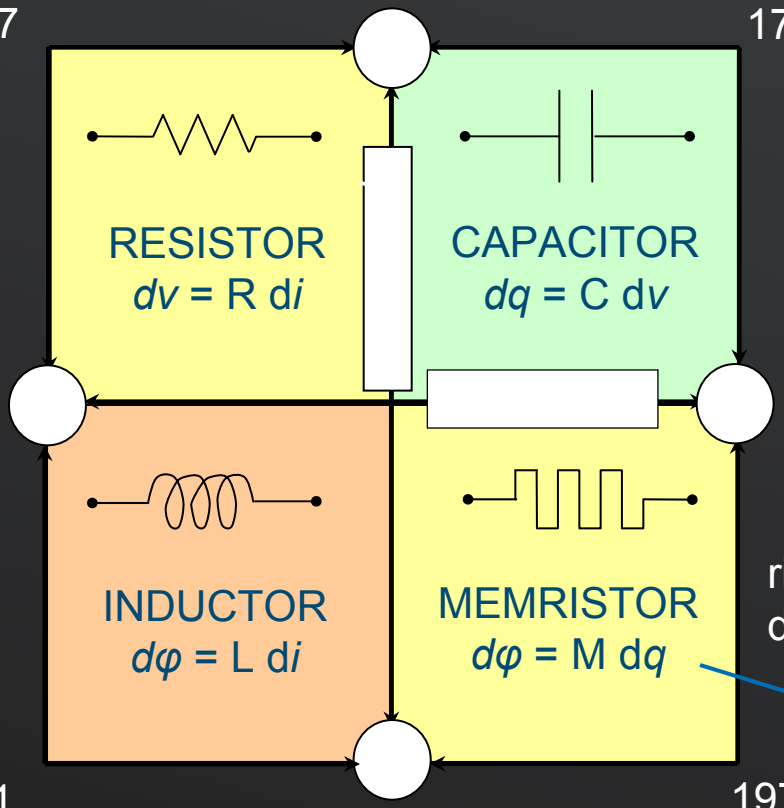
CPU Total: 2.56TFLOPS

PCI –E gen2 x16 x2slot/node

GSIC:NVIDIA Tesla S1070GPU

Research

The Prediction of a New Circuit Element: the Memristor



L. O. Chua, *IEEE Trans. Circuit Theory* **18**, 507 (1971)

rigorous definition

$$v(t) = R[w, i(t)]i(t)$$

Quasi-static conduction eq.-
R depends on state variable *w*

$$\frac{dw(t)}{dt} = f[w, i(t)]$$

Dynamical equation –
Evolution of state in time

First Hybrid CMOS-Memristor Chip

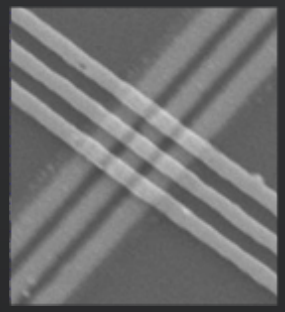
Issues that had to be overcome:

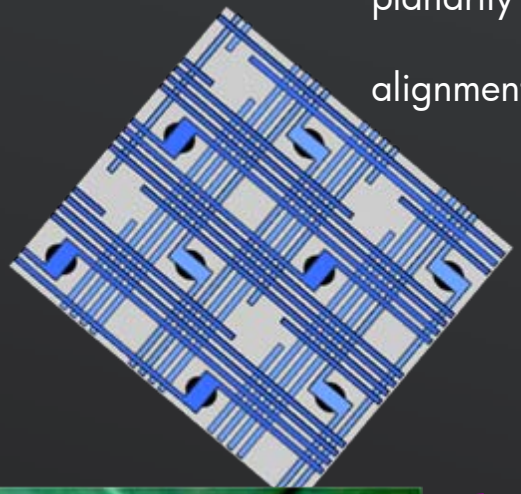
planarity

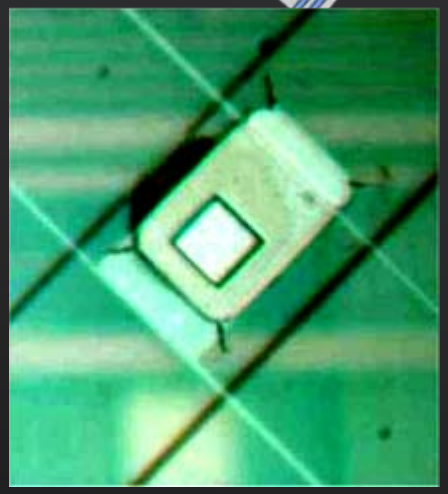
alignment of fine features

3x3 100nm nanowire

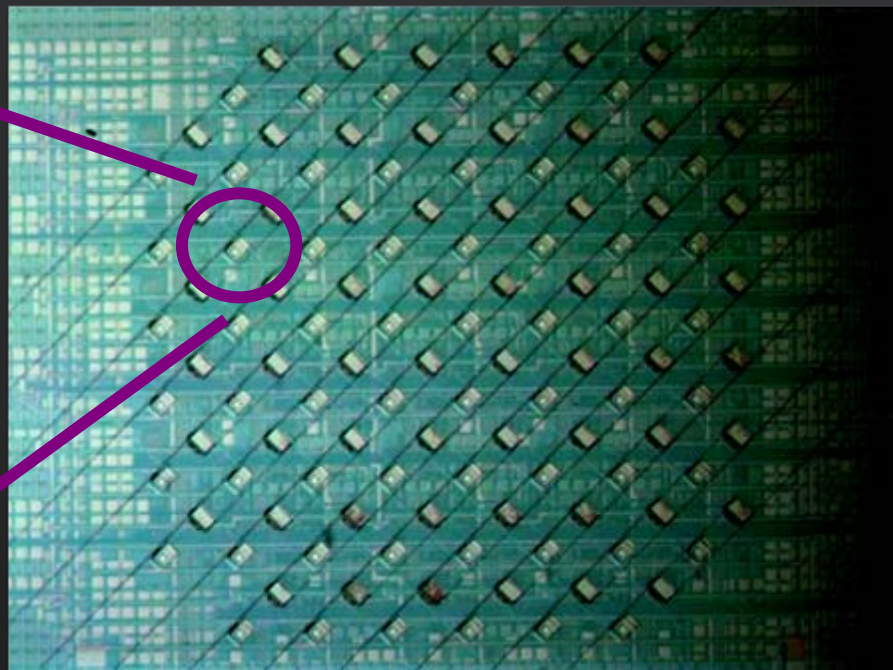
Crossbar junctions







CMOS chip with memristive devices

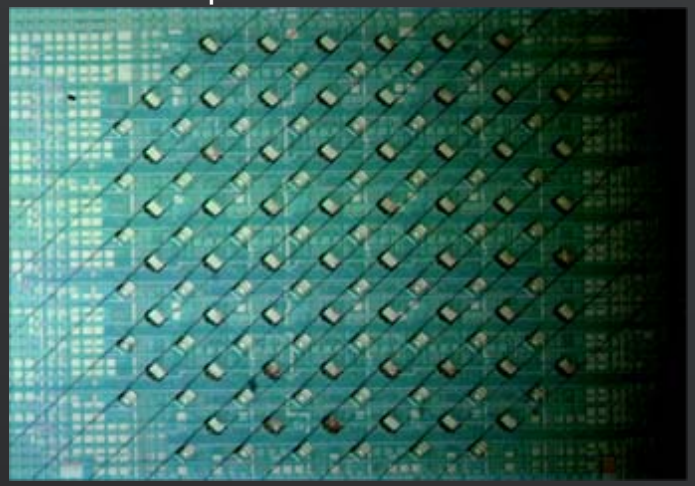


Connecting the CMOS layer with the nanowire crossbar junctions

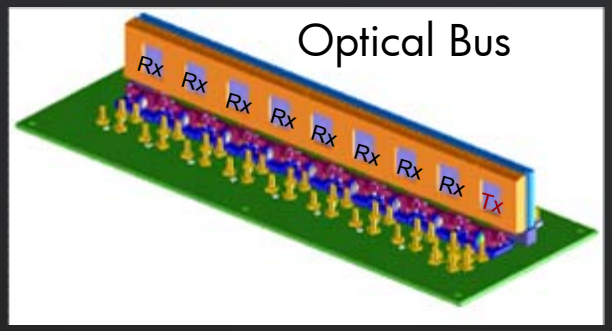
Long-term Trends in HPC: Examples of HP Labs Innovation

- Capacity - Memristor (short for memory resistor)
  - Scales to extremely high density (many terabits/sq cm)
  - Non-volatile – essentially infinite data retention time
  - Reasonably fast (ns) and low energy (pJ)

CMOS chip with memristive devices



- Bandwidth – Photonics
  - High bandwidth, and highly energy efficient
  - Photonic interconnects between systems available now
  - Long term research leading to photonic interconnects within systems and chips



www.hpl.hp.com

Final Trend: Towards Exascale Systems (Looking at MTBF of minutes…)

Leveraging 3D PCRAM Technologies to Reduce Checkpoint Overhead for Future Exascale Systems

Xiangyu Dong[†‡], Naveen Muralimanohar[†], Norm Jouppi[†], Richard Kaufmann[†], Yuan Xie[‡]
[†]Hewlett-Packard Labs, [‡]Pennsylvania State University
[†]Email: {xiangyu.dong,naveen.muralimanohar,norm.jouppi,richard.kaufmann}@hp.com
[‡]Email: {xydong,yuanxie}@cse.psu.edu

ABSTRACT

The scalability of future massively parallel processing (MPP) systems is being severely challenged by high failure rates. Current hard disk drive (HDD) checkpointing results in overhead of 25% or more at the petascale. With a direct correlation between checkpoint frequencies and node counts, novel techniques that can take more frequent checkpoints with minimum overhead are critical to implement a reliable exascale system. In this work, we leverage the upcoming *Phase-Change Random Access Memory* (PCRAM) technology and propose a hybrid local/global checkpointing mechanism.

After a thorough analysis of MPP systems failure rates and fail-

the failure rate growth in future systems.

To tolerate the rising failure rate and reduce its impact on workload running time, modern MPP systems are equipped with a centralized non-volatile storage system (typically built with arrays of disks) that takes frequent synchronized checkpoints of every node in the system. However, the current approach has many serious limitations. First, the design of using a single centralized medium storing all checkpoints is inherently not scalable; second, as the number of compute nodes increases and the size of applications grow, the performance overhead of conventional techniques can reach an unacceptable level. A recent study by Oldfield *et al.* [3] showed a 1-petaFLOPS system can potentially take more than 50% performance hits because of frequent checkpointing operations. There-

http://www.cs.utah.edu/~naveen/papers/sc09.pdf

Attend HP-CAST Hamburg, May 28-29 Worldwide User Group Conference

HP-CAST

HP Consortium for Advanced Scientific and Technical Computing
Word-Wide User Group Meeting
Scalable Computing Infrastructure (ISS/SCI) Organization
InterContinental Hotel, Fontenay 10, 20354 Hamburg, Germany
May 28th – 29th, 2010

HP-CAST 14

World-wide User Group Conference with Participation of
NTIG (Nordic Technical Interest Group) & HP-CAST IBÉRICA
Draft Agenda V2.1p

Thursday, May 27th – Registration & Get-Together

17:00 – 22:00	***Registration***	
19:00 – 22:00	***HP-CAST Welcome Reception***	***All Attendees***

Friday, May 28th – Conference

HP-CAST – History and Outlook

HP-CAST 1 Dallas (first one after HP/Compaq & Board merger)

HP-CAST 2 Brisbane, Australia

HP-CAST 3 Pittsburgh @SC04 (new "light" structure introduced)

HP-CAST 4 Krakow, Poland

HP-CAST 5 Seattle @SC05 (new Board elected)

HP-CAST 6 Seoul, Korea

HP-CAST 7 Tampa @SC06 (first HP-CAST "light" with tutorials)

HP-CAST 8 Karlsruhe, Germany

HP-CAST 9 Reno @SC07

HP-CAST 10 Singapore

HP-CAST 11 Austin @SC08

HP-CAST 12 Madrid, Spain

HP-CAST 13 Portland @SC09

HP-CAST 14 Hamburg @ISC10

HP-CAST 15 New Orleans @SC10 - November 12 and 13 !

Thank You