Advances in HP Servers with Integrated NVIDIA GPUs

NVIDIA GPU Technology Workshop, Singapore

Ed Turkel

Group Manager, HPC Segment Product Management

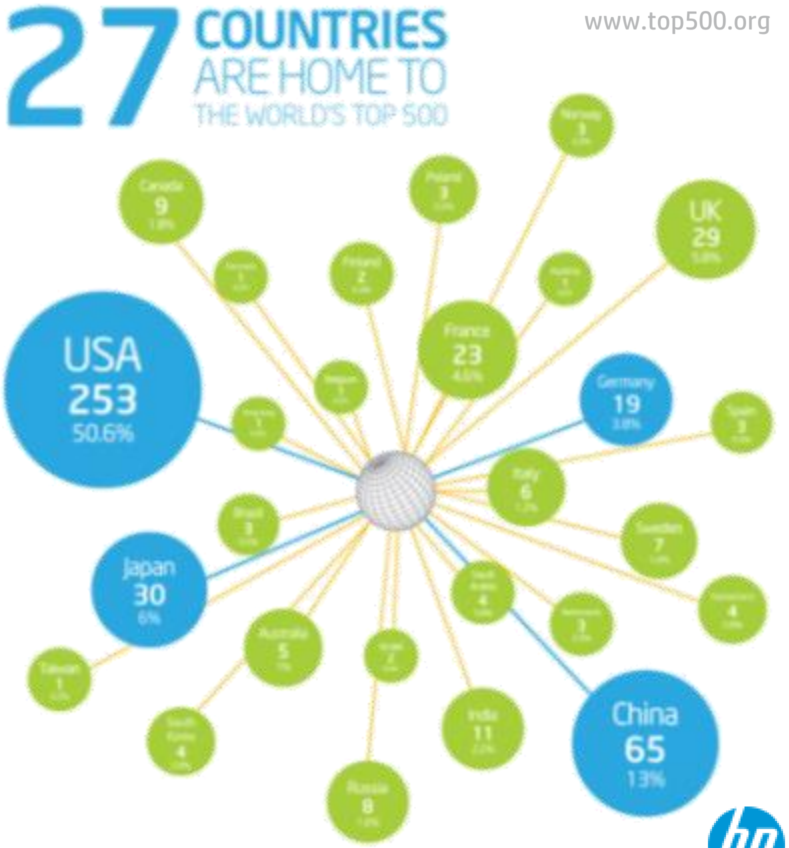
July 8, 2014

To out-compute is to out-compete

Why High Performance Computing is so important

- Firmly linked to **economic competitiveness** as well as scientific advances
  - **97% of companies** that had adopted supercomputing said they **could no longer compete or survive** without it

- Worldwide political leaders increasingly recognize this trend
  - Enables not only enterprise but also **national competitiveness**



www.top500.org

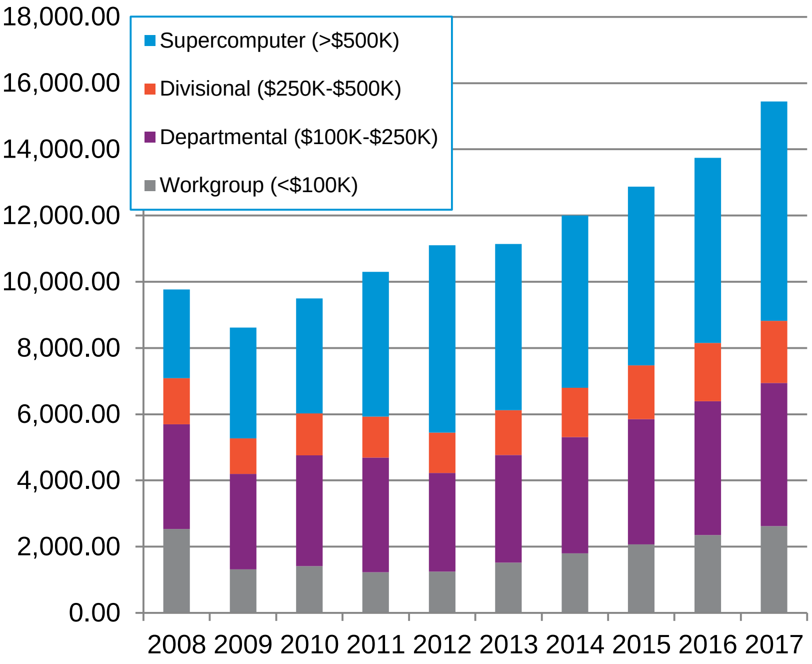
Top trends in High Performance Computing

The global economy in HPC is growing

- Compound annual growth rate (CAGR) of 7.3% over the 2013-2017 forecast period with revenues to exceed $15 billion in 2017

Major challenges

- Constantly growing demand for compute performance
- Power, cooling, real estate, system management
- Storage and data management continue to grow in importance
- Software hurdles continue to grow
- The worldwide petascale race is at full speed
- Big data and accelerators are hot relatively new technologies



Source: IDC 2014

Solving global problems requires greater…

- Computer-Aided Engineering
- Electronic Design Automation

- Research & Development
- Life Sciences
- Pharmaceutical

- Geophysical Sciences
- Energy Research & Production
- Meteorological Sciences

- Entertainment
- Media Production
- Visualization & Rendering

- Government
- Academia

- Financial Services

Performance

Efficiency

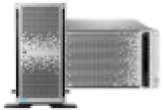
Accessibility

HP Servers with Integrated NVIDIA GPUS

Workload-optimized portfolio for better business outcomes

For core business applications

HP MicroServer


HP ProLiant ML

HP ProLiant DL

Intelligence to increase productivity

For mission-critical environments

HP ProLiant scale-up

"DragonHawk"

HP Integrity blades & Superdome

HP Integrity NonStop

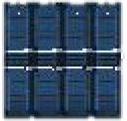
Availability to function in real-time

Common modular compute architecture

For Big Data, **HPC,** and web scalability

HP ProLiant SL

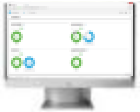
HP Moonshot


HP Apollo

Density and efficiency to scale rapidly

For virtualized and cloud workloads

HP BladeSystem


HP OneView

Convergence to accelerate IT service delivery

Global support and services | Best-in-class partnerships | Converged solutions

Breakthrough performance for blazing fast results

NEW

NEW HP ProLiant DL580 Gen8 Server

30x faster

transaction processing

4S Processor	Memory	I/O Expansion	Smart Array	Internal Storage
Intel® Xeon® E7-4800/8800 v2	3TB* max memory (6TB later)	9 PCI-e Gen3	12Gbps SAS	10 SFF Drives
2X	1.5X	2.7X	2X	1.2X

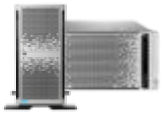
* Up to 6TB post-launch with 64GB DIMMs

Optimized for acceleration (K6000, K40c)

Workload-optimized portfolio for better business outcomes

For core business applications

HP MicroServer


HP ProLiant ML

HP ProLiant DL

Intelligence to increase productivity

For mission-critical environments

HP ProLiant scale-up

"DragonHawk"

HP Integrity blades & Superdome

HP Integrity NonStop

Availability to function in real-time

Common modular compute architecture

For Big Data, **HPC,** and web scalability

HP ProLiant SL

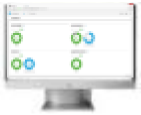
HP Moonshot

HP Apollo

Density and efficiency to scale rapidly

For virtualized and cloud workloads

HP BladeSystem

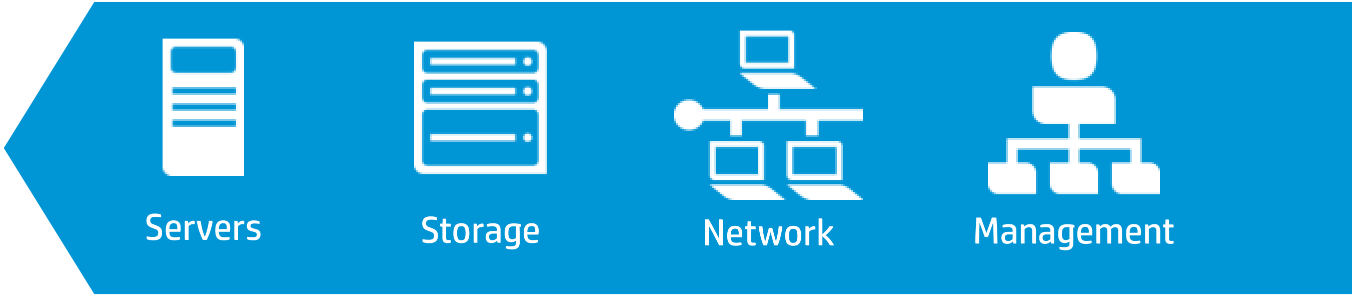

HP OneView

Convergence to accelerate IT service delivery

Global support and services | Best-in-class partnerships | Converged solutions

HP Ultimate Converged Infrastructure

A complete HPC cluster in a single blade enclosure



HP ProLiant WS460c Gen8 Graphics Server Blade

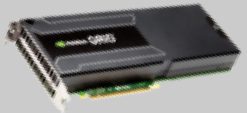
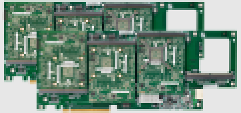
Built from the world's leading server blade BL460c Gen8, and enhanced with high-performance professional graphics accelerators, HP ProLiant WS460c Gen8 Graphics Server Blade offers the ideal balance of performance, scalability, and graphics functionality, to make it the gold standard for Client Virtualization platform

Key workloads include:

- Graphics accelerated Virtual Desktop Infrastructure (VDI) hosting
- Graphics accelerated shared application session hosting
- Dedicated remote workstation for 3D graphics design & analysis
- Natural resource exploration and analysis
- Multi-display remote desktop server for financial services

Broad GPU performance range and density

Best matching of graphics for different user needs and cost requirements

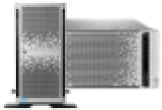
Performance	Card/GPU
Ultra High-end	NVIDIA GRID K2 (2 GPU), Quadro K6000, K5000 
High-end	NVIDIA K4000, 6 x Quadro K3100M (HP MultiGPU) 
Mid/Entry	8 x Quadro 1000M (HP MultiGPU)
	NVIDIA GRID K1 (4 GPU)

Note: NVIDIA GRID GPU and HP MultiGPU graphics available only with Intel Xeon E5-2600 v2 (Ivybridge) processors

Workload-optimized portfolio for better business outcomes

For core business applications

HP MicroServer


HP ProLiant ML

HP ProLiant DL

Intelligence to increase productivity

For mission-critical environments

HP ProLiant scale-up

“DragonHawk”

HP Integrity blades & Superdome

HP Integrity NonStop

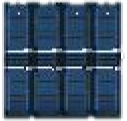
Availability to function in real-time

Common modular compute architecture

For Big Data, **HPC,** and web scalability

HP ProLiant SL

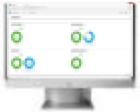
HP Moonshot


HP Apollo

Density and efficiency to scale rapidly

For virtualized and cloud workloads

HP BladeSystem


HP OneView

Convergence to accelerate IT service delivery

Global support and services | Best-in-class partnerships | Converged solutions

Engineered to accelerate innovation

The HP ProLiant SL6500 Scalable System

Scalable performance

- Engineered for massive scale

Maximum efficiency

- Efficient to power, operate and maintain

Operational agility

- Fast adoption, faster time to results

Designed for power and space efficiency to reduce both capital expense and operational expense when deploying systems at scale

Over 225 Tflops
performance in 1 rack

provision 1,000 nodes
less than 30 min.

Cluster arrival to production
in DAYS not months

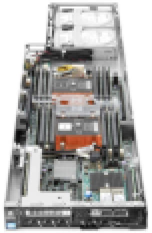
Simple and efficient for highly scalable systems

- More performance per watt and per square foot
- Shared, efficient hot-plug fans
- Shared high-efficiency power supplies
- Optional redundant fans/power supplies
- Less sheet metal and mechanical components

SL230s

CPU Compute optimized

up to 160 CPUs per rack

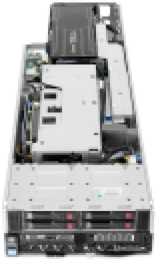


SL250s

Balanced CPU/GPU performance

up to 3 GPUs per server

up to 80 CPUs plus 120 GPUs per rack



SL270s

GPU computing optimized

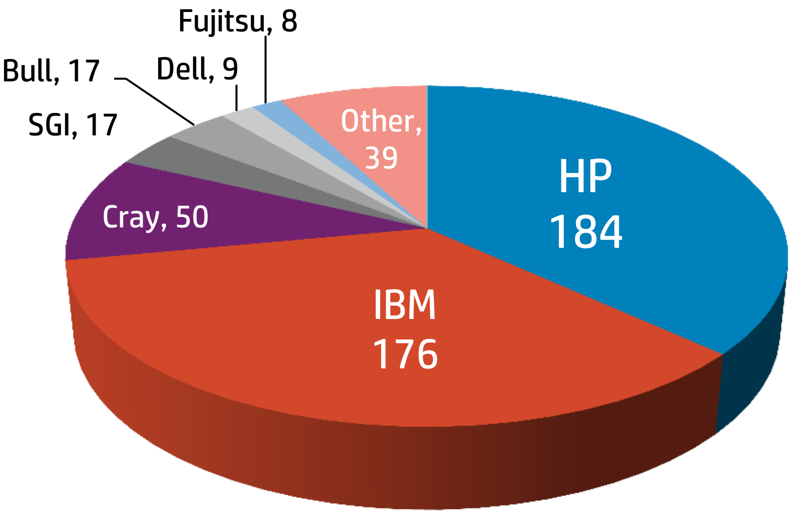
up to 8 GPUs per server

up to 40 CPUs plus 160 GPUs per rack

HP #1 on the TOP500 list

Multiple system with integrated NVIDIA Tesla GPUs

June'14 TOP500 Systems



Tokyo Institute of Technology – "Tsubame 2.5"

- 1408 HP ProLiant SL390s G7 servers, each with three NVIDIA Tesla K20x GPUs, recently upgraded from NVIDIA Tesla M2050 GPUs.
- #13 on the Jun'14 TOP500 list and #8 on the Nov'13 Green500 list, with 5.6PF peak performance and 2.8PF Linpack Rmax, over double the performance of the prior system

Clemson University – "Palmetto 2"

- HP ProLiant SL250s Gen8 servers, each with two NVIDIA Tesla K20 GPUs
- #66 on the Jun'14 TOP500 list, with 739GF peak performance and 551GF Linpack Rmax

University of Southern California – "HPCC"

- HP ProLiant SL250s Gen8 servers, each with two NVIDIA Tesla K20 GPUs
- #71 on the Jun'14 TOP500 list, with 690GF peak performance and 532GF Linpack Rmax

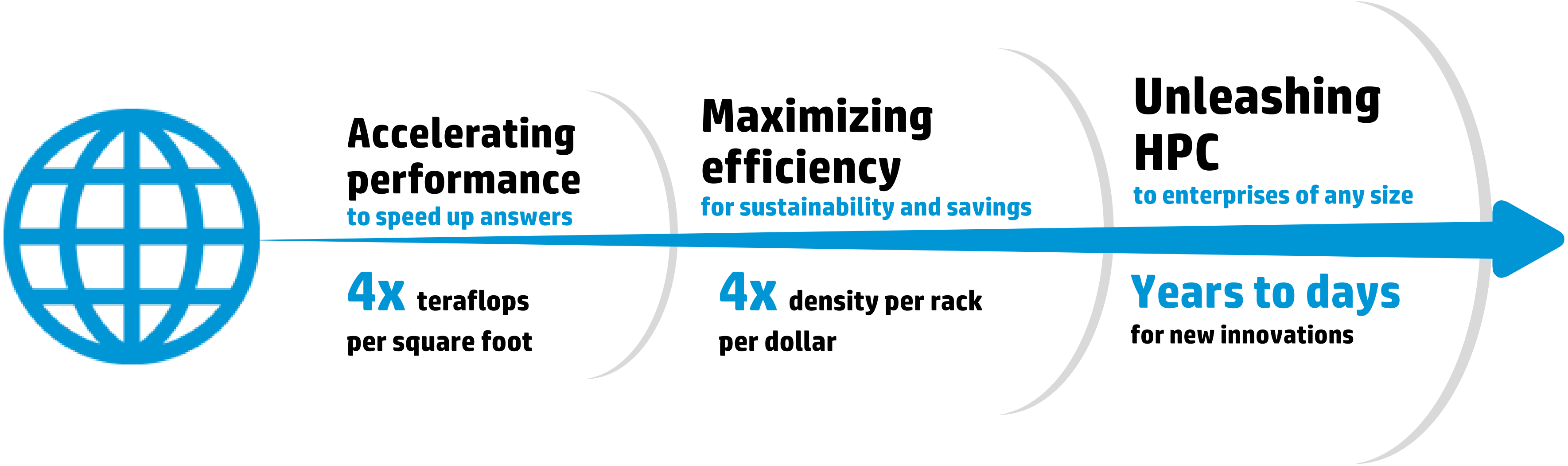
Introducing

HP Apollo System

NEW

Telling a compelling story

Reinventing HPC today to accelerate the world of tomorrow



Accelerating performance
to speed up answers

4x teraflops per square foot

Maximizing efficiency
for sustainability and savings

4x density per rack per dollar

Unleashing HPC
to enterprises of any size

Years to days for new innovations

Introducing

HP Apollo family

High performance computing at rack scale

Maximizing data center efficiencies

NEW

HP Apollo 6000 System

35%

greater **performance** for EDA

$3 Million

savings per 1000 servers over 3 years

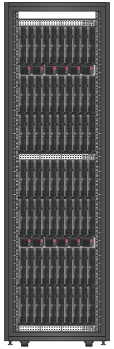
4x

density per rack per dollar

Designed for single threaded HPC workloads such as design automation or financial service risk analysis

The New HP Apollo 6000 System

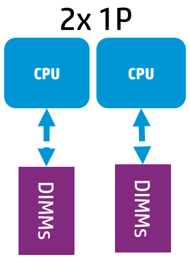
Rack-level shared infrastructure for efficiency and flexibility



Rack scale

- 160 nodes per 48U rack
- 5U chassis (1.0m deep rack)
- 20 nodes per enclosure
- Front service, rear cabled

High performance computing



- Highest frequency per core
- Intel E3-12xx v3 Haswell
  - CPU core generation ahead
- Single-threaded applications
- Max turbo frequency of 4GHz
- Low latency: No 2P cache coherency

First available tray

- ProLiant XL220a Server dual-server tray
- Front serviceable
- Rear cabled solution
- Max power of ~169W per tray
- **2p and 2p+GPUs trays coming soon!**

Shared power & cooling

- Efficient pooled power shelf supports up to 6 chassis
- N, N+1, 2N redundancy configs
- 12 volts DC output with max power of 15.9kW
- Advanced Power Manager

Differentiated: Power shelf, Advanced Power Manager

Rack scale shared infrastructure to get the best performance for your budget

Efficient pooled power

- Power shelf supports up to 6 chassis for rack-level efficiency
- 15.9kW capacity with N, N+1, N+N redundancy

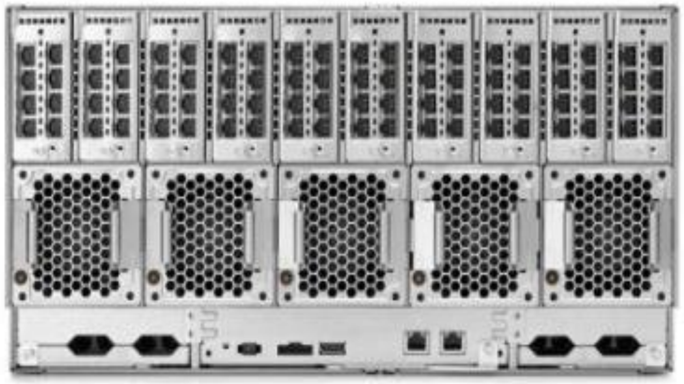
Advanced Power Manager

- **See and manage** shared infrastructure, server, chassis and rack-level power from a single console
- **Simplify, and save** >80% by avoiding spend on serial concentrators, adaptors, cables and switches
- **Flex to meet workload demands** with dynamic power allocation and capping

Differentiated: *Innovation Zone* flexes to meet workload needs

Increase TCO savings with the right connectivity

- The Innovation Zone allows for 2 FlexibleLOMs per tray: InfiniBand, 10GbE, 4-port 1GbE
- Mix and match: Independent I/O modules can be configured differently
- Modify as workload needs change, with flexible inputs and outputs



NEW

Turbo-charging performance to accelerate results

HP Apollo 8000 System

4x

faster molecular simulations

$1 Million

(up to) **energy savings** over 5 years

3,800 tons

removed of CO2 per year

Advancing the science of supercomputing

Apollo 8000 System Technologies

Advancing the science of supercomputing

Dry-disconnect servers

- 100% water cooled components
- Designed for serviceability

Warm water

- Closed secondary loop in CDU
- Isolated and open facility loop

Intelligent Cooling Distribution Unit

- 320 KW power capacity
- Integrated controls with active-active failover

Open door view of 4 compute & redundant CDU racks

Management infrastructure

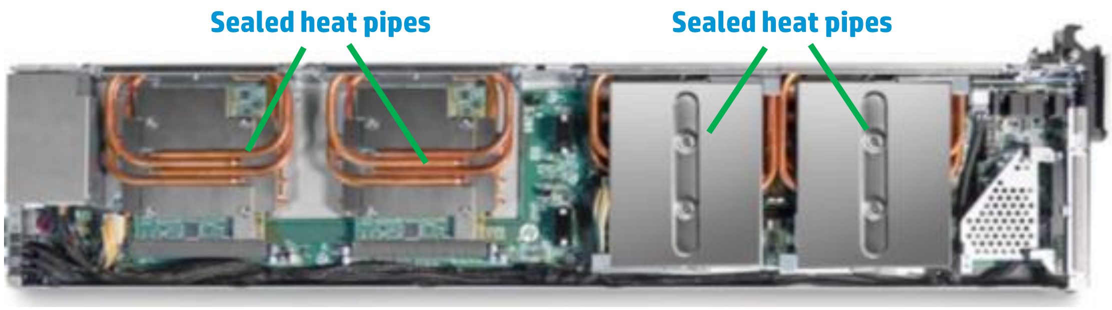
- HP iLO4, IPMI 2.0 and DCMI 1.0
- Rack-level Advanced Power Manager

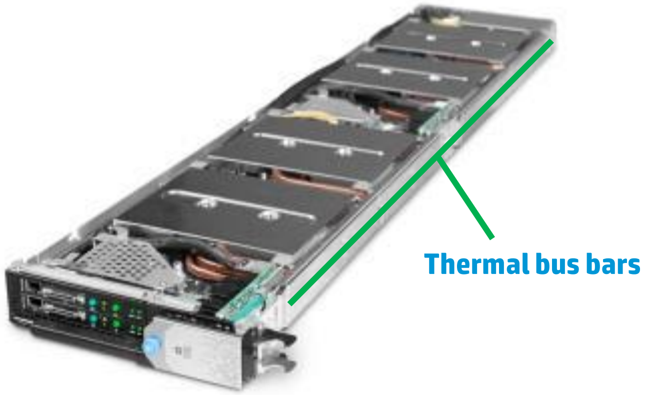
Power infrastructure

- Up to 80kW per rack
- Four 30A 3-phase 380-480VAC

Differentiated: Dry-disconnect servers

New patented technology making a liquid-cooled system as easy to service as air-cooled





- Enables maintenance of servers without breaking a water connection
- Inside the server tray, heat is transferred from components via vapor in **sealed heat pipes**
- **Thermal bus bars** on the side of the compute tray transfer heat to the water wall in the rack
- Water flows through thermal bus bar in the rack from supply-and-return pipes
- Fluid fully contained under vacuum

http://youtu.be/9Ih3R84Corg

Failure is not an option

Efficient liquid cooling without the risk

- **Dry-disconnect servers**: sealed heat pipes cool components
- Facility water **isolated** from IT loop
  - Takes ASHRAE spec water
- Secondary IT loop **vacuum** keeps water in place
- Intelligent Cooling Distribution Unit designed to minimize and **isolate** issues
- Comprehensive **system insight** and management built on Advanced Power Management and smart sensors

World's largest supercomputer dedicated to advancing renewable energy research

National Renewable Energy Laboratory

- **$1 million in annual energy savings**
- 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 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

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

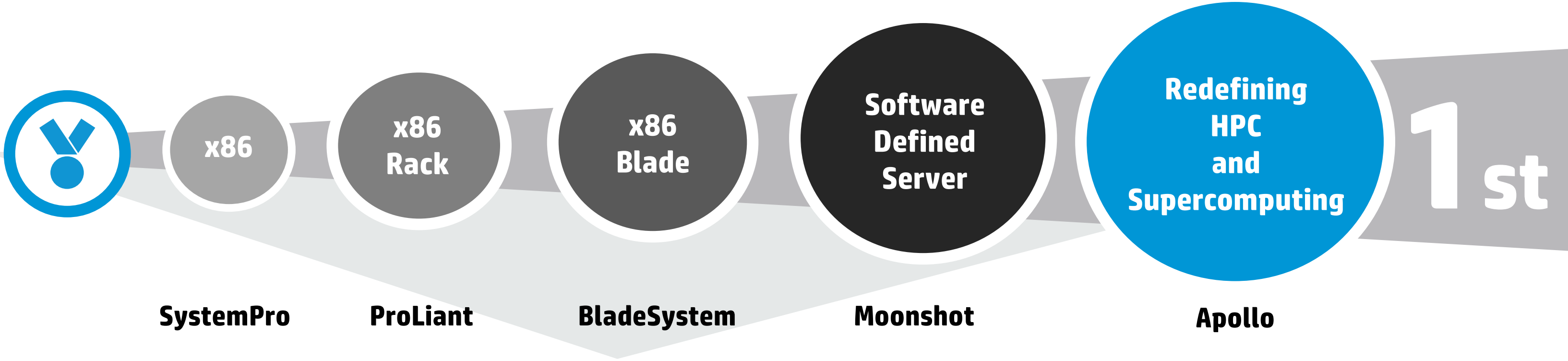
- Increasing research demands, # of advanced calculations
- Energy consumption/sq. meter increased dramatically, 2 megawatts with plans for more
- Building new 400 sq. meter data center
- Expect to reduce 80% of energy costs for computer operation, saving 1.5M krone in operating budget/year

"… 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

History of HP innovations with proven leadership

Defining new markets and business opportunities

Proven HP engineering



x86

x86 Rack

x86 Blade

Software Defined Server

Redefining HPC and Supercomputing

1st

SystemPro

ProLiant

BladeSystem

Moonshot

Apollo

#1 HPC market

NEW

Reinventing HPC today to accelerate the world of tomorrow

Accelerating performance
to speed up answers

4x teraflops
per square foot

Maximizing efficiency
for sustainability and savings

4x density per
rack per dollar

Unleashing HPC
to enterprises of any size

Years to days
for new innovations

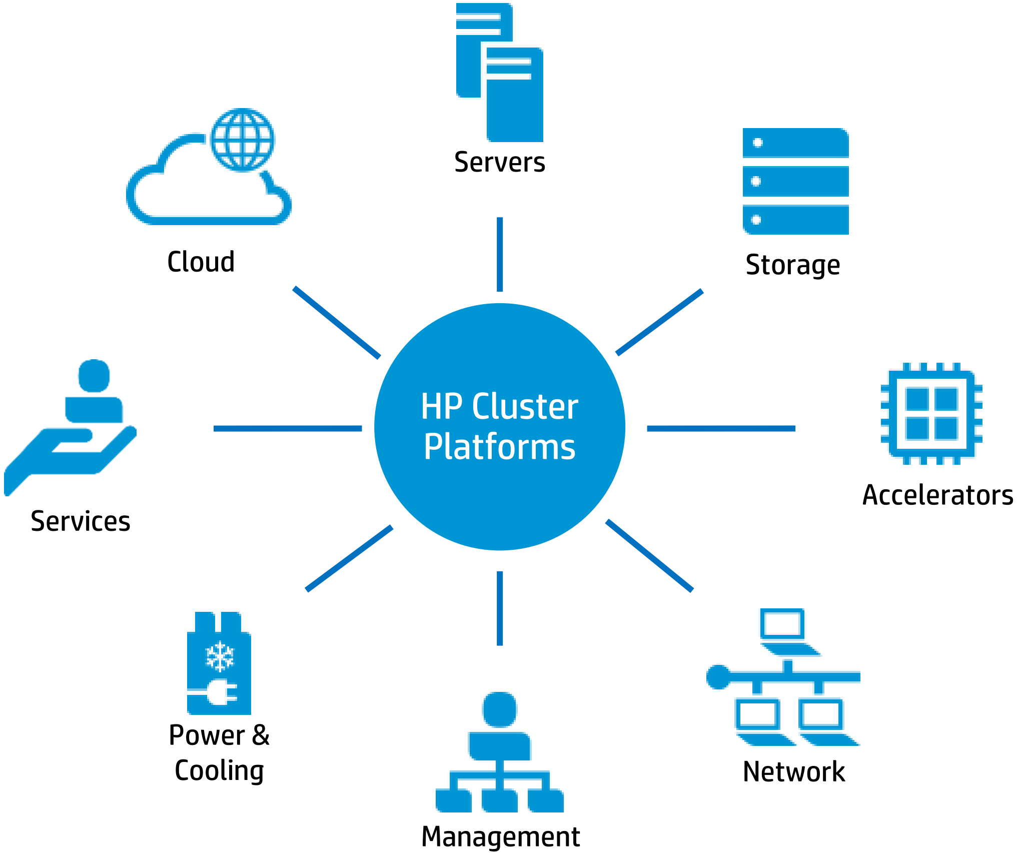
Introducing

HP Apollo Family

Optimizing rack-scale computing for HPC

Thank you

Delivering a complete HPC solution


Servers
Cloud
Storage
HP Cluster Platforms
Services
Accelerators
Power & Cooling
Network
Management