How to build infrastructure for HPC with Huawei

Ivan Krovyakov

IT CTO, Huawei Enterprise Russia

LEADING NEW ICT

HPC Overview

HPC Overview

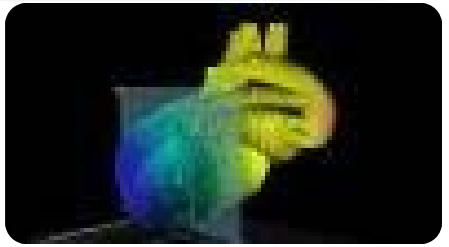
HPC Trends and Challenges

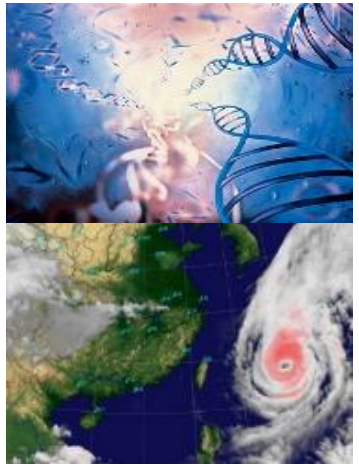
Huawei HPC Solution

Industry Applications Have Different Requirements on HPC

Memory-constrained

- Large memory and high bandwidth
- Application: numerical computing





I/O-intensive

- High-bandwidth and large-capacity storage
- Application: animation rendering and gene analysis

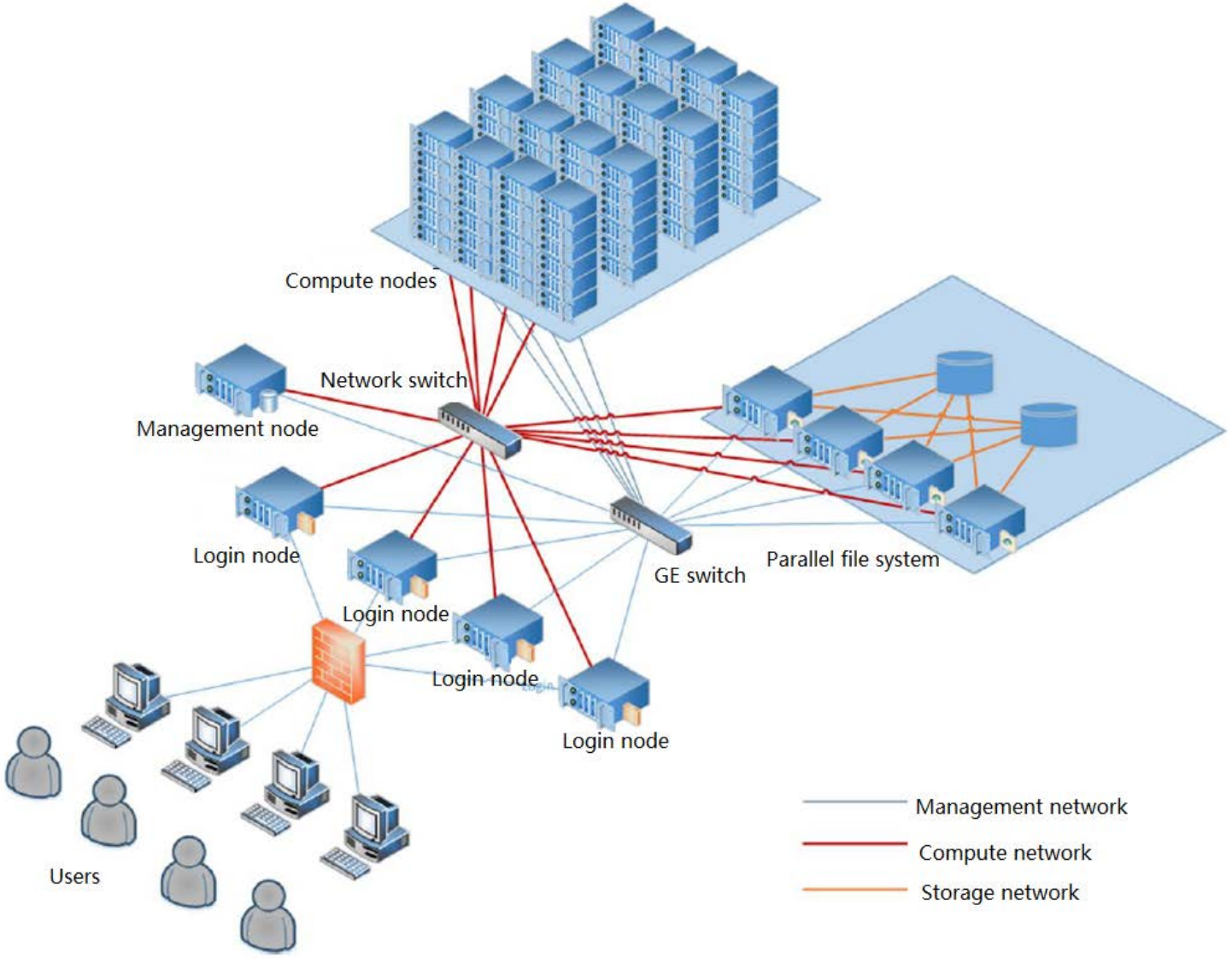
Compute-intensive

- High-speed CPU
- Application: Meteorology, earthquake, and aerospace

Network-intensive

- High-speed network interconnection
- Application: CAE simulation

Typical HPC Cluster Architecture



HPC Trends and Challenges

HPC Overview

HPC Trends and Challenges

Huawei HPC Solution

What Kind of HPC Do Data Centers Need?

Continuous Compute Power Increase

- Moore's law slows down, and heterogeneous computing emerges.
- How to eliminate network and storage bottlenecks?

Lower HPC Use Entry Requirements

- High HPC CAPEX and long construction cycle
- Complicated management, requiring professional skills

Diversified Apps Accelerated

- Different application characteristics
- How to drive traditional and emerging applications?

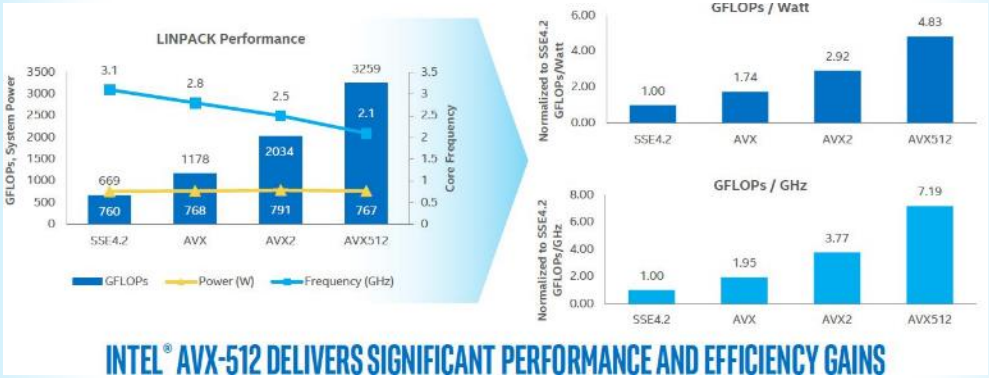
Optimized Computing Perf/Watt Ratio

- Energy consumption continuously increases TCO
- GFlops per watt

x86 Is the Mainstream for HPC, and ARM Has Perf/Watt Advantages

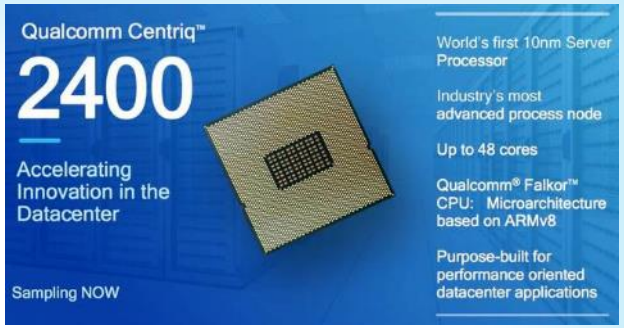
x86 Architecture

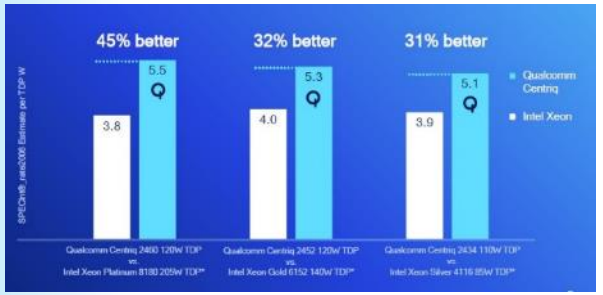
Microarchitecture	Instruction Set	SP FLOPs / cycle	DP FLOPs / cycle
Skylake	Intel® AVX-512 & FMA	64	32
Haswell / Broadwell	Intel AVX2 & FMA	32	16
Sandybridge	Intel AVX (256b)	16	8
Nehalem	SSE (128b)	8	4

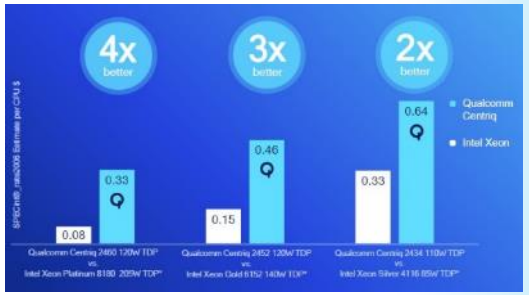


- AVX-512 doubles the floating-point computing capability and improves the computing perf/watt efficiency.
- Integrates the OPA Fabric for better cluster performance.

ARM Architecture

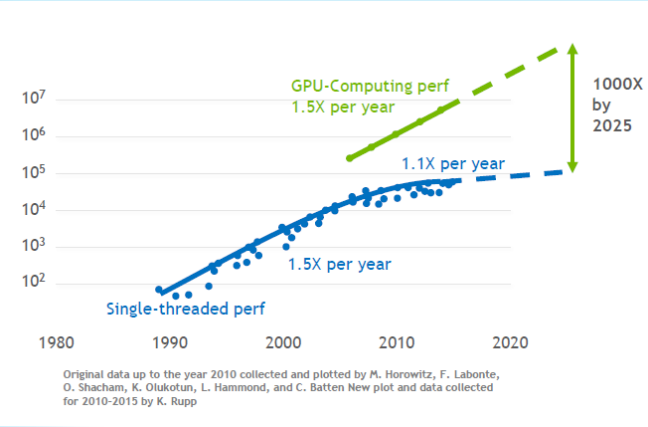


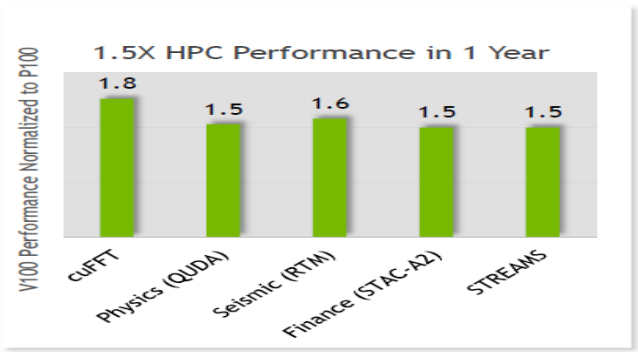




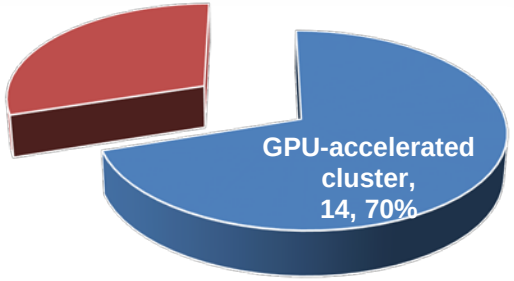
- 64-bit ARMv8 architecture delivers better perf/watt ratios and cost-effectiveness.
- As Thunder X2 and Centriq 2400 enter mass production, more and more mainstream vendors will provide the ARM HPC solution.

AI and Heterogeneous Computing See Rapid Application in HPC

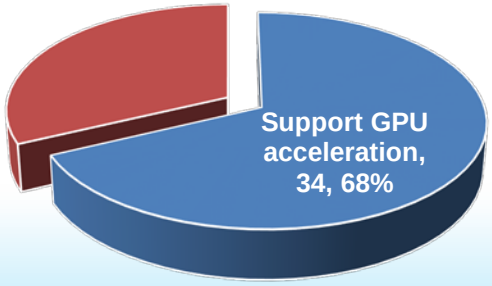




TOP20 HPC cluster with the highest perf/watt ratio



Most commonly used TOP50 HPC applications



HPC Cloud Can Meet the Rapidly Changing Service Requirements

Traditional HPC:

- Asset-heavy, long construction cycle
- Fixed computing scale

Cloud advantages:

- Request and use on demand, enabling rapid application of new technologies
- Flexibly coping with service load bursts

Multi-tenant sharing, pay-as-you-use

- Dynamic application and sharing, secure isolation between tenants
- Rent as per needed, saving investment and shortening the construction cycle

Flexible self-service capabilities

- Provide VMs, cloud bare-metal machines, and computing instances
- Create clusters and deploy HPC applications automatically

Collaboration and sharing

- Data centralization, cross-organization and cross-region collaboration

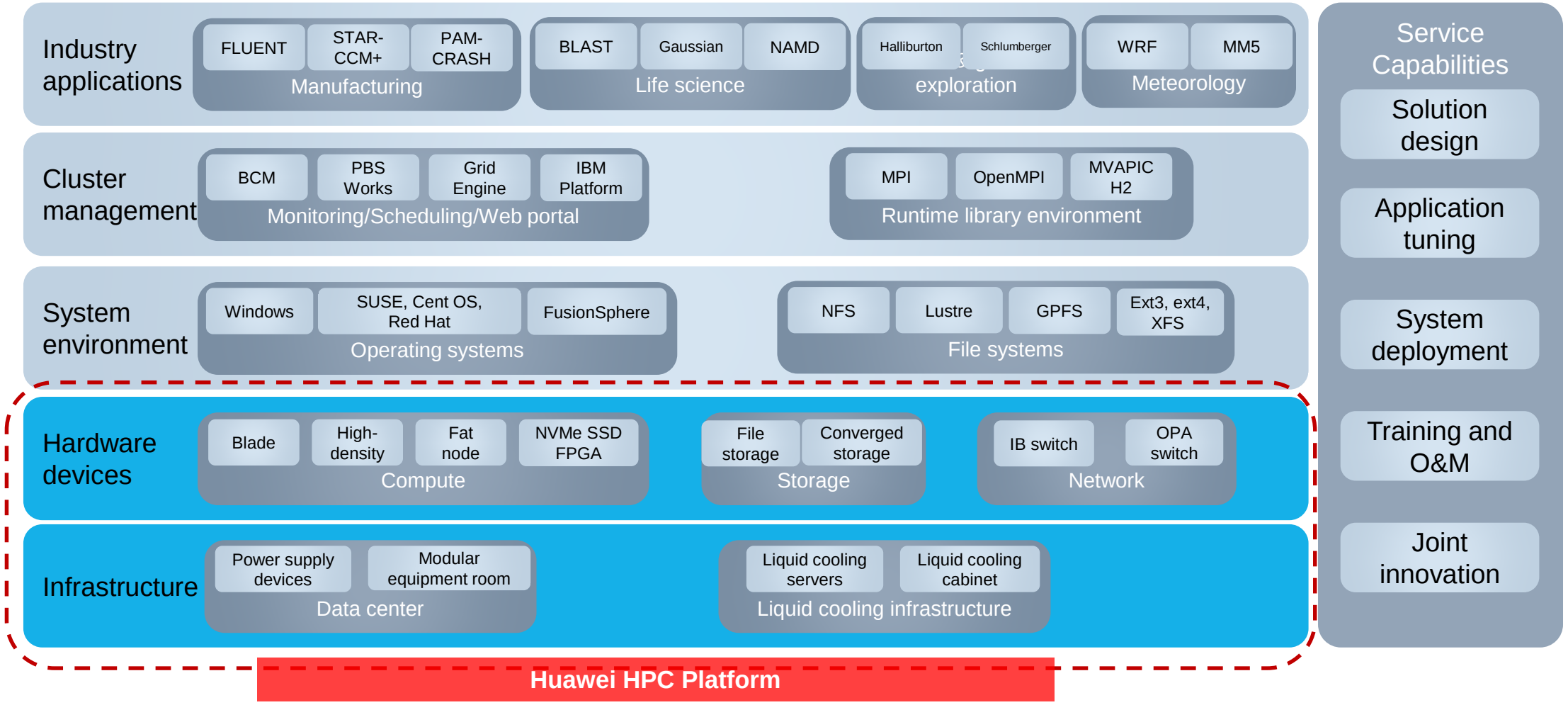
Huawei HPC Solution

HPC Overview

HPC Trends and Challenges

Huawei HPC Solution

Focus on HPC Hardware Platforms and Infrastructure



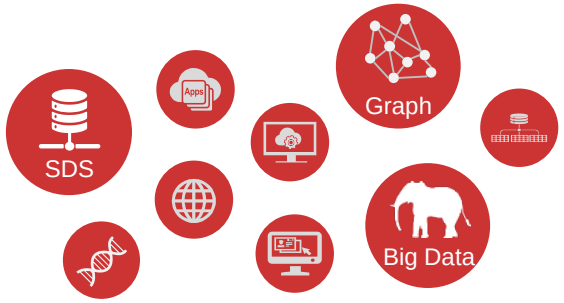
Huawei HPC Advantages

Ultimate Efficiency

Smaller footprint, lower power consumption for higher performance

- E2E engineering design capability
- Efficient and reliable liquid cooling technology
- Integrated delivery and installation

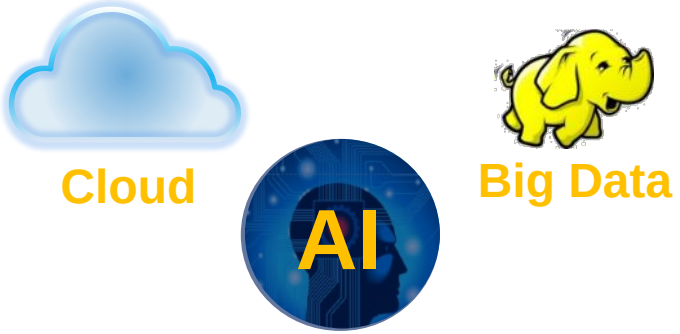
Application-optimized



Ultimate performance optimized for applications

- Flexible modular architecture
- Diversified innovative forms
- Hardware acceleration for in-depth application optimization

Adaptive to Changes



Future-proof HPC converged architecture

- Rapid application of novel technologies
- Multi-purpose HPC system
- HPC combined with cloud

Ultimate Efficient HPC Infrastructure

Computing Platform

Storage, Network, and Management

Liquid Cooling Technology

Huawei HPC Platform - FusionServer V5 Series Servers

Basic Platform

Multi-Node Server

X6000 2U 4-node

- Supports SKU with integrated OPA
- All-flash acceleration

Blade Server

E9000
12U multi-node

- Converged architecture with network switching
- EDR IB & OPA 100G

Rack Server

2488 2U 4-socket

- 4 CPUs on a single node, with simplified network solution
- Balance between compute density and power consumption loads

Extended Platform

Fat Node Server

KunLun
8- to 32-socket

- Multi-core compute on a single node
- Massive in-memory computing

Heterogeneous Server

+
GPU-accelerated server, 4U rack

- Supports multiple CPU:GPU ratios
- Fully modular design

Supports full series of
Intel® Xeon® Scalable processors

X6000 High-Density Server - Fully Upgraded with New Features

New features!

Supports 2.5" NVMe SSDs and 3.5" HDDs
Suitable for different acceleration scenarios

Board-level liquid cooling, PUE ≤ 1.1
Green and energy-saving

Supports SKU with integrated OPA
Boost cluster performance

Supports EDR IB & OPA standard cards
High-speed interconnect

Supports rear-access aggregated management network port
Simplifies cabling

Supports 2 kW/3 kW PSUs and power capping
TDP = 205 W CPU

XH321 V5 Half-Width Dual-Socket Compute Node

- **2 full-series Scalable** Processors, **16** DDR4 DIMMs
- **6 x 2.5" NVMe SSDs** or 3 x 3.5" HDDs + 2 x M.2 SSDs
- **2 x GE + 2 x 10GE** LOM ports, 2 PCIe x16 slots
- Supports air cooling or **liquid cooling**

Liquid cooling!

Up to 72 Nodes in a Single Cabinet

4 x 1U rack servers

=

1 x 2U X6000

E9000 Blade Server - Converged Architecture Computing Platform

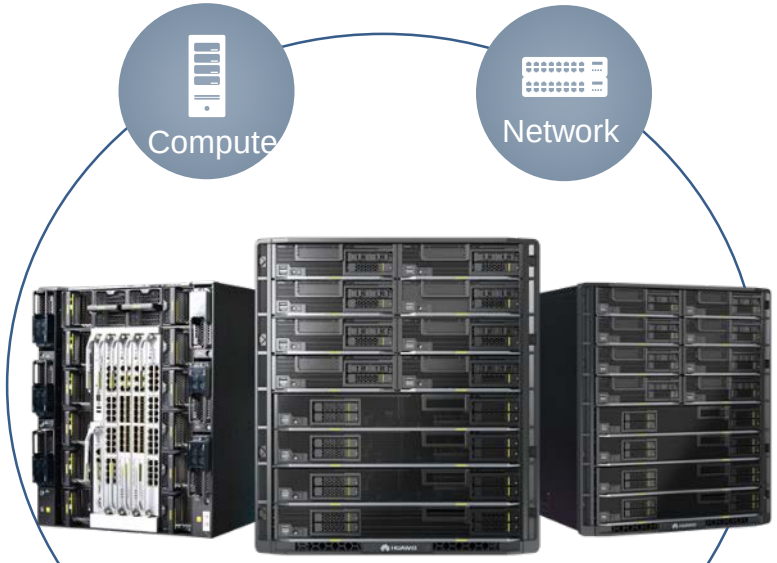
Rich Compute Node Types

Balanced compute node
CH121 V5: half-width, 2 Scalable Processors

GPU-accelerated compute node
CH221 V5: full-width, 2 dual-slot GPUs

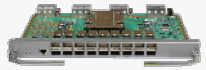
All-flash compute node
CH225 V5: full-width, 12 NVMe/SAS/SATA HDDs/SSDs

- Available in two form factors: half-width and full-width
- Supports CPU, GPU, and all-flash accelerated compute nodes



Powerful Integrated Switch Capabilities

CX620 100G IB EDR Switch module



CX820 100G OPA Switch module

CX320 10GE/40GE Ethernet switch module

- Integrated IB EDR/OPA and Ethernet high-speed switch modules
- 32T passive midplane switching capacity

Advantages of E9000 Blade Nodes and Chassis

Brand New Blade Nodes

CH121 V5 Half-Width 2S Blade

- **2 full-series Scalable** Processors, **24** DDR4 DIMMs
- **2*2.5”NVMe SSDs**
- 1 PCIe x16 slot, supporting air cooling or liquid cooling

Liquid cooling!

New product!

CH242 V5 Full-Width 4S Blade

- **4 full-series Scalable** Processors, **48** DDR4 DIMMs
- **4*2.5”NVMe SSDs**
- 1 PCIe x16 slot

Support Long-Term Evolution Chassis Capabilities

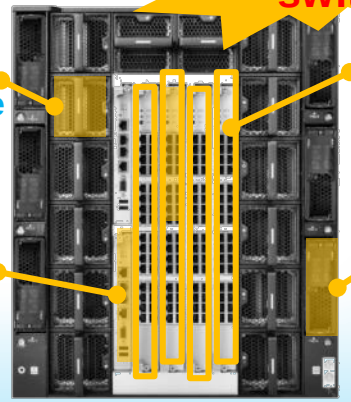
Half-width slot

- 12U chassis, supporting up to **16** half-width or **8** full-width compute nodes
- Supports compute, storage, and GPU nodes

Full-width slot

- Fully modular design
- Redundancy design for key modules:
  - Management modules support **1+1** redundancy
  - PSUs support **N+N** redundancy
  - Fan modules support **N+1** redundancy

Integrated high-speed switch!



G5500 Heterogeneous Computing Platform - 50+TFLOPS on a Single System

1 G560 V5 node: configured with 8 Tesla V100

2 G530 V5 nodes: configured with 8 Tesla V100

- Supports 8 NVIDIA® Tesla® V100/P100/P40
- Supports 1 full-width G560 V5 or 2 G530 V5 nodes
- 2S+24 DIMMs per node
- AI, HPC, database, cloud, and video application acceleration

	For PCIe Servers
Compute	7 TF DP · 14 TF SP · 112 TF DL
Memory	HBM2: 900 GB/s · 16 GB
Interconnect	PCIe Gen3 (up to 32 GB/s)
Power	250W

G5500 Product Portfolio

1 chassis, 3 types of compute nodes, and 4 types of heterogeneous nodes

G5500 full-width model

G560 full-width dual-socket compute node

G560 V5 full-width dual-socket compute node

GP608 heterogeneous node

- **8 V100/P100/P40/P4 GPU cards based on PCIe**

GS608 heterogeneous node

- **8 V100 GPU cards based on NVLink**

New

G5500 half-width model

G530 V5 half-width dual-socket compute node

GP308 heterogeneous node

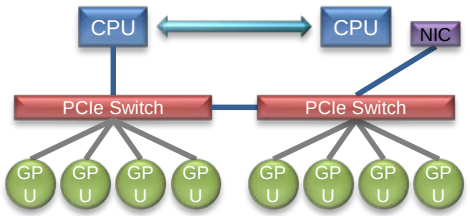
- **4 V100 or 8 V100 GPU cards (150 W)/P4 based on PCIe**

New

GP316 heterogeneous node

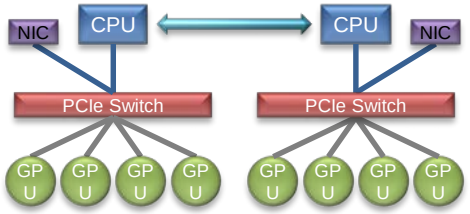
- **16 P4 GPU cards based on PCIe**

G5500 Supports Various CPU:GPU Ratios to Meet Different Workload Requirements



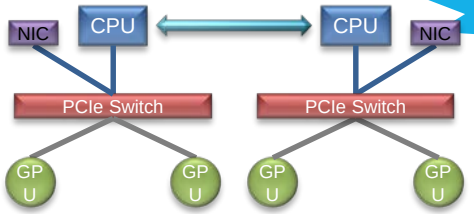
Single-RC topology, 1:8 for CPU:GPU, **AI training**

G560 V5 supports one-click switching!



Dual-RC topology, 1:4 for CPU:GPU, **HPC loose coupling**

This topology is supported by G530 V5.



Dual-RC topology, 1:2 for CPU:GPU, **HPC tight coupling**

Support Multiple Topologies

- One-click topology switching for AI and HPC
- Supports diverse applications and reduces hardware investment

Fully Modular Design

- Decoupled design for CPUs and heterogeneous resources
- Modular design for PSUs, hard drives, and fan modules

KunLun 32S Fat Node Computing Platform

Ultimate Performance

32S/576C, 32 TB
Elastically scale-up architecture

Stability and Reliability

RAS 2.0
Open platform with the highest-level reliability

KunLun Fat Node Computing Platform

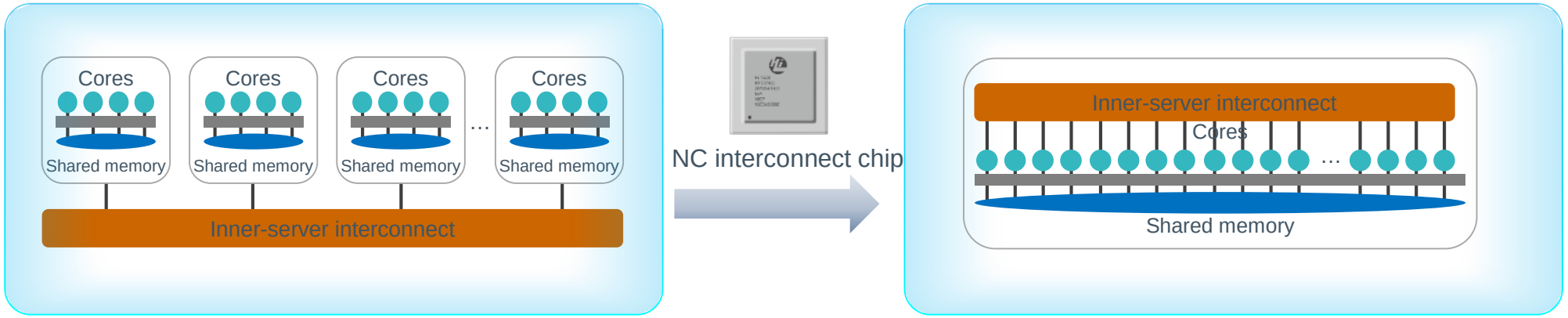
NC Interconnect Chip

- Scale-up
- Fault-tolerant

Management Chip

- Failure Analysis Engine
- Underlying feature security

KunLun Platform - Large-Capacity In-Memory Computing with Low Latency



- Up to 8 CPUs per system
- Milliseconds of latency for inter-server data transmission, including latency from CPU processing, NIC processing, and link transmission

- KunLun 9032 supports 32 CPUs in a single system, with 576 compute cores and memory of up to 24 TB
- CPU high-speed network reduces data transmission latency to nanoseconds, enabling faster service response

Ultimate Efficient HPC Infrastructure

Computing System

Storage, Network, and Management

Liquid Cooling Technology

Huawei OceanStor 9000 NAS High-Performance Storage

OceanStor 9000

Superb Performance

- 400 GB/s throughput; InfoTurbo boosts single-client bandwidth to 2.5 GB/s

Elastic Scalability

Industry No. 1

- Supports flexible scalability of 3–288 nodes; up to 100 PB storage capacity for a single file system

Open Convergence

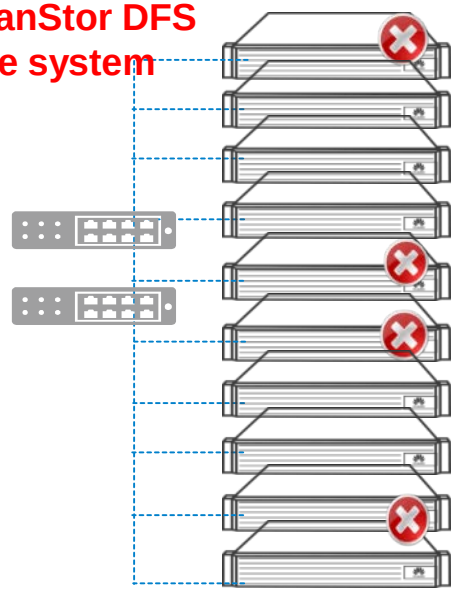
- All-IP architecture and universal hardware architecture, supporting multiple protocol and data types

Key Capabilities of OceanStor 9000 NAS

Fully Symmetric Architecture, Reliable Data Protection

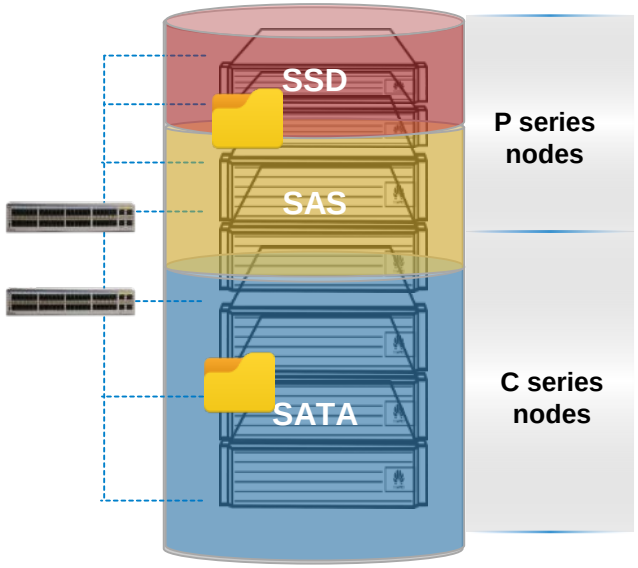
- Use the metadata decentralization design
- Tolerates up to 4 faulty nodes
- Data restores at up to 1 TB per hour

Integrates the OceanStor DFS file system



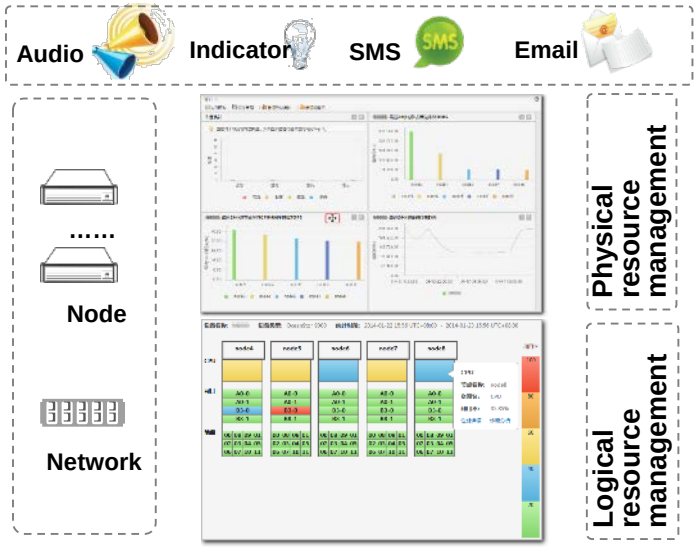
High Bandwidth Throughput and Flexible Expansion

- 1.6 GBps bandwidth per node, linear expansion of capacity and performance
- Flexible combination of various high-performance and archiving nodes
- Intelligent node load scheduling, achieving load balancing



Visualized Management, Easy to Use

- Unified management of storage, analysis, and archiving, and visualized management of physical and logical resources
- Single file system, simplifying management and operations



Prefabricated Data Center

All-In-Cabinet Small-Sized HPC

FusionModule500

FusionModule800

- Cabinet-level deployment, taking only 2 hours to install onsite
- 1–6 cabinets, supporting HPC systems of 10 to 100 TFLOPS

All-In-Room Medium- to Large-Sized HPC

Single row

Double rows

FusionModule2000

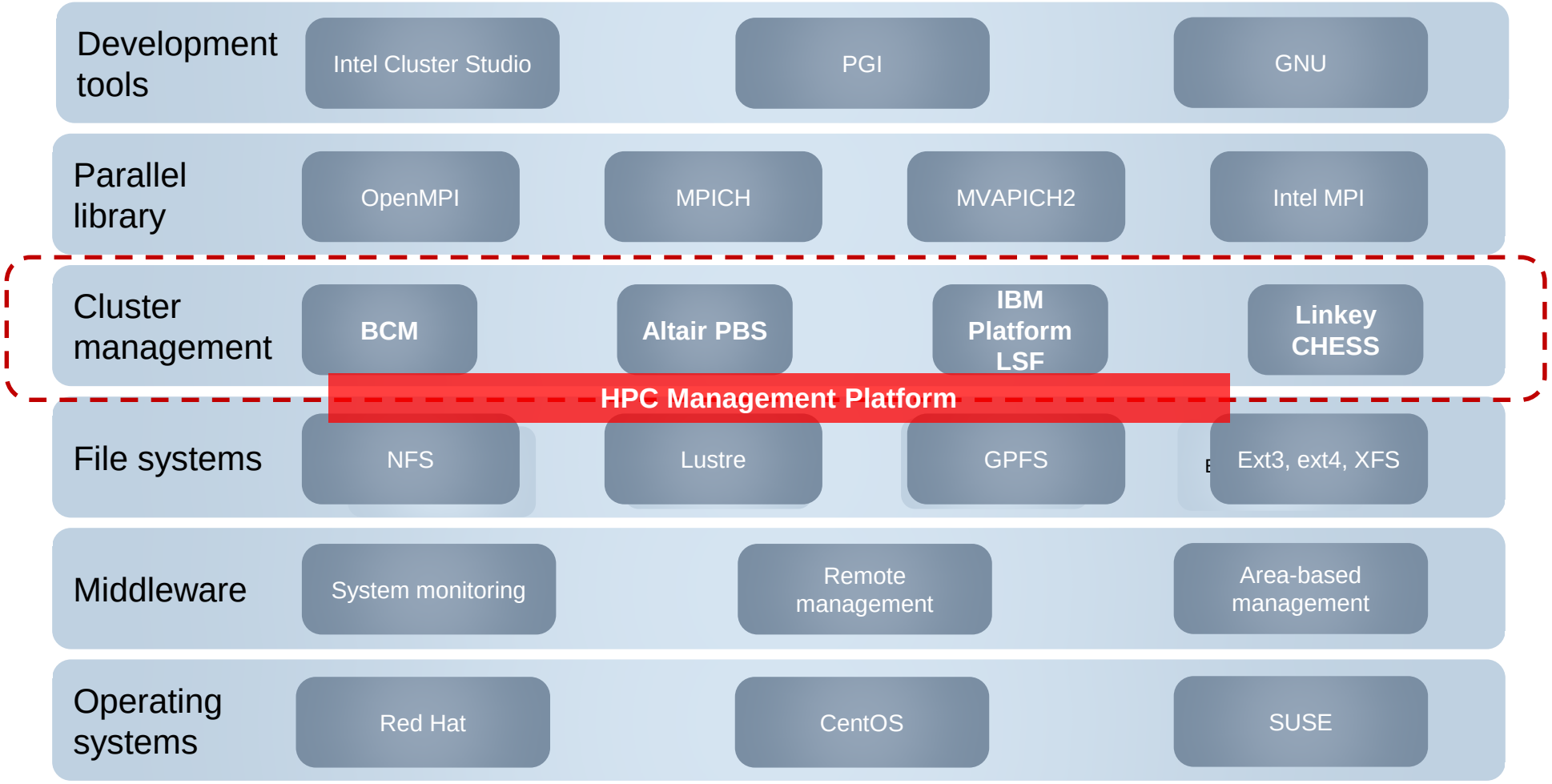
- Supports single- or double-row confined cold/hot channel deployment in an equipment room of 500 m^2 or less
- 2–48 IT cabinets, supporting HPC systems of 100 TFLOPS to 1 PFLOPS

All-In-Container HPC

FusionModule1000A

- Factory prefabricated, pre-verified, and onsite delivery, reducing deployment cycle by 80%
- 8 IT cabinets, supporting HPC systems of 10 to 100 TFLOPS

Huawei HPC Software Solution



Integrate Mainstream Cluster Management Software

Bright Computing

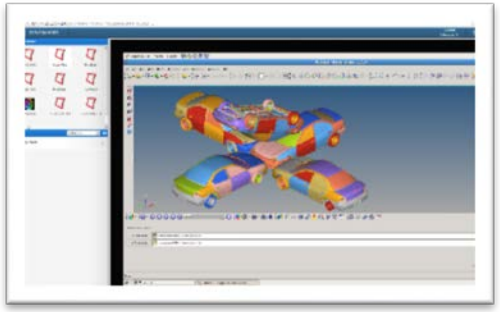
Platform Computing an IBM Company

Altair | PBS Works™

聯科集團 CLUSTERTECH

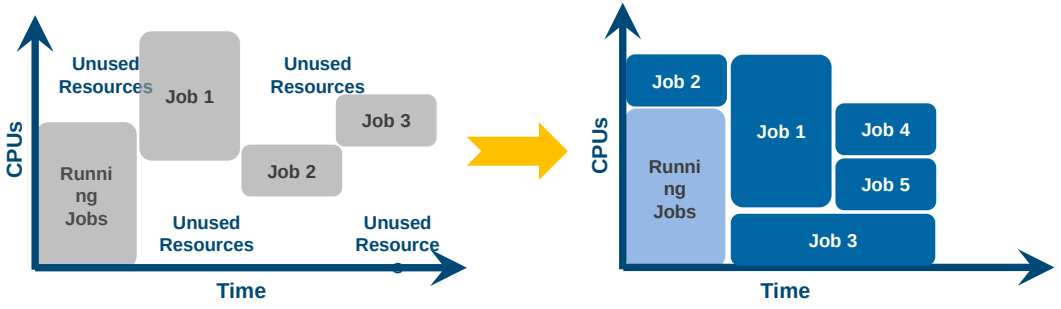
Workload management portal
— Job submitting, management, monitoring, and statistics

Visual portal
— Remote visualization & collaborative design



Simplified Deployment, Visual Monitoring, Efficient Management

Job scheduling
— Efficient resource utilization



Statistics and analysis tool
— Job statistics and report

HUAWEI

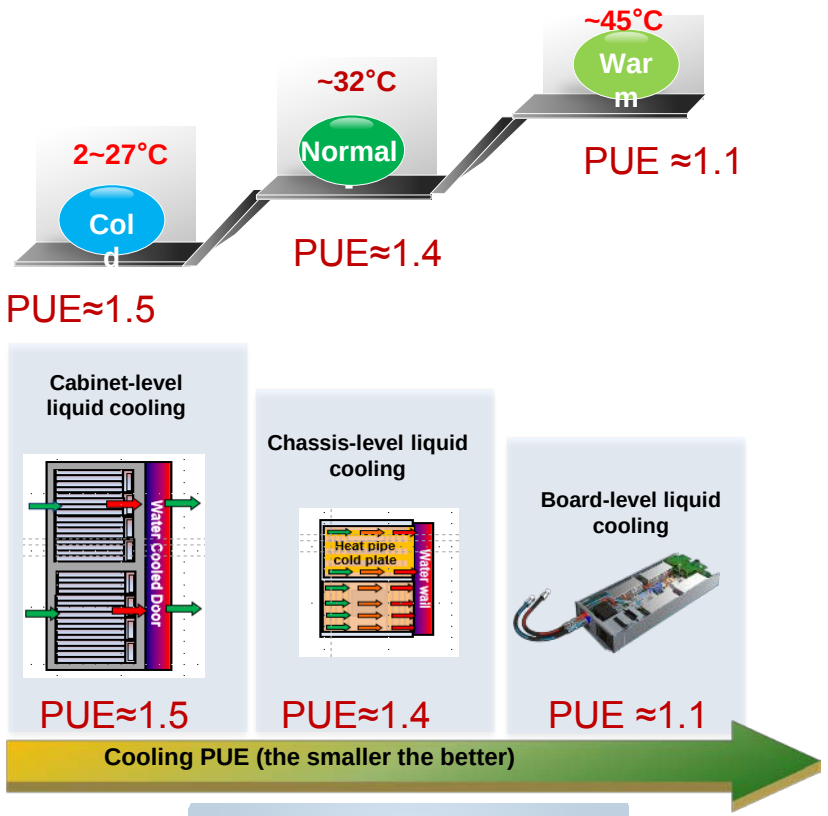
Ultimate Efficient HPC Infrastructure

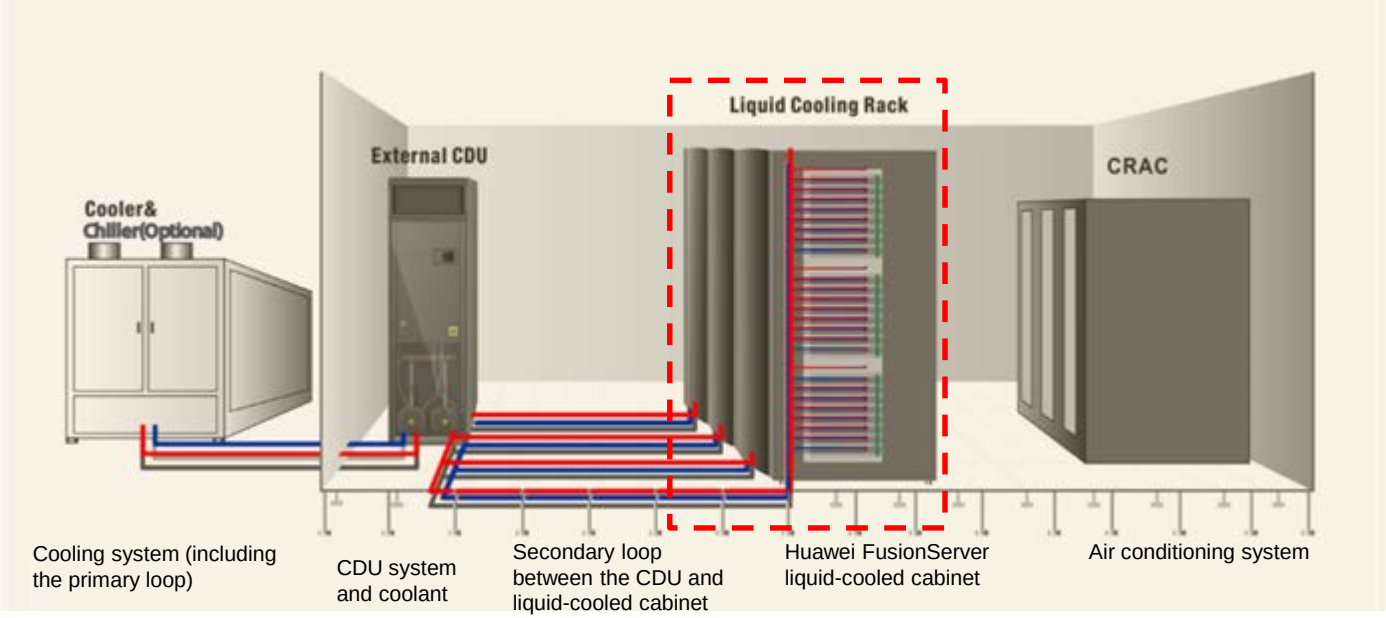
Computing System

Storage, Network, and Management

Liquid Cooling Technology

Huawei Board-Level Liquid Cooling Solution Supports 45ºC Warm Water Cooling





High Reliability

- Integrated cold plate components
- Stringent reliability testing

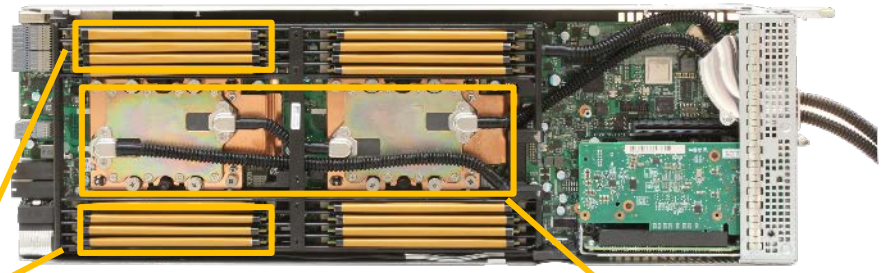
High Perf/Watt Ratio

- Cooling PUE ≤ 1.1
- Supports 45ºC warm water cooling

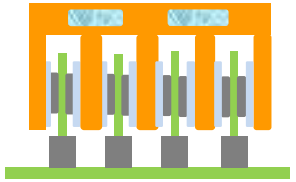
Cost-Effective

- 30% lower TCO

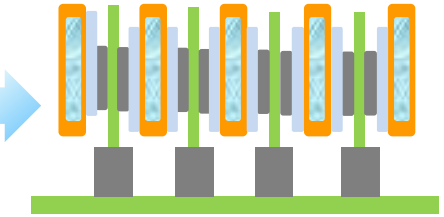
Optimized X6000 Board-Level Liquid Cooling Design



Memory cooling board design



Inter-DIMM water flow design

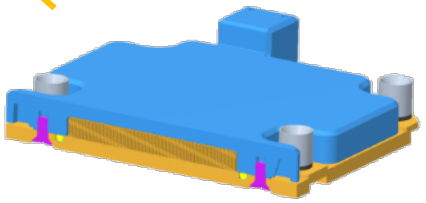


Multi-channel inter-DIMM water flow

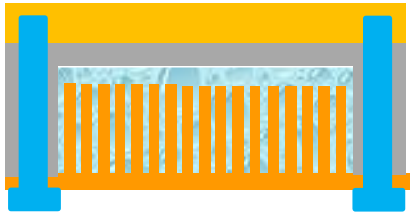
Thermal resistance reduced 65%⬇

Fence-style fixture design

Area of contact with air cooling reduced 80%⬆



Metallic reinforced plate for CPU Cooling



Internal serrated micro-channel design

Metallic reinforced plate

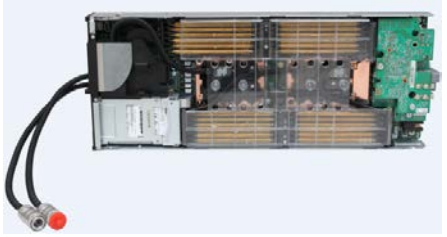
Deformation reduced by 90%⬇

Optimized serrated design

Energy efficiency up by 10%⬆

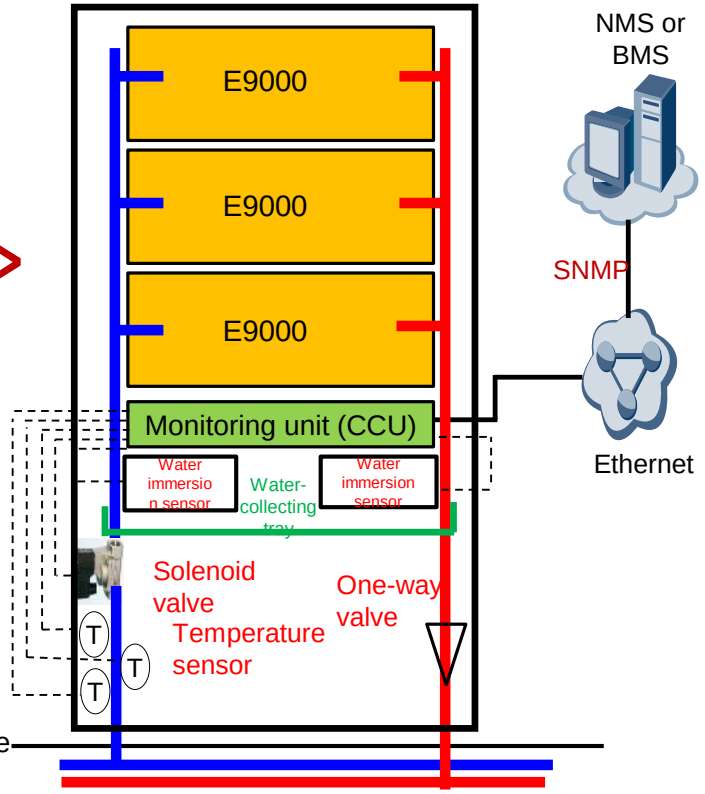
High Reliability Design Ensures Long-Term Liquid Cooling Operation

Cooling for CPU、RAM & VRD



CH121L V5 Blade node

XH321L V5 High-density node



Modular design: modular cold plate, supporting for blade and high-density server type

Abnormality monitoring: leakage prevention, temperature monitoring, water control

System test: 217 reliability tests

Application-oriented

Industry Solutions

System Requirements of Mainstream CAE Application Software

Crash simulation software requirements:

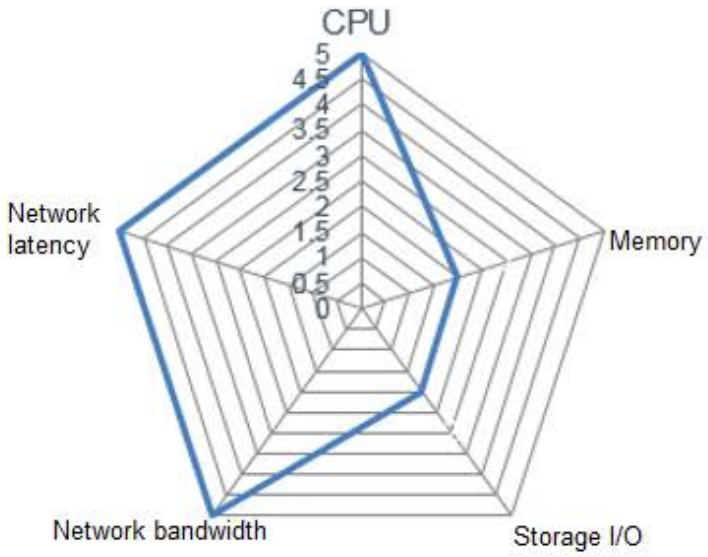
- Two x86 DC/QC processors
- Memory 2-4GB/core
- Local storage 1-2 disks or Shared FS
- Low disk I/O
- Network IB FDR/EDR

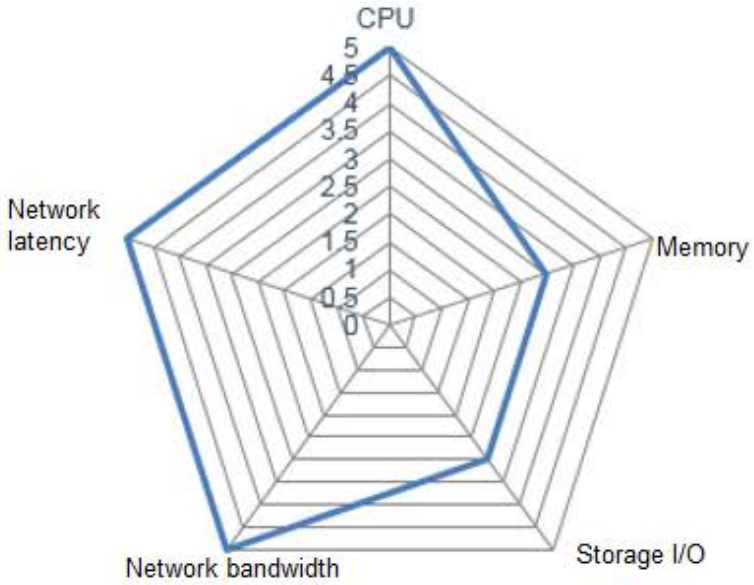
CFD simulation software requirements:

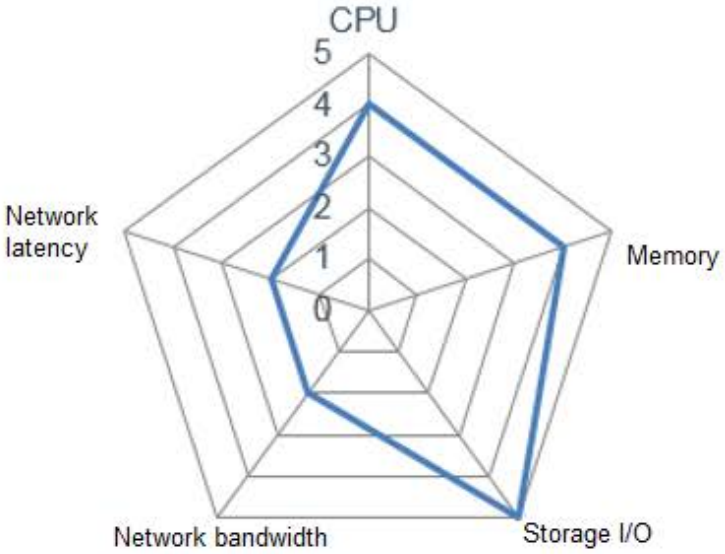
- Two x86 DC processors
- Memory 4-8GB/core
- Local storage 1-2 disks or Shared FS
- High disk I/O
- Network IB FDR/EDR

NVH and Structures simulation software requirements:

- Two RISC, IA64, x86 DC processors
- Memory 8+GB/core
- Local storage 4-12 disks or Shared FS
- Very High disk I/O
- Network Ethernet



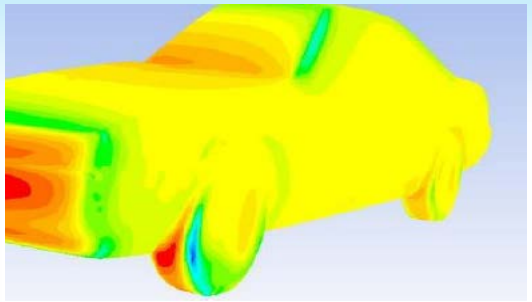




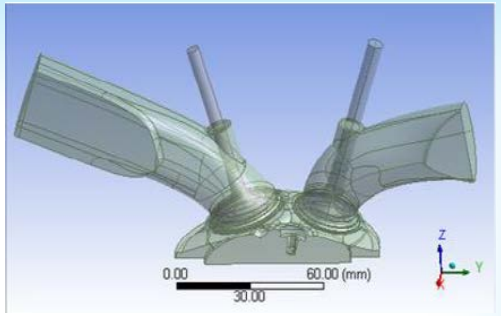
Fluent-based Simulation Performance Optimization

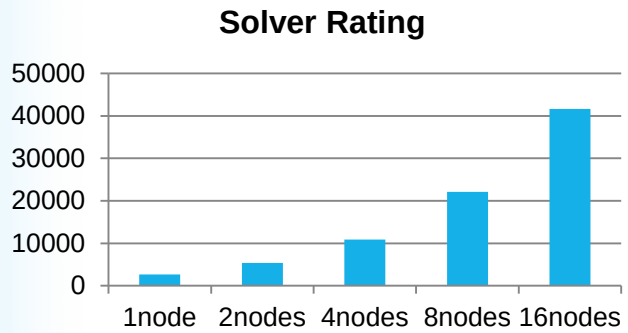
SYSTEM UNDER TEST	HUAWEI FusionServer X6800 High-density Server (Including 16 piece XH620 v3 server nodes)
CAE Application	ANSYS Fluent version 17.2
PROCESSORS	2x Intel Xeon E5-2680 V4 (14 core, 2.4GHz)
MEMORY	8x 16GB DDR4-2400MHz
NETWORK	56G FDR IB network
HARD DRIVE	2x 300G 10K RPM
OS	Red Hat Enterprise Linux 6.7
MPI	Intel MPI 5.0.3, Open MPI 1.6.5, IBM Platform MPI 9.1.3.1

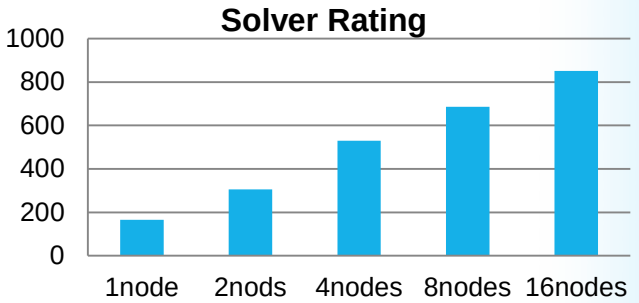
External airflow disturbance model

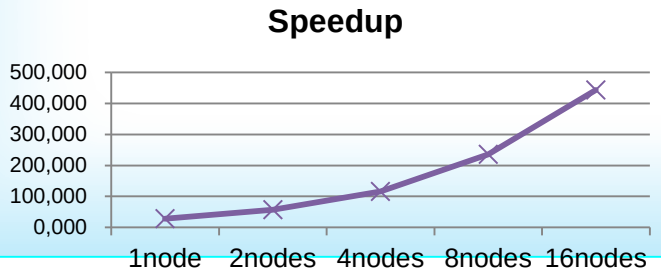


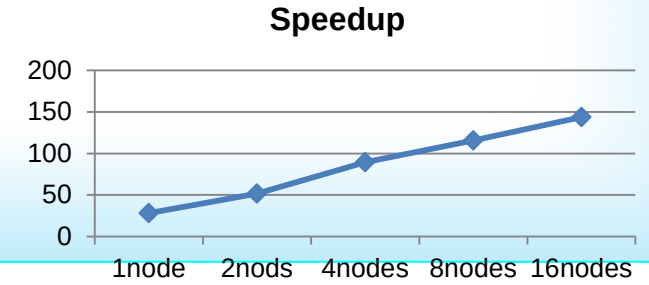
4-stroke spray-guided gasoline direct injection model











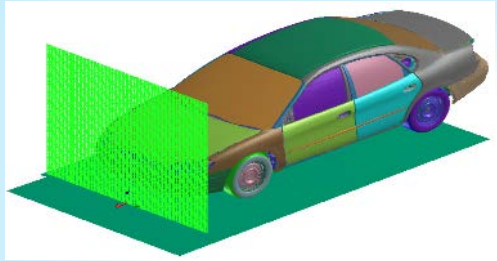
Data source: *Huawei-ANSYS Fluent Performance Test White Paper*

PAM-Crash-based Simulation Performance Optimization

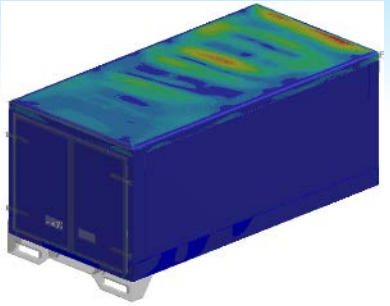
Virtual Performance Solution

SYSTEM UNDER TEST	HUAWEI FusionServer E9000 Blade Server (Including 16x CH121 v3 computing nodes)
CAE Application	ESI Pam-Crash 2015.0
PROCESSORS	2x Intel Xeon E5-2697A V4
MEMORY	8x 16GB DDR4-2400MHz
NETWORK	56G FDR IB network
HARD DRIVE	2x 300G 15K RPM
STORAGE/FILESYSTEM	Huawei Oceanstor 9000 Parallel File System Storage
OS	Red Hat Enterprise Linux 7.1 (kernel 3.10.0-229.el7)
MPI	IBM Platform MPI 9.1.2

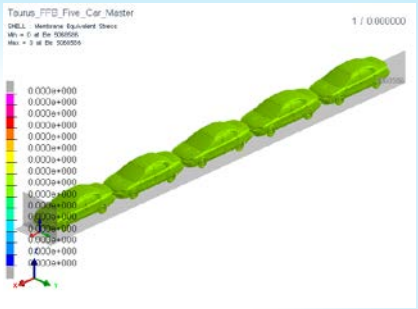
Rigid wall collision

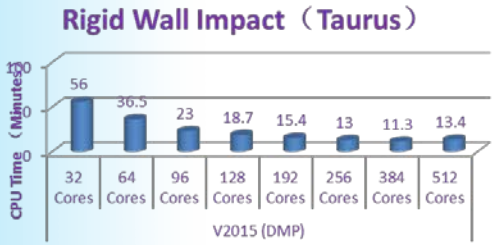


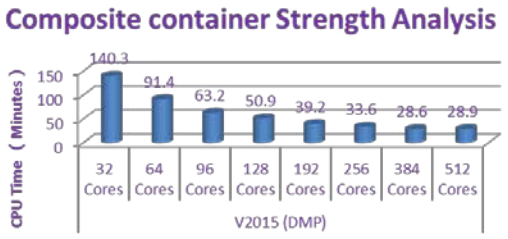
Strength analysis

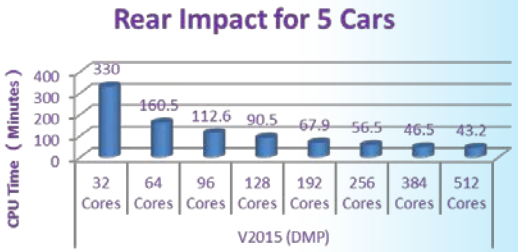


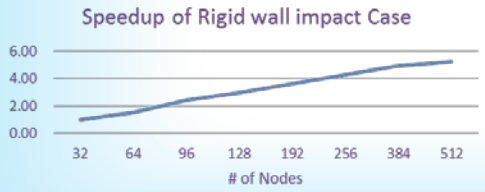
5-car rear collision

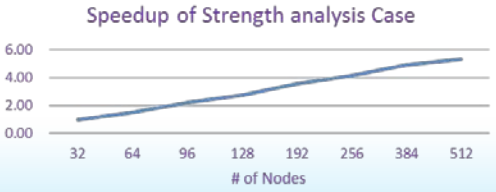


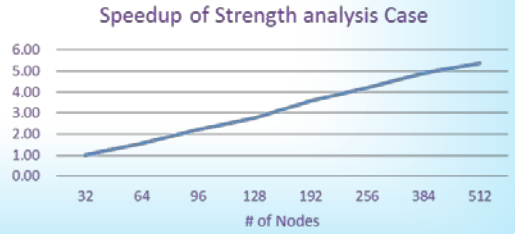










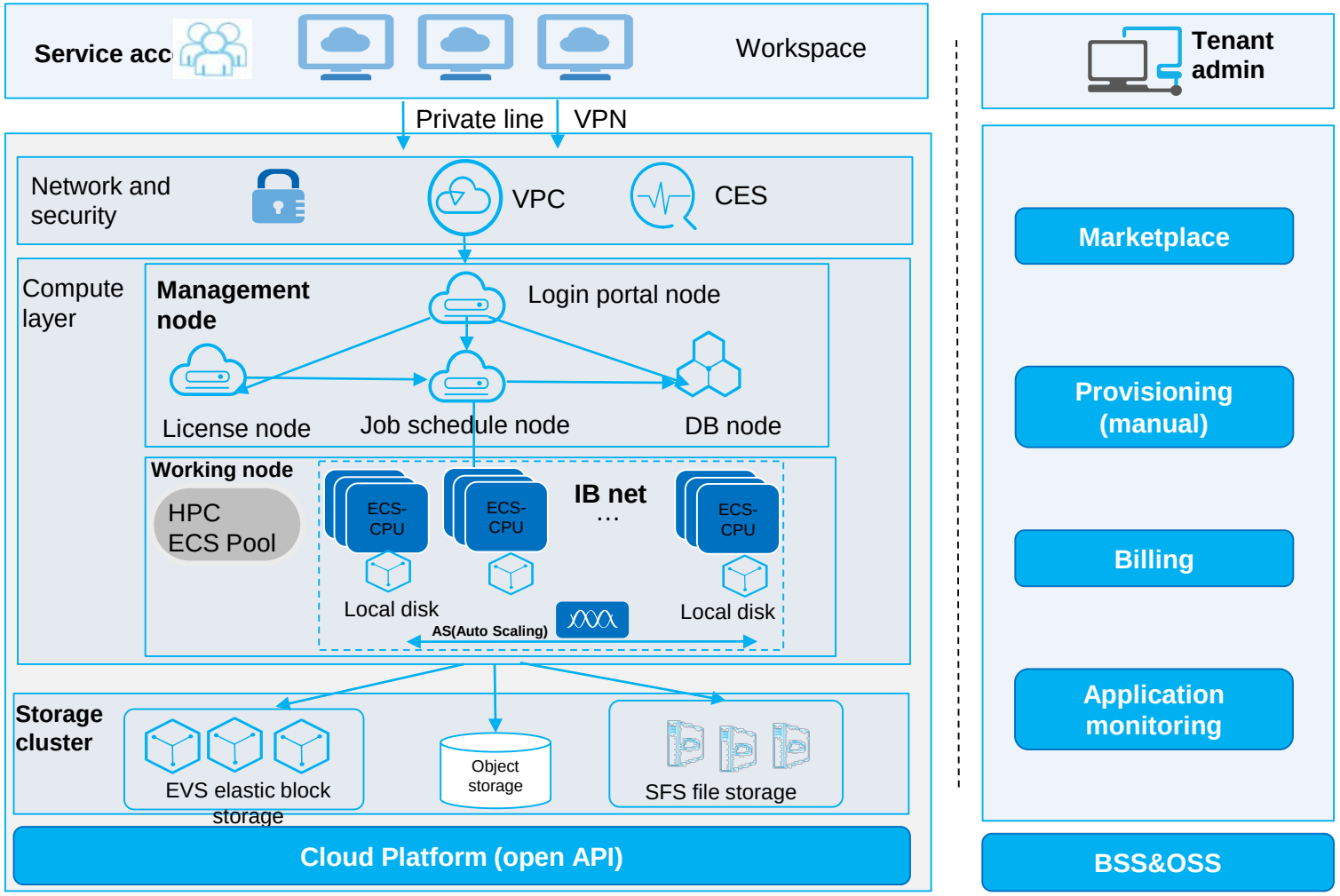


Data source: Huawei-ESI PAM-Crash Performance Test White Paper

Cloudification

HPC Solution

Huawei HPC Cloud Architecture



High Performance

- Bare-Metal Service (BMS) with large-specifications VMs
- 100G IB service network
- Nvidia P100 GPU acceleration
- FPGA data preprocessing
- Lustre parallel file system based on IB

Openness

- Open APIs to avoid lock-ins and facilitate migration
- Open and collaborative ecosystem

High Security

- Secure access via private lines and VPN
- Security isolation via VPC and security groups

Huawei HPC Cloud Key Capabilities: Continuously Building HPC Instances

LEADING NEW ICT

Key Capabilities Meet the Requirements of HPC Major Service Scenarios for Working Nodes

Design simulation cloud

Scientific computing cloud

Energy exploration cloud

Large-Specifications VMs

64vCPU+1TB RAM

- 3.2 TB SSD local disk
- 100G IB network
- 10G SR-IOV network enhancement

BMS *

Bare metal + SDI
Shared storage

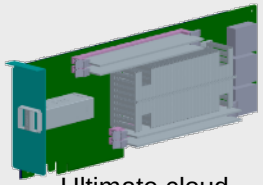
- No virtualization loss; automatic provisioning
- Up to 96 cores and 4 TB memory supported by bare metal
- Supports high-speed channels (100IB or GE) to better suit load scenarios
- More local disks, up to 8 P100 GPU cards

GPU Compute Instance*

Nvidia P100
GPU Acceleration

- Graphics acceleration GPU instance: M60
- Compute acceleration GPU instance: up to 8 P100
- Supports GPU P2P for peer-to-peer communication with higher bandwidth and lower latency
- Supports 100G IB interconnect

FPGA Instance *



Ultimate cloud acceleration
Data preprocessing

High-end FPGA (Xilinx) + PCIe Fabric (X16) + local NVMe SSD + 1.2 TB silicon optical interconnect (direct connection between FPGA cards)

*** Released in 2017 H2**

HUAWEI

Key Capabilities of HPC Cloud Match Requirements of Various Business Scenarios

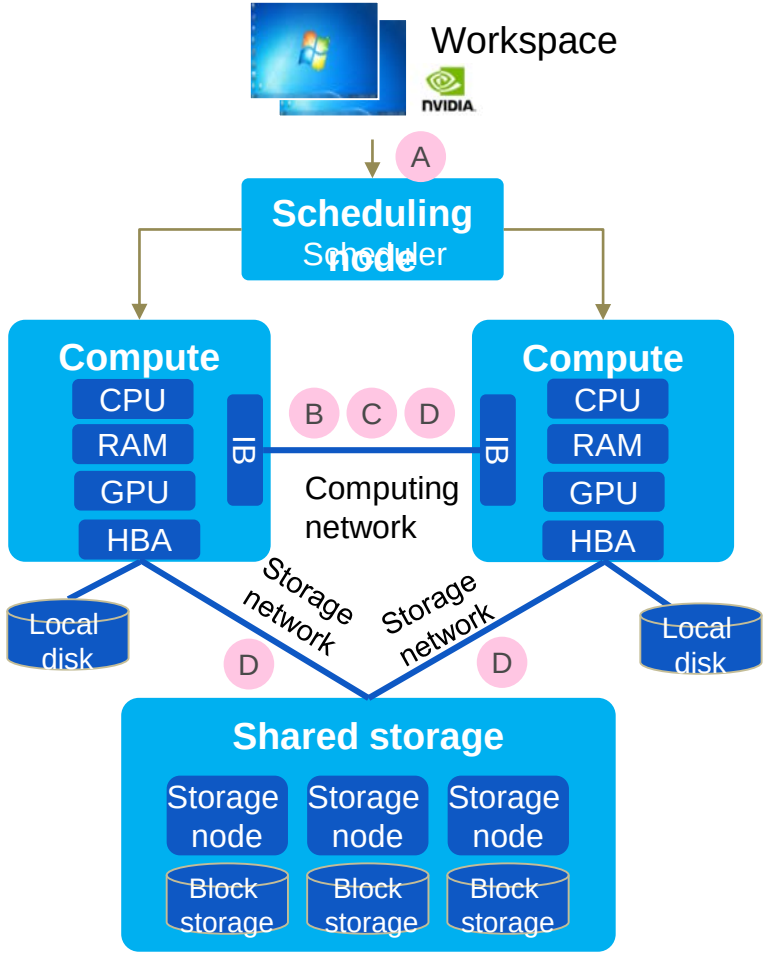
LEADING NEW ICT

CADaaS scenario

CPU: 2-8 Core
16-64GB memory
SSD or 15k Local Disk
The most powerful M60 GPU in the industry used for graphics rendering

Tightly coupled cluster computing

Computing network with low latency (2 μs)
100G IB network



High-performance Lustre parallel file system cluster

Loosely coupled grid computing

Moderate requirements on computing networks and storage networks (> 20 μs)

Data-intensive computing

Shared storage read/write:
High bandwidth (up to 50 GB) and low latency (< 5 ms)

HUAWEI

Success Cases

Win-Win, Open HPC Ecosystem

Customer Benefits

Industry Applications

Community/Organization

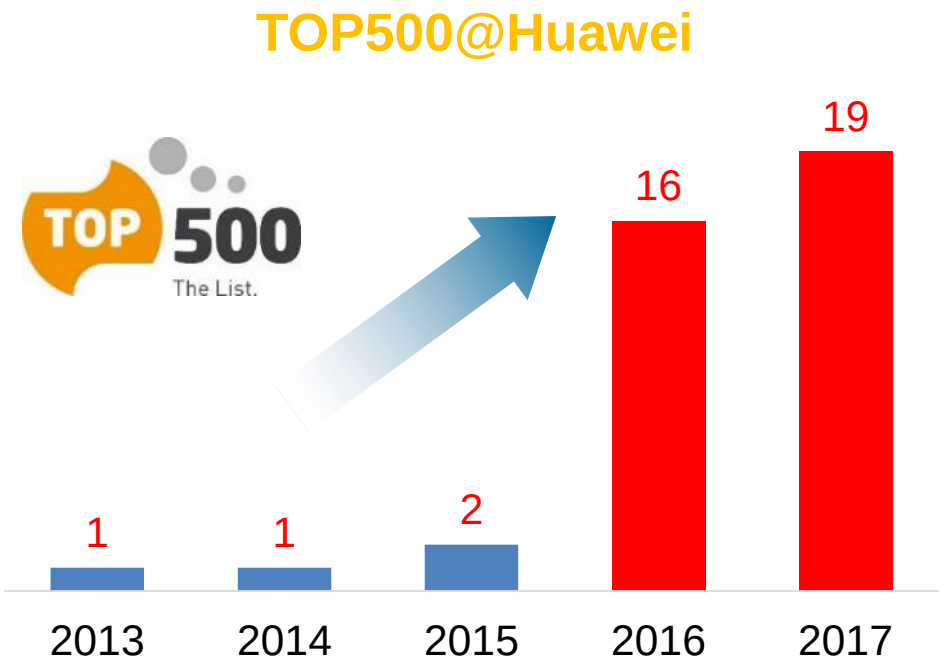
Management Software

Solution Integration

System Environment

Hardware/Infrastructure

Huawei's HPC Market Influence Increases Continuously



Top Universities

Top Manufacturers

Top Research Institutions

Top Supercomputing Centers

Data source: https: //www.top500.org/lists/2017/11/

Rich Experience in HPC Industry Project Construction

Supercomputing Centers & Organizations

CERN
PSNC
Deltares
Enabling Delta Life
INFN
Istituto Nazionale di Fisica Nucleare
CRS4
IDEAS BECOME LIFE
British Antarctic Survey
NATURAL ENVIRONMENT RESEA
PEMEX
Aramco
compute canada
PetroChina
TP
TÜRKİYE PETROLLERİ
ILLUMINATION MACGUFF
DIGITAL DOMAIN MEDIA GROUP

Universities

UNIVERSITY of TASMANIA
MAX-PLANCK-GESELLSCHAFT
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
Agency for Science, Technology and Research
SINGAPORE
BIBLIOTHECA ALEXANDRINA
مكتبة الإسكندرية
TÜBİTAK
ULAKBİM

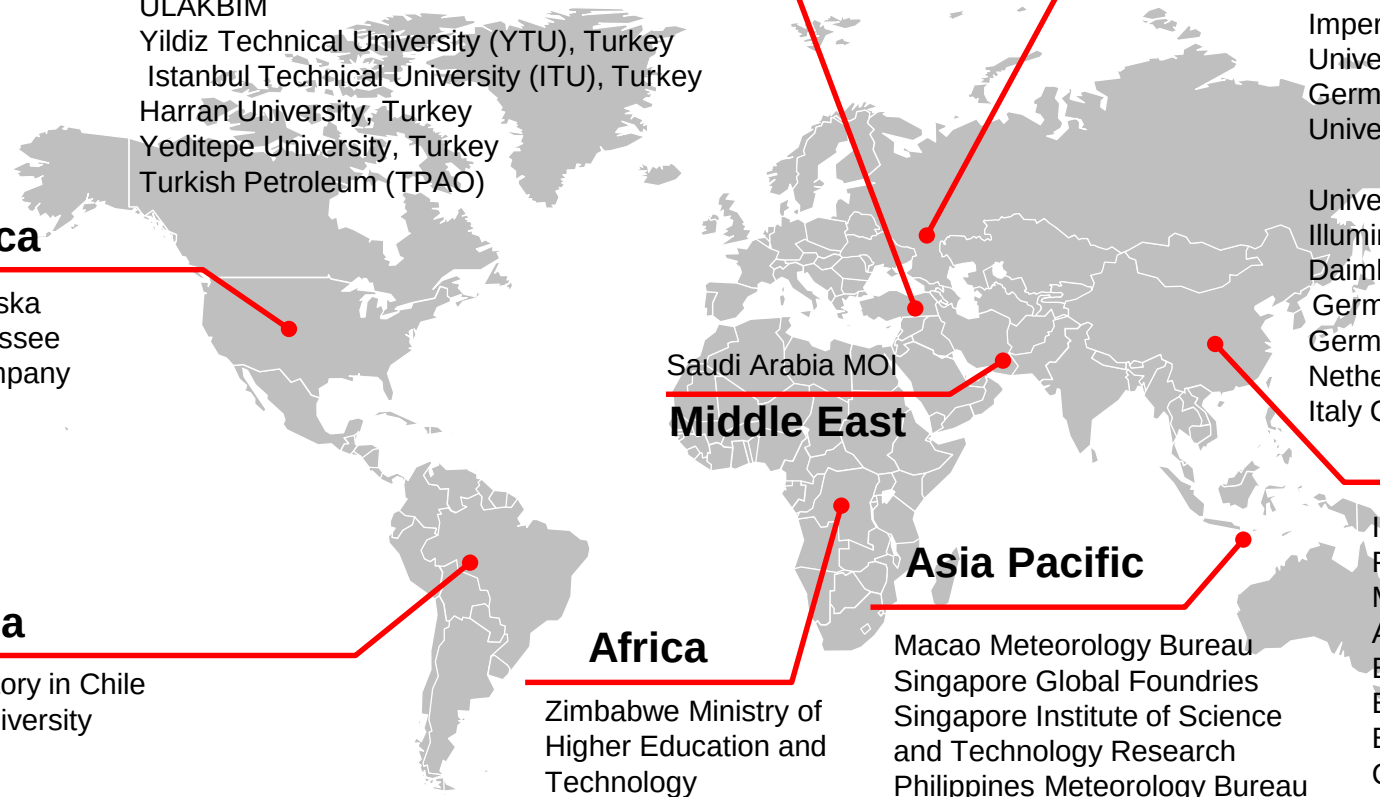
Manufacturing/Automobile Enterprises

DTU
MAN
Audi
BYD
unesp
UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
İTÜ
SMIC

Oil & Gas, Media, Meteorology

HUAWEI

HPC Installation Experience in Industries Worldwide



Central Asia

ULAKBIM
Yildiz Technical University (YTU), Turkey
Istanbul Technical University (ITU), Turkey
Harran University, Turkey
Yeditepe University, Turkey
Turkish Petroleum (TPAO)

Europe

UK Newcastle University
Imperial College London
University of Hamburg, Germany
University of Lubeck, Germany
University of Burgos, Spain
Illumination Mac Guff
Daimler Mercedes-Benz, Germany
German Ilya Ehrenburg
Netherlands Deltares
Italy CNR
University of Warsaw, Poland
Poland PCSS
University of Gdańsk, Poland
University of Silesia, Poland
Poland Cyfronet
Poland Qumak University
Saint-Petersburg State University, Russia

North America

University of Nebraska
University of Tennessee
Digital Domain Company

Middle East

Saudi Arabia MOI

China

Institute of Disaster Prevention, China Meteorological Administration
Environmental Protection Bureau, Hebei Province
Beijing Institute of Data Communication
Beijing Jiaotong University
Beijing University of Aeronautics and Astronautics
Southwest University
Capital Medical University
China Electric Power Research Institute
China Meteorology Bureau
Shanghai Observatory
ZhongXin Biotechnology Shanghai
Bureau of Geophysical Prospecting INC.
Tsinghua University
Beijing Genomics Institute
BGP

Latin America

CASSAC Observatory in Chile
Brazil Mckinsey University
Cuba oil CUPET
Venezuela PDVSA
Ministry of water resource, Mexico
Ministry of Agriculture, Mexico

Africa

Zimbabwe Ministry of Higher Education and Technology Development
South Africa CHPC

Asia Pacific

Macao Meteorology Bureau
Singapore Global Foundries
Singapore Institute of Science and Technology Research
Philippines Meteorology Bureau (phase 1)
Singapore A-Star
University of Victoria
The University of Queensland
Deakin University

University Waterloo Cluster Launch in 2017

SHARCNET @SHARCNET Following

wha
wei

10:36 A

SHARCNET @SHARCNET

The answer is our new cluster Gra
insta

SHAR
what h

1:47 PM

2 Likes

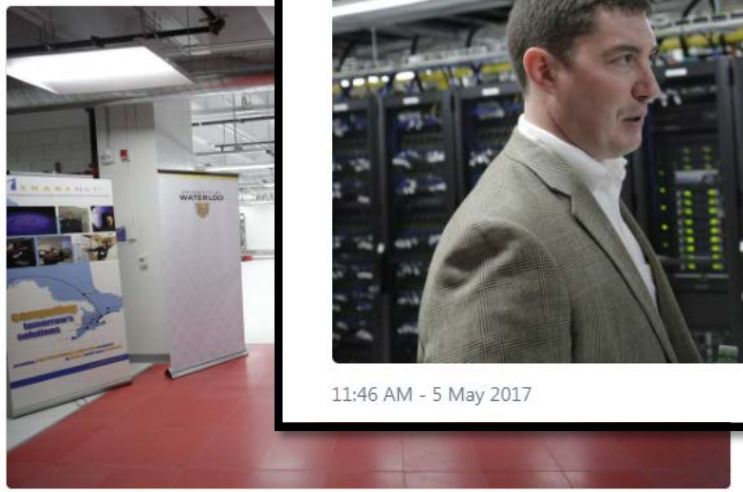
UWaterloo News @UWaterlooNews

@UWa
superc
#super

10:49 AM - 5

Compute Canada @ComputeCanada

Here is the big re
good looking dat
@SHARCNET @Ir
looking powerful



11:44 AM - 5 May 2017

University Waterloo @UWaterloo

#SuperComputerGraham will be h
MC and will be Canada's largest ar
powerful super computer with 1k
33k CPU cores.

11:46 AM - 5 May 2017

Compute Canada @ComputeCanada ·May 5
9/10 Using OpenStack, **#supercomputerGraham** massive storage and batch #HPC offers great cloud computing environment.
1 2

Compute Canada Retweeted

Compute Canada @ComputeCanada ·May 5
7/10 **#supercomputerGraham** is built for #BigData. Can support researchers who are collecting, analyzing, or sharing immense volumes of data.
2 1

Kevin Retweeted

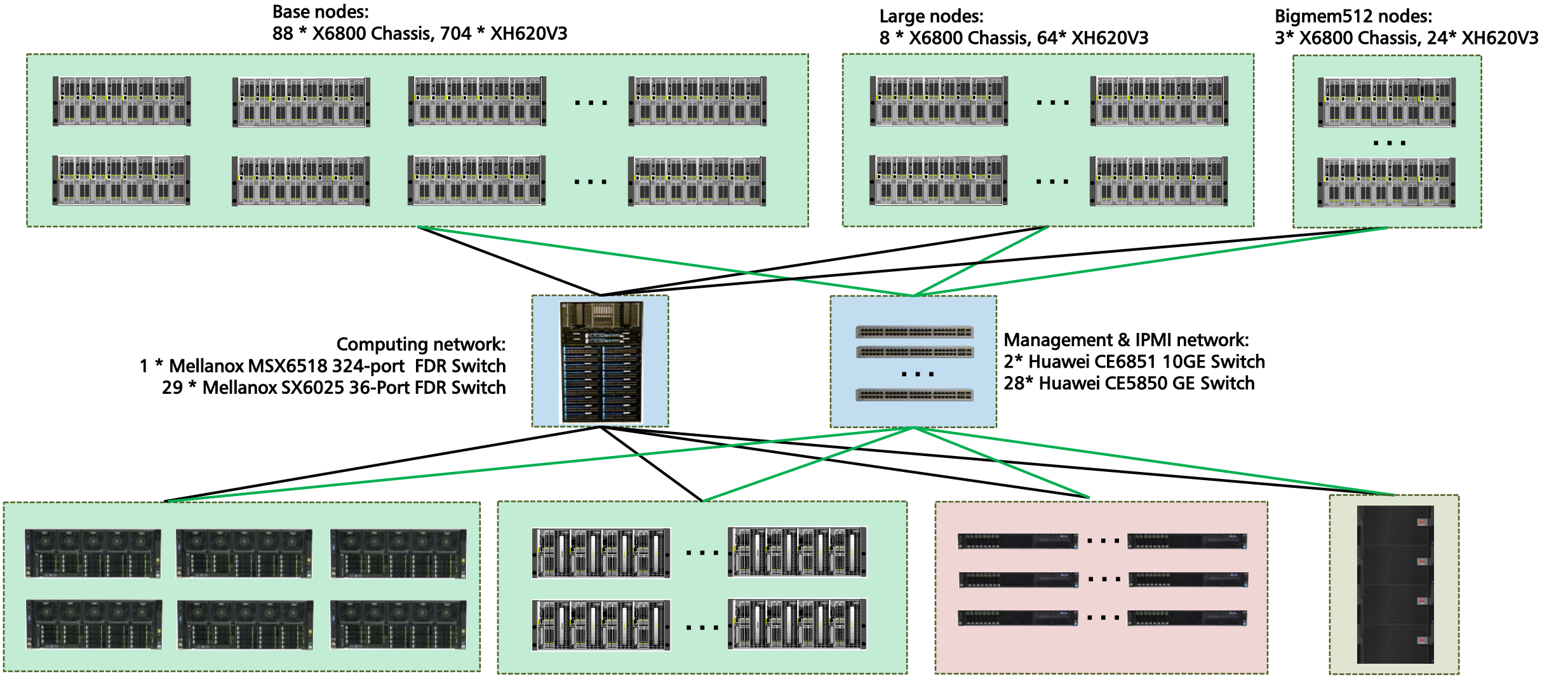
University Waterloo @UWaterloo ·May 5
#UWaterloo & @ComputeCanada to launch **#supercomputerGraham**, the most powerful computer at any Canadian university! ow.ly/ES7O30brhTb
1 17 20

Compute Canada @ComputeCanada ·May 5
1/10 **#supercomputerGraham** @uofwaterloo is Canada's newest supercomputer w/expanded resources for researchers across the country.
1

Graham HPC Cluster @ UWaterloo: Overview

- High-density Servers
- Storage, Switches, Management systems
- 30+ Cabinets
- Liquid cooling
- 33,000 compute cores
- 1,228 TFLOPS (1.2 PFLOPS)

Graham HPC Cluster @ UWaterloo: Internals



THANK YOU