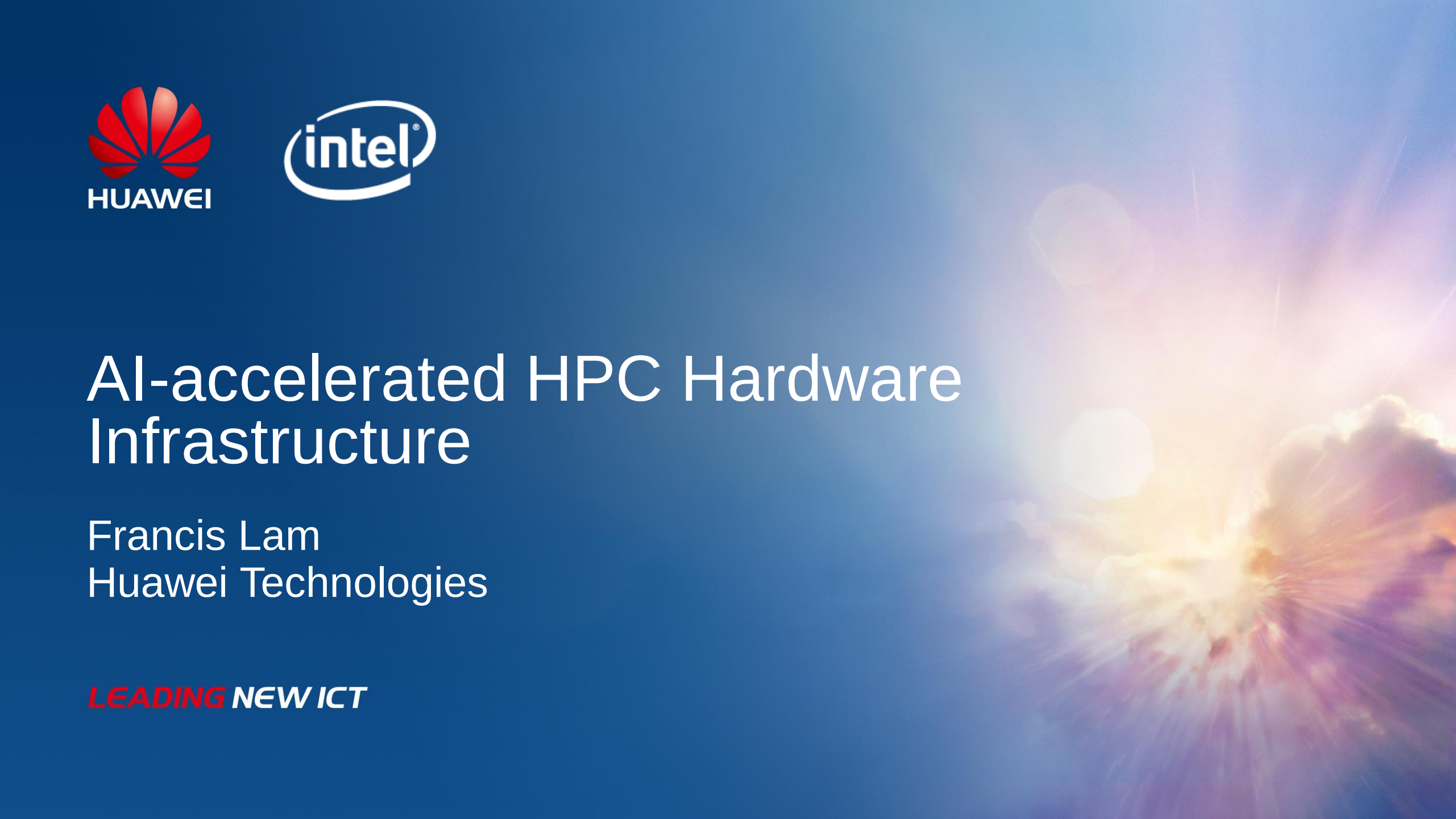


AI-accelerated HPC Hardware Infrastructure

Francis Lam
Huawei Technologies

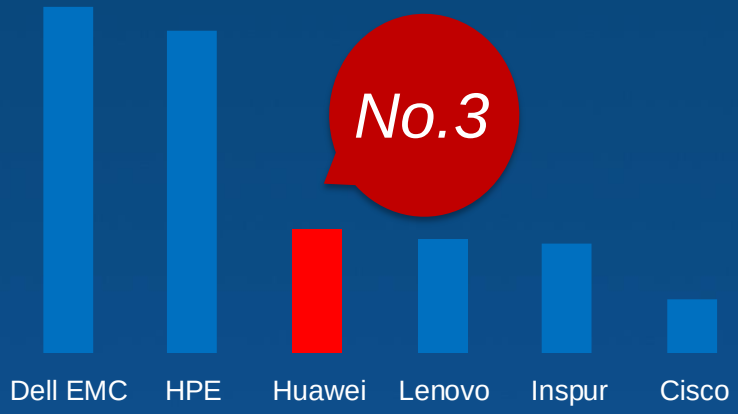
LEADING NEW ICT

Contents

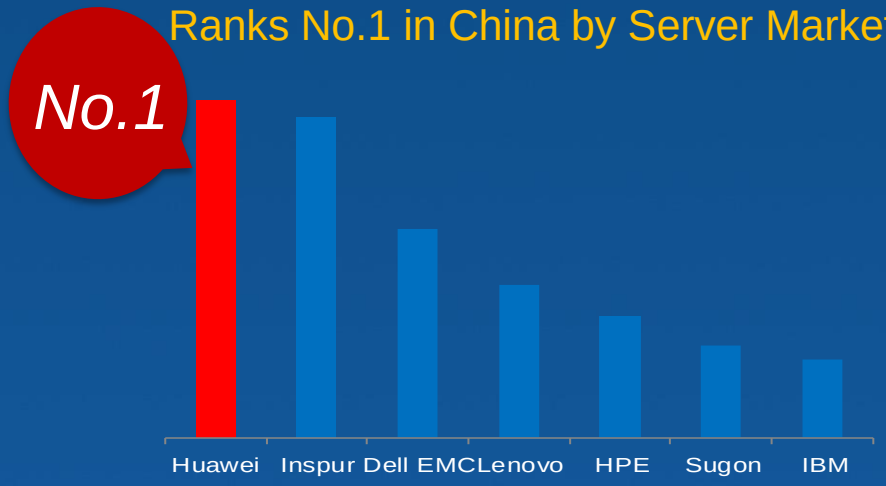
HUAWEI

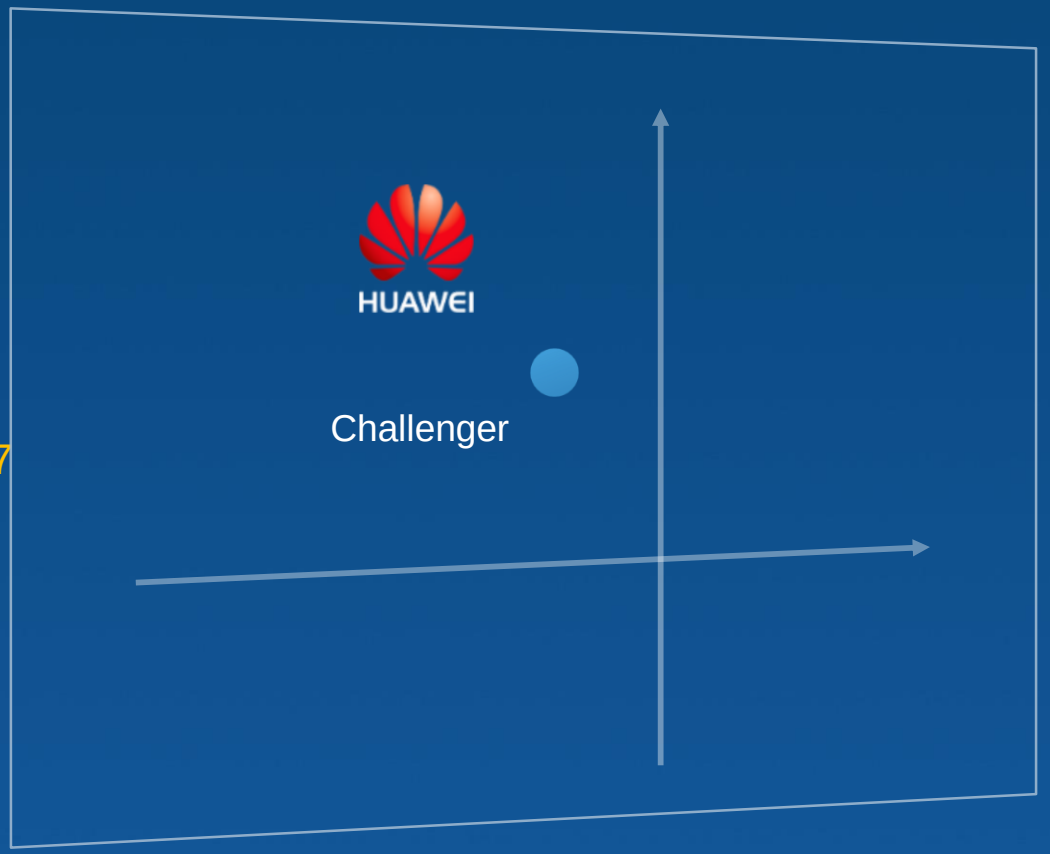
A Rapidly Growing Challenger in Server Market

Ranks No.3 Globally by Server Shipment in 2016Q4-Q3 2017



Ranks No.1 in China by Server Market Share in 2016Q4-Q3 2017





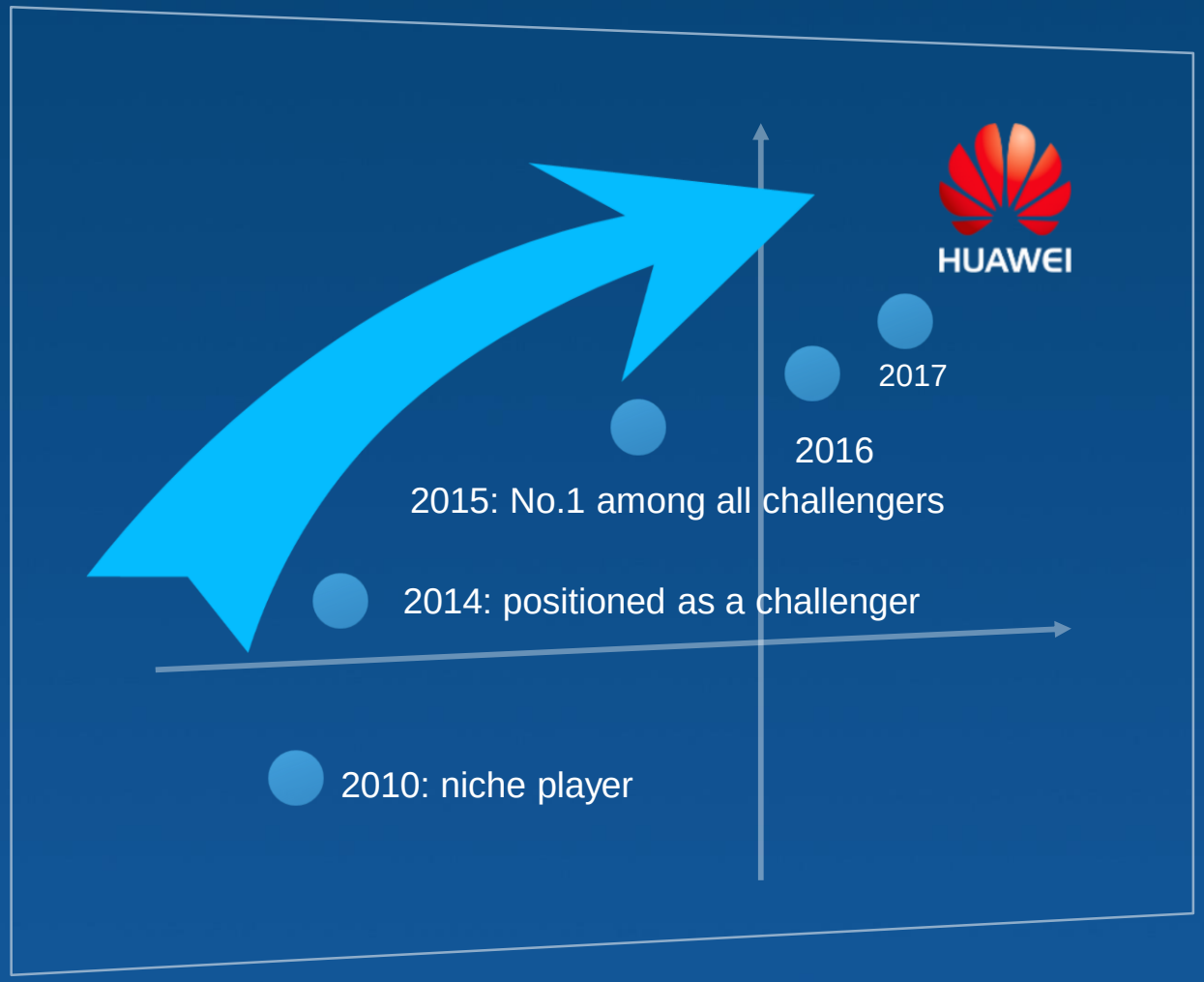
Gartner's Magic Quadrant for Servers

Source: Worldwide Server Market Report in 2016Q4~Q3 2017, Gartner

HUAWEI

A Leader in Storage Market

Source: External Storage Market Share Report in 2Q17, Gartner

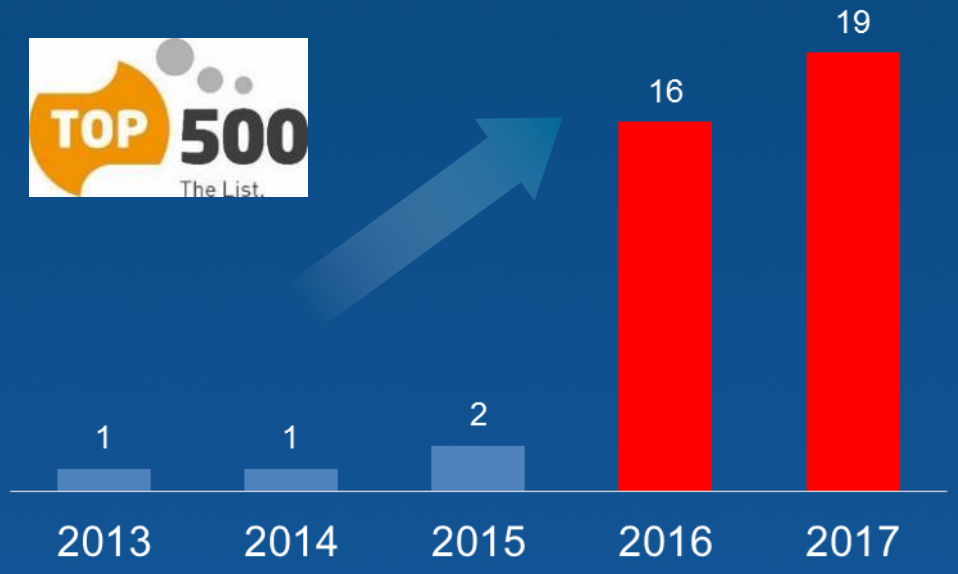


Gartner's Magic Quadrant for Storage

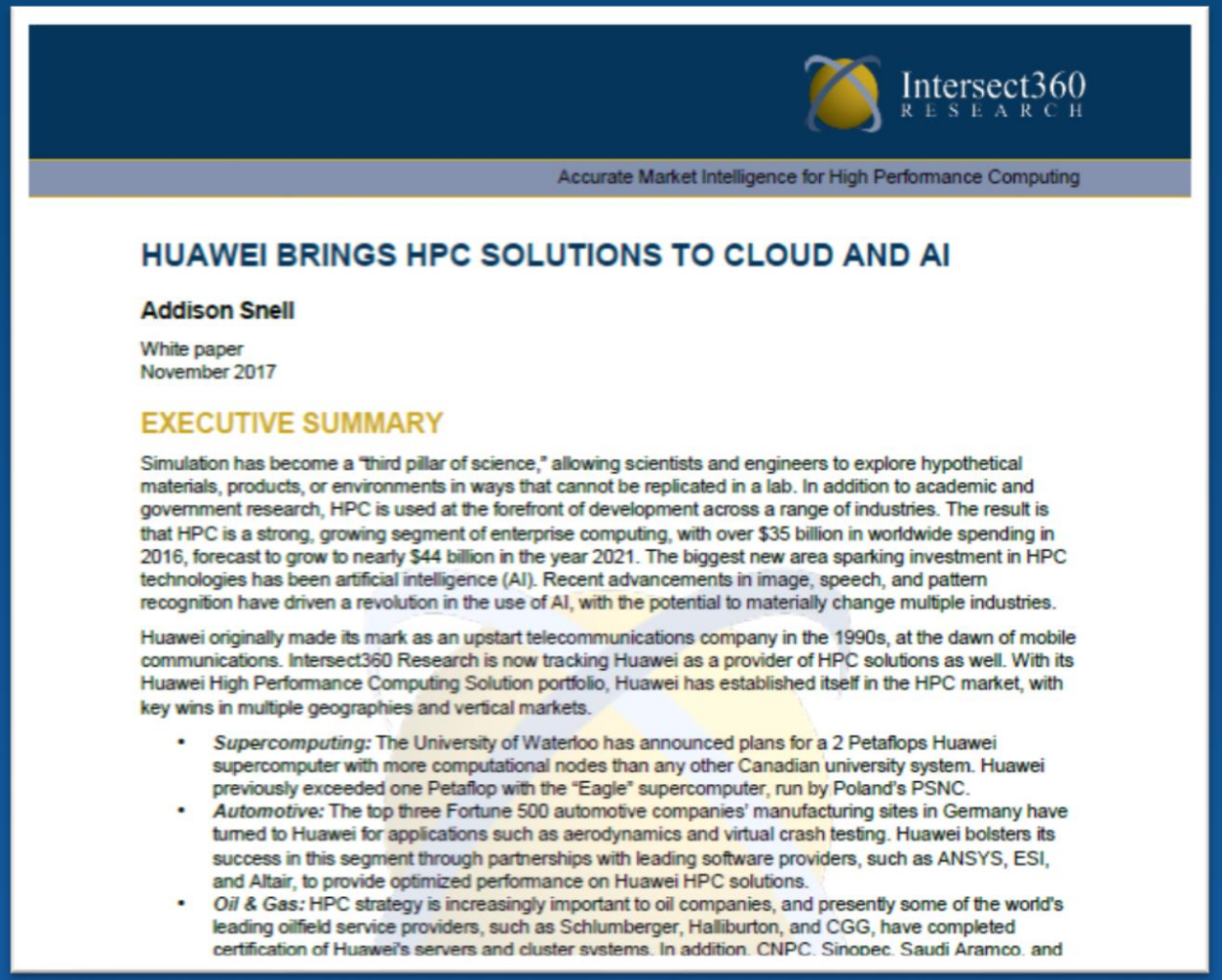
HUAWEI

Growing HPC Influence

TOP500@Huawei



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



Intersect360 RESEARCH

Accurate Market Intelligence for High Performance Computing

HUAWEI BRINGS HPC SOLUTIONS TO CLOUD AND AI

Addison Snell

White paper
November 2017

EXECUTIVE SUMMARY

Simulation has become a "third pillar of science," allowing scientists and engineers to explore hypothetical materials, products, or environments in ways that cannot be replicated in a lab. In addition to academic and government research, HPC is used at the forefront of development across a range of industries. The result is that HPC is a strong, growing segment of enterprise computing, with over $35 billion in worldwide spending in 2016, forecast to grow to nearly $44 billion in the year 2021. The biggest new area sparking investment in HPC technologies has been artificial intelligence (AI). Recent advancements in image, speech, and pattern recognition have driven a revolution in the use of AI, with the potential to materially change multiple industries.

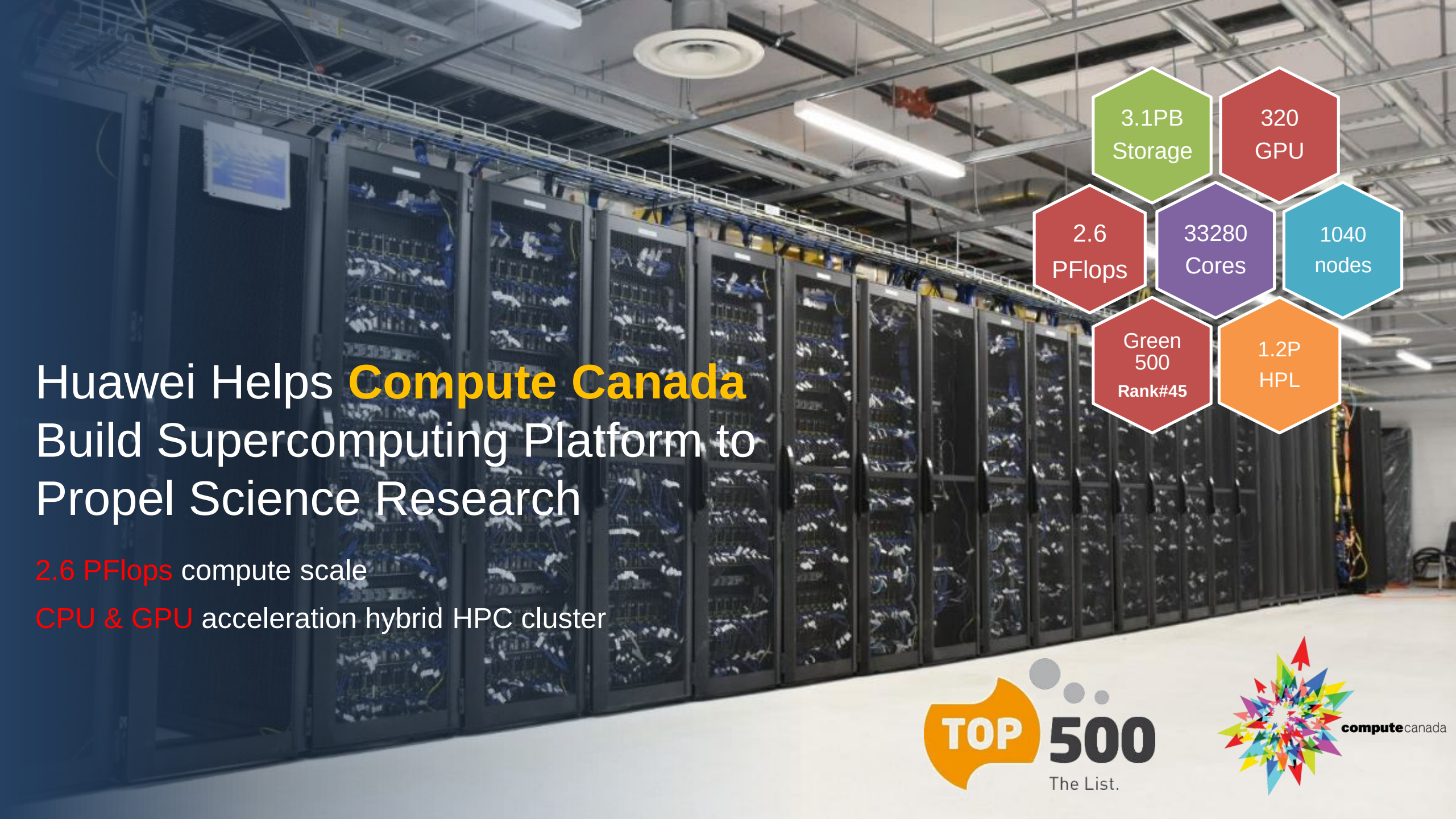
Huawei originally made its mark as an upstart telecommunications company in the 1990s, at the dawn of mobile communications. Intersect360 Research is now tracking Huawei as a provider of HPC solutions as well. With its Huawei High Performance Computing Solution portfolio, Huawei has established itself in the HPC market, with key wins in multiple geographies and vertical markets.

- ***Supercomputing:*** The University of Waterloo has announced plans for a 2 Petaflops Huawei supercomputer with more computational nodes than any other Canadian university system. Huawei previously exceeded one Petaflop with the "Eagle" supercomputer, run by Poland's PSNC.
- ***Automotive:*** The top three Fortune 500 automotive companies' manufacturing sites in Germany have turned to Huawei for applications such as aerodynamics and virtual crash testing. Huawei bolsters its success in this segment through partnerships with leading software providers, such as ANSYS, ESI, and Altair, to provide optimized performance on Huawei HPC solutions.
- ***Oil & Gas:*** HPC strategy is increasingly important to oil companies, and presently some of the world's leading oilfield service providers, such as Schlumberger, Halliburton, and CGG, have completed certification of Huawei's servers and cluster systems. In addition, CNPC, Sinopec, Saudi Aramco, and

HUAWEI

PSNC Builds an Energy-efficient Data Center with Huawei HPC Solution

- 1040 nodes, with compute performance of 1.37 PFlops
- Energy consumption down by 90%, saving electricity costs by 38%
- Liquid-cooled cluster, with cooling PUE of 1.05



Huawei Helps **Compute Canada** Build Supercomputing Platform to Propel Science Research

2.6 PFlops compute scale

CPU & GPU acceleration hybrid HPC cluster

TOP 500
The List.

computecanada

Huawei Servers Empower **Daimler-Benz** with an Automobile Fluid Simulation HPC System

HPC cluster application performance up by 30%

Compute density doubled

Huawei Helps **Volkswagen Group** Build a Car Collision Simulation Design Platform

Collision test cycle slashed by 10%

Huawei Helps CERN Build a Science Cloud — the "Helix Nebula"

Compute power up by 30%

O&M costs down by 67%

GLOBAL FOUNDRIES

Agency for Science, Technology and Research
SINGAPORE

NUS National University of Singapore

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Trusted Partner in Deploying HPC Infrastructure

Universities

Manufacturing/
Automobile Enterprises

Oil & Gas, Media,
Meteorology

Supercomputing Centers
& Scientific Institutions

HUAWEI

Contents

A New Starting Point in Computing

10x heterogeneous vs universal computing

100x dedicated vs universal computing

AI

Universal → Heterogeneous

1M+ nodes

1000x data processing capabilities (TB → PB)

Cloud Computing

Single server → Resource pool

100bn connections (by 2025)

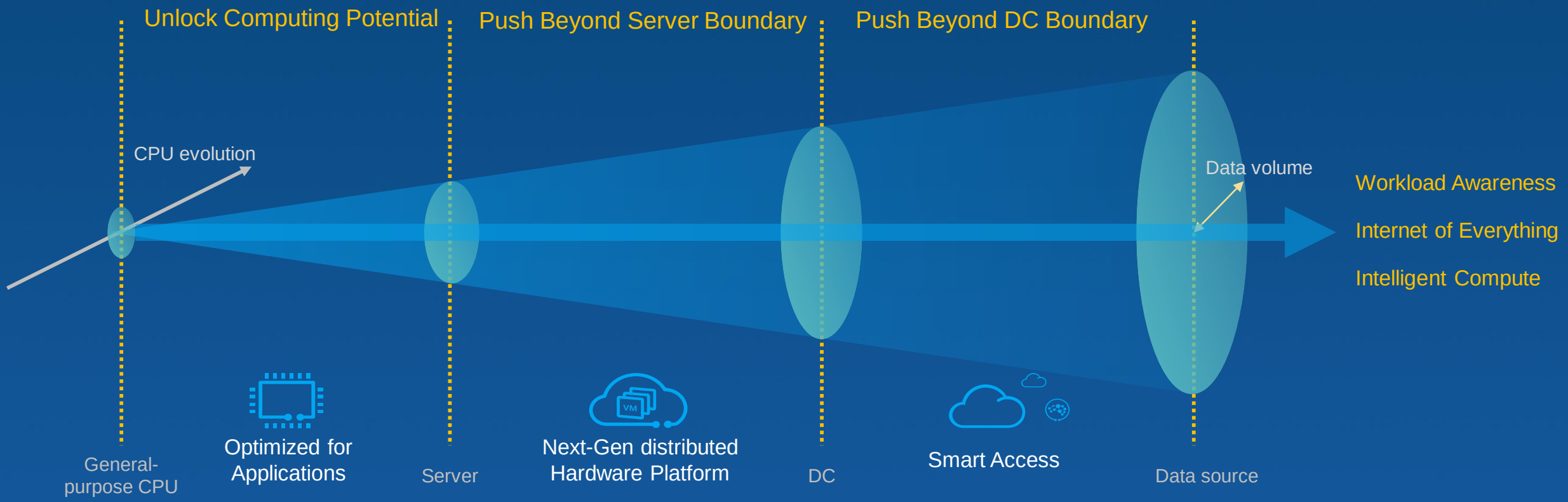
50% data processed on the edge

Smart Edge

Centralized → Distributed

Boundless Computing Strategy

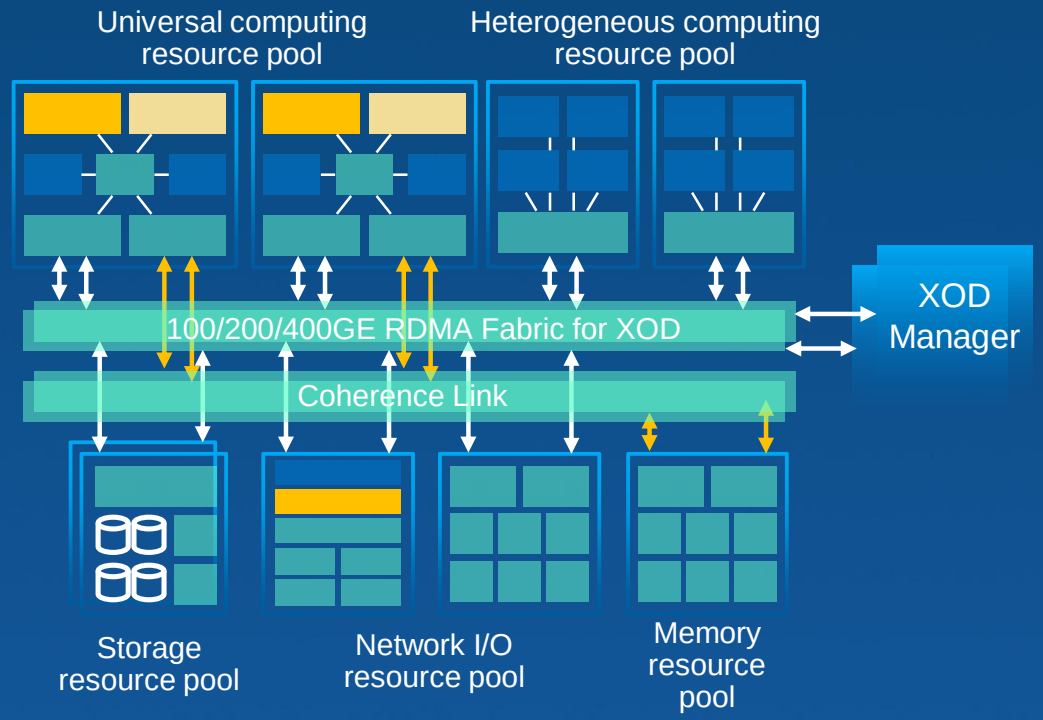
Re-imagine computing for a fully connected world.



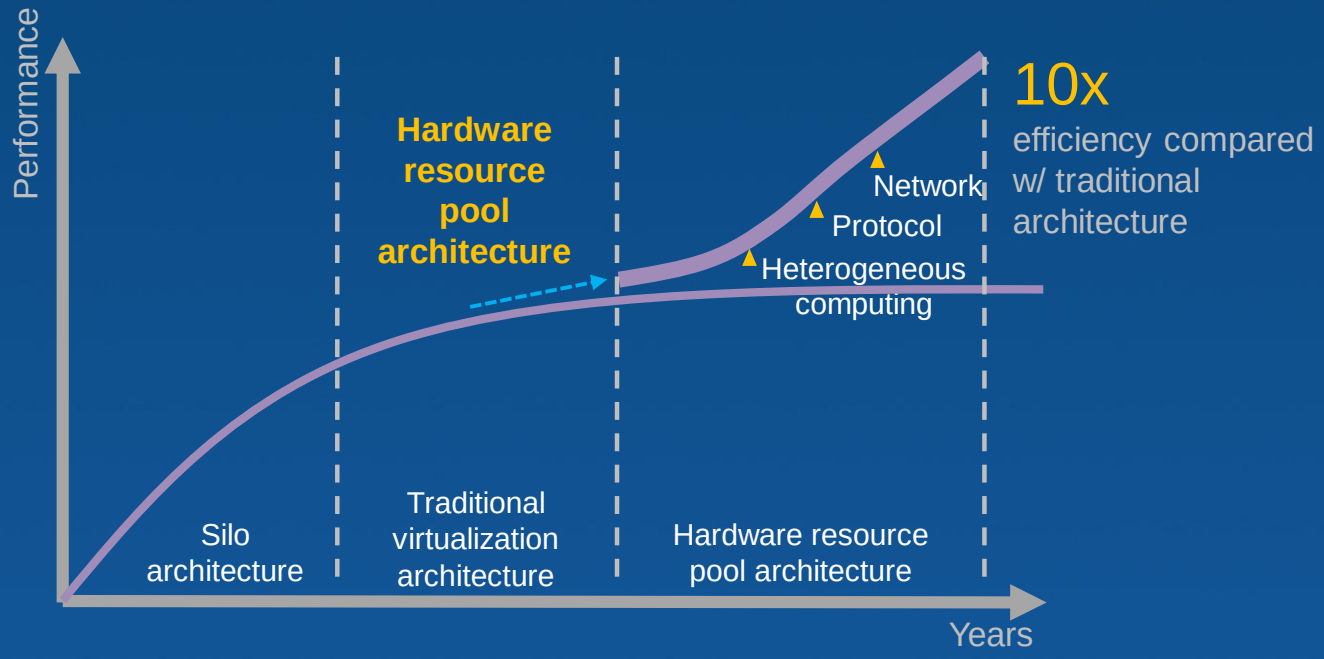
HUAWEI

Push Beyond the Boundary

Enable cloud usage with ultimate accessibility and convenience for DCs.



Hardware resource pooling enables provisioning on demand.



Hardware resource pooling combines with chip capabilities to boost overall efficiency by 10x.

HUAWEI

Contents

Strategic partnership with selected Tier 1 international carriers in public cloud development

Global DCs & Network Nodes

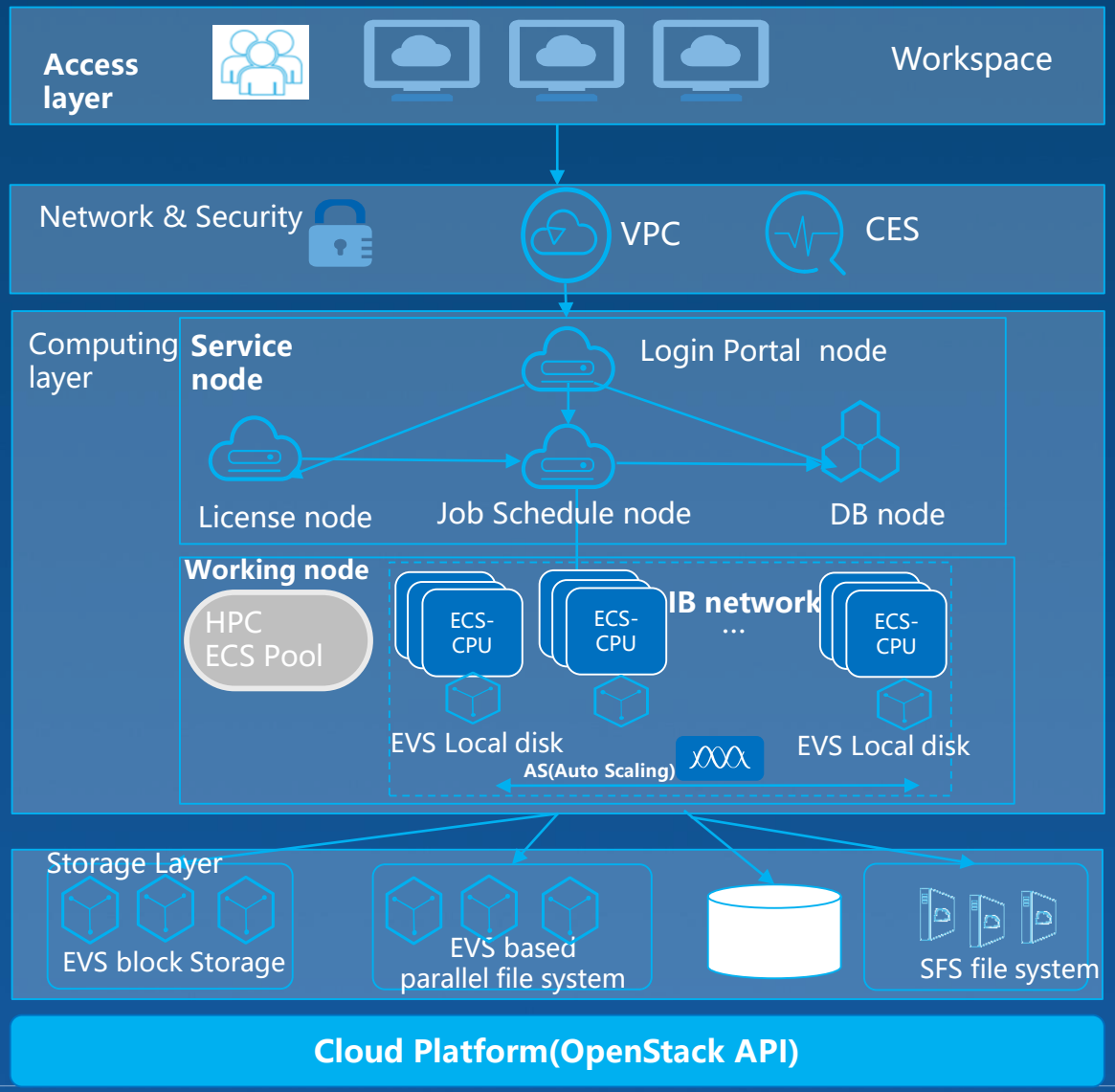
5000+ Salesman

10k+ Engineers

4 + Yrs Public Cloud Operation Experience

10k+ Public Cloud Customers from Top500 to SMB

HPC Cloud： True High Performance Computing Infrastructure



Bare metal service

Bare metal + SDI
Shared storage

VMs with high specifications

128vCPU+4TB RAM

100G IB Service Network
2μs ,Low Latency

GPU acceleration

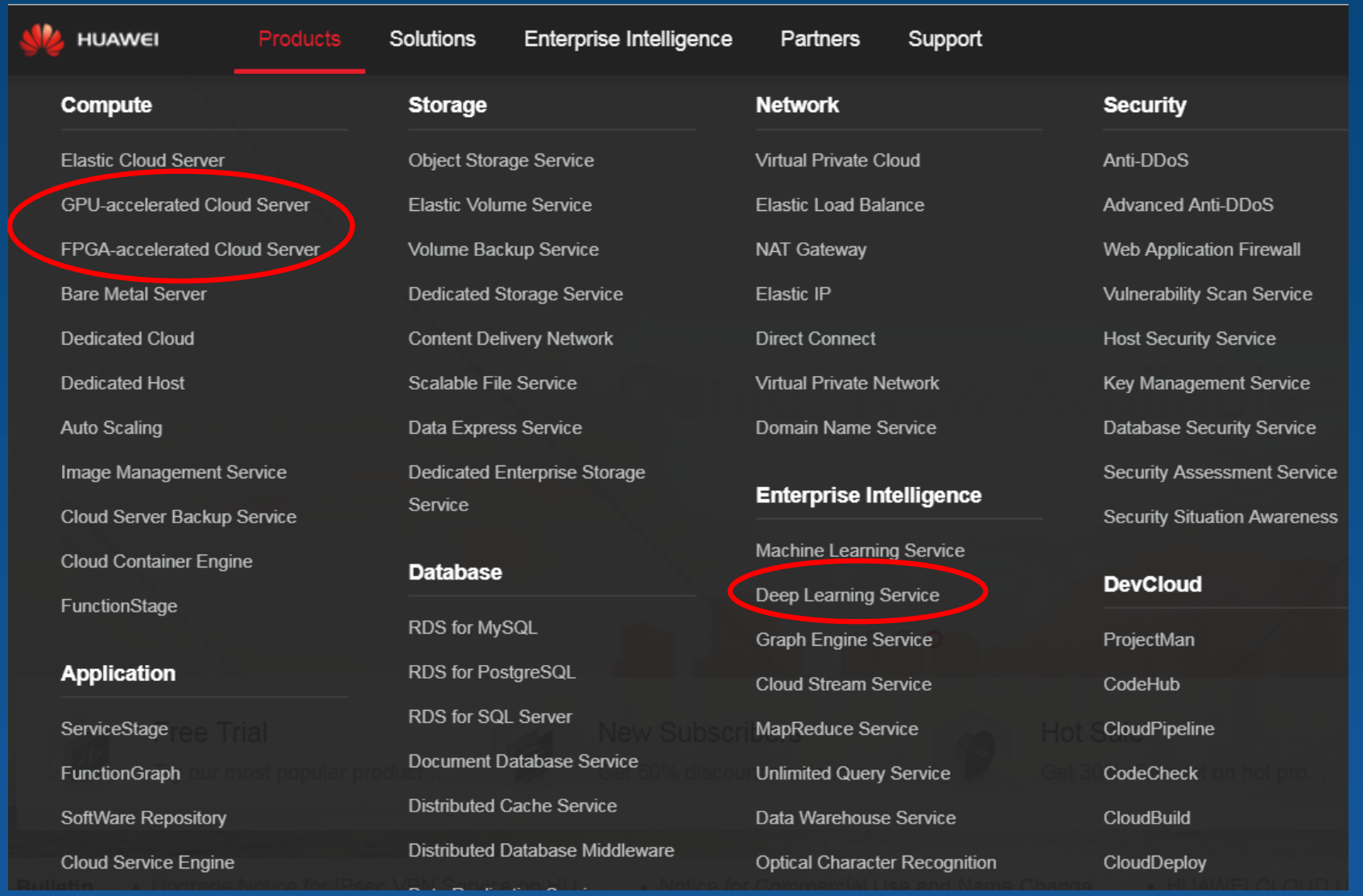
Nvidia P100
GPU Acceleration

FPGA data pre-processing

Optimal cloud acceleration and data pre-processing

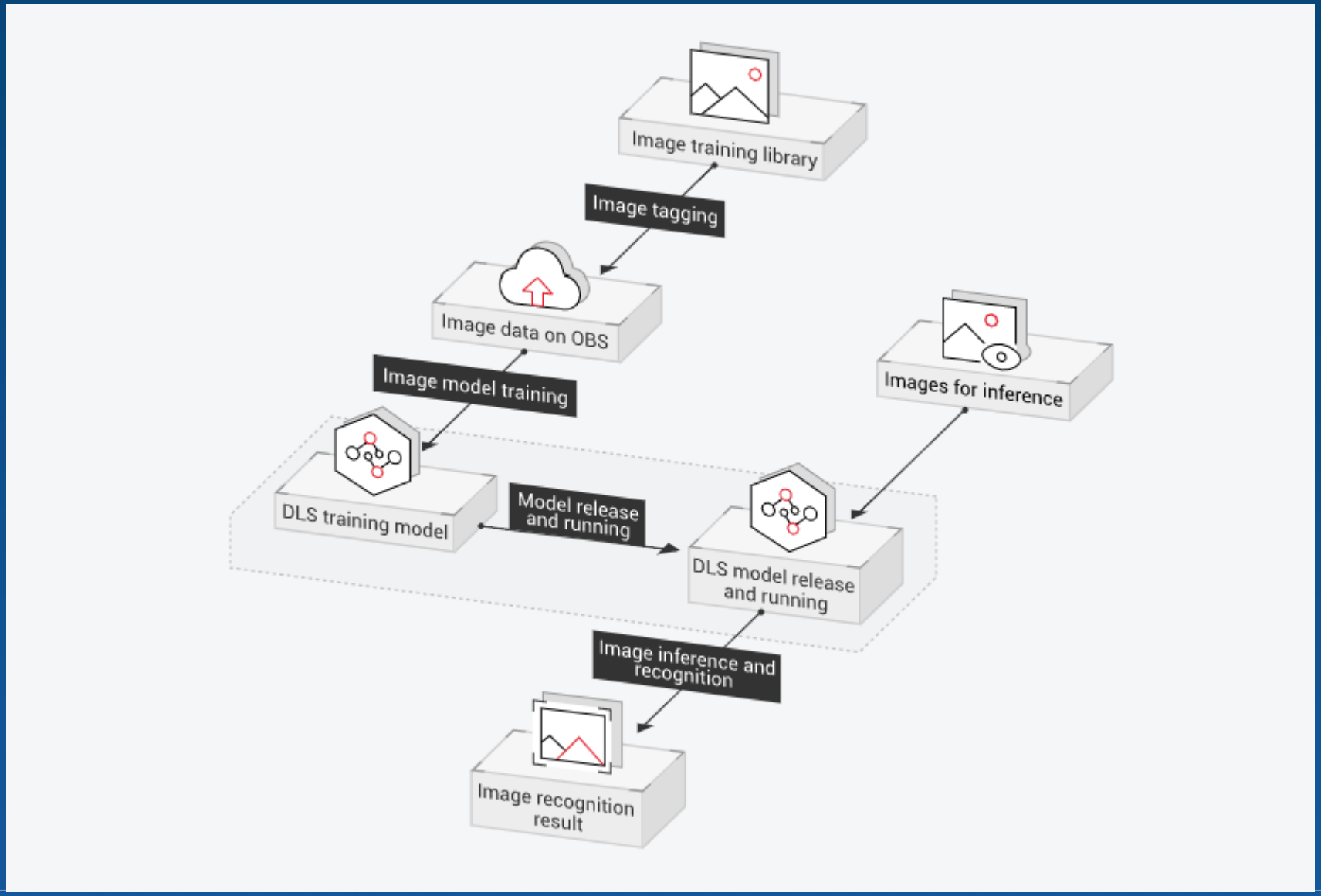
HUAWEI

Huawei Enterprise Cloud Services

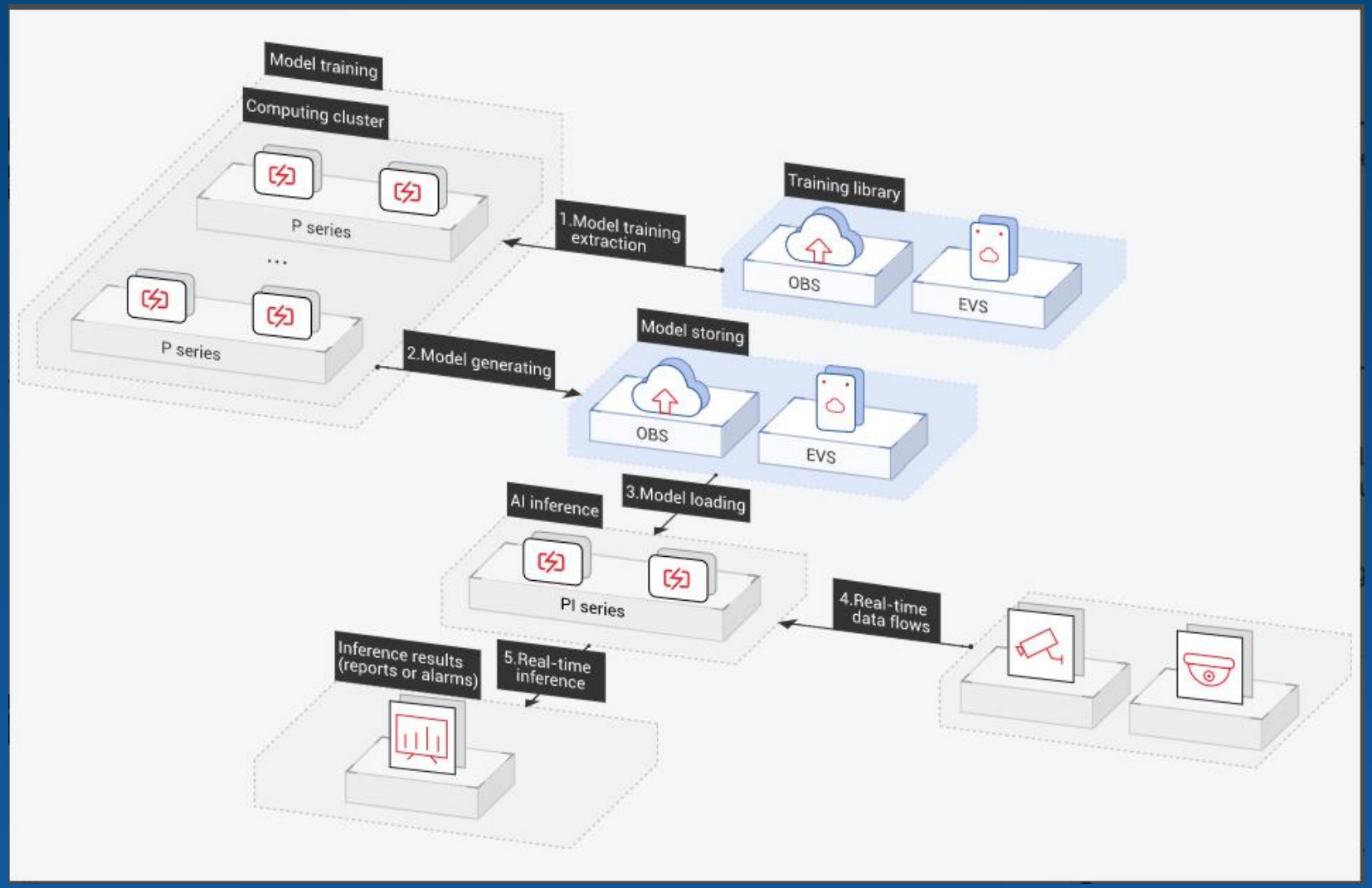


HUAWEI

Huawei Enterprise Cloud DLaaS



Huawei Enterprise Cloud GPUaaS



New Huawei Cloud Services Accelerating AI and HPC

P1 GPU Calculate Accelerated Cloud Server

8*P100

NVLink

Upcoming

Bandwidth 32G↗160G

FusionServer G5500 (8 x P100)

PI1 GPU Inference Accelerated Cloud Server

16*P4

(2017Q4)

16*22T INT8 computing power

FusionServer G5500 (32 x P4)

FP1 FPGA Accelerated Cloud Server

8*VU9P Ultra SCALE

200G MESH

30+ Partners

30+ Accelerate IP

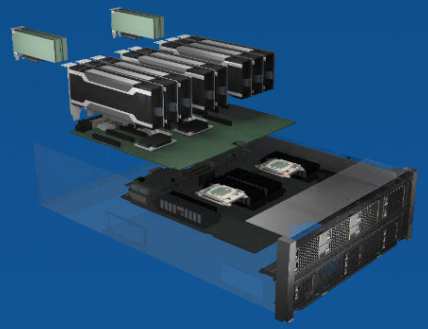
FusionServer G5500 (8 x FPGA-VU9P)

ATLAS Heterogeneous Computing Platform

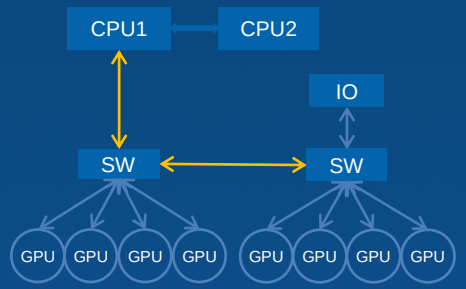
Heterogeneous resource pool

One-click topology switching

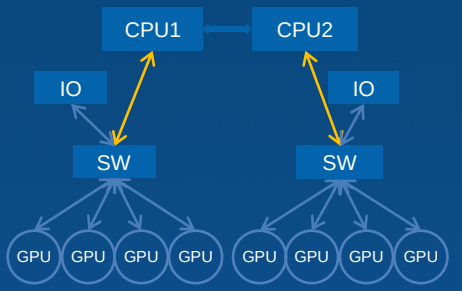
Flexible Modular Design



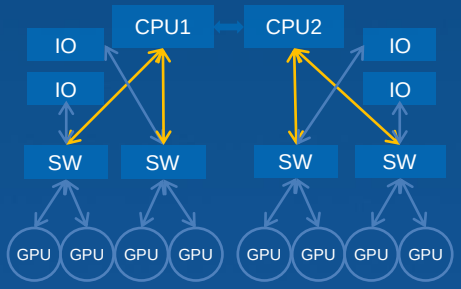
FusionServer G5500



Topo1: Single RC-AI Training



Topo2: Balanced –HPC, Cloud

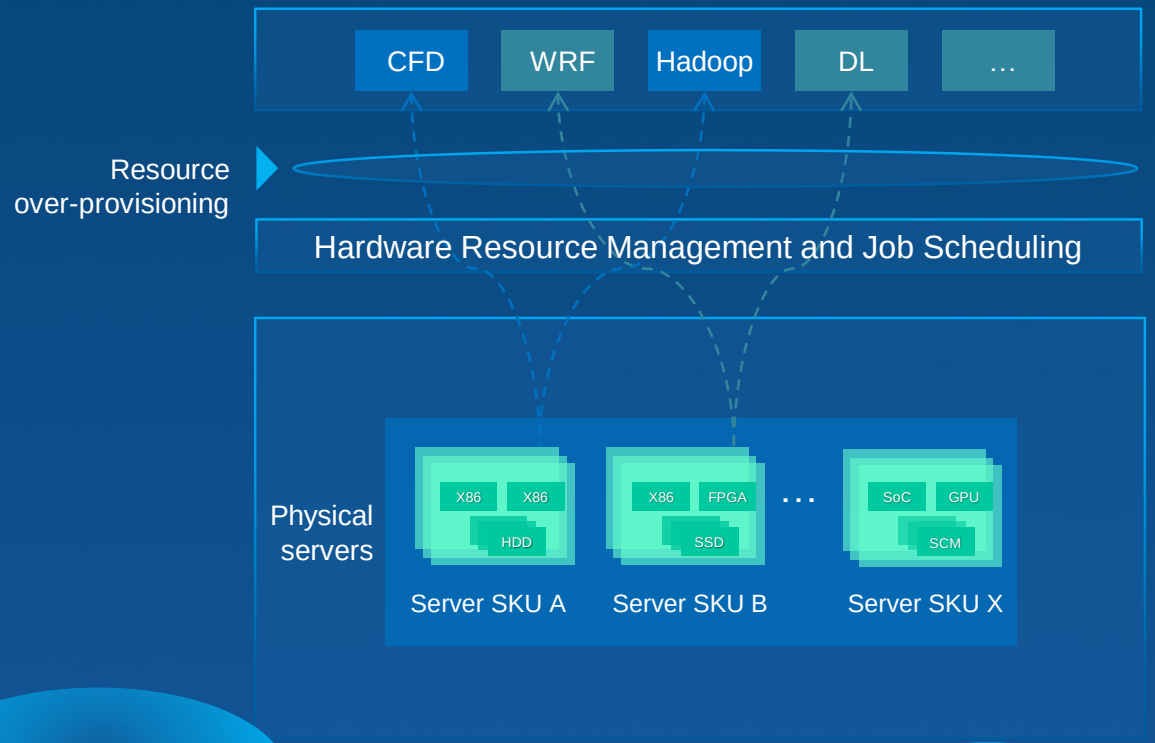


Topo3: High BW-HPC

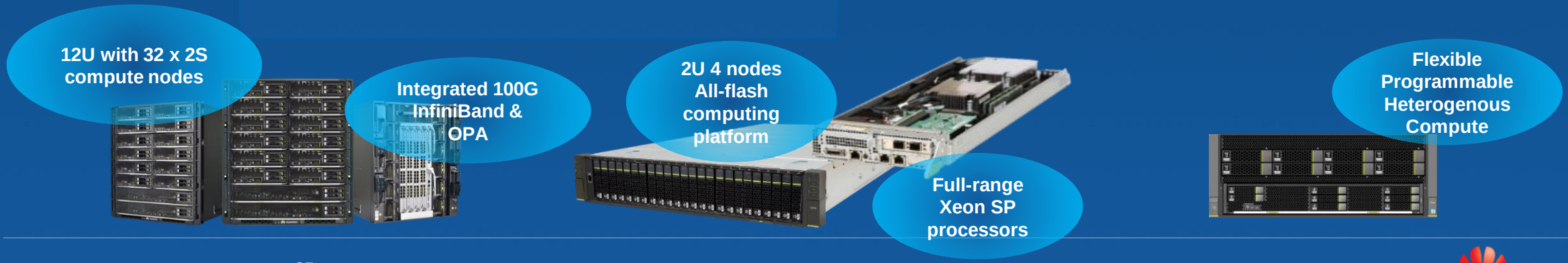
...

HUAWEI

Integrating HPC and AI – Hardware Perspective

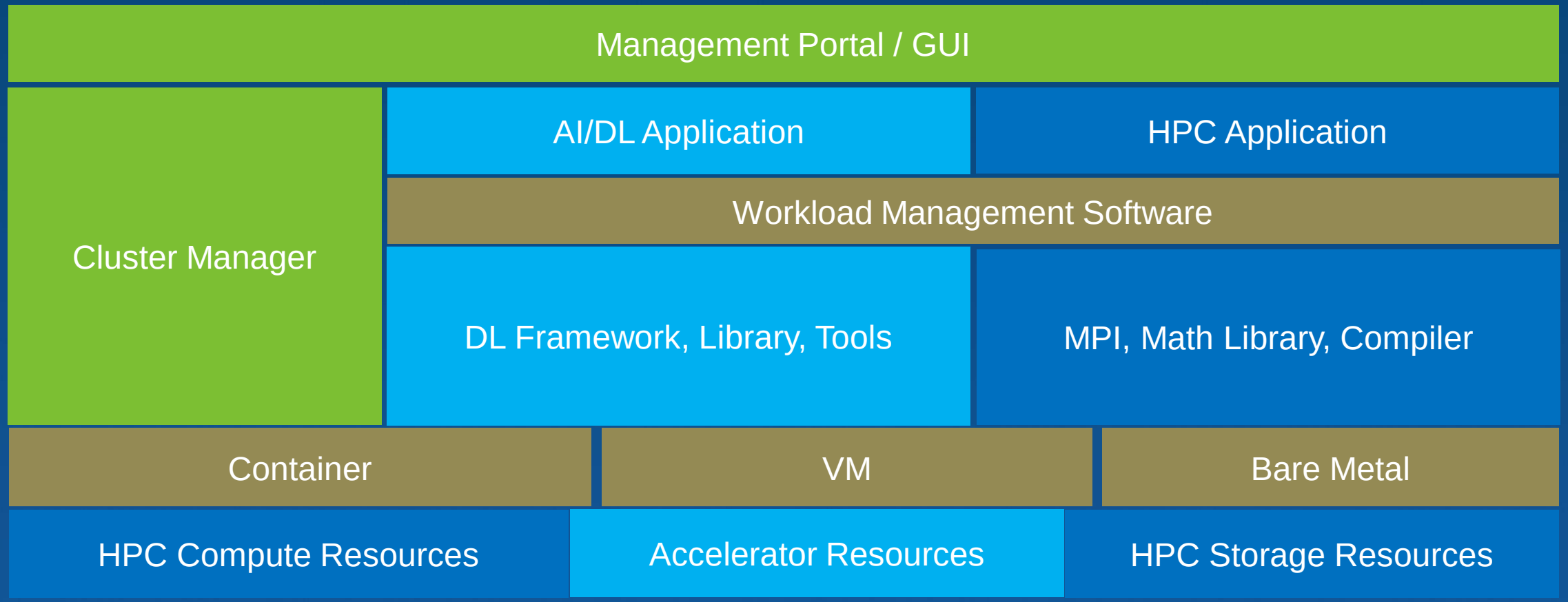


- Built world class HPC and AI hardware platform enabling higher degree of workload optimization
- Built ecosystem of solutions providers to facilitate advanced HPC deployment in both academic and industry
- Leverage diversified set of technology to advance HPC



HUAWEI

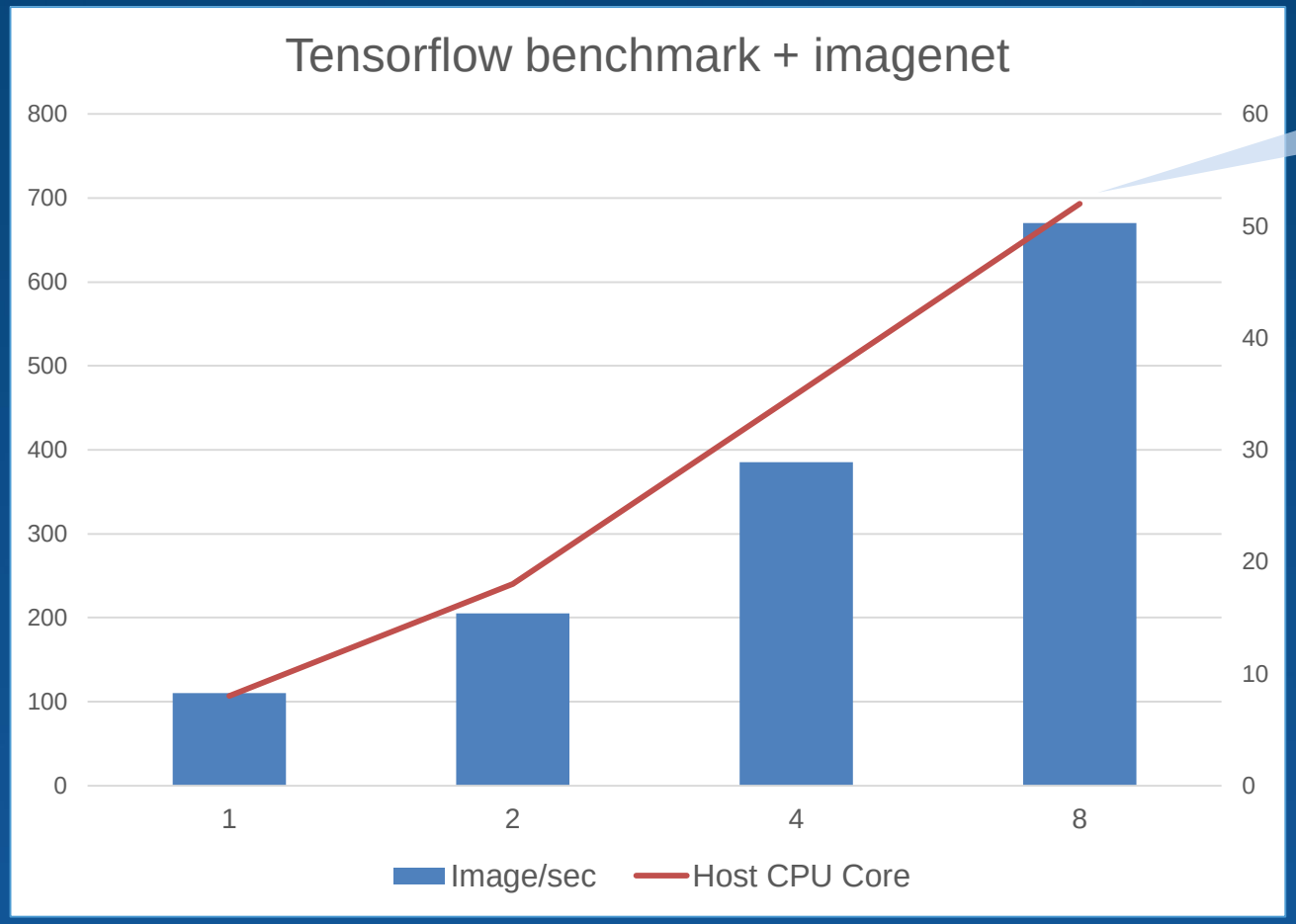
Integrating HPC and AI – Software Perspective



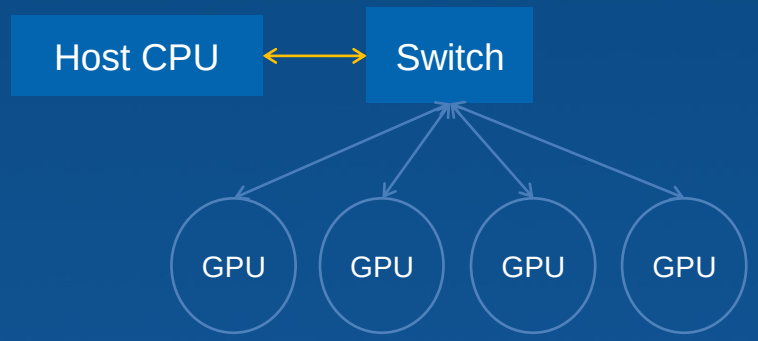
Contents

HUAWEI

Diversified Workloads Demand Hardware Optimization



High Host CPU Consumption

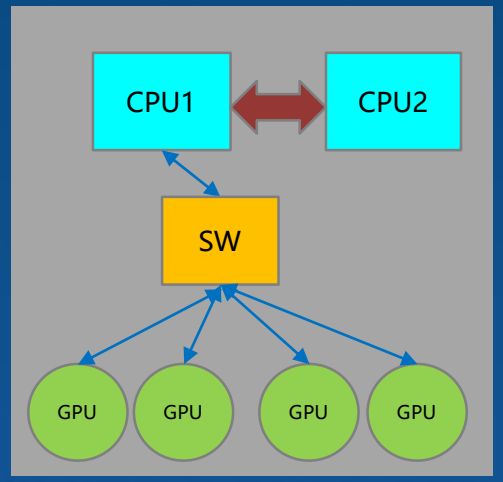


Accelerator hardwired to Host CPU

HUAWEI

Different algorithms Required Varying GPU Resources

Fix GPU Configuration



VS

Actual Workloads

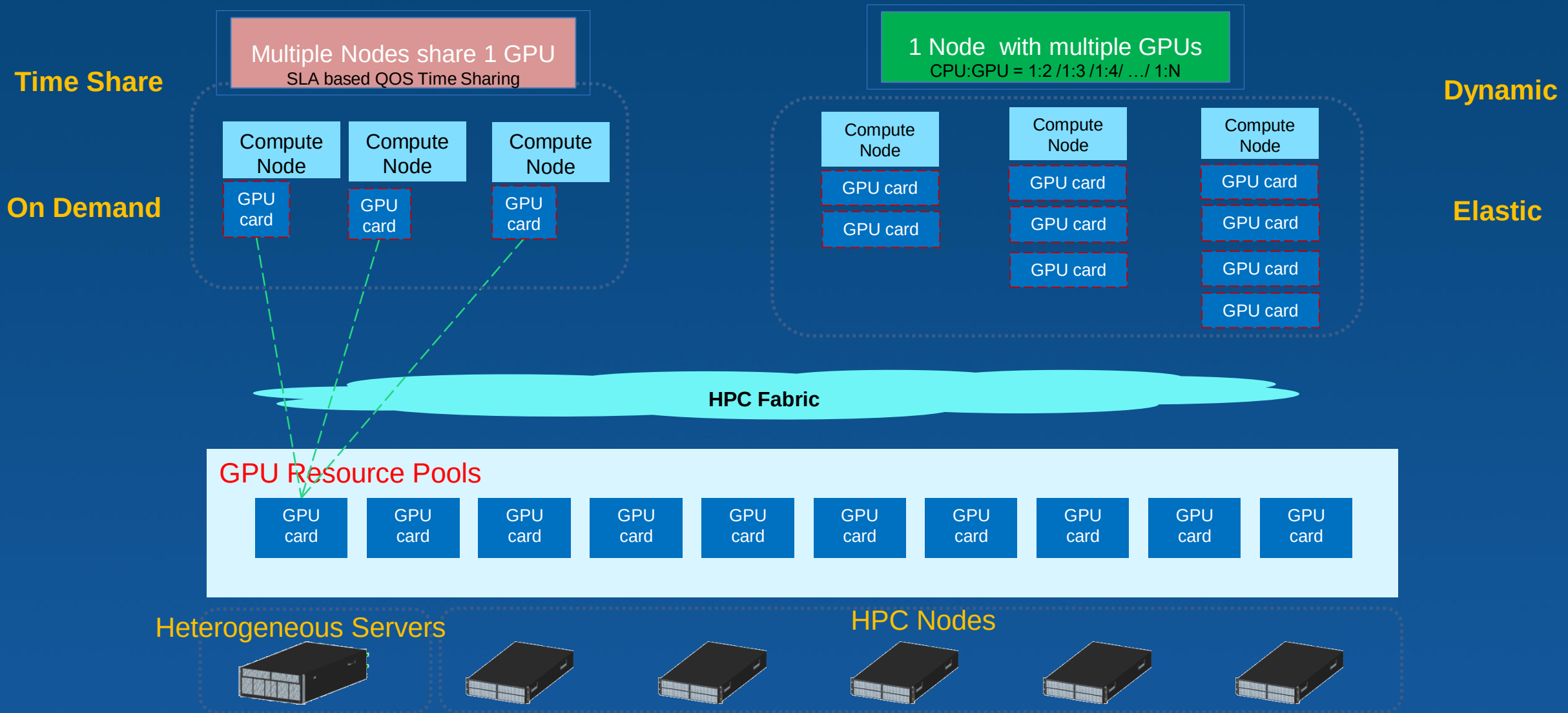
Algorithms	Ideal CPU：GPU
License Plate Recognition	1:2
Vehicle/Cyclist/Pedestrian Detection	1:4
Vehicle Structural Inspection	1:6

Example: Safe City Applications

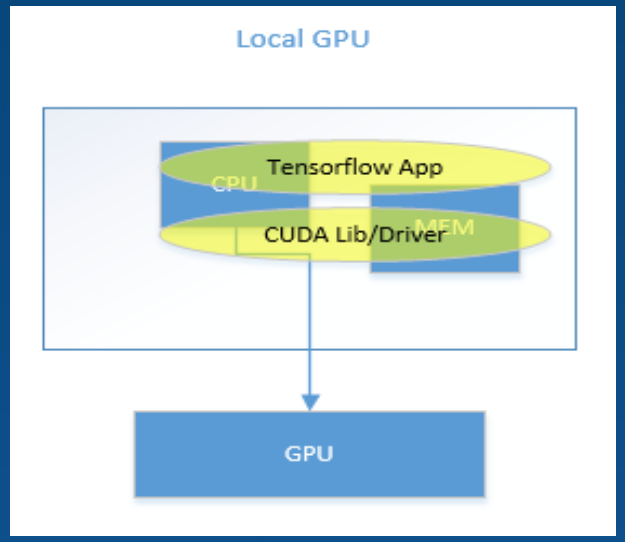
Wasted 2 GPU

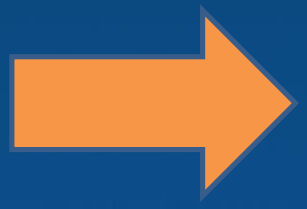
Short 2 GPU

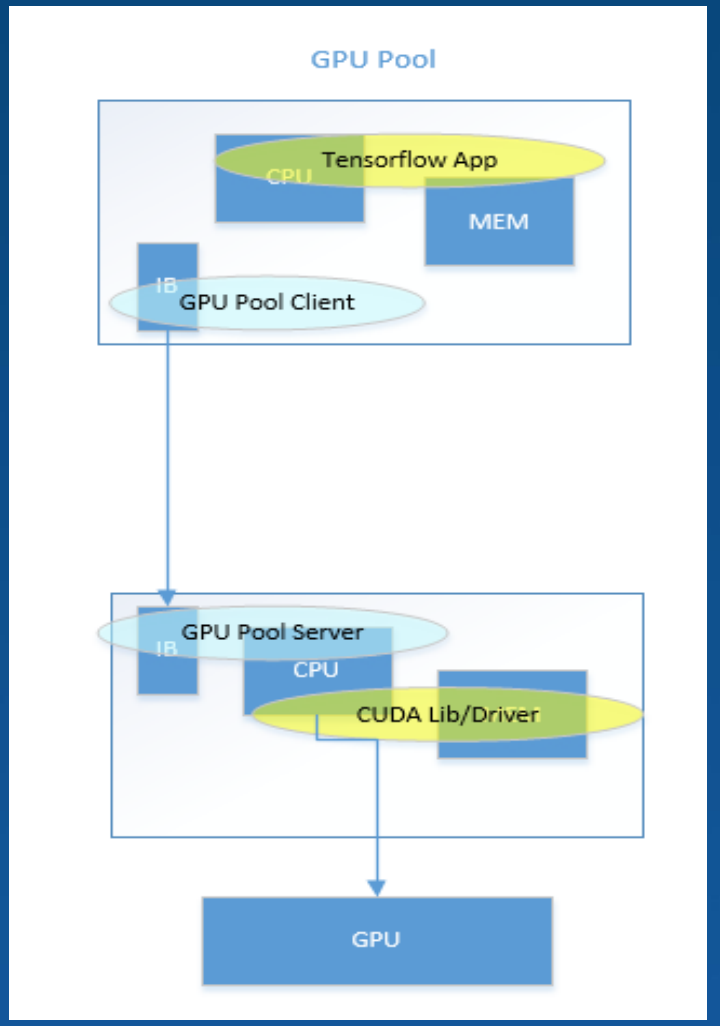
Maximize Utilization with GPU Resource Pool



GPU Resource Pool Improves Resource Utilization

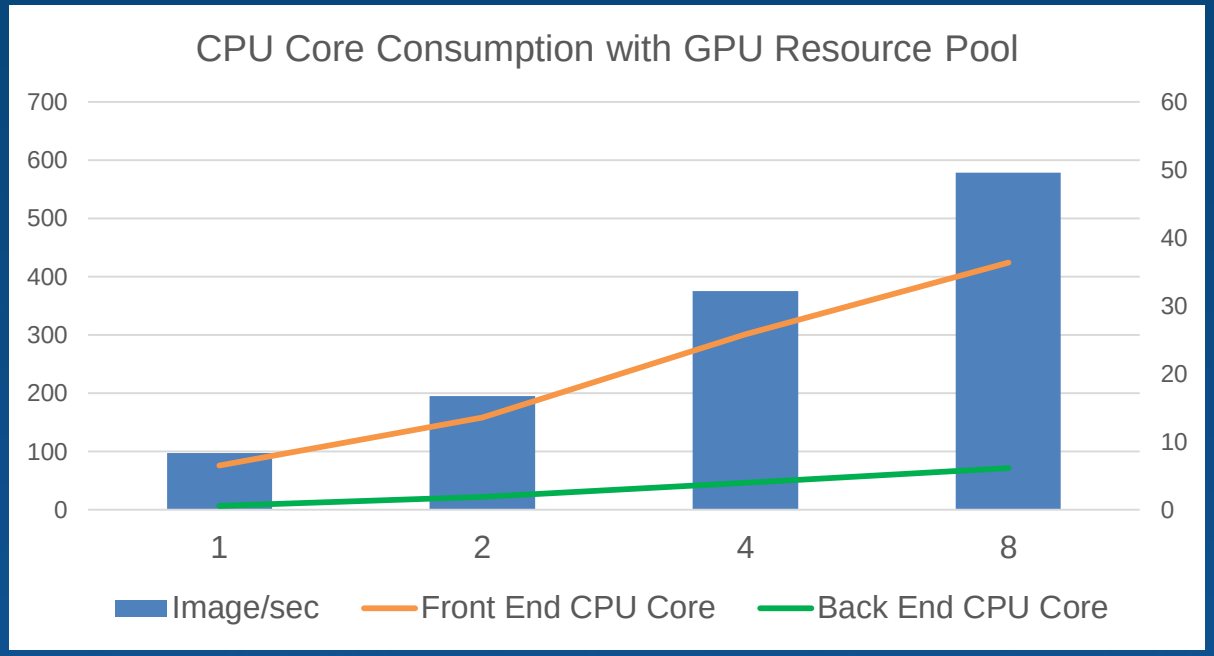


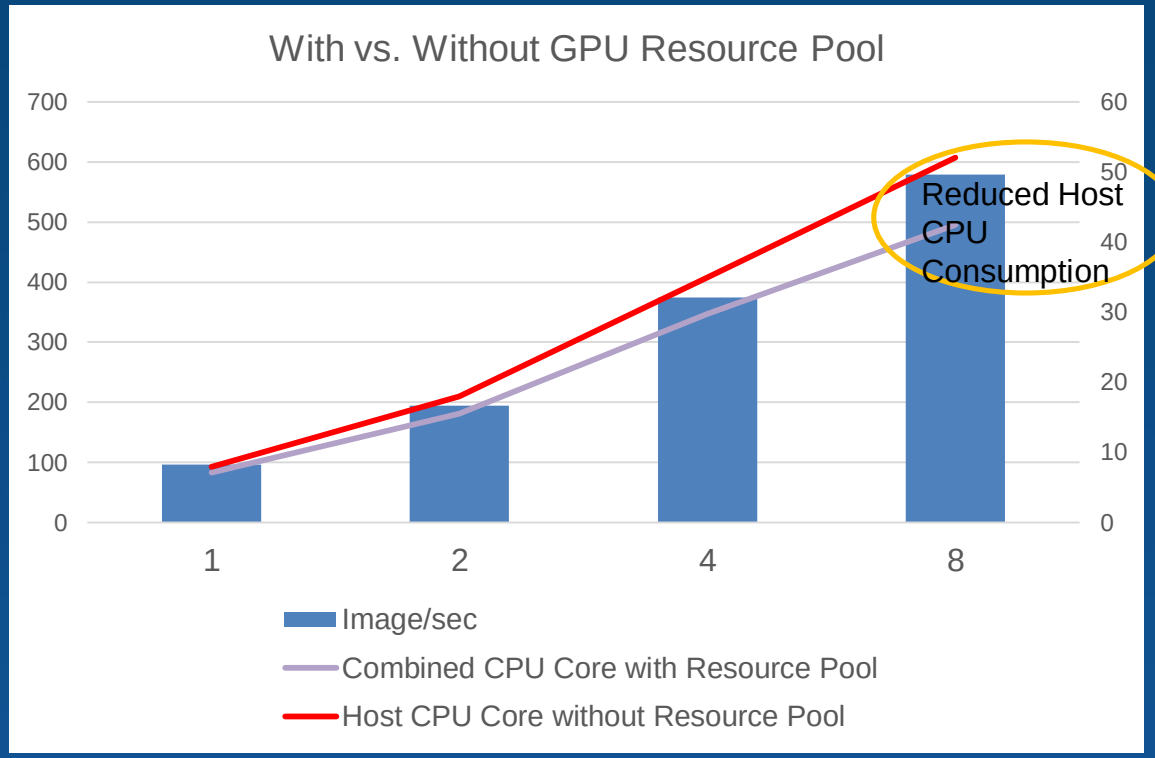


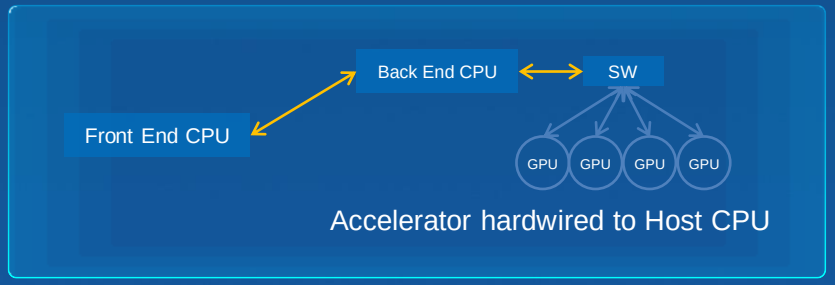


- More effectively sharing GPU resources among tasks
- Reduce compute resource consumption for intensive workload

Adapting to Workloads with Accelerator Resource Pool

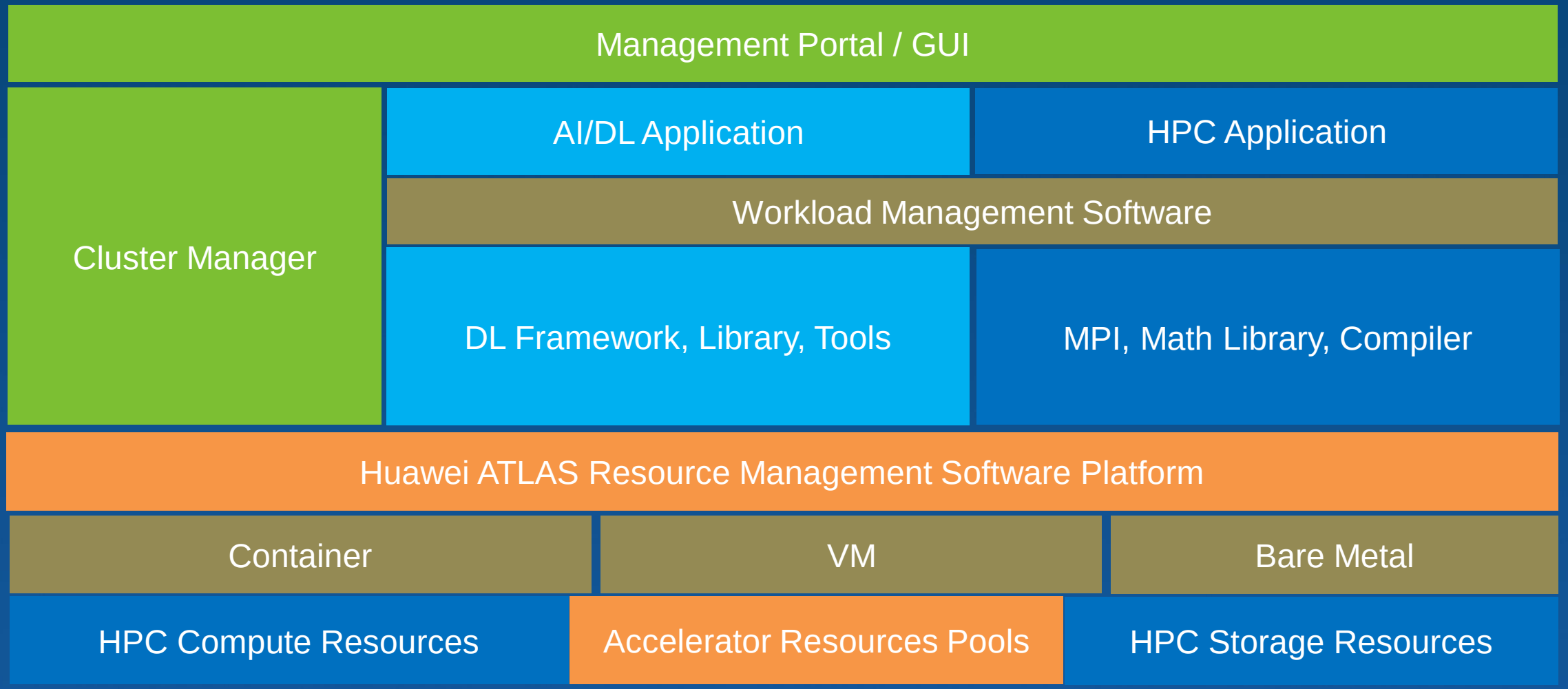




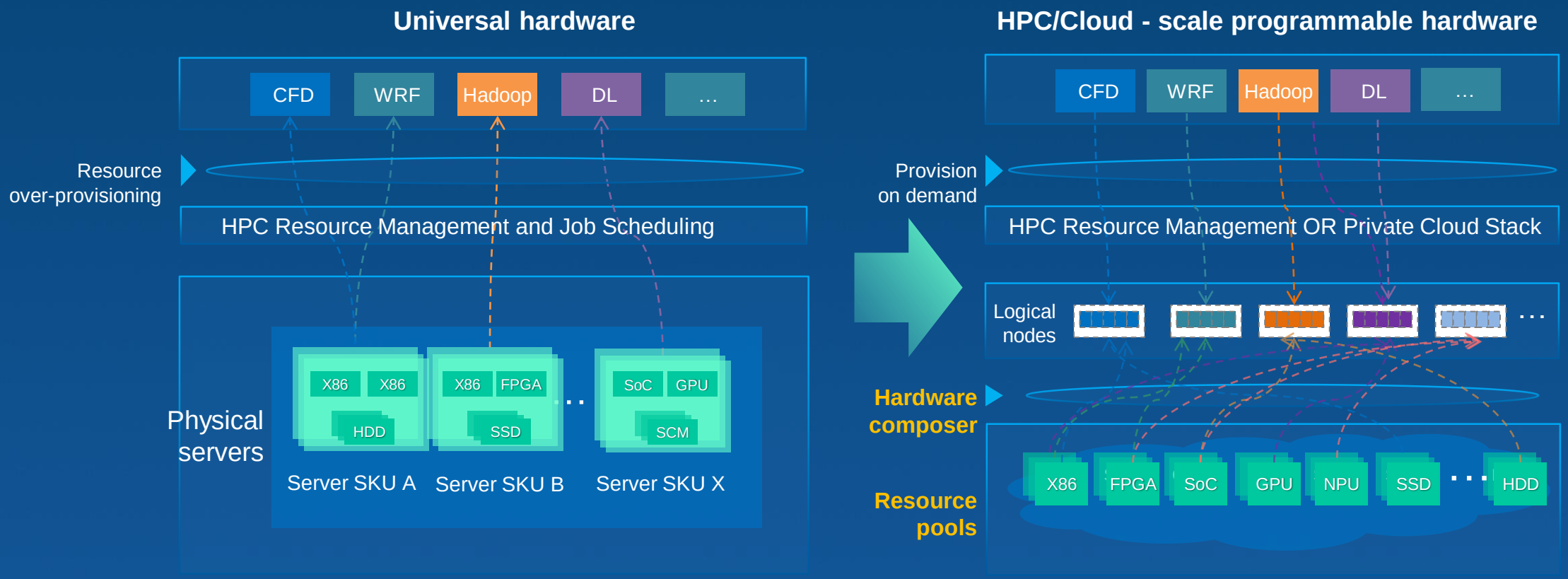


HUAWEI

Unified HPC and AI architecture



ATLAS Programmable Hardware Architecture

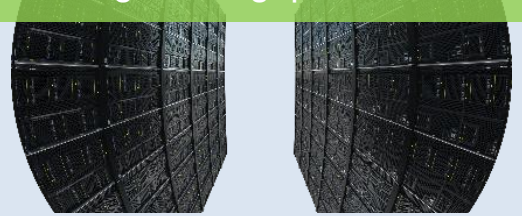


- Flexibly programmable, low-latency converged network
- Everything On Demand (XOD) protocol enables on-demand smart converged resource pool
- Smart converged resource pool management, automatic composition: provision servers of any types flexibly according to service requirements

AI From Core To Edge

Data Center
AI Training/Inference/HPC

- Data set training
- Image, voice, video structured query
- Cluster size, large throughput



FusionServer G5500
Data Center Heterogeneous Server

Edge Computing
Branch Inference

- Identify, reason, solve problems
- Unstructured data processing
- Real-time accurate response

FusionServer G2500
Intelligent Edge Server

Terminal

Upcoming Thin
Intelligent Edge Server

Huawei HPC Advantages

End-to-end Efficiency

Smaller footprint, lower power,

Higher performance

- E2E energy-saving design
- Efficient and reliable liquid-cooling technology
- Integrated delivery and installation

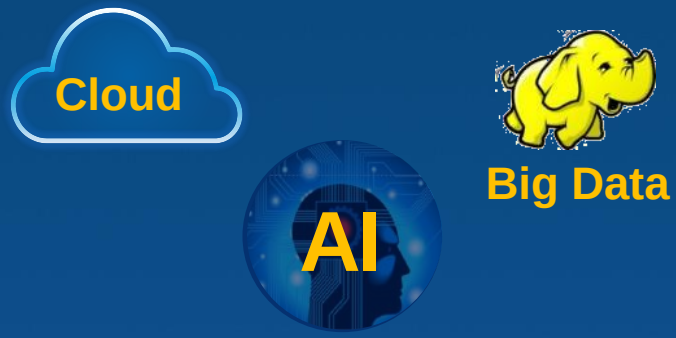
Workload Optimized

Application-optimized

Unrivaled performance

- Flexible modular architecture
- Diversified innovations
- In-depth application-optimized hardware acceleration

Adaptive to Changes



Future-ready

HPC converged architecture

- Rapid adoption of emerging technologies
- Versatile HPC system
- HPC and cloud perfect synergy

Thank You