
Open. Together.
OCP
GLOBAL
SUMMIT

John Hu

Vice President
Inspur Group, Inspur

OCP GLOBAL SUMMIT

OCP SUMMIT | 2019 OCP Global Summit | March 14–15, 2019
San Jose Convention Center | San Jose, California

inspur

Open Innovation Driven by AI.

Dynamic, accelerated solutions for the open data center

AI for Open Rack and OCP Hardware

John Hu
VP/CTO, Inspur Group

Open Platform, Open Eco-system

Embrace Open Sourcing

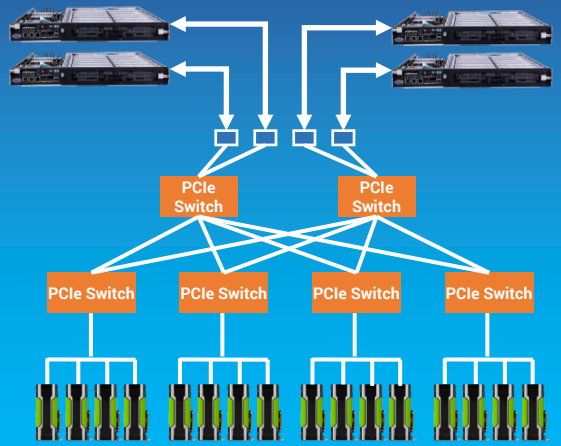
Innovate Open Technology

JDM for Open Design

SR-AI Rack: 1st Rack Server with Resource Pooling for AI Compute

SR-AI RACK
(SN3410M5)

Flexible GPU pooling solution with 1-4 GPU nodes; designed to run various AI applications such as AI cloud, deep learning for training, and image recognition.



PCIe Fabric Interconnected Infra

GPU/FPGA/MIC Compatible

64 x GPU/FPGA/MIC Expandability

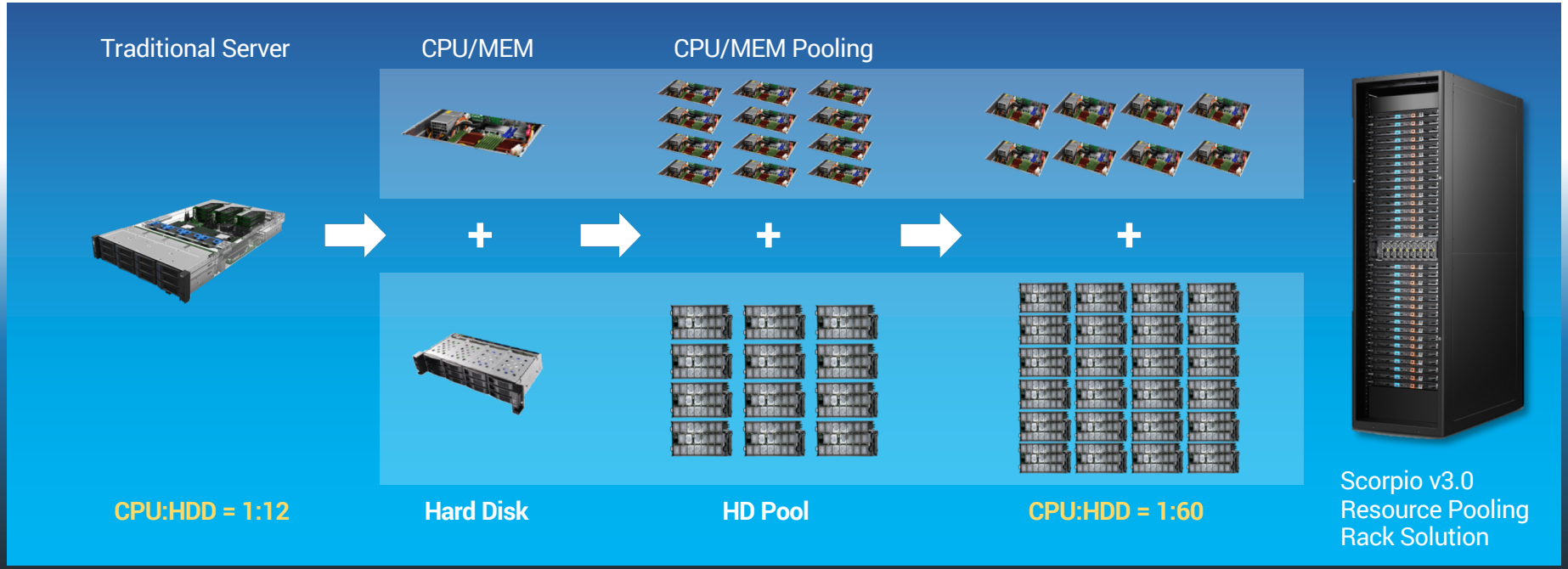
512TFlops Peak Performance

2-4X Comm Bandwidth Boost

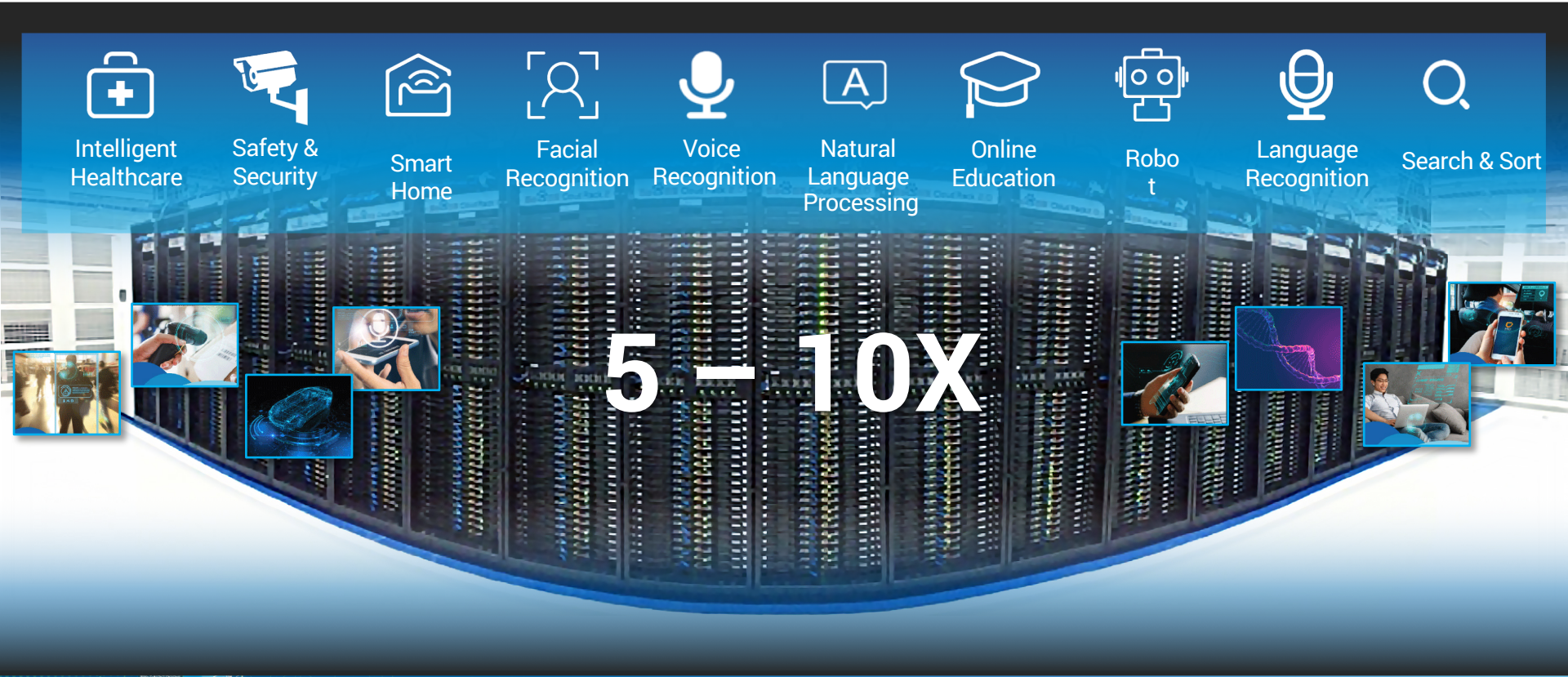
50% Reduced comm latency

Resource Pooling Infrastructure to Enable Equipment's Physical Layer Elasticity and Optimize Cost ⬇ 31%

Decoupling Compute and Storage ▶ Pooling ▶ Flexible Configuration



Efficiency Improvement



High-Density Cloud-Optimized Platform by Inspur

In Process

inspur intel

- A lead cloud-optimized platform design partner with Intel
- Cloud workload optimization system solutions
- Joint GTM and 1st High-density Cloud-optimized Platform Contributor to open source community

High-Density Cloud-Optimized Platform

NF8260M5 Flexible High Density 2U 4-Socket Server

Validated for the future generation Intel® Xeon™ Scalable Processor (Codename Cascade Lake), optimized with Intel® Optane™ DC persistent memory

Designed and optimized for cloud Infrastructure-aaS, Function-aaS, and Bare-Metal-aaS solutions

Operational Efficiency

- High density 4-socket design offers double digit % TCO savings compares to 2-socket in same generation
- Offset processor placement for efficient cooling
- Front hot-swap accessibility U.2 drives, PCIe and OCP 2.0/3.0 module choice

Power Saving

	Fan Power Saving	System Total Saving	Power Saving over 3 Years
2 * 2S vs **1 * 4S**	30watt	87watt	87watt * $6 = $522

Increased VM Density, Cloud-Optimized Performance

2U2S in TWO 42U Rack

2U2S in TWO 42U Rack
80 CPU

64 ports switch
40 chassis
40 servers
480 DIMMs
80 Boot Drives
40/80 NICs
80 Power Supplies

V
S

2U4S in ONE 42U Rack

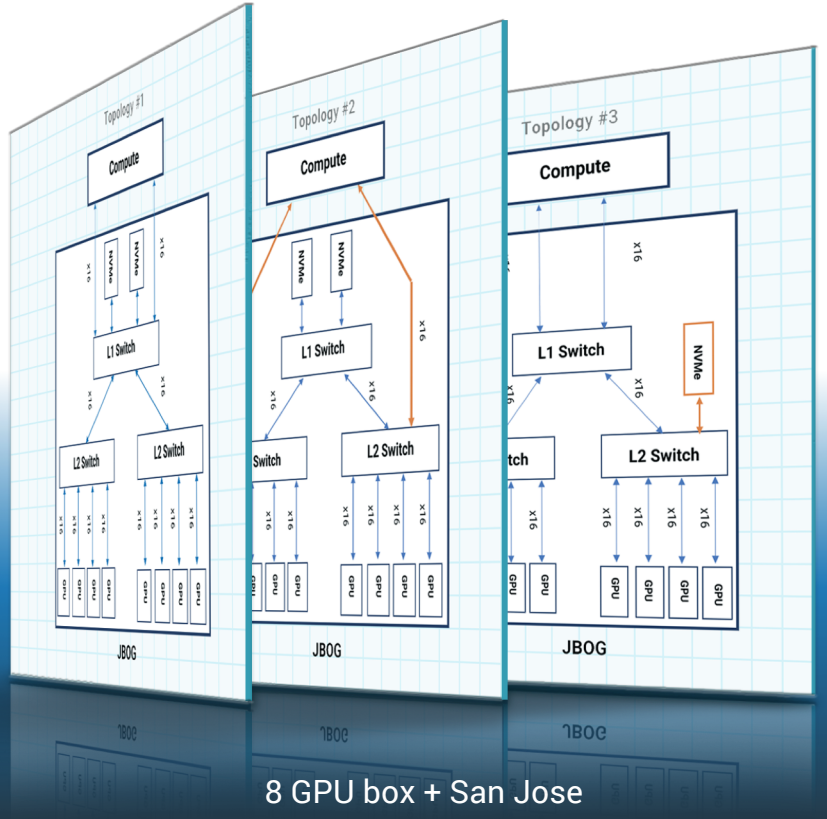
2U4S in ONE 42U Rack
80 CPU

64 ports switch
20 chassis
20 servers
480 DIMMs
40 Boot Drives
20/40 NICs
40 Power Supplies

Additional saving on Network Fabric & Boot Device will further lower the TCO

New JBOG Contribution to OCP

- Contribution and OCP Accepted recognition In Process
- New contribution to OCP for 2nd choice of 8 GPU box solution (XM2)
- Two new topology contributed to OCP for different AI applications



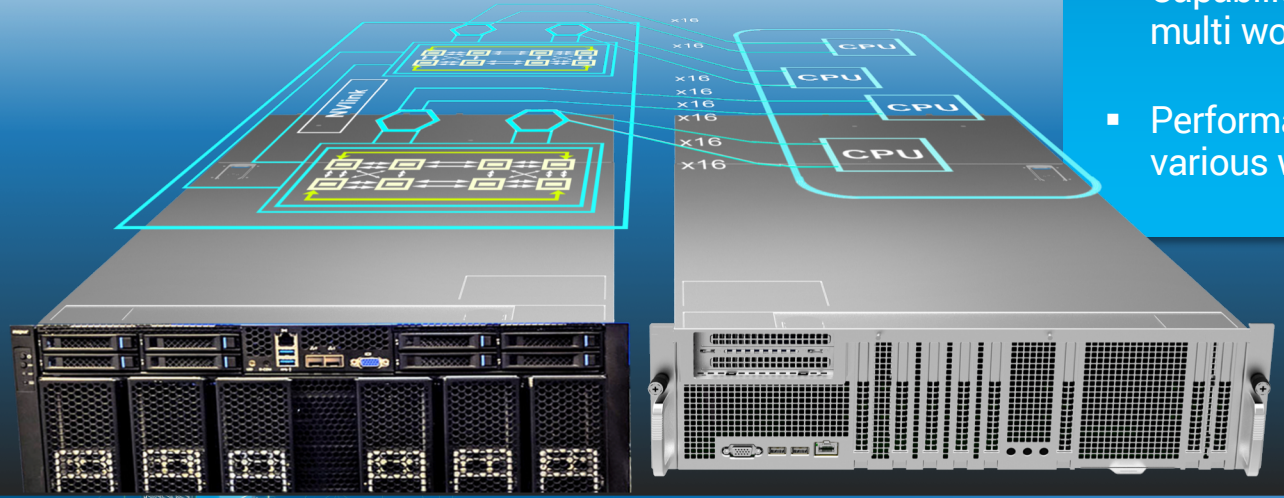
8 GPU box + San Jose

Advanced New Architecture for AI / First 16 GPU Box to Market

4-Socket Olympus: Contribution and OCP Accepted recognition In Process

- First 16 GPU Box launch to market
- Most powerful in computing and GPU combined for deep learning
  - 80 CPU cores : 16 GPUs
- Capability to maximize throughput with multi workload
- Performance increase 20%-60% under various workload with 16 GPU

Topology for the most powerful 16 GPU and computing combined



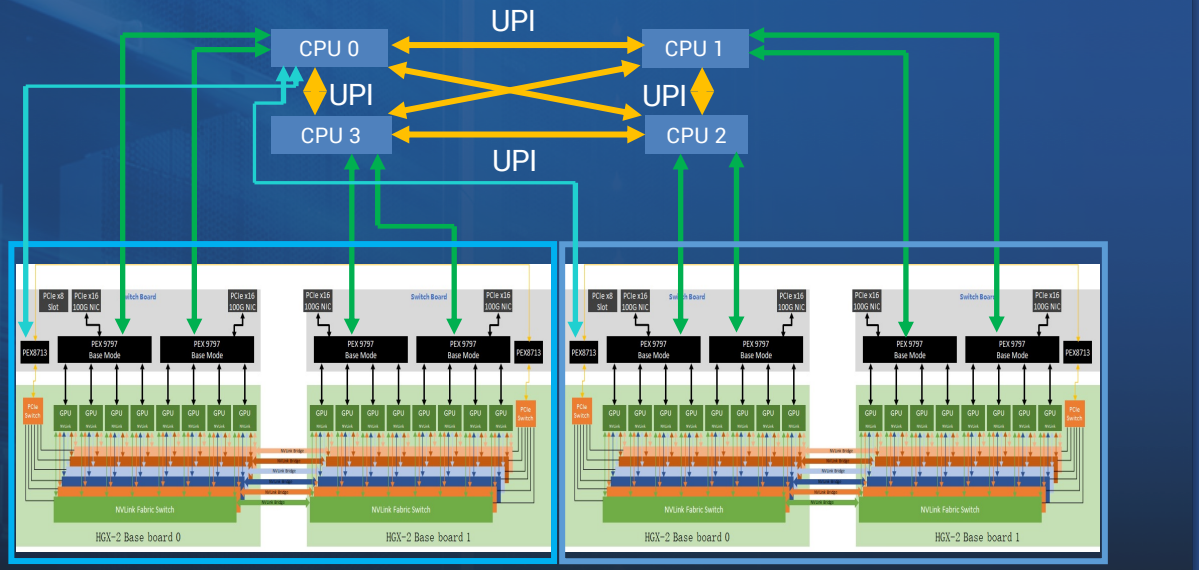
Inspur 4-Socket + GPU Box (16 GPUs)

Most Efficient 32 GPU Topology

4-Socket Olympus and 32 GPU

4-Socket Olympus
Contribution and OCP Accepted recognition In Process

Inspur 4-Socket + 2*16GPU Box (32 GPUs)



NF5888M5 with 16*GPUs

NF5888M5 with 16*GPUs

Most efficient 32 GPU topology vs. InfiniBand

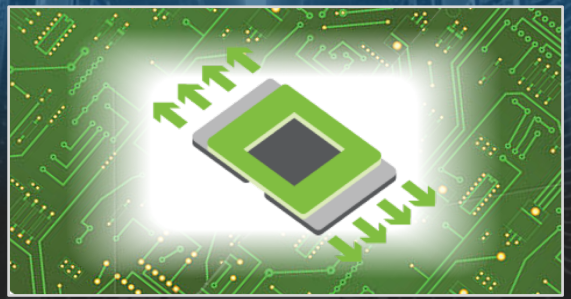
Open Together

OCP Community

- Embrace open collaboration
- Commit to OCP community

OCP OAM Project

- Collaborate on the module
- Plan in place to support baseboard and chassis design



SPEC_ML

- Establish unified & credible benchmarks for ML
- Chair the committee with 12 other members

As a top 3 worldwide server vendor, Inspur is fully committed to “Open Together”.

Moving You Forward