



Inspur SC19 OCP AI Solution Presentation

Gavin Wang Customer Product Manager

About Inspur

- Inspur Server Retains Fastest Growth in 1HY19



World No.3 in shipment and revenue for 5 consecutive quarters



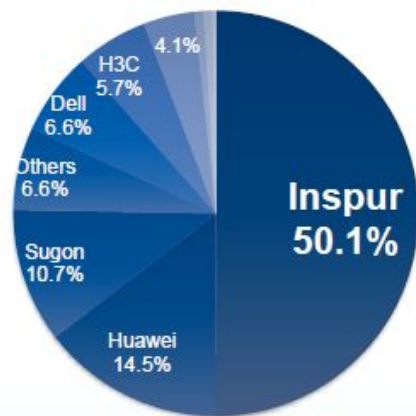
Server Revenue 60%+ CAGR in past 5 years based on Inspur' financial report







The only OEM that retains positive growth in 1H19

- AI Market Leadership

2018 AI Server Market Share by Vendor



#1 Inspur AI sever shipment in China, with MS 50.1%, growth by 103.6% vs 2017

 E2E Vertical Solution	AI Service (for AI Transformational)	AI Solution GPU/Phi/FPGA		
 Optimized Framework	AI Solution (for AI Vertical Leader)	Caffe-MPI		
 Comprehensive Management Suite	AI System (for AI Vertical Leader)	AIStation	T-Eye	
 Cutting-edge Hardware Product	AI Server (for AI Hyperscaler)	GPU Server	Xeon Phi Server	FPGA Accelerator

2019/1- Inspur named as SPEC ML 1st Chairman



Worldwide Server Unit Shipments in Q2, 2019

Ranking	Vendor	2Q19/2Q18 Unit Growth
1	Dell	-16.8%
2	HPE	-5.8%
3	Inspur	14.6%
4	Lenovo	-19.2%
5	ODM direct	-7.3%

Worldwide Server Vendor Revenue in Q2, 2019

Ranking	Vendor	Growth Rate
1	Dell	-13%
2	HPE	-3.6%
3	Inspur	32.3%
4	Lenovo	-21.8%
5	ODM Direct	-22.9%

Powerful AI Servers

Training

AGX-2



World's highest density 2U8GPU server with NVLink-Empowered

FP5295G2



OpenPower CPU & GPU NVLink Tight Coupling

NF5488M5



World's first NVSwitch-empowered 4U8GPU Server, Optimized for Parameter Transfer

AGX-5



One of the Most Powerful AI Server 2 PetaFLOPS Tensor Computing Performance

Inference

NF5468M5



Optimized for AI Inference Support up to 20*T4/MLU100/NNP-I Extreme storage of 384TB

I48



Optimized for High Density Inference Computing accelerated by latest VNNI; Up to 448 CLX CPU cores in 4U chassis

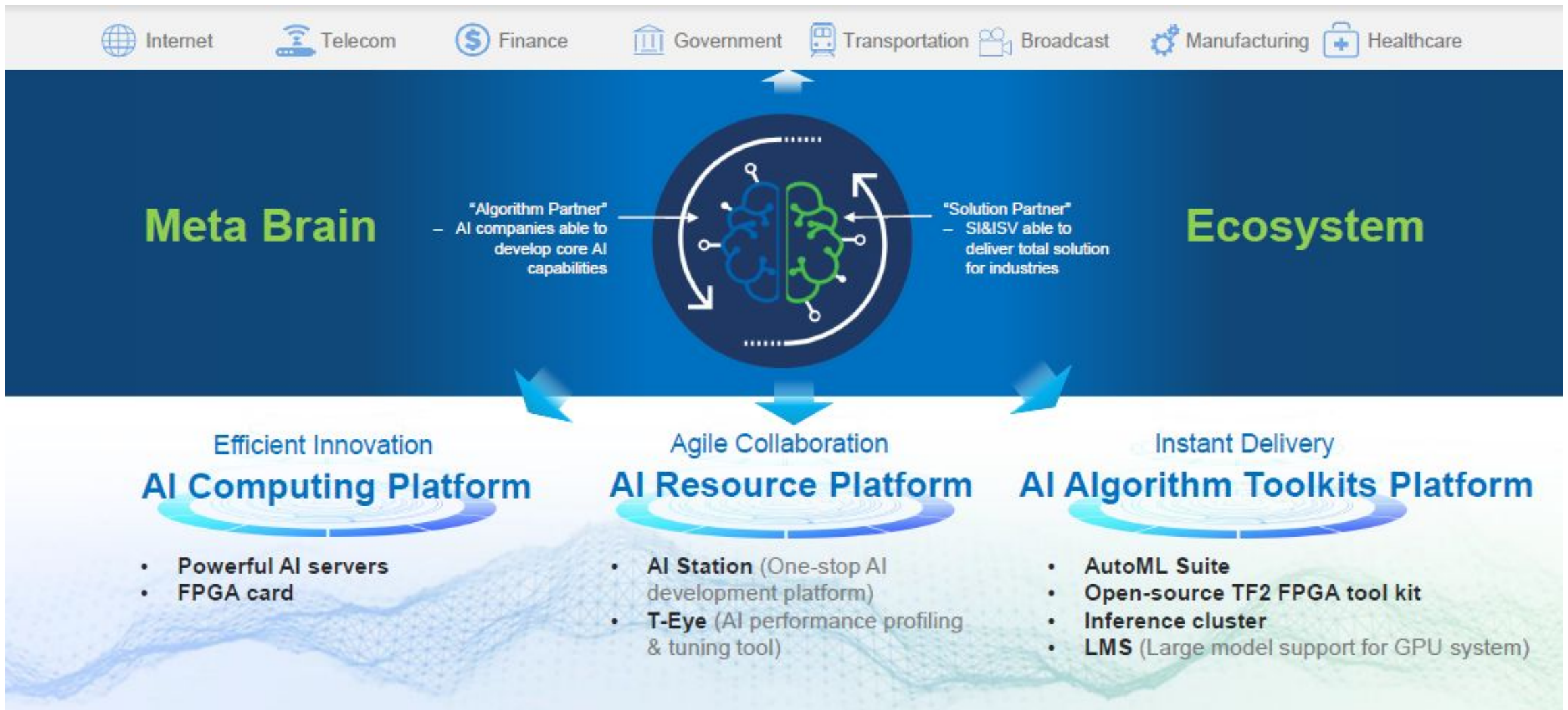
Edge

NE5260M5



Edge AI Computing Support 2*FHFL/4*FHHL PCIe GPU Wall-mounted

Inspur Meta Brain Eco-Plan: Join to from great power and make difference



Cluster Cloud Empowered by Inspur AI Station Solutions

AIStation Case Study: Video Processing

60+ Training Servers, 500+ GPUs; 6 Algorithm Teams, 70+ Members

AI Station Development Platform

Task-Balance Strategy: S-3 H-10 ; Resource-Quota Strategy: 24
Resource-Group Strategy: P100, V100 ; GPU Sharing Strategy: 6

Sharing Strategy: 32 GPUs support simultaneous development and testing for 70 people;


Task Queuing: nights and holidays are taken advantage of, obtaining **20% improvement** in utilization;

Dynamic Allocation: Developers work on 4-5 tasks simultaneously, reducing the development period to **1/3**;

Resource Utilization: improves **from 70% to 90%**.

P100_share	P100	V100	2080Ti	1080TI_dev
Quantity: 96GPU Sharing: 3 Purpose: Training Users: ALL Quota: None SSD Caching	Quantity: 120GPU Sharing: 0 Purpose: Training Users: Behavior Analysis Quota: 8 SSD Caching	Quantity: 64GPU Sharing: None Purpose: Training Users: ALL Quota: None SSD Caching	Quantity: 120GPU Sharing: None Purpose: Robots, Face Recognition Users: ALL Quota: 8 SSD Caching	Quantity: 32GPU Sharing: 6 Purpose: Development & Testing Users: All Quota: None SSD Caching

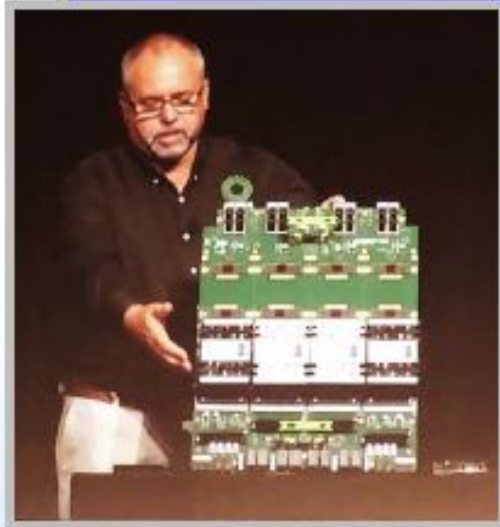
Leading OAI (Open Accelerator Infrastructure) Design



Open Accelerator Infrastructure

- Open Accelerator Module (OAM) spec released
- Universal Baseboard (UBB) spec in development
- Inspur showing first UBB today
- Prototype samples today at the Summit
- Inspur announces developer systems support

Open. Together.



OCP CTO Bill Carter show Inspur UBB on OCP Global Summit

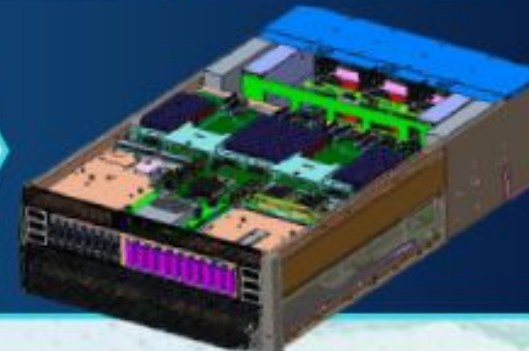


**Baidu & Inspur
X-Man 4.0**

Announced in OCP summit,
Amsterdam



**Inspur
21" OAM reference
Platform**

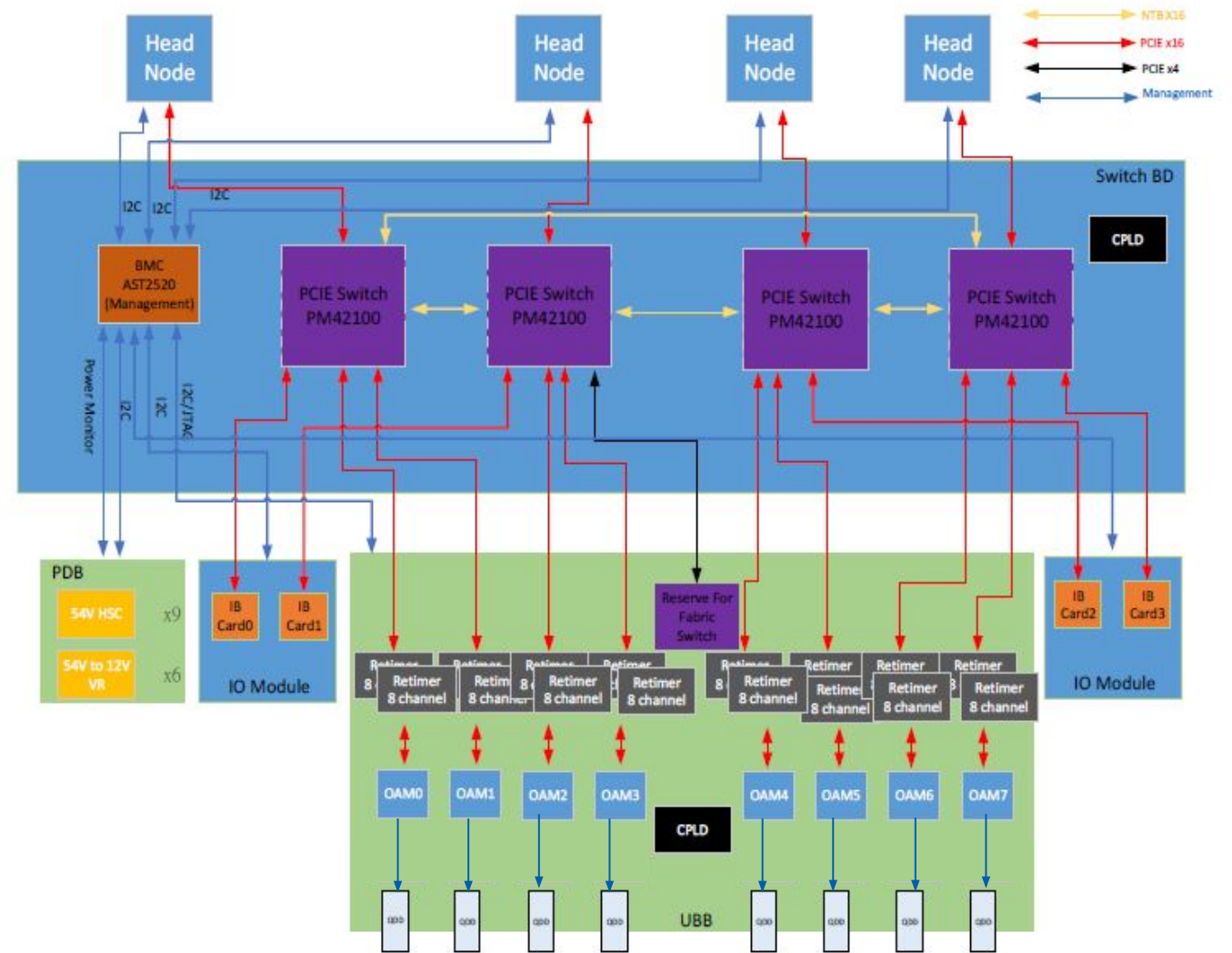


**NF5488M6
6U 8x AI Acceleration Module**

Inspur AI Server for Open Compute Object(OCP)



- 21-inch 3OU System
- Flexible JBOG Solution
- 8*OAM Modules with Universal Backplane
- Fully Connect Sale out
- 8* PCIE x16 High IO Expansion
- Deep Learning Training & Training Cluster

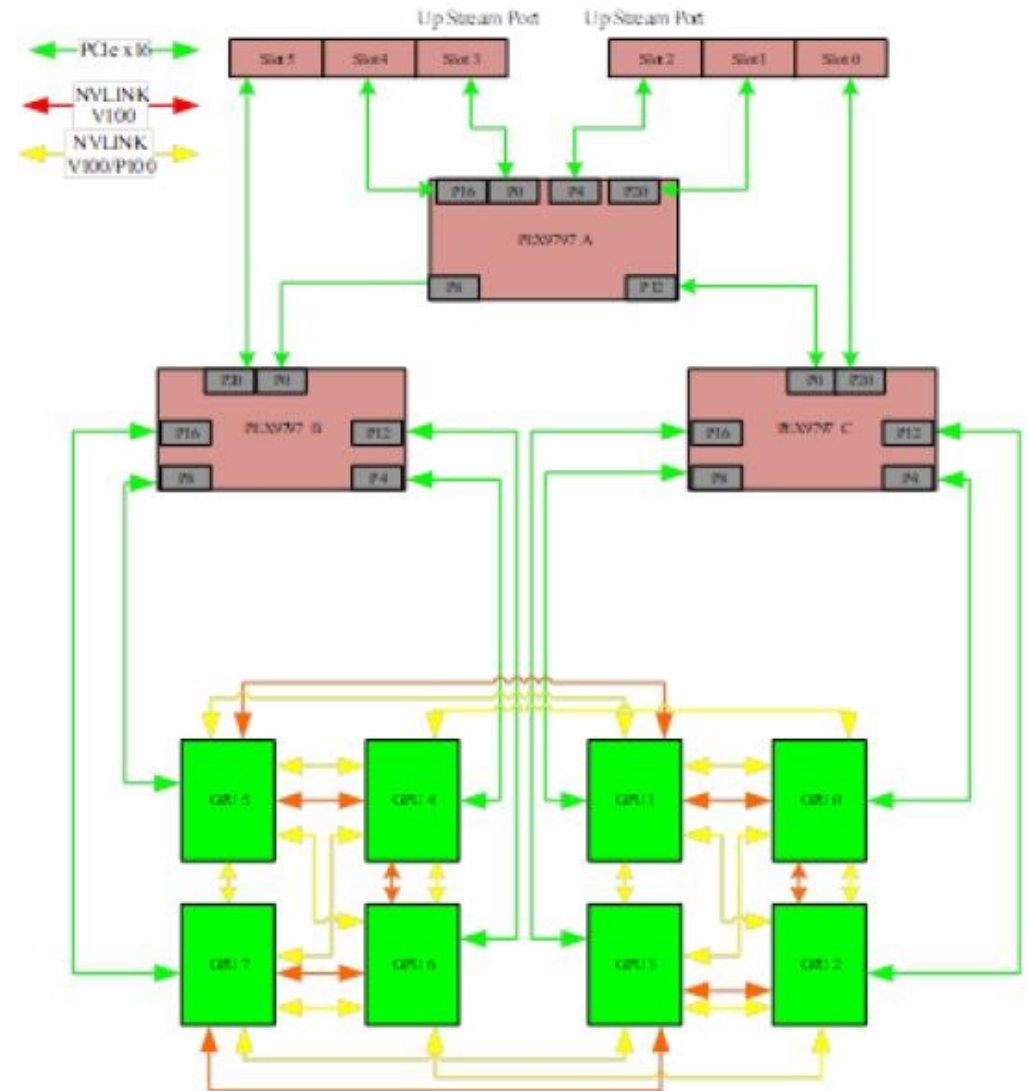


Inspur AI Server for Open Compute Project(OCP)

Mission Bay Project



- 21-inch 30U JBOG System
- Flexible topology for different application
- 8* NVidia Tesla V100
- 6* PCIE3.0 x16 FHHL for IO Expansion
- Open Rack V2 Compatible



Inspur AI Server for Open Compute Project(OCP)

Whistler Project



- 19-inch 3U 4-Socket Head node
- Olympus Architecture Compatible
- 4* Xeon Processors with 24* DIMMs
- 10* PCIE3.0 x16 slots for IO Expansion
- 1600W PSU 2+2 Redundancy

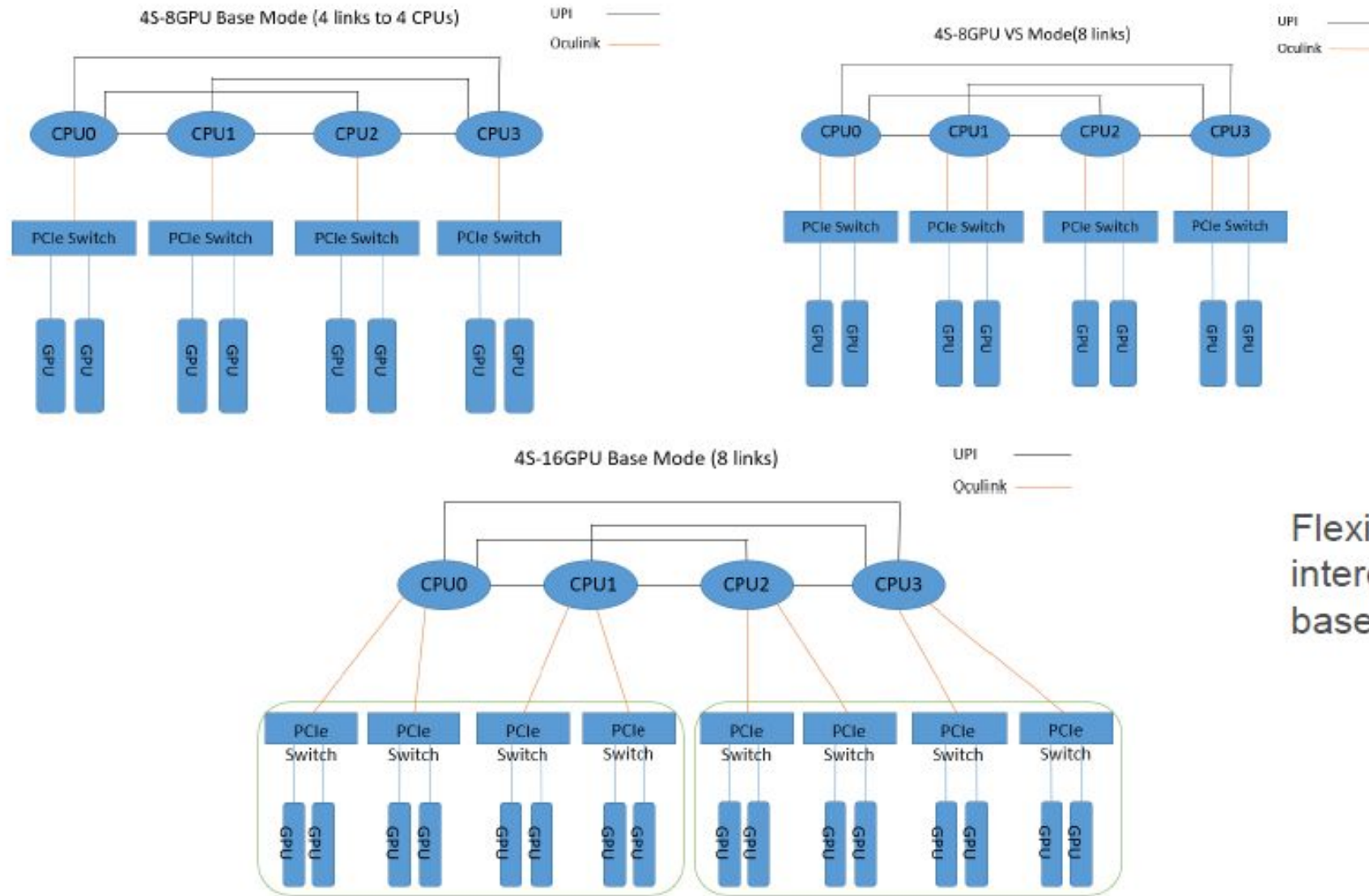


Olympus 4S system can serve as head node to connect with 1, 2 or 4 GPU expansion boxes (such as HGX-1).

Bring up to 8* PCIe x16 links for CPU-GPU communications. Great scale-up capacity for large neural network models.

Support more CPU cores and memory capacity. Improved 42% training performance on certain deep learning framework comparing to 2S.

Inspur AI Server for Open Compute Project(OCP)



Flexible CPU-GPU
interconnection topologies
based on different workloads

Thank you