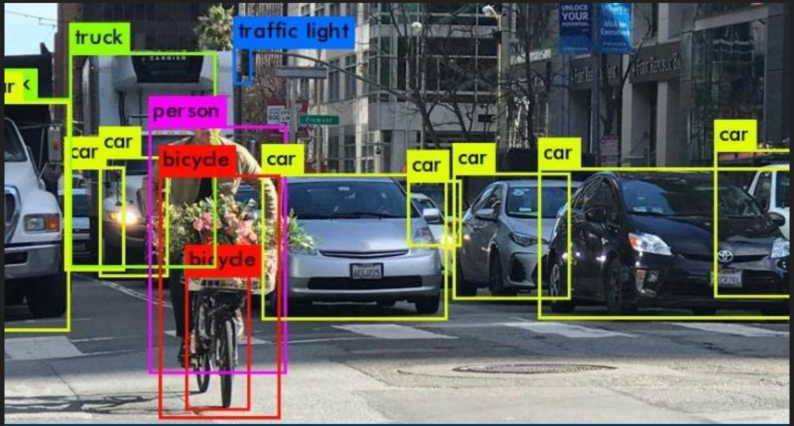
inspur

NF5468M6 Introduction

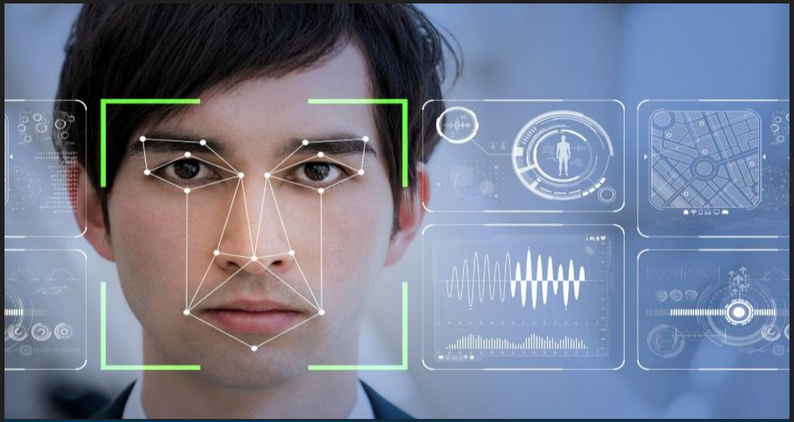
May 2021

Multiple AI Models Today



Objective Detection

SSD, Mask-RCNN, Faster-RCNN

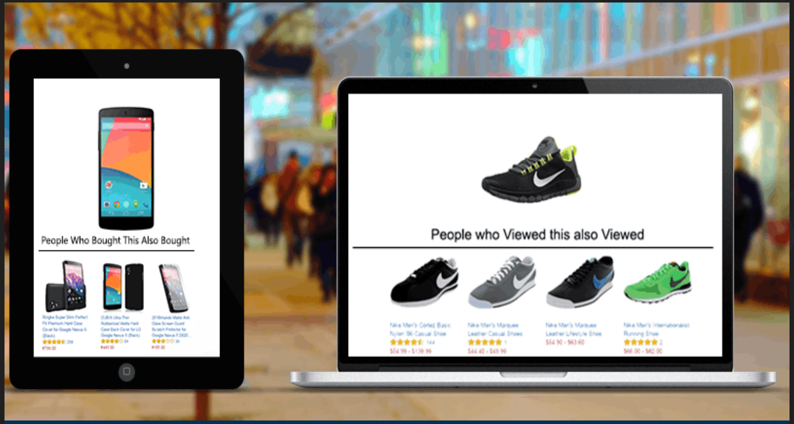


Computer Vision

ResNet, VGG, Inception, ResNext

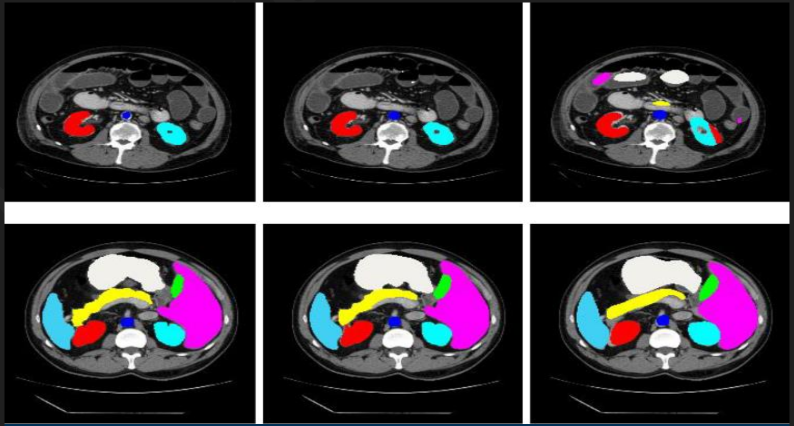
Natural Language Processing

GPT3, GPT-2, BERT


People Who Bought This Also Bought
People who Viewed this also Viewed

Recommendation

DLRM



Medical Image Segmentation

3D UNET

Speech to Text

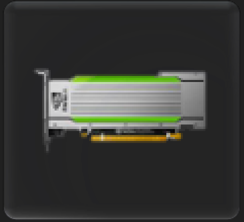
RNNT

Multiple AI Computing Power

NVIDIA

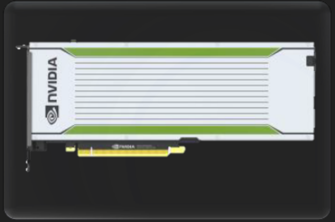
NVIDIA AMPERE

PCIe Dual-slot **A100**



PCIe Dual-slot
T4\A10\A30

NVIDIA QUADRO



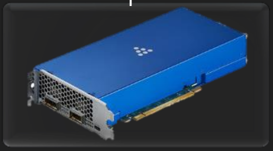
PCIe Dual-slot
RTX6000\A40

AMD

PCIe Dual-slot **MI50**

PCIe Dual-slot **MI100**

intel



PCIe Dual-slot Habana
Gaudi

PCIe **FPGA Card**

GRAPHCORE

PCIe Dual-slot **C2 Card**

XILINX

PCIe Dual-slot **FPGA Card**

NF5468M6

Adaptive PCIe Server tailored for AI Inference and multiple AI applications, such as intelligent video processing, cloud gaming, autonomous driving simulation and graphic rendering.

4U 4/8*A100, 16*A10 or 20*T4 PCIe GPU System with Application-driven Topology

NF5468M6: Elastic Cloud GPU Server

-3 SKUs for Various Applications-

Product	SPEC	Positioning	Highlights
NF5468M6-P	Form factor：**4U 8*PCIe** Storage：12x 3.5” or 24x 2.5” +2*M.2	Mainstream adaptive PCIe server for AI Training/Inference/ Graphic Rendering with 8 GPU	• High scalability, support 8 FHFL DW PCIe cards and 4 HHHL SW cards. • One-click topology switching • Multi/Single Host configuration • Large storage capacity
NF5468M6-T	Form factor：**4U 4*PCIe** Storage：12x 3.5” or 16x 2.5” +2*M.2	Cost-effective PCIe server for light-weight AI Training/Inference with 4 GPU	• CPU-GPU pass-through design, high P2P communication performance • No PCIe switch, low TCO
NF5468M6-V	Form factor：**4U 16*PCIe** Storage：12x 3.5” or 24x 2.5” +2*M.2	High throughput AI Server for intensive AI Inference/ IVA/ Cloud Gaming with up to 20*GPU	• High scalability, support 16 FHFL SW PCIe cards or 20 HHHL SW PCIe cards. • Large storage capacity

NF5468M6 Specification

Model	NF5468M6-P	NF5468M6-T	NF5468M6-V
Height	4U		
CPU	2* new generation Intel® Xeon® IceLake scalable processors, TDP 270W		
GPU	Support 8* A100, A30, A40, MI100, etc. FHFL DW PCIE GPU cards Rear supports up to 4 PCIe4.0 x16 slots	Supports 4* A100, A30, A40, MI100, etc. FHFL DW PCIE GPU cards Rear supports 2 PCIe 3.0 x8 and 2 PCIe4.0 x8 slots	Support 16* A10, etc. FHFL SW GPU cards Rear supports up to PCIe4.0 x16 slots
Chipset	Intel® C621A series chipset ((LBG-R)		
Memory	32* DDR4 3200MHz RDIMM		
Internal PCIE	Support up to 2* internal standard Raid card		
Front I/O	2* USB 3.0, 1* VGA, 1* RJ45 serial port		
Rear I/O	1* serial port, 2* USB 3.0, 1* RJ45 management port, 1个OCP3.0(support NCSI)		
Storage	24* 2.5” or 12* 3.5” SAS/SATA drives(up to 8* NVME SSD), 2* M.2 SATA SSD	16* 2.5” or 12* 3.5” SAS/SATA drives(up to 2* NVME SSD), 2* M.2 SATA SSD	24* 2.5” or 12* 3.5” SAS/SATA drives(8* NVME SSD), 2* M.2 SATA SSD
RAID	Optional support RAID 0, 1, 10, 5, 50, 6, 60, etc., support Cache super capacitor protection, provide RAID state migration, RAID configuration memory		
OS	Microsoft Windows Sever、Red Hat Enterprise Linux、Ubuntu Linux、CentOS, etc. mainstream OS		
Cooling	N+1 Redundant system cooling fan		
Power	4* 1600W/2000W/2200W/3000W 80Plus platinum PSU, supports 2+2 redundancy		
Size (W*H*D)	483mm * 175.5mm * 830mm		
Temperature	5 - 35°C / 41°F - 95°F		
Full Load Weight	≤85kg		

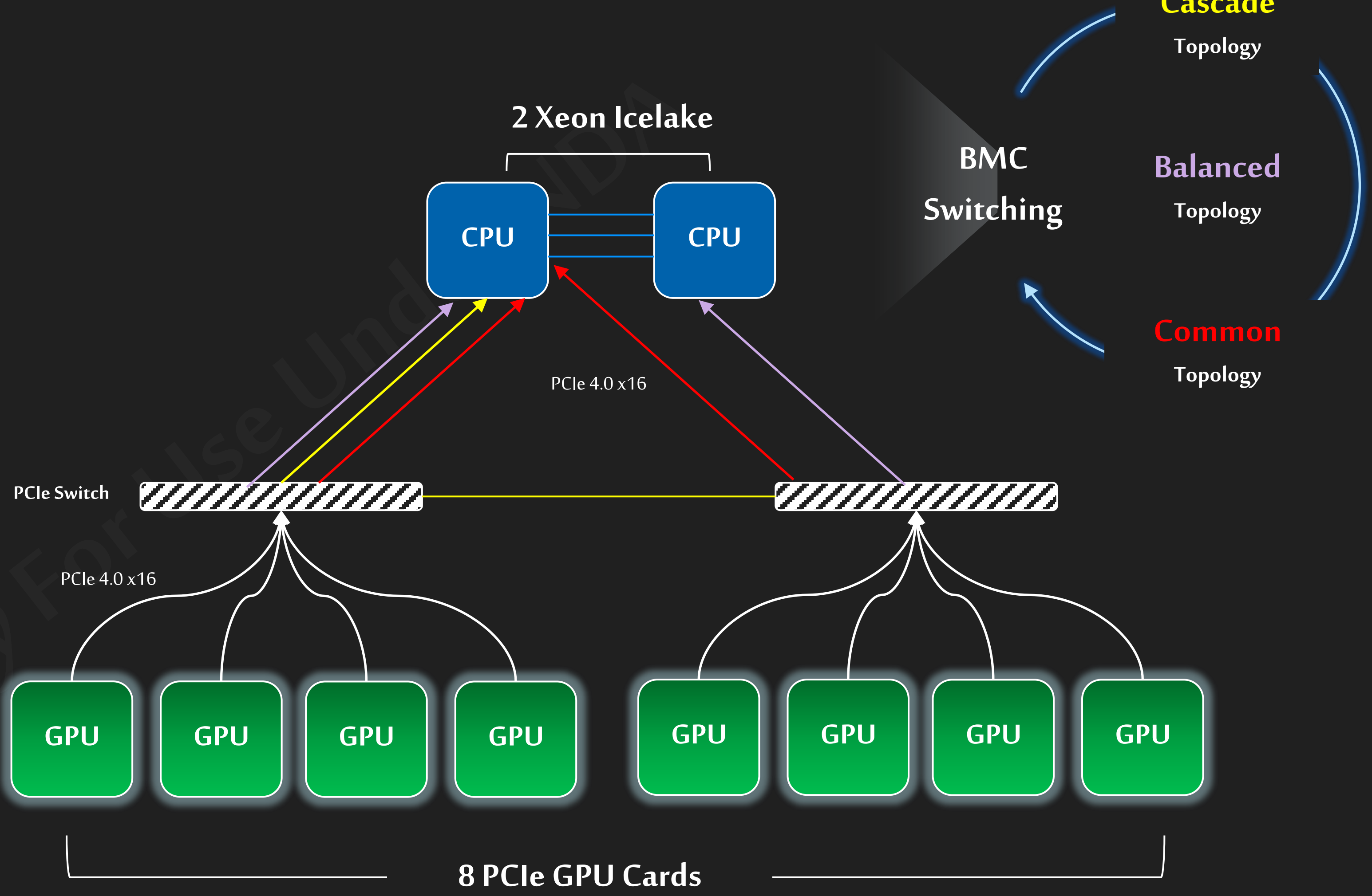
NF5468M6-P : Mainstream Adaptive PCIe Server

POC MP

NF5468M6-P

2021/02 2021/06

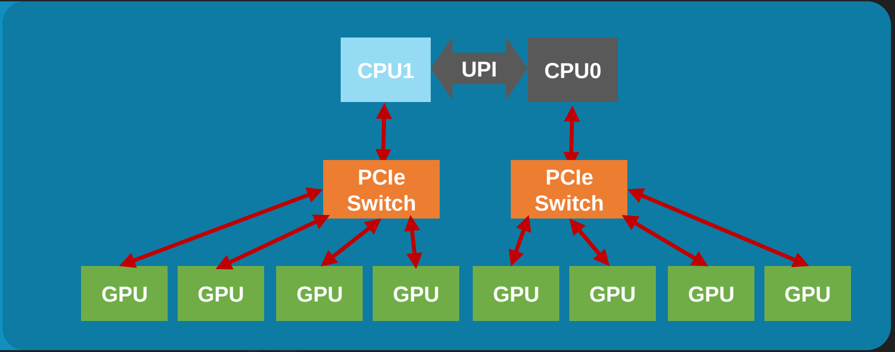
- 2 Intel Xeon ICX CPU, PCIe Gen4, Up to 205W
- 8 PCIe A100 / MI / RTX GPUs
- 32 DDR4 3200MT/s RDIMMs//LRDIMMs
- Config1: 24*2.5" SAS/SATA (Up to 8*NVMe)
  Config2: 12*3.5" SAS/SATA (Up to 8*NVMe)
- 1* OCP 3.0 + 4 * PCIe4.0 x16 (IB) slots
- (2+2) 1600/2000/2200/3000W PSU



NF5468M6-P: Optimize AI Computing Resource Allocation

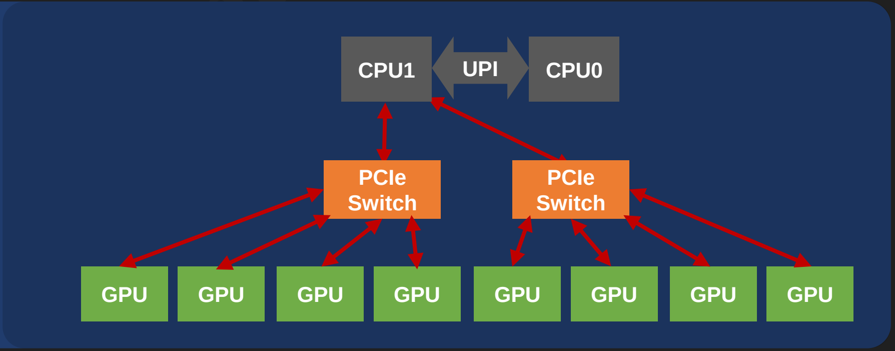
Balance

- Suitable for GPU pass-through virtualization
- GPU cloud application
- Small and medium-sized deep learning training
- Inference, public cloud and HPC scenarios



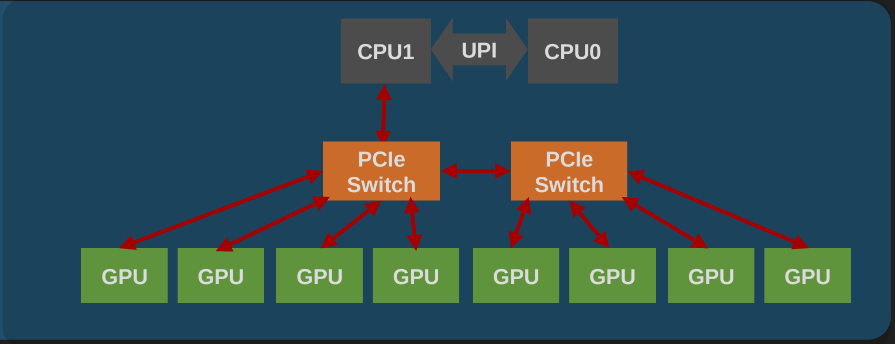
Common

- Excellent AI training performance
- GPU P2P communication
- Most deep learning application scenarios



Cascade

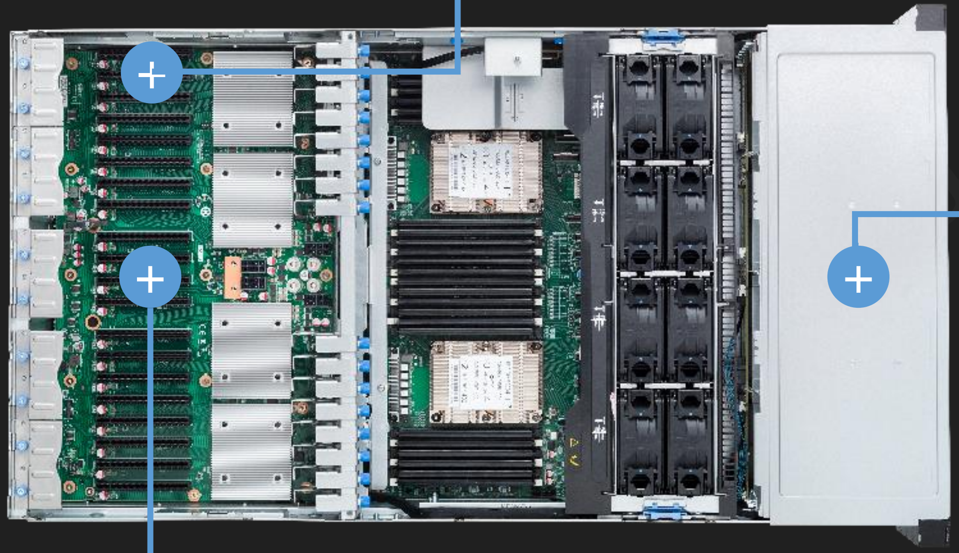
- Some AI training models have the best performance
- GPU P2P communication
- Large-scale deep learning application scenarios



NF5468M6-V: High-throughput PCIe Server

NF5468M6-V

- 2* Intel® Xeon® IceLake scalable processors
- 24* 2.5″/12* 3.5″drives
- 16* A10 (T4 Next)
- 32* DDR4-3200
- GPU Upstream communication bandwidth convergence ratio 4:1



Ultra video acceleration and AI inference ability

16* A10 GPU

GPU Upstream communication bandwidth convergence ratio 4:1

Large-capacity local storage

24* 2.5″/12* 3.5″drives

Significantly save the cost of video & image data storage

Rich IO extension

Built-in support 2* standard RAID card dedicated slots

4* PCIEx16 slots at the rear, support 200G high-speed network

POC MP

NF5468M6-V: System Topology

NF5468M6-V

2021/09 2021/10

2 Intel Xeon ICX CPU, PCIe Gen4, Up to 205W

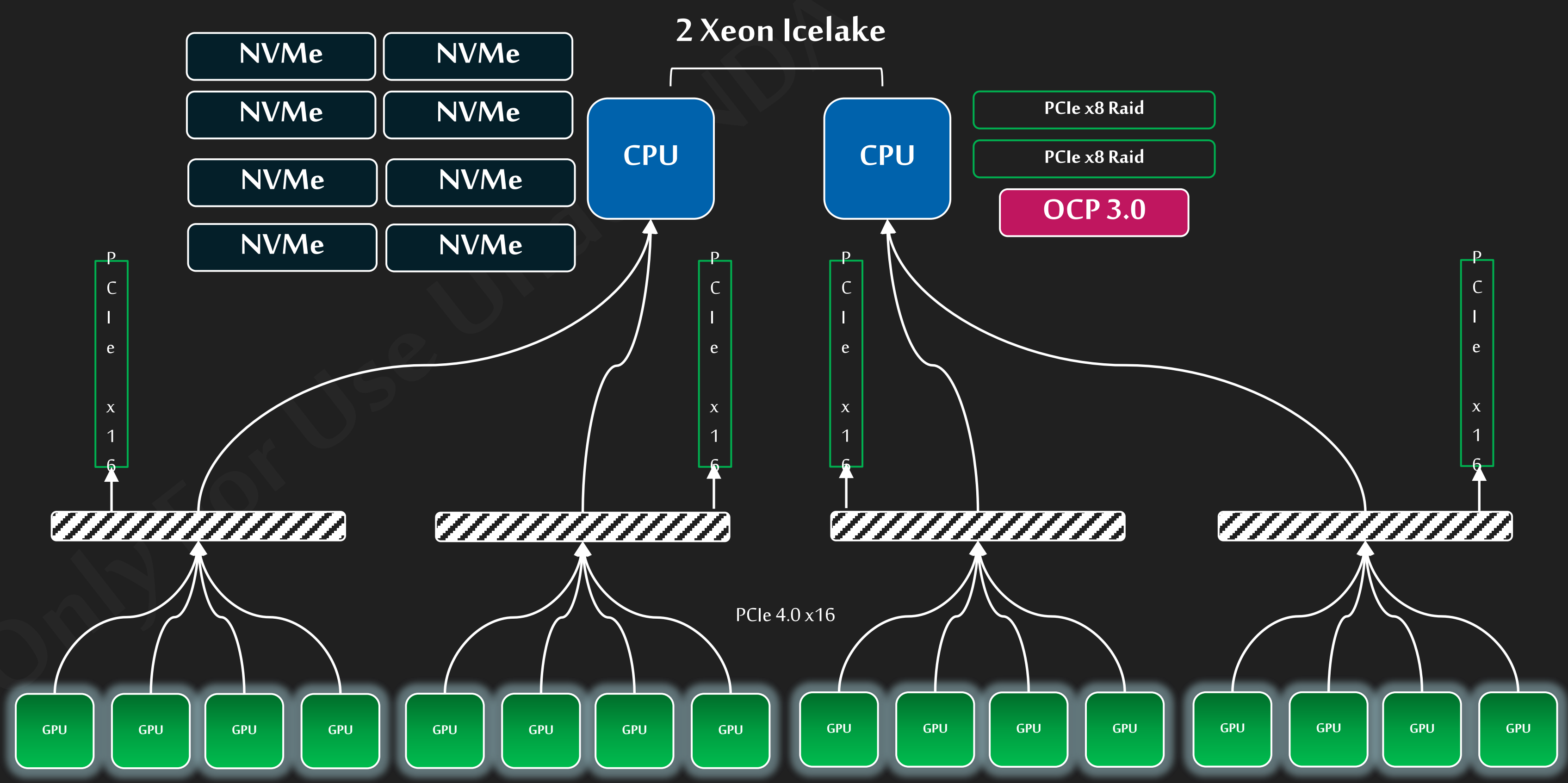
16 PCIe T4 / Single-slot GPUs

32 DDR4 3200MT/s RDIMMs//LRDIMMs

Config1: 24*2.5" SAS/SATA (Up to 8*NVMe)
Config2: 12*3.5" SAS/SATA (Up to 8*NVMe)

1* OCP 3.0 + 4 * PCIe4.0 x16 (IB) slots

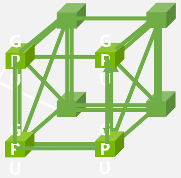
(2+2) 1600/2000/2200/3000W PSU



NF5468M6-T: Cost-effective PCIe Server

NF5468M6-T

- 2* Intel® Xeon® Scalable Processor
- 16* 2.5”/12* 3.5” hard drives
- Up to 4 FHFL GPU cards
- 32×DDR4-3200
- CPU-GPU pass-through design
- The most cost-effective

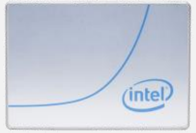


Low latency, high bandwidth

CPU-GPU pass-through design

High IO expansion

2*PCIe 4.0 x8+2*PCIe 3.0 x8



Data acceleration

2*M.2+2*NVMe SSD+10*HDD

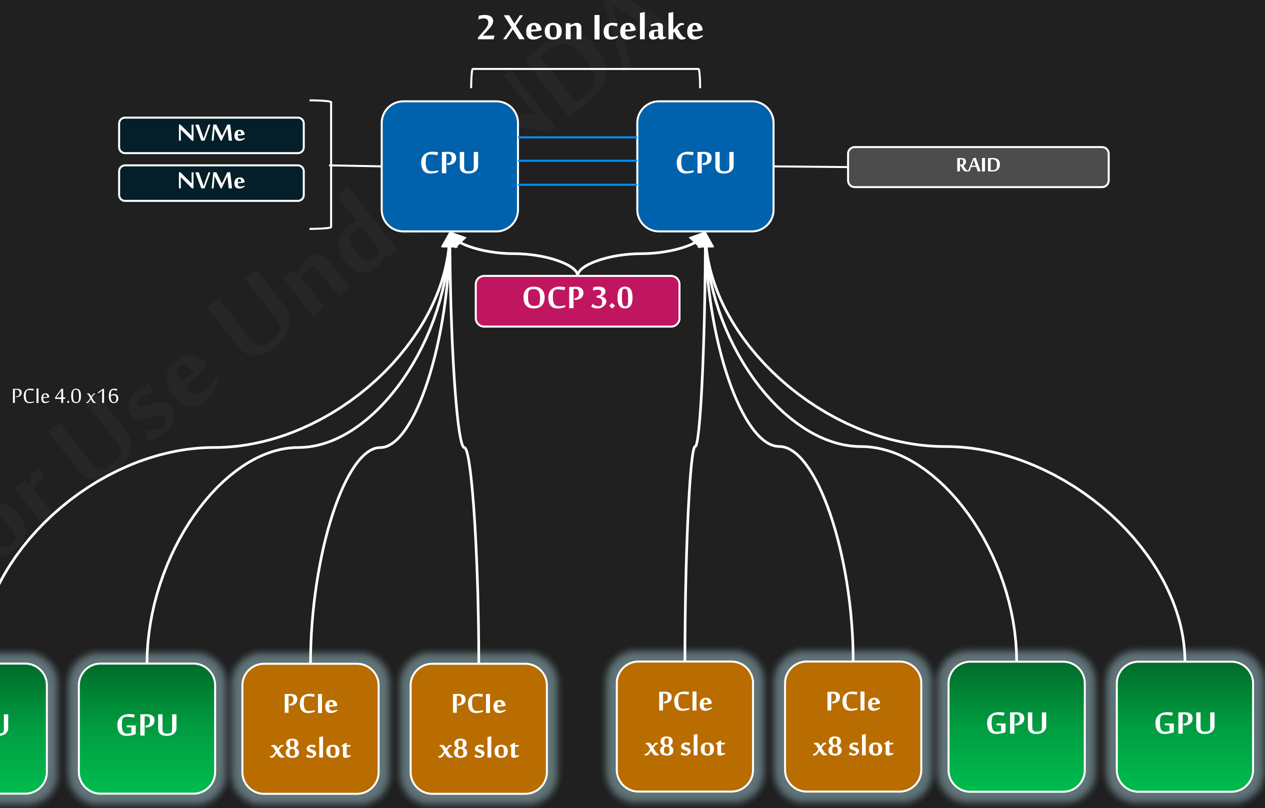
POC MP

NF5468M6-T: System Topology

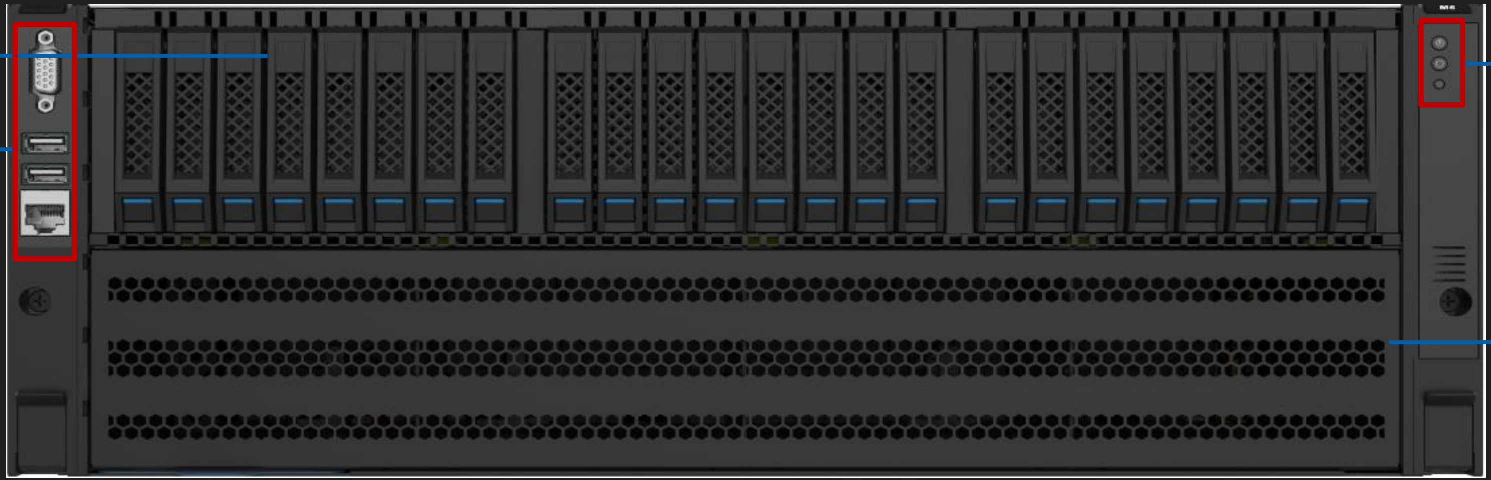
NF5468M6-T

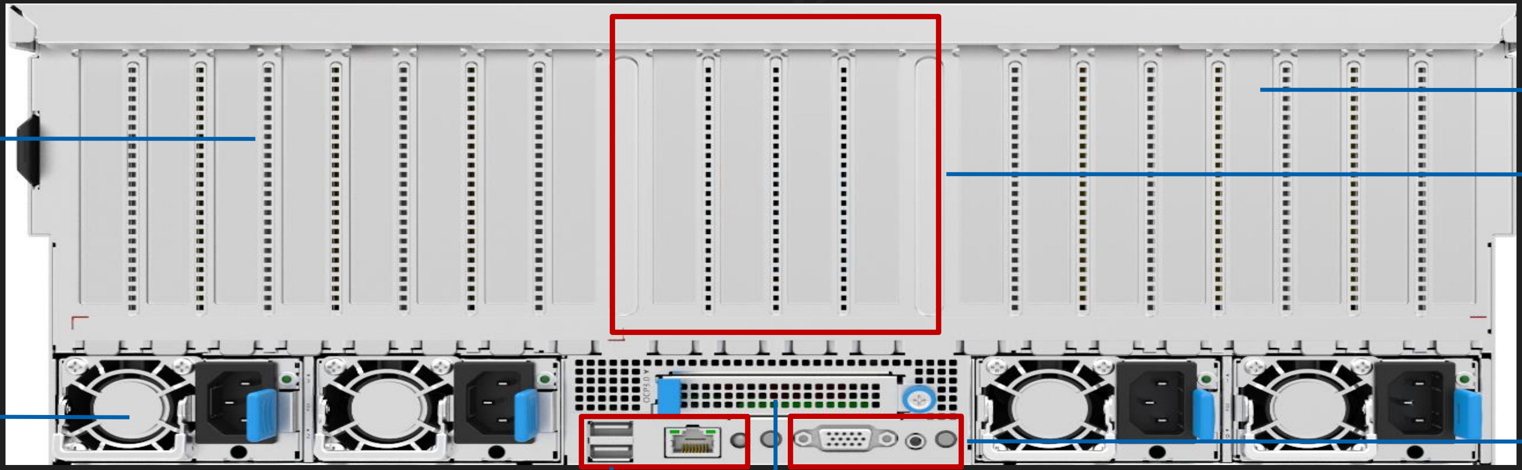
2021/02 2021/06

- 2 Intel Xeon ICX CPU, PCIe Gen4, Up to 205W
- 4 PCIe A100 / MI / RTX GPUs
- 32 DDR4 3200MT/s RDIMMs//LRDIMMs
- Config1: 24*2.5" SAS/SATA (Up to 8*NVMe)
  Config2: 12*3.5" SAS/SATA (Up to 8*NVMe)
- 1* OCP 3.0 + 4 * PCIe4.0 x16 (IB) slots
- (2+2) 1600/2000/2200/3000W PSU

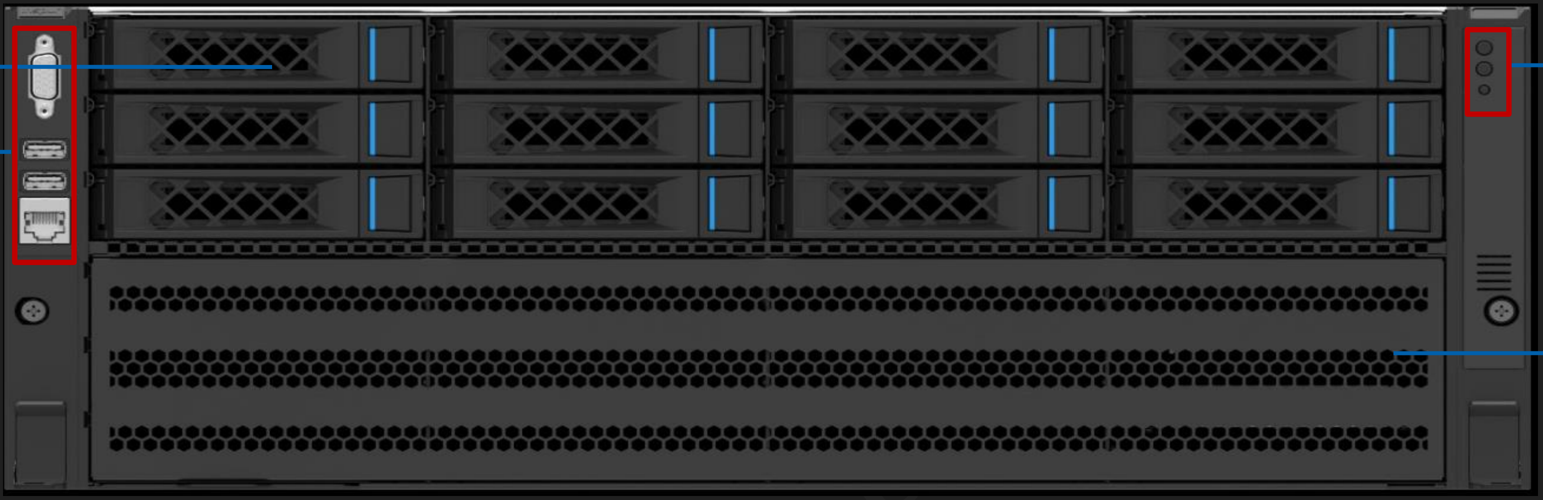


NF5468M6 System View———24* 2.5″ model





NF5468M6 System View———12* 3.5” model

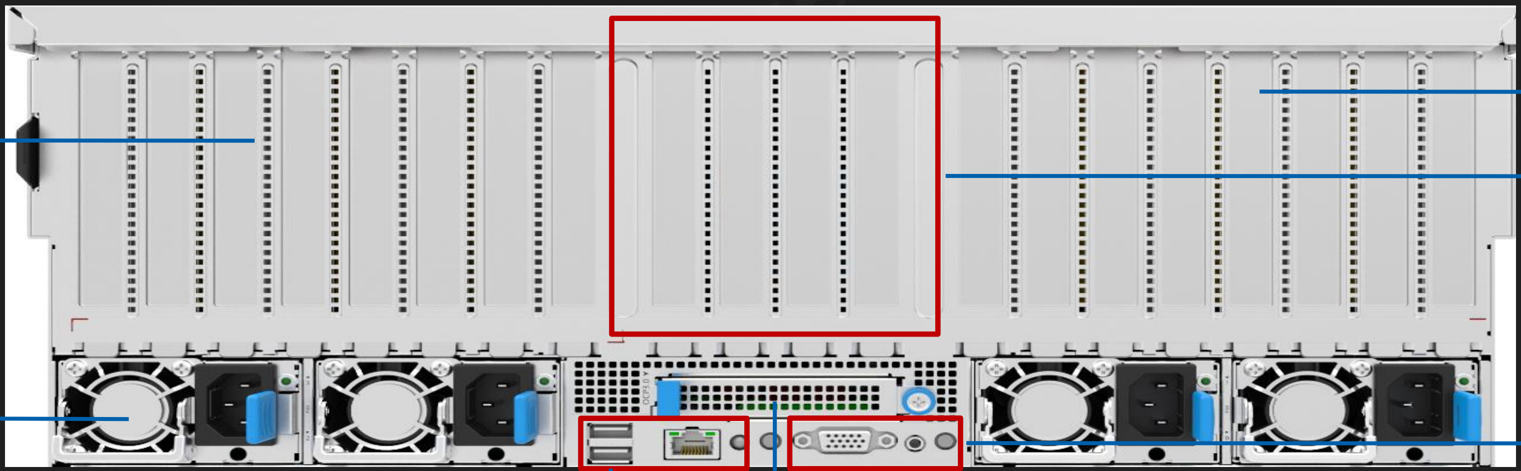


3.5” drives

Front I/O
VGA/USB

UI:PWR,UID,RST

RAID



GPU

GPU

NIC

PSU

VGA/UID/BMC

RJ45(IPMI)
2X USB3.0

OCP 3.0

Thank You

May 2021