Design GPU Systems for Hyperscalers, Diverse AI Applications and Open Compute standard datacenters

Nick Yan

PDT Manager of AI Product Line of Inspur
Inspur is a leading cloud computing and AI computing data center infrastructure provider

Top 3 server vendor according to Gartner and IDC

AI full-stack solution provider

Design GPU Systems for versatile scenarios
End to End Computing AI Product Portfolio

SC 2018 · Colorado
AGX-5

AI Training
8U 16x V100, NVSwitch
World’s highest density 2U server of 8 highest performance GPUs.

GTC2019 · San Jose
NF5488M5

AI Training
4U 8x V100, NVSwitch
Industry - First AI Server 8 V100 GPU with NVSwitch Enabled

IPF2018 · Beijing
NF5468M5

AI Cloud/Inference
4U 8x V100/4U 16x T4
Elastic GPU server designed for AI cloud.

ISC2017 · Frankfurt
GX4

PCI-E Pooling
2U 4x GPU BOX
Flexible Expansion, available for 2-16 GPU cards extendibility.

GTC2019 · San Jose
NE5260M5

Edge AI
2U 2x V100/ 6x T4
Design for Edge Computing

HyperScaler

New Edge Usage
Creating World’s Most Powerful & Reliable System

Nvidia’s HGX

High Volume Open Standard Motherboard

World Class Reliable & High Performance
Pushing the Envelope With HyperScaler

4 socket Platforms on Project Olympus
End to End Computing AI Product Portfolio

AGX-5
AI Training
8U 16x V100, NVSwitch
World’s highest density 2U server of 8 highest performance GPUs.

NF5488M5
AI Training
4U 8x V100, NVSwitch
Industry - First AI Server
8 V100 GPU with NVSwitch Enabled

NF5468M5
AI Cloud/Inference
4U 8x V100/4U 16x T4
Elastic GPU server designed for AI cloud.

AGX-2
AI Training
2U 8x V100/NVLINK
Minimum Size
Maximum Performance
NVIDIA® NVLink™ Enabled

NE5260M5
Edge AI
2U 2x V100 / 6x T4
Design for Edge Computing

HyperScaler

New Edge Usage
AI Training Infrastructure AGX-5 Overview

**HGX’s Wave “Zero” Partner**
Leading OEM partner to design HGX-2 Solution

**Volume Ramp Choice by HyperScalor**
8U with 850mm Depth
Up to 5x AGX-5 within 42U rack space

**Proven Common Building Blocks (CBB)**
Leverage High Volume Motherboard with Nvidia’s HGX-2 to create a super reliable system

**Hyper Redundancy Design**
Up to (2+2) *2 PSU Redundancy Design
Active parts are all Hot-swappable

AGX-5
The Most Powerful / Dense AI Server
AI Training Infrastructure NF5488M5 Overview

Full Speed on GPU-to-GPU communication
- NVIDIA® NVSwitch, 2.4TB/s Aggregate Bandwidth
- GPU-GPU bandwidth 300 GB/s

Build-in Server Node with NVMe Drives
- Full function server node with 2x Xeon-SP with 3x UPI
- Up to 8x NVMe SFF drives

Balance I/O Design
- NUMA balance I/O with 3x PCIe slot from each CPU

World Class Power & Cooling Efficiency
- Best AC-DC Power Conversion Efficiency
- Optimal Air cooling Efficiency
AI Inference Infrastructure NF5468M5 Overview

- **World’s Dense Inferencing Server**: Up to 20x PCIe x16 slots
- **HyperScaler Thermal Quality**: Xeon Motherboard & GPU Board are isolated to create an "non-shadow" thermal design
- **Design with Flexibility**: Support both V100 and T4. Each slot has full PCIe x16 bandwidth
- **Serviceability for Mass Deployment**: Most active components are designed to be hot-swappable in order to reduce service downtime
AI Training Infrastructure AGX-2 Overview

**High Density**
2U 8GPUs highest density

**Superb Performance**
960 Tensor FLOPS, 376 TOPS on INT8.
NVIDIA® NVLink™ 2.0 ready

**Flexible Topology**
10 Topologies of GPU for various applications.

**High Speed Connection**
Up to 400G RDMA InfiniBand, optimized for low latency HPC, AI cluster

Minimum Size. Maximum Performance
2U 8GPU Server with NVIDIA® NVLink™ Enabled
Edge Application is Growing, AI included
Edge AI Infrastructure NE5250M5&NE5260M5 Overview

**World’s First Edge with GPU computation**
Up to 2x V100 GPU card for Edge Training
Up to 6x T4 GPU cards for Edge Inferencing/Video Transcoding

**Super Compact Design for Rack and Edge**
430mm dept., Front service-able

**Uncompromised Xeon & Storage Support**
Support up to 2x Xeon-SP, 205Watt
16x DIMM slots
6x H/S SFF drive

**Open & Application Focus**
Compliant to OTII (Open Telecom IT Infrastructure)
Perfect for NFVi, Composable Infrastructure
Flexible Edge Work On-Demand

6x T4 or 2x V100

nvidia
Market Leadership in GPU-focus System Design

HyperScaler Design Capability

High Performance & Most Reliable Systems

Pushing AI computation with 4 Socket Motherboard

End to End Computation - From Data Center to Edge
Thank You!