GPU-ACCELERATED WORKLOADS NEEDING DEEPER NETWORK INTEGRATION

TELCO

GPU-ACCELERATED 5G SIGNAL PROCESSING

EDGE

AI OVER 5G

CONFIDENTIAL AI

DATA CENTER

MULTINODE AI TRAINING
NVIDIA H100 CNX
Converged Accelerator

H100 Tensor Core GPU
- 80GB HBM2e memory, > 2.0 TB/s
- Up to 7 MIG instances

ConnectX-7
- 400 Gb/s
- Network/security accelerators

PCI Gen5, 128 GB/s

2 slot FHFL

350 W
FASTER DATA SPEEDS WITH CONVERGED ARCHITECTURE
Ideal for Mainstream Servers

H100 and ConnectX-7
Bandwidth limits and contention due to host PCIe system

H100 CNX
Dedicated PCIe Gen5 link enables high transfer rate and predictable latency
H100 CNX USE CASES

5G SIGNAL PROCESSING
- High performance in commodity servers

AI ON 5G
- 5G processing and AI services on a single commodity server

MULTI-NODE TRAINING
- Excellent performance and scalability

Any GPU-accelerated, I/O intensive application can benefit

Availability
- Mass Production in Q3
- NVIDIA-Certified Servers in Q4
HIGHER PERFORMANCE FOR GPU-ACCELERATED I/O INTENSIVE WORKLOADS
Up to 1.5 faster on large-scale training*

Challenges with discrete accelerators
- Bound by speed of host PCIe backplane
- Contention due to bottleneck at CPU
- Number of PCIe lanes limits additional devices

Benefits of Converged Accelerator
- PCIe Gen5 speed regardless of host
- Tight coupling of GPU and network avoids contention
- Scale up capabilities with fewer devices

*assumes a cluster of 16 hosts, 4 GPUs each
H100CX KEY BENEFITS

**BETTER IO PERFORMANCE**
High-speed, contention-free GPU-Network data transfer

**OPTIMIZED DESIGN**
Inherently balanced scale-out architecture

**COST SAVINGS**
High performance with mainstream servers

**APPLICATION-READY**
Acceleration benefits without application modification
NVIDIA H100 CNX - H100 AND CONNECTX-7 CONVERGED ACCELERATOR
Unprecedented performance for GPU-powered, IO-intensive workloads

**BETTER IO PERFORMANCE**
High-speed, contention-free GPU-Network data transfer

**OPTIMIZED DESIGN**
Inherently balanced scale-out architecture

**COST SAVINGS**
High performance with mainstream servers

**APPLICATION-READY**
Acceleration benefits without application modification

- Multi-node Training
- 5G vRAN
- AI on 5G

---

*AI on 5G*