Micas AI-Fabric Solution

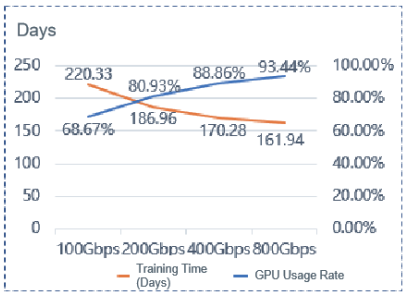
Solution Background

As ChatGPT applications continue to gain immense popularity, companies worldwide have been racing to introduce their own large-scale model products or services. These models are trained with parameter scales ranging from hundreds of billions to trillions, and the underlying computing capability has reached an impressive scale of tens of thousands of GPUs.

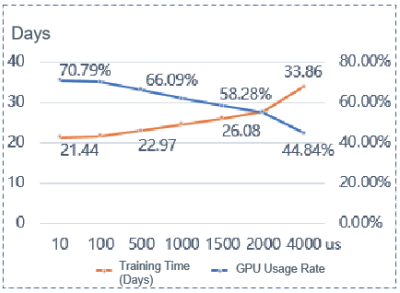
Computing power network, as the backbone infrastructure for supporting these large-scale model services, are facing with significant challenges on flexibly supporting cluster networking ranging from thousands to tens of thousands of GPUs, ensuring continuous and stable operation of services without interruption, and improving the computing power efficiency per unit cost in the service clusters.

In large-scale AI clusters, network factors affecting AI training efficiency include access bandwidth, bandwidth utilization, dynamic delay, and packet loss rate.

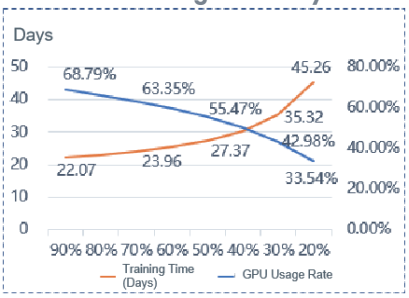
The influence of access bandwidth on training efficiency



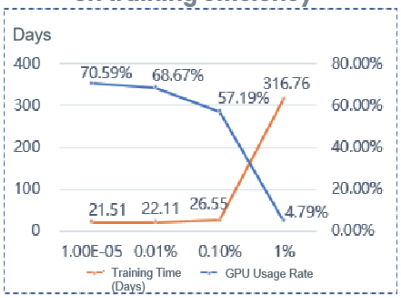
Effect of dynamic delay on training efficiency



The effect of bandwidth utilization on training efficiency



Effect of packet loss rate on training efficiency



Reference: Deepak Narayanant*, Mohammad Shoeybit, Jared Caspert, Patrick ILeGresleyt,Mostofa Patwary, Vijay Korthikantit, Dmitri Vainbrandi, Prethvi KashinkuntitJulie, Bernauert, Bryan Catanzarot, Amar Phanishayee*, Matei Zaharria*NVIDIA * Stanford University

Understanding the significance of network in the AI-Generated Content (AIGC) scenario, Micas has pioneered the **AI-Fabric solution.** The solution offers high throughput, large bandwidth, and high availability. It can be applied to various service scenarios like big data processing, machine learning, and AIGC. It helps customers build NIC-level intelligent computing centers and supports the rapid growth of AI services.

Solution Highlights

1. Computing cluster scalable to tens of thousands of GPUs based on POD expansion

- A single-tier network supports 3456 GPUs.
- POD-based flexible scalability is supported.
- A three-tier network supports up to 17,000 GPUs.

2. High bandwidth utilization based on cell forwarding

- The data stream is split into equal-length cells and balanced across all links.
- No manual intervention is required for optimization, and the business applications are unaware of any changes being made.

3. High service throughput based on VOQ + credit mechanism

- Based on per-service port (GPU) queue back pressure, independent scheduling of communication between different GPUs prevents mutual interference.
- Compared to InfiniBand's per-hop credit approach, it achieves global stability.
- Based on end-to-end credit mechanism, it ensures high efficiency, low latency, and lossless forwarding in the network.

4. Uninterrupted service operation using comprehensive software and hardware solutions

- The distributed OS control plane eliminates the risk of centralized failures and reduces the fault domain.
- Ramon achieves a latency of below 500 ns by device ID-based high-speed routing.

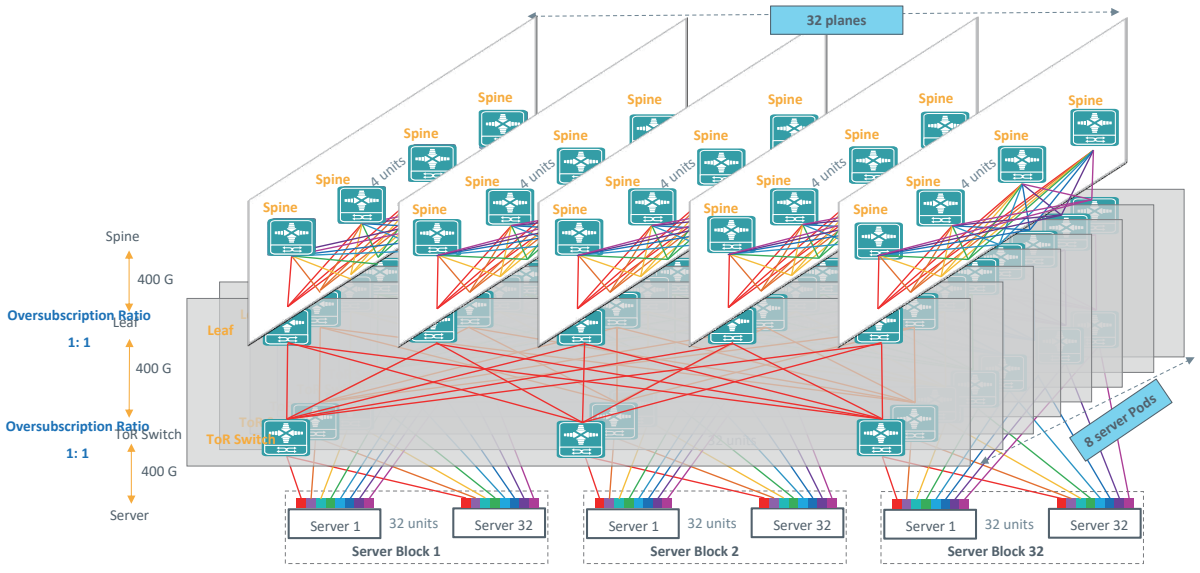
Solution Architecture

Based on the current J2C + Ramon chip

- 17,000 x 400G (17,000 GPU)

Based on the next-gen J3 + Ramon3 chip

- 32,000 x 400G (32,000 GPU)



Customer Benefits

1. Hyper-scale computing power clusters

- Large-scale computing power clusters supporting 17,000 GPUs
- Support for networks of different sizes for flexible expansion;

2. Continuous stable operation

- Based on a distributed operating system control plane, it reduces the fault domain and ensures continuous and stable operation of computing power applications.

3. Maximized computing power efficiency

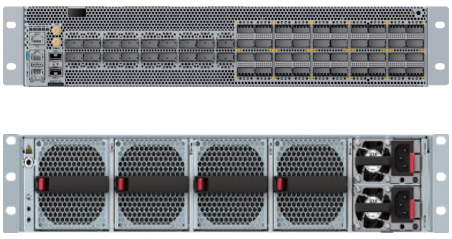
- Using the cell slicing technology, it optimizes link load balancing, improves bandwidth utilization, and reduces the completion time of service flows.
- By adopting Virtual Output Queue (VOQ) and Credit mechanisms, it achieves lossless data transmission in the fabric, resolves the problem of packet loss and retransmission, ensures continuous high-throughput forwarding of services, and enhances GPU utilization in computing power clusters.

Featured Products

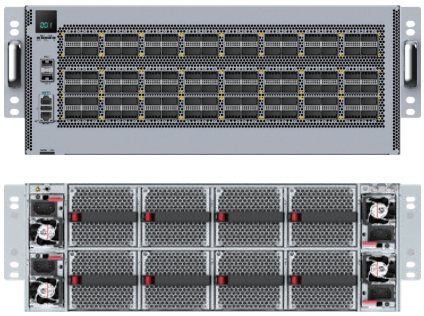
Product Role	Product Name	Description	Chip Solution
NCP	M2-S6820-18QC40F1 (Prototype)	2 U product, 18 × 400G service ports 40 x 200G Fabric inline ports	J2C+
NCF	M2-X56-96F1 (Prototype)	4 U product, 96 x 200G Fabric inline ports	Ramon

Front Panel

Rear Panel



M2-S6820-18QC40F1



M2-X56-96F1

For more information

+1(669)666-7653 | info@micasnetworks.com

250 W Tasman Dr. Ste 170, San Jose, CA 95134

www.micasnetworks.com or contact your local Micas Networks sales representative