MetalSoft Reference Architecture

Flexible, On-Prem, GenAI Infrastructure

Executive Summary

The GPUs used for Generative AI training, fine tuning or inference are generally expensive resources and to ensure their efficient utilization they must often be shared between various groups leading to frequent changes of configuration. Time to value is also essential leading to the need to automate the management of these clusters.

LLM model training requires many GPUs. The GPT3 model from 2020 required 355 GPU-years on Nvidia V100. The next-generation LLMs likely require over 30,000 GPUs of computing power to finish training within a reasonable time[1].

The networking setup of these clusters can be very different as the clos design doesn't work very well. Each GPU has its own network connection, and each server typically has 8 GPUs leading to a very difficult configuration to manage manually.

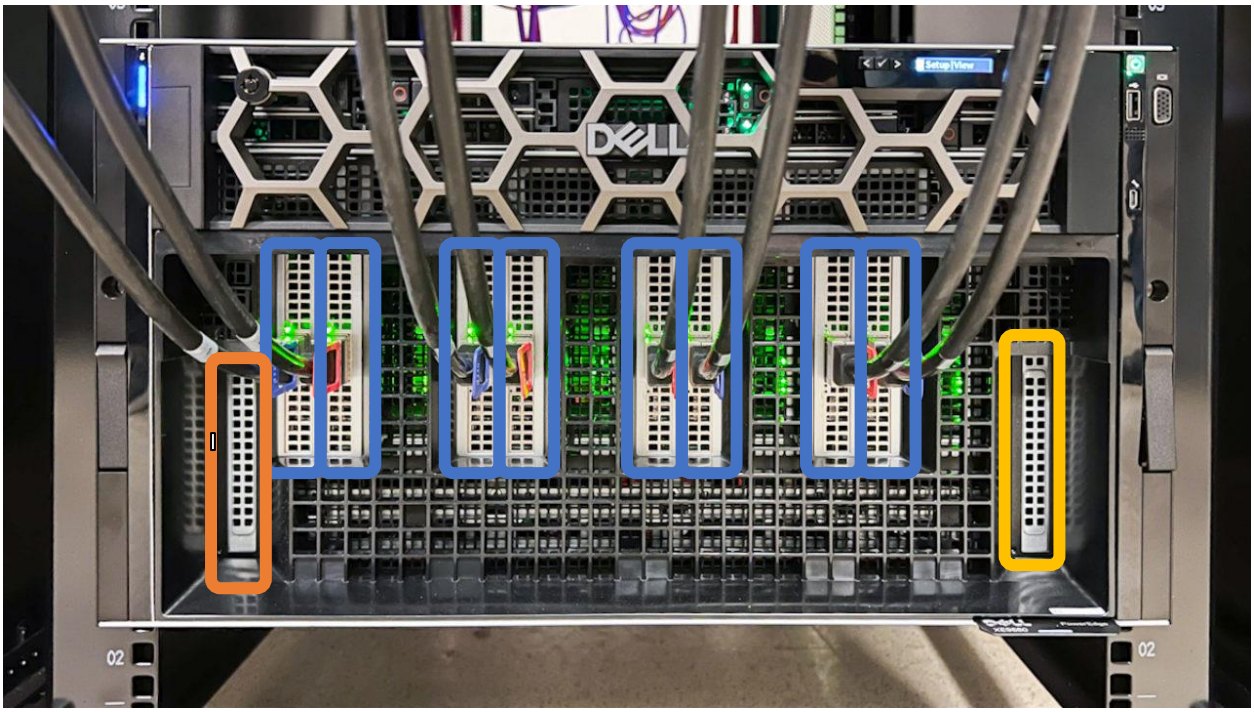


Fig 1: XE9680 server with 8 GPUs, 8 backend connections, two slots available for OOB and

[1] **Weiyang et All. MIT, Meta- Optimized Network Architectures for Training Large Language Models With Billions of Parameters** https://people.csail.mit.edu/ghobadi/papers/rail_llm_hotnets_2023.pdf

Frontend.

The MetalSoft Data Center-as-Code software enables end-to-end lifecycle management of the servers, the network and the storage resources related to GPU intensive infrastructures enabling a GPU Cloud on-premises platform. This also includes the initial discovery and stand-up of the cluster, which is fully automated.

This reference architecture presents how a network with hundreds of GPU servers will look from a network perspective and from an automation perspective. The setup is vendor-agnostic and can be built with any vendor supported by MetalSoft.

This reference architecture also uses a new method to cut costs while preserving network performance by using a rail-based network design instead of a traditional any-to-any clos network which is up to 75% more cost effective while delivering comparable performance.

Backend Network

State of the art setups require each GPU (in this reference architecture A100 or H100) to have its own 400Gbps connection with a multi-level clos network. However, recent research[1] shows that LLM-optimized network setups that uses a designed called “Rail-based networking”. In this setup, instead of a multi-level clos network interconnecting all nodes, each GPU is connected to its own isolated clos network. GPU1 of all servers are connected to “Rail 1”, GPU2 of all servers is connected to “Rail 2” and so forth. This can save up to 75% of switch and transceiver costs while preserving or improving performance.

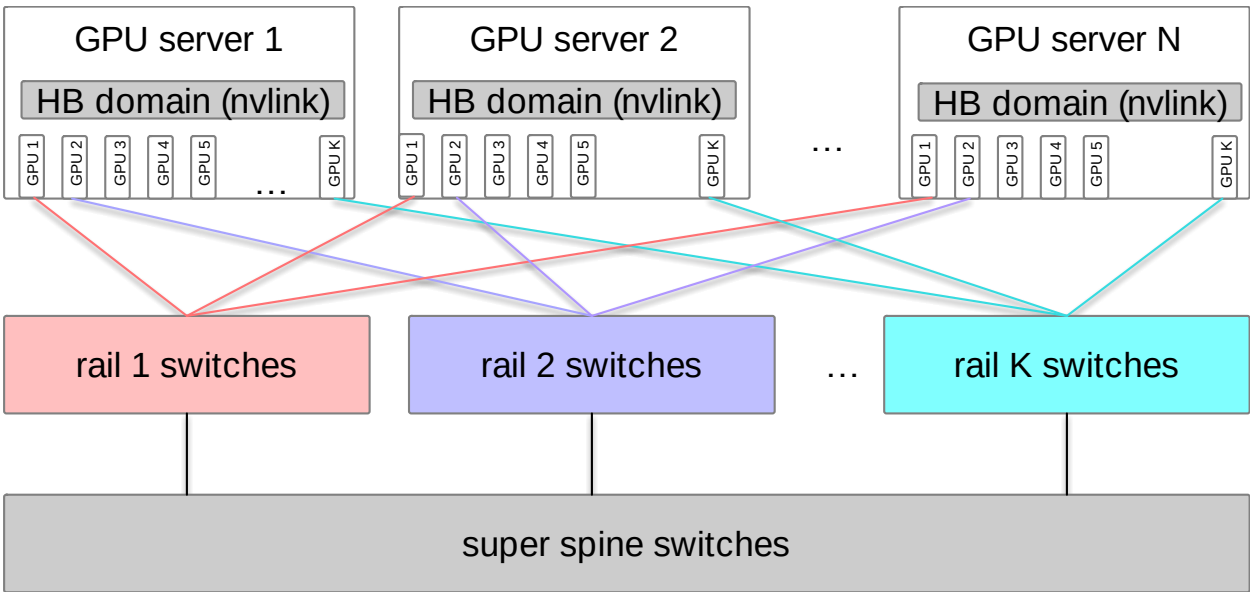


Fig 2: Backend network

[1] [1] **Weiyang et All. MIT, Meta- Optimized Network Architectures for Training Large Language Models With Billions of Parameters** https://people.csail.mit.edu/ghobadi/papers/rail_llm_hotnets_2023.pdf

Note that this approach uses ROCEv2/RDMA which allows each GPU to directly communicate on the backend network bypassing the CPU.

Each “Rail” network is an independent clos network:

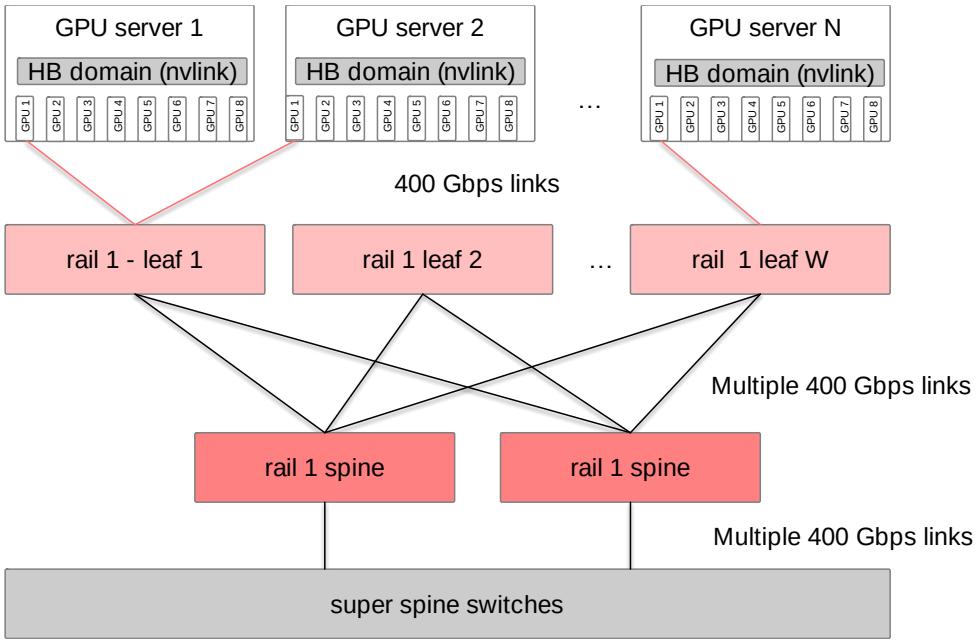


Fig 3: Backend network – single rail

Note that we suggest using a single 400Gbps NIC per GPU to avoid LACP or other forms of hashing-based load balancing which are often sub-optimal given a relatively low number of flows and to also reduce the transceiver costs.

For a 16384 GPU cluster (2048 nodes), for each “rail”, 86 leaf switches (Dell Z9332F-ON) each using 24 interfaces for downlinks and 8 for uplinks and 12 spine switches. For 8 GPU nodes there will be 8 such “rails”. This will mean a total of 784 switches. Other designs might also be used but we chose a relatively low-radix switch to reduce ECMP hashing related issues.

Frontend Network

The control traffic which is less intensive is usually done via separate, 10Gbps links leading to two leaf switches that are usually connected to a “spine” layer.

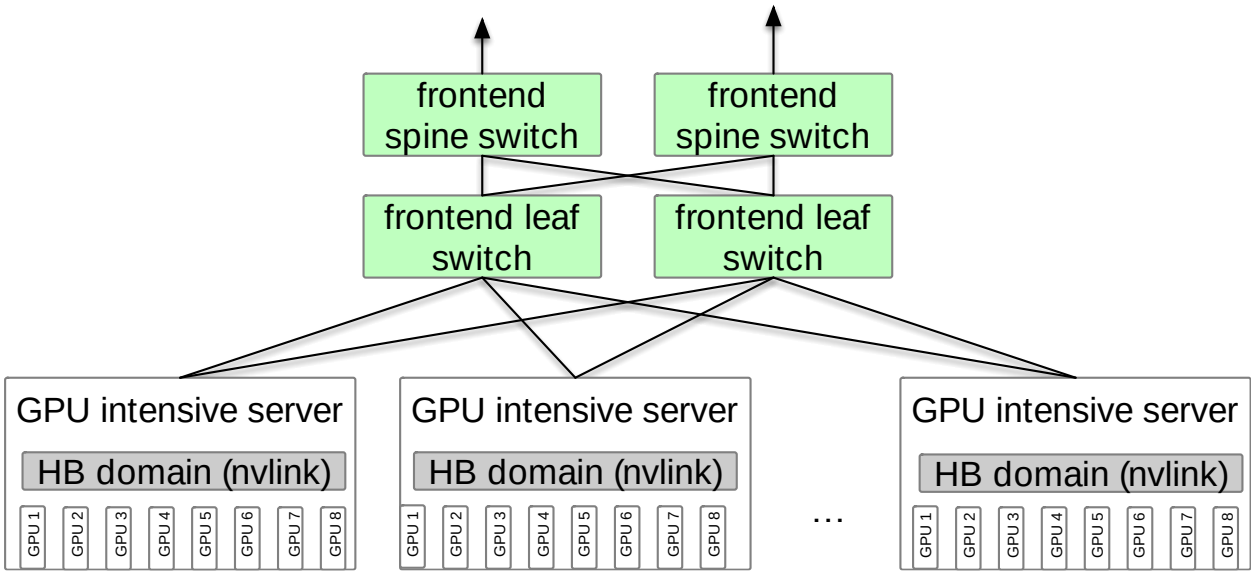


Fig 4: Spine-leaf with n GPU intensive servers, such as H100s.

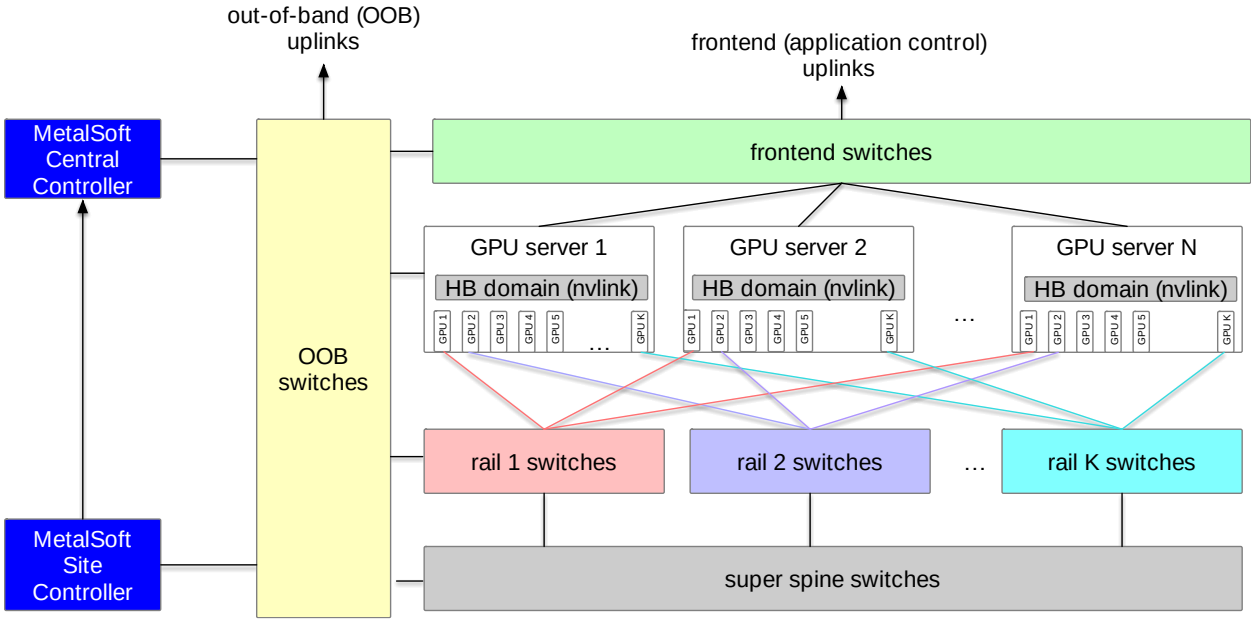
This layer can be scaled to 100 Gbps links if storage traffic is also required.

Management Network (Out of Band)

To manage the entire solution, deploy operating systems, perform storage allocations, configure LANs, and more, we recommend running MetalSoft out of band (OOB).

The control plane of this solution is a small scale, non-redundant deployment of MetalSoft. The MetalSoft Central Controller is the component that offers the APIs to the users while the site controller is the component that interacts with the equipment. A MetalSoft Central Controller

can manage multiple such Pods, in multiple locations via local site controllers.



If a redundant setup is required, a larger cluster can be configured with a minimum of 3 nodes, which can be virtual machines. In many situations these are running in a separate virtualized cluster.

To enable zero touch provisioning, the MetalSoft site controller must be able to receive DHCP broadcast traffic from the management interfaces, thus either L2 connectivity (same VLAN), or DHCP relay must be configured.

The MetalSoft solution can configure the Management switch as well, only the Central Controller and site controller combination is required to stand-up the cluster.

Example configuration with Dell hardware (but other configurations can be built with other vendors):

- Dell XE9680 with:
  - 8 x NVIDIA H100 700W 80GB SXM5 / A100 80GB SXM4 GPUs
  - 2TB RAM
  - 8x 1.6TB Samsung E3.S NVMe PM1743
  - 8 x Mellanox ConnectX-7 400Gbps Single port 900-9X766-003N-SQ0 OSFP
- "Rail" leaf switch Dell Z9332F-ON 32 x 400GbE QSFP56-DD ports
- "Rail spine" & "Super spine" switch Dell Z9664F-ON 64 x 400GbE QSFP56-DD ports
- "Frontend" leaf switch Dell S5296F-ON 96 10Gbps ports 8 QSFP28
- "Frontend" spine switch Dell S5232-ON 32 QSFP 28

References

Weiyang et All. MIT, Meta- Optimized Network Architectures for Training Large Language Models With Billions of Parameters
https://people.csail.mit.edu/ghobadi/papers/rail_llm_hotnets_2023.pdf

Dell RDMA over Converged Ethernet (RoCE) Best Practices
https://infohub.delltechnologies.com/l/dell-enterprise-sonic-networking-use-case-guidebook/rdma-over-converged-ethernet-roce/

Nvidia RDMA over Converged Ethernet (RoCE)
https://docs.nvidia.com/networking/display/mlnxofedv461000/rdma+over+converged+ethernet+(roce)#src-12013422_safe-id-UkRNQW92ZXJDb252ZXJnZWRFdGhlcm5ldChSb0NFKS1Sb0NFTEFHKENvbm5lY3RYLTMvQ29ubmVjdFgtM1Bybyk