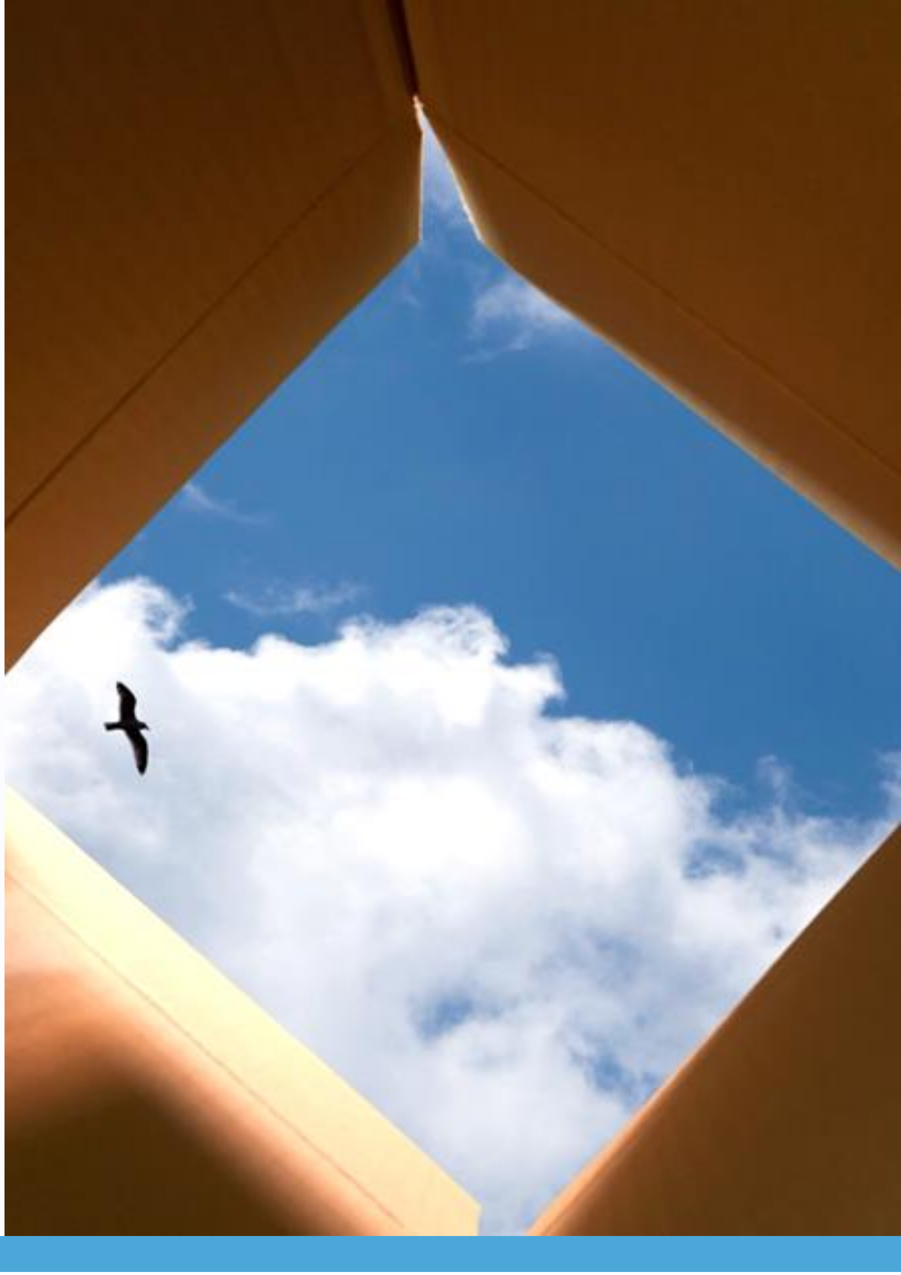
The Scaling Dilemma: Interconnects for AI Infrastructure

Konstantin Rygol – Manager, AI Team – krygol@gigaio.com



Background LLMs

- Large Language Models (LLMs) have emerged as a critical workload
- LLM Training and Inference both run on GPU clusters
  - 8xGPUs per node (Nvidia's HGX/AMD's MI300X Platform)
  - NVlink/Infinity Fabric for intranode 'scale-up'
  - RoCE or InfiniBand for internode 'scale-out'
- Performance is often bottlenecked by GPU-based collective communication
  - Collective communication libraries (NCCL/RCCL) use a multi-ring based algorithm

Scale UP vs Scale Out

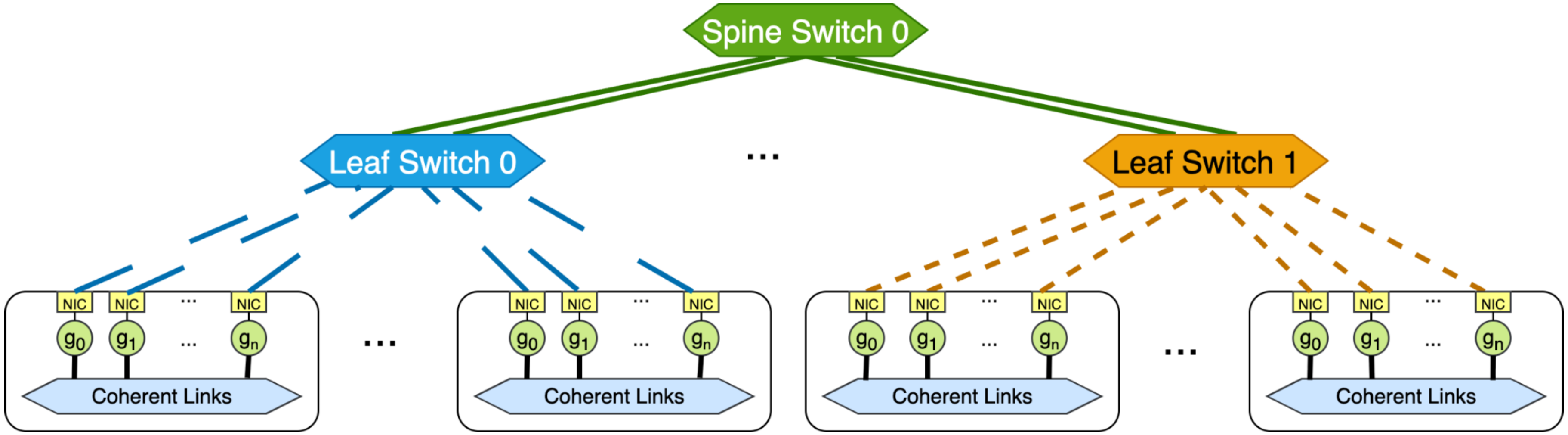
- Scale Up
  - Easier to deploy server, easier to deploy software
  - Easier to maintain software
  - Is usually limited to a certain size of compute resources
- Scale Out
  - Used when Scaling UP is not sufficient anymore
  - Requires Orchestration
  - Requires concepts for distributed maintenance of software
  - Requires multi node software stacks and tuning of the interconnect

⇨ Scale up as long as possible

Background PCIe vs InfiniBand/ROCE

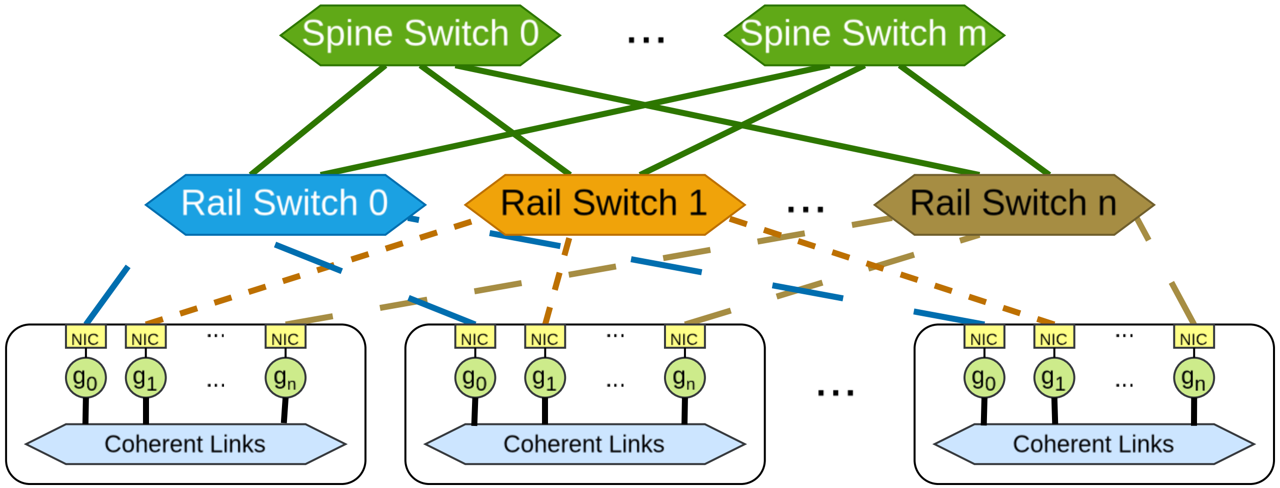
- InfiniBand/ROCE
  - High bandwidth up to 400Gb/s
  - Microsecond Latency
  - Requires kernel-space-drivers – additional tuning might be required
  - Switched fabric great variety of topologies
- PCIe
  - High bandwidth up to 512Gb/s
  - Nanosecond Latency
  - Expose through standard OS
  - Tree topology limited flexibility - Crosslinks are not easily achievable

Background Fat Tree



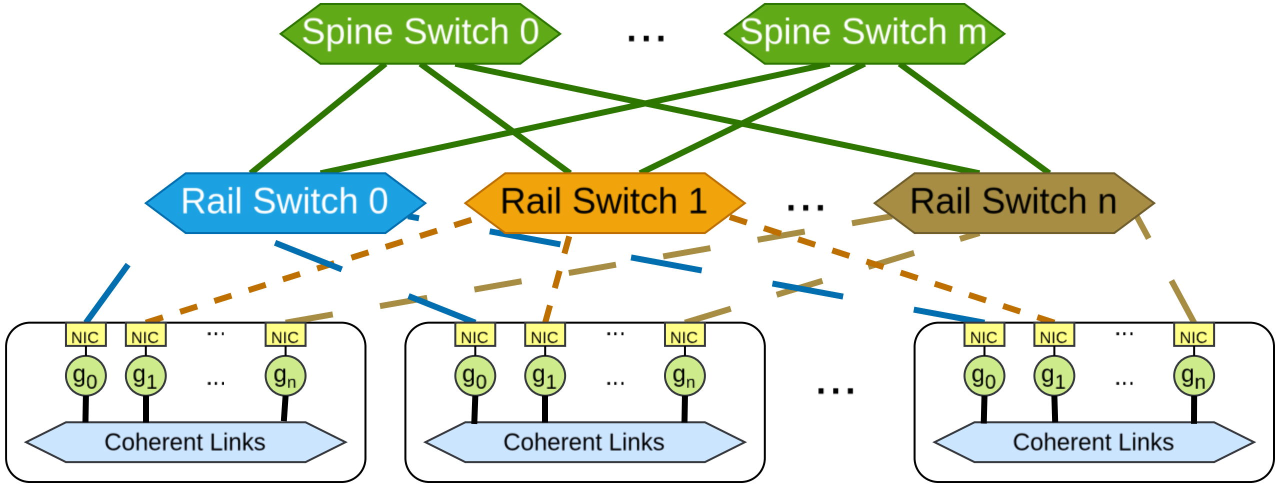
- Fat Tree
  - Legacy network architecture stemming from multi tenant HPC environments
  - First level of parallelization is within a node on the coherent links
  - The second level of parallelization is among all nodes under the same leaf switch
  - For the third level of parallelization date is routed through the spine switch
  - The leaf-spine uplink ratio determines the blocking factor
  - For large deployments it requires many switch hops and becomes expensive

Background Rail Optimized Networks



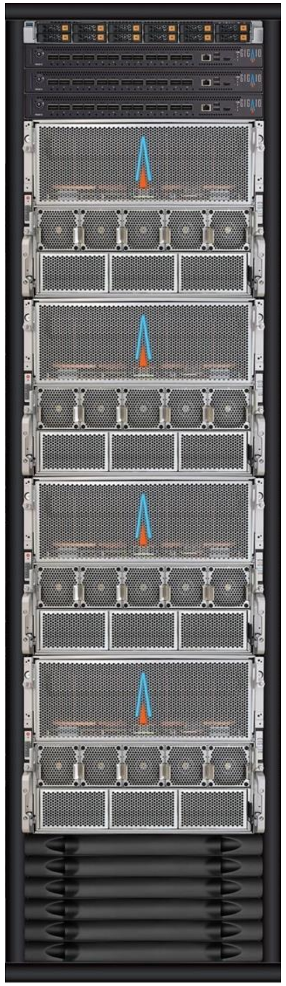
- At a high-level, clusters have a fixed number of GPUs per node, connected by a high-bandwidth GPU<->GPU links, and multiple NICs per node
- Rail-Optimized Networks have emerged as an LLM/AI focused topology for RoCE/InfiniBand
  - For a Node with n NICs, define rails {0, …, n-1} where NIC i is associated with rail i
  - All NICs in a rail are attached to the same leaf switch
  - A spine-layer can provide all-to-all connectivity

Background Rail Optimized Networks



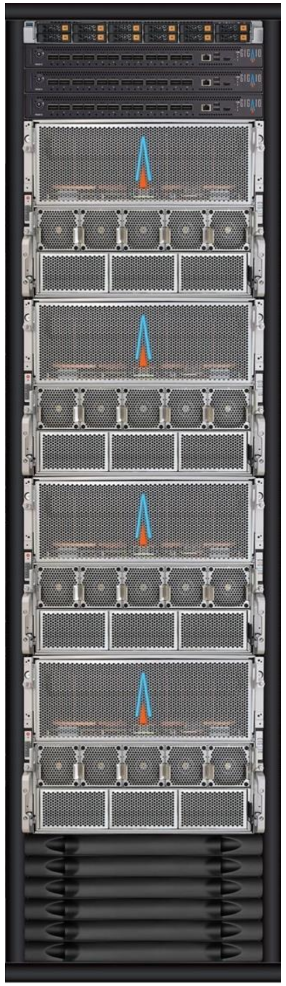
- NCCL/RCCL schedules multiple rings
  - All cross-node traffic takes a single switch-hop
  - Cross rail traffic happens over NVLINK/XGMI
  - Schedule multiple rings on independent rails
  - Leverage PCIe P2P tools like GDRDMA

Composable Disaggregated Infrastructure and FabreX



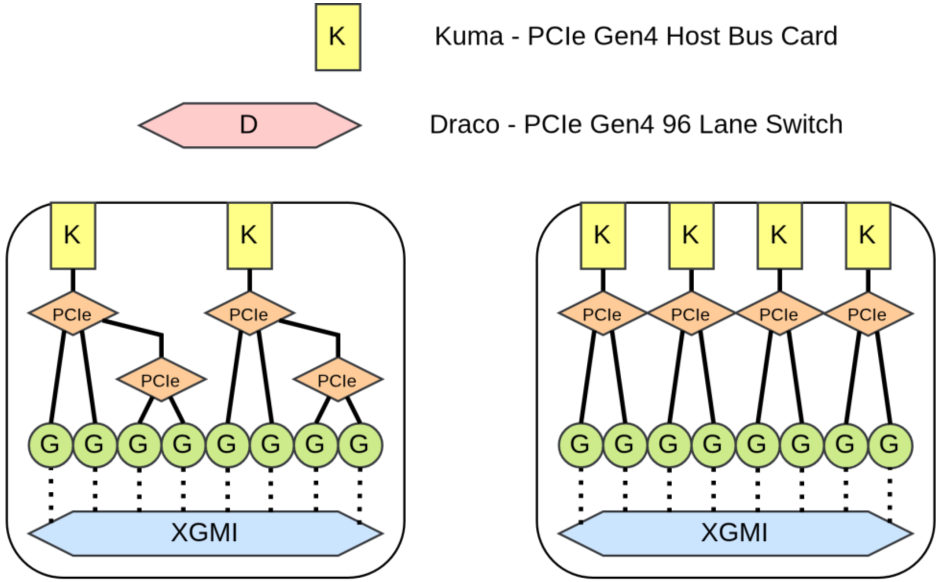
- Modern GPU clusters are disaggregated systems
  - GPU nodes have a fixed number of GPUs with a scale-up network like NVLink or XGMI
  - Remote nodes are accessed over a 'scale-out' network, like RoCE or InfiniBand
- GPU Clusters are complicated to manage
  - They require distributed file-system (Luster, BeeGFS, etc...)
  - Job schedulers (Slurm, Kubernetes, etc...)
  - Multi-Node programming models (MPI, SHMEM, etc...)

Composable Disaggregated Infrastructure and FabreX



- Composable Disaggregated Infrastructure (CDI) adds more GPUs to the 'scale-up' network
- We use FabreX, which is conceptually a 'PCIe Network’
- Breaks up a node into individual resources
  - The 'Host' contains the CPU and RAM
  - GPUs are housed in Accelerator Pooling Appliances (APA)
- Top-of-rack PCIe switches connect resources together

Composable Disaggregated Infrastructure and FabreX

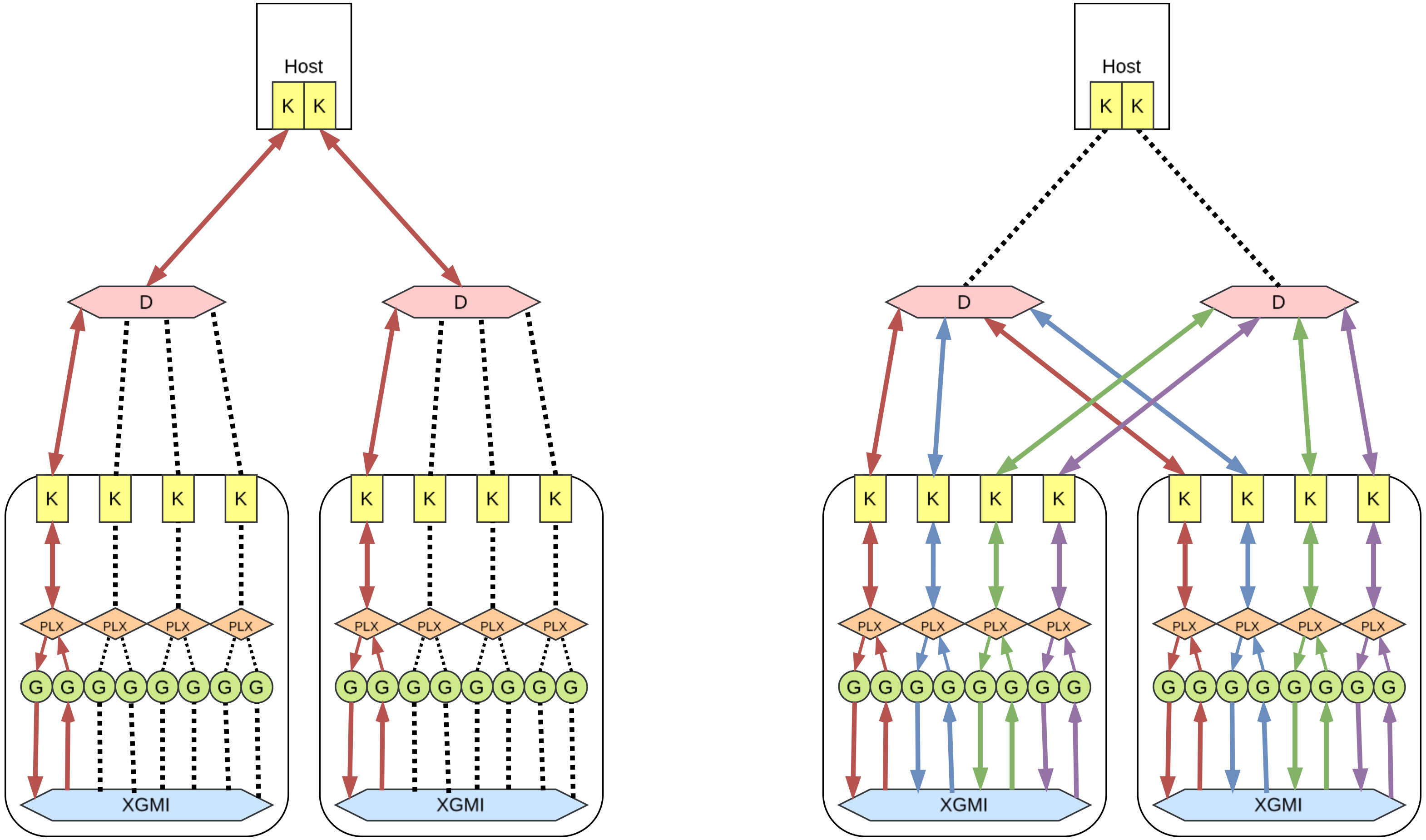


MI300X Columba - Cascade Configuration

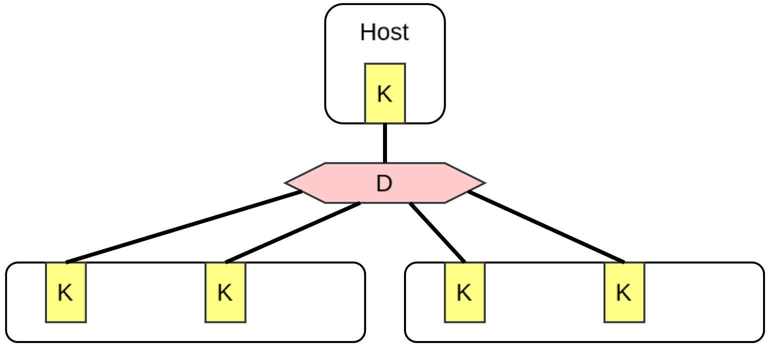
MI300X Columba - Flat Configuration

- Used FabreX PCIe Gen4 platform
- *Kuma card*, converts a PCIe slot to Mini SAS HD cables
- *Draco* switch, up to 6 PCIe Gen4x16 connections
- AMD MI300X APA, codename *Columba*, comes in two configurations:
  - *Cascade*: 2-*Kumas* per box
  - *Flat*: 4-*Kumas* per box
- *What is an optimal PCIe topology for LLMs?*

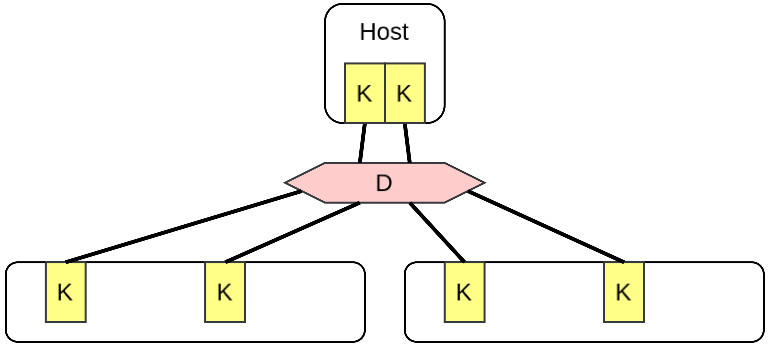
Standard Vs Rail-Optimized PCIe Network



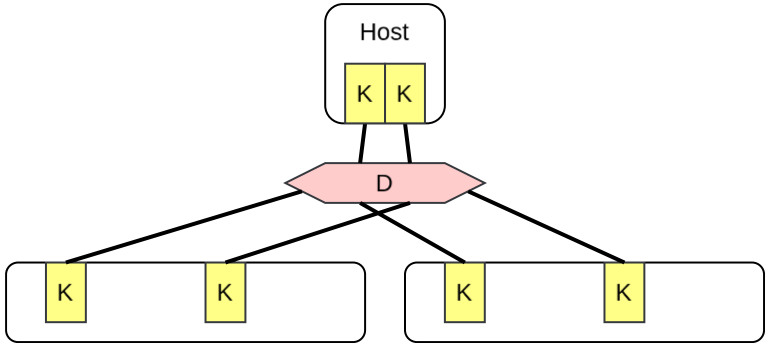
Five Topologies we Investigated



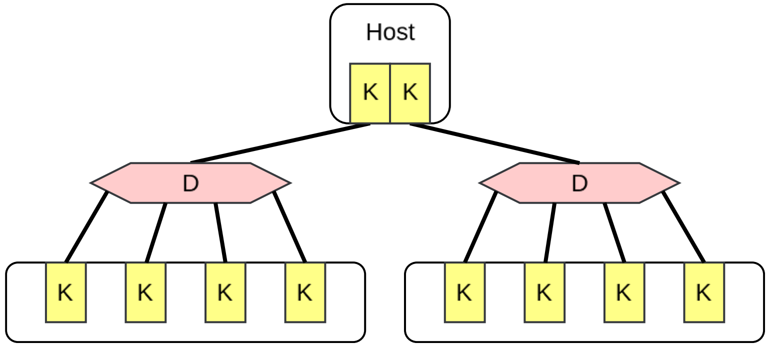
1s-1rp-std-cascade



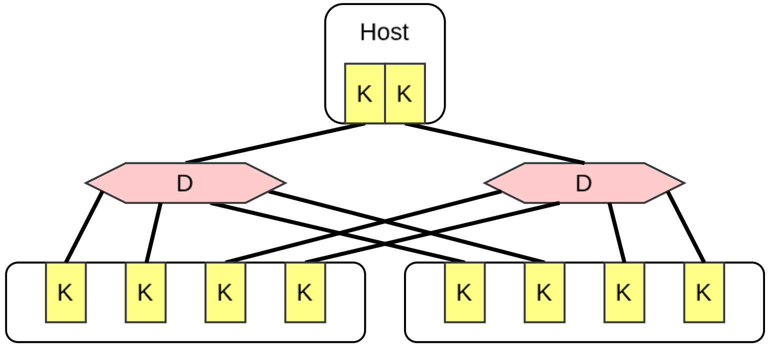
1s-2rp-std-cascade



1s-2rp-split-cascade



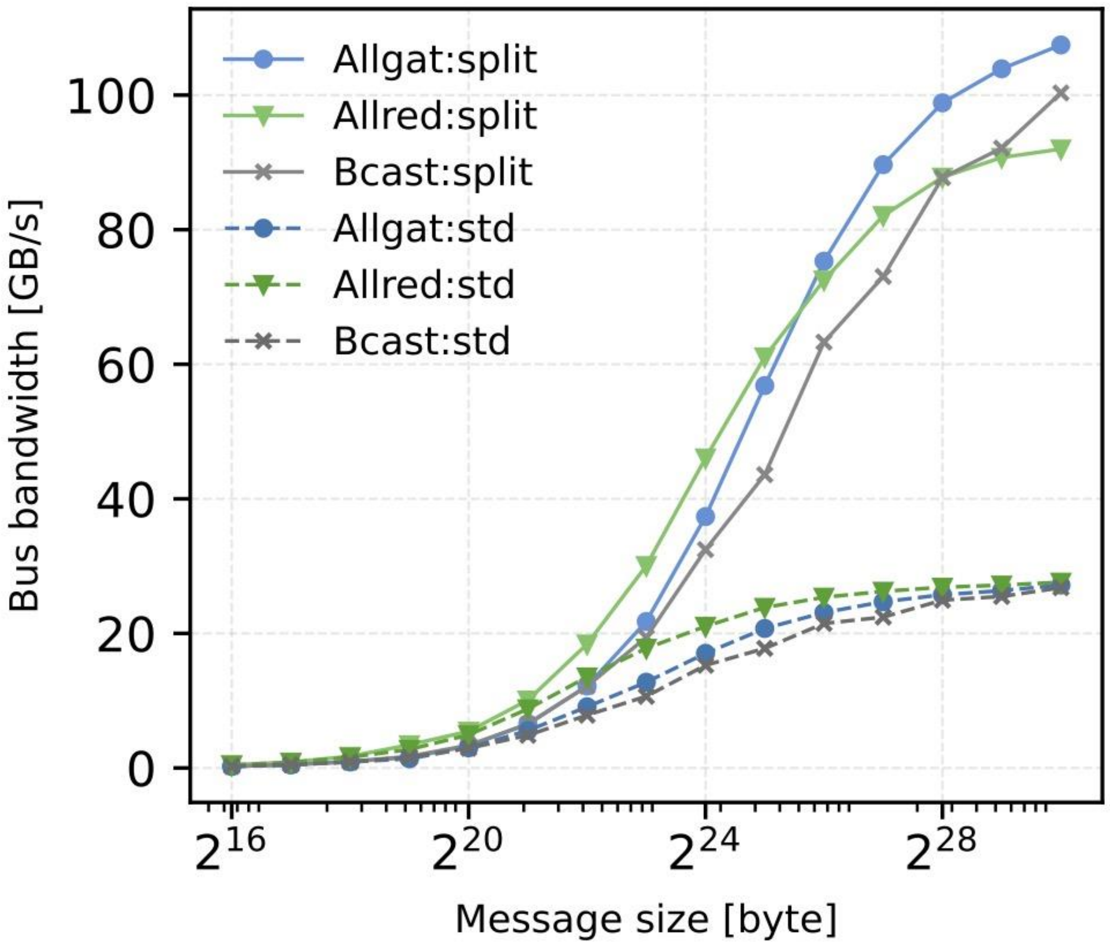
2s-2rp-std-flat



2s-2rp-split-flat

Results Default vs Split Topology

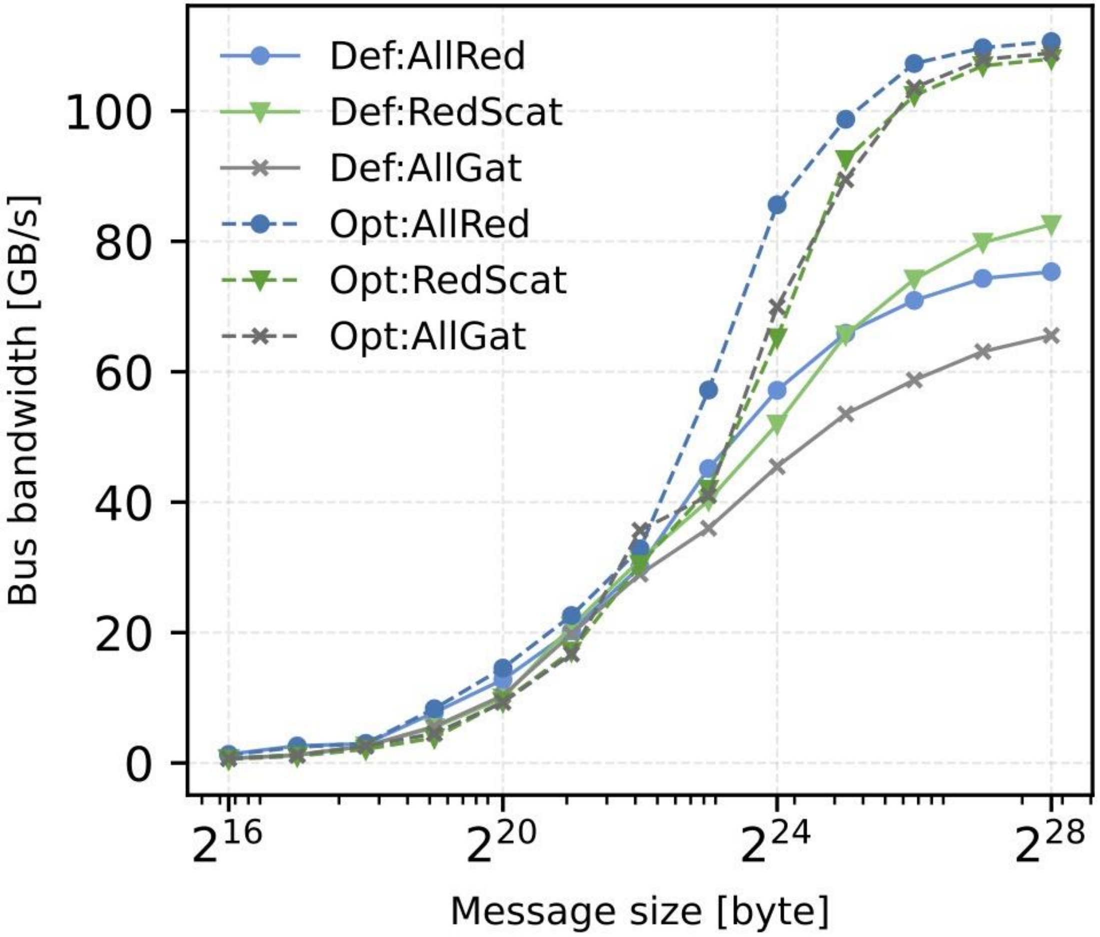
Bus bandwidth 2s-2rp-std-flat topology vs the 2s-2rp-split-flat topology



GB/second	All_Gather	All_Reduce	Broadcast
1s-2rp-split-cascade	56	54	54
1s-1rp-std-cascade	56	54	54
1s-2rp-std-cascade	25	25	25
2s-2rp-split-flat	**92**	**100**	**107**
2s-2rp-std-flat	27	27	27

Results Default vs Optimized RCCL Rings

Bus bandwidth Default vs the optimized RCCL rings on the 2s-2rp-split-flat



GB/second	Default	Optimized	Relative Change
All_Reduce	75	88	17%
Reduce_Scatter	83	108	30%
All_Gather	66	99	50%

Results LLM Training Performance

GPT-NeoX training comparing RCCL v6.1 vs Optimized RCCL on 2s-2rp-split-flat

Samples/second	ROCm-v6.1	Optimized	Relative change
1.3B z-1	78.83	88.42	10.81%
1.3B z-3	85.52	88.72	3.27%
13B z-1	12.61	14.33	12.01%
13B z-3	15.80	16.24	2.72%

TFLOPs/GPU	ROCm-v6.1	Optimized	Relative change
1.3B z-1	120.0	134.6	10.85%
1.3B z-3	130.6	135.0	3.26%
13B z-1	175.8	199.8	12.01%
13B z-3	220.3	226.5	2.74%

- ZeRO optimizations from Deepspeed
  - ○ Z-1 uses All_Gather and All_Reduce for gradients + weight updates
  - ○ Z-3 adds Broadcasts for model weights, resulting in ~40% more communication

Results Inference Performance

Llama-3.1-405B inference performance, output tokens per seconds, with VLLM

Topology	8-GPU TP	16-GPU TP
1s-2rp-std-cascade	24.58	15.86
1s-2rp-split-cascade	24.62	25.38
2s-2rp-split-flat	24.60	28.66

- Evaluated Llama-405B inference performance as a web-server using VLLM
- One process launches the server
  - Tested 16-GPU and 8-GPU tensor parallelism across different topologies
- Separate process issues request and measures the time to completion
  - 1024 random tokens are sent, 128 tokens are generated

Future Work

- Heterogenous Deployments of FPGAs, ASICs and GPUs
- PCIe Scale-up + RoCE/IB Scale-out
  - FabreX provides high performance/low latency scale up combined with ROCE/IB scale out/storage access
- PCIe Gen 5 studies
  - Microbenchmarks have shown that collective bus-bandwidth has doubled vs Gen4, as expected

SC25 BOF: The Future of Open Interconnects for AI

GigaIO AI Team,

https://ieeexplore.ieee.org/document/11018266

Benjamin Kitor – bkitor@gigaio.com

Konstantin Rygol – krygol@gigaio.com

www.gigaio.com

GIGAIO