enfabrica

ACF-S: A Novel Approach to High-Performance Data Movement in AI Compute Fabrics

September 10, 2024

Rochan Sankar, Enfabrica

AI HARDWARE & EDGE AI SUMMIT

:: what we are about

enfabrica

:: mission

redefine networking for accelerated computing to deliver peak performance, resiliency and node scale

:: team

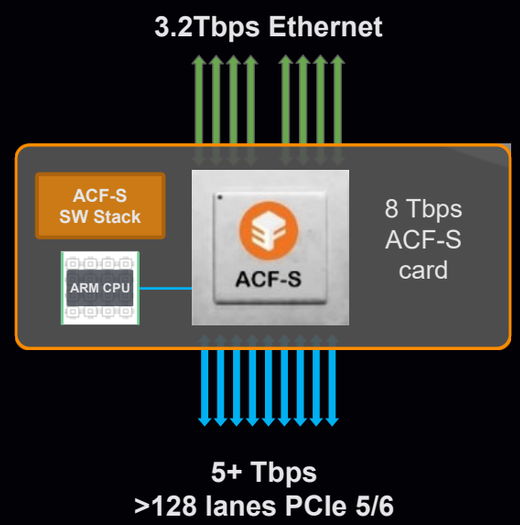
120+ engineers

previously built high-performance NICs, switches, routers, TPUs, graphics, host networking stacks

:: product

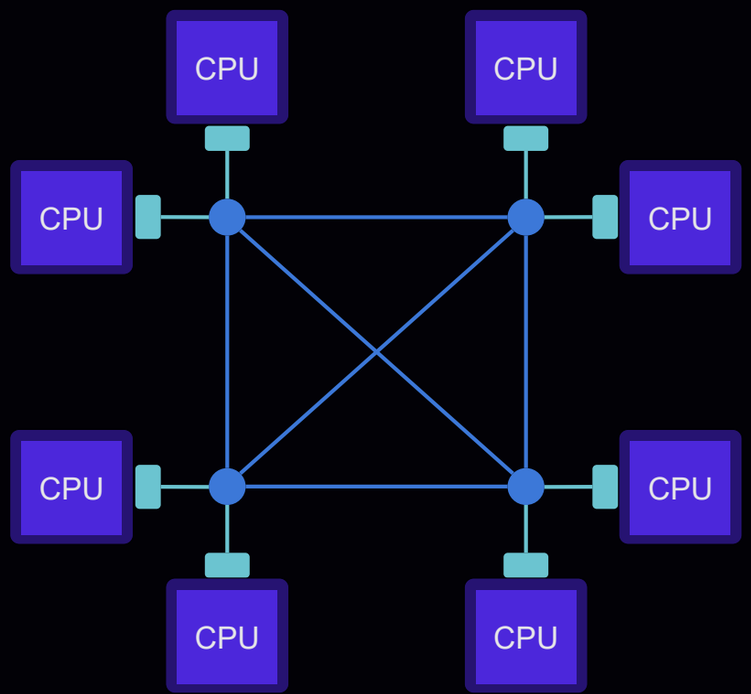
accelerated compute fabric superNIC (ACF-S)

1st chip codename ‘millennium’ @ **8 Tbps** bandwidth



:: a systems perspective

:: scale-up supercomputing // mainframe, ccNUMA



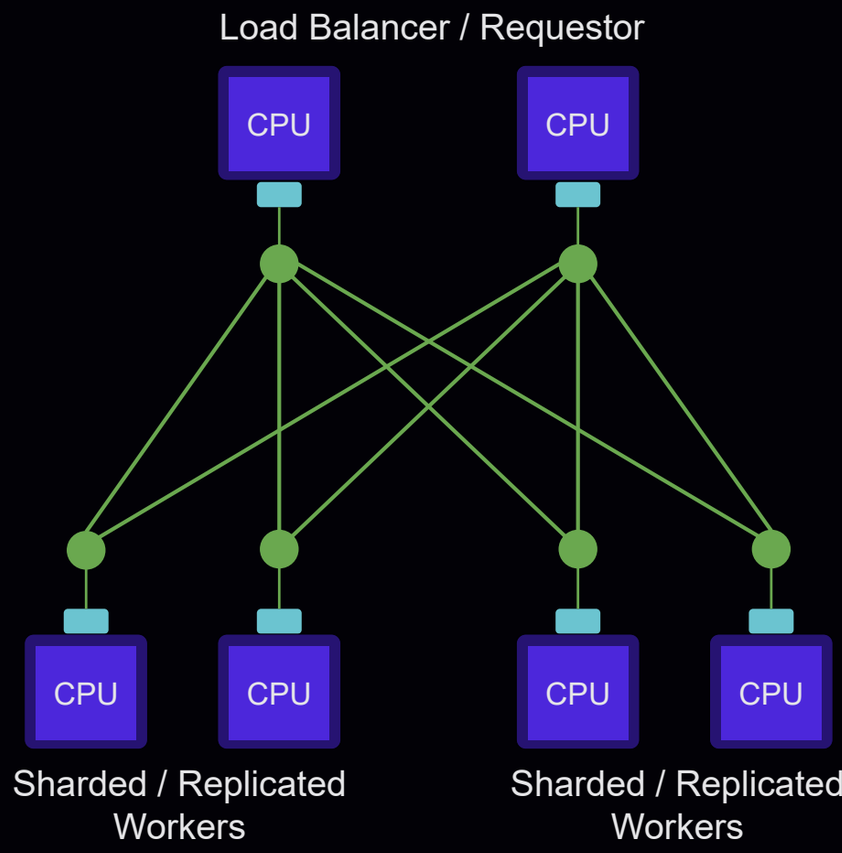
All blue links are IPC communication

Fully coherent memory system operating on a “large” problem by sharding computation

Worker nodes synchronize state and move memory closer using O(100ns) latency IPC transactions

Communication protocols deeply embedded in the processor to enable “transparent” communication

:: hyperscale cloud // the rise of scale-out computing



Client-server design, built for extreme, resilient application scaling

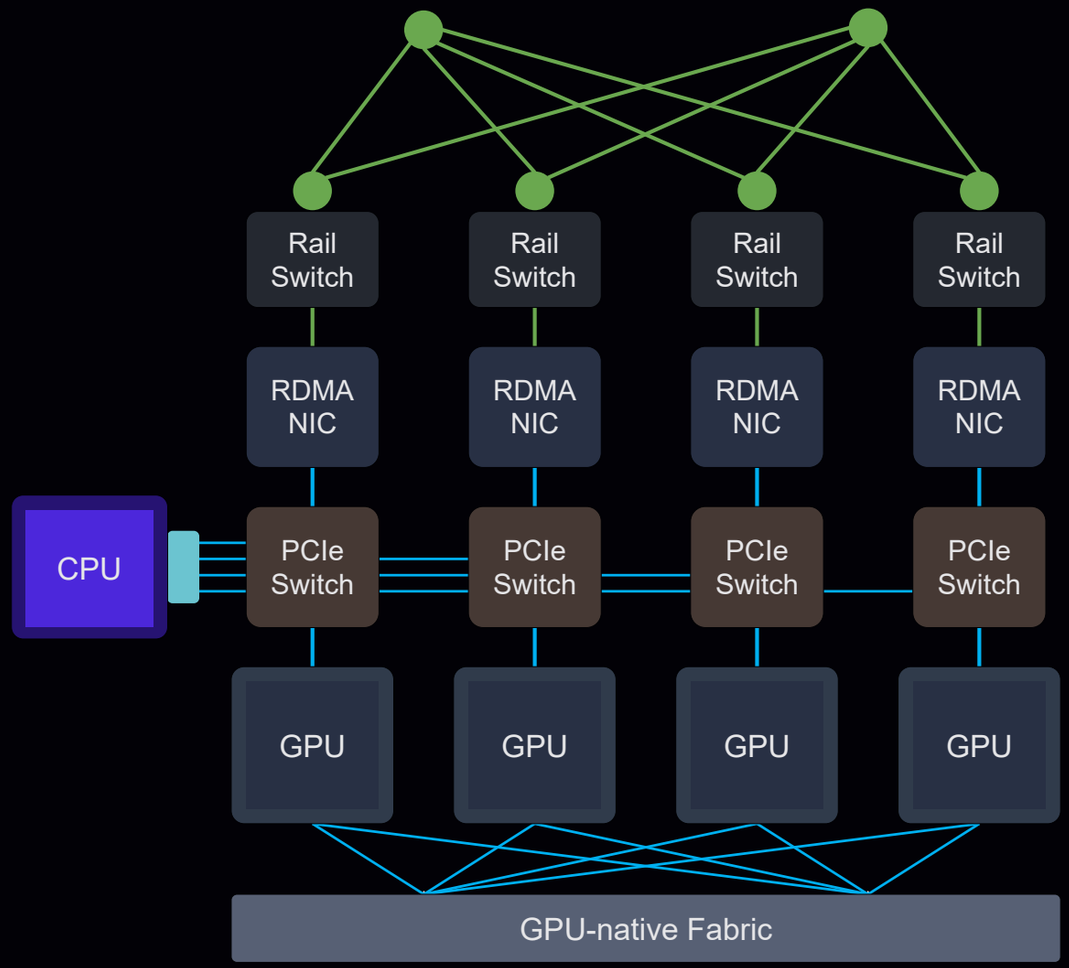
All communication uses retargetable, resilient, software managed RPCs (request-response)

Workers and data pipelines are imminently reconfigurable

Distributed, heterogenous compute nodes with

- high aggregate network throughput
- high tolerance to latency (10s - 1000s of microseconds)
- No shared-fate

:: data center AI, ML systems // super, meet hyper



Modern infrastructure systems for genAI demand:

- Tight performance of a supercomputer

plus

- Elastic scaling of compute threads and resiliency of a cloud scale system

Dual mode communication / programming models:

Scale-up :: Low level (like CUDA, the new assembly language)

Scale-out :: High level, structured kernels for computation and communication (CCLs, MPI)

:: hyperscale AI system communication challenges

Expanding IPC, RPC domains

- Scale-up defines largest tensor / vector op
- Scale-out defines largest bulk data movement
- Disparate fabrics scale disparately

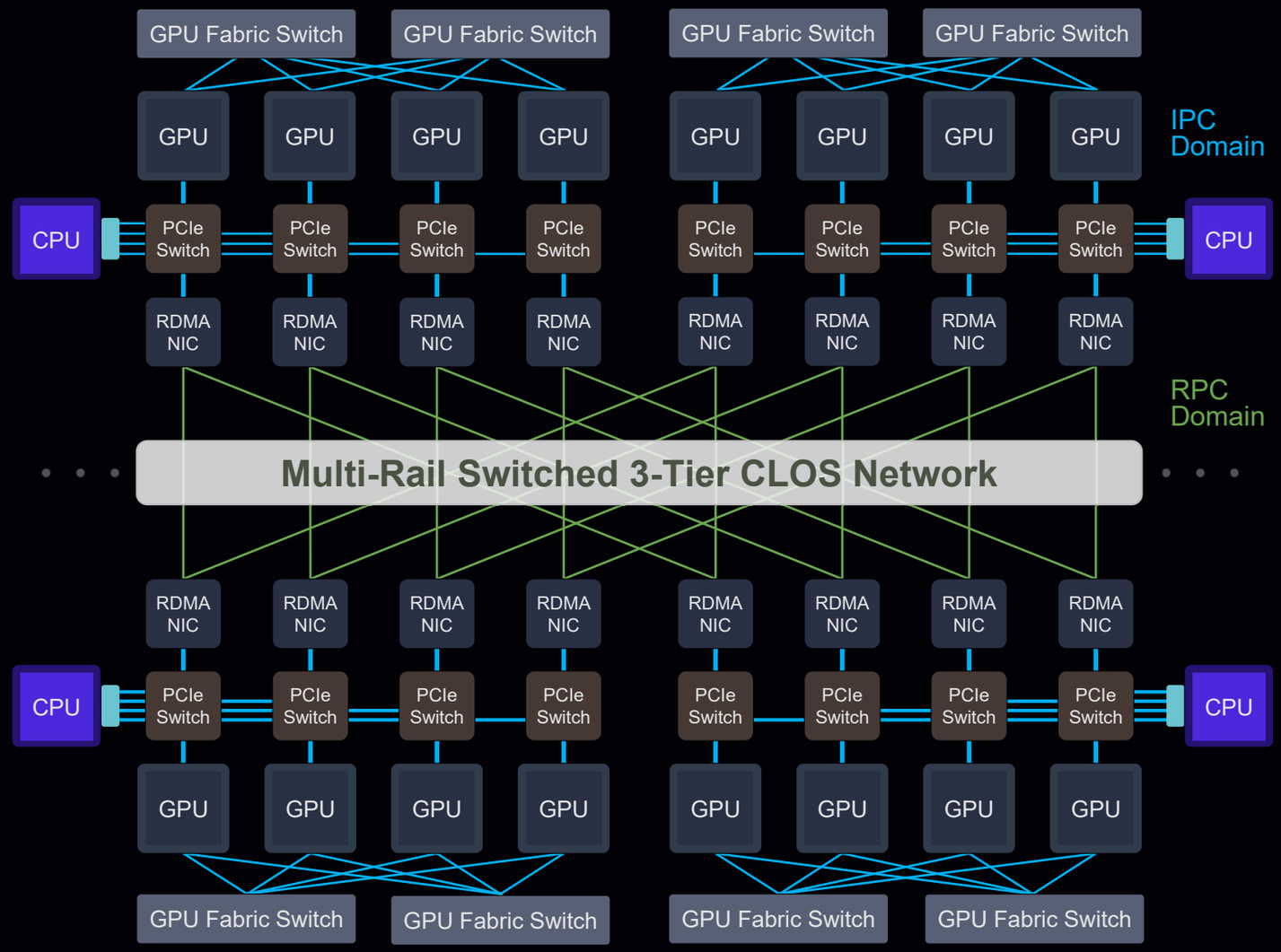
Imbalanced I/O burst bandwidth

Hardwired communication types, sizes

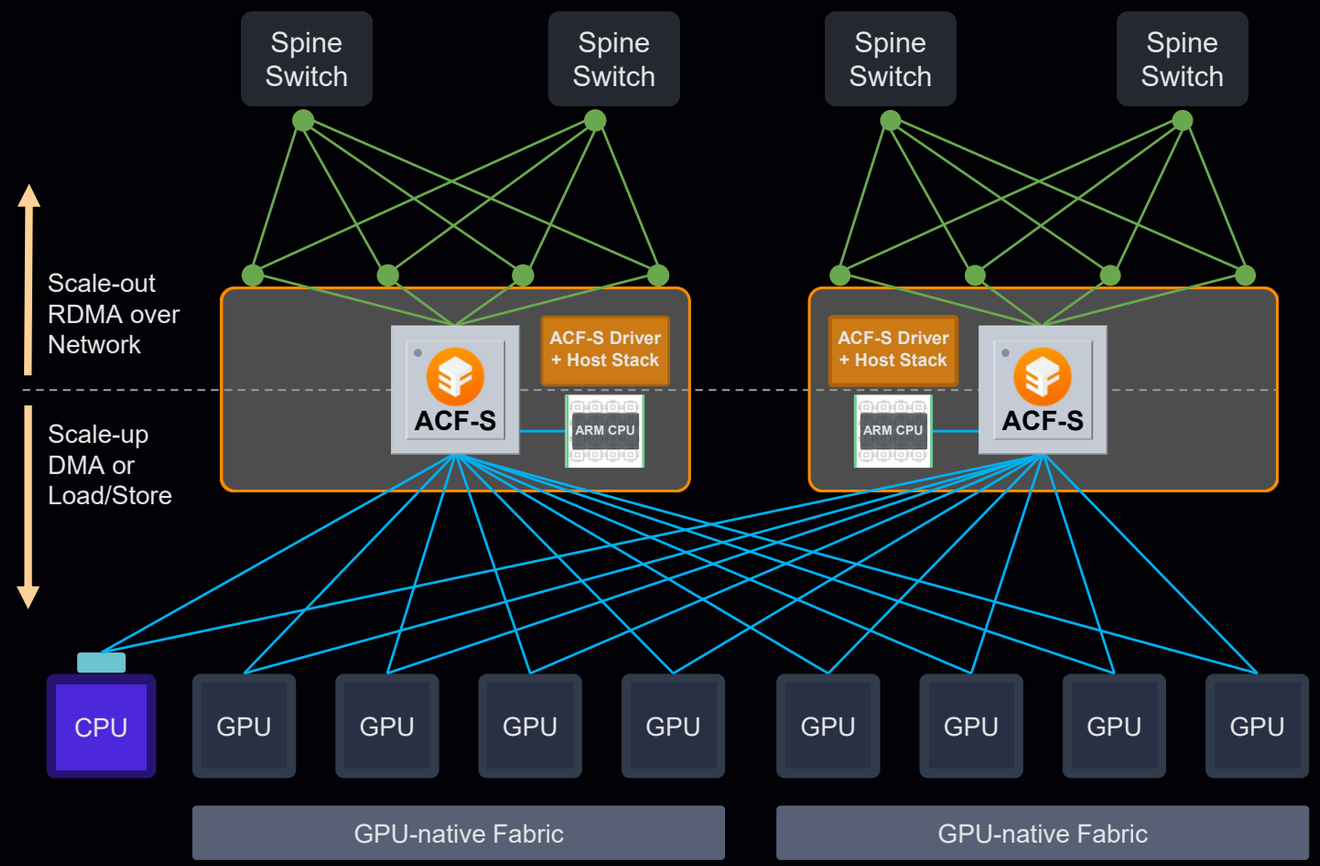
- IPC to infinity is infeasible
- Tensor parallel entirely over RPC is not efficient

Fabric scaling is rapidly reducing MFU

- Larger GPU collectives = incast
- More network tiers = latency, jitter, contention
- Network failures strand or stall GPU compute



:: refactoring the communication endpoint architecture



Systems need a high-bandwidth, fat-tree 'superNIC' standard product that :

Has scale up (IPC) interfaces AND scale out (RPC) interfaces

Has interfaces to speak memory

- *i.e., buffers, virtual addresses, DMA*

Has interfaces to speak packets

- *e.g., RoCE, Congestion Control, ECMP*

Allows software to program how memory transfers convert to network packets

- *And vice versa*

Elastically binds bandwidths of memory interfaces to network interfaces & vice versa

- *e.g., saturate 2x400G accelerator interfaces from sprayed, routed 8x100G Ethernet ports*

:: make scale-up, scale-out fabrics two sides of the same coin

Increased GPU I/O burst bandwidth

Fewer memory copies; land data anywhere in multi-GPU node memory map

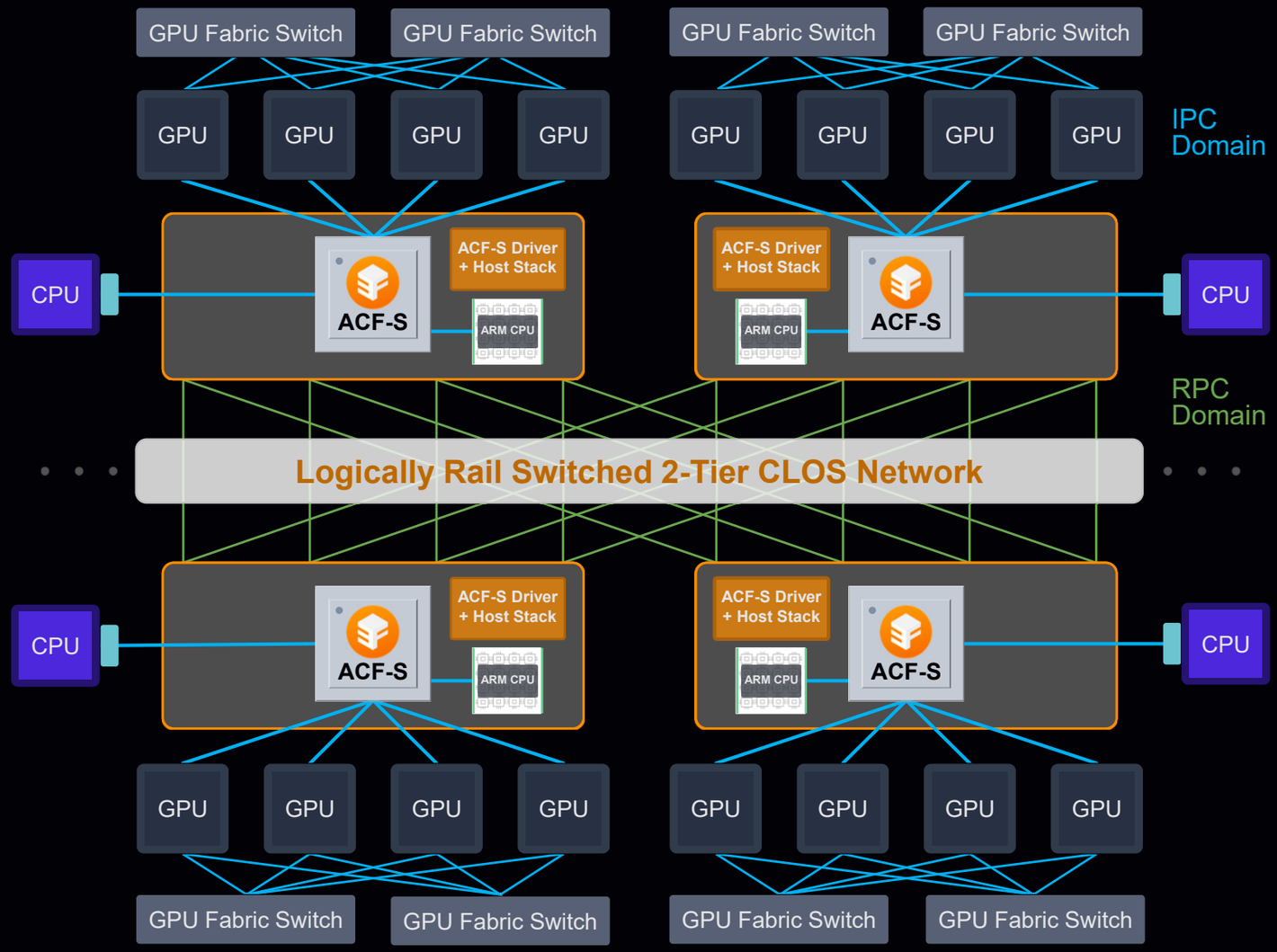
GPU collectives scale in both directions without performance cliffs

Shared endpoint absorbs GPU incast

Fewer network tiers, lower latency, reduced cabling complexity

Failure on NIC port/cable does not brick the GPU

… And the programming model didn't change

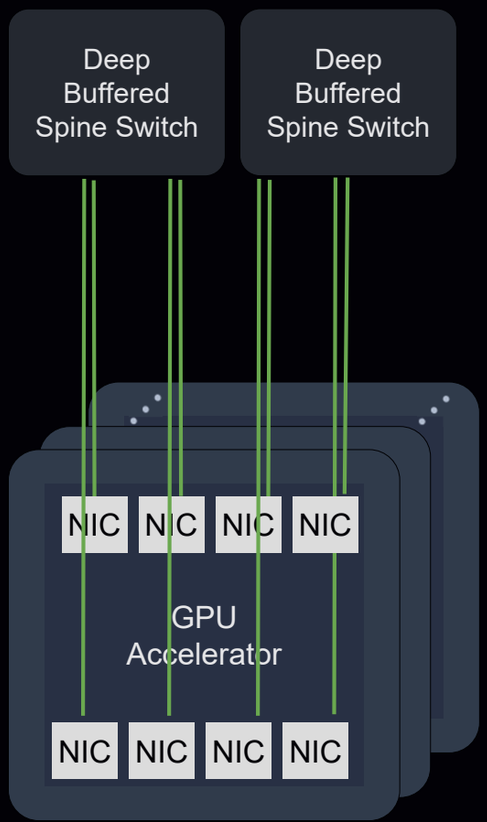


AI HARDWARE & EDGE AI SUMMIT

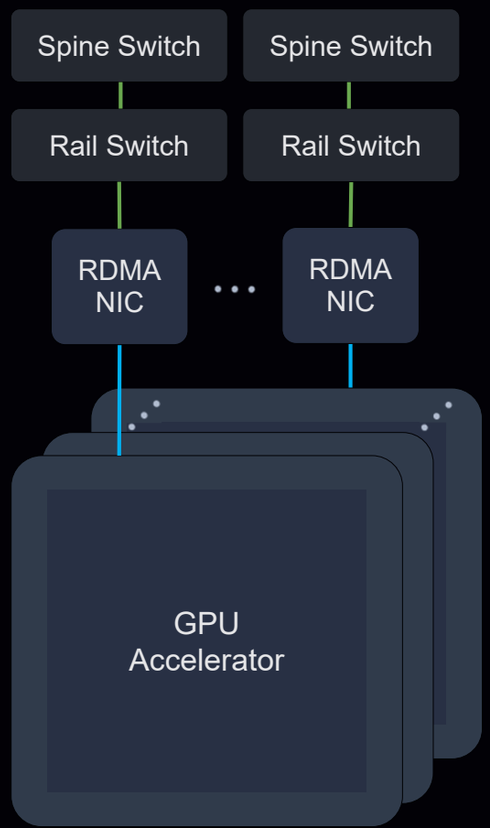
:: a chip level perspective

AI HARDWARE & EDGE AI SUMMIT

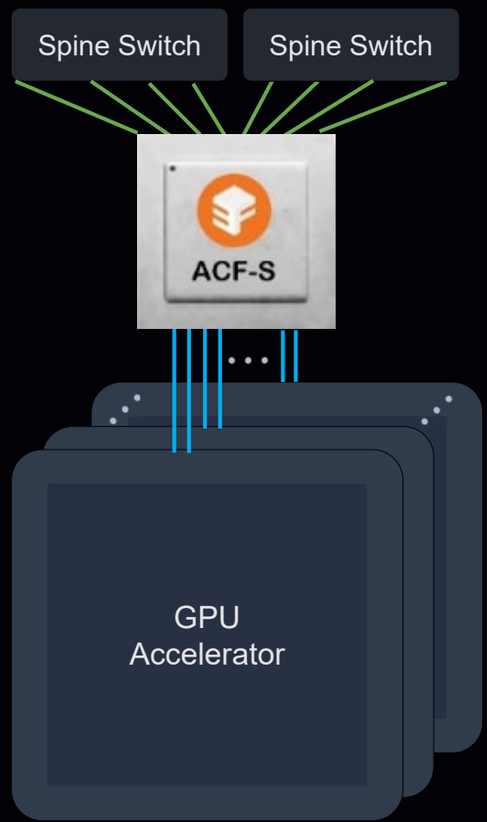
:: three approaches to accelerator networking silicon



Micro NIC

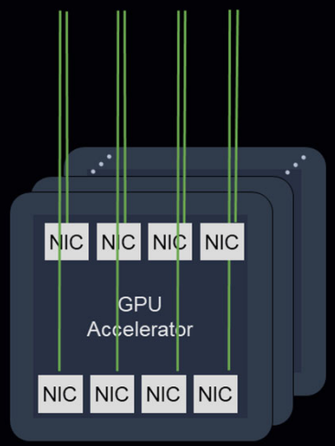


Traditional NIC



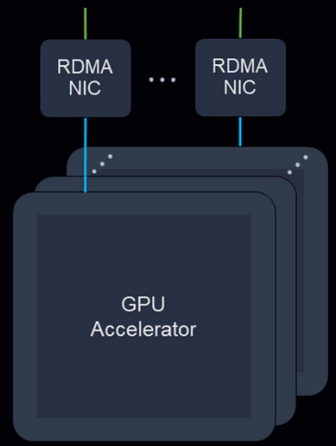
Mega NIC

:: three approaches to accelerator networking silicon



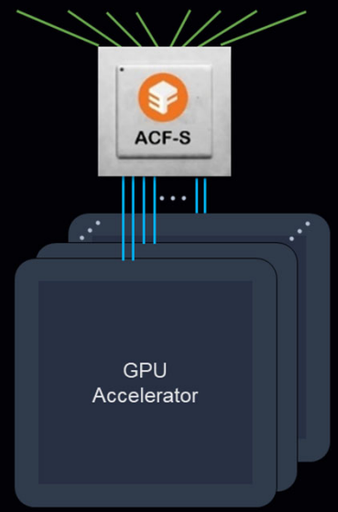
Micro NIC

- 400/800G single port per instance
- Switching & resiliency scheme is left to GPU to solve
- No decoupling of network fabric from compute
- Reduces FLOPS, memory per node
- Little to no adaptability of Transport / CC
- Significant chip/SW/system integration to productize



Traditional NIC

- 400/800G, single or few ports
- No internal switching, low network resiliency
- Commercial coupling of fabric w/ compute
- Maintains usable FLOPS, memory per node
- Low Transport/CC adaptability, limited configurability
- Current market incumbent solution



Mega NIC

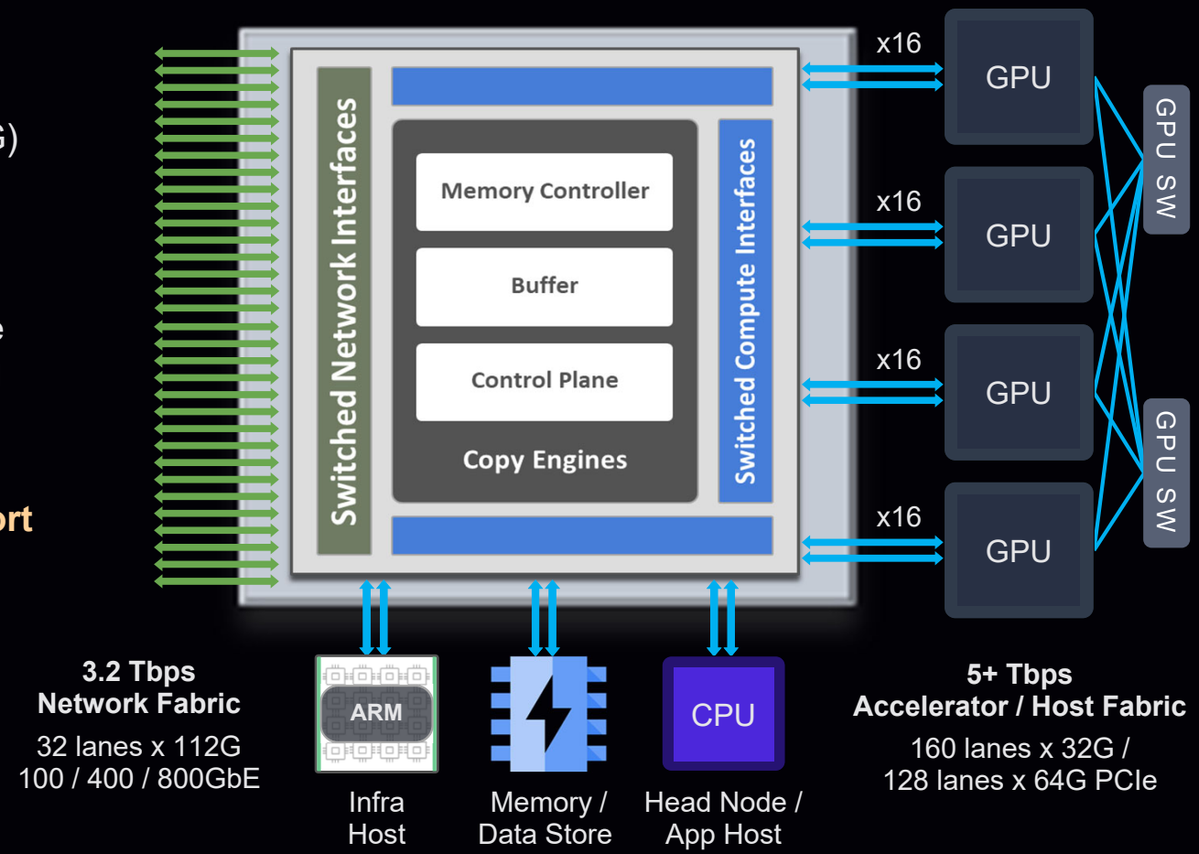
- Multi Terabit, high port count 400/800G shared across GPUs
- Fully switched across all ports w/ high network resiliency
- Enables any choice of compute
- Maximizes usable FLOPS, memory per node
- Highly adaptable transport/CC, SW-defined
- HW/SW interface drop-in to incumbent solution

:: world's first megaNIC chip // 8-Tbps acf-s "millennium"

Resilient network:
high port radix over fat (800G)
or skinny (100G) links

Full Router:
consolidates NIC-TOR-PCIe
fabrics with precise steering
to/from queue pairs

User programmable transport
on scalable infra host cores
at aggregate line rate PPS

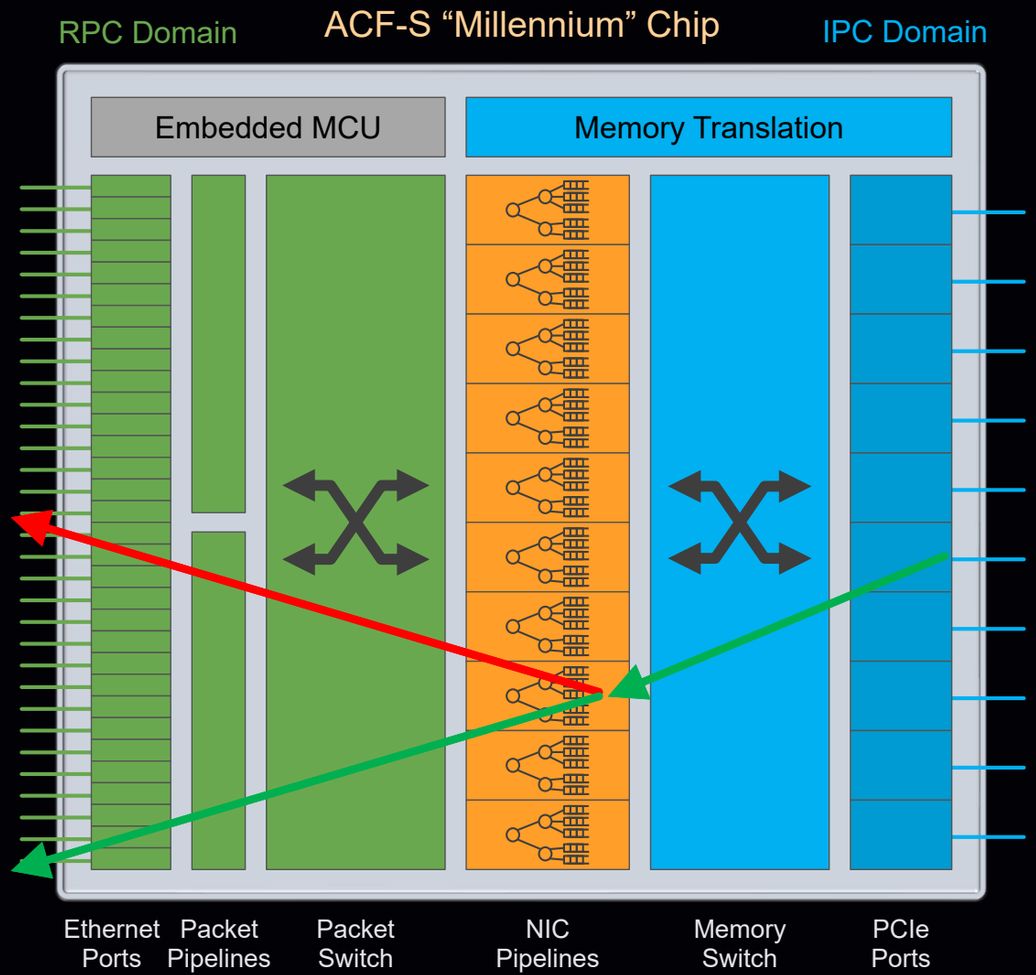


Delivers elastic, peak aggregate
3.2 Tbps bandwidth to accelerator

Multi-planar internal switch fabric
absorbs GPU incast by design

Composable IPC domain:
40K copy engines / data movers

:: redefining NIC chip functionality for AI clusters



Fully switched, bandwidth-rich, path-rich IO on both sides

- 192 serdes, 3.2Tbps Ethernet, 5Tbps+ PCIe
- Reduces overall system serdes and interface power
- Flatter, ultra-resilient network design

Aggregated, shared, rail-able system NIC across GPUs

- Slices, steers traffic anywhere to precisely land headers, payloads
- High performance on-chip memory translation
- More efficient, scalable software stack & use of chip resources

User SW-programmable transport & congestion control loop

- On top of linerate per-packet actions and flow control in HW

GPU queue-aware traffic engineering on lossy network

- Resilient Message Multipathing for path distribution
- Large elastic buffer & early signaling protects against GPU incast & slow receiver

:: acf-s solutions to AI scaling challenges

AI HARDWARE & EDGE AI SUMMIT

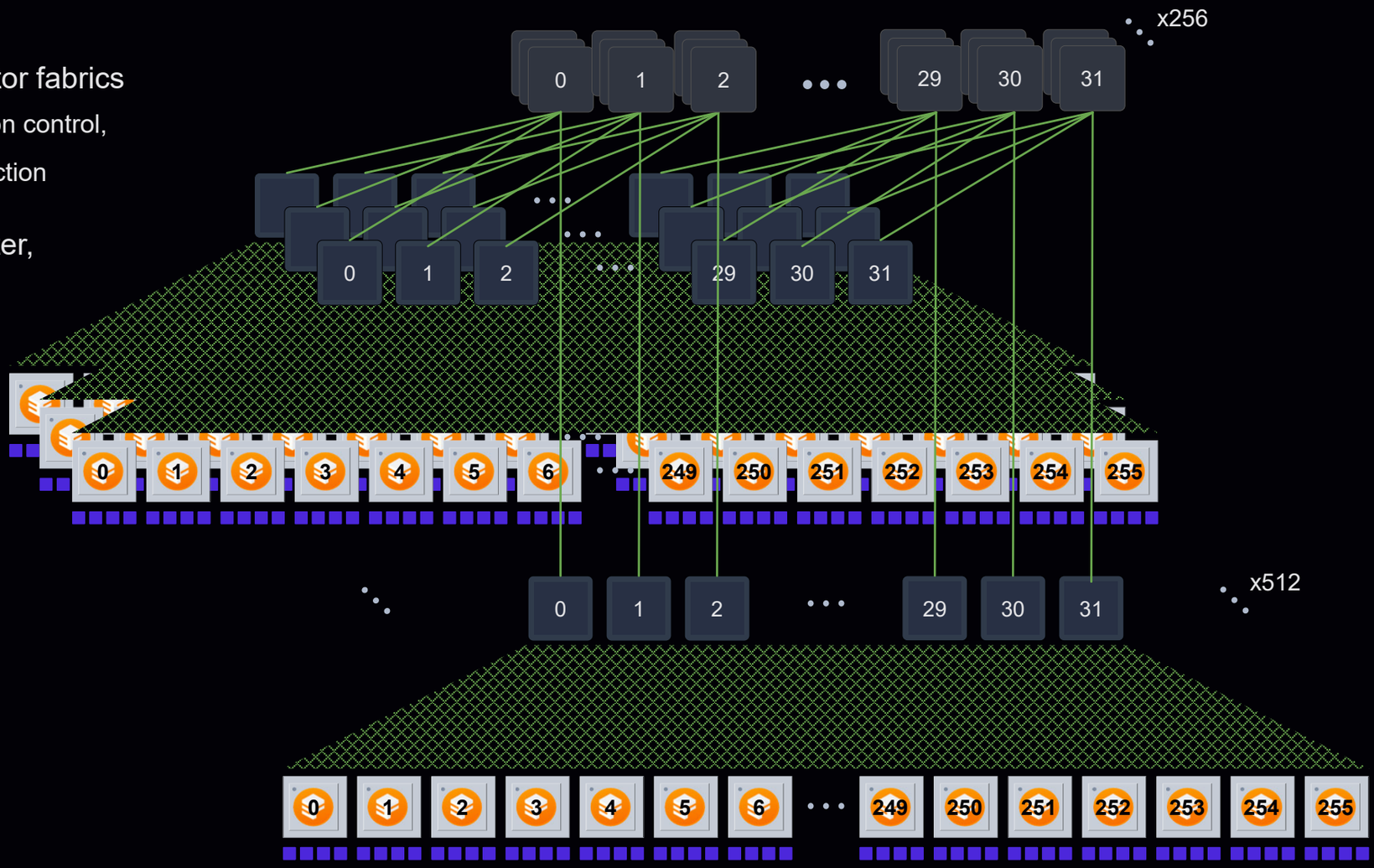
:: 524,288 accelerators in 2 layers of switching

ACF-S enables wider, low-hop accelerator fabrics

With software-defined transport & congestion control,

programmable data distribution & path selection

Flatter networks reduce cost, latency, jitter, congestion points

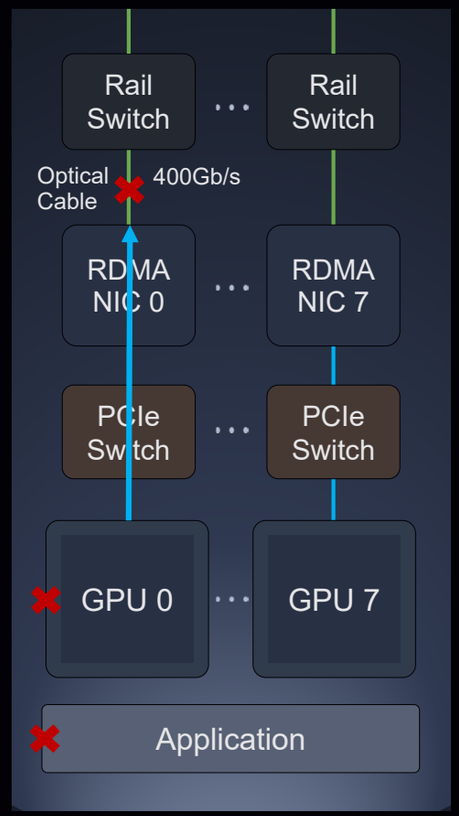


Per rail: 256x 51.2T switches
Total: 16k 51.2T switches

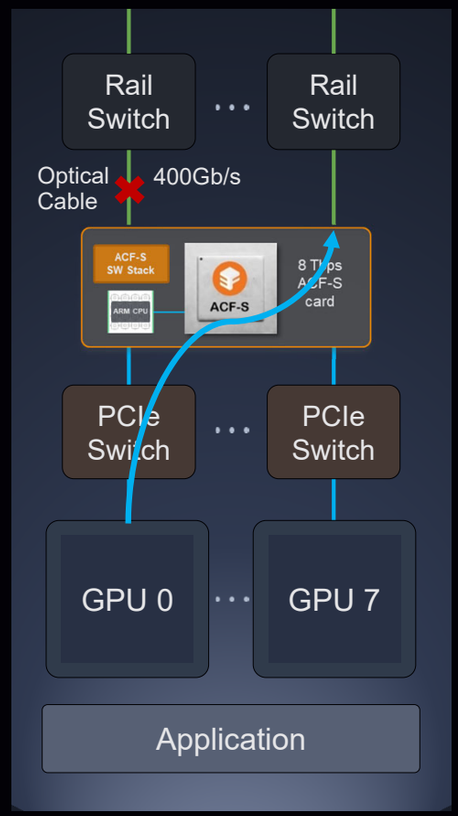
512x instances of the cluster building block

128k x ACF SuperNICs
512k accelerators

:: acf-s radix provides multipath resilience



Traditional GPU-NIC rail



ACF-S logical rail

Calculated time to first job failure

GPU cluster size	Standard GPU NIC-rail	ACF-S logical rail
16,384	160 min	6.48E+57 hours
32,768	80 min	3.24E+57 hours
65,536	40 min	1.62E+57 hours
114,688	**23 min**	**9.26E+56 hours**
524,288	5 min	2.02E+56 hours

* Assume optical cable MTBF of **5 years**
Does not assume vendor-specific CCL level backdoors

Every GPU sees every ACF-S uplink

Every ACF-S uplink is an independently routed and traffic engineered interface

Provides resilience in HW to cable/port failures with only incremental bandwidth degradation

ACF-S stack rebalances paths with workload awareness

:: acf-s solves for AI system efficiency metrics that matter

Typical LLM MFU (Model FLOPs Utilization) ~ 30-50%

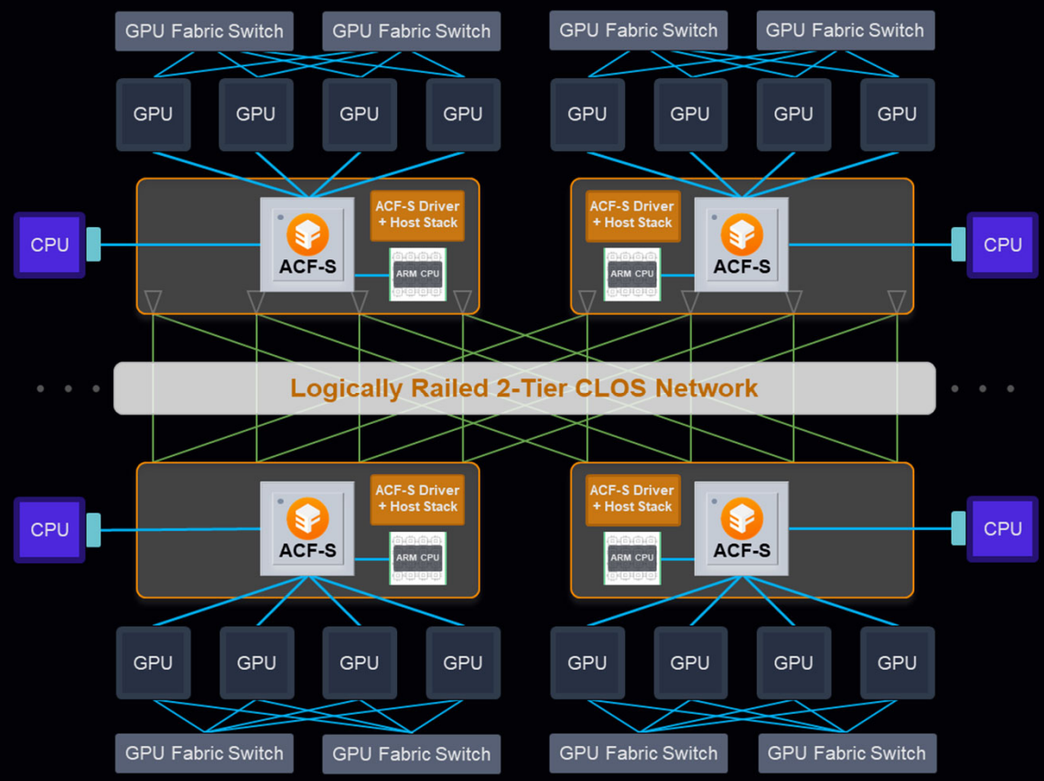
Hardware FLOPs Utilization (HFU) ~15% higher than LLM MFU

- On clusters of H100 systems* ranging from 128 – 512 GPUs
- Includes data store (checkpoint), network & recomputation overheads

What do we think the HFU efficiency will be at 100K+ GPUs?

AI Networking Silicon Evolution must:

- Maximize FLOPs and HBM in accelerator real estate
- Have lightweight, low latency, low overhead scale-up (IPC)
- Have “unlimited”, resilient, flat network scale-out (RPC)
- Use wide, shared, efficient network silicon to distribute traffic (megaNIC)
- Use SW-defined fabric transport, path selection, congestion control to adapt to workloads and avoid hardwiring infrastructure to current models



Reference :github.com/mosaicml/llm-foundryblob/main/scripts/train/benchmarking

:: Thank You

AI HARDWARE & EDGE AI SUMMIT