enfabrica

UNLEASH THE REVOLUTION
IN NEXT-GEN COMPUTING

ACF-S: An 8-Terabit / sec SuperNIC for High-Performance Data Movement in AI & Accelerated Compute Networks

Hot Chips 2024

August 27, 2024

Shrijeet Mukherjee, Enfabrica
Thomas Norrie*, OpenAI

** work previously done at Enfabrica*

:: mission

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

:: team

started 2020

120+ engineers

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

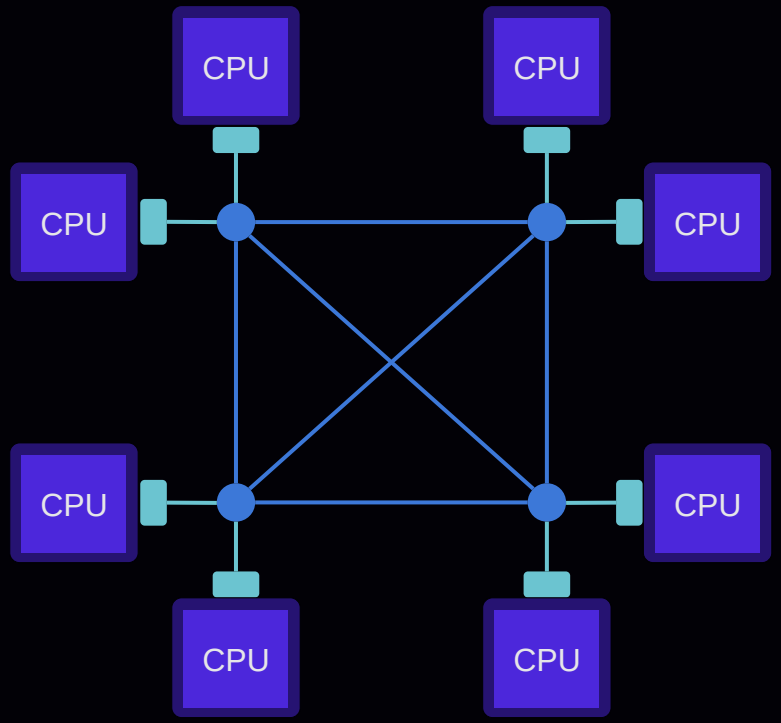
:: product

accelerated compute fabric superNIC (ACF-S)
1st chip codename ‘millennium’ @ 8Tbps bandwidth

enfabrica

HOT CHIPS

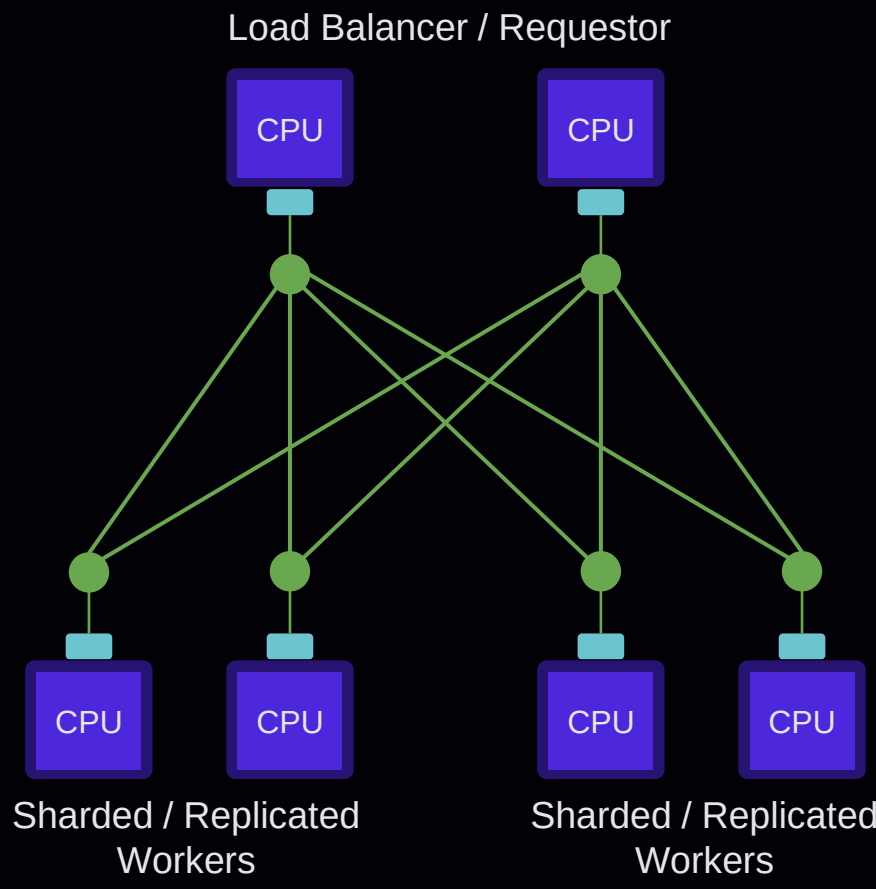
:: scale-up supercomputing // mainframe, ccNUMA



All blue links are IPC communication

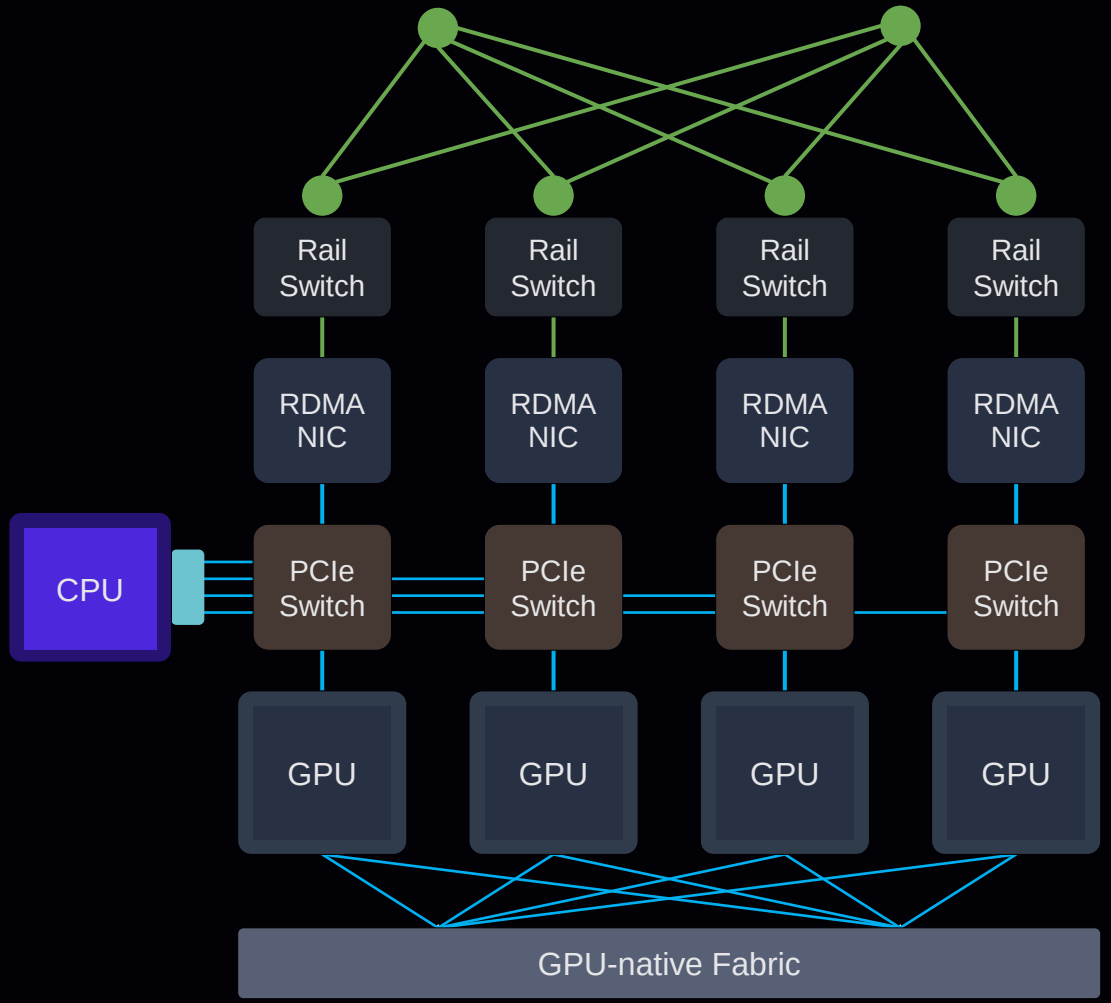
- Fully coherent memory system operating on a “large” problem by sharding computation
- CPUs synchronize state and move memory closer using IPC transactions with latencies in nanoseconds
- Communication protocols deeply embedded in the processor to enable “transparent” communication

:: borg // the rise of scale-out computing



- Client-server design, built for extreme, resilient application scaling
- All communication uses retargetable, resilient software managed RPCs
- Workers and data pipelines are imminently reconfigurable
- Heterogenous elements with high aggregate bandwidth needs and high tolerance to latency (microseconds to milliseconds)

:: hyperscale AI / ML systems // super, meet borg



- A modern, truly scalable solution demands:
  - Tight performance of a supercomputer

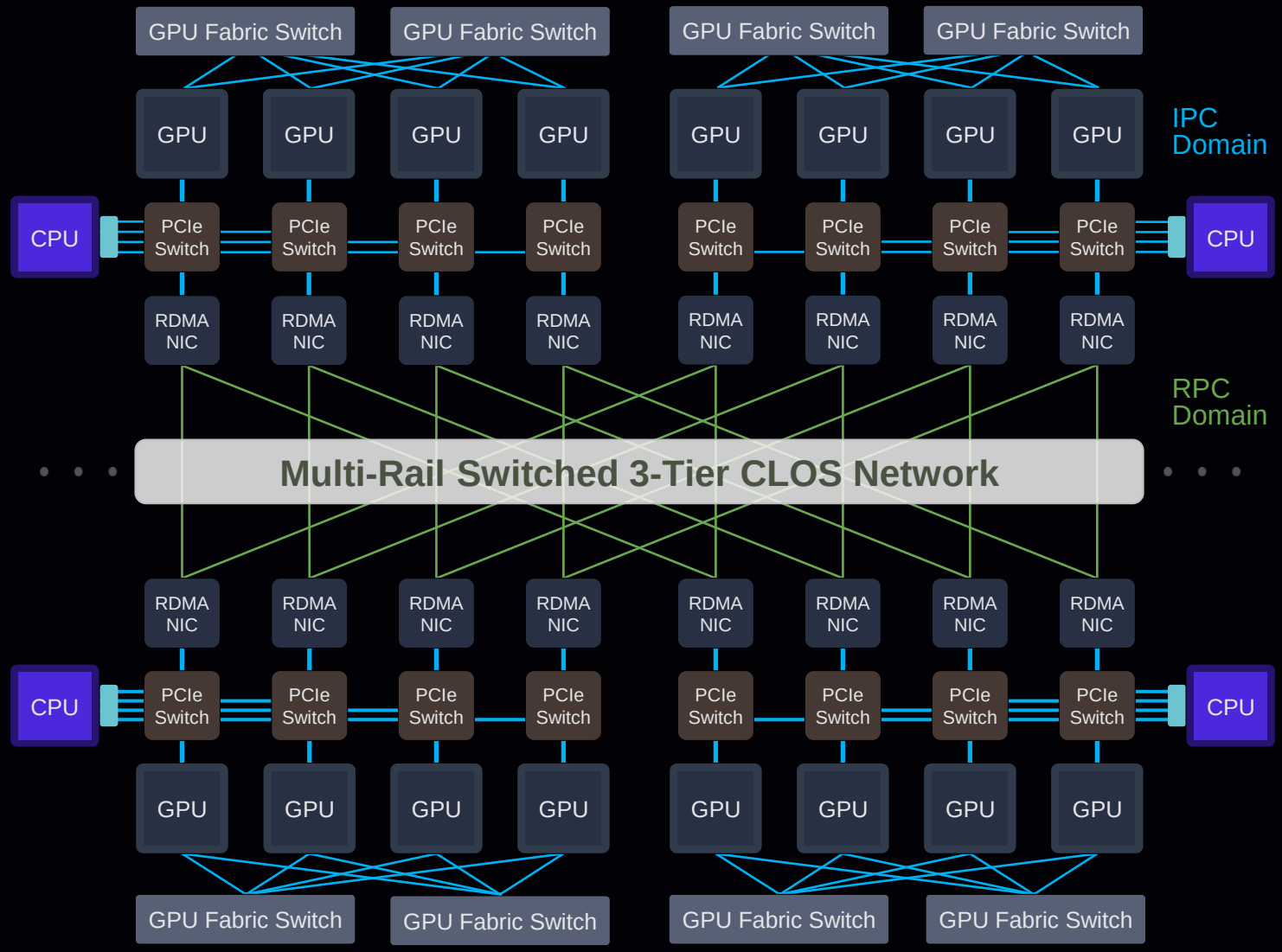
    plus
  - Elastic scaling of compute threads and resiliency of a cloud scale system
- Dual mode communication models:

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

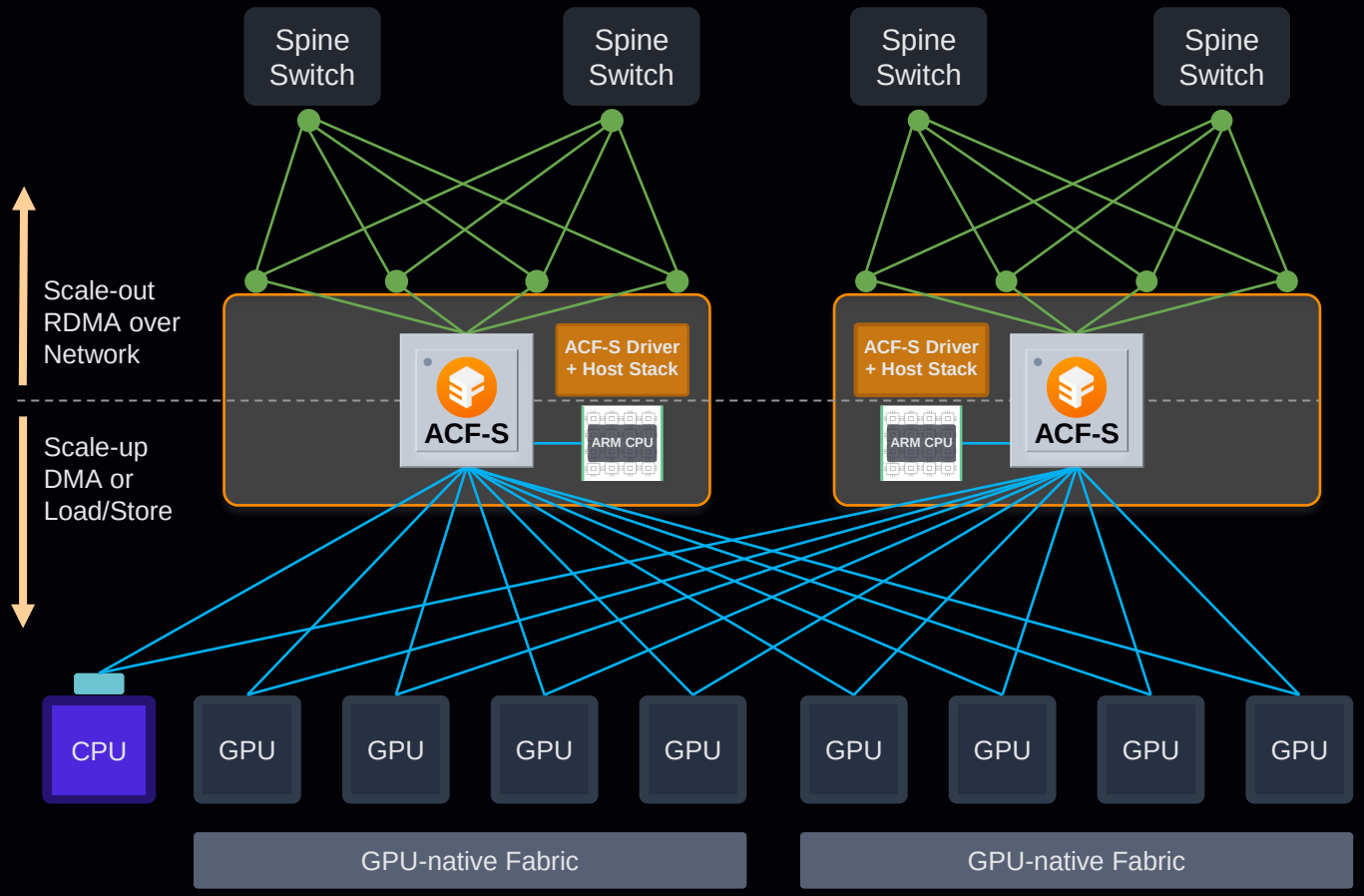
  Scale-out :: High level, structured kernels for computation and communication (CCLs)

:: hyperscale AI system attributes & challenges

- Expanding IPC, RPC domains
  - Scale-up defines the largest tensor / vector op
  - Scale-out defines the largest bulk data movement
  - Disparate fabrics scale disparately
- Imbalanced burst bandwidth
- Hardwired communication types, sizes
  - IPC to infinity is infeasible
  - Tensor parallel ops entirely over RPC is inefficient
- Fabric scaling is reducing MFU
  - More network tiers = more latency, jitter, contention
  - Network incast, failures strand or stall compute



:: refactoring the communication endpoint architecture



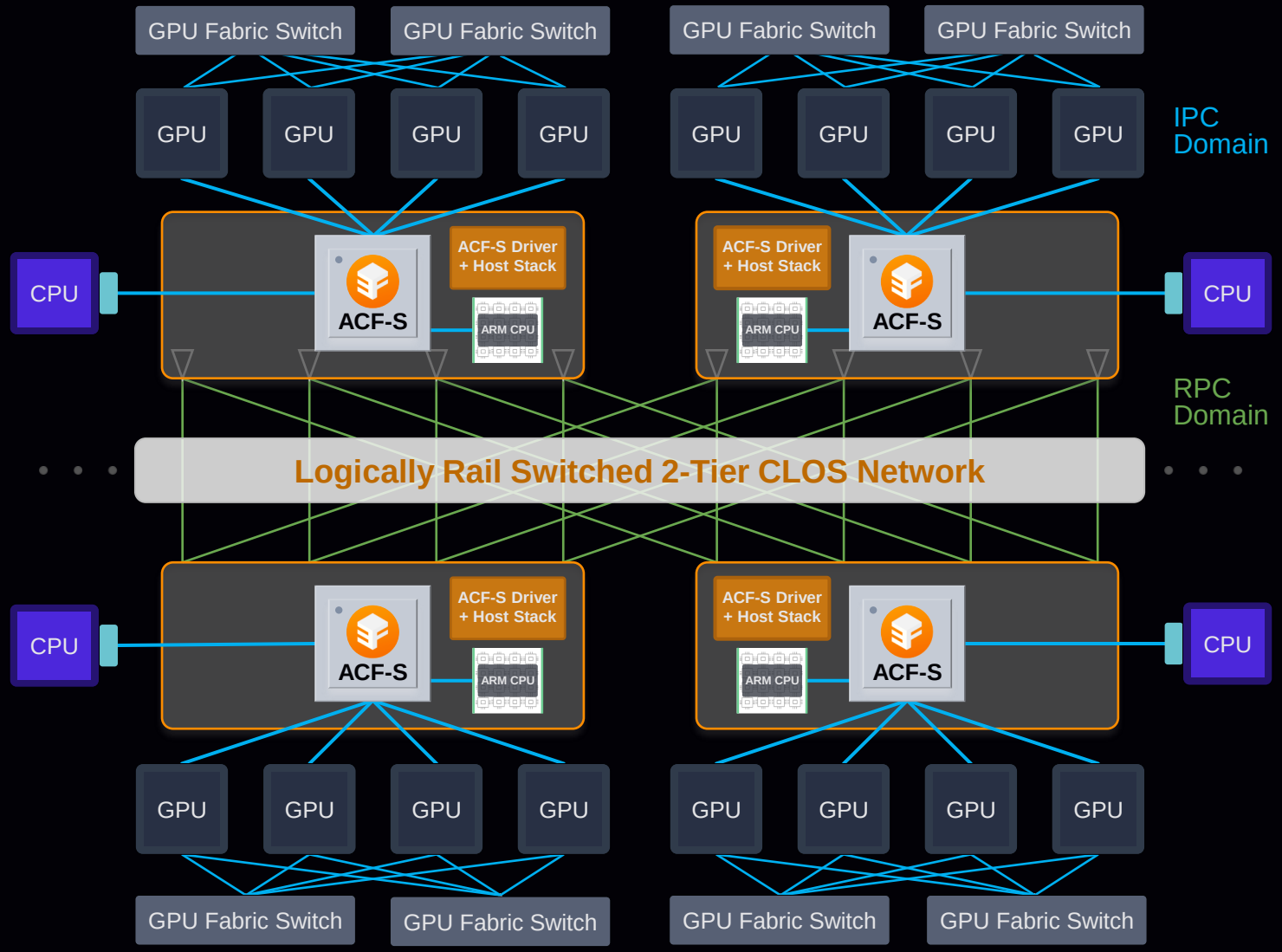
- High-bandwidth, fat-tree ‘superNIC’ that
  - Has scale up interfaces AND scale out interfaces
  - Has interfaces to speak memory

    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 headers / payloads and vice versa
  - Elastically binds memory interfaces to network interfaces and vice versa

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

:: unify scale-up / scale-out fabric communication

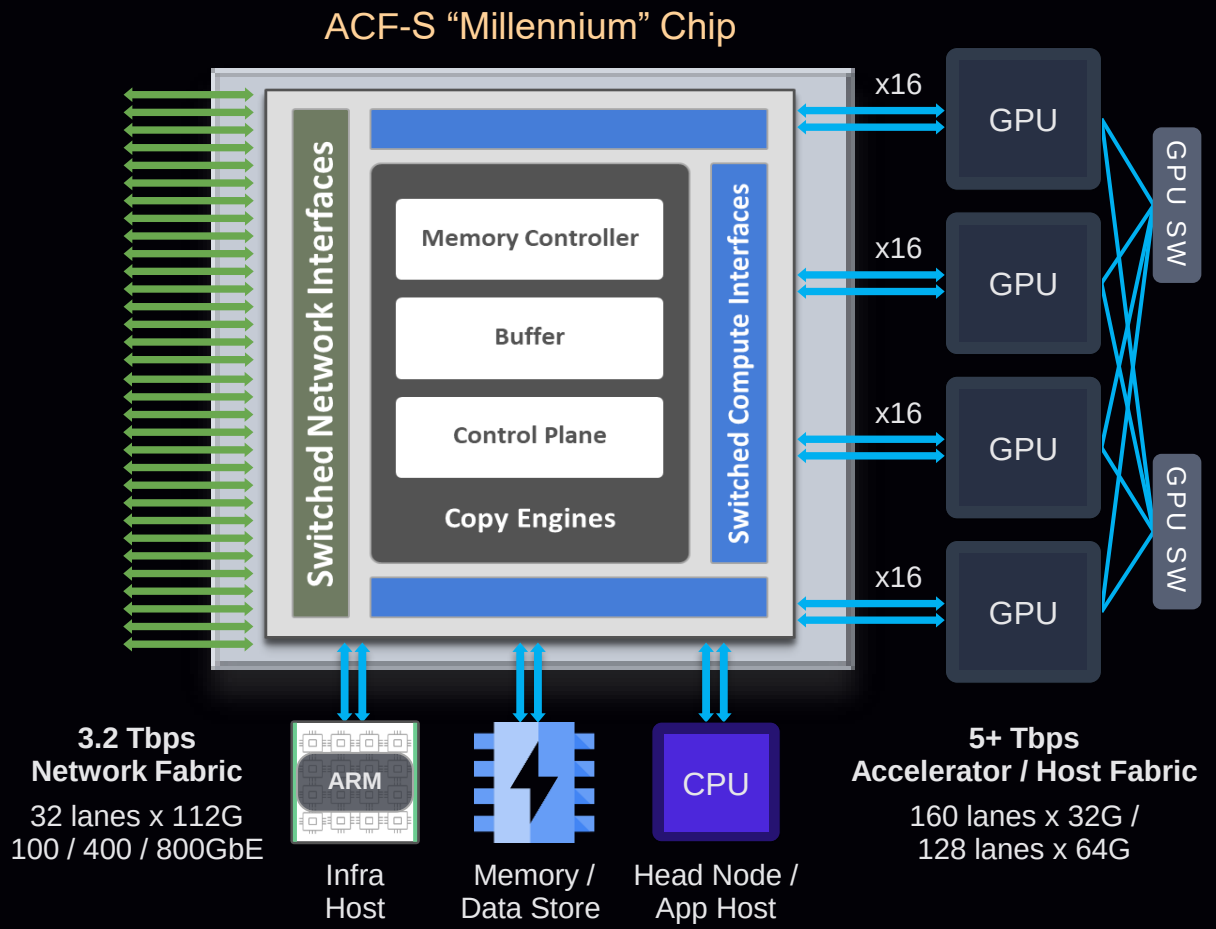


:: acf-s “millennium” // 8-Tbps accelerator network scaling element

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

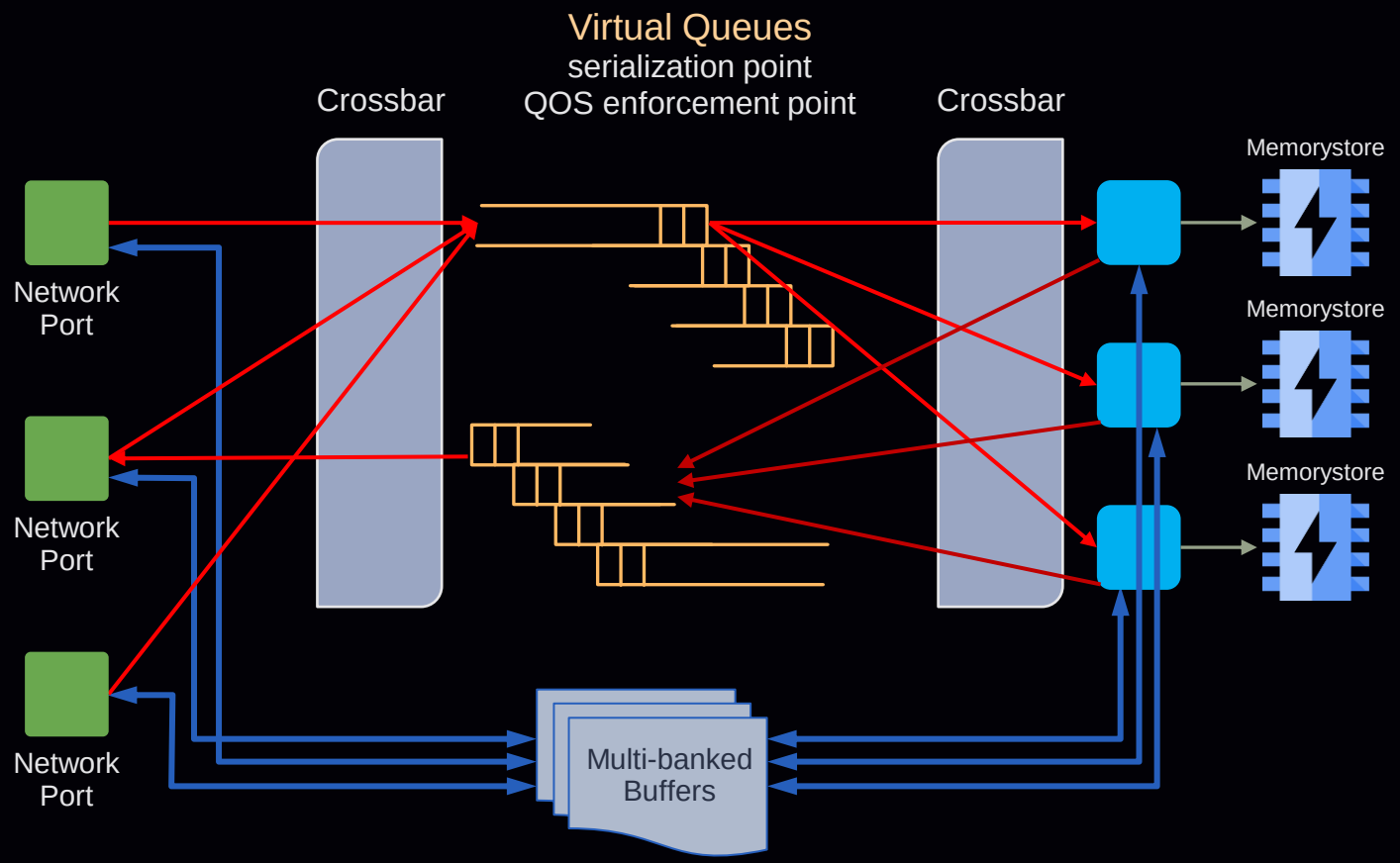


Delivers elastic, peak aggregate
3.2 Tbps bandwidth to accelerator

Multi-planar internal switch fabric
absorbs GPU incast by design

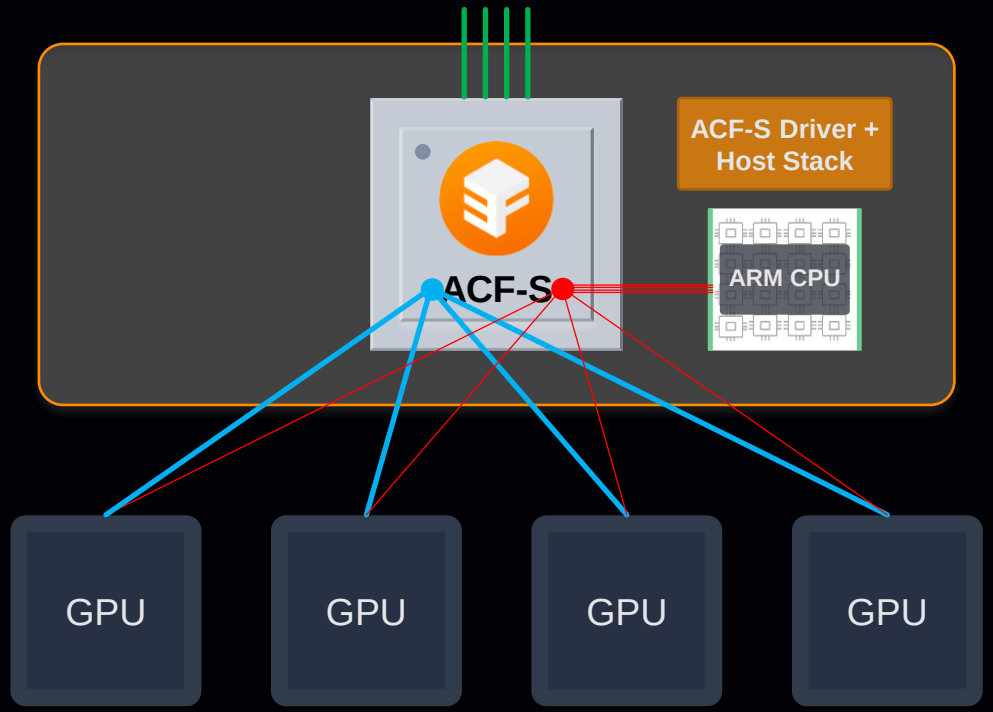
Composable IPC domain:
40K copy engines / data movers

:: data movement in acf-s hardware



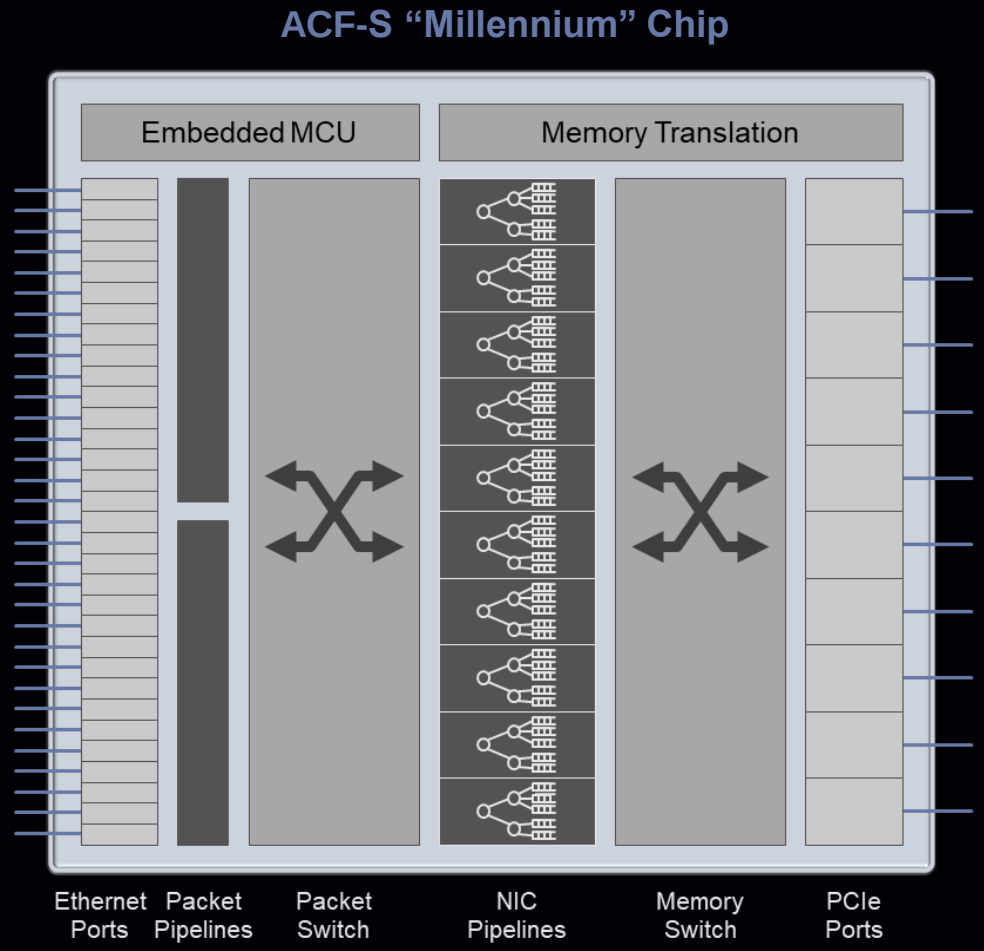
- Powerful “unlimited” data movers which gang together PCIe or Ethernet interfaces
- Enables usage of dense, skinny network links to improve latency and resiliency
- Virtual Queues in data movers enable matching data bandwidth to any multiple of interface bandwidth (e.g., avoid incast congestion)
- Virtual queues enable automatic detection of slow receiver (Rx direction) or network scheduling (Tx direction)

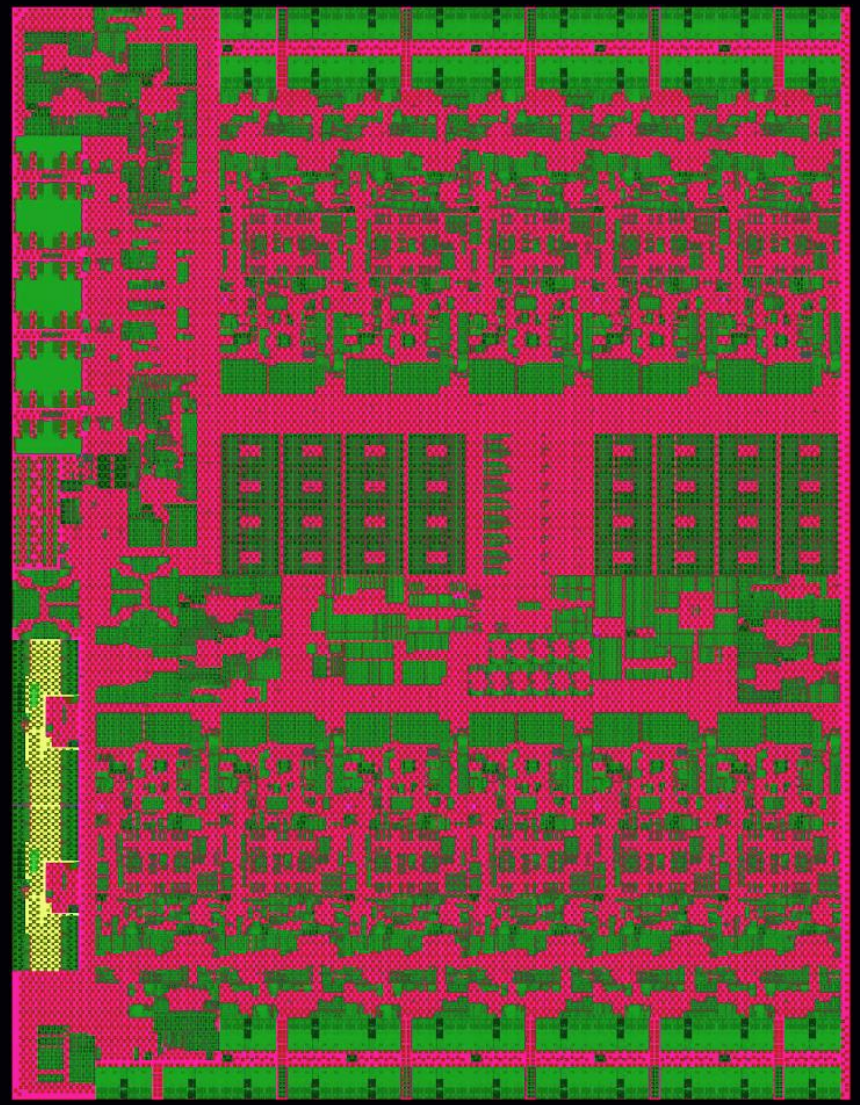
:: programmable transport in acf-s software



- Runs on attached “infra host” CPU cores
  - Optimizes over full surface area of PCIe and Ethernet using descriptor fields that make precise port decisions
  - PPS performance can be scaled
  - Any port can be a transport processor
- Millennium hardware supports
  - Per packet pacing, steering, and congestion reaction
  - Per packet global address translation
  - Ultra-high efficiency send and receive coalescing
  - Fast, precise flow control
- Infra host SW handles low-frequency-per-message ops
  - Protocol state machine, routing policy, send rate and higher-level congestion avoidance

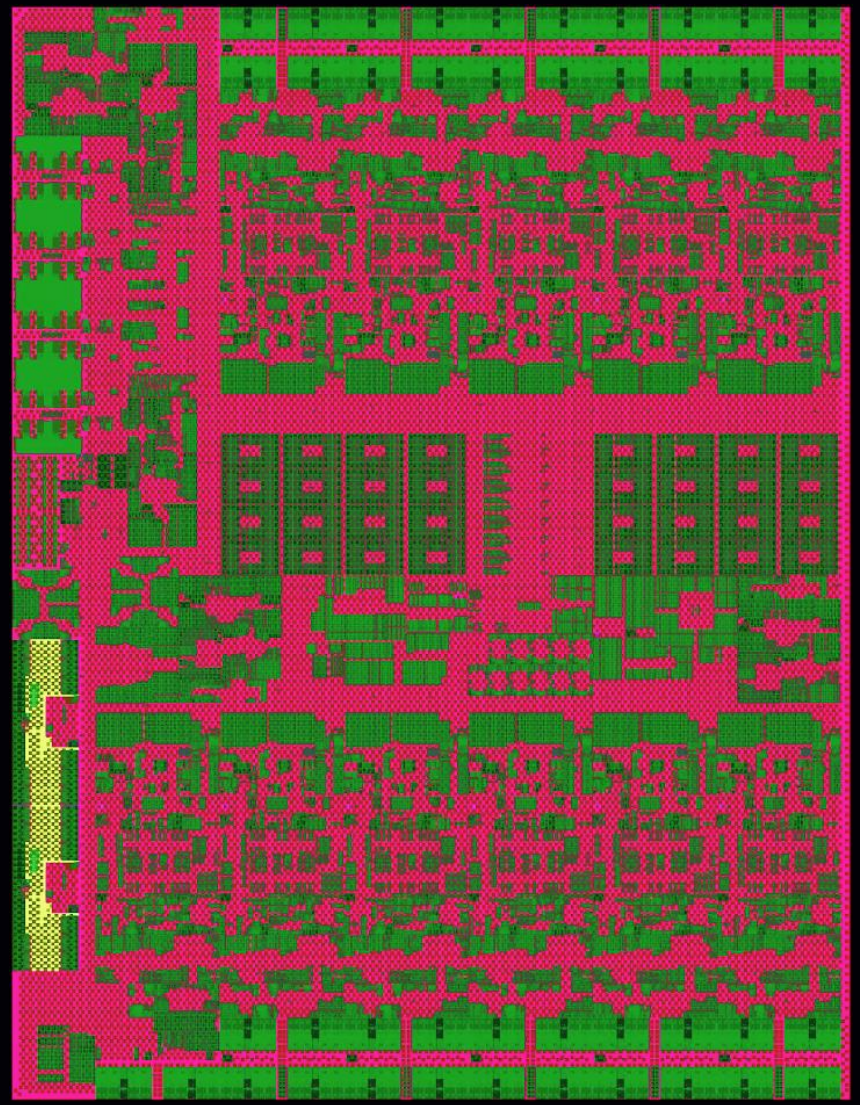
:: acf-s millennium // logical & physical view



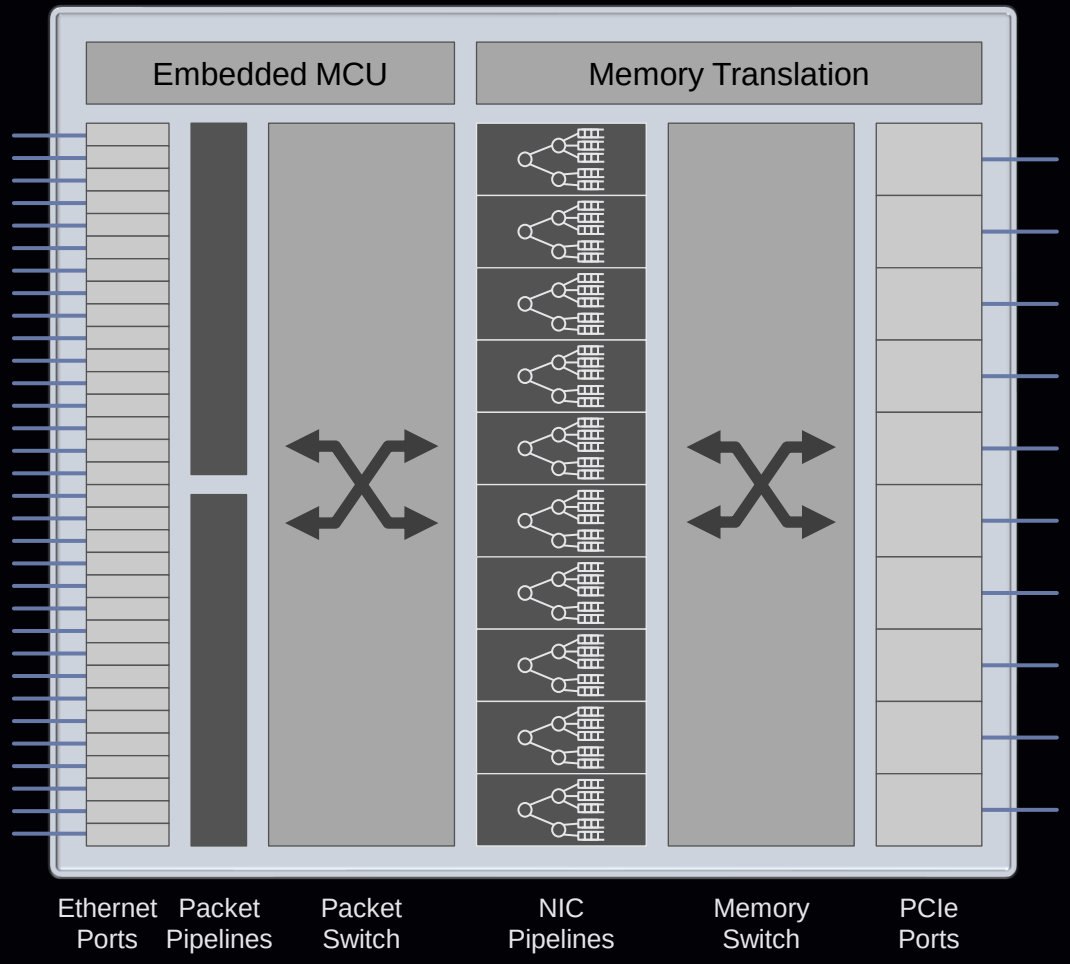


:: acf-s millennium silicon

Millennium ASIC by the Numbers	
Process	5nm, 15-layer metal
Transistor Count	47 Billion (30B non-SRAM)
On-Chip Memory	2446 Mbits
Package	67.5mm x 67.5mm HFCBGA 1mm pitch, 4288 ball count
Core Voltage	0.75V core
Power	250 W typical
Temp	0 – 70 C Ambient

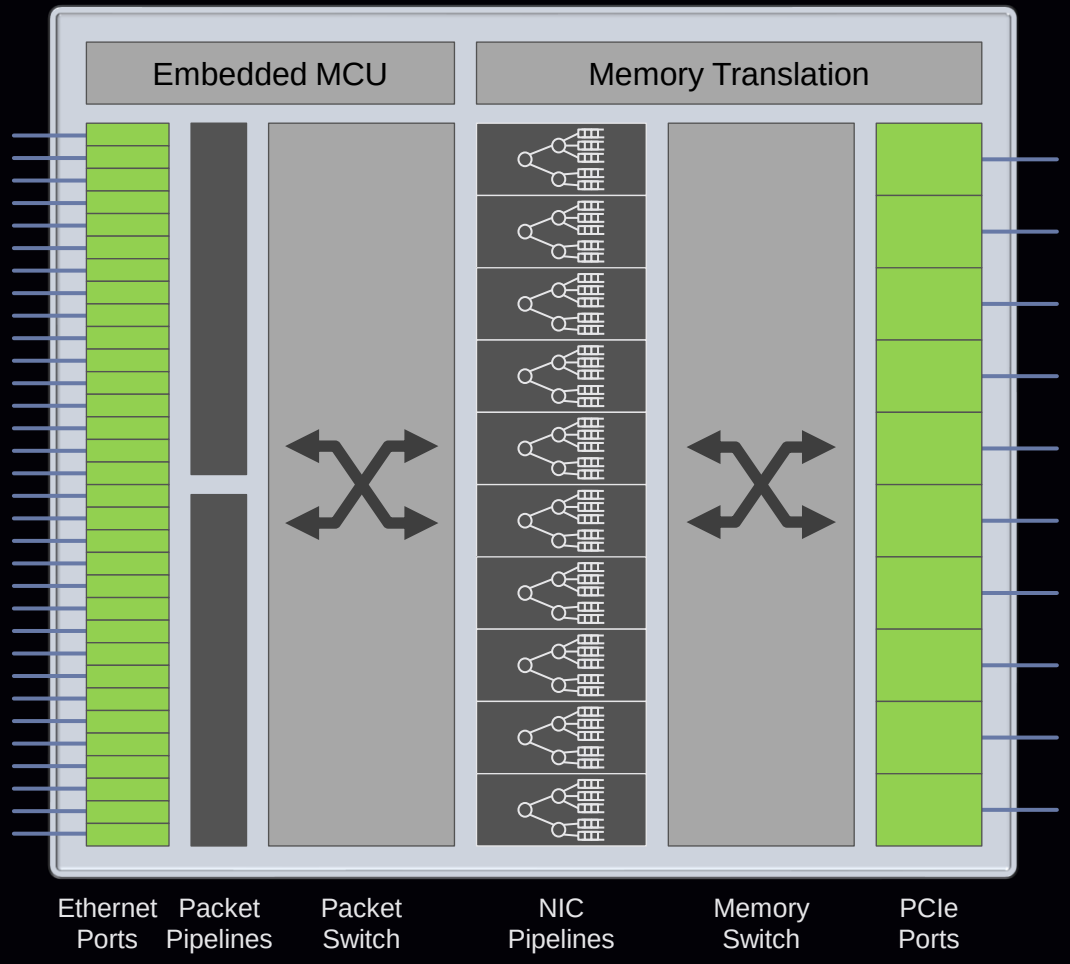


:: how millennium is built differently



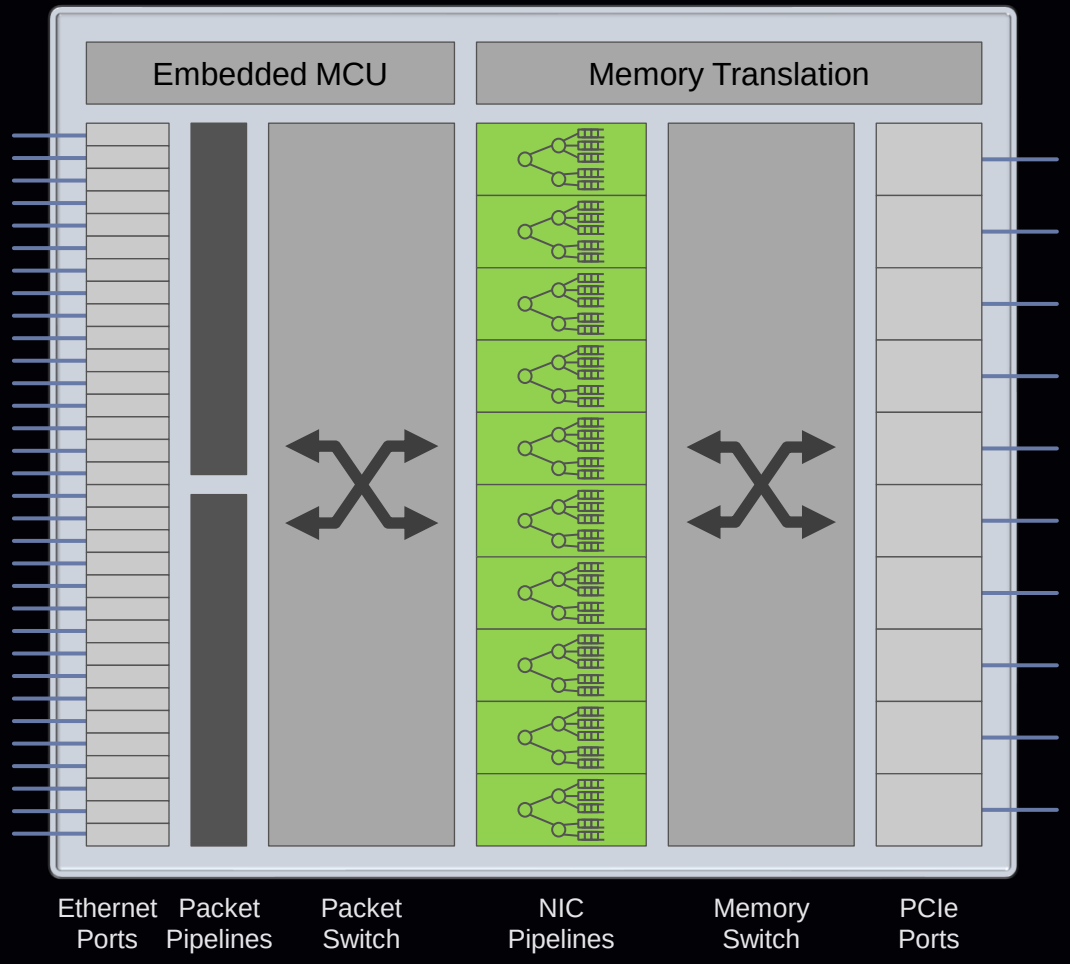
1. Higher chip IO density
2. NICs enmeshed in crossbars
3. Scalable memory translation
4. Shared flow buffer and packet processing

:: higher chip io density



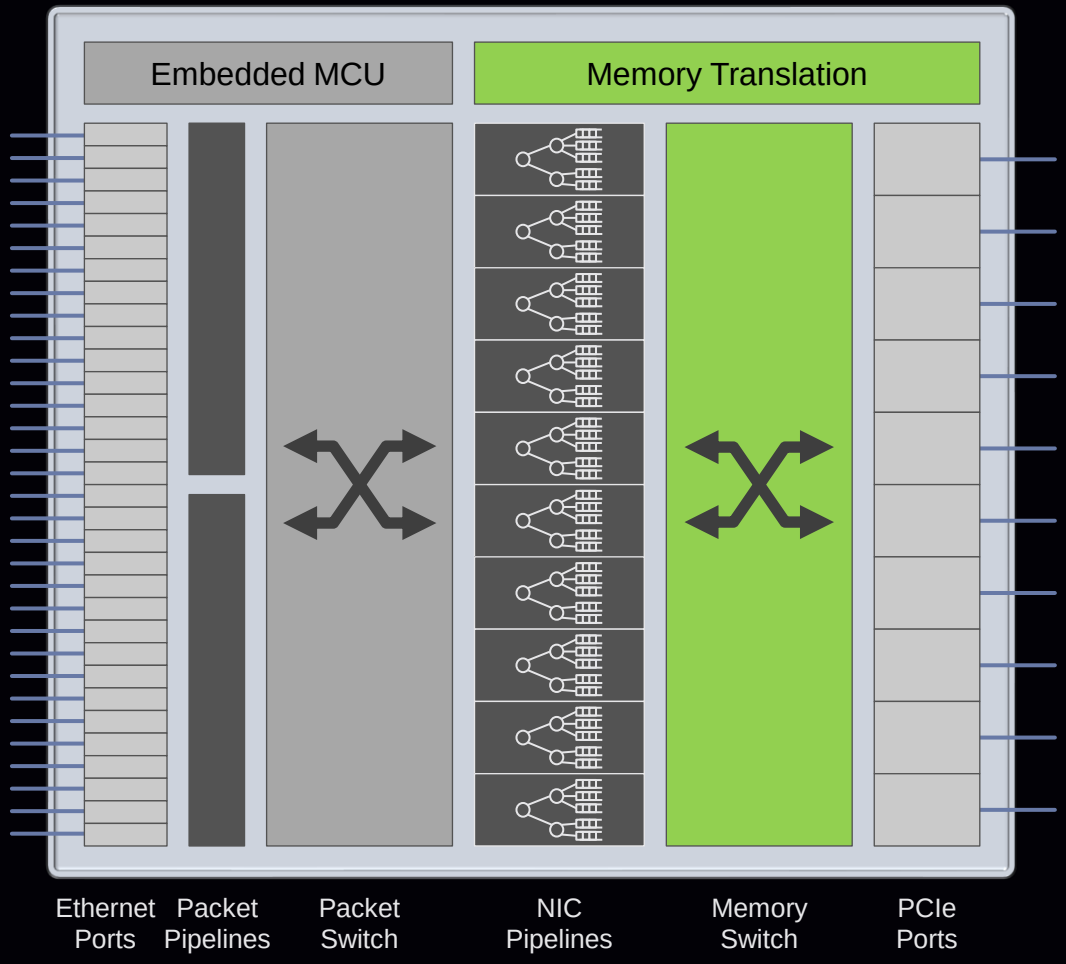
- Bandwidth-rich, path-rich, multi-accelerator superNIC
  - Better bandwidth balancing, congestion management, resilience
- On-die aggregation reduces serdes at system level, reducing power
- Ethernet network interfaces
  - 32 lanes x 112G PAM4
  - 4 x 800GbE, 8 x 400GbE, 16 x 200GbE, or 32 x 100GbE
- Accelerator / host interfaces
  - 160 lanes x 32G NRZ
  - 10x PCIe 5.0 x16 links with CXL 2.0 support
  - 8x PCIe 6.0 x16 links (revision)

:: nics enmeshed in crossbars



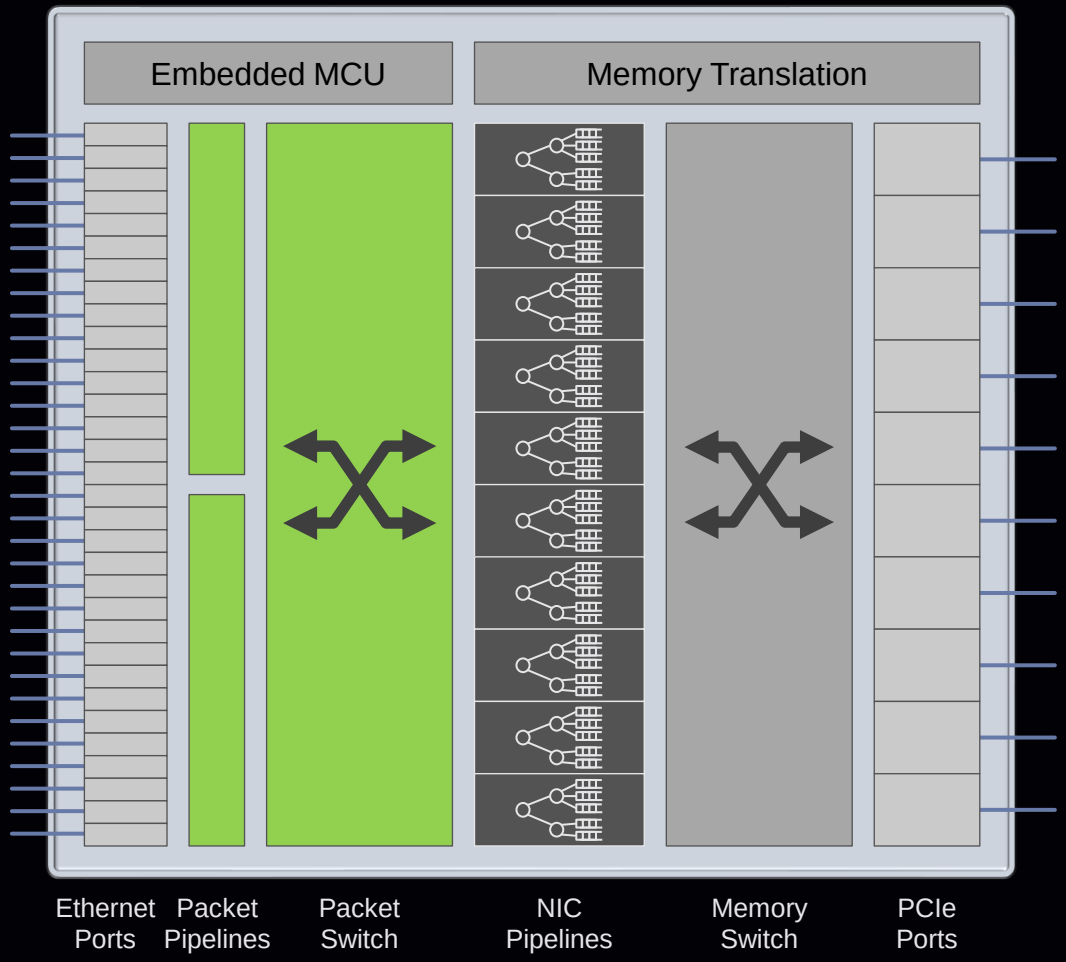
- NIC pipelines are surrounded by switches to both IO planes
- Millennium refactors the NIC from a feedforward protocol converter to multi-planar fabric switching element
- Provides any-byte-to-any-byte movement across all ports
- Flexible packet slicing combined with precise placement into shaped memory buffers with scatter-gather lists
- Millennium is a true 3.2T superNIC to the transport host which can now assign bandwidth between ports to optimize against workload and topology

:: scalable memory translation



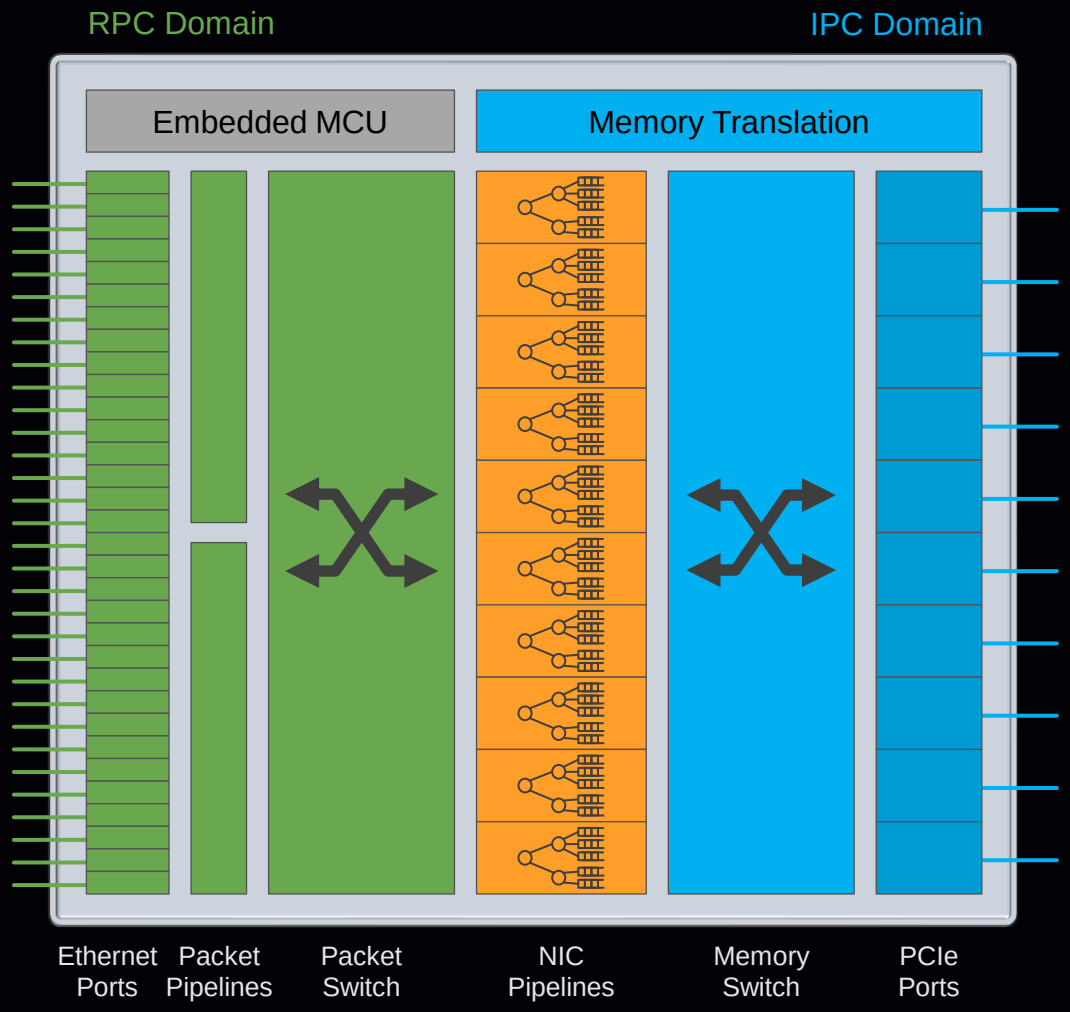
- Problem
  - No global memory addressing since endpoints can have distinct memory spaces
  - Host IOMMU is a bottleneck for this wide high-bandwidth accelerator system
- Solution
  - Built-in high performance memory translation engine with flatter, cached structures
  - Provides access policies and enforcement
  - Enables an external “invisible” transport host to manage memory movement across all endpoints

:: shared flow buffer and packet processing



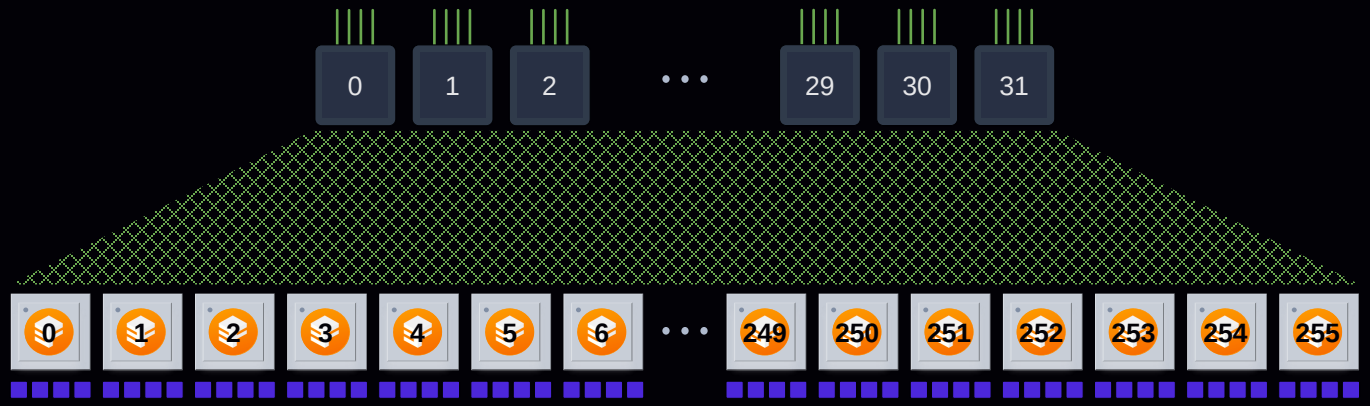
- Large shared packet buffer provides
  - Burst absorption to avoid packet drops and improve performance stability
  - More time for congestion detection and early signaling
- Shared packet pipelines provide
  - Very high packet processing rate (2 Billion PPS)
  - Efficient sharing of programmable tables
- Aggregation gives wider visibility for better control
  - Early congestion detection for fast flow stall
  - NIC egress pipelines can look ahead to switch egress Q depths

:: bringing it all together

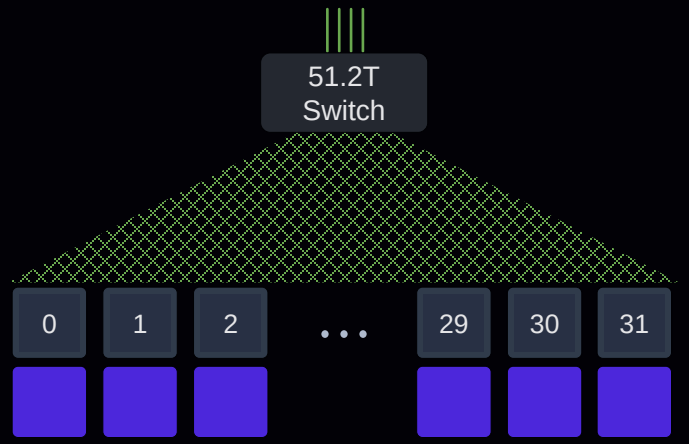


- Lots of IO !
  - Reduces system serdes and power
  - Enables resilient network design
- Array of NIC pipelines enmeshed by high-bandwidth packet and memory switching planes
  - 3.2Tbps system NIC that can slice traffic anywhere
  - Flatter networks
- Clean global addressing model
  - Enabling user-programmable transport host
- Aggregates resources across all NICs
  - Less replicated state / resources, higher efficiency and scalability
  - Early signaling and better congestion properties

:: enables wide, low-hop accelerator fabrics



1024 GPUs in a single, fully bisectional switching layer using ACF-S



Only 32 GPUs using standard NIC-TOR design

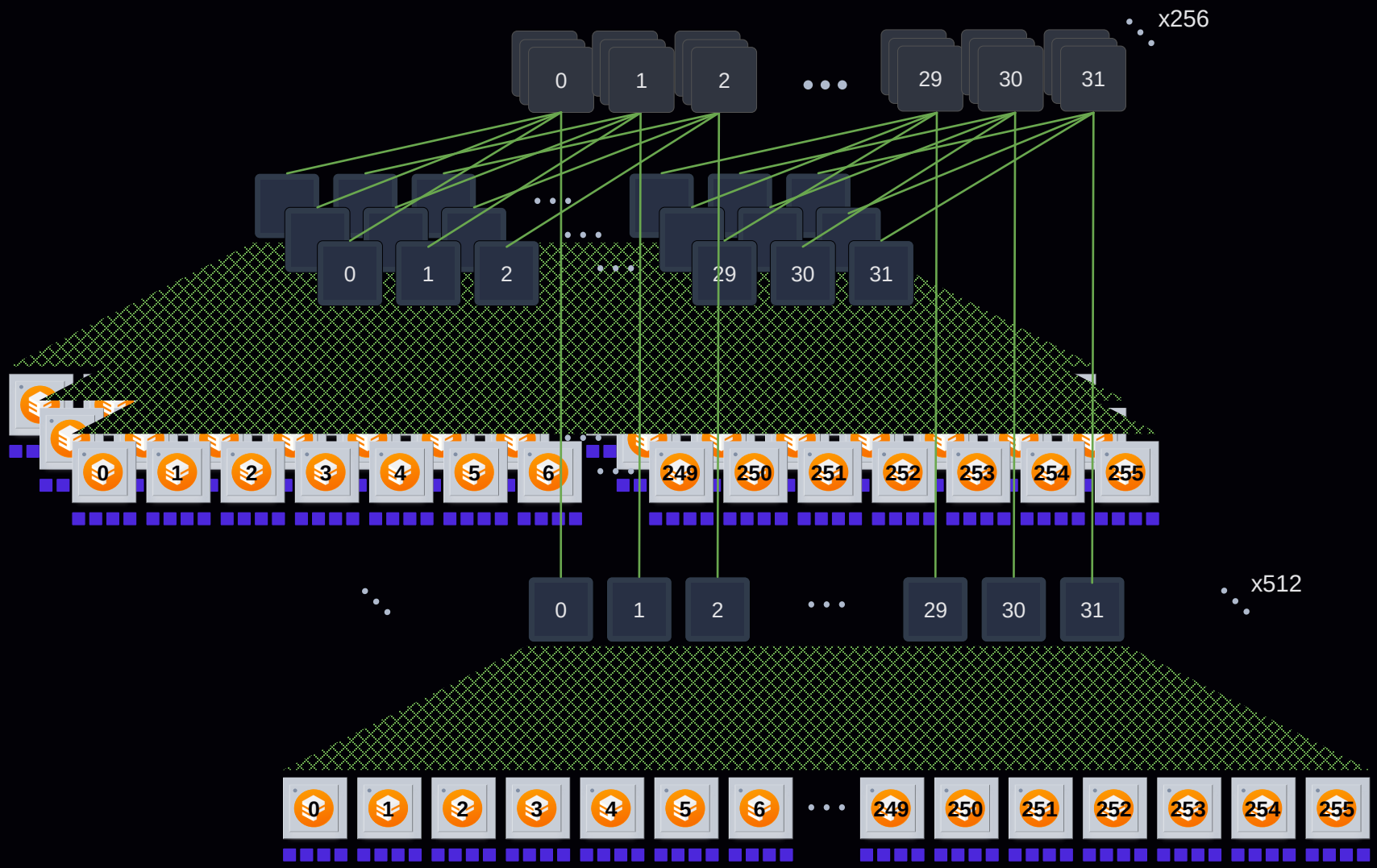
- ACF-S enables 32x network reach for flatter networks
- First 4x reach enabled by:
  - 3.2T wide surface (vs 800G aggregate)
  - Shared across GPUs due to full switch
  - Software-defined path selection
- Next 8x reach enabled by:
  - 100G uplinks vs 800G uplinks
  - Programmable data distribution and congestion management with software-defined transport
- Flatter networks reduce cost, latency, congestion points

:: 524,288 accelerators in 2 layers of switching

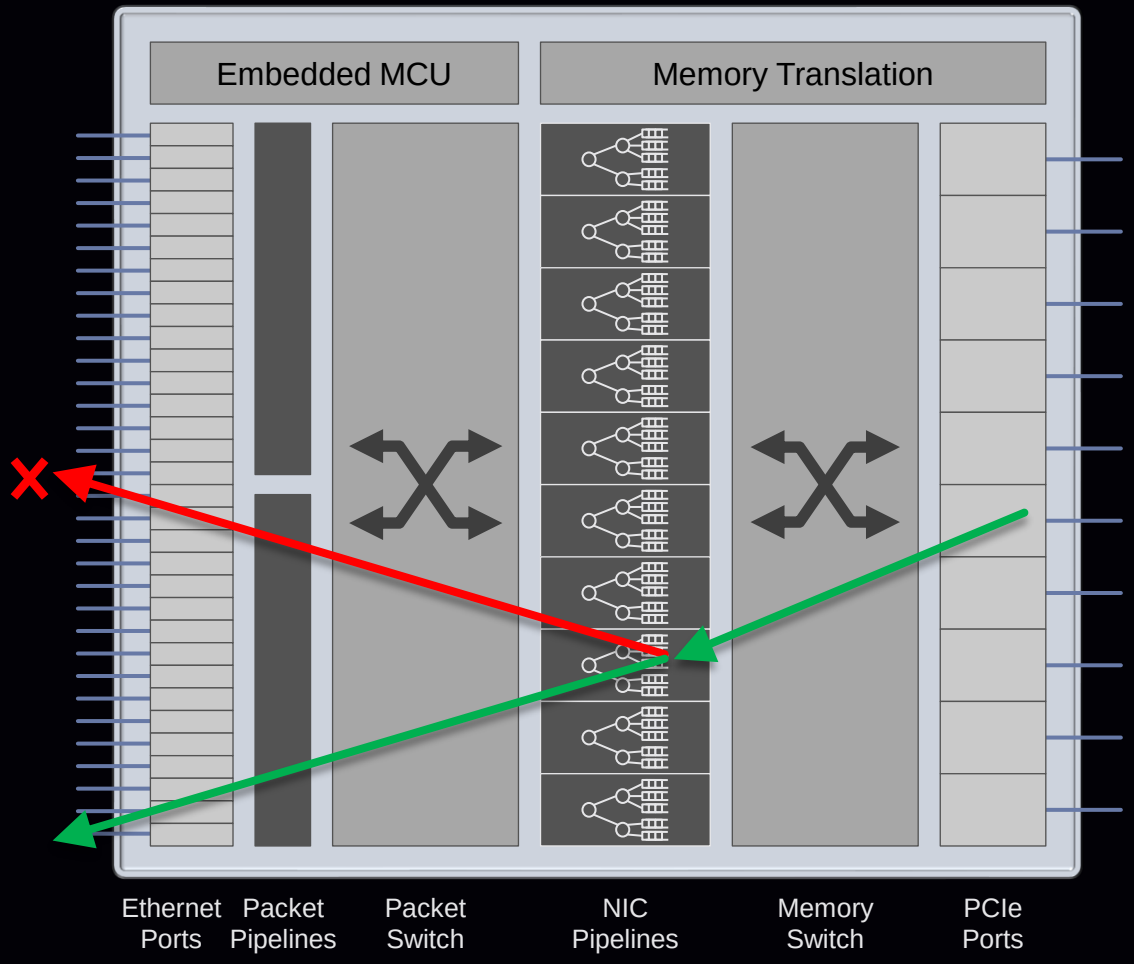
Per rail: 256x 51.2T switches
Total: 16k 51.2T switches

512x instances of the cluster building block

128k x ACF SuperNICs
512k accelerators



:: multiport acf-s gives multipath resilience



- Every GPU sees every uplink!
  Unlike a traditional structure where NIC ←→ ToR path is a single point of failure
- Provides resilience to cable and port failures all the way to the leaf of the network
- Software-defined transport is aware of error state and reacts to dynamic failures under the hood, rebalancing with workload awareness
- High-radix means link loss has very gradual impact on aggregate available uplink bandwidth

:: acf-s mindset for the future

- Let the compute accelerator maximize compute
  - Optimize silicon real estate for FLOPs
  - Scale-up IO: High-efficiency, lightweight, higher baud rate >PCIe
  - Scale-out at the switch: bridge into sophisticated, highly scalable, resilient protocols at the system switch
  - Aggregate by merging RDMA NIC into system switch
- Build useful tiers of headless memory in scale-out fabric
  - Better perf/$ for inference context caches and RAG
- Smart, SW-defined traffic balancing
  - Higher-radix, flatter networks, better performance & resilience
  - Higher-scaling topologies other than full-CLOS
- Optimize software transport with awareness of workload, topology, and telemetry to increase resilience and uptime

enfabrica

:: Thank You