

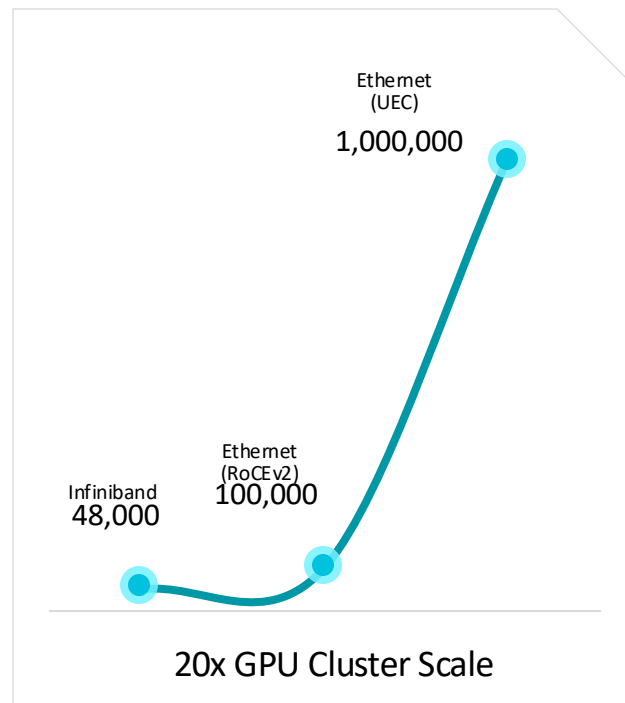
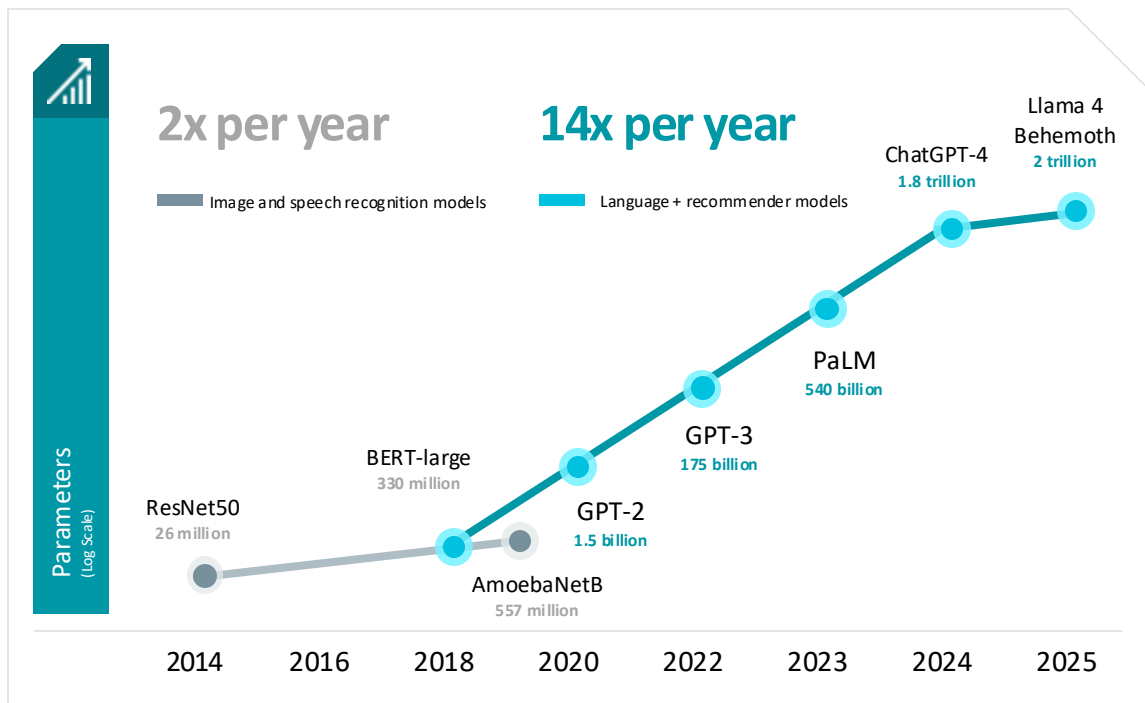


Using P4 NICs for resilient scale-out

GPU Interconnect



The need for Scale-out AI fabric



Traits of Scale-out GPU Interconnect - UEC

Multipathing

Utilizing full bisectonal bandwidth



Congestion Control

Handle lossy networks



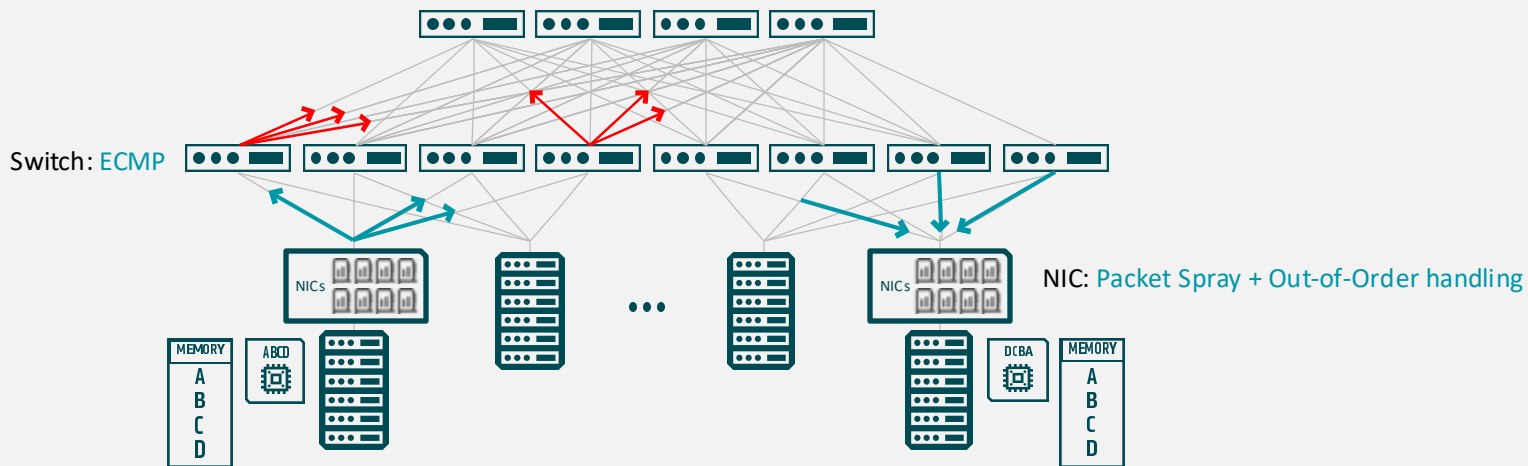
Loss identification

Selective acknowledgement and retransmission



TCO

Maximize compute by eliminating inefficiencies



The challenge of AI fabric scale

100K AI Cluster Key Components



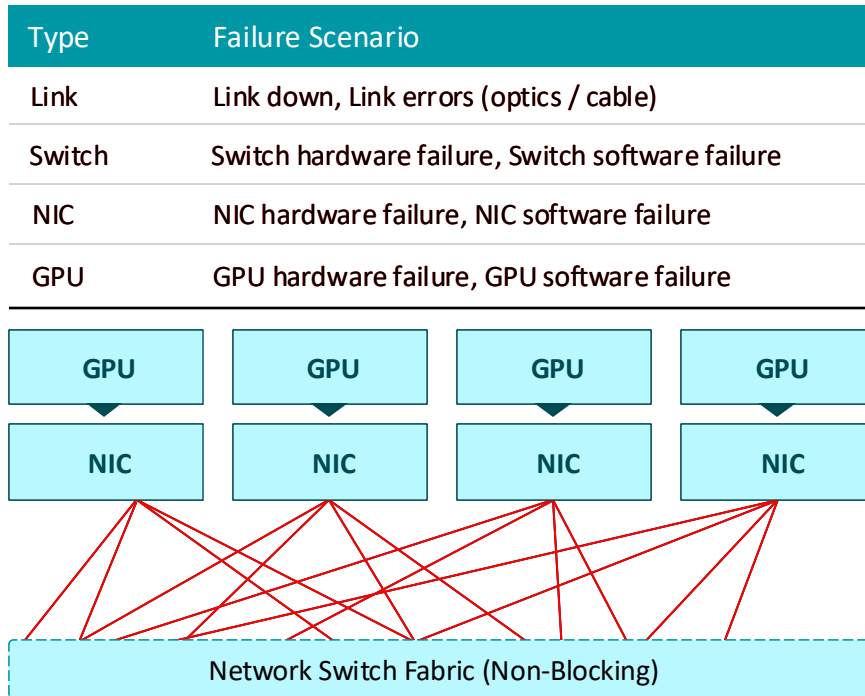
Components	Quantity	Comments
GPU	100K	Typical GPU
Back End NIC	100K	1:1 (GPU:NIC)
GPU Servers	13K	8 GPUs/server
Network Switches	1.2K	512x100GbE ports
Optical Cables	600K+	2-tier design
Transceivers	600K+	QSFP
Racks	~1.6K	64 GPUs/rack

* About 78% of unexpected interruptions were attributed to confirmed or suspected hardware issues.

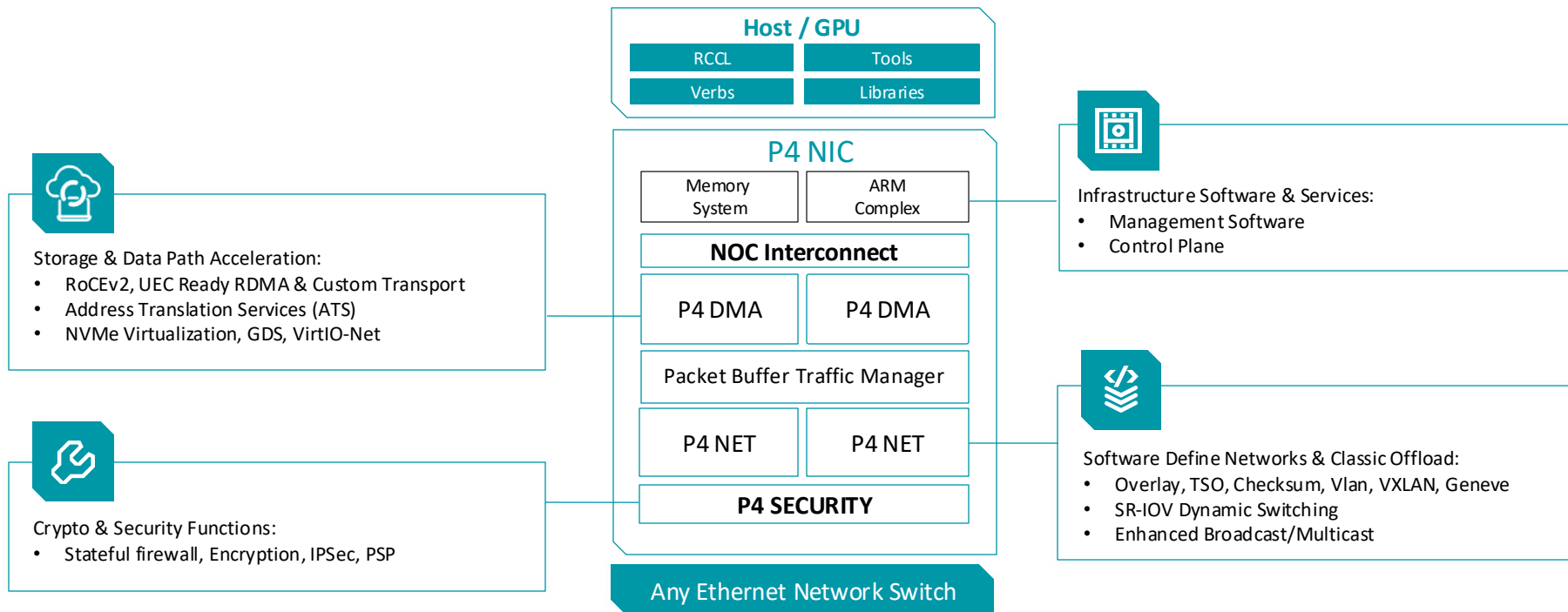
Infrastructure resiliency is not optional



* <https://arxiv.org/pdf/2407.21783> (The Llama 3 Herd of Models)



Why P4 – Modular System Architecture

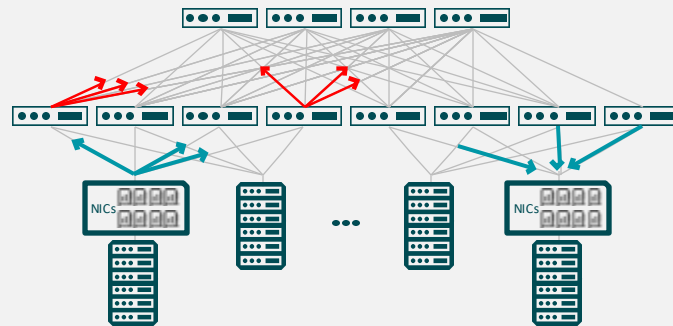
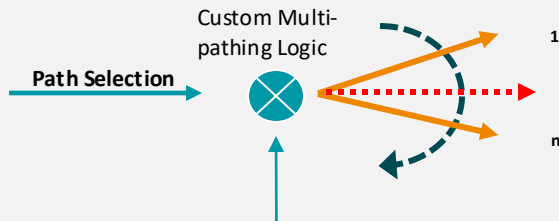


Multi-pathing using P4-based NICs

Multi-pathing is a proven way network failures, but multipathing method varies

- Entropy Values (EVs) based path selection: uses ECMP in the network
- Source routing at the sender NIC/GPU
- Custom Spray (programmable EV discussion in UEC) etc.

P4 offers a programmable approach to solving this challenge



Multi-Plane using P4 NICs

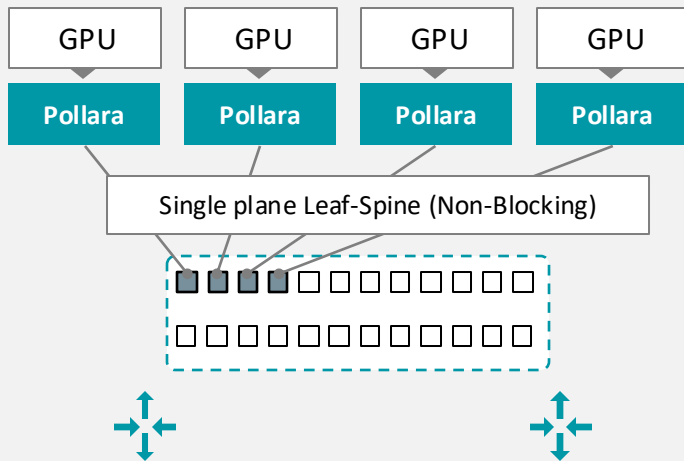


Multi-Plane helps in fault isolation and failures (switch/link/NIC)

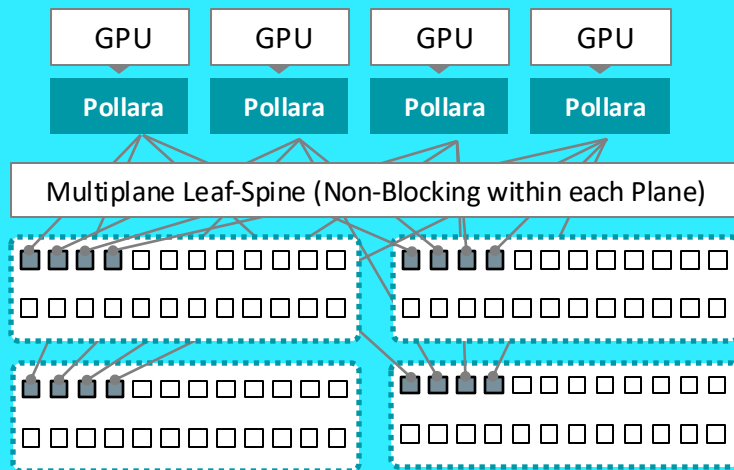


P4-based NIC can implement plane-identification and fast switchover to alternate planes

Standard 400G NIC (single plane)



Multi-plane (4 plane)



* <https://arxiv.org/pdf/2505.09343> (Insights into DeepSeek-V3: Scaling Challenges and Reflections on Hardware for AI Architectures)

Path Probes using P4 NICs



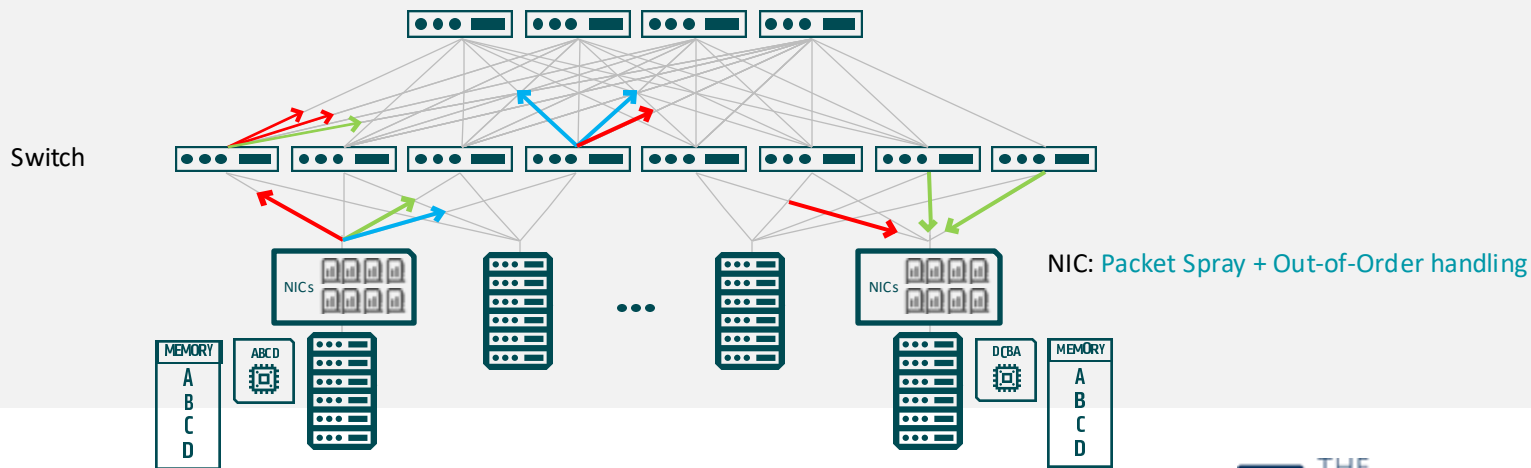
Datapath protocols have aggressive retransmits (given 400/800G speeds)



Identifying failures quicker within 2xRTT is crucial i.e. path failures must be detected in datapath



P4-based NIC can send path probes for specific paths for quicker detection and remediation



Source Routing using P4 NICs



AI applications usually result in pre-determined communication patterns amongst GPUs/NICs

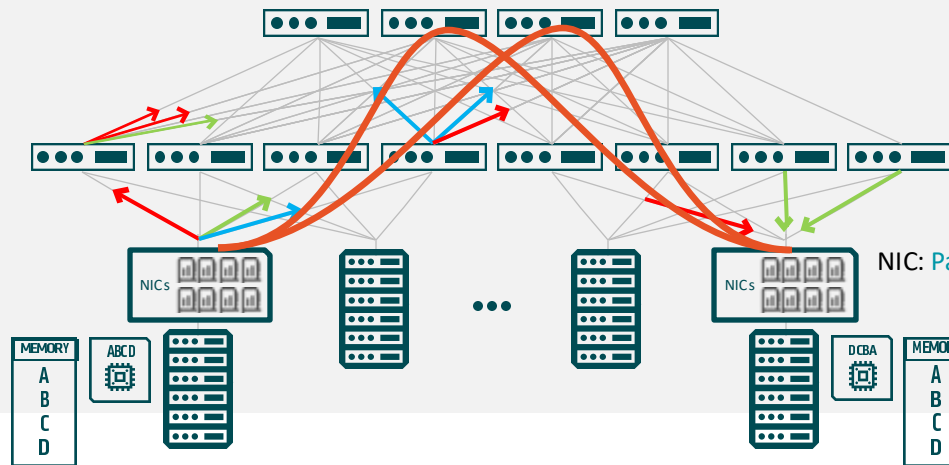


This creates an opportunity to have a pre-determined paths in the network for non-overlapping and most efficient communication in a given network topology



P4 Programmable datapath can adapt to a non-traditional source routing to take advantage of GPU-GPU communication patterns

Switch: Segment Routing



NIC: Packet Spray + Out-of-Order handling





Summary & Call to Action



P4 was invented for doing data/packet processing in datapath



AMD has innovated to do message processing in addition to this and offer a full GPU to GPU scale-out interconnect



AMD Pensando™ Pollara 400 AI NIC has proven the strength of P4 in unforgiving network circumstances, offering resiliency at scale



[Validated Reference Guide](#)



Where to find additional information
<https://www.amd.com/pensando>

