Fabric resiliency at scale

Spectrum-X enabled fabric for AI training

OCP GLOBAL SUMMIT | OCT 15-17, 2024
SAN JOSE, CA

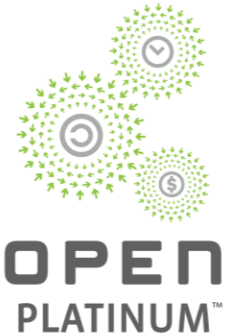
Networking

NETWORKING

Fabric resiliency at scale

Jeff Tantsura , Omer Shabtai

NVIDIA



When wires get tired, Global AR keeps AI infra team hired

- Large AI workloads are sensitive to fabric failures
  - Long running jobs
  - Large scale and tightly coupled jobs
  - Demanding high bisectional BW

- Each link / switch failure can cause job crush
  - Losing all training progress until last check point
    - minutes to hours
  - Large job re-initialization time

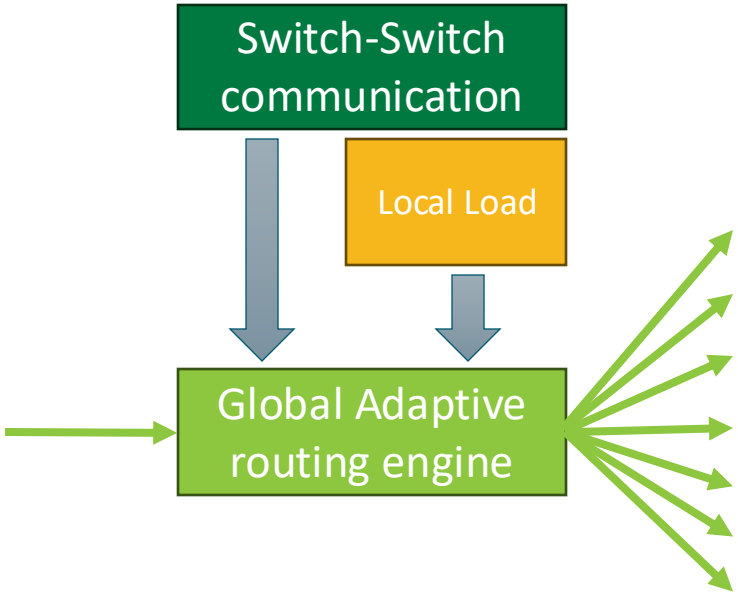
- Performance drop due to fabric a-symmetry
  - A Non-linear, where a single link down will severely harm the training speed

Resiliency KPIs special in AI training

1. Convergence time
    - AI workloads are tightly coupled and can reach significant scale
        - Fabric issues cause Job failure (during Llama3 training – 35 link/switch failed)
        - Collective can fail after several seconds of bad connectivity ➔ On a failure the fabric must converge as fast as possible

2. Performance guaranties at steady state
    - Job can progress only as fast as the weakest link
        - Measure P01 BW under full bisection traffic

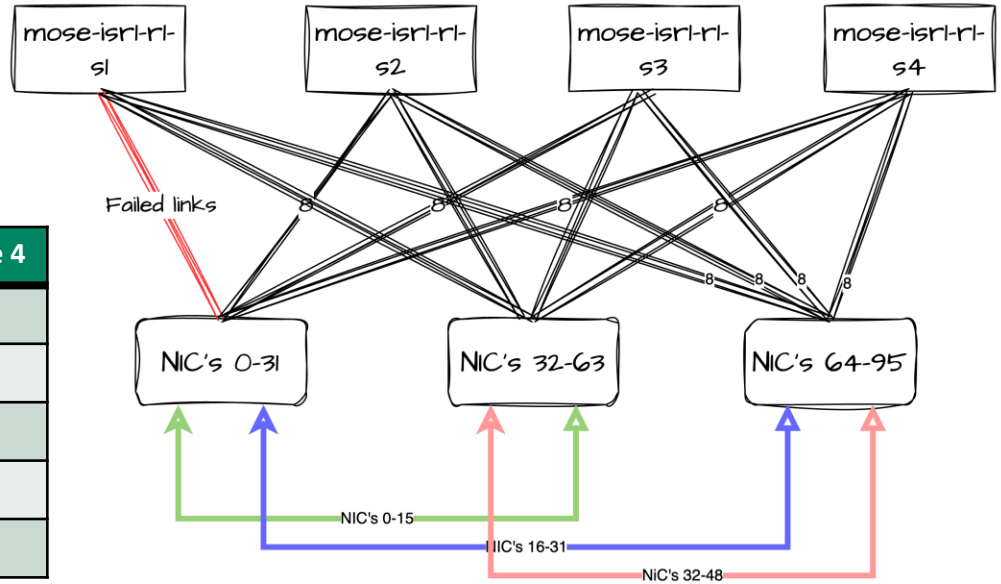
High level overview



SETUP – 3X32 ENDPOINTS

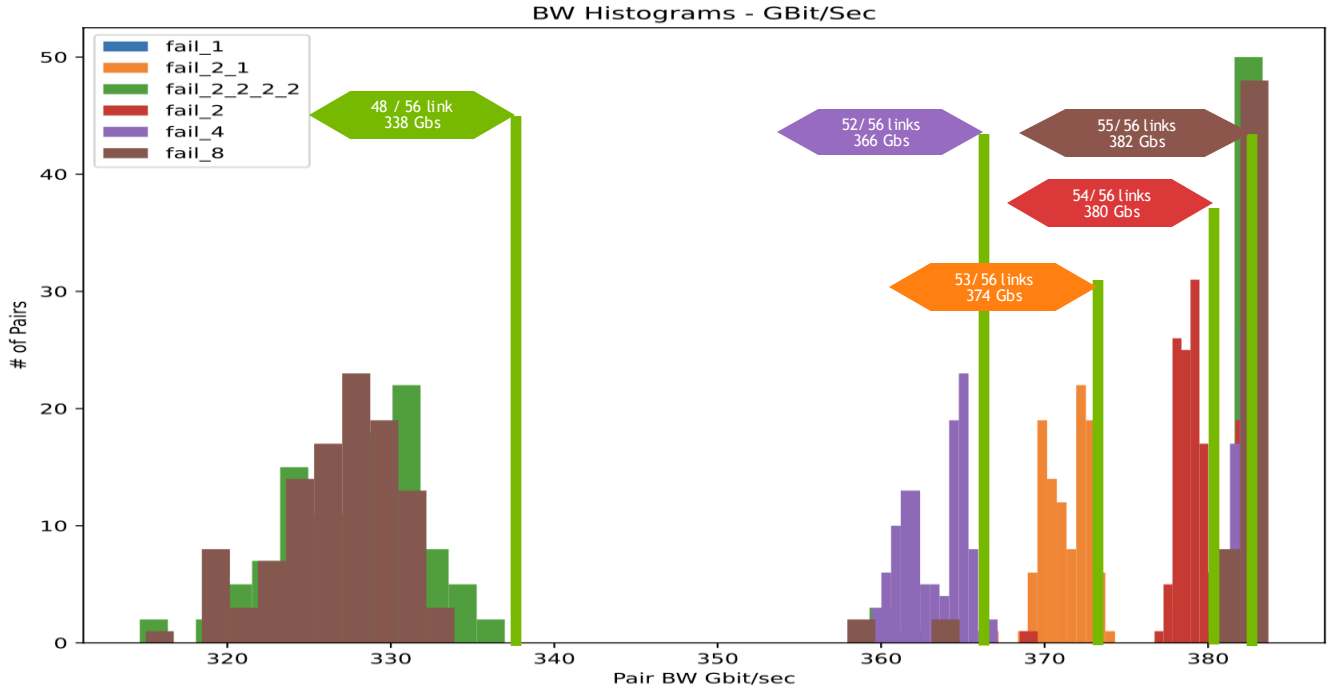
- Leaves: 3x16 Nodes – 3x32 NIC's
- Spines: 4 spines, 8 parallel links
- Traffic: 3x16 RDMA pairs, Bi-directional traffic
- Failures Configuration:

Test case	Spine 1	Spine 2	Spine 3	Spine 4
1 Failure	7	8	8	8
2 Failures	6	8	8	8
4 Failures	4	8	8	8
2+1 Failure	6	7	8	8
2+2+2+2	6	6	6	6



Bisection BW under link Failures

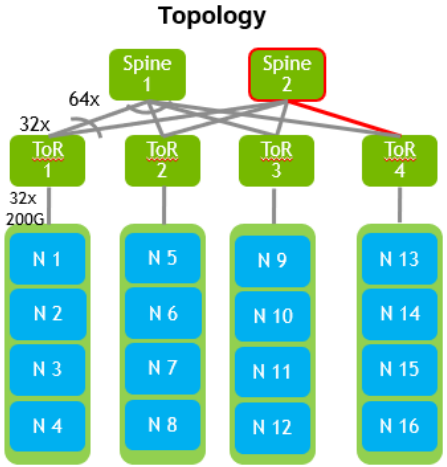
3 SU * 28 Nodes, 4 spines, 14 parallel links

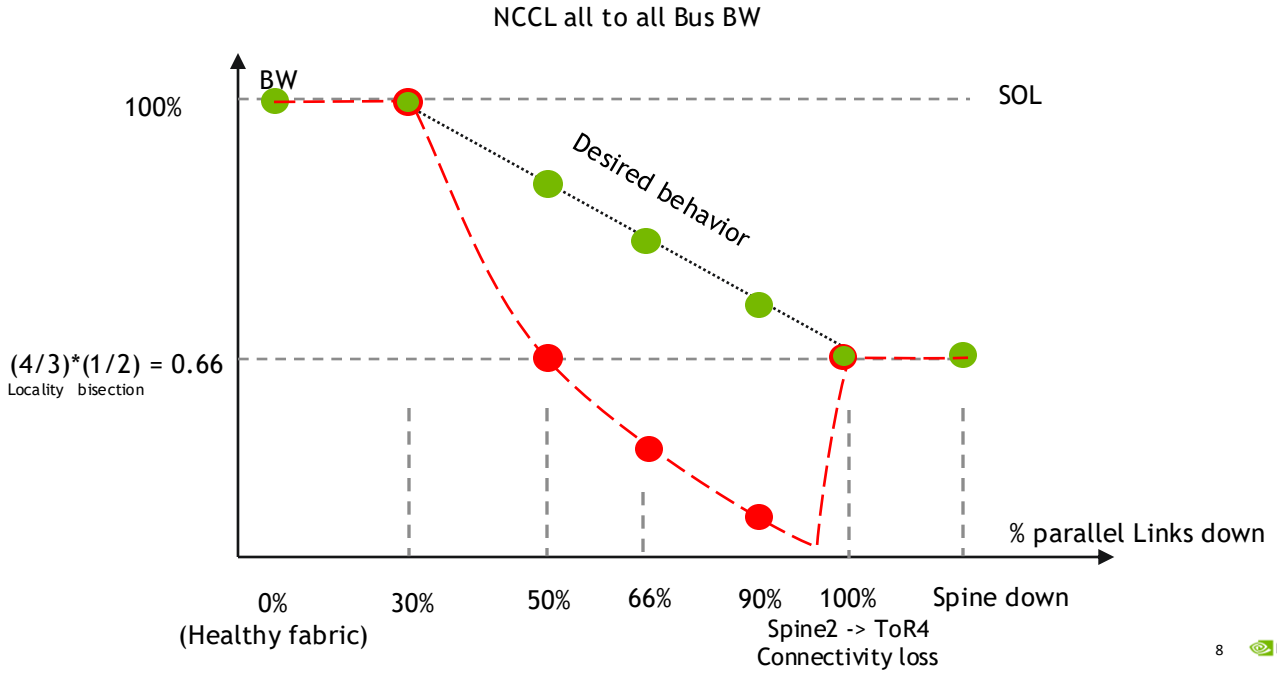


Resiliency - challenges

All to All Collective performance for asymmetric fabric

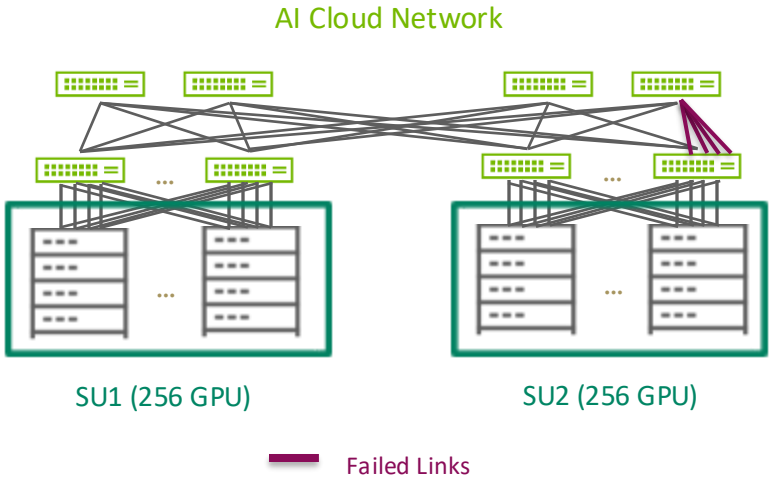
- Baseline - Non-linear degradation grows with a-symmetry
- On a large scale topology with few parallel links – much worse

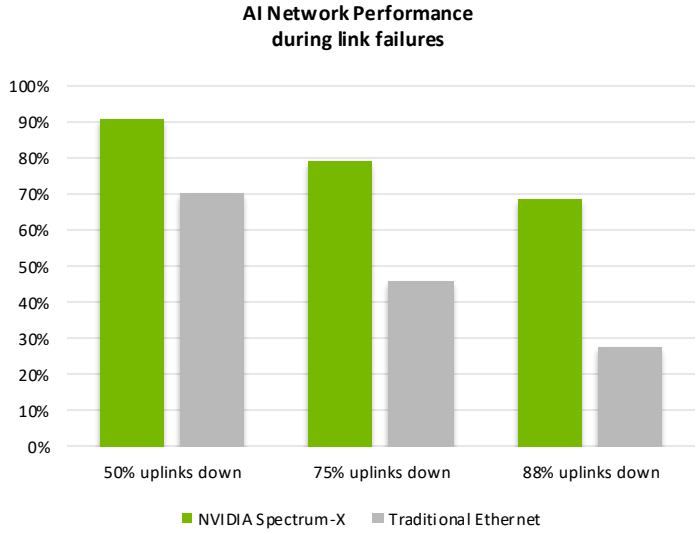




Resilient Adaptive Routing Performance

Link Failures on Traditional Ethernet Led to Outsized Drop in AI Performance





Spectrum-X utilizes Global AR to rebalance NCCL flows and avoid failed paths

Call to Action

- Start AI Cloud POC with Spectrum-X
- Use NVIDIA AIR Simulation environment with SONiC today

- More Information on Spectrum-X
  - Spectrum-X Platform - https://www.nvidia.com/en-us/networking/spectrumx/
  - Spectrum-X Video - https://www.youtube.com/watch?v=nKqfi3q4S5I
  - BlueField DPU - https://www.nvidia.com/en-us/networking/products/data-processing-unit/
  - Spectrum Switches - https://www.nvidia.com/en-us/networking/ethernet-switching/
  - NVIDIA SAI - https://developer.nvidia.com/networking/ethernet-switch-sdk

Open Discussion