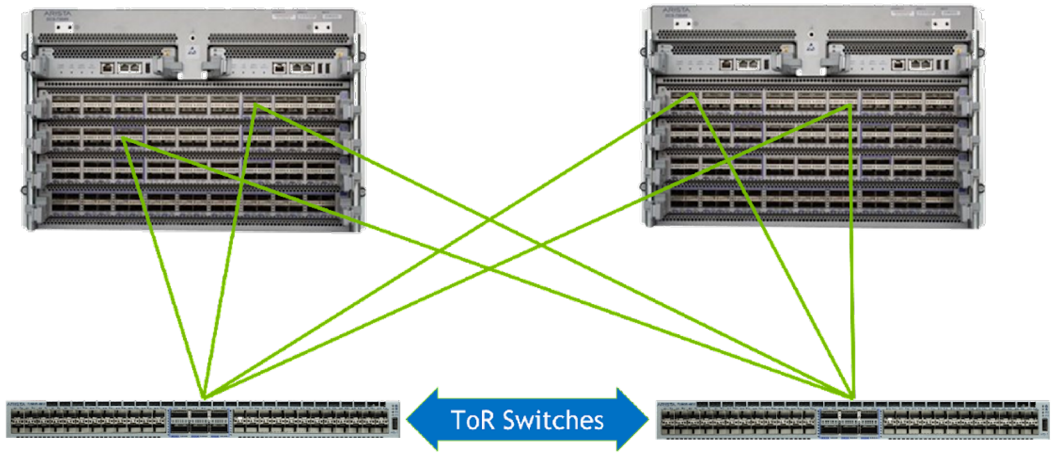
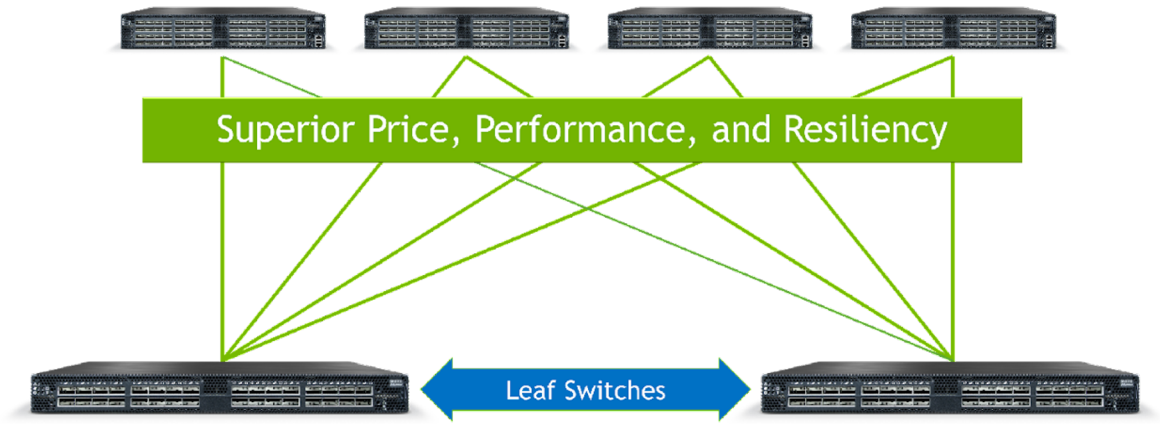
Adaptive Routing For AI & Storage

David Iles

Sr. Director at NVIDIA

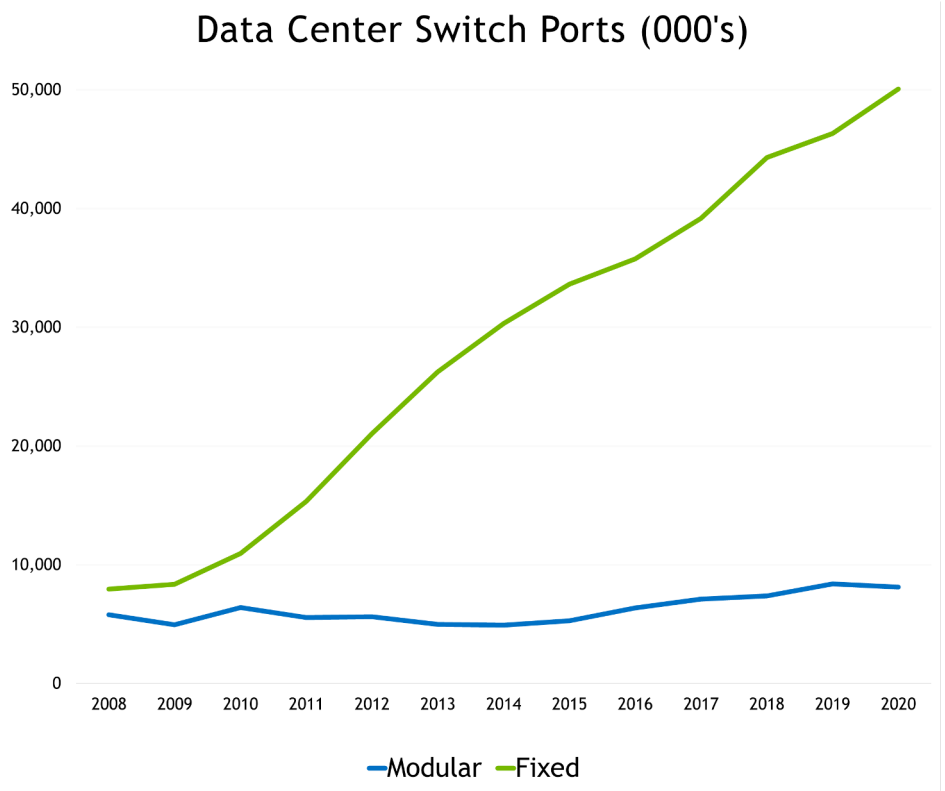
Leaf/Spine Networks

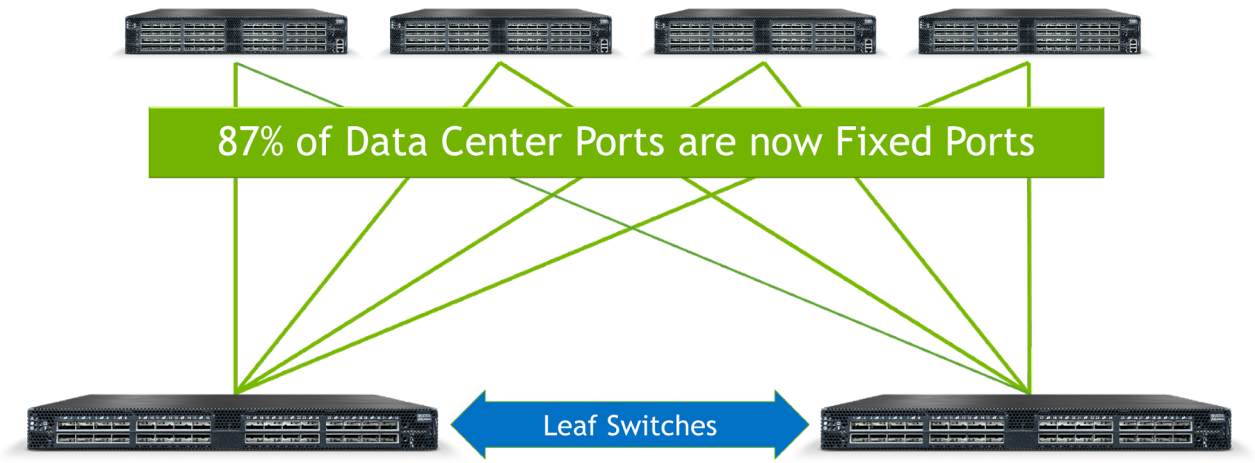




Scale Up Network → Scale Out Network

From Modular to Fixed Port Switches





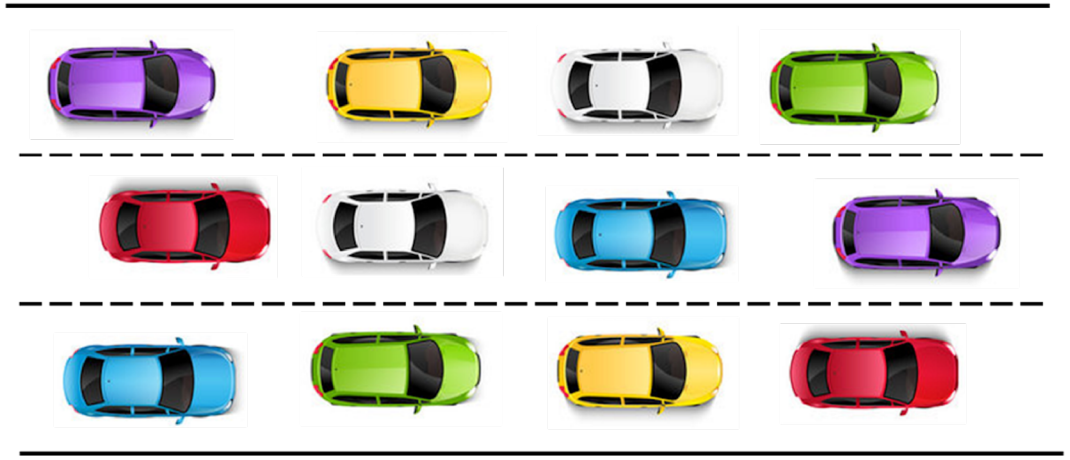
Source: Crehan - Data Center Ethernet Switch Market Share 2021Q1

OPEN Compute Project®

AI and Storage Network Traffic

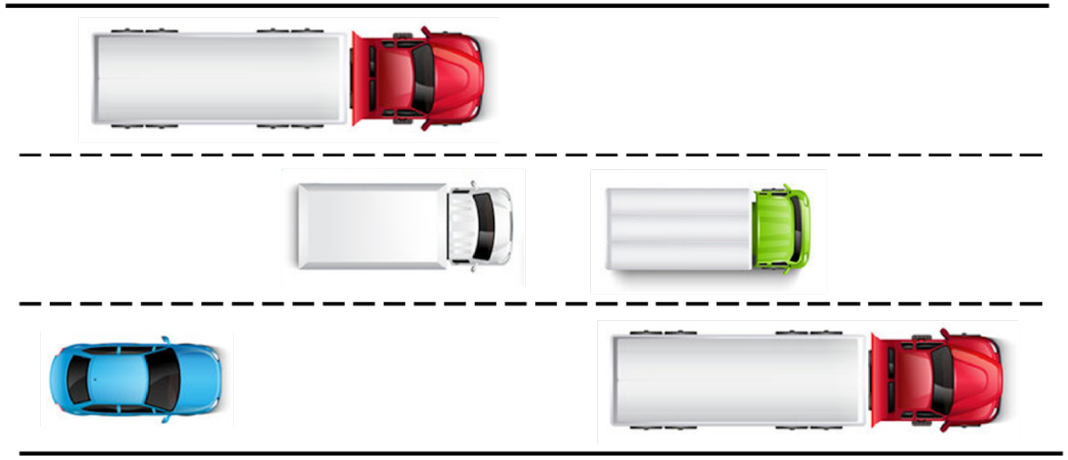
Traditional Network Traffic

Small Packets · Lots of flows · Low Variance



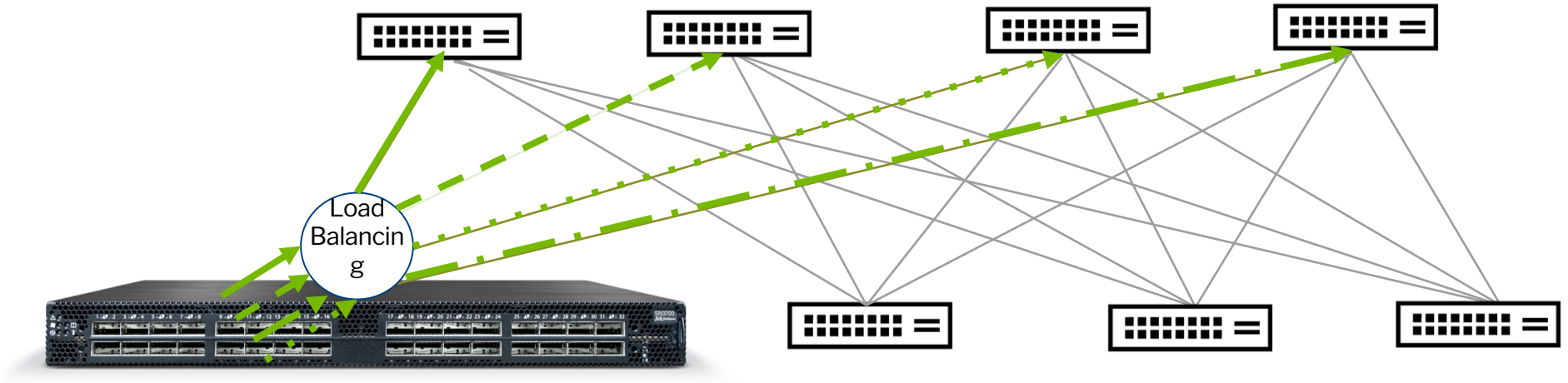
Modern Application Traffic

Very Large Packets · Few flows · High Variance



Connect. Collaborate.
Accelerate.

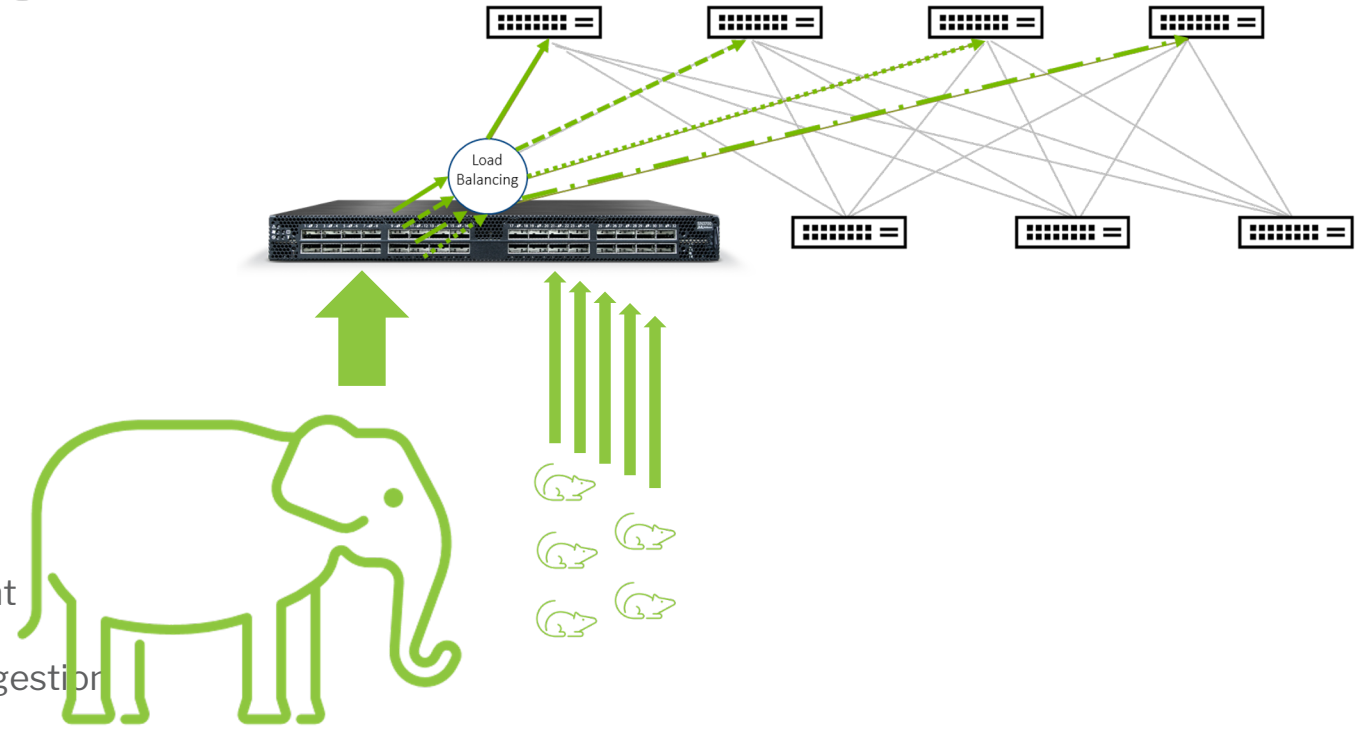
Adaptive Routing from NVIDIA



Traditional ECMP	Adaptive Routing
Static hashing	Congestion based port selection
Independent of traffic conditions	Flowlet-aware: eliminates out-of-order Packets
Bigger flows = higher chance for congestion	Multi-vendor friendly
High tail latency	RoCE OOO placement for highest efficiency

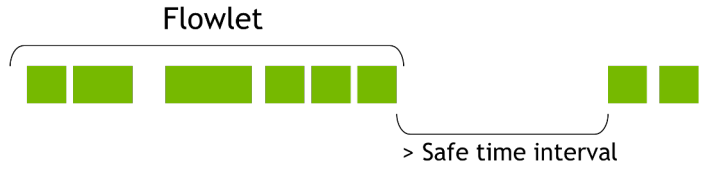
Adaptive Routing

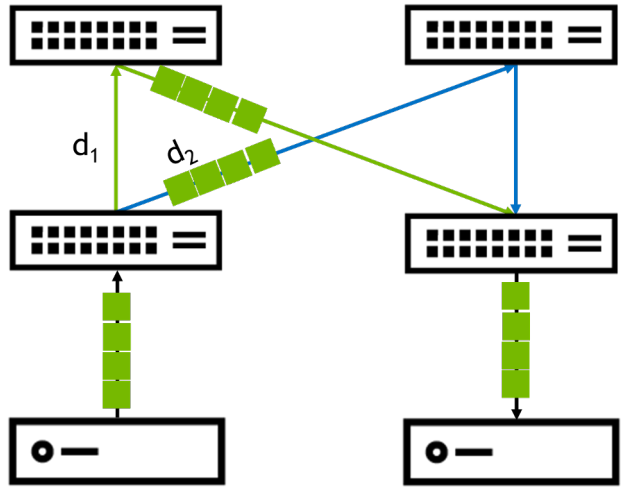
- Elephant flows
  - High-bandwidth
  - Long-lived
  - Most of AI & Storage traffic
- Mice flows
  - Low-bandwidth
  - Short-lived
  - Latency sensitive
- Chance of leaf/spine congestion increases as # of Elephant flows increase
- Adaptive Routing adjusts Elephant flow path to avoid congestion
- Spectrum HW Elephant Identification Engine
  - Adaptive Routing can be used just on Elephant flows
  - Mice flows are already well distributed



Flowlet-aware Adaptive Routing

- Long-lived flows broken into flowlets
  - Each flowlet can be routed independently
- Flow table track active flow → link
- Flow link association changes allowed when safe
  - Safe period is configurable
  - Result: Avoid out of order
- Flowlet resolution 1us – 10min

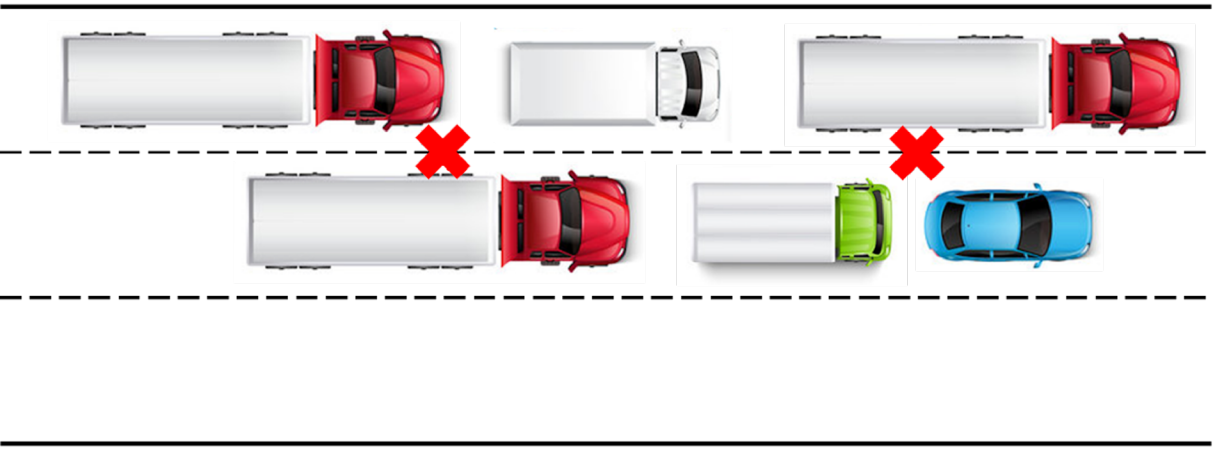




Adaptive Routing

Without Adaptive Routing

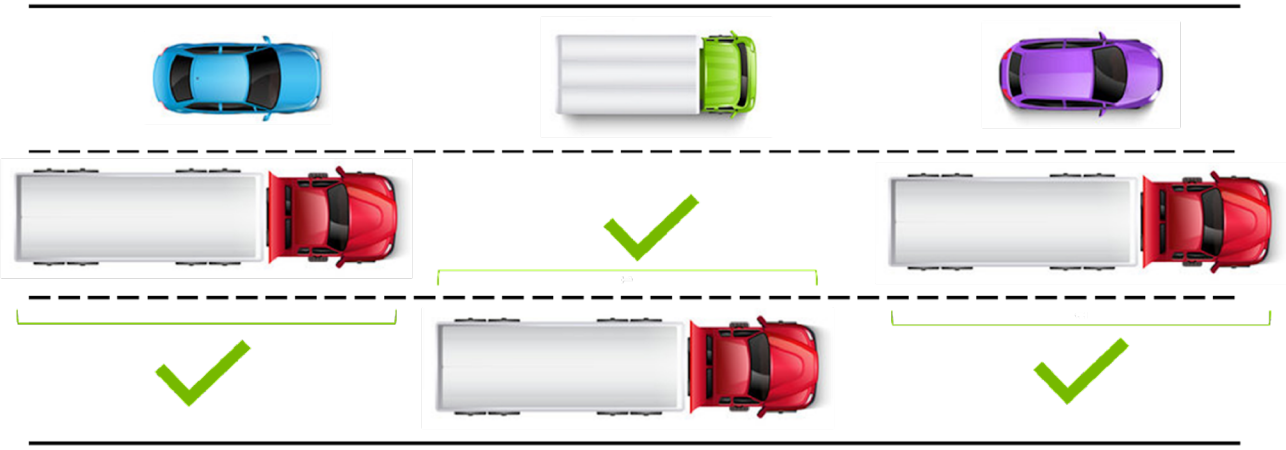
Elephant flows conflict for port access, resulting in congestion



Cars and large trucks prevent merging without smart use of highway lanes

With Adaptive Routing

Elephant flows are dynamically routed while Mice flows are unaffected



Cars zip along while trucks at full speed maintain gaps between each other for merging

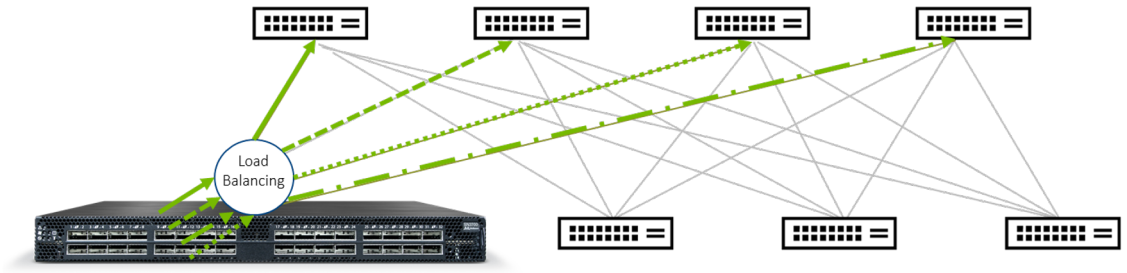
Adaptive Routing

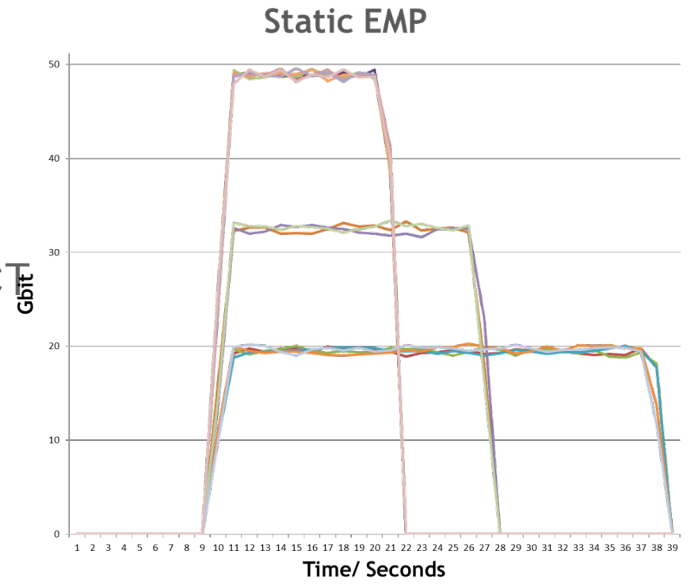
Performance Test Results

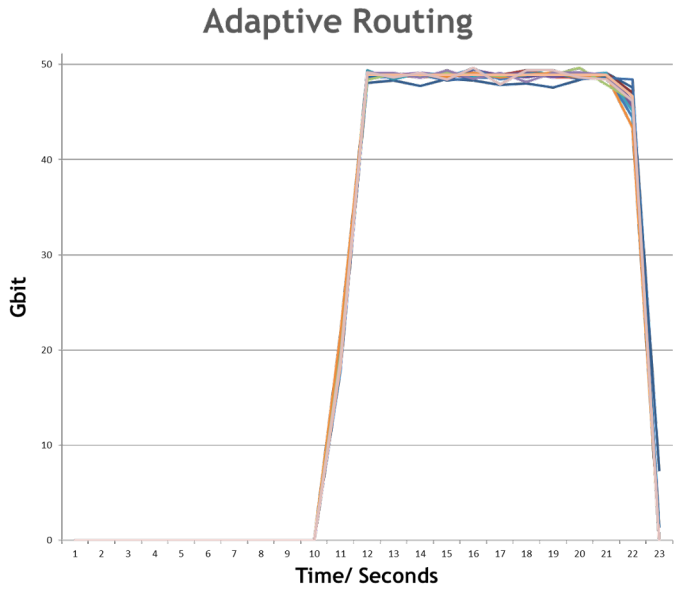
Connect. Collaborate. Accelerate.

Adaptive Routing Test with Traffic Generator

- Single switch transmitting to spines using ECMP
- 16 large-size flows, 50G host links
- Benefits
  - Consistent FCT across senders, lower 99% FCT
  - Higher goodput, shorter completion time

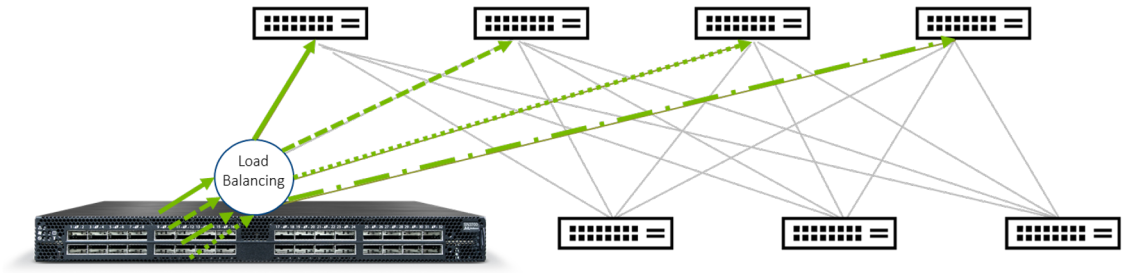


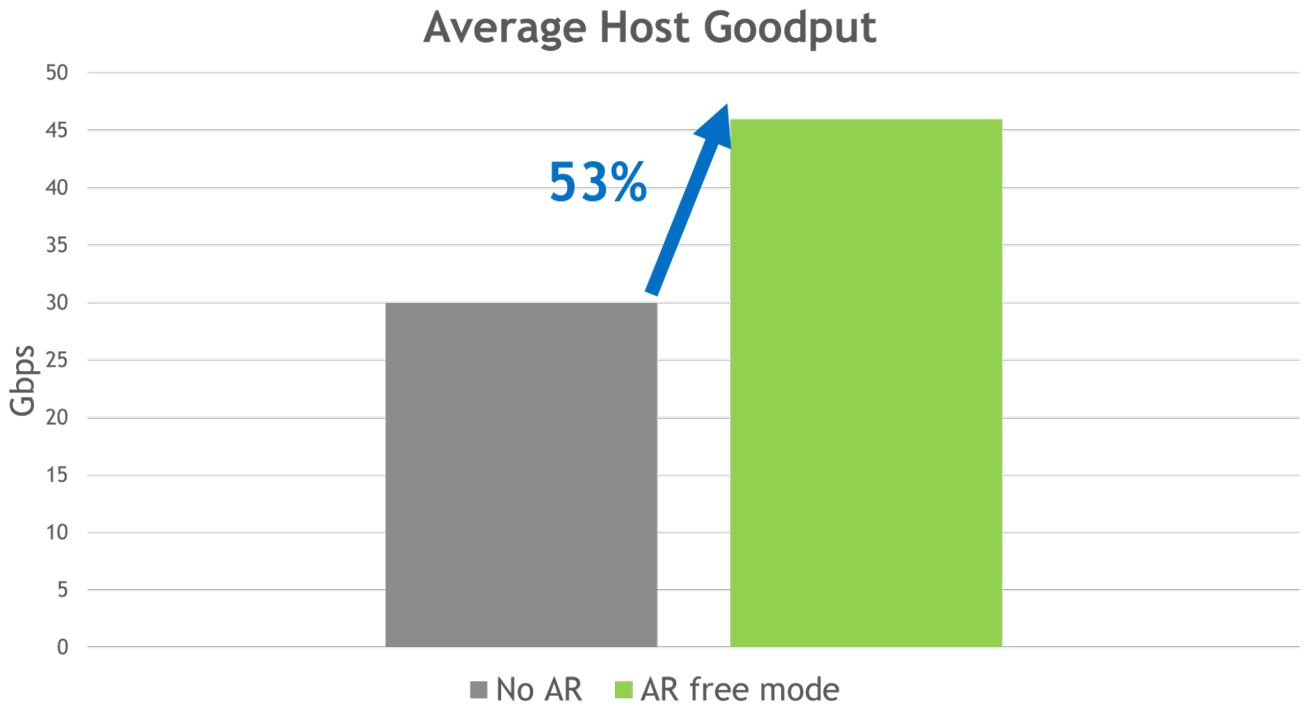




Adaptive Routing Test with Traffic Generator

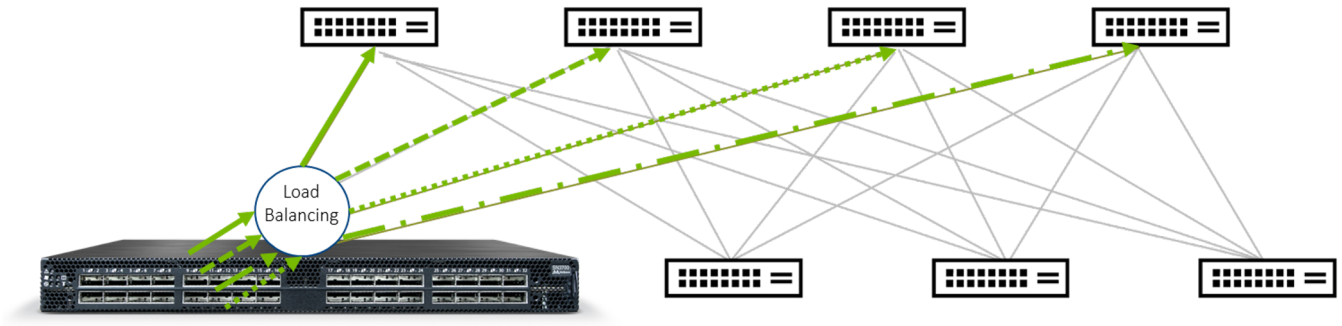
- Single switch transmitting to spines using ECMP
- 16 large-size flows, 50G host links
- Benefits
  - Consistent FCT across senders, lower 99% FCT
  - Higher goodput, shorter completion time

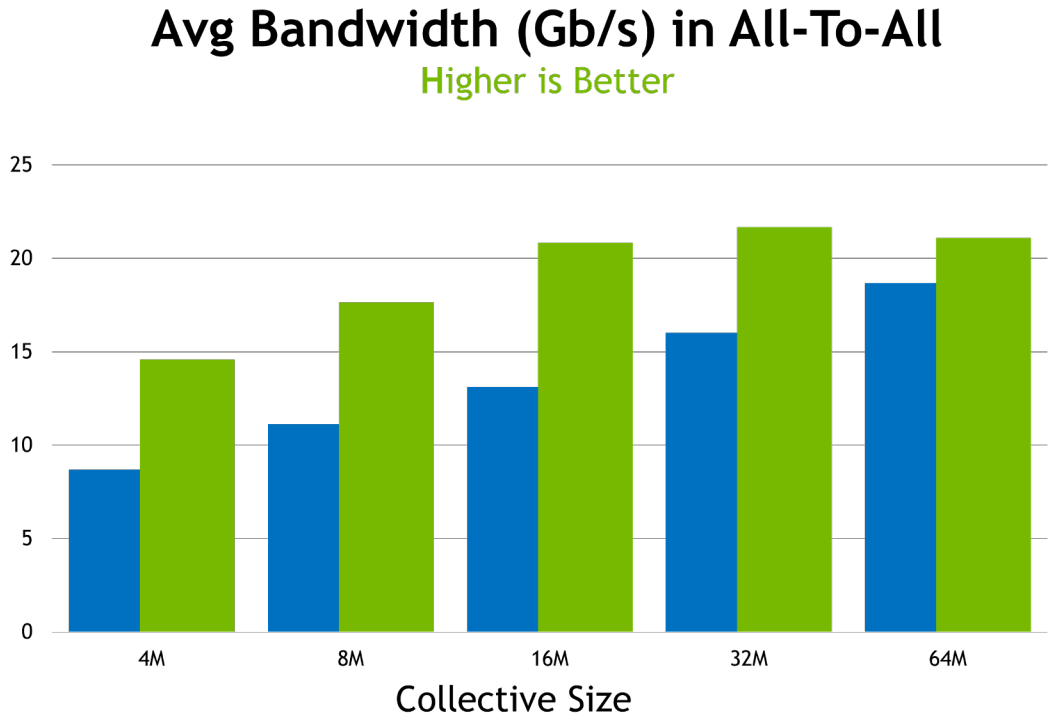




Connect. Collaborate.
Accelerate.

Real World AI Traffic Test (All-to-All)



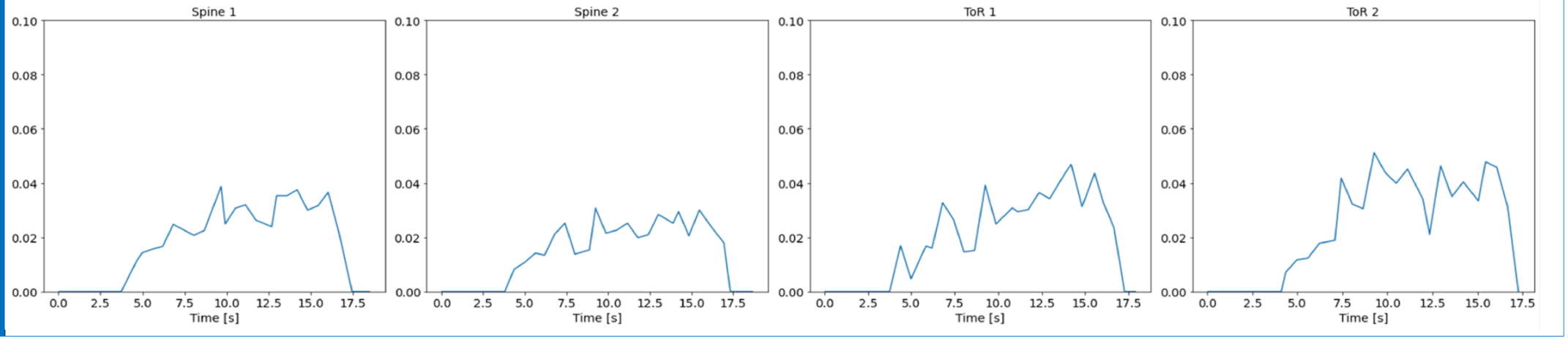


Setup: Spectrum + ConnextX-6Dx +CPU + UCX PARAM communications micro-benchmark

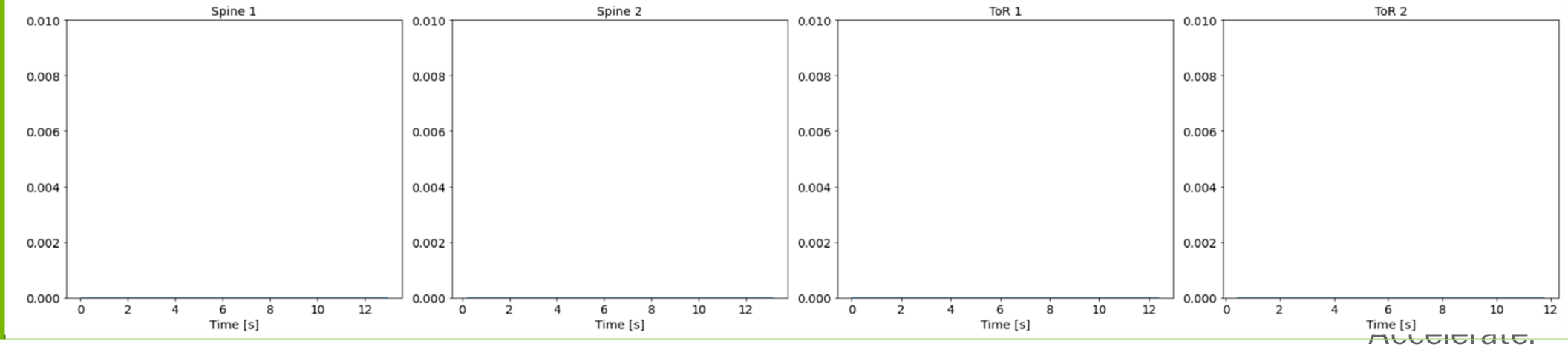
Connect. Collaborate. Accelerate.

AI Traffic Test (All-to-All) Reduction in Buffer

Static ECMP

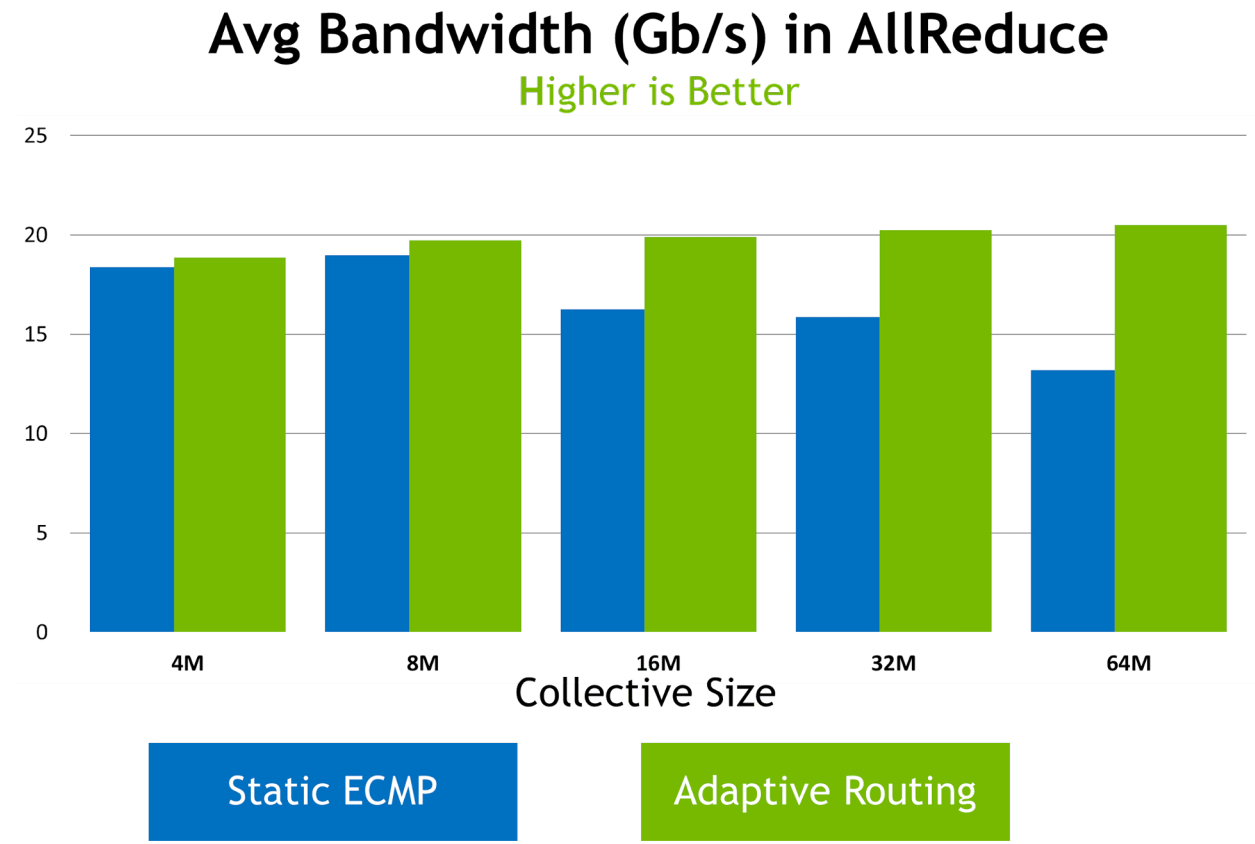


Adaptive Routing



Real World AI Traffic Test (All-Reduce)

- 1.5x Higher Bandwidth with AR
- Tested on CPUs
- NCCL uses Double Binary Trees or Ring



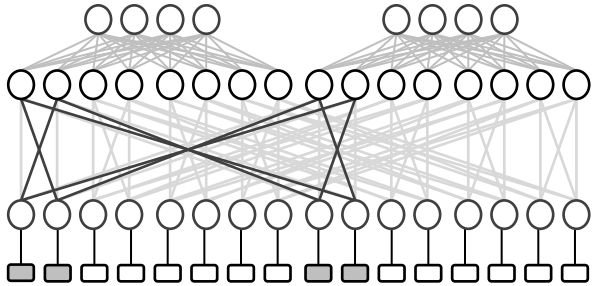
Setup: Spectrum + ConnextX-6Dx +CPU + UCX PARAM communications micro-benchmark

Connect. Collaborate. Accelerate.

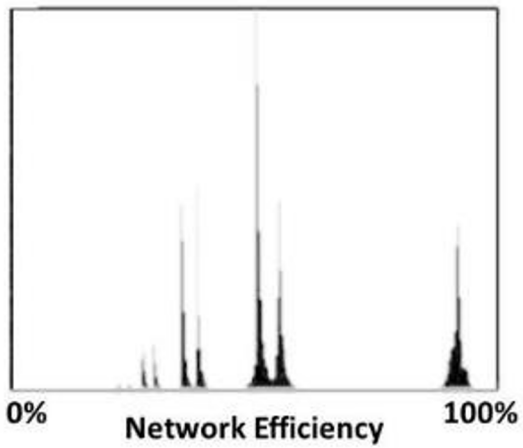
Adaptive Routing In Production

The Design, Deployment, and Evaluation of the CORAL Pre-Exascale Systems

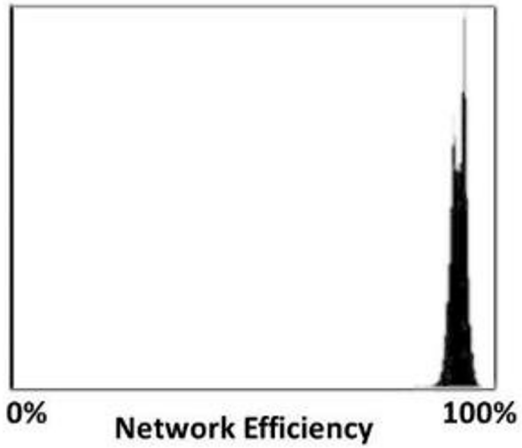
Sudharshan S. Vazhkudai†, Bronis R. de Supinski‡, Arthur S. Bland†, Al Geist†, James Sexton*, Jim Kahle*, Christopher J. Zimmer†, Scott Atchley†, Sarp Oral†, Don E. Maxwell†, Veronica G. Vergara Larrea†, Adam Bertsch‡, Robin Goldstone‡, Wayne Joubert†, Chris Chambreau‡, David Appelhans*, Robert Blackmore*, Ben Casses‡, George Chochia*, Gene Davison*, Matthew A. Ezell†, Tom Gooding*, Elsa Gonsiorowski‡, Leopold Grinberg*, Bill Hanson*, Bill Hartner*, Ian Karlin‡, Matthew L. Leininger‡, Dustin Leverman†, Chris Marroquin*, Adam Moody‡, Martin Ohmacht*, Ramesh Pankajakshan‡, Fernando Pizzano*, James H. Rogers†, Bryan Rosenburg*, Drew Schmidt†, Mallikarjun Shankar†, Feiyi Wang†, Py Watson‡, Bob Walkup*, Lance D. Weems‡, Junqi Yin†
† Oak Ridge National Laboratory, ‡ Lawrence Livermore National Laboratory, * IBM
{vazhkudaiss@ornl.gov, bronis@llnl.gov}



mpiGraph: Static vs. Adaptive Routing



Static Routing



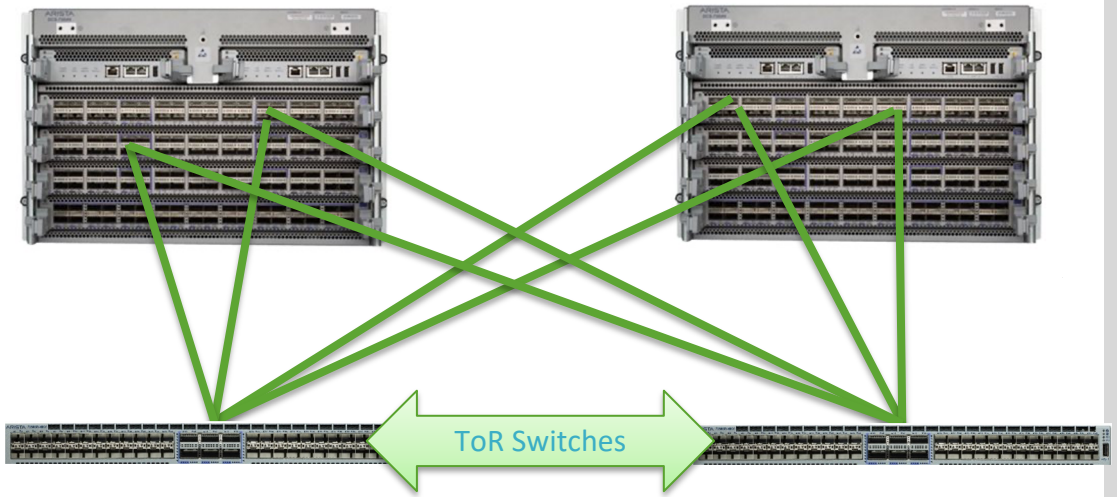
Adaptive Routing

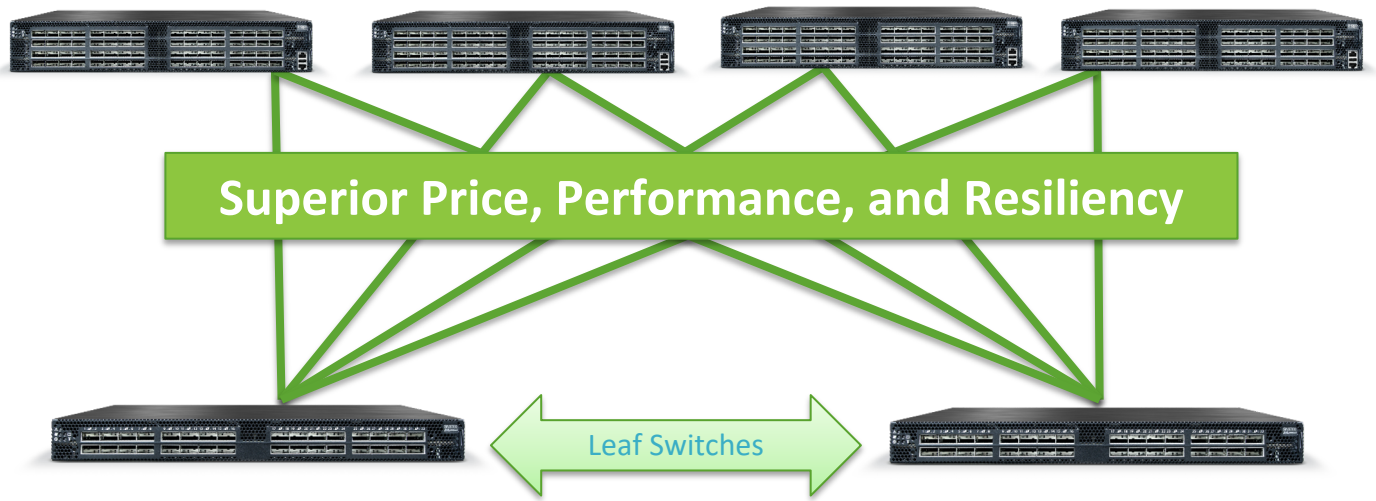
Thank You!

Connect. Collaborate. Accelerate.

OPEN Compute Project®

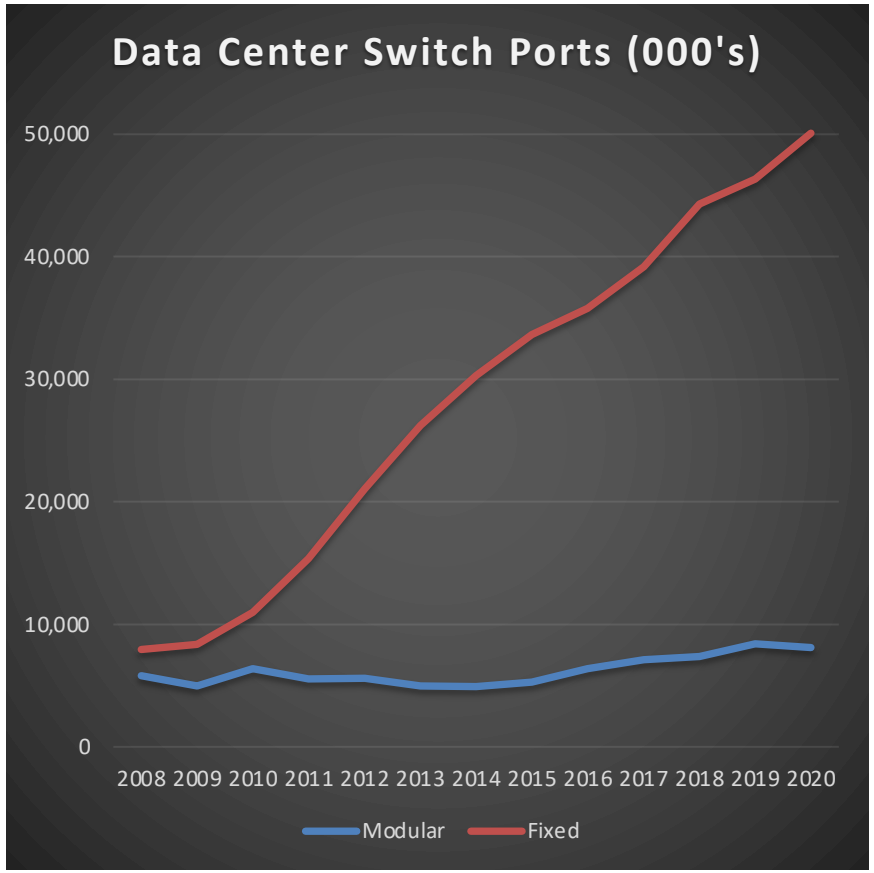
Leaf/Spine Networks



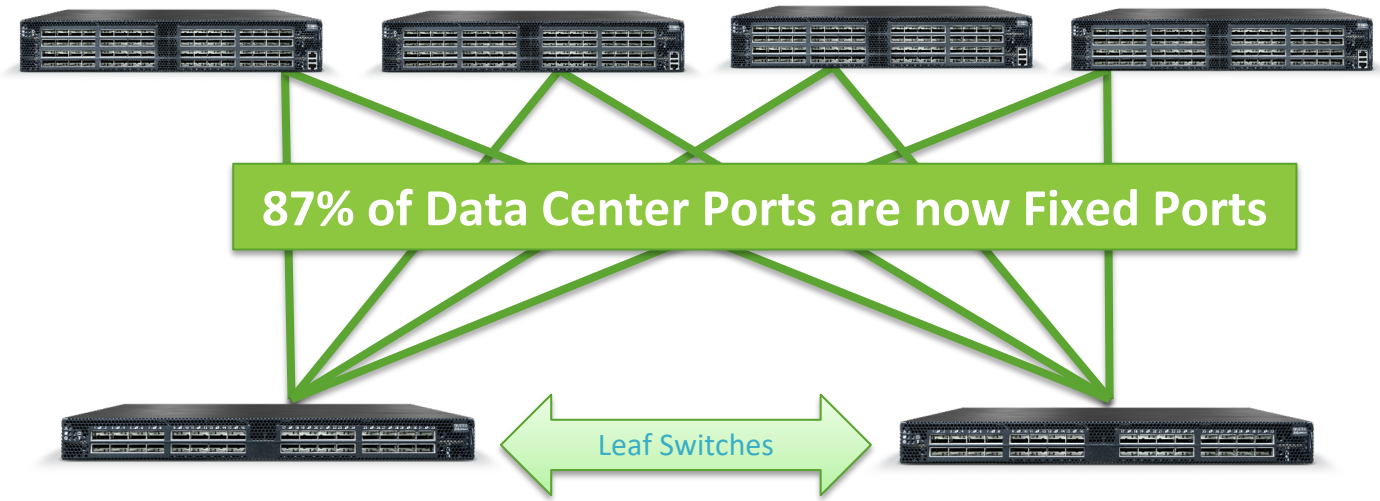


Scale Up Network → **Scale Out Network**

From Modular to Fixed Port Switches



Source: Crehan – Data Center Ethernet Switch Market Share 2021Q1



Connect. Collaborate.
Accelerate.