Optical Shuffle Architectures for Large AI Networks

Jose M. Castro
Distinguished Fiber R&D Engineer
Fiber R&D Labs

IEEE 802.3 New Ethernet Applications Ad Hoc Ethernet for AI Assessment, February 2026

PANDUIT™

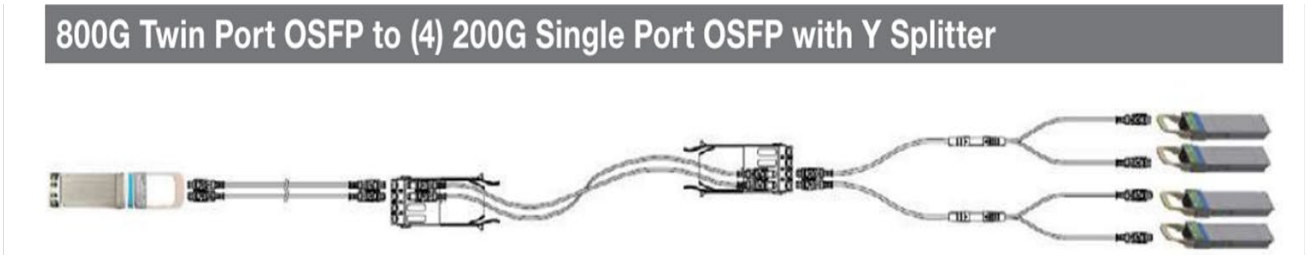
Background

- Flatter network architectures require less switches and transceivers per XPU reducing power consumption and communication latency.
- Flatter networks rely on lane shuffling (meshing) using optical breakouts.
- A breakout lets you re-map lanes so one transceiver can connect to multiple endpoints (across tiers) instead of a single peer.
  - —Example: 1.6T = 8×200G lanes; each lane can be distributed to different switch ports.
- Shuffling optical lanes manually during deployment is unfeasible due to the large number of connections.
  - —Moreover, next-gen transceiver will require optical link training, further constraining lane assignment among PMDs.

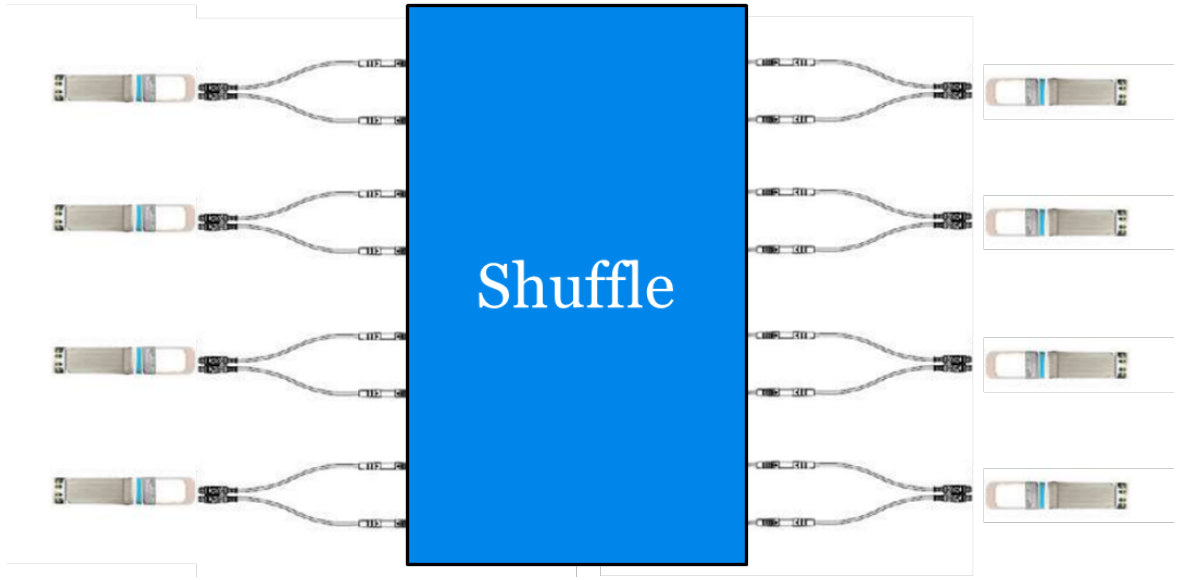
DR Transceivers Optical Lane Breakouts

Without Optical Lane Breakouts

Optical Lane Breakouts



Optical Lane Shuffles



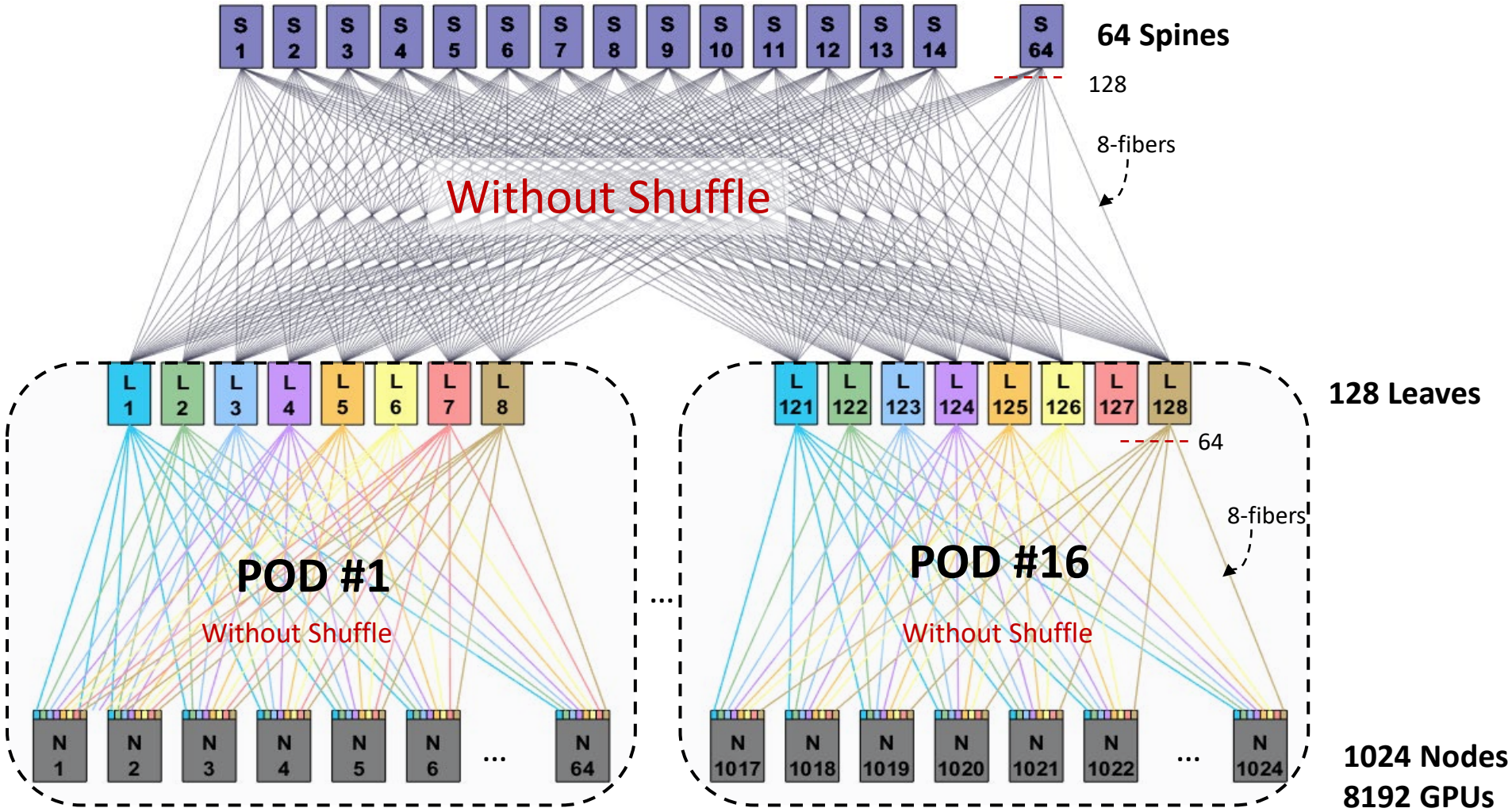
Lane Breakout & Shuffling Benefits in AI Fabrics

- Lane shuffling is being adopted in AI networks to improve performance while lowering cost and power.
  - Lane breakouts enable finer multipath load balancing (packet spraying) improving network reliability.
  - A 2-tier leaf–spine network saves two switch hops compared to a 3-tier leaf–spine–core network, reducing the network contribution to tail latency.
  - A 2-tier network typically requires 40% fewer switches and 50% fewer fabric transceivers than a 3-tier network.
    - Including GPU–leaf optics, the total transceiver reduction is ~33%

Switch Radix	Servers/ POD	GPUs/POD	SW Quad Ports (800G)	POD Breakouts	Leaf-Spine Breakouts	Max # Leaves	Max. # Spines	Max # PODs	Max # GPUs	Notes
512	**64**	**512**	**128**	**1**	**1**	**128**	**64**	**16**	**8192**	No Lane Breakouts
512	**64**	**512**	**128**	**1**	**4**	**512**	**256**	**64**	**32768**	**Breakouts between Switches**
512	**256**	**2048**	**128**	**4**	**4**	**512**	**256**	**64**	**131072**	**Full breakouts**
1024	64	512	128	1	1	128	64	16	8192	No Lane Breakouts
1024	64	512	128	1	8	1024	512	128	65536	**Breakouts between Switches**
1024	512	4096	128	8	8	1024	512	128	524288	**Full breakouts**

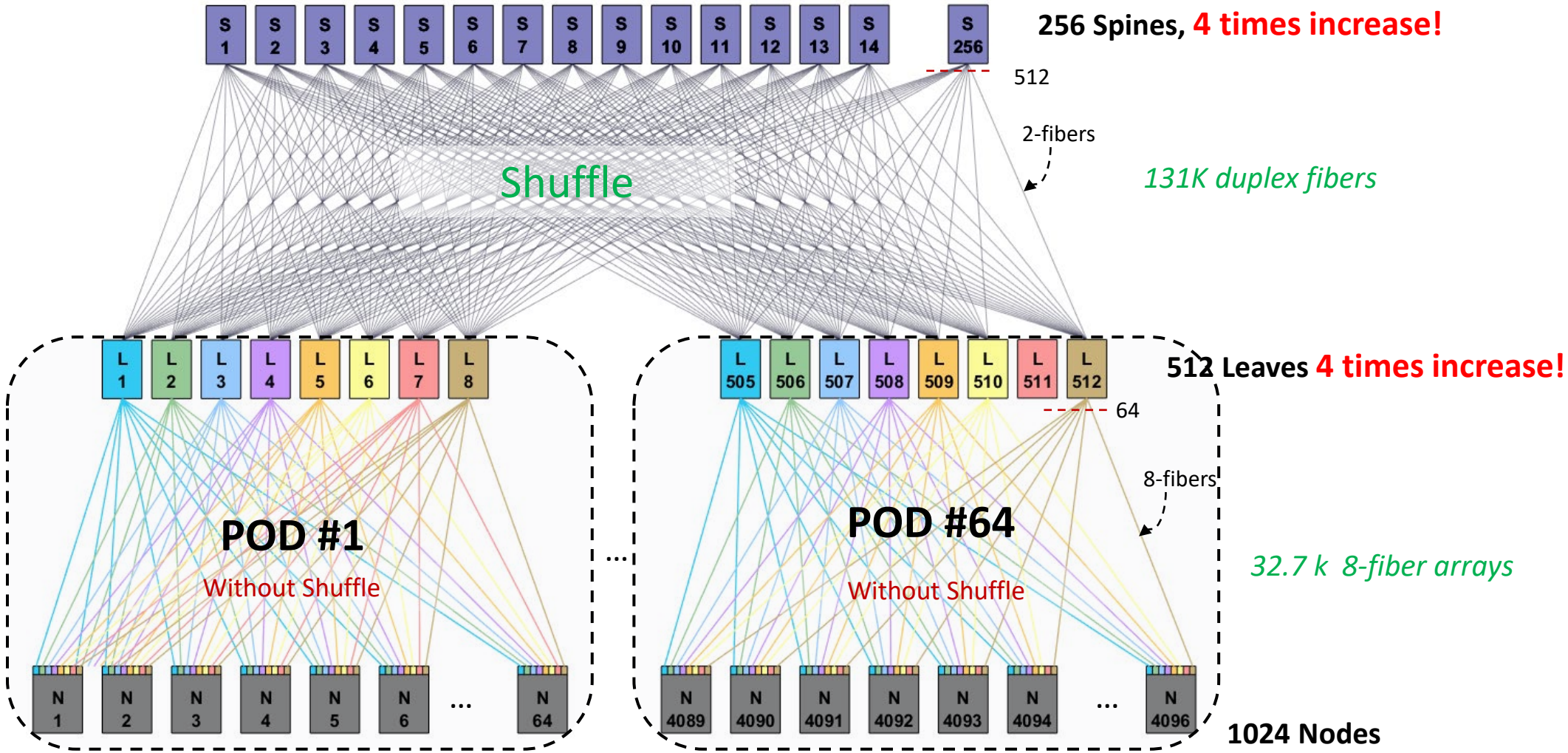
AI Cluster with 8192 GPUs

N: Node with 8 GPUs
L: Leaf , S: Spine
All switches have a radix of 512



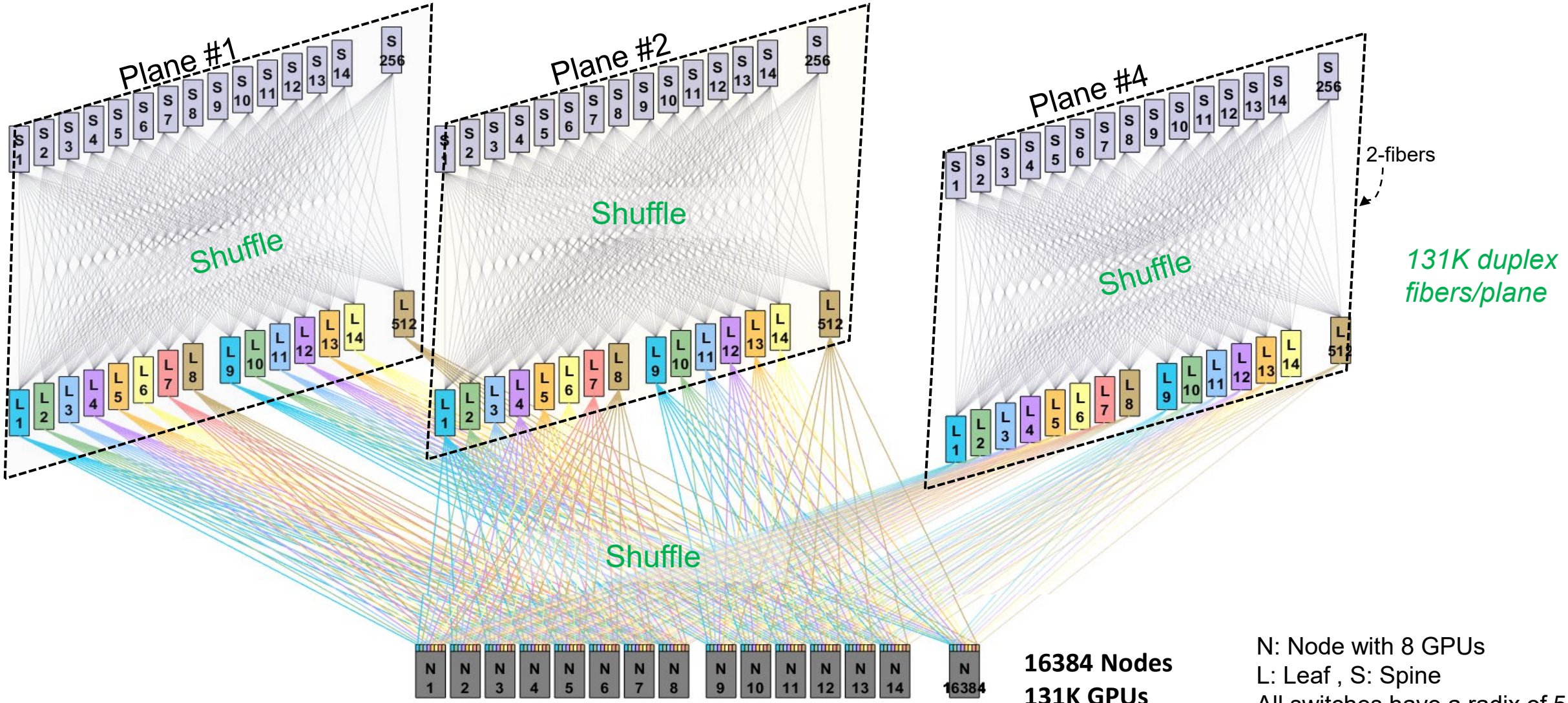
AI Cluster with 32768 GPUs

N: Node with 8 GPUs
L: Leaf , S: Spine
All switches have a radix of 512



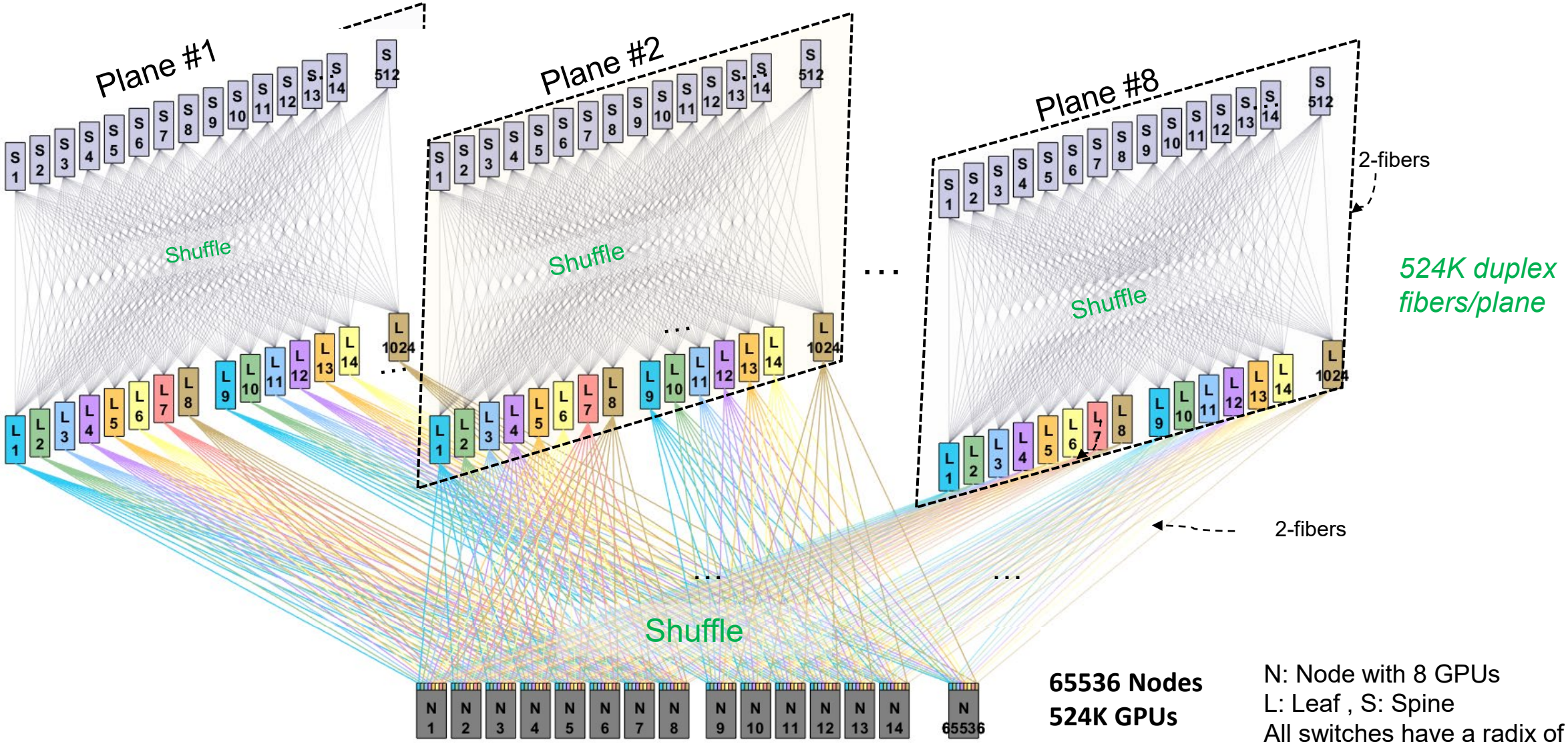
AI Cluster with 131K GPUs

Adding Node to Leaf Lane Breakouts increases the number of GPUs 4 times

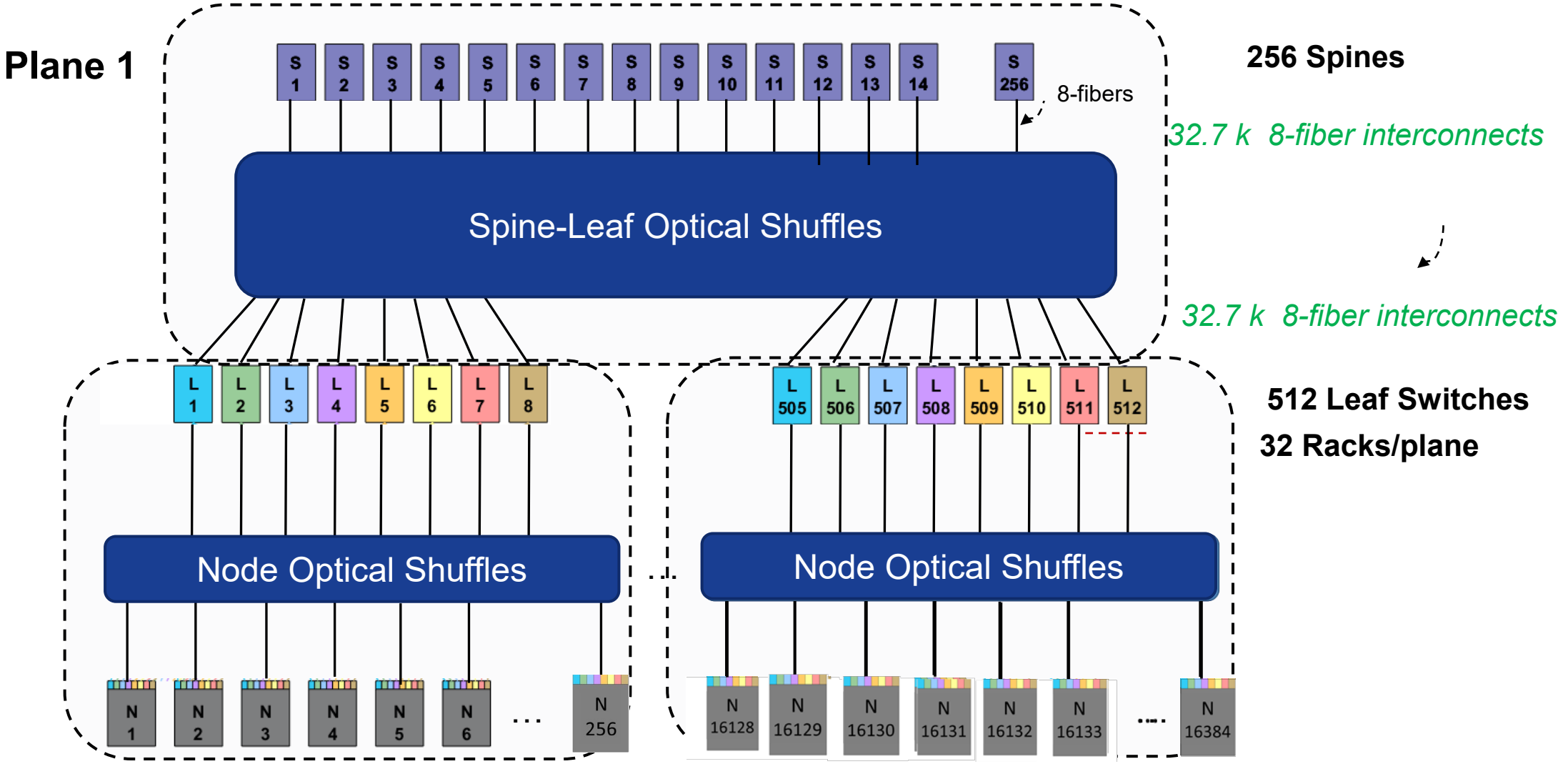


AI Cluster with 524K GPUs

Adding Node to Leaf Lane Breakouts increases the number of GPUs 4 times

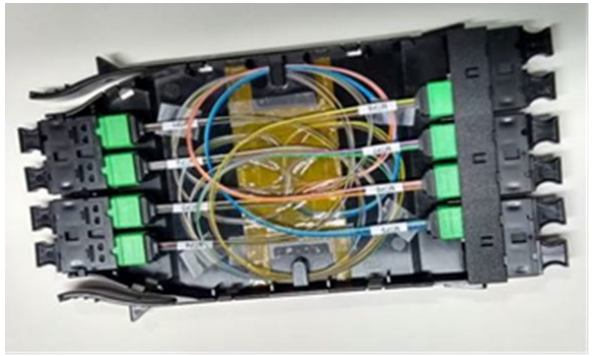


Shuffle Modules Reduce Deployment Complexity



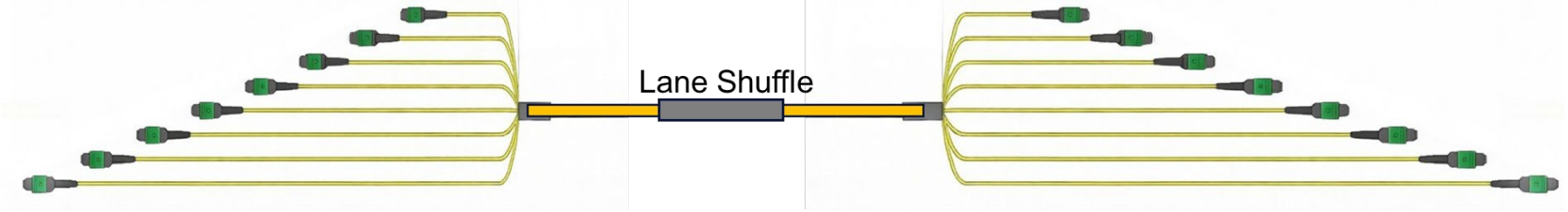
Optical Shuffle Implementation Options

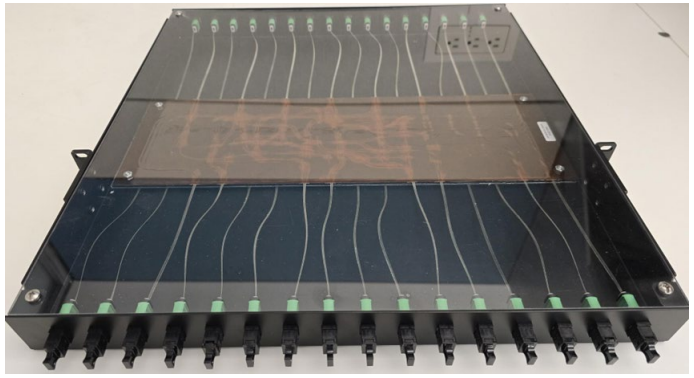
- Shuffle harnesses
  - —Integrated breakout + shuffle junction in the cables staggered legs to land on different switches/ports
  - —Saves rack space but can be harder to troubleshoot.
- Mesh cassettes (incl. small optical flex circuits)
  - —Small form factor; modular, easy to install and troubleshoot, but for larger networks, space use can be inefficient.
- Shuffle boxes / modules
  - —Higher density than cassettes; simplifies installation; relatively easy to replace if damaged.



Mesh Cassettes

Shuffle harnesses

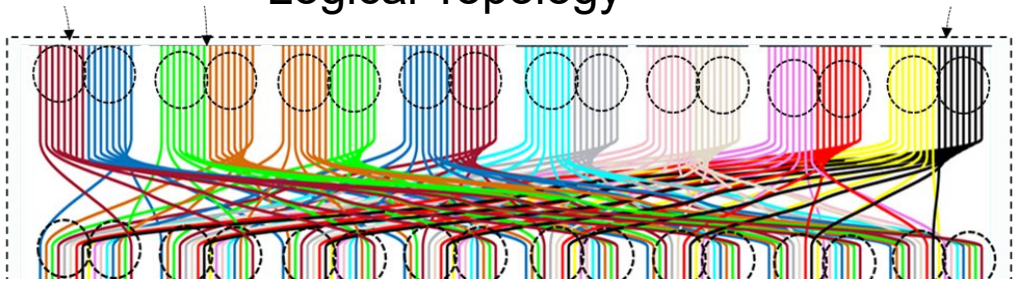




Shuffle boxes/Modules

Logical to Physical Implementation

Logical Topology



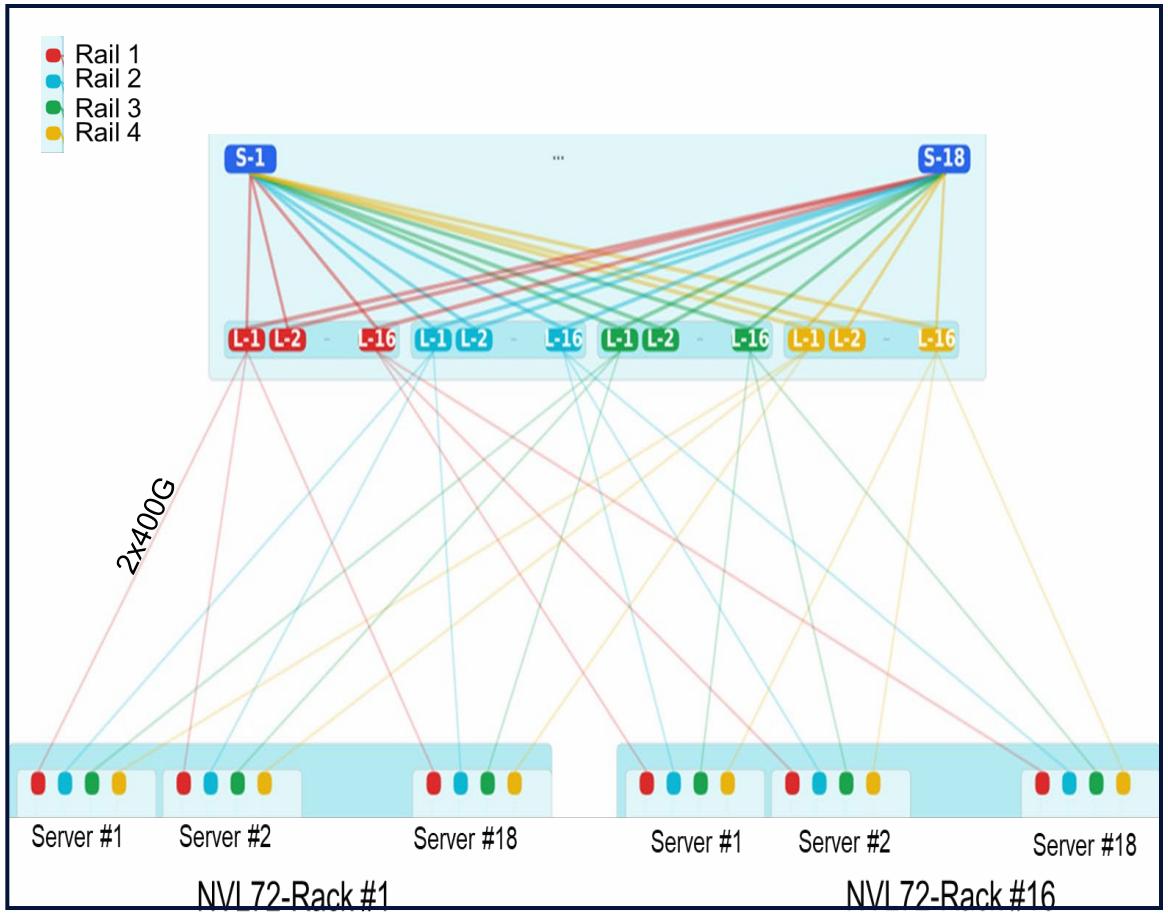
Manual or Automated (robotic)

Large Format Optical Flex Circuit

https://www.usconec.com/featured-products/ez-shuffle

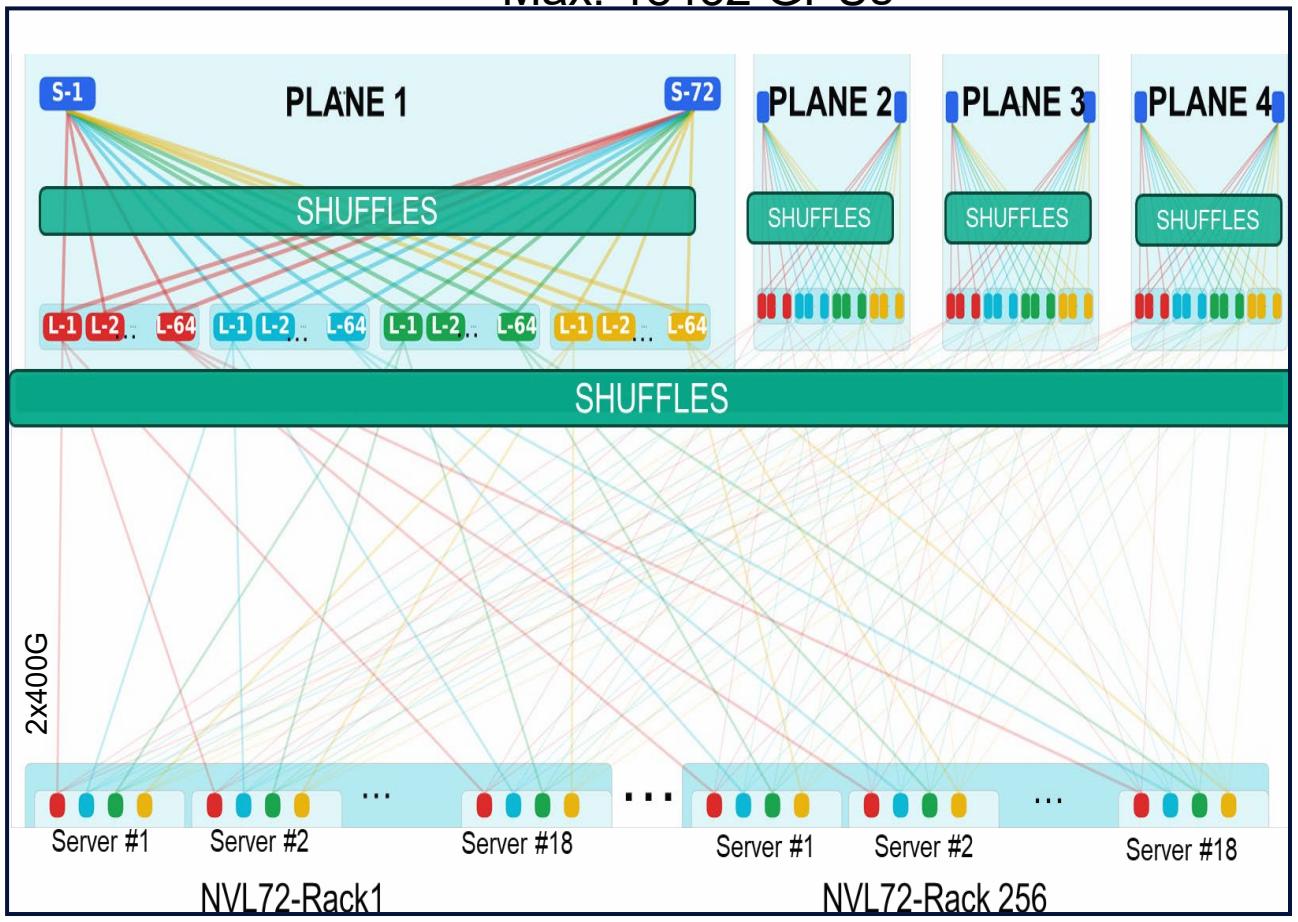
Shuffle-Based Scaling for NVL72 GB300

Without Shuffles:
Max: 1152 GPUs



64 Leaf + 18 Spines (SN5600),
NVL72 with ConnectX8

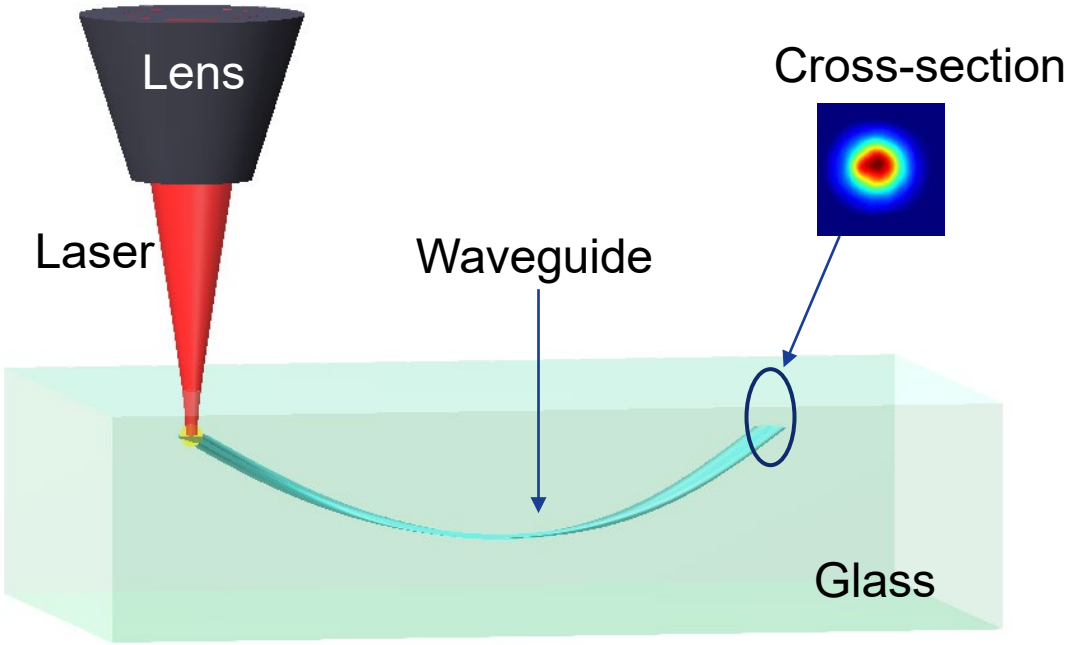
Using Shuffles:
Max: 18432 GPUs



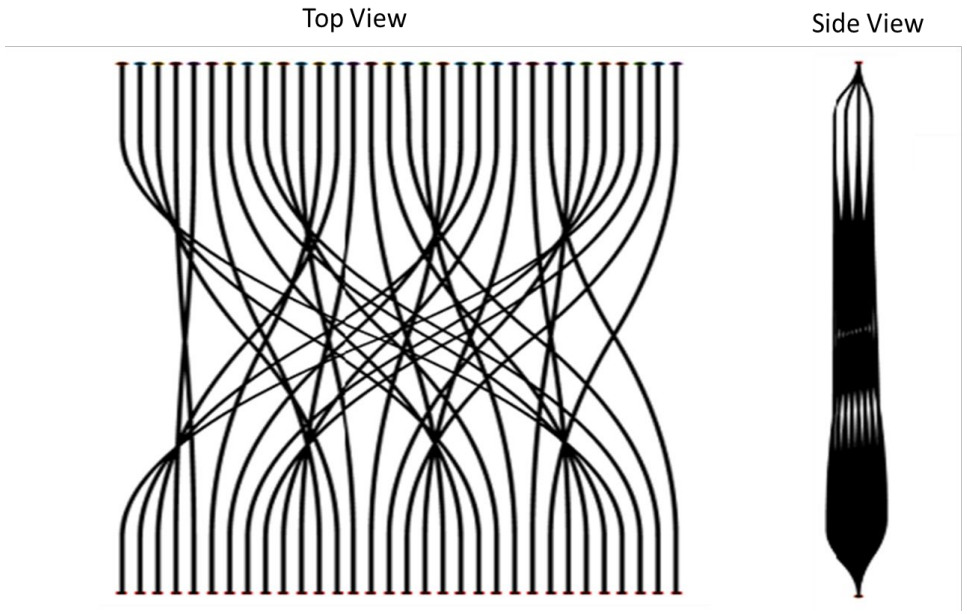
1024 Leaf + 288 Spines (SN5600),
NVL72 with ConnectX8

Novel Shuffle Technology: 3D Fabric Embedded in Glass

- Directly written 3D waveguides on glass minimize crossovers, crosstalk, and fabrication time by eliminating masking, exposure, and but produce higher losses than optical flex–based shuffle modules.

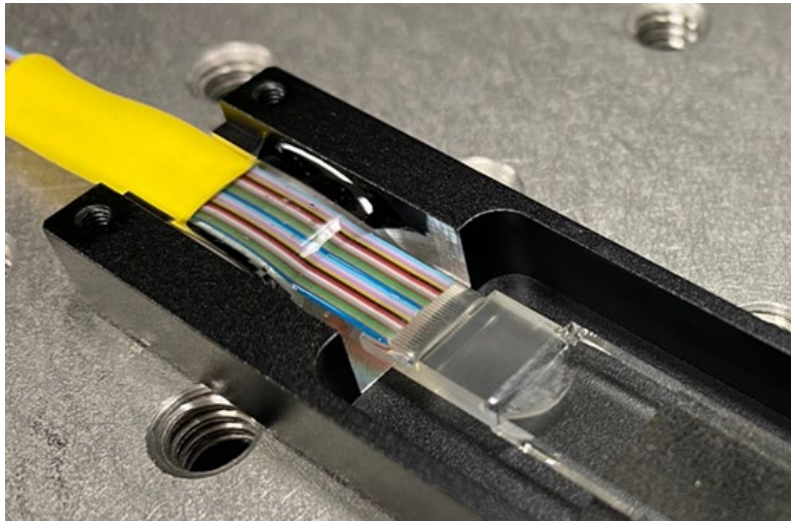


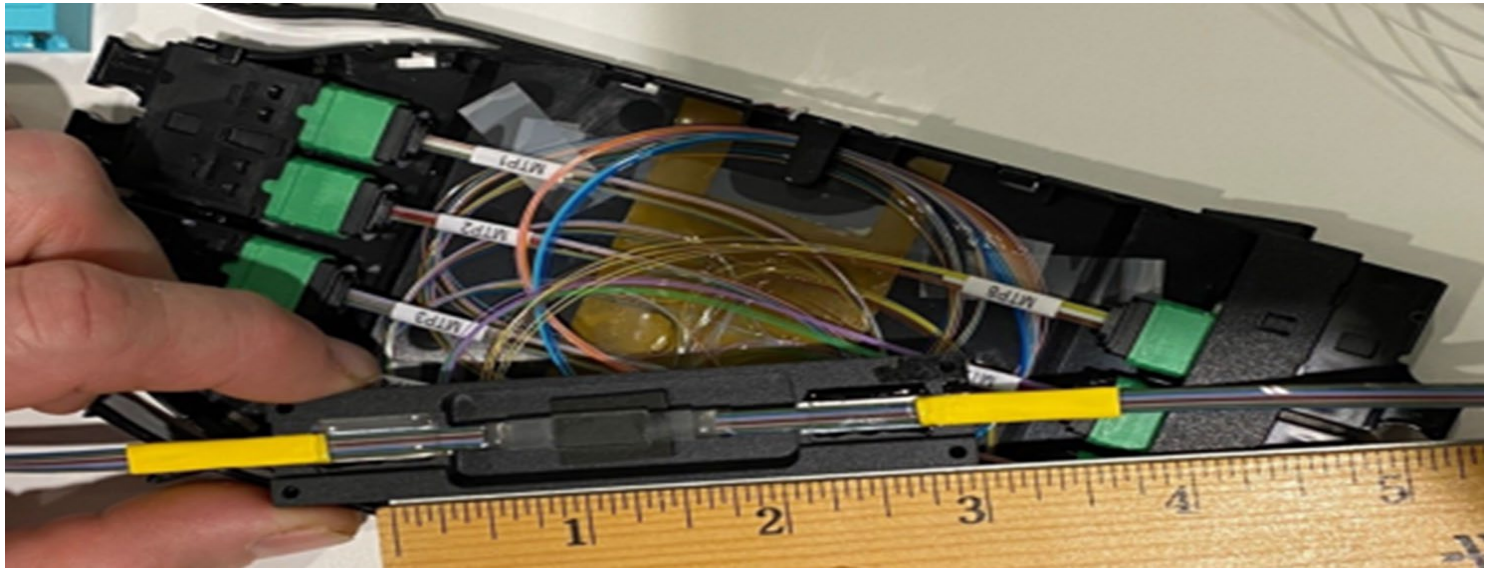
Fabrication Technology



3D Shuffle Design

Shuffle Technology: 3D Fabric embedded in Glass





-For additional information, please see J. Castro et. al., OFC 2025'Photonic Chip with Embedded Data Center Fabric," J. Castro et. al., OFC 2025.

Summary and Discussion

- Optical shuffles will enable AI to scale to a massive number of GPU.
  - —2-Layer fabric with several 100s of thousands GPUs.
- Simplification of deployment
  - —Enable direct well documented connections from node or switches ports
  - —Distribute the GPU signals to many switches increasing path diversity
- Reduce latency and provide path diversity
  - —Distribute the GPU signals to many switches improving entropy, useful for packet spraying.
- Reduce networking cost and power consumption
  - —More GPUs for a given allocated power.
  - —40% fewer switches, 33% fewer transceivers, for similar number of GPUs.

QUESTIONS?