Architectural Trends in Optical Interconnects for the AI Scale-Up Fabric

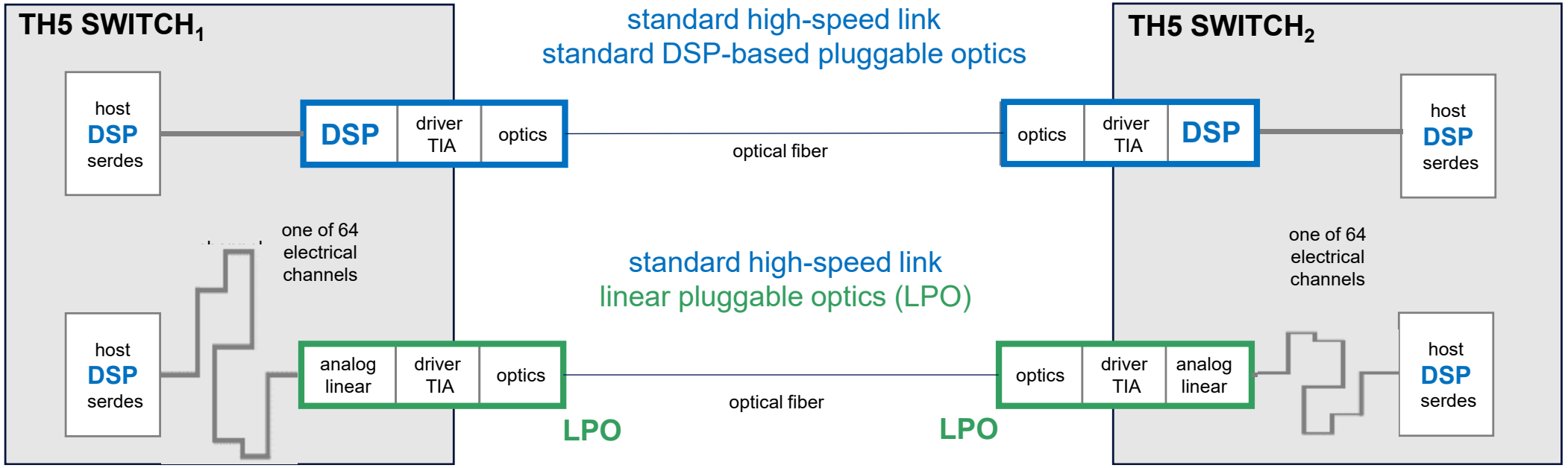
Mohammad Alavirad, PhD
Sr. Principal Research Architect
July 2026

DELL Technologies

Agenda

- Where are the LPOs?
  - LPO’s path to market starts at the NIC.
- CPO, NPO or LPO?
  - They are all linear
  - For AI fabric scale-up and scale-out, there exists a continuum of options.
- For AI scale-up and scale-out, which connectivity paradigms are breaking?
  - Weakening paradigms.
- Scaling-up the scale-up fabric and see what breaks
- Modern copper
- Transmission media bandwidth density
- The shuffle and implications for DACs, AOCs, ACCs, and AECs
- Modern optics
- Optical disaggregation
- 448G

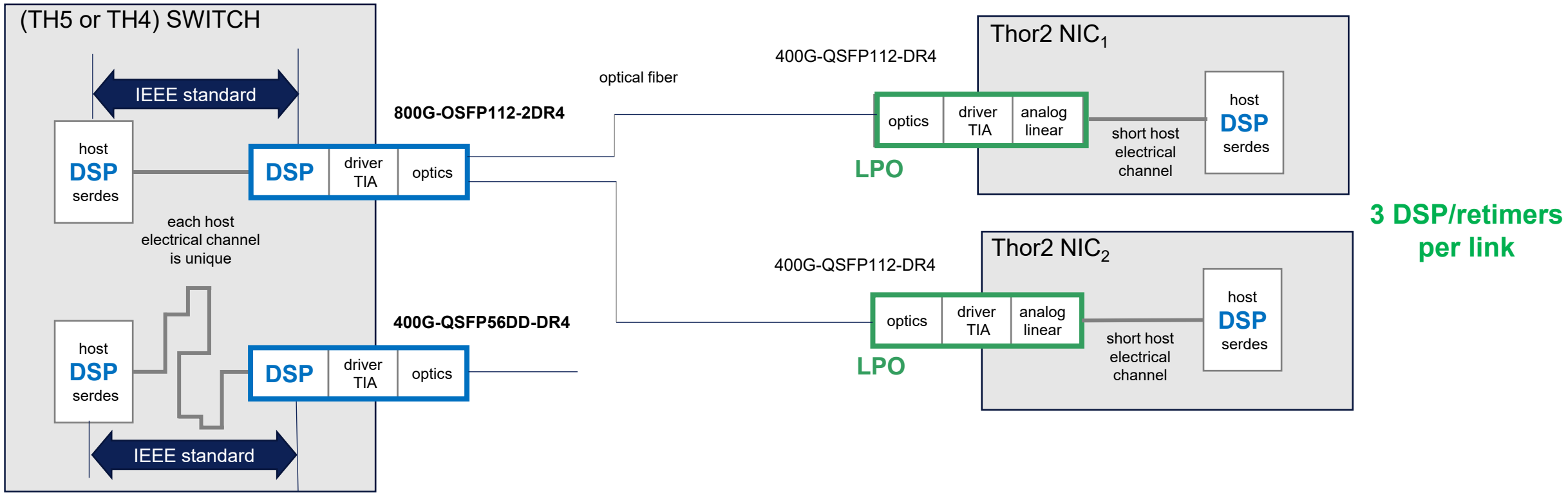
What are the linear pluggable optics (LPOs)?



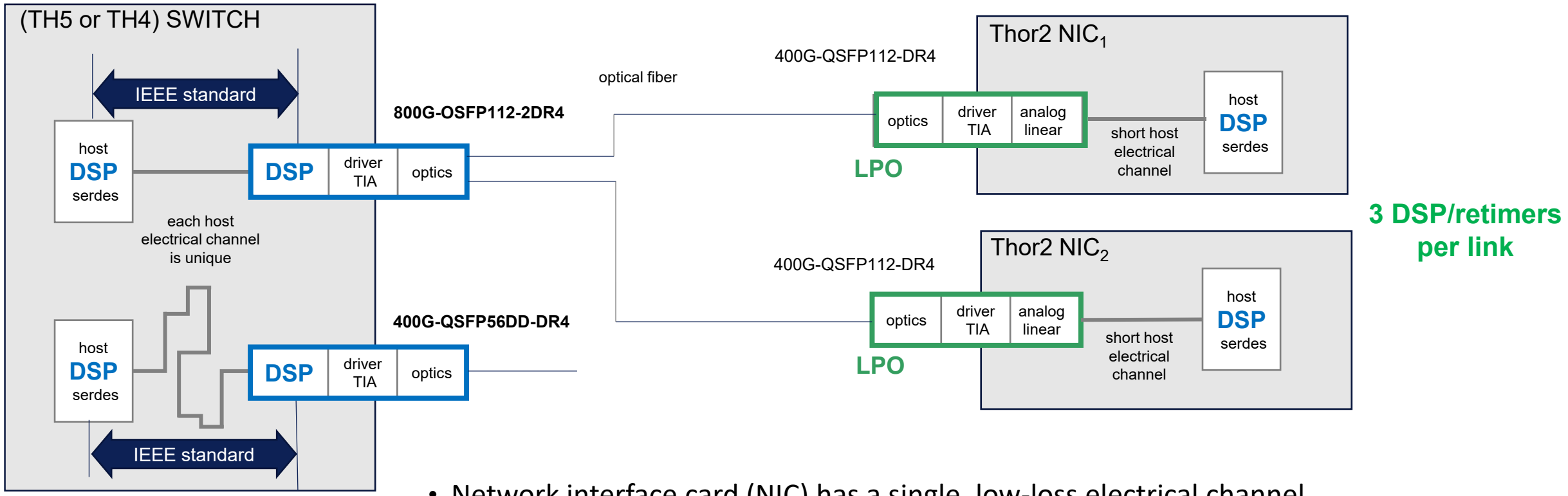
LPO = lower-power dissipation, lower-cost, lower latency, ***higher reliability***.

OFC 2023 – LPO *birthplace,* but questions were raised.

First mover – LPO in NIC only

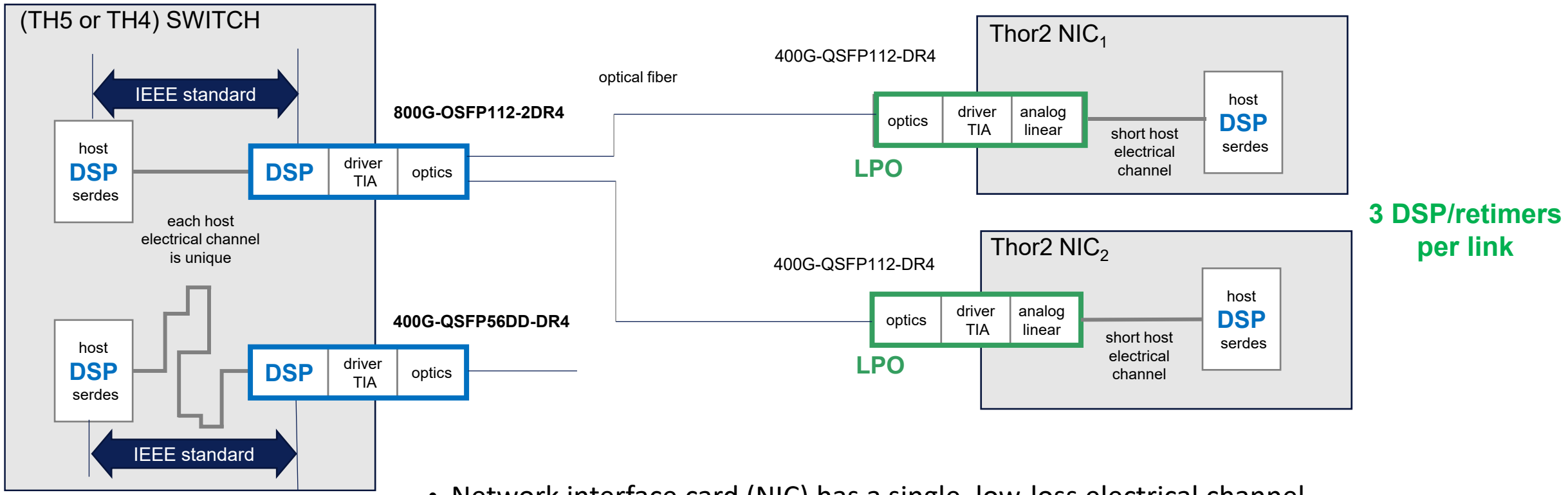


First mover – LPO in NIC only



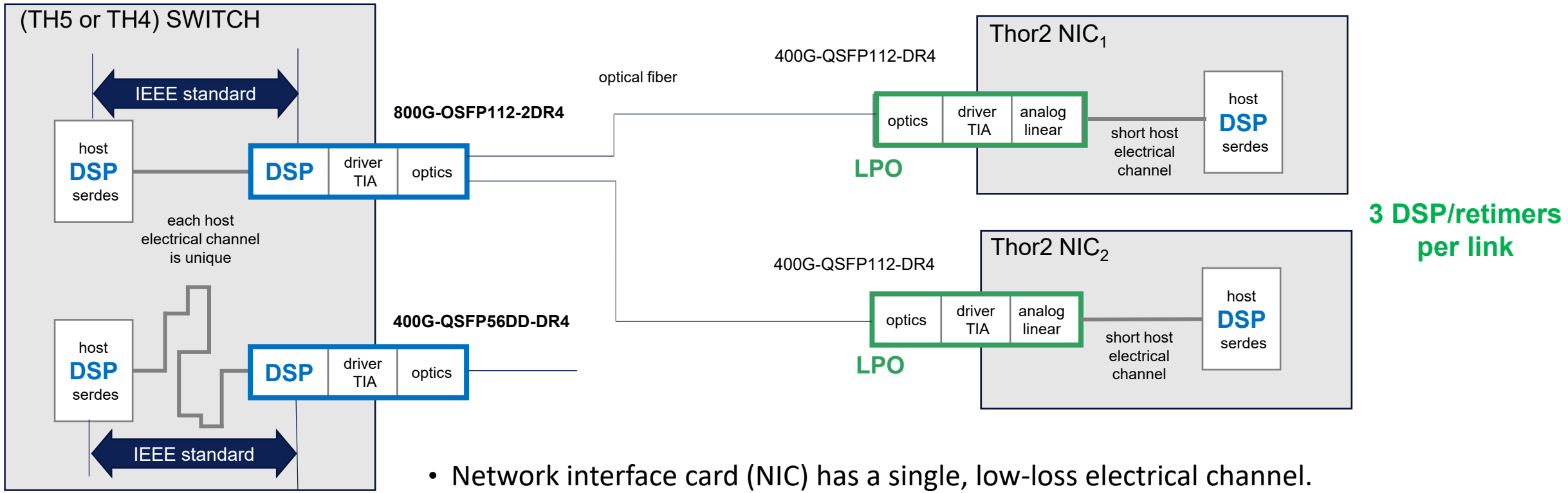
- Network interface card (NIC) has a single, low-loss electrical channel.
  - Retimed optical transceivers in switch make all switch ports look identical to the NIC.

First mover – LPO in NIC only



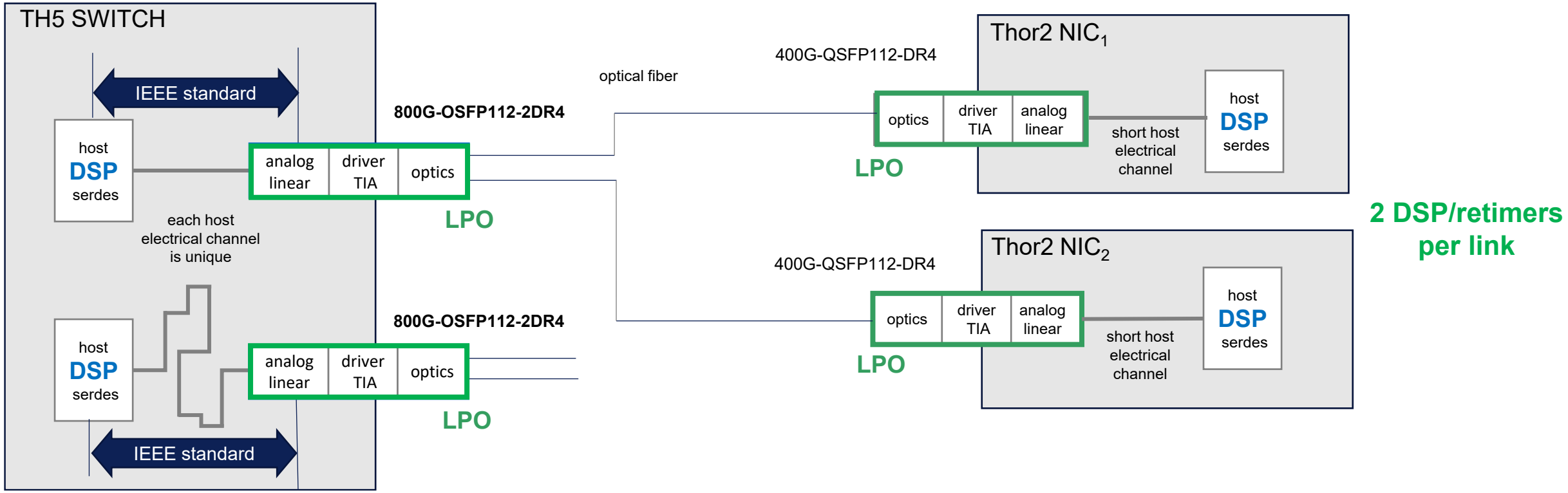
- Network interface card (NIC) has a single, low-loss electrical channel.
  - Retimed optical transceivers in switch make all switch ports look identical to the NIC.
- LPO at the NIC is a bigger *win* than LPO at the switch.
  - The NIC / server environment is more electrical-power, and thermally challenged that present switch environments.

First mover – LPO in NIC only

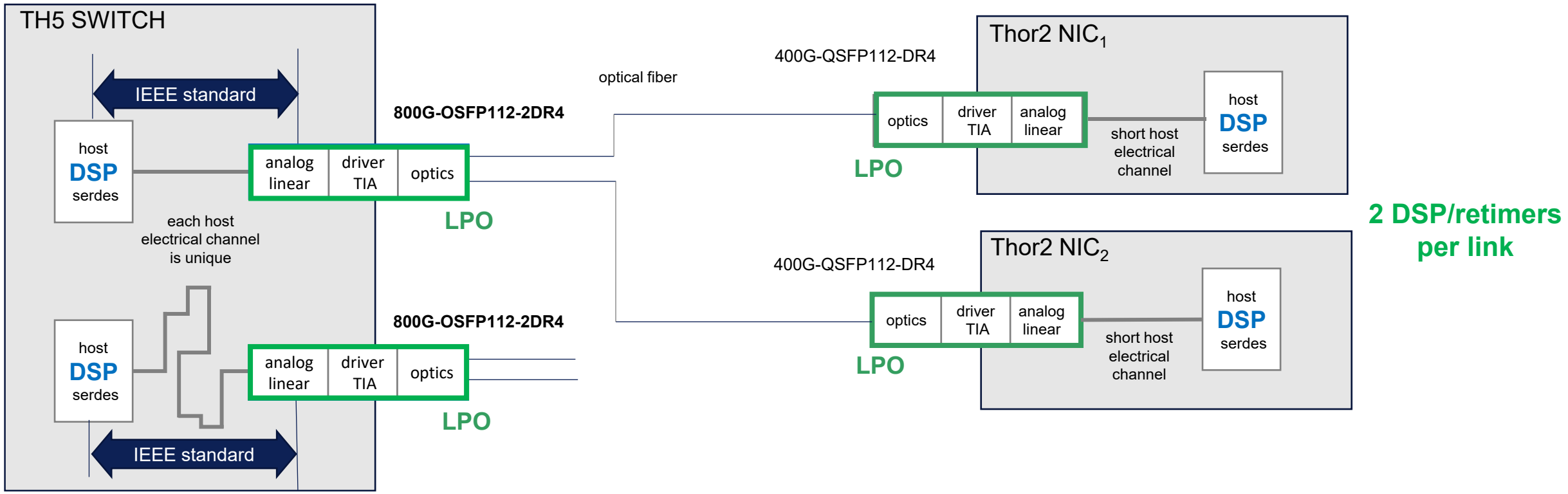


- Network interface card (NIC) has a single, low-loss electrical channel.
  - Retimed optical transceivers in switch make all switch ports look identical to the NIC.
- LPO at the NIC is a bigger *win* than LPO at the switch.
  - The NIC / server environment is more electrical-power, and thermally challenged that present switch environments.
- System vendor ensures the DSP-to-DSP link by bundling switch/NIC optics, NIC/server.

Next mover – LPO in ToR switch and NIC

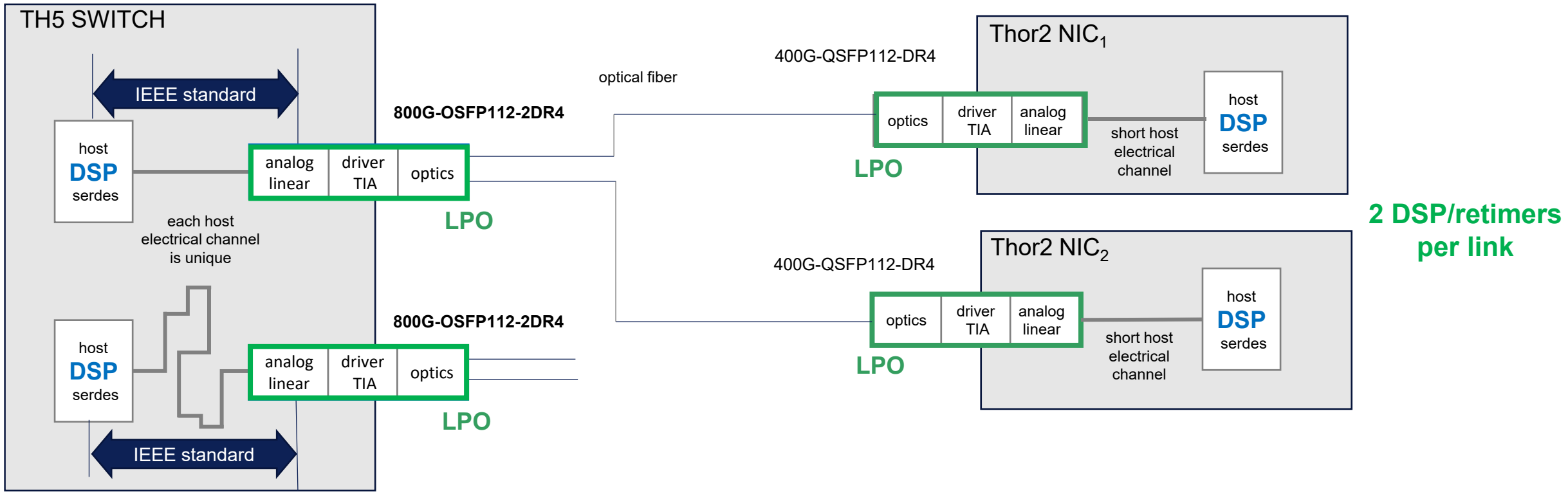


Next mover – LPO in ToR switch and NIC



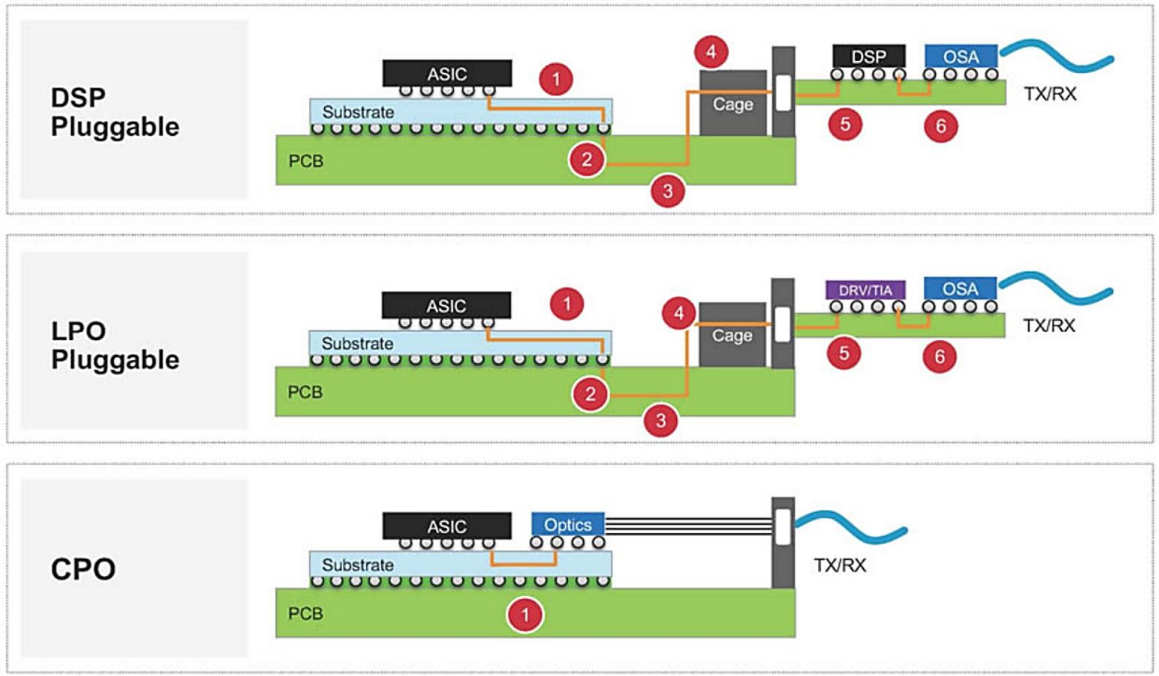
- System vendor ensures the DSP-to-DSP link by bundling switch, switch optics, NIC optics, NIC/server.

Next mover – LPO in ToR switch and NIC



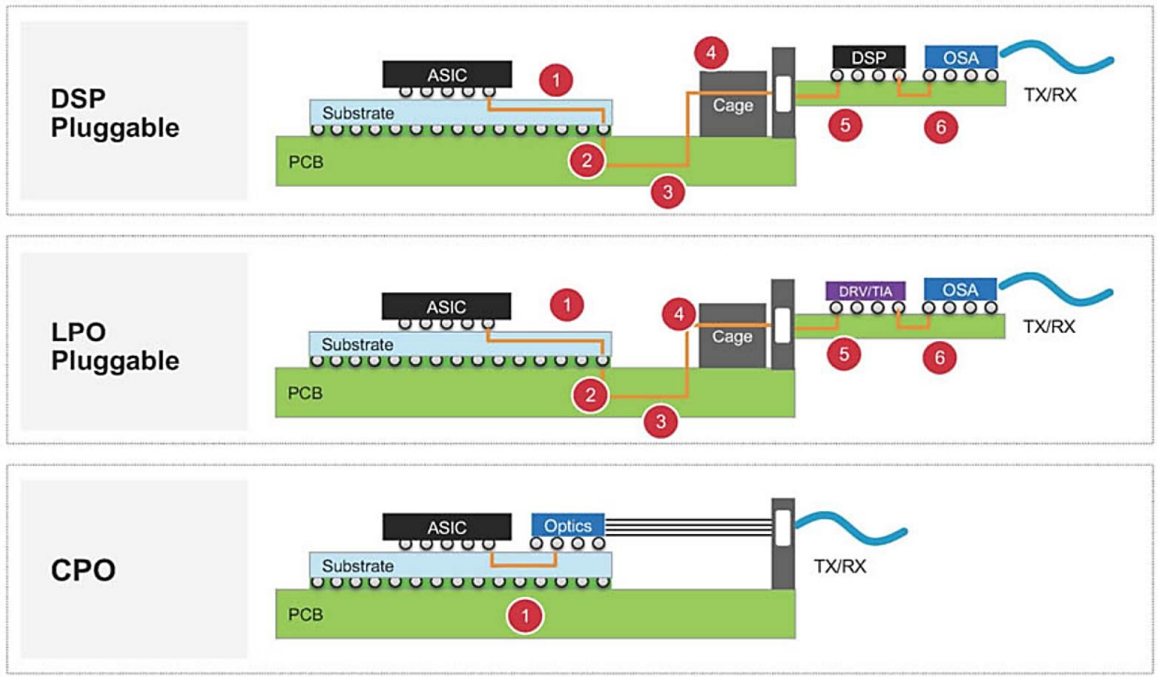
- System vendor ensures the DSP-to-DSP link by bundling switch, switch optics, NIC optics, NIC/server.
- Standards alone may not be enough for overall system optimization.

Weakening paradigm: CPO, NPO or LPO?



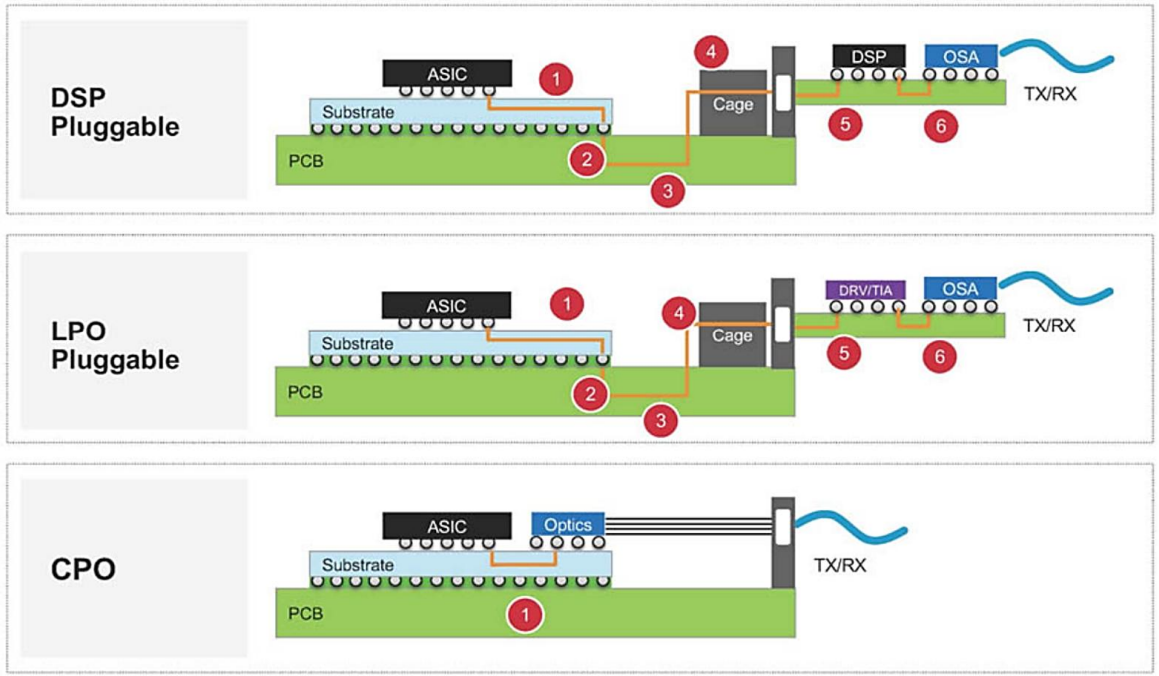
Weakening paradigm: CPO, NPO or LPO?

- This seems be the wrong question.
  - Connectivity solution space forms a continuum (in many dimensions)



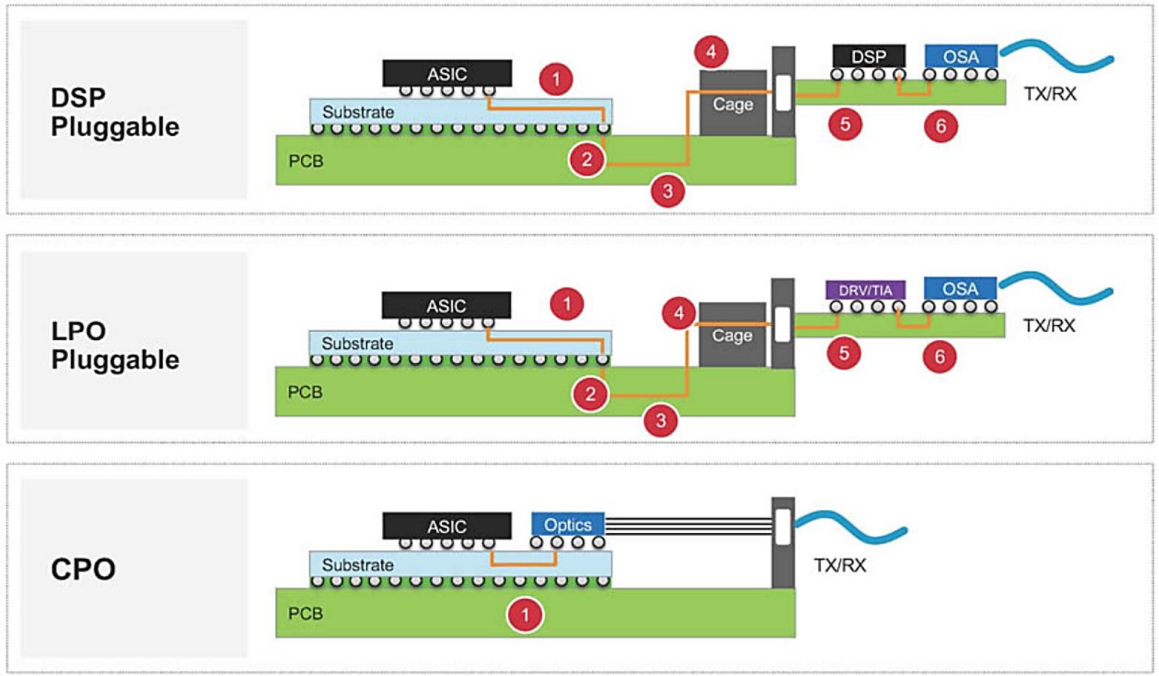
Weakening paradigm: CPO, NPO or LPO?

- This seems be the wrong question.
  - Connectivity solution space forms a continuum (in many dimensions)
- Industry focus on “standard” pizza box ToR switches has led to poor vision.



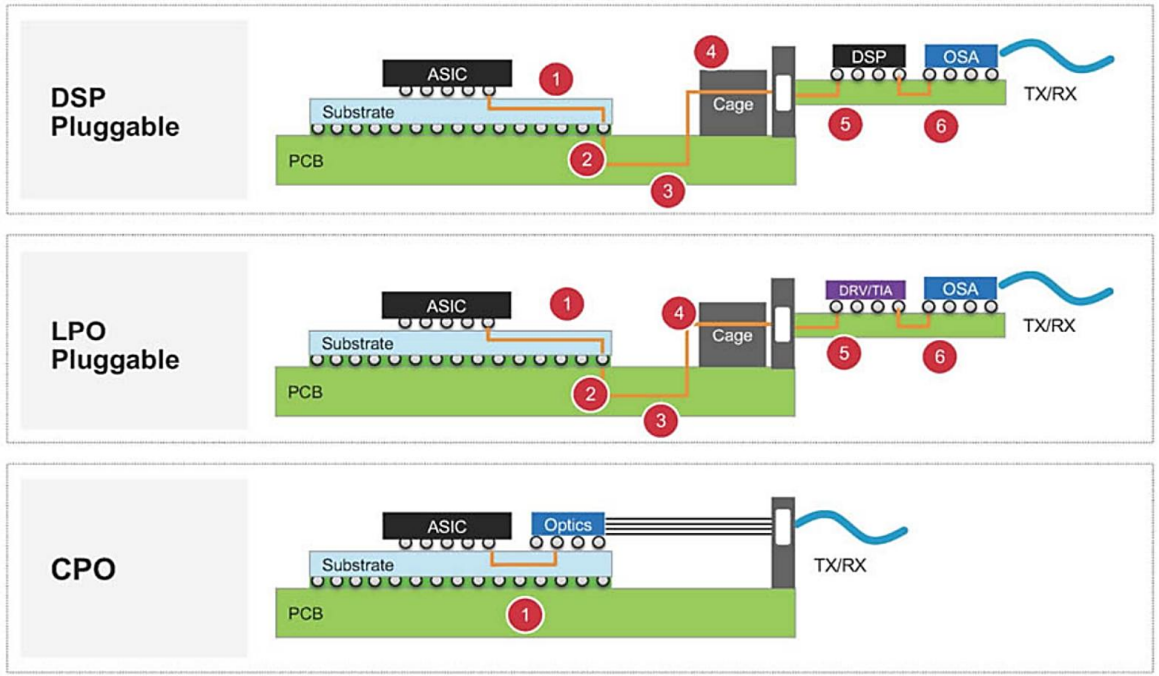
Weakening paradigm: CPO, NPO or LPO?

- This seems be the wrong question.
  - Connectivity solution space forms a continuum (in many dimensions)
- Industry focus on “standard” pizza box ToR switches has led to poor vision.
- Is LPO in the NIC really a
  - pluggable NPO?



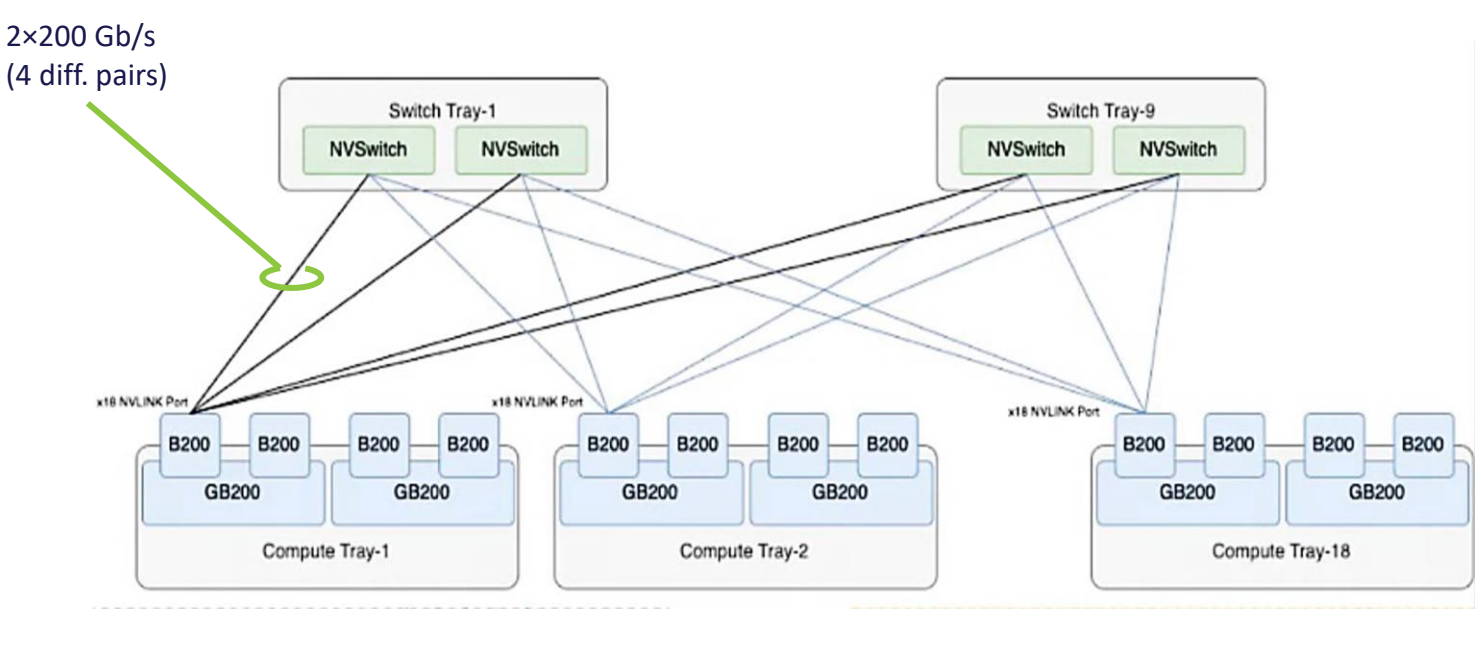
Weakening paradigm: CPO, NPO or LPO?

- This seems be the wrong question.
  - Connectivity solution space forms a continuum (in many dimensions)
- Industry focus on “standard” pizza box ToR switches has led to poor vision.
- Is LPO in the NIC really a
  - pluggable NPO?
- Is a NIC (or GPU) really a
  - Near face-plate DSP-based serdes looking for a pluggable NPO?



Scale-up (exemplary): NVL72

Every GPU is connected to another GPU through the switching layer
Each switch is connected to all GPUs
Each GPU is connected to all switches



18 switches
(144 × 200 Gb/s ports per switch)

5,184 diff. pairs =
5,184 cables per rack

72 GPUs
(36 × 200 Gb/s ports per GPU)

NVL72

Thought experiment: NVL72 (×2)

Every GPU is connected to another GPU through the switching layer

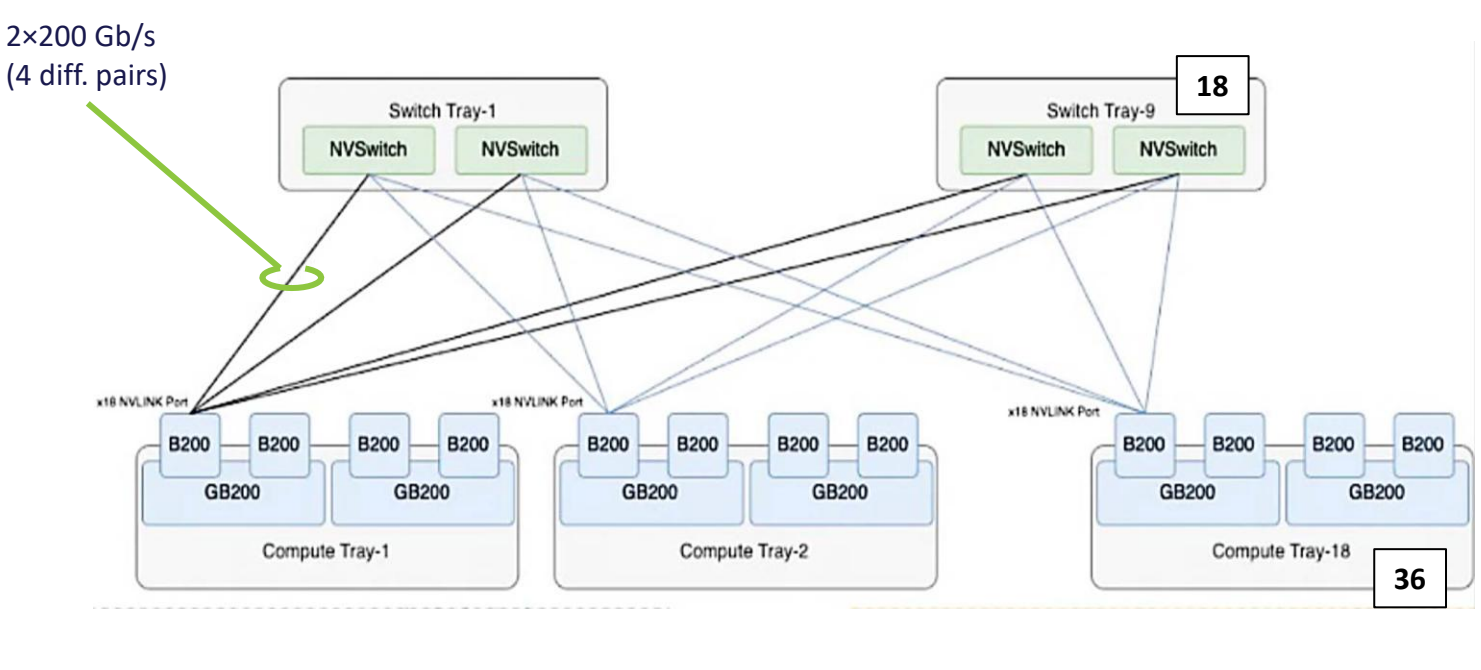
Each switch is connected to all GPUs

Each GPU is connected to all switches

Double number of GPUs
Double radix of each GPU
Double number of switches
Double radix of each switch

_____

Quadruple number of connections



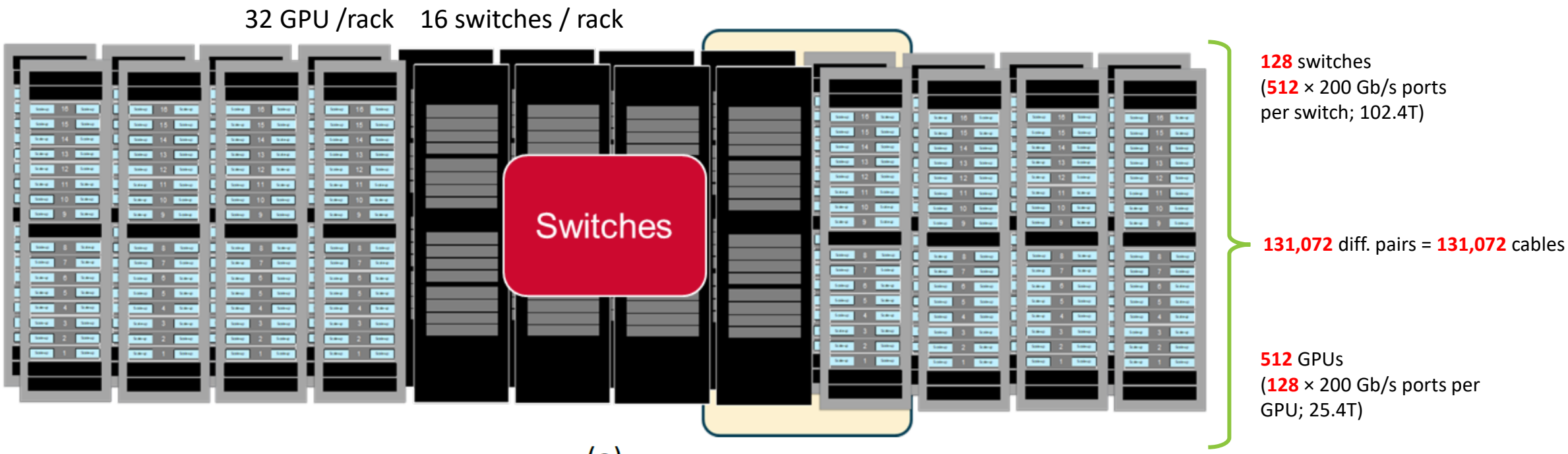
18 switches
(**288** × 200 Gb/s ports per switch)

20,736 diff. pairs = **20,736** cables per rack(s)

144 GPUs
(**72** × 200 Gb/s ports per GPU; 14.4T)

Thought experiment: 512 GPU cluster

and shifting from 9vidia math to base2 math



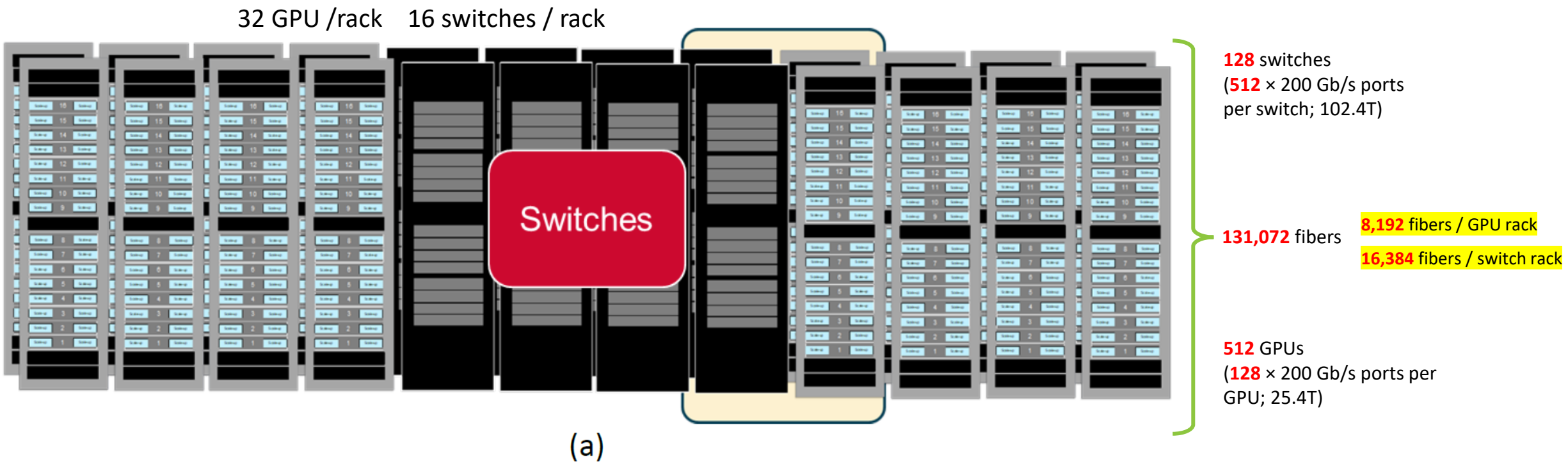
(a)

Figure 4-2 - (a) Physical Diagram of 512 All-to-All GPU Cluster with optical connectivity

From Optical Interconnecting Forum contribution **oif2025.304.05**: *Compute Optics Interface (COI): Energy Efficient Photonic Interconnects for AI Compute*

Thought experiment: 512 GPU cluster – optical fiber!

and shifting from 9vidia math to base2 math
+ 200GBASE-DR optics



(a)

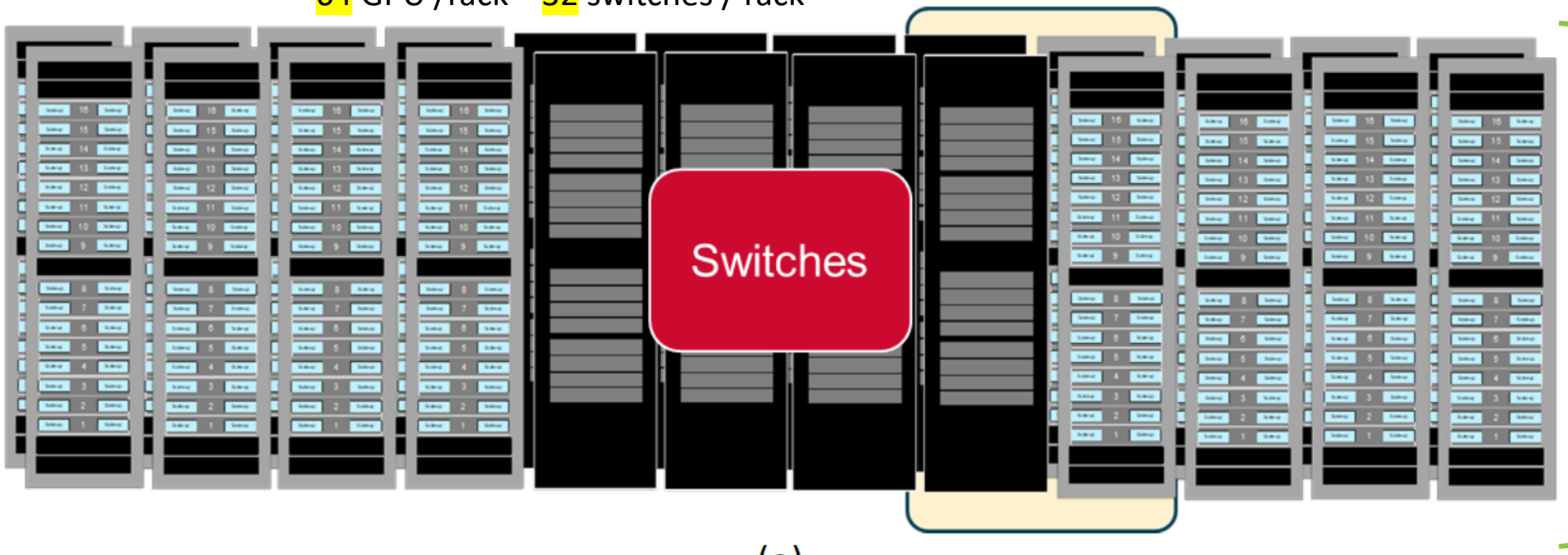
Figure 4-2 - (a) Physical Diagram of 512 All-to-All GPU Cluster with optical connectivity

From Optical Interconnecting Forum contribution **oif2025.304.05**: *Compute Optics Interface (COI): Energy Efficient Photonic Interconnects for AI Compute*

Thought experiment: 512 GPU cluster – optical fiber!

and shifting from 9vidia math to base2 math
+ 200GBASE-DR optics
+ (realistic) denser racks

64 GPU /rack 32 switches / rack



128 switches (**512** × 200 Gb/s ports per switch; 102.4T)

131,072 fibers

16,384 fibers / GPU rack

32,768 fibers / switch rack

512 GPUs (**128** × 200 Gb/s ports per GPU; 25.4T)

(a)

Figure 4-2 - (a) Physical Diagram of 512 All-to-All GPU Cluster with optical connectivity

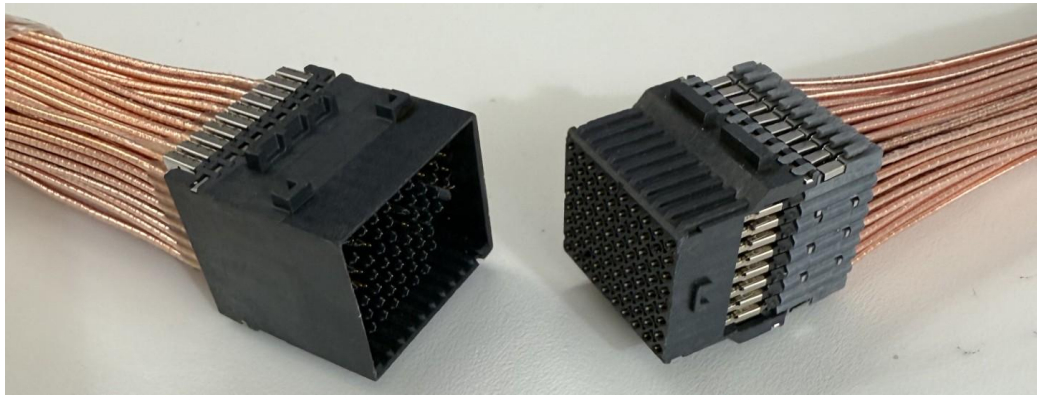
From Optical Interconnecting Forum contribution **oif2025.304.05**: *Compute Optics Interface (COI): Energy Efficient Photonic Interconnects for AI Compute*

What got hard in the thought experiment?

- 200 Gb/s passive copper limited to ~ 1 meter.
  - *Use modern copper tools*
- All-to-all connectivity.
  - *Preconfigured* cable management tool = ***shuffle***
- 10K+ fibers per rack.
  - Raise *bandwidth density* along the data path

Modern copper

Passive copper



Amphenol

Linear signal regeneration

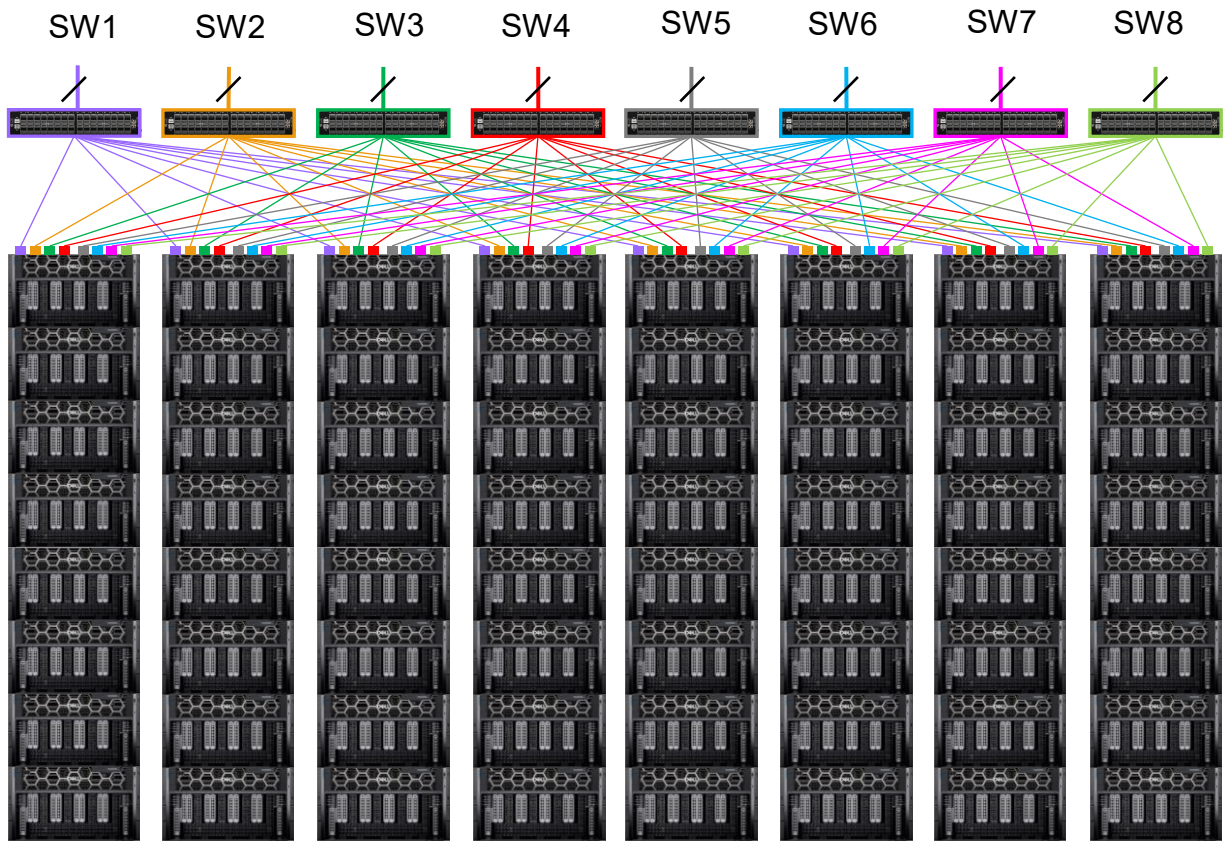
- extends copper reach (@ 200G to ~ 3-4 m)
- low-cost (not a DSP-based retimer)
- low-power (not a DSP-based retimer)
  - ~500 mW / [200 Gb/s lane]
- compact
- smart (improved ease of use)
- linear active copper cables
- on-board applications

SEMTECH MACOM nubis COMMUNICATIONS ciena

Shuffle – managing all-to-all connectivity at scale
AI-fabric scale-up

Exemplary rail-based AI-fabric scale up

Each rack has 64 NICs connecting to the eight switches



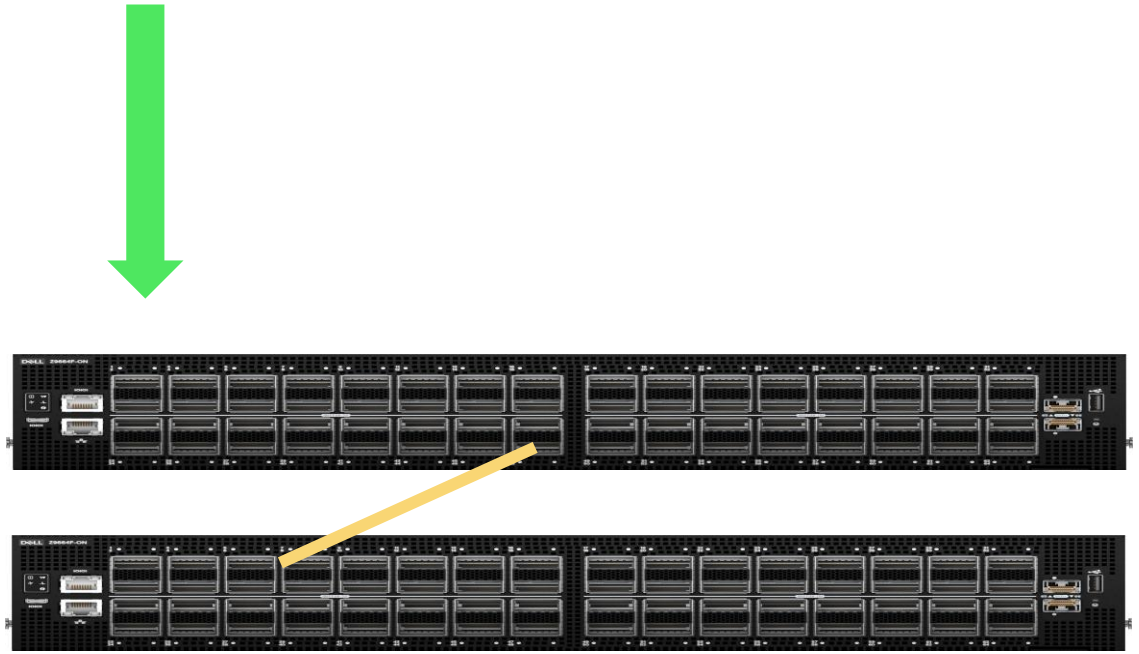
- In each rack, there are eight server shelves, each with eight NICs connecting to eight switches. Each of the eight NIC/GPUs in the shelf must connect to a different switch.
- It is impossible to do this with all passive copper due to NIC / switch distances.
- Except for a few corner cases, all NIC ⇔ switch connections need to be optical

Shuffle – managing all-to-all connectivity at scale
Building networks

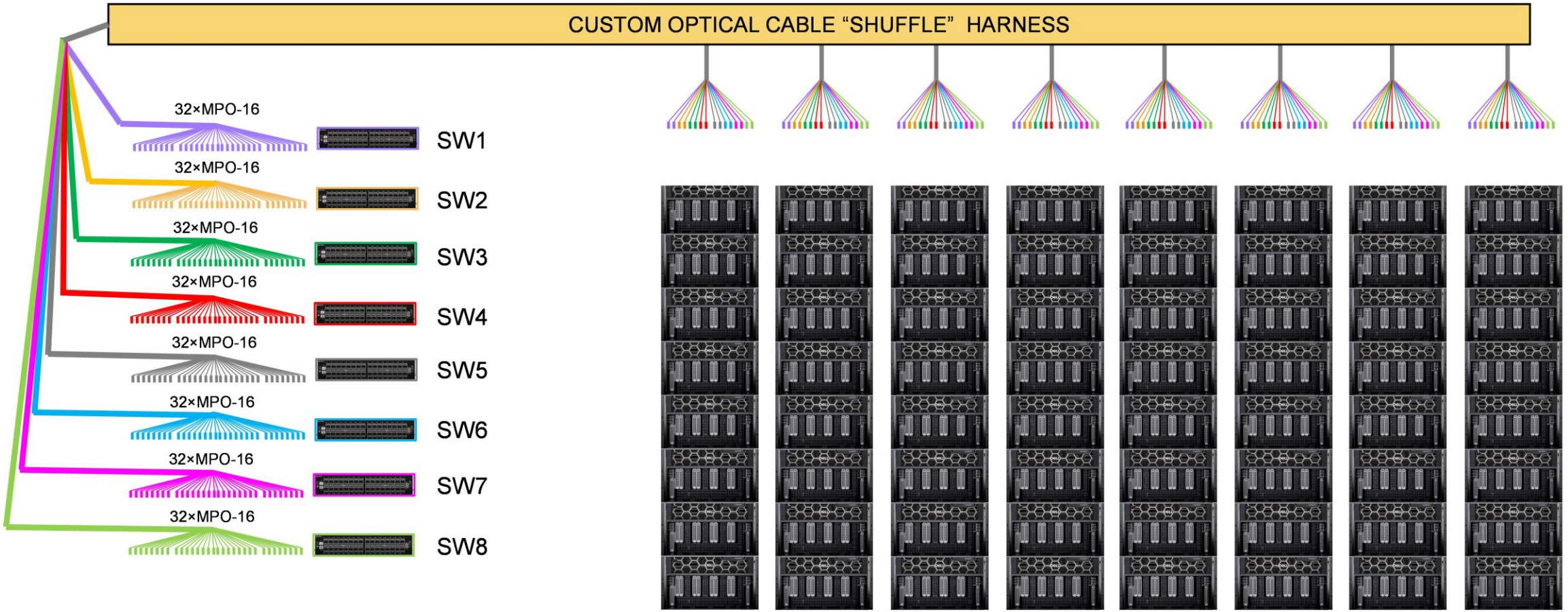
Traditional approach

point-to-point connections added one at a time.

manual cable routing one cable at a time

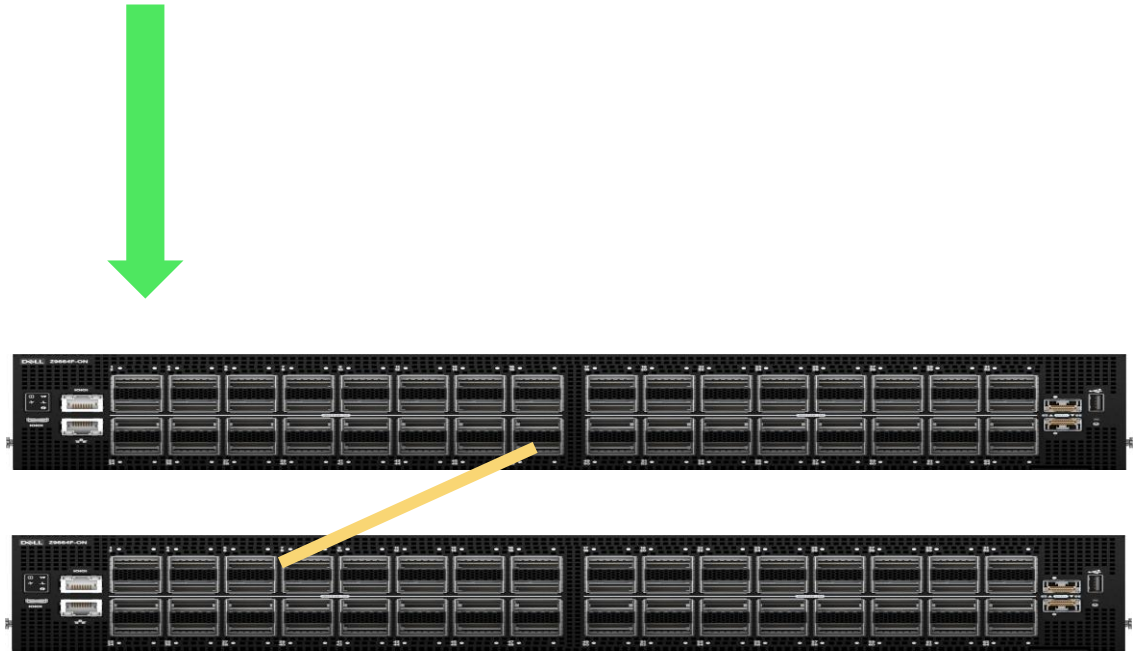


Shuffle – managing all-to-all connectivity at scale
Shuffle in action for rail-based scale-up



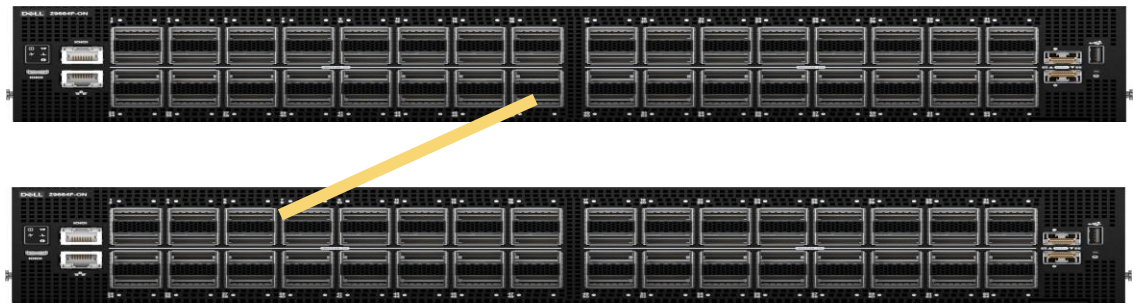
Shuffle – obsoletes traditional DAC, AOC, AEC, ACC
Building networks

Traditional approach
point-to-point connections added one at a time.
manual cable routing one cable at a time



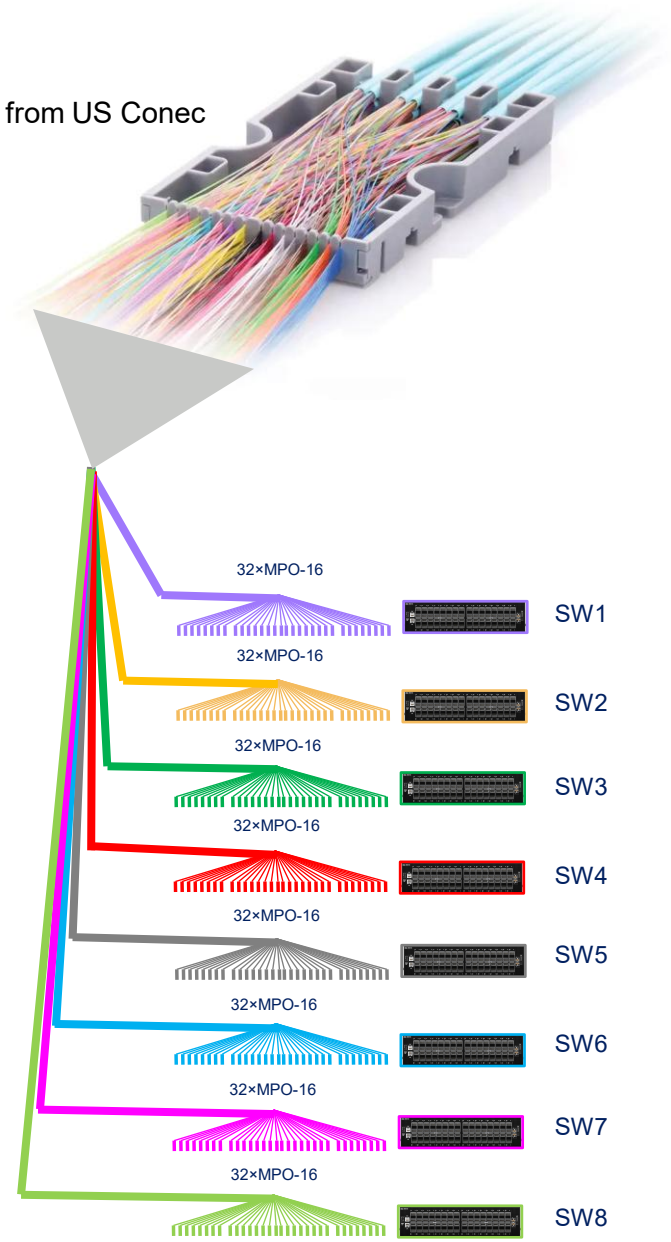
Shuffle – obsoletes traditional DAC, AOC, AEC, ACC
Building networks

Traditional approach
point-to-point connections added one at a time.
manual cable routing one cable at a time



Rack-scale AI fabric
build everything all at once

EZ Shuffle™ from US Conec

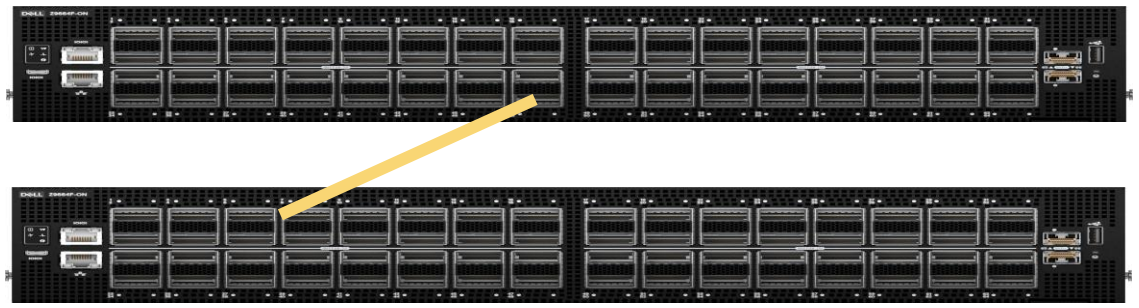


connectivity managed via pre-configured structured cabling cassette

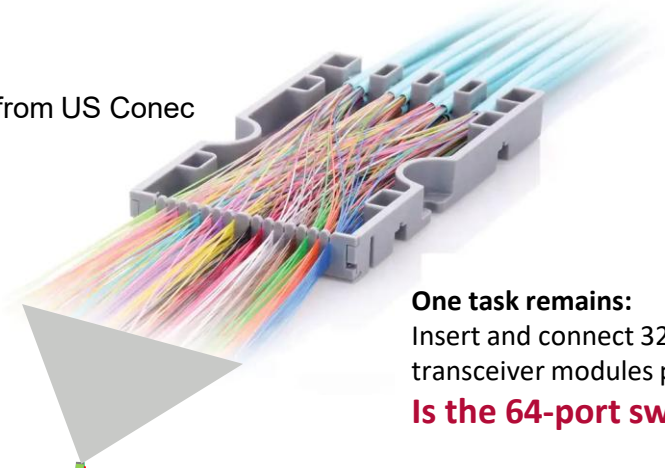
Shuffle – obsoletes traditional DAC, AOC, AEC, ACC
Building networks

Traditional approach
point-to-point connections added one at a time.
manual cable routing one cable at a time

Rack-scale AI fabric
build everything all at once



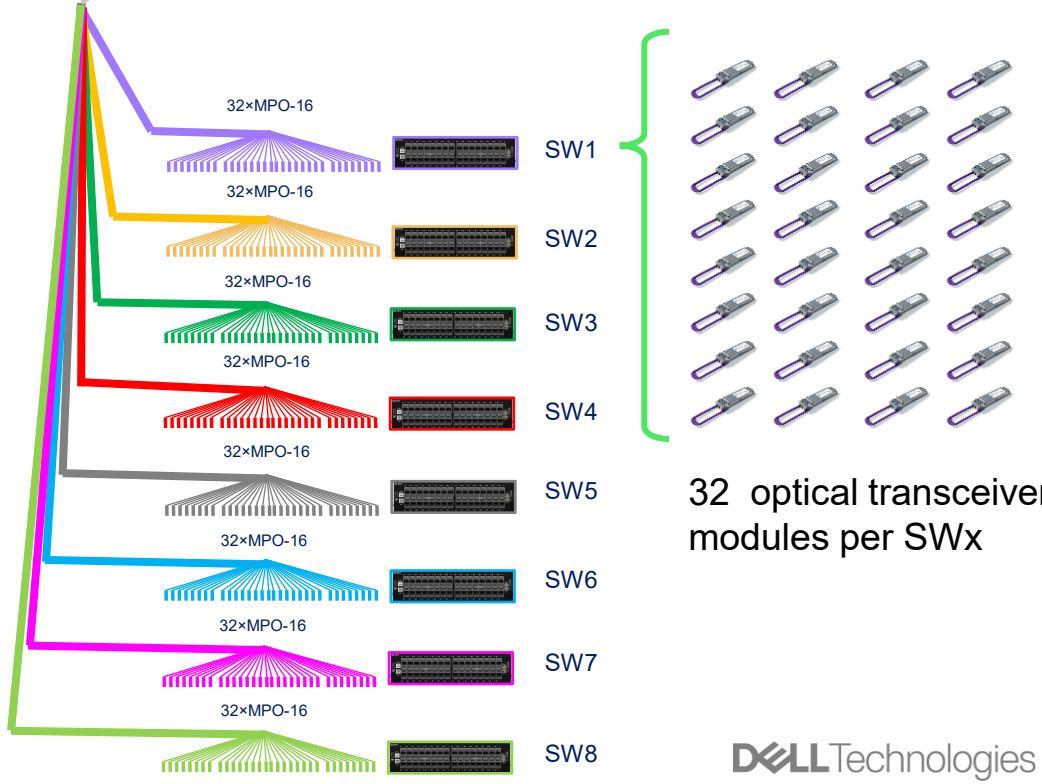
EZ Shuffle™ from US Conec



connectivity managed via pre-configured structured cabling cassette

One task remains:
Insert and connect 32 separate pluggable optical transceiver modules per switch

Is the 64-port switch paradigm useful

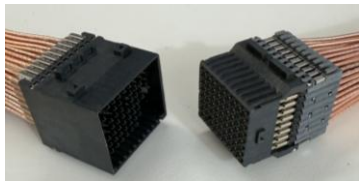


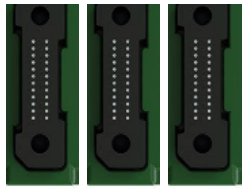
32 optical transceiver modules per SWx

Weakening paradigm: the 32-, 64-. 128-port switch?

- Segmenting bandwidth into 8-lane physical ports seems arbitrary for AI-fabrics.
- MAC speed >> serdes speed is not necessarily useful.
- Certain AI-fabric applications really want a single multi-lane connector from a high-radix switch.
- Does it make sense for pluggable modules to be used for both passive copper and active optics?
- Do pluggable modules make sense, and if so, should they have 32, 64, 128, 256 or more lanes?

Bandwidth density matters at the front panel





	OSFP & OSFP XD	Paladin HD2	MPO-32	(3×) MMC-32
Media	Optical or Electrical	Electrical (passive)	Optical	Optical
Diff pairs/fibers	8×2/16×2 (Tx/Rx)	Variable 4/8 dp high, 2-12 columns	16x2× fiber (Tx/Rx)	16×2 (Tx/Rx)
Density	22.3/11.1 mm^2/dp	8.5 mm^2/dp	3.8 mm^2/dp	1.3 mm^2/dp
Faceplate dimension	24.3×17.2 mm	22.6×28.7 mm (8×10 dp)	12.8×9.4 mm	4.2×9.6 mm

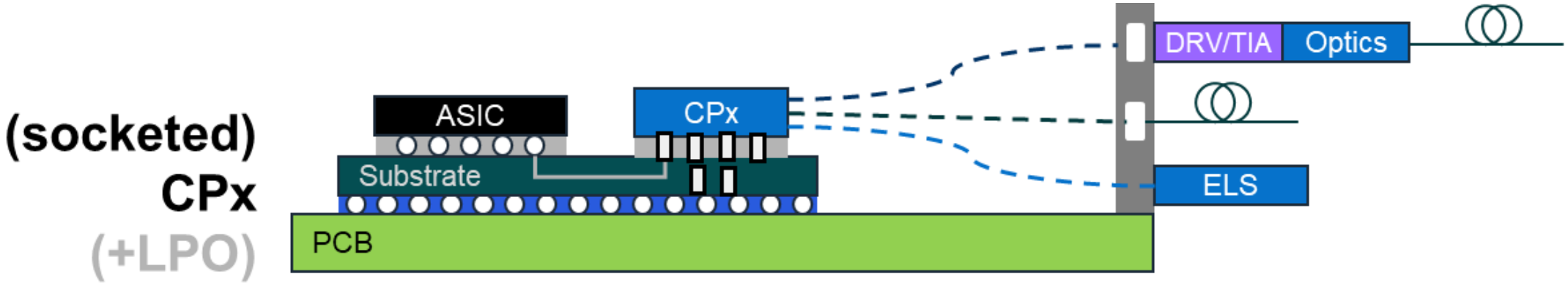
***Note**: 64-lane pluggables upcoming*

Modern optics

- Continuum of solutions from pluggables to 3D co-packaged optics (CPO)
  - Silicon photonics
  - Indium phosphide
  - Thin film electro optical crystals
    - lithium niobate, barium titanate
  - VCSELs
  - µLEDs
- Wide and slow *vs.* narrow and fast

Optical disaggregation – CPx

- The community will benefit from the disaggregation of optics from xPU ASICs.



448G

- Optimal modulation format for 448G over copper (PAM6/8) is not optimal for 448G over optical fiber.
- Move to 448G serdes will drive optics into the AI scale-up.

AI Fabric Domains

- Physical connectivity and connectivity packaging is not one-size-fits all
- Different system and solution designs may require very different solutions

Focus

	Scale In	Scale Up	Scale Out	Scale Across
Physical Reach	1mm → 1cm → 1m In-package or chip-chip on system board	1m → 10m Within chassis/rack/row (current NVL72 is 1 rack: <2m)	~10m → 2km Within a single DC Hall	100m → 100 km Across DC halls Metro/WAN Cluster
Latency	ns → tens of ns	100ns → few µs	few µs → 100s µs (depends on topology)	100µs → 10ms
Bandwidth	~ 10s TB/s per chip	1.8 TB – 3.6 TB per GPU	400Gb - 1.6Tb per port	100Gb – 1.6Tb per port
Scale Comments	Very high bandwidth density; limited by die edge & package routing	All-to-all connectivity @ rack	Huge bisection bandwidth Massive aggregate capacity	Lower bandwidth density May connect subset of GPUs
Power/bit	~0.5 pJ/bit	~5 pJ/bit	~ 15-20 pJ/bit	> 20 pJ/bit
Cost/(Gb/s)	$0.005/(Gb/s)	$0.05 / (Gb/s)	$0.45 / (Gb/s)	$1-10 / (Gb/s)
Optical Role	Optics usually **NOT** required	Copper Primarily *optics and alternatives increasingly attractive as speed increases*	Pluggable Optics Pluggable Copper sometimes	Pluggable Optics Coherent Pluggable Optics DWDM for longer distances

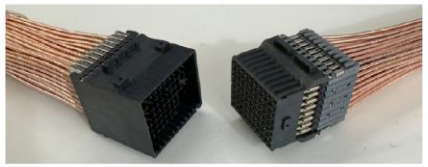
Note: WAN connectivity is a separate consideration

Problems with current Copper and Optics

Copper

- ☺ The Good: Well-understood, passive, low-cost, **highly reliable**
- ☹ The Bad: Physics limits distance at higher bandwidth
- ☠ The Ugly: Physical size and weight of wires is large

Passive copper direct attach cable (DAC)



Paladin high-density connector (Amphenol) (used in NVL72)

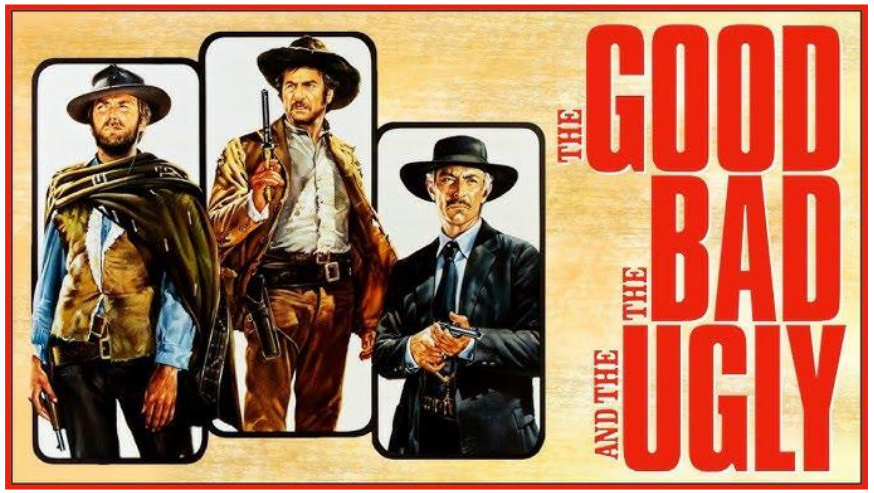
Optics – traditional pluggable

- ☺ The Good: No problem with distance even at 400G serial (if using InP or TFLN signaling technology), fibers are small and light
- ☹ The Bad: Huge change in size (for scale-up), cost, power
- ☠ The Ugly: **Poor reliability** compared to copper

octal small form pluggable (OSFP) optical module

Alternatives to traditional approaches being explored

- *These are promising but immature*
- Radio-mm wave
- µLEDs
- Advanced copper



Physical signaling technology vs. packaging (*no clear winner yet*)

physical signaling technology	fiber / distance	packaging								power consumption (per Gb/s)	bandwidth	comments	time-to-market*	leading technology companies
		high-density pluggable	DSP pluggable	LRO / RTLR	LPO	NPO	CPO (2D)	CPO (2.5 D)	CPO (3D)					
silicon photonics	single-mode > 2 km @ 200G	YES	YES	YES	YES	YES	YES	YES	YES	higher (1)	Good @224G No good @448G	well known technology requires ELS	now	NVIDIA, BRCM, Intel, TSMC, Ayar, Lightmatter
InP (EML)	single-mode > 2 km @ 200G	YES	YES	YES	YES	YES	perhaps	perhaps	NO	higher(0.9)	Good @ 448G	well known technology	now	Coherent, Lumentum
TFLN (thin film $LiNbO_3$)	single-mode > 2 km @ 200G	YES	YES	YES	YES	YES	perhaps	NO	NO	low (0.8)	Good @ 448G	commercial ready? requires ELS for CPO	1 year	Hyperlight
other TF crystals	single-mode > 2 km @ 200G	YES	YES	YES	YES	YES	perhaps	NO	NO	low (0.7)	Good @ 448G	not commercial ready requires ELS for CPO	2 years	Lumiphase
VCSELs	multi-mode 30-50 m @ 200G	YES	YES	YES	YES	YES	YES	perhaps	NO	lower (0.45)	Good @224G ? @448G	well known technology	now	Coherent, Broadcom
µ-LEDs	multimode array 10 m @ 10G	YES	YES	always	always	YES	likely	perhaps	NO	lowest (0.15)	~ 10G	not commercial ready	2 years	Avicena, Mojo, Credo
radio / mm-waves	dielectric λ-guide 10m?	YES	YES	?	?	likely	perhaps	perhaps	NO	low (?)	several 100s of Gb/s	not commercial ready	2 years	AttoTude, point2

For silicon photonics **well known technology** means > 10M pluggable units shipped
For InP and VCSELs **well known technology** means > 100M DSP pluggable units shipped each
*TTM = time-to-market in standard (OSFP) pluggable module

Summary

- Where are the LPOs?
  - LPO path to market starts at the NIC.
- CPO, NPO or LPO?
  - It's all linear
  - For AI fabric scale-up and scale-out, there exists a continuum of options.
- For AI scale-up and scale-out, which connectivity paradigms are breaking?
  - Weakening paradigms
    - CPO, NPO or 8/16-lane pluggables are only options
    - 32-, 64-, 128-receptical face-plates are needed for high radix switches.
    - Geometric constraints
    - Others…
- Optical disaggregation
- Road towards 448G

Thank you!

DELL Technologies