Sustainable Optics for Scaling AI

Andreas Bechtolsheim

Arista Networks

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

Expected AI Fabric Bandwidth Growth

	2022	2024	2026	2028
Bandwidth/XPU	3200	6400	12800	25600
1600 Ports / XPU	2	4	8	16
XPUs Shipped [M]	3	5	8	10
Total 1600 Ports [M]	6	20	64	160

Note: Bandwidth numbers shown are Uni-directional. For Bi-Directional numbers multiply by 2X

Nearly a 10-fold increase in bandwidth from 2024 to 2028

Passive Copper Cables in the Rack

Passive Copper Cables have many advantages

- Lower power, lower cost and higher reliability than optics
- Main limitation is 1m reach at 224G-PAM4

High performance SERDES is the key enabler

- Passive copper cables is a very important use case
- This will not change for the foreseeable future

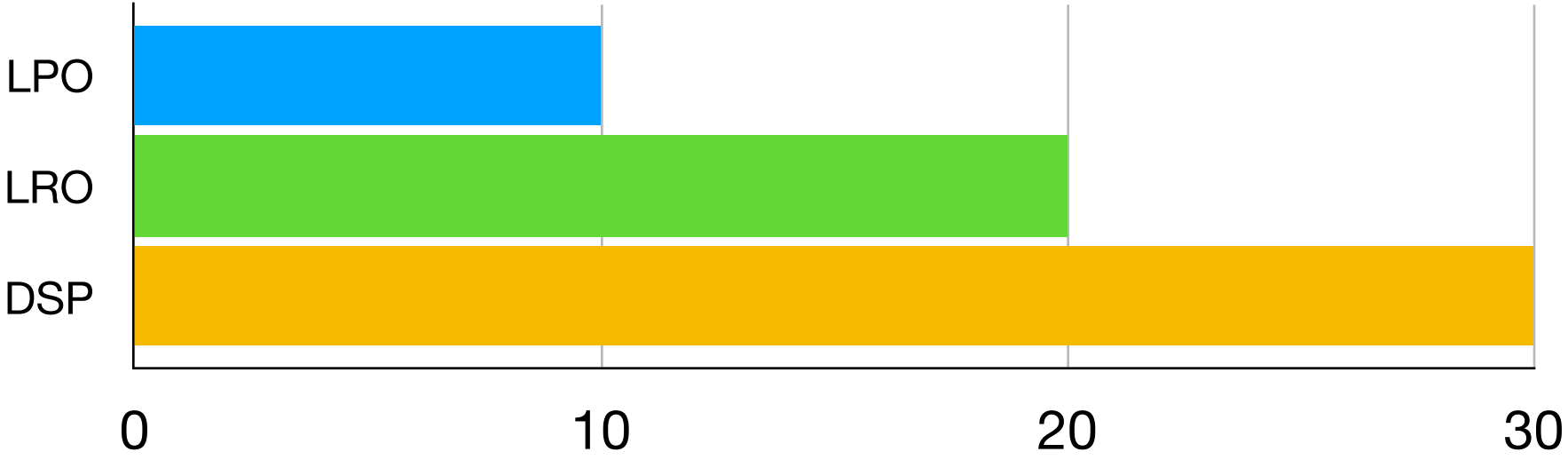
Copper cables within the rack are well proven

Example: Nvidia NVL72 Rack 72 GB200

PHY	PHY Pwr / Port [W]	PHY Pwr / Rack [W]	Total Rack Power [W]	Relative Power [%]
Passive Copper	0	0	120,000	100%
DSP Optics	30	19,440	139,440	116.2%

Copper backplane substituted 648 1600G Optics
DSP Optics would have added 20KW in power,
the equivalent of 12 GB200 chips

One Big Problem: Power for 1600G Optics



Measured Power per Prototype 1600W-DR8 Modules

Next-Gen Scale-Up Domains: 256 to 512 XPUs

Multi-Rack Designs, even with next-gen 400KW Racks

- 4 Racks with 128 3KW XPUs per Rack
- Passive Copper cables cannot meet reach

Need Lower Power Interconnects for Scale-Up

- Reach of at least 10 meter for 512 node cluster
- Includes “slow-and-wide” Optics, uLEDs, uWave

Need lower-power solutions in particular for Scale-Up

Sustainable Optics for Next-generation AI Clusters

Interconnect	Reach	Power per Side [W]	Power per XPU [W]	Power per 1M XPU [MW]	% of XPU Power
DSP Optics	10km	30	960	960	32.0%
LRO Optics	1km	20	640	640	21.3%
LPO Optics	500m	10	320	320	10.7%
uLED/uWave	10m	8	256	256	8.5%
Active Copper	3m	5	160	160	5.3%
Passive Copper	1m	0	0	0	0.0%

Assumes single-tier Scale-Up fabric with 3KW 2027 XPU with 25.6T Fabric I/F = 32 1600G Optics per XPU

Stargate Size AI Datacenter with 1M XPUs

Configuration	DSP Optics	LPO Optics	Delta	LPO/DSP %
1M XPUs	3 GW	3 GW	0	
32M Optics	960 MW	320 MW	640 MW	-66.7%
TOTAL Power	3960 MW	3320 MW	640 MW	-16%
Optics Power/XPU %	32%	11%		-66%
Power/Y @ 0.25/kWh	$8.67B	$7.27B	$1.4B	-16%
Power over 5 Years	$43.35B	$36.35B	$7B	-16%

$70B in power savings over 5 years per 10M XPUs

Call to Action: Solve the Interconnect Power Problem!

- Need to accelerate adoption of Linear Optics
  - Power consumption of 1600G DSP Optics not practical for AI
- Need lower-power interconnect technologies for Scale-Up
  - A large % of the volume will be 10m reach
- Timeline for qualifying 1600G solutions is 2025
  - 1600G will be ramp to high volume in 2026
- All 224G Switch and System Designs in-flight
  - No realistic opportunity for CPO
- 448G SERDES Transition is next
  - Electrical Channel extremely challenging
- Nobody said it was going to be easy
  - No one ever said it would be this hard

Thank you!