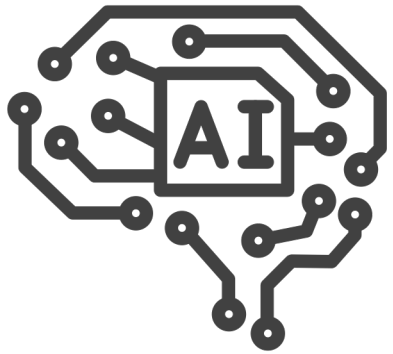
THE GENERATIVE AI AND THE FUTURE OF CLOUD AND DATA-CENTER

Thitipat J.
thitipat@juniper.net

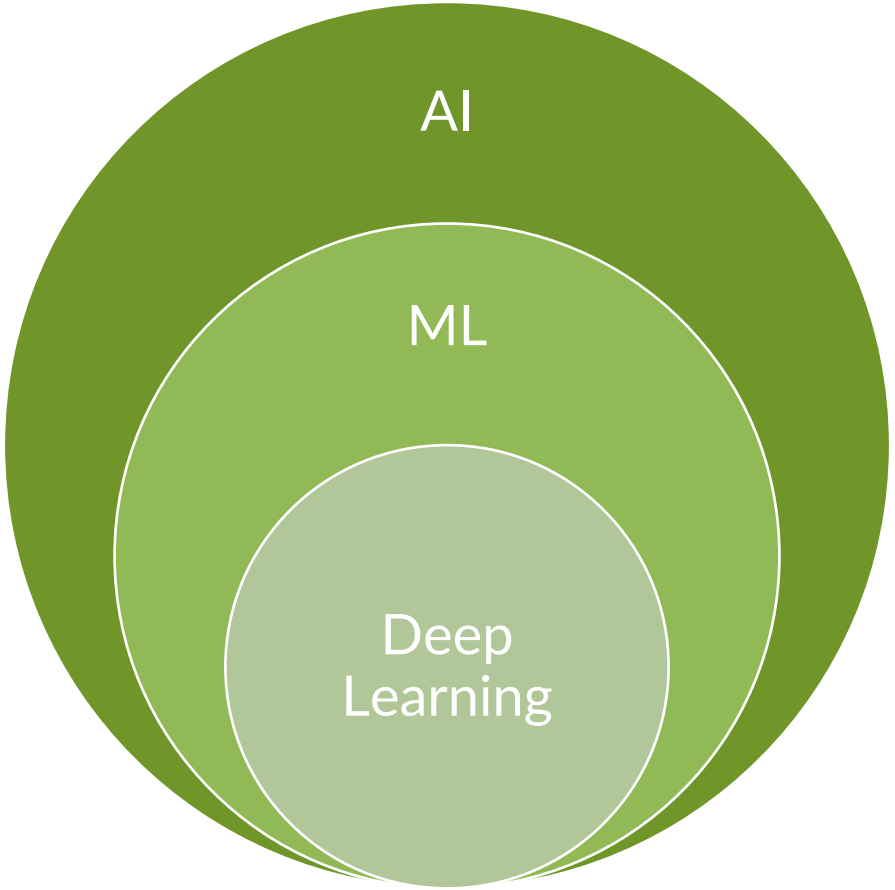
JUNIPER NETWORKS | Driven by Experience™

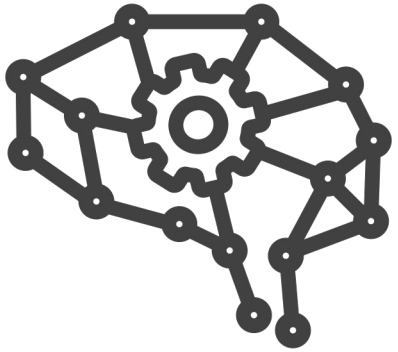
AI/ML?

What is Machine Learning?



Artificial Intelligence
Is a discipline

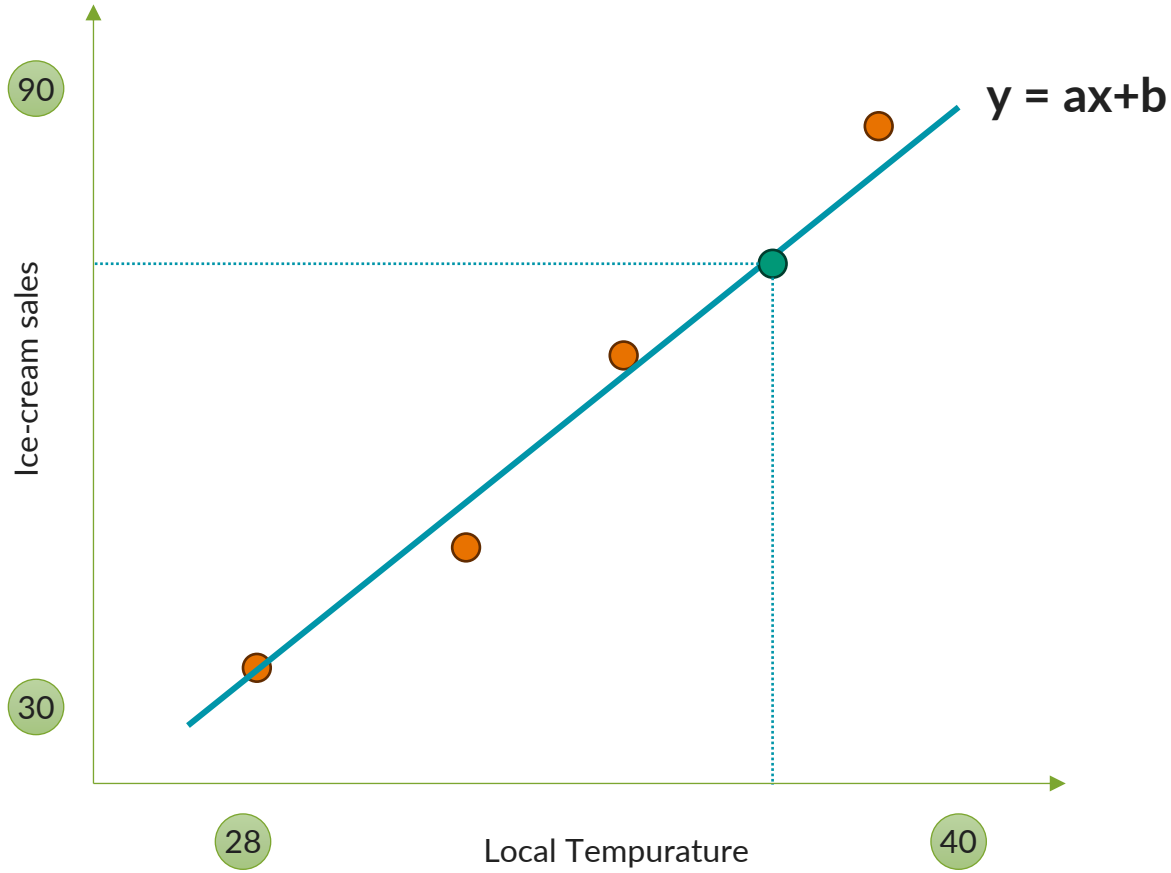




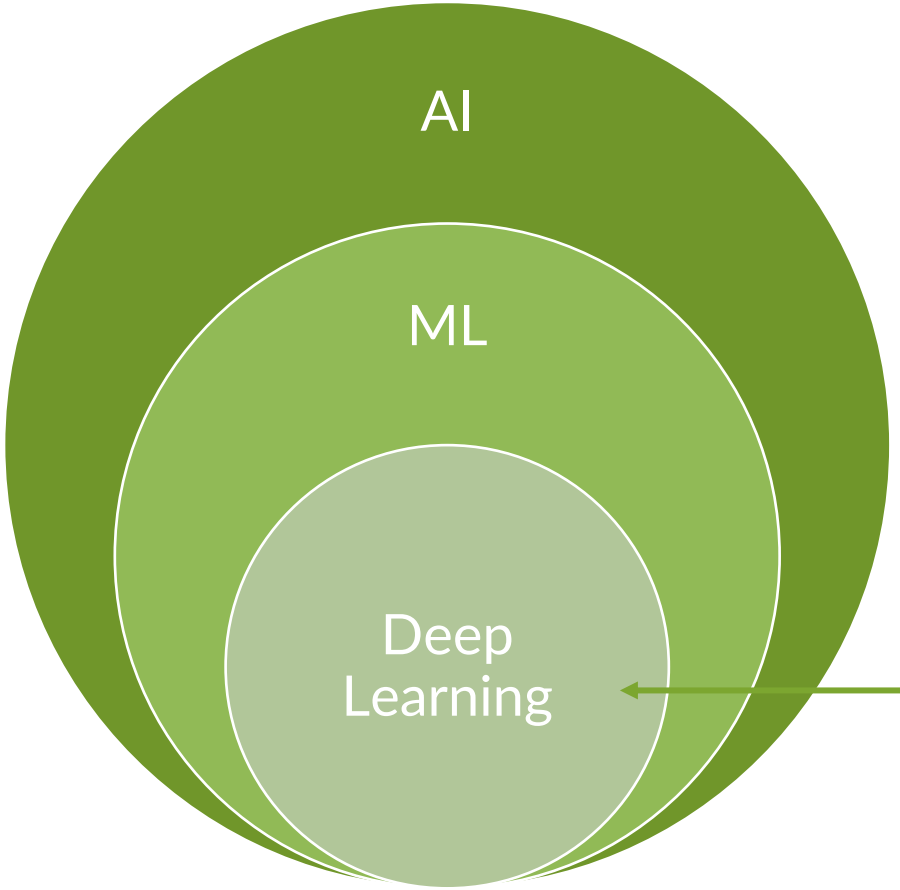
Machine Learning
Is a subfield

Supervised learning

Implies the data is already labeled



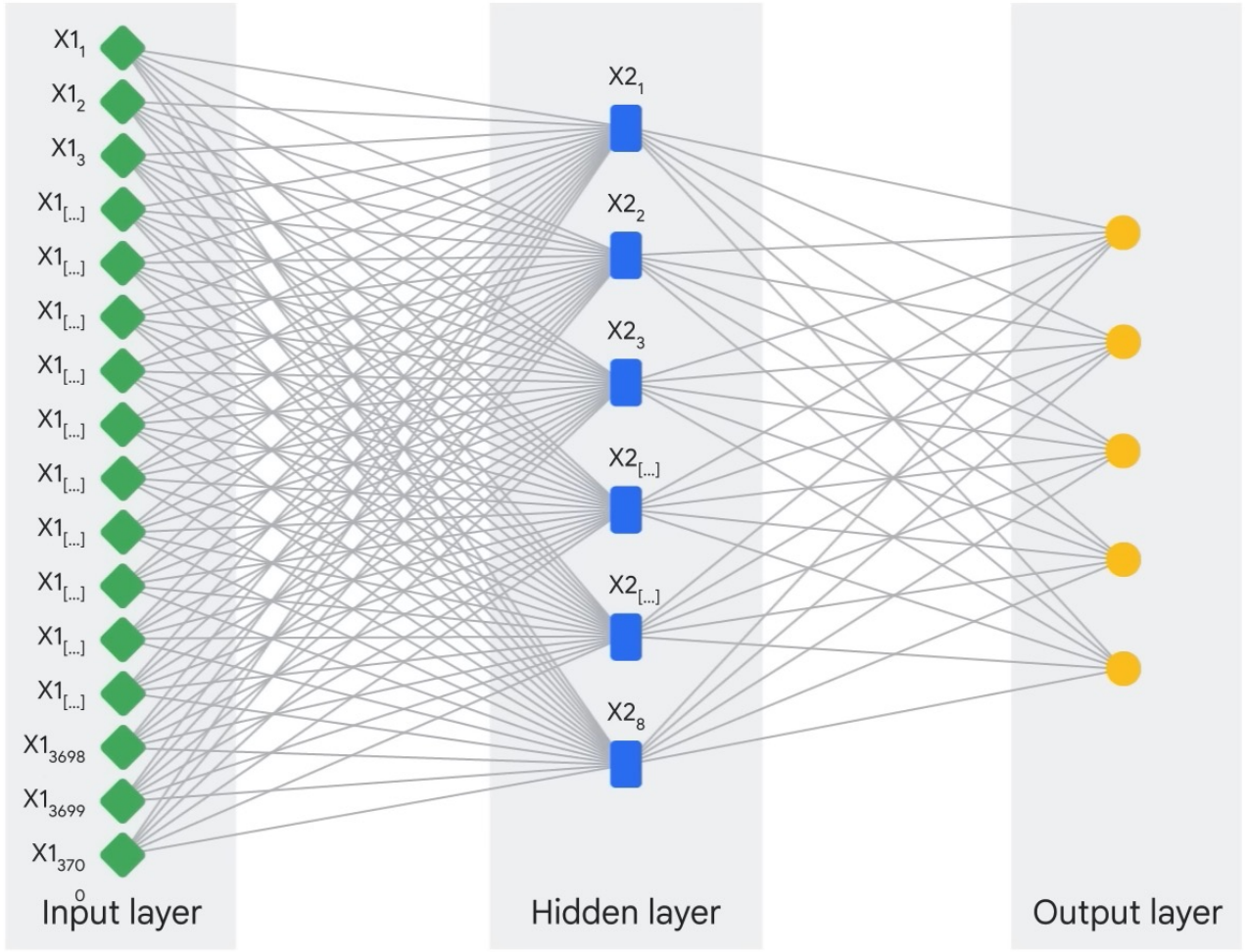
What is Machine Learning?



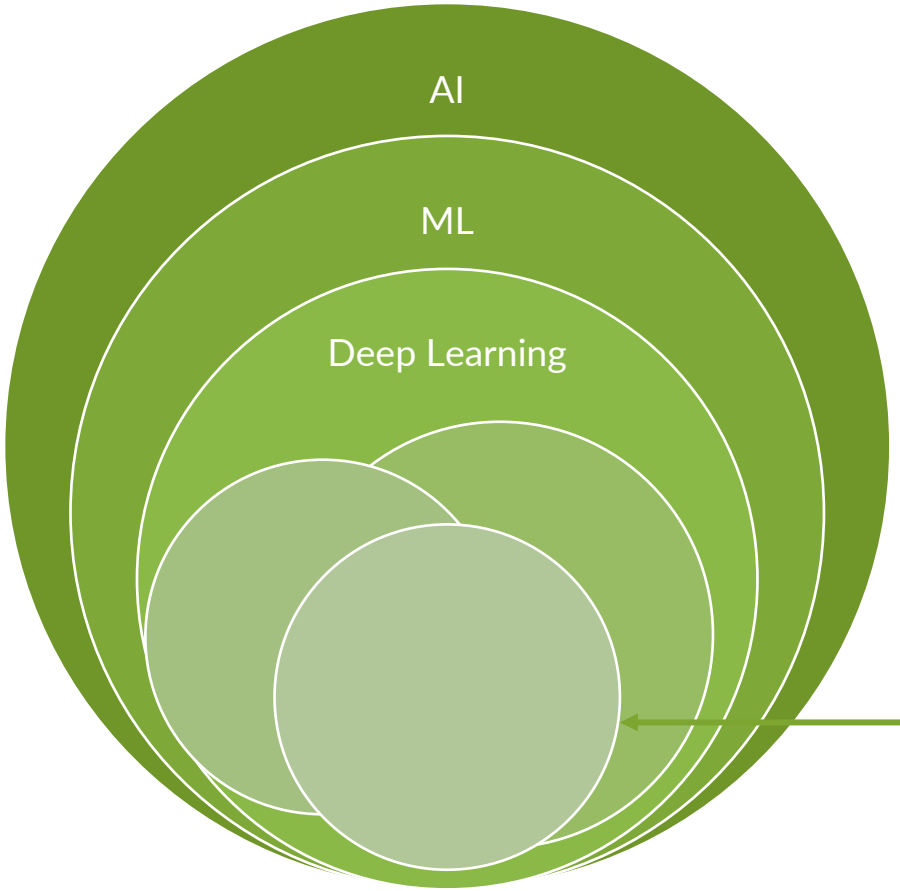
Machine Learning

- Supervised learning
- Unsupervised learning
- Reinforcement learning
- Deep learning

Deep learning uses Artificial Neural Networks – allowing them to process more complex patterns than traditional machine learning



What is Machine Learning?



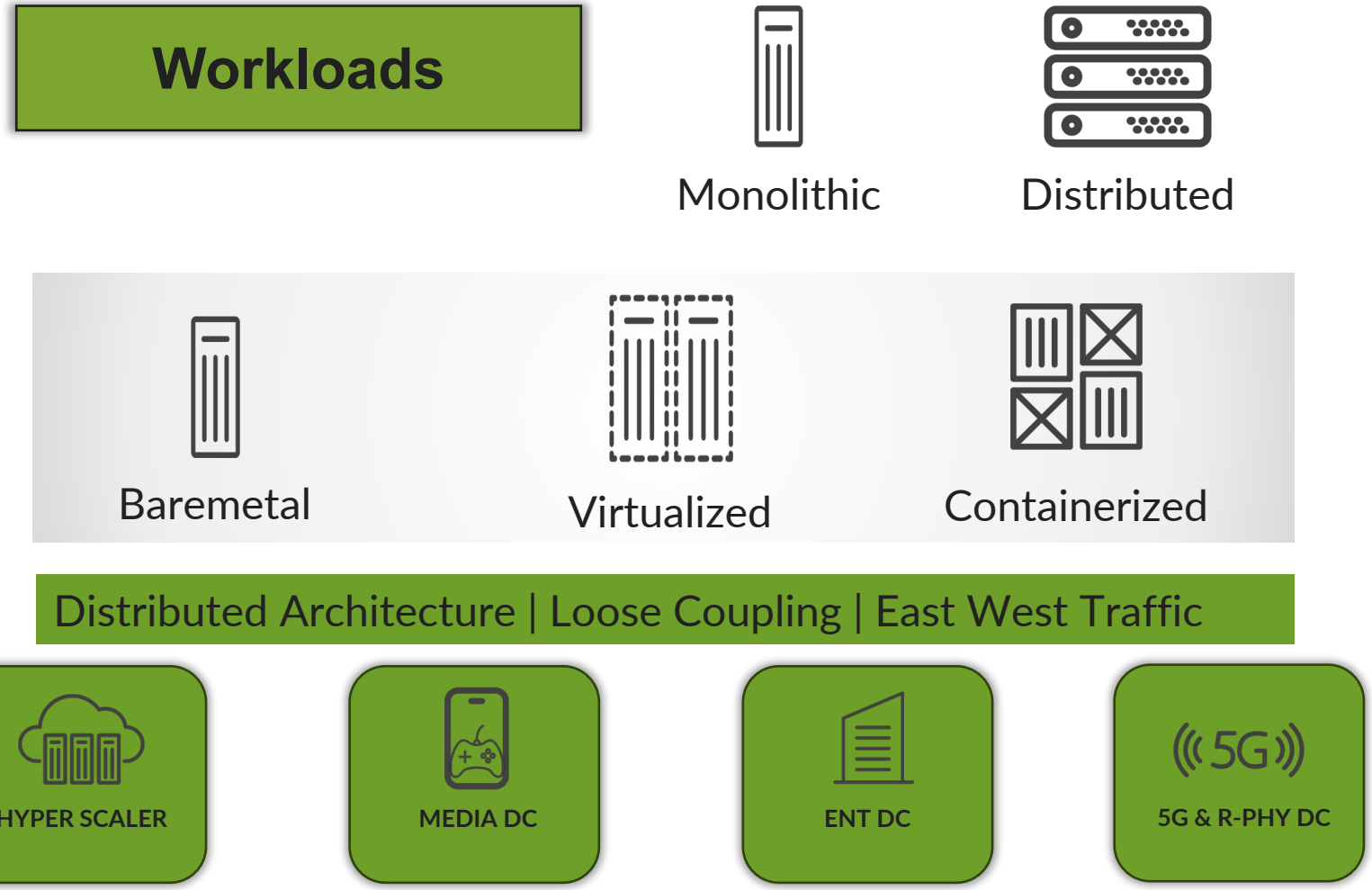
Deep Learning

- Generative AI
- Large Language Model

MODERN DATA-CENTER WORKLOAD

JUNIPER NETWORKS

Existing Workloads



Modern Workloads

GPU/TPU Acceleration

Parallel computing across servers

Disk read/write speed improvement

NVMe (ROCEv2/NVMeTCP)

Analytical / Training / Storage

Existing vs Modern Workloads

Existing Workloads

- Heterogenous applications
- High number of tenants
- Workloads are loosely coupled
- GPU/TPU requirement is relatively less
- Relatively less throughput

Modern Workloads (AI/ML)

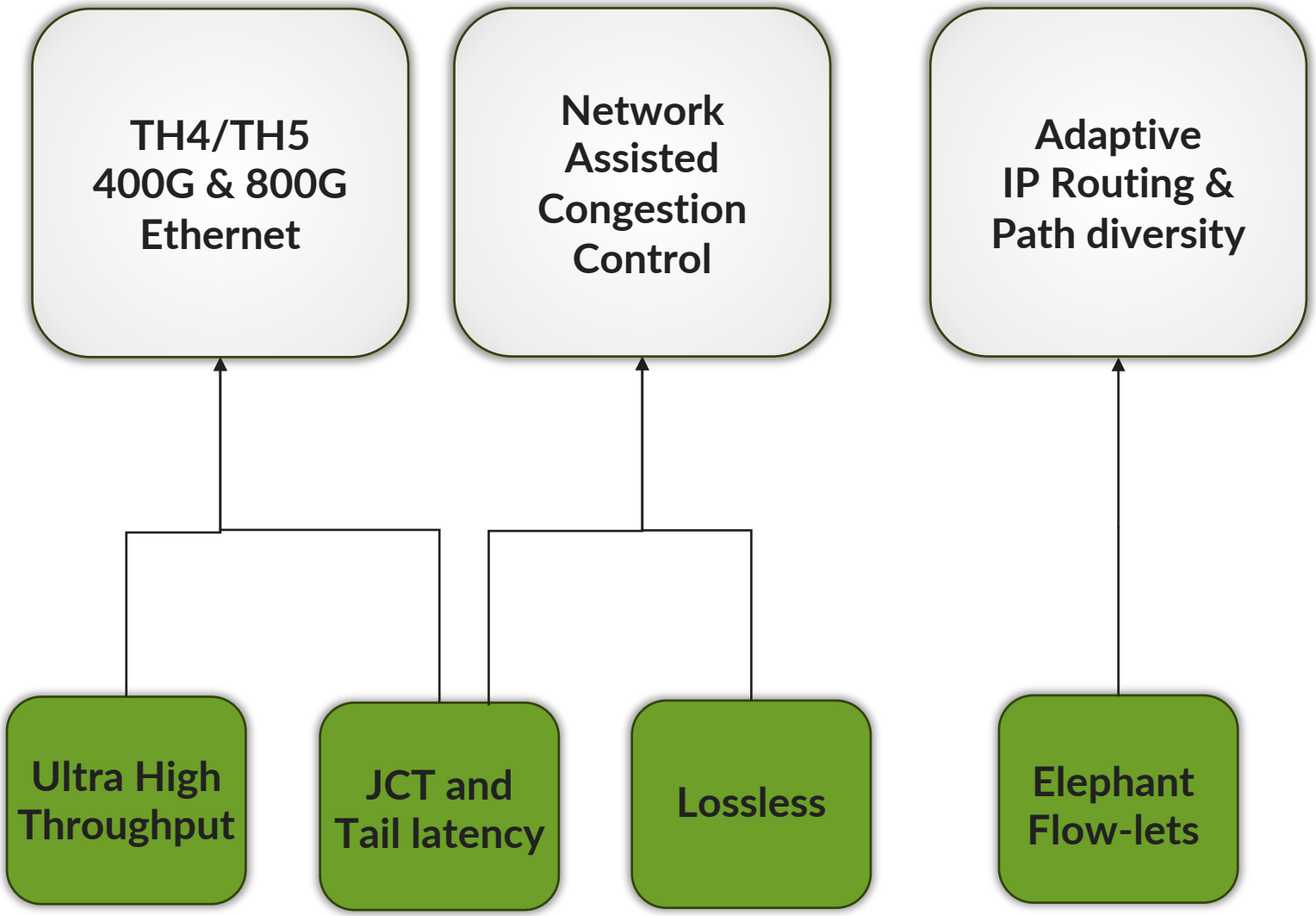
- Large computing problems
- Low number of tenants
- Workloads are tightly coupled
- GPU/TPU requirement is high
- Very high throughput

Modern DC Requirements

- Ultra High Throughput
- JCT and Tail latency
- Elephant Flow-lets
- Optimized Lossless
- Distributed Security *

New Product Mapping

New Juniper products



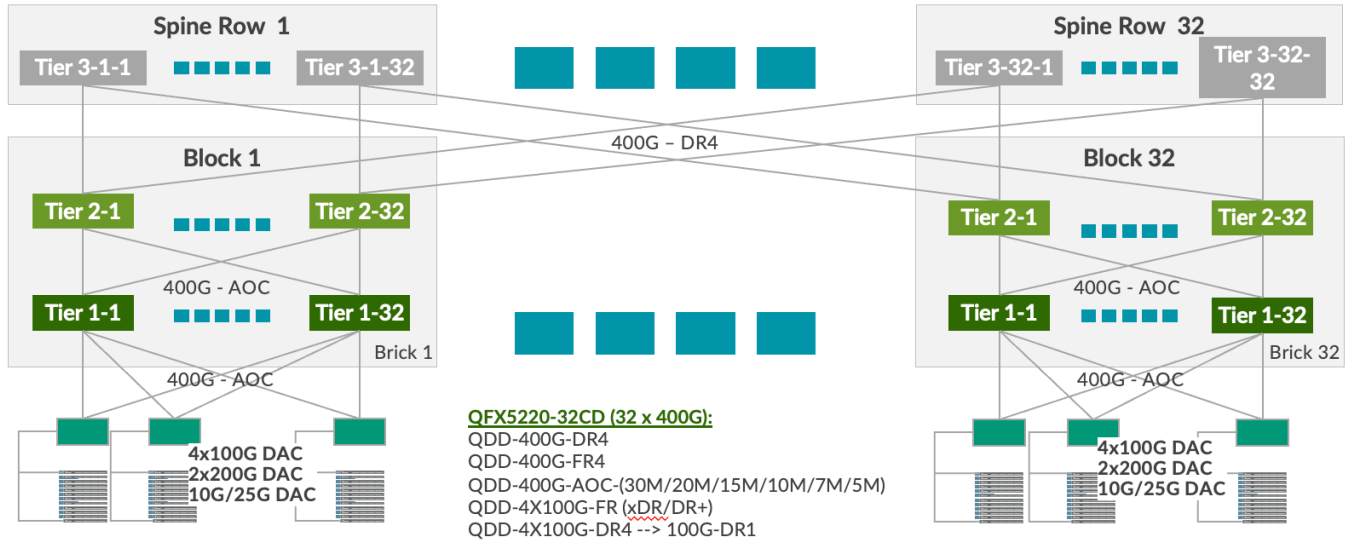
Requirements

Modern Workloads Requirements

High Throughput and Density	• Increase port speed • Increase port density
Reduced Job Completion Time (JCT)	• 1:1 subscription fabric design • Reduce latency (e.g. cut through mode,TH-F1)
Efficient Load Balancing	• Dynamic/Global Load Balancing • IP ECMP (64/128)
Reliable Transit	• ROCEv2 (PFC-IP/ECN) + (Source Flow Control, Congestion Isolation) • Sub-second convergence time • Deep buffer [TBD for AI/ML fabric]
Zero trust security	• MacSec and overlay encryptions VxLAN-Sec • DDoS-protection
Juniper Apstra™ Intent-based operations	• Automated Deployments • Easy scale-out and scale-in • Telemetry, xFlow, and closed loop automation

Hyper-scalar's DC

Architecture	• Multi-tier Native IP Fabric
Platforms	• QFX5220-32CD/128C (EVO) • QFX5210-64C (Junos) • QFX5200-32C (Junos)
Key Feature	• eBGP IPv4/IPv6 • BGP unnumbered underlay • Dynamic Load Balancing • IP ECMP 64 • TCP AO • ISSU
Scaling	• 1000 * BGP peers • 350 to 700K LPM • Shallow buffers: 64MB/128MB
Interoperability	• Cisco • Arista



High-Level Explanation

- Multitier T0->T2 IP CLOS architecture
- Bandwidth at competitive cost
- Block level architecture
- Network virtualization done at server level

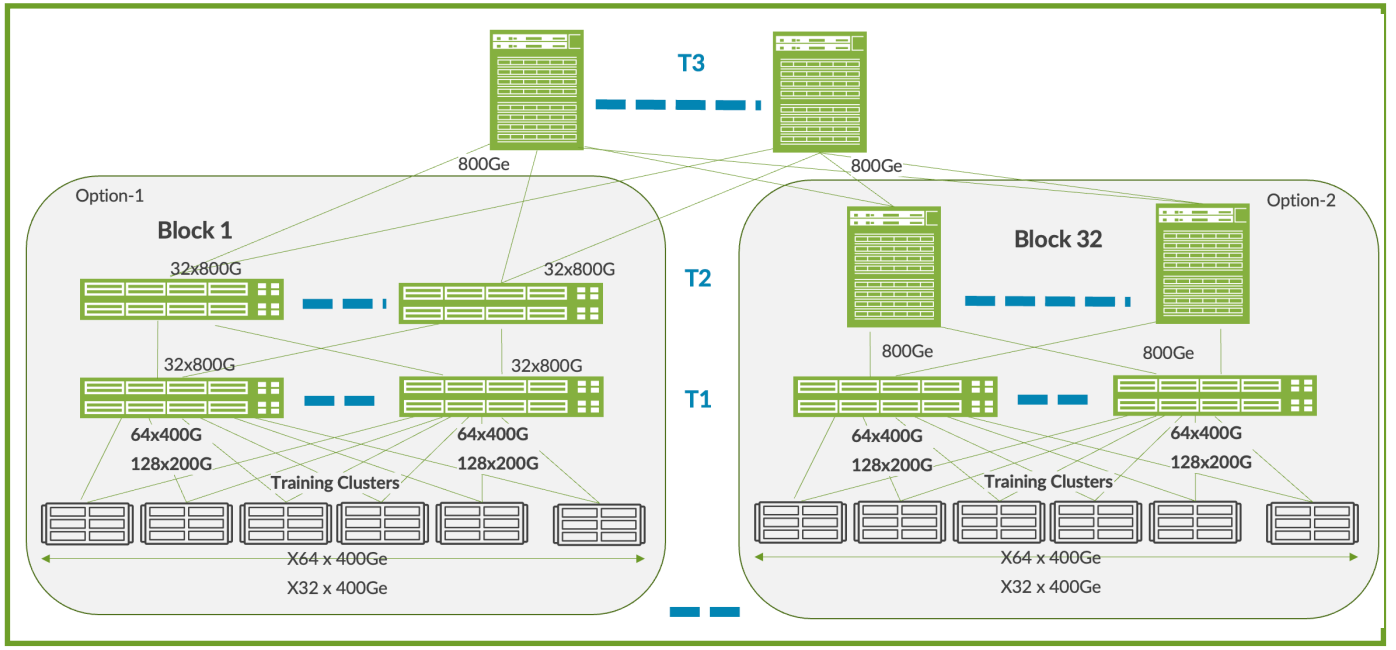
AI/ML block connect

Architecture
- Multi-tier IP Fabric

Platforms
- QFX5240 – TH5 *
- QFX5230-64cd
- PTX10008 [BX based] to reduce the number of nodes at the T2 level

Key Feature
- BGP unnumbered underlay
- Dynamic Load Balancing
- IP ECMP x 128
- ROCE V2

Scaling



High-Level Explanation
- Block connect approach
- Spine row connects to the external world
- 1:1 or 2:1 oversubscription

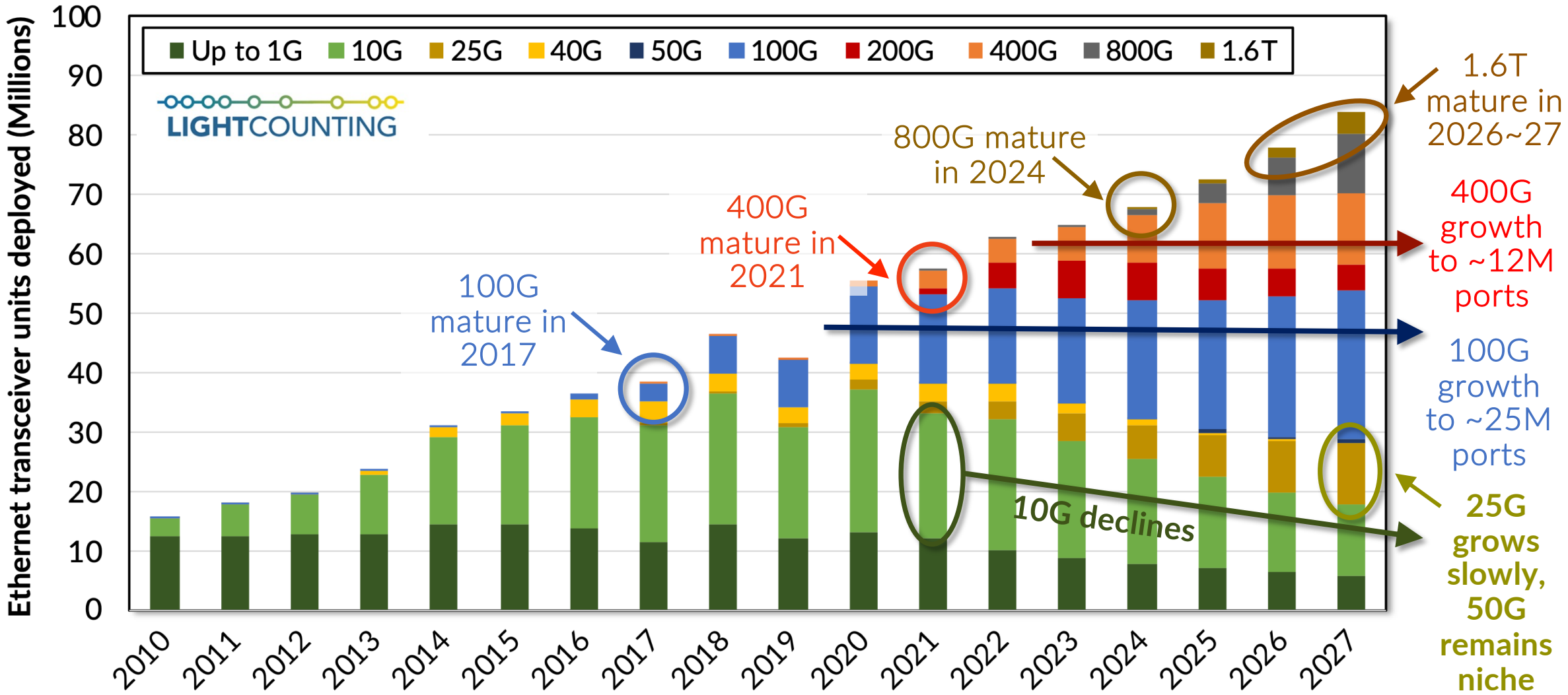
Interoperability
- Arista
- Cisco

THE EVOLUTION

JUNIPER NETWORKS

The Ethernet (R)evolution

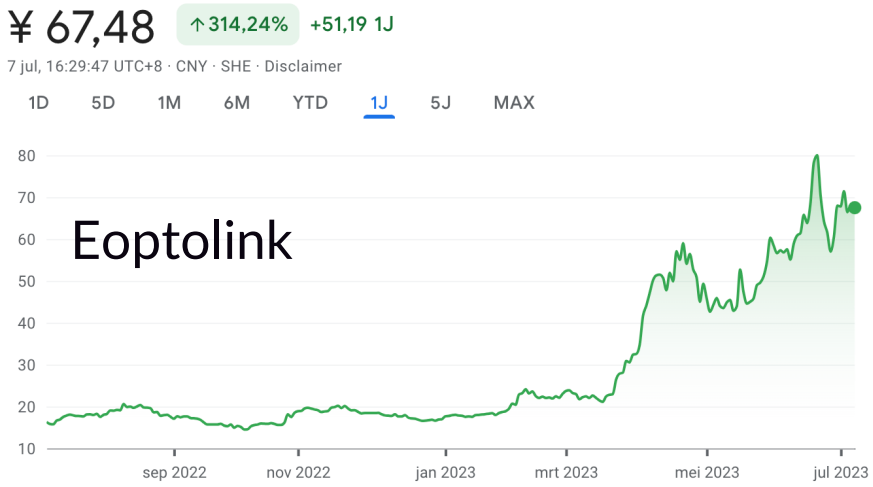
Past, present and future of ethernet transceiver sales across the industry



Adapted from Lightcounting, September 2022 High Speed Ethernet Optics Report

What's driving the need for 800GE/1.6TE?

The AI/ML Goldrush



- AI/ML "goldrush" heavily impacts major optics vendors, as they are expected to benefit from the increases use of 800G and beyond optics in large AI/ML clusters.

What's driving the need for 800GE/1.6TE?

The AI/ML Goldrush

- **AI/ML clusters require A LOT of bandwidth:**
  - Latest generation of GPUs (Nvidia Hopper) use up to 3.6 Tbps as GPU-to-GPU interconnect for shared memory access.
  - Front-end network of a high-end GPU server such as the Nvidia DGX H100 with 8 GPUs has 10 x 400G network interfaces (InfiniBand or Ethernet).
- **AI/ML clusters traditionally use InfiniBand:**
  - Better control over tail-end latency with (very) large flows going over the fabric.
  - Hyperscalers prefer to adopt Ethernet instead:
    - Better scalability to larger clusters
    - Better suited for multi-tenant clouds running many different applications.
    - Latecy can be controlled by packet spraying and re-ordering [1]

Nvidia DGX H100

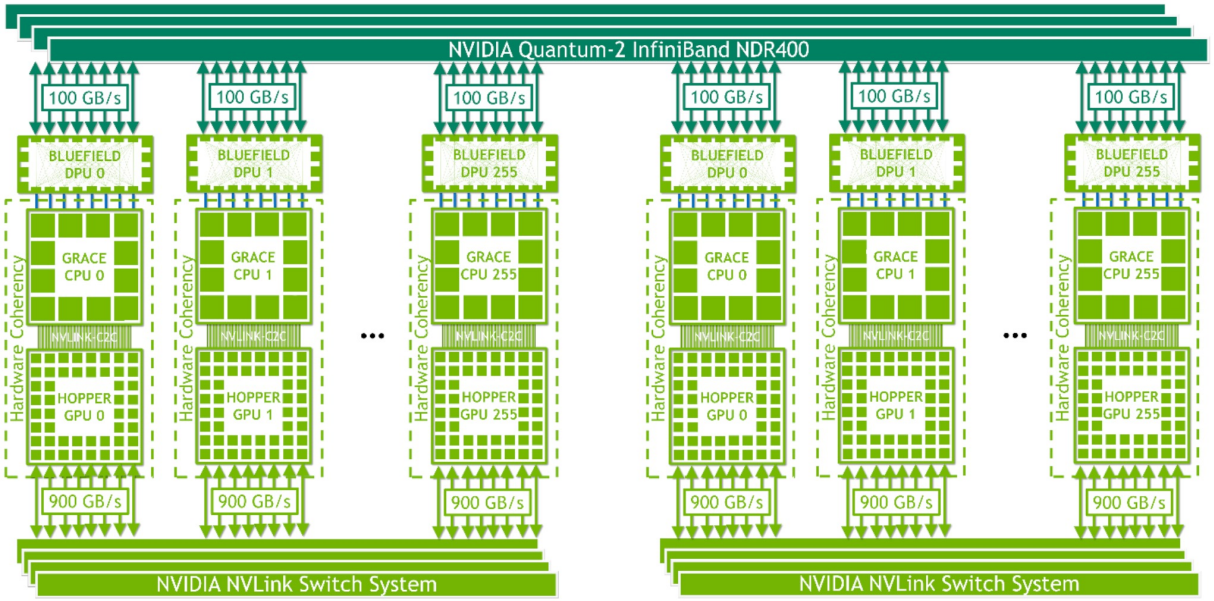
https://www.nvidia.com/en-us/data-center/dgx-h100/

[1] https://nvdam.widen.net/s/6lmkmc8lqg/nvidia-spectrum-x-whitepaper

What's driving the need for 800GE/1.6TE?

The AI/ML Goldrush

- **AI/ML clusters for large language models take connectivity to even more extreme levels:**
  - Nvidia DGX GH200 has back-end network to interconnect 256 GPUs with 7.2 Tbps of GPU-to-GPU bandwidth and 921.6 Tbps bi-sectional BW.
  - Shared memory access creates shared-GPU-memory space of 144 TB.



Nvidia DGX GH200

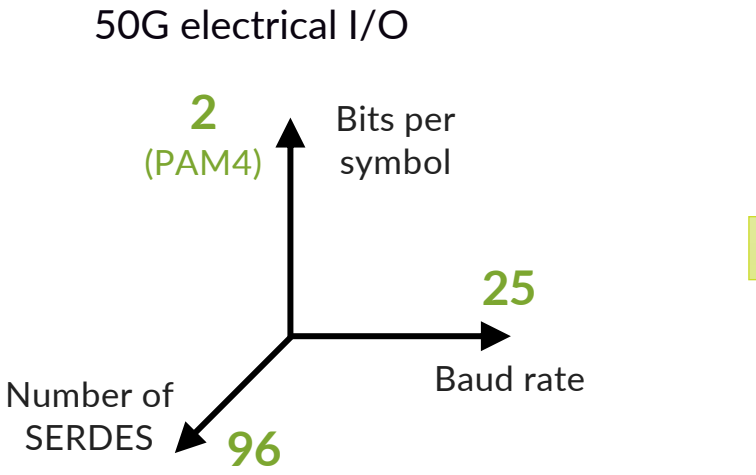
https://hc34.hotchips.org/assets/program/conference/day2/Network%20and%20Switches/NVSwitch%20HotChips%202022%20r5.pdf
https://developer.nvidia.com/blog/nvidia-grace-hopper-superchip-architecture-in-depth/

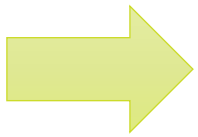
The evolution from 400G to 800G

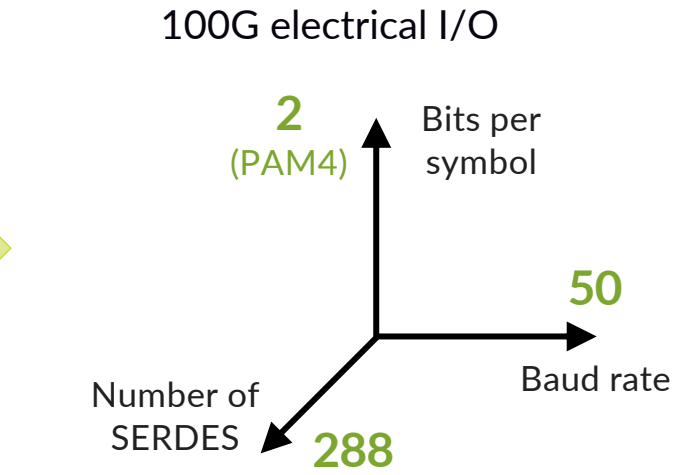
800G adoption on routers and switches

Evolution to 100G Electrical I/O

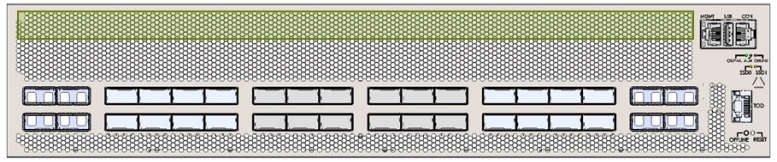
- **Industry is evolving from 50G to 100G electrical I/O, and number of SERDES per PFE increases:**



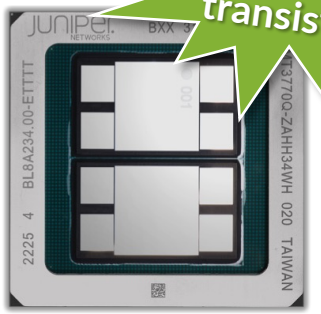




PTX10001-36MR
24 x 400GE

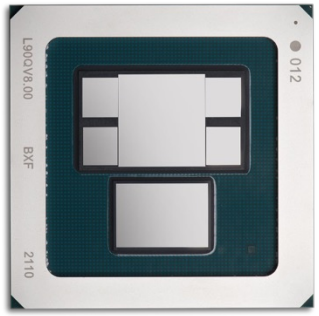


NG compact PTX
36 x 800G



Juniper Express 5
(28.8T, BXX)

288 x 100G



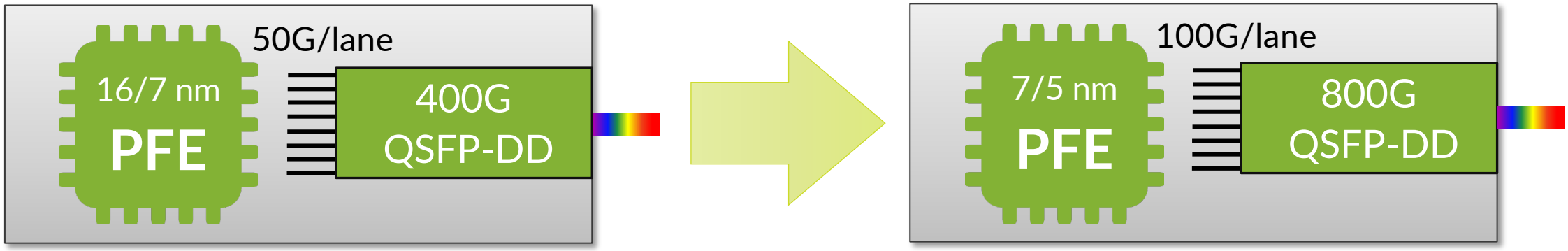
Juniper Express 5
(14.4T, BXF)

144 x 100G

For more details: Chang-Hong Wu, “Juniper’s Express 5: A 28.8Tbps Network Routing ASIC and Variations”, https://hc34.hotchips.org

800G adoption on routers and switches

Evolution to 100G Electrical I/O



8 x 50G PAM4
Electrical I/O

4 x 100G PAM4
Optical I/O

8 x 100G PAM4
Electrical I/O

8 x 100G PAM4
Optical I/O

100G SERDES

53 Gbaud PAM4 (106.25 Gbit/s/lane) using KP4 FEC

8x100G PAM4 optical I/O

Backwards compatible with mainstream 100G/400G optics

100/400GE break-out

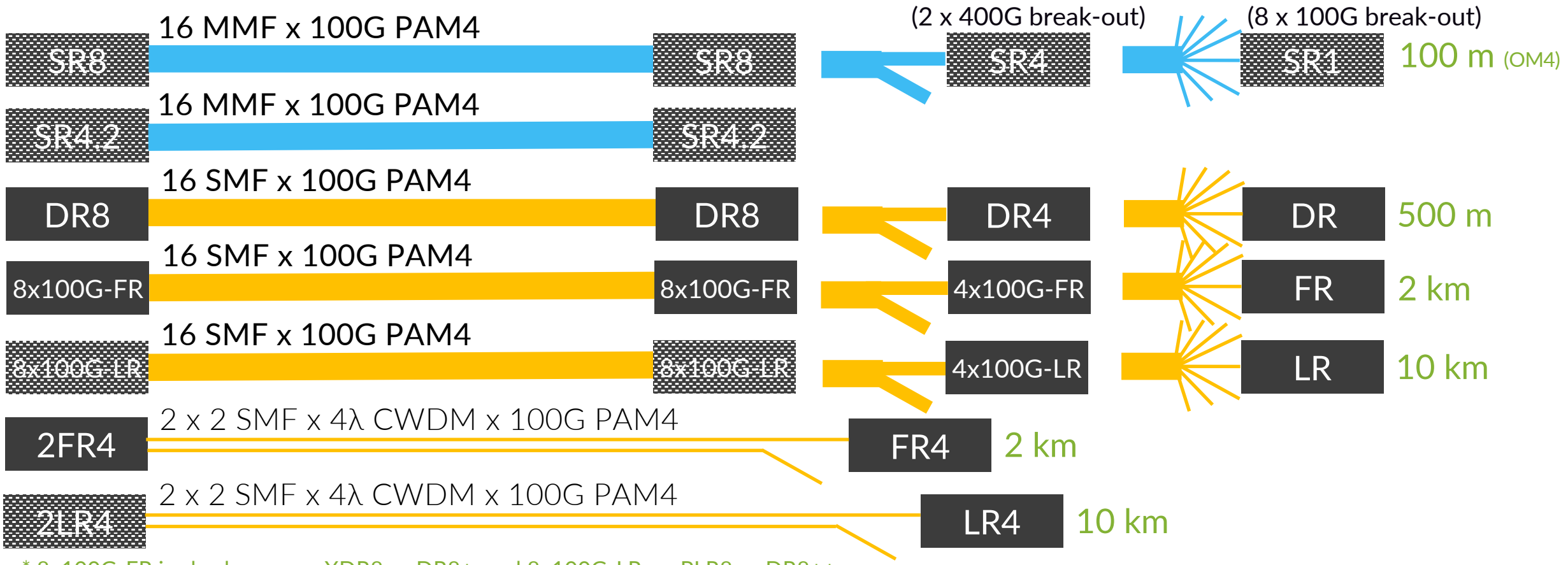
Highly optimized for 8x100GE and 2x400GE break-out

- **The adoption of 100G serial electrical I/O is the key building block for high-density 100GE/400GE-optimized routing and switching platforms**

JUNIPER NETWORKS

800GE optical client interfaces

100G optical I/O extends the life cycle of 100GE & 400GE

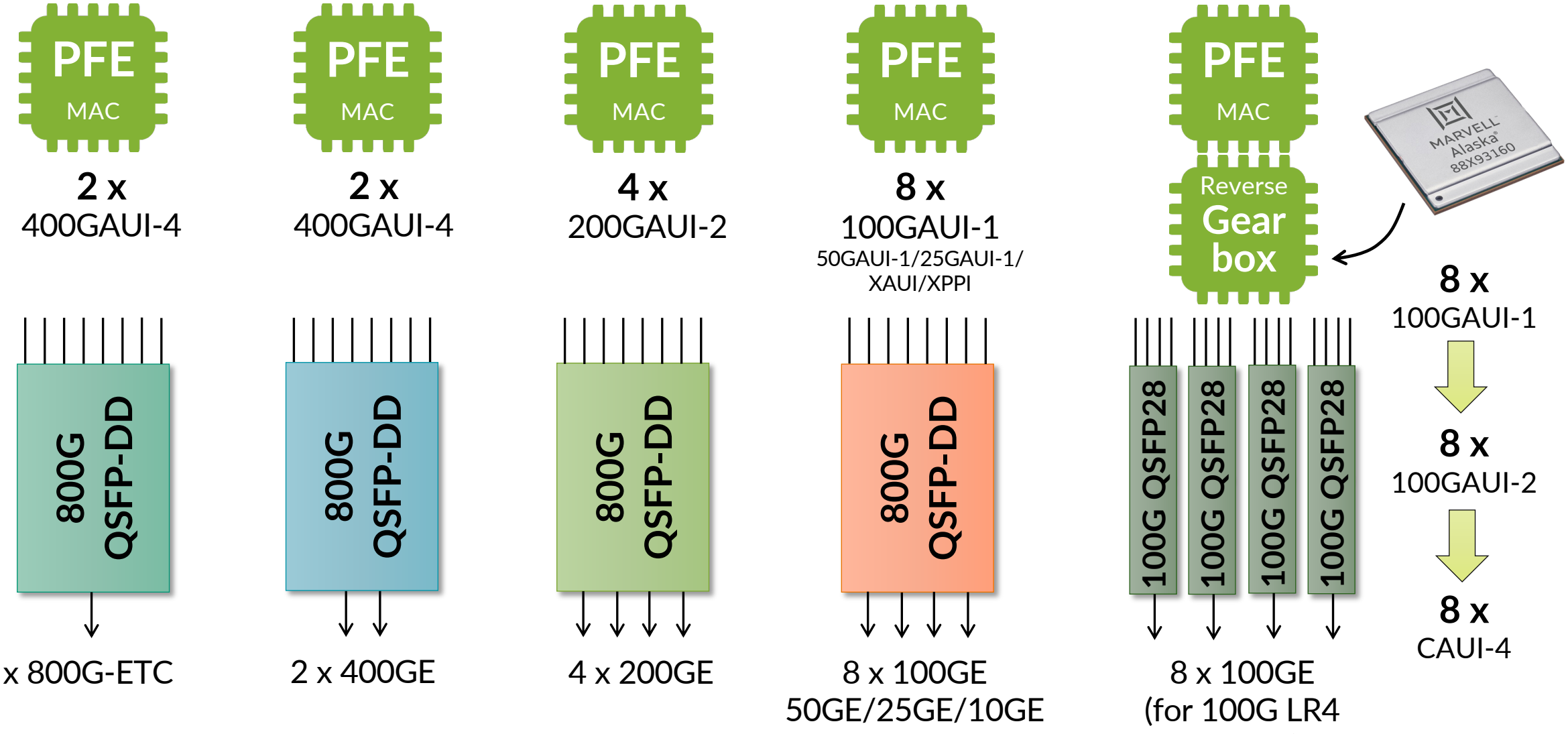


* 8x100G-FR is also known as XDR8 or DR8+, and 8x100G-LR as PLR8 or DR8++

- **Today's mainstream 100G/400G optics, i.e. 100G DR/FR/LR and 400G DR4/FR4/LR4 are forward compatible with 800G break-out**

Break-out options for 800G ports

Increased fan-out to support high-radix architectures



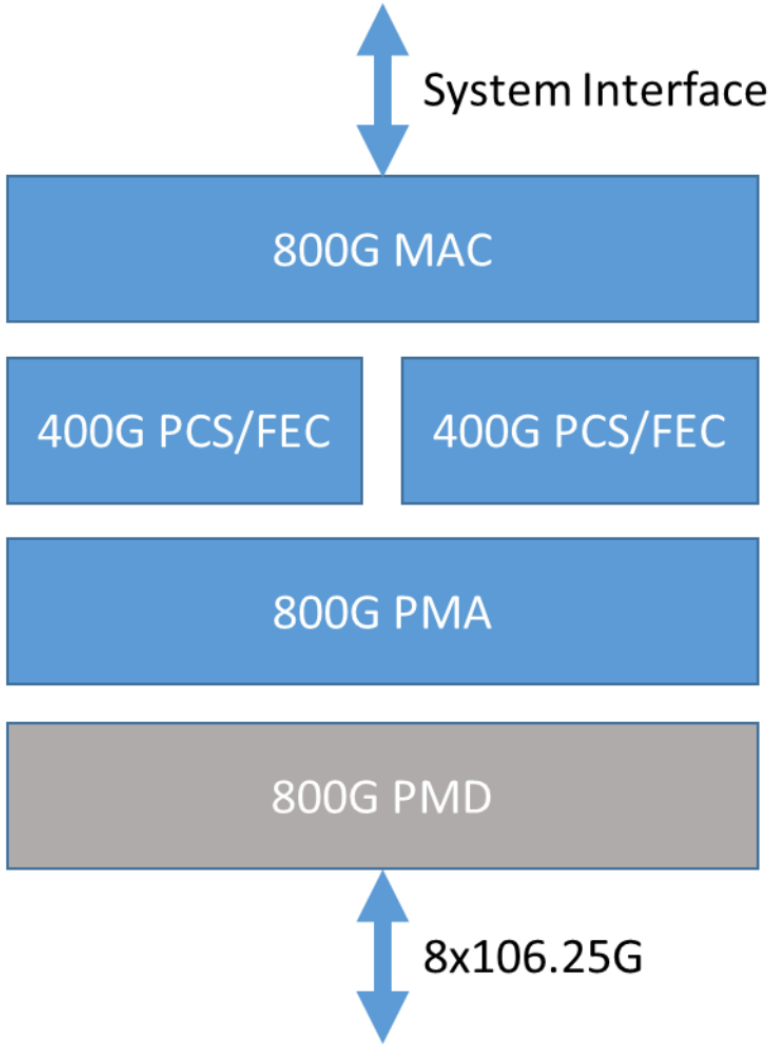
JUNIPER NETWORKS

800G & 1.6T standardization

800G EthernEt “Time to Market”

ETC specification for 800G

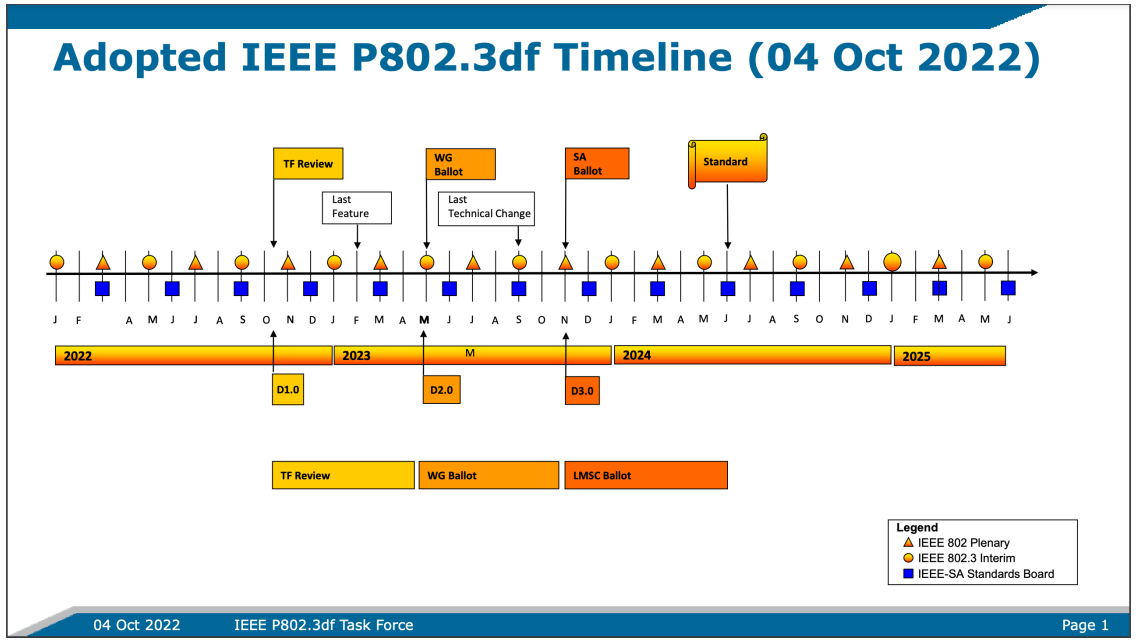
- **Data centers are now starting to deploy 800G ports:**
  - Using e.g switch silicon such as Broadcom Tomahawk 4
  - Initially mainly for use as 2 x 400GE
- **800G Ethernet MSA specification released in 2020 by the Ethernet Technology Consortium (ETC)*:**
  - ETC specification doubles bandwidth of 400GE to support 800G clear channel.
  - Re-uses the PCS/FEC specification from 400GBASE-R.
  - Effectively 2 x 400G PCS in parallel, i.e. 2 x (16 x 25G) PCS lanes
- **Routers & switches supporting 800GE ports will start to become available in 2023~24**
  - Including Juniper Express 5 (BX) and Broadcom Jericho 3



* The ETC was previously known as the 25 Gigabit Ethernet Consortium
https://ethernettechnologyconsortium.org/wp-content/uploads/2020/03/800G-Specification_r1.0.pdf

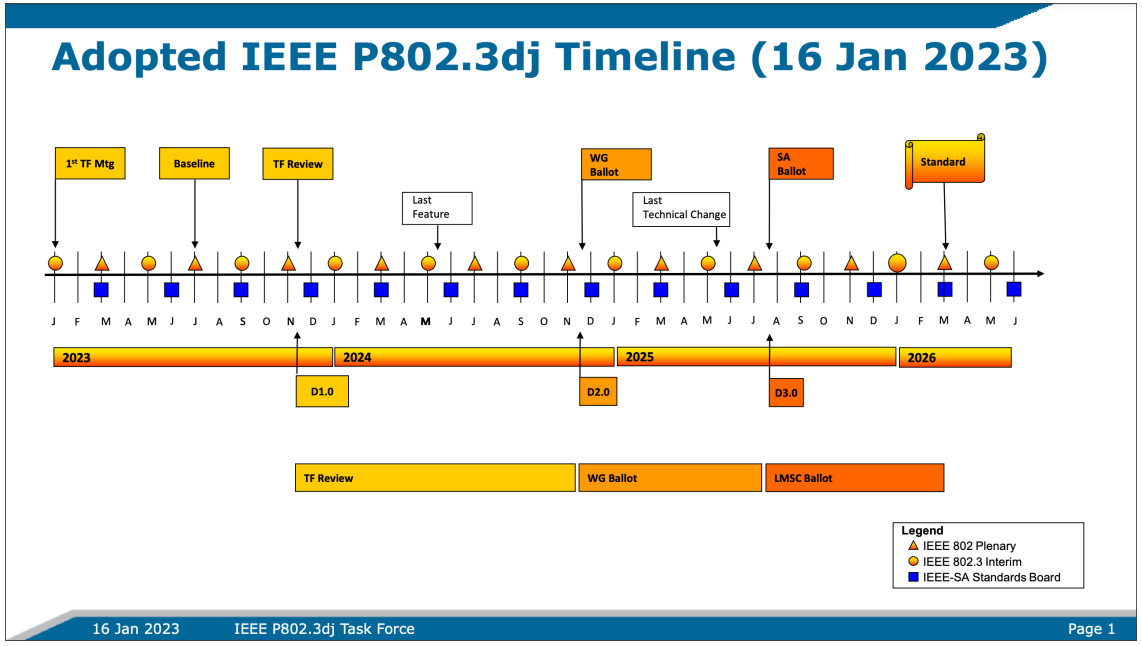
800GE & 1.6TE standardization

IEEE 802.3df & 802.3dj



IEEE 802.3df
focussed on 800GE with 100G electrical I/O
(Mid 2024)

IEEE 802.3dj
focussed on 800GE / 1.6TE with 200G electrical I/O
(early 2026)



- **New Ethernet standards require fundamentally new component and system technologies to build consensus for standard with long-term relevance**