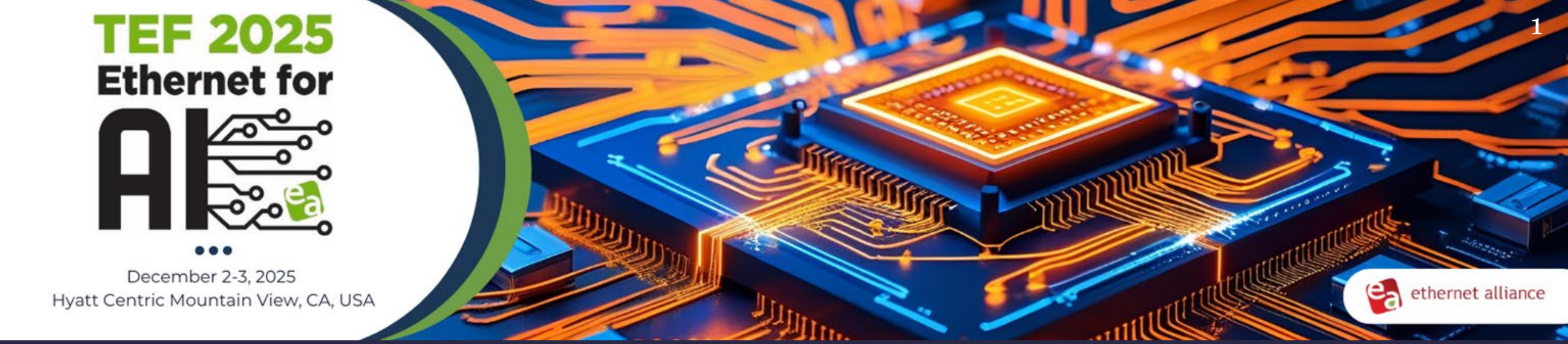


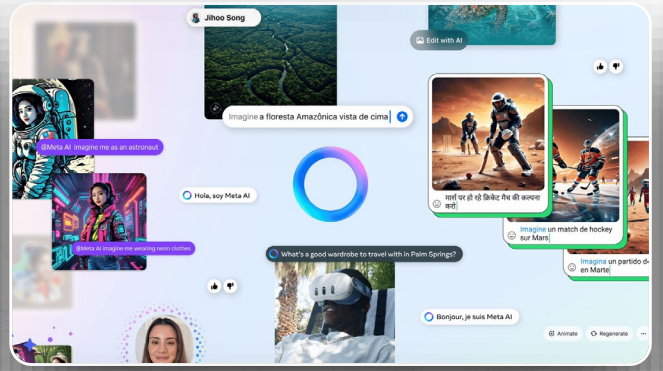
Transitioning to 400G SerDes: Key Drivers and System Design Implications for Future AI Workloads

Halil Cirit, AI Architect & META

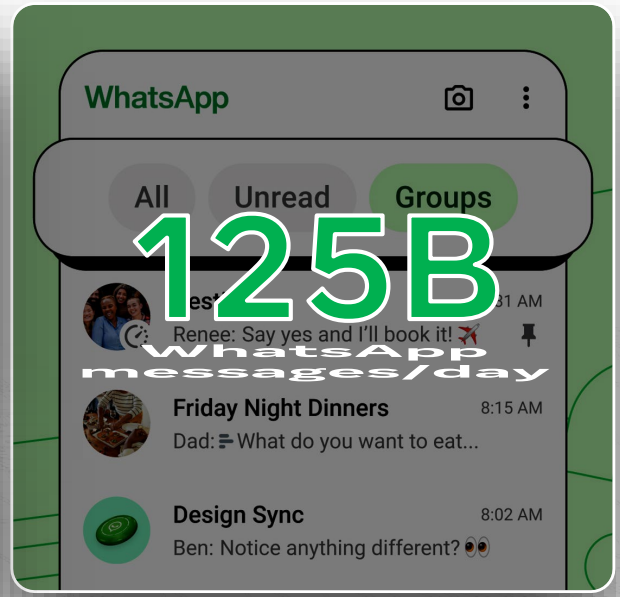
This presentation has been developed within the Ethernet Alliance, and is intended to educate and promote the exchange of information. Opinions expressed during this presentation are the views of the presenters, and should not be considered the views or positions of the Ethernet Alliance

Keynote Speaker

Seasoned SerDes and System expert with over 19 years of experience in the semiconductor industry. Since November 2018, has been with Meta (formerly Facebook), where he is responsible for the SerDes and Systems Interconnect, focusing on next-generation high-speed interfaces. Prior to his current role, Halil served as a Senior Manager at Inphi Corporation where he led the development of 56/112 Gb/s PAM4 DSP chips. Prior to that, he worked on 25 Gb/s Ethernet receivers at Broadcom Inc. and high-speed I/O design at NVIDIA.



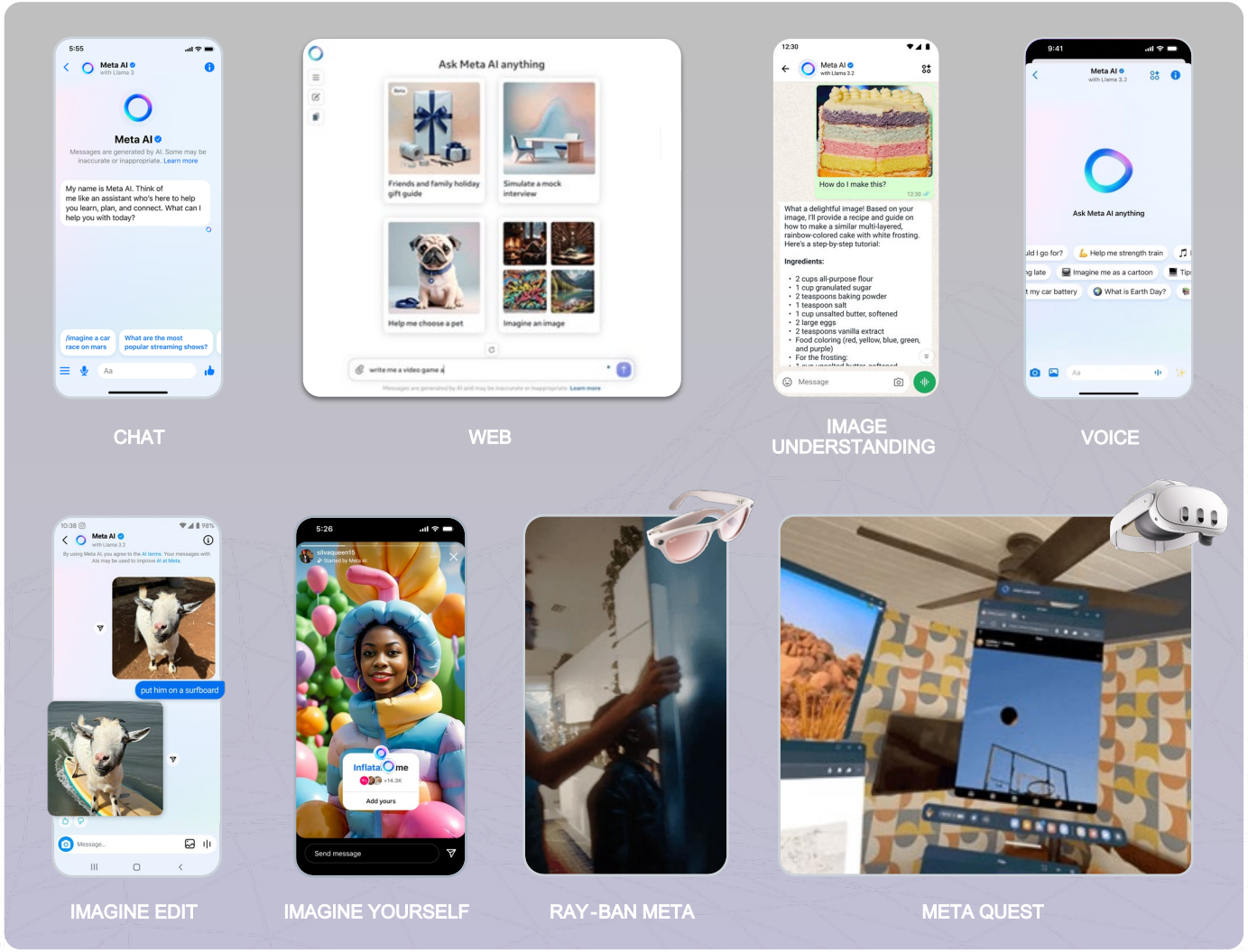
200B+
content recommendations/day



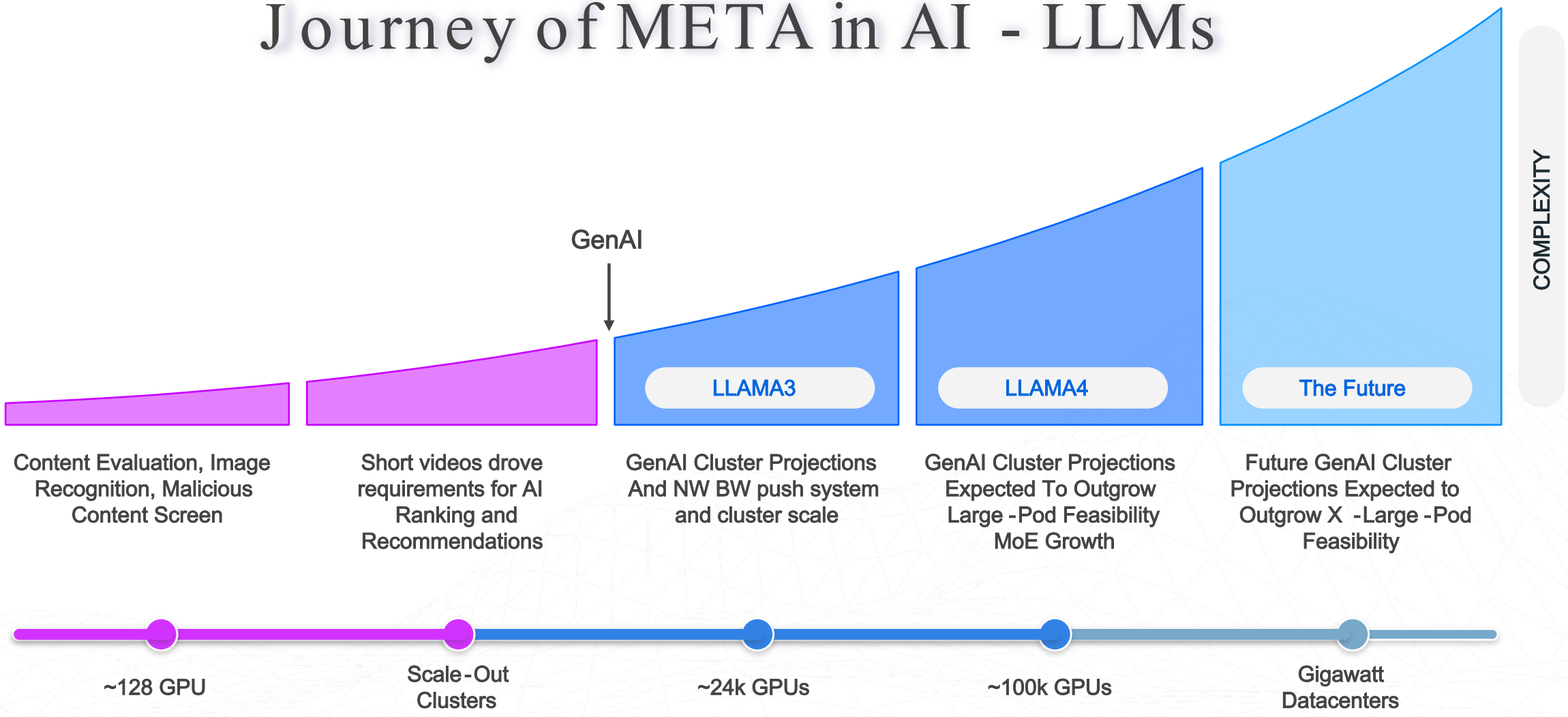
TEF 2025
Ethernet for AI

Built with Llama

Meta AI, launched 1.5 years ago, is available across our apps and on the web. It is more creative, smarter — and used by 1B people every month.

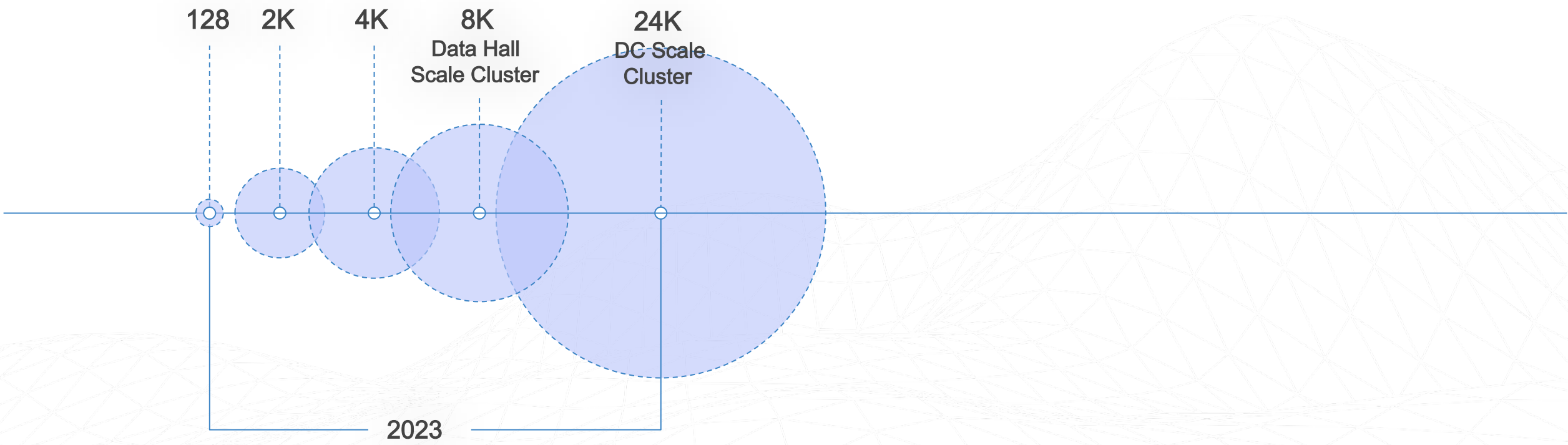


Journey of META in AI - LLMs

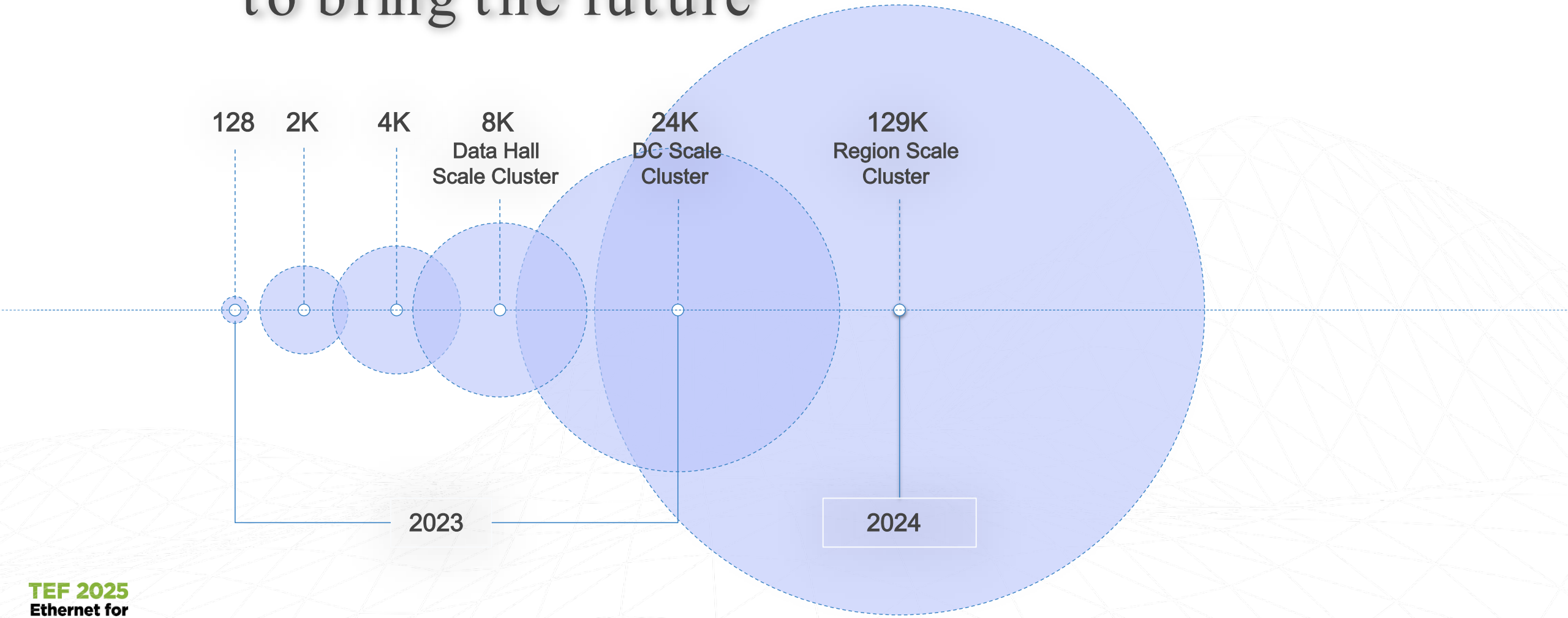


AI SYSTEM INNOVATION

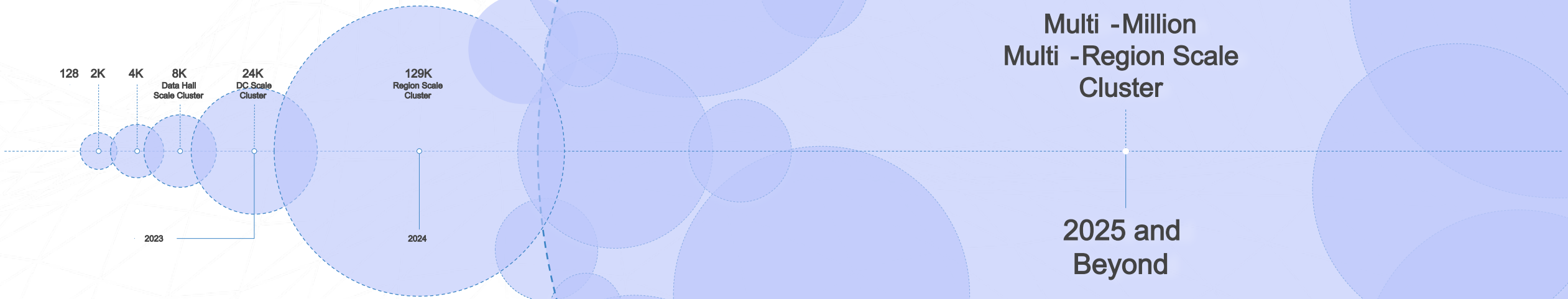
We are building large clusters to bring the future



We are building large clusters to bring the future



We are building large clusters to bring the future



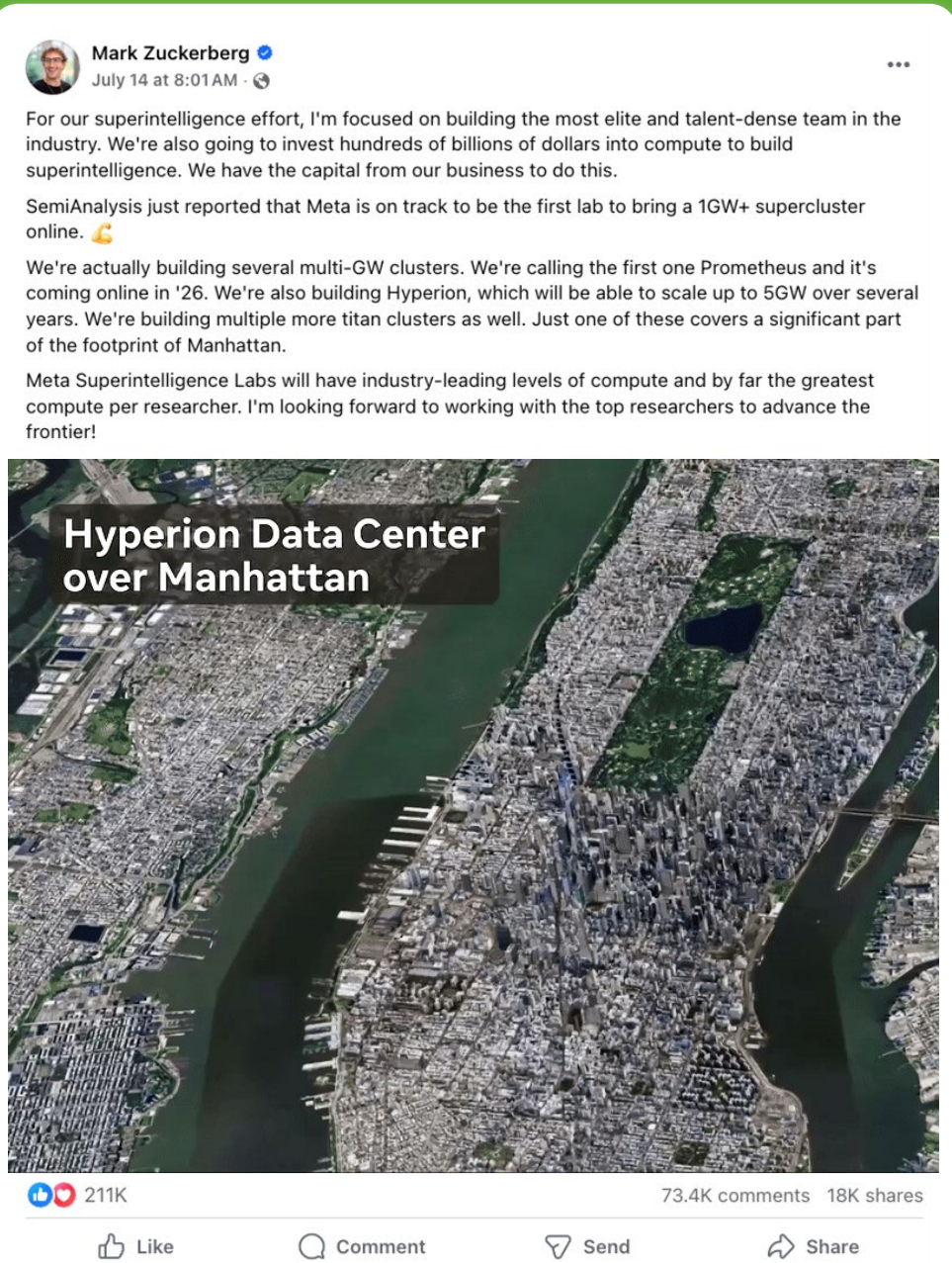
TEF 2025
Ethernet for AI

Talking about the massiveness…

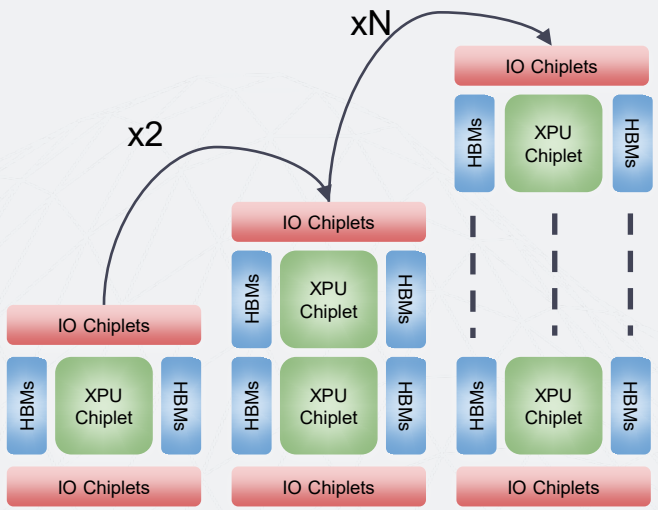
- **Currently Developing - Prometheus** : 1GW+ cluster by 2026
- **Our Big Goal - Hyperion** : 5GW over the next few years

TEF 2025
Ethernet for AI

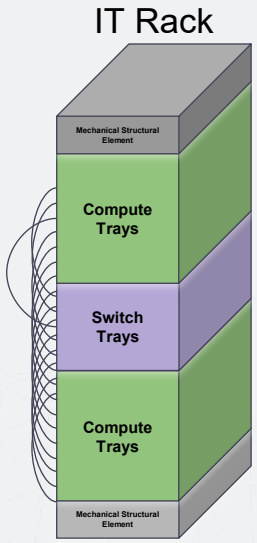
Manhattan shapes our dreams in the skyline; Hyperion powers them in the network. In the end, it's all about our dreams.



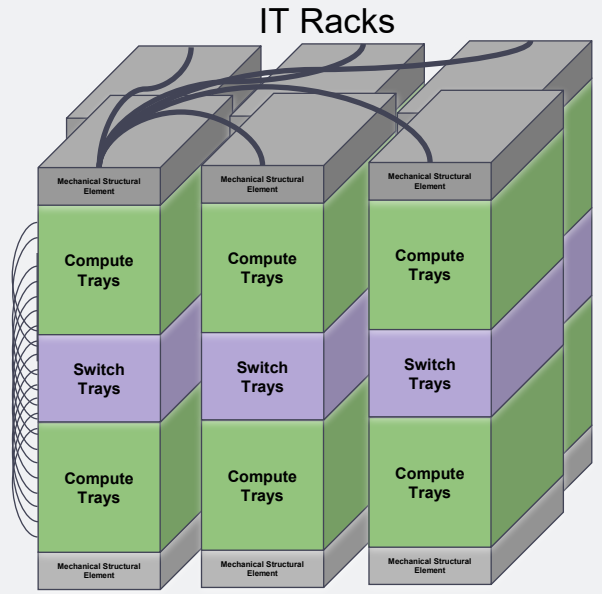
Scaling to Reach Connected Networks



Scale-in
Interconnect
(One Tray)



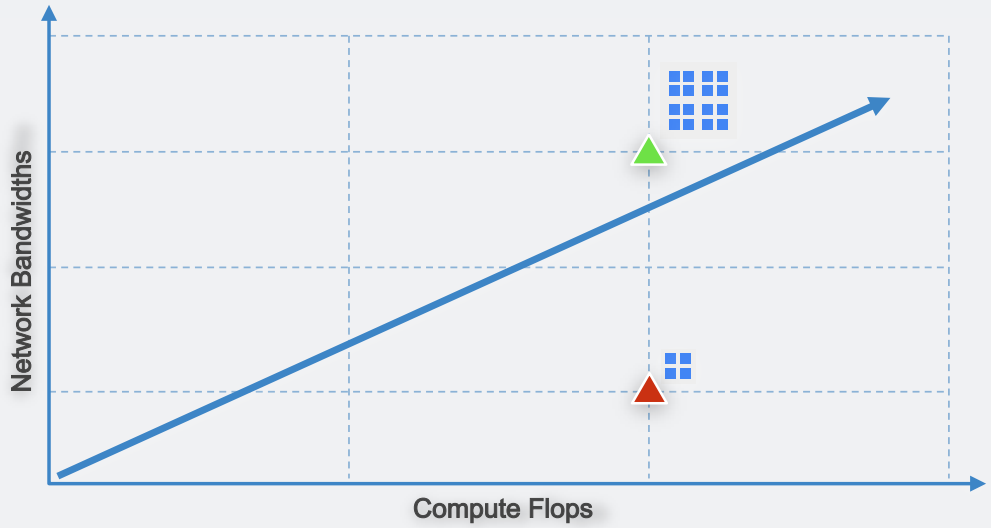
Scale-Up
Interconnect
(Many Trays = One Rack)



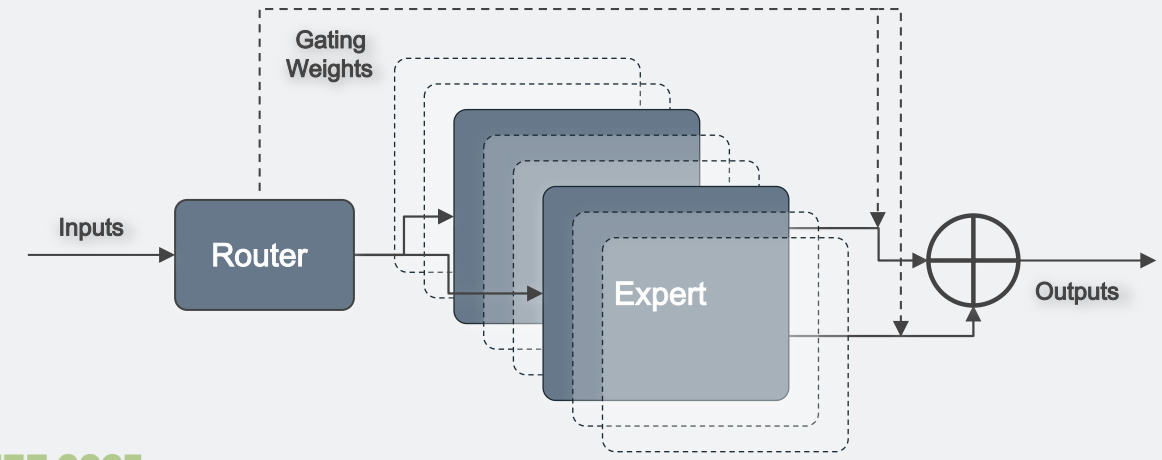
Scale-out
Interconnect
(Many Racks)

TEF 2025
Ethernet for AI

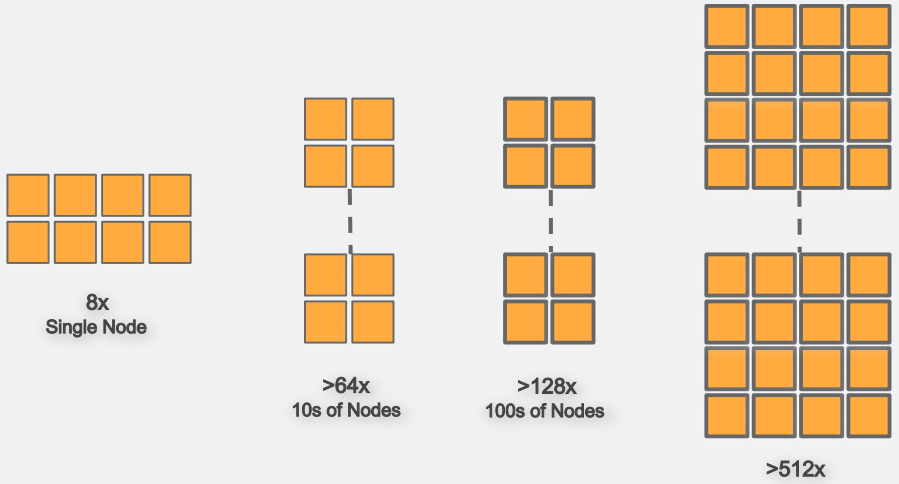
Scale Up domains are important for next generation data centers



Balanced designs prevent stranded resources



Scaling number of experts in LLMs



SIlicon itself is growing in size to have more computing power

TEF 2025
Ethernet for AI

Higher number of accelerators is required for performance

Basic Idea: More nodes in racks?

ORv3 HPR
MP

≤ 72 Accelerators

Cabled Backplane
48 VDC / ± 400 VDC
Air / Liquid Cooled
IT/Power Rack Single -Wide

ORW
2026 -Q3

≤ 144 Accelerators

Cabled Backplane
48 VDC / ± 400 VDC
Air / Liquid Cooled
IT Rack Double -Wide

What is next?
2027 -Q3

≥256 Accelerators

Greater than 900kW
± 400 VDC
Primarily Liquid Cooled
IT Rack Size TBD

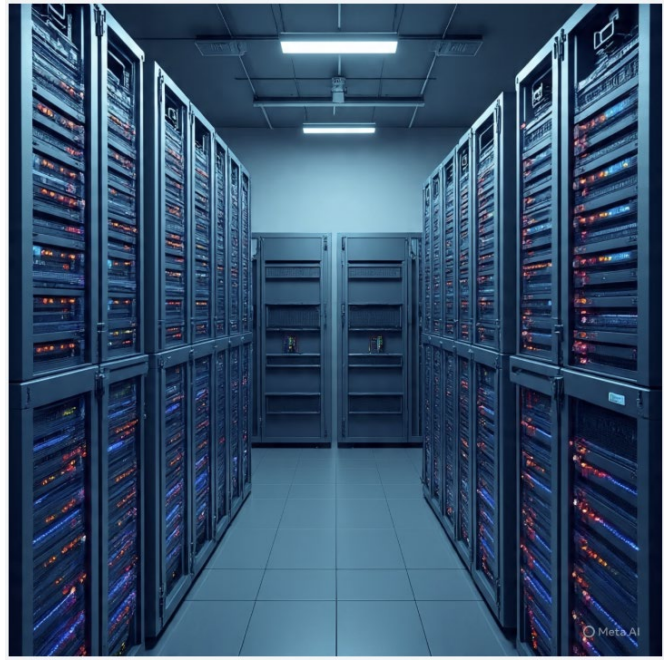
TEF 2025
Ethernet for AI

Bigger racks brings more challenges

Hard to design and manufacture

Expensive to transport

Operationally challenging

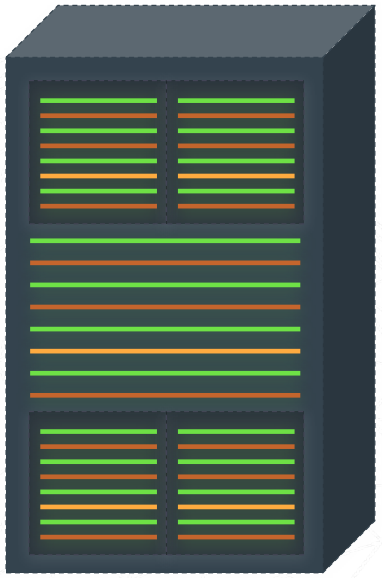


Requiring DC infra upgrades

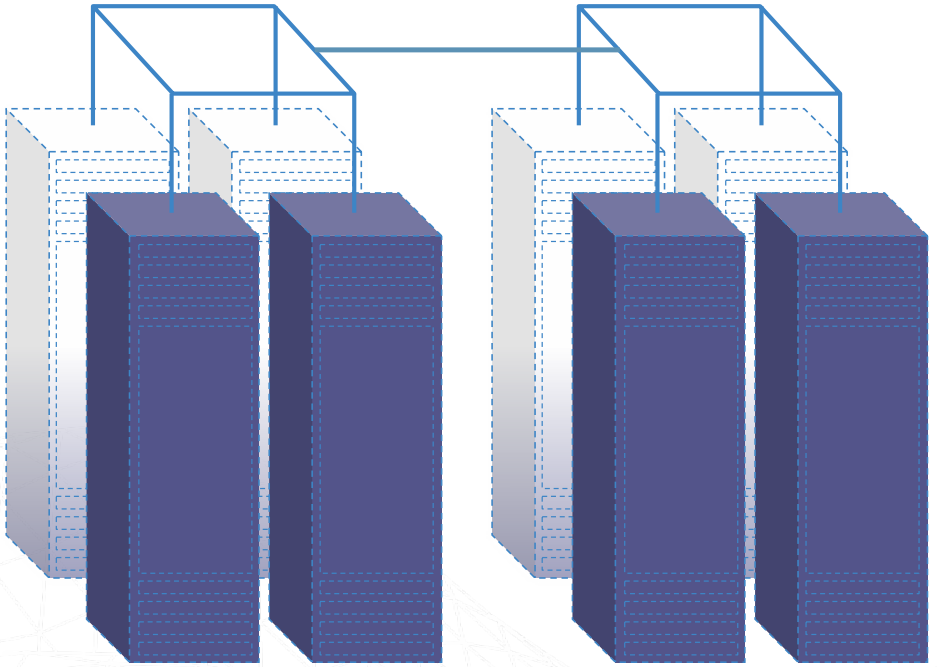
Next Idea: Improving rack scale systems

TEF 2025 Ethernet for AI

The future is disaggregation



BFR
(Backplane View)



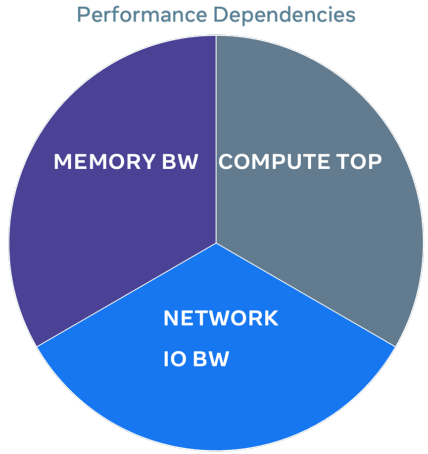
ORV3 ?
Optically interconnected
lower density racks

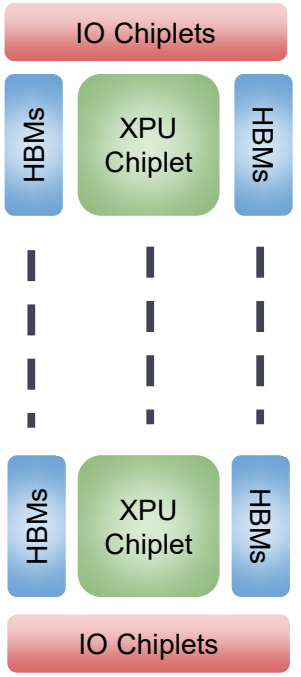
TEF 2025
Ethernet for AI

Next Idea: Improving rack scale systems

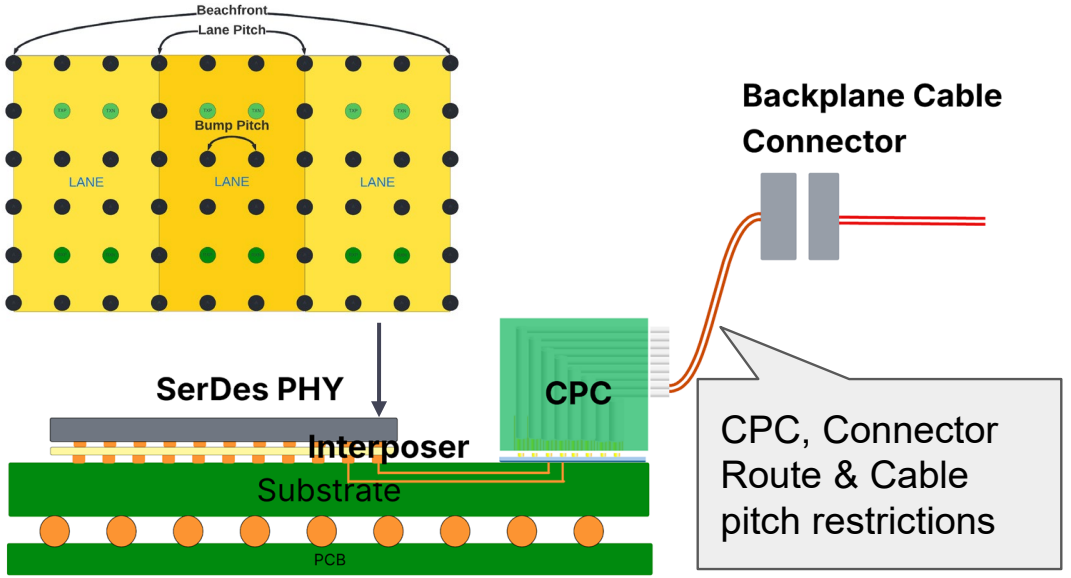
Beachfront Bottleneck of Scale in Systems

- AI system performance is dependent on 3 major pillars which need to improve all together
- Memory and Compute takes full benefit (Linear scaling) of advanced nodes while SerDes design can't
- Beachfront is the major limitation on scaling the Network IO BW





xN Computing
xN Memory BW
x1 Network IO BW



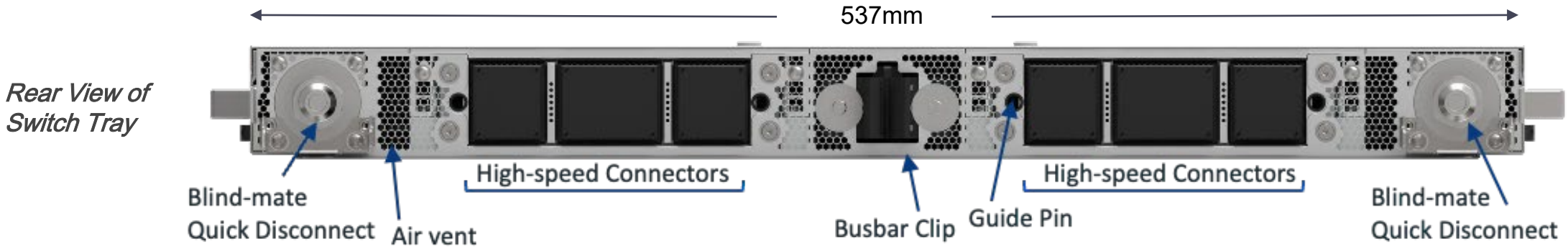
CPC = Co-Packaged Copper

TEF 2025
Ethernet for
AI

Bandwidth Bottleneck in Rack-Scale Systems

Busy Backplane Beachfront

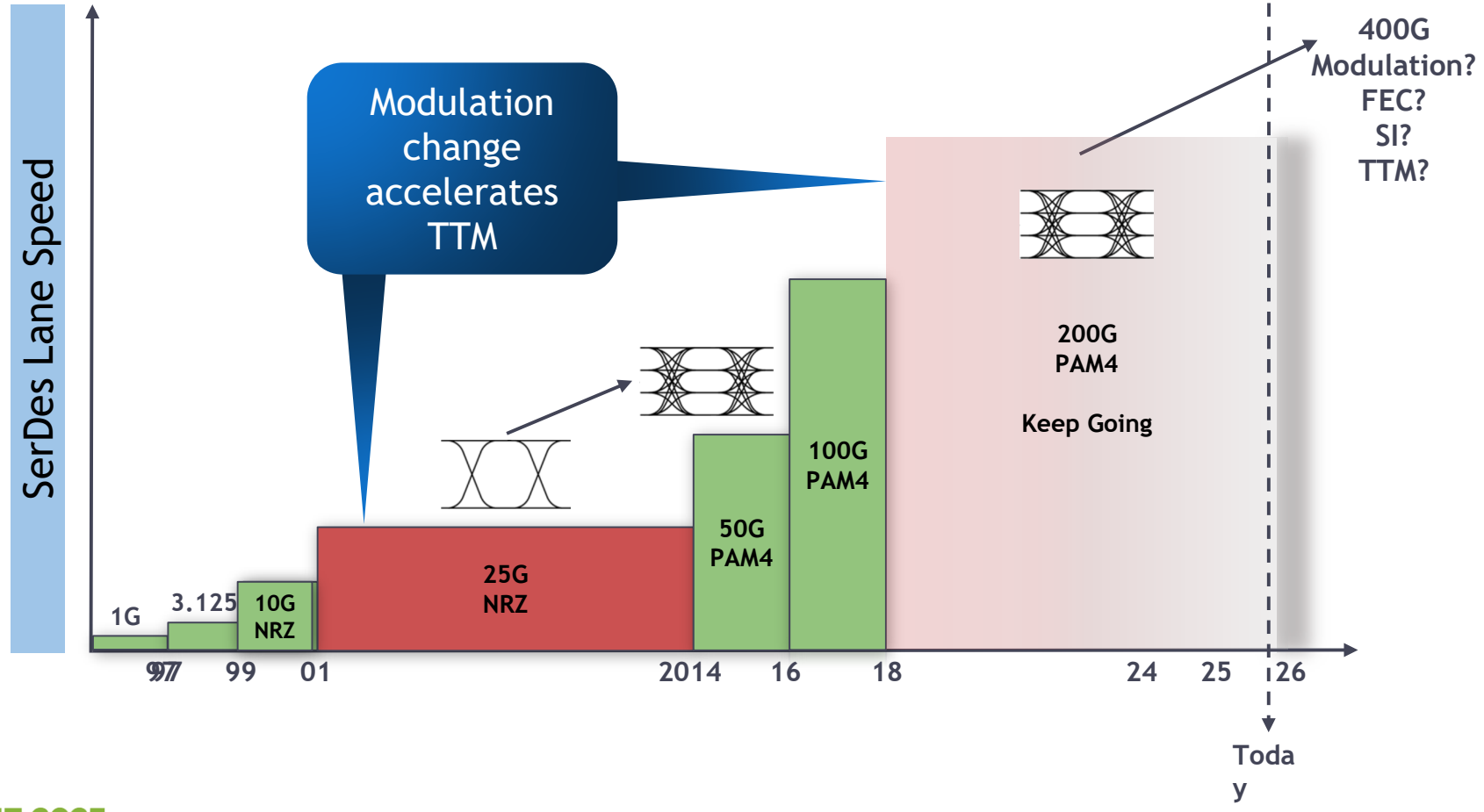
- Crowded backplane beachfront space usage



- 1,024~1,152 differential pairs per OU (ORv3 / 600mm) cross-sectional area for cable backplanes
  - Enables 102.4-115.2T of bandwidth per OU in an ORv3 / 600mm tray with 200G SerDes
  - *Typically a bottleneck for NW switch trays*
- Space required for different needs including connector alignment

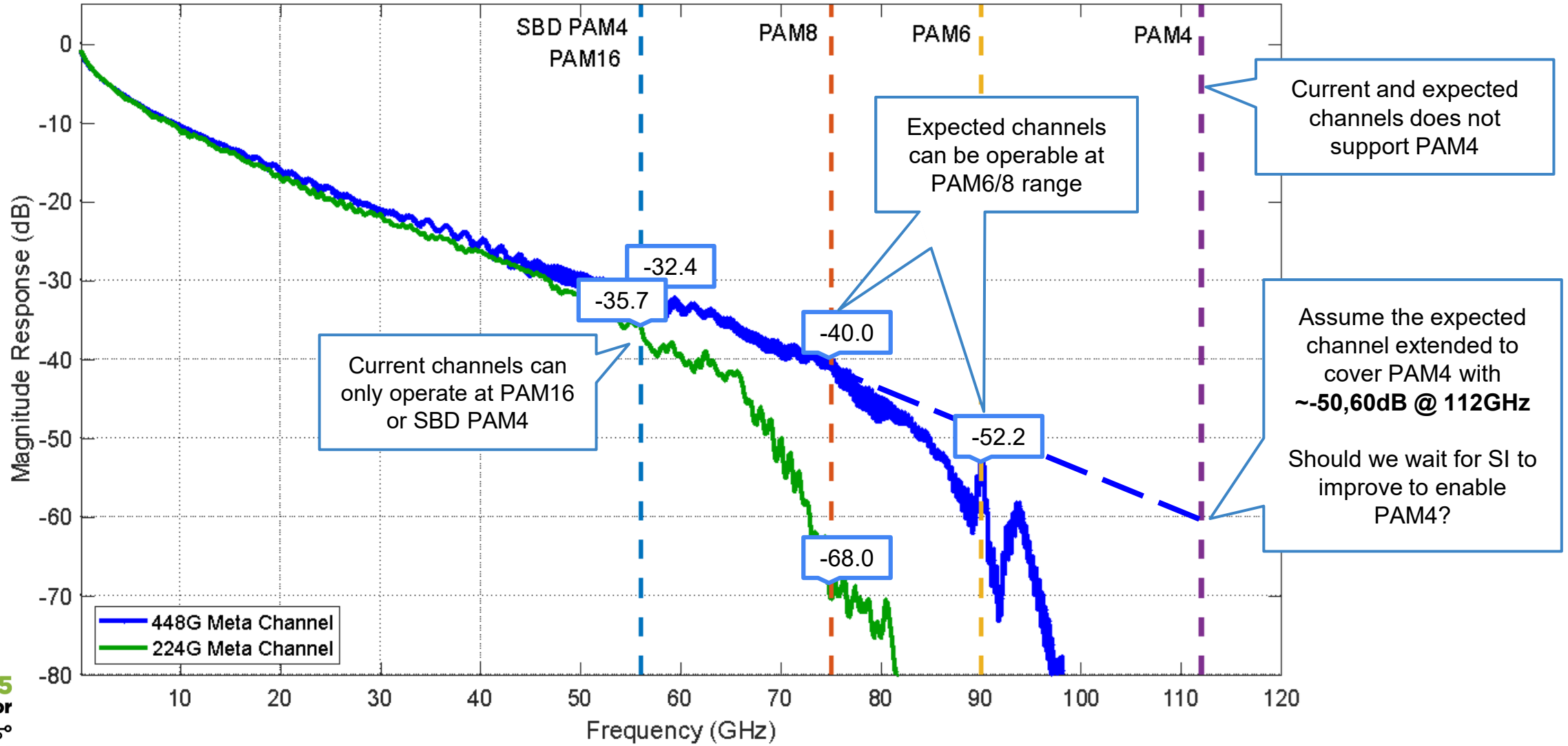
Higher data rate per lane required for aggregate bandwidth scaling

Overcoming Beachfront Bottleneck



- Signaling rate doubled every two years, except 25G and 200G where modulation and FEC are at their limits
- NRZ→PAM4 gave two more generations
- 400G now needs a new breakthrough
- Expected stronger FEC & higher order modulation
- Hyperscalers involvement is needed to accelerate the 400G TTM

BW Problem of Current Connectors & Cables



Accelerating Progress with Collaboration: Vendors Side

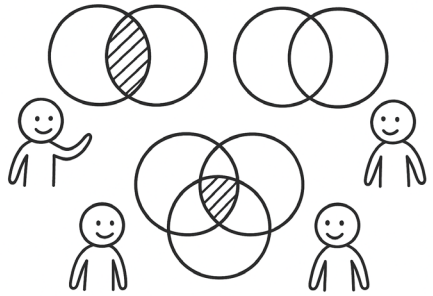
Closing the link WITHOUT a retimer

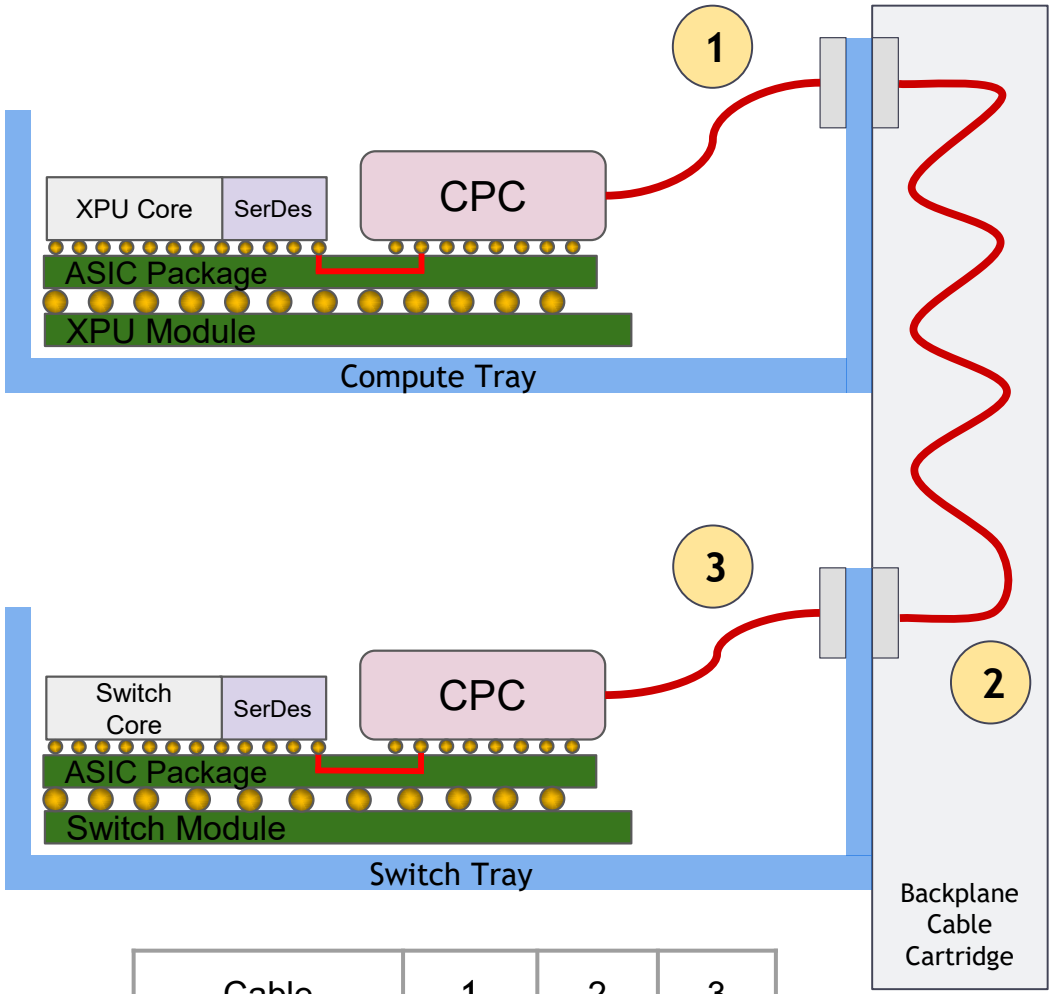
- Low power consumption
- Manageable latency

Fostering close active collaboration among

- SerDes vendors
- Package manufacturers
- Connector manufacturers
- Cable suppliers
- Rack architects

is highly needed





Cable	1	2	3
≈Length (cm)	15	140	70

Openness Wins: Standards Side Works

- Standards-based framework
- Open discussions
- Data sharing environment

Openness over Lock -in

Compatibility First

Enable the Stack

IEEE Advancing Technology for Humanity

ethernet alliance

ESUN

ULTRA ACCELERATOR LINK™

Ultra Ethernet Consortium

OIF

TEF 2025 Ethernet for AI

Conclusion

AI-Driven Data Center Growth is Happening FAST

- AI integration is accelerating across daily life
- Hyperscalers are expanding data centers for greater scale and efficiency
- Rising AI demand pressures interconnects to scale:
  - In, Up, Out, Accross

Need for Next-Gen Speeds

- Transition to 400G essential:
  - Electrical communication
  - Optical communication
- Not only for enabling higher performance, but for better efficiency, and improved cost-effectiveness

TEF 2025 Ethernet for AI

QUESTIONS?