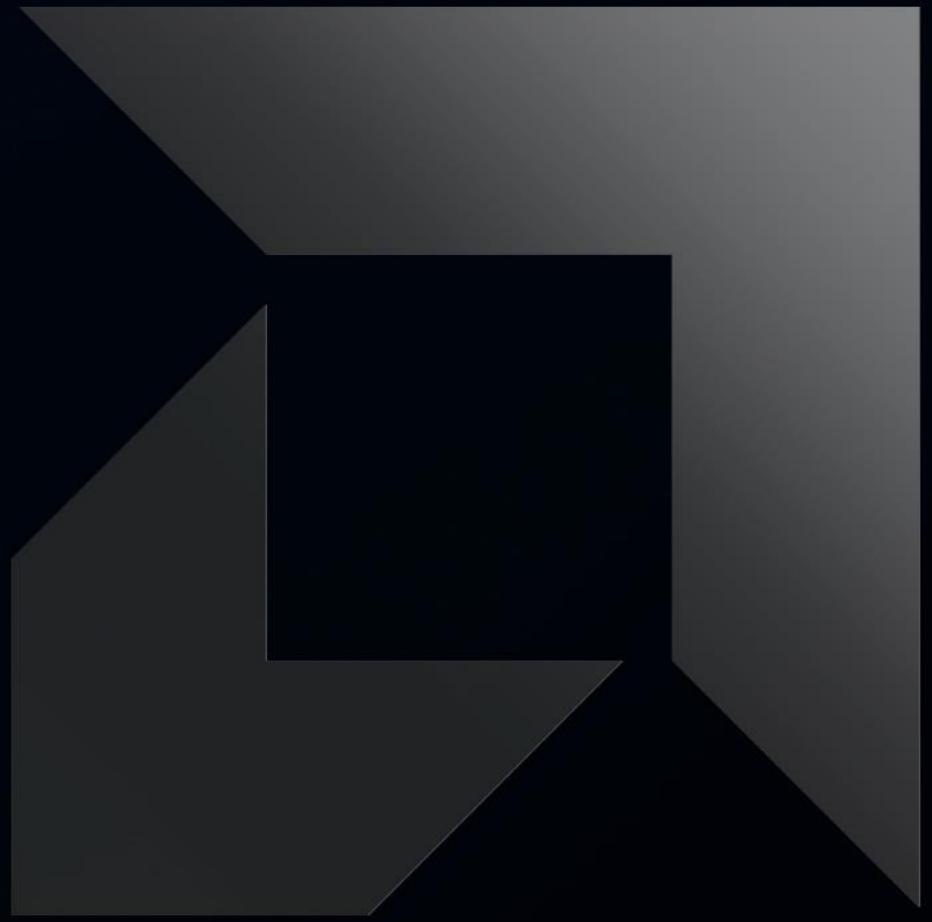


The Evolution of Ethernet Fabrics for HPC & AI with Ultra Ethernet

HOTI2023 (August 24, 2023)

Robert Hormuth | AMD Corporate Vice President, Datacenter Solutions Group Architecture and Strategy

AMD

together we advance_

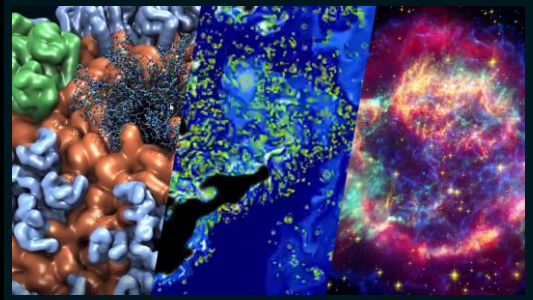
This presentation contains forward-looking statements concerning Advanced Micro Devices, Inc. (AMD) such as the features, functionality, performance, availability, timing and expected benefits of AMD Instinct™ accelerators; AMD Instinct™ MI300A accelerators; AMD Instinct™ MI300X accelerators; AMD EPYC™ “Siena” CPU; AMD DPU, GPU and CPU architecture roadmaps; and expected AMD data center energy efficiencies, which are made pursuant to the Safe Harbor provisions of the Private Securities Litigation Reform Act of 1995. Forward-looking statements are commonly identified by words such as "would," "may," "expects," "believes," "plans," "intends," "projects" and other terms with similar meaning. Investors are cautioned that the forward-looking statements in this presentation are based on current beliefs, assumptions and expectations, speak only as of the date of this presentation and involve risks and uncertainties that could cause actual results to differ materially from current expectations. Such statements are subject to certain known and unknown risks and uncertainties, many of which are difficult to predict and generally beyond AMD's control, that could cause actual results and other future events to differ materially from those expressed in, or implied or projected by, the forward-looking information and statements. Investors are urged to review in detail the risks and uncertainties in AMD’s Securities and Exchange Commission filings, including but not limited to AMD’s most recent reports on Forms 10-K and 10-Q. AMD does not assume, and hereby disclaims, any obligation to update forward-looking statements made in this presentation, except as may be required by law.

AMD does not assume, and hereby disclaims, any obligation to update forward-looking statements made in this presentation, except as may be required by law.

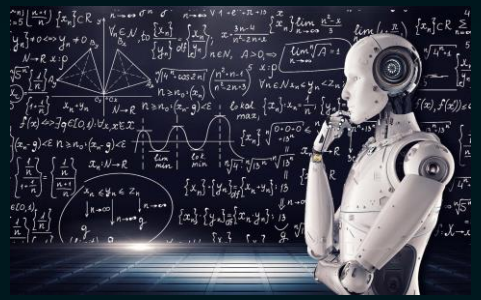
Future Computing Scenarios 2030+

Deeper Perception

Real World Simulation & Exploration



Smarter and more Inclusive AI



Enhanced User Experiences

Data-driven Business Innovation

Sustainability & Efficiency

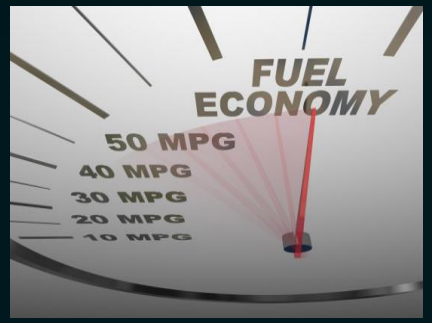
Subscription Economy & Institutional Disaggregation

Data Sovereignty & Security

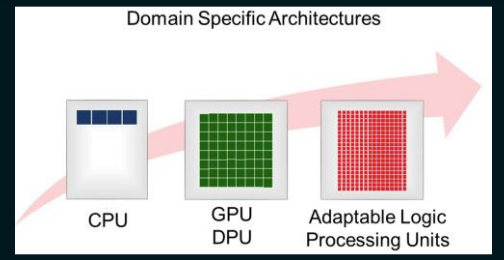
Technology Requirements for Future Compute Scenarios

High Performance

Efficiency



Optimized Architectures

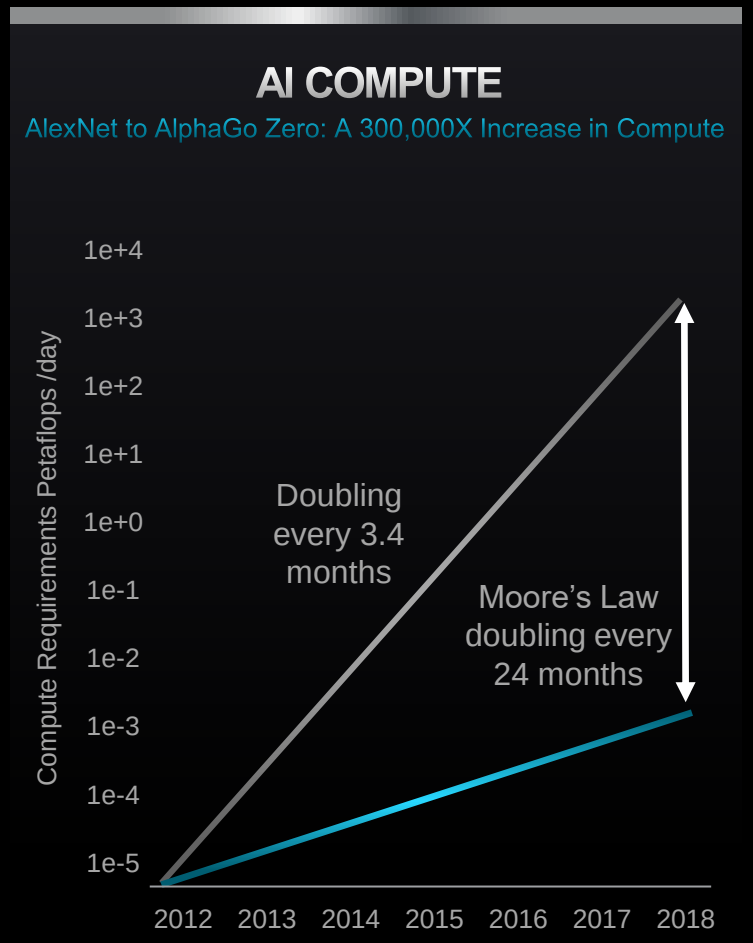


Software & Ecosystem

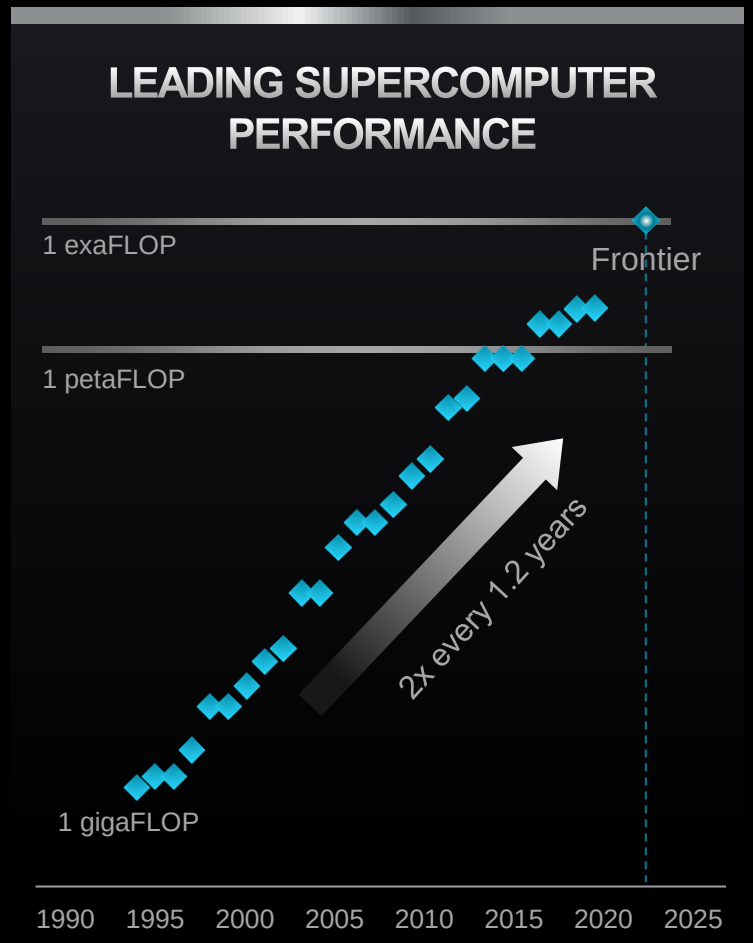
Security

Networking

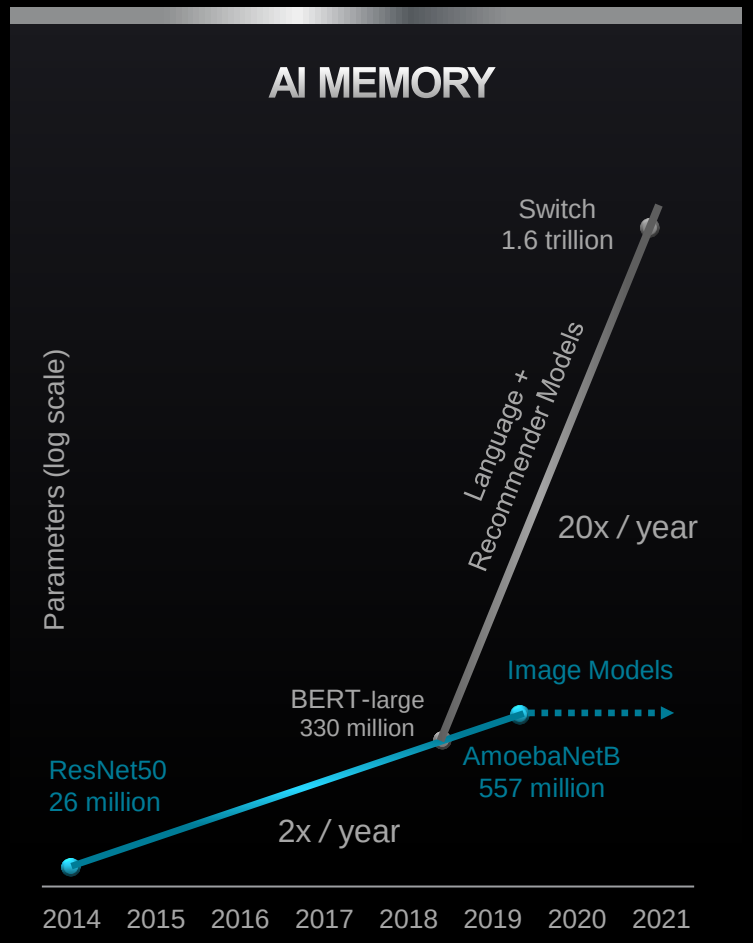
AI & HPC Driving Data Center Compute & Memory Growth



Source: https://blog.openai.com/ai-and-compute



Source: https://www.top500.org



Source: TechInsights (2022)

HPC and AI are driving unique challenges

OPTIMAL EFFICIENCY THROUGH DOMAIN-SPECIFIC MODULAR FABRICS

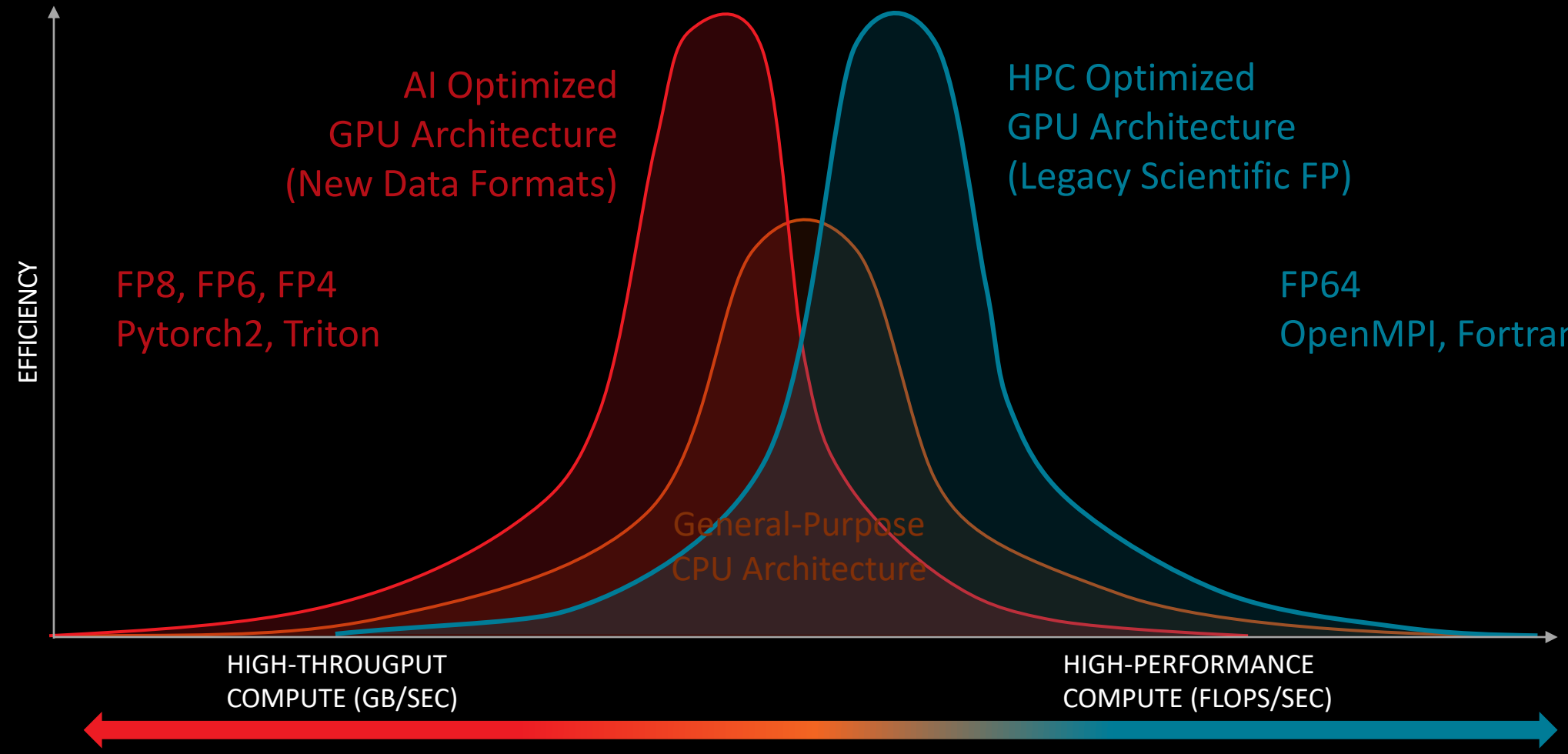


Chart for illustrative purposes

The Future of Computing is Heterogeneous

General Purpose Compute	Adaptive SoC / FPGA	Accelerators	GPU Compute	Custom SoC	ASIC
					
AMD Ryzen™ and **AMD EPYC™ CPUs**	AMD Versal™ Adaptive SoCs	AMD Pensando™ DPU AMD Alveo™ Accelerators	**AMD Instinct™ Accelerators**	AMD custom gaming SOCs	**Various AI/ML Compute**

ADAPTIVE | **SPECIALIZED**

AMD EPYC™ Family

"GENOA"

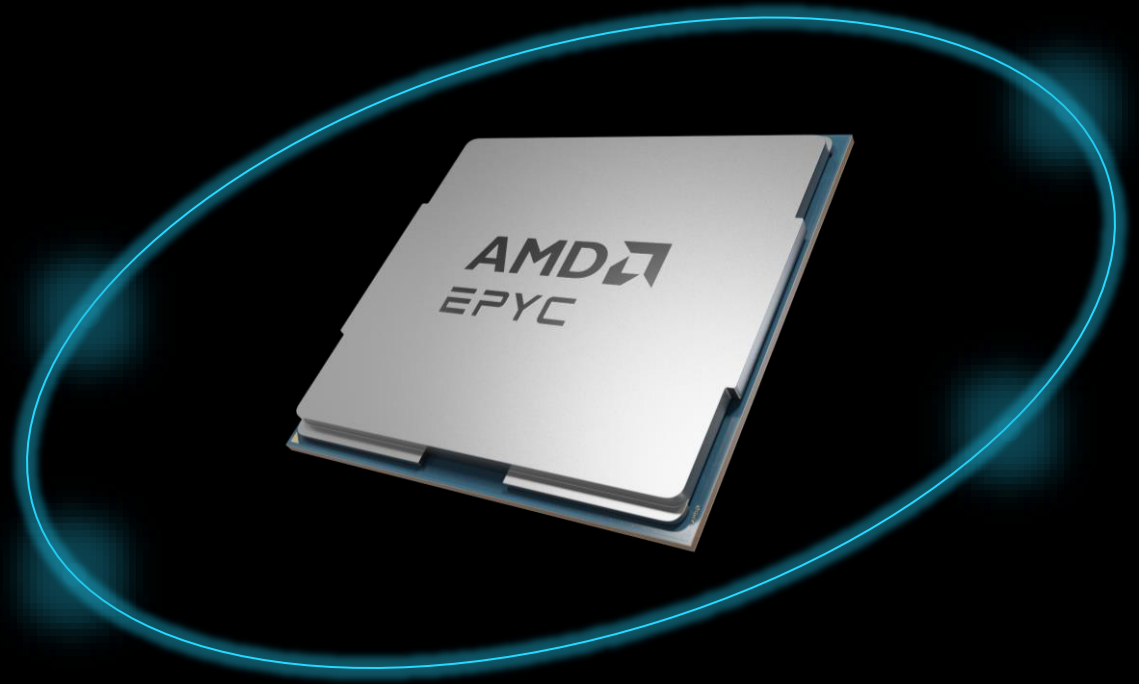
Leadership Performance & Performance per Core

WORKLOADS: BROAD SPECTRUM

"BERGAMO"

Highest Thread Density

WORKLOADS: CLOUD NATIVE & FLOP INTENSIVE



"GENOA-X"

Highest Cache

WORKLOADS: TECHNICAL COMPUTE

"SIENA" – 4Q23

Perf per Watt Optimized

WORKLOADS: LOW POWER FORM FACTORS

Leadership 5nm Process Node

Up to 128 High-Performance "Zen4" / "Zen4c" Cores

Common ISA & software

Validated with AI ISV Solutions

ThirdAI

PCIe® 5.0 & CXL™ Memory Expansion

Advanced Security Features

AMD Instinct™ Family

SAMPLING NOW

"MI300A"

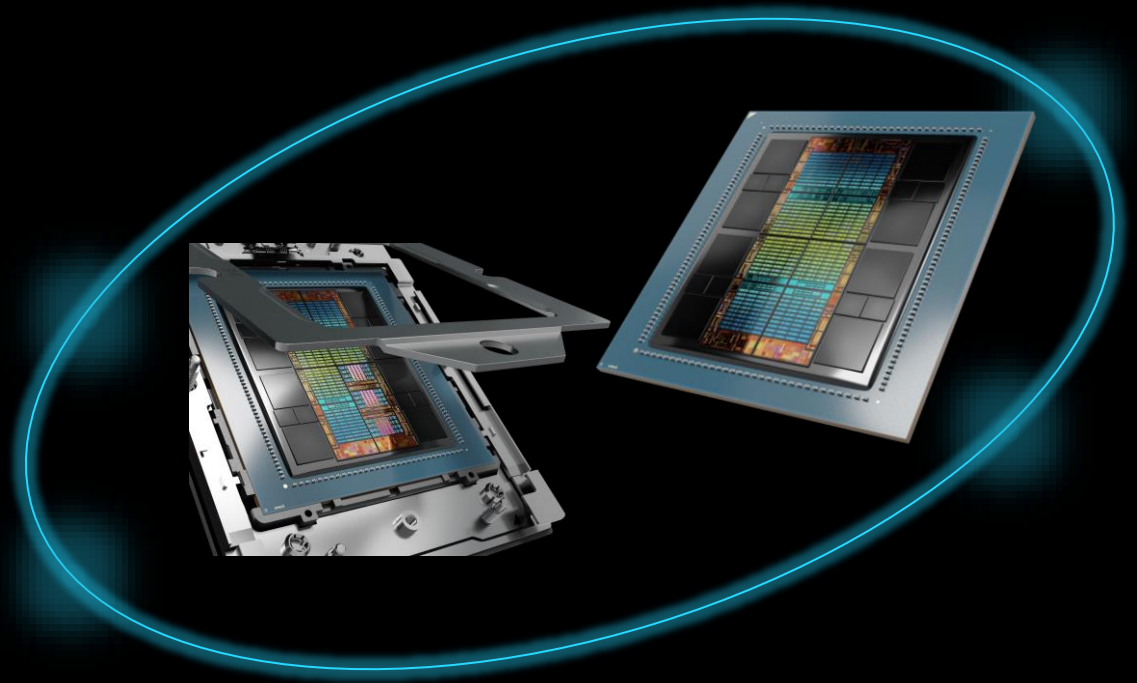
World's first APU accelerator for AI and HPC

ZEN 4

24 CPU Cores

128 GB HBM3

Shared Memory
CPU + GPU



"MI300X"

Leadership generative AI accelerator

5.2 TB/s Memory Bandwidth

192 GB HBM3

896 GB/s Infinity Fabric™ Bandwidth

AMD CDNA 3 Next-Gen Accelerator Architecture

5nm and **6**nm Process Technology

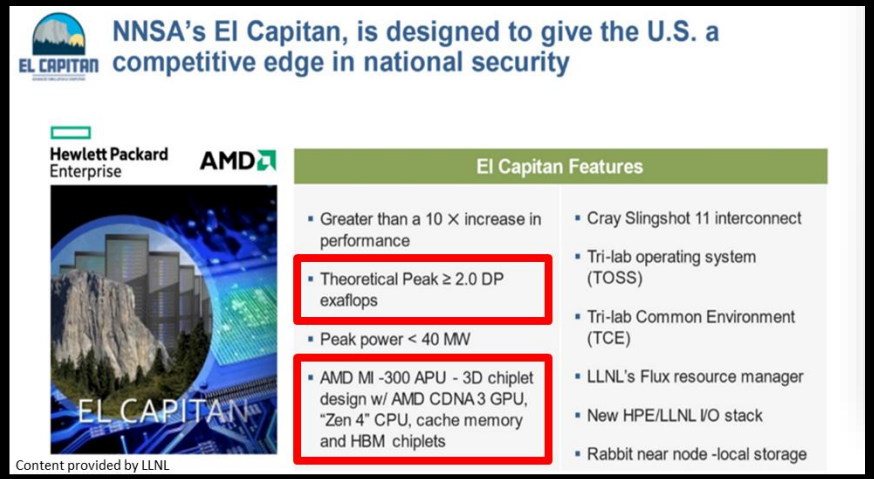
AMD ROCm

[Public]

AMD
EPYC | INSTINCT

LEADING THE EXASCALE ERA

- Powering World's #1 Supercomputer
  First to break Exascale barrier
- Powering 70% of World's Top 10 Green Supercomputers
  7 of top 10 most efficient systems rely on AMD
- Powering World's #1 HPL-MxP Supercomputer
  9.95 Eflops on HPL-MxP Mixed-Precision Benchmark



Source: TOP500, Green500 and HPL-MxP lists, as of June 2023

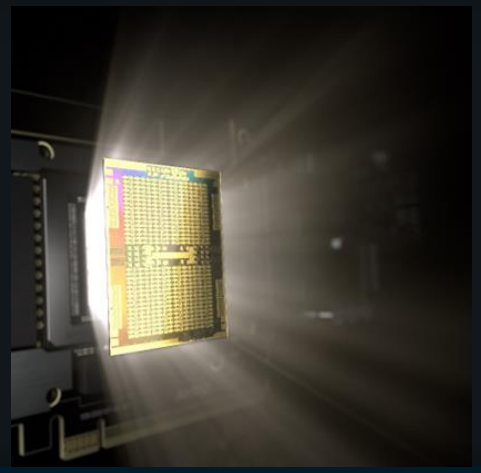
AI/HPC Fabric Classification

Frontend Network

- Traditional Data Center Network
- Ethernet + Every Protocol
- TCP (Internet scale)
- Software Defined Networking

GPU Scale-out RDMA

- High Bandwidth
- Ethernet + MPI Transport
- HPC Scale (100K Nodes)
- New topologies (optimize hops)

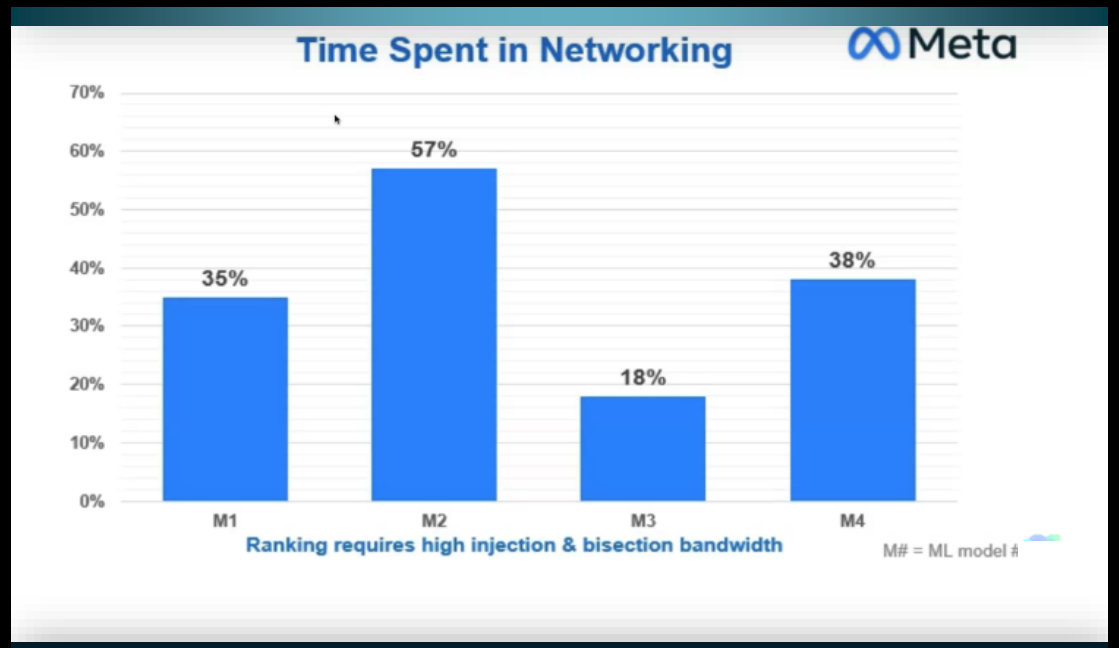


GPU Rack Scale

- Ultra Bandwidth (Memory Speeds)
- DMA & Load/Store
- Extended Node Scale (low latency)
- Next step in NUMA for Memory Scale

Ethernet Enables Advanced Distributed Systems

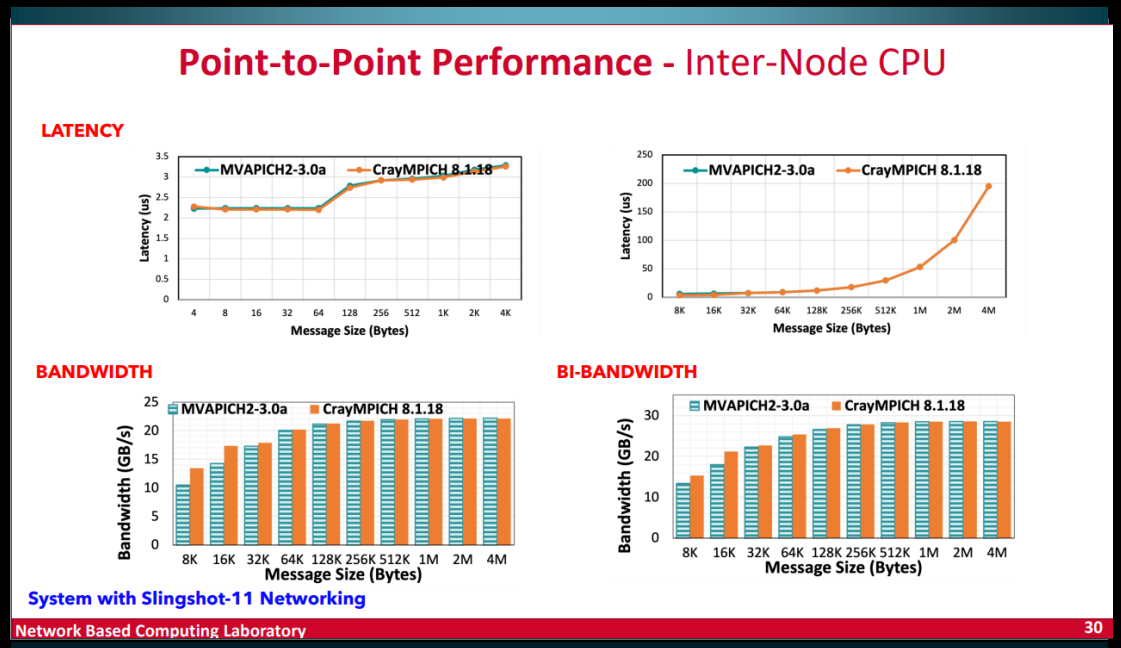
Distributed System Fabric Evolution



GPU Scale-out AI

- Fast moving codebase
- Network infrastructure BW greater than HPC
- New balance of compute, network and memory BW
- New benchmarks

https://youtu.be/miv5PExXTmc?t=782

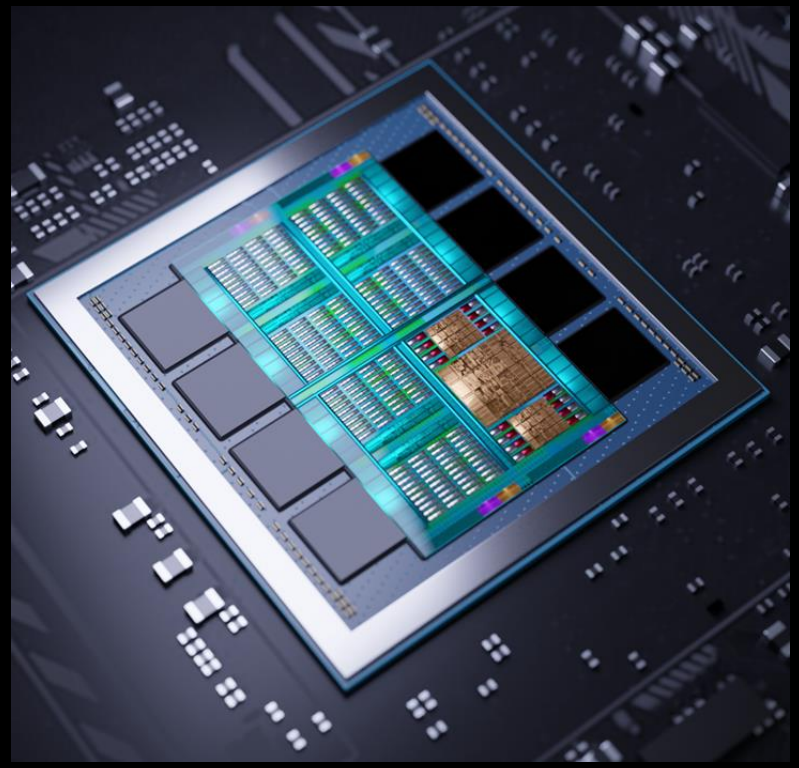


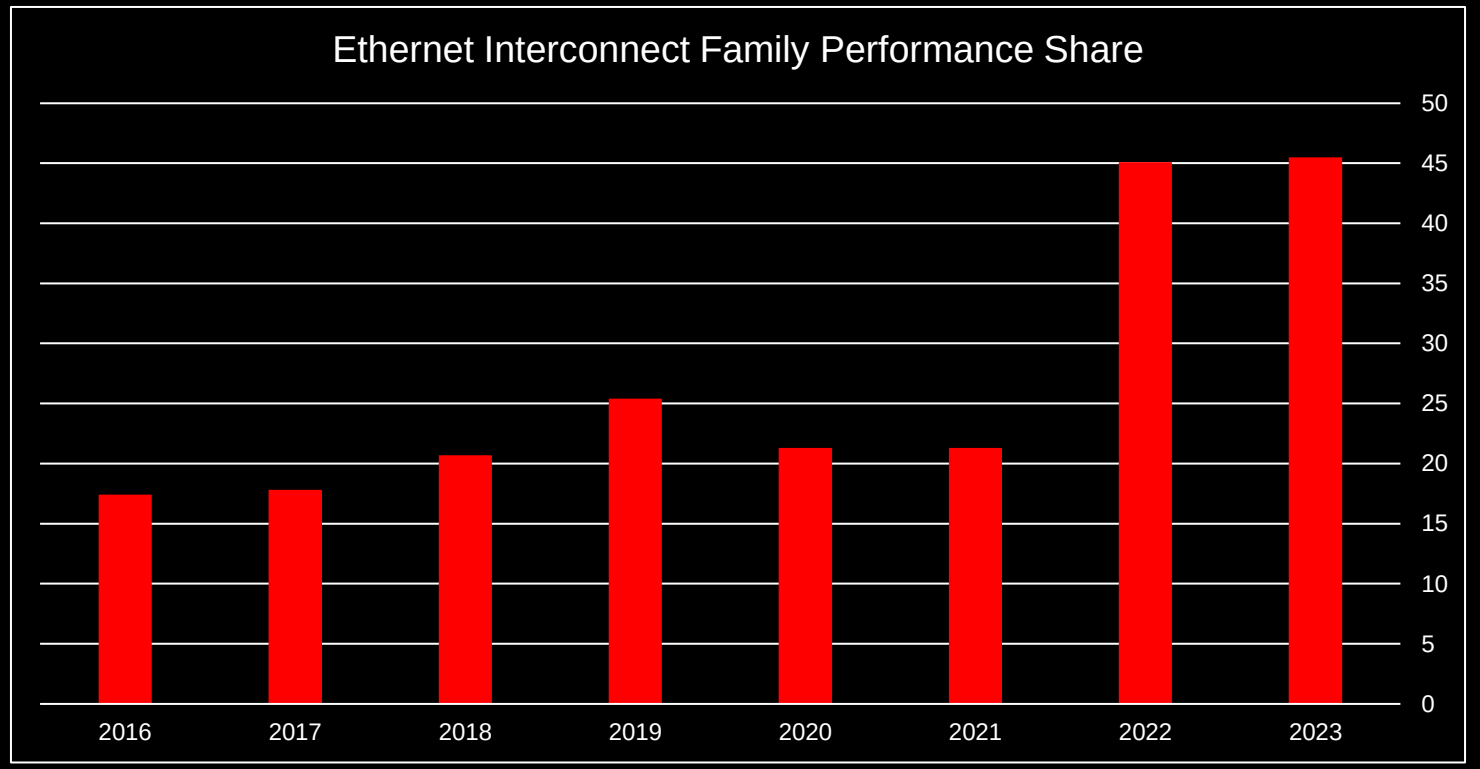
GPU Scale-out HPC

- Decades of highly tuned codes
- Dynamic communication patterns
- MPI Infrastructure
- OSU Benchmarks capture dependency for small messages

https://mvapich.cse.ohio-state.edu/static/media/talks/slide/kawthar-slingshot-osu-booth-sc22_2.pdf

Growth of Ethernet in HPC





Source:
https://www.top500.org
https://top500.org/statistics/list/

Ethernet is #1 in TOP500 list

+

Ethernet is prominent for Datacenter Networks

The Grand Challenges Ahead

1 Limited Scale

2 Network Under-Utilization

3 High Tail Latencies

4 Delayed Congestion Response

5 Fragmented Telemetry Schemes

6 Rigid Ordering Requirements

Ultra Ethernet Consortium

An Ethernet-based, open, interoperable, high performance, full-communications stack architecture to meet the growing network demands of AI & HPC at scale

J Metz, Ph.D
Chair, Ultra Ethernet Consortium

INTRODUCING: THE PROMISE OF ULTRA ETHERNET

https://*ultra*ethernet.org/

THE NEW ERA NEEDS A NEW NETWORK

Ultra Ethernet

As **performant** as a supercomputing interconnect

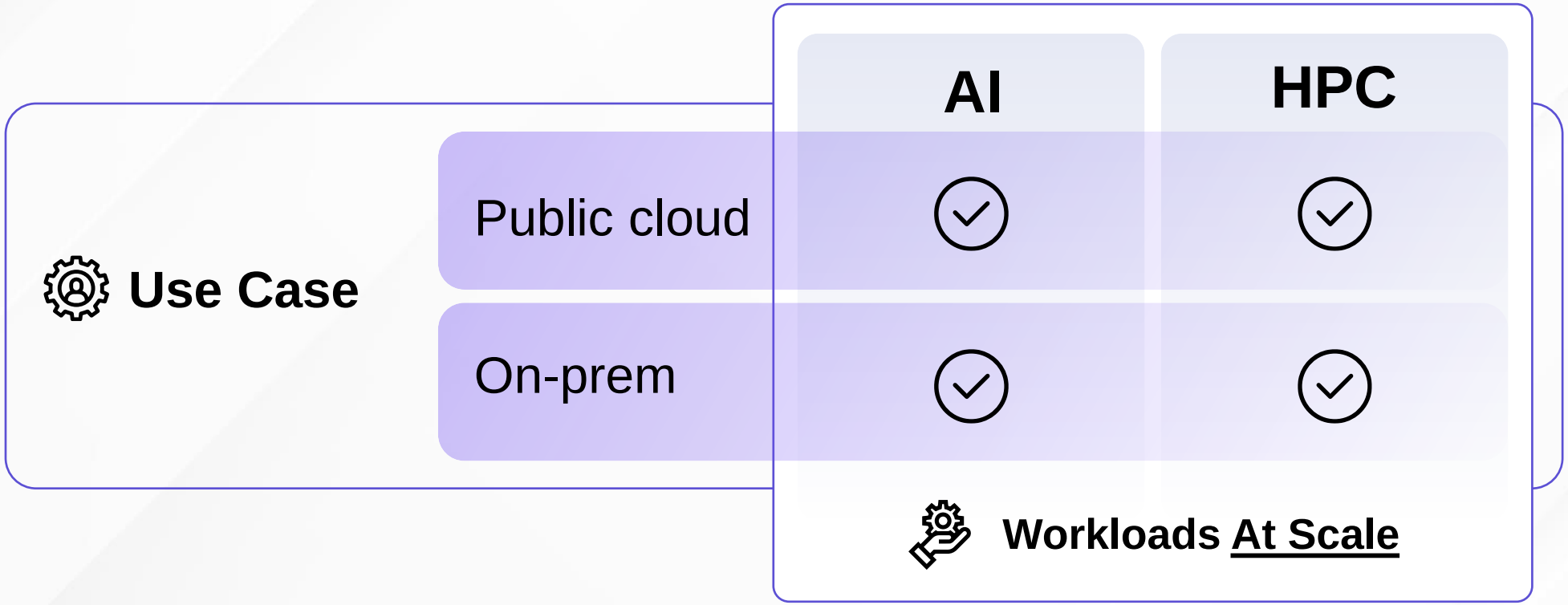
As **ubiquitous** and **cost-effective** as Ethernet

As **scalable** as a cloud data center

FOUNDING MEMBERS

AMD
ARISTA
BROADCOM
CISCO
EVIDEN an atos business
Hewlett Packard Enterprise
intel
Meta
Microsoft

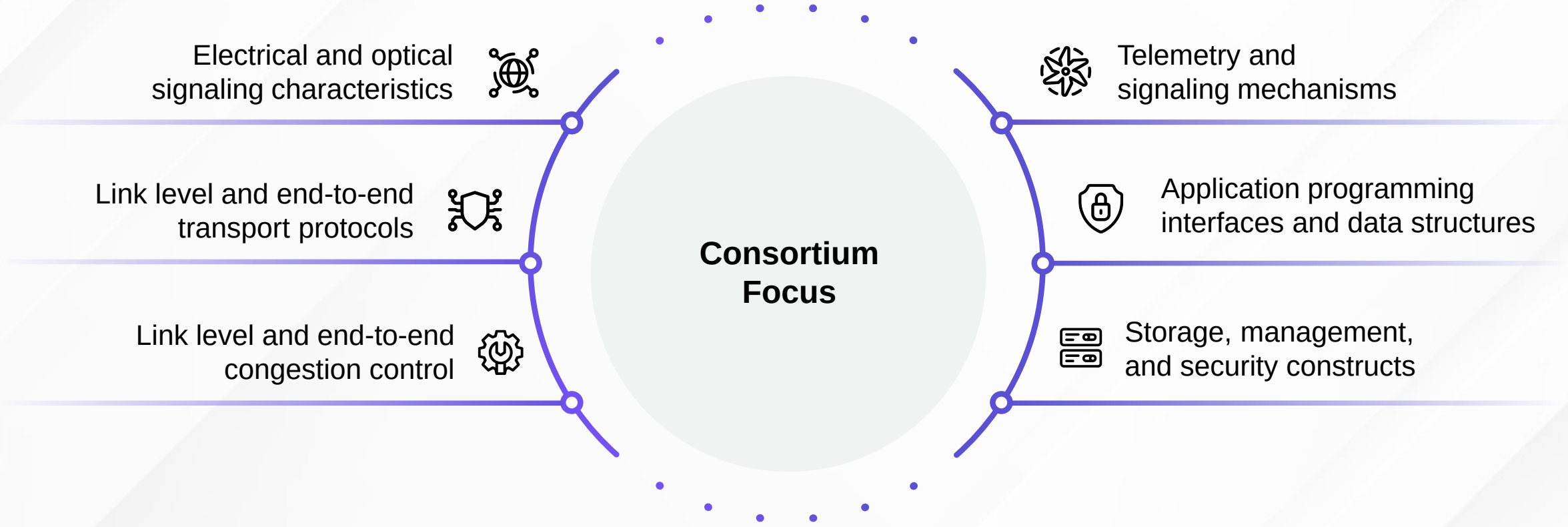
TARGET DEPLOYMENT MODELS / USE CASES



Profiles defined for AI and HPC use cases

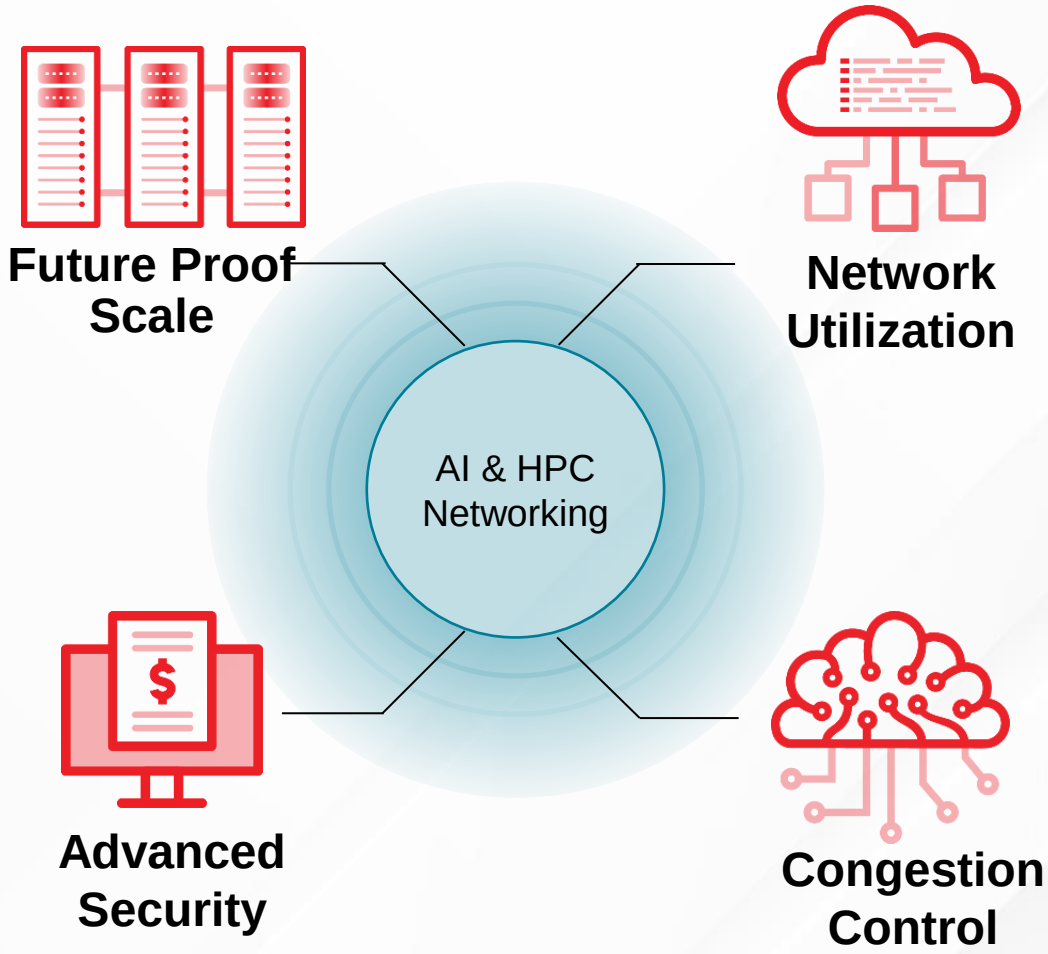
TECHNICAL GOALS

Open specifications, APIs, source code for optimal performance of AI and HPC workloads at scale.



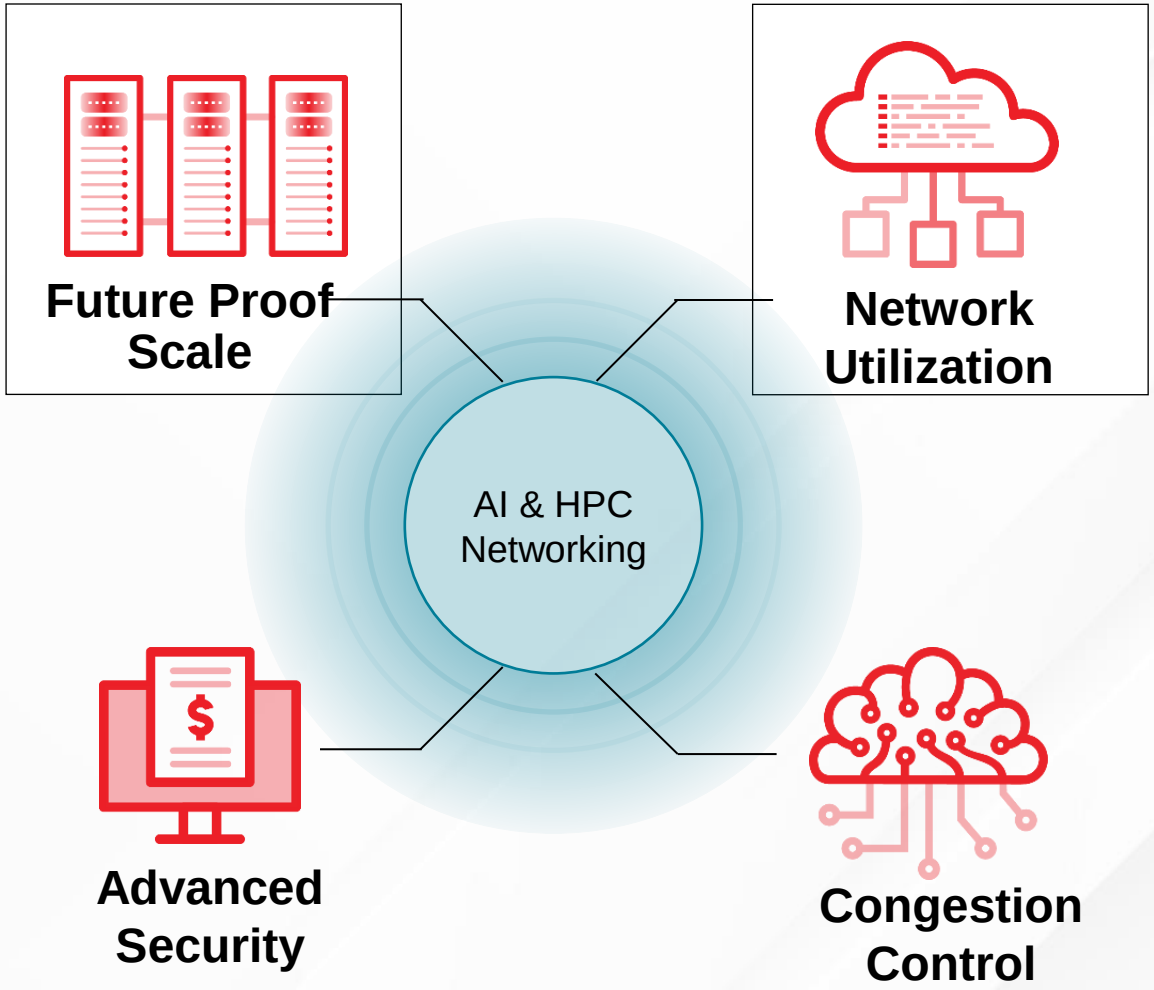
UEC TRANSPORT ADDRESSES GRAND CHALLENGES

- Future proof system scale with 1M endpoints
- Improved network utilization with multi-path routing
- Lower tail latency with flexible packet ordering
- Security built-in from the beginning
- AI and HPC congestion control require faster response times
- End-To-End telemetry provides improved network visibility



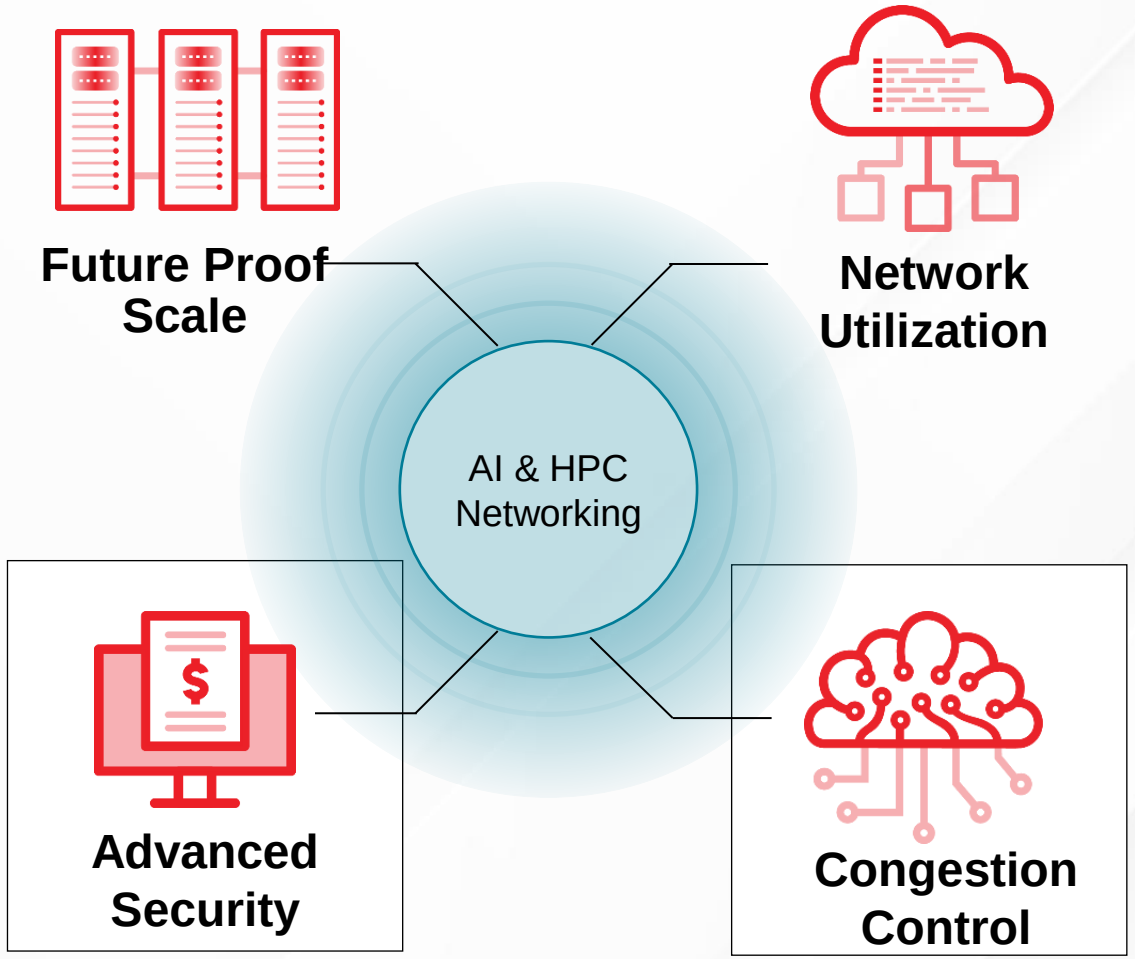
FUTURE PROOF SYSTEM SCALE & NETWORK UTILIZATION

- Determinism and predictability become more difficult as systems grow
  - New methods needed to achieve holistic stability & visibility
- Next phase in the technology evolution is for every flow to simultaneously use all paths to the destination, a technique known as “packet spraying”
  - Every flow simultaneously access all paths
  - Achieves more balanced use of entire network
- From Rigid to Flexible Ordering
  - Rigid ordering enables "go-back-n" recovery and in-order delivery, but restricts network utilization and increases tail latencies
  - Flexible ordering enables packet-spraying in bandwidth-intensive collective operations; eliminates to reorder packets
  - Supports modern APIs that relax the packet-by-packet ordering requirements for applications where it’s critical to curtail tail latencies



ADVANCED SECURITY, CONGESTION CONTROL & TELEMETRY

- Advanced Security
  - Encryption support that doesn't balloon the session state in hosts and network interfaces
  - Similar conditions in AI and HPC
- Congestion
  - Must work with packet spraying
  - Must coordinate with scheduling algorithms on sending host
- Telemetry
  - Congestion information originating from the network can advise the participants of the location and cause of the congestion
  - Robust end-to-end telemetry enables optimized congestion control algorithms
  - Shortening the congestion signaling path and providing more information to the endpoints allows more responsive congestion control



LEARN MORE AND JOIN THE MOVEMENT AT

www.ultraethernet.org 

AMD and UEC

“At AMD, we are proud to be founding members and key contributors of the UEC. We believe it is the natural progression of our core AMD products, like AMD EPYC™ server processors and AMD Instinct™ MI Series accelerators as well as our growing Alveo SmartNIC portfolio. With all these great products to choose from, we need a robust, tuned network to connect them. “ – Mark Papermaster | EVP & CTO | AMD

https://community.amd.com/t5/corporate/amd-founding-member-of-the-ultra-ethernet-consortium-uec/ba-p/619208

AMD & UEC Building for the Future

No Performance Collapse

- High Performance and Scalability
- Future-ready RDMA
- Advanced Telemetry

Security & Reliability

- Line Rate Performance with Encryption
- Data-In-Use

AI & HPC Frameworks

- Advanced Distributed Parallelism
- Optimized Network Operation
- Support for Heterogeneous Accelerators

Advanced Congestion Control | **Zero-Trust Security** | **Open Software Platform**


AMD
together we advance_