Future Architecture

James Sexton
IBM Fellow
Future Computing Systems
sextonjc@us.ibm.com

Challenges and Opportunities

New Applications of Computing:

- Availability of Data
- AI allows disruptive value extraction from data
- AI to unleash a new level of productivity
- Quantum approaching
- Previously intractable problems now within reach

Explosion of New Approaches and Technologies:

- Multichip Modules
- Hierarchical Memories
- 3D packaging
- Analog Computing
- Optics
- Wafer Scale
- Heterogeneous Elements
- …

Sustainability and Climate:

- Estimates that data centers will consume up to 20% of global power generation by 2030
- Semiconductor Industry under intense pressure to reduce environmental and climate impact

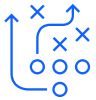
Geopolitical Impacts:

- Need to protect supply chain
- Need to develop regional sovereignty

Future Systems

Why

Challenge is to exploit emerging technologies to radically transform computing in support of knowledge-driven analysis, exploration, reasoning, discovery, and decision making for consumer, enterprise, research, and government uses



How

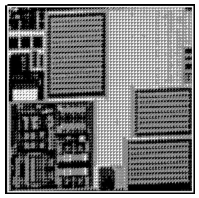
Essential to assemble industry collaborations to create a complete, open, secure, competitive ecosystem that integrates innovation in silicon technology, processors and accelerators, memory, fabric, network, and software to support ever expanding demands on computing resources, to address sustainability.

What

The solutions that will be created will provide an omnipresent set of capabilities deployable and accessible on-premise, in private and public clouds, and at the edge. The impact of these new solutions will be to accelerate the evolution to knowledge-based economies and will be fundamentally transformative

Technical Approach

Develop and prove open standards to support easy, efficient, dynamic assembly of heterogeneous elements

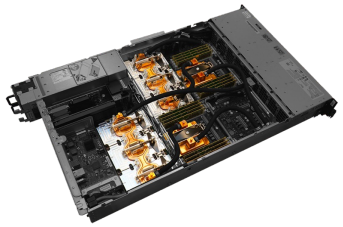


In Module

Technology, Packaging, and Protocols to support assembly of heterogeneous chiplets in module

Target integration of:
- Compute elements
- IO elements
- In Module Memory elements

Focus Areas
- Memory Interfaces
- IO Protocols including CXL
- UCIe
- SMP Protocols



On Planar

Fabric connecting modules and memory

Target
- electrical and optical interfaces
- Communications and memory protocols

Focus Areas:
- CXL Protocols
- Power, Packaging, Cooling

In System

Enable for dynamic assembly of heterogeneous components at scale and performance

Target
- Compute in Data
- Security
- Scalability
- Performance
- Composability
- Memory Pools

Focus Areas:
- CXL comm fabric
- Ultra ethernet
- Memory Pools
- CPU Pools
- Accelerator Pools
- Storage Pools

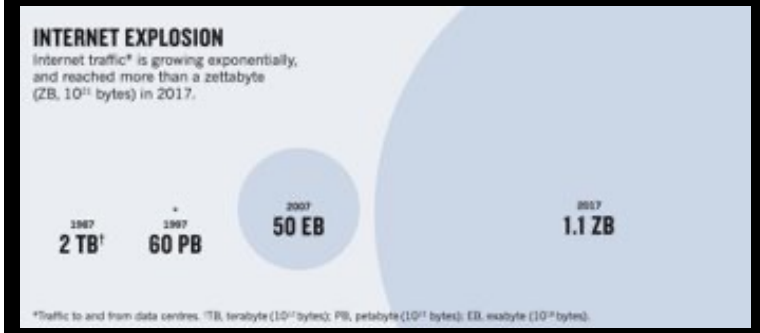
Trends
(Fl)Ops

- Top500 Latest News (May 13, 2024)
  - Top 3 systems: ORNL(AMD), ANL(INTEL), Microsoft(NVIDIA)
    - But now niche, AI Systems dominating
  - Emergence of NVIDIA Grace+Hopper+Mellanox (single supplier)
  - Slowing rate of systems replacement
  - Reduction in growth rate of performance
  - New workloads (AI) disrupting design

The Computer Energy Problem

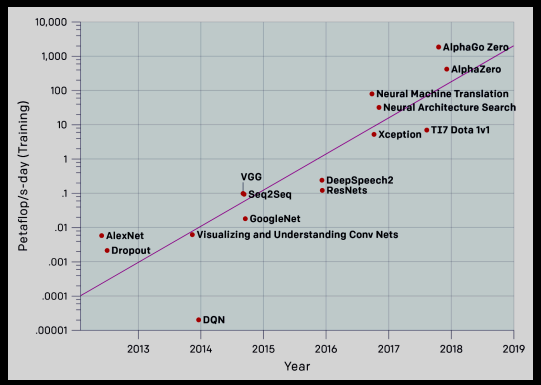
We are at an inflection point :

1) Demand is growing at exponential scale



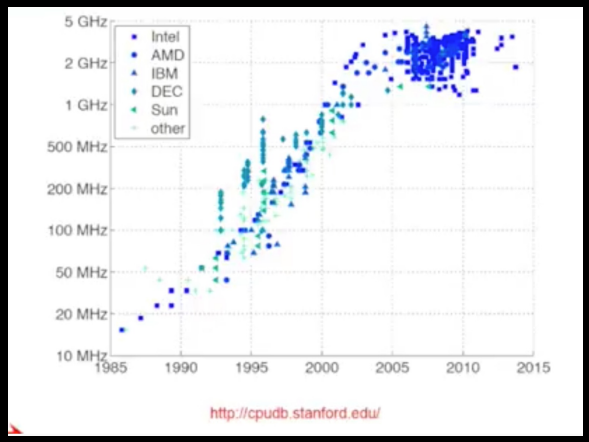
How to stop data centers from gobbling up the world's electricity
https://www.nature.com/articles/d41586-018-06610-y

2) The emergence of energy-demanding workloads(AI)



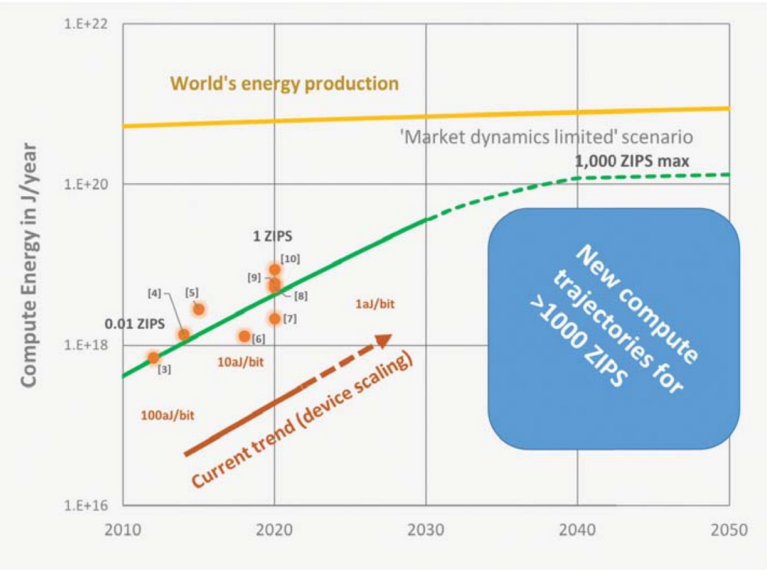
AI power consumption **doubles** every **3-4 months**
*** Green AI, R. Schwartz, J. Dodge, N. A. Smith, O. Etzioni 2019**

3) The end of Dennard Scaling means we can't keep up



Source: Tamar Eilam IBM

Ever rising energy demands for computing vs. global energy production is creating new risk, and new opportunities for **radically different computing paradigms to drastically improve energy efficiency**



31% per year

energy consumption increase trend for hyperscalers in North America

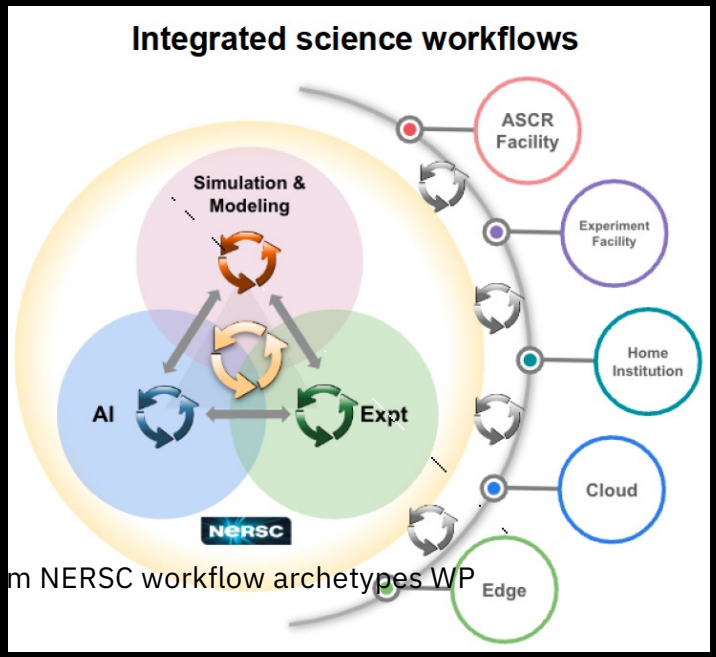
>10%

of the world's power will be consumed by hyperscalers by 2030

Source: Tamar Eilam IBM

Trends

Heterogeneous Workflows / AI – Science Example Replicated in all domains – Automotive, Enterprise, Health, …



m NERSC workflow archetypes WP

Emerging themes

- couple AI training / inference
- "classic" simulation for data generation
- (tightly) couple external data gathering

… across a heterogeneous, distributed, computing environment

Heterogeneous System Architecture
Design Principles

Resource Pools

- CPU
- memory
- accelerators
- network interconnect

1st level interconnect fabric

- open, standard-based
- coherent interconnect, load/store semantics
- >1k endpoints

2nd level network interconnect

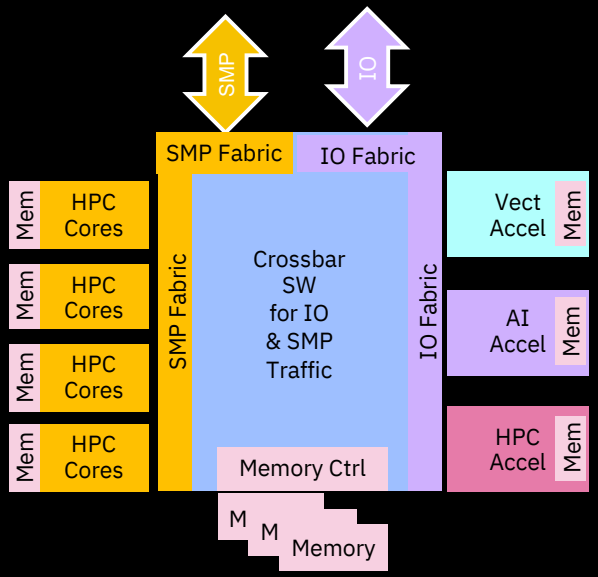
- open, standard-based, scalable network
- connect to storage
- connect to service-oriented partition
- (connect to quantum)

security

- zero trust
- firmware -> applications

configurable / composable / modular upgrade

Open Chiplet-based Ecosystem Supports Heterogeneity

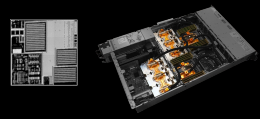


Full-stack Design

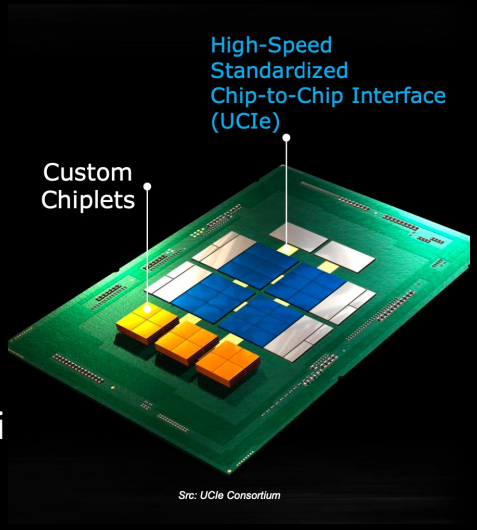
- **Compilation / Debugging / Profiling Tools**
- **Application Domain Libraries**
- **Communication Libraries**
- **Cluster Mgmt / Orchestration / Schedulers**
- **Operating System**
- **Memory System**
- Virtualization / Composability / Security (eg, Attestation)
- Interconnect Fabric Protocol Layer – SMP/IO/Network
- Interconnect Fabric Physical – Crossbar SW
- Communication Links – Standard High-perf IO-Links
- Packaging – Integrating Heterogeneous Chiplets
- Chiplet Design/Fabrication – CPU Cores / Accelerators

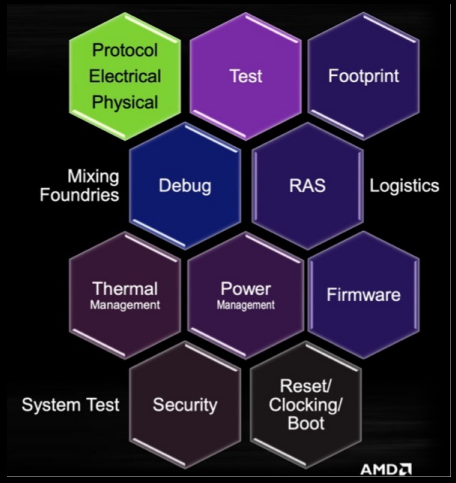
Chiplet-based Modules w/ Processors & Accelerators

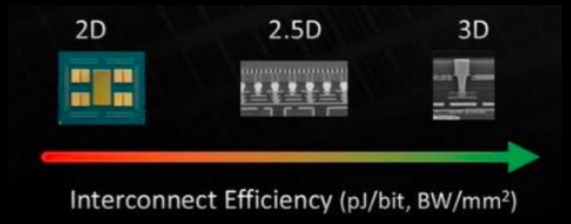
UCIe & PCIe @ Physical



- UCIe as the standard for heterogeneous integration … incl 3rd party chiplets
- offering chiplets to the open chiplet ecosystem
- 2.5D and 3D stacking of chiplets enables architectural performance gains (compute elements and memory) whi lowering total communication energy
- opportunity for co-packaged optics
- PCIe as module-level interconnect



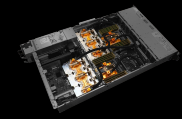


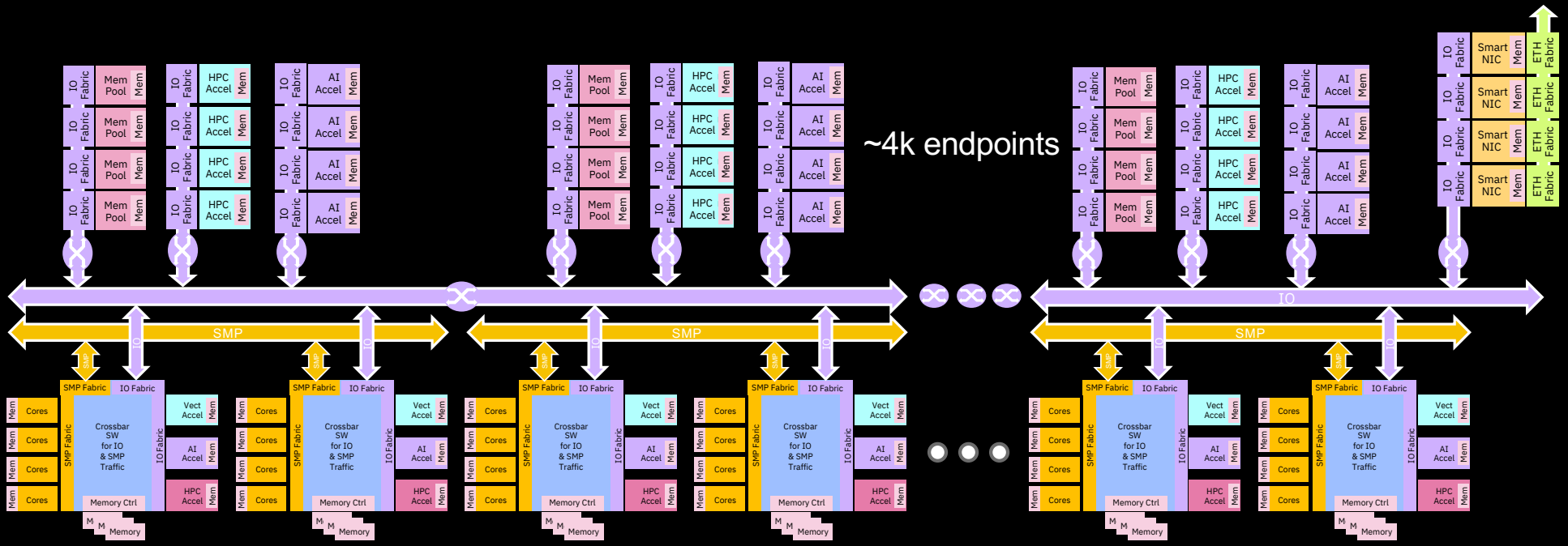


	DIMMS	2.5D Micro-bumps (HBM)	3D Hybrid Bond
pj/bit	~12	~3.5	~0.2

CXL as 1st Level Interconnect

CXL 3.0+





- adding discrete IO-attached accelerators to nodes
- flexibility

- adding IO-fabric switches for high-performance interconnect
- up to 4k end-points

IBM Proposals for CXL Extensions – Included in CXL 3.0

(Basis of a new fabric that can support composable infrastructure)

Enhancements for CXL Fabric Management

- Needed for consistent management of Fabric and data center network
- Support for OpenFabrics Universal Fabric Manager (in proposal stage)
- Dynamic composability using Hot Add / Removal of endpoints

Enhanced Routing for a CXL Fabric

- Needed to increases scope of CXL Fabric (up to 4096 endpoints)
  - CXL currently only support 16 Host (4-bit ID)
- Proposal is to increase ID space to 16-bits
  (12-bit Port ID / 4-bit Logical Device ID)

Host to Host and Device to Device communication

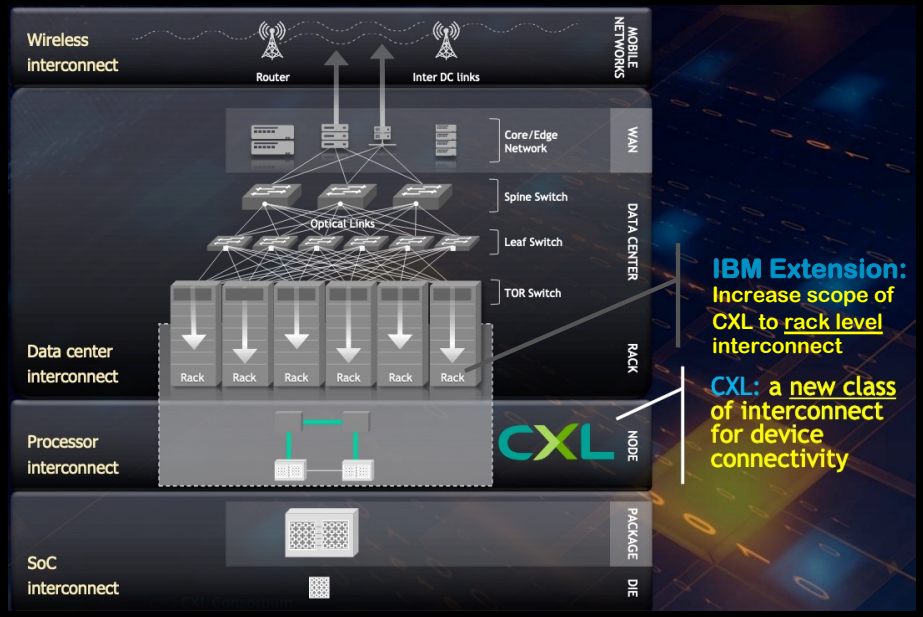
- Needed to allow communications between systems / devices on Fabric
- Extensions for Non-transparent Bridging
  - Compatible with the today's support in Linux
- Potential for creating a single I/O space across the fabric

Enable Larger sized (128B & 256B) read and write transactions

- Needed to improve efficiency after adding larger ID space

Security Enhancements Extensions

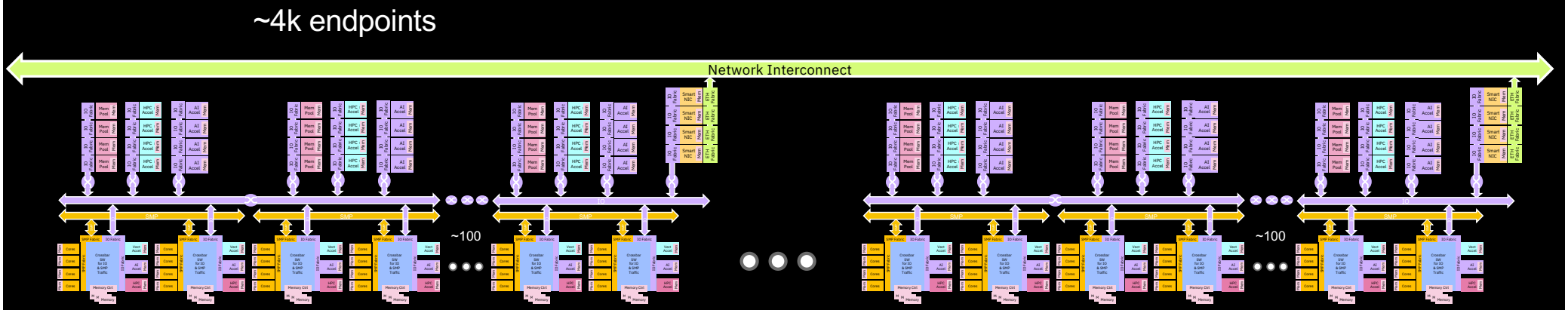
- Needed to improve fabric security for composability
- Current CXL security based on PCIe Secure Channels Integrity and Data Encryption (IDE)
  - Limited to link level (Hop by Hop) IDE which requires fabric to be in trust domain
- Extend CXL security (if possible) to provide endpoint to endpoint security (comparable to PCIe Selective IDE)



2nd Level Network Interconnect
Requirements

- Standards based: IEEE, IBTA, IETF, …
- Converged/inclusive: HPC, Big Data analytics, AI workloads
- Latency: low microsec’s, tightly bound tail latency
- Scalable, robust and cost effective: Efficient to 100’000s of nodes
- Rich set of in-network features (collectives etc.)
- Configuration-free, robust congestion management
- Traffic management: flow class performance isolation, SLA’s
- Predictable performance to support resource dis-aggregation and dynamic compute/storage system composability.

Standards-based Ecosystem
2nd Level Network Interconnect



HPC/AI System

- network interconnecting 10s - 100s of HPC pods
- support HPC, LM-training, & LM-inference

Today's Ecosystem

- Infiniband
- RoCE with proprietary extensions on SmartNIC
- Slingshot, etc (Hyperscalers)

HPC / AI Systems Tomorrow

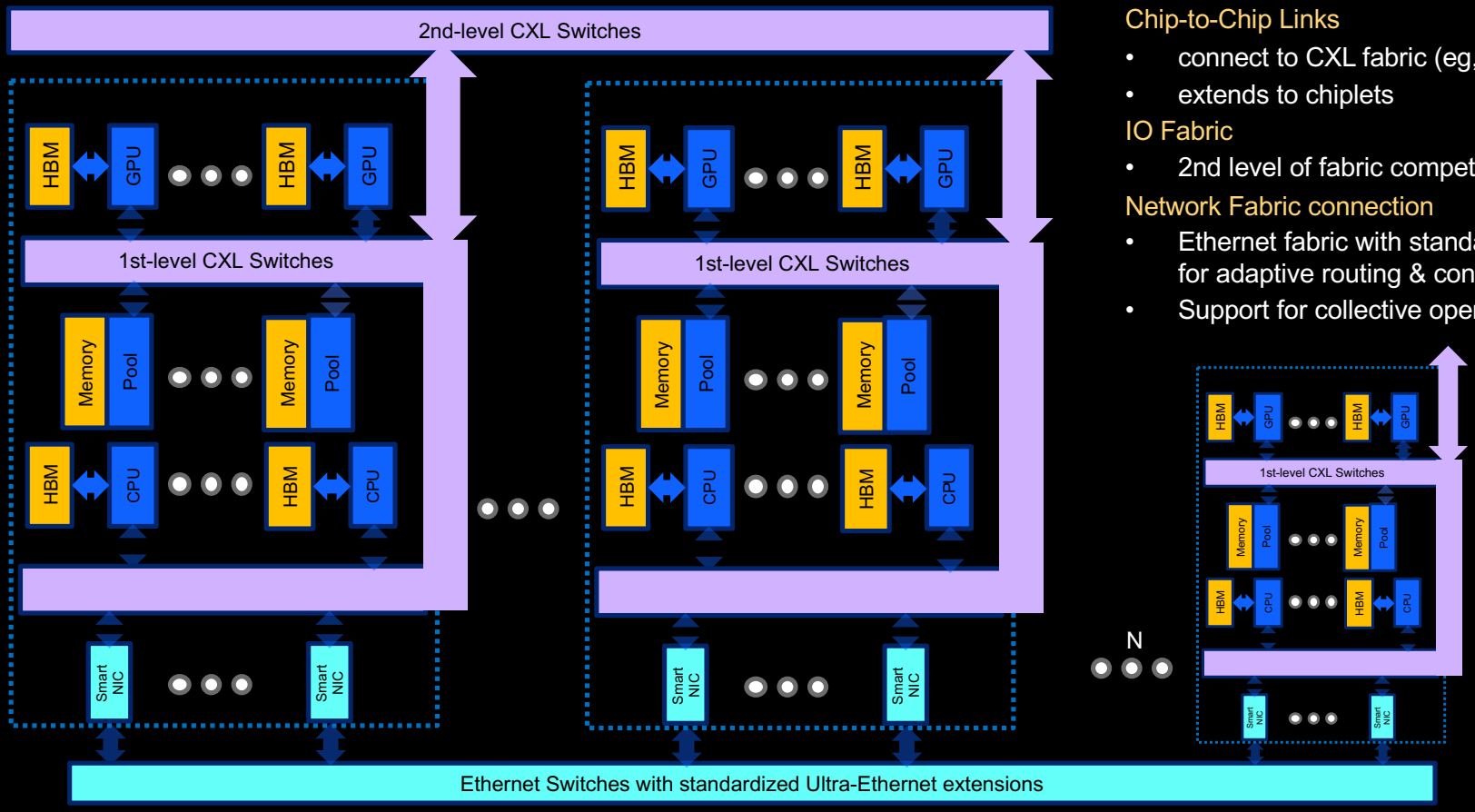
Competitive Standards-based Ecosystem

Memory BW
- HBM attached to CPU/GPU
- CXL Memory pools shared across node or pod

Chip-to-Chip Links
- connect to CXL fabric (eg, CPU/GPU)
- extends to chiplets

IO Fabric
- 2nd level of fabric competitive w/ NVIDIA

Network Fabric connection
- Ethernet fabric with standardized extensions for adaptive routing & congestion control
- Support for collective operations



17

Secure

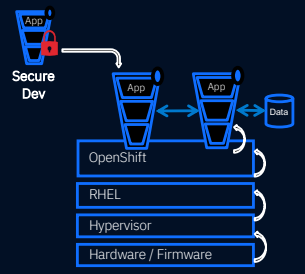
- Holistic Security Model spanning data, compute, network, storage and software:

- Encryption at rest and in flight.

- Can support virtualization and isolation for enclaves as needed.

Integrity and Attestation

A system for holistic integrity management for platform and applications, for clients and providers



Confidential Computing

Solve the security challenges for data

Software Supply Chain

Prevent integrity attacks on software artifacts in software supply chain, provide discoverability and provenance tracking

Compliance Automation

From regulation to controls, gap analysis, threat management intelligence & risk-based compliance management

DevSecOps

Code risk analyzer - automated security and compliance based on static assets in git repository

Challenges and Opportunities

- Power Requirements growing unsustainably
- Emergence of special hardware for specific tasks
- AI Devices, Cereberas, NextSilicon, Graphcore, …
- New packaging methods, new memory technologies, new fabric technologies
- New computing models: Quantum, wafer-scale, analog, in-memory
- No single supplier(?) can deliver all elements.
- Standards essential to support integration

- **Expect increasing heterogeneity**
  - Even within a single system can foresee sub-sections with different capabilities
- **Expect increasing complexity of workflows**
- **Expect increasing need for / opportunity for composability**
  - Within Module
  - Within System
  - Across Systems
- **Expect increasing use of AI in all aspects of the computational analysis.**
  - For deployment, execution, optimization of workflows

Critical Challenge: how do we develop an open ecosystem accessible to many and prevent lock-in to vertically integrated foundry ecosystems or proprietary full stack providers.

Thank You!