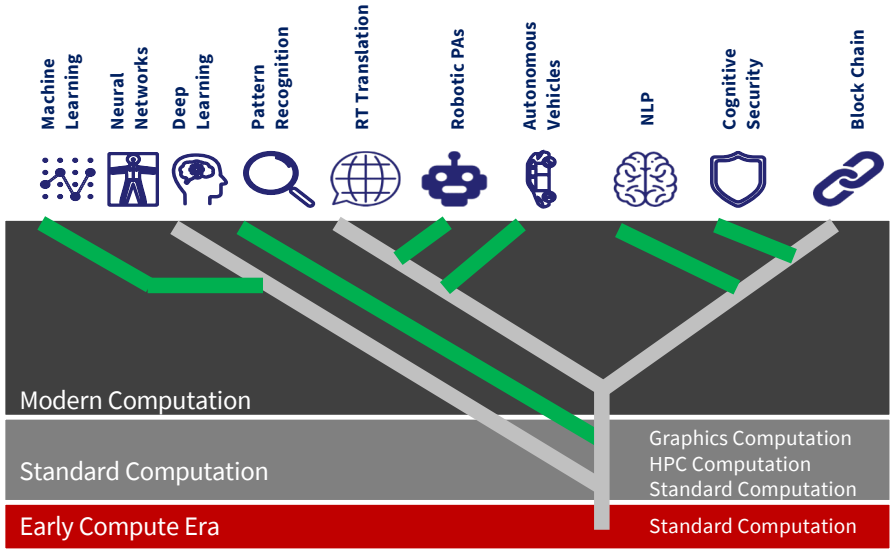
January 24 - 26, 2023
DoubleTree by Hilton San Jose
ChipletSummit.com

The Open Chiplet Ecosystem: Accelerating AI Hardware

Arvind Kumar
IBM Research

Historic Opportunity for HI and AI

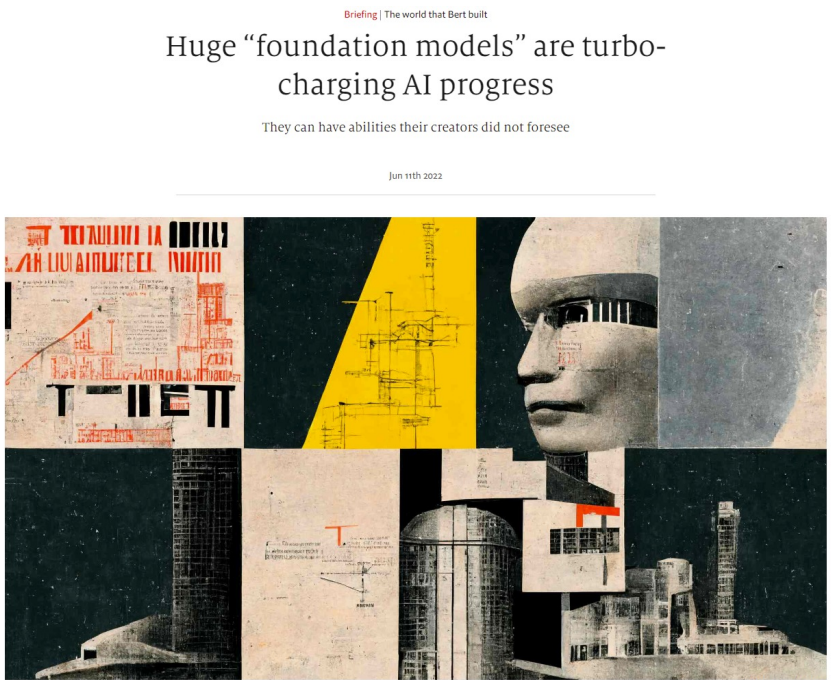
from ODSA:



AI and Machine-learning and data-heavy workloads have exploded in past 5 years and will diversify as new applications are discovered constantly…

All images from Creative Commons

from *The Economist*:

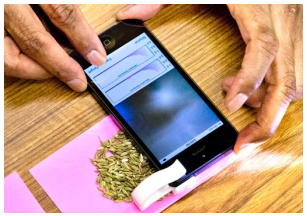


Trends for growing and diverse AI workloads:

- Offer as a Service (**aaS)**
- Time to Market is critical
- Keep infrastructure cost low

Diverse AI Use Cases

IoT / Sensor

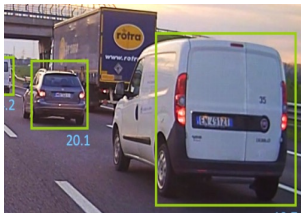


< 100 mW
1-5 cores

Mobile

250 mW to <2W

Automotive



20-50W

Data Center

75-300W
Tens to hundreds of cores

Edge compute → Compute → Data Center

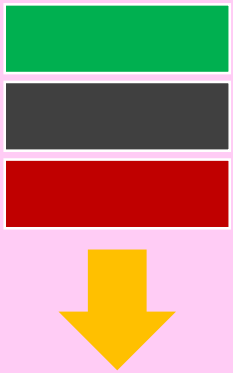
AI workloads	Mainly inference	Training and inference
Core count	Small	Large (even larger for training!)
Technology	Low power/older node	High performance/leading node
Power envelope	Low	High
Latency / throughput	Latency critical	Both may be important
Packaging	Low cost	Advanced packaging to support high BW

Diverse AI use cases with varying core count, core functionality, memory, and power envelope requirements.

Requires domain-specific architectures

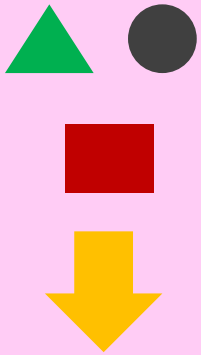
from ODSA:

Logic Disaggregation



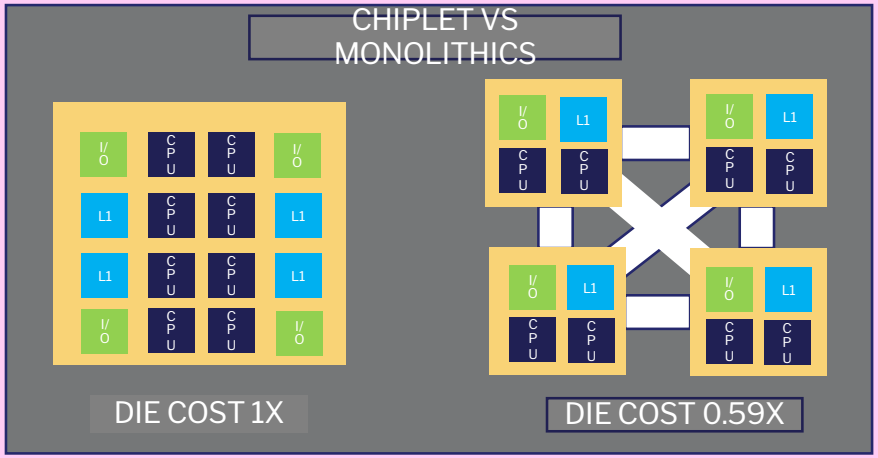
Improve yield and simplify/relax design requirements

IO Disaggregation



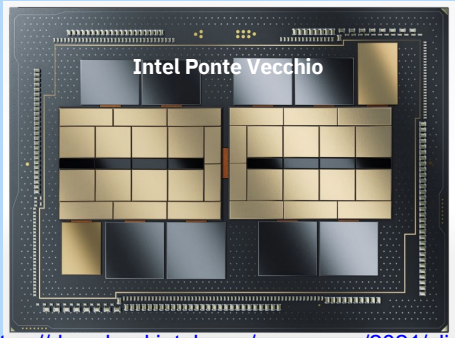
Right functionality in right silicon node

PROVEN EXISTING BUSINESS MODELS

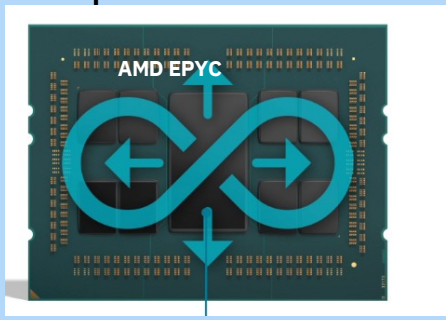


[L. Su, IEDM'17]

recent chiplet industry examples:

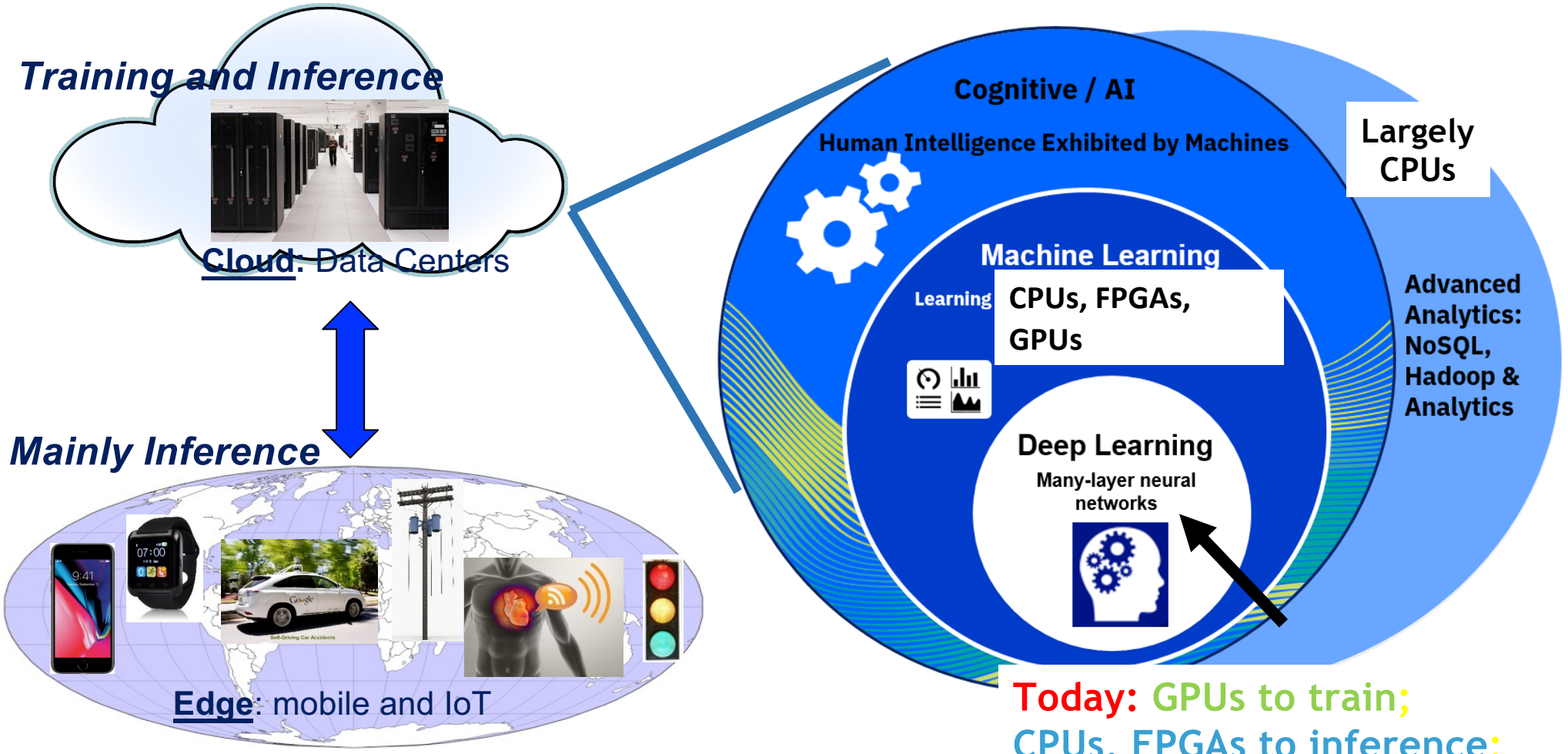


https://download.intel.com/newsroom/2021/client-computing/intel-architecture-day-2021-presentation.pdf



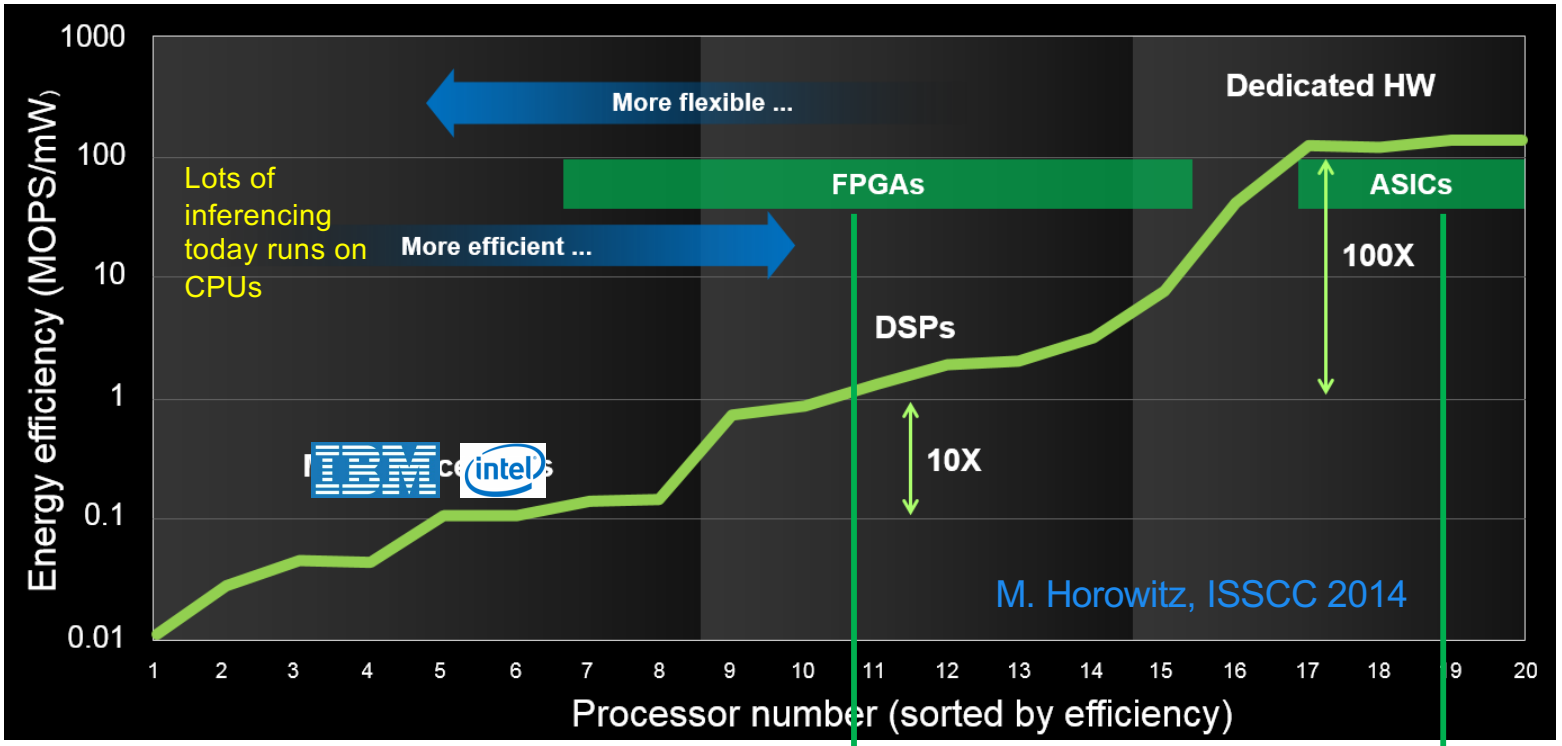
https://www.amd.com/en/technologies/infinity-architecture

Processing AI Workloads



Today: GPUs to train;
CPUs, FPGAs to inference;
Race to ASICs

ASIC Development Costs



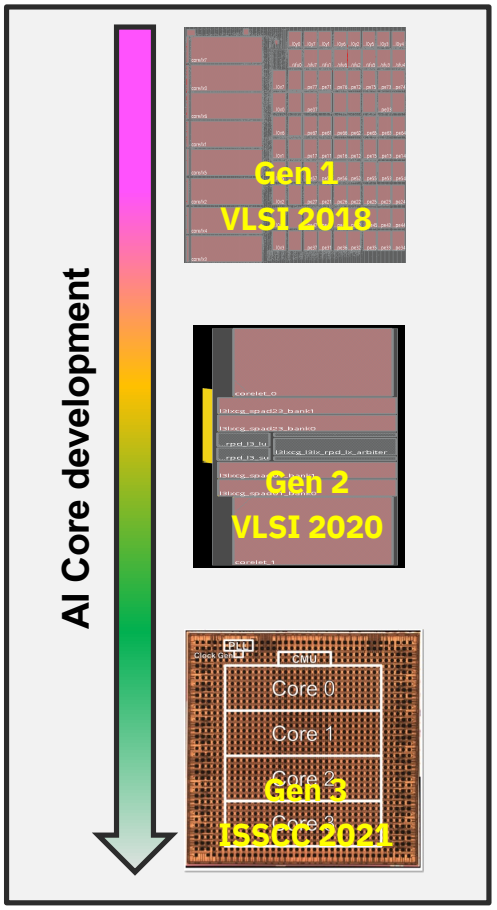
Low NRE
More flexibility
Lower efficiency

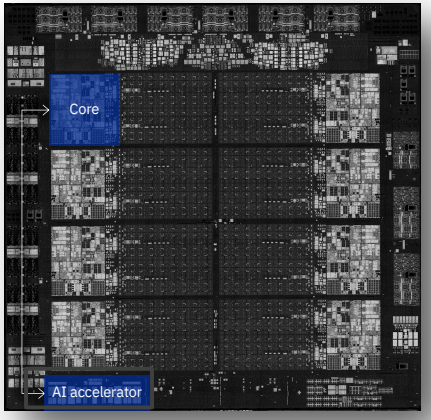
High NRE
Less flexibility
Higher efficiency

Not feasible to build a new ASIC for each new AI application

Can chiplets lower the barrier and help bridge this gap?

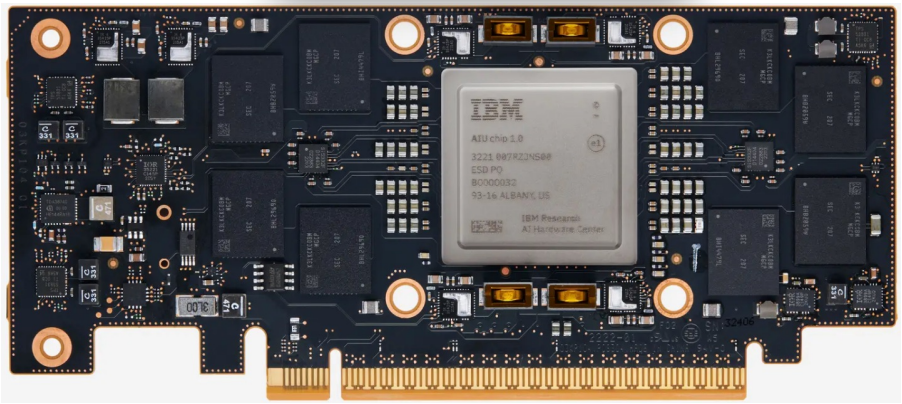
IBM's AI Journey and Path Forward





IBM Telum
(announced 2021, production soon)
Single AI core
on IBM Z mainframe chip
optimized for enterprise workloads

Ultra-low latency



IBM AIU
(announced 2022)
32 AI cores
on PCIe card

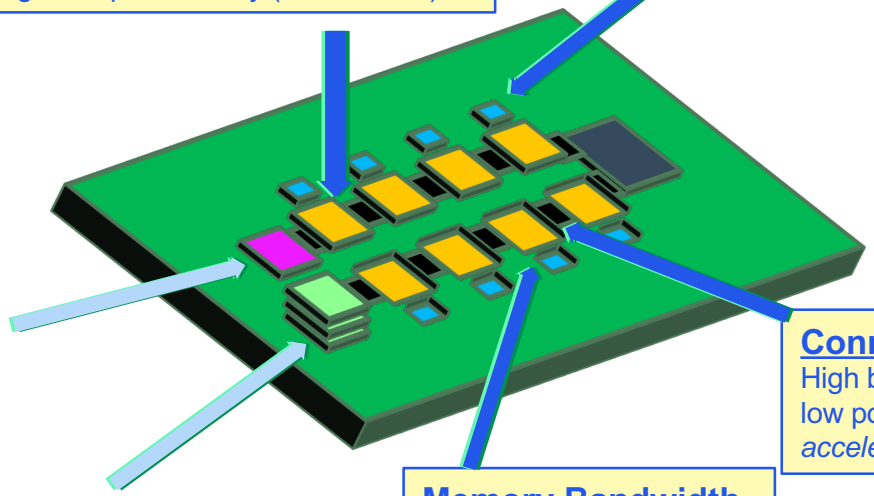
High compute intensity

How can an Open Chiplet Ecosystem accelerate our future development?
- Broaden application space
- Reduce time to market

Chiplet-Based Platform for AI

Compute
High compute density (TOPS/mm2)

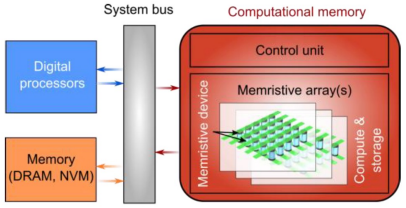
Memory Capacity
High capacity memory near processors



Connectivity
High bandwidth,
low power interfaces for
accelerator-to-accelerator

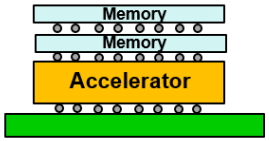
Memory Bandwidth
High bandwidth,
low power interfaces for
accelerator-to-memory

Specialized accelerators
(Example: in memory compute)



Le Gallo *et al.*, 2017

3d memory stacks



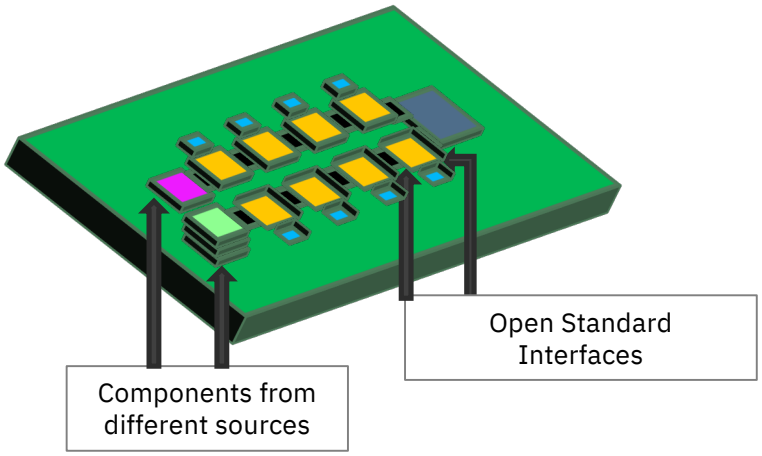
Open Chiplet Ecosystem

Current State

- Single/limited sourcing
- Proprietary interfaces
- **Can't mix and match**

Desired State

- Diversity of components
- Open-standard interfaces
- **Can mix and match**



Open Standard Interfaces

Metrics
- Bandwidth (Gbps/mm)
- Area efficiency (Gbps/mm2)
- Trace length (mm)
- Energy efficiency (pJ/bit)

Properties
- Easy to implement
- Portable across nodes
- Scalable with packaging technology

Examples today
- Bunch of Wires
- OpenHBI
- UCIe

Pre-Conference Tutorial D – Interfaces: Atom Watanabe et al.,
"Chiplet on Advanced Packaging –Integration Approaches, Electrical Interfaces, and System Realization Opportunities"

Takeaway Messages

What can chiplets and an Open Chiplet Ecosystem do for AI?

Cost:

- Yield and node optimization (reduced cost)
- Quicker re-spins (shorter design time)
- Reuse of existing IP

Today's ASIC paradigm does not enable AI variants fast enough

Performance:

- More Si compute density
- Scaling is slowing and more expensive

Applications:

- Modularity/Composability/Scalability
- Rich application space through components from different sources

Chiplets can enable diverse AI workloads aaS
Benefits will extend to other compute tasks


IBM