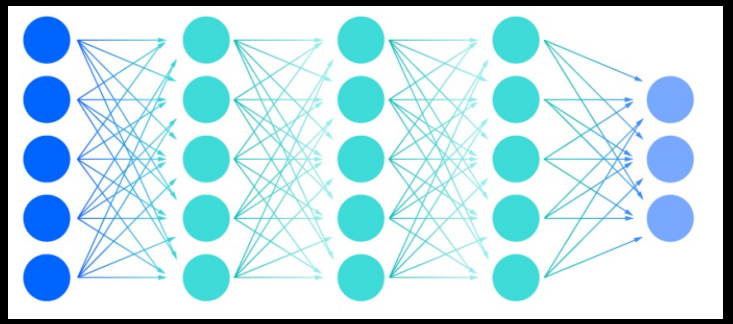
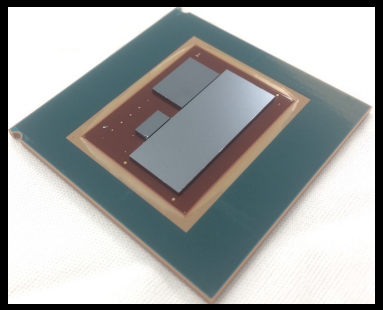
IBM **Semiconductors**

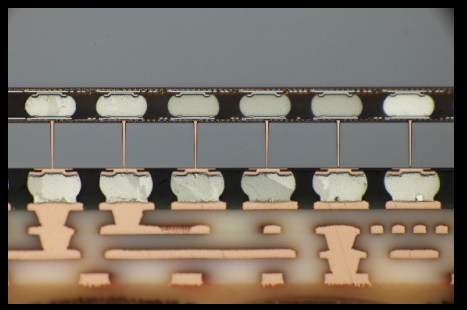
Meeting the Explosive Demands of AI with Chiplet Architectures



Arvind Kumar, PhD
Principal Researcher and Manager
IBM Research

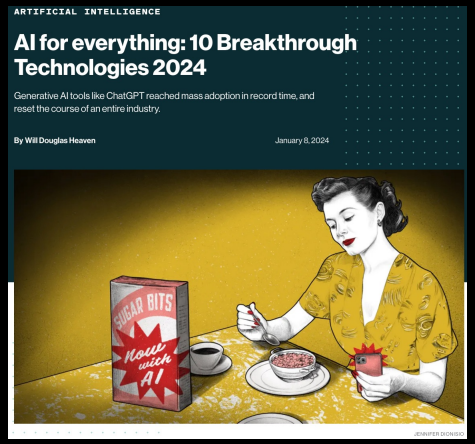
March 19, 2024

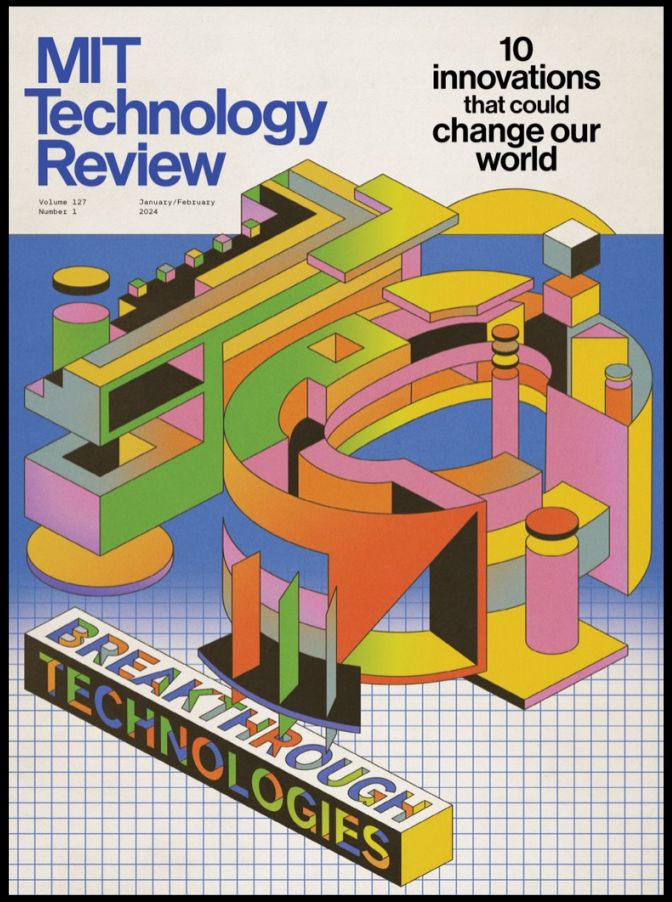




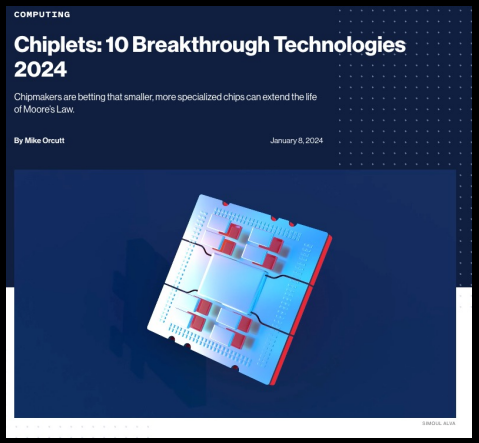
A Historic Opportunity for AI and HI

#1: AI



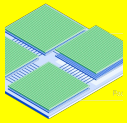


#6: Chiplets



Meeting the Explosive Demands of AI with Chiplet Architectures

Artificial Intelligence

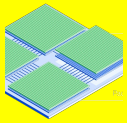


Chiplet Architectures

AI Applications enabled by Chiplet Architectures

Meeting the Explosive Demands of AI with Chiplet Architectures

Artificial Intelligence



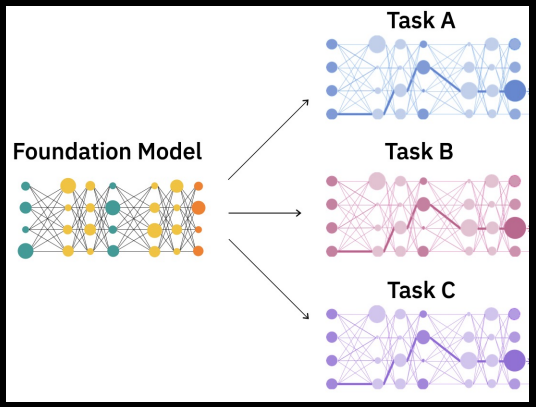
Chiplet Architectures

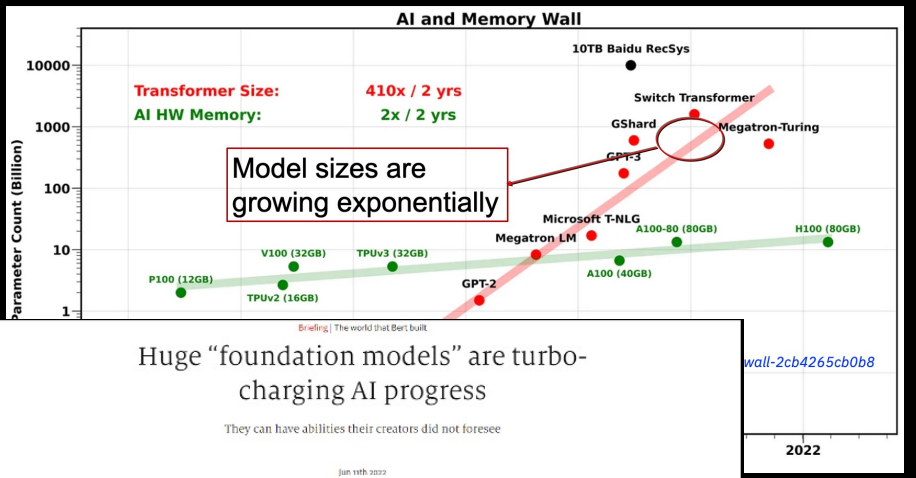
AI Applications enabled by Chiplet Architectures

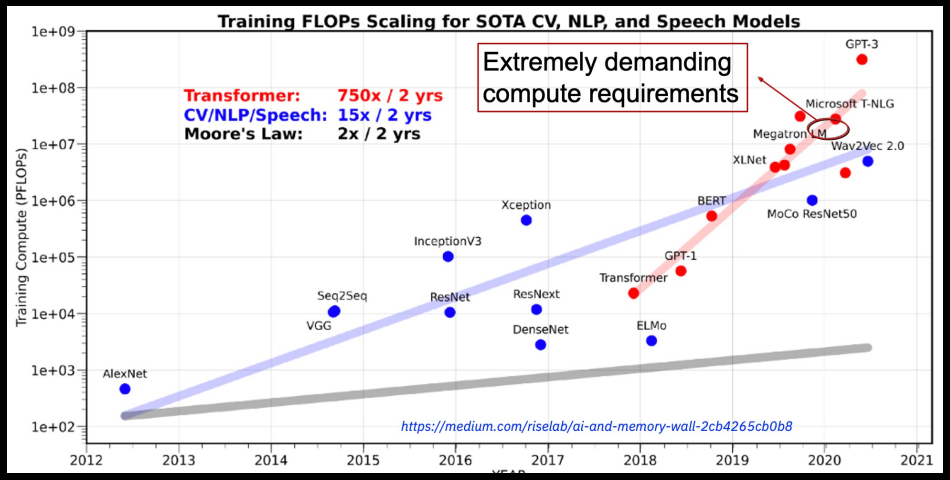
Foundation Models: A Gamechanger for AI

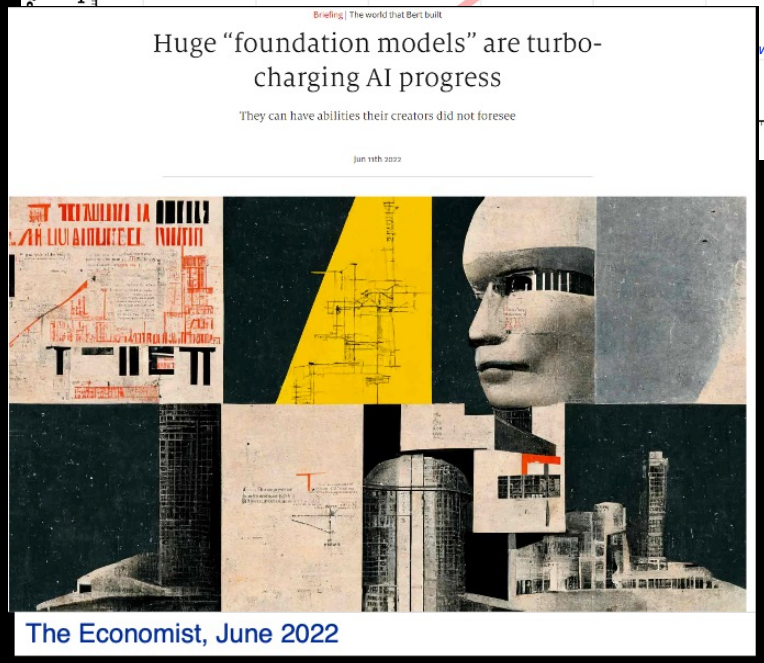
Foundation Model

AI model trained on large amounts of unlabeled data that can be adapted easily to new use cases

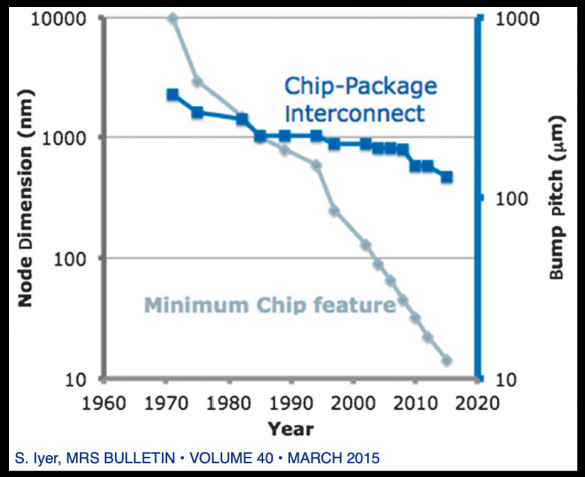




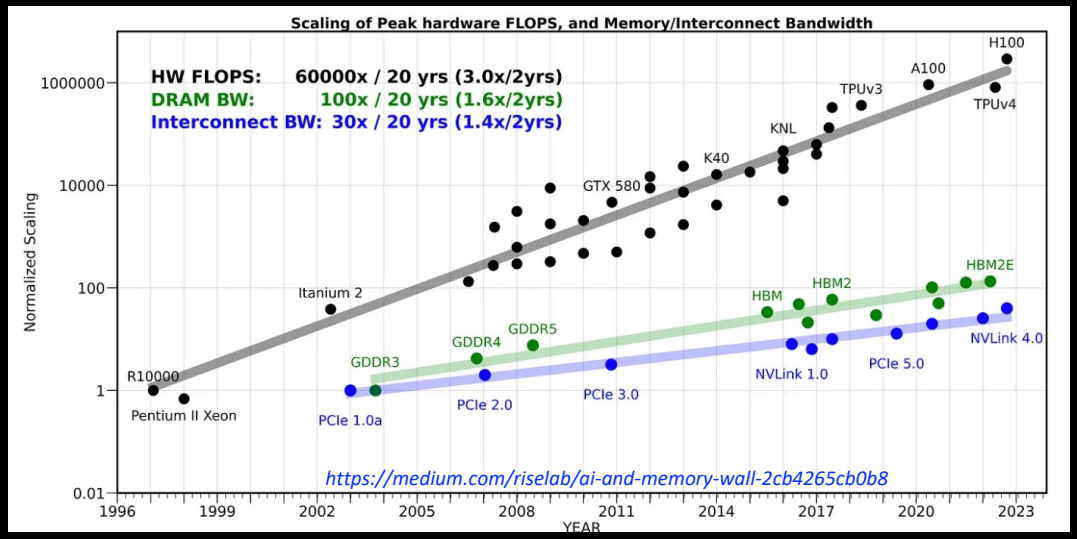




Scaling trends



Packaging scaling has been much slower than logic scaling

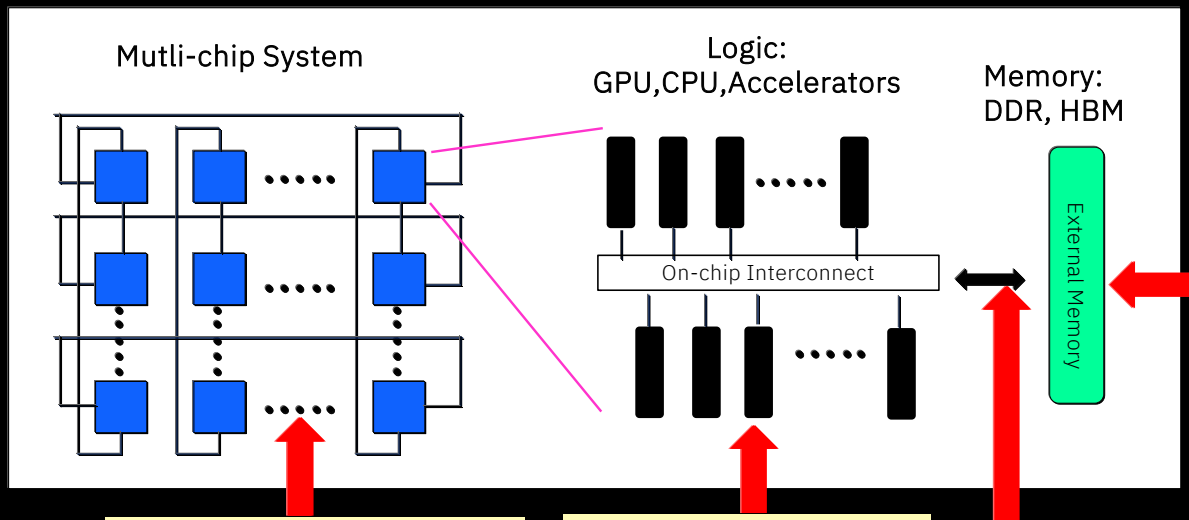


Mismatch between compute and bandwidth gains can lead to low overall system utilization

Continued AI progress is unsustainable without significant hardware and software innovation

Need a new paradigm to meet the demands of data-intensive AI workloads

4 Key AI System Requirements



Memory Capacity
Large capacity required for
- Weights for Large Models
- Activations
- Training-specific
  - Gradients
  - Optimizer States

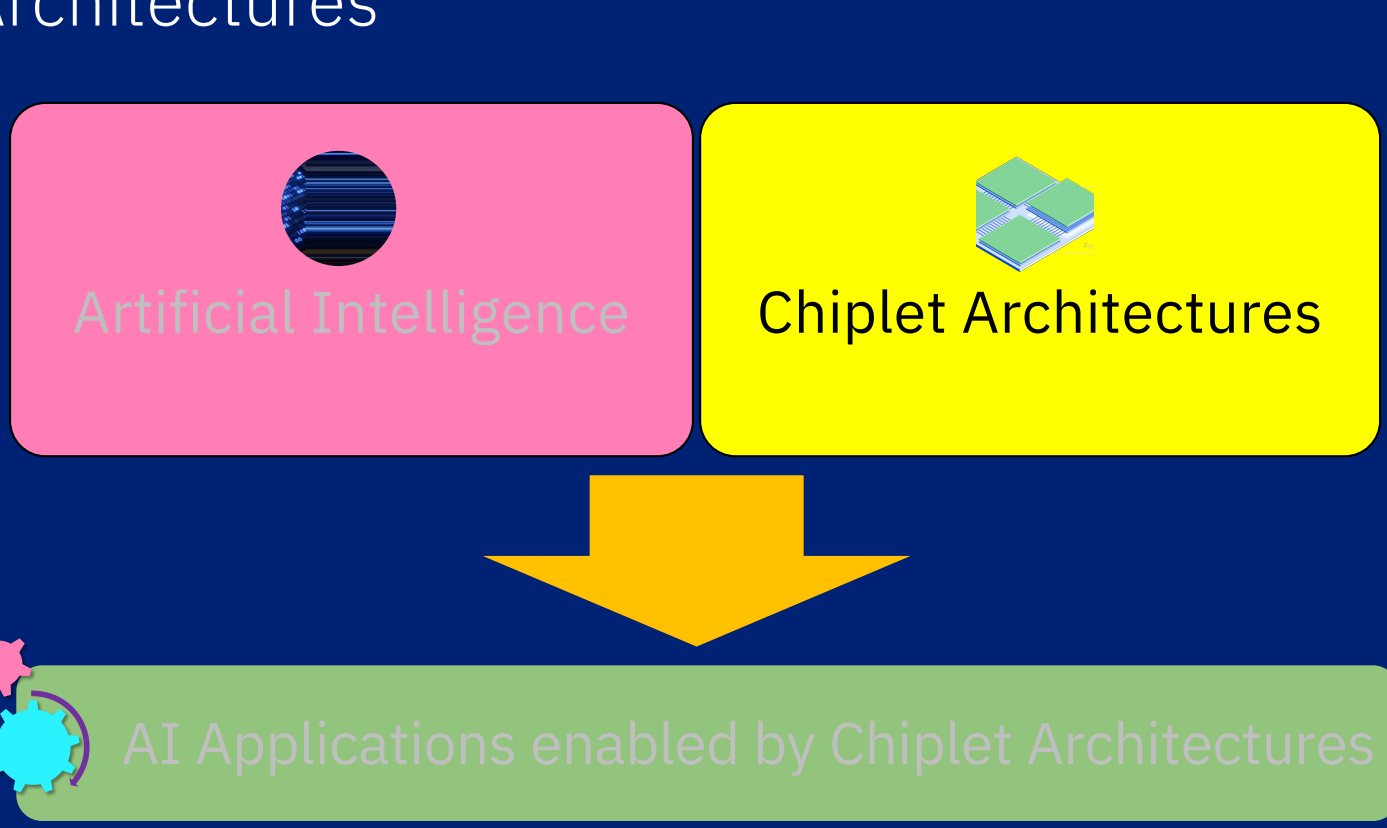
Connectivity
High *accelerator-to-accelerator* bandwidth required to pass activations and exchange gradients for training

Compute
High TOPS for matrix and other compute-intensive operations.

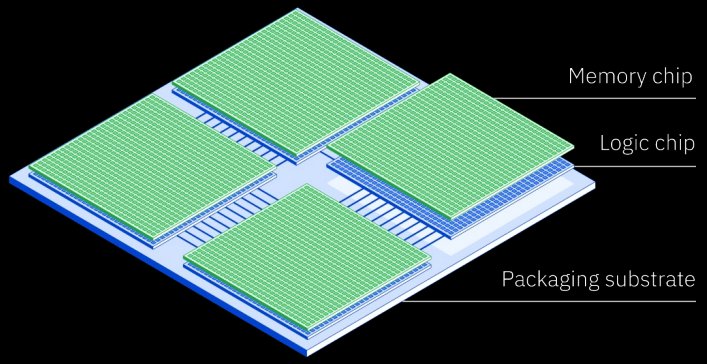
Acceleration leveraging special properties of AI workloads

Memory Bandwidth
High *accelerator-to-memory* bandwidth required to keep cores busy

Meeting the Explosive Demands of AI with Chiplet Architectures



Chiplet technologies: Benefits for Foundation Model accelerators



Compute	Memory	Connectivity	Flexibility
High TOPS/volume to improve compute	High capacity for very large FMs	High bandwidth to move data	Enable variants, fast upgrades
↓	↓	↓	↓
Multi-chiplet platform	3d stacked memory	High interconnect density fabric	Mix-and-match architectures

2.xD and 3D Heterogeneous Integration solutions are a promising path to support the demands of Foundation Models

IBM's Advanced Packaging / HI Infrastructure

Lab to Fab!

Watson Research Center – MRL
Broad Research Capabilities

- IBM **Microelectronics Research Lab** (MRL) has a long-standing capability in **advanced packaging**.
- Full process flow capability from **materials synthesis** all the way through **assembly and test**
- Ability to execute bleeding edge packaging including **design and characterization**

IBM Albany Research
Wafer-Based HI R&D

- **Class 1000 cleanroom** space upgraded to support new state of the art **HI Line**
- Focused on enabling **3DHI technology with hybrid bonding, HD substrate enablement** and **DBHi bridge technology**
- **Transfer resulting technologies** to **IBM Bromont** for IBM and OEM use

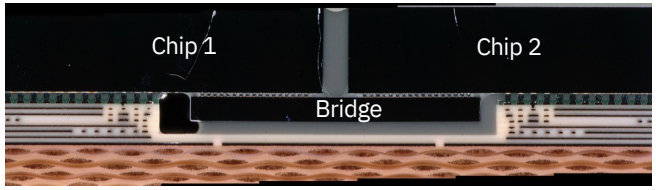
IBM Bromont / C2MI
Development and HVM

- 50 years of packaging manufacturing experience serving both IBM and external customers (majority) for **advanced flip chip, SiP** and **test production**
- Partnered with the **C2MI technology incubator**
- Massive expansion planned in both capability and capacity supported by **Canadian Government funding**

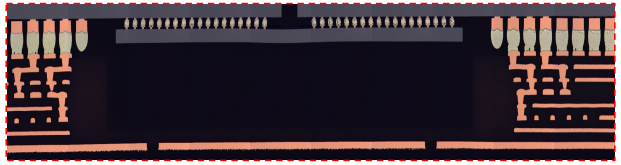
IBM Chiplet Technology Research: 2.XD Integration

DBHi (Si Bridge)

Gen 1

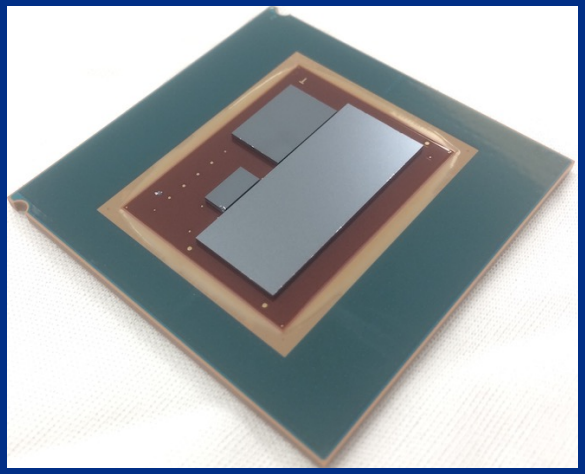


Gen 2



- Thin Bridge
- Direct Assembly
- Compatible with standard substrates

High Density Interconnect Substrates

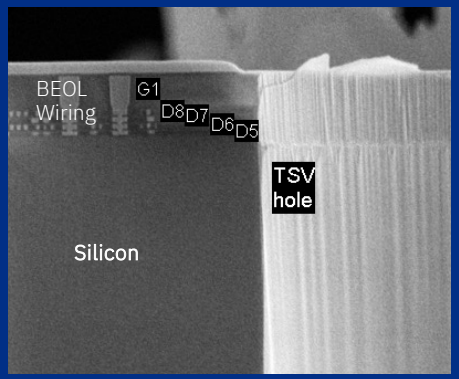


- Advanced Substrate, FOWLP, and Organic Interposer options
- Chips last integration
- Large form factor
- High reliability

IBM Chiplet Technology Research: 3D Integration

3D Enabling Technologies

TSV Late Integration

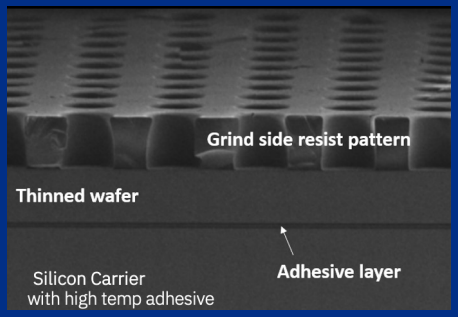


High current carrying capability

Die-to-wafer Hybrid Bonding

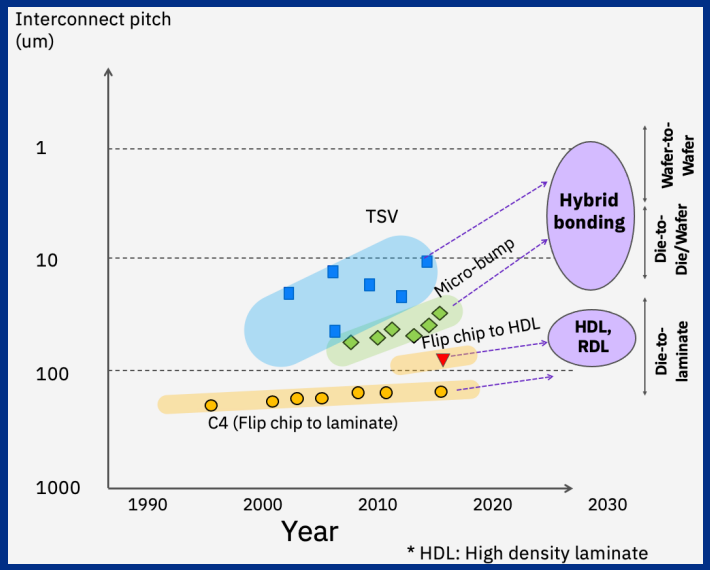
Ultrahigh interconnect density

Wafer Thinning & Backside Process



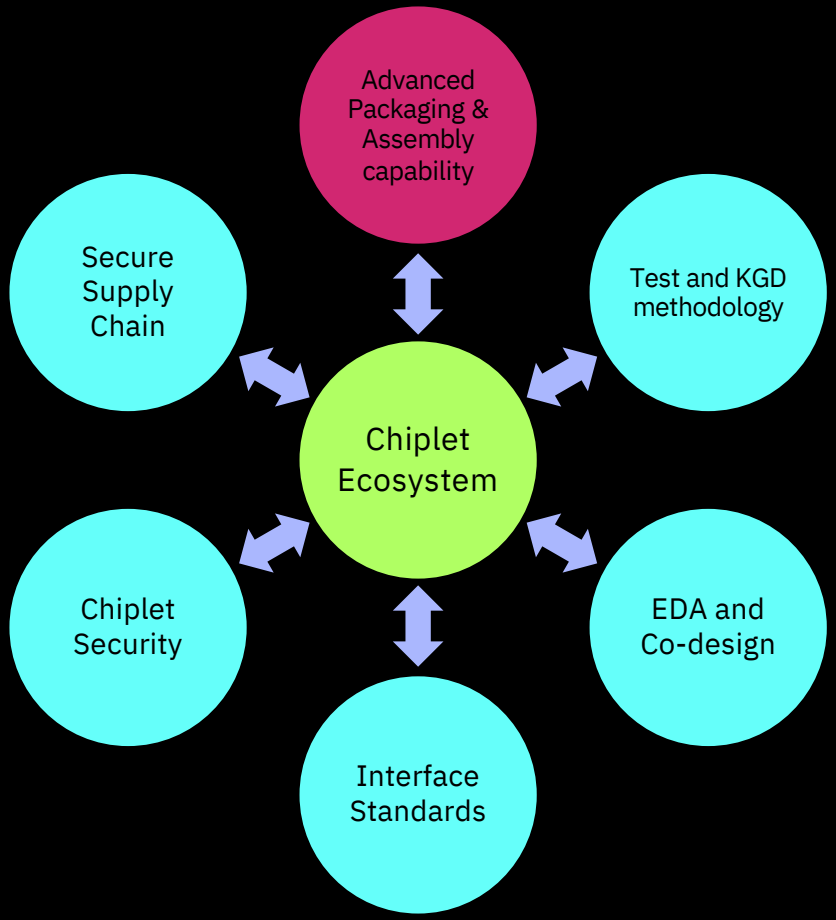
Lower warpage & Si tool compatibility

Benefits for AI Architectures



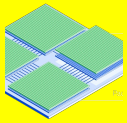
- Much higher bandwidth vs. off-chip memory
- Much lower power vs. off-chip memory
- 3D stacked memory is like on-chip memory
- Lower overall form factor

Building a Chiplet Ecosystem



Meeting the Explosive Demands of AI with Chiplet Architectures

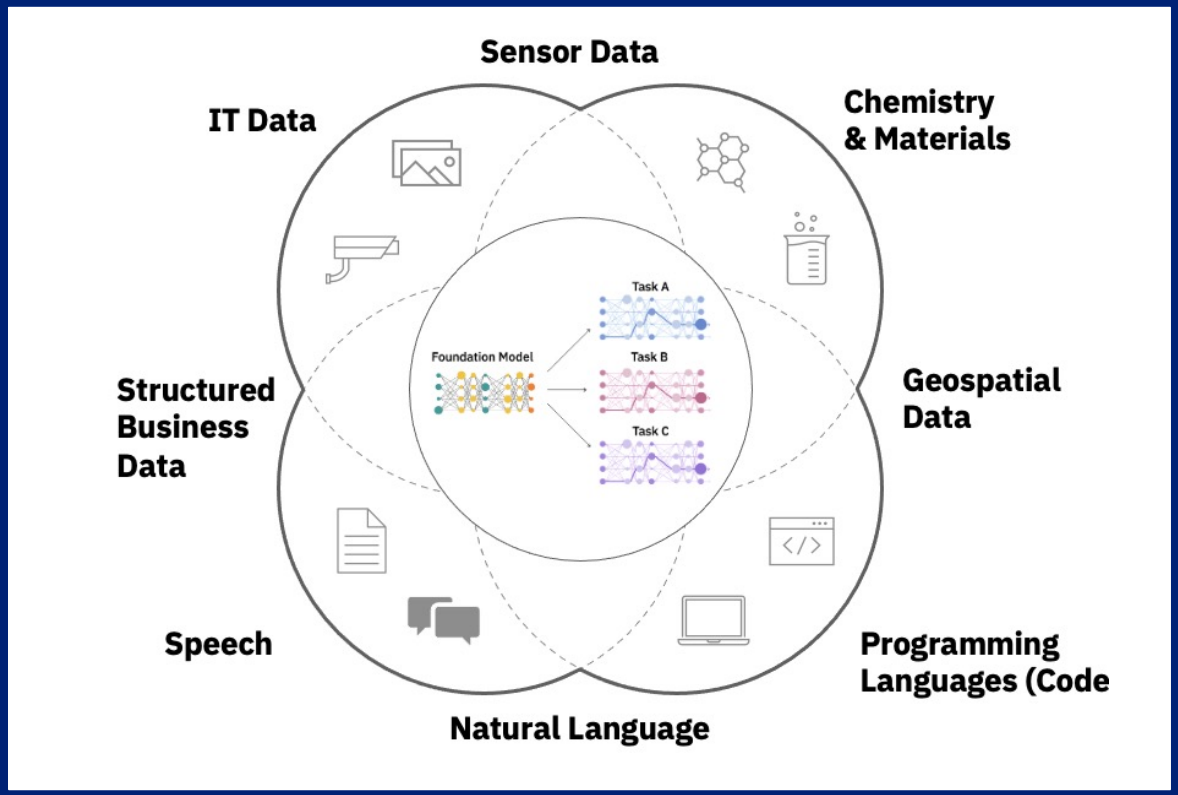
Artificial Intelligence



Chiplet Architectures

AI Applications enabled by Chiplet Architectures

Application 1: Chiplet Architectures for Enterprise Computing



- Many domains have unlabeled data that could be used to train custom foundation models
- Can solve business problems that were previously considered intractable

IBM Artificial Intelligence Unit (AIU)

SoC implements IBM's leadership innovations in low-precision AI arithmetic and algorithms

- Optimized for enterprise AI workloads with foundation models
- Accelerates inferencing and fine-tuning for generative AI
- Supports FP16, FP8, Int8, Int4 data formats
- Plugs into a standard PCIe slot

On-Prem IBM Research AIU Infrastructure

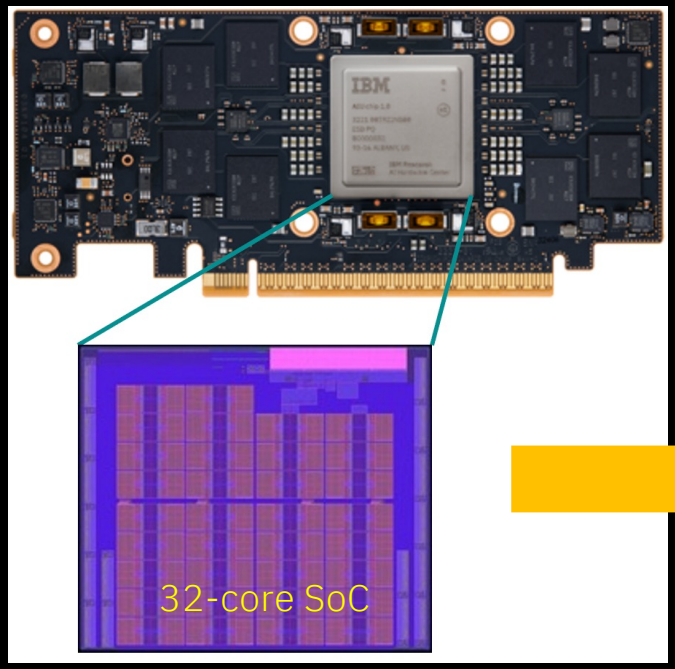
ThinkLab Deployment (Yorktown Heights)

- Powered by IBM Research AIU
- Scalable system architecture
- Highly power efficient

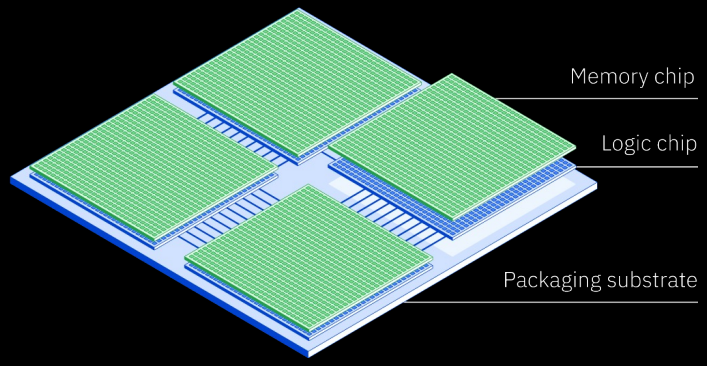
Performance example:
For data pre-process workloads, measured comparable throughput as GPUs while consuming less power

AIU of the future

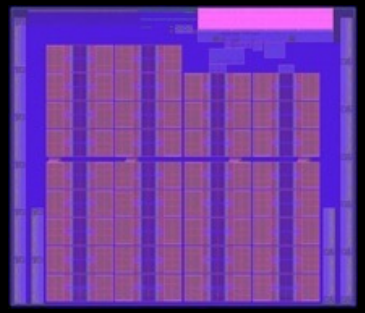
AIU today:
Monolithic SoC



Future AIU:
Chiplets on SiP

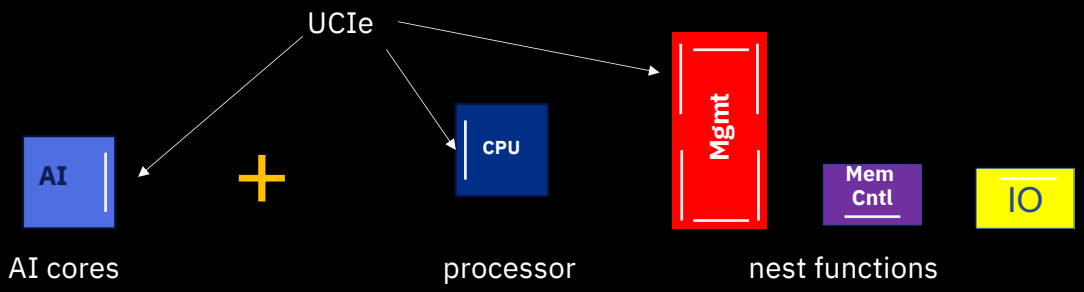


SoC Disaggregation



32 AI cores + nest

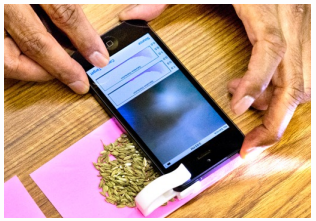
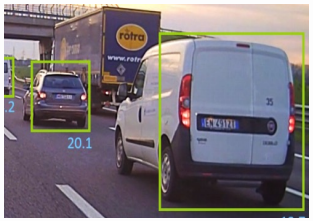
Chiplet Factoring



Cost benefits	• Create variants right-sized for application • Lower cost and faster time-to-market	• Node optimization • Potential for chiplet reuse
Performance benefits	• More AI cores per volume vs monolithic design	• Lower power for data transfers

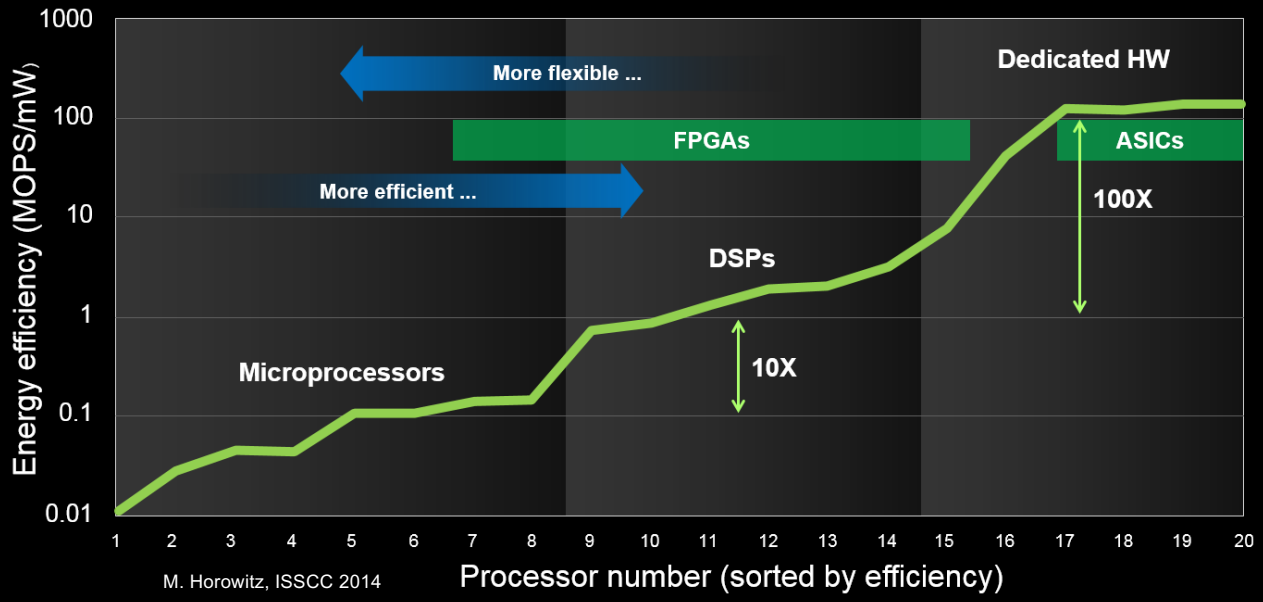
Example:
3D stack derivative

Application 2: Chiplet Architectures for Edge Computing

IoT / Sensor	Mobile	Automotive	Data Center
			
< 100 mW	0.25-2W	20-50W	75-300W
Edge compute	**Compute**		**Data Center**

Chiplets enable heterogeneous architectures for new targeted applications!

Efficiency benefits of customization



Not feasible to build a new ASIC for each application

Can chiplets lower this barrier and help bridge the gap between flexibility and efficiency?

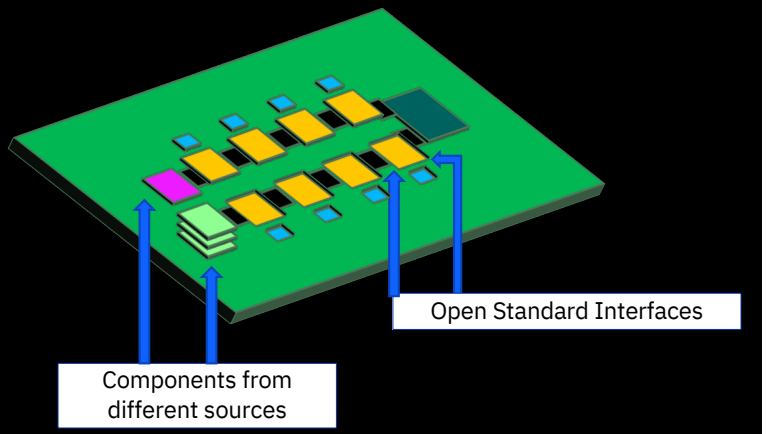
Domain-specific architectures with Open Standards

Current State

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

Desired State

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



Metrics

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

Open Standard Interfaces

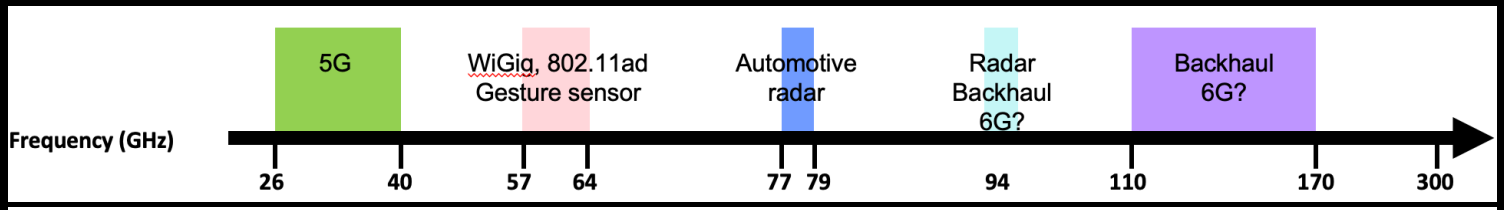
Properties

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

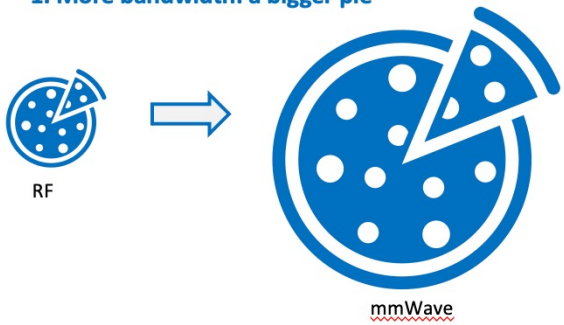
Examples today

- BoW (Bunch of Wires)
- OpenHBI
- UCIe

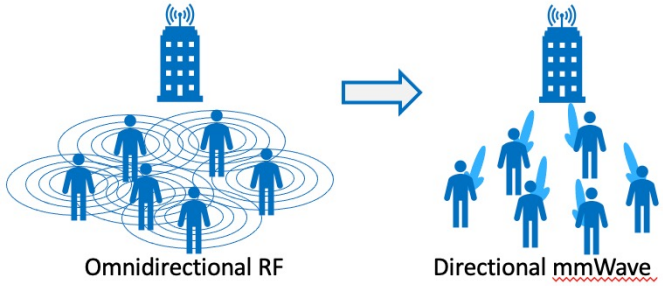
mmWave and Heterogeneous Integration



1. More bandwidth: a bigger pie



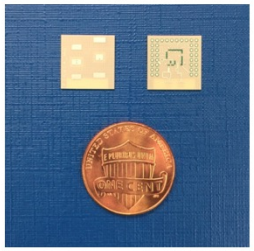
2. Less interference: directionality



3. Miniaturization: smaller wavelength

SDR with 4 antennas
~0-6GHz

60-GHz TRX with 4 antennas
B. Sadhu et al, JSSC 2020



4. Unique properties for sensing

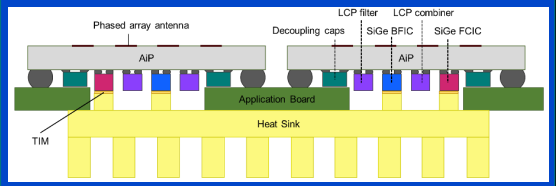
- Transmission through fog, rain, snow
- De-polarization effects
- Micro-doppler
- …

Derived from *B. Sadhu, ISSCC 2022 Tutorial*

Vertical integration from Antennas to AI

A Set of Factors Creating a Key Opportunity

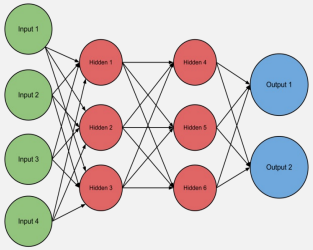
Availability of complex, silicon-based, mmWave multi-antenna systems



D. Liu et al., ECTC 2023

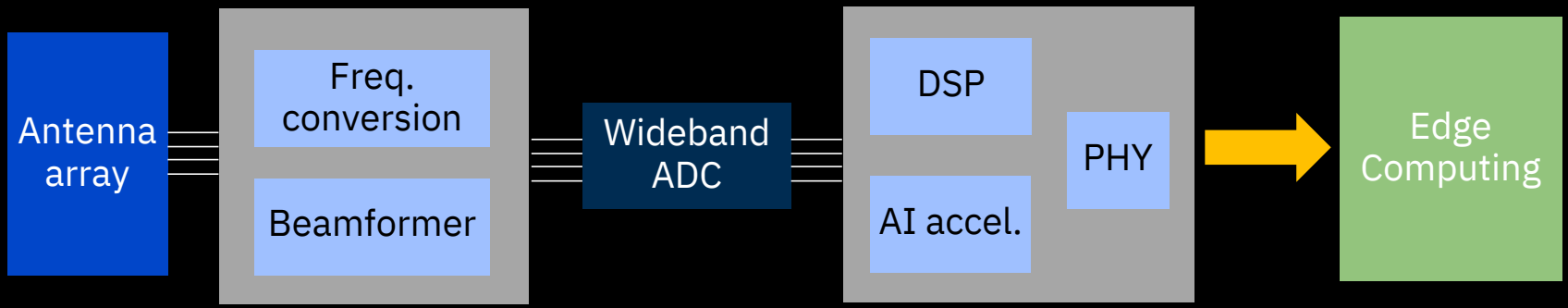
Growth in demand and infrastructure for ubiquitous communications and sensing

Growth in AI capabilities



Vertically integrated "Antennas to AI" systems for communications and sensing

High-level architecture of an ‘Antennas-to-AI’ platform

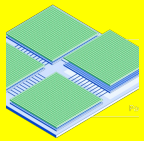


Promise of mmWave systems for emerging commercial and DoD communications/sensing applications can be fully realized by

- AI-driven adaptation and data interpretation
- Heterogeneous integration of RF, mixed-signal, and digital ICs

Let's create ↷

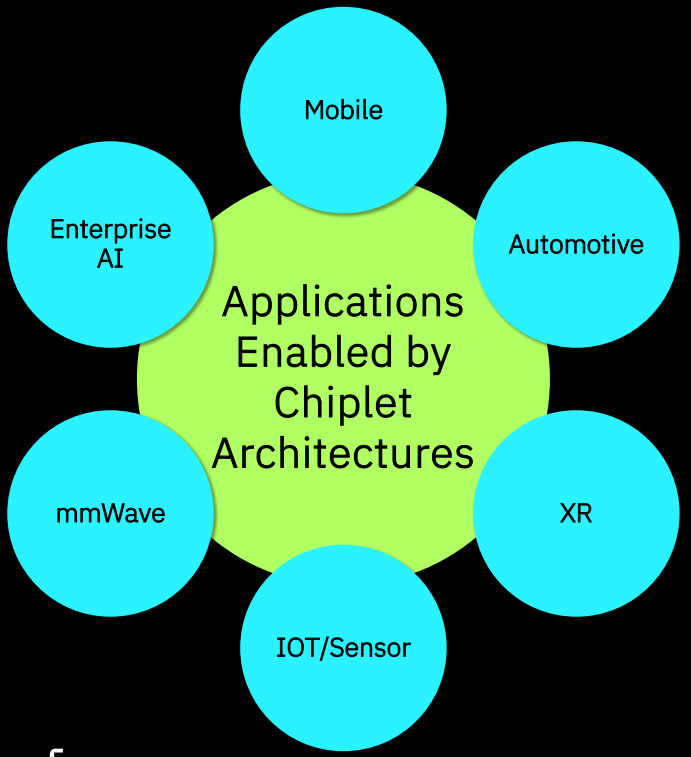
Artificial Intelligence
Very demanding Compute, Memory, and Connectivity requirements



Chiplet Architectures

Promising path for AI:

IBM investing in key technologies to support chiplet architectures



the future of AI Hardware