Data Domain-Specific Architecture with CXL: Breaking the Memory Wall

Jin Kim, CEO/MetisX

MetisX
Xcelerate Your Intelligence

Memory and Beyond Memory

Flash Memory Summit

Can memory become
the center of a computer system?

AI is based on a vast amount of data,
Data resides in memory.

Computational storages are great.
But significant abstraction : file system

MEMORY = Passive?

Drastic interface changes beyond byte/block
seems quite challenging

Jin Kim, Memory Solution Architect (~'2022)

15+ years of developing SSDs
Led Next Gen. Solution Architecture (both HW, SW)
Former Vice President at Semiconductor Giant

MetisX

The Emergence of CXL and Computational Memory

Flash Memory Summit

CXL Compute Express Link™

Data-centric computing based on CXL!

Capacity expansion based on SSDs.
Peta-byte scale memory?
Computational storage without a file system?

Less overheads for computation offloading due to cache-coherency

CXL is flexible enough to empower memory smarter.

No drastic interface changes:
PCIe-based, byte-addressable

Jin Kim, CXL-based Fabless Startup CEO (‘2022~)
Developing CXL Computational Memory
Studying/Enhancing Large-scale Data Acceleration
CEO at MetisX

Domain-Specific Architecture

Hennessy & Patterson
Computer Architecture: A Quantitative Approach, 6th Ed.
Ch 7. Domain-Specific Architecture

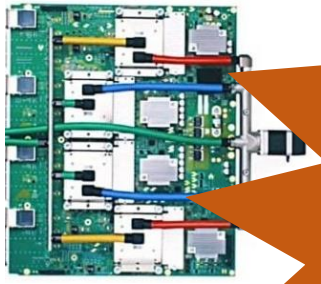
End of Moore's Law &
Dennard Scaling

Minor twists to existing cores: 10% improvements only
Order-of-Magnitude improvements while offering programmability

Need a drastic change in computer architecture:
Domain-Specific Architecture

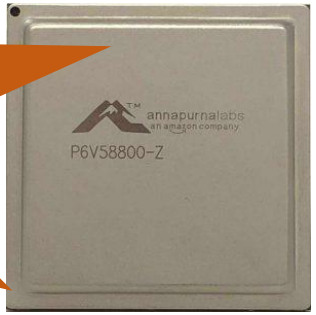
NVIDIA GPU/Tensor Core

Google TPU



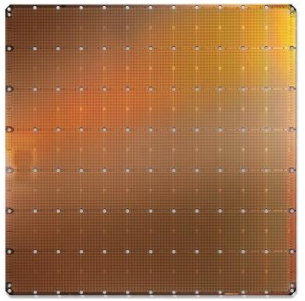
AI Domain-Specific Architectures

AWS Inferentia



Meta MTIA v1

Cerabras



Domain-Specific Architecture

Hennessy & Patterson
Computer Architecture: A Quantitative Approach, 6th Ed.
Ch 7. Domain-Specific Architecture

End of Moore's Law &
Dennard Scaling

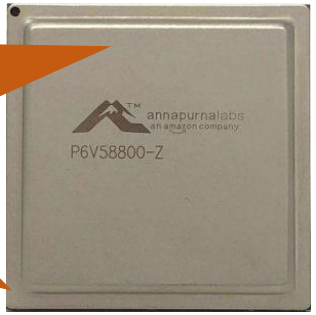
Minor twists to existing cores:
Order-of-Magnitude improvements
programmability

Need a drastic change in computer architecture:
Domain-Specific Architecture

NVIDIA GPU/Tensor Core

Google

AWS Inferentia

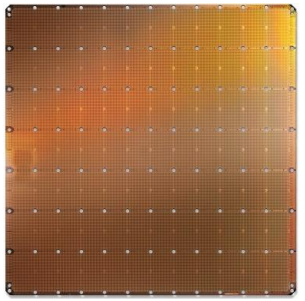


Do we only have AI-related Problems?

AI
Domain-Specific
Architectures

Meta MTIA v1

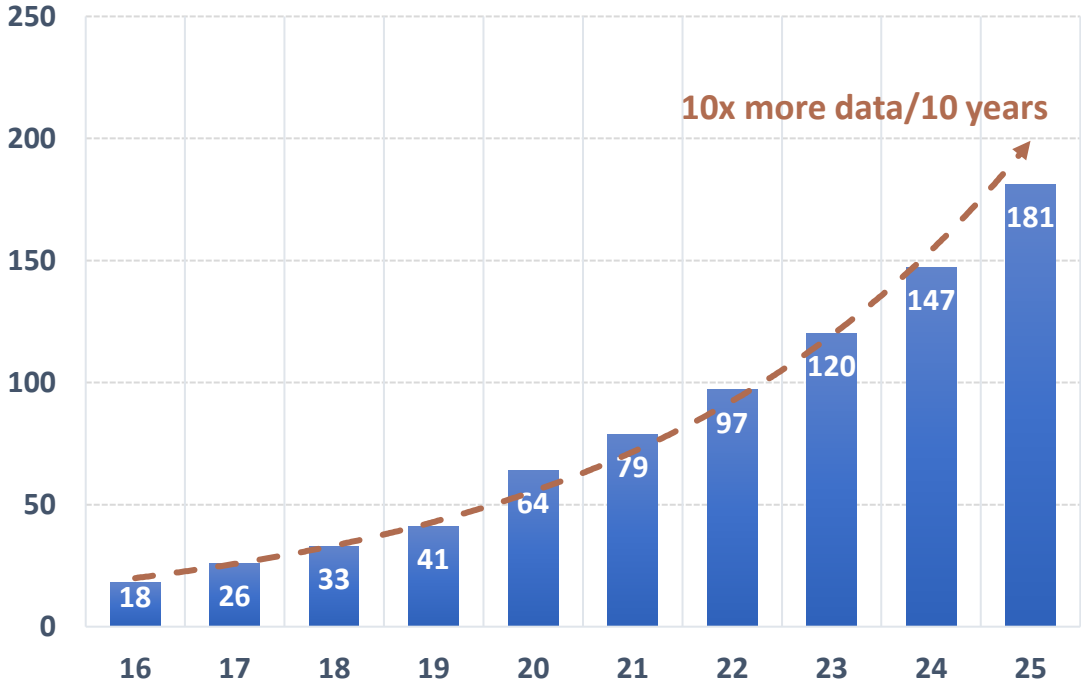
Cerabras



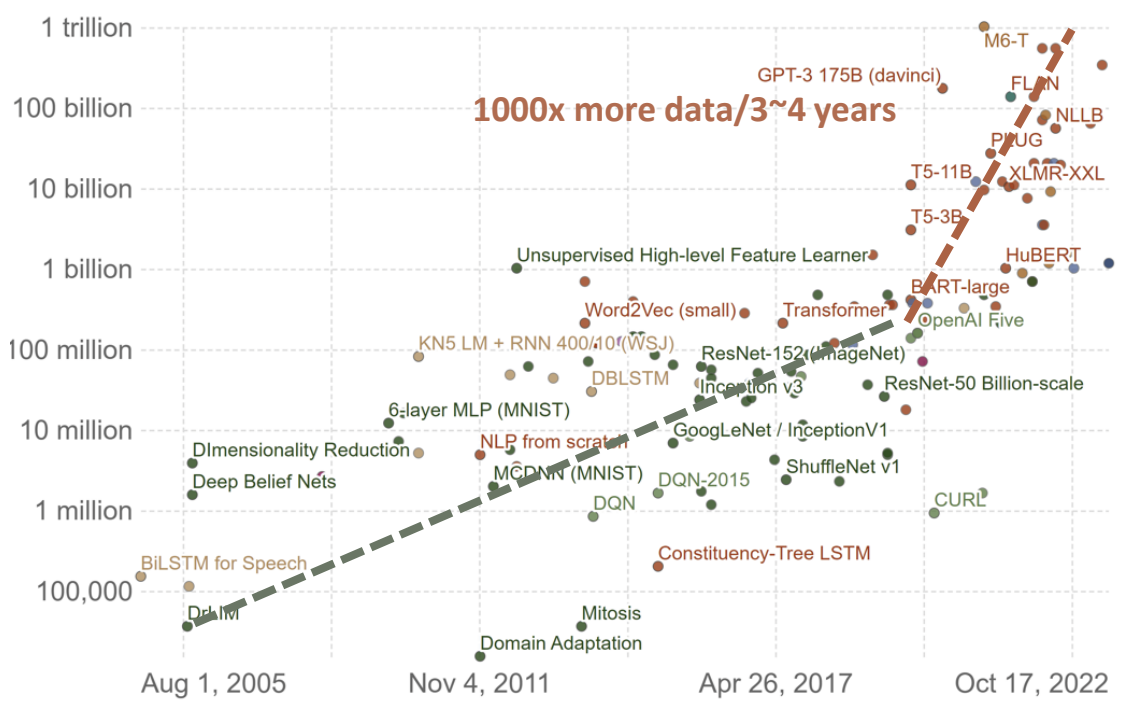
Data and AI Model Growth: Data Domain Problems = Memory Wall

Flash Memory Summit

Data Growth Trend



AI Model Growth Trend



"We need to refine and process such a vast amount of data to create value from it."

"AI revolution is primarily based on data. Moreover, recent LLMs utilize an enormous number of parameters."

MetisX

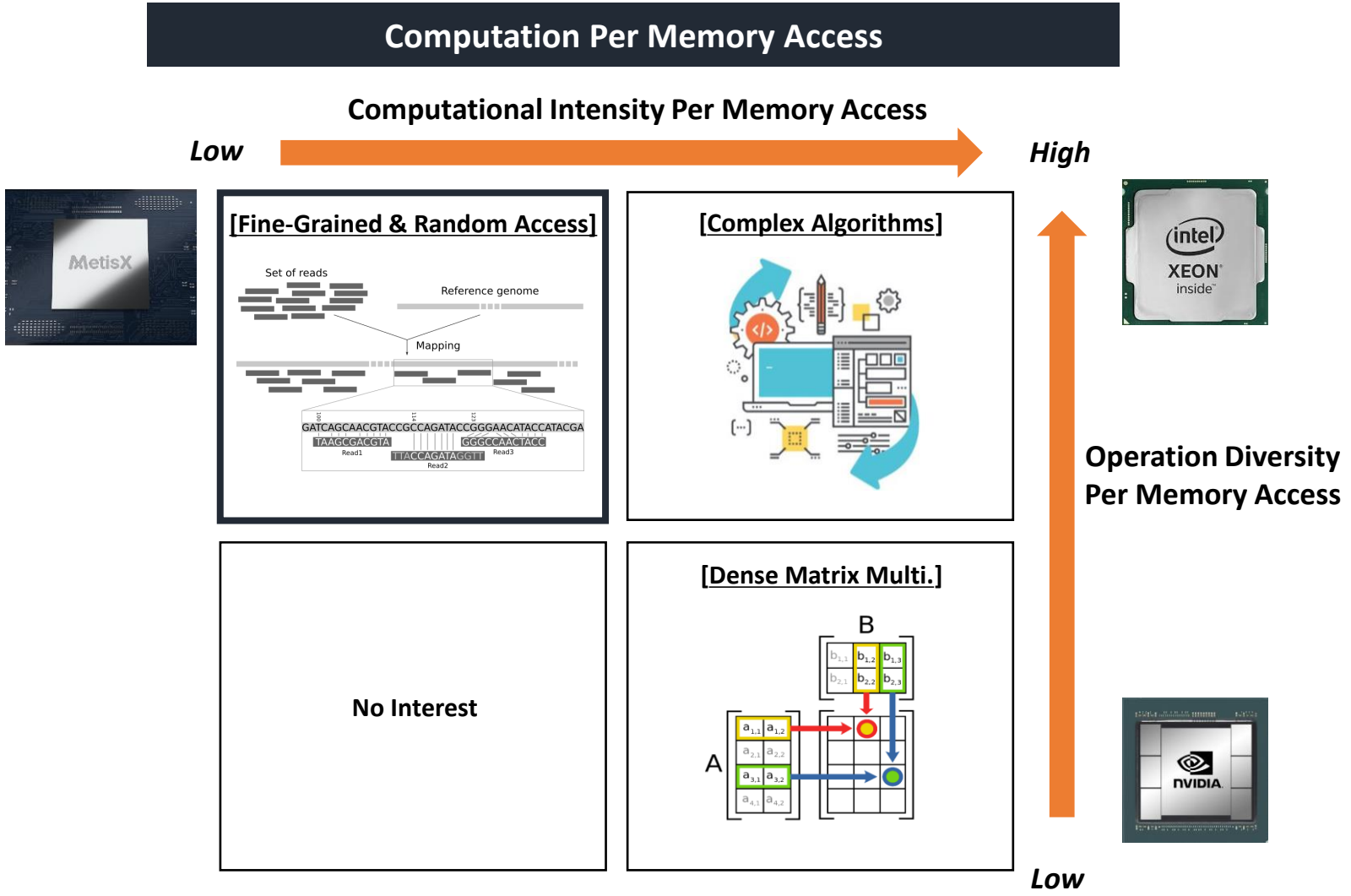
Where Computational Memory Should Focus On

Large-scale Data Processing

Large-scale Data Sets
Memory Latency/Bandwidth Bounded
Highly Parallelizable
Relatively Low Arithmetic
with Several Conditional Branch Operations

Applications

Vector Databases for AI
OLAP Database for Data Analytics
Graph Databases for Social Networks
DNA Analysis for Bioinformatics
Data Compression
……



Where Computational Memory Should Focus On

Flash Memory Summit

Computation Per Memory Access

Larg

Memory L

with Several Co

OLAP

Graph Databases for Social Networks

DNA Analysis for Bioinformatics

Data Compression

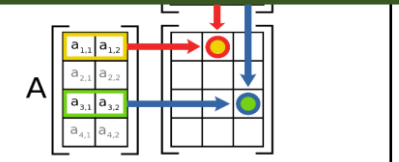
……

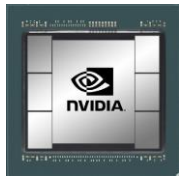
"Memory should be general, even in the computational context.
Acceleration must not be HW-defined but should be fast and efficient!"

Data Domain-Specific Architecture

1. *SW-Defined acceleration to utilize memory b/w far beyond its maximum extent*
2. *Highly Parallelized/scalable architecture*
3. *Host-Device flat shared virtual memory space*

No Interest





Operation Diversity Per Memory Access

Low

MetisX

MetisX Data Domain-Specific Architecture: Hardware Perspective

Flash Memory Summit

SW-Defined Acceleration Mechanism for the Computational Memory
Vertically Optimized Architecture: No Waste of Cycles by Eliminating Any Redundancies
: Data Domain-Specific CXL Differentiations, Cores, Caches, NoC Bus, and Memory Subsystem

CXL Differentiations

- CXL Protocols : .io, .mem, .cache
- Cache-Coherence
- SSD Expansions

Data Domain-Specific Many Cores (~1000s)

- Optimized for Handling Memory
- RISV-V ISA-based Microprocessor
- Extremely Parallelized, Scalable Architecture

Data Domain-Specific Multi-layer Cache with TLB

- Workload-aware Cache
- Supporting TLB
- CXL Cache-coherence
- Full Range Reorder
- Flexible Parameters

Data Domain-Specific NoC Bus

- Vertically-Optimized NoC: 3 types of Bus
- Minimizing Packetizing, Reordering Overhead
- Optimized Decoder and X-Bar Architecture

Data Domain-Specific Memory Subsystem

- Optimized DRAM Scheduling
- Small Area, Low Power without Redundancy
- Co-Optimized with System-Level Cache/Bus
- Enhanced DRAM RAS

MetisX Data Domain-Specific Architecture: Software Perspective

SW-Defined Acceleration Framework for Development and Execution
Acceleration Proof-of-Concept for Large-scale Data Programming Applications in Data Center
: Bioinformatics, Databases, AI, and others (Compression, Homomorphic Encryption, …)

SW Dev. Framework

- C/C++ Compiler
- Offloading Framework
- Acceleration Library

System-Level Simulator

- Architecture Exploration (HW-SW Co-architecting)
- HW, SW Validation
- Easy/Faster Offloading SW Development

Data Domain-Specific Applications: Bioinformatics

- NGS DNA Sequencing Acceleration (BWA, Variant Calling)
- Protein Sequencing
- Dozens of hours → Dozens of minutes

Data Domain-Specific Applications: Databases

- Query Engine Offloading (Spark, PostgreSQL, …)
- Graph database offloading (Neo4j, …)
- Transaction Offloading

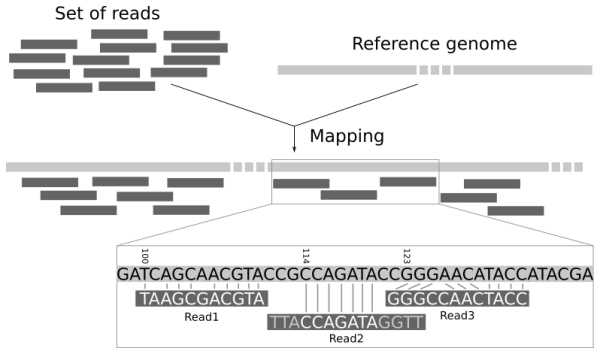
Data Domain-Specific Applications: AI

- Vector Database Acceleration for LLM
- Vector Embedding Acceleration for DLRM

CXL Computational Memory Applications (1/2)

Flash Memory Summit

NGS DNA Analysis

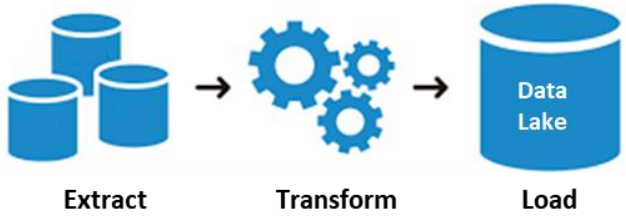


NGS DNA Analysis is transforming the healthcare and medical systems around us. **Human DNA has hundreds of GB of data**.

The current solution based on CPU takes **dozens of hours** to align one human DNA.

CXL computational memory can significantly enhance process performance **in just a matter of minutes.**

Data Analytics

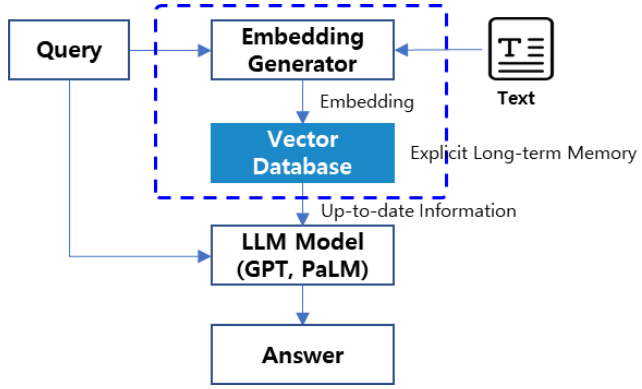


A large volume of data needs to be **processed to create value** from it even before AI training/interference.

Scale-out database clusters like Spark are extensively used in ETL. These clusters typically **consist of numerous servers**.

By offloading the analytics query engine to computational memory, we could significantly **reduce the cluster size**.

AI Vector Databases



Recent LLMs utilize **vector databases** to retrieve updated information after training.

To curb the rapid increase in model size, **vector databases are expected to be utilized more intensively.**

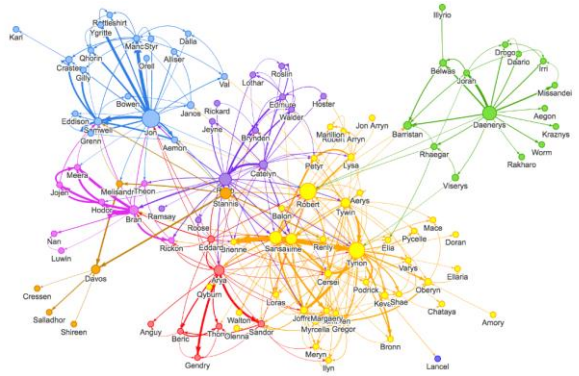
The acceleration of vector databases in memory can play a crucial role in the advancement of LLMs.

MetisX

CXL Computational Memory Applications (2/2)

Flash Memory Summit

Graph Databases

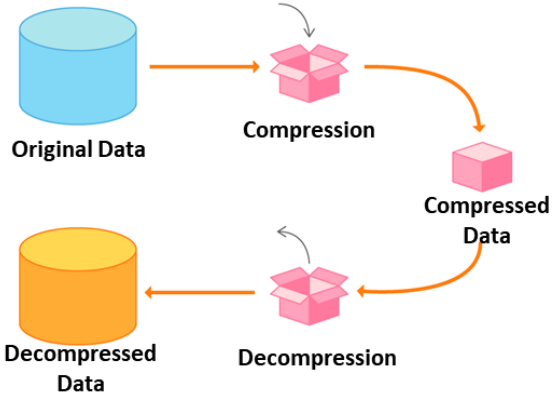


Graph databases are extensively used in social networks **handling enormous amounts of data based on nodes and relationships.**

Graph algorithms mostly involve traversing the relationships between nodes.
The key is **to traverse pointers in parallel**.

Many small cores with memory-optimized architecture are much more suitable for handling pointer traversing than CPUs.

Data Compression

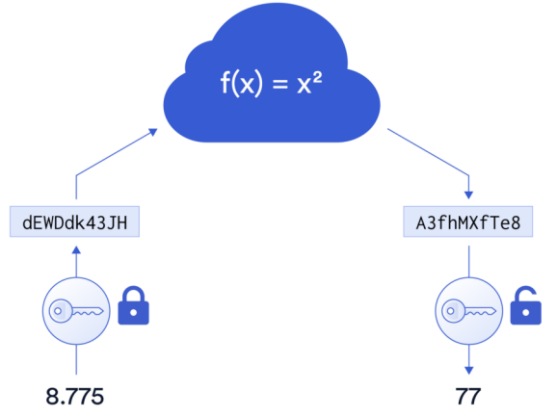


Almost every **data processing inevitably involves compression**. The data growth makes it even more unavoidable.

Using CPU for compression lacks efficiency while leveraging **dedicated accelerators introduces complexity and data copying**.

By offloading the compression to memory, it becomes possible **to compress and decompress data faster and seamlessly during data processing**.

Homomorphic Encryption



Homomorphic encryption is expected to be **crucial to handling sensitive and important** data. It could be an essential part of data processing

While there're advanced algorithms like CKKS, they're still **slow to process practically on CPUs**.

By embedding homomorphic encryption into computational memory, it is possible to **seamlessly perform data processing**.

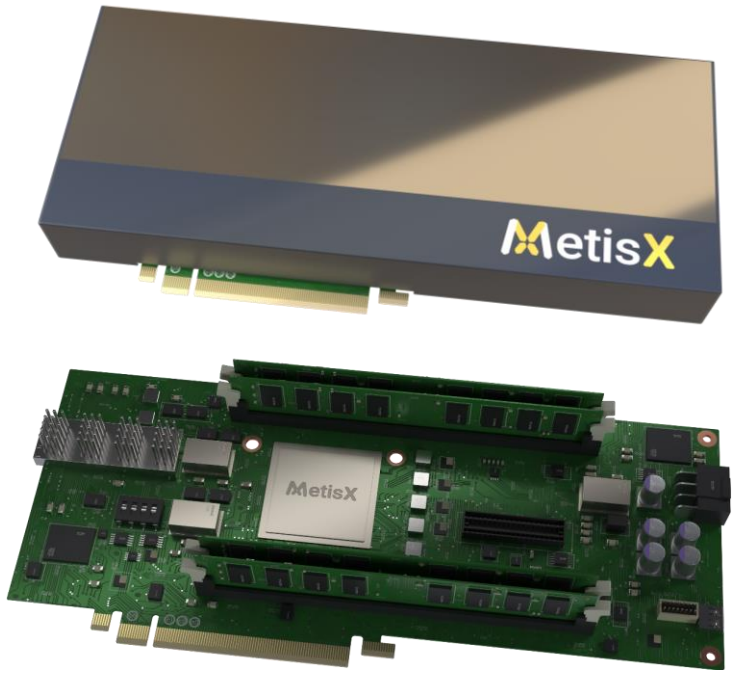
MetisX

MetisX CXL Computational Memory

Flash Memory Summit

MetisX CXL Computational Memory

"By Leveraging Data Domain-Specific Architecture and CXL"



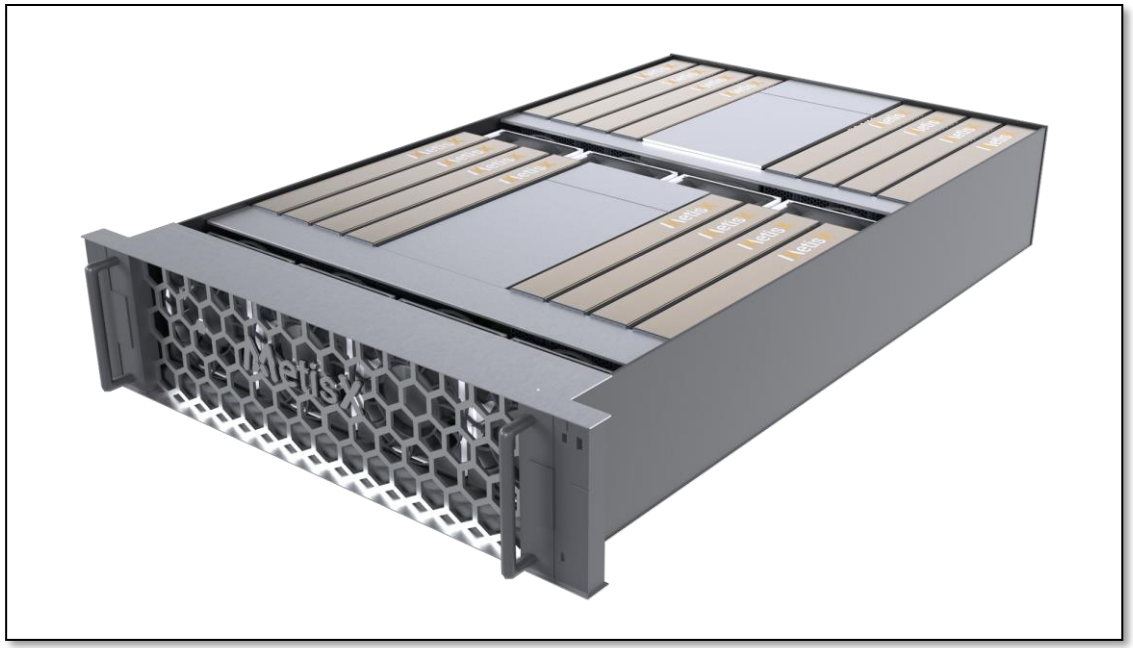
Key Differentiations

- **Cache Coherent Large-Scale Data Acceleration using a SW-defined Method**
  w/ 1000s of MetisX Optimized Cores, Caches, Memory Subsystems
  : AI, DB, Vector Embedding, Graph, NGS DNA Analysis, Compression, etc.
- **Petabytes-Scale Infinite Memory with NVMe SSD Expansion**
- **Strong Software Framework for Application Offloading**
- **Enhanced Reliability with Multi-Symbol Correction for DRAM**
- **Supporting the Expansion of Accelerators like NPU**
  : The Integrated AI Acceleration Solution for LLM

MetisX CXL Computational Memory Pool and Infinite Memory Pool

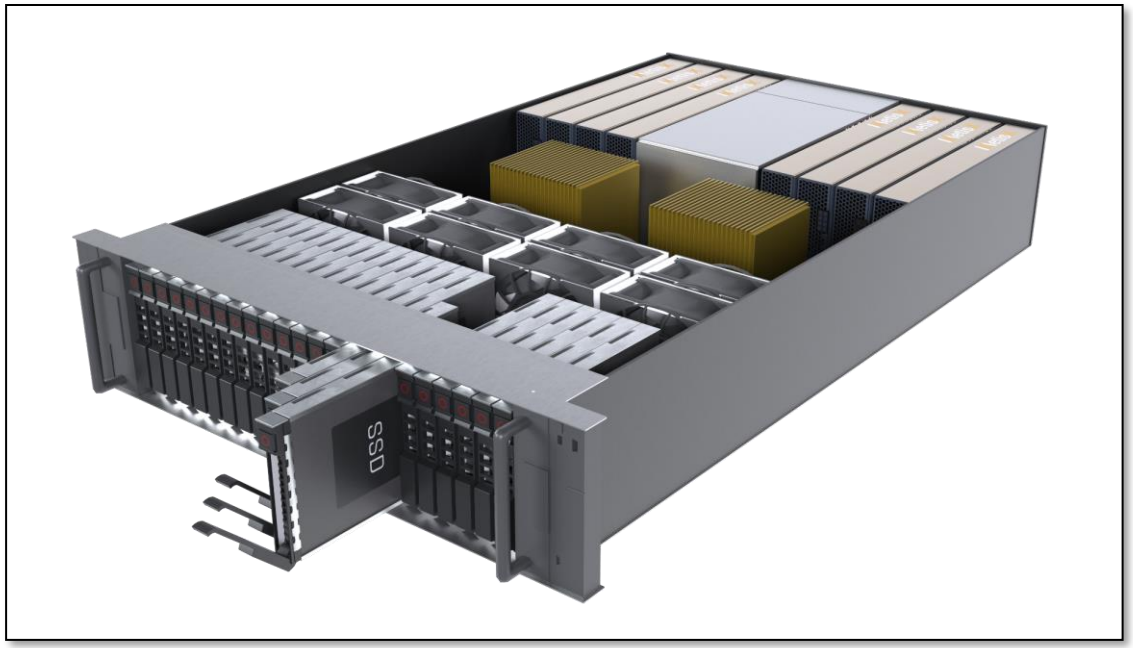
Flash Memory Summit

CXL Computational Memory Pool



- **Disaggregated Memory Pool with Large-scale Data Acceleration**
- **Eliminating Long Latencies from the Far-Memory**

CXL Petabyte-Scale Infinite Memory Pool



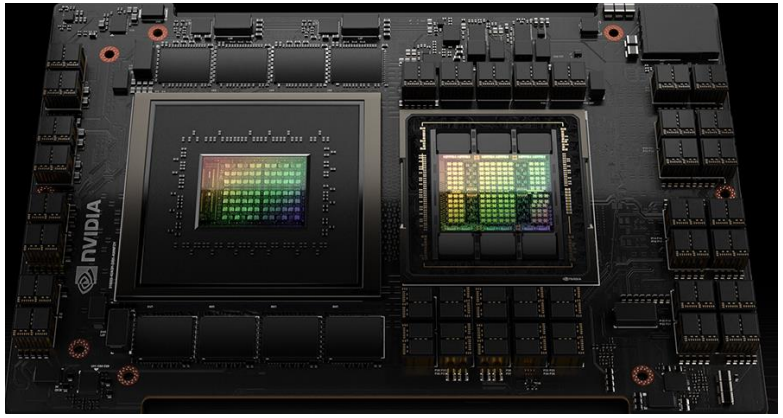
- **Disaggregated Petabytes-scale Memory Pool with Acceleration Sophisticated Workload Management**
  : Latency/Capacity Tiering, Thin Provisioning, Workload Isolation, Nonvolatility

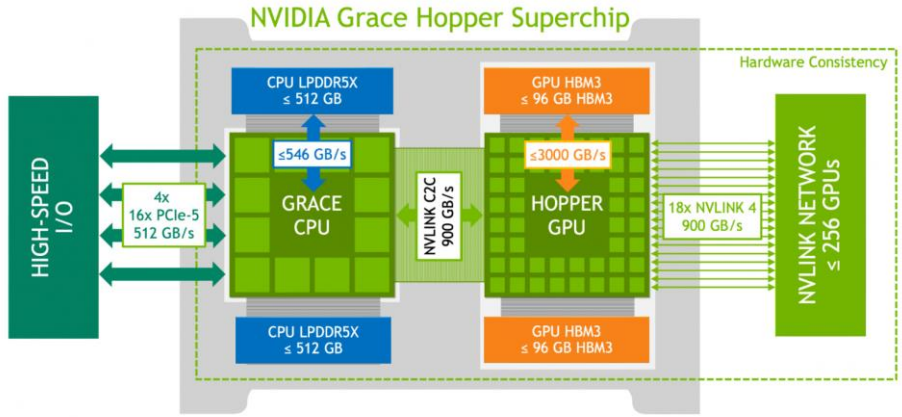
MetisX

Integrated AI-Memory Acceleration Solution for LLM

Flash Memory Summit

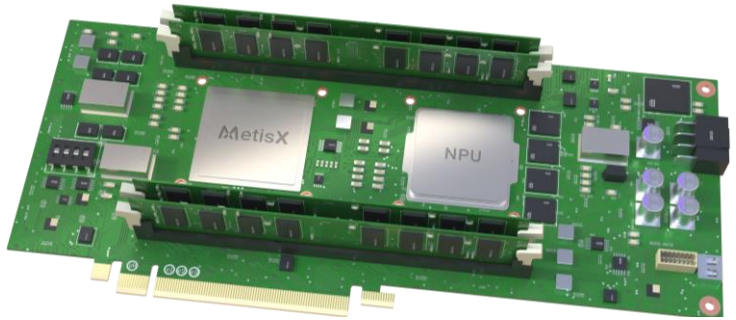
NVIDIA Grace Hopper (GH200)

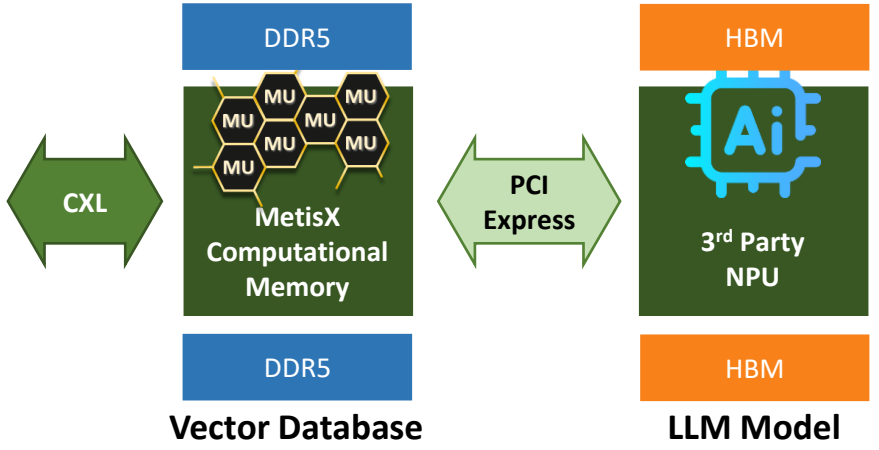




Integrated AI Acceleration Solution for LLM

"Modern LLMs can be highly optimized through integration with computational memory that supports vector database acceleration."





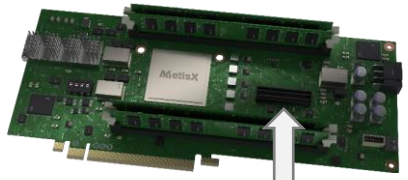
MetisX

CXL-based Data-centric Computing World

"To overcome the memory wall, we don't need another data accelerator. We need to empower memory to be more intelligent."

CXL Computational Memory

Large-Scale Data Acceleration



PCIe

Petabytes-Scale Memory Expansion

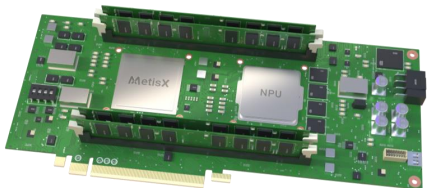
SSD

Integrated AI Acceleration Solution for LLM

CXL

CPU

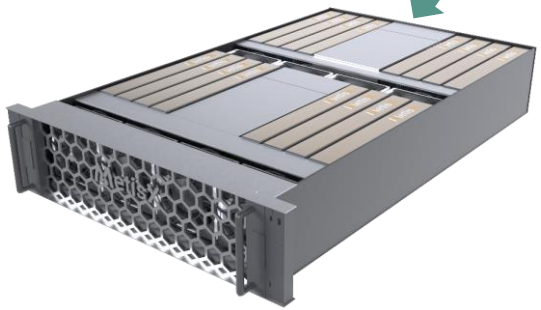
CXL

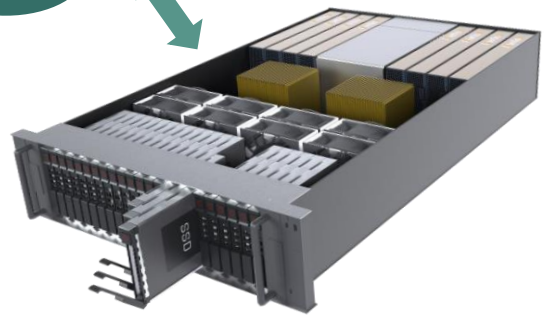


3rd Party NPU Expansion

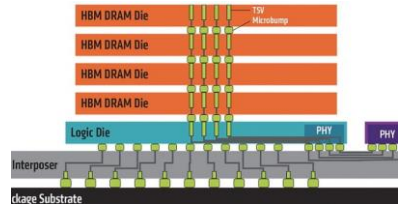
Rack-scale Pooled Memory

CXL Fabric





HBM Base Die Integration



Data DSA IPs on HBM base die for the Computational HBM

CXL Computational Memory Pool

CXL Petabyte-Scale Infinite Memory Pool

MetisX

Thank You

Please Visit MetisX Booth at #1046, the Yellow One.
Contact: Jin Kim jin.kim@metisx.com

More empirical presentation from MetisX
SARC-303-2, Thursday, Aug 10th, 11:00~12:05 PM
"Data Acceleration Approaches on the CXL Memory", Harry Kim, CPO/MetisX