

# Data Acceleration Approaches on the CXL Memory

Harry Kim / CPO



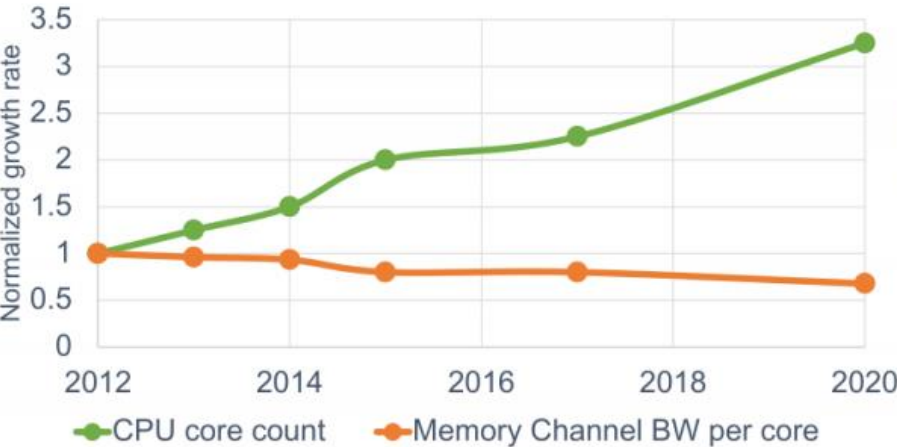
© 2023 MetisX Co., Ltd. All rights reserved.

# Need More Memory & More Performance

Data is Growing, Core Count is Increasing, We can Attach More Memory through CXL.  
How about the Overall System Performance?

## Memory BW Limit

### Increasing Core Counts Drives Growth



SERVER

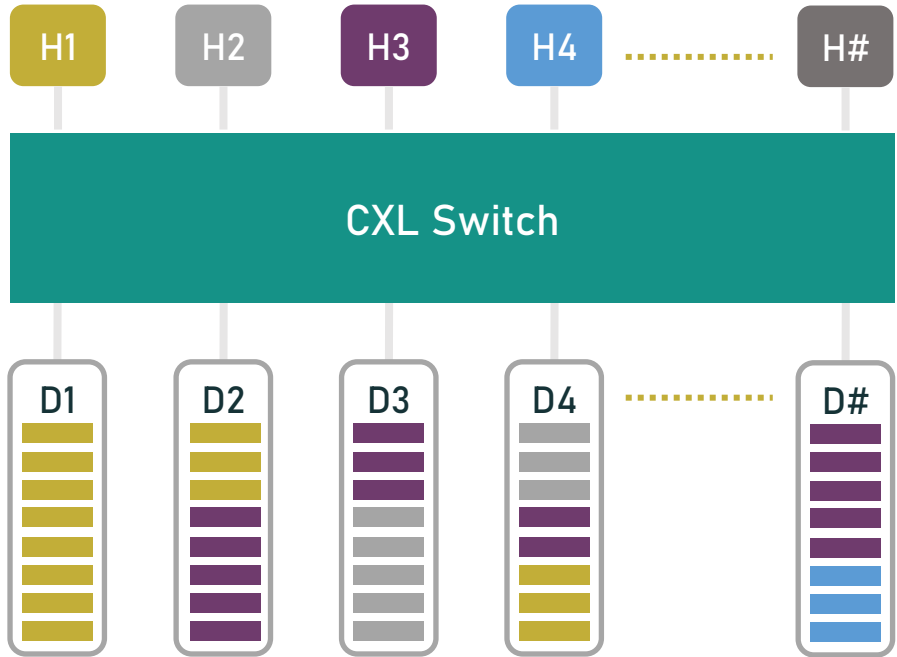
Increasing core counts driving memory demand

- Increased Bandwidth
- Increased Capacity



OPEN POSSIBILITIES.

## CXL Memory Expansion



DIMM-based memory subsystem will reach its limit by 2025

**Need Beyond Just Another CXL Memory, Capacity Expansion is not Enough.**

**Performance Scaling and Integration is the Key.**



## What We Want

- ✓ Large Memory
- ✓ High Bandwidth, Low Latency
- ✓ Better Performance
- ✓ Power Efficient



## What the Real Challenges are

- ✓ Slow DRAM Density Scaling
- ✓ Physical Limit on Interface Speed-up
- ✓ Additional Latency/Power over Far Memory
- ✓ Unexpected Performance Impact



## Comprehensive Solutions

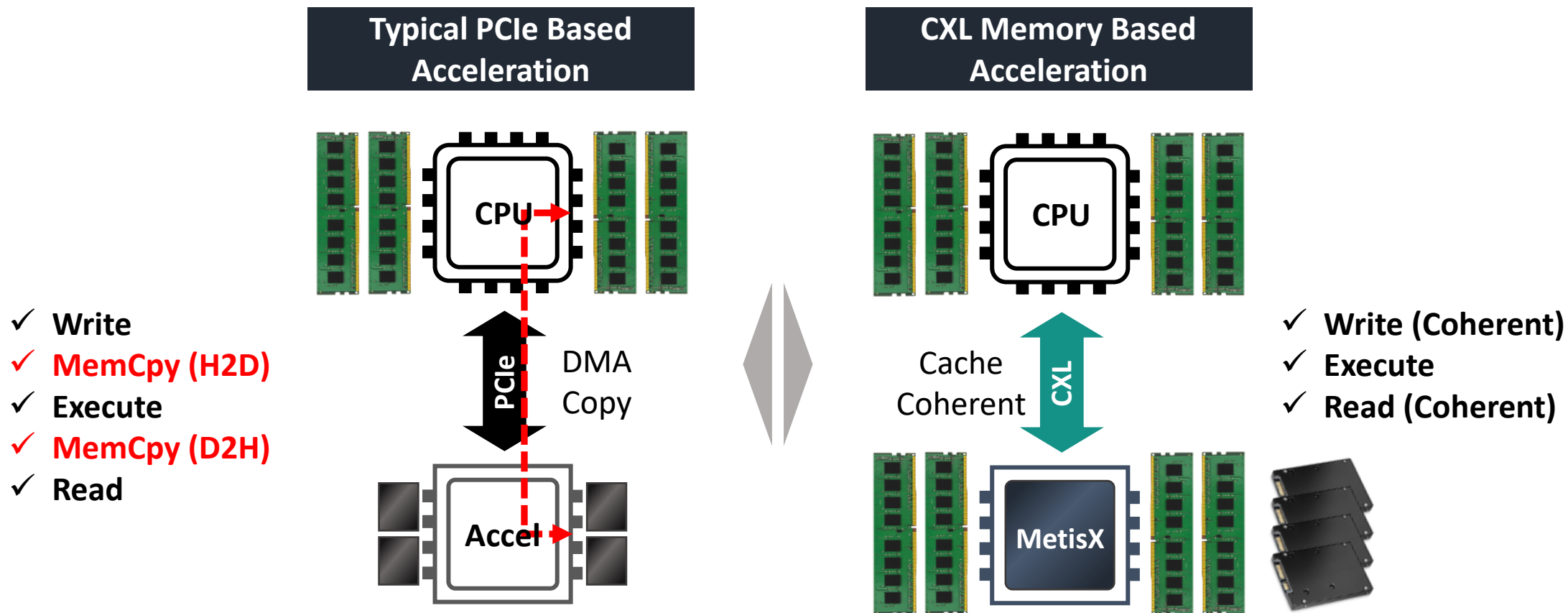
- ✓ Memory Scales with Computing Capability
- ✓ Accelerate-in-Memory for Less Data Movement
- ✓ Easy and Intuitive Architecture
- ✓ Integration with Production Ecosystem



# Differentiation of CXL Memory Based Acceleration

Explicit Data Movement is NOT Necessary.

Fine-grained Offloading is Possible since NO Copy Overhead.



No need to Copy, Data is Already There.

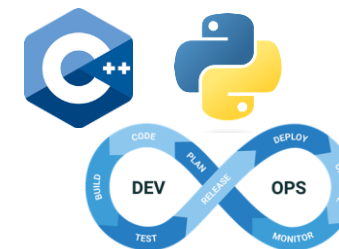
Easy to Use : Support Popular Languages, Intuitive Memory Model, Common Execution Model

→ Continuous Integration of the SW Development/Deployment Cycle

## SW-Defined

C/C++ and Support for Python/Java/+ Binding

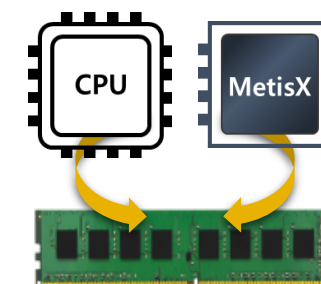
- ✓ Not Limited to a Specific Function/Format
- ✓ Easy Integration for Development and Deployment



## Memory Model

Host-Device Shared Virtual Memory w/ Fine-grained Cache Coherency

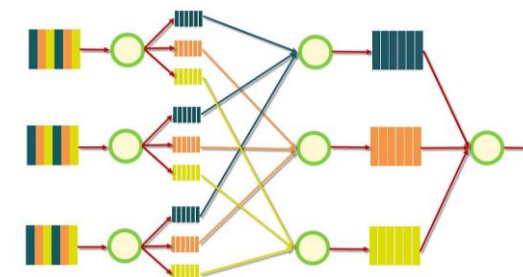
- ✓ Unified Virtual Addressing
- ✓ Simple Coherent Memory View, NUMA-aware



## Execution Model

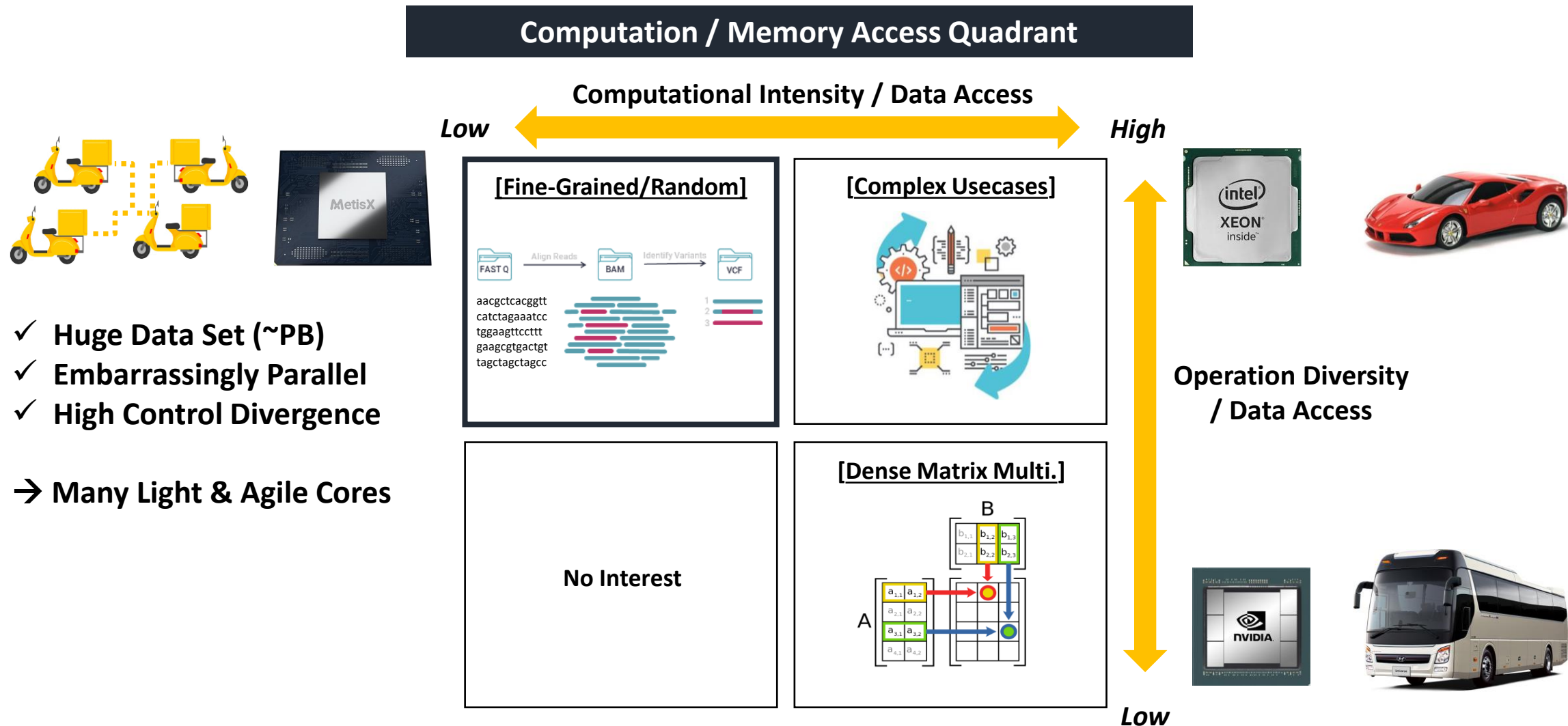
Parallel Big Data Processing Framework

- ✓ MapReduce Framework for ManyCore to Work in Parallel
- ✓ Data Xfer and Message Passing Interface



# Data Domain Specific Architecture

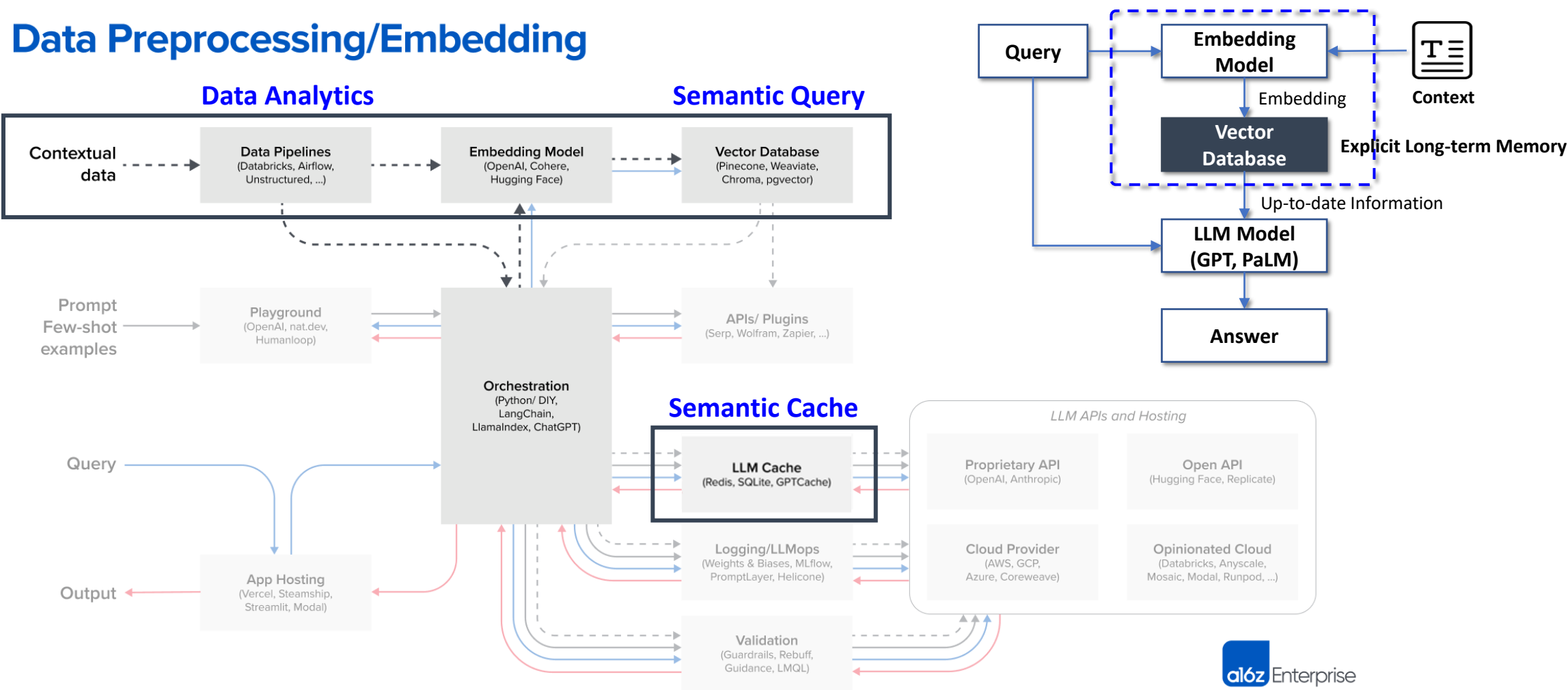
Memory Hungry Applications need Data Domain Specific Architecture.



AI needs More Data Accelerations.

→ Scalable In-Memory Databases for High Performance, The More The Better.

## Data Preprocessing/Embedding





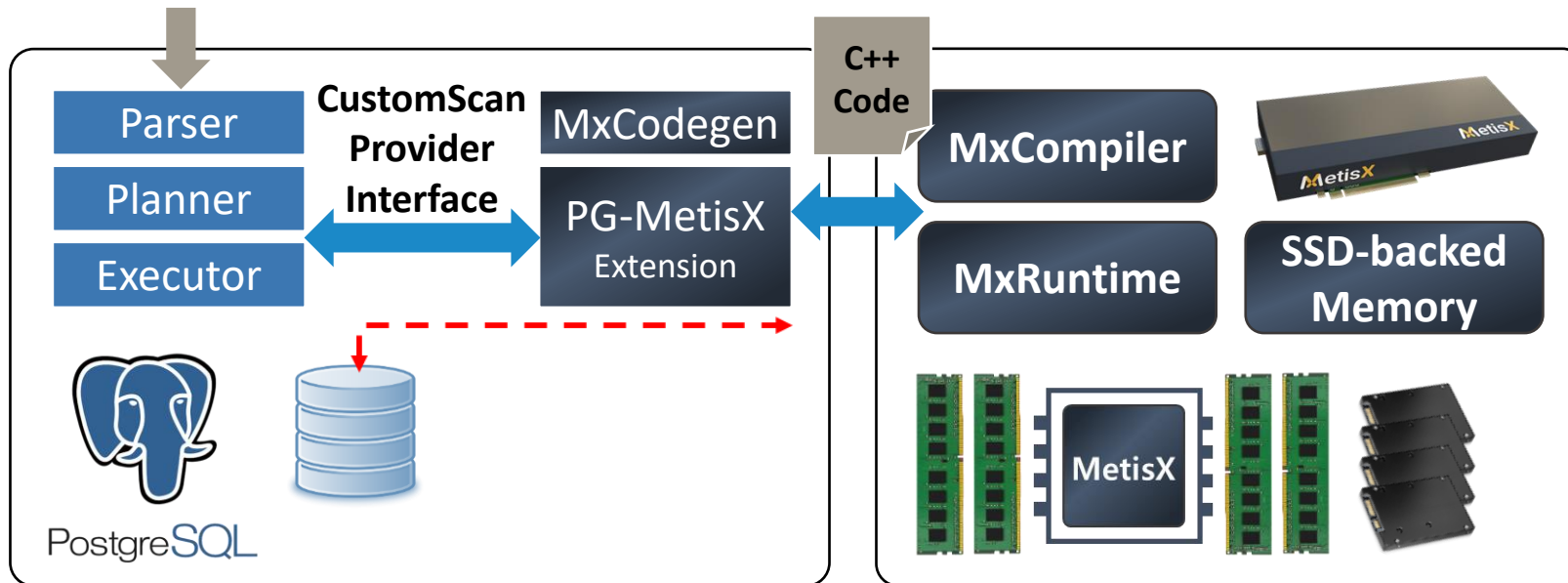
## Database Query Acceleration with CXL Computational Memory

→ ETL(Extract/Transform/Load), Data Pre-processing, Big Data Analytics

## MetisX Runtime Seamlessly Integrates SSD-backed Memory

→ Data Resides in CXL Memory, Reducing the Need for Data Movement

### 【 Query Acceleration with MetisX Computational Memory 】



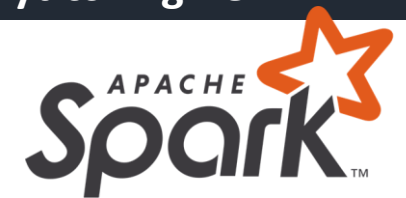
Short Queries can be Accelerated as well  
by Removing Data Movement Overhead.

### 【 For Big Data Analytics 】

Columnar Format for Analytics



Big Data Analytics Engine

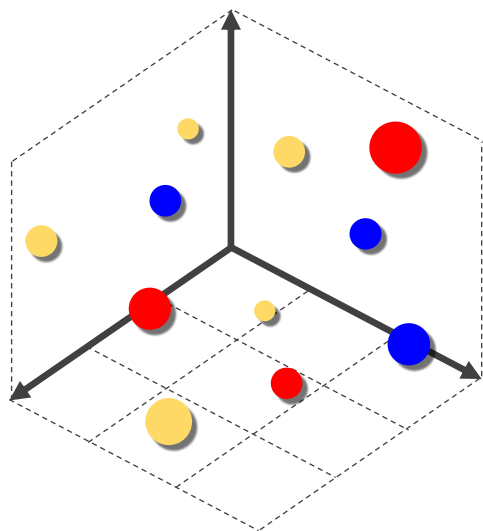




Vector Database is an Essential component for operating modern Large Language Models(LLMs), Requiring Significant Memory Capacity and Fast Response Time.

- Distance in the Vector Space implies Semantic Similarity
- Hard to Achieve the Performance with Traditional Compute Architecture (CPU/GPU)

## High-Dim Vector Space



More Than Thousands of Dimensions

## Vector Indexing Graph

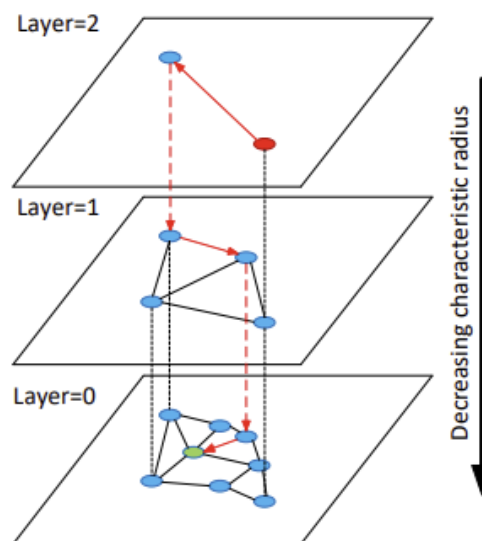


Fig. 1. Illustration of the Hierarchical NSW idea. The search starts from an element from the top layer (shown red). Red arrows show direction of the greedy algorithm from the entry point to the query (shown green).

(Yu. A. Malkov, et al. Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs)

## VectorDB Semantic Query

- ✓ Large and Growing Data Sets
- ✓ Random Pointer Traversal
- ✓ High Performance Requirement
- ✓ In-Memory Database

→ CXL Computational Memory

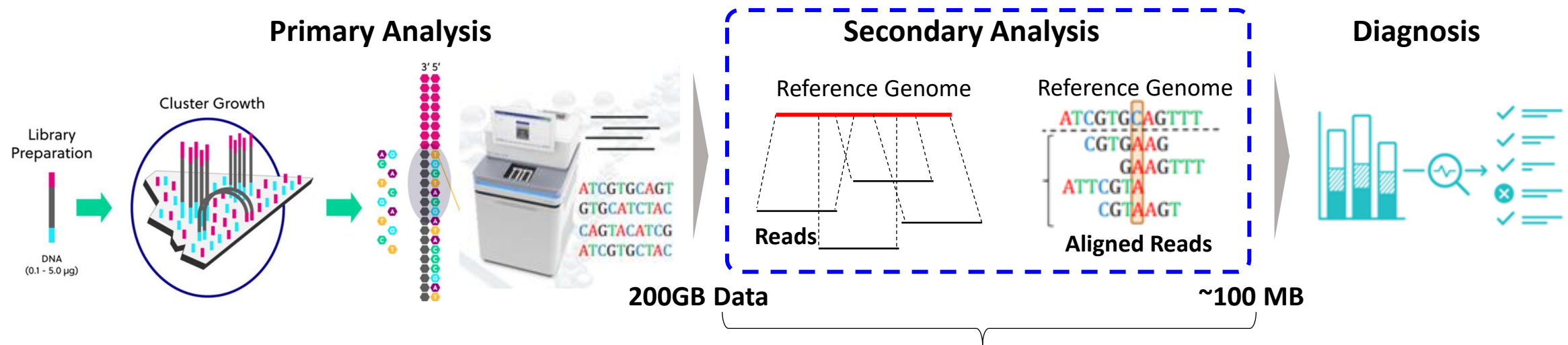
**Call for Collaboration!**

# NGS(Next Generation Sequencing) Acceleration

NGS-based DNA analytics is expected to be a groundbreaking turning point in healthcare, pharmaceuticals, and life sciences.

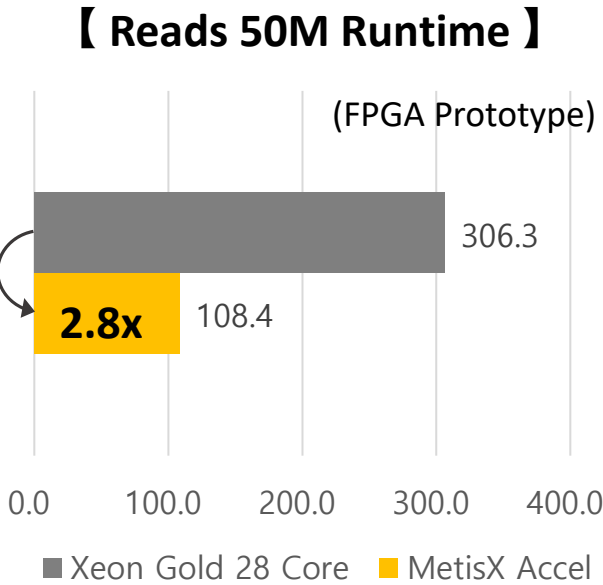
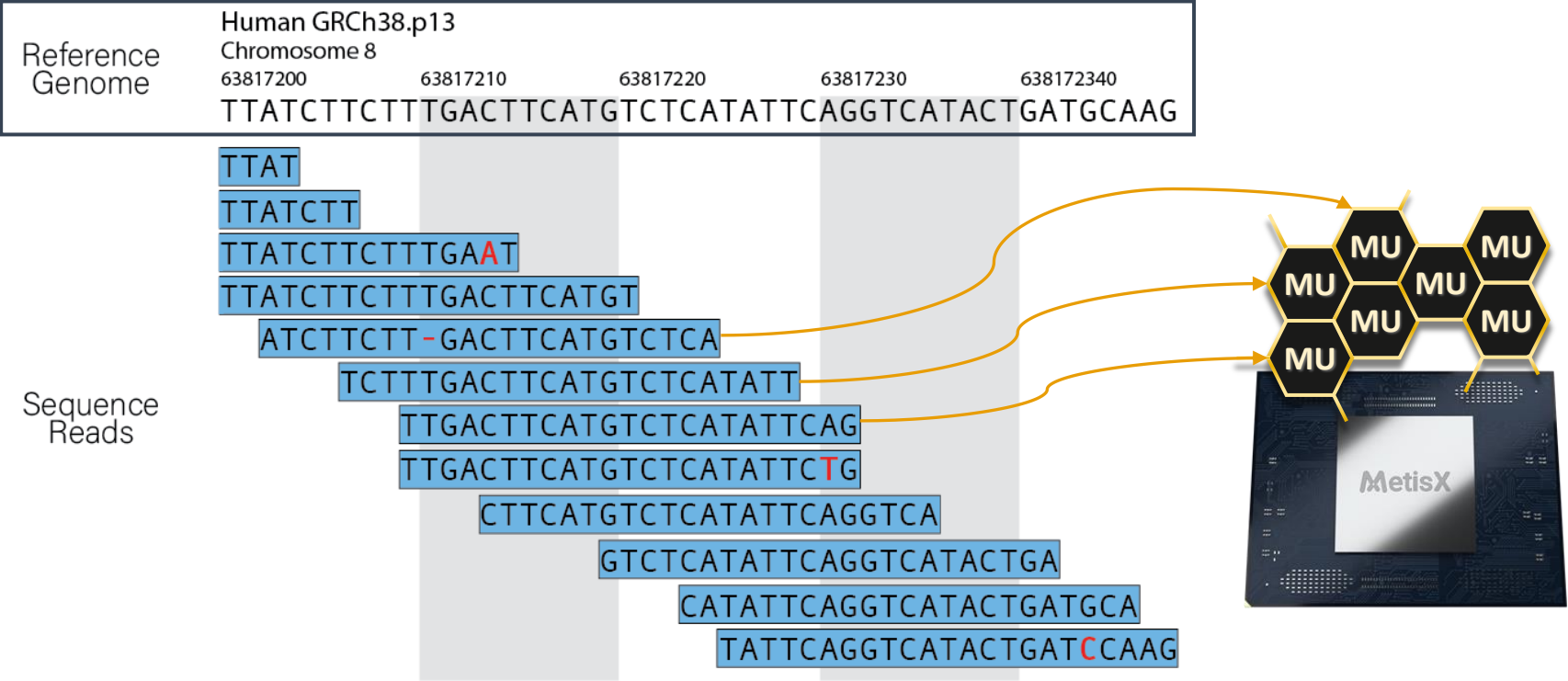
DNA Sample → Library Prep → Read 100~200 Sequence(Optics) → Align with Reference Genome

## NGS DNA Analytics Pipeline



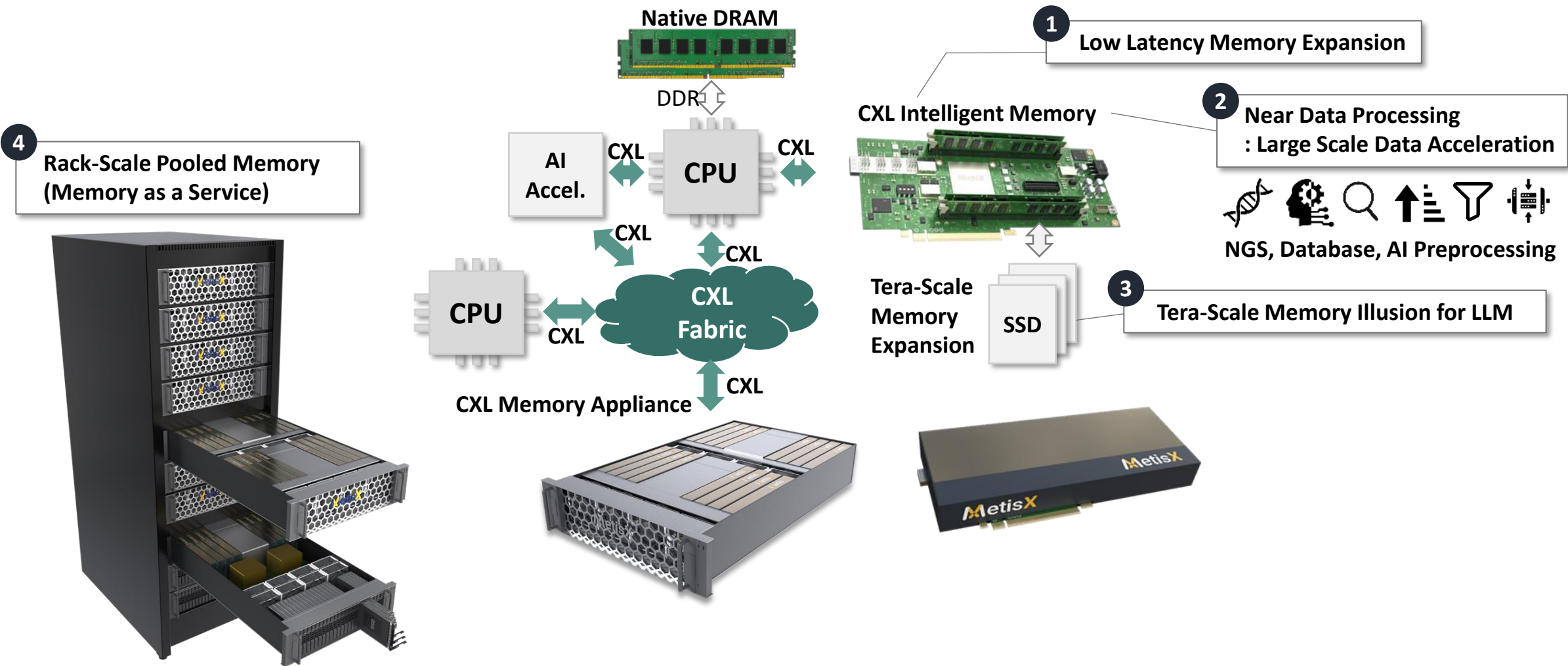
Approximated String-Matching Problem : Difference indicates Genotype/Disease or Error

- Need to Align more than 1B Short Reads for 3B Reference Positions
- Memory and Compute Intensive, Parallel Nature



**50x Improvement with ASIC Projection**

MetisX aims to address the challenges of data explosion by leveraging CXL technology and its own data domain specific architecture.



*Thank You*

Contact Harry Kim [harry.kim@metisx.com](mailto:harry.kim@metisx.com)

Visit **MetisX** booth **#1046**