CXL Computational Memory for the Future of AI Data Centers

JIN KIM / CEO, MetisX Co., Ltd.

BACKGROUND

THE ERA OF AI

THE CORE IDEAS BEHIND
MODERN NEURAL NETWORKS
HAVE NOT CHANGED
SUBSTANTIALLY SINCE THE 1980s.

MOST OF THE IMPROVEMENT
IN NEURAL NETWORK
PERFORMANCE CAN BE
ATTRIBUTED TO 2 FACTORS,

1. LARGER DATASETS
2. POWERFUL COMPUTER

- YOSHUA BENGIO

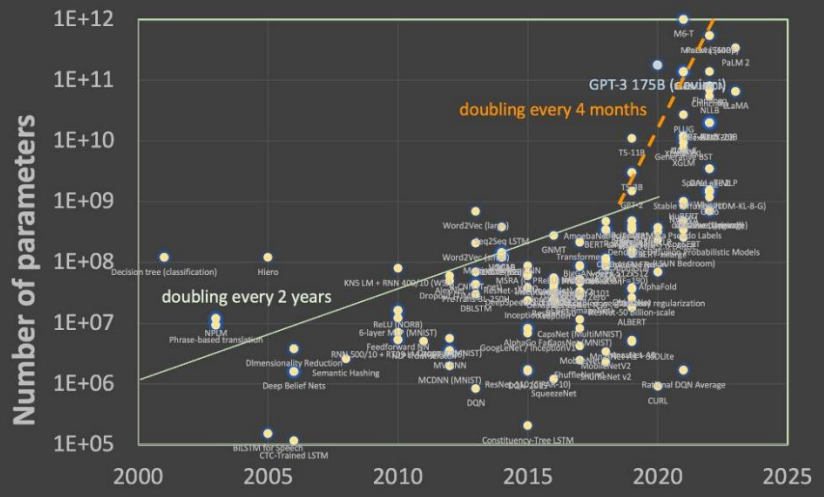
IT IS ALL ABOUT DATA AND DATA RESIDES IN MEMORY

10 YEARS OF GARTNER TOP 10 STRATEGIC TECH TRENDS

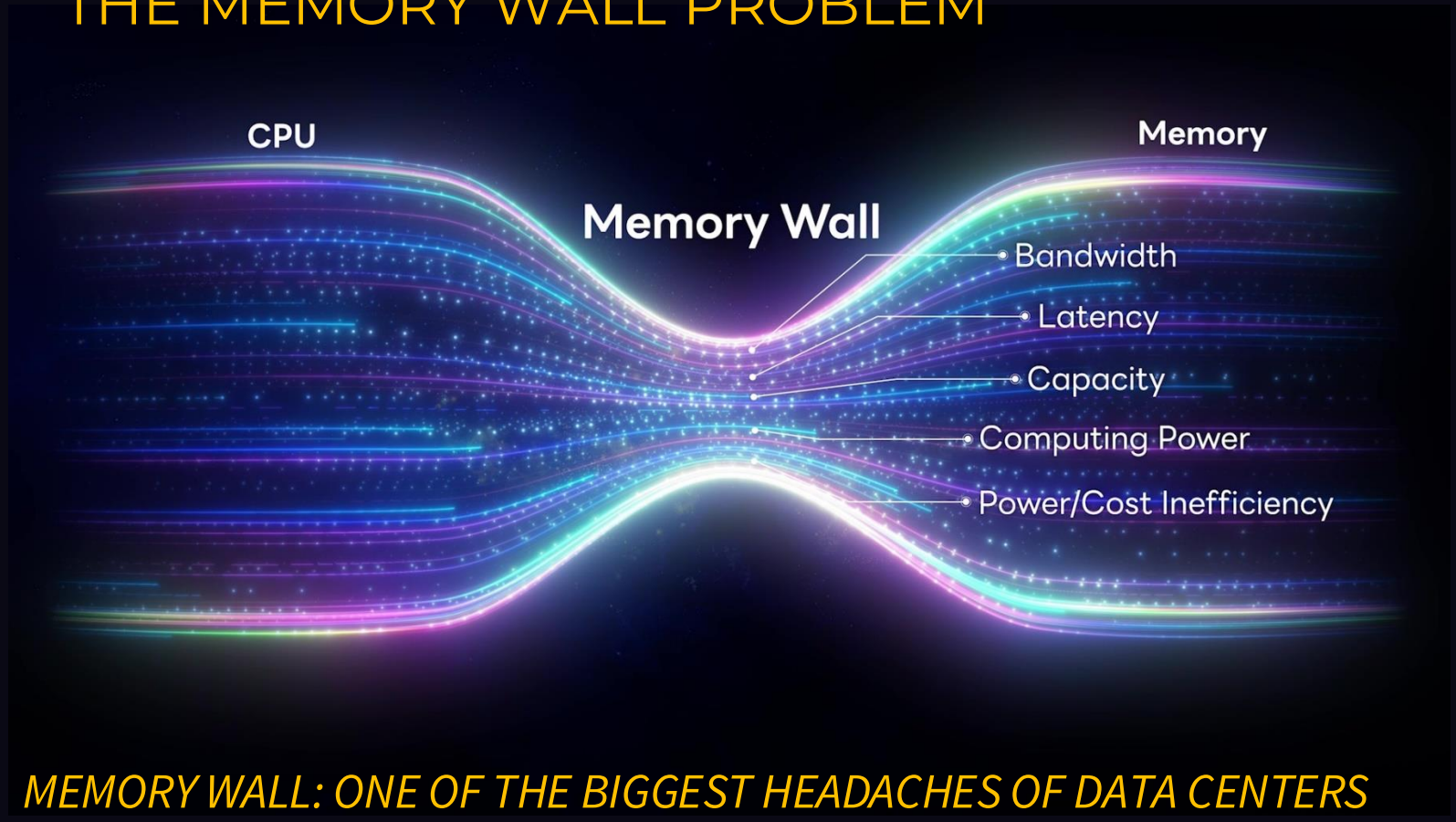
'15	'16	'17	'18	'19	'20	'21	'22	'23	'24
Computing Everywhere	Device Mesh	AI	AI Foundation	Hyper-automation	Autonomous Things	Internet of Behaviors	Generative AI	Digital Immune System	AI Trust
IoT	Ambient User Experience	Intelligent Apps	Intelligent Analytics	Multi-experience	Augmented Analytics	Total Experience	Data Fabric	Applied Observability	CTEM
3D Printing	3D Printing	Intelligent Things	Intelligent Things	Democracy, 2020 style	AI-driven Development	PEC	Distributed Enterprise	AI Trust	Sustainable Technology
Advanced Analytics	Information of Everything	VR/AR	Digital Twins	Augmentation	Digital Twins	Distributed Cloud	Cloud-Native Platform	Industry Cloud Platform	Platform Engineering
Context-rich System	AI	Digital Twins	Cloud to the Edge	Transparency& Privacy	Empowered Edge	Anywhere Operations	Autonomic System	Platform Engineering	AI-Augmented Development
Smart Machines	Autonomous Agents	Blockchain	Conversational Platform	Empowered Edge	Immersive Technologies	Cybersecurity Mesh	Decision Intelligence	Wireless-Value Realization	Industry Cloud Platform
Cloud/Client Computing	Adaptive Security	Conversational System	Immersive Experience	Distributed Cloud	Blockchain	Composable Business	Composable Applications	Super apps	Intelligent Applications
SW-Defined Infra	Advanced System Archi.	Mesh App	Block Chain	Autonomous Things	Smart Spaces	AI Engineering	Hyper-automation	Adaptive AI	Democratized Generative AI
Web-scale IT	Mesh App.	Digital Tech. Platform	Event-driven	Block Chain	Digital Ethics	Hyper-automation	PEC	Metaverse	Augmented Connected WF
Risk-based Security	IoT	Adaptive Security	Adaptive Risk and Trust	AI Security	Quantum Computing	-	AI Engineering	Sustainable Technology	Machine Customers

PROBLEM

DATA IS RAPIDLY INCREASING,
THE INCREASE IN AI MODEL SIZES IS MUCH MORE RAPID



WE NEED TO BREAK
THE MEMORY WALL PROBLEM

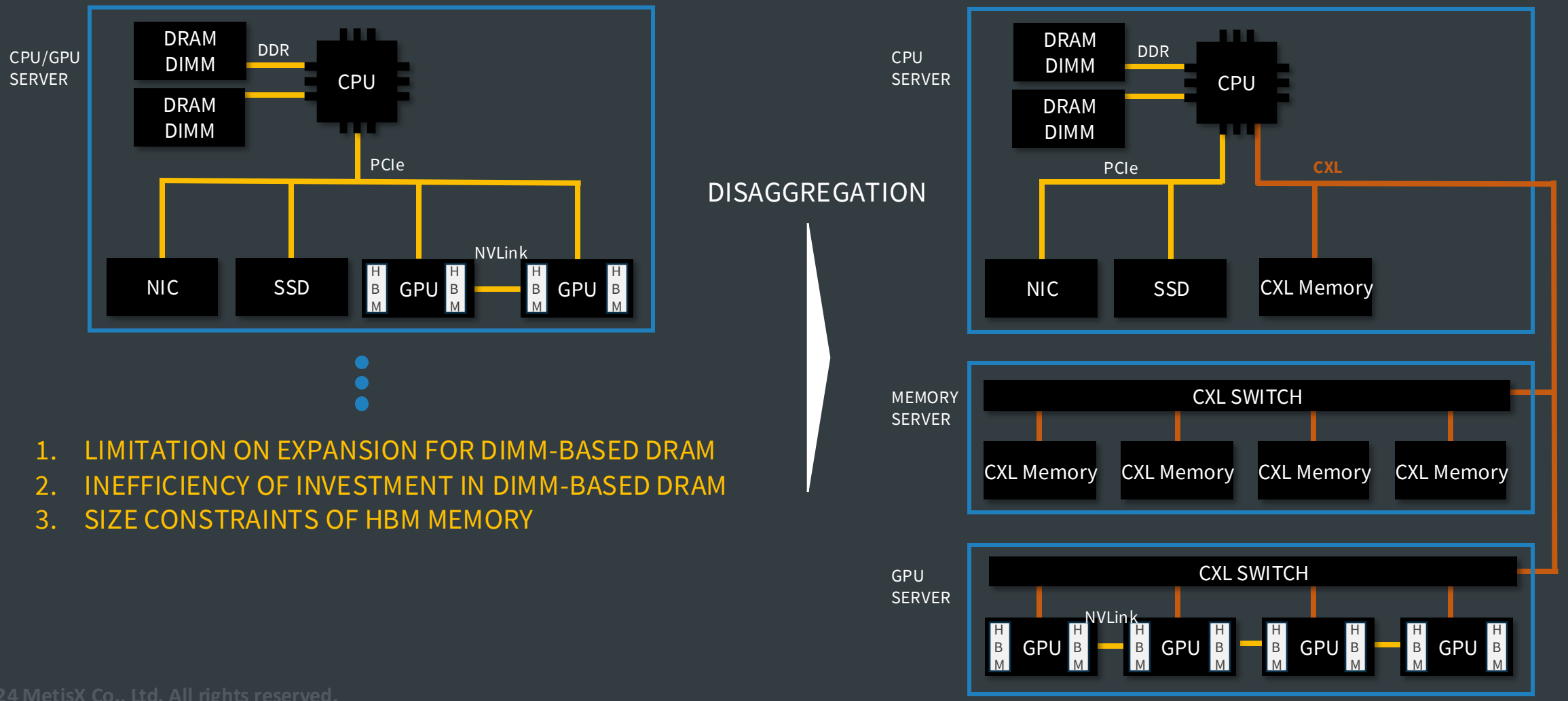


MEMORY WALL: ONE OF THE BIGGEST HEADACHES OF DATA CENTERS

- *SIZE: CHALLENGING TO KEEP UP WITH THE RAPID DATA INCREASE*
- *PERFORMANCE: DATA MOVEMENT IS THE BIGGEST BOTTLENECK*
- *COST: MEMORY IS REALLY EXPENSIVE (HALF OF THE SERVER COSTS)*
- *UTILIZATION: MEMORY UTILIZATION IS VERY LOW*

THE FUTURE OF DATA CENTERS

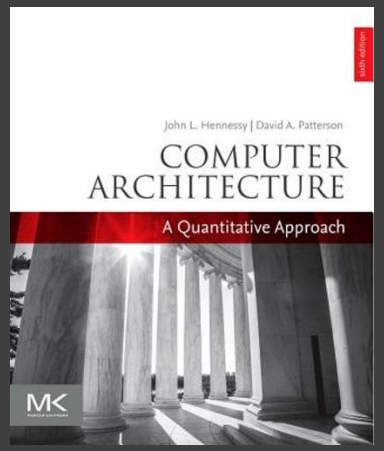
DATA CENTER INFRA WILL BE SIGNIFICANTLY ENHANCED WITH CXL, BUT THE BIGGEST BOTTLENECK WOULD BE DATA MOVEMENT



1. LIMITATION ON EXPANSION FOR DIMM-BASED DRAM
2. INEFFICIENCY OF INVESTMENT IN DIMM-BASED DRAM
3. SIZE CONSTRAINTS OF HBM MEMORY

DATA DOMAIN-SPECIFIC

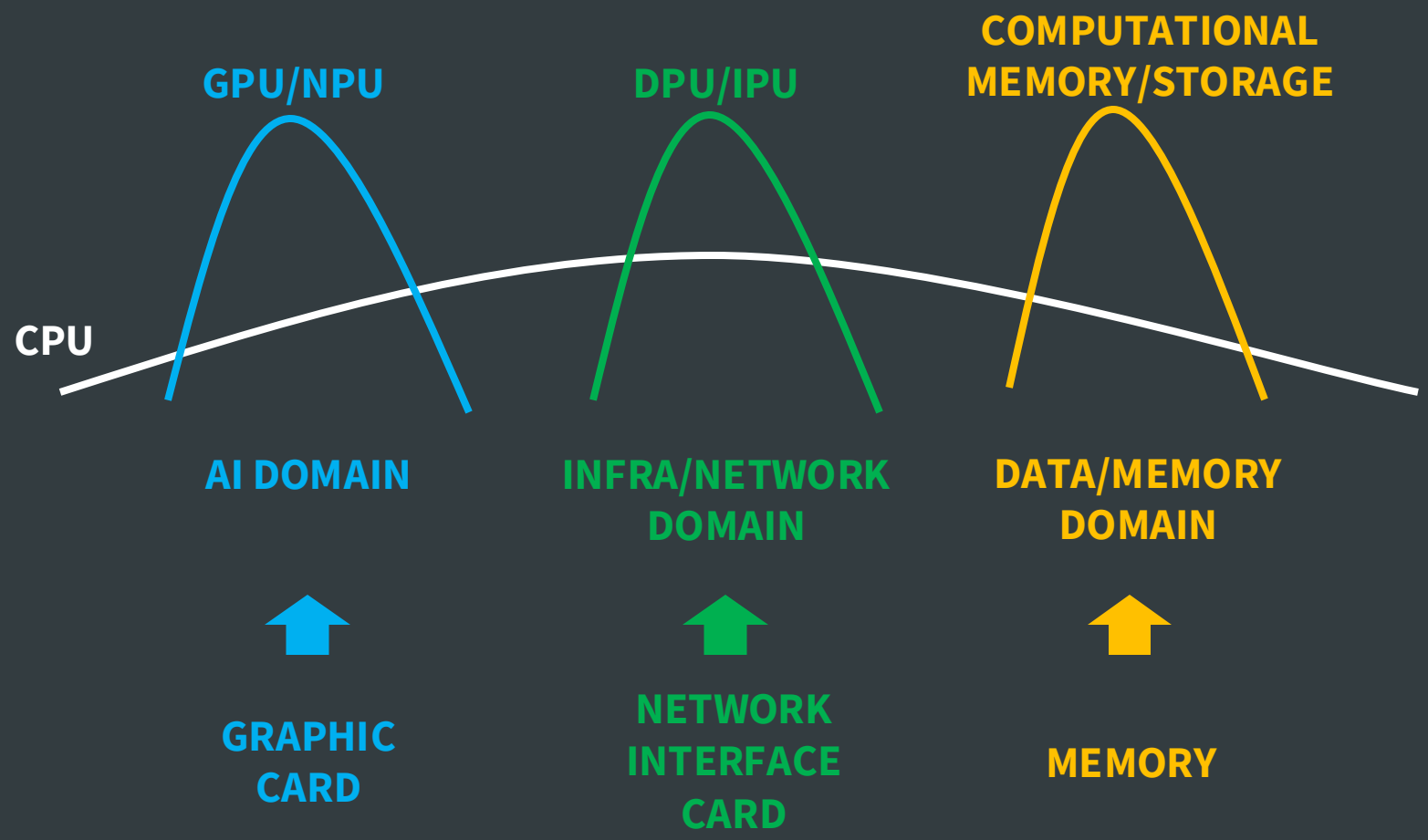
DOMAIN-SPECIFIC ARCHITECTURE



MOORE'S LAW IS DEAD
WE NEED A DRASTIC CHANGE IN COMPUTER ARCHITECTURE FROM GENERAL PURPOSE TO DOMAIN-SPECIFIC

- HENNESSY & PATTERSON

CXL COMPUTATIONAL MEORY IS FOCUSING ON DATA DOMAIN-SPECIFIC ARCHITECTURE



DATA DOMAIN-SPECIFIC

DATA DOMAIN PROBLEMS

WORKLOAD CHARACTERISTICS

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

- Vector Databases For Gen AI
- Scale-out Databases
- Graph Databases
- DNA Analysis

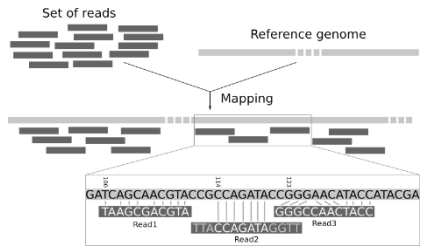
COMPUTING/MEMORY SWEET SPOTS

PROBLEM DOMAIN CLASSIFICATION

Operational Intensity Per Memory Access

Low → *High*

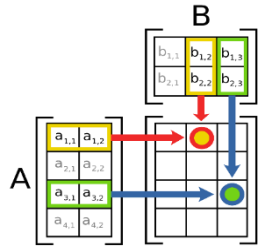
DATA-RELATED DOMAIN



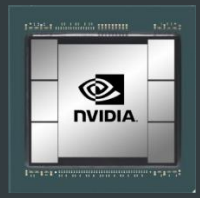
GENERAL

NO INTEREST

AI-RELATED DOMAIN

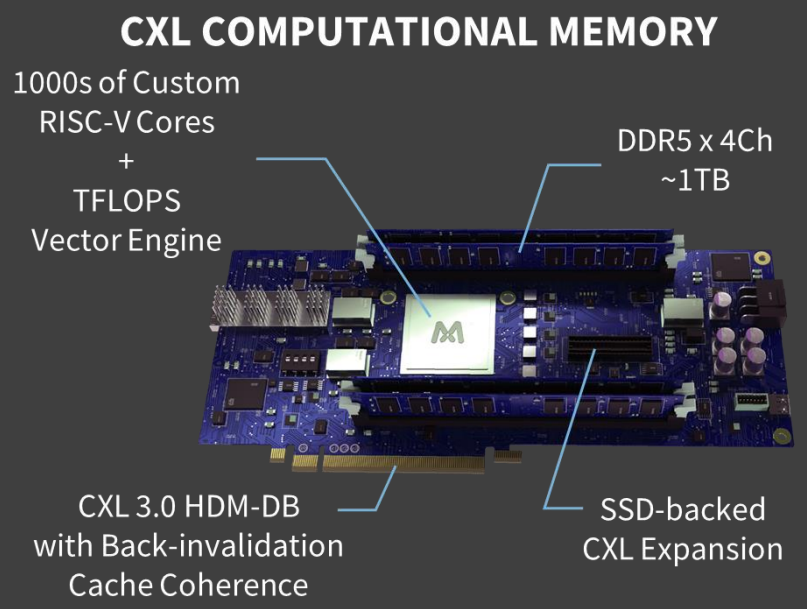


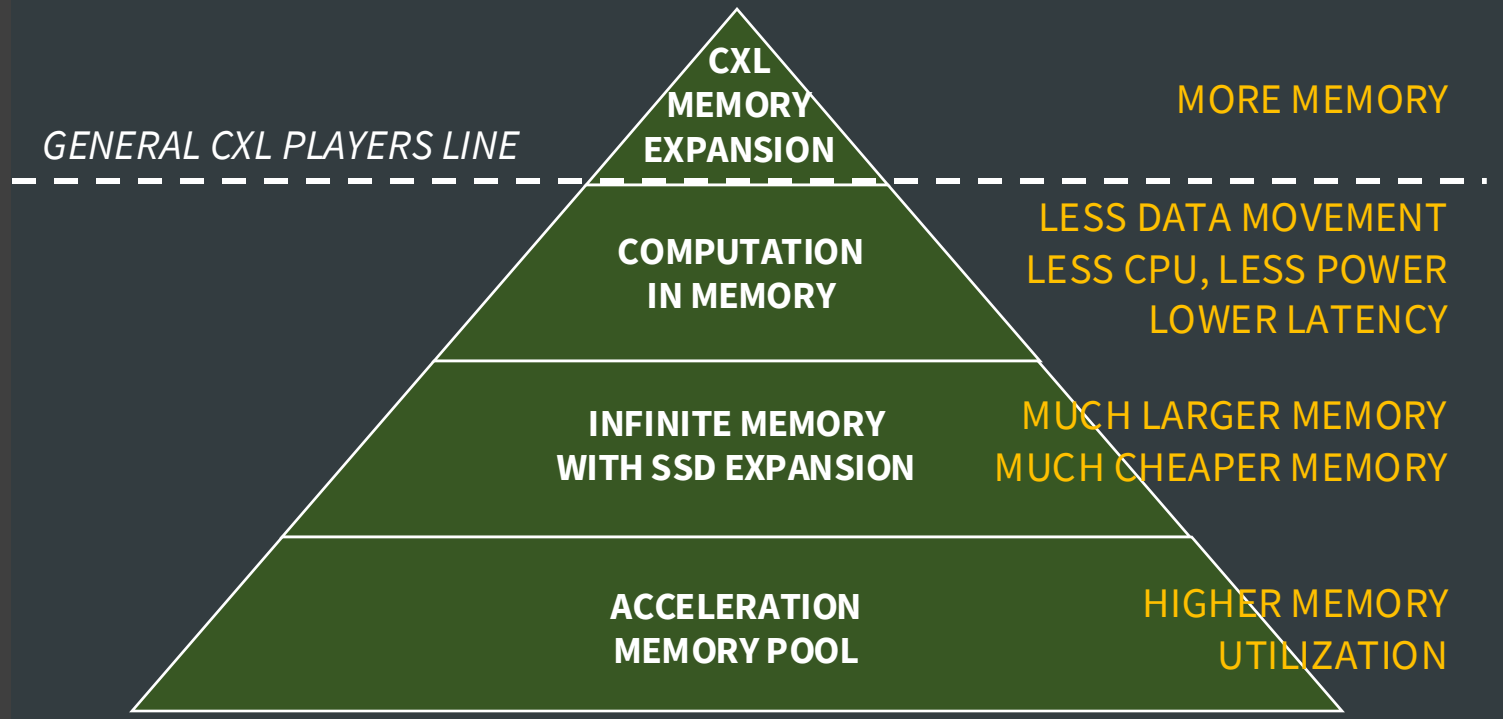
Operational Diversity Per Memory Access



CXL COMPUTATIONAL MEMORY

BEYOND JUST ANOTHER CXL MEMORY EXPANDER



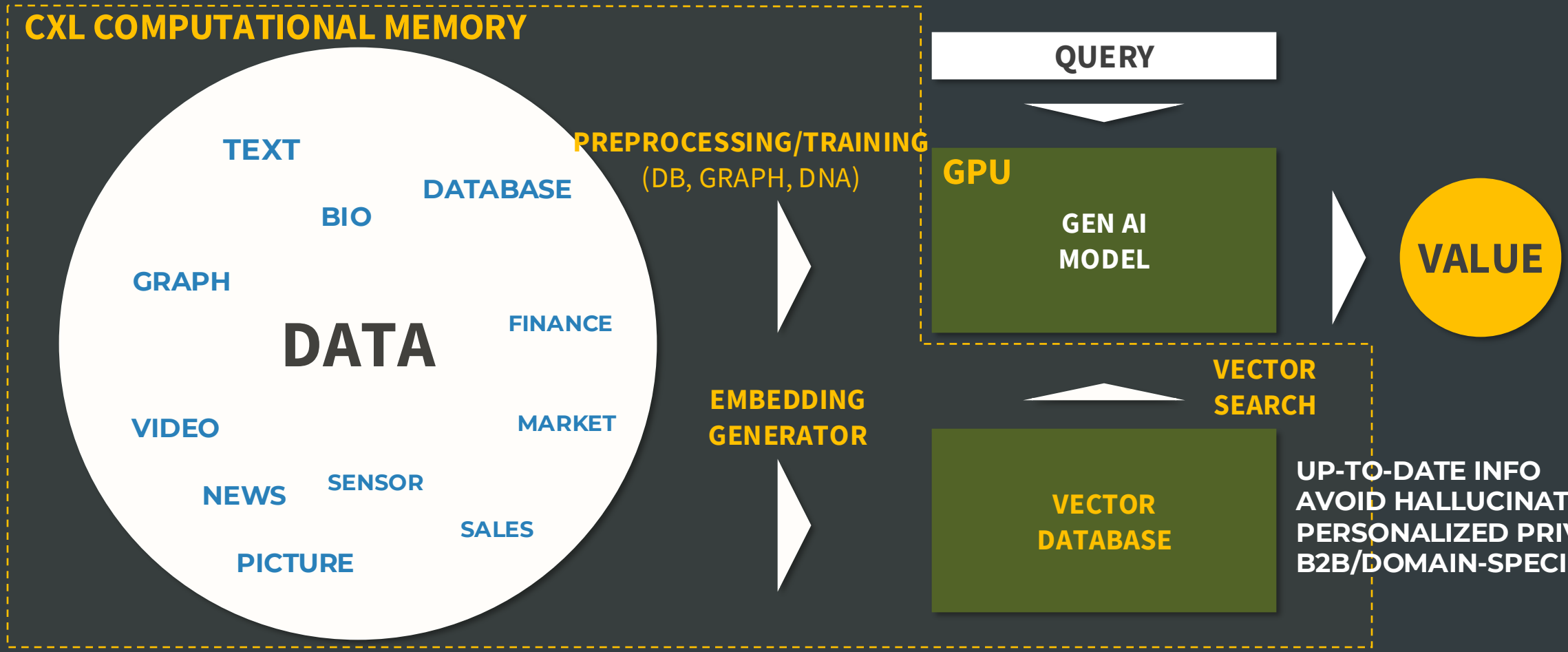


MUCH LARGER, CHEAPER, AND INTELLIGENT MEMORY SOLUTIONS FOR DRAMATICALLY REDUCING TCO IN DATA CENTERS

AI DATA PIPELINE

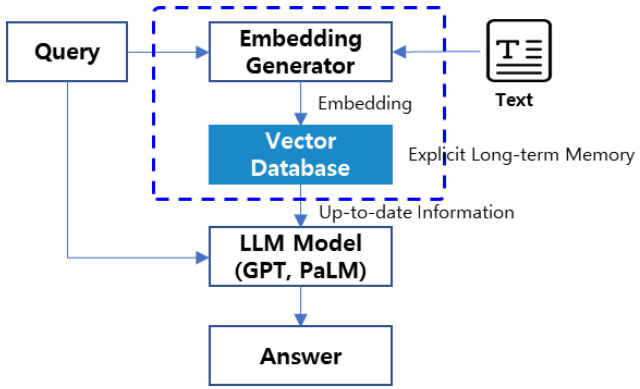
AI IS THE PROCESS OF FINDING VALUE IN DATA

CXL COMPUTATIONAL MEMORY CAN ACCELERATE MOST AI DATA PIPELINES FOR DATA-DRIVEN AI



COMPUTATIONAL MEMORY APPLICATIONS

LLM VECTOR DATABASES

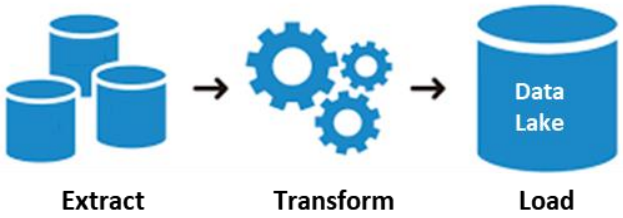


Recent LLMs utilize **vector databases** to retrieve updated information after training.

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.

SCALE-OUT DATABASES

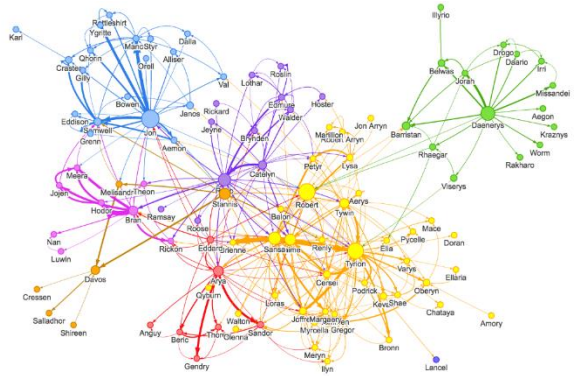


A large volume of data needs to be **processed to create value** from it even before AI training/interference.

Scale-out database clusters like Spark, Databricks, Snowflake 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**.

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.

THANK YOU

jin.kim@metisx.com

http://metisx.com

https://www.linkedin.com/company/metisx/

Visit MetisX Booth # 734