Introduction to Memory-centric AI Rack System (MARS)

[Executive Sessions]

Hoshik Kim

Senior Vice President and Fellow

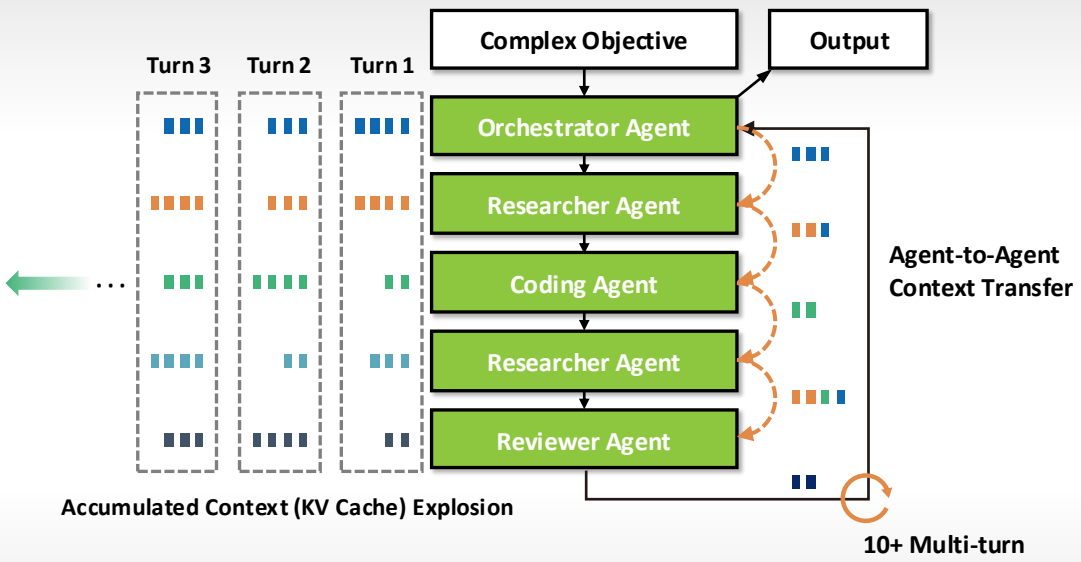
Memory Systems Research

SK hynix, Inc.

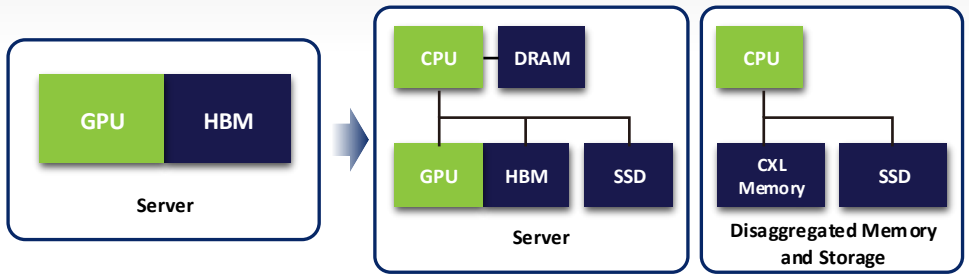


Memory Wall Challenge

- Memory demands in AI continues to grow rapidly, while scaling GPU memory faces physical and economic constraints
- Unlocking the next level of AI performance requires rethinking how data is stored, accessed, and shared
- Compute systems must be re-architected around disaggregated memory/SSD system



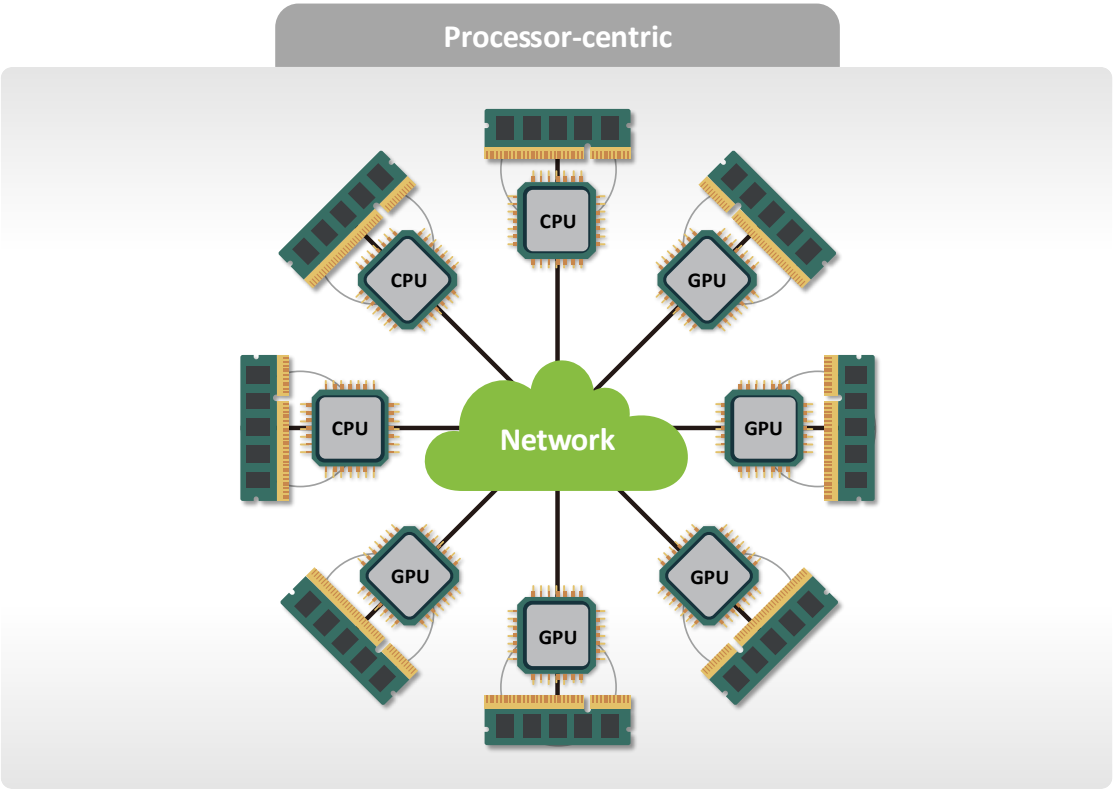
AI Researcher Agent Workflow → Exponential KV cache Growth

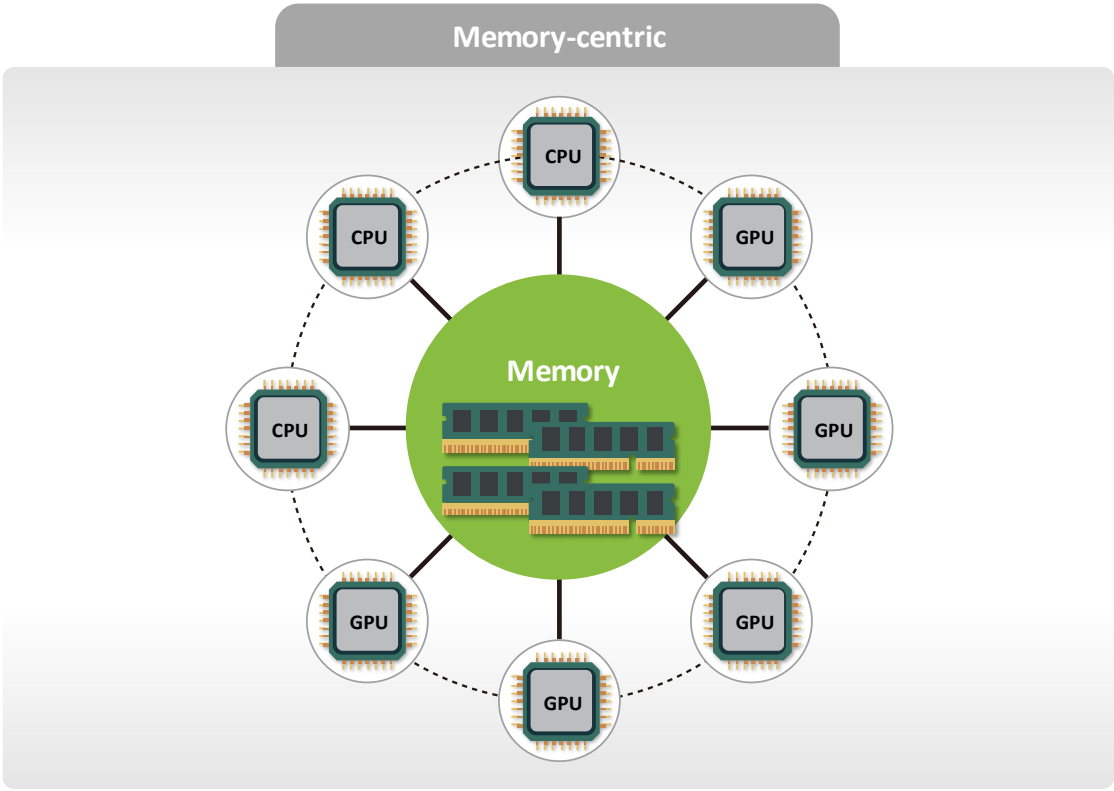


- Model Size Growing
- Context Length Growing
- Multi-Turn Conversation
- More Concurrent Users and Sessions

A Paradigm Shift: From Processor-centric to Memory-centric

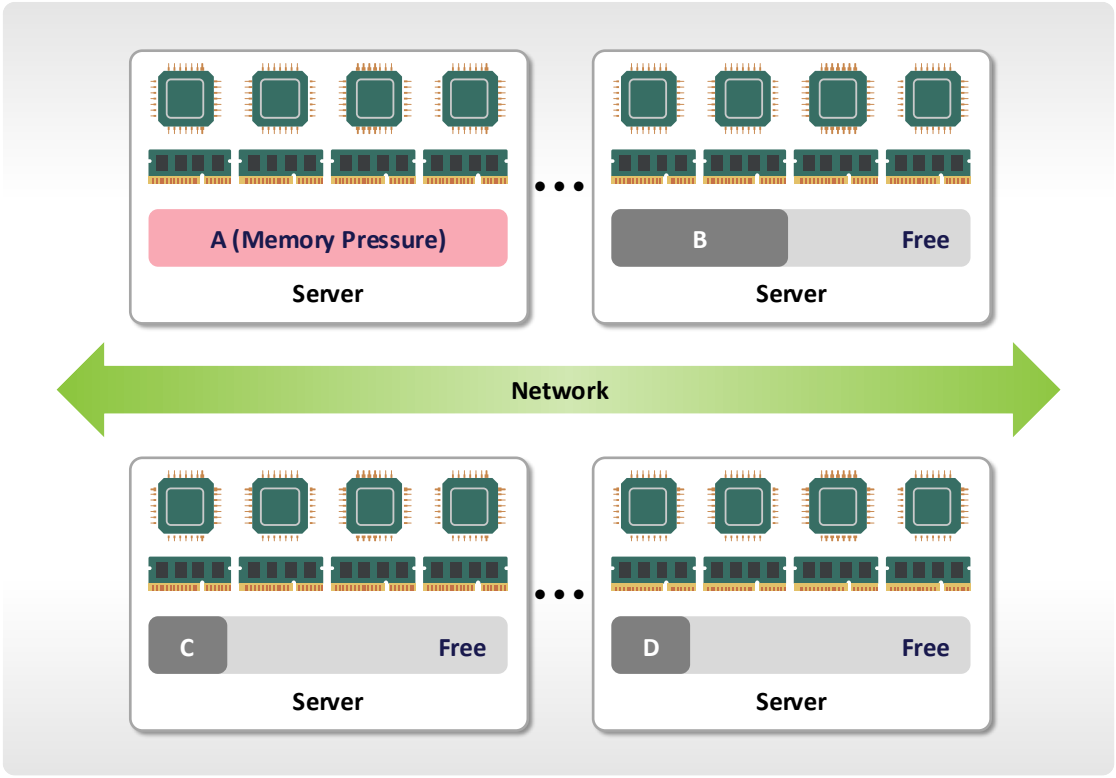
- Processor-centric: CPUs/GPUs perform computation while memory acts merely as a subordinate support
- Memory-centric: A large, shared memory serves multiple compute units, enabling more flexible access to data





Problems in Processor-centric Architecture

- Idle usage on one server cannot easily relieve pressure on another

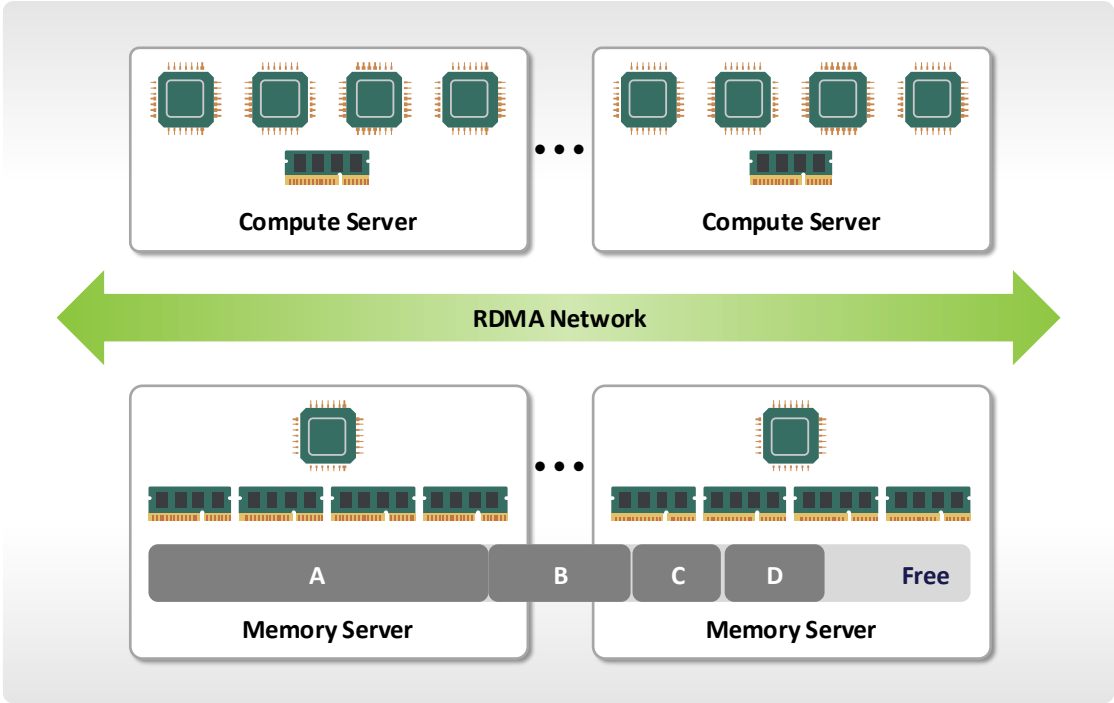


- **Resource Imbalance**
  - Memory exhausted in some nodes, idle in others
- **Rigid Scalability**
  - Must buy entire new server node to add memory
- **High TCO**
  - GPU over-provisioning for memory access drives up costs
- **Data Silos & Latency**
  - Slow network transfers and redundant copying between nodes

- Enough memory in the cluster. Not enough where it is needed.

Breaking Server Limits with Disaggregated Memory

- Memory disaggregation decouples memory from traditional CPU/GPU server constraints, creating a dedicated, massive "Memory Pool.“

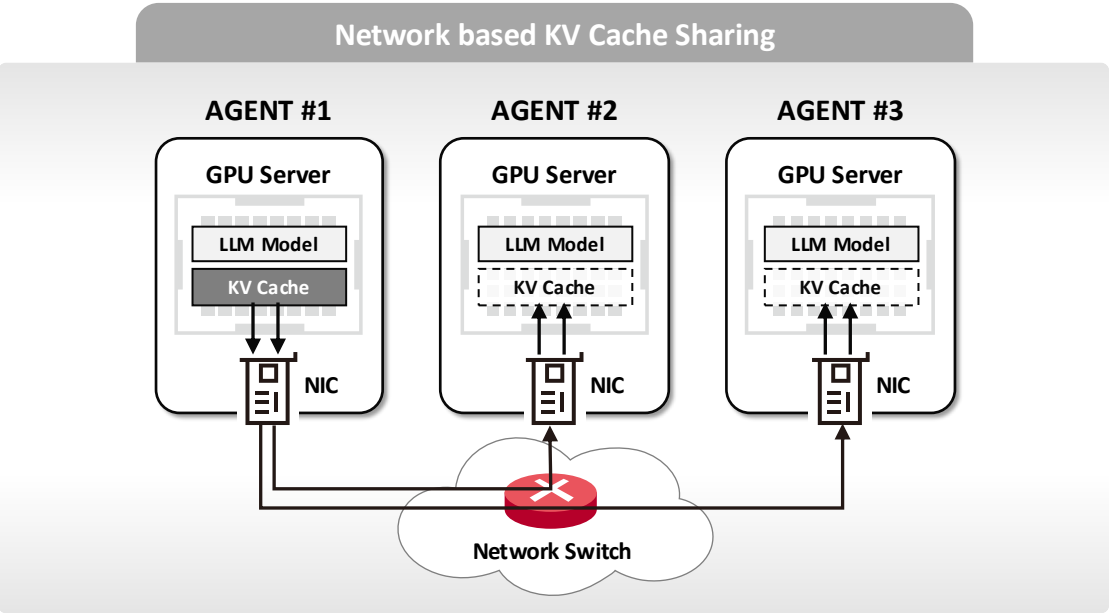


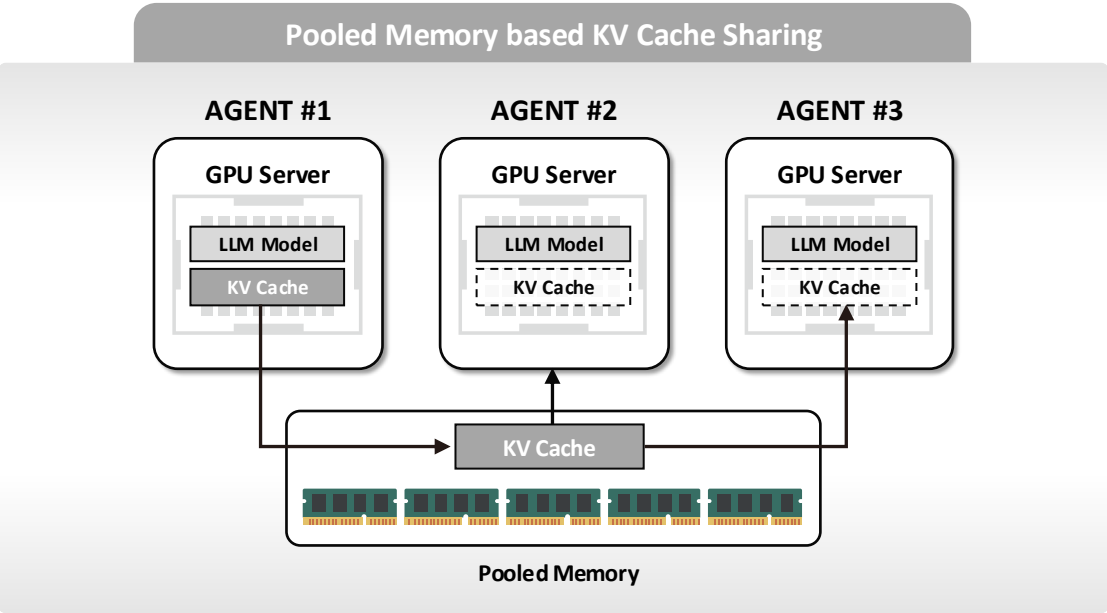
- **Higher Memory Utilization**
  - Use pooled capacity to balance demand across servers
- **Independent Scaling**
  - Add memory without adding CPU/GPUs
- **Flexible Capacity Allocation**
  - Assign and reclaim capacity as demand changes
- **Potential TCO Savings**
  - Reduce excess provisioning and unnecessary server purchases

- From per-server headroom to pool-level efficiency.

High-Speed Data Sharing via Memory Pools

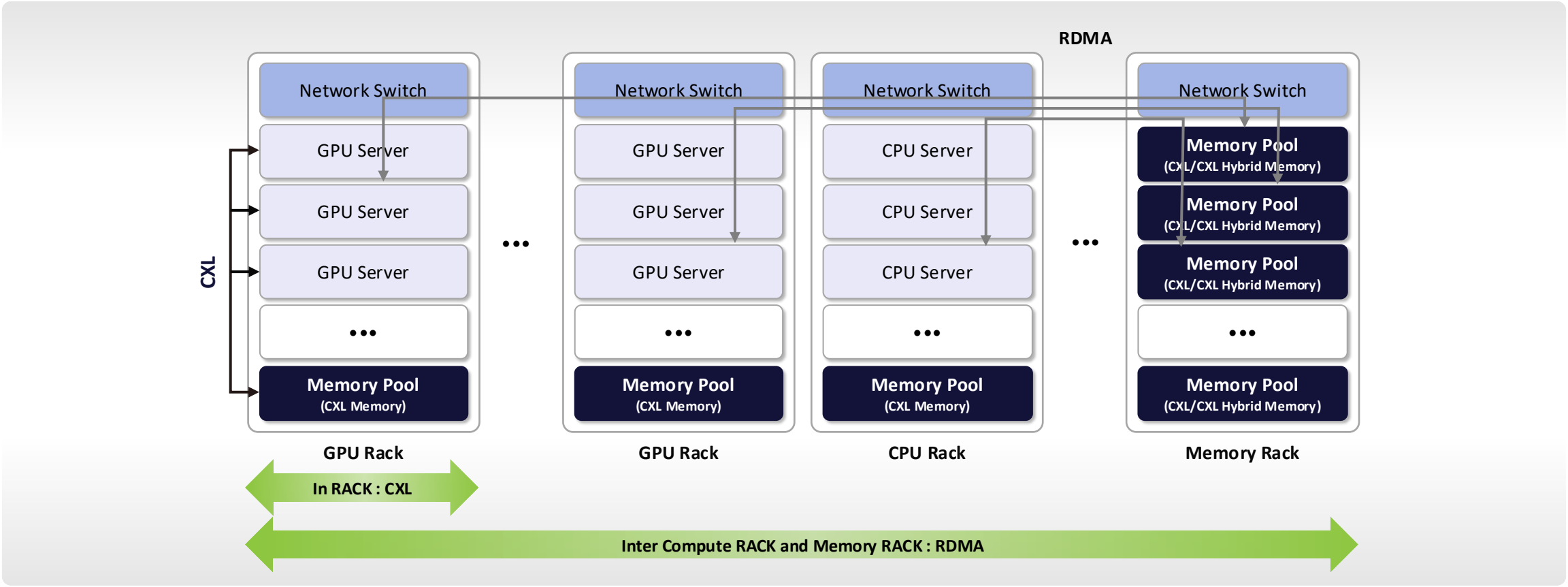
- Multiple AI accelerators (GPU/NPU) directly share large LLM models/KV caches or pipeline results within the pool
  - Reduce communication overhead: Replace explicit server-to-server data transfers with memory access over a CXL fabric
  - Streamline AI pipelines: One accelerator writes its output to shared memory, the next accesses that data without an additional application-level transfer





Multi-rack Disaggregated Memory System

- Memory-centric AI Rack System (MARS): A disaggregated rack-scale memory system that coordinates compute and memory as a shared infrastructure



MARS Software Stack

- We are co-developing a variety of software solutions with ecosystem partners to enable MARS systems

dThread Los Alamos NATIONAL LABORATORY Carnegie Mellon University	**MPI** Pacific Northwest NATIONAL LABORATORY UCMERCED	**Inter-Node Lock** TEXAS The University of Texas at Austin ILLINOIS	**Memory Copy API** (GPU-Mem Pool) VIRGINIA TECH	**Memory Allocator** VIRGINIA TECH	**RPC** UC San Diego
CCL UCMERCED Georgia Tech	**Streaming** IOWA	**NATS** Stanford	**In-Memory Store** (KV, FS, Object Intef.) IOWA STATE UNIVERSITY WISCONSIN UNIVERSITY OF WISCONSIN-MADISON	**GNN Storage** VIRGINIA TECH PennState UC Irvine	**GNN Tool** TEXAS STATE UNIVERSITY
Metadata Server WILLIAM & MARY CHARTERED 1693	**KV cache Store** VIRGINIA TECH	**KV cache Tiering** UC San Diego University of Colorado Boulder	**Graph RAG** Georgia Tech Stanford		

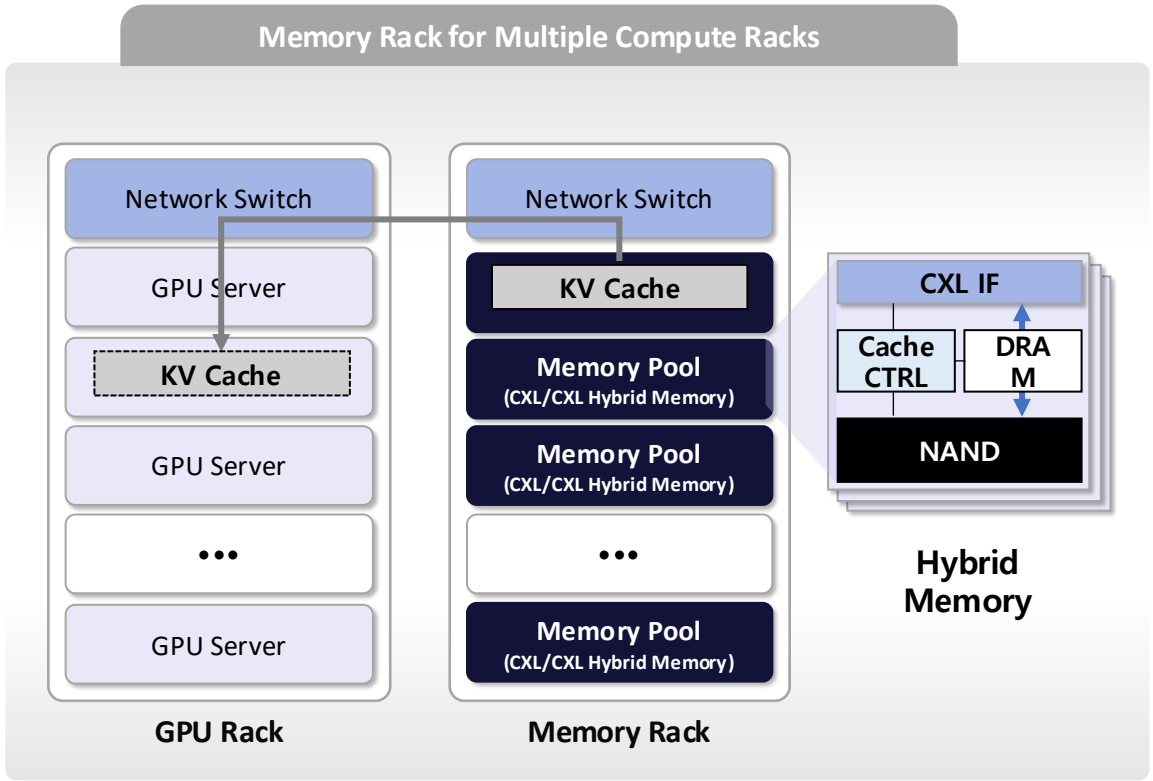
Breaking AI's 3 Bottlenecks: Capacity, Latency, and Complexity

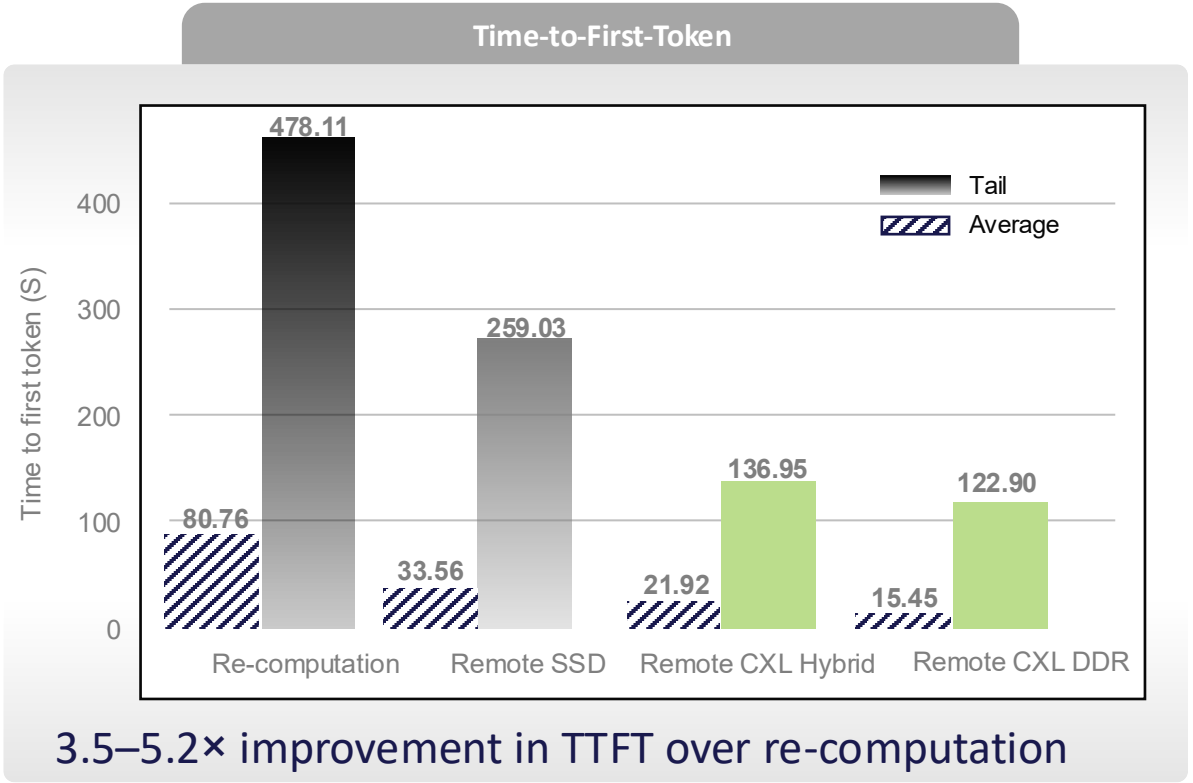
- Memory-centric AI rack system structurally resolves hardware limitation facing next-generation AI inferencing services

Capacity	**Extend available memory beyond GPU memory through a shared memory pool**
Latency	**Boosts inter-node data streaming, enabling real-time interaction**
Data Management Complexity	**Merges partitioned data into a memory pool for low latency data retrieval**

Multi-Turn AI Inference and KV Cache Acceleration

- Multi-turn agentic AI services cause KV cache to explode, exhausting GPU memory and stalling systems
- Fully manage KV cache in a memory pool can increase system throughput and provide reliable Time-to-First-Token (TTFT)

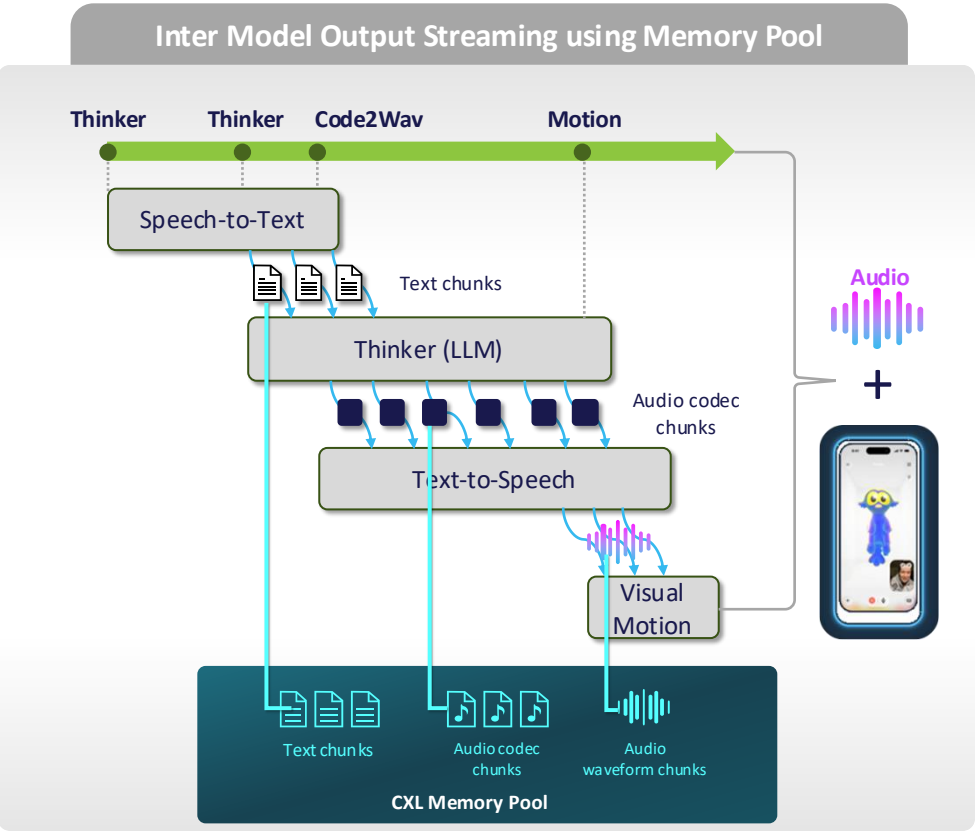


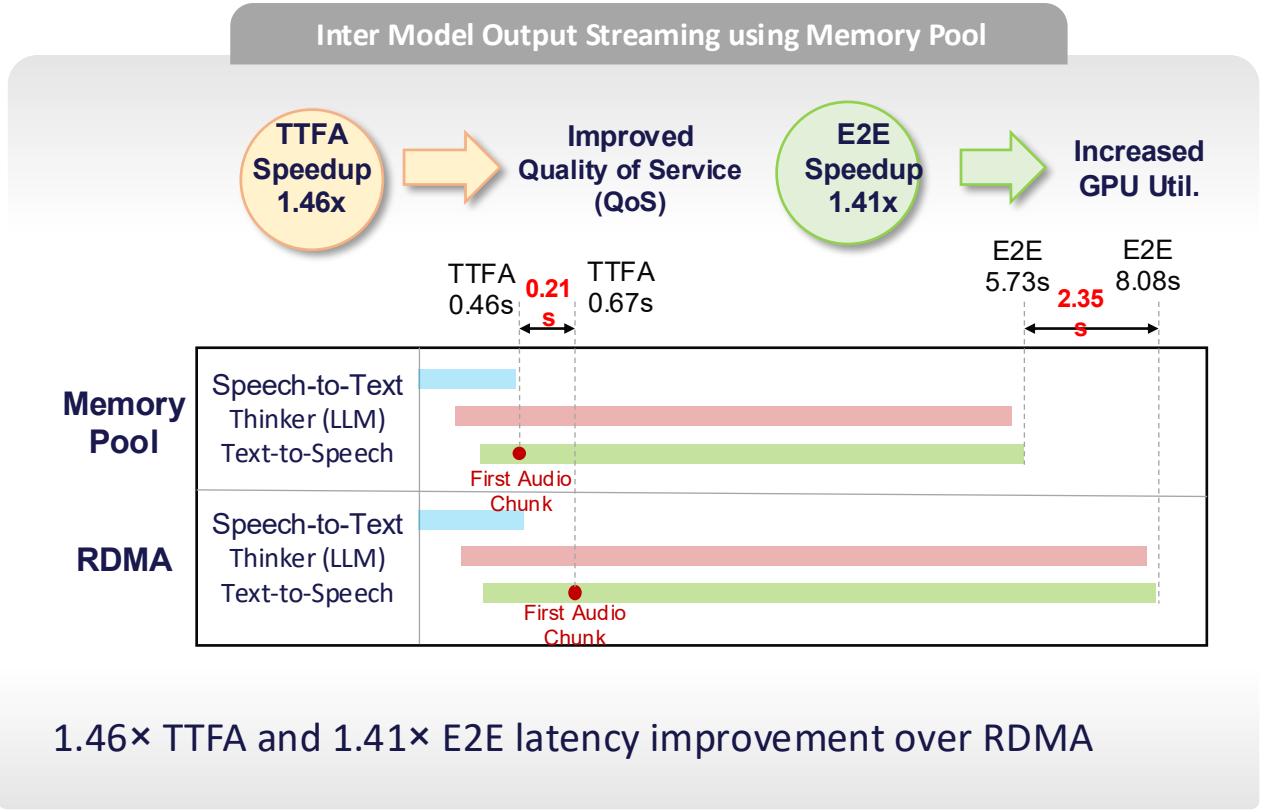


3.5–5.2× improvement in TTFT over re-computation

Low-Latency Voice Interaction for Personal AI Agent Service

- Traditional voice pipelines (Audio Encoder → LLM → Text To Speech (TTS)) suffer from latency due to data copying and network overhead between models, causing unnatural pauses
- All AI models in the voice pipeline operate on the same Shared Memory region within the Disaggregated Memory Pool

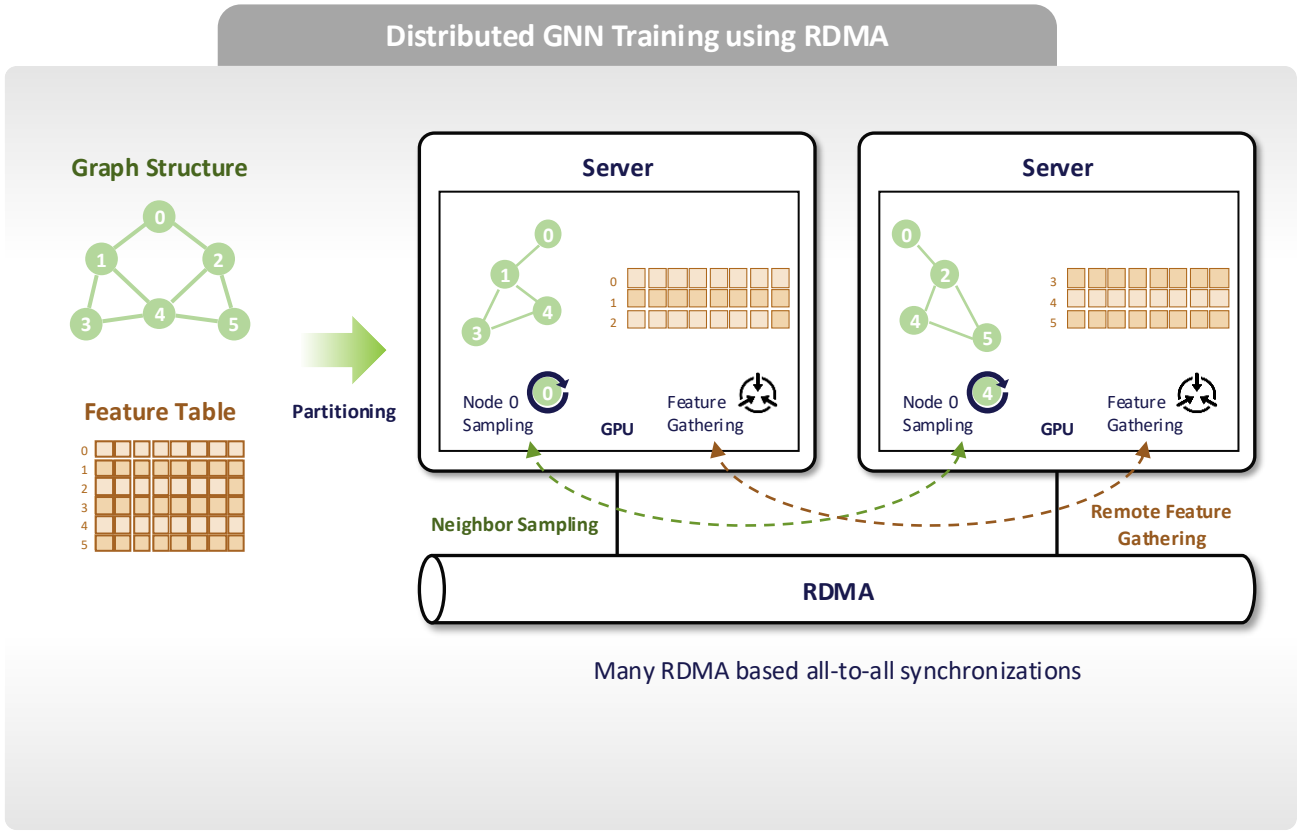


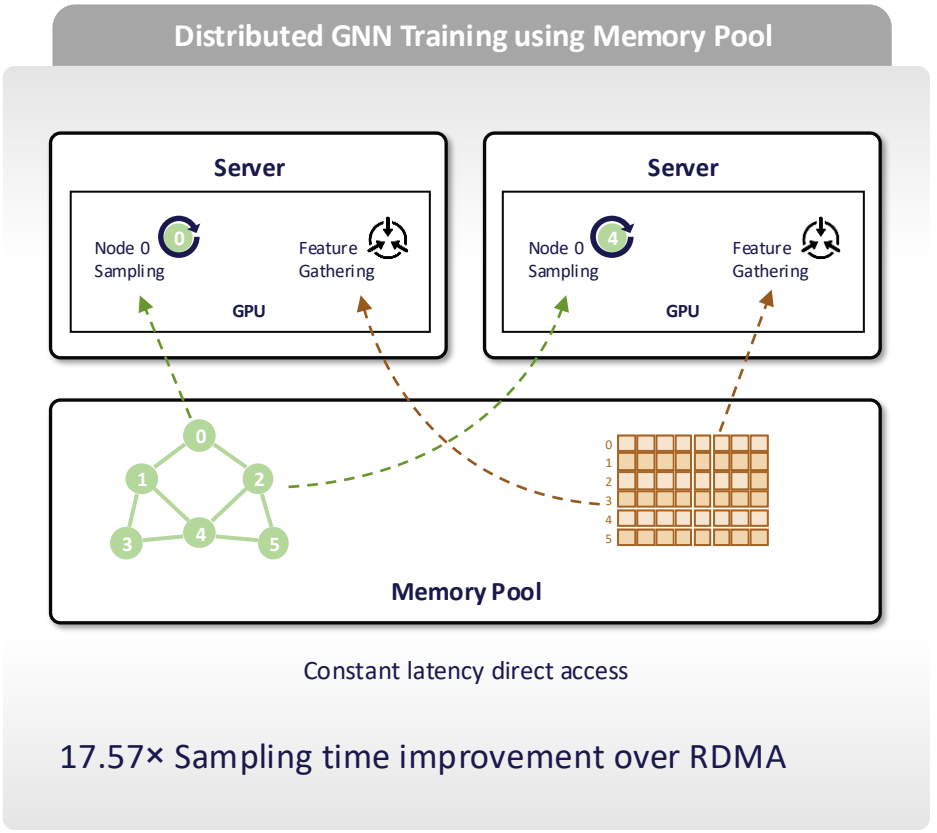


1.46× TTFA and 1.41× E2E latency improvement over RDMA

Unified Graph Sampling for Graph Neural Network Training

- A trillion-scale graph data are partitioned across multiple servers, leading to frequent communication during sampling
- Offloading graph/feature data to memory pool can eliminate sampling synchronization and improve data read latency





17.57× Sampling time improvement over RDMA

Toward a New Foundation for AI Infrastructure

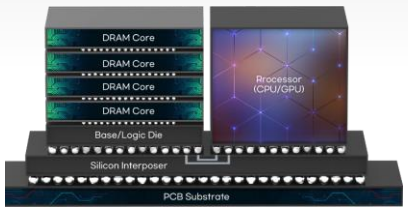
- **Memory-Centric AI Rack Systems (MARS)** structurally resolve hardware limitations facing next-gen AI services
  - **Capacity expansion** for **multi-turn AI inference (Multi-Turn conversation, Multi-Turn Coding Agent)**
  - **Low latency** for **personal AI agent**
  - **Unified data management** for **Graph Neural Network Training**
- MARS are no longer an option but **a requirement for operating massive AI systems**
- MARS supporting **customized pilot tests** for clients and expanding the ecosystem

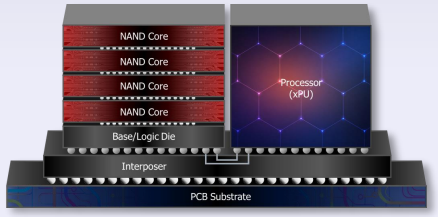
One More Thing …

HBF Extends HBM-class Bandwidth to Flash-scale Capacity

- A New AI memory layer between HBM and eSSD

HBF: Bridging the gap between HBM and eSSD







	HBM	**HBF**	eSSD
Bandwidth	>1x	**0.5x**	<0.1x
Capacity	1x	**14x**	>500x
pJ/b	1x	**3.7x**	6.9x

Energy Efficiency & Bandwidth → **Capacity**

Capacity Scaling

Extend beyond HBM capacity limits

Bandwidth at Scale

Maintain high-speed access as capacity grows

Data Movement Reduction

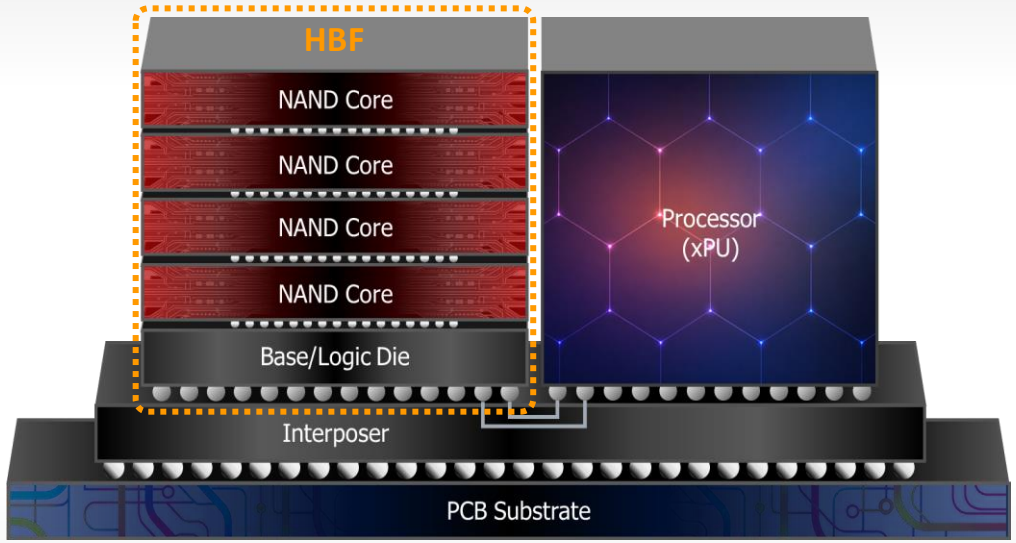
Keep more AI data closer to compute

A New AI Memory Tier Between HBM and eSSD

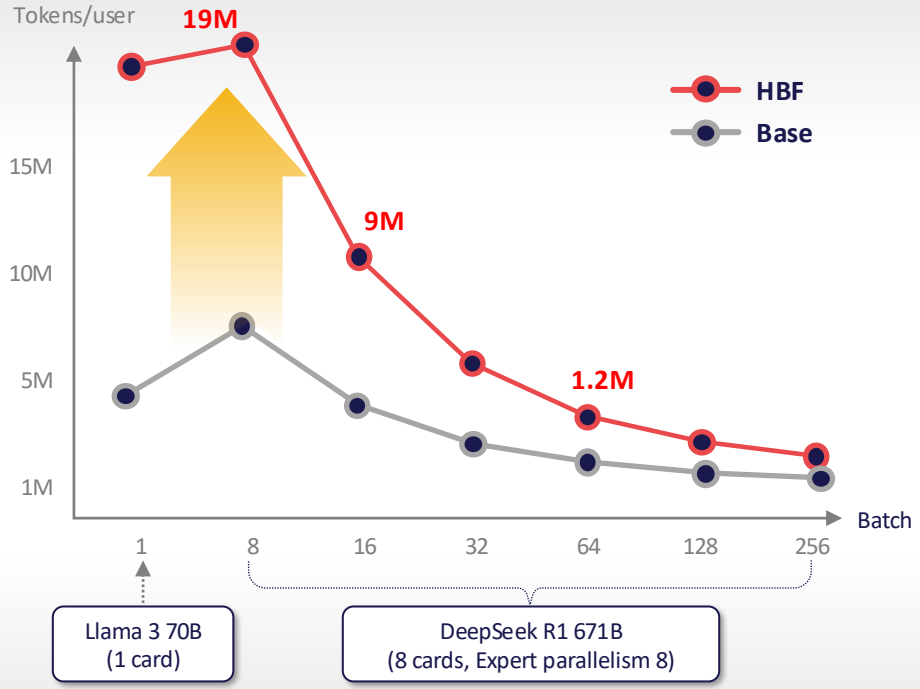
HBF Breaks the Memory Capacity Wall

- HBF provides x1/2 Bandwidth and x10 Capacity of HBM to break the Capacity Wall

1/2x Bandwidth, 10x Capacity of HBM



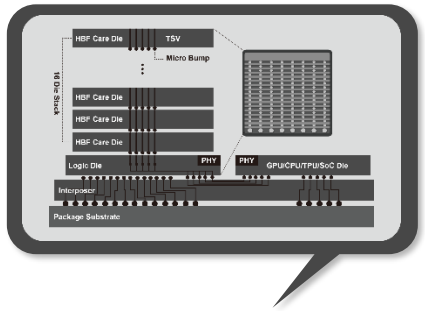
Breaking the Capacity Wall



*Base card capacity is 288GB and HBF card 3TB

Building the Standard and Ecosystem for HBF

HBF Introduction @SanDisk



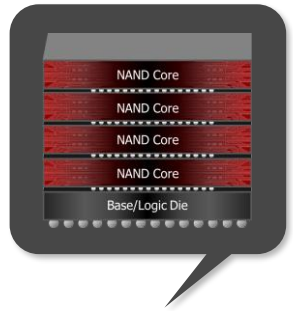
MOU

SK hynix
SANDISK

Consortium Kickoff

SK hynix
SANDISK
Google
tenstorrent
OPEN Compute Project
HIGH BANDWIDTH FLASH(HBF) HIGH-LEVEL SPECIFICATION
VERSION 1.0

HBF Cube



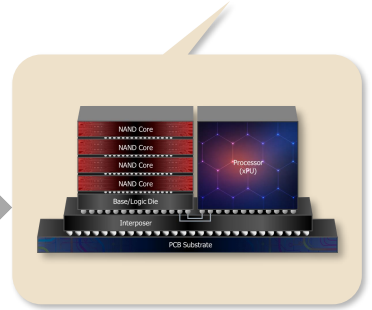
SanDisk HBF Introduction	MOU Signed	Consortium Kick-off w/ Spec α	Public Release	Spec 1.0 Release	HBF Cube Available	TBD
‘25.2	‘25.8	‘26.2	‘26.7	‘27.1 (Expected)	‘28.1Q	



v 0.7 Public Release

Now

Collaboration with partners



HBF Product

Thank You!

