MEMORY COUPLED COMPUTE: INNOVATING THE FUTURE OF HPC AND AI

Dr. Samantika Sury

Vice President and Chief Hardware Architect of HPC

SAIT (Samsung Advanced Institute of Technology), Samsung Electronics

Systems Architecture Lab

SAMSUNG

ACKNOWLEDGMENTS

- SAIT Systems Architecture Lab
  - Alan Gara, David Lombard, Rolf Riesen, Bob Wisniewski

- SAIT Computing Platform Lab
  - Wooseok Chang, Youngjun Hong, Sehwan Lee, Seungwon Lee, Seungwook Lee, Junho Song, Eunsoo Shim, Sehyun Yang

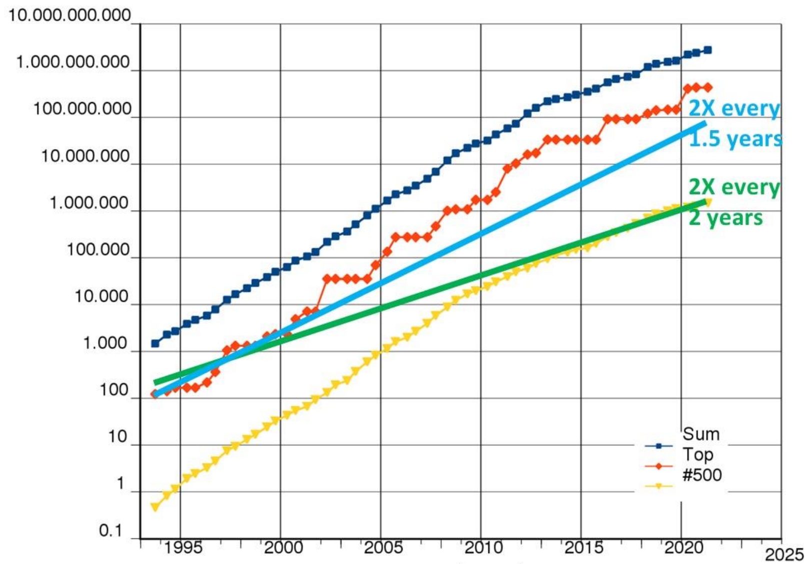
SYSTEMS ARCHITECTURE LAB

- Vision
  - To develop the most innovative technologies for future HPC and AI systems

- Strategy
  - Break through the Memory and Communication walls
  - Significantly increase the memory bytes/flop ratio with Memory Coupled Compute
  - Significantly increase network bytes/flop ratio with Supernodes
  - Drive innovations in system-level energy and cost-efficiency

DISCONTINUITIES

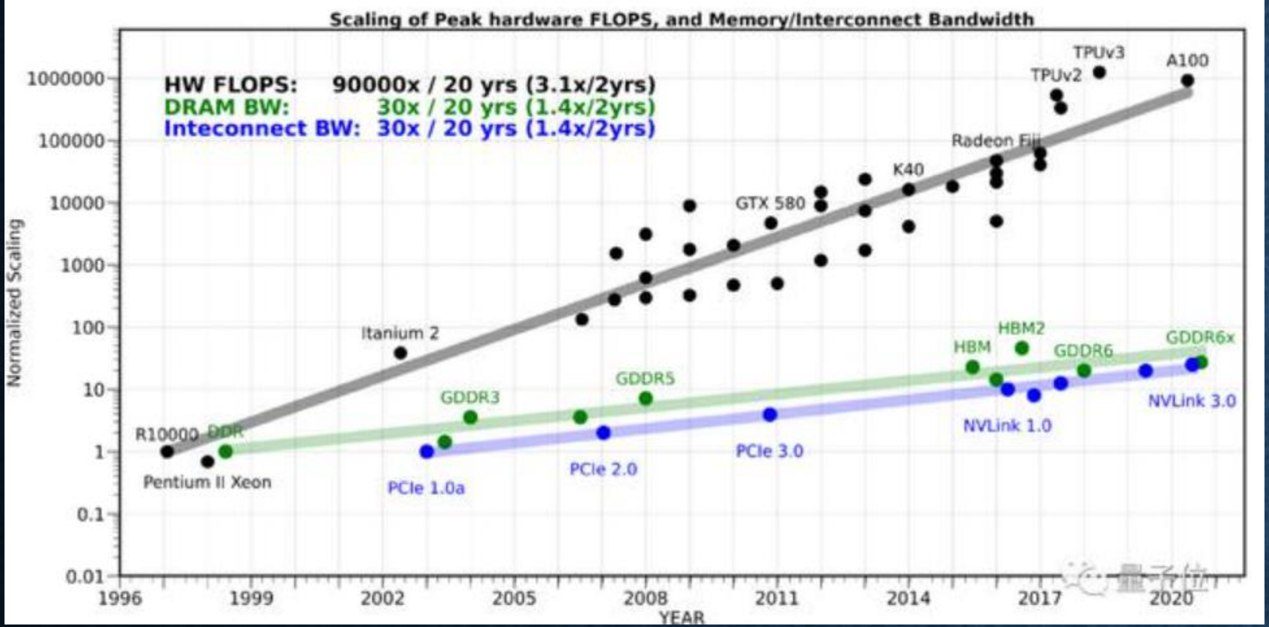
- Vectors (Cray)
- Microprocessors (Beowulf)
- Multicore, multithread (x86/ Power)
- Massive parallelism (Blue Gene)
- Heterogeneity (GPUs)
- Memory Coupled Compute
  - The next discontinuity
  - **Tight-coupling of compute, memory and communication**

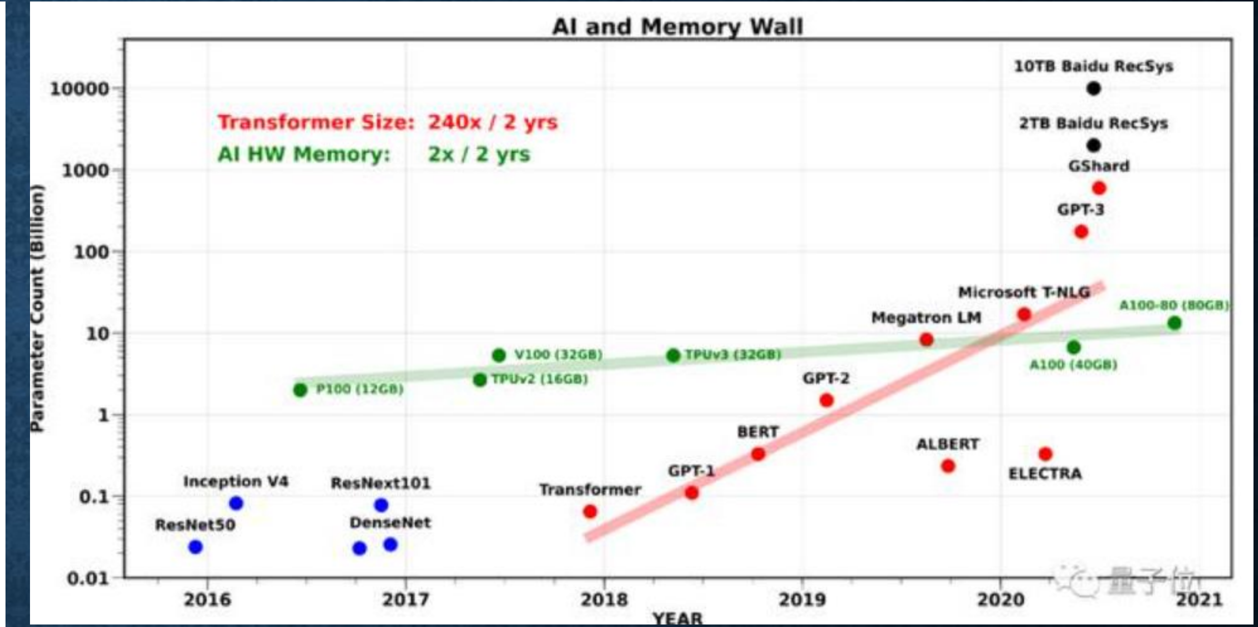


- Source: Wikipedia.org

Discontinuities are often driven by cost !

THE MEMORY AND COMMUNICATION WALL IS GETTING HIGHER





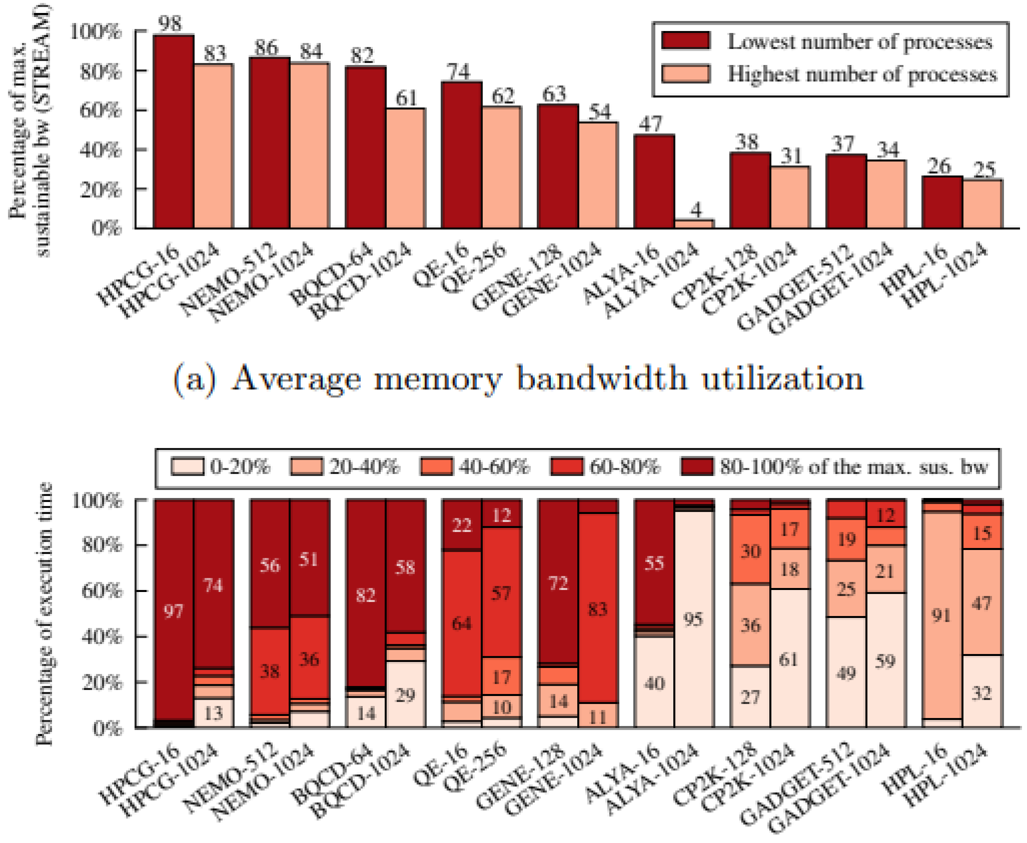
- https://daydaynews.cc/en/science/the-biggest-obstacle-to-ai-training-is-not-computing-power.html

- Modeling and simulation applications are memory bandwidth limited
- AI, and some mod/sim applications are communication bandwidth limited

SAMSUNG

MANY APPLICATIONS ARE MEMORY BOUND

- Increasing divergence between compute and memory
- Increasing number of memory-bound phases or full applications.
- Increasing memory performance → Effective way to improve mod/sim app performance

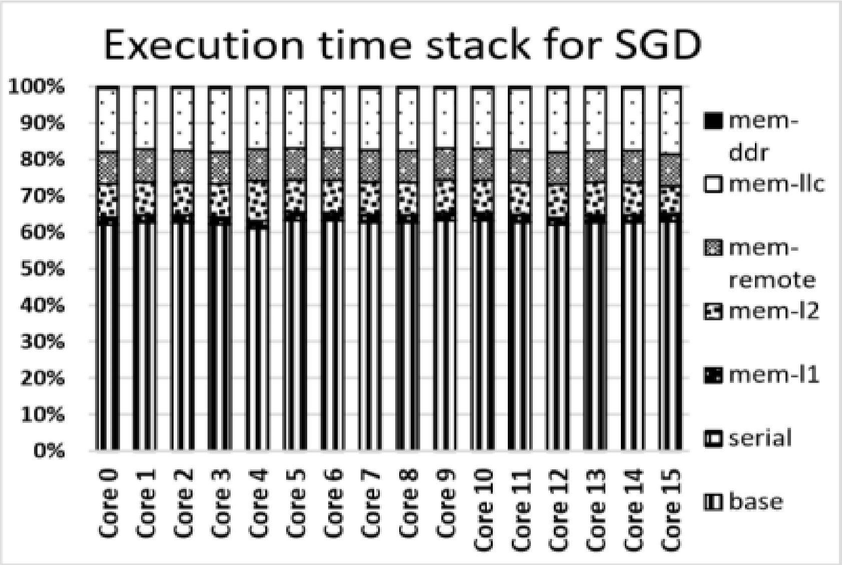


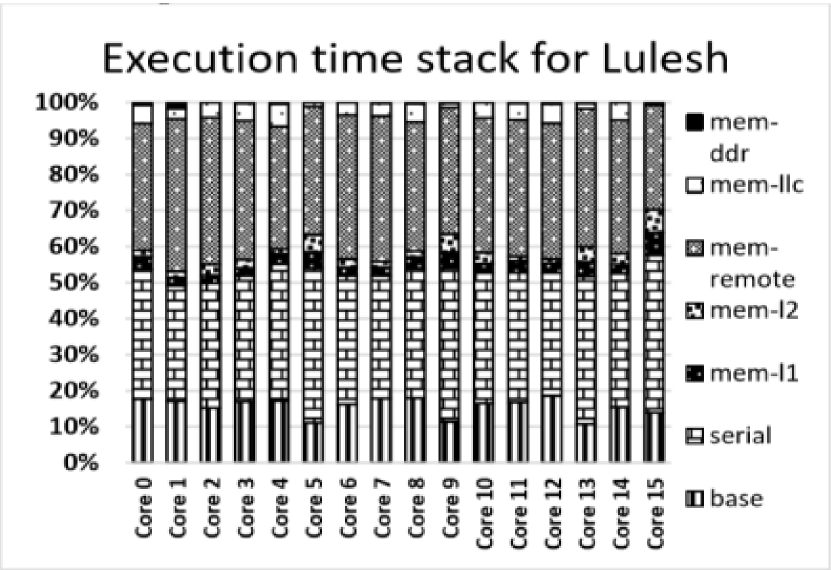
(a) Average memory bandwidth utilization

(b) Memory bandwidth utilization on burst granularity

HPC Benchmarking: Scaling Right and Looking Beyond the Average, Milan Radulović et. al., International Conference on Parallel and Distributed Computing, 2018

IMPACT OF MEMORY PERFORMANCE





Stochastic Gradient Descent used in Machine Learning Algorithms

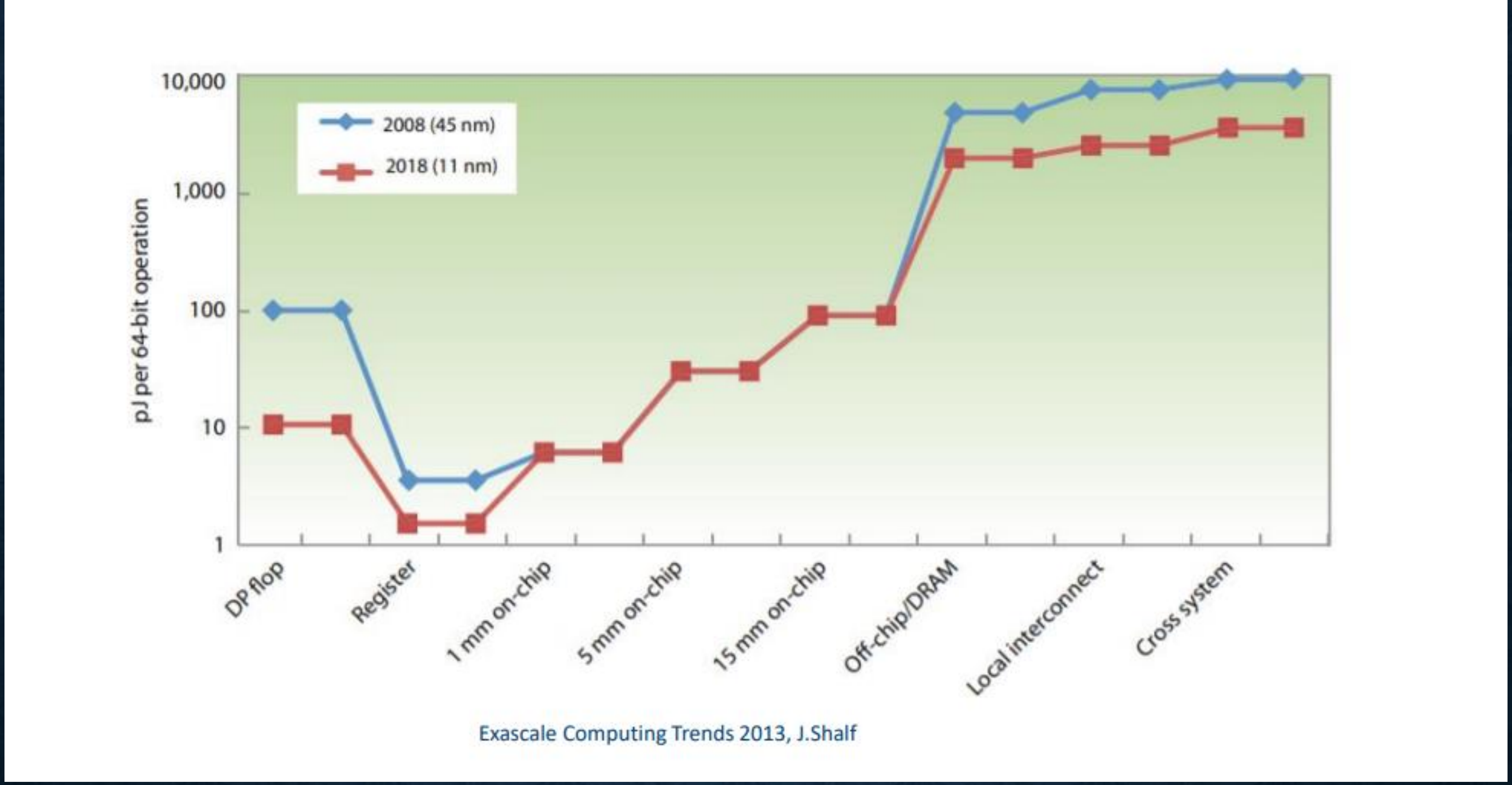
Hydrodynamics code used in Classical HPC

Why are we spending so many cycles communicating data?

IEEE AICCSA 19: CONCORD: Improving COmmuNication using COnsumeR-Count Detection Farah Fargo, Shobha Vissapragada, Samantika Sury

SAMSUNG

COMPUTE EFFICIENCY != COMMUNICATION EFFICIENCY



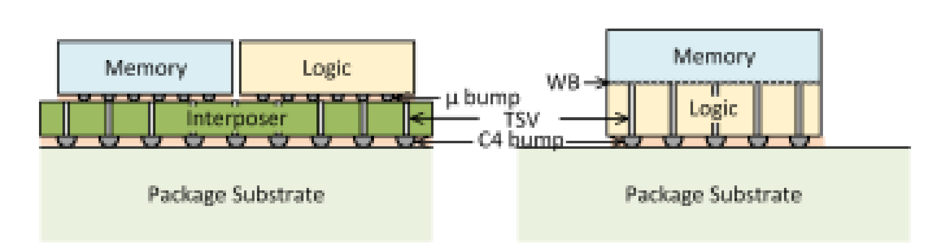
Exascale Computing Trends 2013, J.Shalf

- Exascale era → dominated by compute
- Post-exascale era → dominated by data movement

ATTACKING THE MEMORY WALL

- 2.5D (Processing near memory)
  - Current deployed technology
  - HBM co-packaged with compute
- PIM (Processing in memory)
  - Closest possible to memory
  - Current constraints limit functionality
- 3D (Memory Coupled Compute)
  - Reduces power consumption and latency
  - More efficient packaging than 2.5D

Closer coupling of compute with memory



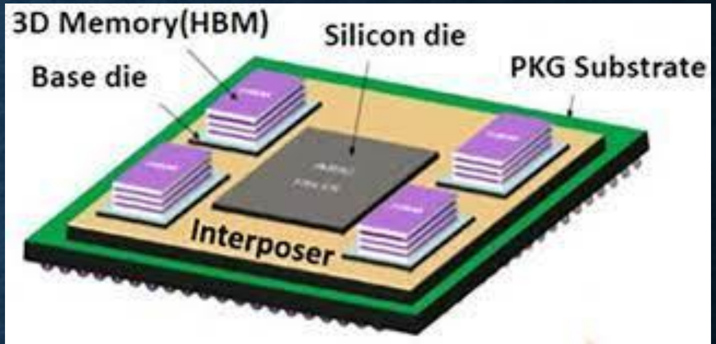
e.g. 3D systolic ML accelerators in IEEE Journal on Exploratory Solid-State Computational Devices and Circuits – June 2021

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4556747
https://people.inf.ethz.ch/omutlu/pub/ProcessingDataWhereItMakesSense_micpro19-invited.pdf

2.5D OPPORTUNITIES AND CHALLENGES

+ Significant improvement over 2D/DDR

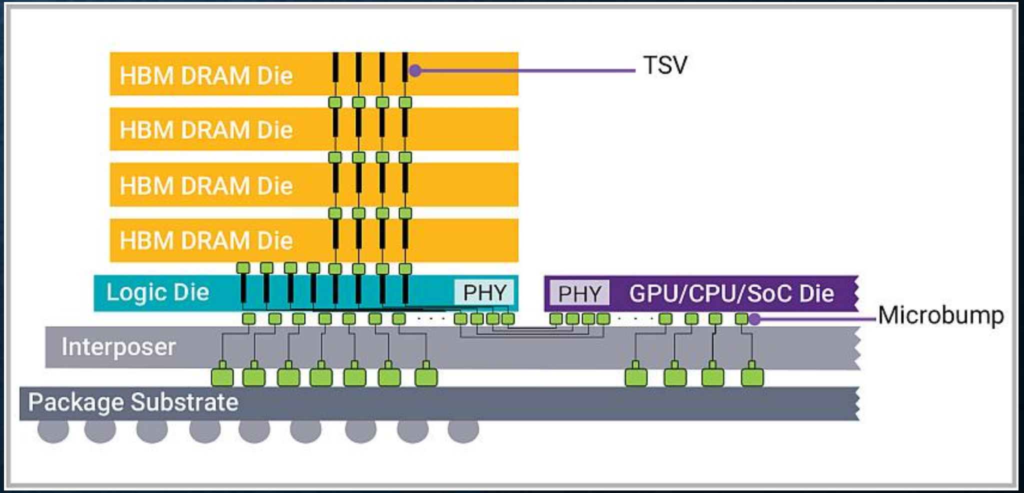
- Higher bandwidth
- Latency on par
- Flexible SOCs



- Substrate and connections can be expensive

- Requires off die logic-mem connection

- Off-die signals require more power and area
- Die crossings to get to HBM
- Data access latency often dominated by SOC size → leads to tradeoff
- BW limited by pin limitations and PCB wires



SAMSUNG

PIM OPPORTUNITIES AND CHALLENGES

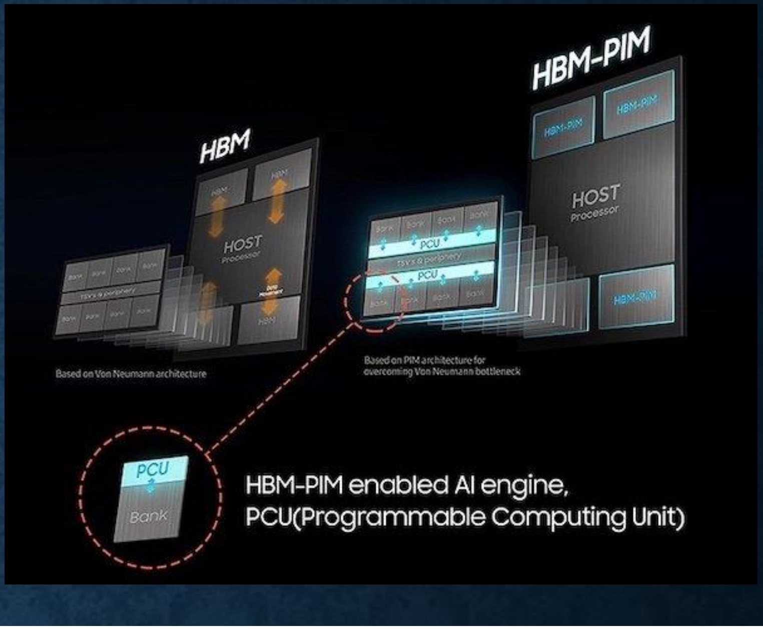
- HBM bandwidth is not enough for many ML workloads
  - BLAS-1 (AXPY) and BLAS-2 (GEMV) get memory bound

+ Most energy efficient compute
  - ALUs and mem on same die → minimal data movement
  - DRAM-optimized AI engine inside memory bank
  - **Samsung's HBM-PIM → >1 TFLOPS of embedded computing power**

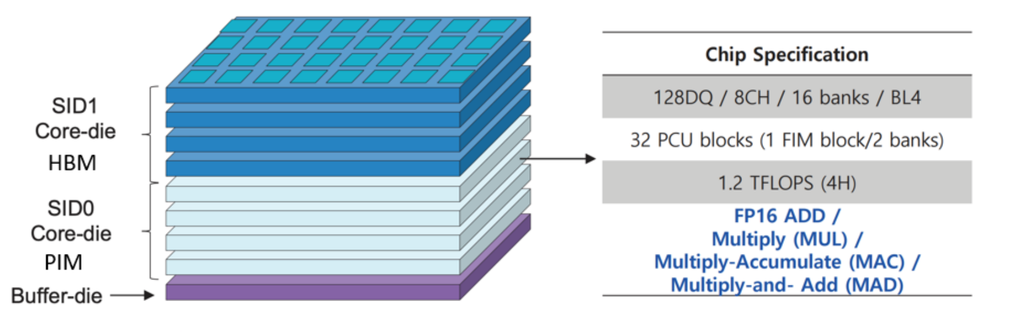
+ Rapid integration into existing systems
  - 16-wide SIMD engine
  - Ease of software (Native and direct execution)

- The type of operations are constrained

- ALUs reduce available memory capacity or increase die area



HBM-PIM enabled AI engine,
PCU(Programmable Computing Unit)



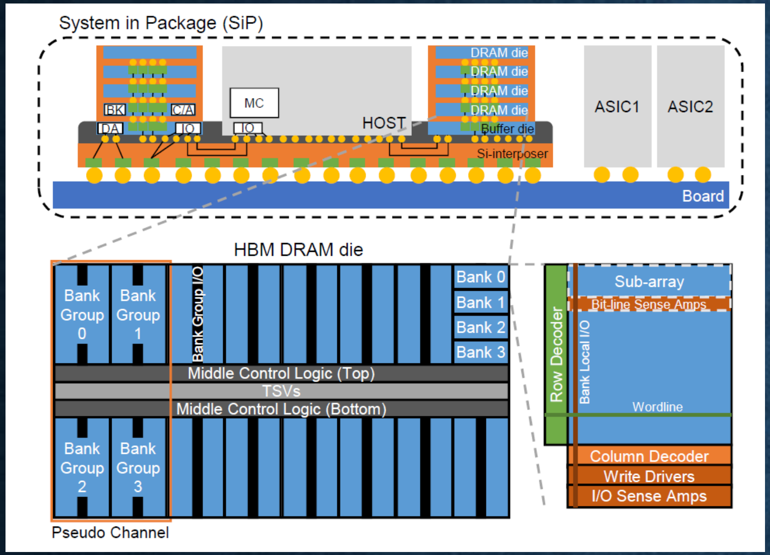
Chip Specification
128DQ / 8CH / 16 banks / BL4
32 PCU blocks (1 FIM block/2 banks)
1.2 TFLOPS (4H)
FP16 ADD / Multiply (MUL) / Multiply-Accumulate (MAC) / Multiply-and- Add (MAD)

AXDIMM

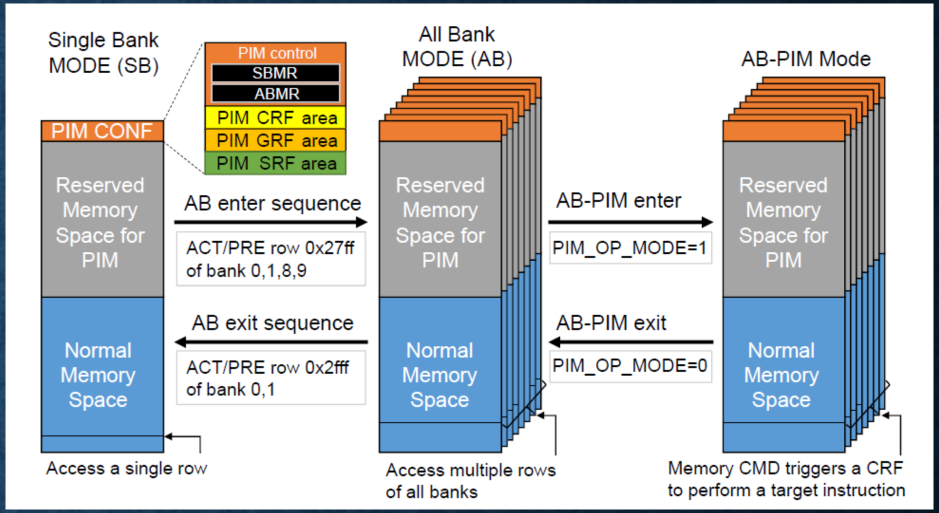
Kwon et al., A 20nm 6GB Function-In-Memory DRAM, Based on HBM2 with a 1.2TFLOPS Programmable Computing Unit Using Bank-Level Parallelism, for Machine Learning Applications, ISSCC 2021

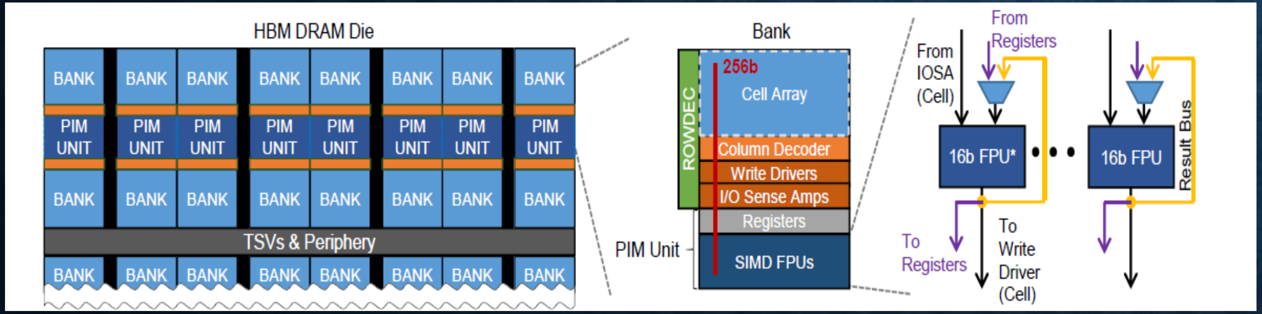
REAL WORLD PIM-HBM ORGANIZATION

Support DRAM and PIM-HBM mode for versatility



Classic HBM die organization





Bank coupled with PIM and simple PIM datapath

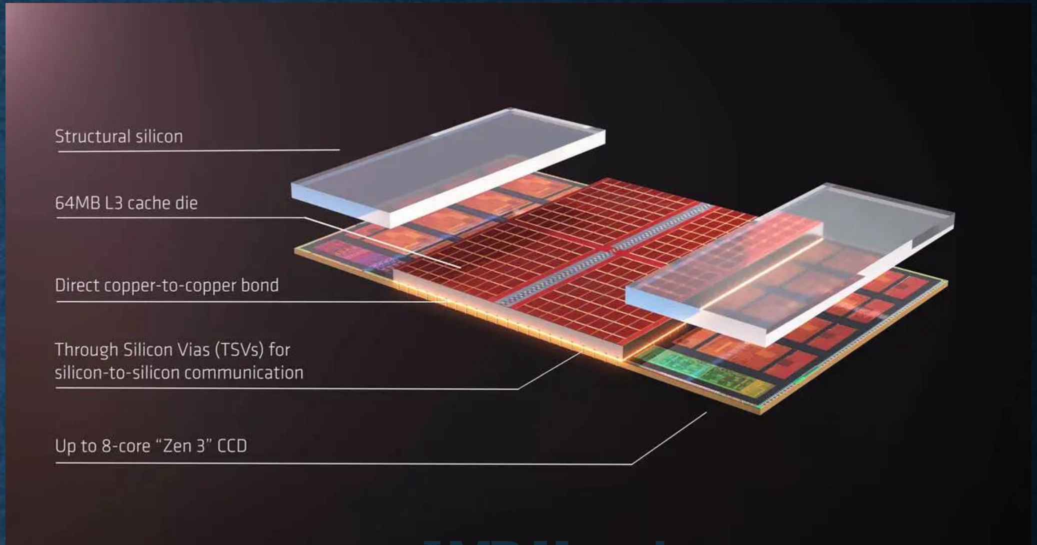
- Exploit bank-level parallelism
- PIM-HBM operation modes: single bank (SB), all-bank (AB) , all-bank-PIM
- AB mode: PIM-HBM 8x higher bw
- AB-PIM : AB+ PIM instruction
- PIM supports RISC-like 32-bit instructions → 9 total instructions
- Support for TensorFLow and Pytorch
- AXDIMM and LPDDR5-PIM extensions

Lee et al., Hardware Architecture and Software Stack for PIM Based on Commercial DRAM Technology, ISCA 2021
Aquabolt-XL HBM- PIM Hotchips 33
https://news.samsung.com/global/samsung-brings-in-memory-processing-power-to-wider-range-of-applications

12

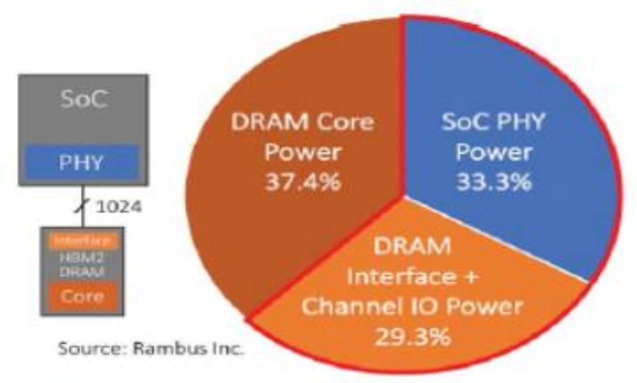
MEMORY COUPLED COMPUTE

- True 3D stacking
- Performance and Power Efficiency
  - Less distance to move data
  - Fine-grained power sharing
- General purpose logic
  - What compute?
    - Energy-efficient cores / AI Accelerators
  - How much compute?
    - Thermal constraints can limit this (< 95^0c) (www.cs.utah.edu/wondp/eckert.pdf)
- Highly configurable bytes/flop
- Can be complemented by PIM


Structural silicon
64MB L3 cache die
Direct copper-to-copper bond
Through Silicon Vias (TSVs) for silicon-to-silicon communication
Up to 8-core "Zen 3" CCD

AMD V-cache

www.pcworld.com/article/394653/amd-v-cache-for-ryzen-everything-you-need-to-know.html



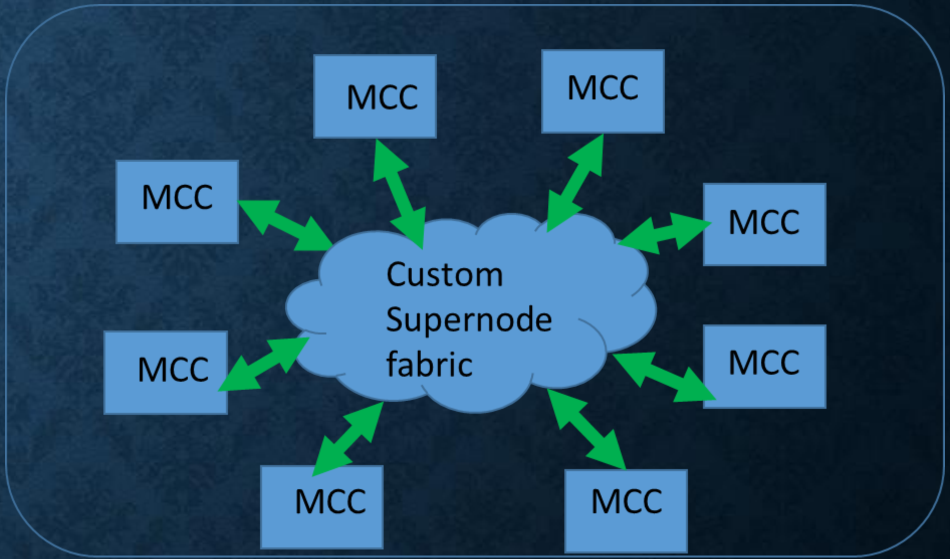
Source: Rambus Inc.

Data Movement: 62.6% of Power

SAMSUNG

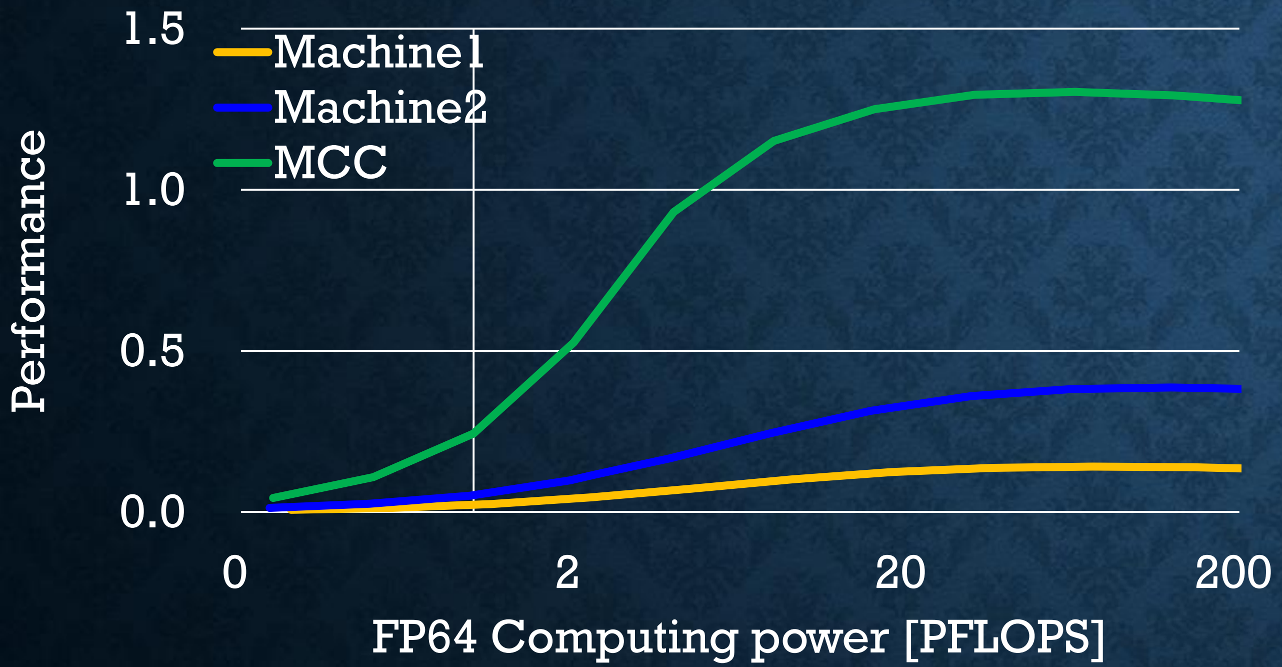
ATTACKING THE COMMUNICATION WALL

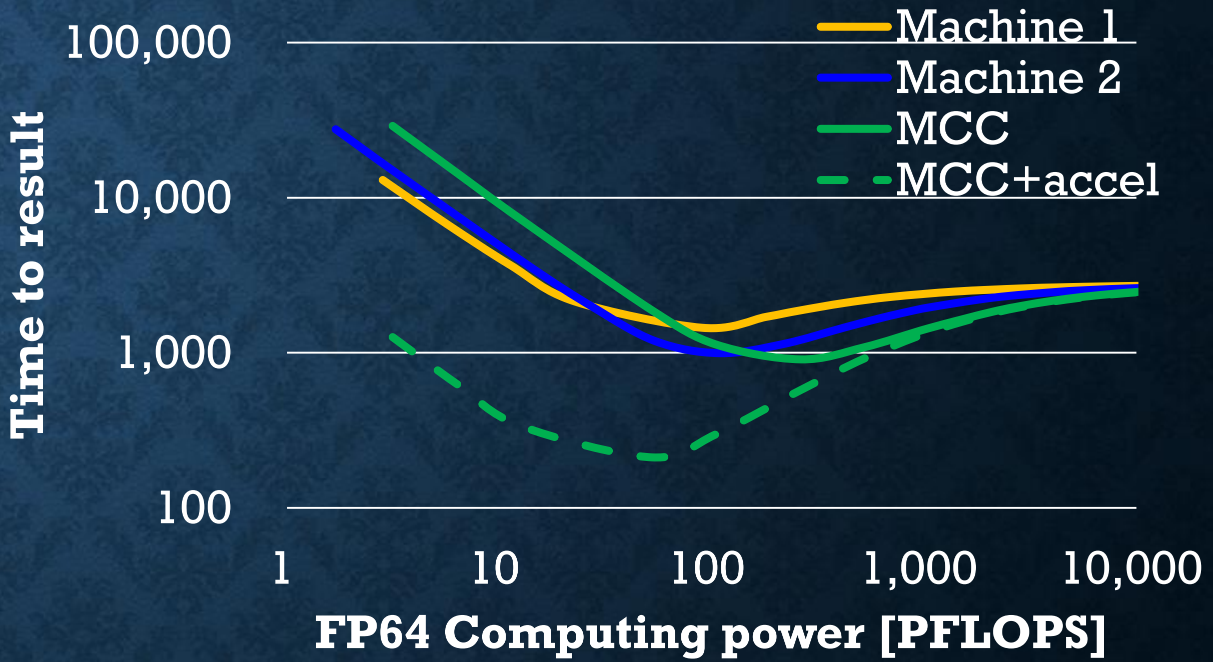
- Closer coupling of compute with memory and communication
  - ➢ Cost-efficient performance and power sharing
- Memory Coupled Compute Packaging → Higher Communication Efficiency and Perf
  - ➢ High point-to-point and all-to-all bandwidth
- Large Supernodes with productive programming model
  - ➢ Valuable to AI models for large reductions and large data exchanges



SAMSUNG

BENEFITS FOR A CLASSICAL HPC AND AN AI TRAINING APP

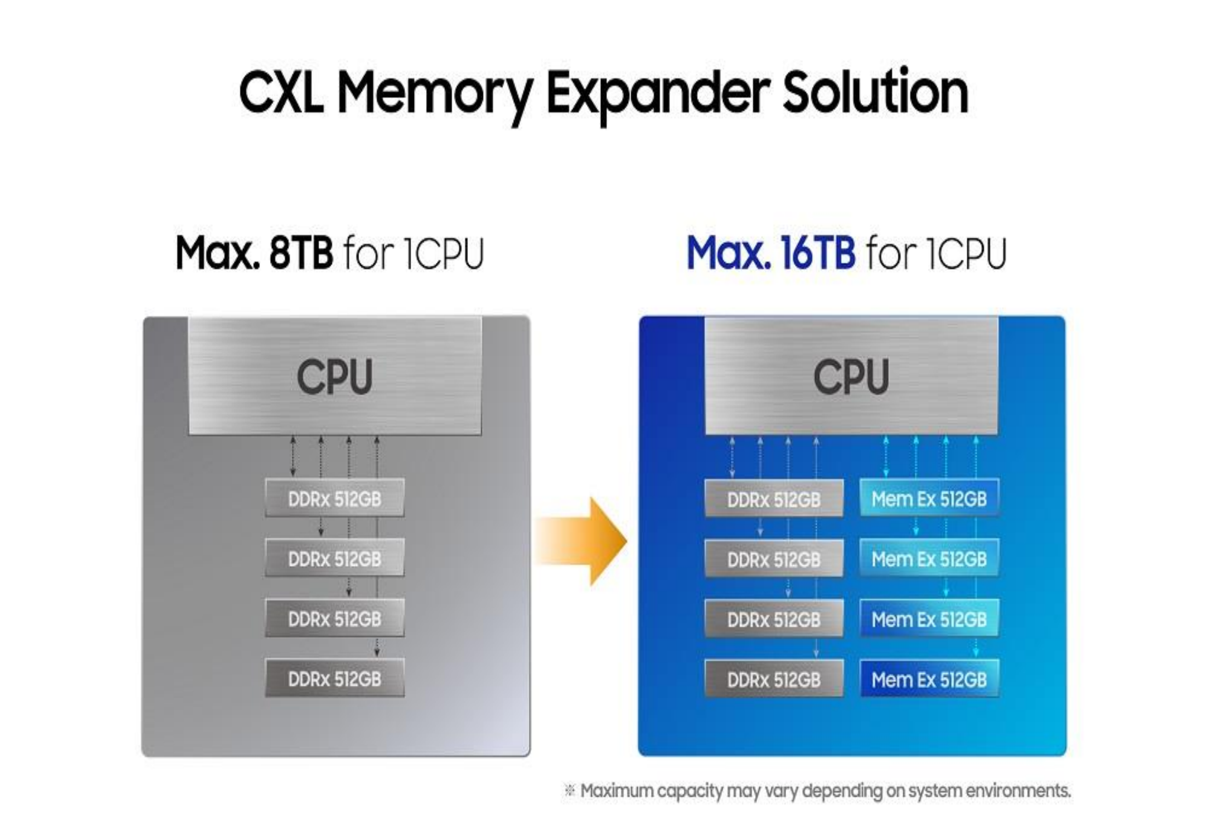




- Memory and communication bound classical HPC code
  - Y axis performance: higher is better

- Communication bound BF16 hungry multi-T AI app
  - Y axis time: lower is better

ADDRESSING MEMORY CAPACITY



CXL Memory Expander Solution

Max. 8TB for 1CPU

CPU

DDRx 512GB

DDRx 512GB

DDRx 512GB

DDRx 512GB

Max. 16TB for 1CPU

CPU

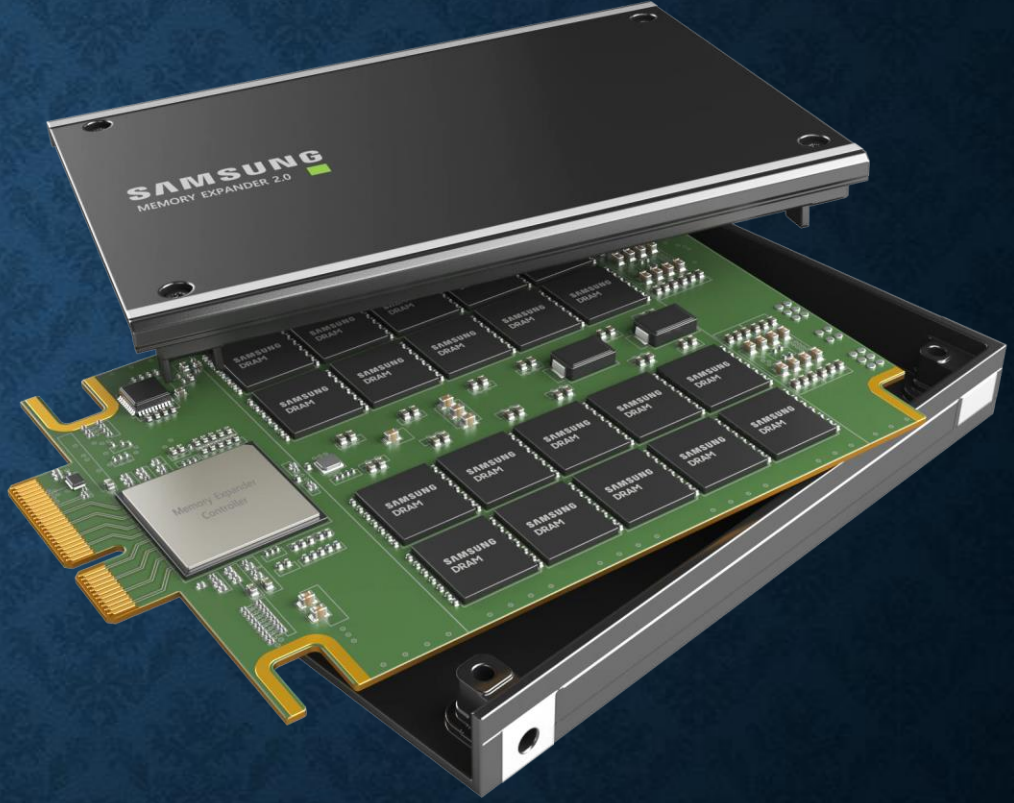
DDRx 512GB | Mem Ex 512GB

DDRx 512GB | Mem Ex 512GB

DDRx 512GB | Mem Ex 512GB

DDRx 512GB | Mem Ex 512GB

※ Maximum capacity may vary depending on system environments.



https://news.samsung.com/global/samsung-electronics-introduces-industrys-first-512gb-cxl-memory-module - May 2022

Productized CXL-DRAM based on PCIe5.0 with compatible software toolkit for hetero memory management
*Need CXL support from CPU

MEMORY COUPLED COMPUTE SUMMARY

- Tight coupling of compute, memory and communication
  - Advantages: Cost, Energy efficiency, Performance
- Samsung is the world leader in memory and silicon technology
  - Well positioned to drive this new technology
- CXL-based memory solutions can help address capacity concerns
- Come innovate the future with us!

THANK YOU

s.sury@samsung.com