Memory Coupled Compute Innovating the Next Discontinuity

Dr. Robert W. Wisniewski

Senior Vice President, Chief Architect of HPC

Head of Samsung's SAIT Systems Architecture Lab

SAIT (Samsung Advanced Institute of Technology) Samsung Electronics

SAMSUNG SAIT

Acknowledgments

- SAIT Systems Architecture Lab
  - Rohit Bhatia, Eric Borch, Ralph Castain, Sourav Chakraborty, Yong Chen, Jai Dayal, Aditya Deshpande, Praveen Francis, Manisha Gajbe, Alan Gara, Sharon Hall, Doug Joseph, Kelly Kim, Patrick LaFratta, David Lombard, Zaid McKie-Krisberg, James Loo, Alfredo Metere, Ali Raza, Rolf Riesen, Arun Rodrigues, Nick Romero, Samantika Sury, Andy Tauferner, Casey Thielen, Matt Turner, Jim Wayda, Robert Wisniewski, Matthew Wolf, Tina Zou, ++

- SAIT Korea
  - Byungwoo Bang, Wooseok Chang, Changkyu Choi, Youngjun Hong, Kyounghoon Kim, Sang Joon Kim, Taeksoo Kim, Soonwan Kwon, Sehwan Lee, Seungwon Lee, Seungwook Lee, Wonyong Lee, Junho Song, Eunsoo Shim, Sehyun Yang, SeokYoung Yoon, ++

Global SAIT (Samsung Advanced Institute of Technology) Labs

- SAIT was established in 1987 as a corporate R&D Center
  - Founding Philosophy: "Boundless Research for Breakthroughs"

History of SAIT Supercomputing

Since 1992, SAIT has provided supercomputing infrastructure for all Samsung employees


Cray YMP
('92.06 ~ '98.11)
1.33 GFlops


HP Exemplar S/V Class
('98.02 ~ '01.01)
51.2 GFlops


HP N Class
('01.01 ~ '02.09)
70.4 GFlops


HP Superdome
('02.10 ~ '05.10)
112 GFlops


Samsung Server BU
('05.08 ~ '10.05)
608 GFlops


HP Integrity SUN V40Z
('05.11 ~ '09.03)
284.8 GFlops


HP DL160/DL380
('09.04 ~'12.04)
9.7 TFlops


HP SL390/DL380
('11.06 ~)
66 TFlops


HP DL360/380 SL230/250/270
('13.10 ~)
140TFlops

Samsung Developed


HP SL270 XL170 ML350
('16.1 ~)
219TFlops


HP Apollo XL230K XL270D
('17.12 ~)
203TFlops


Cray CS500 HP XL270D
('18.12 ~)
191TFlops

Currently Running

High Performance Computing → Grid Computing → High Throughput Computing → Machine Learning

	1992	1995	2000	2002	2004	2005	2006	2007	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2021
CPU	1.3 GF	20.5 GF	222 GF	465 GF	1.0 TF	1.8 TF	1.8 TF	3.9 TF	12.3 TF	24 TF	90 TF	175 TF	301 TF	379 TF	505 TF	505 TF	568 TF	614 TF	4.2 PF
GPU											8 TF	117 TF	203 TF	495 TF	823 TF	962 TF	2 PF	10 PF	63 PF


Intel Paragon
('95.08 ~ '99.11)
19.2 GFlops


Alpha-11
('00.05 ~ '05.11)
170.7 GFlops


HP Superdome
('02.12 ~ '05.12)
112 GFlops


HP Integrity
('04.04 ~ '07.04)
192 GFlops


SUN V40Z HP Integrity
('06.01 ~ '09.03)
284.8 GFlops


SUN V4600 HP DL360
('07.05 ~ '10.5)
1.4 TFlops


HP DL160/380
('10.06 ~ '13.10)
14 TFlops


HP DL360/SL270
('12.07 ~)
95TFlops


HP SL250/270 DL360/380
('14.10 ~)
144TFlops

Samsung Developed

'16.11~ : 2,046 units
'17.12~ : 3,116 units
'18.12~ : 5,876 units

Currently Running

TOP 500 The List
11th ('21) → 15th ('22)

THE GREEN 500
2nd ('21) → 6th ('22)

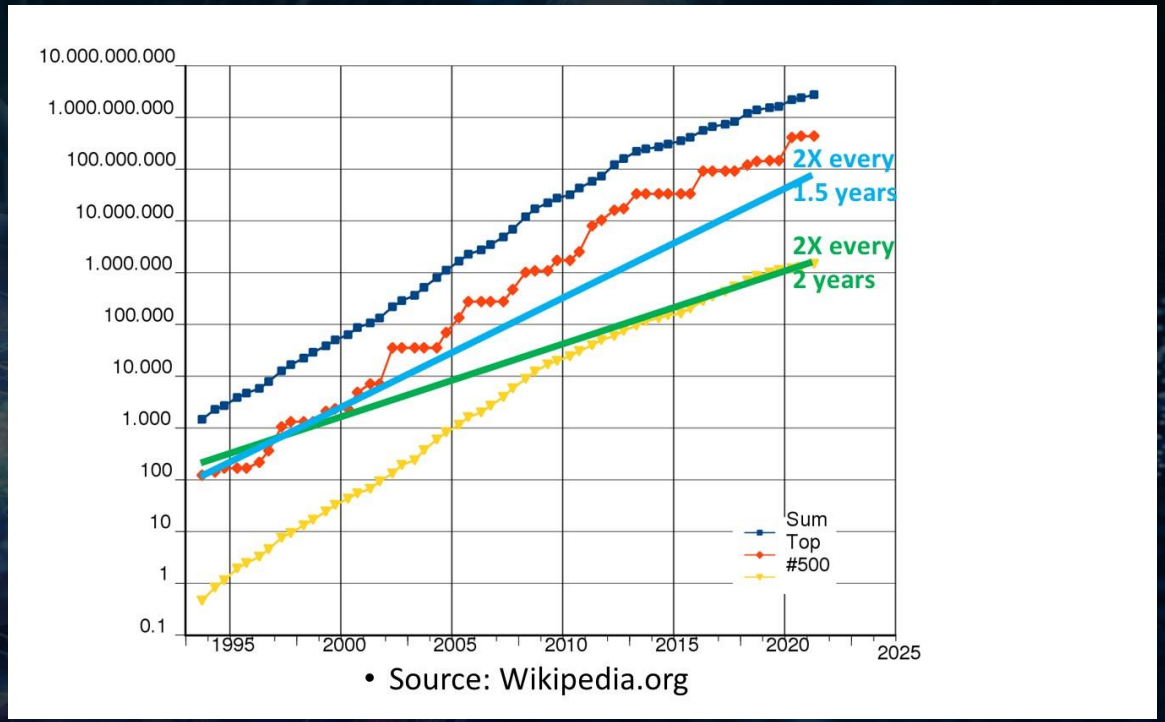
SAMSUNG SAIT

Systems Architecture Lab

- Vision
  - To develop the most innovative technologies for future HPC and AI systems
- Strategy
  - To break through the memory wall by significantly increasing the memory byte/flop ratio and reducing the power per bit with memory coupled compute
  - To break through the communication wall with high network byte/flop ratio utilizing memory coupled compute efficiencies and novel fabric technologies

Discontinuities

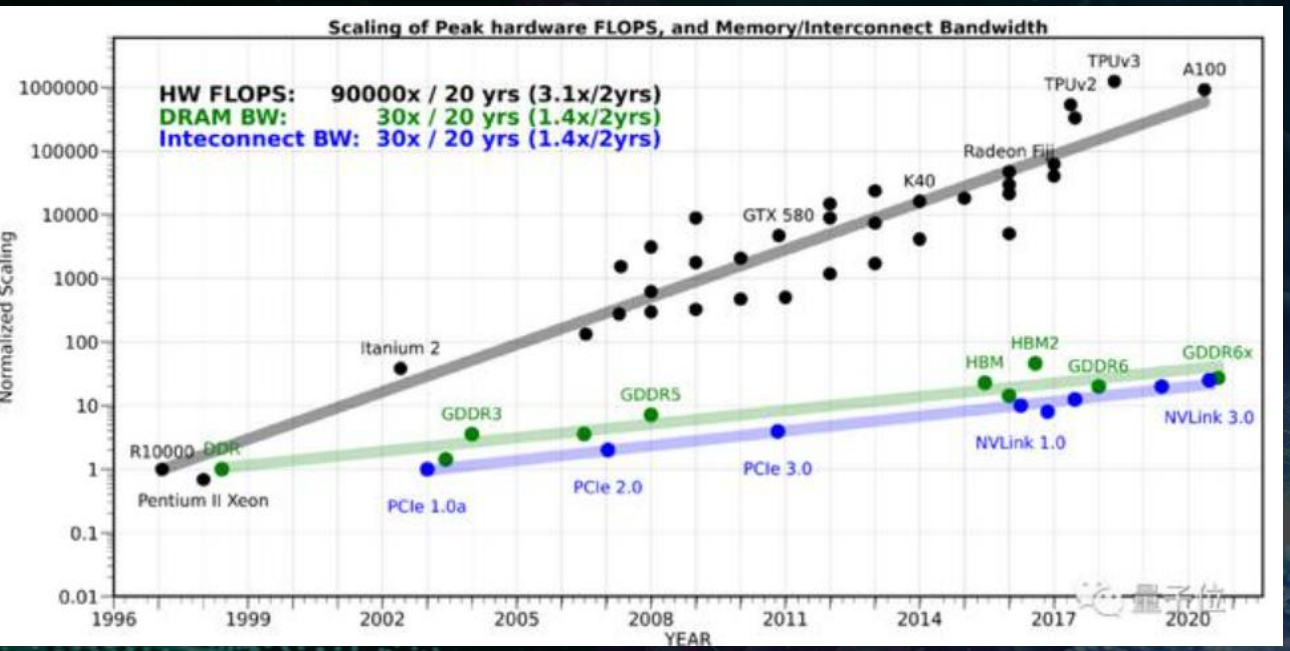
- Vectors (Cray)
- Microprocessors (Beowulf)
- Multicore, multithread (x86/ Power)
- Massive parallelism (Blue Gene)
- Heterogeneity (GPUs)
- Memory coupled compute
  - The next discontinuity
  - Innovate the future collaboratively

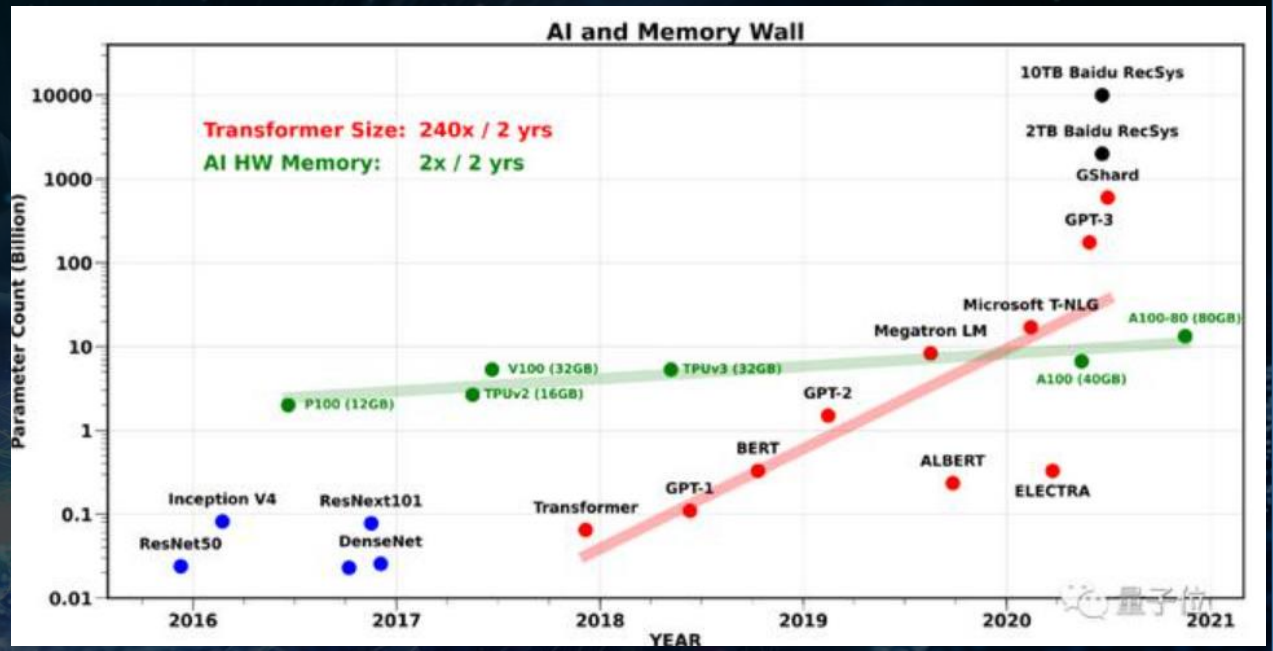


System Overview

- Key Innovations
  - Memory Coupled Compute (3D integration of memory with processors)
  - Supernodes (large and high-performing globally accessible memory)
  - Productive and tight-coupling of mainstream cores with accelerators
  - Extreme system-level energy-efficiency
- Key Goals
  - HPCG performance and energy efficiency
  - Green500

The Memory and Communication Wall is getting Higher





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

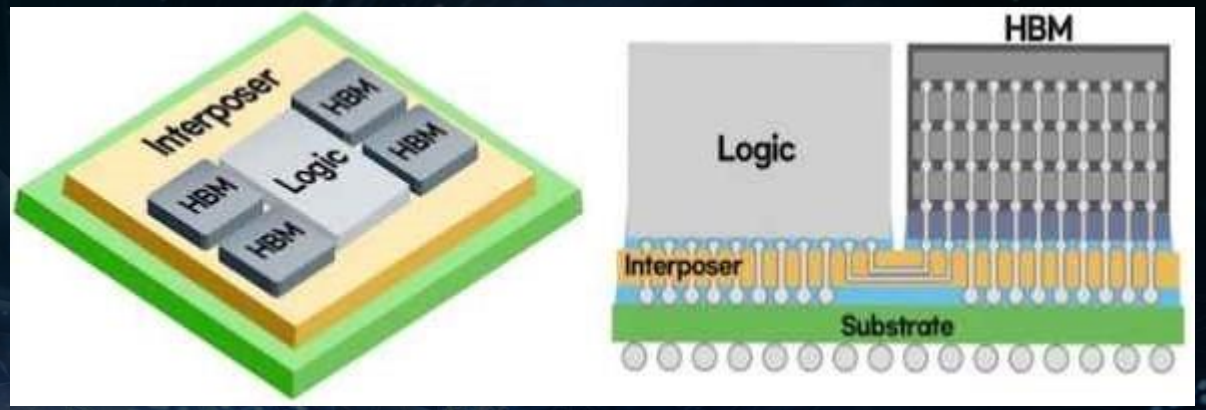
- Modeling and simulation applications are memory bandwidth limited
- AI, and some mod/sim, applications are communication bandwidth limited

Attacking the Memory Wall

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

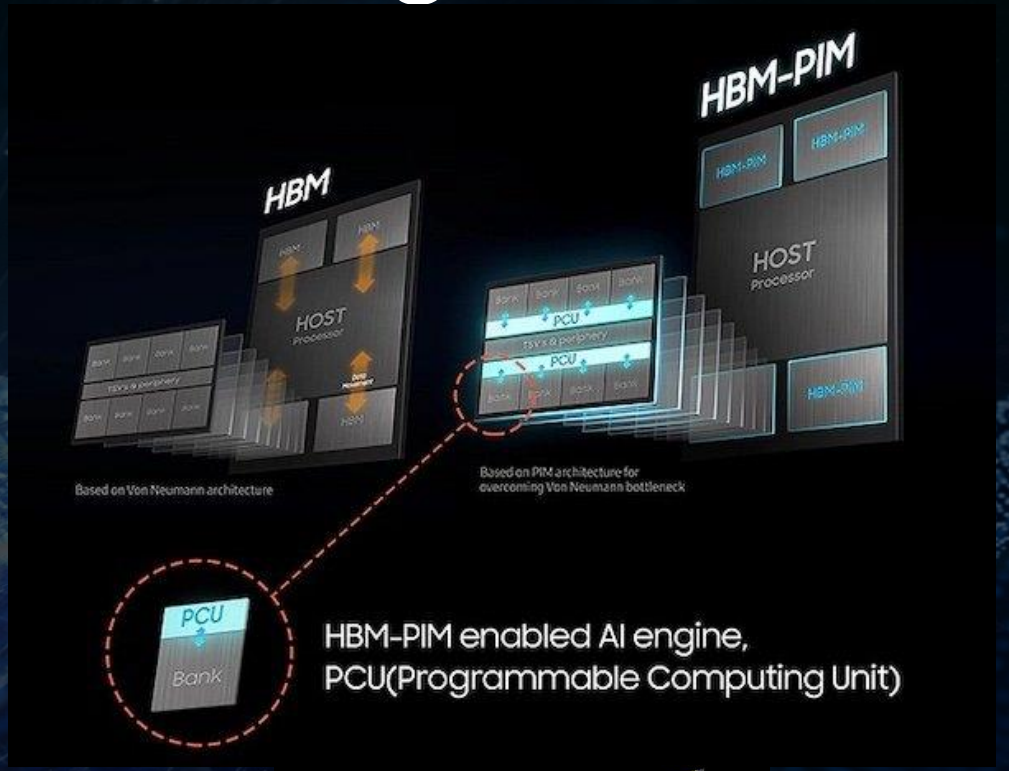
2.5D Opportunities and Challenges

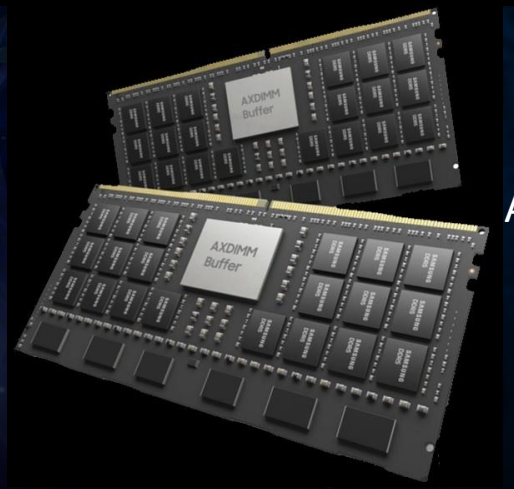
- Significant improvement over DDR
  - Bandwidth is higher
  - Latency on par
- Substrate and connections can be expensive
- Requires off die connection from logic to HBM
  - Off-die signals require more power
  - Takes die area to connect the wires



PIM Opportunities and Challenges

- Most energy efficient compute
  - ALUs on same die as memory cells
  - Data movement is minimal
- The type of operations are constrained
  - ALUs reduce memory are or increase die area
- The operations are synchronous
  - If conforming to JEDEC standard



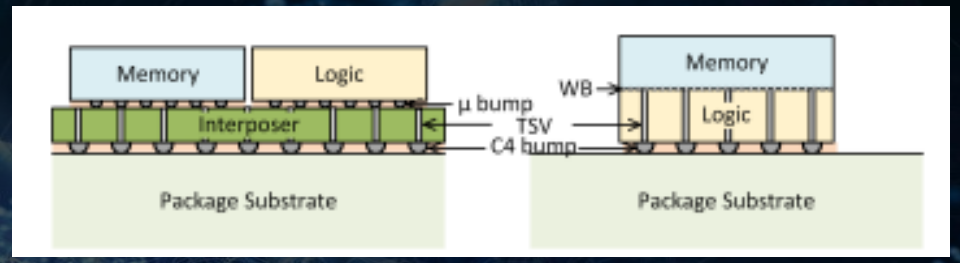


AXDIMM

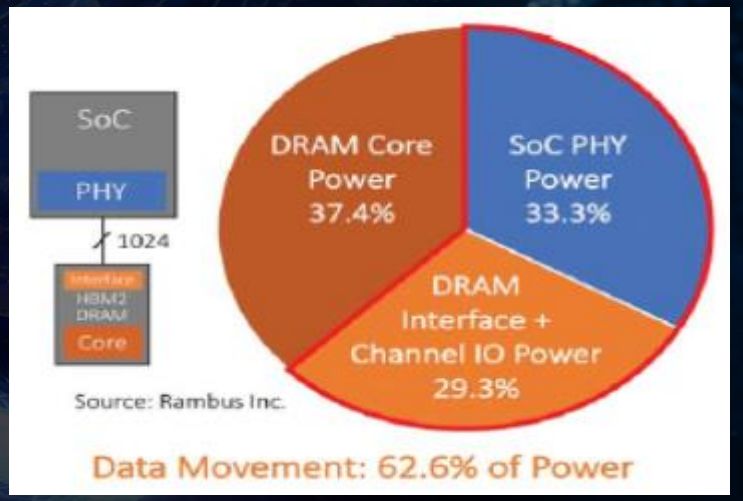
Memory Coupled Compute

- Improves power efficiency
  - Data moves less
- Reduces latency
  - Data travels less distance
- Allows general purpose logic
- Key decisions
  - What compute
    - Just CPU? Heterogeneous?
    - Keep the programming model productive
  - How much compute
    - Provides opportunity for high B/F ratio

Closer coupling of compute with memory

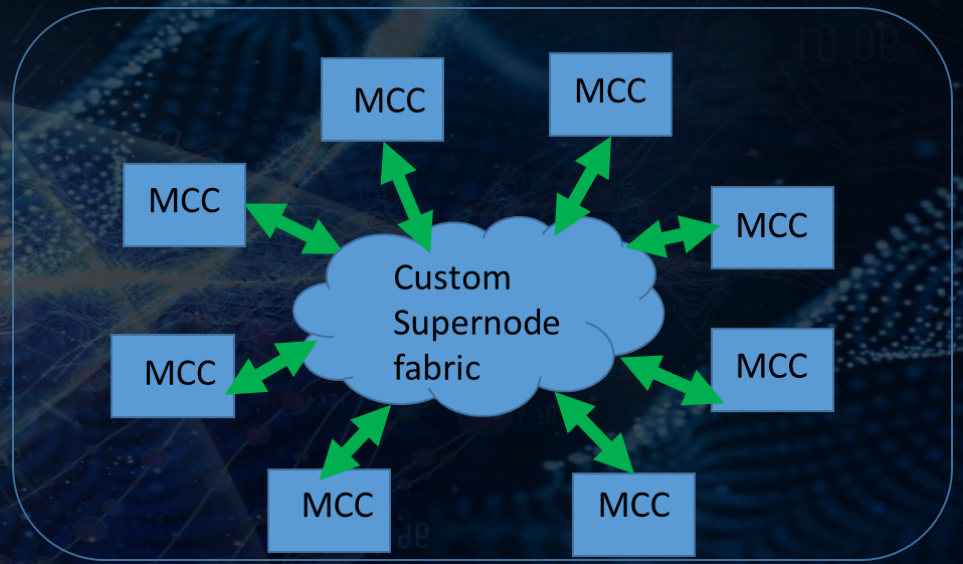


e.g. 3D systolic ML accelerators in IEEE Journal on Exploratory Solid-State Computational Devices and Circuits – June 2021

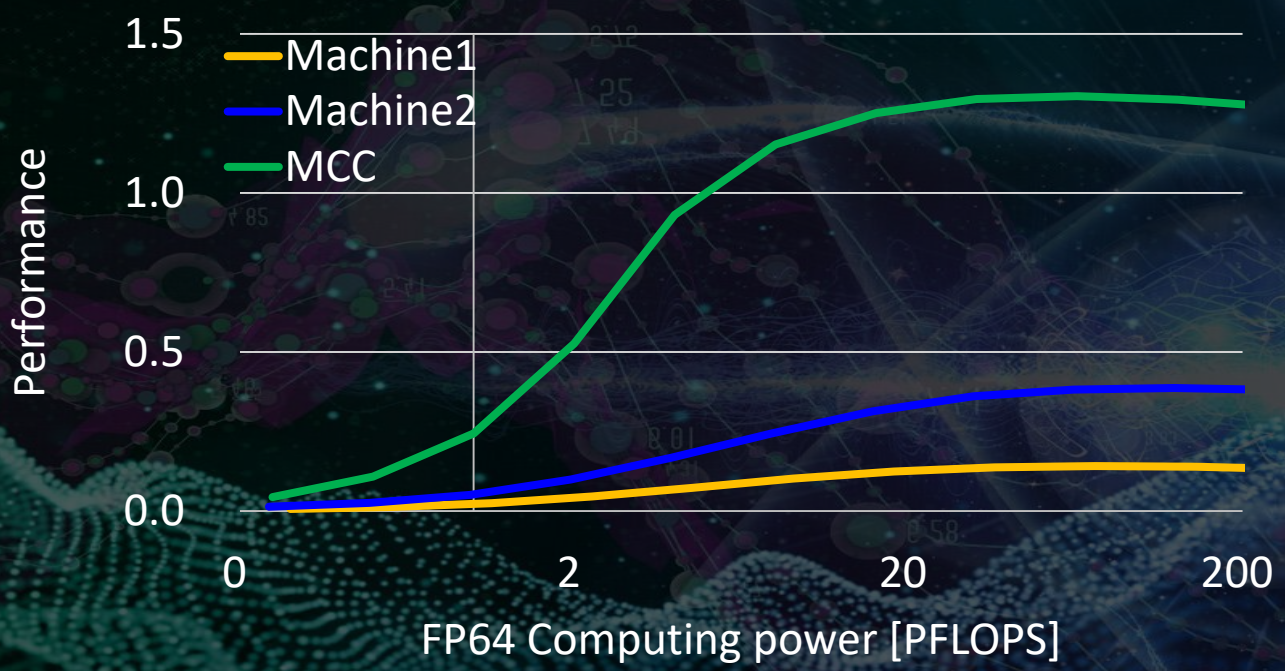


Attacking the Communication Wall

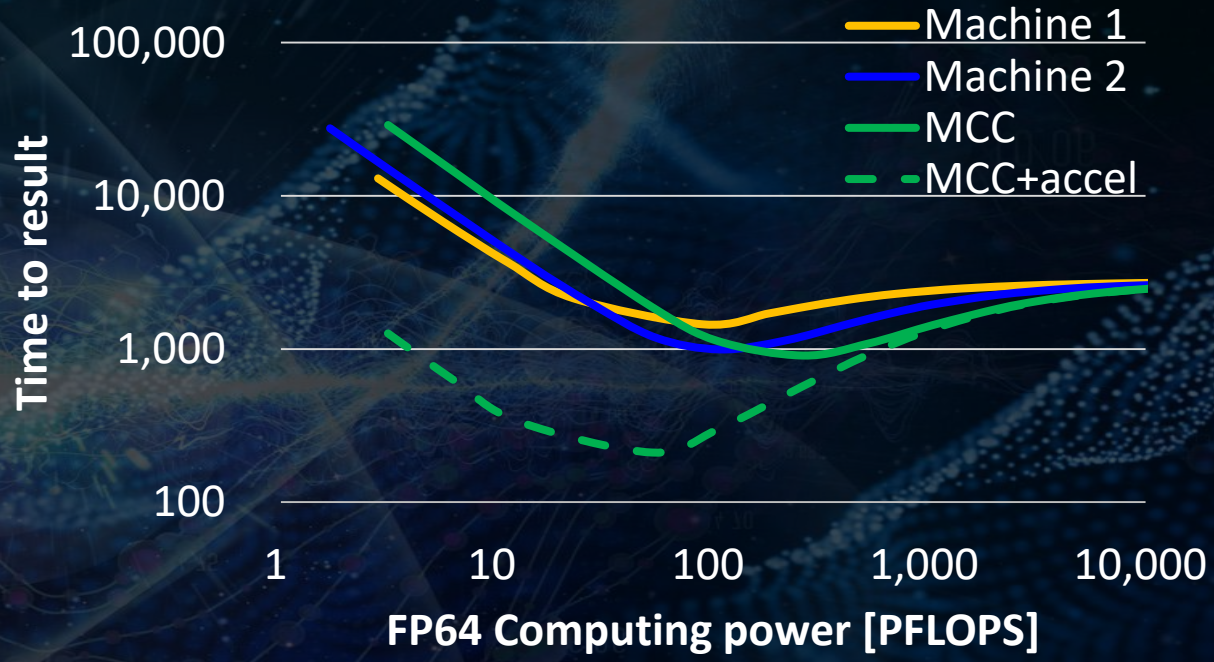
- Closer coupling of compute with memory and communication
  - Cost-efficient performance and power sharing
- Memory Coupled Compute packaging → Higher Communication Performance, driving efficiency
  - High point-to-point and all-to-all bandwidth
- Large supernodes with productive programming model
  - Valuable to AI models for large reductions and large data exchanges, parallel FFT



Benefits for a Classical HPC and an AI Training Application

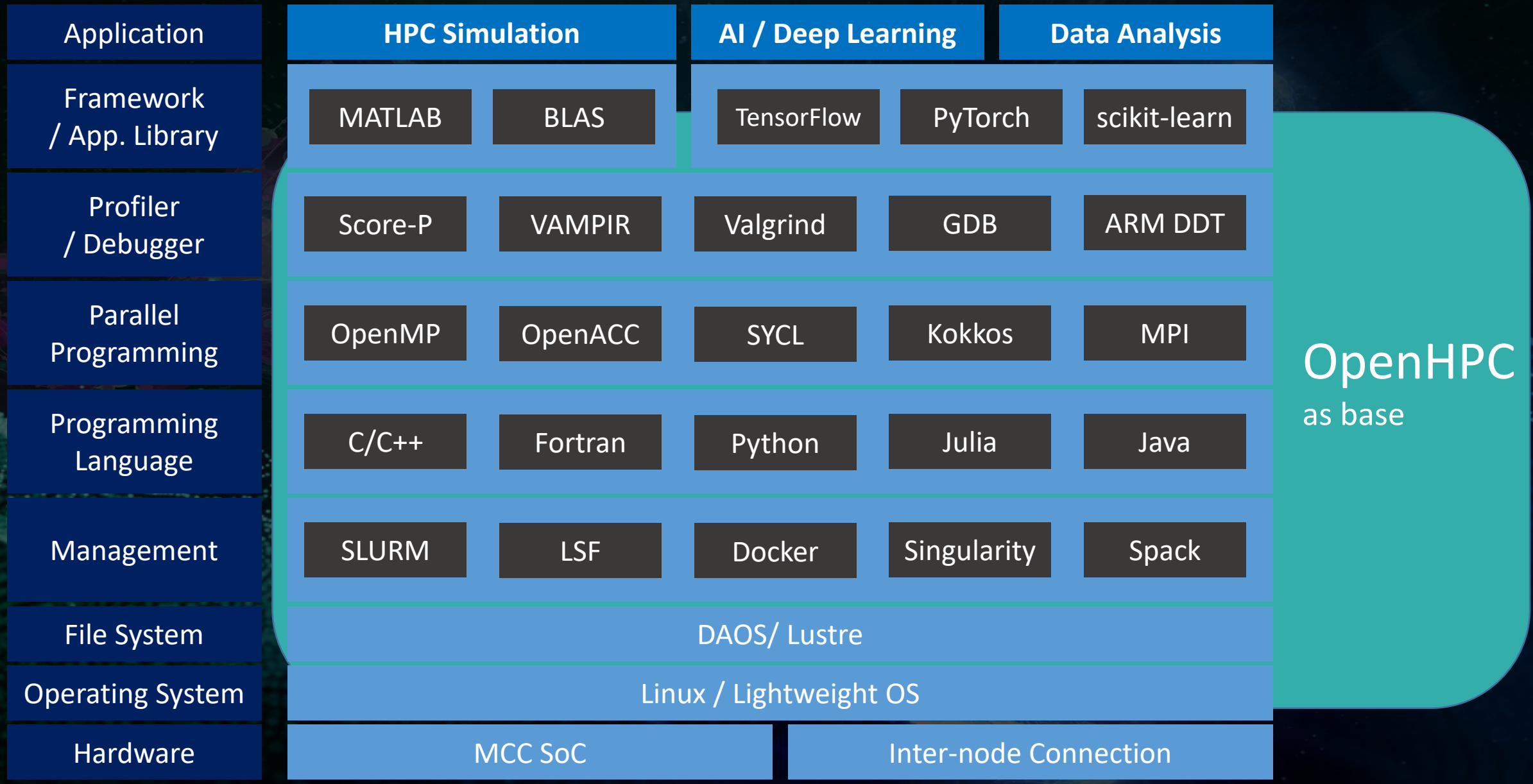


- Memory and communication bound classical HPC code
  - Y axis performance: higher is better



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

Standard Productive Software Stack



Software Innovations and Directions

- Standard productive programming model
  - Coherent shared memory within a node including potential accelerators
  - Globally accessible memory between groups of nodes
    - Exploring innovative memory models leveraging co-designed hardware mechanisms
- Scalable and efficient messaging leveraging the communication innovation
- Linux compatibility
  - Lightweight kernel for ultra scalability and high performance
- Unified control system
  - Providing single source of information and manageability
- Storage system with high bandwidth and efficient checkpoints
- Optimized AI and Big Data frameworks

The Importance of Co-Design

- Goal is to build a broadly compelling MCC
- Internal machine requirements represent one design point
- Look to utilize input from various key architectural design points
  - Traditional HPC
  - AI
  - Supernode connectivity
  - System fabric tapering
  - Etc., etc., etc..

Innovating the Next Discontinuity

- The time is right to innovate the next discontinuity
  - Vision: In the future memory coupled compute will be ubiquitous
- Samsung is the world leader in memory and silicon technology
  - Well positioned to drive the vision
- Integration of memory and compute will allow future optimizations
- Our investigations are focused on AI and HPC systems
- Come innovate the future with us


Thank
You