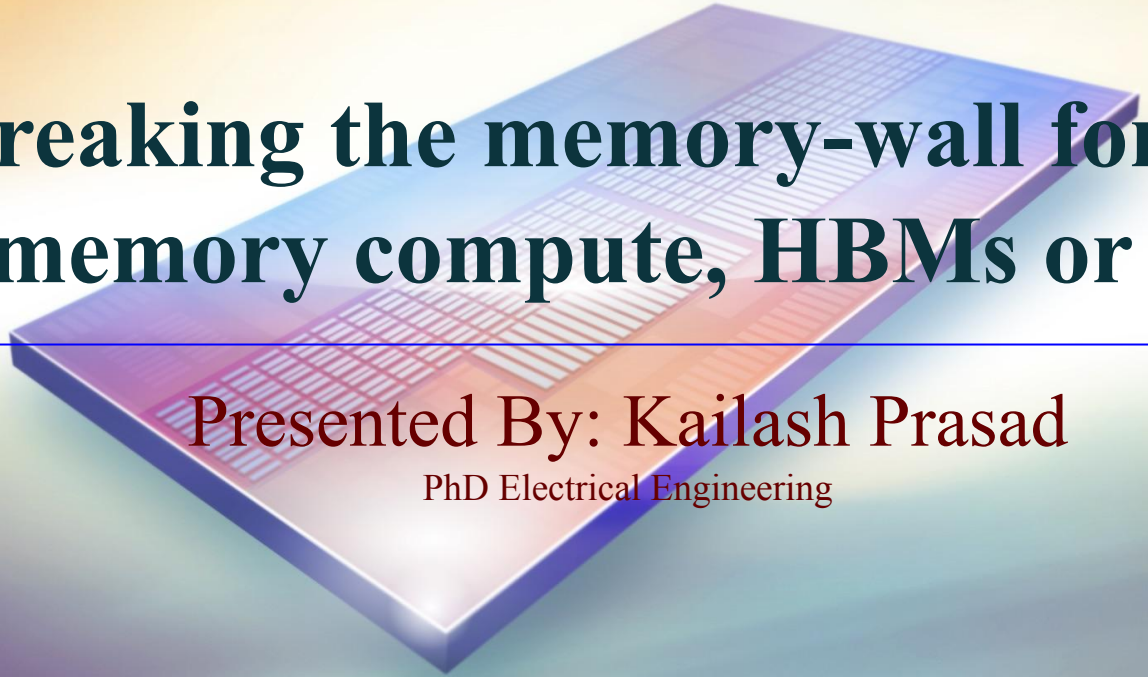


Breaking the memory-wall for AI: In-memory compute, HBMs or both?

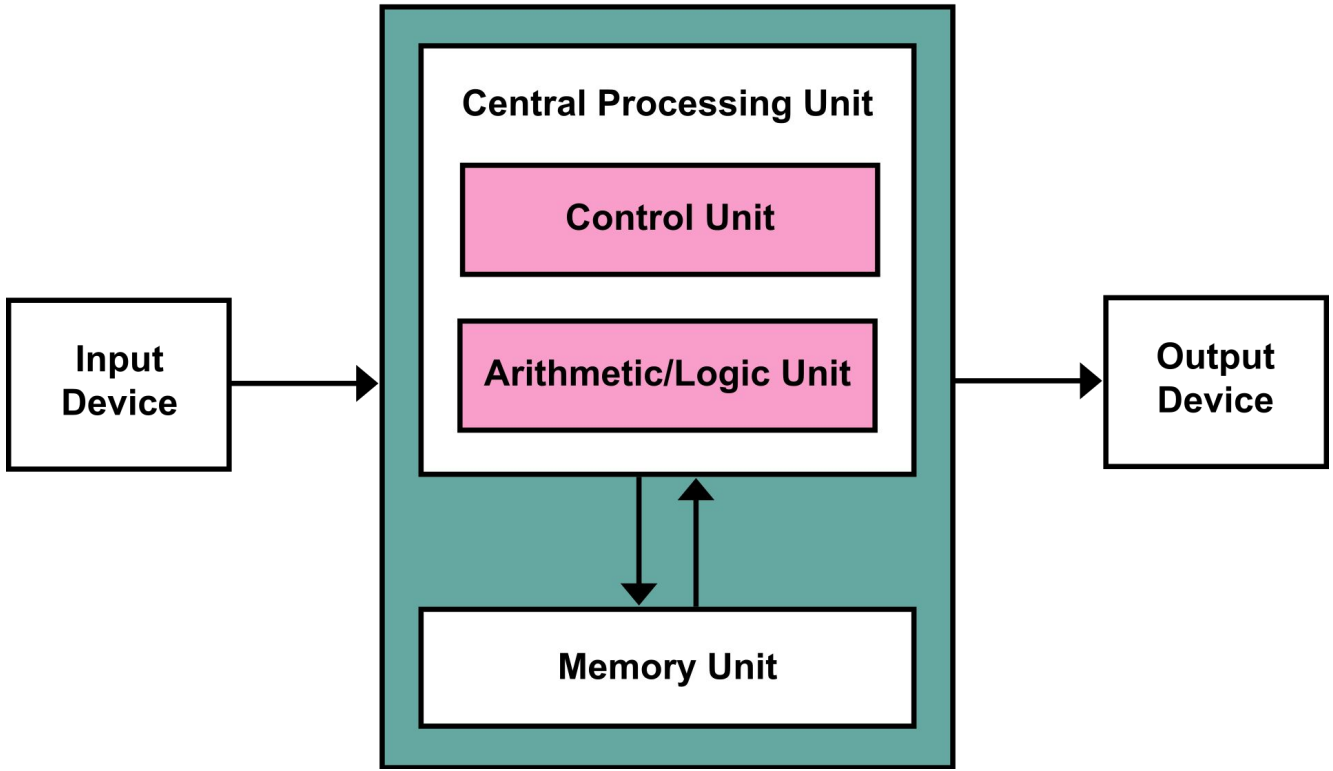
Presented By: Kailash Prasad

PhD Electrical Engineering

Outline

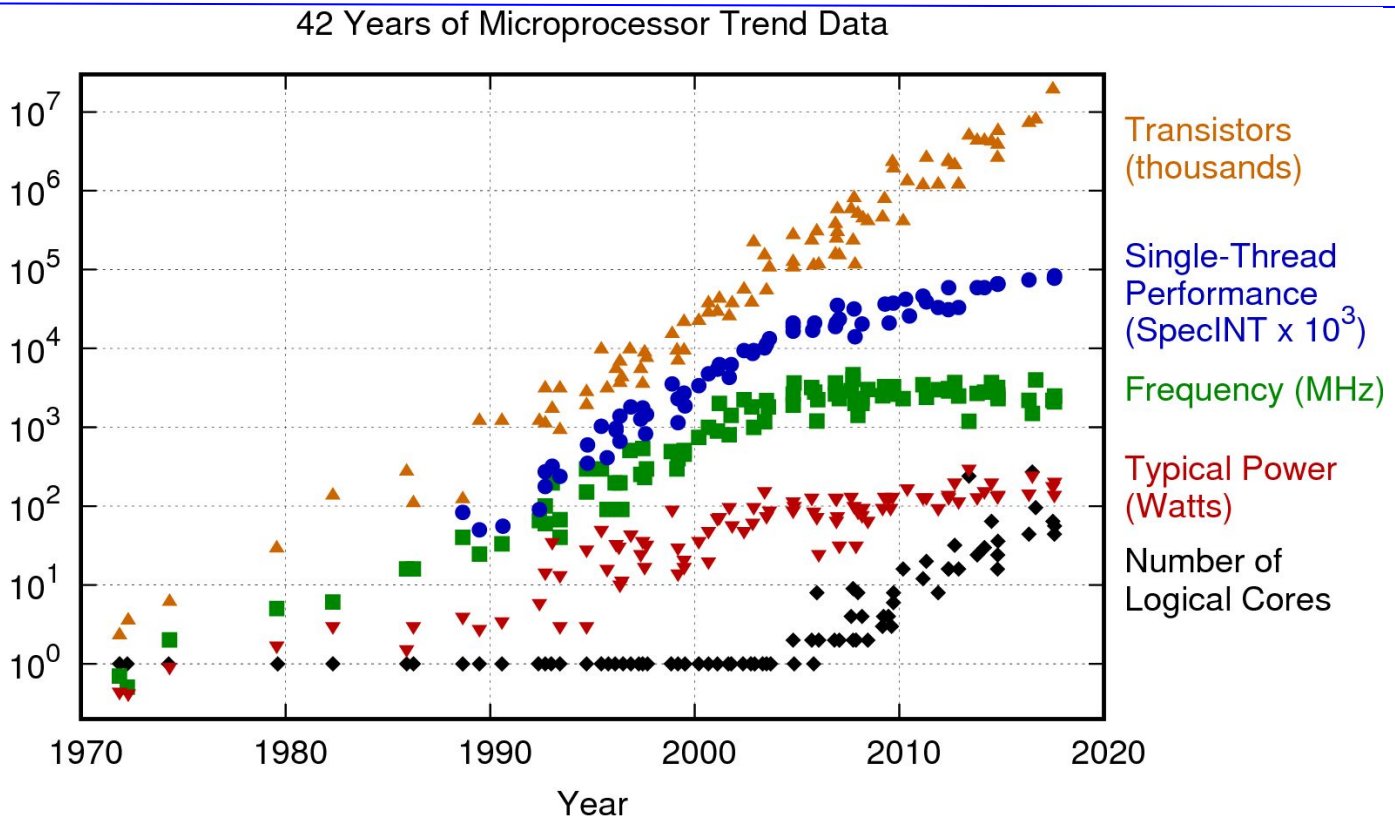
- Motivation
- HBM - High Bandwidth Memory
- IMC - In Memory Computing
- In-memory compute, HBMs or both?
- Conclusion

Von - Neumann Architecture



Source: https://en.wikipedia.org/wiki/Von_Neumann_architecture#/media/File:Von_Neumann_Architecture.svg

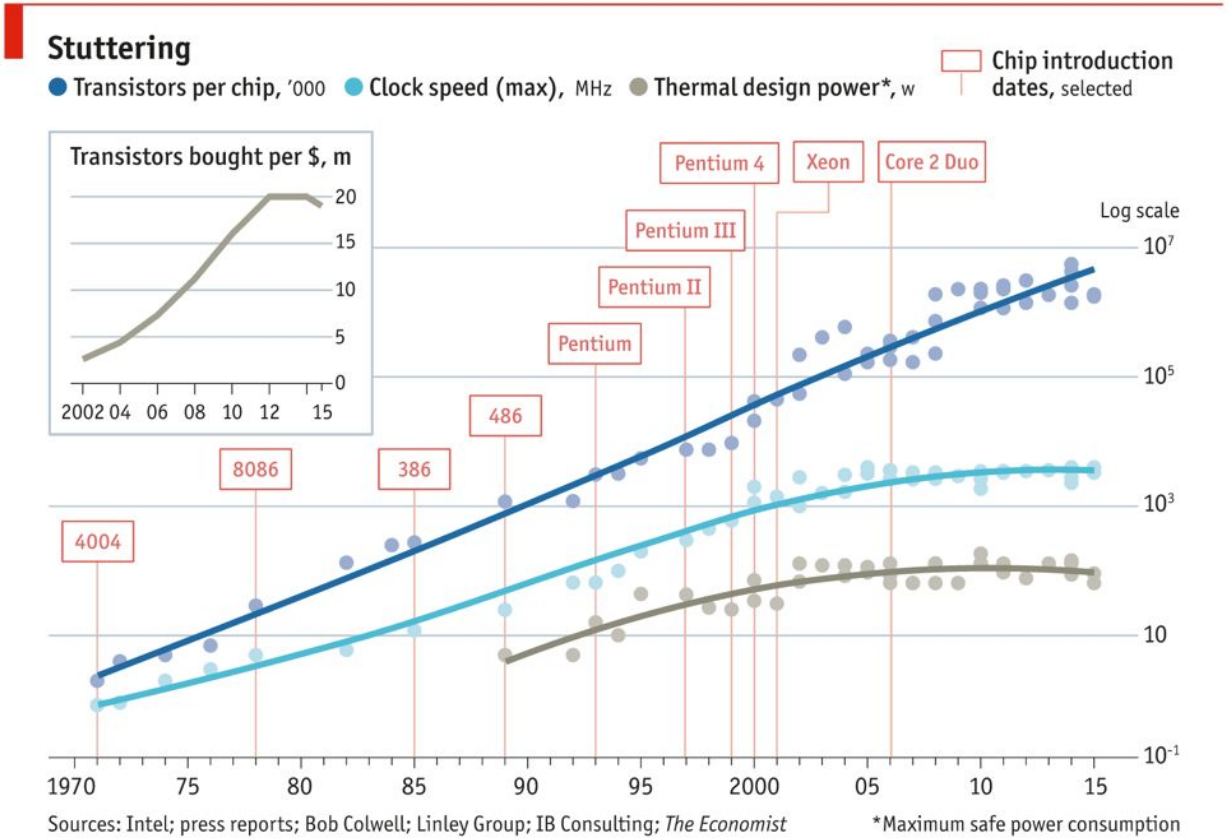
Moore's Law



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

Source: https://www.researchgate.net/figure/Evolution-of-electronic-CMOS-characteristics-over-time-41-Transistor-counts-orange_fig2_325420752

Power Wall



Source: https://medium.com/artis-ventures/omicia-moores-law-the-coming-genomics-revolution-2cd31c8e5cd0

Hwang's Law

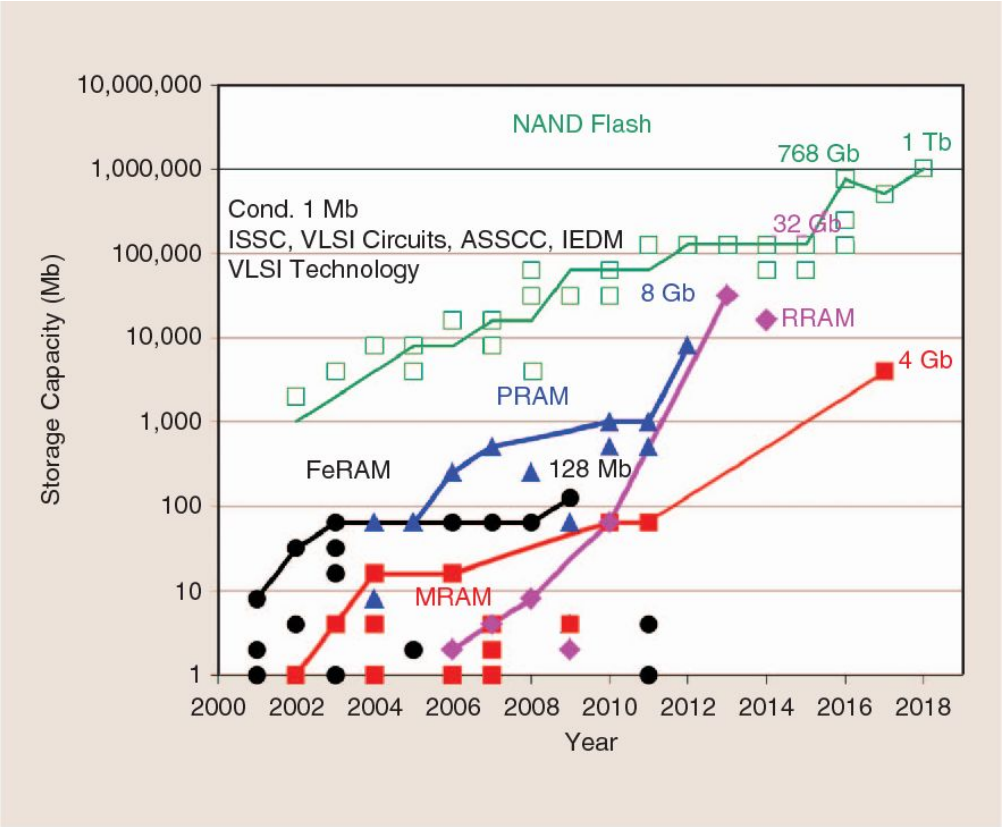
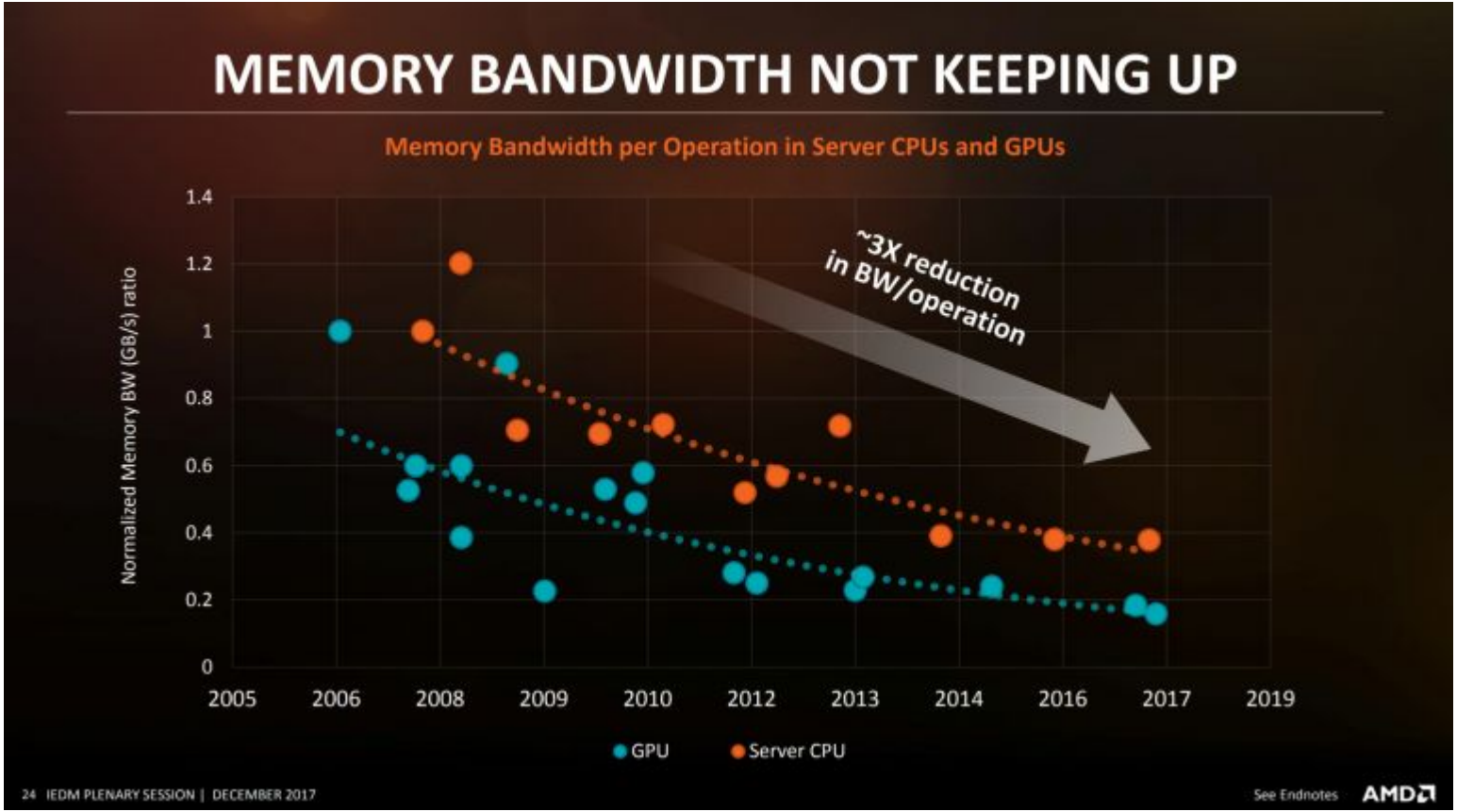


FIGURE 30: Memory-capacity trends for emerging NVMs. ASSCC: IEEE Asian Solid-State

Hwang's Law:
Chip capacity will double every year

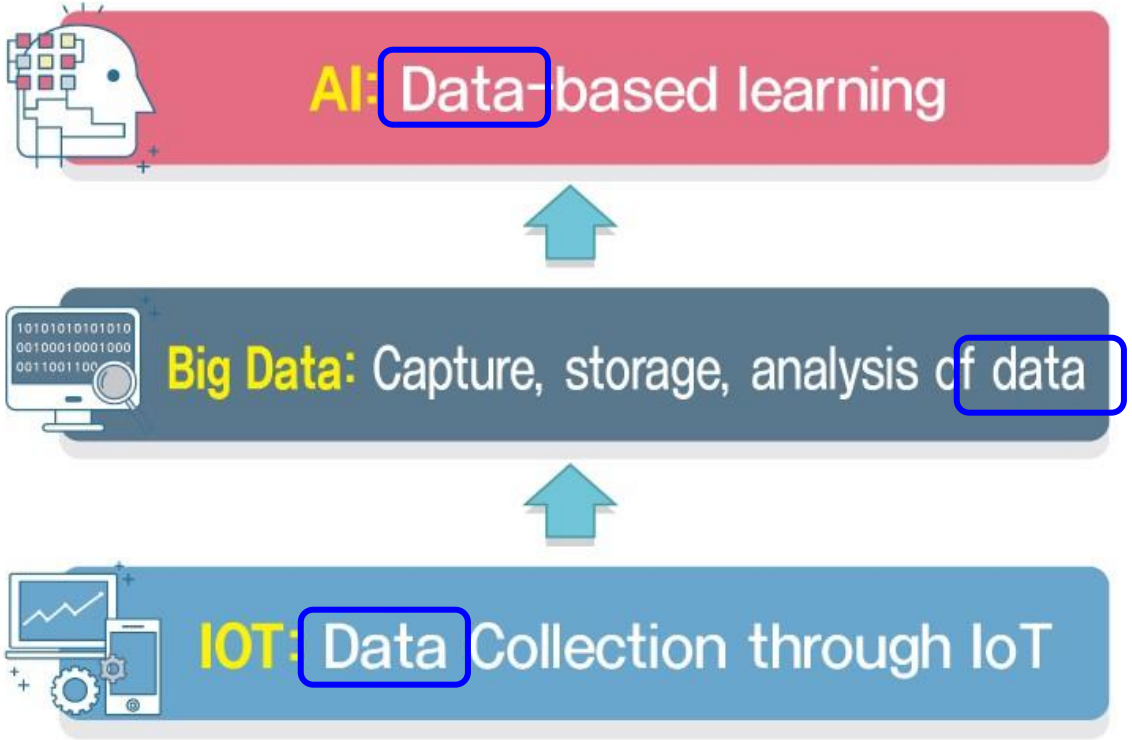
Source: K. C. Smith, A. Wang and L. C. Fujino, "Through the Looking Glass: Trend Tracking for ISSCC 2012," in IEEE Solid-State Circuits Magazine

Memory Wall



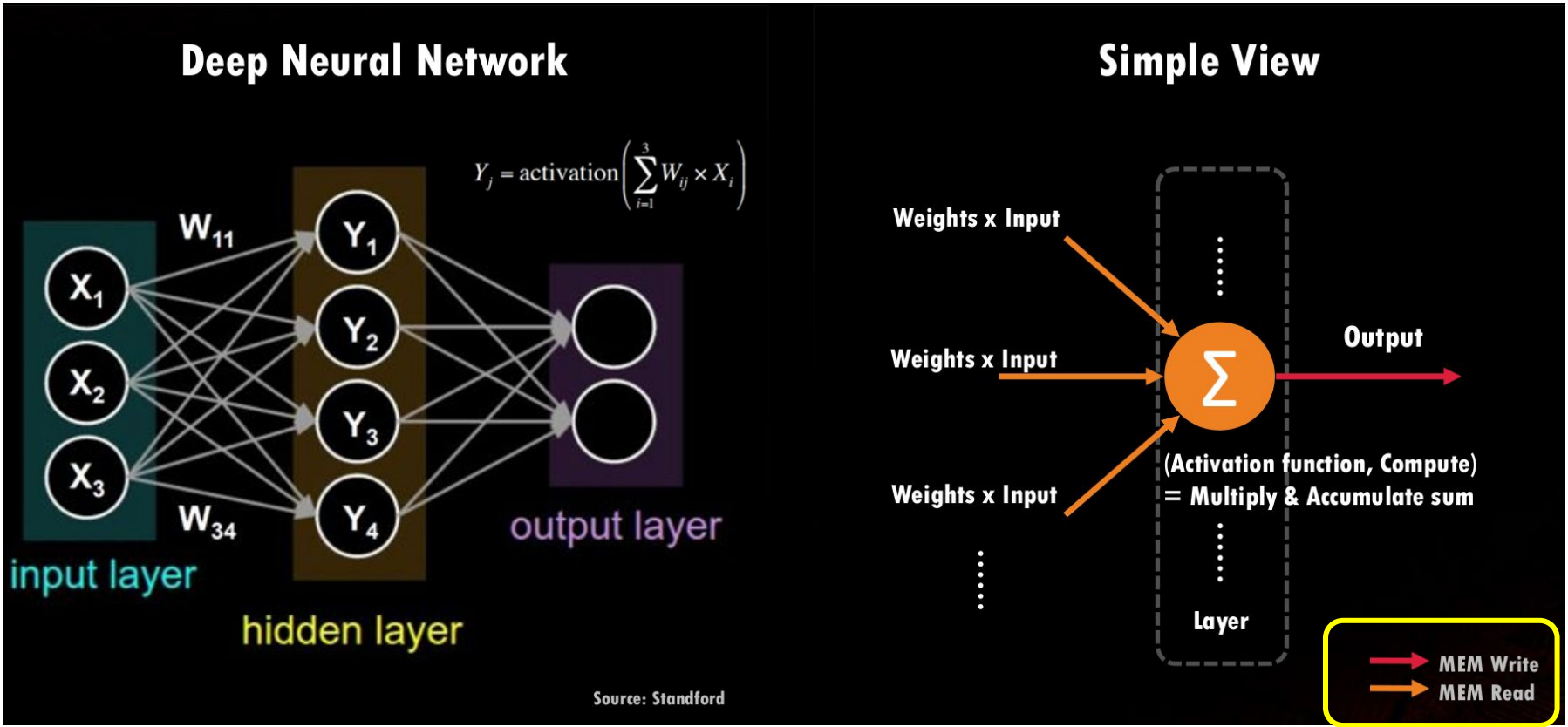
Source: http://on-demand.gputechconf.com/gtc/2018/presentation/s8949

Why this is Important?



Source: https://hackernoon.com/the-most-promising-internet-of-things-trends-for-2018-10a852ccd189

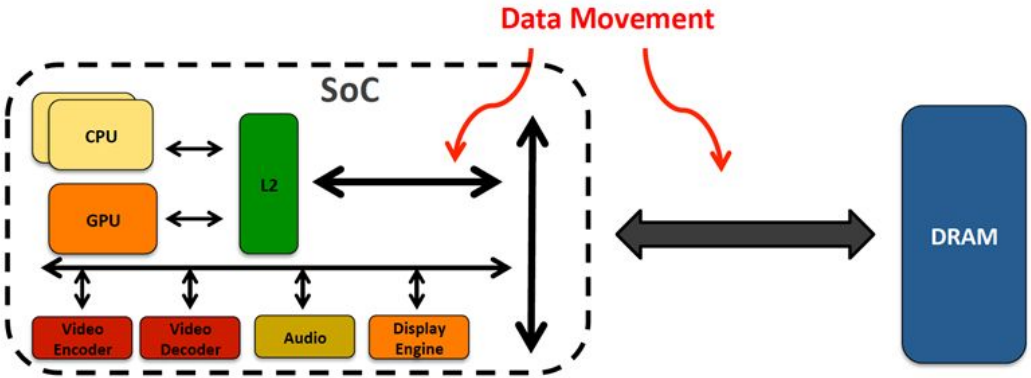
Deep Neural Network



Source: http://on-demand.gputechconf.com/gtc/2018/presentation/s8949

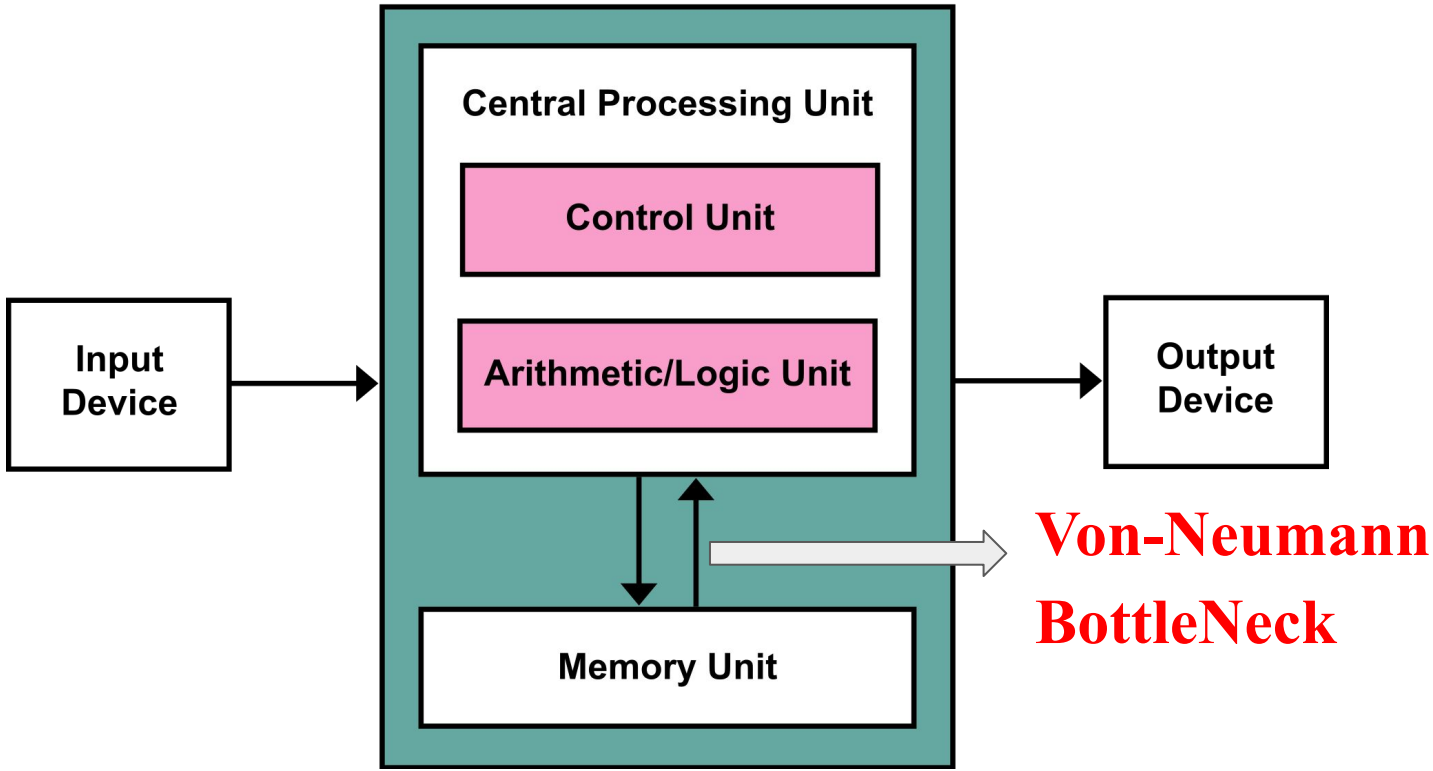
Data Movement vs. Computation Energy

- Data movement is a major system energy bottleneck
  - Comprises **41% of mobile system energy** during web browsing [2]
  - Costs **~115 times** as much energy as an ADD operation [1, 2]



[1]: Reducing data Movement Energy via Online Data Clustering and Encoding (MICRO'16)
[2]: Quantifying the energy cost of data movement for emerging smart phone workloads on mobile platforms (IISWC'14)

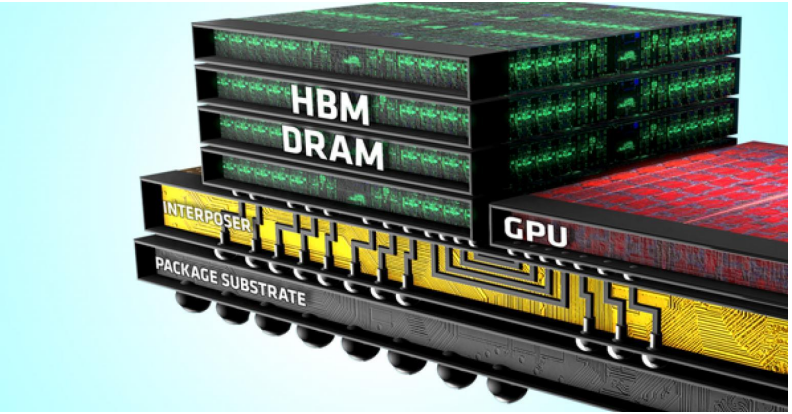
Problem - Von Neumann Bottleneck



Source: https://en.wikipedia.org/wiki/Von_Neumann_architecture#/media/File:Von_Neumann_Architecture.svg

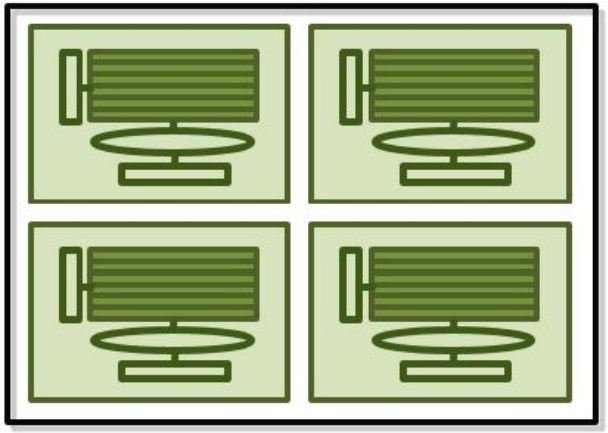
Solutions

Increase Memory Bandwidth



High Bandwidth Memory

Reduce Data Movement

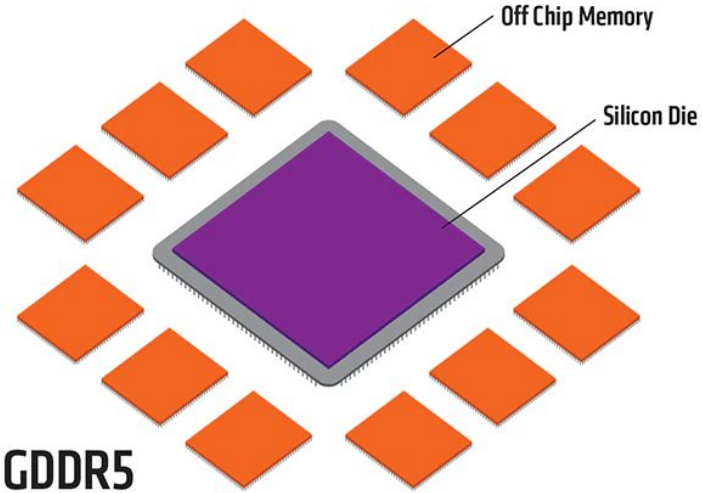


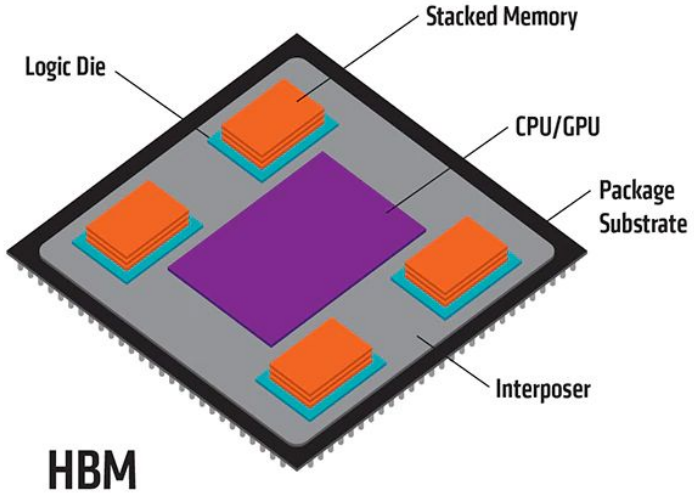
In Memory Computing

Source: https://www.electronicdesign.com/memory/qa-taking-closer-look-amd-s-high-bandwidth-memory

High Bandwidth Memory

High Bandwidth Memory





Source: https://www.amd.com/system/files/49010-high-bandwidth-memory-hbm-1260x709.jpg

Device and Working

- HBM is a new type of memory chip
  - Low power consumption
  - Ultra-wide communication lanes.
- Multiple DRAM dies stacked in a single package connected by **TSV** and **Microbumps**
- Connected to the processor unit directly via **Interposer Layer**
- Developed by **AMD** and **SK Hynix**
- **JEDEC** industry standard in October 2013

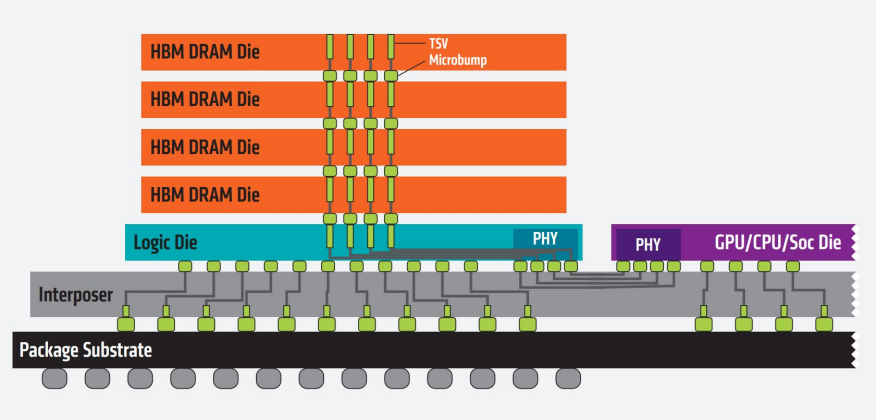
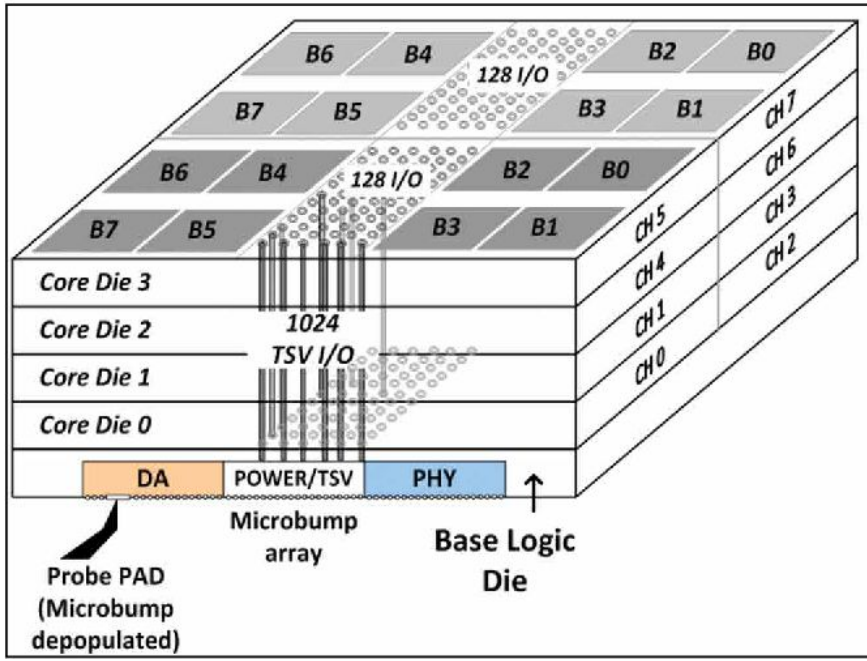
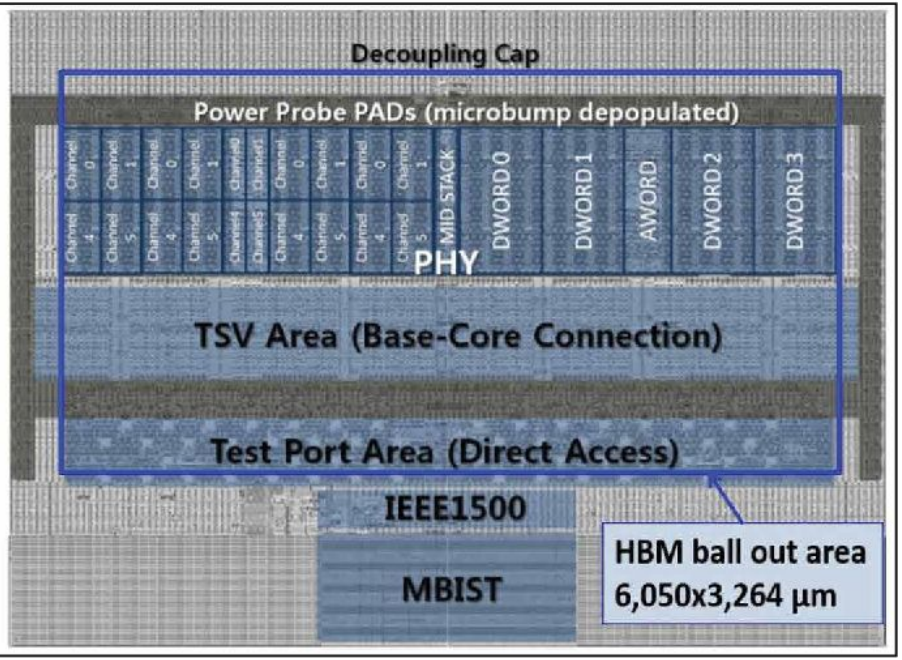


Image Source: https://www.amd.com/system/files/49010-high-bandwidth-memory-hbm-1260x709.jpg

HBM Architecture





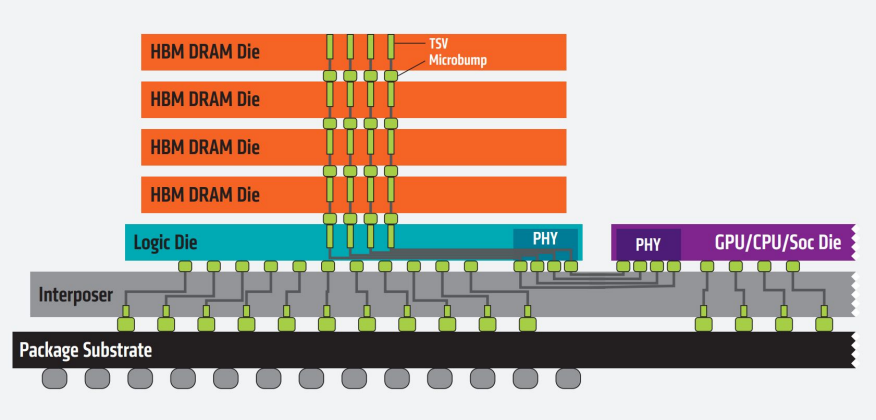
Example consisting of four DRAM core dies and one base logic die

A photo of the base die

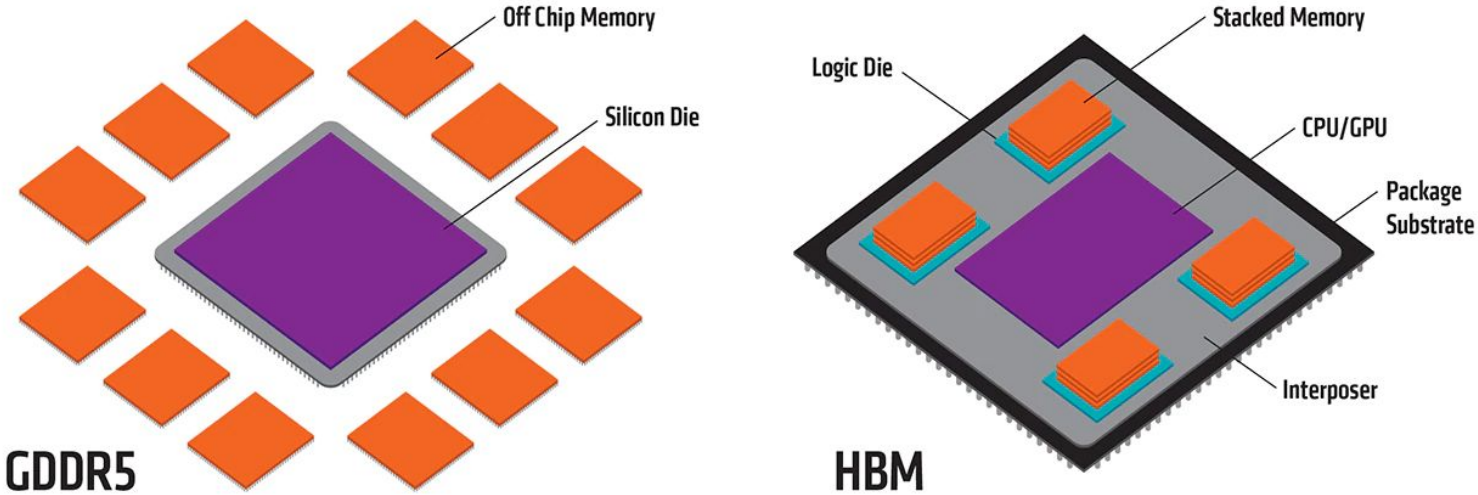
Source:H. Jun, S. Nam, H. Jin, J. Lee, Y. J. Park and J. J. Lee, "High-Bandwidth Memory (HBM) Test Challenges and Solutions," in IEEE Design & Test

Benefits over State of Art Memories

- Very **High Bandwidth**
- **Lower Effective Clock** Speed
- Smaller Package
- **Shorter Interconnect** Wires
- **Lower Power** Consumption

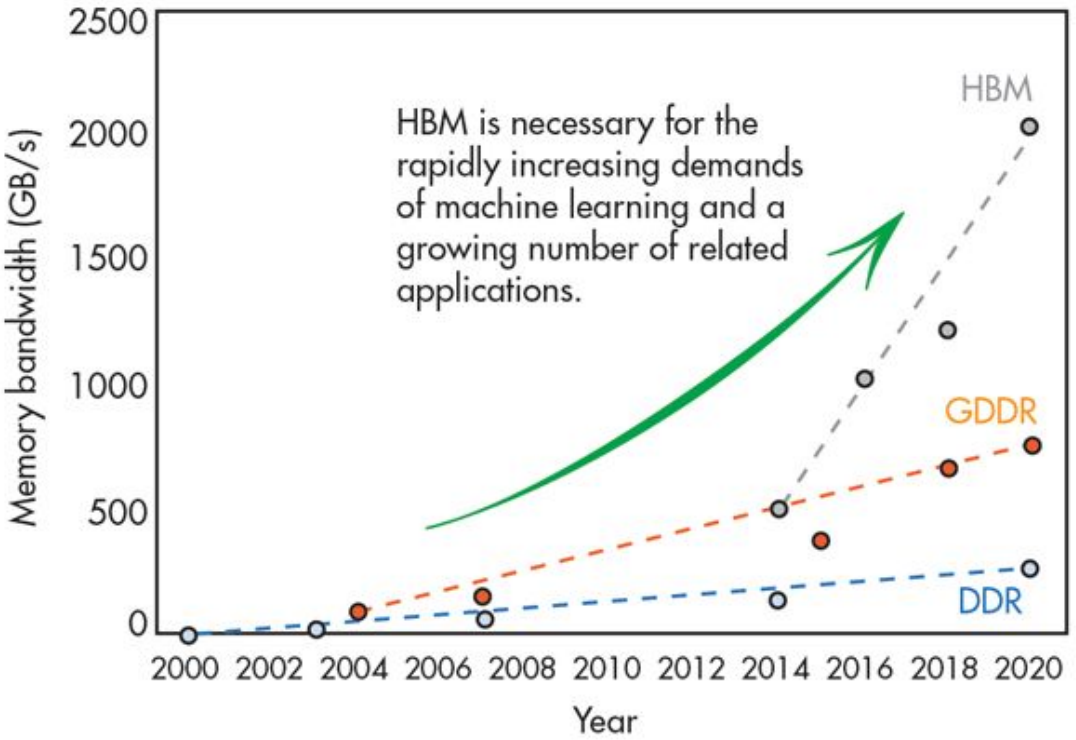


Area Savings



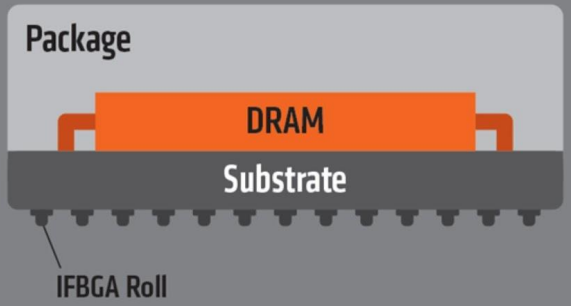
Source: https://www.amd.com/system/files/49010-high-bandwidth-memory-hbm-1260x709.jpg

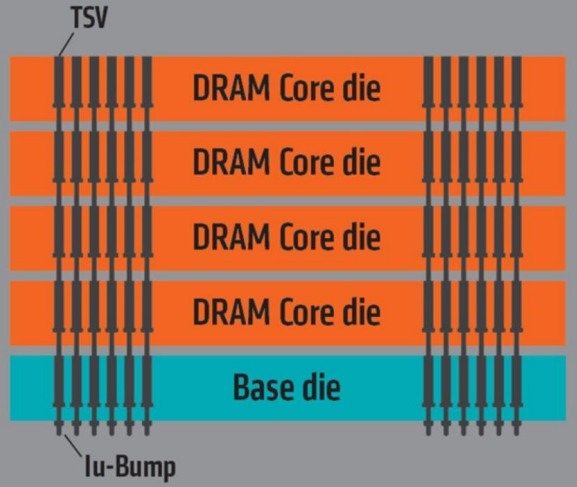
Bandwidth Improvement in HBM



Source: https://www.electronicdesign.com/industrial-automation/high-bandwidth-memory-great-awakening-ai

Comparison - GDDR5 Vs HBM





GDDR5	Per Package	HBM
32-bit	Bus Width	1024-bit
Up to 1750MHz (7GBps)	Clock Speed	Up to 500MHz (1GBps)
Up to 28GB/s per chip	Bandwidth	>100GB/s per stack
1.5V	Voltage	1.3V

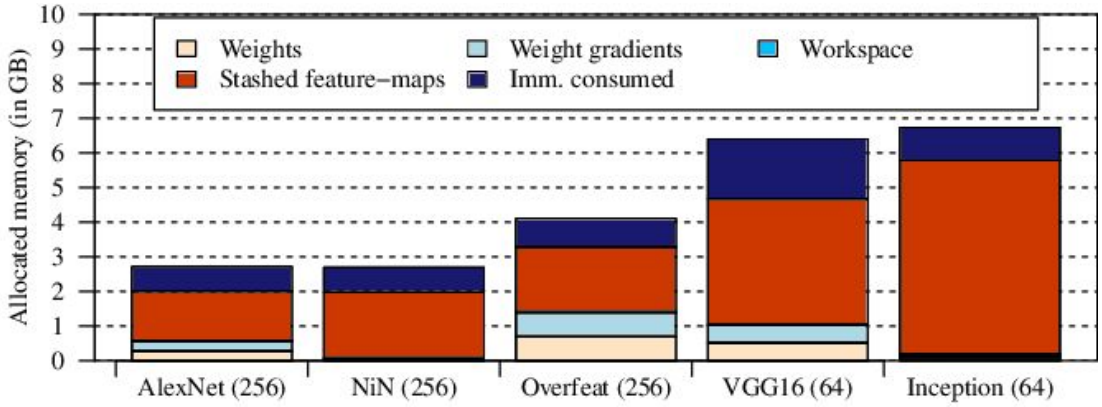
Source: https://www.amd.com/system/files/49010-high-bandwidth-memory-hbm-1260x709.jpg

Comparison - GDDR5 Vs HBM

Memory	GDDR5	GDDR5X	HBM	HBM2
Manufacturer	Samsung, Hynix, Elpida	Micron	Hynix, Samsung	Samsung, Hynix
Appearance	Square / Rectangular Chip	Square / Rectangular Chip	Cube / Cuboid	Cube / Cuboid
Maximum Capacity	8GB per Die	16GB per Die	1GB per Stack	4GB / 8GB per Stack
Maximum Speed	8 Gbps	10 to 14 Gbps (16 Gbps in future)	1 Gbps	2.4 Gbps
Bus Width	32-bit per chip	64-bit per chip	1024-bit per stack	1024-bit per stack or more
Power Consumption	Low	Same / Lower than GDDR5	Lower than GDDR5 and GDDR5X	Lower than HBM
Graphics Cards Used in	Many Graphics Cards from budget, mid-range to high-end e.g. GT 740, GTX 1070, RX 480 etc.	GeForce GTX 1080, GTX 1080 Ti, GTX 1060, Nvidia Titan X (Pascal)	Radeon R9 Fury X, Radeon Pro Duo	Nvidia Tesla P100, Nvidia Quadro GP100, Radeon RX Vega 56, Radeon RX Vega 64, Nvidia Titan V, AMD Radeon VII

HBM's for AI

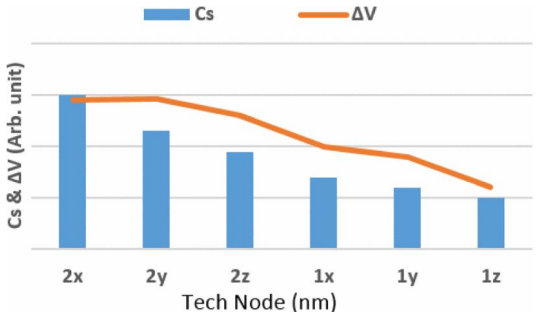
- High Bandwidth
- High Capacity
- Large Data Transfer Rate



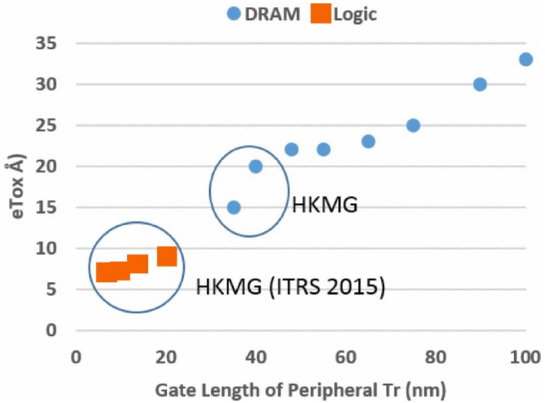
Source:A. Jain, A. Phanishayee, J. Mars, L. Tang and G. Pekhimenko, "Gist: Efficient Data Encoding for Deep Neural Network Training," 2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA), Los Angeles, CA, 2018, pp. 776-789.

Challenges

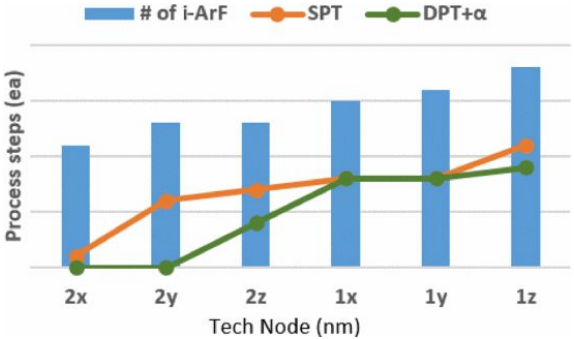
Retention Time and Sense Margin



Peripheral Transistor Performance



Chip Cost Issue and Evolutionary Approaches



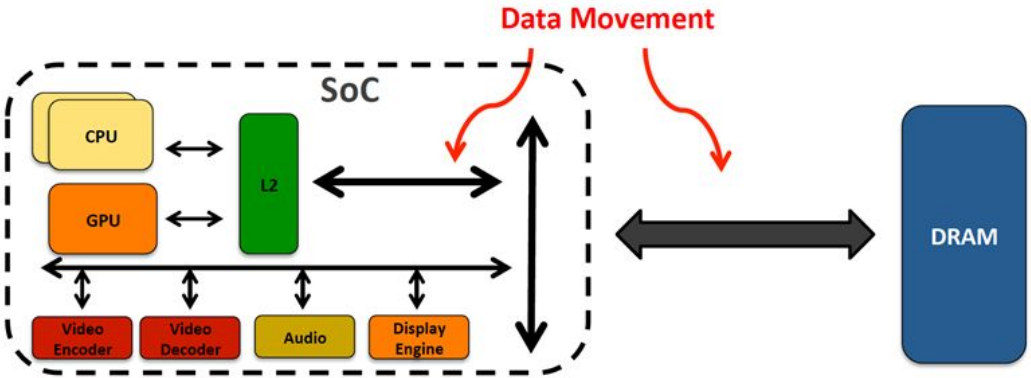
Limited Capacity

Source: S. Lee, "Technology scaling challenges and opportunities of memory devices," 2016 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2016, pp. 1.1.1-1.1.8.

In Memory Computing

Data Movement vs. Computation Energy

- Data movement is a major system energy bottleneck
  - Comprises **41% of mobile system energy** during web browsing [2]
  - Costs **~115 times** as much energy as an ADD operation [1, 2]



[1]: Reducing data Movement Energy via Online Data Clustering and Encoding (MICRO'16)
[2]: Quantifying the energy cost of data movement for emerging smart phone workloads on mobile platforms (IISWC'14)

We Need A Paradigm Shift To …

- Enable computation **with minimal data movement**
- Compute **where it makes sense** (where data resides)
- Make computing architectures more **data-centric**

Two Examples

Data Copy

RowClone: Fast and Energy-Efficient In-DRAM Bulk Data Copy and Initialization

Vivek Seshadri vseshadr@cs.cmu.edu
Yoongu Kim yoongukim@cmu.edu
Chris Fallin* cfallin@c1f.net
Donghyuk Lee donghyuk1@cmu.edu
Rachata Ausavarungnirun rachata@cmu.edu
Gennady Pekhimenko gpekhime@cs.cmu.edu
Yixin Luo yixinluo@andrew.cmu.edu
Onur Mutlu onur@cmu.edu
Phillip B. Gibbons† phillip.b.gibbons@intel.com
Michael A. Kozuch† michael.a.kozuch@intel.com
Todd C. Mowry tcm@cs.cmu.edu
Carnegie Mellon University †Intel Pittsburgh

Bitwise Computation

Ambit: In-Memory Accelerator for Bulk Bitwise Operations Using Commodity DRAM Technology

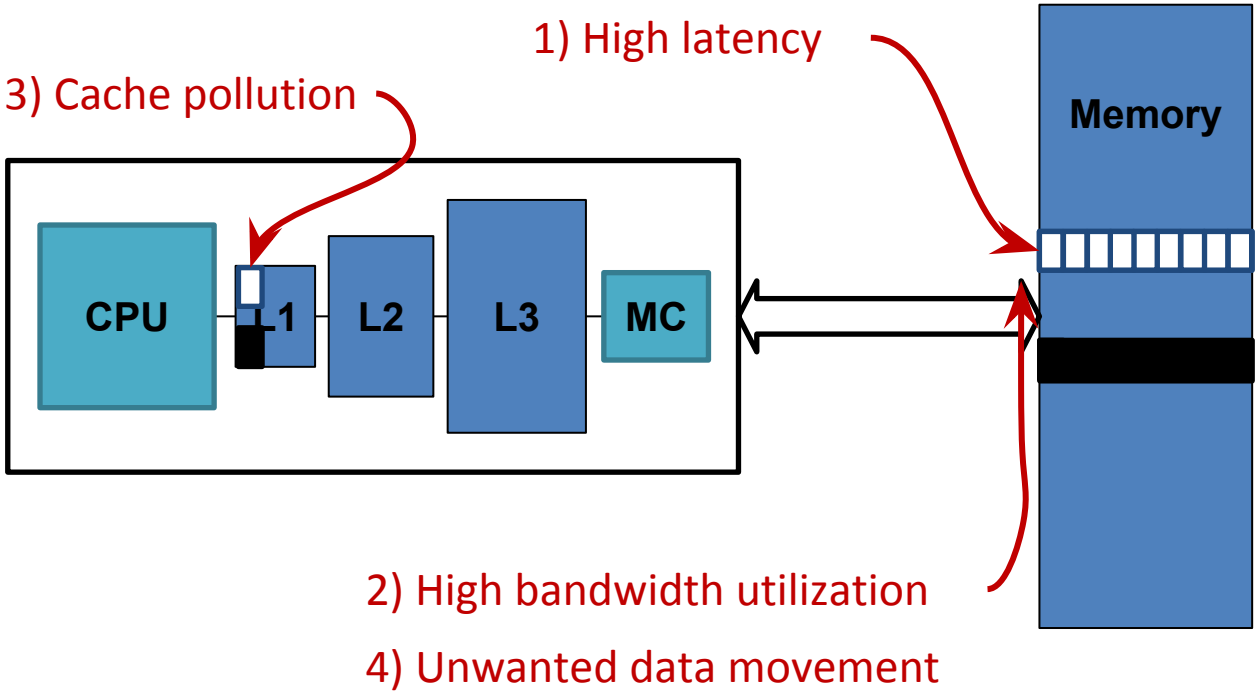
Vivek Seshadri[1,5] Donghyuk Lee[2,5] Thomas Mullins[3,5] Hasan Hassan[4] Amirali Boroumand[5]
Jeremie Kim[4,5] Michael A. Kozuch[3] Onur Mutlu[4,5] Phillip B. Gibbons[5] Todd C. Mowry[5]

[1]Microsoft Research India [2]NVIDIA Research [3]Intel [4]ETH Zürich [5]Carnegie Mellon University

RowClone: Fast and energy-efficient in-DRAM bulk data copy and initialization

Seshadri et al., "RowClone: Fast and Efficient In-DRAM Copy and Initialization of Bulk Data," MICRO 2013.

Today's Systems: Bulk Data Copy

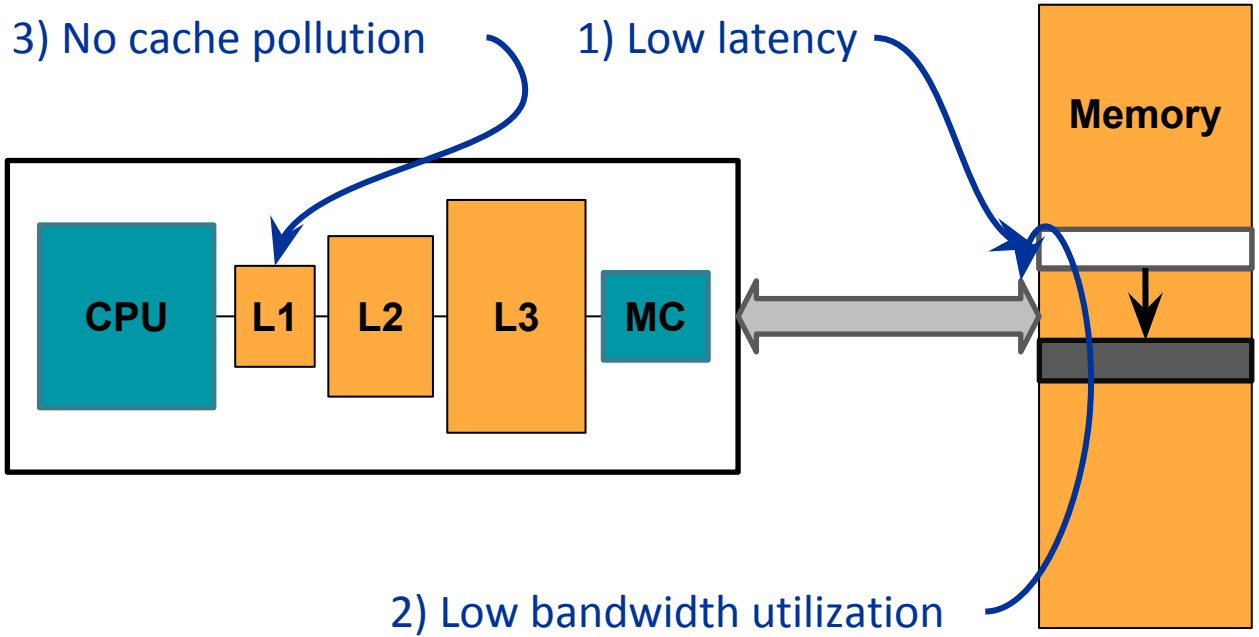


1046ns, 3.6uJ (for 4KB page copy via DMA)

Seshadri et al., "RowClone: Fast and Efficient In-DRAM Copy and Initialization of Bulk Data," MICRO 2013.

Source: Mutlu, O. (2019). Processing Data Where It Makes Sense in Modern Computing Systems: Enabling In-Memory Computation.

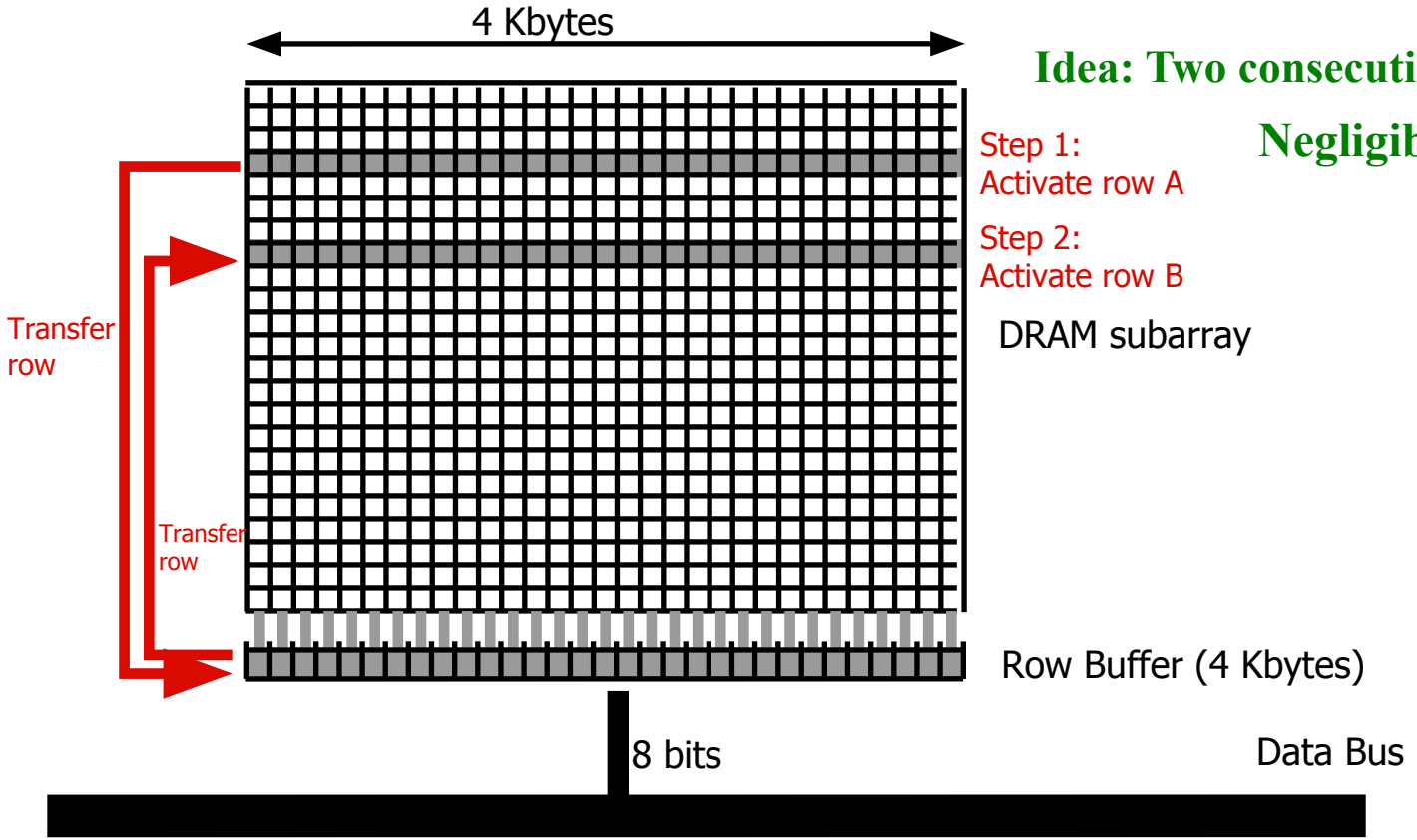
Future Systems: In-Memory Copy



1046ns, 3.6uJ 90ns, 0.04uJ

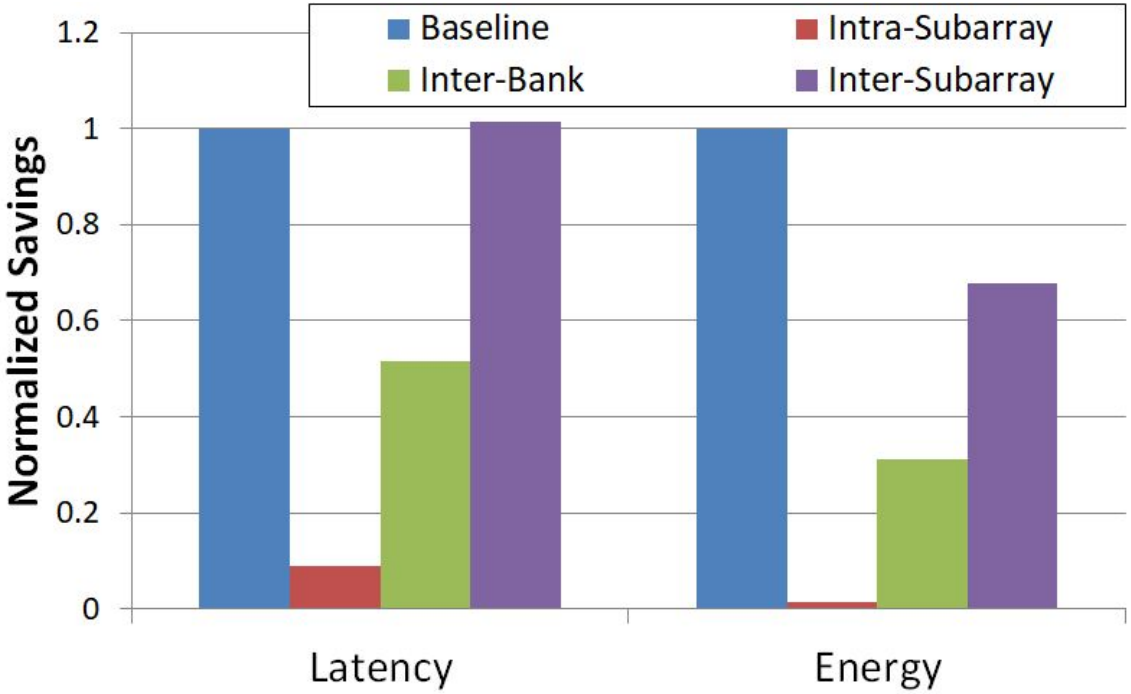
Source:Mutlu, O. (2019). Processing Data Where It Makes Sense in Modern Computing Systems: Enabling In-Memory Computation.

RowClone: In-DRAM Row Copy



Source: Mutlu, O. (2019). Processing Data Where It Makes Sense in Modern Computing Systems: Enabling In-Memory Computation.

RowClone: Latency and Energy Savings

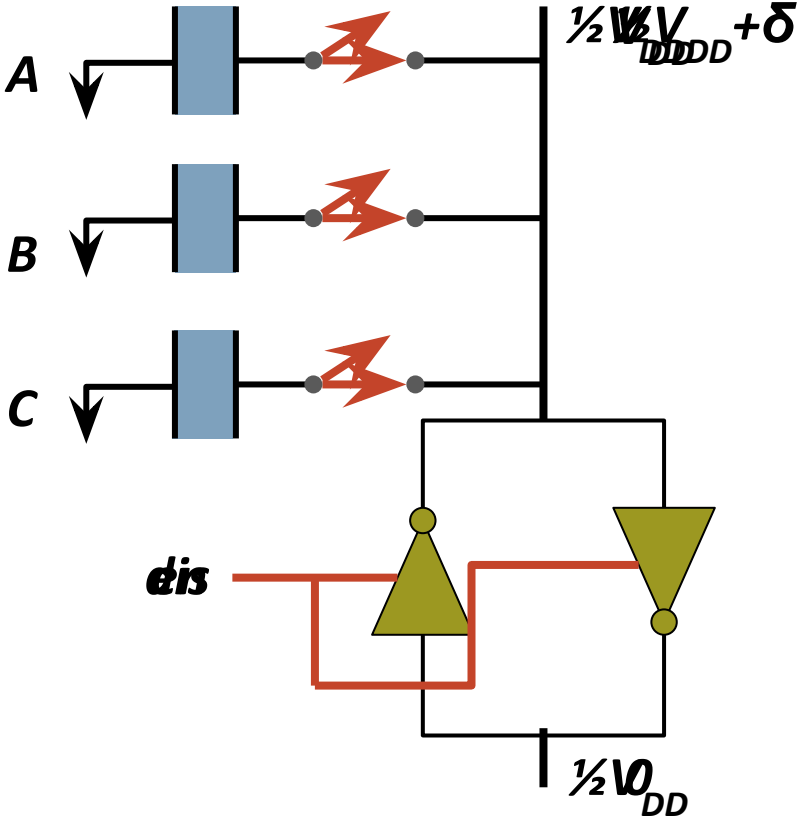


Source: Seshadri et al., "RowClone: Fast and Efficient In-DRAM Copy and Initialization of Bulk Data," MICRO 2013.

Ambit: In-Memory Accelerator for Bulk Bitwise Operations using Commodity DRAM Technology

Seshadri+, "Ambit: In-Memory Accelerator for Bulk Bitwise Operations using Commodity DRAM Technology," MICRO 2017.

In-DRAM AND/OR: Triple Row Activation

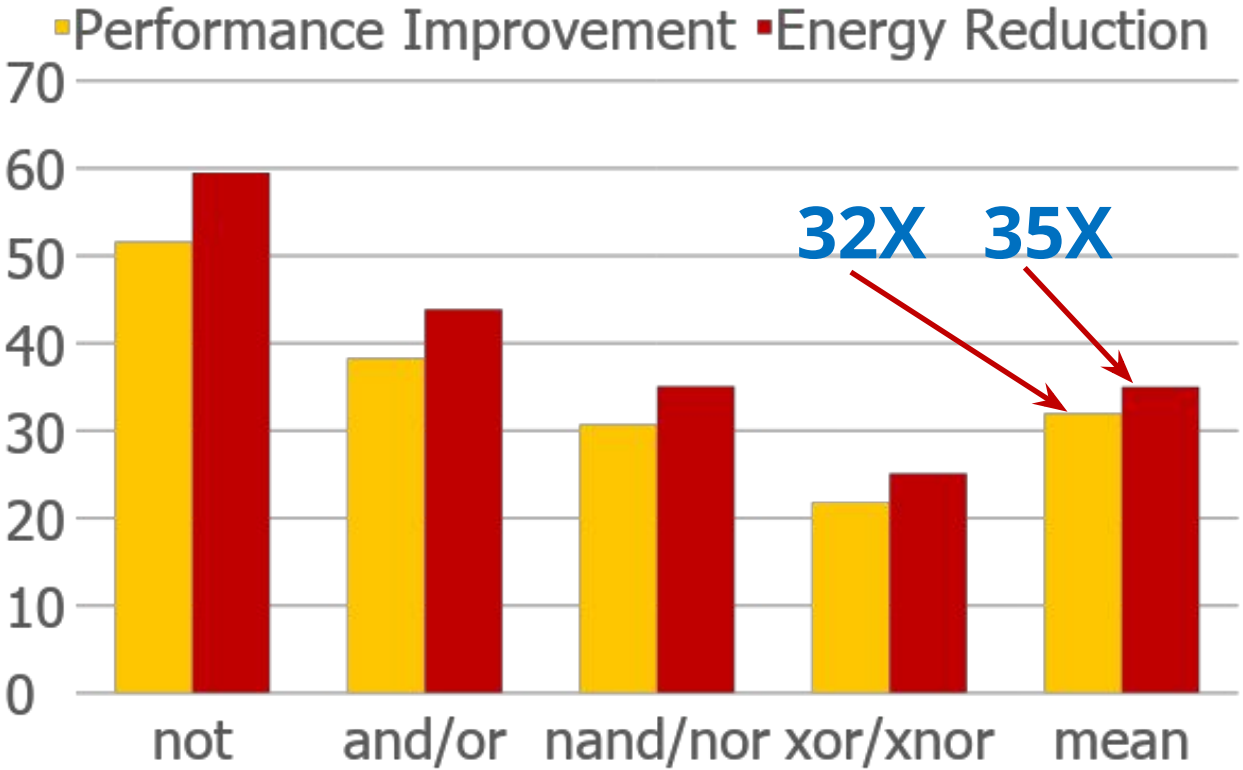


Final State
AB + BC + AC

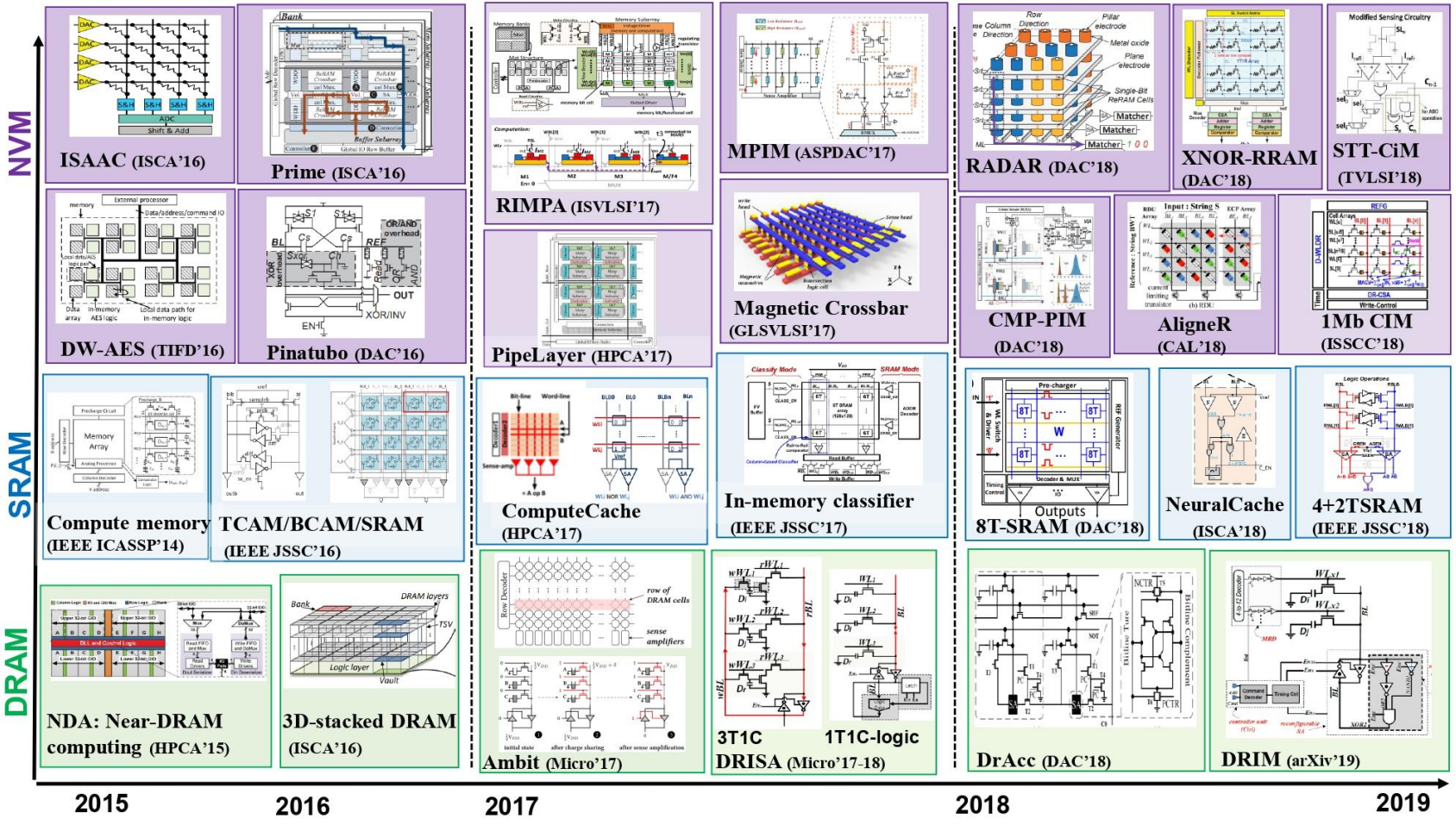
C(A + B) + ~C(AB)

Source: Seshadri+, "Fast Bulk Bitwise AND and OR in DRAM", IEEE CAL 2015.

Ambit vs. DDR3: Performance and Energy

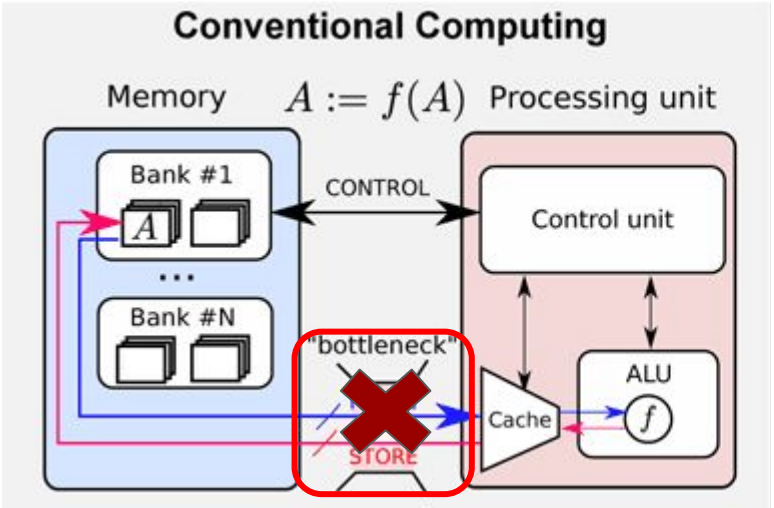


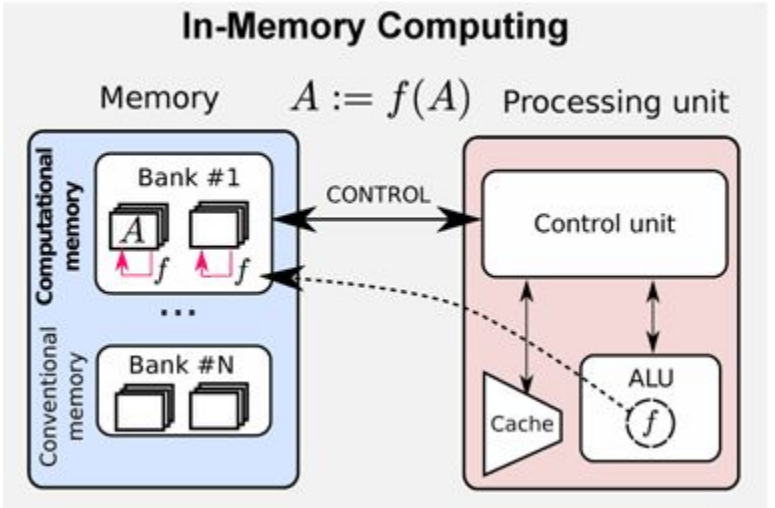
Source: Seshadri+, "Ambit: In-Memory Accelerator for Bulk Bitwise Operations using Commodity DRAM Technology," MICRO 2017.



Source: https://dfan.engineering.asu.edu/in-memory-computing/

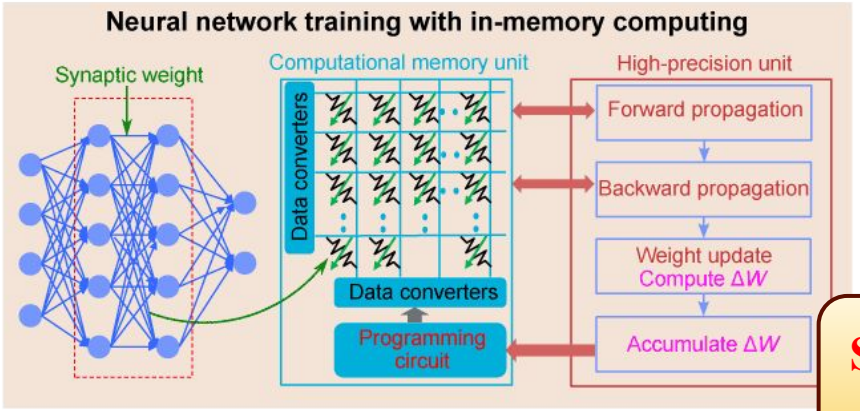
In Memory Computing for AI

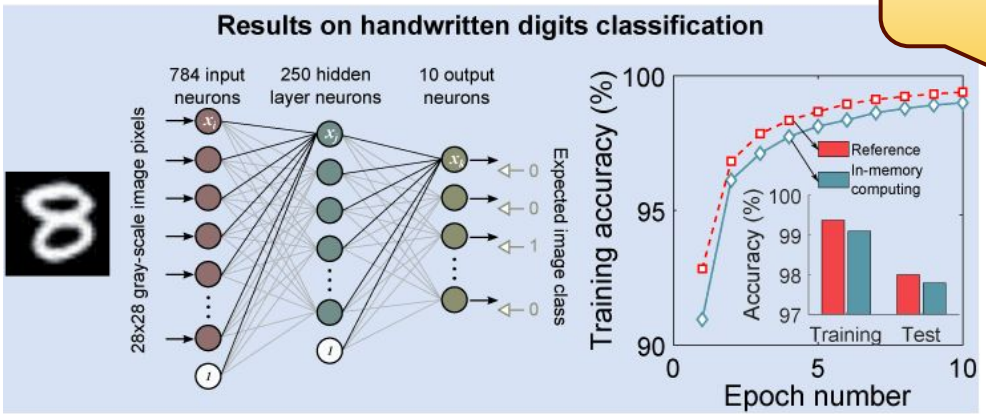




Source: A. Sebastian et al.: "Temporal correlation detection using computational phase-change memory", Nature Communications 8, 1115, 2017.

In Memory Computing for AI





Slight Accuracy Drop

Source: S.R. Nandakumar et al.: "Mixed-precision architecture based on computational memory for training deep neural networks", in Proc. of the IEEE International Symposium on Circuits and Systems (ISCAS), 1-5, 2018.

Challenges - Software and Architecture

- **Functionality of and applications & software** for In Memory Computing
- **Ease of programming** (interfaces and compiler/HW support)
- **System support:** coherence & virtual memory
- **Runtime and compilation systems** for adaptive scheduling,
- **Data mapping**, **access/sharing control**
- **Infrastructures** to assess benefits and feasibility

Challenges - Circuits and Devices

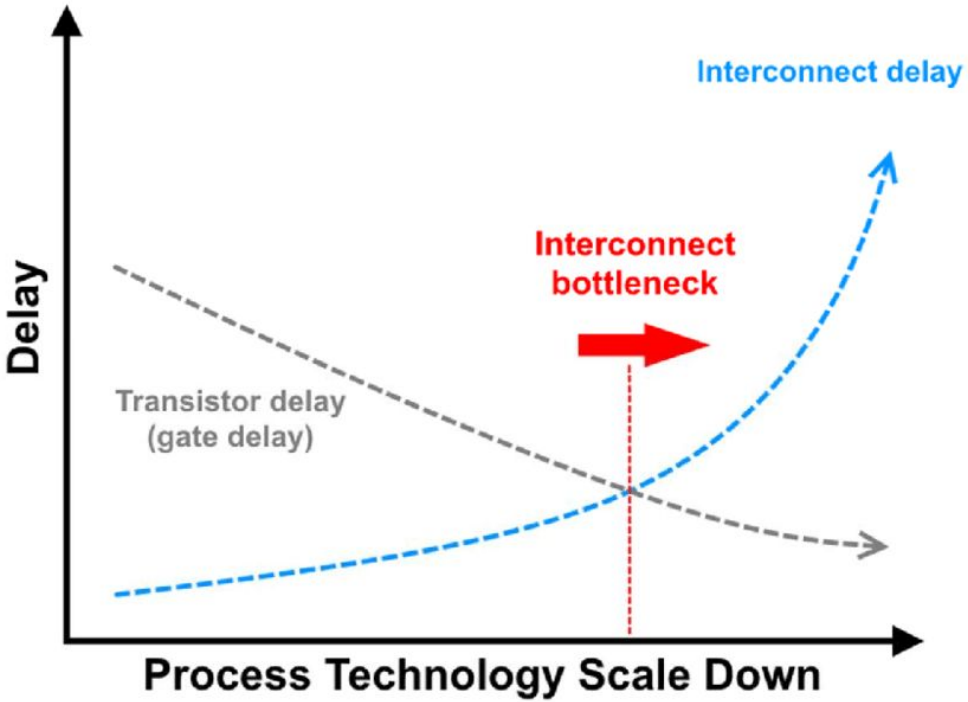
- **Process Variations**
  - Not all Memory cells are **created equal**
  - Some cells have **higher /lower capacitance**
    - affects the reliability of operation
- **Processing Using Memory**
  - Exploits **analog operation** of underlying technology
  - Process variation can introduce failures
- **Challenges**
  - How to **design architecture to reduce impact of variations**
  - How to **test for failures**
- **Intelligent Memory Controllers**
  - **ECC Circuits**

HBM or In Memory Computing?

Both.

Better Way: In Memory Compute in HBM

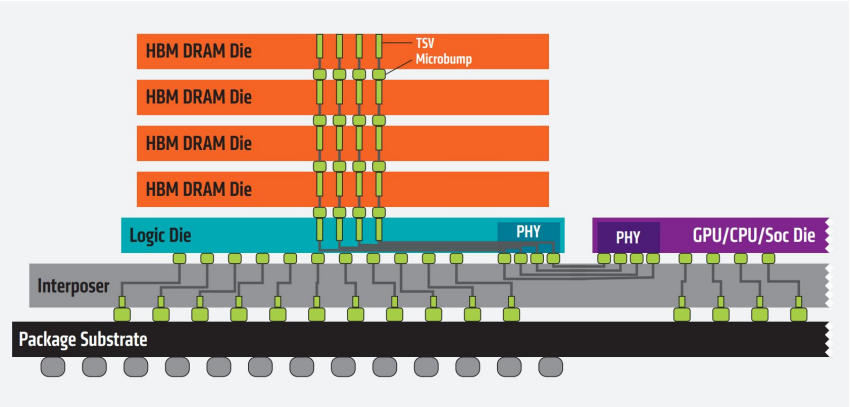
Why?



Source: Grzela, Tomasz. (2015). Comparative STM-based study of thermal evolution of Co and Ni germanide nanostructures on Ge(001).

Why?

- Interconnects
  - **Large Power consumption**
    - Capacitance
    - Metal lines
- **Large Interconnect Delay** compared to Gate delay
- **Logic Layer** beneath DRAM dies
  - Support In Memory Computing



Conclusion

In Memory Compute in HBM.

Some Recent works on IMC in HBM

A 3D-Stacked Logic-in-Memory Accelerator for Application-Specific Data Intensive Computing

Qiuling Zhu, Berkin Akin, H. Ekin Sumbul, Fazle Sadi, James C. Hoe, Larry Pileggi, Franz Franchetti
Dept. of Electrical and Comp. Eng., Carnegie Mellon University, Pittsburgh, PA, USA
Email: {qiulingz,bakin,hsumbul,fsadi,jhoe,pileggi,franzf}@ece.cmu.edu

Session 10: Memory Architectures GLSVLSI'18, May 23-25, 2018, Chicago, IL, USA

A Compiler for Automatic Selection of Suitable Processing-in-Memory Instructions

Hameeza Ahmed, Paulo C. Santos†, João P. C. Lima‡, Rafael F. Moura†,
Marco A. Z. Alves‡, Antônio C. S. Beck†, Luigi Carro†
Dep. of Computer and Information Systems Eng. – NED University – Karachi, Pakistan
†Informatics Institute – Federal University of Rio Grande do Sul – Porto Alegre, Brazil
‡Department of Informatics – Federal University of Paraná – Curitiba, Brazil
Email: hameeza@neduet.edu.pk †{pcssjunior, jplima, rfmoura, caco, carro}@inf.ufrgs.br ‡{mazalves}@inf.ufpr.br

Towards Near-Data Processing of Compare Operations in 3D-Stacked Memory

Palash Das and Hemangee K. Kapoor
Department of CSE, IIT Guwahati, Guwahati, Assam 781039, India
{palash.das, hemangee}@iitg.ernet.in

IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 30, NO. 3, MARCH 2019 589

Exploiting Parallelism for CNN Applications on 3D Stacked Processing-In-Memory Architecture

Yi Wang, *Member, IEEE*, Weixuan Chen, Jing Yang, and Tao Li, *Fellow, IEEE*

Design space exploration for PIM architectures in 3D-stacked memories

João Paulo C. de Lima
Fed. University of Rio Grande do Sul
Porto Alegre, RS, Brazil

Paulo Cesar Santos
Fed. University of Rio Grande do Sul
Porto Alegre, RS, Brazil

Marco A. Z. Alves
Fed. University of Paraná
Curitiba, PR, Brazil

Antonio C. S. Beck
Fed. University of Rio Grande do Sul
Porto Alegre, RS, Brazil

Luigi Carro
Fed. University of Rio Grande do Sul
Porto Alegre, RS, Brazil

Conclusion

- Von-Neumann Bottleneck
- Solutions
  - High Bandwidth Memory
  - In Memory Computing
- In Memory Computing in HBM's
  - Increase Bandwidth
  - Reduce Data Movement

References

1. Moore, Gordon E. "Cramming more components onto integrated circuits." (1965): 114-117.
2. Chu, Yaohan, ed. High-level language computer architecture. Academic Press, 2014.
3. Seshadri, Vivek, et al. "Ambit: In-memory accelerator for bulk bitwise operations using commodity DRAM technology." Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture. ACM, 2017.
4. Seshadri, Vivek, et al. "RowClone: fast and energy-efficient in-DRAM bulk data copy and initialization." Proceedings of the 46th Annual IEEE/ACM International Symposium on Microarchitecture. ACM, 2013.
5. Seshadri, Vivek, et al. "Fast bulk bitwise AND and OR in DRAM." IEEE Computer Architecture Letters 14.2 (2015): 127-131.
6. Wang, Shibo, and Engin Ipek. "Reducing data movement energy via online data clustering and encoding." The 49th Annual IEEE/ACM International Symposium on Microarchitecture. IEEE Press, 2016.
7. H. Jun et al., "HBM (High Bandwidth Memory) DRAM Technology and Architecture," 2017 IEEE International Memory Workshop (IMW), Monterey, CA, 2017, pp. 1-4.
8. Mohsen Imani, Saransh Gupta, Yeseong Kim, and Tajana Rosing. 2019. FloatPIM: in-memory acceleration of deep neural network training with high precision. In Proceedings of the 46th International Symposium on Computer Architecture (ISCA '19). ACM, New York, NY, USA, 802-815.
9. S. Lee, "Technology scaling challenges and opportunities of memory devices," 2016 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2016, pp. 1.1.1-1.1.8.

Thank You