# Heterogenous Integration for Artificial Intelligence

Saibal Mukhopadhyay, S. Yalamanchili and Madhavan Swaminathan

**Students:**
H. M. Torun
Y. Long
B. Mudassar
C. S. Nair
B. H. DeProspo
D. Kim

**Faculty:**

M. Kathaperumal
V. Smet

*Ref: S. Mukhopadhyay, Y. Long, B. Mudassar, C. S. Nair, B. H. DeProspo, H. M. Torun, M. Kathaperumal, V. Smet, D. Kim, S. Yalamanchili and M. Swaminathan, "Heterogonous Integration for Artificial Intelligence: Challenges and Opportunities", IBM Journal of R&D, 2019.*

# Outline

❑Bandwidth Limitation of Neural Networks

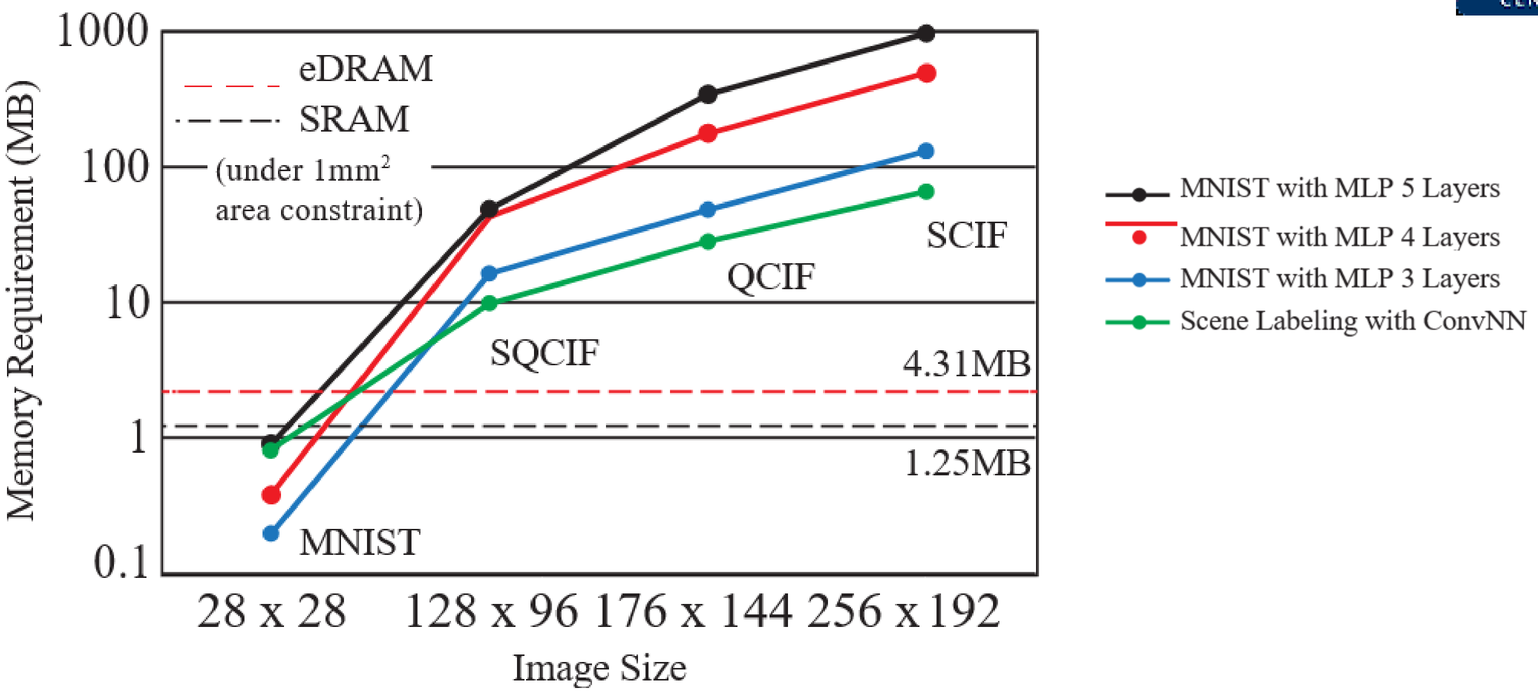
❑High Bandwidth Memory with Silicon and Glass Interposer

- ❑Electrical Analysis and Optimization

❑Architectural Alternatives to HBM for Neural Networks

❑Summary

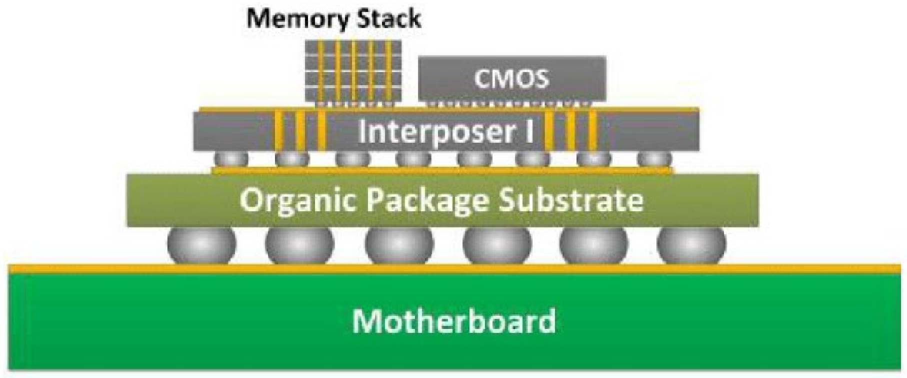
# Bandwidth Limitation of NNs



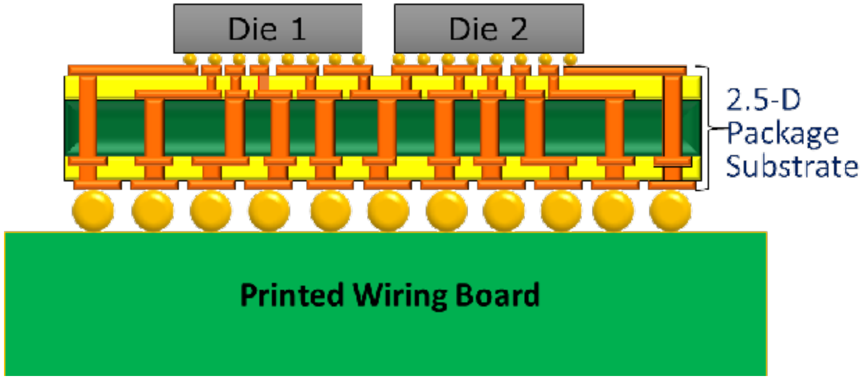
- Neural networks can be processed highly parallel.
- Overall performance is generally limited by insufficient memory bandwidth and latency.
- On-chip memory bandwidth is not enough for processing large image files.
- The burden is on off-chip bandwidth between logic and memory.

# Glass vs Silicon 2.5D Interposer: A Comparison

<u>2.5D Silicon Interposer Package</u>



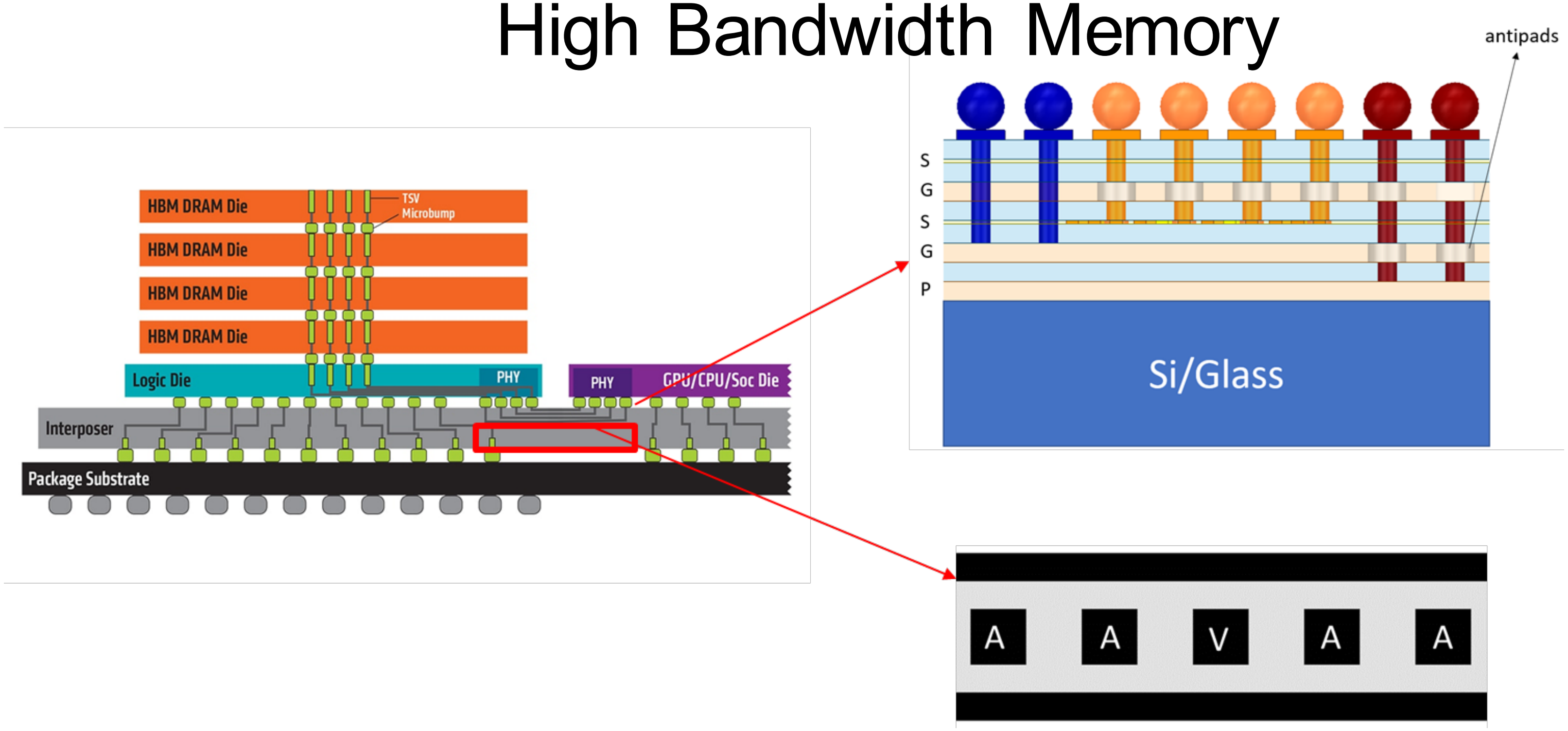
<u>2.5D Glass Interposer Package</u>



- ❑ Commonly used integration technology for AI applications is High Bandwidth Memory (HBM)
  - ❑ Heterogenous integration of GPU and memory dies on an interposer
- ❑ RDL layers interconnect a large GPU die (20 mm x 20 mm) to multiple HBMs (7mm x 5mm)
- ❑ We compare silicon and glass interposer technologies for such scheme.
- ❑ Glass interposer has several advantages:
  - ❑ Low Dk polymers instead of $SiO_2$.
  - ❑ High aspect ratio lines.
  - ❑ Eliminates the need for organic package.

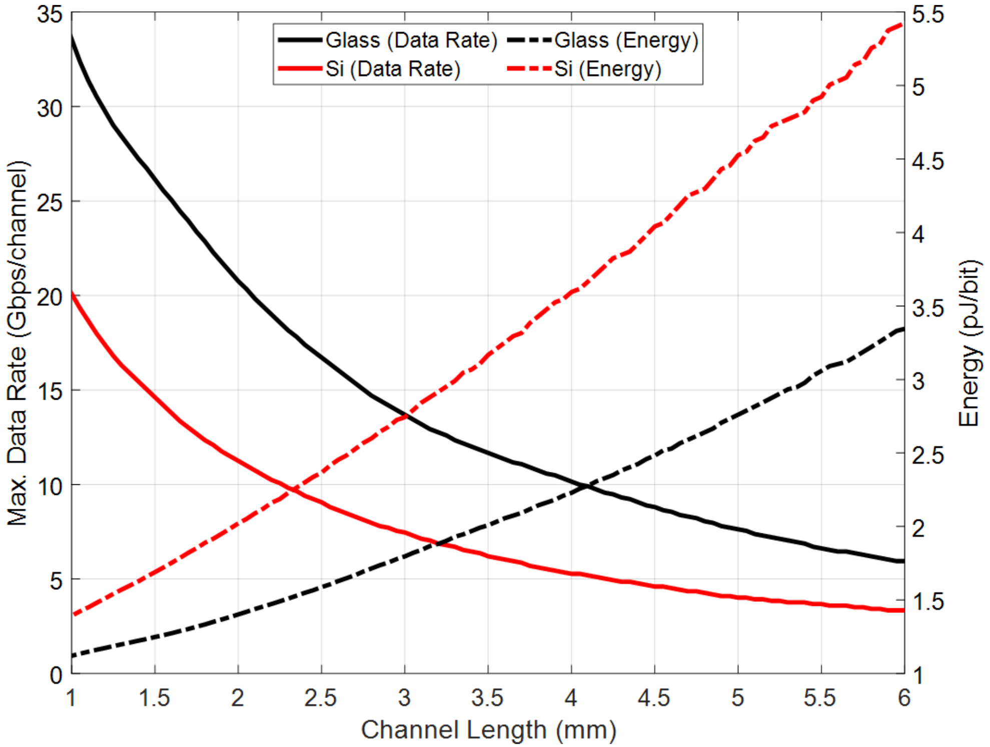
*Ref: Heterogeneous Integration for Artificial Intelligence: Challenges and Opportunities, IBM Journal of Research & Development, 2019*

# High Bandwidth Memory



- A signaling channel between GPU and memory.
- Interconnects lengths are between 1-6 mm.
- Current technologies allow for 55 $\mu m$ bump pitch.
  - RDL Density: 200 IOs/mm/layer to 400 IOs/mm/layer
- Intermetal dielectric material used in simulations:
  - For Si interposer: $SiO_2$ ($\varepsilon_r = 3.9$)
  - For Glass interposer: Low Dk Polymer ($\varepsilon_r = 2.4$)
- Commercially used data rate per signal line for silicon interposer: 2 Gbps
  - Can we go beyond 2 Gbps per line?

# Glass vs Silicon for HBM

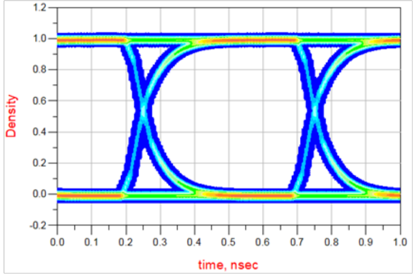


- ❑ Interconnects on both silicon and glass interposer have W/S/AR = 2μm/2μm/1
- ❑ For an eye height of 800 mV for 6 mm channel, maximum achievable data rates are:
  - ❑ Si interposer: 3.2 Gbps
  - ❑ Glass interposer: 6 Gbps
- ❑ Aggregate total bandwidth: 1.63 TB/s and 3.07 TB/s for silicon and glass interposer
  - ❑ 4 memory stacks, each with 8 channels that contain 128-bit data interface
- ❑ Glass: 3.3 pJ/bit per transmitter @ 6 Gbps; Si: 5.4 pJ/bit per transmitter @ 3.2 Gbps

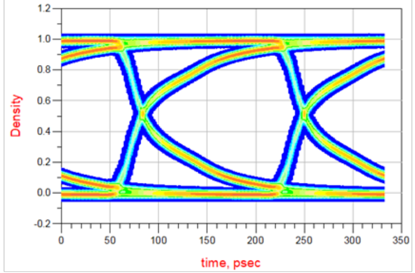
# Glass vs Silicon for HBM

2 Gbps | 6 Gbps | 10 Gbps | 16 Gbps

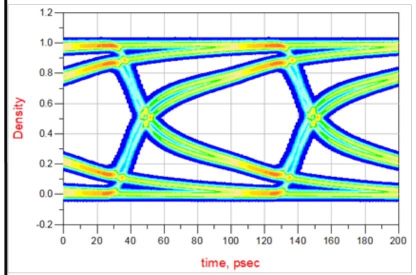
Glass



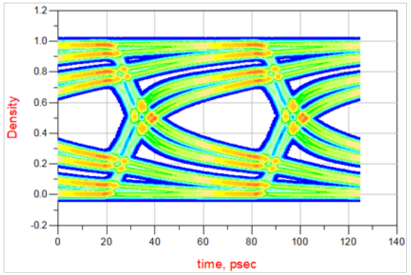
2 Gbps: Eye Width = 97 %UI
Eye Height = 0.91 V



6 Gbps: Eye Width = 93.5 %UI
Eye Height = 0.78 V

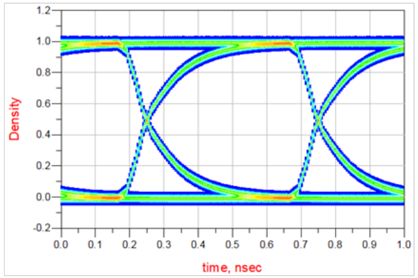


10 Gbps: Eye Width = 85.5 %UI
Eye Height = 0.59 V

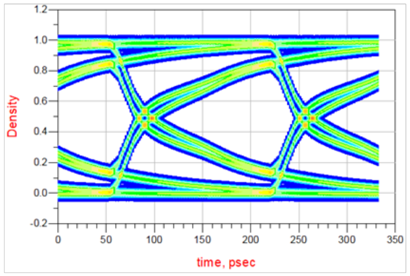


16 Gbps: Eye Width = 73.5 %UI
Eye Height = 0.38 V

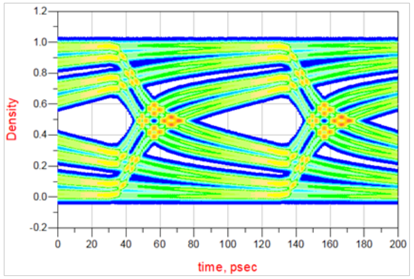
Silicon



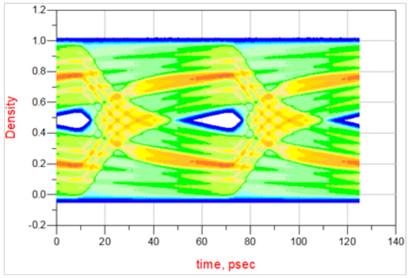
2 Gbps: Eye Width = 95.5 %UI
Eye Height = 0.89 V



6 Gbps: Eye Width = 77.5 %UI
Eye Height = 0.55 V



10 Gbps: Eye Width = 55.5 %UI
Eye Height = 0.25 V



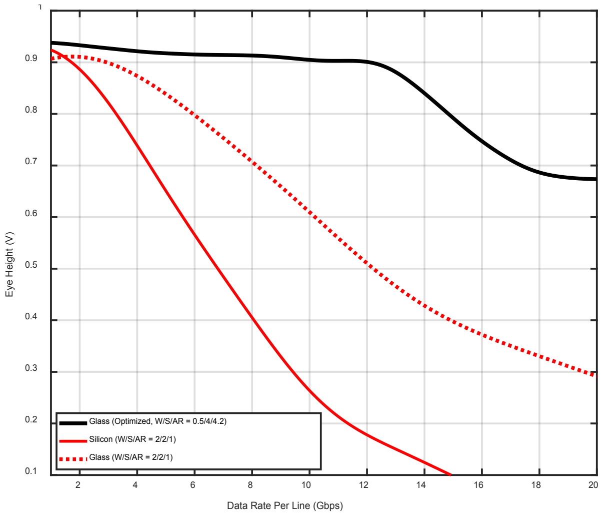
16 Gbps: Eye Width = 24.5 %UI
Eye Height = 0.07 V

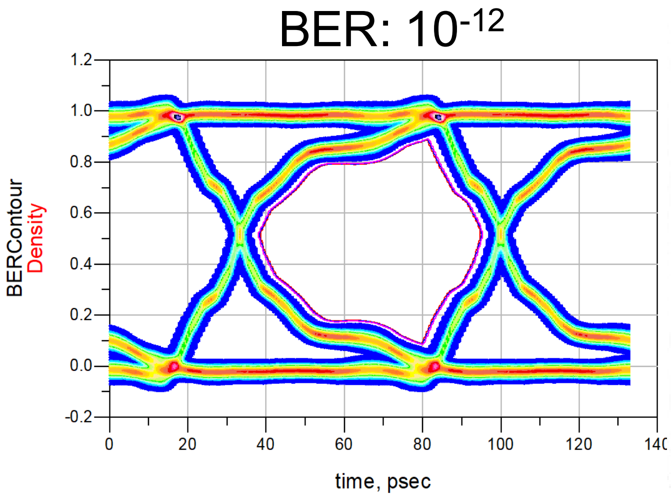
Simulation settings:

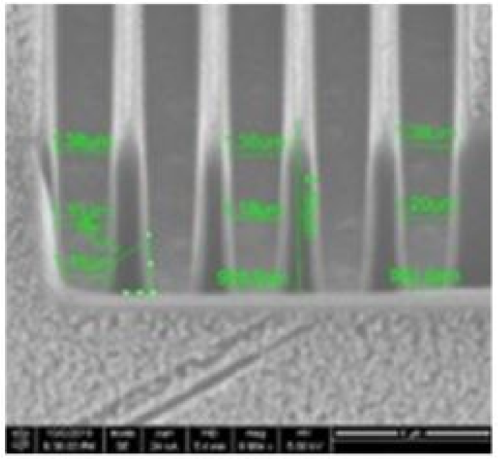
- Statistical simulation at BER = 1E-12
- Rise/Fall Time = 10 %UI
- TX Impedance = 50 Ω
- RX termination at $\mu$-bumps
- Channel length = 6 mm
- Bump Pitch/Diameter = 55/20 um
- Line Width/Spacing/Thickness = 2 um

Low Dk polymers used in glass interposer enables significant signal integrity improvement!

# Optimizing Aspect Ratio using Machine Learning for Glass Interposer



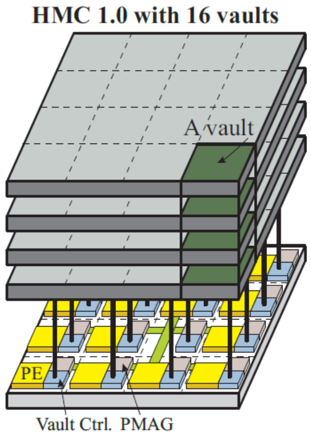


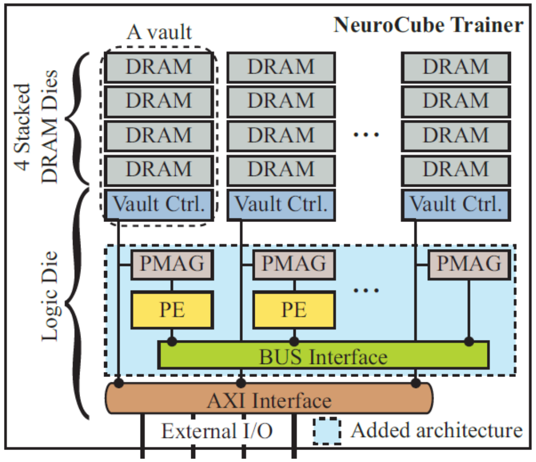


- ❑ Fabrication of high aspect ratio (AR) lines enables flexibility for interconnect design.
- ❑ Multiple trade-offs must be considered to determine the optimal line characteristics
  - ❑ High AR decreases R, but increases L & C as well as mutual C.
- ❑ We use ML based optimization to determine optimal trade-off.
  - ❑ Under the constraint that routing density is fixed to 333 signal lines per layer.
- ❑ Optimized interconnects on glass:
  - ❑ 15 Gbps per signal line (8.19 TB/s total bandwidth) at 2.6 pJ/bit.
  - ❑ Interconnect geometry → W/S/AR: 0.5/4/4.2

# Neurocube using Hybrid Memory Cube Architecture

D. Kim et. Al, ISCA 2016

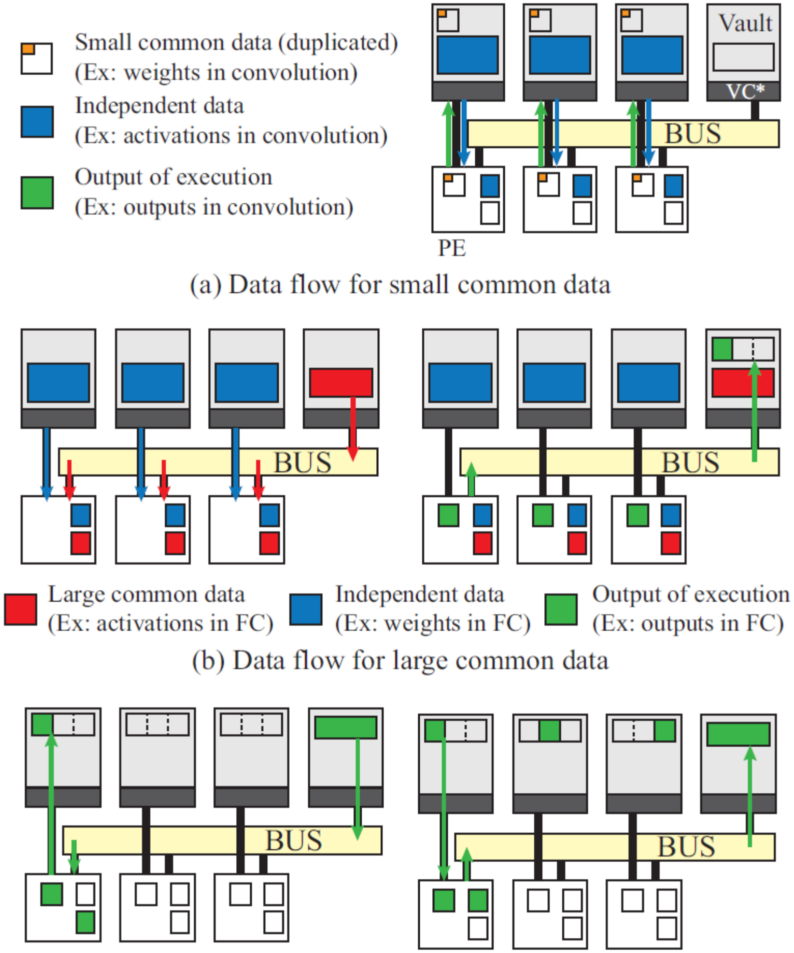




| | DDR3 [33] | Wide I/O 2 [34] | HBM [35] | HMC-Ext [36] | HMC-Int [36] |
|---|---|---|---|---|---|
| Interface | 2D | 3D | 2.5D | 3D | 3D |
| Max. # Channels | 2 | 8 | 8 | 8 | 16 |
| Word size | 64 bit | 128 bit | 128 bit | 32 bit | 32 bit |
| Peak B.W.† | 12.8 GBps | 6.4 GBps | 16 GBps | 40 GBps | 10 GBps |
| $t_{CL} + t_{RCD}$ | 25 ns | N/A | N/A | 27.5 ns [37] | 27.5 ns [37] |
| Operating Voltage | 1.5 V, 1.35 V [33] | 1.1 V [34] | 1.2 V [38] | 1.2 V [39] | 1.2 V [39] |
| Energy | 70 pJ/bit [40] | N/A | N/A | 10 pJ/bit [39] | 3.7 pJ/bit [39] |

- Micron's Hybrid Memory Cube (HMC): Stack DRAM dies & single base logic die through TSVs.
  - Enables parallel access to memory for high performance.
- Neurocube integrates a logic layer within the 3D high-density memory package of HMC.
  - Heterogeneous data flow architecture for different data types/sizes.
  - Logic and memory dies can be fabricated using different process technologies.

# Neurocube: Communication Arch. and Performance

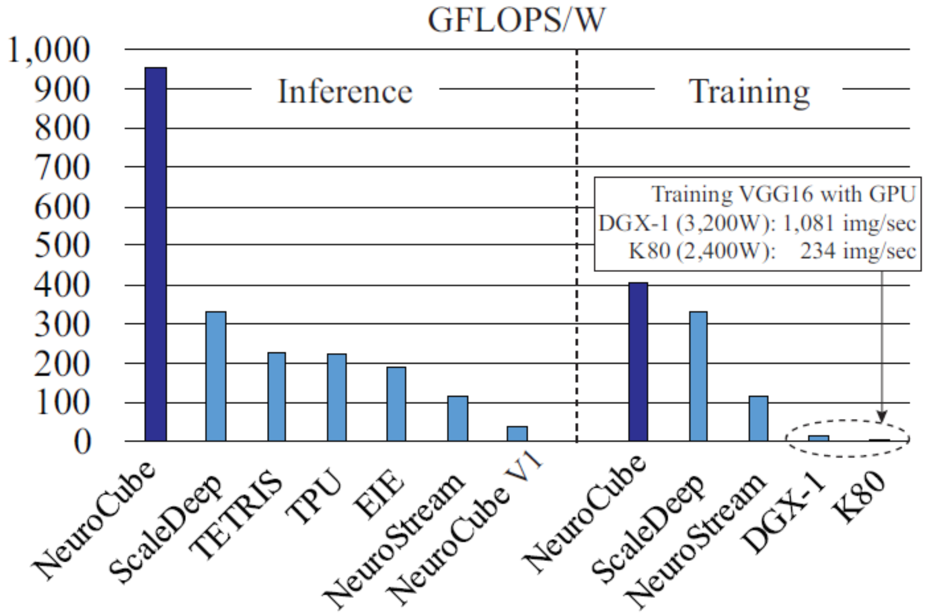


(a) Data flow for small common data

(b) Data flow for large common data

(c) Partitioning

(d) Merging



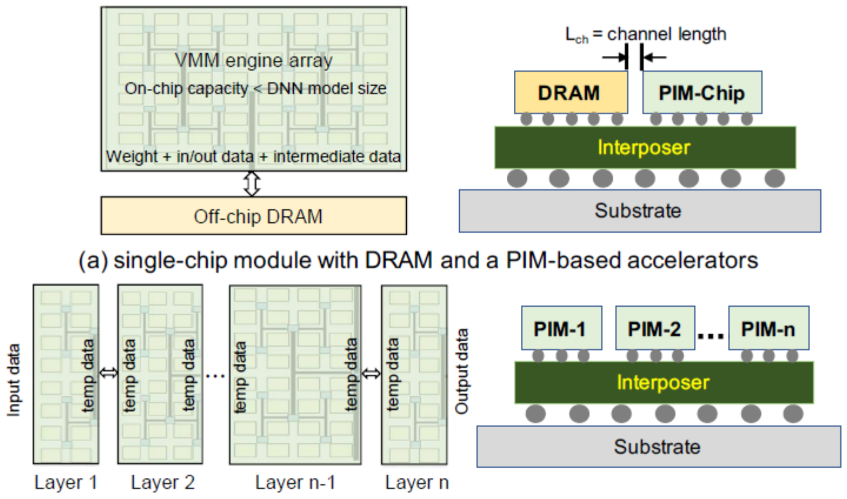
Training VGG16 with GPU
DGX-1 (3,200W): 1,081 img/sec
K80 (2,400W): 234 img/sec

- Heterogeneous data flow architecture of the Neurocube enables significantly improved throughput compared to GPUs.

D. Kim et. Al, IEEE TCAD 2018

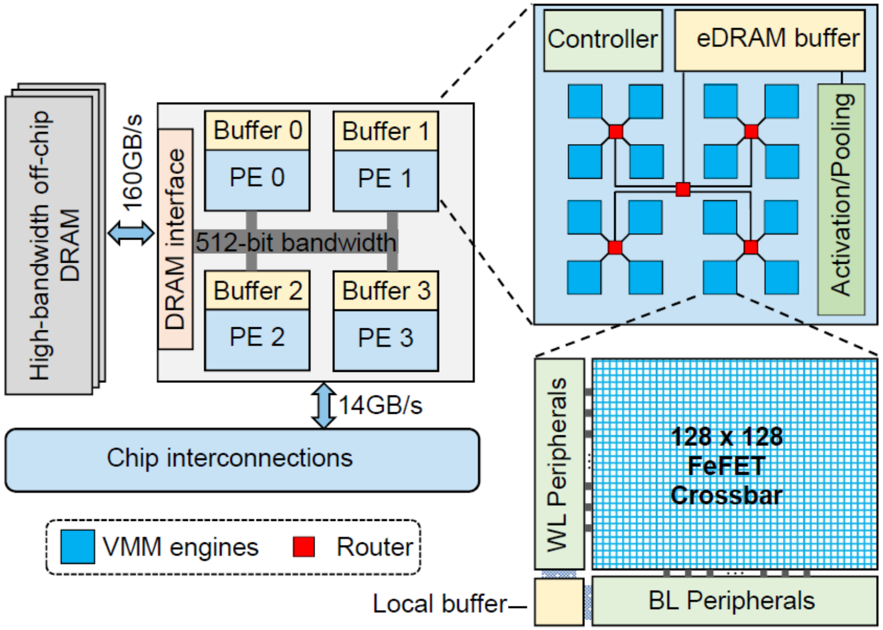
# Processing in Memory (PIM) based Accelerator

Y. Long, et. al, ICCAD 2019



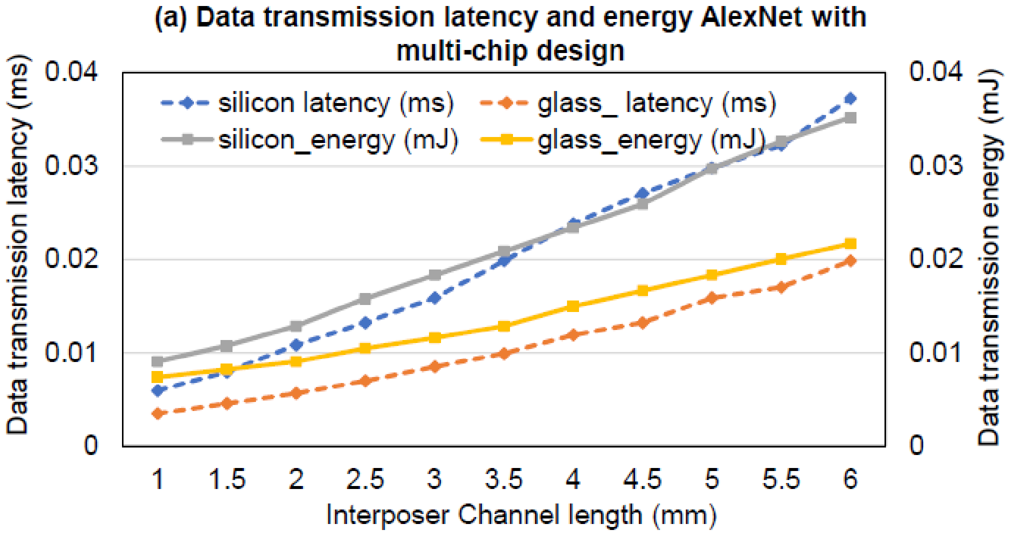
(a) single-chip module with DRAM and a PIM-based accelerators

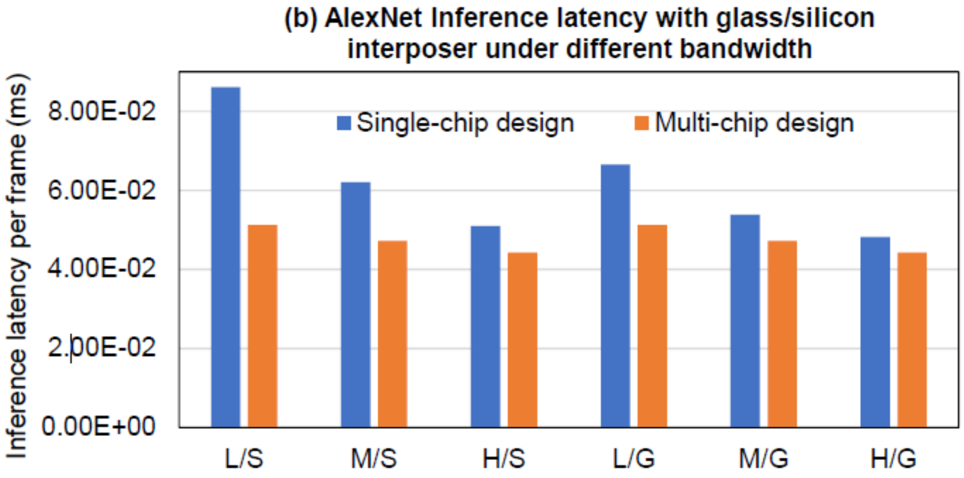
(b) multi-chip-module with PIM-based accelerators



- NMP is still bounded by logic-centric computation where data and logic are separated.
- More aggressive approach is referred to as PIM.
  - Direct computation inside memory.
- For very complex/deep networks, not feasible to map the whole network to on-chip memory.
- Multi-chip PIM architecture integrated on interposer is considered.
  - Each chip is responsible for computation for one-layer in a deep network.

# Case Study: PIM with Glass and Silicon





- Performance of multi-chip design greatly depends on interconnect characteristics.
- For AlexNet architecture and for varying channel lengths:
  - Si Interposer:

    1.6 TB/s – 10.2 TB/s throughput

    1.4 pJ – 5.4 pJ per bit energy
  - Glass Interposer:

    3.1 TB/s – 17.4 TB/s throughput

    1.1 pJ – 3.3 pJ per bit energy
- As the channel length increases, performance gain through multi-chip design increases.

# Summary

- Discussed potential of heterogeneous integration for next gen. energy-efficient AI/ML platforms.
- Packaging technology has a significant role on achievable throughputs.
- Glass interposer-based designs showed superior performance compared to Silicon for HBM.
  - ~2X higher bandwidth at ~2X reduced energy per bit.
- Near memory processing architectures shows great potential for energy-efficiency.
  - Heterogeneous integration of CMOS and non-volatile memory within a 3D stack.
- Processing-in-memory architectures can provide orders of magnitudes gains in efficiency.
- Future research is required to transform the potential of heterogeneous integration for real-systems.