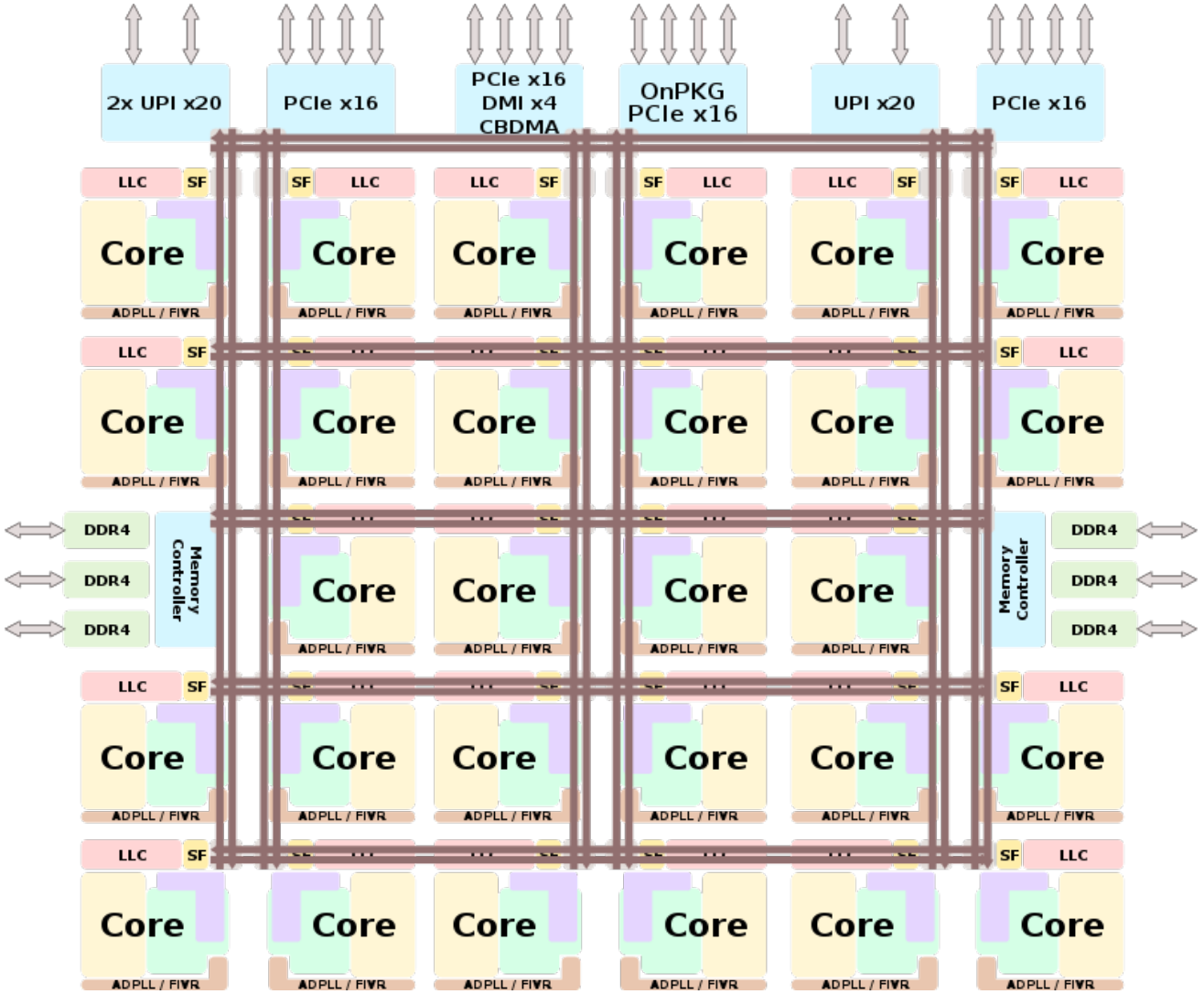
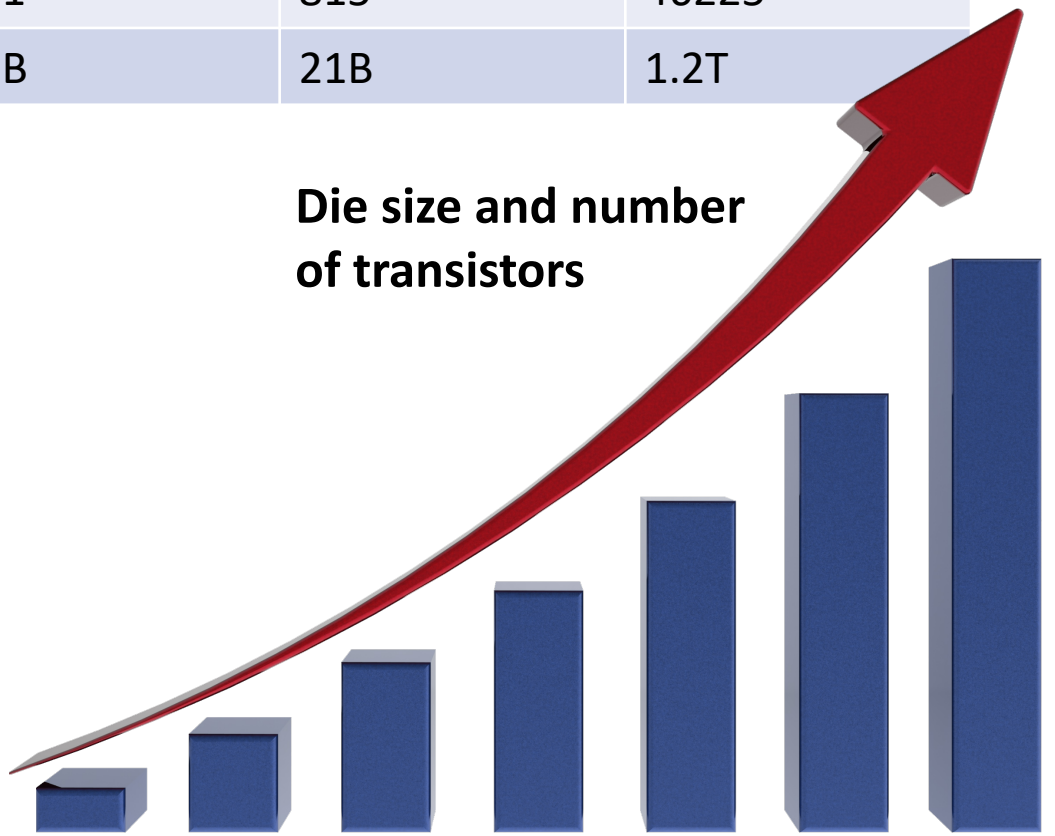
Current architectures

- Current architectures are mostly planar, i.e., 2D
- Not scalable with core counts
- Communication is difficult
- Limited floorplan choices
- More silicon, more probability of failure
  - Not economical to build massive architectures



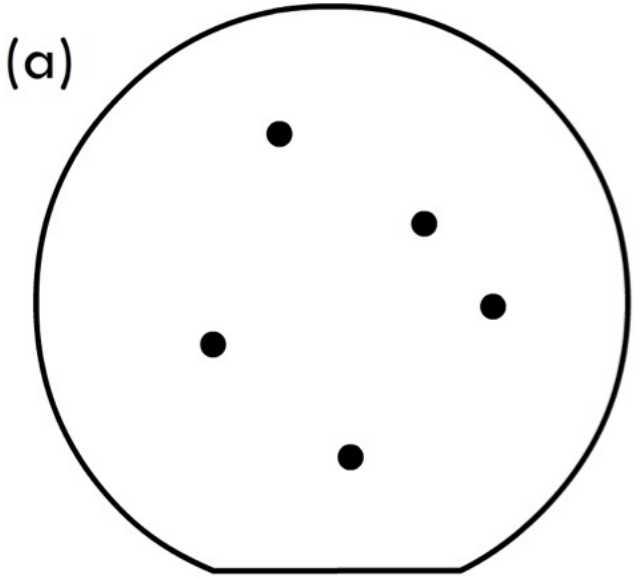
Trends in existing architectures

Product name	Intel Pentium IV	Intel I5-3570K	Nvidia GTX 1080Ti	Nvidia Tesla V100	Cerebras AI engine
Release year	2000	2012	2017	2017	2019
Die size (mm^2)	217	160	471	815	46225
#Transistors	42M	1.4B	12B	21B	1.2T

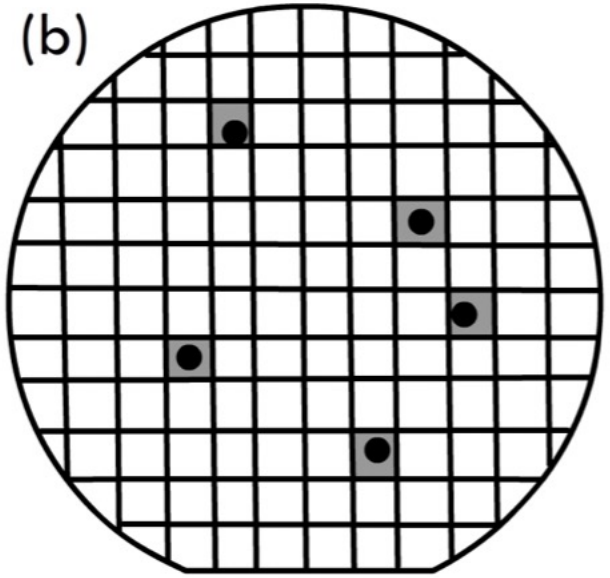


- Transistor counts are likely to increase
- Massive architectures are not practical
  - Not economical, more chance of faults, poor performance

Defects in massive architectures

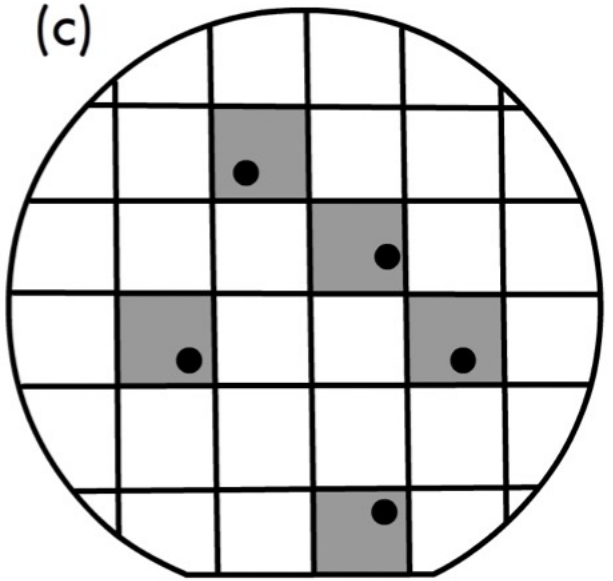


Defects = 5



Defects = 5

$$\text{Yield} = \frac{138 - 5}{138} = 96\%$$



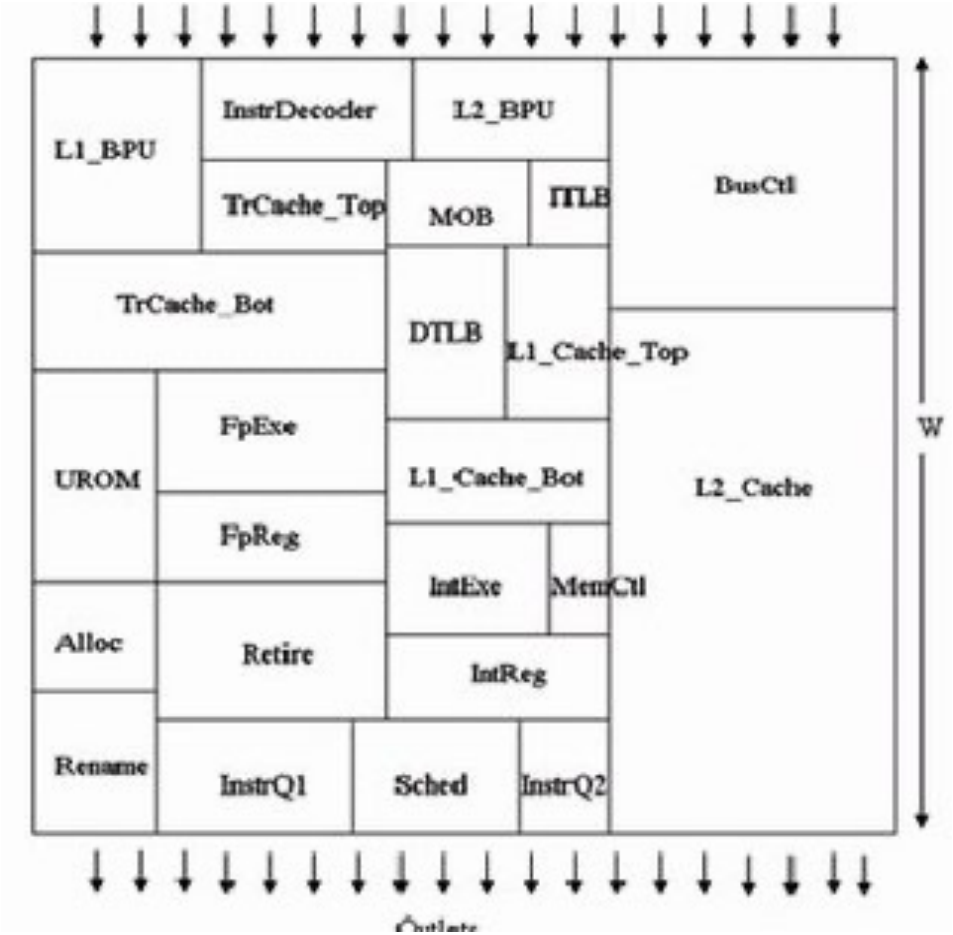
Defects = 5

$$\text{Yield} = \frac{16 - 5}{16} = 69\%$$

- **Low yield is not economical**
- **Cost of defective parts will be passed on to consumer**

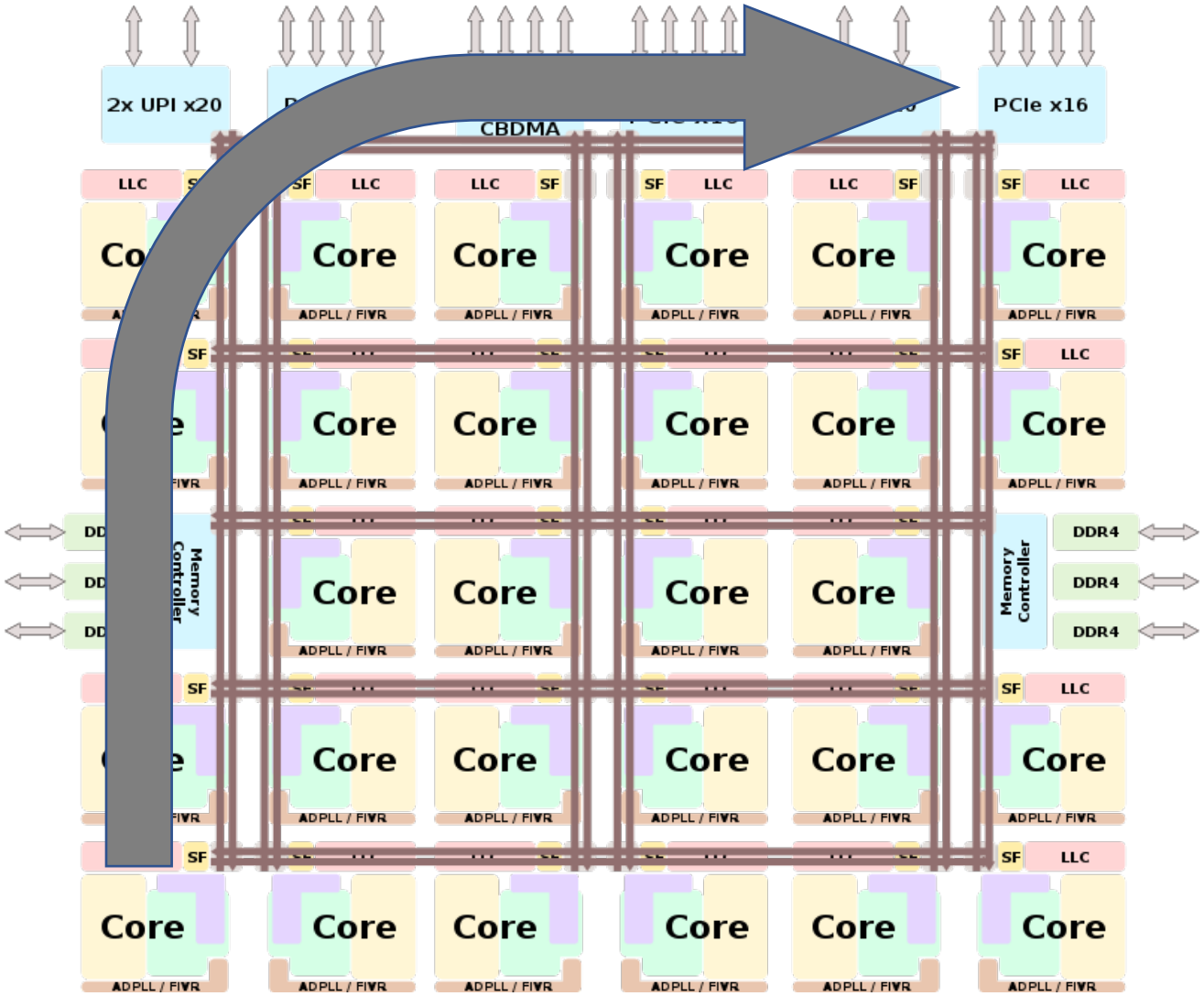
Floor planning in 2D architectures

- Limited size of planar die
- Limited floor planning choices
- Often leads to sub-optimal power-performance-area trade-offs
- More area = more cost to consumer



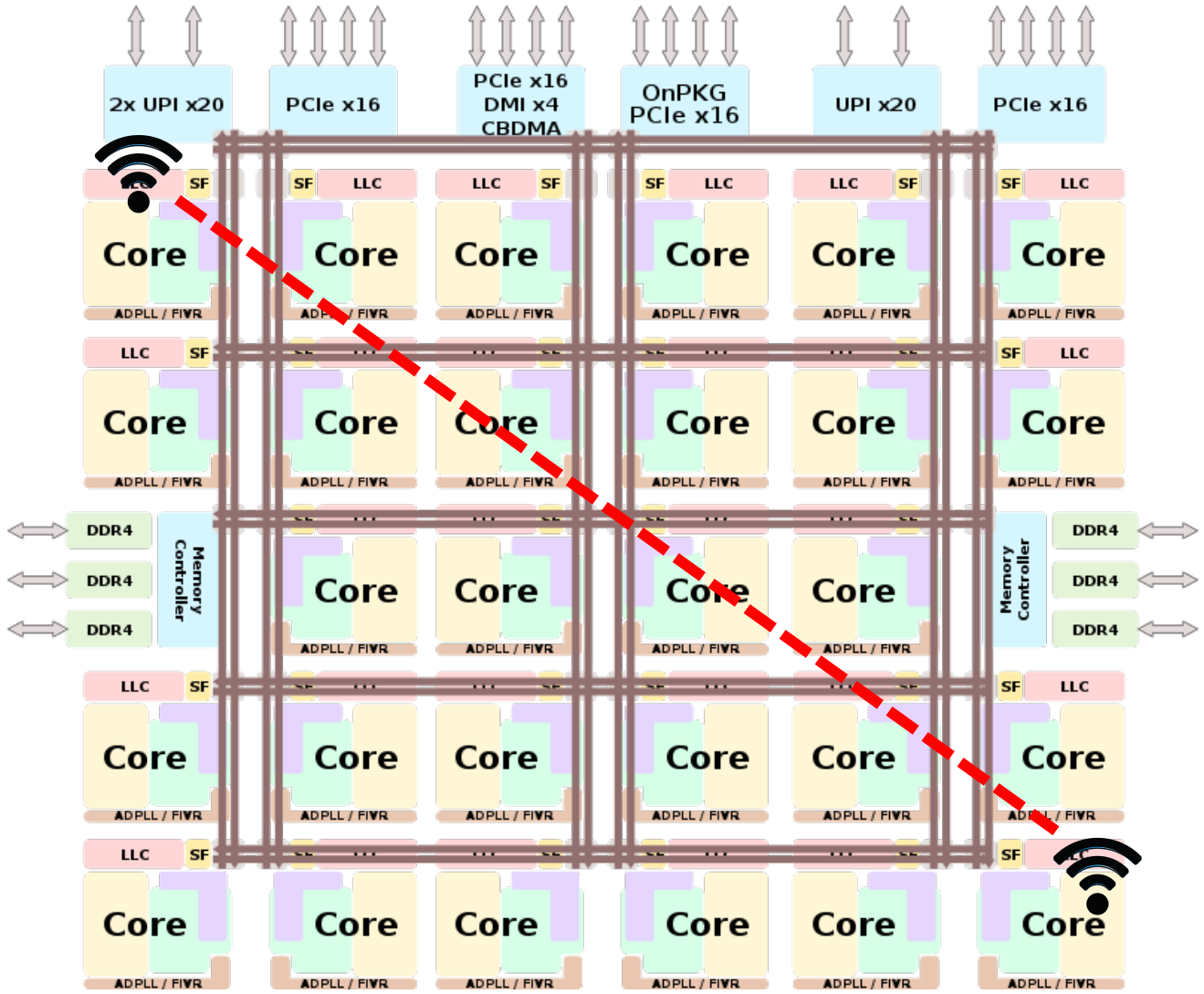
Communication in 2D architectures

- Not scalable with number of cores
- Physical distances are huge
- Data transfer takes many cycles
  - Affects performance

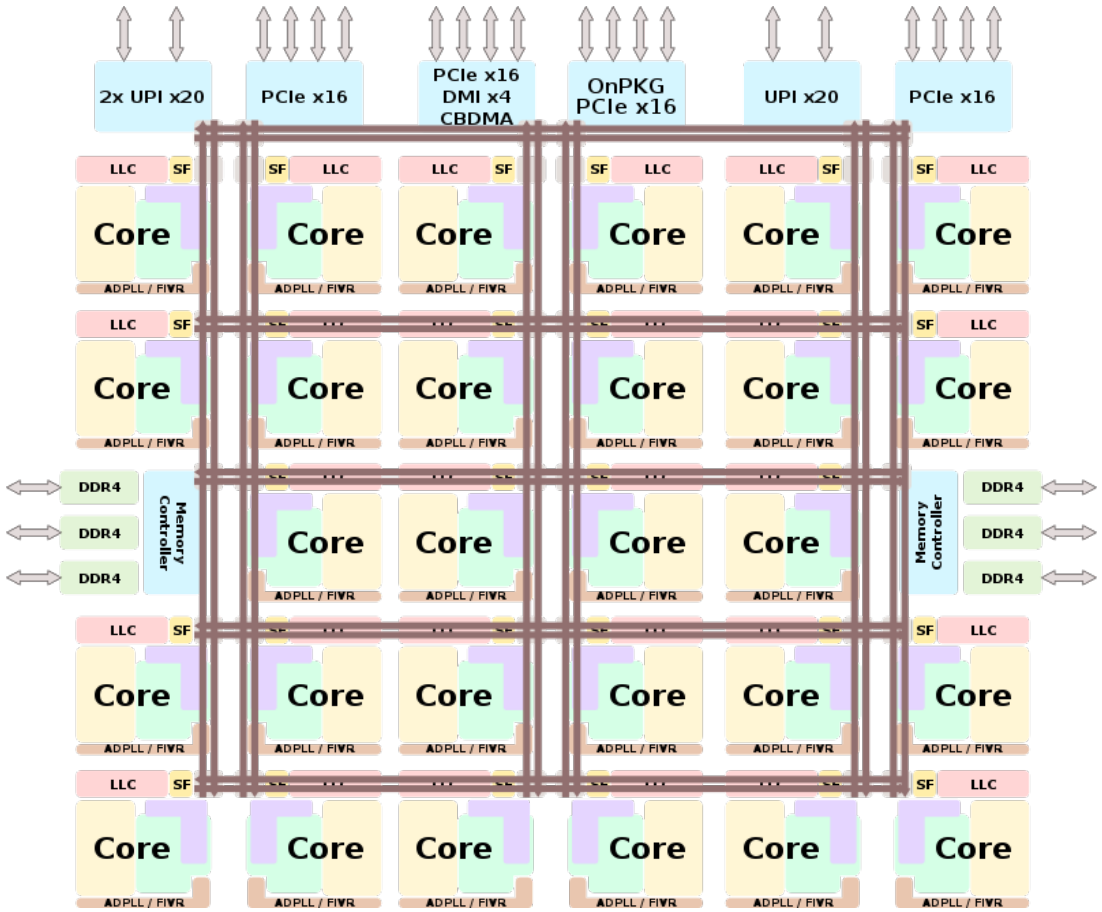


What about better NoC designs?

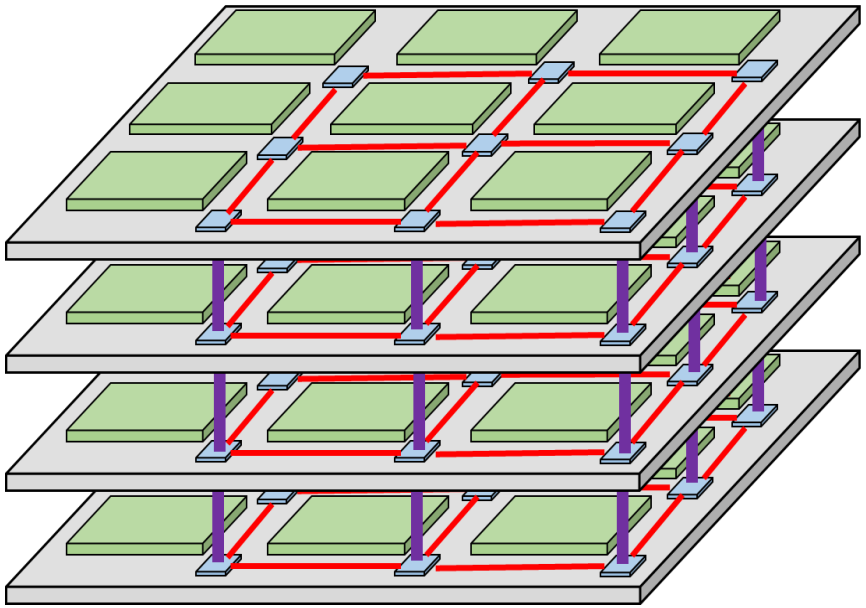
- New NoC technologies can be used
- Wireless, Photonics, etc.
- Single-hop long-range communication
- Challenges:
  - CMOS compatibility
  - Low bandwidth
  - Reliability, etc.



3D architectures (1)

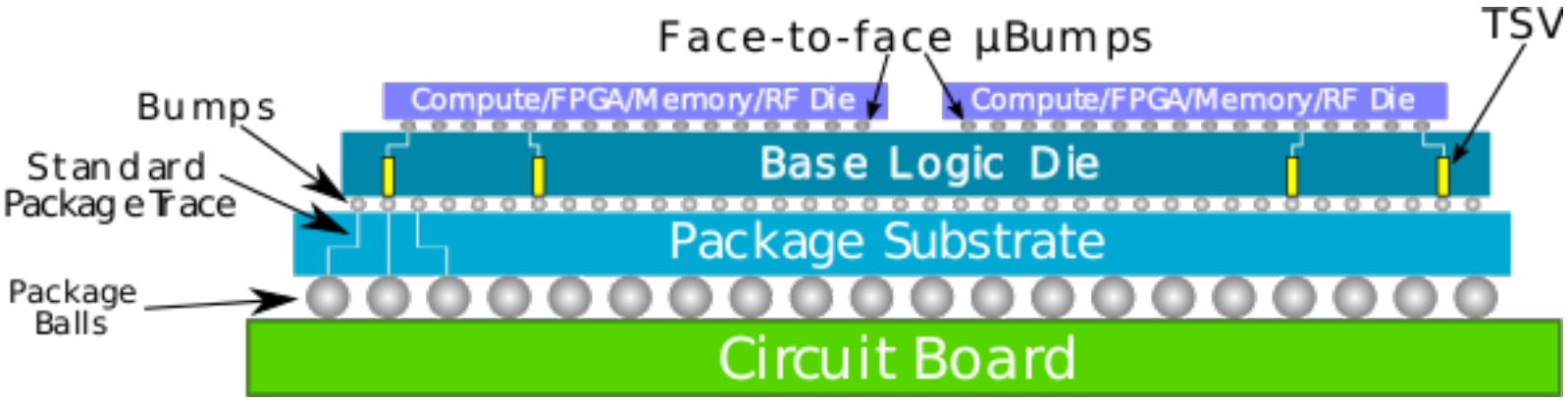


2D architecture
36 core, N^2 area



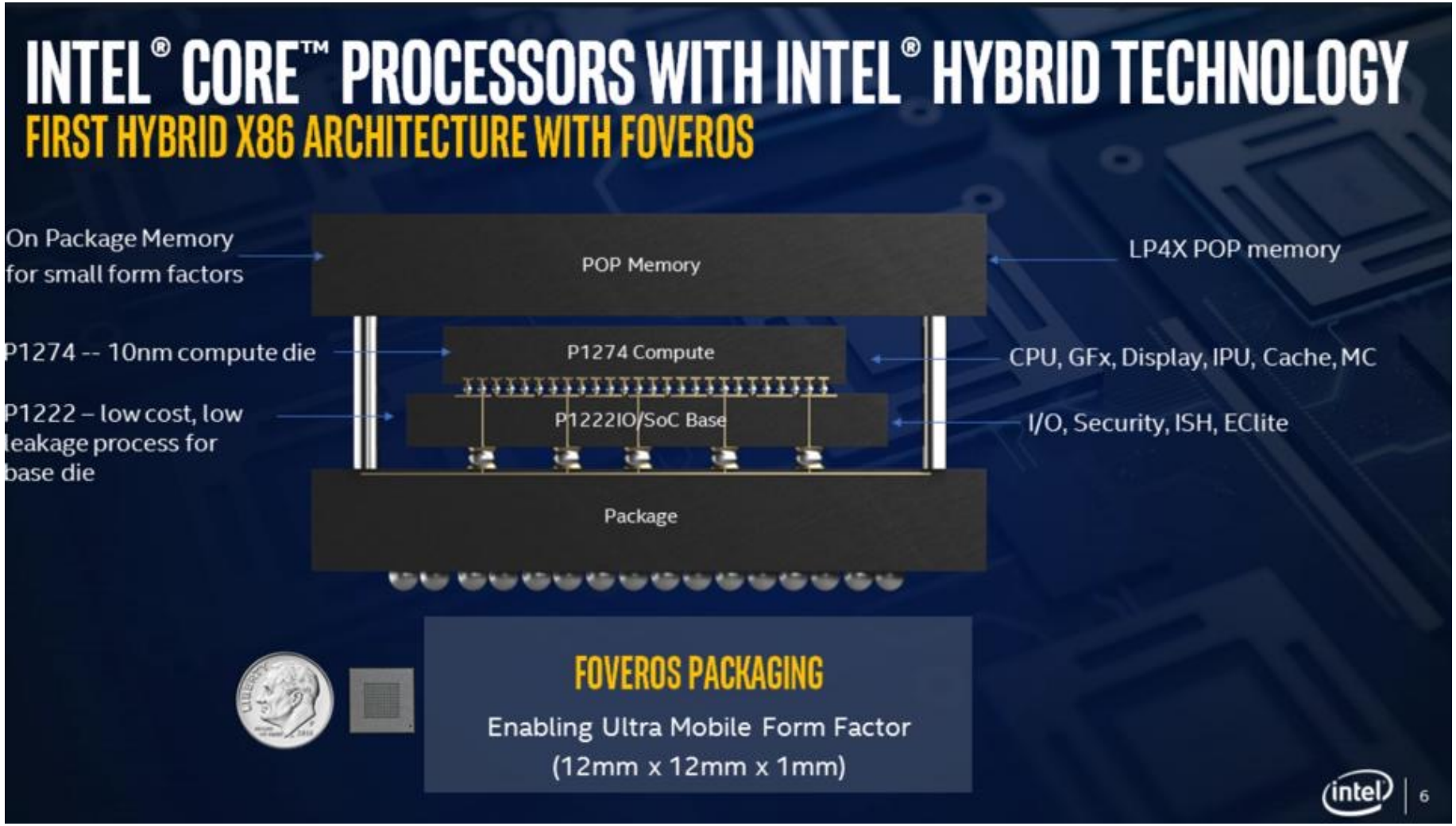
3D architecture
36 core, $N^2/4$ area

3D architectures (2)



- Lower area footprint
  - Smaller devices
- Heterogeneous integration
- Lower end-to-end physical distance
  - Easier communication
  - High-performance, Energy-efficient

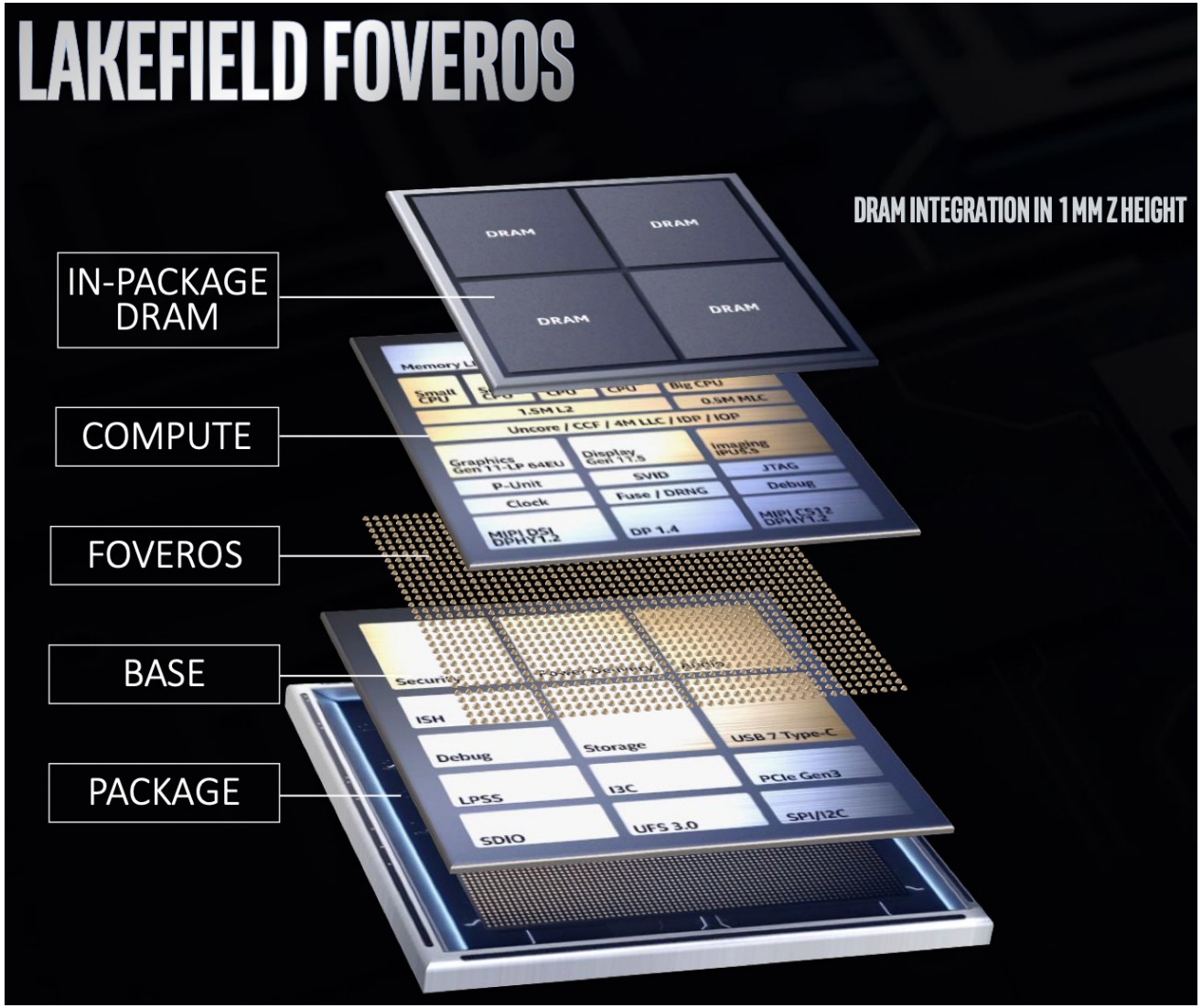
Intel Foveros



https://newsroom.intel.com/wp-content/uploads/sites/11/2019/08/Intel-Lakefield-HotChips-presentation.pdf

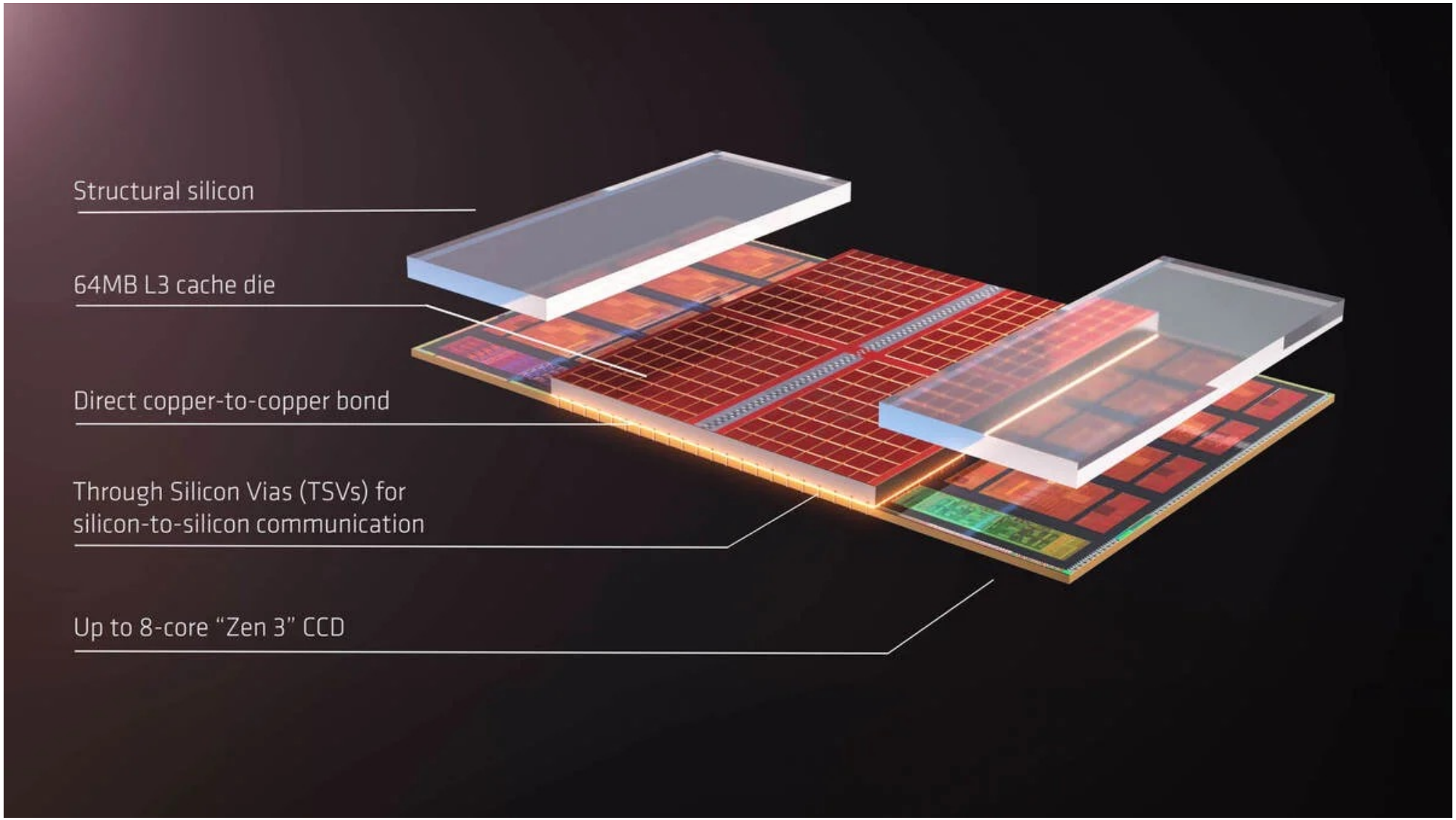
https://www.forbes.com/sites/antonyleather/2020/06/10/intel-announces-first-10nm-hybrid-processors-with-foveros-3d-chip-stacking-tech/?sh=6087325f1284

Intel Lakefield



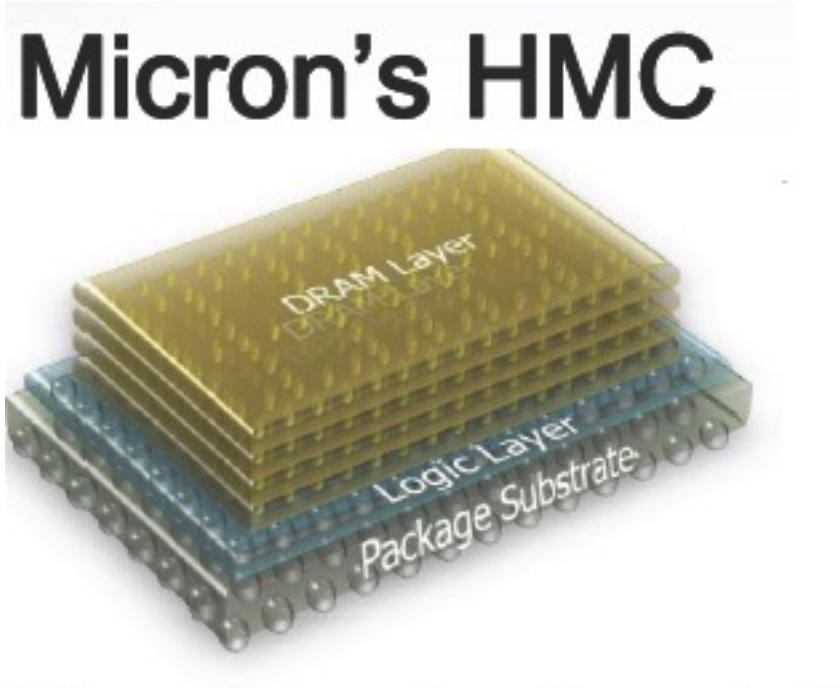
https://newsroom.intel.com/wp-content/uploads/sites/11/2019/08/Intel-Lakefield-HotChips-presentation.pdf

AMD Zen3

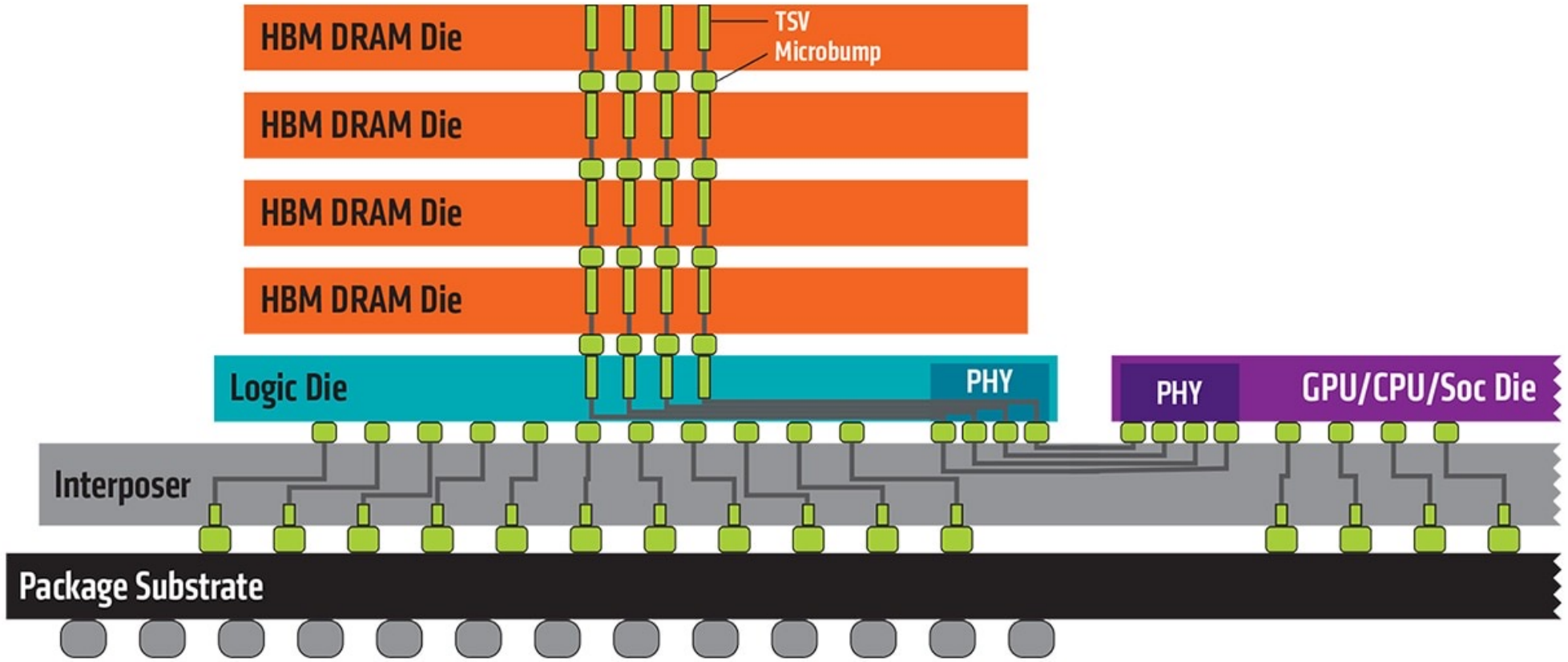

Structural silicon
64MB L3 cache die
Direct copper-to-copper bond
Through Silicon Vias (TSVs) for
silicon-to-silicon communication
Up to 8-core “Zen 3” CCD

Hybrid Memory Cube

- Stack more DRAM layers in 3D
- Increase memory capacity
- Logic layer implements simple logic
- High bandwidth memory access
- Used in many commercial products
- *Simple processing units can be added*



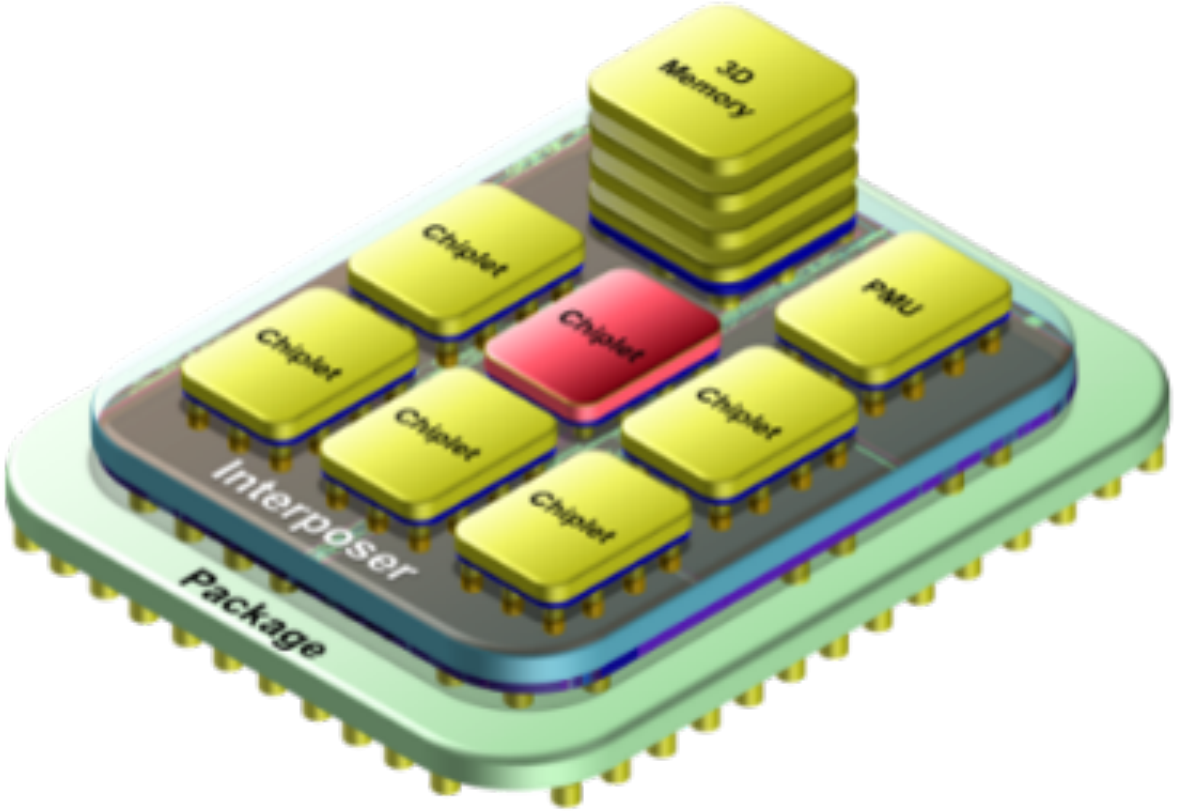
High Bandwidth Memory



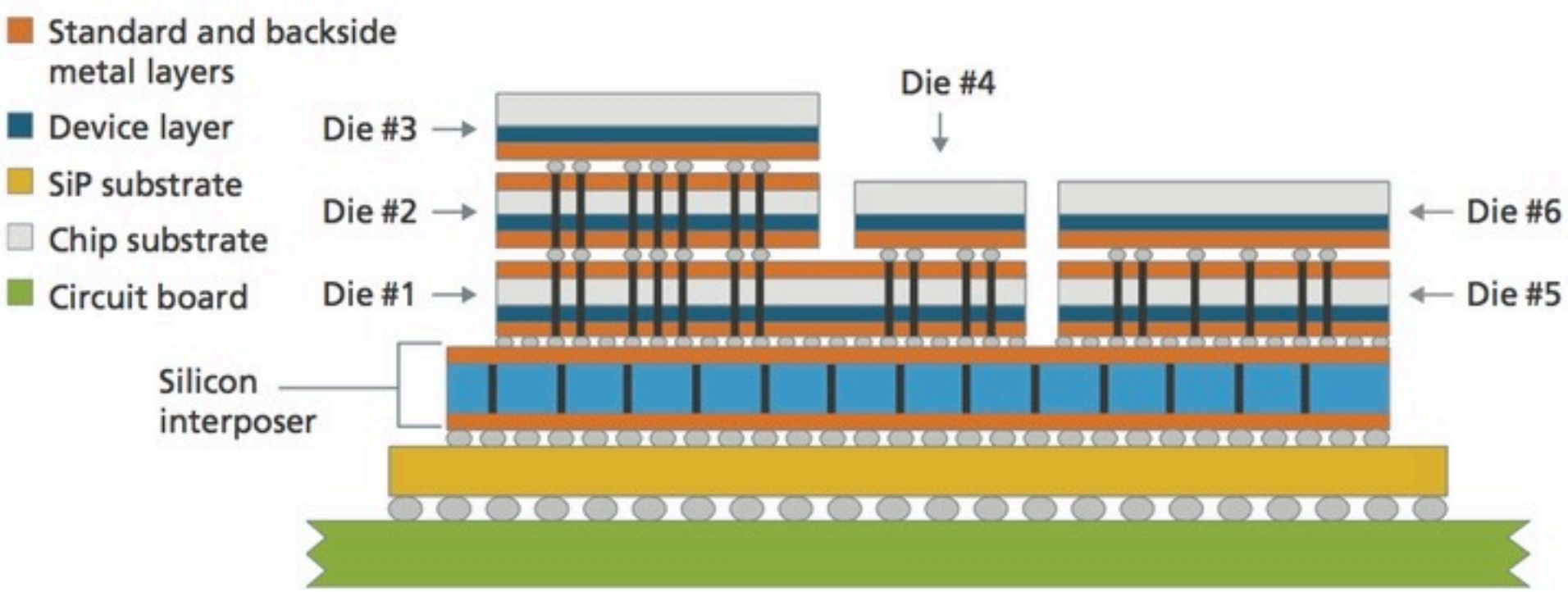
- First HBM memory chip was produced by SK Hynix in 2013
- The first devices to use HBM were the AMD Fiji GPUs in 2015

2.5D architectures

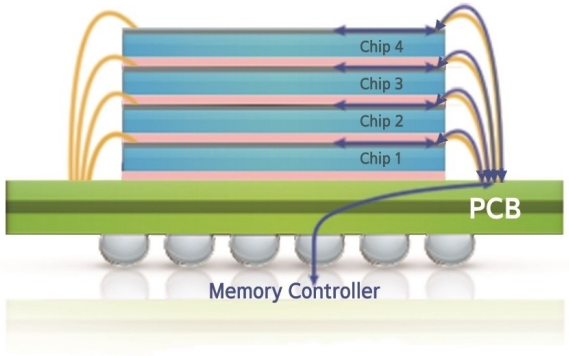
- In-between 2D and 3D architectures
- Mix of planar and 3D architectures
- More scalable than 2D



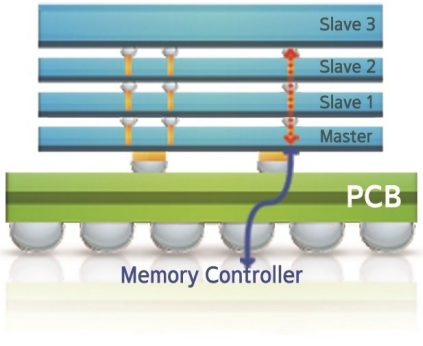
Generic 2.5D/3D architectures



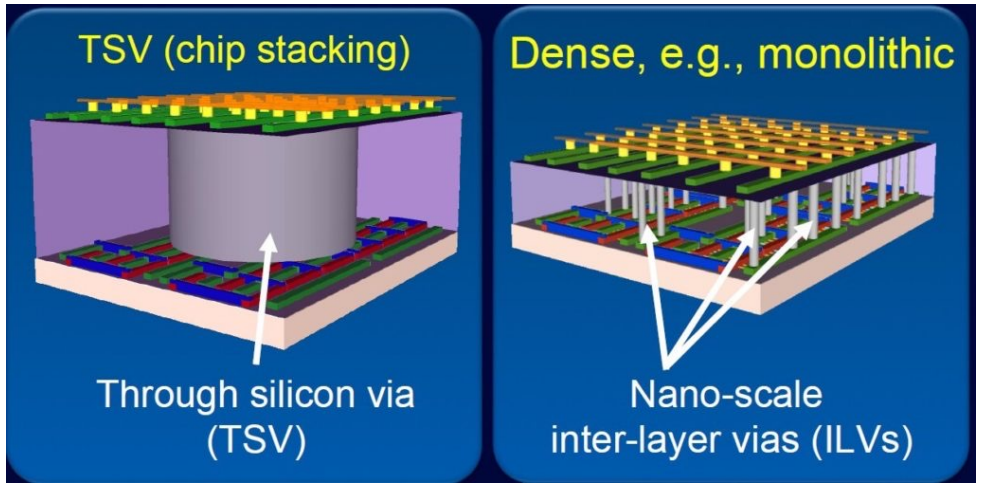
Types of 3D architectures

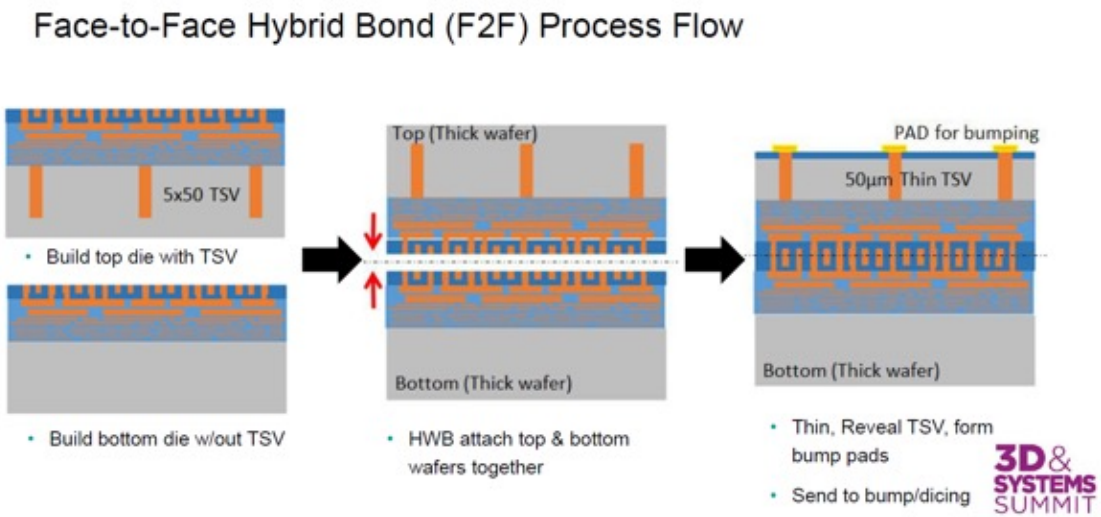


Wire Bonding Technology



TSV Technology

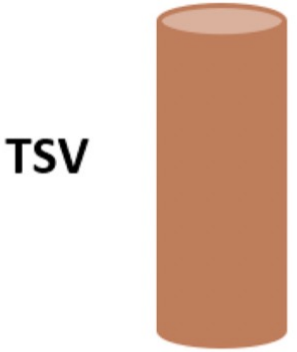




https://www.businesswire.com/news/home/20191007005298/en/Samsung-Electronics-Develops-Industrys-First-12-Layer-3D-TSV-Chip-Packaging-Technology

What is TSV?

- Group of metal wires
- Thicker than normal planar wires
- Establish vertical connectivity



Diameter: 4~8 μm
Depth: 20~50 μm
Pitch: 8~16 μm

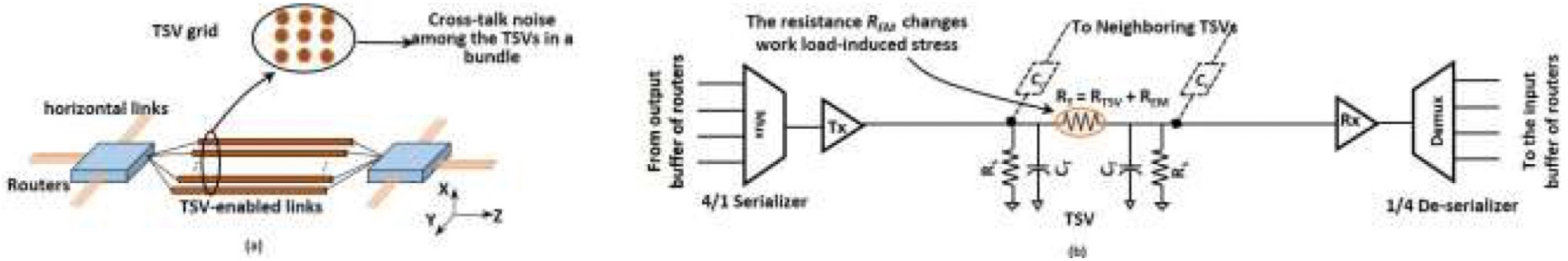
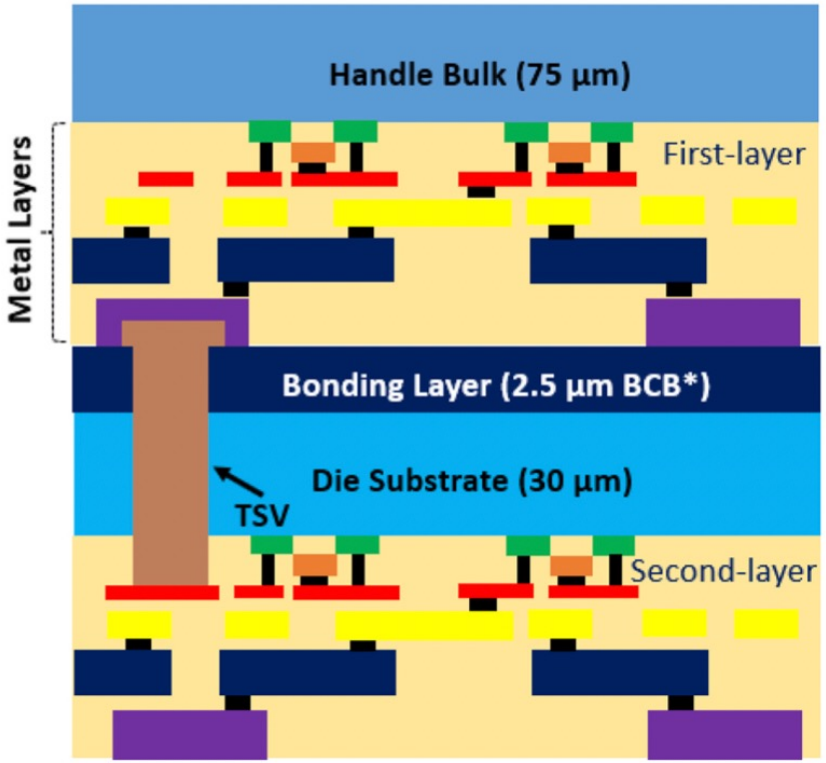


Fig. 2:(a) A TSV-based vertical link between two routers from adjacent layers. (b) A simple model of TSV-based VL to study the workload-induced stress. The TSV is connected to a serializer and a de-serializer.

TSV-based 3D



TSV-based 3D IC

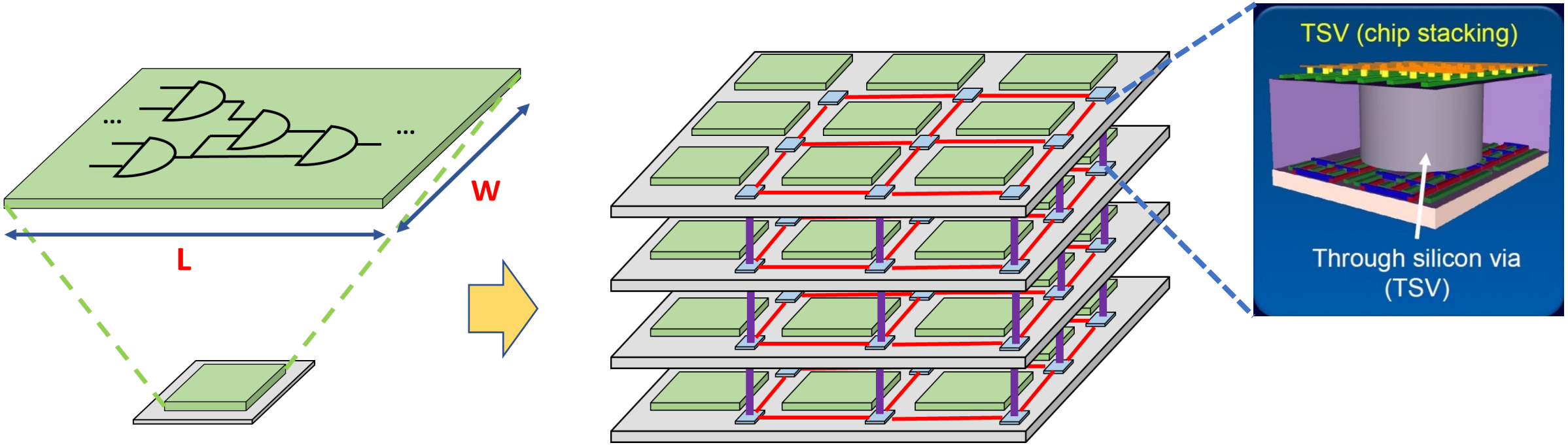
Table 1: The different materials used in the layers, their thermal conductivities and vertical thicknesses

Layer/Structure	Material	Thermal Conductivity (W/m-K)	Vertical Thickness
Monolithic			
Handle Bulk	Silicon	141	$75\mu m$
ILD (Inter-tier)	SiO_2	1.38	$100nm$
TSV-based			
Handle Bulk	Silicon	141	$75\mu m$
Die0 Substrate	Silicon	141	$30\mu m$
Bonding Layer	BCB	0.29	$2.5\mu m$
TSV	Copper	401	$30\mu m$
TSV-bump	Solder	50	$2.5\mu m$

- Transistors on silicon layers
- Bonding layer acts as glue
  - Benzocyclobutene (BCB)
- Overall, several um thick

S. K. Samal, et. al., "Fast and accurate thermal modeling and optimization for monolithic 3D ICs," 2014 51st ACM/EDAC/IEEE Design Automation Conference (DAC), 2014, pp. 1-6

TSV-based manycore design



- Conventional hardware design is planar
  - Sub-optimal power-performance
- Planar logic blocks stacked physically to create 3D

M3D-based 3D

MIV

Diameter: ~50 *nm*
Depth: ~100 *nm*
Pitch: 44 *nm*

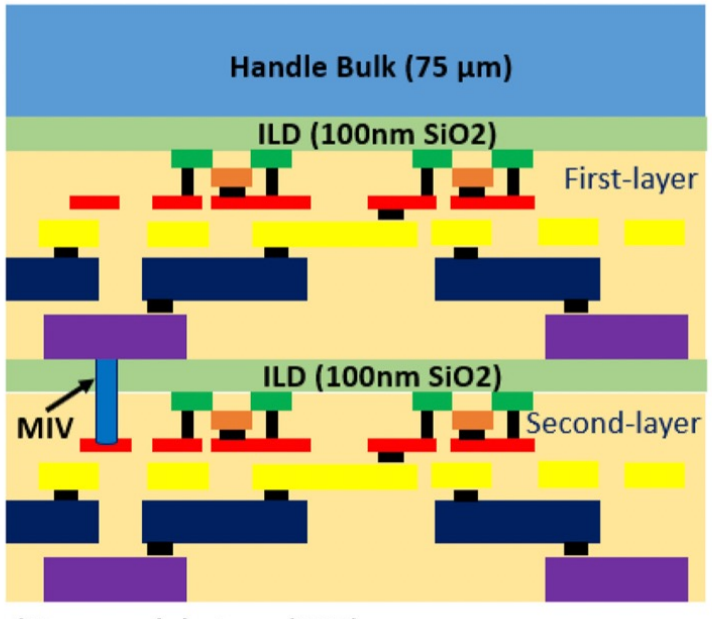
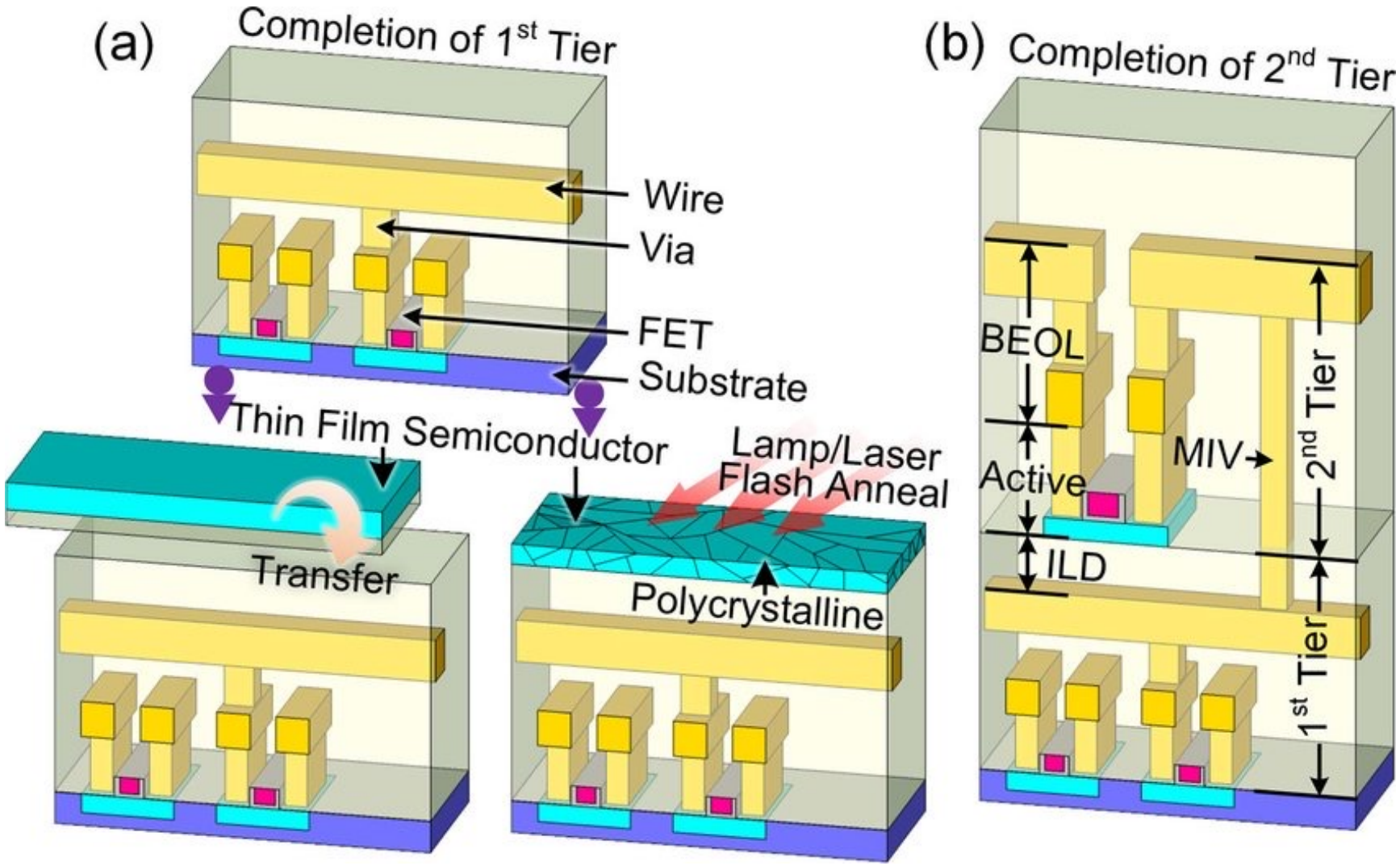


Table 1: The different materials used in the layers, their thermal conductivities and vertical thicknesses

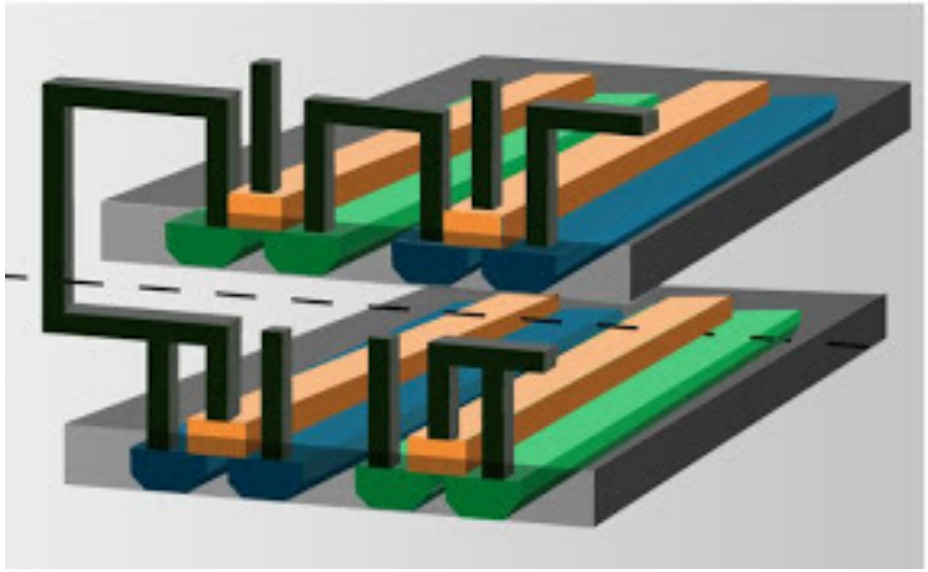
Layer/Structure	Material	Thermal Conductivity (W/m-K)	Vertical Thickness
Monolithic			
Handle Bulk	Silicon	141	$75\mu m$
ILD (Inter-tier)	SiO_2	1.38	$100nm$
TSV-based			
Handle Bulk	Silicon	141	$75\mu m$
Die0 Substrate	Silicon	141	$30\mu m$
Bonding Layer	BCB	0.29	$2.5\mu m$
TSV	Copper	401	$30\mu m$
TSV-bump	Solder	50	$2.5\mu m$

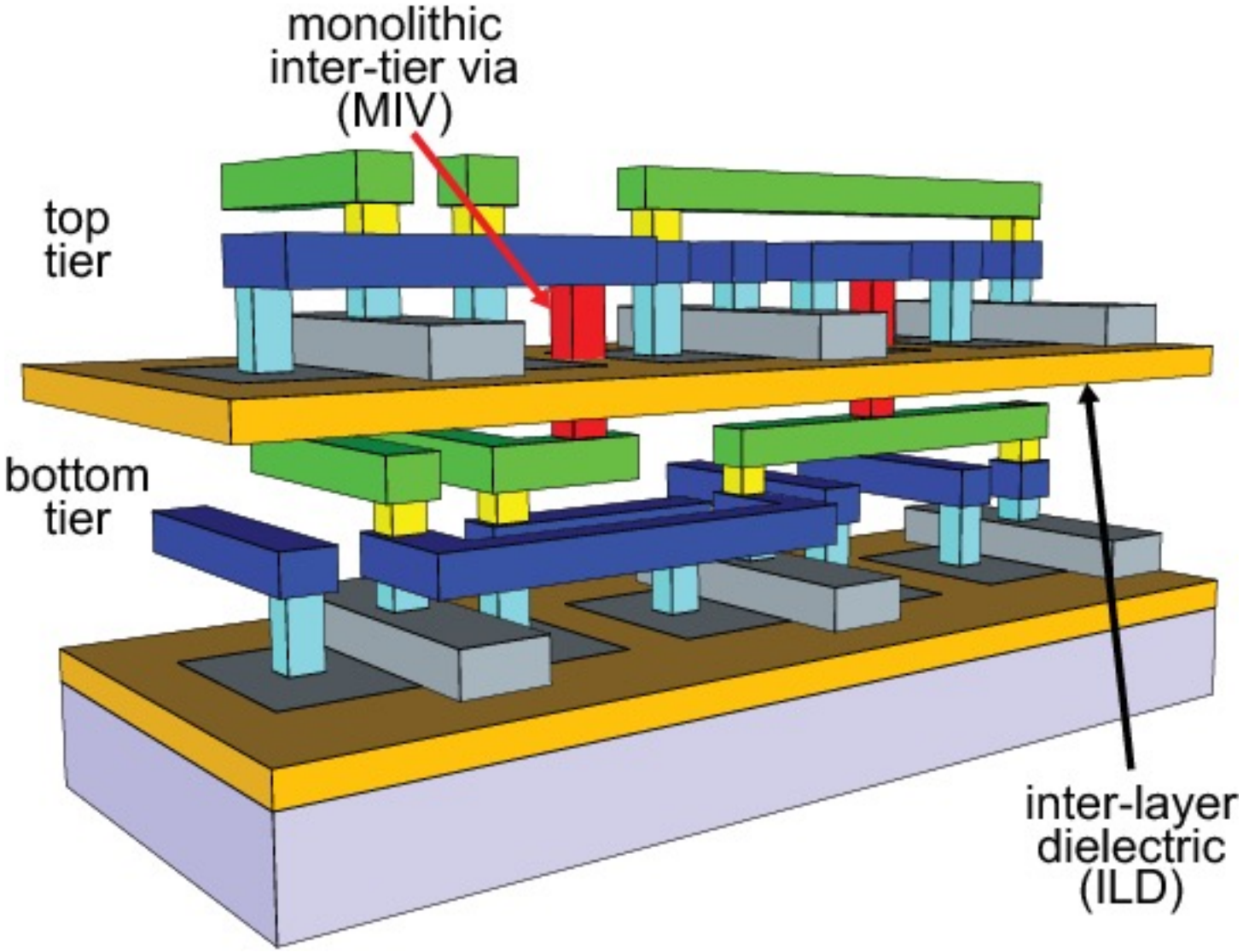
- **No Bonding layer…!!!**
- **Upper layer is built over the previous layer**
- **Significantly thinner**

Manufacturing M3D-based 3D

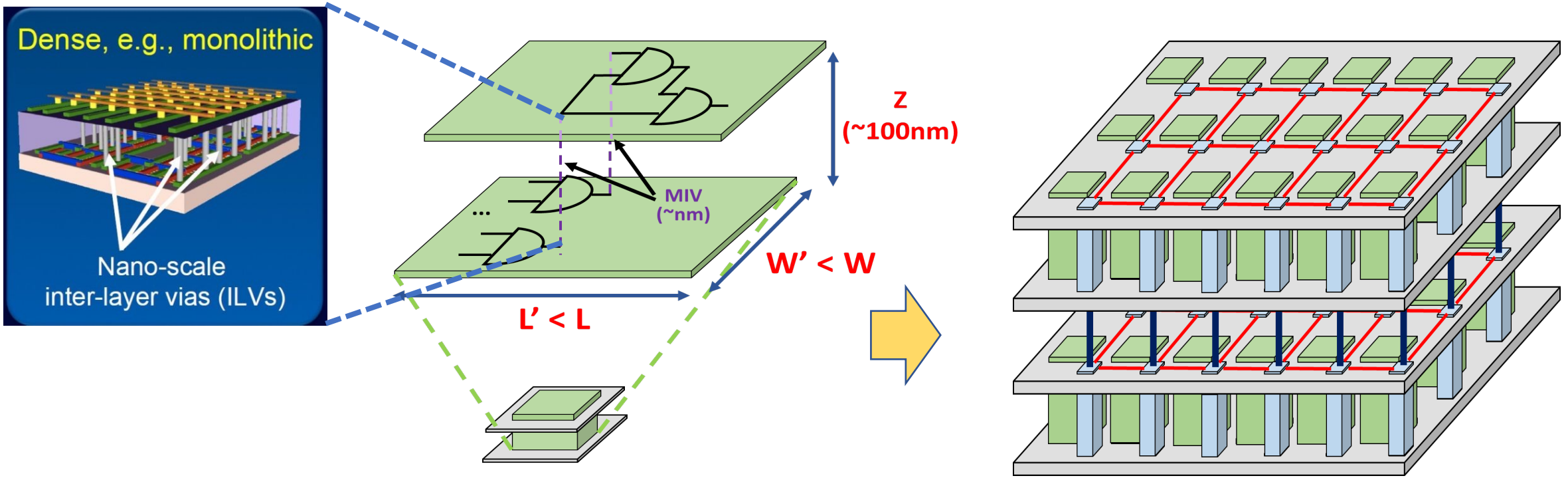


M3D IC




monolithic
inter-tier via
(MIV)
top
tier
bottom
tier
inter-layer
dielectric
(ILD)

M3D-based manycore design



- M3D enables 3D hardware blocks
  - CPU, GPU, ReRAM, etc.
  - Significantly less area overhead
  - Higher performance and low power

Benefits of 3D: Heterogeneity

- Heterogeneous architectures
  - Different types of cores
  - May be incompatible to make in one chip
- Different layers can have different technologies
- Customized architectures

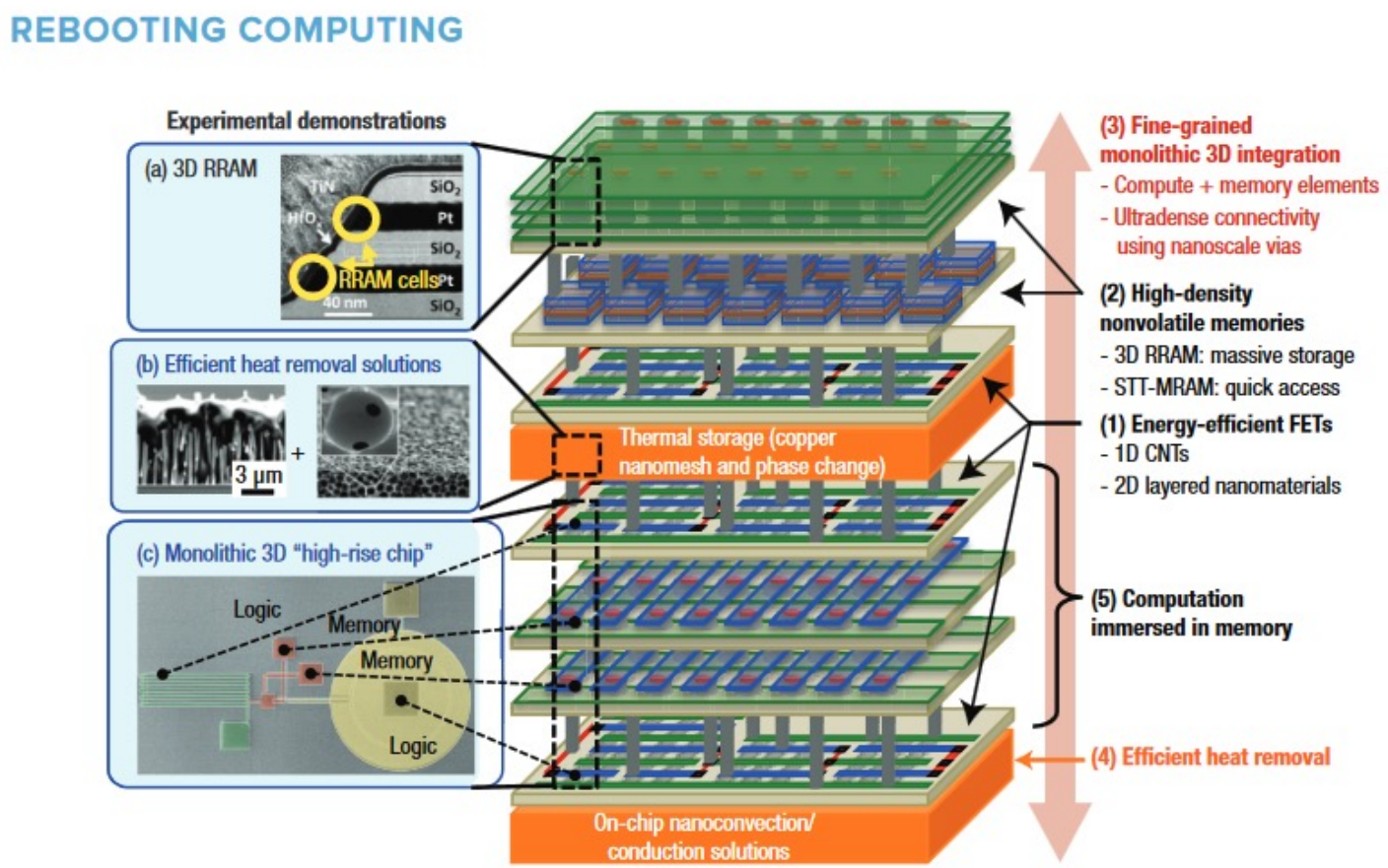
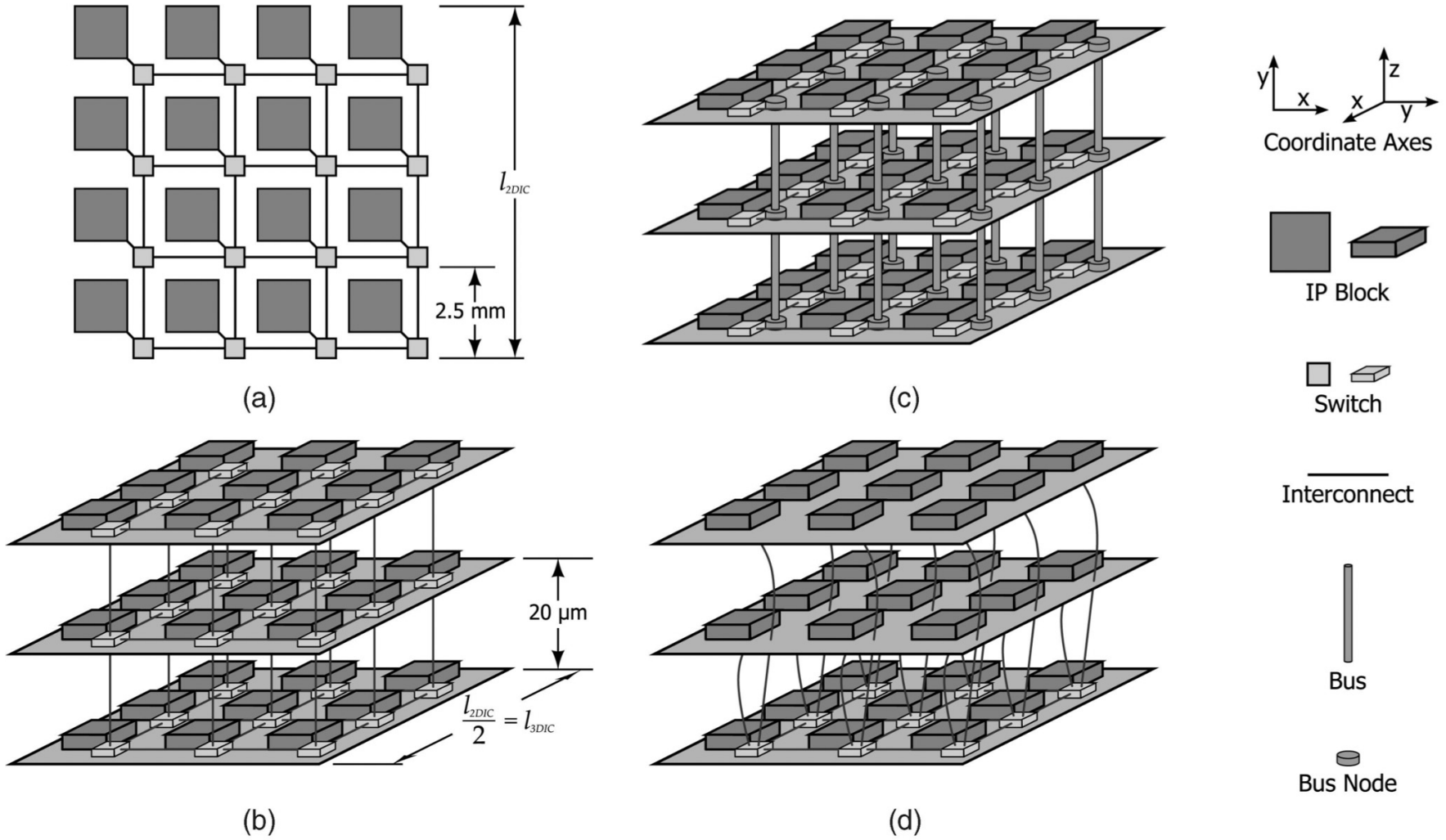


FIGURE 1. Monolithically integrated 3D system enabled by Nano-Engineered Computing Systems Technology (N3XT). On the right are the five key N3XT components. On the left are images of experimental technology demonstrations: (a) transmission electron microscopy (TEM) of a 3D resistive RAM (RRAM) for massive storage, (b) scanning electron microscopy (SEM) of nanostructured materials for efficient heat removal (left: microscale capillary advection; right: copper nanomesh with phase-change thermal storage), and (c) SEM of a monolithic 3D chip for high-performance and energy-efficient computation. CNTs: carbon nanotubes, FETs: field-effect transistors, and STT-MRAM: spin-transfer torque magnetic RAM.

Mohamed M. Sabry Aly. 2019. N3XT Monolithic 3D Energy-Efficient Computing Systems. In Proceedings of the 2019 on Great Lakes Symposium on VLSI (GLSVLSI '19). Association for Computing Machinery, New York, NY, USA, 463. DOI:https://doi.org/10.1145/3299874.3319489

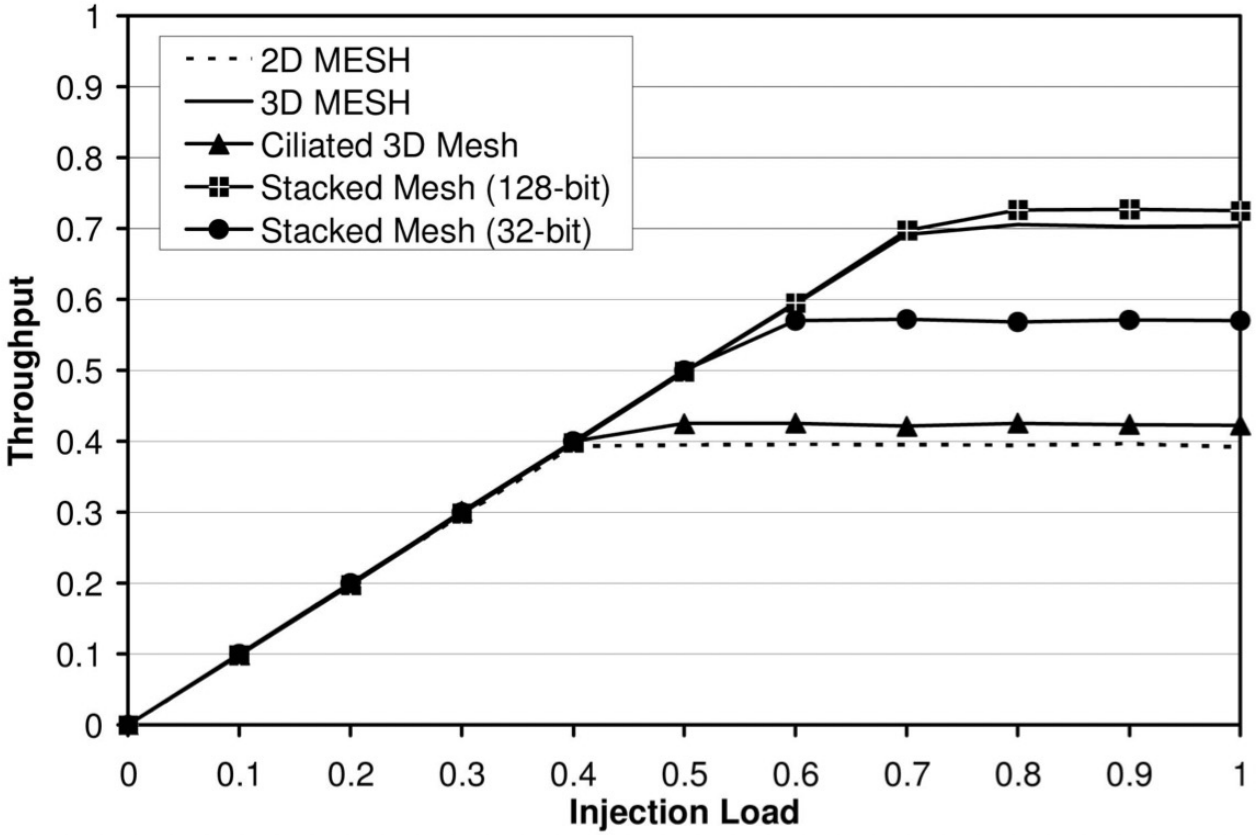
3D NoC-enabled manycore design

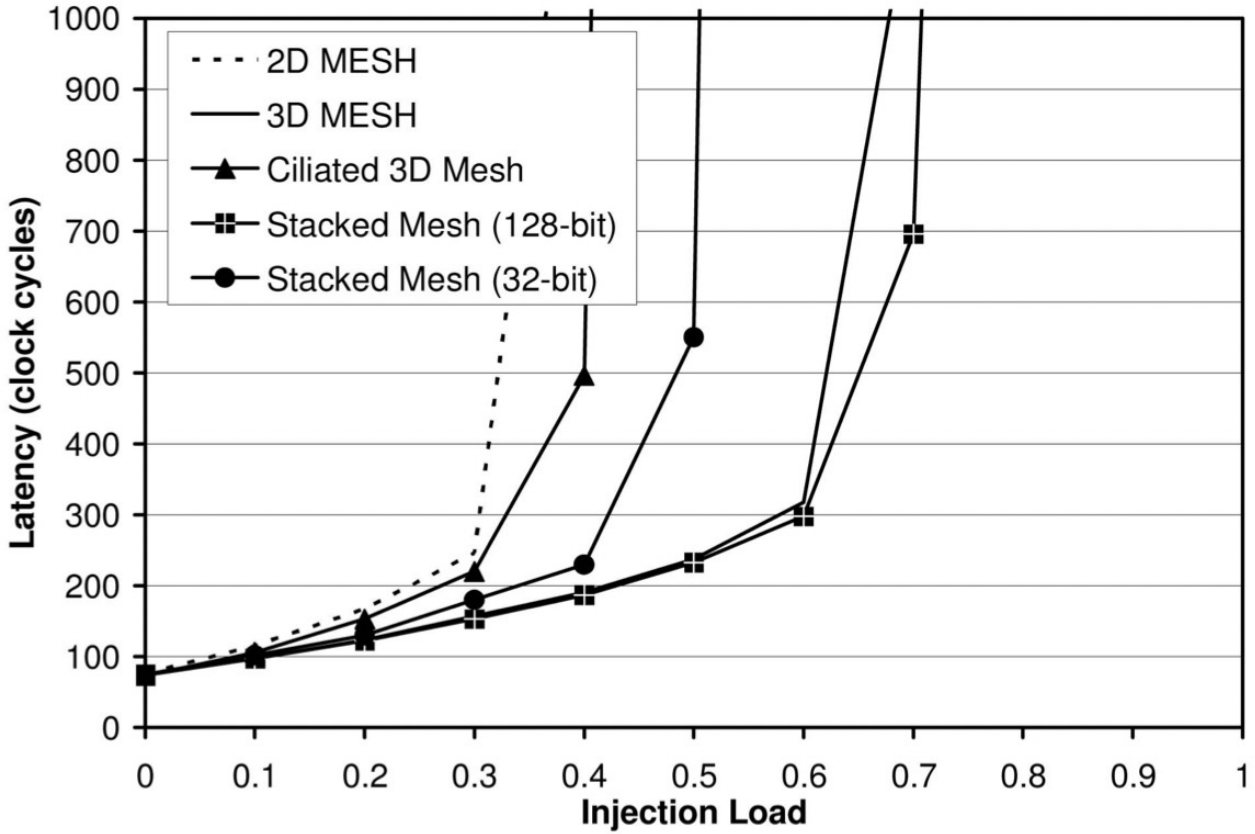


(a) Two-dimensional mesh. (b) Three-dimensional mesh. (c) Stacked mesh. (d) Ciliated 3D mesh

B. S. Feero and P. P. Pande, "Networks-on-Chip in a Three-Dimensional Environment: A Performance Evaluation," in IEEE Transactions on Computers, vol. 58, no. 1, pp. 32-45, Jan. 2009, doi: 10.1109/TC.2008.142.

Power, Performance benefits (1)

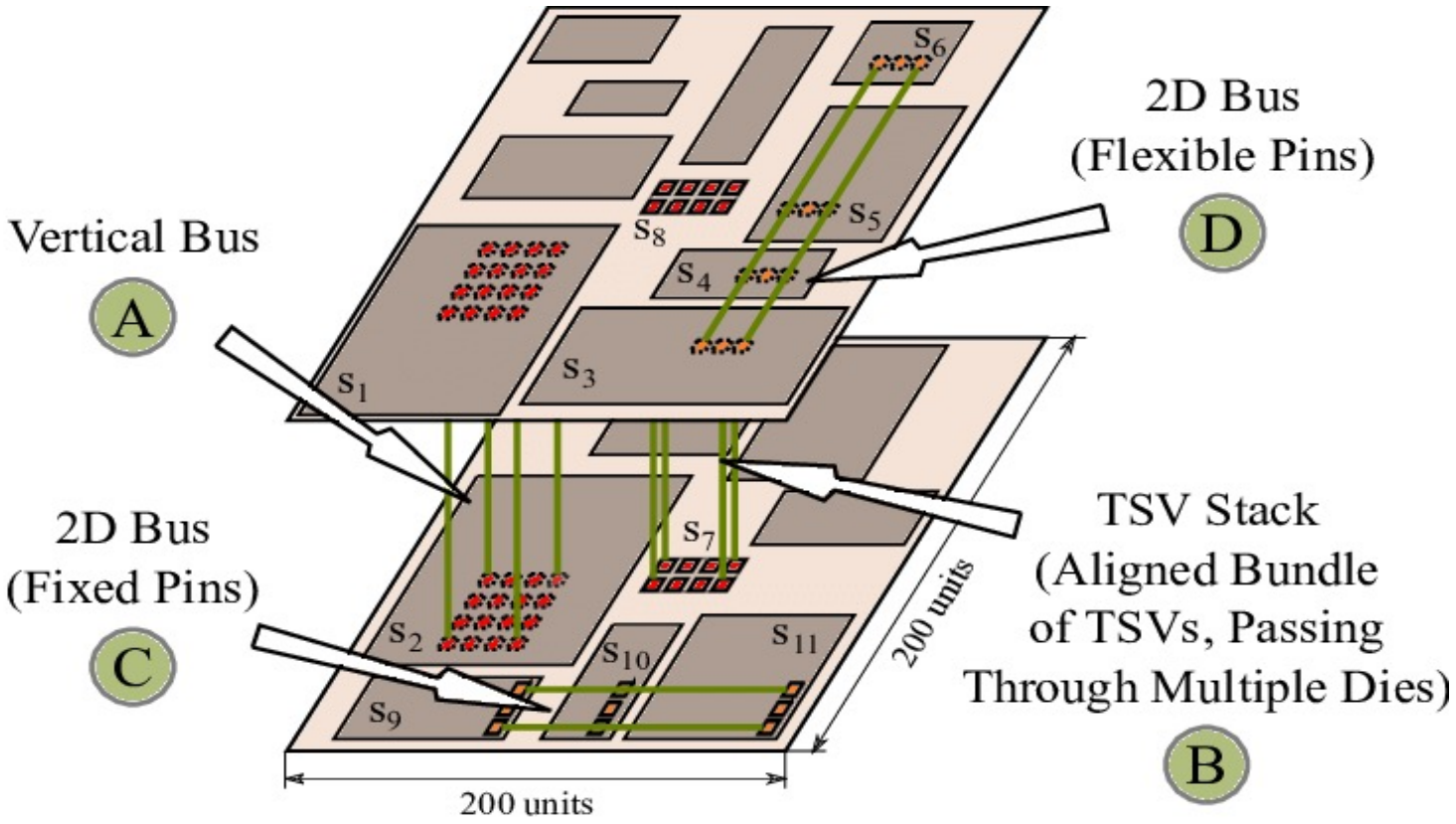




- **3D has higher throughput and lower latency at high injection loads**

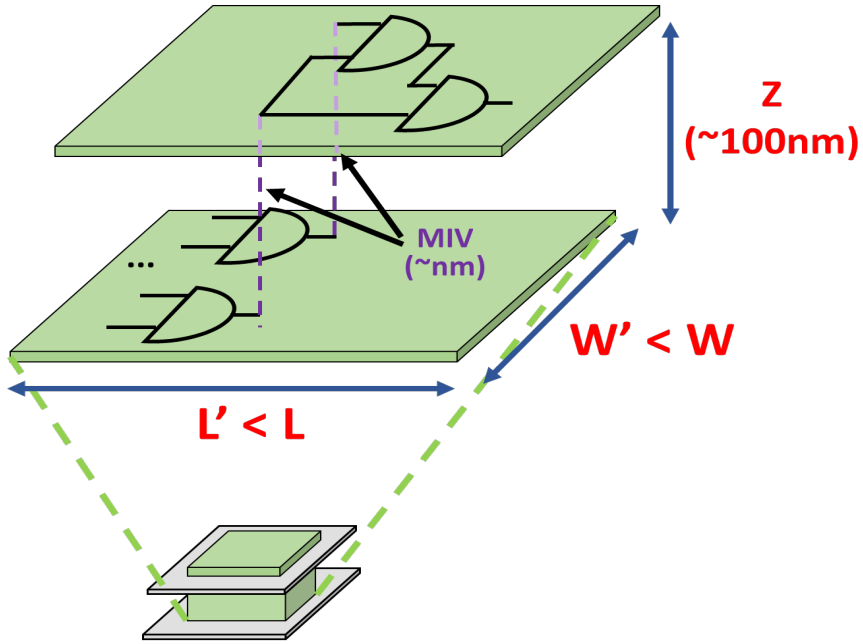
B. S. Feero and P. P. Pande, "Networks-on-Chip in a Three-Dimensional Environment: A Performance Evaluation," in IEEE Transactions on Computers, vol. 58, no. 1, pp. 32-45, Jan. 2009, doi: 10.1109/TC.2008.142.

Floorplanning in 3D



- One additional degree of freedom
- More choices for floorplanning
- Better power-performance

M3D GPU



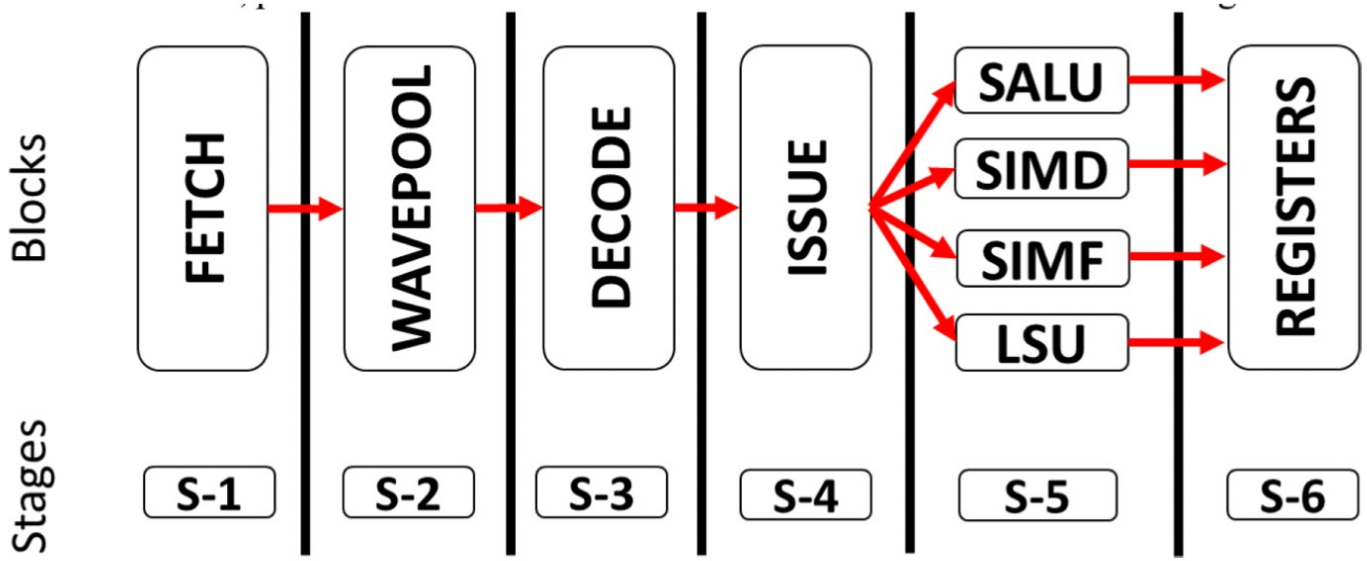


Fig. 3. Basic execution pipeline of a GPU core.

- 3D logic blocks
  - 3D adder, ALU, etc.
  - Smaller footprint
  - Shorter wires

Power, Performance benefits (2)

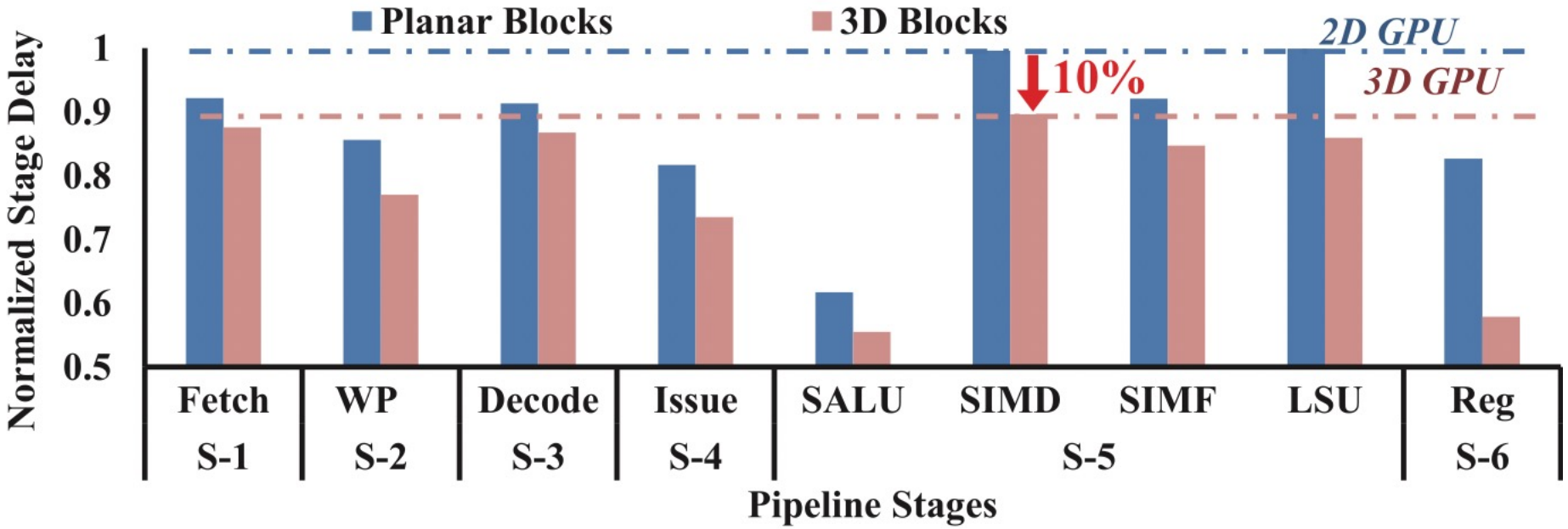
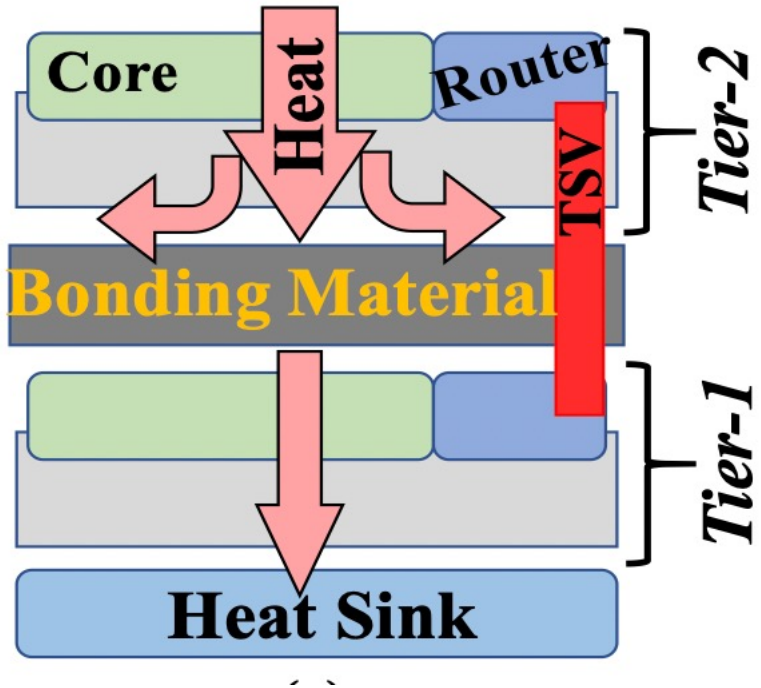


Fig. 6. Normalized pipeline stage latencies of the planar and M3D GPU core.

- GPU has pipelined implementation
- Slowest stage determines clock frequency
- M3D GPU is 10% faster
  - No design changes made

A. I. Arka et. al., HeM3D: Heterogeneous Manycore Architecture Based on Monolithic 3D Vertical Integration. ACM Trans. Des. Autom. Electron. Syst. 26, 2, Article 16 (February 2021), 21 pages

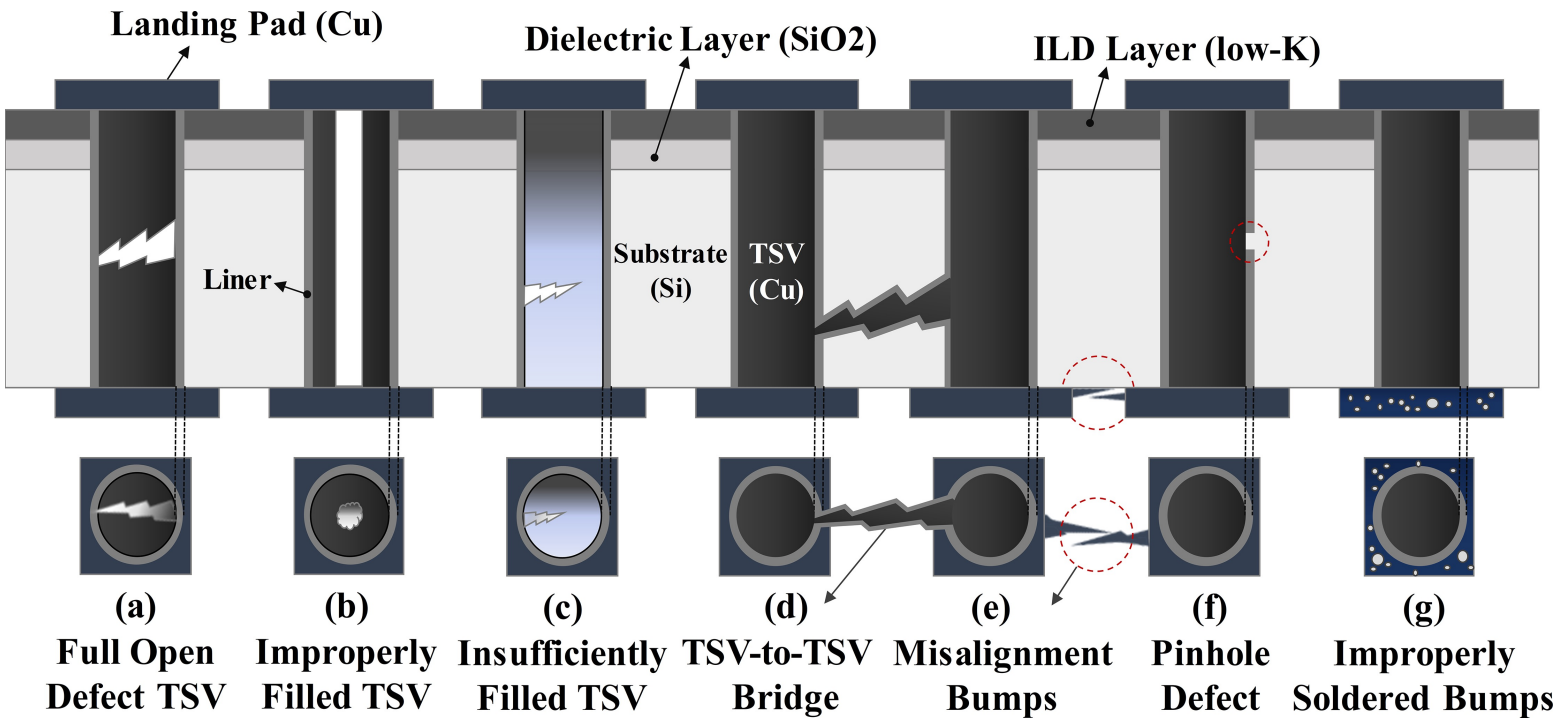
Challenges with 3D architectures: Heat



- **Heat dissipation is difficult**
- **Distance from heat sink**
- **Bonding material has poor thermal conductivity**

A. I. Arka et al., “HeM3D: Heterogeneous Manycore Architecture Based on Monolithic 3D Vertical Integration“, in ACM TODAES 21
Lee Y-w, Lim H, Seo S, Cho K, Kang S (2019) A low-cost concurrent TSV test architecture with lossless test output compression scheme. PLoS ONE 14(8): e0221043.

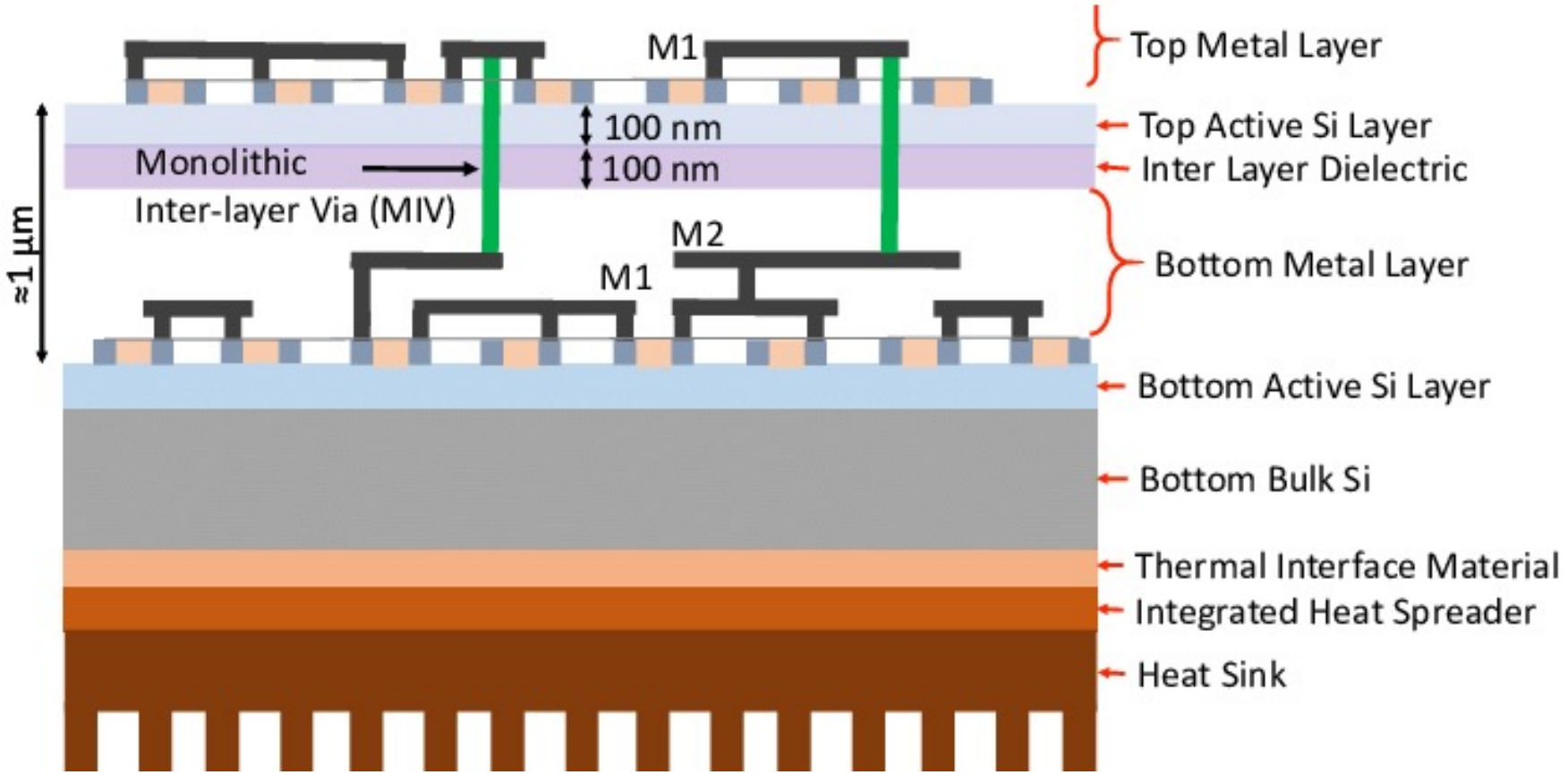
Fabrication challenges



- Fabrication challenges
  - Not as mature as 2D architectures

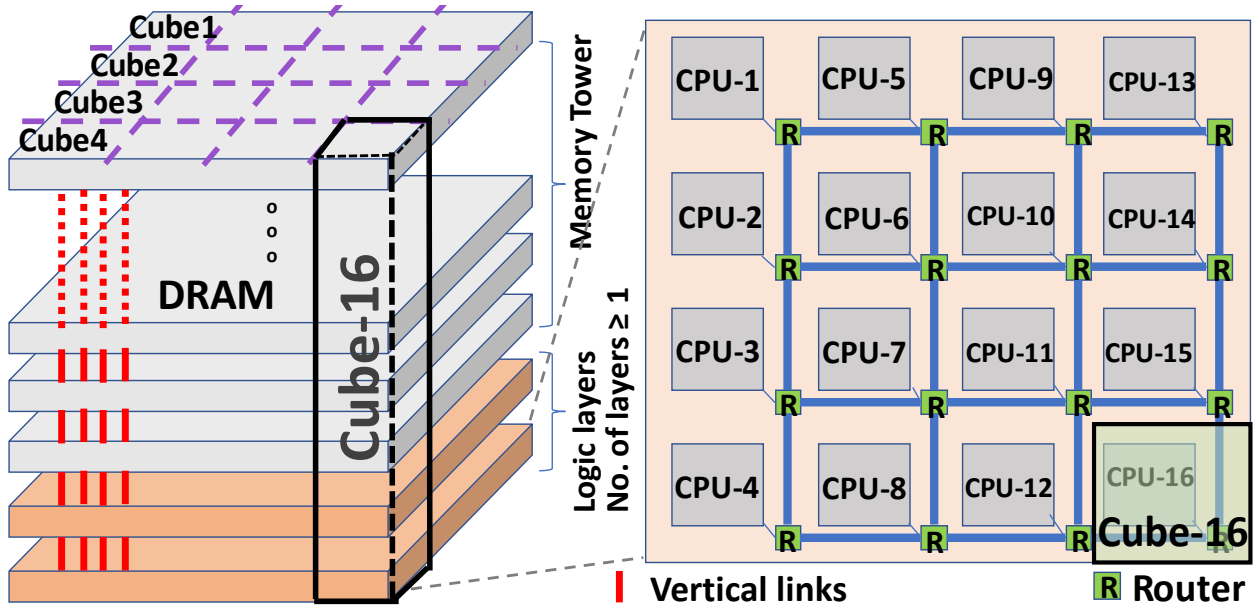
Process variation

- **Upper layer material deposited over lower**
- **High/Low temperature**
- **Wires/Transistors get slower as a result**



Processing-in-memory

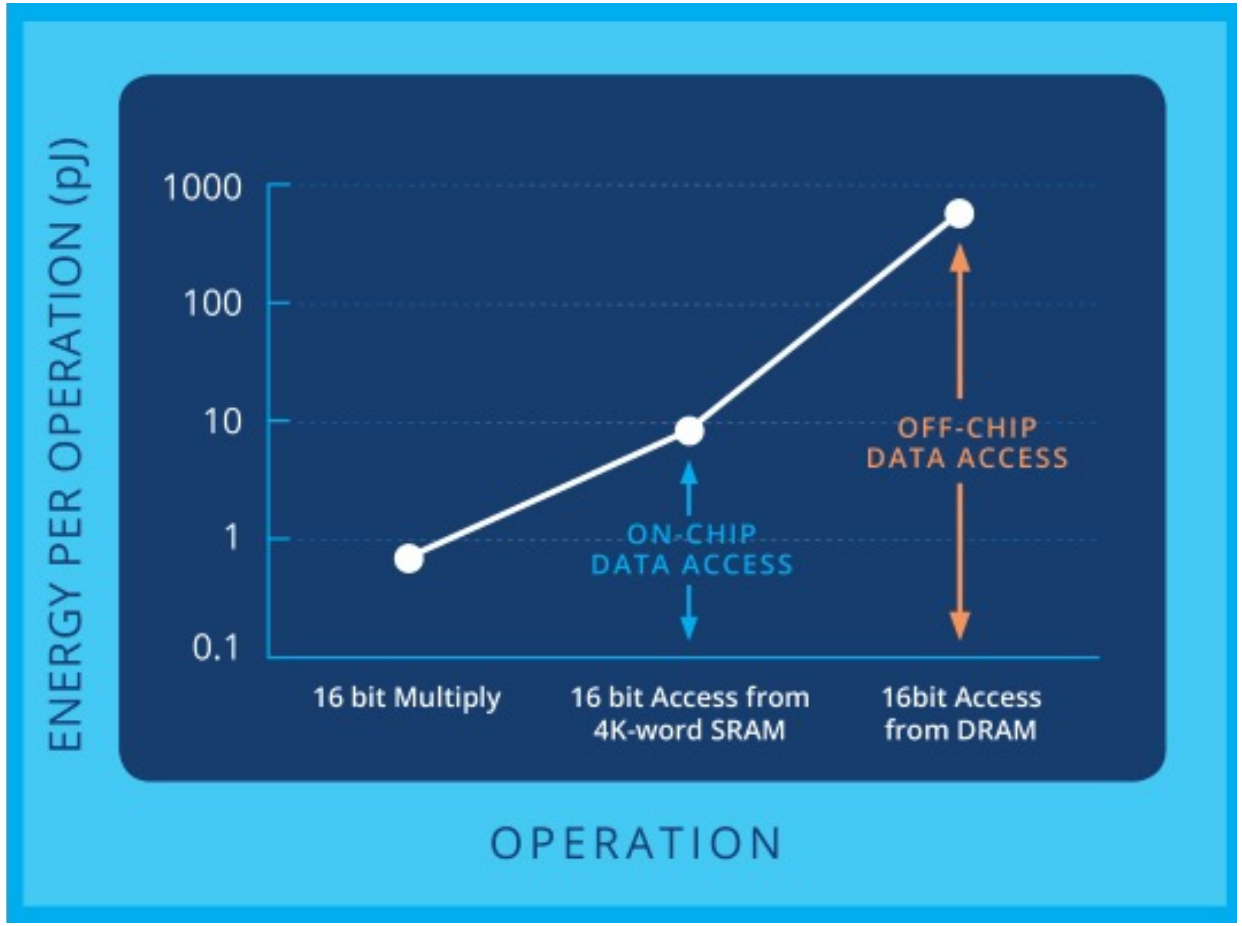
- Also called ”Near memory computing”
- Bring computation close to the data (i.e., memory)
- Low latency
- High throughput
- Energy efficiency

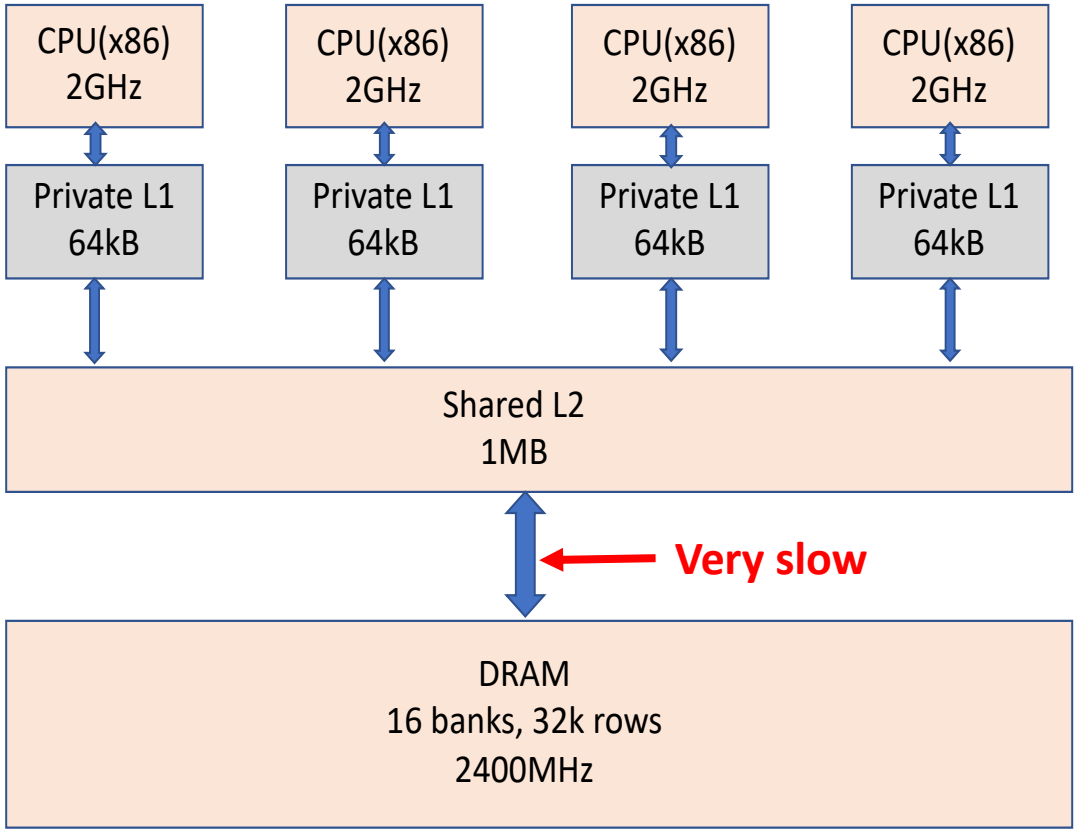


B. K. Joardar, P. Ghosh, P. P. Pande, A. Kalyanaraman, and S. Krishnamoorthy. 2019. NoC-enabled software/hardware co-design framework for accelerating k-mer counting. In Proceedings of the 13th IEEE/ACM International Symposium on Networks-on-Chip (NOCS ‘19). Association for Computing Machinery, New York, NY, USA, Article 4, 1–8. DOI:https://doi.org/10.1145/3313231.3352367

A. Pattnaik et al., "Scheduling techniques for GPU architectures with processing-in-memory capabilities," 2016 International Conference on Parallel Architecture and Compilation Techniques (PACT), 2016, pp. 31-44, doi: 10.1145/2967938.2967940.

Why PIM?


ENERGY PER OPERATION (pJ)
1000
100
10
1
0.1
ON-CHIP DATA ACCESS
OFF-CHIP DATA ACCESS
16 bit Multiply
16 bit Access from 4K-word SRAM
16bit Access from DRAM
OPERATION


CPU(x86) 2GHz
CPU(x86) 2GHz
CPU(x86) 2GHz
CPU(x86) 2GHz
Private L1 64kB
Private L1 64kB
Private L1 64kB
Private L1 64kB
Shared L2 1MB
Very slow
DRAM
16 banks, 32k rows
2400MHz

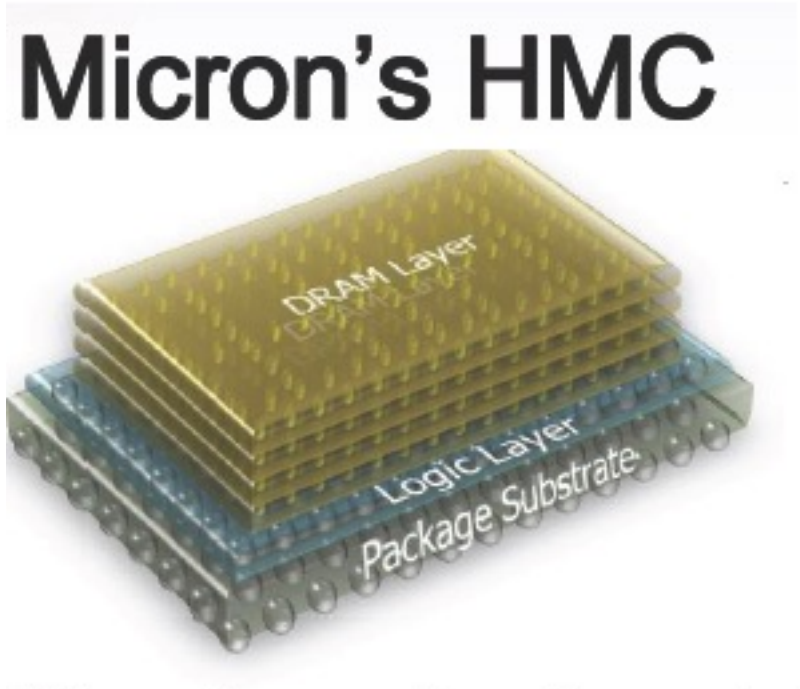
Why 3D PIM?

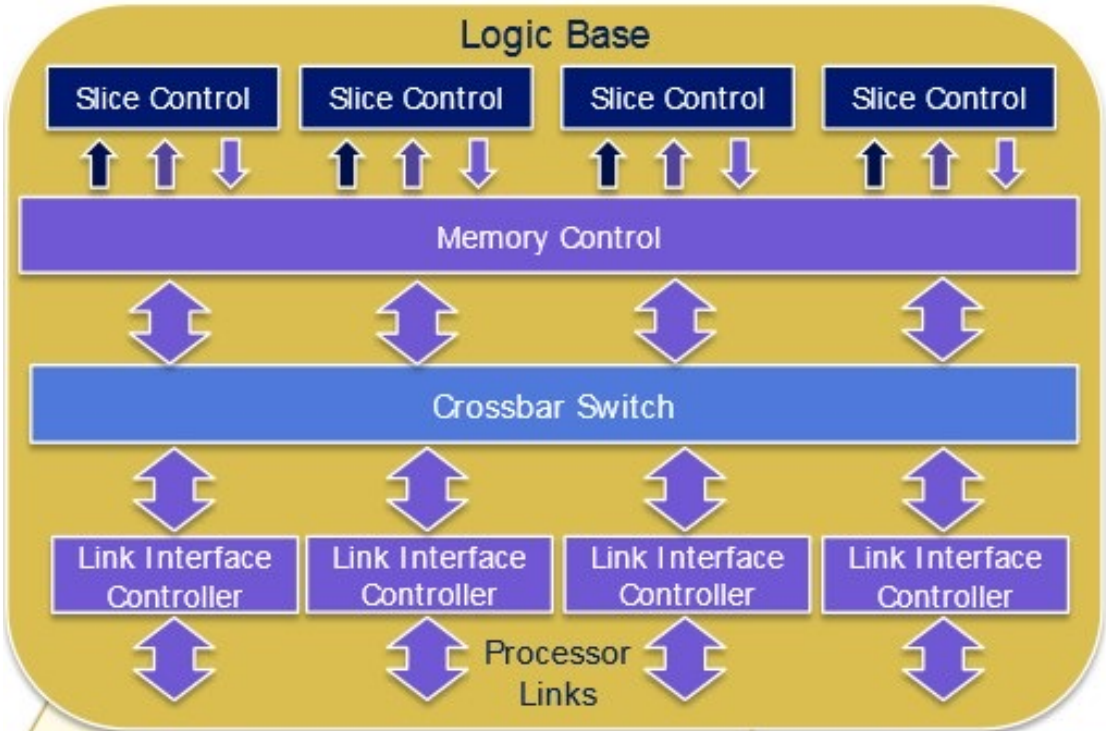
- **2.6X faster memory access**

HOST: L2 cache refill latency - (Size: 256 B) [Total: 102.3 ns]			PIM: 4B read access latency (no caches) [Total: 39.1 ns]		
Membus	1 Cycle@2 GHz	Flit:64 b	PimBus	1 Cycle@1 GHz	Flit:32 b
SMCController	8 Cycles@2 GHz	Pipeline Latency [21]	SMCXbar	1 Cycle@ 1GHz	Flit:256 b [8]
SERDES	1.6 ns	SER = 1.6 DES = 1.6 [14]	Vault Ctrl. Front-end	4 Cycles@1.2 GHz	[8]
Packet Transfer	13.6 ns	16 Lanes@10 Gb/s, 128 b hdr [2]	tRCD	13.75 ns	Activate [14]
PCB Trace Latency	3.2 ns + 3.2 ns	Round Trip	tCL	13.75 ns	Issue Read Com-mand [14]
SMCXBar	1 Cycle@1 GHz	Flit:256 b [8]	tBURST	3.2 ns	1 Beat [19]
VaultCtrl.frontend	4 Cycles@1.2 GHz	[8]	Vault Ctrl. Back-end	4 Cycles@1.2 GHz	[8]
tRCD	13.75 ns	Activate [14]	PimBus	1 Cycle@1 GHz	Flit:32 b
tCL	13.75 ns	Issue Read Command [8]			
tBURST	25.6 ns	256 B Burst [19]			
VaultCtrl.backend	4 Cycles@1.2 GHz	[8]			
SERDES	1.6 ns	SER = 1.6 DES = 1.6 [14]			
Packet Transfer	13.6 ns	16 Lanes@10 Gb/s, 128 b hdr [2]			
SMCController	1 Cycles@2 GHz	Pipeline Latency [21]			
Membus	1 Cycle@2 GHz	Flit:64 b			

Erfan Azarkhish, Davide Rossi, Igor Loi, and Luca Benini. 2016. Design and Evaluation of a Processing-in-Memory Architecture for the Smart Memory Cube. In Proceedings of the 29th International Conference on Architecture of Computing Systems -- ARCS 2016 - Volume 9637. Springer-Verlag, Berlin, Heidelberg, 19-31

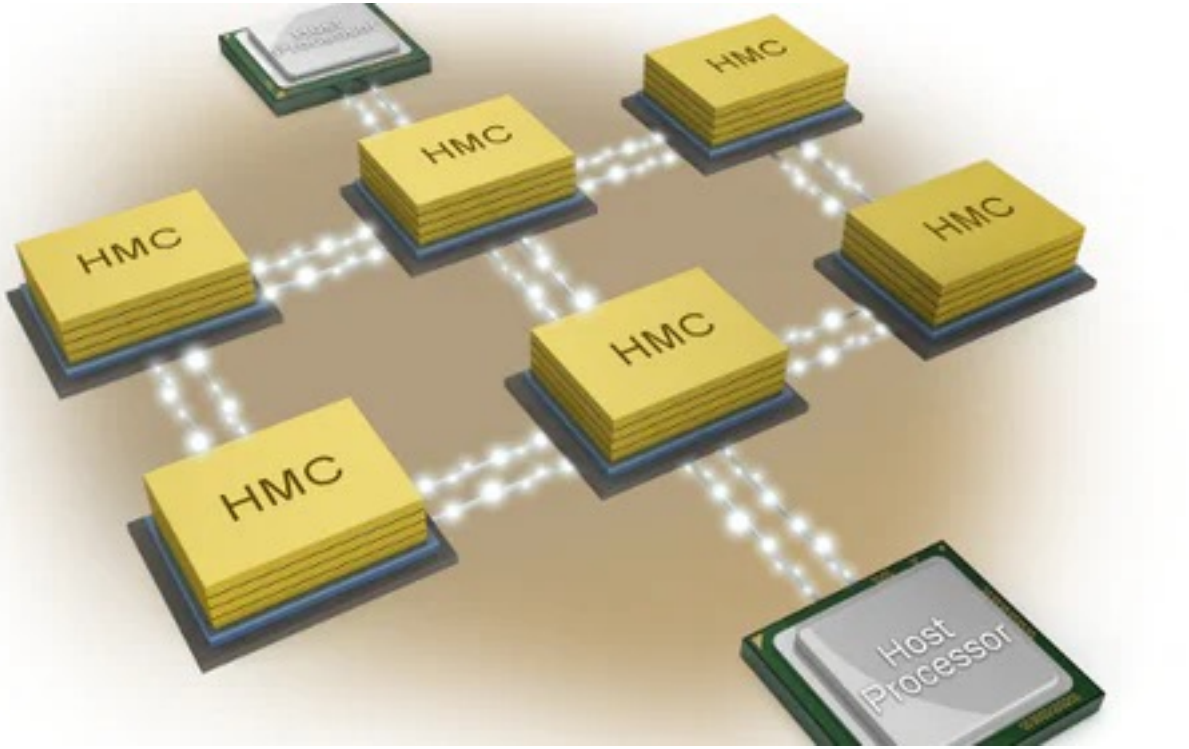
Micron’s HMC





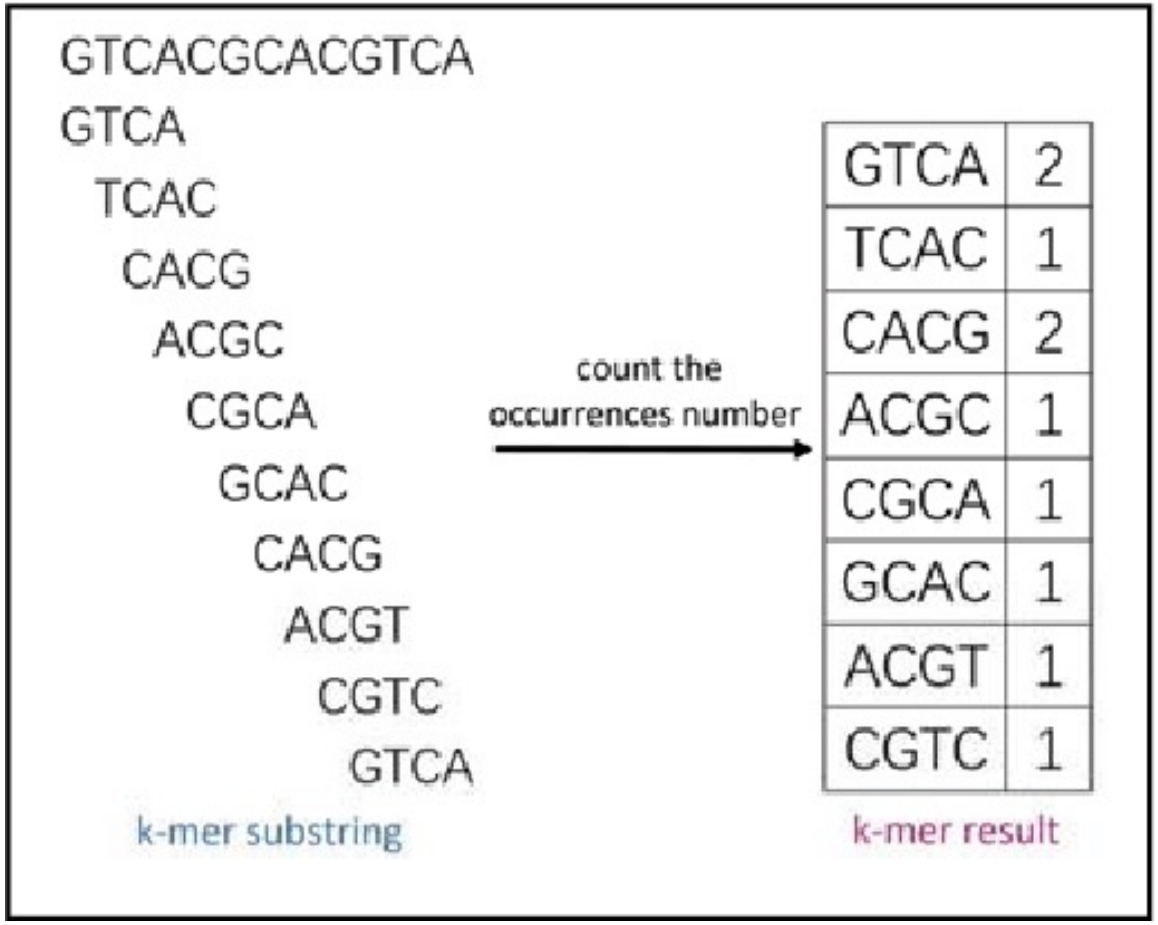
- HMC has one logic layer with many DRAM layers on top
- Logic layer includes memory access logic
- Logic layer is simple to avoid high temperature

Connecting HMCs



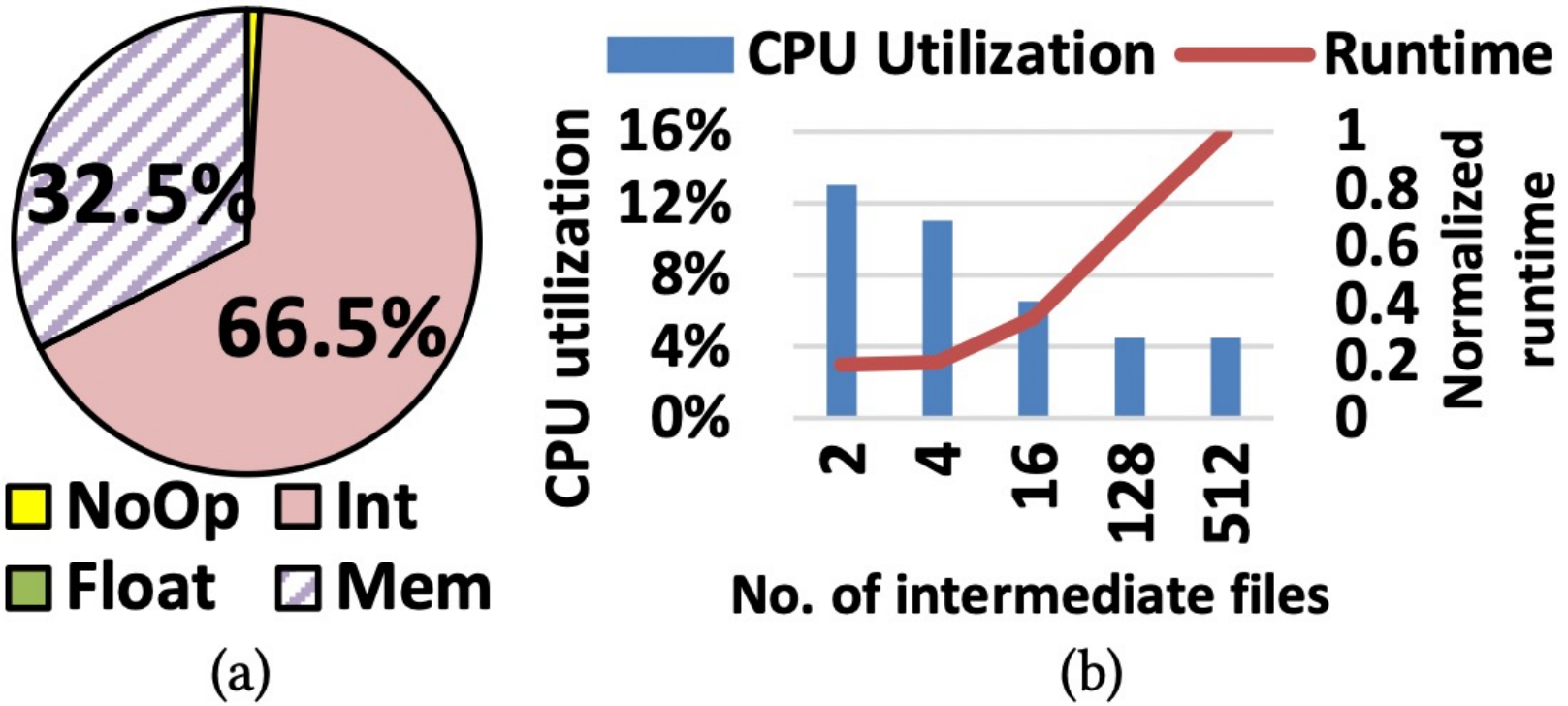
- HMCs can be connected in many ways to create new architectures
- These external links are slower than intra-HMC links
- Mapping, storing data is important for high performance

K-mer counting



- **K-mer = string of length k**
- **Goal: Given a genetic sequence, prepare a histogram of all possible k-mers**

Heavy memory access



(a) (b)

Fig. 2: Gerbil: (a) Instruction types, and (b) CPU utilization and runtime (normalized) for varying number of intermediate files

Effect of PIM

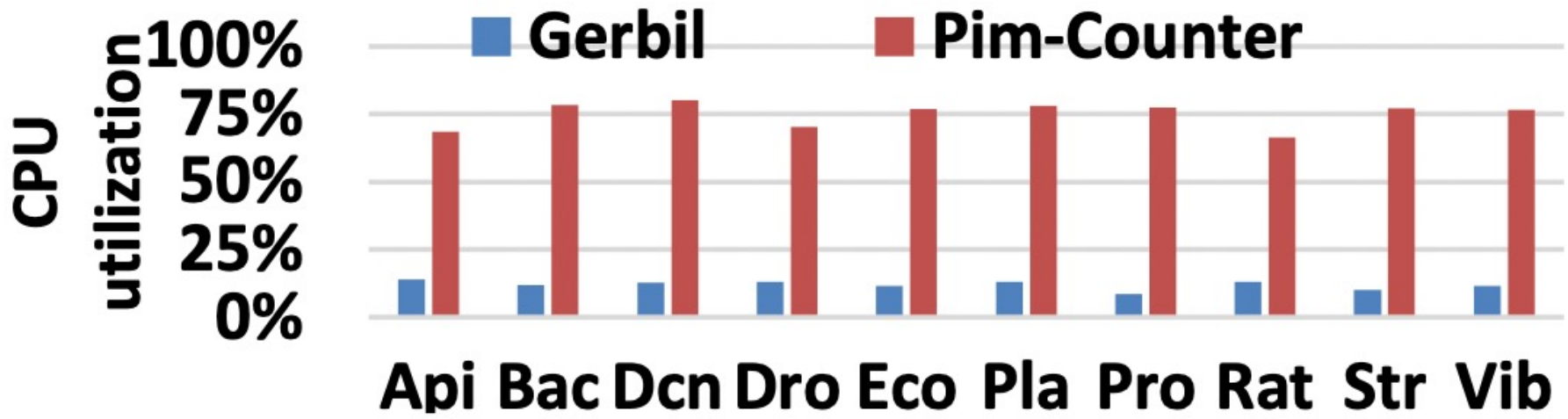
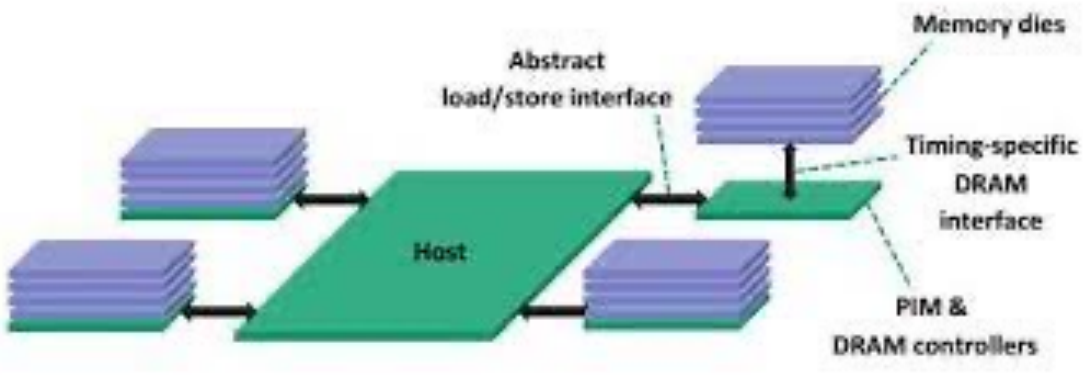


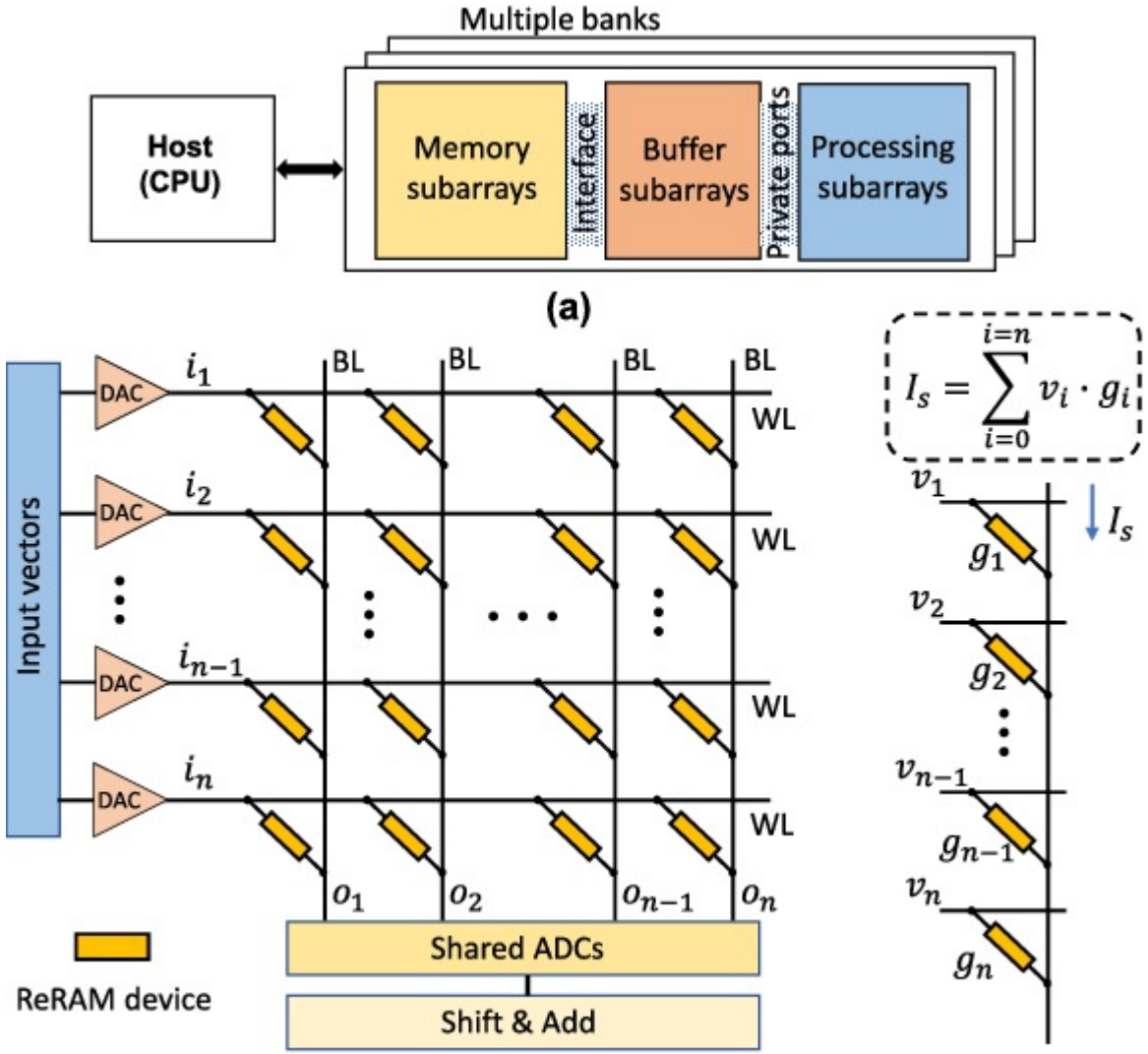
Fig 8: Average CPU utilization in Gerbil and PIM-Counter

- PIM can accelerate memory-intensive applications (like k-mer counting)
- CPU utilization improves significantly
- Up to 4X improvement

Other PIM variants



2.5D PIM

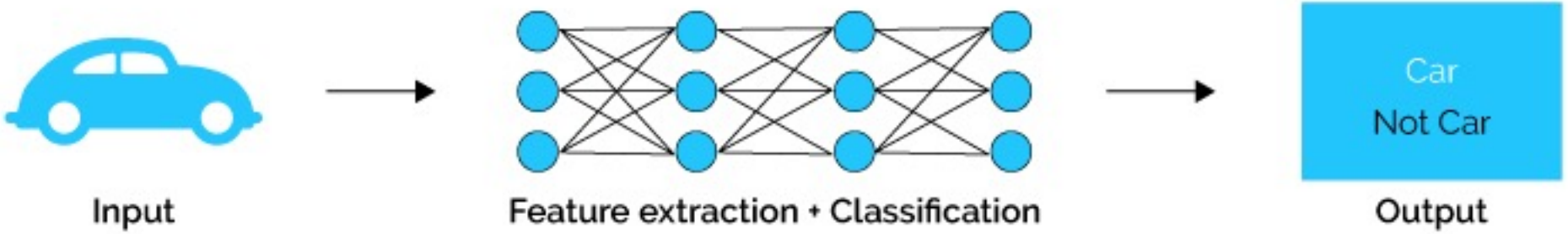


Non-Volatile memory

Other applications suited for PIM

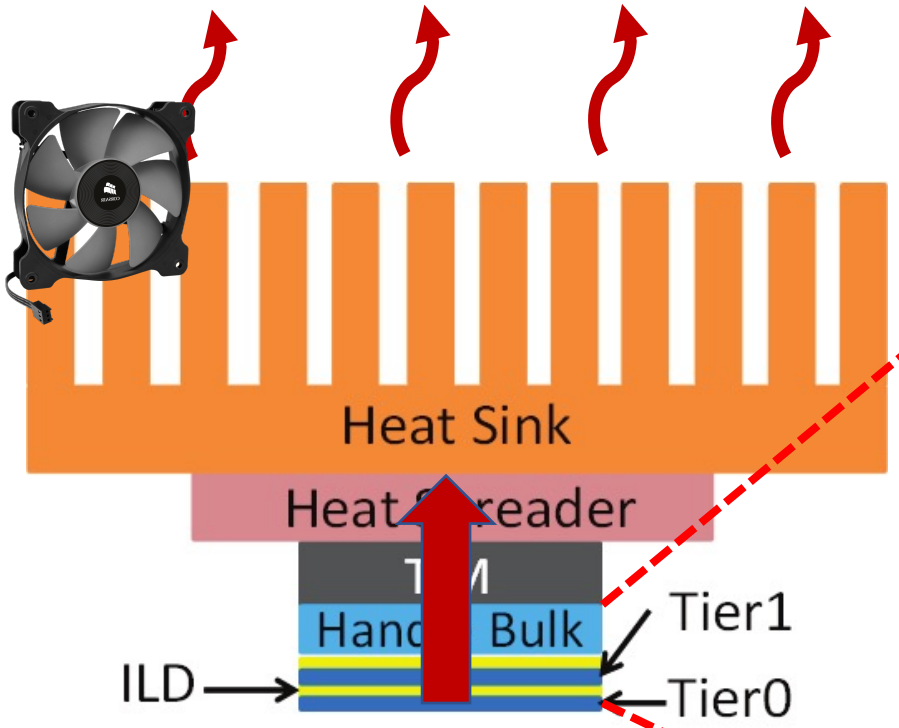
Genetics

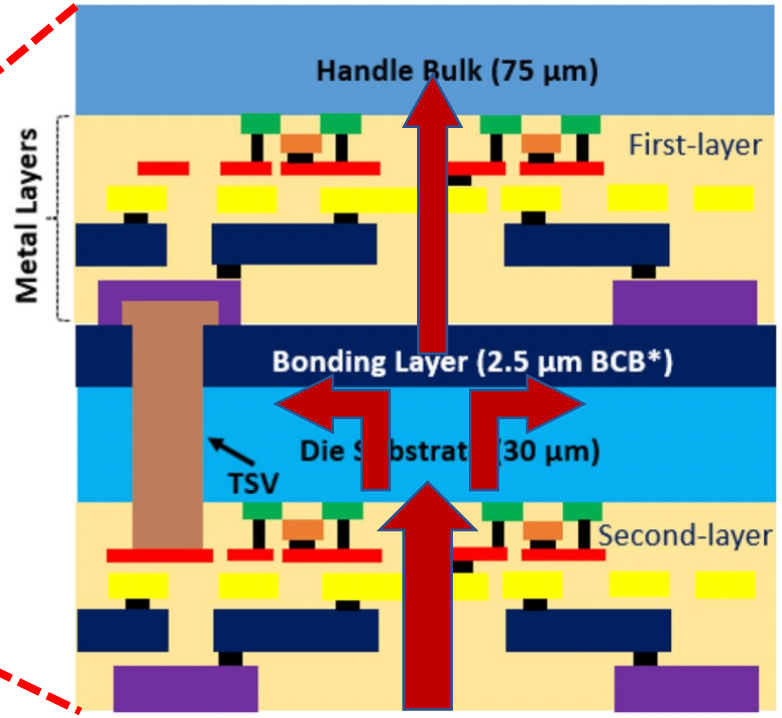
Graph Analytics



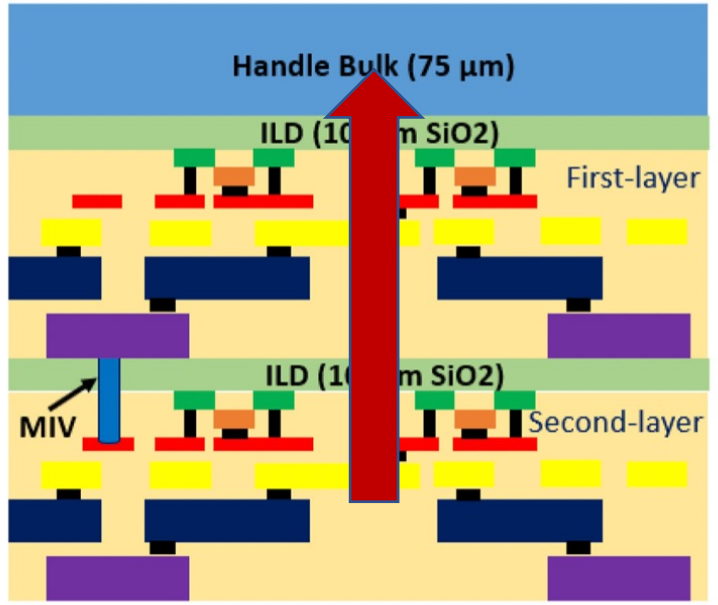
Machine Learning

Temperature Dissipation in 3D





TSV-based 3D IC



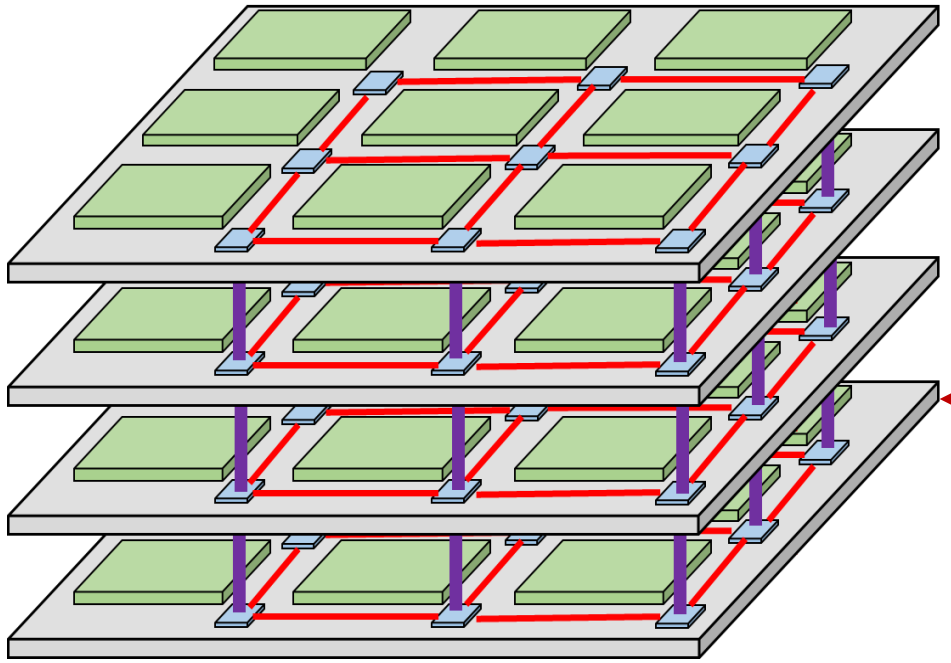
Monolithic 3D IC

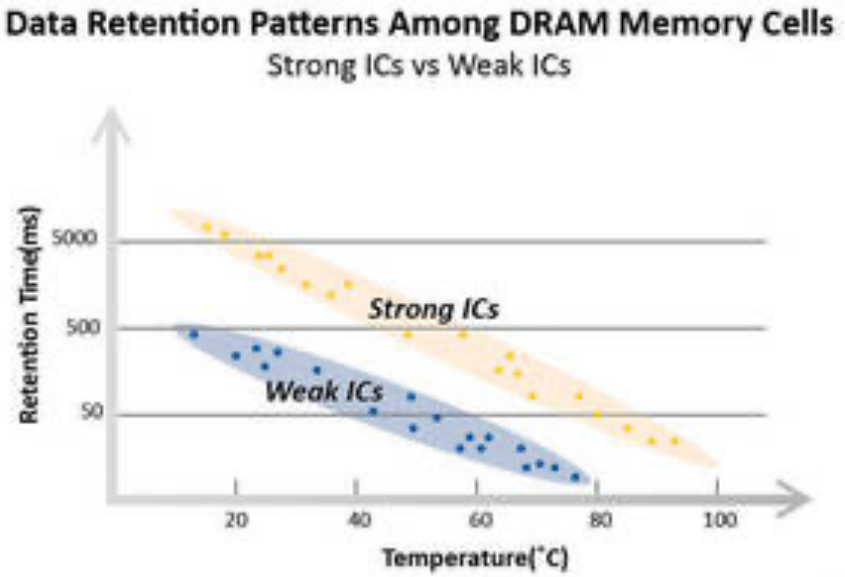
- **Heat dissipation is a problem in 3D architectures**
- **Long distance to heat sink**

Dongjin Lee, Sourav Das, Janardhan Rao Doppa, Partha Pratim Pande, and Krishnendu Chakrabarty. 2018. Performance and Thermal Tradeoffs for Energy-Efficient Monolithic 3D Network-on-Chip. ACM Trans. Des. Autom. Electron. Syst. 23, 5, Article 60 (October 2018), 25 pages

Samal, Sandeep & Panth, Shreepad & Samadi, Kambiz & Saedi, Mehdi & Du, Yang & Lim, Sung. (2014). Fast and Accurate Thermal Modeling and Optimization for Monolithic 3D ICs. Proceedings - Design Automation Conference. 10.1145/2593069.2593140.

Thermal in 3D manycore design





- Layers far from the sink are hotter
- DRAM is sensitive to heat
- High heat can damage the chip
- Can we address this by task mapping?

Heat map for 3D architectures

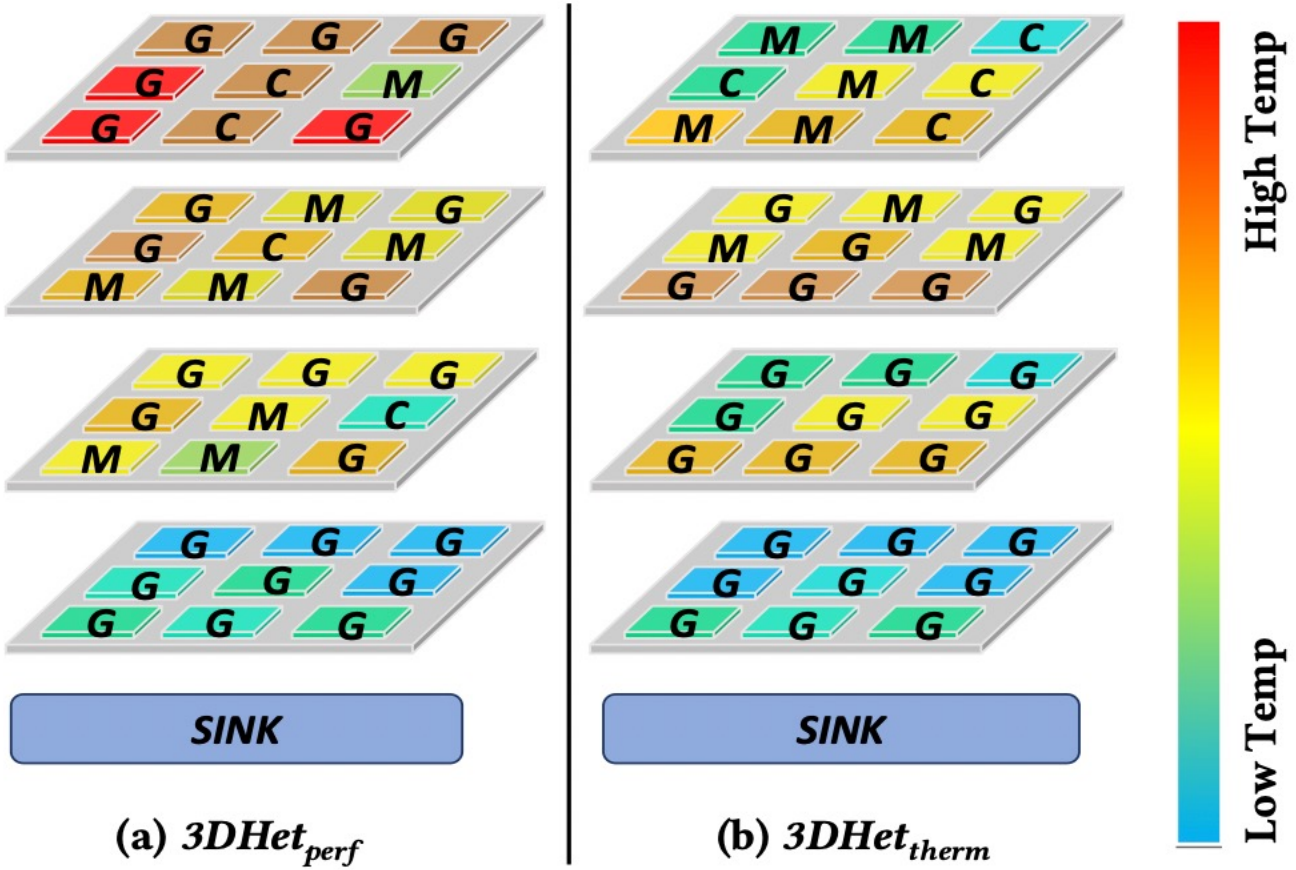
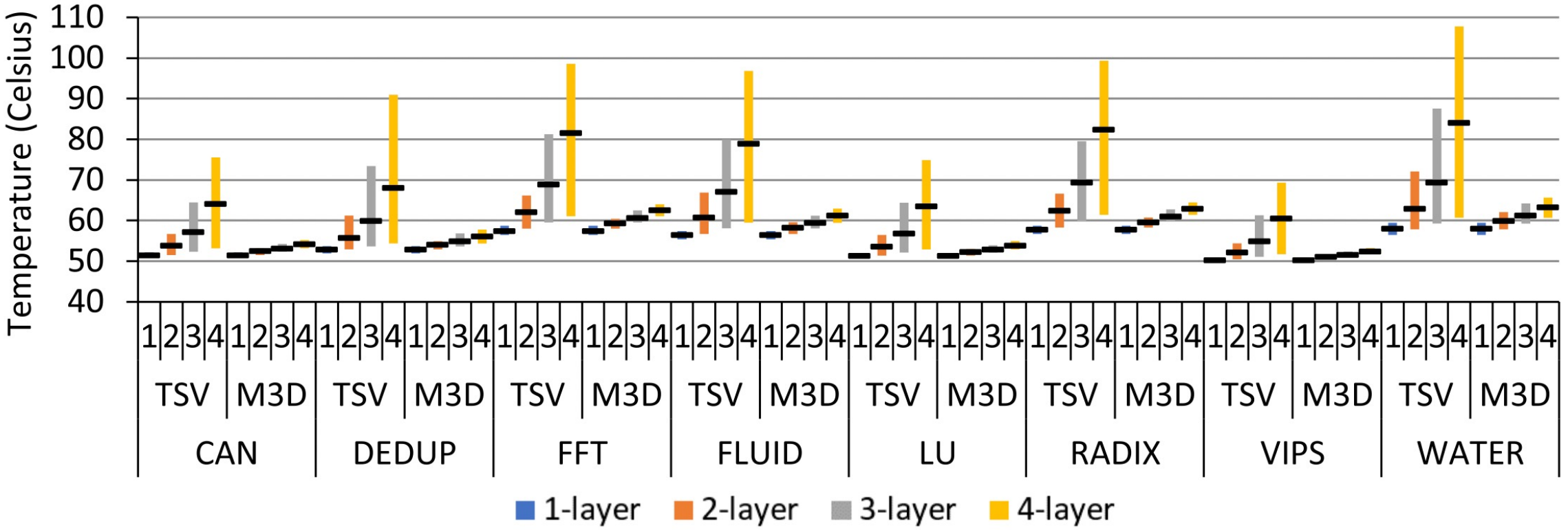


Figure 9: Tile placement and thermal map for (a) *3DHe* (b) *3DHet$_{therm}$* (C: CPU core, G: GPU core, M: MC).

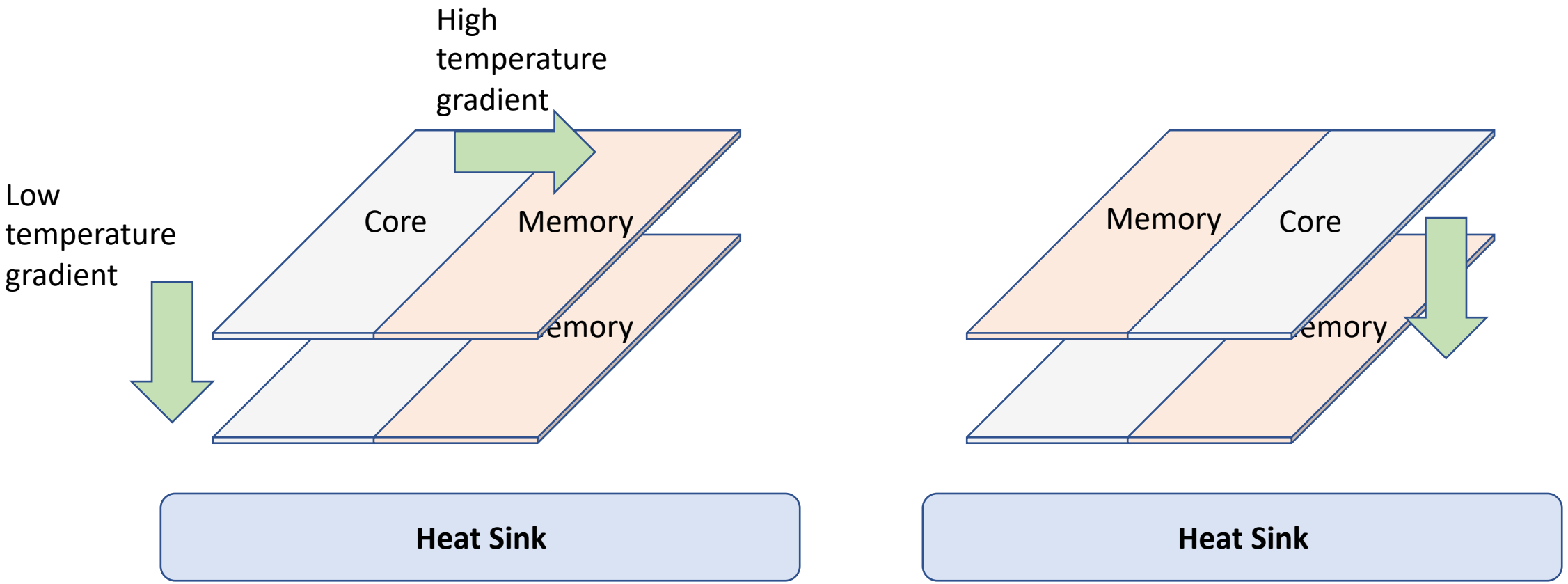
Temperature in 3D



- **Layer farthest from sink is hottest**
- **Higher temperature for TSVs**

Dongjin Lee, Sourav Das, Janardhan Rao Doppa, Partha Pratim Pande, and Krishnendu Chakrabarty. 2018. Performance and Thermal Tradeoffs for Energy-Efficient Monolithic 3D Network-on-Chip. ACM Trans. Des. Autom. Electron. Syst. 23, 5, Article 60 (October 2018), 25 pages

RDS floorplanning (1)



- **Cores dissipate more power than memory**
  - **Cores are hotter**
- **Increase temperature gradient towards sink**

RDS floorplanning (2)

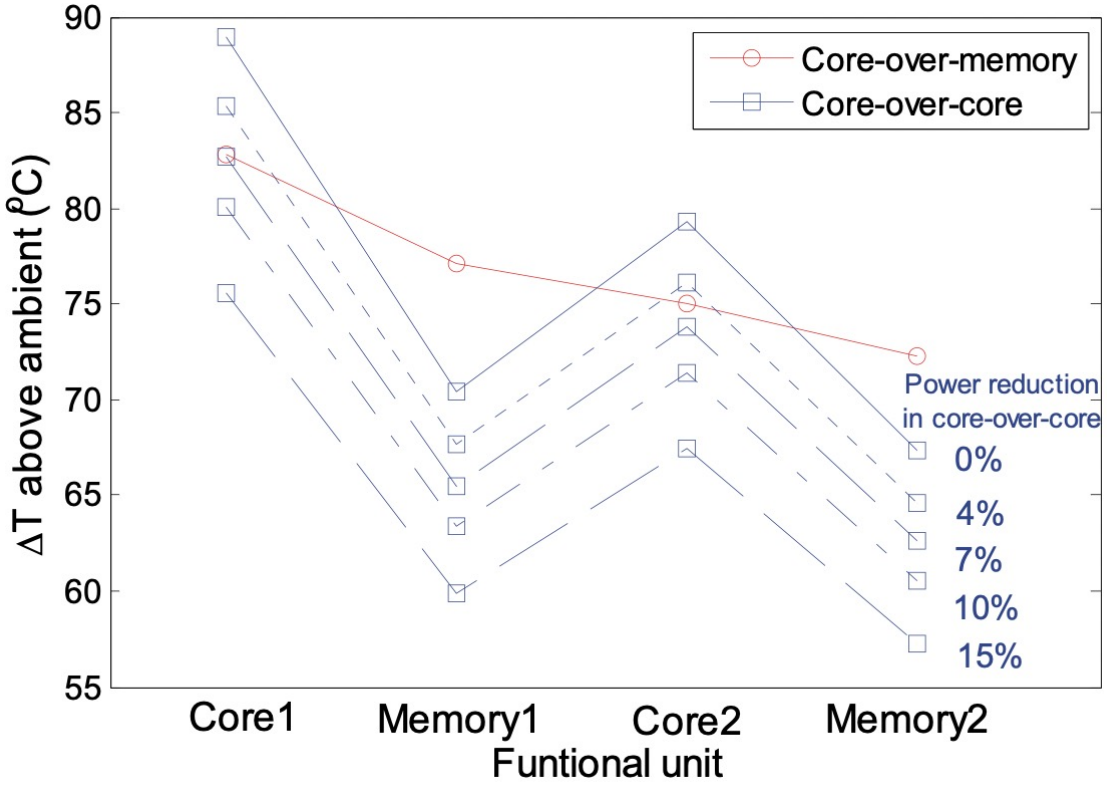
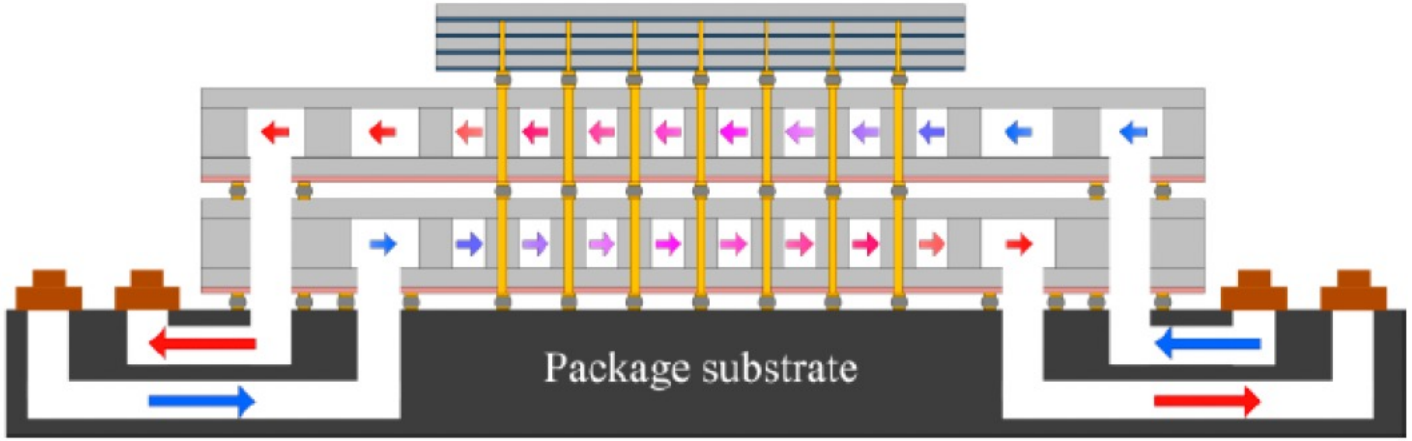


Figure 9: Temperature rise in core-over-core and core-over-memory configurations for the 3D dual-core processor with a power dissipation of 12 watts per die (equivalently in one core and one memory) and 80% of power dissipated in the cores.

- **Temperature reduced by floor planning only**
  - **Max temperature**
- **No architectural changes made**
- **Can be used to design cooler 3D manycore systems**

S. M. Alam, R. E. Jones, S. Pozder and A. Jain, "Die/wafer stacking with reciprocal design symmetry (RDS) for mask reuse in three-dimensional (3D) integration technology," 2009 10th International Symposium on Quality Electronic Design, 2009, pp. 569-575, doi: 10.1109/ISQED.2009.4810357.

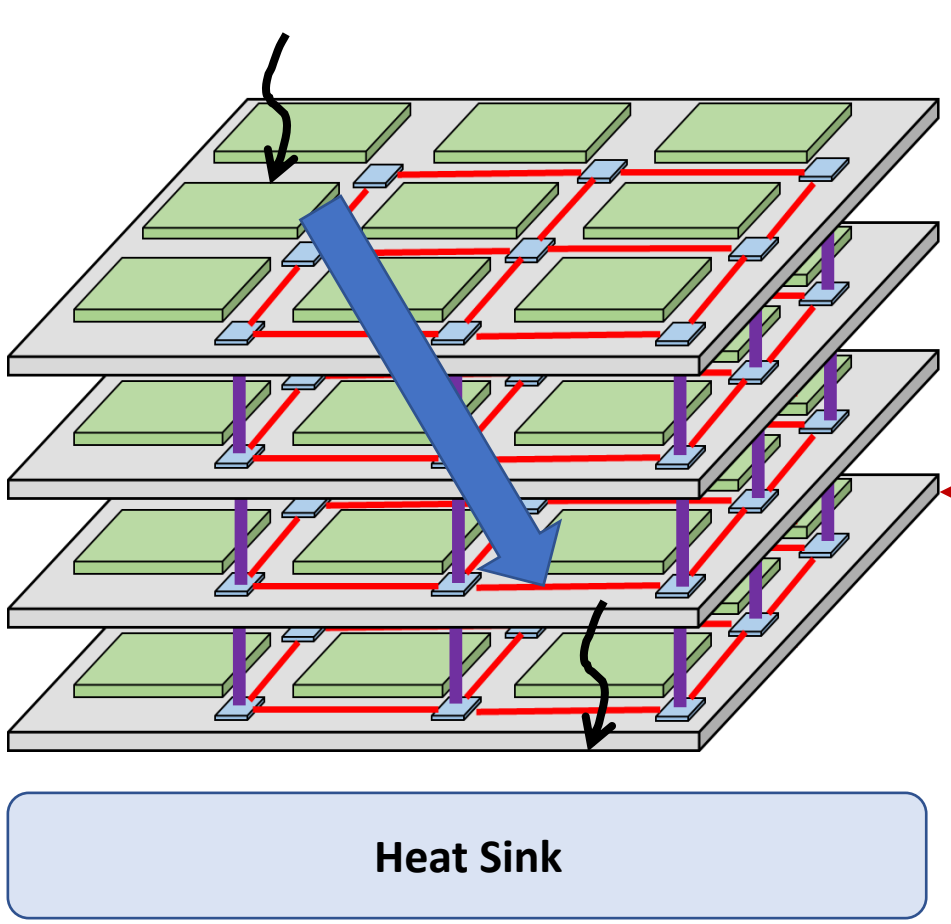
Micro-fluidic cooling



- **Liquid based cooling instead of air cooling**
- **Liquid moves inside the 3D chip**

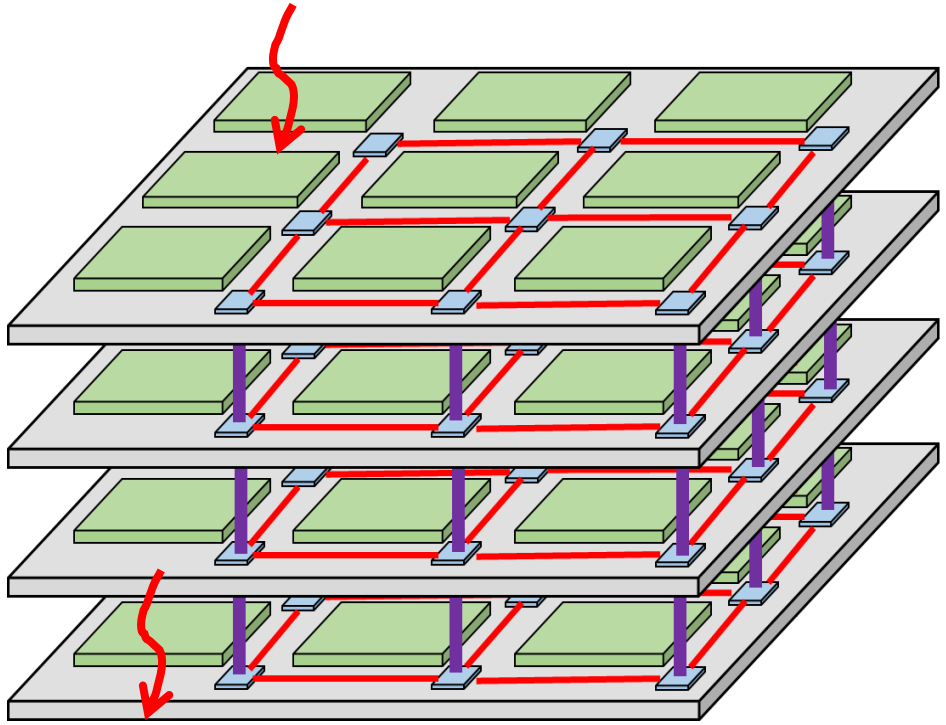
Zheng, L.; Zhang, Y.; Bakir, M.S. Design, fabrication and assembly of a novel electrical and microfluidic I/Os for 3-D chip stack and silicon interposer. In Proceedings of the Electronic Components and Technology Conference, Las Vegas, NV, USA, 28–31 May 2013.

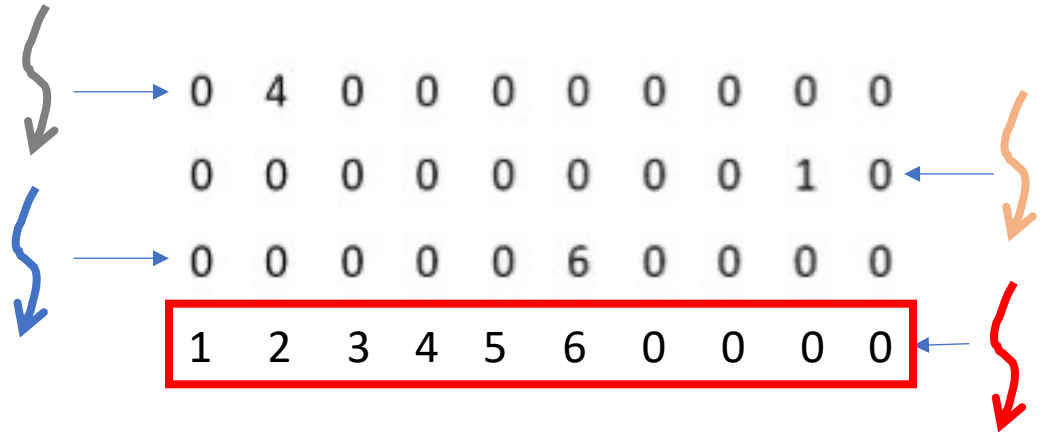
Reducing temperature by task migration



- Move task to cooler region
- Performance overheads and not always feasible
  - Uniform utilization

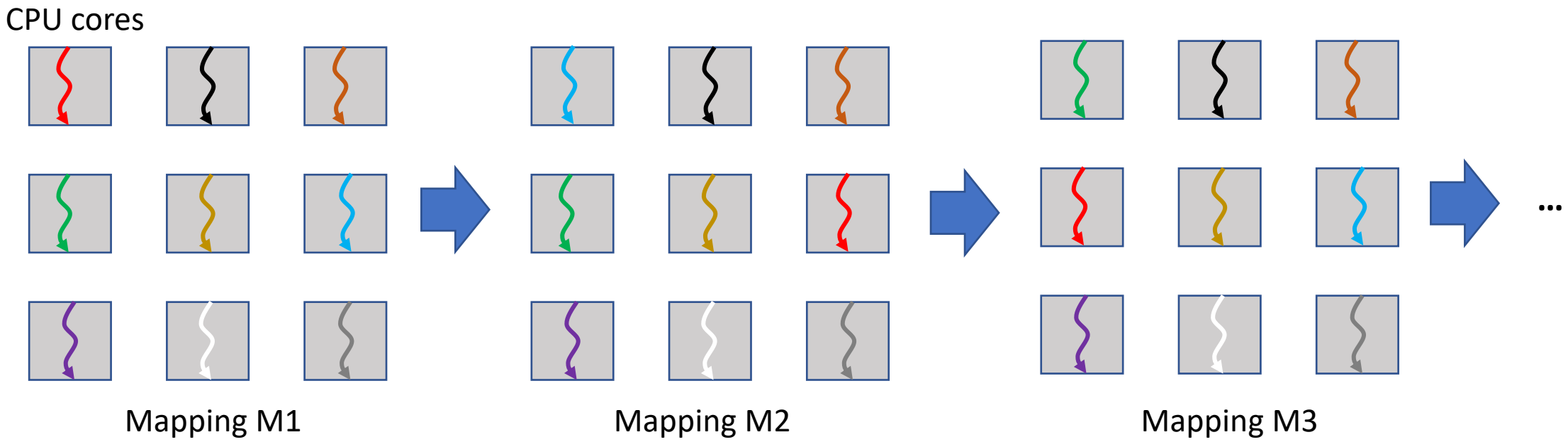
Temperature-aware mapping (1)

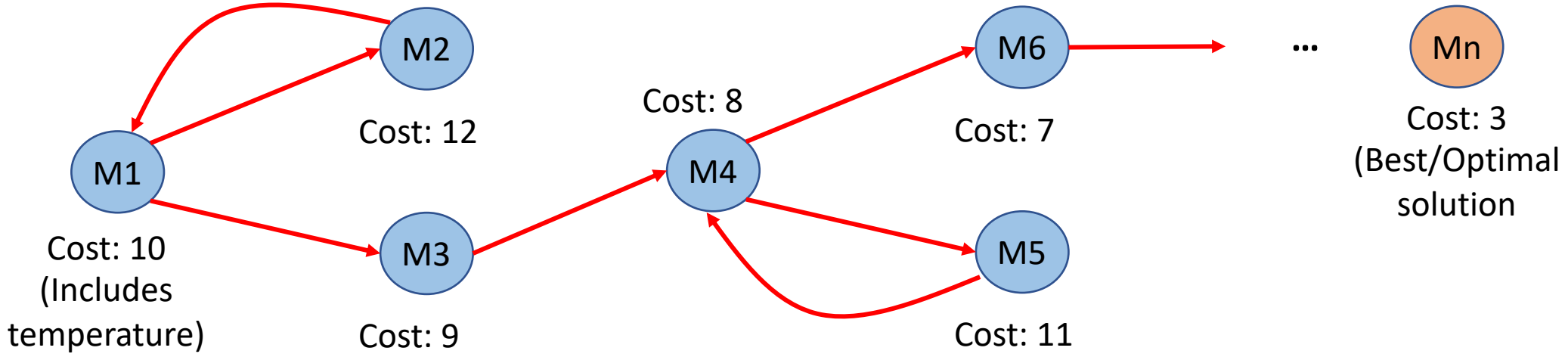




- **Map compute-intensive tasks near the sink**
  - **These tasks generate more heat**
  - **Easily dissipated due to sink**

Temperature-aware mapping





Effect of task mapping & floor planning on temperature

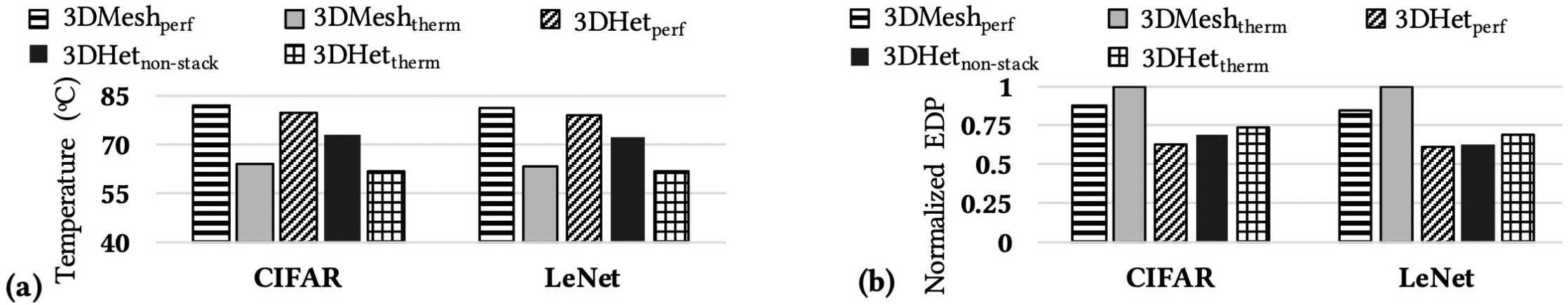


Figure 7: Temperature profile and EDP for different NoCs: (a) Maximum system temperature and (b) network EDP.

- **By suitably mapping tasks to cores, we can reduce temperature**
- **High-power consuming tasks should be mapped near the sink**
- **Up to 15C cooler**

B. K. Joardar et al., "3D NoC-enabled heterogeneous manycore architectures for accelerating CNN training: Performance and thermal trade-offs," 2017 Eleventh IEEE/ACM International Symposium on Networks-on-Chip (NOCS), 2017, pp. 1-8.