

# PDSC2: Introduction to High Bandwidth Memory

Marc Greenberg, Principal/CEO  
Marc Greenberg Consulting LLC  
[marc@marcgreenberg.com](mailto:marc@marcgreenberg.com)

# Abstract

In this Professional Development Series session, you'll learn about key aspects of High Bandwidth Memory (HBM): What is HBM, a short history of HBM, why is HBM important right now, how Large Language Models (LLMs) and Generative AI are driving demand for HBM technology, comparison of HBM with other popular memory types (DDR, LPDDR and GDDR), a high level view of HBM architecture, PCB and package requirements to implement chips deploying HBM, a view of the market for HBM and the chips that use it, and a review of public information on the future development of high bandwidth memories.



# Brief Bio and Disclosures

- Past Endeavors:

- Group Director, Product Marketing, Cadence, 2017-2023
- Director, Product Marketing, Synopsys, 2012-2017
- Director, Product Management, Cadence, 2010-2012
- Director, Technical Marketing, Denali Software, Inc. 2003-2010
- Various roles including IP Procurement Manager, Motorola Semiconductor (SPS) 1993-2003

} All focused on  
Memory and Storage

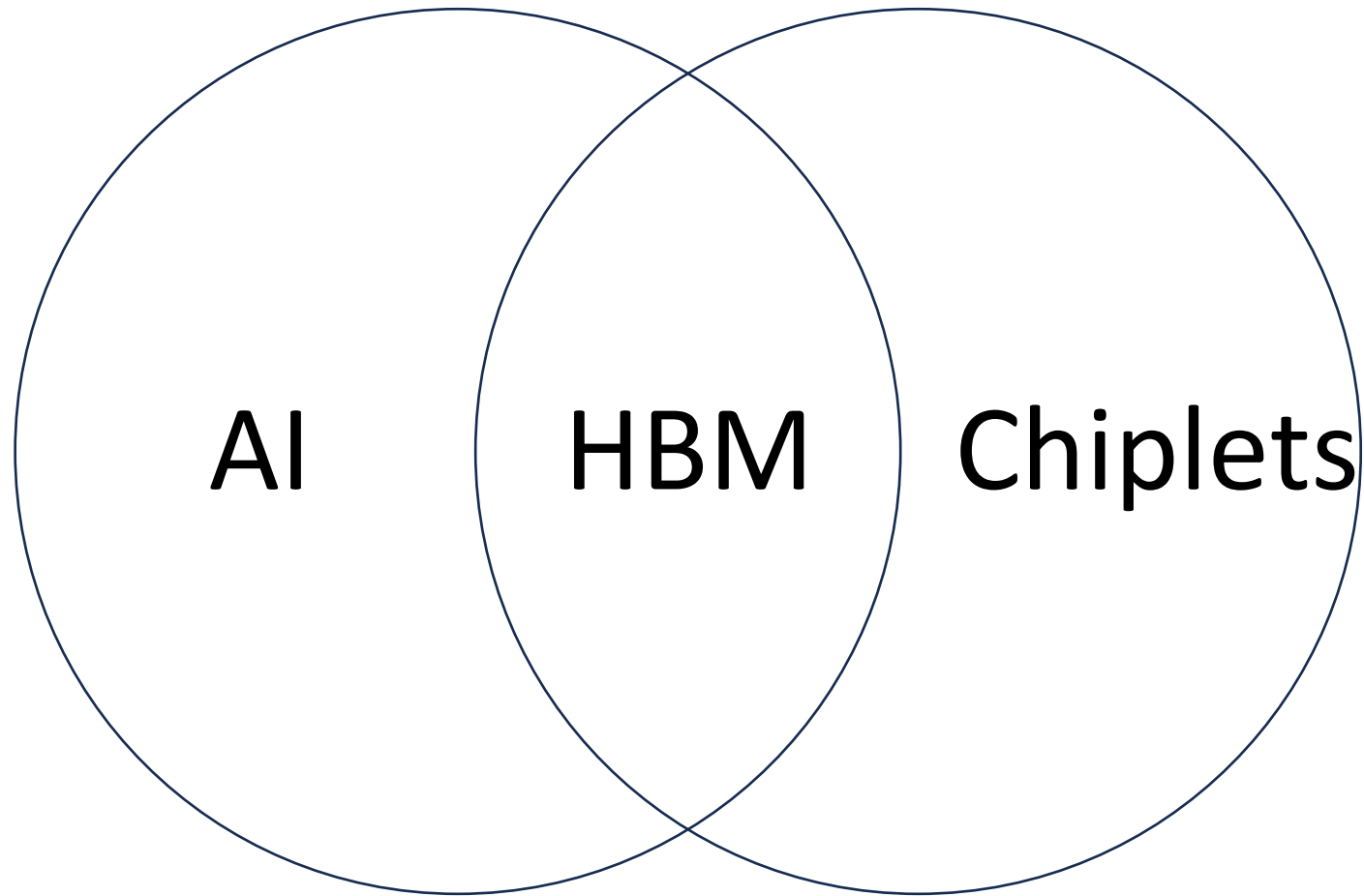
- Current Endeavors:

- VP Product, Cassia.ai
- Director, Strategic Alliances, Blue Cheetah
- Vice-Chair, non-public JEDEC Task Group
- Consultant, The Six Semiconductor
- Leading an undisclosed AI project as part of Marc Greenberg Consulting LLC
- Occasional advisor to investment firms etc
- I own stock in Cadence and Cassia.ai, and I own Marc Greenberg Consulting LLC

All material presented here is my own opinion and does not necessarily represent the position or opinion of any of my clients or any third party



# Why are we here?



AI



# Where we were vs where we're going

Please confirm your identity

Photo 3 of 5



2011 Facebook  
facial recognition  
attempt

This appears to be:

- Owen [redacted]
- Colin [redacted]
- Craig [redacted]
- Troy [redacted]
- Gordon [redacted]
- Greg Th [redacted]



2024 OpenAI Sora show reel

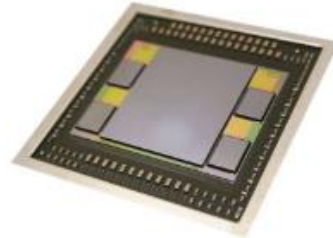
<https://youtu.be/2fAPgOCjToA?si=YtHCU1bytGEKhdwe>



# What is HBM used for?

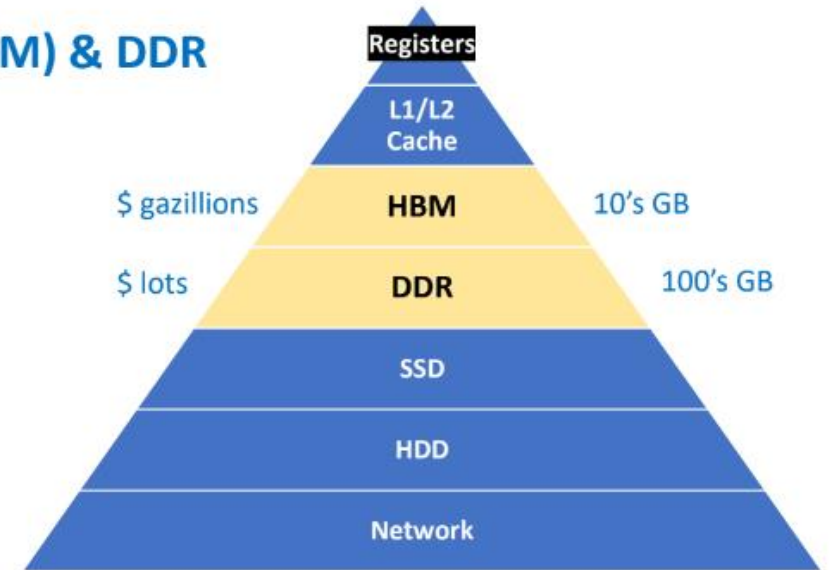
- L4 cache (picture at right)
- Streaming buffer in networking applications
- AI Accelerator

## High Bandwidth Memory (HBM) & DDR



It's not either/or

It's in addition to



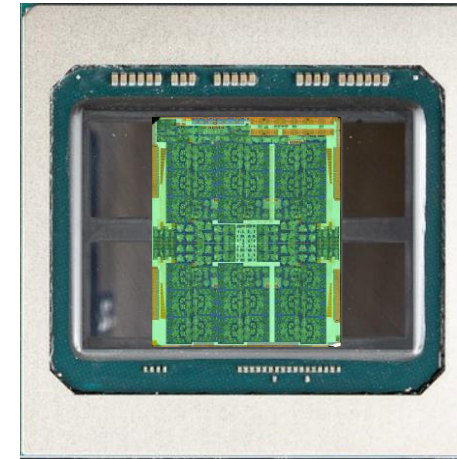
Source: Bill Gervasi, an hour or two ago



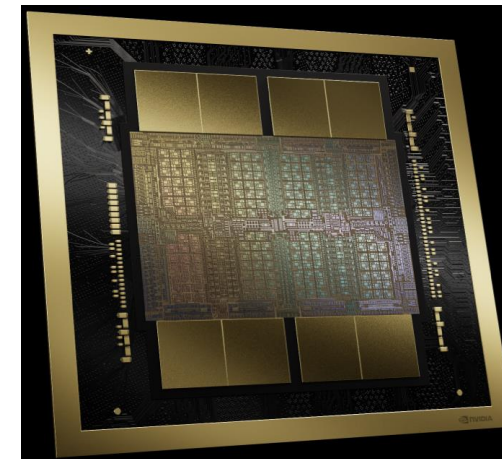
# Introduction to AI chips (more to come later)

- The chips that operate most of the big server-based AI hardware are giant math machines
- Parallelism, specialization and a shift in the types of AI that they run has enabled very large computing machines to be developed

<https://images.nvidia.com/content/pdf/tesla/whitepaper/pascal-architecture-whitepaper.pdf>  
<https://resources.nvidia.com/en-us-blackwell-architecture?ncid=no-ncid>  
<https://www.flickr.com/photos/130561288@N04/albums/72177720295479734/with/51867067870>



NVIDIA P100  
2016  
16nm  
60 SM units  
Quad HBM2  
21 TFLOPs (FP16)  
700w TDP



NVIDIA Blackwell  
2024  
4nm  
144 SM units  
576 Tensor Cores  
Dual-quad HBM3E  
2500 TFLOPs (FP16)  
700w TDP



# The compute demand for AI is insatiable

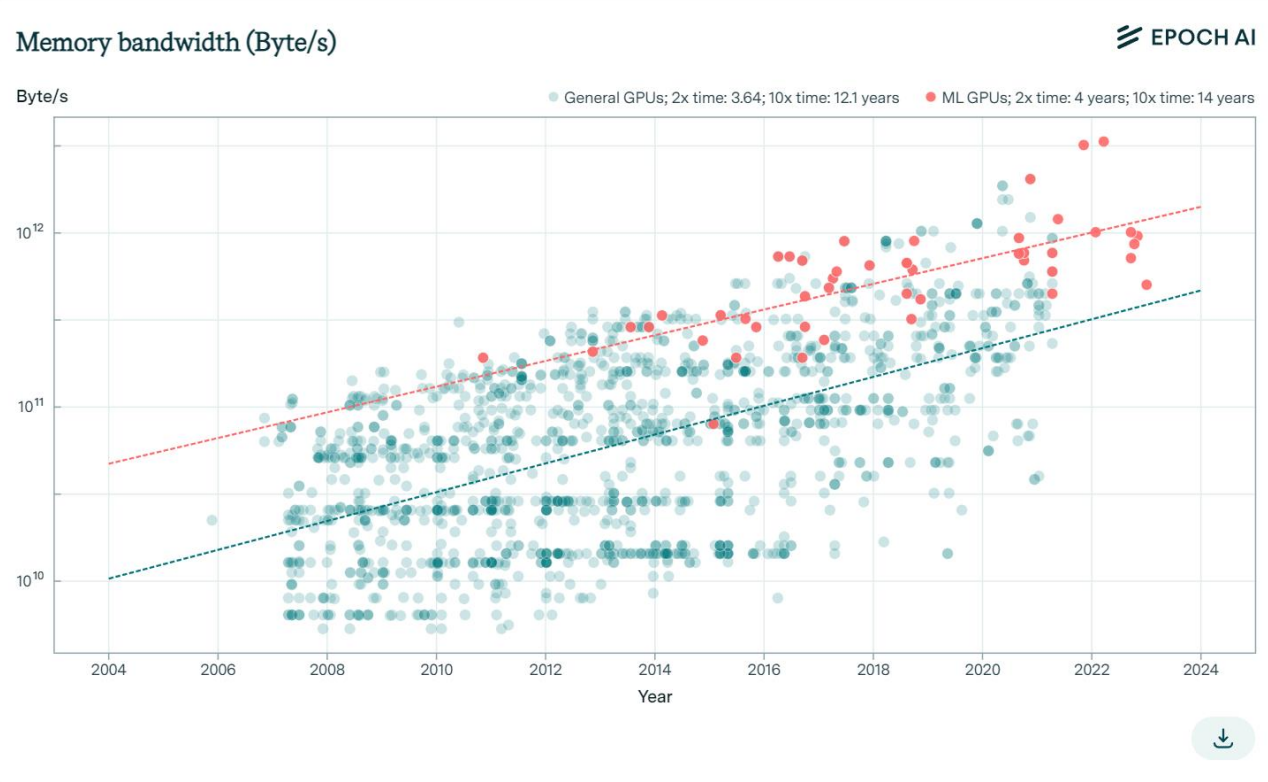
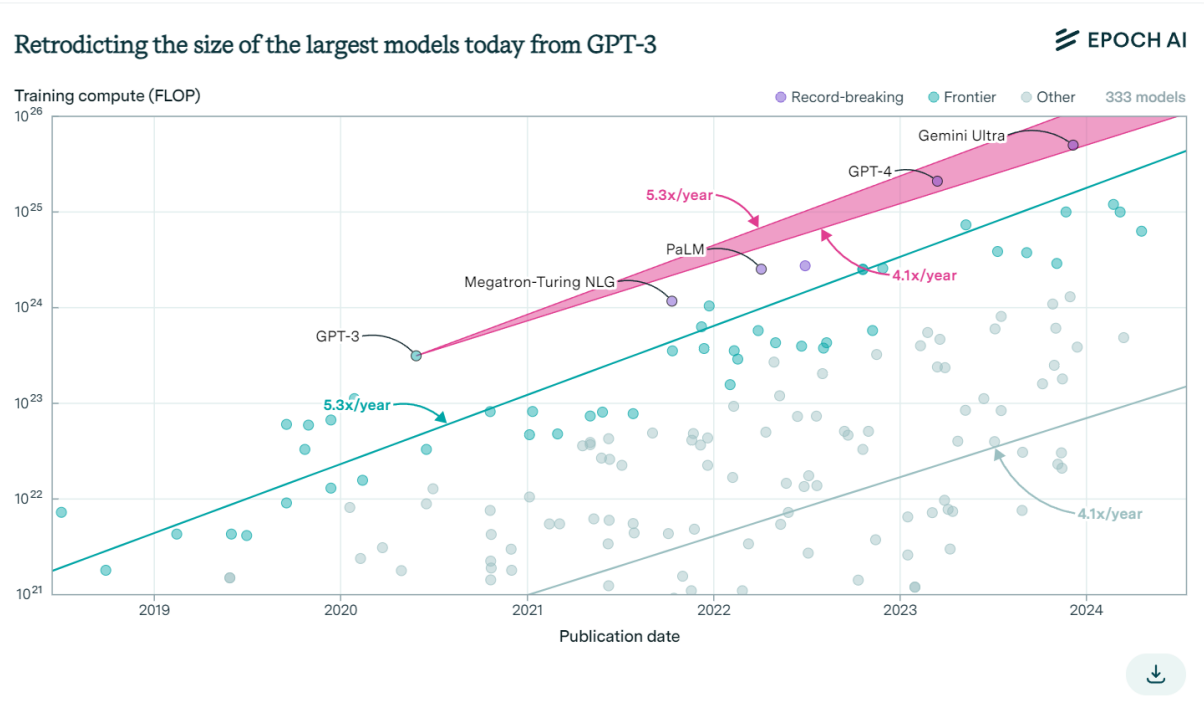


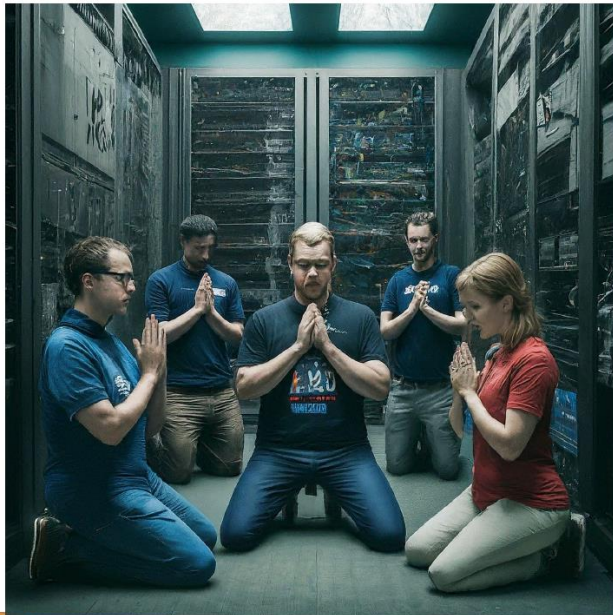
Figure 6: The historical trend for all and frontier models is used to extrapolate the training compute of GPT-3 to predict the compute of the largest models today. Note that many record-breaking systems lie close to the extrapolated line, including Megatron-Turing NLG 540B, PaLM 540B, GPT-4, and Gemini Ultra.



<https://epochai.org/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year>  
<https://epochai.org/blog/trends-in-machine-learning-hardware>

©2024 Marc Greenberg Consulting, LLC  
All Rights Reserved

# And it's expensive



ALPHAWAVE SEMI

Alphawave Semi™. All Rights Reserved.



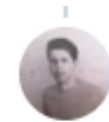
February 6-8, 2024  
Santa Clara Convention Center  
ChipletSummit.com



Tony Chan Carusone, Alphawave, Chiplet Summit Keynote 2024,  
<https://epochai.org/blog/how-much-does-it-cost-to-train-frontier-ai-models>  
x.com

Amortized hardware and energy cost to train frontier AI models over time

EPOCH AI



Sam Altman ✓  
@sama

Follow ...

we will have to monetize it somehow at some point;  
the compute costs are eye-watering

11:38 PM · Dec 4, 2022

# But memory bandwidth and capacity are not keeping up

	Specification and unit	Growth rate Doubling time	Datapoint of highest performance Metric prefix	N
Computational Performance	FLOP/s (FP32)	2x every 2.3 [2.1; 2.6] years	~90 TFLOP/s (NVIDIA L40)	45
	FLOP/s (tensor-FP32)	NA <sup>1</sup>	~495 TFLOP/s (NVIDIA H100 SXM)	7
	FLOP/s (tensor-FP16)	NA	~990 TFLOP/s (NVIDIA H100 SXM)	8
	OP/s (INT8)	NA	~1980 TOP/s (NVIDIA H100 SXM)	10
Computational price-performance	FLOP per \$ (FP32)	2x every 2.1 [1.6; 2.91] years	~4.2 exaFLOP per \$ (AMD Radeon RX 7900 XTX)	33
Computational energy-efficiency	FLOP/s per Watt (FP32)	2x every 3.0 [2.7; 3.3] years	~302 GFLOP/s per W (NVIDIA L40)	43
Memory capacity	DRAM capacity (Byte)	2x every 4 [3; 6] years	~128 GB (AMD Radeon Instinct MI250X)	47
Memory bandwidth	DRAM bandwidth in Byte/s	2x every 4 [3; 5] years	~3.3 TB/s (NVIDIA H100 SXM)	47
Interconnect bandwidth	Chip-to-chip communication bandwidth (Byte/s)	NA	~900 GB/s (NVIDIA H100)	45

Note this is a GDDR6-based card. GPGPU vs Tensor is important.

Note this is a GDDR6-based card. GPGPU vs Tensor is important.

**Table 1: Key performance trends.** All estimates are computed only for ML hardware. Numbers in brackets refer to the [5; 95]-th percentile estimate from bootstrapping with 1000 samples. OOM refers to order of magnitude, and N refers to the number of observations in our dataset. Note that performance figures are for dense matrix multiplication performance.

<https://epochai.org/blog/trends-in-machine-learning-hardware>

©2024 Marc Greenberg Consulting, LLC

All Rights Reserved



# Amdahl's law

## Definition [\[edit\]](#)

Amdahl's law can be formulated in the following way:<sup>[3]</sup>

$$S_{\text{latency}}(s) = \frac{1}{(1-p) + \frac{p}{s}}$$

where

- $S_{\text{latency}}$  is the theoretical speedup of the execution of the whole task;
- $s$  is the speedup of the part of the task that benefits from improved system resources;
- $p$  is the proportion of execution time that the part benefiting from improved resources originally occupied.

Furthermore,

$$\begin{cases} S_{\text{latency}}(s) \leq \frac{1}{1-p} \\ \lim_{s \rightarrow \infty} S_{\text{latency}}(s) = \frac{1}{1-p} \end{cases}$$

shows that the theoretical speedup of the execution of the whole task increases with the improvement of the resources of the system and that regardless of the magnitude of the improvement, the theoretical speedup is always limited by the part of the task that cannot benefit from the improvement.

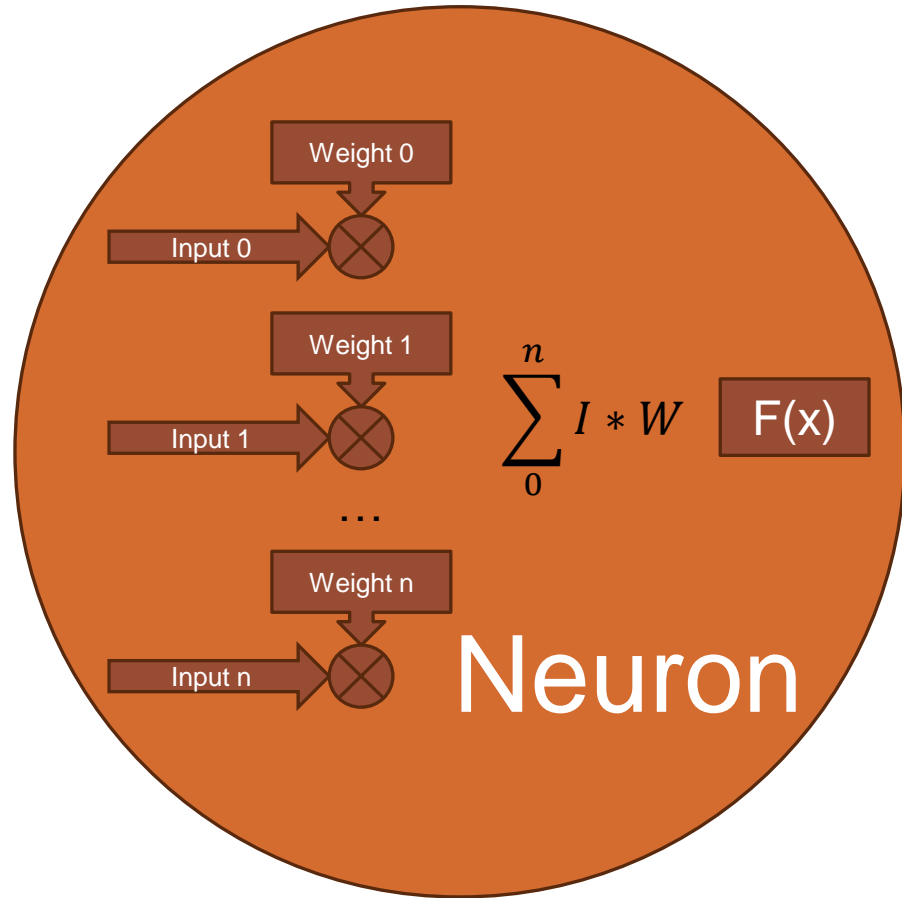
Amdahl's law applies only to the cases where the problem size is fixed. In practice, as more computing resources become available, they tend to get used on larger problems (larger datasets), and the time spent in the parallelizable part often grows much faster than the inherently serial work. In this case, [Gustafson's law](#) gives a less pessimistic and more realistic assessment of the parallel performance.<sup>[4]</sup>

- Translation: If you increase the compute without increasing the memory bandwidth, then the theoretical speedup will be limited by the memory bandwidth

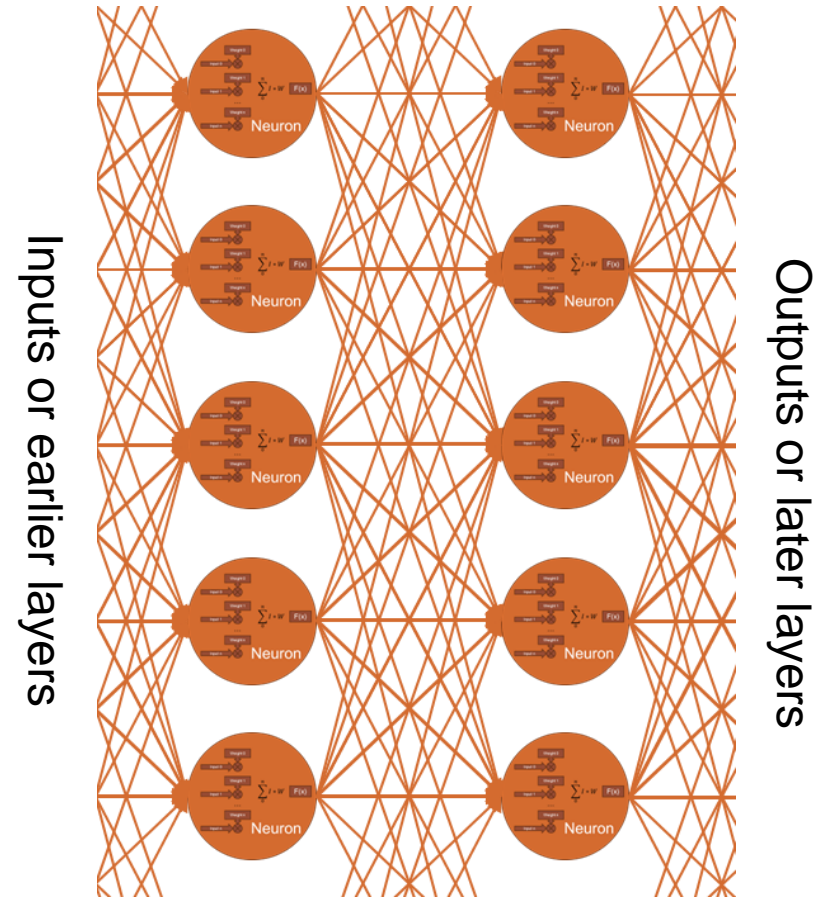
[https://en.wikipedia.org/wiki/Amdahl%27s\\_law](https://en.wikipedia.org/wiki/Amdahl%27s_law)



# Generalized Neuron Behavior



Neural Network (section of larger network)



# Systolic Arrays: The efficient heart of a TPU

- The generalized term “Systolic Array” is the technique used in almost all Tensor Processing units
  - Google “TPU”
  - NVIDIA “Tensor Core”
    - contained within “Streaming Multiprocessors” - SM
  - AMD CDNA “Matrix Cores”
    - contained within “Accelerator Complex Dies” – XCDs
  - Tenstorrent “Tensix” Cores
  - etc

FCCM'96 -- IEEE Symposium on FPGAs for Custom Computing Machines  
April 17-19, 1996, Napa, CA

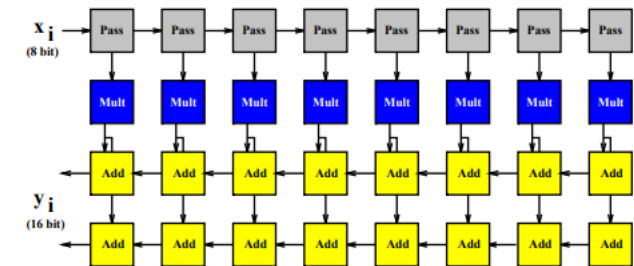


Figure 5: Systolic Convolution Implementation

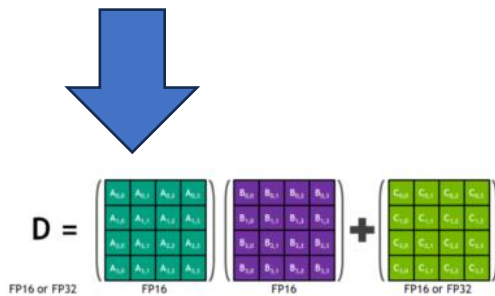
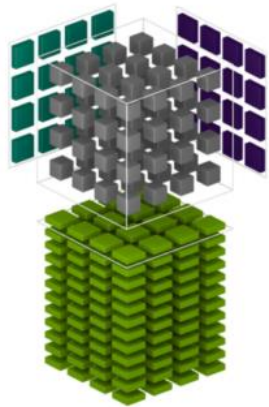
Systolic arrays are not new... here's one I worked on 30 years ago.  
The name “Systolic Array” was coined in 1979 but a WWII code-breaking machine used the same technique



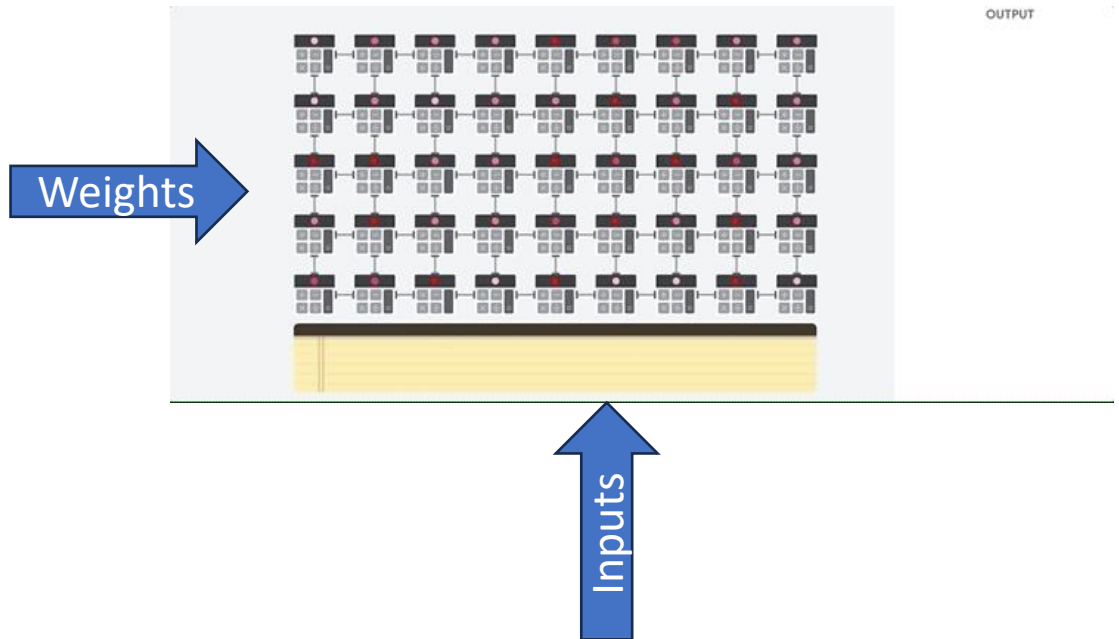
# Generalized Neural Network Behavior

- Arrange all the inputs and weights into a matrix, then multiply and accumulate the results using a systolic array

$$\sum_0^n I * W$$



Systolic array (animated cartoon)



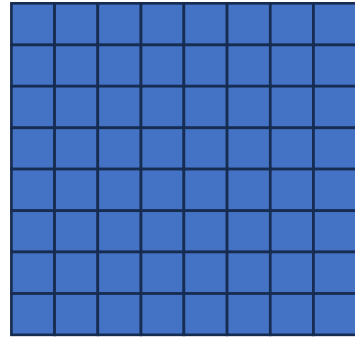
# Scalar, Vector, Matrix, Tensor



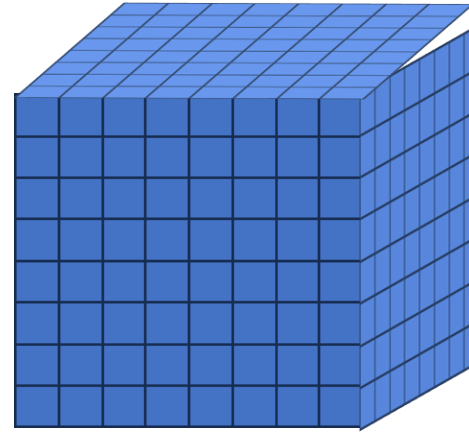
Scalar  
0-way  
Character



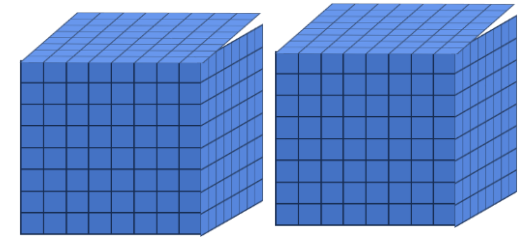
Vector  
1-way tensor  
Word



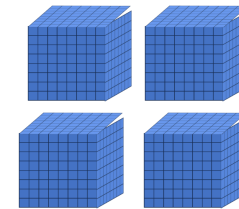
Matrix  
2-way tensor  
Page



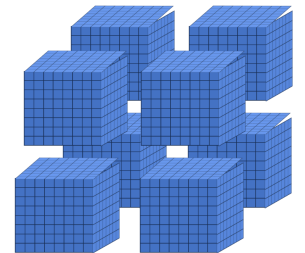
Tensor  
3-way tensor  
Book



4-way tensor  
Bookshelf



5-way  
Bookcase

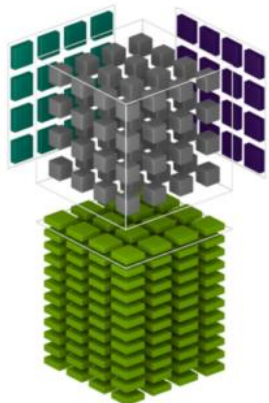
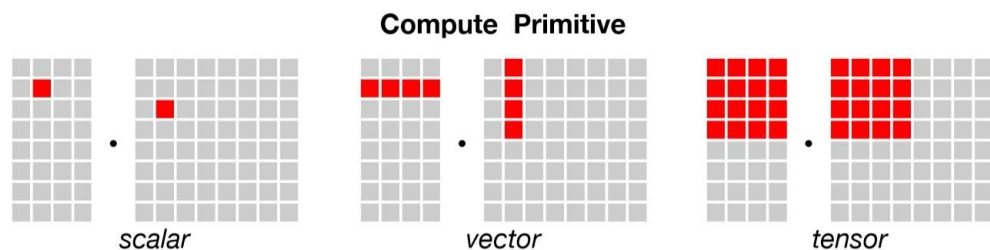


6-way  
Library





# How matrix/tensor math is done by CPU

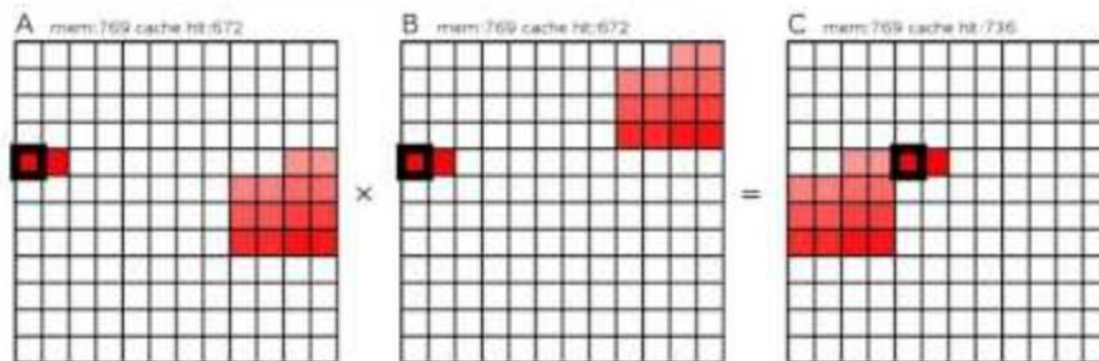


## Tensor math

$$D = \begin{pmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{pmatrix} + \begin{pmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \\ B_{41} & B_{42} & B_{43} \end{pmatrix} = \begin{pmatrix} C_{11} & C_{12} & C_{13} & C_{14} \\ C_{21} & C_{22} & C_{23} & C_{24} \\ C_{31} & C_{32} & C_{33} & C_{34} \\ C_{41} & C_{42} & C_{43} & C_{44} \end{pmatrix}$$

FP16 or FP32      FP16      FP16      FP16 or FP32

## Matrix multiplication: Tiled, B transposed



Totals: mem:2307 cache hits:2080 ≈ 90%



<https://youtu.be/aMvCEEBIBto?si=2olAEufVXcVh8Kc1>

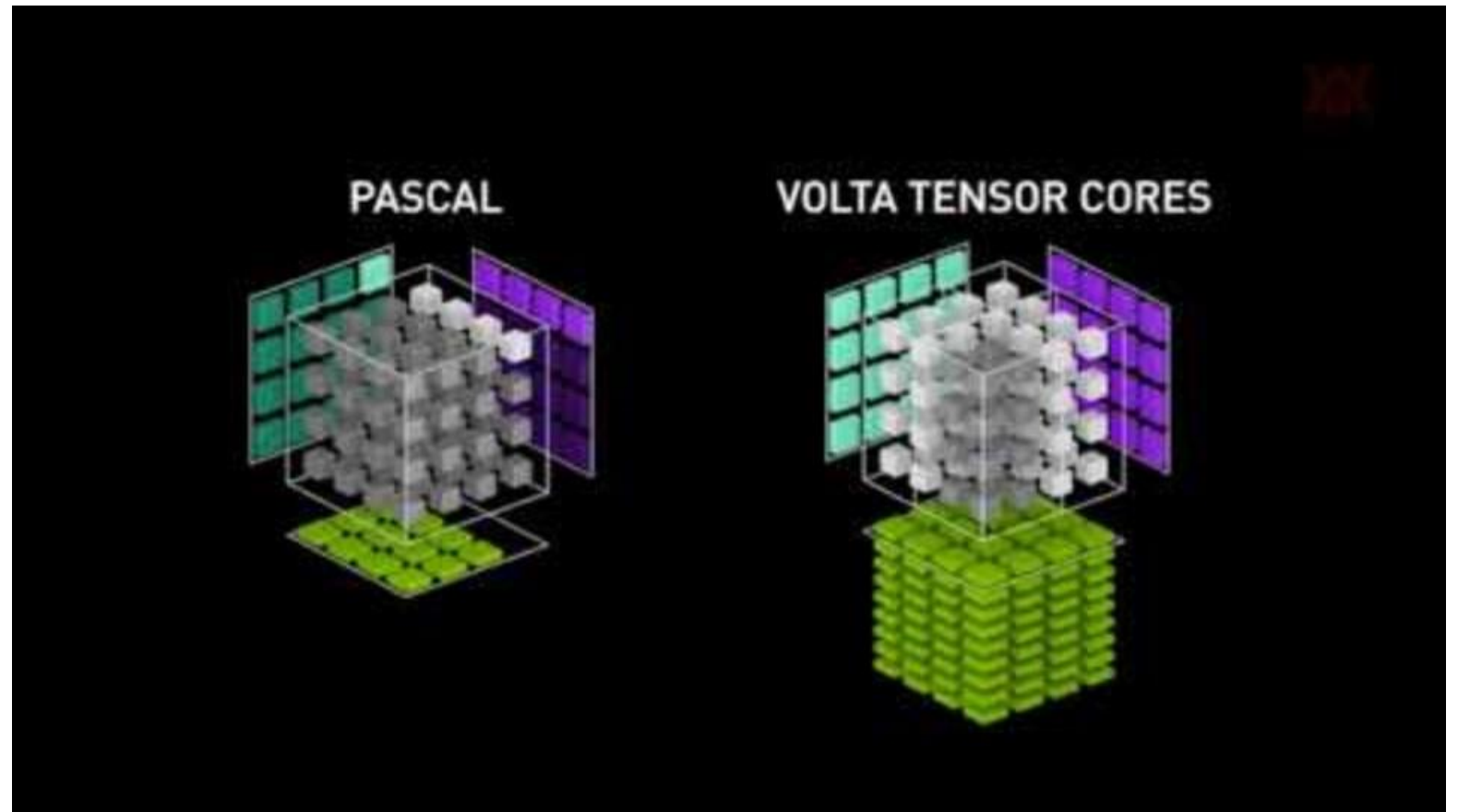


# Vector math by GPU, Tensor math by TPU

Pascal = GPGPU doing vector math

Volta = GPGPU+Tensor unit, tensor unit doing math

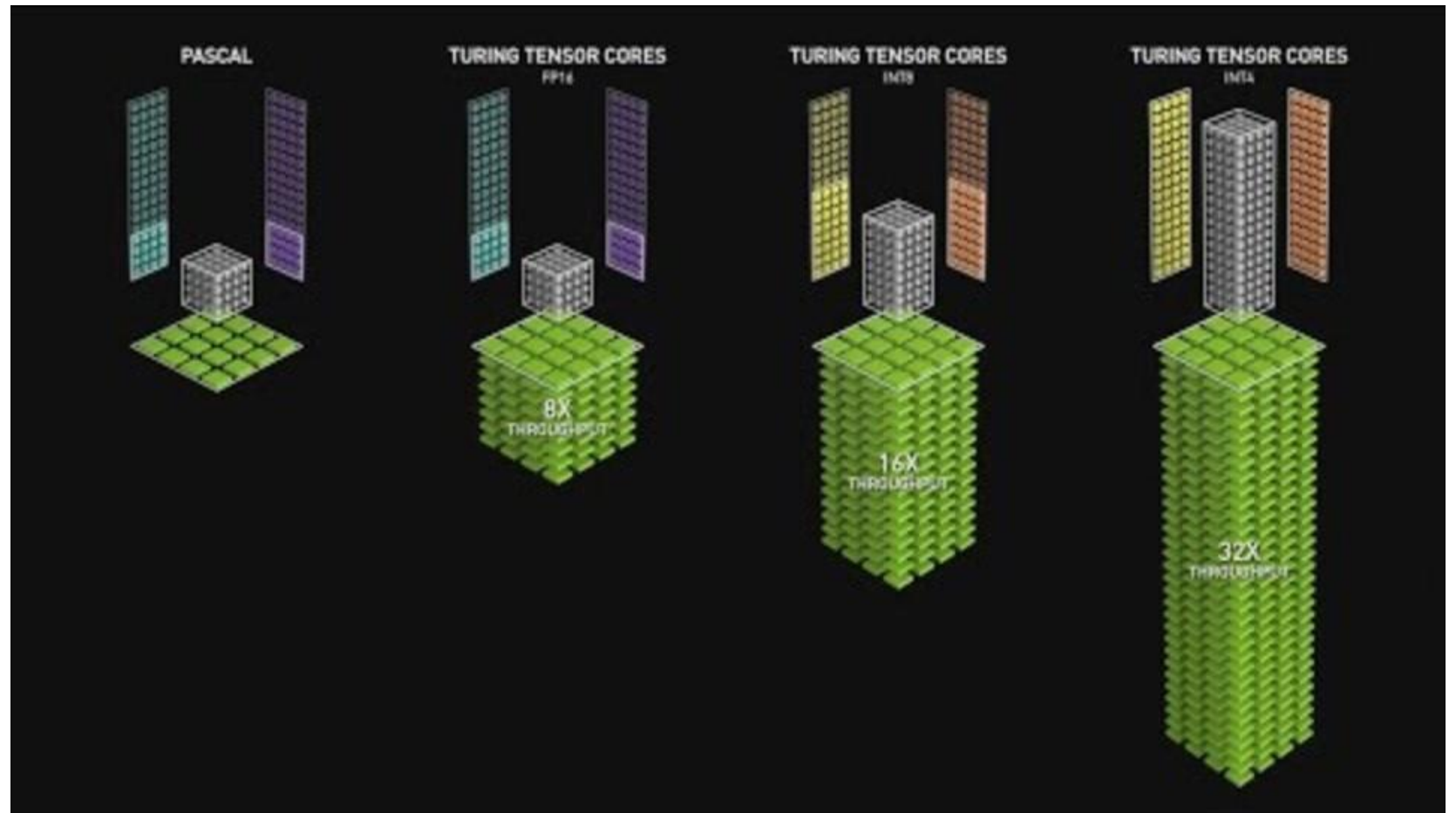
Later animation is effect of multi-precision (for inference)



<https://youtu.be/aWzxnj8JczM?si=ZVN9bGszp4HKf5Vd>

©2024 Marc Greenberg Consulting, LLC  
All Rights Reserved

# Tensor Math (again)



• <https://www.youtube.com/watch?v=yyR0ZoCeBO8>

©2024 Marc Greenberg Consulting, LLC  
All Rights Reserved



# More TPU every generation

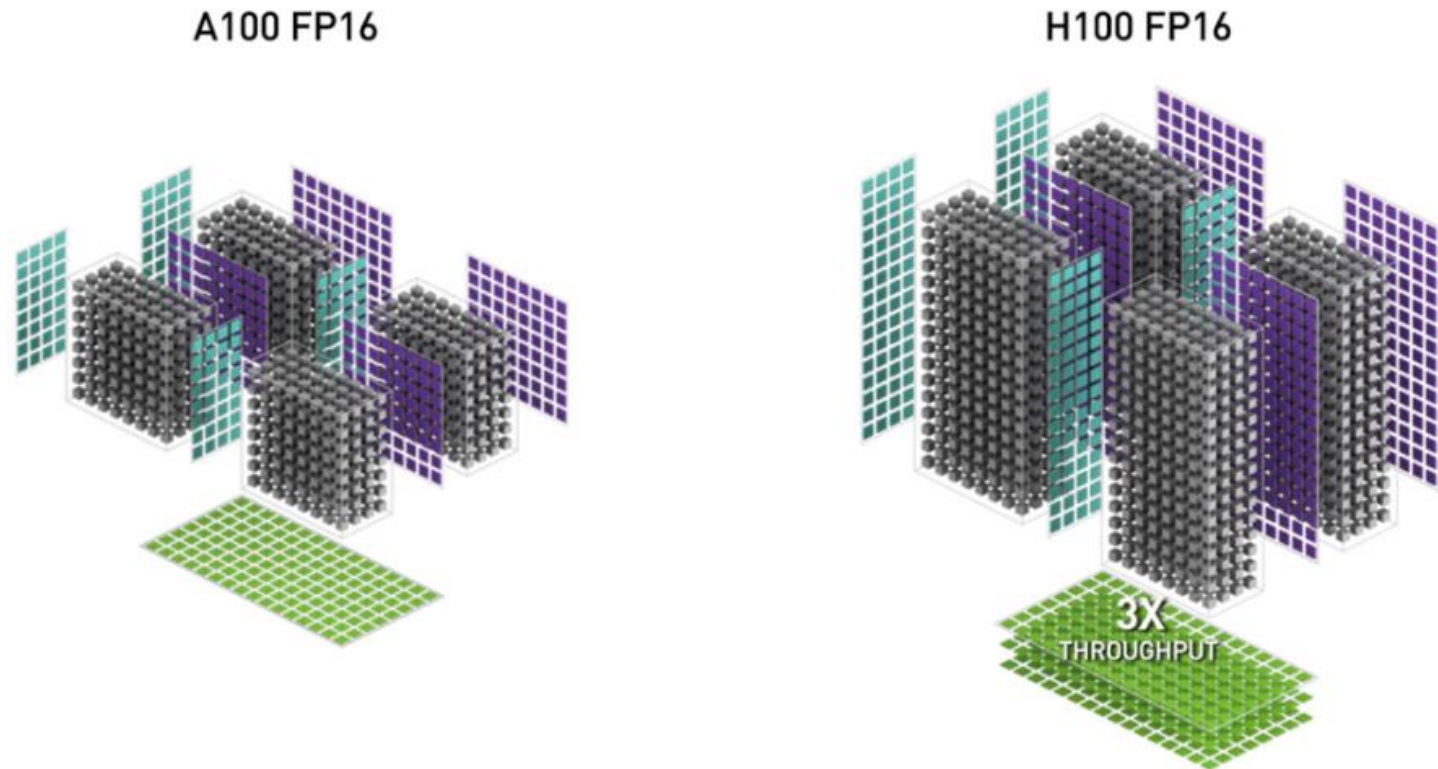


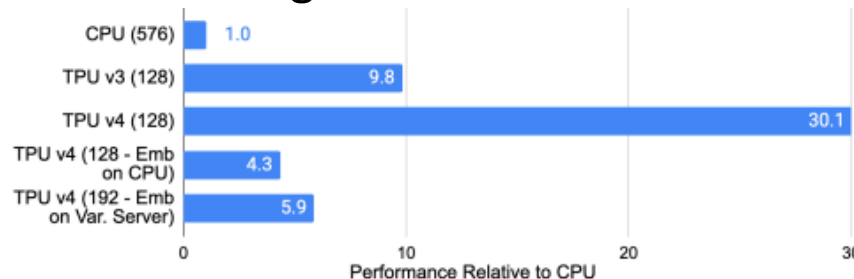
Figure 5. H100 FP16 Tensor Core has 3x throughput compared to A100 FP16 Tensor Core

<https://developer.nvidia.com/blog/nvidia-hopper-architecture-in-depth/>



# The magic of the TPU

- Table 4 compares key features of TPU v3 and TPU v4. Manufactured in 7 nm instead of 16 nm, TPU v4 has twice the matrix multipliers (enabled by the increased process density) and an 11% faster clock—this drives the 2.2X gain in peak performance. About 40% of the performance/Watt improvement was from technology and the rest was from design improvements (e.g., balancing the pipeline, implementing clock gating). The HBM memory bandwidth is 1.3x higher.



<https://arxiv.org/pdf/2304.01433> source: Google

**Table 4: TPU v4 and TPU v3 [26] features. Measured power is for the ASIC and HBM running production applications.**

	Google TPUv4	TPUv3
Production deployment	2020	2018
Peak TFLOPS	275 (bf16 or int8)	123 (bf16)
Clock Rate	1050 MHz	940 MHz
Tech. node, Die size	7 nm, <600 mm <sup>2</sup>	16 nm, < 700 mm <sup>2</sup>
Transistor count	22 billion	10 billion
Chips per CPU host	4	8
TDP	N.A.	N.A.
Idle, min/mean/max power	90, 121/170/192 W	123, 175/220/262 W
Inter Chip Interconnect	6 links @ 50 GB/s	4 links @ 70 GB/s
Largest scale configuration	4096 chips	1024 chips
Processor Style	Single Instruction 2D Data	Single Instruction 2D Data
Processors / Chip	2	2
Threads / Core	1	1
SparseCores / Chip	4	2
On Chip Memory	128 (CMEM) + 32 MiB (VMEM) + 10 MiB (spMEM)	32 MiB (VMEM) + 5 MiB (spMEM)
Register File Size	0.25 MiB	0.25 MiB
HBM2 capacity, BW	32 GiB, 1200 GB/s	32 GiB, 900 GB/s

Figure 9 below shows performance of an internal production recommendation model (DLRM0, see Sections 7.8 and 7.9) across the two TPU generations for 128 chips. The standalone CPU configuration has 576 Skylake sockets (400 for learners and 176 for variable servers). The bottom two bars show TPU v4 without SC, where the embeddings are placed in CPU memory. The “Emb on CPU” bar places embeddings in CPU host memory and the “Emb on Variable Server” bar places embeddings on 64 external variable servers. TPU v3 is faster than CPUs by 9.8x. TPU v4 beats TPU v3 by 3.1x and CPUs by 30.1x. When embeddings are placed in CPU memory for TPU v4, performance drops by 5x–7x, with bottlenecks due to CPU memory bandwidth.



# A rapid shift in capability (NVIDIA 2022)

Introduction of  
TPU

Accelerator Model	K80	P100	P100	P100	V100	V100	A100	A100	A100	H100	H100
GPU	2 * GK210B	GP100	GP100	GP100	GV100	GV100	GA100	GA100	GA100	GH100	GH100
Bus	PCI-E 3.0	PCI-E 3.0	PCI-E 3.0	SXM	PCI-E 3.0	SXM2/3	PCI-E 4.0	SXM4	SXM4	PCI-E 5.0	SXM5
GDDR5 or GDDR6/HBM2 Memory	24 GB	12 GB	16 GB	16 GB	16/32 GB	16/32 GB	40 GB	40 GB	80 GB	80 GB	80 GB
<b>Performance / Watt</b>											
F8 Efficiency (Gigaops/Watt)	-	-	-	-	-	-	-	-	-	9,142.9	5,714.3
INT8 Efficiency (Gigaops/Watt)	-	-	-	-	224.0	209.3	3,120.0	3,120.0	3,120.0	-	-
FP16 TC, FP32 ACC Efficiency (Gigaops/Watt)	-	-	-	-	448.0	416.7	1,560.0	1,560.0	1,560.0	4,571.4	2,857.1
FP16 Efficiency (Gigaops/Watt)	-	74.8	74.8	70.7	100.5	104.7	195.0	195.0	195.0	274.3	171.4
FP32/TF32 Efficiency (Gigaops/Watt)	29.1	37.2	37.2	35.3	50.2	52.3	48.8	48.8	48.8	137.1	85.7
FP64 Efficiency (Gigaops/Watt)	9.7	18.8	18.8	17.7	25.0	26.0	48.8	48.8	48.8	55.7	27.9
<b>\$ / Performance</b>											
Street Price, Single Unit	<b>\$400</b>	<b>\$600</b>	<b>\$2,200</b>	<b>\$1,100</b>	<b>\$7,500</b>	<b>\$2,500</b>	<b>\$6,000</b>	<b>\$12,500</b>	<b>\$15,000</b>	<b>\$17,500</b>	<b>\$19,500</b>
\$ / FP8 Teraflops	-	-	-	-	-	-	-	-	-	\$5.47	\$4.88
\$ / INT8 Teraops	-	-	-	-	\$133.93	\$39.81	\$4.81	\$10.02	\$12.02	-	-
\$ / FP16 TC, FP32 ACC Teraflops	-	-	-	-	\$66.96	\$20.00	\$9.62	\$20.03	\$24.04	\$10.94	\$9.75
\$ / FP16 Teraflops	-	\$32.09	\$117.65	\$51.89	\$267.86	\$79.62	\$76.92	\$160.26	\$192.31	\$182.29	\$162.50
\$ / FP32/TF32 Teraflops	\$45.77	\$64.52	\$236.56	\$103.77	\$597.13	\$159.24	\$19.23	\$40.06	\$48.08	\$56.09	\$62.50
\$ / FP64 Teraflops	\$137.46	\$127.66	\$468.09	\$207.55	\$1,201.92	\$320.51	\$307.69	\$641.03	\$769.23	\$897.44	\$1,000.00
<b>\$ / Performance / Watt</b>											
\$ / FP8 Teraops / Watt	-	-	-	-	-	-	-	-	-	\$191	\$3.41
\$ / INT8 Teraops / Watt	-	-	-	-	\$33.48	\$11.94	\$1.92	\$4.01	\$4.81	-	-
\$ / FP16 TC, FP32 ACC Teraflops / Watt	-	-	-	-	\$16.74	\$6.00	\$3.85	\$8.01	\$9.62	\$3.83	\$6.83
\$ / FP16 Teraflops / Watt	-	\$8.02	\$29.41	\$15.57	\$74.64	\$23.89	\$30.77	\$64.10	\$76.92	\$63.80	\$113.75
\$ / FP32/TF32 Teraflops / Watt	\$13.73	\$16.13	\$59.14	\$31.13	\$149.28	\$47.77	\$123.08	\$256.41	\$307.69	\$127.60	\$227.50
\$ / FP64 Teraflops / Watt	\$41.24	\$31.91	\$117.02	\$62.26	\$300.48	\$96.15	\$123.08	\$256.41	\$307.69	\$314.10	\$700.00

Multiple X every generation

TPU with >10x cost efficiency compared to GPU operations

TPU with ~20X \$/FLOPs/Watt compared to GPU operations

As usual items in bold red italics are estimations by *The Next Platform*.

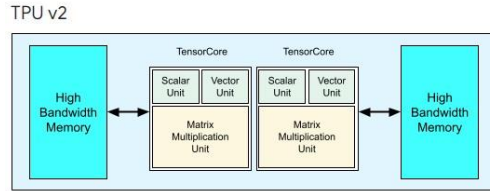
<https://www.nextplatform.com/2022/05/09/how-much-of-a-premium-will-nvidia-charge-for-hopper-gpus/>



# And growing fast (Google TPU example)

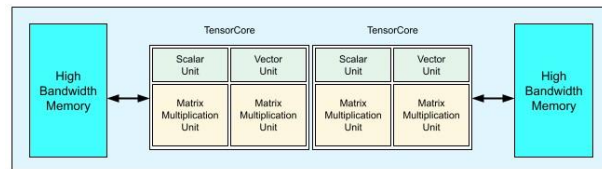
TPU Size

16K



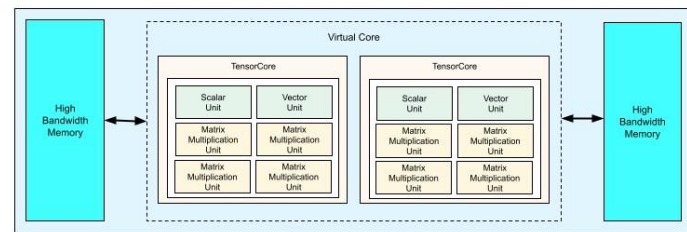
TPU v3

32K



TPU v4

64K



128K Trillium (Nextplatform prediction)

Google TPU Compute Engines	TPU v1	TPU v2	TPU v3	TPU v4	TPU v5p	"Trillium" TPU v6	TPU v3 Over	TPU v4 Over	TPU v5p Over	TPU v6 Over
							TPU v2	TPU v3	TPU v4	TPU v5e
First Deployed	Q2 2015	Q3 2017	Q4 2018	Q4 2021	Q4 2023	<b>Q4 2024</b>				
ML Inference	Yes	Yes	Yes	Yes	Yes	Yes				
ML Training	No	Yes	Yes	Yes	Yes	Yes				
Chip Process	28 nm	16 nm	16 nm	7 nm	<b>5 nm</b>	<b>4 nm</b>				
Transistors	3 B	9 B	10 B	<b>31 B</b>	???	???	1.11	3.10	???	???
Die Size	330 mm <sup>2</sup>	625 mm <sup>2</sup>	700 mm <sup>2</sup>	<b>780 mm<sup>2</sup></b>	<b>700 mm<sup>2</sup></b>	<b>790 mm<sup>2</sup></b>	1.12	1.11	<b>0.90</b>	<b>2.26</b>
Clock Speed	700 MHz	700 MHz	940 MHz	<b>1,050 MHz</b>	<b>2,040 MHz</b>	<b>2,060 MHz</b>	1.34	1.12	<b>1.94</b>	<b>1.18</b>
TensorCores Per Chip	1	2	2	2	2	2				
MXU Matrix Size/Core	1 * 256x256	1 * 128x128	2 * 128x128	4 * 128x128	4 * 128x128	<b>4 * 256x256</b>	1.00	1.00	1.00	2.00
Dataflow SparseCores	-	-	-	4	4	4				
On Chip Cache Memory	28 MB	32 MB	32 MB	32 MB	48 MB	???	1.00	1.00	1.50	???
Off Chip HBM Memory	8 GB	16 GB	32 GB	32 GB	95 GB	32 GB	2.00	1.00	2.97	2.00
HBM Memory Bandwidth	<b>300 Gb/sec</b>	700 GB/sec	900 GB/sec	1,228 GB/sec	2,765 GB/sec	1,640 TB/sec	1.29	1.36	2.25	2.00
INT8 Peak Teraflops	92	-	-	275	918	1,852	-	-	3.34	4.70
BF16 Peak Teraflops	-	46	123	137.5	459	926	2.67	1.12	3.34	4.70
Precision	INT8	BF16	BF16	BF16/INT8	BF16/INT8	<b>BF16/INT8</b>				
ICI Links * Speed Gb/sec	-	4 * 496	4 * 656	6 * 448	<b>6 * 800</b>	<b>4 * 800</b>	1.32	1.02	1.79	2.00
Interconnect Topology	-	2D Torus	2D Torus	3D Torus	3D Torus	<b>2D Torus</b>				
Chip Idle Watts	28	53	84	170	???	???				
Max Measured Watts	???	???	262	192	???	???				
Chip TDP Watts	75	280	450	<b>300</b>	???	???				
Chips Per CPU Host	4	4	4	4	<b>8</b>	???				
Max Chips Per Pod	-	256	1,024	4,096	8,960	256	4.00	4.00	2.19	1.00
Peak Petaflops Per Pod	-	12	126	1,126	8,225	474	10.70	8.94	7.30	4.70
All-Reduce Bandwidth Per Pod	-	<b>120 TB/sec</b>	340 TB/sec	1,100 TB/sec	???	???	2.83	3.24	???	???
Bisection Bandwidth Per Pod	-	2 TB/sec	6.4 TB/sec	24 TB/sec	???	???	3.20	3.75	???	???
<b>Pricing Per TPU Chip, Lowest US Region Pricing</b>										
Preemptible Spot	-	<b>\$0.45</b>	<b>\$0.60</b>	<b>\$0.97</b>	<b>\$2.10</b>	<b>\$1.25</b>	1.33	1.61	2.17	<b>2.08</b>
Preemptible Spot For Three Months	-	<b>\$972.00</b>	<b>\$1,296.00</b>	<b>\$2,086.56</b>	<b>\$4,536.00</b>	<b>\$2,700.00</b>	1.33	1.61	2.17	<b>2.08</b>
Preemptible Spot Price/Peak Perf	-	<b>\$21.13</b>	<b>\$10.54</b>	<b>\$7.59</b>	<b>\$4.94</b>	<b>\$1.46</b>	0.50	0.72	0.65	<b>0.44</b>
On Demand Per Hour	-	<b>\$1.50</b>	<b>\$2.00</b>	<b>\$3.22</b>	<b>\$4.20</b>	<b>\$2.50</b>	1.33	1.61	1.30	<b>2.08</b>
On Demand For Three Months	-	<b>\$3,240.00</b>	<b>\$4,320.00</b>	<b>\$6,955.20</b>	<b>\$9,072.00</b>	<b>\$5,400.00</b>	1.33	1.61	1.30	<b>2.08</b>
On Demand Price/Peak Perf	-	<b>\$70.43</b>	<b>\$35.12</b>	<b>\$25.29</b>	<b>\$9.88</b>	<b>\$2.92</b>	0.50	0.72	0.39	<b>0.44</b>
1 Year CUD Per Hour	-	<b>\$0.95</b>	<b>\$1.26</b>	<b>\$2.03</b>	<b>\$2.94</b>	<b>\$1.75</b>	1.33	1.61	1.45	<b>2.08</b>
1 Year CUD Cost	-	<b>\$8,283.87</b>	<b>\$11,045.16</b>	<b>\$17,782.71</b>	<b>\$25,772.04</b>	<b>\$15,340.50</b>	1.33	1.61	1.45	<b>2.08</b>
1 Year CUD Price/Peak Perf	-	<b>\$180.08</b>	<b>\$89.80</b>	<b>\$64.66</b>	<b>\$28.07</b>	<b>\$8.28</b>	0.50	0.72	0.43	<b>0.44</b>
3 Year CUD Per Hour	-	<b>\$0.68</b>	<b>\$0.90</b>	<b>\$1.45</b>	<b>\$1.89</b>	<b>\$1.13</b>	1.33	1.61	1.30	<b>2.08</b>
3 Year CUD Cost	-	<b>\$5,917.05</b>	<b>\$7,889.40</b>	<b>\$12,701.93</b>	<b>\$16,567.74</b>	<b>\$9,861.75</b>	1.33	1.61	1.30	<b>2.08</b>
3 Year CUD Price/Peak Perf	-	<b>\$128.63</b>	<b>\$64.14</b>	<b>\$46.19</b>	<b>\$18.05</b>	<b>\$5.32</b>	0.50	0.72	0.39	<b>0.44</b>

Growing fast



<https://www.nextplatform.com/wp-content/uploads/2022/10/google-tpuv4-v3-v2-chip-block-diagrams.jpg>

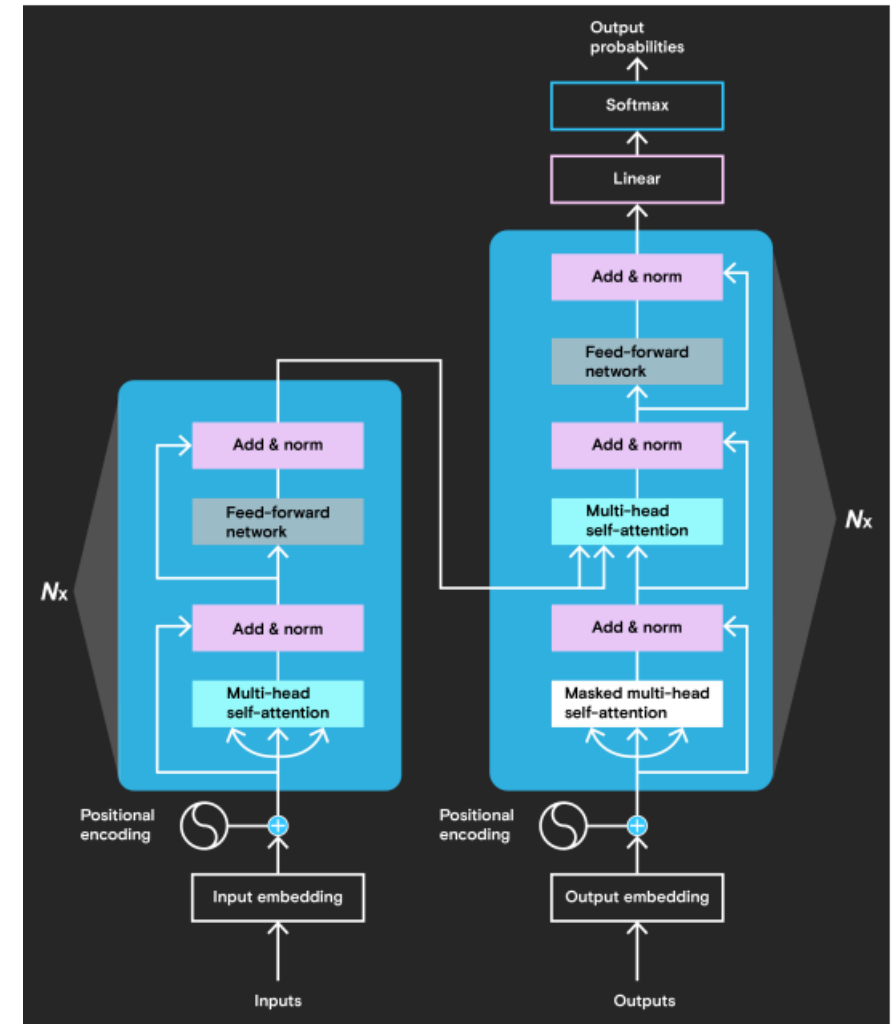
As usual items in bold red italics are estimations by *The Next Platform*. Inference-oriented TPUs removed for clarity

<https://www.nextplatform.com/2024/06/10/lots-of-questions-on-googles-trillium-tpu-v6-a-few-answers/>

©2024 Marc Greenberg Consulting, LLC  
All Rights Reserved

# Generative AI is iterative

- To make things extra fun, LLMs are iterative
- Demonstrative example:
  - We're going to make a children's story about (subject 1) and (subject 2)
  - Each of you is an LLM tasked with adding one word to the end of the story





# Section Summary

- The processors used in AI are powerful math machines
- Matrix Multiply-Accumulate (MAC) is the fastest growing part of most AI / ML Chip Architectures
  - often doubling from generation to generation
- Dramatic increase in math capability drives increase in memory bandwidth demand
- Size of generative AI models and iteration drives increase in memory capacity
- HBM is today's solution for solving memory bandwidth challenges (at a cost, which we'll talk about later)



# Shameless Plug

- I'm also VP Product for Cassia.ai
- We've constructed a TPU that is 33% smaller and improves TOPS/w by 2.5x compared with traditional techniques by using our technology
- This could also mean 2.5x less power or 2.5x more TPU operations within the same power envelope
- Ask me later

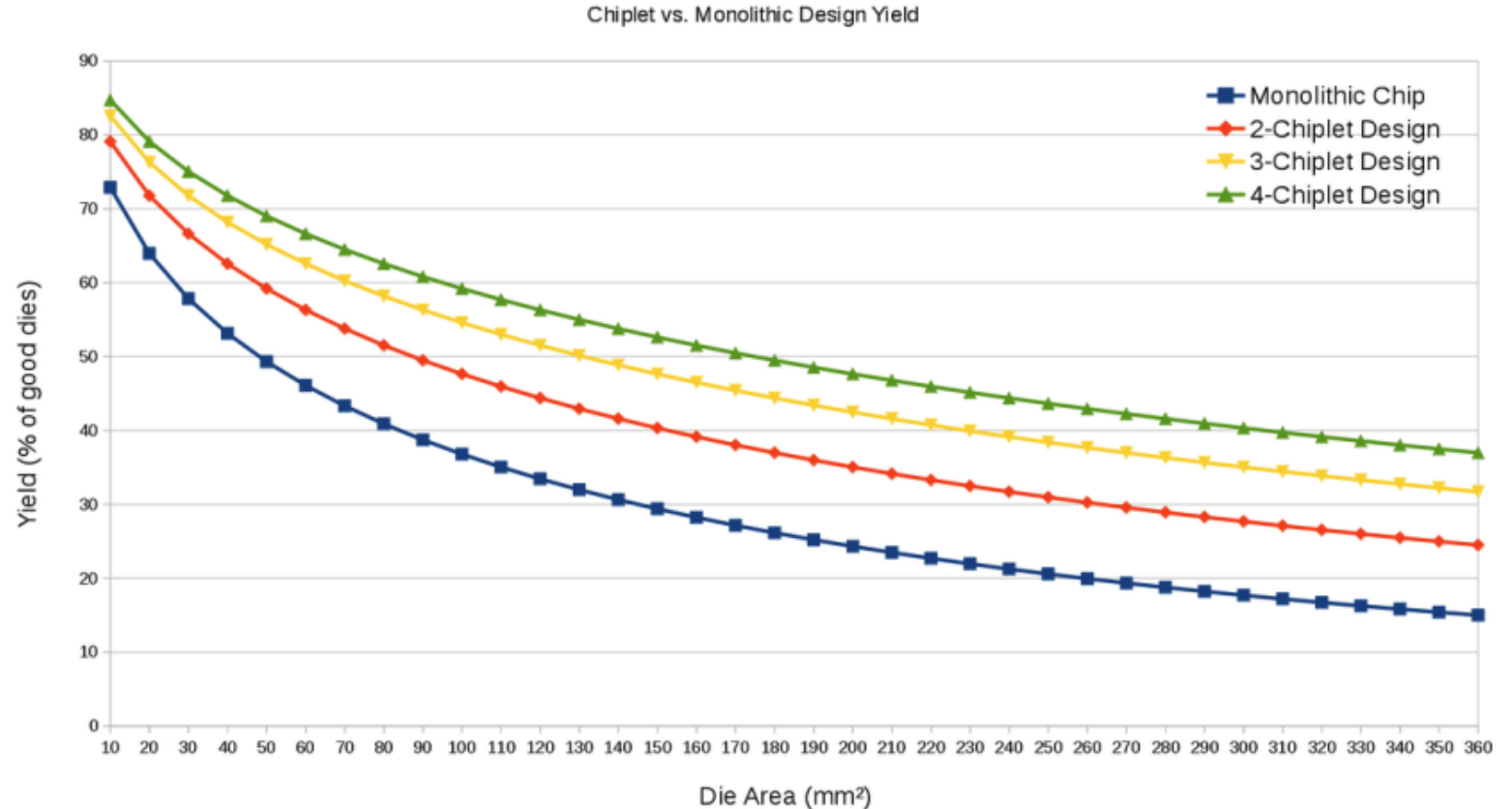


# Chipllets



# The problems with really big monolithic chips

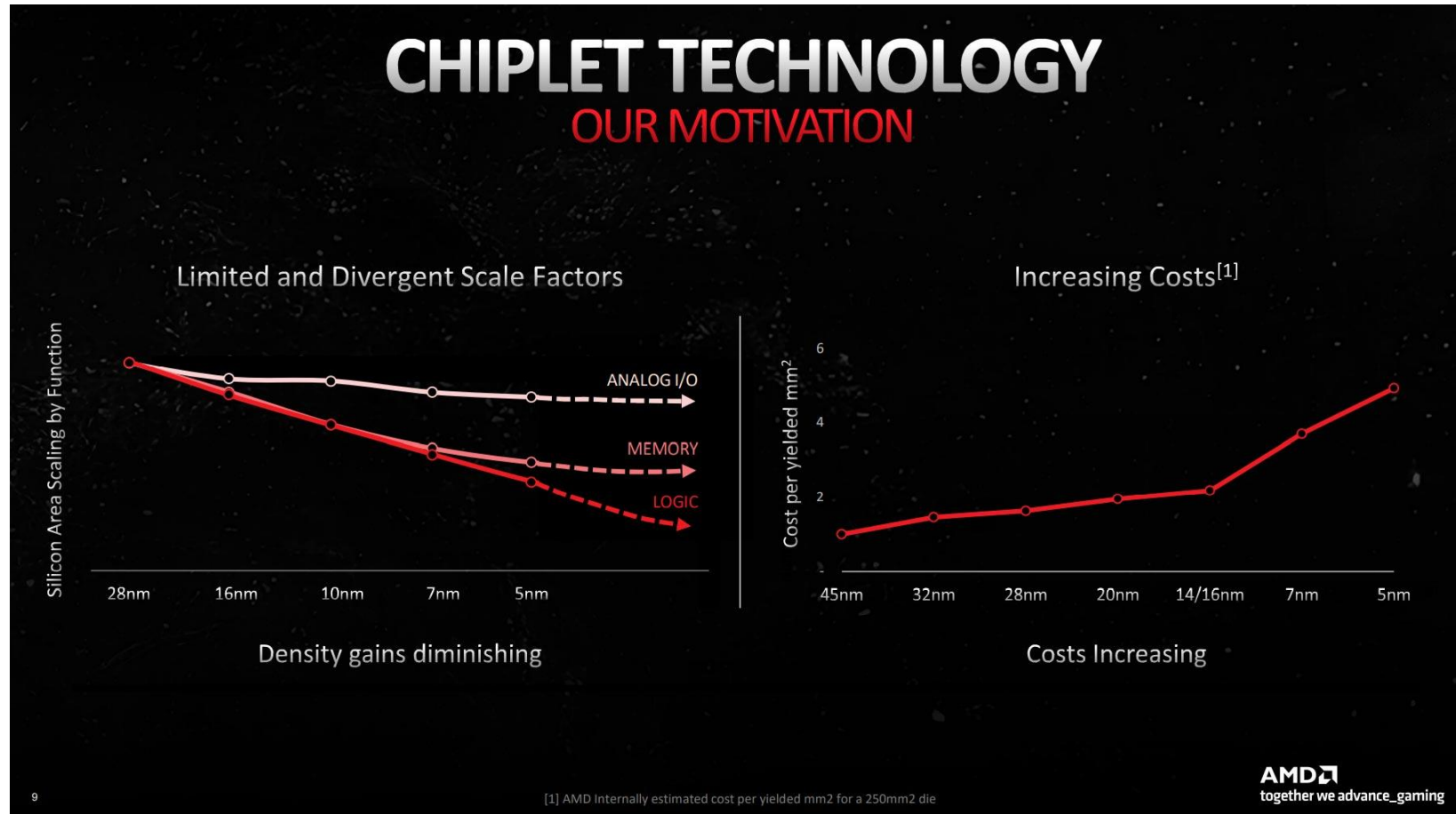
- Yield
- Reticle limit
- Thermal
- Scaling of different parts of the chip
- Cost per transistor



<https://medium.com/@marcussl.chan/chiplets-why-it-can-solve-the-slowng-of-moores-law-651ed53f413d>



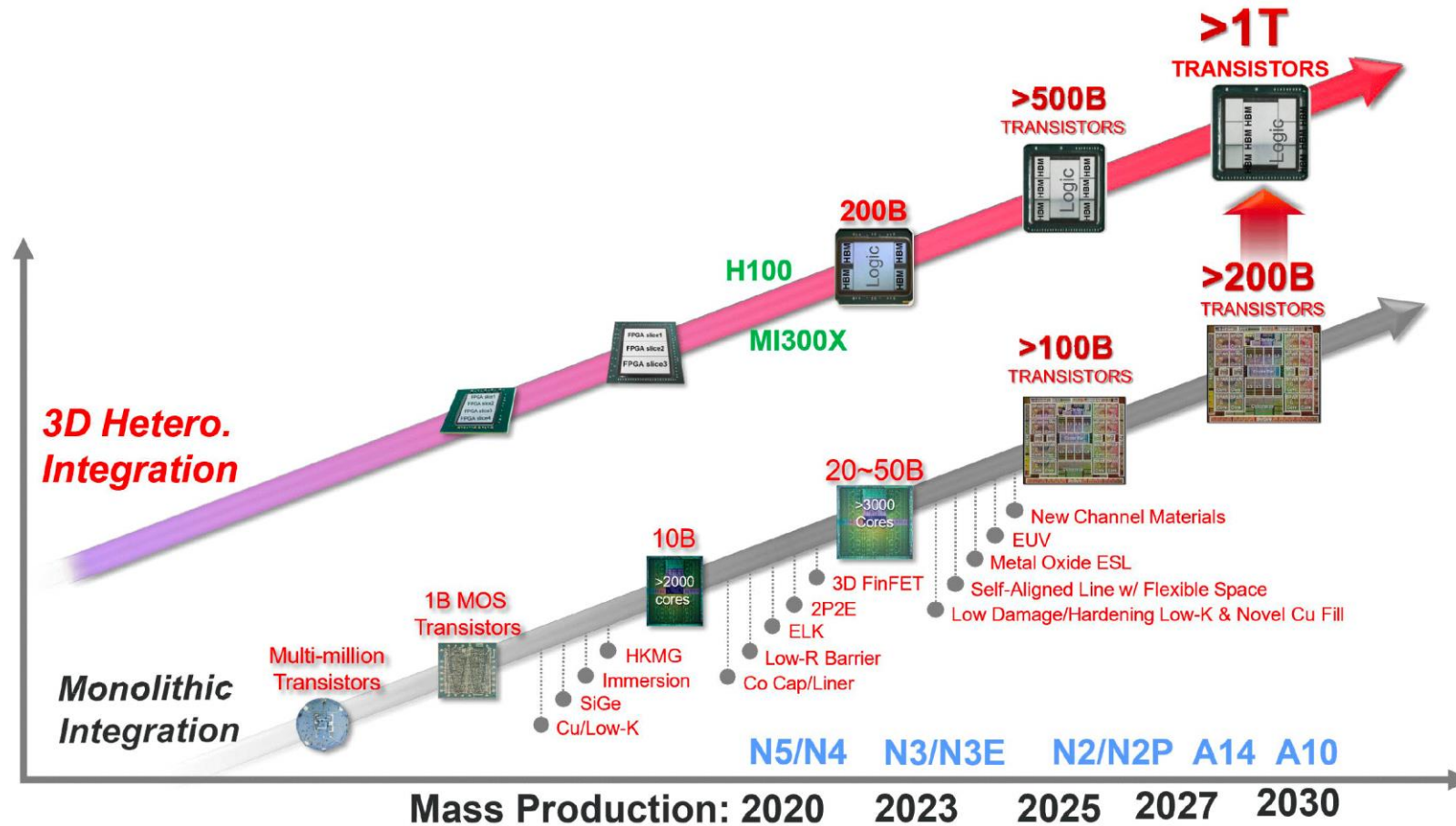
# Chiplet Economy



<https://www.club386.com/amd-radeon-rx-7900-xtx-review-rise-of-the-chiplets/>

©2024 Marc Greenberg Consulting, LLC  
All Rights Reserved

# Chiplet enables extreme integration



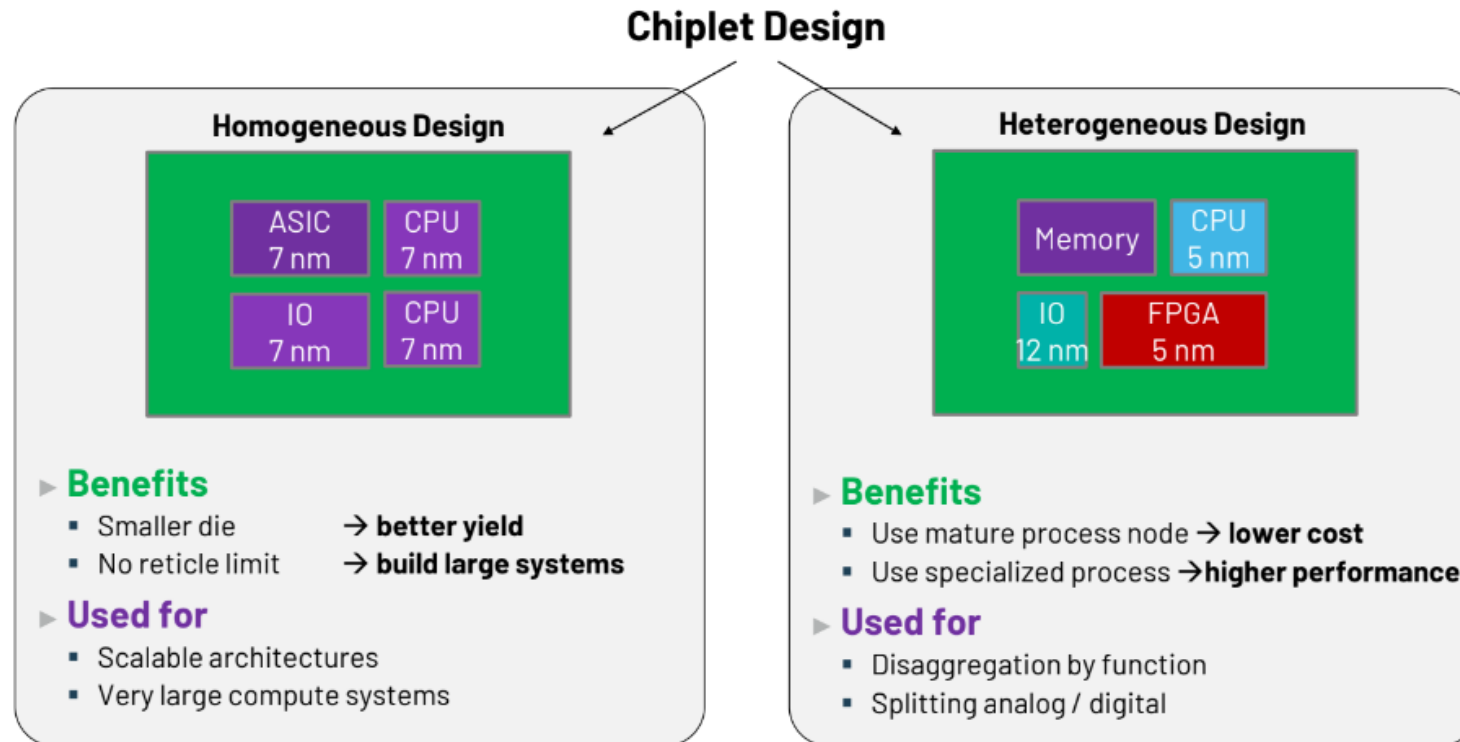
TSMC slide from IEDM conference foresees advancements in packaging technologies. (Image credit: TSMC)

Source: <https://www.tomshardware.com/tech-industry/manufacturing/tsmc-charts-a-course-to-trillion-transistor-chips-eyes-monolithic-chips-with-200-billion-transistors-built-on-1nm-node>



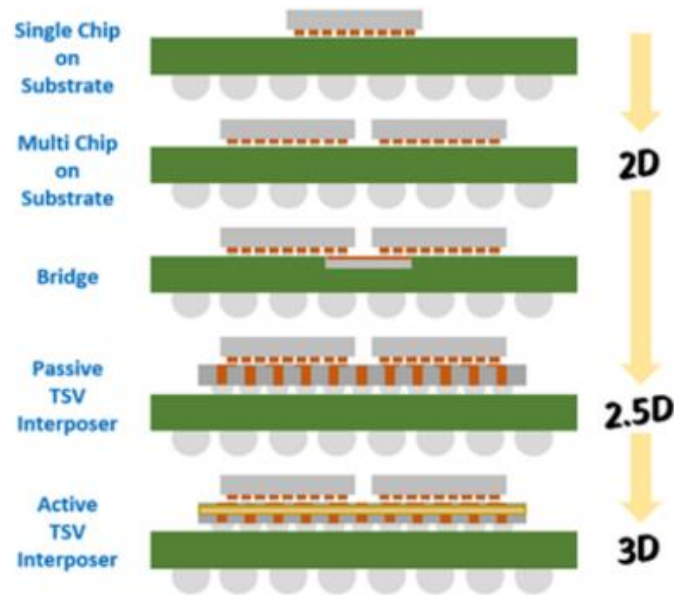
# What is a chiplet?

## Chiplet Architecture



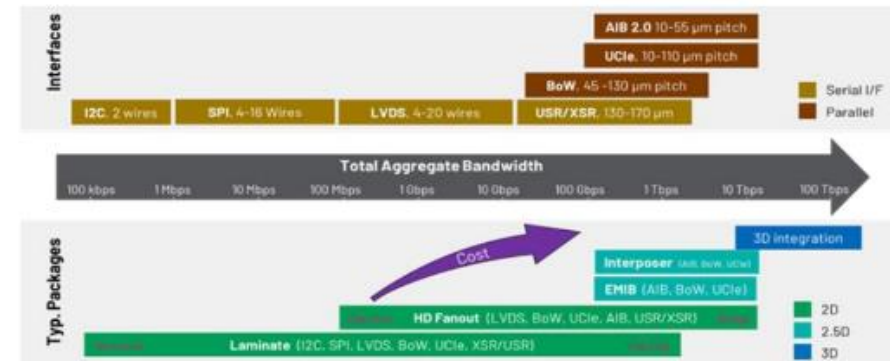
# What is a chiplet?

## Packaging Choices



### ► Packaging and D2D interface are connected

- Bump pitch and escape pitch → via pad, L/S
- Depths of signal bumps → # of layers
- Needed insertion loss → dielectric choice
- Maximum distance between chiplets



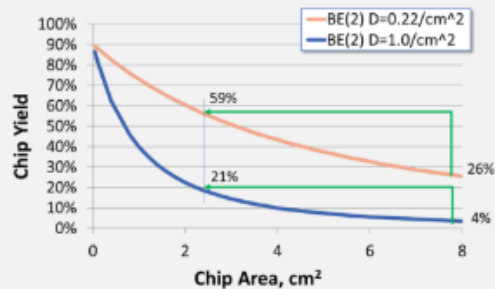


# What is a chiplet?

## Three Financial Benefits of Chiplets

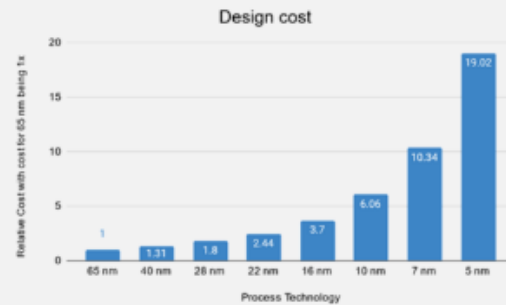
### Yield Advantage

- Smaller die have higher yield
- This can off-set higher package and test cost for large systems
- See [ODSA cost model](#)



### Design Cost Advantage

- Newer processes are more expensive to design on
- Shifting some functions to older nodes reduces cost



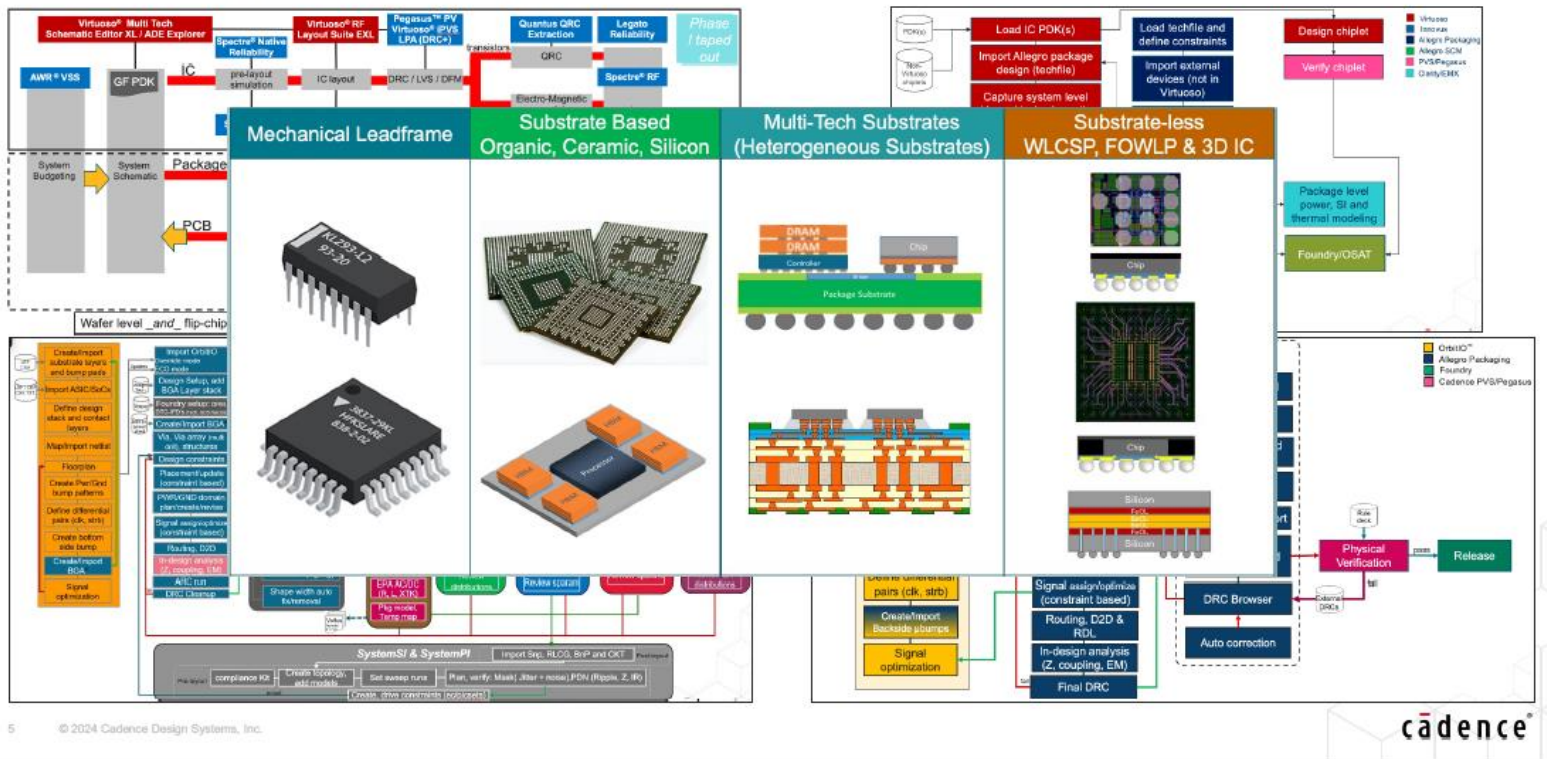
### Early Entry Advantage

- Chiplet reuse accelerates TTM
- This can improve returns by:
  - Higher market share
  - Longer product life-cycle
  - Higher profit margin



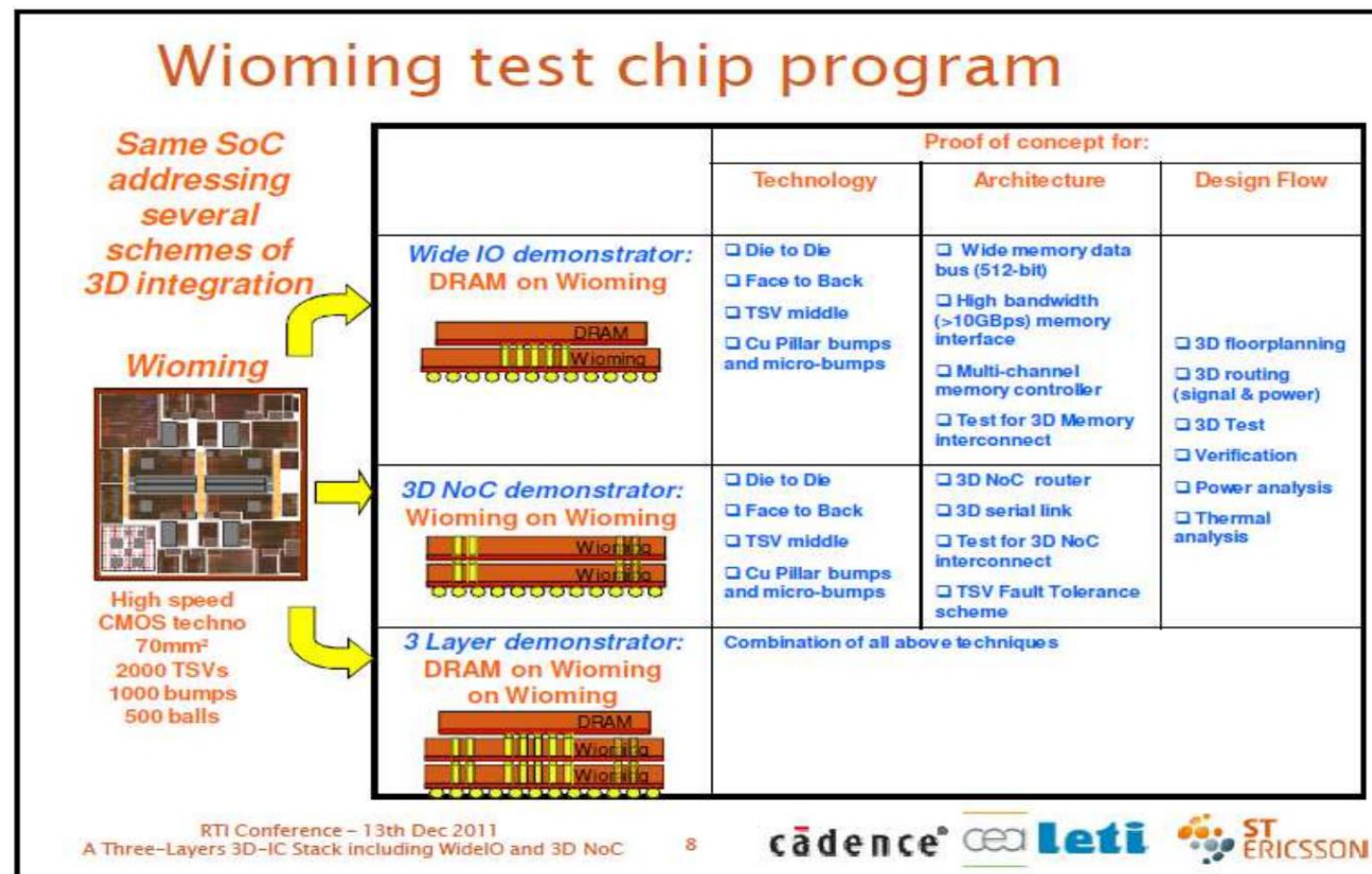
# There are many ways of doing chiplets

## The Wild West of Semiconductor Packaging 100s of ways to package devices



# History of 2.5D/3D Stacked DRAMs

- First standards-based DRAM on Logic Chip: Wioming using 1<sup>st</sup> gen WideIO
- WideIO Goal: Reduce power, increase performance, reduce PCB area
- What actually happened: Package-on-package of LPDDRx



<https://www.ieee-edps.com/archives/2012/c/1800greenberg.pdf>

Enough Talk! Practical Approaches to 3-D IC – TSV/Silicon Interposer and Wide I/O Implementation from People Who Have Been There and Done That, presented by Frank Lee of TSMC and Marc Greenberg of Cadence Design Systems – at 49<sup>th</sup> DAC, June 2012

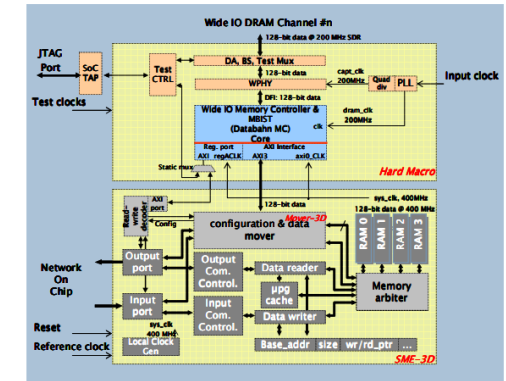


# History of 2.5D/3D Stacked DRAMs

- First standards-based DRAM on Logic Chip: Wioming using 1<sup>st</sup> gen WideIO
- WideIO Goal: Reduce power, increase performance, reduce PCB area
- What actually happened: Package-on-package of LPDDRx

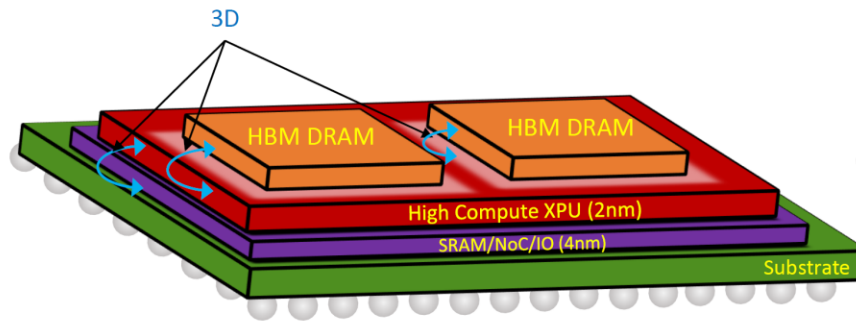
## Wide IO SME architecture overview

- Wide IO Memory Controller (Cadence DENALI)
  - Compliant with DRAM specification for Wide IO from JEDEC (<http://www.jedec.org/>)
  - High performance, and advanced low-power features
  - First deliveries to 3D-IC Wioming ST-Ericsson/LETI project
- Wide IO PHY Interface
  - 200MHz, 128 bit, SDR
  - ~1200 TSVs, μbuffers and μbumps
  - Also integrates ESD protections for DRAM
- Specific Design for Wide IO Testability Integration
  - Boundary scan, direct access, stuck-at, memory bist, PLL test
- Smart Memory Engine
  - Data transfer handling between Wide IO, SRAM and ANoC
  - Integration within ANoC
  - Up to 3.2GB/s data bandwidth



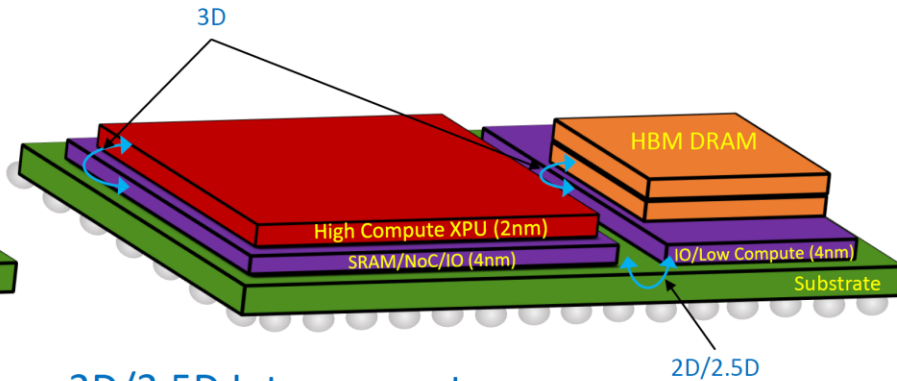
# Why not just put the memory on top?

## 3D vs. 2D/2.5D Interconnects to Solve AI Bottlenecks



### 3D Interconnects

- Highest D2D bandwidth
- XPU uses most advanced/expensive node at highest compute density
- Heat limits compute density under DRAM
  - Wastes valuable XPU silicon area (e.g. 2nm)



### 2D/2.5D Interconnects

- High enough D2D bandwidth for most applications
- Base die uses cheaper N-1/N-2 FinFET node, but advanced enough for most logic functions
- Easier thermal management
  - Best use of valuable XPU silicon area

3D & 2D/2.5D all necessary to develop the optimal solutions



# First HBM Chip: AMD Fiji (2014 production)


- 28nm
- 596mm<sup>2</sup>
  - 1011mm<sup>2</sup> interposer
- 4x HBM (1)
- 512GB/s



# ... and what it took to build it

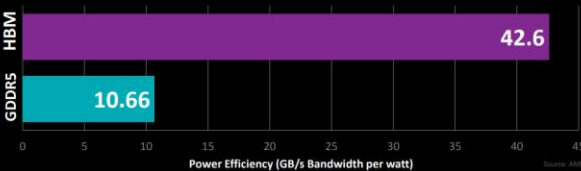
- “>15 Prototypes over 8.5 years” starting ~2008

**“Fiji” Chip**



**HIGH-BANDWIDTH MEMORY**

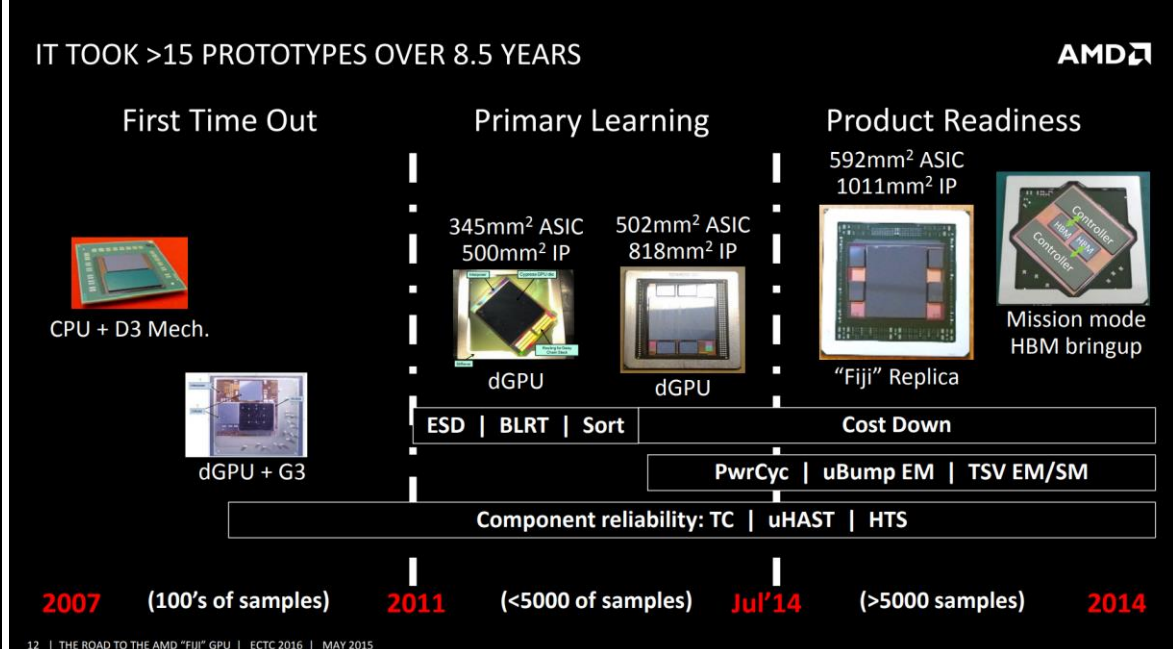
- ▲ Initiated with several DRAM partners 7 years ago
- ▲ SKhynix is in production supporting “Fiji”
- ▲ Benefits
  - 4096-bit memory interface with four stacks creating 512GB/s of bandwidth
  - 60% higher memory bandwidth<sup>6</sup> for 60% less power<sup>7</sup> than GDDR5
  - 4X Bandwidth per watt improvement from Radeon™ R9 290X
- ▲ Also required functional prototyping



Memory Type	Power Efficiency (GB/s Bandwidth per watt)
HBM	42.6
GDDR5	10.66

10 | THE ROAD TO THE AMD “FIJI” GPU | ECTC 2016 | MAY 2015

**IT TOOK >15 PROTOTYPES OVER 8.5 YEARS**



**First Time Out**  
CPU + D3 Mech.  
dGPU + G3

**Primary Learning**  
345mm<sup>2</sup> ASIC  
500mm<sup>2</sup> IP  
502mm<sup>2</sup> ASIC  
818mm<sup>2</sup> IP  
dGPU  
dGPU

**Product Readiness**  
592mm<sup>2</sup> ASIC  
1011mm<sup>2</sup> IP  
“Fiji” Replica  
Mission mode HBM bringup

**ESD | BLRT | Sort** | **Cost Down**

**PwrCyc | uBump EM | TSV EM/SM**

**Component reliability: TC | uHAST | HTS**

**2007** (100’s of samples) | **2011** (<5000 of samples) | **Jul’14** (>5000 samples) | **2014**

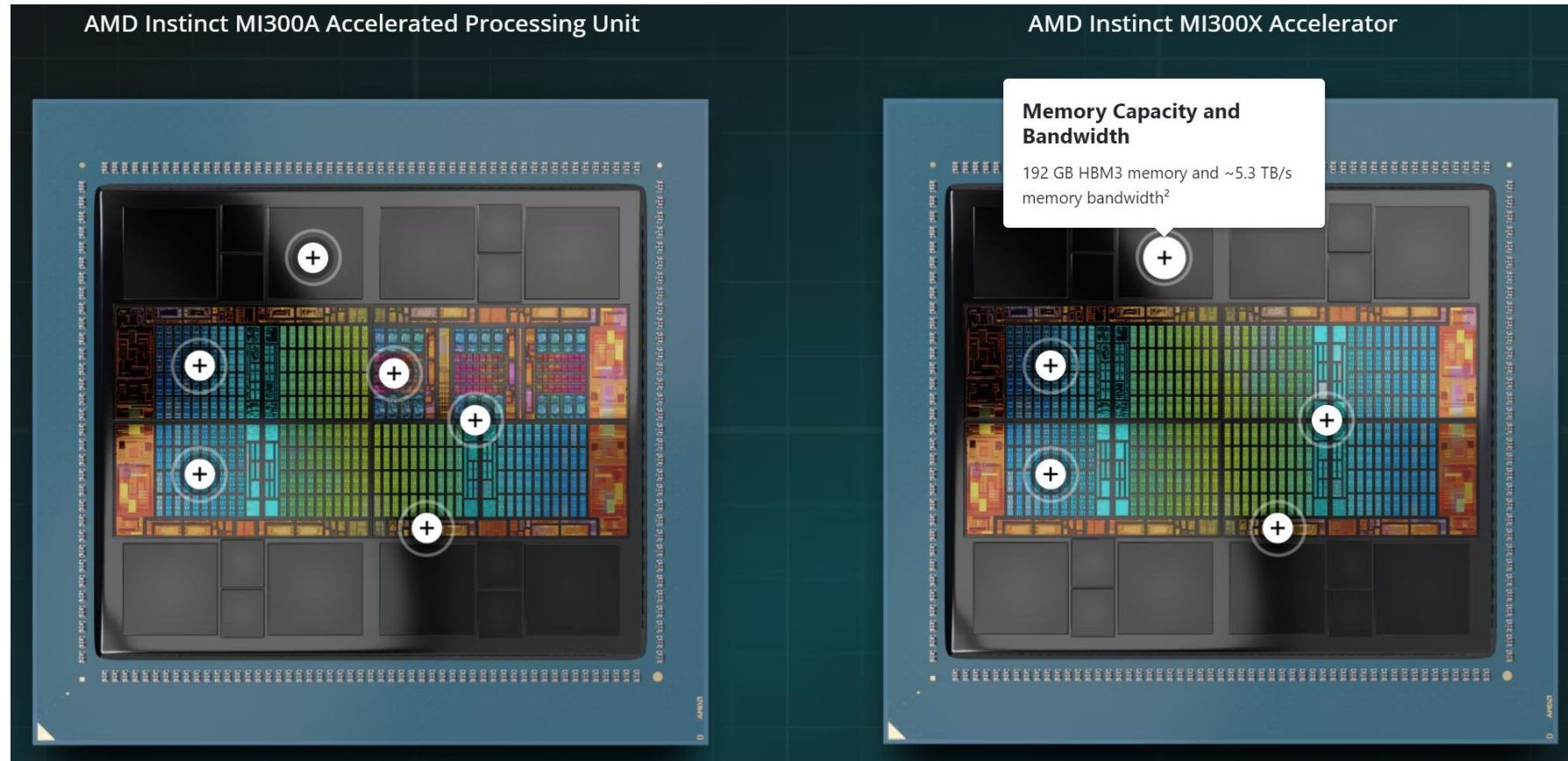
12 | THE ROAD TO THE AMD “FIJI” GPU | ECTC 2016 | MAY 2015



[https://www.ectc.net/files/66/5/66thECTC\\_Panel\\_BlackAMD.pdf](https://www.ectc.net/files/66/5/66thECTC_Panel_BlackAMD.pdf)

©2024 Marc Greenberg Consulting, LLC  
All Rights Reserved

# AMD continues in the chiplet direction



<https://www.amd.com/en/technologies/cdna.html>

©2024 Marc Greenberg Consulting, LLC  
All Rights Reserved





# Mix and match, partitioning

AMD Official Use Only - General

## AMD Instinct™ MI300A Memory System

13 chiplets as a single APU

- Four IOD, Three CCD and Six XCD
- Infinity Fabric AP and 3D packaging

128 Ch Fine-grained Interleaved

AMD Infinity Cache™

- 256 MB at 17 TB/s peak BW
- XCD Bandwidth amplification
- HBM power reduction
- Multi-XCD and CCD cache coherence
- Prefetcher for CPU memory latency

Unified HBM and Infinity Cache

- CCD and XCD data sharing
- Reduced data movement
- Simplified programming

The diagram illustrates the MI300A memory system with 13 chiplets. On the left, there are three CCD chiplets, six XCD chiplets, and four IOD chiplets. Each chiplet is connected to a central vertical bus (red, yellow, green, blue). On the right, there are four HBM stacks. The bus connects to the HBM stacks through the IOD chiplets. The bus is labeled with 'TSV' and 'CM' on the left side.

AMD together we advance.

10 | AMD INSTINCT™ MI300 ARCHITECTURE BRIEFING | UNDER EMBARGO UNTIL DECEMBER 6, 2023

AMD Official Use Only - General

## AMD Instinct™ MI300X Memory System

12 chiplets as a single device

- Four IOD and Eight XCD
- Infinity Fabric AP and 3D packaging

128 Ch Fine-grained Interleaved

AMD Infinity Cache™

- 256 MB at 17 TB/s peak BW
- XCD Bandwidth amplification
- HBM power reduction
- Multi-XCD cache coherence

The diagram illustrates the MI300X memory system with 12 chiplets. On the left, there are eight XCD chiplets and four IOD chiplets. Each chiplet is connected to a central vertical bus (red, yellow, green, blue). On the right, there are four HBM stacks. The bus connects to the HBM stacks through the IOD chiplets. The bus is labeled with 'TSV' and 'CM' on the left side.

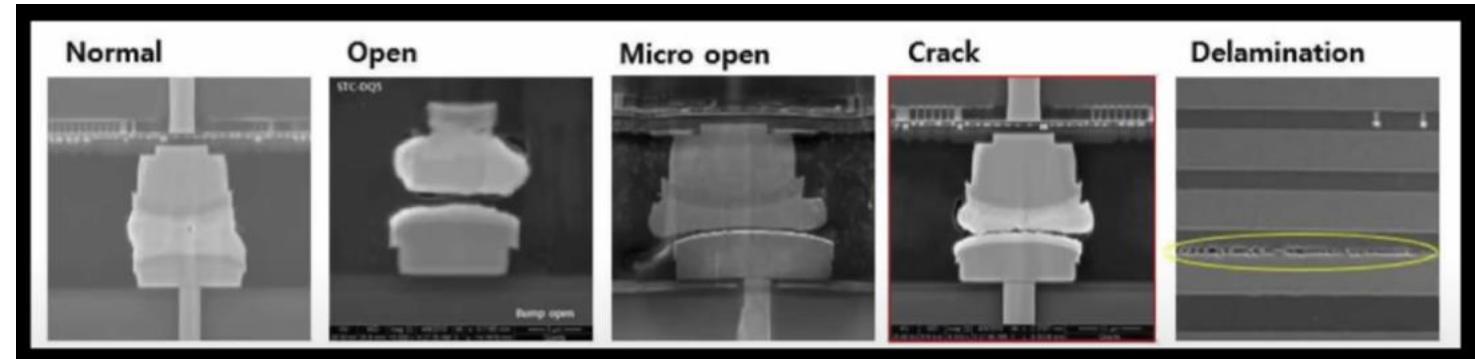
AMD together we advance.

9 | AMD INSTINCT™ MI300 ARCHITECTURE BRIEFING | UNDER EMBARGO UNTIL DECEMBER 6, 2023

<https://www.club386.com/amd-instinct-mi300-architecture-speaks-to-massive-ai-performance/>



# HBM Isn't easy



Tech Industry > Artificial Intelligence

(provider) GPUs and HBM3 memory caused half of failures during (user) training — one failure every three hours for (user) 16,384 GPU training cluster

News By Anton Shilov published July 27, 2024

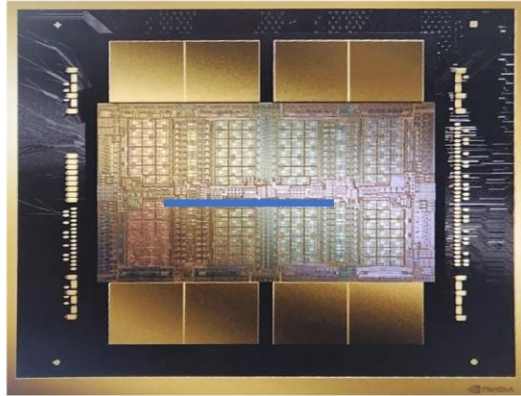
But (user) knows how to mitigate the issues.

Assumptions:  
16,384 GPUs  
6 HBM per GPU  
8Gbps per pin  
1024 pins per HBM  
= $8 \times 10^{17}$ bps \* 10800s  
=1 error per  $8.6 \times 10^{21}$  bits

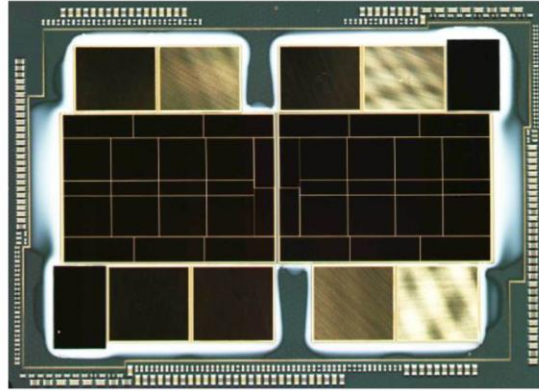


<https://www.tomshardware.com/>

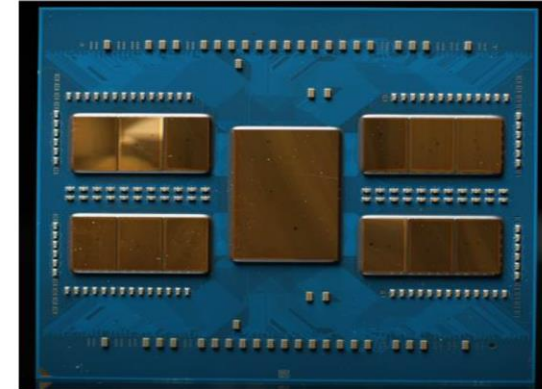
# Gratuitous Die Photos



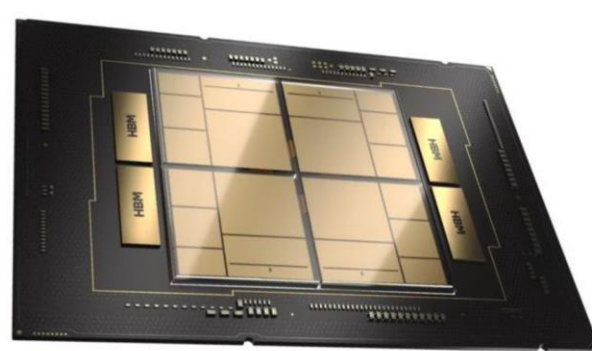
Nvidia Blackwell



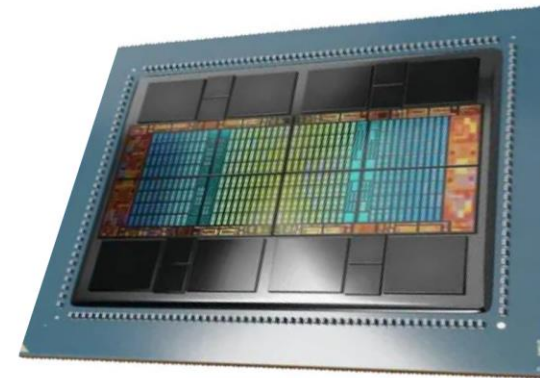
Intel Ponte Vecchio



AMD EPYC



Intel Sapphire Rapids



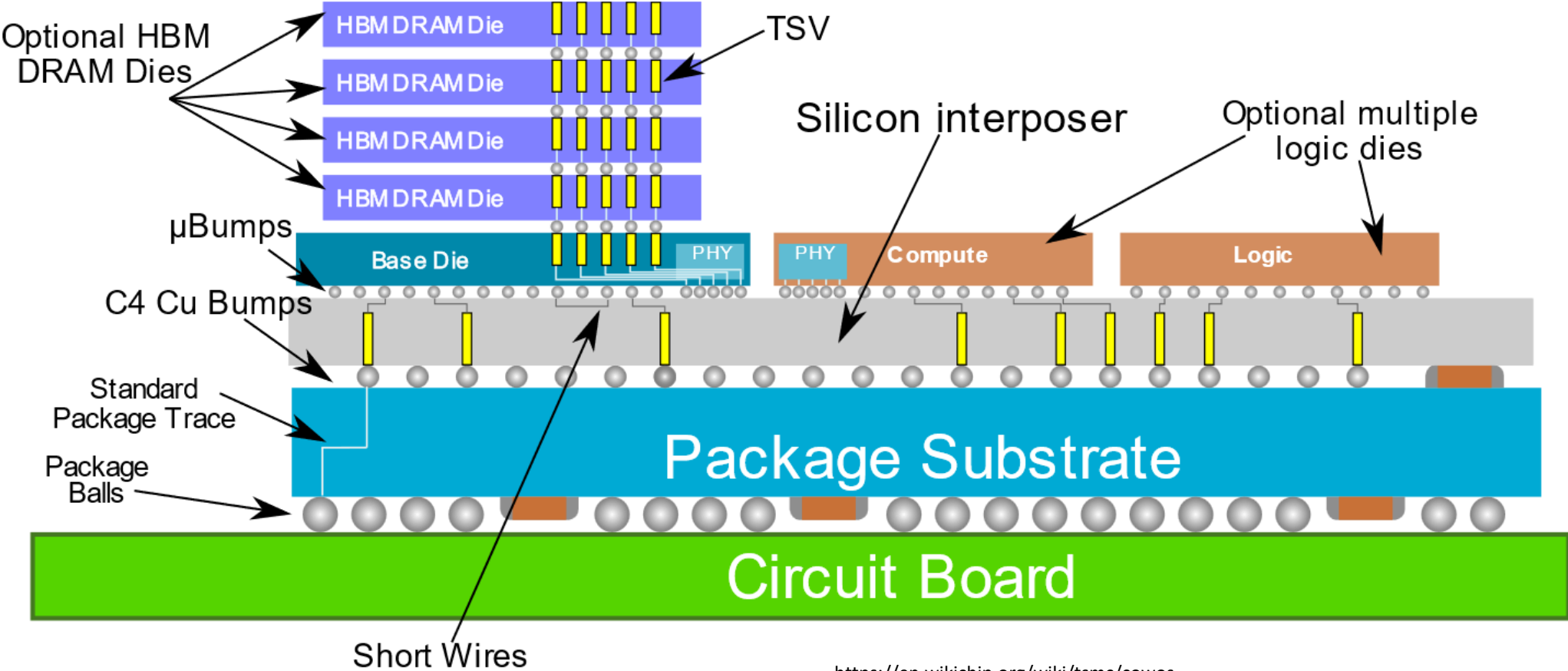
AMD MI350



# HBM



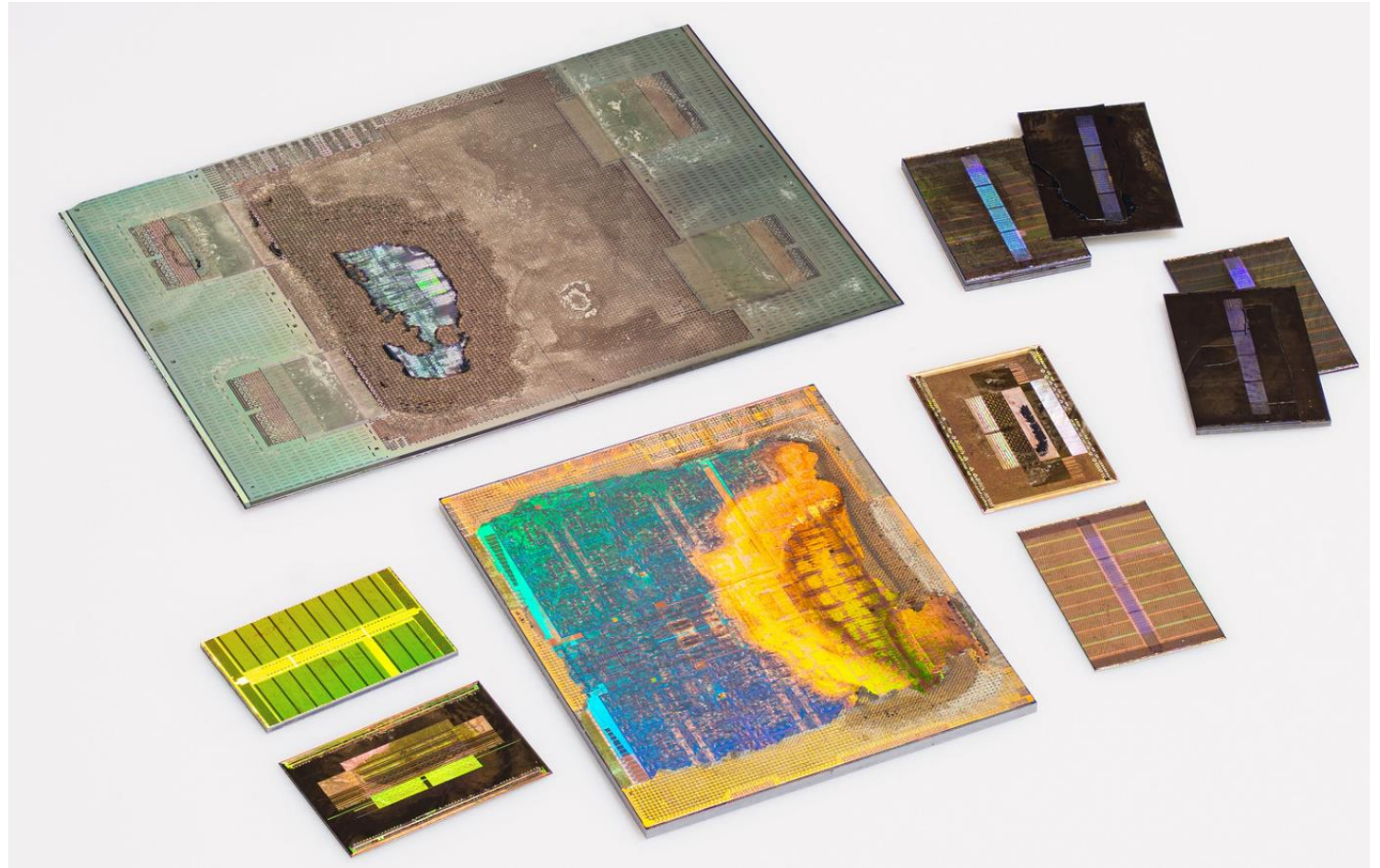
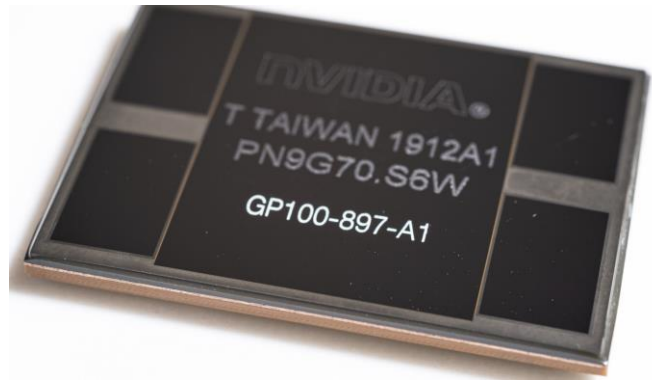
# HBM Cross-section



<https://en.wikichip.org/wiki/tsmc/cowos>



# Some disassembled HBM photos



<https://www.flickr.com/photos/130561288@N04/albums/72177720295479734/with/52207241684>

©2024 Marc Greenberg Consulting, LLC  
All Rights Reserved



# AMD MI300 Chipllet stackup

[AMD Official Use Only - General]

## AMD Instinct™ MI300 Family

### 3.5D Advanced Packaging

Unique hybrid architecture  
to support 3.5D TDP

3D Hybrid Bonded Architecture compute density and perf/W

2.5D Architecture for IOD-IOD and HBM3 integration

Large module on substrate

AMD  
together we advance\_

16 | AMD INSTINCT™ MI300 ARCHITECTURE BRIEFING | UNDER EMBARGO UNTIL DECEMBER 6, 2023



<https://www.servethehome.com/amd-instinct-mi300x-gpu-and-mi300a-apus-launched-for-ai-era/amd-instinct-mi300-family-architecture-chip-stack/>

©2024 Marc Greenberg Consulting, LLC  
All Rights Reserved

# Fundamentals of HBM Operation

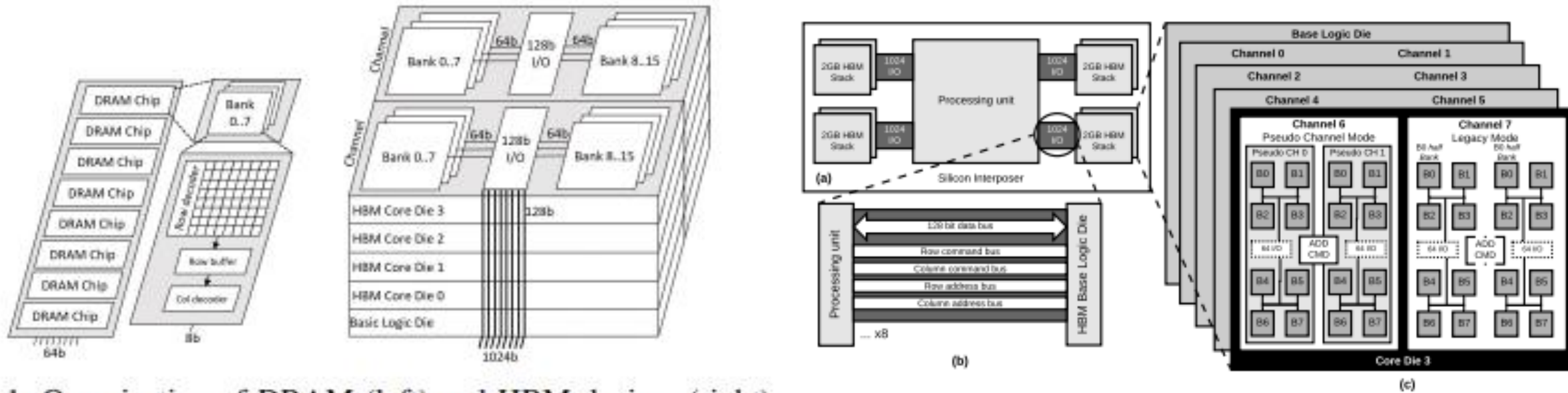


Fig. 1: Organization of DRAM (left) and HBM devices (right).

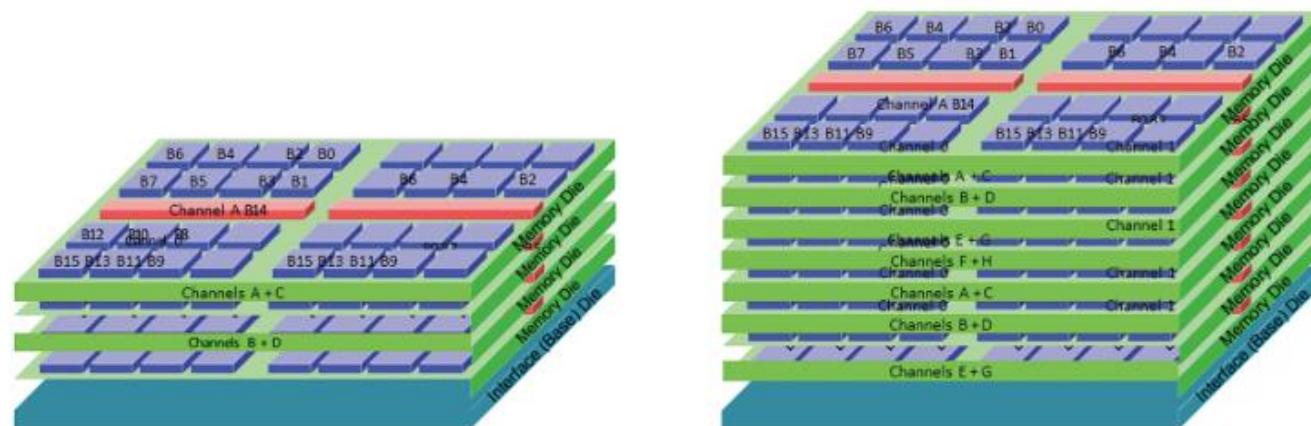
[https://www.ece.mcmaster.ca/faculty/hassan/assets/publications/hbm\\_iccad2021.pdf](https://www.ece.mcmaster.ca/faculty/hassan/assets/publications/hbm_iccad2021.pdf)





# HBM3 Banks and Channels/Pseudochannels

- 8 DWORD channels for data
  - 128 bits wide, divided into 2 64-bit Pseudochannels
  - Total 1024 Data pins
- AWORD for Command/Address
- BL2
- 64 banks per channel



<https://www.2cm.com.tw/2cm/zh-tw/tech/7C42130C5D1645C8884E53E62E27533E>

Massively Parallel Memory Architecture



# HBM3 RAS Features

- Parity
- Redundancy and remapping
  - LANE\_REPAIR
  - SOFT\_LANE\_REPAIR
  - HARD\_LANE\_REPAIR
- Loopback
- ECC / On-die ECC (Symbol based – SEV signal gives ECC status)
  - Auto ECS
- Test (IEEE 1500)



# Summary Section



# Financial Math

- 2022 cost analysis
- HPC Data center cost HBM vs DDR5
- A word on HBM Economics:
  - If you assemble a module with HBM, the HBM inventory might be yours



Technical Research Study | High-Bandwidth Memory (HBM) Architecture CPU Revolutionizes High-Performance Computing (HPC)

Table 7 | Equivalent node costs (rounded to nearest dollar)

Node Component	2S Intel® Xeon® Max Series Processor with Zero DIMMs	2S AMD EPYC™ 7773X Processor with 32 GB DIMM
A. Total server (node) cost (from Table 2)	\$31,400	\$24,136
B. Equivalent nodes (from Table 6)	33	100
C. Equivalent node cost (C = A x B)	\$1,036,200	\$2,413,600
	57% less	-

This comparison tells us that the cost of a 100-node server cluster powered by Intel Xeon Max Series processors will be 57 percent less than the cost of a 100-node server cluster run by AMD EPYC 7773X processors. However, it will still deliver the same performance. The performance of the Intel Xeon Max Series processors helps significantly lower TCO at scale because so many fewer servers are needed.

The biggest assumption in our analysis was the equivalent nodes of 100 to 33. Even if the equivalent nodes were 100 to 50 or 100 to 75, it would still cost less to run a server cluster powered by Intel Xeon Max Series processors, as compared to a server cluster run by AMD EPYC 7773X processors.

<https://www.prowesscorp.com/wp-content/uploads/2022/12/220126-Intel-HBM-Architecture-CPU-Revolutionizes-HPC-technical-research-study.pdf>

# GDDR vs LPDDR vs HBM

Property	HBM4	HBM3E	HBM3	HBM2E	GDDR7	GDDR6	LPDDR5X	DDR5
Max capacity per stack, chip or module	36-64GB	36GB	24GB	16GB	3GB	2GB	16GB	2048GB
Data Transfer Rate	6.4GT/s	8.8GT/s	6.4GT/s	3.6GT/s	32GT/s	24GT/s	9.6GT/s	8.4GT/s
Max stack	16	12	12	8	1*	1*	8	8-16
Interface Width	2048	1024	1024	1024	32	32	32-64	64
Signaling	Undisclosed	NRZ	NRZ	NRZ	PAM3	NRZ	NRZ	NRZ
I/O Voltage		1.1v	1.1v	1.2v	1.2v	1.2-1.35v	0.4v	1.1v
Bandwidth per stack, chip or module	1500-2000GB/s	1200GB/s	819GB/s	406GB/s	128GB/s	96GB/s	77GB/s	67GB/s

\* Designed for clamshell implementation on PCB generating a virtual 2-stack

Mostly derived from <https://www.embedded.com/wp-content/uploads/sites/2/2024/01/memory-bandwidth-table-2023-002.jpg>

©2024 Marc Greenberg Consulting, LLC

All Rights Reserved

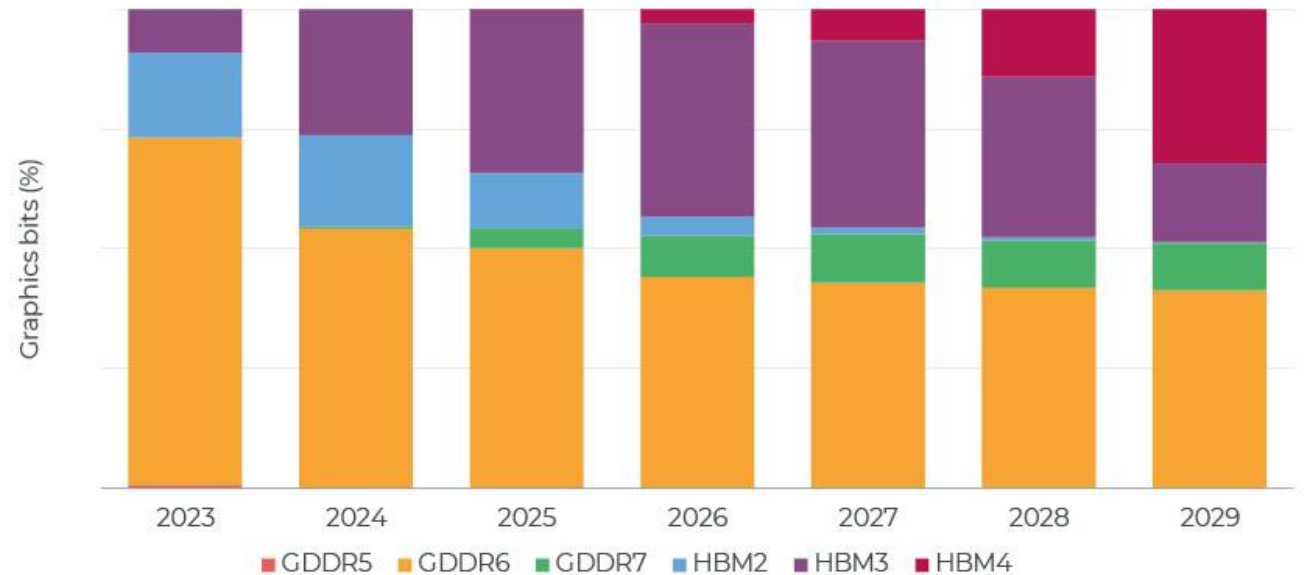


# GPU memory mix

- Forecast rapid growth of HBM4
- Bit split evenly between GDDR and HBM

## 2023-2029 graphics & AI DRAM mix of technology

(Source: DRAM Market Monitor Q2 2024, Yole Intelligence, June 2024)

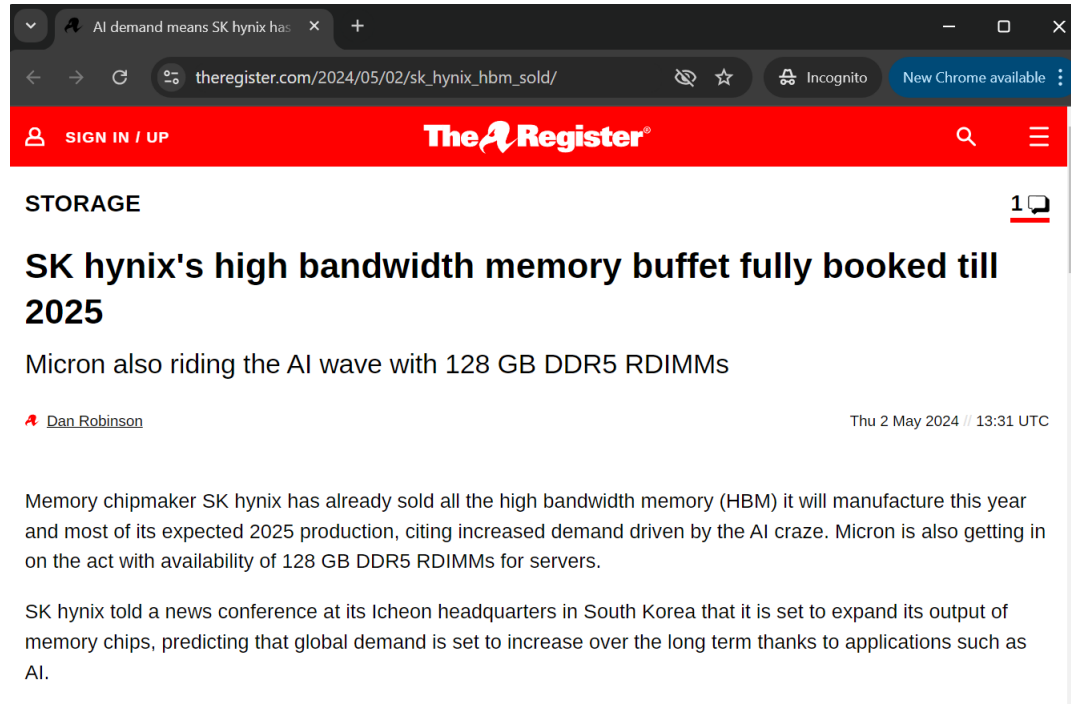


© Yole Intelligence 2024

<https://www.yolegroup.com/product/monitor/dram-market-monitor/>



# Supply Challenges



AI demand means SK hynix has

theregister.com/2024/05/02/sk\_hynix\_hbm\_sold/

**The Register**

**STORAGE**

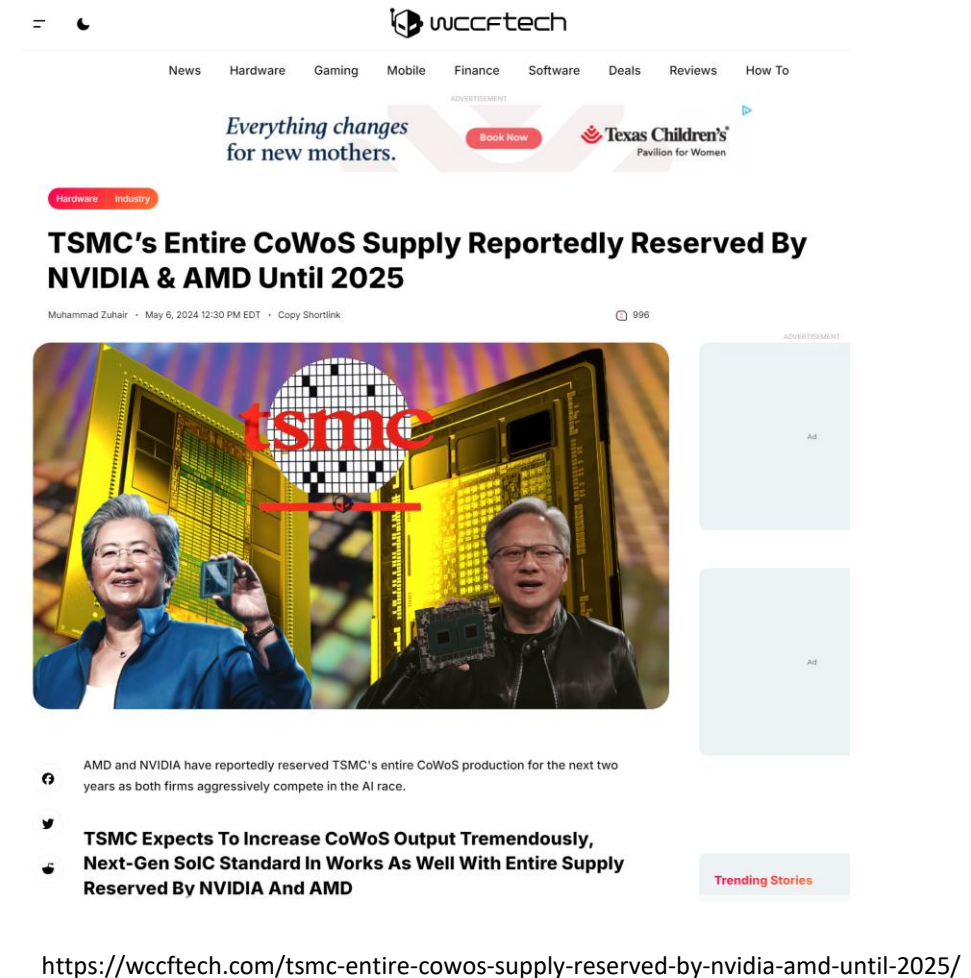
## SK hynix's high bandwidth memory buffet fully booked till 2025

Micron also riding the AI wave with 128 GB DDR5 RDIMMs

**Dan Robinson** Thu 2 May 2024 // 13:31 UTC

Memory chipmaker SK hynix has already sold all the high bandwidth memory (HBM) it will manufacture this year and most of its expected 2025 production, citing increased demand driven by the AI craze. Micron is also getting in on the act with availability of 128 GB DDR5 RDIMMs for servers.

SK hynix told a news conference at its Icheon headquarters in South Korea that it is set to expand its output of memory chips, predicting that global demand is set to increase over the long term thanks to applications such as AI.



wccfttech

News Hardware Gaming Mobile Finance Software Deals Reviews How To

ADVERTISEMENT


Everything changes for new mothers. [Book Now](#) **Texas Children's Pavilion for Women**

Hardware Industry

## TSMC's Entire CoWoS Supply Reportedly Reserved By NVIDIA & AMD Until 2025

Muhammad Zuhair • May 6, 2024 12:30 PM EDT • Copy Shortlink 996

ADVERTISEMENT



ADVERTISEMENT

AMD and NVIDIA have reportedly reserved TSMC's entire CoWoS production for the next two years as both firms aggressively compete in the AI race.

**TSMC Expects To Increase CoWoS Output Tremendously, Next-Gen SoIC Standard In Works As Well With Entire Supply Reserved By NVIDIA And AMD**

Trending Stories

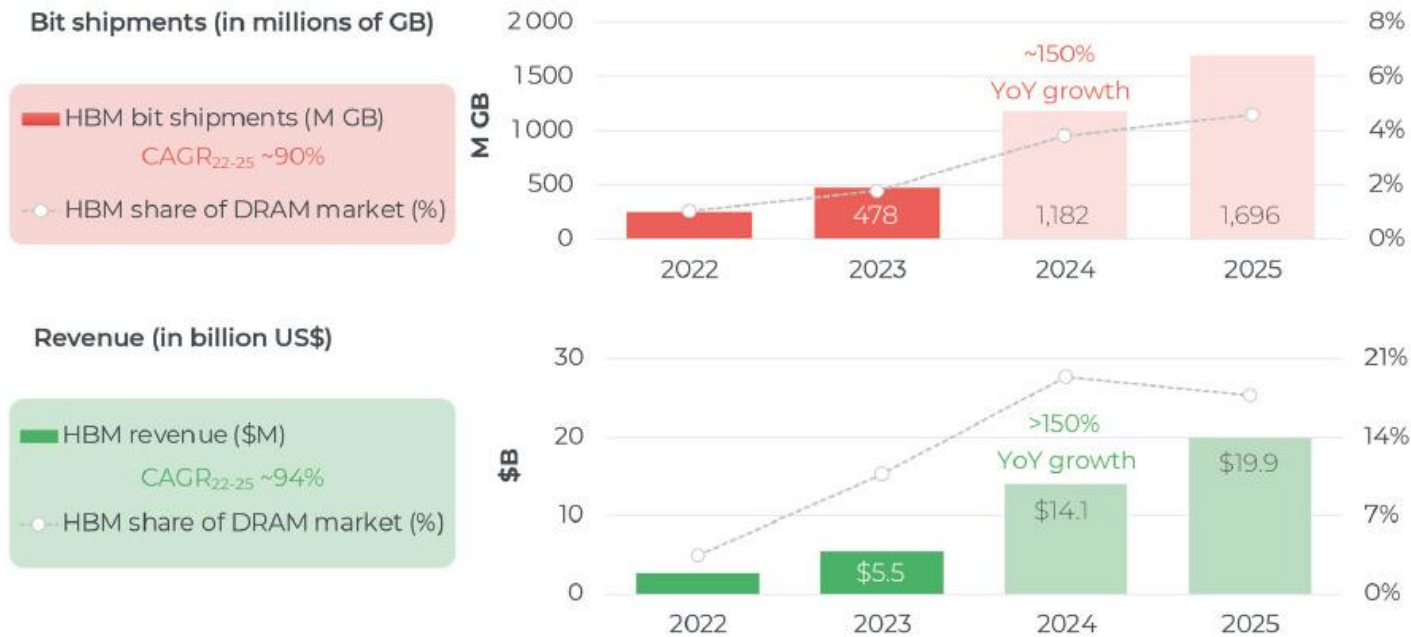
<https://wccfttech.com/tsmc-entire-cowos-supply-reserved-by-nvidia-amd-until-2025/>



# HBM market size

## 2022-2025 High Bandwidth Memory (HBM) market evolution

(Source: Next-Generation DRAM 2024 – Focus on HBM and 3D DRAM, Yole Intelligence, January 2024)



5% of bits shipped

15-20% of revenue



© Yole Intelligence 2024

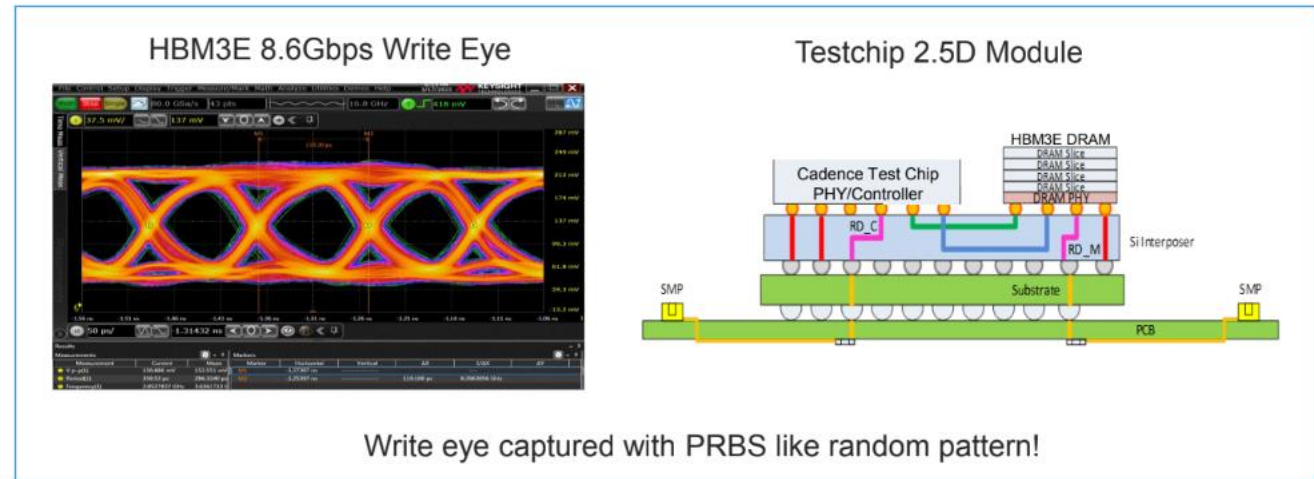
©2024 Marc Greenberg Consulting, LLC  
All Rights Reserved



# IP Availability

- IP availability generally in nodes from 7nm to =<3nm
- Multiple suppliers
  - IP Vendors
  - ASIC vendors

## HBM3E PHY at 8.4Gbps Write Eye Diagram Industry's fastest HBM3E



- HBM3E 8.4Gbps at core voltage – No Overdrive needed
- Full 2.5D stack design by Cadence with reference design for customers

© 2024 Cadence Design Systems, Inc. All rights reserved.

cadence

Example: Cadence, used with permission



# HBM4

- Announced standard changes:
- 2x channels
- 24Gb and 32Gb die configurations
- 4, 8, 12, and 16 high TSV stacks
- speed pins up to 6.4Gbps with discussion ongoing for higher frequencies



## JEDEC Approaches Finalization of HBM4 Standard, Eyes Future Innovations

ARLINGTON, Va., USA – July 10, 2024 – JEDEC Solid State Technology Association, the global leader in the development of standards for the microelectronics industry, today announced it is nearing completion of the next version of its highly anticipated High Bandwidth Memory (HBM) DRAM standard: HBM4. Designed as an evolutionary step beyond the currently published HBM3 standard, HBM4 aims to further enhance data processing rates while maintaining essential features such as higher bandwidth, lower power consumption, and increased capacity per die and/or stack. These advancements are vital for applications that require efficient handling of large datasets and complex calculations, including generative artificial intelligence (AI), high-performance computing, high-end graphics cards, and servers.

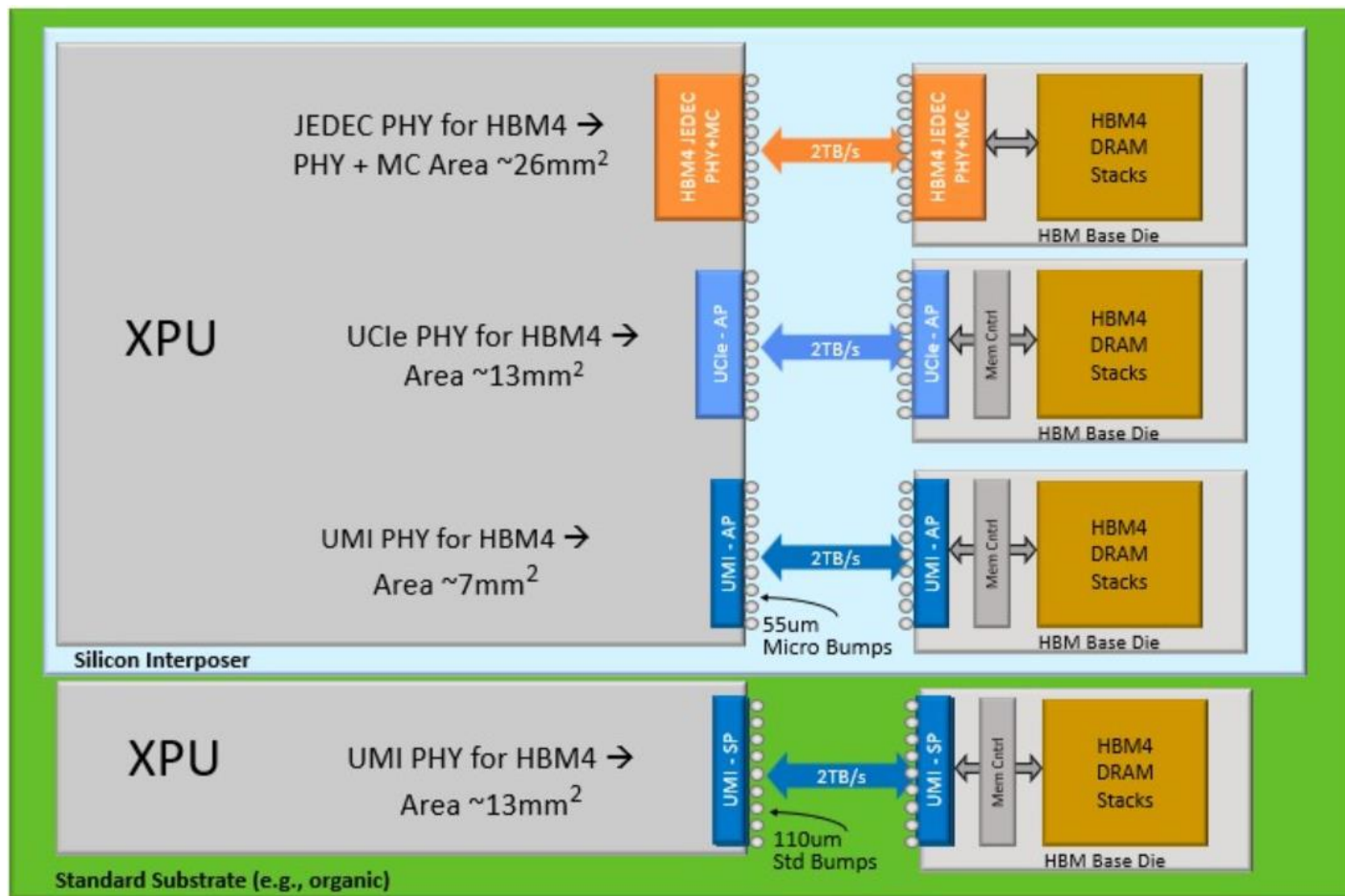
HBM4 is set to introduce a doubled channel count per stack compared to HBM3, with a larger physical footprint. To support device compatibility, the standard ensures that a single controller can work with both HBM3 and HBM4 if needed. Different configurations will require various interposers to accommodate the differing footprints. HBM4 will specify 24 Gb and 32 Gb layers, with options for supporting 4-high, 8-high, 12-high and 16-high TSV stacks. The committee has initial agreement on speeds bins up to 6.4 Gbps with discussion ongoing for higher frequencies.

JEDEC encourages companies to join and help shape the future of JEDEC standards. Membership grants access to pre-publication proposals and provides early insights into active projects like HBM4. [Discover the benefits of membership and join today.](#)

*JEDEC standards are subject to change during and after the development process, including disapproval by the JEDEC Board of Directors.*

<https://www.jedec.org/news/pressreleases/jedec-approaches-finalization-hbm4-standard-eyes-future-innovations>

# What could be in the future: different I/O



<https://www.nextplatform.com/2024/03/28/how-to-build-a-better-blackwell-gpu-than-nvidia-did/>



# What could be in the future: PIM

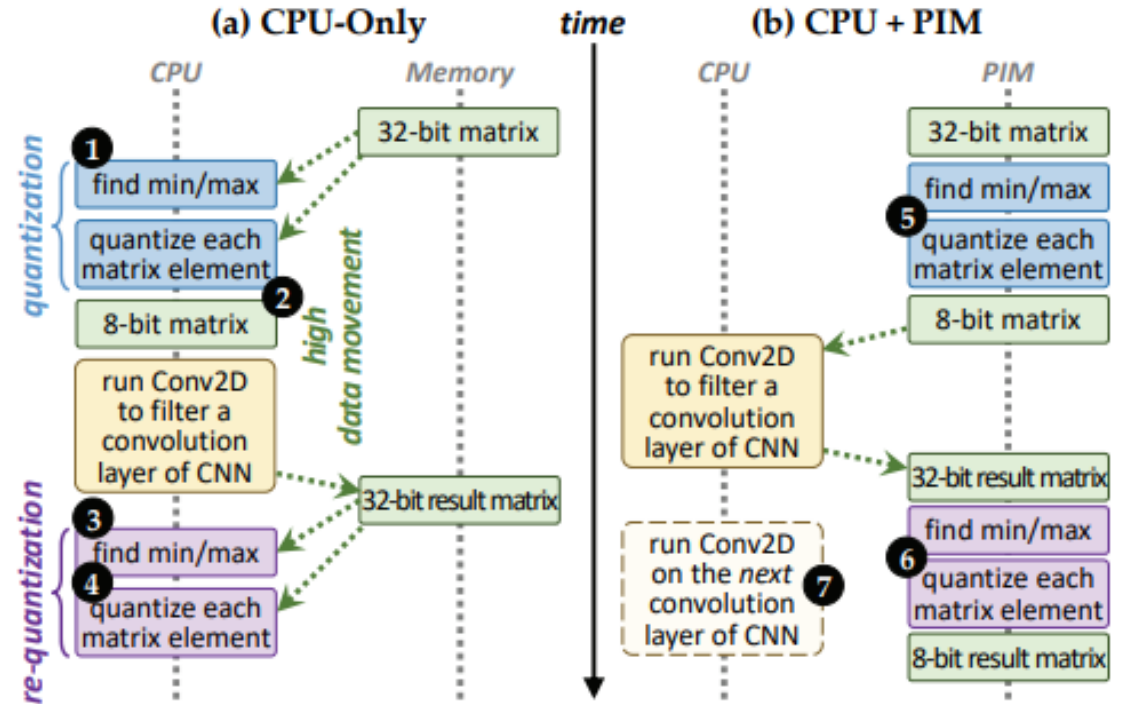
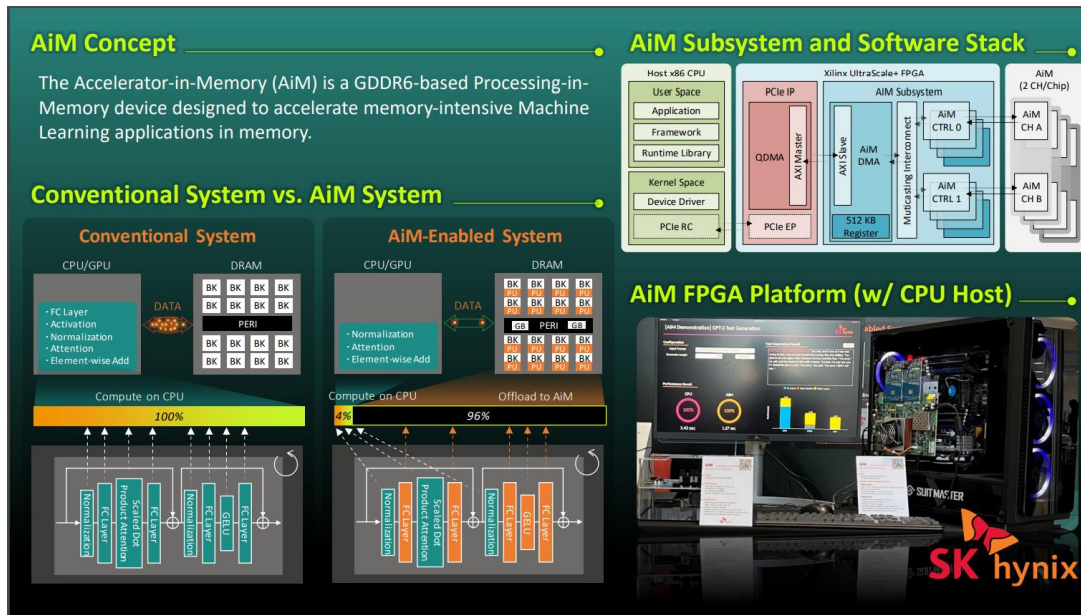


Figure 8. Quantization on (a) CPU vs. (b) PIM.

<https://www.pdl.cmu.edu/PDL-FTP/associated/asplos18-pim-final.pdf>



# Summary

- HBM, Chiplet, AI technology are intrinsically linked
- A robust ecosystem for all the components is available
- A roadmap for higher levels of memory bandwidth is assured
  
- Questions?      [marc@marcgreenberg.com](mailto:marc@marcgreenberg.com)

