Suhas Kamath
DS 221: INTRODUCTION TO SCALABLE SYSTEMS
October 25, 2025

Parallel Programming - Written Assignment

Memory Bandwidth Evolution in Architectures

Moore's law, as originally stated in 1965, observed that the number of transistors on a microchip doubles approximately every two years [1]. This exponential growth in computational density has been a primary driver of performance gains for decades.

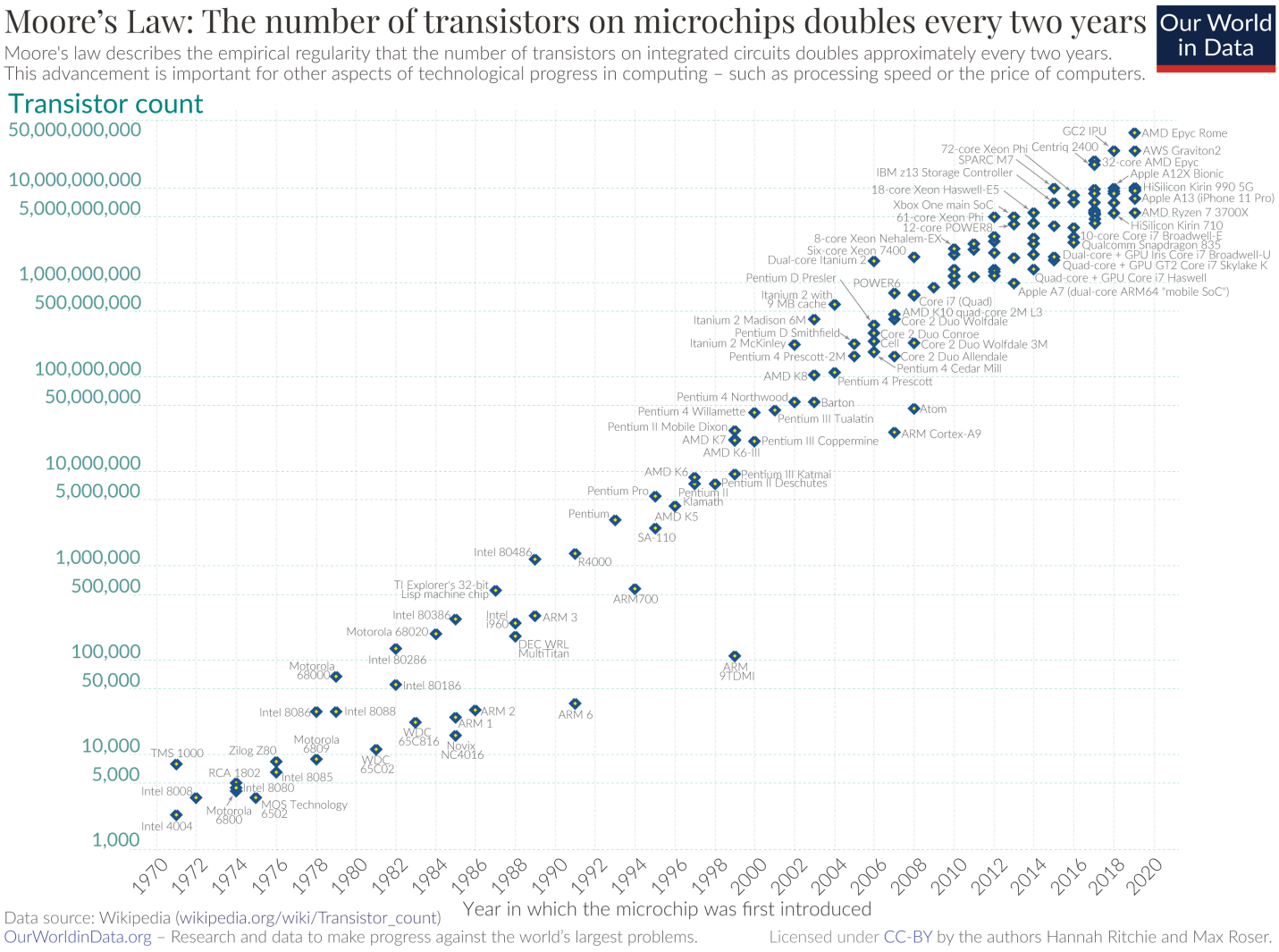

Figure 1. A graph depicting the increase in transistor count against the year. [2]

In stark contrast, memory bandwidth has improved at a much slower, linear rate of approximately 7% per year [3]. This growing disparity creates the "memory wall"—a critical performance bottleneck where the processor is significantly faster than the memory system supplying its data. Consequently, powerful computational cores are often left idle, "starved" for data, which limits the real-world performance of the entire system.

To quantitatively analyze this trend, we use several evaluation metrics:

1) **Memory Bandwidth (GB/s)**: This metric denotes the theoretical peak rate at which data can be read from or written to memory, measured in Gigabytes per second. It is a fundamental measure of a memory subsystem's throughput.

2) **Bandwidth-per-TFLOP**: This metric evaluates the balance of a processor's architecture, particularly for GPUs and HPC accelerators. It quantifies how much memory bandwidth (in GB/s) is available for every trillion floating-point operations per second (TFLOP) the processor can execute. A declining value across hardware generations indicates that compute capabilities are outpacing memory bandwidth, increasing the risk of bottlenecks for memory-intensive applications.

$$\text{Bandwidth-per-TFLOP} = \frac{\text{Memory Bandwidth (GB/s)}}{\text{Peak TFLOPs}}$$

3) **Compute-to-Memory Ratio (Arithmetic Intensity)**: This ratio quantifies the number of arithmetic operations performed per byte of data moved from memory. It is a property of both the algorithm and the hardware. Algorithms with high arithmetic intensity are compute-bound and can effectively utilize powerful processors, while those with low intensity are memory-bound and are limited by the system's memory bandwidth. The following formula can be used to calculate the compute-to-memory ratio:

$$\text{Compute-to-Memory (TFLOPs/GB)} = \frac{1}{\text{Bandwidth-per-Compute (GB/TFLOP)}}$$

Typically, raw bandwidth (GB/s) is the primary metric for general-purpose CPU memory, while the latter two ratios are crucial for analyzing the performance of massively parallel GPUs.

CPU Memory Trends. Over the past decade, CPU memory has evolved across three main technologies, each targeting different performance segments.

- **DDR4** (Double Data Rate 4th generation)
- **DDR5** (Double Data Rate 5th generation)
- **HBM** (High Bandwidth Memory)

DDR4, standardized by JEDEC[1] in 2012, became the workhorse for servers and consumer PCs [4]. It supports up to 64GB per DIMM (Dual In-line Memory Module) and offers a peak bandwidth of 25.6 GB/s per channel at its maximum standard speed of 3200 MT/s [5]. In a typical dual-channel consumer system, this provides $\approx$ 51.2 GB/s of total bandwidth.

DDR5 was standardized by JEDEC in 2020 and represented a major architectural shift [6]. By using two smaller 32-bit sub-channels per DIMM, it improves channel efficiency and signal integrity. The development of 32Gb monolithic DRAM chips by companies like Samsung enables the future manufacturing of massive 128GB consumer DIMMs and server DIMMs up to 1TB [7]. At a standard speed of 6400 MT/s, DDR5 delivers a peak bandwidth of 51.2 GB/s per channel, doubling that of DDR4. High-end servers with an 8-channel memory configuration can achieve a total system bandwidth of over 400 GB/s.

HBM is a paradigm shift, moving memory from external DIMM slots directly onto the CPU package, typically serving as a large L4 cache or a separate memory region [8]. This is achieved by stacking DRAM dies vertically. SK Hynix was the pioneer of this technology [9]. Intel's Xeon Max Series CPUs (Sapphire Rapids) integrate 64GB of HBM2e that can be configured in three distinct modes [10]:

- **Cache Mode:** HBM acts as a large, fast L4 cache.
- **Flat Mode:** HBM and DDR5 are presented to the OS as two separate NUMA nodes. This allows developers to explicitly allocate data to either the high-speed HBM or the high-capacity DDR memory, depending on the specific application.
- **HBM-Only Mode:** The system runs entirely on the on-package HBM.

[1] JEDEC is an organization that develops open standards and publications in the microelectronics industry.

As shown below, on-package HBM provides a significant bandwidth advantage over traditional DDR5 memory:

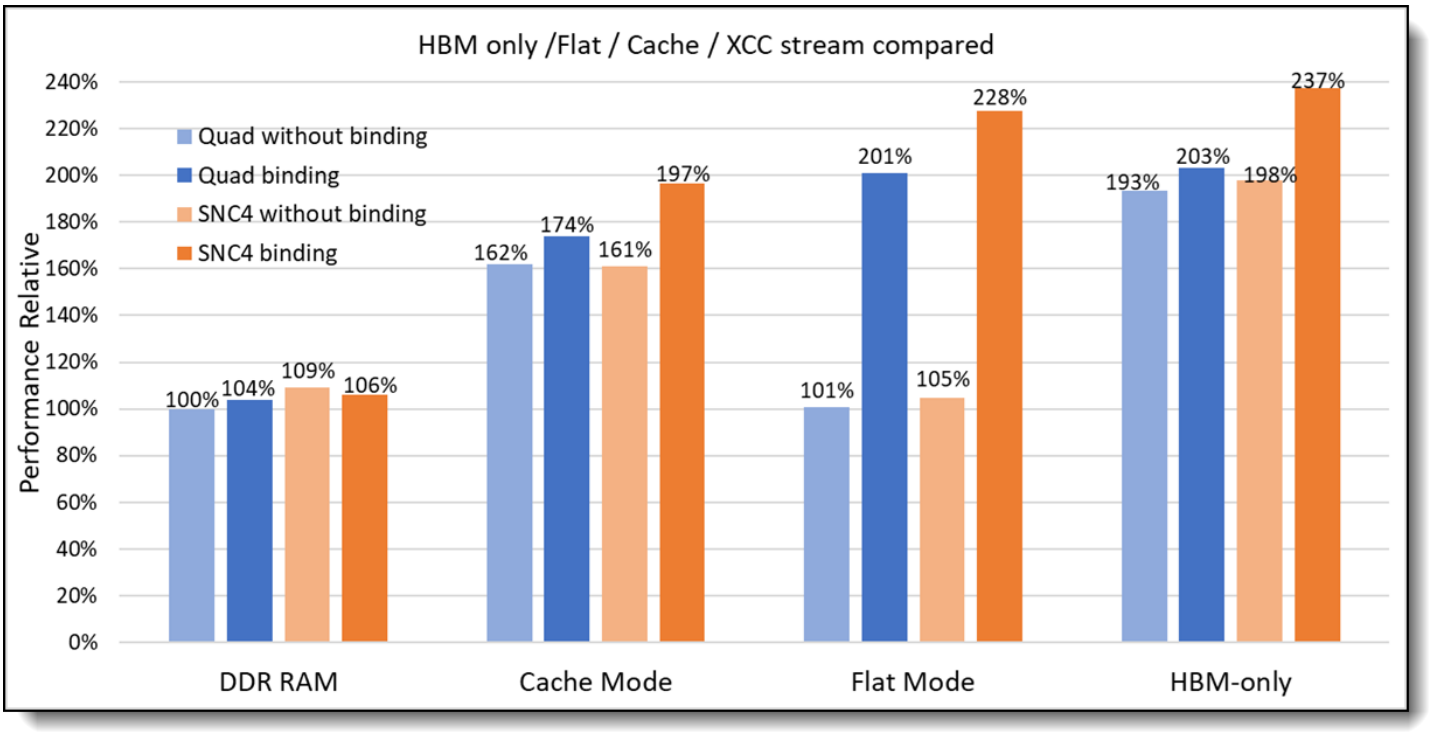

FIGURE 2. Memory bandwidth of HBM and DDR Memory [10]

Overall, peak server memory bandwidth has increased from 200 GB/s (8-channel DDR4) to over 1 TB/s (HBM2e), a 5-fold increase. In the same 10-year period, assuming the common simplification of Moore's Law where compute doubles every two years, CPU compute has increased by a factor of $2^{(10/2)} = 2^5 = 32$. This widening gap highlights the persistent challenge of the memory wall.

GPU Memory Trends. The evolution of GPU memory has been even more aggressive to feed their massively parallel compute engines. This has led to two parallel development tracks: high-speed GDDR for consumer cards and ultra-wide HBM for data center accelerators.

- **GDDR5, GDDR5X, GDDR6, GDDR6X, GDDR7**
- **HBM, HBM2, HBM2E, HBM3, HBM3E, HBM4**

The architectural trade-offs between GDDR and HBM are significant. GDDR focuses on achieving high data rates per pin (measured in Gbps) over a relatively narrow memory bus (e.g., 256-bit). In contrast, HBM uses a much wider bus (e.g., 1024-bit) but at lower clock speeds. This is made possible by stacking DRAM dies vertically using Through-Silicon Vias (TSVs) and placing them very close to the GPU on a silicon interposer.

The technical specifications of these technologies are summarized below:

TABLE 1. Table depicting various GPUs released over the past decade, along with their specifications.

Type	**Year**	**Memory Bandwidth**	**Bandwidth per Compute (GB/T-FLOP)**	**Compute to Memory (TFLOPs/GB)**	**Example GPU**
GDDR-5	2008 [11]	243 GB/s [12]	0.0458	21.83	NVIDIA® Quadro P4000
GDDR-5X	2016 [13]	432 GB/s [14]	0.0360	27.78	NVIDIA® Quadro P6000
GDDR-6	2018 [15]	768 GB/s [16]	0.0198	50.51	NVIDIA® RTX A6000
GDDR-6X	2020 [17]	1008 GB/s [18]	0.0122	81.97	NVIDIA® GeForce RTX 4090
GDDR-7	2022 [19]	1792 GB/s [20]	5.64×10^{-3}	177.30	NVIDIA® GeForce RTX 5090
HBM	2013 [9]	512 GB/s [21]	0.0595	16.81	AMD® Radeon R9 Fury X 4G
HBM-2	2016 [22]	1044 GB/s [23]	0.0708	14.12	AMD® Radeon Instinct MI60
HBM-2E	2019 [24]	2039 GB/s [24]	0.105	9.52	NVIDIA® A100 Tensor Core GPU
HBM-3	2022 [25]	5.3 TB/s [26]	0.0324	30.86	AMD® Instinct™ MI300X Accelerator
HBM-3E	2023 [27]	8 TB/s [28]	0.0133	75.19	NVIDIA® HGX B200
HBM-4	2025 [29]	-	-	-	-

HBM-4 is a very recent standard, and hence no GPU have been released to the public under this standard.

Again in the above table, it is glaringly visible that memory is way behind compute. In other words, all of the above GPUs are not able to read a single byte per FLOP. This essentially leads to starving, where the fast processors are waiting for the data to be loaded from memory.

Conclusion. In conclusion, we have witnessed great evolution in memory technologies over the past decade. CPU evolution, however, always tends to outperform the memory evolution. CPUs evolve with respect to Moore's law, i.e. exponentially, whereas CPU memory evolves incrementally. Computer scientists, however, have tried to solve this problem by means of multi-leveled cache, and other similar technologies.

Chiplet Architecture and Heterogeneous Integration

A chiplet can be considered as a piece of a computer chip that has the ability to perform a specialised function efficiently. Several such chiplets can be combined to form a processor. Such processors are commonly known as Systems-on-Chip, or SoC for short. This idea is in opposition to the traditional "monolithic" processors, which comprise of a single piece of silicon, which is very generalised, i.e. it can perform multiple tasks, but none of the tasks can be optimized due to its very nature.

During the manufacturing process of a CPU or GPU, a large "wafer" [2] is basically cut into processors. If a wafer has a defect, we shall not be able to use that particular section of the wafer, and we shall have to "cut around" the wafer. Some examples of wafer damages are as follows. Due to this, we end up with several small pieces, and very few large pieces. This ends up raising the price of the larger processors, and also leads to wastage of the smaller pieces.

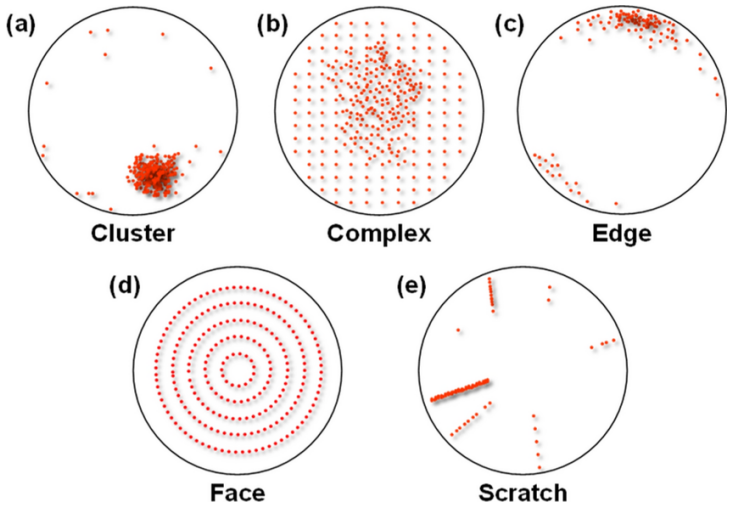

Figure 3. Various types of wafer damage. [30]

Using the chiplet approach can save on both money and time. Because each chiplet is specialised, it shall be able to do its assigned task much faster than a general-purpose chip. [31]

AMD. AMD has their proprietary chipset ecosystem. AMD has the following chiplet types in its ecosystem:

- **I/O Capable Chiplets**: These are the interface between the SoC and I/O devices.
- **Accelerator Chiplets**: These chiplets are used to offload tasks from the main processing cores, thereby improving overall performance. Some examples of such cores are:
  - Custom Arithmetic Engines: These chiplets are used to accelerate arithmetic-heavy algorithms.
  - Compression or encoding engines: These chiplets assist in compression or encoding of data for better storage or transmission.
  - Networking engines: These chiplets specialise in network communication, and do tasks such as checking CRCs, encapsulating data, and other important tasks required in network communication.
- **Compute Cores**: These chiplets are the ones that actually perform the main computation. These cores include CPU cores and GPU cores.

To integrate these chiplets, the following technologies are used by AMD:

- High-Speed Interconnects (Infinity Fabric ™): This high performance fabric ensure optimal data transfer.
- Advanced Memory Interfaces: The entire SoC access the main memory efficiently using high-performance memory interfaces such as DDR or HBM.

[2] A wafer is a thin, circular slice of silicon.

- Datacenter Reliability and Services: The TPA and the entire SoC operate with exceptional uptime, stability, and error correction capabilities.

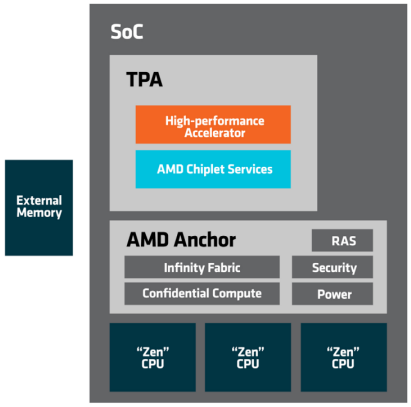

FIGURE 4. A pictoral depiction of AMD's chiplet ecosystem.

Intel. Intel, similar to AMD, also has its own chipset ecosystem. Intel uses the following chipsets:

- **Compute Chiplets**: These are the primary processors.
- **I/O Tiles**: These chiplets handle I/O functionality. They can be built on an older process nodes and can be reused from previous generations to save on cost and development time.
- **Base Tiles**: As the name suggests, these are a base for the other chiplets. A base tile contains logic and memory for tasks such as data caching and routing information between cores and I/O. Similar to I/O tiles, these tiles can also be manufactured on a previous generation node.

To connect the various chiplets, Intel uses the Advanced Interface Bus (AIB) technology. AIB is a physical layer specification, which has two implementations. AIB Base is intended for lighter applications and hence requires lesser circuitry. On the other hand, AIB Plus handles higher speed reliably. This is made possible for using Single Data Rate (SDR) for AIB Base and Double Data Rate (DDR) for AIB Plus.

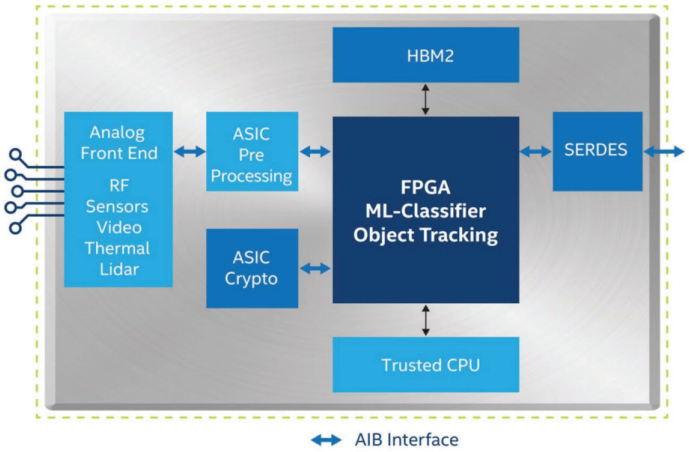

FIGURE 5. A pictoral depiction of Intel's chiplet ecosystem [32].

NVIDIA. NVIDIA employs two different chiplet-based design methodologies. The first one is basically having two of the exact same chip, connected over a high-speed link. A primary example of this is the NVIDIA Blackwell Architecture [33]. Another methodology is similar to what AMD and Intel are doing. Several different chiplets, which are each optimised for a special purpose, are combined into a single package. An example of this is NVIDIA Grace Hopper Superchip [34].

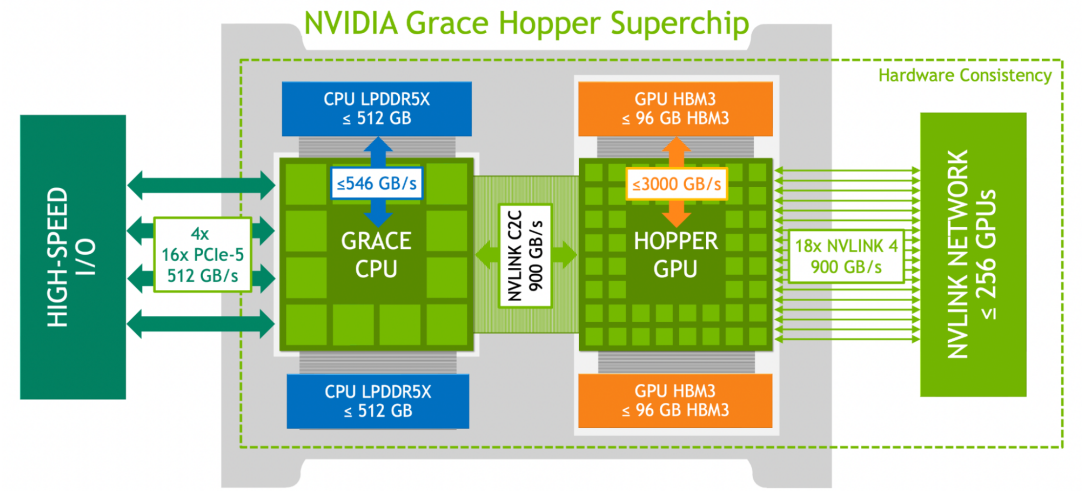

FIGURE 6. A pictorial representation of NVIDIA's Grace Hopper Architecture [34].

Again, when it comes to integration of chiplets, NVIDIA has two technologies. The first one is NVLink. NVLink is NVIDIA's proprietary, high-speed and power-efficient interconnect system [35]. Alternatively, NVIDIA also uses advanced chip packaging systems, namely TSMC Chip-on-Wafer-on-Substrate, i.e. CoWoS. CoWoS is an advanced 2.5-D chip packing technology developed by NVIDIA's foundry partner, TSMC [36].

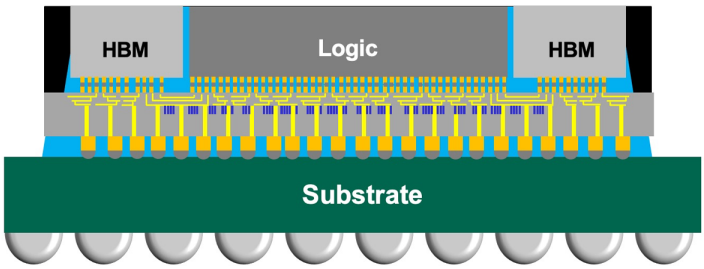

FIGURE 7. A pictorial depiction of CoWoS [36].

References

[1] G. E. Moore, "Cramming more components onto integrated circuits," 1965.

[2] H. Ritchie and M. Roser, "Moore's law: The number of transistors on microchips doubles every two years," 2020.

[3] A. Gholami, Z. Yao, S. Kim, C. Hooper, M. W. Mahoney, and K. Keutzer, "Ai and memory wall," 2024.

[4] JEDEC Solid State Technology Association, "JEDEC Announces Publication of DDR4 Standard." https://www.jedec.org/news/pressreleases/jedec-announces-publication-ddr4-standard, Sept. 2012. Press Release, JESD79-4.

[5] D. Wang, "Why migrate to DDR4?," Dec. 2013.

[6] R. Smith, "DDR5 Memory Specification Released: Setting the Stage for DDR5-6400 And Beyond," *AnandTech*, July 2020. Archived on the Wayback Machine.

[7] Samsung Electronics, "Samsung Electronics Unveils Industry's First and Highest-Capacity 12nm-Class 32Gb DDR5 DRAM, Ideal for the AI Era," *Samsung Global Newsroom*, Sept. 2023. Press Release.

[8] Anton Shilov, "Intel Confirms On-Package HBM Memory Support for Sapphire Rapids," *Tom's Hardware*, dec 2020.

[9] SK Hynix Inc., "These are the records of SK hynix's growth since 1983 until today." https://www.skhynix.com/company/UI-FR-CP05/.

[10] Sam Kuo and Jimmy Cheng, "Implementing High Bandwidth Memory and Intel Xeon Processors Max Series on Lenovo ThinkSystem Servers." https://lenovopress.lenovo.com/lp1738-implementing-intel-high-bandwidth-memory, June 2023.

[11] Alexandru Pancescu, "Samsung Pushes The GDDR5 Standard Forward," *Softpedia*, jul 2007.

[12] Nvidia Corporation, "Data Sheet: Quadro P4000," jun 2018.

[13] Emily Desjardins, "JEDEC Announces Publication of GDDR5X Graphics Memory Standard," *JEDEC*, jan 2016.

[14] Nvidia Corporation, "Data Sheet: Quadro P6000," may 2018.

[15] Samsung Corporation, "Samsung's 16Gb GDDR6 Memory Powers Latest NVIDIA Quadro Professional Graphics Solution," *Samsung*, aug 2018.

[16] Nvidia Corporation, "NVIDIA RTX A6000 datasheet," jan 2021.

[17] Anton Shilov, "Micron Reveals GDDR6X Details: The Future of Memory, or a Proprietary DRAM?," *Tom's Hardware*, sep 2020.

[18] Nvidia Corporation, "NVIDIA ADA GPU ARCHITECTURE," 2023.

[19] Bogdan Solca, "Samsung announces 36 Gbps GDDR7 memory standard, aims to release V-NAND storage solutions with 1000 layers by 2030," *NotebookCheck*, jun 2022.

[20] Nvidia Corporation, "NVIDIA RTX BLACKWELL GPU ARCHITECTURE," 2025.

[21] Micro-Star International Co., Ltd., "Radeon R9 Fury X 4G," 2023.

[22] Samsung Corporation, "Samsung Begins Mass Producing World's Fastest DRAM – Based on Newest High Bandwidth Memory (HBM) Interface," *Samsung*, jan 2016.

[23] TechPowerUp, "AMD Radeon Instinct MI60," nov 2018.

[24] Nvidia Corporation, "NVIDIA A100 TENSOR CORE GPU," jun 2021.

[25] JEDEC Solid State Technology Association, "JEDEC Publishes HBM3 Update to High Bandwidth Memory (HBM) Standard." https://www.jedec.org/news/pressreleases/jedec-publishes-hbm3-update-high-bandwidth-memory-hbm-standard, Jan. 2022.

[26] Advanced Micro Devices, Inc. (AMD), "AMD INSTINCT™ MI300X ACCELERATOR," jun 2021.

[27] Gavin Bonshor, "NVIDIA Unveils Updated GH200 'Grace Hopper' Superchip with HBM3e Memory, Shipping in Q2'2024," *AnandTech*, aug 2023.

[28] Lenovo Group Limited, "ThinkSystem NVIDIA HGX B200 180GB 1000W GPU," jul 2025.

[29] JEDEC Solid State Technology Association, "High Bandwidth Memory (HBM4) DRAM." https://www.jedec.org/standards-documents/docs/jesd270-4, Apr. 2025.

[30] S. Misra, D. Kim, J. Kim, W. Shin, and C. Kim, "A voting-based ensemble feature network for semiconductor wafer defect classification," *Scientific Reports*, vol. 12, 09 2022.

[31] D. D. Sharma, "Universal chiplet interconnect express (ucie)®: Building an open chiplet ecosystem," 2022.

[32] David Kehlet, "Accelerating Innovation Through A Standard Chiplet Interface: The Advanced Interface Bus (AIB),"

[33] Nvidia Corporation, "NVIDIA Blackwell Architecture."

[34] Nvidia Corporation, "NVIDIA Grace Hopper Superchip Architecture In-Depth."

[35] Nvidia Corporation, "NVIDIA® NVLink TM High-Speed Interconnect: Application Performance."

[36] Taiwan Semiconductor Manufacturing Company Limited (TSMC or Taiwan Semiconductor), "CoWoS®."