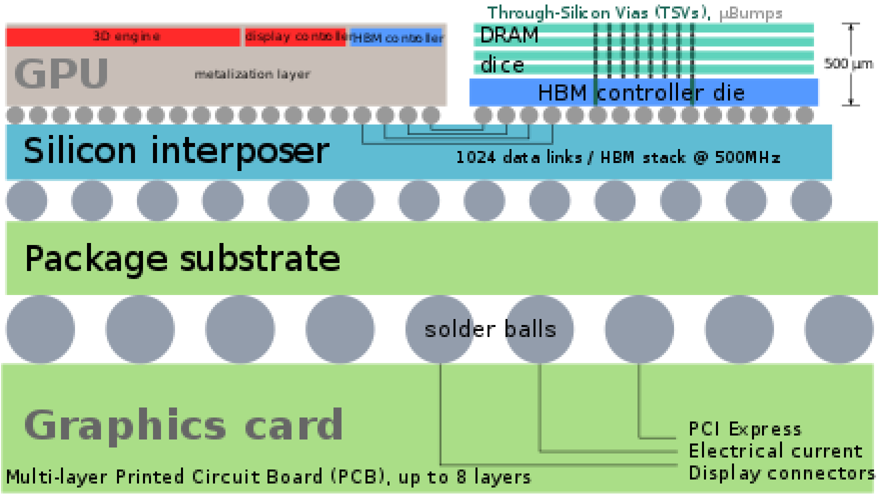
WIKIPEDIA

High Bandwidth Memory

High Bandwidth Memory (HBM) is a high-speed computer memory interface for 3D-stacked SDRAM from Samsung, AMD and SK Hynix. It is used in conjunction with high-performance graphics accelerators, network devices and in some supercomputers. (Such as the NEC SX-Aurora TSUBASA and Fujitsu A64FX)[1] The first HBM memory chip was produced by SK Hynix in 2013,[2] and the first devices to use HBM were the AMD Fiji GPUs in 2015.[3][4]



Cut through a graphics card that uses High Bandwidth Memory. See the through-silicon vias (TSV).

High Bandwidth Memory has been adopted by JEDEC as an industry standard in October 2013.[5] The second generation, HBM2, was accepted by JEDEC in January 2016.[6]

Contents

Technology

HBM achieves higher bandwidth while using less power in a substantially smaller form factor than DDR4 or GDDR5.[7] This is achieved by stacking up to eight DRAM dies (thus being a Three-dimensional integrated circuit), including an optional base die (often a silicon interposer[8][9]) with a memory controller, which are interconnected by through-silicon vias (TSVs) and microbumps. The HBM technology is similar in principle but incompatible with the Hybrid Memory Cube interface developed by Micron Technology.[10]

HBM memory bus is very wide in comparison to other DRAM memories such as DDR4 or GDDR5. An HBM stack of four DRAM dies (4-Hi) has two 128-bit channels per die for a total of 8 channels and a width of 1024 bits in total. A graphics card/GPU with four 4-Hi HBM stacks would therefore have a memory bus with a width of 4096 bits. In comparison, the bus width of GDDR memories is 32 bits, with 16 channels for a graphics card with a 512-bit memory interface.[11] HBM supports up to 4 GB per package.

The larger number of connections to the memory, relative to DDR4 or GDDR5, required a new method of connecting the HBM memory to the GPU (or other processor).[12] AMD and Nvidia have both used purpose-built silicon chips, called *interposers*, to connect the memory and GPU. This interposer has the added advantage of requiring the memory and processor to be physically close, decreasing memory paths. However, as semiconductor device fabrication is significantly more expensive than printed circuit board manufacture, this adds cost to the final product.

Interface

The HBM DRAM is tightly coupled to the host compute die with a distributed interface. The interface is divided into independent channels. The channels are completely independent of one another and are not necessarily synchronous to each other. The HBM DRAM uses a wide-interface architecture to achieve high-speed, low-power operation. The HBM DRAM uses a 500 MHz differential clock CK_t / CK_c (where the suffix "_t" denotes the "true", or "positive", component of the differential pair, and "_c" stands for the "complementary" one). Commands are registered at the rising edge of CK_t, CK_c. Each channel interface maintains a 128-bit data bus operating at double data rate (DDR). HBM supports transfer rates of 1 GT/s per pin (transferring 1 bit), yielding an overall package bandwidth of 128 GB/s.[13]

HBM2

The second generation of High Bandwidth Memory, HBM2, also specifies up to eight dies per stack and doubles pin transfer rates up to 2 GT/s. Retaining 1024-bit wide access, HBM2 is able to reach 256 GB/s memory bandwidth per package. The HBM2 spec allows up to 8 GB per package. HBM2 is predicted to be especially useful for performance-sensitive consumer applications such as virtual reality.[14]

On January 19, 2016, Samsung announced early mass production of HBM2, at up to 8 GB per stack.[15][16] SK Hynix also announced availability of 4 GB stacks in August 2016.[17]

HBM2E

In late 2018, JEDEC announced an update to the HBM2 specification, providing for increased bandwidth and capacities.[18] Up to 307 GB/s per stack (2.5 Tbit/s effective data rate) is now supported in the official specification, though products operating at this speed had already been available. Additionally, the update added support for 12-Hi stacks (12 dies) making capacities of up to 24 GB per stack possible.

On March 20, 2019, Samsung announced their Flashbolt HBM2E, featuring eight dies per stack, a transfer rate of 3.2 GT/s, providing a total of 16 GB and 410 GB/s per stack.[19]

August 12, 2019, SK Hynix announced their HBM2E, featuring eight dies per stack, a transfer rate of 3.6 GT/s, providing a total of 16 GB and 460 GB/s per stack.[20][21] On 2 July 2020, SK Hynix announced that mass production has begun.[22]

HBMnext

In late 2020, Micron unveiled that the HBM2E standard would be updated and alongside that they unveiled the next standard known as HBMnext. Originally proposed as HBM3, this is a big generational leap from HBM2 and the replacement to HBM2E. This new VRAM will come to the market in the Q4 of 2022. This will likely introduce a new architecture as the naming suggests.

While the architecture might be overhauled, leaks point toward the performance to be similar to that of the updated HBM2E standard. This RAM is likely to be used mostly in data center GPUs.

History

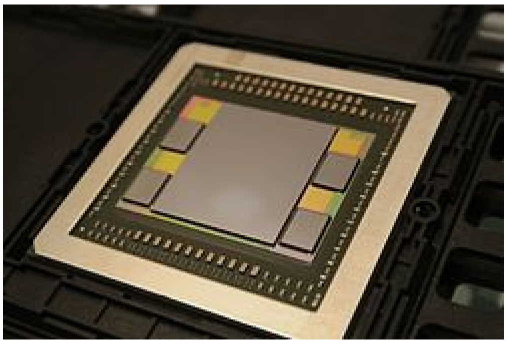
Background

Die-stacked memory was initially commercialized in the flash memory industry. Toshiba introduced a NAND flash memory chip with eight stacked dies in April 2007,[23] followed by Hynix Semiconductor introducing a NAND flash chip with 24 stacked dies in September 2007.[24]

3D-stacked random-access memory (RAM) using through-silicon via (TSV) technology was commercialized by Elpida Memory, which developed the first 8 GB DRAM chip (stacked with four DDR3 SDRAM dies) in September 2009, and released it in June 2011. In 2011, SK Hynix introduced 16 GB DDR3 memory (40 nm class) using TSV technology,[2] Samsung Electronics introduced 3D-stacked 32 GB DDR3 (30 nm class) based on TSV in September, and then Samsung and Micron Technology announced TSV-based Hybrid Memory Cube (HMC) technology in October.[25]

Development

The development of High Bandwidth Memory began at AMD in 2008 to solve the problem of ever-increasing power usage and form factor of computer memory. Over the next several years, AMD developed procedures to solve die-stacking problems with a team led by Senior AMD Fellow Bryan Black.[26] To help AMD realize their vision of HBM, they enlisted partners from the memory industry, particularly Korean company SK Hynix,[26] which had prior experience with 3D-stacked memory,[2][24] as well as partners from the interposer industry (Taiwanese company UMC) and packaging industry (Amkor Technology and ASE).[26]


AMD Fiji, the first GPU to use HBM

The development of HBM was completed in 2013, when SK Hynix built the first HBM memory chip.[2] HBM was adopted as industry standard JESD235 by JEDEC in October 2013, following a proposal by AMD and SK Hynix in 2010.[5] High volume manufacturing began at a Hynix facility in Icheon, South Korea, in 2015.

The first GPU utilizing HBM was the AMD Fiji which was released in June 2015 powering the AMD Radeon R9 Fury X.[3][27][28]

In January 2016, Samsung Electronics began early mass production of HBM2.[15][16] The same month, HBM2 was accepted by JEDEC as standard JESD235a.[6] The first GPU chip utilizing HBM2 is the Nvidia Tesla P100 which was officially announced in April 2016.[29][30]

Future

At Hot Chips in August 2016, both Samsung and Hynix announced the next generation HBM memory technologies.[31][32] Both companies announced high performance products expected to have increased density, increased bandwidth, and lower power consumption. Samsung also announced a lower-cost version of HBM under development targeting mass markets. Removing the buffer die and decreasing the number of TSVs lowers cost, though at the expense of a decreased overall bandwidth (200 GB/s).

See also

- Stacked DRAM
- eDRAM
- Chip stack multi-chip module
- Hybrid Memory Cube: stacked memory standard from Micron Technology (2011)

References

1. ISSCC 2014 Trends (http://isscc.org/doc/2014/2014_Trends.pdf) Archived (https://web.archive.org/web/20150206093927/http://isscc.org/doc/2014/2014_Trends.pdf) 2015-02-06 at the Wayback Machine page 118 "High-Bandwidth DRAM"
2. "History: 2010s" (https://www.skhynix.com/eng/about/history2010.jsp). *SK Hynix*. Retrieved 8 July 2019.
3. Smith, Ryan (2 July 2015). "The AMD Radeon R9 Fury X Review" (http://www.anandtech.com/show/9390/the-amd-radeon-r9-fury-x-review). Anandtech. Retrieved 1 August 2016.
4. Morgan, Timothy Prickett (March 25, 2014). "Future Nvidia 'Pascal' GPUs Pack 3D Memory, Homegrown Interconnect" (http://www.enterprisetech.com/2014/03/25/future-nvidia-pascal-gpus-pack-3d-memory-homegrown-interconnect/). EnterpriseTech. Retrieved 26 August 2014. "Nvidia will be adopting the High Bandwidth Memory (HBM) variant of stacked DRAM that was developed by AMD and Hynix"
5. High Bandwidth Memory (HBM) DRAM (JESD235) (http://www.jedec.org/standards-documents/results/jesd235), JEDEC, October 2013
6. "JESD235a: High Bandwidth Memory 2" (https://www.jedec.org/news/pressreleases/jedec-updates-groundbreaking-high-bandwidth-memory-hbm-standard). 2016-01-12.
7. HBM: Memory Solution for Bandwidth-Hungry Processors (http://www.setphaserstostun.org/hc26/HC26-11-day1-epub/HC26.11-3-Technology-epub/HC26.11.310-HBM-Bandwidth-Kim-Hynix-Hot%20Chips%20HBM%202014%20v7.pdf) Archived (https://web.archive.org/web/20150424141343/http://www.setphaserstostun.org/hc26/HC26-11-day1-epub/HC26.11-3-Technology-epub/HC26.11.310-HBM-Bandwidth-Kim-Hynix-Hot%20Chips%20HBM%202014%20v7.pdf) 2015-04-24 at the Wayback Machine, Joonyoung Kim and Younsu Kim, SK Hynix // Hot Chips 26, August 2014
8. https://semiengineering.com/whats-next-for-high-bandwidth-memory/
9. https://semiengineering.com/knowledge_centers/packaging/advanced-packaging/2-5d-ic/interposers/
10. Where Are DRAM Interfaces Headed? (http://www.eetimes.com/author.asp?section_id=36&doc_id=1321783) Archived (https://web.archive.org/web/20180615032452/http://www.eetimes.com/author.asp?section_id=36&doc_id=1321783) 2018-06-15 at the Wayback Machine // EETimes, 4/18/2014 "*The Hybrid Memory Cube (HMC) and a competing technology called High-Bandwidth Memory (HBM) are aimed at computing and networking applications. These approaches stack multiple DRAM chips atop a logic chip.*"
11. Highlights of the HighBandwidth Memory (HBM) Standard (http://www.cs.utah.edu/events/thememoryforum/mike.pdf). Mike O’Connor, Sr. Research Scientist, NVidia // The Memory Forum – June 14, 2014
12. Smith, Ryan (19 May 2015). "AMD Dives Deep On High Bandwidth Memory – What Will HBM Bring to AMD?" (http://www.anandtech.com/show/9266/amd-hbm-deep-dive). Anandtech. Retrieved 12 May 2017.

13. "High-Bandwidth Memory (HBM)" (https://www.amd.com/Documents/High-Bandwidth-Memory-HBM.pdf) (PDF). AMD. 2015-01-01. Retrieved 2016-08-10.
14. Valich, Theo (2015-11-16). "NVIDIA Unveils Pascal GPU: 16GB of memory, 1TB/s Bandwidth" (http://vrworld.com/2015/11/16/nvidia-unveils-pascal-gpu-16gb-of-memory-1tbs-bandwidth/). *VR World*. Retrieved 2016-01-24.
15. "Samsung Begins Mass Producing World's Fastest DRAM – Based on Newest High Bandwidth Memory (HBM) Interface" (https://news.samsung.com/global/samsung-begins-mass-producing-worlds-fastest-dram-based-on-newest-high-bandwidth-memory-hbm-interface). *news.samsung.com*.
16. "Samsung announces mass production of next-generation HBM2 memory – ExtremeTech" (http://www.extremetech.com/extreme/221473-samsung-announces-mass-production-of-next-generation-hbm2-memory). 19 January 2016.
17. Shilov, Anton (1 August 2016). "SK Hynix Adds HBM2 to Catalog" (http://www.anandtech.com/show/10527/sk-hynix-adds-hbm2-4-gb-memory-q3). Anandtech. Retrieved 1 August 2016.
18. "JEDEC Updates Groundbreaking High Bandwidth Memory (HBM) Standard" (https://www.jedec.org/news/pressreleases/jedec-updates-groundbreaking-high-bandwidth-memory-hbm-standard-0) (Press release). JEDEC. 2018-12-17. Retrieved 2018-12-18.
19. "Samsung Electronics Introduces New High Bandwidth Memory Technology Tailored to Data Centers, Graphic Applications, and AI | Samsung Semiconductor Global Website" (https://www.samsung.com/semiconductor/insights/tech-leadership/samsung-electronics-introduces-new-high-bandwidth-memory-technology-tailored-to-data-centers-graphic-applications-and-ai/). *www.samsung.com*. Retrieved 2019-08-22.
20. "SK Hynix Develops World's Fastest High Bandwidth Memory, HBM2E" (https://www.skhynix.com/eng/pr/pressReleaseView.do?seq=2809&offset=1). *www.skhynix.com*. August 12, 2019. Retrieved 2019-08-22.
21. "SK Hynix Announces its HBM2E Memory Products, 460 GB/S and 16GB per Stack" (https://www.techpowerup.com/258194/sk-hynix-announces-its-hbm2e-memory-products-460-gb-s-and-16gb-per-stack).
22. Ryan Smith (2 July 2020). "SK Hynix: HBM2E Memory Now In Mass Production" (https://www.anandtech.com/show/15892/sk-hynix-hbm2e-memory-now-in-mass-production). *Anandtech.com*. Retrieved 2 July 2020.
23. "TOSHIBA COMMERCIALIZES INDUSTRY'S HIGHEST CAPACITY EMBEDDED NAND FLASH MEMORY FOR MOBILE CONSUMER PRODUCTS" (https://web.archive.org/web/20101123023805/http://www.toshiba.com/taec/news/press_releases/2007/memy_07_470.jsp). *Toshiba*. April 17, 2007. Archived from the original (http://www.toshiba.com/taec/news/press_releases/2007/memy_07_470.jsp) on November 23, 2010. Retrieved 23 November 2010.
24. "Hynix Surprises NAND Chip Industry" (http://www.koreatimes.co.kr/www/news/biz/2007/09/123_9628.html). *Korea Times*. 5 September 2007. Retrieved 8 July 2019.
25. Kada, Morihiro (2015). "Research and Development History of Three-Dimensional Integration Technology" (https://books.google.com/books?id=JaUvCwAAQBAJ&pg=PA15). *Three-Dimensional Integration of Semiconductors: Processing, Materials, and Applications*. Springer. pp. 15–8. ISBN 9783319186757.
26. High-Bandwidth Memory (HBM) from AMD: Making Beautiful Memory (https://www.youtube.com/watch?v=se9TSUfZ6i0s), AMD
27. Smith, Ryan (19 May 2015). "AMD HBM Deep Dive" (http://www.anandtech.com/show/9266/amd-hbm-deep-dive). Anandtech. Retrieved 1 August 2016.
28. [1] (https://www.amd.com/en-us/press-releases/Pages/new-era-pc-gaming-2015jun16.aspx) AMD Ushers in a New Era of PC Gaming including World's First Graphics Family with Revolutionary HBM Technology
29. Smith, Ryan (5 April 2016). "Nvidia announces Tesla P100 Accelerator" (http://www.anandtech.com/show/10222/nvidia-announces-tesla-p100-accelerator-pascal-power-for-hpc). Anandtech. Retrieved 1 August 2016.
30. "NVIDIA Tesla P100: The Most Advanced Data Center GPU Ever Built" (http://www.nvidia.com/object/tesla-p100.html). *www.nvidia.com*.

31. Smith, Ryan (23 August 2016). "Hot Chips 2016: Memory Vendors Discuss Ideas for Future Memory Tech – DDR5, Cheap HBM & More" (http://www.anandtech.com/show/10589/hot-chips-2016-memory-vendors-discuss-ideas-for-future-memory-tech-ddr5-cheap-hbm-more). Anandtech. Retrieved 23 August 2016.
32. Walton, Mark (23 August 2016). "HBM3: Cheaper, up to 64GB on-package, and terabytes-per-second bandwidth" (https://arstechnica.com/gadgets/2016/08/hbm3-details-price-bandwidth/). Ars Technica. Retrieved 23 August 2016.

External links

- High Bandwidth Memory (HBM) DRAM (JESD235) (http://www.jedec.org/standards-documents/results/jesd235), JEDEC, October 2013
- Lee, Dong Uk; Kim, Kyung Whan; Kim, Kwan Weon; Kim, Hongjung; Kim, Ju Young; et al. (9–13 Feb 2014). "A 1.2V 8Gb 8-channel 128GB/s high-bandwidth memory (HBM) stacked DRAM with effective microbump I/O test methods using 29nm process and TSV". *2014 IEEE International Solid-State Circuits Conference – Digest of Technical Papers*. IEEE (published 6 March 2014): 432–433. doi:10.1109/ISSCC.2014.6757501 (https://doi.org/10.1109%2FISSCC.2014.6757501). S2CID 40185587 (https://api.semanticscholar.org/CorpusID:40185587).
- HBM vs HBM2 vs GDDR5 vs GDDR5X Memory Comparison (https://graphicscardhub.com/gddr5-vs-gddr5x-vs-hbm-vs-hbm2/)

Retrieved from "https://en.wikipedia.org/w/index.php?title=High_Bandwidth_Memory&oldid=981217039"

This page was last edited on 1 October 2020, at 00:09 (UTC).