HBM (High Bandwidth Memory) DRAM Technology and Architecture

Hongshin Jun, Jinhee Cho, Kangseol Lee, Ho-Young Son, Kwiwook Kim, Hanho Jin, Keith Kim
SK hynix
Icheon, Gyeonggi-do, Korea

***Abstract*— HBM (High Bandwidth Memory) is an emerging standard DRAM solution that can achieve breakthrough bandwidth of higher than 256GBps while reducing the power consumption as well. It has stacked DRAM architecture with core DRAM dies on top of a base logic die, based on the TSV and die stacking technologies. In this paper, the HBM architecture is introduced and a comparison of its generations is provided. Also, the packaging technology and challenges to address reliability, thermal dissipation capability, maximum allowable package sizes, and high throughput stacking solutions are described. Test technology and testability features are discussed for KGSD and 2.5D SiP.**

***Index Terms*—HBM, Stacked DRAM, TSV, Micro-bump, 2.5D SiP, 3D IC, High bandwidth DRAM, Low Power DRAM, 1500**

I. Introduction

As processor performance increases through the use of additional cores and higher clock frequencies, external DRAM performance becomes a bottleneck for the system performance. It has been a very difficult challenge to develop a memory solution that can remove the gaps between processor memory bandwidth requirements and actual bandwidth performance with only commodity DRAM solutions. These days, not only CPU cores but also other types of computing resources such as GPUs and specialized accelerators are increasingly being used for more parallel computations and better power efficiency. Other system requirements such as lower power, smaller form factor, higher speed, higher density also drive the development of new DRAM solutions.

The industry has requested JEDEC, the leading standardization organization for semiconductor memories, to develop a high-bandwidth memory (HBM) solution [1], utilizing the latest advancements in IC packaging technologies which are Through Silicon Via (TSV) and die stacking. The main goal is to provide enough bandwidth to meet their performance targets for their high-performance computing (HPC), networking, and graphics applications. In response, DRAM vendors and SoC makers have collaborated to produce a standardized HBM in the form of a known-good stacked die (KGSD) which is a 3D IC using TSV technology.

The HBM is a very attractive solution for high performance system designers because it can provide a scalability of memory capacity, smaller footprints, lower power consumption. In the following sections, HBM architectures is introduced and its generations are compared. Its packaging technology and challenges are discussed. Furthermore, test technology challenges and testability features for KGSD and 2.5D SiP are elaborated.

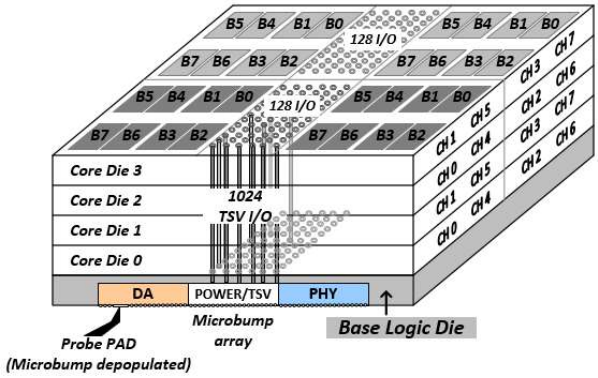
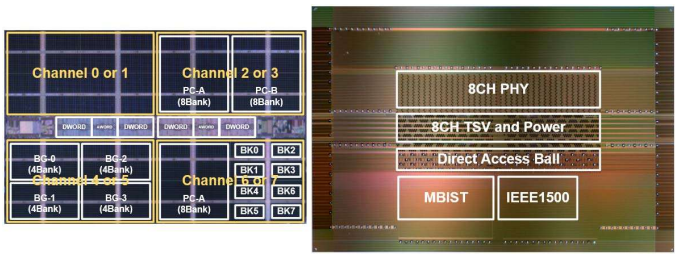


Fig. 1. HBM Stacked-DRAM Architecture

II. High-Bandwidth Memory (HBM) Architecture

The fundamental structure of HBM is composed of a base logic die at the bottom and stacked core DRAM dies, which are interconnected by TSVs as shown in Fig. 1 [2][3]. The power and ground have common planes to support all of the eight channels. In the heterogeneous HBM structure, the core dies have a conventional DRAM architecture with TSV interfaces. The base die has I/O buffers and inevitable test logic. By using stacked-DRAM, TSV, micro-bump, and 2.5D package technologies, HBM offers improved capacity, bandwidth, and power efficiency compared with conventional DRAMs.



(a) (b)

Fig. 2. Die photos of HBM2: (a) core, (b) base

978-1-5090-3274-7/17/$31.00 ©2017 IEEE

A. Core DRAM Die

A core DRAM die has 2n-prefetch with minimum access granularity of 16 bytes per channel. The core architecture of HBM2 is similar to the conventional DRAM which has memory cell arrays and peripheral logic as shown in Fig. 2 (a) [4]. HBM2 has multiple 8 Gb core DRAM dies and 8 channels of 128 I/Os. Depending on the density of HBM2 either 2 GB or 4/8 GB, each slice of core DRAM die needs to have 2 or 4 channels and 8 or 16 independent banks, respectively. Each channel of a core DRAM die has an independent address and data TSV's with point-to-point (P2P) connections from the base die to support independent channel operations. The peripheral area is dedicated to AWORD, DWROD, and TSV. AWORD is assigned for column commands, row commands, and address control. HBM uses semi-independent row- and column-command interfaces, which allow RAS and CAS commands in parallel. DWORDs are assigned for data transport. The array of TSVs for power, ground and signals are assigned in the peripheral area as well.

B. Base Logic Die

The base logic die of HBM is composed of PHY, TSV, DFT logic, and direct-access (DA) in Fig. 2 (b). The PHY block is for the main interface between HBM DRAM and the memory controller in the host ASIC. It has a total of 8 channels where a channel consists of an AWORD (address/command buffers) and four channel-interleaved DWORDs (data buffers). A total 8 of AWORDs and 32 DWORDs are located in the PHY area. The center area is reserved for TSV's that deliver signals, power, and ground to stacked core dies. The area between PHY and TSV is filled with signal lines for 1024 bit data and decoupling capacitance. The lower area is for test logic and DA ports.

Some micro-bumps are depopulated for probe test pads. They are used for DA pads, which are the main interface between the automatic test equipment (ATE) and stacked-DRAM. Outside of the main HBM ballout, more test power probe pads and decoupling capacitors are located at the chip boundary. MBIST (Memory BIST) and IEEE1500 blocks are located at the bottom, which are implemented with RTL design methodology.

The PHY on the base logic die communicates with the PHY on the host SoC through the interconnections on a silicon interposer. The interposer connections have large resistance and capacitance, causing high power consumption and ISI (Inter-Symbol Interference). The PHY allows a significant reduction of the interconnection length and CIO so that the power and speed benefits in the 2.5D SiP can be maximized.

C. Comparison of HBM Generations

The HBM is a standard firstly defined by JEDEC in October 2013. Since then, HBM2 has been developed whereas HBM3 is still under discussion. Table 1 shows the key feature comparisons among generations. The HBM1 offers around 128 GB/s bandwidth, which stacks four 2 Gb core dies with 2 channels operating at 1 Gbps data rate per pin. Supply voltages of HBM1, VDDC, VDDQ, and VPP are 1.2V, 1.2V, and 2.5V, respectively. HBM1 supports only legacy mode operations. In the HBM2, the bandwidth has been improved to 256 GBps or higher. Its configurations support 2, 4, or 8 core dies in a stack. Also new features have been introduced in HBM2, such as pseudo channel and implicit precharge operations, as well as ECC storage. Pseudo channel mode is a very important improvement in HBM2. In the pseudo channel mode, each 128-bit channel can operate as two separate pseudo channels of 64 bits. A pseudo channel consists of 4 bank groups where each group has 4 banks. Two pseudo channels in a channel share a common AWORD, but have separate 16 banks and execute commands individually. By using the pseudo channel mode, HBM2 can achieve optimized command bandwidth, decreased latency, and higher effective data bandwidth. The HBM3 specification is still under discussion in JEDEC, the goal being a dramatic improvement in memory density, bandwidth, and power efficiency. There are requests to double the density of core dies from 8 Gb to 16 Gb, and to feature 4Hi, 8Hi, or higher stacks, and to support 0.4V VDDQL to reduce IO power dramatically, and to increase peak bandwidth 2x compared to HBM2. The base die for HBM1 and HBM2 has been successfully developed and fabricated in DRAM process technologies. However for HBM3, to meet the higher speed and lower power requirements, the base die may need to adopt some logic process technologies such as high-k metal gates and low-k IMD (Inter-Metal Dielectric) schemes on top of the pure DRAM technology.

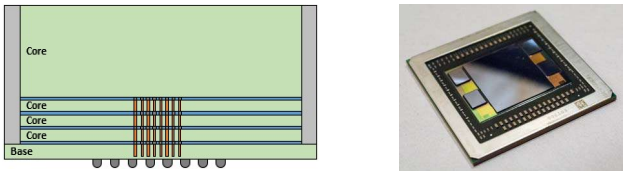
TABLE I. KEY FEATURE COMPARISONS OF HBM GENERATIONS

Item	HBM1	HBM2	HBM3(Expect)
VDDC/VDDQ/VPP	1.2V/1.2V/2.5V	1.2V/1.2V/2.5V	less
Density / slice	2Gb	8Gb ~	16Gb ~
Data Rate / Pin	1Gbps	2Gbps ~	3.2Gbps ~
# of Stack/Chip	4	2/4/8	4/8/12/16
BandWidth / Chip	128GB/s	256GB/s ~	410GB/s ~
Operation Mode	Legacy	Legacy Pseudo Channel	Pseudo Channel
# of IO and CH	128IO/CH, 8CH	64IO/P-CH, 16P-CH	Similar or more
Banks	8Banks/CH	4Banks/BG, 4BGs/P-CH	Similar or more
IO Interface	CMOS	CMOS	Low voltage swing interface
New Functions	-	ECC storage Pseudo Channel Implicit precharge	Under discussion (RAS)

III. HBM PACKAGE STRUCTURE AND CHALLENGES

An HBM package consists of one base die at the bottom and multiple core DRAM dies on the top. The dies are connected with several thousands of TSVs and micro-bumps. The size of the core dies are a slightly smaller than the base die, and the surroundings are encapsulated with an epoxy compound material for the side mold. To better dissipate the heat from the base die, HBM cubes have no over molds and the

top die is exposed with silicon. Fig. 3 shows the schematic diagram of HBM KGSD and 2.5D SiP with GPU and 4 HBM cubes on a large size interposer die from AMD [5].



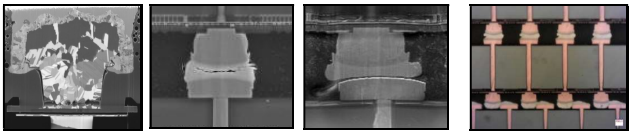
(a) (b)

Fig. 3. (a) HBM KGSD Schematic Diagram (b) 2.5D SiP with four HBM cubes (Courtesy:AMD)

From a packaging point of view, there are four major challenges not only for the robust premium solution in high-end applications but also for the cost-effective and widely extensive ones in mid-end applications: package reliability, thermal dissipation capability, maximum allowable package size versus memory die shrinkage roadmap, and high throughput chip stacking to reduce manufacturing cost.

A. Package reliability

Package reliability is mainly dependant upon the micro-bump joints at chip-to-chip interconnections in 3D-IC packages using TSVs. Thermal-compression bonding (TCB) is a common method to stack up multiple chips with TSVs and micro-bumps, but its rapid bonding process may cause poor quality joints. Insufficient bonding time can lead to abnormal bump joints and various failure modes such as non-wet, brittle intermetallic compound (IMC) formation, bump cracking, head-in-pillar (HiP) joints, and so on [6-11]. These joints could potentially cause critical impacts not only on assembly yield but also on long-term reliability performance such as during thermal cycling or during high temperature storage tests. Fig. 4 shows the major failure modes in micro-bump joints of 3D-IC packages.



(a) (b) (c) (d)

Fig. 4. Example of various micro-bump joint failures (a) bump pop-up (b) bump crack (c) underfill trap, (d) misalignment

B. Thermal dissipation capability

Due to the heat delivered from an SoC and a base die of an HBM cube, upper core dies can be overheated during operations and DRAM cells can be thermally damaged. Therefore, the thermal dissipation capability of an HBM cube is very important, and proper thermal design techniques should be considered such as thermal conductive underfill material and thermal dummy bumps to reduce the junction temperature effectively. In addition, thickness tolerance between an SoC and HBM cubes is another concern. Considering the thickness differences, TIM and heat sink design may be necessary.

C. Maximum allowable package size vs. memory die shrinkage

HBM package size cannot be standarized in JEDEC, due to the fact that DRAM vendors are unable to agree on a uniform HBM size because it is directly related to manufacturing cost. Moreover, with each generaion they will need to convert the previous implementation into next generation with smaller DRAM technology nodes and then the core die size will be decreased. The increase in the side mold width due to core die size shrinkage can result in warpage behavior changes at wafer level or at die level.

D. High throughput chip stacking

Thermal compression bonding using pre-applied underfill is a commonly used method in thin die stacking. However, it is done by individual die bonding and thus long stacking times lead to low productivity. Many research engineers and developers are evaluating new stacking methods to minimize stacking time such as gang bonding in a vertical or horizontal plane, and mass reflow bonding with molded underfill or with capillary underfill. Total stacking die counts and chip size could be critical factors in the next generation stacking technology since it is obviously a key process in determining overall package quaility, reliability, assembly yield, manufacturing cost, and so on. As described in Table 1, HBM3 requires 12-16 die stacks. Since the maximum package height cannot exceed 775 um of Si die thickness, the device quality and package level reliability are very important in terms of thinner die handling.

Another challenge facing HBM products is combining 2.5D SiP with a large size interposer and SoC dies. There are various kinds of 2.5D SiP structure platforms such as chip-first COWOS (Chip On Wafer On Substrate) [12][13], chip-last COWOS, and COCOS (Chip On Chip On Substrate) based on Si interposers and EMIB (Embedded Multi-die Interconnect Bridge) [14] using embedded Si bridge chips and organic substrates. Unique and various 2.5D SiP structures may lead to many different requirements for HBM. Currently HBM industry does not yet have consolidated HBM mechanical specifications or requirements and this poses an essential challenge for the future.

IV. HBM AND 2.5D TESTING

The molded KGSD structure of HBM introduces many test challenges to fulfill the DRAM quality requirements for applications. At the wafer level, core DRAM wafers are tested simply following the conventional DRAM test flow: WFBI, hot/cold test, and repair. The base wafer test covers IEEE 1500 testing, scan testing, and high-speed PHY testing. After stacking, the KGSD wafer is not in a packaged form. This leads to many test challenges, including TSV testing, dynamic burn-in stress testing, warpage handling, and speed testing through DA pads. DFT solutions should be introduced to apply dynamic stress more efficiently to DRAM cells at the KGSD wafer level, which would replace package level burn-in stress. TSV open/short test and repair schemes are also very important to secure higher stacking yields. Debug capabilities which can

pin point failing TSV locations are also critical for failure analysis and further improvements [15].

One of the unique test challenges with HBM is that there is no reliable test solution to touch the thousands of micro-bumps. [16] All the tests should be done through a limited number of DA ports. For an efficient PHY I/O testing through DA ports, EXTEST_TX/RX can be used for DC I/O testing by utilizing the capacitance of the micro-bumps. Also at-speed internal loopback through TX and RX is applied. After the singulation of KGSD, optical inspection steps screen structural failures of micro-bumps such as: absence, dislocation, or tilting.

HBM is delivered in KGSD form and then integrated into a 2.5D SiP structure. HBM specifications define DFT features that can be used to test and repair interposer interconnections and HBM cells after 2.5D SiP assembly. IEEE 1500-based instructions such as EXTEST_TX/_RX, DWORD/AWORD_MISR, MISR_MASK, MBIST, SOFT/HARD_LANE_REPAIR, and SOFT/HARD_REPAIR are available already. The test solutions have been implemented and proven in successful HBM mass productions.

V. Conclusion

HBM is a breakthrough 3D stacked DRAM solution for high bandwidth, low power, and high capacity applications in a small form factor, like graphics, HPC, and networking. TSV and die stacking technologies enable an HBM architecture which has multiple core DRAM dies stacked on a base logic die, supporting 256 GBps bandwidth with thousands of DQ I/Os running at 2 Gbps. Some challenges and solutions for package and testing have been discussed. As 2.5D SiP eco-systems mature among HBM vendors, SoC foundries, OSAT, IP providers, and EDA vendors, the HBM solution will penetrate into major DRAM sectors, starting from high end graphics and HPC applications and expanding to main computing and console applications.

References

[1] JEDEC Standard High Bandwidth Memory (HBM) DRAM Specification, JESD235A, 2015.

[2] D. Lee, et al., "A 1.2V 8Gb 8-channel 128GB/s High-Bandwidth Memory (HBM) Stacked DRAM with effective Micro-Bump I/O Test Methods using 29nm process and TSV," ISSCC Dig. Tech. Papers, pp. 432-433, Feb. 2014.

[3] D. Lee, et al., "An exact measurement and repair circuit of TSV connections for 128GB/s high-bandwidth memory (HBM) stacked DRAM," IEEE Symposium on VLSI Circuits, pp. 1-2, June 2015.

[4] J. Chern, et al., "A 1.2V 64Gb 8-Channel 256GB/s HBM DRAM with Peripheral-Base-Die Architecture and Small-Swing Technique on Heavy Load Interface," ISSCC Dig. Tech. Paper, pp. 318-319, Feb. 2016.

[5] M. Alfano, et al., "Unleashing Fury: A New Paradigm for 3-D Design and Test," IEEE Design & Test, Volume 34, Issue 1, Feb. 2017.

[6] H.Y. Son, et al., "Mechanical and Thermal Characterization of TSV Multi-Chip Stacked Packages for Reliable 3D IC Applications", *Proceedings in ECTC* (2016)

[7] K. Sakuma, et al., "An Enhanced Thermo-compression Bonding Process to Address Warpage in 3D Integration of Large Die on Organic Substrates", *Proceedings in ECTC* (2015)

[8] Y.C. Liang, et al., "Side Wall Wetting Induced Void Formation due to Small Solder Volume in Microbumps of Ni/SnAg/Ni upon Reflow" *Proceedings in ECS Solid State Lett* (2012)

[9] H.Y. Son, et al., "Reliability Studies on Micro-Bumps for 3-D TSV Integration", *Proceedings in ECTC* (2013)

[10] H. Liu, et al., "Effect of IMC Growth on thermal cycling reliability of micro solder bumps" *Proceedings in EPTC* (2013)

[11] W.S. Kwon, et al., "Enabling a Manufacturable 3D Technologies and Ecosystem using 28nm FPGA with Stack Silicon Interconnect Technology", *Proceedings in IMAPS* (2013)

[12] Mike Ma, et al., "The Development and Technological Comparison of Various Die Stacking and Integration Options with TSV Si Interposer", ECTC (Electronic Components and Technology Conference) 2016

[13] Suresh Ramalingam, "Stacked Silicon Interconnect Technology (SSIT) Qualification – Requirements and Tools, 3D Stress Workshop at SEMATECH, 2011

[14] Ravi Mahaja, et al., "Embedded Multi-die Interconnect Bridge (EMIB) - A high density, high bandwidth packaging interconnect", ECTC (Electronic Components and Technology Conference) 2016

[15] H. Jun et al., "High-Bandwidth Memory (HBM) Test Challenges and Solutions", IEEE Design & Test, Volume 34, Issue 1, Feb. 2017.

[16] E.J. Marinissen et al., "Direct Probing on Large-Array Fine-Pitch Micro-Bumps of a Wide-I/O Logic-Memory Interface," Proc. International Test Conference, pp. 1-10, 2014, DOI 10.1109/TEST.2014.7035314