Design Considerations of HBM Stacked DRAM and the Memory Architecture Extension

Dong Uk Lee, Kang Seol Lee, Yongwoo Lee, Kyung Whan Kim, Jong Ho Kang, Jaejin Lee, Jun Hyun Chun

SK Hynix, Icheon, Korea

***Abstract* — Recently, the 3D stacked memory, which is known as HBM (high bandwidth memory), using TSV process has been developed. The stacked memory structure provides increased bandwidth, low power consumption, as well as small form factor. There are many design challenges, such as multi-channel operation, microbump test and TSV connection scan. Various design methodology make it possible to overcome the difficulties in the development of TSV technology. Vertical stacking enables more diverse memory architecture than the flat architecture. The next generation of HBM focuses on not only the bandwidth but also the system performance enhancement by adopting pseudo channel and 8-Hi stacking. The architecture applied to the second generation HBM are introduced in this paper.**

***Index Terms* — High-Bandwidth, Stacked Memory, TSV, Microbump, Multi-Channel.**

I. Introduction

DRAM scaling is becoming more difficult to achieve, when it comes to the density and the bandwidth [1][2]. Although, many applications, such as HPC(high performance computing), TbE(terabit Ethernet), graphics memory and high-end client applications, need more than 128GB/s ~1TB/s bandwidth. It is difficult to get high bandwidth using conventional package, such as FBGA. That needs many additional circuitries due to the high speed per pin. Recently, the memory I/O power portion becomes more dominant in the system power [3]. High speed operation and large form factor with many memory chips per system make a lot of power consumption, which degrade the system performance.

New solution has been developed using wide I/O with lower speed [4]. Memory can be directly integrated on the interposer or host chip, and communicate with the memory controller through fine-pitch wide I/O[5]-[7], which gives compact solution for the system. The 2.5D connection using the interposer with wide I/O slightly increases the trace length between chips, but it is good solution for heat dissipation. Another advantage of 2.5D solution is that the host chip does not need TSV process since the interposer has its own TSVs.

Fig. 1 shows the KGS(known good stack) of HBM DRAM. Heterogeneous stacking process using the CoW (chip-on-wafer) technology has various advantages such as the cost efficiency and the parallel testing capability over chip-on-chip stacking process of memory interface.

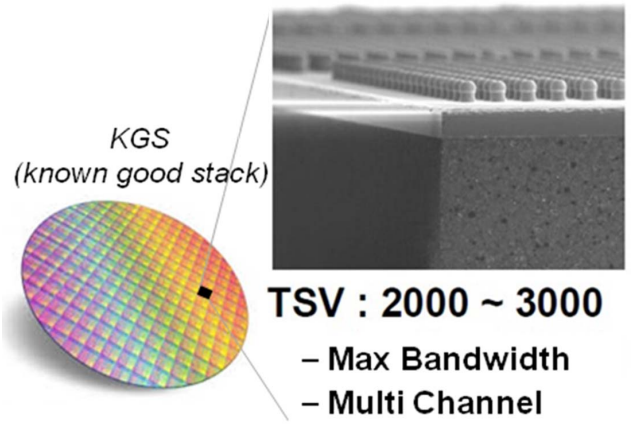


Fig. 1. Known good stack of HBM DRAM using CoW(chip-on-wafer) process.

However, microbump and TSV has inherent problem, which is the lack of testability. Determining known good stack is the dominant factor of mass production in respective of the cost and the reliability.

Achieving the core memory bandwidth is another problem. To increase core memory bandwidth, the chip needs enough sense-amplifiers and many internal I/O, which cause area penalty. The reduction of the number of banks and increased page size per bank can compensate the area penalty. However, these approaches are not a good solution for the system performance, because the IR drop caused by the increased page size makes it difficult to provide enough on-chip supply voltage. Chip stacking technology with vertical memory architecture can mitigate these problems. Multi-stacked DRAM using point-to-point TSV connection multiply the memory bandwidth by the number of core DRAM slices. In section IV, the features of next generation HBM are introduced. The vertical connection enables various memory architectures. It is possible for the memory system to increase the number of banks and sub-channels by using 3D memory structure.

II. HBM Stacked Memory

The structure of HBM, that consists of 4-Hi(high) core DRAM and base logic die at the bottom [8], is described in Fig. 2. Chip-on-wafer HBM has three steps process

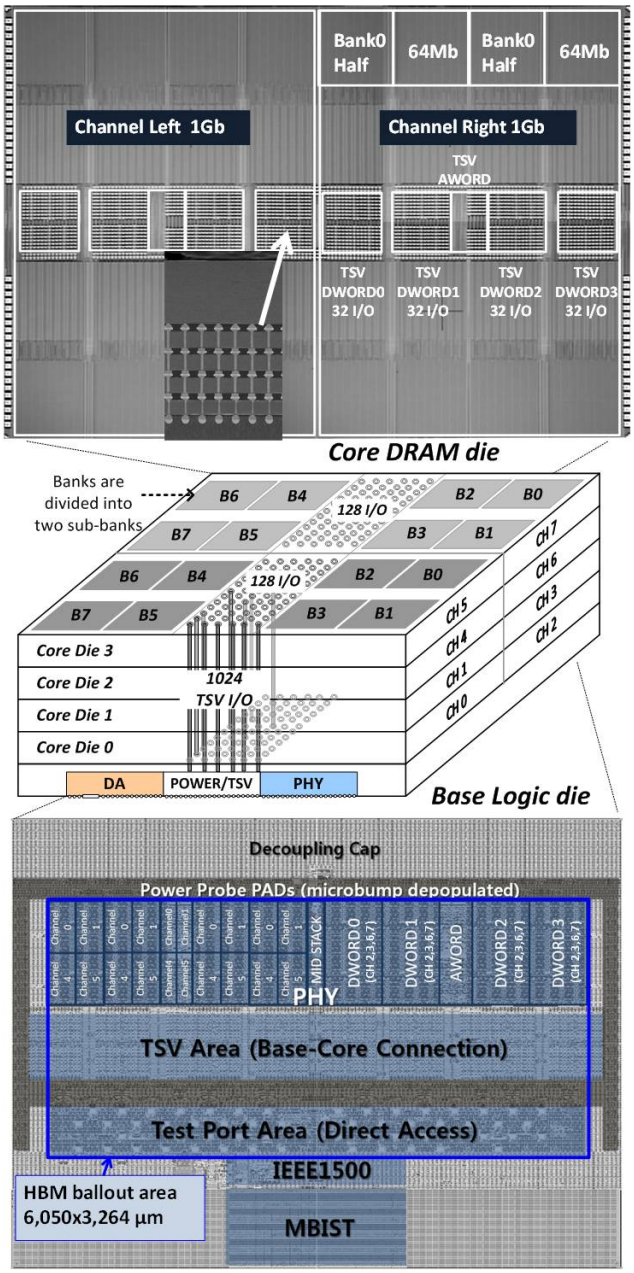


Fig. 2. 8Gb(1Gb/channel) HBM Stacked DRAM architecture and micrograph with TSV structure.

before sawing for assembly, which is stacking dies on the wafer, flipping, and testing.

A. Core DRAM Die

The core DRAM consists of 2 channels, where each channel has 1Gb density with 128 I/O and 8 banks. Each core DRAM die has the isolated address and data TSV per channel with P2P (point-to-point) connection. Core DRAM has 2n-prefetch with minimum access granularity of 32 bytes [9]. Architecture is similar to graphics DRAM [10], in addition, it has 2 channel per slice. For the addresses and the global I/O to the base die connection, there is TSV area in the center peripheral area of channel. A bank is divided into 2 sub-banks due to the large I/O numbers. The sub-bank consists of 64M DRAM cells which includes 8K word lines and 8K bit lines. Total 256 global I/O per channel are connected to the TSV. The TSV area for power supply is also dedicated for this area.

HBM uses semi-independent row- and column-command interfaces, which allow RAS and CAS commands in parallel. Similar to LPDDR3 [11], DDR interface is used as well, for reducing pin count. Wide I/O memory has large I/O current consumption as well as simultaneous switching noise. DBI AC function in HBM is important for reducing these effects [12]. The single bank refresh function and tRAS(bank active-to-precharge) counter function is added for increasing effective system throughput. It is still necessary for increased stack yield that the core die have wafer probe PAD at the edge of chip. But the wafer probe PAD area and test logic should be minimized because it has no functionality after stacking.

B. Base Logic Die

The bottom of Fig. 2 is the layout which describes the base logic die architecture. The size of the ballout in the center is about 6 mm by 3 mm. PHY (Physical layer) is the main interface between DRAM and memory controller. PHY has total of 8 channels where a channel consists of an AWORD and four channel-interleaved DWORD. The center area is allocated for TSV's that deliver signals and power to the core die. Because there are fine-pitch 1024 PHY and TSV I/O, it is difficult to test with conventional method, such as direct I/O probing. A direct-access (DA) port at the bottom is for the stack test using the depopulated microbump. MBIST(Memory BIST) and IEEE1500, RTL based circuit [13] for the purpose of on-chip and off-chip test are located at the bottom. For the RAS(reliability, availability, serviceability) feature, cell repair after assembly, microbump lane repair and burn-in function in CoW level are also included[14].

Current based CMOS output driver is used for the off-chip driving. In case of homogeneous stack without base die, the C_{IO} would be greatly increased because the TSV I/O is the final I/O of the chip and the load capacitance is multiplied by the number of core slices. The interposer has also large R and C, causing high power consumption and ISI. Therefore, the usage of base die reduces C_{IO} significantly (for example, from 1.6pF to 0.4pF) and the reduced distance of PHY between DRAM and memory controller decreases the total interface power [15][16].

III. Design Considerations

In designing the stacked memory with microbump, there are three major considerations. Those are the indirect microbump test using IEEE1500 or DA port, TSV connection measurement and the power distribution of the stacked memory for the simultaneous channel operations.

A. High Speed Microbump I/O Test

One of the limitations in HBM is that I/O buffer cannot be directly tested because the pitch of the microbump is 45x55μm. Besides the functionality as a buffer die, the usage of the base logic die is to support various functions for indirect microbump test. External loopback function between memory and controller is designed for I/O link test and training after assembly [17]. In the system-in-package level, as depicted in Fig. 3, DRAM and memory controller has its own loopback register. The loopback function and register is designed for the checking I/O after assembly. IEEE1500 based on serial port test is also useful for the devices with many I/O. SiP level DRAM cell repair using serial port reduces the chip fail rate in system level. From the cost point of view, the significant problem in CoW test procedure for KGS(known good stack) is how to guarantee it before assembly. Pad-loopback methods for HBM are proposed to guarantee the memory function after assembly. Pad-loopback is the method of parallel write and read operation via microbump [18] using DA port. The signals are driven into the PAD, using the test TX for address and output TX for the write. In such a way, the microbump I/O can be in operation at the similar condition to that of after-assembly. In Fig. 3, the read data as well as the write data can be captured by DA_DQS using appropriate timing control. In the write operation, DA_write data are driven into MUX and TX and finally microbump. DA_DQS can also be driven through test TX. The write DQS in HBM is unidirectional, and write DQS need test TX. In the read operation, the data latched by delayed DA_DQS can be safely stored or compressed in loopback register. In case the phase of DA_DQS is located between the RL+tDQSCK and RL+tDQSCK+UI, the pass/fail zone of the data can be measured by adjusting DA_DQS delay. The stacked wafer test shmoo is shown in Fig. 4.

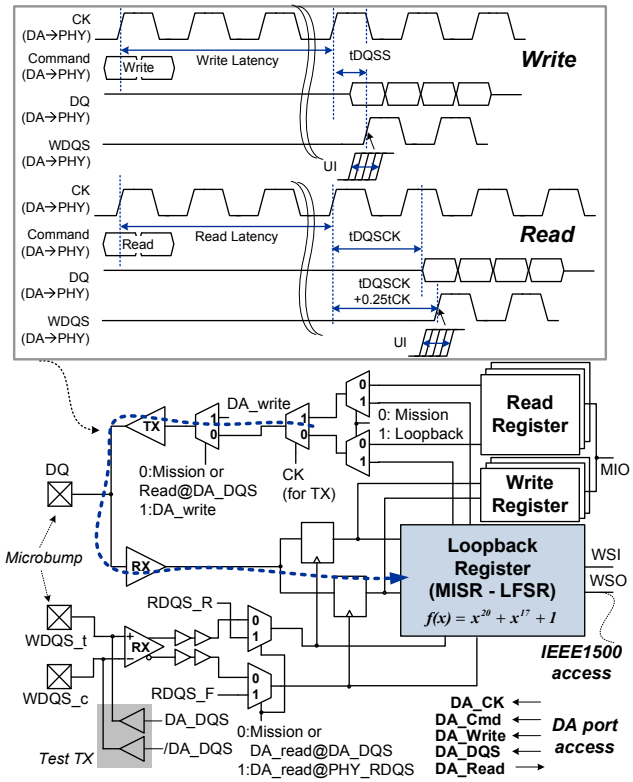


Fig. 3. Microbump test circuit before assembly using pad-loopback and MISR register.

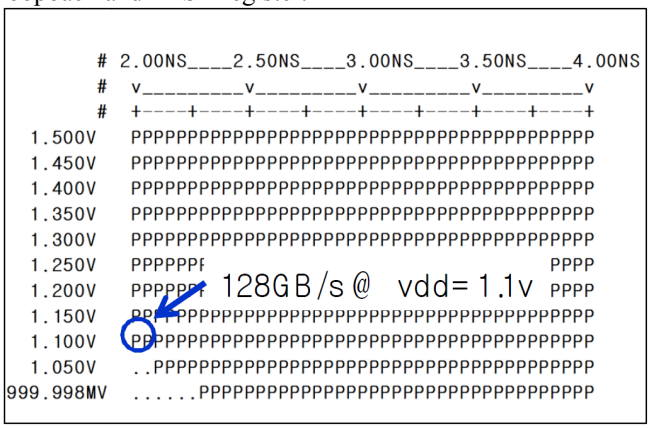
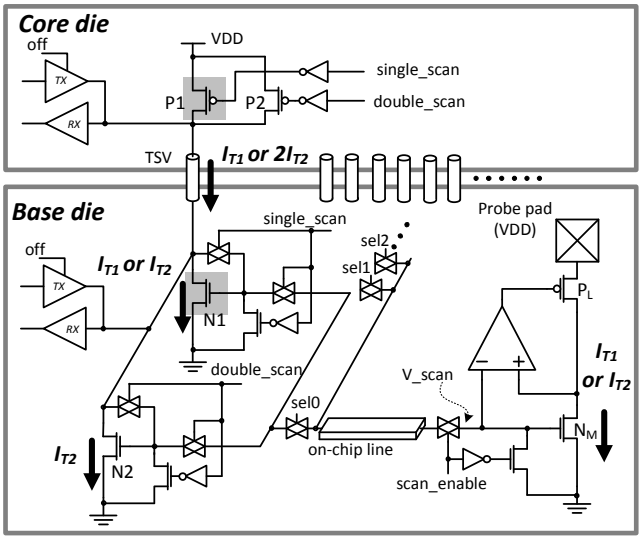


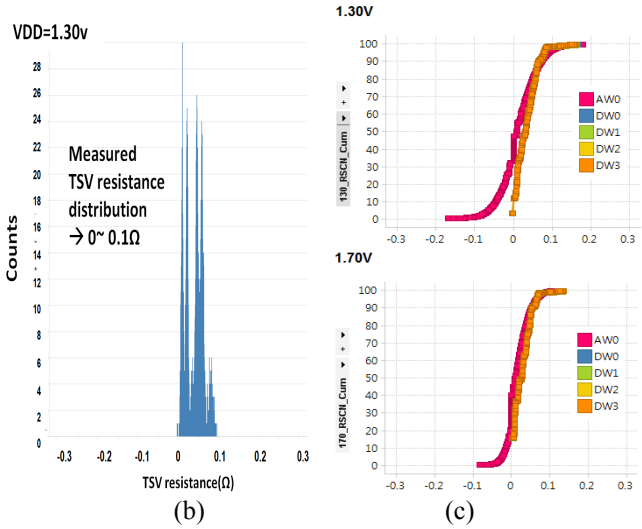
Fig. 4. Stacked wafer test results(gapless read operation)

B. TSV Current Scan and Repair

Conventional TSV resistance measurement method has the limited resolution, in other words, it has the resolution of several hundreds of ohms. Because the average resistance of TSV connections without defect has the milliohm scale resolution [19], it is difficult to distinguish flawless or not using conventional method. Either small crack in TSV or mis-alignment of TSV makes changes of resistance in a small portion, which causes the potential reliability failure [20], but that cannot be detected by conventional method. The double sampling can solve this problem [21]. In Fig. 5 (a), when single_scan signal is enabled, I_{T1} current flows through core die P1, TSV and base die N1. That current is copied using current mirror N_M which minimizes the effect of on-chip line resistance. Although, it is difficult to measure the exact TSV resistance because tester offsets, transistor resistances and their PVT offset are included in this current elements. When both single_scan and double_scan are enabled, P1,



(a)



(b) (c)

Fig. 5. (a)TSV scan method using the double sampling and (b) TSV resistance measurement distribution (c) cumulative distribution.

$$R_{TSV} = \frac{I_{T1} - I_{T2}}{I_{T1} I_{T2}} V_{DD} \qquad \textbf{(1)}$$

P2, N1 and N2 are turned on. The current flow I_{T2} would be precisely the double suppose that the TSV has nearly zero impedance. In this case, the current at the probe pad would be the same, because V_scan level maintain the same irrespective of double_scan enabling. If the TSV does not have near-zero impedance but has some defects, the current at probe pad will have different values according to the double_scan enable. The TSV resistance value can be calculated by extracting the single_scan current I_{T1} and double_scan current I_{T2} as the formula (1). There are more measurement offsets in AWORD than in DWORD, because the PMOS in AWORD is only for the scan and is small size. The Fig. 5(c) shows the cumulative distributions of TSV resistance. The diversity of AWORD TSV distributions compared with DWORD is shown, because the transistor effect is more dominant. Transistor offsets can be reduced by increasing supply voltage level (from 1.3v to 1.7v). TSV repair is performed by the signal multiplexers and chain type shift registers. The shift register based logic is used for the soft TSV repair, which informs the cause of the failure and confirms the chip operation before e-fuse rupture.

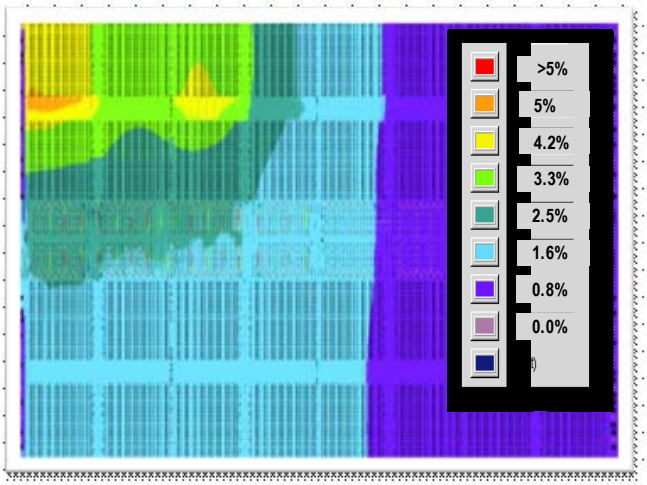


Fig. 6. Power distribution analysis (4 stack simultaneous write operation, VDD IR drop)

C. Power Distribution Analysis

For reducing chip failure rate by IR drop without large area penalty, it is necessary that the power TSV are allocated considering total current profile. On-chip IR drop is significant because of the large power consumption. Stacked memory has the advantage in that TSV (extremely low resistance is measured) can be allocated in the area of peak IR drop. In 4-Hi stack case, the IDD4W is one of the worst conditions in respective of core die IR drop. The simulation results of 1Gb/ch HBM is shown. In the IDD4W case, there are local write drivers of each bank that provide current into BLSA. In the IDD4W operation, 256 local write drivers are driving simultaneously per channel. Based on PDN(power distribution network) model and the IDD4W value of power supply of 1.2V, the max VDD drop(AC+DC) is about 5%, VSS bounce is 7%, as depicted in Fig. 6. Adding decoupling capacitance or power TSV near the Y decoder will reduce IR drop, but it has area penalty. Peak IR drop in a specific bank is caused by gapless data write in a minimum tCCD. Bank group operation can mitigate the peak IR drop, because the bank access is restricted and the current consumption is distributed to wider area.

IV. Memory Architecture for the System Performance

HBM has the great capability in the bandwidth, the low power consumption (energy efficiency), and the various memory architectures. In the 2nd generation HBM, the bandwidth requirement is doubled, that is, 256GB/s. The bandwidth limitation is caused by tCCD(column-to-column access) in most of the DRAM. There are two approaches to overcome the tCCD limitation (about minimum 2ns in the commodity DRAM). The first approach is increasing prefetch, that is, internal parallel I/O operation. The architecture with the increased prefetch needs the multiplied number of BLSA and local and global I/O lines. Therefore, increasing prefetch to get the higher bandwidth makes the page size of the memory larger, as well as more power consumption of the row activation.

Bank group [22], that is adopted to increase bandwidth in the 2nd generation HBM, is another method to mitigate the tCCD limitation without increasing prefetch. The tCCD limitation is overcome by interleaving the bank access in system level. The data of a different bank group can be accessed to the timing tCCDS(about 1ns). But the bank access to the same bank group is still restricted to tCCDL requirement. The number of banks and the number of bank groups should be larger to prevent the effective bandwidth degradation.

After assembly, HBM is located close to the host CPU or GPU in the SiP level, as depicted in Fig. 7. Most of the memory application of HBM is the near memory that is used as the high bandwidth application, rather than the high density application such as the DRAM module. In the near memory applications, random row access for the close page access is more important than the open page access. However, most of the DRAM has the restriction of core parameter, such as tRRD(row-to-row access) and tFAW(four active window), which is closely related to the DRAM power budget (such as IR drop in section III), the page size and the memory architecture. Several DRAM architectures have been developed to overcome the scaling limit and to get the effective data transfer [23]-[27].

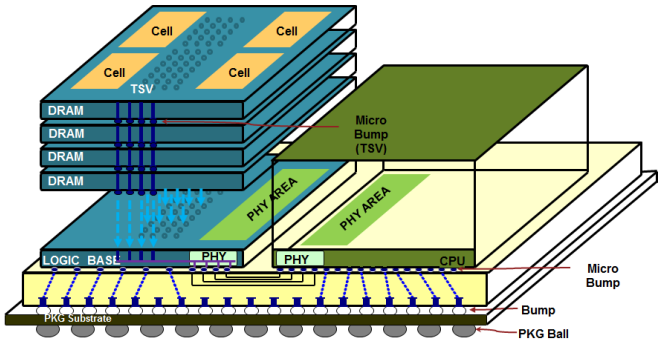


Fig. 7. The diagram of SiP using HBM stacked memory on the interposer(2.5D) .

HBM is interfaced with SoC frequently compared with DDR4. The number of access of CPU is more often than conventional DRAM. As CPU clock frequency speed is getting higher, typical row access latency (tRRD /tFAW /tRC) should be reduced. But if we reduce the row active-to-active timing, the row activation current is increased. The increased current consumption in proportion to the reduced latency degrades energy efficiency. Therefore, it is necessary to find alternative solution to reduce the activation current without die area overhead.

On the other hand, in perspective of the DRAM

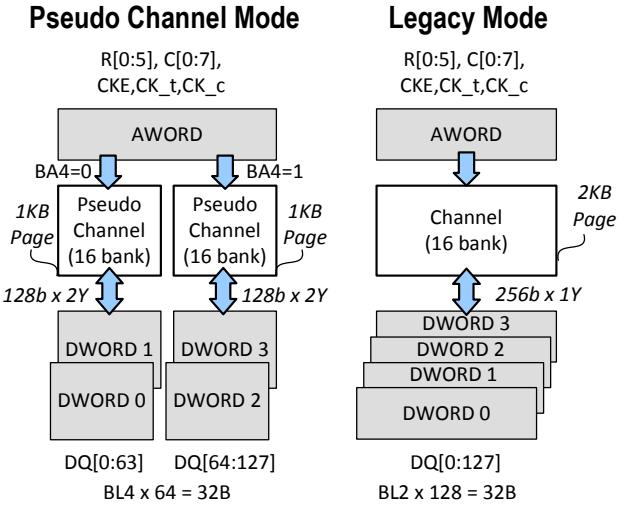


Fig. 8. The concept of pseudo channel mode (128b x 2 subsequent Y addresses) and the comparison with legacy mode(256b x 1 Y address).

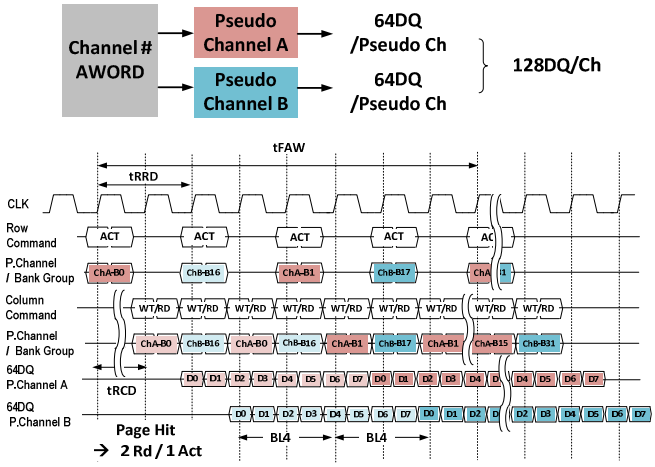


Fig. 9. Timing diagram of the pseudo channel mode operation

architecture, there is global I/O sharing problem, that is, every bank should share the same global I/O and RX/TX. In HBM, there are 256 global I/O(128 I/O multiplied by 2bit prefetch) per channel. Comparing to that of DDR4, which has only 64 global I/O, it is difficult that every bank share 256 global I/O, from the standpoint of the speed of I/O and the area overhead.

The new architecture is proposed for solving these problems, which is the pseudo channel for HBM, as illustrated in Fig. 8. The pseudo channel architecture is one of solution for mitigating tRRD/tFAW, as well as the solution for the limitation of the number of banks. HBM is composed of 8 channels with 128 I/O each. According to the increased density and large IO numbers, a bank has been divided into two sub banks as described in section II. Therefore, a channel can be composed of 2 pseudo channels without much additional area penalty. Small peripheral logic area is increased due to the number of banks. Each pseudo channel share AWORD, but has separated banks, independent 64 I/O and BL4(2 subsequent of 2n-prefetch) is needed as well to meet access granularity.

High bandwidth memory has difficulty in reducing tFAW due to page size. In the pseudo channel architecture, each pseudo channel has reduced page size, that is, a half page in a channel. As a result, HBM can mitigate tRRD, from 2 clock to 4 clock, tFAW from 8 clock to 16 clock. The adoption of pseudo channel also increases the number of banks and the decrease in page size, which make the shorter core timing. There is another advantage of the pseudo channel architecture in random row access. The column access timing is increased twice due to the burst length increase, therefore, the required row access timing for random access can also be the twice value, as described in Fig. 9. In case of BL4 (2 subsequent of 2n-prefetch), random row access is more flexible without losing I/O bandwidth due to longer burst length.

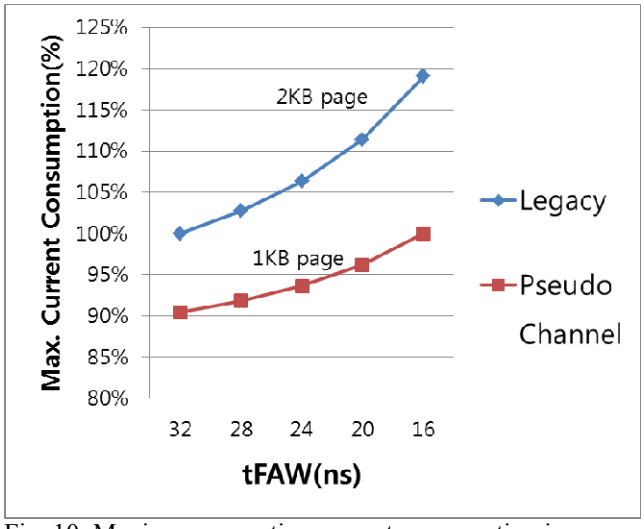


Fig. 10. Maximum operating current consumption increase versus tFAW

The estimation of current consumption vs tFAW is described in Fig. 10. Maximum current, which is IDD7, can be defined as IDD4R*bus_utilization + tRC/tRRD*row_activation_current. Row activation current is also defined as IDD0-tRAS/tRC*IDD3N-tRP/tRC*IDD2N. In the legacy mode, there is 2KB page size per row active. The estimated current in IDD4R condition of 256GB/s HBM is about 6~8A dependent on PVT variation without the interposer load. If we assume in simulation there are 75% bus used by read operation, there is 20% overall increase in the case of tFAW reduction from 32ns to 16ns. Therefore, the 20% increase of IR drop raises the fail rate. But, in the pseudo channel mode, there is not much current increase because the page size is reduced to a half.

HBM has the architecture that has multiplied bandwidth as the number of stacked dies increase. But the bandwidth is limited by the number of PHY and TSV I/O, which is 1024. If HBM uses 8-Hi stack structure, DRAM core bandwidth could be the double of 4-Hi stack structure. However, some system needs the memory density increase without the increase of the bandwidth or the number of channels. In the stacked memory, 8-Hi can be used for the increase of the number of banks rather than the bandwidth or channel. The compensation the gap between core

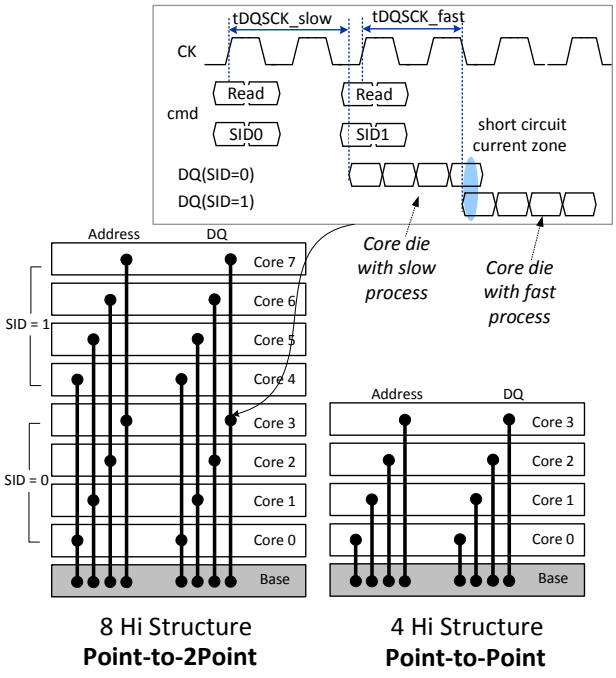


Fig. 11. 8-Hi and 4-Hi stacked DRAM structure.

bandwidth and I/O bandwidth can be made by increasing the effective bandwidth.

There are the increase in the density and the number of banks (from 32 to 64 per channel) in 8-Hi stack memory shown in Fig. 11. It is similar to the rank concept in DRAM module. The upper 4 stack core and the lower 4 stack core can be partitioned by SID (Stack ID) as shown in Fig. 11. The address and data are connected by point-to-2point configuration. Because each of the die has different process skew, read output timing from source clock (tDQSCK) cannot be matched. Various methods have been developed for the compensation of the TSV output timing [28][29]. However, in HBM, it is difficult that the read operation between lower die (SID=0) and the upper die (SID=1) has gapless operation, because the TSV I/O speed is relatively fast compared to conventional DRAM. The timing mismatch between the upper and the lower stack causes a large short circuit current. There should be the data gap of 1 or 2 cycles for the safe operation.

Another approach to organize the 8-Hi stack as P2P structure for gapless operation is the I/O width reduction, for example, from 128 I/O to 64 I/O per channel SID. In case that the banks of the upper stack and the lower stack are activated at the same time, the reduced I/O has the same effect as the twice of the prefetch, for example, from prefetch-2 to prefetch-4. The upper and the lower stack also use simultaneous column access at that condition. But it also needs SERDES in each I/O and the doubled prefetch make the page size larger. In this approach, the page size per channel is 4KB, and that makes large row activation current. Organizing the upper and the lower stack as each pseudo channel can be another solution for reducing page size, but it also has a lot of area penalty. It is possible to make various architectures using vertical connections.

The brief target SPEC comparison is shown in Table I. Compared to HBM 1st gen., HBM 2nd gen. has the twice of the target bandwidth, the density per channel and the number of banks are quadrupled. The page size is reduced to a half by using the pseudo channel memory architecture.

V. Conclusion

There are many design and test challenges in developing stacked-DRAM with microbump interface. Stacked HBM with four memories has more than four times higher bandwidth than ever developed component DRAM. It is shown that the first generation standard 8 Gb HBM operates at 128 GB/s at 1.2 V. The measurements of TSV resistance of HBM DRAM using current scan method is described, which shows that the measurement results of TSV resistance is under 0.1 ohm. Using 3D stack, the vertical structure enables the more diverse memory architecture than the 2D flat architecture. The next generation of HBM focuses on not only the bandwidth but also the system performance enhancement by adopting pseudo channel and 8-Hi stacking.

TABLE I
TARGET SPEC COMPARISON

Organization	HBM 1ST GEN	HBM 2ND GEN
Microbump ballmap	6.05mm x 3.26mm	
Microbump pitch	48 μm x 55 μm	
Supply voltage	VDD=1.2v, VDDQ=1.2v, VPP=2.5v	
Organization	8 channel x 8 banks x 128 I/O	8 channel x 16 banks x 2 P.C x 64 I/O
Density	1Gb / channel	4Gb / channel
Addresses	RA<0:12>, CA<0:5>, BA<0:2>	RA<0:13>, CA<0:5>, BA<0:4>
Page size	2KB/bank	1KB/bank (/P.C.)
Bank group	No	4 bank x 4 group (/P.C.)
Output Driver (C_{IO})	6mA~12mA (0.4pF)	6mA~18mA (0.4pF)
8-Hi stack	No	Yes
Max. data rate	1.0 Gb/s/pin	2.0 Gb/s/pin
Target Bandwidth	128GB/s	256GB/s

ACKNOWLEDGEMENT

The authors thank to Jang Ryul Kim of the product enabling team, Yun Hyeong Hoe of the product test team, many engineers of the package design team, the product application team and the device engineering team.

REFERENCES

[1] E. Doller, et al., "DataCenter 2020: Near-memory Acceleration for Data-oriented Applications," *Symposium on VLSI Circuits Digest of Technical Papers*, June 2014.

[2] J. Y.-C. Sun, "System Scaling and Collaborative Open Innovation," *Symposium on VLSI Technology Digest of Technical Papers*, pp.2-7, June 2013.

[3] K. T. Malladi, et al., "Towards Energy-Proportional Datacenter Memory with Mobile DRAM", *Proc. of International Symposium of Computer Architecture (ISCA)*, pp. 37-48, June 2012

[4] J. Kim, C. S. Oh, et al., "A 1.2V 12.8GB/s 2Gb Mobile Wide-I/O DRAM with 4×128 I/Os Using TSV-Based Stacking," *IEEE J. Solid-State Circuits*, vol. 47, no. 1, pp. 107–116, Jan. 2012.

[5] L. Madden, E. Wu, et al., "Advancing High Performance Heterogeneous Integration Through Die Stacking," *European Solid-State Device Research Conference (ESSDERC)*, pp. 18-24, Sept. 2012.

[6] D. Maheshwari, "Memory and System Architecture for 400Gb/s Networking and Beyond," *ISSCC Dig. Tech. Papers*, pp. 116-117, Feb. 2014.

[7] M.-S. Lin. C.-C. Tsai, et al., "An extra low-power 1Tbit/s bandwidth PLL/DLL-less eDRAM PHY using 0.3V low-swing IO for 2.5D CoWoS application," *Symposium on VLSI Circuits Digest of Technical Papers*, pp.C16-C17, June 2013.

[8] D. U. Lee, K. W. Kim, et al., "A 1.2 V 8 Gb 8-Channel 128 GB/s High-Bandwidth Memory (HBM) Stacked DRAM With Effective I/O Test Circuits," *IEEE J. Solid-State Circuits,* vol. 50, no. 1, pp. 191–203, Jan. 2015

[9] JEDEC Standard High Bandwidth Memory (HBM) DRAM Specification, 2013.

[10] T.-Y. Oh, Y.-S. Sohn, et al., "A 7 Gb/s/pin 1 Gbit GDDR5 SDRAM With 2.5 ns Bank to Bank Active Time and No Bank Group Restriction," *IEEE J. Solid-State Circuits*, vol. 46, no. 1, pp. 107–118, Jan. 2011.

[11] JEDEC Standard LPDDR3 SDRAM Specification, 2012

[12] T. M. Hollis, et al., "Data Bus Inversion in High-Speed Memory Applications," *IEEE Transactions on Circuits and Systems II*, vol.56, no. 4, pp. 300-304, Apr. 2009.

[13] Y.-C. Lai, S.-Y. Huang, "Robust SRAM Design via BIST-Assisted Timing-Tracking (BATT)," *IEEE J. Solid-State Circuits*, vol.44, no.2, pp. 642-648, Feb. 2009

[14] S. Ohbayashi, M. Yabuuchi, et.al., "A 65 nm embedded SRAM with wafer level burn-in mode, leak-bit redundancy and Cu e-trim fuse for known good die," *IEEE J. Solid-State Circuits*, vol. 43, no. 1, pp. 96–108, Jan. 2008.

[15] S. Takaya, M. Nagata, et al., "A 100GB/s Wide I/O with 4096b TSVs Through an Active Silicon Interposer with In-Place Waveform Capturing," *ISSCC Dig. Tech. Papers*, pp. 434-435, Feb. 2013.

[16] Y. Liu, W. Luk, et al., "A Compact Low-Power 3D I/O in 45nm CMOS," *ISSCC Dig. Tech. Papers*, pp. 142-143, Feb. 2012.

[17] B. Keller, T. Bartenstein, "Use of MISRs for Compression and Diagnostics," *Proc. International Test Conference (ITC)*, pp. 735-743, Nov. 2005.

[18] A.L.S. Loke, B.A. Doyle, et.al, "Loopback architecture for wafer-level at-speed testing of embedded HyperTransportTM processor links," *Custom Integrated Circuits Conference(CICC)*, pp.605 – 608, Sept. 2009.

[19] J. S. Park, et al., "PDN Impedance Modeling and Analysis of 3D TSV IC by Using Proposed P/G TSV Array Model Based on Separated P/G TSV and Chip-PDN Models" *IEEE Transactions on Components, Packaging, and Manufacturing Technology*, Vol. 1, No. 2, pp.208 – 219, Feb. 2011

[20] C. Y. Lee, et al., "TSV Technology and Challenges for 3D Stacked DRAM," *Symposium on VLSI Technology Digest of Technical Papers*, June 2014.

[21] D. U. Lee, et al., "An exact measurement and repair circuit of TSV connections for 128GB/s high-bandwidth memory(HBM) stacked DRAM, " *Symposium on VLSI Circuits Digest of Technical Papers*, June 2014

[22] S. Bae et al., "A 60nm 6Gb/s/pin GDDR5 Graphics DRAM with Multifaceted Clocking and ISI/SSN-Reduction Techniques," *ISSCC Dig. Tech. Papers*, pp. 278-279, Feb. 2008.

[23] Shiratake, S, et al., "A pseudo multi-bank DRAM with categorized access sequence," *Symposium on VLSI Circuits Digest of Technical Papers*, pp.127-130, June 1999.

[24] G. Atwood, et al., "A semiconductor memory development and manufacturing perspective," *European Solid-State Device Research Conference (ESSDERC)*, Sept. 2014

[25] R. Kho et al., "75nm 7Gb/s/pin 1Gb GDDR5 graphics memory device with bandwidth-improvement techniques," *IEEE J. Solid-State Circuits*, vol. 45, no. 1, pp. 120–133, Jan. 2010.

[26] Yoongu Kim, et al., "A case for exploiting subarray-level parallelism (SALP) in DRAM," *International Symposium on Computer Architecture (ISCA)*, pp. 368-379, 2012

[27] Donghyuk Lee, et al., "Tiered-latency DRAM: A low latency and low cost DRAM architecture," *International Symposium on High Performance Computer Architecture*, pp. 615-626, 2013

[28] R. Oh, et al., "Design Technologies for a 1.2V 2.4Gb/s/pin High Capacity DDR4 SDRAM with TSVs," *Symposium on VLSI Technology Digest of Technical Papers*, June 2014.

[29] S.-B. Lim, et al., "A 247 μW 800 Mb/s/pin DLL-Based Data Self-Aligner for Through Silicon via (TSV) Interface," *IEEE J. Solid-State Circuits*, vol. 48, no. 3, pp. 711–722, Mar. 2013