High Bandwidth Memory (HBM) and High Bandwidth NAND (HBN) with the Bumpless TSV Technology

Koji Sakui1,2)  
1) Tokyo Institute of Technology, IIR, WOW Alliance  
Yokohama, Japan sakui.k.aa@m.titech.ac.jp  
2) Honda Research Institute Japan Co., Ltd.  
Wako-shi, Japan koji.sakui@jp.honda-ri.com  

Takayuki Ohba1)  
1) Tokyo Institute of Technology, IIR, WOW Alliance  
Yokohama, Japan  
ohiba.t.ac@m.titech.ac.jp

Abstract — This paper proposes a fundamental architecture for the High Bandwidth Memory (HBM) with the bumpless TSV for the Wafer-on-Wafer (WOW) technology. The bumpless interconnects technology can increase the number of TSVs per chip with fine pitch of TSVs, and reduce the impedance of the TSV interconnects with no bumps. Therefore, a further higher speed and higher density HBM can be realized. Also, the High Bandwidth NAND (HBN), which can read and program by plane instead of by line by using the bumpless TSV, has been proposed.

Keywords—wafer-on-wafer; bumpless; thinning; TSV; high density integration; HBM; HBN

I. INTRODUCTION

TSVs with micro-bumps are conventionally used for the High Bandwidth Memory (HBM) [1-4]. However, there are several issues using the micro-bumps. One major problem is that it should be difficult for even the HBM to catch up with the speed of the GPU or CPU. For example, the Pascal of NVIDIA is 1TB/s, so that four sets of the HBM with 256GB/s are forced to use for the operation of the Pascal. The GPU/CPU vendors are challenging their products speed up much more, like 2TB/s and 4TB/s, for emphasizing the AI system, such as an advanced automated driving. However, the HBM speed should be getting almost saturated to 341GB/s [4].

On the other hand, the 3D stacking, in combination with the conventional 2D integration, has been studied extensively [5-8], because it is well understandable that the conventional two-dimensional (2D) scaling should be forced to face with a severe economic crisis due to the expensive lithography process and facilities required. However, reducing the TSV pitch and impedance has not been focused on in details. This paper is devoted to increasing the number of TSVs per chip with fine pitch of TSVs, and reducing the impedance of the TSV interconnects with no bumps [9-14]. As a result, a further higher speed and higher density HBM, and the High Bandwidth NAND (HBN) by using the bumpless TSV, can been proposed [15-16].

II. BENEFITS OF 3D STACKED MEMORIES WITH BUMPLESS TSV

1. Shorten Physical Wiring Length

Fig. 1 illustrates the wiring length comparison between the 2D and 3D. In the 3D layout, the block-to-block length is 10 ~ 100 μm, while that of the 2D layout is 1 mm ~ 1 cm.

2. Thin Die Thickness

Fig. 2 shows the die thickness comparison between the 2D and 3D. Because the bump is very large, the die thickness of at least 30 μm is needed for the robust stiffness. As a result, the stacked chip pitch is something like 100 μm. If 16 chips are stacked, the total height should be 1600 μm. On the contrary, thanks to the bumpless technology, the chip pitch becomes one tenths. Therefore, the total height of the 16 chip stacked case would be only 160 μm.
3. **Low Impedance Interconnects**

   The total length for the bumpless interconnect is 10 $\mu$m, while that for the bump interconnect is 100 $\mu$m, as shown in Fig. 3. Therefore, the impedance of the bumpless TSV is drastically reduced to at least one tenth of that of the bump TSV.

Fig. 3 Comparison between bump and bumpless interconnects.

4. **Massive Parallelism Bumpless TSVs**

   With respect to the bump TSV, the bump pitch is limited by the bump-to-bump short failure due to the stress migration, as shown in Fig. 4. Furthermore, the non-contact open failure, and Back-End-of-Line destruction would be occurred. Thereby, the TSV pitch cannot be tightened, so that the number of channels is limited. Conversely, the bumpless technology can tighten the TSV pitch, which increases the number of channels.

Fig. 4 Massive parallelism bumpless TSVs.

5. **Low Heat Generation**

   The bumpless technology can lower the interconnection thermal resistance, as shown in Fig. 5 [17]. The Cu TSVs penetrate all layers and act like a thermal “highway”. In the bumpless stacked system, the total volume can be small, so that the density of the interconnect metal becomes high. Therefore, the total thermal resistance by the bumpless TSVs becomes very small. Fig. 6 shows a temperature rise simulation result. While the temperature rise of layer 0 of the micro-bump type is 20 degree C, that of the bumpless type is only 5.8 degree C, so that the bumpless type allows a sophisticated design that can reduce the temperature rise into 30%. As a result, the eight die stacked HBM can be designed with this bumpless technology, and no modification in the refresh time of layer-by-layer is needed. However, the conventional TSVs with micro-bumps have a temperature rising problem, which should be more significant according as the number of stacked dies increases over four dies.

Fig. 5 Thermal design and reduction.

Fig. 6 Temperature rise simulation result.

III. **COMPETITIVE 3D STRUCTURE**

1. **HBM: High Bandwidth Memory (RAM)**

   The first target for the competitive 3D structure enables the 8-die stacks with bumpless TSVs, as illustrated in Fig. 7. By increasing the number of channels and lowering the TSV impedance, an ultra-high bandwidth of 1, 4, and 8 TB/s should be achieved.

Fig. 7 HBM with bumpless TSVs.

The first target for the competitive 3D structure enables the 8-die stacks with bumpless TSVs, as illustrated in Fig. 7. By increasing the number of channels and lowering the TSV impedance, an ultra-high bandwidth of 1, 4, and 8 TB/s should be achieved.

Fig. 8 shows the roadmap for the HBM data bandwidth. Thanks to the parallelism enhancement due to the increase in the
number of I/O’s, the bandwidth of the bumpless HBM is expected to be ever-increasing. With respect to the I/O power consumption, the first target of the bumpless HBM is one thirtieths that of the current HBM2 [18], as shown in Fig. 9. Fig. 10 compares the HBM2 with the bump and the bumpless HBM with respect to their data bandwidth and I/O buffer power, according as the I/O number. The bumpless HBM can drastically enhance the I/O parallelism, which should achieve the ultra-high speed of 8.5TB/s with the relatively low I/O buffer power of 5.4W.

2. HBN: High Bandwidth NAND (ROM)~ Word Plate Access ~

In the conventional word line access scheme, if we would make the through-put double, double page buffer, or half bit line pitch, is needed [19]. Of course, additional page buffers in CUA, CMOS Under Array, and divided bitlines in sub-arrays might have a solution. But both the CUA area and the number of sub-arrays have limitations.

At VLSI Technology ’07, Toshiba proposed an original 3D NAND at first, as shown in Fig. 11 [20]. Multiple CGs and Lower SGs should be merged in each plate for reducing the number of HV-driver transistors, as well as for tightening the pillar pitch. Right now, the poly silicon word plate is replaced by the damascene tungsten metal gate. But the basic 3D NAND structure of the merged multiple CGs and Lower SG is the same as the original one. In the past, a future 3D stacked memory design was issued as patents, as presented in Fig. 12 [21]. Thanks to increasing the number of TSVs, a part of a peripheral circuit of the first memory chip can be located in the second memory chip. An example shows the stacked DRAM in combination with the NAND Flash memory.

Figs. 13 and 14 have proposed the HBN (High Bandwidth NAND) with multiple BL layers. The read and program speed of the conventional NAND Flash was limited by the number of the page buffers, because the memory cell data were read or programmed through one line of the page buffers. Thanks to the 3D NAND original architecture, a word line has expanded to a word plate, so that the 3D NAND can read and program by plane instead of by line.
WOW technology with bumpless interconnects using TSVs for three-dimensional stacking in wafer form has been described. An optimized thinning wafer thickness of 4 μm can increase the number of TSVs per chip with fine pitch of TSVs, and reduce the impedance of the TSV interconnects with no bumps. Therefore, a further higher speed and higher density HBM can be realized. Also, the High Bandwidth NAND (HBN), which can read and program through-silicon via and die stacking technologies for microsystems-integration, in IEEE IEDM Tech. Dig., pp. 495-498, Dec. 2008.

According as the number of the stacked memory chips is increased, the total memory density should be huge like an enterprise. Therefore, as an example, an AI robotic bee, which has a CPU, ultra-small enterprise, HBM, HBN, and sensors, should be eventually realized in 50mm³ with 0.5mW for human assistant, as proposed in Fig. 15.

Fig. 15 AI Robotic Bee (50mm³, 0.5mW) for a human assistant.

REFERENCES


