Extended Abstracts of the 2020 International Conference on Solid State Devices and Materials, VIRTUAL conference, 2020, pp145-146

Bumpless Build Cube (BBCube) using Wafer-on-Wafer (WOW) Technology with 3D-manner Redundancy Scheme

Shinji Sugatani, Norio Chujo, Koji Sakui, Hiroyuki Ryoson, Tomoji Nakamura, and Takayuki Ohba

Tokyo Institute of Technology, IIR, WOW Alliance.
Nagatsuda-cho, Midori-ku, Yokohama 226-8503, Japan
Phone: +81-45-924-5866 E-mail: sugatani.s.aa@m.titech.ac.jp

Abstract

An application of vertically replaceable memory block architecture scheme, hereinafter referred to as "3D-manner redundancy" for BBCube is presented. Productivity of better than current Known Good Die (KGD) stacking process will be shown, which leads to conclude that wafer level fabrication is possible. Superior energy efficiency to the conventional HBM structure will be shown, which has been resulted from lower TSV capacitance by ultra-thinning technique of silicon, and bumpless feature, combined with higher parallelism by denser TSVs.

1. Introduction

We have presented a high parallelism stacked DRAM called BBCube, fabricated by WOW technology, which provides higher density and lower capacitance TSVs [1].

To realize the wafer-on-wafer fabrication, it is inevitable to investigate defect management design, especially for random defect, since the probability of randomly defective portions being included in the module stack, must not be eliminated. In this study, we present 3D-manner redundancy [2], applied for the configuration of the first generation BBCube.

2. Application of 3D-manner redundancy for BBCube

Configuration of BBCube

Fig. 1 shows the configuration and structural hierarchy of BBCube, which is consisted of 16 tiles. Each tile includes at least four banks with 1024 data width I/Os. Here, we assume 16 pairs of sub-arrays with extra sub-arrays for individual-die-in-each-layer basis (2D-manner) redundancy, which is combined with 3D-manner redundancy.

Totally 9 stacks, consisted of 8 stacks and redundantly added 1 stack, are used, to accumulate required 32 fine banks from sparing resources of 36 banks. All banks are completely equivalent with each other, so that they are mutually replaceable across the stacked layers.

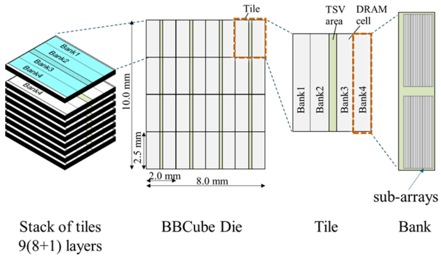


Fig. 1 Configuration and structural hierarchy of BBCube. Banks consisted of sub-arrays for 2D-manner redundancy, and stacked tiles of 9 layers for 3D-manner redundancy are illustrated.

Yield calculation

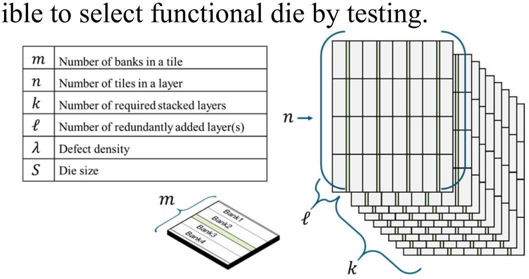
The following formula (2) presents the model for yield of BBCube, Y_{BBCube} with 3D-manner redundancy scheme illustrated in this study. We assume a Poisson distribution model for the random defect yield.

$$Y_{single\ die} = \exp(-\lambda S) \quad \cdots (1)$$

$$Y_{BBCube} = \left[\sum_{i=k\times m}^{(k+\ell)\times m} \{ ({}_{(k+\ell)\times m}C_i) \cdot Y^i \cdot (1-\mathrm{Y})^{(k+\ell)\times m-i} \} \right]^n \quad \cdots (2)$$

, where each parameter is described in Fig. 2. When the single layer die yield $Y_{single\ die}$ is given as formula (1), bank yield is calculated as $Y_{Bank} = \exp[-\lambda S/(n \times m)]\,, Y \equiv Y_{Bank}$.

Total yield of the tile is calculated by summing for all product of bank yield and defect rate weighted by number of its combination. Random defect yield of targeted tile is calculated as a term in brackets of formula (2), so that the yield of the whole BBCube system can be attained, as the tile yield to the power of the number of tiles. In comparison with KGD process, BBCube yield is equal to $Y_{single\ die}$, because it is possible to select functional die by testing.



m	Number of banks in a tile
n	Number of tiles in a layer
k	Number of required stacked layers
ℓ	Number of redundantly added layer(s)
λ	Defect density
S	Die size

Fig. 2 BBCube stack configuration and symbols for calculation.

Yield comparison

As indicated in Fig. 3(a), BBCube by WOW process shows better yield than that of KGD stacking case, for all single layer die yield. The reason is as follows. For KGD case, if the required number of banks is not available in a silicon die, the die is forced to be disposed of in vain. On the contrary, for WOW case, when there is an error of banks, it is possible to spare the necessary number of banks from other layers in the total stack.

For productivity comparison, we need to consider area penalty of redundancy schemes. 3D-manner redundancy consumes 9 wafers to realize the function of 8 wafers. Therefore, the area penalty is 12.5%, hence, in the case that single die yield is greater than 87.5%, KGD process seems to be more productive. But to realize such excellent single die yield, area penalty of more than or equal to 12.5% is necessary for 2D-manner redundancy. Moreover, for almost all practical range of die yield without redundancy, to achieve more than 99% yield, 2D-manner redundancy costs 12.5% or more area penalty than 3D-manner redundancy, as illustrated in Fig. 3(b). For 3D-manner redundancy, more than 50% of single layer die yield is enough to achieve more than 99% BBcube yield.

Therefore, WOW process with 3D-manner redundancy provides productivity, better than KGD stacking case.

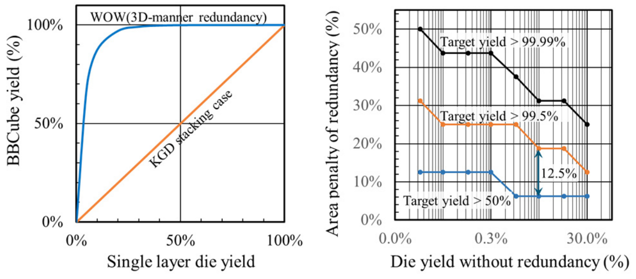


Fig. 3(a)

Fig. 3(b)

Fig. 3(a) BBCube yield comparison between WOW and KGD cases. Fig. 3(b) Area penalty of 2D-manner redundancy for target yields.

3D-manner redundancy at sub-array level

In the discussion so far, we assume that each bank provides 1024 width data to I/Os in the tile, so that mutual compatibility among the banks is guaranteed. Though, tiles of BBCube are already well-fine-grained, partition into narrower pseudo banks from a 1024 data width bank should be considered for better energy efficiency [3]. We investigated if we could use sub-array level, which is next hierarchy below bank level in BBCube configuration.

In general, sub-array level is used for 2D-manner redundancy, to achieve excellent yield for KGD process. With certain amount of area penalty, near100% yield can be obtained as illustrated in Fig. 3(b), assuming single bank, equipped with 16 sub-arrays and redundant sub-arrays. One extra sub-array costs 6.25% area penalty.

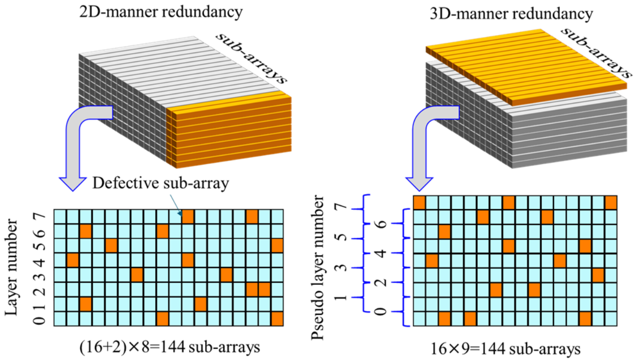


Fig. 4 In case that 2 redundant sub-arrays give near 100% yield, stack of single bank tiles with 1 extra layer tile includes 16 defective sub-arrays at maximum.

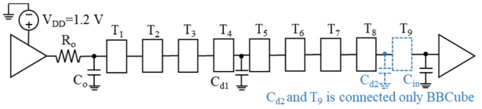
In case of 2 redundant sub-arrays give near 100% yield, maximum number of defected sub-arrays in the stack of tiles, must be 16. Thus, (16+2) ×8, equal to 144 sub-arrays must include 128 fine sub-arrays. Also 144, equal to 16×9 sub-arrays must include 128 fine sub-arrays. This means, neighboring 3 physical layers should include less than or equal to 16 defected sub-arrays. As a result, 8 pseudo layer assignment out of 9 physical layers is capable as illustrated in Fig. 4.

Such “2 more sub-arrays are enough” case is realized, and equivalently realized by yield improvement activities and more nested redundancy. Transferred data from replaced sub-arrays need to be bypassed across physical layers, in front of the data multiplexers which provide bank data to I/O buffers.

Accordingly, sub-array level 3D-manner redundancy is feasible, bank configuration in a tile should be flexible for energy efficiency optimization.

Power efficiency of BBCube

Eye diagram simulation results in Fig. 5 show the power efficiency of BBCube, compared with current HBM structure. I/O power consumption of one thirtieth is realized, with the same band width of 3.2 [Gb/s]. Total bandwidth of BBCube is 300% higher than HBM with 87% less power of I/Os, while prior work of bumpless scheme reports less aggressive energy efficiency and bandwidth improvement [4]. It seems to be that, the capacitance of TSVs remains high, because ultra-thinning of silicon substrate is not performed.



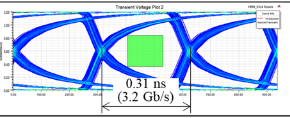
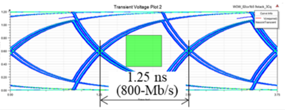
	Eye pattern	Power
HBM 8 stacks 2 drops		2142 μW (incl. pre-buf.)
BBCube 9 stacks 3 drops		71.5 μW (4 signals)

Fig. 5 Eye diagram simulation result of BBCube and HBM. Power efficiency of BBCube is 30 times better than current HBM structure.

3. Conclusions

Excellent performance of BBCube due to WOW technology, and application of 3D-manner redundancy for BBCube have been presented.

The wafer-on-wafer fabrication is realized with the support of 3D-manner redundancy scheme, leads to conclude BBCube as next system scaling enabler.

Acknowledgements

This study was carried out in WOW Alliance program of the Tokyo Institute of Technology.

References

[1] N.Chujo, et al., *the IEEE Symp. on VLSI Technology Dig.*, TH1.3, Jun. 2020.

[2] S. Sugatani et al., *the IEEE Trans. on Electron Devices, Special Section on ESSDERC 2020*, Nov. 2020, to be published.

[3] Mike O’Connor et al., 2017 50*th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*.

[4] C.H.Tsai, *the IEEE Symp. on VLSI Technology Dig.*, TH1.1, Jun. 2020.