**Terraced Memory Cube**

**CC Kuo, Percy Soto and DerChang Kau**

High Bandwidth Memory (HBM) plays a crucial role in enhancing AI accelerators due to its 3D stacking and in-package memory integration. A 3D stacking architecture vertically links data path using high density Through-Silicon Vias (TSVs) to widen data width with lower energy consumption. HBM addresses the memory wall bottleneck by offering significantly higher bandwidth than mainstream DRAM. However, the shared TSV footprint in every tier of memory blows up chip size and reduces array efficiency, leading to waste of Si area and process yield drop. Besides, pre-defined interconnection with built-in TSVs also limit flexibility and scalability. The oversized active Si interposer subject to package technology has very low silicon utilization. In addition, the 3D stacking architecture presents significant thermal dissipation challenge with most power consumed at the bottom of the 3D stack. In the paper, a simple scalable vertical interconnect scheme, called Terraced Memory Cube (TMC), is disclosed (Figure 1). The terraced chiplet cubing architecture (TCC) is created to address the cost and thermal issue in HBM. A TMC-LPW consists of three building blocks – Wide-IO DRAM with LPDDR interface in a package of TCC (Figure 2). TCC technology disaggregates vertical connectivity from active silicon in 3D stacking. Wide-IO improves parallelism of mainstream DRAM products and provides better array efficiency than a HBM die with on par bandwidth. LPDDR interface deploys the simplest connectivity for energy efficient high data rate transfer without needing a base die. Three TMC mockups, 16, 32 and 64GB, are modeled in D1c/D1γ node for circa’27.   All mockups exhibit superior density and cost in capacity and bandwidth than a 64GB HBM4+ with substantial margin (Table 1).

Table 1: Performance and cost benchmark – 16,32 and 64GB TMCs vs. HBM in circa’27

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Component** | **Capacity  [GB]** | **Die Density [Gb]** | **# of Die  in a stack** | **Norm.  GB/mm2** | **Norm. BW/mm2** | **Norm. $/GB** | **Norm.  $/BW** |
| 3DS HBM4+ | 64 | 32 | 17 | 100% | 100% | 100% | 100% |
| TMC LPW6 | 64 | 32 | 16 | 139% | 102% | 56% | 56% |
| TMC LPW6 | 32 | 32 | 8 | 139% | 63% | 49% | 49% |
| TMC LPW6 | 16 | 16 | 8 | 132% | 116% | 44% | 22% |

Figure 2: A TMC example for 3D stack of Low Power wide-IO (LPW) DRAM chips.



Figure 1: TMC – terraced stacking of memory chips interconnected vertically using low cost TSV chips to widen data width in a TCC construction.