Packaging for off-the-shelf commercial NAND Flash Memory chips

Suin Yi

Electrical & Computer Engineering Department
Texas A&M University, College Station

Analog in-memory-computing mainly by vector-matrix-multiplication acceleration has emerged recently due to the overwhelming computation demand that cannot be accommodated by existing digital computers including NVIDIA's GPUs such as H100 and B200. Non-volatile memories such as resistive random access memories (RRAMs) and phase change memories (PCMs) have been actively explored to showcase the feasibility and the projected energy efficiency as well as the throughput much improved over digital computers.

Now it is time for the technology to the market. NAND Flash memory has recently reached 28.5 Gb/mm^2 by Samsung's QLC (Quadratic Level Cell) 3D NAND [1], which is equivalent to 275 billion artificial synapses present in a 36 mm^2-sized chip. This is a 10,000× improvement compared to 3 million artificial synapses realized by RRAM cells in NeuRRAM chip [2].

In order to utilize the abundant artificial synapses readily available in the commercial off the shelf (COTS) 3D NAND Flash chips, newly designed peripheral circuitries such as transimpedance amplifiers (TIAs) and analog-to-digital converters (ADCs) are necessary. Tiny pitches of the bit lines (BLs) and the common source lines (CSLs) of sub 40 nm and 2 μm, respectively, need to be carefully handled for the wafer bonding between the cell wafer and the peripheral circuit wafer.

Recent advances of wafer bonding for COTS 3D NAND Flash memory products pioneered by Yangtze Memory Technologies Corp. (YMTC) [3] will be reviewed for the packaging infrastructure currently in use by NAND Flash market as well as the one for Flash-based VMM accelerators.

[1] 2024 ISSCC, Jung et al., 'A 280-Layer 1Tb 4b/cell 3D-NAND Flash Memory with a 28.5Gb/mm2 Areal Density and a 3.2GB/s High-Speed IO Rate'
[2] 2022 Nature, W. Wan et al., 'A compute-in-memory chip based on resistive random-access memory'
[3] 2022 VLSI Technology, Z. Huo et al., 'Unleash Scaling Potential of 3D NAND with Innovative Xtacking® Architecture'