4. Design and Implementation of 3D-DRAM Memory Controllers

To fully exploit the high bandwidth offered by Through-Silicon Vias (TSVs) in 3D-stacked DRAM systems, a hierarchical DRAM controller must be carefully designed to match the architectural characteristics of the memory stack. Unlike conventional 2D-DRAM systems—where the number of I/O pins is inherently limited—3D-DRAM leverages TSV-based vertical interconnects to achieve shorter interconnection lengths, lower signal latency, and significantly higher bandwidth. Moreover, by stacking the memory directly on top of the accelerator die, data can be delivered with minimal latency, enabling the processing elements (PEs) of the accelerator to operate closer to their theoretical peak performance and effectively alleviating the long-standing memory wall bottleneck.

However, the very wide I/O interface enabled by TSVs introduces new architectural and microarchitectural design challenges. Chief among these is how to fully utilize the available bandwidth while balancing power, thermal, and timing constraints. The memory controller, situated between the logic die and the 3D-DRAM stack, becomes the key component in orchestrating this data movement. Its design must account for factors such as traffic scheduling, bank-level parallelism, refresh management, and quality-of-service (QoS) policies to prevent underutilization of the TSV channels.

Therefore, the development of a dedicated 3D-DRAM memory controller—optimized for ultra-wide I/O operation and tight integration with on-chip accelerators—is critical to unlocking the full potential of 3D-stacked DRAM technology. A thorough analysis of the controller’s architecture, timing behavior, and interaction with the DRAM banks is essential to ensure that the system achieves both maximum throughput and energy efficiency under real-world workloads.

4.1 Overview of Hierarchical 3D-DRAM Controller