**2.1 Basic DRAM Structure & Architecture Overview**

To meet the demands of modern deep learning (DL) workloads, on-chip memory stacking has emerged as a promising solution. Several innovative 3D DRAM architectures have been developed to demonstrate the viability of stacked memory designs. These architectures leverage three-dimensional (3D) integrated circuit technology, utilizing through-silicon vias (TSVs) to vertically interconnect multiple layers. This vertical stacking significantly reduces memory access latency and increases bandwidth by overcoming the limitations of conventional I/O pins. In the following, we explore a range of 3D-DRAM architectures and examine the critical role of TSV technology in their implementation.

2.1.1 DRAM Architecture

A typical DRAM chip is organized hierarchically into multiple **banks**, each of which contains several **subarrays** and shares a central **I/O peripheral** interface for data access. As shown in Figure 2.1.1, four banks are connected to a shared I/O peripheral that handles communication over a common bus. Within each bank, **subarrays** are stacked vertically and managed by peripheral logic such as local decoders and sense amplifiers.

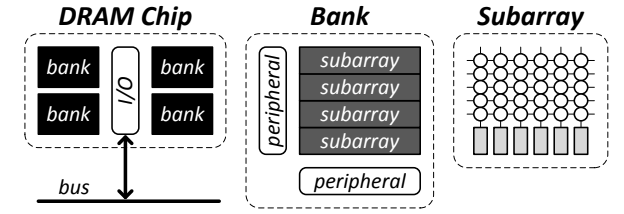
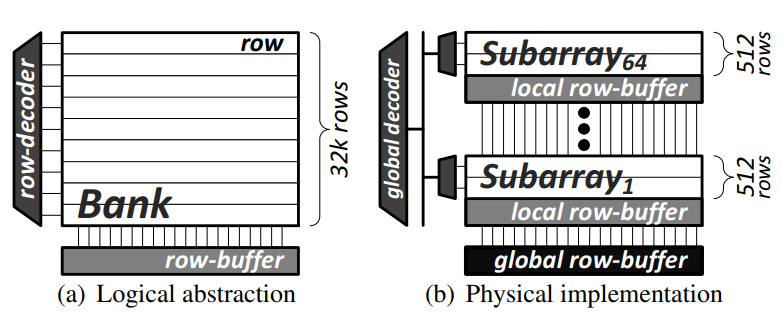


Figure 2.1.2 provides both a **logical abstraction** and a **physical implementation** of a DRAM bank. Logically, a bank is composed of thousands of rows (e.g., 32K rows), which are addressed through a row decoder and accessed via a global row buffer. Physically, these rows are distributed across multiple **subarrays** (e.g., 64 subarrays with 512 rows each). Each subarray contains a **local row buffer** that temporarily holds the accessed row data. During an **activation operation**, data from the local row buffer is transferred to the **global row buffer**, enabling subsequent read or write operations. This hierarchical buffering improves both area efficiency and access latency.



Each subarray is further divided into **mats**, where every mat contains its own set of sense amplifiers and wordlines, forming the basic units of data storage and access.

一張含有 文字, 圖表, 行, 方案 的圖片

自動產生的描述

2.1.2 DRAM Command & Essential Operations

The following figure explain the basic DRAM access operations

**1. Row Activation (ACTIVATE)**  
Before accessing a row, the DRAM bank must be in the **precharged state**, where all bitlines are equilibrated to **1/2VDD**. Upon receiving an **ACTIVATE** command with a specific row address, the corresponding **wordline** is elevated to **VPP**, connecting the memory cells of the target row to the bitlines. This connection causes a slight voltage disturbance on the bitlines, depending on whether the cells are charged (1) or discharged (0). The **sense amplifiers** in the **row buffer** detect this perturbation and amplify it to full logic levels, latching the row data. During this sensing phase, cell contents are temporarily undefined. Once amplification completes, the original data is restored to the cells. The total duration of this operation is defined by the timing parameter **tRAS** (Row Access Strobe), typically around **35 ns**.

**2. Column Access (READ/WRITE)**  
After activation, the **memory controller** issues a **READ** or **WRITE** command with a column address to access specific data within the activated row. A timing delay, denoted as **tRCD** (Row to Column Delay), typically **~15 ns**, must elapse between the ACTIVATE and column command. This delay allows the row buffer to fully capture and stabilize the data. If a subsequent access targets the same row, it can proceed immediately with a READ/WRITE command, avoiding a new activation and thus reducing latency.

**3. Bank Precharge (PRECHARGE)**  
To access a different row within the same bank, the current row must first be **closed** by returning the bank to the **precharged state**. This process involves two steps: the active **wordline is lowered** to disconnect the memory cells, and the **bitlines are re-equalized** to **1/2VDD**. The time required for this operation is captured by the **tRP** (Row Precharge) parameter, typically around **15 ns**.

一張含有 文字, 字型, 圖表, 行 的圖片

自動產生的描述

**4. Refresh(REF)**

DRAM requires periodic refresh operations to preserve data integrity, as each cell gradually leaks charge over time. These refreshes momentarily block access to the corresponding banks or rows, making it essential for the memory controller to schedule them efficiently without significantly degrading performance.

**2.1.3 DRAM Timing Constraints**

The execution of DRAM operations is strictly governed by a set of timing constraints, which are essential for maintaining correct and reliable memory behavior. These constraints are determined by the internal architecture of the DRAM device and are specified by the DRAM manufacturer based on worst-case conditions, ensuring robust operation under all supported scenarios.

As shown in Figure X, each timing parameter defines the minimum delay between specific command sequences—such as tRCD (ACTIVATE to READ), tRTP (READ to PRECHARGE), or tWR (WRITE to PRECHARGE)—and must be respected by the memory controller to prevent data corruption and maintain signal integrity across banks, rows, and I/O channels.

A memory controller is responsible for enforcing all of these timing constraints while simultaneously striving to maximize performance, typically through techniques such as request reordering, bank-level parallelism, and row-buffer locality optimization.

The table below summarizes additional critical timing constraints, including **tRC** (row-to-row activation delay), **tRP** (precharge-to-activate delay), **tRTW** and **tWTR** (read/write turnaround times), as well as **CL** (CAS latency) and **CWL** (CAS write latency), which represent the delay from column commands to data transfer.一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

**2.2 TSV technology and state-of-art 3D-DRAM** **Arhictecture**

3D-stacked DRAM uses vertical Through-Silicon Vias (TSVs) to connect memory layers, allowing data to move quickly between layers and to the processor. This design provides much higher internal bandwidth than 2D DRAM, which relies on slower, horizontal connections and external channels with limited bandwidth. Compared to 2D DRAM, 3D-stacked DRAM offers faster data access, uses less power, and takes up less space—making it better suited for high-performance systems like AI processors, GPUs, and data centers.

2.2.1 State-of-Art High Bandwidth Memory(HBM)

Figure X is the basic structure of HBM2 and the basic Dataflow of HBM2, where the dataflow of a read operation is being described. Such technology enables more bandwidth for DRAM.

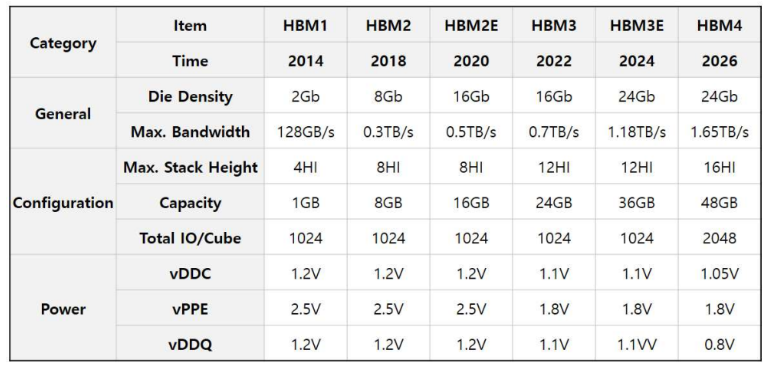
.**一張含有 文字, 螢幕擷取畫面, 平行, 設計 的圖片

自動產生的描述**

Figure X: Architecture of state-of-art 3D-DRAM HBM2 Chip

The figure highlights the progression of HBM technologies from HBM1 to HBM4, showing significant improvements in memory capacity, bandwidth, and energy efficiency. Over time, each new generation has increased die density, expanded stack height, and doubled or even tripled bandwidth. For example, HBM4 offers higher stack heights and over ten times the bandwidth of HBM1, while also reducing operating voltage levels to improve power efficiency.

These advancements demonstrate the rapid evolution of **3D-stacked DRAM** technologies, driven by growing demands in AI, HPC, and data center applications. As bandwidth continues to rise and access energy decreases, integrating the latest HBM technologies becomes essential for achieving high-performance, energy-efficient memory systems in next-generation computing platforms.



**Figure X**: *Evolution of High Bandwidth Memory (HBM) generations.*

The figure illustrates the relative energy consumption for various memory access patterns. It shows that 3D DRAM consumes significantly less energy than 2D DRAM, particularly for random access. This demonstrates the energy efficiency advantage of 3D-stacked DRAM, making it a favorable choice for modern high-performance and low-power computing systems.

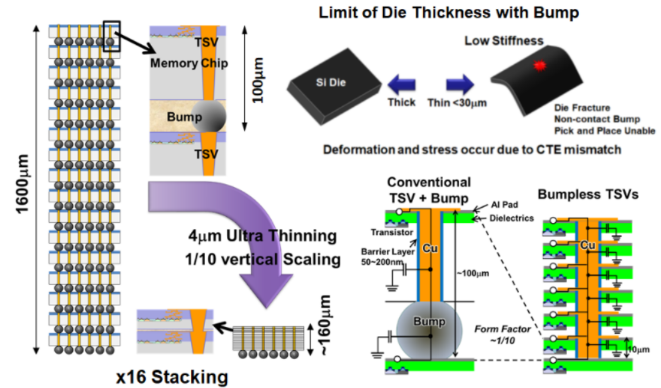
**一張含有 文字, 字型, 螢幕擷取畫面, 數字 的圖片

自動產生的描述**

**Figure X**: *Energy comparison between 2D and 3D DRAM operations.*

2.2.2 Very High Bandwidth Memory Key Technologys

One of the key enabling technologies in 3D-stacked DRAM is the use of Through-Silicon Vias (TSVs)—vertical electrical interconnects that pass through silicon wafers or dies. TSVs play a critical role in enhancing performance, energy efficiency, and scalability by enabling high-bandwidth, low-latency communication between stacked memory layers. This vertical integration not only reduces interconnect capacitance and power consumption but also increases storage density through compact and efficient memory stacking. The figure shows the 16 stacking bumpless TSV interconnection technology



The Wafer-on-wafer hybrid bonding process, where a DRAM wafer is directly stacked on top of a logic wafer. By aligning and bonding the two wafers face-to-face, the communication distance between memory and processing units is significantly reduced.

**一張含有 螢幕擷取畫面, 圖表, 設計, 圖解 的圖片

自動產生的描述**

This tightly coupled stacked structure enables low-latency data transfer and improved energy efficiency, which are essential features of modern 3D-DRAM architectures. The packaging also includes thermal management components such as a heat sink and thermal adhesive, ensuring reliable operation in high-performance systems.

一張含有 圖表, 圖畫, 文字, 寫生 的圖片

自動產生的描述

The integration of TSVs relies on advanced bonding techniques, including copper-to-copper (Cu-Cu) bonding, Cu-DBI (Direct Bond Interconnect), and hybrid wafer bonding. In Cu-Cu bonding, deep vias are etched, filled with copper, and planarized before being bonded to adjacent layers. Hybrid bonding, a more sophisticated method, simultaneously bonds wafers at both the dielectric and metal layers, improving electrical performance and reliability.

???? ??, ?????? ???

???????

Stacking the DRAM logic die onto the logic die enables higher bandwidth for the Memory System. The figure shows a state of art Processing Near Memory system where HBM DRAM arrays are stacked directly on top of the logic die to enable the fast access and high bandwidth for the Match Engine IP blocks. illustrates the full chip package, including **heat sink**, **thermal adhesive**, and **substrate**, showing how the entire processing near memory system is assembled. The figure shows the Processing Near Memory Engine for Recommendation System using 3D Logic-to-DRAM Hybrid Bonding

**一張含有 文字, 螢幕擷取畫面, 圖表, 地圖 的圖片

自動產生的描述**

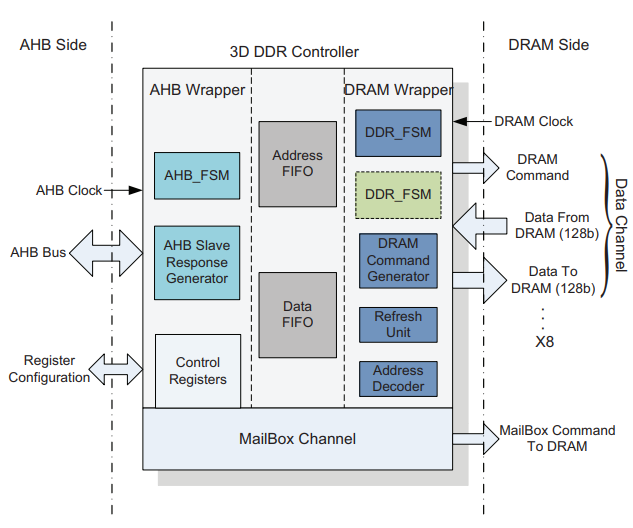
[] 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process Near-Memory Engine for Recommendation System

2.3 Previous Work of 3D-DRAM Controllers

To fully exploit the performance and efficiency benefits of 3D-stacked DRAM architectures, it is essential to design a specialized DRAM controller that is tailored to the unique characteristics of vertically integrated memory systems. Unlike conventional memory controllers, which are optimized for planar (2D) DRAM, controllers for 3D-DRAM must efficiently manage the increased internal bandwidth, reduced latency, and tighter coupling between memory and logic layers. Such controllers must be capable of handling the high-throughput demands of stacked memory, supporting advanced features such as fine-grained access scheduling, thermal-aware data management, and optimized TSV utilization. By aligning the controller architecture with the physical and functional layout of 3D-DRAM, the memory system can achieve significant improvements in bandwidth utilization, power efficiency, and overall performance.

2.3.1 Customized DRAM Controller for 3D-DRAM[]

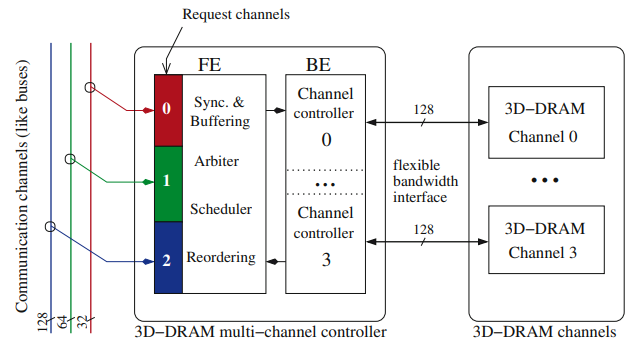
Tao Zhang et al. proposed a 3D DDR memory controller integrated within a 3D-stacked System-on-Chip (SoC) architecture, specifically designed to support real-time multimedia applications such as H.264 video encoding. The design leverages the high bandwidth enabled by Through-Silicon Vias (TSVs) and adopts a parallel access policy to overcome traditional I/O pin limitations. By exploiting the parallelism of multiple memory channels, the proposed controller demonstrates the feasibility and effectiveness of utilizing 3D-stacked DRAM to achieve high data throughput and energy efficiency. The parallel access mechanism enables concurrent memory requests across different channels, further enhancing memory system performance for bandwidth-intensive applications.



**[] Architecture of Parrallel Access Policy Memory Controller**

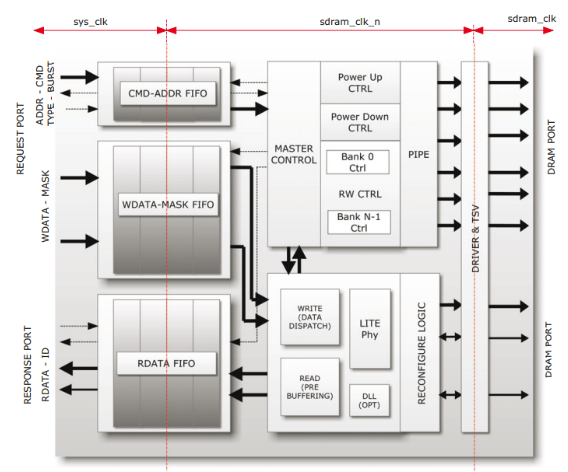
2.3.2 Energy Efficient DRAM Subsystem for 3D SoC

Christian Weis et al. proposed a 3D-DRAM memory subsystem architecture composed of a Memory Channel Controller and multiple 3D-DRAM channels. The design is divided into a front-end and a back-end. The front-end handles incoming memory requests using FIFO buffers, along with arbitration and scheduling mechanisms. The back-end consists of channel controllers that manage command generation and communication for each 3D-DRAM channel.



[] Logical View of DRAM Memory Subsystem

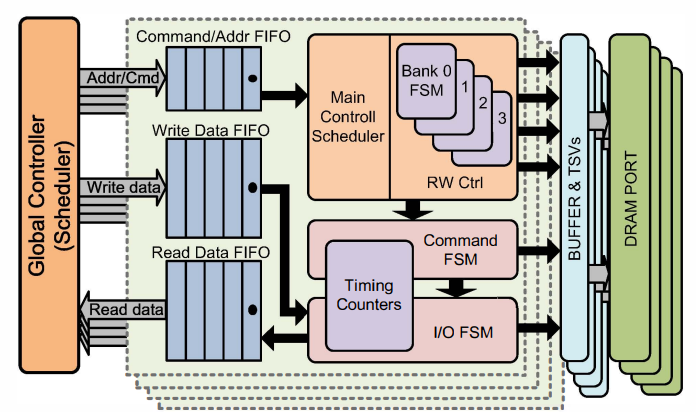
Each channel controller connects to its corresponding DRAM channel through a flexible interface that supports a wide range of bandwidths and burst lengths. This adaptable interface allows the subsystem to operate efficiently across different configurations. The overall system demonstrates low power consumption and high flexibility, making it well-suited for memory-intensive applications that require scalable performance and energy efficiency.



**[] Architecture of Channel Memory Controller**

2.3.3 Thermal-Aware Memory Management Unit of 3DStacked DRAM for 3D High Definition (HD) Video

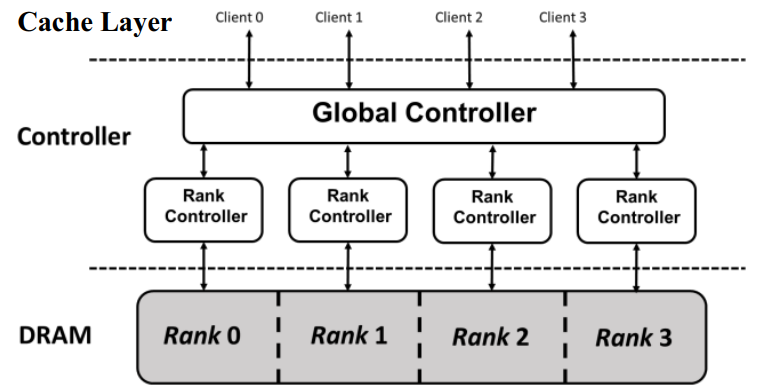
Chih-Yuan Chang et al. proposed a hierarchical memory controller architecture for 3D-stacked DRAM, aimed at supporting high-definition (HD) video system processors. The design features a two-level DRAM controller tailored to the structure of 3D-stacked memory, allowing efficient use of independent channels and the high bandwidth provided by the stacked architecture. To address the thermal challenges common in 3D-DRAM systems, the controller includes a mechanism that automatically adjusts the refresh rate based on temperature. Additionally, it incorporates an efficient address translator and a dynamic, thermal-aware timing control to improve overall reliability and performance.



[]Architecture of the Thermal aware 3D-DRAM Controller

2.3.4 WIDE-I/O 3D-STAKED DRAM CONTROLLER FOR NEAR-DATA PROCESSING SYSTEM

Yu-Hsuan Lin et al. proposed a hierarchical 3D-DRAM memory controller consisting of a global controller and local rank controllers, designed to enhance energy efficiency and performance. The controller adaptively adjusts I/O width, employs command rescheduling and rank interleaving for multi-user parallelism, and integrates self-refresh for power savings. Additionally, near-data processing reduces data movement overhead. This design achieves notable improvements in bandwidth, execution time, and energy consumption.

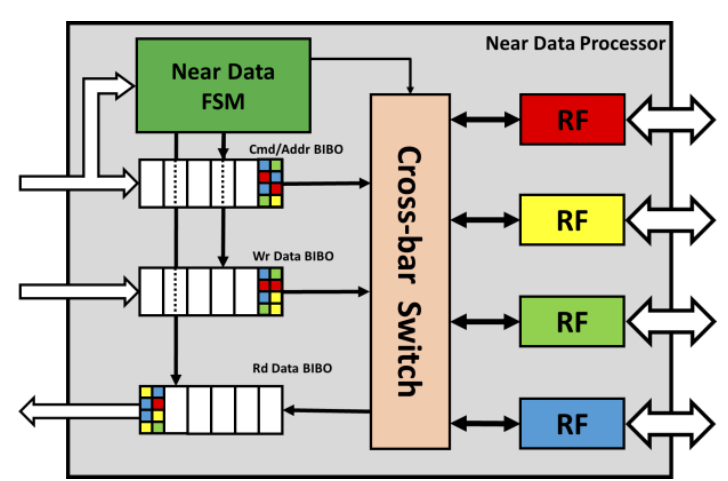


[]Hierarhical Memory Management Unit

To support processing-near-memory (PNM) applications, a near-data processor is introduced, designed to enable efficient and low-latency data exchanges between different memory ranks. The architecture consists of a Near Data Finite State Machine (FSM) that manages control flow and command handling. It interfaces with three Block-In Block-Out (BIBO) FIFOs for command/address, write data, and read data buffering. These FIFOs serve as temporary storage and help decouple timing between memory accesses and internal processing.

A key component of the architecture is the cross-bar switch, which enables dynamic routing of data between I/O ports and multiple register files (RFs). This allows for fast, parallel access and supports efficient data exchange across memory ranks with flexible and scalable communication.

By placing the processor close to memory and using this cross-bar and FIFO-based architecture, the system reduces data transfer latency and energy consumption, making it highly effective for data-intensive workloads in 3D-stacked DRAM environments.



[]Architecture of Near-Data Processor