3. Architectural Simulator of 3D-DRAM Memory system

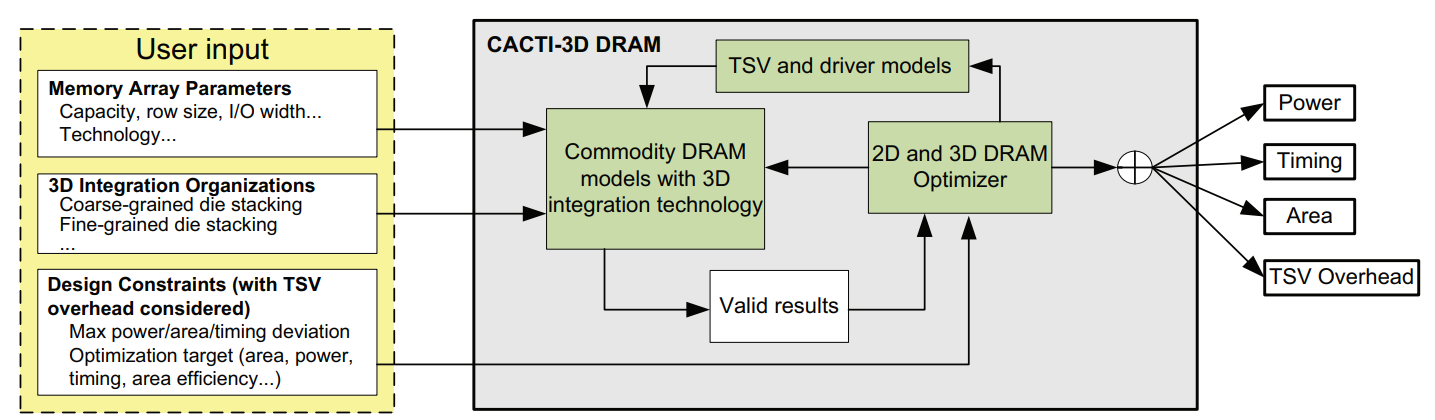
3.1 Motivation for Simulator-Based Study

As 3D-stacked DRAM technologies become attractive for their potential to significantly improve memory bandwidth and efficiency, architectural exploration of 3D-DRAM memory controllers and DRAM organizations becomes increasingly critical. To evaluate design choices at this level, a cycle-accurate simulator is essential so accurately capturing detailed interactions within the memory controller and 3D-DRAM layers can be done. To overcome this, my work focuses on developing an architectural simulator specifically for 3D-DRAM memory controllers and DRAM devices. Such a simulator is vital for fast and agile design space exploration, enabling faster design space exploration of 3D-DRAM Architecture and its correseponding memory system.

3.2 Toolchain Overview: CACTI-3DD, Ramulator2 & 3D-DRAM System Establishment Flow

3.2.1 CACTI-3DD 3D Die-stacked DRAM Architectural Modeling Tool

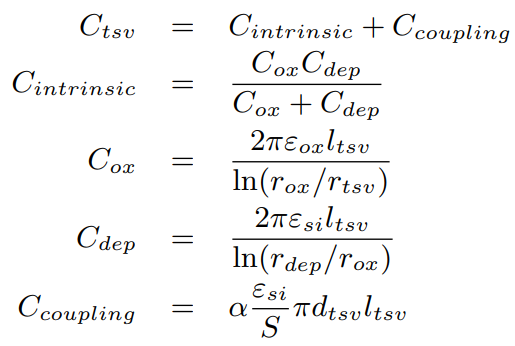
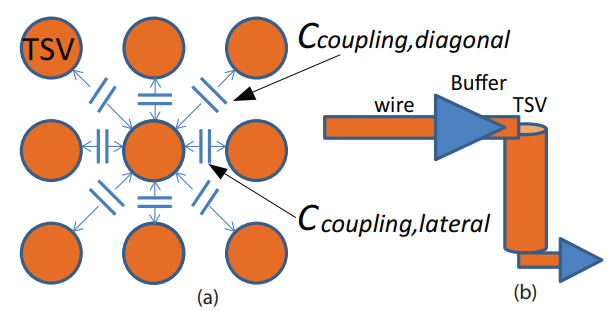
CACTI-3DD is an architecture-level modeling tool specifically developed for 3D die-stacked DRAM memory systems. Developed by HP Labs and extended from the well-established CACTI-7 framework originally created by Norman Jouppi et.al, CACTI-3DD is tailored to capture the unique characteristics of 3D-DRAM. It places particular emphasis on through-silicon via (TSV) modeling and the architectural implications of vertical die stacking. By incorporating detailed models of TSVs and their drivers, CACTI-3DD enables accurate analysis of area, power, and timing in 3D-DRAM designs.



[]

It integrates detailed models of TSVs and their drivers into the traditional planar DRAM modeling flow, enabling accurate analysis of 3D-stacked memory configurations. Given input parameters such as technology node, die and row sizing, and other key 3D-DRAM characteristics, CACTI-3DD outputs critical design metrics including area, power, and timing.

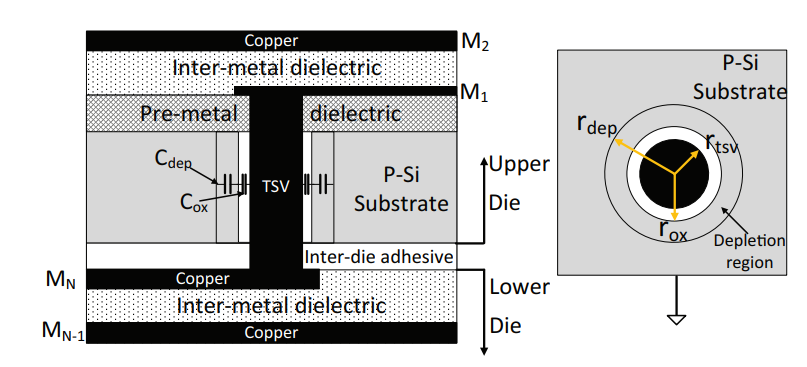
Furthermore, the tool supports the exploration of various 3D integration organizations, allowing users to evaluate different stacking architectures and their impact on system-level constraints. This makes CACTI-3DD an essential tool for early-stage design space exploration of 3D-DRAM systems.



[]

To enable accurate modeling of 3D-stacked DRAM architectures, CACTI-3DD is integrated into the simulation toolchain to generate key design specifications for the target 3D-DRAM system. As through-silicon vias (TSVs) are critical enablers of 3D die stacking, CACTI-3DD provides detailed and analytical models of TSV-based vertical interconnects.

Specifically, it accounts for TSV resistance and capacitance, and estimates the resulting power consumption, area overhead, and timing characteristics, including both TSVs and their associated driver circuits. CACTI-3DD models "via-first" TSVs, which are fabricated prior to front-end-of-line (FEOL) silicon processing.

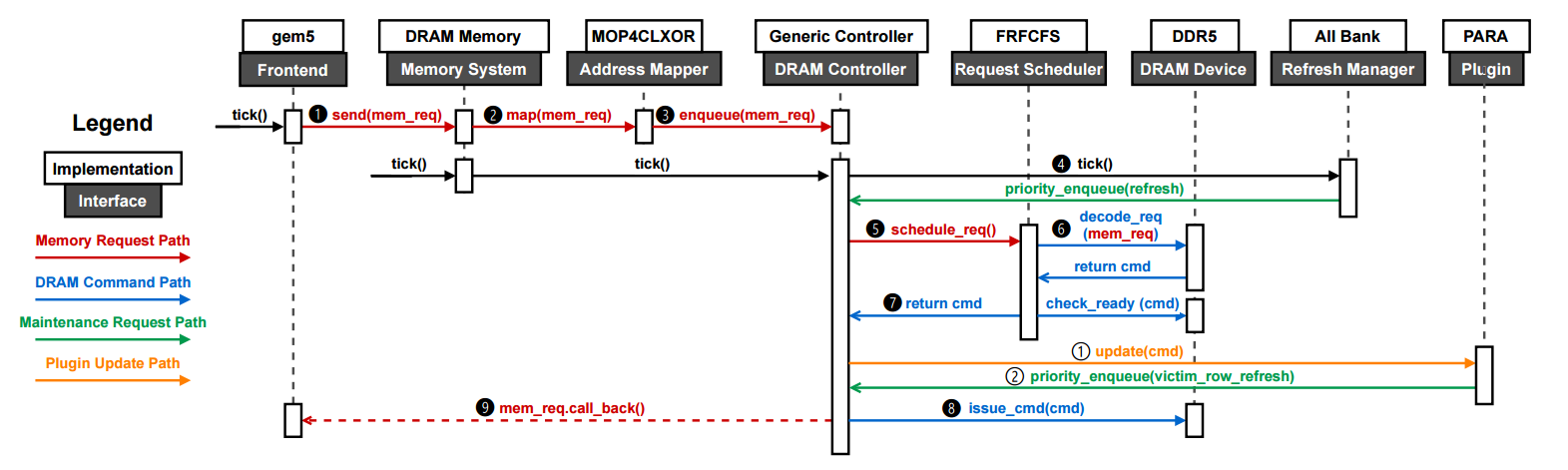


[]

The geometry and technology parameters used are based on International Technology Roadmap for Semiconductors (ITRS) projections, encompassing both global interconnect-level TSVs—larger, sparser TSVs that span more dies—and intermediate interconnect-level TSVs—smaller, denser TSVs traversing fewer dies. This level of detail is essential for capturing the performance and energy implications of 3D-DRAM integration early in the design phase.

3.2.2 Ramulator2 DRAM System Architectural Simulator

**Ramulator2** is a highly extensible architectural DRAM simulator developed by the SAFARI Research Group. It is designed to support rapid prototyping and evaluation of memory system designs, enabling researchers and system architects to explore and implement intrusive architectural changes across the entire DRAM subsystem. Ramulator2 follows a modular design approach, separating the memory system into clearly defined interfaces and components. This modularity allows for flexible integration and fast experimentation with various memory controller and DRAM design schemes.

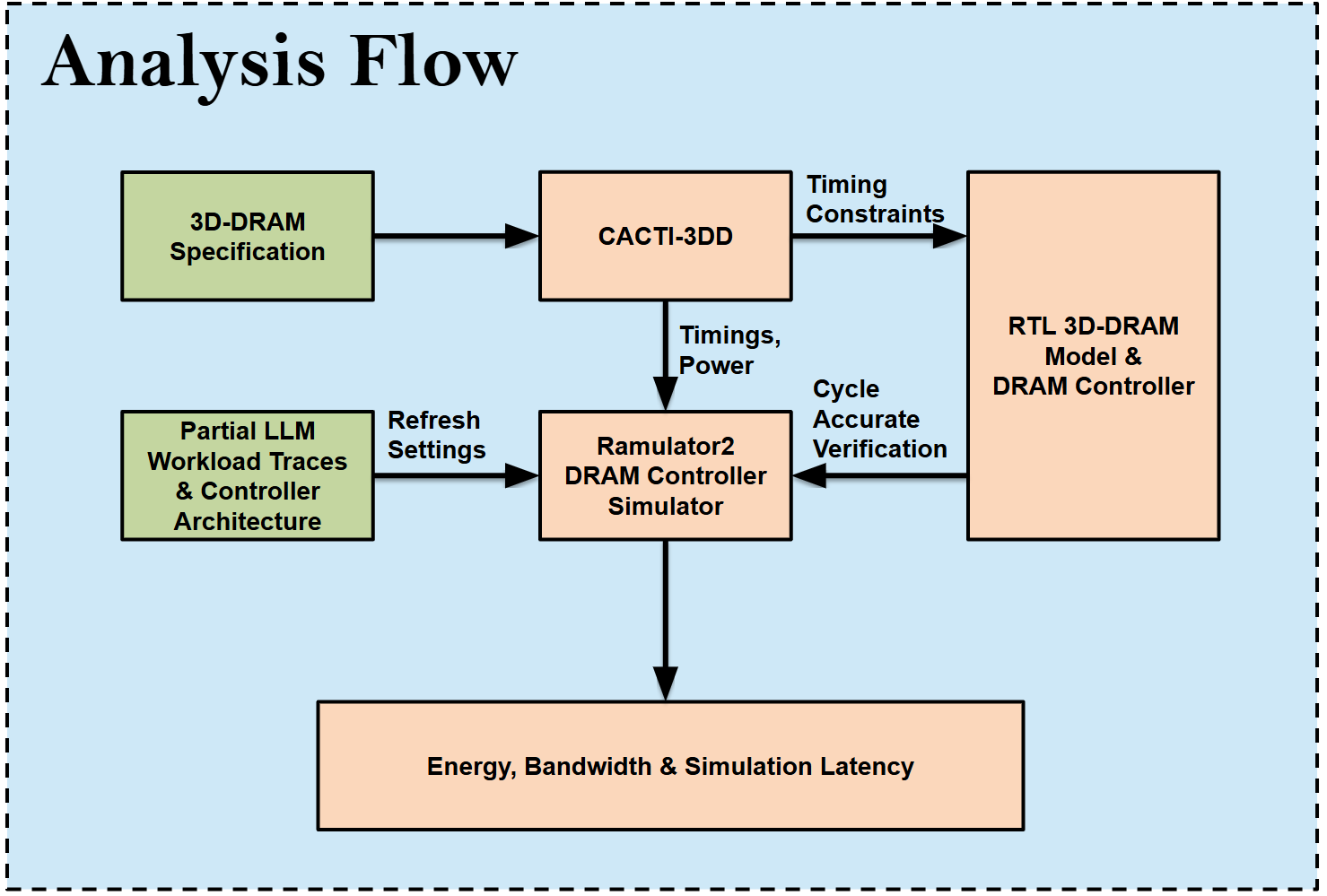


[]Sequence Diagram of Ramulator2 given as an Example of DDR5

Written in C++20, Ramulator2 can function as a standalone simulator using memory traces or serve as a memory system library embedded within larger system-level simulators. It maintains high simulation accuracy while offering faster performance compared to traditional cycle-accurate DRAM simulators. The tool has been rigorously validated and is widely used for architectural exploration of next-generation DRAM technologies.

3.2.3 3D-DRAM System Simulator & Analysis Flow

To support rigorous architectural exploration of user-defined 3D-DRAM designs, a complete analysis flow is established using modified versions of CACTI-3DD and Ramulator2. This flow[] evaluates system-level impacts including bandwidth, energy, and latency based on frontend workload traces and 3D-DRAM specifications. CACTI-3DD generates timing and power characteristics, which are fed into Ramulator2 for cycle-accurate simulation of the DRAM controller. To ensure fidelity, the architecture-level model is validated against an RTL 3D-DRAM model and memory controller, enabling accurate design space exploration and optimization prior to implementation.



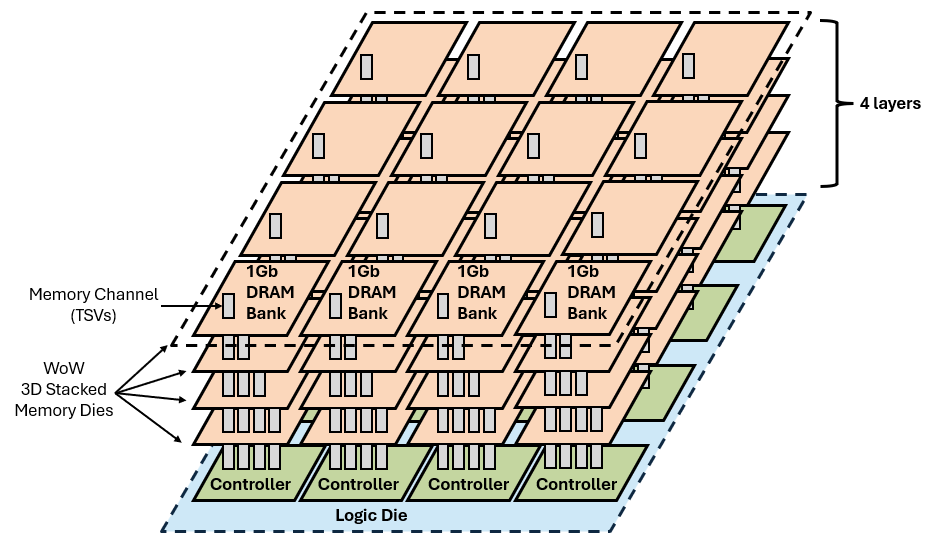
[] 3D-DRAM System Analysis Flow

3.3 Establishment of 3D-Stacked DRAM Model

In the development of a 3D-stacked DRAM memory system, it is essential to construct a detailed and cycle-accurate RTL model to support rigorous simulation and validation. A key requirement in this process is defining the critical timing parameters that govern the behavior of the 3D-DRAM architecture. These timing constraints—such as activation, precharge, and refresh latencies—are architecture-dependent and must be accurately derived to ensure realistic modeling. To this end, **CACTI-3DD** is employed to generate the necessary timing, power, and area specifications based on the target 3D-DRAM configuration.

To validate and support the architectural-level simulator, an RTL model of the estimated 3D-DRAM is developed. This model ensures cycle-level accuracy and provides a ground truth for verifying timing behavior and controller interactions. The 3D-DRAM model is constructed by adapting and extending an existing RTL model of **Micron DDR3 SDRAM**, modifying its internal bank logic and structural organization to reflect the characteristics of the target 3D-stacked architecture. The resulting RTL model serves as a foundational component for accurate simulation and further hardware implementation.

3.3.1 3D-DRAM Architecture & DRAM Logic Die Modification

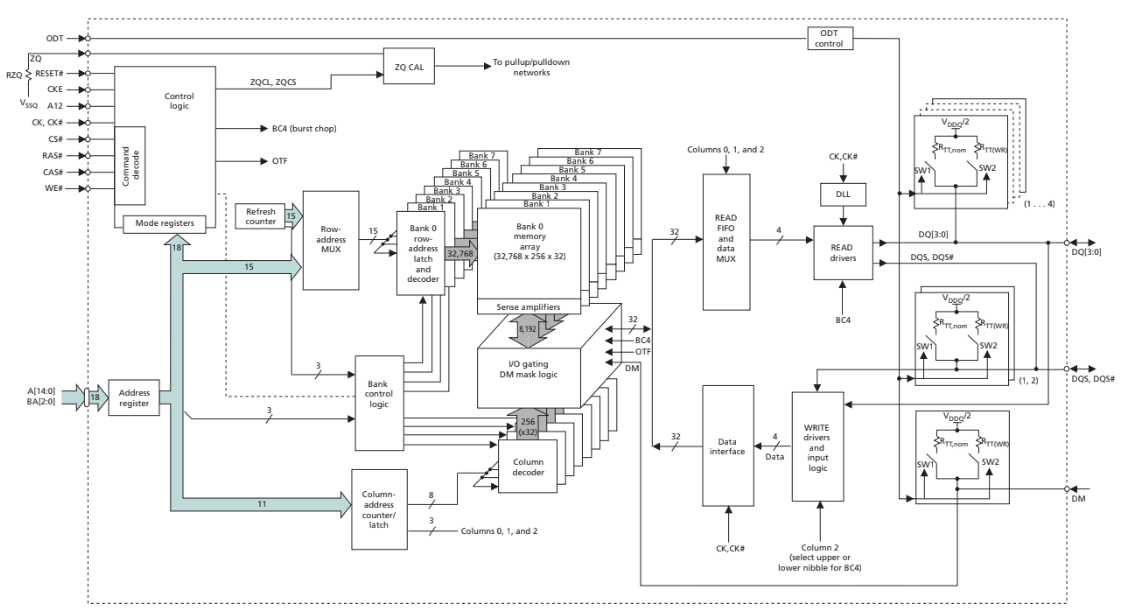


[] The Estimated Very High Bandwidth 3D-DRAM Architecture

To establish the 3D-DRAM model and fully leverage the high bandwidth and low energy consumption benefits of through-silicon vias (TSVs) for vertical data transmission, we employ Wafer-on-Wafer (WoW) stacking technology. This approach enables multiple DRAM memory dies to be vertically integrated, significantly increasing data transfer capability between layers. In this architecture, each DRAM bank has a dedicated TSV connection directly linked to the underlying memory controller on the logic die. This direct vertical interconnect forms the foundation of a very high-bandwidth(VHB) 3D-DRAM system.

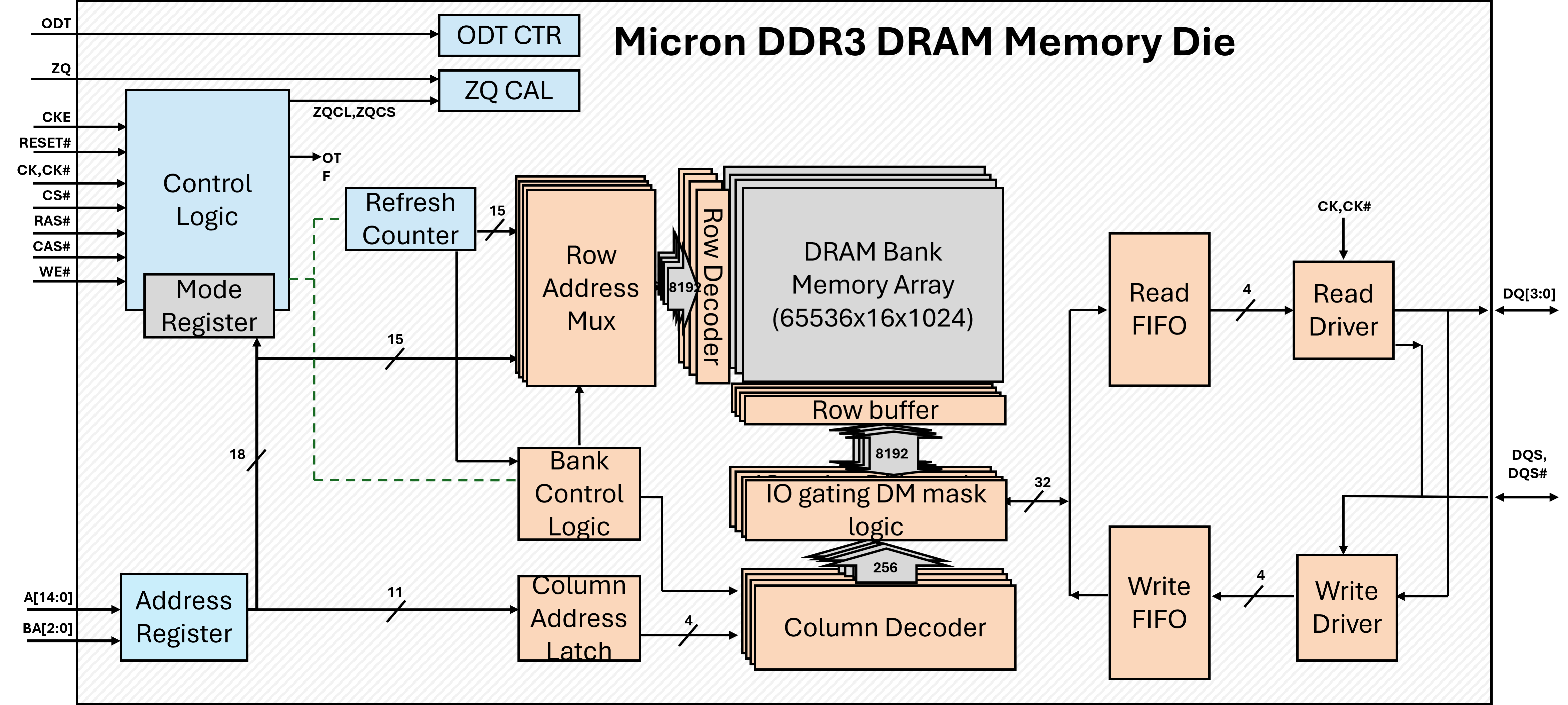
For modeling purposes, we first design and evaluate a single vertical vault—a stack of DRAM banks connected through TSVs to a dedicated controller. The validated vault model is then extended and replicated to form the complete 4×4 DRAM bank configuration shown in the figure, representing the full 3D-stacked memory system.

3.3.2 3D-DRAM Vertical Vault & Bank Modeling



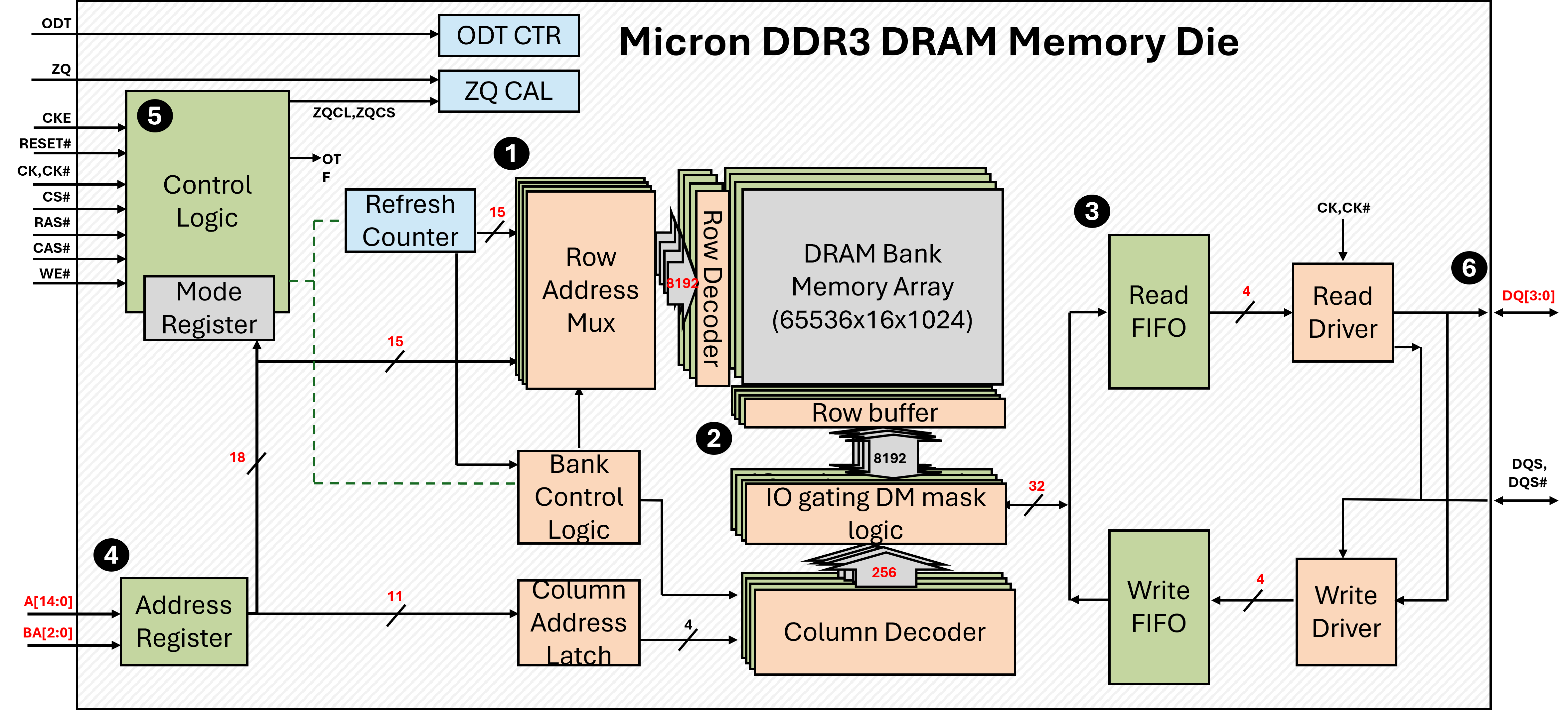
[] The Micron DDR3 Model

To model a single vertical vault of a 3D-DRAM system, the design process begins with creating a single-bank 3D-DRAM model derived from the Micron DDR3 SDRAM RTL architecture []. The original multi-bank structure, along with its associated control logic, is simplified by removing all but one bank, retaining only the essential components required for operation—such as address decoding, sense amplifiers, I/O gating, refresh logic, and read/write interfaces.



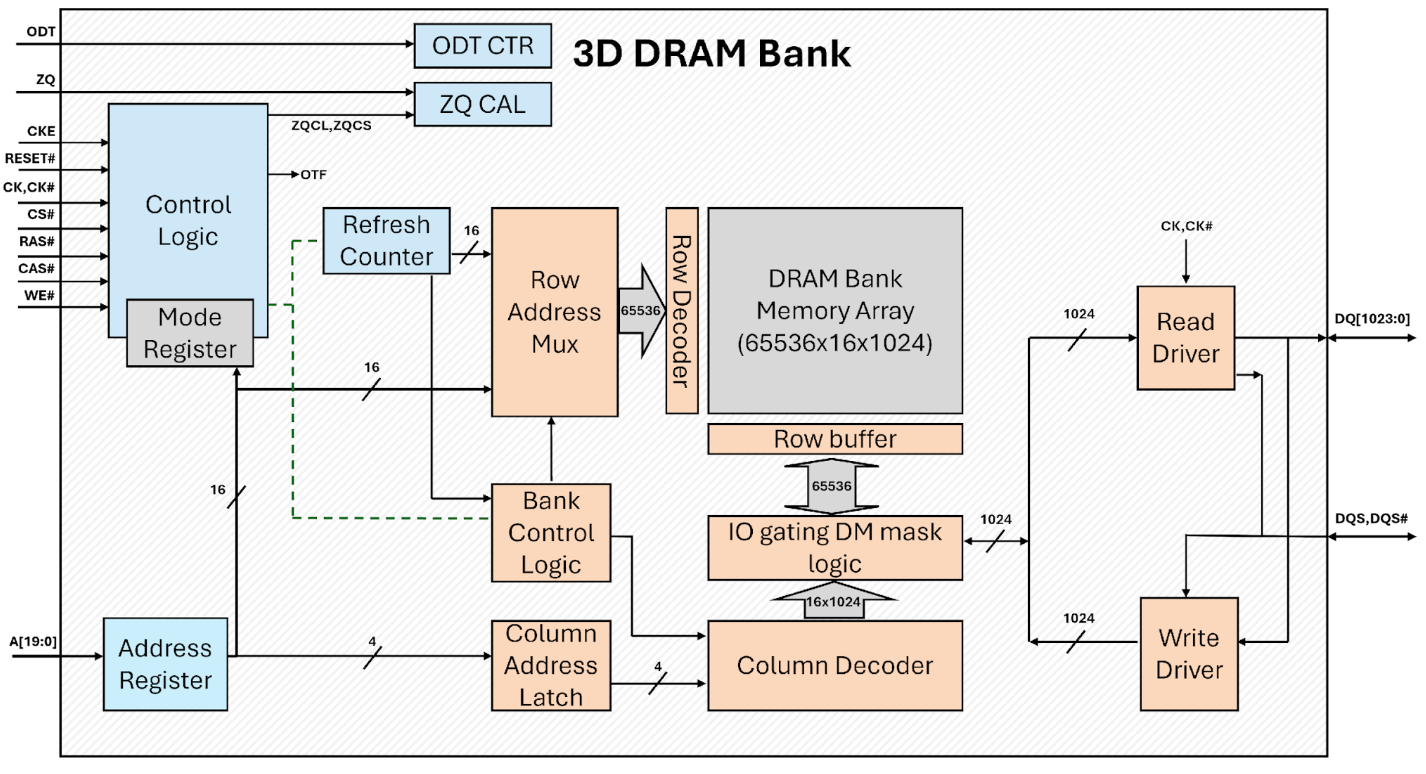
[] Simplified Micron DDR3 Memory Die Block Diagram

This single-bank configuration can then be easily integrated into a vertical vault, where multiple such banks are stacked and interconnected through Through-Silicon Vias (TSVs) to a dedicated vault controller on the logic die. The procedures of modification of a Micron DDR3 DRAM Memory Die to the 3D-DRAM bank Die will be illustrated.



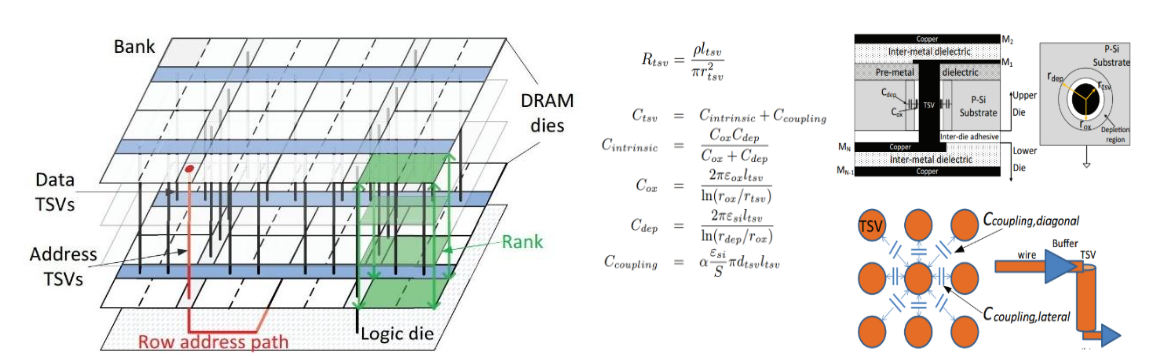
The modification includes: (1) the removal of additional banks, since we only want to start with a single bank configuration. (2) Removing the additional row buffers, IO gating logic and modification of Row buffer sizes. (3) Removing of the read fifo, since original the DDR3 memory reads 32bits data and store it within the read fifo for an 8 cycles bursts due to the pin limits. However; in the 3D-DRAM stacked design, high bandwidth TSV does not have the pin limits problem, thus we remove the fifos away

(4) Modification of the address registers such that the address decoder of the bank can match the correct specification. (5) Modification of control logic such that burst chop related command can be removed. (6) Changing the datapath of Read & Write IO to 1024bits to match the requirement.



[] The modified 3D-DRAM Bank Die from Micron DDR3 Die

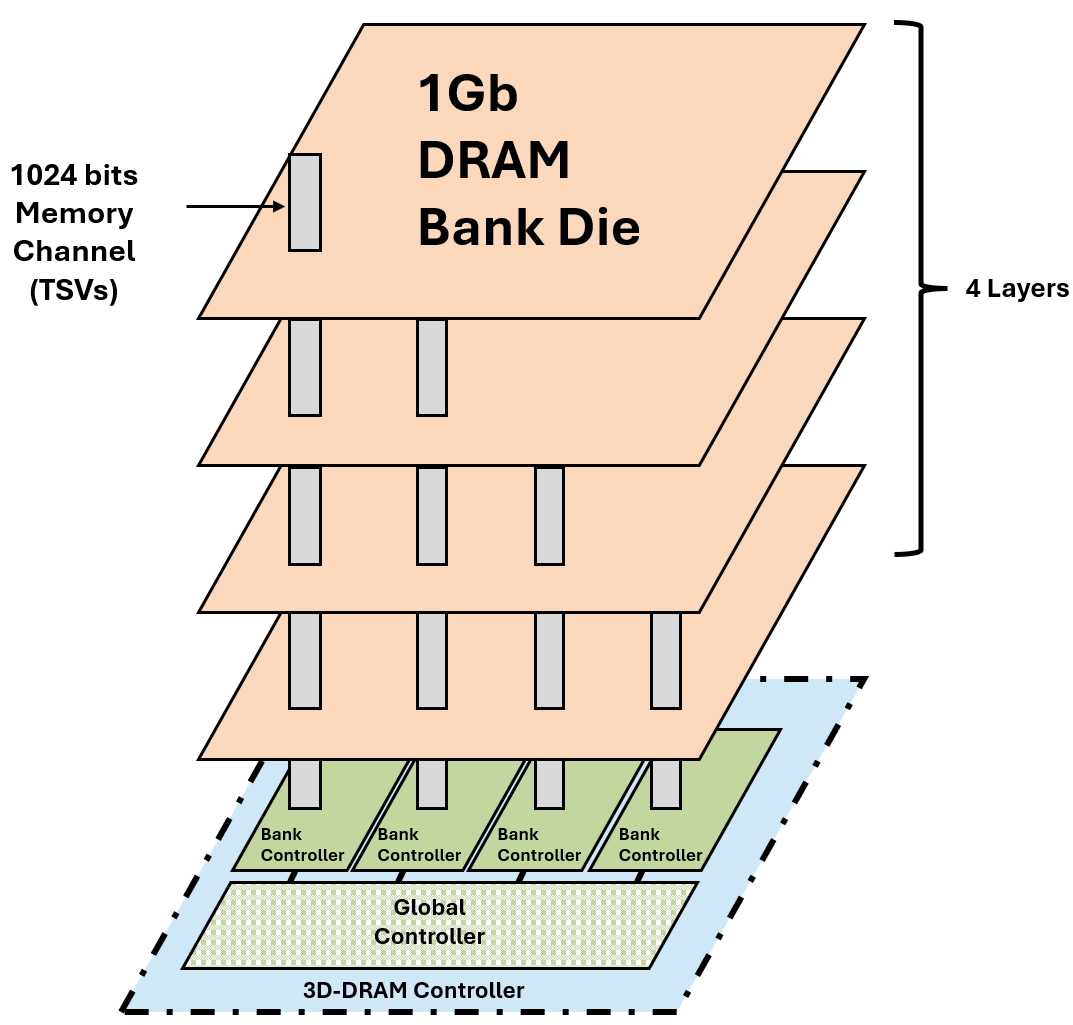
Using ITRS process parameters as the design foundation and the RC circuit model embedded within CACTI-3DD, the critical timing constraints of stacked 3D-DRAM banks are identified. To achieve a more accurate delay and power estimation, the geometric parameters of intermediate-level TSVs are derived from ITRS projections, enabling detailed analysis of TSV delay characteristics. Furthermore, a fine-grained data mapping technique is employed to increase the effective bandwidth of the stacked 3D-DRAM, thereby optimizing data access efficiency and overall system performance.



[] Fine-grained data mapping & Via First TSV Modeling

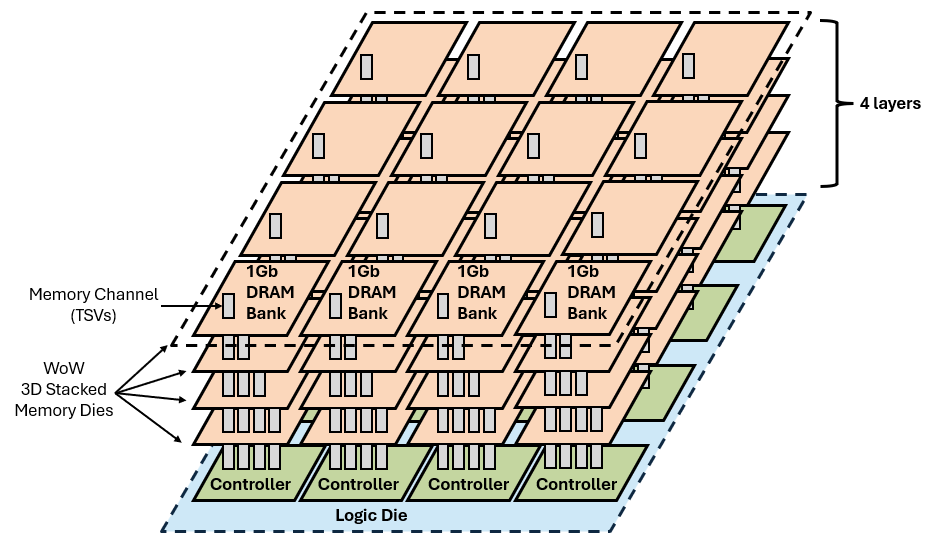
Once the single-bank model is functionally validated, it is replicated vertically to form the complete vault structure, which can be scaled up to the full 4×4 bank 3D-DRAM system. This validated RTL model then serves as the golden reference for developing and verifying the cycle-accurate architectural simulator used in broader system-level performance and energy evaluations.

3.4 3D-DRAM Architecture & Specification



[] The Proposed Vertical Vault of the 3D-DRAM Slice

The target 3D-DRAM architecture is derived by adapting the Micron DDR3 behavioral model and substituting its native timing constraints and operational characteristics with those obtained from CACTI-3DD analysis. Using these updated parameters, a 3D-DRAM slice is synthesized by vertically integrating multiple bank-level models, enabling an accurate representation of the stacked memory hierarchy and its performance characteristics.



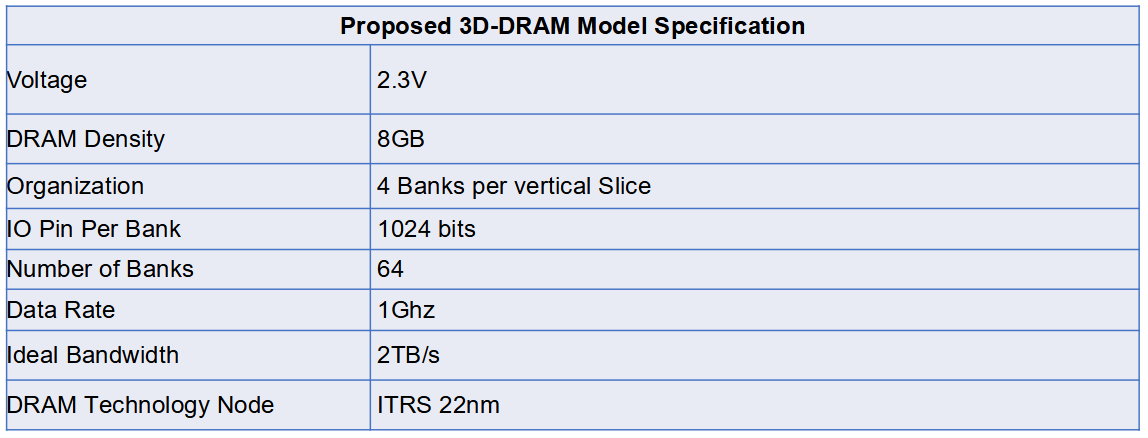
[] Proposed 3D-DRAM Architecture

The specification of the proposed 3D-stacked DRAM [†] is summarized in Table []. The architecture consists of 16 vertical vaults, each directly connected to a dedicated DRAM controller located on the logic die layer through Through-Silicon Vias (TSVs). This configuration enables extremely high memory bandwidth by providing a wide and direct data path between the stacked memory layers and the controller.

Each vault features 4 × 1024-bit TSV I/O channels, delivering a highly parallel data interface between the DRAM banks and the controller. Within each vault, the DRAM array is organized into multiple banks, with each bank having a 1 Gb capacity. Across all vaults, the system comprises 64 banks in total, resulting in an overall memory capacity of 8 GB which is sufficient to store the required model weights for targeted workloads such as large-scale AI inference.

Operating at a clock frequency of 1 GHz, the aggregated TSV bandwidth achieves an ideal peak throughput of 2 TB/s, enabling the memory system to sustain data-intensive operations with minimal latency.

This combination of high capacity, ultra-wide I/O, and low-latency vertical interconnect makes the proposed 3D-DRAM architecture highly suitable for bandwidth-bound workloads and next-generation high-performance computing systems.

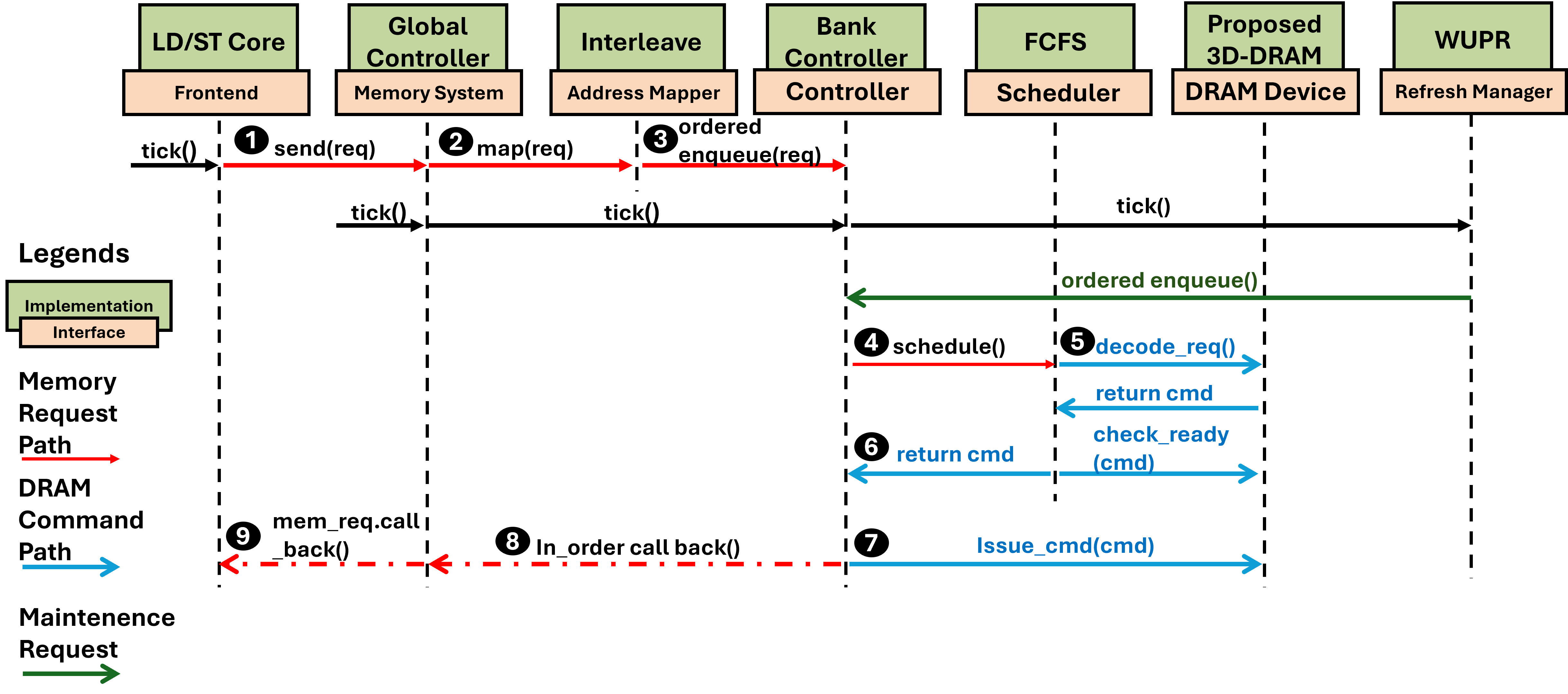


[]Table

3.5 Architectural Simulator For 3D-DRAM Memory System

To enable rapid design space exploration and facilitate tight integration with the upper-level accelerator core architectural simulator, a cycle-accurate 3D-DRAM memory system architecutral Simulator is implemented in C++ on top of the Ramulator2 framework. The figure below presents the sequence diagram of the designed architectural simulator.

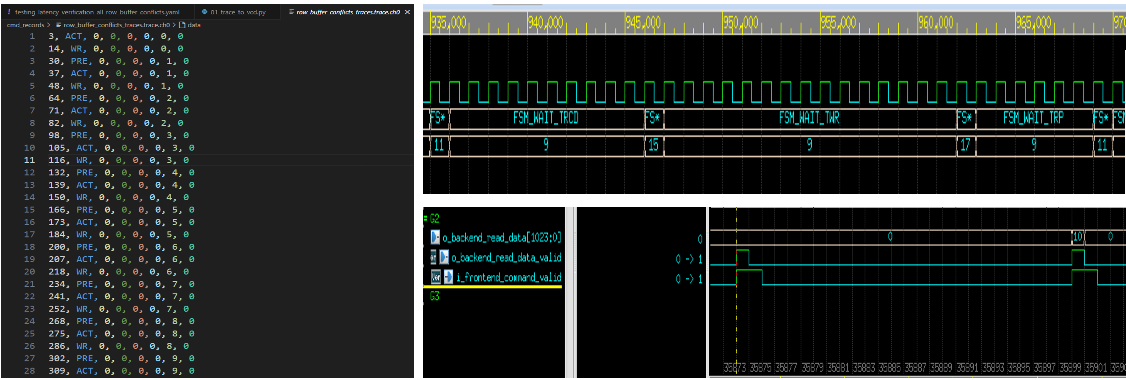
By extending Ramulator2, the proposed 3D-DRAM model incorporates its specific timing constraints, access behaviors, energy model, I/O width, refresh latencies, and refresh intervals into the simulator’s configuration options. The resulting 3D-DRAM memory system architecture can be easily integrated into, and modified within the system-level 3D-DRAM architectural simulator.



[] Sequence Diagram of the 3D-DRAM Architectural Simulator

Furthermore, a load–store (LD–ST) trace mechanism is introduced to more accurately capture the access patterns of the upper accelerator core. C++ based models of the Global Controller and Bank Controller are also implemented as an architectural model of the intended RTL model to support and exploit the distributed features of the proposed 3D-DRAM architecture is explained in Chapter 4. Write Updated Partial Refresh is also implemented on the architectural simulator to perform design space exploration would be further explained in Chapter 5.

Cycle-accurate verification is conducted by generating the cycle accurate traces from C++ model then compare the waveform genreated by the RTL 3D-DRAM system for cycle accurate behaviour ensuring that the simulated behavior of the C++ architectural model aligns with that of the RTL 3D-DRAM memory system. Events queue and fine-tuning of the timing behaviour is needed to simulate the additional latency taken by FSM state transition and fifo delay within the RTL model.



[]Generated Traces & the Waveform Verification Process

This alignment enables fine-grained tuning and performance analysis to be performed on the architectural simulator before committing design choices to RTL implementation. Such a workflow significantly reduces both design time and simulation overhead, ensuring that informed trade-offs and optimizations can be made early in the development process.