**1. Introduction & Motivation**

**1.1 Motivation**

The rapid advancement of large language models (LLMs) and diffusion-based generative models has brought significant breakthroughs in AI, but also introduced massive demands on memory bandwidth, capacity, and energy efficiency. Traditional 2D-DRAM systems are increasingly unable to meet these requirements due to limited bandwidth, high access energy, and refresh overhead.

With the emergence of 3D-stacked DRAM—which offers higher bandwidth and lower access energy through vertical silicon integration—a promising opportunity arises for accelerating AI workloads. Moving DRAM closer to the processing logic, or even integrating it on-chip, helps eliminate I/O pin limitations and significantly boosts memory performance. 3D systems with on-chip DRAM are viewed as a key solution to overcome memory scaling limits and unlock higher throughput. However, this vertical stacking fundamentally changes the interaction between memory and logic, requiring a rethinking of memory interface design.

In particular, refresh becomes an even more critical challenge in 3D-DRAM due to higher cell density and thermal constraints, leading to increased performance and energy penalties. This thesis proposes a memory management unit (MMU) designed specifically for 3D-stacked DRAM, addressing key challenges such as address mapping, access coordination, and refresh overhead. The proposed design aims to fully utilize the advantages of 3D on-chip memory and improve system efficiency for next-generation AI accelerators.

**1.2 Research Goals & Major Contributions**

This thesis makes the following key contributions to the design and optimization of memory systems for AI workloads using 3D-stacked DRAM:

1. Establishment of a 3D-Stacked DRAM Model & the architectural Simulator For this 3D-DRAM system

A 3D-stacked DRAM architecture is developed along with a corresponding cycle-accurate simulator. The model captures the hierarchical structure, internal bank organization, and Through-Silicon Via (TSV) timing characteristics, providing a foundation for evaluating memory access behavior, controller policies, and refresh mechanisms.

1. Design of a 3D-Stacked DRAM Memory Controller  
   A novel memory controller is proposed to exploit the vertical and hierarchical organization of 3D-stacked DRAM. The architecture consists of a global controller that interfaces with the accelerator core and multiple bank-level controllers that manage access to individual DRAM banks through Through-Silicon Via (TSV) interconnects. This layered control structure enables efficient address mapping and bank-level coordination, improving bandwidth and reducing access latency.
2. Write Updated Partial Refresh (WUPR) Scheme  
   A refresh reduction mechanism, Write Updated Partial Refresh (WUPR), is proposed to mitigate the high refresh overhead in 3D-stacked DRAM. WUPR selectively refreshes only the memory rows that have been recently written, while safely skipping unwritten rows with minimal overhead. This approach significantly reduces refresh energy and improves memory availability, without compromising data reliability.

**1.3 Thesis Organization**

The organization of this thesis is as follows:

Chapter 2 reviews related research on DRAM Architecture and memory system architectures, establishing the background for 3D-stacked DRAM System

Chapter 3 presents the design flow for constructing a bank-level 3D-DRAM timing model based on the Micron DDR3 specification. It also details the development of a cycle-accurate architectural simulator that incorporates timing constraints derived from this model. This simulator serves as a foundation for fast design space exploration of 3D DRAM systems.

Chapter 4 describes the design of a hierarchical 3D-stacked DRAM controller. The controller consists of bank-level controllers for managing vertical banks and a global controller responsible for address mapping and coordination among bank controllers.

Chapter 5 introduces the Write Updated Partial Refresh (WUPR) scheme, a low-overhead refresh mechanism that selectively refreshes written rows while skipping unwritten ones. This approach alleviates the refresh bottleneck, improves DRAM bandwidth, and enhances overall efficiency.

Chapter 6 concludes the thesis with a summary of key contributions and discusses potential directions for future research.