Demystifying the Characteristics of 3D-Stacked Memories: A Case Study for the Hybrid Memory Cube (HMC)

Ramyad Hadidi, Bahar Asgari, Burhan Ahmad Mudassar, Saibal Mukhopadhyay, Sudhakar Yalamanchili, and Hyesoon Kim

IISWC’17

3D-Stacking Technology
Provides opportunities & novel features
- 3D-DRAMs:
  - Provide higher bandwidth and density
  - Enable lower power consumption
  - Motivate processing-in-memory
HMC is an example of such memories.

New Considerations
- New internal organization
- New thermal behavior
- New latency and bandwidth hierarchy
- New packet-switched interface

Hybrid Memory Cube (HMC)

HMC 1.1 (Gen2): 4GB size

Experimental Setup
Pico SC6 Mini
EX700 Backplane
ACS10 Module
4GB HMC 1.1

Experimental Setup II
FPGA frequency: 187.5 MHz

Modified GUPS (giga updates per second) benchmark
Apply different masks to addresses

Bandwidth
Accessing 4 banks saturates 1 vault bandwidth. External bandwidth is saturated at 4 vaults.

Temperature (read only)

Latency Deconstruction
Access patterns affect temperature.

Latency Deconstruction Summary

Conclusions
- Mixing read and write requests and using large request sizes lead to effective use of bi-directional bandwidth.
- Distributing accesses prevents internal bottlenecks and exploits bank-level parallelism.
- Controlling the request rate to avoid high latency.
- Employing fault-tolerant mechanisms and using proper cooling solutions enables temperature-sensitive operations to reach a higher bandwidth.
- Reducing latency overhead of the infrastructure will greatly benefit latency.