Performance Measurement and Modeling for a Custom Memory Cube Research Platform

Yu Zou1, Gongyu Wang1, Vinayak Deshpande1, Suvrat Tedia2, Herman Lam1, Alan George1

NSF Center for High-Performance Reconfigurable Computing (CHREC),

Department of Electrical and Computer Engineering,

University of Florida, Gainesville, FL

1{zou, deshpande, hlam, wangg}@chrec.org 2suvrat17@ufl.edu

# INTRODUCTION

Big data era is driving a stricter requirement for data-intensive workloads. 3D-stacked memory, such as HMC, offers researchers much higher memory bandwidth and better power efficiency to accelerate memory access. Based on base logic layer, integrated with stacked DRAMs, replacing fixed control logic with a self-defined data-processing logic gives researchers more design space to explore near-memory-processing, which moves computation from conventional computing units to storage units. Many architectures, algorithms, and memory internal connection topologies [1][2][3] have been developed towards next-generation CMC-enabled in-memory processing. However, most of existing works are based on software-simulated 3D-stacked DRAM models, such as CACTI-3DD and gem5 [4]. In this extended abstract, based on our previous work about hardware-in-the-loop CMC research platform [5], we present a methodology to measure and model performance of both HMC and custom logic. Details on the performance measurement and modeling are discussed in Section 2. In Section 3, we report our progress and current already measured performance.

# MEASUREMENT AND MODELING

## Research Platform

From the aspect of either a CMC custom logic designer or a HMC developer, a notional CMC research platform should have the ability to provide measured performance of certain critical points of the designed CMC architecture. These performance data can benefit design space exploration no matter in CMC logic design or in HMC design. As shown in Figure **1**, we defined 5 parameters (A, B, C, D, and E) that we think critical for CMC researches and 2 ancillary performance points (F and G). The definition of each parameter is explained in the below:

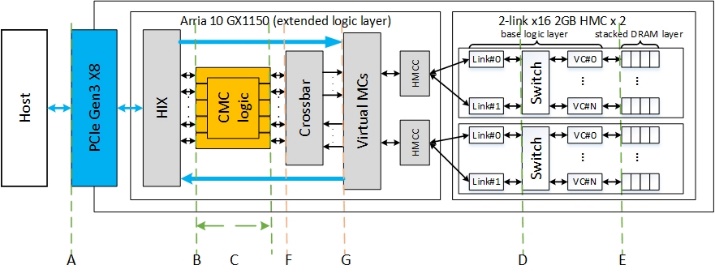


Figure 1. Concept diagram of CMC research platform

* A - round-trip CMC latency + PCIe data transfer delay
* B - round-trip CMC latency
* C - CMC custom logic latency
* D - round-trip latency between HMC switch and vault controller
* E - round-trip 3D-stacked DRAM latency
* F - round-trip latency of FPGA access to HMC
* G - round-trip latency of FPGA memory controller’s access to HMC

A, B, C parameters can help to monitor and improve the custom logic design. D and E parameters can help to monitor latencies caused by HMC access, and even to compare latencies caused by different types of 3D-stacked DRAMs in the future, not only limited to HMC.

## Performance Measurement and Modeling

In this section, we describe the methodology to provide performance parameters by combining modeling and measurement.

As shown in Figure **1**, A is measured in host C code, B, C, and F parameters are measured inside HT hardware code [5] on FPGA. G is measured in the performance monitor instantiated in infrastructure. The hardware performance monitor is modified based on Micron’s original code. D is calculated by using a mathematical model and measured performance of G point. Latency of HMCC and latency of HMC link can be found respectively in HMCC specification and HMC specification [7][8].

Since D and E are inside HMC, the delayed latency between D and E should be fixed, which means the latency of E point can be measured using current HMC simulators, such as HMC-Sim 2.0 [6], and be integrated into the research platform.

# PROGRESS AND FUTURE WORK

The work of performance modeling and measurement is still at its early stage. The progress we have made includes: (1) measurement-enabled research platform setup; (2) performance measurement for parameters A, B, C, and F. The hardware performance monitor for G is being modified. We have conducted experiments to use the platform to monitor the performance of single-memory-operation benchmarks. The results of single-memory-read and single-memory-write operations are shown in Table **1**.

Table 1. Single-memory-operation benchmarking results on the research platform

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Operations | A (ns) | B (cycles) | C (cycles) | F (cycles) |
| Single memory read | 1058559 | 168.4414 | 4 | 164.4414 |
| Single memory write | 1057302 | 175.6587 | 4 | 171.6587 |

# ACKNOWLEDGMENTS

This research was supported by CHREC members and by the I/UCRC Program at NSF under Grant No. IIP-1161022. We are grateful for the feedback, collaboration, and support provided by Laboratory of Physical Sciences, Micron, and Lawrence Livermore National Laboratory.

# REFERENCES

1. Gokhale, Maya, Scott Lloyd, and Chris Hajas. "Near memory data structure rearrangement." Proceedings of the 2015 International Symposium on Memory Systems. ACM, 2015.
2. Akin, Berkin, Franz Franchetti, and James C. Hoe. "Data reorganization in memory using 3d-stacked dram." ACM SIGARCH Computer Architecture News. Vol. 43. No. 3. ACM, 2015.
3. Ahn, Junwhan, et al. "A scalable processing-in-memory accelerator for parallel graph processing." ACM SIGARCH Computer Architecture News. Vol. 43. No. 3. ACM, 2015.
4. Farmahini-Farahani, Amin, et al. "Drama: An architecture for accelerated processing near memory." IEEE Computer Architecture Letters 14.1 (2015): 26-29.
5. Wang, Gong, et al. “A research platform for custom memory cube.” Workshop on Modeling & Simulation of Systems and Applications. University of Seattle, Seattle, WA, Aug. 2016.
6. Leidel, John D., and Yong Chen. "HMC-Sim 2.0: A Simulation Platform for Exploring Custom Memory Cube Operations." (2016).
7. Hybrid Memory Cube Controller IP Core User Guide. https://www.altera.com/content/dam/altera-www/global/en\_US/pdfs/literature/ug/ug\_hmcc.pdf
8. HMC Specification 1.0. http://hybridmemorycube.org/files/SiteDownloads/HMC\_Specification%201\_0.pdf