Performance Modeling and Measurement for a Custom Memory Cube Research Platform

Yu Zou, Gongyu Wang, Herman Lam, Vinayak Deshpande, Suvrat Tedia, Alan George

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

Department of Electrical and Computer Engineering,

University of Florida, Gainesville, FL

{zou, wangg, hlam, deshpande, suvrat17, george}@chrec.org

# INTRODUCTION

Big data era is putting an increasingly strict performance and power requirements on data-intensive workloads. The emerging 3D-stacked memory devices, such as the hybrid-memory cube (HMC), offer much higher memory bandwidth and better power efficiency, which have reignited the decade-old research on near-memory processing to integrate customized data-processing logic into the based logic layer of HMC, forming a custom memory cube (CMC), and thus moves computation from conventional computing units to storage units. Many architectures, algorithms, and memory-internal connection topologies [1][2][3] have been developed towards this next-generation CMC-enabled in-memory processing. However, existing works are based on software-simulated 3D-stacked DRAM models [2][4]. In this extended abstract, we present a methodology to model and measure performance of both HMC and custom logic, based on our previous work on hardware-in-the-loop CMC research platform [5]. Details on the performance modeling and measurement are discussed in Section 2. In Section 3, we report our progress and current already measured performance.

# MODELING AND MEASUREMENT

As shown in Figure 1, we identify 5 measurement points (A, B, C, D, and E), within the notional CMC architecture, that are critical for design-space exploration (DSE) of CMC architectures and apps. (For simplicity of explanation, we use *latency A* to represent round-trip latency of point A, and use *latency A\_B* to represent delay between point A and point B in the following.) Definitions of measurement parameters are shown in below:

* *latency A*: round-trip CMC latency + data transfer delay
* *latency B*: round-trip CMC latency
* *latency C*: CMC custom logic latency
* *latency D*: round-trip latency of HMC internal data access
* *latency E*: round-trip latency of DRAM layers

Next, we describe the methodology to provide performance parameters by combining modeling and measurement. Measurement points are mapped from the CMC architecture to the platform. Mapped measurement points are shown in Figure 2. Two ancillary measurement points (F and G) are also added. Definitions of relating parameters are shown in below:

* *latency F*: round-trip latency of FPGA’s access to HMC
* *latency G*: round-trip latency of FPGA memory controller’s access to HMC

As shown in Figure 2, is measured in host C code. *Latency B*, *C*, and *F* are measured in HT hardware code [5] on the FPGA. *Latency G* is measured in the performance monitor instantiated inside the low-level infrastructure. The hardware performance monitor is modified based on Micron’s original code. *Latency D* is calculated by using a mathematical model, and transformed to real *latency D* on the notional architecture.

Latency of HMCC and latency of HMC link can be found respectively in HMCC specification [6] and HMC specification [7]. *Latency D\_E* can be measured using current DRAM simulators and be integrated into the research platform.

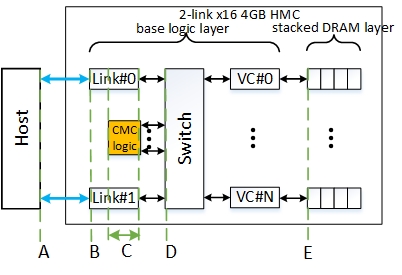


Figure 1. Concept diagram of notional CMC architecture

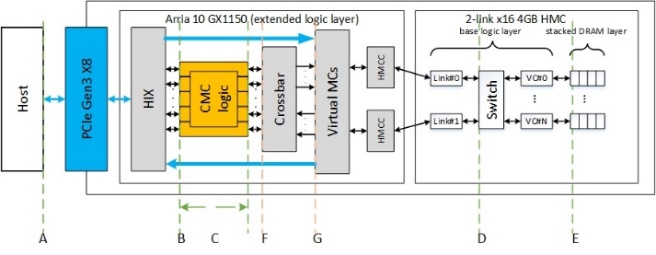


Figure 2. Concept diagram of CMC research platform

# 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 points A, B, C, and F. The hardware performance monitor for *latency 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. In the future, we will continue the performance measurement work and use the research platform to test different CMC logics.

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Operations | A | B | C | F |
| Single memory write | 1058559 | 168.4414 | 4 | 164.4414 |
| Single memory write | 1057302 | 175.6587 | 4 | 171.6587 |

\* A – ns; B – cycles; C – cycles; F - cycles

# 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, Gongyu, et al. “A research platform for custom memory cube.” Workshop on Modeling & Simulation of Systems and Applications. University of Washington, Seattle, WA, Aug. 2016.
6. 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
7. HMC Specification 1.0. http://hybridmemorycube.org/files/SiteDownloads/HMC\_Specification%201\_0.pdf