Shared DRAMCache Management for Integrated Heterogeneous Systems

ABSTRACT

This document is intended to serve as a sample for submissions to ISCA 2016. We provide some guidelines that authors should follow when submitting papers to the conference. In an effort to respect the efforts of reviewers and in the interest of fairness to all prospective authors, we request that all submissions follow the formatting and submission rules detailed below.

1. INTRODUCTION

The remarkable advances in computing power of the modern microprocessor over the last few decades can predominantly be attributed to Moore's Law [] and advances in manufacturing technology that has allowed shrinking of transistor sizes. Miniaturization of transistors has allowed addition of specialized on-chip hardware circuitry for acceleration units. Koomey et. al in 2010 found that the amount of computation that could be done per unit of energy doubled about every 18 months. However to reach exascale and beyond requires a thousand fold decrease in energy consumed per flop computed. Graphics processing units (GPUs) have transformed from fixed function hardware to a far more general compute platform. Compared with multi-core CPUs, GPGPU computing offers the potential for better performance at lower energy. Traditionally these processors have had their own independent memory systems. To take advantage of GPUs, the CPU must copy data to the GPUs memory. This data movement is inefficient and adds to energy expended and added latency as the transfer happens over slower PCIe bus. Complex programming models further impede expansion of the workloads that benefit from GPGPU computing. Therefore, processor manufacturers including AMD, Intel[1], and NVIDIA beginning to integrate CPUs and GPUs on the same silicon chip thus yielding better efficiency by sharing memory interface, power delivery and cooling infrastructure. Secondly this provides a shared virtual allowing pointer sharing semantics possible to be deferenced on CPU and the GPU which simplified programming. Further shared physical address space reduces GPU initialization time and enables several high level languages to also take advantage of the parallel processing in concert with the CPU.

Parallely, DRAM memory speeds have not kept pace commensurate to CPU speeds and this coupled with a limited growth in pin counts has led to memory and bandwidth wall for off-chip DRAM systems. The advent of die-stacking technology [7] provides a way to integrate disparate silicon die of NMOS DRAM chips and CMOS logic chips with better interconnects. The implementation is accomplished either by 3D vertical stacking of DRAM chips using through-silicon vias (TSV) interconnects or horizontally/2.5D stacking on a interposer chip as depicted in Figure 1. This allows the addition of a sizable DRAM chip close to processing cores. The onchip DRAM memory can provide anywhere from a couple of hundreds of megabytes to a few gigabytes of storage at high bandwidths of 400GB/s compared to the 90GB/s of DDR4 bandwidth. The better interconnect also lowers the latency of access by around 20-25% compared to off-chip memory. This on-chip DRAM capacity has been advocated to be used as a large last level cache which is transparent to software in several works in literature. In this context, throughput oriented GPUs with high MLP and bandwidth requirement can benefit from the high bandwidth capabilities, meanwhile latency sensitive CPUs applications can benefit from reduced latency of data access from the DRAM Cache thus improving the overall system performance. The stacked DRAM Cache also reduces energy consumed per access for the overall system.

However, this introduces complexity in managing shared system resources, which we mitigate with

2. MOTIVATION

Papers must be submitted in printable PDF format and should contain a maximum of 11 pages of single-spaced, two-column text, not including references. You may include any number of pages for references, but see below for more instructions.

- If you are using LaTeX [2]to typeset your paper, then we strongly recommend that you use the template provided. **Please set your document to Times font (`\usepackage{mathptmx}`) and do not play with interline spacing**.
- If you are using a different software package to typeset your paper, then please adhere to the guidelines mentioned in the table below. **You must use 10pt Times font or larger**.

Please ensure that you include page numbers with your submission. This makes it easier for the reviewers to refer to

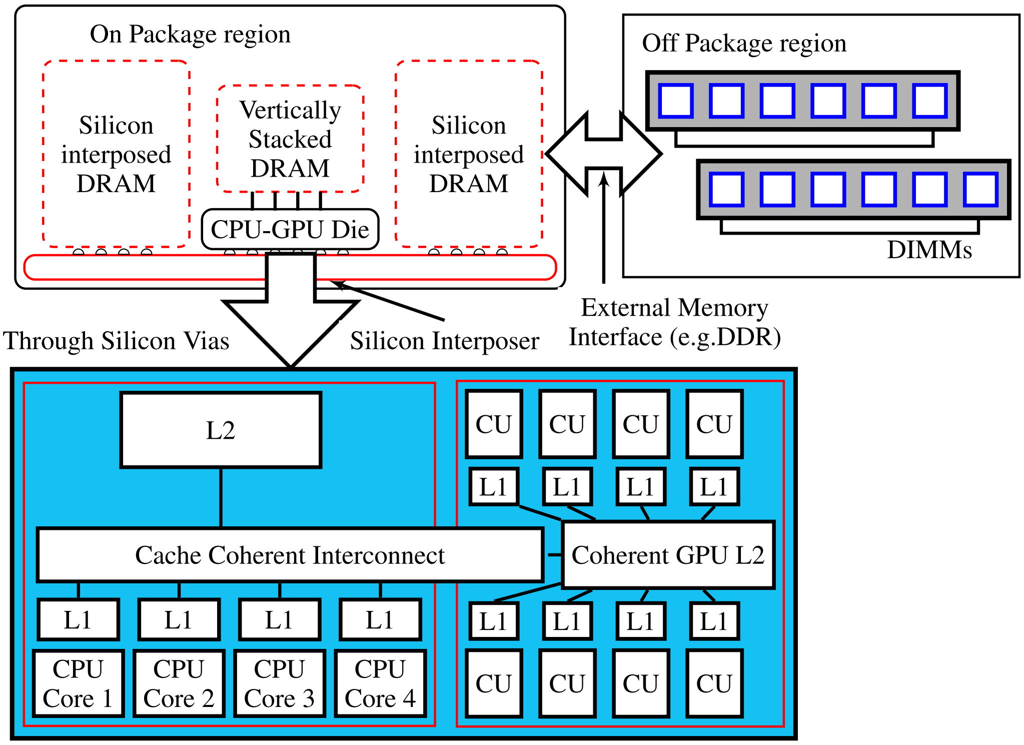

Figure 1: Architecture of a Integrated Heterogeneous System

Table 1: Formatting guidelines.

Field	Value
Page limit	11 pages w/o refs.
Paper size	US Letter 8.5x11in
Top margin	1in
Bottom margin	1in
Left margin	0.75in
Right margin	0.75in
Body	2-col., single-spaced
Separation between columns	0.25in
Body font	10pt Times
Abstract font	10pt Times
Section heading font	12pt, bold
Subsection heading font	10pt, bold
Caption font	9pt (minimum), bold
References	8pt, no page limit

different parts of your paper when they provide comments. Please ensure that your submission has a banner at the top of the title page which contains the submission number and a notice of confidentiality.

3. CHAINING MECHANISM

3.1 Author List

Reviewing will be double-blind; therefore, please do not include any author names on any submitted documents except in the space provided on the submission form. You must also ensure that the metadata included in the PDF does not give away the authors. If you are improving upon your prior work, refer to your prior work in the third person and include a full citation for the work in the bibliography. For example, if you are building on your own prior work in the papers [**?**, **?**], you would say something like: “While prior work did X, Y, and Z [3, 4], this paper additionally does W, and is therefore much better.” Do not omit or anonymize references for blind review. There is one exception to this for your own prior work that appeared in IEEE CAL, workshops without archived proceedings, etc., as discussed later in this document.

Recall that, per IEEE authorship guidelines, it is not acceptable to award honorary authorship or gift authorship. Please keep these guidelines in mind while determining the author list of your paper. Also please note that addition/removal of authors once the paper is accepted will have to be approved by the Program Chair.

3.2 Figures and Tables

Ensure that the figures and tables are legible. Please also ensure that you refer to your figures in the main text. Many reviewers print the papers in gray-scale. Therefore, if you use colors for your figures, ensure that the different colors are highly distinguishable in gray-scale.

4. EXPERIMENTAL SETUP & METHODOLOGY

We evaluate the performance of chaining using multi-programmed SPEC 2006 applications coupled with a Rodinia application that contains the GPU phase of execution. We use Rodinia [5] applications that are modified to elide the memcpy api calls to run with unified virtual and physical address spaces. These workloads are run on a cycle accurate simulator gem5-gpu [4] which is configured to simulate cache coherent unified CPUs and GPUs using the VI_Hammer protocol. The cache hierarchy has per SM private GPU L1 that are non-inclusive of the shared GPU L2 cache and can hold stale data. However, GPU L2 cache is coherent with all levels of the CPU hierarchy. The DRAM Cache we evaluate here is the first level shared cache between the 2 split cache hierarchies of CPUs and GPUs while they have a shared level of cache within themselves. We fast-forward the initialization phase of the workloads up until just before the launch of the first kernel of the GPU program. We ensure that each core executes atleast 2 Billion instructions and each 4 core workload executes 20 billion instructions and each 16 core workload executes x Billion on average in the fast-forward phase. We do this by letting adding no-ops to the Rodinia benchmarks for the duration until the initialization of the SPEC programs is complete.

5. RESULTS

Authors must register all their conflicts on the paper submission site. Conflicts are needed to ensure appropriate assignment of reviewers. If a paper is found to have an undeclared conflict that causes a problem OR if a paper is found to declare false conflicts in order to abuse or game the review system, the paper may be rejected.

Please declare a conflict of interest with the following people for any author of your paper. A conflict occurs in the following cases:

1. Between advisor and advisee forever.
2. Between family members forever.
3. Between people who have collaborated in the last 5 years. This collaboration can consist of a joint research or development project, a joint paper, or when there is direct funding from the potential reviewer (as opposed to company funding) to an author of the paper. Co-participation in professional activities, such as tutorials or studies, is not cause for conflict. When in doubt, the author should check with the PC chair.
4. Between people from same institution or who were in the same institution in the last 5 years.
5. Between people whose relationship prevents the reviewer from being objective in his/her assessment

"Service" collaborations, such as co-authoring a report for a professional organization, serving on a program committee, or co-presenting tutorials, do not themselves create a conflict of interest. Co-authoring a paper that is a compendium of various projects with no true collaboration among the projects does not constitute a conflict among the authors of the different projects.

We hope to draw most reviewers from the PC and the ERC, but others from the community may also write reviews. Please declare all your conflicts (not just restricted to the PC and ERC). When in doubt, contact the Program Chair.

6. RELATED WORK

By submitting a manuscript, the authors guarantee that the manuscript has not been previously published or accepted for publication in a substantially similar form in any conference, journal, or the archived proceedings of a workshop (e.g., in the ACM digital library); but see exceptions below. The authors also guarantee that no paper that contains significant overlap with the contributions of the submitted paper will be under review for any other conference or journal or an archived proceedings of a workshop during the review period. Violation of any of these conditions will lead to rejection.

The only exceptions to the above rules are for the authors' own papers in (1) workshops without archived proceedings such as in the ACM digital library (or where the authors chose not to have their paper appear in the archived proceedings), or (2) venues such as IEEE CAL where there is an explicit policy that such publication does not preclude longer conference submissions. In all such cases, the submitted manuscript may ignore the above work to preserve author anonymity. This information must, however, be provided to the Program Chair who will make this information available to reviewers if it becomes necessary to ensure a fair review. As always, if you are in doubt, it is best to contact the Program Chair.

Finally, we also note that the ACM Plagiarism Policy covers a range of ethical issues concerning the misrepresentation of other works or one's own work; please consult it carefully.

7. REFERENCES

[1] "Intel graphics opencl." `https://software.intel.com/en-us/node/540387`.

[2] L. Lamport, *LaTeX: A Document Preparation System*. Reading, Massachusetts: Addison-Wesley, 2nd ed., 1994.

[3] A. J. Cheng-Chieh Huang, Vijay Nagarajan, "Dca: a dram-cache-aware dram controller," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, November. 2016.

[4] J. Power, J. Hestness, M. Orr, M. Hill, and D. Wood, "gem5-gpu: A heterogeneous cpu-gpu simulator," *Computer Architecture Letters*, vol. 13, Jan 2014.

[5] S. Che, M. Boyer, J. Meng, D. Tarjan, J. W. Sheaffer, S.-H. Lee, and K. Skadron, "Rodinia: A benchmark suite for heterogeneous computing," in *Proceedings of the 2009 IEEE International Symposium on Workload Characterization (IISWC)*, IISWC '09, (Washington, DC, USA), pp. 44–54, IEEE Computer Society, 2009.