Thesis for The Degree of Licentiate of Engineering

High Performance Hybrid Memory Systems with 3D-stacked DRAM

Evangelos Vasilakis

Division of Computer Engineering
Department of Computer Science & Engineering
Chalmers University of Technology
Gothenburg, Sweden, 2019

High Performance Hybrid Memory Systems with 3D-stacked DRAM

Evangelos Vasilakis

Advisor: Ioannis Sourdis, Prof. at Chalmers University of Technology
Co-Advisor: Vassilios Papaefstathiou, Post Doc. researcher at FORTH-ICS
Co-Advisor: Pedro Trancoso, Prof. at Chalmers University of Technology
Examiner: Ulf Assarsson, Prof. at Chalmers University of Technology
Discussion Leader: Yale Patt, Prof. at The University of Texas at Austin

Technical Report No 197L
ISSN 1652-876X
Department of Computer Science & Engineering
Division of Computer Engineering
Chalmers University of Technology
Gothenburg, Sweden

This thesis has been prepared using LaTeX.
Printed by Chalmers Reproservice,
Gothenburg, Sweden 2019.

Abstract

The bandwidth of traditional DRAM is pin limited and so does not scale well with the increasing demand of data intensive workloads limiting performance. 3D-stacked DRAM can alleviate this problem providing substantially higher bandwidth to a processor chip. However, the capacity of 3D-stacked DRAM is not enough to replace the bulk of the memory and therefore it is used either as a DRAM cache or as part of a flat address space with support for data migration. The performance of both above alternative designs is limited by their particular overheads. In this thesis we propose designs that improve the performance of hybrid memory systems in which 3D-stacked DRAM is used either as a cache or as part of a flat address space with data migration. DRAM caches have shown excellent potential in capturing the spatial and temporal data locality of applications, however they are still far from their ideal performance. Besides the unavoidable DRAM access to fetch the requested data, tag access is in the critical path adding significant latency and energy costs. Existing approaches are not able to remove these overheads and in some cases limit DRAM cache design options. To alleviate the tag access overheads of DRAM caches this thesis proposes *Decoupled Fused Cache* (DFC), a DRAM cache design that fuses DRAM cache tags with the tags of the on-chip Last Level Cache (LLC) to access the DRAM cache data directly on LLC misses. Compared to current state-of-the-art DRAM caches, DFC improves system performance by 6% on average and by 16-18% for large cacheline sizes. Finally, DFC reduces DRAM cache traffic by 18% and DRAM cache energy consumption by 7%. Data migration schemes have significant performance potential, but also entail overheads, which may diminish migration benefits or even lead to performance degradation. These overheads are mainly due to the high cost of swapping data between memories which also makes selecting which data to migrate critical to performance. To address these challenges of data migration this thesis proposes *LLC guided Data Migration* (LGM). LGM uses the LLC to predict future reuse and select memory segments for migration. Furtermore, LGM reduces the data migration traffic overheads by not migrating the cache lines of memory segments which are present in the LLC. LGM outperforms current state-of-the art migration designs improving system performance by 12.1% and reducing memory system dynamic energy by 13.2%.

Keywords

Hybrid memory systems, 3D-stacked DRAM, DRAM caches, Data migration

Acknowledgment

A big thanks to my advisor Yiannis for putting up with me for all this time and for guiding me from the beginning of this endeavour. A big thanks also to my co-advisors, Vassilis and Pedro for their valuable help throughout my studies so far. Without all of you this work would not have been possible. Also thanks to all the good people at Chalmers for making a wonderful environment to work in.

This work is supported by the European Commission under the Horizon 2020 Program through the ECOSCALE (grant agreement 671632) and SHARCS (grant agreement 644571) projects as well as by the European Research Council (ERC) under the MECCA project (Contract No. 340328).

List of Publications

Appended publications

This thesis is based on the following publications:

[A] Evangelos Vasilakis, Vassilis Papaefstathiou, Pedro Trancoso and Ioannis Sourdis
"Decoupled Fused Cache: Fusing a Decoupled LLC with a DRAM Cache"
ACM Transactions on Architecture and Code Optimization (TACO), January 2019.

[B] Evangelos Vasilakis, Vassilis Papaefstathiou, Pedro Trancoso and Ioannis Sourdis
"LLC-guided Data Migration in Hybrid Memory Systems"
International Parallel and Distributed Processing Symposium (IPDPS) 2019.

Other publications

The following publications were also published during my PhD studies. However, they are not appended to this thesis because their contents are overlapping of not related to the thesis.

[a] Evangelos Vasilakis, Vassilis Papaefstathiou, Pedro Trancoso and Ioannis Sourdis
"FusionCache: using LLC Tags for DRAM Cache"
Design, Automation and Test in Europe (DATE) 2018.

[b] Alirad Malek, Evangelos Vasilakis, Vassilis Papaefstathiou, Pedro Trancoso, Ioannis Sourdis
"Odd-ECC: On-demand DRAM Error Correcting Codes"
4th Annual International Symposium on Memory Systems (MEMSYS) 2017

[c] Evangelos Vasilakis, Ioannis Sourdis, Vassilis Papaefstathiou, Antonis Psathakis and Manolis Katevenis
"Modeling Energy-Performance Tradeoffs in ARM big.LITTLE Architectures"
27th International Symposium on Power and Timing Modeling (PATMOS) 2017

Contents

Chapter 1

Introduction

The performance of computer systems is largely dominated by their memory hierarchy [1]. Besides latency, memory bandwidth can be a limiting factor for many workloads running on Chip Multiprocessors (CMPs) [2–5]. On one hand, data intensive applications as well as the large number of cores and specialized accelerators integrated on a chip increase the demand for higher data rates. On the other hand, memory bandwidth is pin limited [2, 6] and is therefore more difficult to scale [5].

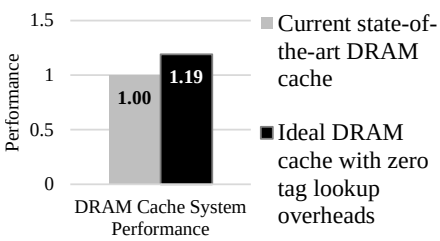
3D-stacking technology can be used to increase memory bandwidth. In particular, 3D-stacked DRAM can be placed near the processor die offering substantially higher bandwidth than off-chip DRAM. 3D-stacked DRAM though has limited capacity and often needs to be complemented with a larger off-chip DRAM that has however lower bandwidth. Currently, there are two general approaches to integrate both 3D-stacked DRAM and off-chip DRAM in a system: the first one is a flat address-space memory system supporting migration between the two types of DRAM [7–11]; the second one uses 3D-stacked DRAM as a cache [12–26].

DRAM caches have shown excellent potential in capturing the spatial and temporal data locality of applications, however they are still far from their ideal performance as explained in the next section.

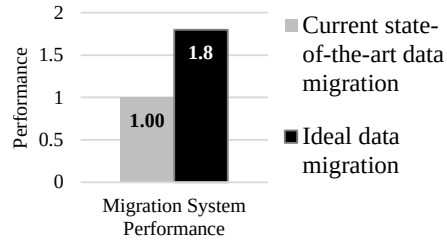
As opposed to DRAM caches, data migration keeps 3D-stacked DRAM capacity available to the system. This means that data cannot be just copied to 3D-stacked DRAM as in caches, but instead have to be swapped which incurs double the overhead of copying. To amortize the overheads of swapping, it is important to select for migration only a subset of the accessed data; preferably the ones with the highest potential for future reuse. Reducing the swapping overheads as well as selecting the most promising data to migrate are critical factors for the perfomance of systems that support data migration.

This thesis addresses the above problems proposing new, more efficient designs for DRAM caches and data migration in a system that utilizes 3D-stacked DRAM in addition to conventinal off-chip DRAM.

The rest of this introductory Chapter is organized as follows: The problem statement is presented in Section 1.1 followed by a discussion of the objectives of this thesis in Chapter 1.2. Finally, the contributions of this thesis are summarized in Section 1.3.



(a) Performance of a system with current state-of-the-art DRAM cache [19] and with an ideal DRAM cache with zero tag lookup overheads.



(b) Performance of a system with current state-of-the-art data migration scheme [9] and ideal system where all data are in 3D-stacked DRAM without any overheads.

Figure 1.1: Potential performance of DRAM caches and Data migration.

1.1 Problem Statement

The performance of systems that use DRAM caches as well as systems with data migration is limited by their particular overheads. For DRAM cache designs, these overheads are largely related to the management of the tags and for data migration schemes, these overheads are related to the migration traffic. These are the core problems addressed in this thesis and are explained in more detail bellow.

1.1.1 Tag Lookups in DRAM Caches

The DRAM cache tag access latency affects performance and depends on the tag organization and management. Each design choice comes with different tradeoffs that are tightly related to the DRAM cache line size. Smaller DRAM cache lines offer more flexibility and more efficient use of the cache bandwidth when the application is characterized by low spatial locality. Larger DRAM cache lines offer better prefetching and overall better performance when the workloads exhibit spatial locality. On the other hand, smaller DRAM cache lines require more tag storage than larger ones for the same cache size making it infeasible to store them on chip. Even for larger DRAM cache lines, the cost of storing the tags on chip is not negligible and it could otherwise be utilized for a larger on-chip Last Level Cache (LLC). Storing the DRAM cache tags in DRAM is more space efficient and also allows for smaller DRAM cache lines but it results in substantially higher tag access latency as well as increased 3D-stacked DRAM traffic.

Figure 1.1a shows the performance overhead of tag lookups in the current state-of-the-art DRAM cache design [19]. An ideal DRAM cache that performs tag lookups with zero latency would have 19% better performance. This performance gap presents an opportunity for improving existing DRAM cache designs.

1.1.2 Data Movement Overheads and Data Selection for Migration

Data migration differs from caching in that it does not waste the capacity of the 3D-stacked from the memory system. To achieve that, data migration schemes need to swap memory segments from 3D-stacked to off-chip DRAM instead of just copying as is for caching. Swapping however, requires double the memory traffic of copying. Migration traffic competes directly with processor memory requests for bandwidth and increases the queuing latency, especially in off-chip DRAM, which is the bottleneck. To increase the performance of data migration it is important to reduce the migration traffic overheads as well as to select data with good potential for future reuse.

Figure 1.1b shows that an ideal system where all data are always found (with zero overheads) in the high bandwidth 3D-stacked DRAM could achieve 1.8× better performance than current state-of-the-art data migration scheme [9]. This significant gap in performance is due to the migration overheads as well as due to sub-optimal data selection of existing data migration schemes. This thesis aims to bridge this gap by improving these aspects of data migration.

1.2 Thesis Objectives

The aim of this thesis is to improve the performance of systems that use DRAM caches as well as systems that employ data migration by reducing their respective overheads described above. Bellow follows a more detailed description of the objectives for each of the two design alternatives along with some related work and the approach pursued in this thesis.

1.2.1 Minimizing the Tag lookup Overheads in DRAM Caches

The first thesis objective is to improve the performance of systems that use 3D-stacked DRAM as a cache by minimizing their tag lookup overheads. The main idea behind this is to:

- *Store information about the location of DRAM cache lines in the tag array of the on-chip Last Level Cache (LLC) so as to access the data in the DRAM cache directly after LLC misses.*

Related Work: Several designs have been proposed aiming to reduce the DRAM cache tag access latency, however they are not able to nullify it and some of them introduce significant constraints to the system. One such design employs an on-chip SRAM cache of the DRAM cache tags [19]. This reduces the average DRAM cache tag lookup latency however it adds a constant delay to every DRAM cache access for accessing the tag-cache and more on-chip resources are occupied for caching the DRAM cache tags. Another technique places the DRAM cache data addresses directly in the TLB entries [21]. Every TLB entry would then have information about the location of the respective page in the DRAM cache. However, this requires fixing the DRAM cache line size to the Operating System (OS) page size, which can be inefficient for

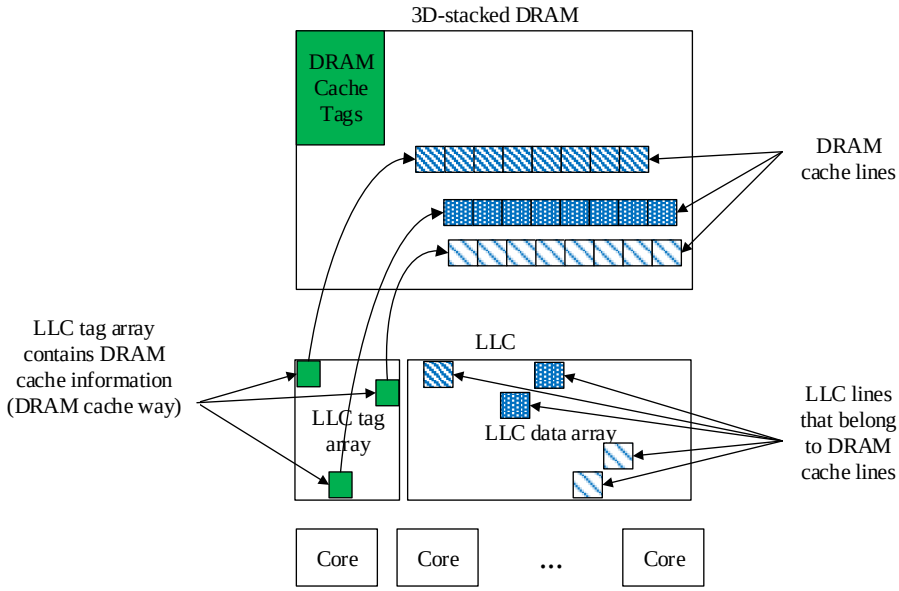


Figure 1.2: Decoupled Fused Cache (DFC) overview.

applications with low spatial locality and wasteful in terms of off-chip bandwidth and DRAM cache space. The inefficiencies of this approach would be even more evident in systems that use super-pages/huge-pages [27–29]. Other techniques such as *Alloy Cache* and *Compound Access Scheduling* collocate DRAM cache data and tags in the same DRAM row to allow faster accesses [13, 24]. These designs either require a direct mapped cache organization or customizing the cache associativity and cacheline size to the DRAM row size. Such restrictions can impact the hit rate or waste DRAM cache capacity.

In summary, although existing DRAM cache designs reduce the tag lookup latency, they do so by either introducing a constant latency to all accesses, as in the case of tag-caches, or by severely limiting critical DRAM cache parameters such as cache line size and associativity, ultimately limiting the performance of DRAM caches.

Thesis Approach: To minimize the tag lookup overheads for DRAM caches this thesis proposes *Decoupled Fused Cache* (DFC), described in *paper A* of this thesis. DFC is a new DRAM cache architecture that mitigates the cost of accessing the DRAM cache tags while enforcing minimal design restrictions. Figure 1.2 provides a conceptual overview of our proposal. *Decoupled Fused Cache* (DFC) takes advantage of the redundancy in the tags within the LLC as well as across the LLC and DRAM cache tag arrays and uses the LLC tag-array to store information about the location of data in the DRAM cache. In the common case, this allows DFC to access the DRAM cache data array without looking up its tags which are stored in 3D-stacked DRAM. *Decoupled Fused Cache* decouples the location of LLC tags from the location of the LLC lines in the LLC data array in a way that resembles Decoupled Sector Caches [30]. In a nutshell, an LLC tag is associated with a DRAM cache line, which consists of several LLC lines, while the LLC management (validity, dirty, etc.) is

performed (and related information is stored) at LLC line granularity. DFC can support a configurable (at boot time) DRAM cache line size, which is a power-of-two multiple of the LLC line size. In essence the only limitation of DFC is that the DRAM cache lines needs to be at least twice as large as an LLC line.

Contrary to existing work, DFC mitigates the DRAM cache tag access overheads without imposing significant design restrictions. More precisely, DFC does not require any OS support, it does not limit DRAM cache associativity, it does not impose additional overheads in every access, and does not affect LLC performance. Still, DFC offers zero tag access overhead in the common case, and can dynamically (at boot time) support variable DRAM cache line sizes.

1.2.2 Minimizing Migration Overheads and Improving Data Selection

The second objective of this thesis is to improve the performance of systems that use 3D-stacked DRAM as part of a flat address space with data migration. It does so by minimizing the migration traffic overheads and improving the selection of data that are migrated. The ideas behind this objective are to:

- *Use the on-chip LLC to guide data selection for migration based on the observed spatial locality.*
- *Reduce migration traffic by not migrating cache lines already present in the LLC, as they can be written to their new location upon eviction.*

Related Work: There exists a large body of prior work on data migration for hybrid memory systems. The core component of every data migration strategy, is the way a memory segment is selected for migration. Most approaches use counters to keep track of accesses to memory segments [31] or counters for every segment within a group [7, 10]. So far, the most promising approach has been the activity tracking mechanism proposed by Mempod [9], which uses the *Majority Element Algorithm* (MEA) [32]. MEA has been shown to predict the hottest pages within an interval with high accuracy and at minimal hardware cost.

Different approaches trigger migrations in different ways. Many of them do it on time intervals [9, 31], while others do it on an event, e.g. CAMEO migrates at every memory access that is in a far memory [33]. Some approaches trigger migrations when the values of selection counters go beyond some threshold [7, 10].

Another aspect that characterizes the different approaches is whether the migration mechanism is based on software or hardware, or a combination of the two. Some migration mechanisms rely on the OS with some hardware support to identify the working set and orchestrate the migration [31], others only involve the hardware and are transparent to the OS [7, 9, 10, 33].

Overall, current approaches to data migration are far from their ideal performance. Their performance is limited partly due to the increased overheads of data migration and also because of the difficulty to select data to migrate with good potential for future reuse.

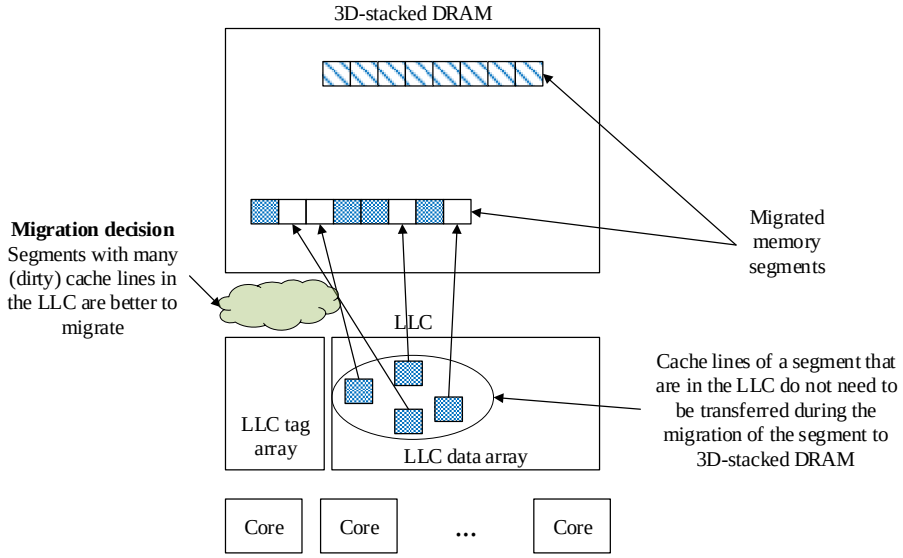


Figure 1.3: LLC-guided data migration (LGM) overview.

Thesis Approach: For data migration between off-chip and 3D-stacked DRAM, this thesis proposes LLC-guided Migration (LGM), a novel scheme for data migration in hybrid memory systems aiming both at improving the selection of migrated data as well as at reducing their traffic overheads. Figure 1.3 provides a conceptual overview of our proposed design, described in paper B of this thesis. Improving the selection of migrated data is achieved by using the LLC to guide the selection of memory segments to be migrated by detecting high spatial and temporal locality. More precisely, the LLC is used to identify memory segments that have a large number of cachelines on-chip. This is an indication for potential future reuse, which gets stronger when these cachelines are dirty. Employing the LLC to select segments for migration ensures that these segments are at that moment –at least partly– present in the last level cache (LLC). This can be used for reducing migration traffic. This is because when a fraction of a memory segment is located in the LLC, it can be omitted from the migration to reduce the migration traffic, as long as the LLC writes it back to memory when evicted.

The main novelty of our approach is the following: Firstly, the migration overheads are reduced by avoiding traffic for cachelines already present in the LLC. Secondly, the quality of selecting of data for migration is improved. Even more important is that segments are selected for migrations when a large fraction of them resides in the LLC, this timing further reduces the migration traffic.

1.3 Contributions

Following the above approaches towards achieving the objectives of this thesis, the following contributions are made and presented in the papers included in the thesis:

1.3.1 DRAM Caches

For DRAM caches we propose Decoupled Fused Cache [34], a new cache hierarchy which:

- Stores information about the contents of the DRAM cache in the LLC to avoid DRAM cache tag lookups for most LLC misses.
- Supports any DC-cacheline size power-of-two multiple of a LLC-cacheline (up to 4KB in our experiments), which is configurable at boot time.
- Improves performance by an average of 55% and 11% compared to a baseline DRAM cache and the current state-of-the-art DRAM cache [19], respectively.
- Reduces DRAM cache traffic by 1/3 and 2/3 compared to a baseline and the current state-of-the-art DRAM cache respectively.
- Reduces DRAM cache energy by by 24.5% versus the current state-of-the-art and by 62% compared to a baseline DRAM cache.

1.3.2 Data Migration

For data migration we propose LLC-guided Data Migration (LGM) [35], a data migration scheme which:

- Employs the LLC to detect locality and leverages it for migrating data with higher potential for reuse.
- Reduces the migration traffic overhead by avoiding to migrate data that reside in the LLC.
- Increases the benefits of the above migration traffic reduction because the selected data are more likely to be in the LLC when migrated.
- Reduces migration traffic to almost half and enables more data to be migrated therefore increasing the ratio of memory requests serviced by the 3D-stacked DRAM.
- Improves performance by 12.1% and reduces memory system dynamic energy by 13.2% compared to the current state-of-the-art [9].

1.4 Thesis Outline

The remainder of this thesis is organized as follows. Chapter 2 presents the design and evaluation of *Decoupled Fused Cache*, A DRAM cache design that uses the LLC tag array to store information about the contents of the DRAM cache. Chapter 3 presents the design and evaluation of *LLC-guided Data Migration in Hybrid Memory Systems*, a data migration scheme that uses the LLC to select memory segments to migrate based on their observed spatial locality and to reduce the migration traffic overheads.