Citation:
Marino, MD and Li, KC (2015) Reducing Memory Controller Transaction Queue Size in Scalable Memory Systems. The 2015 World Congress on Information Technology Applications and Services: Advanced Mobile, Communications, Security, Multimedia, Vehicular, Cloud, IoT, and Computing.

Link to Leeds Beckett Repository record:
https://eprints.leedsbeckett.ac.uk/id/eprint/1869/

Document Version:
Article (Accepted Version)

The aim of the Leeds Beckett Repository is to provide open access to our research, as required by funder policies and permitted by publishers and copyright law.

The Leeds Beckett repository holds a wide range of publications, each of which has been checked for copyright and the relevant embargo period has been applied by the Research Services team.

We operate on a standard take-down policy. If you are the author or publisher of an output and you would like it removed from the repository, please contact us and we will investigate on a case-by-case basis.

Each thesis in the repository has been cleared where necessary by the author for third party copyright. If you would like a thesis to be removed from the repository or believe there is an issue with copyright, please contact us on openaccess@leedsbeckett.ac.uk and we will investigate on a case-by-case basis.

Reducing Memory Controller Transaction Queue Size in Scalable Memory Systems

Mario Donato Marino and Kuan-Ching Li

Independent Researcher,
Italy

Corresponding Author,
Providence University, Taiwan

mario.dmarino@gmail.com
kuancli@gm.pu.edu.tw

Abstract. Scalable memory systems provide scalable bandwidth to the core growth demands in multicores' and embedded systems' processors. In these systems, as memory controllers (MCs) are scaled, memory traffic per MC is reduced, therefore transaction queues become shallower. As a consequence, there is an opportunity to explore transaction queue utilization and its impact on energy. In this paper we propose to evaluating the performance and energy-per-bit impact of the number of entries of the transaction queues along the MCs in these systems. Preliminary results show that reducing 50% of the number of entries, bandwidth and energy-per-bit levels are not practically affected, while if reducing them of 93%, bandwidth is reduced of 91% and energy-per-bit levels are increased of 780%.

Keywords: memory controller, RF, optical, scalable memory system.

1. Introduction

The traditional focus on memory design has switched from frequency scaling to memory scalability. The presence of multiple memory controllers increase the amount of memory parallelism, thus allowing larger memory widths. For example, Wide I/O 2 [15][16] which presents 8 MCs, each one connected to a 128bit-width rank - set of memory banks with data output aggregated and sharing addresses, thus performing a total width of 1024 bits. Furthermore, HyperMemory Cube (HMC) [4] with up to 8 MCs/ranks of individual width of 55bits (total of 440bits, I/O bit rate of 10Gbits/s).

Comparatively to these previously described solutions, advanced memory interfaces use a significant larger number of MCs. For example, optical Corona [2] presents 64 optical-MCs while DIMM Tree [14] up to 64 RFMCs (RF-based memory controllers) - total memory width is estimated about 4096 bits when interfaced to simple double data rate (DDR) memories.

As noticed in [8][17], the amount of memory traffic per channel is reduced as MCs are scaled, i.e., transaction queue utilization is poorly explored. Therefore, there is an opportunity to approach this aspect in terms of bandwidth and power. In this paper, we propose the following contributions to advance the state of art in scalable memory systems: (i) determination of the bandwidth impact when employing shallower transaction queues; (ii) determination of energy impact with reduced-size queues.

This paper is organized as follows: Section 2 presents the background and related work, Section 3 discusses about the reduction on transaction queue sizes, section 4 presents the experiments, Section 5 the results, and Section 6 our concluding remarks.

2. Background and Related Work

Recently proposed off-chip memory solutions still employ a larger number of pins, thus restricting MC scalability, i.e., memory width. For instance, Hybrid Memory Cube [4] employs 55 pins and can utilize up to 8 MCs. The maximum aggregated bandwidth in HMC is 320 GB/s while each I/O-link presents individually 10 Gbit/s. Furthermore, Wide I/O 2 [15][16] employs 128 bits per rank and 8 MCs, thus still MC-count restricted (total width 1024 bits).

RFiop [7] – illustrated in Figure 1a - is an advanced memory solution similar to 2.5D integration on silicon interposer. Originally designed to have 16 RFMCs and 96GB/s using 6GB/s-64bit ranks (total of 1024 bits), assuming that the 32nm-design proposed in [7] is scaled to 22nm, RFiop is likely to achieve 32 RFMCs/ranks and 480GB/s, i.e., significatively larger number of MCs and bandwidth.

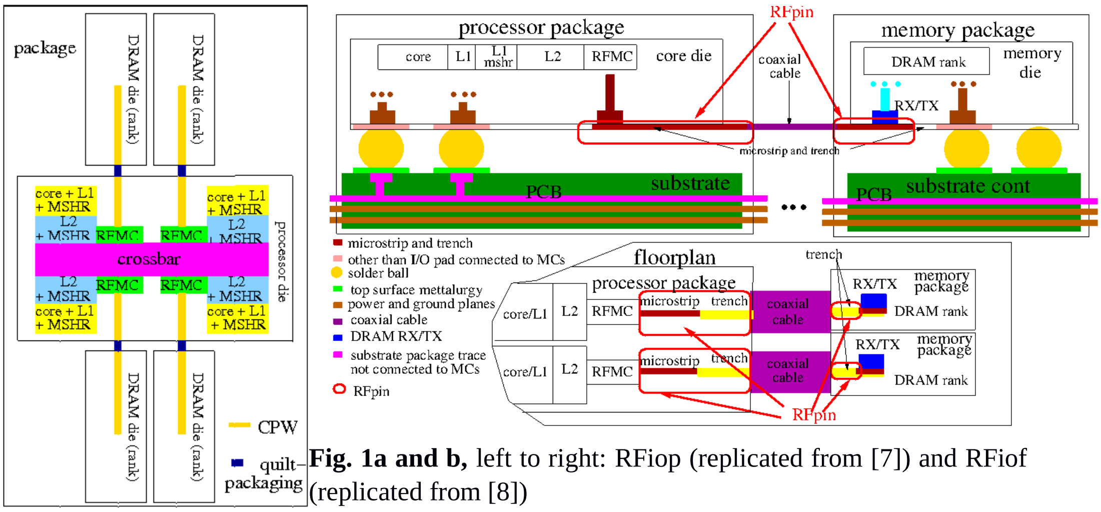
package
DRAM die (rank)
DRAM die (rank)
core + L1 + MSHR
L2 + MSHR
RFMC
crossbar
processor die
CPW
quilt-packaging
processor package
core
L1
L1 mshr
L2
RFMC
core die
RFpin
coaxial cable
microstrip and trench
memory package
DRAM rank
memory die
RX/TX
PCB
substrate
substrate cont
microstrip and trench
other than I/O pad connected to MCs
solder ball
top surface mettalurgy
power and ground planes
coaxial cable
DRAM RX/TX
substrate package trace not connected to MCs
RFpin
floorplan
processor package
core/L1
L2
RFMC
microstrip
trench
coaxial cable
RX/TX
memory package
DRAM rank
RFpin

Fig. 1a and b, left to right: RFiop (replicated from [7]) and RFiof (replicated from [8])

RFiof [8] is an advanced off-chip memory solution – illustrated in Figure 1b - presents similar MC-scalability bandwidth benefits of optical-based interfaces. RFiof is designed to scale to 32 RFMCs and 32 GB/s, using 10.8GB/s ranks. Given its low pin usage and assuming the replacement of its interface with a more conventional RF-interface (FR-board as in [14]), this technology can be scaled to use 64 RFMCs and ranks of 17.2 GB/s and likely to achieve 1024GB/s bandwidth (total width of 4096 bits), i.e., a very significative bandwidth magnitude.

According to [8], as MCs are scaled, memory traffic per channel is likely smaller. We also observe that this effect is present when when scaling multiple MCs in embedded systems such as in [17]/

All previously mentioned systems where a larger degree of MC-scaling is present, lower transaction queue utilization is likely to happen, which turns into an interesting challenge to be approached.

3. Shallower Transaction Queues

Depending on the amount of traffic, all transaction queue entries are not fully utilized, thus having performance implications. For example, for the workloads being utilized, if average utilization of the queues corresponds to 50% of the queue size, it is likely that queue size can be similarly reduced while not affecting bandwidth.

In order to mitigate the transaction queue sub-utilization, shallower transaction queues could be utilized. In this case, under the same amount of traffic and due to the employment of reduced queue sizes, the ratio between the total number of utilized entries and the total queue size is increased, i.e., a better utilization is obtained. We evaluate the performance and energy impact of this aspect in next section.

Table 1. Architectural Parameters and Benchmarks

Architectural Parameter	Description
Core	4.0 GHz, OOO, multicore, 32 cores, 4-wide issue, turnament branch predictor
Technology	22 nm
L1 cache	32kB dcache + 32 kB icache; associativity = 2, MSHR = 8, latency = 0.25 ns
L2 cache	1MB/per core ; associativity = 8, MSHR = 16; latency = 2.0 ns
RF-crossbar	latency = 1 cycle, 80GB/s
RFMC-transaction queue	1 to 32 RFMCs; 1-16 transaction queue entries, 1 RFMC/core, 2.0GHz, on-chip, buffer size = 32/MC, close page mode, interleaving memory addresses along RFMCs
Memory rank	DDR3-1333MT/s, 1 rank/MC, 1GB, 8 banks, 16384 rows, 1024 columns, 64 bits, Micron MT41K128M8 [20], tras=26.7cycles, tcas=trcd=8cycles
RF interconnection length size, delay	2.5 cm, 0.185ns
Benchmark	32 threads
STREAM [9]	4 Mdoubles per core, 2 iterations, read:write = 2.54:1
pChase [13]	64 MB/thread, 3 iterations, random, read:write=158:1

4. Experimental Section and Methodology

In order to evaluate the performance and energy-per-bit impact of the transaction queue size, we combine detailed accurate simulators using the methodology developed in [6]: upon benchmark execution of a multicore model in M5 [12], memory transactions are generated and captured by DRAMsim [3] (set with 32 RFMCs, which represent a large number of MCs). DRAMsim responds to M5 with the result of each memory transaction.

The multicore model employs a 4.0-GHz-4-wide out-of-order (OOO) core. We used Cacti [1] to obtain cache latencies and 1 MB/core L2 caches, which are interconnected via a 80GB/s-RF-crossbar with 1-cycle latency - adopting same timing settings of [5][10]: 200ps of TX-RX delays, plus the rest of the cycle to transfer 64 Bytes using high speed and modulation. RF timing settings include low bit error rate (BER) and RF-transmission delays.

The baseline configuration has 32 RFMCs, each queue with 16 entries, while having RFMCs at 2.0GHz (half of processor clock frequency). The parameters and memory-bound benchmarks [9][13] employed in this experimentation are listed in Table 1. Each RFMC is assumed to be connected to one rank in order to extract its maximum bandwidth. We vary the number of entries of each MC (from 16 to 1 entry) in order to capture the behavior of bandwidth and energy-per-bit.

5. Results

Figure 2a illustrates the results of the bandwidth experiments: as transaction queues are reduced from 16 to 1 entry, we obtain a bandwidth reduction of up to 65% for pChase and 91% for STREAM (average of STREAM benchmarks).

Figure 2b illustrates the related rank energy-per-bit results: as transaction queue sizes are reduced from 16 to 1 entry, average energy-per-bit levels increase up to 123% for pChase and 780% for STREAM.

Comparing these two figures, it is interesting to notice that 4 and 8 entries (up to 50% of the total number of entries – 16 entries) have equivalent performance to 16 entries while presenting similar energy-per-bit-levels than 16 entries, which demonstrate the advantage of smaller transaction queues. Therefore, by having medium-size ones the bandwidth/energy efficiency of the memory system is improved. For an aggressive reduction, bandwidth/energy are significantly affected.

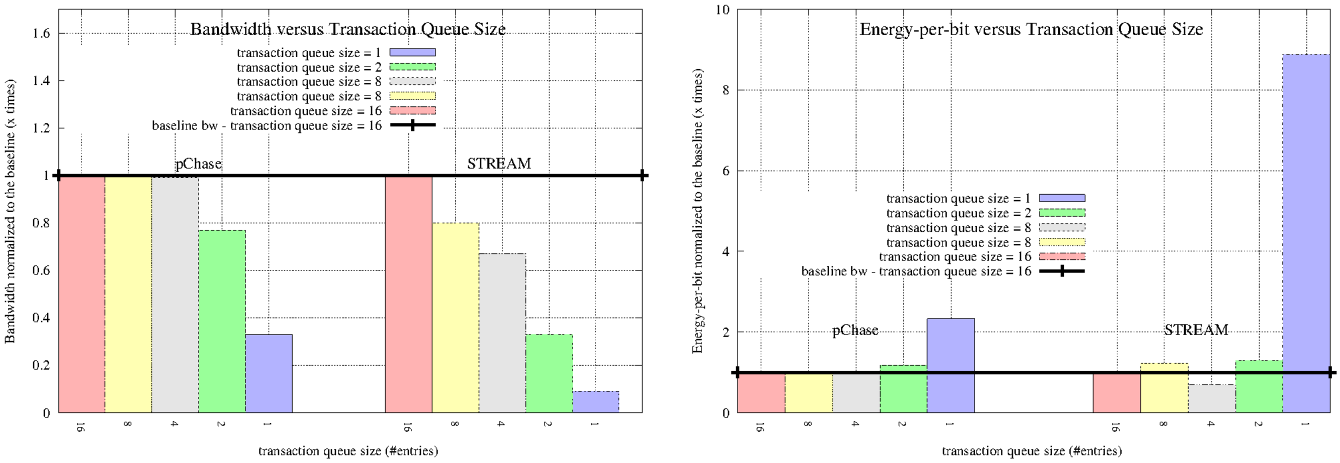

Fig 2a and b, left to right: bandwidth and rank energy versus transaction queue size

6. Conclusions

In this paper we have performed an initial evaluation of the bandwidth and energy behavior when reducing transaction queue sizes in scalable memory systems. These preliminary results show that using medium-size transaction queues, bandwidth and energy-per-bit levels are still interesting, thus leading to a higher efficiency.

Given these results we propose as future plans the evaluation of the MC power with smaller transaction queues and the extension of this evaluation with scientific benchmarks. Furthermore, we propose to evaluate the MC power effects when having shallower queues. Moreover, we also are considering the utilization of low power DDRs and evaluate the impact of those in performance when combined to shallower transaction queues.

References

[1] CACTI 5.1. Accessed Date: 09/10/2014; http://www.hpl.hp.com/techreports/2008/HPL200820.html.

[2] D. Vantrease et al. Corona: System Implications of Emerging Nanophotonic Technology. In ISCA, pages 153–164, DC, USA, 2008. IEEE.

[3] David Wang et al. DRAMsim: a memory system simulator. SIGARCH Comput. Archit. News, 33(4):100–107, 2005.

[4] Hybrid Memory Cube Specification 1.0. Accessed date: 03/03/2014; http://www.hybridmemorycube.org/.

[5] M. Frank Chang et al. CMP Network-on-Chip Overlaid With Multi-Band RF-interconnect. In HPCA, pages 191–202, 2008.

[6] Marino, M. D. On-Package Scalability of RF and Inductive Memory Controllers. In Euromicro DSD, IEEE, 2012.

[7] Marino, M. D. RFiop: RF-Memory Path To Address On-package I/O Pad And Memory Contro ller Scalability. In ICCD, 2012, Montreal, Quebec, Canada. IEEE, 2012.

[8] Marino, M. D. RFiof: An RF approach to the I/O-pin and Memory Controller Scalability for Off-chip Memories. In CF, May 14-16 , Ischia, Italy. ACM, 2013.

[9] McCalpin, J. D. Memory Bandwidth and Machine Balance in Current High Performance Computers. IEEE TCCA Newsletter, pages 19–25, December 1995.

[10] M.C.F. Chang et al. Power reduction of CMP communication networks via RF-interconnects. In MICRO, pages 376–387, Washington, USA, 2008. IEEE.

[11] Micron manufactures DRAM components and modules and NAND Flash. Accessed date: 11/10/2014 ; http://www.micron.com/.

[12] Nathan L. Binkert et al. The M5 Simulator: Modeling Networked Systems. IEEE Micro, 26(4):52–60, 2006.

[13] The pChase Memory Benchmark Page. Accessed date: 08/09/2014 ; http://pchase.org/.

[14] Kanit et al. Therdsteerasukdi. The dimm tree architecture: A high bandwidth and scalable memory system. In ICCD, pages 388–395. IEEE, 2011.

[15] JEDEC Publishes Breakthrough Standard for Wide I/O Mobile DRAM. Accessed date: 11/28/2014 ; http://www.jedec.org/.

[16] Wide I/O 2, Hybrid Memory Cube (HMC) Memory Models Advance 3D-IC Standards. Accessed date: 12/05/2014 ; http://www.cadence.com/Community/blogs/ii/archive/2013/08/06/wide-io-2-hybrid-memory-cube-hmc-memory-models-advance-3d-ic-standards.aspx.

[17] Marino, M. D., Li, K.C. Insights on Memory Controller Scaling in Multicore Embedded Systems . International Journal of Embedded Systems, 6(4), 2014.