A Comparative Study of Predictable DRAM Controllers

Real-time embedded systems require hard guarantees on task Worst-Case Execution Time (WCET). For this reason, architectural components employed in real-time systems must be predictable, i.e., allow the derivation of tight bounds on worst-case latency. However, it has been established that COTS DRAM controllers yield extremely pessimistic latency bounds, leading to high WCET and rendering them unusable for real-time systems. As a result, the research community has produced several predictable memory controller designs that provide tighter worst-case latency. The proposed controllers significantly differ in terms of system model, memory configuration, arbitration and command scheduling, latency bounds, general performance, and simulation environment. Due to such differences and the complexity of evaluating designs, no controller has been properly compared against other proposed solutions, making it difficult to assess the contribution of each approach. To bridge this gap, this paper provides the first comprehensive evaluation of state-of-the-art predictable DRAM controllers. We first propose a categorization of available controllers based on key architectural characteristics and real-time guarantees. We then introduce an analytical performance model to compare controllers based on worst-case latency. Finally, we develop a common simulation platform, conduct extensive evaluation of all state-of-the-art controllers, and discuss findings and recommendations.

1. INTRODUCTION

Modern real-time hardware platforms use memory hierarchies consisting of both on-chip and off-chip memories that are specifically designed to be predictable. A predictable memory hierarchy simplifies worst-case execution time (WCET) analysis to compute an upper bound on the memory access latency incurred by a task's execution. This is essential in computing a task's overall WCET, which is used to ensure that temporal requirements of the task are never violated. These memory hierarchies are specially designed for real-time embedded systems because conventional memory hierarchies are optimized to improve average-case performance, yielding overly pessimistic WCET bounds [Kim et al. 2014]. Consequently, there has been considerable interest in the real-time research community in designing memory hierarchies to produce tight WCET bounds while delivering a reasonable amount of performance.

As of late, the community has focused in making accesses to off-chip dynamic random-access memories (DRAM)s predictable. This need arose because the data requirement demands from modern real-time applications greatly exceeded the capacity available solely with on-chip memories. However, commercial-off-the-shelf (COTS) DRAMs are unsuitable for real-time embedded systems because their controllers are optimized to improve average-case performance; thus, rendering either pessimistic WCET bounds or even no upper bound [Wu et al. 2013]. As a result, we have witnessed several innovations in DRAM memory controller (MC) design in recent years [Paolieri et al. 2009; Hassan et al. 2015; Li et al. 2014; Wu et al. 2013; Ecco and Ernst 2015; Krishnapillai et al. 2014; Ecco et al. 2014; Jalle et al. 2014; P et al. 2011]. Each of these designs trade-off predictability with performance, but they also make it difficult to compare against each other. This is because the authors of these works use different system models, assumptions, memory configurations, arbitration and command scheduling algorithms, benchmarks, and simulation environments. A designer or company wishing to adopt one of these DRAM MCs for their real-time application would have virtually no scientific method to judiciously select the one that best suits the needs of their ap-

plication. Moreover, researchers producing novel DRAM MC designs are also unable to effectively compare against prior state-of-the-arts. We believe that this is detrimental to future progress in the research and design of DRAM MCs, and its adoption into main-stream hardware platforms for real-time embedded systems.

To address this issue, in this paper we develop a methodology that enables comparing predictable MCs, and we provide a comprehensive evaluation of the state-of-the-art predictable DDR SDRAM MCs. More in details, we provide the following main contributions: 1) We discuss a characterization of existing predictable MCs based on their key architectural characteristics and real-time properties; 2) We introduce an analytical performance model that enables a quantitative comparison of existing MCs based on their worst-case latency; and 3) We develop a common evaluation platform to provide a fair, standardized experimental comparison of the analyzed MCs. Based on this platform, we carry out an extensive simulation-based evaluation using embedded benchmarks, and provide insights into the advantage and disadvantages of different controller architectures. In particular, we expose and evaluate essential trade-offs between latency bounds provided to real-time tasks and average memory bandwidth offered to non real-time tasks.

Our source-code for managing and simulating all considered MC designs is available at [Authors removed for double-blind review 2015]. To the best of our knowledge, this is the first work that enables comparing performance of all state-of-the-art architectural solutions for predictable scheduling of DRAM operations [1]. A key result of the evaluation is that the relative performance of different predictable controllers is highly influenced by the characteristics of the employed DDR memory device; hence, controller and device should be co-selected.

The rest of the paper is organized as follows. Section II provides the required background on Dual Data Rate (DDR) DRAM. Section III discusses the structure of predictable memory controllers and their key architectural characteristics. Section IV presents related work in general and the evaluated predictable MCs in details, based on the introduced architectural characteristics. Section V provides the analytical model of worst-case latency. Section VI discusses our evaluation platform, provides extensive evaluation of all covered predictable controller, and discusses findings and recommendations. Finally, Section VII provides concluding remarks.

2. DRAM BACKGROUND

We begin by providing key background details on double data rate synchronous dynamic RAM (DDR SDRAM). Most recent predictable MCs are based on JEDEC DDR3 devices. In this evaluation, we focus on both DDR3 and its currently available successor standard, DDR4. There are other standards exist, but are not included in this work. Note that we only consider systems with a single memory channel, i.e., a single MC and command/data buses. In general, from an analysis point of view, if more than

[1]Note that our evaluation mainly focuses on solutions for scheduling of DRAM commands, i.e., at the MC back-end level. We are not concerned with scheduling of memory requests at the core, cache or interconnection level.

one channel is present, then each channel can be treated independently; hence, all discussed predictable MCs are single-channel. Optimization of static channel assignment for predictable MCs is discussed in [Gomony et al. 2013].

2.1. DRAM Organization

A DRAM chip is a 3-dimensional array of memory cells organized in banks, rows, and columns. A DRAM chip consists of 8 (DDR3/DDR4) or 16 (DDR4) banks that can be accessed simultaneously, but share the same command/address and data bus. Each bank is further organized into rows and columns. Every bank contains a row-buffer, which temporarily stores the most recently accessed row of memory cells. Data can only be retrieved once the requested row is placed in the row-buffer. This makes subsequent accesses to the same row (row locality) quicker to access than different rows. A memory module, used in a typical computer system comprises either one or multiple independent sets of DRAM chips connected to the same buses. Each memory set is also known as a rank. Figure 1 shows an overview of a DRAM memory module with N ranks, where each rank includes 8 DRAM chips. In this example, each chip has an 8 bits data bus, and 8 chips are combined to form an overall data bus with width $W_{BUS} = 8 \cdot 8 = 64$ bits for the whole module. While each rank can be operated independently of other ranks, they all share the same address/command bus, used to send memory commands from the MC to the device, as well as the same data bus.

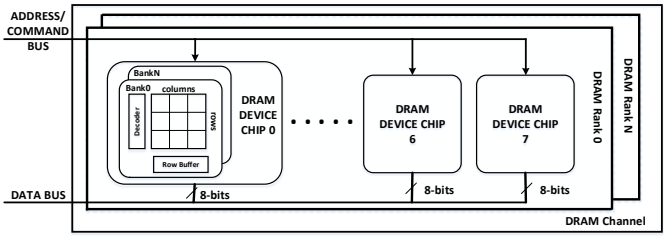

Fig. 1: Architecture of Memory Controller and DRAM memory module.

2.2. DRAM Commands and Timing Constraints

The commands pertinent to memory request latency are as follows: ACTIVATE (ACT), READ (RD), READA (RDA), WRITE (WR), WRITEA (WRA), PRECHARGE (PRE) and REFRESH (REF). Other power related commands are out of the scope of this paper. Each command has some timing constraints that must be satisfied before the command can be issued to the memory device. A simplified DRAM state diagram, presented in Figure 2, shows the relationship and timing constraints between device states and commands. We report the most relevant timing constraints for DDR3-1600H and DDR4-1600K in Table I, which are defined by the JEDEC standard [JEDEC 2008] .

The ACT command is used to open (retrieve) a row in a memory bank into the row-buffer. The row remains active for accesses until it is closed by a PRE command. PRE is used to deactivate the open row in one bank or in all the banks. It writes the data in the row-buffer back to the storage cells; after the PRE, the bank(s) will become available for another row activation after t_{RP}. Once the required row is opened in the row-buffer, after t_{RCD}, requests to the open row can be performed by issuing CAS

commands: reads (RD) and writes (WR). Since the command bus is shared, only one command can be sent to the device at a time. If a request accesses a different row in the bank, a PRE has to be issued to close the open row. In the case of auto precharge, a PRE is automatically performed after a RD (RDA) or WR (WRA) command. Finally, a REF command needs to be issued periodically (t_{REFI}) to prevent the capacitors that store the data from becoming discharged. REF can only be issued once the device is in Idle mode for at least t_{RP} after all the banks are precharged. After the refresh cycles (t_{RFC}) complete, all the banks will be in the precharged (idle) state.

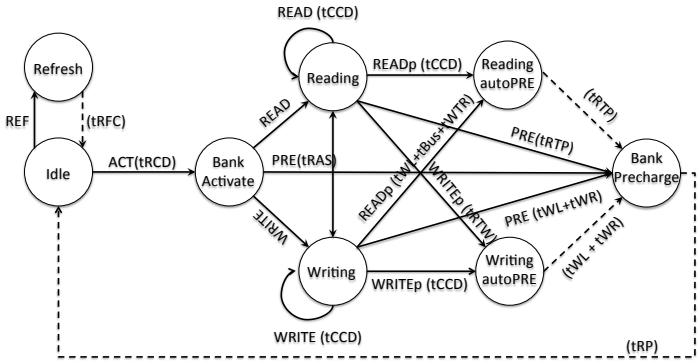

Fig. 2: DRAM Operation State Machine.

Table I: JEDEC Timing Constraints.

JEDEC Specifications (cycles)			
Parameters	Description	DDR3-1600H	DDR4-1600K
t_{RCD}	ACT to RD/WR delay	9	11
t_{RL}	RD to Data Start	9	11
t_{RP}	PRE to ACT Delay	9	11
t_{WL}	WR to Data Start	8	9
t_{RTW}	RD to WR Delay	7	8
t_{WTR}	WR to RD Delay	6	6
t_{RTP}	Read to PRE Delay	6	6
t_{WR}	Data End of WR to PRE	12	12
t_{RAS}	ACT to PRE Delay	28	28
t_{RC}	ACT-ACT (same bank)	37	39
t_{RRD}	ACT-ACT (diff bank)	5	4
t_{FAW}	Four ACT Window	24	20
t_{BUS}	Data bus transfer	4	4
t_{RTR}	Rank to Rank Switch	2	2
t_{RFC}	Time required to refresh	$160ns$	$160ns$
t_{REFI}	Refresh period	$7.8us$	$7.8ns$

A DDR device is named in the format of DDR(generation)-(data rate)(version) such as DDR(3)-(1600)(H). In each generation, the supported data rate varies. For example, for DDR3 the data rate ranges from 800 to 2133 MegaTransfers(MT)/s, while for DDR4 the rate starts from 1600 and goes up to 2400MT/s. Note that since the device operates at double data rate (2 data transfers every clock cycle), a device with 1600MT/s is clocked at frequency of 800MHz. Devices operating in the same speed with lower version letter can execute commands faster than devices with higher version.

Based on the timing constraints in Table I, we make the following three important observations. 1) While the operation of banks can be in parallel, command and data

must still be serialized because the MC is connected with the memory devices using a single command and a single data bus. One command can be transmitted on the command bus every clock cycle, while each data transmission (read or write) requires $t_{BUS} = 4$ clock cycles. In this paper, we use a burst length of 8 since it is supported by both JEDEC DDR3 and DDR4 devices. 2) Since consecutive requests targeting the same row in a given bank do not require ACT and PRE commands, they can proceed faster than requests targeting different rows; timing constraints that are required to precharge and reactivate the same bank (t_{RC}, t_{RAS}, t_{WR}, t_{RTP} and t_{RP}) are particularly long. 3) Switching between requests of different types (read/write) incurs extra timing constraints in the form of a read-to-write (t_{RTW}) and write-to-read ($t_{WL} + t_{BUS} + t_{WTR}$) switching delays between CAS commands. Such constraints only apply to CAS commands targeting banks in the same rank; for CAS commands targeting banks in different ranks, there is a single, short timing constraint (t_{RTR}) between data transmissions, regardless of the request type.

3. MEMORY CONTROLLER DESIGN

Based on the background on DDR DRAM provided in Section 2, we now discuss the design of the memory controller. In particular, in Section 3.1 we describe a common architectural framework that allows us to categorize different MCs based on their key architectural features.

3.1. Hardware Architecture

A DRAM memory controller is the interface to the DRAM memory module and governs access to the DRAM device by executing the requests as required by the timing constraints of the DRAM specification. In doing so, the MC performs four essential roles: address mapping, request arbitration, command generation, and command scheduling as shown in Figure 3.

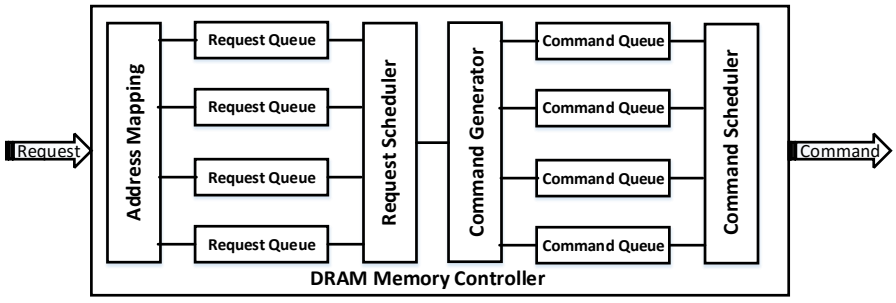

Fig. 3: Architecture of Memory Controller.

— Memory Address Mapping: Address mapping decomposes the incoming physical address of a request into rank, bank, row, and column bits. The address translation determines how each request is mapped to a rank and bank. There are two main classes of mapping policies.

(1) Interleaved-Banks: each requestor can access any bank or rank in the system. This policy provides maximum bank parallelism to each individual requestor,

but suffers from row interference since different requestors can cause mutual interference by closing each other's row buffers. Hence, predictable MCs using interleaving-banks also employ close-page policy, which ignores row locality. COTS MCs also typically employ interleaved-banks because in the average case, the row interference is often limited.

(2) Private-Banks: each requestor is assigned its own bank or set of banks. This allows a predictable MC to take advantage of row locality, since the behavior of one requestor has no impact on the row buffer of other requestors' banks. As a downside, the performance of a requestor executing alone is negatively impacted, since the number of banks that it can access in parallel is reduced. Sharing data among requestors also becomes more complex [ZP 2013]. Finally, the number of requestors can be a concern due to the limited number of ranks and banks. For example, a DDR3 memory supports only up to 4 ranks and 8 banks per rank, but a multi-core architecture may have 16 or more memory requestors.

— Request Arbitration: While a MC only needs to schedule individual commands to meet JEDEC timing constraints, in practice all considered MCs implement an additional *front-end request scheduler* that determines the order in which requests are processed. We consider three main arbitration schemes:

(1) (Non-work Conserving) Time Division Multiplexing (TDM): under TDM, each requestor is assigned one or more slots, and its requests can only be serviced during assigned slot(s). If no request can be served during the assigned slot, then the slot is wasted.

(2) Round Robin (RR) and Work-conserving TDM: compared to non-work conserving TDM, unused slots are assigned to the next available requestor.

(3) First-Ready First-Come-First-Serve (FR-FCFS): COTS MCs generally implement some variation of FR-FCFS scheduling to improve the memory bandwidth. This scheme prioritizes requests that target an open row buffer over requests requiring row activation; open requests are served in first-come-first-serve order. FR-FCFS controllers always implement an open-page policy. As shown in [Wu et al. 2013], if the MC does not impose any limit to the number of reordered requests, no upper bound on request latency can be derived. Therefore, based on experimental evaluation, the analysis in [Kim et al. 2014] derives a latency bound assuming that at most 12 requests within a bank can be reordered ahead of a request under analysis.

For general purpose system, the write operations are not in the critical path, therefore, some MCs provide high priority for read requests and write requests can be served when there is no read operation. Most real-time MCs treat these two type of requests equally and providing individual latency or the maximum between the two. For the MCs evaluated in this work, we take the maximum latency among the read and write requests as the worst case request latency.

— Command Generation: Based on the request type (read or write) and the state of the memory device, the command generation module generates the actual memory commands. The commands generated for a given request depend on the row policy used

by the MC and the number of CAS commands needed by a request; this is determined by the data size of the request and the size of each memory access. For instance, for a $W_{BUS} = 16$ bits, each operation transfers 16 bytes, thus requiring 4 accesses for a 64 bytes request; whereas for $W_{BUS} = 64$ bits, only one access per request would be needed. The commands for a request can be generated based on two critical parameters introduced in [Li et al. 2014]: the number of interleaved banks (BI) and the burst count for one bank (BC). The BI determines the number of banks accessed by a request and the BC determines the number of CAS commands generated for each bank. The value for BI and BC depends on the request size and data bus width. Predictable MCs cover the whole spectrum between close-page policy, open-page policy and combined hybrid approaches.

(1) Open-Page: allows memory accesses to exploit row locality by keeping the row accessed by a previous request available in the row-buffer for future accesses. Hence, if further requests target different column cells in the same row opened in the row-buffer, then the command generator only needs to generate the required number of CAS commands, incurring minimum access latency. Otherwise, if the further requests target different rows, the command generator needs to create a sequence of commands PRE+ACT and required CAS which results in longer latency.

(2) Close-Page: transitions the row-buffer to an idle state after every access completes by using auto-precharge READ/WRITE commands. Hence, subsequent accesses place data into the row-buffer using an ACT command prior to performing the read or write operation. The command generator only needs to create a sequence of ACT and CAS commands. While this does not exploit row locality, all requests incur the same access latency making them inherently predictable. Furthermore, the latency of a request targeting a different row is shorter under close-page policy since the pre-charge operation is carried out by the previous request.

(3) Hybrid-Page: is a combination of both open and close policy for large requests that require multiple memory accesses (CAS commands). The CAS commands for one request can be a sequence of a number of CAS commands to leverage the benefit of row locality, followed by a CASp command to precharge the open buffer.

— Command Scheduler: The command scheduler ensures that queued commands are sent to the memory device in the proper order while honouring all timing constraints. Apart from the page policy, we find that the biggest difference between predictable MCs is due to the employed command scheduling policy.

(1) Static: Controllers using static command scheduling schedule groups of commands known as bundles. Command bundles are statically created off-line by fixing the order and time at which each command is issued. Static analysis ensures that the commands meet all timing constraints independently of the exact sequence of requests serviced by the MC at run time. Static command scheduling results in a simpler latency analysis and controller design, but can only support

close-page policy since the controller can not distinguish the row state at run time.

(2) Dynamic: These controller schedule commands individually. The command arbiter must include a complex sequencer unit that tracks the timing constraints at run time, and determines when a command can be issued. Dynamic command scheduling allows the controller to adapt to varying request types and bank states; hence, it is often used in conjunction with open-page policy.

Except serving the commands for a memory request, a memory controller is responsible for refreshing the DRAM device. The refresh strategy is different for memory controller with different page policy because the refresh command requires all the banks to be precharged before it can be issued. Refresh delay is generally limited to 1% - 5% of total task memory latency [Akesson et al. 2007] and can be easily incorporated in WCET analysis, see [Wu et al. 2013] for example.

3.2. Other Features

Outside of the architectural alternatives discusses in Section 3.1, there are a few additional key features that distinguish MCs proposed in the literature. First of all, in some system, requests generated by different requestors can have **varying request sizes**. For example, a processor generally makes a memory request in the size of a cache line, which is 64 bytes in most modern processors. On the other hand, an I/O device could have memory requests up to KBs. Some MCs are able to natively handle requests of different sizes at the command scheduler level; as we will show in our evaluation, this allows to trade-off the latency of small requests versus the bandwidth provided to large requests. Other MCs handle only fixed-size requests, in which case large requests coming from the system must be broken down into multiple fixed-size ones before they are passed to the memory controller.

Requestors can be further differentiated by their **criticality** (temporal requirement) as either hard real-time (HRT) or soft real-time (SRT). Latency guarantees are the requirement for HRTs, while for SRT, a good throughput should be provided while worst-case timing is not crucial. In the simplest case, a MC can support mixed-criticality by assigning higher static priority to critical requests over non-critical ones at both the request and command scheduling level. We believe that all predictable MCs can be modified to use the fixed priority scheme. However, some controllers are designed to support mixed-criticality by using a different scheduling policy for each type of request.

Finally, as we described in the DRAM background, a memory module can be constructed with a number of **ranks**. In particular, a DDR3/DDR4 memory module can have up to 4 ranks. However, only some controllers distinguish between requests targeting different ranks in the request/command arbitration. Since requests targeting different ranks do not need to suffer the long read-to-write and write-to-read switching delays, such controllers are able to achieve tighter latency bounds, at the cost of needing to employ a more complex, multi-rank memory device.

4. RELATED WORK

Since memory is a major bottleneck in almost all computing systems, we have observed a lot of research efforts to address this problem. These efforts inspired many novel memory controller designs in the recent years, which we categorize into two main groups. Both groups attempt to address the shortcomings of the commonly-deployed FR-FCFS; though, each group focuses on different aspects. The first group investigate the effect of FR-FCFS on conventional high-performance multi-core platforms. In these platforms, FR-FCFS aims to increase DRAM throughput by prioritizing ready accesses to an already-open row (row htis). This behaviour may lead to unfairness across different running applications. Applications with large number of ready accesses are given higher priority; thus, other applications are penalized and may even starve. Researchers attempt to solve these problem by proposing novel scheduling mechanisms such as ATLAS [Kim et al. 2010], PARBS [Mutlu and Moscibroda 2008], TCM [Kim et al. 2010], and most-recently BLISS [Subramanian et al. 2016]. The common trend amongst all these designs is promoting application-aware memory controller scheduling.

On the other hand, fairness is not an issue for real-time predictable memory controllers. In fact, prioritization is commonly adopted by those controllers to favor critical cores for instance. However, FR-FCFS is not also a perfect match for those controllers, yet for a different reason. Critical applications executing on real-time systems must have bounded latency. Because of the prioritization nature of FR-FCFS, it does not guarantee a bound of the memory latency. Many works have been proposed to provide guaranteed memory latency bounds [Ecco and Ernst 2015; Ecco et al. 2014; Goossens et al. 2013; Reineke et al. 2011; Kim et al. 2015; Hassan et al. 2015; Jalle et al. 2014; Paolieri et al. 2009; Li et al. 2014; Wu et al. 2013; Krishnapillai et al. 2014], of which, we consider [Ecco and Ernst 2015; Ecco et al. 2014; Hassan et al. 2015; Jalle et al. 2014; Paolieri et al. 2009; Li et al. 2014; Wu et al. 2013; Krishnapillai et al. 2014] in this comparative study. We briefly discuss the efforts we did not consider in the study a long with the reasons for this decision. Afterwards, we discuss the MCs we consider in the study in details in Section 4.1.

Goossen et al. proposed a mixed-row policy memory controller [Goossens et al. 2013] that leaves the row open for a fixed amount before closing it. This method requires aggressive memory accesses to take advantage of the open row access windows, which normally suit for out-order memory accesses. In our simulations, we proposed that every request is in order which may not fully utilize the open windows. Reineke et al. designed the PRET controller [Reineke et al. 2011] with private bank mapping. This work is not taken into account because it relies on a specific precision-timed architecture (PTARM [I et al. 2010]), which makes the controller incompatible with standard cache based architectures. Kim et al. [Kim et al. 2015] designed a mixed criticality with private bank mapping. The design is very similar to ORP except that it assumes a non-critical requestor can share the same bank with the critical requestor with lower priority in the command level.

4.1. Predictable Memory Controller Analysis

In this section, we summarize the state-of-the-art predictable DRAM memory controllers [Ecco and Ernst 2015; Ecco et al. 2014; Hassan et al. 2015; Jalle et al. 2014; Paolieri et al. 2009; Li et al. 2014; Wu et al. 2013; Krishnapillai et al. 2014] described in the literature. In general, we consider as predictable all MCs that are composable. A MC is composable if requestors cannot affect the temporal behaviour of other requestors [Akesson and Goossens 2012]. This implies that applications running on different cores have independent temporal behaviours, which allows for independent and incremental system development [Kim et al. 2014]. However, we notice that composability is used with two different semantics in the literature, which we term analytically and run-time composable. A MC is **analytically composable** if it supports an analysis that produces a predictable upper bound on request latency that is independent of the behavior of other requestors. A MC is **run-time composable** if the run-time memory schedule for a requestor is independent of the behavior of other requestors. Run-time composability implies analytical composability, but not vice-verse. All the selected designs are analytical composable, however (non work-conserving) TDM is the only arbitration that supports run-time composability, potentially at the cost of degrading average-case performance. In Table II, we classify each MC based on its architectural design choices (address mapping, request arbitration, page policy and command scheduling) and additional features (variable request size, mixed criticality, and rank support). Note: the Dirc request scheduler passes the request on top of a request queue to the backend.

	AMC	PMC	RTMem	ORP	DCmc	ReOrder	ROC	MCMC
Req. Size	N	Y	Y	N	N	N	N	N
Mix-Criti.	Fix Prio.	Req. Sched.	N	N	Fix Prio.	N	Ranks	Fix Prio.
Rank	N	N	N	N	N	Y	Y	Y
Addr. Map.	Intlv.	Intlv.	Intlv.	Priv.	Priv.	Priv.	Priv.	Priv.
Req. Sched.	RR	WC TDM	WC TDM	Dirc	RR	Dirc	Dirc	TDM
Page Policy	Close	Hybrid	Hybrid	Open	Open	Open	Open	Close
Cmd. Sched.	Static	Static	Dyn.	Dyn.	Dyn.	Dyn.	Dyn.	Static

Table II: Memory Controllers Summary

4.1.1. Analyzable MC (AMC). AMC [Paolieri et al. 2009] is the first design for predictable MC which employs the simplest scheduling scheme: static command scheduling with close-page policy is used to construct off-line command bundles for read/write requests.

4.1.2. Programmable MC (PMC). PMC [Hassan et al. 2015] also employs a static command scheduling strategy with four static command bundles based on the minimum request size in the system. For a request size that can be completed within one bundle, PMC uses close-page policy. However, PMC divides larger requests into multiple bundles using open-page policy. PMC also employs an optimization framework to generate an optimal work-conserving TDM schedule. The framework supports mixed-criticality systems, allowing the system designer to specify requirements in terms of either maximum latency or minimum bandwidth for individual requestors. The generated TDM schedule comprises several slots, and requestors are mapped to slots based on an assigned period.

4.1.3. Dynamic Command Scheduling MC (RTMem). RTMem [Li et al. 2014] is a memory controller back-end architecture using dynamic command scheduling and can be combined with any front-end request scheduler; we decided to implement work-conserving TDM to better compare against other predictable MCs. RTMem accounts for variable request size by decoding each size into a number of interleaved banks (BI) and a number of operations per bank (BC) based on a pre-defined table. The BI and BC values are selected off-line to minimize the request latency.

4.1.4. Private Bank Open Row Policy MC (ORP). To the best of our knowledge, ORP [Wu et al. 2013] is the first example of a predictable MC using private bank and open-page policy with dynamic command scheduling. Latency bounds are derived assuming that the number of close-row and open-row requests for an application are known, for example based on static analysis [Bourgade et al. 2008]. The MC uses a complex FIFO command arbitration to exploit maximum bank parallelism, but still essentially guarantees RR arbitration for fixed-size critical requests.

4.1.5. Dual-Criticality MC (DCmc). Similar to ORP, DCmc [Jalle et al. 2014] uses a dynamic command scheduler with open page policy, but it adds support for mixed-criticality and bank sharing among requestors. Critical requestors are scheduled according to RR, while non-critical requestors are assigned lower priority and scheduled according to FR-FCFS. The controller supports a flexible memory mapping; requestors can be either assigned private banks, or interleaved over shared sets of banks. Our evaluation considers the private-bank configuration since it minimizes latency bounds for critical requestors.

4.1.6. Rank Switching, Open-row MC (ROC). ROC [Krishnapillai et al. 2014] improves over ORP using multiple ranks to mitigate the t_{WTR} and t_{RTW} timing constraints. As noted in Section 2, such timing constraints do not apply to operations targeting different ranks. Hence, the controller implements a two-level request arbitration scheme for critical requestors: the first level performs a RR among ranks, while the second level performs a RR among requestors assigned to banks in the same rank. ROC's rank-switching mechanism can support mixed-criticality applications by mapping critical and non-critical requestors to different ranks. FR-FCF, can be applied for non-critical requestors.

4.1.7. Read/Write Bundling MC (ReOrder). ReOrder[Ecco and Ernst 2015; Ecco et al. 2016] improves ORP and ROC by employing CAS reordering techniques to reduce the access type switching delay. It uses dynamic command scheduler among all the three DRAM commands such that round-robin for ACT and PRE commands, and read/write command reorder for the CAS command. The CAS arbiter keeps track the serviced requestor and the previous issued CAS type (Read/Write) in order to make decision on the next scheduled CAS command. The CAS reorder can eliminate repetitive CAS switching timing constraint if the Read and Write commands are scheduled alternatively.

4.1.8. Mixed Critical MC (MCMC). MCMC [Ecco et al. 2014] uses a similar rank-switching mechanism as in ROC, but applies it to a simpler scheduling scheme using

static command scheduling with close-page policy. TDM arbitration is used to divide the timeline into a sequence of slots alternating between ranks. Each slot is assigned to a single critical requestor and any number of non-critical requestors; the latter are assigned lower priority. The slot size can be minimized by using a sufficient number of ranks to mitigate the t_{WTR}/t_{RTW} timing constraints and a sufficient number of slots to defeat the intra-bank timing constraints. As with TDM arbitration, the main drawback of this approach is that bandwidth will be wasted at run-time if no requestor is ready during a slot.

4.2. Analytical Worst-Case Memory Access Latency

As discussed in Section 4.1, all considered predictable MCs are analytically composable. In particular, all authors of cited papers provide, together with their MC design, an analytical method to compute a worst case bound on the maximum latency suffered by memory requests of a task running on a core under analysis, which is considered one of the memory requestors. This bound depends on the timing parameters of the employed memory device, any other static system characteristics (such as the number of requestors), and potentially the characteristics of the tasks (such as the row hit ratio), but does not depend on the activity of the other requestors. To do so, all related work assume a task running on a fully timing compositional core [Wilhelm et al. 2009], such that the task can produce only one request at a time, and it is stalled while waiting for the request to complete. The worst-case execution time (WCET) of the task is then obtained as the computation time of the task with zero-latency memory operations plus the computed worst-case total latency of memory operations. Note that in general no restriction is placed on soft or non real-time requestors, i.e., they can be out-of-order cores or DMA devices generating multiple requests at a time.

In the rest of this section, we seek to formalize a common expression to compute the memory latency induced by different predictable controllers. Inspired by the WCET derivation method detailed in [Wu et al. 2013; Kim et al. 2014], we shall use the following procedure: 1) for a close page controller, we first compute the worst case latency $Latency^{Req}$ of any request generated by the task, assuming that the request is not interrupted by a refresh procedure. This is because refreshes are infrequent but can stall the memory controller for a significant amount of time; hence, including the refresh time in the latency bound would produce an extremely pessimistic bound. Assuming that the task under analysis produces NR memory requests, the total memory latency can then be upper bounded by $NR \cdot Latency^{Req}$ plus the total refresh delay for the whole task, which can be tightly bounded by the procedure in [Wu et al. 2013; Kim et al. 2014]. 2) For an open page controller, we compute worst case latencies $Latency^{Req-Open}$ and $Latency^{Req-Close}$ for any open and close request, respectively. Assuming that the task has row hit ratio HR, we can then simply follow the same procedure used for close page controllers by defining:

$$Latency^{Req} = Latency^{Req-Open} \cdot HR + Latency^{Req-Close} \cdot (1 - HR). \quad (1)$$

Based on the discussion above, Equations 2 and 3 summarize the per-request latency for a close page and an open page MC, respectively, where HR is the row hit ratio of the task and $REQr$ is either the number of requestors in the same rank as the requestor

under analysis (for controllers with rank support), or the total number of requestors in the system (for controllers without rank support).

$$Latency^{Req} = BasicAccess + Interference \cdot (REQr - 1) \tag{2}$$

$$\begin{aligned} Latency^{Req} = &(BasicAccess + RowAccess \cdot (1 - HR)) + \\ &(Interference + RowInter \cdot (1 - HR)) \cdot (REQr - 1) \end{aligned} \tag{3}$$

In the proposed latency equations, we factored out the terms HR and $REQr$ to represent the fact that for all considered MCs, latency scales proportionally to $REQr$ and $(1 - HR)$. The four latency components, $BasicAccess$, $RowAccess$, $Interference$ and $RowInter$, depend on the specific MC and the employed memory device, but they also intuitively represent a specific latency source. For a close page controller, $BasicAccess$ represents the latency encountered by the requests itself, assuming no interference from other requestors; note that since predictable MCs treat read and write operations in the same way, their latency is similar and we thus simply consider the worst case among the two. $Interference$ instead expresses the delay caused by every other requestor on the commands of the request under analysis. For an open page controller, $BasicAccess$ and $Interference$ represent the self-latency and interference delay for an open request, while $RowAccess$ and $RowInter$ represent the additional latency/interference for a close request, respectively. We will make use of this intuitive meaning to better explain the relative performance of different MCs in Section 5. When the number of requestors in each rank is the same for rank support MCs, the expression can be rearranged to be a function of total number of requestors in the system REQ, instead of using the requestors per rank $REQr$. The expression is demonstrated in Equation 4.

$$\begin{aligned} Latency^{Req} &= (BasicAccess + RowAccess \cdot (1 - HR)) + (Interference + RowInter \cdot (1 - HR)) \cdot (\frac{REQ}{R} - 1) \\ &= (BasicAccess - Interference \cdot \frac{(R-1)}{R}) + (RowAccess - RowInter \cdot \frac{(R-1)}{R}) \cdot (1 - HR) + \\ &(\frac{Interference}{R} + \frac{RowInter}{R} \cdot (1 - HR)) \cdot (REQ - 1) \end{aligned} \tag{4}$$

We tabulate the values of these four latency terms for all covered MC in Table III. Equations are derived based on the corresponding worst case latency analysis for each MC; we refer the reader to [Paolieri et al. 2009; Hassan et al. 2015; Jalle et al. 2014; Wu et al. 2013; Krishnapillai et al. 2014; Ecco et al. 2014; Kim et al. 2014] for detailed proofs of correctness and tightness evaluation. We provide the detail proof of the each MC expression in [omitted for double-blind review 2016]. While the numeric values in Table III are specific for a DDR3-1600H memory device, the general equations and related observations hold for all considered memory devices. In the table, the BI and BC are referred in Section 3.1 . R represents the number of ranks used in the memory module.

Finally, note that a composable analytical bound for FR-FCFS scheduling with private bank partition has been proposed in [Kim et al. 2014]. However, we believe that such a bound is generally over-pessimistic to be usable in practice since the interference

Table III: MC General Equation Components ($\mathcal{K}(cond)$ equals 1 if $cond$ is satisfied and 0 otherwise.)

	$RowInter$	$Interference$	$BasicAccess$	$RowAccess$
AMC	NA	$(15 \cdot \mathcal{K}(BI = 8) + 42) \cdot BC$	$(15 \cdot \mathcal{K}(BI = 8) + 42)$	NA
PMC RTMem	NA	$\mathcal{K}(BC = 1) \cdot ((15 \cdot \mathcal{K}(BI = 8) + 42)) + \mathcal{K}(BC > 1) \cdot ((4 \cdot BC + 1) \cdot BI + 13 + 4 \cdot \mathcal{K}(BI = 8))$	$\mathcal{K}(BC = 1) \cdot ((15 \cdot \mathcal{K}(BI = 8) + 42)) + \mathcal{K}(BC \neq 1) \cdot ((4 \cdot BC + 1) \cdot BI + 13 + 4 \cdot \mathcal{K}(BI = 8))$	NA
DCmc	0	$28 \cdot BC$	$13 \cdot BC$	18
ORP	7	$13 \cdot BC$	$19 \cdot BC + 6$	27
ReOrder	$7 + 2R$	$8R \cdot BC$	$(8R + 12) \cdot BC + 13$	$31 + 2R$
ROC	$3 \cdot R + 6$	$(3 \cdot R + 12) \cdot BC$	$\left(3 \cdot R + 24\right) \cdot BC + 6$	$3 \cdot R + 27$
MCMC	NA	$Slot \cdot R \cdot BC$ Where $Slot = \begin{cases} 42/PE & if(REQr \leq 6) \wedge (R \leq 2) \\ 9 & if(R = 2) \wedge (REQr > 6) \\ 7 & Otherwise \end{cases}$	$Slot \cdot R \cdot BC + 22$	NA
FR-FCFS	0	$224 \cdot BC$	$24 \cdot BC$	18

component value is much higher than the other predictable MCs as shown in Table III; hence, in the context of this paper we deem the memory controller with FR-FCFS arbitration to be non-predictable.

5. EXPERIMENTAL EVALUATION

We next present an extensive experimental evaluation of all considered predictable MCs. We start by discussing our experimental framework in Section 5.1. Then, we show results in terms both of simulated latencies and analytical worst case bounds in Section 5.2. In particular, we start by showing execution time results for a variety of memory-intensive benchmarks. We then compare the worst case latencies of the MCs based on the number of requestors and row hit ratio, as modelled in Section 4.2. At last, we evaluate both the latency and bandwidth available to requestors with different properties (request sizes and criticality) and on different memory modules (data bus width and speed). Finally, based on the obtained results, Section 5.3 provides a discussion of the relative merits of the different MCs and their configurations.

5.1. Experimental Framework

The way the discussed MCs have been evaluated in their respective papers is widely different in terms of selected benchmarks and evaluation assumptions such as the operation of the frontend, the behavior of the requestors, and the pipelining through the controller. The consequence is that it is not possible to directly compare the evaluation results of these designs with each other. Therefore, we strive to create an evaluation environment that allows the community to conduct a fair, and comprehensive comparison of existing predictable controllers. We select the EEMBC auto benchmark suite [Poovey 2007] as it is representative of actual real-time applications. Using the benchmark, we generate memory traces using MACsim architectural simulator [Kim et al. 2012]. The simulation uses a x86 CPU clocked at 1GHz with private 16KB level 1, 32KB level 2 and 128KB level 3 caches. The output of the simulation is a memory trace containing a list of accessed memory addresses together with the memory request type (read or write), and the arrival time of the request to the memory controller. In Table IV, we present the information for memory traces with bandwidth higher than 150MB/s, which can stress the memory controller with intensive memory accesses. We

provide the computation time of each application without memory latency, the total number of requests and the open request (row hit) ratio. An essential note is related

Table IV: EEMBC Benchmark Memory Traces.

Benchmark	Computation Time (ns)	Number of Requests	Bandwidth (MB/s)	Row Hit Ratio
a2time	660615	2846	275	0.35
cache	1509308	5503	233	0.18
basefp	1051300	3336	202	0.30
irrflt	1022514	3029	189	0.33
aifirf	1035458	2765	170	0.40
tblook	1152044	2865	159	0.35

to the behaviour of the processor. As discussed in Section 4.2, to obtain safe WCET bounds for hard real-time tasks, all related work assume a fully timing compositional core [Wilhelm et al. 2009]. Therefore, we decided to run the simulations under the same assumption: in the processor simulation, traces are first derived assuming zero memory access latency. The trace is then fed to a MC simulator that computes the latency of each memory request. In turn, the request latency is added to the arrival time of all successive requests in the same trace, meaning a request can only arrive to the memory controller after the previous request from the same requestor has been complete. This represents the fact that the execution of the corresponding application would be delayed by an equivalent amount on a fully timing compositional core.

Each of the considered MCs is designed with a completely different simulator, which varies in simulation assumption as we described above, and simulation model such as event-driven or cycle-accurate. In this case, it is very difficult to fairly evaluate the performance of these controllers by running individual simulators. Therefore, we have implemented all the designs based on a common simulation engine, which allows us to realize each memory controller by specifying their memory address mapping, request scheduler, command generator, and command scheduler.

In this way, we can guarantee that all designs are running with same memory device, same type of traces, same request interface and no delay through the memory controller. We configured each controller for best performance; AMC, PMC, RTMem are allowed to interleave up to the maximum number of banks per rank (8) based on the request size and the data bus width. ROC and MCMC are configured to use up to 4 ranks. In DCmc, we assume no bank sharing is allowed between HRT requestors. For all analyses and simulations, we use the timing constraints of DDR3-1600H 2GB device provided in Ramulator. We did not include the impact of refresh to allow a simpler comparison with the analytical per-request latency bounds, which do not include refresh time as discussed in Section 4.2; in any case, note that the the total refresh time is a small portion of the execution time of a task, as described in Section 2.

5.2. Evaluation Results

5.2.1. Benchmark Execution Times. We demonstrate the worst case execution time in Figure 4 for all the selected memory intensive benchmarks. In all experiments in this section, unless otherwise specified, we set up the system with 8 requestors (REQs), where REQ0 is considered as the requestor under analysis and is executing one benchmark. The other REQs are executing synthetic memory intensive trace to maximize the interference. We also assume 64 bytes requests with a bus size $W_{BUS} = 64$ bits. For

controllers using multiple ranks (ReOrder, ROC and MCMC), requestors are evenly split among ranks, leading to 4 requestors per rank with 2 ranks, and 2 requestors per rank with 4 ranks. When measuring the execution time of the benchmark, the simulation will be stopped once all the requests in REQ0 have been processed by the memory controller. The execution time of each benchmark is normalized based on the analytical bound of AMC. The color bar represents simulated execution time for the requestor (benchmark) under analysis and the T-sharp bar represents the analytical worst case execution time.

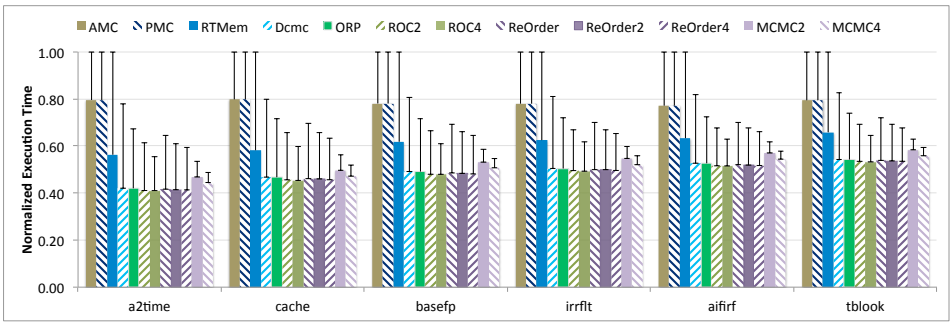

Fig. 4: EEMBC Benchmark WCET with 8 64B REQs and 64bit Data Bus

To best demonstrate the performance for each MC, in the rest of the evaluation, we use the benchmark with highest bandwidth **a2time**, and we plot the worst case per-request latency $Latency^{Req}$, so that results are not dependent on the computation time of the task under analysis. For the analytical case, $Latency^{Req}$ is derived according to either Equation 2 or 3, while in the case of simulations, we simply record either the maximum latency of any request (for close page controllers) or the maximum latencies of any open and any close request (for open page controllers), so that $Latency^{Req}$ can be obtained based on Equation 1.

5.2.2. Number of Requestors. In this experiment, we evaluate the impact of the number of requestors on the analytical and simulated worst case latency per memory request of REQ0. Figure 5 and 6 shows the latency of a close request and an open request as the number of requestors varies from 4 to 16[2]. Furthermore, in Table VI we show the analytical equation components for all MCs [3]. We make the following observations: 1) For interleaved banks MCs (AMC, PMC, and RTMem), latency increases exactly proportionally to the number of requestors: *Interference* is equal to *RowInterference*. The latency components are also larger than other controllers. This is because these MCs implement scheduling at the request level through an arbitration between requestors. In this case, one requestor gets its turn only when other previously scheduled requestors complete their requests. The timing constraint between two requests is bounded by the re-activation process of the same bank, which is the longest constraint among all others. Therefore, increasing the number of requestors has a large effect on

[2]Note that since ORP and DCmc assign one REQ per bank and use a single rank, for the sake of fair comparison we assume they can access 16 banks even when using DDR3.
[3]Note that since the request size is 64 bytes and the data bus width is 64 bit, each request can be served by one CAS command with a burst length of 8. Therefore, the parameter BI and BC is set to 1.

Table V: WC Latency Components with BI=1, BC=1

	AMC/PMC /RTMem	DCmc	ORP	ReOrder1	ReOrder2	ReOrder4	ROC2	ROC4	MCMC2	MCMC4
Interference (per REQ)	42	28	13	8	8	8	9	6	9	7
RowInterfer (per REQ)	NA	0	7	9	6	4	6	5	NA	
BasicAccess	42	13	25	33	33	33	27	24	31	29
RowAccess	NA	18	27	33	30	28	27	25	NA	

the latency. 2) Bank privatized MCs (DCmc, ORP, ReOrder, ROC and MCMC) are less affected by the number of requestors because each requestor has its own bank and it only suffers interference from other requestors on different banks. The timing constraints between different banks are much smaller than constraints on the same bank. Dynamic command scheduling is used in DCmc, ORP, ReOrder and ROC to schedule requestors at the command level. Increasing the number of requestors increases the latency for each command of a request, therefore, the latency for a request also depends on the number of commands it requires. For example, a close request in open page MCs can suffer interference from other requestors for PRE, ACT and CAS commands. MCMC uses fixed TDM to schedule requestors at the request level. Increasing the number of requestors increases the number of TDM slots one requestor suffers. 3) MCs that are optimized to minimized read-to-write and write-to-read penalty (ReOrder, ROC and MCMC) have much lower interference, especially for open requests, compared to other controllers. For close requests, MCMC achieves significantly better results than other controllers, since it does not suffer extra interference on PRE and ACT commands.

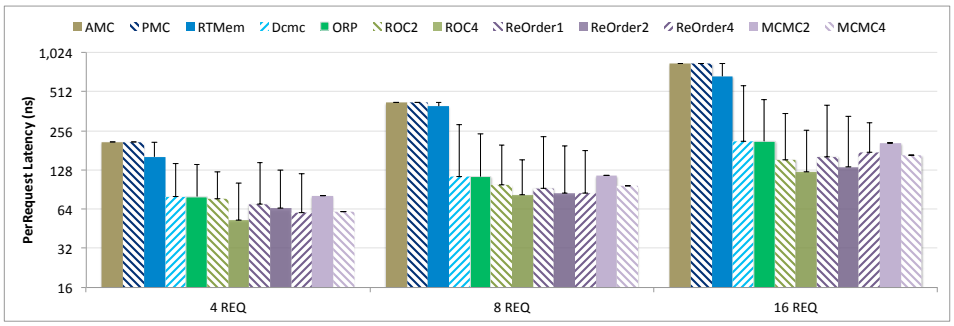

Fig. 5: WC Latency per Close Request of REQ0 wtih 64Bit Data Bus.

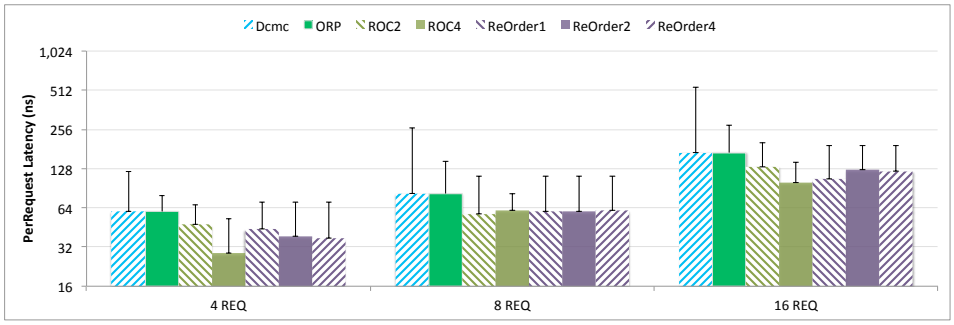

Fig. 6: WC Latency per Open Request of REQ0 wtih 64Bit Data Bus.

5.2.3. Row Locality. As we described in Section 3, the row hit ratio is an important property of the task running as a requestor for MCs with open page policy. In this experiment, we evaluate the impact of row hit ratio on the worst case latency of open page MCs ORP, DCmc, ReOrder and ROC. In order to maintain the memory access pattern, and change the row hit ratio, we synthetically modify the request address to achieve row hit ratio from 0% to 100%. Instead of showing the worst latency for both close and open requests, we take the average latency of the application as the general expression proposed in Section 4.2. As expected, in Figure 7 we observe that both the analytical latency bound and the simulated latency decrease as the row hit ratio increases. The impact of row hit ratio can be easily predicted from the equation based on the $RowAccess$ and $RowInter$ components.

Fig. 7: Average Request Latency for Open Page MCs

5.2.4. Data Bus Width. In this experiment, we evaluate the request latency by varying the data bus width W_{BUS} from 32 to 8 bits. Using smaller data bus width, same size of request is served with either interleaving more banks or multiple accesses to the same bank. The commands generated by the bank privatized MCs depend on the applied page policy. For open page private MCs (DCmc, ORP, ReOrder, and ROC), a PRE+ACT followed by a number of CAS commands are generated for a close request. On the other hand, MCMC needs to perform multiple close page operations, and each request needs multiple TDM rounds to be completed. The analytical and simulated worst case latency per request is plotted in Figure 8, while Table VI shows the analytical components as a function of the number of interleaved banks BI for interleaved MCs and number of consecutive accesses to the same bank BC for private bank MCs; for example, with 64 bytes request size and 8 bits data bus, interleaved banks MCs interleave through BI=8 banks and bank privatized MCs require BC=8 accesses to the same bank. We can make

Table VI: WC Latency Components with 8 REQ($Ex = 15 \cdot \mathcal{K}(BI = 8)$)

	AMC/PMC /RTMem	DCmc	ORP	ROC2	ROC4	ReOrder1	ReOrder2	ReOrder4	MCMC2	MCMC4
Interference (per REQ)	$42 + Ex$	$28BC$	$13BC$	$9BC$	$6BC$	$8BC$	$8BC$	$8BC$	$9BC$	$7BC$
RowInter (per REQ)	NA	0	7	4	3	9	6	4	NA	
BasicAccess	$42 + Ex$	$13BC$	$19BC$+6	$21BC$+6	$18BC$+6	$33BC$	$33BC$	$33BC$	31	29
RowAccess	NA	18	27	27	26	33	30	28	NA	

the following observations: 1) The analytical bound for MCs with interleaving bank

mapping is not affected by a size of the data bus of 32 or 16 bits because the activation window for the same bank (t_{RC}) can cover all the timing constraints for accessing up to 4 interleaved banks. In the case of 8 bits width, the MCs interleave over 8 banks, resulting in 36% higher latency because of the timing constraints between the CAS commands in the last bank of a request and the CAS command in the first bank of next request (such as t_{WTR}). Interleaved Banks MCs can process one request faster by taking benefit of the bank parallelism for a request, hence leading to better access latency. 2) Both the analytical bound and the simulation result for MCs with private bank increase dramatically when the data bus width gets smaller; both *Interference* and *BasicAccess* are linear or almost linear with BC, given that each memory request is split into multiple small accesses. However, *RowInter* and *RowAccess* are unchanged, since the row must be opened only once per request.

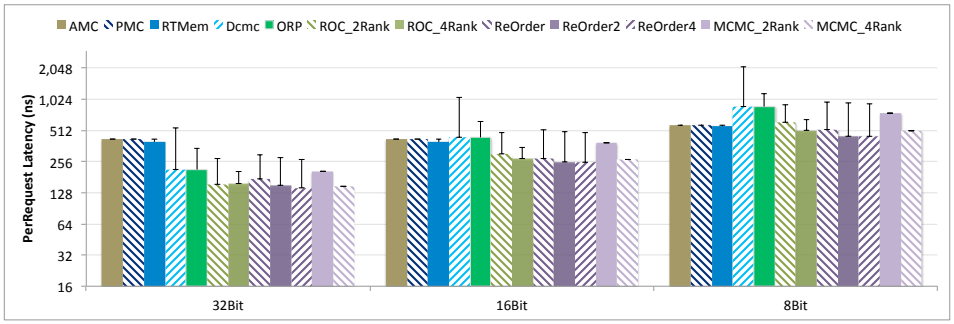

Fig. 8: Worst Case Latency per Request of REQ0 with 8 REQs.

5.2.5. Memory Device. The actual latency (ns) of a memory request is determined by both the memory frequency and the timing constraints. In general, the length of timing constraints in number of clock cycles increases when the memory device gets faster. Each timing constraint has different impact on MCs designed with different techniques. For example, the 4-Activation window (t˙FAW) has impact on interleaved banks MCs if one request needs to interleave over more than 4 banks, and affects private bank MCs only if there are more than 4 requestors in the system. In this experiment, we look at the impact of memory devices on both the analytical and simulated worst case latency. We run each MC with memory devices from DDR3-1066E to DDR3-2133L which cover a wide range of operating frequencies. We also show the difference between devices running in same frequency but different timing constraints such as DDR3-1600K and DDR3-1600H. Figure 9 represents the latency per request for each MC, and it shows that as the frequency increases, the latency decreases for all the MCs with private bank mapping and very small change to MCs with interleaved bank. This is because the interleaved banks MCs are bounded by the re-activation window to the same bank, which does not change much with the operating frequency.

5.2.6. Large Request Size. In this experiment, we consider different request sizes. We configure the system to include four requests with request size of 64 bytes (simulating a typical CPU), and four requestors generating large requests with a size of 2048

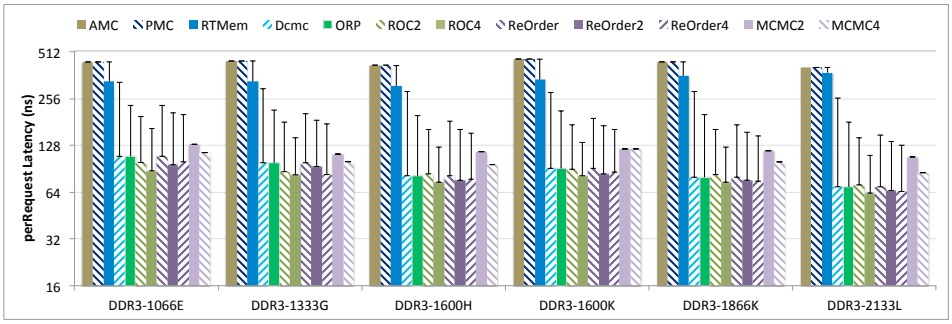

Fig. 9: Worst Case Latency per Request of REQ0

bytes (simulating a DMA device). RTMem and PMC are the only controllers that natively handle varying request sizes. RTMem has a lookup table of BI and BC based on the request size, while PMC uses a different number of scheduling slots for requestors of different types. Overall, we employ the following configuration: (1) AMC interleaves 4 banks with auto-precharge CAS commands, given that interleaving can go up to 4 banks without any delay penalty, and performs multiple interleaved accesses based on the request size; (2) PMC changes the scheduling slot order for different requestor types to trade-off between latency and bandwidth; (3) RTMem changes the commands pattern for large requests; (4) private bank MCs (ORP, DCmc, ReOrder, ROC and MCMC) do not differentiate the request size, and each large request is served as a sequence of multiple acccesses, similarly to the previous experiment with small data bus width. The configuration for AMC, RTMem and PMC is shown in Table VII. PMC executes all the predefined slot sequences in the configuration and repeats the same order after all the sequences are processed. In details, the number in each sequence is the requestor ID and the order in the sequence is the order of requestor arbitration. In short, PMC_1 assigns one slot per requestor; PMC_2 assigns double the number of slots to small requests compared to large requests; and PMC_4 assigns to small requests four times the number of slots.

Table VII: Large Request Configuration

AMC	PMC1	PMC2	PMC3	RTMem4/8	RTMem8/4
BI=4 BC=8	[0 1 2 3 4 5 6 7]	[0 1 2 3 4 5] [0 1 2 3 6 7]	[0 1 2 3 4] [0 1 2 3 5] [0 1 2 3 6] [0 1 2 3 7]	BI=4 BC=8	BI=8 BC=4

The worst case latency per request for REQ0 with 64 bytes request is shown in Figure 10 and the bandwidth of REQ 7 with 2048 bytes request is shown in Figure 11. For

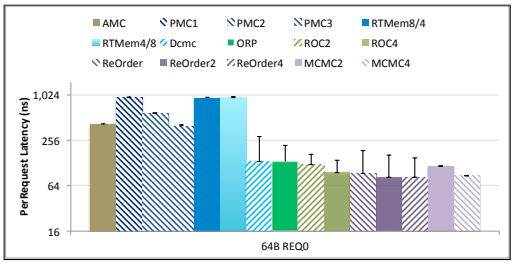

Fig. 10: WC Latency of REQ0 with 64B

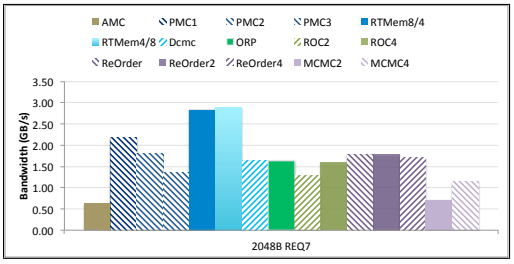

Fig. 11: Bandwidth of REQ7 with 2048B

private bank MCs, the access latency for small requests is not affected by the large requests because all the requestors are executed in parallel and the interference is only caused by memory commands instead of the requests. AMC is not affected because the slot for each requestor is the same based on the configuration. On the other hand, the bandwidth for the large requestor is low compared to MCs that take the request size into consideration. RTMem can switch the command pattern for a large request. The latency for small requestor is slightly higher when the large request is configured as [BI=4, BC=8] comparing to [BI=8 and BC=4]. However, the bandwidth is slightly increased. Based on the arbitration scheme of PMC, the latency for small request and bandwidth for large requestor is greatly affected. The trade of between the latency and bandwidth is very obvious.

5.2.7. Mixed Criticality. In this case of mixed criticality system, the system is configured with 8 HRT REQs as before, but on top of that, there are another 8 SRT REQs in the system. We can observe how much impact the SRT REQs can have on the HRT REQs and the performance of SRT requestor in each MC. AMC, DCmc, and MCMC assign priority to HRT over SRT requestors. ROC assigns different ranks to HRT and SRT REQs. Note that all open page MCs have been configured to apply FR-FCFS for SRT REQs to maximize the bandwidth [4]. ROC and MCMC requires specific assignments of HRT and SRT requestors to individual ranks; the employed configurations are detailed in Table VIII. Results are plotted in terms of latency for HRT REQ0 in Figure 12 and

Table VIII: Mixed Critical System Configuration for Multi-Rank MCs

	ROC2 ReOrder2	ROC4_1 ReOrder4_1	ROC4_2 ReOrder4_2	MCMC2	MCMC4
R0	8HRT	4HRT	3HRT	4HRT+4SRT	2HRT+2SRT
R1	8SRT	4HRT	3HRT	4HRT+4SRT	2HRT+2SRT
R2	NA	4SRT	2HRT	NA	2HRT+2SRT
R3	NA	4SRT	8SRT	NA	2HRT+2SRT

bandwidth for SRT REQ8 in Figure 13. For MCs that do not differentiate HRT and

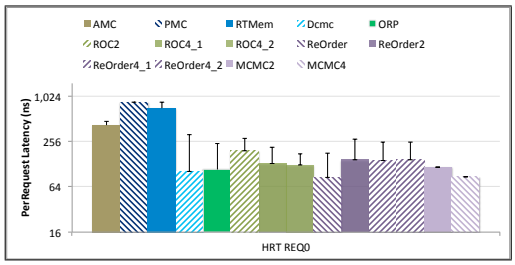

Fig. 12: WC Latency of of HRT REQ0

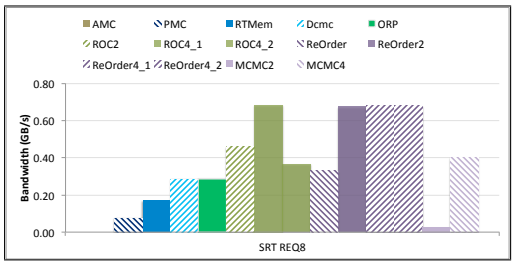

Fig. 13: Bandwidth of SRT REQ8

SRT REQs, the latency is the same as having 16 REQs in the system. The analytical latency of a HRT request for AMC and DCmc is increased due to the possibility of scheduling one SRT requestor before a HRT requestor. ROC can trade off the latency

[4]DCmc [Jalle et al. 2014], ROC [Krishnapillai et al. 2014] and ReOrder [Ecco et al. 2016] specifically mention such policy for SRT REQs. We have extended ORP and ReOrder with 1 rank to support the same configuration for the sake of fair comparison. We do not apply such policy to close page MCs since it would not yield any benefit.

for HRT and the bandwidth for SRT by allocating requestors in different ranks. In general, open page MCs perform much better in terms of available bandwidth for SRT REQs compared to close page MCs, since they can take advantage of row hits and requests reordering in the average case. In particular, note that while MCMC on 2 ranks has the second lowest analytical latency for HRT REQs, it provides almost no bandwidth to SRT REQs. This is because the slot size for MCMC on 2 ranks is fairly large, leading to low memory utilization.

5.3. Discussion

Based on the obtained results, we now summarize the key takeaways of the evaluation.

5.3.1. Memory Configuration. The characteristics of the employed memory module: data bus width, memory device speed, and number of ranks, have a significant impact on the relative performance of the tested MCs. Out of the three main characteristics, the data bus width seems by far the most important. A memory device with smaller data bus can be better utilized by MCs with interleaved banks because each request can be served by accessing a number of banks in parallel. On the other hand, private banks MCs can have better memory access latency when wider data bus is used, where a memory request can be served with less accesses to the same bank. In general, private banks MCs perform better for bus width of 32 bits and above, while interleaved banks MCs pull ahead at widths of 16 bits and below. At the same time, it is important to recognize that bus width is the major factor in the cost of the main memory subsystem: while doubling the data bus width or doubling the number of ranks both requires doubling the number of DRAM chips, an enlarged data bus width also requires adding extra physical pins to the memory controller, which can be expensive. In addition, private banks MCs show moderate improvements in latency on faster devices, and significant improvements in both latency and bandwidth from increasing the number of ranks (see also Section 5.3.2). However, the impact of faster memories is negligible for interleaved banks MCs since the bounding constraint of re-activation to the same bank is almost constant through all devices.

In summary, based on performance alone, we believe that interleaved banks MCs are suitable for simple microcontrollers, employing small bus width of 8 or 16 bits and slow, single rank devices, while private banks MCs allow improved performance at higher cost on more complex systems. However, outside of the performance/cost trade off, it is also important to recognize that private banks MCs impose a more complex system configuration: main memory must be partitioned among requestors. Note that if data must be shared among multiple HRT requestors, such data can be allocated to a shared bank [Jalle et al. 2014], but the resulting latency bound for accesses to shared data then becomes similar to AMC as the controller cannot avoid row conflicts.

5.3.2. Write-Read Switching. Among private banks MCs, the latency bounds for Re-Order, ROC and MCMC are generally significant better than ORP and Dcmc: this is because the arbitration schemes used by the former are designed to minimize the impact of the long read-to-write and write-to-read switching delays, either by reordering CAS commands, or by switching between ranks. Among the three MCs, MCMC shows

the smallest latencies, followed by ROC and ReOrder. Given that the main difference between MCMC and ROC is the page policy (close vs open), a relevant takeaway is that based on current analysis technology, there seems to be no advantage in employing open page policy for latency minimization: based on Table VI, for open requests ROC performs slightly better than MCMC, but it suffers a heavy penalty hit for close requests. This is because MCMC can construct an efficient, TDM-like memory schedule that effectively pipelines the delays suffered by the PRE, ACT and CAS commands, while the analysis for open page controllers requires adding the interference on PRE, ACT and CAS: again looking at Table VI, ROC and MCMC have the same *Interference* term, but ROC suffers from an additional *RowInter* term which adds an extra 55-66% latency for close page accesses.

On the other hand, among the three MCs, ReOrder has the least requirements in terms of system configuration, since it does not require extra ranks or complex requestors to rank assignments. It also offers average bandwidth for SRT and large size requestors comparable to ROC. MCMC shows poor performance in terms of provided bandwidth to both SRT and large size requestors, especially for the 2 ranks case, due to poor average memory utilization and close page policy. It also imposes the most constraints on the system by requiring TDM arbitration: a SRT requestor cannot be assigned to more than one slot, meaning than with 8 HRT requestors, no SRT requestor could consume more than 1/8 of the provided throughput under any circumstance. This could be a significant issue for devices such as GPU which can typically saturate memory bandwidth even when running alone. Finally, we need to notice that the evaluation has been conducted using the t_{RTR} (rank-to-rank switching) timing constraint suggested by Ramulator, which is 2 for all devices. For memory modules with larger values of t_{RTR} [Ecco et al. 2016], the performance of both ROC and MCMC would rapidly drop, since the *Interference* term for both MCs cannot be smaller than $t_{BUS} + t_{RTR}$.

5.3.3. Latency and Bandwidth Trade-offs. When a system is characterized by different size of requests or mixed temporal criticality requirements, a trade-off between latency and bandwidth must be considered by the designer as shown in the experiments in Section 5.2.6 and 5.2.7. In general, PMC appears more suitable for handling system with various request sizes because it can be explicitly configured to handle the trade off. RTMem provides the best bandwidth to large requests, but it does so at the cost of increasing latency for small requests compared to AMC by 100%. For SRT requestors, the fixed priority mechanism employed by AMC, Dcmc, ORP, ReOrder and MCMC can strongly limit the bandwidth of SRT requestors depending on the workload of the HRT requestors; in general, no guarantee can be made on minimum bandwidth offered to SRT requestors. Apart from PMC, ROC can also provide guaranteed bandwidth to SRT requestors by allocating them to dedicated ranks, at the cost of increased latency for HRT requestors.

5.3.4. Analytical Bounds vs Simulation Results. We can make three important observations regarding the difference between the analytical latency and the simulated worst case latency in the provided experiments: (1) they are identical for MCs with static command scheduling and close page policy (AMC, PMC, MCMC) because the schedule slot is calculated based on the worst timing constraints in all situations; (2) they have

slight difference for the only MC with dynamic command scheduling and close page (RTMem) because the scheduler can differentiate the type of commands and the location the command targets. The opportunity for the worst case scenario to happen is highly depending on the actual memory request pattern; (3) they have relative large difference for MCs with dynamic command scheduling and open page policy (DCmc, ORP, ReOrder and ROC). We believe this indicates that the analyses for these controllers are fundamentally pessimistic, especially for close page accesses. As noted in Section 5.3.2, the analysis derives the bound by adding together the maximum delays suffered by each command of a request, but this cannot happen in reality: if a request suffers maximum interference on its ACT command, then it should not be able to suffer maximum interference on its CAS command as well (see also [Yun et al. 2015] for an in-depth discussion on the problem). Hence, we believe it is important to focus on deriving tighter analysis for MCs with dynamic command scheduling. An approach based on model-checking is proposed in [Li et al. 2016] and applied to RTMem, but its high computational complexity seem to make it inapplicable to large number of requestors and open page MCs.

6. CONCLUSIONS

The performance of real-time multicore systems can be highly impacted by the behavior of the memory controller. A large number of design proposal [Hassan et al. 2015; Li et al. 2014; Wu et al. 2013; Krishnapillai et al. 2014; Ecco et al. 2014; Jalle et al. 2014] for predictable DRAM controllers have been recently proposed in the literature; however, due to the complexity of comparing multiple controllers on an even ground, there is a significant lack of experimental evaluation. This paper has attempted to bridge such gap by both comparing state-of-the-art predictable controllers based on key configuration parameters, and by proposing an experimental and analytical evaluation based on memory traces generated using EEMBC benchmarks. We believe our results show that there is no universally better controller; rather, the choice of controller should be guided by the desired memory configuration, analytical guarantees and application characteristics.

REFERENCES

B Akesson and K Goossens. 2012. *Memory Controllers for Real-Time Embedded Systems*. Springer.

B Akesson, K Goossens, and M Ringhofer. 2007. Predator: a predictable SDRAM memory controller. In *Hardware/software codesign and system synthesis*. 251–256.

Authors removed for double-blind review. 2015. *Predictable DRAM Controllers Evaluation Code*. Technical Report. University of Waterloo. URL removed for double-blind review.

R Bourgade, C Ballabriga, H Cass, C Rochange, and P Sainrat. 2008. Accurate analysis of memory latencies for WCET estimation. In *16th International Conference on Real-Time and Network Systems*.

L Ecco and R Ernst. 2015. Improved DRAM Timing Bounds for Real-Time DRAM Controllers with Read/Write Bundling. In *Real-Time Systems Symposium*. 53–64.

L Ecco, A Kostrzewa, and R Ernst. 2016. Minimizing DRAM Rank Switching Overhead for Improved Timing Bounds and Performance. In *28th Euromicro Conference on Real-Time Systems*.

L Ecco, S Tobuschat, S Saidi, and R Ernst. 2014. A Mixed Critical Memory Controller Using Bank Privatization and Fixed Priority Scheduling. In *Embedded and Real-Time Computing Systems and Applications*.

M. Gomony, B. Akesson, and K. Goossens. 2013. Architecture and optimal configuration of a real-time multi-channel memory controller. In *Design, Automation Test in Europe Conference Exhibition*. 1307–1312.

S Goossens, B Akesson, and K Goossens. 2013. Conservative Open-page Policy for Mixed Time-Criticality Memory Controllers. In *Design, Automation and Test in Europe Conference*.

M Hassan, H Patel, and R Pellizzoni. 2015. A Framework for Scheduling DRAM Memory Accesses for Multi-Core Mixed-time Critical Systems. In *Real-Time and Embedded Technology and Applications Symposium*.

Liu I, Reineke J, and Lee E. 2010. A PRET Architecture Supporting Concurrent Programs with Composable Timing Properties. In *Signals, Systems and Computers (ASILOMAR)*. 2111–2115.

J Jalle, E Quiñones, J Abella, L Fossati, M Zulianello, and F Cazorla. 2014. A Dual-Criticality Memory Controller (DCmc): Proposal and Evaluation of a Space Case Study. In *Real-Time Systems Symposium*. 207–217.

DDR3 SDRAM JEDEC. 2008. JEDEC JESD79-3B. (2008).

H Kim, D Broman, EA Lee, and M Zimmer. 2015. A predictable and command-level priority-based DRAM controller for mixed-criticality systems. In *21st Real-Time and Embedded Technology and Applications Symposium*. 317–326.

H Kim, D de Niz, B Andersson, M Klein, O Mutlu, and R Rajkumar. 2014. Bounding memory interference delay in COTS-based multi-core systems. In *19th Real-Time and Embedded Technology and Applications Symposium*.

H Kim, J Lee, NB Lakshminarayana, J Lim, and T Pho. 2012. MacSim: Simulator for Heterogeneous Architecture. (2012).

J Kim, M Yoon, Bradford R, and Lui S. 2014. Integrated Modular Avionics (IMA) Partition Scheduling with Conflit-Free I/O for Multicore Avionics Systems. In *Computer Software and Applications Conference*.

Yoongu Kim, Dongsu Han, Onur Mutlu, and Mor Harchol-Balter. 2010. ATLAS: A scalable and high-performance scheduling algorithm for multiple memory controllers. In *6th International Symposium on High-Performance Computer Architecture*. IEEE, 1–12.

Y Kim, M Papamichael, O Mutlu, and M Harchol-Balter. 2010. Thread cluster memory scheduling: Exploiting differences in memory access behavior. In *43rd Annual IEEE/ACM International Symposium on Microarchitecture*. IEEE, 65–76.

Y Krishnapillai, ZP Wu, and R Pellizoni. 2014. ROC: A Rank-switching, Open-row DRAM Controller for Time-predictable Systems. In *Euromicro Conference on Real-Time System*.

Y Li, B Akesson, and K Goossens. 2014. Dynamic Command Scheduling for Real-Time Memory Controllers. In *Euromicro Conference on Real-Time Systems*.

Y Li, B Akesson, K Lampka, and K Goossens. 2016. Modeling and Verification of Dynamic Command Scheduling for Real-Time Memory Controllers. In *Real-Time and Embedded Technology and Applications Symposium (RTAS)*.

Onur Mutlu and Thomas Moscibroda. 2008. Parallelism-aware batch scheduling: Enhancing both performance and fairness of shared DRAM systems. In *ACM SIGARCH Computer Architecture News*, Vol. 36. IEEE Computer Society, 63–74.

omitted for double-blind review. 2016. Technical report for generalizing memory request worst case latency equation. (2016).

Rosenfeld P, Cooper-Balis E, and Jacob B. 2011. DRAMSim2: A Cycle Accurate Memory System Simulator. (2011), 16–19.

M Paolieri, E Quiñones, F J Cazorla, and M Valero. 2009. An analyzable memory controller for hard real-time CMPs. *Embedded Systems Letters, IEEE* (2009).

J Poovey. 2007. Characterization of the EEMBC benchmark suite. *North Carolina State University* (2007).

J Reineke, I Liu, H D. Patel, S Kim, and E Lee. 2011. PRET DRAM Controller: Bank Privatization for Predictability and Temporal Isolation. In *Hardware/software codesign and system synthesis*. 99–108.

L Subramanian, D Lee, V Seshadri, H Rastogi, and O Mutlu. 2016. BLISS: Balancing Performance, Fairness and Complexity in Memory Access Schedyuling. *IEEE Transactions on Parallel and Distributed Systems* (2016).

R. Wilhelm, D. Grund, J. Reineke, M. Schlickling, M. Pister, and C. Ferdinand. 2009. Memory Hierarchies, Pipelines, and Buses for Future Architectures in Time-Critical Embedded Systems. *Computer-Aided Design of Integrated Circuits and Systems* 28, 7 (2009), 966–978.

ZP Wu, Y Krish, and R Pellizzoni. 2013. Worst case analysis of DRAM latency in multi-requestor systems. In *Real-Time Systems Symposium*.

H Yun, P Pellizzoni, P Kumar, and K Valsan. 2015. Parallelism-Aware Memory Interference Delay Analysis for COTS Multicore Systems. In *27th Euromicro Conference on Real-Time Systems*. 184–195.

Wu ZP. 2013. *Worst Case Analysis of DRAM Latency in Hard Real Time Systems*. Master's thesis. University of Waterloo.

A. APPENDIX A

This appendix discussed the analytical latency used in the evaluation for ReOrder controller [Ecco and Ernst 2015]. The authors of [Ecco and Ernst 2015] make a different assumption on the arrival time of CAS commands compared to our work. In particular, in Lemma 1 in [Ecco and Ernst 2015], the authors show that the worst case latency for a RD command happens when a request is served at the beginning of a round, and a new request arrives immediately after the data is transferred as shown in Figure 14. The worst case latency is then derived as follows:

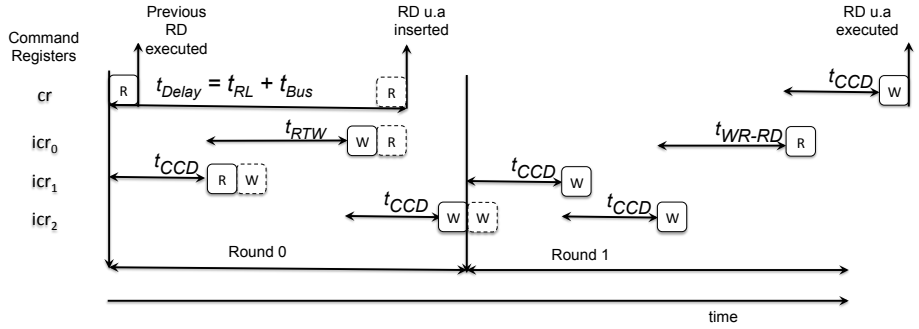

Fig. 14: Worst Case Execution Pattern in Burst Mode

$$L^{RD} = (t_{Round0} - t_{delay})^{+} + t_{Round1}, \tag{1}$$

where:

$$t_{Round0} = (t_{RTW} - 1) + (N - 2) \cdot t_{CCD}, \tag{2}$$

$$t_{Round1} = (N - 1) \cdot t_{CCD} + t_{WR-to-RD}, \tag{3}$$

$$t_{delay} = t_{RL} + t_{Bus}. \tag{4}$$

The $t_{WR-to-RD}$ refers to the delay from a WR command to RD command, equivalent as $t_{WL} + t_{Bus} + t_{WTR}$. However, this is the worst case pattern only when requests are sent in burst, such that one request arrives immediately after the previous one from the same requestor. We argue that if the ready time of the RD command is not known, which is the assumption that we make in this paper, then the worst case access pattern should be derived according to Figure 15, where a RD becomes ready and inserted into the command register just after the type switching within a round. In the example, $Round0$ begins at time 0, with pending WR commands in icr_0 and icr_1. The current bundling-type is Read because the previous round last issued a RD from icr_2. However, once $Round0$ starts, because there is no RD command in the command registers and there are pending WR in icr_0 and icr_1, then the bundling type must switch to WR.The scheduler starts executing WR commands after checking timing constraints. Once the scheduler starts waiting for switching delay (t_{RTW}), cr becomes pending with a RD. Since by this time the bundling type has been switched to WR, the RD request of cr cannot be serviced in the current round. Scheduling $Round0$ is complete once all WR

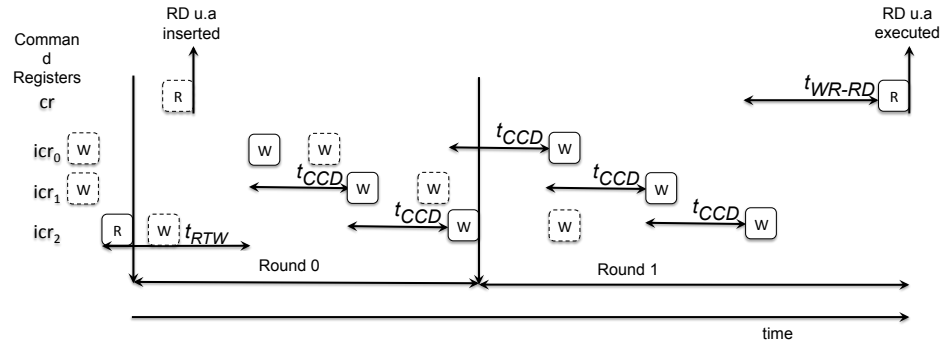

Fig. 15: Worst Case Execution Pattern in Non-Burst Mode

commands are executed. Then a new round $Round1$ starts and resets the served flags without changing the bundling type, as there are still pending WR commands in icr_0, icr_1 and icr_2. After the WR commands are executed, cr will finally be served after the bundling type is switched back to RD. Therefore, the RD under analysis can be blocked twice by each competing command register. The worst case latency for a RD is then given as follows:

$$L^{RD} = t_{Round0} + t_{Round1}, \tag{5}$$

where:

$$t_{Round0} = (t_{RTW} - 1) + (N - 2) \cdot t_{CCD}, \tag{6}$$

$$t_{Round1} = (N - 1) \cdot t_{CCD} + t_{WR-to-RD}. \tag{7}$$

We can observe that the only difference between the two modes is that when requests are executing in burst, the intra-bank timing constraints of a request can be substracted from the delay in $Round0$. However, without such assumption, a command can become ready at any time. Then the worst case we described in Figure 15 can occur, leading to a higher latency. The same assumption and latency computation can also be applied to WR command. This assumption can also be applied when there are multiple ranks. The remaining delay components used to derive the worst-case latency for a request are not affected.

B. EQUATION FORMAT

In this section, we demonstrate the essential steps to convert the request latency equations proposed in the related works to the general expression we showed in this paper. We assume that the total number of requestors N is in power of two, and the DDR device used as an example is DDR3-1600H. Since the latency of a request depends on many conditions, we use the notation $\mathcal{K}(cond)$ such that it equals 1 if $cond$ is satisfied and 0 otherwise

B.1. AMC, RTMem, PMC

Since AMC is analyzed using DDR2 device and many features in DDR3 is not considered such as the $tFAW$ and 8 banks, we apply the analysis in PMC when BC = 1.

RTMem applies dynamic scheduling for commands of request and computes the worst case execution time based on the memory access pattern, but under the worst case access pattern, any consecutive request can access a different row in the same bank, therefore, the latency equation is the same as the PMC. These controllers handle large requests differently, as the described in the following: 1) As AMC applies close-page policy with auto-precharge for every CAS command, a large request that requires accesses to more than 8 banks will be divided into BC several small requests. Each small request can maximally interleave over 8 banks. Therefore, every small request can be delayed by $REQr - 1$ other requestors in the system. 2) PMC first interleaves through 8 banks, and if the request is larger than one access to each bank, another set of access to each bank will be performed. and 3) RTMem has dynamic setting for BI and BC. PMC and RTMem issue all the commands for a large request in a sequence, then switch to schedule another request.

We first demonstrate the latency equation for each bundle in PMC. The bundles must satisfied the following three conditions between two consecutive requests: 1) minimum interval time between two consecutive ACTs to the same bank t_{IB},) the delay caused by ACT commands and 3) the switching delay caused by CAS commands.

We first present the equation for Bundle1. Condition1 must be satisfied by Equation 1 for a read request and 2 for a write

$$t_{ACTrd} = \max(t_{RCD} + \max(t_{BUS} + t_{RTP}) + t_{RP}, t_{RC}) = \max(9 + 6 + 9, 37) = 37 \quad (1)$$

$$t_{ACTwr} = \max(t_{RCD} + t_{WL} + t_{BUS} + t_{WR} + t_{RP}, t_{RC}) = \max(9 + 8 + 4 + 12 + 9, 37) = 42 \quad (2)$$

Condition2 must be satisfied by Equation 3

$$t_{ACTfaw} = \begin{cases} 0 & \text{if } BI < 4; \\ t_{FAW} = 24 & \text{if } 4 \leq BI < 8; \\ 2 \cdot t_{FAW} = 48 & \text{if } BI = 8; \end{cases} \quad (3)$$

Condition3 must be satisfied by Equation 4 for a request request followed by write and Equation 5 for a write request followed by read.

$$t_{CASrd} = t_{RTW} + \begin{cases} t_{RRD} \cdot (BI - 1) = 5 \cdot (BI - 1) + 6 \leq 21 & \text{if } BI \leq 4; \\ t_{FAW} + t_{RRD} \cdot (BI - 5) = 45 & \text{if } BI = 8; \end{cases} \quad (4)$$

$$t_{CASwr} = t_{WL} + t_{BUS} + t_{WTR} + \begin{cases} t_{RRD} \cdot (BI - 1) = 5 \cdot (BI - 1) + 18 \leq 33 & \text{if } BI \leq 4; \\ t_{FAW} + t_{RRD} \cdot (BI - 5) = 57 & \text{if } BI = 8; \end{cases} \quad (5)$$

As a result,

$$t_{Bundle1} = \max(t_{ACTrd}, t_{ACTwr}, t_{ACTfaw}, t_{CASrd}, t_{CASwr}) = \begin{cases} 42 & \text{if } BI \leq 4; \\ 57 & \text{if } BI = 8; \end{cases} \quad (6)$$

The latency for bundle1 can be further presented as $t_{Bundle1} = 42 + 15 \cdot \mathcal{K}(BI = 8)$

The other PMC bundles are presented as the following:

$$t_{Bundle2} = tRCD + tCCD + \begin{cases} t_{RRD} \cdot (BI - 1) = 5 \cdot (BI - 1) + 13 & \text{if } BI \leq 4; \\ t_{RRD} \cdot (BI - 1) + (t_{FAW} - 4 \cdot t_{RRD}) = 5 \cdot (BI - 1) + 17 & \text{if } BI = 8; \end{cases} \quad (7)$$

The latency for bundle2 can be further presented as $t_{Bundle2} = 5 \cdot BI + 8 + 4 \cdot \mathcal{K}(BI = 8)$

$$t_{Bundle3} = t_{CCD} \cdot BI = 4 \cdot BI \tag{8}$$

$$t_{Bundle4} = t_{CCD} \cdot (BI - 1) + \begin{cases} (t_{RTW} - t_{RCD})^+ == 4 \cdot (BI - 1) & \text{if read;} \\ t_{WL} + t_{BUS} + t_{WTR} - t_{RCD} = 4 \cdot (BI - 1) + 9 & \text{if write;} \end{cases} \tag{9}$$

The latency for bundle4 can be further presented as $t_{Bundle4} = 4 \cdot BI + 5$

In summary, the latency for AMC is presented in Equation 10. We assume that a request arrive right after the previous request has data transferred, therefore the request under analysis suffers the auto-precharge delay first before delayed by other requestors.

$$\begin{aligned} L^{AMC} &= t_{Bundle1} \cdot (REQr - 1) + t_{Bundle1} \cdot REQr \cdot (BC - 1) + t_{Bundle1} \\ &= t_{Bundle1} \cdot REQr - t_{Bundle1} + t_{Bundle1} \cdot REQr \cdot BC - t_{Bundle1} \cdot REQr + t_{Bundle1} \\ &= t_{Bundle1} \cdot BC \cdot (REQr - 1) + t_{Bundle1} \cdot (BC - 1) + t_{Bundle1} \\ &= t_{Bundle1} \cdot BC \cdot (REQr - 1) + t_{Bundle1} \cdot BC \end{aligned} \tag{10}$$

The latency equation can be divided into the general expression components

$$\begin{aligned} Interference &= t_{Bundle1} \cdot BC \\ &= (42 + 15 \cdot \mathcal{K}(BI = 8)) \cdot BC \end{aligned} \tag{11}$$

$$basicAccess = t_{Bundle1} \cdot BC \tag{12}$$

The latency for PMC and RTMem can be presented as a combination of the four bundles based on the BI and BC. Because each request will be complete, and the request under analysis also takes a slot to process data, therefore, the latency from request arrive to data transfer is: The length of one request slot can be computed in Equation 13

$$\begin{aligned} D^{PMC} &= t_{Bundle1} \cdot \mathcal{K}(BC = 1) + \\ &\quad \mathcal{K}(BC > 1) \cdot (t_{Bundle2} + t_{Bundle3} \cdot (BC - 2) + t_{Bundle4}) \\ &\quad = \mathcal{K}(BC = 1) \cdot (42 + 15 \cdot \mathcal{K}(BI = 8)) + \\ &\quad \mathcal{K}(BC > 1) \cdot (5 \cdot BI + 8 + 4 \cdot \mathcal{K}(BI = 8) + (4 \cdot BI) \cdot (BC - 2) + 4 \cdot BI + 5) \\ &\quad = \mathcal{K}(BC = 1) \cdot (42 + 15 \cdot \mathcal{K}(BI = 8)) + \\ &\quad \mathcal{K}(BC > 1) \cdot (4 \cdot \mathcal{K}(BI = 8) + (4 \cdot BC + 1) \cdot BI + 13) \end{aligned} \tag{13}$$

$$L^{PMC} = (REQr - 1) \cdot D^{PMC} + D^{PMC} \tag{14}$$

The latency equation can be divided into the general expression components

$$Interference = BasicAccess = D^{PMC} \tag{15}$$

B.2. MCMC

Because MCMC is scheduled with non-work conserving TDM. The length of the slots must cover all the intra-bank timing constraints of re-activating a same bank and the

CAS switching delay between banks in the same rank. The slot can be computed based on the following three conditions: 1) reactivate the same bank; 2) CAS switching delay covered by rank switching; and 3) the minimum rank switching delay.

$$slot = \max \begin{cases} \lceil \frac{D^{sameBank}}{N} \rceil = \lceil \frac{42}{N} \rceil \\ \frac{\max(t_{RTW}, t_{WL}+t_{BUS}+t_{WTR})}{R} = \frac{18}{R} \\ \max(t_{BUS}+t_{RTR}, t_{RL}-t_{WL}+t_{BUS}+t_{RTR}, t_{WL}-t_{RL}+t_{BUS}+t_{RTR}) = \max(6,7,5) \end{cases} \tag{16}$$

If the request is large and divided into several small request of one memory access, then the latency equation can be represented in Equation 17. The first access of a large request can miss its own slot, and the following small request does not miss the slot, and delayed by all other requestors.

$$\begin{aligned} L_{LrgReq} &= slot \cdot REQ \cdot BC + t_{RCD} + t_{RL} + t_{BUS} \\ &= slot \cdot REQr \cdot R \cdot BC + 22 \\ &= slot \cdot R \cdot BC \cdot (REQr - 1) + slot \cdot R \cdot BC + 22 \end{aligned} \tag{17}$$

The interference and basic access can be represented as

$$Interference = slot \cdot R \cdot BC \tag{18}$$

$$BasicAccess = slot \cdot ((R-1) \cdot BC + 1) + 22 \tag{19}$$

B.3. ORP

We will refer to the equations in the original paper and convert into the expression to better show the insight of parameters change. The notation Eq represents the equations used in the original paper. According to Eq3, the latency of PRE command can be presented in the following

$$L^{PRE} = REQr - 1 = 1 \cdot (REQr - 1) \tag{20}$$

According to Eq5, the latency of ACT command can be presented in the following

$$\begin{aligned} L^{ACT} &= (t_{FAW} - 4 \cdot t_{RRD}) + \lfloor \frac{REQr-1}{4} \rfloor \cdot t_{FAW} + ((REQr-1) mod 4) \cdot t_{RRD}) \\ &= t_{RRD} \cdot (REQr - 1) + \lceil \frac{REQr-1}{4} \rceil \cdot (t_{FAW} - 4 \cdot t_{RRD}) \\ &= t_{RRD} \cdot (REQr - 1) + (REQr + 2) \\ &= (t_{RRD} + 1) \cdot (REQr - 1) + 3 \\ &= 6 \cdot (REQr - 1) + 3 \end{aligned} \tag{21}$$

Since we are assuming that number of requestors is in the power of 2. Therefore $\lfloor \frac{N-1}{2} \rfloor = \frac{N-2}{2}$ and $\lceil \frac{N-1}{2} \rceil = \frac{N}{2}$. According Equation 10 and 11 in [Wu et al. 2013], the latency is shown as the time from when the CAS command becomes ready and the data is transferred. In our expression, the interference represents the time from the

CAS is ready until it is issued. Therefore, tR/WL is subtracted from both equations.

$$\begin{aligned} L^{RD} &= \lfloor \frac{REQr-1}{2} \rfloor \cdot (t_{WTR} + t_{RTW}) + \lceil \frac{REQr-1}{2} \rceil \cdot (t_{WL} + t_{BUS}) + (t_{WTR} + t_{RL} + t_{BUS}) \\ &= \frac{REQr-2}{2} \cdot (13) + \frac{REQr}{2} \cdot (12) + (19) \\ &= 12.5 \cdot (REQr-1) + 18.5 < 13 \cdot (REQr-1) + 19 \end{aligned} \tag{22}$$

$$\begin{aligned} L^{WR} &= \lfloor \frac{N}{2} \rfloor \cdot (t_{RTW} + t_{WTR}) + \lceil \frac{N}{2} \rceil \cdot (t_{WL} + t_{BIS}) \\ &= \frac{N}{2} \cdot (13) + \frac{N}{2} \cdot (12) \\ &= 12.5 \cdot (REQr-1) + 12.5 < 13 \cdot (REQr-1) + 12.5 \end{aligned} \tag{23}$$

The delay before the request can be processed as the following

$$D^{Close} = t_{WR} + t_{RP} + t_{RCD} = 30 \tag{24}$$

$$D^{OpenRD} = t_{WTR} = 6 \tag{25}$$

$$D^{OpenWR} = 0 \tag{26}$$

Convert into the general expression, read request has larger latency comparing to write. If a large request requires BC number of memory accesses. The latency can be represented in Equation 27 for open request and 28 for a close request.

$$\begin{aligned} L_{OpenRD} &= D^{OpenRD} + L^{RD}) \cdot BC \\ &= 6 + (13 \cdot (REQr-1) + 19) \cdot BC \\ &= 13 \cdot BC \cdot (REQr-1) + 19 \cdot BC + 6 \end{aligned} \tag{27}$$

$$\begin{aligned} L_{CloseRD} &= D^{Close} + L^{PRE} + L^{ACT} + L^{RD} \cdot BC \\ &= 30 + (6+1) \cdot (REQr-1) + 3 + 13 \cdot BC \cdot (REQr-1) + 19 \cdot BC \\ &= 27 + 7 \cdot (REQr-1) + 13 \cdot BC \cdot (REQr-1) + 19 \cdot BC + 6 \end{aligned} \tag{28}$$

Based on the latency equation, the equation can be broken into the following terms

$$Interference = 13 \cdot BC \tag{29}$$

$$RowInter = 7 \tag{30}$$

$$BasicAccess = 19 \cdot BC + 6 \tag{31}$$

$$RowAccess = 27 \tag{32}$$

B.4. DCmc

According to Equations 2,4,11,12, 13 and 23 in [Jalle et al. 2014], we can express the components as the following

$$D^{PRE} = 1 \tag{33}$$

$$D^{CAS} = \max(tWL + tBus + tWTR, tRTW) = \max(18, 7) = 18 \tag{34}$$

$$D^{ACT} = \max(tRRD, tFAW - 3 \cdot tRRD) = \max(5, 24 - 15) = 9 \tag{35}$$

$$L^{CloseRD} = tRP + tRCD + tRL + tBus = 31 \tag{36}$$

$$L^{OpenRD} = tRL + tBus = 13 \tag{37}$$

The latency equation for an open and a close request can be presented in Equations 38 and 39.

$$\begin{aligned} L^{DCmc}_{Open} &= (L^{OpenRD} + (REQr - 1) \cdot (D^{PRE} + D^{ACT} + D^{CAS})) \cdot BC \\ &= 13 \cdot BC + 28 \cdot (REQr - 1) \cdot BC \end{aligned} \tag{38}$$

$$\begin{aligned} L^{DCmc}_{Close} &= L^{CloseRD} + L^{OpenRD} \cdot (BC - 1) + (REQr - 1) \cdot (D^{PRE} + D^{ACT} + D^{CAS}) \cdot BC \\ &= 31 + 13 \cdot (BC - 1) + 28 \cdot (REQr - 1) \cdot BC \\ &= 18 + 13 \cdot BC + 28 \cdot (REQr - 1) \cdot BC \end{aligned} \tag{39}$$

Convert into the general expression

$$Interference = (D^{PRE} + D^{ACT} + D^{CAS}) \cdot BC = 28 \cdot BC \tag{40}$$

$$RowInter = 0 \tag{41}$$

$$BasicAccess = L^{OpenRD} = 13 \cdot BC \tag{42}$$

$$RowAccess = L^{CloseRD} - L^{OpenRD} = 18 \tag{43}$$

B.5. ROC

The bus conflict delay was defined in Equation 3 in [Krishnapillai et al. 2014] as the following

$$\begin{aligned} \alpha_{PA}(K) &= K + \lceil \frac{K}{t_{BUS} - 1} \rceil \\ &= K + \frac{K}{t_{BUS} - 1} + \frac{t_{BUS} - 2}{t_{BUS} - 1} \\ &= K + \frac{K}{3} + \frac{2}{3} = \frac{4}{3} \cdot K + \frac{2}{3} \end{aligned} \tag{44}$$

Based on this simplified equation, Equation 4 in [Krishnapillai et al. 2014] can be computed as Equation 45.

$$\begin{aligned} t_{IP} &= \alpha_{PA}(R \cdot REQr) - 1 \\ &= (R \cdot REQr) \cdot \frac{4}{3} + \frac{2}{3} - 1 \\ &= \frac{4 \cdot R}{3} \cdot (REQr - 1) + \frac{4 \cdot R - 1}{3} \end{aligned} \tag{45}$$

The worst case value for t_{IA} of Equation 5 in [Krishnapillai et al. 2014] can be broken into the following format

$$\begin{aligned}\delta_{IA} &= \alpha_{PA}(R) - 1 \\ &= \frac{4R}{3} + \frac{2}{3} - 1 \\ &= \frac{4 \cdot R - 1}{3}\end{aligned} \tag{46}$$

$$\begin{aligned}t_{IA} &= \lceil \frac{REQr - 1}{4} \rceil \cdot (t_{FAW} - 4 \cdot t_{RRD}) + (REQr - 1) \cdot t_{RRD} + REQr \cdot \delta_{IA} \\ &= \lceil \frac{REQr - 1}{4} \rceil \cdot (24 - 5 \cdot 4) + 5 \cdot (REQr - 1) + REQr \cdot \frac{4 \cdot R - 1}{3} \\ &= 4 \cdot (\frac{REQr - 1}{4} + \frac{3}{4}) + 5 \cdot (REQr - 1) + \frac{4 \cdot R - 1}{3} \cdot REQr \\ &= (REQr - 1) + 3 + 5 \cdot (REQr - 1) + \frac{4 \cdot R - 1}{3} \cdot (REQr - 1) + \frac{4 \cdot R - 1}{3} \\ &= (1 + 5 + \frac{4 \cdot R - 1}{3}) \cdot (REQr - 1) + 3 + \frac{4 \cdot R - 1}{3} \\ &= (\frac{4 \cdot R + 17}{3}) \cdot (REQr - 1) + \frac{4 \cdot R + 8}{3}\end{aligned} \tag{47}$$

We then put the value of timing constraints into Equation 7, 8, 9 and 10 to get the following equations

$$\begin{aligned}t_{WRD} &= \max(R \cdot (t_{BUS} + t_{RTR}), t_{WTR} + t_{RL} + 2 \cdot t_{BUS} + t_{RTR} - 1) \\ &= \max(R \cdot 6, 6 + 9 + 8 + 2 - 1) \\ &= \max(6 \cdot R, 24)\end{aligned} \tag{48}$$

Since $R \leq 4$ for ROC, then $6 \cdot R \leq 24$, therefore, $t_{WRD} = 24$.

$$\begin{aligned}t_{RWD} &= \max(R \cdot (t_{BUS} + t_{RTR}), t_{RTW} + t_{WL} - t_{RL} + t_{BUS} + t_{RTR} - 1) \\ &= \max(R \cdot 6, 7 + 8 - 9 + 4 + 2 - 1) \\ &= \max(6 \cdot R, 11)\end{aligned} \tag{49}$$

Since $R \geq 2$ for ROC, then it always hold that $6 \cdot R \geq 11$, therefore, $t_{RWD} = 6 \cdot R$.

$$\begin{aligned}t_{RD} &= \max(t_{RL} + t_{BUS} - 1 + R \cdot (t_{BUS} + t_{RTR}), t_{WTR} + t_{RL} + 2 \cdot t_{BUS} + t_{RTR} - 1) \\ &= \max(12 + R \cdot 6, 6 + 9 + 8 + 2 - 1) \\ &= \max(12 + 6 \cdot R, 24)\end{aligned} \tag{50}$$

Since $R \geq 2$ for ROC, then it always hold that $12 + 6 \cdot R \geq 24$, therefore, $t_{RD} = 12 + 6 \cdot R$.

$$t_{WD} = t_{RL} + t_{BUS} - 1 + R \cdot (t_{BUS} + t_{RTR}) = 12 + 6 \cdot R \tag{51}$$

As we assume that REQr is even, $\lceil \frac{REQr-1}{2} \rceil = \frac{REQr}{2}$ and $\lfloor \frac{REQr-1}{2} \rfloor = \frac{REQr-2}{2}$. With this simplified equations, we can derive the latency for read and write as the following

$$\begin{aligned} t^{Write} &= \lceil \frac{REQr-1}{2} \rceil \cdot t_{RWD} + \lfloor \frac{REQr-1}{2} \rfloor \cdot t_{WRD} + t_{RD} \\ &= \frac{REQr}{2} \cdot (6 \cdot R) + \frac{REQr-2}{2} \cdot (24) + 12 + 6 \cdot R \\ &= 3 \cdot R \cdot REQr + 12 \cdot (REQr - 2) + 12 + 6 \cdot R \\ &= 3 \cdot R \cdot (REQr - 1) + 3R + 12 \cdot (REQr - 1) - 12 + 12 + 6 \cdot R \\ &= (3 \cdot R + 12) \cdot (REQr - 1) + 9 \cdot R \end{aligned} \tag{52}$$

$$\begin{aligned} t^{Read} &= \lceil \frac{REQr-1}{2} \rceil \cdot t_{RWD} + \lfloor \frac{REQr-1}{2} \rfloor \cdot t_{WRD} + t_{RD} \\ &= \frac{REQr}{2} \cdot 24 + \frac{REQr-2}{2} \cdot (6 \cdot R) + 12 + 6 \cdot R \\ &= 12 \cdot REQr + (REQr - 2) \cdot 3 \cdot R + 12 + 6 \cdot R \\ &= 12 \cdot (REQr - 1) + 12 + 3 \cdot R \cdot (REQr - 1) - 3 \cdot R + 12 + 6 \cdot R \\ &= (3 \cdot R + 12) \cdot (REQr - 1) + 3 \cdot R + 24 \end{aligned} \tag{53}$$

If the request is divided into BC small request, then the read (read latency is larger) latency can be presented in

$$\begin{aligned} L^{Read}_{Open} &= t_{WTR} + (t^{Read}) \cdot BC \\ &= 6 + ((3 \cdot R + 12) \cdot (REQr - 1) + 3 \cdot R + 24) \cdot BC \\ &= 6 + (3 \cdot R + 12) \cdot BC \cdot (REQr - 1) + (3 \cdot R + 24) \cdot BC \end{aligned} \tag{54}$$

$$\begin{aligned} L^{Read}_{Close} &= t_{WR} + t_{RP} + t_{RCD} + t_{IP} + t_{IA} + (L^{Read}_{Open} - t_{WTR}) \\ &= (30 - t_{WTR}) + \frac{4 \cdot R}{3} \cdot (REQr - 1) + \frac{4 \cdot R - 1}{3} + (\frac{4 \cdot R + 17}{3}) \cdot (REQr - 1) + \frac{4 \cdot R + 8}{3} + L^{Read}_{Open} \\ &= (\frac{4 \cdot R}{3} + \frac{4 \cdot R + 17}{3}) \cdot (REQr - 1) + (\frac{4 \cdot R - 1}{3} + \frac{4 \cdot R + 8}{3}) + L^{Read}_{Open} \\ &= (\frac{8 \cdot R + 17}{3}) \cdot (REQr - 1) + \frac{8 \cdot R + 7}{3} + 24 + L^{Read}_{Open} \end{aligned} \tag{55}$$

Based on the latency equation, we can break the equation into the individual terms

$$Interference = (12 + 3R) \cdot BC \tag{56}$$

$$RowInter = \frac{8 \cdot R + 17}{3} < (3R + 6) \tag{57}$$

$$BasicAccess = (3 \cdot R + 24) \cdot BC + 6 \tag{58}$$

$$RowAccess = \frac{8 \cdot R + 7}{3} + 24 = 3R + 27 \tag{59}$$

B.6. rankReOrder

Summarize the switching time between different ranks for CAS commands $t_{WRRD} = t_{WL} - t_{RL} + t_{BUS} + t_{RTR} = 8 - 9 + 4 + 2 = 5$

$t_{WRWR} = t_{BUS} = 4$
$t_{RDWR} = t_{RL} - t_{WL} + t_{BUS} + t_{RTR} = 7$
$t_{RDRD} = t_{BUS} + t_{RTR} = 6$
$WTR = t_{WL} + t_{BUS} + t_{WTR} = 18$

Considering the non-burst mode demonstrated in Appendix A, the command can arrive 1 CYCLE after the switching to different type round (tRTW should be considered) and suffer 2(REQr-1) write commands

$$\begin{aligned}
L^{CAS} &= switch_i + ccds_i + switch_{i+1} + ccds_{i+1} - 1 \\
&= \max(t_{RDRD} \cdot (R-1) + t_{RDWR}, tRTW) + (R-1) \cdot t_{WRWR} + R \cdot t_{CCD} \cdot (REQr-2) + \\
&\max(t_{WRWR} \cdot (R-1) + t_{WRRD}, WTR) + (R-1) \cdot t_{RDRD} + (R-1) \cdot t_{CCD} \cdot (REQr-2) + t_{CCD} \cdot (REQr - \\
&= 6 \cdot (R-1) + 7 + 4 \cdot (R-1) + 4R \cdot (REQr-2) + \\
&18 + 6 \cdot (R-1) + 4(R-1) \cdot (REQr-2) + 4 \cdot (REQr-1) - 1 \\
&= 16(R-1) + 8R \cdot (REQr-2) - 4 \cdot (REQr-2) + 4 \cdot (REQr-1) + 24 \\
&= 16R - 16 + 8R \cdot (REQr-1) - 8R - 4REQr + 8 + 4REQr - 4 + 24
\end{aligned} \tag{60}$$

Since $\max(t_{RDRD} \cdot (R-1) + t_{RDWR}, tRTW) = t_{RDRD} \cdot (R-1) + t_{RDWR} = 6 \cdot (R-1) + 7$ and $\max(t_{WRWR} \cdot (R-1) + t_{WRRD}, WTR) = WTR = 18$, remains the same for less than 4 ranks. The equation applies for 1, 2 and 4 ranks.

The other interference for close request comes from PRE and ACT. For the bus conflict, we assume that PRE and ACT can suffer 1 cycle.

$$\begin{aligned}
L^{PRE} &= (REQr-1) \cdot (1+1) + (R-1) \cdot REQr = 2(REQr-1) + R \cdot REQr - REQr \\
&= REQr - 2 + R \cdot REQr = (R+1)(REQr-1) + (R-1)
\end{aligned} \tag{61}$$

$$\begin{aligned}
L^{ACT} &= \lceil \frac{REQr-1}{4} \rceil \cdot (t_{FAW} - 4 \cdot t_{RRD}) + (REQr-1) \cdot (t_{RRD}+1) + (R-1) \cdot REQr \\
&= (t_{RRD} + 1 + R - 1) \cdot (REQr-1) + (R-1) + 4 \cdot (\frac{REQr-1}{4} + \frac{3}{4}) \\
&= (t_{RRD} + R) \cdot (REQr-1) + (R-1) + (REQr-1) + 3 \\
&= (t_{RRD} + R + 1) \cdot (REQr-1) + (R-1) + 3 \\
&= (6+R) \cdot (REQr-1) + (R+2)
\end{aligned} \tag{62}$$

We first combine the equations to show the worst case latency for a read request with BI = 1 and BC = $\frac{requestsize}{dataperaccess}$.

$$\begin{aligned}
L^{OpenR} &= (L^{CAS} + t_{RL} + t_{BUS}) \cdot BC \\
&= 8R \cdot BC(REQr-1) + (8R+25) \cdot BC
\end{aligned} \tag{63}$$

$$\begin{aligned}
L^{CloseR} &= t_{WR} + L^{PRE} + t_{RP} + L^{ACT} + t_{RCD} + L^{OpenR} \\
&= 30 + (R+1)(REQr-1) + (R-1) + (6+R) \cdot (REQr-1) + (R+2) + L^{OpenR} \\
&= (7+2R) \cdot (REQr-1) + (30 + 2R + 1) + L^{OpenR}
\end{aligned} \tag{64}$$

Convert into the general expression

$$Interference = 8R \cdot BC \tag{65}$$

$$RowInter = 7 + 2R \tag{66}$$

$$BasicAccess = (8R + 25) \cdot BC \tag{67}$$

$$RowAccess = 31 + 2R \tag{68}$$