A Memory Scheduling Infrastructure for Multi-core Systems with Re-programmable Logic

Denis Hoornaert
Technical University of Munich, Germany

Shahin Roozkhosh
Boston University, USA

Renato Mancuso
Boston University, USA

Abstract

The sharp increase in demand for performance has prompted an explosion in the complexity of modern multi-core embedded systems. This has lead to unprecedented temporal unpredictability concerns in Cyber-Physical Systems (CPS). On-chip integration of programmable logic (PL) alongside a conventional Processing Systems (PS) in modern Systems-on-Chip (SoC) establishes a genuine compromise between specialization, performance, and re-configurability. In addition to typical use-cases, it has been shown that the PL can be used to observe, manipulate, and ultimately manage memory traffic generated by a traditional multi-core processor.

This paper explores the possibility of PL-aided memory scheduling by proposing a Scheduler In-the-Middle (SchIM). We demonstrate that the SchIM enables transaction-level control over the main memory traffic generated by a set of embedded cores. Focusing on extensibility and reconfigurability, we put forward a SchIM design covering two main objectives. First, to provide a safe playground to test innovative memory scheduling mechanisms; and second, to establish a transition path from software-based memory regulation to provably correct hardware-enforced memory scheduling. We evaluate our design through a full-system implementation on a commercial PS-PL platform using synthetic and real-world benchmarks.

2012 ACM Subject Classification Computer systems organization → Real-time system architecture

Keywords and phrases MPSoC, FPGA, Memory Scheduling

Digital Object Identifier 10.4230/LIPIcs.CVIT.2016.23

Funding *Denis Hoornaert*: Denis Hoornaert was supported by an Alexander von Humboldt Professorship endowed by the German Federal Ministry of Education and Research.
Renato Mancuso: The material presented in this paper is based upon work supported by the National Science Foundation (NSF) under grant number CCF-2008799. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the NSF.

1 Introduction

It is undeniable that the massive increase in expectation on the performance of next-generation cyber-physical systems has deeply impacted the way we design modern embedded and real-time systems. High-resolution, high-bandwidth sensors such as lidars, and depth cameras on the one hand, and data-intensive processing workload such as machine-learning applications on the other hand, have exacerbated the push for high-performance embedded platforms. Following this performance *moving target*, chip manufactures have significantly scaled up clock speeds, CPU count, and heterogeneity. For instance, the on-chip integration of powerful graphic processing units (GPUs) has been the characterizing factor in the NVIDIA Tegra series of embedded systems-on-a-chip (SoC).

In this context, an embedded architectural paradigm that is surging in popularity among manufacturers, researchers, and industry practitioners is the PS-PL organization. This class of embedded platforms integrates on the same die (1) traditional full-speed embedded CPUs and (2) programmable logic constructed using field-programmable gate array (FPGA) technology. This organization naturally defines two macro-domains, namely the Processing System (PS) and the Programmable Logic (PL), hence the name. PS-PL platforms establish a good trade-off between specialization, raw performance, and mission-specific re-configurability. The current generation of commercially available PS-PL platforms is dominated by ARM-based products offered by, most notably, Intel [12] and Xilinx [38]. A pilot large-scale, high-performance PS-PL system is the Enzian platform [3] being rolled out by ETH Zurich[1]. Furthermore, a RISC-V-based solution has been recently made available by Microsemi with their PolarFire SoC [18].

From a real-time perspective, the co-existence of traditional CPUs and a tightly-coupled block of PL has more profound implications than expected. Clearly, it is possible to define custom accelerators in PL and to relieve the main CPUs of some of the heavy data-processing workload. However, more interestingly, recent studies have highlighted the possibility of using the PL also as a way to manage the memory traffic originated from the main CPUs [13, 29]. Such a possibility opens the doors for memory traffic inspection and control at the level of individual transactions; which in turn promises to unlock provable determinism for the real-time workload.

In this paper, we embrace the concept of PL-aided memory traffic management and propose an infrastructure to develop, test and evaluate memory scheduling policies. Specifically, we propose a component, called the Scheduler In-the-Middle—or SchIM, for short—that can be instantiated in the PL to enforce a set of configurable scheduling policies on individual memory transactions generated by the CPUs in the PS.

The overarching goal of the proposed SchIM is twofold. First, we want to provide a playground for researches to test promising novel memory scheduling ideas for multi-core platforms, much like LITMUS^{RT} [7] fostered research on CPU scheduling techniques. Second, we want our SchIM to act as an intermediate stepping stone for industrial applications where strong determinism over memory performance is required. The SchIM can be used to analyze the behavior of realistic workload in a multitude of what-if memory management use-cases. We note that such kind of analysis was previously possible only through full-system simulation or by synthesizing the entire SoC on FPGA—that is, with a soft-core implementation.

In short, this paper makes the following contributions. (1) We demonstrate that a configurable module could be interposed between the cores and the memory controller to perform transaction-level scheduling in commercial PS-PL platforms; (2) we propose a design for a memory scheduling infrastructure that focuses on extensibility and runtime reconfigurability; (3) we address important issues to correctly account and regulate CPU-generated traffic when a shared last-level cache is present; (4) we design and implement two pilot memory scheduling policies as a proof-of-concept on the potential of our SchIM; and (5) we perform a full system integration and implementation on a commercial PS-PL embedded platform to evaluate the behavior of the SchIM with synthetic and realistic workload.

[1] Also see `http://enzian.systems/`

2 Related Work

There is a broad consensus that memory resources represent the main performance bottleneck in modern multi-core processors. The observation has sparked a host of research works addressing the problem from multiple angles [17]. In this context, the works representing the inspiration for our SchIM fall in two macro-categories, namely **hardware-based** and **software-based** techniques for main memory traffic management.

The first category includes a large body of works aimed at achieving better and/or more predictable performance by advancing novel hardware redesigns. The works in [22–24] strive to construct high-performance and fair memory schedulers. The addition of software-controlled memory deadlines and transactional semantics where explored in [33] and [10], respectively. Next, the work by Åkesson et al. [1, 2] and Paolieri et al. [25] attains timing predictability through careful scheduling of SDRAM commands. Finally, the MEDUSA DRAM controller [9, 34] implements a two-tiers scheduler at the DRAM controller to ensure predictability when accessing memory areas where access time strongly impact application performance. Finally, the hardware designs proposed in [8, 26, 43] put their emphasis on main memory bandwidth partitioning; clever dynamic pipelining is further explored in [20] to better balance average performance and determinism.

Among the software-based techniques are the mechanisms that stemmed from MemGuard, originally proposed in [42] and that rely on broadly available performance counters to regulate the bandwidth extracted by individual CPUs. Later extensions to jointly consider regulation and cache partitioning [39] and to expose control over memory bandwidth as a lockable resource [40] were proposed. Software-based memory throttling has also been implemented at the hypervisor-level [21, 30]. Remarkably, the work in [30] combines regulation mechanisms for CPU and embedded accelerators through the ARM QoS extensions [4].

In addition to the two categories surveyed above, perhaps the most closely related works are those that explored memory isolation techniques in PS-PL platforms. The work in [11] demonstrated that the PL-side can be used to define private memory storage, control, and bus units to strongly isolate high-criticality workload. A number of techniques developed as part of the FRED framework [6] put an emphasis on memory traffic arbitration and management for in-PL accelerators [27, 28]. The AXI HyperConnect [27] is perhaps the component most similar to the SchIM in terms of high-level design. However, both are substantially different as the SchIM is designed to manage embedded CPUs' memory traffic.

Compared to the literature reviewed above, what sets this work apart are the following aspects. (1) Our SchIM applies to existing PS-PL commercial systems without introducing any hardware modification; (2) it allows management in the PL of memory traffic originated by the embedded CPUs residing in the PS; (3) it provides the framework to test the feasibility and performance of custom memory scheduling policies; and (4) it is designed such that multiple schedulers can coexist, be activated, and configured at runtime.

3 Background Concepts

In this section, we introduce some fundamental concepts necessary to understand the overall system design and the class of platforms targeted by this work.

3.1 Hybrid Multi-Core Platforms with Programmable Logic

This work targets the aforementioned class of embedded multi-core platforms with programmable logic—i.e., PS-PL platforms. In such platforms, the PS encompasses a multi-core processor with a multi-level cache hierarchy and a main memory (DRAM) controller. A simplified block diagram for a reference PS-PL organization is illustrated in Fig. 1. The figure considers a platform with four CPUs denoted as C_0, C_1, C_2, and C_3.

A key feature in PS-PL platforms is the presence of high-performance communication channels between the two domains. These come in the form of data exchange interfaces and interrupt lines. Data exchange channels follow a master-slave paradigm. Specifically, high-performance masters (HPM, Fig. 1①) and high-performance slaves (HPS, Fig. 1②) send and receive transactions to and from the PL, respectively. Additionally, there exist programmable interrupt request (IRQ) lines (see Fig. 1③) that can be driven by the PL and are connected to the interrupt controller (Fig. 1④) inside the PS. As we discuss in Section 5.7, the presence of PS-PL interrupt lines is crucial to building PL-assisted memory traffic regulation.

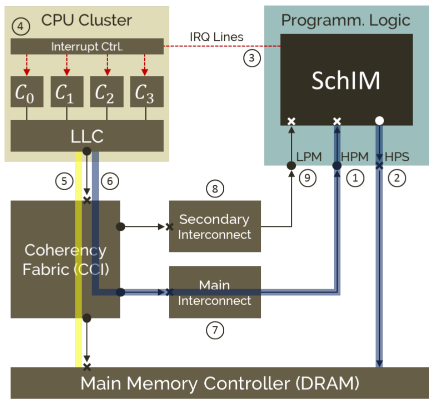


Figure 1 PS-PL interconnect block diagram.

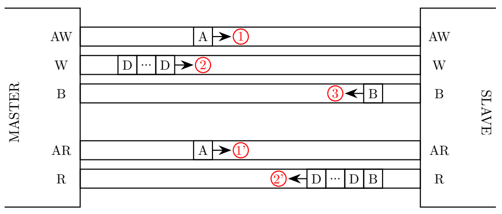
Note also that there might exist PS-PL data ports that are routed through a secondary interconnect (Fig. 1⑧). These can generally sustain less throughput compared to HPS ports; hence we refer to them as low-performance masters (LPM, Fig. 1⑨). LPM ports are useful to perform memory-mapped configuration of PL modules.

3.2 Programmable Logic In-the-Middle

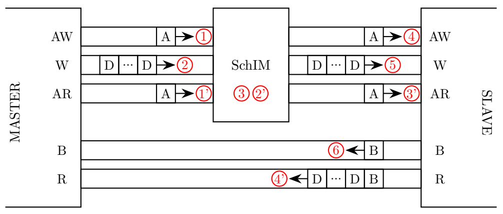
In this work, we leverage the ability to route main memory traffic originated by the CPUs through the PL. This technique is known as Programmable Logic In-the-Middle, or PLIM for short. PLIM was originally proposed in [29]. To fully grasp how PLIM can be achieved, one needs to understand how memory accesses are routed in PS-PL platforms.

Any CPU-generated memory access that results in an LLC miss is routed directly to main memory if its physical address falls within the aperture, say the address range $[A, B]$ handled by the DRAM controller. We refer to this as the *normal route*, depicted in Fig. 1⑤ and highlighted in yellow.

Conversely, generic memory access resulting from an LLC cache miss will be sent on an HPM port if the corresponding physical address falls within another range, say $[C, D]$. One can then insert (1) a lightweight layer of virtualization to map all the physical addresses of a guest OS to the PL, i.e., to fall in the range $[C, D]$; and (2) an address translator in the PL that re-bases request physical addresses to access main memory and relays back the data payload to the requesting CPU(s). In other words, one can find a constant k such that $C = A + k$. Then, the translator in the PL, upon receiving any request at address $x \in [C, D]$ will issue a main memory request at the address $(x - k)$ through the HPS port and provide the response to the CPU. The PLIM technique introduces a secondary memory route for reaching the DRAM, called the *PL loop-back*, or simply *loop-back*, which is highlighted in blue in Fig. 1⑥. Memory transactions on the loop-back route typically traverse the main interconnect, as depicted in Fig. 1⑦. The advantage of PLIM is that transactions on the loop-back route can be inspected, blocked, re-routed, and in general managed by custom re-programmable logic. Importantly, switching from the direct to the loop-back route can



(a) Standard AXI communication Scheme



(b) AXI communication mediated with SchIM.

be done dynamically at runtime so that the overhead of PLIM can be avoided if deemed detrimental for the application under analysis.

In this paper, we leverage the PLIM approach to perform memory scheduling, hence, we call our module the Scheduler In-the-Middle, or SchIM for short.

3.3 Advanced eXtensible Interface (AXI)

The vast majority of PS-PL platforms currently available are ARM-based. This is also the case for the platform we used for our evaluation, namely the Xilinx Zynq UltraScale+ MPSoC. Thus, we briefly introduce the communication protocol used for on-chip communication in ARM-based SoCs, namely the Advanced eXtensible Interface (AXI). The AXI is an open specification bus protocol [5] used for high-bandwidth data exchanges between on-chip subsystems — such as cache controllers, memory controllers, DMAs, PL modules. It is also used in the PS-PL platforms of reference to exchange data on the HPM and HPS ports.

The AXI protocol is based on the master-slave duality. A master AXI interface can initiate transactions toward a connected slave interface. The latter responds master-initiated requests. Masters and the slaves communicate with each other through five different channels named AW (address write), W (write), B (write acknowledgment), AR (address read) and R (read), as illustrated in Fig. 2a.

A write transaction begins with an address phase ① where the channel AW is used to transmit the transaction's meta-data, such as the destination address, the transaction ID, and the cacheability attributes the type/length of the burst, and so on. Upon completing this phase, follows the data phase ②, which consists of the transmission of the data payload to be written through the W channel. The response phase ③ concludes a successful write transaction and occurs on the B channel.

The transmission of a read transaction is carried out in a similar way. The address phase ①’ is transmitted through the equivalent AR channel and is directly followed by the data phase ②’. A response initiated by the slave follows where the read data is transferred over the R channel. The protocol is asynchronous because different phases of different transactions can interleave on any AXI bus segment. Hence, multiple outstanding transactions can be emitted by a single master and the receipt of out-of-order responses is possible.

4 Design Goals and Overview

In this section, we introduce the proposed SchIM design and describe the overarching goals of this work. We then provide a bird's-eye view of the SchIM organization and principles of operation.

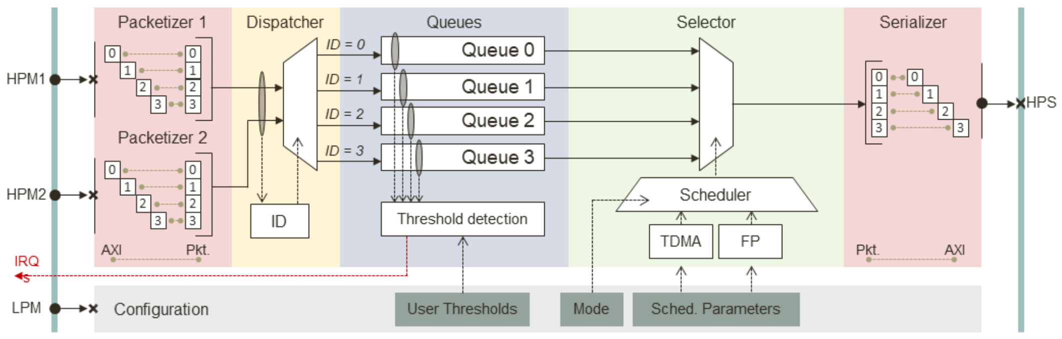


Figure 3 SchIM internal organization connected to the PS via the HPM, LPM and HPS ports.

4.1 Design Goals

As briefly surveyed in Section 2, there have been numerous proposals for better memory controllers and approaches to manage memory traffic in modern multi-core embedded platforms. With respect to the existing literature, the purpose of this work is twofold. First, we want to demonstrate that scheduling CPU-originated memory traffic at the granularity of individual transactions is possible in PS-PL platforms. Second, and more importantly, we want to provide an infrastructure that is generic and extensible enough for the broader research community to adopt and foster a new chapter on PL-assisted memory scheduling. With this in mind, we establish the following goals.

Extensible memory scheduling infrastructure. First and foremost, the SchIM has been designed with modularity and extensibility in mind. We separate the functionalities that concern handling, queuing, selection, and forwarding of memory requests inside our infrastructure. Moreover, we design our SchIM to be able to support multiple memory scheduling policies simultaneously. A simple, standardized interface is provided to define new memory scheduling policies without impacting the design of the rest of the SchIM. We discuss in Section 5.5 the generic interface provided by the SchIM to implement a new memory scheduling policy.

Runtime configuration and transparency. We want the SchIM to be a robust supporting infrastructure to evaluate, compare, and contrast memory scheduling policies. As such, we strive to provide (1) runtime reconfigurability and (2) operational transparency. It is possible to rapidly identify desirable configuration parameters by allowing memory scheduling policies to be switched at runtime. Besides, an adopted policy can be tuned according to the workload criticality and memory intensiveness. For this purpose, the SchIM exposes a memory-mapped configuration interface where all the operational parameters can be changed at runtime. At the same time, we want to ensure that the applications and the (real-time) operating system under analysis need not be modified to use the SchIM. Hence, we propose using a thin virtualization layer to selectively route memory traffic through the SchIM without changes to the binary of OS kernel and applications.

Realistic performance with experimental policies. One of the limiting factors of research on memory scheduling policies is the ability to construct evidence of performance improvements with the realistic workload. Proposing a new memory scheduling policy is traditionally done with either a simulated setup or with a full-system soft-core implementation. Both cases have their drawbacks. The former gives a great deal of flexibility but achieving clock-level accuracy requires simulating many components the SoC whose details might not be publicly available. In addition, simulated setups that propose custom hardware designs cannot be directly adopted on real platforms without being first synthesized in hardware. Full soft-core-based SoC implementations suffer from two shortcomings. First, they run at relatively low frequencies and thus can extract only a fraction of the available DRAM bandwidth. Secondly, they are typically based on processors IPs that do not feature the same Instructions Set Architecture (ISA) as widely available COTS, which further limits the practical impacts of these works.

As reported in , re-routing the traffic of the core cluster through the PL-side comes at a cost in terms of extra latency and reduced bandwidth. Nonetheless, as PS-PL platforms mature and the interplay of PL and memory resources improves, a SchIM-like design could be the way to go for mission-reconfigurable, upgradable embedded systems.

4.2 Design Overview

As previously mentioned, the SchIM leverages the PLIM approach. CPU-originated main memory transactions are re-routed through the programmable logic and scheduled by the SchIM according to a flexible and configurable policy. The result is that the timing of memory transactions generated by real-time applications can be carefully determined and reasoned upon. Because the SchIM follows a PLIM approach, transactions can be selectively sent to the SchIM for scheduling. However, it is always possible to dynamically exclude the SchIM and route transactions directly to the main memory. Toward this paper's incentive, we consider a setup in which SchIM handles all the CPU-generated memory transactions.

Fig. 1 provides an overview of the location of the SchIM in the reference platform, while its internal organization is visible in Fig. 3. Application memory requests reach the SchIM the aforementioned HPM ports. Without loss of generality, we consider a SchIM instance with two arrival lanes, which are labeled as $\texttt{HPM}_1$ and $\texttt{HPM}_2$ in Fig. 3. The SchIM then forwards the received transactions towards main memory through the `HPS` interface. A more detailed view of the SchIM module is provided in Fig. 3 where the same convention is used to identify input and output ports. In addition, as shown in Fig. 3, a fourth `LPM` port is used to configure the SchIM from the PS.

The SchIM is composed of a number of sub-modules grouped into three different domains, namely (i) the *interfacing domain*, (ii) the *queuing domain*, and (iii) the *scheduling domain*.

The interfacing domain encompasses the sub-modules to interface the core logic of the SchIM with the rest of the system using the AXI protocol. This is comprised of three sub-modules. These are (i) the *packetizer(s)*, (ii) the *serializer*, and (iii) the previously mentioned *configuration* interface.

The PS-facing end of the **packetizer** offers an AXI slave port to accept new incoming transactions. Upon receipt, this module transforms each transaction into an equivalent *packet* that can be queued and scheduled by SchIM. Packetization of AXI transactions is necessary to be able to store transactions that are serial by nature. A standard AXI transaction is composed of one address phase (AR or AW channel) followed by a data phase (R or W channel), which can be itself composed of multiple successive bursts.

In many ways, the **serializer** is the dual module of the packetizer. Its purpose is to transform the packets that encode CPU-generated memory requests back into AXI-compliant transactions. As such, the serializer offers a master port to the rest of the system to be routed to the main memory controller.

The queuing domain handles how packets are stored between receipt and re-trasnmission. This domain is comprised of (i) the *dispatcher* module, (ii) the *transaction queues*, and (iii) the *selector* module.

The use of **multiple transaction queues** is necessary to differentiate the traffic of the CPUs and perform scheduling. As such, the SchIM associates a queue to each of the active cores — four in the platform of reference. The queues implemented in the SchIM not only act as a holding space for in-flight memory transactions. They also (a) provide information to the scheduling domain regarding their current state, and (b) they can generate a congestion control signal to the associated CPU core.

Congestion control is vital because memory transactions originated at the LLC controller follow the same route to the SchIM regardless of the originating CPU. The total number of outstanding transactions that the cores can emit exceeds the queuing elements' capacity on the loop-back route. Hence, priority inversion arises if a low-priority CPU's memory traffic is (temporarily) held. Latter is due to the uncontrolled queue buildup, which provokes a head-of-line blockage. Importantly, what described is true also for the normal route and it is a direct consequence of the best-effort nature of traditional multi-core memory buses. The SchIM allows the user to specify a configurable threshold on the occupancy of the queues that, when reached, issues a regulation signal to the corresponding CPU. We describe in greater detail how congestion control was implemented on the target platform in Section 5.7.

As suggested by Fig. 3, transactions are categorized and enqueued based on the source of traffic. The **dispatcher** module performs the matching between an incoming transaction and the destination queue. Similarly, transactions are dequeued by the **selector** module and sent directly to the output of the SchIM following the scheduling domain's resolutions.

The scheduling domain encompasses all the sub-modules that enable arbitration of transactions issued by the different cores of the PS. The modules in this domain are intended to be generic for extensibility, albeit the first set of two template schedulers is provided as a proof of concept. The scheduling policies currently implemented in the SchIM are Fixed Priority (FP) and Time Division Multiple Access (TDMA). Each of the parameters required by the implemented policies — such as the priorities and the periods — can be adjusted at runtime via the configuration interface.

The FP scheduler allows associating a priority value to each of the transaction queues. Pending transactions at the queues are then forwarded out of the SchIM following the user-defined priority order. The TDMA scheduler allows associating a transmission time slot to each of the queues expressed in PL clock cycles. The module then builds a schedule by concatenating the per-core slots so that only pending transactions from one queue at a time are forwarded by the SchIM.

5 SchIM Design and Implementation

A full-system implementation was carried out on a Xilinx ZCU102 development system, which is based on a Xilinx Zynq UltraScale+ XCZU9EG PS-PL SoC. The PS comprises four ARM Cortex-A53 CPUs that share a unified 1 MB LLC. The PS includes a DDR4-2666 controller connected to a 4 GB DDR4 memory module. There are two high-performance master interfaces (HPM1 and HPM2); and a third interface routed through the low power domain (LPM). The PL is capable of driving up to 16 interrupt requests lines towards the PS interrupt controller. We hereby provide key details on the operation of our SchIM in the target platform. These include complementary software stack, memory traffic accounting, regulation to prevent head-of-line blocking, and programming model.

5.1 Software Stack

As mentioned in Section 4.1, we want to ensure that the SchIM can be used with no modification to the OS and the applications under analysis. For this reason, we rely on a thin virtualization layer that can be used to redirect memory traffic from the direct route to the loop-back route (see Section 3.2). For this purpose, we use the open-source Jailhouse [16] partitioning hypervisor[2] Jailhouse does not boot the target machine. Instead, it relies on a standard Linux kernel to perform the initial boot sequence. When enabled from a Linux driver, Jailhouse dynamically virtualizes the original OS. In line with its partitioning-only philosophy, Jailhouse has a small footprint and enforces virtualization-aided partitioning of essential resources like CPUs, interrupts, main memory, I/O devices. It does not perform any virtual-CPU scheduling.

Following Jailhouse's nomenclature, a resource partition is called a *cell*, while guest OS's are referred to as *inmates*. An inmate can be either a bare-metal application, an RTOS or a full-fledged OS like Linux. Jailhouse uses ARM hardware Virtualization Extensions (VE) to offer a set of Intermediate Physical Address (IPA) to its inmates that is compatible with the way they have been compiled. Jailhouse then maps IPA ranges of different cells to configurable Physical Addresses (PAs) — stage-2 translation. By changing the configured stage-2 mapping, it is possible to dynamically re-route via the loop-back the memory traffic generated by each inmate.

As described below, some modifications were necessary to the mainline Jailhouse code for our full system implementation[3].

5.2 Altered communication scheme

In order to achieve the objective of re-ordering transactions, one must alter the standard AXI communication scheme explained in the Section 3.3. To this end, the SchIM is interposed between the master (HPM) and the slave (HPS) as depicted in Fig. 2b. As shown in Fig. 2b, only the phases initiated by the masters (i.e., address phase on AW and AR and the data phase on W) are intercepted for re-ordering by the SchIM. The introduction of the SchIM has a direct consequence on the overall communication scheme. Unlike the response phases on channels R and B that remain unchanged, the address and write data phases are handled following a store-and-forward scheme. Consequently, a write transaction will start exactly as in the standard AXI scheme with its address phase ① and data phase ②. These two transactions are buffered within the SchIM's queues (③) and only relayed following the internal memory scheduler's logic. This release of the transaction leads to the initialization of two new addresses and data phase ④, and ⑤. Finally, the response phase ⑥ goes directly from the slave to the master without being intercepted. For read transactions, the same modifications apply to the address phase ①' which is buffered (②') for some time before being re-emitted in ③'. Just like for write acknowledgments writing, the read response phase ④' is not intercepted by the SchIM.

5.3 Queueing Domain

At the heart of the queueing domain, lies the queues. They work as FIFOs. However, instead of inserting the new data at the back of the queue, the new data is always inserted as close

[2] The source code is available at `https://github.com/siemens/jailhouse.git`.
[3] The modified Jailhouse sources are available at `https://github.com/rntmancuso/jailhouse-rt`.

as possible to the front of the queue. This mechanism helps avoiding gaps within the queues prevents the loss of few clock cycles that would be required to move the data from the back to the front. From the authors' experiments, saving clock cycles in SchIM is vital to keep the final bandwidth as high as possible.

Furthermore, the queues have been designed to deal with three constraints. Firstly, the queues store both read and write packets such that the order at which transactions arrived is guaranteed. This implies that all the queue slots have the same size regardless of whether they contain read or write packets. Secondly, due to the altered communication scheme (see Section 5.2), each slot needs to be large enough to store both the address phase payload and the corresponding data of an AXI write transaction (678 bits). The depth of each queue is determined by considering the worst-case scenario. The latter consists of having to handle the maximum number of outstanding read and write transactions simultaneously. Our SchIM instance on the considered Xilinx UltraScale+ platform was configured with queues that are 16 slots in-depth. Indeed, the HPM ports in this platform cannot handle more than eight transactions of each type [37].

5.4 LLC-SchIM Interface and Traffic Accounting

As illustrated in Fig. 1, the considered system features an LLC shared between the four cores of the PS. For a non-cacheable read (resp., write) memory access, which CPU represents the source of the traffic is carried in the ID bits of the corresponding AR (resp., AW) AXI transaction. But for cacheable memory accesses, which is the norm for application workload, this is not the case. This is mainly because cache controllers typically use a write-back strategy. In this case, a read or write cache miss causes up to two events: (1) a cache refill and (2) a cache eviction. The cache refill is carried out with a read AXI transaction. If the line being evicted was previously written (dirty), then the eviction causes a write AXI transaction. It follows that, while read AXI transactions have an easily identifiable source, write transactions do not. Indeed, a CPU x might be causing the eviction of a line previously allocated and modified by CPU y. Hence, accounting (and scheduling) the resulting write transaction as if it originated from CPU x would be incorrect.

To ensure fair accounting for both read and write traffic, we rely on cache partitioning through coloring. As studied in a number of previous works, cache coloring is easy to implement at the hypervisor level [15, 21, 32]. In our system setup, we leverage the support Jailhouse already provides. The standard support has been extended to support booting a Linux inmate over colored memory. Cache partitioning allows us to establish a 1-to-1 relationship between any read/write transaction traversing the SchIM and the originating CPU. Moreover, with cache coloring in place, the SchIM uses the color bits in the address of the memory transactions (AR and AW channels) — instead of the AXI ID bits — to differentiate between the traffic of the various cores.

Finally, recall that the SchIM forwards transactions between HPM and HPS ports. These ports follow the asynchronous AXI protocol that allows issuing multiple outstanding AR and AW transactions. The protocol dictates that any outstanding transaction must have a unique AXI ID. This property is crucial to be able to match received responses with outstanding requests. Unfortunately, a potential mismatch between the bit-width of the AXI ID emitted at the HPM ports and the bit-width of AXI ID accepted by the HPS ports. For instance, in the platform of reference, the HPMs emit 16-bit AXI IDs, while the HPS AXI ID bit-width is 6 bits. Therefore, the SchIM also acts as an AXI ID translator.

5.5 Scheduling Interface and Implemented Policies

All the memory schedulers included in the scheduling domain share a common interface to ease the integration of a new scheduler. In terms of input signals, a generic scheduler module must define (1) a manual reset signal that can be triggered through the configuration port; (2) a vector of bits where each bit indicates whether the associated queue is empty; and (3) a signal indicating if the last scheduled transaction as been consumed. Alongside these inputs, the scheduling modules also have access to all the configuration registers listed in Table 1. In terms of outputs a SchIM scheduler must define (1) a signal to the selector indicating the queue considered for scheduling; and (2) a signal stating whether the current scheduling decision is valid. We hereby review the initial set of memory scheduling policies implemented in the SchIM.

5.5.1 Fixed Priority

The FP scheduling module aims at enforcing strict prioritization of cores' memory traffic. The priority ordering is explicitly defined by the user through the configuration port. While the SchIM instance used in this paper only has four queues, 16 different levels of priority are offered as the considered platform supports up to 16 different colors. This is useful if an hypervisor that supports vCPU scheduling is used. In this case, the SchIM allows assigning different priorities to different partitions sharing the same physical CPU. The core-to-priority assignment must be strict, meaning that two cores cannot be assigned the same priority.

The FP scheduling module only needs two pieces of information. That is (1) the priority associated with each queue and (2) whether a given queue contains at least one buffered transaction. The module logic always selects the queue with the highest priority. Lower priority queues are considered when higher priority queues do not have transactions. This is done by internally setting the user-defined priority of a queue as 0 when the corresponding queue is empty.

5.5.2 Time Division Multiple Access

The TDMA memory scheduler is a non-work conserving policy that operates by defining a per-core time *slot* during which the core has exclusive access to main memory. The slots are expressed in PL clock cycles, to maximize granularity. The configuration port can be used to specify and change the slots specifications at runtime.

The implementation of the module uses a counter register to track the time elapsed in the current TDMA primary frame — defined as the sum of all the cores' slots. It is reset to 0 at the beginning of a new major frame. Using the time-tracking register, the module determines to which core the current slot belongs, and forwards the information to the queue selector. This is done by summing up the length of all the previous slots, and determining if the current time falls within the interval of the considered core's slot.

5.6 Programming Model

The parameters that compose the programming interface of the SchIM are summarized in Table 1. The `base` address referenced in the table can be set when the SchIM is deployed in the PL. By default, this is set to `0x800000000`. All the parameter registers are 32 bit wide, except for the priorities of the FP scheduler. In this case, the priority values are encoded using 8 bits. The last "Mode" register allows a user to select the active memory scheduler.

Table 1 Available SchIM configuration registers.

Parameter	Associated Core	Address
TDMA slots	C_0	`base+0x00`
	C_1	`base+0x04`
	C_2	`base+0x08`
	C_3	`base+0x0C`
User Thresholds	C_0	`base+0x10`
	C_1	`base+0x14`
	C_2	`base+0x18`
	C_3	`base+0x1C`
FP Priorities	C_0 | C_1 | C_2 | C_3	`base+0x20`
Reserved		
Mode	N/A	`base+0x38`

5.7 PL-to-PS Feedback

Each of the HPM ports interfacing the PS and the PL sides (HPM1 and HPM2) have two dedicated queues for read and write transactions. Since transactions are being buffered inside SchIM as well as in these port buffers, head-of-line blocking can happen. Head-of-the-line blocking is harmful for performance; and can cancel out the benefits of transaction scheduling performed by the SchIM. For instance, in the case of a non work-conserving policy (e.g., TDMA), if the HPM port queue gets filled with transaction coming for the same core, no other transaction will be able to reach the SchIM and thus be considered for scheduling. This implies that no transaction would be scheduled until the end of the active core's TDMA slot. On the other hand, for work-conserving policies (e.g., FP) in the presence of head-of-line blocking, the decisions being taken by SchIM would directly depend on the order at which transactions are emitted by the HPM port buffer.

In both cases, one must prevent the cores from saturating the HPM port buffers. In order to avoid such situation, we implemented a feedback scheme aimed at slowing down the cores when necessary. As we mentioned in the context of Fig. 3, the SchIM's queues are associated a programmable threshold. Whenever the queue occupancy reaches (or exceeds) the associated threshold, a per-core interrupt line is asserted from the PL to the PS side. When received, the interrupt is treated by the platform software as a *fast interrupt request* (FIQ) and directly handled by the hypervisor—invisible to any guest OS. The advantage of using FIQs instead of regular IRQs is the significantly reduced handling latency [31]. Minor modifications to the TrustZone monitor were necessary to correctly configure FIQ handling. To minimize overhead, the installed FIQ handler only executes two assembly instructions. These are (1) a `dsb` memory barrier that stops the core until all the outstanding memory transactions have been completed, and (2) a `eret` instruction to exit the FIQ context. There is not need to save/restore any register because FIQs have banked syndrome/status registers and because no general purpose register is modified in the handler.

Ideally, the available space in the HPM buffers should be shared evenly between the cores. Since each HPM port has a buffer with a depth of 8+8 transactions, each core should occupy at most 2 slots in each buffer. Unfortunately, our experiments highlighted that the control over amount of transactions buffered by each core is imperfect. Often times, the selected threshold is exceeded by up to two transactions. This is the main reason why we propose a dual-ported SchIM which uses both the available HPM ports. Indeed, by assigning two cores on each of the ports, the ideal threshold on maximum amount buffered transactions can be doubled. The increase provides enough room to compensate for imperfections in the micro-regulation performed with PL-to-PS FIQ delivery.

6 Evaluation

The present section aims at evaluating the behavior of the SchIM on the target platform, its overhead and benefits. First, in subsection 6.1, we review our experimental setup. Thereafter, we assess the overhead introduced by the SchIM in Section 6.2. Section 6.3 explores the impact of the PL-to-PS feedback on the control and the performance. In Section 6.4, an in-depth analysis of the SchIM's behavior is presented. Finally, an evaluation of the temporal behavior of a set of real-world benchmarks operating through the SchIM is provided in Section 6.5.

6.1 Experimental Setup

The SchIM has been evaluated using synthetic benchmarks (or *Memory Bombs*), real benchmarks selected from the San Diego Vision Benchmark Suite (SD-VBS) [35] and a combination of the two. Specifically, seven memory-intensive benchmarks have been selected, i.e. *stitch*, *texture synthesis*, *disparity*, *tracking*, *localization*, *mser* and *sift*. For our runs, we have considered all the intermediate input sizes ranging from SQCIF (128×196 pixels) to VGA (640×480 pixels). When running any benchmark, we use the cache coloring mechanism implemented in the Jailhouse hypervisor [32] to partition the LLC evenly amongst the 4 cores and to prevent our measurements from being affected by inter-core cache interference. As a result, each benchmark operates on 1/4 of the total cache space—256 KB. As extensively discussed in [14, 41], it is also important to avoid inter-core DRAM bank conflicts, which can cause the arbitrary re-ordering of transactions originating from different cores. This is accomplished by (1) configuring the DRAM controller to disable DRAM bank interleaving; and (2) by performing static cache bleaching [11, 29] at the SchIM's output to re-compact accesses to colored pages into contiguous DRAM accesses. In this platform, there are a total of 16 DRAM banks of 256 MB each. Thanks to bleaching, we can assign the full size of 4 banks (i.e., 1 GB) to each core, instead of being restricted to only 1/4 of that due to non-overlapping color and bank address bits.

To evaluate the capabilities of the SchIM, two memory routes for the traffic generated by the cores are compared. The first serves as baselines, whereas, the last one is the one under analysis and involves the SchIM module. The first path consists in the cores directly accessing the main memory. As illustrated in Fig. 1, the traffic simply goes through the *Main Interconnect* before arriving at the DDR controller. This path is referred to as the *normal route*. Secondly, we consider the case where the SchIM module is deployed and in use to schedule memory traffic generated by the CPUs in the PL. Cores 0 and 1 target HPM1 aperture, while cores 2 and 3 target HPM2. In our analysis, the SchIM is used in all the available modes, i.e., FP and TDMA.

Note that in the case of the *normal route*, combining both a strict cache partitioning and strict bank partitioning could not be applied. In fact, as a direct consequence of the address coloring and in the absence of a bleacher, only 1/16 of each 1 GB wide memory allocation can be used by each core. The resulting reduced space of 64 MB is not enough for running Linux. Consequently, in the case of the *normal route*, the cores have been split into two groups of two, where each group targets independent sets of banks. This configuration allows the cache to be partitioned using address cooring.

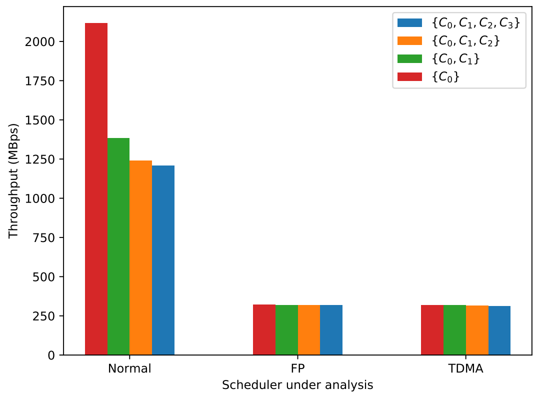


Figure 4 Bandwidth in MBps for different path under increasing set of cores contending.

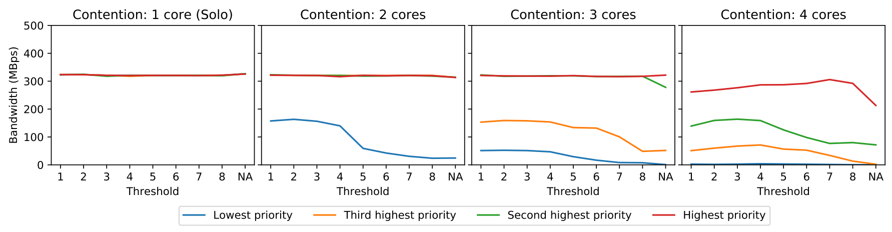
6.2 Platform Capabilities and performance degradation

Intuitively and as discussed in [29], redirecting the traffic coming from the cores to the PL side incurs a performance hit. In spite of the lower frequency at which the SchIM operates (250 MHz), the theoretical throughput when using both the HPM lanes should be around 8 GBps. We observe, however, that the achievable throughput through the HPM ports is a fraction of what we measured by accessing the main memory through the *normal route* (2116.5 MBps and 1207.41 MBps for solo and full contention by 3 other cores, respectively). We further provide a discussion on the bandwidth drop when transactions are routed through the PL in Section. For the sake of completeness, we quantify in Fig. 4 the maximum bandwidth achieved through the PL — and hence through the SchIM. Nevertheless, it is important to remember that the absolute figures are strictly platform dependent.

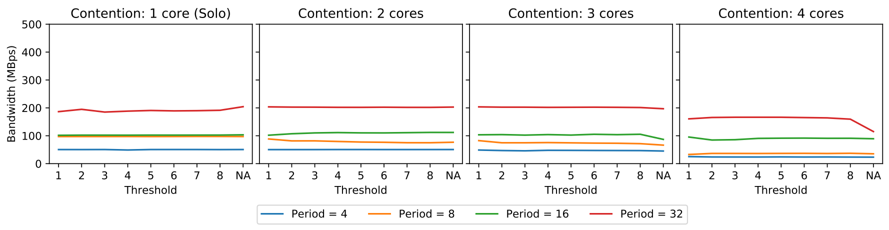
In Fig. 4, we have computed the throughput of one *core under analysis*, here core 0 (noted C_0) when a synthetic memory-intensive application is deployed on an increasing number of cores denoted with the same notation. The first bar cluster ("Normal") refers to the throughput measured via the normal route. The other two clusters capture the observed bandwidth when traffic is routed through and managed by the SchIM. One cluster is provided for each of the implemented memory scheduling policies, namely — from left to right — FP and TDMA. As expected, there is a sharp reduction (around 75%) in terms of absolute bandwidth. Importantly, however, two aspects need to be highlighted. First, the bandwidth achieved through the SchIM is still remarkably high and allows studying the behavior of the realistic workload under custom memory scheduling policies, which is the primary goal of this research. Second, it emerges that the implemented FP and TDMA policies are capable of protecting the core under analysis from inter-core interference, while this is not the case when going through the normal route

6.3 PL-to-PS feedback performance impact

As mentioned in Section 5.7, the PL-to-PS feedback enables our SchIM to regulate the HPM ports buffer occupancy to prevent head-of-line blocking. Since this feedback directly throttles the desired core, the selection of an adequate threshold is important to preserve the balance between control and performance. Therefore, in Fig. 5, we have explored the sensitivity to the threshold for each of the proposed schedulers under different levels of contention. The thresholds in use range from 1 to 8 and even include the case where the feedback mechanism



(a) Threshold-Bandwidth relationship curves for the FP scheduler



(b) Threshold-Bandwidth relationship curves for the TDMA scheduler

Figure 5 Figures showing the impact of the threshold in use on the final bandwidth experinced by the cores for the offered schedulers

is disabled (noted *NA*). The contention is created by up to four co-running cores emitting write transactions. For each parameter applied to a scheduler (i.e., fixed priority or TDMA slot), the co-running cores are assigned the most demanding parameters available (i.e., the highest priority for FP or the biggest TDMA slot).

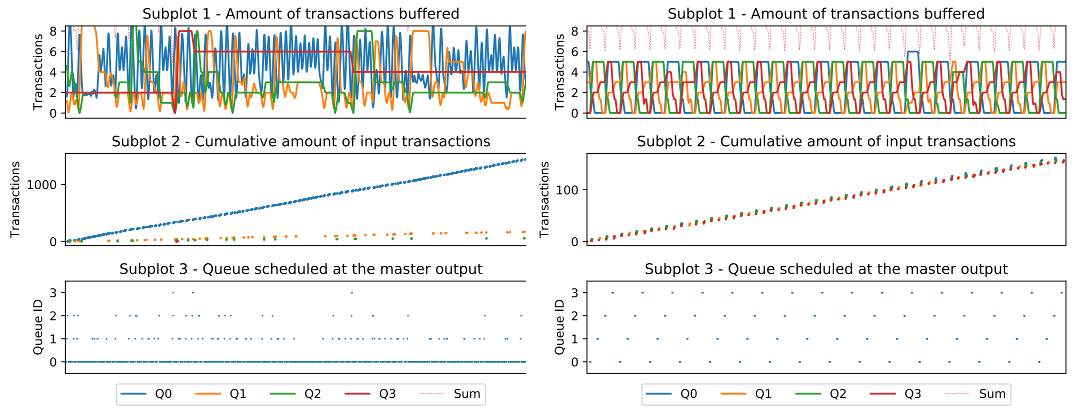
In the case of the FP scheduler (Fig. 5a), one can observe that when running alone, the threshold has no influence on the throughput. However, as soon as co-runners are added, the cores start to experience a decrease in throughput. Fig. 5b shows that the TDMA scheduler is not impacted considerably by the threshold with respect to the throughput. Globally, the scheduler manages to preserve a constant throughput regardless of the contention and the assigned slot.

Nonetheless, under high contention, one can observe that the throughput of each core is affected. The fourth inset of Fig. 5a and Fig. 5b illustrate the importance of the threshold and the PL-to-PL feedback mechanism as a a considerable drop of throughput can be observed for the highest priority of FP and for a TDMA period of 32.

Considering these experiments, setting the threshold to four for all the schedulers seems to bring the best trade-off between control and performance. However, this value cannot be blindly applied to all cases as this experiment is performed for a sequential and contiguous access pattern.

6.4 Internal Behaviour of SchIM

The next objective is to verify the correct behavior of the schedulers at the granularity of a clock cycle by observing the inputs, the outputs and the internal signals and registers of the SchIM module. This is made possible thanks to the *Integrated Logic Analyzer* (or ILA) provided by Xilinx [36]. The latter IP can be directly implemented on the PL side, alongside the SchIM, and is able to probe the signals and to store them in a local memory. For this experiment, a group of relevant internal signals have been probed and captured during a window of 16384 contiguous clock cycles. Then, the information has been extracted by post-processing the data. To characterize the behavior of the two different policies, the



(a) FP with ordering $C_0 \succ C_1 \succ C_2 \succ C_3$ **(b)** TDMA with slots of 256 clock cycles

Figure 6 Trace snapshots of SchIM for FP (6a), TDMA (6b)

ILA has been instrumented to collect (i) the amount of transactions being buffered in the queues at each clock cycle (inset 1 in Fig. 6a and Fig. 6b) (ii) the rate at which queues receive new transactions from the cores cluster (inset 2 in Fig. 6a and Fig. 6b) and (iii) the queues ID of each transaction forwarded by the SchIM module (inset 3 in Fig. 6a and Fig. 6b).

For the Fixed Priority trace snapshot displayed in Fig. 6a, the following strict priority ordering has been considered: $C_0 \succ C_1 \succ C_2 \succ C_3$ where the $\succ$ operator means that the left argument has a strictly higher priority than the right argument. In this experiment, a regulation threshold of 3 for each core has been used. As emphasized by the inset 2 in Fig. 6a, the FP scheduler is able to prioritize the traffic of one core at the expense of the others according to the priorities assignment. Furthermore, one can observe that the rate at which the queues receive new transactions from their associated core is proportional to the priority level in the priority ordering. Finally, the third inset in Fig. 6a confirms the correct behavior of the FP policy.One can see that the cores with the highest priority also feature the highest density of transactions at the output of the SchIM.

The trace snapshot displayed in Fig. 6b has been obtained by configuring the SchIM module in TDMA mode. For the sake of clarity, a slot of 256 clock cycles has been set for each core. Besides, the threshold of each core has been set to 4 to create sharp transitions. The insets 2 and 3 of Fig. 6b clearly show the behavior expected from a TDMA schedule. In fact, one can clearly see in the latter that transactions originating from one core are only being repeated out of the SchIM module during a well-defined and periodic time slot of 256 clock cycles. In the inset 2 of Fig. 6b, we can observe a similar pattern, with transactions arriving only during the TDMA slot associated with their queue (and indirectly core). Globally, the rate at which queues receive transactions is steady and constant.

6.5 Memory Isolation

On the platform considered for this set of experiments, the Xilinx ZCU102 development board, we denote three main sources of inter-core performance interference: (1) cache contention, (2) DRAM bank conflicts, and (3) the congestion and saturation of the memory controller. Despite their orthogonality, the two first sources are tackled respectively via the integration of page coloring in the hypervisor and static bleaching in the SchIM. On the other hand, since the SchIM provides fine-grained control over the timing and ordering of transactions

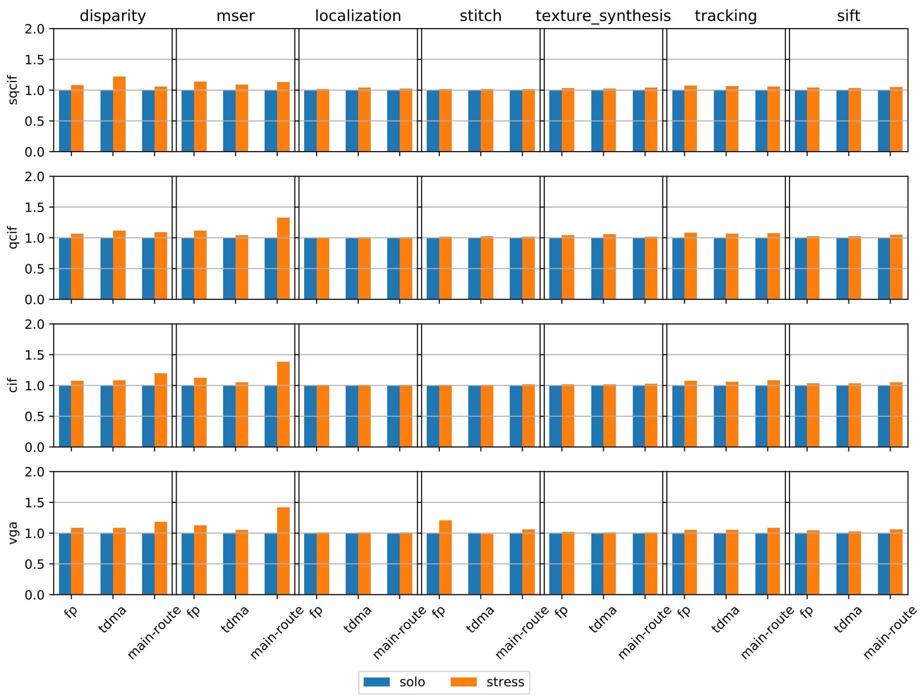


Figure 7 Normalized execution time for each benchmark and input size for *Solo* and *Stress*. Each column denotes a given benchmark of the SD-VBS suite, while each row denotes a specific input size (in increasing order from top to bottom).

originating from the application cores as they reach the memory controller. Thus, the SchIM brings memory bandwidth management into the PL, and provides not only regulation but a generic infrastructure to experiment with custom bandwidth management techniques, both work-conserving and non-work-conserving.

The evaluation setup considered for this experiment is identical to the one presented in Section 6.1. The routes going through the PL and using our SchIM (i.e., FP and TDMA) benefit from both cache partitioning and bank partitioning. On the other hand, the *normal route* uses cache partitioning and sees its cores divided into two sets targeting each a different group of private banks.

To evaluate the capability of our SchIM with respect to its ability to ensure performance isolation between the cores, a set of experiments involving SD-VBS benchmarks were designed. Here, we compare the execution time of an application on a given core when running alone (referred to as *Solo*) and when running alongside interfering synthetic benchmarks (write memory bombs) on all the other cores (referred to as *Stress*). For each combination of a route to main memory (i.e., the *normal route* or the *SchIM route*) and scheduler, the result obtained for *Stress* is normalized with respect to the equivalent configuration in *Solo*. The results obtained on the considered benchmarks are listed in Fig. 7. The results in the Fig. 7 are the aggregation (arithmetic average) of 30 different runs in the same configuration. Each bar cluster of the Fig. 7 insets represents one of the aforementioned configuration for *Solo* and *Stress*. The height of each bar denotes its normalized execution time.

For this set of experiments, the FP scheduler was configured such that the core under analysis (i.e., the one running the benchmark) has the highest priority and a threshold of 8. The other cores are assigned lower priorities and thresholds matching their priority order (i.e., 4, 2, 1). Under TDMA scheduling, the core under analysis has a slot of 512 clock cycles and a threshold of 14 while the co-runners are assigned slots of 32 and 16 clock cycles with thresholds of 4 and 1.

The *normal route* is used as a baseline for this experiment because no scheduling is performed in this configuration. The Fig. 7 highlights the sensitivity of both *disparity* and *mser* to inter-core interference on the *normal route*. This is especially the case for large input sizes such as *cif* and *vga*. On the other hand, *texture synthesis* and *localization* do not suffer from inter-core interference. Globally, the TDMA scheduler always manages to preserve the isolation of the core, having execution times under *Stress* similar or smaller than the *normal route*. This is particularly visible for *qcif*, *cif* and *vga* input sizes of *disparity* and *mser*. Similarly, the FP scheduler is also capable of ensuring sound isolation of the core under analysis.

7 Discussion and Limitations

By design, the PLiM module proposed in this paper, the SchIM, centralizes the memory traffic and its scheduling. A centralized design makes sense on the specific target platform because there exist only one memory controller and thus a single path between the LLC and the DRAM controller. In systems where multiple paths between the processing units and the memory controllers exist, for instance when multiple controllers and channels are present, a decentralized design is to be preferable to better exploit the available memory parallelism. In such platforms, a possible avenue could be instantiating multiple SchIM modules, roughly one per channel, and introducing appropriate out-of-band signaling between the modules for coordination off the critical path.

As we mentioned in Section 6.1, our setup includes the Jailhouse partitioning hypervisor. While the SchIM module does not strictly require the PS side to use a hypervisor, Jailhouse has been extensively used for the evalution as it provides convenient features to control physical memory allocation. For instance, the support for page coloring has been used to both partition the LLC space and to easily identify the owner of each memory transactions in the SchIM (as presented in Section 5.4). However, instead of enforcing cache partitioning, one could instead identify the ownership of memory transactions by extracting a different subset of address bits. For instance, if the physical memory allocated to different partitions is not interleaved, then the most significant bits of the address can be used to perform traffic accounting. In addition, the IPA address virtualization is convenient to transparently redirect the memory traffic of the application partitions through the PL side, even if they are initially booted through the normal route. Finally, the cores throttling mechanism (see Section 5.7) via the FIQs can be implemented at EL3 (Secure Monitor) or in the individual guest OS's instead (EL1). Implementing FIQ handling in the hypervisor (EL2), however, has the advantage of not requiring any change in the guest OS's, as well as not requiring a full switch into secure mode compared to an implementation at EL3.

On the same note, provided that the FIQ lines are not used by the inmates, the feedback regulation mechanism is entirely transparent to the guest OS's (or even for bare-metal applications) and introduces minimum overhead. The Linux kernel do not use FIQs, and the same goes for typical RTOS's. Nonetheless, it must be acknowledged that defining a FIQ handler to be used for CPU throttling might interfere with (and be interfered by) the latency of FIQ handling in guest OS's that rely on the same functionality. This is mainly because FIQ handling is non-preemptive. We also recognize that the PL-to-PS feedback mechanism is relatively coarse. Inset 1 of Fig. 6b highlights this problem. Even though all the queues have been assigned a threshold of 4, the threshold is often exceeded. The worst-case being queue 3 exceeding the threshold by 2 on the right-hand side of the plot. This problem can be attributed to the reaction time of the FIQ routine, and to the fact that jumping to the FIQ handler itself might cause a few memory transactions depending on the cache state. Currently, the thresholds used for FIQ-based regulation require to be fine-tuned manually by the user. Future extensions of the SchIM will explore the implementation of schedulers capable of dynamically adapting the thresholds to maximize performance and improve isolation.

The loss in bandwidth caused by routing transactions through the PL is important and a serious drawback against the adoption of the SchIM. Our experiments in Section 6.2 have shown that rerouting the traffic through the PL has a cost. As illustrated in Fig. 4, up to 2100 MBps can be extracted from the *normal route* whereas any route through the PL only achieves around 320 MBps. In contrast, a back-of-the-envelope calculation reveals that for a PL operating at 250 MHz (the SchIM frequency), and with a bus width of 128 bits, a full-duplex throughput of approximately 3.7 GBps can be sustained. This calculation is in line with the reported throughput in an experiment conducted in [19], in which PL-originated transactions targeting the DRAM passed through the one of the HP ports. This suggests that the PL-to-DRAM route can sustain a much higher throughput than what has been experimentally observed in our evaluation setup, where transactions originate from the PS side. In light of thse considerations, we can conclude that the source of the bandwidth loss can be imputed to the bus segments connecting the CPU cluster to the HPM ports. A focused study is necessary to narrow down the exact reason for the performance drop. Nonetheless, vendor-imposed bandwidth throttling, PS-to-PL clock-domain crossing delays, and shallow FIFOs at the HPM ports and/or at the main PS-side interconnect represent plausible reasons. We anticipate that due to the platform-specific nature of this issue, the raw performance of the SchIM will substantially vary across different SoCs.

8 Conclusion

In the present article we introduced the SchIM, a memory transactions scheduler framework that can be integrated with commercially available platforms featuring a tightly coupled processing system and programmable logic. A full-system implementation in a commercially available PS-PL platform has been detailed, which encompasses the accompanying software stack and the platform-specific integration steps have been detailed in as well as advanced scheduling techniques are few of many possible future directions.

Through a set of experiments, we assessed the capabilities of the framework and demonstrated the correct behavior of the proposed scheduling policies, namely Fixed Priority, Time Division Multiple Access and Traffic Shaping. Finally, we showed using a suite of real-world benchmarks that the SchIM is capable of enforcing strong temporal isolation despite heavy memory contention.

The authors see the proposed SchIM as a stepping stone to propose, test and validate novel memory scheduling policies to be tested on embedded platforms with realistic performance and complex workload. For this reason, the SchIM has been designed to be open-source and with extensibility in mind. Especially, we strongly envision that the SchIM could represent a stepping-stone toward profile-based memory traffic scheduling.

References

1 B. Akesson. *Predictable and composable system-on-chip memory controllers*. PhD thesis, 2010. doi:10.6100/IR658012.

2 B. Akesson, K. Goossens, and M. Ringhofer. Predator: a predictable SDRAM memory controller. In *Proceedings of the 5th IEEE/ACM international conference on Hardware/software codesign and system synthesis*, pages 251–256, 2007.

3 G. Alonso, T. Roscoe, D. Cock, M. Ewaida, Kaan Kara, Dario Korolija, D. Sidler, and Ze ke Wang. Tackling hardware/software co-design from a database perspective. In *Conference on Innovative Data Systems Research (CIDR)*, Amsterdam, Netherlands, Jan. 2020.

4 ARM. ARM® CoreLink™ QoS-400 Network Interconnect Advanced Quality of Service, 2013. Accessed on 09.01.2020.

5 ARM. AMBA AXI and ACE Protocol Specification. Technical report, 2019. URL: https://static.docs.arm.com/ihi0022/g/IHI0022G_amba_axi_protocol_spec.pdf.

6 A. Biondi, A. Balsini, M. Pagani, E. Rossi, M. Marinoni, and G. Buttazzo. A framework for supporting real-time applications on dynamic reconfigurable FPGAs. In *2016 IEEE Real-Time Systems Symposium (RTSS)*, pages 1–12, 2016. doi:10.1109/RTSS.2016.010.

7 J. M. Calandrino, H. Leontyev, A. Block, U. C. Devi, and J. H. Anderson. LITMUS^{RT} : A testbed for empirically comparing real-time multiprocessor schedulers. In *2006 27th IEEE International Real-Time Systems Symposium (RTSS'06)*, pages 111–126, 2006. doi:10.1109/RTSS.2006.27.

8 F. Farshchi, Qijing Huang, and H. Yun. BRU: Bandwidth regulation unit for real-time multicore processors. *2020 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pages 364–375, 2020.

9 F. Farshchi, P. Kumar, R. Mancuso, and H. Yun. Deterministic Memory Abstraction and Supporting Multicore System Architecture. In Sebastian Altmeyer, editor, *30th Euromicro Conference on Real-Time Systems (ECRTS 2018)*, volume 106 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 1:1–1:25, Barcelona, Spain, July 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. URL: http://drops.dagstuhl.de/opus/volltexte/2018/9001, doi:10.4230/LIPIcs.ECRTS.2018.1.

10 C. Ferri, A. Marongiu, B. Lipton, R. Bahar, T. Moreshet, L. Benini, and M. Herlihy. SoC-TM: integrated HW/SW support for transactional memory programming on embedded MPSoCs. In *Proceedings of the seventh IEEE/ACM/IFIP international conference on Hardware/software codesign and system synthesis*, pages 39–48, 2011.

11 G. Gracioli, R. Tabish, R. Mancuso, R. Mirosanlou, R. Pellizzoni, and M. Caccamo. Designing mixed criticality applications on modern heterogeneous MPSoC platforms. In *31st Euromicro Conference on Real-Time Systems (ECRTS 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

12 Intel, Corp. Intel's Stratix 10 FPGA: Supporting the smart and connected revolution, October 2016. Accessed on 09.01.2020. URL: https://newsroom.intel.com/editorials/intels-stratix-10-fpga-supporting-smart-connected-revolution/.

13 A. K. Jain, S. Lloyd, and M. Gokhale. Microscope on memory: MPSoC-enabled computer memory system assessments. In *2018 IEEE 26th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, pages 173–180, 2018. doi:10.1109/FCCM.2018.00035.

14 H. Kim, D. de Niz, B. Andersson, M. Klein, O. Mutlu, and R. Rajkumar. Bounding memory interference delay in COTS-based multi-core systems. In *2014 IEEE 19th Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pages 145–154, 2014. doi:10.1109/RTAS.2014.6925998.

15 H. Kim and R. Rajkumar. Real-time cache management for multi-core virtualization. In *2016 International Conference on Embedded Software (EMSOFT)*, pages 1–10, 2016.

16 J. Kiszka, V. Sinitsin, H. Schild, and contributors. Jailhouse Hypervisor. Accessed on 09.01.2020. URL: ttps://github.com/siemens/jailhouse.

17 C. Maiza, H. Rihani, J. Rivas, J. Goossens, S. Altmeyer, and R. Davis. A Survey of Timing Verification Techniques for Multi-Core Real-Time Systems. *ACM Comput. Surv.*, 52(3), June 2019. `doi:10.1145/3323212`.

18 Microsemi — Microchip Technology Inc. PolarFire SoC - Lowest Power, Multi-Core RISC-V SoC FPGA, July 2020. Accessed on 09.01.2020. URL: `https://www.microsemi.com/product-directory/soc-fpgas/5498-polarfire-soc-fpga`.

19 S. Min, S. Huan, M. El-Hadedy, J. Xiong, D. Chen, and W. Hwu. Analysis and optimization of I/O cache coherency strategies for SoC-FPGA device. In *2019 29th International Conference on Field Programmable Logic and Applications (FPL)*, pages 301–306, 2019. `doi:10.1109/FPL.2019.00055`.

20 R. Mirosanlou, M. Hassan, and R. Pellizzoni. DRAMbulism: balancing performance and predictability through dynamic pipelining. In *2020 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pages 82–94, 2020. `doi:10.1109/RTAS48715.2020.00-15`.

21 P. Modica, A. Biondi, G. Buttazzo, and A. Patel. Supporting temporal and spatial isolation in a hypervisor for ARM multicore platforms. In *2018 IEEE International Conference on Industrial Technology (ICIT)*, pages 1651–1657, 2018.

22 O. Mutlu and T. Moscibroda. Stall-time fair memory access scheduling for chip multiprocessors. In *40th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO 2007)*, pages 146–160. IEEE, 2007.

23 O. Mutlu and T. Moscibroda. Parallelism-aware batch scheduling: Enhancing both performance and fairness of shared DRAM systems. In *2008 International Symposium on Computer Architecture*, pages 63–74. IEEE, 2008.

24 K. Nesbit, N. Aggarwal, J. Laudon, and J. Smith. Fair queuing memory systems. In *2006 39th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'06)*, pages 208–222. IEEE, 2006.

25 M. Paolieri, E. Quinones, F. Cazorla, and M. Valero. An analyzable memory controller for hard real-time CMPs. *IEEE Embedded Systems Letters*, 1(4):86–90, 2009.

26 N. Rafique, W. Lim, and M. Thottethodi. Effective management of DRAM bandwidth in multicore processors. In *16th International Conference on Parallel Architecture and Compilation Techniques (PACT 2007)*, pages 245–258. IEEE, 2007.

27 F. Restuccia, A. Biondi, M. Marinoni, G. Cicero, and G. Buttazzo. AXI HyperConnect: A predictable, hypervisor-level interconnect for hardware accelerators in FPGA SoC. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6, 2020. `doi:10.1109/DAC18072.2020.9218652`.

28 F. Restuccia, M. Pagani, A. Biondi, M. Marinoni, and G. Buttazzo. Is your bus arbiter really fair? restoring fairness in AXI interconnects for FPGA SoCs. *ACM Trans. Embed. Comput. Syst.*, 18(5s), October 2019. `doi:10.1145/3358183`.

29 S. Roozkhosh and R. Mancuso. The potential of programmable logic in the middle: Cache bleaching. In *26th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS 2020)*, Sydney, Australia, April 2020.

30 P. Sohal, R. Tabish, U. Drepper, and R. Mancuso. E-WarP: a system-wide framework for memory bandwidth profiling and management. In *41st IEEE Real-Time Systems Symposium (RTSS 2020)*, Houston, TX, USA, Dec. 2020.

31 ST Microelectronics Inc. Real-time performance using FIQ interrupt handling in SPEAr MPUs, January 2010. Accessed on 10.01.2020.

32 M. Solieri T. Kloda, R. Mancuso, N. Capodieci, P. Valente, and M. Bertogna. Deterministic Memory Hierarchy and Virtualization for Modern Multi-Core Embedded Systems. In *25th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS 2019)*, pages 1–14, Montreal, Canada, April 2019. `doi:10.1109/RTAS.2019.00009`.

33 H. Usui, L. Subramanian, K. Chang, and O. Mutlu. Dash: Deadline-aware high-performance memory scheduler for heterogeneous systems with hardware accelerators. *ACM Transactions on Architecture and Code Optimization (TACO)*, 12(4):1–28, 2016.

34 P. Valsan and H. Yun. MEDUSA: A predictable and high-performance DRAM controller for multicore based embedded systems. In *2015 IEEE 3rd International Conference on Cyber-Physical Systems, Networks, and Applications*, pages 86–93. IEEE, 2015.

35 S. K. Venkata, I. Ahn, D. Jeon, A. Gupta, C. Louie, S. Garcia, S. Belongie, and M. B. Taylor. SD-VBS: The san diego vision benchmark suite. In *2009 IEEE International Symposium on Workload Characterization (IISWC)*, pages 55–64, 2009.

36 Xilinx. Integrated Logic Analyzer v6.2 LogiCORE IP Product Guide. Technical report, 2016. URL: `https://www.xilinx.com/support/documentation/ip_documentation/ila/v6_2/pg172-ila.pdf`.

37 Xilinx. Zynq UltraScale+ Device Technical Reference Manual. Technical report, 2019. URL: `https://www.xilinx.com/support/documentation/user_guides/ug1085-zynq-ultrascale-trm.pdf`.

38 Xilinx, Inc. Zynq UltraScale+ MPSoC - All Programmable Heterogeneous MPSoC, August 2016. Accessed on 09.01.2020. URL: `https://www.xilinx.com/products/silicon-devices/soc/zynq-ultrascale-mpsoc.html`.

39 M. Xu, L. T. X. Phan, H. Choi, Y. Lin, H. Li, C. Lu, and I. Lee. Holistic resource allocation for multicore real-time systems. In *2019 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pages 345–356, 2019. `doi:10.1109/RTAS.2019.00036`.

40 H. Yun, W. Ali, S. Gondi, and S. Biswas. BWLOCK: A Dynamic Memory Access Control Framework for Soft Real-Time Applications on Multicore Platforms. *IEEE Transactions on Computers*, 66(7):1247–1252, 2017.

41 H. Yun, R. Mancuso, Z. P. Wu, and R. Pellizzoni. Palloc: DRAM bank-aware memory allocator for performance isolation on multicore platforms. In *2014 IEEE 19th Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pages 155–166, 2014. `doi:10.1109/RTAS.2014.6925999`.

42 H. Yun, G. Yao, R. Pellizzoni, M. Caccamo, and L. Sha. MemGuard: Memory bandwidth reservation system for efficient performance isolation in multi-core platforms. In *2013 IEEE 19th Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pages 55–64, 2013.

43 Y. Zhou and D. Wentzlaff. MITTS: Memory inter-arrival time traffic shaping. *ACM SIGARCH Computer Architecture News*, 44(3):532–544, 2016.