

MemorEDF

Authors omitted for review.

Abstract—TBD

Index Terms—component, formatting, style, styling, insert

I. INTRODUCTION

This paper makes the following contributions: 1) We demonstrate that without any modification to the SoC circuitry, a configurable module could be interposed between the core and memory controller to perform traffic shaping. The proposed module operates at the granularity of transactions based on their interarrival time. 2) We perform transaction-level memory scheduling in configurable hardware. Various scheduling policies implemented in the module and could be configured at run-time.

3) A groundbreaking view on memory access scheduling, namely MemorEDF, has been proposed. In this method, the memory is viewed as a single resource

4) We provide a run-time profiling interface to log the transaction-level behavior of an application. Recorded data exploited to make and enforce efficient scheduling decisions for EDF and LLF policies.

5) This work implement a sandbox framework to test scheduling platforms on real hardware. Newly proposed scheduling techniques can be evaluated against the actual device, without rewiring the platform.

6) We implement and evaluate a full-stack design that includes the memory scheduler hardware module. Using this implementation, we analyze several well-known scheduling algorithms, namely, Earliest Deadline First (EDF), Least Laxity First (LLF), Time-division multipleaccess (TDMA).

II. RELATED WORK

Ask Tomasz and Gero for their ECRTS article

We should look at this paper "BRU: Bandwidth Regulation Unit for Real-Time Multicore Processors"

Citation list:

- PREM
- MemGuard
- PLIM
- Three Phase Model
- FRED Framework from Sant'Anna

III. SYSTEM BACKGROUND

A. Programmable Logic in the Middle(PLIM)

Here we detail the necessary background on interposing a module between a traditional multi-core processor and main memory.

Commercially available SoCs integrate a traditional embedded multi-core processor system (PS) and a block of

programmable logic (PL) with high-performance PS-PL communication interfaces. High-performance masters (HPM) and high-performance slaves (HPS) send and receive transactions to and from the PL, respectively. However, HPS's are not the only interface mediums for PL-PS communications. The chip has (limited number of) programmable interrupt lines present connected from the PL to the generic interrupt controller (GIC) inside the PS, whereas the GIC is a primary resource for managing interrupts sent to the processors. Like any other interrupt source in the system, a unique ID number identifies each of the PL-PS interrupts lines. Hence, by providing proper handlers associated with specific interrupt ID lines and unmasking them, PL can request a desired service routine from the PS. Nevertheless, as PS is buttressed by Arm Cortex-A53 Based Application Processing Unit (APU), the interrupt controller supports security extension to manage secure and non-secure interrupts grouped by ARM TrustZone.

The underlying mechanism is the ability to intercept memory transactions originated from the processors inside the PS, at the PL. Intercepted transactions are then forwarded from the PL again toward the memory controller inside the PS. The primary mechanism of PS-PL and PL-PS redirection of a transaction is called the Memory Loop-Back. Loop-Back is done through address bit manipulation of the transaction such that it falls in the range of the target HPM(HPS) for the PS-PL(PL-PS) interception. In this way, the main memory content is accessed, but through a programmable environment. It is possible to act on the characteristics of the traffic that now traverses the PL. For example, in the PL, it is possible to direct the transaction to arbitrary modules before, eventually, redirecting it back to PS and the memory controller, ultimately.

This provides a unique capability of manipulating individual memory transactions. Hence, by sitting between CPUs and main memory, PLIM is exploited to perform memory scheduling. The idea is further leveraged by designing a configurable memory scheduler in the middle, namely, SchIM.

First and foremost, SchIM follows a selective and dynamic re-routing policy settled at the run-time, meaning that provided a proper routing configuration interface, SchIM can be bypassed if unacceptable overhead for some applications is detected by going through the PL following the Loop-Back routing strategy. Additionally, SchIM, provides a safe production-ready environment on COTS platforms where one can enforce and test any proposed scheduling policy at the level of the transaction altogether by the real hardware. Memory scheduling traditionally done via hardware modifications at the controller level, or solely via software. Providing SchIM, we are now able to transcend this on intact COTS to conduct an analysis of feasibility, challenges and performance.

In this paper we provide a several proof-cases where SchIM is programmed to enforce a several elected scheduling policies of Fixed Priority, TDMA, and Memguard.

B. Jailhouse, the partitioning Hypervisor

Jailhouse is a well-known open-source partitioning hypervisor¹ based on Linux, with a prominent levity due to focusing on software-based partitioning of only a handful of essential resources, i.e., on-chip resources like CPUs, memory regions, I/O devices, and skipping any virtual scheduling.

In jailhouses nomenclature, Virtual Machines (VMs) are referred to as inmate cells where Jailhouse already supports colored mappings. An inmate can be either a bare-metal application or any operating system like Linux, each programmed to see a contiguous Intermediate Physical Address (IPA) space. At the second stage, Jailhouse maps IPAs of different cells to Physical Addresses (PAs) with a configurable color specification. By doing so, Jailhouse supports the notion of temporal partitioning wherein, by assigning a collection of non-overlapping sets of processors to colored inmates, those cells cannot access a (physical) address beyond their own address space resulting in interference prevention.

Jailhouse gets enabled on top of a Linux, called root-cell, and uses a configuration file associated with it to split off parts of the system's resources and assign them to desired cells. Last but not least, Jailhouse is especially useful since one can circumvent the TrustZone switches in any ARM-adapted OS (e.g., Linux), exploiting the hypervisor calls.

- 1) *PS-PL SoCs:*
- 2) *Memory-Loopback:*
- 3) *Cache Bleaching:*
- 4) *Transaction-level Inspection:*

C. Cache and DRAM temporal partitioning

- 1) *Partitioning:*
- 2) *Run-time Zero-Copy Recoloring support:*

D. Advanced eXtensible Interface (AXI)

The Advanced eXtensible Interface (or AXI for short) is a widely spread open specification bus protocol proposed by ARM [1] and exploited by Xilinx to interface both the PL side and the PS side.

The AXI protocol is based on the master-slave duality. Typically, the former is in charge of instantiating transactions directed to any of the slaves composing the system. On the other hand, the latter, does not emit any transaction but receive the requests emitted by a master and answer them. The masters and the slaves communicate between each other through five different channels named AW, W, B, AR and R as illustrated in figure 1. A write transaction will start first by its address phase ①. That is, the transmission through the channel AW of meta-data regarding the transaction such as the destination address, the transaction ID, the amount of bursts and so on. Upon the completion of this phase, follows the data phase ② which, as its name suggests it, consists in the transmission of

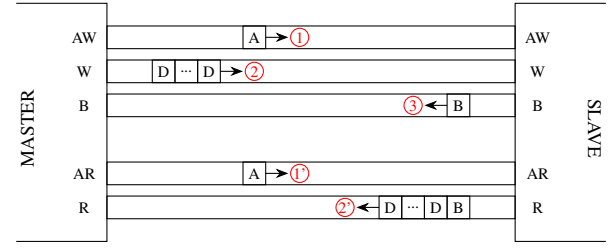


Fig. 1: Caption

the payload itself through the W channel. Thereafter, comes the response phase ③ performed on the B channel. This phase, the only one being initiated by the slave, is used to inform the master whether the transaction has been completed successfully. The transmission of a read transaction is carried out in a similar way. In fact, as for the writing, the address phase ① is transmitted through the equivalent channel called AR and is directly followed by the data phase ②. However, unlike the writing scheme, the data being fetched, the data phase is instantiated by the slave. The reading response phase is performed simultaneously and is thus merge within the R channel.

The protocol is said to be asynchronous as transaction behaves similarly to packets each having a dedicated ID. Hence, multiple outstanding transactions can be emitted successively by a single master which can managed them in an out-of-order manner.

IV. HIGH-LEVEL OVERVIEW

In this section, we provide a high-level description of the organization of the proposed scheduler in-the-middle. As previously mentioned, the SchIM leverages the PLIM approach originally proposed in [3]. Indeed, 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. It is always possible, however, to dynamically exclude the SchIM and instead route transactions directly to main memory. For the purpose of this paper, however, we mainly consider a setup in which all the CPU-generated memory transactions are handled by the SchIM.

Figure ?? provides an overview of the location of the SchIM within the main components of the target platform. The highlighted **COLOR** path depicts the route that memory requests originated by the CPUs follow when the SchIM is used. Application memory requests can reach the SchIM through multiple interfaces. Without loss of generality, we consider a SchIM instance with two arrival lanes, which are labeled as AXI₁ and AXI₂ in Figure ?. The SchIM then forwards the received transactions towards main memory through the AXI₃ interface. A more detailed view of the SchIM module

¹The source code is available at <https://github.com/siemens/jailhouse.git>.

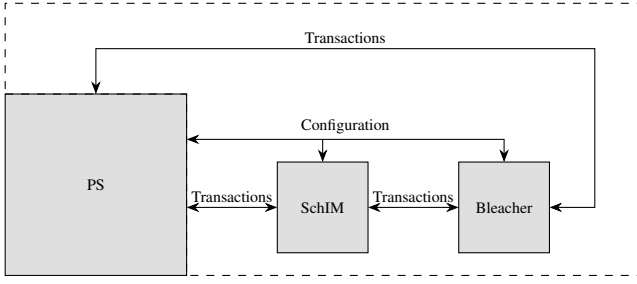


Fig. 2: High-level block design of the proposed SchIM module. *RM: we need to replace this picture with one that provides more details about the main blocks of the platforms. We also do not need to include the bleacher as it is not fundamental for the operation of the SchIM.*

is provided in Figure 3 where the same convention is used to identify input and output ports. In addition, as shown in Figure 3, a fourth AXI_4 port is used to configure the SchIM module from the PS.

A. Micro-architecture

The SchIM module 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 in charge of interfacing the core logic of the SchIM with the rest of the system using the AXI protocol. This domain is comprised of three sub-modules. These are (i) the *packetizer*, (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 the rest of the SchIM. Packetization of AXI transactions is necessary to be able to store transactions that are serial by nature. Indeed, 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-transmission. 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 CPU cores 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

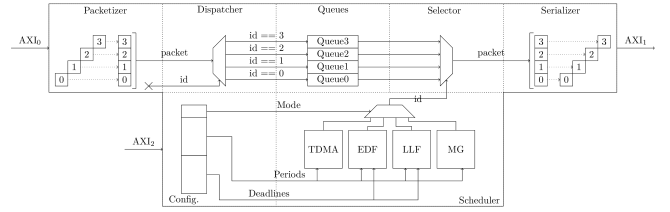


Fig. 3: Caption

information to the scheduling domain regarding their current state, and (b) they can generate a congestion control signal to the associated CPU core.

As suggested by figure 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 decisions of the scheduling domain.

Additional important details about the categorization and congestion control mechanisms enacted in the queuing domain are provided in Section ??.

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 a first set of three template schedulers is provided as a proof of concept. The scheduling policies currently implemented in the SchIM are Fixed Priority (FP), Time Division Multiple Access (TDMA), and Budget-based Traffic Shaping (TS). Each of the parameters required by the implemented policies — such as the priorities, the periods, and the budgets — can be adjusted at run-time 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. Finally, with the TS scheduler, it is possible to associate a maximum rate at which transactions from each queue are forwarded by the SchIM.

V. SCHIM IMPLEMENTATION

The present section exposes how SchIM interacts with the remaining of the systems in V-A, how its internal logic enforces transaction scheduling policies in IV-A and finally, an example of the transaction life cycle within the SchIM module is provided in V-C.

A. Altered communication scheme

In order to achieve the objective of re-ordering transactions, one must alter the standard AXI communication scheme explained in the subsection III-D. To this end, we instantiate a

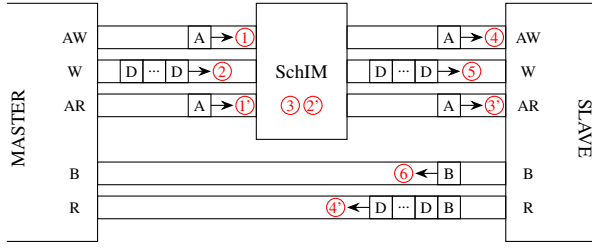


Fig. 4: Caption

in-between the master and the slave a pass-through module named SchIM as depicted in figure 4. This module is the one in charge of the re-ordering of the transactions and does so by intercepting on the fly the transactions emitted by the masters before they reach the desired slaves. As shown in the figure 4, only the phases instantiated by the masters (i.e. address phase on AW and AR and the data phase on W) are intercepted for re-ordering by SchIM. The introduction of SchIM has a direct consequence on the overall communication scheme. Indeed, while the response phases on channels R and B remain unchanged, the address and data phases are duplicated. 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 module in ③ and only repeated when decided by the latter's internal logic. This release of the transaction leads to the initialisation of two new address and data phase ④ and ⑤. Finally, the response phase ⑥ goes directly from the slave to the master without being intercepted. The same modifications apply to the transmissions of read transactions as the address phase ①' is being buffered in ②' for some time before being re-emitted in ③'. As for the writing, the response phase ④' is not intercepted by SchIM.

B. Embedded Schedulers

1) *Fixed Priority*: The Fixed Priority scheduling module aims at enforcing prioritization of the traffic of the cores. The priority ordering is explicitly defined by the user through the configuration port. While SchIM only has four queues, 16 different levels of priority are offered. Note that, as it stands, the ordering is strict, meaning that two cores cannot be assigned with the same priority.

The FP scheduling module only needs two information: the priority associated with each queue and whether a given queue contains at least one buffered transaction. Intuitively, the module logic consists in always considering the highest priority hence, a argmax function is implemented. However, lower priority queues must also be considered when higher priority queue do not have transactions. Consequently, the user defined priority is temporarily considered as 0 when its corresponding queue does not contain transaction.

2) *Time Division Multiple Access*: The Time Division Multiple Access (or TDMA) scheduling module provided by SchIM is a non-work conserving policy that splits the bus usage between the cores for given periods (also referred to as

slots). The proposed embedded TDMA module enables the user to specify a different TDMA slot size for each core. The periods are expressed in clock cycles, enabling a fine grained granularity. The configuration port enables the end-user to specify and change the periods at runtime.

The implementation of the modules relies on a single register used as a counter. In fact, we use this register to count the time elapsed in the current TDMA hyper-period (i.e., the sum of all the cores period) and is reset to 0 once this hyper-period reached. Alongside this register, a logic is there to determine in which core slot the counter actually is and to forward the information to the queue selector. The logic is able to determine the core to schedule by summing the period of each previous cores. Provided that the current value of the counting register is contain between the sum of the previous periods and the sum of the previous periods and the current one, the information forwarded to the remaining of the system will be the core id corresponding to the interval.

DH: @DH TODO improve description!

3) *MemGuard*: The proposed MemGuard transaction scheduling policy is inspired by the Software-based bus regulation technique ???. The latter has been proved to ensure memory isolation for all the cores involved. Unlike the original MemGuard, the proposed version does not rely on memory budgets and replenishment periods. Instead, it enforces inter-arrival time (i.e. a guaranteed minimal period) between two consecutive transactions emitted by a given core. This distinguishes our approach from [2].

This scheduling module is implemented as follows. For each of the queues to schedule, the module has a register counting the time elapsed since the start of the period. Once this counter has reached the period set by the user (through the configuration port), the module checks if the queue corresponding to the core contains any transaction. In the case where a transaction is available in the corresponding queue, the latter is forwarded to the output of SchIM (i.e., the serializer) and the counting register is reset to 0. Otherwise, the counting register is blocked to the desired period until a transaction is available for scheduling in the corresponding queue, leading to the counting register to be reset to 0. Any tie between two cores is solved using a fixed priority arbitration defined by the user thanks to the configuration port.

C. Transactions Life Cycle

Let us consider a system with four cores c_0, \dots, c_3 sending transactions $T = \{t_0, t_1, \dots, t_n\}$ to the SchIM module. Consequently, the latter boasts four queues (noted $Q = \{q_0, q_1, q_2, q_3\}$) buffering the transactions under the form of packets $P = \{p_0, p_1, \dots, p_n\}$ where $p_i = \text{Packetizer}(t_i) \forall i \in [0 : n]$.

In the present example, we will assume t_1 as being the transaction under analysis. The latter is emitted by c_2 in direction of the SchIM module. The packetizer receives this transaction and, once the AXI protocol completed, transform it into an equivalent packet $p_1 = \text{Packetizer}(t_1)$. Following

this transformation, the newly created packet is forwarded to the dispatcher which, thanks to the emitter's id embedded within the transaction, is re-routed to the corresponding queue q_2 (since emitted by c_2). After the insertion of p_1 in q_2 , the state of the queuing domain is as follows: q_0 has two packets p_0 and p_k and q_2 only has p_1 . At this point, q_0 is considered for scheduling by the scheduling domain. In consequence, p_0 is forwarded to the serializer through the selector. Simultaneously to the reception of the packet by the serializer, the latter receives an activation signal from the scheduling domain informing the serializer that the packet is valid and that a transaction can be started. Similarly to the packetizer, the serializer will transform the packet p_0 back to its initial AXI transaction form $t_0 = \text{Serializer}(p_0)$. Thereafter, once the t_0 has been sent, the serializer will inform the scheduling domain via a signal, that he is ready to accept the next packet as input. Upon the reception of this signal, the scheduling domain will both re-direct the latter to the queue of the previous packet to indicate that it has been consumed and change the selected queue according to the scheduling policy so that the first packet of this queue can be forwarded to the serializer through the selector module. In the present example, the "consumed" signal forwarded by the scheduler is sent to q_0 which is then empty. At this instant, two scenarios are possible:

- 1) q_0 is still considered for scheduling following the selected scheduling policy. Therefore, as q_0 is empty, it outputs an "empty" signal received by the scheduling domain. The latter then decides to not send any activation signal to the serializer because there is nothing left to transmit in the selected queue. In other words, the access to the main memory is being stalled on purpose by the scheduling policy i.e. the scheduling policy is not work conserving. For instance, such a scenario could happen in the case of TDMA or if all the queues are empty. The logic will resume as soon as the selected queue is filled.
- 2) q_2 is now considered instead of q_0 for scheduling. In this case, the "consumed" signal is repeated to q_0 while the queue ID changes in order to select q_2 . This results in the packet contained inside q_1 to be forwarded to the selector.

VI. PL-TO-PS FEEDBACK

While the combination of the PS and the PL sides offer great opportunities, it also comes with challenges. In fact, each of the HPM ports interfacing the PS and the PL sides (HPM0 and HPM1) 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-the-line blocking can happen. Head-of-the-line blocking can be harmful for performance or simply cancel all the efforts put in place to enforce transaction reordering and core isolation. 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 TDMA slot of the core, leading to a worst case scenario of having to wait for a full TDMA hyperperiod. On the other hand, for work-conserving policies (e.g., FP), the decisions being taken by SchIM would be totally dependent on the order at which transactions are being emitted by the HPM port buffer, cancelling the policy.

In both cases, one might wish to prevent cores from saturating the HPM port buffers. In order to avoid such situation, we implemented a feedback scheme aiming at slowing down the cores. More accurately, the SchIM architecture presented in V has been extended with each of the queues being associated with a comparator comparing the current amount of buffered transactions with a threshold defined by the user through the configuration port. If the threshold is reached, an interrupt is sent at the same clock cycle to the PS side. Each core has a dedicated interrupt line considered as a FIQ. The FIQ is handled by a routine in the hypervisor that set a DBS instruction (or Data Barrier Synchronisation) in order to make sure that will not perform any new transaction as long as all the pending transaction have not been served.

Ideally, these buffers should be shared evenly amongst the cores. In our case, since each HPM port has a buffer with a depth of 8 for each type of transactions, each core should occupy at most 2 slots in each buffer. Unfortunately, from our experiments, the control of the exact amount of transactions for each core in the buffers is coarsened. Most of the time, the threshold will be exceeded by one or two transactions. By adding another slave port on which we connect the second available HPM port (namely, HPM1) and assigning two cores two each of the two ports, the ideal amount of transaction from each core being buffered can be doubled, letting us with a bigger margin of error. This is the reason why SchIM includes two slave ports even though one is enough.

DH: Explain how, intuitively, we can choose the values for each proposed Scheduling policies.

VII. EVALUATION

We have performed a full system implementation on a Xilinx ZCU102 development board, featuring a Xilinx Zynq UltraScale+ XCZU9EG SoC. For this work, we utilize a combination of synthetic and real benchmarks. Our synthetic benchmark.

Talk about IO intensive and memory intensive benchmarks to be designed.... We also include a study of the behavior of real applications from the San Diego Vision Benchmark Suite (SD-VBS) [4], which comes with multiple input sizes....

DH: Maybe we could use AES from MiBench ?

As mentioned earlier, the high-performance master ports, namely HPMs, serve as a gateway from the PS to the PL. We implemented our scheduler IP, SchIM, responding under the physical addresses of HPM0 right after the PS. By doing so, any transaction toward the PL has to go through the

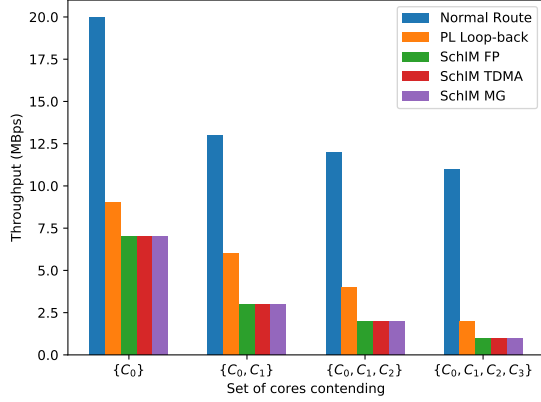


Fig. 5: Board bandwidth

TODO: outdated version. Has to be repeated for cached transactions only with the latest version of SchIM.

SchIM. Hence, all memory requests yield to the scheduling policy being enforced by the scheduler. Design features of SchIM implementation and configuration interface are detailed in section ??.

At the preprocessing stages, an AXI Performance Monitor (APM) unit connected to the bus interface between the HPM0 and the scheduler. It is crucial to affirm that this medium does not affect the transaction flow toward the PL. APMs are powerful tools available in the PS and instantiable in the PL capable of measuring primary performance metrics (for AXI4, AXI4-Lite, or AXI4-Stream-based systems) such as bus latency for specific master/slave or amount of memory traffic for the particular duration. Moreover, APM offers the functionality of logging the necessary information about data transfers between any master and slave communicating in the AXI protocol. In this work, we program APMs to profile the behavior of the benchmarks to analyze the task's deadline. SchIM employs profiled information on task deadlines to make scheduling decisions (e.g., enforce EDF policy) for hard real-time jobs.

A. Performance degradation

B. Platform Capabilities

Here, discuss Fig.5

Broad evaluation of the bandwidth offer by the ZCU102 depending on the route considered.

C. Internal Behaviour of SchIM

Here, show and emphasize on the behaviour of SchIM with Fig.6

Here, we use the trace snapshots

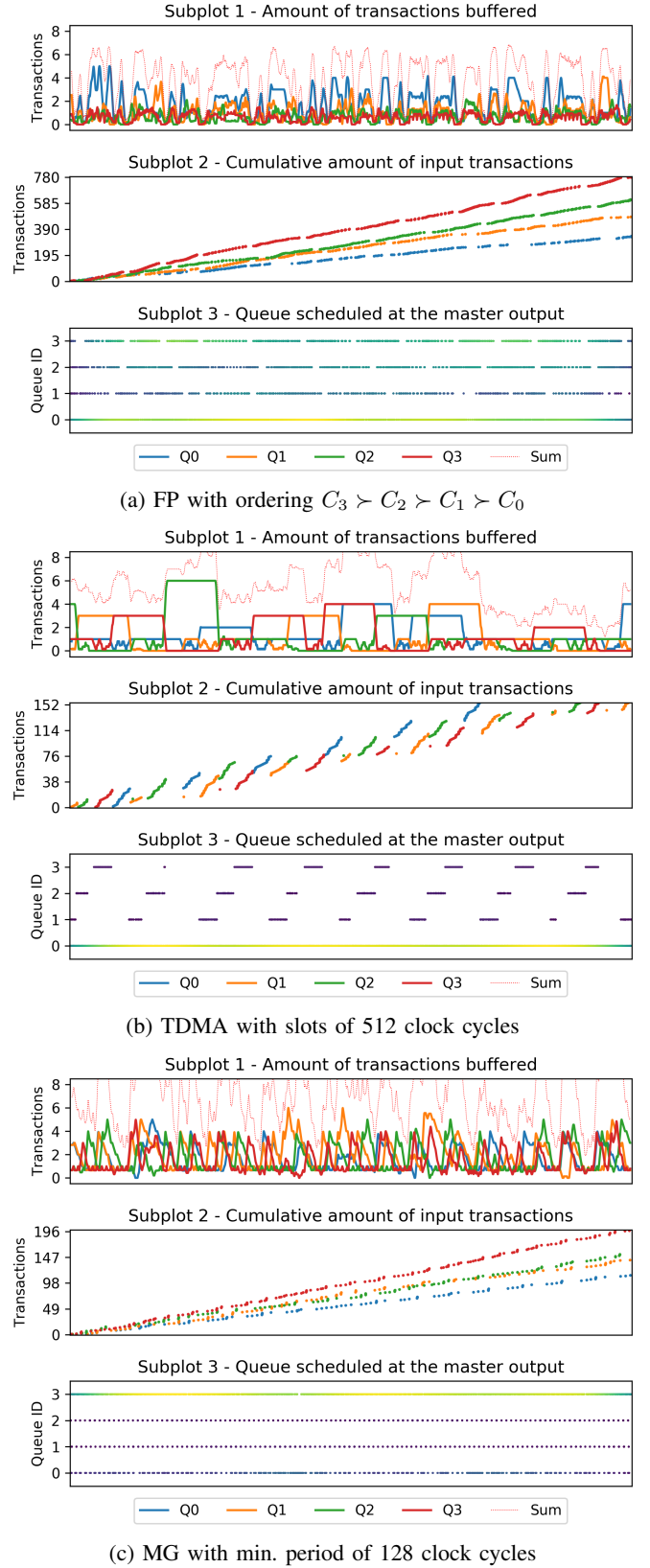


Fig. 6: Trace snapshot of SchIM for the three implemented policies

TODO: The plots have to be reworked.

D. Memory Isolation

Here, prove that SchIM enables us to isolate the cores. We need a bunch of benchmarks competing with mem-bombs on the remaining cores. We will try:

- FP: The benchmark running alone with the highest priority VS. the same setup with mem-bombs
- TDMA: The benchmark running alone with all the cores having a slot of the same size VS. the same setup with mem-bombs
- MG: The benchmark running alone with all the cores having a given periodicity VS. the same setup with mem-bombs

VIII. OVERHEAD

DH: Is this section still relevant ?

A. Discussion

IX. CONCLUSION

DH: As stressed by @Renato, we must mention that SchIM is the building brick for profile-based Bus scheduling/regulation!

REFERENCES

- [1] ARM, "Amba axi and ace protocol specification," Tech. Rep., 2019. [Online]. Available: https://static.docs.arm.com/ih0022/g/IHI0022G_amba_axi_protocol_spec.pdf
- [2] F. Farshchi, Q. Huang, and H. Yun, "Bru: Bandwidth regulation unit for real-time multicore processors," *2020 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pp. 364–375, 2020.
- [3] R. M. S. Roozkhosh, "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, conference.
- [4] 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)*, 2009, pp. 55–64.