The Best of All Worlds: Improving Predictability at the Performance of Conventional Coherence with *No* Protocol Modifications

Salah Hessien
McMaster University, Canada
salahga@mcmaster.ca

Mohamed Hassan
McMaster University, Canada
mohamed.hassan@mcmaster.ca

***Abstract*—Tasks in modern embedded systems such as automotive and avionics communicate among each other using shared data towards achieving the desired functionality of the whole system. In commodity platforms, cores communicate data through the shared memory hierarchy and correctness is maintained by a cache coherence protocol. Recent works investigated the deployment of coherence protocols in real-time systems and showed significant performance improvements. Nonetheless, we find these works to suffer from two main drawbacks. 1) They suffer from significant latency delays due to coherence interference. 2) They require amendments to existing coherence protocols. This represents a significant obstruction hindering the industry adoption of these proposals since it requires to re-verify the coherence protocol. Coherence verification is considered one of the most complex challenges in computer architecture, which makes it inconceivable for chip manufacturers to adopt modifications to their already verified protocols that they have stable for decades.**

In this work, we propose PISCOT: a predictable and coherent bus architecture that (i) provides a considerably tighter bound compared to the state-of-the-art predictable coherent solutions ($4\times$ tighter bounds in a quad-core system). (ii) It does so with a negligible performance loss compared to conventional high-performance architecture coherence delays (less than 4% for SPLASH-3 benchmarks). This improves average performance by up to $5\times$ ($2.8\times$ on average) compared to its predictable coherence counterpart. Finally, (iii) it achieves that without requiring any modifications to conventional coherence protocols.

I. Introduction

Multi-core platforms are the norm nowadays in all computing systems, and real-time embedded systems are no exception. Multi-core platforms are envisioned to be the solution for the increasing computational and data demands in modern real-time embedded systems such as those deployed in automotive, avionics, and Internet-of-things (IoT). Nonetheless, multi-core platforms bring their own challenges. One of the biggest challenges is the interference among various cores in the system while competing to access shared hardware resources such as memory buses, shared caches, and off-chip memories. This interference hinders the system analyzability since the execution time of a task on one core now depends on the run-time behavior of tasks running on other cores. In order to provide the timing guarantees mandated by the real-time tasks, the hardware itself must be predictable such that the delays resulting from the aforementioned interference can be analytically bounded. To address this challenge, several efforts have been proposed to provide predictable memory buses [1]–[3], shared caches [4]–[8], and off-chip memories [9]–[11].

Despite being effective in managing the timing interference, most of these solutions assume that tasks are completely isolated with no communication among each other. We find this assumption to limit the applicability of these solutions in practical embedded systems, which require inter-task communication such as those deployed in automotive [12], and avionics [13]. Consequently, recent approaches investigated the communication among tasks through shared data [14]–[25]. Among these approaches, this paper is focusing on enabling tasks to communicate and share data by deploying hardware cache coherence, which is the approach followed by [17]–[25]. This is because cache coherence is the most commonly followed approach by commodity multi-core platforms [26], it improves overall system performance, and it does not impose any restrictions on the embedded legacy software or the operating system. In spite of their performance benefits, previous predictable coherence works [18], [21] suffer from two drawbacks. 1) They require major modifications to commodity coherence protocols (and hence, the hardware cache controllers). Those modifications are difficult to adopt by industry because of the significant time and intellectual effort required to implement and verify coherence protocols [27], [28]. 2) They suffer from extremely pessimistic worst-case latencies (WCLs) that reach to thousands of cycles for a single memory request to the shared cache as we explain in detail in Section III. Motivated by these limitations, this paper proposes a solution to allow for coherent sharing of data in multi-core real-time systems without requiring any changes to the coherence protocol, while notoriously reducing the WCL upon accessing the cache hierarchy. This is achieved by proposing PISCOT as a memory bus arbiter resembling the following contributions.

Contributions. 1) We study the deployment of coherence protocols using traditional predictable bus arbiters and investigate the sources of significant latency increase due to coherence interference (Section III). Our study shows that most of the traditional arbitration schemes widely used in the real-time embedded systems domain are data-sharing oblivious. Therefore, to enable coherent data sharing while minimizing

the coherence delays, we need a novel bus architecture that accommodates for this sharing by design. 2) Motivated by this observation, we propose PISCOT, a predictable and coherent bus architecture that substantially reduces coherence delays, while improving overall system performance (Section IV). This is achieved by decoupling the data responses from their coherence requests, implementing a split-transaction interconnect with two separate buses. While the coherence requests are arbitrated using Time Division Multiplexing (TDM) to ensure predictability, the data responses are managed in a First Come First Serve (FCFS) fashion to increase average performance. Balancing the trade-off between predictability and performance is one of the main requirements of modern embedded systems. Unlike existing solutions, PISCOT does not require any modifications to the underlying coherence protocol. This is key since modifications to coherence protocols are both hard to be adopted by commercial chips and hard to verify [22], [27], [28]. PISCOT should be implemented as a predictable hardware arbiter managing accesses to the shared memory. Although implementing a bus arbiter in conventional systems will definitely require hardware modifications, two main differences are worth noting. 1) It is a must to deploy a predictable bus arbitration to ensure timing guarantees anyway regardless of supporting coherence or not. 2) Verifying a bus arbiter is a notoriously easier task than a coherence protocol, which makes it more appealing to adopt by industry. 3) We conduct a detailed timing analysis for the latency suffered by any memory request. The analysis provides an analytical bound that guarantees the system predictability (Section V). The derived bounds are $4\times$ tighter than the state-of-the-art predictable coherent buses [18], [20], [21] for a quad-core system. 4) We deploy PISCOT in two different cache architectures currently adopted by commercial embedded systems. In the first, cores communicate only through the shared cache, while in the second, there is a direct cache-to-cache communication bus to increase efficiency. 5) We evaluate PISCOT with both the representative SPLASH-3 benchmarks as well as synthetic benchmarks. Comparisons with existing solutions show that PISCOT achieves up to $5\times$ better performance ($2.8\times$ on average), while increasing memory bandwidth utilization by $12\times$ on average across the SPLASH-3 benchmarks.

II. Cache Coherence: A background

One of the key contributions behind PISCOT is that it offers predictable and coherent data-sharing without the need to apply any changes to the coherence protocol itself. In this paper, we exemplify by integrating PISCOT with the foundational Modified-Shared-Invalid (MSI) protocol. Despite of its simplicity, MSI is the foundation of coherence protocols deployed in most existing architectures such as the MESIF protocol in Intel's i7 and the MOESI protocol in AMD's Opteron [29]. Figure 1 delineates the complete state diagram of MSI. For MSI, there are three stable states for any cache line in a core's private cache; *modified* (M): meaning that the cache line is valid and modified (i.e. written), *shared* (S): meaning that it is valid but only read, and *invalid* (I): indicating a cache line that either does not exist in the private cache or has a stale data (a cache miss). A cache line can be in the S state in multiple cores' private caches. On the other hand, to maintain data correctness, only one core can have a cache line in the M state at any time, while all other cores will have this line in the I state. If a core has a load (store) miss to a cache line, it will issue a GetS() (GetM()) coherence message on the bus, and once it receives the data in its private cache, it moves to the S (M) state. A load to a cache line in the S or M state will be a cache hit. A store to a cache line in the M state is also a cache hit. Contrarily, a store to a cache line in the S state has to broadcast a coherence message on the bus (either a GetM() or an Upg() based on the deployed coherence protocol details) to inform other cores that might be in S state to invalidate their lines. A core with a cache line in the S or M state that observes in the bus a GetM() message from another core to the same line (called OtherGetM()) has to move to the I state. If the core was in M state it also has to send the updated data to the shared memory and/or the requesting core based on whether there is a communication interconnect between private caches as we discuss later in this section. A core with the cache line in the M state upon observing a GetS() of another core (OtherGetS()) has to send the updated data similar to the previous situation, while moving to the S state.

Transient States. The aforementioned transitions between states do not usually happen atomically. They are usually interrupted by other requests from other cores as requests to the memory bus from different cores are allowed to interleave (i.e. there can be multiple pending requests at the same time) to increase system performance. For instance, a request can be pending for data to be fetched from the main memory; hence, the system allows for other younger requests to proceed if their data is already ready to increase overall throughput. During these interruptions of a request, the cache line may need to change its state to keep track of the updated coherence events on the bus, and this is the rule of *transient states*. Generally, a cache line moves to one or multiple transient state(s) in its journey from one stable state to another. In the interest of this paper, we classify transient states into four distinct categories.

- **Waiting for data and message states:** A core in these transient states has a pending request that is not granted access to the bus yet by the arbiter. Once the request message is issued on the bus, due to the reorderings and delays that can happen in the bus and its non-atomic nature, the core can first see either its coherence message or the requested data.
- **Waiting for data states:** These states indicate that the core has already observed its coherence message but is yet waiting for data.
- **Waiting for message states:** A core will be in one of these states if it receives its data before observing its coherence message.
- **Response to other requests states:** While a core is in one of the aforementioned three categories, it can observe requests from other cores (OtherGetM() or OtherGetS()) to the same

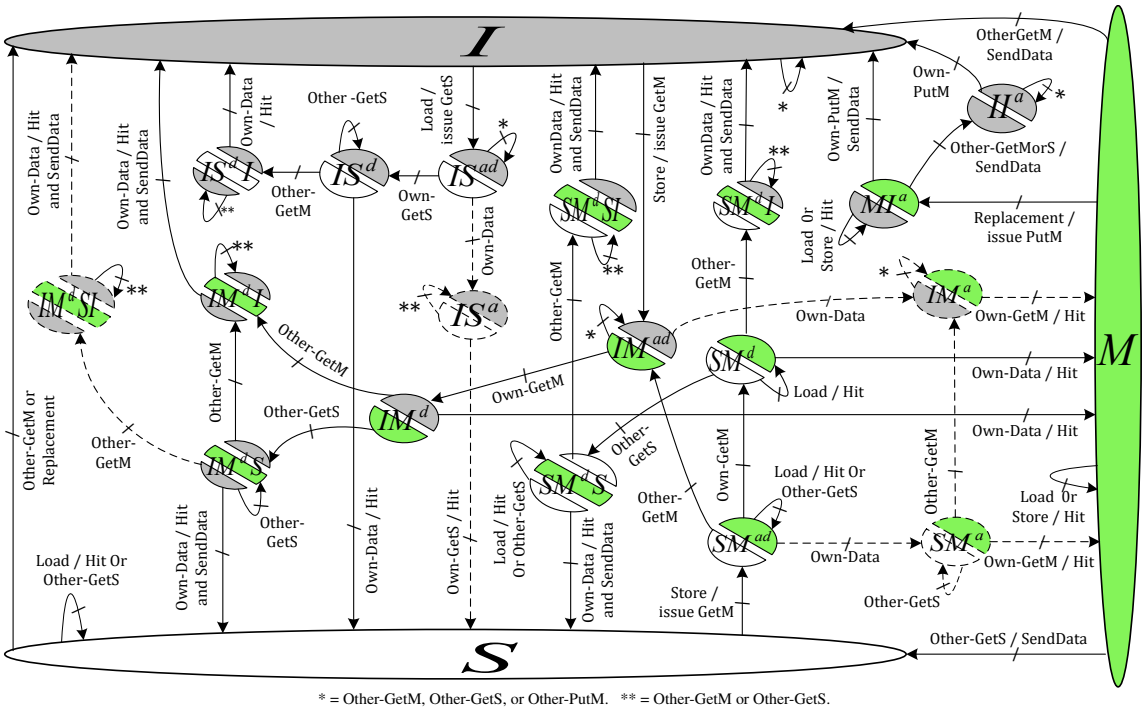


Fig. 1: MSI Coherence Protocol with transient states. States and transitions dashed are those that can be removed from the original MSI protocol under PISCOT.

cache line. Hence, it may need to move to another transient state to acknowledge the receiving of this request.

To illustrate the four categories, assume a load request to a cache line that is in the I state. Once the request misses in its private cache, the core queues a GetS() message to its local buffer waiting to be granted access to the bus by the arbiter. In this case (as Figure 1 shows), the line will move to the IS^{ad} state, which indicates that the core is waiting both to observe its message and receive the data. Afterwards, if the core observes its coherence message on the bus, it will change its state to IS^{d} indicating that it is now waiting for data. On the other hand, if it receives the requested data before observing its coherence message, it has to change its state to IS^{a} and wait for its message to appear on the bus. This is necessary since these broadcasted messages are the contract between all cores guaranteeing that they all observe changes to cache lines in the same order; otherwise, data inconsistencies will exist among cores. An example state from the fourth category happens if the core while in the IS^{d} state, observes an OtherGetM() to the same cache line. As a result, it has to move to the $IS^{d}I$ state. This state indicates that the core after receiving its requested data and conducting its load operation, has to invalidate its cache line since there is another pending store request from another core to the same cache line.

Cache-to-Cache Communication. Two architectural models are considered with regard to how data is transferred among cores' private caches. The first model covers architectures that do not employ a direct cache-to-cache interconnect. In this case, the owner core always has to send the data to the shared memory (such as the last-level cache (LLC)). Afterwards, the shared memory sends this data to the requesting core. The second model represents architectures that support direct cache-to-cache communication. In this model, the owner core sends the data directly to the requesting core. In addition, if the requesting core's message was a GetS() (meaning that it is a load request), the owner also has to send the data to the shared memory since the shared memory will be the owner in this case. In both models, if there is no owner core (i.e. no core has the requested line in the modified state), the shared memory is the owner and it is responsible for sending the data to the requesting core.

III. Motivation

Three main observations motivate this work. 1) Exiting solutions supporting coherence data sharing in commodity platforms are designed for performance. Accordingly, they provide no timing guarantees, and thus, cannot be safely used in real-time systems. 2) Traditional real-time arbiters designed for predictability are not considering data sharing among tasks, and it has been shown that even when using such predictable arbiters, they can lead to unpredictable behaviors when considering such sharing using coherence [18]. 3) Recent solutions that support coherence sharing of data are building on top of these traditional arbitration schemes. This leads to two significant drawbacks in these solutions. First, despite achieving predictability, the guaranteed latency bounds are notoriously large (in the range of thousands of cycles for a single request) [18], [21] which can be infeasible for systems with

TABLE I: Positioning PISCOT compared to existing approaches.

Approaches	Arbiter	Shared Data Support	Coherence Protocol	Predictability	Examples
COTS platforms baseline	High performance	✓✓	✓✓	✗	FCFS [30], [31], split-transaction [32]–[35], priority-based [31], [36]
Traditional Real-Time Arbitration	Predictable by-design	✗(not data-aware)	✗	✓✓	TDM: [1], [3], [37], RR: [38], Harmonic RR (HRR): [39], weighted RR: [2]
Data-Aware Arbitration	builds on traditional arbitration	✓✓	✓(requires coherence modifications)	✓(with significant latency bounds)	PMSI [18], CARP [21], HourGlass [19], PENDULUM [20]
PISCOT	Predictable split-transaction	✓✓	✓✓(with no changes to existing protocols)	✓✓(with tight latency bounds)	–

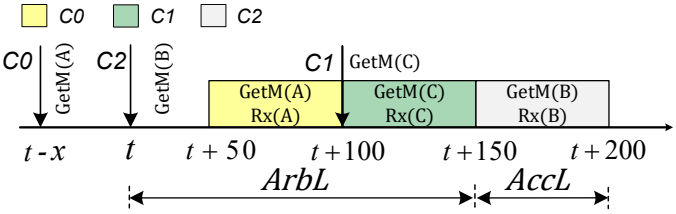


Fig. 2: Traditional TDM arbitration with no shared data.

tight timing requirements. Second, they support coherence by proposing amendments to existing coherence protocols, which handicaps their adoption by industry in commercial platforms. We summarize these observations in Table I, and we discuss them in details in the following subsections.

A. Commodity Performance-Oriented Arbitration

Arbitration among different requests in COTS platforms is usually realized using a high-performance arbiter that favors system performance over other metrics such as fairness and predictability. Such arbiter prioritizes requests based on their arrival time (age-based priority), where older requests are serviced before younger ones. A common example of such arbiter is the First-Come First-Serve (FCFS) scheme [30], [31]. Such arbitration is not predictable since it provides no latency guarantees upon accessing the shared memory. This is because one core can have a request that is pending (theoretically) forever, while other cores are saturating the queues. In addition to age-based arbitration, some COTS platforms also deploy another level of fixed-priority arbitration to give higher-priority for requests from a certain processor. This also entails no guarantees are granted to lower-priority requests. A final observation about COTS arbiters is that for cache coherent systems, the bus is usually implemented as a split-transaction interconnect to increase system performance by concurrently handling both coherent requests (messages) and data responses [27], [32]–[35]. For instance, the ARM Corelink CCI550 dictates separate channels for snooping requests and their corresponding responses [34]. Similarly, the Intel's QPI designates different virtual channels to data and coherence messages [35].

B. Traditional Real-Time Arbitration

In multi-core real-time systems, access to the shared memory (e.g. the Last-Level Cache (LLC)) is managed through a predictable arbiter such as (TDM) [1], [3], [37], and Round Robin (RR) [38]. Considering the TDM arbitration example depicted in Figure 2, a request suffers a maximum latency of one TDM period before it is granted access to the bus. For a system with N cores, this is $N \cdot S$ cycles, where S is the slot width in cycles. This occurs when the requesting core just misses its own slot. Please note that throughout this section, we denote a core as Cx, where x is the core index. The GetM(B) from $C2$ in Figure 2 is an example of such a request, where it arrives to the private cache controller at timestamp t. Assuming that $C2$ just missed its own slot, it waits until $t+150$ to gain access to the bus. Since the system in Figure 2 has three cores, this is equivalent to a one TDM period of 3 slots assuming that the slot width allows for only one memory transfer (one request) and is 50 cycles. Once granted access to the bus, the request conducts its memory transfer consuming an extra slot (50 cycles) and finishes at $t + 200$.

The big limitation of this analysis is that it only applies if cores do not share data. In the example in Figure 2, all the cores request to access different cache lines. Consequently, the shared memory is able to respond with the correct data in the request's same slot. Unfortunately, this does not apply if cores are allowed to share data. It has been shown by [18] that shared data can lead to unpredictable behavior even when deploying a predictable arbitration such as TDM.

C. Coherent Shared-Data Aware Predictable Arbitration

To guarantee predictability while allowing coherent sharing of data, several recent arbitration solutions have been proposed [18], [20], [21], [25]. All these solutions assume a variant of the TDM arbitration scheme and propose coherence protocol as well as architectural changes to support predictability. Despite showing that coherence can lead to significant performance improvements in data-sharing real-time systems, they incur significant WCL bounds. To illustrate this drawback, Figure 3 delineates the TDM behavior for the same system in Figure 2 but with assuming that cores can share data, and hence, they issue requests to the same cache line, A. The example follows the protocol guidelines from PMSI [18]. It is clear from Figure 3 the significant added latency due to the coherence interference on the shared data. The request under analysis (GetM(A) from $C2$) in this case has to wait for every other core to receive the data of cache line A, conduct the store operation, and then write it back to the shared memory. Since the slot width of the TDM allows for only one memory transfer, and every core gets one slot per TDM period, every core now requires two TDM periods to conduct the aforementioned operation. As a result, $C2$'s GetM(A) request waits until timestamp $t + 1050$ in Figure 3

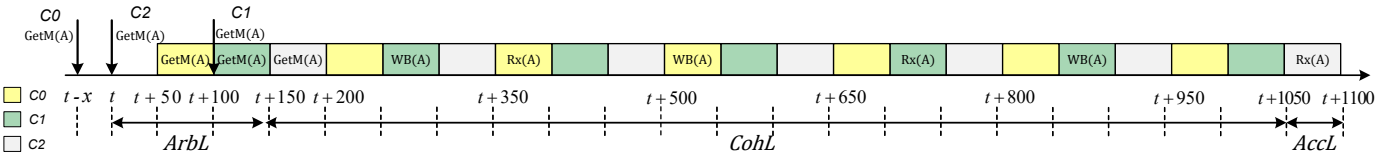


Fig. 3: TDM-based coherence approach [18]. Initially, C1 owns A in the M state.

before it can start receiving its requested data. Formally, for a system with N cores and a TDM arbitration with shared data, a request has to wait for up to $(2 \cdot N^2 + 2 \cdot N) \cdot S$ before it can start transferring its requested data [18]. The other existing solutions while supporting systems with mixed criticalities [20], [21], this comes at the expense of incurring even larger WCL than PMSI if all cores have the same criticality. The DISCO solution in [25] improves the WCL bounds by requiring a special handling of writes compared to reads.

It is worth noting that in Figure 3 it might seem that there are many idle slots, and thus, this large latency can be completely avoided using a work-conserving schedule. However, this is not true since there can be requests from other cores in the system that utilize these slots. They are not shown in Figure 3 for simplicity. For example, $C0$ receives its requested data at timestamp $t + 400$. Thus, it can issue another request afterwards in its coming slots. Clearly, in an out-of-order architecture, more pending memory requests can also co-exist in the system.

Two key observations we make in this paper about the existing predictable cache-coherent TDM-based solutions. First, their previously highlighted large WCLs are mainly because they inherit the scheduling paradigm of traditional real-time arbiters (such as TDM in this case but the argument applies to other arbiters such as RR). This paradigm when applied to systems with shared data, it couples two different types of communication into the same bus arbitration. Namely, it couples both coherence messages and data transfers and schedules them using the same bus arbitration, which is inherited from traditional non-data-sharing TDM schedules. This in addition to the fact that the TDM slot has to accommodate for at least one memory transfer to be efficient to service ready memory requests, leading to the excessively large memory delays when introducing data sharing. Second, they impose certain modifications to the coherence protocol to enable predictability. As previously discussed, modifications to coherence protocols are highly costly in terms of verification and are thus inconceivable to adopt by industry.

Based on these observations, PISCOT targets to enable data sharing in real-time systems, while significantly reducing the associated coherence delays by decoupling the two different communication types. This is achieved by using a split-bus architecture, where requests (through coherence messages) and responses (i.e. data transfers) are issued in different buses and are managed using different arbitration mechanisms. In addition, PISCOT does not impose any changes to existing coherence protocols; therefore, disburden system designers from the need to re-verify the coherence protocol.

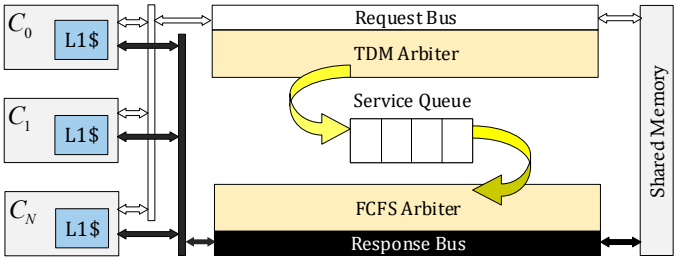


Fig. 4: PISCOT architecture.

IV. Proposed Solution

In this section, we detail the architectural details of PISCOT, which Figure 4 delineates its high-level modules. Compared to the solutions discussed in Section III and highlighted in Table I, PISCOT makes multiple architecture decisions to take into account predictability by design, while maintaining a high average-case performance.

- PISCOT's architecture migrates from the traditional arbitration schemes considered by the community (such as TDM and RR) to a split-transaction bus interconnect that connects private caches and the shared memory as Rule 1 explains.

  ***Rule** 1:* PISCOT implements a split-transaction bus through deploying two buses: a `Request Bus` and a `Response Bus`. The `Request Bus` is responsible for broadcasting the coherence messages initiating memory requests, while the `Response Bus` transfers data as a response to these requests.
- Aiming at performance, the `Request Bus` and the `Response Bus` operate in parallel. On the other hand, to simplify system analysis and maintain predictability, both buses communicate through only one module: the `Service Queue`. Requests broadcasted on the `Request Bus` are buffered into the `Service Queue` until they are selected by the `Response Bus`'s arbiter.
- Unlike conventional solutions that use high-performance arbiters at the expense of predictability (e.g. FCFS), the `Request Bus` in PISCOT is managed using a TDM arbiter to predictably manage interference among different cores (Rule 2). To increase system performance, a work-conserving TDM is deployed, where at any slot, if the dedicated core does not have a ready request, the arbiter picks the next core with a pending request instead of leaving the slot idle as in traditional non work-conserving TDM.

  ***Rule** 2:* PISCOT manages the `Request Bus` using a work-conserving TDM arbiter.

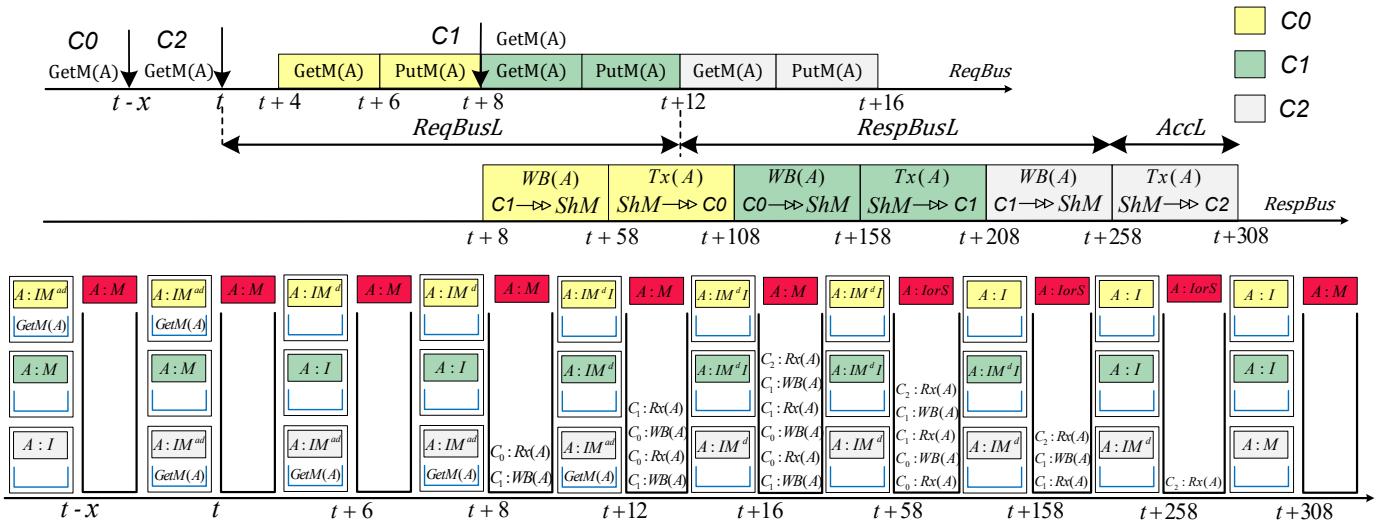


Fig. 5: An illustrative example for the operation of PISCOT. Latency components are for the getM(A) request from $C2$. At different time instances: the bottom of the figure shows the state of the private core's cache line (left side), shared memory state (marked in red), and the `Service Queue` contents on the right side.

- The `Response Bus`'s arbiter implements a First-Come First-Serve (FCFS) scheduler, and thus, serves requests based on their arrival time on the `Service Queue` (Rule 3). The oldest request will be at the head of the queue, and therefore, is serviced first by the FCFS response arbiter. Once selected by the FCFS arbiter, the requested data is transferred on the `Response Bus` to the requesting core's private cache, the request message is removed from the `Service Queue`, then the core proceeds with its load/store operation indicating that the request is successfully finished.

  ***Rule** 3:* PISCOT manages the `Response Bus` using a FCFS arbiter.
- PISCOT supports out-of-order execution and allows cores to have multiple outstanding requests. Nonetheless, according to Rule 4, those requests from a certain core would remain in its local buffer and will not be picked by the TDM arbiter if the core already has one request in-service (i.e. queued in the `Service Queue`). The rationale for this is to limit the coherence interference among cores such that a request from any core can suffer interference due to a maximum of only one request from each other core, which leads to tightening worst-case latencies and minimizing interference from other cores compared to the conventional MSI protocol with FCFS split-transaction bus as we detail in the latency analysis in Section V.

  ***Rule** 4:* PISCOT supports OOO architectures by allowing cores to issue multiple outstanding requests. However, to limit coherence interference, it only services at most one request from any given core at a time.

A. Illustrative Example

To better explain the operation of PISCOT, we use the same example from Section III for a system with three cores: $C0$–$C2$ and delineates PISCOT's behavior in Figure 5. The example focuses on a single cache line A, which is assumed to be initially owned by $C1$. At timestamp $t-x$, a store request to A from $C0$ misses in its private cache (it was originally in `I` state). As a result, a GetM(A) message is placed in its cache controller's local buffer waiting for $C0$'s slot on the request bus. The line state changes in the private cache from `I` to $\texttt{IM}^{ad}$ waiting for its message to appear on the bus. The same situation occurs for $C2$ at timestamp t. At $t+4$, $C0$ is granted a slot by the TDM arbiter and its request is issued on the `Request Bus`. The coherence message is assumed to consume two cycles to be broadcasted. Accordingly, $C0$ observes its OwnGetM(A) on the bus and move to $\texttt{IM}^{d}$ while waiting for data. On the other hand, once $C1$ observes $C0$'s GetM(A) (OtherGetM(A)) and since $C1$ is the owner of A, it responds with placing the updated data in its local buffer to be written back to the shared memory (timestamp $t+6$) and moves to `I` state. In addition, two actions are pushed into the `Service Queue` as a result of $C0$'s request. This is because $C1$ has to write back its updated A first to the shared memory and then the shared memory will send the data to $C0$; these are indicated in Figure 5 in the `Service Queue` as $C1$:WB(A) and $C0$:Rx(A), respectively.

Simultaneously at $t+8$, a GetM(A) request from $C1$ arrives and is issued on the `Request Bus` immediately since it is $C1$'s slot. Similar to what happened during $C0$'s slot, $C1$ moves to the $\texttt{IM}^{d}$ state and two actions are pushed into the `Service Queue`: $C0$:WB(A) and $C1$:Rx(A). The reason for this is that $C0$ should obtain its requested data first, according to the FCFS schedule, conduct its store operation, and write back the updated data to the shared memory before $C1$ can proceed with its GetM(A) request. For the same reason, $C0$ moves to the $\texttt{IM}^{d}\texttt{I}$ state. This is indicated at timestamp $t+12$. Now, $C2$ is finally granted access to the `Request`

Bus and issues its request. Similar events to those during $C1$'s slot occur with the difference that $C1$ is the owner responsible to write-back A before the shared memory sends it to $C2$ according to the FCFS order. For the Response Bus, it services requests in the Service Queue in order of their arrival as previously explained. Assuming that one data transfer requires 50 cycles, it finishes the data transfer of $C1$'s WB(A) to shared memory at $t + 58$. $C0$'s Rx(A) from shared memory at $t + 108$, performs its store operation and places the new data in its local buffer and moves to I state. $C0$'s WB(A) to shared memory finishes at $t + 158$. $C1$'s Rx(A) from shared memory at $t + 208$, performs its store operation and places the new data in its local buffer and moves to I state. $C1$'s WB(A) to shared memory finishes at $t + 258$, and finally $C2$ receives A from shared memory at $t + 308$.

Comparing this with the behavior of PMSI adopting the traditional TDM bus in Figure 3, it shows the clear advantage of PISCOT that reduces the total latency of the same sequence of memory requests by 792 cycles (from $t + 1100$ to $t + 308$). More detailed comparisons on the effect of both WCL as well as average performance are introduced in Section VI.

B. Satisfying Coherence Predictability Invariants

Coherence protocols can generally lead to unpredictable behaviors if not carefully managed. In addition, previous works have shown that combining conventional coherence protocols with traditional predictable arbiters also breaks system's predictability [18]. Since we claim that PISCOT indeed achieves predictability by utilizing conventional coherence while deploying the proposed split-transaction predictable arbiter, we believe it is necessary to elaborate more on how PISCOT achieves this predictability. Authors of [18] introduced 6 invariants that they stated that they must be satisfied to ensure predictability in the existence of coherence. We now show how PISCOT, unlike PMSI [18], is satisfying those invariants without the need to modify the coherence protocol. This discussion also illustrates the novel operation of PISCOT compared to traditional predictable arbiters such as TDM when tasks can share data. For inclusiveness, we state each invariant and then prove how PISCOT satisfies it. We prove each case by contradiction starting with a hypothesis that PISCOT breaks such invariant and then show that this contradicts PISCOT's operation explained at the beginning of this section.

***Invariant** 1:* A predictable bus arbiter must manage coherence messages on the bus such that each core may issue a coherence request on the bus if and only if it is granted an access slot to the bus.

***Lemma** 1:* PISCOT satisfies Invariant 1.

Proof: The proof is trivial since allowing a core to send a request without being granted access by the arbiter contradicts with PISCOT's TDM arbiter at the Request Bus. ■

***Invariant** 2:* The shared memory services requests to the same line in the order of their arrival to the shared memory.

***Lemma** 2:* PISCOT satisfies Invariants 2.

Proof: Let Req_i and Req_j be two requests to the same cache line such that Req_i arrived to the shared memory first. Assume that the shared memory serviced Req_j before Req_i such that Invariant 2 is broken. (1)

Now considering PISCOT's operation, Req_i will arrive to the shared memory first only if it is broadcasted on the Request Bus first. Hence, Req_i arriving at the shared memory first indicates that it has been queued into the Service Queue ahead of Req_j. Now, according to the Response Bus's FCFS, Req_i must be serviced before Req_j. (2)

(1) and (2) contradicts, which completes the proof. ■

***Invariant** 3:* A core responds to coherence requests in the order of their arrival to that core.

***Lemma** 3:* PISCOT satisfies Invariant 3.

Proof: Let $Req_i(A)$ and $Req_j(B)$ be two requests to cache lines A and B respectively that are owned by Core Ck such that Ck observes $Req_i(A)$ first. To break Invariant 3, PISCOT has to service $Req_j(B)$ before $Req_i(A)$. (1)

Now, according to PISCOT's operation, a core responds to a request for a cache line that it owns by placing the data immediately in its local buffer. Additionally, a WB action is queued into the Service Queue along with its initiating coherence message of the request itself during the same Request Bus's TDM slot. For instance, at time $t + 8$ in Figure 5, $C0$'s GetM(A) message resulted in pushing two actions to the Service Queue: 1) $C1$ has to write back A (WB(A)) first and only afterwards 2) $C2$ can receive its requested data (RX(A)) from shared memory. Since Ck observes $Req_i(A)$ first, it mandates under PISCOT that $Req_i(A)$ was issued in the Request Bus before $Req_j(B)$. Additionally, since requests are queued in the Service Queue based on their appearance timestamp on the Request Bus, it mandates that $Req_i(A)$ and its corresponding WB(A) are queued in the Service Queue ahead of $Req_i(B)$ and its WB(B). Finally, according to the Response Bus's FCFS policy, $Req_i(A)$ will get its data before $Req_i(B)$. (2)

(1) and (2) contradicts, which completes the proof. ■

***Invariant** 4:* A write request from a core that is a hit to a non-modified line in its private cache has to wait for the arbiter to grant this core an access to the bus.

***Lemma** 4:* PISCOT satisfies Invariant 4.

Proof: Let $Req_i(A)$ be a write request from core Ck to line A that Ck has in the S state in its private cache. To break Invariant 4, PISCOT shall allow $Req_i(A)$ to hit in the private cache and execute the operation silently without waiting for any permission from the bus arbiter. (1)

According to PISCOT's coherence protocol inherited from conventional MSI (Figure 1), A store to a cache line in S state has to issue a getM() coherence message and wait in the SM^{ad} state. Afterwards, this message is only issued on the bus once its core is granted access according to the Request Bus's TDM schedule. (2)

(1) and (2) contradicts, which completes the proof. ■

***Invariant** 5:* A write request from a core that is a hit to a non-modified line, A, in its private cache has to wait until all waiting cores that previously requested A get an access to A.

***Lemma** 5:* PISCOT satisfies Invariant 5.

Proof: Let cache line A to be initially in the S state in core Cj's private cache. Let also $Req_i(A)$ be a read request from core Ci to cache line A that is broadcasted on the bus at time $t1$. Then, assume that Cj at time $t1+\delta$ (where $\delta > 0$) has a store request $Req_j(A)$ to A. To break Invariant 5, assume that $Req_j(A)$ is serviced before $Req_i(A)$. (1)
However, from Lemma 4, it follows that $Req_j(A)$ has to wait for Cj's TDM slot on the `Request Bus` to broadcast a $GetM(A)$ message on the bus before it can proceed with its store operation. Assume that this happens at time $t2$. Since $Req_j(A)$ arrived at $t + \delta$, it follows that $t2 \geq t1 + \delta$. As a result and from Lemma 2, $Req_i(A)$ request is serviced before $Req_j(A)$ since $t2 > t1$. (2)
(1) and (2) contradicts, which completes the proof. ■

***Invariant** 6:* Each core has to deploy a predictable arbitration between its own generated requests and its responses to requests from other cores.

***Lemma** 6:* PISCOT satisfies Invariant 6.

Proof: Assume a system with N cores C0 to CN such that one core Ci, $0 \leq i \leq N$, has a request to service from the memory, say Req_i, while all the other $N-1$ cores keep issuing requests to cache lines that are modified (owned) by Ci. To break Invariant 6, Ci keeps servicing these requests and is not granted a guaranteed time at all where it can finish its Req_i request. (1)
Now, we discuss how PISCOT schedules these requests. First, each core can only issue requests during its dedicated TDM slot (Lemma 1). Second, an owner core responds to requests from another core immediately during this other core slot and not its own slot (Lemma 3). Accordingly, for our dictated scenario, Req_i has a guaranteed time slot to be issued on the `Request Bus`. Finally, since the `Response Bus` services requests in their order on the `Service Queue`, Req_i is guaranteed to finish its data transfer once all requests in front of it in the `Service Queue` finish their transfers. Now, it remains to show that the number of these requests is bounded. According to the operation described at the beginning of this section, PISCOT only allows a maximum of one request from any core at any time in the `Service Queue`. As a result, Req_i cannot have more than $N-1$ requests ahead of it `Service Queue`, which guarantees it a bound on the time it can be serviced (Section V provides a detailed latency analysis to derive these bounds). (2)
For now, (2) clearly contradicts (1), which completes the proof. ■

V. Analytical Worst-Case Latency

We derive the WCL suffered by any single request to the cache hierarchy that is managed by PISCOT. In doing so, we will use Figure 5, where the GetM(A) from $C2$ is the request under analysis or rua. As previously explained, the system in Figure 5 has three cores. As the figure illustrates, upon the arrival of the rua at timestamp t, there is a pending request from $C0$ to the same cache line A, which is initially owned by $C1$ in the M state. Generally, from its arrival to the private cache controller buffer until it completely receives the requested data, a request suffers from three different latency components. Namely, it suffers from latency due to arbitration on the request bus, denoted as $ReqBusL$, latency due to arbitration on the response bus, denoted as $ResBusL$, and finally the latency needed to transfer its data from the memory denoted as $AccL$. The $AccL$ depends on the time required to access the shared memory and transfer one cache line to the requesting core's private cache. Now, we derive the worst-case latency of each of the other two components.

***Lemma** 7:* **Worst-Case Request-Bus Latency ($ReqBusL^{WC}$).** For a system with N cores, a request has to wait for a maximum of $ReqBusL^{WC}$ cycles as calculated in Equation 1 before it is granted access to the request bus, where S^{Req} is the TDM slot width of the request bus in cycles.

$$ReqBusL^{WC} = N \cdot S^{Req} \tag{1}$$

Proof: Recall that the request bus is managed using a TDM arbiter. In the worst case, the rua arrives such that its core has just missed its own slot. Since we have N cores and each core is allocated one TDM slot of width S^{Req}, the rua has to wait for $N \cdot S^{Req}$ cycles before its corresponding core gets another slot. ■
In Figure 5, $S^{Req} = 4$ cycles and $N = 3$; thus, the GetM(A) from $C2$ waits until $t + 12$ to gain access to the bus.

***Lemma** 8:* **Worst-Case Response-Bus Latency ($ResBusL^{WC}$).** For a system with N cores, a request has to wait for a maximum of $ResBusL^{WC}$ cycles from its arrival time to the `Service Queue` before it can start receiving its requested data. $ResBusL^{WC}$ is calculated by Equation 2, where S^{Res} is the time required to conduct one memory transfer on the response bus.

$$ResBusL^{WC} = (2 \cdot N - 1) \cdot S^{Res} \tag{2}$$

Proof: Recall that the `Response Bus` services requests that arrive to the `Service Queue` from the `Request Bus` in a FCFS fashion. In addition, PISCOT allows each core to have at most one request in the `Service Queue` at any given time. Accordingly, the rua waits in worst-case for a request from every other core to get serviced. Moreover, in worst-case, each request can require two memory transfers. This is because each request can be modified by another core and hence requires a write-back before the shared memory can send the updated data to the requesting core. Since we have $N-1$ other cores, this consumes a total of $(N-1) \cdot 2 \cdot S^{Res}$. Finally, the rua itself in worst-case requires a write-back before it can start transferring its own data, which consumes an additional S^{Res}. This leads to $ResBusL^{WC} = (N-1) \cdot 2 \cdot S^{Res} + S^{Res}$ or $(2 \cdot N - 1) \cdot S^{Res}$. ■
In Figure 5, where $S^{Res} = 50$ cycles, the GetM(A) from $C2$ incurs a $ResBusL$ from $t+8$ to $t+258$, which is 250 cycles.

***Lemma** 9:* **Total Request Worst-Case Latency ($TotL^{WC}$).** For a system with N cores, the maximum total latency that a request can encounter from its arrival time to its private cache

controller before it can start receiving its requested data can be calculated as:

$$TotL^{WC} = N \cdot (S^{Req} + 2S^{Res}) \quad (3)$$

Proof: Since $TotL^{WC} = ReqBusL^{WC} + RespBusL^{WC} + accL$, the proof directly follows from Lemmas 7 and 8, and the fact that $accL = S^{Res}$ per definition. ■

A. Direct Cache-to-Cache Communication

In this case, only one response slot is needed for any request as Lemma 10 proves. Therefore, the total request WCL for such architecture reduces to the value in Lemma 11.

Lemma 10: **Worst-Case Response-Bus Latency with Cache-to-Cache Support ($ResBusL^{WC}_{C2C}$).** For a system with N cores that supports direct communication among cores' private caches, the maximum latency a request can suffer from its arrival time to the global response queue before it can start receiving its requested data can be calculated as in Equation 4, where S^{Res} is the time required to conduct a memory transfer on the response bus.

$$ResBusL^{WC}_{C2C} = (N-1) \cdot S^{Res} \quad (4)$$

Proof: The proof directly follows from the proof of Lemma 8, with the exception that only one response slot is required per core instead of two as follows. For any request, there are three possibilities. 1) A core requests to read from or write to a cache line that is up-to-date at the shared memory. In this case, the shared memory transfers this line to the requesting core. 2) A core requests to write to a line that is modified by another core. Thus, the owner core has to send this line to the requesting core. Since the latter is going to update the line, the shared memory does not need to receive the line at the moment. 3) A core requests to read from a line that is modified by another core. In this case, the owner has to send this line to both the requesting core and the shared memory. However, since the architecture supports cache-to-cache communication, the data can be sent to both at the same slot. This proves that under all these possibilities, only one response slot is needed instead of two compared to Lemma 8. In conclusion, the $ResBusL^{WC}_{C2C} = (N-1) \cdot S^{Res}$. ■

Lemma 11: **Total Request Worst-Case Latency with Cache-to-Cache Support ($TotL^{WC}_{C2C}$).** For a system with N cores that supports direct communication among cores' private caches, the maximum total latency that a request can encounter from its arrival time to its private cache controller before it can start receiving its requested data can be calculated as:

$$TotL^{WC} = N \cdot (S^{Req} + S^{Res}) \quad (5)$$

Proof: The proof directly follows from summing the latency components in Lemmas 7 and 10, and the $AccL$. ■

B. Total Task's Worst-Case Memory Latency

The latencies derived so far are concerned with a single memory request. However, to derive the total task's WCET, the total memory latency, $WCML$, has to be obtained and then added to the worst-case computation time, $WCCT$, such that:

$$WCET = WCCT + WCML \quad (6)$$

Let WCL_{Req} to be the per-request WCL to differentiate it from the total $WCML$, where WCL_{Req} is either the $TotL^{WC}$ in Lemma 9 if no cache-to-cache is supported, or the $TotL^{WC}_{C2C}$ in Lemma 11 otherwise. We now show different approaches to utilize this WCL_{Req} to derive $WCML$.

1) Using total number of requests: The first approach directly obtains $WCML$ through Equation 7, where $NReq$ is the worst-case total number of issued memory requests by the task. $NReq$ can be obtained by statically analyzing the task in isolation [2].

$$WCML = NReq \times WCL_{Req} \quad (7)$$

2) Distinction between private and shared data: Although the bound provided in Equation 7 is safe, it is rather pessimistic. This is because it assumes that all requests are misses, while in reality some of the requests will hit in the private caches and thus suffer a much less latency than WCL_{Req}. One challenge in data-sharing systems is that whether a task access to shared data hits or misses in the private cache depends on the access pattern of competing tasks, entailing that no reasoning can be made about whether this access hits or misses in the private cache by statically analyzing the task in isolation. Even worse, since shared cache lines can conflict with private lines in the core's private cache and hence evict each other, no analysis can be applied to access to private data as well. In this case, Equation 7 applies. In contrast, if private and shared data are isolated from each other; for instance, by mapping them to different cache sets, tighter memory latency bounds can be obtained for requests to the private data. Assuming this isolation, a task's hit ratio to the private data obtained from analyzing the task in isolation still holds upon interference from co-running other tasks. As a result, in such system, we can obtain the $WCML$ as in Equation 8, where $NReq^{priv}$ is the number of requests to private data, among them $NReq^{priv}_{hit}$ are hits in the private cache, and $NReq^{priv}_{miss}$ are misses. L_{hit} is the hit latency of the private cache and $NReq^{shrd}$ is the number of requests to shared data. Since $L_{hit} << WCL_{req}$ (L_{hit} is one or two cycles in modern architectures), the $WCML$ bound in Equation 8 is generally tighter than that of Equation 7. The actual values depend on the ratio of requests to private and shared data, and hence, is application dependent.

$$\begin{aligned} WCML =& NReq^{priv}_{hit} \times L_{hit} \\ &+ (NReq^{priv}_{miss} + NReq^{shrd}) \times WCL_{Req} \end{aligned} \quad (8)$$

C. Replacement of Dirty Cache Lines

The analysis in Lemmas 7–11 assumes that when a request misses in the private cache, it is sent directly to the bus arbiter to fetch the requested data. However, it is possible that the requested cache line is mapped to an entry that already has a valid data of another cache line. This is called a cache conflict.

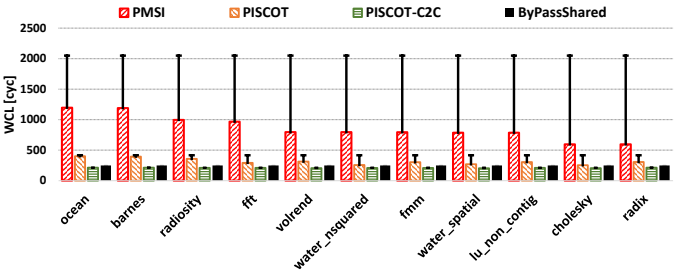


Fig. 6: Per-request WCL for SPLASH-3 suite.

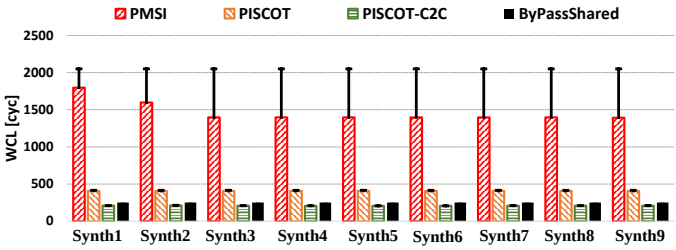


Fig. 7: Per-request WCL for the synthetic workloads.

In this case, the previous cache line is to be evicted from the private cache and the requested cache line is to be fetched to the same entry. If the evicted cache line has modified data, it has to be written first to the shared memory; otherwise, this data is going to be lost. This adds an extra latency of one memory transfer (or S^{Res}) for each miss request in the worst case. In other words, this adds $N \times S^{Res}$ to the latencies in Lemmas 9 and 11. However, assuming that every request is going to an eviction to a modified line is overly pessimistic and a tighter bound can be obtained as follows.

1) Total number of writes: Since the additional latency component is caused only upon evicting a dirty cache line, the total number of these replacements is bounded by the total number of write requests of the task, $WReq$. Accordingly, the effect of the replacement is better to be considered at the task level by updating Equation 7 to:

$$WCML = NReq \times WCL_{Req} + WReq \times S^{Res} \quad (9)$$

2) Distinction between private and shared data: Moreover, if the isolation between private and shared data discussed in Section V-B2 is adopted, the delay effects of replacement can be further reduced. This is because the number of replacements happening withing private data can also be obtained from analyzing the task using existing static analysis tools. Therefore, integrating the effect of replacements in Equation 8 leads to the $WCML$ in Equation 10, where $NRepl^{priv}$ is the worst-case number of dirty cache line replacements within private data, $WReq^{shrd}$ is the worst-case number of write requests to shared data.

$$WCML = NReq_{hit}^{priv} \times L_{hit} + NRepl^{priv} \times S^{Res} + (NReq_{miss}^{priv} + NReq^{shrd}) \times WCL_{Req} + WReq^{shrd} \times S^{Res} \quad (10)$$

VI. Evaluation

We develop an open-source simulation framework [1] to evaluate the performance of PISCOT and compare it with state-of-the-art solutions. The simulation environment consists of a multi-core system with configurable number of cores and cache organization. The simulator parameters are chosen to emulate the behavior of quad-core system running at 2.5 GHz with out-of-order pipelines, 8 kB direct-mapped L1 per-core private cache, and a 4 MB 8-ways set-associative L2 shared cache across all cores. Both L1 and LLC have a cache line size of 64 bytes. Furthermore, each core and LLC/shared memory are embedded with cache controller units which implement the MSI coherence state machine as described in section II. The collection of these coherence controllers are connected to the memory bus using bi-directional FIFO queues which are used for buffering incoming and outgoing messages generated by the coherence protocol. For PISCOT, the request and the response buses are split and operate independently. The former uses work-conserving TDM arbitration amongst cores with a slot width of 4 cycles (S^{Req} = 4 cycles), while the latter services the responses in FCFS fashion assuming the access latency to the LLC is fixed equals to 50 cycles (S^{Res} = 50 cycles). Accesses that hit in the L1 consume one cycle. We use a perfect LLC cache similar to existing works [18], [21] to avoid extra delays from accessing off-chip DRAM to measure only coherence and memory bus latencies. The DRAM access overheads can be computed using other approaches such as [9], [10], [40], and they are additive [41] to the latencies derived in this work.

The address translation between virtual CPU address and physical memory address is disabled such that all memory addresses generated by the cores are physical memory addresses. The maximum number of pending requests ($N_{Pending}$) a core can issue is set to 4 requests. This allows the core to issue multiple memory requests in parallel. The private cache controller has to track all pending requests issued on the bus and stall the core pipeline if it reaches to the maximum $N_{Pending}$. In addition, the controller needs to ensure that there is no more than one coherent message issued on the bus or in its local buffer in case of multiple cache misses occur on the same cache line. ***Benchmarks.*** We use the SPLASH-3 [42] benchmark suite since it is a representative of multi-threaded applications with shared data. In addition, we craft 9 synthetic workloads to stress the behavior of the evaluated approaches. All the synthetic workload resemble the maximum data sharing among cores (all lines are shared). They only differ in their memory intensity and the read/write ratio.

A. Per-Request Worst-Case Latency

Figures 6 and 7 depict the WCL for any request to the cache hierarchy for both SPLASH-3 benchmarks and the synthetic workloads, respectively. The figures show both the analytical WCL bounds (T bars) and the observed (experimental) WCL

[1] https://gitlab.com/FanosLab/piscot-and-msi-split-bus

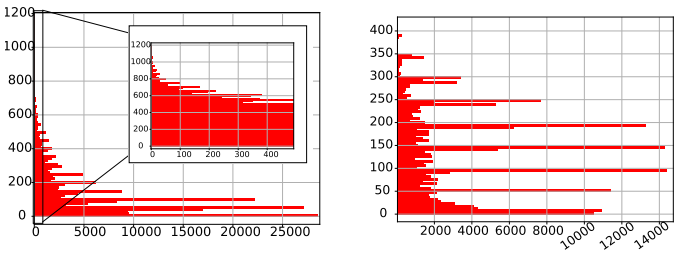
(colored solid bars). We compare the WCL of the two PISCOT schemes (where PISCOT-C2C is the one supporting cache-to-cache communication) with PMSI and not caching the shared data (BypassShared) approaches in [5], [43]. From this experiment, we make the following observations. 1) For both PISCOT and PISCOT-C2C, all the observed WC latencies are always within the analytical worst-case latency bounds. 2) PISCOT shows up to $4.9\times$ improvement in the analytical WCL compared to PMSI. The analytical WCL of PMSI is 2050 cycles compared to 416 and 216 cycles in PISCOT and PISCOT-C2C, respectively. 3) Compared to PMSI, the observed WCLs in PISCOT and PISCOT-C2C achieve up to $2.74\times$ and $4\times$ tighter bounds on average across benchmarks, respectively.

4) PMSI incurs a large gap between experimental and analytical WCLs. In the SPLASH-3 benchmarks (Figure 6), this gap ranges from 70% (barnes and ocean) and reaches up to $3.4\times$ (cholesky and radix). This is because PMSI's analytical WCL assumes a pathological worst-case scenario that is hard to construct in real applications. On the other hand, PISCOT achieves a tighter bound for the derived WCL. PISCOT achieves this tightness by enforcing FCFS arbitration policy on the response bus.

To further investigate the behavior of PISCOT and conventional split-transaction MSI, Figure 8 plots the observed latencies for requests for one of the BMs (Ocean) (others show similar behavior) for both solutions. As the figure illustrates, for MSI 8a, it shows a huge latency variability. Although most of the requests finish relatively fast, there are requests that their latencies reach up to 1200 cycles. On the other hand, PISCOT encounters less variability (Requests are suffering between $0-400$ cycles and all latencies under PISCOT operation are lower than its corresponding analytical bounds, which confirms the predictability of PISCOT.

B. Total Worst Case Latency

In this experiment we are interested in calculating the total memory WCL suffered by the total number of memory requests generated by a core during a period of time t. Figure 9 shows both the analytical bound for the total WCL derived by Equation 7 (T bars) and the observed total latencies (colored solid bars). Furthermore, the observed one is decomposed to its sub-components: a) the request bus arbitration latency, b) the response bus memory transfer latency, c) the hit latency in the core's private cache, and d) the write-backs latency due to replacement. From Figure 9, we conclude the following observations. 1) The response bus latency component dominates the total WCL for all applications. For instance, the total observed response latency reach up to $8\times$ (barnes and volrend) and $4.3\times$ on average larger than the replacement latency. This emphasises the conclusion we made in Section V-C that the effect of the eviction delays should be considered at the task-level and not the request-level. 2) Since SPLASH-3 applications exhibit a reduced ratio of writes compared to reads, they do not stress the difference between PISCOT and PISCOT-C2C in the observed response bus latency. Therefore, to further show this effect, we execute synthetic experiments



(a) MSI (b) PISCOT

Fig. 8: Request latency histogram for the Ocean benchmark with the no cache-to-cache transfer architecture. y-axis is the latency in cycles and x-axis is the number of requests encounter this latency.

using the synthetic benchmarks that are used to generate WCL in Figure 7 except that we change the percentage of the CPU memory write request to 50% of total memory requests. The results show that PISCOT-C2C achieves up to $1.74\times$ ($1.56\times$ on average) higher bandwidth compared to PISCOT.

C. Average-Case Performance

Figure 10 shows the slowdown of PISCOT and PMSI compared to the conventional MSI with split-transaction FCFS bus. PMSI's slowdown is $2\times$ on average (and up to $4.3\times$) across all benchmarks. This is due to the coupling of coherence and data transfer on the same TDM bus as explained in Section III in addition to the enforced protocol changes. Authors of [18] compared PMSI with an MSI+conventional TDM arbiter, for which they reported that PMSI showed only a 45% slowdown. Recall here we consider MSI+split-transaction bus. These results combined emphasise our observation that the split-bus architecture can significantly increase performance compared to the traditionally considered bus architectures by the real-time community. On the other hand, Figure 10 shows that PISCOT achieves comparable results with slowdown in the range of $1\% - 4\%$. This is clearly a negligible cost for achieving timing predictability with tight latency bounds.

We also observe that PISCOT improves the bandwidth utilization for the SPLASH-3 benchmarks by $12\times$ compared to PMSI (results are not shown due to space limitation). This significant improvement is because PMSI adopting the traditional TDM, which wastes many bus slots in only issuing coherence requests as we detailed in Section III. On the other hand, PISCOT maximizes bus utilization by splitting the coherence and response into two buses with two different slot widths and arbitration.

VII. Related Work

Towards adopting multi-core platforms in real-time embedded systems, several proposals are introduced to predictably manage shared hardware components among cores [1]–[9]. Among these, two lines of work are closely related to this paper, memory bus arbitration, and cache coherence.

Predictable Bus Arbitration. The memory bus in a multi-core platform is one of the main sources of interference, which

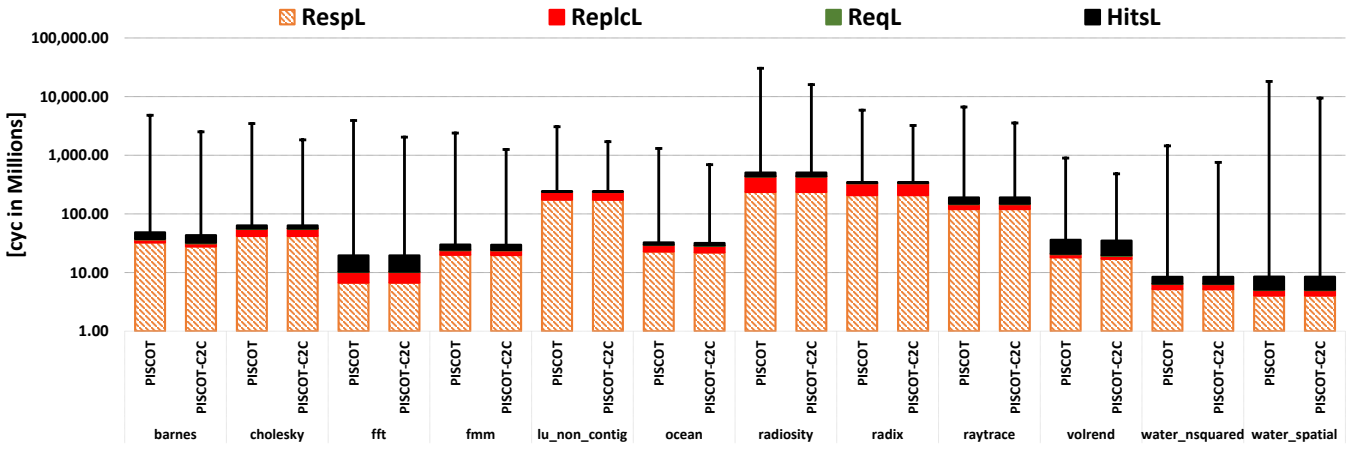


Fig. 9: Total observed and analytical memory latency of Splash-3 benchmarks. Values in y-axis are in log scale.

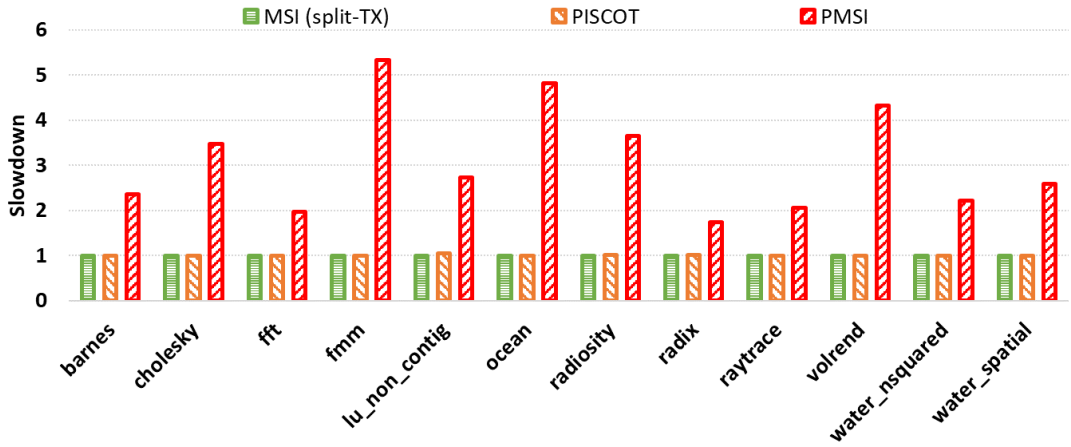


Fig. 10: Execution time slowdown compared to conventional MSI protocol with FCFS split-transaction bus.

was found to solely increase the Worst-Case Execution Time (WCET) by up to 44% in a quad-core system [44]. to address this challenge, researchers proposed predictable bus arbitration schemes. This includes: Time Division Multiplexing (TDM) [1], [3], [37], Round Robin (RR) [38], Harmonic RR (HRR) [39], and weighted [2] arbitration schemes. Unlike all these works, which focus only on timing interference assuming that tasks do not share data, PISCOT is a coherent bus that takes into account the cache coherence traffic and proposes a split-transaction architecture, where coherence requests and their responses are decoupled and arbitrated separately to increase system performance, while offering predictability.

Predictable Cache Coherence. There are multiple recent efforts to enable predictable sharing of data among real-time tasks through cache coherence [17]–[25]. PISCOT differentiates itself from these works by enabling predictable cache coherence through its bus arbitration architecture without requiring any changes to the coherence protocol (such as in [18]–[21]), the operating system's scheduler such as in [17], or the legacy software. The work in [23] focuses on formal modelling of cache coherence interference effects using an abstract model of existing commodity bus architectures. However, commercial architectures are not designed in the first place to be predictable, and thus, provide only very pessimistic bounds if any. PISCOT, in contrast, is a predictable split-transaction coherent bus architecture that significantly reduces WCLs, while maintaining high average performance.

VIII. Conclusion

PISCOT is a predictable and coherent bus architecture that provides significantly tighter bounds than existing predictable coherence protocols (4× tighter memory latency bounds for a quad-core system) with a performance near to that achieved by conventional high-performance arbiters (less than 4% overhead), which is 5× (2.8× on average) better performance compared to the state-of-the-art predictable coherence solutions. PISCOT achieves this by decoupling the data responses from their coherence requests utilizing a split-transaction predictable bus arbiter. PISCOT can be realized without any modifications to the coherence protocol or cache controller.

REFERENCES

[1] B. Cilku, B. Frömel, and P. Puschner, "A dual-layer bus arbiter for mixed-criticality systems with hypervisors," in *IEEE International Conference on Industrial Informatics (INDIN)*, 2014.

[2] M. Hassan and H. Patel, "Criticality- and requirement-aware bus arbitration for multi-core mixed criticality systems," in *IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, 2016.

[3] F. Hebbache, M. Jan, F. Brandner, and L. Pautet, "Shedding the shackles of time-division multiplexing," in *IEEE Real-Time Systems Symposium (RTSS)*, 2018.

[4] G. Gracioli, A. Alhammad, R. Mancuso, A. A. Fröhlich, and R. Pellizzoni, "A survey on cache management mechanisms for real-time embedded systems," *ACM Comput. Surv.*, 2015.

[5] D. Hardy, T. Piquet, and I. Puaut, "Using bypass to tighten WCET estimates for multi-core processors with shared instruction caches," in *IEEE Real-Time Systems Symposium (RTSS)*, 2009.

[6] N. C. Kumar, S. Vyas, R. K. Cytron, C. D. Gill, J. Zambreno, and P. H. Jones, "Cache design for mixed criticality real-time systems," in *IEEE International Conference on Computer Design (ICCD)*, 2014.

[7] M. Schoeberl, W. Puffitsch, and B. Huber, "Towards time-predictable data caches for chip-multiprocessors," in *Springer International Workshop on Software Technolgies for Embedded and Ubiquitous Systems (IFIP)*, 2009.

[8] R. Mancuso, R. Dudko, E. Betti, M. Cesati, M. Caccamo, and R. Pellizzoni, "Real-time cache management framework for multi-core architectures," in *IEEE 19th Real-Time and Embedded Technology and Applications Symposium (RTAS)*, 2013.

[9] D. Guo, M. Hassan, R. Pellizzoni, and H. Patel, "A comparative study of predictable dram controllers," *ACM Transactions on Embedded Computing Systems (TECS)*, 2018.

[10] M. Hassan and R. Pellizzoni, "Bounding dram interference in cots heterogeneous mpsocs for mixed criticality systems," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2018.

[11] M. Hassan, "On the off-chip memory latency of real-time systems: Is ddr dram really the best option?" in *IEEE Real-Time Systems Symposium (RTSS)*, 2018.

[12] A. Hamann, D. Dasari, S. Kramer, M. Pressler, and F. Wurst, "Communication centric design in complex automotive embedded systems," in *29th Euromicro Conference on Real-Time Systems (ECRTS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[13] M. Younis and M. Aboutabl, "Communication handling in integrated modular avionics," Oct. 3 2002, uS Patent App. 09/821,601.

[14] M. Becker, D. Dasari, B. Nicolic, B. Akesson, V. Nélis, and T. Nolte, "Contention-free execution of automotive applications on a clustered many-core platform," in *IEEE Euromicro Conference on Real-Time Systems (ECRTS)*, 2016.

[15] M. Chisholm, N. Kim, B. C. Ward, N. Otterness, J. H. Anderson, and F. D. Smith, "Reconciling the tension between hardware isolation and data sharing in mixed-criticality, multicore systems," in *IEEE Real-Time Systems Symposium (RTSS)*, 2016.

[16] N. Kim, M. Chisholm, N. Otterness, J. H. Anderson, and F. D. Smith, "Allowing shared libraries while supporting hardware isolation in multi-core real-time systems," in *IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, 2017.

[17] G. Gracioli and A. A. Fröhlich, "On the design and evaluation of a real-time operating system for cache-coherent multicore architectures," *ACM SIGOPS Oper. Syst. Rev.*, 2015.

[18] M. Hassan, A. M. Kaushik, and H. Patel, "Predictable cache coherence for multi-core real-time systems," in *IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, 2017.

[19] N. Sritharan, A. M. Kaushik, M. Hassan, and H. Patel, "Hourglass: Predictable time-based cache coherence protocol for dual-critical multi-core systems," 2017.

[20] ——, "Enabling predictable, simultaneous and coherent data sharing in mixed criticality systems," 2019.

[21] A. M. Kaushik, P. Tegegn, Z. Wu, and H. Patel, "Carp: A data communication mechanism for multi-core mixed-criticality systems," in *IEEE Real-Time Systems Symposium (RTSS)*, 2019.

[22] A. Bansal, J. Singh, Y. Hao, J.-Y. Wen, R. Mancuso, and M. Caccamo, "Cache where you want! reconciling predictability and coherent caching," *arXiv preprint arXiv:1909.05349*, 2019.

[23] N. Sensfelder, J. Brunel, and C. Pagetti, "Modeling cache coherence to expose interference," in *31st Euromicro Conference on Real-Time Systems (ECRTS 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[24] ——, "On how to identify cache coherence: Case of the nxp qoriq t4240," in *32nd Euromicro Conference on Real-Time Systems (ECRTS 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

[25] M. Hassan, "Discriminative coherence: Balancing performance and latency bounds in data-sharing multi-core real-time systems," in *Euromicro Conference on Real-Time Systems (ECRTS)*, 2020, pp. 1–22.

[26] M. M. MARTIN, M. D. HILL, and D. J. SORIN, "Why on-chip cache coherence is here to stay," *Communications of ACM*, 2012.

[27] D. J. Sorin, M. D. Hill, and D. A. Wood, "A primer on memory consistency and cache coherence," *Synthesis Lectures on Computer Architecture*, 2011.

[28] F. Pong and M. Dubois, "A new approach for the verification of cache coherence protocols," *IEEE Transactions on Parallel and Distributed Systems*, 1995.

[29] J. L. Hennessy and D. A. Patterson, *Computer architecture: a quantitative approach*. Elsevier, 2011.

[30] M. S. Khaira, "Fast first-come first served arbitration method," Nov. 12 1996, uS Patent 5,574,867.

[31] W. Bain Jr and S. Ahuja, "Performance analysis of high-speed digital buses for multiprocessing systems," in *Proceedings of the 8th annual symposium on Computer Architecture*, 1981, pp. 107–133.

[32] M. A. Fischer, "Fair arbitration technique for a split transaction bus in a multiprocessor computer system," Nov. 15 1988, uS Patent 4,785,394.

[33] A. Singhal, B. Liencres, J. Price, F. M. Cerauskis, D. Broniarczyk, G. Cheung, E. Hagersten, and N. Agarwal, "Implementing snooping on a split-transaction computer system bus," Nov. 2 1999, uS Patent 5,978,874.

[34] ARM, "ARM CoreLink CCI-550 Cache Coherent Interconnect, Technical Reference Manual," 2015. [Online]. Available: https://static.docs.arm.com/100282/0001/corelink_cci550_cache_coherent_interconnect_technical_reference_manual_100282_0001_01_en.pdf

[35] D. Ziakas, A. Baum, R. A. Maddox, and R. J. Safranek, "Intel® quickpath interconnect architectural features supporting scalable system architectures," in *IEEE Symposium on High Performance Interconnects*, 2010.

[36] F. Poletti, D. Bertozzi, L. Benini, and A. Bogliolo, "Performance analysis of arbitration policies for soc communication architectures," *Design Automation for Embedded Systems*, 2003.

[37] T. Kelter, H. Falk, P. Marwedel, S. Chattopadhyay, and A. Roychoudhury, "Bus-aware multicore wcet analysis through tdma offset bounds," in *Euromicro Conference on Real-Time Systems (ECRTS)*, 2011.

[38] M. Paolieri, E. Quiñones, F. J. Cazorla, G. Bernat, and M. Valero, "Hardware support for wcet analysis of hard real-time multicore systems," *ACM SIGARCH Computer Architecture News*, 2009.

[39] M.-K. Yoon, J.-E. Kim, and L. Sha, "Optimizing tunable wcet with shared resource allocation and arbitration in hard real-time multicore systems," in *IEEE Real-Time Systems Symposium (RTSS)*, 2011.

[40] M. Hassan and R. Pellizzoni, "Analysis of memory-contention in heterogeneous cots mpsocs," in *Euromicro Conference on Real-Time Systems (ECRTS)*, 2020.

[41] H. Yun, R. Pellizzoni, and P. K. Valsan, "Parallelism-aware memory interference delay analysis for cots multicore systems," in *Euromicro Conference on Real-Time Systems (ECRTS)*, 2015.

[42] C. Sakalis, C. Leonardsson, S. Kaxiras, and A. Ros, "Splash-3: A properly synchronized benchmark suite for contemporary research," in *IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2016.

[43] B. Lesage, D. Hardy, and I. Puaut, "Shared Data Caches Conflicts Reduction for WCET Computation in Multi-Core Architectures." in *International Conference on Real-Time and Network Systems*, 2010.

[44] R. Pellizzoni, B. D. Bui, M. Caccamo, and L. Sha, "Coscheduling of cpu and i/o transactions in cots-based embedded systems," in *IEEE Real-Time Systems Symposium (RTSS)*, 2008.