Cache Coherency in ARMv8-A for Cross-Architectural DSM Systems

Zhengyi Chen

4th Year Project Report
Computer Science
School of Informatics
University of Edinburgh

2024

Abstract

To be done…

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Zhengyi Chen*)

Acknowledgements

Jordanian River to the Mediterranean Sea, maybe…

Contents

Chapter 1

Introduction

Though large-scale cluster systems remain the dominant solution for request and data-level parallelism [23], there have been a resurgence towards applying HPC techniques (e.g., DSM) for more efficient heterogeneous computation with tighter-coupled heterogeneous nodes providing (hardware) acceleration for one another [9, 38, 30]. Orthogonally, within the scope of one motherboard, *heterogeneous memory management (HMM)* enables the use of OS-controlled, unified memory view across both main memory and device memory [22], all while using the same libc function calls as one would with SMP programming, the underlying complexities of memory ownership and data placement automatically managed by the OS kernel. However, while HMM promises a distributed shared memory approach towards exposing CPU and peripheral memory, applications (drivers and front-ends) that exploit HMM to provide ergonomic programming models remain fragmented and narrowly-focused. Existing efforts in exploiting HMM in Linux predominantly focus on exposing global address space abstraction to GPU memory – a largely non-coordinated effort surrounding both *in-tree* and proprietary code [14, 1]. Limited effort have been done on incorporating HMM into other variants of accelerators in various system topologies.

Orthogonally, allocation of hardware accelerator resources in a cluster computing environment becomes difficult when the required hardware accelerator resources of one workload cannot be easily determined and/or isolated as a "stage" of computation. Within a cluster system there may exist a large amount of general-purpose worker nodes and limited amount of hardware-accelerated nodes. Further, it is possible that every workload performed on this cluster asks for hardware acceleration from time to time, but never for a relatively long time. Many job scheduling mechanisms within a cluster *move data near computation* by migrating the entire job/container between general-purpose and accelerator nodes [50, 44]. This way of migration naturally incurs large overhead – accelerator nodes which strictly perform computation on data in memory without ever needing to touch the container's filesystem should not have to install the entire filesystem locally, for starters. Moreover, must *all* computations be performed near data? *Adrias*[40], for example, shows that RDMA over fast network interfaces (25 Gbps $\times$ 8), when compared to node-local setups, result in negligible impact on tail latencies but high impact on throughput when bandwidth is maximized.

This thesis paper builds upon an ongoing research effort in implementing a tightly coupled cluster where HMM abstractions allow for transparent RDMA access from accelerator nodes to local data and migration of data near computation, leveraging different consistency model and coherency protocols to amortize the communication cost for shared data. More specifically, this thesis explores the following:

- The effect of cache coherency maintenance, specifically OS-initiated, on RDMA programs.
- Discussion of memory models and coherence protocol designs for a single-writer, multi-reader RDMA-based DSM system.

The rest of the chapter is structured as follows:

- We identify and discuss notable developments in software-implemented DSM systems, and thus identify key features of contemporary advancements in DSM techniques that differentiate them from their predecessors.
- We identify alternative (shared memory) programming paradigms and compare them with DSM, which sought to provide transparent shared address space among participating nodes.
- We give an overview of coherency protocol and consistency models for multi-sharer DSM systems.
- We provide a primer to cache coherency in ARM64 systems, which *do not* guarantee cache-coherent DMA, as opposed to x86 systems [55].

1.1 Experiences from Software DSM

A majority of contributions to software DSM systems come from the 1990s [5, 11, 27, 26]. These developments follow from the success of the Stanford DASH project in the late 1980s – a hardware distributed shared memory (specifically NUMA) implementation of a multiprocessor that first proposed the *directory-based protocol* for cache coherence, which stores the ownership information of cache lines to reduce unnecessary communication that prevented previous multiprocessors from scaling out [33].

While developments in hardware DSM materialized into a universal approach to cache-coherence in contemporary many-core processors (e.g., *Ampere Altra*[2]), software DSMs in clustered computing languished in favor of loosely-coupled nodes performing data-parallel computation, communicating via message-passing. Bandwidth limitations with the network interfaces of the late 1990s was insufficient to support the high traffic incurred by DSM and its programming model [57, 36].

New developments in network interfaces provides much improved bandwidth and latency compared to ethernet in the 1990s. RDMA-capable NICs have been shown to improve the training efficiency sixfold compared to distributed *TensorFlow* via RPC, scaling positively over non-distributed training [28]. Similar results have been observed for *APACHE Spark* [37] and *SMBDirect* [34]. Consequently, there have been a resurgence of interest in software DSM systems and programming models [43, 10].

1.1.1 Munin: Multi-Consistency Protocol

Munin[11] is one of the older developments in software DSM systems. The authors of Munin identify that *false-sharing*, occurring due to multiple processors writing to different offsets of the same page triggering invalidations, is strongly detrimental to the performance of shared-memory systems. To combat this, Munin exposes annotations as part of its programming model to facilitate multiple consistency protocols on top of release consistency. An immutable shared memory object across readers, for example, can be safely copied without concern for coherence between processors. On the other hand, the *write-shared* annotation explicates that a memory object is written by multiple processors without synchronization – i.e., the programmer guarantees that only false-sharing occurs within this granularity. Annotations such as these explicitly disables subsets of consistency procedures to reduce communication in the network fabric, thereby improving the performance of the DSM system.

Perhaps most importantly, experiences from Munin show that *restricting the flexibility of programming model can lead to more performant coherence models*, as exhibited by the now-foundational *Resilient Distributed Database* paper [59] which powered many now-popular scalable data processing frameworks such as *Hadoop MapReduce* [3] and *APACHE Spark* [4]. "To achieve fault tolerance efficiently, RDDs provide a restricted form of shared memory [based on]. . . transformations rather than. . . updates to shared state" [59]. This allows for the use of transformation logs to cheaply synchronize states between unshared address spaces – a much desired property for highly scalable, loosely-coupled clustered systems.

1.1.2 Treadmarks: Multi-Writer Protocol

Treadmarks[5] is a software DSM system developed in 1996, which featured an intricate *interval*-based multi-writer protocol that allows multiple nodes to write to the same page without false-sharing. The system follows a release-consistent memory model, which requires the use of either locks (via `acquire`, `release`) or barriers (via `barrier`) to synchronize. Each *interval* represents a time period in-between page creation, `release` to another processor, or a `barrier`; they also each correspond to a *write notice*, which are used for page invalidation. Each `acquire` message is sent to the statically-assigned lock-manager node, which forwards the message to the last releaser. The last releaser computes the outstanding write notices and piggy-backs them back for the acquirer to invalidate its own cached page entry, thus signifying entry into the critical section. Consistency information, including write notices, intervals, and page diffs, are routinely garbage-collected which forces cached pages in each node to become validated.

Compared to *Treadmarks*, the system described in this paper uses a single-writer protocol, thus eliminating the concept of "intervals" – with regards to synchronization, each page can be either in-sync (in which case they can be safely shared) or out-of-sync (in which case they must be invalidated/updated). This comes with the following advantage:

- Less metadata for consistency-keeping.
- More adherent to the CPU-accelerator dichotomy model.

- Much simpler coherence protocol, which reduces communication cost.

In view of the (still) disparate throughput and latency differences between local and remote memory access [10], the simpler coherence protocol of single-writer protocol should provide better performance on the critical paths of remote memory access.

1.1.3 Hotpot: Single-Writer & Data Replication

Newer works such as *Hotpot*[52] apply distributed shared memory techniques on persistent memory to provide "transparent memory accesses, data persistence, data reliability, and high availability". Leveraging on persistent memory devices allow DSM applications to bypass checkpoints to block device storage [52], ensuring both distributed cache coherence and data reliability at the same time [52].

We specifically discuss the single-writer portion of its coherence protocol. The data reliability guarantees proposed by the *Hotpot* system requires each shared page to be replicated to some *degree of replication*. Nodes who always store latest replication of shared pages are referred to as "owner nodes", which arbitrate other nodes to store more replications in order to reach the degree of replication quota. At acquisition time, the acquiring node asks the access-management node for single-writer access to shared page, who grants it if no other critical section exists, alongside list of current owner nodes. At release time, the releaser first commits its changes to all owner nodes which, in turn, commits its received changes across lesser sharers to achieve the required degree of replication. These two operations are all acknowledged back in reverse order. Once all acknowledgements are received from owner nodes by commit node, the releaser tells them to delete their commit logs and, finally, tells the manager node to exit critical section.

The required degree of replication and logged commit transaction until explicit deletion facilitate crash recovery at the expense of worse performance over release-time I/O. While the study of crash recovery with respect to shared memory systems is out of the scope of this thesis, this paper provides a good framework for a **correct** coherence protocol for a single-writer, multiple-reader shared memory system, particularly when the protocol needs to cater for a great variety of nodes each with their own memory preferences (e.g., write-update vs. write-invalidate, prefetching, etc.).

1.1.4 MENPS: A Return to DSM

MENPS[19] leverages new RDMA-capable interconnects as a proof-of-concept that DSM systems and programming models can be as efficient as *partitioned global address space* (PGAS) using today's network interfaces. It builds upon *TreadMark*'s [5] coherence protocol and crucially alters it to a *floating home-based* protocol, based on the insight that diff-transfers across the network is comparatively costly compared to RDMA intrinsics – which implies preference towards local diff-merging. The home node then acts as the data supplier for every shared page within the system.

Compared to PGAS frameworks (e.g., MPI), experimentation over a subset of *NAS Parallel Benchmarks* shows that MENPS can obtain comparable speedup in some of

the computation tasks, while achieving much better productivity due to DSM's support for transparent caching, etc. [19]. These results back up their claim that DSM systems are at least as viable as traditional PGAS/message-passing frameworks for scientific computing, also corroborated by the resurgence of DSM studies later on[40].

1.2 PGAS and Message Passing

While the feasibility of transparent DSM systems over multiple machines on the network has been made apparent since the 1980s, predominant approaches to "scaling-out" programs over the network relies on the message-passing approach [54]. The reasons are twofold:

1. Programmers would rather resort to more intricate, more predictable approaches to scaling-out programs over the network [54]. This implies manual/controlled data sharding over nodes, separation of compute and communication "stages" of computation, etc., which benefit performance analysis and engineering.
2. Enterprise applications value throughput and uptime of relatively computationally inexpensive tasks/resources [23], which requires easy scalability of tried-and-true, latency-inexpensive applications. Studies in transparent DSM systems mostly require exotic, specifically-written programs to exploit global address space, which is fundamentally at odds in terms of reusability and flexibility required.

1.2.1 PGAS

Partitioned Global Address Space (PGAS) is a parallel programming model that (1) exposes a global address space to all machines within a network and (2) explicates distinction between local and remote memory [16]. Oftentimes, message-passing frameworks, for example *OpenMPI*, *OpenFabrics*, and *UCX*, are used as backends to provide the PGAS model over various network interfaces/platforms (e.g., Ethernet and Infiniband)[53, 48].

Notably, implementation of a *global* address space across machines on top of machines already equipped with their own *local* address space (e.g., cluster nodes running commercial Linux) necessitates a global addressing mechanism for shared/shared data objects. DART[61], for example, utilizes a 128-bit "global pointer" to encode global memory object/segment ID and access flags in the upper 64 bits and virtual addresses in the lower 64 bits for each (slice of) memory object allocated within the PGAS model. A *non-collective* PGAS object is allocated entirely local to the allocating node's memory, but registered globally. Consequently, a single global pointer is recorded in the runtime with corresponding permission flags for the context of some user-defined group of associated nodes. Comparatively, a *collective* PGAS object is allocated such that a partition of the object (i.e., a sub-array of the repr) is stored in each of the associated node – for a k-partitioned object, k global pointers are recorded in the runtime each pointing to the same object, with different offsets and (intuitively) independently-chosen virtual addresses. Note that this design naturally requires virtual addresses within each node to be *pinned* – the allocated object cannot be re-addressed to a different virtual

address, thus preventing the global pointer that records the local virtual address from becoming spontaneously invalidated.

Similar schemes can be observed in other PGAS backends/runtimes, albeit they may opt to use a map-like data structure for addressing instead. In general, despite both PGAS and DSM systems provide memory management over remote nodes, PGAS frameworks provide no transparent caching and transfer of remote memory objects accessed by local nodes. The programmer is still expected to handle data/thread movement manually when working with shared memory over network to maximize their performance metrics of interest.

1.2.2 Message Passing

Message Passing remains the predominant programming model for parallelism between loosely-coupled nodes within a computer system, much as it is ubiquitous in supporting all levels of abstraction within any concurrent components of a computer system. Specific to cluster computing systems is the message-passing programming model, where parallel programs (or instances of the same parallel program) on different nodes within the system communicate via exchanging messages over network between these nodes. Such models exchange programming model productivity for more fine-grained control over the messages passed, as well as more explicit separation between communication and computation stages within a programming subproblem.

Commonly, message-passing backends function as *middlewares* – communication runtimes – to aid distributed software development [54]. Such a message-passing backend expose facilities for inter-application communication to frontend developers while transparently providing security, accounting, and fault-tolerance, much like how an operating system may provide resource management, scheduling, and security to traditional applications [54]. This is the case for implementing the PGAS programming model, which mostly rely on common message-passing backends to facilitate orchestrated data manipulation across distributed nodes. Likewise, message-passing backends, including RDMA API, form the backbone of many research-oriented DSM systems [19, 25, 10, 29].

Message-passing between network-connected nodes may be *two-sided* or *one-sided*. The former models an intuitive workflow to sending and receiving datagrams over the network – the sender initiates a transfer; the receiver copies a received packet from the network card into a kernel buffer; the receiver's kernel filters the packet and (optionally) [49] copies the internal message into the message-passing runtime/middleware's address space; the receiver's middleware inspects the copied message and performs some procedures accordingly, likely also involving copying slices of message data to some registered distributed shared memory buffer for the distributed application to access. Despite it being a highly intuitive model of data manipulation over the network, this poses a fundamental performance issue: because the process requires the receiver's kernel AND userspace to exert CPU-time, upon reception of each message, the receiver node needs to proactively exert CPU-time to move the received data from bytes read from NIC devices to userspace. Because this happens concurrently with other kernel and userspace routines in a concurrent system, a preemptable kernel may incur significant

latency if the kernel routine for packet filtering is pre-empted by another kernel routine, userspace, or IRQs.

Comparatively, a "one-sided" message-passing scheme, for example RDMA, allows the network interface card to bypass in-kernel packet filters and perform DMA on registered memory regions. The NIC can hence notify the CPU via interrupts, thus allowing the kernel and the userspace programs to perform callbacks at reception time with reduced latency. Because of this advantage, many recent studies attempt to leverage RDMA APIs for improved distributed data workloads and creating DSM middlewares [37, 28, 19, 25, 10, 29].

1.3 Consistency Model and Cache Coherence

Consistency model specifies a contract on allowed behaviors of multi-processing programs with regards to a shared memory [42]. One obvious conflict, which consistency models aim to resolve, lies within the interaction between processor-native programs and multi-processors, all of whom needs to operate on a shared memory with heterogeneous cache topologies. Here, a well-defined consistency model aims to resolve the conflict on an architectural scope. Beyond consistency models for bare-metal systems, programming languages [8, 7, 39, 45] and paradigms [5, 25, 10] define consistency models for parallel access to shared memory on top of program order guarantees to explicate program behavior under shared memory parallel programming across underlying implementations.

Related to the definition of a consistency model is the coherence problem, which arises whenever multiple actors have access to multiple copies of some datum, which needs to be synchronized across multiple actors with regards to write-accesses [42]. While less relevant to programming language design, coherence must be maintained via a coherence protocol [42] in systems of both microarchitectural and network scales. For DSM systems, the design of a correct and performant coherence protocol is of especially high priority and is a major part of many studies in DSM systems throughout history [11, 5, 47, 19, 15].

1.3.1 Consistency Model in DSM

Distributed shared memory systems with node-local caching naturally implies the existence of the consistency problem with regards to contending read/write accesses. Indeed, a significant subset of DSM studies explicitly characterize themselves as adhering to one of the well-known consistency models to better understand system behavior and to provide optimizations in coherence protocols [5, 26, 11, 19, 56, 10, 31], each adhering to a different consistency model to balance between communication costs and ease of programming.

In particular, we note that DSM studies tend to conform to either release consistency [5, 19, 11] or weaker [26], or sequential consistency [12, 56, 31, 17], with few works [10] pertaining to moderately constrained consistency models in-between. While older works, as well as works which center performance of their proposed DSM systems over

existing approaches [19, 10], favor release consistency due to its performance benefits (e.g., in terms of coherence costs [19]), newer works tend to adopt stricter consistency models, sometimes due to improved productivity offered to programmers [31].

	Sequential	TSO	PSO	Release	Acquire	Scope
Home; Invalidate	[31, 17, 60]			[52, 19]	[24]	[26]
Home; Update						
Float; Invalidate				[19]		
Float; Update						
Directory; Inval.	[56]					
Directory; Update						
Dist. Dir.; Inval.	[12]		[10]	[11]	[11, 5]	
Dist. Dir.; Update				[11]		

Table 1.1: Coherence Protocol vs. Consistency Model in Selected Disaggregated Memory Studies. "Float" short for "floating home". Studies selected for clearly described consistency model and coherence protocol.

We especially note the role of balancing productivity and performance in terms of selecting the ideal consistency model for a system. It is common knowledge that weaker consistency models are harder to program with, at the benefit of less (implied) coherence communications resulting in better throughput overall – provided that the programmer could guarantee correctness, a weaker consistency model allows for less invalidation of node-local cache entries, thereby allowing multiple nodes to compute in parallel on (likely) outdated local copy of data such that the result of the computation remains semantically correct with regards to the program. This point was made explicit in *Munin* [11], where (to reiterate) it introduces the concept of consistency "protocol parameters" to annotate shared memory access pattern, in order to reduce the amount of coherence communications necessary between nodes computing in distributed shared memory. For example, a DSM object (memory object accounted for by the DSM system) can be annotated with "delayed operations" to delay coherence operations beyond any write-access, or shared without "write" annotation to disable write-access over shared nodes, thereby disabling all coherence operations with regards to this DSM object. Via programmer annotation of DSM objects, the Munin DSM system explicates the effect of weaker consistency in relation to the amount of synchronization overhead necessary among shared memory nodes. To our knowledge, no other more recent DSM works have explored this interaction between consistency and coherence costs on DSM objects, though relatedly *Resilient Distributed Dataset (RDD)* [59] also highlights its performance and flexibility benefits in opting for an immutable data representation over disaggregated memory over network when compared to contemporary DSM approaches.

1.3.2 Coherence Protocol

Coherence protocols hence becomes the means over which DSM systems implement their consistency model guarantees. As table 1.1 shows, DSM studies tends to implement write-invalidated coherence under a *home-based* or *directory-based* protocol

framework, while a subset of DSM studies sought to reduce communication overheads and/or improve data persistence by offering write-update protocol extensions [11, 52].

1.3.2.1 Home-Based Protocols

Home-based protocols define each shared memory object with a corresponding "home" node, under the assumption that a many-node network would distribute home-node ownership of shared memory objects across all hosts [26]. On top of home-node ownership, each mutable shared memory object may be additionally cached by other nodes within the network, creating the coherence problem. To our knowledge, in addition to table 1.1, this protocol and its derivatives had been adopted by [20, 35, 26, 43, 52, 19].

We identify that home-based protocols are conceptually straightforward compared to directory-based protocols, centering communications over storage of global metadata (in this case ownership of each shared memory object). This leads to greater flexibility in implementing coherence protocols. A shared memory object at its creation may be made known globally via broadcast, or made known to only a subset of nodes (0 or more) via multicast. Likewise, metadata storage could be cached locally to each node and invalidated alongside object invalidation or fetched from a fixed node with respect to one object. This implementation flexibility is further taken advantage of in *Hotpot*[52], which refines the "home node" concept into *owner node* to provide replication and persistence, in addition to adopting a dynamic home protocol similar to that of [19].

1.3.2.2 Directory-Based Protocols

Directory-based protocols instead take a shared database approach by denoting each shared memory object with a globally shared entry describing ownership and sharing status. In its non-distributed form (e.g., [56]), a global, central directory is maintained for all nodes in network for ownership information: the directory hence becomes a bottleneck for imposing latency and bandwidth constraints on parallel processing systems. Comparatively, a distributed directory scheme may delegate responsibilities across all nodes in network mostly in accordance to sharded address space [25, 10]. Though theoretically sound, this scheme performs no dynamic load-balancing for commonly shared memory objects, which in the worst case would function exactly like a non-distributed directory coherence scheme. To our knowledge, in addition to table 1.1, this protocol and its derivatives had been adopted by [11, 5, 51, 18, 25].

1.3.3 DMA and Cache Coherence

The advent of high-speed RDMA-capable network interfaces introduce introduce opportunities for designing more performant DSM systems over RDMA (as established in 1.2.2). Orthogonally, RDMA-capable NICs on a fundamental level perform direct memory access over the main memory to achieve one-sided RDMA operations to reduce the effect of OS jittering on RDMA latencies. For modern computer systems with cached multiprocessors, this poses a potential cache coherence problem on a local level

– because RDMA operations happen concurrently with regards to memory accesses by CPUs, which stores copies of memory data in cache lines which may [32, 55] or may not [21, 13] be fully coherent by the DMA mechanism, any DMA operations performed by the RDMA NIC may be incoherent with the cached copy of the same data inside the CPU caches (as is the case for accelerators, etc.). This issue is of particular concern to the kernel development community, who needs to ensure that the behaviors of DMA operations remain identical across architectures regardless of support of cache-coherent DMA [13]. Likewise existing RDMA implementations which make heavy use of architecture-specific DMA memory allocation implementations, implementing RDMA-based DSM systems in kernel also requires careful use of kernel API functions that ensure cache coherency as necessary.

1.3.4 Cache Coherence in ARMv8-A

We specifically focus on the implementation of cache coherence in ARMv8-A. Unlike x86 which guarantees cache-coherent DMA [55, 13], the ARMv8-A architecture (and many other popular ISAs, for example *RISC-V*) *does not* guarantee cache-coherency of DMA operations across vendor implementations. ARMv8 defines a hierarchical model for coherency organization to support *heterogeneous* and *asymmetric* multi-processing systems [6].

Definition 1 (cluster)**.** A *cluster* defines a minimal cache-coherent region for Cortex-A53 and Cortex-A57 processors. Each cluster usually comprises of 1 or more core as well as a shared last-level cache.

Definition 2 (sharable domain)**.** A *sharable domain* defines a vendor-defined cache-coherent region. Sharable domains can be *inner* or *outer*, which limits the scope of broadcast coherence messages to *point-of-unification* and *point-of-coherence*, respectively.

Usually, the *inner* sharable domain defines the domain of all (closely-coupled) processors inside a heterogeneous multiprocessing system (see 5); while the *outer* sharable domain defines the largest memory-sharing domain for the system (e.g. inclusive of DMA bus).

Definition 3 (Point-of-Unification)**.** The *point-of-unification* (*PoU*) under ARMv8 defines a level of coherency such that all sharers inside the **inner** sharable domain see the same copy of data.

Consequently, *PoU* defines a point at which every core of a ARMv8-A processor sees the same (i.e., a *unified*) copy of a memory location regardless of accessing via instruction caches, data caches, or TLB.

Definition 4 (Point-of-Coherence)**.** The *point-of-coherence* (*PoC*) under ARMv8 defines a level of coherency such that all sharers inside the **outer** sharable domain see the same copy of data.

Consequently, *PoC* defines a point at which all *observers* (e.g., cores, DSPs, DMA engines) to memory will observe the same copy of a memory location.

1.3.4.1 Addendum: *Heterogeneous* & *Asymmetric* Multiprocessing

Using these definitions, a vendor could build *heterogeneous* and *asymmetric* multiprocessing systems as follows:

Definition 5 (Heterogeneous Multiprocessing)**.** A *heterogeneous multiprocessing* system incorporates ARMv8 processors of diverse microarchitectures that are fully coherent with one another, running the same system image.

Definition 6 (Asymmetric Multiprocessing)**.** A *asymmetric multiprocessing* system needs not contain fully coherent processors. For example, a system-on-a-chip may contain a non-coherent co-processor for secure computing purposes [6].

1.3.5 ARMv8-A Software Cache Coherence in Linux Kernel

Because of the lack of hardware guarantee on hardware DMA coherency (though such support exists [46]), programmers need to invoke architecture-specific cache-coherency instructions when porting DMA hardware support over a diverse range of ARMv8 microarchitectures, often encapsulated in problem-specific subroutines.

Notably, kernel (driver) programming warrants programmer attention to software-maintained coherency when userspace programmers downstream expect data-flow, interspersed between CPU and DMA operations, to follow program ordering and (driver vendor) specifications. One such example arises in the Linux kernel implementation of DMA memory management API [41][1]:

Definition 7 (DMA Mappings)**.** The Linux kernel DMA memory allocation API, imported via

```
1 #include <linux/dma-mapping.h>
```

defines two variants of DMA mappings:

- *Consistent* DMA mappings:

  They are guaranteed to be coherent in-between concurrent CPU/DMA accesses without explicit software flushing. [2]

- *Streaming* DMA mappings:

  They provide no guarantee to coherency in-between concurrent CPU/DMA accesses. Programmers need to manually apply coherency maintenance subroutines for synchronization.

Consistent DMA mappings could be trivially created via allocating non-cacheable memory, which guarantees *PoC* for all memory observers (though system-specific fastpaths exist).

[1] Based on Linux kernel v6.7.0.

[2] However, it does not preclude CPU store reordering, so memory barriers remain necessary in a multiprocessing context.

On the other hand, streaming DMA mappings require manual synchronization upon programmed CPU/DMA access. Take single-buffer synchronization on CPU after DMA access for example:

```
/* In kernel/dma/mapping.c */
void dma_sync_single_for_cpu(
    struct device *dev,          // kernel repr for DMA device
    dma_addr_t addr,             // DMA address
    size_t size,                 // Synchronization buffer size
    enum dma_data_direction dir, // Data-flow direction
) {
    /* Translate DMA address to physical address */
    phys_addr_t paddr = dma_to_phys(dev, addr);

    if (!dev_is_dma_coherent(dev)) {
        arch_sync_dma_for_cpu(paddr, size, dir);
        arch_sync_dma_for_cpu_all(); // MIPS quirks...
    }

    /* Miscellaneous cases...*/
}
```

```
/* In arch/arm64/mm/dma-mapping.c */
void arch_sync_dma_for_cpu(
    phys_addr_t paddr,
    size_t size,
    enum dma_data_direction dir,
) {
    /* Translate physical address to (kernel) virtual address */
    unsigned long start = (unsigned long)phys_to_virt(paddr);

    /* Early exit for DMA read: no action needed for CPU */
    if (dir == DMA_TO_DEVICE)
        return;

    /* ARM64-specific: invalidate CPU cache to PoC */
    dcache_inval_poc(start, start + size);
}
```

This call-chain, as well as its mirror case which maintains cache coherency for the DMA device after CPU access:

```
dma_sync_single_for_device(struct device *, dma_addr_t, size_t,
↪  enum dma_data_direction)
```

, call into the following procedures, respectively:

```c
/* Exported @ arch/arm64/include/asm/cacheflush.h */
/* Defined @ arch/arm64/mm/cache.S */
/* All functions accept virtual start, end addresses. */

/* Invalidate data cache region [start, end) to PoC.
 *
 * Invalidate CPU cache entries that intersect with [start, end),
 * such that data from external writers becomes visible to CPU.
 */
extern void dcache_inval_poc(
    unsigned long start, unsigned long end
);

/* Clean data cache region [start, end) to PoC. ??
 *
 * Write-back CPU cache entries that intersect with [start, end),
 * such that data from CPU becomes visible to external writers.
 */
extern void dcache_clean_poc(
    unsigned long start, unsigned long end
);
```

1.3.5.1 Addendum: `enum dma_data_direction`

The Linux kernel defines 4 direction `enum` values for fine-tuning synchronization behaviors:

```c
/* In include/linux/dma-direction.h */
enum dma_data_direction {
    DMA_BIDIRECTION = 0, // data transfer direction uncertain.
    DMA_TO_DEVICE = 1,   // data from main memory to device.
    DMA_FROM_DEVICE = 2, // data from device to main memory.
    DMA_NONE = 3,        // invalid repr for runtime errors.
};
```

These values allow for certain fast-paths to be taken at runtime. For example, `DMA_TO_DEVICE` implies that the device reads data from memory without modification, and hence precludes software coherence instructions from being run when synchronizing for CPU after DMA operation.

1.3.5.2 Use-case: Kernel-space *SMBDirect* Driver

An example of cache-coherent in-kernel RDMA networking module over heterogeneous ISAs could be found in the Linux implementation of *SMBDirect*. *SMBDirect* is an

extension of the *SMB* (*Server Message Block*) protocol for opportunistically establishing the communication protocol over RDMA-capable network interfaces [58].

We focus on two procedures inside the in-kernel SMBDirect implementation:

1.3.5.2.1 Before send: `smbd_post_send`

`smbd_post_send` is a function downstream of the call-chain of `smbd_send`, which sends SMBDirect payload for transport over network. Payloads are constructed and batched for maximized bandwidth, then `smbd_post_send` is called to signal the RDMA NIC for transport.

The function body is roughly as follows:

```
1  /* In fs/smb/client/smbdirect.c */
2  static int smbd_post_send(
3      struct smbd_connection *info, // SMBDirect transport context
4      struct smbd_request *request, // SMBDirect request context
5  ) {
6      struct ib_send_wr send_wr; // Ib "Write Request" for payload
7      int rc, i;
8
9      /* For each message in batched payload */
10     for (i = 0; i < request->num_sge; i++) {
11         /* Log to kmesg ring buffer... */
12
13         /* RDMA wrapper over DMA API1 */
14         ib_dma_sync_single_for_device(
15             info->id->device,       // struct ib_device *
16             request->sge[i].addr,   // u64 (as dma_addr_t)
17             request->sge[i].length, // size_t
18             DMA_TO_DEVICE,          // enum dma_data_direction
19         );
20     }
21
22     /* Populate `request`, `send_wr`... */
23
24     rc = ib_post_send(
25         info->id->qp, // struct ib_qp * ("Queue Pair")
26         &send_wr,     // const struct ib_recv_wr *
27         NULL,         // const struct ib_recv_wr ** (err handling)
28     );
29
30     /* Error handling... */
31
32     return rc;
33 }
```

Line 13 writes back CPU cache lines to be visible for RDMA NIC in preparation for

DMA operations when the posted *send request* is worked upon.

1.3.5.2.2 Upon reception: `recv_done` `recv_done` is called when the RDMA subsystem works on the received payload over RDMA.

Mirroring the case for `smbd_post_send`, it invalidates CPU cache lines for DMA-ed data to be visible at CPU cores prior to any operations on received data:

```
/* In fs/smb/client/smbdirect.c */
static void recv_done(
    struct ib_cq *cq, // "Completion Queue"
    struct ib_wc *wc, // "Work Completion"
) {
    struct smbd_data_transfer *data_transfer;
    struct smbd_response *response = container_of(
        wc->wr_cqe,           // ptr: pointer to member
        struct smbd_response, // type: type of container struct
        cqe,                  // name: name of member in struct
    ); // Cast member of struct into containing struct (C magic)
    struct smbd_connection *info = response->info;
    int data_length = 0;

    /* Logging, error handling... */

    /* Likewise, RDMA wrapper over DMA API1 */
    ib_dma_sync_single_for_cpu(
        wc->qp->device,
        response->sge.addr,
        response->sge.length,
        DMA_FROM_DEVICE,
    );

    /* ... */
}
```

Chapter 2

Software Coherency Latency

Coherency must be maintained at software level when hardware cache coherency cannot be guaranteed for some specific ISA (as established in subsection 1.3.5). There is, therefore, interest in knowing the latency of coherence-maintenance operations for performance engineering purposes, for example OS jitter analysis for scientific computing in heterogeneous clusters and, more pertinently, comparative analysis between software and hardware-backed DSM systems (e.g. [40, 56]).

The purpose of this chapter is hence to provide a statistical analysis over software coherency latency in ARM64 systems by instrumenting hypothetical scenarios of software-initiated coherence maintenance in ARM64 test-benches.

The rest of the chapter is structured as follows:

- **Experiment Setup** covers the test-benches used for instrumentation, including the kernel version, distribution, and the specifications of the instrumented (bare-metal/virtual) machine.
- **Methodology** covers the kernel module and workload used for instrumentation and experimentation, including changes made to the kernel, the kernel module, and userspace programs used for experimentation.
- **Results** covers the results gathered during instrumentation from various test-benches, segmented by experiment.
- **Discussion** identifies key insights from experimental results, as well as deficiencies in research method and possible directions of future works.

2.1 Experiment Setup

2.1.1 QEMU-over-x86: `star`

The primary source of experimental data come from a virtualized machine: a virtualized guest running a lightly-customized Linux v6.7.0 preemptive kernel with standard non-graphical Debian 12 distribution installed to provide userspace support. The specifics of this QEMU-emulated ARM64 test-bench, running atop of an x86-64 host PC, is at 2.1.

Processors	3x QEMU virt-8.2 (2-way SMT; emulates Cortex-A76)					
CPU Flags	fp	asimd	evtstrm	aes	pmull	sha1
	sha2	crc32	atomics	fphp	asimdhp	cpuid
	asimdrdm	lrcpc	dcpop	asimddp		
NUMA Nodes	1: $\{P_0, \ldots, P_5\}$					
Memory	4GiB					

Table 2.1: Specification of `star`

Table 2.2: Specification of Host

2.1.2 *Neoverse N1*: `rose`

2.2 Methodology

2.2.1 Exporting `dcache_clean_poc`

2.2.2 Kernel Module: `my_shmem`

2.2.3 Instrumentation: `ftrace` and *eBPF*

2.2.4 Userspace Programs

2.3 Results

2.3.1 Controlled Allocation Size; Variable Page Count

2.3.2 Controlled Page Count; Variable Allocation Size

2.4 Discussion

Chapter 3

DSM System Design

Chapter 4

Summary

Bibliography

[1] URL: https://www.phoronix.com/search/Heterogeneous%20Memory%20Management.

[2] URL: https://uawartifacts.blob.core.windows.net/upload-files/Altra_Max_Rev_A1_DS_v1_15_20230809_b7cdce449e_424d129849.pdf.

[3] URL: https://hadoop.apache.org/.

[4] URL: https://spark.apache.org/.

[5] Cristiana Amza et al. "Treadmarks: Shared memory computing on networks of workstations". In: *Computer* 29.2 (1996), pp. 18–28.

[6] ARM. *ARM® Cortex®-A Series Programmer's Guide for ARMv8-A*. 2015. URL: https://developer.arm.com/documentation/den0024/a.

[7] Hans J Boehm and Lawrence Crowl. *C++ Atomic Types and Operations*. 2007. URL: https://www.open-std.org/jtc1/sc22/wg21/docs/papers/2007/n2427.html.

[8] *BS ISO/IEC 9899:2011: Information technology. Programming languages. C*. eng. 2013.

[9] Javier Cabezas et al. "GPU-SM: shared memory multi-GPU programming". In: *Proceedings of the 8th Workshop on General Purpose Processing using GPUs*. 2015, pp. 13–24.

[10] Qingchao Cai et al. "Efficient distributed memory management with RDMA and caching". In: *Proceedings of the VLDB Endowment* 11.11 (2018), pp. 1604–1617.

[11] John B Carter, John K Bennett, and Willy Zwaenepoel. "Implementation and performance of Munin". In: *ACM SIGOPS Operating Systems Review* 25.5 (1991), pp. 152–164.

[12] David Chaiken, John Kubiatowicz, and Anant Agarwal. "LimitLESS directories: A scalable cache coherence scheme". In: *Proceedings of the Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*. ASPLOS IV. Santa Clara, California, USA: Association for Computing Machinery, 1991, pp. 224–234. ISBN: 0897913809. DOI: 10.1145/106972.106995. URL: https://doi.org/10.1145/106972.106995.

[13] Jonathan Corbet. 2021. URL: https://lwn.net/Articles/855328/.

[14] Jonathan Corbet. *Heterogeneous memory management meets EXPORT_SYMBOL_GPL()*. 2018. URL: https://lwn.net/Articles/757124/.

[15] Maria Couceiro et al. "D2STM: Dependable distributed software transactional memory". In: *2009 15th IEEE Pacific Rim International Symposium on Dependable Computing*. IEEE. 2009, pp. 307–313.

[16] Mattias De Wael et al. “Partitioned global address space languages”. In: *ACM Computing Surveys (CSUR)* 47.4 (2015), pp. 1–27.

[17] Zhuocheng Ding. “vDSM: Distributed Shared Memory in Virtualized Environments”. In: *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*. 2018, pp. 1112–1115. DOI: 10.1109/ICSESS.2018.8663720.

[18] Noel Eisley, Li-Shiuan Peh, and Li Shang. “In-network cache coherence”. In: *2006 39th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO’06)*. IEEE. 2006, pp. 321–332.

[19] Wataru Endo, Shigeyuki Sato, and Kenjiro Taura. “MENPS: a decentralized distributed shared memory exploiting RDMA”. In: *2020 IEEE/ACM Fourth Annual Workshop on Emerging Parallel and Distributed Runtime Systems and Middleware (IPDRM)*. IEEE. 2020, pp. 9–16.

[20] Brett Fleisch and Gerald Popek. “Mirage: A coherent distributed shared memory design”. In: *ACM SIGOPS Operating Systems Review* 23.5 (1989), pp. 211–223.

[21] Davide Giri, Paolo Mantovani, and Luca P Carloni. “NoC-based support of heterogeneous cache-coherence models for accelerators”. In: *2018 Twelfth IEEE/ACM International Symposium on Networks-on-Chip (NOCS)*. IEEE. 2018, pp. 1–8.

[22] Mark Harris. *Unified memory for cuda beginners*. 2017. URL: https://developer.nvidia.com/blog/unified-memory-cuda-beginners/.

[23] John L Hennessy and David A Patterson. *Computer architecture: a quantitative approach*. Elsevier, 2011.

[24] Stephen Alan Holsapple. *DSM64: A Distributed Shared Memory System in User-Space*. California Polytechnic State University, 2012.

[25] Yang Hong et al. “Scaling out NUMA-aware applications with RDMA-based distributed shared memory”. In: *Journal of Computer Science and Technology* 34 (2019), pp. 94–112.

[26] Weiwu Hu, Weisong Shi, and Zhimin Tang. “JIAJIA: A software DSM system based on a new cache coherence protocol”. In: *High-Performance Computing and Networking: 7th International Conference, HPCN Europe 1999 Amsterdam, The Netherlands, April 12–14, 1999 Proceedings 7*. Springer. 1999, pp. 461–472.

[27] Ayal Itzkovitz, Assaf Schuster, and Lea Shalev. “Thread migration and its applications in distributed shared memory systems”. In: *Journal of Systems and Software* 42.1 (1998), pp. 71–87.

[28] Chengfan Jia et al. “Improving the performance of distributed tensorflow with RDMA”. In: *International Journal of Parallel Programming* 46 (2018), pp. 674–685.

[29] Stefanos Kaxiras et al. “Turning Centralized Coherence and Distributed Critical-Section Execution on their Head: A New Approach for Scalable Distributed Shared Memory”. In: *Proceedings of the 24th International Symposium on High-Performance Parallel and Distributed Computing*. HPDC ’15. Portland, Oregon, USA: Association for Computing Machinery, 2015, pp. 3–14. ISBN: 9781450335508. DOI: 10.1145/2749246.2749250. URL: https://doi.org/10.1145/2749246.2749250.

[30] Ahmed Khawaja et al. “Sharing, Protection, and Compatibility for Reconfigurable Fabric with {AmorphOS}”. In: *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*. 2018, pp. 107–127.

[31] Sang-Hoon Kim et al. “DeX: Scaling Applications Beyond Machine Boundaries”. In: *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*. 2020, pp. 864–876. DOI: 10.1109/ICDCS47774.2020.00021.

[32] Toddj Kjos et al. *Hardware cache coherent input/output*. eng. PALO ALTO, 1996.

[33] Daniel Lenoski et al. “The stanford dash multiprocessor”. In: *Computer* 25.3 (1992), pp. 63–79.

[34] Feng Li et al. “Accelerating relational databases by leveraging remote memory and RDMA”. In: *Proceedings of the 2016 International Conference on Management of Data*. 2016, pp. 355–370.

[35] Kai Li and Richard Schaefer. “Shiva: An operating system transforming a hypercube into a shared-memory machine”. In: (1989).

[36] Honghui Lu et al. “Message passing versus distributed shared memory on networks of workstations”. In: *Supercomputing’95: Proceedings of the 1995 ACM/IEEE Conference on Supercomputing*. IEEE. 1995, pp. 37–37.

[37] Xiaoyi Lu et al. “Accelerating spark with RDMA for big data processing: Early experiences”. In: *2014 IEEE 22nd Annual Symposium on High-Performance Interconnects*. IEEE. 2014, pp. 9–16.

[38] Jiacheng Ma et al. “A hypervisor for shared-memory FPGA platforms”. In: *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*. 2020, pp. 827–844.

[39] Jeremy Manson and Brian Goetz. 2004. URL: https://www.cs.umd.edu/~pugh/java/memoryModel/jsr-133-faq.html.

[40] Dimosthenis Masouros et al. “Adrias: Interference-Aware Memory Orchestration for Disaggregated Cloud Infrastructures”. In: *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE. 2023, pp. 855–869.

[41] David S Miller, Richard Henderson, and Jakub Jelinek. *Dynamic DMA mapping Guide*. 2024. URL: https://www.kernel.org/doc/html/v6.7/core-api/dma-api-howto.html.

[42] Vijay Nagarajan et al. *A primer on memory consistency and cache coherence*. Springer Nature, 2020.

[43] Jacob Nelson et al. “{Latency-Tolerant} software distributed shared memory”. In: *2015 USENIX Annual Technical Conference (USENIX ATC 15)*. 2015, pp. 291–305.

[44] SeungYong Oh and JongWon Kim. “Stateful Container Migration employing Checkpoint-based Restoration for Orchestrated Container Clusters”. In: *2018 International Conference on Information and Communication Technology Convergence (ICTC)*. 2018, pp. 25–30. DOI: 10.1109/ICTC.2018.8539562.

[45] *Ordering in core::sync::atomic - Rust*. 2024. URL: https://doc.rust-lang.org/core/sync/atomic/enum.Ordering.html.

[46] Neil Parris. *Extended system coherency: Cache Coherency Fundamentals*. 2013. URL: https://community.arm.com/arm-community-blogs/b/architectures-

and-processors-blog/posts/extended-system-coherency---part-1---cache-coherency-fundamentals.

[47] Christian Pinto et al. "Thymesisflow: A software-defined, hw/sw co-designed interconnect stack for rack-scale memory disaggregation". In: *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE. 2020, pp. 868–880.

[48] *Platform-Specifc Notes*. 2023. URL: https://chapel-lang.org/docs/platforms/index.html#.

[49] The FreeBSD Project. *FreeBSD manual pages*. 2021. URL: https://man.freebsd.org/cgi/man.cgi?query=bpf&manpath=FreeBSD+14.0-RELEASE+and+Ports.

[50] Manuel Rodríguez-Pascual et al. "Job migration in hpc clusters by means of checkpoint/restart". In: *The Journal of Supercomputing* 75 (2019), pp. 6517–6541.

[51] Ioannis Schoinas et al. "Sirocco: Cost-effective fine-grain distributed shared memory". In: *Proceedings. 1998 International Conference on Parallel Architectures and Compilation Techniques (Cat. No. 98EX192)*. IEEE. 1998, pp. 40–49.

[52] Yizhou Shan, Shin-Yeh Tsai, and Yiying Zhang. "Distributed Shared Persistent Memory". In: *Proceedings of the 2017 Symposium on Cloud Computing*. SoCC '17. Santa Clara, California: Association for Computing Machinery, 2017, pp. 323–337. ISBN: 9781450350280. DOI: 10.1145/3127479.3128610. URL: https://doi.org/10.1145/3127479.3128610.

[53] *upcc.1*. 2022. URL: https://upc.lbl.gov/docs/user/upcc.html.

[54] Maarten Van Steen and Andrew S Tanenbaum. *Distributed systems*. Maarten van Steen Leiden, The Netherlands, 2017.

[55] Arjan van de Ven. *Background on ioremap, cacheing, cache coherency on x86*. 2008. URL: https://lkml.org/lkml/2008/4/29/480.

[56] Qing Wang et al. "Concordia: Distributed Shared Memory with In-Network Cache Coherence". In: *19th USENIX Conference on File and Storage Technologies (FAST 21)*. USENIX Association, Feb. 2021, pp. 277–292. ISBN: 978-1-939133-20-5. URL: https://www.usenix.org/conference/fast21/presentation/wang.

[57] Paul Werstein, Mark Pethick, and Zhiyi Huang. "A performance comparison of dsm, pvm, and mpi". In: *Proceedings of the Fourth International Conference on Parallel and Distributed Computing, Applications and Technologies*. IEEE. 2003, pp. 476–482.

[58] Xelu86 et al. *SMB Direct*. 2024. URL: https://learn.microsoft.com/en-us/windows-server/storage/file-server/smb-direct.

[59] Matei Zaharia et al. "Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing". In: *9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*. San Jose, CA: USENIX Association, Apr. 2012, pp. 15–28. ISBN: 978-931971-92-8. URL: https://www.usenix.org/conference/nsdi12/technical-sessions/presentation/zaharia.

[60] Jin Zhang et al. “Giantvm: A type-ii hypervisor implementing many-to-one virtualization”. In: *Proceedings of the 16th ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments*. 2020, pp. 30–44.

[61] Huan Zhou et al. “DART-MPI: An MPI-based implementation of a PGAS runtime system”. In: *Proceedings of the 8th International Conference on Partitioned Global Address Space Programming Models*. 2014, pp. 1–11.

Appendix A

First appendix

A.1 First section

Any appendices, including any required ethics information, should be included after the references.

Markers do not have to consider appendices. Make sure that your contributions are made clear in the main body of the dissertation (within the page limit).