Herding Cats - Modelling, simulation, testing, and data-mining for weak memory

Alglave, J; Maranget, L; Tautschnig, M

Herding cats
Modelling, simulation, testing, and data-mining for weak memory

Jade Alglave, University College London
Luc Maranget, INRIA
Michael Tautschnig, Queen Mary University of London

We propose an axiomatic generic framework for modelling weak memory. We show how to instantiate this framework for SC, TSO, C++ restricted to release-acquire atomics, and Power. For Power, we compare our model to a preceding operational model in which we found a flaw. To do so, we define an operational model that we show equivalent to our axiomatic model.

We also propose a model for ARM. Our testing on this architecture revealed a behaviour later acknowledged as a bug by ARM, and more recently 31 additional anomalies.

We offer a new simulation tool, called herd, which allows the user to specify the model of his choice in a concise way. Given a specification of a model, the tool becomes a simulator for that model. The tool relies on an axiomatic description; this choice allows us to outperform all previous simulation tools. Additionally, we confirm that verification time is vastly improved, in the case of bounded model-checking.

Finally, we put our models in perspective, in the light of empirical data obtained by analysing the C and C++ code of a Debian Linux distribution. We present our new analysis tool, called mole, which explores a piece of code to find the weak memory idioms that it uses.

1. INTRODUCTION

There is a joke where a physicist and a mathematician are asked to herd cats. The physicist starts with an infinitely large pen which he reduces until it is of reasonable diameter yet contains all the cats. The mathematician builds a fence around himself and declares the outside to be the inside. Defining memory models is akin to herding cats: both the physicist's or mathematician's attitudes are tempting.

Recent years have seen many formalisations of memory models emerge; see for example [Adir et al. 2003; Arvind and Maessen 2006; Boehm and Adve 2008; Chong and Ishtiaq 2008; Boudol and Petri 2009; Sarkar et al. 2009; Alglave et al. 2009; Alglave et al. 2010; Batty et al. 2011; Sarkar et al. 2011; Alglave et al. 2012; Sarkar et al. 2012; Alglave 2012; Mador-Haim et al. 2012; Boudol et al. 2012]. Yet we feel the need for more work in the area of *defining* models. There are several reasons for this.

On the hardware side, all existing models of Power (some of which we list in Fig. I) have some flaws (see Sec. 2). This calls for reinvestigating the model, for the sake of repairing it of course, but for several other reasons too, that we explain below.

One important reason is that Power underpins C++'s *atomic* concurrency features [Boehm and Adve 2008; Batty et al. 2011; Sarkar et al. 2012]: implementability on Power has had a major influence on the design of the C++ model. Thus modelling flaws in Power could affect C++.

Another important reason is that a good fraction of the code in the wild (see our experiments in Sec. 9 on release 7.1 of the Debian Linux distribution) still does not use the C++ atomics. Thus, we believe that programmers have to rely on what the hardware does, which requires descriptive models of the hardware.

On the software side, recent work shows that the C++ model allows behaviours that break modular reasoning (see the *satisfaction cycles* issue in [Batty et al. 2013]), whereas Power does not, since it prevents *out of thin air* values (see Sec. 4). Moreover, C++ requires the irreflexivity of a certain relation (see the HBVSMO axiom in [Batty et al. 2013]), whereas Power offers a stronger acyclicity guarantee, as we show in this paper.

Ideally, we believe that these models would benefit from stating *principles* that underpin weak memory as a whole, not just one particular architecture or language. Not only would it be aesthetically pleasing, but it would allow more informed decisions on the design of high-level memory models, ease the conception and proofs of compilation schemes, and allow the reusability of simulation and verification techniques from one model to another.

Models roughly fall into two classes: *operational* and *axiomatic*. Operational models, e.g. the Power model of [Sarkar et al. 2011], are abstractions of actual machines, composed of idealised hardware components such as buffers and queues. Axiomatic models, e.g. the C++ model of [Batty et al. 2011], distinguish allowed from forbidden behaviours, usually by constraining various relations on memory accesses.

We now list a few criteria that we believe our models should meet; we do not claim to be exhaustive, nor do we claim that the present work fully meets all of them, although we discuss in the conclusion of this paper to what extent it does. Rather, we see this list as enunciating some wishes for works on weak memory (including ours of course), and more generally weak consistency, as can be found for example in distributed systems.

***

Stylistic proximity of models, whether hardware (e.g. x86, Power or ARM) or software (e.g. C++), would permit the statement of general principles spanning several models of weak memory. It should be easier to find principles common to Power and C++, amongst others, if their respective models were described in the same terms.

Rigour is not our only criterion: for example, all the recent Power models enjoy a considerable amount of rigour, yet are still somewhat flawed.

Concision of the model seems crucial to us: we want to specify a model concisely, to grasp it and modify it rapidly, without needing to dive into a staggering number of definitions.

One could tend towards *photorealistic* models and account for each and every detail of the machine. We find that operational models in general have such traits, although some do more than others. For example, we find that the work of Sarkar et al. [Sarkar et al. 2011; Sarkar et al. 2012] is too close to the hardware, and, perhaps paradoxically, too precise, to be easily amenable to pedagogy, automated reasoning and verification. Although we do recognise the effort and the value of this work, without which we would not have been able to build the present work, we believe that we need models that are more *rationally descriptive* (as coined by Richard Bornat). We go back to the meaning of "rational" at the end of the introduction.

Efficient simulation and verification have been relatively neglected by previous modelling work, except for [Mador-Haim et al. 2010; Alglave et al. 2013; Alglave et al. 2013]. These works show that simulation [Mador-Haim et al. 2010] and verification [Alglave et al. 2013] (for bounded model-checking) can be orders of magnitude faster when it relies on axiomatic models rather than operational ones.

Yet operational models are often considered more intuitive than axiomatic models. Perhaps the only tenable solution to this dilemma is to propose both styles in tandem, and show their equivalence. As exposed in [Hoare and Lauer 1974], *"a single formal definition is unlikely to be equally acceptable to both implementor and user, and [...] at least two definitions are required, a constructive one [...] for the implementor, and an implicit one for the user [...]."*

Soundness w.r.t. hardware is mandatory regardless of the modelling style. Ideally, one would prove the soundness of a model w.r.t. the formal description of the hardware,

e.g. at the RTL level [Gordon 2002]. However, we do not have access to these data because of commercial confidentiality. To overcome this issue, some previous work has involved experimental testing of hardware (see e.g. [Collier 1992; Sarkar et al. 2009; Alglave et al. 2012; Sarkar et al. 2011; Mador-Haim et al. 2012]), with increasing thoroughness over time. [1]

A credible model cannot forbid behaviours exhibited on hardware, unless the hardware itself is flawed. Thus models should be extensively tested against hardware, and retested regularly: this is how we found flaws in the model of [Sarkar et al. 2011] (see Sec. 2). Yet, we find the experimental flaws themselves to be less of an issue than the fact that the model does not seem to be easily fixable.

Adaptability of the model, i.e. setting the model in a flexible formalism, seems crucial, if we want stable models. By stable we mean that even though we might need to change parameters to account for an experimental flaw, the general shape of the model, its principles, should not need to change.

Testing (as extensive as it may be) cannot be sufficient, since it cannot guarantee that an unobserved behaviour might not be triggered in the future. Thus one needs some guarantee of *completeness* of the model.

Being in accord with the architectural intent might give some guarantee of completeness. We should try to create models that respect or take inspiration from the architects' intents. This is one of the great strengths of the model of [Sarkar et al. 2011]. However, this cannot be the only criterion, as the experimental flaw in [Sarkar et al. 2011] shows. Indeed the architects' intents might not be as formal as one might need in a model, two intents might be contradictory, or an architect might not realise all the consequences of a given design.

Accounting for what programmers do seems a sensible criterion. One cannot derive a model from programming patterns, since some of these patterns might rely on erroneous understanding of the hardware. Yet to some extent, these patterns should reflect part of the architectural intent, since systems programmers or compiler writers communicate relatively closely with hardware designers.

Crucially, we have access to open-source code, as opposed to the chips' designs. Thus we can analyse the code, and derive some common programming patterns from it.

***

Rational models is what we advocate here. We believe that a model should allow a *rational explanation* of what programmers can rely on. We believe that by balancing all the criteria above, one can provide such a model. This is what we set out to do in this paper.

By rational we mean the following: we think that we, as architects, semanticists, programmers, compiler writers, are to understand concurrent programs. Moreover we find that to do so, we are to understand some particular patterns (e.g. the message passing pattern given in Fig. 8, the very classical store buffering pattern given in Fig. 14, or the controversial load buffering pattern given in Fig. 7). We believe that by being able to explain a handful of patterns, one should be able to generalise the explanation and thus to understand a great deal of weak memory.

To make this claim formal and precise, we propose a generic model of weak memory, in axiomatic style. Each of our four axioms has a few canonical examples, that should be enough to understand the full generality of the axiom. For example, we believe

[1] Throughout this paper, we refer to online material to justify our experimental claims; the reader should take these as bibliography items, and refer to them when details are needed.

that our NO THIN AIR axiom is fully explained by the load buffering pattern of Fig. 7. Similarly our OBSERVATION axiom is fully explained by the message passing, write to read causality and Power ISA2 patterns of Fig. 8,11 and 12 respectively.

On the modelling front, our main stylistic choices and contributions are as follows: to model the propagation of a given store instruction to several different threads, we use only one memory event per instruction (see Sec. 4), instead of several subevents (one per thread for example, as one would do in Itanium [itanium 2002] or in the Power model of [Mador-Haim et al. 2012]). We observe that this choice makes simulation much faster (see Sec. 8).

To account for the complexity of write propagation, we introduce the novel notion of *propagation order*. This notion is instrumental in describing the semantics of fences for instance, and the subtle interplay between fences and coherence (see Sec. 4).

We deliberately try to keep our models concise, as we aim at describing them as simple text files that one can use as input to an automated tool (e.g. a simulation tool, or a verification tool). We note that we are able to describe IBM Power in less than a page (see Fig. 38).

Outline: we present related works in Sec. 2. We describe our new generic model of weak memory in Sec. 4, and show how to instantiate it to describe Sequential Consistency (SC) [Lamport 1979], Total Store Order (TSO) (used in Sparc [sparc 1994] and x86's [Owens et al. 2009] architectures) and C++ restricted to release-acquire atomics.

Sec. 5 presents examples of the semantics of instructions, which are necessary to understand Sec. 6, where we explain how to instantiate our model to describe Power and ARMv7. We compare formally our Power model to the one of [Sarkar et al. 2011] in Sec. 7. To do so, we define an operational model that we show equivalent to our axiomatic model.

We then present our experiments on Power and ARM hardware in Sec. 8, detailing the anomalies that we observed on ARM hardware. We also describe our new herd simulator, which allows the user to specify the model of his choice in a concise way. Given a specification of a model, the tool becomes a simulator for that model.

Additionally, we demonstrate in the same section that our model is suited for verification by implementing it in the bounded model-checker CBMC [Clarke et al. 2004] and comparing it with the previously implemented models of [Alglave et al. 2012] and [Mador-Haim et al. 2012].

In Sec. 9, we present our mole analysis tool, which explores a piece of code to find the weak memory behaviours that it contains. We detail the data gathered by mole by analysing the C and C++ code in a Debian Linux distribution. This gives us a pragmatic perspective on the models that we present in Sec. 4. Additionally, mole may be used by programmers to identify areas of their code that may be (unwittingly) affected by weak memory, or by static analysis tools to identify areas where more fine-grained analysis may be required.

Online companion material: we provide the source and documentation of herd at http://diy.inria.fr/herd. We provide all our experimental reports w.r.t. hardware at http://diy.inria.fr/cats. We provide our Coq scripts at http://www0.cs.ucl.ac.uk/staff/j.alglave/cats. We provide the source and documentation of mole at http://diy.inria.fr/mole, as well as our experimental reports w.r.t. release 7.1 of the Debian Linux distribution.

2. RELATED WORK

Our introduction echoes position papers by Burckhardt and Musuvathi [Burckhardt and Musuvathi 2008], Zappa Nardelli et al. [Zappa Nardelli et al. 2009] and Adve and Boehm [Adve and Boehm 2010;

Boehm and Adve 2012], which all formulate criteria, prescriptions or wishes as to how memory models should be defined.

Looking for general principles of weak memory, one might look at the hardware documentation: we cite Alpha [alpha 2002], ARM [arm 2008], Intel [intel 2009], Itanium [itanium 2002], IBMPower [ppc 2009] and Sun [spa 1994]. Ancestors of our SC PER LOCATION and NO THIN AIR axioms (see Sec. 4) appear notably in Sun and Alpha's documentations.

We also refer the reader to work on modelling particular instances of weak memory, e.g. ARM [Chong and Ishtiaq 2008], TSO [Boudol and Petri 2009] or x86 [Sarkar et al. 2009; Owens et al. 2009], C++ [Boehm and Adve 2008; Batty et al. 2011], or Java [Manson et al. 2005; Cenciarelli et al. 2007]. We go back to the case of Power at the end of this section.

In the rest of this paper, we write TSO for Total Store Order, implemented in Sparc TSO [sparc 1994] and Intel x86 [Owens et al. 2009]. We write PSO for Partial Store Order and RMO for Relaxed Memory Order, two other Sparc architectures. We write Power for IBM Power [ppc 2009].

Collier [Collier 1992], Neiger [Neiger 2000], as well as Adve and Gharachorloo [Adve and Gharachorloo 1995] have provided general overviews of weak memory, but in a less formal style than one might prefer.

Steinke and Nutt provide a unified framework to describe consistency models. They choose to express their models in terms of the view order of each processor, and describe instances of their framework, amongst them several classical models such as PRAM [Lipton and Sandberg 1988] or Cache Consistency [Goodman 1989]. Meyer et al. later showed how to instantiate Steinke and Nutt's framework to express TSO and PSO. However, it is unclear to us whether the view order method à la Steinke and Nutt is expressive enough to describe models without store atomicity, such as Power and ARM, or indeed describe them easily.

Rational models appear in Arvind and Maessen's work, aiming at weak memory in general but applied only to TSO [Arvind and Maessen 2006], and in Batty et al.'s [Batty et al. 2013] for C++. Interestingly, Burckhardt et al.'s work on distributed systems [Burckhardt et al. 2013; Burckhardt et al. 2014] belongs to the same trend.

Some works on weak memory provide simulation tools: the ppcmem tool of [Sarkar et al. 2011] and Boudol et al.'s [Boudol et al. 2012] implement their respective operational model of Power, whilst the cppmem tool of [Batty et al. 2011] enumerates the axiomatic executions of the associated C++ model. Mador-Haim et al.'s tool [Mador-Haim et al. 2012] does the same for their axiomatic model of Power. MemSAT has an emphasis towards the Java memory model [Torlak et al. 2010], whilst Nemos focusses on classical models such as SC or causal consistency [Yang et al. 2004], and TSOTool handles TSO [Hangal et al. 2004].

To some extent, decidability and verification papers [Gopalakrishnan et al. 2004; Burckhardt et al. 2007; Atig et al. 2010; Bouajjani et al. 2011; Atig et al. 2011; Atig et al. 2012; Bouajjani et al. 2013; Kuperstein et al. 2010; Kuperstein et al. 2011; Liu et al. 2012; Abdulla et al. 2012; Abdulla et al. 2013] do provide some general principles about weak memory, although we find them less directly applicable to programming than semantics work. Unlike our work, most of them are restricted to TSO, or its siblings PSO and RMO, or theoretical models.

Notable exceptions are [Alglave et al. 2013; Alglave et al. 2013] which use the generic model of [Alglave et al. 2012]. The present paper inherits some of the concepts developed in [Alglave et al. 2012]: it adopts the same style in describing executions of programs, and pursues the same goal of defining a generic model of weak memory. Moreover, we adapt the tool of [Alglave et al. 2013] to our new models, and reuse the diy testing tool of [Alglave et al. 2012] to conduct our experiments against hardware.

Yet we emphasise that the model that we present here is quite different from the model of [Alglave et al. 2012], despite the stylistic similarities: in particular [Alglave et al. 2012] did not have a distinction between the OBSERVATION and PROPAGATION axioms (see Sec. 4), which were somewhat merged into the *global happens-before* notion of [Alglave et al. 2012].

model	style	comments
[Adir et al. 2003]	axiomatic	based on discussion with IBM architects; pre-cumulative barriers
[Alglave et al. 2009]	axiomatic	based on documentation; practically untested
[Alglave et al. 2012; Alglave and Maranget 2011]	single-event axiomatic	based on extensive testing; semantics of lwsync stronger than [Sarkar et al. 2011] on r+lwsync+sync, weaker on mp+lwsync+addr
[Sarkar et al. 2011; Sarkar et al. 2012]	operational	based on discussion with IBM architects and extensive testing; flawed w.r.t. Power h/w on e.g. mp+lwsync+addr-po-detour (see Fig. 36 and http://diy.inria.fr/cats/pldi-power/#lessvs) and ARM h/w on e.g. mp+dmb+fri-rfi-ctrlisb (see http://diy.inria.fr/cats/pldi-arm/#lessvs)
[Mador-Haim et al. 2012]	multi-event axiomatic	thought to be equivalent to [Sarkar et al. 2011] but not experimentally on e.g. mp+lwsync+addr-po-detour (see http://diy.inria.fr/cats/cav-power)
[Boudol et al. 2012]	operational	semantics of lwsync stronger than [Sarkar et al. 2011] on e.g. r+lwsync+sync
[Alglave et al. 2013]	operational	equivalent to [Alglave et al. 2012]

Table I. A decade of Power models in order of publication

A decade of Power models is presented in Tab. I. Earlier work (omitted for brevity) accounted for outdated versions of the architecture. For example in 2003, Adir et al. described an axiomatic model [Adir et al. 2003], *"developed through [. . .] discussions with the PowerPC architects"*, with outdated *non-cumulative barriers*, following the pre-PPC 1.09 PowerPC architecture.

Below, we refer to particular weak memory behaviours which serve as test cases for distinguishing different memory architectures. These behaviours are embodied by *litmus tests*, with standardised names in the style of [Sarkar et al. 2011], e.g. mp+lwsync+addr. All these behaviours will appear in the rest of the paper, so that the novice reader can refer to them after a first read through. We explain the naming convention in Sec. 4.

In 2009, Alglave et al. proposed an axiomatic model [Alglave et al. 2009], but this was not compared to actual hardware. In 2010, they provided another axiomatic model [Alglave et al. 2010; Alglave et al. 2012], as part of a generic framework. This

model is based on extensive and systematic testing. It appears to be sound w.r.t. Power hardware, but its semantics for lwsync cannot guarantee the mp+lwsync+addr behaviour (see Fig. 8), and allows the r+lwsync+sync behaviour (see Fig. 16), both of which clearly go against the architectural intent (see [Sarkar et al. 2011]). This model (and the generic framework to which it belongs) has a provably equivalent operational counterpart [Alglave et al. 2013].

In 2011, Sarkar et al. [Sarkar et al. 2011; Sarkar et al. 2012] proposed an operational model in collaboration with an IBM designer, which might be taken to account for the architectural intent. Yet, we found this model to forbid several behaviours observed on Power hardware (e.g. mp+lwsync+addr-po-detour, see Fig. 36 and http://diy.inria.fr/cats/pldi-power/#lessvs). Moreover, although this model was not presented as a model for ARM, it was thought to be a suitable candidate. Yet, it forbids behaviours (e.g. mp+dmb+fri-rfi-ctrlisb, see Fig. 32 http://diy.inria.fr/cats/pldi-arm/#lessvs) that are observable on ARM machines, and claimed to be desirable features by ARM designers.

In 2012, Mador-Haim et al. [Mador-Haim et al. 2012] proposed an axiomatic model, thought to be equivalent to the one of [Sarkar et al. 2011]. Yet, this model does not forbid the behaviour of mp+lwsync+addr-po-detour (see http://diy.inria.fr/cats/cav-power), which is a counter-example to the proof of equivalence appearing in [Mador-Haim et al. 2012]. The model of [Mador-Haim et al. 2012] also suffers from the same experimental flaw w.r.t. ARM hardware as the model of [Sarkar et al. 2011].

More fundamentally, the model of [Mador-Haim et al. 2012] uses several write events to represent the propagation of one memory store to several different threads, which in effect mimics the operational transitions of the model of [Sarkar et al. 2011]. We refer to this style as *multi-event axiomatic*, as opposed to *single-event axiomatic* (as in e.g. [Alglave et al. 2010; Alglave et al. 2012]), where there is only one event to represent the propagation of a given store instruction. Our experiments (see Sec. 8) show that this choice impairs the simulation time by up to a factor of ten.

Later in 2012, Boudol et al. [Boudol et al. 2012] proposed an operational model where the semantics of lwsync is stronger than the architectural intent on e.g. r+lwsync+sync (like Alglave et al.'s [Alglave et al. 2010]).

3. PREAMBLE ON AXIOMATIC MODELS

We give here a brief presentation of axiomatic models in general. The expert reader might want to skip this section.

Axiomatic models are usually defined in three stages. First, an *instruction semantics* maps each instruction to some mathematical objects. This allows us to define the *control flow semantics* of a multi-threaded program. Second, we build a set of *candidate executions* from this control-flow semantics: each candidate execution represents one particular data-flow of the program, i.e. which *communications* might happen between the different threads of our program. Third, a *constraint specification* decides which candidate executions are valid or not.

We now explain these concepts in a way that we hope to be intuitive. Later in this paper, we give the constraint specification part of our model in Sec. 4, and an outline of the instruction semantics in Sec. 5.

Multi-threaded programs, such as the one given in Fig. 1, give one sequence of *instructions* per thread. Instructions can come from a given assembly language instruction set, e.g. Power ISA, or be pseudo-code instructions, as is the case in Fig. 1.

In Fig. 1, we have two threads T_0 and T_1 in parallel. These two threads communicate via the two memory locations x and y, which hold the value 0 initially. On T_0 we have a store of value 1 into memory location x, followed in program order by a store of value 1 into memory location y. On T_1 we have a load of the contents of memory location y into

mp	
initially `x=0; y=0`	
T_0	T_1
(a) `x` $\leftarrow$ `1`	(c) `r1` $\leftarrow$ `y`
(b) `y` $\leftarrow$ `1`	(d) `r2` $\leftarrow$ `x`

Fig. 1. A multi-threaded program implementing a message passing pattern

register r_1, followed in program order by a load of the contents of memory location x into register r_2. Memory locations, e.g. x and y, are shared by the two threads, whereas the registers are private to the thread holding them, here T_1.

The snippet in Fig. 1 is at the heart of a message passing pattern, where T_0 would write some data into memory location x, then set a flag in y. T_1 would then check if he has the flag, then read the data in x.

Control-flow semantics The instruction semantics, in our case, translates instructions into *events*, that represent e.g. *memory or register accesses* (i.e. reads and writes from and to memory or registers), *branching decisions* or *fences*.

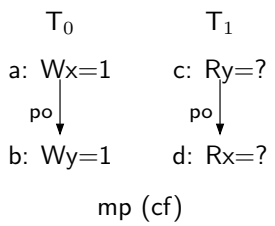


Fig. 2. Control-flow semantics for the message passing pattern of Fig. 1

Consider Fig. 2: we give a possible control-flow semantics to the program in Fig. 1. To do so, we proceed as follows: each store instruction, e.g. `x` $\leftarrow$ `1` on T_0, corresponds to a write event specifying a memory location and a value, e.g. Wx=1. Each load instruction, e.g. `r1` $\leftarrow$ `y` on T_1 corresponds to a read event specifying a memory location and a undetermined value, e.g. Ry=?. Note that the memory locations of the events are determined by the program text, as well as the values of the writes. For reads, the values will be determined in the next stage.

Additionally, we also have fictitious write events Wx=0 and Wy=0 representing the initial state of x and y, that we do not depict here.

The instruction semantics also defines *relations* over these events, representing for example the *program order* within a thread, or *address, data or control dependencies* from one memory access to the other, via computations over register values.

Thus in Fig. 2, we also give the program order relation, written po, in Fig. 1, which lifts the order in which instructions have been written to the level of events. For example, the two stores on T_0 in Fig. 1 have been written in program order, thus their corresponding events Wx=1 and Wy=1 are related by po in Fig. 2.

We are now at a stage where we have, given a program such as the one in Fig. 1, several *event graphs*, such as the one in Fig. 2. Each graph gives a set of events representing accesses to memory and registers, the program order between these events, including branching decisions, and the dependencies.

Data-flow semantics The purpose of this step is to define which *communications*, or *interferences*, might happen between the different threads of our program. To do so, we need to define two relations over memory events: the *read-from* relation rf, and the *coherence order* co.

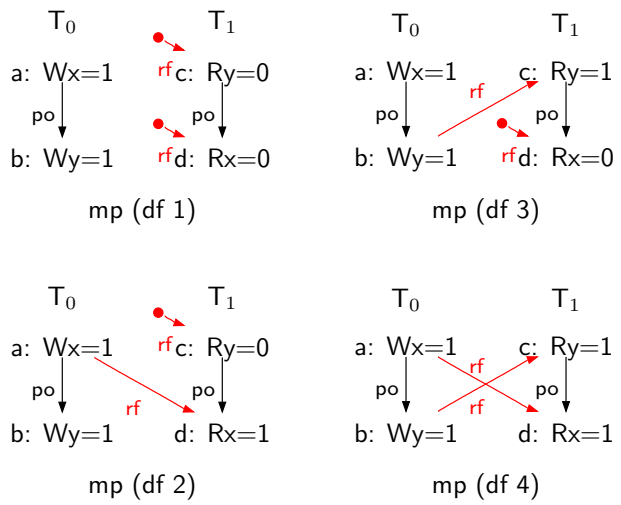


Fig. 3. One possible data-flow semantics per control-flow semantics as given in Fig. 2

The read-from relation rf describes, for any given read, from which write this read could have taken its value. A read-from arrow with no source, as in the top left of Fig. 2, corresponds to reading from the initial state.

For example in Fig. 3, consider the drawing at the bottom left-most corner. The read c from y takes its value from the initial state, hence reads the value 0. The read d from x takes its value from the update a of x by T_0, hence reads the value 1.

The coherence order gives the order in which all the memory writes to a given location have hit that location in memory. For example in Fig. 3, the initial write to x (not depicted) hits the memory before the write a on T_0, by convention, hence the two writes are ordered in coherence.

We are now at a stage where we have, given a program such as the one in Fig. 1, several event graphs as given by the control-flow semantics (see Fig. 2), and for each of these, a read-from relation and a coherence order describing the communications across threads (see Fig. 3).

Note that for a given control-flow semantics there could be several suitable data-flow semantics, if for example there were several writes to x with value 1 in our example: in that case there would be two possible read-from to give a value to the read of x on T_1.

Each such object (see Fig. 3, which gathers events, program order, dependencies, read-from and coherence, is called a *candidate execution*. As one can see in Fig. 3, there can be more than one candidate execution for a given program.

Constraint specification Now, for each candidate execution, the constraint specification part of our model decides if this candidate represents a valid execution or not.

Traditionally, such specifications are in terms of acyclicity or irreflexivity of various combinations of the relations over events given by the candidate execution. This

means for example that the model would reject a candidate execution if this candidate contains a cycle amongst a certain relation defined in the constraint specification.

For example in Fig. 3, the constraints for describing Lamport's Sequential Consistency [Lamport 1979] (see also Sec. 4.8) would rule out the right-most top candidate execution because the read from x on T_1 reads from the initial state, whereas the read of y on T_1 has observed the update of y by T_0.

4. A MODEL OF WEAK MEMORY

We present our axiomatic model, and show its SC and TSO instances. We also explain how to instantiate our model to produce C++ R-A, i.e. the fragment of C++ restricted to the use of release-acquire atomics. Sec. 6 presents Power.

The inputs to our model are candidate executions of a given multi-threaded program. Candidate executions can be ruled out by the four axioms of our model, given in Fig. 5: SC PER LOCATION, NO THIN AIR, OBSERVATION and PROPAGATION.

4.1. Preliminaries

Before explaining each of these axioms, we define a few preliminary notions.

Conventions: in this paper, we use several notations on *relations* and *orders*. We denote the transitive (resp. reflexive-transitive) closure of a relation r as r^+ (resp. r^*). We write $\mathsf{r}_1;\mathsf{r}_2$ for the sequence of two relations r_1 and r_2, i.e. $(x,y) \in (\mathsf{r}_1;\mathsf{r}_2) \triangleq \exists z.(x,z) \in \mathsf{r}_1 \wedge (z,y) \in \mathsf{r}_2$. We write $\mathrm{irreflexive}(\mathsf{r})$ to express the irreflexivity of r, i.e. $\neg(\exists x.(x,x) \in \mathsf{r})$. We write $\mathrm{acyclic}(\mathsf{r})$ to express its acyclicity, i.e. $\neg(\exists x.(x,x) \in \mathsf{r}^+)$.

A *partial order* is a relation r that is *transitive* (i.e. $\mathsf{r} = \mathsf{r}^+$), and irreflexive. Note that this entails that r is also acyclic. A *total order* is a partial order r defined over a set $\mathbb{S}$ that enjoys the *totality property*: $\forall x \neq y \in \mathbb{S}.(x,y) \in \mathsf{r} \vee (y,x) \in \mathsf{r}$.

Executions are tuples $(\mathbb{E}, \mathsf{po}, \mathsf{rf}, \mathsf{co})$, which consist of a set of *events* $\mathbb{E}$, giving a semantics to the instructions, and three relations over events: po, rf and co (see below).

Events consist of a unique identifier (in this paper we use lower-case letters, e.g. a), the thread holding the corresponding instruction (e.g. T_0), the line number or program counter of the instruction, and an *action*.

Actions are of several kinds, which we detail in the course of this paper. For now, we only consider read and write events relative to memory locations. For example for the location x we can have a read of the value 0 noted $\mathrm{R}x = 0$, or a write of the value 1, noted $\mathrm{W}x = 1$. We write $\mathrm{proc}(e)$ for the thread holding the event e, and $\mathrm{addr}(e)$ for its location.

Given a candidate execution, the events are determined by the program's instruction semantics – we give examples in Sec. 5.

Given a set of events, we write $\mathsf{WR}, \mathsf{WW}, \mathsf{RR}, \mathsf{RW}$ for the set of write-read, write-write, read-read and read-write pairs respectively. For example $(w,r) \in \mathsf{WR}$ means that w is a write and r a read. We write $\mathsf{po} \cap \mathsf{WR}$ for the write-read pairs in program order, and $\mathsf{po} \setminus \mathsf{WR}$ for all the pairs in program order except the write-read pairs.

Relations over events: the *program order* po lifts the order in which instructions have been written in the program to the level of events. The program order is a total order over the memory events of a given thread, but does not order events from different threads. Note that the program order unrolls the loops and determines the branches taken.

The *read-from* rf links a read from a register or a memory location to a unique write to the same register or location. The value of the read must be equal to the one of the write. We write rfe (external read-from) when the events related by rf belong to distinct

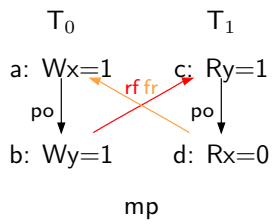
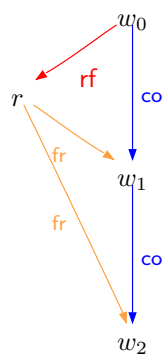


Fig. 4. Message passing pattern

threads, i.e. $(w, r) \in \mathsf{rfe} \triangleq (w, r) \in \mathsf{rf} \wedge \mathrm{proc}(w) \neq \mathrm{proc}(r)$. We write rfi for internal read-from, when the events belong to the same thread.

The *coherence* order co totally orders writes to the same memory location. We write coi (resp. coe) for internal (resp. external) coherence.

We derive the *from-read* fr from the read-from rf and the coherence co, as follows:



That is, a read r is in fr with a write w_1 (resp. w_2) if r reads from a write w_0 such that w_0 is in the coherence order before w_1 (resp. w_2). We write fri (resp. fre) for the internal (resp. external) from-read.

We gather all *communications* in $\mathsf{com} \triangleq \mathsf{co} \cup \mathsf{rf} \cup \mathsf{fr}$. We give a glossary of all the relations which we describe in this section in Tab. II. For each relation we give its notation, its name in English, the directions (i.e. write W or read R) of the source and target of the relation (column "dirns"), where to find it in the text (column "reference"), and an informal prose description. Additionally in the column "nature", we give a taxonomy of our relations: are they fundamental execution relations (e.g. po, rf), architectural relations (e.g. ppo), or derived (e.g. fr, hb)?

Reading notes: we refer to orderings of events w.r.t. several relations. To avoid ambiguity, given a relation r, we say that an event e_1 is *r-before* another event e_2 (or e_1 is an *r-predecessor* of e_1, or e_2 is *r-after* e_1, or e_2 is *r-subsequent*, etc) when $(e_1, e_2) \in \mathsf{r}$.

In the following we present several examples of executions, in the style of [Sarkar et al. 2011]. We depict the events of a given thread vertically to represent the program order, and the communications by arrows labelled with the corresponding relation. Fig. 4 shows a classic *message passing* (mp) example.

This example is a communication pattern involving two memory locations x and y: x is a message, and y a flag to signal to the other thread that it can access the message.

notation	name	nature	dirns	reference	description
po	program order	execution	any, any	§Relations over events	instruction order lifted to events
rf	read-from	execution	WR	§Relations over events	links a write w to a read r taking its value from w
co	coherence	execution	WW	§Relations over events	total order over writes to the same memory location
ppo	preserved program order	architecture	any, any	§Architectures	program order maintained by the architecture
ffence, ff	full fence	architecture	any, any	§Architectures	e.g. sync on Power, dmb and dsb on ARM
lwfence, lwf	lightweight fence	architecture	any, any	§Architectures	e.g. lwsync on Power
cfence	control fence	architecture	any, any	§Architectures	e.g. isync on Power, isb on ARM
fences	fences	architecture	any, any	§Architectures	union of some (depending on the architecture) of the fence relations, e.g. ffence, lwfence, cfence
prop	propagation	architecture	WW	§Architectures	order in which writes propagate, typically enforced by fences
po-loc	program order restricted to the same memory location	derived	any, any	§SC PER LOCATION	$\{(x,y) \mid (x,y) \in \mathsf{po} \wedge \mathrm{addr}(x) = \mathrm{addr}(y)\}$
com	communications	derived	any, any	§Relations over events	$\mathsf{co} \cup \mathsf{rf} \cup \mathsf{fr}$
fr	from-read	derived	RW	§Relations over events	links a read r to a write w' co-after the write w from which r takes its value
hb	happens before	derived	any, any	§NO THIN AIR	$\mathsf{ppo} \cup \mathsf{fences} \cup \mathsf{rfe}$
rdw	read different writes	derived	RR	Fig. 27	two threads; first thread holds a write, second thread holds two reads
detour	detour	derived	WR	Fig. 28	two threads; first thread holds a write, second threads holds a write and a read

Table II. Glossary of relations

T_0 writes the value 1 to memory at location x (see the event a). In program order after a (hence the po arrow between a and b), we have a write of value 1 to memory at location y. T_1 reads from y (see the event c). In the particular execution shown here, this read takes its value from the write b by T_0, hence the rf arrow between b and a. In program order after d, we have a read from location x. In this execution, we suppose that this event d reads from the initial state (not depicted), which by convention sets the values in all memory locations and registers to 0. This is the reason why the read d has the value 0. This initial write to x is, by convention, co-before the write a of x by T_0, hence we have an fr arrow between d and a.

Note that, in the following, even if we do not always depict all of the program order, a program order edge is always implied between each pair of events ordered vertically below a thread id, e.g. T_0.

Convention for naming tests: we refer to tests following the same convention as in [Sarkar et al. 2011]. We roughly have two flavours of names: classical names, that are abbreviations of classical litmus test names appearing in the litterature; and systematic names, that describe the accesses occurring on each thread of a test.

Classical patterns, such as the message passing pattern above, have an abbreviated name: mp stands for "message passing", sb for "store buffering", lb for "load buffering", wrc for "write-to-read causality", rwc for "read-to-write causality".

When a pattern does not have a classical name from the literature, we give it a name that simply describes which accesses occur: for example 2+2w means that the test is made of two threads holding two writes each; w+rw+2w means that we have three threads: a write on a first thread, a read followed by a write on a second thread, and then two writes on a last thread.

Finally when we extend a test (e.g. rwc, "read-to-write causality") with an access (e.g. a write) on an extra thread, we extend the name appropriately: w+rwc (see Fig. 19). We give a glossary of the test names presented in this paper in Tab. III, in the order in which they appear; for each test we give its systematic name, and its classic name (i.e. borrowed from previous works) when there is one.

Note that in every test we choose the locations so that we can form a cycle in the relations of our model: for example, 2+2w has two threads with two writes each, such that the first one accesses e.g. the locations x and y and the second one accesses y and x. This precludes having the first thread accessing x and y and the second one z and y, because we could not link the locations on each thread to form a cycle.

Given a certain pattern such as mp above, we write mp+lwfence+ppo for the same pattern, but where the first thread has a lightweight fence lwfence between the two writes and the second thread maintains its two accesses in order thanks to some *preserved program order* mechanism (ppo, see below). We write mp+lwfences for the mp pattern with two lightweight fences, one on each thread. We sometimes specialise the naming to certain architectures and mechanisms, as in mp+lwsync+addr, where lwsync refers to Power's lightweight fence, and addr denotes an *address dependency* — a particular way of preserving program order on Power.

Architectures are instances of our model. An architecture is a triple of functions $(\mathsf{ppo}, \mathsf{fences}, \mathsf{prop})$, which specifies the *preserved program order* ppo, the *fences* fences and the *propagation order* prop.

The preserved program order gathers the set of pairs of events which are guaranteed not to be reordered w.r.t. the order in which the corresponding instructions have been written. For example on TSO, only write-read pairs can be reordered, so that the preserved program order for TSO is $\mathsf{po} \setminus \mathsf{WR}$. On weaker models such as Power or ARM, the preserved program order merely includes *dependencies*, for example address dependencies, when the address of a memory access is determined by the value read by a preceding load. We detail these notions, and the preserved program order for Power and ARM, in Sec. 6.

The function ppo, given an execution $(\mathbb{E}, \mathsf{po}, \mathsf{co}, \mathsf{rf})$, returns the preserved program order. For example, consider the execution of the message passing example given in Fig. 8. Assume that there is an address dependency between the two reads on T_1. As such a dependency constitutes a preserved program order relation on Power, the ppo function would return the pair (c, d) for this particular execution.

Fences (or *barriers*) are special instructions which prevent certain behaviours. On Power and ARM (see Sec. 6), we distinguish between control fence (which we write cfence), lightweight fence (lwfence) and full fence (ffence). On TSO there is only one fence, called mfence.

In this paper, we use the same names for the fence instructions and the relations that they induce over events. For example, consider the execution of the message pass-

classic	systematic	diagram	description
coXY		Fig. 6	coherence test involving an access of kind X and an access of kind Y; X and Y can be either R (read) or W (write)
lb	rw+rw	Fig. 7	load buffering i.e. two threads each holding a read then a write
mp	ww+rr	Fig. 8	message passing i.e. two threads; first thread holds two writes, second thread holds two reads
wrc	w+rw+rr	Fig. 11	write to read causality i.e. three threads; first thread holds a write, second thread holds a read then a write, third thread holds two reads
isa2	ww+rw+rr	Fig. 12	one of the tests appearing in the Power ISA documentation [ppc 2009] i.e. write to read causality prefixed by a write, meaning that the first thread holds two writes instead of just one as in the wrc case
2+2w	ww+ww	Fig. 13(a)	two threads holding two writes each
	w+rw+2w	Fig. 13(b)	three threads; first thread holds a write, second thread holds a read then a write, third thread holds two writes
sb	wr+wr	Fig. 14	store buffering i.e. two threads each holding a write then a read
rwc	w+rr+wr	Fig. 15	read to write causality three threads; first thread holds a write, second thread holds two reads, third thread holds a write then a read
r	ww+wr	Fig. 16	two threads; first thread holds two writes, second thread holds a write and a read
w+rwc	ww+rr+wr	Fig. 19	read to write causality pattern rwc, prefixed by a write i.e. the first thread holds two writes instead of just one as in the rwc case
iriw	w+rr+w+rr	Fig. 20	independent reads of independent writes i.e. four threads; first thread holds a write, second holds two reads, third holds a write, fourth holds two reads

Table III. Glossary of litmus tests names

ing example given in Fig. 8. Assume that there is a lightweight Power fence lwsync between the two writes a and b on T_0. In this case, we would have $(a, b) \in$ lwsync.[2]

The function fences returns the pairs of events in program order which are separated by a fence, when given an execution. For example, consider the execution of the message passing example given in Fig. 8. Assume that there is a lightweight Power fence lwsync between the two writes on T_0. On Power, the fences function would thus return the pair (a, b) for this particular execution.

[2]Note that if there is a fence "fence" between two events e_1 and e_2 in program order, the pair of events (e_1, e_2) belongs to the eponymous relation "fence" (i.e. $(e_1, e_2) \in$ fence), regardless of whether the particular fence "fence" actually orders these two accesses. For example on Power, the lightweight fence lwsync does not order write-read pairs in program order. Now consider the execution of the store buffering idiom in Fig. 14, and assume that there is an lwsync between the write a and the read b on T_0. In this case, we have $(a.b) \in$ lwsync. However, the pair (a, b) would not be maintained in that order by the barrier, which we model by excluding write-read pairs separated by an lwsync from the propagation order on Power (see Fig. 17 and 18: the propagation order prop contains lwfence, which on Power is defined as lwsync $\setminus$ WR only, not lwsync).

$$
\begin{aligned}
\text{Input data:}\quad & (\mathsf{ppo}, \mathsf{fences}, \mathsf{prop}) \text{ and } (\mathbb{E}, \mathsf{po}, \mathsf{co}, \mathsf{rf}) \\
(\textsc{sc per location})\quad & \mathrm{acyclic}(\mathsf{po\text{-}loc} \cup \mathsf{com}) \text{ with} \\
& \mathsf{po\text{-}loc} \triangleq \{(x,y) \in \mathsf{po} \wedge \mathrm{addr}(x) = \mathrm{addr}(y)\} \\
& \mathsf{fr} \triangleq \{(r, w_1) \mid \exists w_0.(w_0, r) \in \mathsf{rf} \wedge (w_0, w_1) \in \mathsf{co}\} \\
& \mathsf{com} \triangleq \mathsf{co} \cup \mathsf{rf} \cup \mathsf{fr} \\
(\textsc{no thin air})\quad & \mathrm{acyclic}(\mathsf{hb}) \text{ with} \\
& \mathsf{hb} \triangleq \mathsf{ppo} \cup \mathsf{fences} \cup \mathsf{rfe} \\
(\textsc{observation})\quad & \mathrm{irreflexive}(\mathsf{fre}; \mathsf{prop}; \mathsf{hb}^*) \\
(\textsc{propagation})\quad & \mathrm{acyclic}(\mathsf{co} \cup \mathsf{prop})
\end{aligned}
$$

Fig. 5. A model of weak memory

The propagation order constrains the order in which writes are propagated to the memory system. This order is a partial order between writes (not necessarily to the same location), which can be enforced by using fences. For example on Power, two writes in program order separated by an lwsync barrier (see Fig. 8) will be ordered the same way in the propagation order.

We note that the propagation order is distinct from the coherence order co: indeed co only orders writes to the same location, whereas the propagation order can related writes with different locations through the use of fences. However both orders have to be compatible, as expressed by our PROPAGATION axiom, which we explain next (see Fig. 5 and Fig. 13(a)).

The function prop returns the pairs of writes ordered by the propagation order, given an execution. For example, consider the execution of the message passing example given in Fig. 8. Assume that there is a lightweight Power fence lwsync between the two writes on T_0. On Power, the presence of this fence forces the two writes to propagate in the order in which they are written on T_0. The function prop would thus return the pair (a, b) for this particular execution.

4.2. Axioms of our framework

We can now explain the axioms of our model (see Fig. 5). For each example execution that we present in this section, we write in the caption of the corresponding figure whether it is allowed or forbidden by our model.

4.3. SC PER LOCATION

SC PER LOCATION ensures that the communications com cannot contradict po-loc (program order between events relative to the same memory location), i.e. $\mathrm{acyclic}(\mathsf{po\text{-}loc} \cup \mathsf{com})$. This requirement forbids exactly the five patterns (as shown in [Alglave 2010, A.3 p. 184]) given in Fig. 6.

The pattern coWW forces two writes to the same memory location x in program order to be in the same order in the coherence order co. The pattern coRW1 forbids a read from x to read from a po-subsequent write. The pattern coRW2 forbids the read a to read from a write c which is co-after a write b, if b is po-after a. The pattern coWR forbids a read b to read from a write c which is co-before a previous write a in program order. The pattern coRR imposes that if a read b reads from a write a, all subsequent

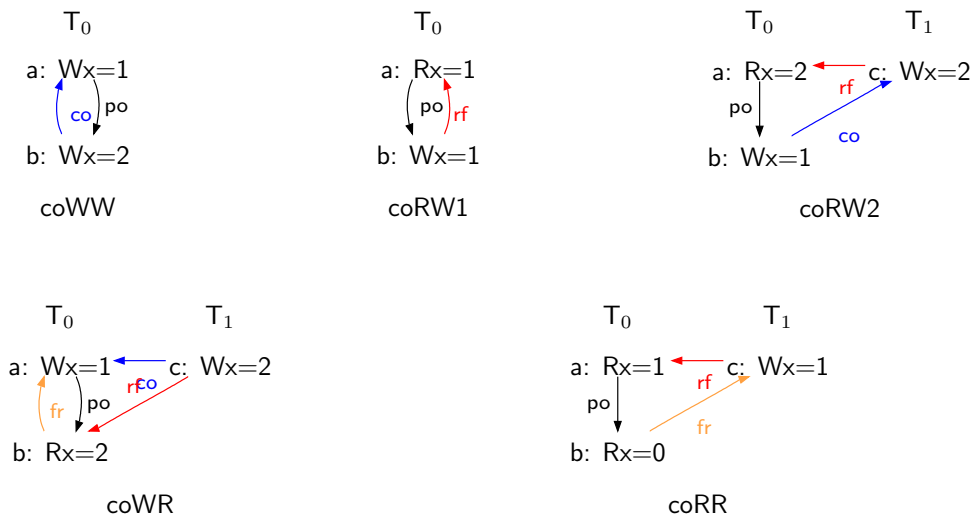


Fig. 6. The five patterns forbidden by SC PER LOCATION

reads in program order from the same location (e.g. the read c) read from a or a co-successor write.

4.4. NO THIN AIR

NO THIN AIR defines a *happens-before* relation, written hb, defined as ppo ∪ fences∪rfe, i.e. an event e_1 happens before another event e_2 if they are in preserved program order, or there is a fence in program order between them, or e_2 reads from e_1.

NO THIN AIR requires the happens-before relation to be acyclic, which prevents values from appearing out of thin air. Consider the load buffering pattern (lb+ppos) in Fig. 7.

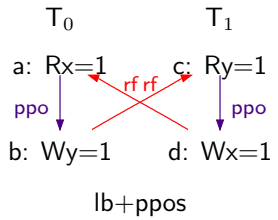


Fig. 7. The load buffering pattern lb with ppo on both sides (forbidden)

T_0 reads from x and writes to y, imposing (for example) an address dependency between the two accesses, so that they cannot be reordered. Similarly T_1 reads from y and (dependently) writes to x. NO THIN AIR ensures that the two threads cannot communicate in a manner that creates a happens-before cycle, with the read from x on T_0 reading from the write to x on T_1, whilst T_1 reads T_0's write to y.

Using the terminology of [Sarkar et al. 2011], we say that a read is *satisfied* when it binds its value; note that the value is not yet irrevocable. It becomes irrevocable

when the read is *committed*. We say that a write is *committed* when it makes its value available to other threads.

Our happens-before relation orders the point in time where a read is satisfied, and the point in time where a write is committed.

The pattern above shows that the ordering due to happens-before applies to any architecture that does not speculate values read from memory (i.e. values written by external threads as opposed to the same thread as the reading thread), nor allows speculative writes (e.g. in a branch) to send their value to other threads.

Indeed, in such a setting, a read such as the read a on T_0 can be satisfied from an external write (e.g. the write d on T_1) only after this external write has made its value available to other threads, and its value is irrevocable.

Our axiom forbids this pattern regardless of the method chosen to maintain the order between the read and the write on each thread: address dependencies on both (lb+addrs), a lightweight fence on the first and an address dependency on the second (lb+lwfence+addr), two full fences (lb+ffences). If however one or both pairs are not maintained, the pattern can happen.

4.5. Observation

OBSERVATION constrains the value of reads. If a write a to x and a write b to y are ordered by the propagation order prop, and if a read c reads from the write of y, then any read d from x which happens after c (i.e. $(c, d) \in$ hb) cannot read from a write that is co-before the write a (i.e. $(d, a) \notin$ fr).

4.5.1. Message passing (mp) A typical instance of this pattern is the message passing pattern (mp+lwfence+ppo) given in Fig. 8.

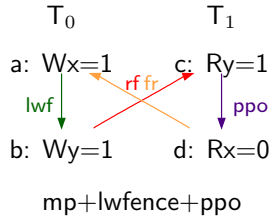


Fig. 8. The message passing pattern mp with lightweight fence and ppo (forbidden)

T_0 writes a message in x, then sets up a flag in y, so that when T_1 sees the flag (via its read from y), it can read the message in x. For this pattern to behave as intended, following the message passing protocol described above, the write to x needs to be propagated to the reading thread before the write to y.

TSO guarantees it, but Power or ARM need at least a lightweight fence (e.g. lwsync on Power) between the writes. We also need to ensure that the two reads on T_1 stay in the order in which they have been written, e.g. with an address dependency.

The protocol would also be ensured with a full fence on the writing thread (mp+ffence+addr), or with two full fences (mp+ffences).

More precisely, our model assumes that a full fence is at least as powerful as a lightweight fence. Thus the behaviours forbidden by the means of a lightweight fence are also forbidden by the means of a full fence. We insist on this point since, as we shall see, our ARM model does not feature any lightweight fence. Thus when reading

an example such as the message passing one in Fig. 8, the "lwf" arrow should be interpreted as any device that has at least the same power as a lightweight fence. In the case of ARM, this means a full fence, i.e. dmb or dsb.

From a micro-architectural standpoint, fences order the propagation of writes to other threads. A very naive implementation can be by locking a shared bus; modern systems feature more sophisticated protocols for ordering write propagation.

By virtue of the fence, the writes to x and y on T_0 should propagate to T_1 in the order in which they are written on T_0. Since the reads c and d on T_1 are ordered by ppo (e.g. an address dependency), they should be satisfied in the order in which they are written on T_1. In the scenario depicted in Fig. 8, T_1 reads the value 1 in y (see the read event c) and the value 0 in x (see the read event d). Thus T_1 observes the write a to x after the write b to y. This contradicts the propagation order of the writes a and b enforced by the fence on T_0.

We note that the Alpha architecture [alpha 2002] allows the pattern mp+fence+addr (a specialisation of Fig. 8). Indeed some implementations feature more than one memory port per processor, e.g. by the means of a split cache [Howells and MacKenney 2013].

Thus in our mp pattern above, the values written on x and y by T_0 could reach T_1 on different ports. As a result, although the address dependency forces the reads c and d to be satisfied in order, the second read may read a stale value of x, while the current value of x is waiting in some queue. One could counter this effect by synchronising memory ports.

4.5.2. Cumulativity To explain a certain number of the following patterns, we need to introduce the concept of *cumulativity*.

We consider that a fence has a cumulative effect when it ensures a propagation order not only between writes on the fencing thread (i.e. the thread executing the fence), but also between certain writes coming from threads other than the fencing thread.

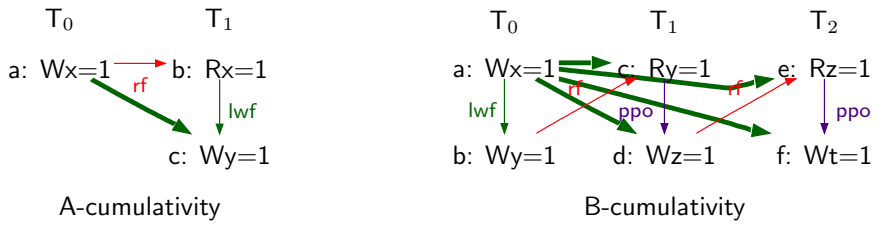


Fig. 9. Cumulativity of fences

A-cumulativity More precisely, consider a situation as shown on the left of Fig. 9. We have a write a to x on T_0, which is read by the read b of x on T_1. On T_1 still, we have an access e in program order after b. This event e could be either a read or a write event; it is a write c in Fig. 9. Note that b and e are separated by a fence.

We say that the fence between b and e on T_1 is *A-cumulative* when it imposes that the read b is satisfied before e is either committed (if it is a write) or satisfied (if it is a read). Note that for b to be satisfied, the write a on T_0 from which b reads must have been propagated to T_1, which enforces an ordering between a and e. We display this ordering in Fig. 9 by the mean of a thicker arrow from a to c.

The vendors documentations describe A-cumulativity as follows. On Power, quoting [ppc 2009, Book II, Sec. 1.7.1]: "[the group] A [of T_1] includes all [...] accesses by any [...] processor [...] [e.g. T_0] that have been performed with respect to [T_1] before the memory barrier is created." We interpret "performed with respect to [T_1]" as a write (e.g. the write a on the left of Fig. 9) having been propagated to T_1 and read by T_1 such that the read is po-before the barrier.

On ARM, quoting [ARM 2009, Sec. 4]: "[the] group A [of T_1] contains: All explicit memory accesses [...] from observers [...] [e.g. T_0] that are observed by [T_1] before the dmb instruction." We interpret "observed by [T_1]" on ARM as we interpret "performed with respect to [T_1]" on Power (see paragraph above).

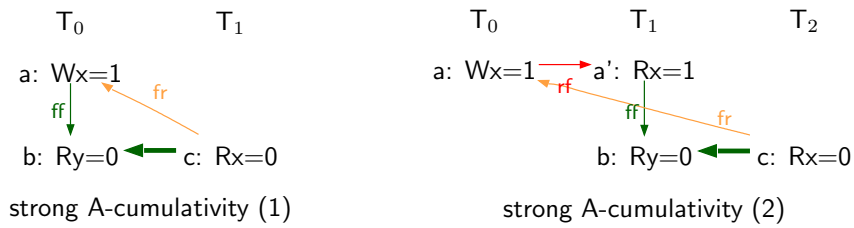


Fig. 10. Strong A-cumulativity of fences

Strong A-cumulativity Interestingly, the ARM documentation does not stop there, and includes in the group A "[a]ll loads [...] from observers [...] [(e.g. T_1)] that have been observed by any given observer [e.g. T_0], [...] before [T_0] has performed a memory access that is a member of group A."

Consider the situation on the left of Fig. 10. We have the read c on T_1 which reads from the intial state for x, thus is fr-before the write a on T_0. The write a is po-before a read b, such that a and b are separated by a full fence. In that case, we count the read c as part of the group A of T_0. This enforces a (strong) A-cumulativity ordering from c to b, which we depict with a thicker arrow.

Similarly, consider the situation on the right of Fig. 10. The only difference with the left of the figure is that we have one more indirection between the read c of x and the read b of y, via the rf between a and a'. In that case too we count the read c as part of the group A of T_1. This enforces a (strong) A-cumulativity ordering from c to b, which we depict with a thicker arrow.

We take strong A-cumulativity into account only for full fences (i.e. sync on Power and dmb and dsb on ARM). We reflect this in Fig. 18, in the second disjunct of the definition of the prop relation ($\mathsf{com}^*; \mathsf{prop\text{-}base}^*; \mathsf{ffence}; \mathsf{hb}^*$).

B-cumulativity Consider now the situation shown on the right of Fig. 9. On T_0, we have an access e (which could be either a read or a write event) in program order before a write b to y. Note that e and b are separated by a fence. On T_1 we have a read c from y, which reads the value written by b.

We say that the fence between e and b on T_0 is *B-cumulative* when it imposes that e is either committed (if it is a write) or satisfied (if it is a read) before b is committed. Note that the read c on T_1 can be satisfied only after the write b is committed and propagated to T_1, which enforces an ordering from e to c. We display this ordering in Fig. 9 by the mean of a thicker arrow from a to c.

This B-cumulativity ordering extends to all the events that happen after c (i.e. are hb-after c), such as the write d on T_1, the read e on T_2 and the write f on T_2. In Fig. 9, we display all these orderings via thicker arrows.

The vendors documentations describe B-cumulativity as follows. On Power, quoting [ppc 2009, Book II, Sec. 1.7.1]: "[the group] B [of T_0] includes all [...] accesses by any [...] processor [...] that are performed after a load instruction [such as the read c on the right of Fig. 9] executed by [T_0] [...] has returned the value stored by a store that is in B [such as the write b on the right of Fig. 9]." On the right of Fig. 9, this includes for example the write d. Then the write d is itself in the group B, and observed by the read e, which makes the write f part of the group B as well.

On ARM, quoting [ARM 2009, Sec. 4]: "[the] group B [of T_0] contains [a]ll [...] accesses [...] by [T_0] that occur in program order after the dmb instruction. [...]" On the right of Fig. 9, this mean the write b.

Furthermore, ARM's group B contains "[a]ll [...] accesses [...] by any given observer [e.g. T_1] [...] that can only occur after [T_1] has observed a store that is a member of group B." We interpret this bit as we did for Power's group B.

4.5.3. Write to read causality (wrc) This pattern (wrc+lwfence+ppo, given in Fig. 11) illustrates the A-cumulativity of the lightweight fence on T_1, namely the rfe; fences fragment of the definition illustrated in Fig. 9 above.

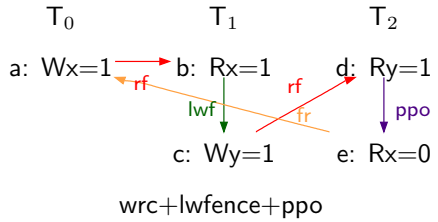


Fig. 11. The write-to-read causality pattern wrc with lightweight fence and ppo (forbidden)

There are now two writing threads, T_0 writing to x and T_1 writing to y after reading the write of T_0. TSO still enforces this pattern without any help. But on Power and ARM we need to place (at least) a lightweight fence on T_1 between the read of T_0 (the read b from x) and the write c to y. The barrier will force the write of x to propagate before the write of y to the T_2 even if the writes are not on the same thread.

A possible implementation of A-cumulative fences, discussed in [Gharachorloo et al. , Sec. 6], is to have the fences not only wait for the previous read (a in Fig. 11) to be satisfied, but also for all stale values for x to be eradicated from the system, e.g. the value 0 read by e on T_2.

4.5.4. Power ISA2 (isa2) This pattern (isa2+lwfence+ppos, given in Fig. 12), distributes the message passing pattern over three threads like wrc+lwfence+ppo, but keeping the writes to x and y on the same thread.

Once again TSO guarantees it without any help, but Power and ARM need a lightweight fence between the writes, and (for example) an address or data dependency between each pair of reads (i.e. on both T_1 and T_2).

Thus on Power and ARM, the pattern isa2+lwfence+ppos illustrates the B-cumulativity of the lightweight fence on T_0, namely the fences; hb* fragment of the definition of cumul illustrated in Fig. 9.

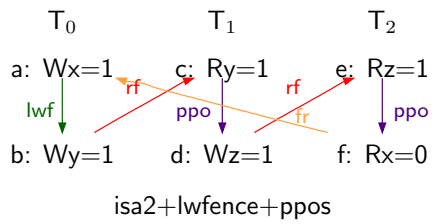


Fig. 12. The pattern isa2 with lighweight fence and ppo twice (forbidden)

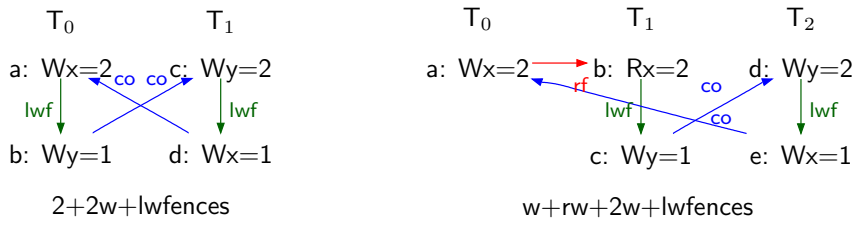
4.6. Propagation

PROPAGATION constrains the order in which writes to memory are propagated to the other threads, so that it does not contradict the coherence order, i.e. $\mathrm{acyclic}(\mathsf{co} \cup \mathsf{prop})$.

On Power and ARM, lightweight fences sometimes constrain the propagation of writes, as we have seen in the cases of mp (see Fig. 8), wrc (see Fig. 11) or isa2 (see Fig. 12). They can also, in combination with the coherence order, create new ordering constraints.

The 2+2w+lwsync pattern (given in Fig. 13(a)) is for us the archetypal illustration of coherence order and fences interacting to yield new ordering constraints. It came as a surprise when we proposed it to our IBM contact, as he wanted the lighweight fence to be as lightweight as possible, for the sake of efficiency. However the pattern is acknowledged to be forbidden. By contrast and as we shall see below, other patterns (such as the r pattern in Fig. 16) that mix the co communication with fr require full fences to be forbidden.

The w+rw+2w+lwfences pattern in Fig. 13(b) distributes 2+2w+lwfences over three threads. This pattern is to 2+2w+lwfences what wrc is to mp. Thus, just as in the case of mp and wrc, the lightweight fence plays an A-cumulative role, which ensures that the two patterns 2+2w and w+rw+2w respond to the fence in the same way.



(a) The pattern 2+2w

(b) The pattern w+rw+2w

Fig. 13. Two similar patterns with lightweight fences (forbidden)

On TSO, every relation contributes to the propagation order (see Fig. 21), except the write-read pairs in program order, which need to be fenced with a full fence (mfence on TSO). Consider the store buffering (sb+ffences) pattern given in Fig. 14.

We need a full fence on each side to prevent the reads b and d from reading the initial state.

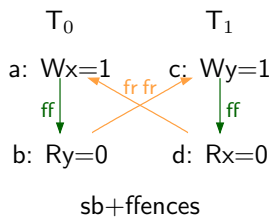


Fig. 14. The store buffering pattern sb with full fences (forbidden)

The pattern sb without fences being allowed, even on x86-TSO, is perhaps one of the most well-known examples of a relaxed behaviour. It can be explained by writes being first placed into a thread-local store buffer, and then carried over asynchronously to the memory system. In that context, the effect of a (full) fence can be described as flushing the store buffer. Of course, on architectures more relaxed than TSO, the full fence has more work to do, e.g. cumulativity duties (as illustrated by the iriw example given in Fig. 20).

On Power, the lightweight fence lwsync does not order write-read pairs in program order. Hence in particular it is not enough to prevent the sb pattern; one needs to use a full fence on each side. The sb idiom, and the following rwc idiom, are instances of the strong A-cumulativity of full fences.

The read-to-write causality pattern rwc+ffences (see Fig. 15) distributes the sb pattern over three threads with a read b from x between the write a of x and the read c of y. It is to sb what wrc is to mp, thus responds to fences in the same way as sb. Hence it needs two full fences too. Indeed, a full fence is required to order the write a

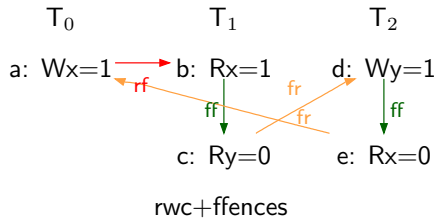


Fig. 15. The read-to-write causality pattern rwc with full fences (forbidden)

and the read c, as lightweight fences do not apply from writes to reads. The full fence on T_1 provides such an order, not only on T_1, but also by (strong) A-cumulativity from the write a on T_1 to the read c on T_1, as the read b on T_1 that reads from the write a, po-precedes the fence.

The last two patterns, r+ffences and s+lwfence+ppo (Fig. 16) illustrate the complexity of combining coherence order and fences. In both patterns, the thread T_0 writes to x (event a) and then to y (event b). In the former pattern r+ffence, T_1 writes to y and reads from x. A full fence on each thread forces the write a to x to propagate to T_1 before the write b to y. Thus if the write b is co-before the write c on T_1, the read d of x on T_1 cannot read from a write that is co-before the write a. By constrast, in the latter

pattern s+lwfence+ppo, T_1 reads from y, reading the value stored by the write b, and then writes to x. A lightweight fence on T_0 forces the write a to x to propagate to T_1 before the write b to y. Thus, as T_1 sees the write b by reading its value (read c) and as the write d is forced to occur by a dependency (ppo) after the read c, that write d cannot co-precede the write a.

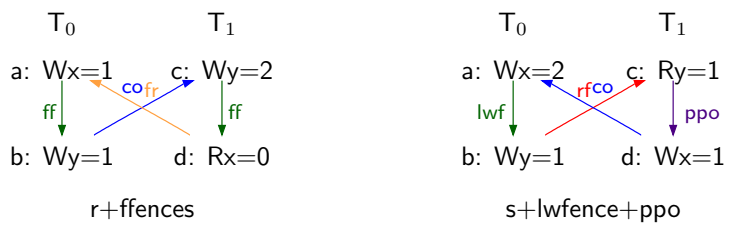


Fig. 16. The patterns r with full fences and s with lightweight fence and ppo (both forbidden)

Following the architect's intent, inserting a lightweight fence lwsync between the writes a and b does not suffice to forbid the r pattern on Power. It comes in sharp contrast with, for instance, the mp pattern (see Fig. 8) and the s pattern, where a lightweight fence suffices. Thus the interaction between (lightweight) fence order and coherence order is quite subtle, as it does forbid 2+2w and s, but not r. For completeness, we note that we did not observe the r+lwsync+sync pattern on Power hardware.

4.7. Fences and propagation on Power and ARM

Recall that architectures provides special instructions called fences to prevent some weak behaviours.

We summarise the fences and propagation order for Power and ARM in Fig. 17 and 18. Note that the control fences (isync for Power and isb for ARM) do not contribute to the propagation order. They contribute to the definition of preserved program order, as explained in Sec. 5 and 6.

Power	$\mathsf{ffence} \triangleq \mathsf{sync}$	$\mathsf{lwfence} \triangleq \mathsf{lwsync} \setminus \mathsf{WR}$	$\mathsf{cfence} \triangleq \mathsf{isync}$
ARM	$\mathsf{ffence} \triangleq \mathsf{dmb} \cup \mathsf{dsb}$	$\mathsf{lwfence} \triangleq \emptyset$	$\mathsf{cfence} \triangleq \mathsf{isb}$

Fig. 17. Fences for Power and ARM

$$\mathsf{hb} \triangleq \mathsf{ppo} \cup \mathsf{fences} \cup \mathsf{rfe} \qquad \mathsf{fences} \triangleq \mathsf{lwfence} \cup \mathsf{ffence} \qquad \mathsf{A\text{-}cumul} \triangleq \mathsf{rfe}; \mathsf{fences}$$

$$\mathsf{prop\text{-}base} \triangleq (\mathsf{fences} \cup \mathsf{A\text{-}cumul}); \mathsf{hb}^*$$

$$\mathsf{prop} \triangleq (\mathsf{prop\text{-}base} \cap \mathsf{WW}) \cup (\mathsf{com}^*; \mathsf{prop\text{-}base}^*; \mathsf{ffence}; \mathsf{hb}^*)$$

Fig. 18. Propagation order for Power and ARM

Fence placement is the problem of automatically placing fences between events to prevent undesired behaviours. For example, the message passing pattern mp in Fig. 8 can be prevented by placing fences between events a and b in T_0 and events c and d in T_1 in order to create a forbidden cycle.

This problem has already been studied in the past, but for TSO and its siblings PSO and RMO (see e.g. [Kuperstein et al. 2010; Bouajjani et al. 2011; Liu et al. 2012; Bouajjani et al. 2013]), or for previous versions of the Power model (see e.g. [Alglave and Maranget 2011]).

We emphasise how easy fence placement becomes, in the light of our new model. We can read them off the definitions of the axioms and of the propagation order. Placing fences essentially amounts to counting the number of communications (i.e. the relations in com) involved in the behaviour that we want to forbid.

To forbid a behaviour that involves only read-from (rf) communications, or only *one* from-read (fr) and otherwise rf communications, one can resort to the OBSERVATION axiom, using the prop-base part of the propagation order. Namely to forbid a behaviour of the mp (see Fig. 8), wrc (see Fig. 11) or isa2 (see Fig. 12) type, one needs a lightweight fence on the first thread, and some preserved program order mechanisms on the remaining threads.

If coherence (co) and read-from (rf) are the only communications involved, one can resort to the PROPAGATION axiom, using the prop-base part of the propagation order. Namely to forbid a behaviour of the 2+2w (see Fig. 13(a)) or w+rw+2w (see Fig. 13(b)) type, one needs only lightweight fences on each thread.

If more than one from-read (fr) occurs, or if the behaviour involves both coherence (co) and from-read (fr) communications, one needs to resort to the part of the propagation order that involves the full fence (indeed that is the only part of the propagation order that involves com). Namely to forbid behaviours of the type sb (see Fig. 14), rwc (see Fig. 15) or r (see Fig. 16), one needs to use full fences on each thread.

Power's eieio and ARM's dmb.st and dsb.st: we make two additional remarks, not reflected in the figures, about write-write barriers, namely the eieio barrier on Power, and the dmb.st and dsb.st barriers on ARM.

On Power, the eieio barrier only maintains write-write pairs, as far as ordinary (*Memory Coherence Required*) memory is concerned [ppc 2009, p. 702]. We demonstrate (see http://diy.inria.fr/cats/power-eieio/) that eieio cannot be a full barrier, as this option is invalidated by hardware. For example the following w+rwc+eieio+addr+sync pattern (see Fig. 19) would be forbidden if eieio was a full barrier, yet is observed on Power 6 and Power 7 machines. Indeed this pattern involves two from-reads (fr), in which case our axioms require us to resort to the full fence part of the propagation order. Thus we take eieio to be a lightweight barrier that maintains only write-write pairs.

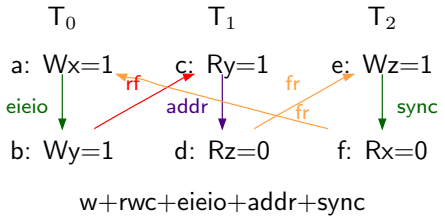


Fig. 19. The pattern w+rwc with eieio, address dependency and full fence (allowed)

The ARM architecture features the dmb and dsb fences. ARM documentation [ARM 2009] forbids iriw+dmb (see Fig. 20). Hence dmb is a full fence, as this behaviour involves two from-reads (fr), and thus needs full fences to be forbidden.

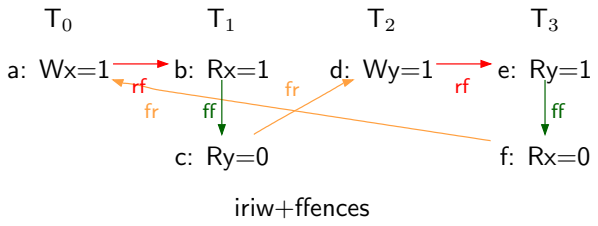


Fig. 20. The independent reads from independent writes pattern iriw with full fences (forbidden)

The ARM documentation also specifies dsb to behave at least as strongly as dmb [arm 2008, p. A3-49]: "*A dsb behaves as a dmb with the same arguments, and also has [. . .] additional properties [. . .]*". Thus we take both dmb and dsb to be full fences. We remark that this observation and our experiments (see Sec. 8) concern memory litmus tests only; we do not know whether dmb and dsb differ (or indeed should differ) when out-of-memory communication (e.g. interrupts) comes into play.

Finally, the ARM architecture specifies that, when suffixed by .st, ARM fences operate on write-to-write pairs only. It remains to be decided whether the resulting dmb.st and dsb.st fences are lightweight fences or not. In that aspect ARM documentation does not help much, nor do experiments on hardware, as all our experiments are compatible with both alternatives (see http://diy.inria.fr/cats/arm-st-fences). Thus we choose simplicity and consider that .st fences behave as their unsuffixed counterparts, but limited to write-to-write pairs. In other words, we assume that the ARM architecture makes no provision for some kind of lightweight fence, unlike Power which provides with lwsync.

Formally, to account for .st fences being full fences limited to write-to-write pairs, we would proceed as follows. In Fig. 17 for ARM, we would extend the definition of ffence to $\mathsf{dmb} \cup \mathsf{dsb} \cup (\mathsf{dmb.st} \cap \mathsf{WW}) \cup (\mathsf{dsb.st} \cap \mathsf{WW})$. Should the option of .st fences being lightweight fences be prefered (thereby allowing for instance the pattern w+rwc+dmb.st+addr+dmb of Fig. 19), one would instead define lwfence as $(\mathsf{dmb.st} \cap \mathsf{WW}) \cup (\mathsf{dsb.st} \cap \mathsf{WW})$ in Fig. 17 for ARM.

4.8. Some instances of our framework

We now explain how we instantiate our framework to produce the following models: Lamport's Sequential Consistency [Lamport 1979] (SC), Total Store Order (TSO), used in Sparc [sparc 1994] and x86's [Owens et al. 2009] architectures, and C++ R-A, i.e. C++ restricted to the use of release-acquire atomics [Batty et al. 2011; Batty et al. 2013].

SC and x86-TSO are described in our terms on top of Fig. 21, which gives their preserved program order, fences and propagation order.

We show that these instances of our model correspond to SC and TSO:

LEMMA 4.1. *Our model of SC as given in Fig. 21 is equivalent to Lamport's SC [Lamport 1979]. Our model of TSO as given in Fig. 21 is equivalent to Sparc TSO [sparc 1994].*

SC: $\mathsf{ppo} \triangleq \mathsf{po}$ $\quad \mathsf{ffence} \triangleq \emptyset$ $\quad \mathsf{lwfence} \triangleq \emptyset$ $\quad \mathsf{fences} \triangleq \mathsf{ffence} \cup \mathsf{lwfence}$
$\mathsf{prop} \triangleq \mathsf{ppo} \cup \mathsf{fences} \cup \mathsf{rf} \cup \mathsf{fr}$

TSO: $\mathsf{ppo} \triangleq \mathsf{po} \setminus \mathsf{WR}$ $\quad \mathsf{ffence} \triangleq \mathsf{mfence}$ $\quad \mathsf{lwfence} \triangleq \emptyset$
$\mathsf{fences} \triangleq \mathsf{ffence} \cup \mathsf{lwfence}$ $\quad \mathsf{prop} \triangleq \mathsf{ppo} \cup \mathsf{fences} \cup \mathsf{rfe} \cup \mathsf{fr}$

C++ R-A $\approx$: $\mathsf{ppo} \triangleq \mathsf{sb}$ (see [Batty et al. 2011]) $\quad \mathsf{fences} \triangleq \emptyset$ $\quad \mathsf{prop} \triangleq \mathsf{hb}$

(we write $\approx$ to indicate that our definition has a discrepancy with the standard)

Fig. 21. Definitions of SC, TSO and C++ R-A

PROOF. *An execution* $(\mathbb{E}, \mathsf{po}, \mathsf{co}, \mathsf{rf})$ *is valid on SC (resp. TSO) iff* $\mathrm{acyclic}(\mathsf{po} \cup \mathsf{com})$ *(resp.* $\mathrm{acyclic}(\mathsf{ppo} \cup \mathsf{co} \cup \mathsf{rfe} \cup \mathsf{fr} \cup \mathsf{fences})$*) (all relations defined as in Fig. 21) by [Alglave 2012, Def. 21, p. 203] (resp. [Alglave 2012, Def. 23, p. 204].)* □

C++ R-A, i.e. C++ restricted to the use of release-acquire atomics, appears at the bottom of Fig. 21, in a slightly stronger form than the current standard prescribes, as detailed below.

We take the *sequenced before* relation sb of [Batty et al. 2011] to be the preserved program order, and the fences to be empty. The happens-before relation $\mathsf{hb} \triangleq \mathsf{sb} \cup \mathsf{rf}$ is the only relation contributing to propagation, i.e. $\mathsf{prop} = \mathsf{hb}^+$. We take the modification order mo of [Batty et al. 2011] to be our coherence order.

The work of [Batty et al. 2013] shows that C++ R-A is defined by three axioms: ACYCLICITY, i.e. $\mathrm{acyclic}(\mathsf{hb})$ (which immediately corresponds to our NO THIN AIR axiom); COWR, i.e. $\forall r. \neg(\exists w_1, w_2.\ (w_1, w_2) \in \wedge(w_1, r) \in \mathsf{rf} \wedge (w_2, r) \in \mathsf{hb}^+)$ (which corresponds to our OBSERVATION axiom by definition of fr and since $\mathsf{prop} = \mathsf{hb}^+$ here); and HBVSMO, i.e. $\mathrm{irreflexive}(\mathsf{hb}^+; \mathsf{mo})$. Our SC PER LOCATION is implemented by HBVSMO for the coWW case, and the eponymous axioms for the coWR, coRW, coRR cases.

Thus C++ R-A corresponds to our version, except for the HBVSMO axiom, which requires the irreflexivity of $\mathsf{hb}^+; \mathsf{mo}$, whereas we require its acyclicity via the axiom PROPAGATION. To adapt our framework to C++ R-A, one simply needs to weaken the PROPAGATION axiom to $\mathrm{irreflexive}(\mathsf{prop}; \mathsf{co})$.

4.9. A note on the genericity of our framework

We remark that our framework is, as of today, not as generic as it could be, for several reasons that we explain below.

Types of events For a given event (e.g. read, write or barrier), we handle only one type of this event. For example we can express C++ when all the reads are acquire and all the writes are release, but nothing more. As of today, we could not express a model where some writes can be relaxed writes (in the C++ sense), and others can be release writes. To embrace models with several types of accesses, e.g. C++ in its entirety, we would need to extend the expressivity of our framework. We hope to investigate this extension in future work.

Choice of axioms We note that our axiom SC PER LOCATION might be perceived as too strong, as it forbids *load-load hazard* behaviours (see coRR in Fig. 6). This pattern was indeed officially allowed by Sparc RMO [sparc 1994] and pre-Power 4 machines [tdf 2002].

Similarly, the NO THIN AIR axiom might be perceived as controversial, as several software models, such as Java or C++, allow certain lb patterns (see Fig. 7).

We also have already discussed, in the section just above, how the PROPAGATION axiom needs to be weakened to reflect C++ R-A accurately.

Yet we feel that this is much less of an issue than having only one type of events. Indeed one can very simply disable the NO THIN AIR check, or restrict the SC PER LOCATION check so that it allows load-load hazard (see for example [Alglave 2012, Sec. 5.1.2]), or weaken the PROPAGATION axiom (as we do above).

Rather than axioms set in stone, that we would declare as the absolute truth, we present here some basic bricks from which one can build a model at will. Moreover, our new herd simulator (see Sec. 8.3) allows the user to very simply and quickly modify the axioms of his model, and re-run his tests immediately, without to have to dive into the code of the simulator. This makes the cost of experimenting with several different variants of a model, or fine-tuning a model, much less high. We give a flavour of such experimentations and fine-tuning in our discussion about the ARM model (see for example Tab. V.)

5. INSTRUCTION SEMANTICS

In this section, we specialise the discussion to Power, to make the reading easier. Before presenting the preserved program order for Power, given in Fig. 25, we need to define *dependencies*. We borrow the names and notations of [Sarkar et al. 2011] for more consistency.

To define dependencies formally, we need to introduce some more possible actions for our events. In addition to the read and write events relative to memory locations, we can now have:

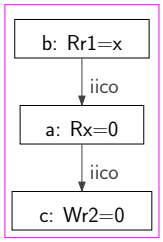
— read and write events relative to *registers*, e.g. for the register r we can have a read of the value 0, noted Rr=0, or a write of the value 1, noted Wr=1;
— branching events, which represent the branching decisions being made;
— fence events, e.g. lwfence.

Note that in general a single instruction can involve several accesses, for example, to registers. Events coming from the same instruction can be related by the relation iico (intra-instruction causality order). We give examples of such a situation below.

5.1. Semantics of instructions

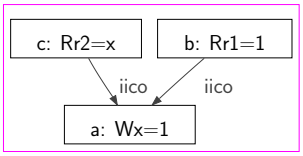
We now give examples of the semantics of instructions. We do not intend to be exhaustive, but rather to give the reader enough understanding of the memory, register, branching and fence events that an instruction can generate, so that we can define dependencies w.r.t. these events in due course. We use Power assembly syntax [ppc 2009].

Here is the semantics of a load "lwz r2,0(r1)" with r1 holding the address x, assuming that x contains the value 0. The instruction reads the content of the register r1, and finds there the memory address x; then (following iico) it reads x, and finds there the value 0. Finally, it writes this value into register r2:



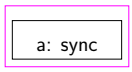
load

Here is the semantics of a store “stw r1,0(r2)” with r1 holding the value 1 and r2 holding the address x. The instruction reads the content of the register r2, and finds there the memory address x. In parallel, it reads the content of the register r1, and finds there the value 1. After (in iico) these two events, it writes the value 1 into memory address x:



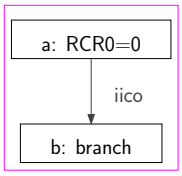
store

Here is the semantics of a “sync” fence, simply an eponymous event:



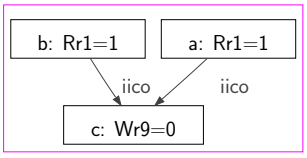
sync fence

Here is the semantics for a branch “bne L0”, which branches to the label L0 if the value of a special register (the Power ISA specifies it to be register CR0 [ppc 2009, Chap. 2, p. 35]) is not equal to 0 (“bne” stands for “branch if not equal”). Thus the instruction reads the content of CR0, and emits a branching event. Note that it emits the branching event regardless of the value of CR0, just to signify that a branching decision has been made:



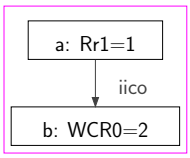
branching

Here is the semantics for a “xor r9,r1,r1”, which takes the bitwise xor of the value in r1 with itself and puts the result into the register r9. The instruction reads (twice) the value in the register r1, takes their xor, and writes the result (necessarily 0) in r9:



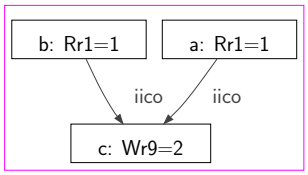
xor

Here is the semantics for a comparison “cmpwi r1, 1”, which compares the value in register r1 with 1. The instruction writes the result of the comparison (2 encodes equality) into the special register CR0 (the same that is used by branches to make branching decisions). Thus the instruction reads the content of r1, then writes to CR0 accordingly:



comparison

Here is the semantics for an addition "add r9,r1,r1", which reads the value in register r1 (twice) and writes the sum into register r9:



addition

5.2. Dependencies

We can now define dependencies formally, in terms of the events generated by the instructions involved in implementing a dependency. We borrow the textual explanations from [Sarkar et al. 2011].

In Fig. 22, we give the definitions of address (addr), data (data), control (ctrl) and control+cfence (ctrl+cfence) dependencies. Below we detail and illustrate what they mean.

$\mathsf{dd\text{-}reg} = (\mathsf{rf\text{-}reg} \cup \mathsf{iico})^+$	Data dependency over registers
$\mathsf{addr} = \mathsf{dd\text{-}reg} \cap \mathsf{RM}$	The last rf-reg is to the address entry port of the target instruction.
$\mathsf{data} = \mathsf{dd\text{-}reg} \cap \mathsf{RW}$	The last rf-reg is to the value entry port of the target store instruction.
$\mathsf{ctrl} = (\mathsf{dd\text{-}reg} \cap \mathsf{RB}); \mathsf{po}$	On Power or ARM, control dependencies targetting a read do not belong to ppo.
$\mathsf{ctrl{+}cfence} = (\mathsf{dd\text{-}reg} \cap \mathsf{RB}); \mathsf{cfence}$	On Power or ARM, branches followed by a control fence (isync on Power, isb on ARM) targetting a read belong to ppo.

Fig. 22. Definitions of dependency relations

In Fig. 22, we use the notations that we have defined before (see Sec. 4 for sets of events. We write "M" for the set of all memory events, so that for example RM is the set of pairs (r, e) where r is a read and e any memory event (i.e. a write or a read). We write "B" for the set of branch events (coming from a branch instruction); thus RB is the set of pairs (r, b) where r is a read and b a branch event.

Each definition uses the relation dd-reg, defined as $(\mathsf{rf\text{-}reg} \cup \mathsf{iico})^+$. This relation dd-reg builds chains of read-from through registers (rf-reg) and intra-instruction causality order (iico). Thus it is a special kind of data dependency, over register accesses only.

We find that the formal definitions in Fig. 22 make quite evident that all these dependencies (addr, data, ctrl and ctrl+cfence) all correspond to data dependencies over

registers (dd-reg), starting with a read. The key difference between each of these dependencies is the kind of events that they target: whilst addr targets any kind of memory event, data only targets writes. A ctrl dependency only targets branch events; a ctrl+cfence also targets only branch events, but requires a control fence cfence to immediately follow the branch event.

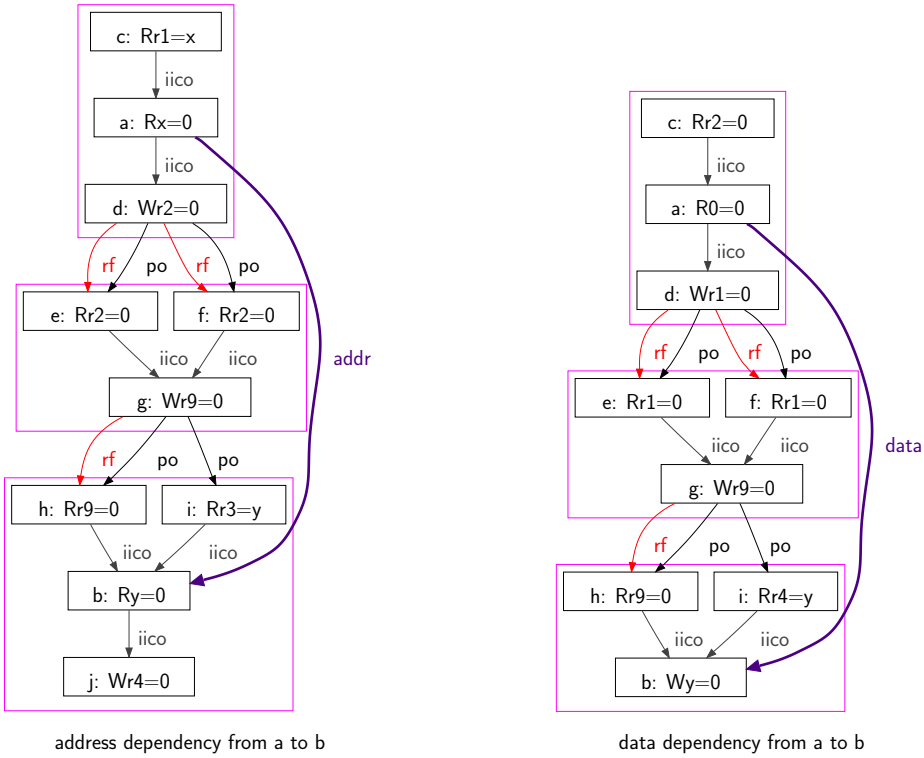


Fig. 23. Data-flow dependencies

5.2.1. Address dependencies Address dependencies are gathered in the addr relation. There is an address dependency from a memory read r to a po-subsequent memory event e (either read or write) if there is a data flow path (i.e. a dd-reg relation) from r to the address of e through registers and arithmetic or logical operations (but not through memory accesses). Note that this notion also includes *false dependencies*, e.g. when xor'ing a value with itself and using the result of the xor in an address calculation. For example, in Power (on the left) or ARM assembly (on the right), the following excerpt

```
(1) lwz r2,0(r1)          ldr r2,[r1]
(2) xor r9,r2,r2          eor r9,r2,r2
(3) lwzx r4,r9,r3         ldr r4,[r9,r3]
```

ensures that the load at line (3) cannot be reordered with the load at line (1), despite the result of the xor at line (2) being always 0.

Graphically (see the left diagram of Fig. 23), the read a from address x is related by addr to the read b from address y because there is a path of rf and iico (through register events) between them. Notice that the last rf is to the index register (here r9) of the load from y instruction.

5.2.2. Data dependencies Data dependencies are gathered in the data relation. There is a data dependency from a memory read r to a po-subsequent memory write w if there is a data flow path (i.e. a dd-reg relation) from r to the value written by w through registers and arithmetic or logical operations (but not through memory accesses). This also includes false dependencies, as described above. For example

```
(1) lwz r2,0(r1)        ldr r2,[r1]
(2) xor r9,r2,r2        eor r9,r2,r2
(3) stw r9,0(r4)        str r9,[r4]
```

ensures that the store at line (3) cannot be reordered with the load at line (1), despite the result of the xor at line (2) being always 0.

Graphically (see the right diagram of Fig. 23), the read a from address x is related by data to the write b to address y because there is a path of rf and iico (through registers) between them, the last rf being to the value entry port of the store.

Remark: Our semantics does not account for conditional execution in the ARM sense (see [arm 2008, Sec. A8.3.8 "Conditional execution"] and [ARM 2009, Sec. 6.2.1.2]). Informally, most instructions can be executed or not, depending on condition flags. It is unclear how to handle them in full generality, both as target of dependencies (conditional load and conditional store instructions); or in the middle of a dependency chain (e.g. conditional move). In the target case, a dependency reaching a conditional memory instruction through its condition flag could act as a control dependency. In the middle case, the conditional move could contribute to the data flow path that defines address and data dependencies. We emphasise that we have not tested these hypotheses.

5.2.3. Control dependencies Control dependencies are gathered in the ctrl relation. There is a control dependency from a memory read r to a po-subsequent memory write w if there is a data flow path (i.e. a dd-reg relation) from r to the test of a conditional branch that precedes w in po. For example

```
(1) lwz r2,0(r1)        ldr r2,[r1]
(2) cmpwi r2,0          cmp r2,#0
(3) bne L0              bne L0
(4) stw r3,0(r4)        str r3,[r4]
(5) L0:                 L0:
```

ensures that the store at line (4) cannot be reordered with the load at line (1). We note that there will still be a control dependency from the load to the store, even if the label immediately follows the branch, i.e. the label L0 is placed between the conditional branch instruction at line (3) and the store.

Graphically (see the left diagram[3] of Fig. 24), the read a from address x is related by ctrl to the write b to address y because there is a path of rf and iico (through registers) between a and a branch event (h in that case) po-before b (some po edges are omitted for clarity).

Such a data flow path between two memory reads is not enough to order them in general. To order two memory reads, one needs to place a *control fence* cfence (isync on Power, isb on ARM, as shown on top of Fig. 25) after the branch, as described below.

5.2.4. Control+cfence dependencies Control+cfence dependencies are gathered in the ctrl+cfence relation. There is such a dependency from a memory read r_1 to a po-subsequent memory read r_2 if there is a data flow path (i.e. a dd-reg relation) from r_1 to the test of a conditional branch that po-precedes a control fence, the fence itself preceding r_2 in po. For example

[3]The diagram depicts the situation for Power; ARM status flags are handled differently.

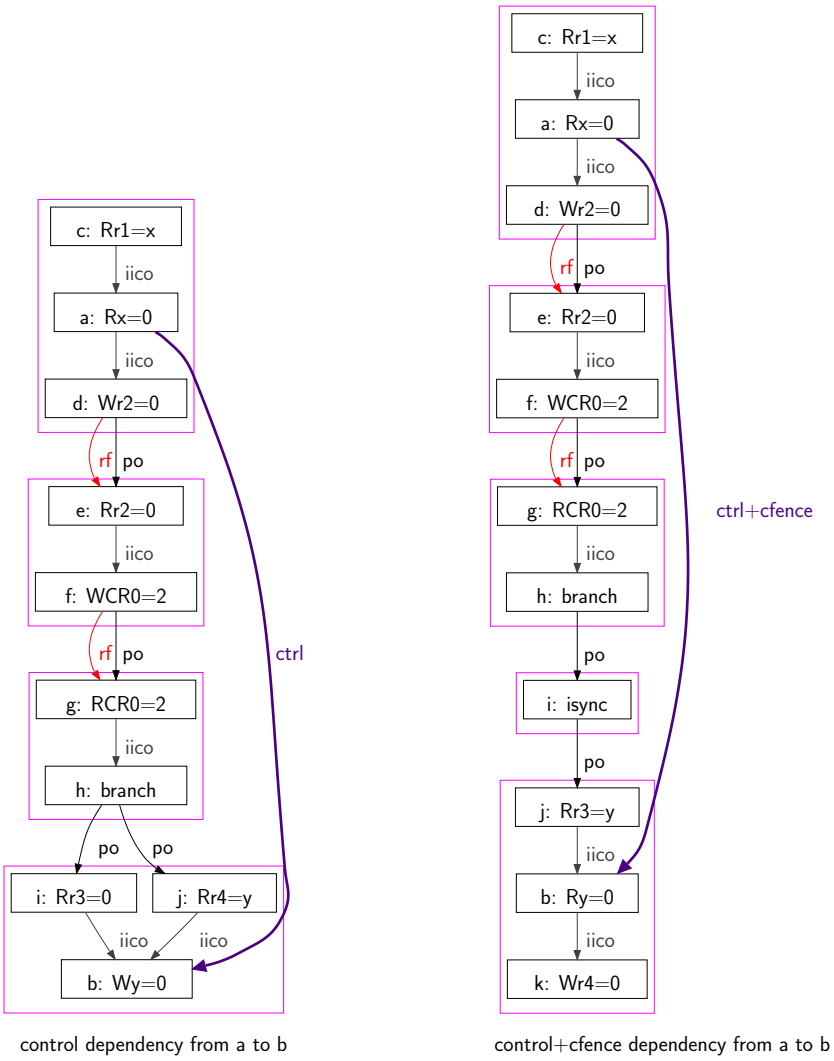


Fig. 24. Control-flow dependencies

(1) lwz r2,0(r1)	ldr r2,[r1]
(2) cmpwi r2,0	cmp r2,#0
(3) bne L0	bne L0
(4) isync	isb
(5) lwz r4,0(r3)	ldr r4,[r3]
(6) L0:	L0:

ensures, thanks to the control fence at line (4), that the load at line (5) cannot be reordered with the load at line (1).

Graphically (see the right diagram[3] of Fig. 24), the read a from address x is related by ctrl+cfence to the read b from address y because there is a path of rf and iico (through registers) between a and a branch event (h here) po-before a cfence (i : isync here) po-before b.

6. PRESERVED PROGRAM ORDER FOR POWER

We can now present how to compute the preserved program order for Power, which we give in Fig. 25. Some readers might find it easier to read the equivalent specification given in Fig. 38. ARM is broadly similar; we detail it in the next section, in the light of our experiments.

To define the preserved program order, we first need to distinguish two parts for each memory event. To name these parts, we again borrow the terminology of the models of [Sarkar et al. 2011; Mador-Haim et al. 2012] for more consistency. We give a table of correspondence between the names of [Sarkar et al. 2011; Mador-Haim et al. 2012] and the present paper in Tab. IV.

present paper	[Sarkar et al. 2011]	[Mador-Haim et al. 2012]
init read $\mathsf{i}(r)$	satisfy read	satisfy read
commit read $\mathsf{c}(r)$	commit read	commit read
init write $\mathsf{i}(w)$	n/a	init write
commit write $\mathsf{c}(w)$	commit write	commit write

Table IV. Terminology correspondence

A memory read r consists of an *init* $\mathsf{i}(r)$, where it reads its value, and a *commit* part $\mathsf{c}(r)$, where it becomes irrevocable. A memory write w consists of an *init* part $\mathsf{i}(w)$, where its value becomes available locally, and a *commit* part $\mathsf{c}(w)$, where the write is ready to be sent out to to other threads.

$$\mathsf{dp} \triangleq \mathsf{addr} \cup \mathsf{data} \qquad \mathsf{rdw} \triangleq \mathsf{po\text{-}loc} \cap (\mathsf{fre};\mathsf{rfe}) \qquad \mathsf{detour} \triangleq \mathsf{po\text{-}loc} \cap (\mathsf{coe};\mathsf{rfe})$$

$$\mathsf{ii}_0 \triangleq \mathsf{dp} \cup \mathsf{rdw} \cup \mathsf{rfi} \qquad \mathsf{ci}_0 \triangleq (\mathsf{ctrl{+}cfence}) \cup \mathsf{detour}$$

$$\mathsf{ic}_0 \triangleq \emptyset \qquad \mathsf{cc}_0 \triangleq \mathsf{dp} \cup \mathsf{po\text{-}loc} \cup \mathsf{ctrl} \cup (\mathsf{addr};\mathsf{po})$$

$$\mathsf{ii} \triangleq \mathsf{ii}_0 \cup \mathsf{ci} \cup (\mathsf{ic};\mathsf{ci}) \cup (\mathsf{ii};\mathsf{ii}) \qquad \mathsf{ci} \triangleq \mathsf{ci}_0 \cup (\mathsf{ci};\mathsf{ii}) \cup (\mathsf{cc};\mathsf{ci})$$

$$\mathsf{ic} \triangleq \mathsf{ic}_0 \cup \mathsf{ii} \cup \mathsf{cc} \cup (\mathsf{ic};\mathsf{cc}) \cup (\mathsf{ii};\mathsf{ic}) \qquad \mathsf{cc} \triangleq \mathsf{cc}_0 \cup \mathsf{ci} \cup (\mathsf{ci};\mathsf{ic}) \cup (\mathsf{cc};\mathsf{cc})$$

$$\mathsf{ppo} \triangleq (\mathsf{ii} \cap \mathsf{RR}) \cup (\mathsf{ic} \cap \mathsf{RW})$$

Fig. 25. Preserved program order for Power

We now describe how the parts of events relate to one another. We do a case disjunction over the part of the events we are concerned with (init or commit).

Thus we define four relations (see Fig. 25): ii relates the init parts of events; ic relates the init part of a first event to the commit part of a second event; ci relates the commit part of a first event to the init part of a second event; cc relates the commit parts of events. We define these four relations recursively, with a least fixpoint semantics; we write r_0 for the base case of the recursive definition of a relation r.

Note that the two parts of an event e are ordered: its init $\mathsf{i}(e)$ precedes its commit $\mathsf{c}(e)$. Thus for two events e_1 and e_2, if for example the commit of e_1 is before the commit of e_2 (i.e. $(e_1, e_2) \in \mathsf{cc}$), then the init of e_1 is before the commit of e_2 (i.e. $(e_1, e_2) \in \mathsf{ic}$). Therefore we have the following inclusions (see also Fig. 25 and 26): ii contains ci; ic contains ii and cc; cc contains ci.

Moreover, these orderings hold transitively: for three events e_1, e_2 and e_3, if the init of e_1 is before the commit of e_2 (i.e. $(e_1, e_2) \in \mathsf{ic}$), which is itself before the init of e_3 (i.e.

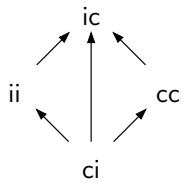


Fig. 26. Inclusions between subevents relations

$(e_2, e_3) \in$ ci), then the init of e_1 is before the init of e_3 (i.e. $(e_1, e_3) \in$ ii). Therefore we have the following inclusions (see also Fig. 25): ii contains (ic; ci) and (ii; ii); ic contains (ic; cc) and (ii; ic); ci contains (ci; ii) and (cc; ci); cc contains (ci; ic) and (cc; cc).

We now describe the base case for each of our four relations, i.e. $\mathsf{ii}_0, \mathsf{ic}_0, \mathsf{ci}_0$ and cc_0. We do so by a case disjunction over whether the concerned events are reads or writes. We omit the cases where there is no ordering, e.g. between two init writes.

Init reads ($\mathsf{ii}_0 \cap \mathsf{RR}$ in Fig. 25) are ordered by (i) the dp relation (which stands for "dependencies"), which gathers address (addr) and data (data) dependencies as defined above; (ii) the rdw relation ("read different writes"), defined as po-loc $\cap$ (fre; rfe).

The (i) case means that if two reads are separated (in program order) by an address , then their init parts are ordered. The micro-architectural intuition is here quite clear: we simply lift to the model level the constraints induced by data-flow paths in the core. The vendors documentations [ppc 2009, Book II, Sec. 1.7.1] and [ARM 2009, Sec. 6.2.1.2] support our intuition, as the following quotes show.

Quoting Power's documentation [ppc 2009, Book II, Sec. 1.7.1]: "[i]f a load instruction depends on the value returned by a preceding load instruction (because the value is used to compute the effective address specified by the second load), the corresponding storage accesses are performed in program order with respect to any processor or mechanism [. . .]." We interpret this bit as a justification for address dependencies (addr) contributing to the ppo.

Furthermore, Power's documentation adds: "[t]his applies even if the dependency has no effect on program logic (e.g., the value returned by the first load is ANDed with zero and then added to the effective address specified by the second load)." We interpret this as a justification for "false dependencies", as described in Sec. 5.2.1.

Quoting ARM's documentation [ARM 2009, Sec. 6.2.1.2]: "[if] the value returned by a read is used to compute the virtual address of a subsequent read or write (this is known as an address dependency), then these two memory accesses will be observed in program order. An address dependency exists even if the value read by the first read has no effect in changing the virtual address (as might be the case if the value returned is masked off before it is used, or if it had no effect on changing a predicted address value)." We interpret this ARM rule as we do for the Power equivalent.

We note however that the Alpha architecture (see our discussion of mp+fence+addr being allowed on Alpha on page 18) demonstrate that sophisticated hardware may invalidate the lifting of core constraint to the complete system.

The (ii) case means that the init parts of two reads b and c are ordered when the instructions are in program order, reading from the same location, and the first read b reads from an external write which is co-before the write a from which the second read c reads, as shown in Fig. 27.

We view this as a side-effect of the write propagation model: as the read b reads from some write that is co-before the write a, b is satisfied before the write a is propagated

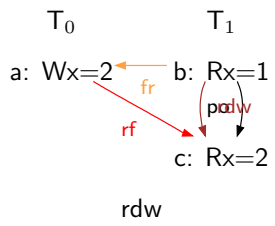


Fig. 27. The read different writes rdw relation

to T_1; furthermore as the read c reads from the write a, it is satisfied after the write a is propagated to T_1. As a result, the read b is satisfied before the read c is.

Init read and init write ($\mathsf{ii}_0 \cap \mathsf{RW}$) relate by address and data dependencies, i.e. dp. This bit is similar to the ordering between two init reads (see $\mathsf{ii}_0 \cap \mathsf{RR}$), but here both address and data dependencies are included. Vendors documentations [ppc 2009, Book II, Sec. 1.7.1] and [arm 2008, Sec. A3.8.2] document address and data dependencies from read to write.

Quoting Power's documentation [ppc 2009, Book II, Sec. 1.7.1]: "[b]ecause stores cannot be performed "out-of-order" (see Book III), if a store instruction depends on the value returned by a preceding load instruction (because the value returned by the load is used to compute either the effective address specified by the store or the value to be stored), the corresponding storage accesses are performed in program order." We interpret this bit as a justification for lifting both address and data dependencies to the ppo.

Quoting ARM's documentation [arm 2008, Sec. A3.8.2]: "[i]f the value returned by a read access is used as data written by a subsequent write access, then the two memory accesses are observed in program order." We interpret this ARM rule as we did for Power above.

Init write and init read ($\mathsf{ii}_0 \cap \mathsf{WR}$) relate by the internal read-from rfi. This ordering constraint stems directly from our interpretation of init subevents (see Tab. IV): init for a write is the point in time when the value stored by the write becomes available locally, while init for a read is the point in time when the read is satisfied. Thus a read can be satified from a local write only once the write in question has made its value available locally.

Commit read and init read or write ($\mathsf{ci}_0 \cap \mathsf{RM}$) relate by ctrl+cfence dependencies. The ctrl+cfence relation models the situation where a first read controls the execution of a branch which contains a control fence that po-precedes the second memory access. An implementation of the control fence could refetch the instructions that po-follows the fence (see for example the quote of ARM's documentation [arm 2008, Sec. A3.8.3] that we give below), and prevent any speculative execution of the control fence. As a result, instructions that follow the fence may start only once the branching decision is irrevocable, i.e. once the controling read is irrevocable.

Quoting Power's documentation [ppc 2009, Sec 1.7.1]: "[b]ecause an isync instruction prevents the execution of instructions following the isync until instructions preceding the isync have completed, if an isync follows a conditional branch instruction that depends on the value returned by a preceding load instruction, the load on which the branch depends is performed before any loads caused by instructions following the isync." We interpret this bit as saying that a branch followed by an isync orders read-read pairs on Power (read-write pairs only need a branch, without an isync). Furthermore the documentation adds: "[t]his applies even if the effects of the "dependency" are

independent of the value loaded (e.g., the value is compared to itself and the branch tests the EQ bit in the selected CR field), and even if the branch target is the sequentially next instruction." This means that the dependency induced by the branch and the isync can be a "false dependency".

Quoting ARM's documentation [arm 2008, Sec. A3.8.3]: "[a]n isb instruction flushes the pipeline in the processor, so that all instructions that come after the isb instruction in program order are fetched from cache or memory only after the isb instruction has completed." We interpret this bit as saying that a branch followed by an isb orders read-read and read-write pairs on ARM.

Commit write and init read ($\mathsf{ci}_0 \cap \mathsf{WR}$) relate by the relation $\mathsf{detour} \triangleq \mathsf{po\text{-}loc} \cap (\mathsf{coe};\mathsf{rfe})$. This means that the commit of a write b to memory location x precedes the init of a read c from x when c reads from an external write a which follows b in coherence, as shown in Fig. 28.

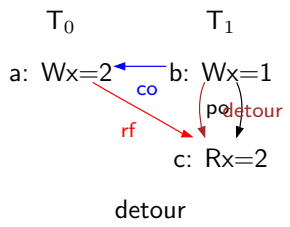


Fig. 28. The detour relation

This effect is similar to the rdw effect (see Fig. 27) but applied to write-read pairs instead of read-read pairs in the case of rdw. As a matter of fact, in the write propagation model of [Sarkar et al. 2011; Mador-Haim et al. 2012], the propagation of a write w_2 to a thread T before T makes a write w_1 to the same address available to the memory system implies that w_2 co-precedes w_1. Thus, if the local w_1 co-precedes the external w_2, it must be that w_2 propagates to T after w_1 is committed. In turn this implies that any read that reads from the external w_2 is satisfied after the local w_1 is committed.

Commits relate by program order if they access the same location, i.e. po-loc. Thus in Fig. 25, cc_0 contains po-loc.

We inherit this contraint from [Mador-Haim et al. 2012]. From the implementation standpoint, the core has to perform some actions so as to enforce the SC PER LOCATION check, i.e. the five patterns of Fig. 6. One way could be to perform these actions at commit time, which this bit of the ppo represents.

However note that our model performs the SC PER LOCATION check independently of the definition of the ppo. Thus, the present commit-to-commit ordering constraint is not required to enforce this particular axiom of the model. Nevertheless, if this bit of the ppo definition is adopted, it will yield specific ordering constraint, such as $\mathsf{po\text{-}loc} \in \mathsf{cc}_0$. As we shall see in Sec. 8.1.2 this very constraint needs to be relaxed to account for some behaviours observed on Qualcomm systems.

In addition, *commit read* and commit read or write ($\mathsf{cc}_0 \cap \mathsf{RM}$) relate by address, data and control dependencies, i.e. dp and ctrl.

Here and below (see Fig. 29) we express "chains of irreversibility", when some memory access depends, by whatever means, on a po-preceding read it can become irrevocable only when the read it depends upon has.

Finally, a read r must be committed before the commit of any access e that is program order after any access e' which is addr-after r, as in the lb+addrs+ww[5] pattern in Fig. 29.

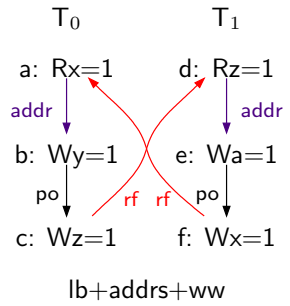


Fig. 29. A variant of the load buffering pattern lb (see also Fig. 7)

In the thread T_0 of Fig. 29, the write to z cannot be read by another thread before the write b knows its own address, which in turn requires the read a to be committed because of the address dependency between a and b. Indeed if the address of b was z, the two writes b and c to z could violate the coWW pattern, an instance of our SC PER LOCATION axiom. Hence the commit of the write c has to be delayed (at least) until after the read a is committed.

Note that the same pattern with data instead of address dependencies is allowed and observed (see http://diy.inria.fr/cats/data-addr).

7. OPERATIONAL MODELS

We introduce here an operational equivalent of our model, our *intermediate machine*, given in Fig. 30, which we then compare to a previous model, the *PLDI machine* of [Sarkar et al. 2011].

7.1. Intermediate machine

Our intermediate machine is simply a reformulation of our axiomatic model as a transition system. We give its formal definition in Fig. 30, which we explain below. In this section, we write $\mathsf{udr}(\mathsf{r})$ for the union of the domain and the range of the relation r. We write $\mathsf{r}{+}{+}[e]$ to indicate that we append the element e at the end of a total order r.

Like the axiomatic model, our machine takes as input the events $\mathbb{E}$, the program order po and an architecture $(\mathsf{ppo}, \mathsf{fences}, \mathsf{prop})$.

It also needs a *path of labels*, i.e. a total order over labels; a label triggers a transition of the machine. We borrow the names of the labels from [Sarkar et al. 2011; Sarkar et al. 2012] for consistency.

We build the labels from the events $\mathbb{E}$ as follows: a write w corresponds to a *commit write* label $\mathrm{c}(w)$ and a *write reaching coherence point* label $\mathrm{cp}(w)$; a read r first guesses from which write w it might read, in an angelic manner; the pair (w, r) then yields two labels: *satisfy read* $\mathrm{s}(w, r)$ and *commit read* $\mathrm{c}(w, r)$.

[5] We note that this pattern does not quite follow the naming convention that we have outlined in Fig. III, but we keep this name for more consistency with previous works, e.g. [Sarkar et al. 2011; Sarkar et al. 2012; Mador-Haim et al. 2012].

Input data: $(\mathsf{ppo}, \mathsf{fences}, \mathsf{prop})$, $(\mathbb{E}, \mathsf{po})$ and p

Derived from p: $\mathsf{co}(\mathbb{E}, \mathsf{p}) \triangleq \{(w_1, w_2) \mid \mathrm{addr}(w_1) = \mathrm{addr}(w_2) \wedge (\mathrm{cp}(w_1), \mathrm{cp}(w_2)) \in \mathsf{p}\}$

COMMIT WRITE

$$\frac{\begin{array}{c}\text{(CW: SC PER LOCATION/coWW)} \quad \text{(CW: PROPAGATION)} \\ \neg(\exists w' \in \mathsf{cw} \text{ s.t. } (w, w') \in \mathsf{po\text{-}loc}) \quad \neg(\exists w' \in \mathsf{cw} \text{ s.t. } (w, w') \in \mathsf{prop}) \\ \text{(CW: fences} \cap \text{WR)} \\ \neg(\exists r \in \mathsf{sr} \text{ s.t. } (w, r) \in \mathsf{fences})\end{array}}{s \xrightarrow{\mathrm{c}(w)} (\mathsf{cw} \cup \{w\}, \mathsf{cpw}, \mathsf{sr}, \mathsf{cr})}$$

WRITE REACHES COHERENCE POINT

$$\frac{\begin{array}{c}\text{(CPW: WRITE IS COMMITTED)} \\ w \in \mathsf{cw} \\ \text{(CPW: po-loc AND cpw ARE IN ACCORD)} \quad \text{(CPW: PROPAGATION)} \\ \neg(\exists w' \in \mathsf{udr}(\mathsf{cpw}) \text{ s.t. } (w, w') \in \mathsf{po\text{-}loc}) \quad \neg(\exists w' \in \mathsf{udr}(\mathsf{cpw}) \text{ s.t. } (w, w') \in \mathsf{prop})\end{array}}{s \xrightarrow{\mathrm{cp}(w)} (\mathsf{cw}, \mathsf{cpw} \mathbin{++} [w], \mathsf{sr}, \mathsf{cr})}$$

SATISFY READ

$$\frac{\begin{array}{c}\text{(SR: WRITE IS EITHER LOCAL OR COMMITTED)} \\ (w, r) \in \mathsf{po\text{-}loc} \vee w \in \mathsf{cw} \\ \text{(SR: PPO/ii}_0 \cap \text{RR)} \quad \text{(SR: OBSERVATION)} \\ \neg(\exists r' \in \mathsf{sr} \text{ s.t. } (r, r') \in \mathsf{ppo} \cup \mathsf{fences}) \quad \neg(\exists w' \text{ s.t. } (w, w') \in \mathsf{co} \wedge (w', r) \in \mathsf{prop}; \mathsf{hb}^*)\end{array}}{s \xrightarrow{\mathrm{s}(w,r)} (\mathsf{cw}, \mathsf{cpw}, \mathsf{sr} \cup \{r\}, \mathsf{cr})}$$

COMMIT READ

$$\frac{\begin{array}{c}\text{(CR: READ IS SATISFIED)} \quad \text{(CR: SC PER LOCATION/coWR, coRW\{1,2\}, coRR)} \\ r \in \mathsf{sr} \quad \mathrm{visible}(w, r) \\ \text{(CR: PPO/cc}_0 \cap \text{RW)} \quad \text{(CR: PPO/(ci}_0 \cup \text{cc}_0) \cap \text{RR)} \\ \neg(\exists w' \in \mathsf{cw} \text{ s.t. } (r, w') \in \mathsf{ppo} \cup \mathsf{fences}) \quad \neg(\exists r' \in \mathsf{sr} \text{ s.t. } (r, r') \in \mathsf{ppo} \cup \mathsf{fences})\end{array}}{s \xrightarrow{\mathrm{c}(w,r)} (\mathsf{cw}, \mathsf{cpw}, \mathsf{sr}, \mathsf{cr} \cup \{r\})}$$

Fig. 30. Intermediate machine

For the architecture's functions ppo, fences and prop to make sense, we need to build a coherence order and a read-from map. We define them from the events $\mathbb{E}$ and the path p as follows:

— $\mathsf{co}(\mathbb{E}, \mathsf{p})$ gathers the writes to the same memory location in the order that their corresponding coherence point labels have in p: $\mathsf{co}(\mathbb{E}, \mathsf{p}) \triangleq \{(w_1, w_2) \mid \mathrm{addr}(w_1) = \mathrm{addr}(w_2) \wedge (\mathrm{cp}(w_1), \mathrm{cp}(w_2)) \in \mathsf{p}\}$;

— $\mathsf{rf}(\mathbb{E}, \mathsf{p})$ gathers the write-read pairs with same location and value which have a commit label in p: $\mathsf{rf}(\mathbb{E}, \mathsf{p}) \triangleq \{(w, r) \mid \mathrm{addr}(w) = \mathrm{addr}(r) \wedge \mathrm{val}(w) = \mathrm{val}(r) \wedge \mathrm{c}(w, r) \in \mathsf{udr}(\mathsf{p})\}$.

In the definition of our intermediate machine, we consider the relations defined w.r.t. co and rf in the axiomatic model (e.g. happens-before hb, propagation prop) to be defined w.r.t. the coherence and read-from above.

Now, our machine operates over a state $(\mathsf{cw}, \mathsf{cpw}, \mathsf{sr}, \mathsf{cr})$ composed of

— a set cw ("committed writes") of writes that have been committed;
— a relation cpw over writes having reached coherence point, which is a total order per location;
— a set sr ("satisfied reads") of reads having been satisfied;
— a set cr ("committed reads") of reads having been committed.

7.1.1. Write transitions The order in which writes enter the set cw for a given location corresponds to the coherence order for that location. Thus a write w cannot enter cw if it contradicts the SC PER LOCATION and PROPAGATION axioms. Formally, a commit write $\mathrm{c}(w)$ yields a commit write transition, which appends w at the end of cw for its location if

— (CW: SC PER LOCATION/coWW) there is no po-loc-subsequent write w' which is already committed, which forbids the coWW case of the SC PER LOCATION axiom, and
— (CW: PROPAGATION) there is no prop-subsequent write w' which is already committed, which ensures that the PROPAGATION axiom holds, and
— (CW: fences $\cap$ WR) there is no fences-subsequent read r which has already been satisfied, which contributes to the semantics of the full fence.

A write can reach coherence point (i.e. take its place at the end of the cpw order) if:

— (CPW: WRITE IS COMMITTED) the write has already been committed, and
— (CPW: po-loc AND cpw ARE IN ACCORD) all the po-loc-previous writes have reached coherence point, and
— (CPW: PROPAGATION) all the prop-previous writes have reached coherence point.

This ensures that the order in which writes reach coherence point is compatible with the coherence order and the propagation order.

7.1.2. Read transitions The read set sr is ruled by the happens-before relation hb. The way reads enter sr ensures NO THIN AIR and OBSERVATION, whilst the way they enter cr ensures parts of SC PER LOCATION and of the preserved program order.

Satisfy read The satisfy read transition $\mathrm{s}(w, r)$ places the read r in sr if:

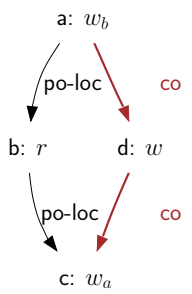
— (SR: WRITE IS EITHER LOCAL OR COMMITTED) the write w from which r reads is either local (i.e. w is po-loc-before r), or has been committed already (i.e. $w \in$ cw), and
— (SR: PPO/$\mathsf{ii}_0 \cap$ RR) there is no (ppo $\cup$ fences)-subsequent read r' that has already been satisfied, which implements the $\mathsf{ii}_0 \cap$ RR part of the preserved program order, and
— (SR: OBSERVATION) there is no write w' co-after w was[6] s.t. $(w', r) \in \mathsf{prop}; \mathsf{hb}^*$, which ensures OBSERVATION.

Commit read To define the commit read transition, we need a preliminary notion. We define a write w to be *visible* to a read r when

— w and r share the same location ℓ;
— w is equal to, or co-after[6], the last write w_b to ℓ that is po-loc-before r, and
— w is po-loc-before r, or co-before the first write w_a to ℓ that is po-loc-after r.

We give here an illustration of the case where w is co-after w_b and before w_a:

[6]Recall that in the context of our intermediate machine, $(w, w') \in$ co means that w and w' are relative to the same address and $(\mathrm{cp}(w), \mathrm{cp}(w')) \in$ p.



visibility of w to r

Now, recall that our read labels contain both a read r and a write w that it might read. The commit read transition $\mathrm{c}(w, r)$ records r in cr when:

— (CR: READ IS SATISFIED) r has been satisfied (i.e. is in sr), and
— (CR: SC PER LOCATION/ coWR, coRW{1,2}, coRR) w is *visible* to r, which prevents the coWR, coRW1 and coRW2 cases of SC PER LOCATION, and
— (CR: PPO/$\mathsf{cc}_0 \cap \mathsf{RW}$) there is no $(\mathsf{ppo} \cup \mathsf{fences})$-subsequent write w' that has already been committed, which implements the $\mathsf{cc}_0 \cap \mathsf{RW}$ part of the preserved program order, and
— (CR: PPO/$(\mathsf{ci}_0 \cup \mathsf{cc}_0) \cap \mathsf{RR}$) there is no $(\mathsf{ppo} \cup \mathsf{fences})$-subsequent read r' that has already been satisfied, which implements the $\mathsf{ci}_0 \cap \mathsf{RR}$ and $\mathsf{cc}_0 \cap \mathsf{RR}$ parts of the preserved program order.

To forbid the coRR case of SC PER LOCATION, one needs to (i) make the read set cr record the write from which a read takes its value, so that cr is a set of pairs (w, r) and no longer just a set of reads; and (ii) augment the definition of $\mathrm{visible}(w, r)$ to require that there is no (w', r') s.t. r' is po-loc-before r yet w' is co-after w . We chose to present the simpler version of the machine to ease the reading.

7.2. Equivalence of axiomatic model and intermediate machine (proof of Thm. 7.1)

We can now state our equivalence result:

THEOREM 7.1. *All behaviours allowed by the axiomatic model are allowed by the intermediate machine and conversely.*

We prove Thm. 7.1 in two steps. We first show that given a set of events $\mathbb{E}$, a program order po over these, and a path p over the corresponding labels such that $(\mathbb{E}, \mathsf{po}, \mathsf{p})$ is accepted by our intermediate machine, the axiomatic execution $(\mathbb{E}, \mathsf{po}, \mathsf{co}(\mathbb{E}, \mathsf{p}), \mathsf{rf}(\mathbb{E}, \mathsf{p}))$ is valid in our axiomatic model:

LEMMA 7.2. *All behaviours allowed by the intermediate machine are allowed by the axiomatic model.*

Conversely, from a valid axiomatic execution $(\mathbb{E}, \mathsf{po}, \mathsf{co}, \mathsf{rf})$, we build a path accepted by the intermediate machine:

LEMMA 7.3. *All behaviours allowed by the axiomatic model are allowed by the intermediate machine.*

We give below the main arguments for proving our results. For more confidence, we have implemented the equivalence proof between our axiomatic model and our

intermediate machine in the Coq proof assistant [Bertot and Casteran 2004]. We give our proof scripts online: http://www0.cs.ucl.ac.uk/staff/j.alglave/cats.

7.2.1. From intermediate machine to axiomatic model (proof of Lem. 7.2) We show here that a path of labels p relative to a set of events $\mathbb{E}$ and a program order po accepted by our intermediate machine leads to a valid axiomatic execution. To do so, we show that the execution $(\mathbb{E}, \mathsf{po}, \mathsf{co}(\mathbb{E}, \mathsf{p}), \mathsf{rf}(\mathbb{E}, \mathsf{p}))$ is valid in our axiomatic model.

Well formedness of $\mathsf{co}(\mathbb{E}, \mathsf{p})$ (i.e. co is a total order on writes to the same location) follows from p being a total order.

Well formedness of $\mathsf{rf}(\mathbb{E}, \mathsf{p})$ (i.e. rf relates a read to a unique write to the same location with same value) follows from the fact that we require that each read r in $\mathbb{E}$ has a unique corresponding write w in $\mathbb{E}$, s.t. r and w have same location and value, and $\mathrm{c}(w, r)$ is a label of p.

The SC PER LOCATION *axiom holds:* to prove this, we show that coWW, coRW1, coRW2, coWR and coRR are forbidden.

coWW is forbidden: suppose as a contradiction two writes e_1 and e_2 to the same location s.t. $(e_1, e_2) \in \mathsf{po}$ and $(e_2, e_1) \in \mathsf{co}(\mathbb{E}, \mathsf{p})$. The first hypothesis entails that $(\mathrm{cp}(e_1), \mathrm{cp}(e_2)) \in \mathsf{p}$, otherwise we would contradict the premise (CPW: po-loc AND cpw ARE IN ACCORD) of the WRITE REACHES COHERENCE POINT rule. The second hypothesis means by definition that $(\mathrm{cp}(e_2), \mathrm{cp}(e_1)) \in \mathsf{p}$. This contradicts the acyclicity of p.

coRW1 is forbidden: suppose as a contradiction a read r and a write w relative to the same location, s.t. $(r, w) \in \mathsf{po}$ and $(w, r) \in \mathsf{rf}(\mathbb{E}, \mathsf{p})$. Thus w cannot be visible to r as it is po-after r. This contradicts the premise (CR: SC PER LOCATION/ coWR, coRW{1,2}, coRR) of the COMMIT READ rule.

coRW2 is forbidden: suppose as a contradiction a read r and two writes w_1 and w_2 relative to the same location, s.t. $(r, w_2) \in \mathsf{po}$ and $(w_2, w_1) \in \mathsf{co}(\mathbb{E}, \mathsf{p})$ and $(w_1, r) \in \mathsf{rf}(\mathbb{E}, \mathsf{p})$. Thus w_1 cannot be visible to r as it is co-after w_2, w_2 itself being either equal or co-after the first write w_a in po-loc after r. This contradicts the premise (CR: SC PER LOCATION/ coWR, coRW{1,2}, coRR) of the COMMIT READ rule.

coWR is forbidden: suppose as a contradiction two writes w_0 and w_1 and a read r relative to the same location, s.t. $(w_1, r) \in \mathsf{po}$, $(w_0, r) \in \mathsf{rf}(\mathbb{E}, \mathsf{p})$ and $(w_0, w_1) \in \mathsf{co}(\mathbb{E}, \mathsf{p})$. Thus w_0 cannot be visible to r as it is co-before w_1, w_1 being itself either equal or co-before the last write w_b in po-loc before r. This contradicts the premise (CR: SC PER LOCATION/ coWR, coRW{1,2}, coRR) of the COMMIT READ rule.

coRR is forbidden: (note that this holds only for the modified version of the machine outlined at the end of Sec. 7.1) suppose as a contradiction two writes w_1 and w_1 and two reads r_1 and r_2 relative to the same location, s.t. $(r_1, r_2) \in \mathsf{po}$, $(w_1, r_1) \in \mathsf{rf}$, $(w_2, r_2) \in \mathsf{rf}$ and $(w_2, w_1) \in \mathsf{co}$. Thus w_2 cannot be visible to r_2 as it is co-before w_1 (following their order in co), and r_1 is po-loc-before r_2. This contradicts the premise (CR: SC PER LOCATION/ coWR, coRW{1,2}, coRR) of the COMMIT READ rule.

The NO THIN AIR *axiom holds:* suppose as a contradiction that there is a cycle in hb^+, i.e. there is an event x s.t. $(x, x) \in ((\mathsf{ppo} \cup \mathsf{fences}); \mathsf{rfe})^+$. Thus there exists y s.t. (i) $(x, y) \in (\mathsf{ppo} \cup \mathsf{fences}); \mathsf{rfe}$ and (ii) $(y, x) \in ((\mathsf{ppo} \cup \mathsf{fences}); \mathsf{rfe})^+$. Note that x and y are reads, since the target of rf is always a read.

We now show that (iii) for two reads r_1 and r_2, having $(r_1, r_2) \in ((\mathsf{ppo} \cup \mathsf{fences}); \mathsf{rfe})^+$ implies $(\mathrm{s}(r_1), \mathrm{s}(r_2)) \in \mathsf{p}$ — we are abusing our notations here, writing $\mathrm{s}(r)$ instead of

$\mathrm{s}(w, r)$ where w is the write from which r reads. From the fact (iii) and the hypotheses (i) and (ii), we derive a cycle in p, a contradiction since p is an order.

Proof of (iii): let us show the base case; the inductive case follows immediately. Let us have two reads r_1 and r_2 s.t. $(r_1, r_2) \in (\mathsf{ppo} \cup \mathsf{fences}); \mathsf{rfe}$. Thus there is w_2 s.t. $(r_1, w_2) \in \mathsf{ppo} \cup \mathsf{fences}$ and $(w_2, r_2) \in \mathsf{rfe}$. Now, note that when we are about to take the SATISFIED READ transition triggered by the label $\mathrm{s}(w_2, r_2)$, we know that the premise (SR: WRITE IS EITHER LOCAL OR COMMITTED) holds. Thus we know that either w_2 and r_2 belong to the same thread, which immediately contradicts the fact that they are in rfe, or that w_2 has been committed. Therefore we have (iv) $(\mathrm{c}(w_2), \mathrm{s}(r_2)) \in \mathsf{p}$. Moreover, we can show that (v) $(\mathrm{s}(r_1), \mathrm{c}(w_2)) \in \mathsf{p}$ by the fact that $(r_1, w_2) \in \mathsf{ppo} \cup \mathsf{fences}$. Thus by (iv) and (v) we have our result.

Proof of (v): take a read r and a write w s.t. $(r, w) \in \mathsf{ppo} \cup \mathsf{fences}$. We show below that (vi) $(\mathrm{c}(r), \mathrm{c}(w)) \in \mathsf{p}$. Since it is always the case that $(\mathrm{s}(r), \mathrm{c}(r)) \in \mathsf{p}$ (thanks to the fact that a read is satisfied before it is committed, see premise (CR: READ IS SATISFIED) of the COMMIT READ rule), we can conclude. Now for (vi): since p is total, we have either our result or $(\mathrm{c}(w), \mathrm{c}(r)) \in \mathsf{p}$. Suppose the latter: then when we are about to take the COMMIT READ transition triggered by $\mathrm{c}(r)$, we contradict the premise (CR: PPO/$\mathsf{cc}_0 \cap \mathsf{RW}$). Indeed we have $w \in \mathsf{cw}$ by $\mathrm{c}(w)$ preceding $\mathrm{c}(r)$ in p, and $(r, w) \in \mathsf{ppo} \cup \mathsf{fences}$ by hypothesis.

The OBSERVATION *axiom holds:* suppose as a contradiction that the relation $\mathsf{fre}; \mathsf{prop}; \mathsf{hb}^*$ is not irreflexive, i.e. there are w and r s.t. $(r, w_2) \in \mathsf{fre}$ and $(w_2, r) \in \mathsf{prop}; \mathsf{hb}^*$. Note that $(r, w_2) \in \mathsf{fr}$ implies the existence of a write w_1 s.t. $(w_1, w_2) \in \mathsf{co}$ and $(w_1, r) \in \mathsf{rf}$. Observe that this entails that r, w_1 and w_2 are relative to the same location.

Thus we take two writes w_1, w_2 and a read r relative to the same location s.t. $(w_1, w_2) \in \mathsf{co}$, $(w_1, r) \in \mathsf{rf}$ and $(w_2, r) \in \mathsf{prop}; \mathsf{hb}^*$ as above. This contradicts the (SR: OBSERVATION) hypothesis. Indeed when we are about to process the transition triggered by the label $\mathrm{s}(r)$, we have $(w_2, r) \in \mathsf{prop}; \mathsf{hb}^*$ by hypothesis, and $(w_1, w_2) \in \mathsf{co}$ by hypothesis.

The PROPAGATION *axiom holds:* suppose as a contradiction that there is a cycle in $(\mathsf{co} \cup \mathsf{prop})^+$, i.e. there is an event x s.t. $(x, x) \in (\mathsf{co} \cup \mathsf{prop})^+$. In other terms there is y s.t. (i) $(x, y) \in \mathsf{co}; \mathsf{prop}$ and (ii) $(y, x) \in (\mathsf{co}; \mathsf{prop})^+$. Note that x and y are writes, since the source of co is always a write.

We now show that (iii) for two writes w_1 and w_2, having $(w_1, w_2) \in (\mathsf{co}; \mathsf{prop})^+$ implies $(\mathrm{cp}(w_1), \mathrm{cp}(w_2)) \in \mathsf{p}^+$. From the fact (iii) and the hypotheses (i) and (ii), we derive a cycle in p, a contradiction since p is an order.

Proof of (iii): let us show the base case; the inductive case follows immediately. Let us have two writes w_1 and w_2 s.t. $(w_1, w_2) \in \mathsf{co}; \mathsf{prop}$; thus there is a write w s.t. $(w_1, w) \in \mathsf{co}$ and $(w, w_2) \in \mathsf{prop}$. Since p is total, we have either the result or $(\mathrm{cp}(w_2), \mathrm{cp}(w_1)) \in \mathsf{p}$. Suppose the latter. Thus we contradict the (CPW: PROPAGATION) hypothesis. Indeed when we are about to take the WRITE REACHES COHERENCE POINT transition triggered by the label $\mathrm{cp}(w)$, we have $(w, w_2) \in \mathsf{prop}$ by hypothesis, and w_2 already in cpw: the hypothesis $(w_1, w) \in \mathsf{co}$ entails that $(\mathrm{cp}(w_1), \mathrm{cp}(w)) \in \mathsf{p}$, and we also have $(\mathrm{cp}(w_2), \mathrm{cp}(w_1)) \in \mathsf{p}$ by hypothesis. Therefore when we are about to process $\mathrm{cp}(w)$ we have placed w_2 in cpw by taking the transition $\mathrm{cp}(w_2)$.

7.2.2. From axiomatic model to intermediate machine (proof of Lem. 7.3) We show here that an axiomatic execution $(\mathbb{E}, \mathsf{po}, \mathsf{co}, \mathsf{rf})$ leads to a valid path p of the intermediate ma-

chine. To do so, we show that the intermediate machine accepts certain paths[8] that linearise the transitive closure of the relation r defined inductively as follows (we abuse our notations here, writing e.g. $\mathrm{c}(r)$ instead of $\mathrm{c}(w, r)$ where w is the write from which r reads):

— for all $r \in \mathbb{E}$, $(\mathrm{s}(r), \mathrm{c}(r)) \in \mathsf{r}$, i.e. we satisfy a read before committing it;
— for all $w \in \mathbb{E}$, $(\mathrm{c}(w), \mathrm{cp}(w)) \in \mathsf{r}$, i.e. we commit a write before it can reach its coherence point;
— forall w and r separated by a fence in program order, $(\mathrm{c}(w), \mathrm{s}(r)) \in \mathsf{r}$, i.e. we commit the write w before we satisfy the read r;
— for all $(w, r) \in \mathsf{rfe}$, $(\mathrm{c}(w), \mathrm{s}(w, r)) \in \mathsf{r}$, i.e. we commit a write before reading externally from it;
— for all $(w_1, w_2) \in \mathsf{co}$, $(\mathrm{cp}(w_1), \mathrm{cp}(w_2)) \in \mathsf{r}$, i.e. cpw and the coherence order are in accord;
— for all $(r, e) \in \mathsf{ppo} \cup \mathsf{fences}$, we commit a read r before processing any other event e (i.e. satisfying e if e is a read, or committing e if e is a write), if r and e are separated e.g. by a dependency or a barrier;
— for all $(w_1, w_2) \in \mathsf{prop}^+$, $(\mathrm{cp}(w_1), \mathrm{cp}(w_2)) \in \mathsf{r}$, i.e. cpw and propagation order are in accord.

Since we build p as a linearisation of the relation r defined above, we first need to show that we are allowed to linearise r, i.e. that r is acyclic.

Linearisation of r: suppose as a contradiction that there is a cycle in r, i.e. there is a label l s.t. $(l, l) \in \mathsf{r}^+$. Let us write S_1 for the set of commit writes, satisfy reads and commit reads, and S_2 for the set of writes reaching coherence points. We show by induction that for all pair of labels $(l_1, l_2) \in \mathsf{r}^+$, either:

— l_1 and l_2 are both in S_1, and their corresponding events e_1 and e_2 are ordered by happens-before, i.e. $(e_1, e_2) \in \mathsf{hb}^+$, or
— l_1 and l_2 are both in S_2 and their corresponding events e_1 and e_2 are ordered by $(\mathsf{co} \cup \mathsf{prop})^+$, or
— l_1 is in S_1, l_2 in S_2, and their corresponding events e_1 and e_2 are ordered by happens-before, or
— l_1 is in S_1, l_2 in S_2, and their corresponding events e_1 and e_2 are ordered by $(\mathsf{co} \cup \mathsf{prop})^+$, or
— l_1 is in S_1, l_2 in S_2, and their corresponding events e_1 and e_2 are ordered by $\mathsf{hb}^+; (\mathsf{co} \cup \mathsf{prop})^+$, or
— l_1 is a satisfy read and l_2 the corresponding commit read.
— l_1 is a commit write and l_2 the write reaching coherence point.

Each of these items contradicts the fact that $l_1 = l_2$: the first two resort to the axioms of our model prescribing the acyclicity of hb on the one hand (NO THIN AIR), and $\mathsf{co} \cup \mathsf{prop}$ on the second hand (PROPAGATION); all the rest resorts to the transitions being different.

We now show that none of the transitions of the machine given in Fig. 30 can block.

[8]The path has to linearise r so that for all writes w_1 and w_2, if $(\mathrm{cp}(w_1), \mathrm{cp}(w_2) \in \mathsf{p}$ then $(\mathrm{c}(w_1), \mathrm{c}(w_2) \in \mathsf{p}$. We refer to this property as "p being fifo'.'In other words, the linearisation must be such that coherence point labels and commit labels are in accord. Note that this does not affect the generality of Lem. 7.3, as to prove this lemma, we only need to find one valid intermediate machine path for a given axiomatic execution; our path happens to be so that coherence point and commit labels are in accord.

COMMIT WRITE *does not block:* suppose as a contradiction a label $\mathrm{c}(w)$ s.t. the transition of the intermediate machine triggered by $\mathrm{c}(w)$ blocks. This means that one of the premises of the COMMIT WRITE rule is not satisfied.

First case: the premise (CW: SC PER LOCATION/coWW) is not satisfied, i.e. there exists w' in cw s.t. $(w, w') \in$ po-loc. Since $(w, w') \in$ po-loc we have $(w, w') \in$ co by SC PER LOCATION. By construction of p, we deduce $(\mathrm{cp}(w), \mathrm{cp}(w')) \in$ p. By p being fifo (see footnote on page 43), we deduce (i) $(\mathrm{c}(w), \mathrm{c}(w')) \in$ p. Moreover if w' is in cw when we are about to process $\mathrm{c}(w)$, then w' has been committed before w, hence (ii) $(\mathrm{c}(w'), \mathrm{c}(w)) \in$ p. By (i) and (ii), we derive a cycle in p, a contradiction since p is an order (since we build it as a linearisation of a relation).

Second case: the premise (CW: PROPAGATION) is not satisfied, i.e. there exists w' in cw s.t. $(w, w') \in$ prop. Since w' is in cw when we are about to process the label $\mathrm{c}(w)$, we have (i) $(\mathrm{c}(w'), \mathrm{c}(w)) \in$ p. Since $(w, w') \in$ prop, we have (ii) $(\mathrm{cp}(w), \mathrm{cp}(w')) \in$ p (recall that we build p inductively; in particular the order of the $\mathrm{cp}(w)$ transitions in p respects the order of the corresponding events in prop). Since we build p so that it is fifo (see footnote on page 43), we deduce from (i) the fact (iii) $(\mathrm{c}(w), \mathrm{c}(w')) \in$ p. From (ii) and (iii), we derive a cycle in p, a contradiction since p is an order (since we build it as a linearisation of a relation).

Third case: the premise (CW: fences $\cap$ WR) is not satisfied, i.e. there exists r in sr s.t. $(w, r) \in$ fences. From $r \in$ sr we deduce $(\mathrm{s}(r), \mathrm{c}(w)) \in$ p. From $(w, r) \in$ fences we deduce (by construction of p) $(\mathrm{c}(w), \mathrm{s}(r)) \in$ p, which creates a cycle in p.

WRITE REACHES COHERENCE POINT *does not block:* suppose as a contradiction a label $\mathrm{cp}(w)$ s.t. the transition of the intermediate machine triggered by $\mathrm{cp}(w)$ blocks. This means that one of the premises of the WRITE REACHES COHERENCE POINT rule is not satisfied.

First case: the premise (CPW: WRITE IS COMMITTED) is not satisfied, i.e. w has not been committed. This is impossible since $(\mathrm{c}(w), \mathrm{cp}(w)) \in$ p by construction of p.

Second case: the premise (CPW: po-loc AND cpw ARE IN ACCORD) is not satisfied, i.e. there is a write w' that has reached coherence point s.t. $(w, w') \in$ po-loc. From $(w, w') \in$ po-loc, we know by SC PER LOCATION that $(w, w') \in$ co. Thus by construction of p, we know (i) $(\mathrm{cp}(w), \mathrm{cp}(w')) \in$ p. From w' having reached coherence point before w, we know (ii) $(\mathrm{cp}(w'), \mathrm{cp}(w)) \in$ p. By (i) and (ii), we derive a cycle in p, a contradiction since p is an order (since we build it as a linearisation of a relation).

Third case: the premise (CPW: PROPAGATION) is not satisfied, i.e. there is a write w' that has reached coherence point s.t. $(w, w') \in$ prop. By construction of p, we deduce (i) $(\mathrm{cp}(w), \mathrm{cp}(w') \in$ p). From w' having reached coherence point before w, we know (ii) $(\mathrm{cp}(w'), \mathrm{cp}(w)) \in$ p. By (i) and (ii), we derive a cycle in p, a contradiction since p is an order (since we build it as a linearisation of a relation).

SATISFY READ *does not block:* suppose as a contradiction a label $\mathrm{s}(w, r)$ s.t. the transition of the intermediate machine triggered by $\mathrm{s}(w, r)$ blocks. This means that one of the premises of the SATISFY READ rule is not satisfied. Note that since $\mathrm{s}(w, r)$ is a label of p, we have (i) $(w, r) \in$ rf.

First case: the premise (SR: WRITE IS EITHER LOCAL OR COMMITTED) is not satisfied, i.e. w is neither local nor committed. Suppose w not local (otherwise we contradict our hypothesis); let us show that it has to be committed. Suppose it is not, therefore we have $(\mathrm{s}(r), \mathrm{c}(w)) \in$ p. Since w is not local, we have $(w, r) \in$ rfe, from which we deduce (by construction of p) that $(\mathrm{c}(w), \mathrm{s}(r)) \in$ p; this leads to a cycle in p.

Second case: the premise (SR: PPO/$\mathrm{ii}_0 \cap$ RR) is not satisfied, i.e. there is a satisfied read r' s.t. $(r, r') \in$ ppo $\cup$ fences. From r' being satisfied we deduce $(\mathrm{s}(r'), \mathrm{s}(r)) \in$ p. From $(r, r') \in$ ppo $\cup$ fences and by construction of p, we deduce $(\mathrm{c}(r), \mathrm{s}(r')) \in$ p. Since $\mathrm{s}(r)$ precedes $\mathrm{c}(r)$ in p by construction of p, we derive a cycle in p, a contradiction.

Third case: the premise (SR: OBSERVATION) is not satisfied, i.e. there is a write w' s.t. $(w, w') \in \mathsf{co}$ and $(w', r) \in \mathsf{prop}; \mathsf{hb}^*$. Since $(w, w') \in \mathsf{co}$, by (i) $(r, w') \in \mathsf{fr}$. Therefore we contradict the OBSERVATION axiom.

COMMIT READ *does not block:* suppose as a contradiction a label $\mathrm{c}(w, r)$ s.t. the transition of the intermediate machine triggered by $\mathrm{c}(w, r)$ blocks. This means that one of the premises of the COMMIT READ rule is not satisfied.

First case: the premise (CR: READ IS SATISFIED) is not satisfied, i.e. r is not in sr. This is impossible since we impose $(\mathrm{s}(r), \mathrm{c}(r)) \in \mathsf{p}$ when building p.

Second case: the premise (CR: SC PER LOCATION/ coWR, coRW{1,2}, coRR) is not satisfied, i.e. w is not visible to r. This contradicts the SC PER LOCATION axiom, as follows. Recall that the visibility definition introduces w_a as the first write to the same location as r which is po-loc-after r; and w_b as the last write to the same location as r which is po-loc-before r. Now, if w is not visible to r we have either (i) w is co-before w_b, or (ii) equal or co-after w_a.

Suppose (i), we have $(w, w_b) \in \mathsf{co}$. Hence we have $(w, w_b) \in \mathsf{co}$, $(w_b, r) \in \mathsf{po\text{-}loc}$ by definition of w_b, and $(w, r) \in \mathsf{rf}$ by definition of $\mathrm{s}(w, r)$ being in p. Thus we contradict the coWR case of the SC PER LOCATION axiom. The (ii) case is similar, and contradicts (coRW1) if $w = w_a$ or (coRW2) if $(w_a, w) \in \mathsf{cw}$.

Third case: the premise (CR: PPO/$\mathsf{cc}_0 \cap \mathsf{RW}$) is not satisfied, i.e. there is a write w' in cw s.t. $(r, w') \in \mathsf{ppo} \cup \mathsf{fences}$. From $w' \in \mathsf{cw}$ we deduce $(\mathrm{c}(w'), \mathrm{c}(r)) \in \mathsf{p}$. From $(r, w') \in \mathsf{ppo} \cup \mathsf{fences}$ we deduce $(\mathrm{c}(r), \mathrm{c}(w')) \in \mathsf{p}$ by construction of p. This leads to a cycle in p, a contradiction.

Fourth case: the premise (CR: PPO/$(\mathsf{ci}_0 \cup \mathsf{cc}_0) \cap \mathsf{RR}$) is not satisfied, i.e. there is a read r' in sr s.t. $(r, r') \in \mathsf{ppo} \cup \mathsf{fences}$. From $r' \in \mathsf{sr}$ we deduce $(\mathrm{s}(r'), \mathrm{c}(r)) \in \mathsf{p}$. From $(r, r') \in \mathsf{ppo} \cup \mathsf{fences}$ we deduce $(\mathrm{c}(r), \mathrm{s}(r')) \in \mathsf{p}$ by construction of p. This leads to a cycle in p, a contradiction.

7.3. Comparing our model and the PLDI machine

The PLDI machine is an operational model, which we describe here briefly (see [Sarkar et al. 2011] for details). This machine maintains a coherence order (a strict partial order over the writes to the same memory location), and, per thread, a list of the writes and fences that have been propagated to that thread.

A load instruction yields two transitions of this machine (amongst others): a *satisfy read* transition, where the read takes its value, and a *commit read* transition, where the read becomes irrevocable. A store instruction yields a *commit write* transition, where the write becomes available to be read, several *propagate write* transitions, where the write is sent out to different threads of the system, and a *reaching coherence point* transition, where the write definitely takes place in the coherence order. We summarise the effect of a PLDI transition on a PLDI state in the course of this section.

We show that a valid path of the PLDI machine leads to valid path of our intermediate machine.

First, we show how to relate the two machines.

7.3.1. Mapping PLDI Objects (labels and states) to intermediate objects We write $\mathsf{pl2l}(l)$ to map a PLDI l to a label of the intermediate machine. For technical convenience we assume a special noop intermediate label such that, from any state s of the intermediate machine, we may perform a transition from s to s via noop .

We can then define $\mathsf{pl2l}(l)$ as being the eponymous label in the intermediate machine if it exists (i.e. for commit write, write reaches coherence point, satisfy read and commit read), and noop otherwise. We derive from this mapping the set $\mathbb{L}_i$ of intermediate labels composing our intermediate path.

We build a state of our intermediate machine from a PLDI state s and an accepting PLDI path p; we write $\mathsf{ps2s}(\mathsf{p}, s) = (\mathsf{cw}, \mathsf{cpw}, \mathsf{sr}, \mathsf{cr})$ for the intermediate state built as follows:

— for a given location, cw is simply the set of writes to this location that have been committed in s;
— we take cpw to be all the writes having reached coherence point in s, ordered w.r.t. p;
— the set sr gathers the reads that have been satisfied in the state s: we simply add a read to sr if the corresponding satisfy transition appears in p before the transition leading to s;
— the set cr gathers the reads that have been committed in the state s.

7.3.2. Building a path of the intermediate machine from a PLDI path A given path p of the PLDI machine entails a run $s_0 \xrightarrow{l_1} s_1 \xrightarrow{l_2} \cdots \xrightarrow{l_n} s_n$ such that $(l, l') \in \mathsf{p}$ if and only if there exist i and j such that $i < j$ and $l = l_i$ and $l' = l_j$.

We show that $\mathsf{ps2s}(s_0) \xrightarrow{\mathsf{pl2l}(l_1)} \mathsf{ps2s}(s_1) \xrightarrow{\mathsf{pl2l}(l_2)} \cdots \xrightarrow{\mathsf{pl2l}(l_n)} \mathsf{ps2s}(s_n)$ is a path of our intermediate machine. We proceed by induction for $0 \leq m \leq n$. The base case $m = 0$ is immediately satisfied by the single-state path $\mathsf{ps2s}(s_0)$.

Now, inductively assume that $\mathsf{ps2s}(s_0) \xrightarrow{\mathsf{pl2l}(l_1)} \mathsf{ps2s}(s_1) \xrightarrow{\mathsf{pl2l}(l_2)} \cdots \xrightarrow{\mathsf{pl2l}(l_m)} \mathsf{ps2s}(s_m)$ is a path of the intermediate machine. We prove the case for $m + 1$. Take the transition $s_m \xrightarrow{l_{m+1}} s_{m+1}$ of the PLDI machine. We prove $\mathsf{ps2s}(s_m) \xrightarrow{\mathsf{pl2l}(l_{m+1})} \mathsf{ps2s}(s_{m+1})$ to complete the induction. There are several cases.

When $\mathsf{pl2l}(l_{m+1}) = \mathsf{noop}$ we have that $\mathsf{ps2s}(s_m) = \mathsf{ps2s}(s_{m+1})$ simply because the PLDI labels that have noop as an image by pl2l do not affect the components cw, cpw, sr and cr of our state.

Only the PLDI transitions that have an eponymous transition in our machine affect our intermediate state. Thus we list below the corresponding four cases.

Commit write: in the PLDI machine, a commit transition of a write w makes this write co-after all the writes to the same location that have been propagated to its thread. The PLDI transition guarantees that (i) w had not been committed in s_1.

Observe that w takes its place in the set cw , ensuring that we modify the state as prescribed by our COMMIT WRITE rule.

Now, we check that we do not contradict the premises of our COMMIT WRITE rule.

First case: contradict the premise (CW: SC PER LOCATION/coWW), i.e. take a write w' in cw s.t. $(w, w') \in$ po-loc. In that case, we contradict the fact that the commit order respects po-loc (see [Sarkar et al. 2011, p. 7, item 4 of § Commit in-flight instruction]).

Second case: contradict the premise (CW: PROPAGATION), i.e. take a write w' in cw s.t. $(w, w') \in$ prop. In that case, $(w, w') \in$ prop guarantees that w was propagated to the thread of w' in s_1. Therefore (see [Sarkar et al. 2011, p. 6, premise of § Propagate write to another thread]), w was seen in s_1. For the write to be seen, it needs to have been sent in a write request [Sarkar et al. 2011, p. 6, item 1 of §Accept write request]; for the write request to be sent, the write must have been committed [Sarkar et al. 2011, p. 8, action 4 of §Commit in-flight instruction]. Thus we contradict (i).

Third case: contradict the premise (CW: fences $\cap$ WR), i.e. take a read r in sr s.t. $(w, r) \in$ fences. Since r is in sr, r is satisfied. Note that the write from which r reads can be either local (see [Sarkar et al. 2011, p. 8, § Satisfy memory read by forwarding an in-flight write directly to reading instruction]) or committed (see [Sarkar et al. 2011, p. 8, § Satisfy memory read from storage subsystem]).

In both cases, the fence between w and r must have been committed (see [Sarkar et al. 2011, p. 8, item 2 of § Satisfy memory read by forwarding an in-flight write directly to reading instruction and item 2 of § Satisfy memory read from storage subsystem]). Thus by [Sarkar et al. 2011, p. 7, item 6 of § Commit in-flight instruction], w has been committed, a contradiction of (i).

Write reaches coherence point: in the PLDI machine, write reaching coherence point transitions order writes following a linearisation of $(\mathsf{co} \cup \mathsf{prop})^+$. Our cw implements that per location, then we make the writes reach coherence point following cw and $(\mathsf{co} \cup \mathsf{prop})^+$.

Observe that a write w reaching coherence point takes its place after all the writes having already reached coherence point, ensuring that we modify the intermediate state as prescribed by our WRITE REACHES COHERENCE POINT rule.

Now, we check that we do not contradict the premises of our WRITE REACHES COHERENCE POINT rule.

First case: contradict the premise (CPW: WRITE IS COMMITTED), i.e. suppose that w is not committed. This is impossible as the PLDI machine requires a write to have been seen by the storage subsystem for it to reach coherence point [Sarkar et al. 2012, p. 3, §Write reaches its coherence point]. For the write to be seen, it needs to have been sent in a write request [Sarkar et al. 2011, p. 6, item 1 of §Accept write request]; for the write request to be sent, the write must have been committed [Sarkar et al. 2011, p. 7, action 4 of §Commit in-flight instruction].

Second case: contradict the premise (CPW: po-loc AND cpw ARE IN ACCORD), i.e. take a write w' in cpw s.t. $(w, w') \in$ po-loc. This means that (i) w' has reached coherence point before w, despite w preceding w' in po-loc. This is a contradiction, as we now explain. If $(w, w') \in$ po-loc, then w is propagated to its own thread before w' [Sarkar et al. 2011, p. 7, action 4 of §Commit in-flight instruction]. Since w and w' access the same address, when w' is propagated to its own thread, w' is recorded as being co-after all the writes to the same location already propagated to its thread [Sarkar et al. 2011, p. 6, item 3 of §Accept write request], in particular w. Thus we have $(w, w') \in$ co. Now, when w' reaches coherence point, all its coherence predecessors must have reached theirs [Sarkar et al. 2012, p. 4, item 2 of §Write reaches its coherence point], in particular w. Thus w should reach its coherence point before w', which contradicts (i).

Third case: contradict the premise (CPW: PROPAGATION), i.e. take a write w' in cpw s.t. $(w, w') \in$ prop. This contradicts the fact that writes cannot reach coherence point in an order that contradicts propagation (see [Sarkar et al. 2012, p. 4, item 3 of §Write reaches coherence point]).

Satisfy read: in the PLDI machine, a satisfy read transition does not modify the state (i.e. $s_1 = s_2$). In the intermediate state, we augment sr with the read that was satisfied.

Now, we check that we do not contradict the premises of our SATISFY READ rule.

First case: contradict the (SR: WRITE IS EITHER LOCAL OR COMMITTED) premise, i.e. suppose that w is neither local nor committed. Then we contradict the fact that a read can read either from a local po-loc-previous write (see [Sarkar et al. 2011, p. 8, item 1 of §Satisfy memory read by forwarding an in-flight write directly to reading instruction]), or from a write from the storage subsytem — which therefore must have been committed (see [Sarkar et al. 2011, p. 8, § Satisfy memory read from storage subsystem]).

Second case: contradict the (SR: PPO/$\mathsf{ii}_0 \cap \mathsf{RR}$) premise, i.e. take a satisfied read r' s.t. $(r, r') \in \mathsf{ppo} \cup \mathsf{fences}$. Then we contradict the fact that read satisfaction follows the

preserved program order and the fences (see [Sarkar et al. 2011, p. 8, all items of both §Satisfy memory read by forwarding an in-flight write directly to reading instruction and §Satisfy memory read from storage subsystem]).

Third case: contradict the (SR: OBSERVATION) premise, i.e. take a write w' in co after w s.t. $(w', r) \in \mathsf{prop}; \mathsf{hb}^*$. Since w and w' are related by co, they have the same location. The PLDI transition guarantees that (i) w is the most recent write to $\mathrm{addr}(r)$ propagated the thread of r (see [Sarkar et al. 2011, p. 6, §Send a read response to a thread]). Moreover, $(w', r) \in \mathsf{prop}; \mathsf{hb}^*$ ensures that (ii) w' has been propagated to the thread of r, or there exists a write e such that $(w', e) \in \mathsf{co}$ and e is propagated to the thread of r. Therefore, we have $(w', w) \in \mathsf{co}$ by [Sarkar et al. 2011, p. 6, item 2 of §Propagate write to another thread].

Therefore by $(w, w') \in \mathsf{co}$, we know that w reaches its coherence point before w'. Yet w' is a co-predecessor of w, which contradicts [Sarkar et al. 2012, p. 4, item 2 of §Write reaches coherence point].

Proof of (ii): we take $(w', r) \in \mathsf{prop}; \mathsf{hb}^*$ as above. This gives us a write w'' s.t. $(w', w'') \in \mathsf{prop}$ and $(w'', r) \in \mathsf{hb}^*$. Note that $(w', w'') \in \mathsf{prop}$ requires the presence of a barrier between w' and w''.

We first remind the reader of a notion from [Sarkar et al. 2011], the *group A of a barrier*: the group A of a barrier is the set of all the writes that have been propagated to the thread holding the barrier when the barrier is sent to the system (see [Sarkar et al. 2011, p. 5,§Barriers (sync and lwsync) and cumulativity by fiat]). When a barrier is sent to a thread, all the writes in its group A must have been propagated to that thread (see [Sarkar et al. 2011, p. 6 item 2 of §Propagate barrier to another thread]). Thus if we show that (a) the barrier between w' and w'' is propagated to the thread of r and (b) w' is in the group A of this barrier, we have our result.

Let us now do a case disjunction over $(w', w'') \in \mathsf{prop}$.

When $(w', w'') \in \mathsf{prop\text{-}base}$, we have a barrier b such that (i) $(w', b) \in \mathsf{fences} \cup (\mathsf{rfe}; \mathsf{fences})$ and (ii) $(b, r) \in \mathsf{hb}^*$. Note (i) immediately entails that w' is in the group A of b. For (ii), we reason by induction over $(b, r) \in \mathsf{hb}^*$, the base case being immediate. In the inductive case, we have a write e such that b is propagated to the thread of e before e and e is propagated to the thread of r before r. Thus, by [Sarkar et al. 2011][p. 6, item 3 of §Propagate write to another thread], b is propagated to r.

When $(w', w'') \in \mathsf{com}^*; \mathsf{prop\text{-}base}^*; \mathsf{ffence}; \mathsf{hb}^*$, we have a barrier b (which is a full fence) such that $(w', b) \in \mathsf{com}^*; \mathsf{prop\text{-}base}^*$ and $(b, r) \in \mathsf{hb}^*$. We now proceed by reasoning over com^*.

If there is no com step, then we have $(w', b) \in \mathsf{prop\text{-}base}^*$, thus w' is propagated to the thread of b before b by the prop-base case above. Note that this entails that w' is in the group A of b. Since b is a full fence, b propagates to all threads (see [Sarkar et al. 2011, premise of §Acknowledge sync barrier]), in particular to the thread of r. Thus by [Sarkar et al. 2011, item 2 of §Propagate barrier to another thread], w' is propagated to r.

In the com^+ case (i.e. $(w', b) \in \mathsf{com}^*; \mathsf{prop\text{-}base}^*$), we remark that $\mathsf{com}^+ = \mathsf{com} \cup \mathsf{co}; \mathsf{rf} \cup \mathsf{fr}; \mathsf{rf}$. Thus since w' is a write, only the cases rf, co and $\mathsf{co}; \mathsf{rf}$ apply. In the rf case, we have $(w', b) \in \mathsf{prop\text{-}base}^*$, which leads us back to the base case (no com step) above. In the co and co; rf cases, we have $(w', b) \in \mathsf{co}; \mathsf{prop\text{-}base}^*$, i.e. there exists a write e such that $(w', e) \in \mathsf{co}$ and $(e, b) \in \mathsf{prop\text{-}base}^*$, i.e. our result.

Commit read: in the PLDI machine, a commit read transition does not modify the state. In the intermediate state, we augment cr with the read that was committed. We now check that we do not contradict the premises of our COMMIT READ rule.

First case: contradict the premise (CR: READ IS SATISFIED), i.e. suppose that the read that we want to commit is not in sr; this means that this read has not been

	Power	ARM
# tests	8117	9761
invalid	0	1500
unseen	1182	1820

Table V. Summary of our experiments on Power and ARM h/w

satisfied. This is impossible since a read must be satisfied before it is committed (see [Sarkar et al. 2011, p. 7, item 1 of §Commit in-flight instruction]).

Second case: contradict the (CR: SC PER LOCATION/ coWR, coRW{1,2}, coRR) premise. This is impossible since the PLDI machine prevents coWR, coRW1 and coRW2) (see [Sarkar et al. 2011, p. 3, §Coherence]).

Third case: contradict the premise (CR: PPO/$\mathsf{cc}_0 \cap \mathsf{RW}$), i.e. take a committed write w' s.t. $(r, w') \in \mathsf{ppo} \cup \mathsf{fences}$. Since r must have been committed before w' (by [Sarkar et al. 2011, p. 7, items 2, 3, 4, 5, 7 of §Commit in-flight instruction]), we get a contradiction.

Fourth case: contradict the premise (CR: PPO/$(\mathsf{ci}_0 \cup \mathsf{cc}_0) \cap \mathsf{RR}$), i.e. take a satisfied read r' s.t. $(r, r') \in \mathsf{ppo} \cup \mathsf{fences}$. Since r must have been committed before r' was satisfied (by [Sarkar et al. 2011, p. 8, item 3 of §Satisfy memory read from storage subsystem and item 3 of §Satisfy memory read by forwarding an in-flight write directly to reading instruction]), we get a contradiction.

8. TESTING AND SIMULATION

As usual in this line of work, we developed our model in tandem with extensive experiments on hardware. We report here on our experimental results on Power and ARM hardware. Additionally, we experimentally compared our model to the ones of [Sarkar et al. 2011] and [Mador-Haim et al. 2012]. Moreover, we developed a new simulation tool called herd[9]. Finally, we adapted the CBMC tool [Alglave et al. 2013] to our new models.

8.1. Hardware testing

We performed our testing on several platforms using the diy tool suite [Alglave et al. 2010; Alglave et al. 2011; Alglave et al. 2012]. This tool generates *litmus tests*, i.e. very small programs in x86, Power or ARM assembly code, with specified initial and final states. It then runs these tests on hardware and collects the memory and register states that it observed during the runs. Most of the time, litmus tests violate SC: if one can observe their final state on a given machine, then this machine exhibits features beyond SC.

We generated 8117 tests for Power and 9761 tests for ARM, illustrating various features of the hardware, e.g. lb, mp, sb, and their variations with dependencies and barriers, e.g. lb+addrs, mp+lwsync+addr, sb+syncs.

We tested the model described in Fig. 5, 18, and 25 on Power and ARM machines, to check experimentally if this model was suitable for these two architectures. In the following, we write "Power model" for this model instantiated for Power, and "Power-ARM model" for this model instantiated for ARM. We give a summary of our experiments in Tab. V.

For each architecture, the row "unseen" gives the number of tests that our model allows but that the hardware does not exhibit. This can be the case because our model is

[9]We acknowledge that we reused some code written by colleagues, in particular Susmit Sarkar, in an earlier version of the tool.

too coarse (i.e. fails to reflect the architectural intent in forbidding some behaviours), or because the behaviour is intended to be allowed by the architect, but is not implemented yet.

The column "invalid" gives the number of tests that our model forbids but that the hardware does exhibit. This can be because our model is too strict and forbids behaviours that are actually implemented, or because the behaviour is a hardware bug.

8.1.1. On Power We tested three generations of machines: Power G5, 6 and 7. The complete logs of our experiments can be found at http://diy.inria.fr/cats/model-power.

Our Power model is not invalidated by Power hardware (there is no test in the "invalid" columns on Power in Tab. V). In particular it allows mp+lwsync+addr-po-detour, which [Sarkar et al. 2011] wrongly forbids, as this behaviour is observed on hardware (see http://diy.inria.fr/cats/pldi-power/#lessvs).

Our Power model allows some behaviours (see the "unseen" columns on Power), e.g. lb, that are not observed on Power hardware. This is to be expected as the lb pattern is not yet implemented on Power hardware, despite being clearly architecturally allowed [Sarkar et al. 2011].

8.1.2. On ARM We tested several system configurations: NVDIA Tegra 2 and 3, Qualcomm APQ8060 and APQ8064, Apple A5X and A6X, and Samsung Exynos 4412, 5250 and 5410. The complete logs of our experiments can be found at http://diy.inria.fr/cats/model-arm. This section about ARM testing is articulated as follows:

— we first explain how our Power-ARM model is invalidated by ARM hardware (see the "invalid" column on ARM) by 1500 tests (see §"Our Power-ARM model is invalidated by ARM hardware"). We detail and document the discussion below at http://diy.inria.fr/cats/arm-anomalies.
— we then propose a model for ARM (see §"Our proposed ARM model").
— we then explain how we tested this model on ARM machines, and the anomalies that we have found whilst testing (see §"Testing our model").

Our Power-ARM model is invalidated by ARM hardware Amongst the tests we have ran on ARM hardware, some unveiled a *load-load hazard* bug in the coherence mechanism of all machines. This bug is a violation of the coRR pattern shown in Sec. 4, and was later acknowledged as such by ARM, in the context of Cortex-A9 cores, in the note [arm 2011].[10]

Amongst the machines that we have tested, this note applies directly to Tegra 2 and 3, A5X, Exynos 4412. Additionally Qualcomm's APQ8060 is supposed to have many architectural similarities with the ARM Cortex-A9, thus we believe that the note might apply to APQ8060 as well. Morever we have observed load-load hazards anomalies on cortex-A15 based systems (Exynos 5250 and 5410), on the cortex-A15 compatible Apple "Swift" (A6X) and on the Krait-based APQ8064, although much less often than on cortex-A9 based systems. Note that we observed the violation of coRR itself quite frequently, as illustrated by the first line of the table in Tab. VI. The second line of Tab. VI refers to a more sophisticated test, coRSDWI (see Fig. 31), whose executions reveal violations of the coRR pattern on the location z. Both tests considered, we observed the load-load hazard bug on all tested machines.

Others, such as the mp+dmb+fri-rfi-ctrlisb behaviour of Fig. 32, were claimed to be desirable behaviours by the designers that we talked to. This behaviour is a variant of the message passing example, with some more accesses to the flag variable y before

[10]We note that violating coRR invalidates the implementation of C++ modification order mo (e.g. the implementation of `memory_order_relaxed`), which explicitely requires the five coherence patterns of Fig. 6 to be forbidden [Batty et al. 2011, p. 6, col. 1].

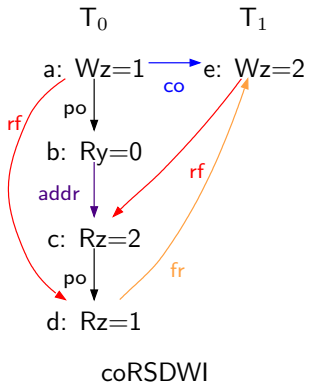


Fig. 31. An observed behaviour that features a coRR violation

	model	machines
coRR	Forbid	Ok, 10M/95G
coRSDWI	Forbid	Ok, 409k/18G
mp+dmb+fri-rfi-ctrlisb	Forbid	Ok, 153k/178G
lb+data+fri-rfi-ctrl	Forbid	Ok, 19k/11G
moredetour0052	Forbid	Ok, 9/17G
mp+dmb+pos-ctrlisb+bis	Forbid	Ok, 81/32G

Table VI. Some counts of invalid observations on ARM machines

the read of the message variable x. We observed this behaviour quite frequently (see the third line of Tab. VI) albeit on one machine (of type APQ8060) only.

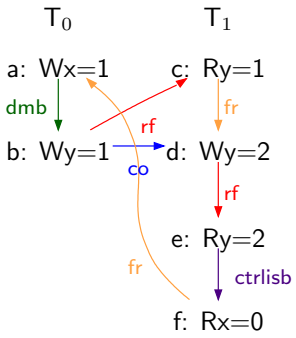
Additionally, we observed similar behaviours on APQ8064 (see also http://moscova.inria.fr/~maranget/cats/model-qualcomm/compare.html#apq8064-invalid). We give three examples of these behaviours in Fig. 33. The first two are variants of the load buffering example of Fig. 7. The last is a variant of the "s" idiom of Fig. 39. We can only assume that these are as desirable as the behaviour in Fig. 32

For reasons explained in the next paragraph, we gather all such behaviours under the name "early commit behaviours" (see reason (ii) in §"Our proposed ARM model").

Our proposed ARM model Our Power-ARM model rejects the mp+dmb+fri-rfi-ctrlisb behaviour via the OBSERVATION axiom, as the event c is ppo-before the event f. More precisely, from our description of preserved program order (see Fig. 25), the order from c to f derives from three reasons: (i) reads are satisfied before they are committed (i(r) precedes c(r)), (ii) instructions that touch the same location commit in order (po-loc is in cc_0), and (iii-a) instructions in a branch that are po-after a control fence (here isb) do not start before the isb executes, and isb does not execute before the branch is settled, which in turn requires the read (e in the diagram) that controls the branch to be committed (ctrl+cfence is in ci_0).

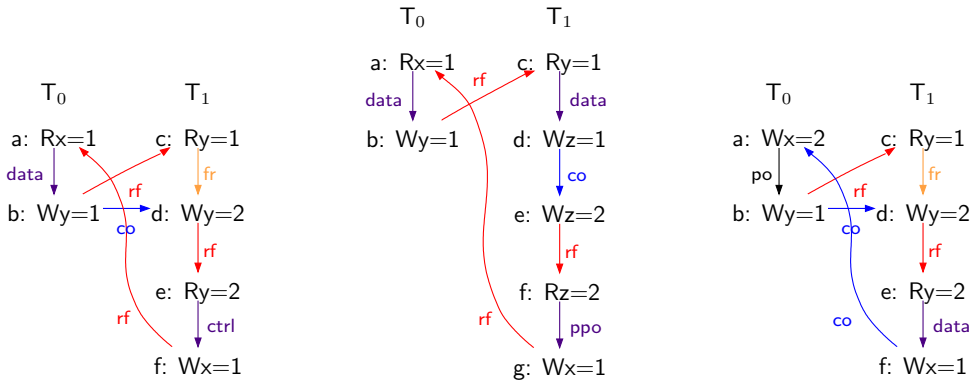
Our Power-ARM model rejects the s+dmb+fri-rfi-data behaviour via the PROPAGATION axiom for the same reason: the event c is ppo-before the event f.

Similarily, our Power-ARM model rejects the lb+data+fri-rfi-ctrl behaviour via the NO THIN AIR axiom, as the event c is ppo-before the event f. Here, still from our description of preserved program order (see Fig. 25) the order from c to f derives from three



mp+dmb+fri-rfi-ctrlisb

Fig. 32. A feature of some ARM machines



lb+data+fri-rfi-ctrl lb+data+data-wsi-rfi-addr s+dmb+fri-rfi-data

Fig. 33. Putative features of some ARM machines

reasons: (i) reads are satisfied before they commit (i(r) precedes c(r)), (ii) instructions that touch the same location commit in order (po-loc is in cc_0), and (iii-b) instructions (in particular store instructions) in a branch do not commit before the branch is settled, which in turn requires the read (e in the diagram) that controls the branch to be committed (ctrl is in cc_0).

Finally, our Power-ARM model rejects the lb+data+data-wsi-rfi-addr via the NO THIN AIR axiom, because because the event c is ppo-before the event g. Again, from our description of preserved program order (see Fig. 25) the order from c to f derives from three reasons: (i) reads are satisfied before they commit (i(r) precedes c(r)), (ii) instructions that touch the same location commit in order (po-loc is in cc_0), and (iii-c) instructions (in particular store instructions) do not commit before a read they depend on (e.g. the read f) is satisfied (addr is in ic_0 because it is in ii_0 and ii_0 is included in ic_0).

The reasons (i) , (iii-a), (iii-b) and (iii-c) seem uncontroversial. In particular for (iii-a), if ctrl+cfence is not included in ppo, then neither the compilation scheme from C++ nor the entry barriers of locks would work [Sarkar et al. 2012]. For (iii-b), if ctrl to a write event is not included in ppo, then some instances of the pattern lb+ppos (see

	Power-ARM	ARM	ARM llh
skeleton	Fig. 5	Fig. 5	Fig. 5s.t. SC PER LOCATION becomes $\mathrm{acyclic}(\mathsf{po\text{-}loc\text{-}llh} \cup \mathsf{com})$, with $\mathsf{po\text{-}loc\text{-}llh} \triangleq \mathsf{po\text{-}loc} \setminus \mathsf{RR}$
propagation	Fig. 18	Fig. 18	Fig. 18
ppo	Fig. 25	Fig. 25 s.t. cc_0 becomes $\mathsf{dp} \cup \mathsf{ctrl} \cup (\mathsf{addr}; \mathsf{po})$	Fig. 25 s.t. cc_0 becomes $\mathsf{dp} \cup \mathsf{ctrl} \cup (\mathsf{addr}; \mathsf{po})$

Table VII. Summary of ARM models

Fig. 7) such as lb+ctrls would be allowed. From the architecture standpoint, this could be explained by value speculation, an advanced feature that current commodity ARM processors do not implement. A similar argument applies for (iii-c), considering this time that lb+addrs should not be allowed by a hardware model. In any case, allowing such simple instances of the pattern lb+ppos would certainly contradict (low-level) programmer intuition.

As for (ii) however, one could argue that this could be explained by an "early commit" feature. For example, looking at mp+dmb+fri-rfi-ctrlisb in Fig. 32 and lb+data+fri-rfi-ctrl in Fig. 33, the read e (which is satisfied by forwarding the local write d) could commit without waiting for the satisfying write d to commit, nor for any other write to the same location that is po-before d. Moreover, the read e could commit without waiting for the commit of any read from the same location that is po-before its satisfying write. We believe that this might be plausible from a micro-architecture standpoint: the value read by e cannot be changed by any later observation of incomming writes performed by load instructions that are po-before the satisfying write d; thus the value read by e is irrevocable.

In conclusion, to allow the behaviours of Fig. 32, we need to weaken the definition of preserved program order of Fig. 25. For the sake of simplicity, we chose to remove po-loc altogether from the cc_0 relation, which is the relation ordering the commit parts of events. This means that two accesses relative to the same location and in program order do not have to commit in this order.[11]

Thus we propose the following model for ARM, which has so far not been invalidated on hardware (barring the load-load hazard behaviours, acknowledged as bugs by ARM [arm 2011] and the other anomalies presented in §"Testing our model", that we take to be undesirable). We go back to the soundness of our ARM model at the end of this section, in §"Remark on our proposed ARM model".

The general skeleton of our ARM model should be the four axioms given in Fig. 5, and the propagation order should be as given in Fig. 18. For the preserved program order, we take it to be as the Power one given in Fig. 25, except for cc_0 which now excludes po-loc entirely, to account for the early commit behaviours, i.e. cc_0 should now be $\mathsf{dp} \cup \mathsf{ctrl} \cup (\mathsf{addr}; \mathsf{po})$. Tab. VII give a summary of the various ARM models that we consider in this section.

Testing our model For the purpose of these experiments only, because our machines suffered from the load-load hazard bug, we removed the read-read pairs from the SC

[11] This is the most radical option; one could choose to remove only $\mathsf{po\text{-}loc} \cap \mathsf{WR}$ and $\mathsf{po\text{-}loc} \cap \mathsf{RR}$, as that would be enough to explain the behaviours of Fig. 32 and similar others. We detail our experiments with alternative formulations for cc_0 at http://diy.inria.fr/cats/arm-anomalies/index.html#alternative. Ultimately we chose to adopt the weakest model since, as we explain in this section, it still exhibits hardware anomalies.

PER LOCATION check as well. This allowed us to have results that were not cluttered by this bug.

We call the resulting model "ARM llh" (ARM load-load hazard), i.e. the model presented in Fig. 5, 18 and 25, where we remove read-read pairs from po-loc in the SC PER LOCATION axiom, thus allowing load-load hazard, and where cc_0 is $\mathsf{dp} \cup \mathsf{ctrl} \cup (\mathsf{addr}; \mathsf{po})$. Tab. VII gives a summary of this model, as well as Power-ARM and our proposed ARM model.

We then compared our original Power-ARM model (i.e. taking literally the definitions of Fig. 5, 18 and 25) and the ARM llh model with hardware:

	ALL	S	T	P	ST	SO	SP	OP	STO	SOP
Power-ARM	37907	21333	842	1133	2471	1130	5561	872	111	4062
ARM llh	1121	105	0	0	0	16	10	460	0	530

Table VIII. Classification of anomalies observed on ARM hardware

More precisely, we classify the executions of both models: for each model we count the number of invalid executions (in the sense of Sec. 4.1). By invalid we mean that an execution is forbidden by the model yet observed on hardware. A given test can have several executions, which explains why the numbers in Tab. VIII are much higher than the number of tests.

The table in Tab. VIII is organised by sets of axioms of our model (note that these sets are pairwise disjoint): "S" is for SC PER LOCATION, "T" for NO THIN AIR, "O" for OBSERVATION, and "P" is for PROPAGATION. For each set of axioms (column) and model (row), we write the number of executions forbidden by said axiom(s) of said model, yet have been observed on ARM hardware. We omit a column (namely O, TO, TP, STP, TOP, STOP) if the counts are 0 for both models.

For example, an execution is counted in the column "S" of the row "Power-ARM" if it is forbidden by the SC PER LOCATION check of Fig. 5 (and allowed by other checks). The column "S" of the row "ARM llh" is SC PER LOCATION minus the read-read pairs. An execution is counted in the column "OP" of the row "Power-ARM" if it is forbidden by both OBSERVATION and PROPAGATION as defined in Fig. 5 (and allowed by other checks); for the row "ARM llh", one needs to take into account the modification to cc_0 mentioned above in the definition of the ppo.

While our original Power-ARM model featured 1500 tests that exhibited 37907 invalid executions forbidden by the model yet observed on ARM machines, those numbers drop to 31 tests and 1121 invalid executions for the ARM llh model (see column "ARM llh", row "ALL"; see also http://diy.inria.fr/cats/relaxed-classify/index.html).

An inspection of each of these anomalies revealed what we believe to be more bugs. We consider the violations of SC PER LOCATION to be particularly severe (see all rows mentioning S). By contrast, the load-load hazard behaviour could be argued to be desirable, or at least not harmful, and was indeed officially allowed by Sparc RMO [sparc 1994] and pre-Power 4 machines [tdf 2002], as we already say in Sec. 4.9.

Fig. 34 shows a violation of SC PER LOCATION. Despite the apparent complexity of the picture, the violation is quite simple. The violation occurs on T_1 wich loads the value 4 from the location y (event f), before writing the value 3 to same location y (event g). However, 4 is the final value of the location y, as the test harness has observed once the test has completed. As a consequence, the event e co-precedes the event f and we witness a cycle $g \xrightarrow{\mathsf{co}} e \xrightarrow{\mathsf{rf}} f \xrightarrow{\mathsf{po\text{-}loc}} g$. That is we witness a violation of the coRW2 pattern (see Sec. 4).

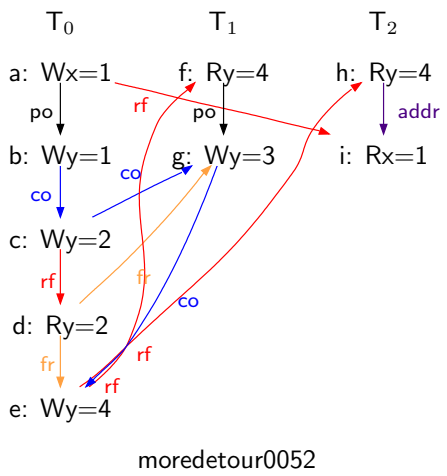


Fig. 34. A violation of SC PER LOCATION observed on ARM hardware

Note that we observed this behaviour rather infrequently, as shown in the fith line of Tab. VI. However, the behaviour is observed on two machines of type Tegra3 and Exynos4412.

In addition to the violation of SC PER LOCATION shown in Fig. 34, we observed the two behaviours given inFig. 35 (rather infrequently, as shown on the last line of Tab. VI, and on one machine only, of type Tegra3), which violate OBSERVATION.

In addition to the violations of SC PER LOCATION shown in Fig. 34, we observed the two behaviours of Fig. 35 (rather infrequently, as shown on the last line of Tab. VI, and on one machine only, of type Tegra3), which violate OBSERVATION.

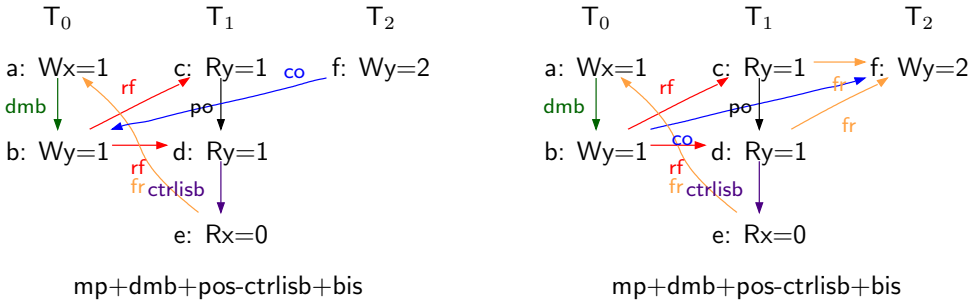


Fig. 35. Two violations of OBSERVATION observed on ARM hardware

The test mp+dmb+pos-ctrlisb+bis includes the simpler test mp+dmb+ctrlisb plus one extra read (c on T_1) and one extra write (f on T_2) of the flag variable y. The depicted behaviours are violations of the mp+dmb+ctrlisb pattern, which must uncontroversially be forbidden. Indeed the only way to allow mp+dmb+ctrlisb is to remove ctrl+cfence from the preserved program order ppo. We have argued above that this would for example break the compilation scheme from C++ to Power (see [Sarkar et al. 2012]).

It is worth noting that we have observed other violations of OBSERVATION on Tegra3, as one can see at http://diy.inria.fr/cats/relaxed-classify/OP.html. For example we have observed mp+dmb+ctrlisb, mp+dmb+addr, mp+dmb.st+addr, which should be uncontroversially forbidden. We tend to classify such observations as bugs of the tested chip. However, since the tested chip exhibits the acknowledged read-after-read hazard bug, the blame can also be put on the impact of this acknowledged bug on our testing infrastructure. Yet this would mean that this impact on our testing infrastructure would reveal on Tegra3 only.

In any case, the interplay between having several consecutive accesses relative to the same location on one thread (e.g. c, d and e on T_1 in mp+dmb+fri-rfi-ctrlisb — see Fig. 32), in particular two reads (c and e), and the message passing pattern mp, seems to pose implementation difficulties (see the violations of OBSERVATION listed in Tab. VIII, in the columns containing "O", and the two examples in Fig. 35).

Remarks on our proposed ARM model Given the state of affairs for ARM, we do not claim our model (see model "ARM" in Tab. V) to be definitive.

In particular, we wonder if the behaviour mp+dmb+fri-rfi-ctrlisb of Fig. 32 can only be implemented on a machine with load-load hazards, which ARM acknowledged to be a flaw (see [arm 2011]), as it involves two reads from the same address.

Nevertheless, our ARM contacts were fairly positive that they would like this behaviour to be allowed. Thus we think a good ARM model should account for it. As to the similar early commit behaviours given in Fig. 33, we can only assume that they should be allowed as well.

Hence our ARM model allows such behaviours, by excluding po-loc from the commit order cc_0 (see Fig. 25 and V). We have performed experiments to compare our ARM model and ARM hardware. To do so, we have excluded the load-load hazard related behaviours.[12]

We give the full comparison table at http://diy.inria.fr/cats/proposed-arm/. As one can see, we still have 31 behaviours that our model forbids yet are observed on hardware (on Tegra 2, Tegra 3 and Exynos 4412).

All of them seem to present anomalies, such as the behaviours that we show in Fig. 34 and 35. We will consult with our ARM contacts for confirmation.

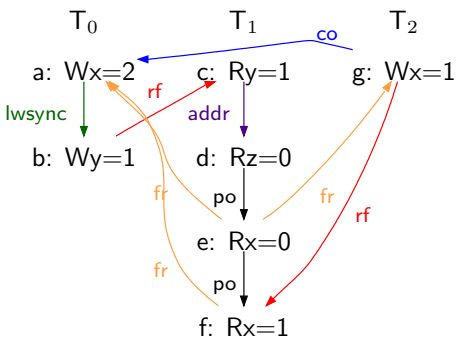
8.2. Experimental comparisons of models

Using the same 8117 and 9761 tests that we used to exercise Power and ARM machines, we have experimentally compared our model to the one of [Sarkar et al. 2011] and the one of [Mador-Haim et al. 2012].

Comparison with the model of [Sarkar et al. 2011]: our experimental data can be found at http://diy.inria.fr/cats/pldi-model. Experimentally, our Power model allows all the behaviours that are allowed by the one of [Sarkar et al. 2011].

We also observe experimentally that our Power model and the one of [Sarkar et al. 2011] differ *only* on the behaviours that [Sarkar et al. 2011] wrongly forbids (see http://diy.inria.fr/cats/pldi-power/#lessvs). We give one such example in Fig. 36: the behaviour mp+lwsync+addr-po-detour is observed on hardware yet forbidden by the model of [Sarkar et al. 2011].

[12]More precisely, we have built the model that only allows load-load hazard behaviours. In herd parlance, this is a model that only has one check: $\mathrm{reflexive}(\mathsf{po\text{-}loc}; \mathsf{fr}; \mathsf{rf})$. We thus filtered the behaviours observed on hardware by including only the behaviours that are not allowed by this load-load hazard model (i.e. all but load-load hazard behaviours). We then compared these filtered hardware behaviours with the ones allowed by our ARM model.



mp+lwsync+addr-po-detour

Fig. 36. A behaviour forbidden by the model of Sarkar et al. but observed on Power hardware

We note that some work is ongoing to adapt the model of [Sarkar et al. 2011] to allow these tests (see http://diy.inria.fr/cats/op-power). This variant of the model of [Sarkar et al. 2011] has so far not been invalidated by the hardware.

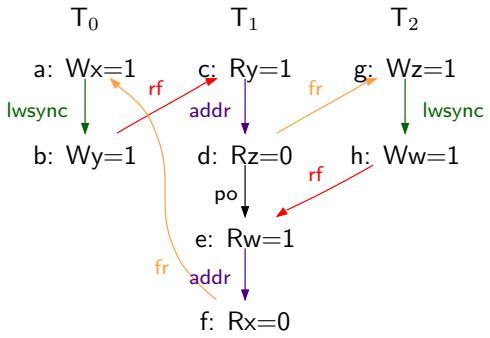
Finally, the model of [Sarkar et al. 2011] forbids the ARM "fri-rfi" behaviours such as the ones given in Fig. 32. Some work is ongoing to adapt the model of [Sarkar et al. 2011] to allow these tests.

Comparison with the model of [Mador-Haim et al. 2012]: our experimental data can be found at http://diy.inria.fr/cats/cav-model. Our Power model and the one of [Mador-Haim et al. 2012] are experimentally equivalent on our set of tests, except for a few tests of similar structure. Our model allows them, whereas the model of [Mador-Haim et al. 2012] forbids them, and they are not observed on hardware. We give the simplest such test in Fig. 37. The test is a refinement of the mp+lwsync+ppo pattern (see Fig. 8). The difference of acceptance between the two models can be explained as follows: the model of [Mador-Haim et al. 2012] does preserve the program order from T_1 initial read c to T_1 final read f, while our model does not. More precisely, the issue reduces to reads d and e (on T_1) being ordered or not. And, indeed, the propagation model for writes of [Mador-Haim et al. 2012] enforces the order, while our definition of ppo does not.

If such a test is intentionally forbidden by the architect, it seems to suggest that one could make the preserved program order of Power (see Fig. 25) stronger. Indeed one could take into account the effect of barriers (such as the one between the two writes g and h on T_2 in the figure above) within the preserved program order.

Yet, we think that one should tend towards more simplicity in the definition of the preserved program order. It feels slightly at odds with our intuition that the preserved program order should take into account dynamic notions such as the propagation order of the writes g and h. By dynamic notions, we here mean notions whose definitions require execution relations such as rf or prop.

As a related side note, although we did include the dynamic relations rdw and detour into the definition of the preserved program order in Fig. 25, we would rather prescribe not to include them. This would lead to a weaker notion of preserved program order, but more stand-alone. By this we mean that the preserved program order would just contain per-thread information (e.g. the presence of a control fence, or a dependency between two accesses), as opposed to external communications such as rfe.



mp+lwsync+addr-bigdetour-addr

Fig. 37. A behaviour allowed by our model while forbidden by [Mador-Haim et al. 2012]

We experimented with a weaker, more static, version of the preserved program order for Power and ARM, where we excluded rdw from ii_0 and detour from ci_0 (see Fig. 25). We give the full experiment report at http://diy.inria.fr/cats/nodetour-model/ On our set of tests, this leads to only 24 supplementary behaviours allowed on Power and 8 on ARM. We believe that this suggests that it might not be worth complicating the ppo for the sake of only a few behaviours being forbidden. Yet it remains to be seen whether these patterns are so common that it is important to determine their precise status w.r.t. a given model.

8.3. Model-level simulation

Simulation was done using our new herd tool: given a model specified in the terms of Sec. 4 and a litmus test, herd computes all the executions allowed by the model. We distribute our tool, its sources and documentation at http://diy.inria.fr/herd.

Our tool herd understands models specified in the style of Sec. 4, i.e. defined in terms of relations over events, and irreflexivity or acyclicity of these relations. For example, Fig. 38 gives the herd model corresponding to our Power model (see Sec. 4 and 6).

We emphasise the concision of Fig. 38, which contains the *entirety* of our Power model.

Language description: we build definitions with `let`, `let rec` and `let rec ... and ...` operators. We build unions, intersections and sequences of relations with "`|`", "`&`" and "`;`" respectively; transitive closure with "`+`", and transitive and reflexive closure with "`*`". The empty relation is "`0`".

We have some built-in relations, e.g. po-loc, rf, fr, co, addr, data, and operators to specify whether the source and target events are reads or writes. For example `RR(r)` gives the relation `r` restricted to both the source and target being reads.

To some extent, the language that herd takes as input shares some similarities with the much broader Lem project [Owens et al. 2011]. However, we merely intend to have a concise way of defining a variety of memory models, whereas Lem aims at (citing the webpage: http://www.cs.kent.ac.uk/people/staff/sao/lem/) *"large scale semantic definitions. It is also intended as an intermediate language for generating definitions from domain-specific tools, and for porting definitions between interactive theorem proving systems."*

```
(* sc per location *) acyclic po-loc|rf|fr|co

(* ppo *)
let dp = addr|data
let rdw = po-loc & (fre;rfe)
let detour = po-loc & (coe;rfe)

let ii0 = dp|rdw|rfi
let ic0 = 0
let ci0 = (ctrl+isync)|detour
let cc0 = dp|po-loc|ctrl|(addr;po)

let rec ii = ii0|ci|(ic;ci)|(ii;ii)
and ic = ic0|ii|cc|(ic;cc)|(ii;ic)
and ci = ci0|(ci;ii)|(cc;ci)
and cc = cc0|ci|(ci;ic)|(cc;cc)
let ppo = RR(ii)|RW(ic)

(* fences *)
let fence = RM(lwsync)|WW(lwsync)|sync

(* no thin air *)
let hb = ppo|fence|rfe
acyclic hb

(* prop *)
let prop-base = (fence|(rfe;fence));hb*
let prop = WW(prop-base)|(com*;prop-base*;sync;hb*)

(* observation *) irreflexive fre;prop;hb*
(* propagation *) acyclic co|prop
```

Fig. 38. herd definition of our Power model

The alloy tool [Jackson 2002] (see also http://alloy.mit.edu/alloy) is closer to herd than Lem. Both alloy and herd allow a concise relational definition of a given system. But while alloy is very general, herd is only targeted at memory models definitions.

Thus one could see herd as a potential front-end to alloy. For example, herd provides some built-in objects (e.g. program order, read-from), that spare the user the effort of defining these objects; alloy would need the user to make these definitions explicit.

More precisely, to specify a memory model in alloy, one would need to explicitly define an object "memory event", for example a record with an identifier, a direction, i.e. write or read, a location and a value, much like we do in our Coq development (see http://www0.cs.ucl.ac.uk/staff/j.alglave/cats).

One would also need to handcraft relations over events (e.g. the program order po), as well as the well-formedness conditions of these relations (e.g. the program order is total order per thread), using first order logic. Our tool herd provides all these basic bricks (events, relations and their well-formedness conditions) to the user.

Finally, alloy uses a SAT solver as a backend, whereas herd uses a custom solver optimised for the limited constraints hat herd supports (namely acyclicity and irreflexivity of relations).

Efficiency of simulation: our axiomatic description underpins herd, which allows for a greater efficiency in the simulation. By contrast, simulation tools based on operational models (e.g. ppcmem [Sarkar et al. 2011], or the tool of [Boudol et al. 2012][13]) are not able to process all tests within the memory bound of 40 GB for [Sarkar et al. 2011] and 6 GB for [Boudol et al. 2012]: ppcmem processes 4704 tests out of 8117; the tool of [Boudol et al. 2012] processes 396 tests out of 518.

Tools based on multi-event axiomatic models (e.g. our reimplementation of [Mador-Haim et al. 2012] inside herd) are able to process all 8117 tests, but require more than eight times the time that our single-event axiomatic model needs.

Tab. IX gives a summary of the number of tests that each tool can process, and the time needed to do so.

tool	model	style	# of tests	(user) time in s
ppcmem	[Sarkar et al. 2011]	operational	4704	14922996
herd	[Mador-Haim et al. 2012]	multi-event axiomatic	8117	2846
—	[Boudol et al. 2012]	operational	396	53100
herd	this model	single-event axiomatic	8117	321

Table IX. Comparison of simulation tools (on Power)

As we have implemented the model of [Mador-Haim et al. 2012][14] and the present model inside herd using the same techniques, we claim that the important gain in run-time efficiency originates from reducing the number of events. On a reduced number of events, classical graph algorithms such as acyclicity test and, more significantly, transitive closure and other fixed point calculations run much faster.

We note that simulation based on axiomatic models outperforms simulation based on operational models. This is mostly due to a state explosion issue, which is aggravated by the fact that Power and ARM are very relaxed architectures. Thus in any given state of the operational machine, there are numerous operational transitions enabled.

We note that ppcmem is not coded as efficiently as it could be. Better implementations are called for, but the distance to herd is considerable: herd is about 45000 times faster than ppcmem, and ppcmem fails to process about half of the tests.

We remark that our single-event axiomatic model also needs several subevents to describe a given instruction (see for example our definition of the preserved program order for Power, in Fig. 25). Yet the opposition between multi-event and single-event axiomatic models lies in the number of events needed to describe the propagation of writes to the system. In multi-event models, there is roughly one propagation event per thread, mimicking the transitions of an operational machine. In single-event models, there is only one event to describe the propagation to several different threads; the complexity of the propagation mechanism is captured through our use of the relations (e.g. rf, co, fr and prop).

We note that single-event axiomatic simulators also suffer from combinatorial explosion. The initial phase computes executions (in the sense of Sec. 4.1) and thus enumerates all possible rf and co relations. However, as clearly shown in Tab. IX, the situation is less severe, and we can still process litmus tests of up to four or five threads.

8.4. Verification of C programs

While assembly-level litmus tests enable detailed study of correctness of the model, the suitability of our model for the verification of high-level programs re-

[13]All the results relative to the tool of [Boudol et al. 2012] are courtesy of Arthur Guillon, who exercised the simulator of [Boudol et al. 2012] on a subset of the tests that we used for exercising the other tools.
[14]The implementations tested in [Mador-Haim et al. 2012] were much less efficient.

mains to be proven. To this effect, we experimented with a modified version of CBMC [Clarke et al. 2004], which is a bounded model-checker for C programs. Recent work [Alglave et al. 2013] has implemented the framework of [Alglave et al. 2010; Alglave et al. 2012] in CBMC, and observed speedups of an order of magnitude w.r.t. other verification tools. CBMC thus features several models, ranging from SC to Power.

In addition, the work of [Alglave et al. 2013] proposes an instrumentation technique, which transforms a concurrent program so that it can be processed by an SC verification tool, e.g. CBMC in SC mode. This relies on an operational model equivalent to the one of [Alglave et al. 2012]; we refer to it in Tab. X under the name "goto-instrument+tool". The advantage of supporting existing tools in SC mode comes at the price of a considerably slower verification time when compared to the implementation of the equivalent axiomatic model within the verification tool, as Tab. X shows.

tool	model	# of tests	time in s
goto-instrument+CBMC (SC)	[Alglave et al. 2012]	555	2511.6
CBMC (Power)	[Alglave et al. 2012]	555	14.3

Table X. Comparison of operational vs. axiomatic model implementation

We adapted the encoding of [Alglave et al. 2013] to our present framework, and recorded the time needed to verify the reachability of the final state of more than 4000 litmus tests (translated to C). As a comparison point, we also implemented the model of [Mador-Haim et al. 2012] in CBMC, and compared the verification times, given in Tab. XI. We observe some speedup with the present model over the implementation of the model of [Mador-Haim et al. 2012].

tool	model	# of tests	time in s
CBMC	[Mador-Haim et al. 2012]	4450	1944
CBMC	present one	4450	1041

Table XI. Comparison of verification tools on litmus tests

We also compared the same tools, but on more fully-fledged examples, described in detail in [Alglave et al. 2013; Alglave et al. 2013]: PgSQL is an excerpt of the PostgreSQL database server software (see http://archives.postgresql.org/pgsql-hackers/2011-08/msg00330.php); RCU is the Read-Copy-Update mechanism of the Linux kernel [McKenney and Walpole 2007], and Apache is a queue mechanism extracted from the Apache HTTP server software. In each of the examples we added correctness properties, described in [Alglave et al. 2013], as assertions to the original source code. We observed that the verification times of these particular examples are not affected by the choice of either of the two axiomatic models, as shown in Tab. XII.

tool	model	PgSQL	RCU	Apache
CBMC	[Mador-Haim et al. 2012]	1.6	0.5	2.0
CBMC	present one	1.6	0.5	2.0

Table XII. Comparison of verification tools on full-fledged examples

9. A PRAGMATIC PERSPECTIVE ON OUR MODELS

To conclude our paper, we put our modelling framework into perspective, in the light of actual software. Quite pragmatically, we wonder whether is it worth going through the effort of defining, on the one hand, then studying or implementing, on the other hand, complex models such as the Power and ARM models that we present in Sec. 6. Are there fragments of these models that are simpler to understand, and embrace pretty much all the patterns that are used in actual software?

For example, there is a folklore notion that iriw (see Fig. 20) is very rarely used in practice. If that is the case, do we need models that can explain iriw?

Conversely, one flaw of the model of [Alglave et al. 2011] (and also of [Boudol et al. 2012]) is that it forbids the pattern r+lwsync+sync (see Fig. 16), against the architect's intent [Sarkar et al. 2011]. While designing the model that we present in the current paper, we found that accounting for this pattern increased the complexity of the model. If this pattern is never used in practice, it might not be worth inventing a model that accounts for it, if it makes the model much more complex.

Thus we ask the following questions: what are the patterns used in modern software? What are their frequencies?

Additionally, we would like to understand whether there are programming patterns used in current software that are not accounted for by our model. Are there programming patterns that are not represented by one of the axioms of our model, i.e. SC PER LOCATION, NO THIN AIR, OBSERVATION or PROPAGATION, as given in Fig. 5?

Conversely, can we understand all the patterns used in current software through the prism of, for example, our OBSERVATION axiom, or is there an actual need for the PROPAGATION axiom too? Finally, we would like to understand to what extent do hardware anomalies, such as the load-load hazard behaviour that we observed on ARM chips (see Sec. 8) impair the behaviour of actual software?

To answer these questions, we resorted to the largest code base available to us: an entire Linux distribution.

What we analysed: we picked the current stable release of the Debian Linux distribution (version 7.1, http://www.debian.org/releases/stable/), which contains more than $17\,000$ software packages (including the Linux kernel itself, server software such as Apache or PostgreSQL, but also user-level software, such as Gimp or Vim).

David A. Wheeler's SLOCCount tool (http://www.dwheeler.com/sloccount/) reports more than 400 million lines of source code in this distribution. C and C++ are still the predominant languages: we found more than 200 million lines of C and more than 129 million lines of C++.

To search for patterns, we first gathered the packages which possibly make use of concurrency. That is, we selected the packages that make use of either POSIX threads or Linux kernel threads anywhere in their C code. This gave us 1590 source packages to analyse; this represents 9.3% of the full set of source packages.

The C language [c11 2011] does not have an explicit notion of shared memory. Therefore, to estimate the number of shared memory interactions, we looked for variables with static storage duration (in the C11 standard sense [c11 2011, §6.2.4]) that were not marked thread local. We found a total of $2\,733\,750$ such variables. In addition to these, our analysis needs to consider local variables shared through global pointers and objects allocated on the heap to obtain an overapproximation of the set of objects (in the C11 standard sense [c11 2011, §3.15]) that may be shared between threads.

A word on C++: the C++ memory model has recently received considerable academic attention (see e.g. [Batty et al. 2011; Sarkar et al. 2012; Batty et al. 2013]). Yet to date even a plain text search in all source files for uses of the corresponding `stdatomic.h`

and `atomic` header files only reveals occurrences in the source code of compilers, but not in any of the other source packages.

Thus practical assessment of our subset of the C++ memory model is necessarily left for future work. At the same time, this result reinforces our impression that we need to study hardware models to inform current concurrent programming.

9.1. Static pattern search

To look for patterns in Debian 7.1, we implemented a static analysis in a new tool called mole. This means that we are looking for an overapproximation of the patterns used in the program. Building on the tool chain described in [Alglave et al. 2013], we use the front-end goto-cc and a variant of the goto-instrument tool of [Alglave et al. 2013], with the new option --static-cycles. We distribute our tool mole, along with a documentation, at http://diy.inria.fr/mole.

9.1.1. Preamble on the goto- tools* goto-cc and goto-instrument are part of the tool chain of CBMC [Clarke et al. 2004], which is widely recognised for its maturity.[15] goto-cc may act as compiler substitute as it accepts the same set of command line options as several C compilers, such as GCC. Instead of executables, however, goto-cc compiles C programs to an intermediate representation shared by the tool chain around CBMC: *goto-programs*. These goto-programs can be transformed and inspected using goto-instrument. For instance, goto-instrument can be applied to insert assertions of generic invariants such as valid pointer dereferencing or data race checks, or dump goto-programs as C code. Consequently we implemented the search described below in goto-instrument, adding the new option --static-cycles.

9.1.2. Cycles Note that an pattern like all the ones that we have presented in this paper corresponds to a cycle of the relations of our model. This is simply because our model is defined in terms of irreflexivity and acyclicity checks. Thus looking for patterns corresponds here to looking for cycles of relations.

Critical cycles Previous works [Shasha and Snir 1988; Alglave and Maranget 2011; Bouajjani et al. 2011; Bouajjani et al. 2013] show that a certain kind of cycles, which we call *critical cycles* (following [Shasha and Snir 1988]), characterises many weak behaviours. Intuitively, a critical cycle violates SC in a minimal way.

We recall here the definition of a critical cycle (see [Shasha and Snir 1988] for more details). Two events x and y are *competing*, written $(x, y) \in \mathsf{cmp}$, if they are from distinct processors, to the same location, and at least one of them is a write (e.g. in iriw, the write a to x on T_0 and the read b from x on T_2). A cycle $\sigma \subseteq (\mathsf{cmp} \cup \mathsf{po})^+$ is critical when it satisfies the following two properties:

— (i) per thread, there are at most two memory accesses involved in the cycle on this thread and these accesses have distinct locations, and
— (ii) for a given memory location ℓ, there are at most three accesses relative to ℓ, and these accesses are from distinct threads ($(w, w') \in \mathsf{cmp}$, $(w, r) \in \mathsf{cmp}$, $(r, w) \in \mathsf{cmp}$ or $\{(r, w), (w, r')\} \subseteq \mathsf{cmp}$).

All the executions that we give in Sec. 4 show critical cycles, except for the SC PER LOCATION ones (see Fig. 6). Indeed a critical cycle has to involve more than one memory location by definition.

Static critical cycles More precisely, our tool mole looks for cycles which:

— alternate program order po and competing accesses cmp,

[15] http://www.research.ibm.com/haifa/conferences/hvc2011/award.shtml

— traverse a thread only once (see (i) above), and
— involve at most three accesses per memory location (see (ii) above).

Observe that the definition above is not limited to the well-known patterns that we presented in Sec. 4. Consider the two executions in Fig. 39, both of which match the definition of a critical cycle given above.

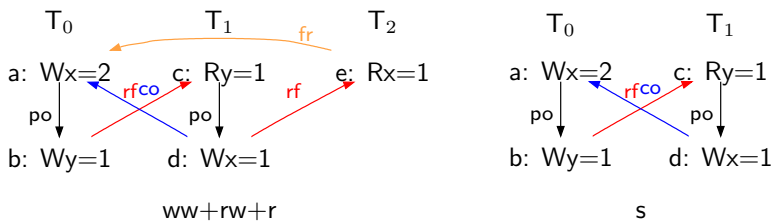


Fig. 39. The pattern s (on the right), and an extended version of it (on the left)

On the left-hand side, the thread T_1 writes 1 to location x (see event d); the thread T_2 reads this value from x (i.e. $(d, e) \in \mathsf{rf}$), before the thread T_0 writes 2 to x (i.e. $(e, a) \in \mathsf{fr}$). By definition of fr, this means that the write d of value 1 into x by T_1 is co-before the write a of value 2 into x by T_0. This is reflected by the execution on the right-hand side, where we simply omitted the reading thread T_2.

Thus to obtain our well-known patterns, such as the ones in Sec. 4 and the s pattern on the right-hand side of Fig. 39, we implement the following reduction rules, which we apply to our cycles:

— $\mathsf{co}; \mathsf{co} = \mathsf{co}$, which means that we only take the extremities of a chain of coherence relations;
— $\mathsf{rf}; \mathsf{fr} = \mathsf{co}$, which means that we omit the intermediate reading thread in a sequence of read-from and from-read, just like in the s case above;
— $\mathsf{fr}; \mathsf{co} = \mathsf{fr}$ which means that we omit the intermediate writing thread in a sequence of from-read and coherence.

We call the resulting cycles *static critical cycles*.

Thus mole looks for all the static critical cycles that it can find in the goto-program given as argument. In addition, it also looks for SC PER LOCATION cycles, i.e. coWW, coRW1, coRW2, coWR and coRR as shown in Fig. 6.

In the remainder of this section, we simply write *cycles* for the cycles gathered by mole, i.e. static critical cycles and SC PER LOCATION cycles.

9.1.3. Static search Looking for patterns poses several challenges, which are also pervasive in static data race analysis (see [Kahlon et al. 2007]):

— identify program fragments that may be run concurrently, in distinct threads;
— identify objects that are shared between these threads.

Finding shared objects may be further complicated by the presence of inline assembly. We find inline assembly in 803 of the packages to be analysed. At present, mole only interprets a subset of inline assembly deemed relevant for concurrency, such as memory barriers, but ignores all other inline assembly.

We can now explain how our pattern search works. Note that our approach does not require analysis of whole, linked, programs – which is essential to achieve scalability to a code base this large. Our analysis proceeds as follows:

1. identify candidate functions that could act as *entry points* for a thread being spawned (an entry point is a function such that its first instruction will be scheduled for execution when creating a thread);
2. group these candidate entry points, as detailed below, according to shared objects accessed – where we consider an object as *shared* when it either has static storage duration (and is not marked thread local), or is referenced by a pointer that is shared;
3. assuming concurrent execution of the threads in each such group, enumerate patterns using the implementation from [Alglave et al. 2013] with a flow-insensitive points-to analysis and, in order to include SC PER LOCATION cycles, weaker restrictions than when exclusively looking for critical cycles;
4. classify the candidates following their patterns (i.e. using the litmus naming scheme that we outlined in Fig. III) and the axioms of the model. The categorisation according to axioms proceeds by testing the sequence of relations occurring in a cycle against the axioms of Fig. 5; we detail this step below.

Note that our analysis does not take into account program logic, e.g. locks, that may forbid the execution of a given cycle. If no execution of the (concurrent) program includes a certain cycle, we call it a *spurious* cycle, and refer to others as *genuine* cycles. Note that this definition is independent of fences or dependencies that may render a cycle forbidden for a given (weak) memory model. Our notion of genuine simply accounts for the feasibility of a concurrent execution. This overapproximation means that any numbers of cycles given in this section cannot be taken as a quantitative analysis of cycles that would be actually executed.

With this approach, instead, we focus on not missing cycles rather than avoiding the detection of spurious cycles. In this sense, the results are best compared to compiler warnings. Performing actual proofs of cycles being either spurious or genuine is an undecidable problem. In principle we could thus only do so in a best-effort manner, akin to all software verification tools aiming at precise results. In actual practice, however, the concurrent reachability problem to be solved for each such cycle will be a formidable challenge for current software verification tools, including several additional technical difficulties (such as devising complex data structures as input values) as we are looking at real-world software rather than stylised benchmarks. With more efficient tools such as the one of [Alglave et al. 2013] we hope to improve on this situation in future work, since with the tool of [Alglave et al. 2013] we managed to verify selected real-world concurrent systems code for the first time.

We now explain these steps in further detail, and use the Linux Read-Copy-Update (RCU) code [McKenney and Walpole 2007] as an example. Fig. 40 shows a code snippet, which was part of the benchmarks that we used in [Alglave et al. 2013], employing RCU. The original code contains several macros, which were expanded using the C preprocessor.

Finding entry points To obtain an overapproximate set of patterns even for, e.g. library code, which does not have a defined entry point and thus may be used in a concurrent context even when the code parts under scrutiny do contain thread spawn instructions, we consider thread entry points as follows:

— explicit thread entries via POSIX or kernel thread create functions;
— any set of functions $f_1, \ldots, f_n$, provided that f_i is not (transitively) called from another function f_j in this set and f_i has external linkage (see [c11 2011, §5.1.1.1]).;
— for mutually recursive functions an arbitrary function from this set of recursive functions.

```
01 struct foo *gbl_foo;
02
03 struct foo foo1, foo2;
04
05 spinlock_t foo_mutex = (spinlock_t) { { .rlock = { .raw_lock = { 0 }, } } };
06
06 void* foo_update_a(void* new_a)
07 {
08   struct foo *new_fp;
09   struct foo *old_fp;
10
11   foo2.a=100;
12   new_fp = &foo2;
13   spin_lock(&foo_mutex);
14   old_fp = gbl_foo;
15   *new_fp = *old_fp;
16   new_fp->a = *(int*)new_a;
16
17   ({ __asm__ __volatile__ ("lwsync" " " : : :"memory");
18      ((gbl_foo)) = (typeof(*(new_fp)) *)((new_fp)); });
19
20   spin_unlock(&foo_mutex);
21   synchronize_rcu();
22   return 0;
23 }
24
25 void* foo_get_a(void* ret)
26 {
26   int retval;
27   rcu_read_lock();
28   retval = ({ typeof(*(gbl_foo)) *_________p1 =
29               (typeof(*(gbl_foo))* )(*(volatile typeof((gbl_foo)) *)&((gbl_foo)));
30               do { } while (0); ; do { } while(0);
31               ((typeof(*(gbl_foo)) *)(_________p1)); })->a;
32   rcu_read_unlock();
33   *(int*)ret=retval;
34   return 0;
35 }
36
36 int main()
37 {
38   foo1.a=1;
39   gbl_foo=&foo1;
40   gbl_foo->a=1;
41
42   int new_val=2;
43   pthread_create(0, 0, foo_update_a, &new_val);
44   static int a_value=1;
45   pthread_create(0, 0, foo_get_a, &a_value);
46
46   assert(a_value==1 || a_value==2);
47 }
```

Fig. 40. Code example from RCU

For any function identified as entry point we create 3 threads, thereby accounting for multiple concurrent access to shared objects only used by a single function, but also for cases where one of the called functions is running in an additional thread.

For RCU, we see several functions (or function calls) in Fig. 40: `main`, `foo_get_a`, `foo_update_a`, `spin_lock`, `spin_unlock`, `synchronize_rcu`, `rcu_read_lock` and `rcu_read_unlock`. If we discard `main`, we no longer have a defined entry point nor POSIX thread creation through `pthread_create`. In this case, our algorithm would consider `foo_get_a` and `foo_update_a` as the only potential thread entry points, because all other functions are called from one of these two, and there is no recursion.

Finding threads' groups Then we form groups of threads using the identified thread entry points. We group the functions f_i and f_j if and only if the set of objects read or written by f_i or any of the functions (transitively) called by f_i has a non-empty intersection with the set for f_j. Note the transitivity in this requirement: for functions f_i, f_j, f_k with f_i and f_j sharing one object, and f_j and f_k sharing another object, all three functions end up in one group. In general, however, we may obtain several groups of such threads, which are then analysed individually.

When determining shared objects, as noted above, pointer dereferencing has to be taken into account. This requires the use of points-to analyses, for which we showed that theoretically they can be sound [Alglave et al. 2011], even under weak memory models. In practice, however, pointer arithmetic, field-sensitivity, and interprocedural operation require a performance-precision trade-off. In our experiments we use a flow-insensitive, field-insensitive and interprocedural analysis. We acknowledge that we may thus still be missing certain cycles due to the incurred incompleteness of the points-to analysis.

For RCU, `main`, `foo_get_a` and `foo_update_a` form a group, because they jointly access the pointer `gbl_foo` as well as the global objects `foo1` and `foo2` through this pointer. Furthermore `main` and `foo_update_a` share the local `new_val`, and `main` and `foo_get_a` share `a_value`, both of which are communicated via a pointer.

Finding patterns With the thread groups established, we enumerate patterns as in [Alglave et al. 2013]. We briefly recall this step here for completeness:

— we first construct one control-flow graph (CFG) per thread;
— then we add communication edges between shared memory accesses to the same object, if at least one of these objects is a write (this is the cmp relation in the definition of critical cycles given at the beginning of this section);
— we enumerate all cycles amongst the CFGs and communication edges using Tarjan's 1973 algorithm [Tarjan 1973], resulting in a set that also contains all critical cycles (but possibly more);
— as final step we filter the set of cycles for those that satisfy the conditions of static critical cycles or SC PER LOCATION as described above.

Let us explain how we may find the mp pattern (see Fig. 8 in Sec. 4) in RCU. The writing thread T_0 is given by the function `foo_update_a`, the reading thread T_1 by the function `foo_get_a`. Now for the code of the writing thread T_0: in `foo_update_a`, we write `foo2` at line 11, then we have an lwsync at line 17, and a write to `gbl_foo` at line 18.

For the code of the reading thread T_1: the function `foo_get_a` first copies the value of `gbl_foo` at line 29 into the local variable `_____p1`. Now, note that the write to `gbl_foo` at line 18 made `gbl_foo` point to `foo2`, due to the assignment to `new_fp` at line 12.

Thus dereferencing `_____p1` at line 31 causes a read of the object `foo2` at that line. Observe that the dereferencing introduces an address dependency between the read of `gbl_foo` and the read of `foo2` on T_1.

Categorisation of cycles As we said above, for each cycle that we find, we apply a categorisation according to the axioms of our model (see Fig. 5). For the purpose of this categorisation we instantiate our model for SC (see Fig. 21). We use the sequence of relations in a given cycle: for example for mp, this sequence is lwsync; rfe; dp; fre. We first test if the cycle is a SC PER LOCATION cycle: we check if all the relations in our input sequence are either po-loc or com. If, as for mp, this is not the case, we proceed with the test for NO THIN AIR. Here we check if all the relations in our sequence match hb, i.e. po $\cup$ fences $\cup$ rfe. As mp includes an fre, the cycle cannot be categorised as NO THIN AIR, and we proceed to OBSERVATION. Starting from fre we find lwsync $\in$ prop (as prop $=$ po $\cup$ fences $\cup$ rf $\cup$ fr on SC), and rfe; dp $\in$ hb*. Thus we categorise the cycle as as a observation cycle. In the general case, we check for PROPAGATION last.

Litmus tests: we exercised mole on the set of litmus tests that we used for exercising CBMC (see Sec. 8.4). For each test we find its general pattern (using the naming scheme that we presented in Sec. 4): for example for mp we find the cycle po; rfe; po; fre. Note that our search looks for memory barriers but does not try to look for dependencies; this means that for the variant mp+lwfence+addr of the mp pattern, we find the cycle lwfence; rfe; po; fre, where the barrier appears but not the dependency.

Examples that we had studied manually before (see [Alglave et al. 2013; Alglave et al. 2013]) include Apache, PostgreSQL and RCU, as mentioned in Sec. 8.4. We analysed these examples with mole to confirm the patterns that we had found before.[16]

In Apache we find 5 patterns distributed over 75 cycles: $4 \times$ mp (see Fig. 8); $1 \times$ s (see Fig. 39); $28 \times$ coRW2, $25 \times$ coWR, and $17 \times$ coRW1 (see Fig. 6).

In PostgreSQL, we find 22 different patterns distributed over 463 cycles. We give the details in Tab. XIII.

In RCU we find 9 patterns in 23 critical cycles, as well as one SC PER LOCATION cycle. We give the details in Tab. XIV. For each pattern we give one example cycle: we refer to the excerpt in Fig. 40 to give the memory locations and line numbers[17] it involves. Note that we list an additional example of mp in the table, different from the one explained above.

9.2. Results for Debian 7.1

We report on the results of running mole on $137\,163$ object files generated while compiling source packages using goto-cc, in $1\,251$ source packages of the Debian Linux distribution, release 7.1.

We provide all the files compiled with goto-cc at http://diy.inria.fr/mole, and present our experimental data (i.e. the patterns per packages) at http://diy.inria.fr/mole.

Our experiment runs on a system equipped with 8 cores and 64 GB of main memory. In our setup, we set the time and memory bounds for each object file subject to static cycle search to 15 minutes and 16 GB of RAM. We spent more than 199 CPU days in cycle search, yet $19\,930$ runs did not finish within the above time and memory limits. More than 50% of time are spent within cycle enumeration for a given graph of CFGs and communication edges, whereas only 12% are spent in the points-to analysis. The remaining time is consumed in generating the graph. The resulting 79 GB of raw data were further processed to determine the results presented below.

[16]In the following we mention patterns that we have not displayed in the paper, and which do not follow the convention outlined in Fig. III: z6.[0-5], 3.2w, or irrwiw. For the sake of brevity, we do not show them in the paper, and refer the reader to the companion webpage: http://diy.inria.fr/doc/gen.html#naming.

[17]Lines 1 and 3 result from initialisation of objects with static storage duration, as prescribed by the C11 standard [c11 2011, §6.7.9].

pattern	# cycles
r (see Fig. 16)	93
w+rr+2w	68
w+rr+wr+ww	62
z6.4	54
sb (see Fig. 14)	37
2+2w (see Fig. 13(a))	25
w+rwc (see Fig. 19)	23
mp (see Fig. 8)	16
w+rw	14
s (see Fig. 39)	14
z6.5	6
w+rw+wr	6
w+rw+2w	4
z6.0	2
wrc (see Fig. 11)	2
lb (see Fig. 7)	2
irrwiw	2
coWR (see Fig. 6)	19
coWW (see Fig. 6)	6
coRW1 (see Fig. 6)	4
coRW2 (see Fig. 6)	4

Table XIII. Patterns in PostgreSQL

pattern	# cycles	memory locations	line numbers
2+2w (see Fig. 13(a))	6	`foo2`, `gbl_foo`	1, 3, 16
3.2w	4	`foo1`, `foo2`, `gbl_foo`	3, 16, 39, 40
w+rr+ww+ww	3	`foo1`, `foo2`, `gbl_foo`	3, 14, 15, 16, 39
z6.5	2	`foo1`, `foo2`, `gbl_foo`	3, 11, 14, 39, 40
r (see Fig. 16)	2	`foo1`, `foo2`	3, 11, 15
mp (see Fig. 8)	2	`foo1`, `gbl_foo`	14, 15, 38, 39
w+rr+ww+wr	2	`foo1`, `foo2`, `gbl_foo`	3, 11, 14, 29, 31, 39
z6.3	1	`foo1`, `foo2`, `gbl_foo`	3, 16, 29, 31
w+rr+2w	1	`foo1`, `gbl_foo`	29, 31, 38, 39
coWW (see Fig. 6)	1	`foo2`	15 16

Table XIV. Patterns in RCU

9.2.1. General results We give here some general overview of our experiments. We detected a total of 86 206 201 critical cycles, plus 11 295 809 SC PER LOCATION cycles. Amongst these, we find 551 different patterns. Fig. 41 gives the thirty most frequent patterns.

The source package with most cycles is mlterm (a multilingual terminal emulator, http://packages.debian.org/wheezy/mlterm, 4 261 646 cycles) with the most frequently occurring patterns iriw (296 219), w+rr+w+rr+w+rw (279 528) and w+rr+w+rw+w+rw (218 061) The source package with the widest variety of cycles is ayttm (an instant messaging client, http://packages.debian.org/wheezy/ayttm, 238 different patterns); its most frequent patterns are z6.4 (162 469), z6.5 (146 005) and r (90 613).

We now give an account of what kind of patterns occur for a given functionality. By functionality we mean what the package is meant for, e.g. web servers (httpd), mail clients (mail), video games (games) or system libraries (libs). For each functionality,

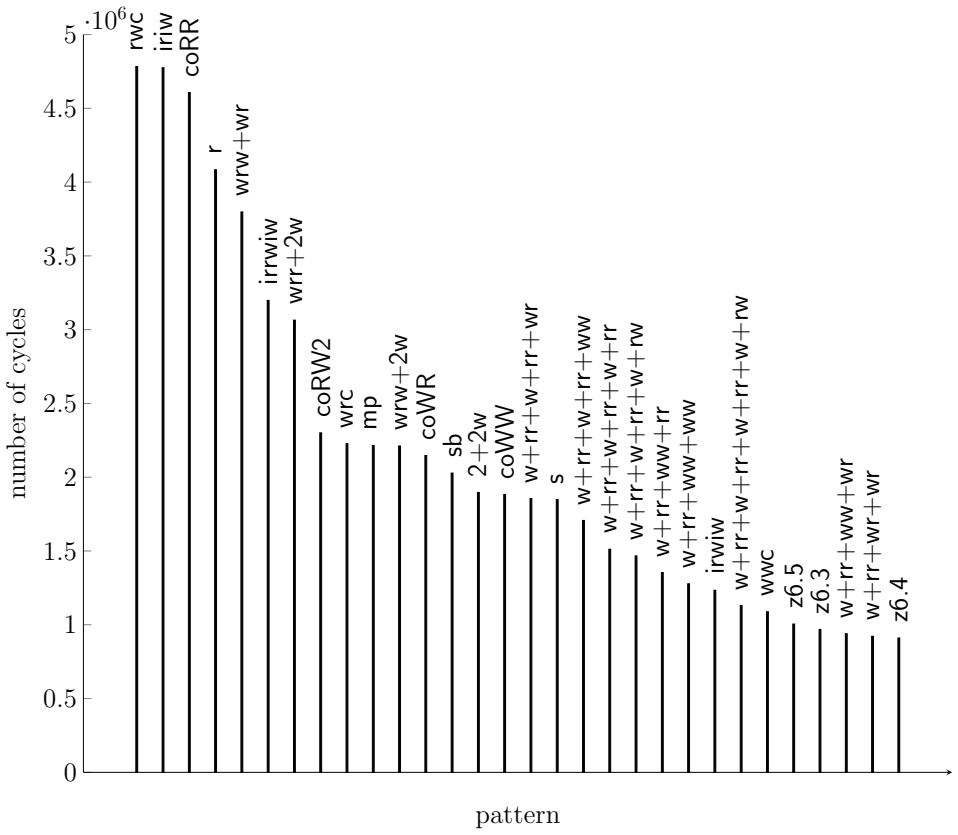


Fig. 41. Thirty most frequent patterns

Tab. XV gives the number of packages (e.g. 11 for httpd), the three most frequent patterns within that functionality with their number of occurrences (e.g. 30 506 for wrr+2w in httpd) and typical packages with the number of cycles contained in that package (e.g. 70 283 for apache2).

function	patterns	typical packages
httpd (11)	wrr+2w (30 506), mp (27 618), rwc (13 324)	libapache2-mod-perl2 (120 869), apache2 (70 283), webfs (27 260)
mail (24)	w+rr+w+rw+ww (75 768), w+rr+w+rr+ww (50 842), w+rr+w+rr+w+rw (45 496)	opendkim (702 534), citadel (337 492), alpine (105 524)
games (57)	2+2w (198 734), r (138 961), w+rr+w+rr+wr (134 066)	spring (1 298 838), gcompris (559 905), liquidwar (257 093)
libs (266)	iriw (468 053), wrr+2w (387 521), irrwiw (375 836)	ecore (1 774 858), libselinux (469 645), psqlodbc (433 282)

Table XV. Patterns per functionality

9.2.2. Summary per axiom Tab. XVI gives a summary of the patterns we found, organised per axioms of our model (see Sec. 4). We chose a classification with respect to SC, i.e. we fixed prop to be defined as shown in Fig. 21. For each axiom we also give some typical examples of packages that feature patterns relative to this axiom.

axiom	# patterns	typical packages
SC PER LOCATION	11 295 809	vips (412 558), gauche (391 180), python2.7 (276 991)
NO THIN AIR	445 723	vim (36 461), python2.6 (25 583), python2.7 (16 213)
OBSERVATION	5 786 239	mlterm (285 408), python2.6 (183 761), vim (159 319)
PROPAGATION	79 974 239	isc-dhcp (891 673), cdo (889 532), vim (878 289)

Table XVI. Patterns per axiom

We now give a summary of the patterns we found, organised by axioms. Several distinct patterns can correspond to the same axiom, e.g. mp, wrc and isa2 all correspond to the OBSERVATION axiom (see Sec. 4). For the sake of brevity, we do not list all the 551 patterns. Fig. 42 gives one pie chart of patterns per axiom.

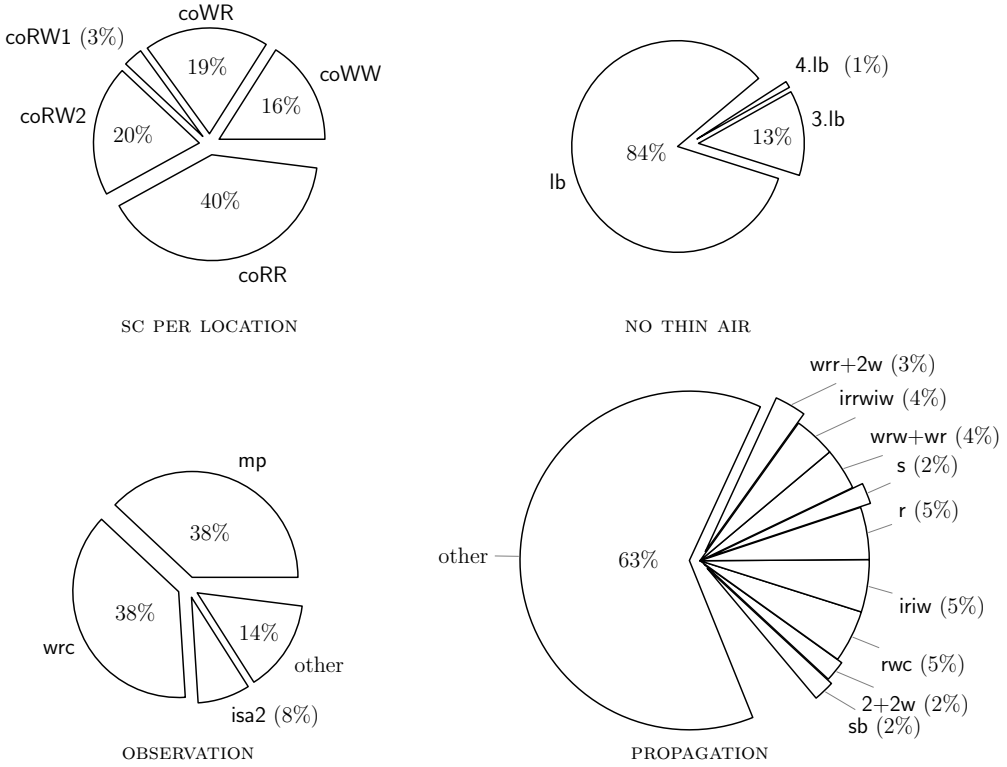


Fig. 42. Proportions of patterns per axiom

Observations We did find 4 775 091 occurrences of iriw (see Fig. 20), which represents 4.897% of all the static cycles that we have found. As such it is the second most frequent pattern detected, which appears to invalidate the folklore claim that iriw is rarely used. It should be noted, however, that these static cycles need not correspond to genuine ones that are actually observed in executions, as discussed further below.

We found 4 083 639 occurrences of r (see Fig. 16), which represents 4.188% of the static cycles that we have found. Observe that r appears in PostgreSQL (see Tab. XIII) and RCU (see Tab. XIV), as well as in the thirty most frequent patterns (see Fig. 41). This seems to suggest that a good model needs to handle this pattern properly.

We also found 4 606 915 occurrences of coRR, which corresponds to the acknowledged ARM bug that we presented in Sec. 8. This represents 4.725% of all the static cycles that we have found. Additionally, we found 2 300 724 occurrences of coRW2, which corresponds to the violation of SC PER LOCATION, observed on ARM machines that we show in Fig. 34. This represents 2.360% of all the static cycles that we have found. These two percentages perhaps nuance the severity of the ARM anomalies that we expose in Sec. 8.

We believe that our experiments with mole provide results that could be used by programmers or static analysis tools to identify where weak memory may come into play and ensure that it does not introduce unexpected behaviours. Moreover, we think that the data that mole gathers can be useful to both hardware designers and software programmers.

While we do provide quantitative data, we would like to stress that, at present, we have little information on how many of the detected cycles are actually genuine. Many of the considered cycles may be spurious, either because of additional synchronisation mechanisms such as locks, or simply because the considered program fragments are not executed concurrently in any concrete execution. Thus, as said above, at present the results are best understood as warnings similar to those emitted by compilers. In future work we will both work towards the detection of spurious cycles, but also aim at studying particular software design patterns that may give rise to the most frequently observed patterns of our models.

We nevertheless performed manual spot tests of arbitrarily selected cycles in the packages 4store, acct and acedb. For instance, the SC PER LOCATION patterns in the package acct appear genuine, because the involved functions could well be called concurrently as they belong to a memory allocation library. An analysis looking at the entire application at once would be required to determine whether this is the case. For other cases, however, it may not at all be possible to rule out such concurrent operations: libraries, for instance, may be used by arbitrary code. In those cases only locks (or other equivalent mutual exclusion primitives) would guarantee data-race free (and thus weak-memory insensitive) operation. The same rationale applies for other examples that we looked at: while our static analysis considers this case in order to achieve the required safe overapproximation, the code involved in some of the iriw cycles in the package acedb or 4store is not obviously executed concurrently at present. Consequently these examples of iriw might be spurious, but we note that no locks or fences are in place to guarantee this.

For the RCU and PostgreSQL examples presented in this paper we use harnesses that perform concurrent execution. For RCU this mimics the intended usage scenario of RCU (concurrent readers and writers), and in the case of PostgreSQL this was modelled after a regression test[18] built by PostgreSQL's developers. Consequently we are able to tell apart genuine and spurious cycles in those cases.

[18] See the attachment at http://www.postgresql.org/message-id/24241.1312739269@sss.pgh.pa.us

For PostgreSQL, the SC PER LOCATION patterns (coWW, coWR, coRW1, coRW2, listed in Tab. XIII) and the critical cycles described in detail in [Alglave et al. 2013] are genuine: one instance of lb (amongst the 2 listed in Tab. XIII) and one instance of mp (amongst the 16 listed).

For RCU, the coWW cycle liste in Tab. XIV and the instance of mp described above (see top of page 67) are genuine – but note that the latter contains synchronisation using lwsync, which means that the cycle is forbidden on Power.

All other cycles are spurious, because the cycle enumeration based on Tarjan's 1973 algorithm does not take into account the ordering of events implied by spawning threads. For example, we report instances of mp in RCU over lines 38 and 39 in function `main` as first thread, and lines 14 and 15 in function `foo_update_a` as second, and seemingly concurrent, thread. As that second thread, however, is only spawned after execution of lines 38 and 39, no such concurrency is possible.

10. CONCLUSION

To close this paper, we recapitulate the criteria that we listed in the introduction, and explain how we address each of them.

Stylistic proximity of models: the framework that we presented embraces a wide variety of hardware models, including SC, x86-TSO, Power and ARM. We also explained how to instantiate our framework to produce a significant fragment of the C++ memory model, and we leave the definition of the complete model (in particular including consume atomics) for future work.

Concision is demonstrated in Fig. 38, which contains the *unabridged* specification of our Power model.

Efficient simulation and verification: the performance of our tools, our new simulator herd (see Fig. IX), and the bounded model-checker CBMC adapted to our new Power model (see Fig. XI), confirm (following [Mador-Haim et al. 2012; Alglave et al. 2013]) that the modelling style is crucial.

Soundness w.r.t. hardware: to the best of our knowledge, our Power and ARM models are to this date not invalidated by hardware, except for the 33 surprising ARM behaviours detailed in Sec. 8. Moreover, we keep on running experiments regularly, which we record at http://diy.inria.fr/cats.

Adaptability of the model was demonstrated by the ease with which we were able to modify our model to reflect the subtle mp+dmb+fri-rfi-ctrlisb behaviour (see Sec. 8).

Architectural intent: to the best of our knowledge, our Power model does not contradict the architectural intent, in that we build on the model of [Sarkar et al. 2011], which should reflect said intent, and that we have regular contacts with hardware designers.

For ARM, we model the mp+dmb+fri-rfi-ctrlisb behaviour which is claimed to be intended by ARM designers.

Account for what programmers do: with our new tool mole, we explored the C code base of the Debian Linux distribution version 7.1 (about 200 millions lines of code) to collect statistics of concurrency patterns occurring in real world code.

Just like our experiments on hardware, we keep on running our experiments on Debian regularly; we record them at http://diy.inria.fr/mole.

***

As future work, on the modelling side, we will integrate the semantics of the lwarx and stwcx Power instructions (and their ARM equivalents ldrex and strex), which are

used to implement locking or compare-and-swap primitives. We expect their integration to be relatively straight forward: the model of [Sarkar et al. 2012] uses the concept of a write reaching coherence point to describe them, a notion that we have in our model as well.

On the rationalist side, it remains to be seen if our model is well-suited for proofs of programs: we regard our experiments w.r.t. verification of programs as preliminary.

ACKNOWLEDGMENTS

We thank Nikos Gorogiannis for suggesting that the extension of the input files for herd should be .cat. We thank Carsten Fuhs (even more so since we forgot to thank him in [Alglave et al. 2013]) and Matthew Hague for their patient and careful comments on a draft. We thank Mark Batty, Viktor Vafeiadis and Tyler Sorensen for comments on a draft. We thank our reviewers for their careful reading, comments and suggestions. We thank Arthur Guillon for his help with the simulator of [Boudol et al. 2012]. We thank Susmit Sarkar, Peter Sewell and Derek Williams for discussions on the Power model(s). Finally, this paper would not have been the same without the last year of discussions on related topics with Richard Bornat, Alexey Gotsman, Peter O'Hearn and Matthew Parkinson.

REFERENCES

1992 and 1994. SPARC Architecture Manual Versions 8 and 9.

2002. Power 4 system microarchitecture. In *IBM J. Res. & Dev. Vol. 46(1)*.

2009. *Power ISA Version 2.06*.

2011. Cortex-A9 MPCore, Programmer Advice Notice, Read-after-Read Hazards. In *ARM Reference 761319*.

2011. Information technology – Programming languages – C. In *BS ISO/IEC 9899:2011*.

April 2008. ARM Architecture Reference Manual (ARMv7). (April 2008).

Parosh Aziz Abdulla, Mohamed Faouzi Atig, Yu-Fang Chen, Carl Leonardsson, and Ahmed Rezine. 2012. Counter-Example Guided Fence Insertion under TSO. In *TACAS*.

Parosh Aziz Abdulla, Mohammed Faouzi Atig, Yu-Fang Chen, Carl Leonardsson, and Ahmed Rezine. 2013. Memorax, a Precise and Sound Tool for Automatic Fence Insertion under TSO. In *TACAS*.

Allon Adir, Hagit Attiya, and Gil Shurek. 2003. Information-Flow Models for Shared Memory with an Application to PowerPC. In *TPDS*.

Sarita Adve and Hans Boehm. 2010. Memory Models: A Case for Rethinking Parallel Languages and Hardware. In *CACM*.

Sarita Adve and Kourosh Gharachorloo. 1995. Shared Memory Consistency Models: A Tutorial. In *IEEE Computer*.

Jade Alglave. 2010. *A Shared Memory Poetics*. Ph.D. Dissertation. Université Paris 7.

Jade Alglave. 2012. A Formal Hierarchy of Memory Models. In *FMSD*.

Jade Alglave, Anthony Fox, Samin Ishtiaq, Magnus Myreen, Susmit Sarkar, Peter Sewell, and Francesco Zappa Nardelli. 2009. The Semantics of Power and ARM Multiprocessor Machine Code. In *DAMP*.

Jade Alglave, Daniel Kroening, John Lugton, Vincent Nimal, and Michael Tautschnig. 2011. Soundness of Data Flow Analyses for Weak Memory Models. In *APLAS*.

Jade Alglave, Daniel Kroening, Vincent Nimal, and Michael Tautschnig. 2013. Software Verification for Weak Memory via Program Transformation. In *ESOP*.

Jade Alglave, Daniel Kroening, and Michael Tautschnig. 2013. Partial Orders for Efficient Bounded Model-Checking of Concurrent Software. In *CAV*.

Jade Alglave and Luc Maranget. 2011. Stability in Weak Memory Models. In *CAV*.

Jade Alglave, Luc Maranget, Susmit Sarkar, and Peter Sewell. 2010. Fences in Weak Memory Models. In *CAV*.

Jade Alglave, Luc Maranget, Susmit Sarkar, and Peter Sewell. 2011. Litmus: Running Tests Against Hardware. In *TACAS*.

Jade Alglave, Luc Maranget, Susmit Sarkar, and Peter Sewell. 2012. Fences in Weak Memory Models (Extended Version). In *FMSD*.

alpha 2002. Alpha Architecture Reference Manual, Fourth Edition. (2002).

ARM. 2009. ARM Barrier Litmus Tests and Cookbook. (Nov. 2009).
Arvind and Jan-Willem Maessen. 2006. Memory Model = Instruction Reordering + Store Atomicity. In *ISCA*.
Mohammed Faouzi Atig, Ahmed Bouajjani, Sebastian Burckhardt, and Madan Musuvathi. 2010. On the Verification Problem for Weak Memory Models. In *POPL*.
Mohammed Faouzi Atig, Ahmed Bouajjani, Sebastian Burckhardt, and Madan Musuvathi. 2012. What's Decidable About Weak Memory Models?. In *ESOP*.
Mohamed Faouzi Atig, Ahmed Bouajjani, and Gennaro Parlato. 2011. Getting Rid of Store-Buffers in the Analysis of Weak Memory Models. In *CAV*.
Mark Batty, Mike Dodds, and Alexey Gotsman. 2013. Library abstraction for C/C++ concurrency. In *POPL*.
Mark Batty, Scott Owens, Susmit Sarkar, Peter Sewell, and Tjark Weber. 2011. Mathematising C++ Concurrency. In *POPL*.
Yves Bertot and Pierre Casteran. 2004. *Interactive Theorem Proving and Program Development: Coq'Art: The Calculus of Inductive Constructions*. Springer Verlag, EATCS.
Hans Boehm and Sarita Adve. 2008. Foundations of the C++ Concurrency Memory Model. In *PLDI*.
Hans Boehm and Sarita Adve. 2012. You don't know jack about shared variables or memory models. In *CACM*.
Ahmed Bouajjani, Egor Derevenetc, and Roland Meyer. 2013. Checking Robustness against TSO. In *ESOP*.
Ahmed Bouajjani, Roland Meyer, and Eike Moehlmann. 2011. Deciding Robustness Against Total Store Ordering. In *ICALP*.
Gérard Boudol and Gustavo Petri. 2009. Relaxed memory models: an operational approach. In *POPL*.
Gérard Boudol, Gustavo Petri, and Bernard Serpette. 2012. Relaxed Semantics of Concurrent Programming Languages. In *Express/SOS*.
Sebastian Burckhardt, Rajeev Alur, and Milo Martin. 2007. CheckFence: Checking Consistency of Concurrent Data Types on Relaxed Memory Models. In *PLDI*.
Sebastian Burckhardt, Alexey Gotsman, and Hongseok Yang. 2013. Understanding eventual consistency. In *MSR TR-2013-39*.
Sebastian Burckhardt, Alexey Gotsman, Hongseok Yang, and Marek Zawirski. 2014. Replicated data types: specification, verification, optimality. In *POPL*.
Sebastian Burckhardt and Madan Musuvathi. 2008. Memory Model Safety of Programs. In *(EC)*^{2}.
Pietro Cenciarelli, Alexander Knapp, and Eleonora Sibilio. 2007. The Java Memory Model: Operationally, Denotationally, Axiomatically. In *ESOP*.
Nathan Chong and Samin Ishtiaq. 2008. Reasoning about the ARM weakly consistent memory model. In *MSPC*.
Edmund M. Clarke, Daniel Kroening, and Flavio Lerda. 2004. A Tool for Checking ANSI-C Programs. In *TACAS*.
William Collier. 1992. *Reasoning About Parallel Architectures*. Prentice-Hall.
K. Gharachorloo, D. Lenoski, J. Laudon, P. Gibbons, A. Gupta, and J. Hennessy. Memory Consistency and Event Ordering in Scalable Shared-Memory Multiprocessors. In *ISCA 1990*.
Jacob Goodman. March 1989. Cache consistency and Sequential consistency. In *TR 61, IEEE Scalable Coherent Interface Group*.
Ganesh Gopalakrishnan, Yue Yang, and Hemanthkumar Sivaraj. 2004. QB or not QB: An Efficient Execution Verification Tool for Memory Orderings. In *CAV*. Springer.
Michael Gordon. 2002. Relating Event and Trace Semantics of Hardware Description Languages. In *Comput. J. 45, 1 (2002), 27-36*.
Sudheendra Hangal, Durgam Vahia, Chaiyasit Manovit, Juin-Yeu Joseph Lu, and Sridhar Narayanan. 2004. TSOTool: A Program for Verifying Memory Systems Using the Memory Consistency Model. In *ISCA*.
C. A. R. Hoare and Peter E. Lauer. 1974. Consistent and Complementary Formal Theories of the Semantics of Programming Languages. In *Acta Inf. 3: 135-153*.
David Howells and Paul E. MacKenney. 2013. Linux Kernel Memory Barriers, 2013 version. (2013). https://www.kernel.org/doc/Documentation/memory-barriers.txt.

intel 2009. Intel 64 and IA-32 Architectures Software Developer's Manual, rev. 30. (March 2009).

2002. A Formal Specification of Intel Itanium Processor Family Memory Ordering. (Oct. 2002). Intel Document 251429-001.

Daniel Jackson. 2002. Alloy: a lightweight object modelling notation. *ACM Trans. Softw. Eng. Methodol.* 11, 2 (2002), 256–290.

Vineet Kahlon, Yu Yang, Sriram Sankaranarayanan, and Aarti Gupta. 2007. Fast and Accurate Static Data-Race Detection for Concurrent Programs. In *CAV*.

Michael Kuperstein, Martin Vechev, and Eran Yahav. 2010. Automatic Inference of Memory Fences. In *FMCAD*.

Michael Kuperstein, Martin Vechev, and Eran Yahav. 2011. Partial-Coherence Abstractions for Relaxed Memory Models. In *PLDI*.

Leslie Lamport. 1979. How to Make a Correct Multiprocess Program Execute Correctly on a Multiprocessor. *IEEE Trans. Comput.* (1979).

Richard J. Lipton and Jonathan S. Sandberg. September 1988. PRAM: a scalable shared memory. In *Princeton University TR CS-TR-180-88*.

Feng Liu, Nayden Nedev, Nedyalko Prisadnikov, Martin Vechev, and Eran Yahav. 2012. Dynamic Synthesis for Relaxed Memory Models. In *PLDI*.

Sela Mador-Haim, Rajeev Alur, and Milo K. Martin. 2010. Generating Litmus Tests For Contrasting Memory Consistency Models. In *CAV*.

Sela Mador-Haim, Luc Maranget, Susmit Sarkar, Kayvan Memarian, Jade Alglave, Scott Owens, Rajeev Alur, Milo Martin, Peter Sewell, and Derek Williams. 2012. An Axiomatic Memory Model for Power Multiprocessors. In *CAV*.

Jeremy Manson, William Pugh, and Sarita Adve. 2005. The Java Memory Model. In *POPL*.

Paul E. McKenney and Jonathan Walpole. 2007. What is RCU, fundamentally? (2007). http://lwn.net/Articles/262464/.

Gil Neiger. October 2000. A Taxonomy of Multiprocessor Memory-Ordering Models. In *Tutorial and Workshop on Formal Specification and Verification Methods for Shared Memory Systems*.

Scott Owens, Peter Böhm, Francesco Zappa Nardelli, and Peter Sewell. 2011. Lem: A Lightweight Tool for Heavyweight Semantics. In *ITP*.

Scott Owens, Susmit Sarkar, and Peter Sewell. 2009. A better x86 model: x86-TSO. In *TPHOL*.

Susmit Sarkar, Kayvan Memarian, Scott Owens, Mark Batty, Peter Sewell, Luc Maranget, Jade Alglave, and Derek Williams. 2012. Synchronising C/C++ and Power. In *PLDI*.

Susmit Sarkar, Peter Sewell, Jade Alglave, Luc Maranget, and Derek Williams. 2011. Understanding Power multiprocessors. In *PLDI*.

Susmit Sarkar, Peter Sewell, Francesco Zappa Nardelli, Scott Owens, Tom Ridge, Thomas Braibant, Magnus Myreen, and Jade Alglave. 2009. The Semantics of x86-CC Multiprocessor Machine Code. In *POPL*.

Dennis Shasha and Marc Snir. 1988. Efficient and Correct Execution of Parallel Programs that Share Memory. In *TOPLAS*.

sparc 1994. SPARC Architecture Manual Version 9. (1994).

Robert Tarjan. 1973. Enumeration of the Elementary Circuits of a Directed Graph. *SIAM J. Comput.* (1973).

Emina Torlak, Mandana Vaziri, and Julian Dolby. 2010. MemSAT: Checking Axiomatic Specifications of Memory Models. In *PLDI*.

Yue Yang, Ganesh Gopalakrishnan, Gary Linstrom, and Konrad Slind. 2004. Nemos: A Framework for Axiomatic and Executable Specifications of Memory Consistency Models. In *IPDPS*.

Francesco Zappa Nardelli, Peter Sewell, Jaroslav Sevcik, Susmit Sarkar, Scott Owens, Luc Maranget, Mark Batty, and Jade Alglave. 2009. Relaxed memory models must be rigorous. In *(EC)*^{2}.