

Proving Skipping Refinement with ACL2s

Mitesh Jain Panagiotis Manolios

Northeastern University

{jmitesh,pete}@ccs.neu.edu *

We describe three case studies illustrating the use of ACL2s to prove the correctness of optimized reactive systems using skipping refinement. Reasoning about reactive systems using refinement involves defining an abstract, high-level *specification* system and a concrete, low-level *implementation* system. Next, one shows that the behaviors of the implementation system are allowed by the specification system. Skipping refinement allows us to reason about implementation systems that can “skip” specification states due to optimizations that allow the implementation system to take several specification steps at once. Skipping refinement also allows implementation systems to *stutter*, *i.e.*, to take several steps before completing a specification step. We show how ACL2s can be used to prove skipping refinement theorems by modeling and proving the correctness of three systems: a JVM-inspired stack machine, a simple memory controller, and a scalar to vector compiler transformation.

1 Introduction

Refinement is a powerful method for reasoning about reactive systems. The idea is that a simple high-level abstract system acts as a specification for a low-level implementation of a concrete system. The goal is then to prove that all observable behaviors of the concrete system are behaviors of the abstract system. It is often the case that the concrete system requires several steps to match one high-level step of the abstract system, a phenomenon commonly known as stuttering. Therefore, notions of refinement usually directly account for stuttering [2, 5, 8]. However, in the course of engineering an efficient implementation, it is often the case that a single step of the concrete system can correspond to several steps of the abstract system, a phenomenon that is dual of stuttering. For example, in order to reduce memory latency and effectively utilize memory bandwidth, memory controllers often buffer requests to memory. The pending requests in the buffer are analyzed for address locality and then at some time in the future, multiple locations in the memory are read and updated simultaneously. Similarly, to improve instruction throughput, superscalar processors fetch multiple instructions in a single clock cycle. These instructions are analyzed for instruction-level parallelism (*e.g.*, the absence of data dependencies), and where possible multiple instructions are executed in parallel, retired in a single clock cycle. In both the above examples, updating multiple locations in memory and retiring multiple instructions in a single clock cycle, results in scenario where a single step in the optimized implementation may correspond to multiple steps in the abstract system. A notion of refinement that only account for stuttering is therefore not appropriate for reasoning about such optimized systems.

In our companion paper [10], we proposed *skipping refinement*, a new notion of correctness for reasoning about optimized reactive systems and a proof method that is amenable for mechanical reasoning. The applicability of skipping refinement was shown using three case studies: a JVM-inspired stack machine, an optimized memory controller, and a vectorizing compiler transformation. In [10] we focused on finite-state models for the systems in the first two case studies and used model-checkers to verify

*This research was supported in part by DARPA under AFRL Cooperative Agreement No. FA8750-10-2-0233, by NSF grants CCF-1117184 and CCF-1319580, and by OSD under contract FA8750-14-C-0024.

skipping refinement. In this paper, we consider their corresponding infinite-state models and prove their correctness in ACL2s, an interactive theorem prover [7]. We also discuss in detail the modeling of vectorizing compiler transformation and its proof of correctness. In Section 2, we motivate the need for a new notion of refinement with an example. In Section 3, we define well-founded skipping simulation. In Section 4 we discuss the three case studies. We end the paper with conclusion and future work in Section 5.

2 Motivating Example

To illustrate the notion of skipping simulation, we consider an example of a discrete-time event simulation (DES) system [10]. An abstract high-level specification of DES is described as follows. Let E be set of events and V be set of variables. Then a state of abstract DES is a three-tuple $\langle t, Sch, A \rangle$, where t is a natural number denoting current time; Sch is a set of pairs (e, t_e) , where e is an event scheduled to be executed at time $t_e \geq t$; A is an assignment to variables in V . The transition relation for the abstract DES system is defined as follows. If at time t there is no $(e, t) \in Sch$, *i.e.*, there is no event scheduled to be executed at time t , then t is incremented by 1. Else, we (nondeterministically) choose and execute an event of the form $(e, t) \in Sch$. The execution of event may result in modifying A and also adding finite number of new pairs (e', t') in Sch . We require that $t' > t$. Finally execution involves removing the executed event (e, t) from Sch .

Now, consider an optimized, concrete implementation of the abstract DES system. As before, a state of the concrete system is a three-tuple $\langle t, Sch, A \rangle$. However, unlike the abstract system which just increments time by 1 when no events are scheduled for the current time, the optimized system uses a priority queue to find the next event to execute. The transition relation is defined as follows. An event (e, t_e) with the minimum time is selected, t is updated to t_e and the event e is executed, as in the abstract DES.

Notice that when no events are scheduled for execution at the current time, the optimized implementation of the discrete-time event simulation system can run faster than the abstract specification system by *skipping* over abstract states. This is not a stuttering step as it results in an observable change in the state of the concrete DES system (t is update to t_e). Also, it does not correspond to a single step of the specification. Therefore, it is not possible to prove that the implementation *refines* the specification using notions of refinement that only allow stuttering [2, 13], because that just is not true. But, intuitively, there is a sense in which the optimized DES system *does* refine the abstract DES system. The notion of skipping refinement proposed in [10] is an appropriate notion to relate such systems: a low-level implementation that can run slower (stutter) or run faster (skip) than the high-level specification.

3 Skipping Refinement

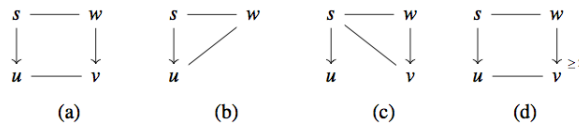
In this section, we first present the notion of well-founded skipping simulation [10]. The notion is defined in the general setting of labeled transition systems (TS) where labeling is on states.¹ We also place no restriction on the state space sizes and the branching factor, and both can be of arbitrary infinite cardinalities. The generality of TS is useful to model systems that may exhibit unbounded nondeterminism, for example, modeling a program in a language with random assignment command $x = ?$, which sets x to an arbitrary integer [3].

¹Note that labeled transition system are also used in the literature to refer to transition systems where transitions (edges) are labeled. However, we prefer to work with TS where states are labeled.

We first describe the notational conventions used in the paper. Function application is sometimes denoted by an infix dot “.” and is left-associative. For a binary relation R , we often use the infix notation xRy instead of $(x,y) \in R$. The composition of relation R with itself i times (for $0 < i \leq \omega$) is denoted R^i ($\omega = \mathbb{N}$ and is the first infinite ordinal). Given a relation R and $1 < k \leq \omega$, $R^{<k}$ denotes $\bigcup_{1 \leq i < k} R^i$ and $R^{\geq k}$ denotes $\bigcup_{\omega > i \geq k} R^i$. Instead of $R^{<\omega}$ we often write the more common R^+ . \uplus denotes the disjoint union operator. Quantified expressions are written as $\langle Qx : r : p \rangle$, where Q is the quantifier (e.g., \exists, \forall), x is the bound variable, r is an expression that denotes the range of x (*true*, if omitted), and p is the body of the quantifier.

Definition 1 A labeled transition system (TS) is a structure $\langle S, \rightarrow, L \rangle$, where S is a non-empty (possibly infinite) set of states, $\rightarrow \subseteq S \times S$ is a left-total transition relation (every state has a successor), and L is a labeling function: its domain is S and it tells us what is observable at a state.

Skipping refinement is defined based on well-founded skipping simulation, a notion that is amenable for mechanical reasoning. This notion allows us to reason about skipping refinement by checking mostly local properties, *i.e.*, properties involving states and their successors. The intuition is, for any pair of states s, w , which are related and a state u such that $s \rightarrow u$, there are four cases to consider (Definition 3): (a) either we can match the move from s to u right away, *i.e.*, there is a v such that $w \rightarrow v$ and u is related to v , or (b) there is stuttering on the left, or (c) there is stuttering on the right, or (d) there is skipping on the left.



Definition 2 (Well-founded Skipping) $B \subseteq S \times S$ is a well-founded skipping relation on a transition system $\mathcal{M} = \langle S, \rightarrow, L \rangle$ iff:

(WFSK1) $\langle \forall s, w \in S : sBw : L.s = L.w \rangle$

(WFSK2) There exist functions, $rankt : S \times S \rightarrow W$, $rankl : S \times S \times S \rightarrow \omega$, such that $\langle W, \prec \rangle$ is well-founded and

$\langle \forall s, u, w \in S : s \rightarrow u \wedge sBw :$

(a) $\langle \exists v : w \rightarrow v : uBv \rangle \vee$

(b) $\langle uBw \wedge rankt(u, w) \prec rankt(s, w) \rangle \vee$

(c) $\langle \exists v : w \rightarrow v : sBv \wedge rankl(v, s, u) < rankl(w, s, u) \rangle \vee$

(d) $\langle \exists v : w \rightarrow^{\geq 2} v : uBv \rangle \rangle$

In the above definition, conditions (WFSK2a) to (WFSK2c) require reasoning only about single step \rightarrow of the transition system. But condition (WFSK2d) requires us to check that there exists a v such that v is *reachable* from w in two or more steps and uBv holds. Reasoning about reachability, in general, is not local. However, for the kinds of optimized systems we are interested in the number of abstract steps that a concrete step corresponds to is bounded by a constant—a bound determined early on in the design phase. For example, the maximum number of abstract steps that a concrete step of a superscalar processor can correspond to is the number of instruction that the designer decides to retire in a single cycle. This is

a constant that is decided early on in the design phase. Therefore, for such systems we can still reason using “local” methods. Furthermore, in the case this constant is a “small” number, condition (WFKS2d) can be checked by simply unrolling the transition relation of the concrete system, an observation that we exploit in our first two case studies. On the other hand, this simplification is not always possible. For example, in the optimized DES system describe above, notice that number of abstract steps that optimized DES can take corresponds to the difference between current time and earliest time an event is scheduled for execution. This difference can not be a priori bounded by a constant.

We now define the notion of skipping refinement, a notion that relates *two* transition systems: an *abstract* transition system and a *concrete* transition system. In order to define skipping refinement, we make use of *refinement maps*, functions that map states of the concrete system to states of the abstract system. Informally, if the concrete system is a skipping refinement of the abstract system then its observable behaviors are also behavior of the abstract system modulo skipping (which includes stuttering). For example, in our running example of DES, if the refinement map is the identity function then it is easy to see that any behavior of the optimized system is a behavior of the abstract system modulo skipping. In practice, the abstract system and the concrete system are described at different levels of abstraction. Refinement maps along with the labeling function enable us to define what is observable at concrete states from the view point of the abstract system.

Definition 3 (Skipping Refinement) Let $\mathcal{M}_A = \langle S_A, \xrightarrow{A}, L_A \rangle$ and $\mathcal{M}_C = \langle S_C, \xrightarrow{C}, L_C \rangle$ be transition systems and let $r: S_C \rightarrow S_A$ be a refinement map. We say \mathcal{M}_C is a skipping refinement of \mathcal{M}_A with respect to r , written $\mathcal{M}_C \lesssim_r \mathcal{M}_A$, if there exists a relation $B \subseteq S_C \times S_A$ such that all of the following hold.

1. $\langle \forall s \in S_C :: sBr.s \rangle$ and
2. B is an WFSK on $\langle S_C \uplus S_A, \xrightarrow{C} \uplus \xrightarrow{A}, \mathcal{L} \rangle$ where $\mathcal{L}.s = L_A(s)$ for $s \in S_A$, and $\mathcal{L}.s = L_C(r.s)$ for $s \in S_C$.

Well-founded skipping gives us a simple proof rule to determine if a concrete transition system \mathcal{M}_C is a skipping refinement of an abstract transition system \mathcal{M}_A with respect to a refinement map r . Given a refinement map $r: S_C \rightarrow S_A$ and relation $B \subseteq S_C \times S_A$, we check the following two conditions: (a) for all $s \in S_C$, $sBr.s$ and (b) if B is a WFSK on the disjoint union of \mathcal{M}_C and \mathcal{M}_A . If (a) and (b) hold, $\mathcal{M}_C \lesssim_r \mathcal{M}_A$. For a more detailed exposition of skipping refinement we refer the reader to our companion paper [10].

Notice that we place no restrictions on refinement maps. When refinement is used in specific contexts it is often useful to place restrictions on what a refinement map can do, *e.g.*, we may require for every $s \in S_C$ that $L_A(r.s)$ is a projection of $L_C(s)$. The generality of refinement map is useful in all three case studies considered in this paper, where a simple refinement map that is a projection function would not have sufficed.

4 Case Studies

We consider three case studies. The first case study is a hardware implementation of a JVM-inspired stack machine with an instruction buffer. The second case study is a memory controller with an optimization that eliminates redundant writes to memory. The third case study is a compiler transformation that vectorizes a list of scalar instructions. For each case study we model the abstract system and the concrete system in ACL2s. We define an appropriate refinement map and prove that the implementation refines the specification using well-founded skipping simulation.

We first briefly list some conventions used to describe the syntax and the semantics of the systems. Adding element e to the beginning or end of a list (or an array) l is denoted by $e::l$ and $l::e$, respectively.

Each transition consists of a *state* :*condition*₁, . . . , *condition*_{*n*} pair above a line, followed by the next state below the line. If a concrete state matches the state in a transition and satisfies each of the conditions, then the state can transition to the state below the line. We formalize the operational semantics of the machines by describing the effect of each instruction on the state of the machine. The proof scripts are publicly available [1].

4.1 JVM-inspired Stack Machine

In this case study, we verify a stack machine inspired by the Java Virtual Machine (JVM). Java processors have been proposed as an alternative to just-in-time compilers to improve the performance of Java programs. Java processors such as JME [9] fetch bytecodes from an instruction memory and store them in an instruction buffer. The bytecodes in the buffer are analyzed to perform instruction-level optimizations *e.g.*, instruction folding. In this case study, we verify BSTK, a simple hardware implementation of part of the JVM. BSTK is an incomplete and inaccurate model of JVM that only models an instruction memory, an instruction buffer and a stack. Only a small subset of JVM instructions are supported (push, pop, top, nop). However, even such a simple model is sufficient to exhibit the applicability of skipping simulation and the limitations of current hardware model-checking tools.

STK is the high-level specification with respect to which we verify the correctness of BSTK, the implementation. Their behaviors are defined using abstract transition systems. The syntax and the operational semantics are shown in Fig. 1.

The state of STK consists of an instruction memory *imem*; a program counter *pc*; and a stack *stk*. An instruction is one of *push*, *pop*, *top*, and *nop*. We use the `listof` combinator in `defdata` to encode the instruction memory as list of instructions and stack as a list of elements [6]. The program counter is encoded as a natural number using the primitive data type `nat`. We then compound these components to encode state of STK using the `defdata` `record` construct. The `defdata` framework introduces a constructor function `sstate`, a set of accessor functions for each field (*e.g.*, `sstate-imem`), a recognizer function `sstatep` identifying the state, an enumerator `nth-sstate` and several useful theorems to reason about compositions of these functions.

```
(defdata el all)

(defdata stack (listof el))

(defdata inst (oneof (list 'pop) (list 'top)
                    (list 'nop) (list 'push el)))

(defdata inst-mem (listof inst))

(defdata sstate (record (imem . inst-mem)
                       (pc . nat)
                       (stk . stack)))
```

STK fetches an instruction from the instruction memory, executes it, increases the program counter, and possibly modifies the stack, as outlined in Fig. 1.

STK fetches an instruction from the *imem*, executes it, increments the *pc* by 1, and possibly modifies the *stk*, as outlined in Fig. 1. Since STK is a deterministic machine, we formalize its transition relation using a function `spec-step`, which uses an auxiliary function `stk-step-inst` to capture the effect of

$stk := [] \mid el :: stk$	(Stack)
$inst := \langle push\ e \rangle \mid \langle pop \rangle \mid \langle top \rangle \mid \langle nop \rangle$	(Instruction)
$imem := [] \mid inst :: imem$	(Program)
$pc := 0 \mid 1 \mid \dots \mid n \mid \dots$	(Program Counter)
$ibuf := [inst_1, \dots, inst_k]$	(Instruction Buffer)
$sstate := \langle imem, pc, stk \rangle$	(STK State)
$istate := \langle imem, pc, ibuf, stk \rangle$	(BSTK State)

STK (\xrightarrow{A}) where $s = \text{capacity of } stk, t = |stk|$

$$\frac{\langle imem, pc, stk \rangle : imem[pc] = \langle push\ v \rangle, t < s}{\langle imem, pc + 1, v :: stk \rangle}$$

$$\frac{\langle imem, pc, stk \rangle : imem[pc] = \langle push\ v \rangle, t = s}{\langle imem, pc + 1, stk \rangle}$$

$$\frac{\langle imem, pc, [] \rangle : imem[pc] = \langle pop \rangle}{\langle imem, pc + 1, [] \rangle}$$

$$\frac{\langle imem, pc, v :: stk \rangle : imem[pc] = \langle pop \rangle}{\langle imem, pc + 1, stk \rangle}$$

$$\frac{\langle imem, pc, stk \rangle : imem[pc] = \langle top \rangle}{\langle imem, pc + 1, stk \rangle}$$

$$\frac{\langle imem, pc, stk \rangle : imem[pc] = \langle nop \rangle}{\langle imem, pc + 1, stk \rangle}$$

$$\frac{\langle imem, pc, stk \rangle : imem[pc] = nil}{\langle imem, pc + 1, stk \rangle}$$

BSTK (\xrightarrow{C}) where $k = \text{capacity of } ibuf, m = |ibuf|$

$$\frac{\langle imem, pc, ibuf, stk \rangle : m < k, imem[pc] \neq \langle top \rangle}{\langle imem, pc + 1, ibuf :: imem[pc], stk \rangle}$$

$$\frac{\langle imem, pc, ibuf, stk \rangle : imem[pc] = \langle top \rangle, \langle ibuf, 0, stk \rangle \xrightarrow{A}^m \langle ibuf, m, stk' \rangle}{\langle imem, pc + 1, [], stk' \rangle}$$

$$\frac{\langle imem, pc, ibuf, stk \rangle : imem[pc] = nil, \langle ibuf, 0, stk \rangle \xrightarrow{A}^m \langle ibuf, m, stk' \rangle}{\langle imem, pc + 1, [], stk' \rangle}$$

$$\frac{\langle imem, pc, ibuf, stk \rangle : m = k, \langle ibuf, 0, stk \rangle \xrightarrow{A}^m \langle ibuf, m, stk' \rangle}{\langle imem, pc + 1, [imem[pc]], stk' \rangle}$$

Figure 1: Syntax and Semantics of Stack and Buffered Stack Machine

executing an instruction on the stack. We are now ready to define the transition system \mathcal{M}_A of STK machine. The set of states S_A in the transition system \mathcal{M}_A is the set of all states satisfying the predicate `sstatep`. Two states $s, u \in S_A$ are related by transition relation \xrightarrow{A} iff it is possible in one step to transition from s to u , i.e., $u = (\text{spec-step } s)$. The labeling function, L_A is the identity function.

```
(defun stk-step-inst (inst stk)
  "returns next state of stk"
  (let ((op (car inst)))
    (cond ((equal op 'push)
           (mpush (cadr inst) stk ))
          ((equal op 'pop)
           (mpop stk))
          ((equal op 'top)
           (mtop stk))
          (t stk))))

(defun spec-step (s)
  "single step of STK machine"
  (let* ((pc (sstate-pc s))
         (imem (sstate-imem s))
         (inst (nth pc imem))
         (stk (sstate-stk s)))
    (if (instp inst)
        (sstate imem (1+ pc) (stk-step-inst inst stk))
        (sstate imem (1+ pc) stk))))
```

The state of BSTK is similar to STK, except that it also includes an instruction buffer `ibuf`. The instruction buffer is encoded as a list of instructions with an additional restriction on its capacity (`ibuf-capacity`). To encode `ibuf` in the `defdata` framework, we have at least two choices. We can use the `oneof` `defdata` construct to encode it as an empty list or list of one, two, or three instructions. Another way is to use the capability of the `defdata` framework to define custom data types. In the later case, we first define a recognizer function `inst-buffp` and an enumerator function `nth-inst-buff`.

```
(defun inst-buffp (l)
  (and (inst-memp l)
        (<= (len l) (ibuf-capacity))))

(defun nth-inst-buff (n)
  (let ((imem (nth-inst-mem n)))
    (if (<= (len imem) (ibuf-capacity))
        imem
        (let ((i1 (car imem))
              (i2 (cadr imem))
              (i3 (caddr imem)))
          (list i1 i2 i3)))))
```

We can now register our custom type `inst-buff` using the `register-custom-type` macro. Once we have registered it as a `defdata` type we can use it just like other type directly introduced using `defdata` construct.

```
(register-custom-type inst-buff :enumerator nth-inst-buff
                             :predicate inst-buffp)
```

We can now define state of BSTK machine using `defdata record` construct.

```
(defdata istate
  (record (imem . inst-mem)
          (pc . nat)
          (stk . stack)
          (ibuf . inst-buff)))
```

BSTK fetches an instruction from the instruction memory, and if the instruction fetched is not *top* and the instruction buffer is not full (function *stutterp* below), it queues the fetched instruction to the end of the instruction buffer and increments the program counter. If the instruction buffer is full, then the machine executes all buffered instructions in the order they were enqueued, thereby draining the instruction buffer and obtaining a new stack. It also updates the instruction buffer so that it only contains just the current fetched instruction. If none of the transition rules match, then BSTK drains the instruction buffer (if it is not empty) and updates the stack accordingly. Since BSTK is also a deterministic machine, we encode its transition relation ($\overset{C}{\rightarrow}$) as the function `impl-step`. Having defined the transition relation and the state of BSTK machine, we can define its transition system \mathcal{M}_C .

```
(defun stutterp (inst ibuf)
  "BSTK stutters if ibuf is not full or the current instruction is not 'top"
  (and (< (len ibuf) (ibuf-capacity))
       (not (equal (car inst) 'top))))
```

```
(defun impl-step (s)
  "single step of BSTK"
  (let* ((stk (istate-stk s))
         (ibuf (istate-ibuf s))
         (imem (istate-imem s))
         (pc (istate-pc s))
         (inst (nth pc imem)))
    (if (instp inst)
        (let ((nxt-pc (1+ pc))
              (nxt-stk (if (stutterp inst ibuf)
                           stk
                           (impl-observable-stk-step stk ibuf)))
              (nxt-ibuf (if (stutterp inst ibuf)
                            (impl-internal-ibuf-step inst ibuf)
                            (impl-observable-ibuf-step inst))))
          (istate imem nxt-pc nxt-stk nxt-ibuf))
        (let ((nxt-pc (1+ pc))
              (nxt-stk (impl-observable-stk-step stk ibuf))
              (nxt-ibuf nil))
          (istate imem nxt-pc nxt-stk nxt-ibuf))))))
```

Before we describe the correctness of BSTK based on skipping refinement, we first discuss why an existing notion of refinement such as stuttering refinement [12] will not suffice. If BSTK takes a step,

which requires it to drain its instruction buffer (the buffer is full or the current instruction fetched is top), then the stack will be updated to reflect the execution of all instructions in `ibuf`, something that is neither a stuttering step nor a single transition of the STK system. Therefore, it is not possible to prove that BSTK refines STK, using stuttering refinement and a refinement map that does not transform the stack.

We now formulate the correctness of BSTK based on the notion of skipping refinement. We show $\mathcal{M}_C \lesssim_r \mathcal{M}_A$, using Definition 2. We define the refinement map, but first we note that we do not have to consider all syntactically well-formed STK states. We only have to consider states whose instruction buffer is consistent with the contents of the instruction memory, so called *good states* [15]. One way of defining a good state is as follows: state s is good iff $pc \geq |ibuf|$ and stepping BSTK from $\langle imem, pc - |ibuf|, [], stk \rangle$ state for $|ibuf|$ steps yields state s , where $|ibuf|$ is number of instructions in the instruction buffer of state s . We define a predicate `good-statep` recognizing a good state and show that the set of good states is closed under the transition relation of BSTK.

```
(defun committed-state (s)
  (let* ((stk (istate-stk s))
         (imem (istate-imem s))
         (ibuf (istate-ibuf s))
         (pc (istate-pc s))
         (cpc (nfix (- pc (len ibuf)))))
    (istate imem cpc stk nil)))

(defun good-statep (s)
  "if state s is reachable from a committed-state in |ibuf| steps"
  (let ((pc (istate-pc s))
        (ibuf (istate-ibuf s)))
    (and (istatep s)
         (>= pc (len ibuf))
         (let* ((cms (committed-state s))
                (s-cms (cond ((endp ibuf)
                              cms)
                             ((endp (cdr ibuf))
                              (impl-step cms))
                             ((endp (cddr ibuf))
                              (impl-step (impl-step cms)))
                             ((endp (cddddr ibuf))
                              (impl-step (impl-step (impl-step cms))))
                (t cms))))
          (equal s-cms s))))))

(defthm good-state-inductive
  (implies (good-statep s)
           (good-statep (impl-step s))))
```

The refinement map `ref-map`, a function from a set of good states to set of abstract states (`sstatep`) is defined as follows.

```
(defun ref-map (s)
  (let* ((stk (istate-stk s))
         (imem (istate-imem s))
         (pc (istate-pc s))
         (ibuf (istate-ibuf s))
         (ibuflen (len ibuf))
         (rpc (nfix (- pc ibuflen))))
    (sstate imem rpc stk)))
```

Given `ref-map`, we define B to be the binary relation induced by it, *i.e.*, sBw iff s is a good state and $w = \text{ref-map}(s)$.

Now observe that when the instruction is full or the current instruction is *top*, one step of BSTK corresponds to largest number of STK steps. In both cases, the BSTK machine executes all instructions in the instruction buffer and if the current instruction is *top*, it executes it as well. The condition WFSK2d in Definition 2 that requires us to reason about reachability, hence can easily be reduced to bounded reachability. Hence, we set $j = k + 2$, where k is the capacity of the instruction buffer, and condition WFSK2d is $\langle \exists v : w \rightarrow^{<j} v : uBv \rangle$.

Since STK and BSTK are deterministic machines and STK does not stutter, we only need to define one rank function, a function from set of good states to non-negative integers.

```
(defun rank (s)
  "rank of an istate s is capacity of ibuf - |ibuf|"
  (- (ibuf-capacity) (len (istate-ibuf s))))
```

With above observations we simplify WFSK2 (Definition 2) to following condition.

For all s, u such that s and u are good states and $u = (\text{ref-map } s)$

$$(\text{ref-map } s) \xrightarrow{A}^{<k+2} (\text{ref-map } u) \vee \\ ((\text{ref-map } u) = (\text{ref-map } s) \wedge (\text{rank } u) < (\text{rank } s)) \quad (1)$$

Notice that since BSTK is deterministic, u is a function of s , so we can remove u from the above formula. Since $k + 2$ is a constant, we can expand out $\xrightarrow{A}^{<k+2}$ using only \xrightarrow{A} instead. We formalize Equation 1 in ACL2s by first defining a function `spec-step-skip-rel`, which takes as input STK states v and w and returns true only if v is reachable from w in $(\text{ibuf-capacity}) + 1$ steps.

```
(defthm bstk-skip-refines-stk
  (implies (and (good-statep s)
                (equal w (ref-map s))
                (equal u (impl-step s))
                (not (and (equal w (ref-map u))
                          (< (rank u) (rank s)))))
            (spec-step-skip-rel w (ref-map u))))
```

Once the definitions were in place, proving `bstk-skip-refines-stk` with ACL2s was straightforward. Next, we evaluated how amenable is SKS for automated reasoning, *i.e.*, using *only* symbolic simulation and no additional lemmas. We model BSTK with instruction buffer capacity of 2, 3, and 4, while no other restrictions were placed on the machines. In particular, the instruction memory (`imem`) and the stack (`stk`) component of the state for BSTK and STK machines are unbounded. The experiments were run on a 2.2 GHz Intel Core i7 with 16 GB of memory. For the BSTK with instruction buffer

capacity of 2 instructions, it took ~ 12 minutes to complete the proof and for a BSTK with instruction buffer capacity of 3 instructions, it took ~ 2 hours. For BSTK with instruction buffer capacity of 4 instructions the proof did not finish in over 3 hours.

4.2 Memory Controller

A memory controller is an interface between a CPU and a memory, and synchronizes communication between them. Designers implement several optimizations in a memory controller to maximize available memory bandwidth utilization and reduce the latency of memory accesses, known bottlenecks in optimal performance of programs. In this case study, we consider OptMEMC, a simple model of such an optimized memory controller. In our simplified model, a CPU is modeled as a list of memory request (*reqs*) and memory as a list of natural numbers (*mem*).

OptMEMC fetches a memory request from location *pt* in a queue of CPU requests, *reqs*. It enqueues the fetched request in the request buffer, *rbuf* and increments *pt* to point to the next CPU request in *reqs*. The capacity of *rbuf* is *k*, a fixed positive integer. If the fetched request is a *read* or the request buffer is full, then before enqueueing the request into *rbuf*, OptMEMC first analyzes the request buffer for consecutive write requests to the same address in the memory (*mem*). If such a pair of writes exists in the buffer, it marks the older write requests in the request buffer as redundant. Then it executes all the requests in the request buffer except the one that are marked redundant. Requests in the buffer are executed in the order they were enqueued. In addition to read and write commands, the memory controller periodically issues a *refresh* command to preserve data in memory. A *refresh* command reads all memory locations and immediately writes them back without modification. Refresh commands are required to periodically reinforce the charge in the capacitive storage cells in a DRAM. In effect, a refresh command leaves the data memory unchanged. We define the function `mrefresh` and prove that the memory is same before and after execution of the `refresh` command. This is the only property of `mrefresh` that we would require.

```
(defthm mrefresh-mem-unchanged
  (equal (mrefresh mem)
         mem))
```

To reason about the correctness of OptMEMC using skipping refinement, we define a high-level abstract system, MEMC, that acts as the specification for OPTMEMC. It fetches a memory request from the CPU and immediately executes the request. The syntax and the semantics of MEMC and OPTMEMC are given in Fig. 2, using the same conventions as described previously in the stack machine section.

We now formulate the correctness of OptMEMC based on the notion of skipping refinement. Let $\mathcal{M}_A = \langle S_A, \xrightarrow{A}, L_A \rangle$ and $\mathcal{M}_C = \langle S_C, \xrightarrow{C}, L_C \rangle$ be transition systems for MEMC and OptMEMC respectively. Like in the previous case study, we encode the state of the machines using `defdata` and formalize the transition relation of OptMEMC and MEMC using a step function that describes the effect of each instruction on the state of the machine. The labeling function L_A and L_C are the identity functions. Given a refinement map `ref-map` : $S_C \rightarrow S_A$, we use Definition 2 to show that $\mathcal{M}_C \lesssim_r \mathcal{M}_A$. As was the case with the previous case study, OptMEMC and MEMC are deterministic machines and MEMC does not stutter. WFSK2 (Definition 2) can again be simplified to Formula 1.

Once the definitions of the transition systems for the two machines were in place, it was straightforward to prove skipping refinement with ACL2s. Like in the previous case study, we also prove the theorem using *only* symbolic execution and no additional lemmas, for configurations of OPTMEMC with buffer capacity of 2 and 3. For OptMEMC with buffer capacity of 2, the final theorem was proved

$mem := [] \mid v :: mem$	(Memory)
$req := \langle write\ addr\ v \rangle \mid \langle read\ addr \rangle \mid \langle refresh \rangle$	(Request)
$pt := 0 \mid 1 \mid \dots \mid n \mid \dots$	(Request Location)
$reqs := [] \mid req :: reqs$	(Requests)
$rbuf := [req_1, \dots, req_k]$	(Request Buffer)
$sstate := \langle reqs, pt, mem \rangle$	(MEMC State)
$istate := \langle reqs, pt, rbuf, mem \rangle$	(OptMEMC State)

MEMC (\xrightarrow{A})

$$\frac{\langle reqs, pt, mem \rangle, reqs[pt] = \langle write\ addr\ v \rangle}{\langle reqs, pt + 1, mem[addr] \leftarrow v \rangle}$$

$$\frac{\langle reqs, pt, mem \rangle, reqs[pt] = \langle read\ addr \rangle}{\langle reqs, pt + 1, mem \rangle}$$

$$\frac{\langle reqs, pt, mem \rangle, reqs[pt] = \langle refresh \rangle}{\langle reqs, pt + 1, mem \rangle}$$

OptMEMC (\xrightarrow{C})

Let $|rbuf| = j$

$$\frac{\langle reqs, pt, rbuf, mem \rangle, \quad j < k, req \neq top}{\langle reqs, pt, rbuf :: reqs[pt], mem \rangle}$$

$$\frac{\langle reqs, pt, rbuf, mem \rangle, \quad reqs[pt] = \langle read\ addr \rangle, \quad \langle rbuf, 0, mem \rangle \xrightarrow{A}^j \langle rbuf, j, mem' \rangle}{\langle reqs, pt, [], mem' \rangle}$$

$$\frac{\langle reqs, pt, rbuf, mem \rangle, \quad j = k, \quad \langle rbuf, 0, mem \rangle \xrightarrow{A}^j \langle rbuf, k, mem' \rangle}{\langle reqs, pt, rbuf :: reqs[pt], mem' \rangle}$$

Figure 2: Syntax and Semantics of MEMC and OptMEMC

in ~ 2 minutes and with OptMEMC buffer capacity of 3, it took ~ 1 hour to prove the final theorem. The proof with buffer capacity of 4 instructions did not finish in over 3 hours.

4.3 Superword Level Parallelism with SIMD instructions

An effective way to improve the performance of multimedia programs running on modern commodity architectures is to exploit Single-Instruction Multiple-Data (SIMD) instructions (*e.g.*, the SSE/AVX instructions in x86 microprocessors). Compilers analyze programs for *superword level parallelism* and when possible replace multiple scalar instructions with a compact SIMD instruction that concurrently operates on multiple data [11]. In this case study, we illustrate the applicability of skipping refinement to verify the correctness of such a compiler transformation.

For the purpose of this case study, we make some simplifying assumptions: the state of the source and target programs (modeled as transition systems) is a three-tuple consisting of a sequence of instructions, a program counter and a store. We also assume that a SIMD instruction simultaneously operates on *two* sets of data operands and that the transformation analyzes the program at a basic block level. Therefore, we do not model any control flow instruction. Fig. 3 shows how two add and two multiply scalar instructions are transformed into corresponding SIMD instructions. Notice that the transformation does *not* reorder instructions in the source program.

$$\begin{array}{l} a = b + c \\ d = e + f \end{array} \rightarrow \begin{array}{l} \boxed{a} \\ \boxed{d} \end{array} = \begin{array}{l} \boxed{b} \\ \boxed{e} \end{array} +_{SIMD} \begin{array}{l} \boxed{c} \\ \boxed{f} \end{array}$$

$$\begin{array}{l} u = v \times w \\ x = y \times z \end{array} \rightarrow \begin{array}{l} \boxed{u} \\ \boxed{x} \end{array} = \begin{array}{l} \boxed{v} \\ \boxed{y} \end{array} \times_{SIMD} \begin{array}{l} \boxed{w} \\ \boxed{z} \end{array}$$

Figure 3: Superword Parallelism

The syntax and operational semantics of the scalar and vector machines are given in Fig. 4, using the same conventions as described previously in the stack machine section. We denote that x, \dots, y are variables with values v_x, \dots, v_y in *store* by $\{(x, v_x), \dots, (y, v_y)\} \subseteq \text{store}$. We use $[(\text{sop } v_x v_y)]$ to denote the result of a scalar operation *sop* and $[(\text{vop } \langle v_a v_b \rangle \langle v_d v_e \rangle)]$ to denote the result of a vector operation *vop*. Finally, we use $\text{store}|_{x:=v_x, \dots, y:=v_y}$ to denote that variables x, \dots, y are updated (or added) to *store* with values v_x, \dots, v_y . Notice that the language of a source program consists of scalar instructions while the language of the target program consists of both scalar and vector instructions. As in the previous two case studies, we model the transition relation of a program (both source and target program) by modeling the effect of an instruction on the state of machines.

We use the translation validation approach to verify the correctness of the vectorizing compiler transformation [4], *i.e.*, we prove the equivalence between a source program and the generated vector program. As in the previous two case studies, the notion of stuttering simulation is too strong to relate a scalar program and the vector program produced by the vectorizing compiler, no matter what refinement map we use. To see this, note that the vector machine might run exactly twice as fast as the scalar machine and during each step the scalar machine might be modifying the memory. Since both machines do not stutter, in order to use stuttering refinement, the length of the vector machine run has to be equal to the run of the scalar machine.

Let $\mathcal{M}_A = \langle S_A, \xrightarrow{A}, L_A \rangle$ and $\mathcal{M}_C = \langle S_C, \xrightarrow{C}, L_C \rangle$ be transition systems of the scalar and vector machines, respectively corresponding to the source and target programs. The vector program is correct iff \mathcal{M}_C refines \mathcal{M}_A . We show $\mathcal{M}_C \lesssim_r \mathcal{M}_A$, using Definition 2. Determining j , an upper-bound on skipping

$loc := \{x, y, z, a, b, c, \dots\}$	(Variables)
$sop := add \mid sub \mid mul \mid and \mid or \mid nop$	(Scalar Ops)
$vop := vadd \mid vsub \mid vmul \mid vand \mid vor \mid vnop$	(Vector Ops)
$sinst := sop \langle z \ x \ y \rangle$	(Scalar Inst)
$vinst := vop \langle c \ a \ b \rangle \langle f \ d \ e \rangle$	(Vector Inst)
$sprg := [] \mid sinst :: sprg$	(Scalar Program)
$vprg := [] \mid (sinst \mid vinst) :: vprg$	(Vector Program)
$store := [] \mid \langle x, v_x \rangle :: store$	(Registers)

Scalar Machine (\xrightarrow{A})	$\frac{\langle sprg, pc, store \rangle, \{ \langle x, v_x \rangle, \langle y, v_y \rangle \} \subseteq store, \quad sprg[pc] = sop \langle z \ x \ y \rangle, \quad v_z = \llbracket (sop \ v_x \ v_y) \rrbracket}{\langle sprg, pc + 1, store _{z:=v_z} \rangle}$
Vector Machine (\xrightarrow{C})	$\frac{\langle vprg, pc, store \rangle, \{ \langle x, v_x \rangle, \langle y, v_y \rangle \} \subseteq store, \quad sprg[pc] = sop \langle z \ x \ y \rangle, \quad v_z = \llbracket (sop \ v_x \ v_y) \rrbracket}{\langle vprg, pc + 1, store _{z:=v_z} \rangle}$ $\frac{\langle vprg, pc, store \rangle, \quad vprg[pc] = vop \langle c \ a \ b \rangle \langle f \ d \ e \rangle, \quad \{ \langle a, v_a \rangle, \langle b, v_b \rangle, \langle d, v_d \rangle, \langle e, v_e \rangle \} \subseteq store, \quad \langle v_c, v_f \rangle = \llbracket (vop \ \langle v_a \ v_b \rangle \langle v_d \ v_e \rangle) \rrbracket}{\langle vprg, pc + 1, store _{c:=v_c, f:=v_f} \rangle}$

Figure 4: Syntax and Semantics of Scalar and Vector Program

that reduces condition WFSK2d in Definition 2 to bounded reachability is simple because the vector machine can perform at most 2 steps of the scalar machine at once; therefore $j = 3$ suffices.

We next define the refinement map. Recall that refinement maps are used to define what is observable at concrete states from viewpoint of the abstract system. Let $sprg$ be the source program and $vprg$ be the compiled vector program. We first define a function pcT that takes as input the vector machine's program counter pc and a vector program $vprg$ and returns the corresponding value of the scalar machine's program counter.

```
(defun num-scalar-inst (inst)
  (cond ((vecinstp inst)
        2)
        ((instp inst)
        1)
        (t 0)))
```

```
(defun pcT (pc vprg)
  "maps values of the vector machine's pc to the corresponding values of
the scalar machine's pc"
  (let ((inst (nth pc vprg)))
```

```

(cond ((or (not (integerp pc))
          (< pc 0))
      0)
      ((zp pc)
       (num-scaler-inst inst))
      (t (+ (num-scaler-inst inst) (pcT (1- pc) vprg))))))

```

We next define a function `scalarize-vprg` that takes as input a vector program `vprg`. It walks through the list of instructions in `vprg` and translates each instruction in one of the following ways: if it is a vector instruction it *scalarizes* it into a list of corresponding scalar instructions, else if it is a scalar instruction it returns the list containing the instruction itself (function `scalarize` below). The result of `scalarize-vprg` is a scalar program. Notice that this function is significantly simpler than the compiler transformation procedure. This is because the complexity of a compiler transformation typically lies in its analysis phase, which determines if the transformation is even feasible, and not in the transformation phase itself.

```

(defun scalarize (inst)
  "scalarize a vector instruction"
  (cond ((vecinstp inst)
        (let ((op (vecinst-op inst))
              (ra1 (car (vecinst-ra inst)))
              (ra2 (cdr (vecinst-ra inst)))
              (rb1 (car (vecinst-rb inst)))
              (rb2 (cdr (vecinst-rb inst)))
              (rc1 (car (vecinst-rc inst)))
              (rc2 (cdr (vecinst-rc inst))))
          (case op
            (vadd (list (inst 'add rc1 ra1 rb1)
                        (inst 'add rc2 ra2 rb2)))
            (vsub (list (inst 'sub rc1 ra1 rb1)
                        (inst 'sub rc2 ra2 rb2)))
            (vmul (list (inst 'mul rc1 ra1 rb1)
                        (inst 'mul rc2 ra2 rb2))))))
        ((instp inst) (list inst))
        (t nil)))

(defun scalarize-vprg-aux (pc vprg)
  "scalarize the vector program from [0,pc]"
  (if (or (not (integerp pc))
          (< pc 0))
      nil
      (let ((inst (nth pc vprg)))
        (cond
          ((zp pc) ;=0
           (scalarize inst))
          (t
           (append (scalarize-vprg (1- pc) vprg) (scalarize inst)))))))

```

```
(defun scalarize-vprg (vprg)
  (scalarize-vprg-aux (len vprg) vprg))
```

The refinement map `ref-map`: $S_C \rightarrow S_A$, now can be defined as follows.

```
(defun ref-map (s)
  (let* ((store (vstate-store s))
         (vprg (vstate-vprg s))
         (isapc (pcT (1- (vstate-pc s)) vprg)))
    (sstate isapc store (scalarize-vprg (len vprg) vprg))))
```

Given `ref-map`, we define B to be the binary relation induced by the refinement map, *i.e.*, sBw iff $s \in S_C$ and $w = (\text{ref-map } s)$. Notice that since the machines do not stutter, WFSK2 (Definition 2) can be simplified as follows. For all $s, u \in S_C$ such that $s \xrightarrow{C} u$:

$$(\text{ref-map } s) \xrightarrow{A}^{<3} (\text{ref-map } u) \tag{2}$$

Since the vector machine is deterministic, u is a function of s , so we can remove u from the above formula, if we wish. Also, we can expand out $\xrightarrow{A}^{<3}$ to obtain a formula using only \xrightarrow{A} instead. We prove the appropriate lemmas to prove the final theorem: vector machine refines scalar machine.

```
(defthm vprg-skip-refines-sprg
  (implies (and (vstatep s)
                (equal w (ref-map s)))
            (spec-step-skip-rel w (ref-map (vec-step s)))))
```

where `vstatep` is the recognizer for a state of vector machine; `vec-step` is a transition function for vector machine; and `spec-step-skip-rel` is a function that takes as input two states of scalar machine and returns true if the second is reachable from the first in less than three steps.

Note that $pcT(pc, vprg)$ can also be determined using a history variable and would be a preferable strategy from verification efficiency perspective.

5 Conclusion and Future Work

In this paper, we used skipping refinement to prove the correctness of three optimized reactive systems in ACL2s. The concrete optimized systems can run “faster” than the corresponding abstract high-level specifications. Skipping refinement is an appropriate notion of correctness for reasoning about such optimized systems. Furthermore, well-founded skipping simulation gives “local” proof method that is amenable for automated reasoning. Stuttering simulation and bisimulation have been used widely to prove correctness of several interesting systems [14, 16, 17]. However, we have shown that these notions are too strong to analyze the class of optimized reactive systems studied in this paper. Skipping simulation is a weaker and more generally applicable notion than stuttering simulation. In particular, skipping simulation can be used to reason about superscalar processors, pipelined processors with multiple instructions completion, without modifying the specification (ISA), an open problem in [16]. We refer the reader to our companion paper [10] for a more detailed discussion on related work.

For future work, we would like to develop a methodology to increase proof automation for proving correctness of systems based on skipping refinement. In [10], we showed how model-checkers can be used to analyze correctness for finite-state systems. Similarly, we would like to use the GL framework [18], a verified framework for symbolic execution in ACL2, to further increase the efficiency and automation.

Acknowledgments

We would like to thank Harsh Raju Chamarthi for help on the proof of vectorizing compiler transformation.

References

- [1] *Skipping Simulation Model*. Available at <http://www.ccs.neu.edu/home/jmitesh/sks>.
- [2] Martín Abadi & Leslie Lamport: *The Existence of Refinement Mappings*. In: *Theoretical Computer Science, 1991*. Available at [http://dx.doi.org/10.1016/0304-3975\(91\)90224-P](http://dx.doi.org/10.1016/0304-3975(91)90224-P).
- [3] Krzysztof R Apt & Gordon D Plotkin: *Countable nondeterminism and random assignment*. Available at <http://dx.doi.org/10.1145/6490.6494>.
- [4] Clark W. Barrett, Yi Fang, Benjamin Goldberg, Ying Hu, Amir Pnueli & Lenore D. Zuck: *TVOC: A Translation Validator for Optimizing Compilers*. In: *CAV, 2005*. Available at http://dx.doi.org/10.1007/11513988_29.
- [5] Michael C. Browne, Edmund M. Clarke & Orna Grumberg: *Characterizing Finite Kripke Structures in Propositional Temporal Logic*. In: *Theoretical Computer Science, 1988*. Available at [http://dx.doi.org/10.1016/0304-3975\(88\)90098-9](http://dx.doi.org/10.1016/0304-3975(88)90098-9).
- [6] Harsh Raju Chamarthi, Peter C. Dillinger & Panagiotis Manolios: *Data Definitions in the ACL2 Sedan*. In: *ACL2 2014*. Available at <http://dx.doi.org/10.4204/EPTCS.152.3>.
- [7] Harsh Raju Chamarthi, Peter C. Dillinger, Panagiotis Manolios & Daron Vroon: *The ACL2 Sedan Theorem Proving System*. In: *TACAS 2011*. Available at http://dx.doi.org/10.1007/978-3-642-19835-9_27.
- [8] Rob J. van Glabbeek: *The Linear Time-Branching Time Spectrum (Extended Abstract)*. In: *CONCUR 1990*. Available at <http://dx.doi.org/10.1007/BFb0039066>.
- [9] David S Hardin: *Real-time objects on the bare metal: an efficient hardware realization of the Java TM Virtual Machine*. In: *ISORC, 2001*, doi:10.1109/ISORC.2001.922817.
- [10] Mitesh Jain & Panagiotis Manolios: *Skipping Refinement*. In: *CAV, 2015*. Available at http://dx.doi.org/10.1007/978-3-319-21690-4_7.
- [11] Samuel Larsen & Saman P. Amarasinghe: *Exploiting superword level parallelism with multimedia instruction sets*. In: *PLDI, 2000*. Available at <http://doi.acm.org/10.1145/349299.349320>.
- [12] P. Manolios: *Mechanical verification of reactive systems*. Ph.D. thesis, University of Texas.
- [13] Panagiotis Manolios: *A Compositional Theory of Refinement for Branching Time*. In: *CHARME 2003*. Available at http://dx.doi.org/10.1007/978-3-540-39724-3_28.
- [14] Panagiotis Manolios: *Correctness of Pipelined Machines*. In: *FMCAD, 2000*. Available at http://dx.doi.org/10.1007/3-540-40922-X_11.
- [15] Panagiotis Manolios & Sudarshan K. Srinivasan: *A computationally effecient method based on commitment refinement maps for verifying pipelined machines*. In: *MEMOCODE, 2005*. Available at <http://dx.doi.org/10.1109/MEMCOD.2005.1487914>.
- [16] Sandip Ray & Warren A. Hunt Jr.: *Deductive Verification of Pipelined Machines Using First-Order Quantification*. In: *CAV 2004*. Available at http://dx.doi.org/10.1007/978-3-540-27813-9_3.
- [17] Sandip Ray & Rob Sumners: *Specification and Verification of Concurrent Programs Through Refinements*. In: *Journal of Automated Reasoning*. Available at <http://dx.doi.org/10.1007/s10817-012-9258-1>.
- [18] Anna Slobodová, Jared Davis, Sol Swords & Warren A. Hunt Jr.: *A flexible formal verification framework for industrial scale validation*. In: *MEMOCODE, 2011*. Available at <http://dx.doi.org/10.1109/MEMCOD.2011.5970515>.