

BasicBlocker: Redesigning ISAs to Eliminate Speculative-Execution Attacks

Jan Philipp Thoma
Ruhr-University Bochum, Germany

Markus Krausz
Ruhr-University Bochum, Germany

Daniel J. Bernstein
University of Illinois at Chicago, USA, and Ruhr-University Bochum, Germany

Jakob Feldtkeller
Ruhr-University Bochum, Germany

Tim Güneysu
Ruhr-University Bochum, Germany

Abstract

Recent research has revealed an ever-growing class of microarchitectural attacks that exploit speculative execution, a standard feature in modern processors. Proposed and deployed countermeasures involve a variety of compiler updates, firmware updates, and hardware updates. None of the deployed countermeasures have convincing security arguments, and many of them have already been broken.

The obvious way to simplify the analysis of speculative-execution attacks is to eliminate speculative execution. This is normally dismissed as being unacceptably expensive, but the underlying cost analyses consider only software written for current instruction-set architectures, so they do not rule out the possibility of a new instruction-set architecture providing acceptable performance without speculative execution. A new ISA requires compiler updates and hardware updates, but those are happening in any case.

This paper introduces BasicBlocker, a generic ISA modification that works for all common ISAs and that removes most of the performance benefit of speculative execution. To demonstrate feasibility of BasicBlocker, this paper defines a BBRISC-V variant of the RISC-V ISA, reports implementations of a BBRISC-V soft core and an associated compiler, and presents a performance comparison for a variety of benchmark programs.

1 Introduction

The IBM Stretch computer in 1961 automatically speculated that a conditional branch would not be taken: it began executing instructions after the conditional branch, and rolled the instructions back if it turned out that the conditional branch was taken. More sophisticated branch predictors appeared in several CPUs in the 1980s, and in Intel’s first Pentium CPU in 1993.

Software analyses in the 1980s such as [12] reported that programs branched every 4–6 instructions. Each branch needed 3 extra cycles on the Pentium, a significant cost on top

of 4–6 instructions, especially given that the Pentium could often execute 2 instructions per cycle. However, speculative execution removed this cost whenever the branch was predicted correctly.

Subsequent Intel CPUs split instructions into more pipeline stages to support out-of-order execution and to allow higher clock speeds. The penalty for mispredictions grew past 10 cycles. Meanwhile the average number of instructions per cycle grew past two, so the cost of each mispredicted branch was more than 20 instructions. Intel further improved its branch predictors to reduce the frequency of mispredictions; see [18].

Today the performance argument for branch prediction is standard textbook material. Accurate branch predictors are normally described as “critical” for performance, “essential”, etc. Deployed CPUs vary in pipeline lengths, but speculative execution is common even on tiny CPUs with just a few pipeline stages, and is universal on larger CPUs.

This pleasant story of performance improvements was then rudely interrupted by Spectre [26], which exploited speculative behavior in various state-of-the-art CPUs to bypass critical security mechanisms such as memory protection, stealing confidential information via hardware-specific footprints left by speculatively executed instructions. This kicked off an avalanche of emergency software security patches, firmware updates, CPU modifications, papers proposing additional countermeasures targeting various software and hardware components in the execution flow, and papers presenting new attacks. Some countermeasures have been broken, and it is difficult to analyze whether the unbroken countermeasures are secure.

1.1 Our Contributions

At this point the security auditor asks “Can’t we just get rid of speculative execution?”—and is immediately told that this would be a performance disaster. *Every* branch would cost P cycles where P is close to the full pipeline length, and would thus cost the equivalent of $P \times I$ instructions where I is the number of instructions per cycle. This extra $P \times I$ -instruction

cost would be incurred every 4–6 instructions. The emergency security patches described above also sacrificed performance, but clearly were nowhere near *this* bad.

We observe, however, that this performance analysis makes an implicit assumption regarding the instruction-set architecture. We introduce an ISA feature, BasicBlocker, that undermines this assumption. BasicBlocker is simple and can be efficiently implemented in hardware. We show how compiler modifications to use BasicBlocker, on top of a CPU without speculative execution, obtain most of the performance benefit that would have been obtained by speculative execution. Eliminating speculative execution from CPUs removes one of the most complicated aspects of an audit of CPU security.

To evaluate performance and demonstrate feasibility of BasicBlocker, we start with an existing compiler and an existing CPU for an existing ISA; we modify all of these to support BasicBlocker; and we compare the performance of the modified CPU to the performance of the original CPU. We selected the RISC-V ISA [4] given its openness, and we selected a soft core (a CPU simulated by an FPGA) to allow evaluations without manufacturing a chip. Full details of our BBRISC-V ISA appear later in the paper.

Our performance results rely on a synergy between changes to the CPU and changes to the compiler, mediated by changes to the ISA. Given that other countermeasures against speculative-execution attacks have also involved modifications to compilers, firmware, and chips, there is no reason to avoid considering changes to the ISA. To improve deployability, we explain how a CPU supporting BasicBlocker can also run code compiled for the old ISA.

1.2 The BasicBlocker Concept in a Nutshell

The P -cycle branch-misprediction cost is the time from early in the pipeline, when instructions are fetched, to late in the pipeline, when a branch instruction computes the next program counter. If a branch passes through the fetch stage and is mispredicted, then the misprediction will not be known until P cycles later, when the next program counter is computed. Every instruction fetched in the meantime then needs to be rolled back.

The implicit assumption is that the ISA defines the branch instruction to take effect starting immediately with the next instruction. This assumption was already challenged by “branch delay slots” on the first RISC architecture in the 1980s; see generally [16]. A branch delay slot means that a branch takes effect only after the next instruction. The compiler compensates by moving the branch up by one instruction, if there is an independent previous instruction in the basic block, the contiguous sequence of instructions preceding the branch. A branch delay slot reduces the cost of a branch misprediction by 1 instruction, and the first RISC CPU pipeline was short enough that this removed any need for branch prediction.

A few subsequent CPUs used double branch delay slots,

reducing the branch-misprediction cost by 2 instructions. Obviously one can define an architecture with $P \times I$ delay slots after each branch, but this raises two issues. First, the compiled code depends on CPU pipeline details that change frequently. Second, a typical program has many short basic blocks that are rarely executed, and each one needs to expand to $P \times I$ instructions, which can raise serious code-size concerns.

In a BasicBlocker ISA, there is a “basic block N ” instruction guaranteeing that the next N instructions¹ will all be executed. These instructions include, optionally, a branch instruction, which takes effect after the N instructions, no matter where the branch is located within the N instructions. The same ISA supports many values of N .

1.3 Organization of the Paper

Section 2 discusses related work, especially alternative countermeasures against microarchitectural attacks. Section 3 introduces the fundamentals of speculation, and establishes definitions and notations that lay the foundation of a formal specification of our concept, which is presented in detail in Section 4. Section 5 explains the implementation of BasicBlocker in the RISC-V ISA and elaborates the support on hardware and software side. Section 6 evaluates performance, and Section 7 concludes.

2 Related Work

Transient-execution attacks, including speculative-execution attacks and faulty-execution attacks, gained widespread attention after the disclosure of Spectre [26] and Meltdown [29]. The attacks in [9, 11, 26, 27, 29, 30, 34, 42–44, 48] have shown many ways that transient execution can undermine “memory protection” and violate basic security assurances. Attacks often target particular CPUs but the general attack ideas apply to most CPUs; for example, [21] demonstrates the traditional Spectre attack on a RISC-V CPU. See [10, 25, 38] for surveys of attack vectors and countermeasures.

A typical speculative-execution attack arranges for mispredicted instructions to see sensitive data. The instructions are eventually rolled back but still leave footprints in the microarchitectural state. The attack inspects the timing of memory accesses to detect these footprints and extract the sensitive data.

One can try to hide the footprints of security-critical applications by regularly flushing caches [3, 8, 45, 49, 52]. However, [47] shows that microarchitectural traces can sometimes be detected in cache timings even after a cache flush. Furthermore, memory timing is only one of many covert channels [38] that need to be analyzed.

One can also try to limit the attacker’s ability to target useful instructions for speculative execution. For example,

¹A variant specifying N bytes of instructions could be better for architectures with variable-length instructions.

Retpoline [41] prevents branch-target injection for indirect calls; however, this stops only a few Spectre variants. By randomizing the branch predictor state, [53] aims to prevent out-of-place mistraining of branch instructions; however, speculative-execution attacks using in-place mistraining or even random mispredictions are still feasible.

2.1 Designing ISAs for Security

There is a long history of security features in ISAs. The most obvious example is memory protection. Memory protection is traditionally viewed as being managed by a few operating-system components, while normal compilers ignore security issues and focus on “user-level” instructions, and higher-level security problems are the responsibility of applications.

This model does not seem to have produced secure systems. Efforts to improve security have often included proposals, and sometimes deployments, of ISA modifications beyond memory protection. For example, when Intel introduced its AES-NI instructions [23], it wrote that the instructions “help in eliminating the major timing and cache-based attacks that threaten table-based software implementations of AES” and “make AES simple to implement, with reduced code size, which helps reducing the risk of inadvertent introduction of security flaws”. Various proposals for ISAs to guarantee some aspects of timing behavior [6, 7, 20], rather than just input-output behavior, are designed to support analyses of security against timing attacks. The recent paper [49] introduces a RISC-V extension to flush microarchitectural state and shows that the extension stops several covert channels.

It is also not a new idea to propose recompiling normal applications to use a modified ISA that enforces security constraints. In the literature on Control Flow Integrity (CFI) [2], HAFIX [14] introduces new call and return instructions to let the CPU enforce a certain type of CFI (subsequently analyzed in [40]), provided that the compiler modifies call instructions and return instructions.

BasicBlocker is similar to AES-NI in that it aims to reduce the performance incentives for a complicated behavior with security risks, i.e., speculative execution in one case and table-based AES implementations in the other case. This is in contrast to memory protection and CFI, which aim to stop various steps of known attacks.

This paper focuses on speculative-execution attacks. It should be possible to similarly address faulty-execution attacks by “preponing” fault detection, removing most of the performance benefit of transient execution after faults, but further investigation of this idea is left to future work.

Another approach to ISA modifications against transient-execution attacks is to explicitly tell the CPU which values are secret, and to limit the microarchitectural operations that can be carried out on secret values [35, 50, 51]. The benchmarks in [50] (also using a modified RISC-V) are not encouraging but the benchmarks in [35] are more encouraging. These ISA

modifications are orthogonal to BasicBlocker.

3 Speculation in Processors

In a pipelined processor, each instruction passes through multiple pipeline stages before eventually retiring. A textbook series of stages is *Instruction Fetch* (IF), *Instruction Decode* (ID), *Execution* (EX), *Memory Access* (MEM) and *Write Back* (WB) [39]. More complex CPUs can have many more stages.

If each stage takes one cycle then a branch instruction will be fetched on cycle n in IF, decoded on cycle $n + 1$ in ID, and executed on cycle $n + 2$ in EX, so at the end of cycle $n + 2$ the CPU knows whether the branch is taken or not. Without branch prediction, IF stalls on cycles $n + 1$ and $n + 2$, because it does not know yet which instructions to fetch after the branch. With branch prediction, IF speculatively fetches instructions on cycles $n + 1$ and $n + 2$, and ID speculatively decodes the first of those instructions on cycle $n + 2$. If the prediction turns out to be wrong then the speculatively executed instructions are *rolled back*: all of their intermediate results are removed from the pipeline.

The functional effects of instructions are visible only when the instructions retire, but side channels sometimes reveal microarchitectural effects of instructions that have been rolled back. As Spectre illustrates, this complicates the security analysis: one can no longer trust a branch to stop the wrong instructions from being visibly partially executed.

The standard separation of fetch from decode also means that *every instruction is being speculatively fetched*. An instruction fetched in cycle n could be a branch (or other control-flow instruction), but the CPU knows this only after ID decodes the instruction in cycle $n + 1$, so IF is speculatively fetching an instruction in cycle $n + 1$. We emphasize that this behavior is present even on CPUs without branch prediction: the CPU cannot know whether the instruction changes the control flow before decoding it.

Disabling all speculative execution thus means that every branch must stall fetching until it is executed, and, perhaps even more importantly, that *every instruction* must stall fetching until it is decoded. BasicBlocker addresses both of these performance problems, as shown below.

3.1 Definitions and Notations

In this section we will introduce some terminology and notation that we will use in the subsequent sections. To formally grasp the notion of speculative execution, we define microarchitectural effects and retired instructions. Other efforts on formalizing transient execution have been made in [5, 13, 22, 46].

Definition 1 (Microarchitectural Effects). *Let C be a processor with stateful resources R_C . We denote the state of $r \in R_C$ as s_r . An instruction i has microarchitectural effects iff i changes*

the state of a resource $r \in R_C$ to a state s'_r so that $s_r \neq s'_r$. We denote this as $i \rightarrow r$.

Definition 2 (Retired Instructions). *The list of retired instructions $W_{p,C}$ for a program p on a processor C holds all instructions $i \in P$ that passed all pipeline stages of C during the execution of p ².*

We further define the instruction stream, that holds all the instructions that caused changes to the state of a processor resource, regardless of their final existence in the program execution. We also give a definition of control flow instructions.

Definition 3 (Instruction Stream). *Let C be a processor with program space $P : \{(i_1, \dots, i_n) | i_x \in I, x \in \mathbb{N}\}$ where I is the set of instructions supported by C . We define the instruction stream of $p \in P$ as an ordered list of instructions of p that caused microarchitectural effects $\vec{I}_p = [i_1, \dots, i_l], l \in \mathbb{N}$.*

Definition 4 (Control Flow Instructions). *Let i_j be an instruction and P the program space. We define i_j as a control flow instruction if there is a program $p \in P, p = (i_1, \dots, i_n)$ in which the semantics of i_j cause the next retired instruction $r \in p$ to be $r \neq i_{j+1}$.*

Hence, control flow instructions are all kind of jumps and branches. We further introduce the notation $\vec{I}_p \setminus i_k$ as an instruction stream, where the k -th instruction has been removed from the instruction stream.

Program code can be divided into *basic blocks*. To do so, all possible control flow paths are mapped into a directed graph called Control Flow Graph (CFG) so that each edge of the CFG represents a control flow from one instruction to the next. If two vertices A and B are connected by a single edge ($A \rightarrow B$) with A having only a single outbound edge and B having only a single inbound edge, the vertices are merged if the two instructions are sequential in memory. This is identical to the textbook definition of CFG, with the difference that we require a basic block (a vertice of the CFG) to be located sequentially in memory, resulting in the termination of a basic block after an unconditional jump or call. Therefore, our definition guarantees that within a basic block, the next instruction can be found at $PC + |i|$, where $|i|$ is the instruction length in bytes.

Furthermore, we define a *sequential* basic block as a basic block that does not contain a control-flow instruction and therefore has only one possible successor which is at the address $Addr_{bb_end} + |i|$. Such a block can occur, for example, if a basic block has multiple input edges but only one output edge and the following basic block has multiple input edges.

Definition 5 (Basic Block). *Let p be a program with $p = (i_1, \dots, i_n)$, where $i_j \in I, j \in \mathbb{N}$ and I is the instruction set. A basic block b_j is a vertex of the (CFG) of p . A basic block is called sequential iff b_j has exactly one outgoing edge (b_j, b_k) and b_k lies directly behind b_j in memory.*

²Processors without pipeline are treated as single stage pipelined.

4 Concept

In this section, we outline the rationale behind our approach as well as the modifications to the ISA that allow the elimination of speculative execution within the microarchitecture. Though we use the RISC-V instruction set in the following examples, our solution is generally applicable to any ISA or processor as motivated in Section 4.6.

4.1 Security Rationale

Our general CPU-design objective follows the “strong t-security” Definition (see Definition 6) and can be summarized as:

The microarchitectural state of a CPU is affected only by instructions that will eventually be retired.

As noted earlier, this paper focuses on speculative-execution attacks, leaving faulty-execution attacks to future work. The goal in this paper is then to ensure that the microarchitectural state is affected only by instructions that will eventually be retired or that raise exceptions. This is “t-security” in Definition 6, parameterized by the set of exceptions.

The most important implication of this goal is that the CPU must *abandon any speculative behavior*. This eliminates a major source of complexity inside the security analysis of modern CPUs.

Abandoning speculative behavior includes abandoning speculative fetching, as fetching affects the state of the instruction cache. We do not want to decide whether speculation in the CPU frontend is exploitable (see [5, 24, 32]); instead we want to remove all speculation so that we do not have to analyze its exploitability.

Definition 6 (t-security). *Let C be a processor with stateful resources R_C and program space $P : \{(i_1, \dots, i_n) | i_x \in I, x \in \mathbb{N}\}$ where I the set of instructions supported by C . We call a program $p \in P$ strong-t-secure if the instruction stream \vec{I}_p of p holds: $\nexists i \in \vec{I}_p : i \rightarrow r \wedge i \notin W$ where $r \in R_C$. We further define p as t-secure on C if any instruction of p that violates the strong-t-security property raises an exception.*

Observation 1. *Speculative fetching is not t-secure on a processor with instruction cache.*

More precisely, let C be a pipelined processor with cache-like, stateful instruction memory M , deterministic speculative fetching at $PC + |i|$ and program space $P : \{(i_1, \dots, i_n) | i \in I, n \in \mathbb{N}\}$ where I the set of instructions supported by C . Assume that I includes a conditional branch instruction. Choose a $p \in P$ such that $i \in p : i \in I_B \wedge t \neq PC + |i|$ where t is the branch target, and assume that p is executed without raising an exception. The CPU fetches i' from $PC + |i| \neq t$ and hence $i' \rightarrow M$ but $i' \notin W$, contradicting the definition of t-security.

We emphasize that the violation of t -security does not solely rely on the presence of an instruction cache but rather any resource whose state is affected during speculative fetching. Examples may be memory controllers and CPU internal counters.

Another important aspect of our concept is that the hardware guarantees the secure execution of any program, independently of measures taken at compile time. We therefore release the software developer from the burden to add cumbersome mitigation techniques. The compiler, however, needs to be enhanced to take advantage of our ISA modifications to improve the performance of programs on the strictly non-speculative hardware. This design principle of hardware security enforcement is defined in Definition 7, which directly leads to Observation 2.

Definition 7 (Hardware secure processor). *Let C be a processor with program space $P : \{(i_1, \dots, i_n) | i_x \in I, x \in \mathbb{N}\}$ where I the set of instructions supported by C . We call C a hardware secure processor if $\forall p \in P : p$ is t -secure.*

Observation 2. *A hardware-secure processor is secure against speculation based transient attacks.*

4.2 Performance Rationale

Disabling speculative fetching and branch prediction, to generically thwart the security issues arising from speculative behavior, is conceptually simple, but is generally believed to incur a severe loss in performance, as explained in Section 1. BasicBlocker addresses this by providing metadata through an ISA modification to assist non-speculative hardware with efficient execution of software programs.

The CPU has a limited view of programs, accessing only a limited number of instructions at a time. With current ISAs, control-flow instructions appear without advance notice, and their result is available only after multiple pipeline stages, even though this result is needed immediately to infer the next instruction.

BasicBlocker takes the concept of basic blocks to the hardware level using custom instructions. At compile time a holistic view of the program is available in form of a control-flow graph, providing the code structure as basic blocks and control-flow changes. BasicBlocker uses the information available at compile time, specifically the length of individual basic blocks, and makes it available to the CPU during execution. This allows a non-speculative CPU to avoid most pipeline stalls, through the advance notice of control flow changes.

4.3 Basic Block Instruction

We introduce a new instruction, called basic block instruction (`bb`), which lays the foundation for BasicBlocker. Enabling

fast but non-speculative fetching requires additional information for the CPU, since normally we know that we can fetch the next instruction only after the prior instruction was decoded and it is ensured that the control flow does not deviate. Hence, normally the fetch unit would have to be stalled until the previous instruction was decoded. To avoid that delay, we define a new invariant that requires each basic block to start with a `bb` instruction that encodes the size of the basic block. Within this basic block, the CPU is allowed to *fast-fetch* instructions, knowing that upcoming instructions can be found in a sequential order in memory and will definitely be executed. That is, since per definition, within the basic block, no control flow changes can occur. The instruction further provides information whether the basic block is *sequential*, stating that the control flow continues with the next basic block in the sequence in memory. If a basic block does not contain a control-flow instruction it is therefore *sequential*. Figure 1 shows the transformation of traditional code (left) to code with `bb` instruction (right). The fetch unit of the CPU is responsible for counting the remaining instructions in a given block and only (fast)-fetch until the end of the basic block. From there, the program continues executing the next basic block which itself starts with a `bb` instruction.

We also modified the behavior of existing control-flow instructions, such as `bne`, `j` and `jlre`. The goal is to give advance notice of upcoming control-flow changes to the CPU. Since the processor knows the number of remaining instructions per basic block, we can schedule control-flow instructions within basic blocks as early as data dependencies allow, and perform the change the control flow at the end of the basic block. This key feature allows the CPU to correctly determine the control flow before the end of the basic block, and renders branch prediction in many cases obsolete.

As a result, the only time that the CPU needs to fetch slowly is at the transition of two basic blocks, because the following `bb` instruction needs to be executed before knowing the size and, hence, being able to continue fast-fetching. To avoid this delay, it is sufficient to add the capability of representing one additional set of basic block information internally and request this information as early as possible. This means that the CPU interposes the `bb` instruction of the next basic block as soon as the next basic block is known, regardless whether there are instructions left in the current basic block or not.

In Figure 2, this principle is illustrated for the code of Figure 1 (right side). The `bb` instruction of the second basic block is fetched as soon as the branch target of `bne` is known. Afterwards, the execution of the first basic block continues. Execution of the second basic block can start as soon as the first basic block is consumed and the size of the second basic block is known (after EX of `bb`). If the current basic block does not contain a control-flow instruction, which is indicated by the *sequential* flag of the `bb` instruction, the CPU can fetch the next `bb` instruction directly. Otherwise, the next `bb` instruction will be fetched after the control-flow instruction

```

; Start of first basic block
add    a5, a0, a4
add    t4, a3, a4
addi   a4, a4, 8
mul    a1, t3, t2
lh     t2, 0(a5)
bne    a4, a6, 80... ; compute branch and
                    change PC
; Start of 2nd basic block
lh     a7, 0(a1)
li     a4, 0
; Start of 3rd basic block
sh     a1, 0(a0)
...

```

```

bb     6, 0 ; first bb, size = 6, not seq
add    a5, a0, a4
add    t4, a3, a4
addi   a4, a4, 8
bne    a4, a6, 80... ; compute branch result
mul    a1, t3, t2
lh     t2, 0(a5) ; change PC after this instr.
bb     2, 1 ; 2nd bb, size = 2, seq
lh     a7, 0(a1)
li     a4, 0
bb     16, 0 ; 3rd bb, size = 16, not seq
sh     a1, 0(a0)
...

```

Figure 1: Example code for the new `bb` instruction. Left: Traditional RISC-V code does not contain information about the size of upcoming basic blocks. The `bne` instruction terminates the first block and conditionally branches. Right: The `bb` instruction gives information about upcoming code parts. The first basic block is terminated by the size given in the line 1 and performs a conditional branch based on the outcome of the `bne` instruction, whose result is already determined earlier.

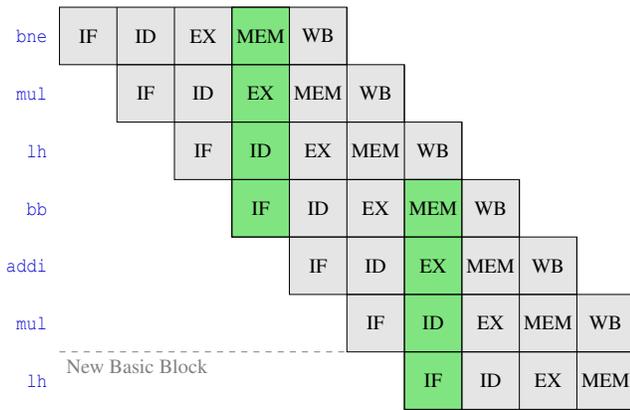


Figure 2: Pipeline diagram for optimal code. The `bb` instruction of the next basic block is fetched as soon as the branch was executed. The branch only takes effect at the end of the current basic block. When the branch instruction is sufficiently early rescheduled, the next basic block can be fetched without stalls.

passes the execution stage.

While the early fetching of the `bb` instruction changes the execution order, it does not affect security or correctness since the instruction is only fetched after the execution path is known for certain.

Even with these changes it is necessary to stall the CPU at the transition of two basic blocks until the size of the new basic block is known. Therefore, this concept works best with software that contains many large basic blocks with multiple opportunities to *reschedule* control-flow instructions at compile time. Software with a large number of small basic

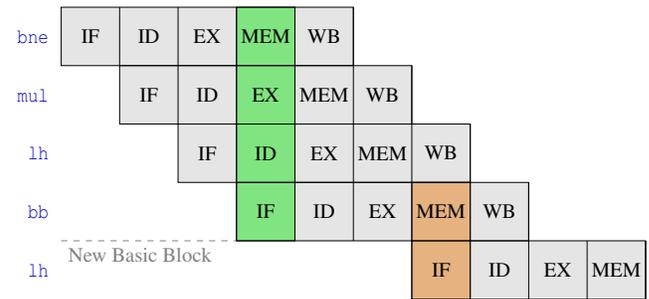


Figure 3: Pipeline diagram for code with non-optimal rescheduling of branch instructions. The next `bb` instruction is not finished with execution when the new basic block begins. The CPU needs to stall until the basic block size is known which is generally after the execution stage.

blocks is therefore less efficient, leading to pipeline stalls as shown in Figure 3.

The worst case is a control-flow instruction that could not be rescheduled, since then the CPU needs to be stalled both for the information from the control-flow instruction as well as from the `bb` instructions. This case is depicted in Figure 4. Note, however, that code with a high number of small basic blocks is also not ideal for traditional speculative CPUs, as a high number of closely spaced branch instructions will inevitably lead to mispredictions and pipeline stalls.

Overall, the rescheduling concept can be imagined as a variably-sized branch delay slot. The advantages of our concept over traditional branch delay slots are twofold:

- The CPU does not need special constructs for the branch delay instructions. At the end of a basic block, the CPU

can simply fetch the instruction at the target address, regardless of the type of instructions that were executed prior. If the basic block was sequential, the target register defaults to $PC + 4$. If any control-flow operations were executed, the target register points to the target address.

- By having a variably-sized branch delay mechanism, the code is compatible to all hardware architectures that support the `bb` instruction. Since the control-flow instructions were rescheduled as early as possible, the code is optimal for those hardware architectures. For fixed size branch delay slots, CPUs with smaller pipelines may introduce unnecessary `nop` instructions.

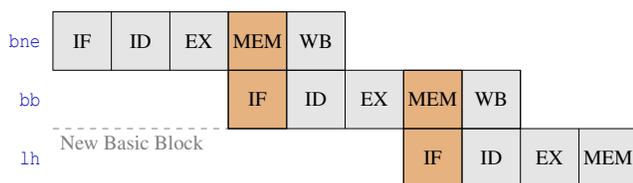


Figure 4: The worst case scenario has a branch instruction at the end of a basic block.

We now define the changes required by BasicBlocker in a more specific way. A processor supporting the `bb` instruction is required to have an instruction counter IC , a target register T , a branch flag B , and an exception flag E , all initialized to 0 on processor reset and used only as defined below. The functional behavior of the `bb` instruction is given in Definition 8, the changes to the control flow in Definition 9 and the behavior that raises an exception in Definition 10.

Definition 8 (BB Instruction). *The `bb` instruction takes a size parameter $n > 0$ and a sequential flag seq , and is executed as follows. If $IC \neq 0$, then $IC \leftarrow 0$ and $E \leftarrow 1$. Otherwise $IC \leftarrow n$; if $seq = 0$ then $B \leftarrow 1$; if $seq = 1$ then $B \leftarrow 0$ and T is set to the address of the $n + 1$ -th instruction following the `bb` instruction.*

Thus, on a functional level, Definition 8 only sets IC , T , B , and E but has no further effect on the execution of a program. The subsequent definitions have further effects.

Definition 9 (BB Delayed Branches). *The execution of non-`bb` instructions is modified as follows:*

- Before every non-`bb` instruction: if $IC > 0$ then $IC \leftarrow IC - 1$.
- During every control-flow instruction: any write to PC is instead written to T if $B > 0$, and is ignored if $B = 0$.
- After every control-flow instruction: if $B = 0$ then $E \leftarrow 1$; otherwise $B \leftarrow B - 1$.

- Subsequently, after every non-`bb` instruction: if $IC = 0$ then $PC \leftarrow T$; and if $IC = 0$ and $B > 0$ then $E \leftarrow 1$.

BasicBlocker raises an exception whenever the `bb` instruction is used in an illegal way.

Definition 10 (BB Exceptions). *After every instruction, an exception is raised if $IC = 0$ and $E \neq 0$.*

In other words, after the n instructions covered by a `bb` instruction, an exception is raised if any of the following occurred:

- $seq = 0$ and there was no control-flow instruction in the n instructions;
- $seq = 0$ and there was more than one control-flow instruction in the n instructions;
- $seq = 1$ and there was a control-flow instruction in the n instructions.

Also, another `bb` instruction within the n instructions immediately raises an exception.

All three definitions are required, in order to add BasicBlocker to an arbitrary ISA. The following extra requirement, a requirement to use `bb` instructions, slightly simplifies the implementation of BasicBlocker, although later we consider dropping this requirement for compatibility.

Definition 11 (BB Required). *In a BasicBlocker CPU with required BB: Before every non-`bb` instruction (and before IC is decremented), an exception is raised if $IC = 0$.*

To achieve an increased performance, an implementation of BasicBlocker can pre-execute `bb` instructions (cf. Figure 2) as defined in Definition 12. This pre-execution affects the microarchitecture and timing but not the ISA semantics.

Definition 12 (BB Prefetching). *A BasicBlocker CPU with prefetching pre-executes a `bb` instruction bb_{i+1} during the execution of a block, indicated by the `bb` instruction bb_i , as soon as any of the following occur:*

- bb_i is resolved if bb_i is sequential.
- the first control flow instruction of the block is resolved if bb_i is not sequential.

This requires an additional register P which holds the values n and seq until execution reaches the instruction following the prefetched `bb` instruction.

If the prefetch address is invalid, or if the prefetch address is valid but the prefetched instruction is not a `bb` instruction, then pre-execution is skipped and does not raise an exception.

4.4 Further optimizations

The above presented concept can be further optimized by providing the information contained in the `bb` instruction as soon as possible using pipeline forwarding. By construction, none of the information contained in the `bb` instructions affects any other element of the CPU than the fetch unit. Hence, it is possible to wire these bits back to the fetch unit directly after the decode stage without further changes to the design. Another clock cycle can be saved by using a bit mask to fast-decode the output of the instruction memory directly, with only marginal overhead.

A significant boost for performance can be achieved by introducing an additional interface to the instruction memory (or cache) that is used to access `bb` instructions. This would allow the fetch unit to request and process `bb` instructions in parallel with the normal instructions and, therefore, eliminate the entire performance overhead that is introduced through the addition of these instructions. Since a basic block contains always at least one instruction additional to the `bb` instruction, this instruction can be fetched before knowing the size of the basic block, without violating the above stated principles.

The worst case scenario with optimizations in place is depicted in Figure 5.

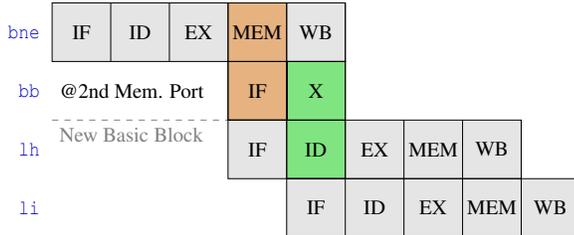


Figure 5: Worst case scenario with optimizations in place. The `bb` instruction is fetched from a second memory port and can therefore be parallelized.

Further optimizations are possible with additional changes to the ISA. For example, the 1-bit sequential flag can be replaced by a multi-bit counter of the number of control-flow instructions in the upcoming block, so (e.g.) `if (a&&bb&&c)` can be expressed as three branches out of a single block. This also changes the branch flag `B` to a multi-bit branch counter.

The idea to announce upcoming control-flow changes early on is also the foundation of hardware loop counters, as already discussed in the literature [17, 33]. Here, the software announces a loop to the hardware, which then takes responsibility for the correct execution.

We can seamlessly support hardware loop counters in our design concept. One new instruction (`lcnt`) is necessary to store the number of loop iterations into a dedicated register. The start and end address of a loop can be encoded into the `bb` instruction, by indicating with two separate flags whether the

corresponding basic block is the start (s-flag) or end (e-flag) block of the loop. This allows the hardware to know the next basic block, as soon as the `bb` instruction of the end block gets executed. The fast execution of nested loops can be supported by adding multiple start and end flags to the `bb` instruction as well as adding multiple registers for the number of loop iterations. A more detailed description of the loop counter integration to our concept can be found in Appendix A.

4.5 Security Considerations

Our strictly non-speculative pipeline design prevents speculative-execution attacks by design as the absence of speculative execution inhibits such leakage through caches and other side channels. As the security against speculative-execution attacks is guaranteed in hardware, the responsibility to defend against these attacks is taken away from the software developer.

We now claim that a processor supporting the `bb` instruction in general does not violate our t-security definition.

Observation 3. *Let C be a hardware secure processor where every instruction has a microarchitectural effect and C_{bb} be the same processor with `bb` instruction as defined in Definitions 8, 9, 10, 11 and 12. Then C_{bb} is a hardware secure processor.*

We support this claim by iteratively adding features from the definitions to the hardware secure processor C and showing that it maintains the hardware security property by showing, that the t-security is not affected.

i) First, let C' be C with functional implementation of `bb` instructions which adds the registers IC , T , B , and E to C and sets them according to Definition 8. Adding this functionality has no influence on retired instructions and does hence not violate the t-security property.

ii) Next, let C'' be C' with the BasicBlocker control flow as defined in Definition 9, raising exceptions according to Definitions 10 and 11. Decrementing the instruction counter IC during every instruction does not violate the t-security property, because even as the decrementation adds a microarchitectural effect to the execution of every instruction, it does not affect the set of retired instructions. Since every instruction of C' already has a microarchitectural effect, t-security is not violated. Similarly, the modified behavior of branches does not violate t-security, as the microarchitectural effect $i \rightarrow PC$ is transformed to $i \rightarrow T$ and $T \rightarrow PC$ which by the rules of transitivity yields $i \rightarrow PC$. The definition of t-security is time invariant and thus, the delay does not affect t-security. All undefined cases raise an exception which is not relevant for t-security.

iii) Let C_{bb} be C'' with prefetching of `bb` instructions as defined in Definition 12. Due to the requirements for prefetching, it is known at this point, that the instruction flow eventually continues at T . The time invariance argument used in *(ii)*

yields that t-security is maintained. Similarly, the addition of the register P does not violate t-security by the arguments of (i). This concludes the argument for Observation 3.

The requirement for every instruction to have a microarchitectural effect is no real restriction, as any state change is a microarchitectural effect (e.g. PC update).

BasicBlocker can also be used as a form of coarse-grained CFI [1, 2], as it only allows control flow changes to beginnings of basic blocks, indicated by a `bb` instruction. This reduces the prospects of success for (JIT-)ROP [36, 37] attacks, as the variety of potential gadgets is reduced. However, it has been shown, that (JIT-)ROP attacks can still be launched with coarse-grained CFI in place [15], even with a reduced amount of gadgets. To support fine-grained CFI the `bb` instruction could easily be extended with a tag.

4.6 Compatibility

For simplicity and comprehension we showed our ISA modification for an in-order, single issue processor with a generic five stage RISC pipeline. One could easily extend our design to larger, more complex CPUs.

Adding support for out-of-order processors is trivial as per design, every instruction that is fetched by the processor will be retired—that is, if none of the instructions raise an exception. Once the CPU fetches the instruction, reordering is permitted as far as functional correctness is ensured. Utilizing the two counter sets, reordering can be done beyond basic block borders if the `bb` instruction of the following basic block has been executed.

Similarly, support for multi-issue pipelines is easy to achieve. Once the `bb` instruction is executed, the CPU may fetch all instructions within the current basic block in an arbitrary amount of cycles. If the successor basic block is known the CPU may fetch instructions from both basic blocks in one cycle. Secondary pipelines may also be useful to pre-execute `bb` instructions for the following basic block in parallel as described earlier. The more parallel pipelines a CPU has, the more important it gets to build software with large basic blocks, as small basic blocks with branch dependent instructions cause a higher performance loss on a multi-issue CPU compared to a single-issue CPU.

Generally, the pipeline length can be flexibly chosen. However, as the CPU needs to wait for results of branch and `bb` instructions, it is desirable to make the results of these instructions available as early as possible.

A major feature of modern systems is the support of interrupts and context switches. We note that our concept does not impede such features; it merely increases the necessary CPU state that needs to be saved in such an event. More specifically, it is necessary to save the already gathered information about the current and upcoming basic blocks as well as the state of the loop counter, in addition to all information usually saved during a context switch.

Our proposal includes one new instruction and a modification to existing control-flow instructions. For easier deployability, it is desirable for a BasicBlocker CPU to be backwards-compatible. One could define new BasicBlocker control-flow instructions separate from the previous control-flow instructions. However, it suffices to interpret a control-flow instruction as having the new semantics if it is within the range of a `bb` instruction, and otherwise as having the old semantics, dropping Definition 11. Legacy code compiled for the non-BasicBlocker ISA will then run correctly, and code recompiled to use `bb` will run correctly with higher performance.

It would also be possible to integrate our solution into a secure enclave by providing a modified fetch unit for the enclave. Security critical applications could be run in the protected enclave while legacy software can be executed on the main processor without performance losses.

5 Implementation

We now give a specific example of BasicBlocker applied to an ISA, by defining BBRISC-V, a BasicBlocker modification of the RISC-V ISA. We further, present a proof-of-concept implementation³ of a BBRISC-V capable soft core and compiler. Our modified ISA further specifies support for hardware loop counters, as proposed in Section 4.4, but the optimal placement of hardware loop counters during compilation is not in the scope of this paper, and therefore not supported by our compiler.

5.1 BBRISC-V ISA

The BasicBlocker modification requires the definition of the `bb` instruction as well as semantic changes to all control flow instructions.

The `bb` instruction does not fit into any of the existing RISC-V instruction types so that we define a new instruction type to achieve an optimal utilization of the instruction bits (Figure 6). This instruction does not take any registers as input but rather parses the information directly from the bitstring. The size is encoded as a 16-bit immediate, enabling basic blocks with up to 65536 instructions. One can split a larger basic block into multiple sequential blocks if necessary. The sequential flag is a one-bit immediate value. The behavior of all RISC-V control-flow instructions (`JAL`, `JALR`, `BEQ`, `BNE`, `BLT`, `BGE`, `BLTU`, `BGEU`) is changed so that they alter the control flow at the end of the current basic block.

We also include hardware loop counters in the BBRISC-V ISA. The `lcnt` instruction sets the number of loop iterations (Figure 6). This I-Type instruction requires a 12 bit immediate value as well as a source and a target register. The counter value is then computed as $cnt = imm + rs.value$ and saved to

³The code is currently under preparation for publication.

Name	Description
Baseline	Standard version without modifications.
BB Info	As in Baseline, but every basic block starts with a <code>bb</code> instruction.
BB Resched	As in BB Info, but with high-priority rescheduling of terminator instructions.

Table 1: Compiled versions used for benchmarking.

Name	BB	SF	Branch Prediction
Simplest	no	no	none
Baseline	no	yes	none
Static	no	yes	backwards taken
Dynamic	no	yes	2-bit BHT
Dyn. Target	no	yes	additional BTB
<i>BasicBlocker</i> (new)	yes	no	none

Table 2: VexRiscv hardware instantiation options. BB: supports `bb` instruction. SF: speculative fetching.

by low-level code transformations just before code emission.

6 Evaluation

To evaluate the performance of BBRISC-V we used Coremark and the Embench IOT [19] benchmark suite. The latter consists of a number of small benchmark programs corresponding to representative real-world workloads.

We compiled three versions of each benchmark program, as listed in Table 1: one without *BasicBlocker*, one with a new compile flag enabling the insertion of `bb` instructions, and one with `bb` plus rescheduling of terminator instructions. Except for these differences, the compiler and compile flags are identical. The compile flags can be found in Appendix B.

We then ran these programs on several variants of VexRiscv, as listed in Table 2. Section 6.1 reports non-*BasicBlocker* performance on four existing VexRiscv variants with speculative fetching and different branch-prediction strategies, and on a purely non-speculative VexRiscv variant. Section 6.2 reports *BasicBlocker* performance, with and without rescheduling, on our new VexRiscv variant supporting *BasicBlocker*.

Figure 8 compares all of these results. To compensate for different runtimes of the benchmarks, all results are normalized to a baseline run of the respective benchmark using the original VexRiscv core with speculative fetching but no branch prediction. The raw benchmark results including the upper and lower quartiles for runtime deviations over 100 executions can be found in Appendix C. As we evaluate the bare metal system, the noise during the measurements is minimal, i.e. the standard deviation is below 1% for all measurements.

Section 6.3 covers various ways to further improve Ba-

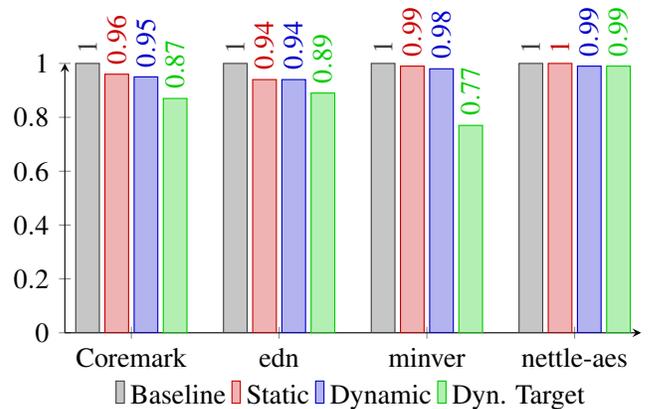


Figure 7: Performance of VexRiscv Branch Predictors for selected programs in cycles, normalized to the no-prediction case.

sicBlocker performance beyond the blue line shown in Figure 8.

6.1 Non-*BasicBlocker* Performance

We first evaluated the effectiveness of each branch predictor, by using the unmodified code of each benchmark and executing it for each of the four branch prediction strategies. In Figure 7, scores for Coremark and selected programs of the Embench suite are plotted in relation to a run with no branch prediction enabled. By means of illustration, we selected the well known Coremark benchmark, *edn*, a vector multiplication program, as an average case, *minver*, a floating point matrix inversion program, as a branch heavy example and *nettle-aes* as a high-performance case.

The performance benefit of branch prediction ranges between 0 and 23%. The best results were obtained by the most complex *dynamic target* branch predictor, with an average of 14% speedup over all tested benchmarks. The highest speedup is achieved by unoptimized code that relies heavily on branches. Highly optimized code, such as cryptographic libraries, are barely affected by branch prediction at all.

The green area in Figure 8 shows the branch-prediction speedup for all of our benchmark programs. The dashed gray line is the time for the baseline, no branch prediction. The bottom of the green area is the time for the best branch predictor. Times are normalized so that the baseline is 1.

We then executed the unmodified code on a strictly non-speculative version of VexRiscv: not just without branch prediction but also without speculative fetching. The result is the red line in Figure 8 (“Simplest”), on average 2.6 times slower than the baseline.

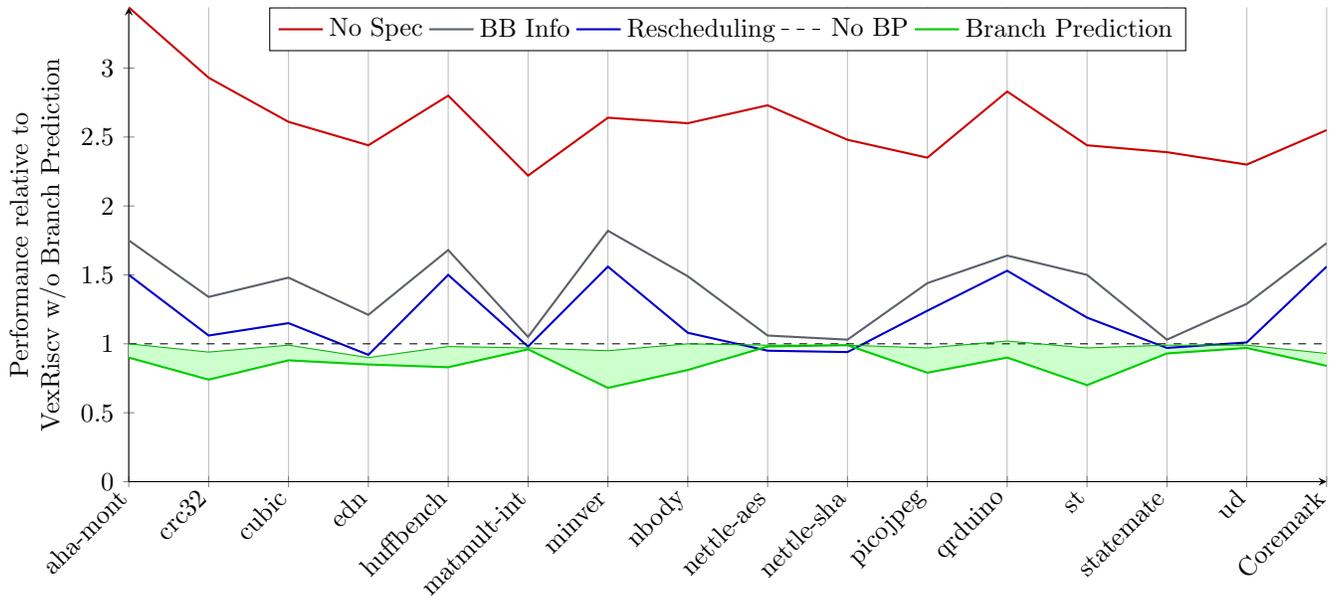


Figure 8: Performance evaluation of BasicBlocker for various benchmarks on VexRiscv. The red line depicts a strictly non-speculative CPU without `bb` instructions. The dashed gray line depicts the baseline processor (VexRiscv - speculative fetching, no branch prediction). The green area visualizes the performance range of the VexRiscv branch predictors.

6.2 BasicBlocker Performance

We then switched to our modified version of VexRiscv, still strictly non-speculative but adding the `bb` instruction. We measured the benchmark programs in two ways: the gray line in Figure 8 (“BB Info”) uses our modified compiler to add `bb` instructions, and the blue line in Figure 8 (“BB Resched”) also uses our modified compiler to reschedule the terminator instruction in each basic block as early as possible.

Most of the non-speculation penalty, compared to the baseline (or to the baseline with branch prediction), is eliminated by switching from Simplest to BB Info. Note that the speedup from Simplest to BB Info varies throughout the different benchmark programs. Generally speaking, this speedup is higher for optimized code, such as *nettle-aes* and *nettle-sha*, than for unoptimized code. The overhead for BB Info compared to the baseline is only 3% for *nettle-sha* whereas it is 82% for *minver*.

We predicted that the size of the executed basic blocks would be a good predictor of performance in this setting, as the fetch unit needs to stall after all basic blocks until the next `bb` instruction is loaded. To check this hypothesis, we conducted a hotspot analysis of the benchmarks to discover the most frequented basic blocks in each program. The distribution of basic block sizes, weighted by the frequency of each block, is shown in Figure 9. It can be seen that well-performing benchmarks have a higher average block size compared to weaker-performing ones. The figures also show that already a few very large basic blocks can lead to decent

performance. As an example, for *nettle-sha* the mean basic block size is 68 instructions even if the median is at only 5 instructions, indicating a wide dispersion. Nevertheless, *nettle-sha* outperforms *crc32* which has a much higher median but lower arithmetic mean.

Switching from BB Info to BB Resched reschedules the terminator instruction in each basic block as early as possible, achieving further performance benefits (blue line in Figure 8). According to our benchmarks, rescheduling improves the performance by 15% on average compared to the version without rescheduling.

For highly branch-dependent code, such as *minver*, we still observe a performance overhead of 56% compared to the baseline CPU. However, less branch-dependent code achieves much better results. Interestingly, some benchmarks surpass the baseline CPU. Most notably, for *nettle-aes* and *nettle-sha*, our BasicBlocker-enabled CPU outperforms the best branch predictor in VexRiscv: it achieves a 3% and 5% speedup respectively, compared to the dynamic target predictor. One should not think that it is impossible for a non-speculative CPU to outperform a speculative CPU: if BasicBlocker allows a branch to be scheduled far enough in advance then it eliminates *all* pipeline stalls for that branch, whereas without BasicBlocker the same branch will sometimes be mispredicted even by a state-of-the-art branch predictor.

To analyze performance in more detail, we collected the average rescheduling parameter per basic block, defined by the average number of instructions following the terminator in-

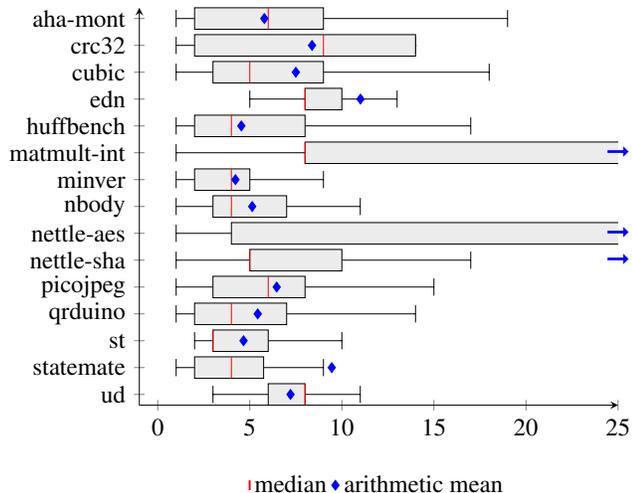


Figure 9: Distribution of basic block sizes (measured in instructions), weighted by the number of invocations derived from the hotspot analysis.

struction in a basic block and weighted it by the frequency of each block according to the hotspot analysis. Its distribution is depicted in Figure 10. Unsurprisingly, the benchmarks with the highest average rescheduling parameter achieved the best speedup. The results show that an average rescheduling by 4-5 instructions is sufficient for performance results that are comparable to the baseline CPU. We emphasize that this number is not the average rescheduling parameter but weighted by the number of invocations of the basic block. As most programs perform some kind of computations in their core functions, rescheduling is usually more easily feasible in these functions than in preamble and IO functions as basic block sizes are normally smaller in the latter. As shown in Figure 10, the *minver* benchmark reschedules terminator instructions in the core basic blocks by an average of 1.8 instructions. As this is not sufficient for seamless execution (at least, when speculative fetching is disabled), the observed performance overhead occurs. At the same time, the *edn* benchmark reschedules terminator instructions by an average of 5.5 instructions, hence achieving close-to-optimal performance.

Obviously, BasicBlocker slightly increases the code size as every basic block is extended by an additional instruction. The overhead directly depends on the average basic block size of a given program. Throughout our benchmarks, the average code size overhead was 17%.

Overall, the performance of Basic Blocker heavily depends on the executed program. We demonstrated that, for sufficiently optimized code, BasicBlocker enabled CPUs can yield performance comparable to traditional CPUs while dismissing any form of speculation. In some cases, BasicBlocker even outperforms sophisticated branch predictors.

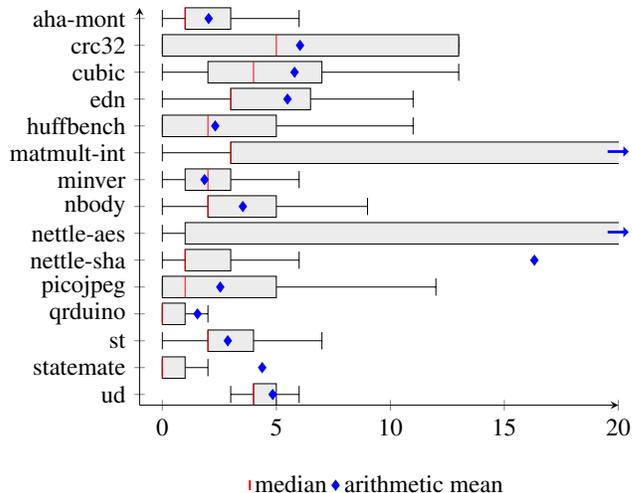


Figure 10: Distribution of instruction rescheduling per basic block, weighted by the number of invocations derived from the hotspot analysis.

6.3 Further Optimizations

We also evaluated some optimization techniques as described in the following sections.

6.3.1 Loop Counters

The BBRISC-V ISA, as defined earlier, supports the use of hardware loop counters. We experimented with a very restricted loop-placement policy in our proof-of-concept compiler. The restricted loop placement works well for the *edn* and *ud* benchmarks and the results demonstrate potential of loop counters as extension to ISAs modified with BasicBlocker.

The *edn* benchmark achieves a 21% speedup compared to the baseline CPU (simple rescheduling led to a 8% speedup; see Figure 8). This is equivalent to a 7% speedup compared to the best branch predictor of VexRiscv (dynamic target). The *ud* benchmark improves with a benchmark score 3% better than the baseline CPU, matching the dynamic target branch predictor.

For other benchmarks, the compiler either could not implement hardware loops or the introduction did not improve the score. For a few benchmarks, the introduction of hardware loops decreased the performance. We suspect that this especially happens for loops with very few iterations that occur within loops with a high number of iterations. Adverse instruction-cache alignment also contributes to cases where the hardware loop counter was not effective. We leave the ideal placement of hardware loops by the compiler as a problem for future work.

6.3.2 Early Branch

The VexRiscv core can be configured to forward the branch result after the execution stage (early branch) compared to the memory stage, thus requiring fewer stall cycles for non-rescheduled control flow instructions. This forwarding technique is already common on many CPUs but especially interesting when it comes to BasicBlocker. Our evaluation revealed that the early branch feature especially improves performance for programs that did not obtain top performance in our previous benchmarks. The reasons are obvious: these programs need to be stalled more frequently, and thus forwarding branch results reduces the amount of stall cycles.

The performance overhead of the *minver* benchmark reduces from 56% to 42% compared to the baseline CPU. The *picojpeg* benchmark improves from 24% to 19% overhead whereas the already well performing *nettle-aes* benchmark achieves only a minimal advantage of less than 1%.

6.3.3 Application-Level Optimizations

We changed the CPU, ISA, and compiler, but we did not change the applications. Some applications run at full speed on BasicBlocker CPUs, and an obvious question for future work is the extent to which other applications can be optimized for the execution paradigm of BasicBlocker CPUs. For example, quicksort has many branches, but it can be replaced by Batcher sort (or radix sort), which has predictable control flow and can be unrolled to use larger basic blocks.

7 Conclusion

Despite the consensus that speculation is inevitable, we demonstrate in this work that a reasonable alternative is possible. With BasicBlocker we propose a novel concept to transport control-flow information from the software to the hardware, enabling practical implementations of non-speculative CPUs. For high-performance code this leads to an increased execution speed compared to execution with the traditional concept of branch prediction, while fully eliminating any speculative-execution attacks.

We showcase our concept by specifying the BBRISC-V ISA, including a concrete implementation of that ISA based on the VexRiscv soft core, accompanied by an optimizing compiler that rests on the LLVM Compiler Framework. We stress, however, that BasicBlocker is a general solution that can be applied to other ISAs as well.

Our prototype implementation shows that our solution yields performance impacts that are similar to existing proposed countermeasures and moreover erases speculative attacks at their root, while alternative solutions often close only individual side channels.

We pointed out additional extensions to our concept and reasons to expect that further work will further improve per-

formance. Notably, hardware loop counters can be seamlessly integrated into our concept as explained in Appendix A, and are even more beneficial than in conventional architectures.

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy - EXC 2092 CASA - 390781972; by the Cisco University Research Program; and by the U.S. National Science Foundation under grant 1913167. “Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation” (or other funding agencies). Date of this document: 31 July 2020.

References

- [1] Martín Abadi, Mihai Budiu, Úlfar Erlingsson, and Jay Ligatti. Control-flow integrity. In *Proceedings of the 12th ACM Conference on Computer and Communications Security, CCS ’05*, page 340–353, New York, NY, USA, 2005. Association for Computing Machinery.
- [2] Martín Abadi, Mihai Budiu, Úlfar Erlingsson, and Jay Ligatti. Control-flow integrity principles, implementations, and applications. *ACM Transactions on Information and System Security (TISSEC)*, 13(1):1–40, 2009.
- [3] Onur Acıçmez, Billy Bob Brumley, and Philipp Grabher. New results on instruction cache attacks. In *International Workshop on Cryptographic Hardware and Embedded Systems*, pages 110–124. Springer, 2010.
- [4] Krste Asanović and David A. Patterson. Instruction sets should be free: The case for RISC-V, 2014. <https://people.eecs.berkeley.edu/~krste/papers/EECS-2014-146.pdf>.
- [5] Musard Balliu, Mads Dam, and Roberto Guanciale. Inspectre: Breaking and fixing microarchitectural vulnerabilities by formal analysis. *arXiv preprint arXiv:1911.00868*, 2019.
- [6] Daniel J. Bernstein. Cache-timing attacks on AES, 2005. <https://cr.yp.to/papers.html#cachetiming>.
- [7] Daniel J. Bernstein. Some small suggestions for the Intel instruction set, 2014. <https://blog.cr.yp.to/20140517-insns.html>.
- [8] Benjamin A Braun, Suman Jana, and Dan Boneh. Robust and efficient elimination of cache and timing side channels. *arXiv preprint arXiv:1506.00189*, 2015.

- [9] Claudio Canella, Daniel Genkin, Lukas Giner, Daniel Gruss, Moritz Lipp, Marina Minkin, Daniel Moghimi, Frank Piessens, Michael Schwarz, Berk Sunar, et al. Fallout: Leaking data on meltdown-resistant CPUs. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 769–784, 2019.
- [10] Claudio Canella, Jo Van Bulck, Michael Schwarz, Moritz Lipp, Benjamin von Berg, Philipp Ortner, Frank Piessens, Dmitry Evtushkin, and Daniel Gruss. A systematic evaluation of transient execution attacks and defenses. In *USENIX Security Symposium*, 2019. extended classification tree at <https://transient.fail/>.
- [11] Guoxing Chen, Sanchuan Chen, Yuan Xiao, Yinqian Zhang, Zhiqiang Lin, and Ten H Lai. SgxPectre: Stealing intel secrets from SGX enclaves via speculative execution. In *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 142–157. IEEE, 2019.
- [12] Douglas W. Clark and Henry M. Levy. Measurement and analysis of instruction use in the VAX-11/780, 1982. <https://dl.acm.org/doi/pdf/10.1145/1067649.801709>.
- [13] Robert J Colvin and Kirsten Winter. An abstract semantics of speculative execution for reasoning about security vulnerabilities. *arXiv preprint arXiv:2004.00577*, 2020.
- [14] Lucas Davi, Matthias Hanreich, Debayan Paul, Ahmad-Reza Sadeghi, Patrick Koeberl, Dean Sullivan, Orlando Arias, and Yier Jin. Hafix: Hardware-assisted flow integrity extension. In *2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, pages 1–6. IEEE, 2015.
- [15] Lucas Davi, Ahmad-Reza Sadeghi, Daniel Lehmann, and Fabian Monrose. Stitching the gadgets: On the ineffectiveness of coarse-grained control-flow integrity protection. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 401–416, 2014.
- [16] John A. DeRosa and Henry M. Levy. An evaluation of branch architectures. In Daniel C. St. Clair, editor, *Proceedings of the 14th Annual International Symposium on Computer Architecture. Pittsburgh, PA, USA, June 1987*, pages 10–16, 1987.
- [17] Scott DiPasquale, Khaled Elmeleegy, CJ Ganier, and Erik Swanson. Hardware loop buffering, 2003.
- [18] Agner Fog. The microarchitecture of Intel, AMD and VIA CPUs: An optimization guide for assembly programmers and compiler makers, 2020. <https://www.agner.org/optimize/>.
- [19] Free and Open Source Silicon Foundation. Embench IOT. <https://www.embench.org/>. Accessed: 2020-05-29.
- [20] Qian Ge, Yuval Yarom, and Gernot Heiser. No security without time protection: We need a new hardware-software contract. In *Proceedings of the 9th Asia-Pacific Workshop on Systems*, pages 1–9, 2018.
- [21] Abraham Gonzalez, Ben Korpan, Ed Younis, and Jerry Zhao. Spectrum: Classifying, replicating and mitigating spectre attacks on a speculating RISC-V microarchitecture. ., 2018.
- [22] Marco Guarnieri, Boris Köpf, Jan Reineke, and Pepe Vila. Hardware-software contracts for secure speculation. *arXiv preprint arXiv:2006.03841*, 2020.
- [23] Shay Gueron. Intel Advanced Encryption Standard (AES) new instructions set, 2010. <https://www.intel.com/content/dam/doc/white-paper/advanced-encryption-standard-new-instructions-set-paper.pdf>.
- [24] Khaled N. Khasawneh, Esmaeil Mohammadian Koruyeh, Chengyu Song, Dmitry Evtushkin, Dmitry Ponomarev, and Nael B. Abu-Ghazaleh. SafeSpec: Banishing the spectre of a meltdown with leakage-free speculation. *CoRR*, abs/1806.05179, 2018.
- [25] Vladimir Kiriansky and Carl Waldspurger. Speculative buffer overflows: Attacks and defenses. *arXiv preprint arXiv:1807.03757*, 2018.
- [26] Paul Kocher, Jann Horn, Anders Fogh, Daniel Genkin, Daniel Gruss, Werner Haas, Mike Hamburg, Moritz Lipp, Stefan Mangard, Thomas Prescher, et al. Spectre attacks: Exploiting speculative execution. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 1–19. IEEE, 2019.
- [27] Esmaeil Mohammadian Koruyeh, Khaled N Khasawneh, Chengyu Song, and Nael Abu-Ghazaleh. Spectre returns! speculation attacks using the return stack buffer. In *12th USENIX Workshop on Offensive Technologies (WOOT 18)*, 2018.
- [28] Chris Lattner and Vikram Adve. LLVM: A compilation framework for lifelong program analysis & transformation. In *International Symposium on Code Generation and Optimization, 2004. CGO 2004.*, pages 75–86. IEEE, 2004.
- [29] Moritz Lipp, Michael Schwarz, Daniel Gruss, Thomas Prescher, Werner Haas, Anders Fogh, Jann Horn, Stefan Mangard, Paul Kocher, Daniel Genkin, et al. Meltdown: Reading kernel memory from user space. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 973–990, 2018.

- [30] Giorgi Maisuradze and Christian Rossow. ret2spec: Speculative execution using return stack buffers. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 2109–2122, 2018.
- [31] Charles Papon. VexRiscv. <https://github.com/SpinalHDL/VexRiscv>. Accessed: 2020-05-28.
- [32] Ivan Puddu, Moritz Schneider, Miro Haller, and Srdjan Capkun. Frontal attack:leaking control-flow in SGX via the CPU frontend. *arXiv preprint arXiv:2005.11516*, 2020.
- [33] Praveen Raghavan, Andy Lambrechts, Murali Jayapala, Francky Catthoor, and Diederik Verkest. Distributed loop controller for multithreading in unithreaded ILP architectures. *IEEE Transactions on Computers*, 58(3):311–321, 2008.
- [34] Michael Schwarz, Moritz Lipp, Daniel Moghimi, Jo Van Bulck, Julian Stecklina, Thomas Prescher, and Daniel Gruss. Zombieload: Cross-privilege-boundary data sampling. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 753–768, 2019.
- [35] Michael Schwarz, Robert Schilling, Florian Kargl, Moritz Lipp, Claudio Canella, and Daniel Gruss. Context: Leakage-free transient execution. *arXiv preprint arXiv:1905.09100*, 2019.
- [36] Hovav Shacham. The geometry of innocent flesh on the bone: Return-into-libc without function calls (on the x86). In *Proceedings of the 14th ACM conference on Computer and communications security*, pages 552–561, 2007.
- [37] Kevin Z Snow, Fabian Monrose, Lucas Davi, Alexandra Dmitrienko, Christopher Liebchen, and Ahmad-Reza Sadeghi. Just-in-time code reuse: On the effectiveness of fine-grained address space layout randomization. In *2013 IEEE Symposium on Security and Privacy*, pages 574–588. IEEE, 2013.
- [38] Jakub Szefer. Survey of microarchitectural side and covert channels, attacks, and defenses. *Journal of Hardware and Systems Security*, 3(3):219–234, 2019.
- [39] Andrew S Tanenbaum. *Structured computer organization*. Pearson Education India, 2016.
- [40] Michael Theodorides and David A. Wagner. Breaking active-set backward-edge CFI. In *2017 IEEE International Symposium on Hardware Oriented Security and Trust, HOST 2017, McLean, VA, USA, May 1-5, 2017*, pages 85–89. IEEE Computer Society, 2017.
- [41] Paul Turner. Retpoline: A software construct for preventing branch-target-injection. URL <https://support.google.com/faqs/answer/7625886>, 2018.
- [42] Jo Van Bulck, Marina Minkin, Ofir Weisse, Daniel Genkin, Baris Kasikci, Frank Piessens, Mark Silberstein, Thomas F Wenisch, Yuval Yarom, and Raoul Strackx. Foreshadow: Extracting the keys to the intel SGX kingdom with transient out-of-order execution. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 991–1008, 2018.
- [43] Stephan van Schaik, Alyssa Milburn, Sebastian Osterlund, Pietro Frigo, Giorgi Maisuradze, Kaveh Razavi, Herbert Bos, and Cristiano Giuffrida. Addendum to RIDL: Rogue in-flight data load. <https://mdsattacks.com/>, 2019.
- [44] Stephan Van Schaik, Alyssa Milburn, Sebastian Österlund, Pietro Frigo, Giorgi Maisuradze, Kaveh Razavi, Herbert Bos, and Cristiano Giuffrida. RIDL: Rogue in-flight data load. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 88–105. IEEE, 2019.
- [45] Venkatanathan Varadarajan, Thomas Ristenpart, and Michael Swift. Scheduler-based defenses against cross-VM side-channels. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 687–702, 2014.
- [46] Marco Vassena, Klaus V Gleissenthall, Rami Gökhan Kici, Deian Stefan, and Ranjit Jhala. Automatically eliminating speculative leaks with blade. *arXiv preprint arXiv:2005.00294*, 2020.
- [47] Pepe Vila, Andreas Abel, Guarnieri, Boris Köpf, and Jan Reineke. Flushgeist: Cache leaks from beyond the flush. *arXiv preprint arXiv:2005.13853*, 2020.
- [48] Ofir Weisse, Jo Van Bulck, Marina Minkin, Daniel Genkin, Baris Kasikci, Frank Piessens, Mark Silberstein, Raoul Strackx, Thomas F. Wenisch, and Yuval Yarom. Foreshadow-NG: Breaking the virtual memory abstraction with transient out-of-order execution. *Technical report*, 2018. See also USENIX Security paper Foreshadow.
- [49] Nils Wistoff, Moritz Schneider, Frank K. Gürkaynak, Luca Benini, and Gernot Heiser. Prevention of microarchitectural covert channels on an open-source 64-bit RISC-V core. *arXiv preprint arXiv:2005.02193*, 2020.
- [50] Jiyong Yu, Lucas Hsiung, Mohamad El Hajj, and Christopher W Fletcher. Data oblivious ISA extensions for side channel-resistant and high performance computing. In *NDSS*, 2019.

- [51] Drew Zagieboylo, G Edward Suh, and Andrew C Myers. Using information flow to design an isa that controls timing channels. In *2019 IEEE 32nd Computer Security Foundations Symposium (CSF)*, pages 272–27215. IEEE, 2019.
- [52] Yinqian Zhang and Michael K Reiter. Düppel: Retrofitting commodity operating systems to mitigate cache side channels in the cloud. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 827–838, 2013.
- [53] Lutan Zhao, Peinan Li, Rui Hou, Jiazhen Li, Michael C Huang, Lixin Zhang, Xuehai Qian, and Dan Meng. A lightweight isolation mechanism for secure branch predictors. *arXiv preprint arXiv:2005.08183*, 2020.

A Loop Counter

Loops are often the execution hotspots in programs and contribute considerably to diverging control flow. Therefore the concept of hardware supported loops can be profitable as already discussed in the literature [17, 33] and implemented in various architectures.

In general, hardware loop counters are realized by a hardware counter which is set by a dedicated instruction with a value representing the maximum trip count for the loop. The trip count must be computable at compile time to be inserted by an immediate value or available in a register at run-time before entering the loop. Information about which instructions are included in the loop is expressed via labels or additional specific instructions. The hardware loop counter decrements the start value after each iteration and induces a branch back to the start of the loop as long as the counter is unequal to zero. This can be done implicitly at the end of the loop or explicitly with an instruction.

Performance improvements by the usage of hardware loops result from reduced instruction size and dedicated loop control logic that does not have to be calculated by the ALU. For our BasicBlocker concept, hardware loops are actually much more valuable for performance when only applied to loops that will not terminate early, because in this case the control flow for all loop iterations is known when entering the loop.

We can seamlessly support hardware loop counters in our design concept, by introducing a new instruction and adding two arguments to the `bb` instruction. The `lcnt` sets the number of loop iterations by storing a specified value into a dedicated register. The start and end address of the loop are encoded into the `bb` instruction, by indicating with two separate flags whether the corresponding basic block is the start or end block of the loop. These two flags in the `bb` instruction are necessary for each loop counter set, which means that the `bb` instruction needs $2n$ bits to support n loop counter sets.

```
bb 2, 1, 00, 00 ; len = 2, seq = 1
```

```
add a0, a0, a1          2
lcnt 3, lc1 ; 3 iterations, set 1  3
bb 2, 0, 01, 01 ; loop start/end  4
add a1, a2, a2         5
mul a2, a1, a2         6
bb 7, 0, 00, 00 ; after loop      7
```

Listing 1: Single basic block loop with 3 iterations in counter set 1; Colors correspond to the execution trace in 2

Listing 1 shows the exemplary use of the hardware loop counter. In line 3, the counter in loop set `ls1` is initialized to 3. The following `bb` instruction has the start- and end flag for loop set 1 enabled which indicates a loop that starts at the beginning of this basic block and stretches until the end of the same basic block. Each bit in the flags represents one loop counter set, allowing nested loops with the same start- or end address and nested loops sharing the same basic block as start or end. It is possible to model loops that stretch across multiple basic blocks by setting the start and end flags in the respective basic blocks accordingly. When the `bb` instruction with the start flag is executed, the current `PC` is saved as start address in the corresponding loop counter set. Simultaneously, the counter value of that set is decremented by one. When the execution reaches the `bb` instruction with the corresponding end flag, the target address (which determines where the CPU continues execution) is set to the corresponding start address if the counter is not zero. Otherwise, the basic block is handled like a normal sequential basic block and the loop will exit.

```
bb 2, 1, 00, 00 ; len = 2, seq. block
bb 2, 0, 01, 01 ; loop: start L1, end L1
add a0, a0, a1
lcnt 2, lc1 ; 2 iterations, set 1
bb 2, 0, 01, 01 ; loop: start L1, end L1
add a1, a2, a2
mul a2, a1, a2
bb 2, 0, 01, 01 ; loop: start L1, end L1
add a1, a2, a2
mul a2, a1, a2
bb 7, 0, 00, 00 ; after loop
add a1, a2, a2
mul a2, a1, a2
```

Listing 2: Execution trace of CPU with color matched instructions to the code sequence in 1.

In Listing 2, the instruction trace of the program snippet from Listing 1 is shown as it is executed by the CPU. Since the first `bb` instruction indicates a sequential basic block, the CPU immediately fetches the `bb` instruction of the next basic block which notifies the fetch unit that the second basic block is the start and end block of the loop. After that, the remaining `add` and `lcnt` instructions are executed to finish the first basic

block. From now on the loop counter determines the execution flow. Since the second basic block is the only basic block of the loop, the `bb` instruction of this block is fetched again, to prepare the second loop round, before the basic block is executed to complete the first round. This happens again until the loop counter is zero, resulting in fetching the last `bb` instruction, to exit the loop, before the last round of the loop is executed. Afterwards the execution continues outside of the loop with the normal instruction flow.

B Compile Flags

The compile flags for the Coremark benchmark are listed in fig. 11 (omitting includes, debug, macros, toolchain paths and flags enabling `bb` instructions).

Flag	Description
O3	Optimization Level 3
march=rv32im	32-Bit RISC-V with IM extensions
mabi=ilp32	Calling convention and memory layout
target=riscv32-unknown-elf	Select target architecture
mno-relax	No linker relaxation
lgcc	Link gcc library
lc	Link C library
nostartfiles	Do not use standard system startup files when linking.
ffreestanding	Only use features available in freestanding environment.

Figure 11: Coremark compile flags.

The compile flags for the Embench benchmarks are listed in fig. 12 (omitting includes, debug, macros, toolchain paths and flags enabling `bb` instructions).

C Raw Benchmark Results

The following tables (figs. 13 to 19) list the mean result of each benchmark as well as the upper and lower quartiles.

Flag	Description
O3	Optimization Level 3
march=rv32im	32-Bit RISC-V with IM extensions
mabi=ilp32	Calling convention and memory layout
target=riscv32-unknown-elf	Select target architecture
mno-relax	No linker relaxation
fno-inline	No function inlining
fno-common	Individual zero-initialized definitions for tentative definitions
fno-strict-aliasing	Disable strict aliasing.

Figure 12: Embench IOT compile flags.

Dynamic Target BP Benchmark			
Benchmark	Quartile	Median	Quartile
Coremark	5175115.75	5175286.50	5175482.00
aha-mont	3638630.50	3638673.00	3638725.25
crc32	3906593.25	3906655.00	3906736.75
cubic	140243.50	140319.50	140398.50
edn	5115438.00	5115498.50	5115674.00
huffbench	3370380.00	3370577.50	3370782.50
matmult-int	4188588.75	4188830.00	4189087.75
minver	355841.75	355906.00	355974.00
nbody	237888.75	237924.00	237950.25
nettle-aes	4838928.00	4839159.00	4839439.25
nettle-sha	5953509.50	5953784.00	5954277.50
picojpeg	5810301.00	5810784.50	5811234.25
qrduino	4952491.75	4952667.50	4952831.00
st	420650.75	420692.50	420752.75
statemate	6987615.00	6988031.00	6988543.50
ud	3302171.00	3302301.50	3302425.25

Figure 13: Benchmark results for the dynamic target branch predictor.

Dynamic BP Benchmark			
Benchmark	Quartile	Median	Quartile
Coremark	5512553.50	5512829.00	5513042.00
aha-mont	3767784.75	3767849.00	3767901.00
crc32	4947929.50	4948035.50	4948146.25
cubic	158541.50	158607.00	158678.00
edn	5424041.50	5424202.00	5424308.00
huffbench	3591732.00	3591902.00	3592047.25
matmult-int	4246322.75	4246702.50	4246892.00
minver	489098.25	489180.50	489250.25
nbody	293125.00	293147.50	293177.50
nettle-aes	4877322.00	4877759.00	4878034.25
nettle-sha	5975359.25	5975804.00	5976259.75
picojpeg	6742694.50	6743166.50	6743513.25
qrduino	5227886.25	5228250.50	5228634.00
st	581956.50	582017.00	582060.50
statemate	7155070.00	7164739.00	7166616.25
ud	3357320.25	3357548.00	3357647.00

Figure 14: Benchmark results for the dynamic branch predictor.

No BP Benchmark			
Benchmark	Quartile	Median	Quartile
Coremark	6153095.50	6153312.50	6153476.50
aha-mont	4029828.75	4029910.00	4029959.00
crc32	5244182.00	5244251.50	5244416.50
cubic	159655.00	159716.50	159812.50
edn	6016278.50	6016387.50	6016540.50
huffbench	4064054.50	4064262.50	4064423.75
matmult-int	4369878.50	4370210.00	4370477.75
minver	521680.50	521731.00	521779.50
nbody	294201.75	294227.50	294250.50
nettle-aes	4949798.00	4950089.00	4950382.00
nettle-sha	6015353.00	6015439.00	6015511.00
picojpeg	7374206.75	7374653.50	7375019.75
qrduino	5494554.75	5494723.00	5494916.25
st	596986.00	597026.50	597088.00
statemate	7521163.75	7521259.00	7521349.25
ud	3413756.75	3413887.00	3413999.25

Figure 16: Benchmark results for original VexRiscv without branch predictor.

Static BP Benchmark			
Benchmark	Quartile	Median	Quartile
Coremark	5707389.50	5707595.00	5707812.00
aha-mont	4011169.50	4011228.50	4011292.25
crc32	4947964.25	4948049.50	4948129.25
cubic	158730.75	158829.50	158909.25
edn	5424664.00	5424760.50	5424861.50
huffbench	3984886.75	3985119.00	3985295.50
matmult-int	4246659.50	4246880.00	4247029.25
minver	496345.75	496406.50	496461.00
nbody	293848.75	293872.50	293893.00
nettle-aes	4900023.75	4900246.00	4900655.75
nettle-sha	5985669.75	5985929.00	5986112.75
picojpeg	7145492.00	7145872.00	7146270.25
qrduino	5622921.50	5623074.00	5623194.75
st	581947.50	581997.50	582049.00
statemate	7446324.25	7446531.50	7446876.50
ud	3372643.50	3372814.50	3372945.00

Figure 15: Benchmark results for the static target branch predictor.

Simplest Benchmark			
Benchmark	Quartile	Median	Quartile
Coremark	15707821.00	15708006.00	15708178.50
aha-mont	13845826.00	13845890.00	13845940.25
crc32	15351114.75	15351222.00	15351352.50
cubic	416312.50	416381.00	416454.25
edn	14654086.00	14654206.00	14654305.50
huffbench	11388765.00	11388930.00	11389122.75
matmult-int	9705278.00	9705636.00	9706117.25
minver	1376270.00	1376335.00	1376398.50
nbody	764324.75	764348.00	764380.00
nettle-aes	13521771.00	13522278.50	13522697.25
nettle-sha	15022300.25	15022661.00	15022907.00
picojpeg	18013845.75	18014134.00	18014463.25
qrduino	15566296.00	15566433.00	15566590.00
st	1458251.75	1458300.00	1458369.00
statemate	17975289.75	17975633.00	17975993.00
ud	7836332.25	7836479.50	7836621.00

Figure 17: Benchmark results for strictly non-speculative VexRiscv processor.

BB Info Benchmark			
Benchmark	Quartile	Median	Quartile
Coremark	10623034.75	10623287.50	10623461.75
aha-mont	7071002.00	7071046.00	7071099.25
crc32	7029190.50	7029325.50	7029424.50
cubic	235699.25	235767.00	235862.50
edn	7308679.50	7308801.50	7308896.50
huffbench	6815852.00	6816058.00	6816280.00
matmult-int	4594765.00	4595057.00	4595333.00
minver	946888.25	946931.00	947003.75
nbody	436978.00	437002.50	437019.25
nettle-aes	5269032.75	5269299.00	5269614.50
nettle-sha	6212643.75	6212790.00	6213101.25
picojpeg	10633148.00	10633705.50	10634233.75
qrduino	8990960.00	8991134.00	8991317.75
st	892666.75	892714.50	892770.00
statemate	7716999.00	7717064.00	7717157.00
ud	4417475.25	4417626.50	4417795.75

Figure 18: Benchmark results for BasicBlocker VexRiscv with BB instructions.

Rescheduling Benchmark			
Benchmark	Quartile	Median	Quartile
Coremark	9595350.25	9595544.00	9595852.00
aha-mont	6050459.75	6050514.50	6050567.25
crc32	5544510.25	5544697.50	5544828.75
cubic	184084.50	184177.00	184288.00
edn	5592789.75	5592957.50	5593105.00
huffbench	6088287.00	6088417.00	6088584.25
matmult-int	4348148.25	4348394.50	4348617.75
minver	816160.50	816273.00	816365.25
nbody	317853.25	317880.00	317907.00
nettle-aes	4737261.25	4737534.50	4737896.75
nettle-sha	5978782.25	5978992.50	5979122.00
picojpeg	9178532.00	9178841.00	9179823.00
qrduino	8396670.00	8396835.00	8396984.25
st	709112.00	709165.00	709209.50
statemate	7334966.00	7335053.50	7335164.75
ud	3438196.00	3438279.50	3438411.00

Figure 19: Benchmark results for BasicBlocker VexRiscv with BB instructions and rescheduling of branches.