Reorder Buffer Contention: A Forward Speculative Interference Attack for Speculation Invariant Instructions

Pavlos Aimoniotis, Christos Sakalis,
Magnus Själander, and Stefanos Kaxiras

Abstract—Speculative side-channel attacks access sensitive data and use transmitters to leak the data during wrong-path execution. Various defenses have been proposed to prevent such information leakage. However, not all speculatively executed instructions are unsafe: Recent work demonstrates that *speculation invariant*instructions are independent of speculative control-flow paths and are guaranteed to eventually commit, regardless of the speculation outcome. Compile-time information coupled with run-time mechanisms can then selectively lift defenses for speculation invariant instructions, reclaiming some of the lost performance. Unfortunately, speculation invariant instructions can easily be manipulated by a form of *speculative interference* to leak information via a new side-channel that we introduce in this paper. We show that *forward* speculative interference where *older* speculative instructions interfere with *younger speculation invariant instructions* effectively turns them into transmitters for secret data accessed during speculation. We demonstrate *forward speculative interference* on actual hardware, by selectively filling the reorder buffer (ROB) with instructions, pushing speculative invariant instructions in-or-out of the ROB *on demand*, based on a speculatively accessed secret. This reveals the speculatively accessed secret, as the occupancy of the ROB itself becomes a new speculative side-channel.

Index Terms—Speculative side-channel attacks, security, spectre, speculative interference

1 Introduction

Speculative side-channel attacks use speculative execution to gain access to information that would otherwise be inaccessible. Speculatively executed instructions are capable of temporarily bypassing hardware or software defenses to gain illegal access to data that are then passed to speculative side-channel instructions, a *transmitter gadget*, capable of leaking those sensitive data to the non-speculative domain. Transmitter gadgets perform an operation that alters the microarchitectural state of the processors, leading to a data leak. A receiver observes the changes in the microarchitectural states and is able to identify leaked data outside of the speculation window.

To tackle this problem several hardware defenses [3], [4], [6], [8], [9], [11], [12], [13], [15] have been proposed, introducing a variety of security guarantees. However, defenses also introduce various levels of complexity and performance overhead. Several hardware defenses rely on techniques that protect instructions while they are speculative, and focus on making them invisible. One example is Delay-on-Miss (DoM) [11]. DoM delays speculative loads that miss in the L1 cache until they become non-speculative, at which point they can be executed safely. Another example is InvisiSpec [13]. InvisiSpec performs speculative loads but keeps the effects of a miss invisible in the cache hierarchy. When the speculation is verified, changes in the memory hierarchy are effected with a visible access.

Hardware defenses, such as DoM and InvisiSpec, add significant performance overhead [11], [13]. For this, Zhao *et al.* proposed InvarSpec [16], a framework that detects and lifts the protection for speculative instructions that become *speculation invariant*. For an instruction to be speculative invariant, its data and control dependencies must be resolved during the speculation window. Such instructions are eventually going to execute with the same operands, even if they are temporarily squashed due to misspeculation, and are, thus, considered safe to execute. Lifting the protection for speculation invariant instructions enables the visible execution of an instruction while it is still under speculation, maintaining the "invisible speculative execution" semantics of defenses such as DoM or InvisiSpec while recovering significant performance lost to these defenses.

In a related development, Behnia *et al.* demonstrate that *Speculative Interference* [5] can break (under some assumptions) the DoM and InvisiSpec defenses. Up until now, the transmitter instructions were considered to be exclusively under speculative execution. With the introduction of Speculative Interference attacks, this has changed. In such an attack, the transmitter instructions are placed before (in program order) the speculation window. Hence, the transmitter instructions can lie outside the protection of DoM or InvisiSpec defenses, as these are engaged only for instructions that follow (in program order) the source-of-speculation instruction(s). Since Speculative Interference is based on the fact that younger speculative instructions can influence the timing of older instructions, it can consequently lead to information leakage even under speculative defense mechanisms [5].

The key insight of our work is that speculation-invariant instructions are susceptible to speculative interference from *older* speculative instructions: *Forward Speculative Interference* (FSI). To clearly differentiate between FSI and the speculative interference from *younger* speculative instructions, we refer to the latter as Backward Speculative Interference (BSI). Using FSI, a new side-channel can be created by manipulating the inclusion or exclusion of speculation-invariant instructions in the reorder buffer (ROB). Other forms of forward interference are also possible and Behnia *et al.* [5] discuss how to delay instruction fetch with reservation station (RS) contention, called $G^I{}_{RS}$ in [5]. However, $G^I{}_{RS}$ concerns blocking of instruction fetch (and the front-end) which affects the I-Cache and is distinctly different from the *ROB-contention* interference discussed here that concerns instruction execution.

We demonstrate FSI with *ROB contention* on actual processors (Intel Sandy Bridge) and show how the ROB can be used as a side-channel. Specifically, we show how, during speculation, we can selectively push in-or-out of the ROB load instructions that are on the—yet unknown—correct path of execution, leading to side-effects that remain observable after the speculation has been resolved. These load instructions would be marked as speculative-invariant by InvarSpec, therefore the InvarSpec framework is susceptible to such a side-channel attack as well.

2 Background

2.1 Delay-on-Miss

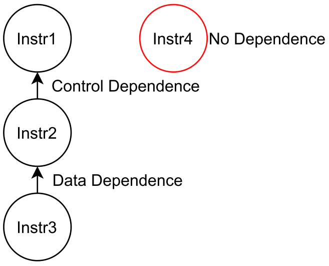
Delay-on-Miss (DoM) is a hardware defense mechanism against speculative side-channel attacks, focusing on side-channels that abuse the memory hierarchy [11]. Consecutively, side-channel attacks that do not focus on the memory hierarchy are outside the scope of DoM and are not hindered by it.

- *Pavlos Aimoniotis, Christos Sakalis, and Stefanos Kaxiras are with Uppsala University, 752 36 Uppsala, Sweden. E-mail: {pavlos.aimoniotis, christos.sakalis, stefanos.kaxiras}@it.uu.se.*
- *Magnus Själander is with the Norwegian University of Science and Technology, 7491 Trondheim, Norway. E-mail: magnus.sjalander@ntnu.no.*

Manuscript received 14 Sept. 2021; accepted 22 Oct. 2021. Date of publication 27 Oct. 2021; date of current version 9 Nov. 2021.
This work was supported in part by Microsoft Research through its EMEA PhD Scholarship Programme under Grant 2021-020 and in part by the Swedish Research Council under Grants 2015-05159 and 2018-05254.
(Corresponding author: Pavlos Aimoniotis.)
Digital Object Identifier no. 10.1109/LCA.2021.3123408

```
if(cond){      // instr1
 si = load i; // instr2
}
a = load si;  // instr3
b = load j;   // instr4
```

(a) Source Code



(b) Instruction dependence graphs

Fig. 1. Dependences related to safe set (SS).

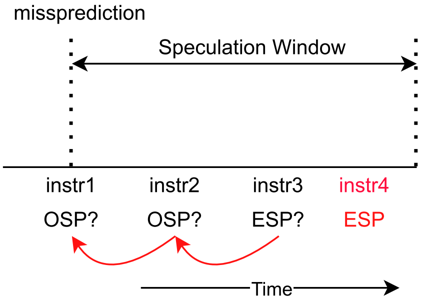


Fig. 2. Speculation invariant timeline: For *instr3* to be considered speculation invariant, *instr2* and *instr1* must reach their OSP. *Instr4* has no dependences, and executes immediately under speculation using InvarSpec framework.

DoM operates on two fundamental principles. First, DoM delays transient loads until they become non-speculative. DoM introduces the concept of *speculative shadows* to efficiently track the speculative state of instructions and discover the earliest time instructions become non-speculative, typically significantly earlier than reaching the commit stage (becoming head of the reorder buffer).

Second, DoM delays only loads that miss in the cache. Because reading data into a cache requires complicated interactions with the rest of the system, it is difficult to hide the side-effects of loads in the memory hierarchy on a cache miss, as demonstrated in prior solutions such as InvisiSpec [13] and Ghost Loads [10]. However, a cache hit requires only small modifications to the cache state (update of the replacement state etc.), which can be easily deferred for when the load is non-speculative. Thus, instead of delaying all loads, DoM allows loads that hit in L1 cache to execute under speculation, while delaying any side-effects until the load becomes non-speculative.

2.2 Speculation Invariance: InvarSpec

InvarSpec is a framework that detects when a speculative instruction becomes *speculation invariant* and upon detection lifts any existing protections for the instruction [16]. InvarSpec consists of two main parts. The first part is a compiler technique that after static analysis generates a *safe set* (SS) for the instructions. The second part is a hardware mechanism that at runtime designates an *execution-safe point* (ESP) according to the SS.

An example of speculation invariance is shown in Fig. 1, where a (*instr3*) has a potential data dependence with *instr2*, and *instr2* has a control dependence with *instr1*. In order for *instr3* to become speculation invariant, it must reach its execution safe point, meaning both *instr1* and *instr2* must reach their *outcome safe point*. Since *instr4* has no data nor control dependencies with any other instruction (its SS is empty) it can execute immediately.

Each instruction has its safe set (SS) defined by the compiler and corresponds to the instruction's control and data dependences on the instructions in the set [16]. The SS is used to determine at runtime when an instruction is ready and safe to execute during speculative execution. An instruction is considered to be speculation invariant when it reaches its execution-safe point (ESP). To reach the ESP, the operands of an instruction must have been finalized. Older instructions that comply with these rules are said to have reached their outcome-safe point (OSP), meaning that their final result will not change, no matter how many future squashes may happen. When everything in the safe set reaches the outcome-safe point, the instruction itself has reached the execution-safe point and the speculative side-channel defense mechanisms can be lifted for the instruction to be executed, even if the speculation has not been verified.

Fig. 2 shows the timeline of an instruction using InvarSpec framework. As a reminder, an instruction is said to have reached its ESP when all its operands reach their OSP. Once the instruction is ready to be executed, even if the speculation has not been resolved, the defense mechanisms are lifted and the instruction executes.

2.3 Backward Speculative Interference

Speculative Interference attacks [5] are able to break defense mechanisms similar to DoM and InvisiSpec. Even though speculative loads are executed invisibly, misspeculated instructions can change the timing of older instructions that *may be outside the protection of DoM or InvisiSpec as non-speculative instructions*. This change can influence the ordering of memory operations that will be committed, setting the fundamentals for a possible attack.

For example, assume that the interference target is a load that takes X cycles before its operand becomes ready. The interference gadget can then use the secret value to selectively add contention in the MSHRs. For example, if the secret is equal to 1, the interference gadget attempts to fill all MSHR entries before the interference target is ready to execute. Otherwise, if the secret is equal to 0, no memory operations are performed by the interference gadget. Once the interference target becomes ready to execute, if the secret was 1 it will be further delayed, otherwise, if the secret was 0, it will be executed unhindered. This difference in behavior can lead to information leakage as it can affect the order of the interference target with respect to other loads, and thus affect the cache replacement state.

3 ROB-Contention: An FSI Attack That Breaks Speculative Invariance

Speculation invariance allows (bound-to-commit) speculative instructions to be executed without defenses before the speculation is verified. In this respect, speculation-invariant instructions behave the same as the corresponding instructions in an unprotected processor.

In *Backward Speculative Interference*, the interference gadget delays the execution of the interference target, a bound-to-commit instruction that is placed *prior* to the speculation. In *Forward Speculative Interference*, the interference gadget instead interferes with a bound-to-commit speculation-invariant instruction, which is executed *while still under speculation*, unprotected by defense mechanisms like DoM [11] or InvisiSpec [13].

While FSI can take many forms, in this paper we introduce a novel side-channel based on manipulating ROB contention. To the best of our knowledge, this has not been explored previously. The ROB side-channel can be used to construct new Spectre [7] variants on unprotected processors, but more importantly, it can break InvarSpec approaches [16] that selectively lift defenses of instructions under speculation. Assuming DoM as the underlying defense mechanism—other defenses, such as InvisiSpec, are similarly susceptible—an FSI ROB-contention attack consists of three parts:

1) A branch predictor that is trained to follow the attack path.
2) A secret that is read from the cache (allowed in DoM) and ROB contention, as a function of the secret value, is added.

```
if(value){ // mispredict - Attack Path

 secret = secret << 10; // Repetition factor

 // Pass secret to ECX and execute rep
 asm("movl␣%0,␣%%ecx" : : "c" (secret));
 asm("rep␣movsb");
}
else { // Normal Path
 t1 = __rdtscp(); // Start measuring latency
 transmitter = probe[0]; // Evaluation
 t2 = __rdtscp(); // End measuring latency
 t = t2-t1;
}

transmitter = probe[0]; // Recovergence Point
```

Fig. 3. Abusing InvarSpec with forward speculative interference using REP instruction.

3) A speculation-invariant target instruction that resides just after the reconvergence point and that is executed with the DoM protections *lifted*. We initialize the speculation-invariant instruction with an empty safe set, i.e., a set that has no dependencies and can execute immediately when it becomes ready.

Depending on the contention-induced delay, and thus on the secret value, the speculation invariant target instruction will be affected in terms of *when it will be ready to execute*. For example, when the secret is equal to 1, we add extra ROB contention, in the form of a loop or a long sequence of spurious instructions. As a result, the ROB is filled with speculative instructions, which prevents the speculation-invariant target instruction from even entering the ROB and executing. On the other hand, the path followed when the secret is 0 behaves normally, enabling the speculation-invariant target instruction to execute when it enters the ROB. Since InvarSpec has lifted the defenses from the instruction, any side-effects caused by its execution will remain observable even after the misspeculation has been detected and squashed, making it possible to infer the secret value outside of the speculative window.

While the FSI ROB-contention attack shares some similarities with the ${G^I}_{RS}$ speculative interference attack, described by Behnia *et al.* [5], it is distinctly different in a number of ways: First, in contrast to ${G^I}_{RS}$, ROB-contention manipulates the execution of bound-to-commit loads (which lie after the reconvergence point) rather than instruction fetch. As such, ROB-contention directly affects mitigations such as DoM or InvisiSpec (when combined with InvarSpec) that aim to protect *data caches* from leaking information, which is not a concern with ${G^I}_{RS}$: ${G^I}_{RS}$ uses the instruction cache as a side-channel—ROB-contention uses the data cache. Second, ${G^I}_{RS}$ must cause a front-end stall to work. ROB-contention works as long as a target instruction is kept just outside the ROB, which does not necessarily mean a front-end stall. For example, if the target instruction is sufficiently far from the reconvergence point, the front end will keep fetching and decoding instructions from the reconvergence point onwards.

The technique of identifying the secret can be thought-of as a version of the *Flush&Reload* attack [14]. It is shown in Fig. 3 and is based on testing if data are cached in the L1 cache or not.

To achieve this, we measure the access time of the speculation-invariant target instruction when the speculation is finally resolved and the execution continues from the correct path. While on the misspeculated attack path, whether the load instruction at the reconvergence point will be executed depends on which path the speculative execution followed, i.e., it depends on if the secret is 0 or 1. Then, on the correct path, the time it takes to execute the load will change depending on if the data was loaded by the attack path, thus making it possible to infer the secret value.

The attack starts by ensuring that the address of the speculation-invariant target instruction is flushed from the cache. If the secret is equal to 1 then the speculative-invariant target instruction is never executed along the incorrect path. Once the speculation is resolved and the correct path is taken a load with the same address as the speculative-invariant target instruction will miss in the cache and experience a long delay. If the secret is equal to 0 then the speculative-invariant target instruction is executed in the incorrect path and the load in the correct path will hit in the cache and experience a short delay.

4 ROB Attack Using REP Instructions

An FSI ROB-contention attack requires filling the ROB with speculative instructions. While either a tight loop, or a long sequence of spurious instructions, fit the bill for this purpose, interestingly, one can achieve the same result with a *single static instruction*. In the x86 ISA, `REP` is a prefix that can be used before string instructions. It creates a single-instruction loop, with the value stored in the `ECX` register acting as the loop counter.

The key property that enables a single `REP` instruction to affect ROB contention is that it unrolls as a μop loop in the microarchitecture, at *decode time* [1]. ROB occupancy becomes a function of `ECX`.

According to empirical studies [1], [2], `REP`-prefixed x86 instructions expand into a number of μops in the ROB.

The following table lists the μop expansion (number of μops generated with `ECX==`n) in the ROB for two typical `REP` instructions and for some well-known microarchitectures—similar expansion takes place for the majority of x86 microarchitectures [1].

Instr./Proc.	Haswell	Broadwell	Skylake	IceLake
`rep movs`	$2n$	$2n$	$2n$	$2n$
`rep lods`	$5n$+12	$5n$+12	$5n$+12	$5n$+12

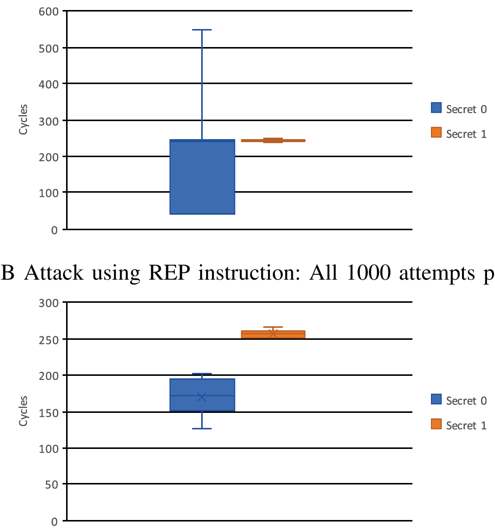
Furthermore, we ascertain that the `REP movs` instruction expands *speculatively* on a Sandy Bridge microarchitecture. We tested this scenario by giving `ECX` various values, after a speculation point, followed by a `REP` instruction (as in the code shown in Fig. 3). By timing the code, we observe that the `REP` instruction, indeed, expands speculatively into a number of μops that is proportional to `ECX`.

To mount a ROB attack with `REP` instructions (Fig. 3), we use the speculatively-accessed secret to update the `ECX` register, which then controls the number of μops that are dispatched to the ROB. To create a large enough repetition factor, we left-shift the secret by, e.g., ten places (if the secret is zero, it does not change). This value is passed to `ECX` which subsequently drives a `REP movs` instruction to *selectively* flood the ROB with up to $2n$ μops.

5 Attack Demo and Experimental Results

We implemented our FSI attack on actual hardware. While DoM defenses and InvarSpec are not implemented, we can see the effects of the attack in an unprotected core, which behaves the same as a protected core with respect to speculative-invariant instructions. We evaluated our results on an Intel® Core™ i7-2600K, which is a Sandy Bridge microarchitecture, running at up to 3.40GHz. The processor has 4 cores (2 SMT threads per core, for 8 threads in total) and 3 cache levels. Each core has a 32KiB L1 Cache and a 256KiB L2 Cache, and all cores share an 8MiB LLC.

The overall structure of the attack demo is illustrated in Fig. 3. We report on the results for the timing-load variant on a real system. Before we follow the attack path, all load addresses are flushed from the cache. The branch predictor is trained so that it



(a) ROB Attack using REP instruction: All 1000 attempts per secret

(b) ROB Attack using REP instruction: Average every 100 attempts

Fig. 4. Speculation Invariant access latency, in cycles, leaking the secret.

will always mispredict and follow the attack path. The secret value is already cached in the L1. Depending on the secret, ROB contention is added, so that `speculation invariant instruction (line 16)` will be delayed. If `secret==1`, delay from ROB contention will be sufficient for speculation to be verified before `speculation invariant instructions (line 16)` executes. If `secret==0`, no delay is applied and `speculation invariant instruction (line 16)` is executed as soon as possible.

Fig. 4a shows the results across all 1000 attempts. Fig. 4b illustrates the average cycles every 100 repetitions. We show, that when repeating the attack, the results diverge, making it easier to identify the secret: An average load when `secret==0` is 170 cycles. On the other hand, when `secret==1` an average load is 260 cycles.

Our results show that, clearly, *forward speculative interference* and ROB-contention work successfully in actual processors, and constitute a new side-channel that can be used to construct Spectre-type attacks. Because the speculation-invariant instructions behave the same as instructions from the re-convergence path in unprotected processors, FSI ROB-contention poses a significant threat when we want to lift defenses for speculation-invariant instructions.

6 Conclusion

In this work, we present a new side-channel, based on ROB contention, and a new speculative execution attack (ROB-contention attack) using this side-channel. The attack is achieved through *FSI*, i.e., speculative instructions interfering with *younger* instructions that are bound to commit regardless of the speculation outcome. For this reason, techniques, such as the InvarSpec framework, that lift the defenses for such bound-to-commit instructions, are susceptible to the same attack and can leak speculatively accessed information.

References

[1] F. Agner, “Instruction tables,” May 2021, [Online]. Available: https://www.agner.org/optimize/instruction_tables.pdf

[2] F. Agner, “The microarchitecture of intel, amd, and via cpus: An optimization guide for assembly programmers and compiler makers,” May 2021, [Online]. Available: https://www.agner.org/optimize/microarchitecture.pdf

[3] S. Ainsworth and T. M. Jones, “MuonTrap: Preventing cross-domain spectre-like attacks by capturing speculative state,” in *Proc. Int. Symp. Comput. Archit.*, 2020, pp. 132–144.

[4] K. Barber, A. Bacha, L. Zhou, Y. Zhang, and R. Teodorescu, “SpecShield: Shielding speculative data from microarchitectural covert channels,” in *Proc. Int. Conf. Parallel Architectural Compilation Techn.*, 2019, pp. 151–164.

[5] M. Behnia *et al.*, “Speculative interference attacks: Breaking invisible speculation schemes,” in *Proc. 26th ACM Int. Conf. Architectural Support Program. Lang. Oper. Syst.*, 2021, pp. 1046–1060.

[6] K. N. Khasawneh, E. M. Koruyeh, C. Song, D. Evtyushkin, D. Ponomarev, and N. Abu-Ghazaleh , “SafeSpec: Banishing the spectre of a meltdown with leakage-free speculation,” in *Proc. ACM/IEEE Des. Automat. Conf.*, 2019, pp. 1–6.

[7] P. Kocher *et al.*, “Spectre attacks: Exploiting speculative execution,” in *Proc. IEEE Symp. Secur. Privacy*, 2019, pp. 19–37.

[8] G. Saileshwar and M. K. Qureshi, “CleanupSpec: An ”undo” approach to safe speculation,” in *Proc. ACM/IEEE Int. Symp. Microarchit.*, 2019, pp. 73–86. [Online]. Available: http://doi.acm.org/10.1145/3352460.3358314

[9] C. Sakalis, S. Kaxiras, A. Ros, A. Jimborean, and M. Själander, “Understanding selective delay as a method for efficient secure speculative execution,” *IEEE Trans. Comput.*, vol. 69, no. 11, pp. 1584–1595, Nov. 2020.

[10] C. Sakalis, M. Alipour, A. Ros, A. Jimborean, S. Kaxiras, and S. Magnus, “Ghost loads: What is the cost of invisible speculation?,” in *Proc. ACM Int. Conf. Comput. Front.*, 2019, pp. 153–163.

[11] C. Sakalis, S. Kaxiras, A. Ros, A. Jimborean, and M. Själander, “Efficient invisible speculative execution through selective delay and value prediction,” in *Proc. Int. Symp. Comput. Archit.*, 2019, pp. 723–735.

[12] M. Taram, A. Venkat, and D. Tullsen, “ Context-sensitive fencing: Securing speculative execution via microcode customization,” in *Proc. Architectural Support Program. Lang. Oper. Syst.*, 2019, pp. 395–410.

[13] M. Yan, J. Choi, D. Skarlatos, A. Morrison, C. W. Fletcher, and J. Torrellas, “InvisiSpec: Making speculative execution invisible in the cache hierarchy,” in *Proc. ACM/IEEE Int. Symp. Microarchit.*, 2018, pp. 428–441.

[14] Y. Yarom and K. Falkner, “FLUSH+ RELOAD: A high resolution, low noise, l3 cache side-channel attack,” in *Proc. USENIX Secur. Symp.*, 2014, pp. 719–732. [Online]. Available: https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/yarom

[15] J. Yu, M. Yan, A. Khyzha, A. Morrison, J. Torrellas, and C. W. Fletcher, “Speculative taint tracking (STT): A comprehensive protection for speculatively accessed data,” in *Proc. ACM/IEEE Int. Symp. Microarchit.*, 2019, pp. 954–968. [Online]. Available: http://doi.acm.org/10.1145/3352460.3358274

[16] Z. N. Zhao *et al.*, “Speculation invariance (invarSpec): Faster safe execution through program analysis,” in *Proc. 53rd Annu. IEEE/ACM Int. Symp. Microarchit.*, 2020, pp. 1138–1152.