

A Software Approach to Defeating Side Channels in Last-Level Caches

Ziqiao Zhou
University of North Carolina
Chapel Hill, NC, USA
ziqiao@cs.unc.edu

Michael K. Reiter
University of North Carolina
Chapel Hill, NC, USA
reiter@cs.unc.edu

Yinqian Zhang
The Ohio State University
Columbus, OH, USA
yinqian@cse.ohio-state.edu

ABSTRACT

We present a software approach to mitigate access-driven side-channel attacks that leverage last-level caches (LLCs) shared across cores to leak information between security domains (e.g., tenants in a cloud). Our approach dynamically manages physical memory pages shared between security domains to disable sharing of LLC lines, thus preventing “FLUSH-RELOAD” side channels via LLCs. It also manages cacheability of memory pages to thwart cross-tenant “PRIME-PROBE” attacks in LLCs. We have implemented our approach as a memory management subsystem called CACHEBAR within the Linux kernel to intervene on such side channels across container boundaries, as containers are a common method for enforcing tenant isolation in Platform-as-a-Service (PaaS) clouds. Through formal verification, principled analysis, and empirical evaluation, we show that CACHEBAR achieves strong security with small performance overheads for PaaS workloads.

Keywords

Cache-based side channel; prime-probe; flush-reload

1. INTRODUCTION

An access-driven side channel is an attack by which an attacker computation learns secret information about a victim computation running on the same computer, not by violating the logical access control implemented by the isolation software (typically an operating system (OS) or virtual machine monitor (VMM)) but rather by observing the effects of the victim’s execution on microarchitectural components it shares with the attacker. Overwhelmingly, the components most often used in these attacks are CPU caches. Early cache-based side channels capable of leaking fine-grained information (e.g., cryptographic keys) across security boundaries used per-core caches (e.g., [28, 10, 34]), though the need for the attacker to frequently preempt the victim to observe its effects on per-core caches renders these attacks

relatively easy to mitigate in software (e.g., [37, 29]). Of more concern are side channels via last-level caches (LLCs) that are shared across cores and, in particular, do not require preemption of the victim to extract fine-grained information from it (e.g., [33, 35, 12, 20]).

Two varieties of LLC-based side channels capable of extracting fine-grained information from a victim have been demonstrated. The first such attacks were of the FLUSH-RELOAD variety [33, 35], which requires the attacker to share a physical memory page with the victim—a common situation in a modern OS, due to shared libraries, copy-on-write memory management, and memory deduplication mechanisms that aim for smaller memory footprints. The attacker first FLUSHes a cache-line sized chunk of the shared page out of the cache using processor-specific instructions (e.g., `clflush` in x86 processors) and later measures the time to RELOAD (or re-FLUSH [9]) it to infer whether this chunk was touched (and thus loaded to the shared cache already) by the victim. More recently, so-called PRIME-PROBE attacks have been demonstrated via LLCs [12, 20]; these do not require page sharing between the attacker and victim. Rather, PRIME-PROBE attacks can be conducted when the two programs share the same CPU cache sets. The attacker PRIMES the cache by loading its own memory into certain cache sets. Later it PROBES the cache by measuring the time to load the same memory into the cache sets and inferring how many cache lines in each cache set are absent due to conflicts with the victim’s execution.

In this paper we propose a software-only defense against these LLC-based side-channel attacks, based on two seemingly straightforward principles. First, to defeat FLUSH-RELOAD attacks, we propose a copy-on-access mechanism to manage physical pages shared across mutually distrusting *security domains* (i.e., processes, containers¹, or VMs). Specifically, temporally proximate accesses to the same physical page by multiple security domains results in the page being copied so that each domain has its own copy. In this way, a victim’s access to its copy will be invisible to an attacker’s RELOAD in a FLUSH-RELOAD attack. When accesses are sufficiently spaced in time, the copies can be deduplicated to return the overall memory footprint to its original size. Second, to defeat PRIME-PROBE attacks, we design a mechanism to manage the cacheability of memory pages so as to limit the number of lines per cache set that an attacker may PROBE. In doing so, we limit the visibility of the attacker into the victim’s demand for memory that maps to that cache set. Of course, the challenge in

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CCS’16 October 24–28, 2016, Vienna, Austria

© 2016 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-4139-4/16/10.

DOI: <http://dx.doi.org/10.1145/2976749.2978324>

¹<https://linuxcontainers.org/>

these defenses is in engineering them to be effective in both mitigating LLC-based side-channels and supporting efficient execution of computations.

To demonstrate these defenses and the tradeoffs between security and efficiency that they offer, we detail their design and implementation in a memory management subsystem called CACHEBAR (short for “Cache Barrier”) for the Linux kernel. CACHEBAR supports these defenses for security domains represented as Linux containers. That is, copy-on-access to defend against FLUSH-RELOAD attacks makes page copies as needed to isolate temporally proximate accesses to the same page from different containers. Moreover, memory cacheability is managed so that the processes in each container are collectively limited in the number of lines per cache set they can PROBE. CACHEBAR would thus be well-suited for use in Platform-as-a-Service (PaaS) clouds that isolate cloud customers in distinct containers; indeed, cross-container LLC-based side channels have been demonstrated in such clouds in the wild [35]. Our security evaluations show that CACHEBAR mitigates cache-based side-channel attacks, and our performance evaluation indicates that CACHEBAR imposes very modest overheads on PaaS workloads.

To summarize, we contribute:

- A novel copy-on-access mechanism to manage physical memory pages shared by distrusting tenants to prevent FLUSH-RELOAD side-channel attacks, and its formal verification using model checking.
- A novel mechanism to dynamically maintain queues of cacheable memory pages so as to limit the cache lines a malicious tenant may access in PRIME-PROBE attacks, and a principled derivation of its parameters to balance security and performance.
- Implementation of both mechanisms in a mainstream Linux operating system kernel and an extensive security and performance evaluation for PaaS workloads.

2. RELATED WORK

Numerous proposals have sought to mitigate cache-based side channels with low overhead through redesign of the cache hardware, e.g., [31, 13, 19]. Unfortunately, there is little evidence that mainstream CPU manufacturers will deploy such defenses in the foreseeable future, and even if they did, it would be years before these defenses permeated the installed computing base. Other proposals modify applications to better protect secrets from side-channel attacks. These solutions range from tools to limit branching on sensitive data (e.g., [4, 5]) to application-specific side-channel-free implementations (e.g., [15]). However, the overheads of these techniques tend to grow with the scope of programs to which they apply and can be very substantial (e.g., [25]).

It is for this reason that we believe that systems-level (i.e., OS- or VMM-level) defenses are the most plausible for deployment in the foreseeable future, and many have been proposed. With attention to cache-based side-channels specifically, several works provide to each security domain a limited number of designated pages that are never evicted from the LLC (e.g., [14, 18]), thereby rendering their contents immune to PRIME-PROBE and FLUSH-RELOAD attacks. These approaches, however, require the application developer to determine what data/instructions to protect and then to modify the application to organize the sensitive content into the protected pages; in contrast, CACHEBAR seeks to protect applications holistically and requires no application modifi-

cations. CACHEBAR also differs in several design choices that free it from limitations of prior approaches (e.g., the limitation of only one protected page per core [14] or dependence on relatively recent, Intel-specific cache optimizations [18]). Other systems-level solutions manage memory so as to partition the use of the LLC by different security domains (e.g., [24, 26]), though these approaches preclude memory-page and CPU-cache sharing entirely and hence can underutilize these resources considerably. Others have suggested disabling or selectively enabling memory sharing [22, 35, 3] for countering various side-channel attacks exploiting shared memory, while stopping short of exploring a complete design for doing so. Our copy-on-access design provides an efficient realization of this idea for addressing FLUSH-RELOAD attacks, and extends this idea with cacheability management for PRIME-PROBE defense, as well.

LLC-based side channels are a particular instance of timing side channels, and so defenses that seek to eliminate timing side channels are also relevant to our problem. Examples include fuzzing real-time sources (e.g., [30]), though this impinges on legitimate uses of real time. Since real-time counters are not the only way to time memory fetches [32], other efforts have sought to eliminate side-channel risks more holistically via altering the CPU scheduler (e.g., [27, 17]) and managing how tenants co-locate (e.g., [16, 36, 2, 17]). In contrast, here we focus specifically on LLC-based side channels (vs. a larger subset of timing side-channels), which again are arguably the most potent known side-channel vectors [33, 35, 12, 20], and restrict our modifications to the memory management subsystem.

3. COPY-ON-ACCESS

The FLUSH-RELOAD attack is a highly effective LLC-based side channel that was used, e.g., by Zhang et al. [35] to mount fine-grained side-channel attacks in commercial PaaS clouds. It leverages physical memory pages shared between an attacker and victim security domains, as well as the ability to evict those pages from LLCs, using a capability such as provided by the `clflush` instruction on the x86 architecture. `clflush` is designed to maintain consistency between caches and memory for write-combined memory [11]. The attacker uses `clflush`, providing a virtual address as an argument, to invalidate the cache lines occupied by the backing physical memory. After a short time interval (the “FLUSH-RELOAD interval”) during which the victim executes, the attacker measures the time to access the same virtual address. Based on this duration, the attacker can infer whether the victim accessed that memory during the interval.

3.1 Design

Modern operating systems, in particular Linux OS, often adopt on-demand paging and copy-on-write mechanisms [7] to reduce the memory footprints of userspace applications. In particular, copy-on-write enables multiple processes to share the same set of physical memory pages as long as none of them modify the content. If a process writes to a shared memory page, the write will trigger a page fault and a subsequent new page allocation so that a private copy of page will be provided to this process. In addition, memory merging techniques like Kernel Same-Page Merging (KSM) [1] are also used in Linux OS to deduplicate identical memory pages. Memory sharing, however, is one of the key factors that enable FLUSH-RELOAD side channel attacks. Disabling

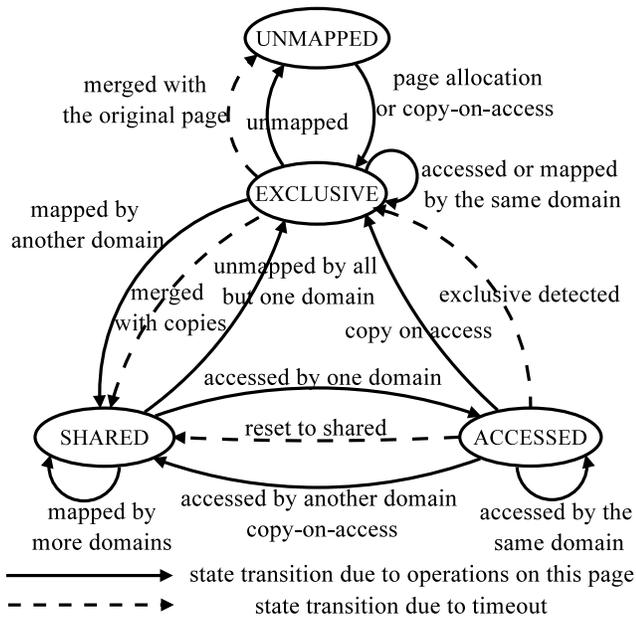


Figure 1: State transition of a physical page

memory page sharing entirely will eliminate FLUSH-RELOAD side channels but at the cost of much larger memory footprints and thus inefficient use of physical memory.

CACHEBAR adopts a design that we call copy-on-access, which dynamically controls the sharing of physical memory pages between security domains. We designate each physical page as being in exactly one of the following states: UNMAPPED, EXCLUSIVE, SHARED, and ACCESSED. An UNMAPPED page is a physical page that is not currently in use. An EXCLUSIVE page is a physical page that is currently used by exactly one security domain, but may be shared by multiple processes in that domain. A SHARED page is a physical page that is shared by multiple security domains, i.e., mapped by at least one process of each of the sharing domains, but no process in any domain has accessed this physical page recently. In contrast, an ACCESSED page is a previously SHARED page that was recently accessed by a security domain. The state transitions are shown in Fig. 1.

An UNMAPPED page can transition to the EXCLUSIVE state either due to normal page mapping, or due to copy-on-access when a page is copied into it. Unmapping a physical page for any reason (e.g., process termination, page swapping) will move an EXCLUSIVE page back to the UNMAPPED state. However, mapping the current EXCLUSIVE page by another security domain will transit it into the SHARED state. If all but one domain unmaps this page, it will transition back from the SHARED state to the EXCLUSIVE state, or ACCESSED state to the EXCLUSIVE state. A page in the SHARED state may be shared by more domains and remain in the same state; when any one of the domains accesses the page, it will transition to the ACCESSED state. An ACCESSED page can stay that way as long only the same security domain accesses it. If this page is accessed by another domain, a new physical page will be allocated to make a copy of this one, and the current page will transition to either EXCLUSIVE or SHARED state, depending on the remaining number of domains mapping this page. The new page will be assigned state EXCLUSIVE.

An ACCESSED page will be reset to the SHARED state if it is not accessed for Δ_{accessed} seconds. This timeout mechanism ensures that only recently used pages will remain in the ACCESSED state, limiting chances for unnecessary duplication. Page merging may also be triggered by deduplication services in a modern OS (e.g., KSM in Linux). This effect is reflected by a dashed line in Fig. 1 from state EXCLUSIVE to SHARED. A page at any of the *mapped* states (i.e., EXCLUSIVE, SHARED, ACCESSED) can transition to UNMAPPED state for the same reason when it is a copy of another page (not shown in the figure).

Merging duplicated pages requires some extra bookkeeping. When a page transitions from UNMAPPED to EXCLUSIVE due to copy-on-access, the original page is tracked by the new copy so that CACHEBAR knows with which page to merge it when deduplicating. If the original page is unmapped first, then one of its copies will be designated as the new “original” page, with which other copies will be merged in the future. The interaction between copy-on-access and existing copy-on-write mechanisms is also implicitly depicted in Fig. 1: Upon copy-on-write, the triggering process will first *unmap* the physical page, possibly inducing a state transition (from SHARED to EXCLUSIVE). The state of the newly mapped physical page is maintained separately.

3.2 Implementation

At the core of copy-on-access implementation is the state machine depicted in Fig. 1.

unmapped \Leftrightarrow exclusive \Leftrightarrow shared. Conventional Linux kernels maintain the relationship between processes and the physical pages they use. However, CACHEBAR also needs to keep track of the relationship between containers and the physical pages that each container’s processes use. Therefore, CACHEBAR incorporates a new data structure, `counter`, which is conceptually a table used for recording, for each physical page, the number of processes in each container that have Page Table Entries (PTEs) mapped to this page.

The `counter` data structure is updated and referenced in multiple places in the kernel. Specifically, in CACHEBAR we instrumented every update of `_mapcount`, a data field in the page structure for counting PTE mappings, so that every time the kernel tracks the PTE mappings of a physical page, `counter` is updated accordingly. The use of `counter` greatly simplifies maintaining and determining the state of a physical page: (1) Given a container, access to a single cell suffices to check whether a physical page is already mapped in the container. This operation is very commonly used to decide if a state transition is required when a page is mapped by a process. Without `counter`, such an operation requires performing reverse mappings to check the domain of each mapping. (2) Given a physical page, it takes N accesses to `counter`, where N is the total number of containers, to tell which containers have mapped to this page. This operation is commonly used to determine the state of a physical page.

shared \Rightarrow accessed. To differentiate SHARED and ACCESSED states, one additional data field, `owner`, is added (see Fig. 2) to indicate the owner of the page (a pointer to a `PID_namespace` structure). When the page is in the SHARED state, its `owner` is NULL; otherwise it points to the container that last accessed it.

All PTEs pointing to a SHARED physical page will have a reserved Copy-On-Access (COA) bit set. Therefore, any

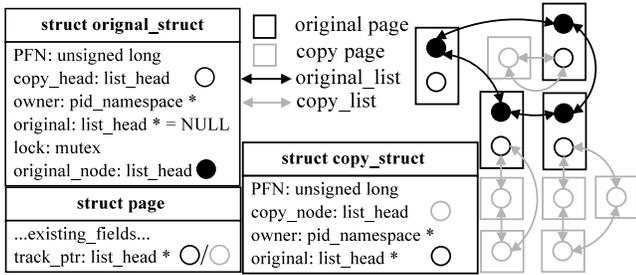


Figure 2: Structure of copy-on-access page lists.

access to these virtual pages will induce a page fault. When a page fault is triggered, CACHEBAR checks if the page is present in physical memory; if so, and if the physical page is in the SHARED state, the COA bit of the current PTE for this page will be cleared so that additional accesses to this physical page from the current process will be allowed without page faults. The physical page will also transition to the ACCESSED state.

accessed \Rightarrow exclusive/shared. If the page is already in the ACCESSED state when a domain other than the owner accesses it, the page fault handler will allocate a new physical page, copy the content of the original page into the new page, and change the PTEs in the accessing container so that they point to the new page. Since multiple same-content copies in one domain burdens both performance and memory but contributes nothing for security, the fault handler will reuse a copy belonging to that domain if it exists. After copy-on-access, the original page can either be EXCLUSIVE or SHARED. All copy pages are anonymous-mapped, since only a single file-mapped page for the same file section is allowed.

A transition from the ACCESSED state to SHARED or EXCLUSIVE state can also be triggered by a timeout mechanism. CACHEBAR implements a periodic timer (every $\Delta_{\text{accessed}} = 1\text{s}$). Upon timer expiration, all physical pages in the ACCESSED state that were not accessed during this Δ_{accessed} interval will be reset to the SHARED state by clearing its owner field, so that pages that are infrequently accessed are less likely to trigger copy-on-access. If an ACCESSED page is found for which its counter shows the number of domains mapped to it is 1, then the daemon instead clears the COA bit of all PTEs for that page and marks the page EXCLUSIVE.

Instead of keeping a list of ACCESSED pages, CACHEBAR maintains a list of pages that are in either SHARED or ACCESSED state, denoted `original_list` (shown in Fig. 2). Each node in the list also maintains a list of copies of the page it represents, dubbed `copy_list`. These lists are attached onto the `struct page` through `track_ptr`. Whenever a copy is made from the page upon copy-on-access, it is inserted into the `copy_list` of the original page. Whenever a physical page transitions to the UNMAPPED state, it is removed from whichever of `original_list` or `copy_list` it is contained in. In the former case, CACHEBAR will designate a copy page of the original page as the new original page and adjust the lists accordingly.

For security reasons that will be explained in Sec. 3.3, we further require flushing the entire memory page out of the cache after transitioning a page from the ACCESSED state to the SHARED state due to this timeout mechanism. This page-flushing procedure is implemented by issuing `clflush`

on each of the memory blocks of any virtual page that maps to this physical page.

State transition upon `clflush`. The `clflush` instruction is subject to the same permission checks as a memory load, will trigger the same page faults, and will similarly set the ACCESSED bit in the PTE of its argument [11]. As such, each FLUSH via `clflush` triggers the same transitions (e.g., from SHARED to ACCESSED, and from ACCESSED to an EXCLUSIVE copy) as a RELOAD in our implementation, meaning that this defense is equally effective against both FLUSH-RELOAD and FLUSH-FLUSH [9] attacks.

Page deduplication. To mitigate the impact of copy-on-access on the size of memory, CACHEBAR implements a less frequent timer (every $\Delta_{\text{copy}} = 10 \times \Delta_{\text{accessed}}$ seconds) to periodically merge the page copies with their original pages. Within the timer interrupt handler, `original_list` and each `copy_list` are traversed similarly to the “ACCESSED \Rightarrow SHARED” transition description above, though the ACCESSED bit in the PTEs of only pages that are in the EXCLUSIVE state are checked. If a copy page has not been accessed since the last such check (i.e., the ACCESSED bit is unset in all PTEs pointing to it), it will be merged with its original page (the head of the `copy_list`). The ACCESSED bit in the PTEs will be cleared afterwards.

When merging two pages, if the original page is anonymous-mapped, then the copy page can be merged by simply updating all PTEs pointing to the copy page to instead point to the original page, and then updating the original page’s reverse mappings to include these PTEs. If the original page is file-mapped, then merging is more intricate, additionally involving the creation of a new virtual memory area (vma structure) that maps to the original page’s file position and using this structure to replace the virtual memory area of the (anonymous) copy page in the relevant task structure.

For security reasons, merging of two pages requires flushing the original physical page from the LLC. We will elaborate on this point in Sec. 3.3.

Interacting with KSM. Page deduplication can also be triggered by existing memory deduplication mechanisms (e.g., KSM). To maintain the state of physical pages, CACHEBAR instruments every reference to `_mapcount` within KSM and updates `counter` accordingly. KSM is capable of merging more pages than our built-in page deduplication mechanisms. However, CACHEBAR still relies on the built-in page deduplication mechanisms for several reasons. First, KSM can merge only anonymous-mapped pages, while CACHEBAR needs to frequently merge an anonymous-mapped page (a copy) with a file-mapped page (the original). Second, KSM may not be enabled in certain settings, which will lead to ever growing `copy_lists`. Third, KSM must compare page contents byte-by-byte before merging two pages, whereas CACHEBAR deduplicates pages on the same `copy_list`, avoiding the expensive page content comparison.

3.3 Security

Copy-on-access is intuitively secure by design, as no two security domains may access the same physical page at the same time, rendering FLUSH-RELOAD attacks seemingly impossible. To show security formally, we subjected our design to model checking in order to prove that copy-on-access is secure against FLUSH-RELOAD attacks. Model checking is an approach to formally verify a specification of a finite-state

concurrent system expressed as temporal logic formulas, by traversing the finite-state machine defined by the model. In our study, we used the Spin model checker, which offers efficient ways to model concurrent systems and verify temporal logic specifications.

System modeling. We model a physical page in Fig. 1 using a byte variable in the PROMELA programming language, and two physical pages as an array of two such variables, named `pages`. We model two security domains (e.g., containers), an attacker domain and a victim domain, as two processes in PROMELA. Each process maps a virtual page, `virt`, to one of the physical pages. The virtual page is modeled as an index to the `pages[]` array; initially `virt` for both the attacker and the victim point to the first physical page (i.e., `virt` is 0). The victim process repeatedly sets `pages[virt]` to 1, simulating a memory access that brings `pages[virt]` into cache. The attacker process FLUSHes the virtual page by assigning 0 to `pages[virt]` and RELOADs it by assigning 1 to `pages[virt]` after testing if it already equals to 1. Both the FLUSH and RELOAD operations are modeled as atomic to simplify the state exploration.

We track the state and owner of the first physical page using another two variables, `state` and `owner`. The first page is initially in the SHARED state (`state` is SHARED), and state transitions in Fig. 1 are implemented by each process when they access the memory. For example, the RELOAD code snippet run by the attacker is shown in Fig. 3. If the attacker has access to the shared page (Line 3), versus an exclusive copy (Line 16), then it simulates an access to the page, which either moves the state of the page to ACCESSED (Line 10) if the state was SHARED (Line 9) or to EXCLUSIVE (Line 14) after making a copy (Line 13) if the state was already ACCESSED and not owned by the attacker (Line 12). Leakage is detected if `pages[virt]` is 1 prior to the attacker setting it as such (Line 19), which the attacker tests in Line 18.

```

1  atomic {
2  if
3  ::(virt == 0) ->
4    if
5    ::(state == UNMAPPED) ->
6      assert(0)
7    ::(state == EXCLUSIVE && owner != ATTACKER) ->
8      assert(0)
9    ::(state == SHARED) ->
10     state = ACCESSED
11     owner = ATTACKER
12    ::(state == ACCESSED && owner != ATTACKER) ->
13     virt = 1 /* copy-on-access */
14     state = EXCLUSIVE
15    fi
16  ::else -> skip
17  fi
18  assert(pages[virt] == 0)
19  pages[virt] = 1
20 }

```

Figure 3: Code snippet for Reload.

To model the dashed lines in Fig. 1, we implemented another process, called *timer*, in PROMELA that periodically transitions the physical page back to SHARED state from ACCESSED state, and periodically with a longer interval, merges the two pages by changing the value of `virt` of each domain back to 0, `owner` to none, and `state` to SHARED.

The security specification is stated as a non-interference property. Specifically, as the attacker domain always FLUSHes

the memory block (sets `pages[virt]` to 0) before RELOADing it (setting `pages[virt]` to 1), if the non-interference property holds, then the attacker should always find `pages[virt]` to be 0 upon RELOADing the page. The model checker checks for violation of this property.

Automated verification. We checked the model using Spin. Interestingly, our first model-checking attempt suggested that the state transitions may leak information to a FLUSH-RELOAD attacker. The leaks were caused by the *timer* process that periodically transitions the model to a SHARED state. After inspecting the design and implementation, we found that there were two situations that may cause information leaks. In the first case, when the timer transitions the state machine to the SHARED state from the ACCESSED state, if the prior owner of the page was the victim and the attacker reloaded the memory right after the transition, the attacker may learn one bit of information. In the second case, when the physical page was merged with its copy, if the owner of the page was the victim before the page became SHARED, the attacker may reload it and again learn one bit of information. Since in our implementation of CACHEBAR, these two state transitions are triggered if the page (or its copy) has not been accessed for a while (roughly Δ_{accessed} and Δ_{copy} seconds, respectively), the information leakage bandwidth due to each would be approximately $1/\Delta_{\text{accessed}}$ bits per page per second or $1/\Delta_{\text{copy}}$ bits per page per second, respectively.

We improved our CACHEBAR implementation to prevent this leakage by enforcing LLC flushes (as described in Sec. 3.2) upon these two periodic state transitions. We adapted our model accordingly to reflect such changes by adding one more instruction to assign `pages[0]` to be 0 right after the two *timer*-induced state transitions. Model checking this refined model revealed no further information leakage.

4. CACHEABILITY MANAGEMENT

Another common method to launch side-channel attacks via caches is using PRIME-PROBE attacks, introduced by Osvik et al. [21]. These attacks have recently been adapted to use LLCs to great effect, e.g., [20, 12]. Unlike a FLUSH-RELOAD attack, PRIME-PROBE attacks do not require the attacker and victim security domains to share pages. Rather, the attacker simply needs to access memory so as to evict (PRIME) the contents of a cache set and later access (PROBE) this memory again to determine (by timing the accesses) how much the victim evicted from the cache set. A potentially effective countermeasure to these attacks, accordingly, is to remove the attacker’s ability to PRIME and PROBE the whole cache set and to predict how a victim’s demand for that set will be reflected in the number of evictions from that set.

4.1 Design

Suppose a w -way set associative LLC, so that each cache set has w lines. Let x be the number of cache lines in one set that the attacker observes having been evicted in a PRIME-PROBE interval. The PRIME-PROBE attack is effective today because x is typically a good indicator of the demand d that the victim security domain had for memory that mapped to that cache set during the PRIME-PROBE interval. In particular, if the attacker PRIMES and PROBES all w lines, then it can often observe the victim’s demand d exactly, unless $d > w$ (in which case the attacker learns at least $d \geq w$).

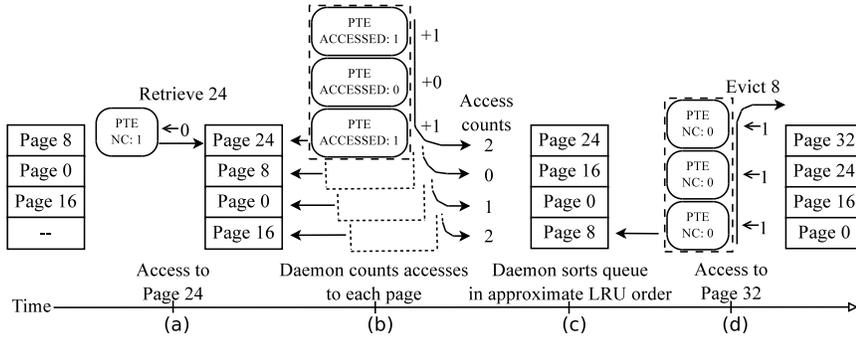


Figure 4: A cacheable queue for one page color in a domain: (a) access to page 24 brings it into the queue and clears NC bit (“←0”) in the PTE triggering the fault; periodically, (b) a daemon counts the ACCESSED bits (“+0”, “+1”) per page and (c) re-orders pages accordingly; to make room for a new page, (d) NC bits in PTEs pointing to the least recently used page are set, and the page is removed from the queue.

Here we propose to periodically and probabilistically re-configure the budget k_i of lines per cache set that the security domain i can occupy. After such a reconfiguration, the attacker’s view of the victim’s demand d is clouded by the following three effects. First, if the attacker is allotted a budget $k_a < w$, then the attacker will be unable to observe any evictions at all (i.e., $x = 0$) if $d < w - k_a$.² Second, if the victim is given allotment k_v , then any two victim demands d, d' satisfying $d > d' \geq k_v$ will be indistinguishable to the attacker. Third, the probabilistic assignment of k_v results in extra ambiguity for the attacker, since x evictions might reflect the demand d or the budget k_v , since $x \leq \min\{d, k_v\}$ (if all x evictions are caused by the victim).

To enforce the budget k_i of lines that security domain i can use in a given cache set, CACHEBAR maintains for each cache set a queue per security domain that records which memory blocks are presently cacheable in this set by processes in this domain. Each element in the queue indicates a memory block that maps to this cache set; only blocks listed in the queue can be cached in that set. The queue is maintained with a least recently used (LRU) replacement algorithm. That is, whenever a new memory block is accessed, it will replace the memory block in the corresponding queue that is the least recently used.

4.2 Implementation

Implementation of cacheable queues is processor micro-architecture dependent. Here we focus our attention on Intel x86 processors, which appears to be more vulnerable to PRIME-PROBE attacks due to their inclusive last-level cache [20]. As x86 architectures only support memory management at the page granularity (e.g., by manipulating the PTEs to cause page faults), CACHEBAR controls the cacheability of memory blocks at page granularity. CACHEBAR uses reserved bits in each PTE to manage the cacheability of, and to track accesses to, the physical page to which it points, since a reserved bit set in a PTE induces a page fault upon access to the associated virtual page, for which the backing physical page cannot be retrieved or cached (if it is not already) before the bit is cleared [11, 23]. We hence use the term *domain-cacheable* to refer to a physical page that is “cacheable” in the view of all processes in a particular security domain, which is implemented by modifying all relevant PTEs (to have no reserved bits set) in the

processes of that security domain. By definition, a physical page that is domain-cacheable to one container may not necessarily be domain-cacheable to another.

To ensure that no more than k_i memory blocks from all processes in container i can occupy lines in a given cache set, CACHEBAR ensures that no more than k_i of those processes’ physical memory pages, of which contents can be stored in that cache set, are domain-cacheable at any point in time. Physical memory pages of which contents can be stored in the same cache set are said to be of the same *color*, and so to implement this property, CACHEBAR maintains, per container and per color (rather than per cache set), one cacheable queue, each element of which is a physical memory page that is domain-cacheable in this container. Since the memory blocks in each physical page map to different cache sets, limiting the domain-cacheable pages of a color to k_i also limits the number of cache lines that blocks from these pages can occupy in the same cache set to k_i .

To implement a non-domain-cacheable memory, CACHEBAR uses one reserved bit, which we denote by NC, in all PTEs within the domain mapped to that physical page. As such, accesses to any of these virtual pages will be trapped into the kernel and handled by the page fault handler. Upon detecting page faults of this type, the page fault handler will move the accessed physical page into the corresponding cacheable queue, clear the NC bit in the current PTE³, and remove a least recently used physical page from the cacheable queue and set the NC bits in this domain’s PTEs mapped to that page. A physical page removed from the cacheable queue will be flushed out of the cache using `clflush` instructions on all of its memory blocks to ensure that no residue remains in the cache. CACHEBAR will flush the translation lookaside buffers (TLB) of all processors to ensure the correctness of page cacheabilities every time PTEs are altered. In this way, CACHEBAR limits the number of domain-cacheable pages of a single color at any time to k_i .

To maintain the LRU property of the cacheable queue, a daemon periodically re-sorts the queue in descending order of recent access count. Specifically, the daemon traverses the domain’s PTEs mapped to the physical frame within that domain’s queue and counts the number having their ACCESSED bit set, after which it clears these ACCESSED bits. It then orders the physical pages in the cacheable queue by this count (see Fig. 4). In our present implementation, this daemon is the same daemon that resets pages from the

²This statement assumes a LRU replacement policy and that the victim is the only security domain that runs in the PRIME-PROBE interval. If it was not the only security domain to run, then the ambiguity of the observable evictions will additionally cause difficulties for the attacker.

³We avoid the overhead of traversing all PTEs in the container that map to this physical page. Access to those virtual pages will trigger page faults to make these updates without altering the cacheable queue.

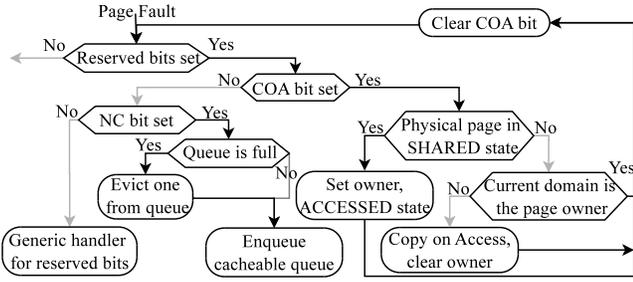


Figure 5: Page fault handler for CacheBar.

ACCESSED state to SHARED state (see Sec. 3), which already checks and resets the ACCESSED bits in copies’ PTEs. Again, this daemon runs every $\Delta_{\text{accessed}} = 1\text{s}$ seconds in our implementation. This daemon also performs the task of resetting k_i for each security domain i , each time it runs.

Interacting with copy-on-access. The cacheable queues work closely with the copy-on-access mechanisms. In particular, as both the COA and NC bits may trigger a page fault upon page accesses, the page handler logic must incorporate both (see Fig. 5). First, a page fault is handled as normal unless it is due to one of the reserved bits set in the PTE. As CACHEBAR is the only source of reserved bits, it takes over page fault handling from this point. CACHEBAR first checks the COA bit in the PTE. If it is set, the corresponding physical page is either SHARED, in which case it will be transitioned to ACCESSED, or ACCESSED, in which case it will be copied and transitioned to either SHARED or EXCLUSIVE. CACHEBAR then clears the COA bit and, if no other reserved bits are set, the fault handler returns. Otherwise, if the NC bit is set, the associated physical page is not in the cacheable queue for its domain, and so CACHEBAR enqueues the page and, if the queue is full, removes the least-recently-used page from the queue. If the NC bit is clear, this page fault is caused by unknown reasons and CACHEBAR turns control over to the generic handler for reserved bits.

4.3 Security

Recall that k_i is the number of cache lines in a certain cache set that is available to domain i for a period. While the budget k_i is in effect, each access to a memory block that maps to this cache set, beyond the in-queue k_i memory blocks, will incur a page fault (because they are all in different pages). Because the page-fault processing time will overwhelm the timing granularity of modern PRIME-PROBE attacks by an order of magnitude, the attacker i realistically needs to restrict himself to accessing k_i pages in his PROBE phase and hence to occupying k_i lines in that cache set.

The security of this design hinges critically on how each k_i is set by the daemon. When k_i is reset, it is drawn from a distribution. In the remainder of this section we present how this distribution is determined.

Suppose there are (at most) m domains on a host that are owned by the attacker—which might be all domains on the host except the victim—and let w be the number of cache lines per LLC set. Below we consider domain 0 to be the “victim” domain being subjected to PRIME-PROBE attacks by the “attacker” domains $1, \dots, m$. Of course, the attacker domains make use of all $\sum_{i=1}^m k_i$ cache lines available to them for conducting their PRIME-PROBE attacks.

Periodically, CACHEBAR draws a new value k_i for each

security domain i . This drawing is memoryless and independent of the draws for other security domains. Let K_i denote the random variable distributed according to how k_i is determined. The random variables that we presume can be observed by the attacker domains include K_1, \dots, K_m ; let $K_a = \min\{w, \sum_{i=1}^m K_i\}$ denote the number of cache lines allocated to the attacker domains. We also presume the attacker can accurately measure the number X of its cache lines that are evicted during the victim’s execution.

Let $\mathbb{P}_d(E)$ denote the probability of event E in an execution period during which the victim’s cache usage would populate d lines (of this color) if it were allowed to use all w lines, i.e., if $k_0 = w$. We (the defender) would like to distribute K_0, \dots, K_m so as to minimize the statistical distance between eviction distributions observable by the attacker for different victim demands d, d' , i.e., to minimize

$$\sum_{0 \leq d < d' \leq w} \sum_x |\mathbb{P}_d(X = x) - \mathbb{P}_{d'}(X = x)| \quad (1)$$

We begin by deriving an expression for $\mathbb{P}_d(X = x)$. Below we make the conservative assumption that all evictions are caused by the victim’s behavior; in reality, caches are far noisier. We first consider the case $x = 0$, i.e., that the attacker domains observe no evictions.

$$\mathbb{P}_d(X = 0 \mid K_0 = k_0 \wedge K_a = k_a) = \begin{cases} 1 & \text{if } w \geq k_a + \min\{k_0, d\} \\ 0 & \text{otherwise} \end{cases}$$

“ $\min\{k_0, d\}$ ” is used above because any victim demand for memory blocks that map to this cache set beyond k_0 will back-fill the cache lines invalidated when CACHEBAR flushes other blocks from the victim’s cacheability queue, rather than evicting others. Since K_0 and K_a are independent,

$$\begin{aligned} \mathbb{P}_d(X = 0) &= \sum_{k_0=0}^d \sum_{k_a=0}^{w-k_0} \mathbb{P}(K_0 = k_0) \cdot \mathbb{P}(K_a = k_a) \\ &+ \sum_{k_0=d+1}^w \sum_{k_a=0}^{w-d} \mathbb{P}(K_0 = k_0) \cdot \mathbb{P}(K_a = k_a) \quad (2) \end{aligned}$$

Note that we have dropped the “ d ” subscripts from the probabilities on the right, since K_0 and K_a are distributed independently of d . And, since K_1, \dots, K_m are independent,

$$\mathbb{P}(K_a = k_a) = \begin{cases} \sum_{k_1+\dots+k_m=k_a} \prod_{i=1}^m \mathbb{P}(K_i = k_i) & \text{if } k_a < w \\ \sum_{k_1+\dots+k_m \geq w} \prod_{i=1}^m \mathbb{P}(K_i = k_i) & \text{if } k_a = w \end{cases} \quad (3)$$

Similarly, for $x \geq 1$,

$$\mathbb{P}_d(X = x \mid K_0 = k_0 \wedge K_a = k_a) = \begin{cases} 1 & \text{if } x+w = k_a + \min\{k_0, d\} \\ 0 & \text{otherwise} \end{cases}$$

and so for $x \geq 1$,

$$\begin{aligned} \mathbb{P}_d(X = x) &= \sum_{k_0=0}^d \mathbb{P}(K_0 = k_0) \cdot \mathbb{P}(K_a = x+w-k_0) \\ &+ \sum_{k_0=d+1}^w \mathbb{P}(K_0 = k_0) \cdot \mathbb{P}(K_a = x+w-d) \quad (4) \end{aligned}$$

From here, we proceed to solve for the best distribution for K_0, \dots, K_m to minimize Eqn. 1 subject to constraints

Eqns. 2–4. That is, we specify those constraints, along with

$$\forall i, i', k : \mathbb{P}(K_i = k) = \mathbb{P}(K_{i'} = k) \quad (5)$$

$$\forall i : \sum_{k_i=0}^w \mathbb{P}(K_i = k_i) = 1 \quad (6)$$

$$\forall i, k_i : \mathbb{P}(K_i = k_i) \geq 0 \quad (7)$$

and then solve for each $\mathbb{P}(K_i = k_i)$ to minimize Eqn. 1.

Unfortunately, solving to minimize Eqn. 1 alone simply results in a distribution that results in no use of the cache at all (e.g., $\mathbb{P}(K_i = 0) = 1$ for each i). As such, we need to rule out such degenerate and “unfair” cases:

$$\forall i : \mathbb{P}(K_i < w/(m+1)) = 0 \quad (8)$$

Also, to encourage cache usage, we counterbalance Eqn. 1 with a second goal that values greater use of the cache. We express this goal as minimizing the earth mover’s distance [6] from the distribution that assigns $\mathbb{P}(K_i = w) = 1$, i.e.,

$$\sum_{k=0}^w (w-k) \cdot \mathbb{P}(K_0 = k) \quad (9)$$

As such, our final optimization problem seeks to balance Eqn. 1 and Eqn. 9. Let constant γ denote the maximum (i.e., worst) possible value of Eqn. 1 (i.e., when $\mathbb{P}(K_i = w) = 1$ for each i) and δ denote the maximum (i.e., worst) possible value of Eqn. 9 (i.e., when $\mathbb{P}(K_i = 0) = 1$ for each i). Then, given a parameter ϵ , $0 < \epsilon < 1$, our optimization computes distributions for K_0, \dots, K_m so as to minimize u subject to

$$u = \frac{1}{\gamma} \left(\sum_{0 \leq d < d' \leq w} \sum_x |\mathbb{P}_d(X=x) - \mathbb{P}_{d'}(X=x)| \right)$$

$$u \geq \frac{1}{\delta(1+\epsilon)} \left(\sum_{k=0}^w (w-k) \cdot \mathbb{P}(K_0 = k) \right)$$

and constraints Eqns. 2–8.

Our evaluation in Sec. 5.2.2 and Sec. 5.3.1 empirically characterizes the security and performance that result from setting $\epsilon = 0.01$ the default setting in CACHEBAR. Of course, other balances could be chosen between these concerns, though as we will see below, this setting achieves convincing security while inducing only a modest performance overhead for most PaaS workloads.

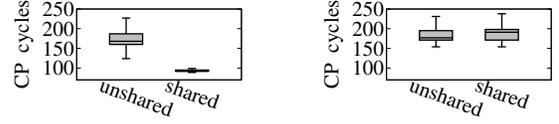
5. EVALUATION

In this section, we evaluate the security and performance of CACHEBAR to validate its design and implementation.

5.1 Setup

Our testbed is a rack mounted DELL server equipped with two 2.67GHz Intel Xeon 5550 processors. Each processor contains 4 physical cores (hyperthreading disabled) sharing an 8MB last-level cache (L3). Each core has a 32KB L1 data and instruction cache and a 256KB L2 unified cache. The rack server is equipped with 128GB DRAM and 1000Mbps NIC connected to a 1000Mbps ethernet.

We implemented CACHEBAR as a kernel extension for Linux kernel 3.13.11.6 that runs Ubuntu 14.04 server edition. Our implementation adds ~ 7000 lines of code to this Linux kernel. We set up containers using Docker 1.7.1.



(a) CACHEBAR disabled (b) CACHEBAR enabled

Figure 6: Reload timings in Flush-Reload attacks on a shared address vs. on an unshared address

5.2 Security Evaluation

We evaluated the effectiveness of CACHEBAR in defending against both FLUSH-RELOAD and PRIME-PROBE attacks.

5.2.1 Flush-Reload Attacks

Although we used `spin` model checker to validate the security of our copy-on-access design (Sec. 3), we empirically tested our implementation to validate its effectiveness. To do so, we constructed a FLUSH-RELOAD covert channel between sender and receiver processes, which were isolated in different containers. Both the sender and receiver were linked to a shared library, `libcrypto.so.1.0.0`, and were pinned to run on different cores of the same socket, thus sharing the same last-level cache. The sender ran in a loop, repeatedly accessing one memory location (the beginning address of function `AES_decrypt()`). The receiver executed FLUSH-RELOAD attacks on the same memory address, by first FLUSHING the memory block out of the shared LLC with an `clflush` instruction and then RELOADING the block by accessing it directly while measuring the access latency. The interval between FLUSH and RELOAD was set to 2500 cycles. The experiment was run for 500,000 FLUSH-RELOAD trials. We then repeated this experiment with the sender accessing an unshared address, to form a baseline.

Fig. 6(a) shows the results of this experiment, when run over unmodified Linux. The three horizontal lines forming the “box” in each boxplot represents the first, second (median), and third quartiles of the FLUSH-RELOAD measurements; whiskers extend to cover all points that lie within $1.5 \times$ the interquartile range. As can be seen in this figure, the times observed by the receiver to RELOAD the shared address were clearly separable from the times to RELOAD the unshared address, over unmodified Linux. With CACHEBAR enabled, however, these measurements are no longer separable (Fig. 6(b)). Certain corner cases are not represented in Fig. 6. For example, we found it extremely difficult to conduct experiments to capture the corner cases where FLUSH and RELOAD takes place right before and after physical page mergers, as described in Sec. 3.3. As such, we rely on our manual inspection of the implementation in these cases to check correctness and argue these corner cases are very difficult to exploit in practice.

5.2.2 Prime-Probe Attacks

We evaluated the effectiveness of CACHEBAR against PRIME-PROBE attacks by measuring its ability to interfere with a simulated attack. Because the machine architecture on which we performed these tests had a w -way LLC with $w = 16$, we limited our experiments to only a single attacker container (i.e., $m = 1$), but an architecture with a larger w could accommodate more.⁴

⁴For example, on an Itanium 2 processor with a 64-way LLC,

In our simulation, a process in the attacker container repeatedly performed PRIME-PROBE attacks on a specific cache set, while a process in a victim container accessed data that were retrieved into the same cache set at the rate of d accesses per attacker PRIME-PROBE interval. The cache lines available to the victim container and attacker container, i.e., k_v and k_a respectively, were fixed in each experiment. The calculations in Sec. 4.3 implied that k_v and k_a could take on values from $\{4, 5, 6, \dots, 14\}$. In each test with fixed k_v and k_a , we allowed the victim to place a demand of (i.e., retrieve memory blocks to fill) $d \in \{0, 1, 2, \dots, 16\}$ cache lines of the cache set undergoing the PRIME-PROBE attack by the attacker. The attacker’s goal was to classify the victim’s demand into one of six classes: NONE = $\{0\}$, ONE = $\{1\}$, FEW = $\{2, 3, 4\}$, SOME = $\{5, 6, 7, 8\}$, LOTS = $\{9, 10, 11, 12\}$, and MOST = $\{13, 14, 15, 16\}$.

To make the attack easier, we permitted the attacker to know k_a ; i.e., the attacker trained a different classifier per value of k_a , with knowledge of the demand d per PRIME-PROBE trial, and then tested against additional trial results to classify unknown victim demands. Specifically, after training a naïve Bayes classifier on 500,000 PRIME-PROBE trials per (d, k_a, k_v) triple, we tested it on another 500,000 trials. To filter out PROBE readings due to page faults, excessively large readings were discarded from our evaluation. The tests without CACHEBAR yielded the confusion matrix in Table 7(a), with overall accuracy of 67.5%. In this table, cells with higher numbers have lighter backgrounds, and so the best attacker would be one who achieves white cells along the diagonal and dark-gray cells elsewhere. As can be seen there, classification by the attacker was very accurate for d falling into NONE, ONE, or LOTS; e.g., $d = 1$ resulted in a classification of ONE with probability of 0.80. Other demands had lower accuracy, but were almost always classified into adjacent classes; i.e., *every* class of victim demand was classified correctly or as an adjacent class (e.g., $d \in$ FEW was classified as ONE, FEW, or SOME) at least 96% of the time.

In contrast, Fig. 7(b) shows the confusion matrix for a naïve Bayes classifier trained and tested using PRIME-PROBE trials conducted with CACHEBAR enabled. Specifically, these values were calculated using

$$\begin{aligned} & \mathbb{P}(\text{class} = c \mid d \in c') \\ &= \sum_{4 \leq k_a, k_v \leq 14} \left(\mathbb{P}(\text{class} = c \mid d \in c' \wedge K_v = k_v \wedge K_a = k_a) \right. \\ & \quad \left. \cdot \mathbb{P}(K_a = k_a) \cdot \mathbb{P}(K_v = k_v) \right) \end{aligned}$$

where **class** denotes the classification obtained by the adversary using the naïve Bayes classifier; $c, c' \in \{\text{NONE}, \text{ONE}, \text{FEW}, \text{SOME}, \text{LOTS}, \text{MOST}\}$; and $\mathbb{P}(K_a = k_a)$ and $\mathbb{P}(K_v = k_v)$ are calculated as described in Sec. 4.3. The factor $\mathbb{P}(\text{class} = c \mid d \in c' \wedge K_v = k_v \wedge K_a = k_a)$ was measured empirically. Though space limits preclude reporting the full class confusion matrix for each k_v, k_a pair, the accuracy of the naïve Bayes classifier per k_v, k_a pair, averaged over all classes c , is shown in Fig. 8. As in Fig. 7, cells with larger values in Fig. 8 are more lightly colored, though in this case, the diagonal has no particular significance. Rather, we would expect that when the attacker and victim are each

CACHEBAR could accommodate $m = 3$ or larger. That said, we are unaware of prior works that have successfully conducted PRIME-PROBE attacks from multiple colluding attackers, which would itself face numerous challenges (e.g., coordinating PROBES by multiple processes).

		Classification by attacker					
		NONE	ONE	FEW	SOME	LOTS	MOST
Victim demand d	NONE	.96	.04	.00	.00	.00	.00
	ONE	.01	.80	.19	.01	.00	.00
	FEW	.00	.16	.50	.30	.04	.00
	SOME	.00	.00	.07	.54	.34	.04
	LOTS	.00	.00	.00	.03	.84	.13
	MOST	.00	.00	.00	.03	.56	.41

(a) Without CACHEBAR

		Classification by attacker					
		NONE	ONE	FEW	SOME	LOTS	MOST
Victim demand d	NONE	.33	.16	.26	.18	.04	.02
	ONE	.16	.36	.19	.19	.06	.04
	FEW	.13	.14	.40	.19	.09	.05
	SOME	.09	.10	.16	.37	.20	.07
	LOTS	.08	.06	.10	.16	.46	.13
	MOST	.10	.07	.18	.18	.18	.29

(b) With CACHEBAR

Figure 7: Confusion matrix of naïve Bayes classifier

		k_v										
		4	5	6	7	8	9	10	11	12	13	14
k_a	4	.18	.17	.17	.17	.17	.17	.17	.17	.36	.22	.33
	5	.19	.17	.30	.32	.27	.27	.20	.26	.33	.46	.39
	6	.17	.31	.24	.18	.21	.17	.20	.27	.43	.39	.41
	7	.17	.33	.22	.22	.19	.31	.33	.33	.46	.48	.54
	8	.33	.35	.32	.23	.43	.37	.43	.42	.32	.38	.49
	9	.20	.26	.31	.28	.44	.38	.34	.34	.46	.39	.56
	10	.41	.31	.27	.35	.50	.55	.53	.31	.53	.50	.62
	11	.45	.45	.40	.45	.47	.54	.54	.57	.67	.50	.50
	12	.55	.50	.59	.63	.49	.48	.54	.49	.56	.58	.57
	13	.55	.53	.68	.68	.54	.65	.52	.56	.57	.66	.66
	14	.53	.56	.45	.65	.46	.62	.48	.68	.55	.57	.53

Figure 8: Accuracy per values of k_v and k_a

limited to fewer lines in the cache set (i.e., small values of k_a and k_v , in the upper left-hand corner of Fig. 8) the accuracy of the attacker will suffer, whereas when the attacker and victim are permitted to use more lines of the cache (i.e., in the lower right-hand corner) the attacker’s accuracy would improve. Fig. 8 supports these general trends.

Returning to Fig. 7(b), we see that CACHEBAR substantially degrades the adversary’s classification accuracy, which overall is only 33%. Moreover, the adversary is not only wrong more often, but is also often “more wrong” in those cases. That is, whereas in Fig. 7(a) shows that each class of victim demand was classified as that demand or an adjacent demand at least 96% of the time, this property no longer holds true in Fig. 7(b). Indeed, the attacker’s *best* case in this regard is classifying victim demand LOTS, which it classifies as SOME, LOTS, or MOST 75% of the time. In the case of a victim demand of MOST, this number is only 47%.

5.3 Performance Evaluation

In this section we describe tests we have run to evaluate the performance impact of CACHEBAR relative to an unmodified Linux kernel. As mentioned previously, we are motivated by side-channel prevention in PaaS clouds, and so we focused our evaluation on typical PaaS applications.

In order to increase server utilization and reduce cost, most public PaaS clouds isolate tenants within the same operating system using Linux containers. While a web application may contain web servers, programming language runtimes, databases, and a set of middleware that enrich its

Table 1: Server+language support in selected PaaS clouds

PaaS cloud	Supported server engines (+ application languages)
AppFog	Tomcat, Apache, Nginx, IIS (Java, Python, PHP, Node.js, Ruby, Go)
Azure	Tomcat, Jetty, Apache, Nginx, GlassFish, Wildfly, IIS (Java, Python, PHP, Node.js, Ruby, .NET)
Elastic Beanstalk	Tomcat, Apache, Nginx, Passenger, Puma, IIS (Java, Python, PHP, Node.js, Ruby, Go, .NET)
Engine Yard	Nginx, Rack, Passenger, Puma, Unicorn, Trinidad (Java, PHP, Node.js, and Ruby)
Google Cloud	JBoss, Wildfly, Tomcat, Apache, Nginx, Zend, Passenger, Mongrel, Thin, IIS (Java, Python, PHP, Node.js, Ruby, ASP.NET, Go)
Heroku	Jetty, Tomcat, Tornado, Nginx, Apache, Mongrel, Thin, Puma, Unicorn, Hypnotoad, Starman, Mongoose, Yaws, Mochiweb (Java, Python, PHP, Node.js, Ruby, Go, Perl, C, Erlang, Scala, Clojure)
HP Stackato	Apache, Apache TomEE, Nginx (Java, Python, PHP, Node.js, Ruby, Perl, Erlang, Scala, Clojure, ASP.NET)
OpenShift	JBoss, Wildfly, Tomcat, Apache, Spring, Tornado, Zend, Vert.x (Java, Python, PHP, Node.js, Ruby, Perl, Ceylon)

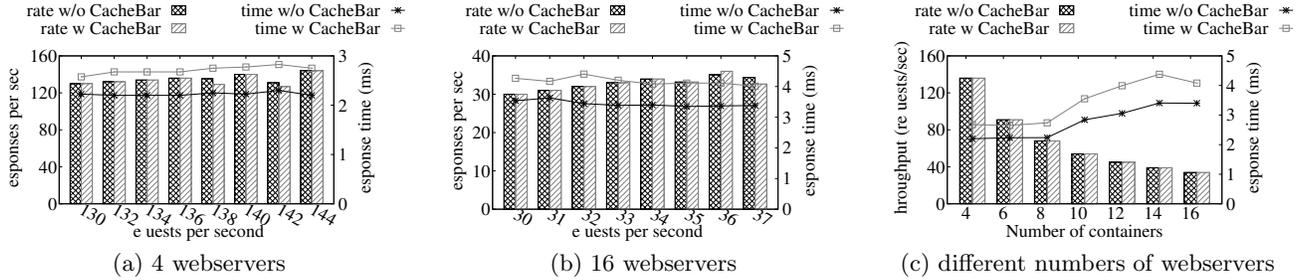


Figure 9: Average throughput and response time per Apache+PHP-FPM server, each in a separate container

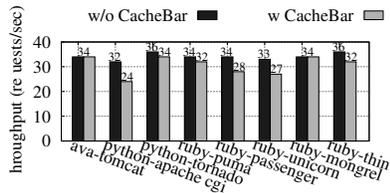


Figure 10: By webserver+language

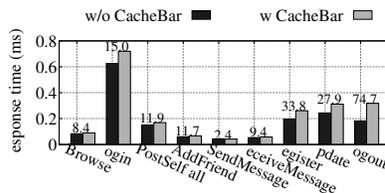


Figure 11: By operation

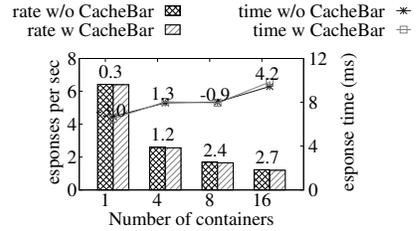


Figure 12: Media streaming

functionality, in all PaaS clouds we have studied, language runtimes and web servers are located on different servers from databases and middleware; web/app servers controlled by different tenants may share the same OS, however. Because users of PaaS clouds do not have the permission to execute arbitrary code on databases and middleware that are typically shared by multiple tenants, the targets of the side-channel attacks we consider in this paper are primarily web servers that supports various language runtimes, which may be co-located with the adversary-controlled malicious web servers on which arbitrary code can be executed. We conducted a survey to understand the popular web/app servers that are used in major PaaS clouds, and the programming languages they support; see Table 1.

5.3.1 Runtime and Throughput Overhead

Our experiments explored CACHEBAR’s performance (1) per the number of container (and webserver) instances; (2) for different combinations of webserver and application language; (3) for complex workloads characteristic of a social networking website; and (4) for media-streaming workloads.

Webserver performance. In the first experiments, each container ran an Apache 2.4.7 web server with PHP-FPM and SSL enabled. We set up one client per server using autobench; clients were spread across four computers, each with the same networking capabilities as the (one) server computer (not to mention more cores and memory than the

server computer), to ensure that any bottlenecks were on the server machine. Each client repeatedly requested a web page and recorded its achievable throughputs and response times at those throughput rates. The content returned to each client request was the 86KB output of `phpinfo()`.

Fig. 9 shows the throughputs and response times when clients sent requests using SSL without reusing connections. In particular, Fig. 9(a) shows the achieved response rates (left axis) and response times (right axis), averaged over all containers, as a function of offered load when there were four containers, and so four web servers). Bars depict average response rates running over unmodified Linux (“rate w/o CACHEBAR”) or CACHEBAR (“rate w CACHEBAR”), and lines depict average response times running over unmodified Linux (“time w/o CACHEBAR”) or CACHEBAR (“time w CACHEBAR”). Fig. 9(b) shows the same information for 16 containers. As can be seen in these figures, the throughput impact of CACHEBAR was minimal, while the response time increased by around 20%. Fig. 9(c) shows this information in another way, with the number of containers (and hence servers) increasing along the horizontal-axis. In Fig. 9(c), each bar represents the largest request rate at which the responses could keep up.

Webserver+language combinations. Next, we selected other common webserver+app-language combinations, namely Java over a Tomcat web server, Python over Apache+cgi, Python over Tornado, and Ruby over Puma. For each con-

figuration, we instantiated 16 containers and set each up to dynamically generate 80KB random strings for clients. We also did tests using another four web servers running the same Ruby application, namely Passenger, Unicorn, Thin, and Mongrel. Fig. 10 shows the throughput that resulted in each case, over Linux and over CACHEBAR. As shown there, the throughput overheads were modest for most of the server+language combinations that we considered. The worst case was Python over Apache+cgi, which suffered a throughput degradation with CACHEBAR of 25%; other degradations were much more modest.

Impact on a more complex workload. To test effects on more complex workloads, we used the webserver instance in CloudSuite [8] that implements a social community website written in PHP over Nginx on our CACHEBAR-protected machine. This implementation queries a MySQL database and caches results using Memcached; in keeping with PaaS architectures, the database and Memcached server were implemented on another machine without protection, since tenants cannot typically execute directly on these machines. We used the Faban tool to generate a mix of requests to the webserver, including `browse` (7.9%), `login` (7.5%), `post` (24.9%), `addFriend` (7.3%), `sendMsg` (44.0%), `register` (0.8%), and `logout` (7.5%). In addition, a background activity happened on the webserver every 10s, which was either `receiveMsg` or `update` with equal likelihood. Fig. 11 shows that the responsiveness of the various common operations suffered little with CACHEBAR, between 2% and 15% overhead. Three operations (`register`, `update`, and `logout`) suffered over 25% overhead, but these operations were rare in the Faban workload (and presumably in practice).

Media streaming in CloudSuite. In addition to the webserver benchmark setup used above, CloudSuite offers a media streaming server running over Nginx that serves 3.1GB static video files at different levels of quality. We set up a client process per server to issue a mix of requests for videos at different quality levels and, through a binary search, to find the peak request rate the server can sustain while keeping the failure rate below a threshold. Fig. 12 shows that CACHEBAR affected this application least of all, in both throughput and response time.

SPEC CPU 2006 benchmarks. For completeness, we measured the impact of CACHEBAR on nine SPEC CPU 2006 benchmarks. Six resulted in reasonable overheads: `hmmr` (13.3% overhead), `gamess` (3.5%), `gromacs` (13.1%), `namd` (14.3%), `povray` (0.4%), and `tonto` (16.8%). However, three exhibited substantially higher overheads: `perlbench` (225%), `bzip2` (76%), and `h264ref` (143%). (Overheads caused by copy-on-access alone were below 5%.) It is not surprising that limiting cache usage using cacheable queue can interfere with some workloads. CACHEBAR is not a panacea and is best suited for the PaaS workloads that formed the core of our evaluation.

5.3.2 CACHEBAR’s Memory Savings

To measure the memory savings that copy-on-access offers over disabling memory sharing between containers, we measured the total unique physical memory pages used across various numbers of web servers, each in its own container, when running over (i) unmodified Linux, (ii) Linux without cross-container memory sharing, and (iii) CACHEBAR-

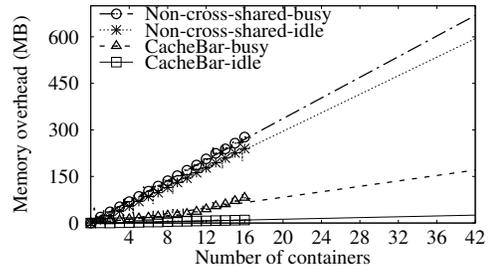


Figure 13: Memory overhead comparison

enabled Linux. We used the system diagnosis tool `smem` for memory accounting, specifically by accumulating the PSS (proportional set size) field output by `smem` for each process, which reports the process’ shared memory pages divided by the number of processes sharing these pages, plus the process’ unshared memory pages and all kernel pages.

Fig. 13 shows the memory overhead of Linux without cross-container sharing and with CACHEBAR, computed by subtracting the memory measured for unmodified Linux from the memory measured for each of these systems. We grew the number of containers to 16 in each case, and then extrapolated to larger numbers of containers using best-fit lines. As can be seen in Fig. 13, the overhead of CACHEBAR is virtually zero (“CACHEBAR-idle”) with negligible query load. “Non-cross-shared-busy” and “CACHEBAR-busy” shows the same measures in an experiment where every fourth server was subjected to a slightly more active load of four requests per second. This was enough to induce CACHEBAR’s copy-on-access mechanism to copy some memory pages. Again, however, the memory overhead of CACHEBAR was much less than of disabling cross-container sharing altogether.

6. CONCLUSION

We have presented two techniques to defend against side-channel attacks via LLCs, namely (i) copy-on-access for physical pages shared among multiple security domains, to interfere with FLUSH-RELOAD attacks, and (ii) cacheability management for pages to limit the number of cache lines per cache set that an adversary can occupy simultaneously, to mitigate PRIME-PROBE attacks. We described the implementation of these techniques in a memory-management subsystem called CACHEBAR for Linux, to interfere with LLC-based side-channel attacks across containers. Using formal analysis (model checking for copy-on-access, and probabilistic modeling for cacheability management), we developed designs that mitigate side-channel attacks in our empirical evaluations. Our experiments also confirmed that the overheads of our approach are modest for PaaS workloads.

Acknowledgments. This work was supported in part by NSF grants 1330599 and 1566444.

7. REFERENCES

- [1] A. Arcangeli, I. Eidus, and C. Wright. Increasing memory density by using KSM. In *Linux Symposium*, 2009.
- [2] Y. Azar, S. Kamara, I. Menache, M. Raykova, and B. Shepard. Co-location-resistant clouds. In *6th ACM Cloud Computing Security Workshop*, 2014.
- [3] E. Bosman, K. Razavi, H. Bos, , and C. Giuffrida. Dedup est machina: Memory deduplication as an

- advanced exploitation vector. In *37th IEEE Symposium on Security and Privacy*, 2016.
- [4] B. Coppens, I. Verbauwhede, K. D. Bosschere, and B. D. Sutter. Practical mitigations for timing-based side-channel attacks on modern x86 processors. In *30th IEEE Symposium on Security and Privacy*, 2009.
- [5] S. Crane, A. Homescu, S. Brunthaler, P. Larsen, and M. Franz. Thwarting cache side-channel attacks through dynamic software diversity. In *ISOC Network and Distributed System Security Symposium*, 2015.
- [6] L. Elizaveta and P. Bickel. The earth mover's distance is the Mallows distance. In *8th International Conference on Computer Vision*, 2001.
- [7] F. J. T. Fábrega, F. Javier, and J. D. Guttman. Copy on write, 1995.
- [8] M. Ferdman et al. Clearing the clouds: a study of emerging scale-out workloads on modern hardware. In *ACM SIGPLAN Notices*, 2012.
- [9] D. Gruss, C. Maurice, and K. Wagner. Flush+Flush: A stealthier last-level cache attack. *CoRR*, abs/1511.04594, 2015.
- [10] D. Gullasch, E. Bangerter, and S. Krenn. Cache games – bringing access-based cache attacks on AES to practice. In *32nd IEEE Symposium on Security and Privacy*, 2011.
- [11] Intel. *Intel® 64 and IA-32 Architectures Software Developer's Manual*, 2010.
- [12] G. Irazoqui, T. Eisenbarth, and B. Sunar. S\$A: A shared cache attack that works across cores and defies VM sandboxing—and its application to AES. In *36th IEEE Symposium on Security and Privacy*, 2015.
- [13] G. Keramidas, A. Antonopoulos, D. N. Serpanos, and S. Kaxiras. Non deterministic caches: A simple and effective defense against side channel attacks. *Design Automation for Embedded Systems*, 12(3), 2008.
- [14] T. Kim, M. Peinado, and G. Mainar-Ruiz. STEALTHMEM: System-level protection against cache-based side channel attacks in the cloud. In *USENIX Security Symposium*, 2012.
- [15] R. Könighofer. A fast and cache-timing resistant implementation of the AES. In *Topics in Cryptology–CT-RSA*, 2008.
- [16] M. Li, Y. Zhang, K. Bai, W. Zang, M. Yu, and X. He. Improving cloud survivability through dependency based virtual machine placement. In *International Conference on Security and Cryptography*, 2012.
- [17] P. Li, D. Gao, and M. K. Reiter. StopWatch: A cloud architecture for timing channel mitigation. *ACM Trans. Information and System Security*, 17(2), 2014.
- [18] F. Liu, Q. Ge, Y. Yarom, F. Mckeen, C. Rozas, G. Heiser, and R. B. Lee. Catalyst: Defeating last-level cache side channel attacks in cloud computing. In *22nd IEEE Symposium on High Performance Computer Architecture*, 2016.
- [19] F. Liu and R. B. Lee. Random fill cache architecture. In *47th Annual IEEE/ACM International Symposium on Microarchitecture*, 2014.
- [20] F. Liu, Y. Yarom, Q. Ge, G. Heiser, and R. B. Lee. Last-level cache side-channel attacks are practical. In *36th IEEE Symposium on Security and Privacy*, 2015.
- [21] D. A. Osvik, A. Shamir, and E. Tromer. Cache attacks and countermeasures: the case of AES. In *Topics in Cryptology–CT-RSA*, 2006.
- [22] R. Owens and W. Wang. Non-interactive OS fingerprinting through memory de-duplication technique in virtual machines. In *30th IEEE International Performance Computing and Communications Conference*, 2011.
- [23] S. Raikin et al. Tracking mechanism coupled to retirement in reorder buffer for indicating sharing logical registers of physical register in record indexed by logical register, 2014. US Patent 8,914,617.
- [24] H. Raj, R. Nathuji, A. Singh, and P. England. Resource management for isolation enhanced cloud services. In *ACM Cloud Computing Security Workshop*, 2009.
- [25] A. Rane, C. Lin, and M. Tiwari. Raccoon: Closing digital side-channels through obfuscated execution. In *24th USENIX Security Symposium*, 2015.
- [26] J. Shi, X. Song, H. Chen, and B. Zang. Limiting cache-based side-channel in multi-tenant cloud using dynamic page coloring. In *Workshops of the 41st IEEE/IFIP International Conference on Dependable Systems and Networks*, 2011.
- [27] D. Stefan, P. Buiras, E. Z. Yang, A. Levy, D. Terei, A. Russo, and D. Mazières. Eliminating cache-based timing attacks with instruction-based scheduling. In *Computer Security–ESORICS*, 2013.
- [28] E. Tromer, D. A. Osvik, and A. Shamir. Efficient cache attacks on AES, and countermeasures. *Journal of Cryptology*, 23(1), 2010.
- [29] V. Varadarajan, T. Ristenpart, and M. Swift. Scheduler-based defenses against cross-VM side channels. In *23rd USENIX Security Symposium*, 2014.
- [30] B. C. Vattikonda, S. Das, and H. Shacham. Eliminating fine grained timers in Xen. In *3rd ACM Cloud Computing Security Workshop*, 2011.
- [31] Z. Wang and R. B. Lee. A novel cache architecture with enhanced performance and security. In *41st IEEE/ACM International Symposium on Microarchitecture*, 2008.
- [32] J. C. Wray. An analysis of covert timing channels. In *1991 IEEE Symposium on Security and Privacy*, 1991.
- [33] Y. Yarom and K. E. Falkner. FLUSH+RELOAD: A high resolution, low noise, L3 cache side-channel attack. In *USENIX Security Symposium*, 2014.
- [34] Y. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart. Cross-VM side channels and their use to extract private keys. In *ACM Conference on Computer & Communications Security*, 2012.
- [35] Y. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart. Cross-tenant side-channel attacks in PaaS clouds. In *ACM Conference on Computer & Communications Security*, 2014.
- [36] Y. Zhang, M. Li, K. Bai, M. Yu, and W. Zang. Incentive compatible moving target defense against VM-colocation attacks in clouds. In *27th IFIP Information Security and Privacy Conference*, 2012.
- [37] Y. Zhang and M. K. Reiter. Düppel: Retrofitting commodity operating systems to mitigate cache side channels in the cloud. In *ACM Conference on Computer & Communications Security*, 2013.