Preliminary RISC-V HFI Spec (v0)

Shravan Narayan, Tal Garfinkel

July 2024

1 Purpose

Hardware-assisted fault isolation (HFI) is a hardware extension that supports in-process isolation (sandboxing) of unmodified native binary code and and also integrates into existing software based fault isolation (SFI [9]) based sandboxing systems, systems such as WebAssembly [2] and Google's Ubercage (Google Chrome's v8 JIT sandbox [8]), to improve their performance, security, and scalability [5].

HFI aims to provide all the capabilities needed for secure sandboxing namely data and control flow isolation as well as complete and efficient mediation of privileged instructions (i.e. system calls). This specification describes the functional properties of HFI on RISC-V, a more complete description of the background and motivation for HFI is presented in a paper published at ASPLOS 2023 [5].

2 Overview

HFI mode. HFI adds a new processor mode (HFI mode). If HFI is enabled, code running on a hart is "sandboxed", i.e, execution is restricted according to: (a) a set of region registers, that grant access to memory (see §6), (b) a register with the sandbox exit handler, where system calls and sandbox exits are redirected to (see §4), and (c) a register with sandbox option flags that modify sandbox semantics (see §3). HFI's mode and region registers operate on a per-hart basis; they are not synchronized across harts.

For convenience, we refer to the code using HFI to sandbox other code as the *sandboxing runtime*. With a few exceptions (§7.2), HFI is enabled when a sandboxing runtime executes an `hfi_enter` instruction, and disabled when sandboxed code executes an `hfi_exit` instruction, which transfers control back to the sandboxing runtime. The runtime is responsible for saving and restoring context appropriately, and can use HFI to multiplex many sandboxes across harts, scheduling them as it sees fit. HFI enables sandboxing through two mechanisms:

Interposition. HFI supports interposition on all paths out of the sandbox including when a sandbox exits (through the `hfi_exit` instruction), system calls — and by extension, signals. Thus, a runtime can use HFI to mediate all control flow and access to sensitive OS resources.

Regions. HFI offers memory and control isolation using a finite set of *regions* — portions of a processes' virtual memory described by *base* (start address for the region) and *bound* (size of the region). By default, an HFI sandbox has no regions defined and thus cannot access memory to either access data or execute code. To grant access, the sandboxing runtime configures regions of memory allowed by using several dedicated instructions (e.g. `hfi_set_region_size`, `hfi_set_region_permission`). Regions come in three types:

- *Implicit data regions:* grant read and/or write memory access to a portion of memory, and are applied to every memory access performed by sandboxed code. For example, if sandboxed code executes an instruction — *load address X into register Y* — HFI will ensure that at least one of the implicit regions has a range that includes X, and then applies the permissions from the first matching region.
- *Implicit code regions:* are similarly used to grant execute permissions, and applied to every instruction fetch a sandboxed program executes.

- *Explicit data regions:* are used to grant read/write access to sandboxed programs, but only through dedicated instructions that perform region-relative memory operations. Specifically, loads and stores to these regions are performed as offsets into the region using new h-prefixed variants the standard RISC-V memory instructions such as `hlw` and `hsw`. The offsets are checked to ensure they remain within the explicit region.

The new instructions introduced by HFI are shown in Figure 1, while the new control and status registers as well as internal registers are shown in in Figure 2.

3 Sandbox Setup

The sandboxing runtime can query if the processor supports HFI by checking the supported extension string from the device tree or the ACPI RISC-V Hart Capabilities Table as is the common practice [3]. Specifically, the string will list "hfi-*version*" if HFI is supported. The HFI version described in this spec is "1".

If HFI is supported, HFI mode is enabled with the `hfi_enter` instruction, and disabled with the `hfi_exit` instruction. The `hfi_enter` instruction has two variants — the first variant takes a single register operand which specifies the sandbox configuration options as a bit vector called `hfi_options_t`, that has the following fields:

- *lock_regions: reg_u1* — Regions configurations are normally specified prior to `hfi_enter`, using instructions such `hfi_set_region_size` (See Section 6). If *lock_regions* is set to *false (0)*, region manipulation instructions can also be called while in HFI mode (i.e., the region configuration can be modified in the sandbox). If *true (1)*, these instructions cannot be called while in HFI mode.

- *redirect_system_calls: reg_u1* — HFI sandboxes can interpose on system calls made by sandboxed code. If *redirect_system_calls* is set to *false (0)*, system calls are unaffected by HFI. If *true (1)*, HFI will redirect system calls (ecall) instructions to the HFI exit handler (which can be set with the `hfi_set_exit_handler` instruction discussed in Section 4).

- *redirect_exits: reg_u1* — HFI sandboxes can interpose on invocations of `hfi_exit` so the application can take control once the sandbox code completes. If *redirect_exits* is set to *false (0)*, `hfi_exit` disables HFI and execution simply falls through to the next instruction. If *true (1)*, `hfi_exit` disables HFI and redirects control flow to the exit_handler (which can be set with the `hfi_set_exit_handler` instruction discussed in Section 4).

- *serialize_enter_exits: reg_u1* — HFI sandboxes can add fences sandbox entries and exit (through the `hfi_enter` and `hfi_exit` instructions), a required step for Spectre protections. If *serialize_enter_exits* is *false (0)*, these instructions do not introduce any serialization or fencing. If *true (1)*, these instructions act as a fence for all memory operations that are run prior to these instructions.

The second variant of `hfi_enter` operates like `hfi_enter` variant 1 plus a jump instruction; it takes the `hfi_options_t` as the first operand and takes the target of the jump in a register as the second operand.

Behavior. The `hfi_enter` instruction sets the value of an internal status register `hfi_usermode_enabled` to 1 on execution, and clears the `hfi_access_violation` CSR indicating that there has not been a violation of HFI's region rules since the last invocation of `hfi_enter`. The second variant of `hfi_enter` additionally sets the PC to the target specified in the second operand after its execution. The `hfi_exit` instruction will set `hfi_usermode_enabled` to 0 on execution, and writes the value of the program counter of the `hfi_exit` instruction to the `hfi_exit_pc` CSR, and sets the `hfi_exit_reason` CSR to 1, indicating the sandbox exit was due to and `hfi_exit` instruction. `hfi_enter` and `hfi_exit` must always act as fence for other HFI instructions, i.e., they only execute after all in-flight HFI instructions have completed.

Faults. The `hfi_enter` instruction will trap if the CPU is already in HFI mode. The `hfi_exit` instruction will trap if the CPU is not in HFI mode.

Design Rationale. Most of the options in `hfi_options_t` are present to support to different use cases. When executed unmodified native code, the sandboxed code is totally untrusted, thus regions are locked,

system calls are redirected, etc. However, when sandboxing code from existing SFI systems such as WebAssembly, it is more efficient/not necessary for HFI to do some checks, since these systems already handle some safety checks in their compiler/runtime. For example, these systems may not want to redirect system calls. Similarly, the choice of whether or not to serialize `hfi_exit` and `hfi_enter` will depend on the threat model that a particular application is trying to enforce (i.e. if they want greater Spectre safety), we leave it to the user to choose whether they wish to opt into this added overhead. Finally, `hfi_enter` is offered in two variants as some sandboxing runtimes may want the option to jump directly to executing sandboxed code after sandboxing is enabled via `hfi_enter`, while other runtimes may want the instruction to fall through to avoid any costs due to extra control flow. The former behavior can difficult to accomplish with the single operand variant of `hfi_enter` as it requires the instruction following `hfi_enter` to be part of the sandbox's code — which may not always be possible. The latter behavior's performance is difficult to achieve with an instruction that includes control flow.

4 Configuring the Exit handler

HFI supports interposition via. redirection on all paths out of the sandbox including sandbox exits (via. `hfi_exit`) and system calls (and by extension signals). As noted, which instructions (system calls and/or `hfi_exit`) are redirected is configured through `hfi_options_t`, that is passed as an operand to `hfi_enter`.

To handle this redirection, an exit handler is setup with the `hfi_set_exit_handler` instruction. This instruction should be invoked prior to `hfi_exit`. `hfi_set_exit_handler` takes one operand, a 64-bit register, that holds the address of the exit handler. The current exit handler can be retrieved via. the `hfi_get_exit_handler` instruction.

Behavior. The `hfi_set_exit_handler` instruction sets the `hfi_exit_handler_reg` CSR to the address specified in its operand.

- If `hfi_exit` is configured to invoke the exit handler, the `hfi_exit` instruction, when executed by sandbox code, will jump to the exit handler *after* execution (during which the `hfi_exit_reason` CSR is set to 1 indicating the exit was due to the `hfi_exit` instruction).
- If system calls are configured to invoke the exit handler, the system call instruction, when executed by sandbox code, will jump to the exit handler *before* execution, set the `hfi_exit_reason` CSR is set to 2 indicating the exit was due to a system call, and disable sandboxing by setting `hfi_usermode_enabled` CSR to 0.

Faults. The `hfi_set_exit_handler` instruction will trap if the CPU is already in HFI mode. If the exit handler is set to a location without code permissions, the behavior of the CPU would be identical to a normal jump to an address without code permissions.

Design Rationale. Sandboxed code will invoke `hfi_exit` to exit the sandbox. We must ensure that the sandboxed code always returns control to the sandboxing runtime after exits, which means, invocation of `hfi_exit` should return control to trusted code. Thus, we have added support for redirecting all invocations of `hfi_exit`.

Sandboxed code can also invoke system calls; since HFI's restrictions don't apply to kernel code, system calls could be used to bypass isolation enforced by the hardware [1]. Thus, trusted code needs the ability to interpose on system calls, so that it can restrict the invocation of unsafe system calls by sandboxed code. Such interposition on system calls could be done using assembly rewriting or using kernel features such as EBPF, however, this is slow and cumbersome. To allow efficient interposition of system calls, HFI provides hardware support to redirect system calls. When enabled, system calls simply act like a jump instruction to the exit handler.

5 New CSRs and internal registers

HFI stores state for the current sandbox in CSRs and internal registers including: (1) the sandbox status (2) the exit handler (3) region configuration (4) the cause and status of HFI induced faults (traps). These are detailed in Figure 2.

The sandbox status, stored in the `hfi_status_reg` CSR, is read-only to userspace software, but writable by kernel code. The `hfi_status_reg` register should support updates via register renaming to support the performance expectations of hardware sandboxing. These values are expected to change frequently (once in few thousand instructions), and serializing this register updates with techniques like score-boarding will hinder practical adoption. The contents of the `hfi_status_reg` register are as follows:

- `hfi_usermode_enabled` (1-bit): is updated automatically during `hfi_enter` (set to true), `hfi_exit` (set to false), and when the exit handler is called (set to false).
- `hfi_exit_reason` (2-bit): indicates the reason for the last sandbox exit. Exit due to `hfi_exit` would leave this set to 1. Exit due to a system call with leave this set to 2.
- `hfi_exit_pc` (60-bit): indicates the PC of the instruction causing the last exit. The first 2-bits and the last 2-bits of the PC are dropped and are assumed to be zero. Thus 60-bits of the PC are stored.

The exit handler of the sandbox is stored in the `hfi_exit_handler_reg` internal register and can be set or read via the `hfi_set_exit_handler` and `hfi_get_exit_handler` instructions. Since the register is not expected to be frequently updated, it's updates may be scoreboarded if needed.

Regions are stored in internal registers. Different version of HFI may require different numbers of these registers, as discussed in §9.2. These registers cannot be directly named, and must instead be modified or accessed through HFI instructions (e.g. `hfi_get_region_size`). Updates to these registers should ideally be supported through register renaming; however, implementations may choose to use the slightly less expensive scheme: In particular, implementations only need to ensure register updates must not fence/serialize when performed with `hfi_usermode_enabled` register is 0 (i.e., we are not running sandboxed code), and defer fencing to when an `hfi_enter` instruction executes, which should wait for all pending updates to these registers to complete. Updates to this register must serialize when performed when the `hfi_usermode_enabled` register is set to 1, which is permitted if the `lock_regions` option of the sandbox is set to 0.

The result of an HFI fault due to an HFI policy violation is stored in the `hfi_fault_status_reg` register. This register stores the precise cause of an HFI induced fault (trap). Since the register is not expected to be frequently updated, it's updates may be scoreboarded if needed. Information in `hfi_fault_status_reg` is readable in user space so the runtime can respond appropriately, and read-write in the kernel so that this state can be saved/restored as part of the process context, as multiple processes may be using HFI, and fault delivery is asynchronous. This register has four fields:

- `hfi_fault_occured` (1-bit): is a bit that indicates that an hfi fault has occurred.
- `hfi_fault_region` (8-bit): indicates which region the fault occured in, if the fault was do to an explicit region trap (out of bounds or insufficient permissions), it will contain the number of the region that caused the fault, i.e., the operand to an h-prefixed instruction. If the fault was caused by an implicit region i.e. insufficient permission to access a matched region,it will contain the number of the region. If the fault was cause because no implicit region matched an operation, it will contain 0.
- `hfi_fault_op` (2-bit): indicates the operation that faulted, this will be a `LOAD_FAULT` or `STORE_FAULT` for a failed load or store, or `FETCH_FAULT` for a failed instruction fetch.
- `hfi_fault_type` (1-bit): indicates the type of fault. If `OUT_OF_BOUNDS`, it indicates that an operation was out of bounds if it was explicit region fault , or that no implicit region matched the operation (in this case `hfi_fault_region` will be equal to zero). If `INSUFFICIENT_PERMISSIONS`, either an explicit region had insufficient permissions for an access, or whatever implicit region matched the operation did not have sufficient permissions.

Design Rationale. The `hfi_enter` and `hfi_exit` instructions modify the various bits of the `hfi_status_reg` register. The updates to this register must be fast to allow rapid entries and exits to HFI mode; thus this must be supported by register renaming. Region registers are updated frequently as well when switching between different sandboxes. While this would ideally support register renaming for efficient register updates, an alternate scheme that ensures a single serialization (on entry into the sandbox) for a batch of region updates prior to entry, would provide adequate performance. Finally, updates to the region configuration within the sandbox should serialize, as an updates during speculative execution may allow Spectre-style attacks to break out of the sandbox.

6 Regions

Regions offer a limited version of the functionality found in traditional segmented memory systems, that control access using <base, bound, permission> tuples to control access to contiguous ranges of memory.

By default, a processor in HFI mode has no access to memory i.e. it cannot read data or run code. To enable sandboxed code to run, a sandbox runtime must explicitly configure regions prior executing `hfi_enter`. In the next sections, we provide a brief overview of the three types of HFI regions, and then specify how to configure these regions.

6.1 Regions Types

HFI offers three region types implicit code, implicit data, and explicit data. Each type is specialized to particular tasks, with the aim of reducing hardware complexity.

Implicit Data and Implicit Code Regions Implicit regions are essential for isolating memory accesses and control flow of unmodified native code, as well as other situations where explicit regions (described next) would be impossible to use. HFI discriminates implicit regions into code and data regions, to keep the control and data pipelines simpler and more efficient. Data regions can grant read and write access and only apply to loads and stores, while code regions apply only to instruction fetches, and can only grant execute permissions.

Implicit data region checks apply to every memory access, and grant access on a first-match basis. For example, if sandboxed code executes an "`lw rd, rs(offset)`" instruction, HFI will check if the address in `rs + offset` is in range in *any* of the implicit regions in parallel. For the first matching implicit region, it will check the permissions to see if reads are allow–if so, it will proceed. If the permission check fails, or if there is no match, HFI will trap. Implicit code regions apply similar checks to code.

Implicit regions perform bounds checks based on *prefix matching*. Concretely, each region specifies a *base_prefix* (the region's base address) and an *lsb_mask*. To check if an address is in bounds, HFI uses the *lsb_mask* to remove the least significant bits of the address, and compares the remaining prefix to *base_prefix*. Implicit regions thus must be power of two sized and aligned — thus, they trade granularity for efficient checking—in particular, checks can be implemented with simple masking operations. Implicit regions checks are not applied to operations on explicit regions, which we discuss next.

Explicit regions. Explicit regions provide region relative addressing, i.e., addressing is always relative to the base of the currently *active region* (by default, explicit data region 1). This offers efficient fine grain control over access to memory within a particular sandbox address or shared buffer. HFI provides two different region sizes (*large/small*), with different granularities. Large regions can address up to 256 TiB (2^{48}) and are sized and aligned to multiples of 64K (2^{16}). Small regions, in contrast, can only address up to 4GiB (2^{32}), but are byte granular in size and alignment.

To access memory through explicit regions, a program must use the h-prefixed variants of normal load and store instructions (`hlw`, `hsw`, `hlh`, etc.). These instruction target the active region, but can be changed to use a different region as described in §9.2). For example, `hlw x1, A(x0)` will succeed if the address being loaded is falls within explicit region 1, and there is a read permission set on that region, otherwise it will fail.

A few notes on size and alignment. Unaligned memory operations that are split by the micro-architecture into independent operations will be checked independently by HFI. If one split faults, the original fused instruction will fault, as will the split instruction that violates HFI bounds. However, the split operation that is within bounds may is allowed to have visible *micro-architectural* side effects within the sandbox.

To safely implement explicit bounds checks, the bounds check must be applied to the operand of an operation (i.e., the address being loaded/stored to) plus the size of the operand, to ensure that longer operations do not exceed bounds checks.

Regions also have minimum sizes. The smallest large explicit region is 64K, there is no minimum size on small explicit regions. The smallest implicit region is 64 bytes. The behavior of regions that do not meet these minimums is undefined.

Design Rationale. The large and small region sizes and alignment constraints on explicit regions allow us to implement explicit regions with a single 32-bit comparator. For small regions, HFI checks the least-significant 32-bits is within bounds, and ensures the top bits are zero. For large regions, HFI will drop the first 16-bits, and compare bits 16-48, while checking the top bits are zero.

While allowing regions that support arbitrary address ranges at with any size and alignment is conceptually simpler than specialized large and small regions, our restrictions allow bounds checking with very simple hardware. HFI's large and small regions constraints can be checked with a single 32-bit comparator, rather than the more costly multiple 64-bit comparators needed to check arbitrary region bounds.

Explicit regions' added granularity is critical for supporting Wasm heaps, which grow in 64K increments [2] — while byte granularity is critical for efficiently sharing individual memory objects and sandboxing legacy code, as existing buffers can be shared in-place changing code or allocators.

6.2 Manipulating Regions

Region state, which is stored in internal registers, can only read or modified HFI instructions. The instructions broadly operate on regions by a region number(`region_number_t`) — a unique number/index assigned to each HFI region on the CPU.

Region number assignments. The version of HFI defined in this spec (hfi1), defines three regions: one explicit data region with a `region_number_t` of 1, one implicit data region of `region_number_t` of 2, and one implicit data region of `region_number_t` of 3. Future versions may define multiple regions of each type (§9.2), and each region will be assigned a unique `region_number_t`.

We now discuss the instructions that can configure these regions:

Setting Region base and size. The following instruction are used to specify the range of memory a region applies to.

```
hfi_set_region_size (region_number_t, reg_u64 base, reg_u64 mask_or_bound)
hfi_get_region_size (region_number_t) -> (reg_u64 base, reg_u64 mask_or_bound)
```

Behaviour. Region sizes and locations are set using the `hfi_set_region_size`. Regions are typically setup prior to entering the sandbox (with `hfi_enter`). If this instruction runs in a sandbox, i.e., `hfi_usermode_enabled` is 1 — which is allowed when `lock_regions` is 0—it must act as a memory fence. All prior memory instructions must complete before executing this instruction and subsequent memory operations should be issued only after the update.

`hfi_set_region_size` has `region_number_t` as it's first operand register, the base of the region as it's second operand register, and the third operand depends on the type of region. If the region is an explicit data region, the third operand should contain the bound/size of the region; if the region is an implicit data or code region, the third operand should contain the mask for the region. The value of the base and bound/mask should additionally conform to the per-region size and alignment requirements in Section 6.1, however, the instruction does not check that operands meet this criteria. If the operands don't meet the criteria, the resulting behavior is undefined.

`hfi_get_region_size` returns region size information. It takes `region_number_t` as it's first register operand, and returns the base and mask/bound in two output registers.

Faults. These instructions fault if the region number specified does not exist (i.e., it is greater than the total number of regions). For efficiency, these instructions should not check whether region locations or sizes are invalid (e.g., the program has specifies an implicit region base that is not aligned to its size); rather the hardware will continue to operate using the provided base and size, although this behavior is to be considered undefined.

Design Rationale. When `hfi_set_region_size` is run prior to `hfi_enter`, it doesn't need to act as a fence as `hfi_enter` can fulfill this purpose. However, when used inside the sandbox, `hfi_set_region_size` must act as a fence, as otherwise inflight memory operations could potentially access memory outside the sandbox when they were issued if a region resized.

Setting Permissions, Enabling/Disabling Regions. The following instructions are used to configure permissions on regions, as well as to enable and disable regions.

`hfi_set_region_permission (permission_set_t, permission_t)`
`hfi_get_region_permission (permission_set_t) -> permission_t`

Behaviour. Region permissions are set using the `hfi_set_region_permission` instruction. This instruction sets the permissions of all regions using a single bit vector. The instruction takes `permission_set_t` as the first argument. This is a 32-bit register which must have the value 0; other values are reserved for future use. The second operand to this instruction is a permission bit-vector `permission_t` encoded as follows:

- Bits 0 to 3 are permissions for explicit data region 1. Bit 0 indicates if the region is enabled (i.e., should be enforced by the sandbox). Bit 1 and 2 indicates whether the region has read and write permissions respectively. Bit 3 indicates if the explicit data region is a large region (i.e., a region with a bound greater than 4GB as explained in Section 6.1.
- Bits 4 to 6 are permissions for implicit data region 1. Bit 4 indicates if the region is enabled. Bit 5 and 6 indicates whether the region has read and write permissions respectively.
- Bits 7 to 8 are permissions for implicit code region 1. Bit 7 indicates if the region is enabled. Bit 8 indicates whether the region has execute permissions.

Faults. This instruction will fault if the `permission_set_t` is not set to 0, or if `hfi_usermode_enabled` and `lock_regions` is set to 1 (when hfi mode is on with region configurations locked). Any unused bits of `permission_t` are ignored; thus setting an unused bit will not fault.

Design Rationale. An alternate design for this instruction would be to split up `hfi_set_region_permission` to operate per-region (by changing the instruction to take `region_number_t` to operate on as the first parameter). However, this would lead to additional overheads in practice. This is because this instruction is primarily used when switching between active sandboxes. In this case, the permissions of all regions would likely need adjusting. If `hfi_set_region_permission` operated per region, then three calls to this instruction would be needed to adjust the permissions of the three regions. In contrast, our design allows this to occur in a single instruction. Additionally, the `permission_set_t` parameter further future-proofs this design by allowing us to modify the format of this instruction for future versions of HFI, without breaking backward compatibility.

Clearing Regions.. To clear region state rapidly e.g. on context switches, HFI offers `hfi_reset_regions`.

Behaviour. This instruction sets the base and bound/mask of all regions to zero (equivalent to calling `hfi_set_region_size` with arguments of 0 on all regions), disables all regions and sets their permissions to zero (equivalent to calling `hfi_set_region_permission` with a permission vector of 0). It also sets the active explicit data region to 1 (this will matter only in future HFI versions which have multiple explicit data regions).

Faults. This instruction faults if `hfi_usermode_enabled` and `lock_regions` is set to 1 (when hfi mode is on with region configurations locked).

Design Rationale. While this instruction can effectively be achieved using combinations of other instructions, unifying this "reset" operation into a single instruction allows optimizing a number of paths. For example, when an application needs to switch between multiple sandboxes, software has to clear the state of the first sandbox before applying the state of the second sandbox. This allows optimizing the first step sequence. This operation is also useful when the OS kernel is switching between two scheduled processes both of which may use HFI. The OS kernel is responsible for saving and restoring each processes' HFI state, and thus can also use this instruction.

6.3 Implementation Considerations

Implementation Semantics and Spectre To ensure Spectre safety, the following guidance is offered for implementer.

For code regions: To ensure security, prefix-checking should be carried out in parallel with the decode stage. If the check finds a matching region with execute permissions, it succeeds, and decode carries on normally. If the check fails, it prevents the decoded micro-ops from entering the pipeline, and instead

translates all instructions into a faulting NOP micro-op. This ensures that instructions that are out-of-bounds are not executed during committed execution, and are also not executed speculatively.

For data regions: Bounds checking, DTLB lookup, and cache index lookups should happen in parallel. One concern here is that, cache state could be modified as a result of secret (out-of-bounds) data. To prevent this sort of side-channel attack all bounds must checks occur *before* the processor resolves the physical address of a memory access. This is secure because the processor can update cache metadata like the LRU bits (for hits) or fetch new data blocks (for misses) only after resolving the physical address. HFI can therefore strictly prevent any metadata updates if there has been a fault.

Note that out-of-bounds address can affect metadata of the DTLB or i-cache — e.g., LRU bits. However, the invariant we guarantee — no secret (data stored outside the boundaries of the region) ever affects architectural state — is still not violated, since we do not allow the *result* of an out-of-bounds memory operation to propagate into any of these structure.

To summarize, HFI's data pipeline is Spectre safe, since the data cache is not updated prior to bounds checks being completed; HFI's control pipeline is safe as bounds checks finish prior to instruction decode which is before the execution of instructions. This approach also helps to guarantee that any code executed as the result of PHT, BTB, and RSB (speculative) predictions are checked prior to execution.

7 Using HFI

Here we explore how HFI's features are used to create sandboxing runtimes in userspace, and the HFI support needed from the operating systems.

7.1 Sandboxing in Userspace

Suppose we have a sandbox runtime (e.g. part of an application implementing a Wasm FaaS server, or a library sandboxing framework [7]), that is ready to create new sandbox. We assume that the runtime has reserved some memory for the sandbox (i.e., the sandbox memory), has placed the input for the sandboxed code in a memory buffer, and the sandboxed code itself is separated from other code and mapped in an isolated contiguous portion of the address space. Our runtime can now take the following steps:

Setting up regions. To start, our runtime sets up access to the code, heap, and input memory so our application has everything it needs once the sandbox starts, it does this using the `hfi_set_region_size`, `hfi_set_region_permission` instructions to setup and enable regions that grant access to the allocated heap and inputs, and the application code. If no code regions are mapped, HFI will immediately trap after `hfi_enter` is called, as the processor will not be able to fetch instructions.

What type of regions the runtime will use, as as how the sandbox is configured will depend on if it is sandboxing a native or SFI (e.g. Wasm) based application.

Sandboxing Native code. When sandboxing native code, the code being sandboxed is entirely untrusted, and thus, it cannot be allowed to modify any of HFI's state, or exit the sandbox in unexpected ways, or perform any other operation that would allow it to violate the sandbox's policy. Thus, the sandbox options flags in `hfi_options_t` that are passed to `hfi_enter` to start the sandbox will: set the `lock_regions` flag to 1 (true), since sandboxed code cannot be trusted to modify regions; set the `redirect_system_calls` flag to 1, as again this code can't be trusted to perform system calls directly; set the `redirect_exit` flag to 1, ensuring all control flow out of the sandbox is redirect to the trusted runtime. The runtime will use an implicit data region to permit the sandboxed code to use of the sandbox memory; the runtime will then ensure the sandboxed code's stack, heap and inputs are part of this implicit region. An implicit code regions will be used to mark the code of the sandbox.

Sandboxing using SFI runtimes. When sandboxing code through SFI runtimes such as a Wasm runtime, sandboxing is handled as a mix of compiler/software-runtime checks as well as hardware checks from HFI. In the scenario, implementer has confidence in the correctness of the SFI's runtime and compiler, and thus their use of HFI is different from sandboxing code, giving it greater flexibility and performance.

For example, on `hfi_enter` it can set the `lock_regions` flag to 0 (false), allowing the compiler/runtime to modify regions with `hfi_set_region_size` and `hfi_set_region_permission` without having to exit the sandbox. This can allow regions to be used more flexibly, e.g. it can load and spilling registers, and never

need to exit the sandbox. The `redirect_system_calls` flag can also be set to 0 (false), as the Wasm compiler disallows direct access to system call instructions, ensuring that any system calls made will come from the trusted runtime, this can eliminate the overhead of unecessary sandbox exits. What the redirect_exit flag will be set to depends on the SFI implementation, it may set this flag to false and opt to let sandbox exit's fall through to minimize overhead, since it can ensure that it knows that it can control whatever instruction follows an `hfi_exit`, or it may opt to set it to true, and setup an exit_handler.

For granting access to memory, the runtime will use still use implicit regions for code, however, an explicit region(s) will be used for the applications heap(s) and inputs, as the SFI compiler can use h-prefixed instructions for these accessing these directly. This allows it to exploit the greater flexibility of explicit regions for sizing and alignment that are necessary to support Wasm and similar systems. Notably, a Wasm runtime in the sandbox may opt to place it's own data into an implicit data region, to ensure Spectre attacks cannot be used to trick the sandbox runtime into leaking its own data.

Saving context. A sandboxing runtime must protect its own execution context such as its stack and contents of CPU registers, before it switches to sandbox code. HFI leaves this mechanism entirely up to software — this flexibility is important for efficiency. For example, if our runtime is running untrusted native code — it will have to use springboards and trampolines [10] — lightweight assembly routines that (1) clear registers and switch to a separate stack prior to executing the sandboxed code and (2) restore these registers after the sandboxed is executed. However, if it is running Wasm code, it could opt to use zero-cost transitions [4] that rely on the compiler to ensure that the sandbox code cannot misuse the stack or scratch registers.

Setting up an exit handler. If our runtime needs an exit handler either to handle `hfi_exit` or redirected system calls, it will need to setup an exit handler with `hfi_set_exit_handler`. This exit handler is implemented as a normal function call in the runtime that takes no arguments. It will query the `hfi_exit_reason` CSR to find out why it was invoked.

Entering the sandbox. Having taken all these steps, our runtime is ready to start the sandbox. Once it calls `hfi_enter`, HFI mode is enabled, and the next instruction that runs will be inside a sandbox.

The exit handler (`hfi_exit` and system calls). When the exit handler is called after the sandbox exits, it will transfer control to the exit handler function in the runtime, which will check a control and status register (CSR) to identify the cause of the exit, and respond appropriately.

For example, for sandbox exits, it will need to save context unless the sandbox execution has completed. Similarly for system calls, it will need to save context, but then also execute whatever additional logic is need to check the parameters of the system call for safety and finally invoking the system call [1].

7.2 OS (and VMM) integration

HFI is designed to require only minimal changes/support from the OS kernel. HFI requires support from OS kernels in two areas:

Handling HFI faults A fault may occur in HFI mode for to two reasons: (1) normal processor traps such as division by zero. (2) HFI policy violations. In both cases, when the processor traps into the kernel, HFI enforcement is disabled so as not to disrupt kernel execution, and the trap is handled through all the normal OS signal mechanisms (This is automatic as HFI hardware checks are only applied to userspace code). When an instruction can traps due to some HFI policy violation, the normal trap mechanism for the current hart is employed. HFI introduces a new trap code, `hfi_fault` to support this.

When an HFI trap occurs, additional details about the cause are stored in the `hfi_fault_status_reg`. A kernel trap handler can immediately read and store this state into the process struct for the current process, so it can be queried by a signal handler. The OS then invokes the standard signal handler registered by the application for memory access violations. The OS must invoke the signal handler with HFI disabled; if the signal handler returns control to the OS, the OS will re-enable HFI prior to resuming the faulting process.

Process scheduling. The kernel must save and restore HFI state (stored in internal registers) when switching between processes/VMs etc. This can be performed using the HFI manipulation instructions `hfi_reset_regions`, `hfi_set_region_permission`, `hfi_set_region_size`, `hfi_get_region_size`, etc. To

know whether HFI mode is currently enabled by the user space process, more privileged code can check the state of the `hfi_usermode_enabled` status register.

8 Instruction Encoding

With the exception of h-prefixed instructions (which we discuss separately), most HFI instructions such as transition instructions, region manipulation instructions have specific behavior and have little-to-no variation. Thus, to minimize opcode usage, these instructions can share a single opcode and simply be distinguished by the funct3 code. This optimization can be hidden from the end user by adding pseudo instructions to assemblers. Additionally some of these instructions can be implemented as pseudo instructions that read and writes to CSRs. To support backward compatible binaries, these instructions should be encoded using instructions that don't fault if not implemented.

H-prefixed instructions which are the HFI variants of memory instructions have very different characteristics. H-prefixed instructions have to mimic native instructions, including variants of the memory instructions (represented by funct3 codes); thus, these instructions need their own opcodes. Furthermore, these instructions should be encoded using op codes that fail if unimplemented; this is because the h-prefixed instructions implement new functionality—relative memory accesses—which makes them different from other HFI instructions which implement restrictions.

As a result of encoding choices, binaries that use HFI's implicit regions would remain backward compatible on CPUs that don't support HFI (albeit without the isolation enforcement). However by disease that use HFI's explicit region's would not be supported in this context.

9 Pending Design Considerations

Here, we include topics that we believe merit further discussion but which we have not fully resolved for inclusion in the specification.

9.1 Streamlining Control Transfers for Native Binaries

When sandboxing unmodified native binaries, we would ideally like control transfers into and out of the sandbox library to require minimal overhead and complexity. With a few small changes, we could make this simpler than what `hfi_enter` and `hfi_exit` offer today.

At present, control transfers require redirecting control flow through small stubs (trampolines) that need to be mapped by the sandbox runtime into the sandboxed library address space, this adds complexity and overhead.

For example, consider a case where a host application has uses a library sandboxed with HFI; the applications want to invoke a function `foo()` in the library To call `foo()`, it will need trampoline code — application code that performs a context switch by saving the current registers, switching the stack register to point to memory inside a region, enabling HFI and transferring control to `foo()`. This is mostly straightforward. However, once `foo` finishes executing (`foo` executes a return instruction), execution would attempt to return to the trampoline code — an operation that would fail as the trampoline code is not part of the sandbox code. Thus the host application, must perform an intermediate step — it must call `foo`, while modifying the return address on the stack to point to a stub within the sandbox, which invokes `hfi_exit` and then returns to the trampoline. We could eliminate the need for this stub by dedicating a bit in the return address (e.g. it's least significant bit, as this should be unused as instructions are at least 16 bit aligned) on the stack that indicates that if HFI is enabled, this return should simply invoke `hfi_exit`. Similar mechanism could also be applied to eliminating the need for trampolines for direct and indirect calls (i.e. callbacks) to host libraries.

9.2 HFI Versions (Profiles)

We aim to support multiple versions (profiles) for HFI to ensure we can support the myriad of RISC-V uses cases from embedded devices to server class CPUs. The two versions we currently aim to support are a

minimal profile aka version 0, and a standard profile aka version 1. The key difference between these two is the number of supported regions. Versions may also be used to incorporate additional features in the future. Software can check what version is supported as described in section 3.

The minimal profile described in this document supports: 1 implicit code region, 1 implicit data region and 1 explicit data region. The additional standard profile supports: 2 implicit code regions, 4 implicit data regions, and 4 explicit data regions. These regions are numbered as follows: explicit data region 1, implicit data region 1 and implicit code region 1 are regions 1, 2, and 3 in both profiles. In the standard profile: explicit data region 2, 3, and 4 are regions 4, 5, and 6 respectively; implicit data region 2, 3, and 4 are regions 7, 8, and 9 respectively; implicit code region 2 is region 10. Additionally the permissions bit vector operand specified in the `hfi_set_region_permission` and `hfi_get_region_permission` instructions is also expanded to accommodate region permissions in the same order.

Design Rationale. Regions exact some cost in terms of circuit area, and differing trade-offs may make sense for different use cases. Obviously more regions facilitate efficient access to more data and code concurrently, and can simplify runtime implementation.

While the minimal profile is limited in the number of regions, this can still offer meaningful benefits for certain use cases without significant concurrency or memory sharing. For example, the Google Chrome browser's Ubercage JIT isolation scheme [8] would be able to leverage this minimal profile for its isolation requirements [6].

The standard profile offers an expanded the number of regions, the particular number of regions was inspired by uses cases such as leveraging WebAssembly for efficient isolation of libraries from applications, and efficient isolation of different clients' code in serverless settings [5]. The main observation here is that there is greater concurrency and sharing is present than simple use cases, but this can nevertheless be handled efficiently with a handful of regions. For uses cases that need additional regions, this can be achieved by spilling and restoring regions similar to how this is done with general-purpose registers.

Two additional instruction are required to support the standard profile with it's multiple explicit regions.

```
hfi_set_curr_explicit_data_region(region_number_t)
hfi_get_curr_explicit_data_region() -> region_number_t
```

These instructions are used to set which explicit region the h-prefixed instructions will utilize.

9.3 HFI in M-mode or S-mode

We plan to add support for HFI in S-mode. Relatively small changes are necessary, however, we have not yet done a full analysis of how privileged instructions are handled. HFI support in m-mode may similarly be possible, but requires additional analysis for hardware implementation details.

```
// Bit vector of sandbox configurations
hfi_options_t: reg_u8
  lock_regions: reg_u1                    // Restrict modifying of regions by sandboxed code.
  redirect_system_calls: reg_u1           // Redirect syscall to the exit_handler.
  redirect_exits: reg_u1                  // Redirect hfi_exit to the exit_handler.
  serialized_enter_exit: reg_u1           // Serialize enter/exit for Spectre protections.

// 1 -> explicit_data, 2 -> implicit_data, 3 -> implicit_code
region_number_t: reg_u32

// Bit vector of region permissions.
permission_t : reg_u8
  r1_enabled : reg_u1, r1_read:  reg_u1, r1_write: reg_u1, r1_is_large: reg_u1, // explicit data region 1
  r2_enabled : reg_u1, r2_read:  reg_u1, r2_write: reg_u1,                     // implicit data region 1
  r3_enabled : reg_u1, r3_exec:  reg_u1,                                      // implicit code region 1

permission_set_t: reg_u32 // For future use. Fixed to 1.

// HFI mode transition instructions
hfi_enter(hfi_options_t)                  // Enter a sandbox with params.
hfi_enter(hfi_options_t, target: reg_u64) // Enter a sandbox with params.
hfi_exit()                                 // Exit the sandbox.

hfi_set_exit_handler(exit_handler_t: reg_u64)       // Set the exit handler.
hfi_get_exit_handler() -> exit_handler_t: reg_u64   // Get the exit handler.

// For explicit data regions set/get the {base_address, bound} acc. to the restrictions
// - base, bound are multiples of 64k for large regions
// - region shouldn't span a 4GiB boundary for small regions
// For implicit regions set/get a {base_prefix, lsb_mask}
hfi_set_region_size(region_number_t, reg_u64 base, reg_u64 mask_or_bound)
hfi_get_region_size(region_number_t) -> (reg_u64 base, reg_u64 mask_or_bound)

// Set/Get region permissions of all regions. permission_set_t is a register for future use, currently it must
↪  be 0.
hfi_set_region_permissions(permission_set_t, permission_t)
hfi_get_region_permission(permission_set_t) -> permission_t

// Set all regions' (size, bound, permissions) and the active explicit data region to 1.
hfi_reset_regions()

// Operations on the active explicit data region. Encoded like standard load, stores.
hlw(...) -> ..., hlh(...) -> ..., hlhu(...) -> ..., hlb(...) -> ..., hlbu(...) -> ..., hli(...) -> ...
hsw(...), hsh(...), hshu(...), hsb(...), hsbu(...), hsi(...)
```

Figure 1: The HFI interface. The functions represent HFI instructions, while the structures represent the parameters passed in, or values returned by the HFI instructions via general purpose registers.

```
hfi_status_t: reg_u64          // Configuration and status register for HFI. Read-only in userspace.
  hfi_usermode_enabled: reg_u1 // Whether HFI mode is enabled. Set to 1 by hfi_enter, and 0 by hfi_exit.
  hfi_exit_reason: reg_u2      // Reason for the last exit (1=> due to hfi_exit inst., 2=> due to system call).
  hfi_exit_pc: reg_u60         // PC[62...2] of the last redirected syscall/hfi_exit. Other bits assumed 0.

implicit_code_region_t:
  base_prefix: reg_u64 // base address prefix
  lsb_mask: reg_u64 // mask for address suffix
  permission_exec: reg_u1 // execute permission

implicit_data_region_t:
  base_prefix: reg_u64 // base address prefix
  lsb_mask: reg_u64 // mask for address suffix
  permission_read: reg_u1 // read permission
  permission_write: reg_u1 // write permission

explicit_data_region_t:
  base_address: reg_u64
  bound: reg_u64
  permission_read: reg_u1 // read permission
  permission_write: reg_u1 // write permission
  is_large_region: reg_u1 // use large/small region. Large regions: base, bound are multiples of 64k. Small
  ↪  regions: region shouldn't span a 4GiB boundary

hfi_fault_t:
    hfi_fault_occured: reg_u1 // true (1) | false (0)
    hfi_fault_region: reg_u8  // implicit region not matched => 0, explicit region fault => region number
    hfi_fault_op: reg_u2 // LOAD_FAULT| STORE_FAULT| FETCH_FAULT
    hfi_fault_type: reg_u1 // OUT_OF_BOUNDS | INSUFFICIENT_PERMISSIONS

// Read-only Control and Status Registers.
hfi_status_reg : hfi_status_t // Bit vector of HFI run status

// Read-write HFI fault register
hfi_fault_status_reg : hfi_fault_t

//Internal registers
hfi_exit_handler_reg : reg_u64 // Bit vector of HFI run status

hfi_implicit_code_region : implicit_code_region_t[1]
hfi_implicit_data_region : implicit_data_region_t[1]
hfi_explicit_data_region : explicit_data_region_t[1]
```

Figure 2: An HFI implementation's control and status registers. The types show the sizes of registers, while the variables show the number of registers needed.

References

[1] R. J. Connor, T. McDaniel, J. M. Smith, and M. Schuchard. PKU pitfalls: Attacks on PKU-based memory isolation systems. In *USENIX Security*, 2020.

[2] A. Haas, A. Rossberg, D. L. Schuff, B. L. Titzer, M. Holman, D. Gohman, L. Wagner, A. Zakai, and J. Bastien. Bringing the web up to speed with WebAssembly. In *PLDI*. ACM, 2017.

[3] R. Jones. Risc-v extensions: what's available and how to find them. `https://research.redhat.com/blog/article/risc-v-extensions-whats-available-and-how-to-find-it/`, Nov. 2023.

[4] M. Kolosick, S. Narayan, C. Watt, M. LeMay, D. Garg, R. Jhala, and D. Stefan. Isolation without taxation: Near zero cost transitions for sfi. In *Proceedings of the ACM SIGPLAN Symposium on Principles of Programming Languages (POPL)*, 2022.

[5] S. Narayan, T. Garfinkel, M. Taram, J. Rudek, D. Moghimi, E. Johnson, C. Fallin, A. Vahldiek-Oberwagner, M. LeMay, R. Sahita, et al. Going beyond the limits of sfi: Flexible and secure hardware-assisted in-process isolation with hfi. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*, pages 266–281, 2023.

[6] S. Groß (saelo). V8 sandbox - hardware support. `https://docs.google.com/document/d/12MsaG6BYRB-jQWNkZiuM3bY8X2B2cAsCMLLdgErvK4c/`, 2024.

[7] The rlbox sandboxing toolkit. `https://rlbox.dev`, 2024.

[8] saelo. V8 sandbox aka. "ubercage". `https://docs.google.com/document/d/1FM4fQmIhEqPG8uGp5o9A-mnPB5BOeScZYpkHjo0KKA8/edit`, 2021.

[9] R. Wahbe, S. Lucco, T. E. Anderson, and S. L. Graham. Efficient software-based fault isolation. In *SOSP*. ACM, 1993.

[10] B. Yee, D. Sehr, G. Dardyk, J. B. Chen, R. Muth, T. Ormandy, S. Okasaka, N. Narula, and N. Fullagar. Native Client: A sandbox for portable, untrusted x86 native code. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, 2009.