NOVA Microhypervisor Interface Specification

Udo Steinberg
Technische Universität Dresden
Operating Systems Group
udo@hypervisor.org

July 13, 2011

Contents

Part I

Introduction

1 System Architecture

The **NOVA OS V**irtualization **A**rchitecture facilitates the coexistence of multiple legacy guest operating systems and a multi-server user environment on a single platform. The core system leverages virtualization technology provided by recent x86 platforms and comprises the Microhypervisor and one or more Virtual-Machine Monitors (VMMs).

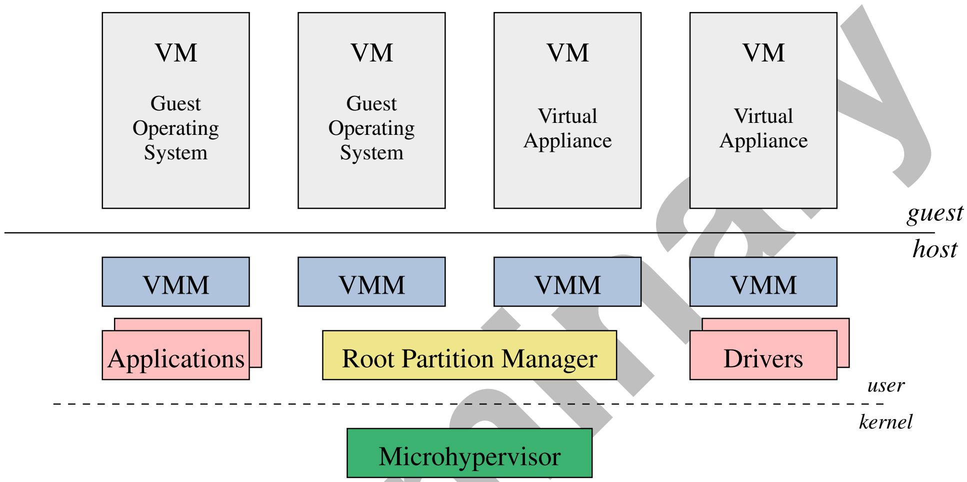

Figure 1.1: System Architecture

Figure 1.1 shows the structure of the system. The microhypervisor is the only component running in privileged root/kernel mode. It isolates the user-level servers, including the virtual-machine monitor, from one another by placing them in different address spaces in unprivileged root/user mode. Each legacy guest operating system runs in its own virtual-machine environment in non-root mode and is therefore isolated from the other components.

Besides isolation, the microhypervisor also provides mechanisms for partitioning and delegation of platform resources, such as CPU time, physical memory, I/O ports and hardware interrupts and for establishing communication paths between different protection domains.

The virtual-machine monitor handles virtualization faults and implements virtual devices that enable legacy guest operating systems to function in the same manner as they would on bare hardware. Providing this functionality outside the microhypervisor in the VMM considerably reduces the size of the trusted computing base for the multi-server user environment and for applications that do not require virtualization support.

The architecture and interfaces of the VMM and the multi-server user environment are not described in this document.

Part II

Basic Abstractions

2 Kernel Objects

2.1 Protection Domain

1. The Protection Domain (PD) is a unit of protection and isolation.
2. A protection domain is referenced by a Protection Domain Capability ($\text{CAP}_{\text{OBJ}_{\text{PD}}}$).
3. A protection domain is composed of a set of spaces that hold capabilities to platform resources or kernel objects that can be accessed by execution contexts within the protection domain. The following spaces are currently defined:
   - Memory Space
   - I/O Space
   - Object Space
4. The memory space of a protection domain holds capabilities to page frames in physical memory.
5. The I/O space of a protection domain holds capabilities to I/O ports.
6. The object space of a protection domain holds capabilities to the following kernel objects:
   - Protection Domain (PD)
   - Execution Context (EC)
   - Scheduling Context (SC)
   - Portal (PT)
   - Semaphore (SM)

2.2 Execution Context

1. The Execution Context (EC) is an abstraction for an activity within a protection domain.
2. An execution context is referenced by an Execution Context Capability ($\text{CAP}_{\text{OBJ}_{\text{EC}}}$).
3. An execution context is permanently bound to the protection domain in which it was created.
4. An execution context may optional have a scheduling context bound to it.
5. There exist two flavors of execution context:
   - Kernel thread
   - Virtual CPU
6. An execution context comprises the following information:
   - Reference to protection domain (2.1)
   - Event Selector Base (SEL_{EVT}) (3.3)
   - Reply capability register (4.1)
   - User Thread Control Block (UTCB) (4.6)
   - Central Processing Unit (CPU) registers (architecture dependent)
   - Floating Point Unit (FPU) registers (architecture dependent)

2.3 Scheduling Context

1. The Scheduling Context (SC) is a unit of dispatching and prioritization.
2. A scheduling context is referenced by a Scheduling Context Capability ($\mathrm{CAP}_{\mathrm{OBJ}_{\mathrm{SC}}}$).
3. A scheduling context is permanently bound to exactly one physical CPU.
4. At any point in time, a scheduling context is bound to exactly one execution context.
5. Donation of a scheduling context to another execution context binds the scheduling context to that other execution context.
6. A scheduling context comprises the following information:
   - Reference to execution context (2.2)
   - Time quantum
   - Priority

2.4 Portal

1. A Portal (PT) represents a dedicated entry point into the protection domain in which the portal was created.
2. A portal is referenced by a Portal Capability ($\mathrm{CAP}_{\mathrm{OBJ}_{\mathrm{PT}}}$).
3. A portal is permanently bound to exactly one execution context.
4. A portal comprises the following information:
   - Reference to execution context (2.2)
   - Message Transfer Descriptor (MTD) (4.4)
   - Entry instruction pointer
   - Portal identifier

2.5 Semaphore

1. A Semaphore (SM) provides a means to synchronize execution and interrupt delivery by selectively blocking and unblocking execution contexts.
2. A semaphore is referenced by a Semaphore Capability ($\mathrm{CAP}_{\mathrm{OBJ}_{\mathrm{SM}}}$).

3 Mechanisms

3.1 Scheduling

The microhypervisor implements a round-robin scheduler with multiple priority levels. Whenever an execution context is ready to execute, the runqueue contains all scheduling contexts bound to that execution context. When an execution context blocks, the microhypervisor removes the corresponding scheduling contexts from the runqueue.

When the microhypervisor needs to make a scheduling decision, it selects the highest-priority scheduling context from the runqueue and dispatches the execution context bound to that scheduling context.

The parameters of a scheduling context influence the scheduling behavior of the system as follows:

- The priority defines the importance of a scheduling context. A higher-priority scheduling context always has precedence and immediately preempts a lower-priority scheduling context.
- The time quantum defines the number of microseconds that the execution context, which is currently bound to the scheduling context, can utilize the CPU when it is dispatched. A dispatched execution context consumes the time quantum of its scheduling context until the quantum reaches zero; at that point the microhypervisor preempts the execution context, replenishes the time quantum of the scheduling context, and makes a scheduling decision.

3.2 Communication

Message passing between protection domains is governed by portals. A portal represents a dedicated entry point into the protection domain to which the portal is bound. An execution context in a protection domain can call any portal for which the protection domain holds a capability. Portal capabilities can be delegated in order to establish cross-domain communication channels.

To initiate a message-passing operation from one protection domain to another, the caller execution context passes a portal capability selector $\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{PT}}}$ to the microhypervisor. The microhypervisor uses the capability selector to look up the portal capability $\mathrm{CAP}_{\mathrm{OBJ}_{\mathrm{PT}}}$ in the object space of the caller protection domain. If the lookup succeeds, the microhypervisor loads the destination protection domain and entry instruction pointer for that domain from the portal.

An arbitrary number of portals can be bound to a callee execution context in a protection domain. The callee provides the stack for handling one incoming request on any of these portals. If the callee is busy handling another request, and both caller and callee are on the same CPU, the caller may optionally lend its scheduling context to the callee to help it run the previous request to completion.

Once the callee is available to handle a new request and a caller exists for any portal bound to the callee, the microhypervisor arranges a rendezvous and transfers the message from the UTCB of the caller to the UTCB of the callee.

If the request established a reply capability for the callee, the callee may subsequently respond directly to the caller through a reply operation without risking to block, because the caller is already waiting for the response.

The following forms of message passing are currently supported:

Nondonating Call

During a nondonating `call`, the caller execution context traverses the destination portal, rendezvouses with a callee execution context and transfers a message to it. The microhypervisor establishes a reply capability for the callee. The caller blocks on the instruction following the hypercall and does *not* donate the current scheduling context to the callee. The callee may later invoke the reply capability to send a response directly to the blocked caller. Upon receiving the response the caller becomes unblocked.

Donating Call

A donating `call` differs from a nondonating call in that the caller donates the current scheduling context to the callee. The donation mechanism implements priority and bandwidth inheritance from the caller to the callee. The caller blocks on the instruction following the hypercall and the callee starts executing immediately. The microhypervisor also establishes a reply capability for the callee. The callee may later invoke the reply capability to send a response directly to the blocked caller. Upon receiving the response the caller becomes unblocked.

Reply

The `reply` operation sends a message back to the caller identified by the reply capability and revokes the reply capability. If the reply capability was established by a donating call, the microhypervisor returns the previously donated scheduling context back to the caller. The callee blocks until the next request arrives.

3.3 Exceptions and Intercepts

When an execution context triggers a hardware exception or VM intercept, the microhypervisor adds the exception number or intercept reason to the Event Selector Base (SEL_{EVT}) of the affected EC. If the resulting capability selector refers to a portal capability $\text{CAP}_{\text{OBJ}_{\text{PT}}}$, the microhypervisor arranges an implicit donating `call` for the execution context through the corresponding portal; otherwise the execution context is shut down.

The entire handling of the exception or intercept is performed using the current scheduling context of the execution context that triggered the event. Furthermore, that execution context remains blocked until the handler has replied with a message to resolve the exception or intercept.

The number of capability selectors used for exception and intercept handling is conveyed in the Hypervisor Information Page (HIP) (6.2). The translation of hardware exception numbers and intercept reasons to capability selectors is described in the processor-specific Application Binary Interface (ABI) (IV).

3.4 Interrupts

The microhypervisor provides a semaphore per Global System Interrupt (GSI). An execution context waits for an interrupt by performing an `sm_ctrl[down]` hypercall to block on the corresponding semaphore. When the interrupt occurs, the microhypervisor issues an `sm_ctrl[up]` operation for the semaphore.

3.5 Capability Delegation

Delegation of capabilities from one protection domain to another is performed during communication. The execution context that sends a message puts typed items in its UTCB, specifying which range of capabilites from the sender's protection domain it wants to delegate to the receiver's protection domain. The receiver specifies in its UTCB, which range of capabilities it is willing to accept and where they should be installed in the receiver's protection domain.

The microhypervisor computes the intersection of the sender and receiver ranges and delegates only those capabilities that are covered by both ranges. The sender may optionally reduce the permissions of the delegated capabilities for the receiver, using the mask field in the Capability Range Descriptor (CRD).

If the capability ranges of the sender and receiver differ in size, the capability hotspot, specified by the sender, is used for disambiguation as illustrated in Figure 3.1.

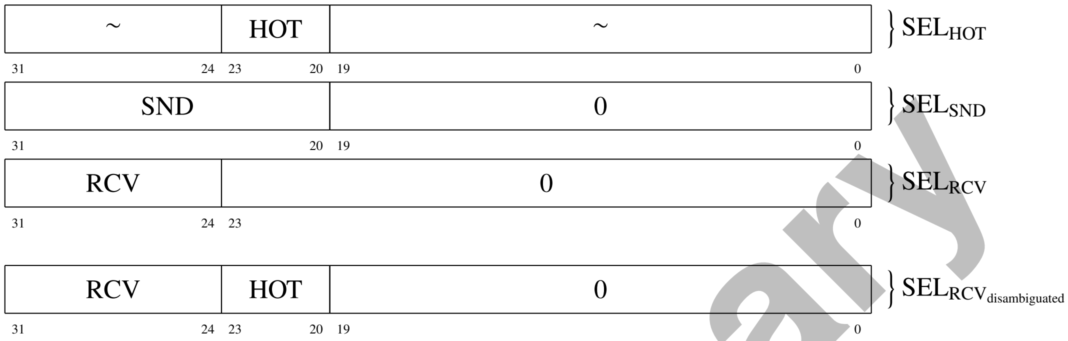

Figure 3.1: Capability Range Disambiguation

In this example, the sender has specified a capability range of order 20, starting at SEL_{SND}, whereas the receiver has specified a capability range of order 24, starting at SEL_{RCV}. There exist 2^4 possible locations in the receiver range, where the sender range could be delegated. Whenever two capability ranges differ in size, the microhypervisor truncates the larger range by taking the ambiguous bits from the capability hotspot.

3.6 Capability Revocation

Accepting a capability delegation constitutes an implicit agreement that the capabilities may be revoked again at any time without the receiver's consent. Revoking a range of capabilities from a protection domain additionally revokes that range from all protection domains that directly or indirectly inherited it from that protection domain.

Part III

Application Programming Interface

4 Data Types

4.1 Capability

A Capability (CAP) is a reference to a kernel object plus associated auxiliary data, such as access permissions. Capabilities are opaque and immutable to the user — they cannot be inspected, modified or addressed directly; instead user programs access a capability via a capability selector (4.2). All capabilities can be delegated and revoked as described in Section 3.5. The following types of capabilities exist:

4.1.1 Null Capability

A Null Capability (CAP_0) does not reference anything and there are no permissions defined.

4.1.2 Memory Capability

A Memory Capability (CAP_{MEM}) references a 4KB page frame. It is stored in the memory space of a protection domain. The capability permissions are defined as follows:

1	1	x	w	r
4	3	2	1	0

r readable if set.

w writable if set.

x executable if set.

4.1.3 I/O Capability

An I/O Capability ($CAP_{I/O}$) references an I/O port. It is stored in the I/O space of a protection domain. The capability permissions are defined as follows:

1	1	1	1	a
4	3	2	1	0

a accessible if set.

4.1.4 Object Capability

An Object Capability (CAP_{OBJ}) references a kernel object. It is stored in the object space of a protection domain. The following types of object capabilities are currently defined:

4.1.4.1 Protection Domain Capability

A Protection Domain Capability ($CAP_{OBJ_{PD}}$) references a protection domain (2.1). The capability permissions are defined as follows:

sm	pt	sc	ec	pd
4	3	2	1	0

pd Hypercall `create_pd` (5.3.1) permitted if set.

ec Hypercall `create_ec` (5.3.2) permitted if set.

sc Hypercall `create_sc` (5.3.3) permitted if set.

pt Hypercall `create_pt` (5.3.4) permitted if set.

sm Hypercall `create_sm` (5.3.5) permitted if set.

4.1.4.2 Execution Context Capability

An Execution Context Capability ($CAP_{OBJ_{EC}}$) references an execution context (2.2). The capability permissions are defined as follows:

1	pt	sc	1	ct
4	3	2	1	0

ct Hypercall `ec_ctrl` (5.4.1) permitted if set.

sc Hypercall `create_sc` (5.3.3) can bind a scheduling context if set.

pt Hypercall `create_pt` (5.3.4) can bind a portal if set.

4.1.4.3 Scheduling Context Capability

A Scheduling Context Capability ($CAP_{OBJ_{SC}}$) references a scheduling context (2.3). The capability permissions are defined as follows:

1	1	1	1	ct
4	3	2	1	0

ct Hypercall `sc_ctrl` (5.4.2) permitted if set.

4.1.4.4 Portal Capability

A Portal Capability ($CAP_{OBJ_{PT}}$) references a portal (2.4). The capability permissions are defined as follows:

1	1	1	1	1
4	3	2	1	0

4.1.4.5 Semaphore Capability

A Semaphore Capability ($\mathrm{CAP}_{\mathrm{OBJ}_{\mathrm{SM}}}$) references a semaphore (2.5). The capability permissions are defined as follows:

1	1	1	dn	up
4	3	2	1	0

up Hypercall `sm_ctrl[up]` (5.4.3) permitted if set.

dn Hypercall `sm_ctrl[down]` (5.4.3) permitted if set.

4.1.5 Reply Capability

A Reply Capability ($\mathrm{CAP}_{\mathrm{RP}}$) references a caller execution context. It is stored in the reply register of an execution context during communication and automatically destroyed when invoked.

4.2 Capability Selector

A Capability Selector (SEL) is a user-visible abstract key for accessing a capability. The capability selector serves as integer index for the memory space, I/O space or object space of a protection domain. All capability selectors that do not refer to capabilities of another type refer to a null capability. For example, in Figure 4.1 capability selector 2 refers to a capability for an execution context.

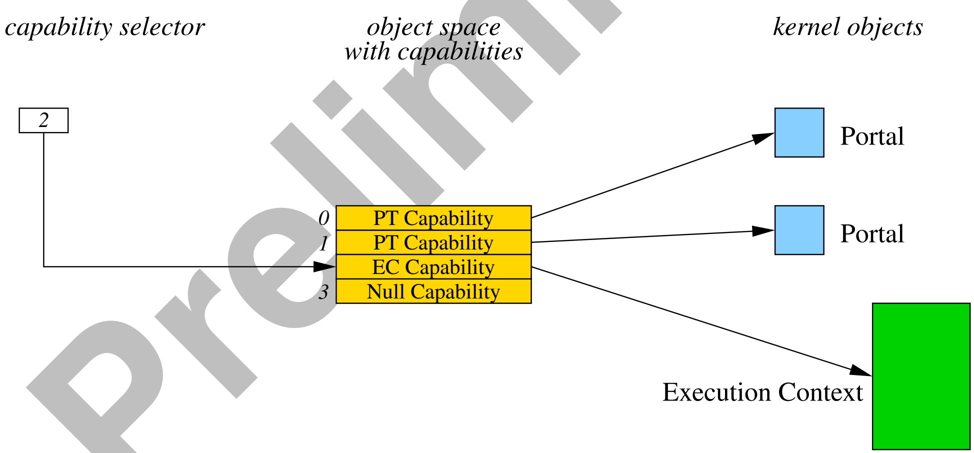

Figure 4.1: Capability Selector

4.3 Capability Range Descriptor

A Capability Range Descriptor (CRD) refers to all capabilities of a particular type in the selector range SEL … $\text{SEL} + 2^{\text{Order}} - 1$. It must be naturally aligned such that $\text{SEL} \equiv 0 \pmod{2^{\text{Order}}}$. During capability delegation, the permissions of the destination capability are computed as the logical AND of the permissions of the source capability and the permission mask from the capability range descriptor.

4.3.1 Null Capability Range Descriptor

A Null Capability Range Descriptor (CRD_0) does not refer to any capabilities.

~	0
31 … 2	1 0

4.3.2 Memory Capability Range Descriptor

A Memory Capability Range Descriptor (CRD_{MEM}) refers to the memory capabilities located within the specified selector range of the memory space. Each memory capability covers 2^{12} bytes of memory.

SEL_{MEM}	Order	Mask	1
31 … 12	11 … 7	6 … 2	1 0

4.3.3 I/O Capability Range Descriptor

An I/O Capability Range Descriptor ($\text{CRD}_{\text{I/O}}$) refers to the I/O capabilities located within the specified selector range of the I/O space.

$\text{SEL}_{\text{I/O}}$	Order	Mask	2
31 … 12	11 … 7	6 … 2	1 0

4.3.4 Object Capability Range Descriptor

An Object Capability Range Descriptor (CRD_{OBJ}) refers to the object capabilities located within the specified selector range of the object space.

SEL_{OBJ}	Order	Mask	3
31 … 12	11 … 7	6 … 2	1 0

4.4 Message Transfer Descriptor

The Message Transfer Descriptor (MTD) is an architecture-specific bitfield that controls the contents of an exception or intercept message. The MTD is provided by the portal associated with the event and conveyed to the receiver as part of the exception or intercept message.

For each bit set to 1, the microhypervisor transfers the processor state associated with that bit to/from the respective fields of the UTCB data area. The layout of the MTD and the fields in the UTCB data area are described in the processor-specific ABI (IV).

4.5 Quantum Priority Descriptor

The Quantum Priority Descriptor (QPD) specifies the priority of a scheduling context and its time quantum in microseconds. It has the following format:

Time Quantum	0	Priority
31 … 12	11 … 8	7 … 0

Figure 4.2: Quantum Priority Descriptor

4.6 User Thread Control Block

Each execution context that acts as a kernel thread has an associated User Thread Control Block (UTCB), which consists of a header area and a data area as illustrated in Figure 4.3.

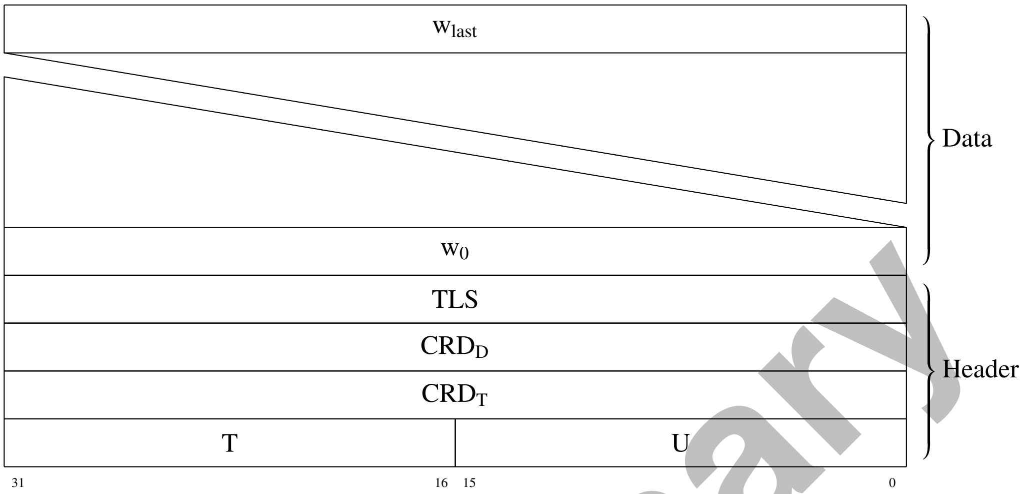

Figure 4.3: User Thread Control Block: General Layout

4.6.1 Header Area

The UTCB header fields are defined as follows:

U

Number of untyped items.

T

Number of typed items.

CRD_T

This capability range descriptor (4.3) specifies a receive window in the memory, I/O, or object space, in which the microhypervisor is allowed to perform capability translations. A null capability range descriptor effectively disables capability translations.

CRD_D

This capability range descriptor (4.3) specifies a receive window in the memory, I/O, or object space, in which the execution context is willing to accept capability delegations. A null capability range descriptor effectively disables capability delegations.

TLS

This field is never written by the microhypervisor and can be used to store thread-local data.

4.6.2 Data Area

The size of the data area is defined by the size of the UTCB minus the size of the header area. An execution context uses its UTCB to send or receive messages, and to transfer typed items during capability delegation. The U and T fields in the UTCB header area define the number of untyped and typed items.

4.6.2.1 Untyped Items

The microhypervisor transfers untyped items from the beginning of the UTCB data area upwards. Each untyped item occupies one word as illustrated in Figure 4.4 For example, during a transfer of u untyped items, the microhypervisor copies words $w_0 \dots w_{u-1}$ from the UTCB data area of the sender to words $w_0 \dots w_{u-1}$ in the UTCB data area of the receiver, without interpreting the contents of these words.

Data		
31	0	

Data	w_0

Figure 4.4: User Thread Control Block: Untyped Item

4.6.2.2 Typed Items

The microhypervisor transfers typed items from the end of the UTCB data area downwards. Each typed item occupies two words. For example, during a transfer of t typed items, the microhypervisor interprets words $w_{last} \dots w_{last-2t+1}$ of the sender's UTCB data area. For each typed item in the sender UTCB, the microhypervisor creates a corresponding typed item in the receiver UTCB. The following typed items are currently defined:

Translate:

31 … 1	0	
~	0	w_{last}
CRD		w_{last-1}

Figure 4.5: User Thread Control Block: Translate Item

If the type of the sender's CRD does not match the type of the receive window CRD_T in the receiver's UTCB header, the receiver obtains a typed item with a null capability range descriptor.

Otherwise, the microhypervisor attempts to translate the capability range specified by the base address and order in the sender protection domain to the corresponding capability range in the receiver protection domain from which it had been originally delegated. If the translation fails, e.g., because the sender range is not derived from the receiver range, the receiver obtains a typed item with a null capability range descriptor. Otherwise the capability range descriptor describes the corresponding range in the receiver and the sender permissions for that range.

Delegate:

31 … 12	11	10	9	8 … 1	0	
SEL_{HOT}	H	G	D	0	1	w_{last}
CRD						w_{last-1}

Figure 4.6: User Thread Control Block: Delegate Item

If the type of the sender's CRD does not match the type of the receive window CRD_D in the receiver's UTCB header, the receiver obtains a typed item with a null capability range descriptor.

Otherwise, the microhypervisor computes the range of capabilities to delegate from the sender to the receiver, using the hotspot SEL_{HOT} for range disambiguation, as described in Section 3.5. The capability range descriptor in the receiver's typed item describes the contents of the receive window.

The root protection domain can control the source of a capability delegation as follows. For other protection domains this bit is ignored.

H If the bit is set, the source is the microhypervisor. Otherwise the source is the protection domain itself.

For memory capability range descriptors (4.3.2), the following bits control which page tables are updated in addition to the host page table. For other capability range descriptors, these bits are ignored.

G The guest page table is updated if the bit is set.

D The DMA page table is updated if the bit is set.

5 Hypercalls

5.1 Definitions

Hypercall Numbers

Each hypercall is identified by a unique number. Figure 5.1 lists the currently defined hypercalls.

Number	Hypercall	Section
0x0	CALL	5.2.1
0x1	REPLY	5.2.2
0x2	CREATE_PD	5.3.1
0x3	CREATE_EC	5.3.2
0x4	CREATE_SC	5.3.3
0x5	CREATE_PT	5.3.4
0x6	CREATE_SM	5.3.5
0x7	REVOKE	5.3.6
0x8	LOOKUP	5.3.7
0x9	EC_CTRL	5.4.1
0xa	SC_CTRL	5.4.2
0xb	SM_CTRL	5.4.3
0xc	ASSIGN_PCI	5.5.1
0xd	ASSIGN_GSI	5.5.2

Figure 5.1: Hypercall Numbers

Status Codes

Figure 5.2 shows the status codes returned to indicate success or failure of a hypercall.

Number	Status Code	Description
0x0	SUCCESS	Successful Operation
0x1	COM_TIM	Communication Timeout
0x2	COM_ABT	Communication Abort
0x3	BAD_HYP	Invalid Hypercall
0x4	BAD_CAP	Invalid Capability
0x5	BAD_PAR	Invalid Parameter
0x6	BAD_FTR	Invalid Feature
0x7	BAD_CPU	Invalid CPU Number
0x8	BAD_DEV	Invalid Device ID

Figure 5.2: Status Codes

5.2 Communication

5.2.1 Call

Synopsis:

```
status = call (SEL_OBJ_PT);
```

Parameters:

$\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{PT}}}$: Target Portal

Flags:

0		DD	DB
3	2	1	0

DB Disable Blocking (0=blocking, 1=nonblocking)

DD Disable Donation (0=dcall, 1=ncall)

Description:

1. If the execution context (2.2), to which the target portal referenced by $\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{PT}}}$ is bound, is busy, the microhypervisor considers the 'disable blocking' flag. If the flag is set, the hypercall returns with a timeout. Otherwise the caller blocks until the callee execution context becomes available.
2. The microhypervisor transfers a message, whose contents is determined by the UTCB, from the caller to the callee.
3. The microhypervisor establishes a reply capability (4.1) in the reply register of the callee. The caller blocks until the callee invokes the reply capability. If the 'disable donation' flag is clear, the current scheduling context, previously bound to the caller, is donated and thereby bound to the callee.

Status:

SUCCESS

Hypercall completed successfully.

COM_TIM

Rendezvous with the callee execution context timed out.

COM_ABT

Operation aborted during execution of the callee execution context.

BAD_CAP

$\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{PT}}}$ did not refer to a Portal Capability ($\mathrm{CAP}_{\mathrm{OBJ}_{\mathrm{PT}}}$).

BAD_CPU

Caller execution context and callee execution context are on different CPUs.

5.2.2 Reply

Synopsis:

```
pid = reply();
```

Description:

1. If the reply register contains a reply capability, the microhypervisor transfers a message, whose contents is determined by the UTCB, to the caller execution context referenced by the reply capability.
2. If the caller had donated its scheduling context to the callee, the microhypervisor binds that scheduling context back to the caller, thereby terminating the donation.
3. The microhypervisor destroys the reply capability by replacing it with a null capability $\mathrm{CAP}_{\emptyset}$.
4. The callee blocks until a subsequent request arrives.

Status:

This hypercall does not return. Instead, when one of the portals bound to the execution context is called, the microhypervisor passes the portal identifier to the execution context, and execution continues at the instruction pointer specified in the called portal.

5.3 Capability Management

5.3.1 Create Protection Domain

Synopsis:

`status = create_pd (`$\text{SEL}_{\text{OBJ}_0}$, $\text{SEL}_{\text{OBJ}_{\text{PD}}}$, CRD_{OBJ}`);`

Parameters:

$\text{SEL}_{\text{OBJ}_0}$: Created PD

$\text{SEL}_{\text{OBJ}_{\text{PD}}}$: Owner PD

CRD_{OBJ}: Initial Portals

Description:

Creates a new protection domain, accounted to the PD specified by $\text{SEL}_{\text{OBJ}_{\text{PD}}}$. Prior to the hypercall, $\text{SEL}_{\text{OBJ}_0}$ must refer to a null capability, and $\text{SEL}_{\text{OBJ}_{\text{PD}}}$ must refer to a protection domain capability with permission bit CAP_{PD} set. The caller PD obtains in place of $\text{SEL}_{\text{OBJ}_0}$ a protection domain capability that refers to the created PD. The microhypervisor delegates the capability range, specified by CRD_{OBJ}, from the caller PD to the created PD.

Status:

SUCCESS

Hypercall completed successfully.

BAD_CAP

$\text{SEL}_{\text{OBJ}_0}$ did not refer to a Null Capability (CAP_0).
$\text{SEL}_{\text{OBJ}_{\text{PD}}}$ did not refer to a Protection Domain Capability ($\text{CAP}_{\text{OBJ}_{\text{PD}}}$).
$\text{CAP}_{\text{OBJ}_{\text{PD}}}$ has insufficient permissions.

5.3.2 Create Execution Context

Synopsis:

`status = create_ec (`$\text{SEL}_{\text{OBJ}_0}$`, `$\text{SEL}_{\text{OBJ}_{\text{PD}}}$`, CPU, UTCB, SP, `$\text{SEL}_{\text{EVT}}$`);`

Parameters:

$\text{SEL}_{\text{OBJ}_0}$: Created EC

$\text{SEL}_{\text{OBJ}_{\text{PD}}}$: Owner PD

CPU: CPU Number

UTCB: Virtual Address: UTCB Pointer

SP: Virtual Address: Stack Pointer

SEL_{EVT}: Event Selector Base

Flags:

0	G
3 … 1	0

G Global Thread (0=local, 1=global)

Description:

Creates a new execution context, accounted to the PD specified by $\text{SEL}_{\text{OBJ}_{\text{PD}}}$, and sets the processor affinity according to CPU. Prior to the hypercall, $\text{SEL}_{\text{OBJ}_0}$ must refer to a null capability, and $\text{SEL}_{\text{OBJ}_{\text{PD}}}$ must refer to a protection domain capability with permission bit CAP_{EC} set. The caller PD obtains in place of $\text{SEL}_{\text{OBJ}_0}$ an execution context capability that refers to the created EC. The microhypervisor binds the execution context to the protection domain referred to by $\text{SEL}_{\text{OBJ}_{\text{PD}}}$ in the caller PD. If the UTCB address is zero, the microhypervisor creates a virtual CPU, otherwise it creates a thread according to the G flag. Local threads cannot have a scheduling context bound to them. Their initial state is as if they had just done a reply hypercall, so they start running when they receive a request on a portal bound to them. Global threads and virtual CPUs generate a startup exception the first time a scheduling context is bound to them.

Status:

SUCCESS

Hypercall completed successfully.

BAD_CAP

$\text{SEL}_{\text{OBJ}_0}$ did not refer to a Null Capability (CAP_0).
$\text{SEL}_{\text{OBJ}_{\text{PD}}}$ did not refer to a Protection Domain Capability ($\text{CAP}_{\text{OBJ}_{\text{PD}}}$).
$\text{CAP}_{\text{OBJ}_{\text{PD}}}$ has insufficient permissions.

BAD_CPU

Invalid CPU number.

BAD_FTR

Virtual CPUs not supported.

BAD_PAR

Invalid UTCB address.

5.3.3 Create Scheduling Context

Synopsis:

status = create_sc (SEL_{OBJ_0}, $SEL_{OBJ_{PD}}$, $SEL_{OBJ_{EC}}$, QPD);

Parameters:

SEL_{OBJ_0}: Created SC

$SEL_{OBJ_{PD}}$: Owner PD

$SEL_{OBJ_{EC}}$: Bound EC

QPD: Quantum Priority Descriptor (4.5)

Description:

Creates a new scheduling context, accounted to the PD specified by $SEL_{OBJ_{PD}}$, and sets the scheduling parameters according to QPD. Prior to the hypercall, SEL_{OBJ_0} must refer to a null capability, $SEL_{OBJ_{PD}}$ must refer to a protection domain capability with permission bit CAP_{SC} set, and $SEL_{OBJ_{EC}}$ must refer to an execution context capability with permission bit CAP_{SC} set. The caller PD obtains in place of SEL_{OBJ_0} a scheduling context capability that refers to the created SC. The microhypervisor binds the scheduling context to the execution context referred to by $SEL_{OBJ_{EC}}$ in the caller PD.

Status:

SUCCESS

Hypercall completed successfully.

BAD_CAP

SEL_{OBJ_0} did not refer to a Null Capability (CAP_0).
$SEL_{OBJ_{PD}}$ did not refer to a Protection Domain Capability ($CAP_{OBJ_{PD}}$).
$SEL_{OBJ_{EC}}$ did not refer to an Execution Context Capability ($CAP_{OBJ_{EC}}$).
$CAP_{OBJ_{PD}}$ or $CAP_{OBJ_{EC}}$ has insufficient permissions.
Binding the scheduling context to the execution context failed.

BAD_PAR

QPD time quantum or priority is zero.

5.3.4 Create Portal

Synopsis:

status = create_pt (SEL_{OBJ_0}, $SEL_{OBJ_{PD}}$, $SEL_{OBJ_{EC}}$, MTD, IP);

Parameters:

SEL_{OBJ_0}: Created PT

$SEL_{OBJ_{PD}}$: Owner PD

$SEL_{OBJ_{EC}}$: Bound EC

MTD: Message Transfer Descriptor (4.4)

IP: Virtual Address: Instruction Pointer

Description:

Creates a new portal, accounted to the PD specified by $SEL_{OBJ_{PD}}$. Prior to the hypercall, SEL_{OBJ_0} must refer to a null capability, $SEL_{OBJ_{PD}}$ must refer to a protection domain capability with permission bit CAP_{PT} set, and $SEL_{OBJ_{EC}}$ must refer to an execution context capability with permission bit CAP_{PT} set. The caller PD obtains in place of SEL_{OBJ_0} a portal capability that refers to the created portal. The microhypervisor binds the portal to the execution context referred to by $SEL_{OBJ_{EC}}$ in the caller PD.

Status:

SUCCESS

Hypercall completed successfully.

BAD_CAP

SEL_{OBJ_0} did not refer to a Null Capability (CAP_0).
$SEL_{OBJ_{PD}}$ did not refer to a Protection Domain Capability ($CAP_{OBJ_{PD}}$).
$SEL_{OBJ_{EC}}$ did not refer to an Execution Context Capability ($CAP_{OBJ_{EC}}$).
$CAP_{OBJ_{PD}}$ or $CAP_{OBJ_{EC}}$ has insufficient permissions.
Binding the portal to the execution context failed.

5.3.5 Create Semaphore

Synopsis:

```
status = create_sm (SEL_OBJ_0, SEL_OBJ_PD, CNT);
```

Parameters:

SEL_{OBJ_0}: Created SM

$SEL_{OBJ_{PD}}$: Owner PD

CNT: Unsigned: Initial Counter Value

Description:

Creates a new semaphore, accounted to the PD specified by $SEL_{OBJ_{PD}}$. Prior to the hypercall, SEL_{OBJ_0} must refer to a null capability, and $SEL_{OBJ_{PD}}$ must refer to a protection domain capability with permission bit CAP_{SM} set. The caller PD obtains in place of SEL_{OBJ_0} a semaphore capability that refers to the created semaphore. The microhypervisor initializes the semaphore counter with the value of CNT.

Status:

SUCCESS

Hypercall completed successfully.

BAD_CAP

SEL_{OBJ_0} did not refer to a Null Capability (CAP_0).
$SEL_{OBJ_{PD}}$ did not refer to a Protection Domain Capability ($CAP_{OBJ_{PD}}$).
$CAP_{OBJ_{PD}}$ has insufficient permissions.

5.3.6 Revoke Capability Range

Synopsis:

```
status = revoke (CRD);
```

Parameters:

CRD: Capability Range Descriptor (4.3)

Flags:

0	SR
3 … 1	0

SR Self Revoke (0=only children, 1=including self)

Description:

Revokes the capabilities within the range specified by the capability range descriptor from all protection domains that directly or indirectly obtained these capabilities through delegation from the calling protection domain. If the self revoke bit is set, the capabilities will also be revoked from the calling protection domain itself. Once all capabilities to a kernel object have been revoked and no references to the kernel object exist anymore, the kernel object will be destroyed. This operation never fails but can take a long time to complete if there are many capabilities to revoke.

Status:

SUCCESS

Hypercall completed successfully.

5.3.7 Lookup Capability Range

Synopsis:

```
status = lookup (CRD);
```

Parameters:

CRD: Capability Range Descriptor (4.3)

Description:

Looks up a range of capabilities in the caller's protection domain. The caller must specify a base address and type in the CRD prior to the hypercall. If a capability exists at the specified address, the microhypervisor returns a completely filled CRD describing the capability range. Otherwise a null capability range descriptor is returned.

Status:

SUCCESS

Hypercall completed successfully.

5.4 Execution Control

5.4.1 Execution Context Control

Synopsis:

status = ec_ctrl ($\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{EC}}}$);

Parameters:

$\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{EC}}}$: Execution Context

Description:

Pends an event for the specified execution context, which causes it to generate a recall exception before its next return from the microhypervisor.

Status:

SUCCESS

Hypercall completed successfully.

BAD_CAP

$\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{EC}}}$ did not refer to an Execution Context Capability ($\mathrm{CAP}_{\mathrm{OBJ}_{\mathrm{EC}}}$).
$\mathrm{CAP}_{\mathrm{OBJ}_{\mathrm{EC}}}$ has insufficient permissions.

5.4.2 Scheduling Context Control

Synopsis:

`status = sc_ctrl (`$\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{SC}}}$`, &Time);`

Parameters:

$\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{SC}}}$: Scheduling Context

Return Values:

Time: Aggregate consumed execution time in microseconds.

Description:

Returns runtime statistics for the specified scheduling context.

Status:

SUCCESS

Hypercall completed successfully.

BAD_CAP

$\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{SC}}}$ did not refer to a Scheduling Context Capability ($\mathrm{CAP}_{\mathrm{OBJ}_{\mathrm{SC}}}$).
$\mathrm{CAP}_{\mathrm{OBJ}_{\mathrm{SC}}}$ has insufficient permissions.

5.4.3 Semaphore Control

Synopsis:

`status = sm_ctrl (`$\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{SM}}}$`);`

Parameters:

SEL$_{\mathrm{OBJ}_{\mathrm{SM}}}$**:** Semaphore

Flags:

3	2	1	0
0		ZC	OP

OP Operation (0=up, 1=down)

ZC Zero Counter (0=decrement, 1=set to zero)

Description:

The *down* operation blocks the calling execution context if the semaphore counter is zero, otherwise the counter is decremented or set to zero, depending on the setting of the ZC bit.

The *up* operation releases an execution context blocked on the semaphore if one exists, otherwise it increments the counter.

Status:

SUCCESS

Hypercall completed successfully.

BAD_CAP

$\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{SM}}}$ did not refer to a Semaphore Capability ($\mathrm{CAP}_{\mathrm{OBJ}_{\mathrm{SM}}}$).

$\mathrm{CAP}_{\mathrm{OBJ}_{\mathrm{SM}}}$ has insufficient permissions.

5.5 Device Control

5.5.1 Assign PCI Device

Synopsis:

status = assign_pci ($\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{PD}}}$, $\mathrm{SEL}_{\mathrm{MEM}_{\mathrm{DEV}}}$, RID);

Parameters:

$\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{PD}}}$: Target PD

$\mathrm{SEL}_{\mathrm{MEM}_{\mathrm{DEV}}}$: PCI Device

RID: Routing Hint

Description:

Assigns the PCI device, named by $\mathrm{SEL}_{\mathrm{MEM}_{\mathrm{DEV}}}$, to the protection domain, named by $\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{PD}}}$. $\mathrm{SEL}_{\mathrm{MEM}_{\mathrm{DEV}}}$ must refer to a memory capability for the memory-mapped PCI configuration space of the device.

Status:

SUCCESS

Hypercall completed successfully.

BAD_CAP

$\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{PD}}}$ did not refer to a Protection Domain Capability ($\mathrm{CAP}_{\mathrm{OBJ}_{\mathrm{PD}}}$).

BAD_DEV

$\mathrm{SEL}_{\mathrm{MEM}_{\mathrm{DEV}}}$ did not refer to a valid PCI device.

5.5.2 Assign Global System Interrupt

Synopsis:

`status = assign_gsi (SEL`$_{OBJ_{SM}}$`, SEL`$_{MEM_{DEV}}$`, CPU, &MSI);`

Parameters:

SEL$_{OBJ_{SM}}$**:** Interrupt Semaphore

SEL$_{MEM_{DEV}}$**:** PCI Device

CPU: CPU Number

Return Values:

MSI: Values to program into the MSI registers of the PCI device to ensure proper operation.

Description:

Routes the GSI, named by SEL$_{OBJ_{SM}}$, to the specified CPU, where it will be signaled on the corresponding interrupt semaphore. For GSIs delivered as MSI, SEL$_{MEM_{DEV}}$ must refer to a memory capability for the memory-mapped PCI configuration space of the device that generates the interrupt. For GSIs delivered through an IOAPIC pin, SEL$_{MEM_{DEV}}$ is ignored.

Status:

SUCCESS

Hypercall completed successfully.

BAD_CAP

SEL$_{OBJ_{SM}}$ did not refer to an Interrupt Semaphore Capability (CAP$_{OBJ_{SM}}$).

BAD_DEV

SEL$_{MEM_{DEV}}$ did not refer to a valid PCI device.

BAD_CPU

Invalid CPU number.

6 Booting

6.1 Root Protection Domain

When the microhypervisor has initialized the system, it creates the root protection domain with a root execution context and a root scheduling context.

6.1.1 Resource Access

Execution contexts in the root protection domain have the special ability to request resources from the microhypervisor during communication, by setting the H-bit in a typed item (4.6.2.2). In addition to memory and I/O ports, the following capabilities can be requested:

Idle Scheduling Contexts

Capability selectors 0 ... n - 1 in the microhypervisor refer to $\text{CAP}_{\text{OBJ}_{\text{SC}}}$ for the idle thread of the respective CPU, where n is the maximum number of supported CPUs, as indicated by the HIP. These capabilities can be used with the `sc_ctrl` hypercall.

Interrupt Semaphores

Capability selectors n ... n + GSI - 1 in the microhypervisor refer to $\text{CAP}_{\text{OBJ}_{\text{SM}}}$ for global system interrupts, where GSI is the maximum number of supported GSIs, as indicated by the HIP. These capabilities can be used with the `sm_ctrl` and `assign_gsi` hypercalls.

6.1.2 Initial Configuration

At bootup the root protection domain is configured as follows:

6.1.2.1 Memory Space

Program Segments

The microhypervisor loads the program segments of the roottask into the memory space as specified by the ELF program headers of the roottask image.

Hypervisor Information Page

The HIP is mapped into the memory space at a specific virtual address that is passed to the root execution context during startup.

UTCB

The UTCB of the root execution context is mapped into the memory space just below the HIP.

All other regions of the memory space are initially empty.

6.1.2.2 I/O Space

The I/O space is initially empty.

6.1.2.3 Object Space

The object space contains the following capabilities:

- Capability selector EXC + 0 refers to the root PD capability.
- Capability selector EXC + 1 refers to the root EC capability.
- Capability selector EXC + 2 refers to the root SC capability.

All other capability selectors refer to null capabilities.

6.2 Hypervisor Information Page

The Hypervisor Information Page (HIP) conveys information about the platform and configuration to the root protection domain. The processor register that contains the virtual address of the HIP during booting is ABI-specific (IV). Figure 6.1 shows the layout of the Hypervisor Information Page. All fields are unsigned values unless stated otherwise.

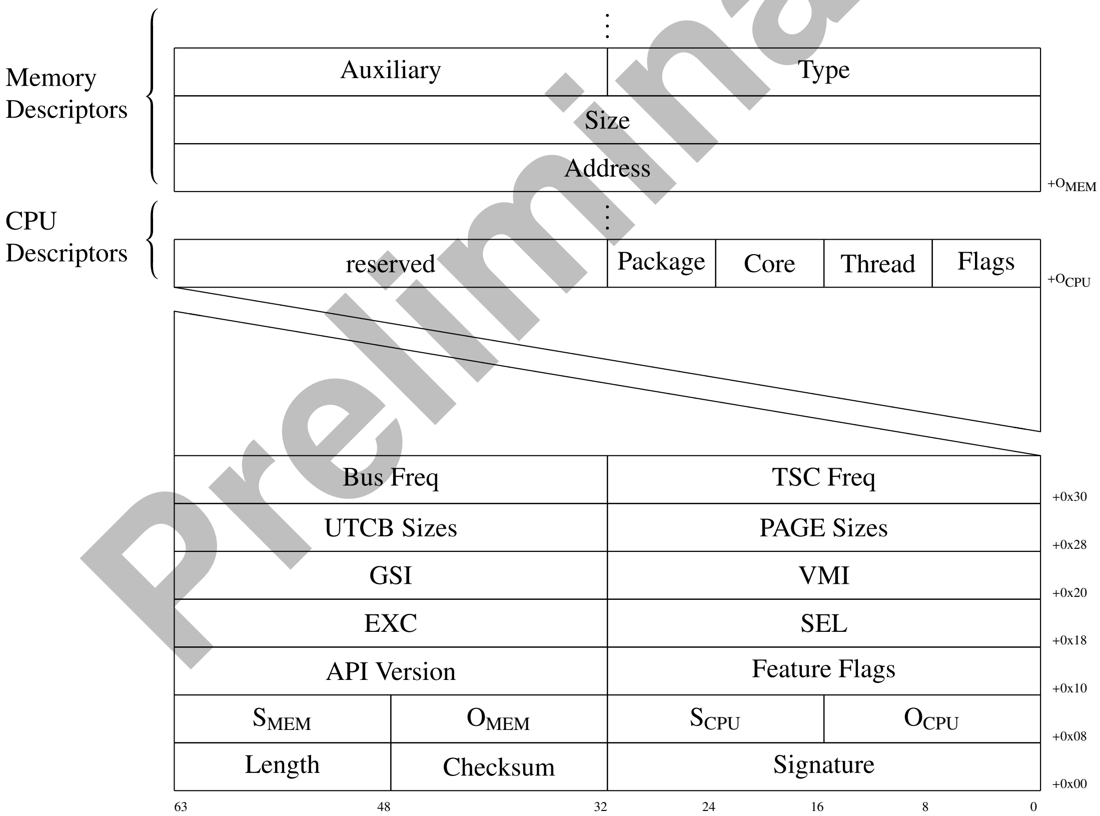

Figure 6.1: Hypervisor Information Page

Signature:

A value of 0x41564f4e identifies the NOVA microhypervisor.

Checksum:

The checksum is valid if 16bit-wise addition the HIP contents produces a value of 0.

Length:

Length of the HIP in bytes. This includes all CPU and memory descriptors.

O_{CPU}:

Offset of the first CPU descriptor in bytes, relative to the HIP base.

S_{CPU}:

Size of one CPU descriptor in bytes.

O_{MEM}:

Offset of the first memory descriptor in bytes, relative to the HIP base.

S_{MEM}:

Size of one memory descriptor in bytes.

Feature Flags:

The microhypervisor supports a particular feature if and only if the corresponding bit in the feature flags is set to 1. The following features are currently defined:

31 … 3	2	1	0
0	SVM	VMX	0

VMX: Intel Virtual Machine Extensions

SVM: AMD Secure Virtual Machine

API Version:

API version number.

SEL:

Number of available capability selectors in each object space. Specifying a capability selector beyond the maximum number supported wraps around to the beginning of the object space.

EXC:

Number of capability selectors used for exception handling (3.3).

VMI:

Number of capability selectors used for virtual-machine intercept handling (3.3).

GSI:

Number of global system interrupts (3.4).

PAGE Sizes:

If bit n is set, the implementation supports memory pages of size 2^n bytes.

UTCB Sizes:

If bit n is set, the implementation supports user thread control blocks of size 2^n bytes.

TSC Freq:

Time Stamp Counter Frequency in kHz.

BUS Freq:

Interconnect Frequency in kHz.

CPU Descriptor

The array of CPU descriptors contains n_{cpu} entries, where

$$n_{cpu} = \frac{O_{MEM} - O_{CPU}}{S_{CPU}}. \tag{6.1}$$

The value of n_{cpu} reflects the maximum number of CPUs supported by the microhypervisor. The array index of a CPU descriptor corresponds to the CPU number that must be specified for certain hypercalls to target that CPU. A CPU can only be used if its descriptor is marked as enabled.

Package, Core, Thread:

CPU multiprocessor topology information.

Flags:

CPU status flags.

7–1	0
0	Enabled

MEM Descriptor

The array of MEM descriptors contains n_{mem} entries, where

$$n_{mem} = \frac{Length - O_{MEM}}{S_{MEM}}. \tag{6.2}$$

Address:

Physical base address of memory region.

Size:

Size of memory region in bytes.

Type:

Type of memory region. Note that the allocates ranges overlap the available ranges.

Type	Description	
-2	Multiboot Module	Allocated ranges
-1	Microhypervisor	
1	Available Memory	Available ranges
2	Reserved Memory	
3	ACPI Reclaim Memory	
4	ACPI NVS Memory	

Auxiliary:

Physical address of command line if type is 'Multiboot Module', reserved otherwise.

Part IV

Application Binary Interface

7 ABI x86-32

7.1 Initial State

Figure 7.1 details the state of the CPU registers when the microhypervisor has finished booting and transfers control to the root protection domain.

Register	Description
CS	Selector=~, Base=0, Limit=0xFFFFFFFF, Code Segment, ro
SS,DS,ES,FS,GS	Selector=~, Base=0, Limit=0xFFFFFFFF, Data Segment, rw
EIP	Address of entry point from ELF header
ESP	Address of hypervisor information page
EAX	Bootstrap CPU number
ECX,EDX,EBX,EBP,ESI,EDI	~
EFLAGS	0x202

Figure 7.1: Initial State

7.2 Event-Specific Capability Selectors

For the delivery of exception and intercept messages, the microhypervisor performs an implicit portal traversal. The destination portal is determined by adding the event number to SEL_{EVT} of the affected execution context.

Exceptions

Number	Exception	Number	Exception	Number	Exception	Number	Exception
0x0	#DE	0x8	#DF[1]	0x10	#MF	0x18	reserved
0x1	#DB	0x9	reserved	0x11	#AC	0x19	reserved
0x2	reserved	0xa	#TS[1]	0x12	#MC[1]	0x1a	reserved
0x3	#BP	0xb	#NP	0x13	#XM	0x1b	reserved
0x4	#OF	0xc	#SS	0x14	reserved	0x1c	reserved
0x5	#BR	0xd	#GP	0x15	reserved	0x1d	reserved
0x6	#UD	0xe	#PF	0x16	reserved	0x1e	STARTUP
0x7	#NM[1]	0xf	reserved	0x17	reserved	0x1f	RECALL

VMX Intercepts

Number	Intercept	Number	Intercept	Number	Intercept
0x0	Exception or NMI[1]	0x15	VMPTRLD	0x2a	reserved
0x1	INTR[1]	0x16	VMPTRST	0x2b	TPR Below Threshold
0x2	Triple Fault[2]	0x17	VMREAD	0x2c	APIC Access
0x3	INIT[2]	0x18	VMRESUME	0x2d	reserved
0x4	SIPI[2]	0x19	VMWRITE	0x2e	GDTR/IDTR Access
0x5	I/O SMI	0x1a	VMXOFF	0x2f	LDTR/TR Access
0x6	Other SMI	0x1b	VMXON	0x30	EPT Violation[2]
0x7	Interrupt Window	0x1c	CR Access[1]	0x31	EPT Misconfiguration[1]
0x8	NMI Window	0x1d	DR Access	0x32	INVEPT
0x9	Task Switch[2]	0x1e	I/O Access[2]	0x33	RDTSCP
0xa	CPUID[2]	0x1f	RDMSR[2]	0x34	VMX Preemption Timer
0xb	GETSEC[2]	0x20	WRMSR[2]	0x35	INVVPID
0xc	HLT[2]	0x21	Invalid Guest State[2]	0x36	WBINVD
0xd	INVD[2]	0x22	MSR Load Failure	0x37	XSETBV
0xe	INVLPG[1]	0x23	reserved	0x38	reserved
0xf	RDPMC	0x24	MWAIT	0x39	reserved
0x10	RDTSC	0x25	MTF	0x3a	reserved
0x11	RSM	0x26	reserved	0x3b	reserved
0x12	VMCALL	0x27	MONITOR	0x3c	reserved
0x13	VMCLEAR	0x28	PAUSE	0xfe	STARTUP
0x14	VMLAUNCH	0x29	Machine Check	0xff	RECALL

SVM Intercepts

Number	Intercept	Number	Intercept	Number	Intercept
0x0–0xf	CR Read	0x6e	RDTSC	0x81	VMMCALL
0x10–0x1f	CR Write	0x6f	RDPMC	0x82	VMLOAD[2]
0x20–0x2f	DR Read	0x70	PUSHF	0x83	VMSAVE[2]
0x30–0x3f	DR Write	0x71	POPF	0x84	STGI
0x40–0x5f	Exception[1]	0x72	CPUID	0x85	CLGI[2]
0x60	INTR[1]	0x73	RSM	0x86	SKINIT[2]
0x61	NMI[1]	0x74	IRET	0x87	RDTSCP
0x62	SMI	0x75	INT	0x88	ICEBP
0x63	INIT[2]	0x76	INVD[2]	0x89	WBINVD
0x64	Interrupt Window	0x77	PAUSE	0x8a	MONITOR
0x65	CR0 Selective Write	0x78	HLT[2]	0x8b	MWAIT
0x66	IDTR Read	0x79	INVLPG	0x8c	MWAIT (cond.)
0x67	GDTR Read	0x7a	INVLPGA	0x8d	reserved
0x68	LDTR Read	0x7b	I/O Access[2]	0x8e	reserved
0x69	TR Read	0x7c	MSR Access[2]	0x8f	reserved
0x6a	IDTR Write	0x7d	Task Switch	0xfc	NPT Fault[2]
0x6b	GDTR Write	0x7e	FERR Freeze	0xfd	Invalid Guest State[2]
0x6c	LDTR Write	0x7f	Triple Fault[2]	0xfe	STARTUP
0x6d	TR Write	0x80	VMRUN	0xff	RECALL

[1] These events do not currently cause a portal traversal, because the microhypervisor handles them internally.
[2] These events are currently force-enabled by the microhypervisor or by hardware.

7.3 UTCB Data Layout

reserved	IDTR Base	IDTR Limit	reserved		+0x110
reserved	GDTR Base	GDTR Limit	reserved		+0x100
reserved	TR Base	TR Limit	TR AR	TR Sel	+0xf0
reserved	LDTR Base	LDTR Limit	LDTR AR	LDTR Sel	+0xe0
reserved	GS Base	GS Limit	GS AR	GS Sel	+0xd0
reserved	FS Base	FS Limit	FS AR	FS Sel	+0xc0
reserved	DS Base	DS Limit	DS AR	DS Sel	+0xb0
reserved	SS Base	SS Limit	SS AR	SS Sel	+0xa0
reserved	CS Base	CS Limit	CS AR	CS Sel	+0x90
reserved	ES Base	ES Limit	ES AR	ES Sel	+0x80
SYSENTER EIP	SYSENTER ESP	SYSENTER CS	DR7		+0x70
CR4	CR3	CR2	CR0		+0x60
TSC Offset		Secondary Exit Ctrl	Primary Exit Ctrl		+0x50
Secondary Exit Qual		Primary Exit Qual			+0x40
EDI	ESI	EBP	ESP		+0x30
EBX	EDX	ECX	EAX		+0x20
Injection Error	Injection Info	Activity State	Interruptibility State		+0x10
EFLAGS	EIP	Instruction Length	MTD		+0x00

7.4 Message Transfer Descriptor

Figure 7.2 illustrates the format of the MTD for exceptions and intercepts, as described in Section 4.4.

0			TSC	STA	INJ	CTRL	QUAL	SYS	DR	CR	IDTR	GDTR	LDTR	TR	CS SS	FS GS	DS ES	EFL	EIP	ESP	BSD	ACDB
22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0

Figure 7.2: Message Transfer Descriptor

Each bit controls the transfer of a subset of the CPU state to/from the respective UTCB fields (7.3). State with access type *r* can be read from CPU into the UTCB. State with access type *w* can be written from the UTCB into the CPU.

Bit	Type	Exceptions	Intercepts
ACDB	rw	EAX, ECX, EDX, EBX	EAX, ECX, EDX, EBX
BSD	rw	EBP, ESI, EDI	EBP, ESI, EDI
ESP	rw	ESP	ESP
EIP	rw	EIP	EIP, Instruction Length
EFL	rw	EFLAGS[3]	EFLAGS
DS ES	rw	≡	DS, ES (Selector, Base, Limit, Access Rights)
FS GS	rw	≡	FS, GS (Selector, Base, Limit, Access Rights)
CS SS	rw	≡	CS, SS (Selector, Base, Limit, Access Rights)
TR	rw	≡	TR (Selector, Base, Limit, Access Rights)
LDTR	rw	≡	LDTR (Selector, Base, Limit, Access Rights)
GDTR	rw	≡	GDTR (Base, Limit)
IDTR	rw	≡	IDTR (Base, Limit)
CR	rw	≡	CR0, CR2, CR3, CR4
DR	rw	≡	DR7
SYS	rw	≡	SYSENTER MSRs (CS, ESP, EIP)
QUAL	r	Exit Qualifications[4]	Exit Qualifications
CTRL	w	≡	Execution Controls
INJ	rw	≡	Injection Info, Injection Error Code
STA	rw	≡	Interruptibility State, Activity State
TSC	rw	≡	TSC Offset

[3] Only the status flags are writable.
[4] The primary exit qualification contains the exception error code. The secondary exit qualification contains the fault address.

7.5 Calling Convention

The following pages describes the calling convention for each hypercall. An execution context calls into the microhypervisor by loading the hypercall identifier and other parameters into the specified processor registers and then executes the *sysenter* instruction.

The hypercall identifier consists of the hypercall number and hypercall-specific flags, as illustrated in Figure 7.3.

flags	number
7 … 4	3 … 0

Figure 7.3: Hypercall Identifier

The status code returned from a hypercall has the format shown in Figure 7.4.

status
7 … 0

Figure 7.4: Status Code

The assignment of hypercall parameters to general-purpose registers is shown on the left side; the contents of the registers after the hypercall is shown on the right side.

Call

$\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{PT}}}[31-8]$/hypercall$_{[7-0]}$	EAX	— **Call** ⟶	EAX	status$_{[7-0]}$
return SP	ECX		ECX	≡
return IP	EDX		EDX	≡
–	EDI		EDI	≡
–	ESI	sysenter	ESI	≡
–	EBX		EBX	≡
–	EBP		EBP	≡
–	ESP		ESP	ECX
–	EIP		EIP	EDX

Reply

hypercall$_{[7-0]}$	EAX	— **Reply** ⟶	EAX	pt→id
return SP	ECX		ECX	≡
–	EDX		EDX	∼
–	EDI		EDI	≡
–	ESI	sysenter	ESI	≡
–	EBX		EBX	≡
–	EBP		EBP	≡
–	ESP		ESP	ECX
–	EIP		EIP	pt→ip

Create Protection Domain

$\mathrm{SEL}_{\mathrm{OBJ}_0[31-8]}$/hypercall$_{[7-0]}$	EAX	— **Create PD** ⟶	EAX	status$_{[7-0]}$
return SP	ECX		ECX	≡
return IP	EDX		EDX	≡
$\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{PD}}}$	EDI		EDI	≡
$\mathrm{CRD}_{\mathrm{OBJ}}$	ESI	sysenter	ESI	≡
–	EBX		EBX	≡
–	EBP		EBP	≡
–	ESP		ESP	ECX
–	EIP		EIP	EDX

Create Execution Context

$\mathrm{SEL}_{\mathrm{OBJ}_0[31-8]}$/hypercall$_{[7-0]}$	EAX	— **Create EC** ⟶	EAX	status$_{[7-0]}$
return SP	ECX		ECX	≡
return IP	EDX		EDX	≡
$\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{PD}}}$	EDI		EDI	≡
$\mathrm{UTCB}_{[31-12]}$/$\mathrm{CPU}_{[11-0]}$	ESI	sysenter	ESI	≡
SP	EBX		EBX	≡
$\mathrm{SEL}_{\mathrm{EVT}}$	EBP		EBP	≡
–	ESP		ESP	ECX
–	EIP		EIP	EDX

Create Scheduling Context

$\mathrm{SEL}_{\mathrm{OBJ}_0[31-8]}$/hypercall$_{[7-0]}$	EAX	— **Create SC** ⟶	EAX	status$_{[7-0]}$
return SP	ECX		ECX	≡
return IP	EDX		EDX	≡
$\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{PD}}}$	EDI		EDI	≡
$\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{EC}}}$	ESI	sysenter	ESI	≡
QPD	EBX		EBX	≡
–	EBP		EBP	≡
–	ESP		ESP	ECX
–	EIP		EIP	EDX

Create Portal

$\mathrm{SEL}_{\mathrm{OBJ}_0[31-8]}$/hypercall$_{[7-0]}$	EAX	— **Create PT** ⟶	EAX	status$_{[7-0]}$
return SP	ECX		ECX	≡
return IP	EDX		EDX	≡
$\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{PD}}}$	EDI		EDI	≡
$\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{EC}}}$	ESI	sysenter	ESI	≡
MTD	EBX		EBX	≡
IP	EBP		EBP	≡
–	ESP		ESP	ECX
–	EIP		EIP	EDX

Create Semaphore

$\mathrm{SEL}_{\mathrm{OBJ}_0[31-8]}$/hypercall$_{[7-0]}$	EAX	**— Create SM ⟶**	EAX	status$_{[7-0]}$
return SP	ECX		ECX	≡
return IP	EDX		EDX	≡
$\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{PD}}}$	EDI		EDI	≡
CNT	ESI	sysenter	ESI	≡
–	EBX		EBX	≡
–	EBP		EBP	≡
–	ESP		ESP	ECX
–	EIP		EIP	EDX

Revoke Capability Range

hypercall$_{[7-0]}$	EAX	**— Revoke ⟶**	EAX	status$_{[7-0]}$
return SP	ECX		ECX	≡
return IP	EDX		EDX	≡
CRD	EDI		EDI	≡
–	ESI	sysenter	ESI	≡
–	EBX		EBX	≡
–	EBP		EBP	≡
–	ESP		ESP	ECX
–	EIP		EIP	EDX

Lookup Capability Range

hypercall$_{[7-0]}$	EAX	**— Lookup ⟶**	EAX	status$_{[7-0]}$
return SP	ECX		ECX	≡
return IP	EDX		EDX	≡
CRD	EDI		EDI	CRD
–	ESI	sysenter	ESI	≡
–	EBX		EBX	≡
–	EBP		EBP	≡
–	ESP		ESP	ECX
–	EIP		EIP	EDX

Execution Context Control

$\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{EC}}[31-8]}$/hypercall$_{[7-0]}$	EAX	**— EC Control ⟶**	EAX	status$_{[7-0]}$
return SP	ECX		ECX	≡
return IP	EDX		EDX	≡
–	EDI		EDI	≡
–	ESI	sysenter	ESI	≡
–	EBX		EBX	≡
–	EBP		EBP	≡
–	ESP		ESP	ECX
–	EIP		EIP	EDX

Scheduling Context Control

$\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{SC}}[31-8]}$/hypercall$_{[7-0]}$	EAX	— **SC Control** ⟶	EAX	status$_{[7-0]}$
return SP	ECX		ECX	≡
return IP	EDX		EDX	≡
–	EDI		EDI	Time$_{[63-32]}$
–	ESI	sysenter	ESI	Time$_{[31-0]}$
–	EBX		EBX	≡
–	EBP		EBP	≡
–	ESP		ESP	ECX
–	EIP		EIP	EDX

Semaphore Control

$\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{SM}}[31-8]}$/hypercall$_{[7-0]}$	EAX	— **SM Control** ⟶	EAX	status$_{[7-0]}$
return SP	ECX		ECX	≡
return IP	EDX		EDX	≡
–	EDI		EDI	≡
–	ESI	sysenter	ESI	≡
–	EBX		EBX	≡
–	EBP		EBP	≡
–	ESP		ESP	ECX
–	EIP		EIP	EDX

Assign PCI Device

$\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{PD}}[31-8]}$/hypercall$_{[7-0]}$	EAX	— **Assign PCI** ⟶	EAX	status$_{[7-0]}$
return SP	ECX		ECX	≡
return IP	EDX		EDX	≡
$\mathrm{SEL}_{\mathrm{MEM}_{\mathrm{DEV}}}$	EDI		EDI	≡
RID	ESI	sysenter	ESI	≡
–	EBX		EBX	≡
–	EBP		EBP	≡
–	ESP		ESP	ECX
–	EIP		EIP	EDX

Assign Global System Interrupt

$\mathrm{SEL}_{\mathrm{OBJ}_{\mathrm{SM}}[31-8]}$/hypercall$_{[7-0]}$	EAX	— **Assign GSI** ⟶	EAX	status$_{[7-0]}$
return SP	ECX		ECX	≡
return IP	EDX		EDX	≡
$\mathrm{SEL}_{\mathrm{MEM}_{\mathrm{DEV}}}$	EDI		EDI	MSI Addr
CPU	ESI	sysenter	ESI	MSI Data
–	EBX		EBX	≡
–	EBP		EBP	≡
–	ESP		ESP	ECX
–	EIP		EIP	EDX

Part V

Appendix

A Acronyms

ABI	Application Binary Interface
CAP	Capability
CAP$_0$	Null Capability
CAP$_{\mathrm{I/O}}$	I/O Capability
CAP$_{\mathrm{MEM}}$	Memory Capability
CAP$_{\mathrm{OBJ}}$	Object Capability
CAP$_{\mathrm{OBJ}_{\mathrm{EC}}}$	Execution Context Capability
CAP$_{\mathrm{OBJ}_{\mathrm{PD}}}$	Protection Domain Capability
CAP$_{\mathrm{OBJ}_{\mathrm{PT}}}$	Portal Capability
CAP$_{\mathrm{OBJ}_{\mathrm{SC}}}$	Scheduling Context Capability
CAP$_{\mathrm{OBJ}_{\mathrm{SM}}}$	Semaphore Capability
CAP$_{\mathrm{RP}}$	Reply Capability
CPU	Central Processing Unit
CRD	Capability Range Descriptor
CRD$_0$	Null Capability Range Descriptor
CRD$_{\mathrm{I/O}}$	I/O Capability Range Descriptor
CRD$_{\mathrm{MEM}}$	Memory Capability Range Descriptor
CRD$_{\mathrm{OBJ}}$	Object Capability Range Descriptor
EC	Execution Context
ELF	Executable and Linkable Format
FPU	Floating Point Unit
GSI	Global System Interrupt
HIP	Hypervisor Information Page
MSI	Message Signaled Interrupt
MTD	Message Transfer Descriptor
IOAPIC	I/O Advanced Programmable Interrupt Controller
IP	Instruction Pointer
PD	Protection Domain

PT	Portal
QPD	Quantum Priority Descriptor
SC	Scheduling Context
SEL	Capability Selector
$\textbf{SEL}_{\text{EVT}}$	Event Selector Base
$\textbf{SEL}_{\text{I/O}}$	I/O Capability Selector
$\textbf{SEL}_{\text{MEM}}$	Memory Capability Selector
$\textbf{SEL}_{\text{MEM}_{\text{DEV}}}$	Memory Capability Selector: Device Configuration Space
$\textbf{SEL}_{\text{OBJ}}$	Object Capability Selector
$\textbf{SEL}_{\text{OBJ}_0}$	Object Capability Selector: Null Capability
$\textbf{SEL}_{\text{OBJ}_{\text{EC}}}$	Object Capability Selector: Execution Context Capability
$\textbf{SEL}_{\text{OBJ}_{\text{PD}}}$	Object Capability Selector: Protection Domain Capability
$\textbf{SEL}_{\text{OBJ}_{\text{PT}}}$	Object Capability Selector: Portal Capability
$\textbf{SEL}_{\text{OBJ}_{\text{SC}}}$	Object Capability Selector: Scheduling Context Capability
$\textbf{SEL}_{\text{OBJ}_{\text{SM}}}$	Object Capability Selector: Semaphore Capability
SM	Semaphore
SP	Stack Pointer
UTCB	User Thread Control Block
VMM	Virtual-Machine Monitor
VM	Virtual Machine