Linux Plumbers Conference

Richmond, Virginia | November 13-15, 2023

Update on RISC-V CoVE

Atish Patra, Ravi Sahita

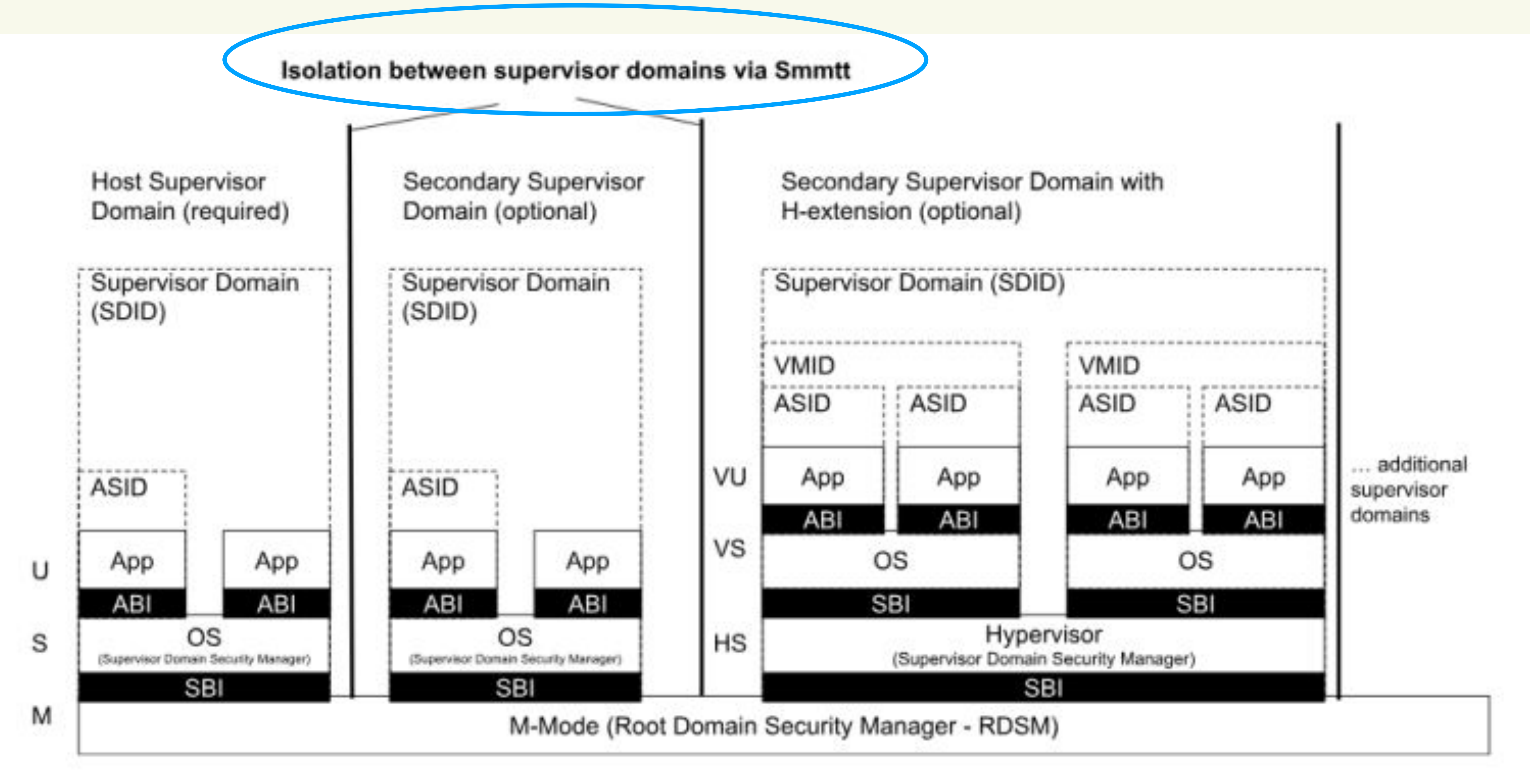
November 14th 2023

RISC-V Supervisor Domains (ISA TG)

- **Platforms supporting multiple tenants (apps, VMs) rely on HW isolation primitives managed by a single host/privileged software**

  -- no separation of TCB

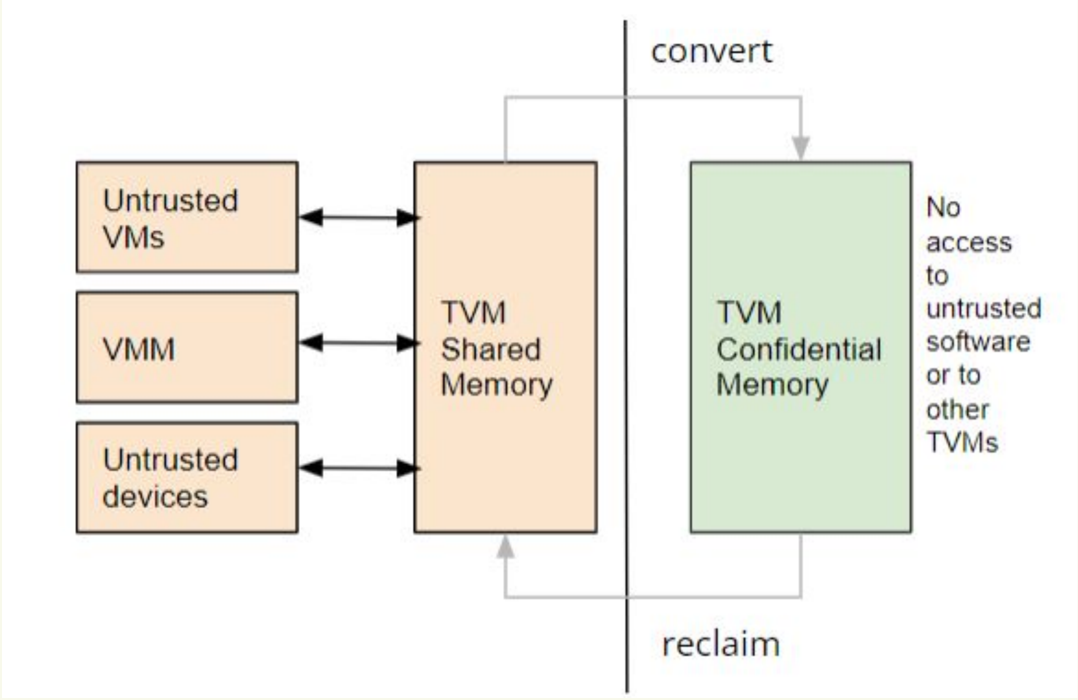
- **RISC-V Supervisor Domains is a priv ISA extension for *isolation between multiple privileged (S/H mode) software execution contexts*, thus enabling differentiated trust models - it entails:**
  - Sdid & Memory Isolation - Smsdid, Smmtt
  - Assigning interrupts to supervisor domains - Smsdia
  - IO-MTT - associating IOMMU & MTT to supervisor domains
  - Metadata attached to address translation, e.g. shared memory - Svpams
  - Secure debug and performance monitoring controls


Isolation between supervisor domains via Smmtt
Host Supervisor Domain (required)
Secondary Supervisor Domain (optional)
Secondary Supervisor Domain with H-extension (optional)
Supervisor Domain (SDID)
Supervisor Domain (SDID)
Supervisor Domain (SDID)
VMID
VMID
ASID
ASID
ASID
ASID
ASID
ASID
App
ABI
App
ABI
App
ABI
App
ABI
App
ABI
App
ABI
App
ABI
App
ABI
OS
SBI
OS
SBI
U
S
M
VU
VS
HS
OS (Supervisor Domain Security Manager)
SBI
OS (Supervisor Domain Security Manager)
SBI
Hypervisor (Supervisor Domain Security Manager)
SBI
M-Mode (Root Domain Security Manager - RDSM)
... additional supervisor domains

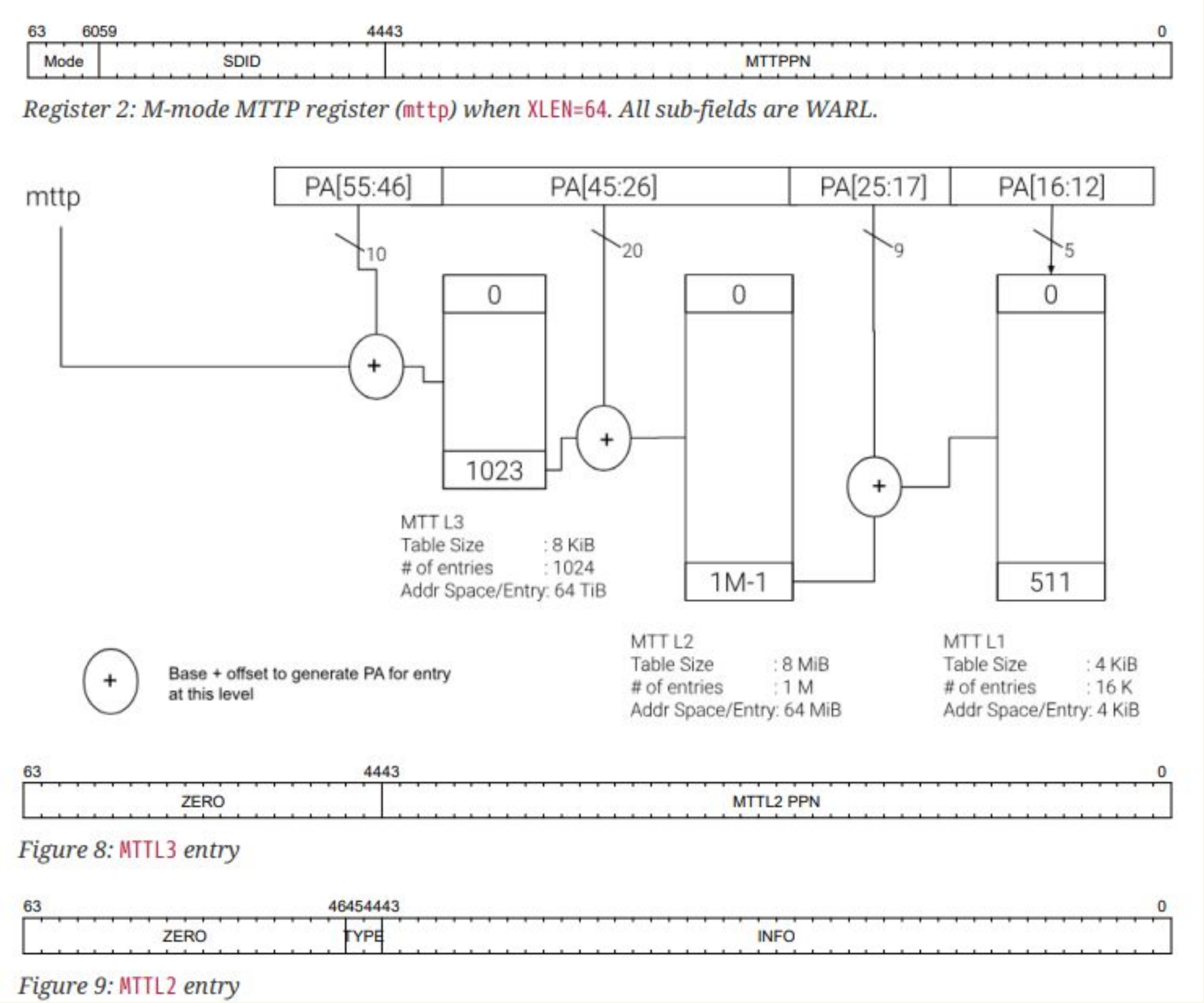
https://github.com/riscv/riscv-smmtt/releases/download/v1.0.3/smmtt-spec.pdf

Smmtt

- **Rich-OS TEEs require *dynamic access-control* of physical memory**
  - Region granularity (4KB and above multiples of page sizes)
  - Regions may be donated by a hosting domain to one or more supervisor domains (access-controlled by Smmtt)
  - Supervisor domain manager can use existing priv ISA (S-stage, G-stage page tables, SPMP) to isolate between its workloads



- Supervisor domain memory accesses may require metadata to be specified with access e.g. to identify encryption contexts - Addressed by *Svpams*



63 60	59 44	43 0
Mode	SDID	MTTPPN

Register 2: M-mode MTTP register (mttp) when XLEN=64. All sub-fields are WARL.

63 44	43 0
ZERO	MTTL2 PPN

Figure 8: MTTL3 entry

63 46	45 44	43 0
ZERO	TYPE	INFO

Figure 9: MTTL2 entry

https://github.com/riscv/riscv-smmtt/releases/download/v1.0.3/smmtt-spec.pdf

Smmtt

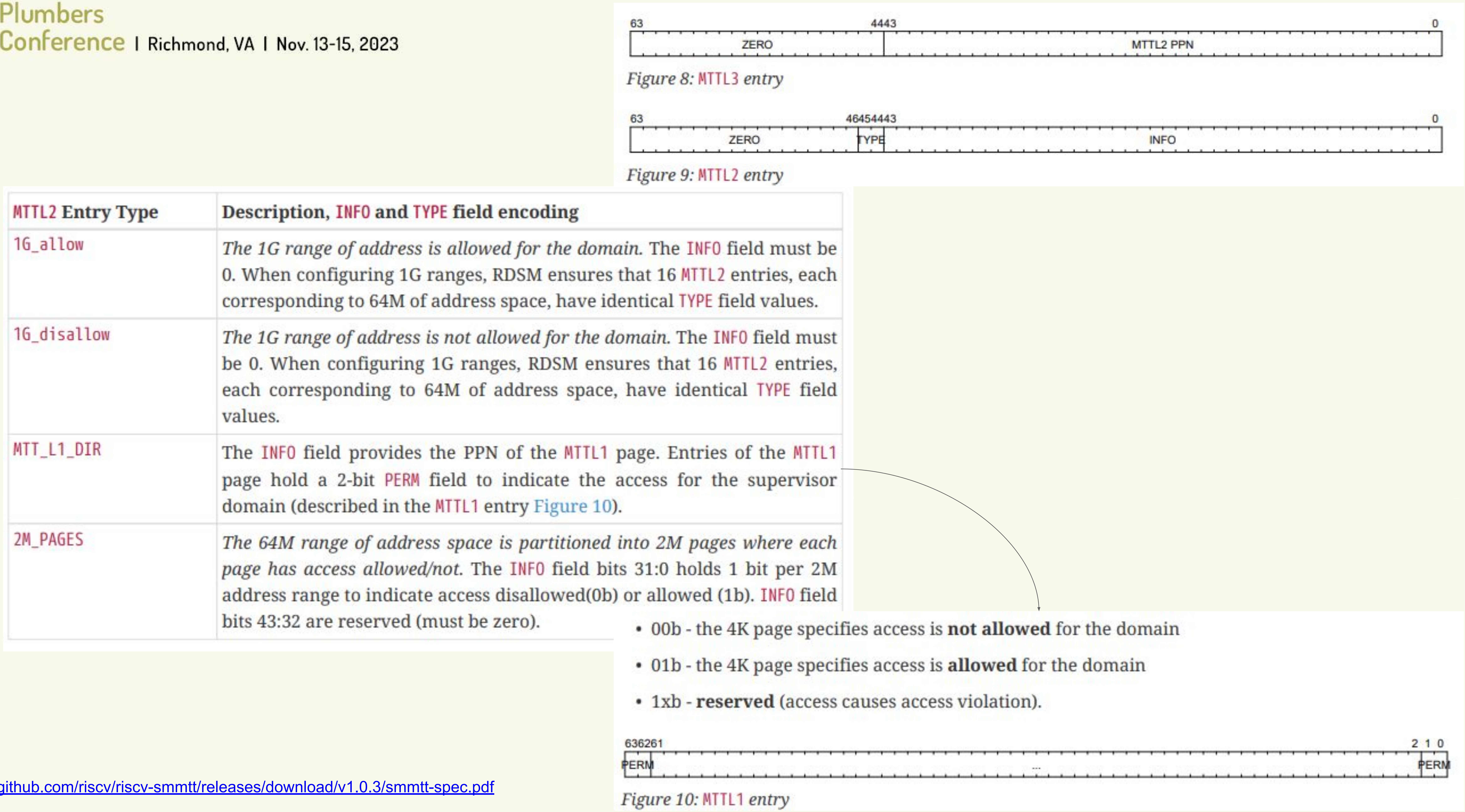

63
4443
0
ZERO
MTTL2 PPN

Figure 8: MTTL3 *entry*

63
46454443
0
ZERO
TYPE
INFO

Figure 9: MTTL2 *entry*

MTTL2 **Entry Type**	**Description,** INFO **and** TYPE **field encoding**
1G_allow	*The 1G range of address is allowed for the domain.* The INFO field must be 0. When configuring 1G ranges, RDSM ensures that 16 MTTL2 entries, each corresponding to 64M of address space, have identical TYPE field values.
1G_disallow	*The 1G range of address is not allowed for the domain.* The INFO field must be 0. When configuring 1G ranges, RDSM ensures that 16 MTTL2 entries, each corresponding to 64M of address space, have identical TYPE field values.
MTT_L1_DIR	The INFO field provides the PPN of the MTTL1 page. Entries of the MTTL1 page hold a 2-bit PERM field to indicate the access for the supervisor domain (described in the MTTL1 entry Figure 10).
2M_PAGES	*The 64M range of address space is partitioned into 2M pages where each page has access allowed/not.* The INFO field bits 31:0 holds 1 bit per 2M address range to indicate access disallowed(0b) or allowed (1b). INFO field bits 43:32 are reserved (must be zero).

- 00b - the 4K page specifies access is **not allowed** for the domain
- 01b - the 4K page specifies access is **allowed** for the domain
- 1xb - **reserved** (access causes access violation).

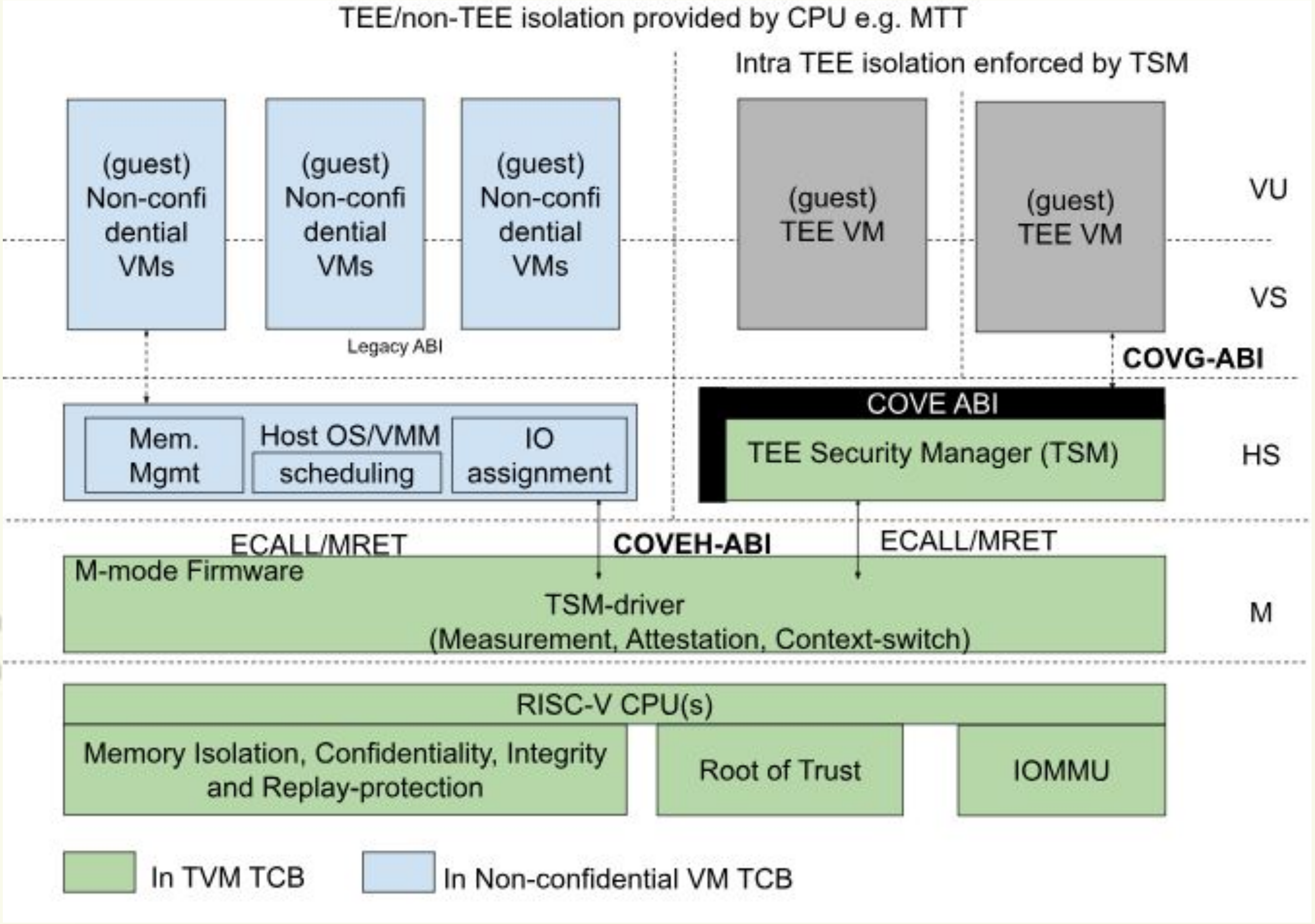
Figure 10: MTTL1 *entry*

https://github.com/riscv/riscv-smmtt/releases/download/v1.0.3/smmtt-spec.pdf

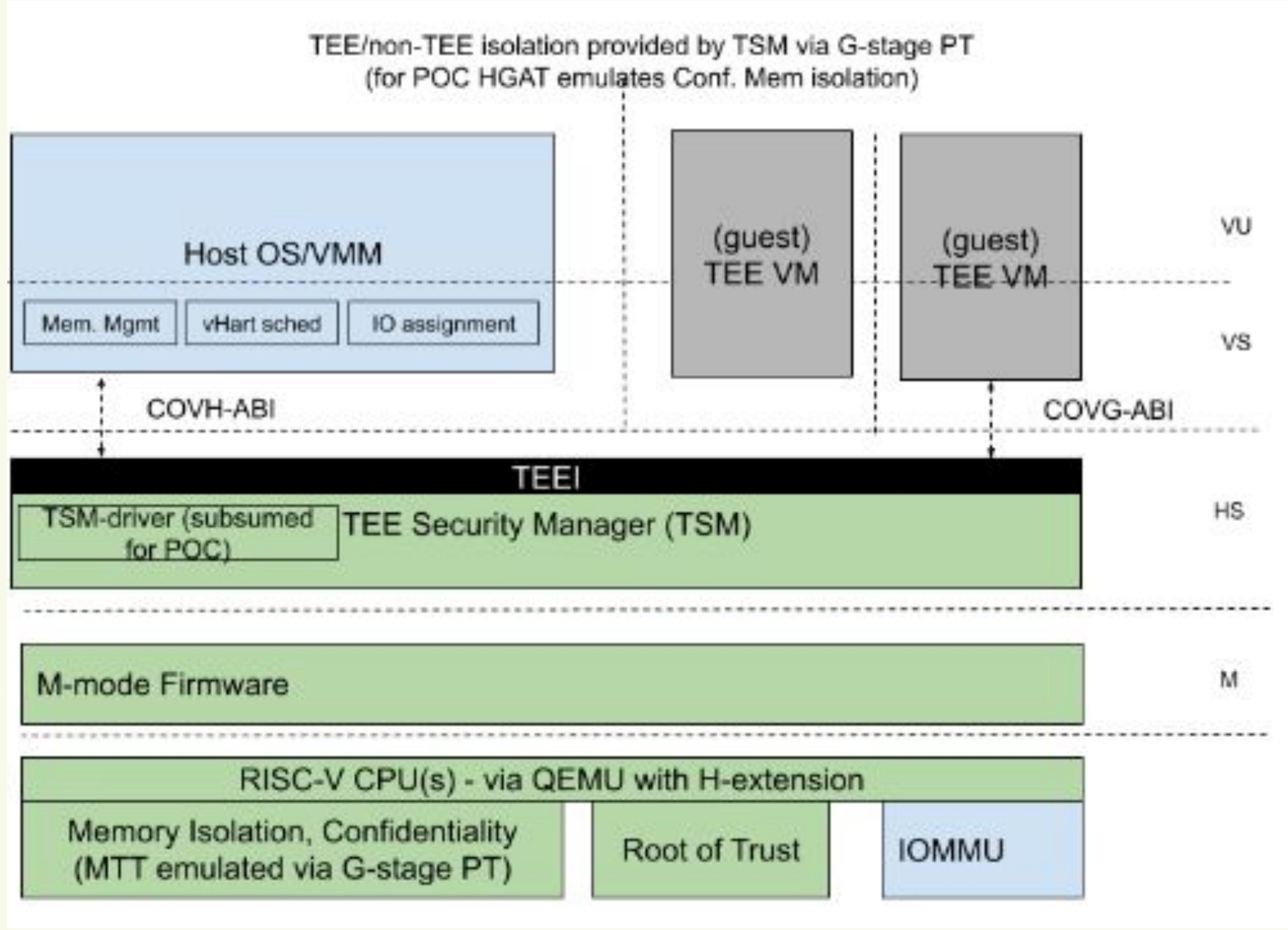
Current POC for CoVE ABI

First POC

Datacenter Confidential Computing … Client/Embedded TEE

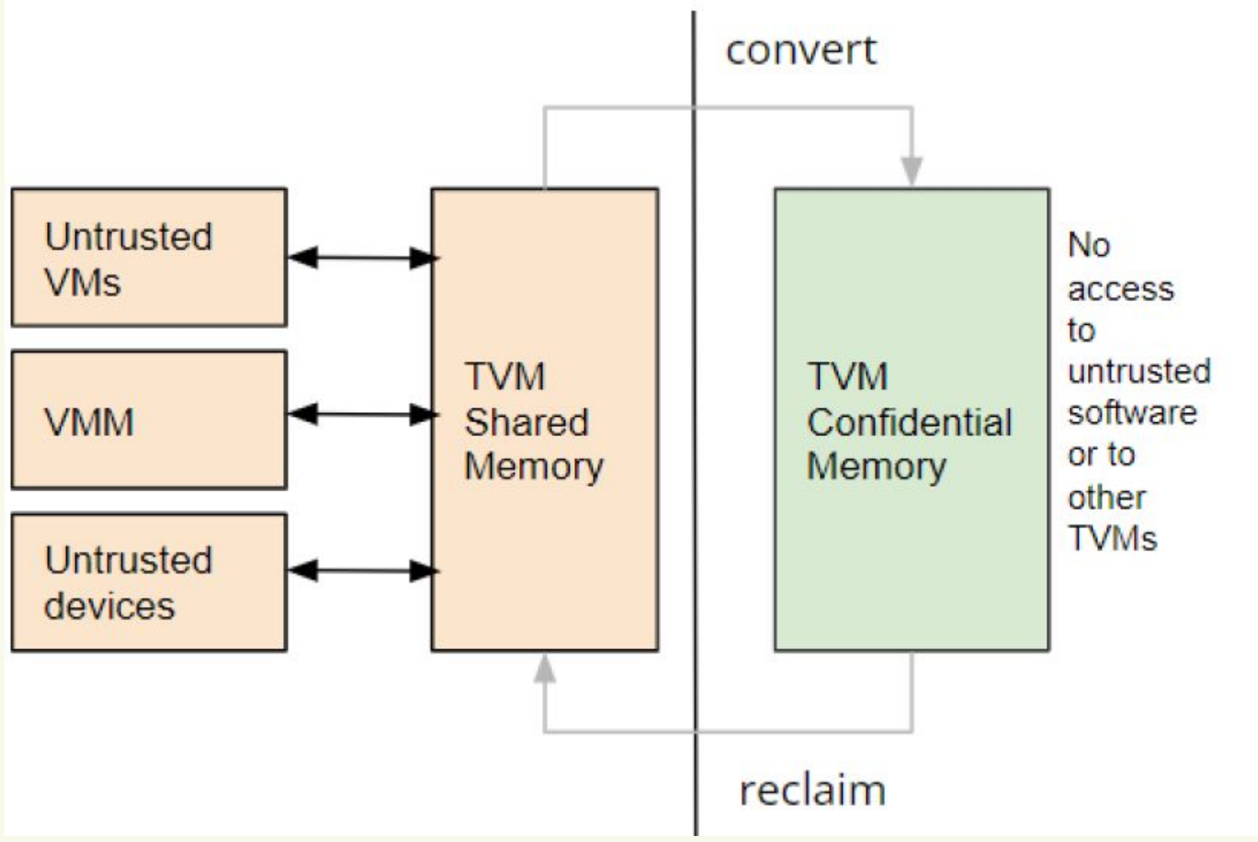

TEE/non-TEE isolation provided by CPU e.g. MTT
Intra TEE isolation enforced by TSM
(guest) Non-confidential VMs
(guest) Non-confidential VMs
(guest) Non-confidential VMs
(guest) TEE VM
(guest) TEE VM
VU
VS
Legacy ABI
COVG-ABI
COVE ABI
Mem. Mgmt
Host OS/VMM scheduling
IO assignment
TEE Security Manager (TSM)
HS
ECALL/MRET
COVEH-ABI
ECALL/MRET
M-mode Firmware
TSM-driver
(Measurement, Attestation, Context-switch)
M
RISC-V CPU(s)
Memory Isolation, Confidentiality, Integrity and Replay-protection
Root of Trust
IOMMU
🟩 In TVM TCB
🟦 In Non-confidential VM TCB

Common CoVE ABI to support both deployment models with incremental ISA requirements


TEE/non-TEE isolation provided by TSM via G-stage PT
(for POC HGAT emulates Conf. Mem isolation)
Host OS/VMM
Mem. Mgmt
vHart sched
IO assignment
(guest) TEE VM
(guest) TEE VM
VU
VS
COVH-ABI
COVG-ABI
TEEI
TSM-driver (subsumed for POC)
TEE Security Manager (TSM)
HS
M-mode Firmware
M
RISC-V CPU(s) - via QEMU with H-extension
Memory Isolation, Confidentiality (MTT emulated via G-stage PT)
Root of Trust
IOMMU

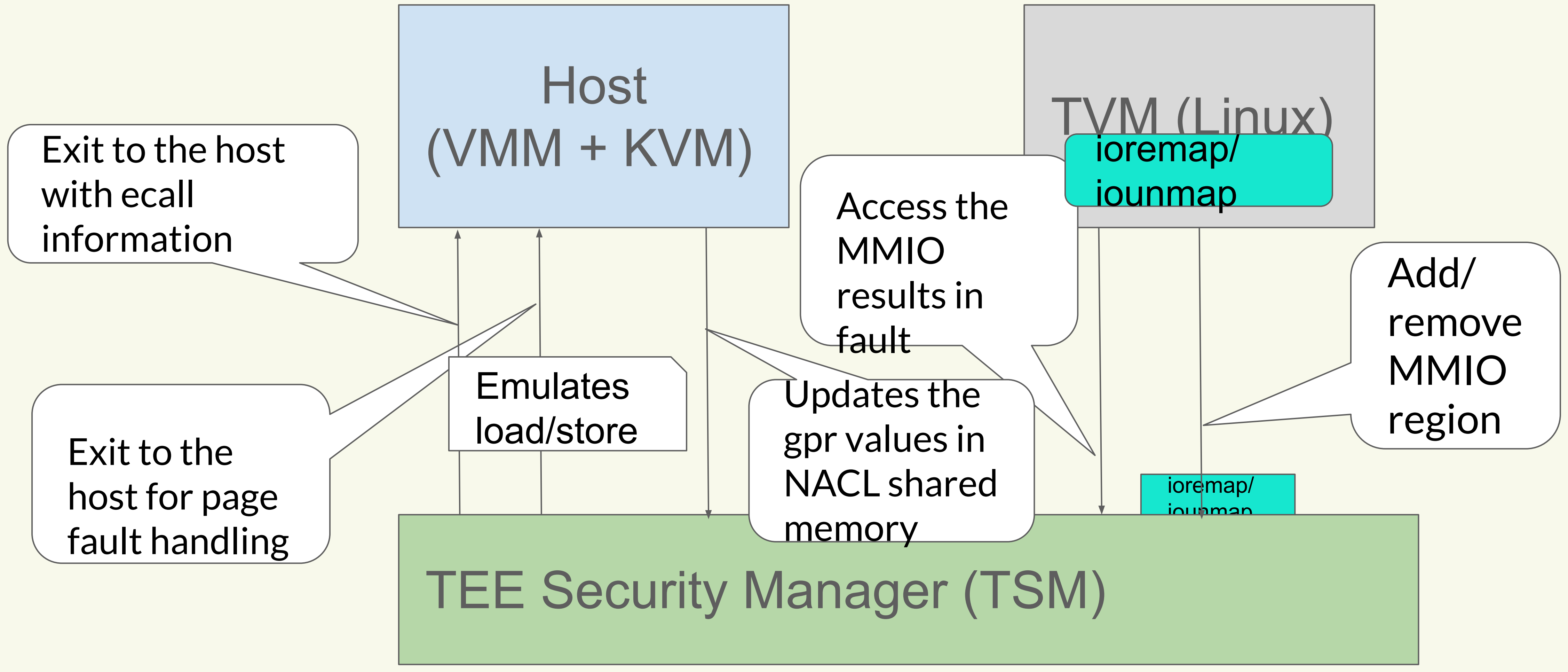
Overall design and current status

- All memory is non-confidential by default
- Explicit memory region conversion marking confidential guest memory region
- Explicit host physical pages conversion via *sbi_covh_convert_pages*
  - Page table management
  - TVM state management
  - On demand zero pages
  - Boot time measured pages (guest code & data payload)
- TLB management by via sbi_covh_[global/local]_fence after conversion
- Once converted, TSM can repurpose them for other TVMs until reclaim
- kvmtool support for now
  - Cross-vm and Qemu-kvm will be supported in the future version
  - guest_memfd implementation once qemu-kvm support is available
- Only one additional ABI added in the first version required by the spec
  - **KVM_RISCV_COVE_MEASURE_REGION**
  - Any unification plans in this area ?



MMIO

- Explicit MMIO region registration from TVM via *sbi_covg_[add|remove]_mmio_region* at runtime
- TSM is notified about MMIO region at runtime
- TSM forwards the fault to the host if request falls within the guest defined region
- Host emulates the MMIO load/store for TVM



Host (VMM + KVM)

TVM (Linux)

ioremap/ iounmap

Exit to the host with ecall information

Access the MMIO results in fault

Add/ remove MMIO region

Emulates load/store

Exit to the host for page fault handling

Updates the gpr values in NACL shared memory

ioremap/ iounmap

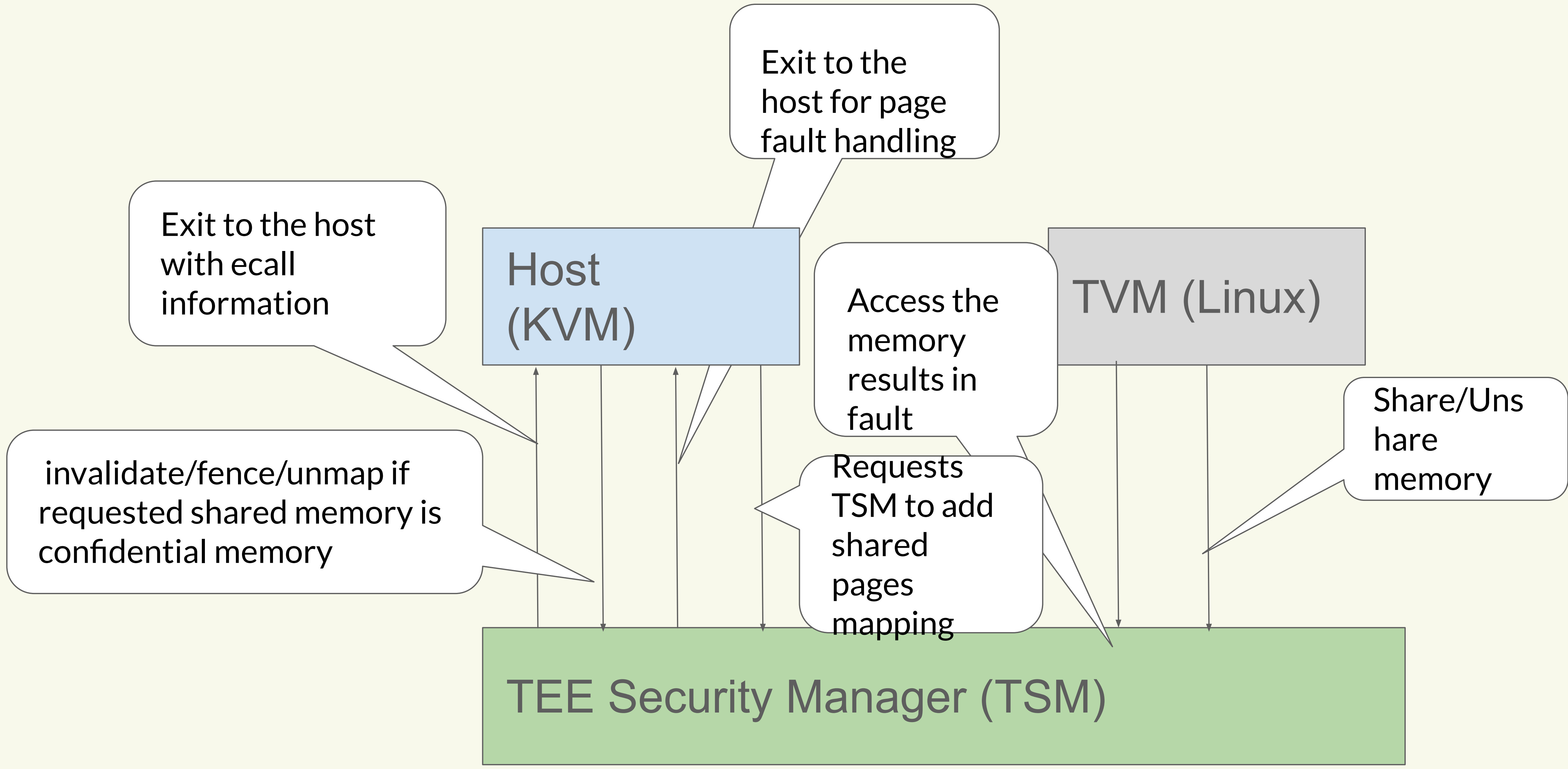
TEE Security Manager (TSM)

Opens

- **Explicit ABI call(current spec) vs PTE bits to indicate I/O pages (suggested on lore)**
- **Device filtering**
  - All io-remapped memory converted to mmio region (implemented in 1st version)
  - Authorized devices via device filtering approach (based on last TDX patches)

Device IO

- Only paravirt I/O devices supported in the first version
- Guest initiates swiotlb bounce buffer sharing via *sbi_covg_[share/unshare]_memory_region*
- Arch hooks under mem_encrypt/decrypt
- On-demand mapping via *sbi_covh_add_tvm_shared_pages* at fault time


Exit to the host for page fault handling
Exit to the host with ecall information
Host (KVM)
Access the memory results in fault
TVM (Linux)
Share/Unshare memory
invalidate/fence/unmap if requested shared memory is confidential memory
Requests TSM to add shared pages mapping
TEE Security Manager (TSM)

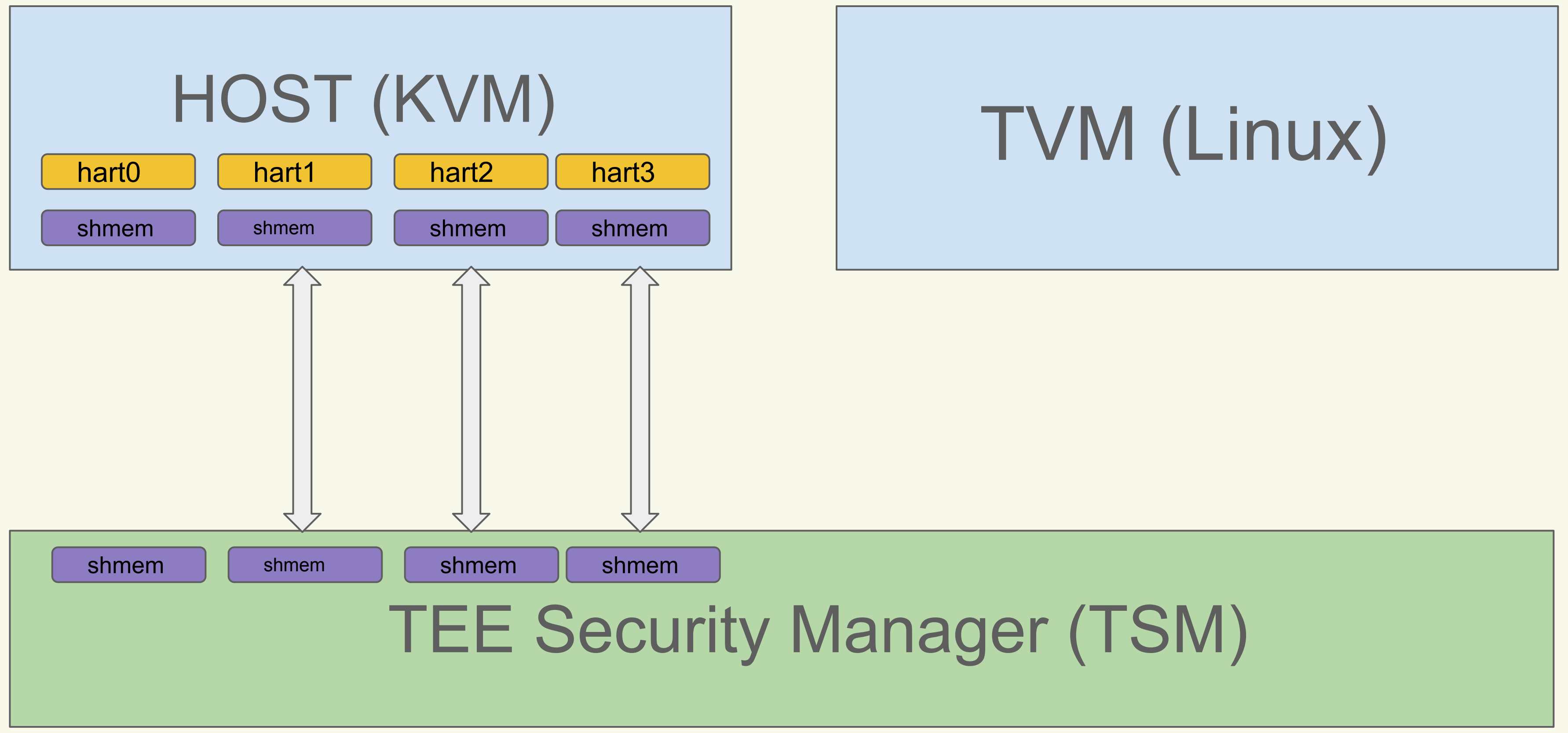
Opens

- **Explicit ABI call (current spec)**

  vs
- **PTE bits to indicate private/shared (suggested on lore)**
  - **See current proposed Svpams extension in Smmtt spec.**
  - **May require additional accept ABI for guest ?**

Shared memory between host and TSM

- Host in VS mode results in traps for accessing hypervisor level CSRs
- Leverages **RISC-V Nested Acceleration (NACL) SBI extension** with some additional security rules
  - Shared memory between the L0(e.g TSM) & L1(KVM) hypervisor per host cpu
- CoVE ABI defines the CSR & GPR state & access available in NACL shared memory
  - Trap related CSRS (htinst, htval)
  - Guest time management CSRs (htimedelta & vstimecmp)
  - Guest interrupt enable state (vsie)
  - GPRs to manage to TVM exits and MMIO loads
- ***TSM Ignores any updates from (untrusted) host to the CSRs and non-writable GPRs***



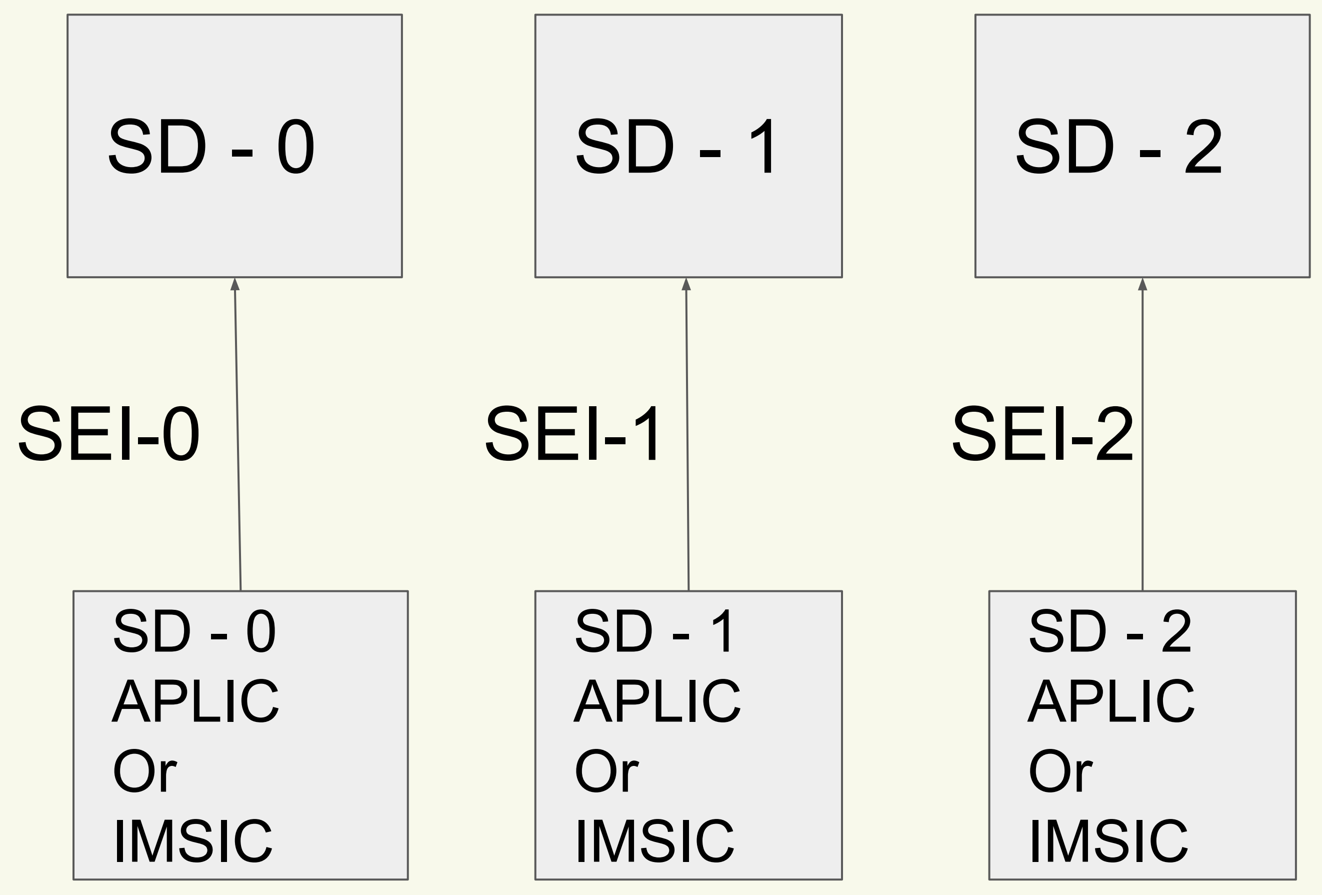
NACL Shared Memory Layout	
Offset	**Description**
0x0000-0x0FFF	Scratch space (4 KB) defined by CoVE to contain GPRs
0x1000 onwards	H extension CSR space (1024 x (XLEN / 8) Bytes)

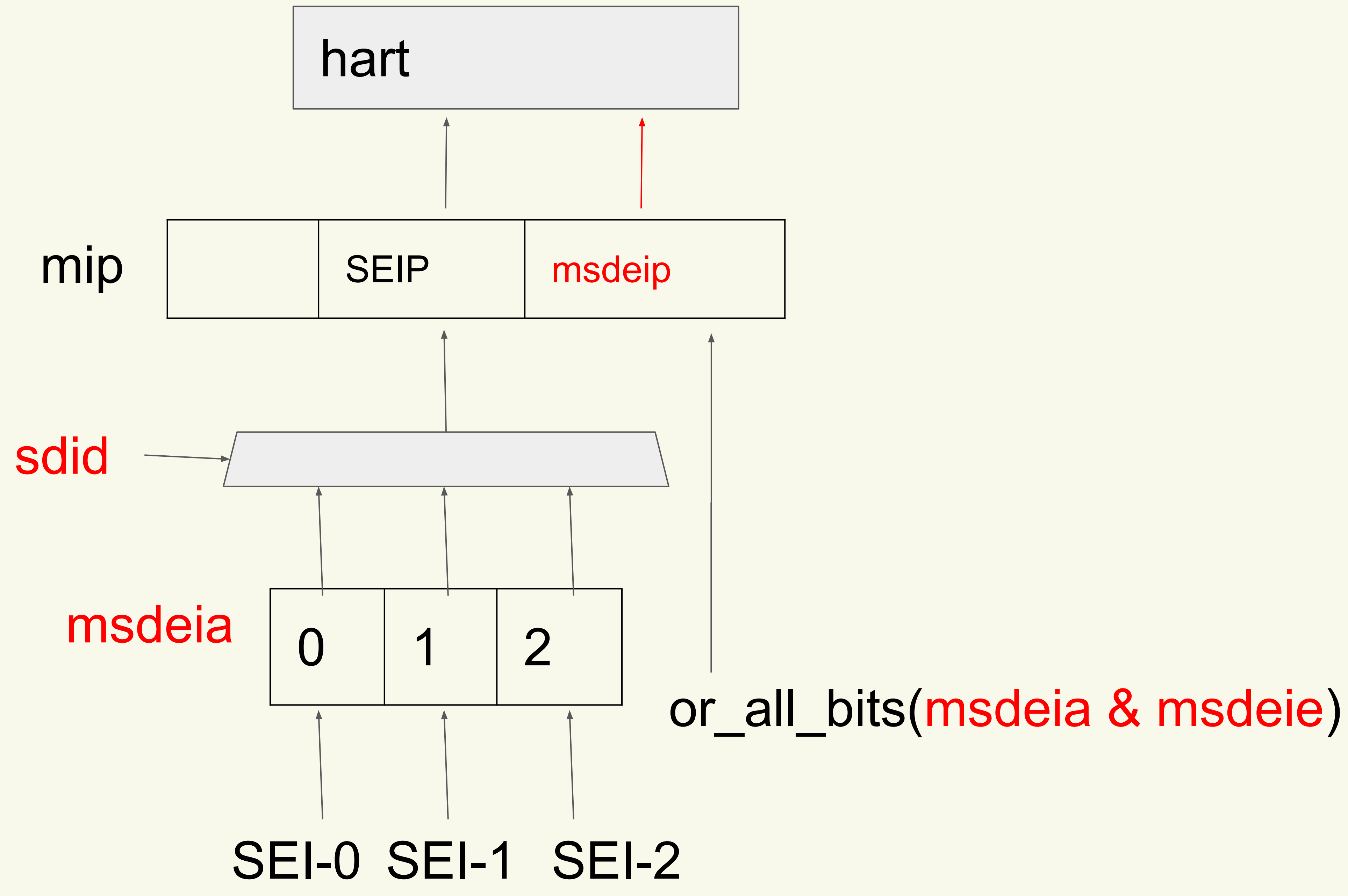
Interrupt assignment to confidential domain

- **Timer Interrupt**
  - Relies on Sstc extension which allows direct timer interrupt injection to guest
  - vstimecmp (guest time compare CSR) read access for host but updates ignored
  - TSM’s responsibility to restore it while switching back to TVM
  - htimedelta updates once at TVM bootime by TSM. Host reads via NACL shared memory

- **IPI/External Interrupt**
  - Only MSI based interrupt via RISC-V AIA specification
  - Direct vs-level interrupts possible by VS interrupt file update
  - Host domain can request confidential domain to inject allowed interrupts to TVM via ***sbi_covi_inject_tvm_cpu***
  - Host is responsible for convert/reclaim of VS interrupt file
  - Each vcpu migration causes interrupt files to be bind/unbind/rebind
  - This may result in co-ordinated TLB invalidation as well
  - (See next slide on supervisor domain S-mode interrupt file assignment)

ISA: Smsdia: Interrupt file assignment to supervisor domains





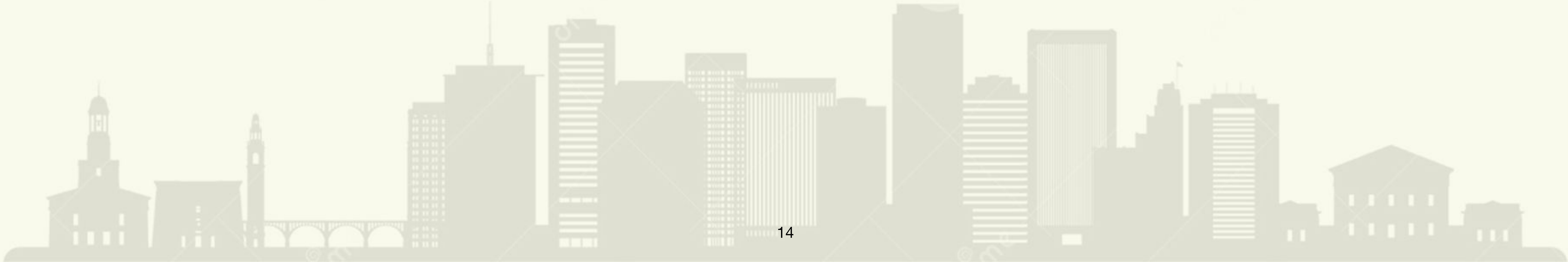
- Supervisor domains need not be aware of other domains
- Each domain is associated with a dedicated IMSIC S-interrupt file or an APLIC
  - Each domain interacts with its interrupt controller instance as defined - no change to interface/behavior
- MTT or PMP used to restrict a supervisor domain (and its devices) to its memory mapped registers
- Wires connected statically or configurable to APLICs

- All SD external interrupts reflected in a new M-mode CSR - msdeia (expected to be a 64 bit CSR allowing for 64 supervisor domains that can have external interrupts assigned)
- SDID selects the external interrupt for the hart from msdeia
- M-mode can be interrupt using a new local interrupt - msdeip - that indicates if *any* of the SD external interrupts are pending - can be masked using new M-mode CSR msdeie
  - M-mode can use this interrupt for scheduling decisions
- IMSIC CSRs: siselect, sireg, stopei - operate on S-mode register file selected per SDID

Q/A & Thanks

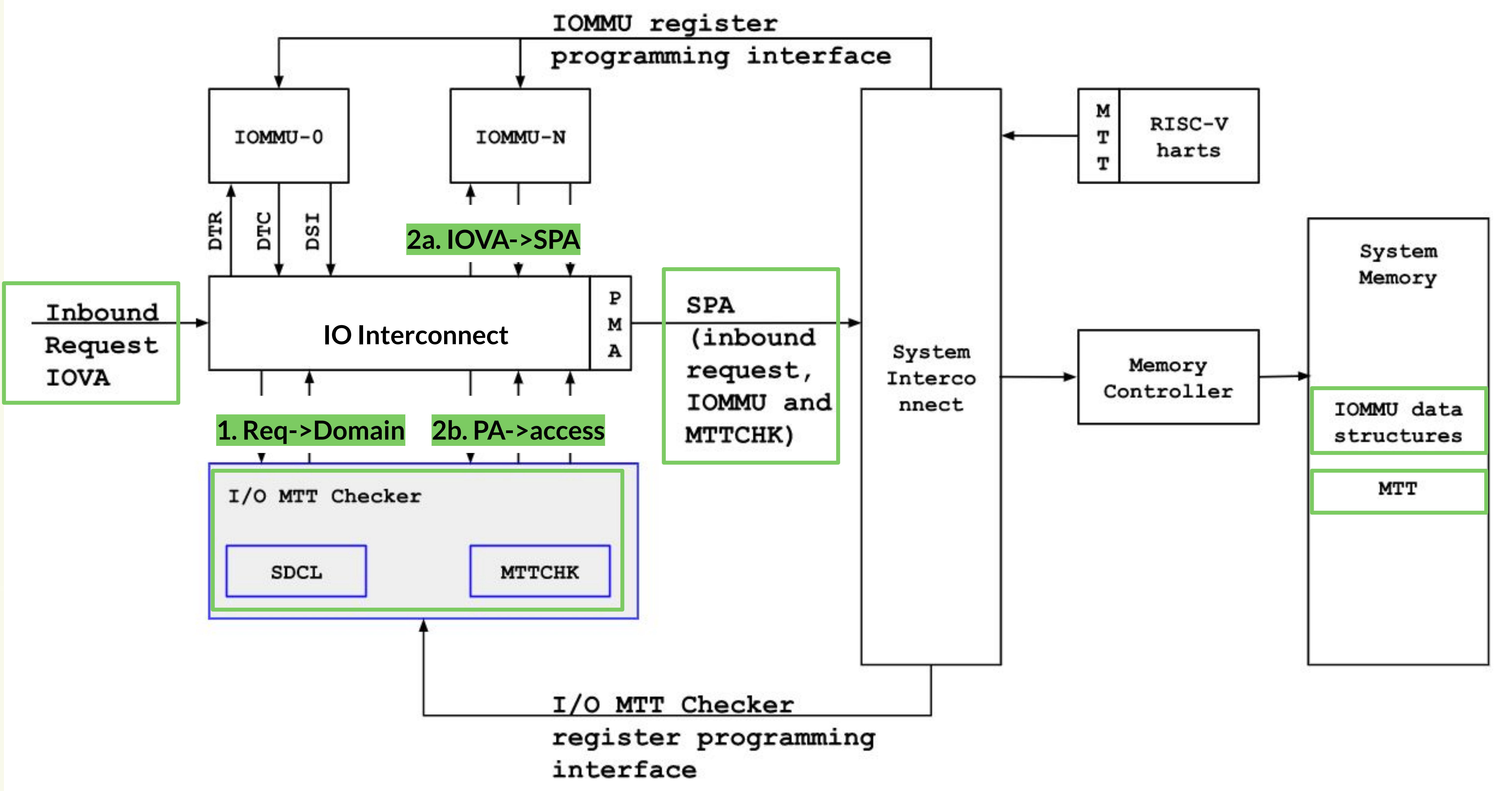
Linux Plumbers Conference

Richmond, Virginia | November 13-15, 2023



Non-ISA: IO-MTT: IO assignment to supervisor domains

- Smmtt TG is discussing an IO-MTT interface to enable device requests to map to a supervisor domain (and hence an IOMMU and an MTT)
- Uses ratified RISC-V IOMMU with MTT checker
  - PA resolved by IOMMU are checked by MTT checker
- RISC-V CoVE-IO TG defining ABI for TEE-IO:
  - Device attestation
  - Device assignment to CoVE TVM


IOMMU register programming interface
IOMMU-0
IOMMU-N
DTR
DTC
DSI
2a. IOVA->SPA
MTT
RISC-V harts
System Memory
Inbound Request IOVA
IO Interconnect
PMA
SPA (inbound request, IOMMU and MTTCHK)
System Interconnect
Memory Controller
IOMMU data structures
MTT
1. Req->Domain
2b. PA->access
I/O MTT Checker
SDCL
MTTCHK
I/O MTT Checker register programming interface