Confidential Virtual Machine Design Survey

Physical Memory Isolation Part

Dingji Li
IPADS, SJTU

2022-03-01

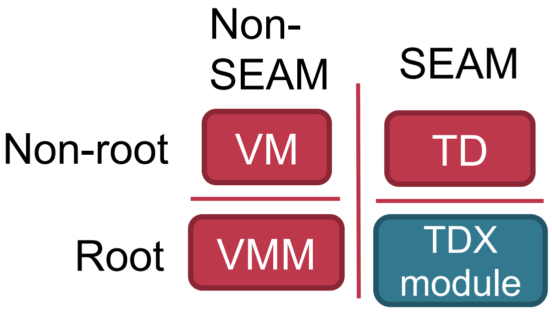
Existing Memory Isolation for CVMs

- **x86-64 platforms**
  - Physical Address Metadata Table (Intel TDX)
  - Reverse Map Table (AMD SEV-SNP)
- **ARMv9 platforms**
  - Granule Protection Table (ARM CCA)

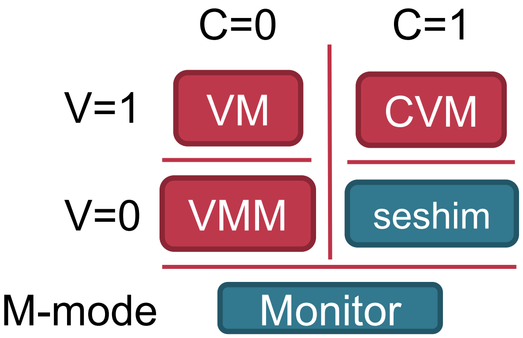
- **What kinds of metadata are necessary?**
- **What can we learn from them?**

Intel TDX

- SEAM mode for TDs and TDX module
- Software TCB: TDX module (root SEAM mode)
- Isolation between SEAM and non-SEAM modes



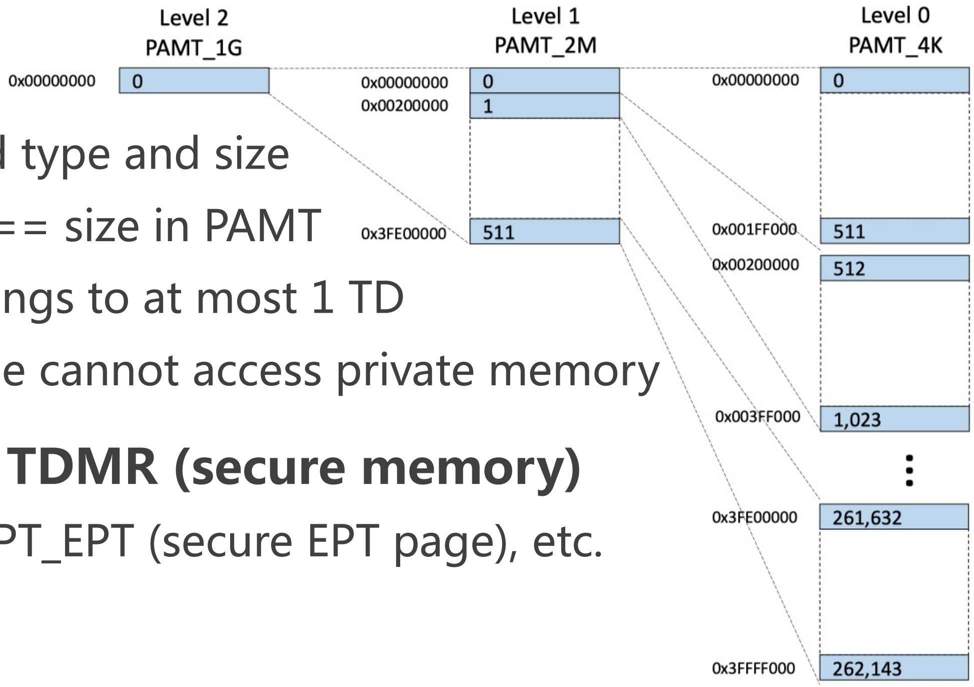
Intel TDX



RISC-V CVM

Physical Address Metadata Table (PAMT)

- **Hardware-enforced properties**
  - Attributes: page has well-defined type and size
  - S-EPT consistency: size in S-EPT == size in PAMT
  - Single TD assignment: page belongs to at most 1 TD
  - Access isolation: non-SEAM mode cannot access private memory
- **Hold metadata of each page in TDMR (secure memory)**
  - Type: PT_REG (TD private page), PT_EPT (secure EPT page), etc.
  - Size: 1GB, 2MB, 4KB
  - Page owner (by only one TD)

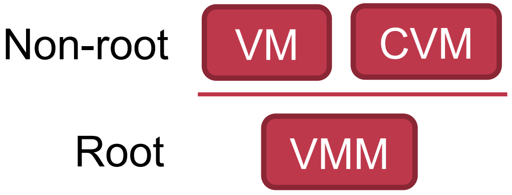


3-level table

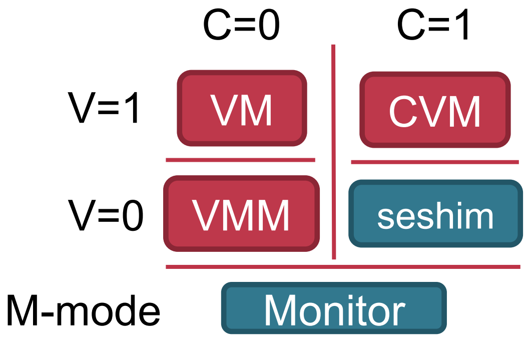
Ref: https://www.intel.com/content/dam/develop/external/us/en/documents/intel-tdx-module-1.5-base-spec-348549001.pdf

AMD SEV-SNP

- No special modes for CVMs
- Pure hardware TCB: AMD-SP
- Isolation between hypervisor/VM and CVMs



AMD SEV-SNP

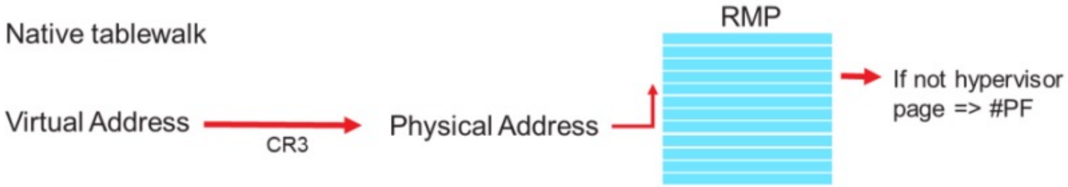


RISC-V CVM

Reverse Map Table (RMP)

- **Hardware-enforced properties**
  - Attributes: page has well-defined type and size
  - NPT consistency: size in NPT == size in RMP
  - Single CVM assignment: ASID + unique HPA → GPA mapping
  - Write isolation: host OS cannot write to private memory
- **RMP entries contain the metadata of each secure page**
  - Page type (host/guest), page size, GPA, ASID, permission, etc.
  - Supported size: 2MB, 4KB

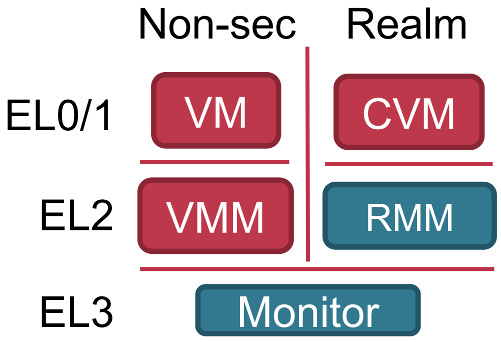
flat table



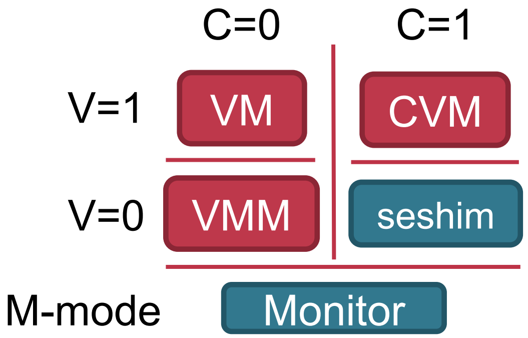
Ref: https://www.amd.com/system/files/TechDocs/24593.pdf

ARM CCA

- **Realm mode for CVMs and Realm manager (RMM)**
- **Software TCB: RMM (R-EL2) & secure monitor (EL3)**
- **Isolation between Non-secure and Realm modes**



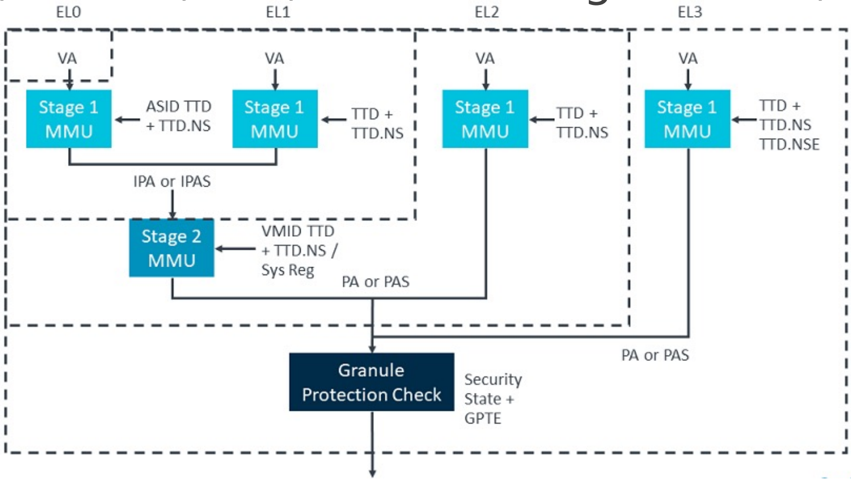
ARM CCA



RISC-V CVM

Granule Protection Table (GPT)

- **Hardware-enforced properties**
  - Access isolation: exclusive access from at most one CPU mode
- **Hold Granule Protection Information (GPI) for each physical page**
  - Granules (GPI for 16 pages) / Contiguous (GPI + contig size)
  - 4-bit GPI: Sec,Non-sec,Root,**Realm** / Contig size: 2MB, 32MB, 512MB



2-level table

Ref: https://developer.arm.com/documentation/den0125/latest/Arm-CCA-Hardware-Architecture

Summary

- **Hardware-enforced page-level isolation at 3rd-stage**
  - Necessary metadata: page size, page type
  - Optional: ownership, Key ID/ASID

Solution	HW Checks
Intel TDX	1. In SEAM mode 2. Correct Key ID
AMD SEV-SNP	1. Correct ASID
ARM CCA	1. In Realm mode

Hardware checks when accessing private memory of CVM

Solution	Ownership checks
Intel TDX	HW: owner in PAMT SW: TDX module
AMD SEV-SNP	HW: ASID in RMP
ARM CCA	SW: Realm manager

Methods to ensure single CVM page assignment

Discussion

- **Simple metadata is sufficient for the 3rd-stage memory isolation**
  - Page size (4KB, 2MB), page type (C, non-C)
  - Reduce hardware complexity (especially at a early stage)
  - Ownership can be achieved by the software TCB (seshim)

- **Should 3rd-stage table be independent from (3rd-stage) PMP?**
  - Both input HPA and output permissions
  - If independent
    - Which should be checked first? Or simultaneously?
    - What about the hardware complexity and memory latency?

Thanks!