Intel® Trust Domain Extensions (Intel® TDX) Module Base Architecture Specification

348549-006US

April 2025

.

Notices and Disclaimers

Table of Contents

SECTION 1: INTRODUCTION AND OVERVIEW

1. About this Document

1.1. Scope of this Document

This document describes the architecture of the Intel® Trust Domain Extensions (Intel® TDX) module, implemented using the Intel TDX Instruction Set Architecture (ISA) extensions, for confidential execution of Trust Domains in an untrusted hosted cloud environment.

This document is part of the **TDX Module Architecture Specification Set**, which includes the following documents:

Table 1.1: TDX Module Architecture Specification Set

Document Name	Reference	Description
TDX Module **Base Architecture Specification**	[TDX Module Base Spec]	Base TDX module architecture overview and specification, covering key management, TD lifecycle management, memory management, virtualization, measurement and attestation, service TDs, debug aspects etc.
TDX Module **TD Migration Architecture Specification**	[TD Migration Spec]	Architecture overview and specification for TD migration
TDX Module **TD Partitioning Architecture Specification**	[TD Partitioning Spec]	Architecture overview and specification for TD Partitioning
TDX Module **TDX Connect Specification**	[TDX Connect Spec]	Architecture overview and specification for TDX Connect
TDX Module **ABI Reference Specification**	[TDX Module ABI Spec]	Detailed TDX module Application Binary Interface (ABI) reference specification, covering the entire TDX module architecture
TDX Module **TDX Connect ABI Reference Specification**	[TDX Connect ABI Spec]	Detailed TDX module Application Binary Interface (ABI) reference specification, covering the TDX connect architecture
TDX Module ABI Reference Tables	[TDX Module ABI Tables]	A set of JSON format files detailing TDX module Application Binary Interface (ABI)
TDX Module ABI Incompatibilities	[TDX Module ABI Incompatibilities]	Description of the incompatibilities between TDX 1.0 and TDX 1.4/1.5 that may impact the host VMM and/or guest TDs

This document is a work in progress and is subject to change based on customer feedback and internal analysis. This document does not imply any product commitment from Intel to anything in terms of features and/or behaviors.

Note: The contents of this document are accurate to the best of Intel's knowledge as of the date of publication, though Intel does not represent that such information will remain as described indefinitely in light of future research and design implementations. Intel does not commit to updating this document in real time when such changes occur.

1.2. Document Organization

The document has the following main sections:

- Section 1 contains an introduction to the document and an overview of the Intel TDX module.
- Section 2 contains the Intel TDX module architecture specification.

1.3. Glossary

Table 1.2: Intel TDX Glossary

Acronym	Full Name	New for TDX	Description
ABI	**Application Binary Interface**	No	A programming interface defined at the binary level (i.e., instruction opcode and CPU registers). The Intel TDX module interface is specified as an ABI.
ACM	**Authenticated Code Module**	No	A code module that is designed to be loaded, verified and executed by the CPU in on-chip memory (CRAM).
ACT	**Access Control Table**	Yes	A table in memory which controls memory access. Each bit in the table represents the state of a 4K memory page: shared or private.
N/A	**Accessible (Memory)**	No	Memory whose content is readable and/or writeable (e.g., TD private memory is accessible to the guest TD).
N/A	**Addressable (Memory)**	No	Memory that can be referred to by its address. The content of addressable memory might not necessarily be accessible (e.g., TDCS is not accessible to the host VMM).
CMR	**Convertible Memory Range**	Yes	A range of physical memory configured by BIOS and verified by MCHECK. MCHECK verification is intended to help ensure that a CMR may be used to hold TDX memory pages encrypted with a private HKID.
N/A	**Enlightened OS**	No	A TD OS is considered enlightened if it is aware that it is running as a TD (see Paravirtualization).
EPxE	**Extended Paging Structures Cache**	No	The CPU's cache of EPT intermediate translations (as opposed to TLB, which caches full LA or GPA to HPA translations).
GPA	**Guest Physical Address**	No	An address viewed as a physical address, from a guest VM's point of view. A GPA is subject to further translation (by EPT) to produce an HPA.
N/A	**Hidden**	No	A resource or a data structure that is not directly addressable by software (except the Intel TDX module).
HKID	**Host Key ID**	Yes	When MKTME is activated, HKID is a key identifier for an encryption key used by one or more memory controllers on the platform.
N/A	**Host VMM**	Yes	The VMM that serves as a host to guest TDs. The term "host" is used to differentiate between the "host VMM" and future VMMs that may be nested within TDs.
HPA	**Host Physical Address**	No	A physical address at the host VMM level. This is the actual physical address used by the hardware (e.g., caches). See also PA.
KD	**Key Domain**	Yes	Represents the control state associated with an ephemeral TDX key resource. Key domains are managed as a resource by the host VMM with the security attributes of the lifecycle of a key domain for Trust Domains (TDs) is enforced by the Intel TDX Module. A TD is assigned to a single Key Domain.
KET	**Key Encryption Table**	Yes	A table held by each MKTME encryption engine, intended for holding encryption key information, indexed by HKID.
KOT	**Key Ownership Table**	Yes	An internal, hidden table held by the Intel TDX module, intended for controlling the assignment of HKIDs to TDs.

Acronym	Full Name	New for TDX	Description
MBZ	**Must Be Zero**	No	Normally used to indicate that reserved fields must contain 0.
MKTME	**Multi-Key TME**	No	This SoC capability adds support to the TME to allow software to use one or more separate keys for encryption of volatile or persistent memory encryption. When used with TDX, it can provide confidentiality via separate keys for memory used by TDs. MKTME can be used with and without TDX extensions.[1]
MRTD	**Measurement of Trust Domain**	Yes	The SHA-384 measurement of a TD accumulated during TD build.
NP-SEAMLDR	**Non-Persistent SEAM Loader**	Yes	An ACM intended to load an Intel P-SEAMLDR module into the SEAM range.
NRX	**Non-Root Extension**	Yes	An extension of the TDX module, which consists of the NRX framework that runs as part of the TDX module in SEAM root mode, and NRX TDs that run in SEAM non-root mode.
P-SEAMLDR	**Persistent SEAM Loader**	Yes	A SEAM module intended to install (load or update) Intel TDX modules into SEAM range.
PA	**Physical Address**	No	The physical address used by the hardware (e.g., caches). See also HPA.
PAMT	**Physical Address Metadata Table**	Yes	An internal, hidden data structure used by the Intel TDX module, which is intended to hold the metadata of physical pages.
PV	**Para-Virtualization**	No	A virtualization technique where the VM can be aware it is being virtualized (as opposed to running directly on hardware).
RTMR	**Run-Time Measurement Register**	Yes	A SHA-384 measurement register that can be updated during TD run-time.
SEAM	**Secure Arbitration Mode**	Yes	See TDX ISA.
SEAMRR	**SEAM Range Register**	Yes	A range register used by BIOS to help configure the SEAM memory range, where the Intel TDX module is loaded and executed.
Service TD	**Service TD**	Yes	A Trust Domain (TD) VM used to provide a dedicated service/utility. Extends the TCB of the tenant TD for which it provides the service. Migration TD (MigTD) is an example Service TD.
SEPT	**Secure EPT**	Yes	An Extended Page Table for GPA-to-HPA translation of TD private HPA. A Secure EPT is designed to be encrypted with the TD's ephemeral private key. SEPT pages are allocated by the host VMM via Intel TDX functions, but their content is intended to be hidden and is not architectural.

[1] In this document, the term "MK-TME" is used to mean both the feature and the encryption engine itself.

Acronym	Full Name	New for TDX	Description
Intel® SGX	**Intel® Software Guard Extensions**	No	An Intel CPU mode and ISA extensions that support operation and management of Intel® SGX enclaves.
SoC	**System on Chip**	No	A whole system, including cores, uncore, interconnects etc., packaged as a single device.
SPA	**System Physical Address**	No	The physical address used by the hardware (e.g., caches). See also HPA.
TD	**Trust Domain**	Yes	Trust Domains (TDs) are designed to be hardware isolated Virtual Machines (VMs) deployed using Intel® Trust Domain Extensions (Intel® TDX).
TD OS	**Trust Domain Operating System**	Yes	The guest operating system that runs in a TD.
TD VM	**TD Virtual Machine**	Yes	Same as TD
N/A	**TD Private Memory (Access)**	Yes	TD Private Memory is designed to hold TD private content, encrypted by the CPU using the TD ephemeral key.
N/A	**TD Shared Memory (Access)**	Yes	TD Shared memory is designed to hold content accessible to the TD and the host software (and/or other TDs). TD shared memory may be encrypted using MKTME keys managed by the VMM.
TDCS	**Trust Domain Control Structure**	Yes	Multi-page control structure for a TD. TDCS pages are allocated by the host VMM via Intel TDX functions, but their content is intended to be non-architectural and not directly accessible to software.
TDCX	**Trust Domain Control Extension**	Yes	4KB physical pages that are intended to hold parts of a multi-page control structure.
TDR	**Trust Domain Root**	Yes	The root control structure for a TD. The TDR page is allocated by the host VMM via Intel TDX functions, but its content is intended to be non-architectural and not directly accessible to software.
TDMR	**Trust Domain Memory Range**	Yes	A range of memory, configured by the host VMM, that is covered by PAMT and is intended to hold TD private memory and TD control structures.
TDVPS	**Trust Domain Virtual Processor State**	Yes	A multi-page structure for holding a TD Virtual CPU (VCPU) state. TDVPS pages are allocated by the host VMM via Intel TDX functions, but their content is intended to be non-architectural and not directly accessible to software.
TDVPR	**Trust Domain Virtual Processor Root**	Yes	A 4KB physical page that is intended to be the root (first) page of a TDVPS.
Intel® TDX	**Intel® Trust Domain Extensions**	Yes	An architecture, based on the TDX Instruction Set Architecture (ISA) extensions and the Intel TDX module, which supports operation and management of Trust Domains.

Acronym	Full Name	New for TDX	Description
TDX ISA	**Intel® TDX Instruction Set Architecture**	Yes	Intel CPU Instruction Set Architecture (ISA) extensions that support the Intel TDX module: an isolated software module that facilitates the operation and management of Trust Domains.
TEE	**Trusted Execution Environment**	No	An isolated processing environment in which software can be securely executed irrespective of the rest of the system.
TME	**Intel® Total Memory Encryption**	No	A memory encryption/decryption engine using an ephemeral platform key designed to encrypt memory contents exposed externally from the SOC.
N/A	**WBINVD Domain**	No	A set of LPs for which a single WBINVD or WBNOINVD instruction, and the TDH.PHYMEM.CACHE.WB interface function, apply.
XFAM	**Extended Features Allowed Mask**	Yes	A mask of CPU extended features (in XCR0 format) that the TD is allowed to use.

1.4. Notation

This section describes the notation used in this document.

1.4.1. Requirement and Definition Commitment Levels

When specifying requirements or definitions, the level of commitment is specified following the convention of RFC 2119: Key words for use in RFCs to indicate Requirement Levels, as described in the following table:

Table 1.3: Requirement and Definition Commitment Levels

Keyword	Description
Must	"**Must**", "**Required**" or "**Shall**" means that the definition is an absolute requirement of the specification.
Must Not	"**Must Not**" or "**Shall Not**" means that the definition is an absolute prohibition of the specification.
Should	"**Should**", or the adjective "**Recommended**", means that there may exist valid reasons in particular circumstances to ignore a particular item, but the full implications must be understood and carefully weighed before choosing a different course.
Should Not	"**Should Not**", or the phrase "**Not Recommended**" means that there may exist valid reasons in particular circumstances when the particular behavior is acceptable or even useful, but the full implications should be understood, and the case must be carefully weighed before implementing any behavior described with this label.
May	"**May**", or the adjective "**Optional**", means that an item is discretionary. An implementation may choose to include the item, while another may omit the same item, because of various reasons.

1.5. References

1.5.1. Intel Public Documents

Table 1.4: Intel Public Documents

Reference	Document	Version & Date
Intel SDM	Intel® 64 and IA-32 Architectures Software Developer's Manual	325462-078US, December 2022
ISA Extensions	Intel® Architecture Instruction Set Extensions and Future Features Programming Reference	319433-047, December 2022
Error Reporting through EMCA2	RAS Integration and Validation Guide for the Intel Xeon Processor – Error Reporting through EMCA Gen 2	April 2015
Key Locker Spec	Intel Key Locker Specification	Sept 2020
FRED	Flexible Return and Event Delivery (FRED)	May 2022
Processor Topology Enumeration	Intel® 64 Architecture Processor Topology Enumeration	337015-002, April 2023

5 ### 1.5.2. Intel TDX Public Documents

Table 1.5: Intel TDX Public Documents

Reference	Document	Version & Date
TDX Web Page	Intel® Trust Domain Extensions (Intel® TDX) **Note:** Most documents below are on this web page	
TDX Overview	An introductory overview of the Intel TDX technology	February 2023
TDX Arch Extensions Spec	A specification of Intel CPU architectural support for Intel TDX	May 2021
TDX Loader Spec	A specification of how a VMM loads the Intel TDX Module on a platform	March 2022
MKTMEi Spec	Intel® Architecture Memory Integrity Specification	Rev. 1.0, March 2020
TDX Module Base Spec	Overview and base architecture specification of the Intel TDX Module version 1.5	March 2023
TD Migration Spec	Overview and architecture specification of the TD Migration feature of the Intel TDX Module version 1.5	March 2023
TD Partitioning Spec	Overview and Architecture Specification for TD partitioning of the TDX Module version 1.5	March 2023
TDX Module ABI Spec	Application Binary Interface (ABI) specification of the Intel TDX Module version 1.5	March 2023
TDX Module ABI Incompatibilities	Description of the incompatibilities between TDX 1.0 and TDX 1.4/1.5 that may impact the host VMM and/or guest TDs	March 2023

Reference	Document	Version & Date
TDX GHCI Spec	Specification of the software interface between the Guest OS (Tenant and Service TD VMs) and the VMM required for enabling Intel TDX version 1.5	July 2022
MigTD Design Guide	A design guide on how to design and implement a Migration TD for TDX 1.5 Live migration	October 2021
TDX Developers Guide	Intel® TDX Guidance for Developers	March 2023
TDX Guest Kernel Hardening	Intel® TDX Guest Kernel Hardening Documentation	March 2023

2. Overview of Intel® Trust Domain Extensions

Intel® Trust Domain Extensions (Intel® TDX) refers to an Intel technology that extends Virtual Machines Extensions (VMX) and Multi-Key Total Memory Encryption (MKTME) with a new kind of virtual machine guest called a **Trust Domain (TD)**. A TD runs in a CPU mode that is designed to protect the confidentiality of its memory contents and its CPU state from any other software, including the hosting Virtual Machine Monitor (VMM), unless explicitly shared by the TD itself.

The TDX solution is built using a combination of Intel® Virtual Machine Extensions (VMX) and Multi-Key Total Memory Encryption (MK-TME), as extended by the **Intel® Trust Domain Extensions Instruction Set Architecture (Intel TDX ISA)**. An attested software module called the **Intel TDX module** is designed to implement the TDX architecture.

The platform is managed by a TDX-aware **host VMM**. As shown in Figure 2.1 below, a host VMM can launch and manage both guest TDs and legacy guest VMs. The host VMM maintains all legacy functionality from the legacy VMs' perspective; it is restricted only with regard to the TDs it manages.

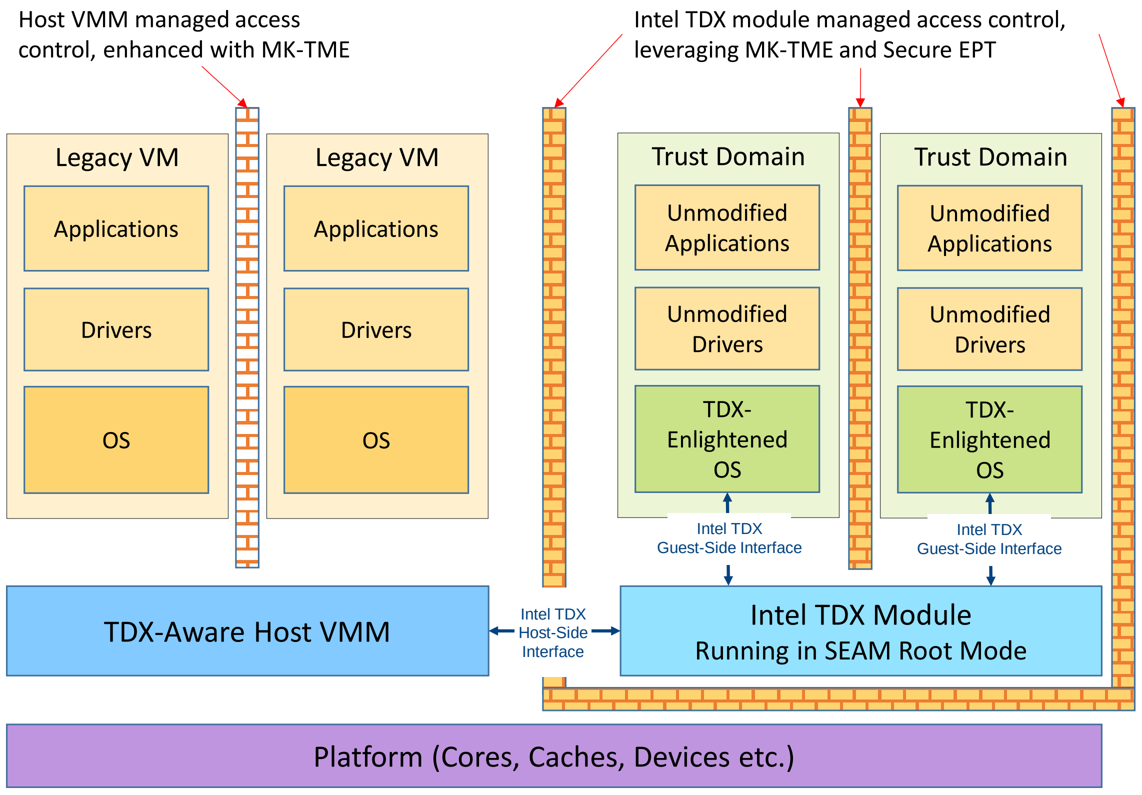

Host VMM managed access control, enhanced with MK-TME
Intel TDX module managed access control, leveraging MK-TME and Secure EPT
Legacy VM
Applications
Drivers
OS
Legacy VM
Applications
Drivers
OS
Trust Domain
Unmodified Applications
Unmodified Drivers
TDX-Enlightened OS
Trust Domain
Unmodified Applications
Unmodified Drivers
TDX-Enlightened OS
Intel TDX Guest-Side Interface
Intel TDX Guest-Side Interface
TDX-Aware Host VMM
Intel TDX Host-Side Interface
Intel TDX Module Running in SEAM Root Mode
Platform (Cores, Caches, Devices etc.)

Figure 2.1: Intel® Trust Domain Extension Components Overview

2.1. Intel TDX Module Lifecycle

2.1.1. Boot-Time Configuration and Intel TDX Module Loading

1. BIOS should activate MKTME with TDX private key IDs, configure the SEAMRR registers and prepares a table of **Convertible Memory Regions (CMRs)** – memory regions that can hold TD-private memory pages.
2. BIOS or OS should then install P-SEAMLDR by launching the NP-SEAMLDR ACM.
3. BIOS or OS can retrieve the trusted platform topology and CMR information, as previously checked by MCHECK, using P-SEAMLDR's SEAMINFO API. Based on the above, the host VMM should then decide on a set of Trust Domain Memory Regions (TDMRs). TDMR is a region of convertible memory that may contain some reserved sub-regions.
4. The host VMM can then load the Intel TDX module using P-SEAMLDR's INSTALL API.

2.1.2. Intel TDX Module Initialization, Enumeration and Configuration

1. After loading the Intel TDX module, the host VMM should call the TDH.SYS.INIT function to globally initialize the module.
2. The host VMM should then call the TDH.SYS.LP.INIT function on each logical processor. TDH.SYS.LP.INIT is intended to initialize the module within the scope of the Logical Processor (LP).

3. The host VMM should then call the TDH.SYS.RD/RDALL or TDH.SYS.INFO function to enumerate the Intel TDX module functionality and parameters; it should retrieve the trusted platform topology and CMR information as previously checked by MCHECK.
4. The host VMM should then call the TDH.SYS.RD/RDALL or TDH.SYS.INFO function to enumerate the Intel TDX module functionality and parameters. If it is not done already, the host VMM can retrieve the trusted platform topology and CMR information, as previously checked by MCHECK, and decide on the set of TDMRs.
5. The host VMM should then call the TDH.SYS.CONFIG function and pass TDMR information with other configuration information. TDH.SYS.CONFIG is intended to check the configuration information vs. the Intel TDX module's trusted internal data.
6. The host VMM should then call the TDH.SYS.KEY.CONFIG function per package. TDH.SYS.KEY.CONFIG is intended to configure a CPU-generated random key that is used as the Intel TDX module's global private key. On platforms with ACT-protected memory, TDH.SYS.KEY.CONFIG also enables ACT memory protection.
7. The host VMM should then use the TDH.SYS.TDMR.INIT function to initialize the TDMRs and their associated control structures.

The Intel TDX module lifecycle is detailed in Chapter 4.

2.2. Guest TD Life Cycle Overview

2.2.1. Guest TD Build

The host VMM can create a new guest TD by allocating and initializing a TD Root (TDR) control structure using the TDH.MNG.CREATE function. As an input to TDH.MNG.CREATE, the host VMM assigns the TD with a memory protection key identifier, also known as a Host Key ID (HKID). The HKID can be used by the CPU to tag memory accesses done by the TD and by the multi-key total memory encryption engines (MKTMEs) to select the encryption/decryption keys – the keys themselves are designed to not be exposed to the host VMM. The VMM should then program the HKID and encryption key into the MKTME encryption engines using the TDH.MNG.KEY.CONFIG function on each package.

Once keys are configured for the TD, the host VMM can build the TD Control Structure (TDCS) by adding control structure pages, using the TDH.MNG.ADDCX function, and initialize using the TDH.MNG.INIT function. It can then build the Secure EPT tree using the TDH.MEM.SEPT.ADD function and add the initial set of TD-private pages using the TDH.MEM.PAGE.ADD function. These pages typically contain Virtual BIOS code and data along with some clear pages for stacks and heap. Most of the guest TD code and data is dynamically loaded at a later stage. The guest TD can extend run-time measurement registers, designed to be securely maintained by the Intel TDX module, for the added contents using the TDG.MR.RTMR.EXTEND function.

The host VMM can then create and initialize TD Virtual CPUs (VCPUs). After creating each VCPU using the TDH.VP.CREATE function, the VMM allocates a set of pages to hold the VCPU state (in a structure called TDVPS) using the TDH.VP.ADDCX function. The host VMM can then initialize the VCPU using the TDH.VP.INIT function.

After the initial set of pages is added and extended, the VMM can finalize the TD measurement using the TDH.MR.FINALIZE function.

2.2.2. Guest TD Execution

The host VMM may enter the TD (launch the TD for the first time or resume a previously intercepted TD execution) using the TDH.VP.ENTER function. The Intel TDX module is designed to load CPU state from the TDVPS structure and perform VM entry to go into SEAM non-root mode.

When TD exit is triggered, the Intel TDX module is designed to save CPU state into the TDVPS structure, load the CPU state saved on TD entry, and switch back to non-SEAM root mode (SEAMRET) at the instruction following SEAMCALL. The VMM can then inspect the TD exit information in General Purpose Registers (GPRs).

2.2.3. Guest TD Management during its Run-Time

During TD lifetime, the VMM might need to dynamically control the TD and manage the resources assigned to it. The Intel TDX module provides the VMM with functions to support scenarios such as:

- Adding and removing TD pages.
- Changing page mapping sizes.
- Reclaiming the HKIDs from a TD and assigning them to another TD.
- Destroying an existing TD.

2.3. Intel TDX Operation Modes and Transitions

The Intel TDX module is designed to provide two main new **logical** modes of operation built upon the new SEAM root and non-root CPU modes added to the Intel VMX architecture: logical TDX Root Mode, and logical TDX Non-Root Mode. Figure 2.2 below shows the Intel TDX logical modes and transitions (in red) on top of the CPU architectural modes.

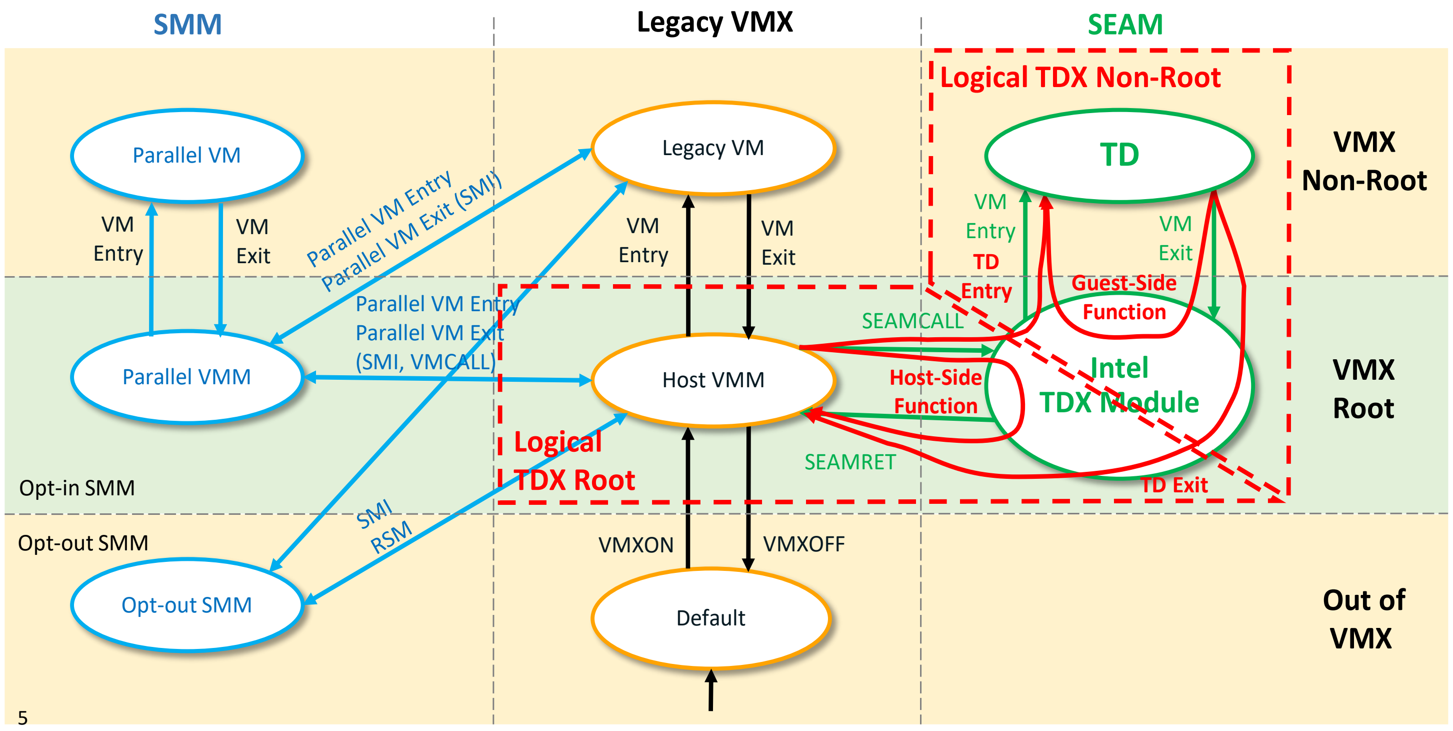


Figure 2.2: Overview of Intel TDX Modes & Transitions based on VMX and SEAM Modes and Transitions

The following table adds more details.

Table 2.1: Overview of Intel TDX Modes

Intel TDX Logical Mode	Intel VMX Mode	SEAM Mode	Description
Logical TDX Root	VMX Root	Non-SEAM (mostly), SEAM Root Mode (during host-side Intel TDX functions execution)	TDX root mode is mostly identical to the legacy VMX root operation mode. It is generally used for host VMM operations. Host-side Intel TDX functions, triggered by SEAMCALL, are provided by the Intel TDX module. Logically, host-side functions run in TDX root mode, though the CPU's SEAM mode is on.
Logical TDX Non-Root	VMX Non-Root (mostly), VMX Root (during guest-side Intel TDX flows execution)	SEAM	TDX non-root mode is used for TD guest operation. TDX non-root operation is similar to legacy VMX non-root operation, with changes and restrictions to better ensure that no other software or hardware has direct visibility of the TD memory and state. The changes in TDX non-root mode vs. legacy VMX non-root operation are implemented by: • The CPU is running in SEAM non-root mode. This modifies the address translation to support Secure EPT and usage of private HKIDs, and it also modifies the VMX operation (entry, exit, etc.). • The Intel TDX module, acting as the root VMM for the guest TD, uses VMX and SEAM to virtualize the CPU behavior and emulate the required TDX non-root behavior.

Intel TDX Logical Mode	Intel VMX Mode	SEAM Mode	Description
			Guest-side Intel TDX flows, triggered by a VM Exit, are provided by the Intel TDX module. Logically, guest-side functions run in TDX non-root mode, though the CPU runs VMX root mode. TDX non-root operation is described in Chapter 11.

Intel TDX transitions between TDX root operation and TDX non-root operation include **TD Entries,** from logical TDX root to logical TDX non-root mode, and **TD Exits** from logical TDX non-root to logical TDX root mode. A TD Exit might be **asynchronous**, triggered by some external event (e.g., external interrupt or SMI) or an exception, or it might be **synchronous**, triggered by a TDCALL(TDG.VP.VMCALL) function.

2.4. Guest TD Private Memory Protection

2.4.1.1. Memory Encryption

The Intel TDX module helps protect guest TD private memory using memory encryption and integrity protection as enabled by the CPU's MKTME and TDX ISA features. The Intel TDX module adds **key management functionality** to help enforce its security objectives.

Memory encryption is designed to be performed by encryption engines that reside at each memory controller. An encryption engine holds a table of encryption keys, known as the Key Encryption Table (KET). An encryption key is selected for each memory transaction based on a **Host Key Identifier (HKID)** that should be provided with the transaction.

In the first generation of MKTME, HKID is "stolen" from the physical address by allocating a configurable number of bits from the top of the physical address. TDX ISA is designed to further partition the HKID space into **shared HKIDs** for legacy MKTME accesses and **private HKIDs** for SEAM-mode-only accesses. Future generations might choose to express HKID differently.

During SEAM non-root operation, memory access can be qualified as either shared or private, based on the value of a new SHARED bit in the Guest Physical Address (GPA). Shared accesses are intended to behave as legacy memory accesses and use the upper bits of the host physical address as an HKID, which must be from the range allocated to legacy MKTME. Private accesses use the guest TD's private HKID.

The host-side Intel TDX functions help provide the means for the host VMM to manage HKID assignment to guest TDs, configure the memory encryption engines, etc., while better assuring proper operation to help maintain the TDX's security objectives. By design, the host VMM does not have access to the encryption keys.

Key management is described in Chapter 4.

2.4.2. Address Translation

Guest Physical Address (GPA) space is divided into private and shared sub-spaces, determined by the SHARED bit of GPA.

As designed, the CPU translates shared GPAs using the Shared EPT, which resides in host VMM memory. The Shared EPT is directly managed by the host VMM – the same as with legacy VMX.

As designed, the CPU translates private GPAs using a separate Secure EPT. The Secure EPT pages are encrypted and integrity-protected with the TD's ephemeral private key. The Secure EPT is not intended to be directly accessible by any software other than the Intel TDX module, nor by any device. Secure EPT can be managed indirectly by the host VMM, using Intel TDX functions. The Intel TDX module helps ensure that the Secure EPT security properties are kept. At the end of translation, the CPU sets the HKID bits in the HPA to the TD's assigned HKID.

TD private memory management is described in Chapter 9.

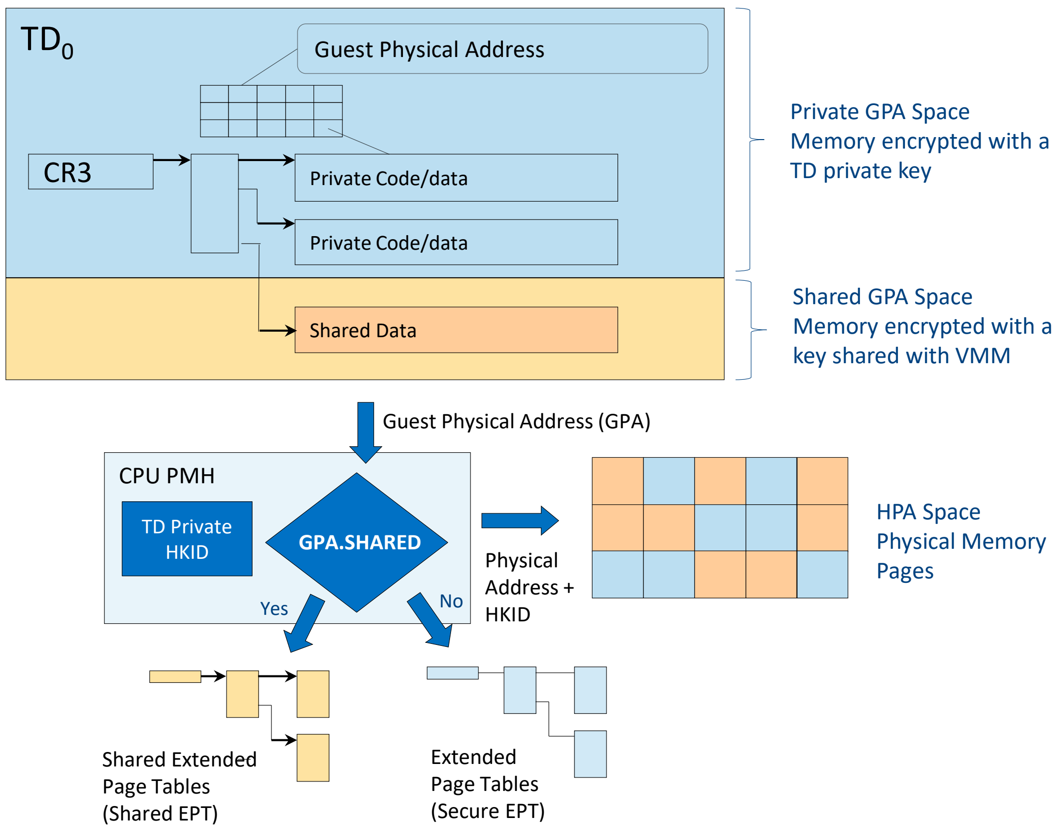


Figure 2.3: Secure EPT Concept

2.5. Guest TD State Protection

Intel TDX helps protect the confidentiality and integrity of a guest TD and the state of its Virtual CPUs (VCPUs) with the following mechanisms:

Protected Control Structures	TD-scope and TD VCPU-scope control structures, which hold guest TD metadata and TD VCPU state, are not directly accessible to any software (besides the Intel TDX module) or devices. As designed, the control structures are encrypted and integrity-protected with a private key and managed by Intel TDX functions. TD control structures are described in Chapter 6.
VCPU State on TD Transitions	On asynchronous TD exits, which usually happen due to external events, the CPU state is saved to the VCPU control structures, and a synthetic state is loaded into the CPU registers. On the following TD Entry, the CPU state is restored from the protected control structures. On synchronous TD-initiated exit, using the TDCALL(TDG.VP.VMCALL) function, selected GPR and XMM state can be passed as-is to the host VMM. On the following TD entry, that state can be passed back as-is to the guest TD.

2.6. Intel TDX I/O Model (w/o TDX Connect)

Note: This section describes I/O support without TDX Connect. For TDX Connect details, see the [TDX Connect Spec].

The TD guest can use the following I/O models:

- Paravirtualized devices
- Paravirtualized devices with MMIO emulation
- Direct assignment of devices to a TD

The Intel TDX architecture does not provide specific mechanisms for trusted I/O. Any integrity or confidentiality protection of data submitted to or received from physical or emulated devices must be done by the guest software using cryptography.

Intel TDX I/O is detailed in Chapter 13.

2.7. Measurement and Attestation

As designed, during TD launch, the initial contents and configuration of the TD are recorded by the Intel TDX module. In addition, run-time measurement registers can be used by the guest TD software, e.g., to measure a boot process. At run-time, the Intel TDX architecture reuses the Intel® Software Guard Extensions (Intel® SGX) attestation infrastructure to provide support for attesting to these measurements as described below.

Intel TDX attestation is intended to be used in two phases:

1. Software within the guest TD can use the TDCALL(TDG.MR.REPORT) function to request the Intel TDX module to generate an integrity protected TDREPORT structure. The Intel TDX ISA provides support for enabling the Intel TDX module to create this structure that includes the TD's measurements, the Intel TDX module's measurements, and a value provided by the guest TD software. This will typically be an asymmetric key that the attestation verifier can use to establish a secure channel or protect sensitive data to be sent to the TD software.
2. An Intel SGX Quoting Enclave, written specifically to support quoting Intel TDX TDs, uses a new ENCLU instruction leaf, EVERIFYREPORT2, to help check the integrity of the TDG.MR.REPORT. If it passes, the Quoting Enclave can use a certified quote signing key to sign a quote containing the guest TD's measurements and the additional data being quoted.

The Quoting Enclave can run anywhere on the platform where Intel SGX is supported.

Note: Running Intel SGX enclaves within a guest TD is not supported.

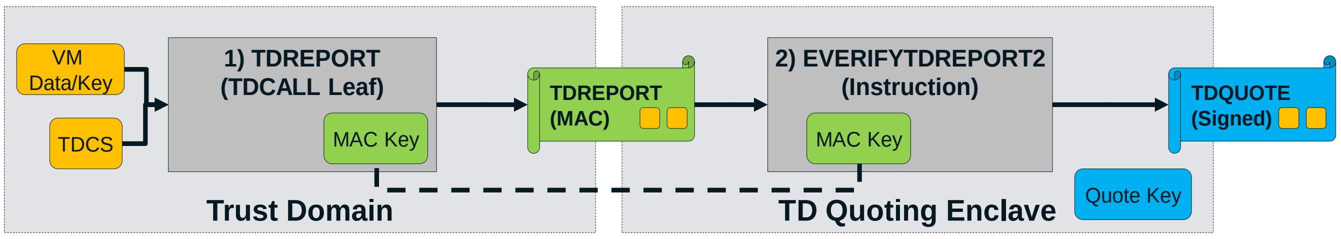


Figure 2.4: TD Attestation

TD measurement and attestation are described in Chapter 12.

2.8. Intel TDX Managed Control Structures

As designed, the Intel TDX module holds and manages a set of control structures that are not directly accessible to software (except the Intel TDX module itself). The controls structures are encrypted with private keys and HKIDs, and their content is only accessible in SEAM mode. Most control structures are addressable by the host VMM, which is responsible for allocating the memory to hold them.

The Intel TDX module uses control structures to help manage TD-private memory, transitions into and out of logical TDX non-root operation (TD entries and TD exits), as well as processor behavior in SEAM non-root operation.

Table 2.2: TDX-Managed Control Structures Overview

Scope	Name	Meaning	Description
Platform	KOT	Key Ownership Table	Designed to control private HKID assignment. KOT is internal to the Intel TDX module, intended not to be directly accessible to any other software.
	PAMT	Physical Address Metadata Table	PAMT is designed to hold metadata of each page in a Trust Domain Memory Range (TDMR). It controls the assignment of physical pages to guest TDs, etc. The PAMT is intended not to be directly accessible to software. It resides in memory allocated by the host VMM on TDX initialization.

Scope	Name	Meaning	Description
Guest TD	TDR	Trust Domain Root	TDR is intended to be the root control structure of a guest TD. It controls the key management and build/teardown process. The TDR is not intended to be directly accessible to software. It resides in memory allocated by the host VMM, via Intel TDX interface functions.
	TDCS	Trust Domain Control Structure	TDCS is intended to control the operation of a guest TD as a whole. The TDCS is not intended to be directly accessible to software. It resides in memory allocated by the host VMM, via Intel TDX interface functions.
	SEPT	Secure EPT	Secure EPT is an Extended Page Table (EPT) tree, managed by the TDX module, and used to help securely manage address translation for the TD private pages. The SEPT is not intended to be directly accessible to software. SEPT pages reside in memory allocated by the host VMM via Intel TDX interface functions.
Guest TD VCPU	TDVPS	Trust Domain Virtual Processor State	The TDVPS helps control the operation and hold the state of a guest TD virtual processor. It holds the TD VMCS and its auxiliary structures as well as other non-VMX control and state fields. The TDVPS is not intended to be directly accessible to software. It resides in memory allocated by the host VMM, via Intel TDX interface functions.

Intel TDX control structures are described in Chapter 6.

2.9. Intel TDX Interface Functions

The Intel TDX module implements functions that are triggered by executing two TDX instructions:

SEAMCALL The instruction used by the host VMM to invoke **host-side TDX interface functions**. The desired interface function is selected by an input operand (**leaf number**, in RAX). Host-side interface function names start with TDH (Trust Domain Host).

TDCALL The instruction used by the guest TD software (in SEAM non-root mode) to invoke **guest-side TDX functions**. The desired interface function is selected by an input operand (**leaf number**, in RAX). Guest-side interface function names start with TDG (Trust Domain Guest).

Intel TDX interface function details are described in the [TDX Module ABI Spec].

2.9.1. Host-Side (SEAMCALL Leaf) Interface Functions

Table 2.3: Host-Side (SEAMCALL Leaf) Interface Functions

Class	Interface Function Name	Leaf #	Description
Intel TDX Module Management	TDH.SYS.CONFIG	45	Globally configure the Intel TDX module
Intel TDX Module Management	TDH.SYS.INFO	32	Get Intel TDX module information
Intel TDX Module Management	TDH.SYS.INIT	33	Globally initialize the Intel TDX module
Intel TDX Module Management	TDH.SYS.KEY.CONFIG	31	Configure the Intel TDX global private key on the current package
Intel TDX Module Management	TDH.SYS.LP.INIT	35	Initialize the Intel TDX module per logical processor
Intel TDX Module Management	TDH.SYS.LP.SHUTDOWN	44	Does nothing; provided for backward compatibility
Intel TDX Module Management	TDH.SYS.RD	34	Read a TDX Module global-scope metadata field
Intel TDX Module Management	TDH.SYS.RDALL	37	Read all host-readable TDX Module global-scope metadata fields

Class	Interface Function Name	Leaf #	Description
Intel TDX Module Management	TDH.SYS.SHUTDOWN	52	Shutdown the Intel TDX module and prepare handoff data
Intel TDX Module Management	TDH.SYS.TDMR.INIT	36	Partially initialize a Trust Domain Memory Region (TDMR)
Intel TDX Module Management	TDH.SYS.UPDATE	53	Populate Intel TDX module state from handoff data
TD Management	TDH.MNG.ADDCX	1	Add a control structure page to a TD
TD Management	TDH.MNG.CREATE	9	Create a guest TD and its TDR root page
TD Management	TDH.MNG.INIT	21	Initialize per-TD control structures
TD Management	TDH.MNG.KEY.CONFIG	8	Configure the TD private key on a single package
TD Management	TDH.MNG.KEY.FREEID	20	Mark the guest TD's HKID as free
TD Management	TDH.MNG.KEY.RECLAIMID	27	Does nothing; provided for backward compatibility
TD Management	TDH.MNG.RD	11	Read TD metadata
TD Management	TDH.MNG.VPFLUSHDONE	19	Check all of a guest TD's VCPUs have been flushed by TDH.VP.FLUSH
TD Management	TDH.MNG.WR	13	Write TD metadata
VCPU Scope	TDH.VP.ADDCX	4	Add a control structure page to a TD VCPU
VCPU Scope	TDH.VP.CREATE	10	Create a guest TD VCPU and its TDVPR root page
VCPU Scope	TDH.VP.ENTER	0	Enter TDX non-root operation
VCPU Scope	TDH.VP.FLUSH	18	Flush the address translation caches and cached TD VMCS associated with a TD VCPU
VCPU Scope	TDH.VP.INIT	22	Initialize the per-VCPU control structures
VCPU Scope	TDH.VP.RD	26	Read VCPU metadata
VCPU Scope	TDH.VP.WR	43	Write VCPU metadata
Physical Memory Management	TDH.PHYMEM.CACHE.WB	40	Write back the contents of the cache on a package
Physical Memory Management	TDH.PHYMEM.PAGE.RDMD	24	Read the metadata of a page in a TDMR
Physical Memory Management	TDH.PHYMEM.PAGE.RECLAIM	28	Reclaim a physical memory page owned by a TD (i.e., TD private page, Secure EPT page or a control structure page)
Physical Memory Management	TDH.PHYMEM.PAGE.WBINVD	41	Write back and invalidate all cache lines associated with the specified memory page and HKID
Private Memory Management	TDH.MEM.PAGE.ADD	2	Add a 4KB private page to a TD during TD build time
Private Memory Management	TDH.MEM.PAGE.AUG	6	Dynamically add a 4KB private page to an initialized TD
Private Memory Management	TDH.MEM.PAGE.DEMOTE	15	Split a 2MB or a 1GB private TD page mapping into 512 4KB or 2MB page mappings respectively
Private Memory Management	TDH.MEM.PAGE.PROMOTE	23	Merge 512 consecutive 4KB or 2MB private TD page mappings into one 2MB or 1GB page mapping respectively
Private Memory Management	TDH.MEM.PAGE.RELOCATE	5	Relocate a 4KB mapped page from its HPA to another
Private Memory Management	TDH.MEM.PAGE.REMOVE	29	Remove a private page from a guest TD
Private Memory Management	TDH.MEM.RANGE.BLOCK	7	Block a TD private GPA range
Private Memory Management	TDH.MEM.RANGE.UNBLOCK	39	Remove the blocking of a TD private GPA range
Private Memory Management	TDH.MEM.RD	12	Read from private memory of a debuggable guest TD
Private Memory Management	TDH.MEM.SEPT.ADD	3	Add and map a 4KB Secure EPT page to a TD
Private Memory Management	TDH.MEM.SEPT.RD	25	Read a Secure EPT entry
Private Memory Management	TDH.MEM.SEPT.REMOVE	30	Remove a Secure EPT page from a TD
Private Memory Management	TDH.MEM.TRACK	38	Increment the TD's TLB tracking counter
Private Memory Management	TDH.MEM.WR	14	Write to private memory of a debuggable guest TD
Measurement and Attestation	TDH.MR.EXTEND	16	Extend the guest TD measurement register during TD build
Measurement and Attestation	TDH.MR.FINALIZE	17	Finalize the guest TD measurement register
Service TD	TDH.SERVTD.BIND	48	Bind a service TD to a target TD
Service TD	TDH.SERVTD.PREBIND	49	Pre-bind a service TD to a target TD
Migration	TDH.MIG.STREAM.CREATE	96	Create a migration stream
Migration Export	TDH.EXPORT.ABORT	64	Abort an export session

Class	Interface Function Name	Leaf #	Description
Migration Export	TDH.EXPORT.BLOCKW	65	Block a TD private page for writing
Migration Export	TDH.EXPORT.MEM	68	Export a list of TD private pages contents and/or cancellation requests
Migration Export	TDH.EXPORT.PAUSE	70	Pause the exported TD
Migration Export	TDH.EXPORT.RESTORE	66	Restore a list of TD private 4KB pages' Secure EPT entry states after an export abort
Migration Export	TDH.EXPORT.STATE.IMMUTABLE	72	Start an export session and export the TD's immutable state
Migration Export	TDH.EXPORT.STATE.TD	73	Export the TD's mutable state
Migration Export	TDH.EXPORT.STATE.VP	74	Export a VCPU mutable state
Migration Export	TDH.EXPORT.TRACK	71	End the current in-order export phase epoch and either start a new epoch or start the out-of-order export phase
Migration Export	TDH.EXPORT.UNBLOCKW	75	Unblock a page that has been blocked for writing
Migration Import	TDH.IMPORT.ABORT	80	Abort an import session
Migration Import	TDH.IMPORT.COMMIT	82	Commit the import session and allow the imported TD to run
Migration Import	TDH.IMPORT.END	81	End an import session
Migration Import	TDH.IMPORT.MEM	83	Import a list of TD private pages contents and/or cancellation requests based on a migration bundle in shared memory
Migration Import	TDH.IMPORT.STATE.IMMUTABLE	85	Start an import session and import the TD's immutable state
Migration Import	TDH.IMPORT.STATE.TD	86	Import the TD's mutable state
Migration Import	TDH.IMPORT.STATE.VP	87	Import a VCPU mutable state
Migration Import	TDH.IMPORT.TRACK	84	End the current in-order import phase epoch and either start a new epoch or start the out-of-order import phase

2.9.2. Guest-Side (TDCALL Leaf) Interface Functions

Table 2.4: Guest-Side (TDCALL Leaf) Interface Functions

Class	Interface Function Name	Leaf #	Description
Intel TDX Module Management	TDG.SYS.RD	11	Read a TDX Module global-scope metadata field
Intel TDX Module Management	TDG.SYS.RDALL	12	Read all gust-readable TDX Module global-scope metadata fields
TD Management	TDG.VM.RD	7	Read a TD-scope metadata field
TD Management	TDG.VM.WR	8	Write a TD-scope metadata field
VCPU Scope	TDG.VP.CPUIDVE.SET	5	Control delivery of #VE on CPUID instruction execution
VCPU Scope	TDG.VP.ENTER	25	Enter L2 VCPU operation
VCPU Scope	TDG.VP.INFO	1	Get TD execution environment information
VCPU Scope	TDG.VP.INVEPT	26	Invalidate cached EPT translations for selected L2 VMs
VCPU Scope	TDG.VP.INVGLA	27	Invalidate cached translations for selected pages in an L2 VM
VCPU Scope	TDG.VP.RD	9	Read a VCPU-scope metadata field
VCPU Scope	TDG.VP.VEINFO.GET	3	Get Virtualization Exception Information for the recent #VE exception
VCPU Scope	TDG.VP.VMCALL	0	Call a host VM service
VCPU Scope	TDG.VP.WR	10	Write a VCPU-scope metadata field
Private Memory Management	TDG.MEM.PAGE.ACCEPT	6	Accept a pending private page into the TD
Private Memory Management	TDG.MEM.PAGE.ATTR.RD	23	Read the GPA mapping and attributes of a TD private page
Private Memory Management	TDG.MEM.PAGE.ATTR.WR	24	Write the attributes of a private page

Class	Interface Function Name	Leaf #	Description
Measurement and Attestation	TDG.MR.REPORT	4	Creates a cryptographic report of the TD
Measurement and Attestation	TDG.MR.RTMR.EXTEND	2	Extend a TD run-time measurement register
Measurement and Attestation	TDG.MR.VERIFYREPORT	22	Verify a cryptographic report of a TD, generated on the current platform
Service TD	TDG.SERVTD.RD	18	Read a target TD metadata field
Service TD	TDG.SERVTD.WR	20	Write a target TD metadata field

3. Software Use Cases

This chapter summarizes the software use cases (also known as software flows) used with the Intel TDX module.

3.1. Intel TDX Module Lifecycle

3.1.1. Intel TDX Module Platform-Scope First-Time Initialization

This sequence is intended to be used by the host VMM to initialize the Intel TDX module at the platform scope.

Table 3.1: Typical Intel TDX Module Platform-Scope First-Time Initialization Sequence

Phase		Intel TDX Function	Scope	Execute On	Description
Boot	1	N/A	Platform	Each core	BIOS configures Convertible Memory Regions (CMRs) and activates MKTME; MCHECK checks them and securely stores the information.
P-SEAMLDR Loading	2	N/A	Platform	One of the BSPs	BIOS or OS launches the NP-SEAMLDR ACM, which loads the Intel P-SEAMLDR module.
Intel TDX Module Loading	3	SEAMLDR.INSTALL	Platform	Each LP, serially	VMM calls the Intel P-SEAMLDR module with "load" scenario to install the first TDX module. The TDX module is installed when SEAMLDR.INSTALL is called on the last LP.
Intel TDX Module Initialization	4	TDH.SYS.INIT	Platform	Any one LP	Perform global initialization of the Intel TDX module.
	5	TDH.SYS.LP.INIT	LP	Each LP	Perform LP-scope, core-scope and package-scope initialization, checking and configuration of the platform and the Intel TDX module.
Enumeration and Configuration	6	TDH.SYS.RD* or TDH.SYS.INFO	Platform	Any initialized LP	Retrieve Intel TDX module information and convertible memory (CMR) information.
	7	TDH.SYS.CONFIG	Platform	Any one LP	Configure the Intel TDX module with TDMR and PAMT setup.
	8	N/A	Package	Each Package	If any MODIFIED cache lines may exist for the PAMT ranges, flush them to memory using, e.g., WBINVD.
	9	TDH.SYS.KEY.CONFIG	Package	Each Package	Configure the Intel TDX global private key used for encrypting PAMT and TDR on the hardware (other TD-scope control structures are encrypted with their respective TD's ephemeral private keys).
	At this point any Intel TDX function may be executed on any LP.				
Memory Initialization	10	TDH.SYS.TDMR.INIT (multiple)	Platform	One or more LPs	Called multiple times to gradually initialize the PAMT structure for each TDMR.
	Once each 1GB block of TDMR has been initialized by TDH.SYS.TDMR.INIT, it can be used to hold TD-private pages.				

3.1.2. Intel TDX Module Shutdown and Update

This sequence is intended to be used by the host VMM to gracefully shut down the Intel TDX module and install a new Intel TDX module. There are 2 scenarios:

- Reload scenario – guest TDs' context and memory are lost.
- Update scenario - guest TDs' context and memory are preserved.

Intel TDX Module Reload

In the reload scenario, the previous TDX module in SEAM range is erased when the next TDX module is installed. Since in this scenario the previous module can't pass any information to the next TDX module, the next TDX module starts afresh, and all guest TDs' context and memory out of SEAM range becomes effectively inaccessible.

Table 3.2: TDX Module Reload Sequence

Phase		Intel TDX Function	Scope	Execute On	Description
P-SEAMLDR	1	SEAMLDR.INSTALL with "load" scenario	Platform	All LPs	Installs the next TDX module, regardless of the previous TDX module.
Next TDX module	The initialization sequence continues in the same way as described in 3.1.1 above, steps 3 to 8.				

3.1.2.2 *Intel TDX Module Update*

In the update scenario, the previous TDX module in SEAM range is not fully erased; the previous TDX module can be asked to leave "handoff data" in a specific location of the SEAM range, so that the next TDX module would be able to initialize itself from this handoff data. The next TDX module can thus keep supporting guest TDs' context and memory.

Table 3.3: TDX Module TD-Preserving Update Sequence

Phase		Intel TDX Function	Scope	Execute On	Description
Previous TDX Module	1	TDH.SYS.SHUTDOWN	Platform	Selected LP	Prepare handoff data and mark the TDX module's global state as "shutdown".
P-SEAMLDR	2	SEAMLDR.INSTALL with "update" scenario	Platform	All LPs, serially	Install the next TDX module without clearing the handoff data.
Next TDX Module	3	TDH.SYS.INIT	Platform	Selected LP	Perform global initialization of the Intel TDX module.
	4	TDH.SYS.LP.INIT	LP	All LPs	Perform LP-scope, core-scope and package-scope initializations, checking and configuration of the platform and the Intel TDX module.
	5	TDH.SYS.UPDATE	Platform	Selected LP	Populate internal variables from handoff data. Mark the TDX module's global state as "ready".
	At this point any Intel TDX function may be executed on any LP (don't call TDH.SYS.CONFIG, TDH.SYS.KEY.CONFIG and TDH.SYS.TDMR.INIT). In particular, TDs created by previous TDX modules can be re-entered.				

3.2. *TD Build*

The following sequence is intended to be used by the host VMM to build a TD.

Table 3.4: Typical TD Build Sequence

	Step		Description	SEAMCALL Leaf Functions
A	**TD Creation and Key Resource Assignment**	1	The host VMM finds/allocates a free HKID for the new TD.	TDH.MNG.CREATE TDH.MNG.KEY.CONFIG
		2	The host VMM allocates a 4KB page in TDMR for the TDR.	
		3	The host VMM creates the new TD by calling the TDH.MNG.CREATE function (passing HPA of the TDR page). This initializes the target TDR page.	

	Step		Description	SEAMCALL Leaf Functions
		4	The TD host VMM configures the MKTME hardware with the TD's private key by calling the TDH.MNG.KEY.CONFIG function on each package.	
		5	At this point, the TD private memory is accessible. The VMM can use Intel TDX interface functions to create control structures and TD private pages as described below.	
B	**TDCS Memory Allocation and TD Initialization**	1	The host VMM allocates multiple 4KB TDCX pages for TDCS. The number of TDCX pages required is enumerated by TDH.SYS.RD* or TDH.SYS.INFO.	TDH.MNG.ADDCX TDH.MNG.INIT
		2	For each TDCX page, the host VMM calls the TDH.MNG.ADDCX function (passing HPA of TDCX) to add the page to the TD.	
		3	The host VMM builds a TD_PARAMS structure. For example, the TD configuration parameters can be obtained from a TD manifest supplied by the TD owner.	
		4	The host VMM calls the TDH.MNG.INIT function (passing the TD_PARAMS structure) to initialize the TD.	
C	**Virtual Processor Creation and Configuration (Executed per each VCPU)**	1	The host VMM allocates target pages for the VCPU's TDVPR and TDCX pages in TDMR in the context of a TD. The number of TDCX pages required is enumerated by TDH.SYS.RD* or TDH.SYS.INFO.	TDH.VP.CREATE TDH.VP.ADDCX TDH.VP.INIT TDH.VP.WR
		2	The host VMM creates a new TD virtual CPU by calling the TDH.VP.CREATE function (passing the HPA of the new TDVPR page and its owner TDR page).	
		3	For each TDCX page, the host VMM calls the TDH.VP.ADDCX function (passing the HPA of the new TDCX page and its parent TDVPR page).	
		4	The host VMM initializes the TD VCPU by calling the TDH.VP.INIT function (passing the HPA of its TDVPR page). It also passes a single 64b parameter that is later passed to the VBIOS in the initial value of RCX. This parameter can be used as a pointer to a configuration structure in shared memory.	
		5	The host VMM allocates Shared EPT for each VP.	
		6	The host VMM uses the TDH.VP.WR function to write to the TD VMCS Shared EPTP field.	
		7	The host VMM may modify a few TD VMCS execution control fields using TDH.VP.WR.	
D	**TD Boot Memory Setup**	1	The host VMM loads the TD boot image to its memory. The boot image contains code and data pages that typically include a virtual BIOS, OS boot loader, configuration, etc.	TDH.MEM.SEPT.ADD TDH.MEM.PAGE.ADD TDH.MR.EXTEND
		2	The host VMM builds the TD Secure EPT by allocating physical pages and calling the TDH.MEM.SEPT.ADD function multiple times.	
		3	The host VMM allocates the initial set of physical pages for the TD boot image and maps them into host address space.	

	Step		Description	SEAMCALL Leaf Functions
		4	For each TD page: 1. The host VMM specifies a TDR as a parameter and calls the TDH.MEM.PAGE.ADD function. It copies the contents from the TD image page into the target TD page which is encrypted with the TD ephemeral key. TDH.MEM.PAGE.ADD also extends the TD measurement with the page GPA. 2. The host VMM extends the TD measurement with the contents of the new page by calling the TDH.MR.EXTEND function on each 256-byte chunk of the new TD page.	
E	**TD Measurement Finalization**	1	The host VMM calls the TDH.MR.FINALIZE function, which finalizes the TD measurement.	TDH.MR.FINALIZE
		2	At this point, the TD is finalized. • Its measurement cannot be modified anymore (except the run-time measurement registers). • TD VCPUs can be entered using SEAMCALL(TDH.VP.ENTER).	

3.3. TD Run Time

3.3.1. Private Memory Management

3.3.1.1. Dynamic Page Addition (Shared to Private Conversion)

5 The following sequence is intended to be used by the host VMM to dynamically add a page to a guest TD.

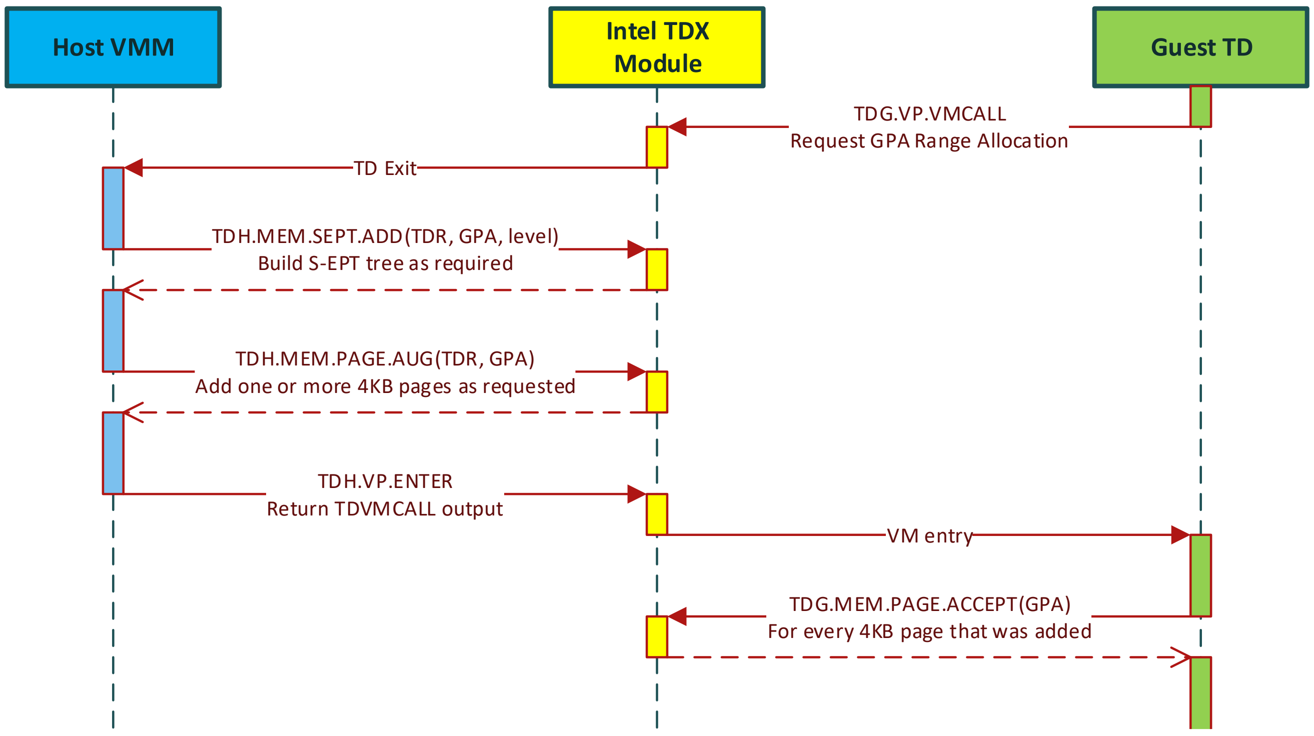

Host VMM
Intel TDX Module
Guest TD
TDG.VP.VMCALL
Request GPA Range Allocation
TD Exit
TDH.MEM.SEPT.ADD(TDR, GPA, level)
Build S-EPT tree as required
TDH.MEM.PAGE.AUG(TDR, GPA)
Add one or more 4KB pages as requested
TDH.VP.ENTER
Return TDVMCALL output
VM entry
TDG.MEM.PAGE.ACCEPT(GPA)
For every 4KB page that was added

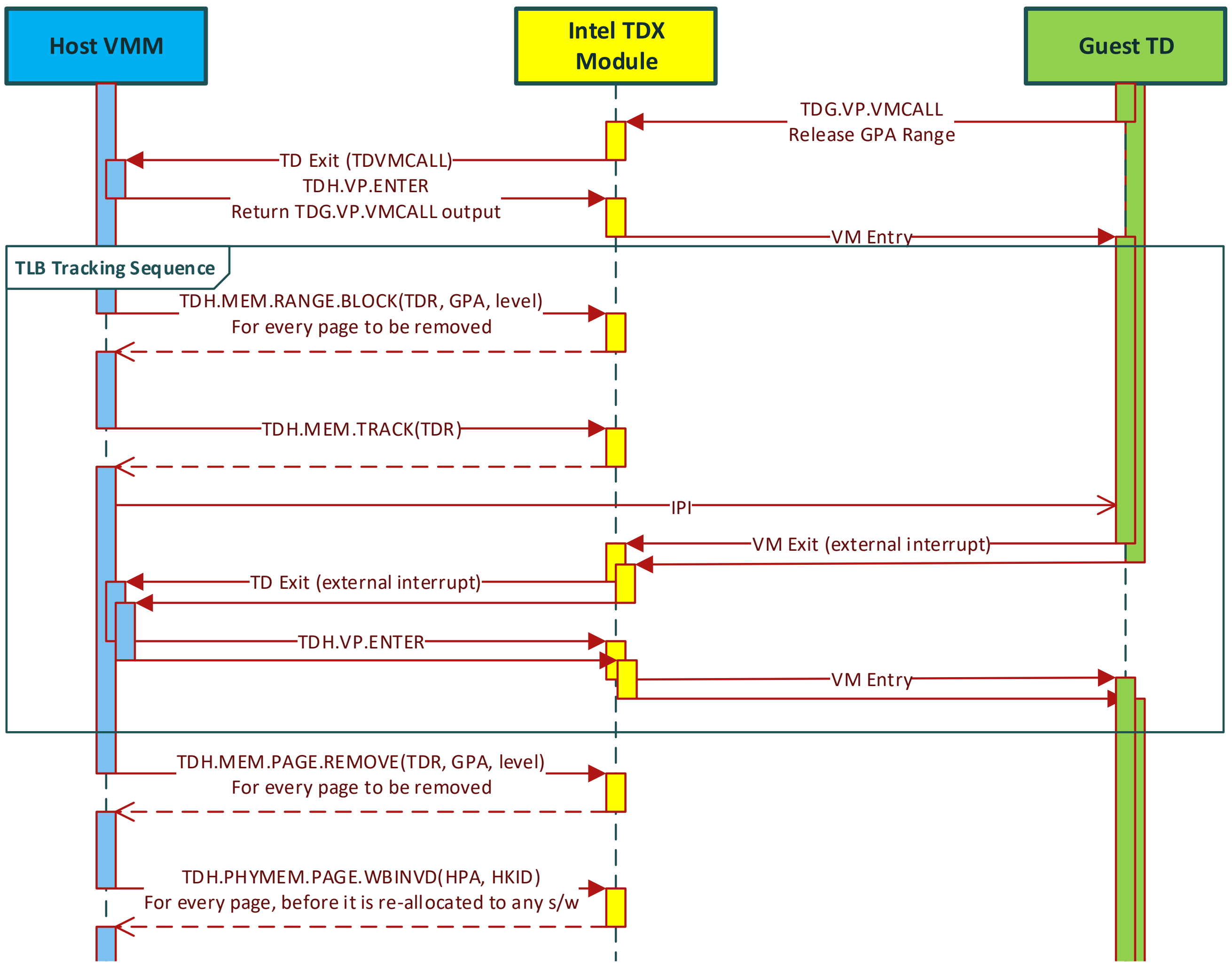
Figure 3.1: Typical Dynamic Page Addition Sequence

Table 3.5: Typical Dynamic Page Addition (Shared to Private Conversion) Sequence

Phase		Side	Intel TDX Function	Scope	Execute On	Description
Allocation Request	1	TD	TDG.VP.VMCALL	TD	Any LP	Optional software protocol: Request GPA range allocation.
Page Addition	2	VMM	TDH.MEM.SEPT.ADD	TD	Any LP	If required, update the Secure EPT.
	3	VMM	TDH.MEM.PAGE.AUG (multiple)	TD	Any LP	Add one or more new 4KB or 2MB private pages.
	At this point, the new page is pending acceptance by the guest TD and cannot be accessed by it yet.					
	4	VMM	TDH.VP.ENTER	TD	Any LP	Optional software protocol: Return TDG.VP.VMCALL result.
Page Acceptance	5	TD	TDG.MEM.PAGE.ACCEPT (multiple)	TD	Any LP	Accept the new pending page(s). The content of each page is zeroed out.
	At this point, the new page can be accessed by the guest TD.					

3.3.12. Dynamic Page Removal (Private to Shared Conversion)

The following sequence is intended to be used by the host VMM to dynamically remove a page from a guest TD.


Host VMM
Intel TDX Module
Guest TD
TDG.VP.VMCALL
Release GPA Range
TD Exit (TDVMCALL)
TDH.VP.ENTER
Return TDG.VP.VMCALL output
VM Entry
TLB Tracking Sequence
TDH.MEM.RANGE.BLOCK(TDR, GPA, level)
For every page to be removed
TDH.MEM.TRACK(TDR)
IPI
VM Exit (external interrupt)
TD Exit (external interrupt)
TDH.VP.ENTER
VM Entry
TDH.MEM.PAGE.REMOVE(TDR, GPA, level)
For every page to be removed
TDH.PHYMEM.PAGE.WBINVD(HPA, HKID)
For every page, before it is re-allocated to any s/w

5

Figure 3.2: Typical Dynamic Page Removal Sequence

Table 3.6: Typical Dynamic Page Removal (Private to Shared Conversion) Sequence

Phase		Side	Intel TDX Function	Scope	Execute On	Description
Ballooning Notification	1	TD	TDG.VP.VMCALL	TD	Any LP	Optional software protocol: Release GPA range.
	2	VMM	TDH.VP.ENTER	TD	Any LP	Optional software protocol: Return TDG.VP.VMCALL result.
TLB Tracking Sequence	3	VMM	TDH.MEM.RANGE.BLOCK (multiple)	TD	Any LP	Block private pages from further address translation.
	4	VMM	TDH.MEM.TRACK	TD	Any one LP	Increment the TD's TLB epoch.
	5	VMM	N/A	TD	Multiple LPs	Send an IPI, causing TD exit on any remote LP associated with a VCPU. Subsequent TDH.VP.ENTER will flush TLB.
Page Removal	6	VMM	TDH.MEM.PAGE.REMOVE (multiple)	TD	Any LP	Clear Secure EPT entry. Non-ACT platforms: Mark the physical page as free. ACT platforms: Flush cache lines of the removed page(s) and fill them with random encrypted data. Then mark the physical page(s) as free and set the ACT bit(s) to shared.
Cache Flushing & Content Init	Before re-allocating any of the removed pages to any use, the host VMM should ensure none of the cache lines of the removed pages are in the MODIFIED state to avoid corruption due to cache line aliasing. This is done using one of the following methods:					
	7a	VMM	ACT platforms: TDH.PHYMEM.PAGE.WBINVD (multiple)	TD	Any one LP	Flush the cache lines of the removed page(s).
	7b	VMM	WBNOINVD	Platform	One LP per WBINVD domain[2]	Globally write back all caches.
	7c	VMM	WBINVD	Platform	One LP per WBINVD domain[3]	Globally write back and invalidate all caches.
	8	VMM	MOVDIR64B	Page	Any LP	Initialize the physical page content for use with a new shared HKID.

[2] Some CPUs may require running WBNOINVD per a set of LPs that is smaller than the set of all LPs in a package.

[3] Some CPUs may require running WBINVD per a set of LPs that is smaller than the set of all LPs in a package.

Page Promotion (Mapping Merge)

Page size promotion is intended to be used by the host VMM to merge 512 pages mapped as 4KB or 2MB into a single page mapped as 2MB or 1GB, respectively. It is detailed in 9.13.2.

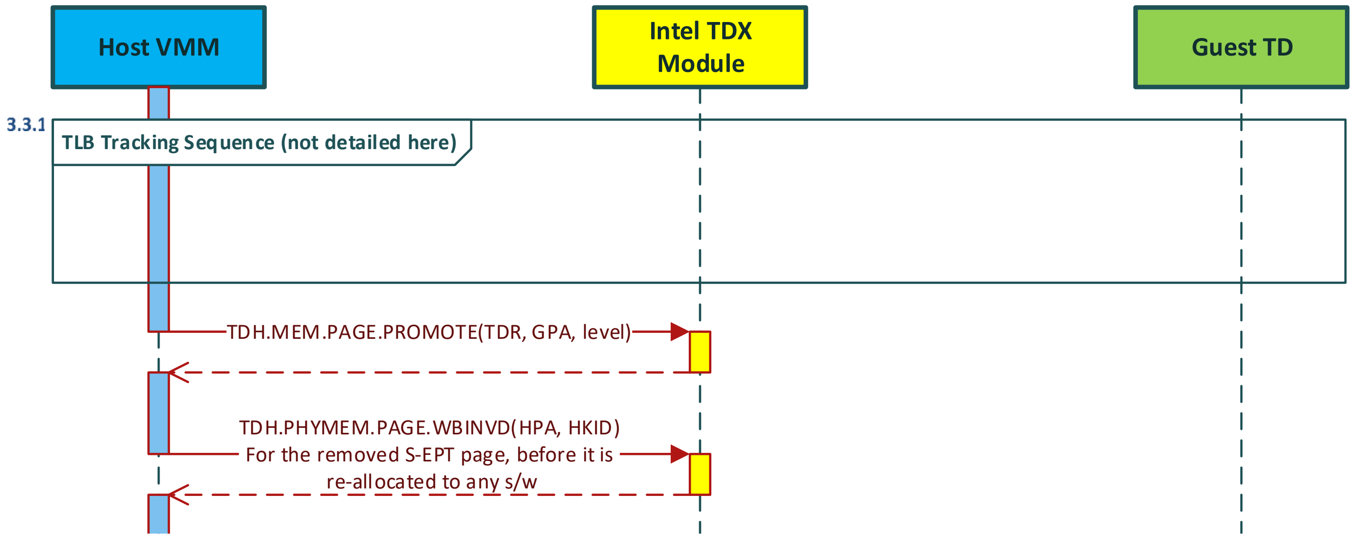

Host VMM
Intel TDX Module
Guest TD
3.3.1
TLB Tracking Sequence (not detailed here)
TDH.MEM.PAGE.PROMOTE(TDR, GPA, level)
TDH.PHYMEM.PAGE.WBINVD(HPA, HKID)
For the removed S-EPT page, before it is re-allocated to any s/w

Figure 3.3: Typical Page Promotion Sequence

Table 3.7: Typical Page Promotion (Mapping Merge) Sequence

Phase		Intel TDX Function	Scope	Execute On	Description
TLB Tracking Sequence	1	TDH.MEM.RANGE.BLOCK	TD	Any LP	Block the GPA range to be merged from further address translation.
	2	TDH.MEM.TRACK	TD	Any one LP	Increment the TD's TLB epoch.
	3	N/A	TD	Multiple LPs	Send an IPI, causing TD exit on any remote LP associated with a VCPU. Subsequent TDH.VP.ENTER will flush TLB.
Promotion	4	TDH.MEM.PAGE.PROMOTE	TD	Any LP	Merge small pages in the GPA range into a large page.
Cache Flushing & Content Init 3.3.1.4.	5	Non-ACT platforms: TDH.PHYMEM.PAGE.WBINVD	TD	Any LP	Flush the removed Secure EPT page's cache lines.
		ACT platforms:			This operation is performed by the TDX module
	6	MOVDIR64B	Page	Any LP	Initialize the physical page content for use with a new shared HKID.

Page Demotion (Mapping Split)

Page size demotion is intended to be used by the host VMM to split a page mapped as 1GB or 2MB into 512 pages mapped as 2MB or 4KB, respectively. It is detailed in 9.13.3.

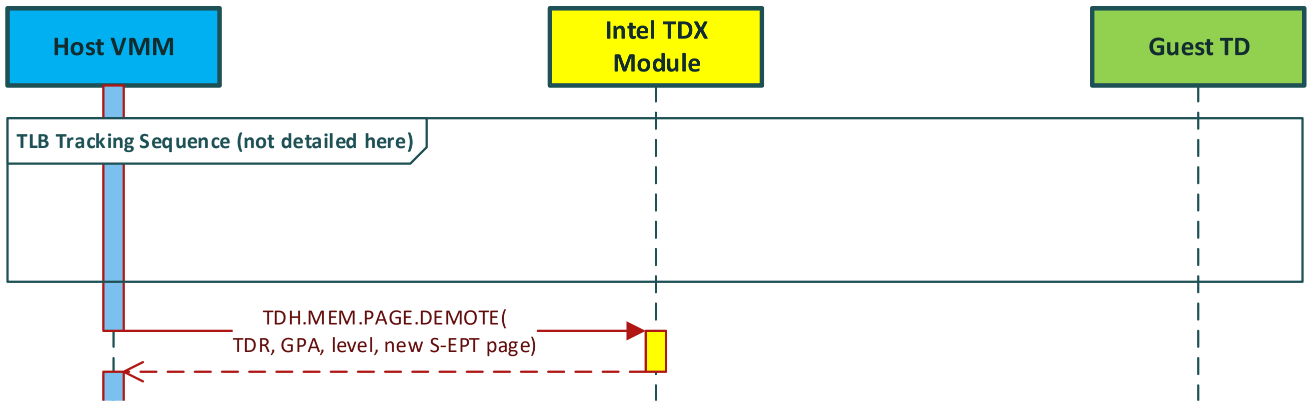

Host VMM
Intel TDX Module
Guest TD
TLB Tracking Sequence (not detailed here)
TDH.MEM.PAGE.DEMOTE(
TDR, GPA, level, new S-EPT page)

Figure 3.4: Typical Page Demotion Sequence

Table 3.8: Typical Page Demotion (Mapping Split) Sequence

Phase		Intel TDX Function	Scope	Execute On	Description
TLB Tracking Sequence	1	TDH.MEM.RANGE.BLOCK	TD	Any LP	Block private large page from further address translation.
	2	TDH.MEM.TRACK	TD	Any one LP	Increment the TD’s TLB epoch.
	3	N/A	TD	Multiple LPs	Send an IPI, causing TD exit on any remote LP associated with a VCPU. Subsequent TDH.VP.ENTER will flush TLB.
Demotion	4	TDH.MEM.PAGE.DEMOTE	TD	Any LP	Split the large page into multiple small pages.

3.3.1.5. *GPA Range Unblock*

GPA range unblock is intended to be used when a range has been blocked, for example, for page removal, but the host VMM decides to cancel the operation. Unblock is detailed in 0.

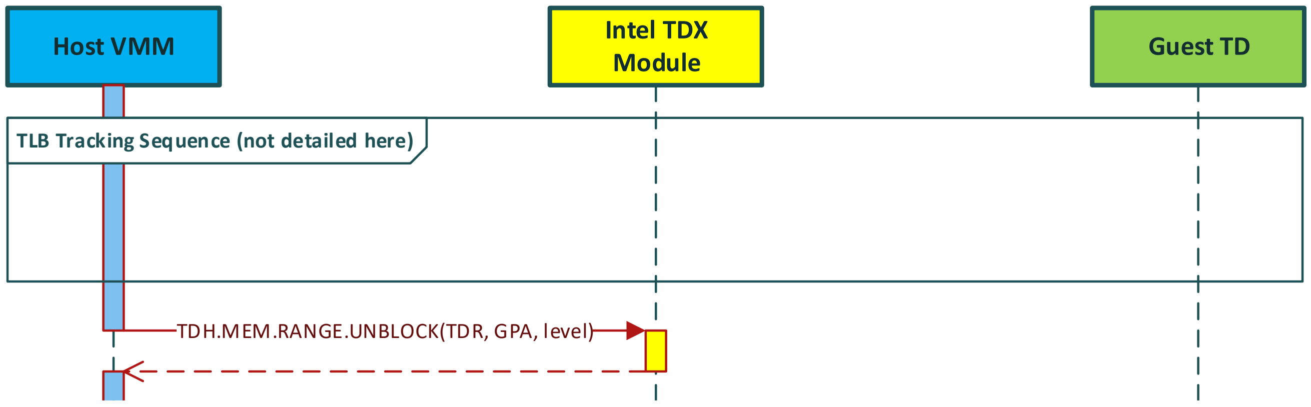

Host VMM
Intel TDX Module
Guest TD
TLB Tracking Sequence (not detailed here)
TDH.MEM.RANGE.UNBLOCK(TDR, GPA, level)

Figure 3.5: Typical GPA Range Unblock Sequence

Table 3.9: Typical GPA Range Unblock Sequence

Phase		Intel TDX Function	Scope	Execute On	Description
TLB Tracking Sequence	1	TDH.MEM.RANGE.BLOCK (multiple)	TD	Any LP	Block private GPA range from further address translation.
	2	TDH.MEM.TRACK	TD	Any one LP	Increment the TD's TLB epoch.
	3	N/A	TD	Multiple LPs	Send an IPI, causing TD exit on any remote LP associated with a VCPU. Subsequent TDH.VP.ENTER will flush TLB.
Unblocking	4	TDH.MEM.RANGE.UNBLOCK	TD	Any LP	Remove the private GPA range blocking.

3.3.2. Guest TD Execution

TD VCPU First-Time Invocation

Table 3.10: Typical TD VCPU First-Time Invocation Sequence

Phase		Side	Intel TDX Function	Scope	Execute On	Description
Entering TD VCPU (First Time)	1	VMM	N/A	LP	LP x	Save VMM LP state not preserved across TD Entry to TD exit.
	2	VMM	TDH.VP.ENTER	VCPU/LP	LP x	Restore the initial LP state, as set by TDH.VP.INIT, from TDVPS and enter SEAM non-root mode.
TD VCPU Initial Execution	TD software (VBIOS) starts execution in 32-bit protected mode with no paging.					
	3	TD	N/A	VCPU/LP	LP x	TD software parses initial information in GPR, builds page tables and switches to 64-bit mode.
	TD software (VBIOS) now executes in 64-bit mode.					
Enumeration	4	TD	TDG.VP.INFO	VCPU/LP	LP x	TD software retrieves basic TD and execution environment information.
	5	TD	TDG.MR.REPORT	VCPU/LP	LP x	TD software retrieves additional TD information.
3.3.2.2.	TD continues execution in SEAM non-root mode.					

TD VCPU Entry, Exit on TDG.VP.VMCALL and Re-Entry

Table 3.11: Typical TD Entry, Exit on TDG.VP.VMCALL and Re-Entry Sequence

Phase		Side	Intel TDX Function	Scope	Execute On	Description
TD Entry	1	VMM	N/A	LP	LP x	Save VMM LP state not preserved across TD Entry to TD exit.
	2	VMM	TDH.VP.ENTER	VCPU/LP	LP x	Restore LP state from TDVPS and enter SEAM non-root mode.
	TD executes in TDX non-root mode.					

Phase		Side	Intel TDX Function	Scope	Execute On	Description
Software Protocol over TDG.VP.VMCALL	3	TD	TDG.VP.VMCALL	VCPU/LP	LP x	Exit SEAM non-root mode, save LP state to TDVPS, and set synthetic state (except most GPRs and all XMMs).
	4	VMM	N/A	LP	LP x	Optionally: Restore VMM LP state saved before TDH.VP.ENTER.
	5	VMM	N/A	LP	LP x	Perform TDG.VP.VMCALL function, as determined by the TD-VMM software contract (out of scope for this document).
	6	VMM	N/A	LP	LP x	Save VMM LP state not preserved across TD Entry to TD exit.
	7	VMM	TDH.VP.ENTER	VCPU/LP	LP x	Restore LP state from TDVPS (except most GPRs and all XMMs). Enter SEAM non-root mode.
	8	TD	N/A	VCPU/LP	LP x	Parse TDG.VP.VMCALL output operands as determined by TD – VMM software contract.
TD Execution	TD continues execution in SEAM non-root mode.					

3.3.2.3. *TD VCPU Entry, Exit on Asynchronous Event and Re-Entry*

Table 3.12: Typical TD Entry, Exit on Asynchronous Event and Re-Entry Sequence

Phase		Side	Intel TDX Function	Scope	Execute On	Description
TD Entry	1	VMM	N/A	LP	LP x	Save LP state not preserved across TD Entry to TD exit.
	2	VMM	TDH.VP.ENTER	VCPU/LP	LP x	Restore LP state from TDVPS. Enter SEAM non-root mode.
	TD executes in TDX non-root mode.					
Async. TD Exit and Re-Entry	3	TD	N/A	VCPU/LP	LP x	Asynchronous events (interrupt, exception, EPT violation, etc.) cause TD exit. Save LP state to TDVPS and set synthetic state.
	4	VMM	N/A	LP	LP x	Restore any required LP state saved by the VMM before TDH.VP.ENTER.
	5	VMM	N/A	LP	LP x	Handle the asynchronous event.
	6	VMM	N/A	LP	LP x	Save VMM LP state not preserved across TD Entry to TD exit.
	7	VMM	TDH.VP.ENTER	VCPU/LP	LP x	Restore LP state from TDVPS and enter SEAM non-root mode.
TD Execution	TD continues execution in SEAM non-root mode.					

3.3.2.4. *Guest-Side Functions*

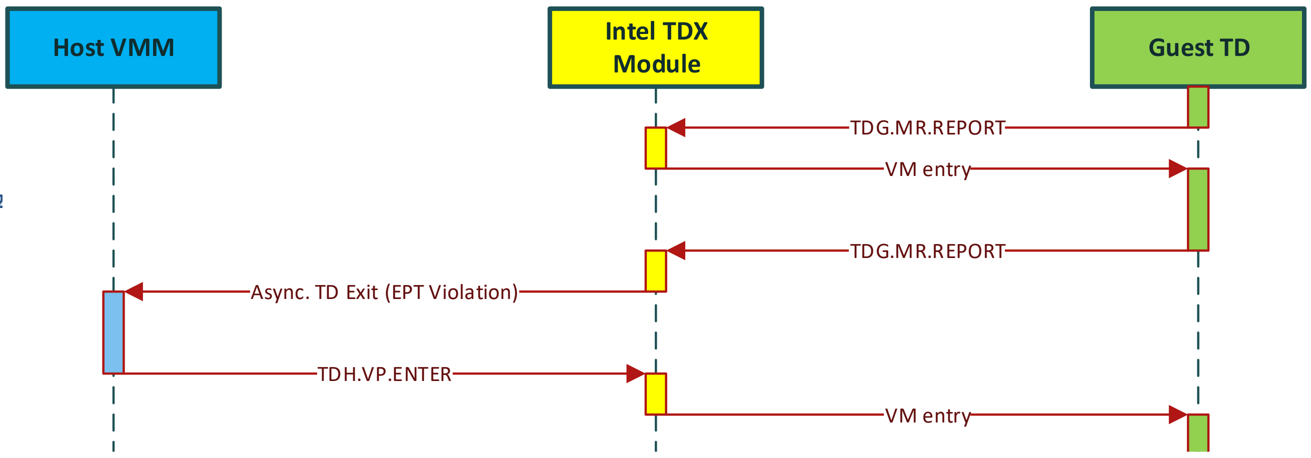


Figure 3.6: Typical Guest-Side Function Sequences

5 Table 3.13: Typical Guest-Side Functions Sequences

Case		Side	Intel TDX Function	Scope	Execute On	Description
Guest-Side Function Returns to Guest TD	TD executes in SEAM non-root mode					
	1	TD	TDG.MR.REPORT	VCPU/LP	LP x	The guest TD VM-exits to the Intel TDX module, which handles the guest-side function and re-enters the TD.
	TD continues execution in SEAM non-root mode					
Guest-Side Function Causes Async. TD Exit	2	TD	TDG.MR.REPORT	VCPU/LP	LP x	The guest TD exits to the Intel TDX module, which handles the guest-side function, but an asynchronous event (e.g., EPT violation, etc.) causes TD exit.
	3	VMM	N/A	LP	LP x	Optional: The host VMM restores the VMM LP state saved before TDH.VP.ENTER.
	4	VMM	N/A	LP	LP x	The host VMM handles the asynchronous event.
	5	VMM	N/A	LP	LP x	The host VMM saves any VMM LP state not preserved across TD Entry to TD exit.
	6	VMM	TDH.VP.ENTER	VCPU/LP	LP x	The Intel TDX module restores LP state from TDVPS and enters SEAM non-root mode.
	TD continues execution in SEAM non-root mode.					

3.3.2.5. *TD VCPU Rescheduling (Migration to Another LP)*

The Intel TDX module is designed to allow a TD VCPU to be associated with at most one LP at any time. The host VMM must explicitly break this association in order to migrate the VCPU to another LP.

10 Table 3.14: Typical VCPU Migration to Another LP Sequence

Phase		Intel TDX Function	Scope	Execute On	Description
Old VCPU→LP Association	1	Any VCPU-specific SEAMCALL leaf	VCPU	Old LP	Any VCPU-specific SEAMCALL leaf (e.g., TDH.VP.INIT, TDH.VP.ENTER, TDH.VP.RD, etc.) creates an association between the current LP and the VCPU.

Phase		Intel TDX Function	Scope	Execute On	Description
Breaking Old VCPU→LP Association	2	TDH.VP.FLUSH	VCPU	Old LP	Break the VCPU-LP association: flush the VCPU's TD VMCS to TDVPS memory and flush the VCPU's TLB ASID.
	At this point the VCPU is not associated with any LP.				
New VCPU→LP Association	3	Any VCPU-specific SEAMCALL leaf	VCPU	New LP	Create a new VCPU-LP association.

3.4. TD Destruction

The following sequence is intended to be used by the host VMM to destroy a TD and reclaim all its resources.

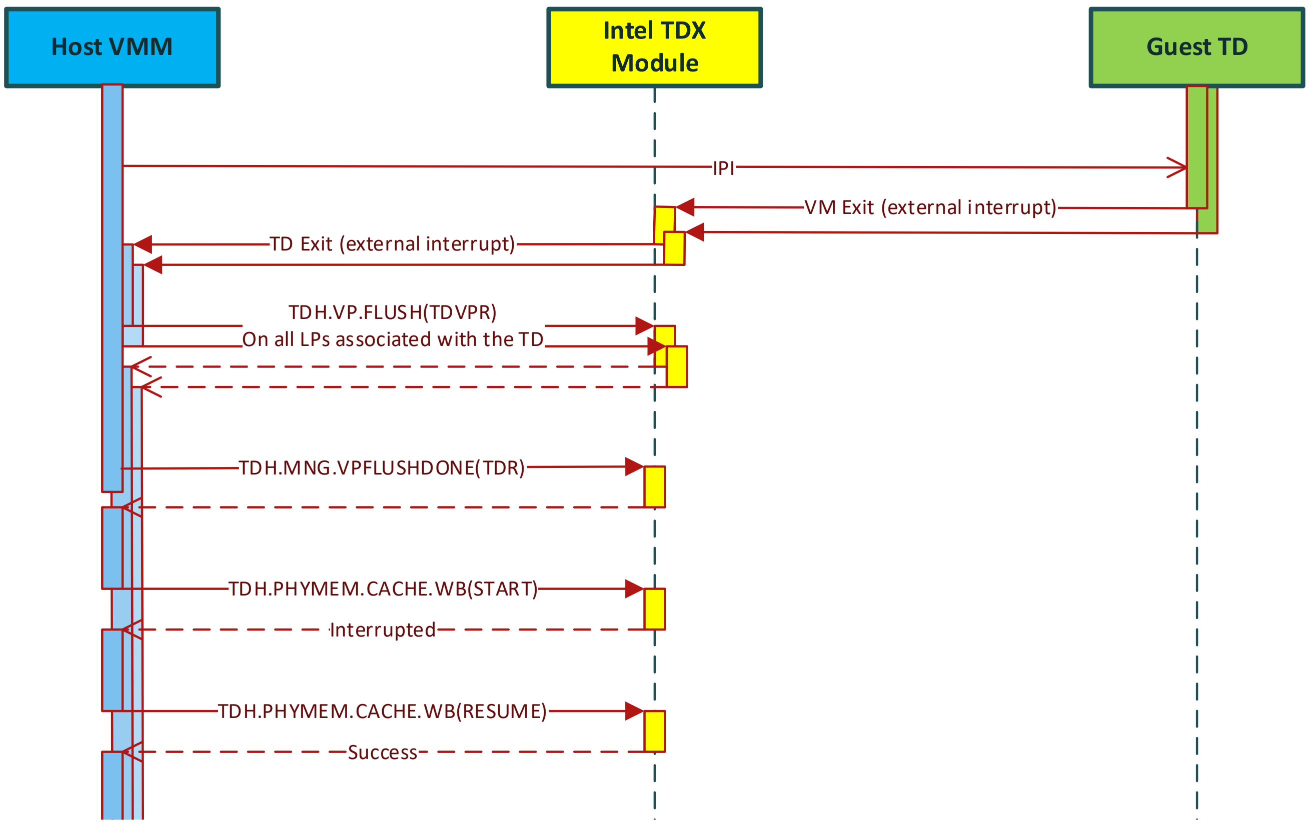

Host VMM
Intel TDX Module
Guest TD
IPI
VM Exit (external interrupt)
TD Exit (external interrupt)
TDH.VP.FLUSH(TDVPR)
On all LPs associated with the TD
TDH.MNG.VPFLUSHDONE(TDR)
TDH.PHYMEM.CACHE.WB(START)
Interrupted
TDH.PHYMEM.CACHE.WB(RESUME)
Success

Figure 3.7: Typical TD Destruction Sequence Step A: Stopping and Flushing Out

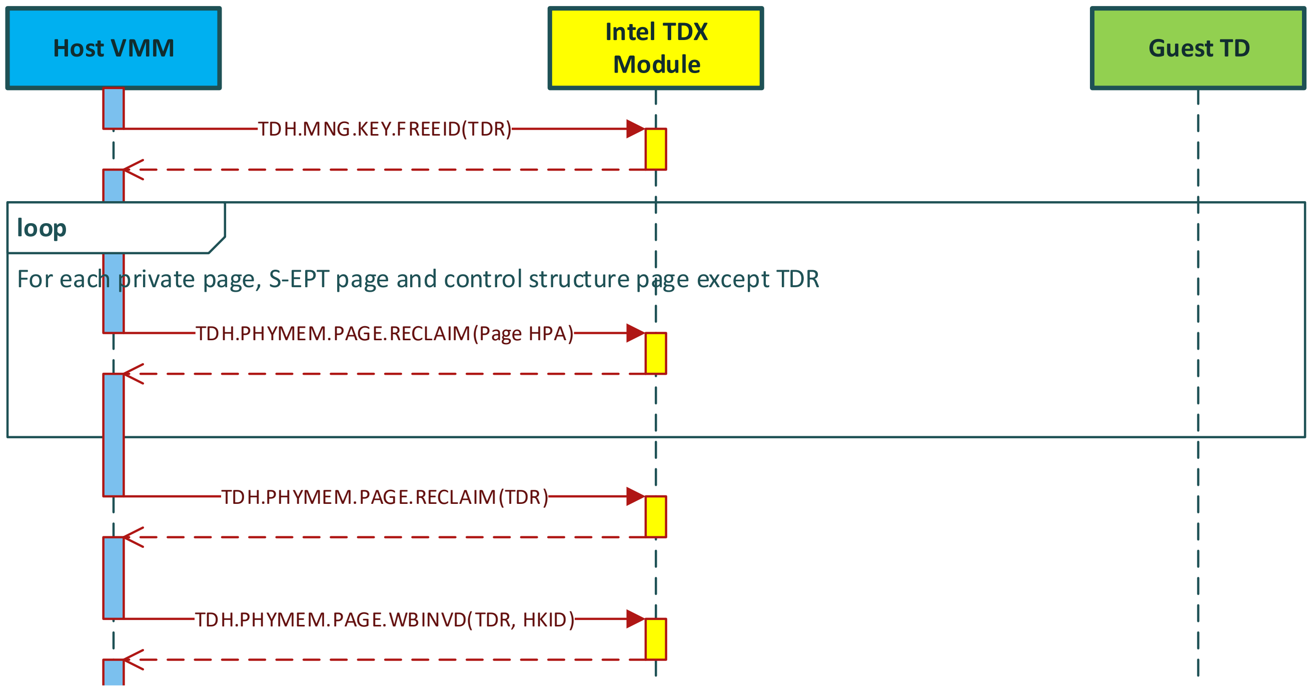

Host VMM
Intel TDX Module
Guest TD
TDH.MNG.KEY.FREEID(TDR)
loop
For each private page, S-EPT page and control structure page except TDR
TDH.PHYMEM.PAGE.RECLAIM(Page HPA)
TDH.PHYMEM.PAGE.RECLAIM(TDR)
TDH.PHYMEM.PAGE.WBINVD(TDR, HKID)

Figure 3.8: Typical TD Destruction Sequence Step B: Resource Reclamation

Table 3.15: Typical TD Destruction Sequence

	Step		Description	SEAMCALL Leaf Functions
A	**TD Stopping and Flushing Out**	1	The host VMM selects a TD to destroy. It sends a virtual interrupt to the TD to shut down gracefully.	TDH.VP.FLUSH TDH.MNG.VPFLUSHDONE TDH.PHYMEM.CACHE.WB
		2	The host VMM broadcasts inter-processor interrupts (IPIs) and must ensure TD exit on all logical processors.	
		3	The host VMM calls the TDH.VP.FLUSH function on all LPs associated with a TD VCPU to flush the TLBs and cached TD VMCS associated with a TD VCPU on those LPs.	
		4	The host VMM calls the TDH.MNG.VPFLUSHDONE function. It checks that the above step executed for all the TD's VCPUs are associated with an LP.	
		5	**Note:** This step may be skipped if TDX_FEATURES0.SKIP_PHYMEM_CACHE_WB (bit 34), readable by TDH.SYS.RD, is 1. The host VMM calls the TDH.PHYMEM.CACHE.WB function on each WBINVD domain to write back to memory the TD contents from all caches. TDH.PHYMEM.CACHE.WB is interruptible by external events. The host VMM should restart it if it indicates it was interrupted, until successfully completed.	
		6	At this point, no address translations or cache lines may exist for this TD except for the TDR page.	
B	**Resource Reclamation**	1	The host VMM calls the TDH.MNG.KEY.FREEID function. It marks the HKID used by the TD as available for other TDs.	TDH.MNG.KEY.FREEID TDH.PHYMEM.PAGE.RECLAIM TDH.PHYMEM.PAGE.WBINVD
		2	For each physical page in TDMR allocated to the TD (TD private pages, Secure EPT pages, and control structures except TDR), the host VMM calls the TDH.PHYMEM.PAGE.RECLAIM function to mark the page as free and initializes its content using MOVDIR64B.	

	Step		Description	SEAMCALL Leaf Functions
		3	The host VMM calls the TDH.PHYMEM.PAGE.RECLAIM function to mark the TDR page as free. The function checks that all other TD physical pages have been reclaimed before.	
		4	Before allocating the reclaimed TDR physical page to any use, the host VMM calls TDH.PHYMEM.PAGE.WBINVD to flush its cache lines and initializes its content using MOVDIR64B.	

SECTION 2:
INTEL TDX MODULE ARCHITECTURE SPECIFICATION

4. Intel TDX Module Lifecycle: Enumeration, Initialization and Shutdown

This chapter discusses the design of the Intel TDX module life cycle: how its capabilities are enumerated by the host VMM, how it is initialized, how it is configured and how it is shut down.

4.1. Overview

TDX module lifecycle use cases are described in 3.1.

4.1.1. Intel TDX Module Lifecycle State Machine

The Intel TDX lifecycle state machine helps track the module's life cycle through the initialization sequence and shutdown.

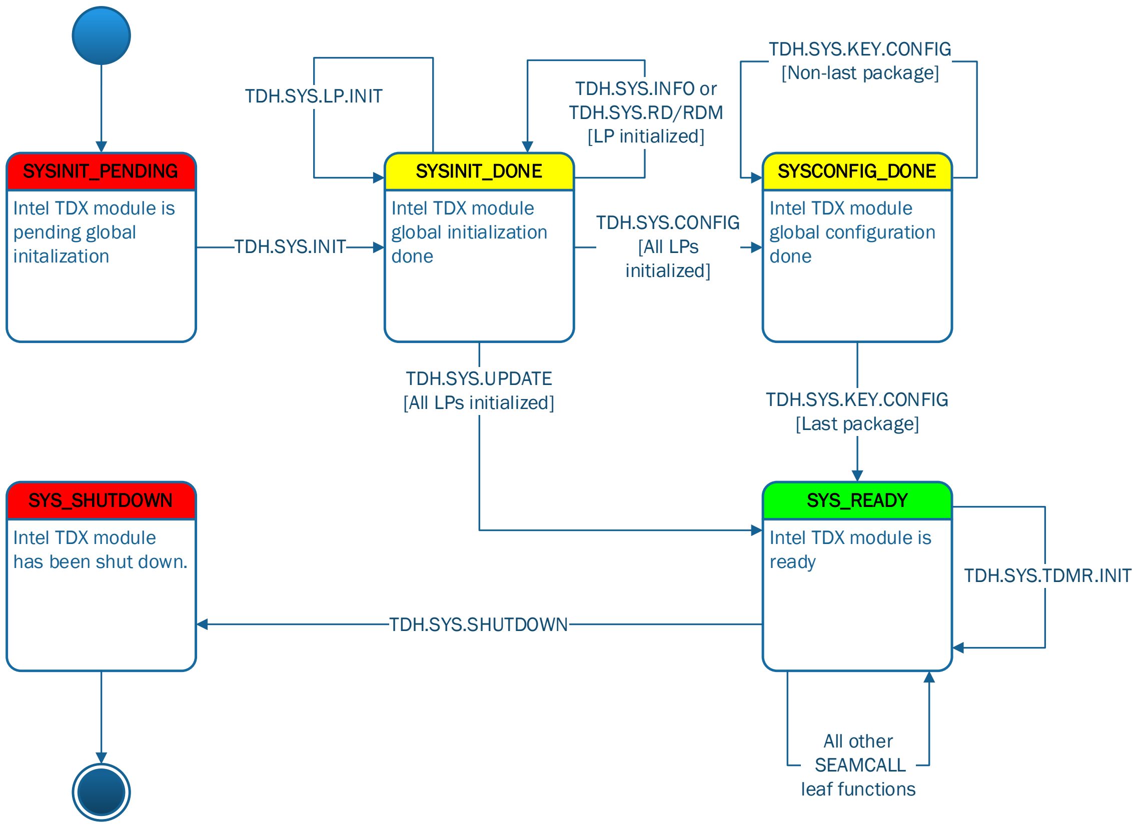

SYSINIT_PENDING
Intel TDX module is pending global initalization
TDH.SYS.INIT
TDH.SYS.LP.INIT
SYSINIT_DONE
Intel TDX module global initialization done
TDH.SYS.INFO or TDH.SYS.RD/RDM [LP initialized]
TDH.SYS.CONFIG [All LPs initialized]
TDH.SYS.KEY.CONFIG [Non-last package]
SYSCONFIG_DONE
Intel TDX module global configuration done
TDH.SYS.UPDATE [All LPs initialized]
TDH.SYS.KEY.CONFIG [Last package]
SYS_SHUTDOWN
Intel TDX module has been shut down.
SYS_READY
Intel TDX module is ready
TDH.SYS.TDMR.INIT
TDH.SYS.SHUTDOWN
All other SEAMCALL leaf functions

Figure 4.1: Intel TDX Module Lifecycle State Machine

Table 4.1: Intel TDX Module Lifecycle States

State Name	Description	Allowed SEAMCALL Leaf Functions
SYSINIT_PENDING	TDH.SYS.INIT has not been called yet.	TDH.SYS.INIT
SYSINIT_DONE	TDH.SYS.INIT has completed successfully. TDH.SYS.LP.INIT must be called on each LP.	TDH.SYS.LP.INIT TDH.SYS.RD* (if current LP has been initialized) TDH.SYS.INFO (if current LP has been initialized) TDH.SYS.CONFIG (if all LPs have been initialized) TDH.SYS.UPDATE (if all LPs have been initialized)
SYSCONFIG_DONE	TDH.SYS.CONFIG has completed successfully. TDH.MNG.KEY.CONFIG must be called on each package.	TDH.SYS.KEY.CONFIG TDH.SYS.RD* TDH.SYS.INFO

State Name	Description	Allowed SEAMCALL Leaf Functions
SYS_READY	The Intel TDX module is ready for use.	Any
SYS_SHUTDOWN	Shutdown operation has been initiated. No new host-side interface functions can be called.	None

4.1.2. Platform Compatibility and Configuration Checking

4.1.2.1. *Overview*

The Intel TDX module is built assuming a certain set of core and platform features. Most platform configuration required to support the Intel TDX module is checked by MCHECK. However, some configurations are designed to be checked by the Intel TDX module. During the initialization process, the Intel TDX module is designed to check that the platform on which it is running is compatible with this core and platform feature set and/or that the same set of features is provided across the platform. Some of the checks are done per core, and some are done per package. Most of the details are part of the Intel TDX module detailed design.

4.1.2.2. *CPU Configuration*

During platform boot, MCHECK verifies all logical CPUs to ensure they meet TDX's security and certain functionality requirements, and MCHECK passes the following CPU configuration information to the NP-SEAMLDR, P-SEAMLDR and the TDX Module:

- Total number of logical processors in the platform.
- Total number of packages installed on the platform.
- A table of per-package CPU family, model and stepping etc. identification, as enumerated by CPUID(1).EAX.

The above information is static and does not change after platform boot and MCHECK run.

Note: TDX doesn't support adding or removing CPUs from TDX security perimeter, as checked by MCHECK. BIOS should prevent CPUs from being hot-added or hot-removed after platform boots.

The TDX module performs additional checks of the CPU's configuration and supported features, by reading MSRs and CPUID information as described in the following sections.

4.1.2.3. *MSR Sampling and Checks*

TDH.SYS.INIT reads and checks the contents of some MSRs. In many cases, the MSR value read by TDH.SYS.INIT is also checked for consistency (i.e., having the same values) by TDH.SYS.LP.INIT. In other cases, TDH.SYS.LP.INIT may perform additional checks.

4.1.2.4. *CPUID Sampling, Checks and Enumeration*

Note: CPUID virtualization is described in 11.11.

The TDH.SYS.INIT and TDH.SYS.LP.INIT functions sample CPUID leaf and sub-leaf return values. This is intended to check compatibility with the Intel TDX module and with any guest TD operation. If any of these checks fail, Intel TDX module initialization is designed to fail.

The TDH.SYS.RD, TDH.SYS.RDALL and TDH.SYS.INFO functions may be called by the host VMM to enumerate the directly configurable and allowable CPUID fields.

4.1.3. Physical Memory Configuration Overview

Configuration of the physical memory available to the Intel TDX module (TDMRs) and its associated metadata (PAMT arrays) is done using the TDH.SYS.CONFIG function. If supported by the TDX module, this function can set dynamic PAMT mode.

4.1.3.1. *Intel TDX ISA Background: Convertible Memory Ranges (CMRs)*

A 4KB memory page is defined as **convertible** if it can be used to hold an Intel TDX private memory page or any Intel TDX control structure pages while helping guarantee Intel TDX security properties (i.e., if it can be **converted** from a Shared page to a private page).

Convertible Memory Ranges (**CMRs**) are defined as contiguous convertible physical address ranges, declared by BIOS. CMRs are checked by MCHECK during platform boot to help ensure their configuration matches TDX security. All memory within each CMR must be convertible and must be present when checked by MCHECK during platform boot. CMRs are static and do not change after platform boot and checking by MCHECK.

Note: The above definition implies that TDX does not support hot-plugin or hot-removal of convertible memory. BIOS should prevent hot removal of convertible memory after platform boot.

CMRs have the following characteristics:

- CMR configuration is “soft” – no hardware range registers are used.
- Each CMR defines a single contiguous physical address range.
- All the memory within each CMR is convertible, and it must comply with the rules checked by MCHECK.
- Each CMR has its own size. CMR size is a multiple of 4KB, and it is **not** required to be a power of two.
- CMRs cannot overlap with each other.
- CMRs must reside within the effective physical address range of the platform (after considering the most significant PA bits stolen for holding Key IDs).
- CMRs are configured at platform scope (no separate configuration per package).
- The maximum number of CMRs is implementation specific. It is not explicitly enumerated; it is deduced from Family/Model/Stepping information provided by CPUID. The current maximum number of CMRs is 32.
- CMRs are available on systems with TDX ISA capabilities, as enumerated by bit 5 of the IA32_VMX_PROCBASED_CTLS3 MSR.
- CMR configuration is checked by MCHECK and cannot be modified afterwards.

MCHECK stores the CMR table, and other platform topology information, in a pre-defined location in SEAM range, so it can be read later and trusted by the P-SEAMLDR module. On TDX module installation, P-SEAMLDR copies MCHECK data to another page in the SEAM range, which is accessible by the Intel TDX module.

4.1.3.2. *TDMRs and PAMT Arrays Configuration*

TDMRs and PAMTs are described in 8.1. This section provides an overview of their configuration and their relationships to CMRs.

By default, the TDX module uses a static PAMT configuration. If supported by the TDX module, it can be configured to use dynamic PAMT allocation.

4.1.3.2.1. Background: Reserved Areas within TDMRs

As described in 8.1, the Intel TDX module physical memory management is done using PAMT Blocks – each holding the metadata of a 1GB block of TDMR. This implies that TDMR granularity must be 1GB.

However, there is a requirement for the host VMM to be able to allocate memory at granularities smaller than 1GB. This is especially important in systems that have a relatively small amount of memory.

To support the two requirements above, the Intel TDX module’s design allows arbitrary reserved areas within TDMRs. Reserved areas are still covered by PAMT. However, during initialization their respective PAMT entries are marked with a PT_RSVD page type, so pages in reserved areas are not used by the Intel TDX module for allocating privately encrypted memory pages (but they can be used for PAMT areas, see below).

If the TDX module has been configured for static PAMT, reserved areas must be aligned to 4KB, and their size must be a multiple of 4KB. Else (dynamic PAMT has been configured), reserved areas must be aligned on 2MB, and their size must be a multiple of 2MB.

Only the non-reserved parts of a TDMR are required to be inside CMRs.

4.1.3.2.2. Background: Three PAMT Areas

As described in 8.1, when static PAMT is used, a logical PAMT Block is composed of 1 PAMT_1G entry, 512 PAMT_2M entries and 512^2 PAMT_4K entries or 512^2 PAMT_PAGE_BITMAP entries, depending on whether static PAMT or dynamic PAMT is used. Thus, the overall size of a PAMT Block, and as a result of the whole PAMT, is not a power of 2.

However, the host VMM may only be able to allocate memory buffers for PAMT in sizes that are power of 2.

To enable this, buffers for PAMT_1G entries, PAMT_2M entries and PAMT_4K entries are allocated separately. As a result, if the host VMM allocates a TDMR whose size is a power of 2, its three respective PAMT areas will also have sizes that are a power of 2.

PAMT areas are required to be inside CMRs because PAMT is encrypted with a private HKID.

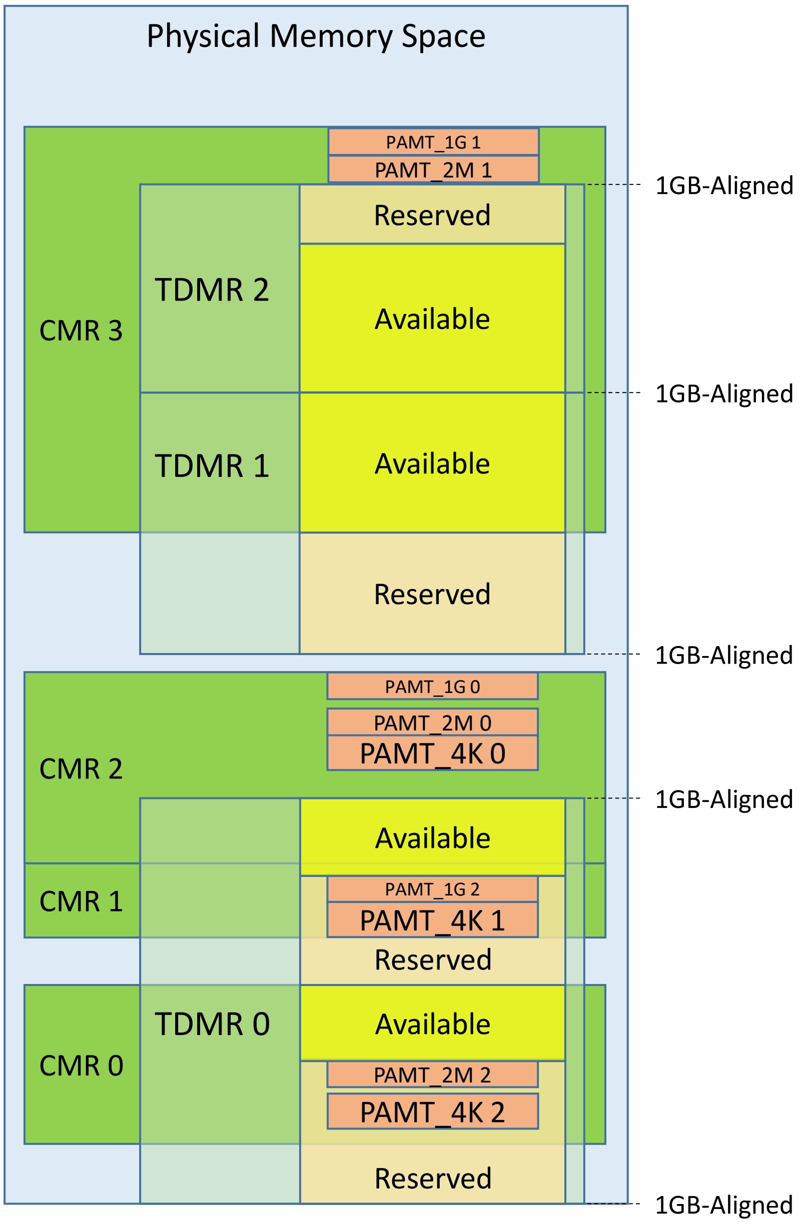


Figure 4.2: Example of Convertible Memory Ranges (CMRs) vs. Trust Domain Memory Regions (TDMRs)

4.1.3.2.3. Configuration Rules

In addition to the rules described in 8.1, the following rules apply to TDMR configuration as related to CMRs:

- Any non-reserved 4KB page within a TDMR must be convertible – i.e., it must be within a CMR.
- Reserved areas within a TDMR need not be within a CMR.

Three PAMT areas must be configured for each TMDR – one for each physical page size controlled by PAMT:

- Area for PAMT_4K (for static PAMT) or PAMT_PAGE_BITMAP (for dynamic PAMT) entries
- Area for PAMT_2M entries
- Area for PAMT_1G entries

PAMT areas have the following attributes:

- A PAMT area size is directly proportional to the TDMR with which it is associated. The size ratio is enumerated by TDH.SYS.RD/RDALL or TDH.SYS.INFO. Note that the size ratio may be different for each of the 3 PAMT array types and for PAMT_PAGE_BITMAP.
- A PAMT area must reside in convertible memory – i.e., each PAMT area page must be a CMR page.
- PAMT areas must not overlap with TDMR non-reserved areas; however, they may reside within TDMR reserved areas (as long as these are convertible).
- PAMT areas must not overlap with each other.

4.1.4. TDX Module Extension Overview

TDX module versions which support some TDX features, such as TDX Connect, implement an extension which includes the following:

Non-Root Extension (NRX)	Extends the functionality of the TDX module, by implementing some of its functionality as one or more guests running in SEAM non-root mode. Technically, each NRX guest is a TD. However, logically and from the usage perspective it is part of the TDX module and is hidden from the host VMM. The NRX TD images are embedded in the TDX module image. The TDX module builds the NRX TDs as part of the module's initialization sequence and calls them when their functionality is required for execution of some TDX module interface functions.
Memory Pool	To support the TDX module extension, the TDX module holds a pool of 4KB physical pages. The host VMM, as part of the TDX module initialization flow, allocates memory for the memory pool, based on the required pool size calculated and provided to the host VMM by the TDX module. All pages in the memory pool are considered equal, e.g., the TDX module does not track their package affinity.

4.2. Intel TDX Module Initialization Interface

4.2.1. Global Initialization: TDH.SYS.INIT

TDH.SYS.INIT is intended to globally initialize the Intel TDX module. It works as follows:

1. Initialize Intel TDX module global data.
2. Sample and check platform features that need to be checked for platform-wide compatibility – i.e., the Intel TDX module supports several options, but they must be the same across platform. These are later checked on each LP.
3. Sample and check the platform configuration on the current LP. For example, TDH.SYS.INIT samples SMRR and SMRR2, checks they are locked and do not overlap any CMR, and stores their values to be checked later on each LP.
4. Set the system state to SYSINIT_DONE.

For a detailed description of TDH.SYS.INIT, see the [TDX Module ABI Spec].

4.2.2. LP-Scope Initialization: TDH.SYS.LP.INIT

TDH.SYS.LP.INIT is intended to perform LP-scope, core-scope and package-scope initialization of the Intel TDX module. It can be called only after TDH.SYS.INIT completes successfully, and it can run concurrently on multiple LPs. At a high level, TDH.SYS.LP.INIT works as follows:

1. Do a global EPT flush (INVEPT type 2).
2. Initialize Intel TDX module LP-scope data.
3. Check features and configuration compatibility and uniformity – once per LP, core or package, depending on the scope of the checked feature or configuration:
   3.1. Check features compatibility with the Intel TDX module.
   3.2. Check configuration uniformity.

For a detailed description of TDH.SYS.LP.INIT, see the [TDX Module ABI Spec].

4.2.3. TDX Module Enumeration: TDH.SYS.RD/RDALL and TDH.SYS.INFO

Once an LP has been initialized, the host VMM can call TDH.SYS.RD, TDH.SYS.RDALL or TDH.SYS.INFO on that LP to help enumerate the Intel TDX module capabilities and platform configuration.

TDH.SYS.RD and TDH.SYS.RDALL are the recommended enumeration methods. They enable the host VMM to read the values of TDX module global metadata fields, enumerating the TDX module capabilities. The list of fields is described in the [TDX Module ABI Spec].

To read all host readable TDX Module fields, the host VMM can invoke TDH.SYS.RDALL. This function returns the information as a metadata list.

To read a single TDX Module field, TDH.SYS.RD can be invoked. It returns the next host-readable field identifier, thus it can also be used to enumerate the TDX Module by calling it in a loop, starting from field identifier value of 0, until it returns a next field identifier value of 0.

TDH.SYS.INFO is provided for backward compatibility with previous TDX module versions:

- Intel TDX module capabilities are enumerated in the returned TDSYSINFO_STRUCT (see the [TDX Module ABI Spec]).
- Convertible Memory Ranges (CMRs), as previously set by BIOS and checked by MCHECK, are enumerated in the returned CMR_INFO table.

For a detailed description of interface functions and metadata fields, see the [TDX Module ABI Spec].

4.2.4. TDH.SYS.CONFIG: TDX Module Global Configuration

After performing global and LP-scope initialization, the host VMM can call TDH.SYS.CONFIG to globally configure the Intel TDX module, providing the following information:

- PAMT mode (static or dynamic).
- **TDMR and PAMT Table,** where each entry contains a TDMR base address, size and corresponding PAMT reserved area base address and size. Refer to 8.1 for definition of TDMRs.
- The **HKID** to be used by the Intel TDX module for its global private key, used for encrypting PAMT and TDRs.

TDH.SYS.CONFIG is also used for enabling some TDX features, such as TDX Connect.

For a detailed description of the table format (TDMR_INFO) and TDH.SYS.CONFIG, see the [TDX Module ABI Spec].

4.2.5. TDH.SYS.KEY.CONFIG: Key Configuration (per Package)

After performing global configuration, the host VMM calls TDH.SYS.KEY.CONFIG to perform package-scope configuration of the Intel TDX module's global private key on the hardware.

For a detailed description of TDH.SYS.KEY.CONFIG, see the [TDX Module ABI Spec].

4.2.6. State Restoration after TD-Preserving TDX Module Update: TDH.SYS.UPDATE

When updating the TDX module, the host VMM can call TDH.SYS.UPDATE after initializing the new TDX module on all LPs, so that the new TDX module will update itself with the handoff data prepared by a call to TDH.SYS.SHUTDOWN on the previous TDX module. This allows "old" TDs, which were created by previous TDX module(s), to keep running under the supervision of the new TDX module.

TDH.SYS.UPDATE is also used for enabling some TDX features, such as TDX Connect, which may be enabled on update.

If TDH.SYS.UPDATE returns successfully, the TDX module is ready, and TDH.SYS.CONFIG and TDH.SYS.KEY.CONFIG cannot (and need not) be called anymore.

If TDH.SYS.UPDATE returns error, then the host VMM can continue with the non-update sequence (TDH.SYS.CONFIG, TDH.SYS.KEY.CONFIG etc.). In this case all existing TDs are lost. Alternatively, the host VMM can request the P-SEAMLDR to update to another TDX module. If that update is successful, existing TDs are preserved.

For a detailed description of TDH.SYS.UPDATE, see the [TDX Module ABI Spec].

4.3. TDMR and PAMT Initialization

TDMR and PAMT initialization procedure is designed to be performed **during VMM run-time**, after VMM boot. The host VMM should be able to work normally while initialization takes place, at any time using memory that has already been initialized. At a high level, TDMR initialization has the following characteristics:

- Initialization is performed gradually.
- Initialization function TDH.SYS.TDMR.INIT adheres to the latency rules of most Intel TDX functions – i.e., they take no more than a predefined number of clock cycles.

- Initialization function TDH.SYS.TDMR.INIT **can run concurrently on multiple LPs** if each concurrent flow initializes a **different TDMR**.
- After each 1GB page of a TDMR has been initialized, that 1GB page becomes available for use by any Intel TDX function that creates a private TD page or a control structure page – e.g., TDH.MEM.PAGE.ADD, TDH.VP.ADDCX, etc.

For each TDMR, the VMM should execute a loop of **TDH.SYS.TDMR.INIT** providing the TDMR start address (at 1GB granularity) as an input.

TDH.SYS.TDMR.INIT initializes an (implementation-defined) number of PAMT entries. The maximum number of PAMT entries to be initialized is designed to avoid latency issues. Initialization uses direct writes (MOVDIR64B).

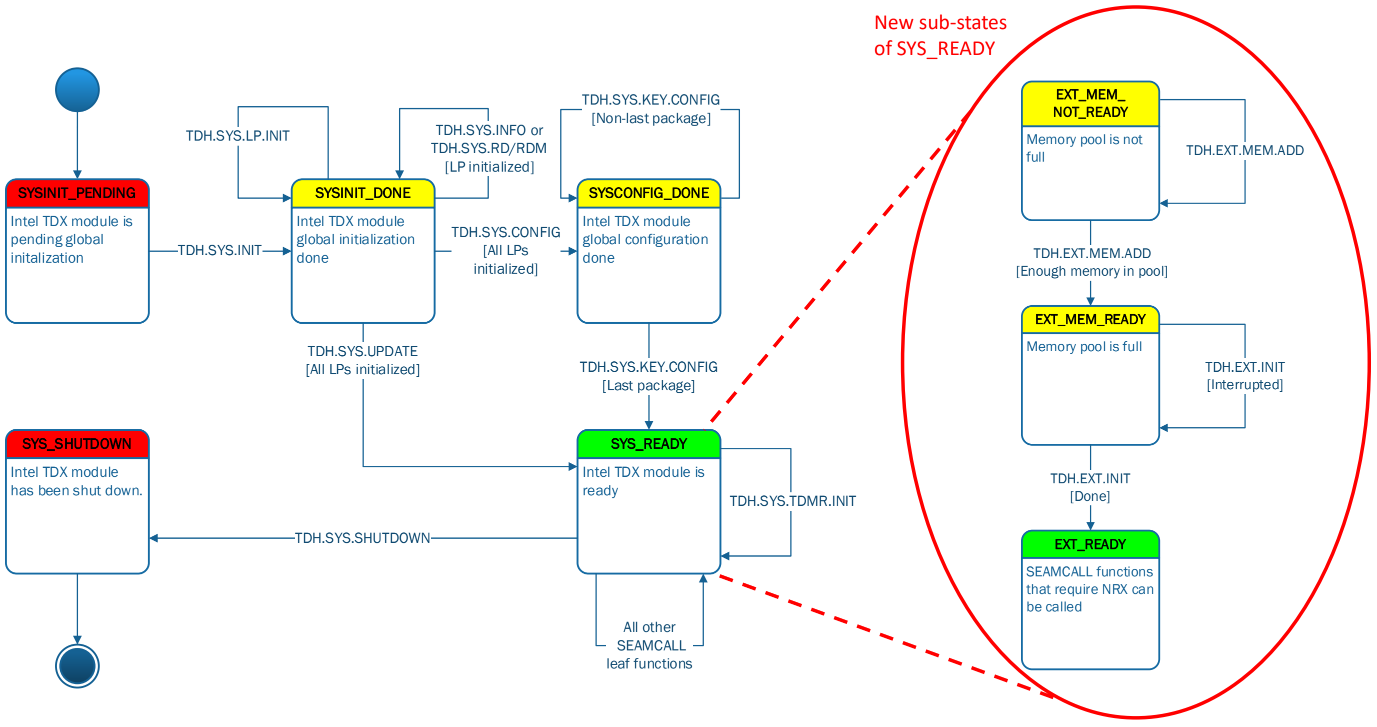
Once the PAMT for each 1GB block of TDMR has been fully initialized, TDH.SYS.TDMR.INIT marks that 1GB block as ready for use; that means 4KB pages in this 1GB block may be converted to private pages – e.g., by TDH.MEM.PAGE.ADD. This can be done concurrently with adding and initializing other TDMRs.

For a detailed description of TDH.SYS.TDMR.INIT, see the [TDX Module ABI Spec].

4.4. TDX Module Extension Initialization

The following sequence is typically used to initialize the TDX module extension:

1. The host VMM configures the desired TDX features, such as TDX Connect. This is done as part of the TDX module initialization sequence or as part of the TDX module update sequence, as an input parameter to TDH.SYS.CONFIG or TDH.SYS.UPDATE respectively.
2. Based on the enabled features, the TDX module checks whether a memory pool is required and if so, calculates its required size.
3. The host VMM reads MEMORY_POOL_REQUIRED_PAGES, the number of missing TDX module's memory pool pages, using TDH.SYS.RD.
4. Once the TDX module has been initialized (TDH.SYS.KEY.CONFIG was called on all packages), the host VMM can call TDH.EXT.MEM.ADD multiple times to add the required number of memory pages to the TDX module's memory pool.
5. The host VMM reads EXT_REQUIRED, which indicates whether the TDX module extension is required to be initialized, using TDH.SYS.RD.
6. If required, the host VMM can then call TDH.EXT.INIT to initialize the TDX module extension.


New sub-states of SYS_READY
SYSINIT_PENDING
Intel TDX module is pending global initalization
TDH.SYS.LP.INIT
TDH.SYS.INIT
SYSINIT_DONE
Intel TDX module global initialization done
TDH.SYS.INFO or TDH.SYS.RD/RDM [LP initialized]
TDH.SYS.CONFIG [All LPs initialized]
TDH.SYS.KEY.CONFIG [Non-last package]
SYSCONFIG_DONE
Intel TDX module global configuration done
TDH.SYS.UPDATE [All LPs initialized]
TDH.SYS.KEY.CONFIG [Last package]
SYS_READY
Intel TDX module is ready
TDH.SYS.TDMR.INIT
All other SEAMCALL leaf functions
SYS_SHUTDOWN
Intel TDX module has been shut down.
TDH.SYS.SHUTDOWN
EXT_MEM_ NOT_READY
Memory pool is not full
TDH.EXT.MEM.ADD
TDH.EXT.MEM.ADD [Enough memory in pool]
EXT_MEM_READY
Memory pool is full
TDH.EXT.INIT [Interrupted]
TDH.EXT.INIT [Done]
EXT_READY
SEAMCALL functions that require NRX can be called

7.

Figure 4.3: TDX Module Extension Initialization as Part of the TDX Module Lifecycle

To avoid long initialization latency, most TDX module interface functions, which are not dependent on the TDX module extension, may be called regardless of the extension initialization. Interface functions defined as dependent on the TDX module extension can only be called after the extension initialization is done; calling such functions before that fails with a TDX_EXT_NOT_INITIALIZED status code.

4.5. Intel TDX Module Shutdown

4.5.1. Shutdown Initiated by the Host VMM (as Part of Module Update)

The host VMM can initiate Intel TDX module shutdown at any time by calling the TDH.SYS.SHUTDOWN function. This is intended for use as part of updating the Intel TDX module without going through a warm or cold reset sequence. TDH.SYS.SHUTDOWN is designed to set state variables to block all SEAMCALL leaf functions.

TDH.SYS.SHUTODWN also prepares handoff data in a designated area in SEAM range called handoff data range. The handoff data contains any TDX module state (in SEAM range) required to preserve old TDs across TDX module updates. This includes (but not limited to) Key Ownership Table (KOT) and TDMR management table. When a new TDX module is installed using the "update" scenario, the P-SEAMLDR module preserves (doesn't wipe-out) the handoff data range. The new TDX module can then update itself (when the TDH.SYS.UPDATE function is called) from that handoff data.

4.5.2. Shutdown Initiated by a Fatal Error

By design, fatal errors during Intel TDX module execution cause an immediate SEAM shutdown. Subsequent SEAMCALLs on any LP fail with a VMfailInvalid indication (RFLAGS.CF set to 1). This situation can only be recovered by a platform reset.

4.6. Intel TDX Module Handoff Data

In order to preserve TDs across Intel TDX module updates, the old TDX module's TDH.SYS.SHUTDOWN function prepares handoff data in the handoff data region, to be consumed by the new TDX module's TDH.SYS.UPDATE function. The handoff data contains all variables maintained inside the SEAM range (i.e., module's global and local data) related to TD management. Note that TD code, data and metadata pages residing out of SEAM range need not be passed, since the DRAM contents and MKTME state are not impacted.

Since it's possible – and normally expected – that the new TDX module would differ from the old TDX module it's replacing, it's necessary that the new TDX module will understand the syntax and semantics of all TD metadata – either in the handoff data range and out of SEAM range – that the old TDX module left behind, including PAMT, Secure EPTs, TDRs, TDCS and TDVPS pages. To support that, a handoff protocol is needed.

The handoff protocol is based on the notion of **Handoff Version** (HV) – an unsigned 16-bit number which identifies the contents, syntax and semantics of all TD metadata fields, in the handoff data region and out of SEAM range, that the old TDX module passes to the new TDX module. The TDX module may understand one or more HVs. This allows the new module to consume handoff data prepared by an older TD module and "upgrade" it with new contents (e.g., put data in TDVPS fields that were previously reserved).

In preparation to TD-preserving TDX module update, the host VMM calls the TDH.SYS.SHUTDOWN function with "requested HV" parameter. If the TDX module understands the requested HV and was not built as "non-downgradable" (see below), then the TDH.SYS.SHUTDOWN *prepares* handoff data with the requested syntax and semantics in the handoff data region; it marks the handoff data region as valid with the requested HV (see [P-SEAMLDR FAS] for the structure of the 64-bit handoff data region's header that contains this information).

After P-SEAMLDR updated the TDX module in SEAM range, the host VMM initializes the new TDX module, and calls its TDH.SYS.UPDATE function which *consumes* the handoff data and marks it as invalid.

Specifically, each TDX module is built with the following constants:

Table 4.2: TDX Module Handoff Constants

Name	Meaning	Description
MODULE_HV	Module Handoff Version	Handoff version that this TDX module works with
MIN_UPDATE_HV	Minimum Updatable Handoff Version	The "oldest" HV this TDX module understands
NO_DOWNGRADE	No-Downgrade Flag	A non-zero value indicates that this TDX module cannot "downgrade" the data is leaves behind to a lower handoff version

The above constants must satisfy the inequality 0 <= MIN_UPDATE_HV <= MODULE_HV. If MIN_UPDATE_HV < MODULE_HV, then this TDX module can consume (in TDH.SYS.UPDATE) older handoff data (i.e., data whose syntax/semantics has lower HV than the syntax/semantics this TDX module was built to work with). In addition, if the NO_DOWNGRADE flag is zero, then this TDX module can generate (in TDH.SYS.SHUTDOWN) older handoff data.

The following table illustrates this protocol with several examples.

Table 4.3: TDX Module Handoff Protocol Examples

Old TDX Module's Parameters	New TDX Module's Parameters	Requested HV to Old TDX Module's TDH.SYS. SHUDOWN	HV of Handoff Data Prepared by Old TDX Module	Module Update by P-SEAMLDR	New TDX Module's TDH.SYS.UPDATE Action
MODULE_HV=10 MIN_UPDATE_HV=10 NO_DOWNGRADE=0	MODULE_HV=10 MIN_UPDATE_HV=10 NO_DOWNGRADE=0	10	10	Installed	Consume as is
MODULE_HV=10 MIN_UPDATE_HV=10 NO_DOWNGRADE=0	MODULE_HV=10 MIN_UPDATE_HV=10 NO_DOWNGRADE=0	11	None (shutdown failed – requested HV is too large)	Not installed (invalid handoff data) Note: the host VMM can install using "Load" scenario	Fail (invalid handoff data) Note: the host VMM should re-configure the new TDX module (TDs are not preserved)
MODULE_HV=10 MIN_UPDATE_HV=10 NO_DOWNGRADE=0	MODULE_HV=10 MIN_UPDATE_HV=10 NO_DOWNGRADE=0	9	None (shutdown failed – requested HV is too small)	Not installed (invalid handoff data) Note: the host VMM can install using "Load" scenario	Fail (invalid handoff data) Note: the host VMM should re-configure the new TDX module (TDs are not preserved)
MODULE_HV=10 MIN_UPDATE_HV=10 NO_DOWNGRADE=0	MODULE_HV=11 MIN_UPDATE_HV=10 NO_DOWNGRADE=0	10	10	Installed	Upgrade to 11 and consume
MODULE_HV=11 MIN_UPDATE_HV=10 NO_DOWNGRADE=0	MODULE_HV=10 MIN_UPDATE_HV=10 NO_DOWNGRADE=0	10	10 (downgraded)	Installed	Consume as is
MODULE_HV=11 MIN_UPDATE_HV=10 NO_DOWNGRADE=0	MODULE_HV=10 MIN_UPDATE_HV=10 NO_DOWNGRADE=0	11	11	Not installed (HV-incompatible module) Note: the host VMM can install using "Load" scenario.	Fail (invalid handoff data) Note: the host VMM should re-configure the new TDX module (TDs are not preserved).
MODULE_HV=11 MIN_UPDATE_HV=10 NO_DOWNGRADE=1	MODULE_HV=10 MIN_UPDATE_HV=10 NO_DOWNGRADE=0	10	Shutdown failure (can't downgrade)	Not installed (invalid handoff data) Note: the host VMM can install using "Load" scenario.	Fail (invalid handoff data) Note: the host VMM should re-configure the new TDX module (TDs are not preserved).

4.7. TDX Module Fatal Error Handling

4.7.1. Overview

A TDX module fatal error happens when the TDX module detects some unexpected behavior. Some examples are:

- An unexpected exception happens during TDX module execution.
- Some sanity check code in the TDX module detects a situation that shouldn't have happened.

A TDX module fatal error condition ends in an unbreakable shutdown, which impacts the current LP and results in a SEAM shutdown, which prevents further SEAMCALLs on any LP.

If supported by the TDX module, the host VMM may configure it to provide diagnostic information in memory and optionally to broadcast a notification interrupt in case they encounter a fatal error, before entering a shutdown state.

When the TDX module detects a fatal error, it does the following:

1. If so configured, log diagnostic information to a shared memory buffer.
2. If so configured, broadcast a notification interrupt to all other LPs.
3. Induce an unbreakable shutdown.

There are some cases, such as poison consumption while in SEAM root mode, which directly result in an unbreakable shutdown and a SEAM shutdown without being able to log diagnostics and/or issue a notification interrupt:

- A poison consumption by the core while the TDX module executes (in SEAM root mode), which may be, e.g., the result of a TDX control structure being overwritten by non-TDX software, causes an immediate unbreakable shutdown and a SEAM shutdown.
- Due to the way the TDX module is implemented, certain unexpected exceptions (#PF and #DF) cause an immediate unbreakable shutdown and a SEAM shutdown.
- Due to the way the TDX module is implemented, diagnostic information logging may not be possible during global initialization (TDH.SYS.INIT) and part of the LP-scope initialization (TDH.SYS.LP.INIT).

4.7.2. FATAL_INFO: Fatal Error Diagnostic Information

FATAL_INFO is a 64-bytes diagnostic information structure, written into memory by the TDX module when it detects a fatal error. The host VMM should only be aware of FATAL_INFO's first byte, which contains state information. Other information in FATAL_INFO is intended for use by Intel to help analyze the root cause of the fatal error.

For a detailed definition of FATAL_INFO, see the [ABI Spec].

4.7.3. Expected Host VMM and BIOS Behavior on TDX Module Fatal Error

Host VMM

1. If a notification interrupt has been configured, then the reception of such an interrupt on a certain LP indicates that the TDX module on another LP may have entered SEAM shutdown.
2. In a SEAM shutdown condition, SEAMCALL on any LP results in a VMfailInvalid (RFLAGS.CF == 1) error. Such error can be caused by multiple other conditions, but for a well-behaving host VMM, this error should only happen due to a SEAM shutdown.
3. If a SEAM shutdown is suspected, the host VMM can check whether FATAL_INFO.STATE indicates a valid fatal error information is available.
4. If FATAL_INFO.STATE does not indicate a valid fatal error information, the host VMM can check the machine check banks to see if a poison was consumed while the TDX module was running due to memory corruption.
5. If no machine check event happened, it is still possible that a SEAM shutdown happened without logging any diagnostic information. Depending on the chosen implementation method, some unexpected exception cases may cause a SEAM shutdown without logging any diagnostics.
6. The host VMM typically logs the fatal error diagnostics information to some persistent storage for later analysis, then reboots the platform.

BIOS

When entering an unbreakable shutdown state, the CPU increments a counter, readable by BIOS using SMM_BLOCKED (MSR 0x4E3). BIOS is expected to use this value in order to avoid waiting indefinitely for LPs that are in an unbreakable shutdown state.

Note: Older CPUs and/or ucode patch versions may fail to increment the SMM_BLOCKED counter. In this case, an unbreakable shutdown typically results in a system hang when BIOS attempts to wait for all LPs.

5. Memory Encryption Key Management

5.1. Objectives

The main goal of Intel TDX key management is to enable the VMM to perform the following:

- Manage HKID space as a limited platform resource, assigning HKIDs to TDs and reclaiming them as required.
- Enable the Intel TDX module to use a global ephemeral key for encrypting its data (e.g., PAMT).
- Enable each TD to use its own ephemeral key.

The Intel TDX interface functions are designed to provide the required building blocks and help ensure that software cannot perform operations that are not compliant with TDX security objectives, as follows:

1. Help ensure that only HKID values that have been configured for TDX private memory encryption keys can be assigned to TDs, and that those HKID values cannot be used by non-TD software or devices.
2. Prevent assignment of the same HKID to more than one TD.
3. At the time an HKID is assigned to a TD, there must be no modified cache lines – at any level, for any core – for that HKID. All such cache lines that may have held modified data have been written to memory (if required). Note that this requirement applies only to TDX private HKID and not to legacy MKTME HKIDs.
4. TD memory may be accessed, and the TD may run, only when the following conditions are met:
   4.1. An HKID has been assigned for the TD's ephemeral key.
   4.2. The encryption key has been configured for all the TD's ephemeral HKID, on all crypto engines, on all packages.

5.2. Background: HKID Space Partitioning

Since the same MKTME encryption engines and the same set of encryption keys are used for legacy MKTME operation and for TDX operation, TDX ISA enables the enumeration and partitioning of the activated HKID space between the two technologies. As designed, the encryption keys and their associated HKIDs are divided into three ranges, as shown in Table 5.1 below. The values of NUM_HKID_KEYS and NUM_TDX_PRIV_KEYS are read from the IA32_MKTME_KEYID_PARTITIONING MSR (0x87).

Private HKIDs and private keys are designed to be fully controlled by the Intel TDX module and are the subject of this chapter.

Note: To be configured for dynamic PAMT (if supported by the TDX module), NUM_HKID_KEYS may be required to be at least a certain minimum. The minimum number of HKID bits for dynamic PAMT is enumerated by MIN_DYNAMIC_PAMT_NUM_HKID_BITS, readable by TDH.SYS.RD*.

Table 5.1: HKID Space Partitioning

	HKID	**Key**
Shared HKIDs	0	Legacy TME key, shared
	1	Legacy MKTME key #1
	2	Legacy MKTME key #2
	…	…
	NUM_HKID_KEYS	Last legacy MKTME key
Private HKIDs	NUM_HKID_KEYS + 1	Private key of a specific TD
	NUM_HKID_KEYS + 2	Private key of a specific TD
	NUM_HKID_KEYS + 3	Private key of a specific TD
	…	…
	NUM_HKID_KEYS + NUM_TDX_PRIV_KIDS	Private key of a specific TD

5.3. WBINVD Domains

Enumeration: TDH.PHYMEM.CACHE.WB is not required if TDX_FEATURES0.SKIP_PHYMEM_CACHE_WB (bit 34), readable by TDH.SYS.RD, is 1.

5.3.1. Overview

TDX memory encryption key management requires flushing caches. The TDH.PHYMEM.CACHE.WB interface function (as well as the CPU instructions WBINVD and WBNOINVD) flush caches in the **WBINVD domain** associated with the LP on which they execute. The extent of each WBINVD domain, i.e., which LPs belong to it, depends on the CPU architecture. For older processors, a WBINVD domains includes all LPs single package. For newer processors, a WBINVD domain may include a group of LPs within a package. TDX operations that involve TDH.PHYMEM.CACHE.WB requires it to be executed on one LP per WBINVD domain in the platform.

5.3.2. Host VMM Enumeration of WBINVD Domains

The host VMM can use the algorithm described in [Processor Topology Enumeration] to enumerate the WBINVD domain on the platform. The following description summarizes the operation.

Do detect the WBINVD domains on the platform, do the following for each LP, identified by its x2APIC ID:

1. Find the last cache level for this LP, by iterating on CPUID(4,N) starting from N=0 until the Cache Type returned in EAX[4:0] is Null (0).
2. If L is the last cache level, then LogicalProcessorsSharingCacheP2, the maximum number of x2APIC IDs sharing this cache is provided in CPUID(4,L).EAX[25:14], rounded up to the nearest power of 2.
3. Calculate CACHE_MASK by ~(LogicalProcessorsSharingCacheP2-1).
4. CACHE_ID is calculated by bitwise-and the current LP's x2APIC ID with CACHE_MASK.
5. If CACHE_ID is new, add it and the associated CACHE_MASK to the list of WBINVD domains.

For any given LP, to determine the associated WBINVD domain, scan the WBINVD domains list, and for each entry:

1. Bitwise-and the current LP's x2APIC ID with CACHE_MASK.
2. If the result equals CACHE_ID, then the current LP belongs to this WBINVD domain.

5.3.3. Enumerating Non-Package WBINVD Domains Support

The TDX module indicates that it supports multiple WBINVD domains per package and is running on a CPU where this is indeed the case, by TDX_FEATURES0.WBINVD_DOMAINS (bit 15), readable by the host VMM using TDH.SYS.RD*.

Note that if TDH.PHYMEM.CACHE.WB is not required, as indicated by TDX_FEATURES0.SKIP_PHYMEM_CACHE_WB (bit 34), then TDX_FEATURES0.WBINVD_DOMAINS (bit 15) is 0.

5.4. Key Management Tables

The CPU and the Intel TDX module maintain several tables for key management. No table is intended to be directly accessible by software; the tables are used by the Intel TDX functions. The tables help the Intel TDX module track the proper operation of the software and help achieve the Intel TDX security objectives.

Table 5.2: Key Management Tables

Table	Scope	Description
Key Encryption Table (KET)	Package	KET is an abstraction of the CPU micro-architectural hardware table for configuring the memory encryption engines. The KET is indexed by HKID. All crypto engines on a package are configured the same way. KET is part of the legacy MKTME architecture. Intel TDX ISA partitions KET to shared and private ranges, as described in 5.2 above. • A KET entry in private HKIDs range is configured per package by the host VMM using the SEAMCALL(TDH.MNG.KEY.CONFIG) function. • A KET entry in the shared HKID range is configured by software per package directly, using the PCONFIG instruction.

Table	Scope	Description
KeyID Ownership Table (KOT)	Platform	KOT is an Intel TDX module hidden table for managing the TDX HKIDs inventory. It is used for assigning HKIDs to TDs, revoking HKIDs from TDs and controlling cache flush. KOT is indexed by HKID. Only the KOT entries in the configured TDX HKIDs range are meaningful.
TD Key Management Fields	TD	TD-scope key management fields are held in TDR. They include the key state, ephemeral private HKID and key information, and a bitmap for tracking key configuration.

Figure 5.1 below provides an abstract, high-level picture of how the tables are related. Detailed discussion is provided in the following sections.

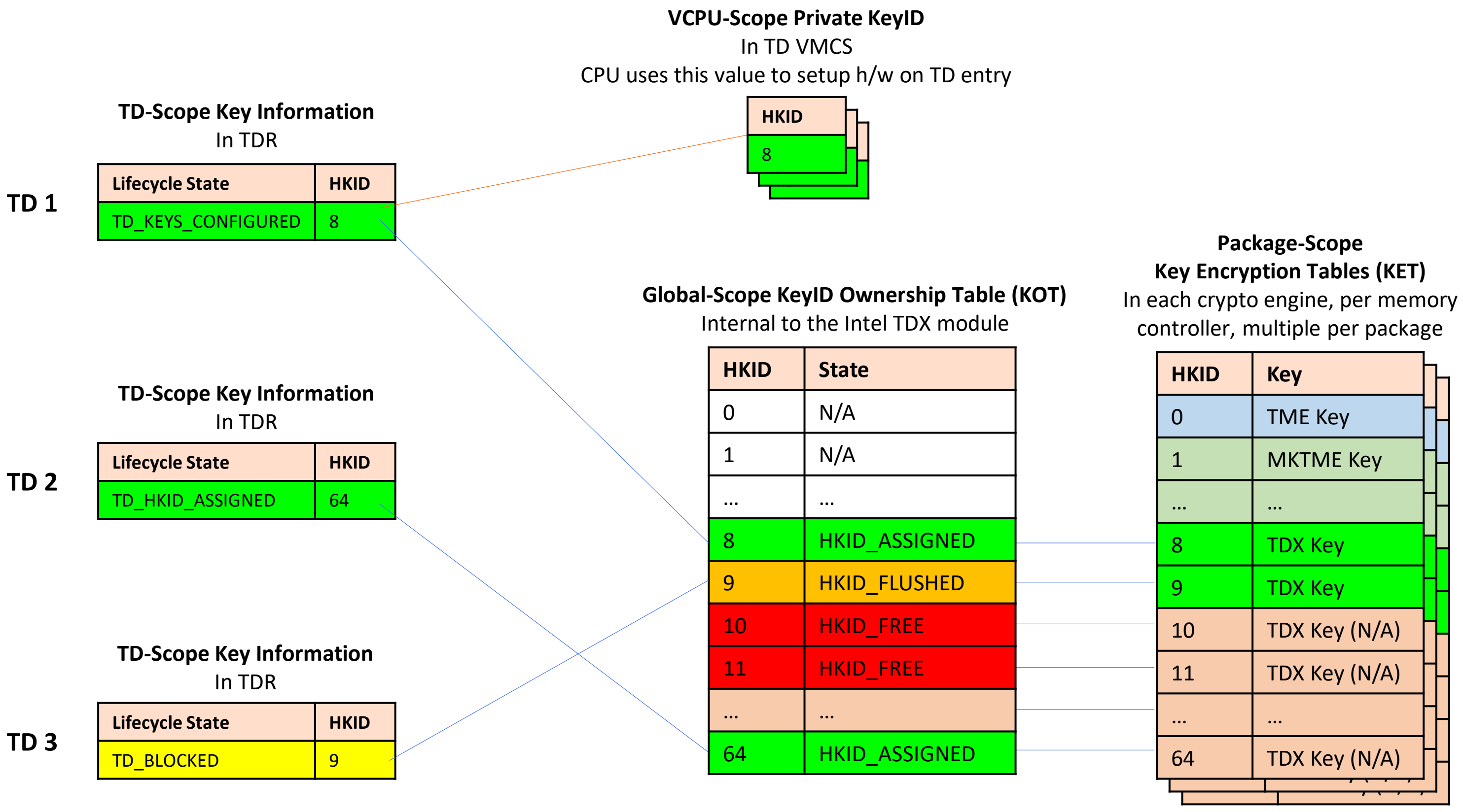


Figure 5.1: Overview of the Key Management State at TD-Scope, LP-Scope, Package-Scope and Global-Scope

5.5. Combined Key Management State

Key management state is composed of two state variables:

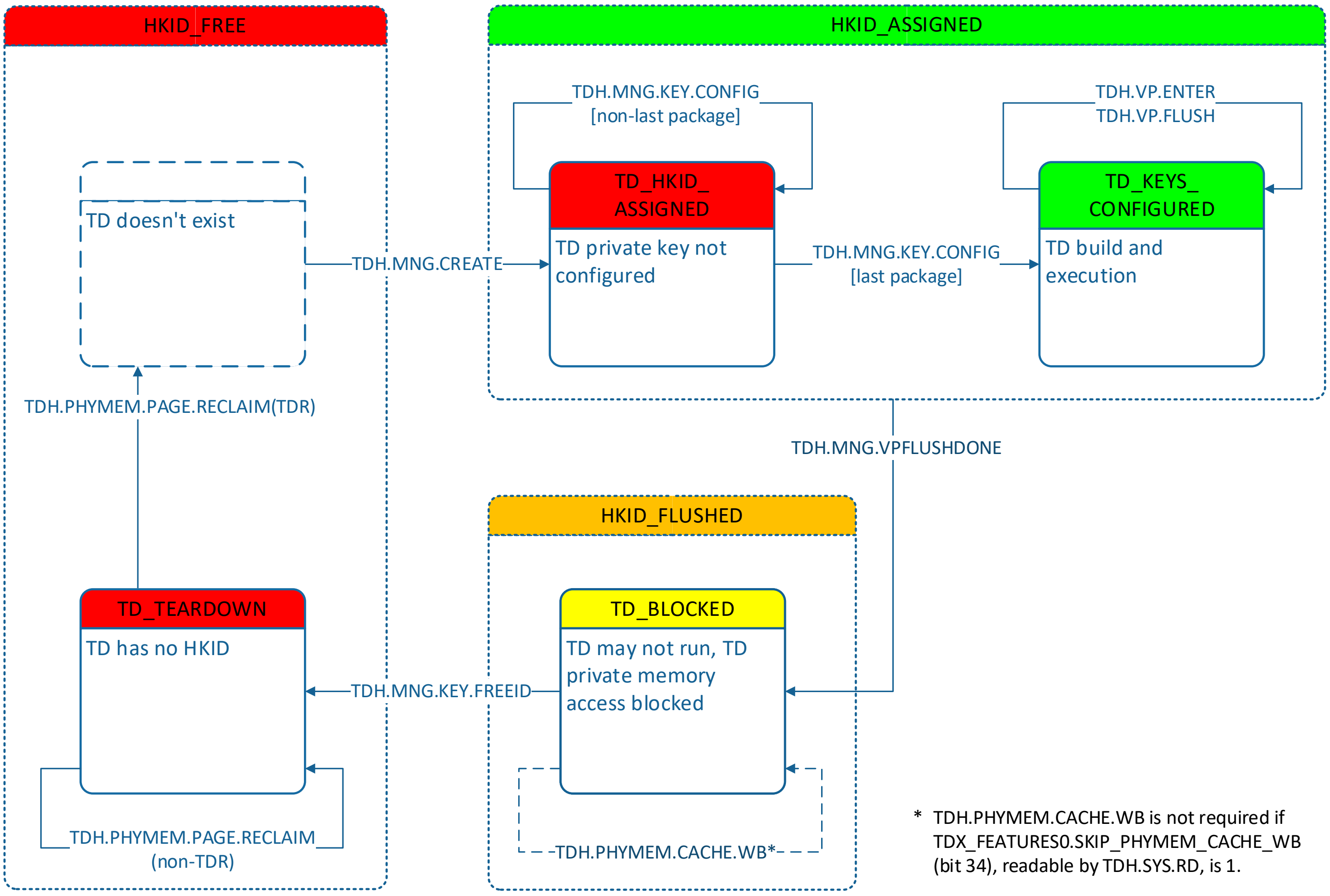
- **Per-HKID KOT Entry State** is designed to control how the inventory of private HKIDs is managed using the KOT.
- **Per-TD Life Cycle State** is designed, among other things, to control how TD keys are configured on the hardware and the process of shutting down a TD.

The combined key management state is intended to affect whether the TD private memory is accessible, whether its contents may be cached, whether private GPA-to-HPA address translations are allowed and whether such translations may be cached.

Table 5.3 below lists the designed combined key management state values and their meaning. Figure 5.2 below shows a simplified diagram of the combined key state. Refer also to the key management sequences described in 5.6.

Table 5.3: Combined TD Key Management States

TD Life Cycle State	KOT Entry (HKID) State	Private Memory Access		S-EPT Translations		Comments
		New	Cached	New	Cached	
N/A	HKID_FREE	No	No	No	No	HKID not assigned to TD
TD_HKID_ASSIGNED	HKID_ASSIGNED	No	No	No	No	TD private key not configured
TD_KEYS_CONFIGURED		TD	TD	TD	TD	TD build and execution
TD_BLOCKED	HKID_FLUSHED	No	TD	No	No	TD private memory access is blocked, TD may not run
TD_TEARDOWN	N/A (HKID_FREE)	No	No	No	No	TD has no HKID
N/A	HKID_RESERVED	Global	Global	N/A	N/A	HKID for Intel TDX global data


HKID_FREE
HKID_ASSIGNED
TD doesn't exist
TDH.MNG.CREATE
TDH.MNG.KEY.CONFIG [non-last package]
TD_HKID_ASSIGNED
TD private key not configured
TDH.MNG.KEY.CONFIG [last package]
TDH.VP.ENTER TDH.VP.FLUSH
TD_KEYS_CONFIGURED
TD build and execution
TDH.PHYMEM.PAGE.RECLAIM(TDR)
TDH.MNG.VPFLUSHDONE
HKID_FLUSHED
TD_TEARDOWN
TD has no HKID
TDH.MNG.KEY.FREEID
TD_BLOCKED
TD may not run, TD private memory access blocked
TDH.PHYMEM.PAGE.RECLAIM (non-TDR)
TDH.PHYMEM.CACHE.WB*
* TDH.PHYMEM.CACHE.WB is not required if TDX_FEATURES0.SKIP_PHYMEM_CACHE_WB (bit 34), readable by TDH.SYS.RD, is 1.

5 Figure 5.2: Simplified Combined TD Key Management State Diagram

Chapter 7 discusses TD life cycle management and zooms-in into the TD_KEYS_CONFIGURED state, detailing its secondary sub-states that control TD operation and TD migration.

5.6. Key Management Sequences

5.6.1. Intel TDX Module Initialization: Setting an Ephemeral Key and Reserving an HKID for Intel TDX Data

This sequence is described as part of the Intel TDX module initialization sequence in 3.1.

5.6.2. TD Creation, Keys Assignment and Configuration

This sequence is intended to be used by the host VMM to create a new TD, select HKIDs from the global pool in KOT and assign them to the TD, and configure the TD keys on the hardware.

Refer also to the software flow discussion in 3.2.

Table 5.4: Typical TD Creation, Keys Assignment and Configuration (TD-Scope and KOT-Scope) Sequence

	Intel TDX Function	Scope	Execute On	Description
1	TDH.MNG.CREATE	TD	One LP	Assign the TD's private HKID.
2	TDH.MNG.KEY.CONFIG	TD	Each package	Configure the TD's random ephemeral key on the package.

5.6.3. TD Keys Reclamation, TLB and Cache Flush

This sequence is intended to be used by the host VMM to reclaim the HKIDs assigned to a TD and return them to the global pool in KOT. At the end of this sequence, the HKIDs should be free to be assigned to another TD.

The cache flush operation is long. Since it is designed to run at global scope and is decoupled from any TD, the host VMM may choose to implement it in a lazy fashion, i.e., wait until a certain number of HKIDs in the KOT pool become RECLAIMED. This is especially important since TDH.PHYMEM.CACHE.WB operates on all cache lines regardless of HKID.

To avoid long latencies, TDH.PHYMEM.CACHE.WB is designed to be interruptible. The host VMM is expected to repeat the execution of this instruction until it returns a success indication.

Refer also to the software flow discussion in 3.4.

Table 5.5: Typical TLB and Cache Flush (TD-Scope and KOT-Scope) Sequence

	Intel TDX Function	Scope	Execute On	Description
As a preparation, the host VMM avoids any VCPU-specific SEAMCALL function (i.e., TDH.VP.ENTER, TDH.VP.INIT, TDH.VP.RD and TDH.VP.WR) and waits until no VCPU is running.				
1	TDH.VP.FLUSH	TD VCPU	One each LP associated with a TD VCPU	Flush the VCPU's TD VMCS to TDVPS memory and flush the VCPU's TLB ASID.
2	TDH.MNG.VPFLUSHDONE	TD, KOT	One LP	Check all the VCPUs have been flushed.
3	TDH.PHYMEM.CACHE.WB	KOT	Each WBINVD domain[4]	**Note:** TDH.PHYMEM.CACHE.WB is not required if TDX_FEATURES0.SKIP_PHYMEM_CACHE_WB (bit 34), readable by TDH.SYS.RD, is 1. Write back cache hierarchy, at least for the HKIDs marked as TLB_FLUSHED. The instruction execution time is long; it is interruptible by external events and may be restarted until completed.

[4] Enumerated by CPU during Intel TDX module initialization, see 4.1.2.4.

	Intel TDX Function	Scope	Execute On	Description
4	TDH.MNG.KEY.FREEID	TD, KOT	One LP	Mark TD’s HKID as FREE.

6. TD Non-Memory State (Metadata) and Control Structures

This chapter discusses the guest TD control structures that hold non-memory state (metadata) and how they are intended to be used during the TD life cycle.

6.1. Overview

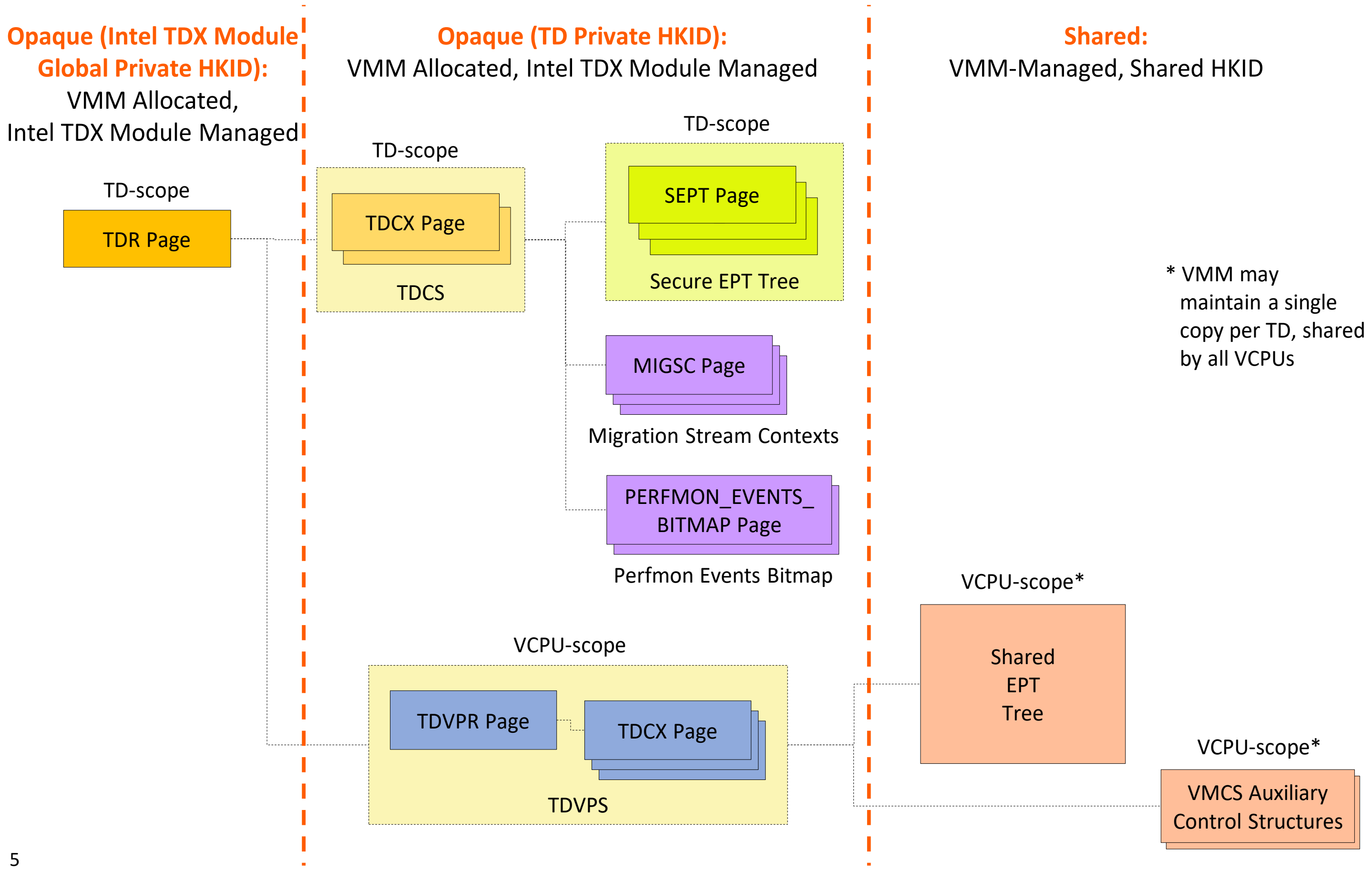

Opaque (Intel TDX Module Global Private HKID):
VMM Allocated, Intel TDX Module Managed
TD-scope
TDR Page
Opaque (TD Private HKID):
VMM Allocated, Intel TDX Module Managed
TD-scope
TDCX Page
TDCS
TD-scope
SEPT Page
Secure EPT Tree
MIGSC Page
Migration Stream Contexts
PERFMON_EVENTS_ BITMAP Page
Perfmon Events Bitmap
VCPU-scope
TDVPR Page
TDCX Page
TDVPS
Shared:
VMM-Managed, Shared HKID
* VMM may maintain a single copy per TD, shared by all VCPUs
VCPU-scope*
Shared EPT Tree
VCPU-scope*
VMCS Auxiliary Control Structures

Figure 6.1: Guest TD Control Structures Overview (Not Including TD Partitioning and TDX Connect)

All guest TD control structures reside in memory pages that are allocated by the host VMM from the pre-configured TDMRs. Guest TD control structure pages are addressable by the host VMM.

6.1.1. Opaque vs. Private vs. Shared Control Structures

Control structures are divided into two classes:

- **Shared control structures** are intended to be directly managed by the host VMM and are encrypted with a shared HKID. The Intel TDX module architecture only describes the shared control structures that might directly impact its operation. The host VMM may hold additional control structures.
- **Private control structures** are mapped to the guest TD's GPA space and are directly accessible by it.
- **Opaque control structures** are not intended to be directly accessible to any software (except the Intel TDX module) or DMA. They are intended to be managed via Intel TDX module functions. Generally speaking, the host VMM is not aware of the exact format of opaque control structures. Opaque control structures' memory pages are intended to be encrypted with a private HKID.

6.1.2. Scope of Control Structures

Guest TD control structures have two possible scopes:

- **TD-scope control structures** are intended to apply for a guest TD as a whole.
- **TD VCPU-scope control structures** are intended to apply for a single virtual CPU of a guest TD.

6.2. TD-Scope Control Structures

TD-scope control structures include TDR and TDCS, discussed below, and Secure EPT, discussed in Chapter 9.

6.2.1. TDR (Trust Domain Root)

TDR is the root control structure of a guest TD. As designed, TDR is encrypted using the Intel TDX global private HKID. It holds a minimal set of state variables that enable guest TD control even during times when the TD's private HKID is not known, or when the TD's key management state does not permit access to memory encrypted using the TD's private key.

TDR occupies a single 4KB naturally aligned page of memory. It is designed to be the first TD page to be allocated and the last to be removed. Its physical address serves as a unique identifier of the TD, as long as any TD page or control structure resides in memory.

At a high level, TDR holds the following information:

- Fields designed to control guest TD build and teardown process.
- Fields designed to manage memory encryption keys.

6.2.2. TDCS (Trust Domain Control Structure)

TDCS is the main control structure of a guest TD. As designed, TDCS is encrypted using the guest TD's ephemeral private key. TDCS is a multi-page logical structure composed of multiple TDCX physical pages.

At a high level, TDCS holds the following information:

- Fields designed to control the TD operation as a whole (e.g., a counter of the number of VCPUs currently running).
- Fields designed to control the TD's execution control (debuggability, CPU features available to the TD, etc.).
- Fields related to TD measurement.
- EPTP: as designed, a pointer (HPA) to the TD's secure EPT root page and EPT attributes.
- MSR bitmaps, designed to be used by all the TD's VCPUs.
- As designed, the secure EPT root page.
- A page filled with zeros, designed to be used in cases where the Intel TDX module needs a read-only constant-0 page encrypted with the TD's private key.

TDCS may hold forward links to the following control structures:

- Secure EPT pages.
- Migration Stream Context (MIGSC) pages.

6.3. TD VCPU-Scope Control Structures

6.3.1. Trust Domain Virtual Processor State (TDVPS)

Trust Domain Virtual Processor State (TDVPS) is the root control structure of a TD VCPU. It helps the Intel TDX module control the operation of the VCPU and holds the VCPU state while the VCPU is not running. TDVPS is a single logical control structure composed of multiple physical 4KB pages.

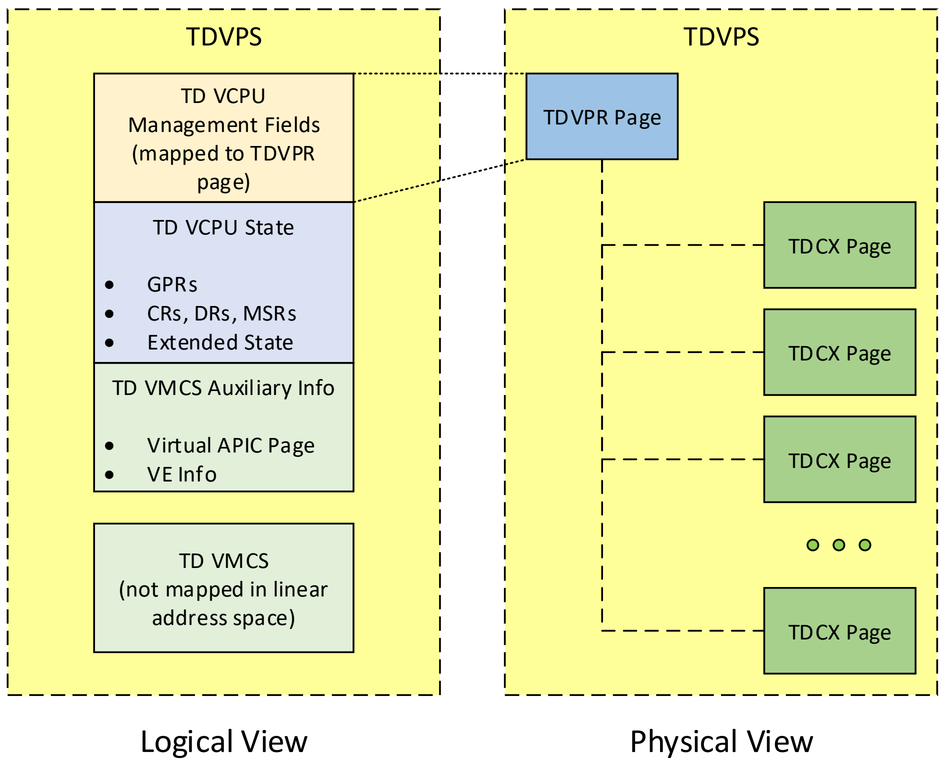


Figure 6.2: High Level Logical and Physical View of TDVPS

6.3.1.1. *Physical View of TDVPS: TDVPR/TDCX*

TDVPS is designed to be opaque to software and DMA access, accessible only by using the Intel TDX module functions. From the VMM perspective, TDVPS is composed of multiple 4KB pages, where each page may reside in arbitrary locations in convertible memory.

Trust Domain Virtual Processor Root (TDVPR) is the 4KB root page of TDVPS. Its physical address serves as a unique identifier of the VCPU (as long as it resides in memory).

Trust Domain Control structure eXtension (TDCX) 4KB pages extend TDVPR to help provide enough physical space for the logical TDVPS structure.

The TDVPR and TDCX pages are designed to be encrypted with the TD's ephemeral private key. They are addressable by the host VMM, which is responsible for allocating memory to hold them.

The required number of 4KB TDVPR/TDCX pages in TDVPS is enumerated to the VMM by the TDH.SYS.RD* or TDH.SYS.INFO function (see 4.2.3).

6.3.1.2. *Logical View of TDVPS*

Logically, TDVPS is organized as a single large data structure. At a high level, it is composed of the following parts:

VMX (with TDX ISA Extensions) Standard Control Structures

- TD VMCS
- TD VMCS auxiliary structures, such as virtual APIC page, virtualization exception information, etc. Note that MSR bitmaps are held as part of TDCS because they are meant to have the same value for all VCPUs of the same TD.

The TDX design does not require some of the VMX control structures (notably, the Shared EPT) to be protected. They are described below.

Proprietary Fields

- TD VCPU Management fields designed to manage the operation of the VCPU
- TD VCPU State fields designed to hold most of the VPCU state (except state that is saved to the TD VMCS) when the VCPU is not running

TDVPS organization and format are detailed in the [TDX Module ABI Spec].

6.3.2. Non-Protected Control Structures: Shared EPT and VMCS Auxiliary Control Structures

Several VMX control structures are directly managed and accessed by the host VMM. These control structures are pointed to by fields in the TD VMCS. The Intel TDX module checks that the pointers conform to the shared-access HPA semantics (see 18.2.1.1).

Non-protected control structures include:

- Shared EPT tree
- Posted interrupt descriptor

6.4. TD Non-Memory State (Metadata) Access Functions

As set of interface functions is provided to enable host VMM and guest TD access to TD non-memory state (metadata). These functions employ **metadata abstraction**, using field code to abstract the actual control structure format. The generic metadata access interface mechanisms are described in 18.6.

Table 6.1: TD Non-Memory State (Metadata) Single Field Access Functions

Side	Scope	Control Structures	Intel TDX Functions
Host VMM (SEAMCALL)	TD	TDR and TDCS	TDH.MNG.RD, TDH.MNG.WR
	VCPU	TDVPS (including TD VMCS)	TDH.VP.RD, TDH.VP.WR
Guest TD (TDCALL)	TD	TDR and TDCS	TDG.VM.RD, TDG.VM.WR
	VCPU	TDVPS (including TD VMCS)	TDG.VP.RD, TDG.VP.WR

Access to control structure fields using the provided interface functions (down to the bit granularity, if required) depends on whether the TD is debuggable (ATTRIBUTES.DEBUG bit is 1) or not.

In many cases, control structure field access means more than just reading or writing the field content. For example:

- When a field that contains an HPA is written, its value is checked not to overlap the SEAMRR range.
- In some cases, there may be inter-dependency between fields. When such fields are written, multiple checks may need to be done, and some actions may need to be taken.
- For some fields, the internal format and/or value may be different than what is visible externally.

For details about the TDX module's metadata access interface, see 18.6.

6.5. Concurrency Restrictions and Enforcement

A general description of concurrency restrictions is provided in 18.1.

Normally, exclusive or shared access is acquired, if needed, for the typically short duration of function flows. A TD VCPU execution is an exception case. Shared access to TDCS and TDVPS is acquired on TD Entry and released on TD Exit. This implies that SEAMCALL(TDH.VP.ENTER) function, all TDCALL functions, and asynchronous TD Exit have implicit shared access to TDCS and TDVPS.

This mechanism helps protect running VCPUs against concurrent functions that may try to change their governing control structures.

7. TD Life Cycle Management

This chapter discusses guest TD life cycle management.

7.1. TD Life Cycle State Machine

The TD Life Cycle state machine controls the overall TD build, run-time and destruction process. It operates in conjunction with the HKID state machine, as described in 5.5. Figure 7.1 below shows the TD Life Cycle state diagram.

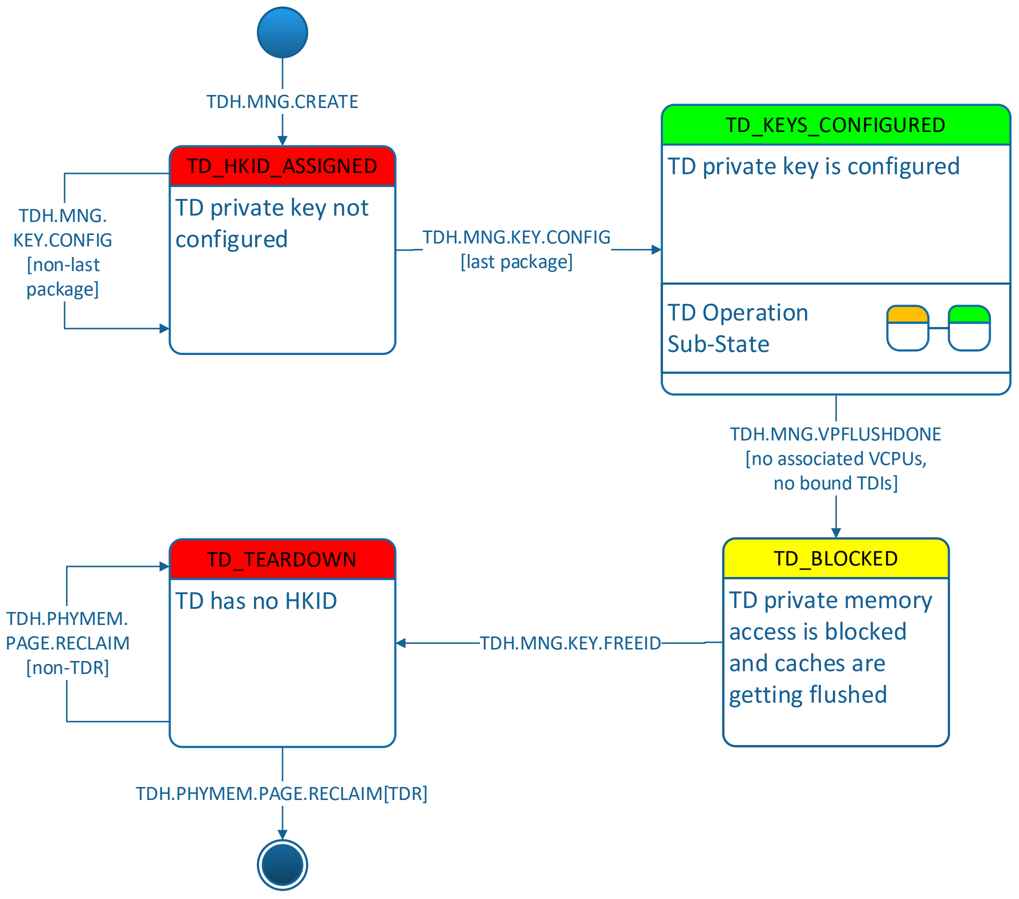

TDH.MNG.CREATE
TD_HKID_ASSIGNED
TD private key not configured
TDH.MNG. KEY.CONFIG [non-last package]
TDH.MNG.KEY.CONFIG [last package]
TD_KEYS_CONFIGURED
TD private key is configured
TD Operation Sub-State
TDH.MNG.VPFLUSHDONE [no associated VCPUs, no bound TDIs]
TD_BLOCKED
TD private memory access is blocked and caches are getting flushed
TDH.MNG.KEY.FREEID
TD_TEARDOWN
TD has no HKID
TDH.PHYMEM. PAGE.RECLAIM [non-TDR]
TDH.PHYMEM.PAGE.RECLAIM[TDR]

Figure 7.1: High-Level TD Life Cycle State Diagram

Most of the TD lifetime is spent in the TD_KEYS_CONFIGURED state. Within that state, a secondary-level state machine controls the overall TD operation and migration.

7.2. OP_STATE: TD Operation Secondary-Level State Machine

The TD Operation state machine controls sub-states of the TD Life Cycle's TD_KEYS_CONFIGURED state. It shown in Figure 7.2 below. This document describes the baseline states: UNALLOCATED, UNINITIALIZED, INITIALIZED and RUNNABLE. Other states and transitions highlighted in red lines support TD migration and are described in the [TD Migration Spec].

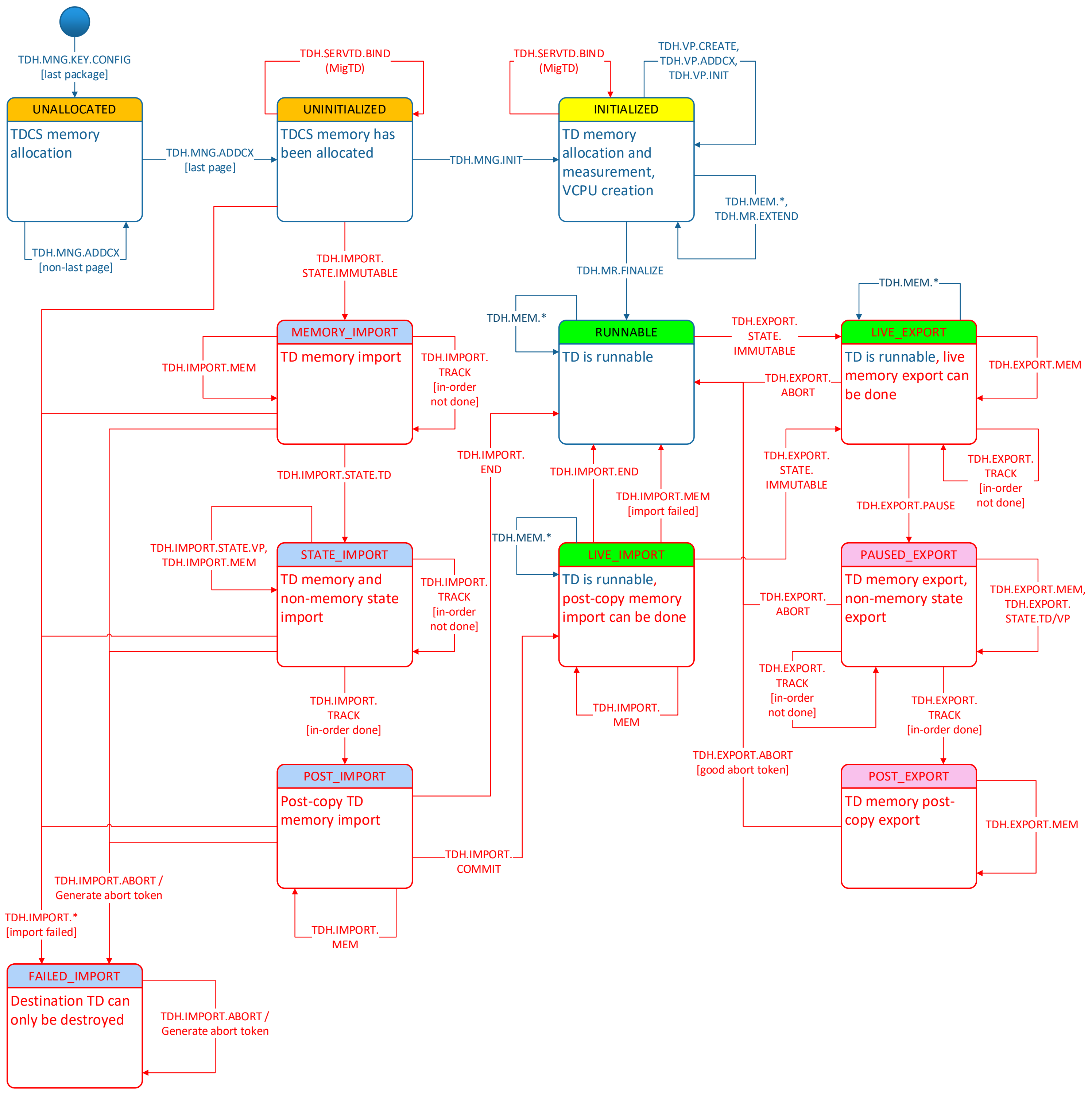

TDH.MNG.KEY.CONFIG [last package]
UNALLOCATED
TDCS memory allocation
TDH.MNG.ADDCX [non-last page]
TDH.MNG.ADDCX [last page]
TDH.SERVTD.BIND (MigTD)
UNINITIALIZED
TDCS memory has been allocated
TDH.MNG.INIT
TDH.SERVTD.BIND (MigTD)
TDH.VP.CREATE, TDH.VP.ADDCX, TDH.VP.INIT
INITIALIZED
TD memory allocation and measurement, VCPU creation
TDH.MEM.*, TDH.MR.EXTEND
TDH.IMPORT. STATE.IMMUTABLE
TDH.MR.FINALIZE
TDH.MEM.*
RUNNABLE
TD is runnable
TDH.EXPORT. STATE. IMMUTABLE
TDH.MEM.*
LIVE_EXPORT
TD is runnable, live memory export can be done
TDH.EXPORT.MEM
TDH.EXPORT. ABORT
MEMORY_IMPORT
TD memory import
TDH.IMPORT.MEM
TDH.IMPORT. TRACK [in-order not done]
TDH.IMPORT. END
TDH.IMPORT.END
TDH.IMPORT.MEM [import failed]
TDH.EXPORT. STATE. IMMUTABLE
TDH.EXPORT. TRACK [in-order not done]
TDH.EXPORT.PAUSE
TDH.IMPORT.STATE.TD
TDH.IMPORT.STATE.VP, TDH.IMPORT.MEM
STATE_IMPORT
TD memory and non-memory state import
TDH.IMPORT. TRACK [in-order not done]
TDH.MEM.*
LIVE_IMPORT
TD is runnable, post-copy memory import can be done
TDH.EXPORT. ABORT
PAUSED_EXPORT
TD memory export, non-memory state export
TDH.EXPORT.MEM, TDH.EXPORT. STATE.TD/VP
TDH.IMPORT. TRACK [in-order done]
TDH.IMPORT. MEM
TDH.EXPORT. TRACK [in-order not done]
TDH.EXPORT. TRACK [in-order done]
TDH.EXPORT.ABORT [good abort token]
POST_IMPORT
Post-copy TD memory import
TDH.IMPORT. COMMIT
POST_EXPORT
TD memory post-copy export
TDH.EXPORT.MEM
TDH.IMPORT.ABORT / Generate abort token
TDH.IMPORT.* [import failed]
TDH.IMPORT. MEM
FAILED_IMPORT
Destination TD can only be destroyed
TDH.IMPORT.ABORT / Generate abort token

Figure 7.2: TD Operation State Machine (Sub-States of TD_KEYS_CONFIGURED)

7.3. TD Creation and Configuration Sequence

The following sequence is intended to be used by the host VMM to create a new TD. Note that only the general aspects of TD creation are described here. Other aspects, such as key management, are described in other chapters.

TD configuration is done by TDH.MNG.INIT. This interface function receives a TD_PARAMS input structure, which contains the following main sections:

- ATTRIBUTES and XFAM, which specify the set of TD attributes (e.g., whether the TD is debuggable) and CPU features the TD may use (e.g., whether AVX2 is available to the TD).
- Other TD configuration parameters, such as the number of L2 VMs the TD contains.

- A set of user-provided measurement fields that will appear in the TD's TDREPORT_STRUCT.
- Virtual CPUID and virtual MSR configuration for the TD.

For a detailed description of TD_PARAMS and its fields, see the [ABI Spec].

Refer also to the software flow discussion in 3.2.

Table 7.1: Typical TD Creation Sequence

	Intel TDX Function	Inputs	Description
1	N/A	N/A	If any MODIFIED cache lines may exist for the physical pages to be written below (TDR, TDCS, Secure EPT root page), flush them to memory using, e.g., CLFLUSH (possibly on multiple LPs). This is required to avoid corruption due to cache line aliasing.
2	TDH.MNG.CREATE	TDR page PA	Create the TDR and generate the TD's random ephemeral key.
3	Multiple	See 5.6.2	Assign an HKID and configure the TD's random ephemeral key on all packages, as described in 5.6.2.
4	TDH.MNG.ADDCX (multiple)	• Owner TDR PA • TDCX page PA	Run multiple times to add the required number of TDCX pages.
5	TDH.MNG.INIT	• Owner TDR PA • TD initialization parameters	Initialize the TD state in TDR and TDCS.
At this point the TD is initialized. Private memory pages can be added as described in Chapter 9. VCPUs can be created and initialized as described below.			

7.4. VCPU Creation and Initialization Sequence

VCPU creation and initialization is only allowed during TD build time.

The following sequence is intended to be used by the host VMM to create a new TD VCPU. After this sequence is done, the TD VCPU may be entered on an LP (assuming other conditions are met).

Refer also to the software flow discussion in 3.2.

Table 7.2: Typical TD VCPU Creation and Initialization Sequence

	Intel TDX Function	Inputs	Description
1	N/A	N/A	If any MODIFIED cache lines may exist for the physical pages to be written below (TDVPR, TDCX), flush them to memory (e.g., using CLFLUSH – possibly on multiple LPs). This is required to avoid corruption from cache line aliasing.
2	TDH.VP.CREATE	• TDVPR page PA • Owner TDR PA	Create the VCPU and its TDVPR page.
3	TDH.VP.ADDCX (multiple)	• TDCX page PA • Parent TDVPR PA	Run multiple times to add the required number of TDCX pages as an extension to a parent TDVPR.
4	TDH.VP.INIT	• TDVPR PA • VMM-provided identifier	Initialize the VCPU state.

	Intel TDX Function	Inputs	Description
5	TDH.VP.WR	• TDVPR page PA • Field code • New field value • Write mask	The host VMM typically writes one or more of the following TD VCPU's VMCS controls: • Shared EPTP • Posted-interrupts descriptor address, posted-interrupts notification vector and process posted interrupt • Bus-lock detection • Notification exiting and notify window For details, see the [TDX Module ABI Spec].

7.5. TD Teardown Sequence

The following sequence is intended to be used by the host VMM to tear down a TD. Note that only the general aspects of TD teardown are described here. Other aspects, such as key management, are described in other chapters. See also the discussion of physical page reclamation in 8.6.

Refer also to the software flow discussion in 3.4.

Table 7.3: Typical TD Teardown Sequence

	Intel TDX Function	Inputs	Description
1	Multiple	See 5.6.3	Reclaim the HKID, and flush TLB and cache, as described in 5.6.3.
2	TDH.PHYMEM.PAGE.RECLAIM (multiple)	TD page or control structure PA	Remove all TD private pages and control structure pages and mark them as PT_NDA in the PAMT.
3	TDH.PHYMEM.PAGE.RECLAIM	TDR PA	Remove the TDR page and mark it as PT_NDA in the PAMT.
4	TDH.PHYMEM.PAGE.WBINVD	TDR PA	Flush MODIFIED cache lines: this is required to avoid corruption due to cache line aliasing. Note that all cache lines for all other TD pages must have been flushed before the TDR page was reclaimed.

8. Physical Memory Management

This chapter describes how the Intel TDX module manages memory as a set of physical pages.

8.1. Trust Domain Memory Regions (TDMRs) and Physical Address Metadata Tables (PAMTs)

Trust Domain Memory Region (TDMR) is defined as a range of convertible memory pages. TDMRs are set by the host VMM, based on the CMR information previously checked by MCHECK.

By default, the TDX module uses static PAMT. With static PAMT, each TDMR is defined as controlled by a (logically) single **Physical Address Metadata Table (PAMT)**, which is composed of 3 PAMT arrays corresponding to the 3 page-mapping sizes: PAMT_1G, PAMT_2M and PAMT_4K. The PAMT structure is discussed in 8.3 below. PAMT tables reside in VMM-allocated memory, and they are designed to be encrypted with the Intel TDX global private HKID. The required size of PAMT memory, as a function of TDMR size, is enumerated to the VMM by TDH.SYS.RD/RDALL or TDH.SYS.INFO.

If supported by the TDX module, it can be configured to use dynamic PAMT. With dynamic PAMT, 4KB PAMT pages which hold PAMT_4K entries are allocated dynamically by the host VMM. A separate bitmap (PAMT_PAGE_BITMAP) is used for controlling the allocation of PAMT pages. That bitmap is statically allocated for each TDMR.

Typically, after the host VMM initializes the Intel TDX module (TDH.SYS.INIT and TDH.SYS.LP.INIT), it configures the TDMRs, the PAMT mode (static or dynamic) and their respective PAMTs using TDH.SYS.CONFIG. It then would gradually initialize the TDMRs using TDH.SYS.TDMR.INIT. For a detailed description of the typical Intel TDX module initialization and configuration sequence, see Chapter 4.

8.2. TDMR Details

The following list includes definitions of the characteristics of a TDMR:

- TDMR configuration is "soft" – no hardware range registers are used.
- Each TDMR defines a single physical address range.
- Each TDMR has its own size which must be a multiple of 1GB. TDMR size is **not** required to be a power of two.
- A TDMR must be aligned on 1GB.
- TDMRs cannot overlap with each other.
- TDMRs may contain reserved areas. This effectively allows the host VMM to flexibly configure TDMRs based on the VMM's own consideration of system memory allocation – without being impacted by the 1GB granularity of the TDMR size.
  - A reserved area must be aligned on 4KB, and its size must be a multiple of 4KB.
  - The number of reserved areas that may be configured per TDMR is enumerated by TDH.SYS.RD/RDALL or TDH.SYS.INFO.
- TDMR memory, except for reserved areas, must be convertible as checked by MCHECK (i.e., every TDMR page must reside within a CMR).
- There is no requirement for TMDRs to cover all CMRs.
- TDMRs are configured at platform scope (no separate configuration per package).
- The maximum number of TDMRs is Intel TDX module implementation specific. It is enumerated to the host VMM using the TDH.SYS.RD/RDALL or TDH.SYS.INFO function, as described below.

8.3. PAMT Details

The Physical Address Metadata Table (PAMT) is designed to track the metadata of every physical page in TDMR. A page metadata includes page type, page size, assignment to a TD, and other attributes.

The PAMT is used by the Intel TDX module to help enforce the following properties:

Page Attributes	A physical page in TDMR has a well-defined set of attributes, such as page type and page size.
Single TD Assignment	A physical page in TDMR can be assigned to at most one TD.
Secure EPT Consistency	The page size of any private TD page, mapped in Secure EPT, matches its page size attribute in PAMT.

8.3.1. Leaf PAMT Entry

Note: The description below is provided at a high level. Implementation details may differ.

A leaf PAMT entry is designed to hold metadata for a single physical page. The page size may be 4KB, 2MB or 1GB depending on the PAMT level (see 8.3.3.2 below).

Table 8.1: High-Level View of a Leaf PAMT Entry

Field	Description
PT	PT indicates the type of page intended to be associated with this PAMT entry. See Table 8.3 below for details.
OWNER	OWNER is designed to contain bits 51:12 of the physical address of the TD's TDR page. This field can be applicable in all cases when a page is assigned to the Intel TDX module at this PAMT level or at a higher level. See Table 8.3 below for details.
BEPOCH	The value of TDCS.TD_EPOCH at the time the TD private page or the SEPT page was blocked by TDH.MEM.RANGE.BLOCK This field is intended to be applicable only if PT is PT_REG, PT_EPT or PT_PR. See 9.7 for a detailed discussion.
MIG_EPOCH	Migration epoch at the time this page was imported This field is used only during TD migration, and is intended to be applicable only if PT is PT_REG. For details, see the TD Migration Spec and the [ABI Spec].
EXPORT_COUNT	TDCS.MIG_COUNT at the time this page was imported This field is used only during TD migration, and is intended to be applicable only if PT is PT_REG. For details, see the TD Migration Spec and the [ABI Spec].

5 ### 8.3.2. Page Type (PT)

Table 8.2 below describes the PAMT entry's Page Type (PT) field. For details, see the [ABI Spec]. Additional PT values exist for TDX Connect; see the [TDX Connect Spec] for details.

Table 8.2: PAMT Entry's Page Type (PT)

Page Type	PAMT Level	Associated TDX Page	Description
PT_NDA	Any	Depending on PT at higher PAMT level (if any)	The physical page is **Not Directly Assigned** to the Intel TDX module at this size (4K, 2M or 1G) and PAMT level. This page may be part of a larger page that is assigned to the Intel TDX module at a higher level, or this page may contain smaller pages that are assigned to the Intel TDX module at lower levels. See Table 8.3 below for details.
PT_RSVD	PAMT_4K	None	The 4KB physical page is reserved for non-TDX usage. The Intel TDX module will not allow converting this page to any other page type. The page can be used by the host VMM for any purpose. PT_RSVD is used for implementing reserved areas within TDMRs. See 4.1.3.2.1 for details.
PT_REG	Any	TD private page	The physical page at this PAMT level (4K, 2M or 1G) holds TD private memory and is mapped in the guest TD GPA space by the Secure EPT.
PT_PR	PAMT_2M, PAMT_1G	TD private page	The physical page at this PAMT level (2M or 1G) holds a removed TD private memory that is waiting for the completion of page removal (by TDH.MEM.PAGE.REMOVE) or reclamation (by TDH.PHYMEM.PAGE.RECLAIM). **Enumeration:** PT_PR applies only to platforms which protect TDX memory using ACT, as enumerated by TDX_FEATURES0.ACT (bit 14).

Page Type	PAMT Level	Associated TDX Page	Description
PT_TDR	PAMT_4K	TDR	TDR control structure page
PT_TDCX	PAMT_4K	TDCX	One 4KB physical page of a multi-page control structure
PT_TDVPR	PAMT_4K	TDVPR	Root page of the multi-page TDVPS control structure
PT_EPT	PAMT_4K	Secure EPT	Secure EPT page
PT_TR	PAMT_4K	-	A physical page which has no current GPA mapping but must be TLB tracked before it can be assigned for any usage **Enumeration:** PT_TR applies only for TDX modules which implement page resize without tracking, as enumerated by TDX_FEATURES0. NON_BLOCKING_RESIZE (bit 35).
PT_PAMT	PAMT_2M (Dynamic PAMT Only)	PAMT	If dynamic PAMT is supported, PT_PAMT is the page type for a non-leaf PAMT entry pointing to a pair of PAMT pages.
TDX Connect Types	-	-	TDX Connect page types are defined in the [TDX Connect Spec]

8.3.3. Static PAMT

8.3.3.1. *Overview*

Static PAMT is the default mode of operation, designed to optimize PAMT entries lookup given a page HPA. With static PAMT, each TDMR is statically associated with three PAMT arrays, all allocated by the host VMM using TDH.SYS.CONFIG. The PAMT arrays' entries are associated with 4BK, 2MB and 1GB physical pages.

8.3.3.2. *Static PAMT Blocks and Static PAMT Arrays*

For each 1GB of TDMR physical memory, there is a corresponding **PAMT Block**. A PAMT Block is **logically** arranged in a three-level tree structure of **PAMT Entries**, as shown in Figure 8.1 below. Levels 0 through 2 (PAMT_4K, PAMT_2M and PAMT_1G) correspond to 4KB, 2MB and 1GB physical TDMR pages, respectively.

Physically, for each TDMR the design includes three arrays of PAMT entries, one for each PAMT level. This aims to simplify VMM memory allocation. A logical PAMT Block has one entry from the PAMT_1G array, 512 entries from the PAMT_2M array, and 512^2 entries from the PAMT_4K array.

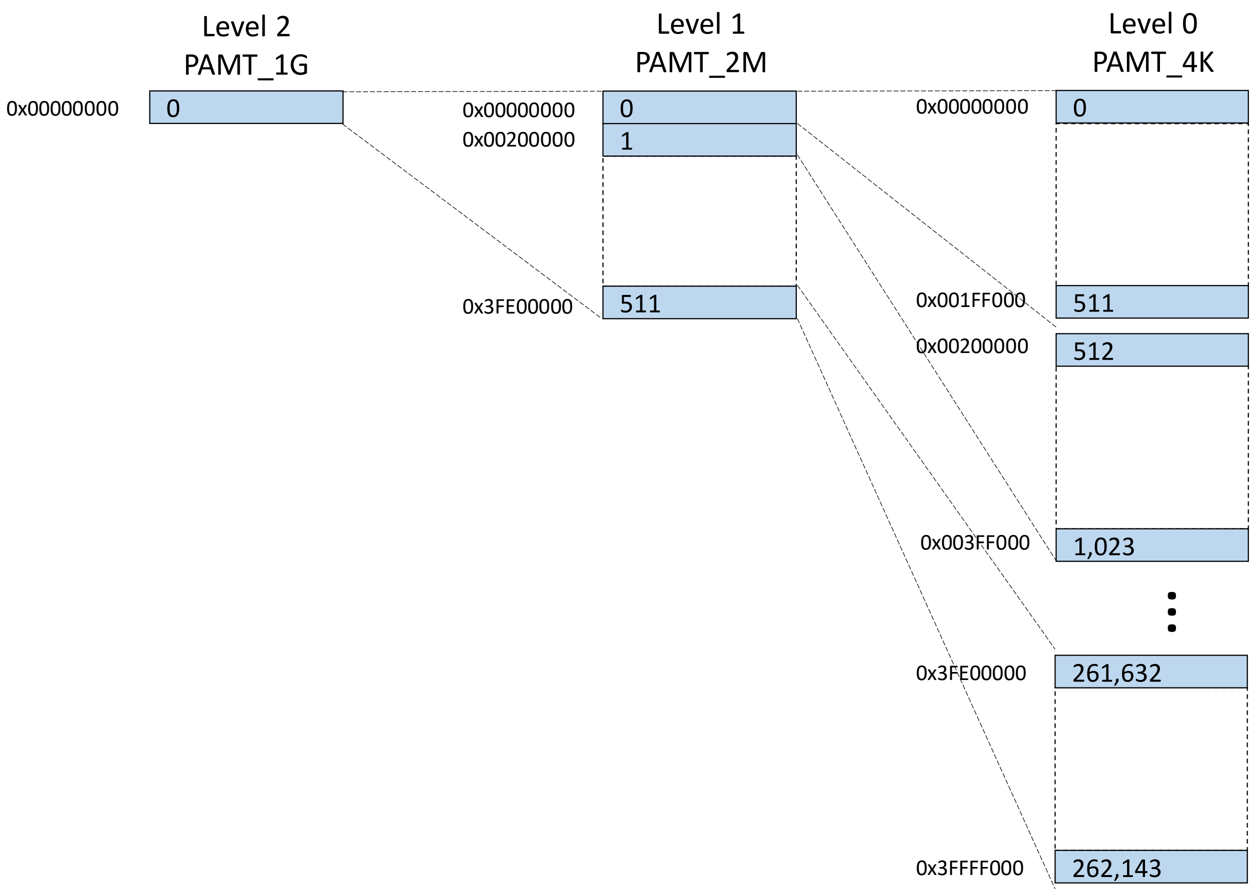


Figure 8.1: Typical Example of a PAMT Block Hierarchy for a 1GB TDMR Block

8.3.3.3. *Static PAMT Hierarchy*

Table 8.3 below shows the page type (PT) of PAMT entries at the three levels of hierarchy, depending on whether the page is assigned to the Intel TDX module manages the page, whether the page is mapped in secure EPT, and the mapping size.

Table 8.3: Static PAMT Hierarchy and Page Types

Logical Page Type	Intel TDX Module Management			PAMT Entry Page Type		
	Assigned to TDX?	**Physical Page Size**	**GPA Mapping Size (Secure EPT Level)**	**PAMT_1G (Level 2)**	**PAMT_2M (Level 1)**	**PAMT_4K (Level 0)**
Reserved	**No**	**4KB**	**N/A**	PT_NDA	PT_NDA	**PT_RSVD**
Free	**No**	**4KB**	**N/A**	PT_NDA	PT_NDA	PT_NDA
	No	**2MB**	**N/A**	PT_NDA	PT_NDA	All PT_NDA
	No	**1GB**	**N/A**	PT_NDA	All PT_NDA	All PT_NDA
Control Structure Page	**Yes**	**4KB**	**None**	PT_NDA	PT_NDA	**PT_TDR, PT_TDCX, PT_TDVPR, PT_EPT**
TD Private Page	**Yes**	**4KB**	**4KB (Level 0)**	PT_NDA	PT_NDA	**PT_REG**
	Yes	**2MB**	**2MB (Level 1)**	PT_NDA	**PT_REG**	All PT_NDA
	Yes	**1GB**	**1GB (Level 2)**	**PT_REG**	All PT_NDA	All PT_NDA
Pending Removal	**Yes**	**2MB**	**2MB (Level 1)**	PT_NDA	**PT_PR**	All PT_NDA
	Yes	**1GB**	**1GB (Level 2)**	**PT_PR**	All PT_NDA	All PT_NDA

Note the following:

- A 4KB page is considered **free** (i.e., not assigned to TDX) if its PAMT.PT at all three PAMT levels is PT_NDA. Any function that attempts to assign an HPA to TDX (e.g., TDH.MEM.PAGE.ADD) is designed to check this.
- In all other cases, PAMT.PT is different than PT_NDA in only one of the three PAMT levels.
- When a page is mapped by Secure EPT at 4KB, 2MB or 1GB GPA mapping size, it is managed by the Intel TDX module as a physical page of the same size. Secure EPT is described in Chapter 9.
- PT_RSVD pages cannot be used by the Intel TDX module. They are used for implementing reserved areas within TDMRs. See 4.1.3.2.1 for details.

8.3.4. Dynamic PAMT

8.3.4.1. *Overview*

The TDX module can be configured to use dynamic PAMT, if it supports that mode. Dynamic PAMT optimizes memory space by requiring PAMT pages holding PAMT_4K entries to be allocated only for 4KB physical memory pages that are assigned to TDX.

Dynamic PAMT is logically organized as a tree structure, similar in concept to page tables, to hold information about physical pages that are allocated to TDX. The tree nodes are pairs of 4KB PAMT pages. Each PAMT page pair holds 512 PAMT entries. Non-leaf PAMT entries point to lower level PAMT page pairs, while leaf PAMT entries are associated with physical pages that are assigned to TDX.

Physically, the PAMT_1G and PAMT_2M levels are statically allocated for each TDMR, same as with static PAMT. Only the PAMT_4K level is dynamically allocated.

PAMT pages, which are 4KB in size, can't be mapped by the dynamic PAMT tree; they are required to build the dynamic part of the tree. Thus, a per-TDMR statically allocated bitmap is used to control the allocation of PAMT pages.

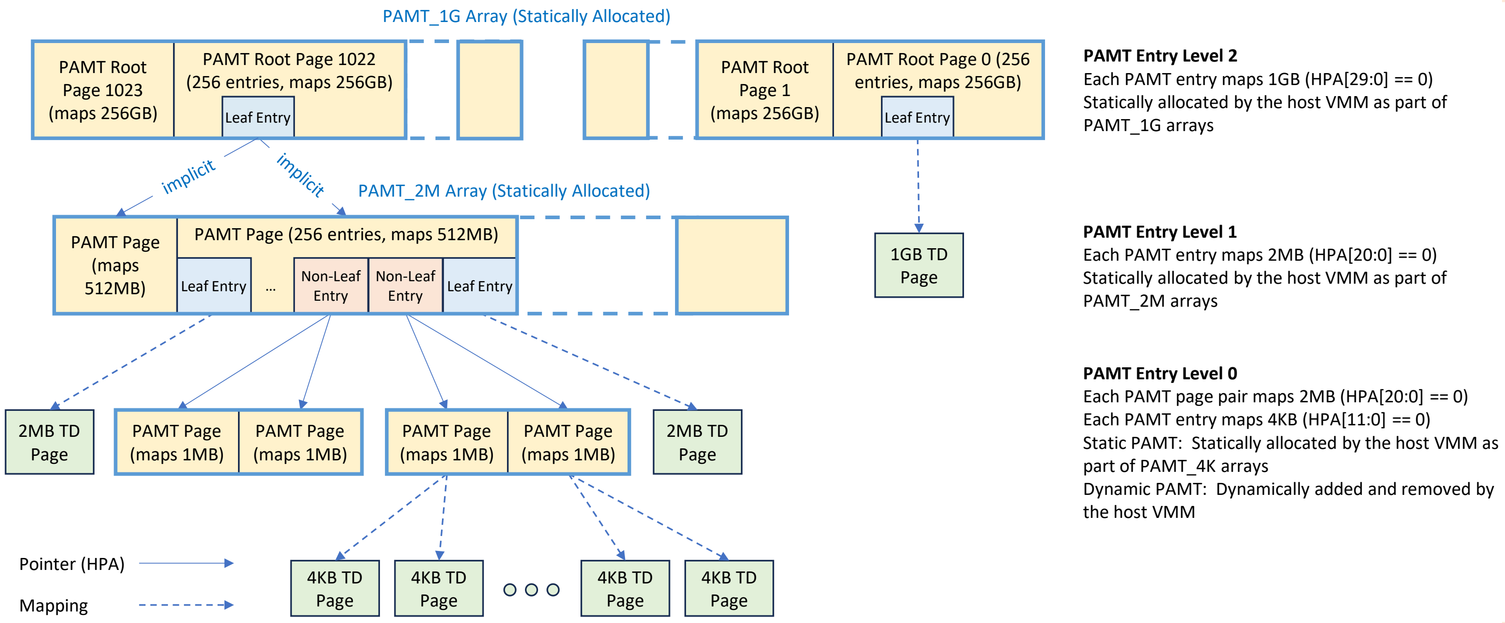

PAMT_1G Array (Statically Allocated)
PAMT Root Page 1023 (maps 256GB)
PAMT Root Page 1022 (256 entries, maps 256GB)
Leaf Entry
PAMT Root Page 1 (maps 256GB)
PAMT Root Page 0 (256 entries, maps 256GB)
Leaf Entry
implicit
implicit
PAMT_2M Array (Statically Allocated)
PAMT Page (maps 512MB)
PAMT Page (256 entries, maps 512MB)
Leaf Entry
…
Non-Leaf Entry
Non-Leaf Entry
Leaf Entry
1GB TD Page
2MB TD Page
PAMT Page (maps 1MB)
PAMT Page (maps 1MB)
PAMT Page (maps 1MB)
PAMT Page (maps 1MB)
2MB TD Page
4KB TD Page
4KB TD Page
4KB TD Page
4KB TD Page
Pointer (HPA)
Mapping
PAMT Entry Level 2
Each PAMT entry maps 1GB (HPA[29:0] == 0)
Statically allocated by the host VMM as part of PAMT_1G arrays
PAMT Entry Level 1
Each PAMT entry maps 2MB (HPA[20:0] == 0)
Statically allocated by the host VMM as part of PAMT_2M arrays
PAMT Entry Level 0
Each PAMT page pair maps 2MB (HPA[20:0] == 0)
Each PAMT entry maps 4KB (HPA[11:0] == 0)
Static PAMT: Statically allocated by the host VMM as part of PAMT_4K arrays
Dynamic PAMT: Dynamically added and removed by the host VMM

Figure 8.2: Dynamic PAMT Tree Example

The table below lists the PAMT tree levels. For a TDX module to be configured for dynamic PAMT, the effective physical address width (excluding HKID bits) must be 48 bits or lower. For platforms with a physical address width of 52 bits, this means that at least 4 bits must be configured as HKID bits. Thus, 4 PAMT levels are required.

Table 8.4: Dynamic PAMT Levels

PAMT Entry Level	PAMT Entry Maps	PAMT Page Pair Maps	PAMT Entry Indexed by HPA Bits
0	4KB	2MB	20:12

PAMT Entry Level	PAMT Entry Maps	PAMT Page Pair Maps	PAMT Entry Indexed by HPA Bits
1	2MB	1GB	29:21
2	1GB	512GB	38:30

8.3.4.2. *PAMT Page Pair*

PAMT pages are always managed as a pair of 4KB physical pages, which are not required to be consecutive in physical address space. Together, a PAMT page pair holds 512 16-byte PAMT entries. Each PAMT entry can be a leaf entry associated with a physical page, or a non-leaf entry, pointing to a lower-level pair of PAMT pages. PAMT pages are encrypted in memory using the TDX module's global private HKID.

PAMT entries within a PAMT page pair are indexed using 9 bits of the HPA, as shown in the table above. The first 256 entries reside in the first PAMT page of the pair; the last 256 entries reside in the second PAMT page of the pair.

Same as with static PAMT, PAMT page pairs at the PAMT_1G and PAMT_2M levels are statically allocated for each TDMR, as part of the PAMT_1G and PAMT_2M arrays of that TDMR.

PAMT page pairs at the PAMT_4K level are dynamically allocated by the host VMM, as needed to map physical pages assigned to TDX, using TDH.PHYMEM.PAMT.ADD. The TDH.MEM.PAGE.DEMOTE interface function also requires the host VMM to allocate a PAMT page pair, as part of its input parameters, when demoting a 2MB page into 512 4KB pages.

PAMT page pairs at the PAMT_4K level (pointed by non-root PAMT_2M entries) can be dynamically removed by the host VMM, if no longer needed, using TDH.PHYMEM.PAMT.REMOVE. The TDH.MEM.PAGE.PROMOTE interface function also removes a PAMT page pair as part of its operation, when promoting 512 4KB pages to a 2MB page.

8.3.4.3. *PAMT Page Bitmap*

Each PAMT page is **pointed** by a **non-leaf** PAMT entry, with **PT = PT_PAMT**. However, since a PAMT page is just 4KB page that may reside at an arbitrary HPA within a TDMR, it also needs to be **mapped** in order to control its allocation and removal. But contrary to other page types, a 4KB PAMT page can't be mapped by normal PAMT_4K entries, since such entries also reside in 4KB PAMT pages that are dynamically allocated and thus may not even exist.

Thus, PAMT pages are mapped by statically allocated per-TDMR PAMT_PAGE_BITMAPs. Bit N in the bitmap corresponds to 4KB page N in the TDMR. A bit value of 0 indicates that the applicable page is not a PAMT page. A value of 1 indicates that the page is either a PAMT page or it is reserved. The combinations of PAMT_PAGE_BITMAP bit value and possible PAMT entry are shown in Table 8.6 below.

8.3.4.4. *Non-Leaf PAMT Entry*

Note: The description below is provided at a high level. Implementation details may differ.

A non-leaf PAMT entry holds the metadata for a pair of lower-level PAMT pages.

Table 8.5: High-Level View of a Non-Leaf PAMT Entry

Field	Description
PT	PT indicates the type of page intended to be associated with this PAMT entry. For a non-leaf entry, PT is PT_PAMT
PAGE_HPA0	HPA of the first PAMT page pointed by this entry
PAGE_HPA1	HPA of the second PAMT page pointed by this entry

8.3.4.5 *Dynamic PAMT Hierarchy*

The table below shows the page type (PT) of PAMT entries and the PAMT_PAGE_BITMAP bit value at the three levels of hierarchy, depending on whether the Intel TDX module manages the page, whether the page is mapped in secure EPT, and the mapping size.

Table 8.6: Dynamic PAMT Hierarchy and Page Types

Logical Page Type	**Intel TDX Module Management**			**PAMT Entry Page Type**			**PAMT_ PAGE_ BITMAP Bit(s)**
	Assigned to TDX?	**Physical Page Size**	**GPA Mapping Size (Secure EPT Level)**	**PAMT_1G (Level 2, Static)**	**PAMT_2M (Level 1, Static)**	**PAMT_4K (Level 0, Dynamic)**	
Reserved	**No**	**4KB**	**N/A**	PT_NDA	PT_NDA	PT_NDA	**1**
Free	**No**	**4KB**	**N/A**	PT_NDA	PT_PAMT	**PT_NDA**	0
	No	**2MB**	**N/A**	PT_NDA	PT_NDA	N/A	All 0
	No	**1GB**	**N/A**	PT_NDA	All PT_NDA	N/A	All 0
PAMT Page	**Yes**	**4KB**	**N/A**	PT_NDA	PT_PAMT	**PT_NDA**	**1**
				PT_NDA	**PT_NDA**	N/A	**1**
Control Structure Page	**Yes**	**4KB**	**None**	PT_NDA	PT_PAMT	**PT_TDR, PT_TDCX, PT_TDVPR, PT_EPT**	0
TD Private Page	**Yes**	**4KB**	**4KB (Level 0)**	PT_NDA	PT_PAMT	**PT_REG**	0
	Yes	**2MB**	**2MB (Level 1)**	PT_NDA	**PT_REG**	N/A	All 0
	Yes	**1GB**	**1GB (Level 2)**	**PT_REG**	All PT_NDA	N/A	All 0
Pending Removal	**Yes**	**2MB**	**2MB (Level 1)**	PT_NDA	**PT_PR**	N/A	All 0
	Yes	**1GB**	**1GB (Level 2)**	**PT_PR**	All PT_NDA	N/A	All 0

8.4. Overview of Memory Protection using Access Control Table (ACT)

Enumeration: Usage of ACT is enumerated by TDX_FEATURES0.ACT (bit 14), readable by TDH.SYS.RD*.

On certain platforms, an Access Control Table (ACT) is used by memory controllers to help protect physical memory marked as TD private from being accessed using shared HKIDs. ACT resides in memory, inside the SEAM range, and thus is protected from software access, except for the TDX module. Each bit in ACT is associated with a 4KB physical memory page:

0: Shared page – memory that can be accessed using shared HKIDs.

1: Private page – memory that can only be accessed by TDs and the TDX module, using private HKIDs.

ACT data is duplicated per memory controller.

The TDX module is responsible for managing the ACT, as follows:

- Initialize the ACT tables.
- Enable the ACT lookup feature in the memory controller.
- Update the table bits whenever a memory page is converted from private to shared and vice versa.
  - When PAMT.PT moves from PT_NDA to any TD private page states, update the ACT page bit to private (1).
  - When PAMT.PT moves from TD private page states to PT_NDA, the update ACT page bit to shared (0).

8.5. Adding Physical Pages

This section discusses various aspects of adding physical pages to TDX, i.e., converting them from shared memory to private memory.

8.5.1. Dynamic PAMT Considerations

If the TDX module is configured for Dynamic PAMT, adding a 4KB physical page requires that the new page HPA will be mapped by PAMT pages. See 8.3.4.2 above for details.

PAMT pages can be added on demand. An interface function which encounters a missing PAMT page pair returns a TDX_MISSING_PAMT_PAGE_PAIR status. The host VMM can add the PAMT page pair and retry the failed function. For details, see the [ABI Spec] definitions for each function which adds 4KB physical pages.

8.5.2. Adding Pages not Mapped to the Guest TD

By design, TD control structure pages TDR, TDCX and TDVPR are not mapped to the guest TD's GPA space, and they are only managed using their HPA. The functions TDH.MNG.CREATE, TDH.MNG.ADDCX, TDH.VP.CREATE and TDH.VP.ADDCX are designed to add 4KB control structure pages PT_TDR, PT_TDCX and PT_TDVPR, respectively. The overall process is described in 7.3 and 7.4.

8.5.3. Adding Pages and Mapping to the Guest TD's GPA

The following page types are associated with a guest TD's GPA:

- Guest TD private pages
- Secure EPT pages are mapped to the guest TD's GPA space.

Those pages are added given their HPA and the required GPA. The functions TDH.MEM.PAGE.ADD, TDH.MEM.PAGE.AUG, TDH.MEM.PAGE.RELOCATE and TDH.IMPORT.MEM add a PT_REG page, and the functions TDH.MEM.SEPT.ADD and TDH.MEM.PAGE.DEMOTE add a 4KB PT_EPT page. TD private memory management functions are described in Chapter 9. This section describes only their physical page management aspects.

8.6. *Reclaiming Physical Pages*

This section discusses various aspects of reclaiming physical pages from TDX, i.e., converting them from private memory to shared memory.

8.6.1. Dynamic PAMT Considerations

PAMT pages can be opportunistically removed. If the TDX module is configured for Dynamic PAMT, when a 4KB physical page is reclaimed the TDX module provides a hint that the PAMT page pair mapping it (see 8.3.4.2 above) may be removed. For details, see the [ABI Spec] definitions for each function which reclaims 4KB physical pages.

8.6.2. Platforms not Using ACT: Required Cache Flush and Initialization by the Host VMM

Once a physical page is reclaimed from a TD, it should be free for use by the host VMM for any purpose, provided that the operations described below are done.

Page Initialization

Before the physical page is used for anything except TD private memory page or TDX control structure page, the host VMM should initialize it using MOVDIR64B. This helps ensure that no content encrypted with a private HKID remains for that physical page, which may result in an integrity violation or TD bit mismatch detection when later being read using a shared HKID.

8.6.3.1.

If the page is to be used as a new TD private memory page or TDX control structure page, this initialization is not required since the TDX module will initialize the page.

8.6.3. Platforms Using ACT: Required Cache Flush, Initialization and ACT Update

ACT Platforms: Overview of the Host VMM Operation

Reclaiming large pages, as part of TDH.MEM.PAGE.REMOVE and TDH.PHYMEM.PAGE.RECLAIM is a long and interruptible operation. See below for details.

8.6.3.2. *ACT Platforms: Overview of the TDX Module Operation*

On platforms which use ACT for memory protection, reclaiming physical pages requires cache lines flushing, page write over and ACT bit clearing. These operations are done by the TDX module as part of the page reclamation sequence.

8.6.3.2.1. Page Overwrite and Cache Flush

To help avoid stability issues caused by cache line aliasing, the TDX module is designed to ensure that no cache lines associated with the reclaimed page are in a Modified state, before the page is reused for any purpose. In addition, the TDX module overwrites the page content with a per-TD random value. This value is generated when the TD is created (TDH.MNG.CREATE). This helps ensure that no cyphertext of known cleartext content is revealed after the page becomes shared.

8.6.3.2.2. Marking the Page as Shared

The TDX module clears the applicable ACT bits to mark the page as shared.

8.6.3.3. *ACT Platforms: Page Reclamation Sequence for Large Pages*

4KB pages are reclaimed directly by the applicable interface functions (e.g., TDH.MEM.PAGE.REMOVE), as discussed in the following sections.

For 2MB and 1GB physical pages, page reclamation may take a long time to run, thus this operation is interruptible and resumable. For backward compatibility, such interface functions (e.g., TDH.MEM.PAGE.REMOVE) have two usage modes, selected by their version number input parameter:

Backward Compatible Mode:	Upon detecting a pending interrupt, the function returns to the host VMM without incrementing RIP. The host VMM is expected to handle the interrupt; typically, when done the interface function natively gets called again.
Explicit Mode:	Upon detecting a pending interrupt, the function returns to the host VMM with a TDX_INTERRUTED_RESUMABLE status. The host VMM is expected to handle the interrupt and then call the interface function again to complete the page.

In both cases, once interrupted the interface functions change the page type to Pending Release (PT_PR). This indicates that the page can no longer be used for its original purpose but has not yet been fully reclaimed.

Once a physical page is reclaimed from a TD, it should be free for use by the host VMM for any purpose.

8.6.4. Reclaiming Pages not Mapped to the Guest TD's GPA Space

There are several cases where pages are not considered as mapped to the guest TD:

- Control structure pages are not mapped to the guest TD.
- In TD_TEARDOWN state, as described below, no mapping is in effect.
- If the TDX module supports non-blocking mapping resize, PT_TR is a special case. It is not mapped in the TD's GPA space, but there may still be TLB entries associated with it.

8.6.4.1. *Reclaiming TD Pages in TD_TEARDOWN State*

As part of the TD teardown process, the host VMM needs to put the TD into a TD_TEARDOWN state, as described in 7.4. This is a non-recoverable state where TD keys have been reclaimed, all address translations and caches have been flushed, and the TD private memory and control structures (except TDR) are no longer accessible.

By design, in the TD_TEARDOWN state, all TD pages are effectively unmapped. Secure EPT is not accessible, and no GPA-to-HPA mapping can be used. The host VMM must treat all the TD private pages and control structure pages as physical memory and reclaim them using the TDH.PHYMEM.PAGE.RECLAIM function in any order, as long as the TDR page is the last one to be reclaimed.

For TDR page, the intention is for the host VMM to call TDH.PHYMEM.PAGE.WBINVD after calling TDH.PHYMEM.PAGE.RECLAIM. This is required to avoid corruption due to cache line aliasing because the TDR page has still been accessed and modified, even when the TD was in TD_TEARDOWN state.

Reclaiming PT_TR Pages in the TD_KEYS_CONFIGURED State

If the TDX module supports non-blocking mapping resize, as enumerated by TDX_FEATURES0.NON_BLOCKING_RESIZE (bit 35), then PT_TR pages can be reclaimed while the TD's lifecycle state is TD_KEYS_CONFIGURED, its normal state while operating. PT_TR pages are former SEPT pages released by TDH.MEM.PAGE.PROMOTE.

In this state, the CPU might hold stale TLB entries associated with the PT_TR page. Thus, the VMM is expected to perform TLB tracking and TDH.PHYMEM.PAGE.RECLAIM checks this.

8.6.4.2. 8.6.5. Reclaiming Physical Pages as Part of TD Private Memory Management

Functions such as TDH.MEM.PAGE.REMOVE and TDH.MEM.PAGE.PROMOTE are designed to remove TD private pages and Secure EPT pages, respectively. By design, they first make sure the pages are no longer accessible using a GPA, then they mark the physical page as free. This is described in Chapter 9; this section only highlights the physical page reclamation.

9. TD Private Memory Management

This chapter described how the Intel TDX module helps manage TD private memory and guest-physical address (GPA) translation.

9.1. Overview

Intel TDX ISA introduced the concept of private GPA vs. shared GPA, depending on the GPA.SHARED bit. In SEAM non-root mode, the controlling VMCS has two EPT pointer fields:

- The legacy EPT pointer is used for translating the guest TD's memory accesses using a private GPA (i.e., GPA.SHARED == 0).
- A new Shared EPT pointer is used for translating the guest TD's memory accesses using shared GPAs (i.e., GPA.SHARED == 1).

A new GPAW execution control determines the position of the SHARED bit in the GPA, and a new HKID execution control defines the HKID used for accessing TD private memory.

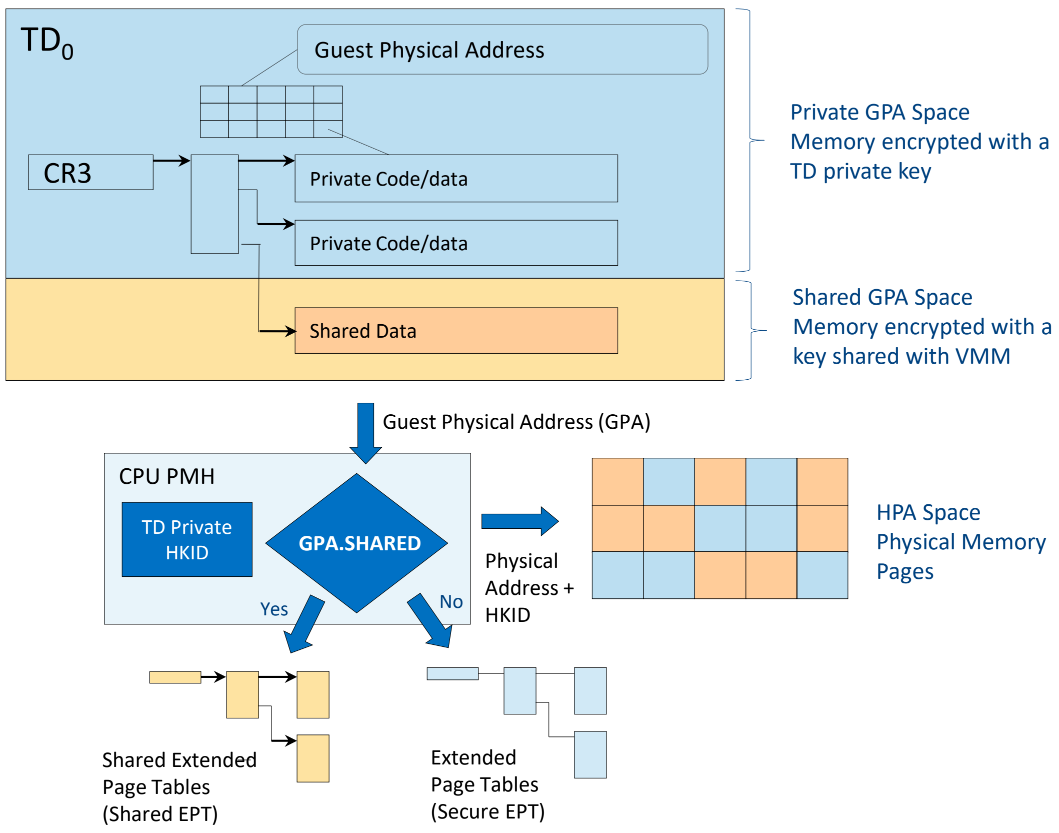


Figure 9.1: Secure EPT Concept

The Intel TDX module maintains a single Secure EPT structure per TD. Secure EPT pages are designed to be opaque; they reside in ordinary memory, and they are encrypted and integrity-protected with the TD's ephemeral private key. The Intel TDX module does not map Secure EPT pages to the guest TD GPA space. Thus, Secure EPT is effectively not accessible by any software besides the Intel TDX module, nor by any devices. Any such access using shared HKID to Secure EPT can lead to data corruption that triggers integrity check failure leading to a machine check fault.

Secure EPT is intended to be managed indirectly by the host VMM using Intel TDX functions. The Intel TDX module helps ensure that the Secure EPT is managed correctly.

The CPU translates shared GPAs using the Shared EPT which resides in host VMM memory. The translation uses a shared HKID, and it is directly managed by the host VMM, just as with legacy VMX.

9.2. Secure EPT Entry

9.2.1. Overview

From the CPU perspective, Secure EPT has the same structure as a legacy VMX EPT.

For the purpose of private memory management, the Intel TDX module holds a state value in each Secure EPT entry. This state value is encoded by multiple bits.

Note: Some Secure EPT entry states are only applicable when certain TDX module features are supported.

Table 9.1: Secure EPT Entry State High Level Description

State Name	Description
FREE	L1 Secure EPT entry does not map a GPA range.
REMOVED	L1 Secure EPT entry is of a removed page
NL_MAPPED	L1 Secure EPT entry maps a private GPA range which is accessible by the guest TD.
NL_BLOCKED	L1 Secure EPT entry maps a private GPA range, but new address translations to that range are blocked.
MAPPED	L1 Secure EPT entry maps a private GPA page which is accessible by the guest TD.
BLOCKED	L1 Secure EPT entry maps a private GPA page but new address translations to that range are blocked.
BLOCKEDW	L1 Secure EPT entry maps a private GPA page, but new address translations for write operations to that range are blocked.
EXPORTED_BLOCKEDW	L1 Secure EPT entry maps a private page that has been blocked for writing and exported.
EXPORTED_DIRTY	L1 Secure EPT entry maps a private page that was exported but is not blocked for writing and its content and/or attributes may have since been modified.
EXPORTED_DIRTY_BLOCKEDW	L1 Secure EPT entry maps a private page that was previously exported, its content and/or attributes may have since been modified and then it was blocked for writing.
PENDING	L1 Secure EPT entry maps a 4KB or a 2MB page that has been dynamically added to the guest TD using TDH.MEM.PAGE.AUG and is pending acceptance by the guest TD using TDG.MEM.PAGE.ACCEPT. This page is not yet accessible by the guest TD.
PENDING_BLOCKED	L1 Secure EPT entry is both pending and blocked.
PENDING_BLOCKEDW	L1 Secure EPT entry is both pending and blocked for writing.
PENDING_EXPORTED_BLOCKEDW	L1 Secure EPT entry is both pending and exported.
PENDING_EXPORTED_DIRTY	L1 Secure EPT entry is both pending and exported and is not blocked for writing.
PENDING_EXPORTED_DIRTY_BLOCKEDW	L1 Secure EPT entry is both pending and exported and is blocked for writing.
REMOVE_IN_PROGRESS	L1 Secure EPT entry maps a private page that is being removed (TDH.MEM.PAGE.REMOVE has been interrupted).
MMIO_MAPPED	L1 Secure EPT entry maps a private MMIO page which is accessible by the guest TD.

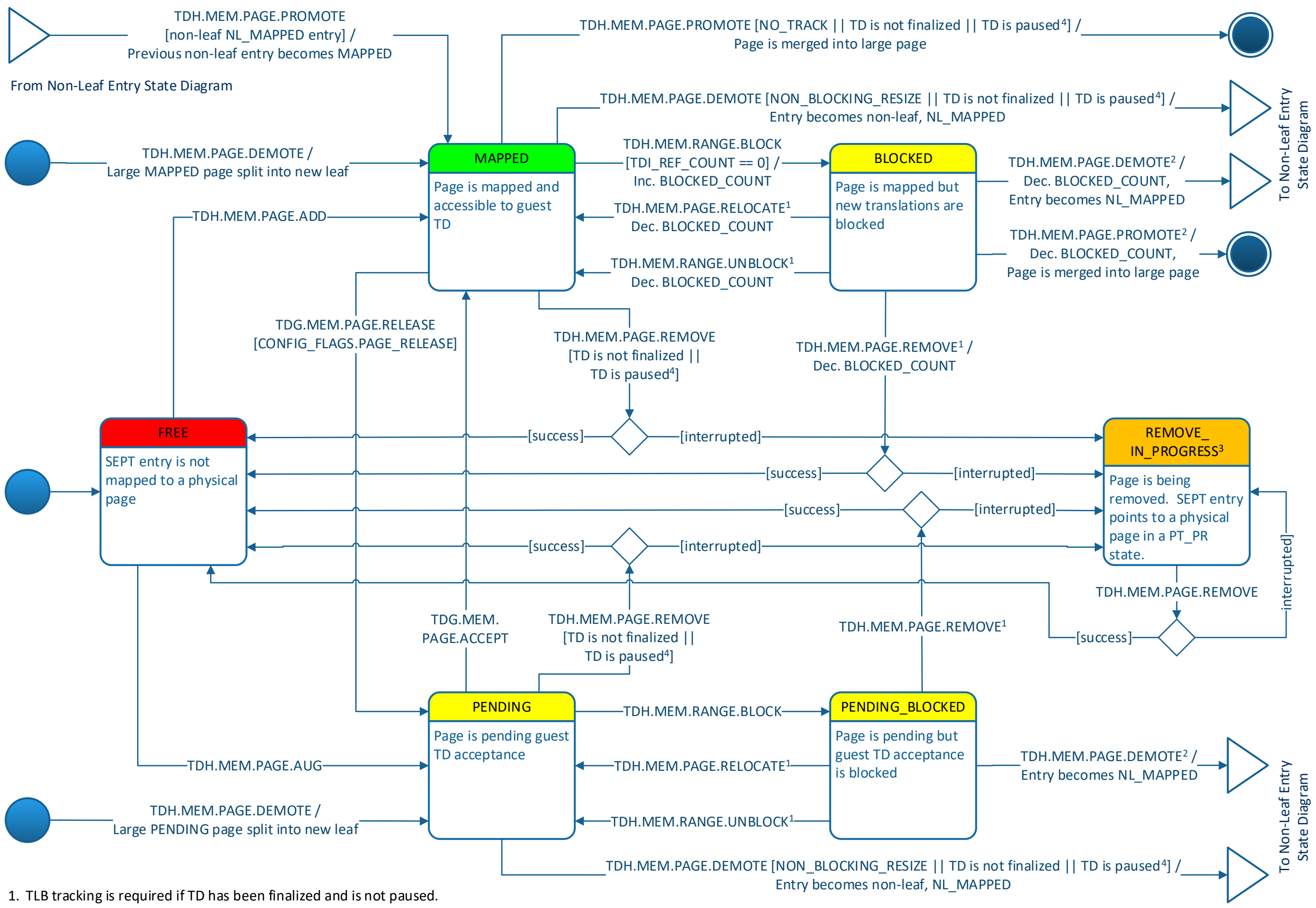
State Name	Description
MMIO_BLOCKED	L1 Secure EPT entry maps a private MMIO page, but new address translations to that page are blocked.
MMIO_PENDING	L1 Secure EPT entry maps a 4KB, 2MB or 1GB MMIO page that is pending acceptance by the guest TD using TDG.MMIO.ACCEPT. This page is not yet accessible by the guest TD.

Secure EPT entry is opaque; the host VMM may not access it directly. The host VMM may read a Secure EPT entry information using the TDH.MEM.SEPT.RD interface function. In addition, multiple other interface functions return the same information in case of an error that is related to a Secure EPT entry. For details, see the [TDX Module ABI Spec].

9.2.2. SEPT Entry State Diagrams

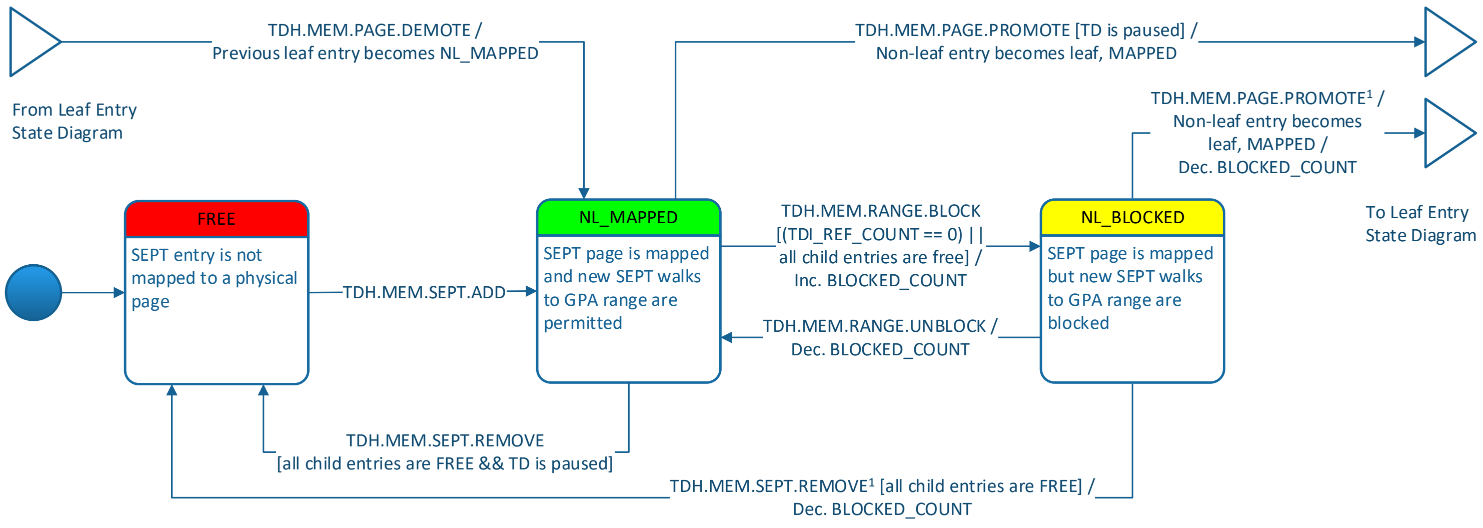
The figures below show partial state diagrams for the basic memory management operation for a leaf and a non-leaf SEPT entry.

Note: The diagrams below are partial. SEPT entry state diagrams for TD migration are provided in the [TD Migration Spec]. SEPT entry state diagrams for TD partitioning are provided in the [TD Partitioning Spec].


TDH.MEM.PAGE.PROMOTE [non-leaf NL_MAPPED entry] / Previous non-leaf entry becomes MAPPED
From Non-Leaf Entry State Diagram
TDH.MEM.PAGE.PROMOTE [NO_TRACK || TD is not finalized || TD is paused4] / Page is merged into large page
TDH.MEM.PAGE.DEMOTE [NON_BLOCKING_RESIZE || TD is not finalized || TD is paused4] / Entry becomes non-leaf, NL_MAPPED
To Non-Leaf Entry State Diagram
TDH.MEM.PAGE.DEMOTE / Large MAPPED page split into new leaf
MAPPED
Page is mapped and accessible to guest TD
TDH.MEM.RANGE.BLOCK [TDI_REF_COUNT == 0] / Inc. BLOCKED_COUNT
BLOCKED
Page is mapped but new translations are blocked
TDH.MEM.PAGE.DEMOTE2 / Dec. BLOCKED_COUNT, Entry becomes NL_MAPPED
TDH.MEM.PAGE.ADD
TDH.MEM.PAGE.RELOCATE1 Dec. BLOCKED_COUNT
TDH.MEM.PAGE.PROMOTE2 / Dec. BLOCKED_COUNT, Page is merged into large page
TDH.MEM.RANGE.UNBLOCK1 Dec. BLOCKED_COUNT
TDG.MEM.PAGE.RELEASE [CONFIG_FLAGS.PAGE_RELEASE]
TDH.MEM.PAGE.REMOVE [TD is not finalized || TD is paused4]
TDH.MEM.PAGE.REMOVE1 / Dec. BLOCKED_COUNT
FREE
SEPT entry is not mapped to a physical page
[success]
[interrupted]
REMOVE_ IN_PROGRESS3
Page is being removed. SEPT entry points to a physical page in a PT_PR state.
TDH.MEM.PAGE.REMOVE
[interrupted]
[success]
TDG.MEM. PAGE.ACCEPT
TDH.MEM.PAGE.REMOVE [TD is not finalized || TD is paused4]
TDH.MEM.PAGE.REMOVE1
PENDING
Page is pending guest TD acceptance
TDH.MEM.RANGE.BLOCK
PENDING_BLOCKED
Page is pending but guest TD acceptance is blocked
TDH.MEM.PAGE.AUG
TDH.MEM.PAGE.RELOCATE1
TDH.MEM.PAGE.DEMOTE2 / Entry becomes NL_MAPPED
TDH.MEM.PAGE.DEMOTE / Large PENDING page split into new leaf
TDH.MEM.RANGE.UNBLOCK1
TDH.MEM.PAGE.DEMOTE [NON_BLOCKING_RESIZE || TD is not finalized || TD is paused4] / Entry becomes non-leaf, NL_MAPPED
To Non-Leaf Entry State Diagram

1. TLB tracking is required if TD has been finalized and is not paused.
2. TLB tracking is required if TD has been finalized and is not paused and non-blocking resize is not supported.
3. This state is applicable only for platforms with ACT-protected memory.
4. Note that if the TD is paused, TDI_REF_COUNT is 0.

Figure 9.2: Secure EPT Leaf Entry Basic Operation Partial State Diagram


TDH.MEM.PAGE.DEMOTE / Previous leaf entry becomes NL_MAPPED
TDH.MEM.PAGE.PROMOTE [TD is paused] / Non-leaf entry becomes leaf, MAPPED
From Leaf Entry State Diagram
TDH.MEM.PAGE.PROMOTE¹ / Non-leaf entry becomes leaf, MAPPED / Dec. BLOCKED_COUNT
To Leaf Entry State Diagram
FREE
SEPT entry is not mapped to a physical page
TDH.MEM.SEPT.ADD
NL_MAPPED
SEPT page is mapped and new SEPT walks to GPA range are permitted
TDH.MEM.RANGE.BLOCK [(TDI_REF_COUNT == 0) | | all child entries are free] / Inc. BLOCKED_COUNT
TDH.MEM.RANGE.UNBLOCK / Dec. BLOCKED_COUNT
NL_BLOCKED
SEPT page is mapped but new SEPT walks to GPA range are blocked
TDH.MEM.SEPT.REMOVE [all child entries are FREE && TD is paused]
TDH.MEM.SEPT.REMOVE¹ [all child entries are FREE] / Dec. BLOCKED_COUNT

1. TLB tracking is required if TD has been finalized and is not paused.

Figure 9.3: Secure EPT Non-Leaf Entry Basic Operation Partial State Diagram

9.3. *Secure EPT Walk*

Host-side (SEAMCALL) Intel TDX functions that manage TD private memory usually accept GPA and Level parameters. They perform a Secure EPT walk which locates the target Secure EPT entry.

If the Secure EPT walk is completed successfully, the Intel TDX function may operate on the located Secure EPT entry. Otherwise, the function typically returns the last visited SEPT entry and its level to the host VMM.

Guest-side (TDCALL) Intel TDX functions typically perform an EPT walk similar to the EPT walk done by the CPU. Only the GPA is provided as an input, and the function may walk the Shared EPT or the Secure EPT, depending on the specific function and the GPA's SHARED bit.

9.4. *Secure EPT Induced TD Exits*

Intel SDM, Vol. 3, 26.2.1 Basic VM-Exit Information

Guest TD memory access to a non-present private GPA causes, in most cases, an asynchronous TD exit with an EPT Violation exit reason. As discussed in 9.2 above, a non-present GPA is any private GPA for which there is either no Secure EPT entry, or the Secure EPT entry is not in the MAPPED state.

Secure EPT-induced TD exits may also be triggered during a guest-side local flow, performing some function on behalf of the guest TD, and executed by the Intel TDX module.

On EPT violation TD exit, VM exit information is provided to the host VMM. This helps the VMM analyze the reason for the EPT violation and take proper action.

Table 9.2: EPT Violation TD Exit Cases and Possible Host VMM Actions

Reason	May be Indicated by	Possible Host VMM Action
Page is not mapped to the TD GPA space	• Exit qualification bits 6:3 value is 0. • Extended exit qualification TYPE (bits 3:0) value is NULL (0). • The host VMM knows, based on its internal data, that either the page or a Secure EPT page that maps it has not been allocated to the TD.	The host VMM may use this as a trigger for dynamic memory allocation (TDH.MEM.PAGE.AUG) or for a post-copy migration import (see [TD Migration Spec]).

Reason	May be Indicated by	Possible Host VMM Action
Page is BLOCKED or GPA range is NL_BLOCKED	• Exit qualification bits 6:3 value is 0. • Extended exit qualification TYPE (bits 3:0) value is NULL (0). • The host VMM knows, based on its internal data, that the page or a Secure EPT page that maps it has been blocked.	The host VMM may resume the TD (TDH.VP.ENTER), possibly after taking some action (e.g., TDH.MEM.PAGE.PROMOTE) for which the page has been blocked.
Page is PENDING or PENDING_EXPORTED_DIRTY	• Exit qualification bits 6:3 value is 0. • Extended exit qualification TYPE (bits 3:0) value is PENDING_EPT_VIOLATION (6).[5] • The host VMM knows, based on its internal data, that the page has been assigned to the TD using TDH.MEM.PAGE.AUG.	This happens if the TD is configured to TD-exit (instead of a #VE) on an EPT violation due to accessing a PENDING page. It normally indicates an error condition; the host VMM may decide to tear the TD down. Configuration is by ATTRIBUTES.SEPT_VE_DISABLE. If CONFIG_FLAGS.FLEXIBLE_PENDING_VE is 1, then the guest TD may select the desired behavior.
Page is blocked for writing (*BLOCKEDW)	• Exit qualification bit 1 value is 1, indicating a write access, and bit 4 is 0, indicating write blocking. • The host VMM knows, based on its internal data, that the page has been blocked for writing using TDH.EXPORT.BLOCKW	The host VMM may unblock the page (TDH.EXPORT.UNBLOCKW). for details, see [TD Migration Spec].
EPT violation during PENDING page acceptance (TDG.MEM.PAGE.ACCEPT)	• Extended exit qualification TYPE (bits 3:0) value is ACCEPT (1). • See the discussion in 9.10 below and the [ABI Spec] for details.	Depending on the information provided in the extended exit qualification, the host VMM may demote the page, add an SEPT page, add a page or retry the operation after the page is not blocked. See the discussion in 9.10 below and the [ABI Spec] definition of TDG.MEM.PAGE.ACCEPT for details.
EPT violation during TDG.MEM.PAGE.ATTR.WR	• Extended exit qualification TYPE (bits 3:0) value is ATTR_WR (5).	Depending on the information provided in the extended exit qualification, the host VMM may demote the page or add an L2 SEPT page. See the [ABI Spec] definition of TDG.MEM.PAGE.ATTR.WR for details.
EPT violation caused by guest-side interface function failure of GPA→HPA translation	• Extended exit qualification TYPE (bits 3:0) value is GPA_DETAILS (2).	Similar to the above cases where the page is not mapped, is blocked or is blocked for writing, except that more information is provided in the extended exit qualification.

By design, since secure EPT is fully controlled by the TDX module, an EPT misconfiguration on a private GPA indicates a TDX module bug and is handled as a fatal error.

[5] Availability of this indication is enumerated by TDX_FEATURES0.PENDING_EPT_VIOLATION_V2 (bit 16), readable by TDH.SYS.RD*.

9.5. *Secure EPT Induced Exceptions*

9.5.1. #PF Exceptions Related to GPA Reserved Bits

Guest TD memory access, with any reserved GPA bit set to 1, causes a #PF exception. See 11.15.1.5 for details.

9.5.2. EPT Violation Mutated into #VE

See 9.10.4 below for details of handling guest TD memory access to a private GPA for which the Secure EPT entry state is PENDING or PENDING_EXPORTED_DIRTY.

For shared GPA, see 11.15.2 for details.

9.6. *Secure EPT Concurrency*

Secure EPT concurrency rules are designed to allow concurrent operations on multiple Secure EPT entries. Secure EPT concurrency is controlled by the following mechanisms:

- An exclusive/shared lock on the whole Secure EPT tree.
- A host-priority mutex on each Secure EPT entry.

Host-Side (SEAMCALL) Interface Functions

- TDX module interface functions that use GPA as an input acquire a lock on the whole Secure EPT tree of the target TD to help prevent changes to the tree while they execute.
  - Interface functions that may impact a whole sub tree of the Secure EPT tree acquire an **exclusive lock** on the Secure EPT tree. These include TDH.MEM.RANGE.BLOCK, TDH.MEM.RANGE.UNBLOCK and TDH.MEM.SEPT.REMOVE.
  - Other interface functions acquire a **shared lock** on the Secure EPT tree.
- Most interface functions that use GPA as an input acquire an **exclusive host-priority lock** on the Secure EPT entry or entries which they use. An exception to this is TDH.MEM.SEPT.RD, which just reads a Secure EPT entry and does not use it to actually access memory.
- Host-side interface functions that obtain exclusive access to the TDR page (and thus, to the whole TD), such as TDH.MEM.PAGE.ADD, are considered as implicitly having exclusive access to the Secure EPT tree and each of its entries.

Guest-Side (TDCALL) Interface Functions

Guest-side TDX module interface functions that need to translate a GPA to an HPA emulate the CPU’s top-down EPT walk operation.

- Guest-side interface functions have no concurrency restrictions on the whole Secure EPT tree.
- Guest-side interface functions that need to manage a Secure EPT entry acquire an **exclusive host-priority lock** on that entry. These include TDH.MEM.PAGE.ACCEPT, TDH.MEM.PAGE.ATTR.RD and TDH.MEM.PAGE.ATTR.WR and TDH.MEM.PAGE.RELEASE.

For details on host-priority concurrency enforcement, see 18.1.4.

9.7. *Introduction to TLB Tracking*

The goal of TLB tracking is to be able to prove (when needed) that no logical processor holds any cached Secure EPT address translations to a given TD private **GPA range**. TLB tracking is required when removing a mapped TD private page (TDH.MEM.PAGE.REMOVE) or when changing the page mapping size (TDH.MEM.PAGE.PROMOTE), etc.

Cached address translations include implicit address translations (TLB) and paging structure translations (PxE) held by the CPU. In addition, GPAs that are translated by the TDX module to HPA and written to VMX control structure fields, to be read by the CPU, are also considered cached address translation.

For TDX Connect, TLB tracking is required for MMIO pages. In addition, a similar mechanism is used for IOMMU TLB tracking, for attached I/O devices. See the [TDX Connect Spec] for details.

Conditions when TLB Tracking is not Required

TLB tracking is not required when the TD's OP_STATE implies that no TD VCPU may run at the time GPA mapping operation modification (e.g., TDH.MEM.PAGE.REMOVE) is done. The only OP_STATE values when TD VCPUs may run are the following:

- RUNNABLE
- LIVE_EXPORT
- LIVE_IMPORT

In addition, TLB tracking is not required if the GPA range's Secure EPT entry state implies that no cached address translations may exist for that Secure EPT entry. This applies to the following SEPT entry states (see the [TDX Module ABI Spec] for details):

- EXPORTED_BLOCKEDW, PENDING_EXPORTED_BLOCKEDW: The page has been exported by TDH.EXPORT.MEM and it is blocked for writing.
- FREE, REMOVED: The page has been removed (e.g., by TDH.MEM.PAGE.REMOVE).

GPA Range TLB Tracking Sequence

This sequence is intended to be used by the host VMM to help guarantee no EPT TLB entries exist to a set of GPA ranges.

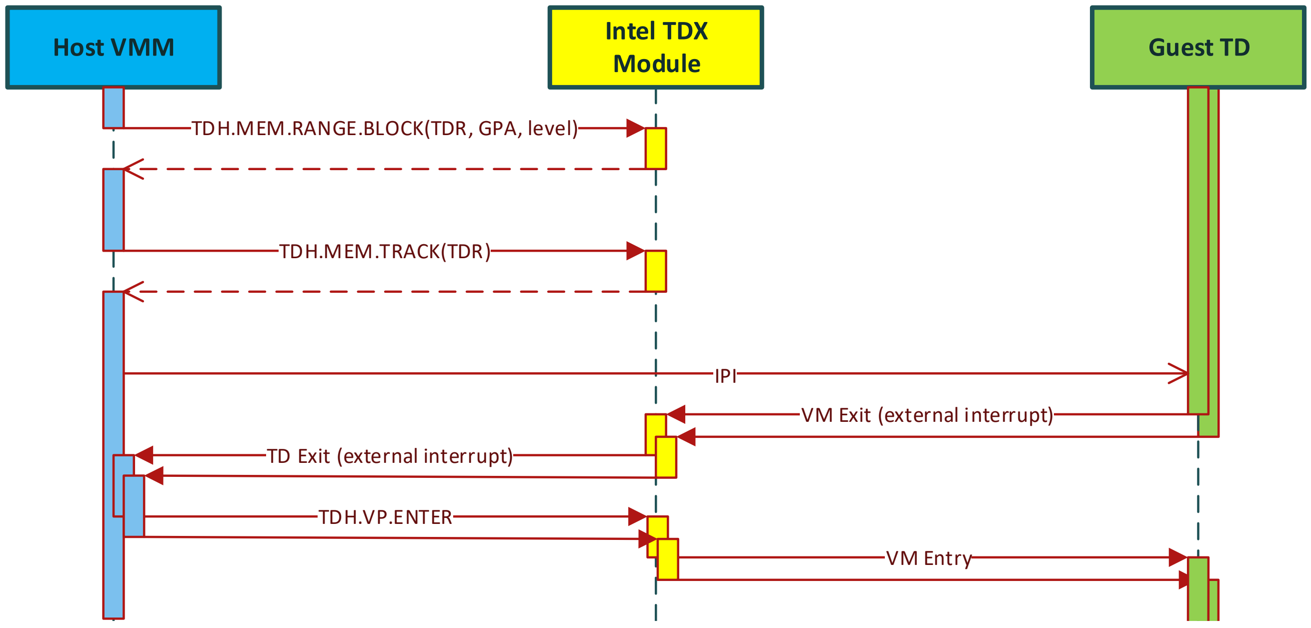

Host VMM
Intel TDX Module
Guest TD
TDH.MEM.RANGE.BLOCK(TDR, GPA, level)
TDH.MEM.TRACK(TDR)
IPI
VM Exit (external interrupt)
TD Exit (external interrupt)
TDH.VP.ENTER
VM Entry

Figure 9.4: Typical TLB Tracking Sequence

The sequence typically includes five steps:

1. Execute TDH.MEM.RANGE.BLOCK on each GPA range, blocking subsequent creation of TLB translation to that range. Note that cached translations may still exist at this stage.
2. Execute TDH.MEM.TRACK, advancing the TD's epoch counter.
3. Send an Inter-Processor Interrupt (IPI) to each Remote Logical Processor (RLP) on which any of the TD's VCPUs is currently scheduled.
4. Upon receiving the IPI, each RLP will TD exit to the host VMM.

When each of the TD VCPUs has been inactive at least once following TDH.MEM.TRACK, the target GPA ranges are considered tracked. Even though some LPs may still hold TLB entries to the target GPA ranges, the following TD entry to each of the TD VCPUs is designed to flush them.

Note: If the host VMM counts the number of active VCPUs, and following TDH.MEM.TRACK this number is 0, the host VMM may skip the IPIs – all VCPUs are already considered tracked.

5. Normally, the host VMM on each RLP will treat the TD exit as spurious and will immediately re-enter the TD.

9.8. Secure EPT Build and Update: TDH.MEM.SEPT.ADD

The host VMM can use the TDH.MEM.SEPT.ADD function to add a Secure EPT page to a guest TD. TDH.MEM.SEPT.ADD inputs are:

- Target TD, identified by its TDR HPA
- Destination physical page for the new Secure EPT table
- Mapping information: GPA and EPT level

At a high level, TDH.MEM.SEPT.ADD works as follows:

1. Check the TD keys are configured.
2. Check the destination physical page is marked as free in the PAMT.
3. Perform a Secure EPT walk to locate the Secure EPT non-leaf entry which will become the parent entry that maps the new Secure EPT page. To help prevent re-maps, TDH.MEM.SEPT.ADD checks the mapping does not already exist, else it aborts the operation.
4. On platforms using ACT-protected memory, mark the new SEPT page's ACT bit(s) as private.
5. Initialize the target page to zero using the target TD's private HKID and direct writes (MOVDIR64B).
6. Update the parent Secure EPT entry to map the page as MAPPED.
7. Update the page's PAMT entry with the PT_EPT page type and the TDR PA as the OWNER.

The Secure EPT's root page (EPML4 or EPML5, depending on whether the host VMM uses 4-level or 5-level EPT) does not need to be explicitly added. It is created during TD initialization (TDH.MNG.INIT) and is stored as part of TDCS. On each VCPU initialization, TDH.VP.INIT copies the address of the Secure EPT root page to the VCPU's TD VMCS's EPTP field clearing the HKID bits to 0[6].

The following example illustrates the build process of a 4-level Secure EPT hierarchy:

1. The host VMM calls TDH.MNG.CREATE(TDR_PA = TDR_0) to create the TD.
2. The host VMM calls TDH.MNG.ADDCX(TDR_PA = TDR_0, DST_PA = TDCX_PAGE_PA) multiple times to allocate pages for TDCS. One of those pages will be used to host the Secure EPT root page D_0.
3. Host VMM calls TDH.MNG.INIT(TDR_PA = TDR_0) to initialize the TD and set an EPML4 page in one of the previously added TDCX pages as the Secure EPT root page. This updates TDCS.EPTP.
4. TDH.VP.INIT of each VPCU copies TDCS.EPTP to the TD VMCS's EPTP field.
5. Host VMM calls TDH.MEM.SEPT.ADD(TDR_PA = TDR_0, DST_PA = D_1, GPA = G_0, LVL= 3) to add an EPDPT page.
6. Host VMM calls TDH.MEM.SEPT.ADD(TDR_PA = TDR_0, DST_PA = D_2, GPA = G_0, LVL= 2) to add an EPD page.
7. Host VMM calls TDH.MEM.SEPT.ADD(TDR_PA = TDR_0, DST_PA = D_3, GPA = G_0, LVL= 1) to add an EPT page.

[6] The CPU adds the TD's private HKID on EPT walks. Having HKID as 0 allows the host VMM to use INVEPT, for managing the usage of shared EPT which shares the ASID with the TD's secure EPT (see).

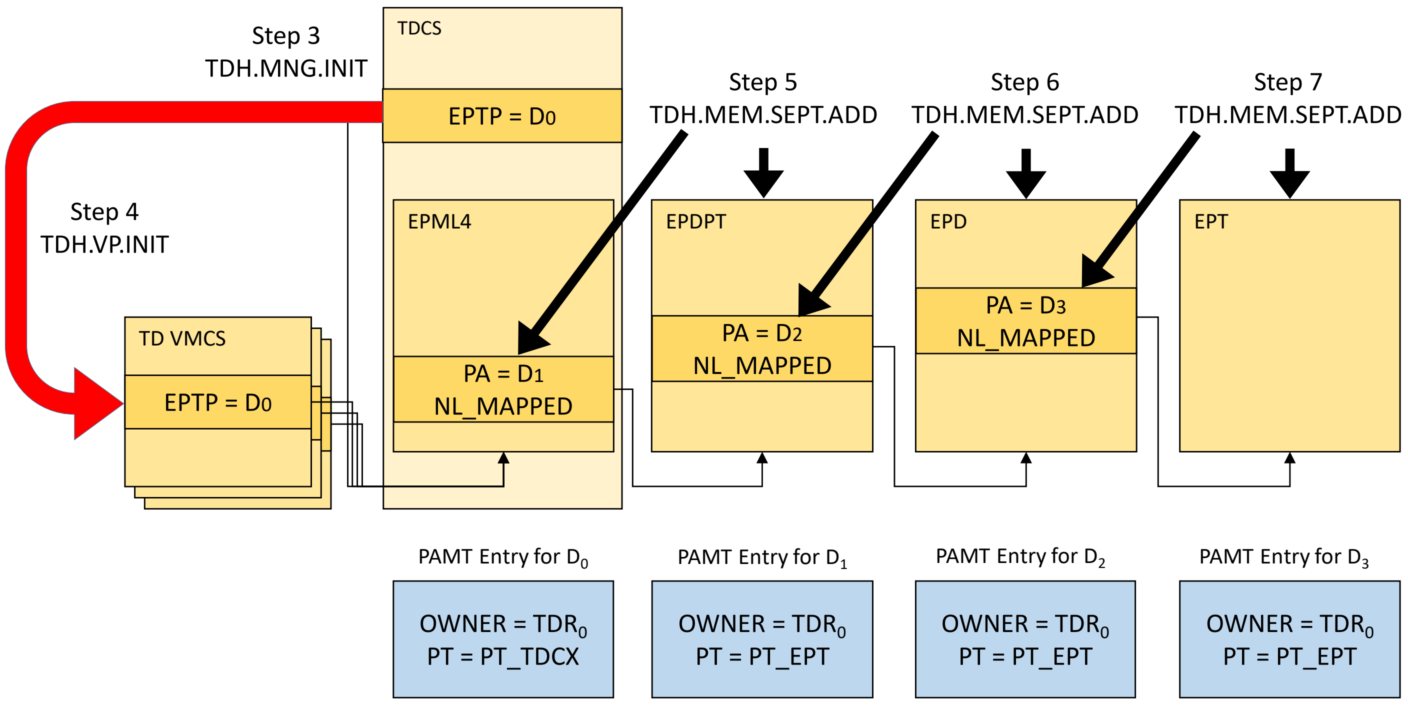


Figure 9.5: Typical Secure EPT Hierarchy Build Process

Dynamic PAMT:	If the TDX module is configured for Dynamic PAMT, adding a 4KB SEPT page requires that the new page HPA will be mapped by PAMT pages. The PAMT pages may be added on demand, based on a status code returned by TDH.MEM.SEPT.ADD. See 8.3.4.2 and the [ABI Spec] for details.
Cache Lines Flushing (Future):	On future platforms, if cache line flushing is required, as enumerated by TDX_FEATURES.CLFLUSH_BEFORE_ALLOC (bit 23), then to help avoid stability issues caused by cache line aliasing, the VMM should ensure that no cache lines associated with the added physical SEPT page are in a Modified state, before calling TDH.MEM.PAGE.AUG. This is typically done by calling TDH.PHYMEM.PAGE.WBINVD.

9.9. Adding TD Private Pages during TD Build Time: TDH.MEM.PAGE.ADD

Adding TD private pages with arbitrary content is allowed only during TD build time (before TDH.MR.FINALIZE). The host VMM adds and maps 4KB private pages to a guest TD using TDH.MEM.PAGE.ADD with the following inputs:

- Target TD, identified by its TDR physical address
- Source page physical address
- Destination page physical address
- Destination page GPA

At a high level, TDH.MEM.PAGE.ADD works as follows:

1. Check the TD has not been initialized.
2. Check the TD keys are configured.
3. Check the destination physical page is marked as free in the PAMT.
4. Perform a Secure EPT walk to locate the parent Secure EPT leaf entry that is going to map the new TD private page. To help prevent re-maps, TDH.MEM.PAGE.ADD checks the mapping does not already exist, else it aborts the operation.
5. On platforms using ACT-protected memory, mark the new private page's ACT bit(s) as private.
6. Copy the source page to the destination page using the target TD's private HKID and direct writes (MOVDIR64B).
7. Update the previously located parent Secure EPT leaf entry to map the page as MAPPED.
8. Update the TD measurement with the new page GPA (as described in 12.2.1).
9. Update the PAMT entry with the PT_REG page type and the TDR PA as the OWNER.

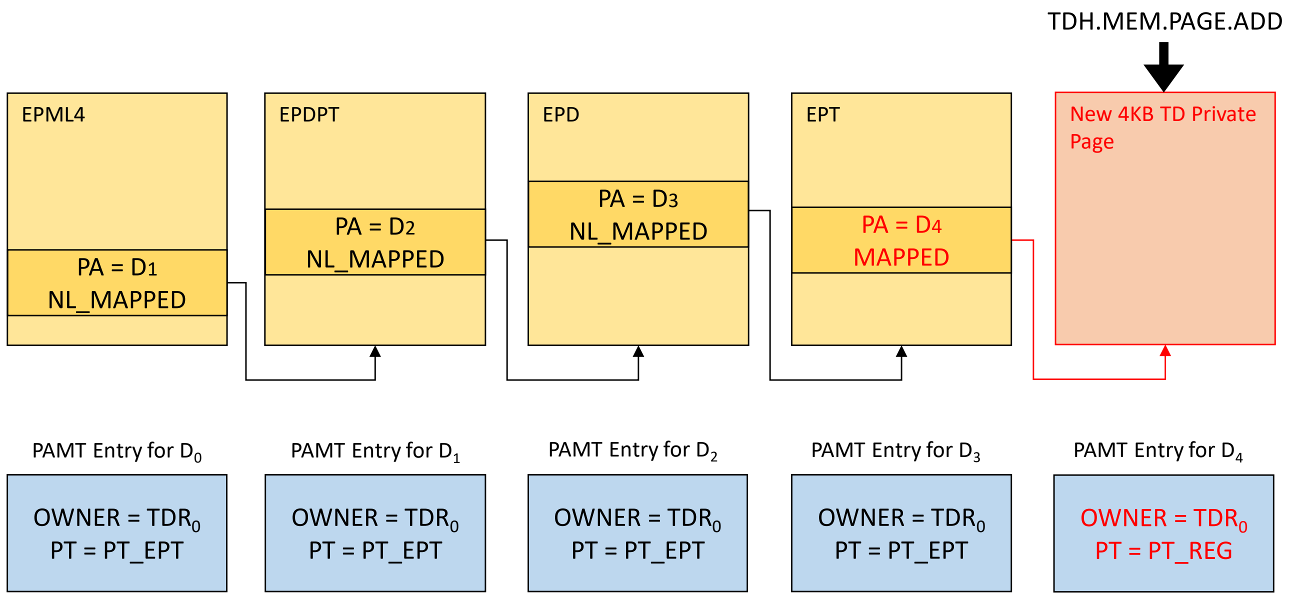


Figure 9.6: Typical Sequence for Adding a TD Private Page during TD Build Time

Dynamic PAMT:	If the TDX module is configured for Dynamic PAMT, adding a 4KB TD private page requires that the new page HPA will be mapped by PAMT pages. The PAMT pages may be added on demand, based on a status code returned by TDH.MEM.PAGE.ADD. See 8.3.4.2 and the [ABI Spec] for details.
Cache Lines Flushing (Future):	On future platforms, if cache line flushing is required, as enumerated by TDX_FEATURES.CLFLUSH_BEFORE_ALLOC (bit 23), then to help avoid stability issues caused by cache line aliasing, the VMM should ensure that no cache lines associated with the added physical page are in a Modified state, before calling TDH.MEM.PAGE.ADD. This is typically done by calling TDH.PHYMEM.PAGE.WBINVD.

9.10. Dynamically Adding TD Private Pages

9.10.1. Overview

Dynamically adding TD private pages after the guest TD has been initialized is typically done as a three-step process:

- The host VMM can update Secure EPT using TDH.MEM.SEPT.ADD and TDH.MEM.SEPT.REMOVE.
- The host VMM adds and maps a 4KB or a 2MB TD private page using TDH.MEM.PAGE.AUG. This page is not measured. The Secure EPT entry state for that added page is PENDING.
- The guest TD must accept the page before it can access it, using TDG.MEM.PAGE.ACCEPT. The page content is zeroed out.

This process is designed to help prevent attacks where the host VMM could remove arbitrary pages from the guest TD's GPA space (using TDH.MEM.PAGE.REMOVE) and replace them with zeroed-out pages.

An attempt by the guest TD to access a page that has been dynamically added by TDH.MEM.PAGE.AUG but has not yet been accepted by TDH.MEM.PAGE.ACCEPT results in either a #VE exception or a TD exit, depending on configuration. See below for details.

Refer also to the software flow described in 3.3.1.1.

9.10.2. PENDING Page Addition by the Host VMM: TDH.MEM.PAGE.AUG

The host VMM can add and map 4KB and 2MB private pages to a guest TD in a non-present and pending state using TDH.MEM.PAGE.AUG, with the following inputs:

- Target TD, identified by its TDR physical address
- Destination page physical address
- Destination page GPA

At a high level, TDH.MEM.PAGE.AUG works as follows:

1. Check the TD keys are configured.
2. Check that the TD has either been initialized (by TDH.MNG.INIT) and no migration session is in progress, or that migration is in progress, but the TD is runnable (live export or import).
3. Check the destination physical page is marked as free in the PAMT.
4. Perform a Secure EPT walk to locate the parent Secure EPT leaf entry that is going to map the new TD private page. To help prevent re-maps, TDH.MEM.PAGE.AUG checks the mapping does not already exist, else it aborts the operation.
5. Update the previously located parent Secure EPT leaf entry to map the page as PENDING.
6. On platforms using ACT-protected memory, mark the new private page's ACT bit(s) as private.
7. Update the PAMT entry with the PT_REG page type and the TDR PA as the OWNER.

Note that TDH.MEM.PAGE.AUG does not need to access the destination page itself; the page is initialized later on by TDG.MEM.PAGE.ACCEPT.

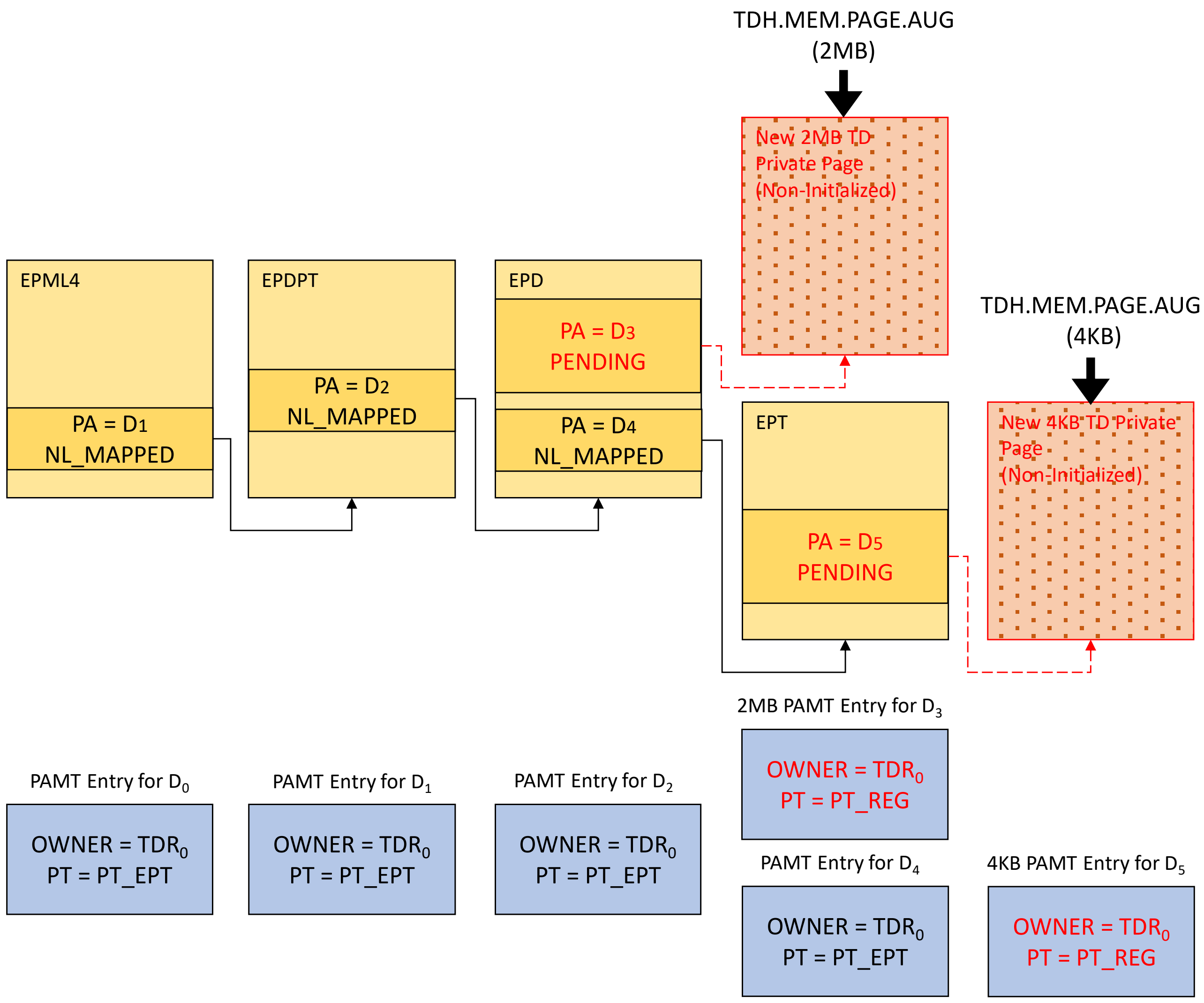


Figure 9.7: Host VMM Adding a 4KB or a 2MB TD Private Page

Dynamic PAMT: If the TDX module is configured for Dynamic PAMT, adding a 4KB TD private page requires that the new page HPA will be mapped by PAMT pages. The PAMT pages may be added on-demand, based on a status code returned by TDH.MEM.PAGE.AUG. See 8.3.4.2 and the [ABI Spec] for details.

Cache Lines Flushing (Future): On future platforms, if cache line flushing is required, as enumerated by TDX_FEATURES.CLFLUSH_BEFORE_ALLOC (bit 23), then to help avoid stability issues caused by cache line aliasing, the VMM should ensure that no cache lines associated with the added physical page are in a Modified state, before calling TDH.MEM.PAGE.AUG. This can be done by calling TDH.PHYMEM.PAGE.WBINVD.

9.10.3. PENDING Page Acceptance by the Guest TD: TDG.MEM.PAGE.ACCEPT

Description

The guest TD can accept a dynamically added 4KB or 2MB page using TDG.MEM.PAGE.ACCEPT with the page GPA and size inputs.

At a high level, TDG.MEM.PAGE.ACCEPT works as follows:

1. Perform a Secure EPT walk to locate the parent Secure EPT leaf entry that maps the TD private page, and handle the walk results as described in the table below.

9.10.3.1.

Table 9.3: TDG.MEM.PAGE.ACCEPT SEPT Walk Cases

SEPT Walk Terminal Entry			TDG.MEM.PAGE.ACCEPT Operation	Typical Software Handling
Level	**Leaf or Non-Leaf**	**State**		
Higher than requested	Leaf	Guest-accessible, i.e., MAPPED or EXPORTED_DIRTY (e.g., 2MB PTE present for a 4KB request).	Return a status code indicating success, with a warning that the page is already present and mapped at a level higher than requested.	Option 1: This is OK, the host VMM did not use the memory released by the TD. Option 2: This is a guest bug; the status code helps debugging it.
		Not guest-accessible and not FREE (e.g., 2MB PTE pending for a 4KB request).	TD exit with EPT violation indicating the error SEPT entry level and state, and the guest-requested accept level. See the [TDX Module ABI Spec].	The host VMM demotes the page to match the requested accept size. It then re-enters the guest TD. TDG.MEM.PAGE.ACCEPT is re-invoked.
	Non-Leaf	Not guest-accessible (e.g., blocked PDE for a 4KB request).	TD exit with EPT violation indicating the error SEPT entry level and state, and the guest-requested accept level. See the [TDX Module ABI Spec].	This may be used as a guest TD request from the host VMM to add a page. The host VMM adds SEPT pages (TDH.MEM.SEPT.ADD) and the requested page (TDH.MEM.PAGE.AUG). It then resumes the guest.
Same as requested	Non-Leaf	Other than FREE (e.g., requested 2MB entry is mapped to an EPT page instead of being a leaf)	Return a status code indicating a size mismatch error.	The guest falls back to accept the range using 4K size.
	Leaf	Guest-accessible, i.e., MAPPED or EXPORTED_DIRTY	Return a status code indicating success, with a warning that the page is already present.	Option 1: This is OK, the host VMM did not use the memory released by the TD. Option 2: This is a guest bug; the status code helps debugging it.
		Not PENDING nor PENDING_EXPORTED_DIRTY	TD exit with EPT violation indicating the error SEPT entry level and state, and the guest-requested accept level. See the [TDX Module ABI Spec].	The host VMM resolves the blocking (e.g., completes the memory management operation that required blocking) and resumes the guest.
		FREE	TD exit with EPT violation indicating the error SEPT entry level and state, and the guest-requested accept level. See the [TDX Module ABI Spec].	This may be used as a guest TD request from the host VMM to add a page. The host VMM adds the requested page (TDH.MEM.PAGE.AUG) and resumes the guest.

SEPT Walk Terminal Entry			TDG.MEM.PAGE.ACCEPT Operation	Typical Software Handling
Level	**Leaf or Non-Leaf**	**State**		
		PENDING	Complete the operation as described below.	Success

If passed:

Note: Since initializing a 2MB page may take a long time, TDG.MEM.PAGE.ACCEPT is interruptible and resumable.

2. If all the above checks pass, loop until done or interrupted:
   2.1. Initialize the next 4KB chunk of the page to zero using the target TD's private HKID and direct writes (MOVDIR64B).
   2.2. If the whole page has been initialized, update the parent Secure EPT entry to set its state to SEPT_PRESENT.
   2.3. Else, if there is a pending interrupt, resume the guest TD without updating RIP and any GPR. The CPU may handle the interrupt, causing a TD exit. When the TD is resumed, TDH.MEM.PAGE.ACCEPT will re-invoked.

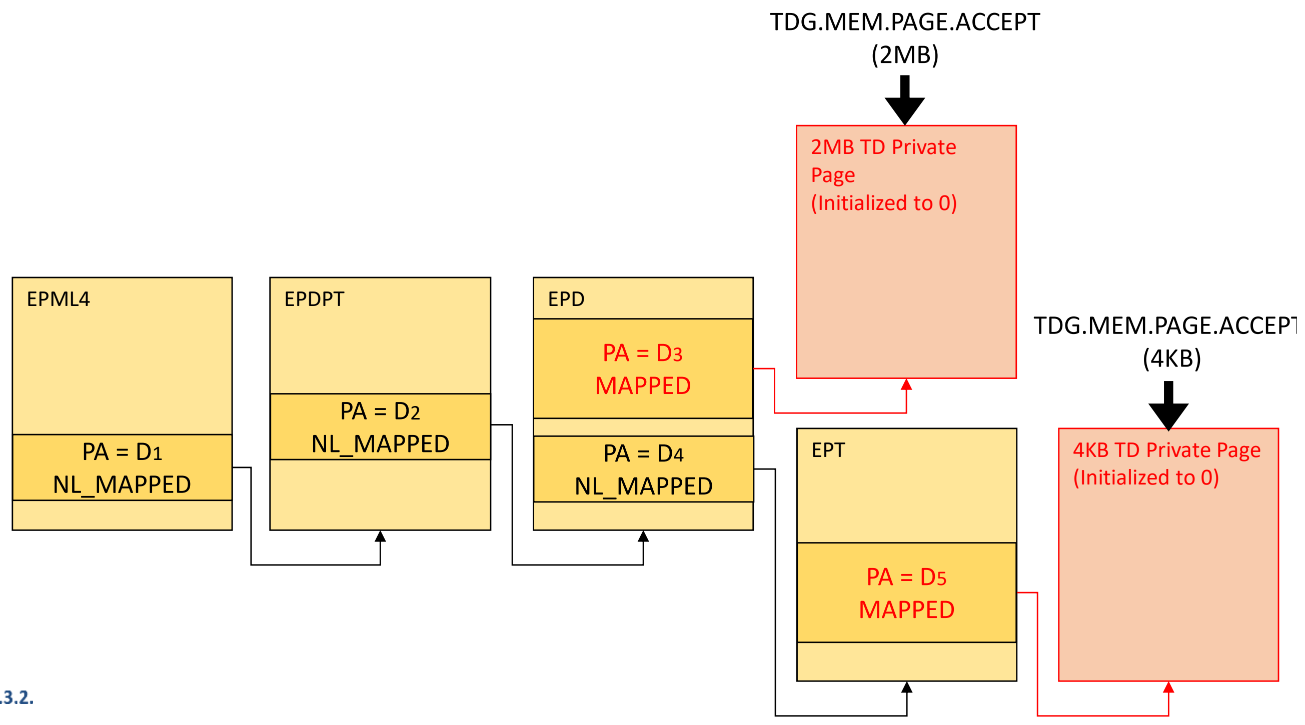

TDG.MEM.PAGE.ACCEPT (2MB)
2MB TD Private Page (Initialized to 0)
TDG.MEM.PAGE.ACCEPT (4KB)
4KB TD Private Page (Initialized to 0)
EPML4
PA = D1 NL_MAPPED
EPDPT
PA = D2 NL_MAPPED
EPD
PA = D3 MAPPED
PA = D4 NL_MAPPED
EPT
PA = D5 MAPPED

Figure 9.8: Guest TD Accepting a 4KB or 2MB Pending TD Private Page

9.10.3.2. *TDG.MEM.PAGE.ACCEPT Concurrency*

Guest-Side

TDG.MEM.PAGE.ACCEPT prevents the guest TD from concurrently accepting the same page by multiple threads. TDG.MEM.PAGE.ACCEPT may also encounter a concurrent host-side operation, such as TDH.MEM.RANGE.BLOCK, that attempts to update the same Secure EPT entry. In such cases, an error is returned to the guest TD, indicating that the Secure EPT entry is busy.

Host-Side

TDG.MEM.PAGE.ACCEPT prevents host-side operations, such as TDH.MEM.RANGE.BLOCK, from concurrently modifying the Secure EPT entry. This is implemented using a host-priority lock, preventing the guest TD from denying service to the host VMM. If a host-side operation fails with a busy indication, the host VMM should retry the operation. For details on host-priority concurrency enforcement, see 18.1.4.

9.10.4. Guest TD (L1) Access to a PENDING Page

The behavior in case of guest TD access to a page in a PENDING or PENDING_EXPORTED_DIRTY page is summarized in the table below. This only applies to L1. L2 VM access to a PENDING page always results in an L2→L1 exit.

- #VE is useful for implementing an accept-on-demand policy. It can be used by the guest TD to trigger a TDG.MEM.PAGE.ACCEPT of the PENDING page.
- TD Exit is useful for guest TD implementations that only map memory that has been accepted into the linear address spaces. For such implementations, an access to a PENDING page indicates a fatal error. The host VMM typically tears the TD down when this happens.

The combinations of configuration flags allow the host VMM to establish a static policy or allow the guests TD to decide on the policy.

Enumeration: Availability of CONFIG_FLAGS.FLEXIBLE_PENDING_VE and TDCS.TD_CTLS.PENDING_VE_DISABLE is enumerated by TDX_FEATURES0.PENDING_EPT_VIOLATION_V2 (bit 16), readable by TDH.SYS.RD*.

Table 9.4: Guest TD (L1) Access to a PENDING Page

Configuration by the Host VMM (TD_PARAMS Input to TDH.MNG.INIT)		Configuration by the Guest TD	Behavior on Guest TD Access of a PENDING Page
ATTRIBUTES. SEPT_VE_DISABLE	**CONFIG_FLAGS. FLEXIBLE_PENDING_VE**	**TDCS.TD_CTLS. PENDING_VE_DISABLE**	
0	0	0	#VE
	1	0	#VE
		1	TD Exit (EPT Violation)
1	0	1	TD Exit (EPT Violation)
	1	0	#VE
		1	TD Exit (EPT Violation)

9.11. Interaction with TDX Connect

If the TD is configured with TDX Connect enabled (CONFIG_FLAGS.TDX_CONNECT is 1), private memory management is restricted to prevent the host VMM from blocking pages while TDIs are attached, which would lead to a silent drop of data. If any TDIs are attached, memory management interface functions are restricted as follows:

- TDH.MEM.RANGE.BLOCK of a leaf SEPT entry is only allowed on PENDING pages.
- TDH.MEM.RANGE.BLOCK of a non-leaf SEPT entry is allowed only if all 512 child SEPT entries are FREE.
- The guest TD must explicitly release a private page using TDG.MEM.PAGE.RELEASE before the host VMM can remove it. Releasing a page puts it in a PENDING state.
- The host VMM can use the non-blocking resize feature of TDH.MEM.PAGE.PROMOTE and TDH.MEM.PAGE.DEMOTE to avoid the need for blocking the GPA range to be promoted or demoted.

See the sections below for more details.

9.12. Releasing a TD Private Page by the Guest TD: TDG.MEM.PAGE.RELEASE

Enumeration: TDX module support of this feature is enumerated by TDX_FEATURES0.PAGE_RELEASE (bit 38), readable using TDG.SYS.RD*.

If supported by the TDX module and allowed by the host VMM by setting CONFIG_FLAGS.PAGE_RELEASE and/or CONFIG_FLAGS.TDX_CONNECT, a guest TD may release a TD private page by TDG.MEM.PAGE.RELEASE, as follows:

1. Before releasing a private page, the guest TD software is expected to ensure all page table TLB translations to the page's GPA are invalidated. This is not enforced by the TDX module.
2. The guest TD then calls TDG.MEM.PAGE.RELEASE, providing the page GPA and the expected mapping size (4KB or 2MB).

   2.1. If the page is mapped at a lower level than requested, the function returns an error code and the actual mapping size. The guest TD may re-invoke TDG.MEM.PAGE.RELEASE specifying the actual mapping size.
   2.2. If the page is mapped at a higher level than requested, this results in an EPT violation TD exit. The host VMM is expected to demote the page, then re-enter the guest TD so TDG.MEM.PAGE.RELEASE is re-invoked.
   2.3. If all checks passed, the TDX module puts the page in a PENDING state and records the current TD epoch in the page metadata for TLB tracking.
3. The guest TD is then expected to notify the host VMM that the page is released, using a software protocol over TDG.VP.VMCALL.
4. The host VMM can remove the page, as with any page in the PENDING state:
   4.1. Call TDH.MEM.RANGE.BLOCK to block the page.
   4.2. Execute the TLB shootdown sequence (TDH.MEM.TRACK and a round of IPIs).
   4.3. Call TDH.MEM.PAGE.REMOVE to remove the page.
5. Alternatively, since the page is in a PENDING state, the guest TD can re-accept it by calling TDG.MEM.PAGE.ACCEPT.

9.13. Page Mapping Resize: Merge and Split

Merging and splitting of the SEPT mapping of 1GB or 2MB GPA ranges is done by TDH.MEM.PAGE.PROMOTE and TDH.MEM.PAGE.DEMOTE.

9.13.1. Overview: Non-Blocking Mapping Resize

Non-blocking mapping resize is designed to support TDX Connect, where no pages must be blocked while TDIs are attached. However, it can be used, if supported by the TDX module, even if TDX Connect is not supported or not enabled for the current TD.

If the TDX module supports non-blocking mapping resize, as enumerated by TDX_FEATURES0.NON_BLOCKING_RESIZE (bit 35), then the following usage model, which eliminates the need to block GPA ranges, is supported:

- No blocking and TLB tracking of the large GPA range to be merged or split is required.
- SEPT pages that are released by TDH.MEM.PAGE.PROMOTE have a new page type: PT_TR.
- The host VMM can keep a pool of such PT_TR pages and use them as inputs to TDH.MEM.PAGE.DEMOTE. TLB track checking of such pages is done when they are used.
- Alternatively, the host VMM can reclaim those pages using TDH.PHYMEM.PAGE.RECLAIM, and use them for any purpose.

9.13.2. Page Merge: TDH.MEM.PAGE.PROMOTE

The host VMM can merge the mapping of 512 consecutive 4KB or 2MB pages into a single 2MB or 1GB page, respectively.

9.13.2.1. *Blocking and TLB Tracking*

Blocking and TLB Tracking is required if the TDX module does not support non-blocking mapping resize, as enumerated by TDX_FEATURES0.NON_BLOCKING_RESIZE (bit 35). Even if the TDX module supports non-blocking mapping resize, the host VMM can choose to first perform the TLB tracking protocol, since this impacts the operation of TDH.MEM.PAGE.PROMOTE as described below.

The host VMM performs the TLB tracking protocol on the large (2MB or 1GB) GPA range:

- The host VMM should first call TDH.MEM.RANGE.BLOCK, which operates on the EPT page for the large range (EPT for 2MB, EPD for 1GB). TDH.MEM.RANGE.BLOCK marks the parent EPT entry for that EPT page as **BLOCKED** and records the TD epoch in the PAMT entry of the EPT page.
- Typically, the host VMM then calls TDH.MEM.TRACK and performs a round of IPIs. After that, there should be no active address translation to the large (2MB or 1GB) GPA range.

Even if the TDX module supports non-blocking mapping resize, the host VMM can choose to first perform the TLB tracking protocol, since this impacts the operation of TDH.MEM.PAGE.PROMOTE as described below.

9.13.2.2. *Promotion*

The actual merge operation is done by TDH.MEM.PAGE.PROMOTE. Figure 9.9 below shows the typical situation before TDH.MEM.PAGE.PROMOTE is called.

Note: For details on TDH.MEM.PAGE.PROMOTE support of partitioned TDs, see the [TD Partitioning Spec].

TDH.MEM.PAGE.PROMOTE has the following inputs:

- The large range GPA
- The large page level (2MB or 1GB)
- If the TDX module supports non-blocking mapping resize, a flag indicating whether TLB tracking check is to be skipped.

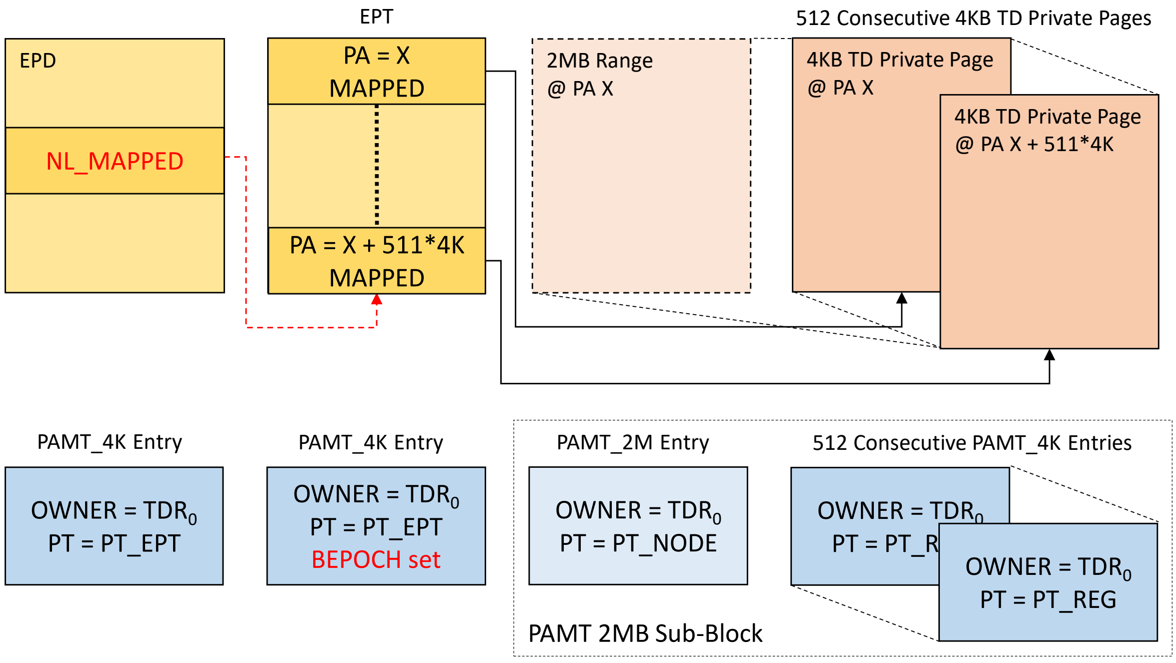

EPD
NL_MAPPED
EPT
PA = X MAPPED
PA = X + 511*4K MAPPED
2MB Range @ PA X
512 Consecutive 4KB TD Private Pages
4KB TD Private Page @ PA X
4KB TD Private Page @ PA X + 511*4K
PAMT_4K Entry
OWNER = TDR$_0$ PT = PT_EPT
PAMT_4K Entry
OWNER = TDR$_0$ PT = PT_EPT BEPOCH set
PAMT_2M Entry
OWNER = TDR$_0$ PT = PT_NODE
512 Consecutive PAMT_4K Entries
OWNER = TDR$_0$ PT = PT_R
OWNER = TDR$_0$ PT = PT_REG
PAMT 2MB Sub-Block

Figure 9.9: Typical State before TDH.MEM.PAGE.PROMOTE of a Range of 512 Consecutive 4KB TD Private Pages

At a high level, TDH.MEM.PAGE.PROMOTE works as follows:

1. If the TDX module does not support non-blocking mapping resize, or the host VMM did not indicate that TLB tracking check should be skipped, check the TLB tracking condition for the large range GPA (i.e., the EPT or EPD page for that range).
2. Check that all 512 entries of that EPT or EPD page are in the MAPPED state and point to leaf pages whose physical address is contiguous within the same 2MB or 1GB range.

If all checks pass, TDH.MEM.PAGE.PROMOTE does the following:

3. Mark all the PAMT_4K or PAMT_2M entries of the small leaf pages (4KB or 2MB, respectively) as PT_NDA.
4. Mark the PAMT_2M or PAMT_1G entry of the merged large (2MB or 1GB, respectively) pages as PT_REG.
5. Set the parent EPT entry to point to the merged large page and mark it as present.
6. If TLB tracking has been done, reclaim the original SEPT physical page as described in 8.6.5.
   6.1. **Platforms not using ACT:** The host VMM should initialize the former EPT or EPD physical page's content before it is reused as a non-private page, as described in 8.6.2.
7. Else (TLB tracking has not been done), mark the original SEPT physical page as PT_TR and record the TD's TD_EPOCH in its PAMT entry. This page can be later reclaimed or used as a new SEPT page input to TDH.MEM.PAGE.DEMOTE.

Figure 9.10 below shows a typical 2MB merged page after TDH.MEM.PAGE.PROMOTE.

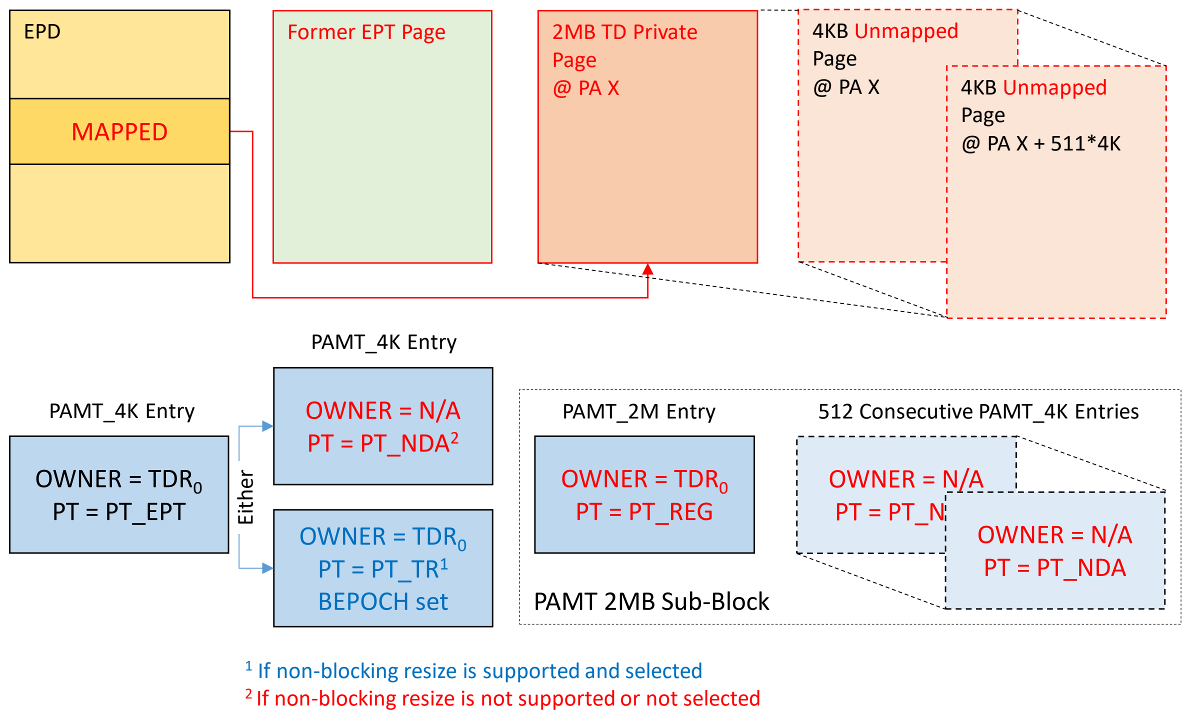

EPD
MAPPED
Former EPT Page
2MB TD Private Page @ PA X
4KB Unmapped Page @ PA X
4KB Unmapped Page @ PA X + 511*4K
PAMT_4K Entry
OWNER = TDR0 PT = PT_EPT
Either
PAMT_4K Entry
OWNER = N/A PT = PT_NDA2
OWNER = TDR0 PT = PT_TR1 BEPOCH set
PAMT_2M Entry
OWNER = TDR0 PT = PT_REG
512 Consecutive PAMT_4K Entries
OWNER = N/A PT = PT_N
OWNER = N/A PT = PT_NDA
PAMT 2MB Sub-Block
1 If non-blocking resize is supported and selected
2 If non-blocking resize is not supported or not selected

Figure 9.10: Typical State of a 2MB TD Private Page after TDH.MEM.PAGE.PROMOTE

Refer also to the software flow described in 3.3.1.3.

Dynamic PAMT: If the TDX module is configured for Dynamic PAMT and the promoted range size is 2MB, TDH.MEM.PAGE.PROMOTE removes that PAMT pages mapping the 512 4KB physical pages. It also returns a hint indicating whether the PAMT pages mapping the removed SEPT pages can be removed. For details, see the [ABI Spec].

9.13.3. Page Split: TDH.MEM.PAGE.DEMOTE

The host VMM can split the mapping of a single 2MB or 1GB page to 512 consecutive 4KB or 2MB pages, respectively.

9.13.3.1. *Blocking and TLB Tracking*

If the TDX module does not support non-blocking mapping resize, as enumerated by TDX_FEATURES0.NON_BLOCKING_RESIZE (bit 35), the host VMM should first perform the TLB tracking protocol on the large (2MB or 1GB) page:

- The host VMM should first call TDH.MEM.RANGE.BLOCK on the large page. TDH.MEM.RANGE.BLOCK marks the parent EPT entry for that page as **BLOCKED** and records the TD epoch in the PAMT entry of the page.
- Typically, the host VMM then calls TDH.MEM.TRACK and performs a round of IPIs. After that, there should be no active address translation to the large (2MB or 1GB) page.

9.13.3.2. *Demotion*

The actual split is done by TDH.MEM.PAGE.DEMOTE. Figure 9.11 below shows the typical situation before TDH.MEM.PAGE.DEMOTE is called.

Note: For details on TDH.MEM.PAGE.DEMOTE support of partitioned TDs, see the [TD Partitioning Spec].

TDH.MEM.PAGE.DEMOTE has the following inputs:

- The large page GPA
- The large page level (2MB or 1GB)
- The physical address of a page that will be used for a new EPT or EPD page. If the TDX module does not support non-blocking mapping resize, this page must be free. Else, the page may also be a PT_TR page, i.e., a former SEPT page converted by TDH.MEM.PAGE.PROMOTE.
- If the TD is partitioned, up to 3 physical addresses of pages that will be used for new L2 EPT or EPD pages.

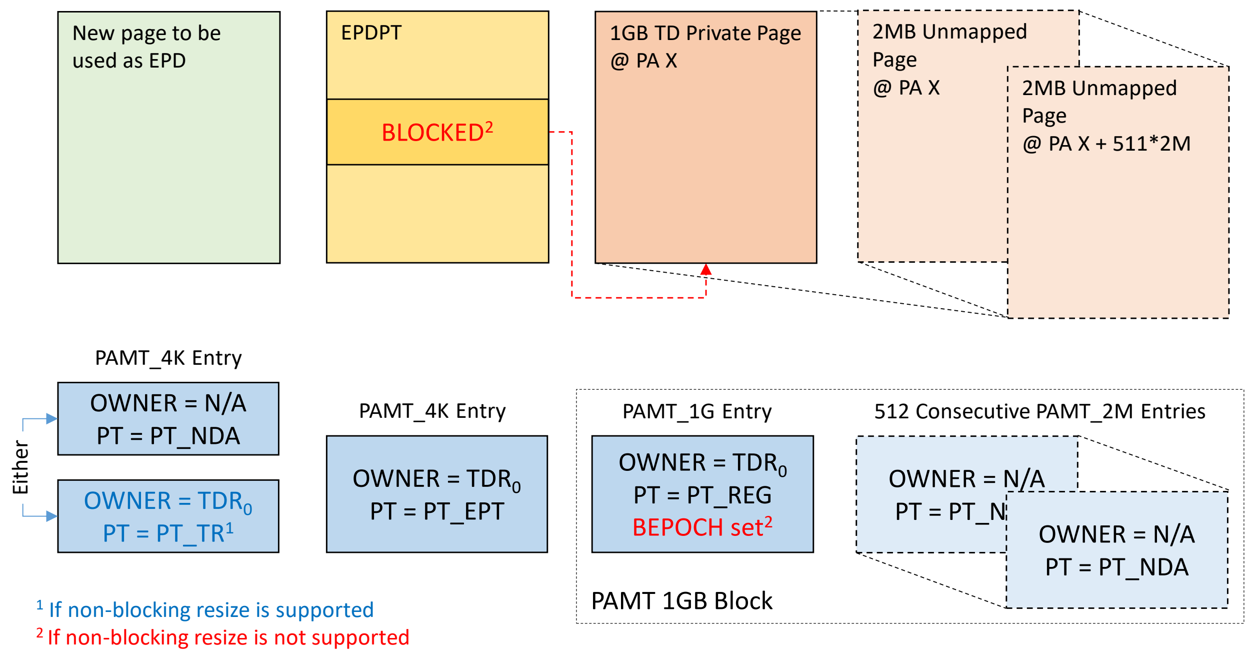


Figure 9.11: Typical State before Calling TDH.MEM.PAGE.DEMOTE on a 1GB Page

At a high level, TDH.MEM.PAGE.DEMOTE works as follows:

1. Check the TLB tracking condition for the large page.
2. Check that the physical page for the new EPT or EPD is either marked as free in the PAMT, or if the TDX module supports non-blocking mapping resize, it is marked as PT_TR and its TLB tracking conditions are met.

If all checks pass, TDH.MEM.PAGE.DEMOTE does the following:

3. Mark the PAMT_2M or PAMT_1G entry of the large (2MB or 1GB respectively) page as PT_NDA.
4. Mark all the PAMT_4K or PAMT_2M entries of the small (4KB or 2MB respectively) consecutive leaf pages as PT_REG.
5. On platforms using ACT-protected memory, mark the new SEPT page's ACT bit(s) as private.
6. Initialize the new SEPT page with 512 EPT entries pointing to the 512 consecutive leaf pages.
7. Mark the new SEPT page's PAMT entry as PT_EPT.
8. Set the parent EPT entry to point to the new EPT or EPD page.

Figure 9.12 below shows the typical state of a 1GB GPA range after TDH.MEM.PAGE.DEMOTE.

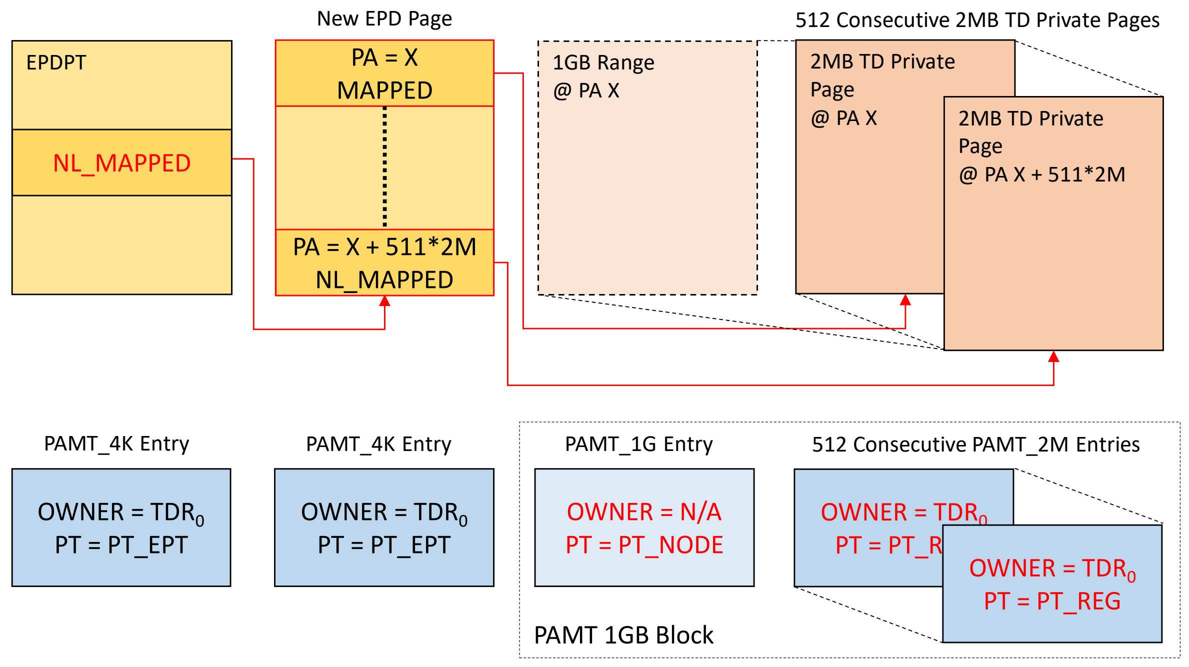


Figure 9.12: Typical State of a 1GB TD Private Range after TDH.MEM.PAGE.DEMOTE

TDH.MEM.PAGE.DEMOTE supports the demotion of PENDING pages.

Refer also to the software flow described in 3.3.1.4.

Dynamic PAMT: If the TDX module is configured for Dynamic PAMT, a new PAMT page pair is required if the page is demoted into 512 4K pages. In addition, adding the new SEPT page HPAs also need be mapped by PAMT pages. Those PAMT pages may be added on-demand, based on a status code returned by TDH.MEM.PAGE.DEMOTE. See 8.3.4.2 and the [ABI Spec] for details.

9.14. Relocating TD Private Pages: TDH.MEM.PAGE.RELOCATE

The host VMM can relocate a 4KB TD private page to another HPA using TDH.MEM.PAGE.RELOCATE. This is useful for, e.g., physical address space de-fragmentation.

If the guest TD's OP_STATE is such that the TD may be running, the host VMM should first perform the TLB tracking protocol on the page. The host VMM should first call TDH.MEM.RANGE.BLOCK on the page. TDH.MEM.RANGE.BLOCK marks the parent EPT entry for that page as **BLOCKED** (if it was MAPPED) or **PENDING_BLOCKED** (if it was PENDING) and records the TD epoch in the PAMT entry of the page.

TDX Connect: If the guest TD has been configured with TDX Connect enabled, TDH.MEM.RANGE.BLOCK on a MAPPED page is only allowed if no TDIs are attached.

Dynamic PAMT: If the TDX module is configured for Dynamic PAMT, relocating to a new 4KB physical page requires that the new page HPA will be mapped by PAMT pages. The PAMT pages may be added on-demand, based on a status code returned by TDH.MEM.PAGE.RELOCATE. In addition, TDH.MEM.PAGE.RELOCATE returns a hint indicating whether the PAMT pages mapping the old 4KB physical page can be removed. See 8.3.4.2 and the [ABI Spec] for details.

Typically, the host VMM then calls TDH.MEM.TRACK and performs a round of IPIs. After that, there should be no active address translation to target page.

The actual relocation is done by TDH.MEM.PAGE.RELOCATE which has the following inputs:

- The page GPA
- The target HPA to which the page will be relocated

At a high level, TDH.MEM.PAGE.RELOCATE works as follows:

1. Check the TD keys are configured.
2. Check the TD has been initialized.
3. Check the target physical page is marked as free in the PAMT.
4. Perform a Secure EPT walk to locate the parent Secure EPT leaf entry that maps the TD private page. Check that the entry has been blocked and get the current HPA.
5. If TLB tracking is required, check the TLB tracking condition for the page.

If all checks pass, TDH.MEM.PAGE.RELOCATE does the following:

6. On platforms using ACT-protected memory, mark the target private page's ACT bit as private.
7. Copy the current physical page to the target physical page using direct writes (MOVDIR64B).
8. Reclaim the old physical page as described in 8.6.5.
9. Mark the PAMT entry of the target page as PT_REG.
10. Update the Secure EPT entry with the new physical page HPA. Set its state to MAPPED or PENDING depending on whether its previous state was BLOCKED or PENDING_BLOCKED, respectively.

Non-ACT Platforms: The host VMM should initialize the old physical page's content before it is reused as a non-private page, as described in 8.6.2.

9.15. Removing TD Private Pages: TDH.MEM.PAGE.REMOVE

The host VMM can remove TD private pages using TDH.MEM.PAGE.REMOVE, freeing them for any use. 4KB, 2MB and 1MB pages can be removed – no demotion is required for large pages.

If the guest TD's OP_STATE is such that the TD may be running, the host VMM should first perform the TLB tracking protocol on the page. The host VMM should first call TDH.MEM.RANGE.BLOCK on the target page.

TDH.MEM.RANGE.BLOCK marks the parent Secure EPT entry for that page as **BLOCKED** (if it was MAPPED) or **PENDING_BLOCKED** (if it was PENDING) and records the TD epoch in the PAMT entry of the page.

Dynamic PAMT: If the TDX module is configured for Dynamic PAMT and the removed page size is 4KB, TDH.MEM.PAGE.REMOVE returns a hint indicating whether the PAMT page pair mapping the removed page can be removed. For details, see the [ABI Spec].

TDX Connect: If the guest TD has been configured with TDX Connect enabled, TDH.MEM.RANGE.BLOCK on a MAPPED page is only allowed if no TDIs are attached. If any TDI is attached, the guest TD should first release the page, by calling TDG.MEM.PAGE.RELEASE as explained in 9.12 above, before the page is blocked and removed by the host VMM.

Typically, the host VMM then calls TDH.MEM.TRACK and performs a round of IPIs. After that, there should be no active address translation to target page.

The actual removal is done by TDH.MEM.PAGE.REMOVE which has the following inputs:

- The page GPA
- The page level (4KB, 2MB or 1GB)

At a high level, TDH.MEM.PAGE.REMOVE works as follows:

1. If TLB tracking is required, check the TLB tracking condition for the page.
2. Check that the mapping size of the page fits the input parameter.

If all checks pass, TDH.MEM.PAGE.REMOVE does the following:

3. Mark the Secure EPT entry for the target page as FREE.
4. Reclaim the physical page as described in 8.6.5.
    4.1. On platforms using ACT-protected memory, this is a long operation which may be interrupted. If interrupted, update the SEPT entry's state to REMOVE_IN_PROGRESS. In this state the physical page is still associated with the GPA, but it's not accessible by the guest.

Non-ACT Platforms: The host VMM should initialize the physical page's content before it is reused as a non-private page, as described in 8.6.2.

Refer also to the software flow described in 3.3.1.2.

9.16. Removing a Secure EPT Page: TDH.MEM.SEPT.REMOVE

The host VMM can remove a Secure EPT page using TDH.MEM.SEPT.REMOVE, freeing it for any use, provided all its entries are FREE.

If the guest TD's OP_STATE is such that the TD may be running, the host VMM should first perform the TLB tracking protocol on the GPA range mapped by the Secure EPT page. The host VMM should first call TDH.MEM.RANGE.BLOCK. TDH.MEM.RANGE.BLOCK marks the parent EPT entry for that page as **BLOCKED** and records the TD epoch in the PAMT entry of the page.

Dynamic PAMT: If the TDX module is configured for Dynamic PAMT, TDH.MEM.PAGE.REMOVE returns a hint indicating whether the PAMT pages mapping the removed SEPT pages can be removed. For details, see the [ABI Spec].

TDX Connect: If the guest TD has been configured with TDX Connect enabled, TDH.MEM.RANGE.BLOCK on the GPA range is allowed only if either no TDIs are attached or if all 512 entries of the applicable Secure EPT page are free (PT_NDA). Note that the second condition is required in any case for the subsequent TDH.MEM.SEPT.REMOVE, as described below.

Typically, the host VMM then calls TDH.MEM.TRACK and performs a round of IPIs. After that, there should be no active address translation to GPA range represented by the Secure EPT page to be removed.

The actual removal is done by TDH.MEM.SEPT.REMOVE which has the following inputs:

- The Secure EPT page GPA
- The Secure EPT level

At a high level, TDH.MEM.SEPT.REMOVE works as follows:

1. If TLB tracking is required, check the TLB tracking condition for the page.
2. Check that the mapping size of the page fits the input parameter.

3. Check that all 512 entries of the Secure EPT page are PT_NDA.

If all checks pass, TDH.MEM.SEPT.REMOVE does the following:

4. Reclaim the physical page as described in 8.6.5.
5. Mark the Secure EPT entry for the Secure EPT page as FREE.

Non-ACT Platforms: The host VMM should initialize the physical page's content before it is reused as a non-private page, as described in 8.6.2.

9.17. Unblocking a GPA Range: TDH.MEM.RANGE.UNBLOCK

The host VMM can unblock previously blocked TD private GPA ranges using TDH.MEM.RANGE.UNBLOCK, returning them to their original state. 4KB, 2MB, 1GB, 512GB or 256TB GPA ranges can be unblocked.

The host VMM should first complete the TLB tracking protocol on the GPA range. It typically calls TDH.MEM.TRACK and performs a round of IPIs. After that, there should be no active address translation to target page.

The actual unblocking is done by TDH.MEM.RANGE.UNBLOCK which has the following inputs:

- The GPA
- The GPA range level (4KB, 2MB, 512GB or 256TB)

At a high level, TDH.MEM.RANGE.UNBLOCK works as follows:

1. Check the TLB tracking condition for the GPA range.
2. Check that the mapping size of the GPA range fits the input parameter.

If all checks pass, TDH.MEM.RANGE.UNBLOCK does the following:

3. Mark the EPT entry for the target GPA as MAPPED (if it was BLOCKED) or PENDING (if it was PENDING_BLOCKED).

Refer also to the software flow described in 3.3.1.5.

10.TD VCPU

This chapter discusses multiple items related to TD VCPUs.

10.1. VCPU Transitions

This section describes transitions between running in the host VMM, TDX module and guest TD VCPU.

Note: Additional transitions that are applicable to partitioned TDs are discussed in the [TD Partitioning Spec].

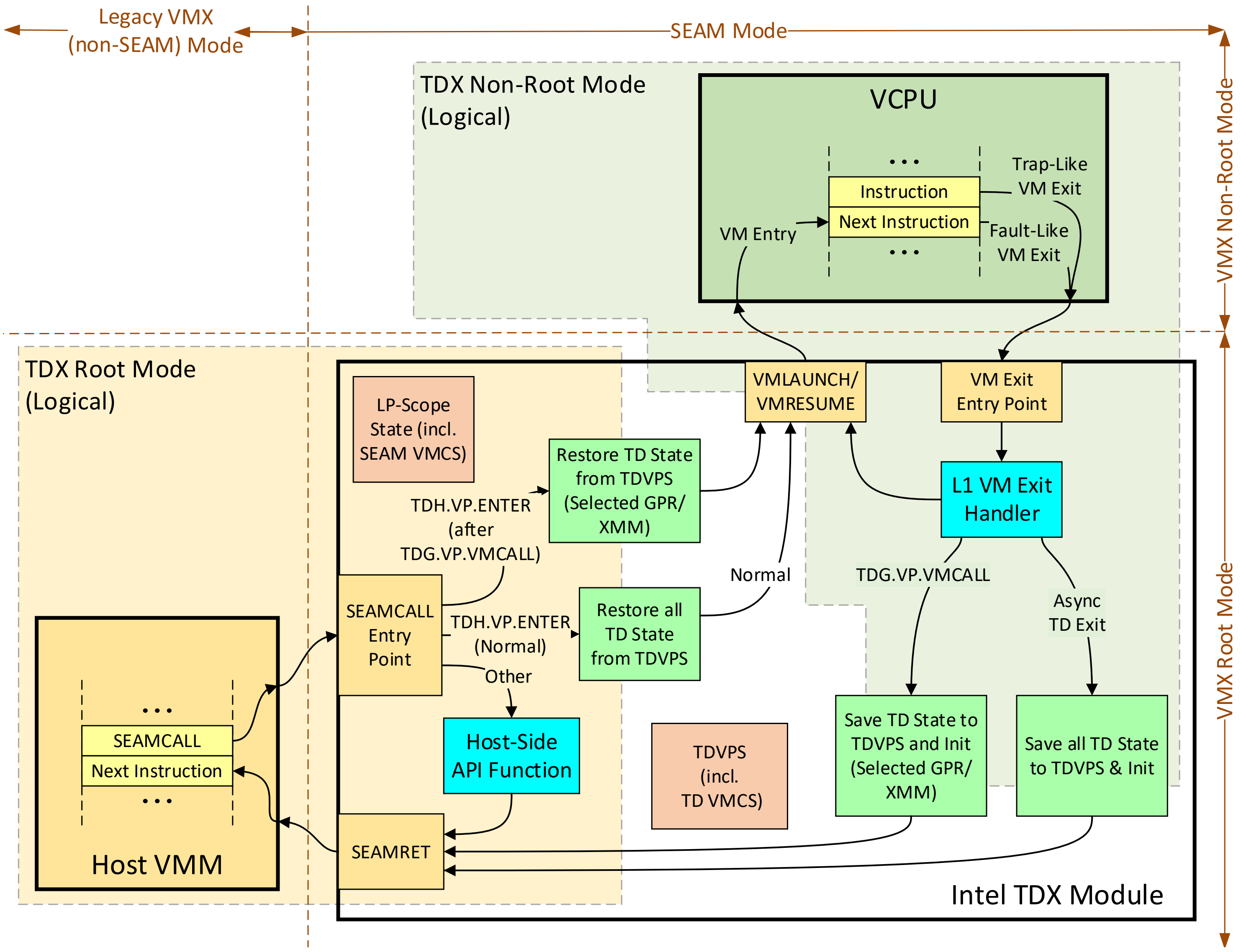


Figure 10.1: TD VCPU Transitions Overview

10.1.1. Initial TD Entry, Asynchronous TD Exit and Subsequent TD Entry

On the initial TD entry to a TD VCPU, the TDX module restores the initial TD VCPU state from TDVPS (including TD VMCS).

Following a successful TDH.VP.ENTER, asynchronous TD exit may happen as a result of events such as interrupts, EPT violations etc. In such case, the TDX module saves the TD VCPU state into TDVPS (including TD VMCS). Most of the host VMM VCPU state that may have been used by the TD is initialized. For a detailed description of VMM state following TDH.VP.ENTER, see the [TDX Module ABI Spec].

On the subsequent TD entry following an asynchronous TD exit, the TDX module restores the TD VCPU state from TDVPS (including TD VMCS). The host VMM does not impact the VCPU state except in one case: a trap-like asynchronous TD exit from a guest-side interface function may indicate that the host VMM can apply a recoverability hint in the following TD entry. In this case, the host VMM provides a recoverability hist to the guest TD, which is combined into the guest-side interface function's completion status returned in RAX.

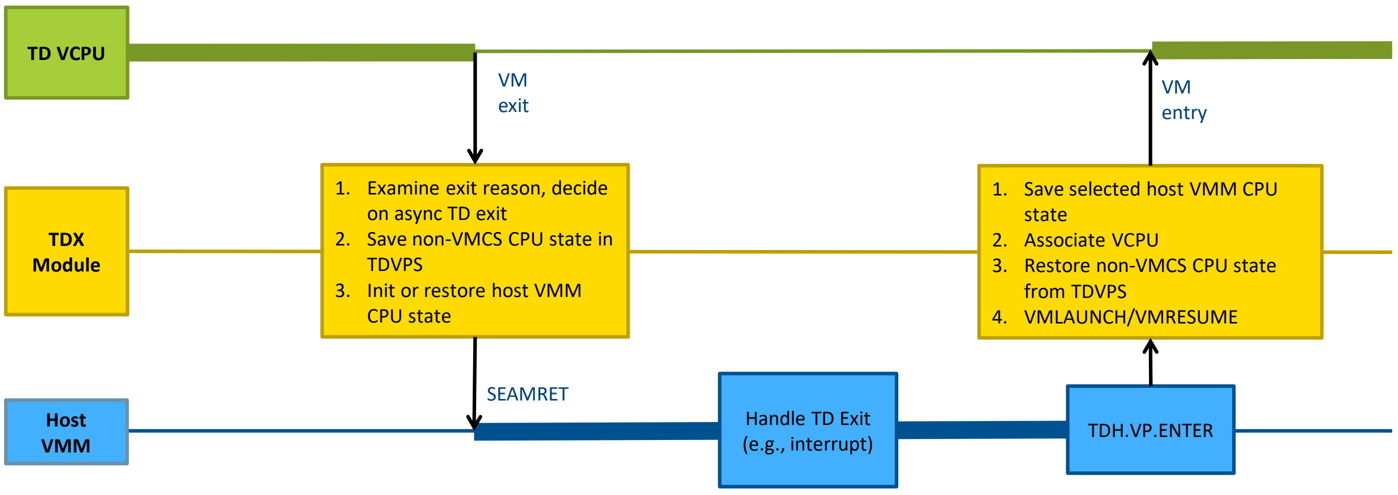

TD VCPU
TDX Module
Host VMM
VM exit
1. Examine exit reason, decide on async TD exit
2. Save non-VMCS CPU state in TDVPS
3. Init or restore host VMM CPU state
SEAMRET
Handle TD Exit (e.g., interrupt)
TDH.VP.ENTER
1. Save selected host VMM CPU state
2. Associate VCPU
3. Restore non-VMCS CPU state from TDVPS
4. VMLAUNCH/VMRESUME
VM entry

Figure 10.2: Example of Asynchronous TD Exit and TD Resumption

10.1.2. Synchronous TD Exit and Subsequent TD Entry

TDG.VP.VMCALL provides a channel for the guest TD to communicate with the host VMM.

The guest TD can initiate a synchronous TD exit by invoking TDG.VP.VMCALL. The RCX input parameter selects the GPRs (from RBX, RDX, RBP, RDI, RSI and R8 through R15) and XMM registers whose value is passed through to the host VMM as the output of TDH.VP.ENTER. RCX itself is passed as-is to the output of TDH.VP.ENTER. Other CPU state components, including GPRs and XMM registers not selected by RCX, are saved in TDVPS and set to fixed values.

On the subsequent TDH.VP.ENTER, the RCX value that was used for TDG.VP.VMCALL selects the GPRs (from RBX, RDX, RBP, RDI, RSI and R8 through R15) and XMM registers whose value is passed through to the guest TD. Other CPU state components, including GPRs and XMM registers not selected by RCX, are restored from RCX.

For details, see the TDH.VP.ENTER and TDG.VP.VMCALL definitions in the [TDX Module ABI Spec].

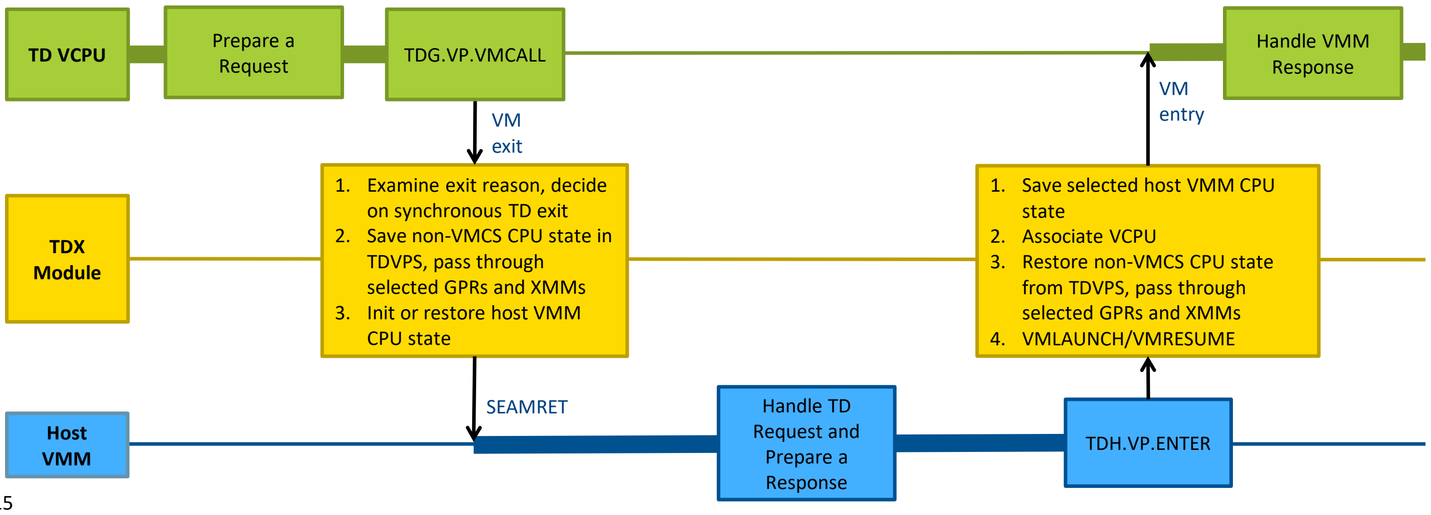

TD VCPU
Prepare a Request
TDG.VP.VMCALL
Handle VMM Response
TDX Module
Host VMM
VM exit
1. Examine exit reason, decide on synchronous TD exit
2. Save non-VMCS CPU state in TDVPS, pass through selected GPRs and XMMs
3. Init or restore host VMM CPU state
SEAMRET
Handle TD Request and Prepare a Response
TDH.VP.ENTER
1. Save selected host VMM CPU state
2. Associate VCPU
3. Restore non-VMCS CPU state from TDVPS, pass through selected GPRs and XMMs
4. VMLAUNCH/VMRESUME
VM entry

Figure 10.3: Example of Synchronous TD Exit and TD Resumption

10.1.3. VCPU Activity State Machine

The VCPU activity state machine, controlled by TDVPS.VCPU_STATE as shown in Table 10.1 below and shown in Figure 10.4 below, helps ensure the following:

- A VCPU can be entered only when its logical TDVPS control structure, composed of TDVPR and TDCX pages, is available in memory and has been initialized by TDH.VP.INIT or successfully imported by TDH.IMPORT.STATE.VP.
- A VCPU can be entered only if its state is consistent (no non-recoverable TD exit happened).
- TD entry is done properly, depending on whether it is the first entry or on the last TD exit type.

Table 10.1: TDVPS.VCPU_STATE Definition

State Name	Description
VCPU_UNINITIALIZED	VCPU has not been initialized yet by TDH.VP.INIT.
VCPU_IMPORT	The VCPU state has been incompletely imported.
VCPU_READY	The VCPU is ready to be executed.
VCPU_ACTIVE	VCPU is active (logical TDX non-root mode) on some LP. For a partitioned TD, a VCPU is considered active regardless of whether it executes in the L1 VM or one of the L2 VMs. For details, see the [TD Partitioning Spec].
VCPU_DISABLED	VCPU is being torn down.

TD Entry and TD Exit transitions normally toggle between the VCPU_READY state and the VCPU_ACTIVE state, except when a non-recoverable VCPU TD Exit (due to a Triple Fault) transitions to a VCPU_DISABLED state.

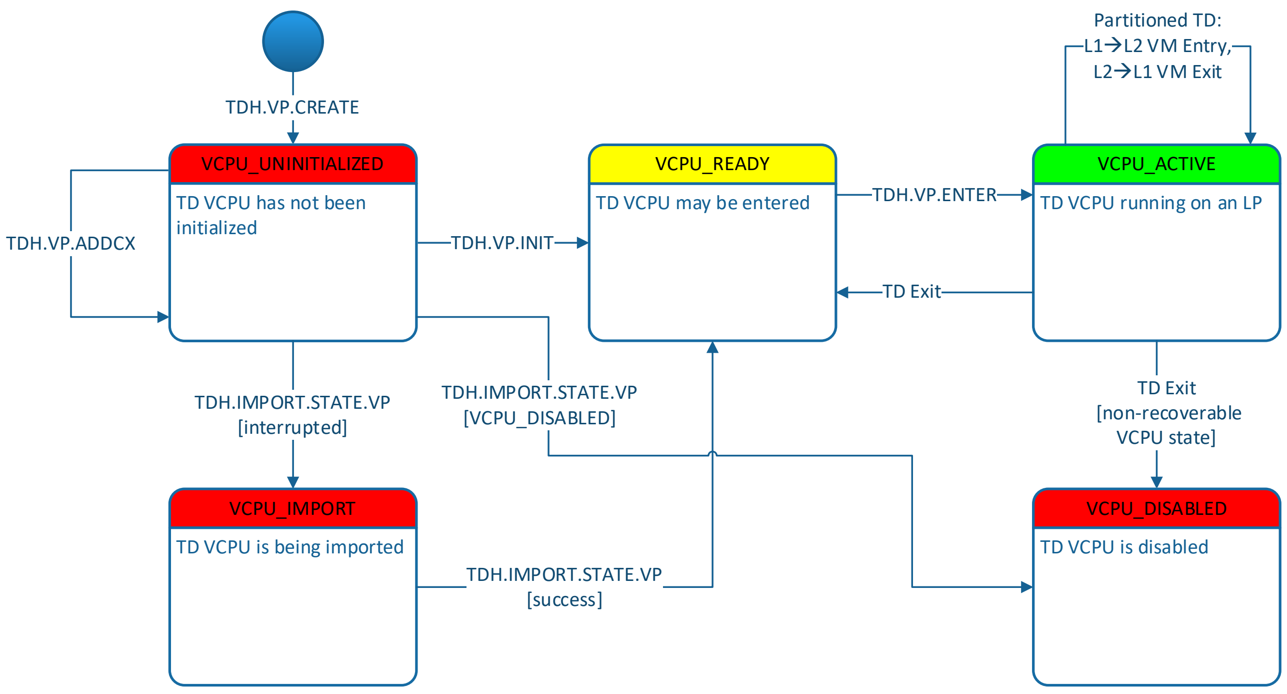


Figure 10.4: VCPU Activity State Machine

LAST_TD_EXIT

In the VCPU_READY and VCPU_IMPORT states, a LAST_TD_EXIT sub-state indicates what was the last TD exit and how a subsequent TD entry should be done.

Table 10.2: TDVPS.LAST_TD_EXIT Definition

Name	Description
ASYNC_FAULT	The last TD exit was due to an asynchronous event (non-TDG.VP.VMCALL) which caused a fault-like exit, i.e., the VCPU state is as if the guest instruction has not been executed. VCPU state has been fully saved on TD exit and will be restored on the next TD entry.
ASYNC_TRAP	The last TD exit was due to an asynchronous event that happened as part of a guest-side interface function (non-TDG.VP.VMCALL) which caused a trap-like exit, i.e., the VCPU state is as if the guest instruction has been executed. VCPU state has been fully saved on TD exit and will be restored on the next TD entry. On the next TD entry, the host VMM provides the guest with a recoverability hint.

Name	Description
TDVMCALL	The last TD exit was due to a TDG.VP.VMCALL. On the next TD entry, most GPR and all XMM state will be forwarded to the guest TD from the host VMM.

CURR_VM

For a partitioned TD, CURR_VM indicates the current VM index for the VCPU. For details, see the [TD Partitioning Spec].

10.2. TD VCPU TLB Address Space Identifier (ASID)

Non-root mode cached address translations are tagged with unique Address Space Identifiers (ASIDs). The goal of TD ASIDs is to reduce the need to flush TLB entries on TD Entry and TD Exit due to the associated performance costs as a result of the flushing.

10.2.1. TD ASID Components

Table 10.3 below shows a high-level view of the components of the TD ASID. The exact structure is micro-architectural.

Table 10.3: TD ASID

Field	Size (Bits)	Description and TDX Usage
SEAM	1	This is an implicit bit 16 of VPID not directly visible to software. It is set to 1 by the CPU in SEAM mode. This bit prevents overlapping with legacy (non-TDX) ASIDs.
VPID	16	The TDX module assigns a platform unique VPID for each TD. If a TD is partitioned, the TDX module assigns a platform unique VPID for each VM in that TD. See the [TD Partitioning Spec] for details.
EPTP	40	Bits [51:12] of the EPTP, which for a TD points to the **Secure** EPT root – HKID bits are cleared to 0 Note that EPTP is unique per TD and is used as an ASID component for **both Secure EPT and Shared EPT** translations caching.
PCID	16	Same as legacy PCID

Note: All VCPUs of the same TD share the same ASID. Consequently, whenever TDH.VP.ENTER is invoked on a certain LP, with a VCPU that is different than the last one that executed on that LP, the TDX module flushes cached TLB translations for the TD, using INVEPT.

10.2.2. INVEPT by the Host VMM for Managing the Shared EPT

The same ASID based on the TD's EPTP is used for caching both secure and shared EPT translations (remember: EPTP is the HPA of the **secure** EPT root page). Thus, to flush shared EPT translations, the host VMM uses INVEPT specifying the TD's EPTP, not its Shared EPTP. The host VMM can obtain the value of EPTP from the TD VMCSs using TDH.VP.RD.

If a TD is partitioned, then to flush shared EPT translations for each L2 VM, the host VMM uses INVEPT specifying that L2 VM's EPTP, not its Shared EPTP. The host VMM can obtain the value of EPTP from the L2 VMCSes using TDH.VP.RD.

An alternative method the host VMM may use is to do TLB tracking similar to how it's done for Secure EPT, i.e., execute TDH.MEM.TRACK and a round of IPI. Contrary to Secure EPT, this is not enforced by the TDX module.

10.3. VCPU-to-LP Association

10.3.1. Non-Coherent Caching

Some TD VCPU states are non-coherently cached. This includes:

- Address translations (TLB/PxE entries) must be explicitly flushed in case they may be stale.
- TD VMCS is cached by the CPU. VMX architecture requires making a VMCS current by VMPTRLD before using it with most VMX instructions, and then explicitly writing it to memory and making it non-current by VMCLEAR before the VMCS memory image can be handled (e.g., by making it current on another LP).

This non-coherent caching implies that some explicit and/or implicit operations are done to help guarantee correctness. This is described in the following sections.

10.3.2. Intel TDX Functions for VCPU-LP Association and Dis-Association

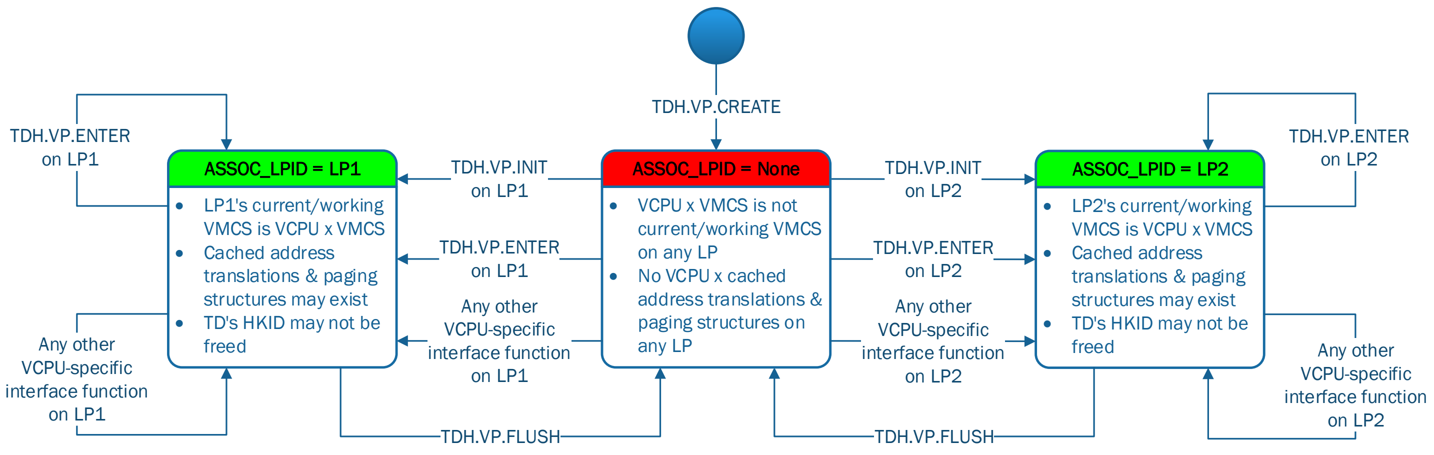

TDH.VP.CREATE
TDH.VP.ENTER on LP1
ASSOC_LPID = LP1
LP1's current/working VMCS is VCPU x VMCS
Cached address translations & paging structures may exist
TD's HKID may not be freed
Any other VCPU-specific interface function on LP1
TDH.VP.INIT on LP1
TDH.VP.ENTER on LP1
Any other VCPU-specific interface function on LP1
ASSOC_LPID = None
VCPU x VMCS is not current/working VMCS on any LP
No VCPU x cached address translations & paging structures on any LP
TDH.VP.INIT on LP2
TDH.VP.ENTER on LP2
Any other VCPU-specific interface function on LP2
ASSOC_LPID = LP2
LP2's current/working VMCS is VCPU x VMCS
Cached address translations & paging structures may exist
TD's HKID may not be freed
TDH.VP.ENTER on LP2
Any other VCPU-specific interface function on LP2
TDH.VP.FLUSH
TDH.VP.FLUSH

Figure 10.5: VCPU Association State Machine

The following Intel TDX module mechanisms are designed to help ensure correct and secure operation:

- TD VCPU to LP association is many-to-one. A TD VCPU can be associated with at most one LP at any given time. An LP may be associated with multiple VCPUs.
- VCPU to LP association is implicitly done by any VCPU-specific SEAMCALL flow, including TDH.VP.ENTER. Those flows check that the VCPU is either already associated with the current LP or is not associated with any LP.
- If the host VMM wishes to associate a VCPU with another LP, it must explicitly flush the VCPU state on the LP currently associated with it using **TDH.VP.FLUSH**. This function flushes TLB for the TD ASID and extended paging structure (EPxE) caches using INVEPT. It flushes the VMCS cache using VMCLEAR. For details, see the [TDX Module ABI Spec].
- If the VMM wishes to reclaim the TD's private HKID, thus making the TDVPS memory inaccessible, it must explicitly flush the VCPU state on the LP currently associated with it. This is described in 5.5.

10.3.3. Performance Considerations

- Migrating VCPUs between LPs is costly. As described above, it involves flushing address translation caches, paging structure caches and VMCS cache. The host VMM should minimize that for best performance.
- Address translation and paging structure caches are flushed at TD-scope on the current LP. This flushing impacts the (possibly non-typical) case where multiple VCPUs of the same TD are associated with a single LP.

11. CPU Virtualization (Non-Root Mode Operation)

This chapter describes how the Intel TDX module virtualizes the CPU to a guest TD.

11.1. Overview: Virtualization vs. Paravirtualization of CPU Features and #VE

This section provides an overview of #VE and paravirtualization. The mechanics of #VE are described in 11.14.

11.1.1. Architectural x86 Virtualization

In most cases, TDX is designed to emulate the x86 architectural behavior of the CPU. Many CPU features are configurable by the host VMM as either available (if the CPU actually supports them) or unavailable. That configuration is reflected to the guest TD by values returned by the CPUID and RDMSR instructions, and the emulated CPU behavior is designed to reflect the architectural CPU behavior – e.g., inject a #GP(0) exception on WRMSR of an MSR that is enumerated by the virtual CPUID values as non-existent.

11.1.2. Paravirtualization and #VE

In many cases, TDX can't emulate some x86 functionality. In such cases, the guest TD may implement that functionality using paravirtualization, possibly in cooperation with the host VMM. To do that, the guest TD is required to implement a **paravirtualization agent** as part of a **#VE exception handler**. On execution of, e.g., an instruction that can't be emulated by the TDX module, a #VE exception is injected to the guest TD. The #VE paravirtualization agent may request information from the host VMM (noting that this information is untrusted) and/or emulate the desired behavior.

If the TDX module supports #VE reduction, then the guest TD may configure CPU virtualization to greatly reduce the number of cases where #VE is injected and needs to be handled by the TD's #VE handler. The guest TD can control this, if desired, for each of multiple CPU features. For details, see 11.2.2 and the [ABI Spec].

11.1.3. #VE for x86 Behavior not Supported by TDX

TDX imposes some restrictions that are not x86-architectural. In such cases, if the guest TD attempts to use a restricted feature, there is no architectural way (such as #GP(0)) to notify it, thus #VE is injected. This typically indicates a guest TD misbehavior.

For example, a TD is not allowed to run in 32-bit protected mode with paging. When a guest TD attempts to set CR0.PE or IA32_EFER.LME to 0, the TDX module injects a #VE.

11.1.4. #VE for TDX-Specific Behavior

#VE is also used by the TDX module to alert the guest TD to some TDX-specific cases, such as an access attempt to a PENDING page. See 9.10.4 for details.

11.2. CPU Virtualization Configuration and Control

11.2.1. Host VMM Configuration of CPU Virtualization

The host VMM configures CPU virtualization at TD initialization time (TDH.MNG.INIT). The table below shows the configurable parameters. For details, see the [ABI Spec] definition of TDH.MNG.INIT and TD_PARAMS.

Table 11.1: Host Configuration of CPU Virtualization

Name	Description	Included in the TD's Attestation
ATTRIBUTES	TD attributes: a bitmap of configurable TD attributes	Yes
XFAM	Extended Features Available Mask: indicates the extended state features allowed for the TD. See 11.8 for details of extended features virtualization.	Yes
Non-Attested Configuration	A set of TD configuration parameters	No

Name	Description	Included in the TD's Attestation
MSR Configuration	Configuration of specific MSRs' virtual values	No
CPUID Configuration	Direct configuration of CPUID leaves/sub-leaves virtualization. See 11.11 for a discussion of CPUID virtualization.	No

11.2.2. Guest TD Control of CPU Virtualization

The guest TD can control some aspects of CPU virtualization, by setting bits in the TDCS.TD_CTLS field using TDG.VM.WR. The table below lists the configurable virtualization features. In some cases, this overrides the host VMM's configuration done at TD initialization time. Availability of each control bit depends on the features supported by the TDX module. For more details, see the definition of TDCS.TD_CTLS in the [ABI Spec].

Table 11.2: Guest TD Control of CPU Virtualization

Name	Description
PENDING_VE_DISABLE	Controls the way guest TD access to a PENDING page is processed
ENUM_TOPOLOGY	Controls the enumeration of virtual platform topology
VIRT_CPUID2	Controls the virtualization of CPUID(2)
REDUCE_VE	Allows the guest TD to control the way #VE is injected by the TDX module on guest TD execution of CPUID, RDMSR/WRMSR and other instructions
LOCK	Controls locking of TD-writable virtualization controls

If the TDX module supports #VE reduction, the guest TD can individually control the virtualization of some CPU features, by setting bits in the TDCS.FEATURE_PARAVIRT_CTRL field using TDG.VM.WR. The guest TD can configure each feature to be emulated as non-supported, or as supported by a paravirtualization agent implemented by the TD (as part of its #VE handler). The table below lists the configurable CPU features. For more details, see the definition of TDCS.FEATURE_PARAVIRT_CTRL in the [ABI Spec].

Table 11.3: Configurable Paravirtualized CPU Feature

Paravirtualized Feature Name & Applicable Linux Kernel Feature Name	Description
CORE_CAPABILITIES (X86_FEATURE_CORE_CAPABILITIES)	Controls IA32_CORE_CAPABILITIES paravirtualization, enumerated by virtual CPUID(7,0).EDX[30] (support IA32_CORE_CAPABILITIES)
DCA (X86_FEATURE_DCA)	Controls Direct Cache Access paravirtualization, enumerated by virtual CPUID(1).ECX[18] (DCA)
EST (X86_FEATURE_EST)	Controls Enhanced Intel SpeedStep technology paravirtualization, enumerated by Virtual CPUID(1).ECX[7] (Enhanced Intel SpeedStep technology)
MCA (X86_FEATURE_MCA)	Controls Machine Check Architecture paravirtualization, enumerated by virtual CPUID(1).EDX[7] (Machine Check Exception) and virtual CPUID(1).EDX[14] (Machine Check Architecture)
MTRR (X86_FEATURE_MTRR)	Controls Memory Type Range Registers paravirtualization, enumerated by virtual CPUID(1).EDX[12] (Memory Type Range Registers)
PCONFIG (X86_FEATURE_PCONFIG)	Controls PCONFIG paravirtualization, enumerated by virtual CPUID(7,0).EDX[18] (PCONFIG)

Paravirtualized Feature Name & Applicable Linux Kernel Feature Name	Description
RDT_A (X86_FEATURE_RDT_A)	Controls RDT-A paravirtualization, enumerated by virtual CPUID(7,0).EBX[15] (RDT-A)
RDT_M (X86_FEATURE_CQM)	Controls RDT-M paravirtualization, enumerated by virtual CPUID(7,0).EBX[12] (RDT-M)
ACPI (X86_FEATURE_ACPI)	Controls Thermal Monitor and Software Controlled Clock Facilities paravirtualization, enumerated by virtual CPUID(1).EDX[22] (ACPI)
TM2 (X86_FEATURE_TM2)	Controls MSR_THERM2_CTL paravirtualization, enumerated by virtual CPUID(1).ECX[8] (TM2)
TME (X86_FEATURE_TME)	Controls Total Memory Encryption paravirtualization, enumerated by virtual CPUID(7,0).ECX[13] (TME_EN)
TSC_DEADLINE (X86_FEATURE_TSC_DEADLINE_TIMER)	Controls IA32_TSC_DEADLINE MSR paravirtualization, enumerated by virtual CPUID(1).ECX[24] (TSC deadline)

The guest TD can also control CPUID virtualization for each VCPU and CPUID leaf/sub-leaf. For details, see 11.11.2.

11.3. Initial Virtual CPU State

Intel SDM, Vol. 3, 10.1.1 Processor State after Reset

11.3.1. Overview

As designed, most of the TD VCPU initial state is the same as the processor architectural state after INIT. However, there are some differences:

- The TD VCPU starts its life in protected (32-bit) non-paged mode, not in real mode. It is allowed only to switch to 64b mode. This impacts the initial state of segment registers, CRs and MSRs. Mode restrictions in SEAM non-root mode are described below.
- The IA32_EFER MSR is initialized to support the CPU modes described below.
- The initial values of some GPRs provide some basic information to the guest TD as described in 11.3.2 below. This information should be sufficient for the vBIOS to set up paging tables and switch as soon as possible to 64b mode, where it can use the TDCALL leaf functions.

See also the TDVPS fields and TD VMCS guest state area in the [TDX Module ABI Spec].

11.3.2. Initial State of Guest TD GPRs

As designed, the following initial state is different than the architectural INIT state:

Table 11.4: Initial Values of GPRs Different from their Architectural INIT Values

Register	Bits	Initial Value
RBX	5:0	GPAW, the effective GPA width (in bits) for this TD (do not confuse with MAXPA) – SHARED bit is at GPA bit GPAW-1 Only GPAW values 48 and 52 are possible.
	63:6	Reserved: set to 0
RCX, R8	63:0	The value of RCX and R8 is provided as an input to TDH.VP.INIT (the same value in both GPRs). No checking is done on this value; the intention is for vBIOS to read RCX immediately after the first TDH.VP.ENTER and use the RCX value appropriately as set by software convention.

Register	Bits	Initial Value
RDX	31:0	Set to the virtualized Family/Model/Stepping returned by CPUID(1).EAX. The value is calculated by TDH.SYS.INIT as to have the minimum Stepping ID across all packages.
	63:32	Reserved: set to 0
RSI	31:0	Virtual CPU index, starting from 0 and allocated sequentially on each successful TDH.VP.INIT
	63:32	Reserved: set to 0
RIP	63:0	Set to 0xFFFFFFF0 (i.e., 4GB - 16B)

11.3.3. Initial State of CRs

As designed, the following initial state is different than the architectural INIT state:

- Virtual CR0 is initialized to 0x0021 – bits PE (0) and NE (5) are set to 1, and all other bits are cleared to 0. See 11.9.1 for details.
- Virtual CR4 is initialized to 0x0040 – bits MCE (6) is set to 1, and all other bits are cleared to 0.

11.3.4. Initial State of Segment Registers

As designed, the following initial state is different than the architectural INIT state:

- CS, DS, ES, FS, GS and SS are initialized with a base of 0 and limit of 0xFFFFFFFF.
- LDTR, TR and GDTR are initialized with a base of 0 and limit of 0xFFFF.
- IDTR is initialized as invalid (limit of 0).

For details, see the [TDX Module ABI Spec].

11.3.5. Initial State of MSRs

As designed, the following initial state is different than the architectural INIT state:

- IA32_EFER is initialized to 0x901 – SCE (bit 0), LME (bit 8) and NXE (bit 11) are set to 1, and all other bits are cleared to 0.

11.4. Guest TD Run Time Environment Enumeration

Guest software can be designed to run either as a TD, as a legacy virtual machine, or directly on the CPU, based on enumeration of its run-time environment. Figure 11.1 below shows a typical flow used by guest software.

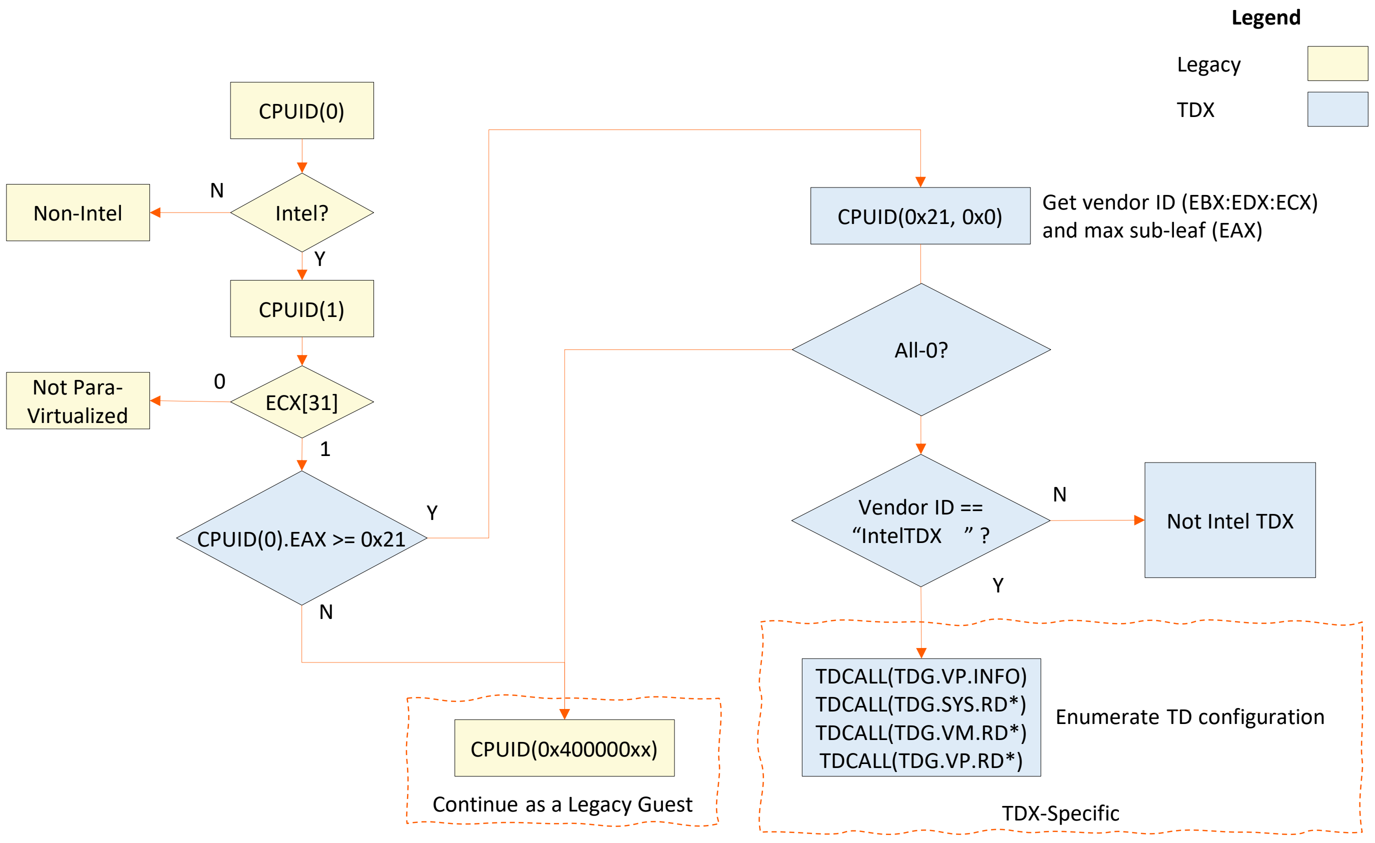


Figure 11.1: Typical Run-Time Environment Enumeration by a Guest TD

CPUID leaf 0x21 emulation is done by the Intel TDX module. Sub-leaf 0 returns the values shown below. Other sub-leaves return 0 in EAX/EBX/ECX/EDX.

Table 11.5: TDX Enumeration by CPUID(0x21,0)

GPR	Value (Hex)	Description
EAX	0x00000000	Maximum sub-leaf number
EBX	0x65746E49	"Inte"
ECX	0x20202020	" "
EDX	0x5844546C	"lTDX"

Once the guest software discovers that it runs as a TD, it can use TDG.VP.INFO to get basic information. It can also use the metadata read functions TDG.SYS.RD*, TDG.VM.RD* and TDG.VP.RD*.

11.5. Uniform VM Virtualization on a Hybrid SOC

Hybrid SOCs contain more than one core type. Some CPU features may be different between core types. However, guest TDs are always virtualized as **uniform VMs**; all VCPUs have the same virtual CPU behavior.

The CPU features available to a guest TD will be, at maximum, the subset of features that are available across all cores on the platform. Main impact is on Perfmon virtualization; for details, see 15.2. Other details are discussed in the following sections.

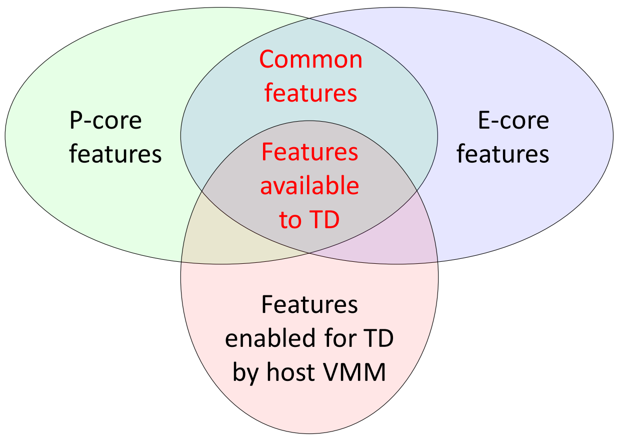


Figure 11.2: Virtualizing a Hybrid SOC as a Uniform TD VM

11.6. CPU Mode Restrictions

Intel SDM, Vol. 3, 2.2	Modes of Operation
Intel SDM, Vol. 3, 9.8.5	Initializing IA-32e Mode
Intel SDM, Vol. 3, 11.5.1	Cache Control Registers and Bits
Intel SDM, Vol. 3, 24.6.6	Guest/Host Masks and Read Shadows for CR0 and CR4

A TD OS running in SEAM non-root mode is required to be a 64-bit OS. The Intel TDX module helps enforce this with the restrictions described below.

Table 11.6: CPU Mode Restrictions in SEAM Non-Root Mode

Restriction	Description
CPU and Paging Modes	In SEAM non-root mode, the CPU is allowed to run in the following modes: • Protected mode (32-bit) with no paging (CR0.PG == 0) • IA-32e mode with 4-level or 5-level paging (CR0.PG == 1), with the sub-modes controlled by CS.L: ○ 64-bit mode ○ Compatibility (32-bit) mode To achieve this, CR0.PE and IA32_EFER.LME are enforced to 1, as described in the following sections.
Execute Disable	When running in IA-32e mode, the PT Execute Disable bit (63) is always enabled. To achieve this, IA32_EFER.NXE is enforced to 1, as described in the following sections.
Caching is Always Enabled	The guest TD runs in Normal Cache Mode. To achieve this, CR0.CD and CR0.NW are enforced to 0, as described in the following sections.

11.7. Instructions Restrictions

11.7.1.1.

The Intel TDX module is designed to block certain instructions from executing in TDX non-root mode. Execution of those instructions results in a VM exit to the Intel TDX module, which then injects an exception to the guest TD. This exception can be #UD, a #GP(0) or, in case where no Intel64 architectural exception can be used, a #VE (described in 11.14).

11.7.1. Unconditionally Blocked Instructions

Instructions that Cause a #UD Unconditionally

- ENCLS, ENCLV

- Most VMX instructions: INVEPT, INVVPID, VMCLEAR, VMFUNC, VMLAUNCH, VMPTRLD, VMPTRST, VMREAD, VMRESUME, VMWRITE, VMXOFF, VMXON
- RSM
- GETSEC
- SEAMCALL, SEAMRET

Instructions that Cause a #VE Unconditionally to Allow Paravirtualization

Guest TD (L1) execution of the following instructions always results in a #VE(NON_CONFIG_PARAVIRT). A #VE handler is expected to paravirtualize the instruction.

Table 11.7: Instructions that Cause a #VE Unconditionally

11.7.

Instruction	Details
String I/O (INS*, OUTS*), IN, OUT	There is no standard x86 way to enumerate as unsupported. If used by any guest, then paravirtualization is necessary.
HLT	There is no standard x86 way to enumerate as unsupported. If used by any guest, then paravirtualization is necessary.
WBINVD	There is no standard x86 way to enumerate as unsupported. If used by any guest, then paravirtualization is necessary.
WBNOINVD	Availability of WBNOINVD is enumerated by virtual CPUID(0x80000008).EBX[9], which is host configurable. However, TDX never allows WBNOINVD to be executed by guest TDs. WBNOINVD causes #VE regardless of virtual CPUID(0x80000008).EBX[9], to allow paravirtualization similar to WBINVD.
INVD	There is no standard x86 way to enumerate as unsupported. If used by any guest, then paravirtualization is necessary.
VMCALL	There is no standard x86 way to enumerate as unsupported. If used by any guest, then paravirtualization is necessary.

11.7.1.3. *Instructions that Cause a #UD or #VE Depending on Feature Enabling, to Allow Paravirtualization*

- PCONFIG (see 11.23)

11.7.1.4. *Other Cases of Unconditionally Blocked Instructions*

- Guest TD execution of ENQCMD results in a #GP(0).
- Guest TD execution of ENQCMDS when CPL is 0 results in a #UD. Otherwise, it results in a #GP(0).

11.7.2. Conditionally Blocked Instructions

Execution of some instructions may be conditionally blocked, depending on which CPU features are configured and available for the TD, as described in the following sections.

11.7.3. Other Exception Cases

In many cases, instructions are not blocked but yet may cause exceptions due to other conditions. For example, the following is a very partial list:

- CPUID may cause a #VE if the CPUID leaf and sub-leaf are not virtualized by the TDX module.
- RDMSR and WRMSR may cause a #GP(0) if an MSR is virtualized as non-existent, or a #VE if an MSR is not virtualized.

11.8. Extended Feature Set

Intel SDM, Vol. 1, 13	Managing State Using the XSAVE Feature Set
Intel SDM, Vol. 3, 13	System Programming for Instruction Set Extensions and Processor Extended State

11.8.1. Allowed Extended Features Control

At the guest TD scope, **TDCS.XFAM (Extended Features Allowed Mask)** is provided as an input during the guest TD build process. XFAM is a 64b mask, using the **state-component bitmap** format used by extended state ISA (XSAVE, XRSTOR, XCR0, IA32_XSS etc.), which specifies the set of extended features the TD is allowed to use.

XFAM is checked to be compliant with the set of extended features supported by the CPU, as enumerated by CPUID and the allowed bit combinations, as shown in Table 11.9 below.

11.8.2. Extended State Isolation

The Intel TDX module helps ensure that any guest TD extended state is saved and isolated from the host VMM across TD exit and entry. It is the VMM's responsibility to save its own extended state across TD entry and exit.

- Before TDH.VP.ENTER, the host VMM should save (e.g., using XSAVES) any extended state that the guest TD VCPU is allowed to use (per XFAM) and the host VMM expects to need after TDH.VP.ENTER is complete.
- The TDH.VP.ENTER function loads the extended state that the TD VCPU is allowed to use, per XFAM, from the VCPU's TDVPS. An exception to this is when TDH.VP.ENTER follows a previous TDG.VP.VMCALL – in the case TDH.VP.ENTER does not load the XMM state (corresponding to XFAM bit 1) from TDVPS; it passes it directly from the host VMM.
- On an asynchronous TD exit, the Intel TDX module saves the extended state that the TD VCPU was allowed to use, per XFAM, to the VCPU's TDVPS. It then clears the extended state.
- On TDG.VP.VMCALL, the Intel TDX module works similarly, but it selectively does not clear some of the XMM register state (corresponding to XFAM bit 1). That XMM state is passed directly to the host VMM.
- On completion of TDH.VP.ENTER (following TD exit), the VMM may restore any extended state that it saved before TDH.VP.ENTER.

11.8.3. Extended Features Execution Control

The Intel TDX module is designed to prohibit the guest TD from using any extended feature not allowed by XFAM. Many extended state features are controlled by XCR0 and IA32_XSS MSR. Other features are controlled by CR4 or by specific MSRs.

Table 11.8: XFAM Interaction with XCR0, IA32_XSS, CR4 and Other MSRs

Register	Description
XCR0 and IA32_XSS MSR	On XSETBV, which attempts to write to XCR0, and on WRMSR of IA32_XSS, the TDX module emulates the architectural behavior of the CPU. The following cases cause a #GP(0): • The new value is not natively valid for XCR0 or IA32_XSS (it sets reserved bits, sets bits for features not recognized by the Intel TDX module, or uses invalid bit combinations). • The new value has any bits set that are not allowed by XFAM.
CR4	On MOV to CR4, the guest TD attempts to set bits not allowed according to XFAM will cause a #GP(0).
Other MSRs	The guest TD attempts to write or read certain MSRs that are not enabled according to XFAM can cause a #GP(0) or a #VE, as described below.

The following table describes how a guest TD executes each of the extended features.

Table 11.9: Extended Features Enumeration and Execution Control

Bits	U/S	Feature	Enumeration[7]	XFAM Value	Description
0	U	FP	Always available	1	Always enabled
1	U	SSE	Always available	1	Always enabled
2	U	AVX	CPUID(0xD, 0x0).EAX[2] CPUID(0x7, 0x0).EBX[2] CPUID(0x7, 0x0).ECX[10:9] CPUID(0x7, 0x1).EAX[5] CPUID(0xD, 0x2).* Specific AVX instructions support is enumerated by other CPUID bits.	0 or 1	Execution is directly controlled by XCR0.
4:3	U	MPX	CPUID(0xD, 0x0).EAX[4:3] CPUID(0x7, 0x0).EBX[14] CPUID(0xD, 0x3).* CPUID(0xD, 0x4).*	00	MPX is being deprecated.
7:5	U	AVX512	CPUID(0xD, 0x0).EAX[7:5] CPUID(0x7, 0x0).EBX[31:30, 28:26, 21, 17:16] CPUID(0x7, 0x0).ECX[14, 12:11, 6, 1] CPUID(0x7, 0x0).EDX[8] CPUID(0x7, 0x1).EAX[5] CPUID(0xD, 0x5).* CPUID(0xD, 0x6).* CPUID(0xD, 0x7).* Specific AVX512 instructions support is enumerated by other CPUID bits.	000 or 111	Execution is directly controlled by XCR0. AVX512 may be enabled only if AVX is enabled – i.e., XFAM[7:5] may be set to 111 only when XFAM[2] is set to 1.
8	S	PT (RTIT)	CPUID(0xD, 0x1).ECX[8] CPUID(0x7, 0x0).EBX[25] CPUID(0x14,*).* CPUID(0xD, 0x8).*	0 or 1	Execution is controlled by IA32_RTIT_CTL. If PT is enabled by XFAM, the guest TD is allowed access to all IA32_RTIT_* MSRs. Otherwise, any access causes #GP(0).
9	U	PK	CPUID(0xD, 0x0).EAX[9] CPUID(0xD, 0x9).*	0 or 1	Execution is controlled by CR4.PKE (bit 22). If PK is disabled by XFAM, the guest TD is disallowed from setting CR4.PKE. An attempt to set this bit causes a #GP(0).
10	S	ENQCMD (PASID)	CPUID(0xD, 0x1).ECX[10] CPUID(0xD, 0xA).*	0	Execution is controlled by IA32_PASID MSR. There is no direct I/O from guest TDs. ENQCMD and ENQCMDS from the guest TD are not supported and cause a #UD or #GP(0) (see 11.7.1.4). Access to IA32_PASID causes a #GP(0).

[7] An extended feature controlled by bits N:M is available if all bits in the range N:M returned by CPUID are set to 1.

Bits	U/S	Feature	Enumeration[7]	XFAM Value	Description
12:11	S	CET	CPUID(0xD, 0x1).ECX[12:11] CPUID(0xD, 0xB).* CPUID(0xD, 0xC).*	00 or 11	Execution is controlled by CR4.CET (bit 23). If CET is disabled by XFAM, the guest TD is disallowed from setting CR4.CET. An attempt to set this bit causes a #GP(0).
13	S	HDC	CPUID(0xD, 0x1).ECX[13] CPUID(0xD, 0xD).*	0	Hardware Duty Cycle is controlled by package-scope IA32_PKG_HDC_CTL and LP-scope IA32_PM_CTL1 MSRs. HDC is disabled. If the TDX module supports #VE reduction and the guest TD has set TDCS.TD_CTLS.REDUCE_VE, guest TD access to the above MSRs causes a #GP(0). Else, it causes #VE(CONFIG_PARAVIRT).
14	S	ULI	CPUID(0xD, 0x1).ECX[14] CPUID(0x7, 0x0).EDX[5] CPUID(0xD, 0xE).*	0 or 1	Execution is controlled by CR4.UINTR (bit 25). If ULI is disabled by XFAM, then the guest TD is not allowed the following: • Setting CR4.ULI. An attempt to set this bit causes a #GP(0). • Access to all IA32_UINTR_* MSRs. Any access causes a #GP(0).
15	S	LBR	CPUID(0xD, 0x1).ECX[15] CPUID(0x7, 0x0).EDX[19] CPUID(0xD, 0xF).* CPUID(0x1C).*	0 or 1	Execution is controlled by IA32_LBR_CTL. If LBR is disabled by XFAM, the guest TD is not allowed the following: • Access to all IA32_LBR_* MSRs. Any access causes a #GP(0). • Setting EN_LBR_LOG (bit 11) in IA32_PERFEVTSELx MSRs. Setting this bit causes #GP(0).
16	S	HWP	CPUID(0xD, 0x1).ECX[16] CPUID(0xD, 0x10).*	0	Execution of Hardware-Controlled Performance State is controlled by IA32_HWP MSRs. This feature is disabled. If the TDX module supports #VE reduction and the guest TD has set TDCS.TD_CTLS.REDUCE_VE, guest TD access to the above MSRs causes a #GP(0). Else, it causes #VE(CONFIG_PARAVIRT).
18:17	U	AMX	CPUID(0xD, 0x0).EAX[18:17] CPUID(0xD, 0x11).* CPUID(0xD, 0x12).* Specific AMX instructions support is enumerated by other CPUID bits.	00 or 11	Advanced Matrix Extensions (AMX) is directly controlled by XCR0.
19	U	APX	CPUID(0x7, 0x1).EDX[21] CPUID(0xD,0).EAX[19]	0 or 1	Execution is controlled by XCR0.APX (bit 19). If APX is disabled by XFAM, the guest TD is disallowed from setting XCR0.APX. An attempt to set this bit causes a #GP(0).
Other	N/A	RESERVED	N/A	0	Reserved

11.8.4. AVX10.2 (Converged Vector ISA)

Converged Vector ISA (AVX10.2) is enumerated by CPUID(0x7, 0x1).EDX[19]. If the TDX module supports AVX10.2, it enumerates the virtual value of that CPUID bit as configurable. The host VMM can configure AVX10.2 support and enumeration, as part of the CPUID configuration during TDH.MNG.INIT.

If the host VMM configures CPUID(0x7, 0x1).EDX[19] as 1, and the CPU supports AVX10.2, then the virtual value of CPUID(0x7, 0x1).EDX[19] is 1:

- AVX10.2 is available to the guest TD.
- Virtual CPUID AVX10.2 leaf CPUID(0x24, 0x0) is enumerated to the guest TD.
- The host VMM can configure the virtual values of AVX10.2 vector width support in CPUID(0x24,0x0).EBX[18:16].

Vector-Extension Packed Matrix Multiplication (VPMM)

VPMM is enumerated by CPUID(0x24, 0x1).ECX[0]. If the TDX module supports VPMM, it enumerates the virtual value of that CPUID bit as configurable. The host VMM can then configure VPMM support and enumeration, as part of the CPUID configuration during TDH.MNG.INIT.

11.8.4.1.

If the virtual value of CPUID(0x24, 0x1).ECX[0] is 1, then VPMM is available to the guest TD.

11.8.5. APX

Enumeration: TDX module support of APX is enumerated by TDX_FEATURES0.APX (bit 28), readable by TDH.SYS.RD*.

Guest TD usage of Advanced Performance Extension (APX) is controlled by the XFAM[19] (APX) bit (see the [TDX Module ABI Spec]). If supported by the TDX module and by the CPU, and the host VMM configured XFAM[19] as 1, then:

- Virtual CPUID enumerates APX availability to the guest TD.
- Guest TDs may enable APX by setting XCR0[19] (APX) bit.
- TDG.VP.VEINFO.GET called with version 2 returns the VM Exit Extended Instruction Information.

11.9. CR Handling

11.9.1. CR0

Intel SDM, Vol. 3, 2.5	Control Registers
Intel SDM, Vol. 3, 23.8	Restrictions on VMX Operation
Intel SDM, Vol. 3, 24.6.6	Guest/Host Masks and Read Shadows for CR0 and CR4
Intel SDM, Vol. 3, 25.6	Unrestricted Guests

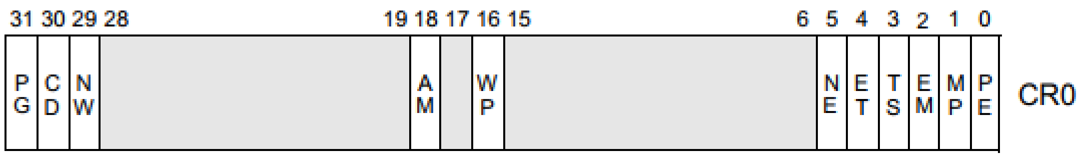


Figure 11.3: CR0

From the guest TD's point of view, as virtualized by the Intel TDX module, CR0 bits PE (0) and NE (5) are always set to 1, and bits NW (29) and CD (30) are always cleared to 0.

Guest TD writes to CR0 are handled by the Intel TDX module as follows:

- Writes to CR0 that are architecturally invalid (such as attempts to set bits that must be 0) or writes to CR0 that set architecturally invalid bit combinations, result in a #GP(0).
- Writes to CR0 that are architecturally invalid but not permitted by the TDX architecture (such as clearing CR0.CD) result in a #VE(UNSUPPORTED_FEATURE).
- Other writings are allowed.

For TD migration, the same rules are used for checking the imported value of guest CR0. Any violation results in a failed import.

11.9.2. CR4

Intel SDM, Vol. 3, 24.6.6 Guest/Host Masks and Read Shadows for CR0 and CR4

11.9.2.1. *CR4 Bits which are Architecturally Virtualized*

For most CR4 bits, the TDX module emulates the x86 architectural behavior of the CPU. If a certain CPU feature is not enabled for the guest TD, the guest TD's attempt to set the corresponding CR4 bit can result in a #GP(0):

1. Depending on the TD's XFAM, guest TD modification of CR4 bits PKE (22), CET (23) and UINTR (25) is prevented. Any guest TD attempt to change those bits results in a #GP(0).
2. If the TD's ATTRIBUTES.KL is 0, guest TD attempts to set bit KL (19) results in a #GP(0). See 11.18 below.
3. If the TD's ATTRIBUTES.PKS is 0, guest TD attempts to set bit PKS (24) results in a #GP(0). See 11.22 below.
4. If the TD's virtual value of CPUID(7,1).EAX[6] (LASS) is 0, the TD is not allowed to use LASS, and guest TD attempts to set bit LASS (27) results in a #GP(0).
5. If the TD's virtual value of CPUID(0x7,1).EAX[17] is 0 (either configured by the host VMM or the CPU does not support FRED), the TD is not allowed to use FRED, and guest TD attempts to set bit FRED (32) results in a #GP(0). See 11.21 below.

In addition, any guest TD attempts to modify any of the architecturally reserved CR4 bits, or to set architectural-invalid bit combinations, can result in a #GP(0).

For TD migration, the same rules are used for checking the imported value of guest CR4. Any violation results in a failed import.

11.9.2.2. *CR4.MCE (Bit 6) Virtualization*

The guest TD's ability to modify the virtual value of CR4.MCE depends on the configuration set by the host VMM and the guest TD. There are multiple configurations, described below.

Note: The real value of CR4.MCE is always set to 1, to allow proper TDX operation, but the guest TD never handles #MC. Machine check events always cause a VM exit to the TDX module and are handled by it. For details, see Ch. 16.

Default Virtualization as Fixed-1

By default, virtual CR4.MCE is fixed at 1, and the guest TD is not allowed to modify it. A guest TD attempt to modify CR4.MCE results in a #VE(CONFIG_PARAVIRT).

Note: If the TD sets TDCS.TD_CTLS.REDUCE_VE to 1, then clears CR4.MCE as described below, then clears TDCS.TD_CTLS.REDUCE_VE, the TD-visible value of CR4.MCE will still be 0.

Architectural, Non-Paravirtualized Virtualization

If the TDX module supports #VE reduction, and the guest TD enables it (by setting TDCS.TD_CTLS.REDUCE_VE to 1) but does not enable MCA paravirtualization (TDCS.FEATURE_PARAVIRT_CTLS.MCA is 0), then the virtual value of CPUID(1).EDX[7] is 0, meaning MCE is disabled. The guest TD is not allowed to modify the value of virtual CR4.MCE from 0 to 1; a guest TD attempt to do so results in a #GP(0).

Note: Virtual CR4.MCE's value is initialized to 1, which is the default value as described above. The guest TD may clear this bit but not set it back to 1 as long as TDCS.TD_CTLS.REDUCE_VE is 1.

Architectural, Paravirtualized Virtualization

If the TDX module supports #VE reduction, and the guest TD enables it (by setting TDCS.TD_CTLS.REDUCE_VE to 1) and enables MCA paravirtualization (by setting TDCS.FEATURE_PARAVIRT_CTLS.MCA to 1), then virtual CPUID(1).EDX[7] is configured by the host VMM.

- If virtual CPUID(1).EDX[7] is 0, meaning MCE is disabled, the behavior is the same as with the non-paravirtualized configuration above. The guest TD is not allowed to modify the value of virtual CR4.MCE from 0 to 1; a guest TD attempt to do so results in a #GP(0).
- Else (virtual CPUID(1).EDX[7] is 1, meaning MCE is enabled), the guest TD is allowed to modify CR4.MCE.

CR4 Bits which are Non-Architecturally Virtualized

From the guest TD's point of view, the following bits are virtualized as fixed-0 by Intel TDX module. Guest TD attempts to modify their values result in a #VE(UNSUPPORTED_FEATURE):

- CR4.VMXE (bit 13)
- CR4.SMXE (bit 14)

11.10. MSR Virtualization

11.10.1. Overview

From the guest TD's point of view, as virtualized by the Intel TDX module, MSRs are divided into the following categories:

- MSRs that are context-switched on TD entry and exit – guest TD access to such MSRs may be full, partial or none
- MSRs that are not context-switched, but guest TD access is read-only
- MSRs that are not context-switched, and are inaccessible to the guest TD

MSR behavior can be either fixed or dependent on the TD configuration via the XFAM, ATTRIBUTES and CPUID configuration parameters. The host VMM has no direct interface to configure specific MSR behavior (e.g., it cannot set a specific MSR to TD exit on write). Instead, guest TD access violations to MSRs can cause a #GP(0) in most cases where the MSR is enumerated as inaccessible by the Intel TDX module via CPUID virtualization. In other cases, guest TD access violations to MSRs can cause a #VE. A guest TD that wishes to access an MSR that is not allowed by the Intel TDX module should do so via explicit requests from the host VMM using TDCALL(TDG.VP.VMCALL).

A detailed list of MSR virtualization is provided in the [TDX Module ABI Spec].

11.10.2. MSR Virtualization Configuration by the Host VMM

For a few MSRs, the virtualized values of some bit field returned to the guest TD when executing RDMSR can be configured by the host VMM. Configuration is done as an input to TDH.MNG.INIT.

Table 11.10: Host VMM Configurable MSR Field Virtualization

MSR Bit Configuration	Description
ALLOW_DIRECT	An MSR bit's virtual value can be configured as follows: • Allowed by the host VMM, i.e., reflects the native value returned by the CPU. • Forced to 0 by the host VMM, regardless of its native values.
FORCE_DIRECT	An MSR bit's virtual value can be configured as follows: • Forced to 1 by the host VMM, regardless of whether or not supported by the CPU. • Allowed by the host VMM, i.e., reflects the native value returned by the CPU.

For details, see the TD_PARAMS definition in the [ABI Spec].

Implication on TD Migration

The virtual MSR values calculated on TDH.MNG.INIT are stored in TDCS. If the TD is migrated, the values are exported by TDH.EXPORT.STATE.IMMUTABLE and checked on import to the destination TD by TDH.IMPORT.STATE.IMMUTABLE to be compatible with the destination platform.

For MSR bits that are configurable as ALLOW_DIRECT, an imported value of 0 is always allowed, regardless of the destination CPU's native value.

For MSR bits that are configurable as FORCE_DIRECT, an imported value of 1 is always allowed, regardless of the destination CPU's native value.

11.10.3. MSR Virtualization Control by the Guest TD

Depending on the TDX features supported by the TDX module, the guest TD may control the virtualization of specific MSRs. Refer to 11.1 for an overview of virtualization and paravirtualization, and to 11.2 for an overview of guest TD controls.

#VE Reduction and Feature Paravirtualization Control

By default, guest TD MSR access of many MSRs results in #VE. The guest TD is expected to implement a paravirtualization agent as part of its #VE exception handler (see 11.1 for an overview).

If the TDX module supports #VE reduction, as indicated by TDX_FEATURES0.VE_REDUCTION (bit 30), readable by TDG.SYS.RD*, the guest TD may set TDCS.TD_CTLS.REDUCE_VE (bit 3) to 1. This greatly reduces the number of cases where an MSR access results in a #VE, by virtualizing many CPU features as unsupported and eliminating the need to paravirtualize them. The guest TD may also control specific CPU features paravirtualization by setting bits in TDCS.FEATURE_PARAVIRT_CTRL.

For details, see 11.1 and the [ABI Spec] definition of TDCS.TD_CTLS, TDCS.FEATURE_PARAVIRT_CTRL and of CPUID virtualization.

11.11. CPUID Virtualization

11.11.1. CPUID Configuration by the Host VMM

For some CPUID leaves and sub-leaves, the virtualized bit fields of CPUID return values (in guest EAX/EBX/ECX/EDX) are configurable by the host VMM. For such cases, the Intel TDX module architecture defines two virtualization types:

Table 11.11: Host VMM Configurable CPUID Field Virtualization

CPUID Field Virtualization	Description	Comments
As Configured	Bit fields for which the host VMM configures the value seen by the guest TD. Configuration is done on TDH.MNG.INIT.	
As Configured (if Native)	Bit fields for which the host VMM configures the value such that the guest TD either sees their native value or a value of 0. Configuration is done on TDH.MNG.INIT.	If a CPUID bit enumerates a CPU feature, and the feature is natively supported, then the feature can either be allowed by the host VMM, or it will be effectively deprecated for the guest TD.

The above CPUID fields can be specified by the host VMM at guest TD initialization time TDH.MNG.INIT using the TD_PARAMS input structure of TDH.MNG.INIT. TDH.MNG.INIT and its input TD_PARAMS structure are described in the [TDX Module ABI Spec]. Configuration is further classified as follows:

Table 11.12: CPUID Configuration by the TD_PARAMS Input of TDH.MNG.INIT

TD_PARAMS Section	Description	Notes
CPUID_CONFIG	Bit fields configurable directly based on a configuration table	Some bit fields are configurable by both CPUID_CONFIG and either XFAM or ATTRIBUTES. See the discussion below.
XFAM	Bit fields configurable based on the guest TD's XFAM XFAM control of extended features virtualization is described in 11.8.	
ATTRIBUTES	Bit fields configurable based on the guest TD's ATTRIBUTES	
Other	Bits fields configurable based on some other field of TD_PARAMS	

A detailed list of CPUID virtualization is provided in the [TDX Module ABI Spec]. For any valid CPUID leaf / sub-leaf combination that is not listed, the Intel TDX module injects a #VE.

The host VMM should always consult the list of directly configurable CPUID leaves and sub-leaves, as enumerated by TDH.SYS.RD/RDALL or TDH.SYS.INFO, described in 4.1.2.4.

11.11.1.1. *Fine Grained Control of CPU Extended Features Enumeration*

Some CPUID bit fields are configurable based on both CPUID_CONFIG and either XFAM or ATTRIBUTES sections of TD_PARAMS. This is intended to support **fine-grained virtualization of sub-features of extended features**. For example:

- The host VMM can configure the TDX module to virtualize some AVX512 as available, but to virtualize other AVX512 instructions as unavailable.
- The host VMM can configure the TDX module to virtualize the Perfmon architectural events support.

This is useful for TD migration, as it allows the host VMM to configure a common subset of supported sub-features.

11.11.1.2. *Configurable Family/Model/Stepping (CPUID(1).EAX) Enumeration*

By default, the CPU's Family/Model/Stepping value, as enumerated to the guest TD by CPUID(1).EAX, is set to the native value of the platform. In a multi-package platform, the stepping value is the minimum of all the packages' native values. If CPUID(1).EAX is enumerated by the TDX module as configurable, then the host VMM can select the default native configuration by setting CPUID(1).EAX in TD_PARAMS.CPUD_CONFIG input to TDH.MNG.INIT to all-0.

If supported by the TDX module, then if the TD is migratable (ATTRIBUTES.MIGRATABLE is 1), the value of CPUID(1).EAX may be configured by the host VMM as described below. This allows the host VMM to create migration pools containing multiple CPU types. As with all configurable features, the host VMM would need to be careful and configure the migratable TDs virtual Family/Model/Stepping values to be recognizable by all platforms in the migration pool.

Enumeration	TDX module's support of CPUID(1).EAX configuration is enumerated by the global metadata field TDX_FEATURES0.FMS_CONFIG (bit 17), which is readable by the host VMM using TDH.SYS.RD*. For details, see the [ABI Spec].

The TDX module allows the host VMM to configure Family/Model/Stepping to one of a specific set of values. Those values are known to the TDX module as CPUs, each conforming to the following criteria:

- There's a TDX module version that supports the CPU denoted by the Family/Model/Stepping value.
- The CPU denoted by the Family/Model/Stepping value is "not newer" than the current (native) CPU at the time of TD initialization, in a sense that it does not implement any security feature that is not implemented by the current CPU. E.g., that CPU does not implement some side channel mitigation that the current CPU does not implement.

The specific list of family/model/stepping values is provided by the following global metadata fields), readable by the host VMM using TDH.SYS.RD*:

ALLOWED_FMS	A list of 32-bit values in CPUID(1).EAX Family/Model/Stepping format. For each, the allowed Family and Model fields are specified. The Stepping value is the maximum allowed value.
DISALLOWED_FMS	A list of 32-bit fields in CPUID(1).EAX Family/Model/Stepping format, listing explicitly disallowed F/M/S settings.

For details, see the [ABI Spec].

11.11.2. Guest TD Control of CPUID Virtualization

11.11.2.1. *Guest TD Control of Specific CPUID Leaves and Sub-Leaves Virtualization*

Depending on the TDX features supported by the TDX module, the guest TD may control the virtualization of specific CPUID leaves and sub-leaves. Refer to 11.1 for an overview of virtualization and paravirtualization, and to 11.2 for an overview of guest TD controls.

11.11.2.1.1. #VE Reduction and Feature Paravirtualization Control

By default, guest TD execution of CPUID with many leaf and sub-leaf numbers result in #VE. The guest TD is expected to implement a paravirtualization agent as part of its #VE exception handler (see 11.1 for an overview).

If the TDX module supports #VE reduction, as indicated by TDX_FEATURES0.VE_REDUCTION (bit 30), readable by TDG.SYS.RD*, the guest TD may set TDCS.TD_CTLS.REDUCE_VE (bit 3) to 1. This greatly reduces the number of cases where a CPUID results in a #VE, by virtualizing many CPU features as unsupported and eliminating the need to paravirtualize them. The guest TD may also control specific CPU features paravirtualization by setting bits in TDCS.FEATURE_PARAVIRT_CTRL.

Guest TD execution of CPUID with a leaf number in the range 0x40000000 - 0x4FFFFFFF always results in a #VE(NON_CONFIG_PARAVIRT). This CPUID leaf range is typically used for guest-host communication by multiple host VMM types; the guest's #VE handler may paravirtualize the host VMM response (e.g., using a hypercall over TDG.VP.VMCALL).

For details, see 11.1 and the [ABI Spec] definition of TDCS.TD_CTLS, TDCS.FEATURE_PARAVIRT_CTRL and of CPUID virtualization.

11.11.2.1.2. Topology Virtualization

Note: Topology virtualization control is a subset of the #VE reduction control described above, which was implemented in earlier TDX module versions. It is available as a separate control for backward compatibility.

By default, virtual topology enumeration by CPUID(0x1F) and CPUID(0xB) is disabled. If the TDX module supports topology virtualization, as indicated by TDX_FEATURES0.TOPOLOGY_ENUM, readable by TDG.SYS.RD*, the guest TD may enable it by setting TDCS.TD_CTLS.ENUM_TOPOLOGY (bit 1) to 1. If the TDX module supports #VE reduction, as indicated by TDX_FEATURES0.VE_REDUCTION (bit 30), the guest TD may set TDCS.TD_CTLS.REDUCE_VE (bit 3) to 1; this will also implicitly set ENUM_TOPOLOGY (bit 1). See 11.12 below for details.

11.11.2.1.3. CPUID(2) (Cache and TLB Information) Virtualization

Note: CPUID(2) virtualization control is a subset of the #VE reduction control described above, which was implemented in earlier TDX module versions. It is available as a separate control for backward compatibility.

By default, cache and TLB enumeration by CPUID(0x2) is disabled. If the TDX module supports CPUID(2) virtualization, as indicated by TDX_FEATURES0.CPUID2_VIRT (bit 29), readable by TDG.SYS.RD*, the guest TD may enable it by setting TDCS.TD_CTLS.VIRT_CPUID2 (bit 2) to 1. If the TDX module supports #VE reduction, as indicated by TDX_FEATURES0.VE_REDUCTION (bit 30), the guest TD may set TDCS.TD_CTLS.REDUCE_VE (bit 3) to 1; this will also implicitly set VIRT_CPUID2 (bit 2).

11.11.2.2. *Per-VCPU Guest TD Control of #VE on CPUID*

11.11.2.2.1. Per-VCPU #VE for all CPUID Leaves and Sub-Leaves

The guest TD may toggle on or off the unconditional injection of #VE(CONFIG_PARAVIRT) on all CPUID leaves and sub-leaves, per VCPU. That can be done in supervisor mode (CPL == 0) and/or user mode (CPL > 0). For example, this enables the TD OS to control CPUID as seen by drivers or by user-level code.

The guest TD may do this by writing to the VCPU-scope metadata fields CPUID_SUPERVISOR_VE and CPUID_USER_VE using TDG.VP.WR.

For backward compatibility, the guest TD may use TDG.VP.CPUIDVE.SET, described in the [TDX Module ABI Spec].

11.11.2.2.2. Per-VCPU #VE for Specific CPUID Leaves and Sub-Leaves

A finer grained control is provided per CPUID leaf and sub-leaf that is virtualized by the TDX module. The guest TD may configure the following, per VCPU:

- #VE(CONFIG_PARAVIRT) injection instead of the normal CPUID virtualization is the guest executed CPUID in supervisor mode (CPL == 0).
- #VE(CONFIG_PARAVIRT) injection instead of the normal CPUID virtualization is the guest executed CPUID in user mode (CPL > 0).

The guest TD may do this by writing to the VCPU-scope metadata field array CPUID_CONTROL using TDG.VP.WR.

Note: This feature is only available for CPUID leaves and sub-leaves that do not inject a #VE if TDCS.TD_CTLS.REDUCE_VE is 0.

11.11.3. CPUID Configuration & Checks at Guest TD Migration

The CPUID virtualization configuration stored in TDCS is exported by TDH.EXPORT.STATE.IMMUTABLE and checked on import to the destination TD by TDH.IMPORT.STATE.IMMUTABLE to be compatible with the destination platform.

CPUID fields that are virtualized as fixed values (defined as "FIXED"), are based on some calculation (defined as "ASSIGNED") or that their value depends on the underlying CPU capabilities (defined as "ALLOWED" or "DIRECT") must retain the same value across migration.

CPUID fields that are virtualized as pass-through (defined as "NATIVE") are considered fixed once exported and are checked for compatibility on import.

11.11.4. CPUID Virtualization for Hybrid SOCs

As described above, a TD is always virtualized as a uniform VM. This implies the following:

- The virtual value of CPUID(7,0).EDX[15], which enumerates the SOC as hybrid, is 0.
- When a TD is initialized on a hybrid SOC, CPUID(0x1A).EAX, which enumerates the core type and native model ID, is virtualized as 0, indicating non-applicable information. When a TD is initialized on a uniform SOC, the virtual value of CPUID(0x1A).EAX reflects its native value.

Other CPUID virtualization cases related to SOC are for performance monitoring (Perfmon), processor trace (PT) and last branch record (LBR). They are described in Ch. 15.

11.12. Platform Topology Virtualization

If supported by the TDX module, it provides the guest TD with virtual platform topology information, configured by the host VMM.

Enumeration: TDX module support for virtual platform topology enumeration is enumerated by TDX_FEATURES0.TOPOLOGY_ENUM, readable by the host VMM and the guest TD using TDH.SYS.RD or TDG.SYS.RD respectively. For details see the [ABI Spec].

11.12.1. Configuration by the Host VMM

As an input to TD initialization (TDH.MNG.INIT), the host VMM can configure the values of the CPUID extended topology leaf (0x1F) and its sub-leaves. Alternatively, the host VMM can configure the CPUID(0x1F) values as all 0; in this case, the virtual values are set to the native CPUID(0x1F) values.

The following conditions must be met:

- CPUID(0x1F) sub-leaves must specify level types (CPUID(0x1F, *).ECX[15:8]) in an ascending order, except the last one.
- The last CPUID(0x1F) sub-leaf must specify level type 0 (INVALID).
- The host VMM must provide configuration for the core level (2).

As an input to each VCPU initialization (TDG.VP.INIT), the host VMM must specify a virtual x2APIC ID for that VCPU. That value must be unique across all VCPUs of the current TD.

The virtual values of CPUID(0xB) are calculated by the TDX module from the configured virtual values of CPUID(0x1F), as described below.

For details, see the [ABI Spec].

11.12.2. Enabling by the Guest TD

See also 11.11.2 for a generic discussion of guest TD control of CPUID virtualization.

By default, virtual topology enumeration is disabled. The guest TD may enable virtual topology enumeration by setting TDCS.TD_CTLS.ENUM_TOPOLOGY (bit 1) to 1 or by setting TDCS.TD_CTLS.REDUCE_VE (bit 3) to 1, if supported. This depends on the following conditions:

- TDX module enumerates this feature as supported, as described above.
- The host VMM properly configured the virtual x2APIC ID for each VCPU. This is indicated by TDCS.TOPOLOGY_ENUM_CONFIGURED, which may be read using TDG.VM.RD.

For details, see the [ABI Spec].

11.12.3. Virtual Topology Information Provided to the Guest TD

The table below shows the virtual topology information, depending on enabling by the guest TD.

Table 11.13: Virtual Topology Information Provided to the Guest TD

CPUID or MSR	TD_CTLS.ENUM_TOPOLOGY = 0	TD_CTLS.ENUM_TOPOLOGY = 1
CPUID(0x1F) (V2 Extended Topology Enumeration)	#VE(CONFIG_PARAVIRT)	EAX, EBX, ECX: Host VMM configured platform topology values EDX: Current VCPU's host VMM configured x2APIC ID
CPUID(0xB) (Extended Topology Enumeration)	#VE(CONFIG_PARAVIRT)	EAX, EBX, ECX: Derived from the CPUID(0x1F) values – see below. EDX: Current VCPU's host VMM configured x2APIC ID
CPUID(1).EBX[31:24]	least significant 8 bits of the VCPU's sequential index	Least significant 8 bits of the VCPU's x2APIC ID
MSR 0x802 (IA32_X2APIC_APICID)	#VE(CONFIG_PARAVIRT)	Current VCPU's host VMM configured x2APIC ID

11.12.3.1. *Derivation of CPUID(0xB) Virtual Values from CPUID(0x1F) Configuration*

Intel SDM, Vol. 2	CPUID

CPUID(0x1F) can enumerate multiple domain levels (Logical Processor, Core, Module, etc.) while CPUID(0xB) can only enumerate the Logical Processor and Core level. The TDX module derives the virtual values of CPUID(0xB) from the configured virtual values of CPUID(0x1F) as follows:

- If a Logical Processor domain has been configured for CPUID(0x1F), then the same values are used as CPUID(0xB)'s Logical Processor domain.
- CPUID(0xB)'s Core domain value is set to the CPUID(0x1F)'s highest configured domain value.

11.13. Interrupt Handling and APIC Virtualization

Intel SDM, Vol. 3, 24.6.8	Controls for APIC Virtualization
Intel SDM, Vol. 3, 31	APIC Virtualization and Virtual Interrupts

11.13.1. Virtual APIC Mode

- Guest TDs must use x2APIC mode. xAPIC mode (using memory mapped APIC access) is not allowed.
- Guest TD attempts to RDMSR or WRMSR the IA32_APIC_BASE MSR cause a #VE(NON_CONFIG_PARAVIRT) to the guest TD. The guest TD cannot disable the APIC.

11.13.2. Virtual APIC Access by Guest TD

Intel SDM, Vol. 3, 30.5	Virtualizing MSR-Based APIC Access

Guest TDs are allowed access to a subset of the virtual APIC registers, which are virtualized by the CPU as described in [Intel SDM, Vol. 3, 30.5]. Access to other registers can cause a #VE. The guest TD is expected to use a software protocol over TDG.VP.VMCALL to request such operations from the host VMM.

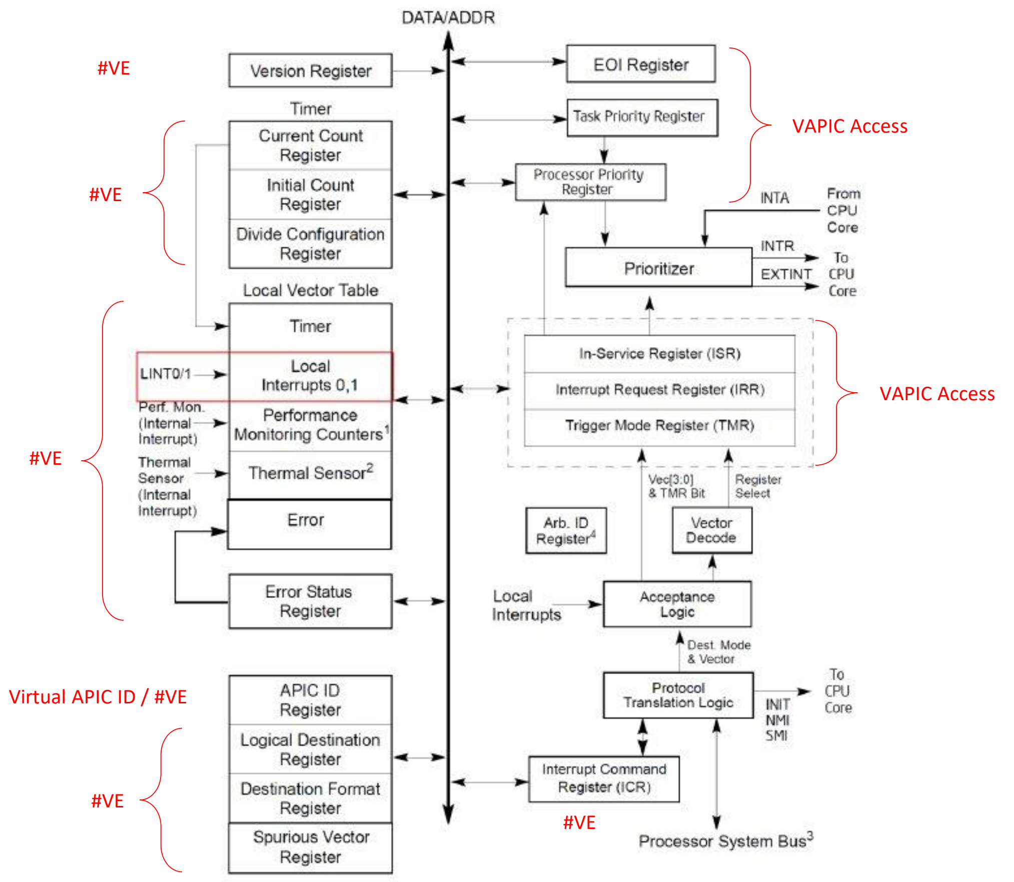


Figure 11.4: Virtual APIC Access by Guest TD

Table 11.14: x2APIC MSRs (0x800 – 0x8FF) Processing

MSR Range	MSR Name(s)	Description	Access	On RDMSR	On WRMSR
0x802	IA32_X2APIC_APICID	APIC ID	RO	If TD_CTLS.ENUM_TOPOLOGY is set, return the virtual x2APIC ID, as configured (see 11.12.3 above) Else, #VE(CONFIG_PARAVIRT)	#GP(0)
0x803	IA32_X2APIC_VERSION	APIC Version	RO	#VE(NON_CONFIG_PARAVIRT)	#GP(0) or #VE[8]
0x808	IA32_X2APIC_TPR	Task Priority	RW	Read from VAPIC page	Write to VAPIC page, TPR virtualization
0x80A	IA32_X2APIC_PPR	Processor Priority	RO	Read from VAPIC page	#GP(0)
0x80B	IA32_X2APIC_EOI	End Of Interrupt	WO	Read from VAPIC page	Write to VAPIC page, EOI virtualization
0x80D	IA32_X2APIC_LDR	Local Destination	RO	#VE(NON_CONFIG_PARAVIRT)	#GP(0) or #VE[8]
0x80F	IA32_X2APIC_SIVR	Spurious Interrupt Vector	RW	#VE(NON_CONFIG_PARAVIRT)	#VE

[8] If the TDX module supports #VE reduction, as enumerated by TDX_FEATTURES0.VE_REDUCTION (bit 30), then this MSR access results in a #GP(0). Else, it results in a #VE.

MSR Range	MSR Name(s)	Description	Access	On RDMSR	On WRMSR
0x810-0x817	IA32_X2APIC_ISR0-IA32_X2APIC_ISR7	In-Service	RO	Read from VAPIC page	#GP(0)
0x818-0x81F	IA32_X2APIC_TMR0-IA32_X2APIC_TMR7	Trigger Mode	RO	Read from VAPIC page	#GP(0)
0x820-0x827	IA32_X2APIC_IRR0-IA32_X2APIC_IRR7	Interrupt Request	RO	Read from VAPIC page	#GP(0)
0x828	IA32_X2APIC_ESR	Error Status	RW	#VE(NON_CONFIG_PARAVIRT)	#VE(NON_CONFIG_PARAVIRT)
0x830	IA32_X2APIC_ICR	Interrupt Command	RW	#VE(NON_CONFIG_PARAVIRT)	#VE(NON_CONFIG_PARAVIRT)
0x82F, 0x832-0x837, 0x83A	IA32_X2APIC_LVT_*	Local Vector Table	RW	#VE(NON_CONFIG_PARAVIRT)	#VE(NON_CONFIG_PARAVIRT)
0x838	IA32_X2APIC_INIT_COUNT	APIC Timer	RW	#VE(NON_CONFIG_PARAVIRT)	#VE(NON_CONFIG_PARAVIRT)
0x839	IA32_X2APIC_CUR_COUNT		RO	#VE(NON_CONFIG_PARAVIRT)	#GP(0) or #VE[8]
0x83E	IA32_X2APIC_DIV_CONF		RW	#VE(NON_CONFIG_PARAVIRT)	#VE(NON_CONFIG_PARAVIRT)
0x83F	IA32_X2APIC_SELF_IPI	Self IPI	WO	Read from VAPIC page	Write to VAPIC page, self-IPI virtualization
0x83B-0x83D	Reserved	N/A	None	#GP(0) or #VE[8]	#GP(0) or #VE[8]
Other	Reserved	N/A	None	#GP(0)	#GP(0)

11.13.3. Implicit APIC Write #VE

The following guest operations result in an APIC write VM exit to the TDX module. The VM exit is trap-like, i.e., it happens after the instruction has been executed:

- WRMSR of IA32_X2APIC_SELF_IPI with EAX[7:4] set to 0, i.e., an interrupt vector value smaller than 16.
- Executing SENDUIPI to send a user-level interrupt.

In all such cases, the TDX module injects a #VE(NON_CONFIG_PARAVIRT) exception back to the guest TD, with the exit reason indicating an APIC write and bits 11:0 of the exit qualification set to the page offset of the write access.

11.13.4. Posted Interrupts

Intel SDM, Vol. 3, 29.6 Posted-Interrupt Processing

Non-NMI interrupt injection into the guest TD by the host VMM or the IOMMU can be done through the posted-interrupt mechanism. If there are pending interrupts in the posted-interrupt descriptor (PID), the VMM can post a self IPI with the notify vector prior to TD entry.

- The posted-interrupt descriptor (PID) resides in a shared page, directly accessible by the host VMM. The VMM must set the TD VMCS's "posted-interrupt descriptor address" control (using the TDH.VP.WR function) to the PA and shared HKID of the posted-interrupt descriptor.
- The host VMM must set the TD VMCS's "posted-interrupt notification vector" control using the TDH.VP.WR function.
- To post pending interrupts in the PID, the host VMM can generate a self IPI with the notification vector prior to TD entry.

When a posted-interrupt notification vector is recognized in SEAM non-root mode, the CPU processes the posted-interrupt descriptor as described in the [Intel SDM].

If needed, the guest TD may use a software protocol over TDCALL(TDG.VP.VMCALL) to ask the VMM to stop interrupt delivery through the PID.

Security Implications of Posted Interrupts

A malicious host VMM or a device may post any virtual interrupt vector in the range 255:31 at any time. The guest TD should be able to process such interrupts without confusing it with a software interrupt that uses the same vector number. The TD's interrupt handler for vector V, which expects a software interrupt, can read the virtual APIC's ISR register by reading the applicable IA32_X2APIC_ISRx MSRs (0x817:0x810). It can check that ISR[V] is indeed 0 for the specific vector.

TD Migration Implications of Posted Interrupts

The TD VMCS posted interrupt execution controls are reset to their initial values when the TD is migrated. The host VMM on the destination platform must set them in order to use posted interrupts.

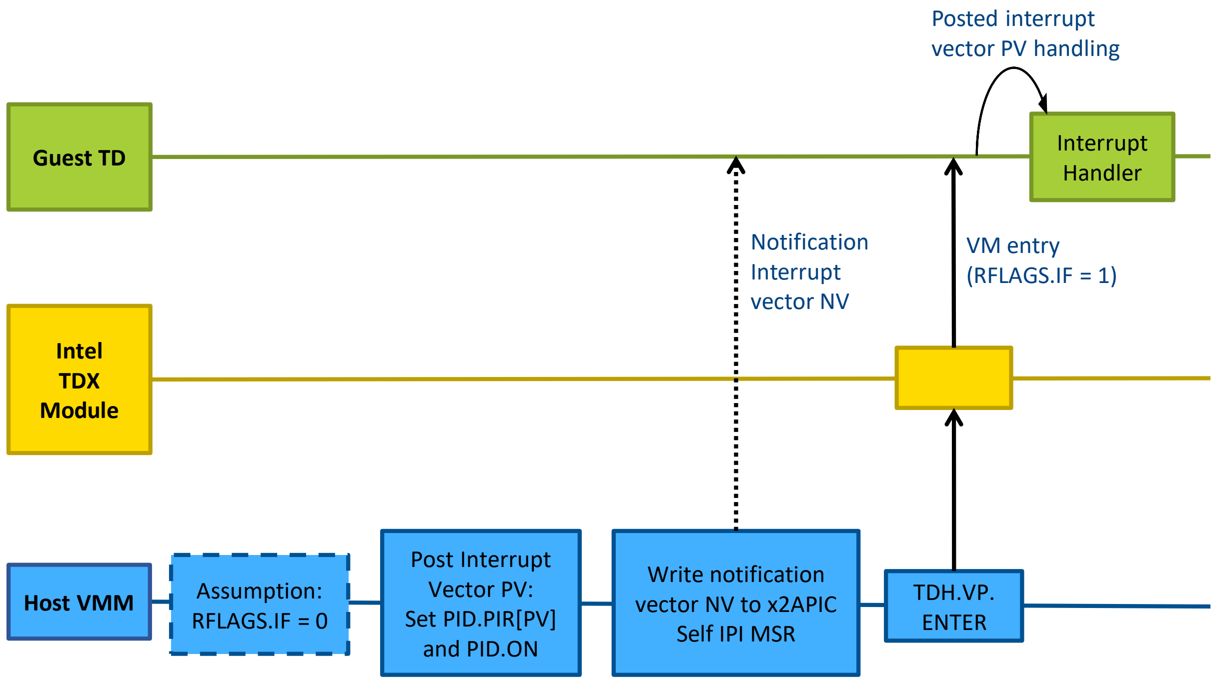

Posted interrupt vector PV handling
Guest TD
Interrupt Handler
Notification Interrupt vector NV
VM entry (RFLAGS.IF = 1)
Intel TDX Module
Host VMM
Assumption: RFLAGS.IF = 0
Post Interrupt Vector PV: Set PID.PIR[PV] and PID.ON
Write notification vector NV to x2APIC Self IPI MSR
TDH.VP. ENTER

Figure 11.5: Typical Sequence for Posted Interrupt Injection to the Current LP

TD Partitioning Implications of Posted Interrupts

Posted interrupts handling for partitioned TDs is discussed in the [TD Partitioning Spec].

11.13.5. Pending Virtual Interrupt Delivery Indication

The host VMM can detect whether a virtual interrupt is pending delivery to a VCPU in the Virtual APIC page, using TDH.VP.RD to read the VCPU_STATE_DETAILS TDVPS field.

The typical use case is when the guest TD VCPU indicates to the host VMM, using TDG.VP.VMCALL, that it has no work to do and can be halted. The guest TD is expected to pass an "interrupt blocked" flag. The guest TD is expected to set this flag to 0 if and only if RFLAGS.IF is 1 or the TDCALL instruction that invokes TDG.VP.VMCALL immediately follows an STI instruction. If the "interrupt blocked" flag is 0, the host VMM can determine whether to re-schedule the guest TD VCPU based on VCPU_STATE_DETAILS.

For further details, see the TDVPS definition in the [TDX Module ABI Spec].

11.13.6. Cross-TD-VCPU IPI

To perform a cross-VCPU IPI, the guest TD ILP should request an operation from the host VMM using TDG.VP.VMCALL. The VMM would then inject the requested virtual interrupt into the guest TD's RLPs using the posted interrupt mechanism, as described in 11.13.4 above. This is an untrusted operation; thus, the TD needs to track its completion.

11.13.7. Virtual NMI Injection

The host VMM can request the Intel TDX module to inject a virtual NMI into a guest TD VCPU using the TDH.VP.WR function, by setting the PEND_NMI TDVPS field to 1. This can be done only when the VCPU is not active (a VCPU can be associated with at most one LP). Following that, the host VMM can call TDH.VP.ENTER to run the VCPU; the Intel TDX module will attempt to inject the NMI as soon as possible.

The host VMM can use TDH.VP.RD to read PEND_NMI and get the status of virtual NMI injection. A value of 0 indicates that virtual NMI has been injected into the guest TD VCPU. The host VMM also may choose to clear PEND_NMI before it is injected.

11.14. Virtualization Exception (#VE)

Intel SDM, Vol. 3, 24.9.4	Information for VM Exits Due to Instruction Execution
Intel SDM, Vol. 3, 25.5.6	Virtualization Exceptions
Intel SDM, Vol. 3, 27.2.5	Information for VM Exits Due to Instruction Execution

The Intel TDX module extends the VMX architectural definition of #VE. It injects #VE into the guest TD in multiple cases where an operation is not allowed by TDX, but an architectural exception (e.g., #GP(0)) is not applicable. Such cases include disallowed instruction executions, disallowed MSR accesses, many CPUID leaves, etc.

The intended usage is for the TDX-enlightened guest TD OS to have a #VE handler. By analyzing the #VE information, the handler would be able to emulate the requested operation for non-enlightened parts of the guest TD – e.g., drivers and applications.

11.14.1. Virtualization Exception Information

The **virtualization-exception information area** (VE_INFO) is maintained as part of TDVPS. It is not intended to be directly accessible by the guest TD. Instead, the #VE information can be retrieved using the **TDG.VP.VEINFO.GET** function (see the [TDX Module ABI Spec]). This is a simple way to help ensure the availability and privacy of this area.

Table 11.15: Virtualization Exception Information Area (VE_INFO), based on [Intel SDM, Vol. 3, Table 24-1]

Section	Field	Offset (Bytes)	Size (Bytes)	Description
Architectural	**EXIT_REASON**	0	4	The value that would have been saved into the VMCS as an exit reason if a VM exit had occurred instead of the virtualization exception.
	VALID	4	4	0 indicates that VE_INFO has no valid contents. The CPU and the Intel TDX module will not update VE_INFO if VALID is not 0. After updating VE_INFO, the CPU and the Intel TDX module write 0xFFFFFFFF to the VALID field.
	EXIT_ QUALIFICATION	8	8	The value that would have been saved into the VMCS as an exit qualification if a VM exit had occurred instead of the virtualization exception.
	GLA	16	8	The value that would have been saved into the VMCS as a guest-linear address if a VM exit had occurred instead of the virtualization exception.
	GPA	24	8	The value that would have been saved into the VMCS as a guest-physical address if a VM exit had occurred instead of the virtualization exception.
	EPTP_INDEX	32	2	The current value of the EPTP index VM-execution control
Non-Architectural **(EXIT_REASON is not EPT Violation)**	**INSTRUCTION_ LENGTH**	Non-arch.	4	The 32-bit value that would have been saved into the VMCS as VM-exit instruction length if a legacy VM exit had occurred instead of the virtualization exception.
	INSTRUCTION_ INFORMATION	Non-arch.	4	The 32-bit value that would have been saved into the VMCS as VM-exit instruction information if a legacy VM exit had occurred instead of the virtualization exception.

Section	Field	Offset (Bytes)	Size (Bytes)	Description
	EXTENDED_ INSTRUCTION_ INFORMATION	Non-arch.	8	The 64-bit value that would have been saved into the VMCS as VM-exit extended instruction information if a legacy VM exit had occurred instead of the virtualization exception. This field is only applicable for TDX Modules and CPUs which support the VM-exit extended instruction information VMCS field.
VE_INFO Category	**CATEGORY**	Non-arch.	1	Category of the VE_INFO – intended to help the guest TD decide how to handle the #VE exception. See the table below for details. **Enumeration:** CATEGORY information is supported if the TDX module supports #VE reduction, as enumerated by TDX_FEATURES0.VE_REDUCTION (bit 30).

The architectural section format for VE_INFO is as defined in the [Intel SDM], and it is used directly by the CPU when it injects a #VE (see 11.14.2 below). VE_INFO can also be used for #VE injected by the Intel TDX module. Some VE_INFO fields are applicable only for some exit reasons.

VE_INFO's non-architectural section is only applicable for TDX-extended #VE (injected by the TDX module), where EXIT_REASON is not EPT violation (48). It should be ignored for EPT violations converted by the CPU to #VE. See below for details.

VE_INFO.VALID is initialized to 0, and it is set to 0xFFFFFFFF when a #VE is injected to the guest TD. When handling a #VE, the guest TD retrieves the #VE information using the **TDG.VP.VEINFO.GET** function (see the [TDX Module ABI Spec]). TDG.VP.VEINFO.GET checks that VE_INFO.VALID is 0xFFFFFFFF. After reading the information, it sets VE_INFO.VALID to 0.

Table 11.16: #VE Category Information

Value	Category	Description	Expected Guest TD Behavior
0x00	**ARCH**	#VE which has been converted from an EPT Violation	Handle the #VE
0x01	**PENDING**	#VE which has been converted from an EPT Violation on a PENDING page	Handle the #VE, e.g., by calling TDG.MEM.PAGE.ACCEPT
0x02	**RESERVED_GPA_BITS**	#VE which has been converted from an EPT Violation due to GPA bits above the MAXGPA range (except the SHARED bit) being set to 1	Handle this as a guest TD error
0x10	**CONFIG_PARAVIRT**	CPU feature configured by the host VMM (via CPUID configuration) to be paravirtualized by the guest TD	Either be prepared to handle the paravirtualization case (triggered by #VE) or treat it as error resulting from incorrect configuration by VMM.
0x11	**NON_CONFIG_PARAVIRT**	CPU feature that must be paravirtualized by the guest TD	Handle the paravirtualization case in #VE handler.
0x80	**UNSUPPORTED_FEATURE**	Guest TD attempted to use an x86 feature which is not supported by TDX	Indicates a bug in guest TD software – not allowed values should not be written

11.14.2. Architectural #VE Injection due to EPT Violations

EPT Violation mutation to #VE is enabled unconditionally for SEAM non-root operation. The Intel TDX module sets the TD VMCS **EPT-violation #VE** VM-execution control to 1.

For shared memory accesses (i.e., when GPA.SHARED == 1), as with legacy VMX, the VMM can choose which pages are eligible for #VE mutation based on the value of the Shared EPTE bit 63.

For private memory accesses (GPA.SHARED == 0), an EPT Violation causes a TD Exit in most cases, except when the Secure EPT entry state is PENDING (an exception to this is described in 11.15.1.5). Secure EPT entry bit 63 is always set to 1; the CPU never directly mutates EPT violations to #VE. However, EPT violations on PENDING pages may be mutated by the TDX module to #VE(PENDING); this is described in 9.5.2 and 9.10.4.

11.14.3. Non-Architectural #VE Injected by the Intel TDX Module

#VE may be injected by the Intel TDX module in several cases:

- Emulation of the architectural #VE injection on EPT violation, done by a guest-side Intel TDX module flow that performs an EPT walk.
- As a result of guest TD execution of a disallowed instruction (see 11.7 above), a disallowed MSR access (see 11.10 above), or CPUID virtualization (see 11.11 above).
- A notification to the guest TD about anomalous behavior (e.g., too many EPT violations reported on the same TD VCPU instruction without making progress). This kind of #VE is raised only if the guest TD enabled the specific notification (using TDG.VM.WR to write the TDCS.NOTIFY_ENABLES field) and when a #VE can be injected. See 17.3 for details.

If, when attempting to inject a #VE, the Intel TDX module discovers that the guest TD has not yet retrieved the information for a previous #VE (i.e., VE_INFO.VALID is not 0), the TDX module injects a #DF into the guest TD to indicate a #VE overrun.

11.15. GPA Space, Secure and Shared Extended Page Tables (EPTs)

EPT is enabled in SEAM non-root mode. SEAM non-root mode uses two EPTs: Secure EPT and Shared EPT.

EPT level is the same for both Secure and Shared EPT. If the guest TD's GPA width is greater than 48 bits (TDCS.GPAW is 1), then 5-level EPT trees are used. Otherwise, 4-level EPT trees can be used.

For further Secure EPT details, refer to Chapter 9.

EPT violations and misconfigurations generally cause a TD Exit, except for the cases described below.

11.15.1. GPA Space Size Configuration and Virtualization

11.15.1.1. *Overview of the GPA Space Size Virtualization Modes*

The host VMM can configure one of the TDX module's options for configuring and virtualizing the GPA space size available to guest TDs:

- No virtualization (native values are used)
- MAXPA (CPUID(0x80000008).EAX[7:0]) virtualization
- MAXGPA (CPUID(0x80000008).EAX[23:16]) virtualization

Specific TDX module releases may not support all the above features; note the enumeration of each feature support as documented below.

In all cases, the TDX module calculates an internal value TDCS.VIRT_MAXPA, which is used for virtualizing the GPA space size; see more details below.

The following table compares the GPA space virtualization modes.

Table 11.17: GPA Space Size Virtualization Modes Comparison

	GPA Space Virtualization Mode		
	None	**MAXPA Virtualization**	**MAXGPA Virtualization**
Virtual CPUID(0x80000008).EAX[7:0]	Native value (at TDH.MNG.INIT time)	Directly configured by the host VMM	Native value (at TDH.MNG.INIT time)
Virtual CPUID(0x80000008).EAX[23:16]	0	0	Configured by the host VMM's GPAW setting

	GPA Space Virtualization Mode		
	None	**MAXPA Virtualization**	**MAXGPA Virtualization**
TDCS.VIRT_MAXPA Used by the TDX Module	Native CPUID(0x80000008).EAX[7:0] value (at TDH.MNG.INIT time)	Set to the configured virtual CPUID(0x80000008).EAX[7:0] value	Set to the configured virtual CPUID(0x80000008).EAX[23:16] value
Exception on GPA bits above physical MAXPA (excl. SHARED bit) being set	#PF(RSVD)	#PF(RSVD)	#PF(RSVD)
Exception on GPA bits above TDCS.VIRT_MAXPA (excl. SHARED bit) being set	#PF(RSVD)	#PF(RSVD)	#VE(RESERVED_GPA_BITS)

11.15.1.2. *MAXPA (CPUID(0x80000008).EAX[7:0]) Virtualization*

Enumeration: TDX module's support of this feature is enumerated by TDX_FEATURES0.MAXPA_VIRT (bit 27), readable using TDH.SYS.RD*.

MAXPA, the number of physical address bits, is enumerated to the guest TD by the virtual value of CPUID(0x80000008).EAX[7:0]. That value can be configured by the host VMM as follows:

- If the host VMM sets TD_PARAMS.CONFIG_FLAGS.MAXPA_VIRT (bit 3) to 0, as an input to TDH.MNG.INIT, then MAXPA virtualization is disabled:
  - Virtual CPUID(0x80000008).EAX[7:0] is set to the native value at the time of TD initialization by TDH.MNG.INIT.
  - TDCS.VIRT_MAXPA (the value used by the TDX module) and virtual CPUID(0x80000008).EAX[23:16] are set depending on the value of TD_PARAMS.CONFIG_FLAGS.MAXGPA_VIRT (bit 4), as described below.
- Else (TD_PARAMS.CONFIG_FLAGS.MAXPA_VIRT (bit 3) is set to 1), then:
  - TD_PARAMS.CONFIG_FLAGS.MAXGPA_VIRT (bit 4) must be 0.
  - If the host VMM configured a virtual CPUID(0x80000008).EAX[7:0] value of 0:
    - TDCS.VIRT_MAXPA (the value used by the TDX module) is set to the smaller value between the native MAXPA value and of the TD's configured GPAW (which can be either 52 or 48).
  - Else, the configured virtual CPUID(0x80000008).EAX[7:0] value is checked as detailed below:
    - The value must not be higher than the native MAXPA (native CPUID(0x80000008).EAX[7:0]).
    - The value must not be higher than the TD's configured GPAW (which can be either 52 or 48).
    - The value must not be lower than the supported minimum value. The host VMM can read that value by reading field MIN_VIRT_MAXPA using TDH.SYS.RD*.

    If all checks pass, the configured virtual CPUID(0x80000008).EAX[7:0] value is used as TDCS.VIRT_MAXPA.

11.15.1.3. *MAXGPA (CPUID(0x80000008).EAX[23:16]) Virtualization*

Enumeration: TDX module's support of this feature is enumerated by TDX_FEATURES0.MAXGPA_VIRT (bit 33), readable using TDH.SYS.RD*.

MAXGPA, the number of guest physical address bits, is enumerated to the guest TD by virtual value of CPUID(0x80000008).EAX[23:16]. That value can be configured by the host VMM as follows:

- If the host VMM sets TD_PARAMS.CONFIG_FLAGS.MAXGPA_VIRT (bit 4) to 0, as an input to TDH.MNG.INIT, then MAXGPA virtualization is disabled:
  - TDCS.VIRT_MAXPA (the value used by the TDX module) and virtual CPUID(0x80000008).EAX[7:0] are set depending on the value of TD_PARAMS.CONFIG_FLAGS.MAXPA_VIRT (bit 3), as described above.
  - Virtual CPUID(0x80000008).EAX[23:16] is set to 0.
- Else (TD_PARAMS.CONFIG_FLAGS.MAXGPA_VIRT (bit 4) is set to 1), then:
  - TD_PARAMS.CONFIG_FLAGS.MAXPA_VIRT (bit 3) must be 0.
  - Virtual CPUID(0x80000008).EAX[7:0] is set to the native value at the time of TD initialization by TDH.MNG.INIT.
  - TDCS.VIRT_MAXPA (the value used by the TDX module) and virtual CPUID(0x80000008).EAX[23:16] are set to the smaller value between the native MAXPA value (native CPUID(0x80000008).EAX[7:0]) and of the TD's configured GPAW (which can be either 52 or 48).

GPA Space Implications of MAXPA and MAXGPA Virtualization

TDCS.VIRT_MAXPA, the value used by the TDX module, is set as described above. The TDX module considers any GPA parameter where bit TDCS.VIRT_MAXPA or higher is 1 as illegal. Among other things, this prevents the host VMM from building Secure EPT with entries associated with a GPA that is above the range allowed by the TDCS.VIRT_MAXPA.

For shared GPA, the host VMM manages the Shared EPT; it is expected to properly virtualize the shared GPA space size. It should never map pages in a GPA range that is not allowed by the TDCS.VIRT_MAXPA setting.

11.15.1.4. If the TDX module supports TDX Connect, then TDH.MEM.SHARED.SEPT.WR is used by the host VMM to set shared GPA entries in Secure EPT pages. The TDX module enforces TDCS.VIRT_MAXPA for the requested shared GPA.

Exceptions Related to GPA Reserved Bits

Address translation with any of the reserved bits of GPA set to 1 causes an exception injection to the guest TD. This includes the following cases:

- GPA bits higher than the SHARED bit are considered reserved and must be 0.

11.15.1.5.

- GPA bits higher than the virtual value of MAXPA, as enumerated to the guest TD by CPUID(0x80000008).EAX[7:0], are considered reserved and must be 0. An exception to this case is the SHARED bit which is never reserved; if TDCS.VIRT_MAXPA is lower than GPAW, e.g., TDCS.VIRT_MAXPA is 46 and the configured GPAW is 48, then the SHARED bit (at position 47) may be set to 1 indicate a shared GPA.
- For L1, the behavior is as follows:
  - If MAXGPA virtualization, as described above, is configured, then:
    - TDCS.VIRT_MAXPA is enumerated to the guest TD by virtual CPUID(0x80000008).EAX[23:16].
    - If any GPA bit not lower than the physical MAXPA, as enumerated by CPUID(0x80000008).EAX[7:0] but excluding the SHARED bit, is set to 1, the injected exception is a #PF with PFEC (Page Fault Error Code) RSVD bit set.
    - Else (GPA bit not lower than the TDCS.VIRT_MAXPA but lower than the physical MAXPA is set to 1), then the injected exception is #VE(RESERVED_GPA_BITS).
  - Else:
    - TDCS.VIRT_MAXPA is enumerated to the guest TD by virtual CPUID(0x80000008).EAX[7:0].
    - The injected exception is a #PF with PFEC (Page Fault Error Code) RSVD bit set.
- For L2, a reserved bits violation causes an L2→L1 exit with an EPT Violation exit reason.

11.15.2. EPT Violation Mutated into #VE

An EPT violation is converted into #VE in the following cases:

- For Secure EPT, see 9.5.2 for details.
- For Shared EPT, if the EPT entry has been configured by host VMM deliver EPT violations to the guest TD as #VE(ARCH) exceptions for usages such as MMIO, as described in 11.14 above.
- On reserved bits violation, if the MAXGPA virtualization mode is configured, as described in 11.15 above.

11.16. Prevention of TD-Induced Denial of Service

VMs, including TDs, can exploit Intel ISA characteristics to cause performance and functional Denial of Service (DOS) to the VMM. The Intel architecture has several mechanisms that help prevent such DOS cases. This section describes how those mechanisms are used in the context of TDX.

11.16.1. Bus Lock Detection by the TD OS

The guest TD OS can enable debug exception traps due to bus locks by setting IA32_DEBUGCTL.BUS_LOCK_DETECT bit (2), which is disabled by default. When enabled, the feature works identically to how it functions in legacy VMX non-root mode or in non-VMX mode. The IA32_DEBUGCTL MSR and DR6 are part of the state that is saved and restored on VM exit and VM entry, respectively. If the delivery of #DB was pre-empted by a trap-like VM exit, then the pending debug exceptions (including due to BUS_LOCK_DETECT if pending) are saved in TD VMCS and restored on subsequent VM Entry. For fault-like VM Exit due to conditions such as EPT violation and EPT misconfiguration that are encountered during execution of an instruction, there is no pending debug exception recorded, including the bus lock debug exception.

11.16.2. Impact of MSR_MEMORY_CTRL (MSR 0x33)

The host VMM can set bits in MSR_MEMORY_CTRL (MSR 0x33, formerly named MSR_TEST_CTRL) to cause exceptions in VMs (including TDs) in case of bus locks:

- Bit 28 (UC_LOCK_DISABLE): If this bit is set to 1, a UC load lock will trigger a fault which depends on the CPU:
  - Older CPUs will generate a #GP(0) fault. This is enumerated by IA32_CORE_CAPABILITIES[4] value of 1 and CPUID(7,2).EDX[6] value of 0.
  - Newer CPUs will generate an #AC(0) fault. This is enumerated by IA32_CORE_CAPABILITIES[4] value of 0 and CPUID(7,2).EDX[6] value of 1.
- Bit 29 (SPLIT_LOCK_DISABLE): If set to 1, a split lock will trigger an #AC fault.

MSR 0x33 is not virtualizable; it is a core-scope MSR and may be modified by the host VMM on one SMT thread while another SMT thread is running a TD VCPU. The TDX module does not allow a guest TD to access this MSR (a #VE(NON_CONFIGURABLE_PARAVIRT) is generated).

To avoid any security issues, **a correctly written TD OS should always be ready to handle #AC and #GP(0) faults** if the TD software might cause UC locks or split locks.

11.16.3. Bus Lock TD Exit

Bus lock TD exit is disabled by default. The host VMM can enable the TD VMCS "bus-lock detection" VM execution control using the TDH.VP.WR function.

Bus Lock VM Exit Reason (74)

If "bus-lock detection" is enabled, then if the processor detects that one or more bus locks were caused by the instruction that was executed, then the processor generates a bus lock VM exit (exit reason 74). This VM exit is trap-like, i.e., it is delivered following the execution of that instruction that caused it. The Intel TDX module then completes a TD exit with the exit information provided in the VM exit.

Bus Lock Detected Bit (26) in VM Exit Reason

If delivery of bus lock VM exit was pre-empted by a higher priority VM exit (e.g., EPT Misconfiguration, EPT Violation, etc.), then the procession sets a "bus lock detected" notification bit (bit 26) in the exit reason. The Intel TDX module reflects this bit to the host VMM on TD exit.

11.16.4. Instruction Timeout TD Exit

Instruction Timeout TD exit is disabled by default. The host VMM can write the TD VCMS "Instruction Timeout Control" and "Instruction Timeout" execution controls using the TDH.VP.WR function. If enabled and configured, then if the processor detects a no-commit case, the processor causes a notification VM exit (exit reason 75) which the Intel TDX module converts to the TD exit.

The conditions that cause an instruction timeout TD exit are the same as those in legacy VMX non-root mode. An example of such a case is the nested #AC exception. If an #AC exception occurs during the delivery of a previous #AC exception, then the CPU may get into an endless loop of #AC without responding to external events.

Bit 0 (VM context invalid) of the exit qualification indicates whether the guest TD context is corrupted and not valid in the TD VMCS. If this bit is set to 1, then it is a non-recoverable situation; thus, the Intel TDX module marks the TD as disabled to help prevent further TD entry. If no TD context corruption occurred (exit qualification bit 0 is cleared to 0), then the TD may be resumed normally.

11.16.5. Denial of Service due to Long Latency Guest-Side Interface Functions

The TDX module limits the rate at which some guest-side interface functions may be called by the guest TD. See 18.7.5 for details.

11.17. Time Stamp Counter (TSC)

Intel SDM, Vol. 3, 10.5.4.1	TSC-Deadline Mode
Intel SDM, Vol. 3, 18.17	Time-Stamp Counter
Intel SDM, Vol. 3, 24.6.5	Time-Stamp Counter Offset and Multiplier
Intel SDM, Vol. 3, 25.3	Changes to Instruction Behavior in VMX Non-Root Operation

11.17.1. TSC Virtualization

For virtual time stamp counter (TSC) values read by guest TDs, the Intel TDX module is designed to achieve the following:

- Virtual TSC values are consistent among all the TD's VCPUs at the level supported by the CPU, see below.
- The virtual TSC value for any single VCPU is monotonously incrementing (except roll over from 2^{64}-1 to 0).
- The virtual TSC frequency is determined by TD configuration.

The host VMM is required to do the following:

- Set up the same IA32_TSC_ADJUST values on all LPs before initializing the Intel TDX module.
- Make sure IA32_TSC_ADJUST is not modified from its initial value before calling SEAMCALL.

The Intel TDX module checks the above as part of TDH.VP.ENTER and any other SEAMCALL leaf function that reads TSC.

The virtualized TSC is designed to have the following characteristics:

- The virtual TSC frequency is specified by the host VMM as an input to TDH.MNG.INIT in units of 25MHz – it can be between 4 and 400 (corresponding to a range of 100MHz to 10GHz).
- The virtual TSC starts counting from 0 at TDH.MNG.INIT time.
- TSC parameters are enumerated to the guest TD by CPUID(0x15).
- Guest TDs are not allowed to modify the TSC. WRMSR attempts of IA32_TIME_STAMP_COUNTER result in a #VE(NON_CONFIGURABLE_PARAVIRT).
- Guest TDs are not allowed to access IA32_TSC_ADJUST because its value is meaningless to them. If the TDX module supports #VE reduction, as enumerated by TDX_FEATURES0.VE_REDUCTION (bit 30), and the guest TD has set TD_CTLS.REDUCE_VE to 1, then WRMSR or RDMSR attempts result in a #GP(0). Else, WRMSR or RDMSR attempts result in a #VE(CONFIGURABLE_PARAVIRT).
- RDTSCP is supported. This instruction returns the contents of the IA32_TSC_AUX MSR in RCX. The Intel TDX module allows the guest TD to access that MSR and context-switches it on TD entry and exit as part of the VCPU state in TDVPS.

11.17.2. TSC Deadline

Guest TDs are not allowed to access the IA32_TSC_DEADLINE MSR directly. Virtualization of IA32_TSC_DEADLINE depends on the virtual value of CPUID(1).ECX[24] bit (TSC Deadline). The host VMM may configure (as an input to TDH.MNG.INIT) virtual CPUID(1).ECX[24] to be a constant 0 or allow it to be 1 if the CPU's native value is 1.

If the TDX module supports #VE reduction, as enumerated by TDX_FEATURES0.VE_REDUCTION (bit 30), and the guest TD has set TD_CTLS.REDUCE_VE to 1, it may control the value of virtual CPUID(1).ECX[24] by writing TDCS.FEATURE_PARAVIRT_CTRL.TSC_DEADLINE. See 11.2.2 for details.

- If the virtual value of CPUID(1).ECX[24] is 0, IA32_TSC_DEADLINE is virtualized as non-existent. WRMSR or RDMSR attempts result in a #GP(0).
- If the virtual value of CPUID(1).ECX[24] is 1, WRMSR or RDMSR attempts result in a #VE(CONFIG_PARAVIRT). This enables the TD's #VE handler to para-virtualize the TSC deadline functionality, e.g., by requesting an (untrusted) service from the host VMM.

11.18. KeyLocker (KL)

Enumeration: TDX module support of KeyLocker is enumerated by the KL bit (31) of ATTRIBUTES_FIXED0/1 fields, readable by TDH.SYS.RD*.

11.18.1. KeyLocker Virtualization

Guest TDs usage of KeyLocker (KL) is controlled by the ATTRIBUTES.KL bit (see the [TDX Module ABI Spec]). When KL is supported by the CPU and ATTRIBUTES.KL is set to 1, the following KL features are available to the guest TD:

- CPUID virtualization enumerates KeyLocker availability to the guest TD. Virtual CPUID(0x19) values can be configured by the host VMM.
- Guest TDs may enable KeyLocker by setting CR4.KL flag.
- Guest TDs may create KL handles using the ENCODEKEY* instructions, use them using AES*KL instructions, and load Internal Wrapping Keys (IWKs) using the LOADIWKEY instruction.

The following KeyLocker features are not supported:

- Guest TDs may not use the KeyLocker backup MSRs.

11.18.2. Host VMM KeyLocker State Restoration after TDH.VP.ENTER

If the host VMM is using KL and the guest TD's ATTRIBUTES.KL is set to 1, the host VMM **must restore its own IWK after every TDH.VP.ENTER** if the IWK has changed. To avoid unnecessary IWK restore, the host VMM can check if IWK has been changed, as follows:

1. After loading a new IWK, encode key 0 and save the resulting handle H.
2. After successful TDH.VP.ENTER to a guest TD which is allowed to use KeyLocker, encode key 0 again.
   2.1. If the resulting handle is the same as H, then the VMM does not need to reload its IWK.
   2.2. Else, the host VMM needs to restore IWK, depending on its type:
      2.2.1. If the IWK is "direct", then IWK restore can be done using LOADIWKEY instruction.
      2.2.2. If the IWK is "randomized", then IWK restore is done using KL backup MSRs.

11.19. Software Code Prefetch

SW code prefetch is enumerated by CPUID(0x7,0x1).EDX[14]. If both the TDX module and the CPU support it, the host VMM can configure the virtual value of CPUID(0x7,0x1).EDX[14] to be 1, as part of the CPUID configuration parameters of TDH.MNG.INIT.

Note: It is the TD's responsibility to execute PREFETCHIT0 and PREFETCHIT1 only if the virtual value of CPUID(0x7,0x1).EDX[14] is 1. This is not enforced by TDX.

11.20. User MSR

Intel Extended Inst.	URDMSR and UWRMSR instruction specifications

CPU support of User MSRs is enumerated by CPUID(0x7,0x1).EDX[15]. If both the TDX module and the CPU support it, the host VMM can configure the virtual value of CPUID(0x7,0x1).EDX[15] to be 1, as part of the CPUID configuration parameters of TDH.MNG.INIT.

- When the User MSR feature is enabled for a guest TD, it can access the IA32_USER_MSR_CTL MSR (0x1C).
- IA32_USER_MSR_CTL is always cleared on TD exit, regardless of whether this feature is enabled for the guest TD.

11.21. FRED

The host VMM can configure FRED as available to the TD, if both the TDX module and the CPU support FRED, by configuring both applicable CPUID bits (CPUID(7, 1).EAX[17] (FRED) and CPUID(7, 1).EAX[18] (LKGS), as 1 as part of the CPUID configuration during TDH.MNG.INIT.

If the virtual values of CPUID(7, 1).EAX[17] (FRED) and CPUID(7, 1).EAX[18] (LKGS) are both 1, then:

- Virtual CPUID values enumerate FRED & LKGS availability to the guest TD.
- Guest TDs may enable FRED by setting CR4[32] (FRED).
- Guest TDs may access FRED MSRs.
- Guest TDs may execute FRED & LKGS ISA.

11.22. Supervisor Protection Keys (PKS)

By design, guest TD usage of Supervisor Protection Keys (PKS) is controlled by the ATTRIBUTES.PKS bit (see the [TDX Module ABI Spec]). When PKS is supported by the CPU and ATTRIBUTES.PKS is set to 1, the following features are available to the guest TD:

- CPUID virtualization enumerates PKS availability to the guest TD.
- Guest TDs may enable PKS by setting CR4.PKS flag.
- Guest TDs may access the PKS state using the IA32_PKRS MSR.

Note: Enumeration of User Mode Protection Keys (PKU) availability to the guest TD is configured as part of the configuration of virtual CPUID(7,0).ECX[3].

11.23. Intel® Total Memory Encryption (Intel® TME) and Multi-Key Total Memory Encryption (MKTME)

Guest TDs may not directly use the Intel TME and MKTME MSRs and the PCONFIG instruction. The Intel TDX module supports para-virtualization of this ISA, as described below.

11.23.1. TME Virtualization

TME is enumerated by CPUID(0x7, 0x0).ECX[13]. The host VMM can configure the virtualization of this bit as enabled or disabled on TDH.MNG.INIT.

If the TDX module supports #VE reduction, as enumerated by TDX_FEATURES0.VE_REDUCTION (bit 30), and the guest TD has set TD_CTLS.REDUCE_VE to 1, it may control the value of virtual CPUID(0x7, 0x0).ECX[13] by writing TDCS.FEATURE_PARAVIRT_CTRL.TME. See 11.2.2 for details.

If enabled, then a guest TD access to the IA32_TME_* MSRs (0x981 – 0x984) causes a #VE, allowing the guest TD's #VE handler to emulate the desired operation. Else, guest TD access to those MSRs causes a #GP(0).

11.23.2. MKTME Virtualization

MKTME is enumerated by CPUID(0x7, 0x0).EDX[18]. The host VMM can configure the virtualization of this bit as enabled or disabled on TDH.MNG.INIT.

If the TDX module supports #VE reduction, as enumerated by TDX_FEATURES0.VE_REDUCTION (bit 30), and the guest TD has set TD_CTLS.REDUCE_VE to 1, it may control the value of virtual CPUID(0x7, 0x0).EDX[30] by writing TDCS.FEATURE_PARAVIRT_CTRL.PCONFIG. See 11.2.2 for details.

If enabled, then the following operations cause a #VE(CONFIG_PARAVIRT). The guest TD's #VE handler may then communicate with the host VMM over TDG.VP.VMCALL to request the desired operation.

- Guest TD access to the IA32_MKTME_PARTITIONING MSR (0x87)
- PCONFIG execution by the guest TD

If the host VMM or guest TD configured CPUID(0x7, 0x0).EDX[18] virtualized value as 0, then:

- Guest TD access to the IA32_MKTME_PARTITIONING MSR (0x87) causes a #GP(0).
- PCONFIG execution by the guest TD causes a #UD.

11.24. Virtualization of Machine Check Capabilities and Controls

Although the guest TD is not allowed to handle machine check events, the following virtualization is used in order to allow possible pare-virtualization behavior, e.g., future handling of MCE by the TD.

By default, the behavior is as follows:

- The values of CPUID(1).EDX[7] (MCE) and CPUID(1).EDX[14] (MCA), as seen by the guest TD, are 1.
- The value of CR4[6] (MCE), as seen by the guest TD, is 1. Guest TD attempt to set this bit to 0 results in a #VE.
- Guest TD accesses to MSRs 0x179 (IA32_MCG_CAP), MSRs 0x17A, 0x17B, 0x4D0 (IA32_MCG_*), MSRs 0x281 through 0x29D (IA32_MCx_CTL2) and MSRs 0x400 through 0x473 (IA32_MCx_*) result in a #VE(CONFIG_PARAVIRT).

If the TDX module supports #VE reduction, as enumerated by TDX_FEATURES0.VE_REDUCTION (bit 30), and the guest TD has set TD_CTLS.REDUCE_VE to 1, it may control the behavior by writing TDCS.FEATURE_PARAVIRT_CTRL.MCA. See 11.2.2 for details.

If TDCS.FEATURE_PARAVIRT_CTRL.MCA is 0 (default), then:

- The values of CPUID(1).EDX[7] (MCE) and CPUID(1).EDX[14] (MCA), as seen by the guest TD, are 0.
- The value of CR4[6] (MCE), as seen by the guest TD, is initialized to 1. The guest TD may clear CR4.MCE but not set it back to 1; an attempt to do so results in a #GP(0).
- Guest TD accesses to MSRs 0x179 (IA32_MCG_CAP), MSRs 0x17A, 0x17B, 0x4D0 (IA32_MCG_*), MSRs 0x281 through 0x29D (IA32_MCx_CTL2) and MSRs 0x400 through 0x473 (IA32_MCx_*) result in a #GP(0).

If TDCS.FEATURE_PARAVIRT_CTRL.MCA is 1, then:

- The values of CPUID(1).EDX[7] (MCE) and CPUID(1).EDX[14] (MCA), as seen by the guest TD, are configured by the host VMM.
- The value of CR4[6] (MCE), as seen by the guest TD, is initialized to 1. If virtual CPUID(1).EDX[7] is 0, the guest TD may clear CR4.MCE but not set it back to 1; an attempt to do so results in a #GP(0. Else, guest TD is allowed to modify CR4.MCE.
- If virtual CPUID(1).EDX[14] is 0, guest TD accesses to MSRs 0x179 (IA32_MCG_CAP), MSRs 0x17A, 0x17B, 0x4D0 (IA32_MCG_*), MSRs 0x281 through 0x29D (IA32_MCx_CTL2) and MSRs 0x400 through 0x473 (IA32_MCx_*) result in a #GP(0). Else, such accesses result in a #VE(CONFIG_PARAVIRT).

11.25. Transactional Synchronization Extensions (TSX)

Intel SDM, Vol. 1, 16	Programming with Intel TSX

The host VMM can enable TSX for a TD by configuring the following CPUID bits as enabled in the TD_PARAMS input to TDH.MNG.INIT:

- CPUID(7,0).EBX[4] (HLE)
- CPUID(7,0).EBX[11] (RTM)

The virtual values of the above bits, as seen by the guest TD, are the bitwise AND of the real values enumerated by the CPU and of the configuration values. To enable TSX for guest TDs, TDX requires the following conditions to be true:

- The virtual values of the HLE and the RTM bits are the same, either 0 or 1.
- The CPU supports the IA32_TSX_CTRL MSR (as enumerated by IA32_ARCH_CAPABILITIES[7]).

Note: If the real value of the HLE bit and the RTM bit are different, the host VMM must configure both virtual values as 0.

If TSX is enabled for the guest TD:

- IA32_TSX_CTRL is accessible by the TD.
- On TD exit:
  - IA32_TSX_CTRL is cleared to 0.
  - On CPUs that support IA32_TSX_STORE_ADDRESS (MSR 0xF3D), as indicated by IA32_PERF_CAPABILITIES.TSX_ADDRESS[18], if the TD's ATTRIBUTES.PERFMON is 1 then IA32_TSX_STORE_ADDRESS is cleared to 0.

  The host VMM is responsible for restoring these MSRs to their desired values, if applicable.

If TSX is disabled for the guest TD:

- CPUID(7,0).EBX bits 4 and 11 are virtualized as 0.
- IA32_TSX_CTRL is virtualized as non-existent: IA32_ARCH_CAPABILITIES bit 7 is virtualized as 0, and TD access results in a #GP(0).
- If IA32_TSX_CTRL is supported by the CPU, then XBEGIN, XEND and XABORT instructions execution by the TD cause a #UD.

11.26. Management of Idle and Blocked Conditions

Intel SDM, Vol. 3, 9.10 Management of Idle and Blocked Conditions

11.26.1. HLT Instruction

HLT executed by a guest TD results in a #VE(NON_CONFIG_PARAVIRT). The TD's #VE handler may notify the host VMM (using TDG.VP.VMCALL), which may schedule other software to execute on the current LP.

11.26.2. PAUSE Instruction and PAUSE-Loop Exiting

Intel SDM, Vol. 3, 25.1.3 Instructions That Cause VM Exits Conditionally

Guest TDs can execute PAUSE. However, modern enlightened guests use a VMM-provided service (hypercall) instead of PAUSE loops – this is the expected usage for Intel TDX.

For TDs running in debug mode (ATTRIBUTES.DEBUG is 1), the host VMM may set the guest TD's "PAUSE-loop exiting" VM-execution control, using TDH.VP.WR.

"PAUSE-loop exiting" allows the VMM to request an exit if the guest (in ring 0) executes PAUSE in a loop (e.g., busy-wait). This is intended to help avoid cases where a guest thread loops, waiting for another thread that is not currently scheduled by the VMM.

11.26.3. MONITOR and MWAIT Instructions

By default, guest TDs are expected not to use MONITOR/MWAIT. The virtual value of CPUID(1).ECX[3] is, by default, 0. Execution of MONITOR or MWAIT by a guest TD results in a #UD exception.

However, the host VMM may configure the guest TD to allow MONITOR/MWAIT, using the CPUID configuration table which is part of the TD_PARAMS input to TDH.MNG.INIT. Configuring the virtual value of CPUID(1).ECX[3] to 1 also enables the TD to execute MONITOR and MWAIT.

11.26.4. WAITPKG: TPAUSE, UMONITOR and UMWAIT Instructions

The host VMM may allow guest TDs to use the TPAUSE, UMONITOR and UMWAIT instructions, if the CPU supports them, by configuring the virtual value of CPUID(7,0).ECX[5] (WAITPKG) to 1 using the CPUID configuration table which is part the TD_PARAMS input to TDH.MNG.INIT. Enabling CPUID(7,0).ECX[5] also enables TD access to IA32_UMWAIT_CONTROL (MSR 0xE1).

If not allowed, then TD execution of TPAUSE, UMONITOR or UMWAIT results in a #UD, and access to IA32_UMWAIT_CONTROL results in a #GP(0).

11.27. Other Changes in SEAM Non-Root Mode

11.27.1. CET

Intel SDM, Vol. 1, 17.2.3 Supervisor Shadow Stack Token

Guest TDs should execute CPUID(7,1) and use the CET_SSS bit value returned in EDX[18] as an indication of whether supervisor shadow stack can be enabled. The TDX module virtualizes CPUID(7,1).EDX[18] as 0 if certain supervisor shadow-stack pushes might cause VM exits, indicating to the guest TD that it should refrain from enabling supervisor shadow stack. For details, see the [Intel SDM].

11.27.2. Tasking

Any task switch results in a VM exit to the Intel TDX module (this is a fixed-1 exit) which then performs a TD exit to the host VMM.

The VMM is expected not to reenter the TD VCPU since this case is non-recoverable; the instruction that caused the task switch (CALL, JMP or IRET) will re-execute and cause another VM exit. If the task switch was incidental to an exception delivery, then the VM entry following TDH.VP.ENTER will reattempt the delivery and cause another task switch VM exit. The expected response from the VMM is to terminate this TD.

12. Measurement and Attestation

12.1. Overview of the Attested Measurements and Configuration Information

The table below summarizes the attested TD measurements and configuration information. For details, see the following sections and the TDINFO_STRUCT definition in the [ABI Spec].

Table 12.1: Attested Measurements and Configuration Information

Field Class	Name	Description
Parameters configured by the host VMM at TD initialization time	**ATTRIBUTES**	TD's ATTRIBUTES
	XFAM	TD's XFAM
	MRCONFIGID	Software-defined ID for non-owner-defined configuration of the guest TD – e.g., run-time or OS configuration
	MRCONFIGSVN	SVN of MRCONFIG[9]
	MROWNER	Software-defined ID for the guest TD's owner
	MROWNERCONFIG	Software-defined ID for owner-defined configuration of the guest TD – e.g., specific to the workload rather than the run-time or OS
	MROWNERCONFIGSVN	SVN of MROWNERCONFIG[10]
Build-time measurement, finalized at the end of TD build	**MRTD**	Measurement of the initial contents of the TD
Build/migration-time measurement register, finalized at the end of TD build and updated on migration	**SERVTD_HASH**	Measurement of the bound service TDs, if any
Run-time measurement registers, updated at run time by the guest TD using TDG.MR.RTMR.EXTEND	**RTMR**	Run-time extendable measurement registers
Fields assigned by the guest TD at run time using TDG.MR.ASSIGNSVNS, based on signature[11]	**MRSIGROOT**	Signer Roots of MRSIGNER
	MRSIGNER	Signer of TD's TDSIGSTRUCT
	PRODID	Product ID
	ISVSVN	SVNs of the TD

[9] Support of MRCONFIGSVN is enumerated by TDX_FEATURES0.SEALING (bit 12), readable by TDG.SYS.RD and TDH.SYS.RD.

[10] Support of MROWNERCONFIGSVN is enumerated by TDX_FEATURES0.SEALING (bit 12), readable by TDG.SYS.RD and TDH.SYS.RD.

[11] Support of TDG.MR.ASSIGNSVNS and the associated fields is enumerated by TDX_FEATURES0.TD_SIGNING_AND_SVN (bit 22), readable by TDG.SYS.RD and TDH.SYS.RD.

12.2. TD Measurement

12.2.1. MRTD: Build-Time Measurement Register

The Intel TDX module measures the TD during the build process. The measurement register TDCS.MRTD is a SHA384 digest of the build process, designed as follows:

- TDH.MNG.INIT begins the process by initializing the digest.
- TDH.MEM.PAGE.ADD adds a TD private page to the TD and inserts its properties (GPA) into the MRTD digest calculation.
- Control structure pages (TDR, TDCX and TDVPR) and Secure EPT pages are not measured.
- For pages whose data contributes to the TD, that data should be included in the TD measurement via TDH.MR.EXTEND. TDH.MR.EXTEND inserts the data contained in those pages and its GPA, in 256-byte chunks, into the digest calculation. If a page will be wiped and initialized by TD code, the loader may opt not to measure the initial contents of the page with TDH.MR.EXTEND.
- The measurement is then completed by TDH.MR.FINALIZE. Once completed, further TDH.MEM.PAGE.ADDs or TDEXTENDs will fail.

MRTD extension by GPA uses a 128B buffer which includes the GPA and the leaf function name for uniqueness.

12.2.2. RTMR: Run-Time Measurement Registers

The RTMR array is initialized to zero on build, and it can be extended at run-time by the guest TD using the TDCALL(TDG.MR.RTMR.EXTEND) leaf. The syntax of the RTMR registers is designed to be similar to that of TPM PCRs, where a register's value after TDG.MR.RTMR.EXTEND(index=i, value=x) is:

```
    RTMR[i] = SHA384(RTMR[i] || x);
```

Four RTMR registers are provided.

Typical expected usage is for TPM emulation during guest TD OS secure boot by the VBIOS.

12.2.3. SERVTD_HASH: Service TDs Measurement Register

See the discussion in 13.2.7 for details.

12.3. Security Version Numbers and Signer Attestation

If supported, guest TDs can have two groups of SVNs: statically set at TD initialization time and dynamically set at TD runtime.

12.3.1. Static SVNs

MRCONFIGSVN and MROWNERCONFIGSVN are set at TD initialization time, as part of the TD_PARAMS input to TDH.MNG.INIT, similarly to their corresponding field MRCONFIG and MROWNERCONFIG.

Enumeration: Support of MRCONFIGSVN and MROWNERCONFIGSVN is enumerated by TDX_FEATURES0.SEALING (bit 12), readable by TDG.SYS.RD and TDH.SYS.RD.

12.3.2. Dynamic SVNs

Dynamic SVNs are assigned at TD runtime; they may be associated with software that is measured during TD boot, for example values in the RTMRs. Dynamic SVNs are set using the TDG.MR.ASSIGNSVNS guest-side interface function, which is provided with a TDSIGSTRUCT containing the SVN and additional information. See the [ABI Spec] for details.

Enumeration: Support of TDG.MR.ASSIGNSVNS and the associated fields is enumerated by TDX_FEATURES0.TD_SIGNING_AND_SVN (bit 22), readable by TDG.SYS.RD and TDH.SYS.RD.

The TD's build time measurement (MRTD) and runtime measurements (RTMRs) can be used to measure software in the TD that composes multiple software layers (for example, BIOS, OS, application apace, etc.). Each layer can be defined to use a subset of the values in MRTD/RTMRs. The TDSIGSTRUCT input to TDG.MR.ASSIGNSVNS specifies which of TDMR and RTMRs is applicable, and a set of PRODID, ISVSVN, SIGNER, and SIGROOT field.

Theoretically, multiple layers can be supported, each with its own dynamic SVN assigned by its own TDSIGSTRUCT. However, TDX currently supports only one SVN layer.

12.4. TD Measurement Reporting

TD attestation is initiated from inside the TD by calling TDG.MR.REPORT and specifying a REPORTDATA value. TDG.MR.REPORT creates a TDREPORT_STRUCT structure containing the following fields:

CPUSVN: Current SVN of the CPU running the TD.

TEE_TCB_INFO: Information about the TDX module running the TD:

- SVN and measurement of the TDX module at the time of TD creation on the current platform.
- SVN of the current TDX module on the current platform.

TDINFO_STRUCT: Information about the guest TD:

- TD measurements and initial configuration of the TD, calculated at TD build finalization time (TDH.MR.FINALIZE).
- Hash of service TDs, calculated at TD build finalization time (TDH.MR.FINALIZE) but updated on TD migration.
- Run-time measurement registers, updated by the guest TD using TDG.MR.RTMR.EXTEND.
- If supported, fields updated by the guest TD using TDG.MR.ASSIGNSVNS.

REPORTDATA: The caller provided REPORTDATA value.

TDREPORT_STRUCT structure and TDG.MR.REPORT are detailed in the [TDX Module ABI Spec].

TDREPORT_STRUCT is HMAC'ed using an HMAC key that is designed to be accessible only to the CPU. This helps protect the integrity of the structure and, by design, can only be verified on the local platform via the TDG.MR.VERIFYREPORT interface function or the SGX ENCLU(EVERIFYREPORT2) instruction. By design, TDREPORT_STRUCT cannot be verified off platform; it first must be converted into signed Quotes, as described in 12.6 below.

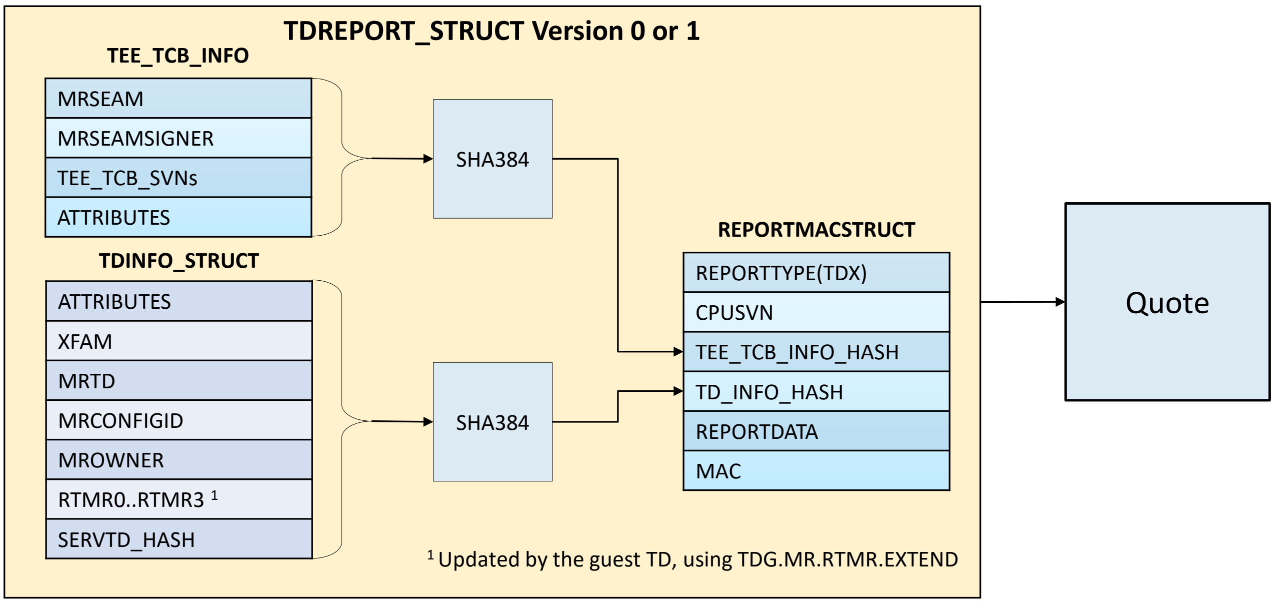


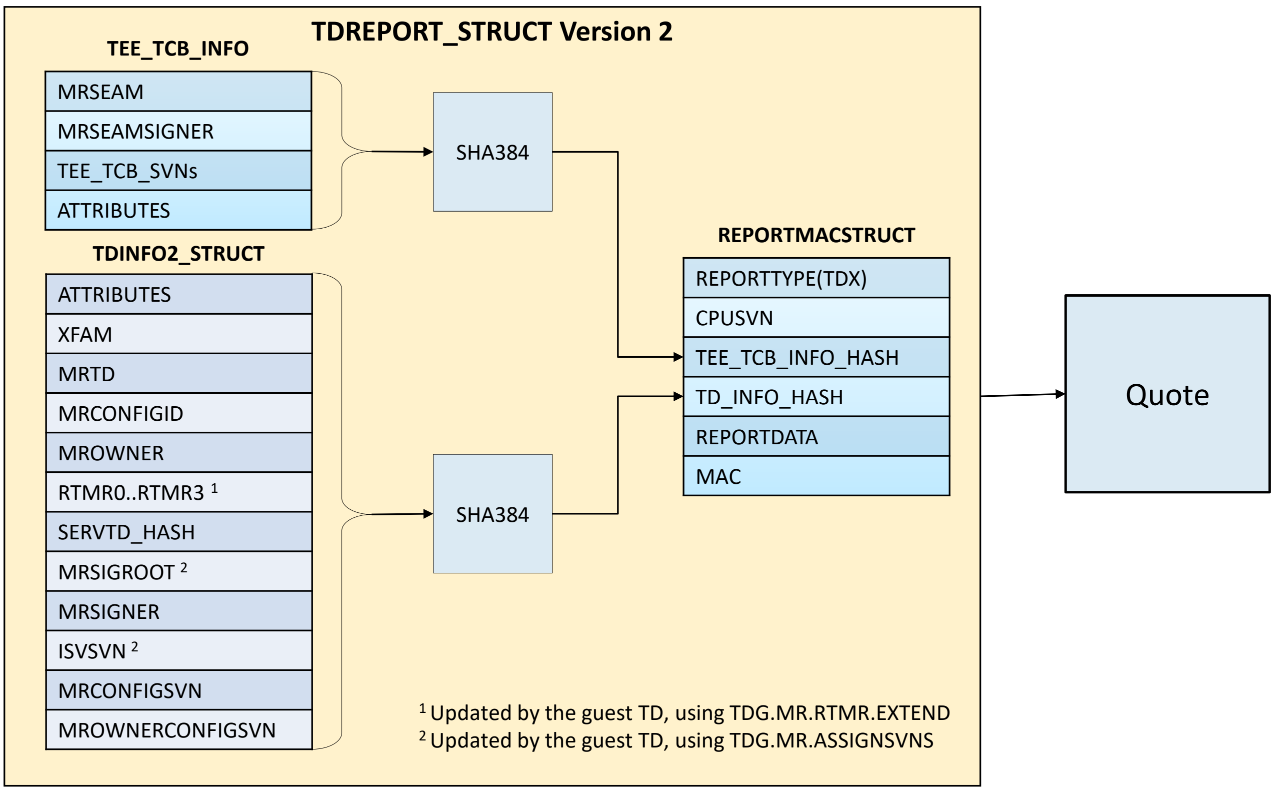
Figure 12.1: TD Measurement Reporting

The REPORTYPE field of REPORTMACSTRUCT indicates that this is a TDX report (TYPE == 0x81) and whether SERVTD_HASH contains a valid hash of service TDs (VERSION == 1) or 0 (VERSION == 0). For details, see Ch. 13 and the [ABI Spec].

TDREPORT_STRUCT Version 2

TDREPORT_STRUCT version 2 is supported if MRCONFIGSVN and MROWNERCONFIGSVN are supported, or if TDG.MR.ASSIGNSVNS is supported.

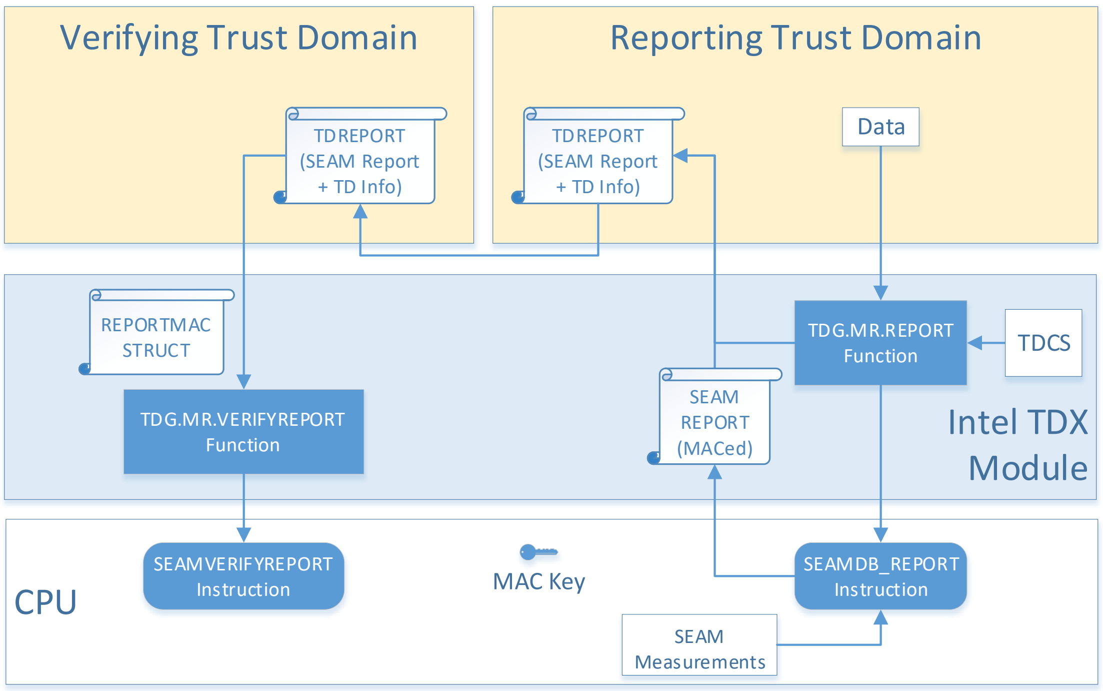
If the host VMM configured MRCONFIGSVN or MROWNERCONFIGSVN to non-0 values, or guest TD called TDG.MR.ASSIGNSVNS with at least one non-0 value, then TDREPORT_STRUCT version 2 is returned by TDG.MR.REPORT, as shown below.



5 **Figure 12.2: TD Measurement Reporting Including TDINFO2_STRUCT**

12.5. Local Report Verification

A TD can verify a report generated by another TD on the same platform using the TDG.MR.VERIFYREPORT interface function. Internally, TDG.MR.VERIFYREPORT executes the SEAMVERIFYREPORT instruction, which uses the same HMAC key, accessible only to the CPU, that was used for generating TDREPORT_STRUCT.



10 **Figure 12.3: High-Level View of Local Report Verification**

Local Report Verification Failure

Local report verification may fail in cases where the MAC key, held by the CPU, has changed between the generation of the TDREPORT by TDG.MR.REPORT and its verification by TDG.MR.VERIFYREPORT. Some examples are:

- After report generation by TDG.MR.REPORT, both the reporting TD and the verifying TD have been migrated to a different platform. Both TDs are not directly aware of the migration.
- After report generation by TDG.MR.REPORT, either the CPU microcode has been updated or the TDX module has been updated using the TD-preserving update process. Both TDs are not directly aware of the TDX module update.

To account for the above cases, the following is recommended in case of local report verification failure. The verifying TD should ask for a fresh report to be generated. The reporting TD should then generate a new report, using TDG.MR.REPORT, and send it to the verifying TD to be verified using TDG.MR.VERIFYREPORT. This can be repeated several times (e.g., 3 – 5 times) and/or coordinated with the host VMM.

12.6. Creating Attestations

12.6.1. Overview

To create a remotely verifiable attestation, the TDREPORT_STRUCT should be converted into a Quote signed by a certified Quote signing key. The following models are supported for creating a Quote:

- Platforms that support Intel SGX can support Quoting Enclaves producing either TDX or SGX Quotes. A TD Quoting Enclave, when available, will produce legacy quotes for TDX.

On platforms that support an enabled Security Engine (ESE/S3M):

- The security engine can be used to create an attestation x509 certificate.
- A Quoting TD can create legacy-style Quotes or x509 certificates. The Quoting TD itself is certified by a Security Engine-based Attestation.

12.6.2. Intel SGX-Based Attestation

The Intel SGX attestation architecture is designed to provide facilities for multiple Quoting Enclaves from multiple providers. This is intended to allow the host to instantiate a Quoting Enclave for Intel SGX attestations and another Quoting Enclave for TD attestation without interference — i.e., each provider can supply its own quoting enclave, and the quoting enclave for Intel SGX and for Intel TDX may be separate; the design does not require the quoting enclave to run inside the TD.

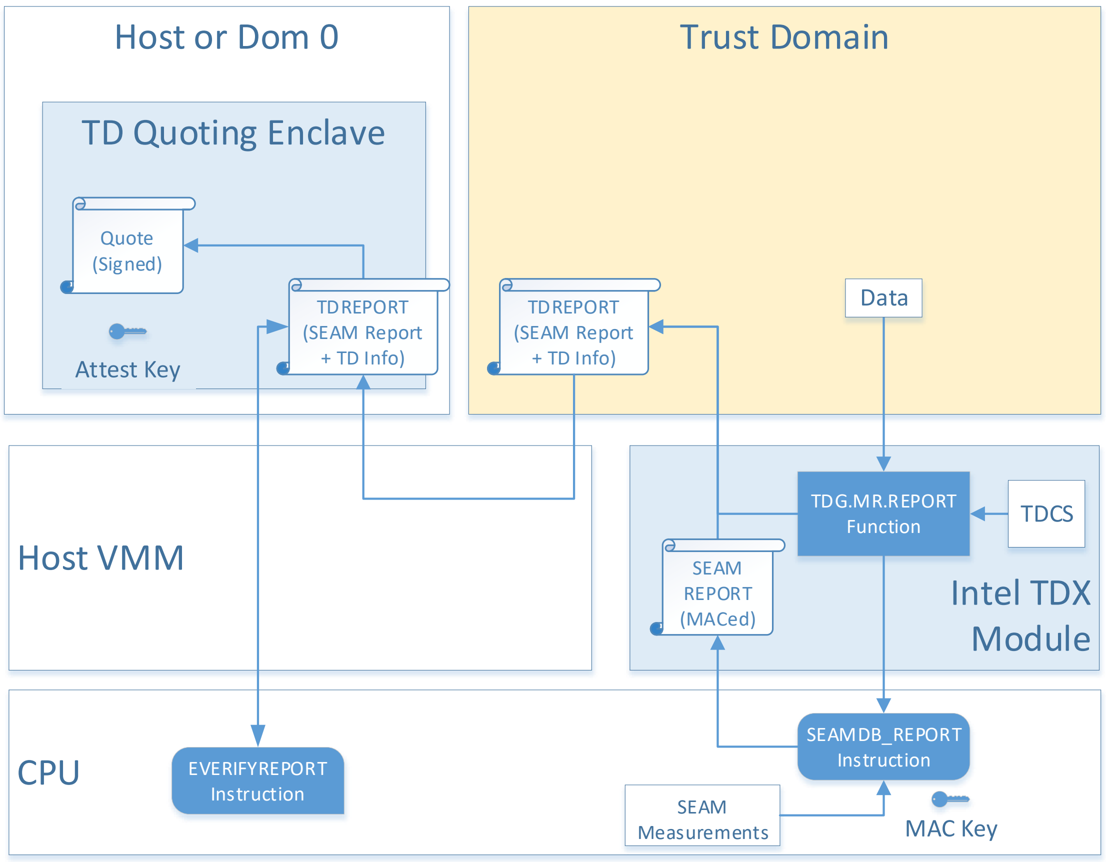


Figure 12.4: High-Level View of the Intel SGX-Based TD Attestation

Quote generation using a quoting enclave is typically performed as follows:

1. Guest TD invokes the TDCALL(TDG.MR.REPORT) function.
2. If the TDX module supports TD-preserving updates, it uses the SEAMOPS(SEAMDB_REPORT) instruction to create MAC'ed TDREPORT_STRUCT with the Intel TDX module measurements from CPU and TD measurements from TDCS. Else, it uses the SEAMOPS(SEAMREPORT) instruction for the same purpose.
3. Guest TD uses TDCALL(TDG.VP.VMCALL) to request that TDREPORT_STRUCT be converted into Quote.
4. The TD Quoting enclave uses EVERIFYREPORT2 to check the TDREPORT_STRUCT. This allows the Quoting Enclave to check the report without requiring direct access to the CPU's HMAC key. Once the integrity of the TDREPORT_STRUCT has been verified, the TD Quoting Enclave signs the TDREPORT_STRUCT body with an ECDSA 384 signing key.

EVERIFYREPORT Failure

Report verification may fail in cases where the MAC key, held by the CPU, has changed between the generation of the TDREPORT by TDG.MR.REPORT and its verification by EVERIFYREPORT. Some examples are:

- After report generation by TDG.MR.REPORT, the reporting TD has been migrated to a different platform. The reporting TD is not directly aware of the migration.
- After report generation by TDG.MR.REPORT, either the CPU microcode has been updated or the TDX module has been updated using the TD-preserving update process. The reporting TD is not directly aware of the TDX module update.

To account for the above cases, the following is recommended in case of report verification failure. The Quoting Enclave should ask for a fresh report to be generated. The reporting TD should then generate a new report, using TDG.MR.REPORT, and send it to the Quoting Enclave to be verified using EVERIFYREPORT. This can be repeated several times (e.g., 3 – 5 times) and/or coordinated with the host VMM.

12.6.2.1. *Quote Signing Key for SGX-Based Attestation*

The Intel SGX infrastructure provides primitives and a certificate infrastructure to allow Quoting Enclaves to certify their own Quoting Keys. The Intel SGX Provisioning Certification Enclave (PCE) uses an Intel-Certified ECDSA-256 signing key to issue certificates to Quoting Enclaves for their attestation keys. Intel offers a service to allow third parties to download these certificates.

Typically, on first launch, the TD Quoting Enclave generates a random ECDSA 384-bit quoting key. It then contacts the Provisioning Certification Enclave which uses its signing key to sign the new quoting key's public key.

Note that the TD Quoting Enclave uses an ECDSA 384 bit key, while the PCE certifies it with an ECDSA-256 key. This is due to the limitations of the SPR platform.

12.6.3. Security Engine-Based Attestation

On supported platforms, security engine-based attestation uses the S3M or ESE for attestation, depending on the platform type as the root for attestation, and a Quoting TD to create attestation at runtime.

A Quoting TD creates attestations similarly to how the SGX Quoting Enclave does. The Quoting TD's attestation key is certified by the security engine, which issues an x509-based attestation.

0. Prior to tenant TDs running, the Quoting TD starts up and generates an Attestation Key. The Quoting TD uses the S3M/ESE to certify its Attestation Pub Key. The certificate chain, rooted in Intel, for the Attestation key will accompany the resulting attestation created by the Quoting TD.
1. Guest TD invokes the TDCALL(TDG.MR.REPORT) function.
2. If the TDX module supports TD-preserving updates, it uses the SEAMOPS(SEAMDB_REPORT) instruction to create MAC'ed TDREPORT_STRUCT with the Intel TDX module measurements from CPU and TD measurements from TDCS. Else, it uses the SEAMOPS(SEAMREPORT) instruction for the same purpose.
3. Guest TD uses TDCALL(TDG.VP.VMCALL) or VSOCK to request that TDREPORT_STRUCT be converted into Quote.
4. The Quoting TD uses TDCALL(TDG.MR.VERIFYREPORT) to verify the TDREPORT_STRUCT. This allows the Quoting TD to verify the report without requiring direct access to the CPU's Report MAC key. Once the integrity of the TDREPORT_STRUCT has been verified, the Quoting TD creates a TD attestation containing the measurements from the TDREPORT_STRUCT body, the QE Certification Data from step 0, and certification information for the ESE/S3M's signing key.

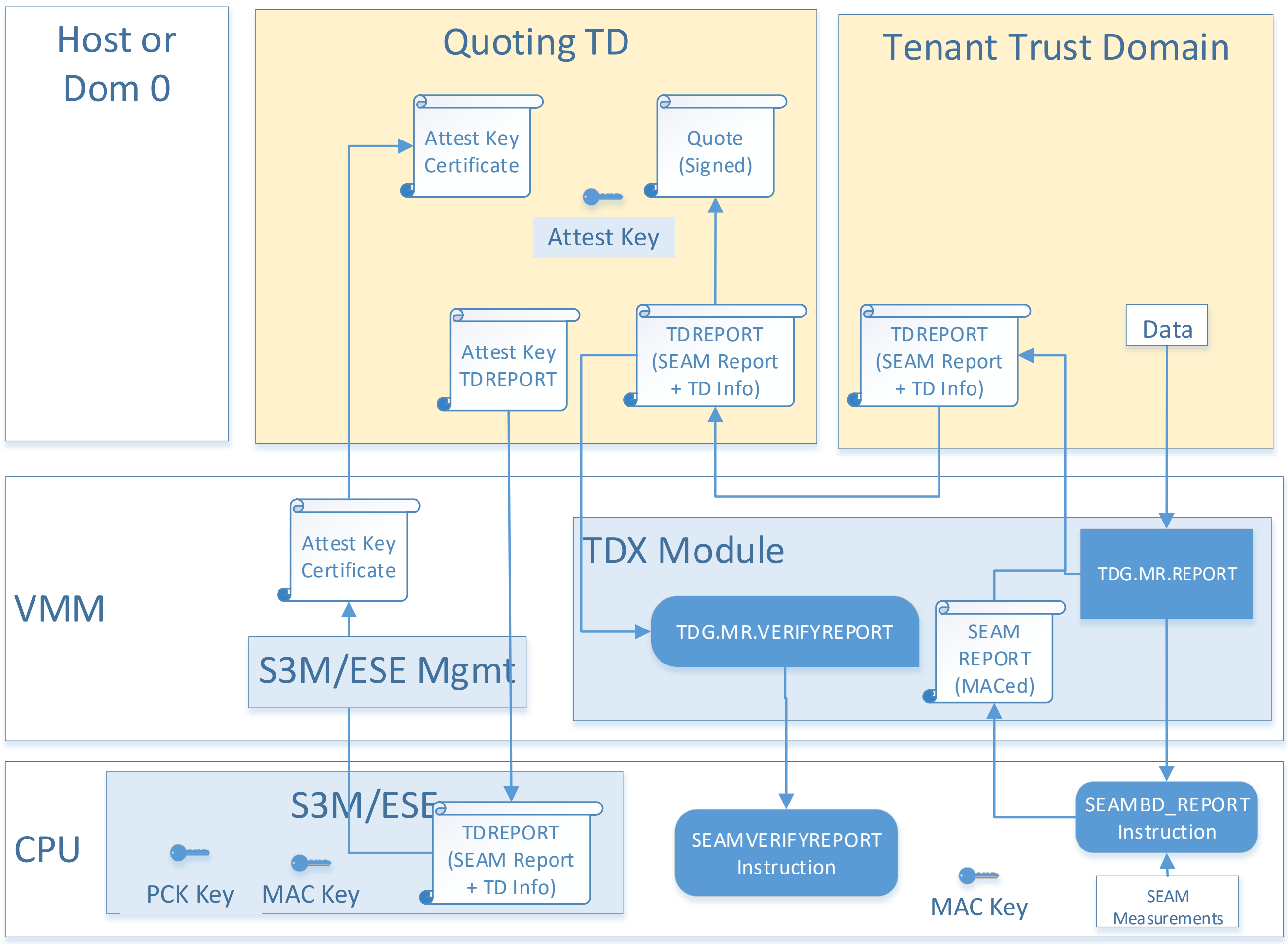


Figure 12.5: Security Engine Based Attestation

12.7. TCB Recovery

The Intel TDX architecture has several levels of TCB:

- CPU HW level, which includes microcode patch, ACMs and PFAT
- Intel TDX module software
- Attestation Enclaves which include the TD Quoting Enclave and Provisioning Certification Enclave

The TCB Recovery story is different for each level. The existing SGX TCB Recovery model for CPU level items applies in the same way with TDX and SGX. The model requires a restart of the platform to take effect. The Intel TDX module can be unloaded and reloaded to reflect an upgraded Intel TDX module. The enclaves can be upgraded at run-time, but if the PCE is upgraded, the design requires a new certificate to be downloaded.

12.7.1. TD Preserving TDX Module Update Implications

TEE_TCB_INFO fields TEE_TCB_SVN and MRSEAM reflect the TDX module at the time of TD creation. TEE_TCB_INFO.TEE_TCB_SVN2 reflects the current TDX module at the time TDG.MR.REPORT is called.

The underlying assumptions are:

- TD preserving update can only happen to a more secure TDX module.
- Microcode updates can only happen to a more secure microcode.

12.8. Sealing

Enumeration: TDX module support of sealing and TDG.MR.KEY.GET is enumerated by TDX_FEATURES0.SEALING (bit 12), readable by TDG.SYS.RD.

The TDG.MR.KEY.GET interface function allows a TD to request a persistent sealing key to be derived for the TD. The key derivation is done according to a key request structure, which specifies a measurement register policy (e.g., which RTMRs are used), TD configuration policy (e.g., which bits of the TD's XFAM and ATTRIBUTES are used) and SVNs.

The sealing key can be used to seal information, i.e., encrypt it so later it can be decrypted by a TD that has a compatible set of measurements, configuration and SVNs as used when deriving the sealing key. TDX sealing keys are not migratable.

For further details, see the [ABI Spec].

13.Service TDs

13.1. Overview

One or more **service TDs** may be bound to a **target TD**. Service TD binding relationship has the following characteristics:

- A service TD has a **type** (SERVTD_TYPE).
- A service TD may **read and/or write certain target TD metadata**. Access permission to target TD metadata fields depends on SERVTD_TYPE.
- **Unsolicited service TD binding** is done without target TD approval. The target TD needs not be aware of the binding.
- The target TD's TDREPORT indicates binding to service TDs.
- The service TD protocol consists of:
  - Binding
  - Metadata access
- Service TD to target TD binding relationship is many-to-many
  - Multiple service TDs of different types may be bound to a single target TD.
  - Multiple target TDs may be bound to a single service TD.
- A service TD may itself be a target TD to other service TDs.

Typical Unsolicited Service TD Binding and Metadata Access Use Case

1. **Optional Pre-Binding:** During target TD build, before calling TDH.MR.FINALIZE, the host VMM calls TDH.SERVTD.PREBIND to write the binding fields (SERVTD_HASH etc.) in the target TD's service TD table.
2. **Binding:** Sometime later, the host VMM calls TDH.SERVTD.BIND to bind the service TD. It gets back a binding handle. The VMM communicates the binding handle, target TD_UUID and other binding parameters to the service TD.
3. **Metadata Access:** The service TD uses TDG.SERVTD.RD/WR* to access target TD metadata.
4. **Rebinding:** May be required due to, e.g., both target TD and service TD have been migrated, or a new service TD instance replaces the original one. The host VMM calls TDH.SERVTD.BIND to rebind the service TD. It gets back a binding handle. The VMM communicates the binding handle, target TD_UUID and other binding parameters to the service TD.

13.2. Service TD Binding

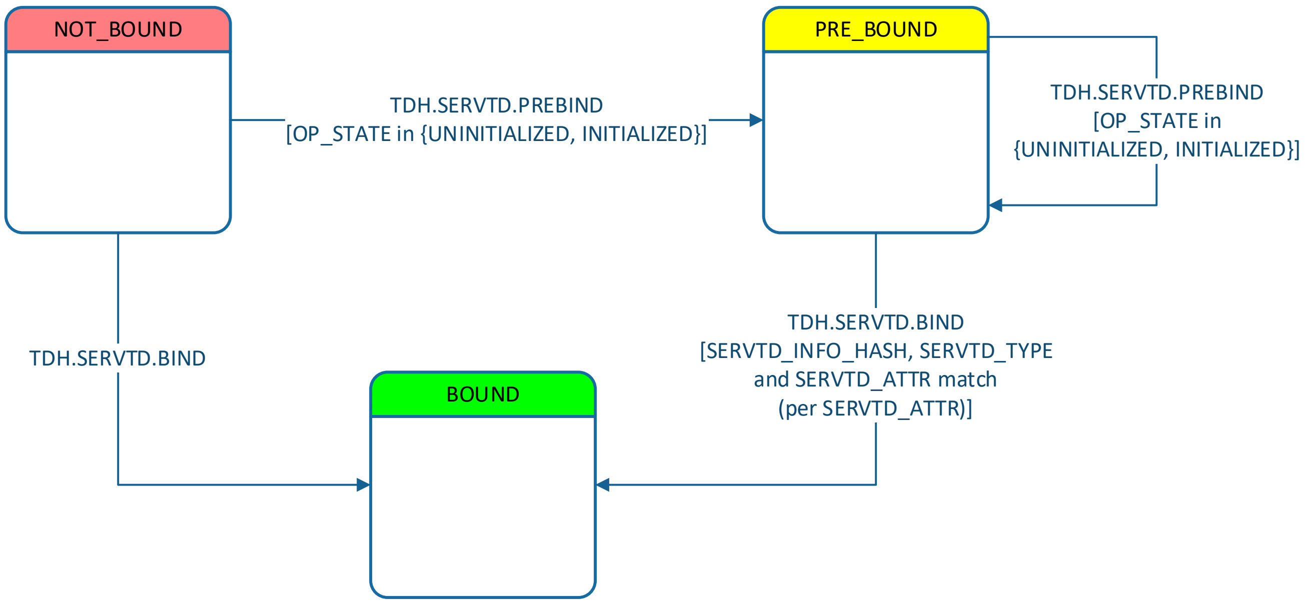


Figure 13.1: Service TD Binding State Machine

13.2.1. Service TD Binding Table in the Target TD's TDCS

The target TD's TDCS holds a service TD binding table. Each row (binding slot) in the table contains the following fields, which are detailed in the following sections:

- SERVTD_BINDING_STATE

- SERVTD_INFO_HASH
- SERVTD_TYPE
- SERVTD_ATTR
- SERVTD_UUID

The available number of slots in the table is enumerated by TDH.SYS.RD*.

13.2.2. SERVTD_BINDING_STATE: Service TD Binding State

SERVTD_BINDING_STATE indicates the state of the service TD binding slot. It has the following values:

Table 13.1: SERVTD_BINDING_STATE Definition

Value	Name	Meaning
0	NOT_BOUND	No service TD is bound. The binding fields in this slot are N/A.
1	PRE_BOUND	No service TD is bound. SERVTD_INFO_HASH, SERVTD_TYPE and SERVTD_ATTR have been set. They will be included in SERVTD_HASH calculation and be checked on any following binding.
2	BOUND	A service TD is bound. SERVTD_UUID, SERVTD_INFO_HASH, SERVTD_TYPE and SERVTD_ATTR have been set and be checked on any following binding. SERVTD_INFO_HASH, SERVTD_TYPE will be included in SERVTD_HASH calculation and be checked on any following binding.

13.2.3. SERVTD_TYPE: Service TD Binding Type

A service TD implements one or more SERVTD_TYPEs. A specific SERVTD_TYPE is specified per binding; the same service TD may be bound multiple times if it implements more than one SERVTD_TYPE.

SERVTD_TYPE controls the following:

- The target TD metadata fields that the service TD may read and/or write.
- Whether or not multiple bindings of this SERVTD_TYPE can exist at the same time for a specific target TD.

SERVTD_TYPE values supported by the TDX module are defined in the [TDX Module ABI Spec].

13.2.4. SERVTD_ATTR: Service TD Binding Attributes

SERVTD_ATTR is a set of service TD binding attributes. Currently, all attribute bits are fixed-0, effectively setting the following attributes. These are used for **Migration TD**, which is currently the only supported service TD type.

Class Binding

Rebinding can be done with any TD with the same SERVTD_INFO_HASH, SERVTD_TYPE and SERVTD_ATTR as the original binding. Those parameters are migrated when the target TD is migrated. SERVTD_UUID is not checked; it is updated by rebinding.

Non-Migratable Binding

Only some of the binding state of the service TD is migrated. A different service TD may be bound at the destination platform, subject to the conditions described below.

SERVTD_BINDING_STATE, SERVTD_TYPE and SERVTD_ATTR are exported as part of the TD's immutable state.

When importing the target TD's mutable state to the destination platform, if the imported SERVTD_BINDING_STATE is PRE_BOUND or BOUND, and there is already a PRE_BOUND or BOUND service TD at this binding slot, then the SERVTD_TYPE must match. The other imported fields for that binding slot are ignored.

If the imported SERVTD_TYPE allows only a single instance of that type, no other service TD slot may have a PRE_BOUND or BOUND service TD of the same SERVTD_TYPE.

SERVTD_HASH is recalculated after all service TD bindings have been imported as part of the immutable TD state.

IGNORE_TDINFO: TDINFO Component Filtering

IGNORE_TDINFO is a bit array which determines which component of the service TD's TDINFO_STRUCT field is included in the calculation of SERVTD_INFO_HASH. For details see 13.2.6 below.

13.2.5. SERVTD_UUID: Service TD Instance Identifier

TD_UUID is a 256-bit random number that serves as a universally unique identifier of a TD. TD_UUID is created by 13.2.4.2. TDH.MNG.CREATE and is stored in the TD's TDR. When a service TD is bound to a target TD, its TD_UUID is stored in the target TD's service TD table slot's SERVTD_UUID field.

13.2.6. Service TD's Binding SERVTD_INFO_HASH Calculation

For the purpose of service TD binding, a SHA384 hash of the service TD's measurable attribute is calculated in a similar way to the calculation done by TDG.MR.REPORT (see 12.4), except that filtering is applied based on the binding SERVTD_ATTR:

- The SERVTD_ATTR.IGNORE_TDINFO selects which TDINFO_STRUCT field is ignored (a value of 0 is used in the calculation).

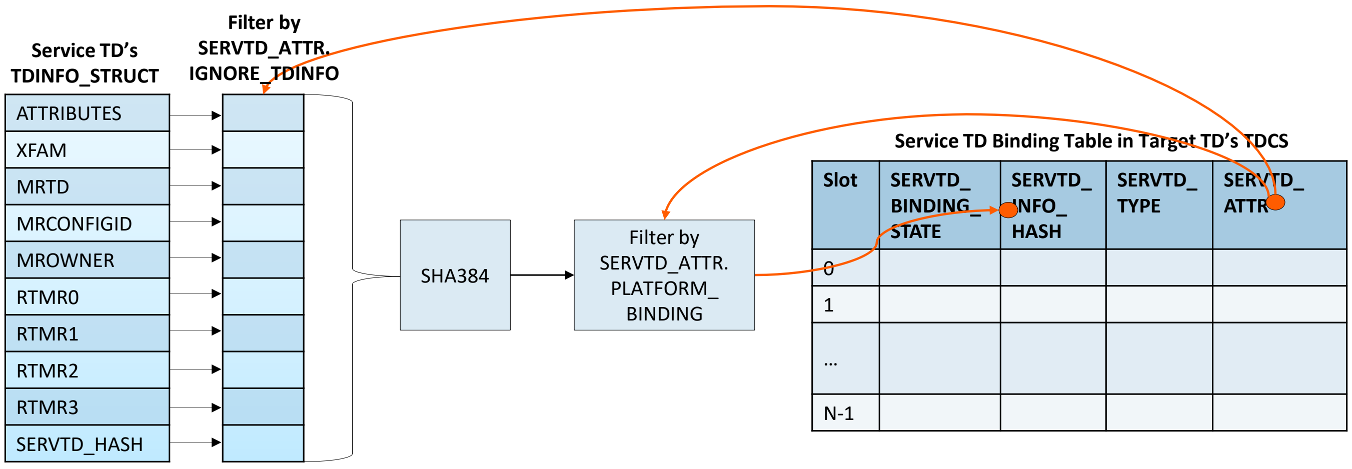

Service TD's TDINFO_STRUCT
Filter by SERVTD_ATTR. IGNORE_TDINFO
ATTRIBUTES
XFAM
MRTD
MRCONFIGID
MROWNER
RTMR0
RTMR1
RTMR2
RTMR3
SERVTD_HASH
SHA384
Filter by SERVTD_ATTR. PLATFORM_ BINDING
Service TD Binding Table in Target TD's TDCS
Slot
SERVTD_ BINDING_ STATE
SERVTD_ INFO_ HASH
SERVTD_ TYPE
SERVTD_ ATTR
0
1
…
N-1

Figure 13.2: SERVTD_INFO_HASH Calculation

13.2.7. Target TD's SERVTD_HASH Calculation

SERVTD_HASH is a single field that summarizes all the service TDs bound or pre-bound to the target TD in an unsolicited 13.2.7.1. mode. SERVTD_HASH is calculated at the end of TD build (by TDH.MR.FINALIZE) and on TD import (by TDH.IMPORT.STATE.IMMUTABLE).

SERVTD_HASH Calculation on Finalization of TD Build

13.2.7.2.

On **TD build,** SERVTD_HASH is calculated by TDH.MR.FINALIZE. At that time, the binding information for all bound or pre-bound service TDs is known.

SERVTD_HASH Calculation on TD Import

On TD import, SERVTD_HASH is recalculated by TDH.IMPORT.STATE.IMMUTABLE. In case of non-migratable service TD binding, the imported binding information is checked but does not replace the existing binding information. E.g., the Migration TD bound on the source platform may have a different INFO_HASH than that of the Migration TD bound on the destination platform. The recalculated SERVTD_HASH reflects the service TDs bound on the destination platform.

The reason for this recalculation is to narrow down the TCB for the migrated TD attestation. E.g., suppose the Migration TD on either or both sides are malicious and can forge any migration information. Even in this case the target TD's attestation is based on information collected by the TDX module. It is independent of any TD and reflects the true identity of the service TDs bound to the target TD.

13.2.7.3. *SERVTD_HASH Calculation Method*

SERVTD_HASH is calculated as follows:

1. Get all service TD binding slots whose SERVTD_BINDING_STATE is not NOT_BOUND.
2. Sort by SERVTD_TYPE as the primary key, SERVTD_INFO_HASH as a secondary key (if multiple service TDs of the same type are bound).
3. Concatenate SERVTD_INFO_HASH, SERVTD_TYPE and SERVTD_ATTR of each slot
4. Concatenate all slots.
5. Calculate SHA384.

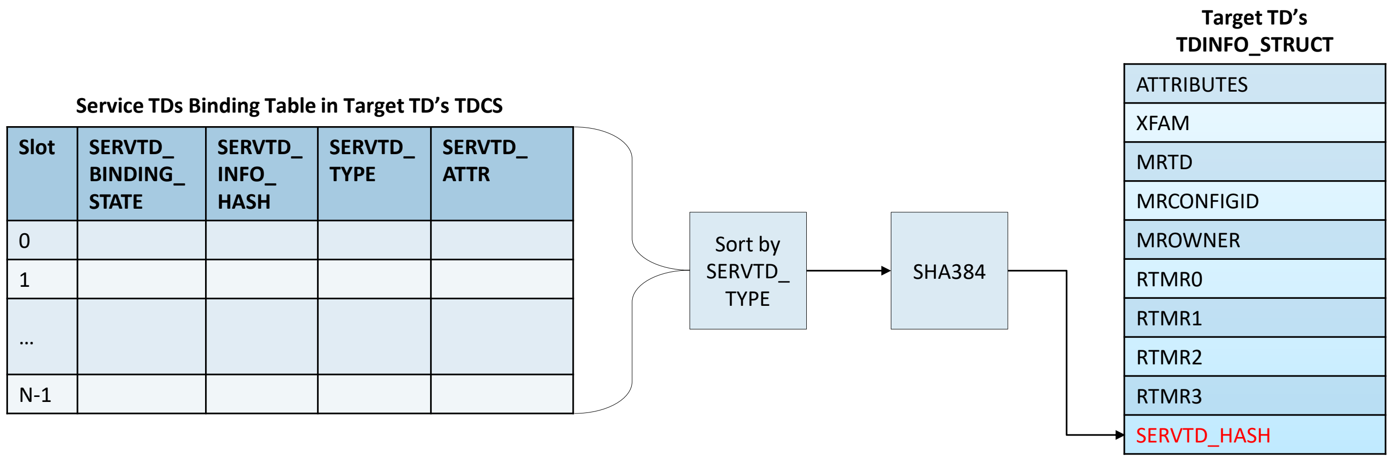


Figure 13.3: SERVTD_HASH Calculation

13.2.8. TDH.SERVTD.PREBIND: Pre-Binding a Service TD

TDH.SERVTD.BIND is used by the host VMM to bind a service TD. It is detailed in the [TDX Module ABI Spec].

Input Operands

- Target TD's TDR HPA
- SERVTD_INFO_HASH
- SERVTD_TYPE
- SERVTD_ATTR
- Service TD Index (slot number in the target TD's binding table)

Operation

- Check that the target TD's measurements have not been finalized (by TDH.MR.FINALIZE).
- Check that no service TD is already bound in the given slot number.
- Store the service TD's SERVTD_INFO_HASH, SERVTD_TYPE and SERVTD_ATTR.

13.2.9. TDH.SERVTD.BIND: Binding a Service TD

TDH.SERVTD.BIND is used by the host VMM to bind a service TD. It is detailed in the [TDX Module ABI Spec].

Binding Scenarios

Initial Binding:	No pre-binding has been done; initial service TD binding can only be done before TDH.MR.FINALIZE of the target TD.
Late Initial Binding:	Pre-binding has been done; initial service TD binding can be done at any time. SERVTD_INFO_HASH and SERVTD_ATTR must match.
Rebinding:	Binding has been done; rebinding conditions depend on SERVTD_ATTR as described before.

Input Operands

- Target TD's TDR HPA
- Service TD's TDR HPA – NULL_PA (-1) if pre-binding is requested
- SERVTD_TYPE

- SERVTD_ATTR
- Service TD Index (slot number in the target TD's binding table)

Output Operands

- Binding Handle (described below)

Operation

- Calculate the service TD's SERVTD_INFO_HASH.
- Check binding conditions vs. the target TD's binding table.
- Store the service TD's SERVTD_INFO_HASH, SERVTD_TYPE, SERVTD_ATTR and SERVTD_UUID in the target TD's binding table.
- Calculate the binding handle as f(service TD's TD_UUID, target TD's TDR HPA, slot number).

13.2.10. Binding Handle

The binding handle is used as a shortcut, to quickly identify both the target TD and the binding slot. It should be noted that the target TD identity is verified by its TD_UUID; the binding handle does not replace it. The binding handle is not a secret.

The binding handle is calculated from the following variables, using a simple addition:

- Least significant 64 bits of SERVTD_UUID – this serves to obfuscate the handle, so the service TD does not use HPA or slot number directly.
- Target TD's TDR HPA (platform-specific unique identifier of the target TD)
- Target TD's binding slot number

Given the handle, the TDX module can reconstruct TDR_HPA and the binding slot number.

The binding handle is platform-specific and must be recreated after migration. This may be triggered when the service TD attempts to access target TD metadata using TDG.SERVTD.RD/WR* and an error is returned.

13.3. Target TD Metadata Access by a Service TD

13.3.1. TDG.SERVTD.RD/WR: Metadata Read/Write Interface Functions

TDG.SERVTD.RD and TDG.SERVTD.WR are similar to other metadata access functions, e.g.:

- Host-side: TDH.MNG.RD/WR
- Guest-side: TDG.VM.RD/WR

Refer to 18.6 for a description of the TDX module metadata interface.

Input Operands

- Target TD_UUID, uniquely identifying the target TD
- Binding handle, identifies the binding slot and a shortcut for identifying the target TD
- A single metadata field ID or metadata field list

Output Operand

- For a single field access: Field value

Operation

1. Calculate the target TD's TDR HPA and binding slot number from the binding handle.
2. Check that the target TD_UUID is the same as specified.
   2.1. A special case (used by Migration TDs) is when the binding had been done on destination platform before the TD was imported. In this case the target TD_UUID is overwritten at the beginning of import, as part of the TD's immutable state import by TDH.IMPORT.STATE.IMMUTABLE. The pre-import TD_UUID is saved in the target TD's TDCS. If the specified target TD_UUID doesn't match the actual value, but matches the pre-import value, a status code is returned to the service TD, with the updated TD_UUID.
3. Get the binding parameters from the target TD's service TD table binding slot.
4. Check that the service TD's TD_UUID is equal to the target TD's bind slot's SERVTD_UUID.
5. Access the metadata (similar to other metadata access operations).

13.3.2. Metadata Access Error Handling

TDG.SERVTD.RD/WR interface functions run in the context of the service TD but access the target TD's control structures. This introduces an opportunity for the service TD to create a denial-of-service to the host VMM, which is handled as described below.

Local Errors (in the Service TD Context)

Local errors that only impact the service TD context are handled normally, as in other TDCALL flows. These include, e.g., the following cases:

- Errors such as incorrect service TD state result in an error code returned to the caller service TD.
- EPT violations when accessing the service TD's memory cause a fault-like TD exit ; The VMM may resolve the situation (e.g., TDH.EXPORT.UNBLOCKW if the service TD is being live-migrated) and resume the service TD.

Cross-TD Errors

Cross-TD errors impact the target TD. For example, errors may happen due to the target TD state, e.g., the target TD may be migrated or may be torn down. The service TD may not be aware of the target TD state when invoking the interface function.

Cross-TD errors cause a trap-like TD exit:

1. TDG.SERVTD.RD/WR* flow sets output operands (e.g., completion status returned in RAX) and advances the virtual CPU state to the next service TD guest instruction, but TD-exits immediately before resuming the guest TD.
2. The host VMM may take action to detect denial of service, e.g., the guest calling TDG.SERVTD.RD/WR* in a tight loop.
3. The host VMM may let the service TD resolve the situation by resuming it, using TDH.VP.ENTER. On TD entry, the service TD gets the status code as returned by TDG.SERVTD.RD*/WR*.

13.3.3. Cross-TD Concurrency Handling: Maintaining Host-Side Priority

13.3.3.1. *Problem Description*

Host VMM access to the target TD has a higher priority than service TD access to that target TD. This helps mitigate denial-of-service cases such as when the service TD loops on TDG.SERV.RD/WR*, locking target TD resources and preventing the host VMM from doing host-side operations that require access to such resources.

Applicable target TD resources are, e.g.:

- TDG.SERV.RD/WR locks the target TD's TDR in a shared mode, to help ensure that the target TD is available throughout the guest-side flow. This may interfere with critical host-side operations (e.g., disabling a TD) that require locking that target TD's TDR in an exclusive mode.
- TDG.SERV.RD/WR locks the target TD's TDCS.OP_STATE to help ensure that OP_STATE doesn't change in a way that prevents access during the guest-side flow. This may interfere with critical host-side operations (e.g., pausing a TD during export) that require locking that target TD's OP_STATE in an exclusive mode.

We currently assume that guest-side flows can only acquire locks in shared mode; thus, they only compete with the host-side flows acquiring locks in exclusive mode.

13.3.3.2. *Solution*

A new HOST_PRIORITY flag is added to shared/exclusive locks protecting resources that may be accessed by the host VMM and a guest service TD. For details, see 18.1.4.

14.I/O Support (without TDX Connect)

This chapter specifies the Intel TDX I/O model (for TDX Module versions and platforms that don't support TDX Connect).

14.1. Overview

Intel TDX architecture does not prescribe a specific software convention to perform I/O from the guest TD. Guest TD providers have many choices to provide I/O to the guest. The common I/O models are emulated devices, para-virtualized devices, SRIOV devices and Direct Device assignments. Guest TD providers can choose to offer combinations of I/O models based on the workload and use case. To virtualize MMIO, the following options can be utilized:

- **Para-Virtualized Drivers** can replace MMIO accesses with TDG.VP.VMCALL to invoke VMM provided MMIO emulation functions.
- **MMIO Emulation by #VE Handlers** can use non-para-virtualized drivers in the guest TD, with the emulation performed by the #VE handler. EPT and #VE mechanisms can be used to reflect violations to the #VE handler in the guest TD on access to virtual MMIO ranges. These violations can invoke VMM-provided MMIO emulation functions through TDG.VP.VMCALL. In this model, the #VE handler is expected to emulate the faulting instruction in the guest TD.

14.2. Paravirtualized I/O

Para-virtualization (e.g., using virtio APIs in KVM, etc.) helps provide a mechanism for the guest TD to use devices on the host machine that are owned and managed by the VMM. The guest TD drivers can use the TDG.VP.VMCALL function to invoke the functions provided by the VMM to perform I/O. The TD drivers must ensure that the data buffers passed to/from functions invoked using TDG.VP.VMCALL are placed in the TD's shared memory space.

14.3. MMIO Emulation and Emulated Devices

An alternate technique that the guest TD may employ to invoke VMM functions for I/O is to emulate MMIO access from legacy device drivers. To support this use model, the VMM may enable reflection of EPT violation to emulated MMIO guest physical addresses as virtualization exceptions (#VE), as described in 11.14. A #VE exception handler in the guest TD OS can emulate the instruction causing the #VE, and as part of the emulation, it can invoke the I/O functions provided by the VMM using TDCALL(TDG.VP.VMCALL). Similar to the paravirtualized I/O model, the TD software must ensure that the data buffers passed to/from functions invoked using TDG.VP.VMCALL are placed in the TD's shared memory space.

14.4. Direct Device Assignment (DDA) and SRIOV

The VMM may assign devices directly to the guest TD. The addresses mapping the MMIO resources of such devices must be mapped in the shared memory space of the TD. When submitting data buffers to these devices, the guest TD must locate the data buffers in shared memory such that the directly assigned device can move data in/out of such buffers using DMA. The data buffers placed in shared memory should be programmed in IOMMU page tables.

The SRIOV virtual function devices assigned to guest TD also follow the DDA guidelines stated above with respect to MMIO and data buffers. The control plane of the virtual function would use the soft or hard mechanism to configure the virtual functions:

- The soft mechanism would use para-virtualization to configure the virtual function.
- The hard mechanism would use hardware mailboxes accessed using MMIO in the shared memory region.

14.5. IOMMU – DMA Remapping

The IOMMU uses the VT-d remapping tables to translate GPA in the DMA from device to HPA. The VT-d remapping tables will reflect the mapping of memory used by I/O devices in the guest TD. The programming of the VT-d remapping tables and management will be done by the VMM.

Only shared GPA memory should be mapped in the VT-d tables:

- If the result of the translation results in a physical address with a TD private key ID, then the IOMMU will abort the transaction and report a VT-d DMA remapping failure.

- If the GPA in the transaction that is input to the IOMMU is private (SHARED bit is 0), then the IOMMU may abort the transaction and report a VT-d DMA remapping failure, even if the translated physical address is with a non-private HKID. This is intended to support debug wherein a TD or VMM could program a bad GPA into the device.

14.6. Shared Virtual Memory (SVM)

Shared Virtual Memory enables applications to access buffers directly accessed by the devices. The VT-d tables help provide the mechanism to map application buffers using the first level and second-level page tables to provide applications access to the same memory accessed by devices.

SVM should be avoided because VT-d tables can only map shared memory.

15. Debug and Profiling Architecture

The Intel TDX module debug architecture includes the following debug facilities:

On-TD Debug: Facilities for debugging a guest TD using software that runs inside the TD

Off-TD Debug: Facilities for debugging a guest TD, configured in debug mode, using software that runs outside the TD

15.1. On-TD Debug

Intel SDM, Vol. 3, 17	Debug, Branch Profile, TSC and Intel Resource Director Technology (Intel RDT) Features

15.1.1. Overview

On-TD debug means that the TD software is using CPU debug capabilities. A debug agent inside the guest TD can use available CPU debug features and – if needed – interact with external debug entities (e.g., a debugger running in a VM on the same platform, or a debugger running on another platform) via standard I/O interfaces. The Intel TDX module is designed to virtualize and isolate TD debug capabilities from the host VMM and software. On-TD debug can be used for production or debug TDs – i.e., regardless of the guest TD's ATTRIBUTES.DEBUG state.

Guest TDs are allowed to use almost all architectural debug features supported by the processor, e.g.:

- Single stepping
- Code, data and I/O breakpoints
- INT3
- Bus lock detection
- DR access detection
- TSX debug

However, the TDX architecture does not allow guest TDs to toggle IA32_DEBUGCTL uncore PMI enabling bit (13).

Guest TDs are allowed to use almost all architectural tracing features, e.g.:

- LBR (if allowed by the TD's XFAM, see 11.8)
- PT (if allowed by the TD's XFAM, see 11.8)
- BTS
- PEBS
- PERF_METRICS

However, the TDX architecture does not allow guest TDs to use BTM.

15.1.2. Generic Debug Handling

15.1.2.1. Context Switch

By design, the Intel TDX module context-switches all debug/tracing state that the guest TD is allowed to use. The host VMM's state of those resources is either restored or initialized following a TD entry/exit. For details, see the [ABI Spec] definition of TDH.VP.ENTER.

15.1.2.2. IA32_DEBUGCTL (MSR 0x1D9) Virtualization

Intel SDM, Vol. 3, 17.4.1	IA32_DEBUGCTL MSR

By design, IA32_DEBUGCTL (MSR 0x1D9) access by the guest TD is restricted as follows:

- Guest TD attempts to set any of the architecturally reserved bits 63:16 and 5:3 result in a #GP(0).
- Guest TD attempts to set bit 14 (FREEZE_WHILE_SMM) to 1 when the virtual value of IA32_PERF_CAPABILITIES[12] (FREEZE_WHILE_SMM_SUPPORTED) is 0 results in a #GP(0). See 15.2 below for Performance Monitoring details.
- Guest TD attempts to set bit 15 (RTM_DEBUG) to 1 when the virtual value of CPUID(7,0).EBX[11] (RTM) is 0 results in a #GP(0).
- Guest TD attempts to set TDX-disallowed values result in a #VE(UNSUPPORTED_FEATURE). This includes the following cases:
  - Enable BTM by setting bits 7:6 to 0x1 (see details in 15.1.3 below).

- Uncore PMI is virtualized as disabled; bit 13 is read as 0 and ignored on write (see 15.5 below).

15.1.3. Debug Feature-Specific Handling

The following table discusses how specific debug features are handled.

Table 15.1: Debug Feature-Specific Handling

Debug Feature	How the Feature is Controlled	Handling
Hardware Breakpoints	• DR7, DR0-3 and DR6	No special handling: DRs are context-switched.
General Detect	• DR7 bit 13 (GD)	No special handling: DR7 is context-switched.
TSX Debug	• DR7 bit 11 (RTM) • IA32_DEBUGCTL bit 15 (RTM)	No special handling: DR7 and IA32_DEBUGCTL are context-switched.
Single Stepping	• RFLAGS bits 18 (Trap Flag) and 16 (Resume Flag) • IA32_DEBUGCTL bit 1 (BTF)	No special handling: RFLAGS and IA32_DEBUGCTL are context-switched.
Bus-Lock Detection	• IA32_DEBUGCTL bit 2 (BUS_LOCK_DETECT)	No special handling: IA32_DEBUGCTL is context-switched.
Software Breakpoints (INT1, INT3)	None	No special handling: software breakpoints are stateless.
Branch Trace Message (BTM)	• IA32_DEBUGCTL bits 6 (TR) and 7 (BTS)	Not allowed: when a guest TD attempts to set IA32_DEBUGCTL[7:6] to 0x1, the Intel TDX module injects a #VE(UNSUPPORTED_FEATURE) (see 15.1.2 above). In debug mode (ATTRIBUTES.DEBUG == 1), the host VMM is allowed to activate BTM by setting the above bits to 0x1.
Branch Trace Store (BTS)	• IA32_DEBUGCTL bits 6 (TR), 7 (BTS), 8 (BTINT), 9 (BTS_OFF_OS) and 10 (BTS_OFF_USR)	No special handling: IA32_DEBUGCTL and IA32_DS_AREA are context-switched. **Notes:** • The guest TD can configure BTS to raise PMI on buffer overflow (by setting BTINT = 1). However, since PMIs are virtualized by the host VMM, the guest TD should be ready to handle spurious, delayed and dropped PMIs. See Perfmon discussion in 15.2 below. • BTS may allow the guest TD to hang the machine if BTS record generation causes a #PF or a #GP(0), because the act of getting to the exception handler may deliver another BTS. **It is highly recommended that the host VMM enables instruction timeout TD exit**.

Debug Feature	How the Feature is Controlled	Handling
Processor Trace (PT)	• IA32_RTIT_* (MSR 0x560 – 0x587) • Requires VMM's consent on TD initialization by setting TD_PARAMS.XFAM[8] to 1 • Setting IA32_RTIT_OUTPUT_BASE with a shared GPA is discouraged[12].	PT state handling as part of the extended feature set state is discussed in 11.8. On hybrid SOCs, CPUID(0x14,0).ECX[31] (IP values contain linear IP) is always virtualized as 0.
Architectural Last Branch Records (LBRs)	• IA32_LBR_CONTROL • Requires VMM's consent on TD initialization by setting TD_PARAMS.XFAM[15] to 1	LBR state handling as part of the extended feature set state is discussed in 11.8. On hybrid SOCs, CPUID(0x1C).EAX[31] (IP values contain linear IP) is always virtualized as 0.
Non-Architectural LBRs	• IA32_DEBUGCTL bit 0 (LBR)	Guest TD attempt to set IA32_DEBUGCTL[0] is ignored by the CPU.

15.2. On-TD Performance Monitoring

Intel SDM, Vol. 3, 18 Performance Monitoring

15.2.1. Overview

If the TDX module supports on-TD performance monitoring, then the host VMM controls whether a guest TD can use the performance monitoring ISA using the TD's ATTRIBUTES.PERFMON bit – part of the TD_PARAMS input to TDH.MNG.INIT (see the [TDX Module ABI Spec]).

By design, if a guest TD is allowed to use performance monitoring, then:

- The guest TD enumerates native architectural Perfmon capabilities via CPUID leaf 0x0A.
- The guest TD is allowed to use all Perfmon ISA. This includes executing the RDPMC instruction and accessing Perfmon MSRs (see 15.2.3 below).
- Perfmon state is context-switched by the Intel TDX module across TD entry and exit transitions.

Context-switching the Perfmon state has a performance impact. TD entry and exit latencies are longer than when a guest TD is not allowed to use Perfmon.

By design, if a guest TD is not allowed to use performance monitoring, then:

- The guest TD enumerates no architectural Perfmon capabilities. CPUID leaf 0x0A returns all 0s.
- The guest TD is not allowed to use Perfmon ISA, including RDPMC.
- Perfmon state is not context-switched across TD entry and exit transitions.

Regardless of Perfmon enabling, per the design:

- IA32_DS_AREA MSR is context-switched across TD entry and exit transitions.
- Counter freeze control (IA32_DEBUGCTL bit 12) is context-switched across TD entry and exit transitions.
- The uncore PMI enable bit (IA32_DEBUGCTL bit 13) is preserved during SEAM mode execution, including Intel TDX module and guest TD execution. This bit is virtualized to the guest TD as 0, and the TD is prevented from setting it. See 15.5 below for details.

See also 15.1 above.

The Intel TDX module is designed to support the following performance monitoring capabilities:

- Architectural performance monitoring version 5, described in [Intel SDM, Vol. 3, 18.2.5)
- Exactly 8 performance monitoring counters (IA32_PMC0 through IA32_PMC7)

[12] The CPU does not allow setting the IA32_RTIT_OUTPUT_BASE value's SHARED bit if it is outside the physical address bits. E.g., when GPAW is 48 and the physical address width is 46. A TD is not typically aware of the physical address width.

- Up to 7 fixed counters (IA32_FIXED_CTR0 through IA32_FIXED_CTR7)
- Some non-architectural MSRs (see 15.2.3 below)
- If the TDX module supports Architectural PEBS (Virtual CPUID(0x23, 0).EAX[5] is configurable), and the CPU supports this feature, the host VMM can enable guest TDs to use Architectural PEBS. The host VMM can limit the TD to using only a subset of the counters with Architectural PEBB. In this case, the host VMM must reset the Architectural PEBS control MSRs for the disallowed counters before calling TDH.VP.ENTER.
  Else (the TDX module does not support Architectural PEBS, virtual CPUID(0x23, 0).EAX[5] is not configurable), guest TDs are not allowed to use this feature. In this case, if the CPU does support Architectural PEBS, the host VMM must reset all Arch PEBS configuration MSRs (see below) before calling TDH.VP.ENTER on a Perfmon-allowed TD.

15.2.2. Performance Monitoring CPUID Virtualization

CPUID(0xA) is the legacy Perfmon leaf. CPUID(0x23) is the new Perfmon leaf, supported if CPUID(7,1).EAX[8] is 1.

Both leaves are virtualized to the guest TD if ATTRIBUTES.PERFMON is 1. CPUID(0x23) is virtualized as 0 if the virtual value of CPUID(7,1).EAX[8] is 0.

TDX does not allow the host VMM to directly configure the virtualization of CPUID(0xA) and CPUID(0x23).

15.2.3. Performance Monitoring MSRs

15.2.3.1. *Overview*

The following tables describe TDX handling of MSRs used by Perfmon:

Table 15.2: Performance Monitoring MSRs

MSR	Comments	Enumeration	Reference
IA32_PMCx	multiple MSRs	x < CPUID(0x0A).EAX[15:8])	
IA32_A_PMCx	Aliases of IA32_PMCx	Same as IA32_PMCx	[Intel SDM, Vol. 3, 18.2.6]
IA32_PERFEVTSELx	multiple MSRs	Same as IA32_PMCx	
IA32_PMC_GPn_CTR	Aliases of IA32_PMCx	(x < CPUID(0x0A).EAX[15:8]) && ((CPUID(7,1).EAX[8] == 0) || (CPUID(0x23,0).EAX[1] == 0) || (CPUID(0x23,1).EAX[x] == 1))	
IA32_PMC_GPn_CFG_A	Aliases of IA32_PERFEVTSELx	Same as IA32_PMC_GPn_CTR	
IA32_FIXED_CTRx	multiple MSRs	(x < CPUID(0x0A).EDX[4:0]) || (CPUID(0x0A).ECX[x] == 1)) The Intel TDX module supports counters 0 through 6, if supported by the CPU.	[Intel SDM, Vol. 3, 18.2.5.2]
IA32_PMC_FXn_CTR	Aliases of IA32_FIXED_CTRx	((x < CPUID(0x0A).EDX[4:0]) || (CPUID(0x0A).ECX[x] == 1)) && ((CPUID(7,1).EAX[8] == 0) || (CPUID(0x23,0).EAX[1] == 0) || (CPUID(0x23,1).EBX[x] == 1)) The Intel TDX module supports counters 0 through 6, if supported by the CPU.	
IA32_FIXED_CTR_CTRL			
IA32_PERF_METRICS		IA32_PERF_CAPABILITIES[15]	

MSR	Comments	Enumeration	Reference
IA32_PERF_CAPABILITIES			
MSR_OFFCORE_RSPx	2 MSRs, non-architectural		
IA32_PERF_GLOBAL_STATUS			
IA32_PERF_GLOBAL_CTRL			
IA32_PERF_GLOBAL_STATUS_RESET IA32_PERF_GLOBAL_STATUS_SET	Command MSRs – not context-switched		
IA32_PERF_GLOBAL_INUSE			
IA32_PMC_GPn_CFG_C	Architectural PEBS event control	CPUID(7,1).EAX[8] && CPUID(0x23,0).EAX[5] && CPUID(0x23,5).EAX[x]	
IA32_PMC_FXn_CFG_C	Architectural PEBS fixed counter control	CPUID(7,1).EAX[8] && CPUID(0x23,0).EAX[5] && CPUID(0x23,5).ECX[x]	
IA32_PEBS_BASE[13] IA32_PEBS_INDEX		CPUID(7,1).EAX[8] && CPUID(0x23,0).EAX[5]	

Table 15.3: Legacy PEBS MSRs

MSR	Comments	Enumeration	Reference
IA32_PEBS_ENABLE	non-architectural	IA32_MISC_ENABLE[12]	
MSR_PEBS_DATA_CFG	non-architectural	IA32_MISC_ENABLE[12]	
MSR_PEBS_LD_LAT	non-architectural	IA32_MISC_ENABLE[12]	
MSR_PEBS_FRONTEND	non-architectural	IA32_MISC_ENABLE[12] Not supported on E-cores	

15.2.3.2.

MSR virtualization is described in 11.10.

New Perfmon MSR Range

A new range of Perfmon MSRs is supported by newer CPUs, which support Perfmon version 6, as enumerated by CPUID(0xA).EDX[7:0]. The new MSR index range, starting at 0x1900, includes (among other) MSRs which are aliases of legacy MSRs.

Table 15.4: New Perfmon MSR Aliases

Legacy MSR	New MSR
IA32_A_PMCn	IA32_PMC_GPn_CTR
IA32_PERFEVTSELn	IA32_PMC_GPn_CFG_A
IA32_FIXED_CTRn	IA32_PMC_FXn_CTR

[13] Setting IA32_PEBS_BASE with a shared GPA is discouraged. The CPU does not allow setting the IA32_PEBS_BASE value's SHARED bit if it is outside the physical address bits. E.g., when GPAW is 48 and the physical address width is 46. A TD is not typically aware of the physical address width.

TDX virtualizes the MSR aliases the same way as it virtualizes the legacy MSRs, since the underlying MSR for an alias index and for its respective legacy index is the same.

Virtualization of Architectural PEBS to TDs

CPU support of Architectural PEBS is enumerated by CPUID(0x23,0).EAX[5]. If the TDX module support Architectural PEBS, the host VMM can configure the virtual value of a Perfmon-allowed TD's CPUID(0x23,0).EAX[5] to be 1, if the CPU supports it, as part of the CPUID configuration parameters of TDH.MNG.INIT.

The Host VMM can allow the TD to use specific PEBS by configuring CPUID(0x23, 5).EAX[31:0] for the programmable counters and CPUID(0x23, 5).ECX[31:0] for the fixed-function counters.

15.2.3.3. Host VMM Behavior

If a TD is allowed to use Performance Monitoring (ATTRIBUTES.PERFMON is 1), but is not allowed to use some PEBS counters, then the host VMM must clear the control MSRs for the disallowed PEBS counters before TD entry. This is enforced by the TDX module on TDH.VP.ENTER.

15.2.4. Performance Monitoring Interrupts (PMIs)

By design, when a guest TD is allowed to use Perfmon, it can also configure the counters to raise PMI on overflow. When such a TD counter overflows, the physical interrupt or an NMI configured by the host VMM into the local APIC is delivered. This interrupt or NMI causes a VM exit, and it is delivered as a TD exit to the host VMM. The host VMM is then expected to inject the PMI into the guest TD, either as a virtual interrupt using the posted interrupt mechanism (see 11.13.4), or as virtual NMI using the NMI injection interface (see 11.13.6).

Since the host VMM is not trusted, the guest TD must be ready to handle spurious, delayed or dropped PMIs. Thus, it is recommended for the guest TD to use PEBS instead of PMIs in order to record TD state at counter overflows.

Uncore PMIs are discussed in 15.5 below.

15.2.5. Perfmon Events Filtering

Perfmon event filtering, if supported by the TDX module, enables the host VMM to specify a set of Perfmon events which the TD is allowed to use.

15.2.5.1. *Enumeration*

Support of Perfmon events filtering is enumerated to the host VMM by TDX_FEATURES0, readable by TDH.SYS.RD*:

- EVENT_FILTERING (bit 24) enumerates support of basic event filtering.
- ENHANCED_EVENT_FILTERING (bit 31) enumerates support of enhanced event filtering.

15.2.5.2. *Background*

Programmable Perfmon counters are configured by the guest TD, using their applicable IA32_PERFEVTSELx MSRs, to counts specific events. An event is identified by the following fields:

Event Select:	8 bits
Unit Mask:	8 bits (UMASK) on CPUs that support Perfmon version 5 or lower, 16 bits (UMASK2/UMASK) on CPUs that support Perfmon version 6.

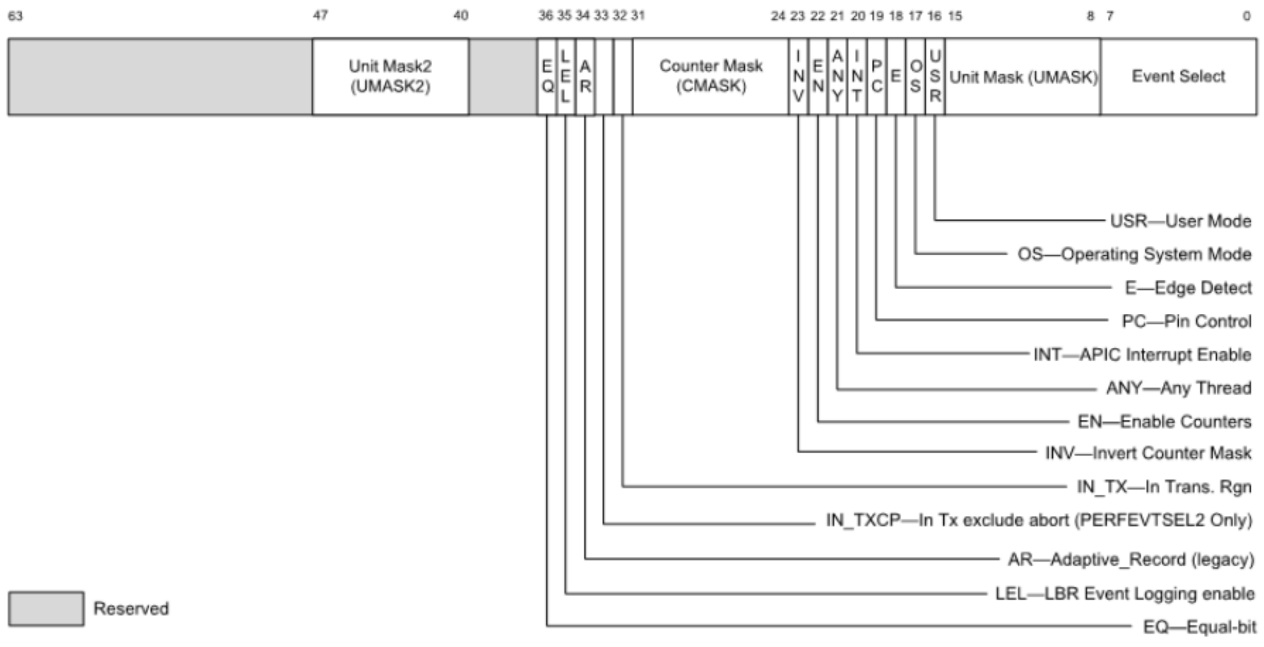


Figure 15.1: Layout of IA32_PERFEVTSELx MSRs

Some of the Perfmon events are architectural; they are enumerated by CPUID(0xA) and CPUID(0x23), as follows:

• CPUID(0xA).EAX[31:24]:	Number of events
• CPUID(0xA).EBX:	Bitmap of **un**supported events
• CPUID(0x23,3).EAX:	Bitmap of supported events

Most events are non-architectural and may vary between CPU models. The list of supported events, per CPU model, is provided by Intel in https://github.com/intel/perfmon.

15.2.5.3. *Event Filtering Configuration and the Filtering Algorithm*

On TD initialization (TDH.MNG.INIT), the host VMM may optionally provide an array of PERFMON_EVENT entries.

Basic Event Filtering

Each PERFMON_EVENT entry specifies Event Select and UMASK values. When the TD writes to an IA32_PERFEVTSELx MSR, the requested Event Select and UMASK values are matched against all PERFMON_EVENT entries. If a match is found, the IA32_PERFEVTSELx MSR is written. Else, the applicable Perfmon counter is disabled.

Enhance Event Filtering

With enhanced events filtering, if supported by the TDX module, each PERFMON_EVENT entry specifies an Event Select and UMASK values to match, a bit mask to mask UMASK before matching, and a flag indicating that the match is negative.

When the TD writes to an IA32_PERFEVTSELx MSR, the TDX module first looks for a **positive match**, where:

- The requested Event Select is equal to the PERFMON_EVENT entry's Event Select.
- The requested UMASK, bit-masked by the PERFMON_EVENT entry's UMASK mask, is equal to the PERFMON_EVENT entry's UMASK.
- The PERFMON_EVENT entry's negative flag is 0.

If a positive match is found, the TDX module first looks for a **negative match**, where:

- The requested Event Select is equal to the PERFMON_EVENT entry's Event Select.
- The requested UMASK, bit-masked by the PERFMON_EVENT entry's UMASK mask, is equal to the PERFMON_EVENT entry's UMASK.
- The PERFMON_EVENT entry's negative flag is 1.

If no negative match is found, then the IA32_PERFEVTSELx MSR is written. Else, the applicable Perfmon counter is disabled.

Guest TD Perspective

If the guest TD executes WRMSR(IA32_PERFEVTSELx) with an EVENT_ID (UMASK and Event Select) value that has not been configured as allowed, the operation appears to complete successfully; there is no error indication. However, the TDX module disables the counter, and no events are counted. If the guest TD executes RDMSR(IA32_PERFEVTSELx), it reads back the value that it wrote to that MSR before.

Perfmon event filtering affects the TD as a whole; L2 VMs are subject to the same restrictions as L1. If required, then L1 15.2.5.4. may implement additional logic by configuring the L2's MSR bitmaps to cause an L2→L1 exit on L2 access to IA32_PERFEVTSELx MSRs.

Perfmon event filtering has no impact on the fixed Perfmon counters. They may be used by the guest TD regardless of the allowed events configuration.

Statistics

The TDX module maintains FILTERED_EVENTS_COUNT, an array of 4 64-bit counters per TD and L1 or L2 VM, which count the number of times the guest TD VM executed WRMSR(IA32_PERFEVTSELx) with an EVENT_ID (UMASK and Event Select) 15.2.5.5. value that has not been configured as allowed. The counters are readable by the host VMM using TDH.MNG.RD.

15.3. Off-TD Debug

A guest TD is defined as **debuggable** if its ATTRIBUTES.DEBUG bit is 1. In this mode, the host VMM can use Intel TDX module functions to read and modify TD VCPU state and TD private memory, which are not accessible when the TD is non-debuggable.

A debuggable TD is, by nature, untrusted. Since the TD's ATTRIBUTES are included in the TDREPORT_STRUCT, the TD's debuggability state is visible to any third party to which the TD attests.

A debuggable TD can't be migrated; its ATTRIBUTES.MIGRATABLE bit must be 0.

The applicable Intel TDX module functions are listed in Table 15.5 below. Note that some of the functions can access non-secret guest TD state regardless of the DEBUG attribute. The lists of state information that can be read and/or written in non-DEBUG and in DEBUG modes are detailed in the referenced sections.

Table 15.5: Off-TD Debug Interface

Intel TDX Function	ATTRIBUTES.DEBUG = 0	ATTRIBUTES.DEBUG = 1
TDH.MNG.RD **TDH.MNG.WR**	N/A	Access secret and non-secret TD-scope state in TDR and TDCS.
TDH.MEM.SEPT.RD	Read Secure EPT entry	Read Secure EPT entry
TDH.VP.RD **TDH.VP.WR**	Access non-secret TD VCPU state in TDVPS (including TD VMCS)	Access secret and non-secret TD VCPU state in TDVPS (including TD VMCS).
TDH.MEM.WR **TDH.MEM.RD**	N/A	Access TD-private memory.
TDH.PHYMEM.PAGE.RDMD	Read page metadata (PAMT information)	Read page metadata (PAMT information).

15.3.1. Modifying Debuggable TD's State, Controls and Memory

When the TD is debuggable, the off-TD debugger can:

- Read and modify TDVMCS fields that contain guest state, VM entry load controls, VM exit save controls, and VM execution controls.
- Read and modify TDVPS fields that contain additional TD VCPU's state (e.g., extended register state).
- Read and modify a per-VCPU copy of the TD's extended feature mask (XFAM), such that more extended register state would be saved to TDVPS on TD exit and restore from TDVPS on TD entry.

This may cause the next VM entry into the TD VCPU to fail due to bad guest state. It may also generate VM exits that wouldn't have happened otherwise (e.g., VM exit due to a #PF within the TD). In non-debuggable TD such VM exits are not expected, and thus treated as fatal TD errors that cause a TD exit with a TDX_NON_RECOVERABLE_TD status. In debuggable TDs, however, such VM exits are expected and cause TD exit.

Specifically, the TDX module handling of TD VM exits works as follows:

1. If this TD VM exit might happen on non-debuggable TDs:
    1.1. Do "standard" handling (may result a TD exit).
    1.2. If an exception is pending to be injected into the TD:
        1.2.1. If the TD is debuggable and its exception bitmap is programmed to intercept that exception:
            1.2.1.1. TD exit to the VMM, as if the exception has been raised during TD execution.
    1.3. Resume the TD (may inject an exception).
2. Else (an unexpected VM exit happened):
    2.1. If the TD is debuggable then TD exit.
    2.2. Else handle this as a fatal TD error. Do non-recoverable TD exit.

In any case, the security of other guest TDs running in production mode is not impacted.

15.3.2. Preventing Guest TD Corruption of DRs

The off-TD (host-side) debugger may need to have full control over guest DRs to help prevent their corruption by the guest TD. To do so, the debugger can do the following:

- Use TDH.VP.WR to set the TD VMCS GUEST_DR7 field's Global Detect bit.
- Set the TD VMCS exception bitmap execution control to intercept debug exceptions.

15.4. Platform-Level Profiling

This section discusses the interoperation of guest TD with platform-level profiling features.

15.4.1. Profiling by IA32_FIXED_CTR1 and IA32_FIXED_CTR2

Intel SDM, Vol. 3, Table 20-2: Association of Fixed-Function Performance Counters with Architectural Performance Events

Enumeration: Availability of the following feature is enumerated by TDX_FEATURES0.FIXED_CTR12_PROF (bit 26), readable by the host VMM using TDH.SYS.RD.

Two of the fixed-function performance monitoring counters continue counting while running in the TDX module. They also continue counting while running is a guest TD, unless the TD is enabled for debugging (ATTRIBUTES.DEBUG is 1) or for performance monitoring (ATTRIBUTES.PERFMON is 1).

The applicable counters are:

IA32_FIXED_CTR1	The number of clock cycles while the logical processor is not in a halt state
IA32_FIXED_CTR2	The number of TSC cycles while the logical processor is not in a halt state and not in a TM stop-clock state

15.4.2. Hardware-Guided Scheduling (HGS+) Profiling

Enumeration: TDX module support of HGS+ profiling is enumerated by ATTRIBUTES_FIXED0.HGS_PLUS_PROF (bit 4) value of 1, readable by TDH.SYS.RD*. This value indicates that the TD's ATTRIBUTES.HGS_PLUS_PROF bit may be set to 1.

If HGS+ profiling is supported, the host VMM can configure guest TDs to either disable HGS+ profiling or enable it during the TD runtime. Enabling HGS+ may impact TD security by, e.g., creating side channels, and thus is reflected in the TD's attestation as part of its ATTRIBUTES. HGS+ profiling is configured by the TD's ATTRIBUTES.HGS_PLUS_PROF bit, provided as an input to TDH.MNG.INIT. HGS+ profiling is always disabled during TDX module runtime, except during TDX module initialization (TDH.SYS.INIT and TDH.SYS.LP.INIT) and during TD entry (TDH.VP.ENTER) into and exit from a TD for which ATTRIBUTES.HGS_PLUS_PROF is set to 1.

15.5. Uncore Performance Monitoring Interrupts (Uncore PMIs)

By design, neither the Intel TDX module itself nor its guest TDs are allowed to use Uncore PMIs. The state of IA32_DEBUGCTL MSR bit 13 (ENABLE_UNCORE_PMI) is preserved across SEAMCALL, SEAM root and non-root mode and SEAMRET, except for very short time periods immediately after SEAMCALL and VM exit.

15.6. Interaction with Core Out-Of-Band (OOB) Telemetry

Enumeration: TDX module support of OOB profiling during TD runtime is enumerated by ATTRIBUTES_FIXED0.PMT_PROF (bit 6) value of 1, readable by TDH.SYS.RD*. This value indicates that the TD's ATTRIBUTES.PMT_PROF bit may be set to 1.

If OOB profiling is supported, the host VMM can configure guest TDs to either disable (default) or enable OOB profiling during TD runtime. This is configured by the TD's ATTRIBUTES.PMT_PROF bit, provided as an input to TDH.MNG.INIT. Enabling PMT_PROF may have security implications due to possible leakage of side channel information.

OOB telemetry is always disabled during runtime of the TDX Module itself.

16. Memory Integrity Protection and Machine Check Handling

16.1. Overview

The Intel TDX module's memory integrity protection and machine check handling are designed to address the following security objectives:

- Corruption of TD private data or Intel TDX module memory must be detectable before the decrypted corrupted data are consumed by the guest TD or by the Intel TDX module.
- To help improve resistance to brute force attacks, software must not be able to **repeatedly** cause memory integrity violations during Intel TDX module or guest TD operation. When an integrity violation is detected, the affected guest TD and the key corresponding to its affected HKID must become unusable for normal operation of the TD – i.e., the TD may only be torn down.
- Any software except guest TD or TDX module must not be able to speculatively or non-speculatively access TD private memory, to detect if a prior corruption attempt was successful in finding an integrity collision or failed and received zero-data.

As a **best effort**, the TDX module is designed to enable limiting the impact of memory integrity violations in a guest TD context to that guest TD, i.e., requiring only that guest TD to be torn down. However, there are cases where memory integrity violations result in an unbreakable shutdown of the LP.

16.2. TDX Memory Integrity Protection Background

16.2.1. Platforms not Using ACT for Memory Protection

16.2.1.1. *Non-ACT Platforms Memory Integrity Protection*

16.2.1.1.1. Non-ACT Platforms: Cryptographic Integrity (Ci) vs. Logical Integrity (Li), MAC and TD Owner

TDX architecture aims to provide resiliency against confidentiality and integrity attacks by software. Towards this goal, the TDX architecture helps enforce the enabling of memory integrity for all private HKIDs. It supports two memory integrity modes that can be configured on the platform:

Cryptographic Integrity (Ci)	Memory content is encrypted and protected by a MAC and a TD Owner bit.
Logical Integrity (Li)	Memory content is encrypted and protected by a TD Owner bit.

In both Ci and Li modes, the memory controllers store a 1-bit **TD Owner** metadata for each cache line. The TD Owner bit is set to 1 for writes with a private HKID and is cleared to 0 for writes with a shared HKID. The TD Owner bit is covered by ECC.

When Ci mode is enabled, the CPU's memory controllers compute a 28-bit integrity check value (**MAC**) for the data (cache line) during writes and store the MAC with the memory as meta-data. The MAC is calculated over the components described in the table below. The MAC is covered by ECC.

Table 16.1: Components for MAC Calculation (Ci Mode)

Component	Description
Ciphertext Data	512 bits of data being written to memory.
Encryption Tweak	128-bit encryption tweak, generated by encrypting the physical address with the 128-bit per-HKID ephemeral AES-XTS tweak key. The tweak key is generated on key configuration (TDH.SYS.KEY.CONFIG and TDH.MNG.KEY.CONFIG).
TD Owner Bit	Indicates that the data was written using a private HKID.
MAC Key	128-bit MAC key, generated by hardware on platform initialization, when BIOS configures the IA32_TME_ACTIVATE MSR.

16.2.1.1.2. Non-ACT Platforms: MAC and TD Owner Update on Memory Writes

The MAC and the TD Owner bit are updated on memory writes by the memory controller per the following criteria:

- If memory write is for a private HKID, the TD Owner bit is set, and integrity information (MAC) is computed and stored as meta-data along with ciphertext in memory.
- Else (write is for a shared HKID), the TD Owner bit is clear, and based on the key configuration, integrity information (MAC) may be stored along with ciphertext in memory.

The state diagram below shows the TD Owner bit state changes due to memory state changes.

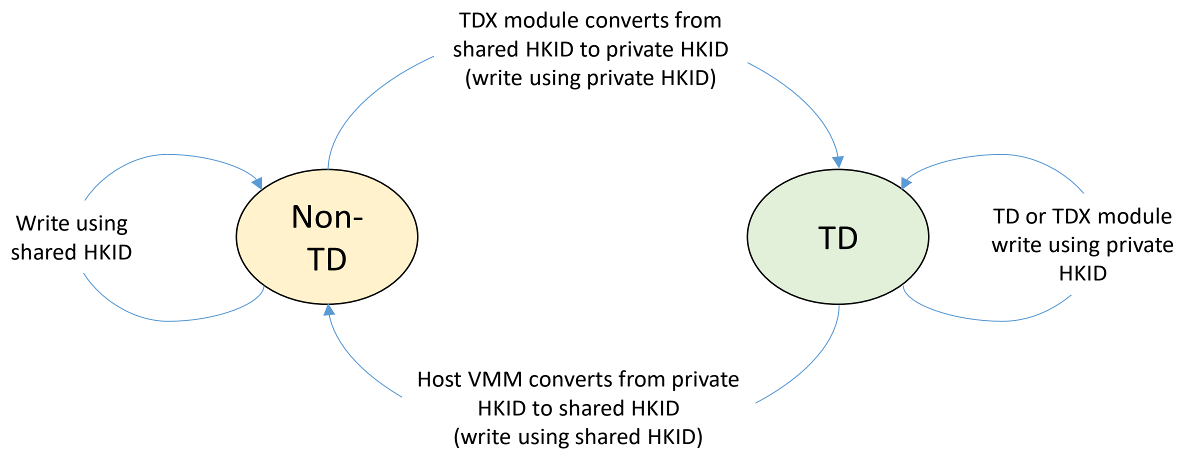

TDX module converts from shared HKID to private HKID (write using private HKID)
Write using shared HKID
Non-TD
TD
TD or TDX module write using private HKID
Host VMM converts from private HKID to shared HKID (write using shared HKID)

Figure 16.1: Non-ACT Platforms: TD Owner Bit Setting on Write

16.2.1.1.3. Non-ACT Platforms Memory Reads: Integrity and TD Owner Bit Checks, Poison Generation and Poison Consumption

On platforms not using ACT for memory protection, checks on memory reads depend on whether Cryptographic Integrity (Ci) is enabled on the platform, or Logical Integrity (Li) is used. This is shown in the tables below.

- When the memory read transaction uses a private HKID, TD Owner bit mismatch and/or integrity check failure (for Ci) result in a new poison generation. An all-0 data is returned, with a poison indication.
- The poison indication is sticky; it is stored back to memory. Subsequent read transactions that read a previously poisoned memory line return a poison indication regardless of the TD Owner bit or integrity checks. A sticky poison indication is cleared when the whole memory line is written; the correct way to do so is by using the MOVDIR64B instruction.
- Any reads of TD private data (TD Owner is 1) done outside SEAM mode (i.e., with a shared HKID) return all-0. This is intended to prevent the host VMM from testing malicious ciphertext for a MAC collision, since the VMM will deterministically see zeroed data in the cache for speculative accesses. No new poison indication is returned; however, a previous poison indication that has been stored in memory may be returned.

Table 16.2: Non-ACT Platforms Checks on Memory Reads in Ci Mode

HKID Type	Integrity Enabled for HKID	TD Owner Bit	Integrity Check	Returned Data	New Poison	Comments
Private	Yes	0	N/A	0	Poison	TD bit mismatch failure may be triggered if the memory was previously written using a shared HKID.
		1	Pass	Decrypted data	None	If the memory line has been previously poisoned, the read transaction may return a poison.
		1	Fail	0	Poison	Integrity check failure may be triggered if the memory was previously written using a different encryption key.
Shared	Yes	0	Pass	Decrypted data	None	If the memory line has been previously poisoned, the read transaction may return a poison.
		0	Fail	0	Poison	Integrity check failure may be triggered if the memory was previously written using a different encryption key.
		1	N/A	0	Poison	TD bit mismatch failure may be triggered if the memory was previously written using a private HKID.
Shared	No	0	N/A	Decrypted data	None	If the memory line has been previously poisoned, the read transaction may return a poison.
		1	N/A	0	None	If the memory line has been previously poisoned, the read transaction may return a poison.

Table 16.3: Non-ACT Platforms Checks on Memory Reads in Li Mode

HKID Type	Integrity Enabled for HKID	TD Owner Bit	Integrity Check	Returned Data	New Poison	Comments
Private	No	0	N/A	0	Poison	TD bit mismatch failure may be triggered if the memory was previously written using a shared HKID.
		1	N/A	Decrypted data	None	If the memory line has been previously poisoned, the read transaction may return a poison.
Shared	No	0	N/A	Decrypted data	None	If the memory line has been previously poisoned, the read transaction may return a poison.
		1	N/A	0	None	If the memory line has been previously poisoned, the read transaction may return a poison.

16.2.1.2. *Non-ACT Platforms Memory Writes: No Integrity nor TD Owner Bit Checks*

On platforms not using ACT for memory protection, the TD Owner bit is not checked on memory writes. It is the responsibility of the host VMM to prevent writing to memory that has been assigned as TD private memory. Failing to do so will result in memory corruption; such corruption will be detected when the guest TD or the TDX module attempts to read that memory, as described above.

The host VMM should always initialize memory that has been used with a private HKID (i.e., TD private memory and TDX control structures), and is about to be used with a shared HKID, using a full line write. The correct way to do so is by using the MOVDIR64B instruction. This helps ensure that the TD Owner bit and any stored poison indication are cleared.

16.2.2. Platforms Using ACT for Memory Integrity Protection

16.2.2.1. *ACT Platforms: Logical Integrity (Li) Provided by an Access Control Table (ACT)*

On platforms using ACT-protected memory, memory integrity protection is provided by a 1-bit **TD Owner** metadata for each 4KB of physical memory. The TD owner bits are stored in Access Control Tables, separate from the memory lines being protected. There is a separate ACT per memory controller; the content of all ACTs is the same, except for a short time during transitions, where the TDX module sets the ACT bits as described below.

TD Owner bits in the ACTs are written by the TDX module on 4KB page conversion between shared and private. They are read by the memory controllers during memory read and write transactions.

16.2.2.2. *ACT Platforms: TD Owner Bit Update on Page Conversion between Shared and Private*

On platforms using ACT-protected memory, page ownership is updated explicitly by the TDX module.

- When the TDX module converts a page from being a shared page to being a private page, it sets the applicable ACT(s)' TD Owner bits to 1.
- When the TDX module converts a page from being a private page to being a shared page, it sets the applicable ACT(s)' TD Owner bits to 0.

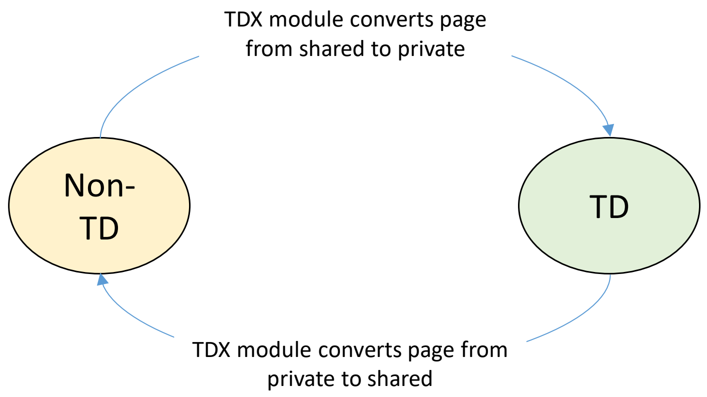

TDX module converts page from shared to private
Non-TD
TD
TDX module converts page from private to shared

Figure 16.2: ACT TD Owner Bit Transition

16.2.2.3. *ACT Platforms Memory Access: TD Owner Bit Checks, Poison Generation and Poison Consumption*

On platforms using ACT-protected memory, checks are done on both memory write and memory reads. This is shown in the tables below.

- Any read of TD private data (TD Owner is 1) done outside SEAM mode (i.e., with a shared HKID) returns all-0. This is intended to prevent the host VMM from testing malicious ciphertext for a MAC collision, since the VMM will deterministically see zeroed data in the cache for speculative accesses.

Table 16.4: ACT Checks on Memory Reads

HKID Type	Integrity Enabled for HKID	TD Owner Bit	Operation	Returned Data	New Poison	Comments
Private	No	0	Read	Decrypted data	None	This case is prevented by the TDX module’s control of Secure EPT
		1	Read	Decrypted data	None	
Shared	No	0	Read	Decrypted data	None	
		1	Read	0	None	

- When a memory write transaction uses a private HKID, no TD Owner bit check is performed. The data is encrypted with the private key and written to memory.
- When a memory write transaction uses a shared HKID, the TD Owner bit is checked to be 0. If not 0, the write is silently dropped.

Table 16.5: ACT Checks on Memory Writes

HKID Type	Integrity Enabled for HKID	TD Owner Bit	Operation	Saved Data	New Poison	Comments
Private	No	0	Write	Encrypted data	None	This case is prevented by the TDX module’s control of Secure EPT
		1	Write	Encrypted data	None	
Shared	No	0	Write	Encrypted data	None	
		1	Write	Dropped	None	

16.2.3. Memory Integrity Error Logging, Machine Checks and Unbreakable Shutdowns

Memory integrity errors that result in poison generation are logged by the memory controller as **UCNA** (uncorrected no-action required) UCR errors which are signaled via CMCI (if CMCI is enabled) or CSMI (if enabled).

On a subsequent consumption (read) of the poisoned data by software, there are two possible scenarios:

Machine Check: In most cases, the core determines that the execution can continue, and it treats poison with fault-like exception semantics signaled as an **MCE** (Machine Check Exception) or **MSMI** (Machine-check System Management Interrupt).

The poison memory address, at a granularity no finer than 32 bytes, is logged in IA32_MCi_ADDRESS MSRs.

Handling of machine check events (MCE or MSMI) when executing in a guest TD (in SEAM non-root mode) and in the Intel TDX module (in SEAM root mode) is described in the following sections.

Unbreakable Shutdown: In some cases, the core determines that execution cannot continue (e.g., long µCode flows), and it goes into an unbreakable shutdown.

An unbreakable shutdown that happens while running in SEAM mode, either in a guest TD or in the TDX module, globally marks TDX as disabled – all subsequent SEAMCALL invocations on any logical processor of the platform lead to a VMfailInvalid error.

16.3. Machine Check Architecture (MCA) Background

Intel SDM, Vol. 3, 15	Machine-Check Architecture

The **machine-check architecture (MCA)** provides a mechanism for detecting and reporting hardware (machine) errors. These include system bus errors, ECC errors, parity errors, cache errors and TLB errors. MCA consists of a set of model-specific registers (MSRs) that are used to set up machine checking, and it includes additional banks of MSRs used for recording errors that are detected.

16.3.1. Uncorrected Machine Check Error

The processor signals the detection of an **uncorrected machine-check error** by generating a **machine-check exception (MCE)**, which is a fault-like exception. An MCA enhancement supports software recovery from certain uncorrected **recoverable** machine check errors. Poisoned cache line consumption by the guest TD is considered such an error. The machine-check exception handler is expected to be implemented in the VMM.

16.3.2. Corrected Machine Check Interrupt (CMCI)

Intel SDM, Vol. 3, 15.5	Corrected Machine Check Error Interrupt

Processors on which TDX will be supported can also report information on corrected machine-check errors and deliver a programmable interrupt for software to respond to MC errors – referred to as **corrected machine-check interrupt (CMCI)**.

CMCI is delivered as a normal interrupt. If delivered during guest TD operation, this interrupt causes a VM exit, and Intel TDX module performs a TD exit to the host VMM. If delivered during Intel TDX module operation, this interrupt remains pending until either SEAMRET to the host VMM or until VM entry to a guest TD.

16.3.3. Machine Check System Management Interrupt (MSMI)

MSMI is part of the Enhanced Machine Check Architecture, Gen. 2 (EMCA2). With EMCA2 enabled, each machine check bank can be configured to assert SMI instead of MCE or CMCI. This is intended to allow the SMM handler to correct the error when possible. For details, see [Error Reporting through EMCA2].

When the processor observes an SMI while a guest TD is running (i.e., SEAM non-root mode) it causes a VM exit to the TDX module with exit reason set to “IO SMI” or “Other SMI” VM exit appropriately. The observed SMI remains pending following the VM exit. The exit qualification bit 0 is set to 1 if the SMI is a machine check initiated SMI (MSMI).

The core ignores MSMI configuration for poison consumption error; they are always reported as MCE.

16.3.4. Local Machine Check Event (LMCE)

Intel SDM, Vol. 3, 15.3.1.5	Enabling Local Machine Check

When system software has enabled LMCE, then hardware will determine if a particular error can be delivered only to a single logical processor, instead of being broadcast to all logical processors. This is the recommended configuration for TDX.

16.4. Recommended MCA Platform Configuration for TDX

The following platform MCA configuration is recommended for TDX:

- LMCE should be enabled, so that machine check events that happen in the scope of a certain logical processor are delivered only to that logical processor.
- EMCA2 should be enabled only if the CPU supports status indication of MCE during non-root SEAM mode execution.

The following sections provide additional details.

16.5. *Handling Machine Check Events during Guest TD Operation*

16.5.1. Machine Check Events Delivered as an #MC Exception

If EMCA2 is not enabled, the machine check event is delivered as an #MC exception. With LMCE enabled, the MCE is delivered only to the logical processor that consumed the poisoned cache line.

The Intel TDX module configures the MCE events when they occur in a TD guest to cause a VM exit to the Intel TDX module. This includes the following cases:

- MCE during guest TD operation
- MCE during a successful VM entry to a guest TD
- MCE during a failed VM exit, where normally execution would remain in the guest TD

The Intel TDX module implements this as follows:

- The Intel TDX module enforces guest TD CR4.MCE to 1.
- The Intel TDX module sets bit 18 (MC) of the TD VMCS Exception Bitmap to 1.

On VM exit, if the exit reason is Exception or NMI (0), the Intel TDX module reads the TD VMCS' VM-exit interruption information to determine if the VM exit was caused by a #MC (18). If so, the Intel TDX module puts the TD in a FATAL state, preventing further TD entries. The TDX module then completes the TD exit flow. The TDH.VP.ENTER outputs indicate the status as TDX_NON_RECOVERABLE_TD_FATAL and provides the exit reason, exit qualification and exit interruption information.

Note: The TDX module does not analyze the MCE to determine its source – whether it's a memory integrity violation or some other event. The TDX module does not read nor write the IA32_MC* MSRs.

Based on the TDH.VM.ENTER outputs (exit reason etc.), the host VMM is expected to understand that a Machine Check event happened, and that the TD should be torn down.

The host VMM can reclaim memory assigned to TDs in a FATAL state using the normal TD teardown flow (TDH.VP.FLUSH, TDH.MNG.VPFLUSHDONE, TDH.PHYMEM.CACHE.WB, TDH.MNG.KEY.FREEID, TDH.PHYMEM.PAGE.RECLAIM).

Note: The host VMM should not attempt to read the poisoned memory locations. Doing so results in a poison consumption and an MCE in the VMM context.

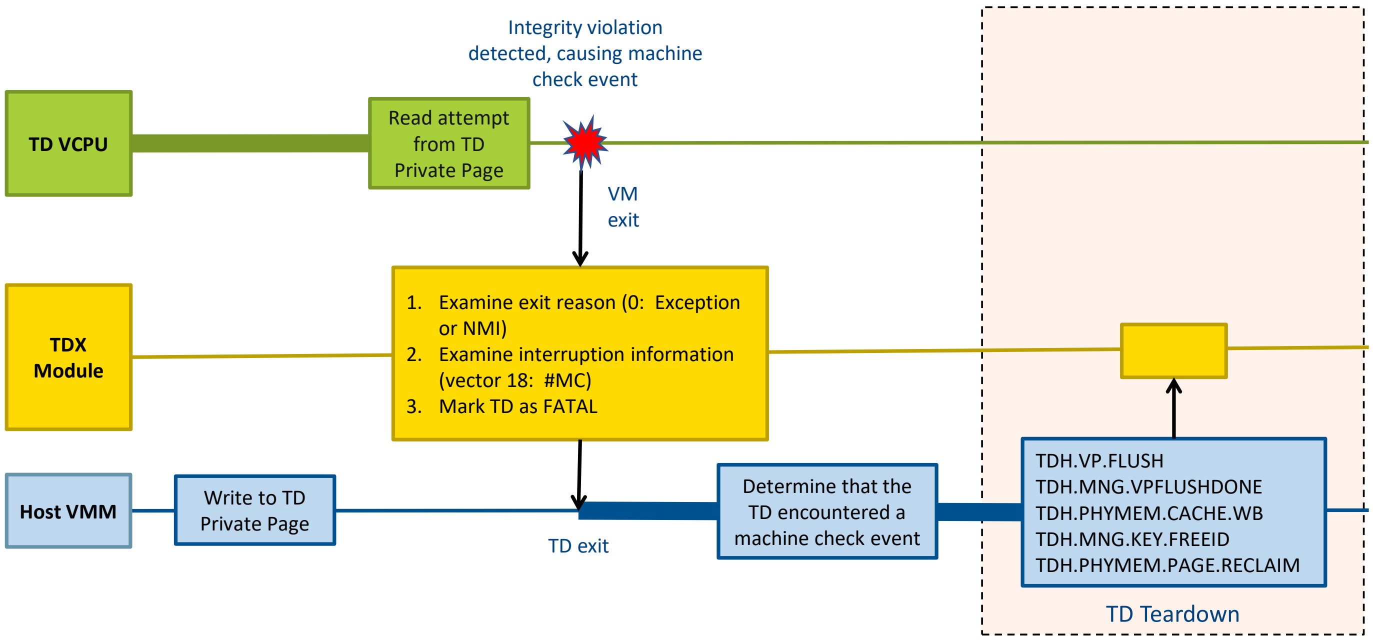

Integrity violation detected, causing machine check event
TD VCPU
Read attempt from TD Private Page
VM exit
TDX Module
1. Examine exit reason (0: Exception or NMI)
2. Examine interruption information (vector 18: #MC)
3. Mark TD as FATAL
Host VMM
Write to TD Private Page
TD exit
Determine that the TD encountered a machine check event
TDH.VP.FLUSH
TDH.MNG.VPFLUSHDONE
TDH.PHYMEM.CACHE.WB
TDH.MNG.KEY.FREEID
TDH.PHYMEM.PAGE.RECLAIM
TD Teardown

Figure 16.3: Example of Handling an MCE in a TD Context

16.5.2. EMCA2: Machine Check Events Delivered as an MSMI

16.5.2.1. *Determining CPU Support*

Host VMM enabling of EMCA2 is only recommended with CPUs that support MCE during non-root SEAM mode, by IA32_MCG_STATUS (MSR 0x17A) SEAM_NR bit (12). Support of this bit is enumerated by IA32_MCG_CAP (MSR 0x179) bit 12.

16.5.2.2. *Pending MSMI Causing a TD Exit*

If EMCA2 is enabled, the machine check event is delivered as an MSMI. With LMCE enabled, the MSMI is delivered only to the logical processor that consumed the poisoned cache line.

Contrary to non-TDX operation, an SMI that occurs in a TD guest does not immediately invoke the SMM handler. Instead, an SMI causes a VM exit to the Intel TDX module and remains pending.

On VM exit, if the exit reason is Other SMI (6), the Intel TDX module reads the TD VMCS' exit qualification bit 0 to determine if the VM exit was caused by a Machine Check that was mutated into an SMI. If so, the Intel TDX module puts the TD in a FATAL state, preventing further TD entries. The TDX module then completes the TD exit flow. The TDH.VP.ENTER outputs indicate the status as TDX_NON_RECOVERABLE_TD_FATAL and provides the exit reason and exit qualification.

Note: The TDX module does not analyze the MCE to determine its source – whether it's a memory integrity violation or some other event.

16.5.2.3. *Operation Following TD Exit*

Once TD exit has completed and the CPU is no longer in SEAM mode, the pending SMI event is taken and the platform's SMM handler is invoked. On RSM, the SMM handler injects an #MC into the host VMM.

If the CPU supports IA32_MCG_STATUS[SEAM_NR], the host VMM's #MC handler can determine that the error happened during guest TD execution if both conditions below are true:

- The error is recoverable, as indicated by IA32_MCi_STATUS[PCC] == 0, and
- The error happened while executing in a TD, as indicated by IA32_MCG_STATUS[SEAM_NR].

The #MC handler should then clear IA32_MCG_STATUS[SEAM_NR]; the CPU doesn't clear it.

If the #MC handler determines that the error happens while executing in the guest TD, then based on the TD exit status the host VMM can then tear down the affected guest TD.

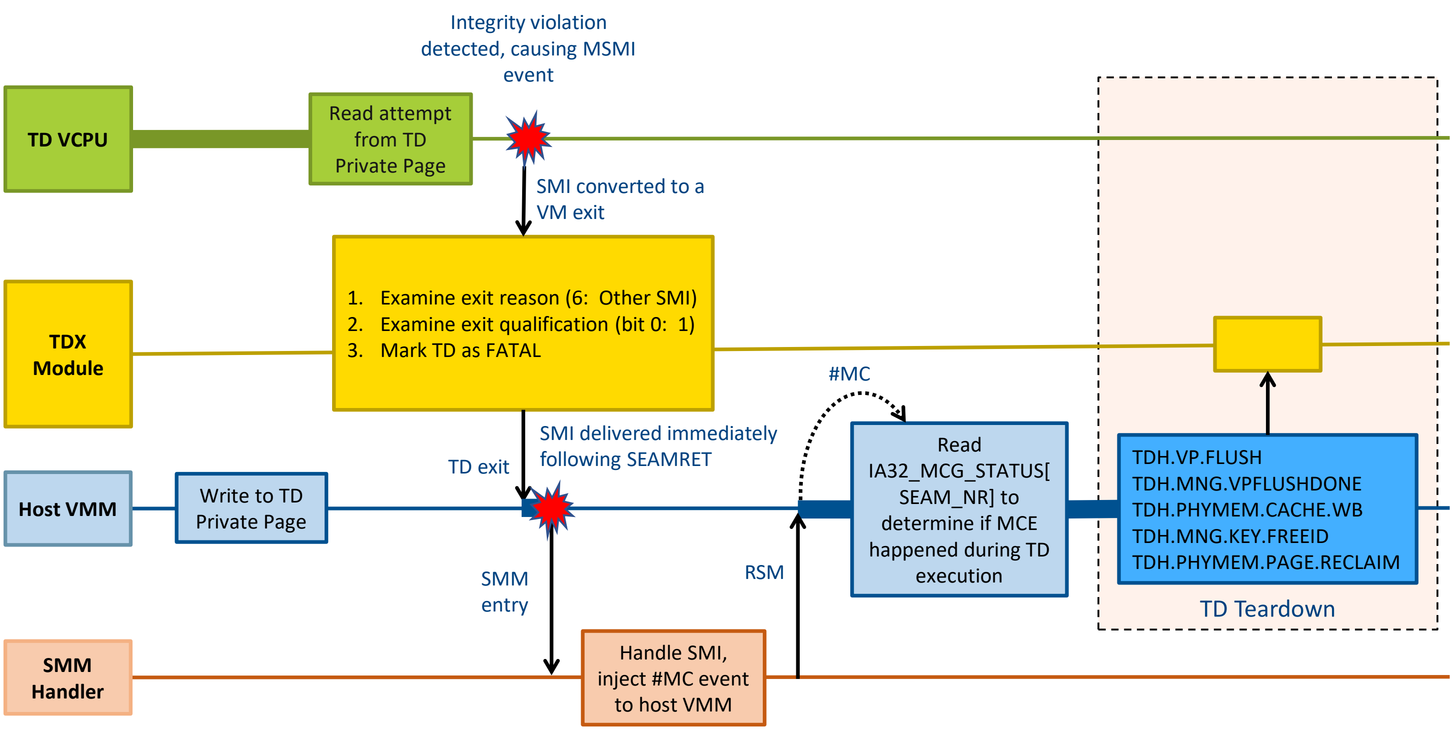

Integrity violation detected, causing MSMI event
TD VCPU
Read attempt from TD Private Page
SMI converted to a VM exit
TDX Module
1. Examine exit reason (6: Other SMI)
2. Examine exit qualification (bit 0: 1)
3. Mark TD as FATAL
#MC
SMI delivered immediately following SEAMRET
TD exit
Host VMM
Write to TD Private Page
Read IA32_MCG_STATUS[SEAM_NR] to determine if MCE happened during TD execution
TDH.VP.FLUSH
TDH.MNG.VPFLUSHDONE
TDH.PHYMEM.CACHE.WB
TDH.MNG.KEY.FREEID
TDH.PHYMEM.PAGE.RECLAIM
TD Teardown
SMM entry
RSM
SMM Handler
Handle SMI, inject #MC event to host VMM

Figure 16.4: Example of Handling an MSMI in a TD Context

16.5.3. LMCE Disabled (Not Recommended)

If LMCE is disabled, then an MCE or MSMI is broadcast to all logical processors on the platform. Any TD that happens to be running will be put in a FATAL state.

Note: The TDX module does not check the MCE details. Any MCE that causes a VM exit from a guest TD is considered fatal to that TD.

16.5.4. Machine Check Events Delivered as a CMCI

CMCI is treated as a normal interrupt, causing an asynchronous TD exit; there's no special handling.

On VM exit, if the exit reason is Exception or NMI (0), the Intel TDX module reads the TD VMCS' VM-exit interruption information to determine if the VM exit was caused by a #MC (18). If not, the Intel TDX module completes the TD exit flow. The TDH.VP.ENTER outputs indicate the status as TDX_SUCCESS and provides the exit reason, exit qualification and exit interruption information.

Based on the TDH.VM.ENTER outputs, the host VMM is expected to process the CMCI interrupt.

16.6. Handling MCE during Intel TDX Module Operation

Any machine check event that occurs during Intel TDX module operation (in SEAM root mode) forces an unbreakable shutdown on a current LP. Shutdown also globally marks TDX as disabled – all subsequent SEAMCALL invocations on any logical processor of the platform lead to a VMfailInvalid error.

17.Side Channel Attack Mitigation Mechanisms

17.1. Checking and Virtualization of CPU Side Channel Protection Mechanisms Enumeration

17.1.1. IA32_ARCH_CAPABILITIES (MSR 0x10A)

On TDX module initialization (TDH.SYS.INIT and TDH.SYS.LP.INIT), the TDX module reads the IA32_ARCH_CAPABILITIES MSR to check the value of multiple bits, indicating whether the CPU is vulnerable to a list of known attacks. The TDX module virtualizes the IA32_ARCH_CAPABILITIES MSR, as seen by guest TDs. Some of the bits are configurable, to allow TD migration between dissimilar platforms. For more information , refer to 11.10.2 and to the [ABI Spec].

Table 17.1: IA32_ARCH_CAPABILITIES MSR Checks and Virtualization

Bit(s)	Name	Native Value Checked on TDX Module Init	Virtual Value as Seen by Guest TDs	Virtual Value Checked on Migration Import
0	RDCL_NO	1	1	Must be 1
1	IBRS_ALL	1	1	Must be 1
2	RSBA	0	0	Must be 0
3	SKIP_L1DFL_VMENTRY	1	1	Must be 1
4	SSB_NO	Same on all LPs	Configurable by the host VMM – can allow to be 1	May be 1 only if native is 1
5	MDS_NO	1	1	Must be 1
6	IF_PSCHANGE_MC_NO	1	1	Must be 1
7	TSX_CTRL	Same on all LPs	Configurable by the host VMM – allowed to be 1 only if CPUID configuration enables TSX	Must be 1 if CPUID configuration enables TSX
8	TAA_NO	1	1	Must be 1
9	RESERVED	Same on all LPs	0	Must be 0
10	MISC_PACKAGE_CTRLS	1	0	Must be 0
11	ENERGY_FILTERING_CTL	1	0	Must be 0
12	DOITM	1	1	Must be 1
13	SBDR_SSDP_NO	1	1	Must be 1
14	FBSDP_NO	1	1	Must be 1
15	PSDP_NO	1	1	1
16	RESERVED	Same on all LPs	0	0
17	FB_CLEAR	Same on all LPs	0	0
18	FB_CLEAR_CTRL	Same on all LPs	0	0
19	RRSBA	Same on all LPs	Configurable by the host VMM – can force to 1	Must be 1 if native is 1
20	BHI_NO	Same on all LPs	Configurable by the host VMM – can allow to be 1	May be 1 only if native is 1
21	XAPIC_DISABLE_STATUS	1	1	1
22	RESERVED	Same on all LPs	0	0
23	OVERCLOCKING_STATUS	Same on all LPs	0	0

Bit(s)	Name	Native Value Checked on TDX Module Init	Virtual Value as Seen by Guest TDs	Virtual Value Checked on Migration Import
24	PBRSB_NO	Same on all LPs	Configurable by the host VMM – can allow to be 1	May be 1 only if native is 1
63:25	RESERVED	Same on all LPs	0	0

17.1.2. CPUID

On TDX module initialization (TDH.SYS.INIT and TDH.SYS.LP.INIT), the TDX module reads some CPUID fields to check the value of multiple bits, indicating whether the CPU is vulnerable to a list of known attacks. The TDX module virtualizes the CPUID values, as seen by guest TDs. Some of the bits are configurable, to allow TD migration between dissimilar platforms. For more information, refer to 11.11 and to the [ABI Spec].

Table 17.2: Checks and Virtualization of Side Channel Related CPUID Fields

Leaf	Sub-Leaf	Reg.	Bit	Name	Verified on TDX Module Init	Virtual Value as Seen by Guest TDs
7	0	EDX	9	MCU_OPT	0	0
			10	MD_CLEAR	1	1
			13	TSX_FORCE_ABORT	0	0
			26	IBRS and IBPB support	1	1
			27	STIBP support	1	1
			28	L1D_FLUSH support	1	1
			29	IA32_ARCH_CAPABILITIES	1	1
			30	IA32_CORE_CAPABILITIES	1	1
			31	SSBD supported	1	1
	2	EDX	0	PSFD supported	1	1
			1	IPRED_CTRL	1	1
			2	RRSBA_CTRL	1	1
			3	DDPD	Same on all LPs	Configurable by the host VMM – can allow to be 1
			4	BHI_NO	1	
			5	MCDT_NO	Same on all LPs	Configurable by the host VMM – can allow to be 1

17.2. Branch Prediction Side Channel Attacks Mitigation Mechanisms

Branch predictions cached by the CPU before entering a guest TD should not impact the behavior of that TD. The Intel TDX module helps ensure that by applying CPU mechanisms to isolate the branch predictions of each guest TD from branch predication done outside its execution.

In a partitioned TD, the L1 VMM is responsible for isolating indirect branch predictors (IBPs) between the L1 VMM and L2 VMMs. The L1 VMM should issue an indirect branch prediction barrier (IBPB) command to the CPU, by writing the IA32_PRED_CMD MSR with the IBPB bit set, immediately before L1→L2 VM entry and immediately after L2→L1 VM exit.

17.3. Single-Step and Zero-Step Attacks Mitigation Mechanisms

17.3.1. Attacks Description

Single-step attacks, zero-step attacks and EPT fault attacks are techniques that provide an adversary with access to a class of powerful, low-noise side channel attacks. They do so by exploiting control over hardware such as fine resolution APIC timers and using TDX module memory management interface functions.

- **Single-Step Attacks** involve timing pin-based events such as interrupts, NMI, SMI and INIT to interrupt the guest TD execution after every instruction executed in the guest TD. This allows the attacker to examine the state of the machine following each instruction execution in interesting regions of code and use side channels to leak information used by that region of code.
- **EPT Fault Attacks** involve causing EPT violations or EPT misconfigurations to infer the control flow of execution inside a guest TD. Such control flow inference coupled with other side channel techniques, such as branch shadowing, can be used as a side channel to leak information from the guest TD.
- **Zero-Step Attacks** involve using an EPT fault on targeted instructions in a guest TD with an intent to glean side channel information from speculative execution past the faulting instruction. Such instructions are called “replay anchors”, as every resumption of the TD execution leads to the same EPT fault and thus the same speculative execution with the same stimulus to be replayed repeatedly, such that noise in side-channel observation of that speculative execution can be reduced.

17.3.2. Mitigation by the TDX Module

The Intel TDX module provides mechanisms to help assist in mitigating single and zero step attacks:

17.3.2.1. *Single-Step Attack Detection and Mitigation*

To help mitigate single-step attacks, the TDX module attempts to detect when a TD VCPU interruption by an interrupt, NMI, SMI or INIT event may indicate a single step attack. An attack is suspected if the interruption happens too soon after TD entry. Two methods are available for detection: instruction counting or heuristics.

Suspected Attack Detection Using Instruction Counting

This attack detection method is applicable if the TDX module implements Instruction-Count Single-Step Defense (ICSSD), as indicated by TDX_FEATURES0.ICSSD, readable by the host VMM using TDH.SYS.RD*. It is used only if the TD is not Perfmon-enabled, i.e., ATTRIBUTES.PERFMON is 0. An interruption is considered far enough from the last TD entry if either of the following conditions is true:

- More than one instruction has been retired since the last TD entry, or
- More than one round of a REP-prefixed instruction has been executed since the last TD entry.

Instruction counting is considered the better method of the two. If the TD’s ATTRIBUTES.ICSSD bit is set, then the TD will only be allowed to execute if instruction counting can be used.

Suspected Attack Detection Using Heuristics

If the instruction counting method can’t be used, either because it is not supported by the TDX module or because the TD is Perfmon-enabled (ATTRIBUTES.PERFMON is 1), then an interruption is considered far enough from the last TD entry if either of the following conditions is true:

- More than enough time (around 2 to 3 usec) has passed since the last TD entry, or
- RIP has changed by more than 32 bytes since the last TD entry, indicating that at least 2 instructions have been retired.

Attack Mitigation

If a suspected attack is detected, the TDX module doesn’t do an immediate asynchronous TD exit. Instead, it provides execution opportunities to the TD VCPU for a small random number of instructions, and only then does it deliver the interruption to the host VMM as an asynchronous TD exit.

17.3.2.2. *Zero-Step Attack Detection and Mitigation*

Suspected Attack Detection

For zero step attacks, the Intel TDX module counts Secure EPT faults that result in a TD exit. After a pre-determined number of such EPT violations occur on the same instruction, the TDX module starts tracking the GPAs that caused Secure EPT faults.

Note: For a partitioned TD, EPT violations that happen in the context of L2 VMs may result in a TD exit or in an L2→L1 exit. Only the TD exit cases are counted for the purpose of zero step attack detection.

Attack Mitigation

Once faulting GPA tracking starts, the TDX module prevents TD entry attempts to the VCPU if the previously faulting private GPAs are not properly mapped in the Secure EPT.

17.3.3. Host VMM Expected Behavior

No change is required to the host VMM's normal memory management behavior:

- The host VMM should block (TDH.MEM.RANGE.BLOCK) TD private pages and remove them (TDH.MEM.PAGE.REMOVE) only after the guest TD has explicitly relinquished the ownership of that page through a software protocol between the VMM and the TD. Such a protocol is implemented by the balloon driver mechanism employed by guest Linux kernel to allow the host VMM to overcommit a guest VM assigned memory.
- The host VMM can block TD private pages and perform the following GPA-to-HPA mapping updates without coordination with the guest TD:
  - Physical page relocation (TDH.MEM.PAGE.RELOCATE)
  - Mapping merge or split (TDH.MEM.PAGE.PROMOTE, TDH.MEM.PAGE.DEMOTE)
  - Unblock (TDH.MEM.RANGE.UNBLOCK)

  An attempt by guest TD VCPU to access such pages while they are blocked results in an EPT violation TD exit. A well-behaved host VMM should not re-enter the TD until the mapping operation is done. Failing to do so will immediately result in another EPT violation and the TD VCPU won't make any progress.

- The host VMM can block TD private pages for writing (TDH.EXPORT.BLOCKW) as part of TD migration. An attempt by the guest TD VCPU to write to such pages while they are blocked for writing results in an EPT violation TD exit. A well-behaved host VMM should not re-enter the TD VCPU before unblocking the page (TDH.EXPORT.UNBLOCKW). Failing to do so will immediately result in another EPT violation and the TD VCPU won't make any progress.

17.3.4. Guest TD Interface and Expected Guest TD Operation

The TDX module provides the guest TD with a notification facility, by which the guest TD can get notified when excessive Secure EPT violations are raised by the same TD instruction. This mechanism allows the guest TD to employ its own policies. The guest TD enables this notification by setting bit 0 of TDCS.NOTIFY_ENABLES field, using TDG.VM.WR. If this bit is set, then when more than a pre-determined number of Secure EPT violations are detected on the same instruction,

- If the EPT violation happened when L1 was running (i.e., the TD is not partitioned, or L1 VMM was running) the TDX module injects a #VE(ARCH) exception, with the EPT violation details.
- If the TD is partitioned, and the EPT violation happened when an L2 VM was running, the TDX module induces an L2→L1 exit. The EPT violation details are provided in the exit information.

As part of its normal memory management behavior, the guest TD should track its GPA space allocation and should only accept (TDG.MEM.PAGE.ACCEPT) PENDING pages that it expects to be added (TDH.MEM.PAGE.AUG) by the host VMM. Failing to do so would make the TD vulnerable to attacks, e.g., the host VMM could zero-out a page by removing it and adding a new one at the same GPA.

Thus, when the guest TD attempts to access a page and a #VE is raised indicating an EPT violation, the expected guest TD's #VE handler behavior is as follows:

- If this page is not known to the guest TD as owned by it, i.e., it was not added at TD build time (TDH.MEM.PAGE.ADD) and has not been added dynamically (TDH.MEM.PAGE.AUG) and accepted (TDG.MEM.PAGE.ACCEPT), the guest TD can accept this page normally.
- Otherwise, this may indicate an attack, and the guest TD can employ its own policy. For example, the guest TD may halt if this page is one of the pages expected to be resident when a security critical workload is executing.

Alternatively, it may signal the current running application so that the application would employ application-specific defenses.

The guest TD's #VE handler, as well as its virtual NMI handler, should not have any secrets that are susceptible to leakage.

The Intel TDX module does not provide protection against attacks when accessing **shared** pages. The guest TD should treat shared memory access as communicating with a potential attacker and not do any secure processing while accessing shared memory.

18. General Aspects of the Intel TDX Interface Functions

18.1. Concurrency Restrictions and Enforcement

18.1.1. Explicit Concurrency Restrictions

Intel TDX functions may specify concurrency restrictions on accessing one or more resources, as described below. In most cases, the restriction applies for the duration of the instruction execution. However, in some cases, the restriction applies for a longer duration. For example, TDH.VP.ENTER requires shared access to the TD-scope logical control structures TDR and TDCS, and it also requires shared access to TDVPS – the VCPU-scope logical control structure which applies during logical TDX non-root operation through TD Exit.

Table 18.1: Concurrency Restrictions of Intel TDX Functions or Flows

Concurrency Restriction	Description	Examples
Exclusive Access	During the period when an LP has exclusive access to a certain resource, any attempt by another LP to concurrently execute an instruction that requires either exclusive or shared access to the same resource will fail.	• TDH.VP.CREATE requires exclusive access to the TDVPR page.
Shared Access	During the period when an LP has shared access to a certain resource, any attempt by another LP to concurrently execute an instruction that requires exclusive access to the same resource will fail. No such restriction exists on another LP that attempts to concurrently execute an instruction that requires shared access.	• TDH.VP.CREATE requires shared access to the TDR page. • TDH.PHYMEM.CACHE.WB requires shared access to the KOT.

Software is expected to comply with the specified concurrency restrictions. The Intel TDX module helps enforce them (using internal locks) for proper TDX operation.

Table 18.2: Concurrency Restrictions Cross-Table

		Logical Processor Y		
	Concurrency Restriction	**Exclusive**	**Shared**	**None**
Logical Processor X	**Exclusive**	Not Allowed	Not Allowed	Allowed
	Shared	Not Allowed	Allowed	Allowed
	None	Allowed	Allowed	Allowed

Intel TDX functions do not wait on a resource that requires exclusive or shared access. If the resource is busy, the function fails immediately.

18.1.2. Implicit Concurrency Restrictions

In some cases, Intel TDX functions and whole flows (e.g., TD Entry through TD Exit) may have **implicit** exclusive or shared access to resources. This means that the access restriction is implied by the architecture, but no direct enforcement is made by the flow itself.

An important case is logical TDX non-root mode. TDH.VP.ENTER acquires shared locks on the TD's TDR and TDCS control structures and on the VCPU's TDVPS control structure. These shared locks are released only on TD exit. Thus, during all the time the LP is in the logical TDX non-root mode, including during TDCALL leaf functions, the LP has implicit shared access to TDVPS, TDR and TDCS.

18.1.3. Transactions

In some cases, Intel TDX module flows update some state as a transaction. They first read the current state, then do some calculations and eventually attempt to update the state using an atomic operation (e.g., LOCK CMPXCHG) to check that the state has not changed and set its new value. If that check fails, an Intel TDX module interface function may fail with a TDX_OPERAND_BUSY status.

18.1.4. Concurrency Restrictions with Host Priority

18.1.4.1. *Overview*

Host priority is a variant on explicit concurrency restrictions, where the host VMM side is given priority over guest TD side. A new HOST_PRIORITY flag is added to locks protecting resources that may be accessed by the host VMM and a guest TD. Both mutexes and shared/exclusive locks can be enhanced with host priority.

18.1.4.2. *Host-Side (SEAMCALL) Operation*

The host VMM is expected to retry host-side operations that fail with a TDX_OPERAND_BUSY status. The host priority mechanism helps guarantee that at most after a limited time (the longest guest-side TDX module flow) there will be no contention with a guest TD attempting to acquire access to the same resource.

Lock operations process the HOST_PRIORITY bit as follows:

- A SEAMCALL (host-side) function that fails to acquire a lock sets the lock’s HOST_PRIORITY bit and returns a TDX_OPERAND_BUSY status to the host VMM. It is the host VMM’s responsibility to re-attempt the SEAMCALL function until it succeeds; otherwise, the HOST_PRIORITY bit remains set, preventing the guest TD from acquiring the lock.
- A SEAMCALL (host-side) function that succeeds to acquire a lock clears the lock’s HOST_PRIORITY bit.

18.1.4.3. *Guest-Side (TDCALL) Operation*

A TDCALL (guest-side) function that attempts to acquire a lock fails if HOST_PRIORITY is set to 1; a TDX_OPERAND_BUSY status is returned to the guest. The guest is expected to retry the operation.

Guest-side TDCALL flows that acquire a host priority lock have an upper bound on the host-side latency for that lock; once a lock is acquired, the flow either releases within a fixed upper time bound or periodically monitors the HOST_PRIORITY flag to see if the host is attempting to acquire the lock.

18.1.4.4. *Host Priority Busy Timeout*

Once a host-side operation failed with a TDX_OPERAND_BUSY status, the host VMM should retry this operation until it no longer fails with the same TDX_OPERAND_BUSY status. Otherwise, the guest TD may be stuck trying to acquire a lock where the HOST_PRIORITY bit is set.

The TDX module implements a timeout mechanism for guest-side host priority lock acquisition failures. If the guest TD loops on a TDCALL (guest-side) function that fails with TDX_OPERAND_BUSY due to HOST_PRIORITY value of 1 for more than a configurable timeout, the TDX module initiates a trap-like TD exit with a TDX_HOST_PRIORITY_BUSY_TIMEOUT status. It is expected that this will never happen with a properly operating host VMM. However, it is still possible for the host VMM to resolve the lock contention by calling the SEAMCALL function that previously failed with a TDX_OPERAND_BUSY status until successful, and then re-entering the guest TD by calling TDH.VP.ENTER with the HOST_RECOVERABILITY_HINT bit cleared to 0. For details, see the [TDX Module ABI Spec].

The host priority timeout’s default value is 1 second. It is configurable between 10 msec to 100 seconds by using TDH.MNG.WR to update TDCS.HP_LOCK_TIMEOUT.

18.2. Memory and Resource Operands Access

Intel SDM, Vol. 3, 11.5.2	Precedence of Cache Controls
Intel SDM, Vol. 3, 11.11	Memory Type Range Registers (MTRRs)
Intel SDM, Vol. 3, 11.12	Page Attribute Table (PAT)

18.2.1. Overview

In this section, we discuss Intel TDX functions' memory and resource operands access from the following perspectives:

- Access semantics (shared, private, opaque and hidden)
- Explicit vs. implicit accesses
- Operand address specification (host-physical address, guest-physical address)
- Actual memory access (read or write) vs. memory reference

18.2.1.1. Access Semantics

Access semantics, as used in this document, convey the intended purpose of the access. Intel TDX functions are designed to use one of the following access semantics when accessing their memory and/or platform resource parameters:

Table 18.3: Access Semantics Definition

Access Semantics	Description	Intel TDX Module Usage
Shared	Memory is accessed using one of the shared HKIDs (in the range 0 to MAX_MKTME_HKIDS - 1). This is mostly used for memory parameters accessed by the VMM.	• Source page of TDH.MEM.PAGE.ADD • Memory operands of TDCALL leaf functions
Private	The memory is mapped in the TD's private GPA space. Memory accessed using the target TD's private HKID (in the range MAX_MKTME_HKIDS - 1 to MAX_HKIDS - 1). Such memory pages can be mapped in the TD's private GPA space.	• TD private pages • Secure EPT pages
Opaque	Memory is addressable by the host VMM, but its content is not directly accessible to software or devices. Memory is encrypted using either the Intel TDX global private key (for TDR) or the TD's ephemeral private key (for other control structures).	• TDR • TDCX • TDVPR
Hidden	Access is to an Intel TDX module internal resource. That resource is not directly addressable as a memory operand to software or devices.	• KOT • WBT

Note that on guest-side (TDCALL) functions, shared vs. private semantics is determined by the GPA provided as an operand to the function. A specific TDCALL leaf function may or may not impose a private or a shared access – e.g., TDG.MEM.PAGE.ACCEPT requires a private GPA, while TDG.MR.REPORT may work with either a private GPA or a shared GPA.

18.2.1.2. Explicit vs. Implicit Access

Explicit memory access is defined as an access where the memory location is provided as explicit operand to an Intel TDX function. The address may be provided directly in a GPR or indirectly via some address field in a software-accessible memory data structure.

The HKID for accessing the memory can be inferred by the instruction – implicitly or explicitly from the explicitly provided access.

Implicit memory access is defined as an access to a platform physical memory address, or to some other resource, that is not passed as an operand of an instruction (either directly or indirectly) but is implied by use of the Intel TDX function. TDX architecture helps guarantee that an implicit access is performed correctly, or a proper error action is taken.

18.2.1.3. *Memory Operand Address Specification*

Host-side Intel TDX functions (SEAMCALL leaf functions) memory operands are specified using their **host-physical address (HPA)**, their **guest-physical address (GPA)**, or when a GPA-to-HPA mapping is done (e.g., TDH.MEM.PAGE.ADD) by **both HPA and GPA**.

In most cases, HPA for private or opaque access semantics must be specified with all HKID bits set to 0.

Guest-side Intel TDX functions (TDCALL leaf functions) memory operands are specified using their **guest-physical address (GPA)**.

18.2.1.4. *Memory Type*

18.2.1.4.1. Memory Type for Private and Opaque Accesses

The memory type for **private** and **opaque** access semantics, which use a private HKID, is WB.

18.2.1.4.2. Memory Type for Shared Accesses

Intel SDM, Vol. 3, 28.2.7.2 Memory Type Used for Translated Guest-Physical Addresses

The memory type for **shared** access semantics, which use a shared HKID, is determined as described below. Note that this is different from the way memory type is determined by the hardware during non-root mode operation. Rather, it is a best-effort approximation that is designed to still allow the host VMM some control over memory type.

- For **shared access during host-side (SEAMCALL) flows**, the memory type is determined by MTRRs.
- For **shared access during guest-side flows (VM exit from the guest TD)**, the memory type is determined by a combination of the Shared EPT and MTRRs.
  - If the memory type determined during Shared EPT walk is WB, then the effective memory type for the access is determined by MTRRs.
  - Else, the effective memory type for the access is UC.

18.2.1.5. *Actual Memory Access vs. Memory Reference*

In some cases, Intel TDX functions only **reference** memory – i.e., use its address, but no actual access is done.

In other cases, Intel TDX functions **access** the memory – i.e., perform read or write (but not execute) operations.

Summary Table

18.2.1.6.

Table 18.4: Memory Access Summary

Explicit/ Implicit	Intel TDX Function	Access Semantics	Address Operand	HKID Derivation	Memory Type	Example
Explicit	Host-Side (SEAMCALL Leaf)	Shared	HPA	Derived HPA operand's HKID bits	From MTRR	SRCPAGE operand of TDH.MEM.PAGE.ADD
		Private	HPA	TD's HKID	WB	Target page of TDH.PHYMEM.PAGE.RECLAIM
			GPA	TD's HKID	WB	CHUNK operand of TDH.MR.EXTEND
			HPA and GPA	TD's HKID	WB	Target page of TDH.MEM.PAGE.ADD
		Opaque	HPA	TD's HKID or Intel TDX global HKID	WB	TDVPR operand of TDADDVPR
	Guest-Side (TDCALL Leaf)	Shared	GPA	From Shared EPT	From Shared EPT and MTRR	REPORTDATA operand of TDG.MR.REPORT
		Private	GPA	TD's HKID	WB	Target page of TDG.MEM.PAGE.ACCEPT
Implicit	All	Private/ Opaque	N/A	TD's HKID or Intel TDX global HKID	WB	TDCS access by TDH.VP.ENTER
		Hidden	N/A	N/A	N/A	KOT access by TDH.MNG.KEY.CONFIG

18.3. Register Operands and CPU State Convention

Intel SDM, Vol. 3, 24.9	VM-Exit Information Fields
Intel SDM, Vol. 3, App. C	VMX Basic Exit Reasons

18.3.1. Overview: Regular vs. Transition Leaf Functions

Intel TDX functions can be divided into transition functions and non-transition functions.

The **non-transition functions** are where SEAMCALL and TDCALL leaf functions behave as emulated CPU instructions from the perspective of the host VMM and the guest TD, respectively. In those cases, the meaning of input and output register operands is straightforward – similar to CPU instructions.

Transition cases are SEAMCALL(TDH.VP.ENTER) and TDCALL(TDG.VP.VMCALL) leaf functions, where a full cycle (until the start of the next instruction) includes TD transitions to the guest TD or host VMM, respectively, and back to the host VMM or guest TD, respectively. In those cases, we look at the functions from the point of view of the caller. The meaning of input and output register operands is more complicated.

Both cases are explained in the following sections and in the function reference sections.

18.3.2. Interface Function Leaf and Version Numbers

Interface functions are selected by a leaf number, provided in RAX. A version number enables supporting multiple versions of the same function, if required for backward compatibility. Unless otherwise specified, the default version number is 0.

Table 18.5: Intel TDX Interface Functions Leaf and Version Numbers in RAX

Bits	Field	Description
15:0	Leaf Number	Selects the SEAMCALL or TDCALL interface function
23:16	Version Number	Selects the SEAMCALL or TDCALL interface function version
63:24	Reserved	Must be 0

18.3.3. CPU State Preservation Convention

18.3.3.1. *TDH.VP.ENTER*

TDH.VP.ENTER is a special case. In addition to explicit output operands, TDH.VP.ENTER is not designed to preserve the extended CPU state that the TD may use according to TDCS.XFAM.

The host VMM is expected to save any state it needs before calling TDH.VP.ENTER. Details are provided in the TDH.VP.ENTER leaf function definition (see the [TDX Module ABI Spec]).

18.3.3.2. *Other Interface Functions*

All Intel TDX functions except TDH.VP.ENTER are designed to preserve the CPU state not explicitly defined as output.

Most interface functions preserve the AVX, AVX2 and AVX512 state. There are some exceptions, as described in the specific function definitions (see the [ABI Spec]):

- TDG.VP.VMCALL may use some XMM registers to pass information to and from the host VMM.
- Some interface functions may reset AVX, AVX2 and AVX512 state and/or the APX state (if the CPU supports it) to the architectural INIT state.

18.3.4. Transition Cases: TD Entry and Exit

18.3.4.1. *TD Entry: TDH.VP.ENTER*

Transfer of Host VMM State to TD Guest

By design, in the case of a TDH.VP.ENTER leaf function that follows a previous TDG.VP.VMCALL, the RCX input parameter of the previous TDG.VP.VMCALL is used as a bitmap. It selects the GPRs (from RBX, RDX, RBP, RDI, RSI and R8 through R15) and XMM registers whose value is transferred to the guest TD as-is. RAX is set to 0. See the TDG.VP.VMCALL description in the [TDX Module ABI Spec].

The rest of the CPU state is restored from the TD VCPU state as saved on TDG.VP.VMCALL.

Output State (Back to the Host VMM)

On completion of TDH.VP.ENTER, a success – indicated by the ERROR bit (RAX[63]) being 0 – means that TD Entry into the TD guest was successful. The TD guest ran for some time and then exited to the Intel TDX module. That exit can be due to execution of TDG.VP.VMCALL) or due to an asynchronous exit (e.g., an EPT Violation). The Intel TDX module then executes SEAMRET, transferring control to the instruction following TDH.VP.ENTER. In this case, the DETAILS field (RAX[31:0]) format is designed to be the same as the VMX **Exit reason**.

If the completion of TDH.VP.ENTER (i.e., exit from the TD guest) was due to TDCALL(TDG.VP.VMCALL), then the RCX input parameter of TDG.VP.VMCALL is designed to be used as a bitmap. It selects the GPRs (from RBX, RDX, RBP, RDI, RSI and R8 through R15) and XMM registers whose value is passed to the host VMM as the output of TDH.VP.ENTER. RCX itself is passed as-is to the output of TDH.VP.ENTER, and RAX[31:0] indicates the **VMCALL** exit reason (see below). See the TDG.VP.VMCALL description in the [TDX Module ABI Spec].

If the completion of TDH.VP.ENTER was due to another reason, then other VMX-like Exit Information fields are provided in other GPRs. Details are provided in the TDH.VP.ENTER leaf function definition (see the [TDX Module ABI Spec]).

By design, any GPRs and extended states that do not return values as described above are set to synthetic values. If the VMM uses any of them, it must explicitly save them before TDH.VP.ENTER and restore them afterward.

18.3.4.2. *TD Synchronous Exit: TDG.VP.VMCALL*

Transfer of TD Guest State to Host VMM

In the case of a TDG.VP.VMCALL leaf function, the RCX input parameter of TDG.VP.VMCALL is designed to be used as a bitmap. It selects the GPRs (from RBX, RDX, RDI, RSI and R8 through R15) and XMM registers whose value is passed to the host VMM as the output of TDH.VP.ENTER. RCX itself is passed as-is to the output of TDH.VP.ENTER.

RAX provides TDH.VP.ENTER completion status (see above). All other CPU state components, including GPRs and XMM registers not selected by RCX, are saved in TDVPS and set to fixed values (see the [TDX Module ABI Spec]). The value of RCX itself is also saved to TDVPS.

Output State (Back to the Guest TD)

On completion of TDG.VP.VMCALL, a success – indicated by the ERROR bit (RAX[63]) being 0 – means that a SEAMRET into the VMM was successful. The VMM ran for some and then executed TDH.VP.ENTER successfully (possibly on another LP). The Intel TDX module executed VMRESUME successfully, transferring control to the instruction following TDCALL.

In this case, the RCX input parameter of TDG.VP.VMCALL is designed to be used as a bitmap. It selects the GPRs (from RBX, RDX, RDI, RSI and R8 through R15) and XMM registers whose value reflects their state as input to TDH.VP.ENTER. All other CPU states, including GPRs and XMM registers not selected by RCX, are restored from TDVPS.

18.4. Interface Function Completion Status

Intel TDX function completion status is returned in RAX. The status is structured to provide as many details to software about error conditions as practically possible. At the same time, the status enables software to ignore details that it does not need. Software may parse the completion status at three detail levels, as described below.

18.4.1. Least Detailed Level: Success/Warning/Error

At this simplest level, software can differentiate between three cases:

Table 18.6: Intel TDX Interface Functions Completion Status in RAX at the Least Detailed Level

RAX Value	Meaning	Description
0	Success	Function completed successfully
Positive (0x00000000_00000001 – 0x7FFFFFFF_FFFFFFFF)	Informational / Warning	Function completed successfully, but with some informational or warning code – e.g., TDH.PHYMEM.PAGE.RECLAIM of a TDCX page that is already not VALID
Negative (0x80000000_00000000 – 0xFFFFFFFF_FFFFFFFF)	Error	Function aborted due to some error

18.4.2. Medium Detailed Level: Class, Recoverability and Fatality

At this level, software can understand the following information:

Table 18.7: Intel TDX Interface Functions Completion Status in RAX at the Medium Detailed Level

Name	Description
CLASS	Class of the function completion status
ERROR	Indicates that the instruction was aborted due to error
NON_RECOVERABLE	Recoverability hint: • In case of an error, it indicates that the error is probably not recoverable. • In case of a TD exit, it indicates that the TD failed in a non-recoverable way.

Name	Description
FATAL	Fatality hint – applicable only for SEAMCALL (TDH.*), indicates that the TD entered a state where it can only be torn down.
HOST_RECOVERABILITY_HINT	As a TDH.VP.ENTER output, indicates a TDCALL (TDG.*) that resulted in a trap-like TD exit for which the host VMM needs to provide a recoverability hint in the following TD entry. As a TDCALL (TDG.*) output, indicates that the host VMM provided a hint that the error is probably not recoverable.

18.4.3. Most Detailed Level

At this level, software can understand more details of an error that happened – e.g., if TDH.VP.ADDCX fails, software may understand if it is due to a wrong number of TDCX pages or due to the VCPU already being initialized.

Refer to the [TDX Module ABI Spec] for a detailed definition of function completion status.

18.5. TD, VM and VCPU Identification

Table 18.8: TD and VCPU Identification

Identifier	Format	Details
TD Handle	TDR HPA	While residing in memory, a TD is uniquely identified by the **TDR page HPA**, serving as the **TD handle** input operand of TDX module host-side interface functions. TDR HPA may change when, e.g., a TD is migrated.
TD Universally Unique Identifier	256-bit integer	TD_UUID serves as a globally unique TD identifier, randomly created when the TD is created. TD_UUID survives migration.
VCPU Handle	TDVPR HPA	While residing in memory, a VCPU is uniquely identified by the **TDVPR page HPA**, serving as the **VCPU handle** input operand of TDX module host-side interface functions. TDVPR HPA may change when, e.g., a TD is migrated.
VCPU Index	16-bit integer	A sequential VCPU index is assigned when the VCPU is created. VCPU index survives migration.
VM Index	16-bit integer	VM index identifies a VM within a TD. • VM index 0 identifies the L1 VMM. • VM indices higher than 0 identify L2 VMs.

18.6. Metadata Access Interface

18.6.1. Introduction

Metadata access interface is the architecture that allows representing TDX metadata, i.e., TD non-memory state and TDX module control state, in a way that is independent of the way it is stored. It does this by hiding the memory format of TDX control structures and allowing abstraction of data, as follows:

- The actual fields stored in the TD control structures may be different than their abstract representation. E.g., a TDVPS field may be provided as a GPA to TDH.VP.WR, while internally stored as an HPA.
- Access to a TD metadata field may **trigger some operation**. E.g., writing the TD VMCS's "posted-interrupt descriptor address" control triggers the verification of related control and may enable posted interrupt handling.
- TD metadata fields may be completely **virtual**, i.e., there may be no actual control structure fields represented by them.

Metadata abstraction is used in the following cases:

- Read of TDX Module information by the host VMM and guest TD using the following SEAMCALL and TDCALL functions:
  - Single Field Read: TDH.SYS.RD, TDG.SYS.RD
  - All Fields Read: TDH.SYS.RDALL, TDG.SYS.RDALL
- Read and write of TDR, TDCS and TDVPS control structures by the host VMM using the following SEAMCALL functions:
  - Single Field Access: TDH.MNG.RD, TDH.MNG.WR, TDH.VP.RD, TDH.VP.WR
- Read and write of TDR, TDCS and TDVPS control structures by the guest TD using the following TDCALL functions:
  - Single Field Access: TDG.VM.RD, TDG.VM.WR, TDG.VP.RD, TDG.VP.WR
- Read and write of TDR and TDCS a service TD using the following TDCALL functions:
  - Single Field Access: TDG.SERVTD.RD, TDG.SERVTD.WR
- For **TD migration**, export and import of TD metadata by the host VMM using the following SEAMCALL functions:
  - State Export: TDH.EXPORT.STATE.IMMUTABLE, TDH.EXPORT.STATE.TD, TDH.EXPORT.STATE.VP
  - State Import: TDH.IMPORT.STATE.IMMUTABLE, TDH.IMPORT.STATE.TD, TDH.IMPORT.STATE.VP

18.6.2. Metadata Fields and Elements

Metadata fields are identified by **field identifiers (MD_FIELD_ID)**. A field identifier contains a **FIELD_CODE** and other information. A detailed description and MD_FIELD_ID values are defined in tables provided in the [TDX Module ABI Spec]. Metadata fields size may be up to 128 bytes.

For the purpose of metadata abstraction interface, fields are divided into multiple **field elements**; the size of each element can be 1, 2, 4 or 8 bytes. Elements in a field have consecutive field codes, incremented by 1 or 2 as encoded in by the field identifier's INC_SIZE.

Figure 18.1 below shows an example of a SHA384 fields (e.g., TDCS.MRCONFIGID), whose size is 48B. When access using the metadata access functions, this field is divided into six 8-byte elements.

	Element 0	**Element 1**	**Element 2**	**Element 3**	**Element 4**	**Element 5**
FIELD_CODE	X	X + 1	X + 2	X + 3	X + 4	X + 5
Content	Bytes 7:0	Bytes 15:8	Bytes 23:16	Bytes 31:24	Bytes 39:32	Bytes 47:40

Figure 18.1: Example of a 48 Byte TDCS.MRCONFIGID Field Composed of Six 8 Byte Elements

A detailed definition of a field identifier is provided in the [TDX Module ABI Spec].

18.6.3. Arrays of Metadata Fields

Metadata fields can be organized in arrays. Figure 18.2 below shows an example of an array of 4 fields, each composed of 1 element. In this case, fields in the array have consecutive field codes, incremented by 1 or 2 as encoded in by the field identifier's INC_SIZE field.

Array Index	**Field Code**	**Content**
0	X + 0	Array[0]
1	X + 1	Array[1]
2	X + 2	Array[2]
3	X + 3	Array[3]

Figure 18.2: Example of an Array of 4 Single-Element Fields

Figure 18.3 below shows an example where each field is composed of multiple elements. TDCS.RTMR is such a case. The base FIELD_ID of each field in the array is incremented by the number of elements in each field, multiplied by 1 or 2 as encoded in by the field identifier's INC_SIZE field.

Array Index	Base FIELD_ID	Element 0's FIELD_ID	Element 1's FIELD_ID	Element 2's FIELD_ID	Element 3's FIELD_ID	Element 4's FIELD_ID	Element 5's FIELD_ID
0	X + 0	X + 0	X + 1	X + 2	X + 3	X + 4	X + 5
1	X + 6	X + 6	X + 7	X + 8	X + 9	X + 10	X + 11
2	X + 12	X + 12	X + 13	X + 14	X + 15	X + 16	X + 17
3	X + 18	X + 18	X + 19	X + 20	X + 21	X + 22	X + 23

Figure 18.3: Example of an Array of Four 48 Byte TDCS.RTMR Fields, Each Composed of 6 Elements

18.6.4. Metadata Field Sequences

Field sequences contain one or more whole metadata fields, each composed of one or more elements. A sequence is composed of a sequence header and one or more values.

- All fields in a sequence have the same CONTEXT_CODE, CLASS_CODE and field size (i.e., the same number of elements and the same element size).
- Each element is a sequence occupies 8 bytes, even if its size is 1, 2 or 4 bytes. When a sequence is used as an output of the TDX module, the upper bytes beyond the element size are zeroed out. When a sequence is used as an input of the TDX module, the upper bytes are ignored.
- The FIELD_CODEs of each element in a sequence are consecutive.
- A field sequence may contain a write mask, which applies to each element value in the sequence. This is applicable when the sequence is used for writing bit fields, e.g., VMCS execution controls.
- A sequence always contains whole fields, i.e., if a field is composed of multiple elements, the sequence contains all of them.

A **field sequence header** contains the initial field code and other information – for a detailed description see the [TDX Module ABI Spec].

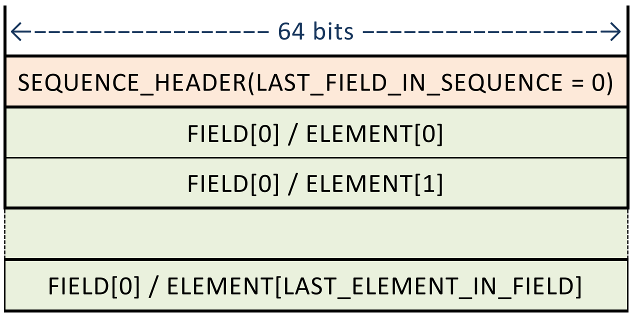

64 bits
SEQUENCE_HEADER(LAST_FIELD_IN_SEQUENCE = 0)
FIELD[0] / ELEMENT[0]
FIELD[0] / ELEMENT[1]
FIELD[0] / ELEMENT[LAST_ELEMENT_IN_FIELD]

Figure 18.4: Example of a Metadata Field Sequence with One Field Composed of Multiple Elements

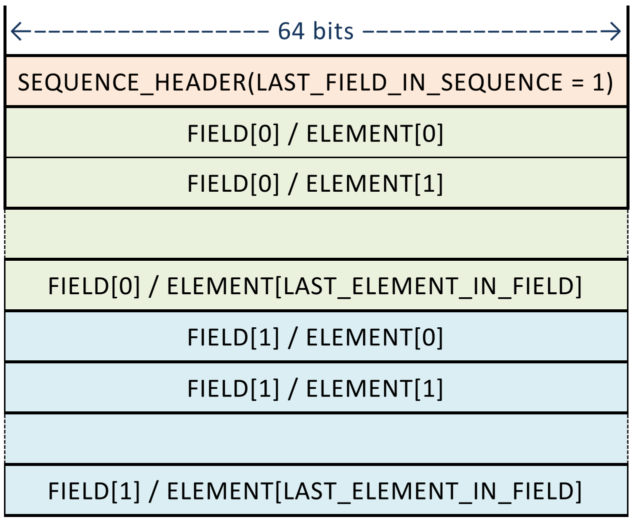

64 bits
SEQUENCE_HEADER(LAST_FIELD_IN_SEQUENCE = 1)
FIELD[0] / ELEMENT[0]
FIELD[0] / ELEMENT[1]
FIELD[0] / ELEMENT[LAST_ELEMENT_IN_FIELD]
FIELD[1] / ELEMENT[0]
FIELD[1] / ELEMENT[1]
FIELD[1] / ELEMENT[LAST_ELEMENT_IN_FIELD]

Figure 18.5: Example of a Metadata Field Sequence with 2 Fields Composed of Multiple Elements

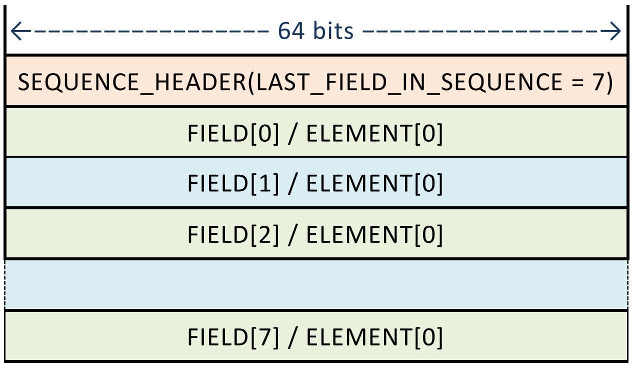


Figure 18.6: Example of a Metadata Field Sequence with 7 Fields Composed of a Single Element

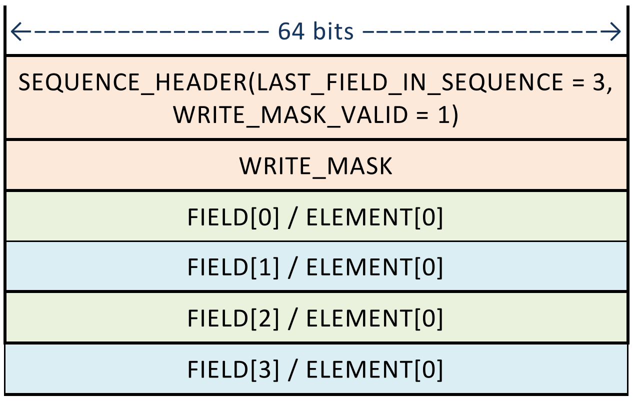


Figure 18.7: Example of a Metadata Field Sequence with a Write Mask

18.6.5. Metadata Lists

A metadata list is composed of a list header and one or more field sequences. The list header specifies list buffer size in bytes and the number of sequences. Metadata lists are used, e.g., for exporting VCPU metadata by THD.EXPORT.STATE.VP.

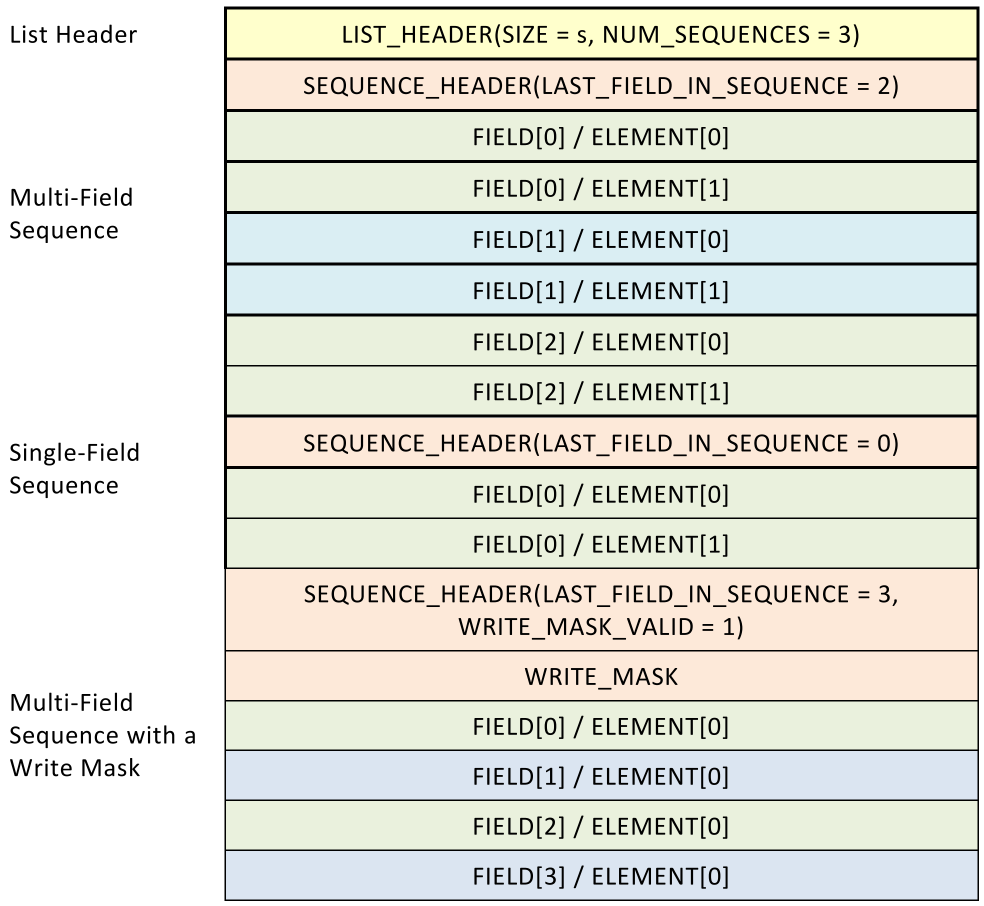


Figure 18.8: Metadata List Example

The metadata list header format is defined in the [TDX Module ABI Spec].

18.7. Interrupt Latency

18.7.1. Introduction

While the TDX module is running, no external events (interrupts, NMI, SMI, INIT) are recognized. To support proper system responsiveness, the TDX module is designed to limit the latency from the time an external event occurs until the host VMM gets control and may respond to that event.

Applicable cases are:

- Events that occur during local host-side flows (i.e., SEAMCALL interface functions except TDH.VP.ENTER).
- Events that occur during TD execution: from the beginning of TDH.VP.ENTER until completion of asynchronous TD exit.
- Events that occur during local guest-side flows (i.e., TDCALL interface functions except TDG.VP.VMCALL).
- Events that occur during synchronous TD exit (i.e., TDG.VP.VMCALL).

18.7.2. Latency of the Intel TDX Interface Functions

There are infrequent cases where the latency of some interface functions may be longer than normal, as listed below.

- Host-side interface functions that are invoked a limited number of times during TDX module lifecycle. The interface functions below are known to have longer than normal latencies:
  - TDH.SYS.INIT
  - TDH.SYS.LP.INIT
  - TDH.SYS.KEY.CONFIG
  - TDH.IDE.STREAM.IDEKMREQ
- Host-side interface functions that are invoked a limited number of times during TD lifetime. The interface functions below are known to have longer than normal latencies:
  - TDH.MNG.KEY.CONFIG

  - TDH.MNG.INIT
  - TDH.VP.INIT
- TDH.VP.ENTER may have a long latency if the single/zero step attack mitigation (described in 17.3) is activated due to a suspected attack.
- Guest-side interface functions that are invoked a limited number of times during TD lifetime. Such interface functions are rate-limited to avoid VM-DOS by the guest TD; see 11.16.5. The interface functions below are known to have longer than normal latencies (they may not be supported by some TDX module versions):
  - TDH.MR.ASSIGNSVNS

18.7.3. Interruptible Host-Side Interface Functions

Some SEAMCALL flows, which have long execution times, are designed to be interruptible. Interruptible flows check for pending events after doing some part of their work. E.g., TDH.EXPORT.MEM, which exports multiple 4KB pages, checks for pending events after processing each page. Detection of pending interrupts, but not of other events (NMI, SMI, etc.), is conditioned by the host VMM's interrupt enabling status (i.e., RFLAGS.IF).

There are two types of interruptible host-side interface functions:

- **Resumable** functions store their intermediate state securely before returning to the host VMM on interruption. They return a TDX_INTERRUPTED_RESUMABLE status. The host VMM is expected to call them again, indicating a resumption. Upon resumption, the resumable function continues from the point where it was interrupted.
- **Restartable** functions don't store any intermediate state before returning to the host VMM on interruption. They return a TDX_INTERRUPTED_RESTARTABLE status. The host VMM is expected to call them again. The resumable function starts from scratch.

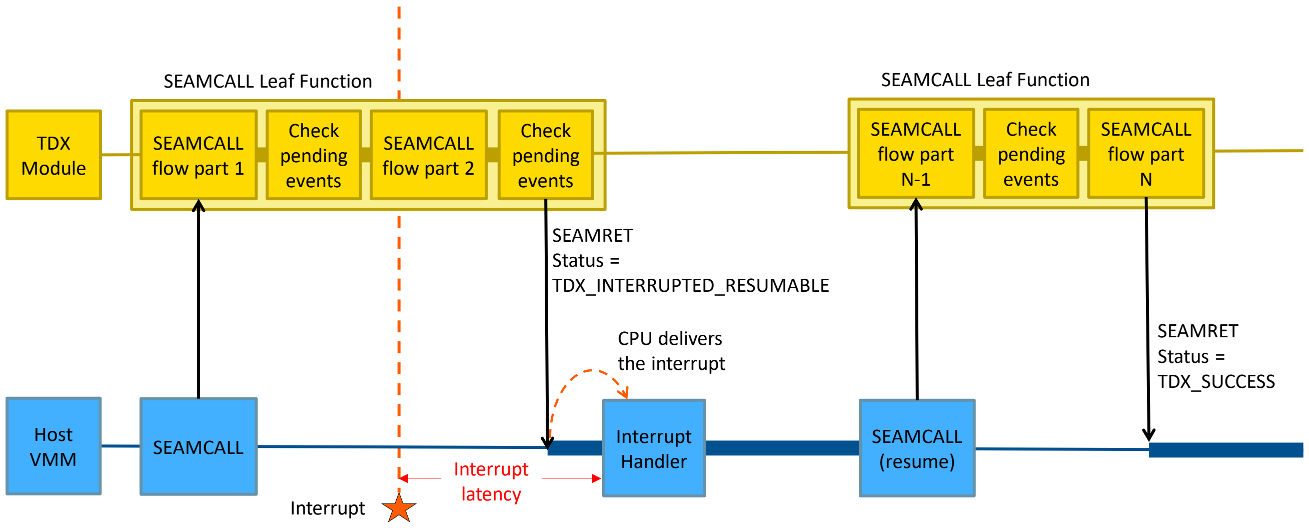


Figure 18.9: Typical Interruptible & Resumable SEAMCALL Leaf Function

18.7.4. Interruptible Guest-Side Interface Functions

Description

Some TCALL flows, which have long execution times, are designed to be interruptible. Interruptible flows check for pending events after doing some part of their work. E.g., TDG.MEM.PAGE.ACCEPT, which may need to initialize a 2MB page, checks for pending events after initializing each 4KB block. Contrary to SEAMCALL flows, detection of pending interrupts is not conditioned by the host VMM's interrupt enabling status (i.e., RFLAGS.IF).

When a pending event is detected, the interruptible TDCALL flow stores its intermediate state securely and resumes the guest TD. The guest virtual state, specifically RIP, is unmodified – except for the interruptibility state (STI Blocking and MOVSS Blocking). Thus, upon resuming the guest TD, the CPU delivered the event. This causes an asynchronous TD exit. The host VMM processes the TD exit information and processes the event, and then re-enters the TD. Since the guest RIP has not changed, the same TDCALL is executed again and the interrupted TDCALL flow is resumed.

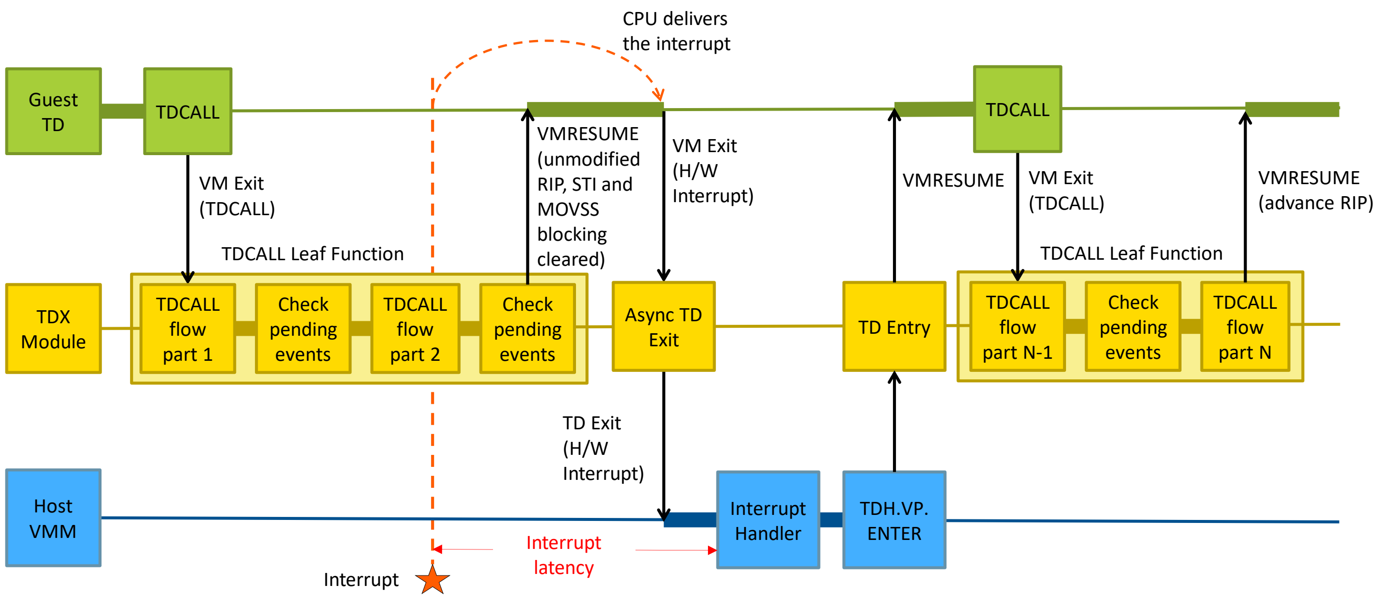


Figure 18.10: Typical Interruptible TDCALL Leaf Function – Hardware Interrupt Example

Posted Interrupts

In case the pending event is a posted interrupt notification, the posted interrupt is delivered by CPU when the guest TD is resumed, and the TD's interrupt handler is called.

Note: Guest TD software should treat guest side interface functions as functions calls, not as a single TDCALL instruction. The guest TD should not immediately precede an TDCALL to an interruptible leaf function with an STI instruction or a MOV to SS instruction. If it does, it should be aware that virtual interrupts will not be blocked until the completion of that TDCAL leaf function.

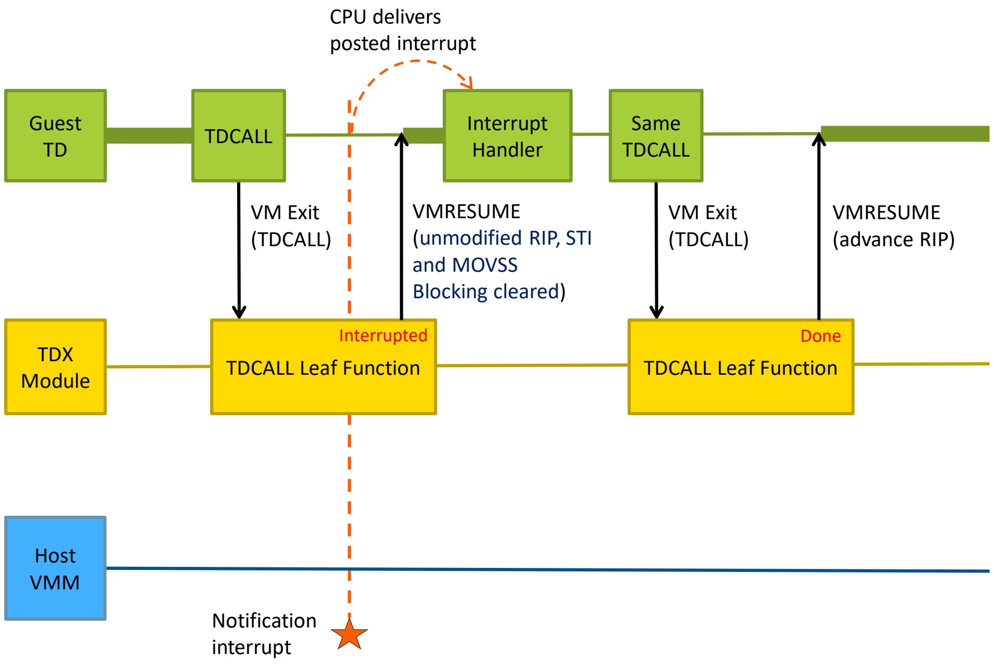


Figure 18.11: Typical Interruptible TDCALL Leaf Function – Posted Interrupt Example

Guest-Visible State when a Guest-Side Function is Interrupted

Interruptible guest-side functions may be designed in either of the following ways:

- For some interface functions, their effect is not visible to the guest TD when the function is interrupted; it becomes visible to the guest TD only when the whole function is completed. An example of such interface function is TDG.MEM.PAGE.ACCEPT. An accepted 2MB page becomes accessible to the guest TD only when TDG.MEM.PAGE.ACCEPT is successfully completed.

- Other interface functions are designed so they execute in parts, and each completed part is visible to the guest TD as soon as it is completed. An example of such interface function is TDG.VP.INVGLA. It works on a list of GLAs; when interrupted, is adjusts the list information provided in RDX to reflect the work completed so far.

18.7.5. Rate-Limited Guest-Side Interface Functions

Some guest-side (TDCALL) interface functions may have longer execution time than the maximum for supporting the interrupt latency requirements but are difficult to design as interruptible. For example, TDG.MR.ASSIGNSVNS uses crypto functions that are difficult to break into smaller units.

For such rate-limited functions, the rate at which the guest TD VCPU can call them is limited by the TDX module. If the guest TD calls such functions before 200usec have passed since the last successful call, the TDX module induces a TD exit with a TDX_TDCALL_RATE_LIMIT status, without modifying the guest CPU state.

The host VMM can then introduce a delay before calling TDH.VP.ENTER to resume the guest TD VCPU. After the following TD entry, the rate-limited function is called again; in this case the TDX module doesn't check the time passed since the last call.

18.8. DRNG Entropy Errors

Multiple TDX module interface functions may use the CPU's digital random number generation (DRNG) facility, via the RDSEED, RDRAND and PCONFIG instructions. Such interface functions may fail on entropy errors due to the inability of the DRNG to generate random values at the requested rate, returning a TDX_RND_NO_ENTROPY or a similar status code. This may happen where, e.g., software running on other LPs in the same package also executes RDSEED at a high rate.