

# Confidential, Attestable, and Efficient Inter-CVM Communication with Arm CCA

Sina Abdollahi\*, Amir Al Sadi\*, Marios Kogias\*, David Kotz<sup>†‡</sup>, Hamed Haddadi\*

\*Imperial College London, London, United Kingdom

{s.abdollahi22, a.al-sadi, m.kogias, h.haddadi}@imperial.ac.uk

<sup>†</sup>Dartmouth College, Hanover, NH, USA

David.F.Kotz@dartmouth.edu

**Abstract**—Confidential Virtual Machines (CVMs) are increasingly adopted to protect sensitive workloads from privileged adversaries such as the hypervisor. While they provide strong isolation guarantees, existing CVM architectures lack first-class mechanisms for inter-CVM data sharing due to their disjoint memory model, making inter-CVM data exchange a performance bottleneck in compartmentalized or collaborative multi-CVM systems. Under this model, a CVM’s accessible memory is either shared with the hypervisor or protected from both the hypervisor and all other CVMs. This design simplifies reasoning about memory ownership; however, it fundamentally precludes plaintext data sharing between CVMs because all inter-CVM communication must pass through hypervisor-accessible memory, requiring costly encryption and decryption to preserve confidentiality and integrity.

In this paper, we introduce *CAEC*, a system that enables protected memory sharing between CVMs. *CAEC* builds on Arm Confidential Compute Architecture (CCA) and extends its firmware to support Confidential Shared Memory (CSM), a memory region securely shared between multiple CVMs while remaining inaccessible to the hypervisor and all non-participating CVMs. *CAEC*’s design is fully compatible with CCA hardware and introduces only a modest increase (4%) in CCA firmware code size. *CAEC* delivers substantial performance benefits across a range of workloads. For instance, inter-CVM communication over *CAEC* achieves up to 209× reduction in CPU cycles compared to encryption-based mechanisms over hypervisor-accessible shared memory. By combining high performance, strong isolation guarantees, and attestable sharing semantics, *CAEC* provides a practical and scalable foundation for the next generation of trusted multi-CVM services across both edge and cloud environments.

**Index Terms**—Confidential Computing, Arm CCA, Trusted Execution Environment, Attestation

## 1. Introduction

Trusted Execution Environments (TEEs) have emerged as a cornerstone for building secure systems in the presence of powerful adversaries. By providing isolated execution, integrity guarantees, and confidentiality for code

and data even against a compromised operating system or hypervisor, TEEs enable developers to deploy sensitive applications on untrusted infrastructure.

Although initially the TEE landscape was rather heterogeneous with different vendors offering different mechanisms, *e.g.*, Intel providing enclaves with SGX [1] and Arm offering physical memory partitioning with TrustZone [2], over the years there is a convergence towards confidential virtual machines (CVMs) as the prevalent abstraction for confidential computing. Currently, major vendors have developed CVM offerings. Technologies such as AMD SEV-SNP (Secure Encrypted Virtualization-Secure Nested Paging) [3] and Intel TDX (Trust Domain Extensions) [4] have already been adopted by cloud providers [5]–[9], while Arm has similarly introduced Confidential Compute Architecture (CCA) [10], [11], which is expected to be deployed across both edge devices and cloud servers.

Despite design variations across vendors, all existing CVM architectures adopt a *disjoint memory model*, in which each CVM owns a protected memory isolated from both the hypervisor and other CVMs [3], [4], [11]. While each CVM may share memory with the hypervisor (*e.g.*, for I/O), it cannot create protected memory regions shared with other CVMs only and not visible to the hypervisor. This design simplifies reasoning about memory ownership and provides strong confidentiality and integrity guarantees for the CVM’s protected memory region.

Such a disjoint memory model though, leads to substantial efficiency losses in CVMs. In practice, this design forces all inter-CVM communication to be exposed to the hypervisor, either through hypervisor-managed services such as virtual sockets (mode (a) in Fig. 1) or through hypervisor-accessible shared memory regions (mode (b) in Fig. 1). Both approaches require expensive encryption and decryption on the CVM sides to preserve confidentiality and integrity when data is transmitted through hypervisor-accessible memory, while the use of confidential paravirtual devices is an open challenge both in terms of performance and correctness [12]–[15]. Furthermore, the disjoint memory model prevents memory savings through deduplication, where identical pages are shared across CVMs. This inefficiency becomes particularly problematic not only in large-scale datacenter environments where DRAM is a major cost factor, but also in memory-constrained edge and embedded devices, where TEEs are increasingly deployed.

Unfortunately, communication and memory inefficien-

<sup>‡</sup>This work was performed while Professor Kotz was in residence at Imperial College London.

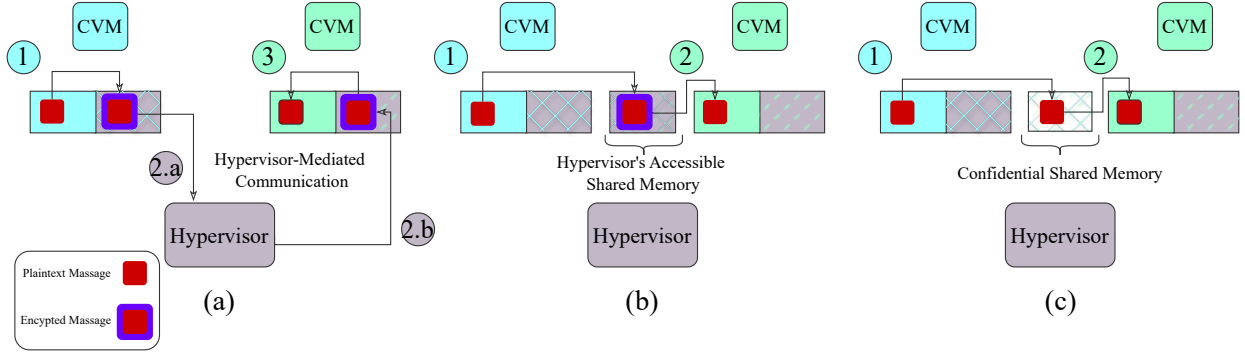


Figure 1: Communication modes between two CVMs. (a) Communication through virtualization services, where data must be encrypted and passed via the hypervisor-mediated service. (b) Communication using shared memory provided by the hypervisor, still requiring encryption as the shared memory is accessible to the hypervisor. (c) *CAEC*, which enables CSM between CVMs. There is no need for encryption as *CAEC* protects the CSM from hypervisor and other CVMs. Cyan: memory regions accessible to the first CVM, Green: memory regions accessible to the second CVM, Purple: memory regions accessible to the hypervisor.

cies substantially affect modern machine learning (ML) and agentic systems, which increasingly dominate confidential computing workloads [5]–[9]. These systems are increasingly becoming compartmentalized across CVMs, for example, splitting the networking and inference stacks into different components to reduce the attack surface [5], [6], [16]. Moreover, ML and agentic systems provide large opportunities for memory deduplication and sharing between CVMs given the overlap in the models and contexts used across different applications, which can be multiple gigabytes.

To address this limitation, in this paper, we introduce Confidential Shared Memory (CSM), a hypervisor-protected (confidential) memory which can be shared between multiple CVMs. Designing a system that (1) enables CSM regions within a confidential computing environment and (2) restricts their use exclusively to mutually attested realms provides CVMs with a protected memory for exchanging plaintext data directly with each other (mode (c) in Fig. 1). This capability substantially improves CPU efficiency and memory utilization of inter-CVM data exchange.

Enabling CSM may appear straightforward, yet it is a challenging undertaking. Depending on the architecture, it may be infeasible without hardware modifications, for example, in AMD SEV-SNP (Section 8.2). Even systems that enable similar memory sharing between enclaves (user-space TEEs) fail to provide flexible and scalable mechanisms that fit the needs of modern ML systems. For instance, Plug-In Enclaves [17] enable read-only shared enclaves, which can be used to effectively share static resources such as libraries in serverless applications. Cerberus [18] introduces formal techniques to verify such a sharing with the same read-only access model. However, both designs are insufficient for applications that require extensive inter-process communication (IPC), where writes must be visible to the other side. Elasticlave [19], on the other hand, supports writable shared memory using RISC-V Physical Memory Protection (PMP). However, the underlying isolation mechanism inherently limits the number of shareable regions: the total number of memory partitions (protected or shared) (e.g., to 16) due to

limitations in the underlying isolation technology (riscv’s Physical Memory Protection), thus, restricting both the number of coexisting enclaves and the number of shared regions.

Even after extending a confidential computing architecture to support CSM across multiple CVMs, protecting the shared memory from attacks by other CVMs remains challenging. Participating CVMs must be able to discover and attest each other while being assured that the shared region cannot be accessed or modified by unauthorized CVMs. Without a principled ownership model and explicit access-control mechanisms, the system risks undermining the confidentiality of the CSM, enabling adversarial CVMs to exploit the CSM management interface to escalate their privileges or gain unauthorized access to shared regions.

**Contribution.** In this paper, we introduce *CAEC*, a system that enables CSM between CVMs in Arm CCA. To the best of our knowledge, *CAEC* is the first approach to support CSM between CVMs. *CAEC* leverages Arm CCA’s RISC architecture to enable CSM without requiring any modifications to the CCA hardware. It provides dynamically manageable memory sharing between CVMs without imposing any limitation on the number of shared regions, either per CVM or system-wide. Through these capabilities, *CAEC* enables efficient plaintext communication and the direct sharing of large resources (e.g., ML models) between mutually attested CVMs. We implement *CAEC* on the latest functional and performance prototypes of Arm CCA. Our evaluation shows that *CAEC* achieves substantial performance improvements in inter-CVM communication and memory sharing such as up to  $209\times$  reduction in CPU cycles during communication. In summary, we claim the following contributions:

- *CAEC* extends the CCA firmware to support the CSM. It introduces a principled ownership model, explicit access-control rules, and attestation extensions that ensure CSM remains protected from the hypervisor and all unauthorized CVMs. Our analysis shows that *CAEC* prevents unauthorized access to CSM while preserving CCA’s security guarantees for all non-CSM regions.

TABLE 1: Memory access control applied by Granule Protection Check (GPC) in CCA

Security State	Normal PAS	Secure PAS	Realm PAS	Root PAS
Normal	✓	✗	✗	✗
Secure	✓	✓	✗	✗
Realm	✓	✗	✓	✗
Root	✓	✓	✓	✓

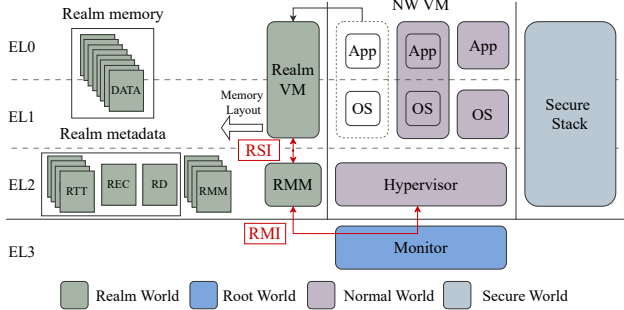


Figure 2: Arm CCA 1.0 software architecture. The hypervisor allocates resources (e.g., memory and CPU) for realm VM but cannot access those resources, as the realm VM is running on the other side of the isolation boundaries.

- *CAEC* integrates cleanly across all layers of the CCA stack, providing flexible CSM management capabilities to realms while retaining the hypervisor’s authority over physical memory management. *CAEC* achieves this with only a 4% increase in firmware size.
- *CAEC* demonstrates significant benefit for communication and data sharing between CVMs. *CAEC* achieves up to a 209× reduction in CPU cycles compared to encryption-based mechanisms over hypervisor-accessible shared memory. Moreover using *CAEC*, big data like a LLM can be shared between two CVMs, resulting in 16.6%–28.3% reduction in the system’s memory footprint.

## 2. Background & Motivation

In this section, we review the foundational concepts required for the remainder of the paper. We begin with an overview of Arm CCA (Section 2.1), followed by a discussion of two key components of the architecture: the Realm Management Monitor (Section 2.2) and the attestation framework (Section 2.3). We conclude this section with the motivation behind this work (Section 2.4).

### 2.1. Arm CCA

Arm CCA [10] is a series of hardware and software extensions for Armv9-A architecture (Fig. 2). Arm CCA extends the Armv9-A architecture with realm world<sup>1</sup> and root world, orthogonal to the existing normal world (NW) and secure world. Each 4 KB frame of physical memory (also referred to as a granule) is tagged with the world it belongs to at any given time. This ownership information is recorded in a structure called Granule Protection Table

1. In some references, the term execution environment is used instead of world; however, in this work, the two terms are used interchangeably.

(GPT). A hardware mechanism called Granule Protection Check (GPC) enforces access restrictions based on both the ownership state of each granule (as recorded in the GPT) and the current processor state. A memory access is permitted only if it complies with the rules defined in Table 1. In particular, when the processor operates in the root world state, it has access to the physical address space (PAS) of all other worlds. When operating in the realm or secure world, the processor can access the normal world’s PAS, but the realm and secure worlds are isolated from each other. The normal world has no access to the PAS of any other world. Arm CCA also leverages the isolation primitives of the Arm architecture—such as exception levels (ELs) and virtualization. Exception levels begin with EL3, the highest privilege level in the system, while EL2–EL0 provide intra-world privilege separation. The architecture support for virtualization includes two stages of address translation: mapping virtual addresses (VAs) to intermediate physical addresses<sup>2</sup> (IPAs) and mapping IPAs to physical addresses (PAs).

Within CCA, the architecture software stack includes the Monitor running at EL3, responsible for initially booting all EL2 components, managing GPTs, and context switching between worlds. The normal world stack consists of a hypervisor operating at EL2, virtual machines (VMs) running at EL1 and EL0, and user-space apps running at EL0. Secure world can host a stack similar to the NW, however, it is usually reserved for vendor specific services, impossible to run third party code. The realm world stack consists of realm VMs (or simply realms) running at EL1 and EL0 and a lightweight firmware known as Realm Management Monitor (RMM). In CCA, the hypervisor retains control over system resources such as CPU cores and physical memory. It can allocate or reclaim resources for realms much like it does for NW VMs. However, in the case of realms, the RMM acts as a trusted mediator between the hypervisor and realms. It validates all hypervisor requests concerning realm resources and proceeds only if they satisfy CCA’s isolation and security requirements.

### 2.2. Realm Management Monitor

The RMM is the trusted firmware component in CCA responsible for coordinating all interactions between the hypervisor and realms. It enforces *isolation* by protecting resources delegated to realms from both the hypervisor and other realms, while maintaining the hypervisor’s capability to *manage* those resources. To achieve this, the RMM exposes Realm Management Interface (RMI) to the hypervisor and the Realm Service Interface (RSI) to realms [20]. The hypervisor can issue RMI commands to request operations on realm resources such as memory delegation or vCPU (virtual CPU) scheduling. However, before executing these operations, the RMM performs a series of validity checks and proceeds only if all validations succeed. The RSI also can be used by realms to access services such as attestation and hypercalls. The RMM manages four types of granules during the lifecycle of each realm (Fig. 2): (1) the Realm Descriptor (RD), which

2. We adopt Arm’s term intermediate physical address (IPA), which is also commonly referred to as guest physical address (GPA).

defines the realm’s general attributes (*e.g.*, address space size); (2) the Realm Execution Context (REC), which stores vCPU-related state (*e.g.*, system registers); (3) the Realm Translation Tables (RTTs), a hierarchical structure maintaining IPA-to-PA mappings and access permissions; and (4) DATA granules, which represent protected memory regions accessible only to realm software.

To allocate a new DATA granule to realm during runtime, the hypervisor must first delegate a granule to the realm world using `RMI_GRANULE_DELEGATE`. It then requests the RMM to create the corresponding mapping in the realm’s protected address space<sup>3</sup> using `RMI_DATA_CREATE_UNKNOWN`. Upon receiving the request, the RMM verifies that the granule has already been delegated to the realm world (*i.e.*, inaccessible to the hypervisor) and that it is not already mapped in the protected address space of another realm, thereby enforcing CCA’s disjoint memory model. After creating the new mapping in the RTT, the granule becomes accessible to realm.

### 2.3. Attestation

In CCA, a realm can obtain an *attestation token* via RSI. The attestation token is a set of claim values and their signature that describe the state of the realm and the platform on which it runs [20], [21]. A token includes claims such as the Realm Initial Measurement (RIM), which captures the realm’s configuration and initial memory contents. Because the RMM computes these claims and signs them as part of the token, a remote verifier (*e.g.*, a realm owner) can (1) validate their authenticity, ensuring that each claim was generated by the trusted RMM and has not been tampered with and (2) check whether these claims match the expected values.

### 2.4. Motivation

The traditional disjoint memory model of CVMs forbids any form of CSM between CVMs. This design is reasonable under the common assumption that CVMs do not trust one another. However, it is too restrictive for emerging workloads, which increasingly require CVMs—while still isolated from each other—to collaborate once attestation confirms that a peer is running an expected and acceptable software stack. These collaborating CVMs may belong to different administrative parties (*e.g.*, agentic systems [22] or collaborative learning [23]), or they may belong to a single organization but be compartmentalized for security (*e.g.*, separating networking and inference stack [6]). Under the current trust model, such workloads cannot efficiently exchange plaintext data or share large identical memory regions (*e.g.*, LLM model weights), even though the shared content itself poses no confidentiality risk. Meeting these emerging demands requires rethinking the architecture of confidential computing: CVMs must be able to share CSM with attested peers while remaining fully protected from untrusted or malicious CVMs.

3. A realm’s address space consists of two halves: the protected half, used to map realm-world granules, and the unprotected half, used to map normal-world granules.

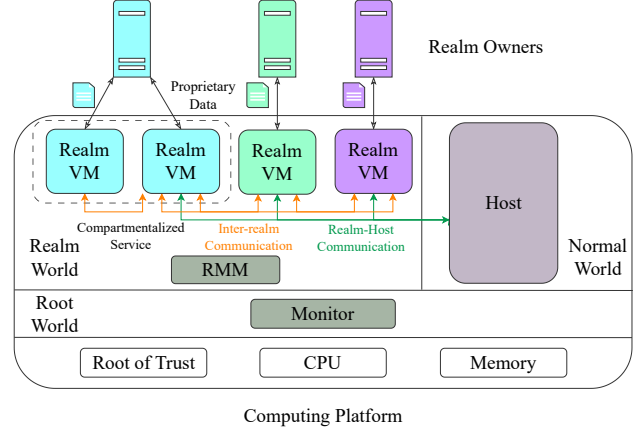


Figure 3: CAEC System Model

**Arm CCA for CSM.** Arm CCA introduces architectural properties that make it a promising candidate for supporting flexible and secure CSM. First, unlike AMD SEV-SNP or Intel TDX, which rely extensively on hardware and microcode extensions, CCA places most of its trusted functionality in firmware (*i.e.*, the RMM) while relying on a small set of hardware mechanisms to enforce isolation [24], [25]. This design provides better flexibility for extending the architecture with new features. For example, CCA uses the GPC hardware mechanism to isolate the realm world from the normal world, but isolation between realms is enforced entirely through RMM-managed checks. Because these checks reside in firmware rather than hardware, they can be extended or relaxed through firmware updates. Similarly, the RMM can incorporate new metadata types and validation rules without requiring hardware or microcode modifications—capabilities that are essential for securely managing CSM regions across realms. Second, in contrast to the physical partitioning mechanism used in Elasticclave [19], CCA’s use of virtualization can be adapted to support an unbounded number of regions, both within a single CVM and across the system, albeit with increased design complexity. Overall, these properties suggest that CCA provides a solid foundation for implementing a scalable and fine-grained CSM mechanism within a confidential computing architecture.

## 3. System & Threat Model

In this section, we present system model (Section 3.1) and threat model (Section 3.2) of CAEC.

### 3.1. System Model

CAEC’s system model is illustrated in Fig. 3. CAEC is designed to operate across both edge platforms (*e.g.*, smartphones) and cloud platforms that support the Armv9-A architecture with CCA extensions. The platform boots the CCA firmware, consisting of the RMM and the Monitor, along with a hypervisor responsible for managing system resources. The hypervisor and all services running under its control are collectively referred to as the host. Realm owners are entities external to the computing platform that provide confidential services, either locally



(when the platform is an edge device) or remotely (in the cloud scenario). Examples of such services include model inference [26]–[28], private LLM agents [14], [22], digital rights management (DRM), and authentication services [29]. Realm owners are mutually untrusted with respect to one another and to the host; each seeks to protect its proprietary data from all other entities. Realm owners deploy realm(s) to deliver their services while preserving the confidentiality of their proprietary data from other entities. A single realm owner may further compartmentalize its service into multiple realms for stronger isolation. Realm owners may agree to collaborate and share data only through *attested realms and explicitly established CSM regions between those attested realms*. The use of realms to deliver confidential services and host proprietary data of external entities (realm owners) adhere to the design principle of CCA [10], [30] and align with other system models such as Android Virtualization Framework [31] and ASTER [32]. *CAEC* assumes that all entities in the system trust the device’s hardware and the CCA firmware. Finally, the hypervisor is assumed to provision the necessary resources (*e.g.*, memory and CPU time) to ensure forward progress of realms.

### 3.2. Threat Model

Following the Arm CCA threat model [30], *CAEC* assumes that the host and all realms are mutually untrusted. For any given realm, both the host and other realms may attempt to compromise the confidentiality or integrity of its protected memory or vCPU state. This includes collaborating realms, which may try to read or modify memory outside the mutually agreed CSM boundaries. A realm may also attempt to obtain additional resources (*e.g.*, granules) beyond those delegated by the host or retain resources longer than permitted, which may result in denial-of-service (DoS) attack against the host or other realms. Introducing CSM support creates additional attack vectors. Because *CAEC* exposes CSM creation and access to existing ones as a generic service available to all realms (including malicious realms), an adversarial realm might attempt to impersonate an attested peer to gain unauthorized access to an existing CSM. Conversely, it may create a fake CSM region and trick other realms into using it, thereby compromising the confidentiality or integrity of shared data. Such an adversarial realm may belong to a competing realm owner or be instantiated by the host itself with the intent of targeting a specific CSM configuration. Physical attacks and microarchitectural side-channel attacks are out of scope. The platform is assumed to support secure boot, ensuring the trusted loading and execution of all EL3 and EL2 components and preventing boot-time compromise. Hardening of higher-level protocols and interfaces used by realms for inter-CVM communication is considered orthogonal to *CAEC*’s contribution.

## 4. CAEC

In this section, we introduce *CAEC*. We begin by describing the setup of each realm in Section 4.1. We then present the full lifecycle of CSM in Section 4.2, and finally detail several key design components, including the realm

identifier (Section 4.3) and physical memory management (Section 4.4).

### 4.1. Realm Setup

At a high level, *CAEC* extends Arm CCA with support for CSM, enabling realms to securely create, manage, and attach to CSM regions. The RMM serves as the trusted core of *CAEC*: it implements validation, enforces access control, and orchestrates all CSM-related operations. *CAEC* introduces new RSI commands to the RMM, which exposes the CSM-related services to realms (see Table 2).

**Ownership Model.** In *CAEC* each CSM region has exactly one creator and lifetime manager, referred to as its Provider realm (P-realm). The P-realm (1) contributes a portion of its protected address space to instantiate the CSM region, (2) grants or revokes other realms’ access to it, and (3) determines the access permissions types (*e.g.*, read-only or read-write) of each participating realm. Realms that later join an existing CSM region are referred to as Consumer realms (C-realms). Crucially, these roles are per-region and not global: a realm may act as a P-realm for some CSM regions and a C-realm for others.

**Allocation Semantics.** *CAEC* preserves the hypervisor’s control over physical memory management for both CSM and non-CSM regions, aligning with Arm CCA’s design principles. Since a CSM region is shared across multiple realms, memory-management flows must diverge from those used for regular private memory of realms. Accordingly, the RMM notifies the hypervisor whenever CSM regions are created, shared, or removed (through vCPU exits, see Table 2). The hypervisor uses this information to populate the P-realm’s CSM region while ensuring that C-realms do not receive additional physical granules beyond what the P-realm explicitly shares (see Section 4.4 for further details). *CAEC*, however, guarantees that once a CSM region is established, all participating realms observe an identical and consistent view of the underlying physical granules.

**Access Policy Table.** The RMM in *CAEC* requires to effectively track the state and ownership of each CSM region. In particular, whenever a realm issues a CSM-related RSI command, the RMM must consult—and potentially update—the corresponding metadata to ensure that all security checks are correctly enforced. In CCA, however, the RMM does not store per-realm metadata directly in its internal memory; instead, it relies on metadata granules that the hypervisor delegates to each realm (*e.g.*, RTT granules). Following this design principle, *CAEC* introduces a new metadata structure, Access Policy Table (APT), associated with each realm. The APT allows the RMM to record metadata for all CSM regions mapped into a realm’s protected address space. APT record one entry per CSM region, which is either P-realm entry or C-realm entry (depending on the role of the realm in that CSM). C-realm entries store information such as base address, size, the identifier of the owner (P-realm), while P-realm entries keep additional metadata such as permission type of C-realms authorized to access that CSM. To prevent race conditions or deadlocks, *CAEC* reuses the RMM’s existing locking mechanisms [33], ensuring that multiple vCPUs cannot access or modify realm’s APT concurrently.

TABLE 2: New and modified commands introduced by *CAEC*

Type	Name	Description
<b>Caller: Hypervisor</b>	<b>Callee: RMM</b>	
New RMI	<code>err = RMI_APT_CREATE(PA<sub>RD</sub>, PA<sub>APT</sub>)</code>	Allocates the granule at physical address PA <sub>APT</sub> for the realm's APT.
New RMI	<code>err = RMI_APT_DESTROY(PA<sub>RD</sub>, PA<sub>APT</sub>)</code>	Destroys the realm's APT. This command is only invoked during realm destruction.
Modified RMI	<code>err = RMI_DATA_CREATE_UNKNOWNS(PA<sub>RD</sub>, PA<sub>D</sub>, IPA<sub>D</sub>)</code>	Maps a data granule D at physical address PA <sub>D</sub> into intermediate physical address IPA <sub>D</sub> . CAEC ensures that if IPA <sub>D</sub> is in a CSM region, new mappings are created for other participating realms.
Modified RMI	<code>err = RMI_DATA_DESTROY(PA<sub>RD</sub>, IPA<sub>D</sub>)</code>	Wipes and unmaps the data granule in IPA <sub>D</sub> . CAEC ensures that if IPA <sub>D</sub> is in a CSM region, it is unmapped for other participating realms.
<b>Caller: P-realm</b>	<b>Callee: RMM</b>	
New RSI	<code>ID<sub>CSM</sub> = RSI_CSM_CREATE(IPA<sub>base-P</sub>, Size<sub>CSM</sub>)</code>	Creates a CSM region beginning at IPA <sub>base-P</sub> with size Size <sub>CSM</sub> . Returns CSM identifier ID <sub>CSM</sub> .
New RSI	<code>ID<sub>CSM-PC</sub> = RSI_CSM_SHARE(ID<sub>CSM</sub>, ID<sub>C</sub>, P)</code>	Shares a CSM region ID <sub>CSM</sub> with a C-realm identified by ID <sub>C</sub> and access permission P. Returns the sharing identifier ID <sub>CSM-PC</sub> .
New RSI	<code>err = RSI_CSM_REVOKE(ID<sub>CSM-PC</sub>)</code>	Revokes access to an already existing sharing identified by ID <sub>CSM-PC</sub> .
New RSI	<code>err = RSI_CSM_DESTROY(ID<sub>CSM</sub>)</code>	Destroys a CSM region identified by ID <sub>CSM</sub> .
<b>Caller: C-realm</b>	<b>Callee: RMM</b>	
New RSI	<code>err = RSI_CSM_RESERVE(ID<sub>CSM-PC</sub>, IPA<sub>base-C</sub>, Size<sub>CSM</sub>)</code>	Reserves a region begins at IPA <sub>base-C</sub> with size Size <sub>CSM</sub> for sharing identified by ID <sub>CSM-PC</sub> .
New RSI	<code>err = RSI_CSM_ATTACH(ID<sub>CSM-PC</sub>)</code>	Maps the CSM with sharing identifier ID <sub>CSM-PC</sub> into the corresponding reserved region, delegating access to the CSM.
New RSI	<code>err = RSI_CSM_DETACH_AND_FREE(ID<sub>CSM-PC</sub>)</code>	Unmaps the CSM with sharing identifier ID <sub>CSM-PC</sub> from the C-realm private address space and frees the previously reserved region.
<b>Caller: RMM</b>	<b>Callee: Hypervisor</b>	
New Exit Reason	<code>REC_EXIT_C_REALM_CSM(IPA<sub>base-C</sub>, Size<sub>CSM</sub>)</code>	Notifies the hypervisor about a new CSM region in the C-realm's protected address space.
New Exit Reason	<code>REC_EXIT_P_REALM_CSM(IPA<sub>base-P</sub>, Size<sub>CSM</sub>)</code>	Notifies the hypervisor about a new CSM region in the P-realm's protected address space.
New Exit Reason	<code>REC_EXIT_REALM_REMOVE_CSM(IPA<sub>base-P</sub>, Size<sub>CSM</sub>)</code>	Notifies the hypervisor to remove a CSM region from realm's protected address space.

The physical address of RD PA<sub>RD</sub> must be provided within each RMI command.

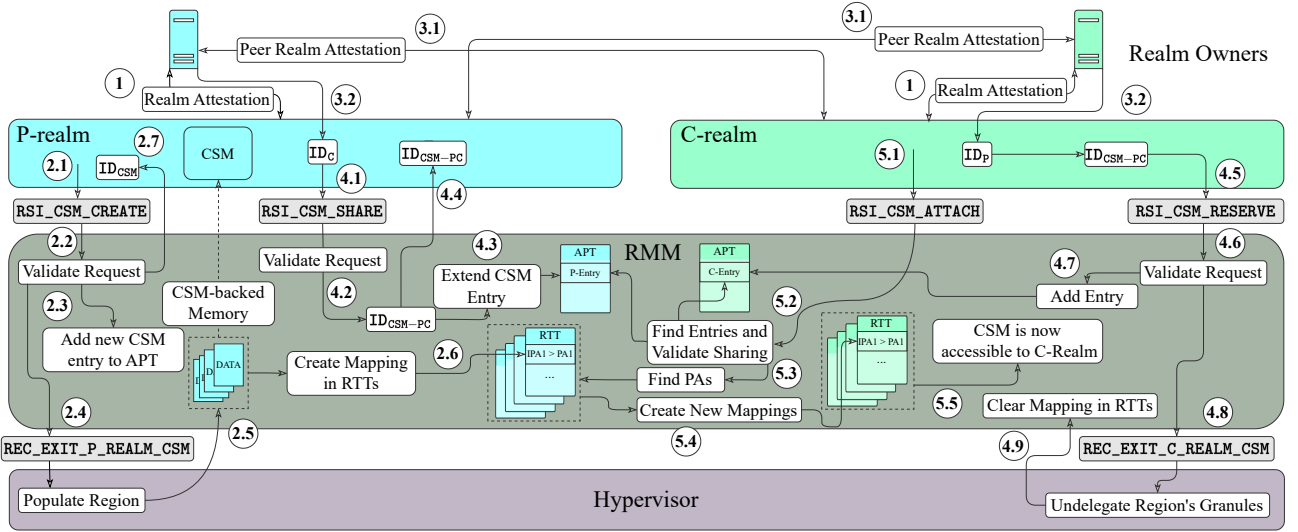


Figure 4: Overview of *CAEC*.

**Realm Initialization.** *CAEC* adapts the realm initialization flow of CCA, in which the hypervisor issues RMI commands [20] to populate the initial realm’s image and delegate realm’s metadata granules. The metadata granule introduced by *CAEC* (*i.e.*, APT) is similarly created and initialized through a new RMI command, `RMI_APT_CREATE` (see Table 2). At this stage, the APT is allocated for each realm within the RMM’s internal structures but does not yet contain any configuration. Once initialization is complete, the hypervisor begins scheduling the vCPUs of each realm. At this point, realms can communicate with external entities, including their respective realm owners. The realms establish TLS channel and prove themselves to their realm owner via attestation (step ① in Fig. 4).

## 4.2. CSM Lifecycle

**Creation.** A realm can create arbitrary CSM regions in its private address space via `RSI_CSM_CREATE`, providing the region’s start address and size (see [Table 2](#)). While *CAEC* does not restrict the number of CSM regions a realm may create, each region must (1) be granule-aligned and (2) not overlap with any existing CSM region in the realm’s private address space. If these conditions are

satisfied, the RMM registers this region within the APT of P-realm, notifies the hypervisor about the new CSM region via vCPU exit `REC_EXIT_P_REALM_CSM`. The hypervisor then populate the entire CSM range within the P-realm’s private address space (see [Section 4.4](#)). For every new delegated granule, the RMM creates the corresponding entry in the P-realm’s RTTs. The RMM later returns a unique region identifier  $ID_{CSM}$  to the realm (steps 2.1) to 2.7 in [Fig. 4](#)).

**Pre-Sharing.** Before sharing CSM, realms must be able to securely identify their peers. To enable this, *CAEC* adapts the attestation primitives of CCA and delegates peer realm attestation to the realm owners. Each realm owner attests the peer realm to verify that the expected software stack is running within it (step ③.1). In *CAEC*, the RMM assigns each realm a unique identifier that is used to refer to that realm during subsequent CSM-related operations. The RMM also reports these identifiers as a separate claim in the attestation token. As a result, realm identifiers can be reliably recovered by realm owners during peer realm attestation. After validating the attestation report (*i.e.*, checking that realm’s claims match the expected values), realm owners distribute the validated peer realm’s identifier to their respective realms (step ③.2). From this

point onward, both realms know each others identifiers. In [Section 4.3](#), we provide a detailed explanation on realm identifier and the way *CAEC* builds it.

**Sharing.** To allow sharing CSM between realms, *CAEC* adopts a simple but strong rule: a realm can attach to an existing CSM region only if both the P-realm and the C-realm explicitly agree to share and attach, respectively. P-realm can initiate sharing by issuing `RSI_CSM_SHARE`, providing the region identifier  $ID_{CSM}$ , the desired access permissions  $P$ , and the identifier of the C-realm  $ID_C$ . The RMM validates these inputs and, upon success, records a sharing identifier in the P-realm's APT entry associated to the CSM region, marking that the P-realm has agreed to share the CSM with the specified C-realm. It returns the sharing identifier  $ID_{CSM-PC}$  to the P-realm (steps (4.1) to (4.4)). The identifier is a deterministic concatenation of the P-realm and C-realm identifiers, combined with a counter to distinguish multiple shared regions between the same pair of realms.

On the C-realm side, it can regenerate the sharing identifier due to its deterministic format. The C-realm must then reserve a region of identical size in its own private address space using `RSI_CSM_RESERVE`. It provides the region's start address, size, and the region identifier  $ID_{CSM}$  to the RMM. The RMM verifies that the reserved range (1) is granule-aligned and (2) does not overlap with any existing CSM region already present in the C-realm's address space. If the checks succeed, the RMM adds a corresponding entry to the C-realm's APT, indicating that the C-realm has agreed to attach to the region shared by the designated P-realm. The RMM then notifies the hypervisor via `REC_EXIT_C_REALM_CSM`, prompting the hypervisor to undelegate any previously assigned granules in the reserved CSM range of the C-realm (steps (4.5) to (4.9)).

**Access.** To gain access to the shared region, the C-realm must finally issue `RSI_CSM_ATTACH` with the sharing identifier  $ID_{CSM-PC}$  provided. The RMM first locates the corresponding entries in both the P-realm's and the C-realm's APTs; the presence of both entries confirms that the P-realm has previously agreed to share the region and the C-realm has agreed to attach. *CAEC* does not require the CSM region to appear at the same base IPA address in each realm's private address space. However, both realms must use an identical size for the region; this ensures that neither realm gains unauthorized access to memory outside the mutually agreed-upon CSM boundaries. If these conditions are satisfied, the RMM begins constructing the required mappings in the C-realm's address space. To achieve that, the RMM walks through the P-realm's RTTs to locate the physical addresses corresponding to the CSM's range and creates equivalent mappings in the C-realm's RTTs (steps (5.1) to (5.5)). During this process, the RMM applies the access permissions previously specified by the P-realm to all new RTT entries of C-realm. Once these mappings are established, the C-realm can access the CSM region according to the assigned permissions.

**Revocation & Destruction.** The P-realm retains its control over sharing of the CSM with the C-realm. It can issue `RSI_CSM_REVOKE`, which removes the mapping from the C-realm's RTT, and clear the C-realm's identifier from the associate APT entry in P-realm APT. A P-realm

can also destroy a CSM region via `RSI_CSM_DESTROY` with the region identifier provided. In return the RMM, repeats the handler of `RSI_CSM_REVOKE` command for every peer C-realm and finally destroys the associated APT entry. The RMM then notifies the hypervisor about the destruction of CSM sharing in the P-realm's private address space via `REC_EXIT_REMOVE_CSM`. The C-realm can also unmap and free its address space from the CSM via `RSI_DETACH_AND_FREE` with the sharing identifier provided. In return, the RMM remove mapping from C-realm's RTT, removes the associated entry in the APT of C-realm, and notifies the hypervisor about the destruction of CSM sharing in the C-realm's private address space via `REC_EXIT_REMOVE_CSM`.

### 4.3. Realm Identifier

In the traditional model of CCA, all RMM services (exposed as RSI and RMI commands) are *local* to a single realm. By contrast, *CAEC* extends the RMM to support services *between* realms, enabling realms to share CSM with one another. This fundamentally requires realms to identify and refer to each other securely. For example, a realm invoking `RSI_CSM_SHARE` must specify the identifier of the target C-realm. The RMM must be able to reliably interpret these identifiers to enforce access control throughout the CSM lifecycle. If realm identifiers were forgeable or manipulable (e.g., by a malicious hypervisor), the security of *CAEC* would be fundamentally compromised. Thus, *CAEC* requires a secure, unforgeable mechanism for realm identification.

**CCA Identifier Model.** The RMM design in CCA provides no such system-wide realm identifier, as the RMM was not designed to support inter-realm operations. The RMM in CCA distinguishes realms solely by the physical address of their RD, which the hypervisor supplies with each RMI invocation. However, this identifier is unsuitable to be used as identifier between realms. First, exposing RD physical addresses as the identifier to entities such as realm owners and their associated realms meaning that they obtain information about the physical memory layout of the system, which violates virtualization principles. Second, since the hypervisor can control physical memory layout, it provides no uniqueness guarantee: a malicious hypervisor could terminate the current realm and instantiate a malicious one at the same RD address, causing both to appear to have the same identifier. This enables a classic time-of-check-to-time-of-use (TOCTOU) attack in which the hypervisor can replace a legitimate realm with a malicious one while preserving the same identifier. In summary, none of the realm-related arguments available in the current CCA design satisfies both essential properties: (1) not leaking system-level information, and (2) being protected from hypervisor manipulation. Candidates such as the RD physical address, the Realm Personalization Value (RPV), and the Virtual Machine ID (VMID) [20] all fail to meet both criteria.

**Realm Identifier in *CAEC*.** The RMM in *CAEC* assigns a unique system-wide identifier to each newly created realm and maintains a registry of these identifiers. These identifiers reveal no information about the underlying system and are safe from hypervisor manipulation, as



each realm receives a fresh identifier upon creation. The RMM further reports realm identifier as a separate claim within each attestation token, creating a cryptographic binding between the identifier and other claims—such as the RIM—within the attestation token. As a result, the realm identifier becomes an attestable property that third parties can verify alongside other claims. The attestation token in *CAEC* can prove (1) the realm’s software and platform configuration (as in *CCA*), and (2) the authenticity of the realm identifier (new in *CAEC*). Consequently, realm owners can use attestation to obtain authenticated identifiers of peer realms and safely distribute them to their own realms for subsequent CSM-related operations.

#### 4.4. Physical Memory Allocation

When a new CSM region is registered, the RMM notifies the hypervisor through the vCPU exits `REC_EXIT_P_REALM_CSM` and `REC_EXIT_C_REALM_CSM`. These notifications allow the hypervisor to finalize the physical memory configuration required for subsequent system operations. If the request corresponds to the creation of a CSM region by the P-realm, the hypervisor delegates granules for the entire CSM range, if they are not already delegated, ensuring that the region is fully populated and accessible to the P-realm. Conversely, if the request corresponds to attaching to an existing CSM region by a C-realm, the hypervisor undelegates any previously delegated granules that fall within the CSM range of that realm’s protected address space. This mechanism guarantees that the CSM region is instantiated exactly once in physical memory, maintaining consistency across participating realms and preventing redundant allocation.

For CSM creation, the hypervisor performs the following steps for each granule in the CSM range. It first issues `RMI_RTT_READ_ENTRY` to check whether the target IPA is already populated. If it is not populated, the hypervisor issues `RMI_GRANULE_DELEGATE` followed by `RMI_DATA_CREATE_UNKNOWN` to delegate a physical granule and create the corresponding mapping in the realm’s RTT, respectively. For CSM attachment, the hypervisor again invokes `RMI_RTT_READ_ENTRY`. If the IPA is already populated, it reclaims the granule using `RMI_DATA_DESTROY` and `RMI_GRANULE_UNDELEGATE`. The hypervisor also ensures that the RTTs of both the P-realm and the C-realm exist for the CSM range. If necessary, it creates them by issuing `RMI_GRANULE_DELEGATE` followed by `RMI_RTT_CREATE`. As outlined in the adversary model (Section 3.2), we assume that the hypervisor correctly allocates all required resources to ensure uninterrupted realm execution. However, failure to allocate these granules does not compromise the security guarantees of *CAEC*.

### 5. Security Analysis

In this section we provide an in-depth analysis of different attacks vectors introduced in threat model (Section 3.2), explaining how *CAEC* remains protected against these threats.

**CSM Protection.** As discussed in Section 2.1, the hardware-enforced GPC mechanism ensures that any access to realm-world memory from either the normal world

or the secure world is strictly prohibited. Since the RMM enforces that all CSM regions reside the realm world, CSM regions are inherently protected from direct access attempts originating from normal world and secure world actors.

*CAEC* defends the CSM against malicious realms through two complementary mechanisms: (1) A system-wide, attestation-integrated realm identifier, and (2) RMM-enforced access control checks within all CSM-related RSI commands. First, reporting realm’s identifier as a separate claim in the attestation token creates a cryptographic binding between realm identifier and its other claims such as RIM. Realm owners verify these claims and proceed only if they match the expected configuration, after which they provision the validated identifier to their associated realm. A malicious realm, whether instantiated by a competing realm owner or by the hypervisor, cannot bypass this attestation check because its software content (and thus its RIM) will not match the expected values. Consequently, such a realm cannot impersonate an attested peer, attach to an existing CSM, or create a fraudulent CSM region to lure honest realms. Second, after attestation, all CSM participation requests are subject to explicit access-control checks performed by the RMM. Both the P-realm and the C-realm must independently reference each other using their attested identifiers when issuing CSM-related RSI commands. This bidirectional referring rule ensures that, although a malicious realm may arbitrarily invoke CSM-related RSI commands, it cannot establish a CSM with a realm to gain access to a portion of its address space, unless the peer realm explicitly agrees.

Finally, the RMM performs required cache and TLB maintenance whenever mappings change. These operations flush stale translations, ensuring that no realm can retain access to the CSM after its permissions have been revoked.

**Non-CSM Protection.** *CAEC* guarantees that neither the P-realm nor the C-realm can access each other’s protected memory outside the agreed-upon CSM region. During the CSM sharing phase, the RMM validates that sharing occurs only at granule-aligned boundaries and for the exact size confirmed by both realms, thereby preventing malicious access outside the CSM.

**Ownership Protection.** When a C-realm issues `RSI_CSM_RESERVE`, the RMM explicitly verifies that the requested region does not overlap with any existing mappings. This restriction prevents a C-realm from delegating access to a CSM region to another realm without the P-realm’s consent. By ensuring that all sharing relationships originate from the P-realm, ownership and access remain fully traceable. Such control is crucial for allowing the P-realm to reason about which realms have visibility into the shared memory and for guaranteeing that, once access is revoked, no realm can retain or reacquire access to the CSM.

**Host Protection.** The RMM in *CAEC* imposes no constraints on how the hypervisor manages CPU scheduling or memory allocation for either CSM or non-CSM regions. As a result, realms cannot exploit *CAEC*’s services to obtain more memory or vCPU resources than those explicitly provisioned by the hypervisor. This design protects



TABLE 3: Line of code added to components in *CAEC*

TCB Component	Lines of Code (Extension)
RMM (v0.5.0) [34]	1062 (4%)
Non-TCB Component	Lines of Code
kvmtool-cca (v3/cca) [35]	247
Linux KVM (v5+v7) [36]	394
CSM Driver	467

the hosting system from DoS attacks originating from malicious or misbehaving realms.

## 6. Implementation

We implement and evaluate *CAEC* with both functional and performance prototypes of CCA. Our implementation includes both host-side and realm-side extensions to support the creation and use of CSM regions.

**Software Stack.** We use unmodified Trusted Firmware-A [37] (v2.11) as the Monitor. We adapt the Trusted Firmware reference implementation of the RMM [34] (v0.5.0), extending it to support the new and modified commands introduced by *CAEC* (Table 2). On the host side, we base our hypervisor on linux-cca [36] (v5+v7) and the virtual machine manager on kvmtool-cca [35] (v3/cca). We extend the Linux KVM (Kernel-based Virtual Machine) module [38], [39] to support the new RMI commands and to delegate physical memory to CSM regions in response to the newly defined vCPU exits from the RMM (Table 2). We modify kvmtool to insert a node into the guest device tree, reserving a portion of its private address space to be used as the CSM. On the guest side, we add CSM driver to linux-cca [36] (v5+v7). This driver discovers the reserved region from the device tree, issues the appropriate RSI commands to create or attach to CSM regions, and exposes CSM to user space as a character device. The same driver is used by both P-realms and C-realms.

Table 3 summarizes the code changes introduced by *CAEC*, measured in lines of code (LoC). *CAEC* adds 1,062 LoC to the RMM (29k LoC), resulting in approximately a 4% increase in its size. The RMM version used in our prototype (v0.5.0) targets CCA v1.0, which lacks planned features such as device assignment and planes (Section 8.1). These forthcoming features are expected to significantly increase the RMM’s size, making *CAEC*’s relative contribution to the TCB even smaller over time. On the host side, *CAEC* contributes 247 LoC to kvmtool, of which 213 LoC implement the PCI device used in our experimental setup (Section 7.3), and 34 LoC correspond to core *CAEC* functionality. *CAEC* also adds 394 LoC to the Linux KVM module. On the guest side, the CSM driver adds 467 LoC.

**Functional Prototype.** At the time of writing, two functional prototypes of Arm CCA are publicly available. Arm’s Fixed Virtual Platform (FVP) and Linaro’s QEMU [40] both provide emulated CCA-compatible hardware. We adopt FVP as our functional prototype because it is Arm’s official release, fully aligned with the CCA specification [11], [41], and widely used in prior work [14], [26], [42]–[44]. FVP models key hardware components of an Arm system, including the processors, cache hierarchy, bus traffic, and memory subsystem. It

is instruction-accurate, meaning that it correctly executes architecturally valid code, but it does not model cycle-accurate timing or the performance characteristics of real processors [41], [45]. Consequently, FVP is well-suited for evaluating the feasibility and functional correctness of *CAEC* on Arm CCA hardware, but it cannot be used for timing or performance measurements.

**Performance Prototype.** At the time of writing, no commercial hardware implements the Arm CCA extensions. For performance prototyping, we therefore adopt OPENCCA [46], the first open-source performance prototype of Arm CCA. Our performance evaluation can be easily reproduced as OPENCCA runs on cost-effective hardware (*i.e.*, Radxa Rock 5B [47]). OPENCCA constructs the realm world as a separate execution environment within the normal world. Because current off-the-shelf hardware lacks essential CCA extensions (*e.g.*, GPC), it cannot provide hardware-enforced isolated realm world. Nevertheless, performance prototyping remains feasible by emulating the realm world in software. This is achieved by splitting the normal-world software stack into two execution domains: one running the standard normal-world environment, and the other hosting the RMM and realm components. A modified Monitor then manages boot and context switching between these domains, effectively emulating transitions between the normal world and the realm world.

We acknowledge that such a prototype cannot fully capture the performance characteristics of real CCA hardware. For example, the absence of GPC in the system registers and memory system, can affect memory access behavior and caching, introducing discrepancies compared to a genuine CCA environment. While we do not claim that our performance results precisely reflect the cost of *CAEC* on production-grade CCA hardware, this setup provides a best-effort approximation until such hardware becomes available.

## 7. Evaluation

In this section, we evaluate *CAEC* by addressing the following key questions:

- **Q1:** Is *CAEC* fully compatible with CCA hardware?
- **Q2:** What is the performance *CAEC* through different tasks?

We answer **Q1** in Section 7.2 and **Q2** in Section 7.3, Section 7.4, and Section 7.5.

### 7.1. Experimental Setting

For functional prototype we set FVP to have two clusters, each with four cores supporting Armv9.2-A and 4GB of RAM. All performance experiments are conducted on a Radxa Rock 5B board [47] equipped with 16GB of RAM and an 8-core processor (4× ARM Cortex-A76 and 4× ARM Cortex-A55). We always create realm VMs with one vCPU and pin each vCPU to a specific core, with highest scheduling priority given to the realm’s vCPU process. For LLM-inference experiments, we use llama.cpp [48] as our inference engine.

## 7.2. Compatibility with CCA hardware

To evaluate the compatibility of *CAEC* with real CCA hardware, we repeated the data-sharing benchmark (Section 7.4) on FVP. We observed no runtime errors or stalls on FVP cores during CSM creation, inference execution, and CSM termination, demonstrating that *CAEC* can operate on Arm CCA-enabled hardware. In all experiments, we use the CSM driver to expose the CSM region to user-space processes.

## 7.3. Communication Benchmark

In this section, we evaluate the effectiveness of *CAEC* for communication between realms. We assume two user-space programs running in separate realms exchange messages via shared memory. The shared memory is provided either by *CAEC* or through a shared region in normal world. In each experiment, one side writes messages into the shared memory, and the other side reads them. For the NW shared memory case, we evaluate two configurations: (1) Encrypted communication, in which two sides employ an encryption/decryption protocol (using mbedTLS or OpenSSL) to each exchanged message; and (2) Plaintext communication, in which no encryption is applied. The encryption-based configuration provides confidentiality and integrity guarantees against an active NW adversary capable of intercepting or modifying shared-memory content. In contrast, when using *CAEC*, encryption is unnecessary because the CSM region is inherently protected from both the NW and other realms.

**Results.** Fig. 5 presents the results for CPU usage, latency (message delivery time), and throughput (data transferred per unit time). Across all metrics, *CAEC* consistently outperforms both encrypted NW-based configurations. Compared to OpenSSL, *CAEC* achieves  $24\times$ – $212\times$  lower latency,  $25\times$ – $209\times$  fewer CPU cycles, and  $26\times$ – $204\times$  higher throughput. Against MbedTLS, *CAEC* delivers  $4.4\times$ – $203\times$  lower latency,  $4.5\times$ – $200\times$  fewer cycles, and  $4.8\times$ – $194\times$  higher throughput. In all cases, the performance gap widens as message size increases. These results highlight the substantial overhead introduced by cryptographic protection when using NW shared memory for inter-realm communication. Because *CAEC* provides system-level protection for the CSM, it eliminates the need for encryption and its associated per-message costs. Fig. 5 also reports the performance of plaintext communication over NW shared memory. As the results indicate, *CAEC* achieves performance equivalent to plaintext NW communication, demonstrating that the previously observed differences between *CAEC* and encrypted modes over NW memory stem entirely from cryptographic processing rather than from differences in memory-access latency, caching effects, or other architectural factors.

For completeness, we note that in the encryption-based modes, each message consists of a small header—containing session\_id and seq—followed by the main payload. The receiver verifies the header for integrity, checks that seq matches the expected value (ensuring ordering and replay protection). Each side allocates a buffer on the shared memory to acknowledge the latest sent/received message. Synchronization between the two

side is achieved through polling over these buffers. Each experiment is repeated 1000 times to compute median values for latency and CPU usage, while the throughput measurement transfers 1000 messages sequentially. Finally, note that as kvmtool does not natively support sharing NW memory pages between two VMs, we extended it by implementing a new PCI device, with a design similar to ivshmem [49] in QEMU.

## 7.4. Data Sharing Benchmark

In this section, we evaluate *CAEC* in the context of sharing LLMs between realms. We assume a setup where multiple realms provide local inference services, each requiring access to a LLM. We measure the minimum RAM required to run ten typical inference in non-conversational mode, without observing a stall or memory-related error. We define the experiment under two configurations: (i) a baseline, where each realm maintains its own instance of the model, and (ii) model sharing between realms enabled by *CAEC*. We repeat these experiments with two and three realms, using models of different sizes (all 8-bit quantized).

**Results.** Table 4 summarizes the results, showing that model sharing between realms reduces the overall system memory footprint by 16.6% to 28.3%, depending on the model size and number of realms. The reduction becomes more pronounced for larger models or when sharing occurs among three realms, demonstrating *CAEC*’s scalability and efficiency in multi-realm deployments. These savings represent a conservative *lower bound*. In practical deployments, the benefits of *CAEC* could be significantly higher because: (1) much larger LLMs may be deployed within realms; and (2) resource sharing can extend beyond LLMs to include read-only user-space binaries and shared packages. All experiments used the same kernel (43 MB) and filesystem (233 MB).

## 7.5. Runtime Cost Benchmark

To evaluate the runtime cost of *CAEC*, we conducted an experiment in which we measure inference latency under two configurations. In the baseline setting, the model (GPT-2) is stored in the realm’s private memory. In the second setting, the model is shared with another realm and stored in the CSM. We repeat each experiment using 10 representative queries in non-conversational mode.

**Results.** The average inference time is 13.4 seconds in both scenarios, showing that executing a model from CSM via the CSM driver achieves native performance and that the same physical pages can safely be shared between two realms without incurring any inference-time overhead, promising for collaborative or resource-constrained deployments.

## 8. Discussion

In this section, we discuss three topics: why *CAEC* remains compatible with (and orthogonal to) future extensions of Arm CCA (Section 8.1); how CSM can be enabled in other confidential computing architectures (Section 8.2); and potential future research directions for *CAEC* (Section 8.3).

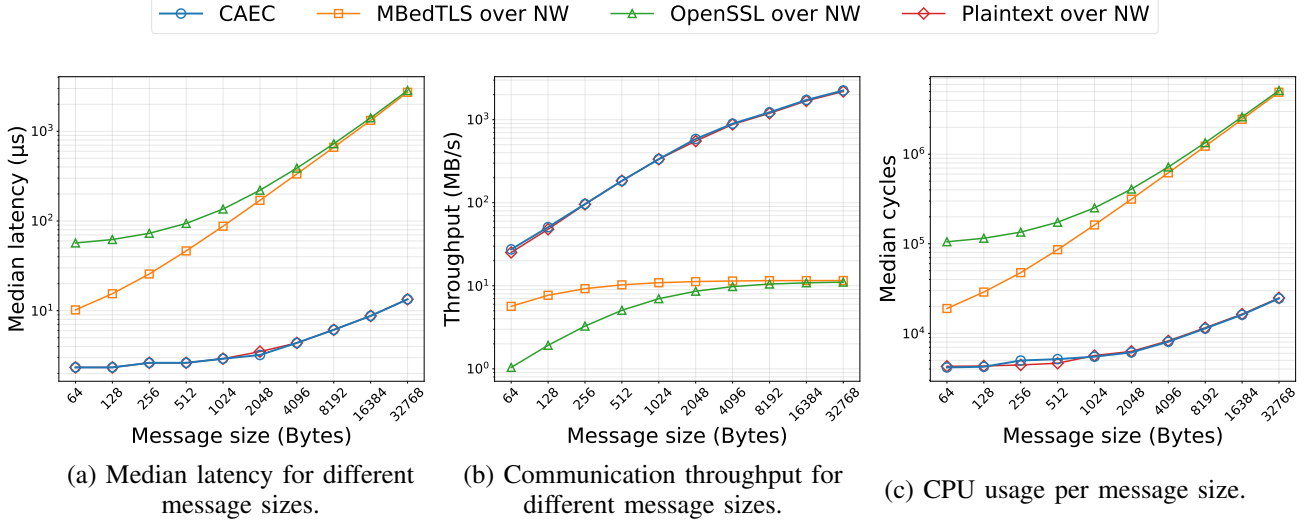


Figure 5: Communication cost between two realms across four modes: plaintext over NW shared memory, encrypted (OpenSSL/MbedTLS) with confidentiality and integrity, and CAEC.

TABLE 4: Memory footprint comparison of inference services deployed in multiple realms. In the baseline, each realm hosts its own LLM; with CAEC, multiple realms share a single model.

Model	Model Size (MB)	2 Realms		3 Realms	
		Memory Footprint (MB) Baseline / CAEC Total (P-realm, C-realm)	Reduction (%)	Memory Footprint (MB) Baseline / CAEC Total (P-realm, 2*C-realm)	Reduction (%)
GPT-2 [50]	177	960 (480, 480) / 800 (490, 310)	16.6%	1440 (480, 960) / 1110 (490, 620)	22.9%
GPT-2 Medium [51]	437	2000 (1000, 1000) / 1580 (1010, 570)	21.0%	3000 (1000, 2000) / 2150 (1010, 1140)	28.3%

### 8.1. Arm CCA Future Extensions

Although CAEC is designed and implemented on top of Arm CCA version 1.0, Arm has recently announced its planned enhancements for CCA version 1.1 [52]. In this section, we describe why CAEC remains both relevant and fully compatible with the upcoming extensions.

**Planes.** Plane extension refers to an architectural extension that enables the decomposition of a realm into multiple EL0&EL1 execution environments called planes. Planes are managed by a privileged software component known as the paravisor, which runs in plane 0. The paravisor is responsible for restricting memory access, emulating interrupts, and managing context switching among the other planes. Planes were originally proposed to augment a realm with kernel-level runtime services that cannot be provided by the untrusted hypervisor, such as access to a Trusted Platform Module (TPM) [53]. However, since all planes work in the same realm’s address space, they can also be configured to share confidential memory. By enabling multi-tenant memory sharing within the realm world, planes can conceptually achieve functionality similar to that of CAEC. Nevertheless, CAEC offers key differences compared to the plane extension, providing a more flexible and decentralized approach. First, the number of inter-CVMs memory sharing CAEC can be dynamically extended at runtime, whereas the number of planes within a realm must be statically defined at boot time. Second, CAEC requires no trusted central entity; realms can independently decide where and when to share memory with each other. In contrast, in the plane extension, all planes must trust plane 0 to enforce isolation

boundaries and manage sharing, introducing a centralized trust dependency that is unsuitable for scenarios involving multi-party collaboration.

**Memory Encryption Context.** The Memory Encryption Context (MEC) extends Arm CCA by introducing hardware-level encryption of memory, providing an additional layer of defense-in-depth [52], [54]. Memory pages tagged with the same encryption context are encrypted using the same encryption key. As proposed by Arm [52], [54], each realm’s memory can be protected under a single encryption context. Nevertheless, MEC also allows multiple encryption contexts to coexist within a single realm’s address space [54]. In other words, a realm vCPU can be configured to access multiple encryption contexts simultaneously. Consequently, a new encryption context can be assigned to each CSM region, while the memories of the P-realm and C-realm(s) remain encrypted under their own respective contexts. Therefore, CAEC is fully compatible with MEC hardware extension and MEC’s trust model, and it can be extended to leverage the MEC’s capabilities.

**Device Assignment.** This extension enhances CCA architecture to enable secure assignment of physical devices to realms, a concept previously explored in the literature for GPUs and other generic devices [42], [44], [55]. Each realm can independently choose whether to allow an off-processor resource, such as an accelerator, to access a region of its address space [52], [56]. Device assignment extends RSI and RMI and introduces an additional type of metadata granules for realms. We were unable to find any interference between the CAEC extension of the RMM



and the device assignment extension to the RMM.

## 8.2. Enabling CSM in Other Architectures

AMD SEV-SNP [3] and Intel TDX [4] are the two major CVM technologies currently deployed by cloud providers, offering security properties comparable to those of realms in Arm CCA. However, the CISC nature of these architectures makes implementing and evaluating *CAEC*'s functionality significantly more complex—or even infeasible. This section outlines the limitations of these platforms and discusses potential modifications required to enable CSM-like functionality.

**AMD SEV-SNP.** Similar to the GPT in Arm CCA, AMD SEV-SNP introduces a system-wide data structure called the Reverse Map Table (RMP), tracking the ownership and attributes of physical pages. Unlike the GPT, which only associates pages' entries with a world state, the RMP enforces stricter control by associating pages' entries with CVM's identifier, known as Address Space Identifier (ASID) [3], [57]. AMD's hardware checks the RMP entries at the end of each memory access, granting access to a page only if the ASID of the currently executing guest matches the ASID stored in the corresponding RMP entry. As a result, two CVMs cannot simultaneously access the same confidential page while the hardware-level ASID check is enforced and the RMP only accepts one ASID for each page, rendering CSM impossible to implement on SEV-SNP without hardware changes.

**Intel TDX.** To the best of our knowledge, Intel TDX [4] imposes no inherent hardware limitation that would prevent the implementation of CSM. However, supporting CSM would require modifications to the TDX Module, which has a comparable rule with the RMM in Arm CCA. TDX maintains a system-wide Physical Address Metadata Table (PAMT) that tracks the ownership and state of every physical page. Whenever a new mapping is established in a CVM's address space, the TDX Module checks the PAMT to ensure that the corresponding physical pages are not already mapped as protected memory of another CVM. This restriction could, in principle, be relaxed to permit shared mappings between CVMs that mutually agree to share memory. The TDX Module also employs Multi-Key Total Memory Encryption (MKTME) [58] to encrypt CVM's memory. MKTME supports memory encryption at page-level granularity, and thus does not impose any inherent restriction on maintaining multiple encrypted regions within a single CVM. For each CVM, the TDX Module requests a unique encryption key—identified by a Host Key Identifier (HKID)—and embeds it within every page table entry of that CVM, ensuring that all protected memory is encrypted under a single key [4], [59]. A CSM-aware TDX implementation would therefore require an additional encryption key dedicated to each new CSM region, with its corresponding page table entries tagged accordingly to preserve compatibility with the existing trust model. We acknowledge that above discussions offer a starting point for adapting CSM in TDX, however, a comprehensive design and concrete implementation merits a separate paper.

Both AMD's SEV-SNP and Intel's TDX implement intra-CVM isolation mechanisms, *e.g.*, the virtual machine

privilege levels in AMD. These isolation mechanisms are similar in nature to planes in CCA, but orthogonal to *CAEC* given that they do not allow sharing across different CVMs and follow a strictly hierarchical privilege model that can be limiting [60].

## 8.3. Future Direction

**Formal Verification.** Arm CCA's use of formal methods for system design and verification distinguishes it from other confidential computing architectures [61], [62]. Although our current work does not provide formal guarantees, an important direction for future research is to apply formal verification techniques to the CSM design within CCA. Building on recent efforts to formally verify shared-memory mechanisms between enclaves [18], extending the existing formal models of Arm CCA to encompass CSM functionality would represent a significant step toward provable security and functional correctness.

**Local Attestation.** As part of establishing the CSM, realms refer to one another solely through attestation-integrated identifiers. In *CAEC*, each realm owner is responsible not only for remotely attesting its own realm but also for attesting peer realms with which it intends to associate within the CSM. While remote attestation between a realm owner and its own realm is unavoidable, it is worth exploring whether *CAEC* could be redesigned to support local attestation between realms. Such a design—especially in scenarios where no secure channel exists between realm owners and peer realms—could significantly improve system performance while still upholding the security properties expected of attested local interactions.

**CVM Signaling over CSM.** *CAEC* enables direct memory sharing between CVMs, but the design and integration of signaling mechanisms over this shared substrate remain underexplored. Developing hypervisor-independent signaling, synchronization, and coordination primitives represents a promising direction for robust inter-CVM interaction. In particular, integrating CSM-backed signaling with established abstractions such as *virtio* could enhance both the portability and extensibility of inter-CVM communication. Event-driven synchronization is also feasible through inter-CVM interrupts, which we found can be implemented using standard KVM interfaces—requiring no changes to the hypervisor or the RMM—much like the mechanism employed by *ivshmem* [49]. However, ensuring strong security, isolation, and provenance guarantees for CSM-based signaling remains an open problem and deserves further investigation.

## 9. Related Works

**CVM Systems.** Cloud providers already offer CVM instances for a wide range of applications [7], [9], with some vendors introducing specialized designs tailored for ML workloads [5], [6]. Recently, new edge-based confidential computing systems have emerged. For instance, Samsung's Islet [63] adopts Arm CCA with a Rust-based RMM, while Aster [32] introduces sandboxed realm abstractions to secure Android applications. Similarly,

Android’s Virtualization Framework (AVF) [31] enables VMs that are isolated from the Android kernel layer while providing protected services to Android applications.

**Communication and Memory Sharing.** A persistent challenge in current CVM architectures is the performance overhead caused by routing I/O and data exchange through the untrusted hypervisor. It has been shown that hypervisor-mediated services impose significantly higher costs in CVMs than in traditional VMs [12]. Directly sharing memory between two isolated environments can substantially improve the performance of data exchange. This concept has been explored in previous work for conventional VMs [64], [65] and enclaves [17]–[19]. Specifically, Plug-in Enclave [17] enables read-only shared enclaves for serverless applications, Cerberus [18] focuses on the formal verification of memory sharing, and Elasticlave [19] explores sharing models and optimization techniques. Sartakov *et al.* [66] further extend this idea by leveraging CHERI capabilities to allow multiple VM-like compartments to share a single physical address space.

**Arm CCA.** As Arm CCA gains adoption, research in this area remains limited but is steadily growing. Li *et al.* [24] proposed a formal verification methodology for the RMM. Beyond verification, early systems research has explored various extensions to CCA. References [42], [44], [67], [68] represent a number of efforts in this space, introducing new features and capabilities to CCA by modifying its trusted components (the RMM and Monitor). SHELTER [67] leverages the GPC mechanism to support user-space enclaves in the NW. ACAI [44] and CAGE [42] address the integration of accelerators into CCA-based systems. Specifically, ACAI enables PCIe accelerators, while CAGE supports the use of integrated GPUs within realms. Portal [68] focuses on secure, high-performance device I/O by enabling direct peripheral access from within a realm on mobile SoCs. Other recent work showcases CCA potential for the next-generation of on device ML. GuaranteTEE [26] is a framework for attestable, privacy-preserving machine learning at the edge. It combines remote attestation and data protection to support collaborative ML inference across devices. Abdollahi *et al.* [14] evaluate the effectiveness of CCA in protecting ML model during inference, validating its applicability for emerging AI workloads.

## 10. Conclusion

In this work, we presented *CAEC*, the first system that enables CSM, a hypervisor-protected (confidential) memory which can be shared between multiple CVMs. *CAEC* extends Arm CCA firmware with a principled ownership model, explicit access-control rules, and attestation extensions that ensure CSM remains inaccessible to the hypervisor and all unauthorized CVMs, while preserving CCA’s security guarantees for non-CSM memory. *CAEC* delivers substantial benefits for communication and data sharing between CVMs. It achieves up to 209× reduction in CPU cycles compared to encryption-based mechanisms over hypervisor-accessible shared memory, and enables sharing of large data objects such as LLMs with 16.6%–28.3% reduction in overall system memory footprint. *CAEC* marks a step toward future compartmentalized and collaborative

multi-CVM systems, where each party can protect its proprietary data while efficiently collaborating with, and providing services to, other parties.

## Acknowledgment

We thank Jon Crowcroft for his invaluable suggestions that helped improve the paper. The research in this paper was supported by the UKRI Open Plus Fellowship (EP/W005271/1 Securing the Next Billion Consumer Devices on the Edge) and an Amazon Research Award “Auditable Model Privacy using TEEs”.

The authors employed ChatGPT models (GPT-5.1, GPT-5, and GPT-4o) as auxiliary tools for editorial support, exploration of related research, and code debugging. All generated material was examined and verified by the authors. The authors take full responsibility for the accuracy, integrity, and originality of the paper and the released code.

## References

- [1] “Intel Software Guard Extensions,” 2025. [Online]. Available: <https://www.intel.com/content/www/us/en/developer/tools/software-guard-extensions/overview.html>
- [2] A. Limited, “Learn the architecture - TrustZone for AArch64,” 2025, accessed Feb 2025. [Online]. Available: <https://developer.arm.com/documentation/102418/latest/>
- [3] “AMD SEV-SNP: Strengthening VM Isolation with Integrity Protection and More,” [Online]. Available: <https://www.amd.com/content/dam/amd/en/documents/epyc-business-docs/white-papers/SEV-SNP-strengthening-vm-isolation-with-integrity-protection-and-more.pdf>
- [4] “Intel® Trust Domain Extensions (Intel TDX),” 2025. [Online]. Available: <https://cdrdv2.intel.com/v1/dl/getContent/690419>
- [5] E. S. GmbH, “The always encrypted ai service,” Mar. 2025, march 7, 2025. [Online]. Available: <https://www.privatemode.ai/>
- [6] C. Renzo, L. d’Aliberti, J. Miles, and J. Kovba, “Large language model inference over confidential data using aws nitro enclaves,” 2024. [Online]. Available: <https://aws.amazon.com/blogs/machine-learning/large-language-model-inference-over-confidential-data-using-aws-nitro-enclaves/>
- [7] “Confidential space security overview,” [Online]. Available: <https://cloud.google.com/docs/security/confidential-space#:~:text=,resource%20is%20protected%20by%20an>
- [8] Apple Security Engineering and Architecture (SEAR), User Privacy, Core Operating Systems (Core OS), Services Engineering (ASE), and Machine Learning and AI (AIML), “Private cloud compute: A new frontier for ai privacy in the cloud,” 2024, march 9, 2025. [Online]. Available: <https://security.apple.com/blog/private-cloud-compute/>
- [9] “Confidential vms on azure,” 2023. [Online]. Available: <https://techcommunity.microsoft.com/blog/windowsplatform/confidential-vms-on-azure/3836282>
- [10] A. Limited, “Arm Confidential Compute Architecture,” 2025, accessed Feb 2025. [Online]. Available: <https://www.arm.com/architecture/security-features/arm-confidential-compute-architecture>
- [11] T. L. Foundation, “Arm Confidential Compute Architecture open-source enablement,” 2025, accessed Feb 2025. [Online]. Available: <https://confidentialcomputing.io/webinars/arm-confidential-compute-architecture-open-source-enablement/>
- [12] M. Misono, D. Stavrakakis, N. Santos, and P. Bhatotia, “Confidential VMs Explained: An Empirical Analysis of AMD SEV-SNP and Intel TDX,” *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 8, no. 3, pp. 1–42, 2024.

- [13] D. Li, Z. Mi, C. Ji, Y. Tan, B. Zang, H. Guan, and H. Chen, "Bifrost: Analysis and optimization of network {I/O} tax in confidential virtual machines," in *2023 USENIX Annual Technical Conference (USENIX ATC 23)*, 2023, pp. 1–15.
- [14] S. Abdollahi, M. Maheri, S. Siby, M. Kogias, and H. Haddadi, "An Early Experience with Confidential Computing Architecture for On-Device Model Protection," *arXiv preprint arXiv:2504.08508*, 2025.
- [15] H. Lefeuvre, D. Chisnall, M. Kogias, and P. Olivier, "Towards (really) safe and fast confidential I/O," in *Proceedings of the 19th Workshop on Hot Topics in Operating Systems*, 2023, pp. 214–222.
- [16] "Advancing security for large language models with nvidia gpus and edgeless systems," 2024. [Online]. Available: <https://developer.nvidia.com/blog/advancing-security-for-large-language-models-with-nvidia-gpus-and-edgeless-systems/>
- [17] M. Li, Y. Xia, and H. Chen, "Confidential serverless made efficient with plug-in enclaves," in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2021, pp. 306–318.
- [18] D. Lee, K. Cheang, A. Thomas, C. Lu, P. Gaddamadugu, A. Vahldiek-Oberwagner, M. Vij, D. Song, S. A. Seshia, and K. Asanovic, "Cerberus: A formal approach to secure and efficient enclave memory sharing," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022, pp. 1871–1885.
- [19] J. Z. Yu, S. Shinde, T. E. Carlson, and P. Saxena, "Elasticlave: An efficient memory model for enclaves," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 4111–4128.
- [20] A. Limited, "Realm Management Monitor Specification," 2025, accessed Feb 2025. [Online]. Available: <https://developer.arm.com/documentation/den0137/1-0cac5/?lang=en>
- [21] M. Sardar, T. Fossati, and S. Frost, "SoK: Attestation in confidential computing," *ResearchGate pre-print*, 2023.
- [22] Y. Wu, F. Roesner, T. Kohno, N. Zhang, and U. Iqbal, "Isolategpt: An execution isolation architecture for llm-based agentic systems," *arXiv preprint arXiv:2403.04960*, 2024.
- [23] D. Chen, A. Dethise, I. E. Akkus, I. Rimac, K. Satzke, A. Koskela, M. Canini, W. Wang, and R. Chen, "Protecting Confidentiality, Privacy and Integrity in Collaborative Learning," *arXiv preprint arXiv:2412.08534*, 2024.
- [24] X. Li, X. Li, C. Dall, R. Gu, J. Nieh, Y. Sait, and G. Stockwell, "Design and verification of the Arm confidential compute architecture," in *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*, 2022, pp. 465–484.
- [25] A. Limited, "Arm Confidential Compute Architecture Software Architecture Guide," 2025, accessed Feb 2025. [Online]. Available: <https://developer.arm.com/documentation/den0127/0200/?lang=en>
- [26] S. Siby, S. Abdollahi, M. Maheri, M. Kogias, and H. Haddadi, "GuaranTEE: Towards Attestable and Private ML with CCA," in *Proceedings of the 4th Workshop on Machine Learning and Systems*, 2024, pp. 1–9.
- [27] Z. Zhang, C. Gong, Y. Cai, Y. Yuan, B. Liu, D. Li, Y. Guo, and X. Chen, "No Privacy Left Outside: On the (In-) Security of TEE-Shielded DNN Partition for On-Device ML," in *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 2024, pp. 52–52.
- [28] M. Moon, M. Kim, J. Jung, and D. Song, "ASGARD: Protecting On-Device Deep Neural Networks with Virtualization-Based Trusted Execution Environments," in *Proceedings 2025 Network and Distributed System Security Symposium*, 2025.
- [29] "Virtual Machine as a core Android Primitive." [Online]. Available: <https://android-developers.googleblog.com/2023/12/virtual-machines-as-core-android-primitive.html>
- [30] A. Limited, "Arm CCA Security Model 1.0," 2025, accessed Feb 2025. [Online]. Available: <https://developer.arm.com/documentation/DEN0096/latest>
- [31] "AVF architecture," 2025, accessed July 2025. [Online]. Available: <https://source.android.com/docs/core/virtualization/architecture#memory-ownership>
- [32] M. Kuhne, S. Sridhara, A. Bertschi, N. Dutly, S. Capkun, and S. Shinde, "Aster: Fixing the Android TEE ecosystem with Arm CCA," *arXiv preprint arXiv:2407.16694*, 2024.
- [33] "RMM Locking Guidelines," 2025, accessed April 2025. [Online]. Available: <https://tf-rmm.readthedocs.io/en/latest/design/locking.html>
- [34] TrustedFirmware, "TF-RMM," 2025, accessed Feb 2025. [Online]. Available: <https://www.trustedfirmware.org/projects/tf-rmm>
- [35] "kvmtool-cca," 2025, accessed Feb 2025. [Online]. Available: [https://gitlab.arm.com/linux-arm/kvmtool-cca/-/tree/cca/v3?ref\\_type=heads](https://gitlab.arm.com/linux-arm/kvmtool-cca/-/tree/cca/v3?ref_type=heads)
- [36] A. Limited, "linux-cca," 2025, accessed Feb 2025. [Online]. Available: <https://gitlab.arm.com/linux-arm/linux-cca/-/commit/fad35572db>
- [37] "TF-A," 2025, accessed Feb 2025. [Online]. Available: <https://www.trustedfirmware.org/projects/tf-a>
- [38] "Kernel-based Virtual Machine," 2025. [Online]. Available: [https://en.wikipedia.org/wiki/Kernel-based\\_Virtual\\_Machine](https://en.wikipedia.org/wiki/Kernel-based_Virtual_Machine)
- [39] C. Dall and J. Nieh, "KVM/ARM: the design and implementation of the Linux ARM hypervisor," *ACM Sigplan Notices*, vol. 49, no. 4, pp. 333–348, 2014.
- [40] Linaro, "qemu," 2025, accessed Feb 2025. [Online]. Available: <https://git.codetinaro.org/linaro/dcap/qemu>
- [41] A. Limited, "Fast Models Fixed Virtual Platforms (FVP) Reference Guide," 2025, accessed Feb 2025. [Online]. Available: <https://developer.arm.com/Tools%20and%20Software/Fixed%20Virtual%20Platforms>
- [42] C. Wang, F. Zhang, Y. Deng, K. Leach, J. Cao, Z. Ning, S. Yan, and Z. He, "CAGE: Complementing Arm CCA with GPU Extensions," in *Network and Distributed System Security (NDSS) Symposium*, 2024.
- [43] Y. Zhang, Y. Hu, Z. Ning, F. Zhang, X. Luo, H. Huang, S. Yan, and Z. He, "SHELTER: Extending Arm CCA with Isolation in User Space," in *32nd USENIX Security Symposium (USENIX Security'23)*, 2023.
- [44] S. Sridhara, A. Bertschi, B. Schlüter, M. Kuhne, F. Aliberti, and S. Shinde, "ACAI: Extending Arm Confidential Computing Architecture Protection from CPUs to Accelerators," in *33rd USENIX Security Symposium (USENIX Security'24)*, 2024.
- [45] A. Limited, "Fast Models Reference Guide," 2025, accessed Feb 2025. [Online]. Available: <https://developer.arm.com/Tools%20and%20Software/Fixed%20Virtual%20Platforms>
- [46] A. Bertschi and S. Shinde, "OpenCCA: An Open Framework to Enable Arm CCA Research," *arXiv preprint arXiv:2506.05129*, 2025.
- [47] "ROCK 5B." [Online]. Available: <https://radxa.com/products/rock5b/>
- [48] G. Gerganov, "llama.cpp," 2023, accessed Feb 2025. [Online]. Available: <https://github.com/ggerganov/llama.cpp>
- [49] Q. Project, "Inter-VM Shared Memory device," 2023. [Online]. Available: <https://www.qemu.org/docs/master/system/devices/ivshmem.html>
- [50] OpenAI Community, "openai-community/GPT2," accessed Feb 2025. [Online]. Available: <https://huggingface.co/openai-community/gpt2>
- [51] —, "openai-community/gpt2-medium," accessed Feb 2025. [Online]. Available: <https://huggingface.co/openai-community/gpt2-medium>
- [52] "Mad24-410 arm confidential compute architecture open-source enablement update," 2024. [Online]. Available: <https://resources.linaro.org/en/resource/rEjhEezEvnNMC3LALzUTrr>
- [53] "Evolution of the arm confidential compute architecture by g. stockwell, n. sample & p. howard — oc3." [Online]. Available: <https://www.youtube.com/watch?v=1AsvIt7bSLY&t=2086s>
- [54] A. Limited, "Introducing Arm Confidential Compute Architecture," 2025, accessed Feb 2025. [Online]. Available: <https://developer.arm.com/documentation/den0125/0300/Overview>



- [55] A. Bertschi, S. Sridhara, F. Groschupp, M. Kuhne, B. Schlüter, C. Thorens, N. Dutly, S. Capkun, and S. Shinde, “Devlore: Extending Arm CCA to Integrated Devices A Journey Beyond Memory to Interrupt Isolation,” *arXiv preprint arXiv:2408.05835*, 2024.
- [56] M. Weidmann, “Arm A-Profile Architecture Developments 2022,” 2022. [Online]. Available: <https://community.arm.com/arm-community-blogs/b/architectures-and-processors-blog/posts/arm-a-profile-architecture-20221>
- [57] “SEV Secure Nested Paging Firmware ABI Specification,” 2025. [Online]. Available: <https://docs.amd.com/v/u/en-US/56860>
- [58] “Intel® Architecture Memory Encryption Technologies Specification,” 2025. [Online]. Available: <https://www.intel.com/content/www/us/en/content-details/679154/intel-architecture-memory-encryption-technologies-specification.html>
- [59] P.-C. Cheng, W. Ozga, E. Valdez, S. Ahmed, Z. Gu, H. Jamjoom, H. Franke, and J. Bottomley, “Intel tdx demystified: A top-down approach,” *ACM Computing Surveys*, vol. 56, no. 9, pp. 1–33, 2024.
- [60] C. Castes, A. Ghosn, N. S. Kalani, Y. Qian, M. Kogias, M. Payer, and E. Bugnion, “Creating trust by abolishing hierarchies,” in *Proceedings of the 19th Workshop on Hot Topics in Operating Systems*, 2023, pp. 231–238.
- [61] X. Li, X. Li, C. Dall, R. Gu, J. Nieh, Y. Sait, G. Stockwell, M. Knight, and C. Garcia-Tobin, “Enabling realms with the arm confidential compute architecture,” 2023.
- [62] A. C. Fox, G. Stockwell, S. Xiong, H. Becker, D. P. Mulligan, G. Petri, and N. Chong, “A Verification Methodology for the Arm® Confidential Computing Architecture: From a Secure Specification to Safe Implementations,” *Proceedings of the ACM on Programming Languages*, vol. 7, no. OOPSLA1, pp. 376–405, 2023.
- [63] “Islet,” 2025, accessed Feb 2025. [Online]. Available: <https://github.com/islet-project/islet>
- [64] S. Sreenivasamurthy and E. Miller, “SIVSHM: Secure inter-vm shared memory,” *arXiv preprint arXiv:1909.10377*, 2019.
- [65] Y. Ren, L. Liu, Q. Zhang, Q. Wu, J. Guan, J. Kong, H. Dai, and L. Shao, “Shared-memory optimizations for inter-virtual-machine communication,” *ACM Computing Surveys (CSUR)*, vol. 48, no. 4, pp. 1–42, 2016.
- [66] V. A. Sartakov, L. Vilanova, D. Eysers, T. Shinagawa, and P. Pietzuch, “CAP-VMs: Capability-Based Isolation and Sharing in the Cloud,” in *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*, 2022, pp. 597–612.
- [67] T. Shen, J. Qi, J. Jiang, X. Wang, S. Wen, X. Chen, S. Zhao, S. Wang, L. Chen, X. Luo *et al.*, “SOTER: Guarding Black-box Inference for General Neural Networks at the Edge,” in *2022 USENIX Annual Technical Conference (USENIX ATC 22)*, 2022, pp. 723–738.
- [68] F. Sang, J. Lee, X. Zhang, and T. Kim, “PORTAL: Fast and Secure Device Access with Arm CCA for Modern Arm Mobile System-on-Chips (SoCs),” in *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2025, pp. 4099–4116.