doi :10.3969/j.issn.1671-1122.2026.05.010

基于可信执行环境的联邦学习平台

李子豪，张锋巍
（南方科技大学计算机科学与工程系，深圳 518055）

摘　要：为评估不同的隐私增强技术在联邦学习中的安全－效率－精度权衡，文章面向典型视觉任务构建了基于可信执行环境的联邦学习平台。该平台以Intel可信域扩展（TDX）和软件防护扩展（SGX）为核心架构，并引入同态加密（HE）与安全多方计算（MPC）作为性能对比基准。在CIFAR-10数据集与ResNet-18模型的高维视觉任务场景下，文章利用该平台进行了对比实验。实验结果表明，在保持基线精度的前提下，基于TDX的方案在提供虚拟机级硬件保护的同时，仅引入约1.3%的端到端时延，综合表现优于SGX、HE与MPC。尽管HE提供了可形式化验证的安全性，但将单轮训练时延与通信开销分别提升至基线的约9倍与21倍，系统负载增加显著；MPC则在时间与通信开销间存在局限。文章明确了各类技术方案的适用边界，对于高维模型的安全聚合场景，TDX是平衡安全需求与性能开销的一个有利选项。

关键词：联邦学习；隐私保护；机密计算；可信执行环境

中图分类号：TP309　**文献标志码：**A　**文章编号：**1671-1122（2026）05-0788-21

中文引用格式：李子豪，张锋巍 . 基于可信执行环境的联邦学习平台 [J]. 信息网络安全，2026，26（5）：788-808.

英文引用格式：LI Zihao, ZHANG Fengwei. TEE-Based Federated Learning Platform[J]. Netinfo Security, 2026, 26(5): 788-808.

TEE-Based Federated Learning Platform

LI Zihao, ZHANG Fengwei
(*Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen* 518055, *China*)

Abstract: This paper presented a confidential federated learning platform (CFLP) based on trusted execution environments (TEEs) for typical vision tasks, aiming to evaluate the security-efficiency-accuracy trade-off of different privacy-enhancing technologies in federated learning. The platform utilized intel trust domain extensions (TDX) and software guard extensions (SGX) as its core architecture, while incorporating homomorphic encryption (HE) and secure multi-party computation (MPC) as performance comparison benchmarks. Systematic comparative experiments were conducted using this platform in

收稿日期：2025-12-25
基金项目：国家自然科学基金 [62372218, U24A6009]
作者简介：李子豪（2003—），男，山东，硕士研究生，CCF会员，主要研究方向为系统安全；张锋巍（1986—），男，湖南，研究员，博士，CCF会员，主要研究方向为可信执行环境、GPU机密计算。
通信作者：张锋巍 zhangfw@sustech.edu.cn

high-dimensional vision task scenarios involving the CIFAR-10 dataset and the ResNet-18 model. The results indicate that, while maintaining baseline accuracy, the TDX-based TEE scheme provided virtual-machine-level hardware protection with only an approximately 1.3% increase in end-to-end latency, outperforming SGX, HE, and MPC in comprehensive performance. Although HE offers formally verifiable security, it increased the single-round training latency and communication overhead to approximately 9 times and 21 times that of the baseline, respectively, resulting in significant computational overhead. MPC exhibited limitations in the trade-off between time and communication costs. This study clarifies the applicable boundaries of various technical solutions, demonstrating that for secure aggregation scenarios involving high-dimensional models, TDX is a favorable option for balancing security requirements and performance overhead.

Key words: federated learning; privacy protection; confidential computing; trusted execution environment

0 引言

在数据驱动决策日益重要的今天，数据的隐私保护需求与合规要求的同步增长[1]加剧了“数据孤岛”效应。这一矛盾对依赖集中数据训练的传统学习范式构成了根本性挑战。联邦学习（Federated Learning，FL）通过其“数据不动、模型动”的核心架构[2]，有效降低了原始数据在传输和汇聚过程中的泄露风险，为解决数据隐私与协同学习之间的矛盾提供了可行路径。然而，联邦学习将潜在的隐私攻击面从原始数据本身转移到了模型训练过程中交换的梯度信息与模型参数上[3]。研究表明，即使是遵守协议的半诚实服务器，也能通过分析模型更新（尤其是梯度）反演出用户的敏感信息[4]。

为了填补联邦学习“使用态数据”（即梯度与参数）的隐私保护缺口，隐私增强技术（Privacy-Enhancing Technologies，PETs）在可信执行环境（Trusted Execution Environment，TEE）或以同态加密（Homomorphic Encryption，HE）与安全多方计算（Secure Multi-Party Computation，MPC）为代表的密码学层面构建安全屏障，成为提升联邦学习隐私保护水平的关键技术方向。尽管多种集成方案被提出，但当前面临的一个关键问题是不同隐私增强技术的引入伴随着显著的系统性能开销与潜在的模型精度损失[5,6]。在典型的视觉识别等任务场景下，何种隐私增强技术能在安全保证、模型性能与计算/通信开销之间达到最佳的平衡，仍是一个开放的、待实证研究的问题[7]。

本文旨在针对“在典型视觉任务环境下，应用主流隐私增强技术是否能以可接受的模型精度损失与性能代价，有效实现其在联邦学习中提供的隐私保护”这一问题，提供实验依据与分析。为此，本文构建了一个面向联邦学习的机密联邦学习平台（Confidential Federated Learning Platform，CFLP），通过对比实验，本文系统性地量化了不同隐私增强技术对联邦学习过程产生的具体影响，初步揭示了其内在的安全—效率—精度权衡关系。希望研究成果能为相关技术方案的选择提供一些参考，并为未来的优化方向提供一些思路。

1 背景知识

本章主要介绍联邦学习平台设计过程中涉及的技术背景。

1.1 联邦学习

联邦学习是一种分布式机器学习范式[2]，旨在解决数据隐私保护和“数据孤岛”问题。其核心思想是让多个参与方在不共享原始数据的情况下，通过共享模型参数或梯度信息来协作训练一个全局模型。联邦学习通过安全协议，确保数据在本地处理，只有模型更新信息在不同参与方之间传输，从而在保护数据隐私的同时实现模型的高效训练。

1.1.1 联邦学习步骤

联邦学习流程如图1所示，联邦学习的训练过程通常包括以下6个步骤[7]。

1）初始化模型：在服务器端初始化全局模型，并将模型分发给各个参与方。

2）本地训练：参与方使用本地数据对模型进行训练，计算梯度或更新模型参数。

3）更新模型：参与方将本地模型的更新信息（如梯度或参数差值）发送至服务器。

4）全局聚合：服务器根据各参与方的更新信息，通过聚合算法（如联邦平均算法[2]）更新全局模型。

5）分发模型：服务器将更新后的全局模型重新分发给各参与方，进入下一轮迭代。

6）终止条件：当满足预设的终止条件（如达到最大迭代次数或模型收敛）时，训练结束。

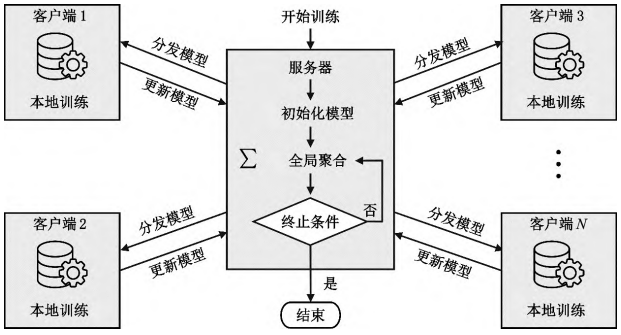


图1 联邦学习流程

1.1.2 横向与纵向联邦学习

联邦学习根据参与方之间的数据分布特点，通常分为横向联邦、纵向联邦与联邦迁移学习三大类[8]。鉴于本文聚焦于多方持有的图像数据特征空间一致但样本不同的场景，本小节重点讨论前两者的区别，并以此定位本文的技术路线。

横向联邦学习又被称为按样本划分（Sample-Partitioned）的联邦学习，适用于各参与方的数据特征空间相同，但样本ID不同的场景[8]。例如，多家不同地区的医院拥有大量患者的医疗影像数据，这些影像数据的特征是相同的（例如，都是28×28像素的CT图像），但覆盖的患者群体（样本）完全不同。横向联邦的目标是在不共享患者数据的前提下，联合这些数据训练一个性能更优的全局诊断模型。其核心技术挑战在于如何安全、高效地聚合各方上传的模型参数（如梯度或权重）。

纵向联邦学习又被称为按特征划分（Feature-Partitioned）的联邦学习，适用于各参与方样本ID大部分重叠，但数据特征空间不同的场景[9]。例如，同一地区的银行和电商平台拥有同一批用户的消费数据，银行拥有用户的信贷历史、收入等金融特征，而电商平台则拥有用户的购买记录、浏览偏好等行为特征。纵向联邦的目标是在保护用户隐私的前提下，联合双方的特征来构建一个更精准的用户信用评分模型。其技术挑战更为复杂，通常涉及加密状态下的实体对齐、中间结果的安全计算等步骤[10]。

图2展示了横向与纵向联邦学习的数据分布特征，为了更清晰地理解二者的差异，总结如表1所示。

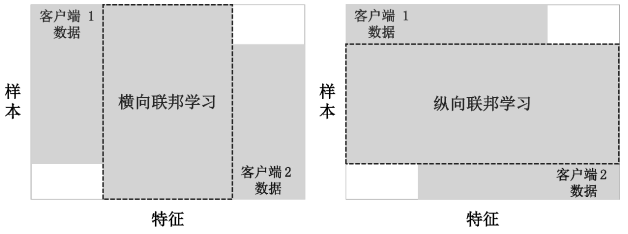


图2 横向与纵向联邦学习的数据分布特征对比

表1 联邦学习两种模式的对比结果

对比维度	横向联邦学习	纵向联邦学习
数据划分	特征相同，样本不同	样本重叠，特征不同
核心挑战	模型参数的安全聚合	加密实体对齐与联合建模
典型场景	不同医院间的医疗影像分析	银行与电商的联合信用评分
聚合内容	完整的模型参数或梯度	加密的中间计算结果

本文锚定横向联邦学习场景，旨在利用其“计算逻辑简单但数据吞吐量巨大”的特性，评估不同隐私增强技术的适用边界。与侧重复杂逻辑交互的纵向联邦学习不同，横向联邦学习的核心挑战在于海量高维模型参数的安全聚合。在此场景下，同态加密与安全多方计算因密文膨胀和多轮交互面临严峻的通信与计算瓶颈；而可信执行环境凭借硬件隔离与近乎原生的计算速度，更适合高吞吐量的聚合需求。因此，选择横向联邦学习既符合视觉任务特征，又能验证TEE相比纯密码学方案在解决高维模型聚合难题时的表现。

1.1.3 通信效率与聚合策略

通信效率和聚合策略是联邦学习中的关键问题。

高效的通信机制能够减少客户端与服务器之间的数据传输量，降低通信成本和延迟[11]。本文采用的联邦平均（Federated Averaging，FedAvg）[2]算法是联邦学习中广泛采用的聚合策略，其核心思想是通过加权平均客户端模型参数来更新全局模型，如公式（1）所示。

$$\theta^{t+1}=\sum_{i=1}^{K}\frac{n_i}{N}\theta_i^{t+1} \quad (1)$$

其中，θ_i^{t+1} 表示第 i 个客户端本地更新后的模型参数，n_i 为客户端本地数据量，$N=\sum_{i=1}^{K}n_i$ 为总数据量。合理的聚合策略可以确保全局模型的更新更加准确和稳定[12]，从而提高联邦学习的整体性能。

为提高通信效率，本平台使用gRPC（Google Remote Procedure Call）框架进行通信。gRPC是由Google开发并开源的一个高性能、跨语言的远程过程调用框架。它构建于HTTP/2协议之上，充分利用了HTTP/2提供的多路复用、头部压缩和双向流等特性，从而实现了低延迟和高吞吐量的通信。gRPC 默认采用 Protocol Buffers（Protobuf）作为其接口定义语言和数据序列化格式。Protobuf是一种语言无关、平台无关的高效结构化数据序列化机制，它通过定义.proto文件来明确服务接口，保证了强类型约束，此外其二进制序列化格式相比于基于文本的 JSON 或 XML具有更高的编解码效率和更小的数据体积。这些特性使 gRPC 成为构建云原生应用、连接多语言微服务以及实现高性能数据交换的理想选择。

如何在保证隐私安全的前提下，优化通信效率和设计有效的聚合策略，是当前联邦学习研究的重要方向[7]。

1.2 可信执行环境

1.2.1 TEE 的基本思想

TEE作为机密计算的核心基础设施，通过硬件辅助的隔离机制为敏感数据提供“计算黑箱”保护[13]。其核心思想是在主处理器内部，通过硬件机制构建一个与上层操作系统，通常被称为富执行环境（Rich Execution Environment，REE），相隔离的安全处理区域。这个隔离区旨在保护其中加载的代码和数据的机密性与完整性，确保它们在运行时（Data-in-Use）免受来自REE（包括内核、驱动，甚至物理外设）的窥探和篡改。可信执行环境的硬件隔离与安全机制示意图如图3所示。

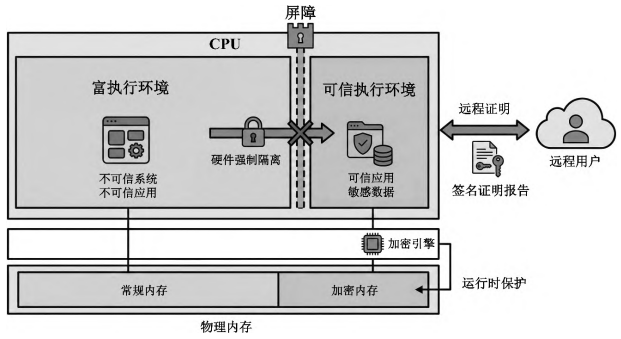


图 3 可信执行环境的硬件隔离与安全机制示意图

TEE技术的实现通常建立在以下3个共通的基本原则之上。

1）硬件强制隔离（Hardware-Enforced Isolation）。TEE的信任根基在于硬件，它利用CPU的特殊指令和访问控制逻辑，在硬件层面划分出安全与非安全两种状态或区域。任何从非安全侧发起的、对安全区域资源的未授权访问都将被硬件直接阻断。这种隔离是强制性的，其安全性不依赖于上层软件的正确性。

2）运行时保护（Runtime Protection）。TEE的首要目标是保护“使用态数据”。为实现此目标，不同TEE采用多样的硬件防护机制。例如，通过严格的内存访问权限控制来阻止非法读写[14]（如ARM TrustZone），或通过内存加密技术确保即使物理内存被访问也无法窃取明文信息[15]，如Intel的软件防护扩展（Software Guard Extensions，SGX）和可信域扩展（Trusted Domain Extensions，TDX）技术，以及AMD的安全虚拟化（Secure Encrypted Virtualization，SEV）技术，其共同点在于都为在隔离区内执行的敏感计算提供了运行时保护。

3）远程证明（Remote Attestation）。为了让外部用户能够信任一个远程的、自己无法物理控制的TEE环境，所有TEE都必须提供远程证明机制。该机制允许隔离区生成一份由硬件签名、能够证明其自身软硬件状态真实性和完整性的加密报告。用户在与TEE交互前，可通过验证这份报告来确认其运行在真实、未被

篡改的TEE环境中，从而建立起远程信任链。

1.2.2 常见 TEE 的安全机制

当前主流TEE技术呈现架构分化特征，其安全机制可归纳为三大技术路线。

1）应用级隔离（Intel SGX）。通过飞地（Enclave）实现进程内可信计算，其安全根基在于内存加密与远程认证两大机制[13,16]。然而，早期SGX版本受限于最大256MB的处理器预留内存，这为大规模模型训练等内存密集型应用带来了性能瓶颈[17]。为突破此限制，新一代SGX2引入了飞地动态内存管理（Enclave Dynamic Memory Management，EDMM）技术。该技术支持 Enclave 内存页在受硬件加密保护的飞地页面缓存（Enclave Page Cache，EPC）与外部非可信内存之间进行动态置换，从而将Enclave的可用逻辑内存扩展，优化了大规模应用的内存容量难题，但当工作集超出物理EPC时，内存分页仍会引入性能开销[18]。

2）系统级隔离（ARM TrustZone）。ARM TrustZone在单一物理处理器内构建了相互隔离的安全世界（Secure World）与非安全世界（Normal World）双执行域[19]。该架构在系统级芯片（System on Chip，SoC）层面实现了物理资源的非对称分区，即安全世界独占访问加密引擎、受保护内存区及安全外设控制器，其可信执行环境运行轻量化可信应用（Trusted Application，TA）；非安全世界则托管常规操作系统与用户应用[20]。硬件内存保护单元和总线过滤器严格拦截跨域内存/外设访问请求，而基于监控模式（Monitor Mode）或安全EL2层级的上下文切换协议由安全监视调用（Secure Monitor Call，SMC）指令触发，确保双域寄存器状态与地址空间的完全隔离[21]。

3）虚拟机级隔离（Intel TDX）。Intel TDX在安全机制上实现了多重突破[16]：（1）通过多密钥全内存加密（Multi-Key Total Memory Encryption，MKTME） 技术实现TB级内存分区加密，每个 Trust Domain 拥有独立加密密钥，可防御物理攻击下的数据泄露；（2）引入硬件级安全协处理器（TDX Module），以安全仲裁模式（Secure-Arbitration Mode，SEAM）将虚拟机监控器（Virtual Machine Monitor，VMM）权限限定为资源调度，阻断其对加密内存的访问路径；（3）采用基于密码的消息认证码（Cipher-Based Message Authentication Code，CMAC）算法实现加密虚拟机状态的完整性迁移；（4）基于硬件度量寄存器（Runtime Measurement Register，RTMR）构建零信任认证链，支持从处理器微码到应用层的逐级可信验证。

与此类似，AMD的SEV-SNP（Secure Encrypted Virtualization-Secure Nested Paging）技术也提供了强大的虚拟机级隔离[22]。（1）通过为每个虚拟机分配独立的高级加密标准（Advanced Encryption Standard，AES）加密密钥，并由板载的AMD 安全处理器（Secure Processor，SP）进行管理，实现了全内存机密性保护；（2）通过引入安全嵌套分页（Secure Nested Paging，SNP）技术和硬件管理的逆向映射表（Reverse Mapping Table，RMP），SEV-SNP在硬件层面强制实施内存完整性保护，有效防御来自恶意虚拟机监视器（Hypervisor）的内存重放、重映射等攻击，同时通过加密虚拟机CPU状态来阻止Hypervisor窥探或篡改运行时上下文，从而将其排除在信任边界之外；（3）通过在虚拟机内部运行的安全服务模块（Secure Service Module，SSM），实现了加密虚拟机状态的完整性迁移，保障了机密负载在跨主机迁移过程中的安全；（4）基于芯片唯一的版本化芯片背书密钥（Versioned Chip Endorsement Key，VCEK）构建了硬件信任根，通过AMD密钥分发服务（Key Distribution Service，KDS）进行远程验证，形成了从芯片固件版本到应用的可验证信任链。

而 Arm的机密计算架构（Confidential Computing Architecture，CCA）则通过底层的架构性革新来实现隔离[23]。（1）CCA引入机密领域管理扩展（Realm Management Extension，RME）硬件特性，通过由领域监控固件（Realm Monitor，RM）管理的颗粒化保护表（Granular Protection Table，GPT），在硬件层面将物理内存划分为根、领域、安全和普通4个世界，从根本上阻止了Hypervisor对领域（Realm）内存的访问；（2）

CCA 将 Hypervisor 的权限严格限定为资源管理者，而将安全策略执行下放给一个体积极小、可验证的领域管理监视器（Realm Management Monitor，RMM），RMM 作为 Hypervisor 和底层硬件之间的安全中介，彻底将其从可信计算基（Trusted Computing Base，TCB）中移除；（3）由于其硬件强制的访问控制模型，领域的状态机密性和完整性在设计上就得到了保障，任何来自外部的非法访问都会被硬件直接阻断；（4）CCA 定义了一套标准的认证令牌（Attestation Token）格式，该令牌包含平台和领域的测量值，由硬件安全模块签名，为远程验证方提供了可验证的证明平台和机密负载完整性的证据。表 2 对以上 TEE 技术的关键差异进行了对比。

表 2 TEE 特性对比

特性	Intel® SGX	Arm TrustZone	Intel® TDX	AMD SEV-SNP	Arm CCA
隔离粒度	进程级 (Enclave)	系统分区级 (Normal / Secure World)	虚拟机级 (Trust Domain)	虚拟机级 (Encrypted VM)	虚拟机级 (Realm)
部署模型	代码重构 / LibOS	专用可信 OS / APP	直接迁移	直接迁移	直接迁移
内存容量	SGX1 受 EPC 限制 SGX2 支持动态内存	受限于系统配置	支持整个 VM 内存	支持整个 VM 内存	支持整个 Realm 内存
可信基 (TCB)	应用代码 SGX 硬件	可信 OS / APP Arm 硬件	客户机操作系统 TDX 硬件	客户机操作系统 AMD-SP 硬件	Realm 操作系统 Arm CCA 硬件 RMM
I/O 模型	OCALL 委托不可信主机	委托给非安全世界的 OS 处理	由客户机操作系统驱动处理	通过共享内存由客户机驱动处理	通过共享内存由客户机驱动处理
性能开销	频繁的 Enclave 转换开销高	世界切换 (World Switch) 开销较高	计算开销低，完整性检查引入额外开销	计算开销低，完整性检查引入额外开销	理论开销低（待大规模硬件验证）
生态	成熟，研究多，已知攻击多	成熟，数十亿移动 /IoT 设备的基础	较新，云支持增长快	已在主流云平台广泛部署	发展初期，硬件生态正在构建

1.2.3 TEE 技术选型的关键维度分析

在机密联邦学习的部署实践中，选择合适的 TEE 技术至关重要[24]。为深入探究不同隔离模型对视觉任务下联邦学习的精度以及开销的影响，本文选取 Intel 公司提供的软件防护扩展（SGX）与信任域扩展（TDX）两种主流且具有代表性的 TEE 技术进行对比研究。

Intel SGX 与 TDX 机密计算架构对比如图 4 所示，这两种技术分别代表了进程级隔离和虚拟机级隔离两种截然不同的设计思路，其本质差异构成了本文对其对比研究的核心基础。SGX 通过 Enclave 在应用进程内创建细粒度保护区，理论上拥有极小的 TCB，但代价是应用与 Enclave 间复杂的交互接口显著扩大了潜在的攻击面[25]。这种精细化的隔离模型直接导致了其高昂的开发与移植成本，通常要求对应用进行深度代码重构，或借助 Gramine 等 LibOS 库运行，这可能引入额外开销[18]。与之形成对比的是，TDX 将隔离粒度提升至整个虚拟机，通过专用的硬件安全协处理器将不可信的 VMM 排除在 TCB 之外，从而有效收敛了攻击风险。更重要的是，TDX 支持直接迁移的部署模式，允许将包含完整操作系统和复杂软件栈的虚拟机镜像直接部署[16]，这对于联邦学习聚合器等复杂应用具有很大优势。在资源与性能维度，二者的权衡同样显著。早期 SGX 版本受限于较小的物理加密内存[17]，新一代的 SGX2 虽通过动态内存管理技术将可用逻辑内存扩展，但在处理大规模梯度聚合等内存密集型任务时，一旦工作集超出物理 EPC，频繁的内存页交换仍会引发性能瓶颈。此外，其频繁的模式切换（ECALLs/OCALLs）也会带来巨大的性能开销[26]。相比之下，TDX 通过保护整个虚拟机内存，可支持更大级别的机密计算容量，为大型 AI 应用提供了更多资源。最后，从生态成熟度看，SGX 历史更久，研究更充分，但暴露的漏洞也更多[25]；TDX 作为新一代架构，虽在快速发展，但公开的实战安全验证少于 SGX。

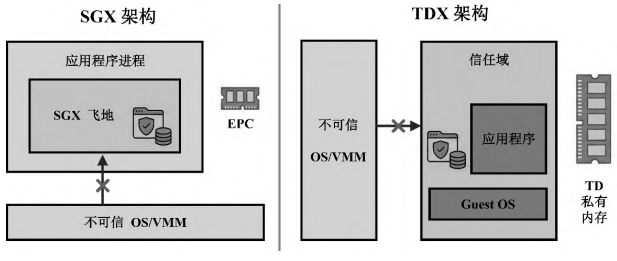


图 4 Intel SGX 与 TDX 机密计算架构对比

综上所述，SGX 和 TDX 在联邦学习场景下呈现出技术权衡关系。SGX 代表了追求最小 TCB 的精细化隔离方案，但需付出高昂的开发成本和忍受严格的资源

限制；而TDX则凭借其虚拟机级隔离、更强的迁移能力和资源可扩展性，更契合联邦学习负载的部署需求。因此，本文对这两种技术进行并列实验，旨在量化评估它们在真实联邦学习场景中的部署便捷性、性能表现及安全稳健性，为行业在不同应用需求下进行技术选型提供参考与优化思路。

1.2.4 与联邦学习的互补性

可信执行环境与联邦学习之间存在着直接的互补性，两者的结合为在不被信任的环境中进行分布式机器学习提供了一种有效的方案。联邦学习作为一种新兴的隐私保护计算框架，其核心思想在于“数据不动模型动”，通过将模型训练任务分发至各个数据持有方，仅在中心服务器聚合加密后的模型参数（如梯度），从而避免了原始敏感数据的集中化存储和传输，在架构层面降低了数据泄露的风险。

然而，联邦学习本身并不能完全解决所有安全和隐私问题。1）联邦学习无法保护计算过程中的数据，即“使用态数据”的机密性。在客户端或服务器端，即使是经过加密的梯度，一旦解密进行计算，就可能暴露在具有特权的操作系统或虚拟机监控器面前，为恶意主机或内部攻击者提供了可乘之机[7]。2）联邦学习的安全性高度依赖于聚合服务器的诚实性，一个恶意的或被攻陷的聚合器可能会篡改全局模型，或通过分析各方上传的梯度，发起模型逆向攻击（Model Inversion Attacks）、成员推断攻击（Membership Inference Attacks）等，从而间接窃取参与方的训练数据信息[4]。TEE技术的引入，为解决联邦学习的上述问题提供了可能。通过其基于硬件的隔离机制，TEE能够在计算平台上创建一个Enclave或信任域（Trust Domain）的黑箱环境。当联邦学习的聚合计算在TEE内部执行时，即使是云服务提供商或系统管理员，也无法访问或篡改正在处理的模型参数，从而为使用态数据提供了机密性保障。此外，TEE的远程证明机制允许参与方在传输模型更新之前，对聚合服务器的执行环境进行密码学验证，确保其上运行的是预期中未经篡改的聚合算法。这种可验证的计算完整性，将对服务器的信任建立在硬件和密码学证明之上，而不是假定其行为诚实。

综上所述，联邦学习在架构层面解决了分布式数据的安全流转问题，而TEE在模型聚合层面保障了核心计算环节的机密性与完整性。这种架构与聚合层面的结合，为构建更为安全可靠的联邦学习系统提供了坚实的技术基础。

2 相关工作

本章旨在全面回顾与可信联邦学习相关的研究工作，首先阐述联邦学习面临的隐私攻击，随后分类讨论主流的隐私增强技术，并在此基础上引出本文的研究动机。

2.1 联邦学习的隐私攻击

联邦学习虽然避免了原始数据集的泄露风险，但模型参数与梯度更新成为新的攻击载体。攻击层面的研究揭示了这一风险[10]。早期的模型逆向攻击已经证明，通过查询模型的输出和置信度，攻击者可以重建出具有代表性的训练数据样本[27]。随着研究的深入，针对联邦学习场景的攻击变得更加直接和高效。HITAJ[4]等人利用生成对抗网络（Generative Adversarial Network，GAN），证实了联邦学习中的恶意参与者仅通过分析共享的模型更新，就能高质量地重构出其他参与方的私有训练图像。他们的实验在MNIST和LFW人脸数据集上的重建精度分别达到了75%和83%，并且这类攻击在半诚实的威胁模型下即可奏效，这表明隐私泄露已成为联邦学习的系统性风险。ZHU[28]等人提出的梯度深度泄露（Deep Leakage from Gradients）攻击表明，即使是单次迭代中泄露的梯度信息，也足以让攻击者近乎完美地逐像素重建出原始的训练图像和标签。NASR[29]等人则对深度学习中的隐私泄露进行了全面分析，系统研究了包括被动和主动白盒攻击在内的多种推理攻击方法，进一步证实了在中心化和联邦学习环境下，模型参数交换过程中存在着严重的隐私泄露隐患。这些研究共同说明，若不加以特殊保护，

联邦学习中的梯度和模型更新将成为新的隐私泄露后门。

2.2 隐私增强技术

为应对联邦学习中的隐私挑战，学术界和工业界探索了多种隐私增强技术（Privacy-Enhancing Technology，PET）以保障模型训练的机密性与完整性，主要可分为差分隐私（Differential Privacy，DP）、密码学方法和可信执行环境三大技术路径。

2.2.1 差分隐私

差分隐私作为一种主流的隐私保护理论，通过向梯度更新中注入受控噪声来提供可量化的隐私保证，如ABADI[30]等人提出的DP-SGD（Differentially Private Stochastic Gradient Descent）算法，以及GHAZI[31]等人对标签差分隐私的探索。其核心优势在于提供了严格的、可数学证明的隐私保护上限。然而，差分隐私的有效性建立在隐私与模型效用的权衡之上，高强度的隐私保护往往以牺牲模型精度为代价，如何在二者之间找到最佳的平衡点是当前研究的难点。

2.2.2 密码学方法

以MPC和HE为代表的密码学方法提供了更强的安全承诺。MPC允许多方在不暴露各自输入的情况下协同计算，例如，MOHASSEL[32]等人设计的SecureML系统，该系统在双服务器模型下，利用安全双方计算（Two-Party Computation，2PC）技术，在不泄露各方数据的情况下高效训练线性回归、逻辑回归和神经网络等模型；而HE则支持在密文上直接进行聚合运算。尽管这些方法能从数学上严格防止信息泄露，但其高昂的计算与通信开销限制了它们在复杂联邦学习任务中的可扩展性。

2.2.3 可信执行环境

TEE可以利用硬件隔离技术创建一个安全区来保护服务器端的聚合过程。近年来，基于TEE的机密计算在AI与机器学习领域获得了广泛关注。

在联邦学习场景下，研究者不仅利用TEE保护聚合过程的机密性，还致力于解决训练的完整性问题。例如，CHEN[33]等人提出一种保证训练完整性的隐私保护联邦学习方案，通过在客户端和服务器端双向部署TEE，确保各参与方都正确执行预设的学习算法，从而有效检测和防御了旨在污染全局模型的恶意更新。

但TEE的安全性高度依赖于底层硬件实现。硬件加密引擎中的设计缺陷可能被恶意主机利用，通过CipherShadow等攻击手段重建Enclave中的机密数据[34]。这印证了对此类环境进行形式化验证以确保其安全承诺的重要性[35]。另外，TEE普遍存在性能瓶颈。这催生了模型分区（Model Partitioning）方案，即仅将模型的“隐私敏感”部分放入TEE，而将其余部分卸载到GPU等非可信硬件上执行。但研究表明，这种分区本身极不安全，卸载到外部的“非敏感”部分依然会泄露大量隐私信息，导致模型窃取（Model Stealing，MS）和成员推理攻击几乎达到白盒水平[36]。

综上所述，现有研究为联邦学习的隐私保护提供了多种技术路径。然而，这些工作大多侧重于理论分析或针对某一种特定技术的优化，缺少在统一的测试环境和评价标准下，对几种主流隐私增强技术进行的系统性、横向的性能对比。这种比较的缺乏，使得开发者在实际应用中难以针对具体需求（如性能开销、部署复杂度）选择合适的技术。

为弥补这一不足，本文构建了基于可信执行环境的机密联邦学习平台（Confidential Federated Learning Platform，CFLP），为TEE（主要是SGX和TDX）、HE、MPC等隐私增强技术提供一个统一的实验和评估环境。本文基于此平台，在典型的视觉识别任务上开展了一系列对照实验，旨在系统性地量化不同技术方案的性能和特点，为在联邦学习场景下选择和部署隐私增强技术提供直接的实证依据。

3 联邦学习平台设计与安全分析

机密联邦学习平台CFLP由联邦模拟平台与联邦实验平台两部分构成。这两部分共同构成一个从算法原型验证到真实部署评估的完整研究工作流，二者在

实现上完全独立，但在工作流上前后承接。

联邦模拟平台的核心意义在于提供一个轻量级、高效率的算法原型验证环境。它采用单进程、共享地址空间的伪分布式架构在本地执行，使研究者能脱离真实网络通信与容器化部署的复杂度，专注于算法本身的性能对比。

联邦实验平台则承接模拟阶段验证过的最优模型及其参数，其定位是评估该算法在集成隐私增强技术后，在真实分布式环境中的性能、开销。它通过Docker和gRPC构建了一个真实的多容器分布式系统，旨在量化隐私技术在实践中引入的真实系统开销。

CFLP通过这两阶段的研究（模拟平台用于算法选型，实验平台用于部署评估），为联邦学习提供了从理论验证到实践部署的全链路评估能力。

3.1 威胁模型与安全假设

本节将详细阐述系统的威胁模型、信任边界划分及核心安全假设。

3.1.1 威胁模型

本文的威胁模型主要针对联邦学习场景下的两类关键安全挑战。1）来自中心服务器的推断攻击，本文将核心敌手定义为“诚实但好奇”（Honest-but-Curious）的服务器。该敌手会忠实地执行联邦学习协议，但会利用其数据中枢的特权地位，尝试分析接收自各客户端的模型更新，以推断用户的本地私有数据。2）来自网络的窃听与篡改攻击，本文假设客户端与服务器之间的通信信道是不可信的，网络路径上的攻击者可能截获并破坏传输中模型参数的机密性与完整性。

3.1.2 信任边界与安全假设

基于上述威胁模型，本文对各组件的信任边界进行了严格划分。系统的TCB由两部分构成：1）参与方的客户端环境，本文假设客户端的硬件、操作系统及本地数据是安全可信的，所有隐私相关的预处理操作均在此完成；2）服务器端的TEE，本文信任TEE硬件本身能为在其内部执行的聚合程序提供可靠的机密性与完整性保护。与此相对，所有TCB之外的组件均被视为不可信，这包括中心服务器的宿主环境（如操作系统、虚拟机监视器）以及连接所有节点的通信网络，它们均是“诚实但好奇”敌手的潜在控制范围。

3.1.3 安全设计与完备性分析

CFLP平台的核心安全设计在于系统的安全性并非依赖于对服务器软件的信任，而是通过构建一个从可信客户端到可信执行环境的端到端加密链路来实现。具体而言，在可信执行环境策略中，模型更新首先在可信的客户端本地完成加密封装；随后，这些密文经由不可信的网络传输到服务器。在TDX模式下，密文直接传入虚拟机；在SGX模式下，则由服务器主进程转发至Enclave。在此过程中，服务器仅扮演无法窥探内容的“邮差”角色；最终，密文被安全地递送到可信的TEE中进行解密和聚合。因此，服务器主进程仅能得到聚合后的全局参数，无法据此推断用户的本地私有数据。通过密码学与硬件隔离技术的结合，CFLP平台确保了即使是特权受损的服务器，也无法危害客户端的数据隐私，从而在定义的威胁模型下，为整个联邦学习流程提供了安全保障。

3.2 联邦模拟平台

为了在集成开销较大的隐私增强技术前确立算法的性能基准，本文首先构建了一个轻量级的联邦模拟平台，其架构如图5所示。该模拟平台旨在脱离真实网络通信与硬件隔离环境的复杂性，专注于验证不同模型架构在联邦环境下的收敛行为。

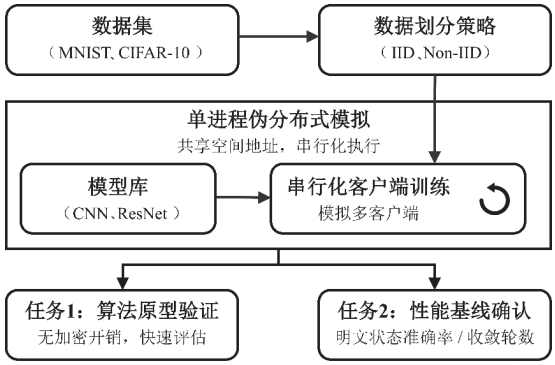


图5 联邦模拟平台架构

联邦模拟平台采用单进程、共享地址空间的伪

分布式架构，通过串行化执行模拟多客户端的并行训练，消除了网络延迟与序列化开销。在功能上，平台集成了从简单的卷积网络到深层残差网络（Residual Network，ResNet）的模型支持，并内置了独立同分布（Independent and Identically Distributed，IID）与包含标签偏斜、数量偏斜在内的非独立同分布（Non-Independent and Identically Distributed，Non-IID）数据划分策略。该模拟平台在本文中承担两项核心任务。

1）算法原型验证：在无加密开销的环境下，快速评估模型对复杂数据分布的适应能力，确定最优超参数。

2）性能基线确立：建立模型在明文状态下的准确率与收敛轮数。这为后续在联邦实验平台中剥离算法本身的影响，精准量化TDX、SGX、HE与MPC引入的系统开销提供了必要的参照系。

3.3 联邦实验平台

为了评估不同隐私增强技术在真实网络环境与系统层面的开销，本文构建了基于 Docker 容器化与 gRPC 通信的联邦实验平台。与模拟平台不同，实验平台侧重于量化分布式环境下的通信延迟、内存占用以及加密计算对系统吞吐量的影响。

3.3.1 通信架构与协议

为了高效、安全地实现聚合算法中频繁的模型参数交换，本平台采用了谷歌开源的高性能远程过程调用框架作为底层通信协议，并在协议层设计实现了以下关键特性，以提升系统的扩展性与鲁棒性。

1）基于多态载荷的统一接口设计。为实现聚合逻辑与隐私增强策略的解耦，平台在通信协议层设计了多态序列化机制。该机制将明文参数、同态加密密文、秘密共享份额以及硬件封装的混合加密包等异构数据格式，封装于统一的消息结构中。这种设计使得通信接口对具体的数据形态“不可见”，不仅降低了不同隐私增强技术集成的复杂度，也为系统未来扩展新型安全协议提供了标准化的通用接口。

2）事件驱动的状态同步机制。针对传统轮询模式造成的网络拥塞与资源空耗问题，平台采用了基于服务端推送的流式同步模型。服务器利用事件驱动机制能够根据训练状态的变更，主动向分布式客户端推送控制指令。这种交互模式相比于高频的轮询，显著消除了空闲周期的通信开销，并能有效避免死锁问题。

3）大规模密文的分块流式传输。针对同态加密等策略引起的载荷膨胀问题，平台引入了分块流式传输。该策略将高维模型参数切分为连续的数据流进行传输，配合传输层的多路复用与压缩特性，有效规避了在带宽受限或内存受限环境下单次传输体积过大导致的内存溢出风险，保障了联邦训练的稳定性。

此外，通信链路启用传输层安全（Transport Layer Security，TLS）协议，确保了模型更新在不可信网络环境中的端到端机密性与完整性。

3.3.2 实验平台设计

联邦实验平台采用星型拓扑结构，以容器化方式构建多容器集群，通过定义服务器、客户端和聚合器的镜像来明确依赖环境，确保实验的可重复性。本平台的设计优先保障实验的可复现性，而非追求生产级的高可用性。因此，系统采用了单协调服务器架构，以确保实验环境的一致性。该服务器节点的并发处理能力依赖于底层的gRPC线程池。联邦实验平台架构如图6所示，包含服务器容器、客户端容器和SGX聚合器3类核心容器组件。

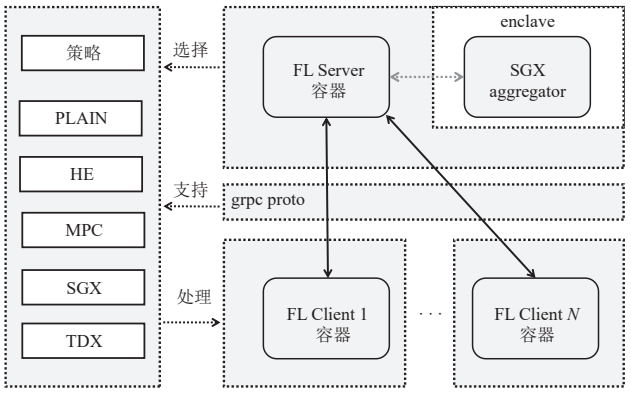


图 6 联邦实验平台架构

1）服务端容器：作为中心协调者，负责从各个客户端收集更新、执行安全聚合算法、评估全局模型性

能以及控制训练流程。在TDX策略下，该容器可直接部署于Intel TDX虚拟机（Trust Domain）中，利用硬件隔离保护聚合过程。

2）客户端容器：模拟独立的联邦学习参与方。每个容器挂载独立的私有数据卷，严守数据不出本地原则。客户端支持GPU加速，负责本地训练及参数的加密封装。

3）SGX 聚合器：专为SGX模式设计的独立容器。该组件运行在Gramine-SGX环境中，负责在Enclave内部执行密钥生成与安全聚合。它通过TCP Socket与服务端容器通信，实现了不可信宿主进程与可信聚合逻辑的解耦。

联邦实验平台的安全机制具体如下。

1）安全边界与运行时保护。如3.1节所述，本文的威胁模型假定服务器端的宿主进程及操作系统均处于非可信空间，而客户端环境被假定为可信。如图6所示，只有核心的聚合计算在TEE中执行。系统不依赖容器隔离来保护运行时数据，而是通过应用层的端到端加密来保护数据，模型更新以密文形式穿透非可信的网络及非可信的服务器宿主进程，最终只在TEE内部解密，确保宿主进程无法窃取客户端的明文梯度。

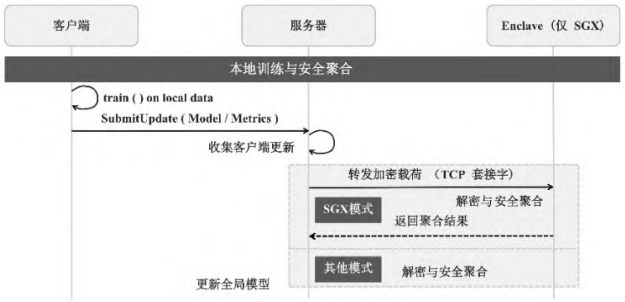
2）身份真实性。实验平台提供了分层级的身份验证机制。服务端身份通过标准TLS加密信道保证，客户端可通过加载根证书验证服务器身份。TEE身份通过远程证明保证，在SGX模式下，服务器会向客户端分发由硬件签名的Quote和Enclave公钥。客户端在生产环境中必须验证Quote的有效性并比对MRENCLAVE值，以确保连接的是未被篡改的聚合程序。

3）数据传输机密性。机密性由两个层面保障：1）信道层支持gRPC-TLS传输加密；2）在应用层，所有隐私策略的模型参数在离开客户端前就已被加密或转换为秘密份额。这意味着，即使TLS信道被攻破，攻击者截获的也只是无法解密的密文，从而确保了核心数据的端到端机密性。

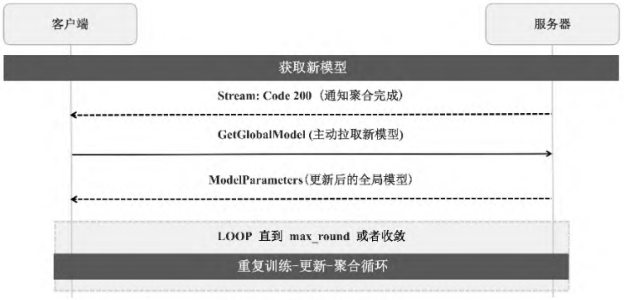
3.4 核心工作流程

本平台的联邦学习生命周期遵循标准的联邦学习流程，完整工作流程如图7所示，主要包含以下3个核心阶段。

a）客户端注册与启动同步



b）本地训练与安全聚合



c）全局模型更新与迭代

图 7 联邦实验平台工作流程

1）客户端注册与启动同步。流程始于初始化阶段，分布式客户端启动后主动向服务端发起注册请求，提交身份标识及本地数据统计信息。服务端在确认达到预设的参与方数量后，向各客户端下发全局配置响应，其中包含当前的隐私增强策略、初始全局模型参数以及必要的密码学材料。为了实现高效的集群协同，建立基于长连接的状态同步通道。服务端利用事件驱动机制，通过流式接口主动向所有已注册的客户端推

送开始训练的指令，从而精确控制分布式节点的同步启动。

2）本地训练与安全聚合。在收到训练指令后，客户端在本地私有数据集上执行模型训练任务，生成梯度或参数更新。随后，系统依据配置的隐私增强策略对更新数据进行安全封装，并通过统一的数据接口提交至服务端。服务端作为协调者，负责收集来自各方的更新载荷。在聚合阶段，系统针对不同的隐私增强策略做出相应的反应。对于SGX硬件隔离模式，服务端将加密载荷通过专用套接字转发至独立的Enclave内部，在硬件保护的内存中完成解密与聚合，确保明文数据对宿主环境不可见。而对于其他模式，服务端则直接在密文域或特定计算域内执行聚合运算，生成新的全局模型。

3）全局模型更新与迭代。当一轮安全聚合完成后，服务端再次通过状态流通道向所有客户端发送聚合完成通知。客户端随即发起请求，主动拉取最新的全局模型参数以覆盖本地状态，为下一轮训练做准备。与此同时，服务端利用全局测试集对新聚合的模型进行性能评估，并记录准确率与收敛情况。这一循环将持续进行，直到达到预设的最大迭代轮次或系统满足收敛条件，此时联邦学习任务正式终止并输出最终实验结果。

3.5 策略详解

3.5.1 BASE 策略

BASE 策略作为本平台的实验对照基准，旨在确立系统在无额外隐私增强技术开销下的性能上限。该模式遵循标准联邦平均（Federated Averaging，FedAvg）流程，模型参数在客户端仅经序列化处理，以明文形式通过安全传输信道提交。服务端直接在非可信内存中执行加权聚合，该策略剔除了应用层加密运算与密文膨胀的影响。通过记录其收敛行为与吞吐量，本研究能够获得算法的原始性能数据，从而为量化后续隐私增强技术的“安全—效率—精度”代价提供精确的参照系。

3.5.2 HE 策略

HE策略采用CKKS（Cheon-Kim-Kim-Song）全同态加密方案，旨在通过其支持单指令多数据（Single Instruction Multiple Data，SIMD）批处理的特性，解决高维模型参数在加密计算中的效率瓶颈。在此模式下，客户端利用向量化编码技术，将大量浮点型权重打包封装至密文块中。为进一步应对同态加密固有的数据膨胀挑战，本平台在通信链路中集成了密文压缩与分块流式传输机制，有效降低了传输带宽占用并规避了大规模参数聚合时的内存溢出风险。服务端在接收密文流的同时在密文域直接执行加法聚合，全程无需解密单个客户端的上传数据，仅在每轮聚合结束后解密全局模型以供分发和评估。该方案通过算法层面的批处理与系统层面的传输优化，在严格保障数据机密性的前提下，显著提升了密文计算的并行效率与系统吞吐量。

3.5.3 MPC 策略

本平台基于Shamir门限机制的秘密共享技术，实现了一种服务器辅助 MPC聚合方案。在此模式下，客户端首先将本地模型参数映射至有限域整数空间，并将其拆分为满足特定恢复阈值的若干秘密份额，随后将这些份额而非原始参数上传。服务器利用秘密共享加法同态特性，直接在份额层面执行高效的向量化聚合运算，最终通过拉格朗日插值法重构全局模型。相比全同态加密，这种基于有限域算术的实现方式规避了高昂的密码学运算，显著降低了计算负载与内存开销。

为适配中心化的实验环境，本文方案在威胁模型上进行了必要的折中。即假设中心服务器遵循“诚实但好奇”原则，服务器持有所有份额并忠实执行聚合逻辑，但不尝试还原单个客户端的私有数据。这种设计虽弱化了理想MPC的去中心化安全假设，但有效剥离了多方协同中的网络通信与状态同步时延。

通过消除网络拓扑的开销，本文方案将评估焦点锁定在秘密共享算法本身的计算复杂度，即有限域算术与多项式插值的开销上。这种单点聚合架构能够模

拟理想网络环境下的计算性能，从而证明即使在网络拓扑最理想的情况下，MPC在高维模型上的计算/通信压力依然巨大。

3.5.4 SGX策略

SGX策略在服务器端构建硬件隔离的Enclave，实现了精细化的进程级隐私保护。为降低应用迁移复杂度，本平台采用Gramine库操作系统作为运行时支撑，将聚合逻辑托管于Enclave内部。在此架构下，仅核心密钥管理与聚合代码运行于受保护的私有内存区域，即便是拥有特权的宿主操作系统，也无法窥探其运行状态，从而显著收敛了系统的TCB。

在信任建立机制上，本平台架构设计支持标准的DCAP（Data Center Attestation Primitives）远程证明流程，但在当前的实验实现中，为聚焦于性能评估并简化部署，本文采用白名单的方式进行本地模拟验证。

针对SGX硬件架构中EPC有限（仅128 MB）这一关键约束，本文方案实施了双重内存优化策略。首先，为压缩单模型的内存占用，客户端在加密上传前将参数由单精度（float32）压缩为半精度（float16）格式，旨在缩小工作集以减少昂贵的页交换开销。此外，为解决多客户端聚合带来的内存峰值叠加问题，本平台采用流式解密聚合机制。宿主进程不会一次性将所有密文载入Enclave，而是通过TCP Socket管道将各客户端的加密载荷逐一推送。Enclave内部采用“即解密—即聚合—即销毁”的流水线模式，每接收一个客户端的密文，便立即在受保护内存中解密、恢复为float32精度并累加至全局模型，随后立即释放该部分临时空间。这种串行化的流式处理大幅降低了Enclave内的峰值内存（所有客户端参数大小之和降至单个客户端参数大小），从而在受限的硬件资源下保障了大规模联邦聚合的稳定性。

3.5.5 TDX策略

TDX策略利用虚拟机级硬件隔离特有的架构透明性，实现了对BASE基线策略的无缝安全增强。系统仅需在BASE的通信流程基础上叠加端到端加密模块与远程证明接口，客户端在上传模型参数前，使用由服务端TD（Trust Domain）生成的公钥执行混合加密，确保数据穿过非受信网络后，仅在受全内存加密（Multi-Key Total Memory Encryption，MKTME）保护的TD内部被解密与聚合。

本文方案对远程证明流程进行了适应性简化，即在本地校验底层驱动生成的Quote报告及硬件度量值，而在验证逻辑上暂时剥离了对在线证书缓存服务（Provisioning Certification Caching Service，PCCS）的依赖。这种设计在最小化系统改造代价的同时，有效构建了基于硬件的可信计算边界。

3.6 安全性分析

本平台严格依据3.1节构建的威胁模型与信任边界进行设计，通过密码学协议与硬件隔离机制的有机结合，在计算、传输及存储的全生命周期内保障联邦学习的安全性。整个安全体系建立在零信任网络与不可信服务端宿主环境的假设之上，核心目标是确保单一客户端的私有梯度信息对除自身以外的任何实体（包括中心服务器管理员）均不可见。

1）传输过程的安全性保障。本平台构建了网络信道与应用数据的双重防御体系，以抵御网络窃听、中间人攻击及恶意篡改风险，确保数据从客户端至聚合终点的端到端安全性。在信道层，系统采用gRPC框架并默认开启TLS加密，通过双向证书验证建立安全的传输隧道，保护所有控制指令与通信载荷不被网络路径上的攻击者截获。此外，平台引入了独立于信道之外的应用层端到端防护机制，确保即使TLS被破解或服务端宿主环境不可信，数据仍处于密文保护之下。具体而言，在TEE模式下，客户端在上传前会执行远程证明，验证服务端是否持有合法的硬件签名（如SGX Quote或TDX Report）以及预期的聚合代码度量值，确认环境可信后，对模型参数进行混合加密。对于HE和MPC模式，模型参数在离开客户端前即被转换为CKKS密文或Shamir秘密份额。这种设计实现了数据穿透不可信网络与非可信操作系统，仅在最终的

安全计算域内才可被解析。

2）聚合过程的安全性保障。针对“诚实但好奇”的服务器威胁模型，平台利用密码学原语与硬件隔离技术，确保聚合逻辑在计算过程中无法泄露单一客户端的梯度隐私。在HE策略中，平台利用CKKS算法的SIMD特性，服务端直接在密文空间执行向量加法聚合，宿主进程仅能接触到语义不可知的密文块。在MPC策略中，平台利用Shamir秘密共享的加法同态特性，设计了基于中心化模拟的聚合协议。尽管服务器接收了所有份额，但严格遵循“诚实但好奇”假设，服务器仅在有限域内对来自不同客户端的份额执行加法聚合，生成全局模型的份额，最后再恢复全局参数。在此协议约束下，服务端仅执行份额层面的代数运算，而不尝试利用持有的份额去重建任何单一客户端的原始明文梯度。对于基于硬件的可信执行环境，平台利用物理隔离边界阻断特权软件的窥探。SGX策略下，宿主进程仅作为不透明的流量转发器，将加密数据流式推送到由Gramine托管的Enclave内部，解密与聚合运算严格限制在受保护的私有内存中进行。TDX策略则将隔离粒度扩展至虚拟机级，通过MKTME技术防止Hypervisor或物理攻击者读取聚合过程中的运行时内存。这4种方案虽然实现路径不同，但均确保了服务端在聚合期间无法获取明文模型更新，仅能获得最终的全局聚合结果。

3）存储过程的安全性保障。在数据存储与管理上，平台确保内存中的数据对服务器宿主而言，仅呈现为密文或分片形式。客户端本地数据安全由其自身保障，符合系统的信任边界划分。对于服务器端，在联邦学习的通信轮次间隙或聚合等待期间，所有缓存的模型更新均保持加密状态。具体而言，HE模式下存储的是CKKS密文对象，MPC模式下存储的是无语义的秘密份额，而在TEE/SGX模式下则是经由AES-GCM封装的二进制加密包。这意味着，即使攻击者获取了服务器操作系统的最高权限（Root）并对内存或磁盘进行转储（Dump），由于缺乏驻留于硬件TEE内部的解密私钥或无法突破同态加密与秘密共享的数学难题，也无法从静态存储中还原出任何单一参与方的隐私数据，从而构筑了最后一道静态防线。

4 实验设置及结果评估

4.1 数据集与模型

4.1.1 MNIST 数据集

实验采用业界标准的手写数字识别数据集MNIST（Modified National Institute of Standards and Technology）[37]。该数据集由Yann LeCun团队于1998年构建，已成为机器学习领域广泛使用的基准数据集之一。如图8所示，原始数据包含70000张28×28像素的灰度图像，涵盖0~9共10类手写数字样本，其官方划分提供60000张训练图像与10000张测试图像。

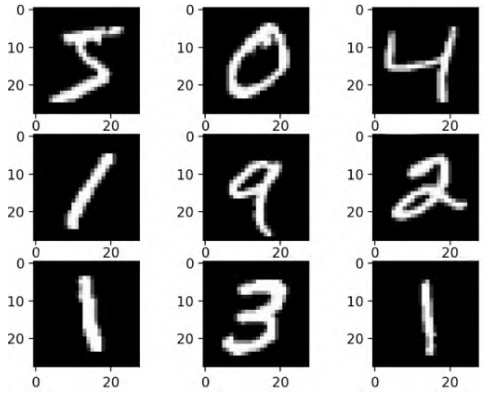


图 8 MNIST 数据集示例

在数据预处理阶段，平台执行标准的归一化操作，将原始像素值映射为标准正态分布，以加速深度学习模型的收敛。为确保评估的客观性，官方提供的10000张测试图像被完整保留并独立保存为全局测试集，仅用于服务端在每轮聚合后评估全局模型的泛化性能，严格与客户端本地训练过程物理隔离。

4.1.2 CIFAR-10 数据集

为验证平台在更复杂视觉任务上的有效性，实验引入了CIFAR-10数据集[38]。如图9所示，该数据集包含 10 个类别（涵盖飞机、汽车、鸟类等通用物体）的60000 张彩色图像，图像尺寸为 32×32 像素，包含RGB三个颜色通道，包括50000张训练图像和10000张测试

图像。

图 9 CIFAR–10 数据集示例

在数据预处理阶段，系统执行标准化的数据处理。首先，将原始图像数据转换为浮点张量，并将像素值映射至标准区间；然后，针对彩色图像的3个通道分别执行标准化处理，依据数据集的全局统计特性（即各通道的均值与标准差）消除不同通道间的数值差异，以加速深度学习模型的收敛速度。最后，官方提供的10000张测试图像在经过同样的预处理后，被独立封存为全局测试集，仅用于服务端在每轮聚合后评估全局模型的泛化准确率。

4.1.3 数据划分

为模拟真实的联邦学习环境并全面评估算法鲁棒性，本文进行联邦数据划分。对于MNIST和CIFAR-10数据集，官方训练集被分配给3个独立的客户端。每个客户端在收到分配的数据后，进一步将其划分为90%的本地训练集和10%的本地验证集，用于本地训练过程中的超参数微调。

针对数据分布异质性，本文实验构建了3种不同的数据分布场景。

1）独立同分布（IID）：将训练样本随机打乱后均匀分配给各客户端，确保每个参与方的数据分布与全局分布在统计学上保持一致。此场景作为理想情况下的性能基准，用于验证算法在无数据异质性干扰下的收敛上限。

2）数量偏斜分布（Non-IID Quantity）：模拟参与方算力与数据收集能力不均的场景。3个客户端的数据量分配比例设定为[0.6, 0.3, 0.1]，即分别持有约60%、30%和10%的训练样本，用于评估聚合算法对贡献度悬殊的客户端的适应性。

3）狄利克雷分布（Non-IID Dirichlet）：为模拟真实世界中复杂且自然的Non-IID场景，本文采用狄利克雷分布（Dirichlet Distribution）对原始数据集进行划分[39]。系统通过设置浓度参数为每个类别生成随机的分配比例，这使得每个客户端虽然理论上拥有所有类别的样本，但各类样本的持有比例存在显著差异。例如，某客户端可能拥有大量某一类别的样本，而另一类别的样本极少。这种划分方式更能反映现实世界中数据自然分布不均的特点。

4.1.4 模型选取

为确保实验结果具有现实指导意义，本文在模型选取上充分考虑从功能验证到真实场景模拟的跨度，选择SimpleCNN与ResNet-18两类架构分别作为轻量级验证与高吞吐压力测试的基准，其差异如表3所示。

表 3 模型参数特征与实验负载对比

模型名称	适配数据集	参数量 / M (10^6)	模型体积 / MB	实验定位
SimpleCNN	MNIST（1×28×28）	0.58	2.3	功能验证
ResNet-18	CIFAR-10（3×32×32）	11.17	44.6	真实场景模拟

针对MNIST任务，本文构建了自定义的SimpleCNN模型，其仅包含两层卷积与两层全连接结构。然而，由于该模型仅有0.58 M参数且计算逻辑过于简单，其较低的负载强度难以有效暴露隐私增强技术在处理复杂任务时的真实开销，因此仅作为验证系统协议连通性的基础对照。

鉴于MNIST任务的计算强度不足以代表现实世界的视觉应用，本文的核心评估主要基于CIFAR-10数据集与ResNet-18模型展开。选择ResNet-18作为基准是由于其作为工业界广泛采用的标准骨干网络，拥有约11.17 M的参数规模与深层残差结构，能够更真实地贴

近生产环境下的计算负载。为适配CIFAR-10的输入特征，本文对模型首层进行了微调（调整卷积核并移除池化层）。这种更贴近现实的配置能够产生足够的通信体积与内存压力，从而避免由于任务过于简单而掩盖了SGX内存分页或同态加密带宽占用等关键性能瓶颈，确保了对各隐私增强技术“安全—效率—精度”权衡的评估具有实际参考价值。

4.2 实验环境

为系统性地评估不同隐私增强技术在联邦学习场景下的性能差异，本文构建了一个包含高性能计算节点与隐私增强计算节点的实验集群。集群内部通过专用局域网互联，实测节点间平均往返通信延迟（Round-Trip Time，RTT）为6.4 ms。实验环境的硬件设施与软件依赖详情如表4所示。

服务端（聚合端）利用宿主机的双重TEE特性（同时支持SGX与TDX）构建了对比环境。为模拟真实的云端机密计算场景，TDX实验被部署于基于KVM的TD中。为严格控制变量以保证公平性，除TDX外的其他实验组均在配置完全一致但未开启TDX特性的普通虚拟机中运行。特别地，SGX实验还受限于128 MB的EPC资源，旨在评估内存受限条件下的系统性能。客户端选用配备高端加速卡的高性能节点，旨在消除本地训练阶段的计算瓶颈，确保实验结果能精准反映不同安全聚合协议带来的通信与计算开销。

表 4 实验环境配置详情

组件	项目	规格参数
服务器宿主机	CPU	Intel® Xeon® Silver 4510 @ 2.40 GHz (12 Physical Cores)
	RAM	128 GB DDR5
	Kernel	Linux 6.8.0-tetd (TDX/SGX Enabled)
	TEE Support	Intel® SGX2 (EPC: 128 MB)& Intel® TDX 1.0
机密虚拟机	vCPU	12 vCPUs (KVM Virtualization)
	RAM	16 GB Allocated
	OS	Ubuntu 24.04.2 LTS (Kernel 6.8.0-generic)
客户端节点	CPU	2 × Intel® Xeon® Silver 4510 (48 Threads)
	RAM	512 GB DDR5
	GPU	NVIDIA H100 PCIe (80 GB HBM3)
软件环境	Framework	PyTorch 2.9.1 (CUDA 12.8), Python 3.12
	TEE Runtime	Gramine 1.9 (for SGX Enclave)
	Deployment	Docker 29.0.4

4.3 实验设置

为系统性地评估不同隐私增强技术在联邦学习场景下的性能差异，本文设计了双层实验架构。首先利用联邦模拟平台确立算法性能基准，然后利用联邦实验平台在真实网络与硬件隔离环境中进行机密计算方案的实测。

4.3.1 联邦模拟平台实验设置

为全面评估平台在不同任务负载与数据分布环境下的性能表现，本文依据任务复杂度与数据异质性两个维度构建了包含集中式训练与联邦学习在内的6组对照实验。在任务维度上，选取MNIST + SimpleCNN（轻量级）与CIFAR-10 + ResNet-18（重量级）分别验证系统原型与高维特征下的负载能力；在分布维度上，首先通过集中式训练确立理论性能上界，进而利用联邦IID场景量化分布式聚合的基准损失，最后引入浓度参数 $\alpha = 0.5$ 的狄利克雷分布构建Non-IID场景以模拟现实中的数据统计异质性。为了更精准地判断各实验组在收敛特性上的差异，本文在模拟阶段并未预设统一的训练轮次，而是采用了自适应的收敛检测机制，通过实时监控模型准确率的变化，自动判定收敛并终止训练。为消除随机性误差，每组实验均独立重复运行5次并取平均值。这种层层递进的实验设计旨在剥离通信与算法因素的干扰，从而为后续量化不同隐私增强技术引入的“安全—效率—精度”开销提供精确的性能参照系。

4.3.2 联邦实验平台实验设置

为量化评估不同隐私增强技术在真实网络环境与硬件隔离条件下的系统开销，本文基于CIFAR-10数据集与ResNet-18模型构建了高维视觉任务场景。在数据分布方面，本文选择了更具挑战性的Non-IID狄利克雷分布（ $\alpha = 0.5$ ）。这是因为狄利克雷分布能够模拟客户端之间显著的标签分布偏移（即统计异质性），更能反映真实联邦场景中数据分布的复杂性，从而对模型的泛化能力与聚合算法的鲁棒性提出了更高的要求。模型选择ResNet-18（参数量约11 M）旨在模拟现实生产环境中的计算负载与通信压力，以充分暴露同态加密与SGX在处理大规模参数聚合时的性能瓶颈。

在此基准任务下，本文设计了BASE（明文基准）、HE（同态加密）、MPC（安全多方计算）、SGX（软件防护扩展）与TDX（信任域扩展）5组对照实验。所有实验组均在统一的客户端配置与网络架构下运行。为统一实验的时间维度以实现系统开销的公平横向对比，本文依据模拟平台的平均收敛轮数（36轮），统一设定训练参数为40轮全局聚合，每轮包含5个本地训练Epoch，以确保各方案在相同的模型收敛状态下进行公平的效率对比。

各实验组的核心差异在于服务端宿主环境的部署架构。

1）BASE、HE与MPC组：服务端部署于普通虚拟机的容器中，分别执行明文、CKKS密文与秘密份额的聚合。该环境运行状态对Hypervisor可见，作为性能评估的基准对照。

2）TDX组：服务端部署于开启了Intel TDX硬件特性的机密虚拟机中。该环境利用MKTME技术实现了虚拟机级别的硬件强制隔离，在计算资源配置（vCPU、内存）与普通虚拟机完全一致的前提下，提供了对宿主不可见的安全执行环境。

3）SGX组：服务端利用Gramine库操作系统将聚合逻辑托管于受SGX硬件保护的Enclave中，实现了进程级别的细粒度隔离。

为了消除网络波动与系统调度带来的随机误差，每组实验均独立重复运行5次，并记录训练总耗时、通信吞吐量等指标的平均值作为最终评估结果。

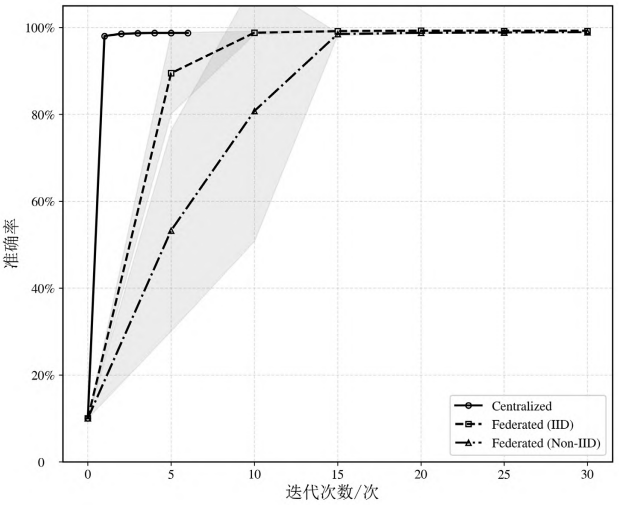
4.4 实验结果分析

4.4.1 联邦模拟平台结果分析

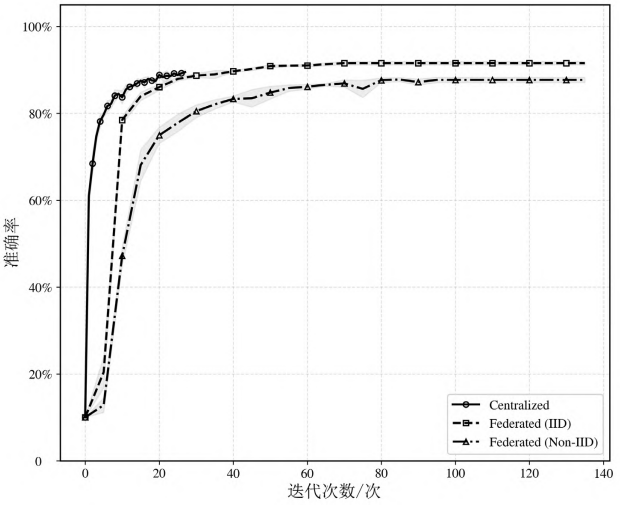
实验结果如表5及图10所示，通过对比不同数据分布下的模型表现，本小节确立了算法在无隐私增强技术开销下的性能基准。表5中，联邦学习客户端本地每训练5个Epoch聚合一次，因此收敛Epoch数 = 联邦学习轮数 ×5，本地—全局精度差是指训练过程中最佳全局模型精度与训练结束时本地模型平均精度之差。

表5 联邦模拟平台实验结果

学习模式	模型架构	数据集	数据分布	准确率	收敛Epoch/轮	通信体积	本地—全局精度差
集中训练	SimpleCNN	MNIST	—	98.82%±0.12%	5.0	—	—
联邦学习			IID	99.28%±0.04%	27.0 (5.4×5)	71.94 MB	0.42%±0.12%
联邦学习			Non-IID	98.94%±0.12%	32.0 (6.4×5)	85.26 MB	3.06%±0.61%
集中训练	ResNet-18	CIFAR-10	—	90.58%±0.43%	38.2	—	—
联邦学习			IID	91.92%±0.26%	82.0 (16.4×5)	4.10 GB	3.68%±0.54%
联邦学习			Non-IID	89.70%±1.03%	180.0 (36.0×5)	8.99 GB	12.80%±1.99%



a）MNIST



b）CIFAR-10

图10 联邦模拟平台收敛图

1）数据分布对收敛效率影响显著，Non-IID场景是主要的性能挑战。实验结果显示，独立同分布（IID）场景下的联邦学习能够达到与集中式训练近乎一致的

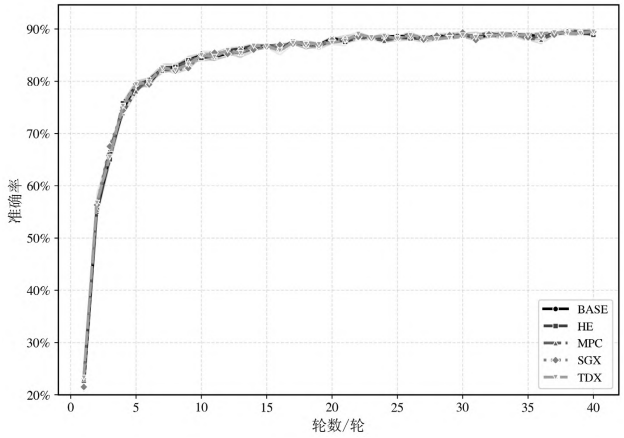
准确率（CIFAR-10上为91.92%），验证了算法的有效性。然而，在更接近真实应用且更具挑战性的Non-IID狄利克雷分布下，ResNet-18模型的全局准确率下降至89.70%，且收敛所需的Epoch数激增至180.0轮（约为数据集中训练的4.7倍）。这表明统计异质性显著加剧了模型聚合的难度，导致通信与计算需求更高。

2）确立了后续隐私增强计算实验的收敛标准与对照基线。模拟结果表明，在CIFAR-10的Non-IID数据分布下，ResNet-18模型大约需要36轮全局聚合（对应180个Epoch）才能达到收敛稳定状态。这为后面实验平台的设置提供标准，即以40轮（略大于平均的36轮）聚合为标准周期，在此基础上量化引入SGX、TDX、HE等隐私增强技术后所增加的系统延迟与吞吐量损耗，从而实现对“安全—效率—精度”权衡的精准评估。

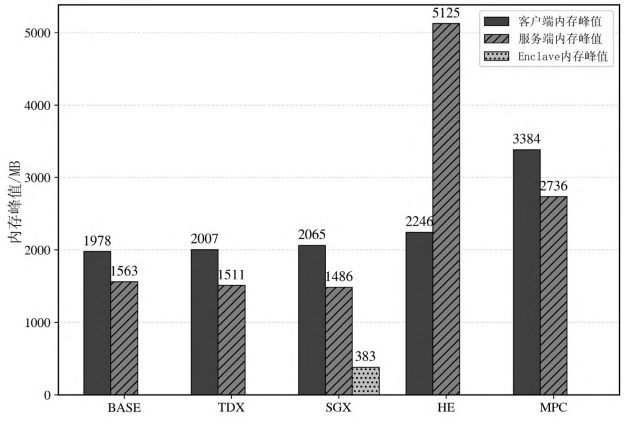
4.4.2 联邦实验平台结果分析

本小节基于CIFAR-10数据集与ResNet-18模型的实验数据，从模型训练的有效性与收敛一致性、系统延迟与计算瓶颈以及通信开销与内存资源3个维度，深入剖析各隐私增强技术在联邦学习场景下的性能表现与内在机理。隐私增强环境下的模型准确率收敛曲线与系统内存峰值开销对比如图11所示，各实验组内存资源消耗峰值与模型模型准确率统计如表6所示。

1）模型训练的有效性与收敛一致性。实验结果证实了引入隐私增强技术并未破坏联邦学习算法的收敛特性。如图11 a）所示，在40轮的训练周期内，BASE、TDX、SGX、HE与MPC这5种策略的模型收敛曲线高度重合。结合表6的准确率统计，各实验组的准确率均稳定在89.50%左右，组间最大差异仅为0.62%。这种高度的一致性源于各方案在数学层面的等价性，TEE方案（SGX/TDX）仅将通用内存中的浮点数运算迁移至硬件隔离的加密内存中执行，其底层的运算逻辑未发生改变，因此不会引入计算误差；而密码学方案中，CKKS同态加密虽引入了微小噪声但仍在深度学习容错范围内，Shamir秘密共享则是基于有限域的精确算术运算。这表明本平台所集成的隐私增强技术成功在不牺牲模型效用的前提下实现了安全聚合。



a）模型准确率收敛曲线



b）内存峰值对比

图 11 隐私增强环境下的模型准确率收敛曲线与系统内存峰值开销对比

表 6 各实验组内存资源消耗峰值与模型准确率统计

策略	客户端内存峰值 /MB	服务端内存峰值 /MB	Enclave内存峰值 /MB	准确率
BASE	1978.5±61.9	1563.5±159.7	—	89.36%±0.44%
HE	2246.0±178.7	5124.8±71.0	—	88.89%±0.28%
MPC	3383.6±232.4	2735.6±129.1	—	89.49%±0.45%
SGX	2064.8±68.2	1485.9±116.1	383.1±13.3	89.51%±0.59%
TDX	2006.6±72.4	1511.3±129.4	—	89.50%±0.28%

2）系统延迟与计算瓶颈。图12的单轮训练耗时分解与表7的详细时间数据揭示了不同技术路线在系统延迟上的显著分层。TDX 策略展现了在大规模负载下的架构优势，其总耗时（31.50 s）仅比明文基线（31.08 s）增加约1.3%，实现了近乎原生的性能表现。这种高效性归功于TDX的架构透明性，利用MKTME引擎在内存控制器层面自动处理加解密，避免了传统

TEE频繁的用户态—内核态上下文切换，仅在虚拟机退出（VM Exit）与远程证明握手时引入了较小的开销。

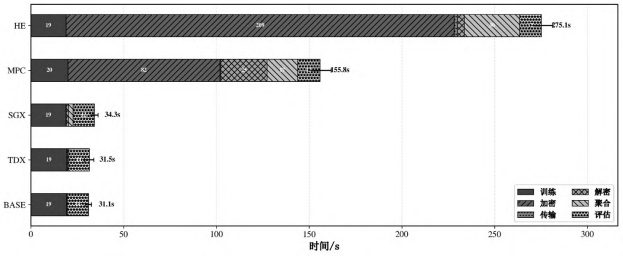


图 12 不同策略在真实网络环境下的单轮训练耗时分解

表 7 不同策略下的单轮训练端到端耗时分解与通信开销

策略	训练 /s	加密 /s	传输 /s	解密 /s	聚合 /s	评估 /s	总耗时 /s	上传量 /MB
BASE	19.20 ± 0.82	0.40 ± 0.07	0.09	0.00 ± 0.00	0.06 ± 0.01	11.33 ± 1.40	31.08	42.7
HE	18.99 ± 2.20	209.22 ± 5.67	1.80	3.60 ± 0.22	29.86 ± 0.23	11.68 ± 0.57	275.15	897.6
MPC	19.97 ± 3.21	81.90 ± 4.92	0.51	24.86 ± 0.47	16.58 ± 0.32	12.02 ± 1.06	155.84	256.0
SGX	18.88 ± 0.89	0.16 ± 0.02	0.04	1.09 ± 0.07	2.55 ± 0.17	11.57 ± 1.67	34.29	21.3
TDX	19.26 ± 2.37	0.45 ± 0.08	0.09	0.59 ± 0.11	0.06 ± 0.01	11.05 ± 0.83	31.50	42.7

相比之下，SGX 策略的总耗时为34.29 s，比基线增加了约10%。这一延迟增量的成因较为复杂，鉴于ResNet-18的参数规模与PyTorch运行时的内存放大效应，实测Enclave内存峰值（383 MB）已显著超过物理EPC容量（128 MB）。这意味着EPC页面换出（EPC Paging）带来的性能惩罚不可避免。性能损耗极低，运行效率与明文训练基本持平，且可提供硬件级安全防护，FP16压缩与流式聚合机制发挥了关键作用，有效降低了活跃工作集（Work Set）的大小，避免了因高频缺页导致的灾难性页面颠簸（Thrashing）。因此，该延迟主要由AES-GCM加解密计算、Enclave边界交互（ECALL/OCALL）以及必要的EPC与系统内存间的数据交换开销共同构成。这一结果不仅揭示了SGX在处理大规模深度学习模型时的硬件内存瓶颈，更论证了在物理EPC受限的情况下，应用层内存优化是缓解硬件局限、维持高性能隐私计算的核心手段。

而纯密码学方案HE与MPC则受限于算术复杂度，如图12所示，HE方案CKKS复杂的多项式乘法与模运算，导致客户端加密耗时高达209 s，成为计算瓶颈。MPC虽通过简化的有限域运算将加密时间缩短至82 s，但仍远逊于硬件方案。

3）通信开销与内存资源。通信开销与内存资源的分析揭示了密码学方案在传输层面的局限性。表7的数据直观地反映了全同态加密固有的密文膨胀难题，HE策略的单轮上传量激增至897.6 MB，是基线的21倍。这是因为CKKS为支持密文域运算，需要维护庞大的多项式系数与模数链，导致承载单一浮点数的密文体积增长，这在现实的广域网环境中将造成巨大的带宽压力。同时，图11 b）的内存峰值对比结果显示，HE在服务端峰值内存消耗高达5125 MB，限制了其在边缘设备上的部署。相比之下，TEE方案在资源消耗上表现优异。TDX与SGX仅对数据进行标准的AES-GCM封装，几乎不增加额外载荷。此外，如表7所示，SGX策略在采用FP16压缩传输优化后，将上传量进一步降低至21.3 MB，且利用流式聚合机制将Enclave进程内存峰值成功控制在383 MB。然而，即使经过此类软件层面的极致优化，物理EPC依然是制约SGX性能上限的关键因素。

综上所述，实验结果清晰地界定了各隐私增强技术在横向联邦学习场景下的适用边界。对于高维深度学习模型的安全聚合，Intel TDX凭借其虚拟机级的硬件隔离与对上层应用的透明性，相比明文基线 BASE 策略，SGX 策略的单轮端到端总耗时由 31.08 s 增至34.29 s，额外开销约为 3.21 s，这表明是当前平衡安全需求与性能开销的最优解。SGX虽然提供了更细粒度的隔离，但在处理大规模模型时面临严峻的I/O瓶颈。而HE与MPC尽管提供了可形式化验证的安全性，但在面对千万级参数的神经网络时，其巨大的计算延迟与通信代价使其难以满足实际生产环境对实时性的要求。

5 结束语

本文针对横向联邦学习中使用态数据的隐私保护与性能权衡难题，构建了机密联邦学习平台CFLP，在

典型视觉任务下系统量化评估了Intel TDX、SGX、HE与MPC这4种主流隐私增强技术。实验结果表明，尽管各方案均能在不损失模型精度的前提下实现安全聚合，但纯密码学方案（HE与MPC）在高维模型场景下面临严峻的计算与通信瓶颈，SGX则受限于物理内存容量约束。相比之下，Intel TDX凭借虚拟机级隔离特性，在仅增加微小系统延迟的情况下提供了硬件级机密性保障与近乎原生的性能体验。本文明确了各技术的适用边界，表明对于大规模深度学习模型的安全聚合，TDX是当前平衡安全需求、性能开销与部署便捷性的最优技术路径，为构建高效可信的联邦学习系统提供了坚实的实证依据。

参考文献：

[1] SHOKRI R, SHMATIKOV V. Privacy-Preserving Deep Learning[C]//ACM. The 22nd ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2015: 1310-1321.

[2] MCMAHAN B, MOORE E, RAMAGE D, et al. Communication-Efficient Learning of Deep Networks from Decentralized Data[C]//PMLR. The 20th International Conference on Artificial Intelligence and Statistics. New York: PMLR, 2017: 1273-1282.

[3] YANG Li, ZHU Lingbo, YU Yueming, et al. Review of Federated Learning and Offensive-Defensive Confrontation[J]. Netinfo Security, 2023, 23(12): 69-90.
杨丽，朱凌波，于越明，等.联邦学习与攻防对抗综述[J].信息网络安全，2023，23（12）：69-90.

[4] HITAJ B, ATENIESE G, PEREZ-CRUZ F. Deep Models under the GAN: Information Leakage from Collaborative Deep Learning[C]//ACM. The 24th ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2017: 603-618.

[5] GILAD-BACHRACH R, DOWLIN N, LAINE K, et al. Cryptonets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy[C]//PMLR. The 33rd International Conference on Machine Learning. New York: PMLR, 2016: 201-210.

[6] SHEN Youren, TIAN Hongliang, CHEN Yu, et al. Occlum: Secure and Efficient Multitasking inside a Single Enclave of Intel SGX[C]//ACM. The Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems. New York: ACM, 2020: 955-970.

[7] KAIROUZ P, MCMAHAN H B, AVENT B, et al. Advances and Open Problems in Federated Learning[J]. Foundations and Trends® in Machine Learning, 2021, 14(1-2): 1-210.

[8] HE Zeping, XU Jian, DAI Hua, et al. A Review of Federated Learning Application Technologies[J]. Netinfo Security, 2024, 24(12): 1831-1844.
何泽平，许建，戴华，等.联邦学习应用技术研究综述[J].信息网络安全，2024，24（12）：1831-1844.

[9] LIU Yang, KANG Yan, ZOU Tianyuan, et al. Vertical Federated Learning: Concepts, Advances, and Challenges[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(7): 3615-3634.

[10] GAO Ying, CHEN Xiaofeng, ZHANG Yiyu, et al. A Survey of Attack and Defense Techniques for Federated Learning Systems[J]. Chinese Journal of Computers, 2023, 46(9): 1781-1805.
高莹，陈晓峰，张一余，等.联邦学习系统攻击与防御技术研究综述[J].计算机学报，2023，46（9）：1781-1805.

[11] BONAWITZ K, IVANOV V, KREUTER B, et al. Practical Secure Aggregation for Privacy-Preserving Machine Learning[C]//ACM. The 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2017: 1175-1191.

[12] LI Yipeng, LYU Xinchen. Convergence Analysis of Sequential Federated Learning on Heterogeneous Data[J]. Advances in Neural Information Processing Systems, 2023, 36: 56700-56755.

[13] NING Zhenyu, ZHANG Fengwei, SHI Weisong. Research on Trusted Execution Environment Based on Edge Computing[J]. Journal of Computer Research and Development, 2019, 56(7): 1441-1453.
宁振宇，张锋巍，施巍松.基于边缘计算的可信执行环境研究[J].计算机研究与发展，2019，56（7）：1441-1453.

[14] ARM Limited. TrustZone for Armv8-A[EB/OL]. [2025-11-31]. https://developer.arm.com/-/media/Arm%20Developer%20Community/PDF/Learn%20the%20Architecture/TrustZone%20for%20Armv8-A.pdf.

[15] COSTAN V, DEVADAS S. Intel SGX Explained[R]. IACR, Cryptology ePrint Archive, Report 2016/086, 2016.

[16] Intel Corporation. Intel Trust Domain Extensions[EB/OL]. (2021-05-01)[2025-12-10]. https://www.intel.com/content/www/us/en/developer/tools/trust-domain-extensions/documentation.html.

[17] MCKEEN F, ALEXANDROVICH I, BERENZON A, et al. Innovative Instructions and Software Model for Isolated Execution[C]//ACM. The 2nd International Workshop on Hardware and Architectural Support for Security and Privacy. New York: ACM, 2013: 10.

[18] TSAI C, PORTER D E, VIJ M. Graphene-SGX: A Practical Library OS for Unmodified Applications on SGX[C]//USENIX. 2017 USENIX Annual Technical Conference. Berkeley: USENIX, 2017: 645-658.

[19] PINTO S, SANTOS N. Demystifying Arm TrustZone: A Comprehensive Survey[J]. ACM Computing Surveys, 2019, 51(6): 1-36.

[20] LIPP M. Cache Attacks and Rowhammer on ARM[D]. Graz: Graz University of Technology, 2016.

[21] ARM LTD. ARM TrustZone for Cortex-A: Technical Reference Manual[R]. Cambridge: ARM Ltd, Revision r1p0, 2022.

[22] LI Mengyuan, ZHANG Yinqian, WANG Huibo, et al. CIPHERLEAKS: Breaking Constant-Time Cryptography on AMD SEV via the Ciphertext Side Channel[C]//USENIX. The 30th USENIX Security Symposium (USENIX Security 21). Berkeley: USENIX, 2021: 717-732.

[23] ZHANG Yiming, HU Yuxin, NING Zhenyu, et al. SHELTER: Extending ARM CCA with Isolation in User Space[C]//USENIX. The 32nd USENIX Security Symposium (USENIX Security 23). Berkeley: USENIX, 2023: 6257-6274.

[24] ZHANG Fengwei, ZHOU Lei, ZHANG Yiming, et al. Trusted

Execution Environment: Status and Prospects[J]. Journal of Computer Research and Development, 2024, 61(1): 243-260.
张锋巍，周雷，张一鸣，等. 可信执行环境：现状与展望[J]. 计算机研究与发展，2024，61（1）：243-260.
[25] VAN BULCK J, PIESSENS F, STRACKX R. SGX-Step: A Practical Attack Framework for Precise Enclave Execution Control[C]//ACM. The 2nd Workshop on System Software for Trusted Execution. New York: ACM, 2017: 1-6.
[26] BAUMANN A, PEINADO M, HUNT G. Shielding Applications from an Untrusted Cloud with Haven[J]. ACM Transactions on Computer Systems (TOCS), 2015, 33(3): 1-26.
[27] FREDRIKSON M, JHA S, RISTENPART T. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures[C]//ACM. The 22nd ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2015: 1322-1333.
[28] ZHU Ligeng, LIU Zhijian, HAN Song. Deep Leakage from Gradients[J]. Advances in Neural Information Processing Systems, 2019, 32: 14400-14409.
[29] NASR M, SHOKRI R, HOUMANSADR A. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-Box Inference Attacks against Centralized and Federated Learning[C]//IEEE. 2019 IEEE Symposium on Security and Privacy (SP). New York: IEEE, 2019: 739-753.
[30] ABADI M, CHU A, GOODFELLOW I, et al. Deep Learning with Differential Privacy[C]//ACM. The 23rd ACM Conference on Computer and Communications Security. New York: ACM, 2016: 308-318.
[31] GHAZI B, GOLOWICH N, KUMAR R, et al. Deep Learning with Label Differential Privacy[J]. Advances in Neural Information Processing Systems, 2021, 34: 27131-27145.
[32] MOHASSEL P, ZHANG Yupeng. Secureml: A System for Scalable Privacy-Preserving Machine Learning[C]//IEEE. 2017 IEEE Symposium on Security and Privacy (SP). New York: IEEE, 2017: 19-38.
[33] CHEN Yu, LUO Fang, LI Tong, et al. A Training-Integrity Privacy-Preserving Federated Learning Scheme with Trusted Execution Environment[J]. Information Sciences, 2020, 522: 69-79.
[34] PENG Wei, LI Yinshuai, ZHANG Yinqian. Shadows in Cipher Spaces: Exploiting Tweak Repetition in Hardware Memory Encryption[C]//USENIX. The 34th USENIX Security Symposium (USENIX Security 25). Berkeley: USENIX, 2025: 5759-5776.
[35] WITHARANA H, WEERASENA H, MISHRA P. Formal Verification of Virtualization-Based Trusted Execution Environments[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2024, 43(11): 4262-4273.
[36] ZHANG Ziqi, GONG Chen, CAI Yifeng, et al. No Privacy Left Outside: On the (In-) Security of TEE-Shielded DNN Partition for On-Device ML[C]//2024 IEEE Symposium on Security and Privacy (SP). New York: IEEE, 2024: 3327-3345.
[37] LECUN Y, CORTES C. The MNIST Database of Handwritten Digits[EB/OL]. (1998-01-01)[2025-01-15]. http://yann.lecun.com/exdb/mnist/.
[38] KRIZHEVSKY A, HINTON G. Learning Multiple Layers of Features from Tiny Images[EB/OL]. (2009-09-01)[2025-12-15]. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.
[39] HSU T M H, QI Huiqi, BROWN M. Measuring the Effects of Non-Identical Data Distribution for Federated Visual Classification[EB/OL]. (2019-09-13)[2025-12-15]. https://arxiv.org/abs/1909.06335.