立项申请

项目主要内容简介 200-

共享智能场景下，涉及GPU的数据交互缺乏保护。本项目针对共享智能场景下深度学习等主流机器学习算法环境的数据安全保护需求，研究基于主流加速硬件GPU与CPU混合异构框架下如何构建可信执行环境（TEE）的技术。分析主流机器学习算法，研究单块及多GPU架构及其工作机理，研究基于Intel SGX安全保护方法，研究TEE+SGX的Graviton方案的性能优化和OpenCL+SGX的技术，实现一套GPU上的SGX安全保护软件。通过以上关键技术突破，形成论文、专利、软件原型等关键技术成果。

项目背景、目的及意义 1100-

背景：

大数据时代无论是金融授信风控或在线营销等都离不开数据。数据质量和数量已成为影响机器学习效果最重要的因素之一，通过数据共享扩充数据量以提升模型效果的需求已经变得越来越强烈。

同时，由于数据安全方面的以下两方面的隐忧，行业应用领域大数据积累形成的多部门数据间没有共享的安全条件。

基于现况，可信执行环境TEE技术如Intel SGX软件扩展指令技术得到了发展。此技术可以用于增强纯CPU主机（非GPU环境下的）的数据安全保护功能，是现有CPU硬件环境操作系统提供安全的方案。基本思路是通过SGX提供指令创建一块飞地内存区域

（只能通过SGX的指令访问），将重要数据存入其中，即使操作系统被攻击，任何其它应用的CPU指令无法访问，从而实现数据保护。

但目前的保护仍有不足，在当前机器学习算法运行的框架环境下，传统CPU计算环境无法满足快速需求， GPU作为运算模块得到广泛应用。GPU的存在扩大了硬件系统的受攻击面，而目前在CPU和GPU交互方面的研究还有待开展。

目的：

本项目的目标是针对共享智能场景下深度学习等主流机器学习算法环境的数据安全保护需求，研究基于主流加速硬件GPU与CPU混合异构框架下如何构建可信执行环境（TEE）的技术。具体如下：

1. 对主流机器学习算法特别是深度学习展开研究，分析算法特点，包括运行机理、数据内容及特征，刻画关键数据保护的安全级别，提出相应级别的安全保护需求；针对各级数据安全保护需求，研究提出TEE的安全保护方法。
2. 对基于单块加速硬件卡（GPU）的简单架构及其工作机理展开研究，分析CPU计算与GPU计算之间协同工作机制和关键信息流数据，研究基于Intel SGX安全保护方法。
3. 对多GPU混合架构下深度学习框架如Tensorflow架构及工作机理展开研究，研究Nvidia GPU的CUDA、通用GPU的OpenCL接口支持下的工作机理和数据信息流，研究多GPU架构下面向深度学习框架的TEE安全保护技术；研究Graviton方案的性能优化技术。

通过以上关键技术突破，形成论文、专利、软件原型等关键技术成果。

意义：

在当前CPU与GPU混合的架构下，操作系统管理GPU是将其作为传统主机上增加的设备来管理的，主机访问设备是通过与设备驱动的交互来实现的。有了GPU之后，GPU内部主要包含内核和GPU存储（称为device内存），而GPU之外的部分就是host主机，即CPU所在环境。

如果一个CUDA开发的应用需要GPU完成计算程序，需要先在设备上分配内存，从主存中装入数据，GPU内核计算处理完毕后，应用把结果复制到主机（host）内存并且释放分配到GPU上的内存。

事实上，当前CUDA应用与GPU设备之间的数据交互过程是缺乏安全保护的。因此，在CPU和GPU混合的架构下，研究构建GPU+TEE环境保护关键数据，对于增强安全性、实现有效保护数据隐私而充分发挥数据真正价值具有重要意义。

项目研究方案 1200-

- **研究目标**

在当前机器学习算法运行的框架环境下，传统CPU计算环境无法满足快速需求，基于GPU的运算得到广泛应用。我们希望通过研究Intel SGX技术原理、分布式机器学习原理、分析SGX开源库的使用、研究GPU CUDA、OpenCL等编程技术，在分布式的机器学习环境下，引入SGX安全保护技术，构建可信执行环境（TEE），实现GPU关键数据的安全保护。

- **主要内容**

SGX（software guard extensions）是Intel公司在2013年推出的指令集扩展，以硬件安全为强制性保障，为用户空间提供可信执行环境。本研究旨在构建基于SGX的TEE环境，进而实现对GPU架构中数据的保护，详细介绍如下：

一是对主流机器学习算法特别是深度学习展开研究，刻画关键数据保护的安全级别，提出相应级别的安全保护需求，研究提出TEE的安全保护方法。

二是对基于单块加速硬件卡（GPU）的简单架构及其工作机理展开研究，分析CPU计算与GPU计算之间协同工作机制和关键信息流数据，研究基于Intel SGX安全保护方法。

三是对多GPU混合架构下深度学习框架如Tensorflow架构及工作机理展开研究，研究nvidia GPU的CUDA、通用GPU的OpenCL接口支持下的工作机理和数据信息流，研究多GPU架构下面向深度学习框架的TEE安全保护技术。研究Graviton方案的性能优化技术。

- **研究方法**

技术路线：

（1）通过调研国内外研究现状，分析现有方案如引力子Graviton方案的可行性；

（2）部署搭建GPU、SGX模拟或真实环境运行主流的深度学习框架，研究在分布式联合学习场景下，单/多节点GPU/CPU混合架构深度学习框架的工作机理与数据流，获取安全保护需求并进行安全保护评价定性定级评估；

（3）攻破SGX在GPU上构建TEE的技术，针对nvidia GPU场景应用，研究CUDA与SGX之间的安全接口实现技术，研究性能优化技术提升Graviton方案性能；

（4）针对支持通用的加速硬件如AMD GPU、Intel GPU以及nvidia GPU场景，研究OpenCL与SGX之间的安全接口实现技术；形成SGX GPU中间件技术成果和论文，并在实际应用开展验证。

GPU-TEE技术方案：

针对GPU架构环境，构建一个基于SGX的TEE环境，具体分为以下三方面：

（1）针对TEE的操作管理，研究原语。

（2）研究GPU-TEE安全共享智能软件架构、研究远程认证、安全区内容管理与安全内容隔离性机制。实现一套通用的GPU-TEE软件。

（3）研究现有Graviton方案的性能优化与OpenCL通用场景的TEE技术。

项目研究条件与创新之处 500-

研究条件：

1. 浙江大学智能计算与系统实验室INCAS-LAB高性能计算与机器学习部分研究基础。
2. 近年来有部分SGX+GPU相关的研究论文和成果，可以给我们提供学习的范本。
3. 成员已有针对nvidia gpu的CUDA编程基础、并行计算方面基础。

尚缺少的条件：

1. 没有OpenCL、C++ AMP的通用gpu编程基础。

解决办法：尽快学习掌握相关的编程语言。

项目创新点：

在当前CPU与GPU混合的架构下，操作系统管理GPU是将其作为传统主机上增加的设备来管理的，主机访问设备是通过与设备驱动的交互来实现的。以nvidia GPU为例，访问GPU的工具是CUDA的API。事实上，当前CUDA应用与GPU设备之间的数据交互过程是缺乏安全保护的。尽管现今已有少数如Graviton引力子的SGX+GPU的安全保护方案，但其只支持nvidia显卡、时间开销较大等局限性，意味着SGX+GPU的方案可优化、提升、拓展的空间很大。

本项目研究如何在CPU和GPU混合的架构下，针对GPU支持的机器学习环境，构建TEE环境保护关键数据，力求在共享智能的安全增强方面有所突破与创新。项目将得到浙江大学智能计算与系统实验室的指导，具有比较高的可行性和广泛的应用前景。

参考资料

GPU架构下分布式机器学习的SGX安全保护技术研究

简介：大数据时代无论是金融授信风控或在线营销等都离不开数据。数据质量和数量已成为影响机器学习效果最重要的因素之一，通过数据共享扩充数据量以提升模型效果的需求也变得越来越强烈。数据共享日益重要，但数据安全难以得到有效保障，出现了数据买卖、泄露和滥用等诸多问题。Intel SGX是Intel CPU上的软件扩展指令集，提供了一种安全的可信执行环境TEE（Trust executive environment），从硬件

层面来保护业务的关键数据，即使操作系统被攻击，也能够确保关键数据的安全。在分布式机器学习环境下，机器学习算法的数据来自多方，不同部门数据需要保护隐私，涉及到使用中用户身份的认证和签名验证。在分布式机器学习的整个架构中，机器学习算法执行的一部分是在CPU上，一部分会在GPU上，当数据从主机内存提交到GPU存储的过程中，可能存在数据隐私泄露。

研究目标及内容：本项目研究Intel SGX技术原理，分布式机器学习原理，分析SGX开源库的使用，研究GPU cuda编程技术，在分布式的机器学习环境下，引入SGX安全保护技术，实现GPU关键数据的安全保护。

1.1. 研究方法

配图说明研究项目拟采用的技术路线与研究方法，指出其相对现有技术方法的不同点，突出其优势。不超过3页。

1.1.1. 技术路线

- 通过调研国内外研究现状，分析现有方案如*引力子Graviton*方案的可行性；
- 部署搭建GPU、SGX模拟或真实环境运行主流的深度学习框架，研究在分布式联合学习场景下，单/多节点GPU/CPU混合架构深度学习框架的工作机理与数据流，获取安全保护需求并进行安全保护评价定性定级评估；
- 攻破SGX在GPU上构建TEE的技术，针对nvidia GPU场景应用，研究CUDA与SGX之间的安全接口实现技术，研究性能优化技术提升Graviton方案性能；
- 针对支持通用的加速硬件如AMD GPU、Intel GPU以及nvidia GPU场景，研究OpenCL与SGX之间的安全接口实现技术；形成SGXGPU中间件技术成果和论文，并在实际应用开展验证。

1.1.2. 研究方法

本研究首先对当前国内外SGX在加速硬件场景的应用研究展开调研，结合分布式机器学习场景，研究实现*引力子*的基础上提出GPUTEE方案，开展性能评估与优化改进，并开展实际验证，总结形成论文专利成果。

1. **分布式机器学习TEE应用方案**
   为了将多方数据在可信执行环境中进行加密共享和融合学习，本方案中数据提供方的原始数据通过认证模块的公钥加密后，将加密数据上传到可信模型平台的可信计算环境（TEE）中，模型平台的认证模块通过私钥对加密数据进行解密后，发送给融合学习模块，对解密后的数据进行机器学习。TEE方案中各参与方加密本地数据，并上传到模型平台。任何一个参与方都可以发起训练任务。模型平台创建可信环境，解密接收到的各个参与方发送来的加密数据，基于解密数据进行模型训练，得到共享模型。最后销毁可信环境，以实现数据安全和隐私保护。整体方案的技术架构如图1所示。

2. **GPUTEE技术方案**
   方案的目标是针对GPU架构环境，构建一个基于SGX的TEE环境，隔离主机不可信软件和其他TEE环境，用于存储保护GPU设备内存、访问命令队列、寄存器等相关资源数据，并利用公私钥对的方式进行访问。通过将密钥放在CPU的TEE中实现保护密钥，使攻击者无法访问安全内容环境的地址空间。
   （1）针对TEE的操作管理，**研究原语**，即1）度量原语，用于给出生成远程可验证的TEE上下文环境状态及平台的摘要或总结；2）安全内存分配和重分配原语，用于使设备驱动能够动态分配和释放内存并不影响安全性。
   （2）研究GPUTEE安全共享智能软件架构，研究远程认证、安全区内容管理与安全内容隔离性机制，实现一套**通用的GPUTEE软件**。具体如下：

   a) 在GPU的SGX远程认证方面，研究构建安全区的访问机制，设计好根签名密钥，签名用非对称密钥生成和签名的算法；

   b) 在SGX安全区的内容管理方面，设计多种不同的访问通道，实现不同安全保护内容管理；针对不同的安全访问通道，对GPU命令处理器进行扩展支持命令的创建、管理和撤销。

   c) 在安全内容的隔离性方面，研究实现包括内存区域访问的安全管理、地址空间、所有权追踪、命令集、引导机制以及内存大页支持等。

   d) 实现基于GPU+SGX的TEE安全共享智能软件架构：研究CUDA等的API实现TEE，针对内存管理、内存数据拷贝、内核调用命令以及数据流管理等方面，引入SGX的安全访问机制，并集成多方机器学习的共享智能安全技术，形成一个统一的共享智能安全软件架构。
   （3）研究现有**Graviton**方案的性能优化与**OpenCL**通用场景的TEE技术。Graviton基于CUDA+Nvidia GPU的TEE，在性能上开销过高超过17%，研究引入并行优化技术提升身份验证过程中频繁使用的加密解密操作性能，使整体性能优化提升Graviton性能不低于10%。研究OpenCL支持通用异构GPU的TEE安全。

研究目标：

本项目的目标是针对共享智能场景下深度学习等主流机器学习算法环境的数据安全保护需求，研究基于主流加速硬件**GPU与CPU混合异构框架**下如何构建**可信执行环境（TEE）**的技术。具体如下：

1. 对主流机器学习算法特别是**深度学习**展开研究，分析算法特点，包括运行机理、数据内容及特征，刻画关键数据保护的安全级别，提出相应级别的安全保护需求；针对各级数据安全保护需求，研究提出TEE的安全保护方法。
2. 对基于**单块加速硬件卡**（GPU）的简单架构及其工作机理展开研究，分析CPU计算与GPU计算之间协同工作机制和关键信息流数据，研究基于Intel SGX安全保护方法。

3. 对**多GPU**混合架构下深度学习框架如Tensorflow架构及工作机理展开研究，研究nvidia GPU的CUDA、通用GPU的OpenCL接口支持下的工作机理和数据信息流，研究多GPU架构下面向深度学习框架的TEE安全保护技术。研究Graviton方案的性能优化技术。
   通过以上关键技术突破，形成论文、专利、软件原型等关键技术成果。

1.1. 项目背景及研究意义

1.2.1 项目背景

大数据时代无论是金融授信风控或在线营销等都离不开数据。数据质量和数量已成为影响机器学习效果最重要的因素之一，通过数据共享扩充数据量以提升模型效果的需求已经变得越来越强烈。然而，当前数据共享面临以下两方面问题：
（1）数据安全风险。数据共享日益重要，但数据安全难以得到有效保障，数据买卖、泄露和滥用等现象严重；
（2）公众和政府日益重视数据隐私保护，数据隐私保护要求被提到了一个新的高度。比如，欧盟2016年通过2018年开始实施的《通用数据保护法案（GDPR）》，明确指出所有与个人相关的信息都是个人数据，数据的使用行为必须有明确授权。
因此，在合规监管大环境下，提出共享智能并开展实践，实现有效保护数据隐私的同时充分发挥数据真正价值，具有重要意义。共享智能的目标是基于数据安全和隐私保护，在多个参与方之间通过共享加密数据或加密机制下的参数交换与优化，使用机器学习建立虚拟共享模型的平台。

1.2.2 研究GPU+TEE的意义

在当前机器学习算法运行的框架环境下，传统CPU计算环境无法满足快速需求，基于GPU的运算得到广泛应用。同时，由于互联网安全，行业应用领域大数据积累形成的多部门数据间的无法共享。可信执行环境TEE技术如Intel SGX软件扩展指令技术的出现，可以用于增强纯CPU主机（非GPU环境下的）的数据安全保护功能，为现有CPU硬件环境操作系统提供安全的方案。基本思路是通过SGX提供指令创建一块飞地内存区域，让后将重要数据存入其中，这个区域只能通过SGX的指令访问，即使操作系统被攻击，任何其它应用的CPU指令无法访问，从而实现数据保护。
在当前CPU与GPU混合的架构下，操作系统管理GPU是将其作为传统主机上增加的设备来管理的，主机访问设备是通过与设备驱动的交互来实现的。有了GPU之后，GPU内部主要包含内核和GPU存储（成为device内存），而GPU之外的部分就是host主机，即CPU所在环境。以nvidia GPU为例，访问GPU的工具是CUDA的API。一个CUDA开发的应用需要GPU完成矩阵相乘，基本工作过程如下：应用首先调用CUDA的API创建一个新的CUDA上下文，接着分配一段内存用于存储输入和输出数据，数据装入后再调用GPU内核完成矩阵相乘，同时传递指向设备内

存的指针和其它内核参数。GPU内核计算处理完毕后，应用把结果复制到主机（host）内存并且释放分配到GPU上的内存。

事实上，当前CUDA应用与GPU设备之间的数据交互过程是缺乏安全保护的。因此，在CPU和GPU混合的架构下，研究构建GPU+TEE环境保护关键数据，对于增强安全性具有重要意义。本项目研究如何在GPU的环境下，针对GPU支持的机器学习环境，构建TEE环境保护关键数据，力求在共享智能的安全增强方面有所突破与创新。