难题1：高效支持批处理可编程自举的全同态加密算法

出题组织：可信理论、技术与工程实验室　　接口专家： Liu Yamin/liuyamin3@huawei.com

技术背景

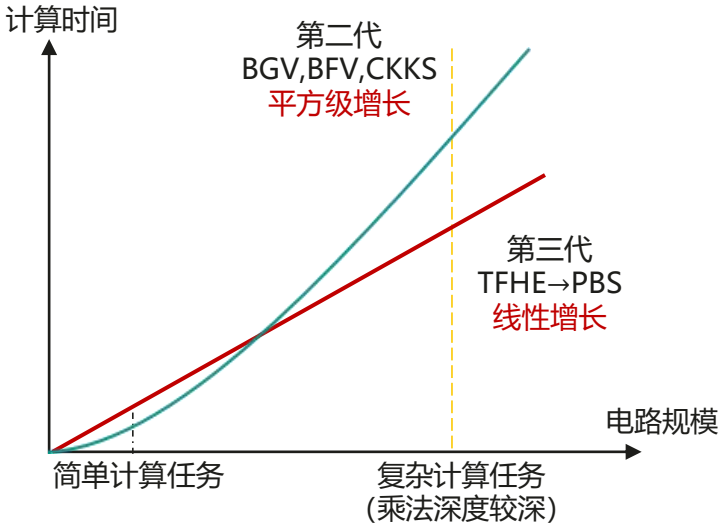
- 数据安全面临挑战，需要既能发挥数据价值、又能保护数据隐私的新型技术和基础设施；
- 全同态加密是隐私计算技术很重要的一个分支，可以实现数据可用、可控、不可见；
- 目前全同态加密的性能还不足，密文计算比明文慢几万到百万倍，但通过硬件加速的潜力较大；
- 第二代、第三代全同态加密方案都存在“偏科”现象，各有擅长和不擅长的计算类型，存在改进空间。

第二代：层次型

- **优势**：支持加法、乘法、批处理，密文/明文尺寸膨胀率较低
- **局限性**：不支持非线性运算，自举效率慢，必须一次性打包上万个元素
- **延迟因子**：在不同场景下差异巨大：100~10亿

第三代：高性能自举

- **优势**：支持加法、数乘、非线性运算
- **局限性**：不支持批处理、能支持的非线性函数尺寸有限、密文/和密钥尺寸太大，硬件加速时数据搬运开销过大
- **延迟因子**：在不同场景下差异很小：10万左右



方案	算术运算	非线性运算	批处理	自举效率	硬件加速
第二代	√	×	√	×	√
第三代	×	√	×	√	×

技术挑战

- **支持批处理可编程自举的全同态加密算法**：当前可编程自举方案只支持对单个消息的处理，如何利用批处理并行处理的特性进一步提升效率是可编程自举走向实用化的核心难题

当前结果

- **第三代TFHE/PBS技术可实现13ms一次自举，**自举的同时支持可编程非线性计算，但其不支持批处理式的自举
- **第二代自举总体效率低效且不支持可编程**，如CKKS方案批处理自举一次需要18s，但其批处理的特性使得均摊下来单个元素自举只需要5ms
- **最新研究尝试用三代的自举技术去加速二代自举**，但本质上在自举时仍需将二代密文提取成三代密文自举，时间复杂度高

技术诉求

设计高效支持批处理可编程自举的全同态加密算法，以同时具备二、三代全同态加密算法的优势，均摊（总时长/批处理数）效率达到10ms量级

方案	算术运算	非线性运算	批处理	自举效率	硬件加速
挑战目标	√	√	√	√	-（不作要求）

参考文献：

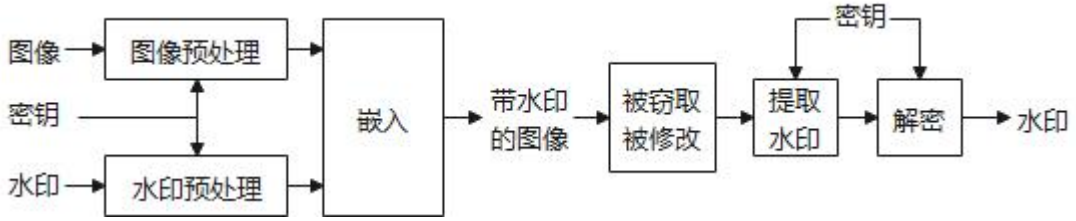
[1] TFHE: Fast Fully Homomorphic Encryption over the Torus.
[2] Programmable Bootstrapping Enables Efficient Homomorphic Inference of Deep Neural Networks
[3] Efficient Bootstrapping for Approximate Homomorphic Encryption with Non-Sparse Keys
[4] General Bootstrapping Approach for RLWE-based Homomorphic Encryption

难题2：面向版权确权场景的可公开鲁棒盲水印技术

出题组织：可信理论、技术与工程实验室　接口专家：周海波/zhouhaibo10@huawei.com

技术背景

水印技术是指在数字内容（如图像）中嵌入一些信息，在发生版权冲突时，能够通过其中的水印，完成数字内容版权归属的确定。



为了保证水印确权的权威性，需要水印算法公开可验证，并且能够抵抗不破坏图像价值前提下的攻击手段。对图像水印的攻击手段可以分为几何攻击和非几何攻击。几何攻击包括裁剪，旋转，缩放，平移等改变图像尺寸的修改方式，非几何攻击包括JPEG压缩，高斯模糊，滤波等调整像素值的修改方式。

技术挑战

- **水印算法的可公开性**：水印算法公开的情况下，敌手可以针对算法的薄弱环节进行攻击，水印的安全性面临挑战。
- **频域水印难以抵抗裁剪攻击**：频域变换相当于对像素矩阵进行一系列矩阵运算，在得到的矩阵中嵌入水印信息，再逆变换回像素矩阵，水印信息就分散的每一个像素中了。裁剪攻击破坏了矩阵运算的运算规则，无法从被修改的内容中提取水印。
- **空域抗裁剪水印鲁棒性低**：空域水印的嵌入通常比较简单，通过增加水印的冗余嵌入或者增加水印强度提高鲁棒性，但是在水印算法公开的前提下，即使增加冗余嵌入，攻击者可以很容易从内容中将水印信息去除。同时，增加水印的嵌入强度会导致对图像的观感产生影响。

当前结果

- **基于置乱的水印保护措施**：在水印嵌入之前首先根据密钥对图像进行置乱，相当于得到一个新的图像，在新的图像上进行水印嵌入，如果不知道置乱用的密钥就不能得到水印。但是这种方式对于可能改变置乱映射的任何攻击都不具有抵抗性，比如裁剪。
- **基于几何不变域的频域水印**：Fourier-Mellin（DFT+LPM）变换得到的是旋转，缩放，平移的不变域，即对图像做的这些操作，在这个域上只是坐标位置的变化。尽管在归一化坐标条件下，可以抵抗一定程度的按原图比例的裁剪，但是这种频域水印方法同样不能抵抗随意的裁剪攻击。
- **基于模板或周期嵌入**：因为裁剪破坏的是图像的一部分，剩余部分是完好的，因此部分水印算法在空域中周期性的嵌入相同的水印，保证在裁剪后还能够进行完整水印信息提取，但是存在安全性和不可见性问题。

技术诉求

- **安全的可公开方案**：版权确权场景，设计满足Kerckhoffs原则的水印方案。
- **抗裁剪能力**：保证非几何攻击的鲁棒性前提下，能够在25%以上的任意裁剪攻击下后，完成水印的提取。并且嵌入水印过程不能影响图像质量，PSNR(峰值信噪比)>40dB 或 SSIM(结构相似性) >95%。

参考文献：

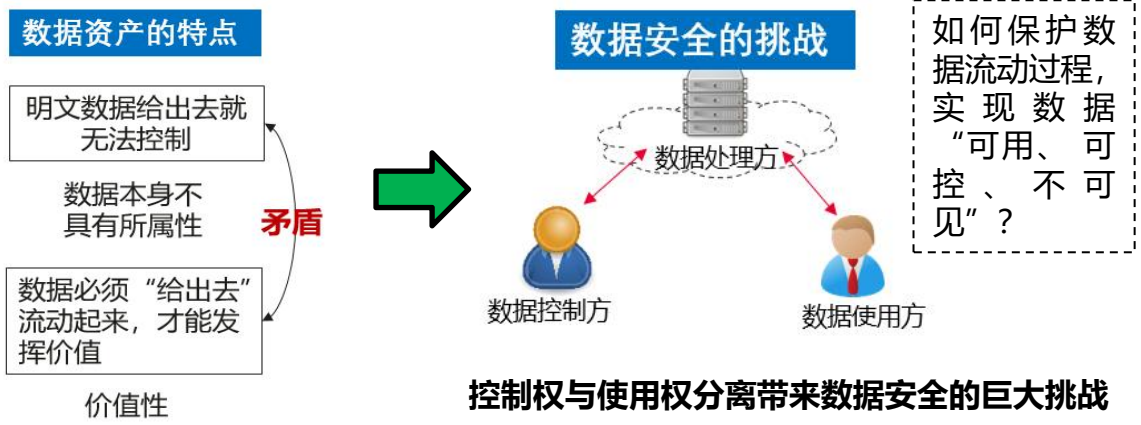
[1] Cryptanalysis and improvement of a chaos-based watermarking scheme, IEEE Access, 2019.
[2] Efficient general print-scanning resilient data hiding based on uniform log-polar mapping, 2010.
[3] Local geometric distortions resilient watermarking scheme based on symmetry. IEEE Transactions on Circuits and Systems for Video Technology, 2021.

难题3：如何高效解决安全多方计算合谋问题

出题组织：可信理论、技术与工程实验室　　接口专家：Yong Li/yong.li1@huawei.com

法规要求:在保障个人数据主权的前提下，促进数据有序流动

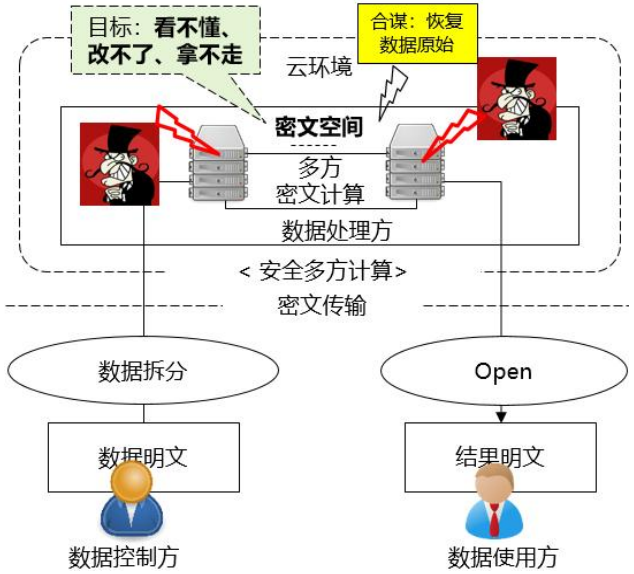
产业需求：数据**共享流通**使数据产生**价值**，我司ICT，CBU，CBG，IoV等产线有数据流通场景，研究一套完善的**安全数据流通解决方案**是企业**迫切需求**



控制权与使用权分离带来数据安全的巨大挑战

安全多方计算是保护数据流动的重要手段之一

如果计算方合谋数据将会被泄露：由于安全多方计算是一个分布式计算方式，各个计算方掌握隐私数据的随机分片，如果计算方合谋将会恢复原始数据，尤其针对云服务场景，难以自证清白，因此计算方法合谋是在MPC部署场景需要考虑的一个安全因素。



挑战：如何防止计算方合谋获取数据，实现自证清白，一直是业界一个难题

技术问题

- **性能损耗较大**：针对安全多方计算合谋问题，一般考虑的是主动安全模式（active security），即dishonest majority （t<n），t是corrupt parties，n是总的计算方个数。由于考虑到t个计算方被攻击者控制，同时还要保证MPC的正确性（correctness），公平性（fairness），输入的独立性（independence of the inputs）及输出可交付性（guaranteed output delivery）的特点，导致性能对比损耗较大（对比半诚实模型性能相差~10+倍）。

当前结果

- **主动安全算法协议**：基于主动安全的算法协议（以BMR和SPDZ为代表），但是并不能完全防止合谋攻击风险，并且性能损失很大；
- **依赖可信环境**：利用可信硬件来执行多方分布式计算，但是需要和可信硬件强绑定和依赖，同时由于可信硬件资源受限，性能也是瓶颈；

技术诉求

- **诉求和目标**： 可以基于加速硬件，MPC协议，以及TEE等技术，在保障主动安全性的前提下，利用硬件加速能力实现高效的多方计算功能，并能防止计算方合谋获取原始数据（这里我们假设计算方不可信）；
- **应用场景**： 密态AI训练和预测场景，针对数据集（cifar10，tiny imagenet，imagenet）的深层CNN网络（VGG16，ResNET 50, ResNET 101, ResNET 152;）
- **性能指标**：同等测试环境，对比半诚实（ semi-honest）的安全环境，性能损耗<50%
- **精度指标**：对比明文AI训练及预测，精度损耗<2%;
- **交付件**：算法及相应的形式化证明，PoC原型以及测试结果；

参考文献：

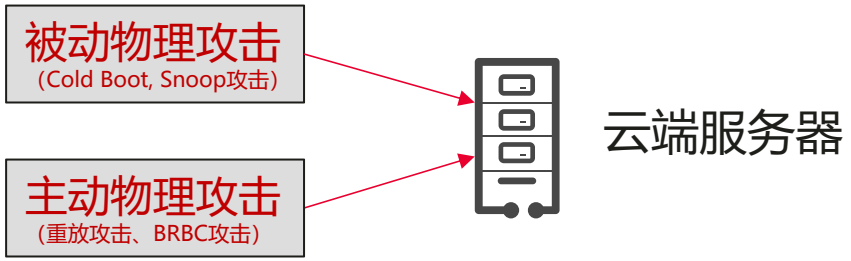
[1] The round complexity of secure protocols (extended abstract). In 22nd ACM STOC
[2] Multiparty computation from somewhat homomorphic encryption. CRYPTO 2012
[3] The price of active security in cryptographic protocols. EUROCRYPT 2020

难题4：低时延对称密码

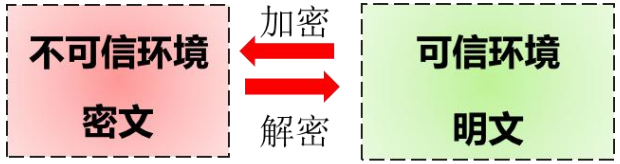
出题组织：可信理论、技术与工程实验室　　接口专家：Huang Tao/huangtao80@huawei.com

技术背景

云计算场景中，远程用户无法物理接触云端设备，需要保障用户数据运行时安全，防御可能的内存物理攻击



解决方案：使用内存加密防御对云上设备物理攻击



业界现有的内存加密方案，主要是基于标准的密码算法，如AES及其各种工作模式：

- Intel SGX：AES-CTR + 基于Carter-Wegman MAC的Merkle Tree，用于小规模内存加密和完整性保护
- Intel TDX: AES-XTS + SHA3-MAC，用于大规模（虚拟机级别）内存加密与完整性保护
- AMD SEV-SNP：AES-GCM，用于保护客户消息（Guest Message）的机密性与完整性

技术挑战

- **对于硬件场景专用的优化设计方法研究仍在相对早期的阶段，不够成熟：**当前低时延算法对于硬件场景专用的优化设计思路仍偏少，主要S盒、线性层等组件的优化，如何充分利用Permutation运算硬件友好的性质或者是基于LRX结构设计低时延算法仍有待探索
- **针对具体低时延场景进行密码算法设计优化：**在cache line加密的场景，单次读写数据量固定长度，不同于通用的算法需要考虑支持各种长度的消息加解密。如何在算法设计上充分利用场景的特殊性质值得进一步探索。

当前结果

- **当前标准密码算法通常是面向通用场景设计，硬件实现时延性能不够理想：**NIST标准加密算法AES使用8-bit的S-盒设计作为非线性运算，会引入较大的时延。SM4的S盒设计存在类似的问题。AES被应用在内存加密时，在数据读写阶段性能会慢5~9倍[1]。
- **低时延密码算法类型需要扩展：**当前低时延密码算法中（可调）分组密码分组长度主要是64/128-bit，缺乏分组为256-bit的算法。(Intel 的K-Cipher[2]可以支持灵活的分组长度，但最近安全性分析受到挑战[3]）。此外，目前仍缺乏专门设计的低时延认证加密算法。

技术诉求

设计新的低时延认证加密算法

应用场景：对于512/1024-bit cache line的加密和完整性保护，单次数据读写带宽为128-bit（对应512-bit cache line）或256-bit（对应1024-bit cache line）

安全要求：

- IV/Nonce重用的情况下仍具有认证安全性（不要求额外的空间存储IV）
- 加密安全性: 128-bit 及以上;
- 完整性保护安全性：支持32-bit 及 64-bit 两种，可选支持更高

时延要求：

- 每128/256-bit（对应512/1024-bit cache line），加解密时延 2~3 Cycle @ 3GHz

参考文献：

[1] Mofrad, Saeid & Zhang, Fengwei & Lu, Shiyong & Shi, Weidong. (2018). **A comparison study of intel SGX and AMD memory encryption technology**. HASP '18: Proceedings of the 7th International Workshop on Hardware and Architectural Support for Security and Privacy.
[2] Kounavis, Michael, Sergej Deutsch, Santosh Ghosh, and David Durham. "**K-cipher: A low latency, bit length parameterizable cipher**." In 2020 IEEE Symposium on Computers and Communications (ISCC), pp. 1-7. IEEE, 2020.
[3] Mahzoun, Mohammad, Liliya Kraleva, Raluca Posteuca, and Tomer Ashur. "**Differential Cryptanalysis of K-Cipher**." In 2022 IEEE Symposium on Computers and Communications (ISCC), pp. 1-7. IEEE, 2022.

难题5：大规模AI模型安全验证关键技术

出题组织：可信理论、技术与工程实验室　　接口专家：Fang Chengfang (fang.chengfang@huawei.com)

技术背景

问题：AI技术能够极大的提升各行业效率，然而AI的安全风险影响着其在关键业务中的广泛部署。其中最主要风险是AI的对抗攻击，即攻击者能够对AI模型输入进行少量扰动就能控制其输出。目前业界通过经验方法来测试业务模型对某些已知攻击的抵抗能力，这些经验方法存在局限性，它们既无法验证未知攻击，也无法覆盖已知攻击的全部扰动，难以全面保证安全性，因此，通过AI模型安全验证提供可证明安全是当前重要问题。

- 关键业务需要对AI模型安全进行严格要求，对模型的可证明安全验证能够消除业务中一个重要的不可控因素，是保障系统安全的重要手段；
- 然而AI模型规模越来越大且与环境之间有大量交互，存在状态空间爆炸问题，验证属于NP困难问题；
- 深度学习AI系统的不可解释性与形式化的可证明性存在矛盾

参考文献：
[1] Algorithms for Verifying Deep Neural Networks. arxiv.1903.06758
[2] The Second International Verification of Neural Networks Competition (VNN-COMP 2021): Summary and Results. arxiv.2109.00498
[3] SoK: Certified Robustness for Deep Neural Networks. SP2023
[4] Principles for Verified Artificial Intelligence, DAC2017

技术挑战

- 如何将AI模型的计算过程转换为可验证、可扩展的表达式，提升大规模模型静态验证的效率和可验证边界，对攻击形式化分析和刻画，并证明模型有效防御对抗攻击
- 如何引入形式化验证方法来对实现对模型的设计和训练，实现模型针对攻击的可验证鲁棒性

当前结果

- 在图像分类领域，部分可验证鲁棒性的模型参数到达百万级； 在目标检测领域，尚未有相关的工作成果
- 可验证鲁棒性：通过有效的鲁棒性训练对AI模型实施强化， 在ImageNet数据集上，smoothed DNN在扰动半径（L2范数）为2的条件下，可验证准确率不超过40%

技术诉求

- 验证效率、模型参数规模和可验证半径的大幅提升：
  实现AI模型安全性验证的效率、规模和可验证边界的大幅提升
- 验证领域突破：
  实现目标检测和人脸识别等模型的鲁棒性验证的突破
- 整体验证框架：
  构建并完善AI系统验证框架，包含AI系统各关键要素的整体验证，增强AI系统安全评估的理论基础，保证实际应用中的系统抗攻击的能力

Thank you.

Bring digital to every person, home and organization for a fully connected, intelligent world.