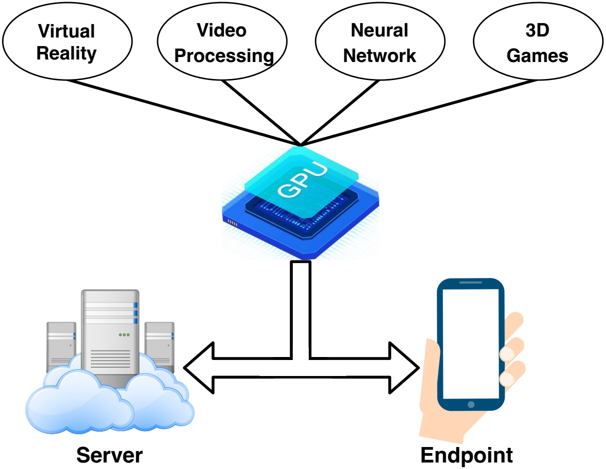
StrongBox: A GPU TEE on Arm Endpoints

Yunjie Deng*, Chenxu Wang*, Shunchang Yu, Shiqing Liu, Zhenyu Ning, Kevin Leach, Jin Li, Shoumeng Yan, Zhengyu He, Jiannong Cao, Fengwei Zhang ✉

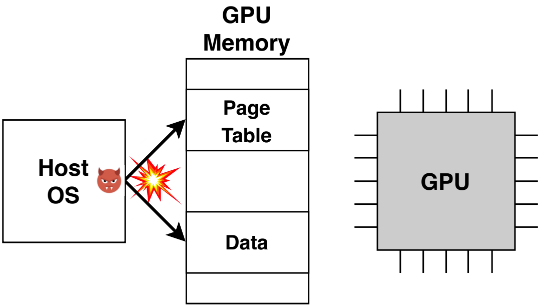
南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY
THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學
湖南大學 HUNAN UNIVERSITY
GUANGZHOU UNIVERSITY
VANDERBILT UNIVERSITY
ANT GROUP

Wide Application of GPU



GPU Security

- Varied **sensitive data** are processed on GPU
  - face, fingerprints, voice ...
- The vulnerable host OS severely threats GPU computing
  - Privileged attackers can directly access the data, or
  - Break the page table isolation between GPU computation

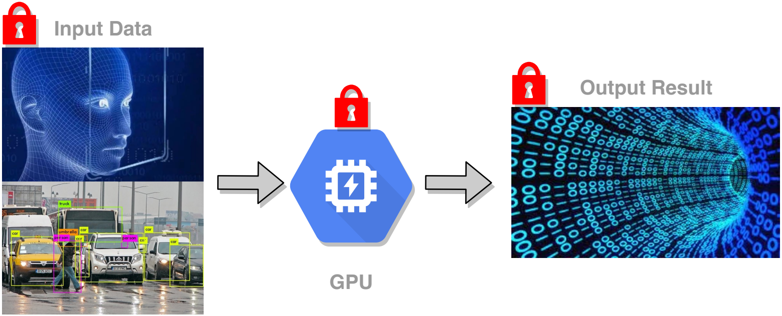


Trusted Execution Environments

- Processor IP developers introduce hardware-assisted **T**rusted **E**xecution **E**nvironment (TEE) for secure data storage and computation
  - Arm TrustZone
  - Intel Software Guard Extensions (SGX)
  - AMD Secure Encrypted Virtualization (SEV)

GPU TEEs

- Secure data transmission between OS and GPU
- Isolate GPU memory and GPU computation

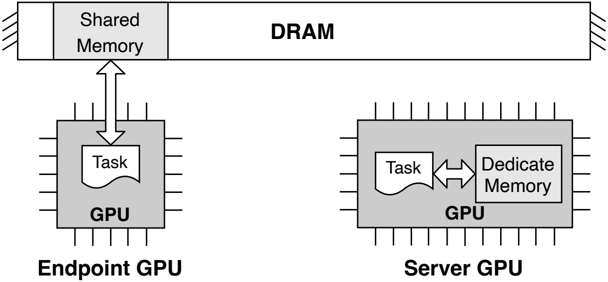


GPU Trusted Execution Environments

- TEEs have participated in secure GPU computing
  - **Graviton**: Trusted Execution Environments on GPUs (OSDI'18)
  - **HIX**: Heterogeneous isolated execution for commodity gpus (ASPLOS'19)
  - **HETEE**: Enabling Rack-scale Confidential Computing using Heterogeneous Trusted Execution Environment (S&P'20)
  - **LITE**: A Low-Cost Practical Inter-Operable GPU TEE (ICS'22)
  - **Secdeep** (IoTDI'21): Secure and Performant On-device Deep Learning Inference Framework for Mobile and IoT Devices
  - ...

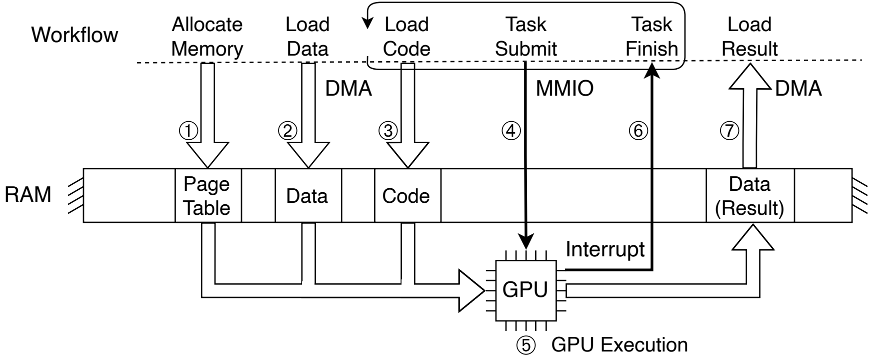
Challenges of Adapting Existing Works to Arm Endpoints

- Architecture
  - CPU Architecture: Intel vs. Arm
  - GPU Architecture: Dedicated-memory GPU vs. Shared-memory GPU



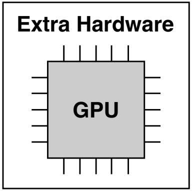
Challenges of Adapting Existing Works to Arm Endpoints

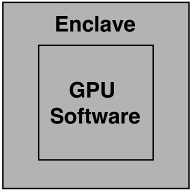
- Architecture
  - A typical workflow on Arm endpoint GPUs



Challenges of Adapting Existing Works to Arm Endpoints

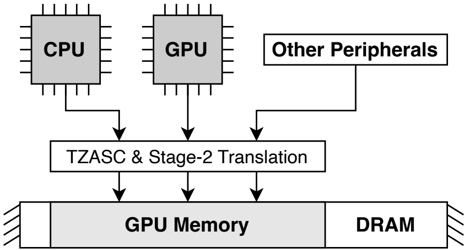
- Compatibility
  - Hardware modification on GPU chips or system architecture
- TCB size
  - Directly porting the vulnerable GPU software stacks into enclave





StrongBox Overview

- Architecture
  - Arm hardware features
    - ★ TrustZone Address Space Controller (TZASC)
    - ★ Stage-2 translation
  - Shared-memory GPU
    - ★ Reserve a memory region for sensitive GPU tasks
    - ★ Protect GPU memory by TZASC and Stage-2 translation



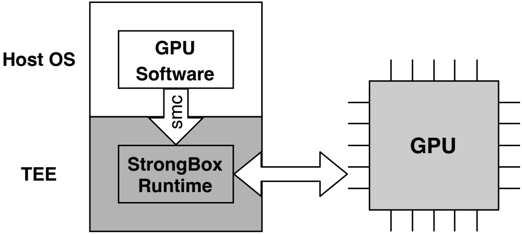
StrongBox Overview: Threat Model and Assumptions

- Compromised GPU software stacks
  - GPU runtime
  - GPU driver
  - Other peripheral drivers
  - System OS
- No hypervisor on Arm endpoints
- *Trusted secure OS and applications
- Out of scope: side-channel attacks, physical attacks, Denial-of-Service

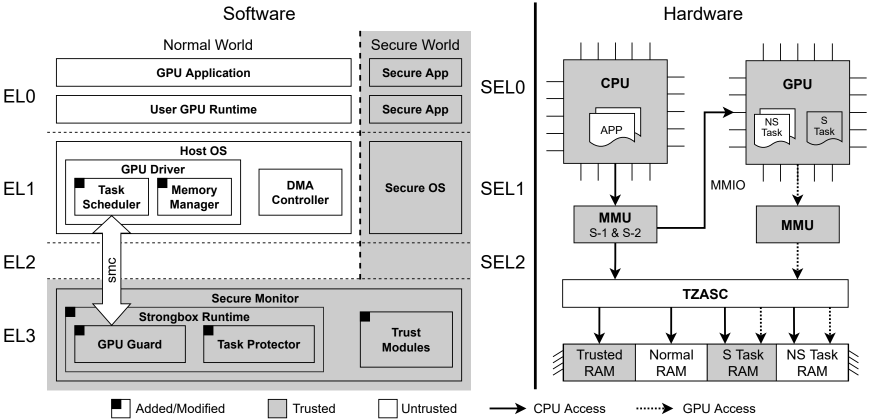
*: Addressed in future works

StrongBox Overview

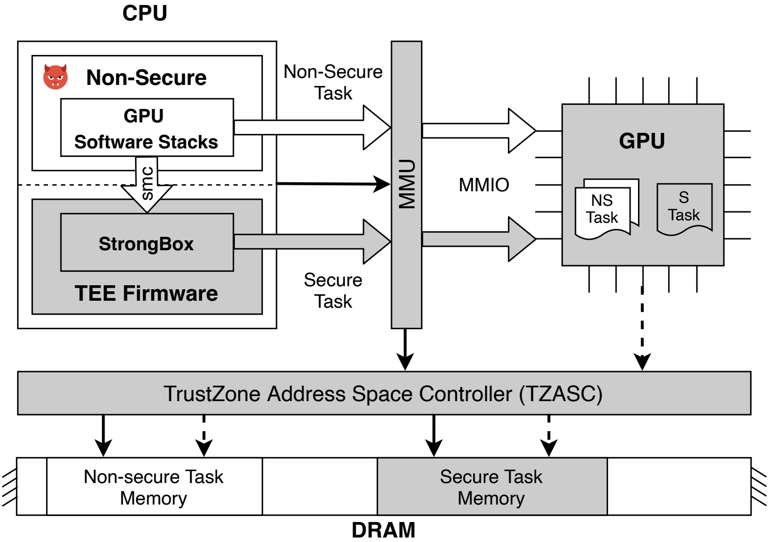
- High Compatibility
  - No hardware modification
- Minimal TCB
  - Reuse GPU software to fulfill functionality
  - Deploy lightweight StrongBox runtime to perform security check for sensitive computation tasks



StrongBox Overview



StrongBox Overview: Secure Tasks and Non-secure Tasks



Design Details

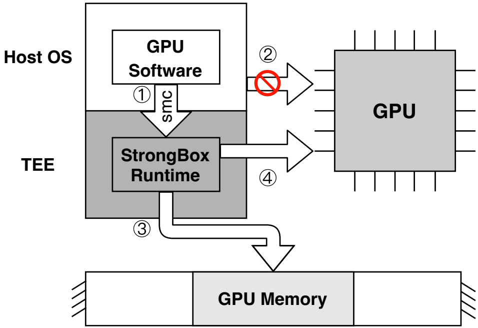
- Isolated Execution Environment
  - Prohibit the attackers access GPU and GPU memory when executing sensitive tasks
- Dynamic and fine-grained GPU memory access control
  - Prohibit the attackers access scattered sensitive data and code
- Reduce performance overhead
  - Optimize the protection overhead on multi-tasks GPU applications

Isolated Execution Environment

- Restrict two modes of data access
  - Host OS to GPU
  - Host OS to shared memory
- Approach
  - Route the control from GPU driver to StrongBox runtime inside TrustZone
  - Manage the access to the shared memory
- Other requirements
  - Small TCB
  - No hardware modification

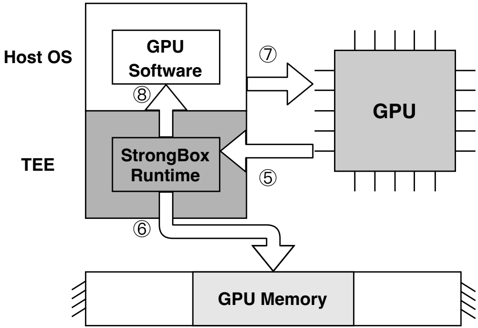
Isolated Execution Environment: Submission

- ①: Route control to StrongBox runtime
- ②: Forbid the Host OS to access GPU
- ③: Protect the sensitive data and code
- ④: Submit computation task to GPU



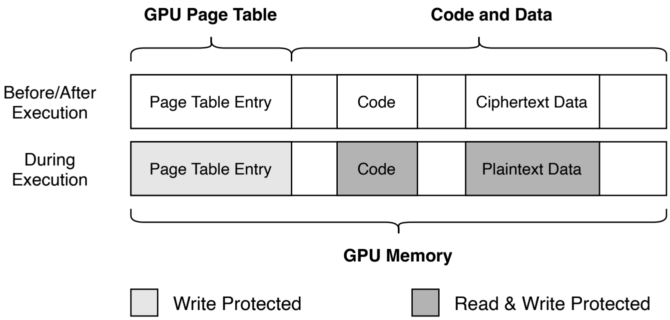
Isolated Execution Environment: Termination

- ⑤: Capture task finish interrupt
- ⑥: Restore the access permission to sensitive data and code
- ⑦: Allow Host OS to access GPU
- ⑧: Route the control to GPU driver

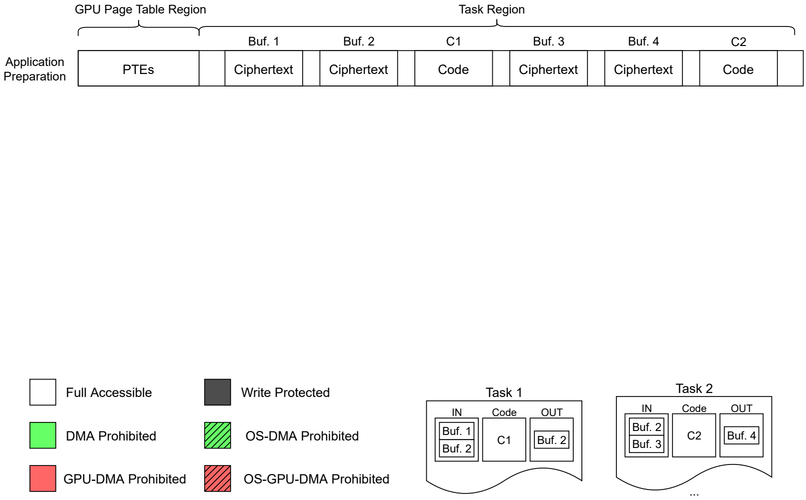


Dynamic and Fine-grained Memory Access Control

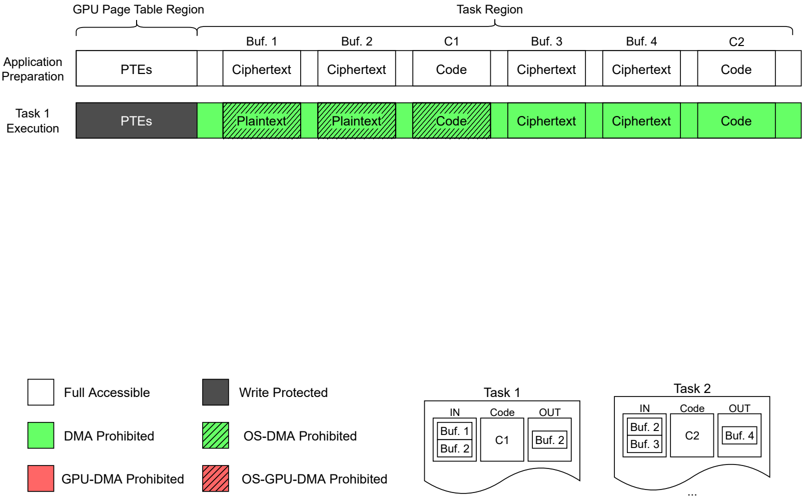
- Dynamic access control
  - Apply the protection to different GPU memory content
- Fine-grained protection
  - Combine Stage-2 translation (page-grained) and TZASC (slot-grained)
  - Prohibit the attackers access scattered sensitive data and code
  - Allow the GPU driver access the remaining non-sensitive region to fulfill functionality



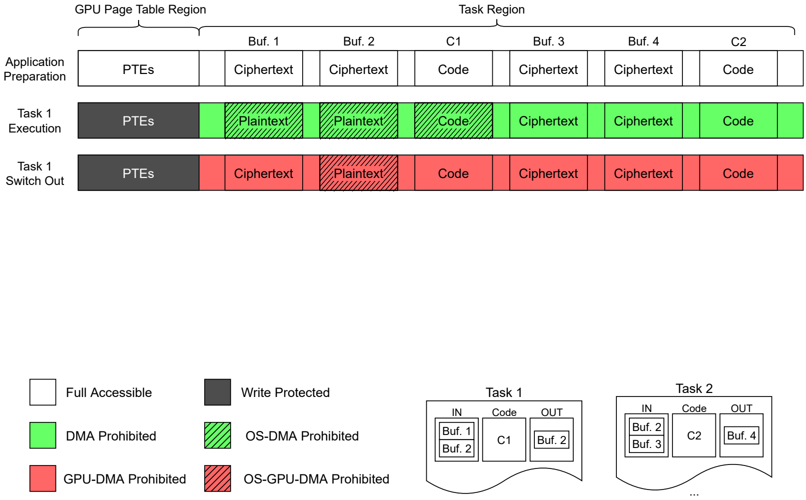
Example of Memory Access Control



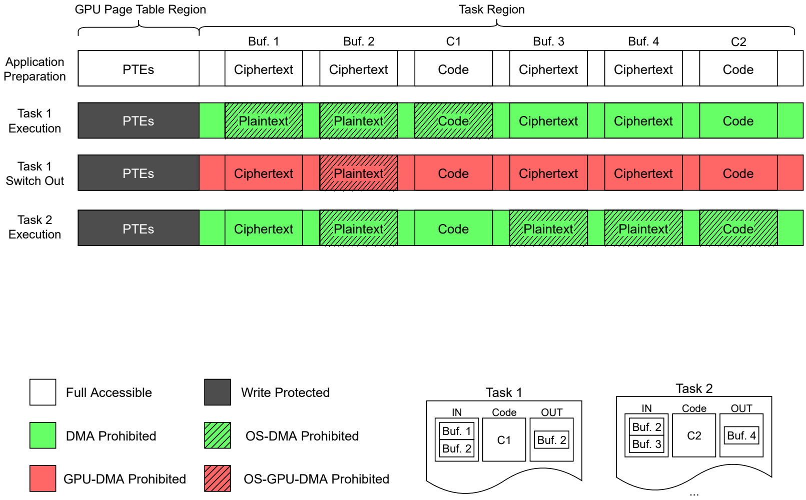
Example of Memory Access Control



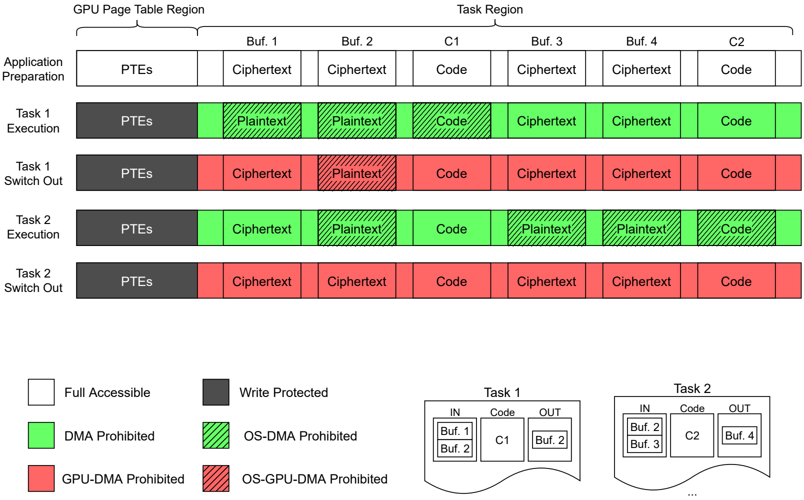
Example of Memory Access Control



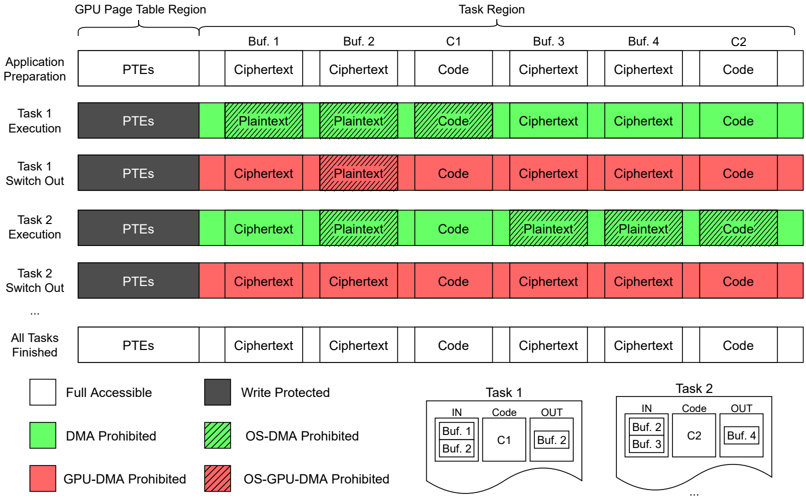
Example of Memory Access Control



Example of Memory Access Control

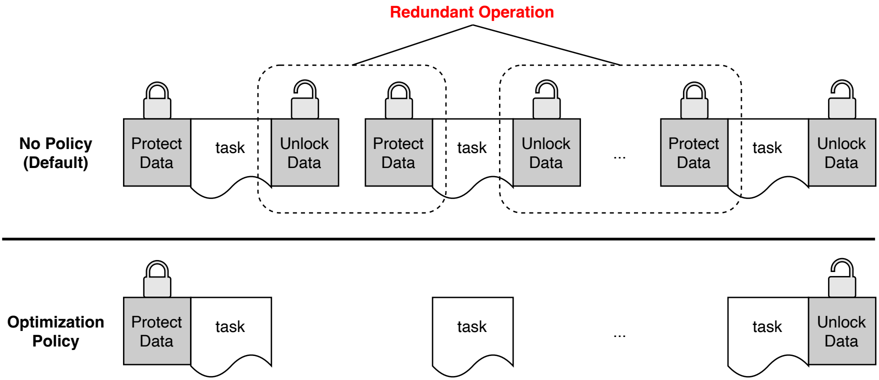


Example of Memory Access Control



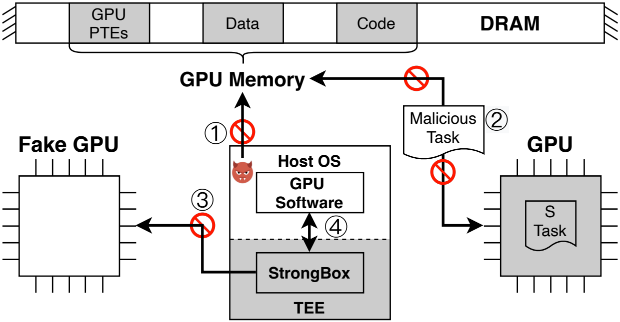
Reduce Performance Overhead

- In multi-task applications, the output of one task can be used as the input of the next task
- Eliminate redundant operations to reduce performance overhead



Evaluation: Security Analysis

- ①: Directly access the sensitive data and code ×
- ②: Attack with malicious tasks ×
- ③: Attack with fake GPU ×
- ④: Attack with compromised GPU software ×



Evaluation: Rodinia Benchmark

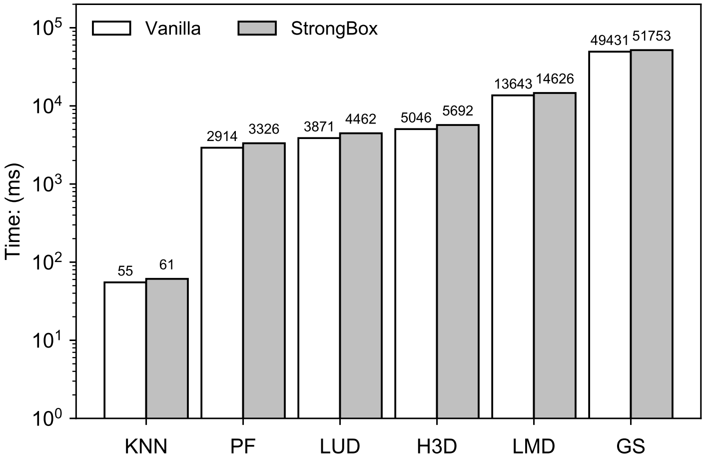


Figure: Evaluation on Rodinia benchmarks (overhead 4.70% - 15.26%).

Evaluation: Optimization

- Optimization on redundant protection

Benchmark		No Optimization		StrongBox	
		TProtect	Total	TProtect	Total
Single Task	KNN	7.31 (11.55%)	63.30	4.86 (7.95%)	61.10
	LMD	1,227.88 (8.27%)	14,854.08	977.46 (6.68%)	14,626.98
Multi Task	PF	3,495.99 (54.50%)	6,414.31	399.48 (12.01%)	3,326.04
	LUD	97,179.42 (95.24%)	102,032.57	338.10 (7.58%)	4,462.57
	H3D	196,457.42 (96.87%)	202,797.82	332.82 (5.85%)	5,692.59
	GS	2,149,460.48 (97.40%)	2,206,881.00	694.52 (1.34%)	51,753.57

Conclusion on StrongBox

- First GPU TEE on Arm Endpoints
  - Ensure secure and isolated computation on Arm endpoint GPUs
  - Entail a minimal TCB to reduce potential attack surface
  - Maintain high compatibility
- Source code
  - https://github.com/Compass-All/CCS22-StrongBox

Thank You