SeDA: Secure and Efficient DNN Accelerators with Hardware/Software Synergy

Wei Xuan, Zhongrui Wang, Lang Feng, Ning Lin, Zihao Xuan, Rongliang Fu, Tsung-Yi Ho, Yuzhong Jiao, Luhong Liang

ACCESS AI Chip Center for Emerging Smart Systems 智能晶片與系統研發中心

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

1. Background

Secure is important for DNN-based Applications:

Attacked auto-vehicles can disrupt critical functions and pose safety risks

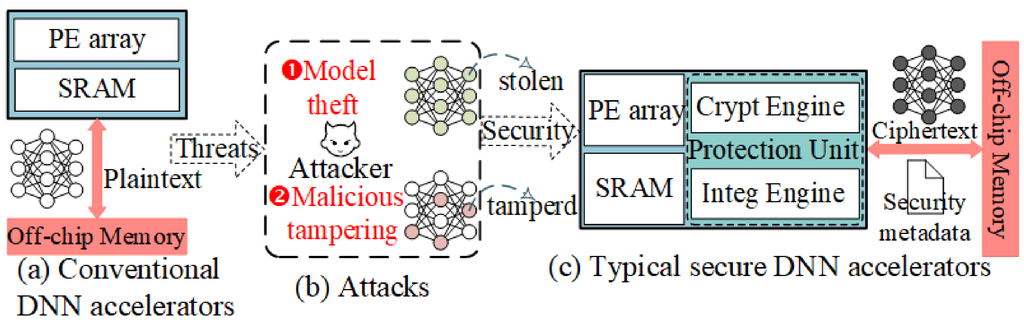
Healthcare cybersecurity endanger patient data, medical devices, and operations

Maliciously hacked AI robots may harm human interests

Malicious data alterations undermine financial market fairness and investor trust

Typical Secure DNN Accelerators:

Threat Model: unsecure off-chip memory and communication buses.
Memory Protection Mechanism: authentication encryption.



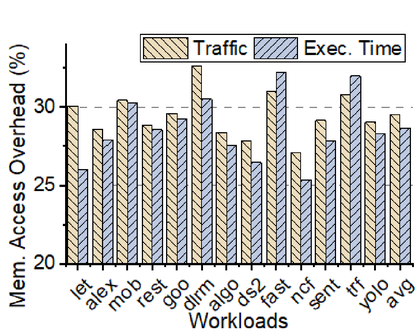
2. Motivations

Costly Off-Chip Memory Access Overhead for Integrity Verification

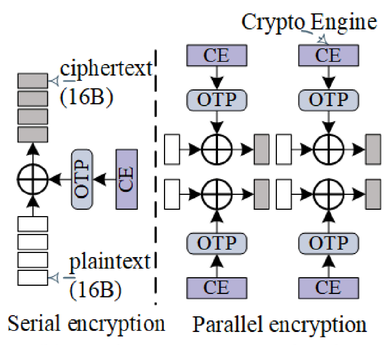
- Merkle Tree (MT) with its variants incurs extra off-chip data access, to defend replay attacks.
- Different cross-layer tiling patterns introduce redundant authentication data read and computation.
- XOR-based MAC scheme may lead to re-permutation attack.

High Hardware Overhead for Confidentiality Protection

- Serial encryption engine provides limited bandwidth.
- Parallel approach incurs amount of hardware overhead.
- Shared one-time pad (OTP) can be vulnerable to the single-element collision attacks.



(a) Memory access overhead

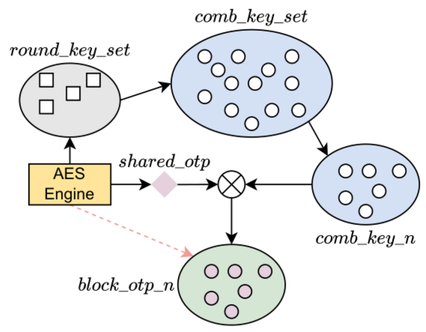


(b) Hardware crypto overhead

3. Proposed SeDA

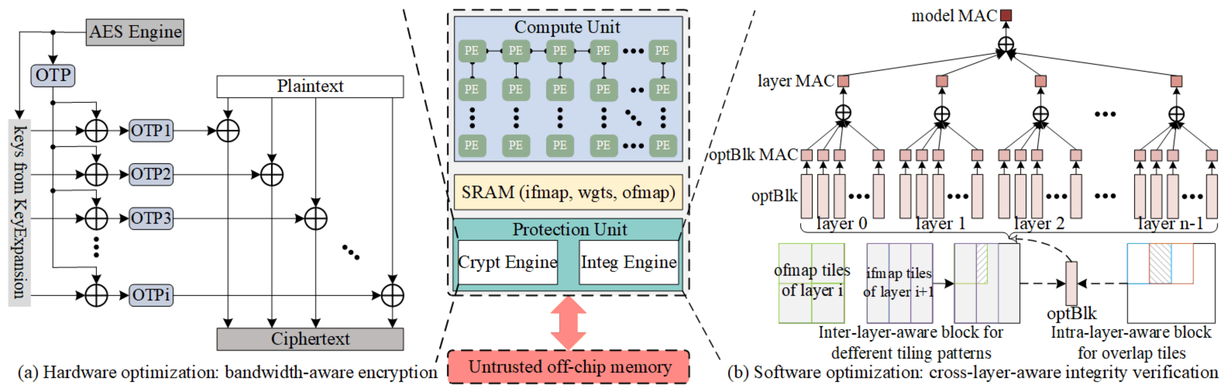
Bandwidth-Aware Encryption Scheme:

- Utilize KeyExpansion of AES to generate multiple unique OTPs
- It can en/decrypt all 128-bit sub-blocks by running one time of AES engine
- Improve the memory protection bandwidth to meet requirements



Multi-level Integrity Verification:

- Explore optimal block (optBlk) considering diverse cross-layer tiling patterns;
- Calculate message authentication code of each optBlk (optBlk MAC);
- Aggregate layer MAC (model MAC) by XORing all optBlk MACs (layer MACs);

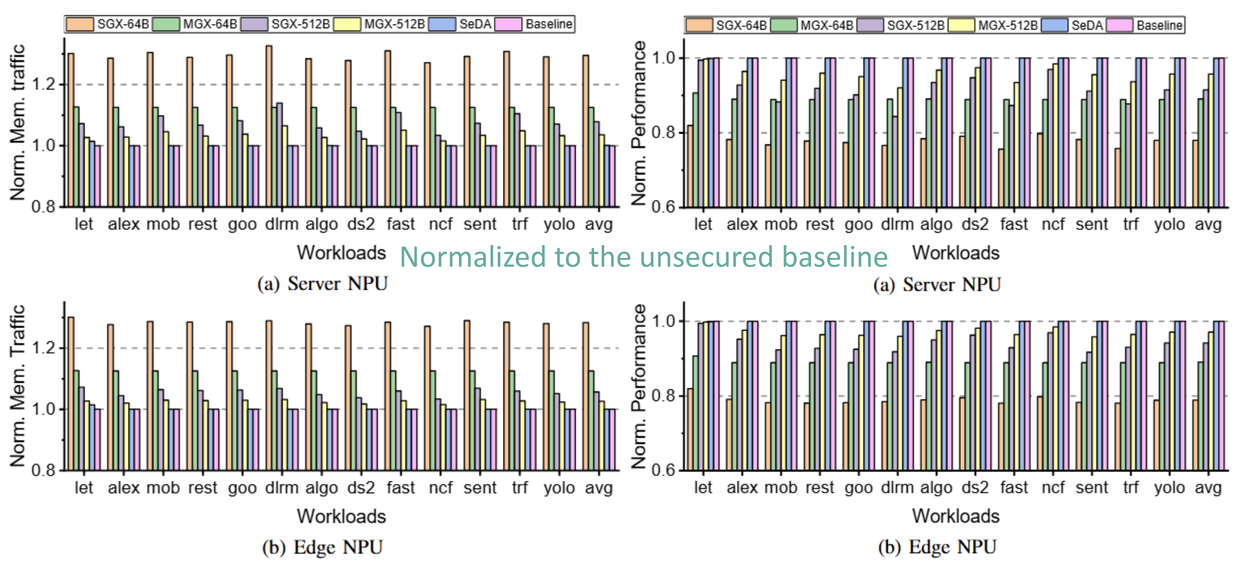


(a) Hardware optimization: bandwidth-aware encryption

(b) Software optimization: cross-layer-aware integrity verification

4. Evaluation

Through systematic experiments, SeDA is shown to mitigate performance overhead by over 12% for both server and edge NPUs, through the implementation of bandwidth-aware encryption and multi-level integrity verification framework.



Protection Scheme	Encryption Granularity	Integrity Granularity	Off-chip Memory Access	DNN Tiling Pattern	Encryption Scalability
SGX-64B	16B	64B	MAC,VN,IT	✗	✗
SGX-512B	16B	512B	MAC,VN,IT	✗	✗
MGX-64B	16B	64B	MAC	✗	✗
MGX-512B	16B	512B	MAC	✗	✗
SeDA	bandwidth-aware	multi-level	minimal to no cost	✓	✓

Acknowledgment: This research was partially supported by the National Key R&D Program of China (Grant No. 2022YFB3608300). It was supported in part by ACCESS – AI Chip Center for Emerging Smart Systems, sponsored by the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government. It was also supported in part by Shenzhen Science and Technology Innovation Commission (Grant No. SGDX20220530111405040), Beijing Natural Science Foundation (Grant No. Z210006), Hong Kong Research Grant Council (Grant Nos. 27209621, 17205922, 17212923), the National Natural Science Foundation of China under Grant 62204111.