Tunable Memory Protection for Secure Neural Processing Units

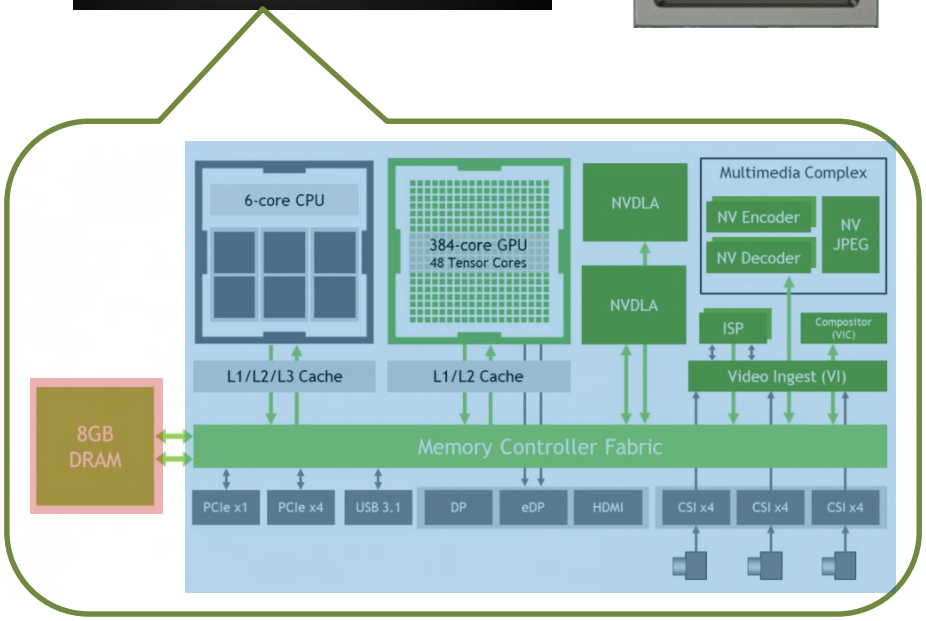
Sunho Lee, Seonjin Na, Jungwoo Kim,
Jongse Park, and Jaehyuk Huh

KAIST School of Computing

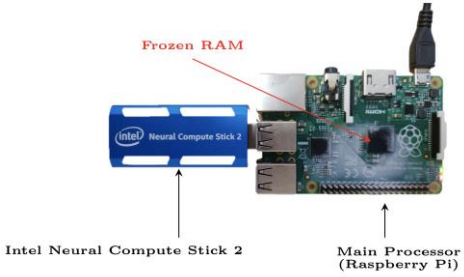
CASYS Computer Architecture and Systems Lab

Why Memory Protection for Integrated-NPU?

Confidentiality and Integrity should be protected



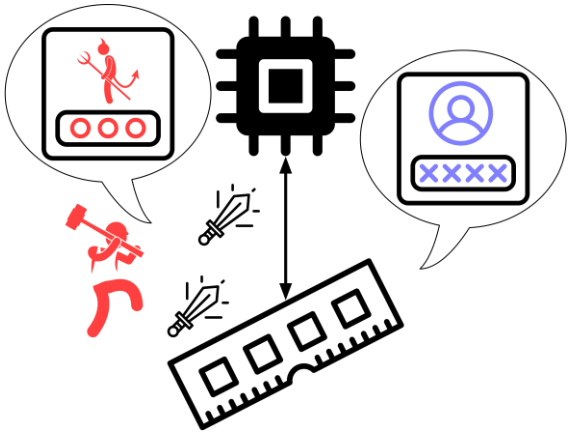
Confidentiality Attack [1]



(a) NCS on Main Processor

(b) Air Duster

Integrity Attack [2]

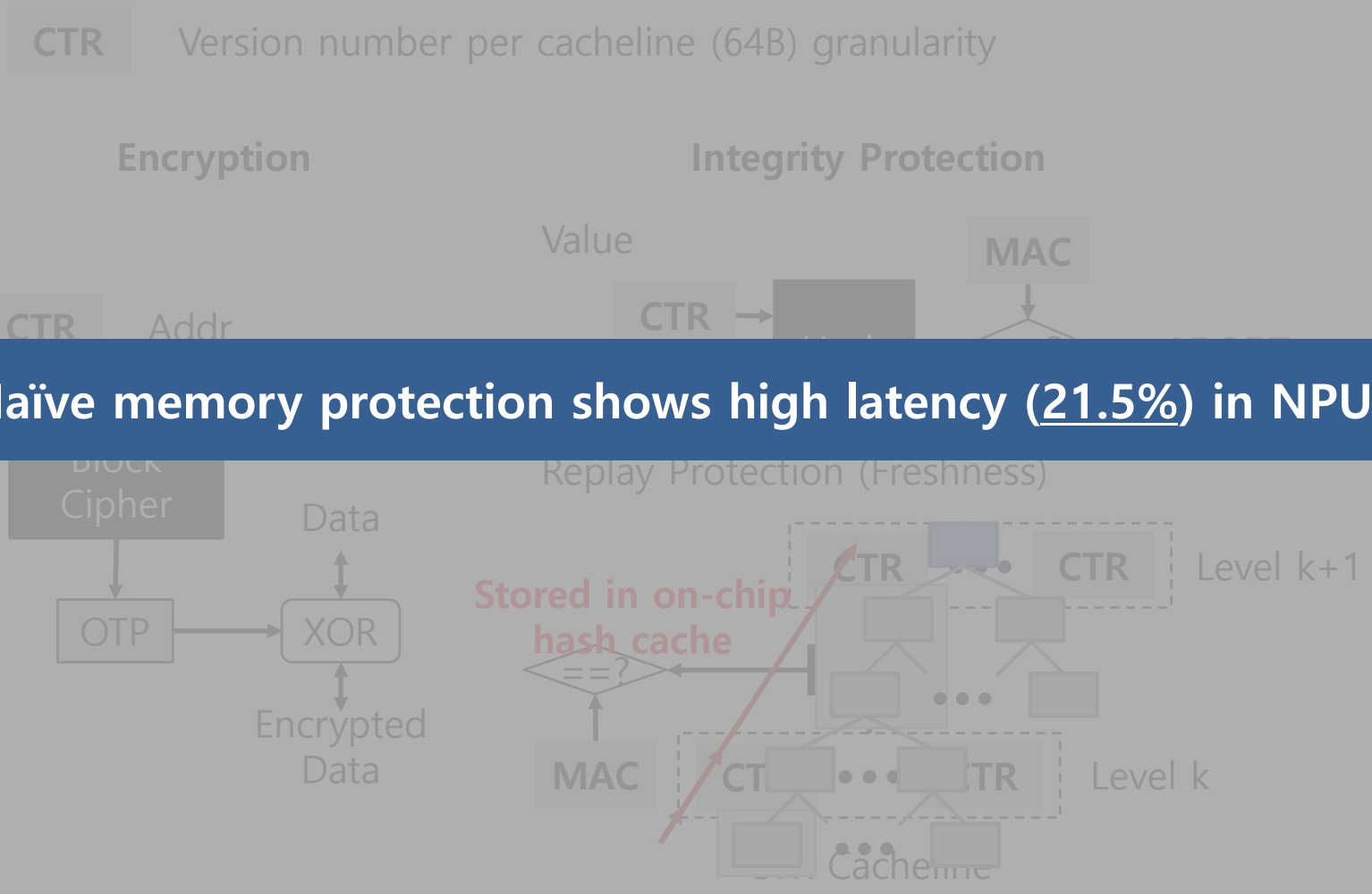


[1] DeepFreeze: Cold Boot Attacks and High Fidelity Model Recovery on Commercial EdgeML Device. (ICCAD, 2021)

[2] Bit-flip attack: Crushing neural network with progressive bit search. (ICCV, 2019)

Traditional CPU-Adopted Memory Protection

Counter-mode encryption and integrity protection

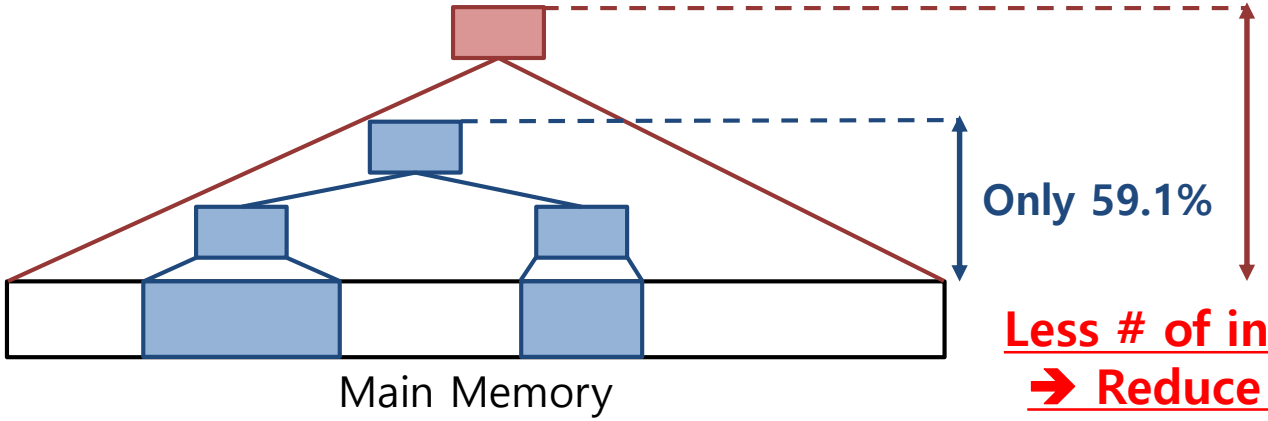


Naïve memory protection shows high latency (21.5%) in NPU

Our Contributions

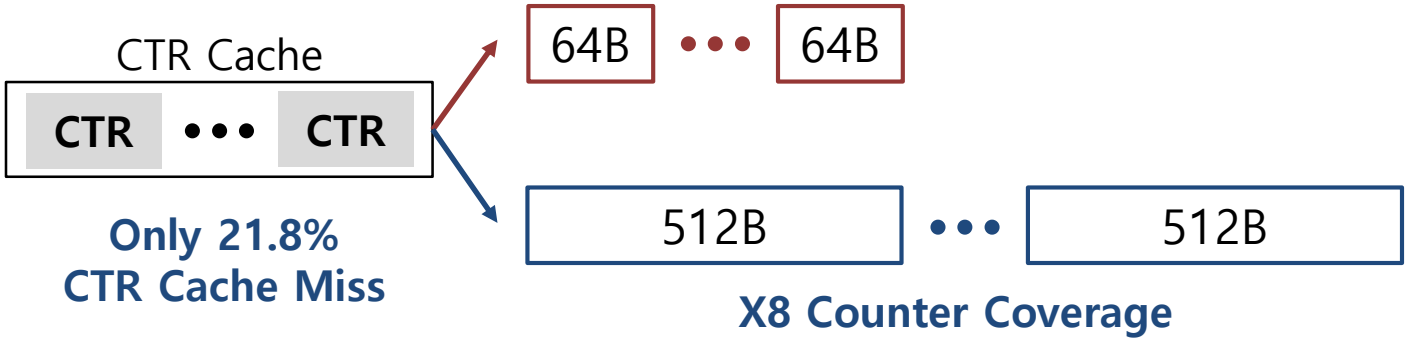
Two major optimizations: Virtual tree and multi-granular counter

Virtual Tree: Only NPU Context



Less # of intermediate nodes
➔ Reduce hash cache miss

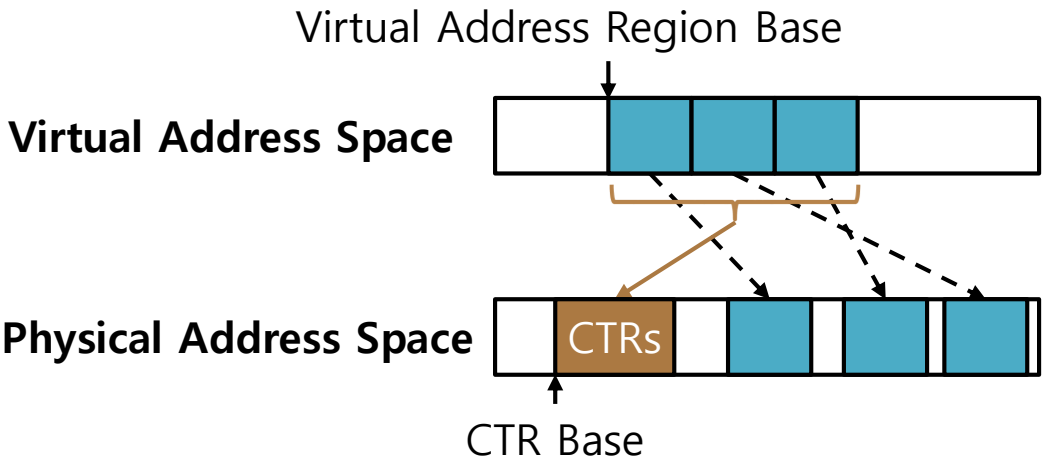
Multi-granular Counter: Leveraging Large-granular NPU Access Pattern

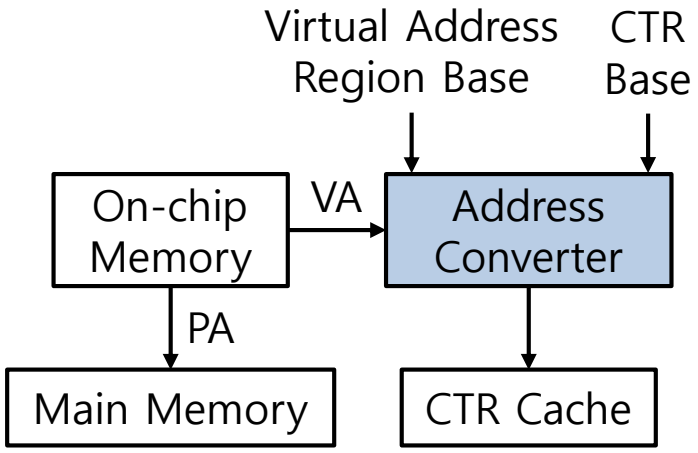


Less # of leaf nodes ➔ Reduce counter cache miss

Virtual Integrity Tree

Mechanism to find counter address



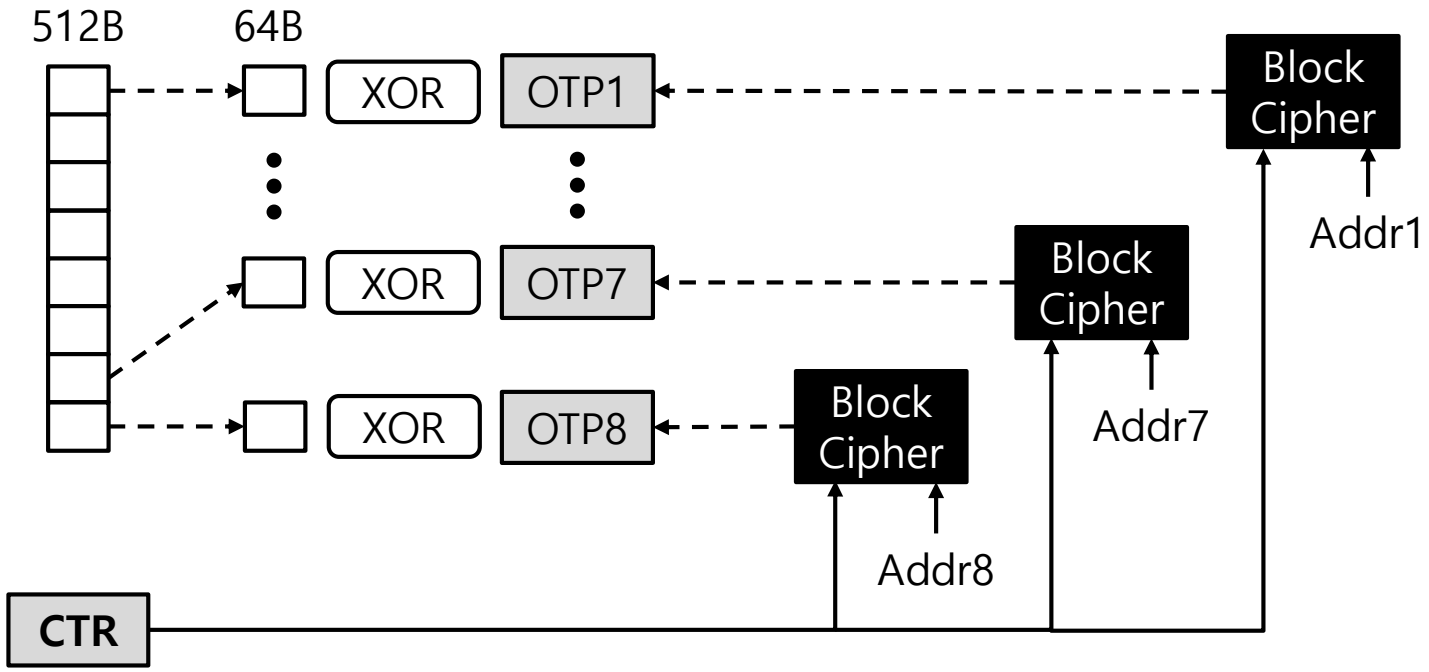


$$CTR\ Address = CTR\ Base + \frac{VA - VA\ Region\ Base}{Granularity}$$

$$Parent\ CTR\ Address = f(CTR\ Address, Tree\ Arity)$$

Multi-granular Counter

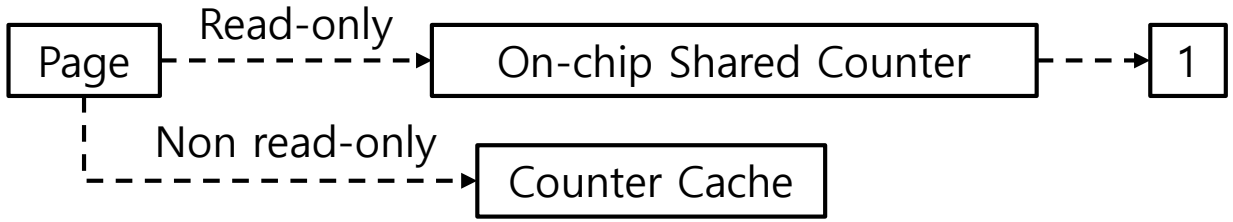
Hardware to support multi-granular counter



Other Optimizations

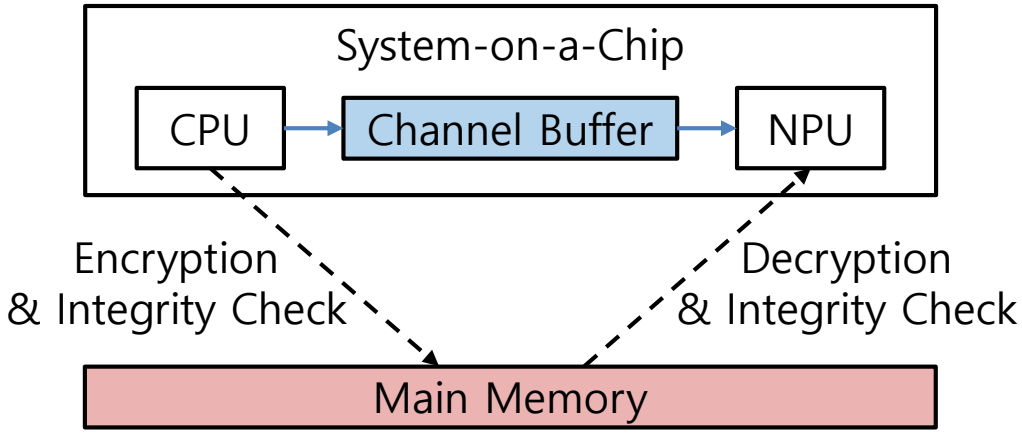
Shared counter for read-only region and direct communication path

Read-only Optimization [1]



Reduce counter cache traffic

Direct Communication Path



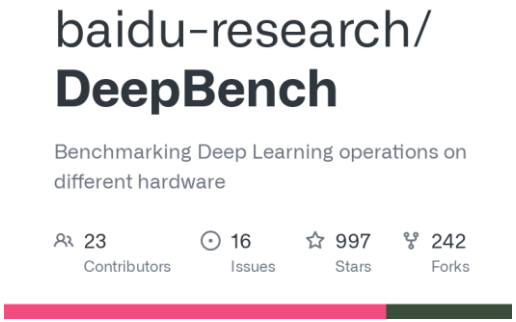
Skip additional memory protection

Why possible?
→ Streaming Nature of NPU Workload

[1] Common Counters: Compressed Encryption Counters for Secure GPU Memory. (HPCA, 2021)

Evaluation Environment

Cycle-level simulation modified from SCALE-Sim [1]

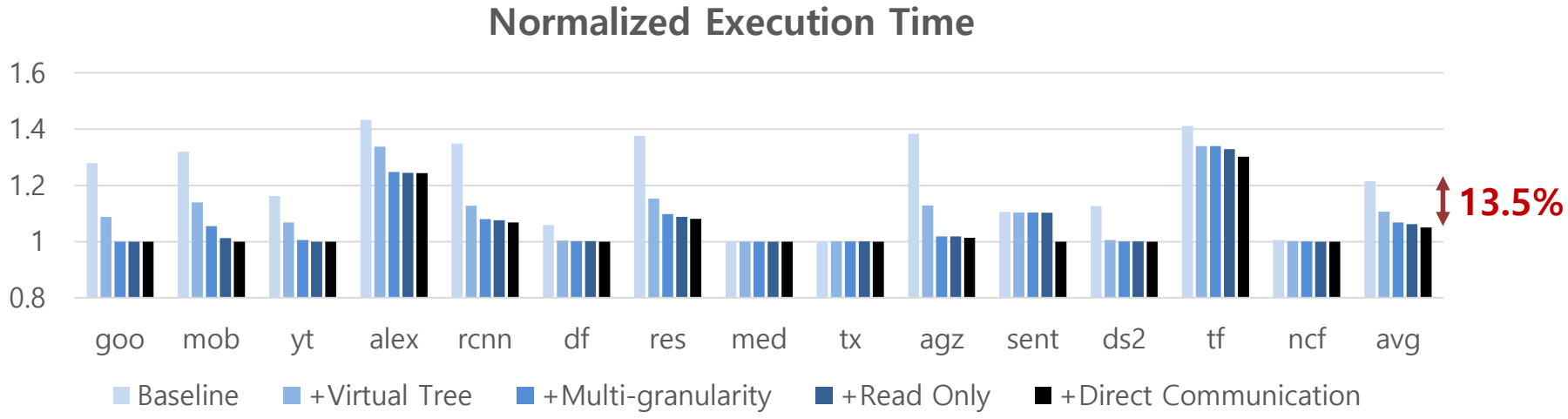


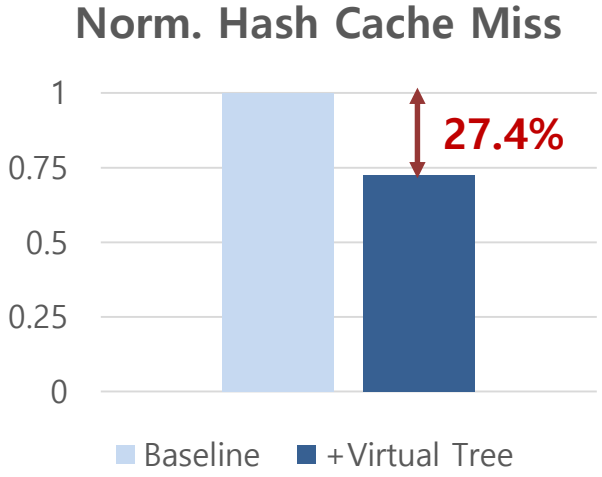
	NVDLA Configuration
PE	16 x 16
Bandwidth	5 GB/s (8 channels)
Frequency	1 GHz (both processor/memory)
SPM	192KB in total
Counter Cache	512B
Hash Cache	2KB
Precision	Int8

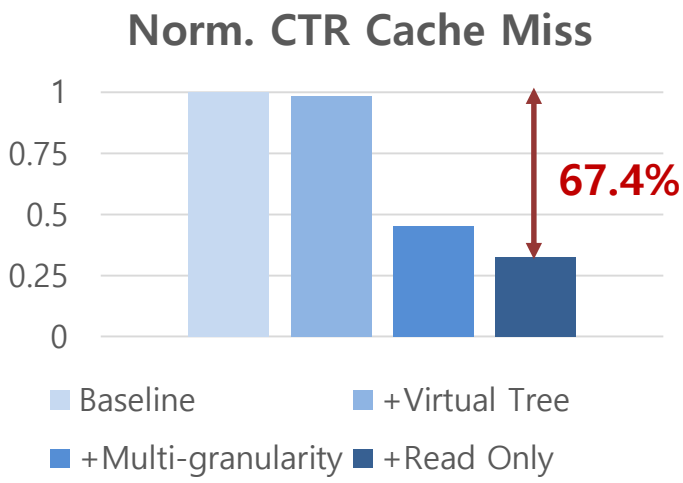
[1] A systematic methodology for characterizing scalability of DNN accelerators using SCALE-Sim (ISPASS 2020)

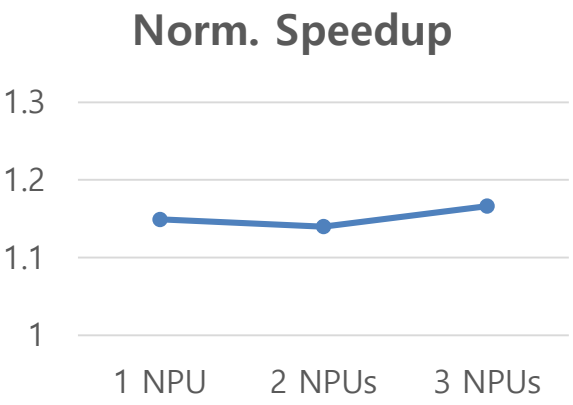
Evaluation Result

Performance improves by four optimizations









Summary

Tunable Memory Protection for Secure Neural Processing Units

- **Result**
  - Secure & efficient memory protection
  - Performance improvement: **13.5%**
- **Challenge**
  - Suboptimal traditional counter-based secure technique
- **Main contribution**
  - Virtual integrity tree
  - Multi-granular counter
- **Additional optimization**
  - Read-only optimization
  - Direct communication channel

Thank you