SecureLoop: Design Space Exploration of Secure DNN Accelerators

Kyungmi Lee, Mengjia Yan, Joel S. Emer*, Anantha P. Chandrakasan

MIT, *MIT/NVIDIA

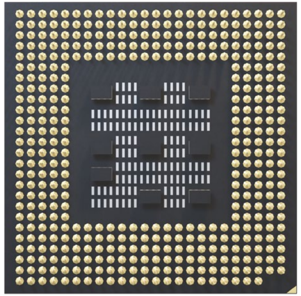
ML Needs Both Security and Performance

Autonomous Driving

Biometric Authentication

Healthcare

Applications require security

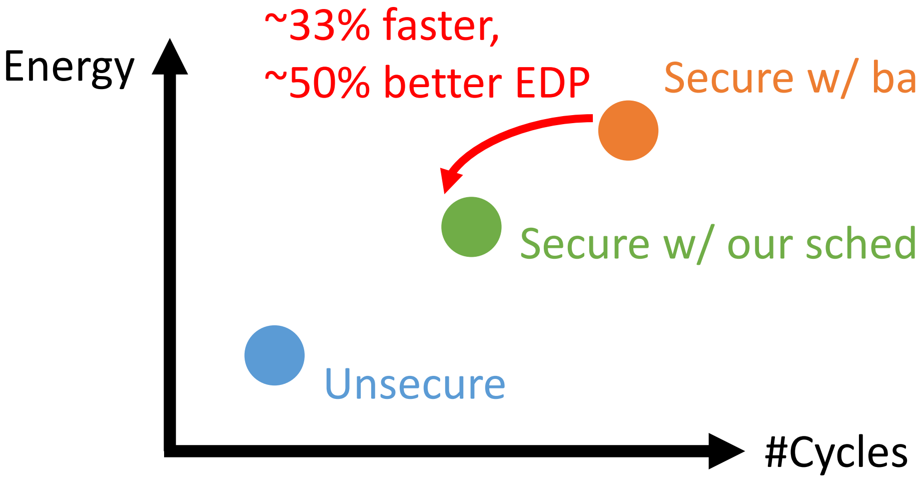


Accelerator design requires performance, area, energy

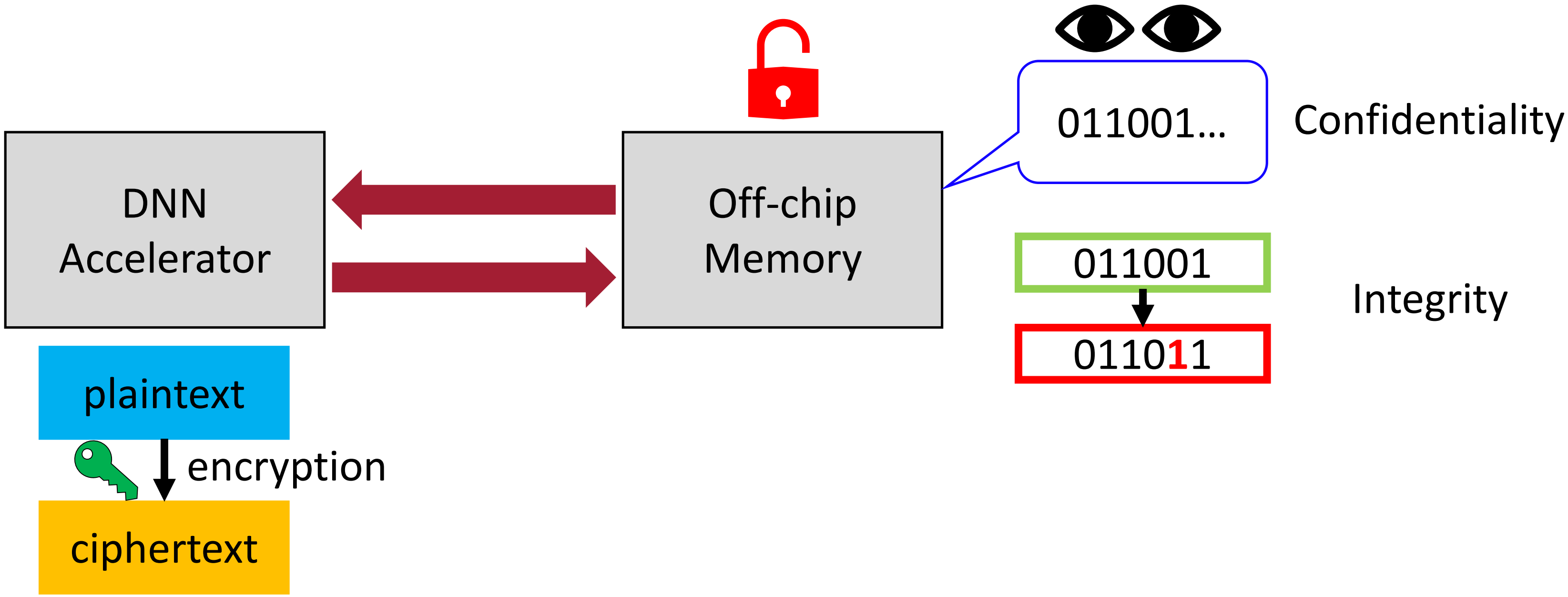
Design space exploration for “secure” DNN accelerators

SecureLoop

A tool for **design space exploration of secure DNN accelerators** equipped with cryptographic engines

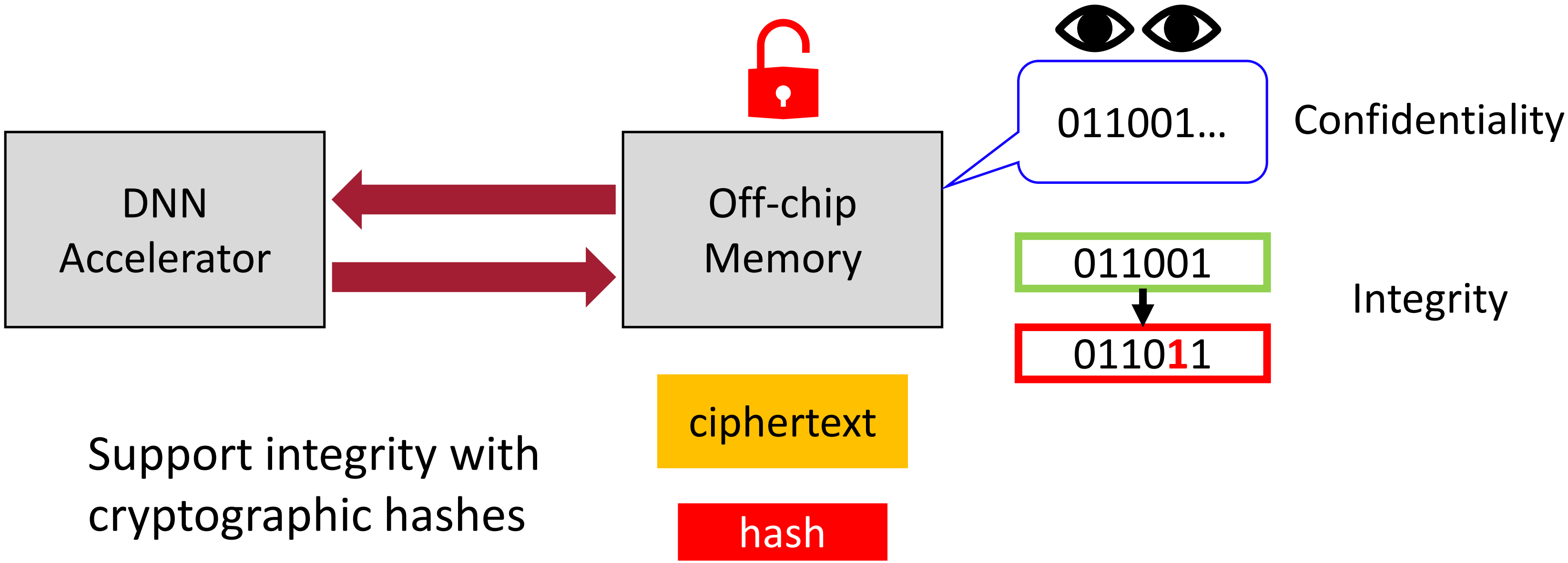


Background: TEE and DNN Accelerators



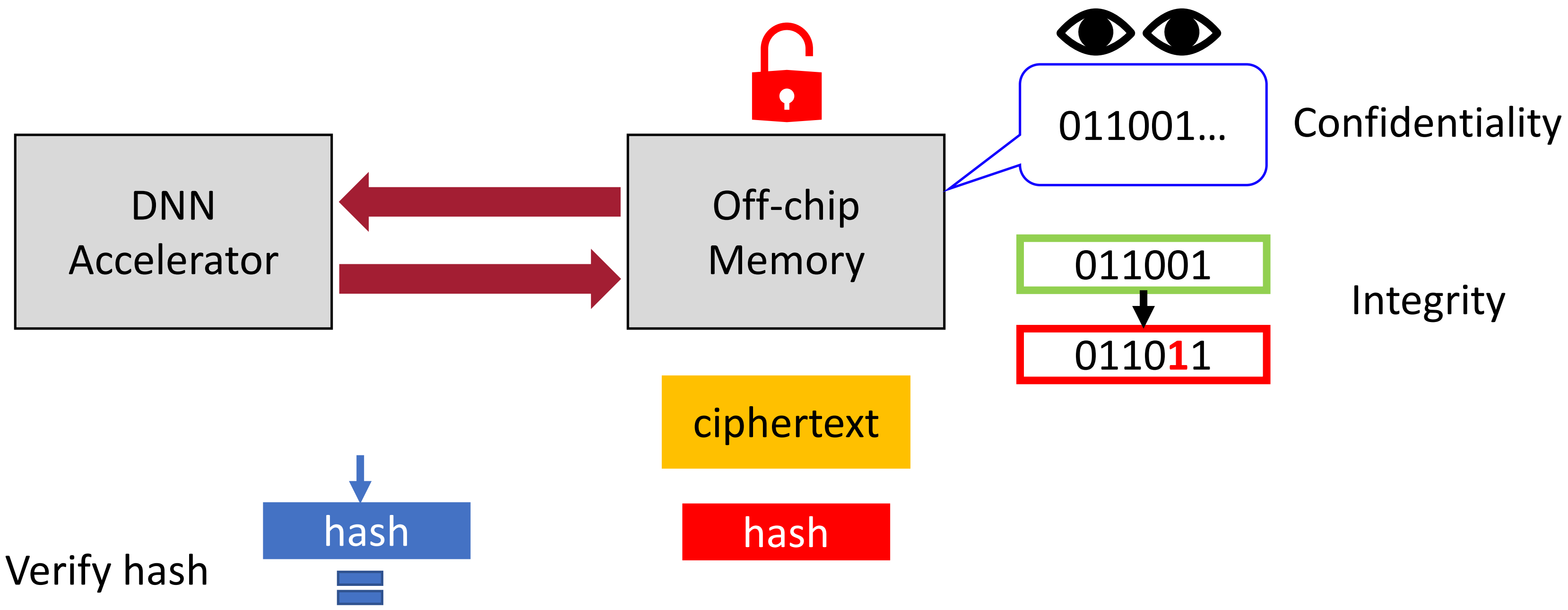
Support confidentiality with cryptographic encryption

Background: TEE and DNN Accelerators



Support integrity with cryptographic hashes

Background: TEE and DNN Accelerators



DNN Accelerator

Off-chip Memory

011001...

Confidentiality

011001

011011

Integrity

ciphertext

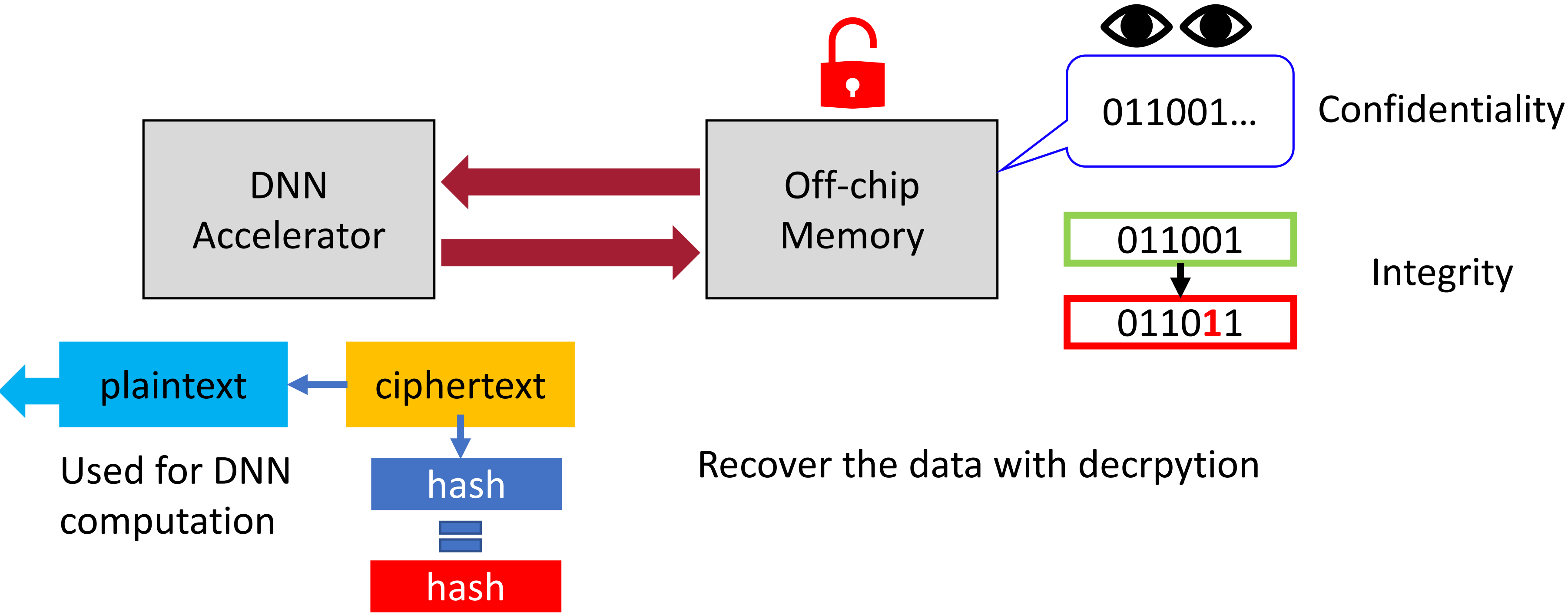
hash

Verify hash

hash

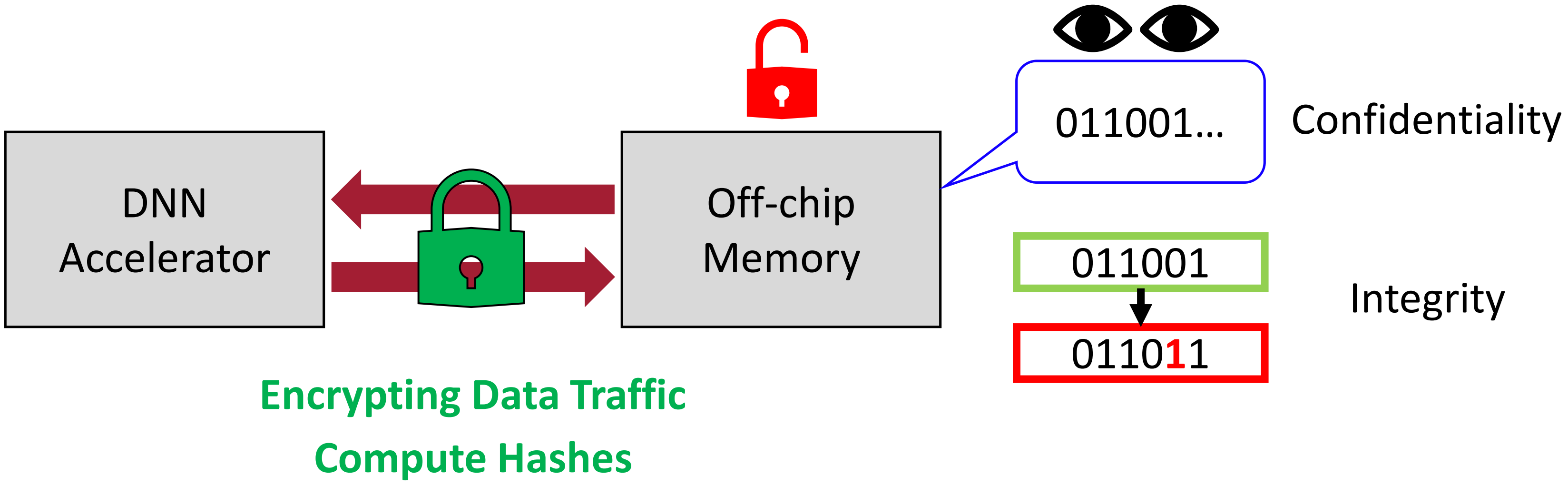
=

Background: TEE and DNN Accelerators



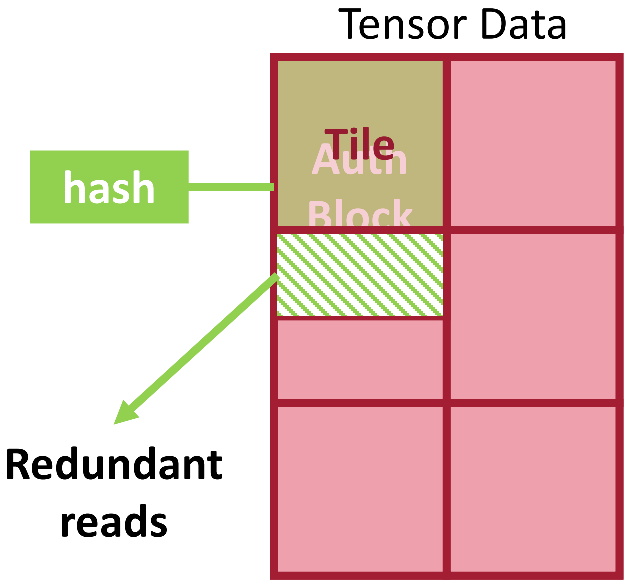
Recover the data with decrpytion

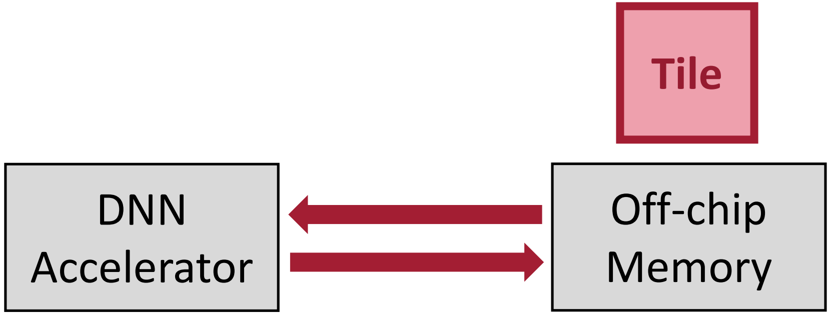
Background: TEE and DNN Accelerators



- Accelerator scheduling has to be coordinated with cryptographic operations

What if Tile != Authentication Block

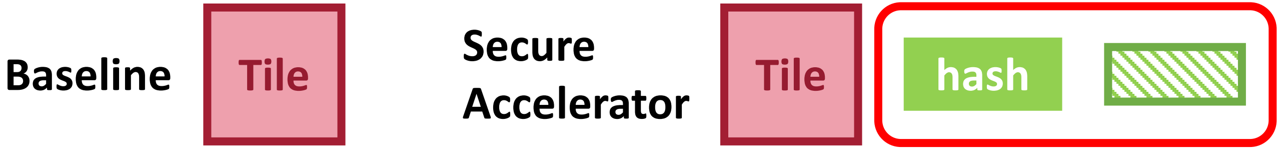




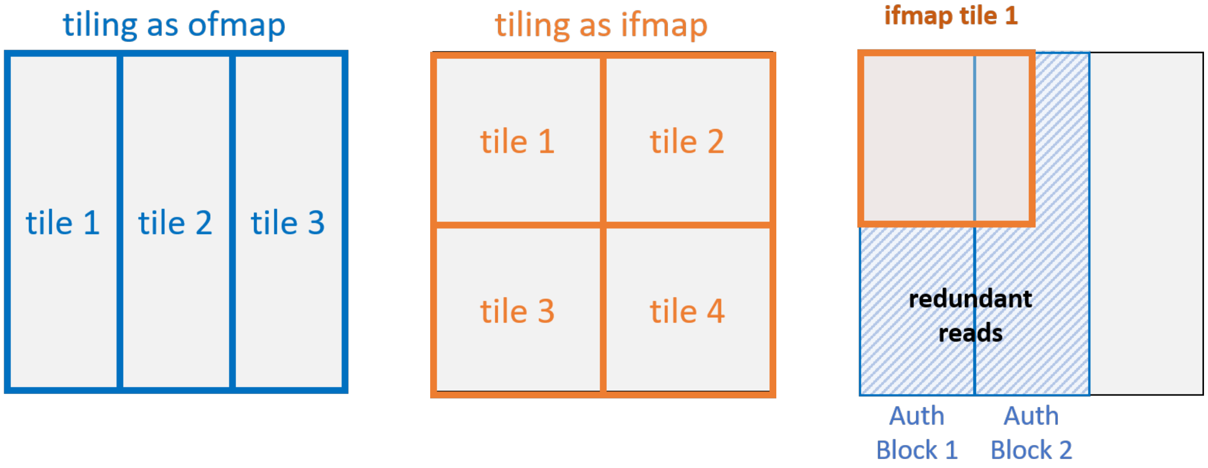
Can’t get a hash only with this ▨ta

Fetch extra “redundant” data for hash

additional off-chip traffic



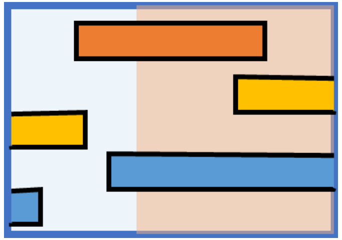
Tile-as-an-AuthBlock is not optimal



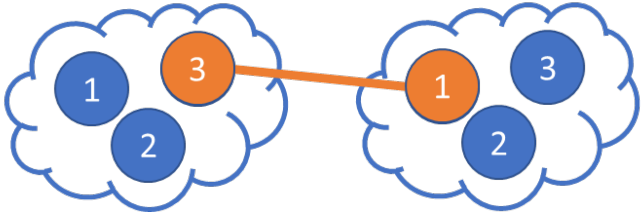
- Option 1: Tile-as-an-AuthBlock based on the output tiling ☹
- Option 2: Rehash between layers ☹

Summary of our technique

Analytical approach to identify the optimal AuthBlock assignment

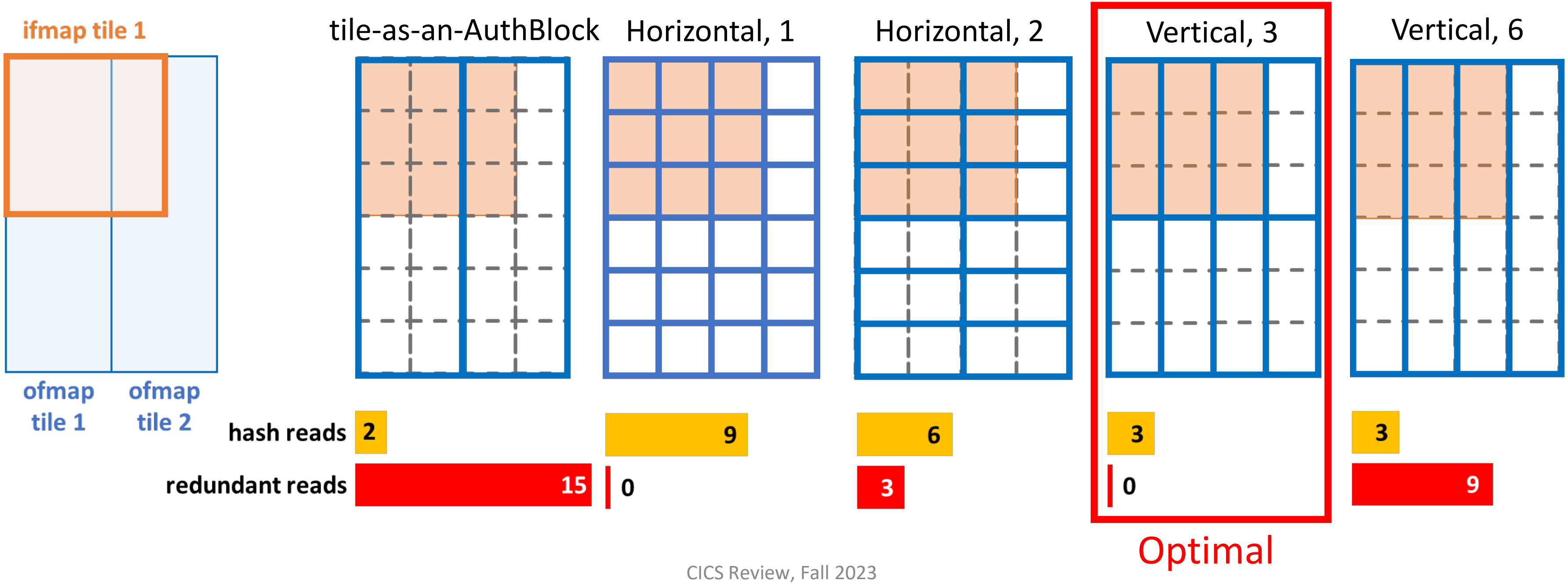


Cross-layer fine tuning from the loopnest schedule



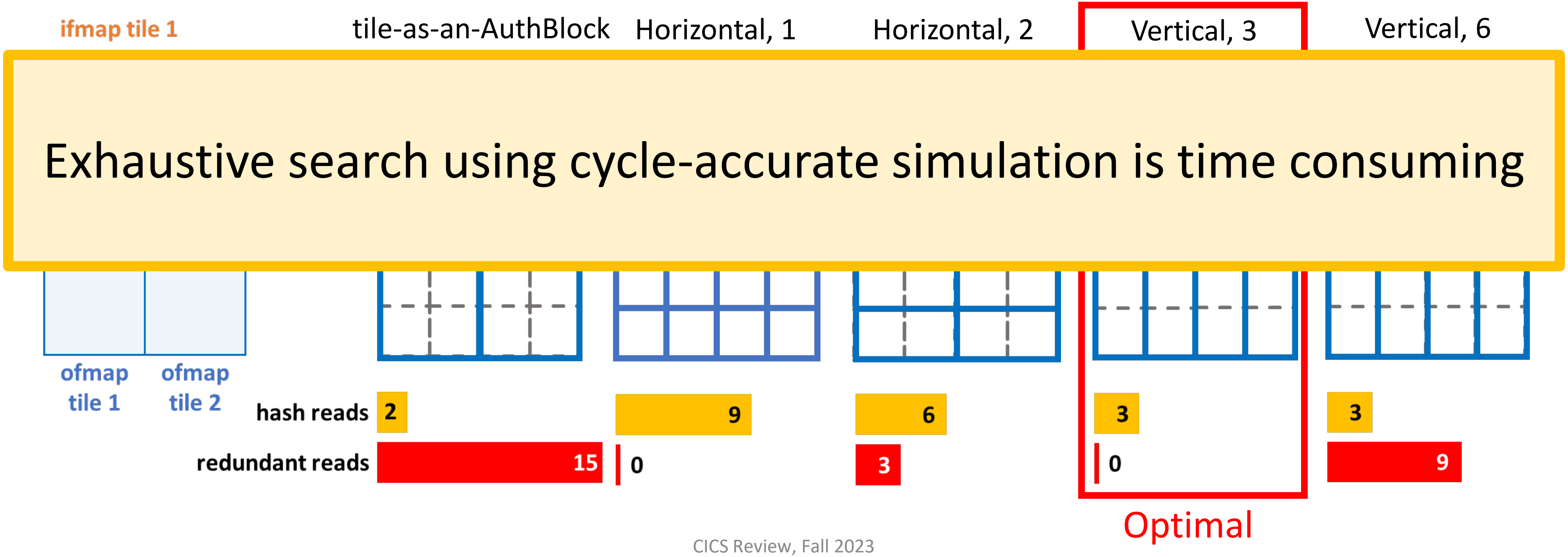
Search space of AuthBlocks is complex

- Find the AuthBlock assignment that minimizes the additional off-chip traffic
- Both size and orientation of AuthBlocks matter



Search space of AuthBlocks is complex

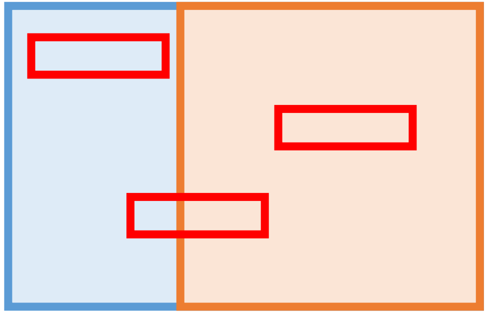
- Find the AuthBlock assignment that minimizes the additional off-chip traffic
- Both size and orientation of AuthBlocks matter



Exhaustive search using cycle-accurate simulation is time consuming

Analytical approach to AuthBlock assignment

- Counting how many AuthBlocks intersect a tile

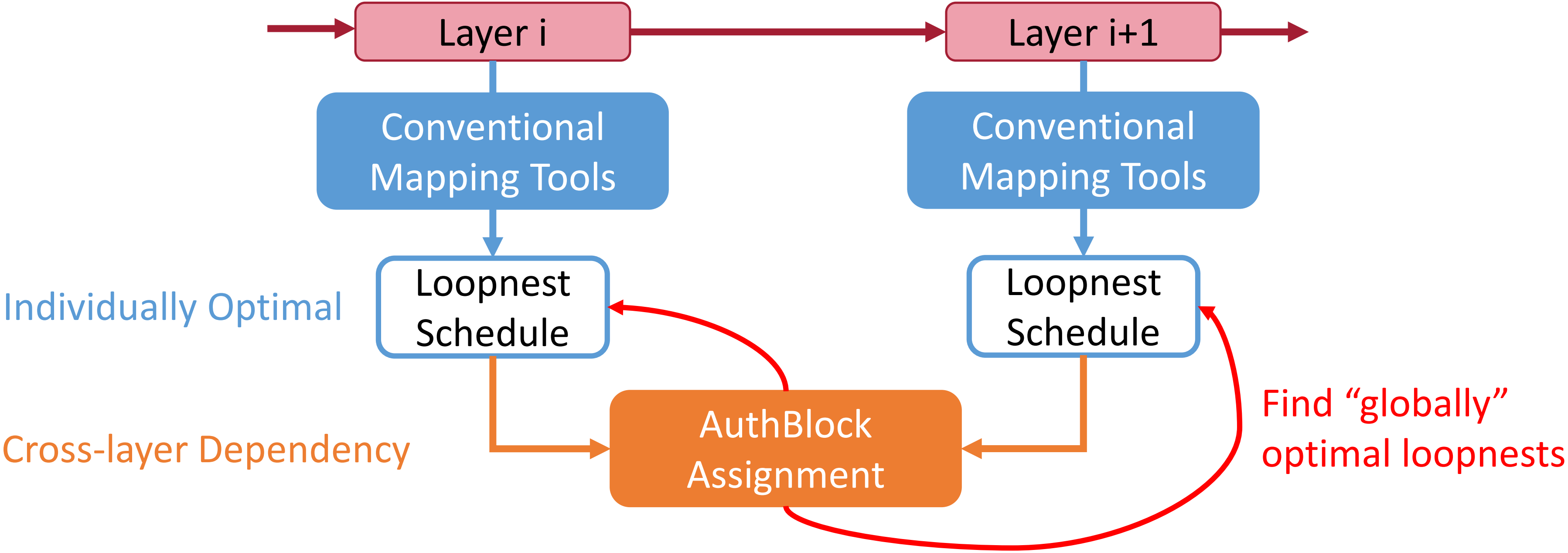


Please refer to the paper for details

Linear congruence problem

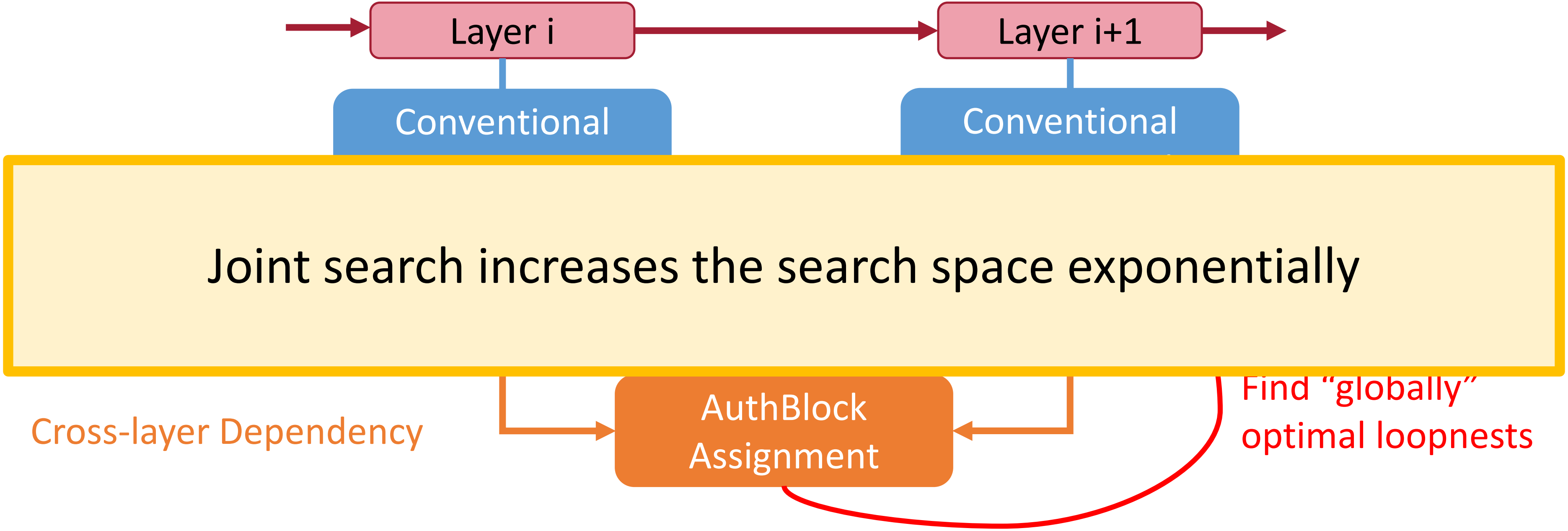
$$u \times k \equiv \min\left(w_i - w_j - u + 1, 0\right)$$
$$, \ldots, w_i - 1 \bmod w_i$$

Cross-layer dependency from the loopnest level



Loopnest schedules optimal for individual layer ↛ globally optimal

Cross-layer dependency from the loopnest level

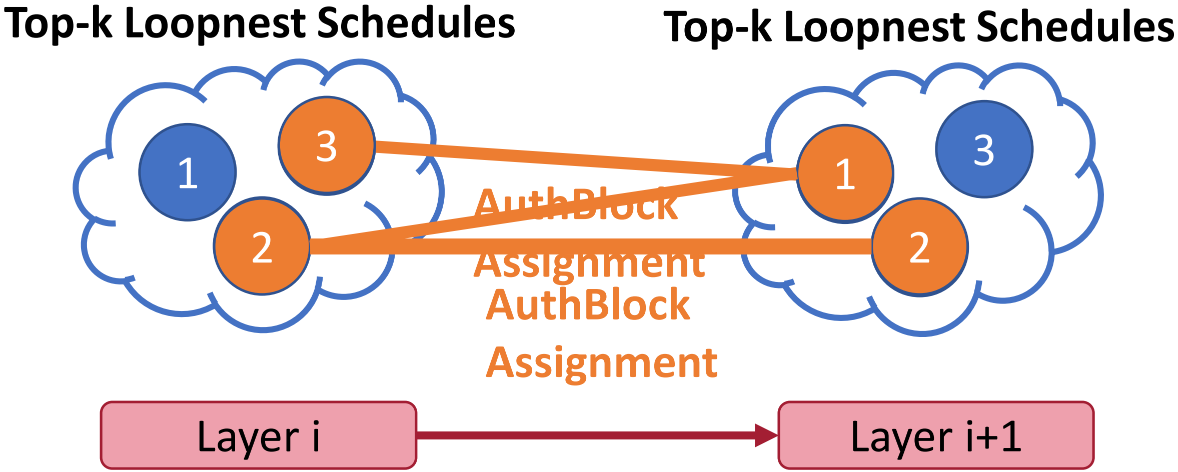


Joint search increases the search space exponentially

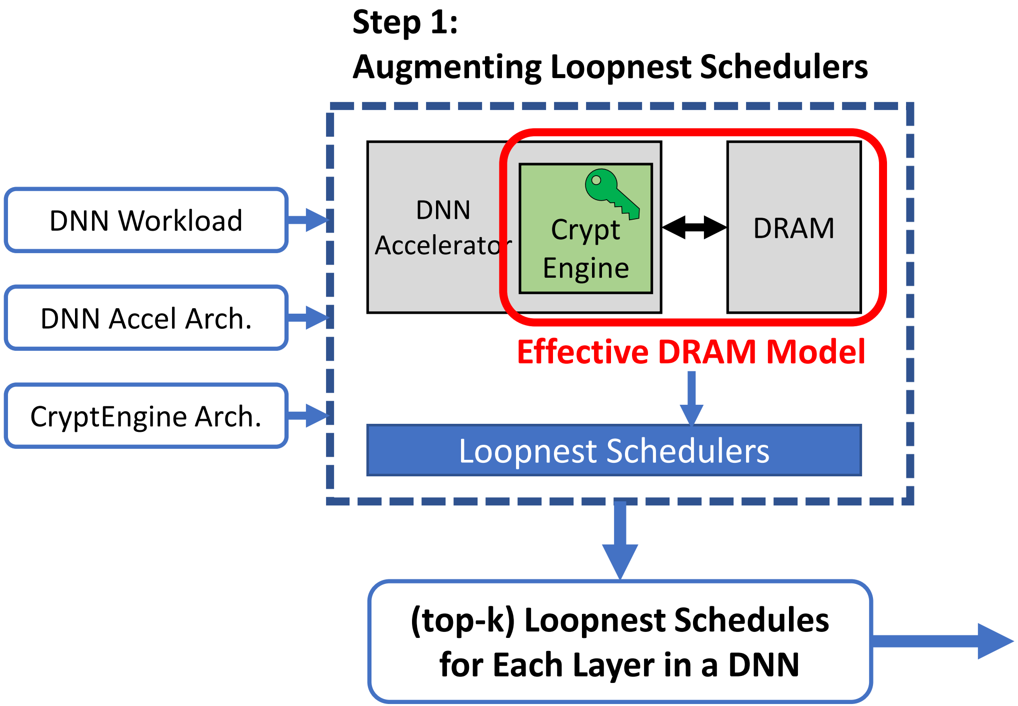
Loopnest schedules optimal for individual layer ↛ globally optimal

Heuristic approach to joint optimization

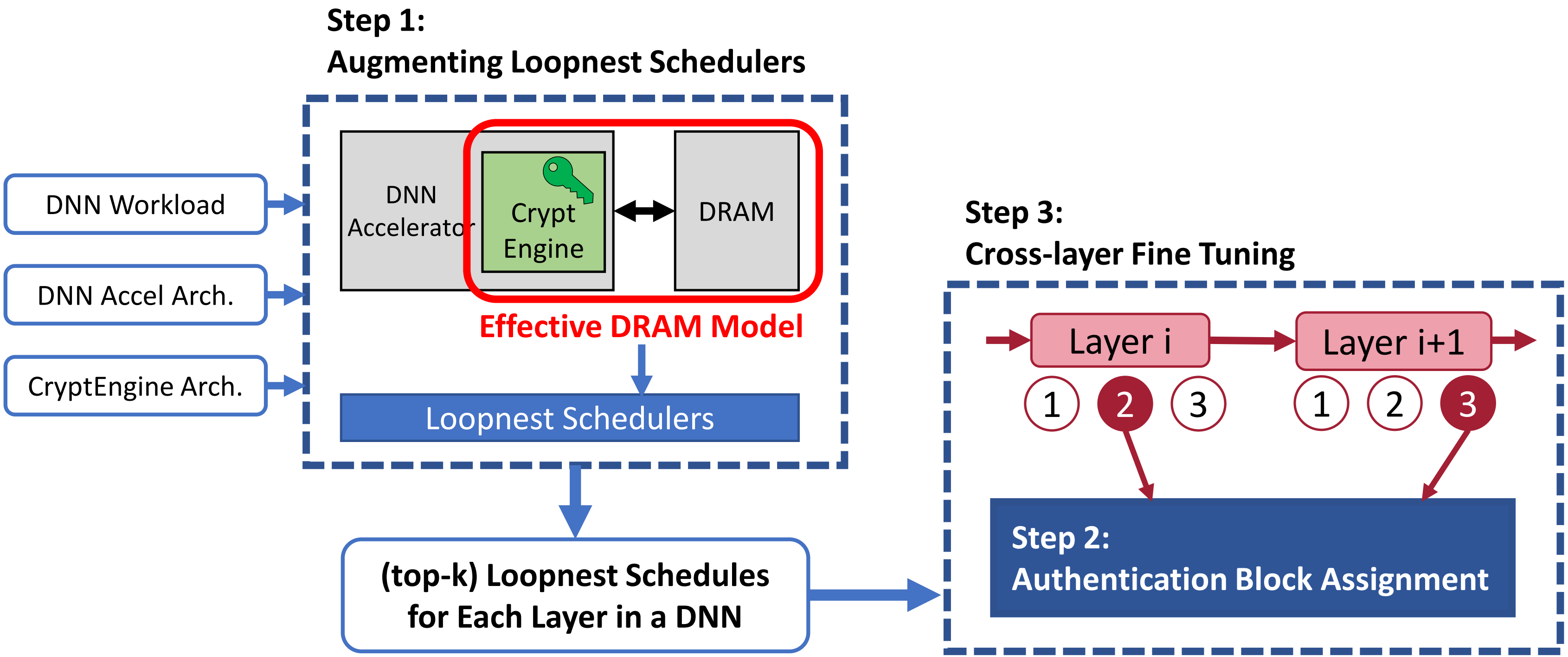
- Simulated annealing to find the approximate solution



SecureLoop: Scheduling Search Engine



SecureLoop: Scheduling Search Engine



Step 1:
Augmenting Loopnest Schedulers

(top-k) Loopnest Schedules for Each Layer in a DNN

Step 3:
Cross-layer Fine Tuning

Step 2:
Authentication Block Assignment

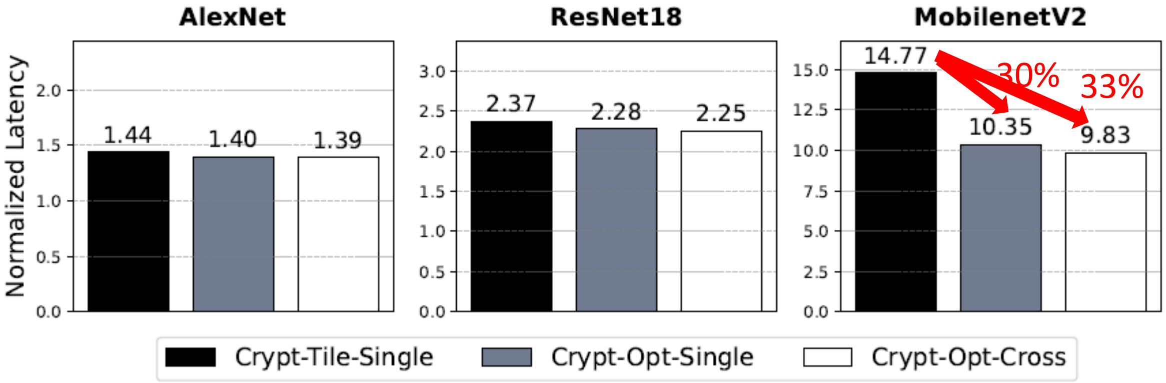
Comparing scheduling algorithms

Setup

- Same hardware
- Mostly conv workloads
- Different scheduling algo.

shallow,
unavoidable rehashing

deeper, ↑ opportunity



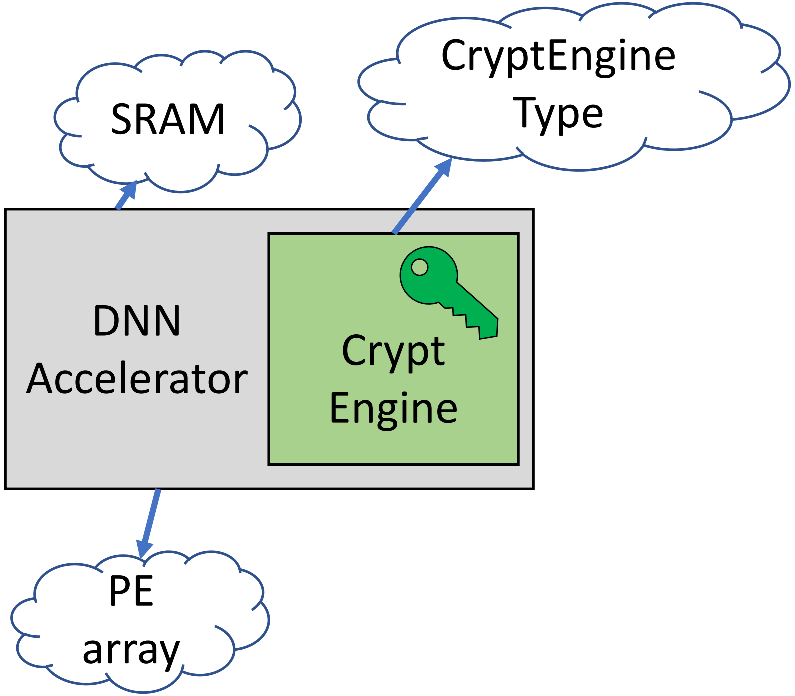
Step 1 only
+ tile-as-an-AuthBlock

Step 1 + Step 2

All three steps

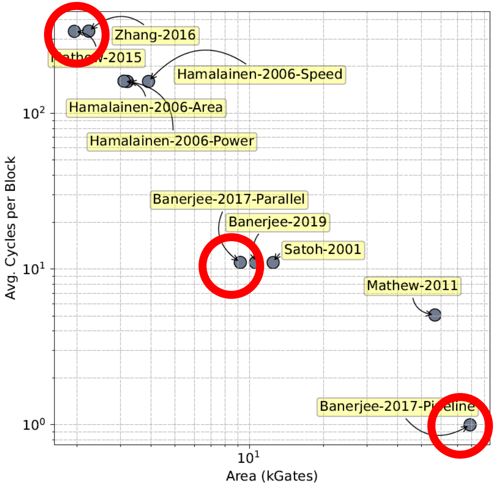
Summary: ~33% faster, ~50% better in EDP compared to the “tile-as-an-AuthBlock”

Case Studies: Design Space Exploration

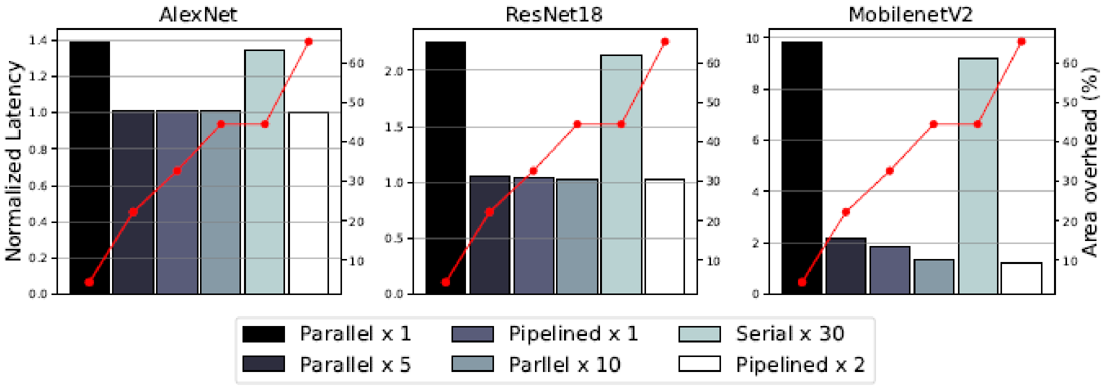


Case Studies: Design Space Exploration

CryptEngine Type



Different CryptEngine architectures and numbers

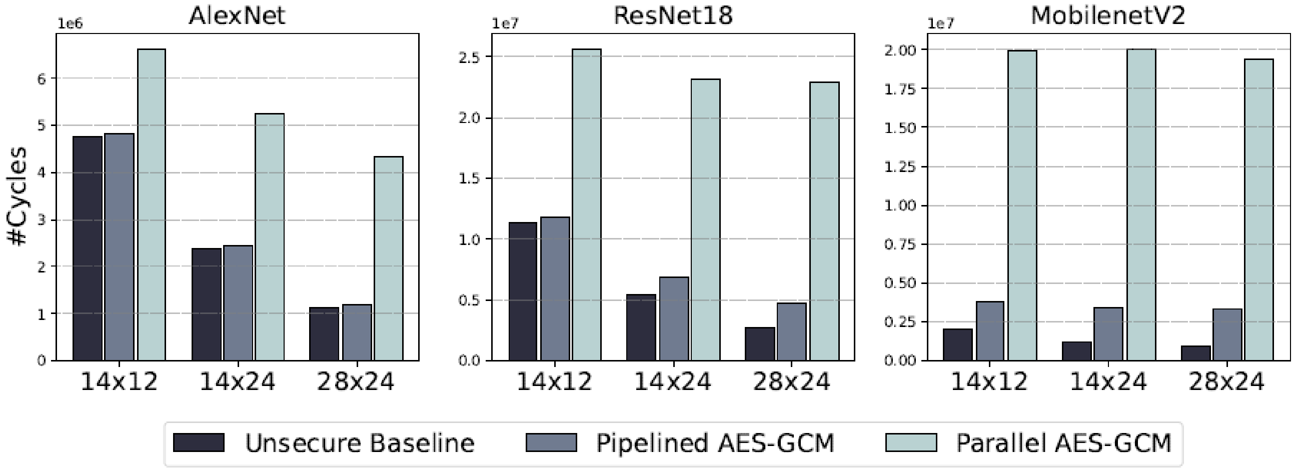


Similar area overhead, but with different performance

Case Studies: Design Space Exploration

PE array

Different PE array sizes with two types of CryptEngines

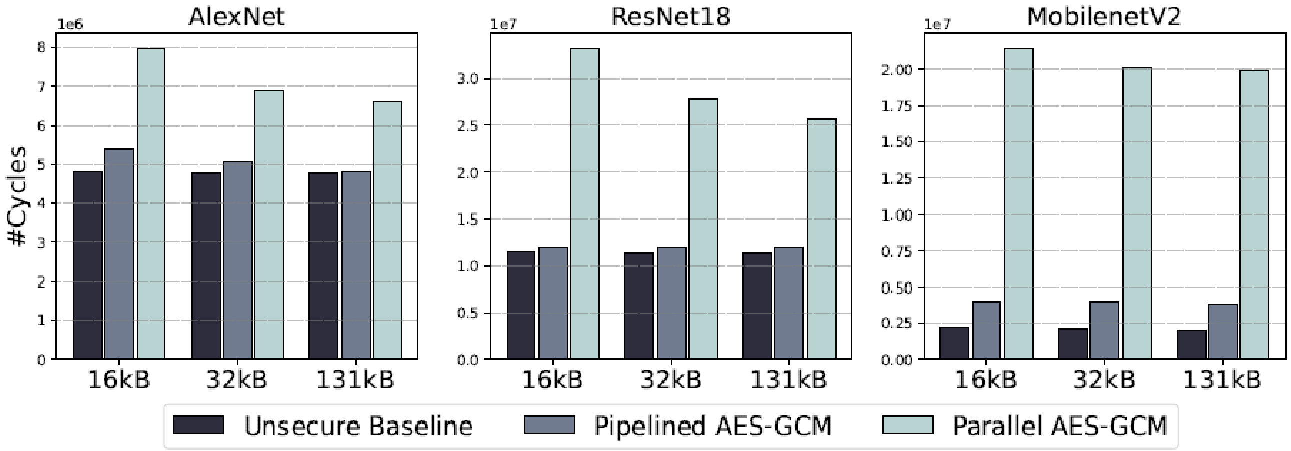


Increasing the PE array size cannot provide speedup when the CryptEngine throughput is low

Case Studies: Design Space Exploration

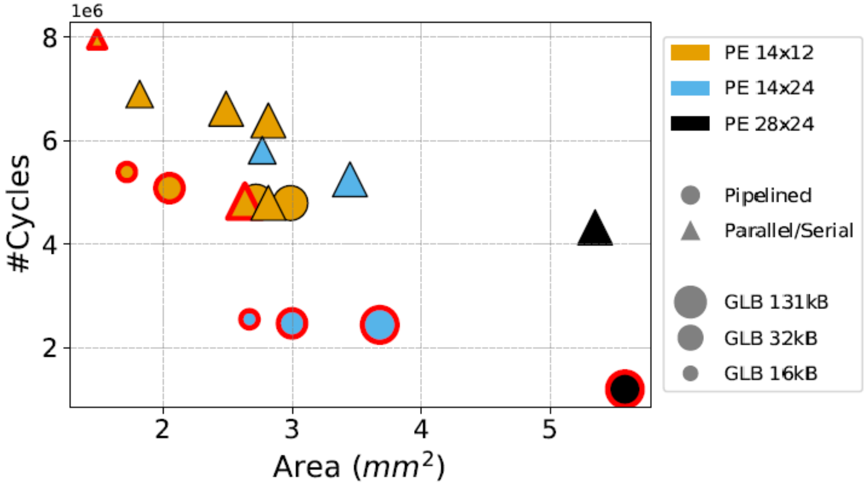
SRAM

Different on-chip SRAM buffer sizes



Reducing SRAM size doesn’t hurt performance
if the CryptEngine throughput is sufficient

Case Studies: Design Space Exploration

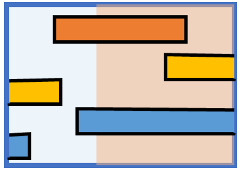


For the similar area, dedicate more area to the CryptEngine and the PE array, while sacrificing the GLB

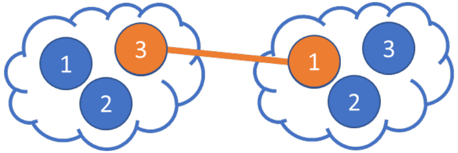
Summary

- Scheduling algorithm considering cryptographic operations

Analytical approach to identify the optimal AuthBlock assignment



Cross-layer fine tuning from the loopnest schedule



- Design space exploration for secure accelerators