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

Applications require security

Design space exploration for “secure” DNN accelerators

Accelerator design requires performance, area, energy
SecureLoop

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

- **Secure w/ baseline scheduling**
  - ~33% faster,
  - ~50% better EDP

- **Secure w/ our scheduling**

- **Unsecure**

Diagram shows:
- Energy vs. #Cycles
- Secure Loop tool comparison with different scheduling strategies.
Background: TEE and DNN Accelerators

Support confidentiality with cryptographic encryption
Background: TEE and DNN Accelerators

Support integrity with cryptographic hashes

Confidentiality

Integrity

DNN Accelerator

Off-chip Memory

Support integrity with ciphertext

hash

011001...

011001

011011
Background: TEE and DNN Accelerators

DNN Accelerator → Off-chip Memory

Confidentiality: 011001...
Integrity: 011011

Verify hash: hash = ciphertext

MICRO 2023
Background: TEE and DNN Accelerators

DNN Accelerator

Off-chip Memory

Confidentiality

Integrity

plaintext

ciphertext

hash

Recover the data with decryption

Used for DNN computation
Accelerator scheduling has to be coordinated with cryptographic operations.
What if Tile != Authentication Block

- Tensor Data
- Redundant reads
- Baseline
- Secure Accelerator

Can’t get a hash only with this data.
Fetch extra “redundant” data for hash.

Addional 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 techniques

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

\[ u \times k \equiv \min(w_i - w_j - u + 1, 0), \ldots, w_i - 1 \mod w_i \]

Please refer to the paper for details.
Cross-layer dependency from the loopnest level

**Cross-layer Dependency**

Loopnest schedules optimal for individual layer $\not\Rightarrow$ globally optimal

**Individually Optimal**

Conventional Mapping Tools

Conventional Mapping Tools

Find “globally” optimal loopnests
Cross-layer dependency from the loopnest level

Joint search increases the search space exponentially

Loopnest schedules optimal for individual layer $\not\Rightarrow$ globally optimal
Heuristic approach to joint optimization

- Simulated annealing to find the approximate solution
SecureLoop: Scheduling Search Engine

Step 1: Augmenting Loopnest Schedulers

- DNN Workload
- DNN Accel Arch.
- CryptEngine Arch.

Loopnest Schedulers

(top-k) Loopnest Schedules for Each Layer in a DNN
SecureLoop: Scheduling Search Engine

Step 1: Augmenting Loopnest Schedulers

- DNN Workload
- DNN Accel Arch.
- CryptEngine Arch.

Effective DRAM Model

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

Step 2: Authentication Block Assignment

Layer i: 1, 2, 3
Layer i+1: 1, 2, 3

Step 3: Cross-layer Fine Tuning

Layer i → Layer i+1
Comparing scheduling algorithms

Setup

- Eyeriss-like + AES-GCM
- Mostly conv workloads
- Different scheduling algo.

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

• 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

  more in paper, including sweeps for different design choices and trade-off curves..