Sealer: In-SRAM AES for High-Performance and Low-Overhead Memory Encryption

Jingyao Zhang*, Hoda Naghibijouybari†, Elaheh Sadredini*

*University of California, Riverside
†Binghamton University
Data Encryption is Crucial for Many Organizations

Hospital
Medical records
Bank
Credit or debit cards
Government
Faces
Motivating Example: Face Recognition

- However, memory and bus are vulnerable
- Advanced Encryption Standard (AES) can provide data confidentiality

Demand for high-performance low-overhead AES
△ **Challenges:** Performance, Area, Security

- **Dedicated hardware engine on chip (JSSC ’11)**
  - Low throughput
  - High area consumption on chip

- **In-memory bulk encryption (DATE ’18)**
  - Low security level
  - High latency
  - Low throughput per unit area

- **Near-memory encryption (ISCA ’17)**
  - More surface exposed to attackers
  - High latency
  - Large capacity overhead
△ **Challenges:** Performance, Area, Security

- Dedicated hardware engine on chip (JSSC ‘11)
  - Low throughput
  - High area consumption on chip

- In-memory bulk encryption (DATE ‘18)
  - Low security level
  - High latency
  - Low throughput per unit area

- Near-memory encryption (ISCA ‘17)
  - More surface exposed to attackers
  - High latency
  - Large capacity overhead

**Demand for low-latency, high-throughput, low-overhead, on-chip AES**
Overview of Our Solution: **Sealer**

- **On-chip Encryption** -> high security level
- **Bitline Computing** -> high throughput
- **Effective Data Organization** -> low area overhead
- **Stage Fusion** -> low latency

[Diagram showing Sealer components and their interactions]
Overview of Our Solution: **Sealer**

- **On-chip Encryption** -> high security level
- **Bitline Computing** -> high throughput
- **Effective Data Organization** -> low area overhead
- **Stage Fusion** -> low latency

*Sealer* can achieve **up to 323x** performance, **91x** throughput-per-area than state-of-the-art.
Sealer: Bitline Computing

- **Bitline Computing** [1]
  - Activate two wordlines simultaneously
  - Inherently perform logic operations
    - NOR
    - AND
  - Additionally support other logic operations
    - XOR
    - 8-bit SHIFT
  - Provide high parallelism

Sealer: Effective Data Organization

- Effective Data Organization
  - Integrate S-box into SRAM
  - Reduce hardware overhead
  - Enable to fuse computation stages

Our design in 256x256 subarray

Tradition design
Sealer: Stage Fusion

- Stage Fusion

AES algorithm flow chart

9x Main Rounds

SubBytes
ShiftRows
MixColumns
AddRoundKey

Substitution (S-box)

Matrix D1
2 3 1 1
1 2 3 1
1 1 2 3
3 1 1 2

Matrix D2

Round Key

S-box
Data
Key
SRAM

In-SRAM S-box

Shift-enabled SA

BLB Vref BL
MUX
Clk
EN

S_{i-8}

Si
**Sealer: Stage Fusion**

- **Stage Fusion**
  - Read and shift one byte
  - Reduce latency

AES algorithm flow chart

AES algorithm flow chart
Sealer: Overall Architecture

- High-Performance and Low-Overhead Memory Encryption

Area

Security

Latency

Throughput
Evaluation Methodology

- NVSim simulator for area consumption
- DESTINY simulator for energy and power consumption
- Cycle numbers for bitline computing are from [1,2]

Baselines:
- On-chip dedicated engines
  - EE-1 [3], EE-2 [4]
- Off-chip in-memory engines
  - AIM-NVM [5], DW-AES [6]
- On-chip in-memory engine (apples-to-apples comparison)
  - AIM-SRAM[5]

[3] Design and implementation of low-area and low-power AES encryption hardware core. In DSD.
[5] Securing emerging nonvolatile main memory with fast and energy-efficient AES in-memory implementation. TVLSI.
Latency

- On-chip dedicated engines are limited by low parallelism
- Architectural contribution
  - Effective data organization > No LUT query
  - Stage fusion -> Data movement reduction
- Technology contribution
  - Frequency

![Latency Chart]

- 24 data blocks
- 192 data blocks

- Normalized Latency

<table>
<thead>
<tr>
<th></th>
<th>Sealer</th>
<th>EE-1</th>
<th>EE-2</th>
<th>DW-AES</th>
<th>AIM-NVM</th>
<th>AIM-SRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>24 data blocks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>192 data blocks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Latency

- On-chip dedicated engines are limited by low parallelism

- Architectural contribution
  - Effective data organization > No LUT query
  - Stage fusion - > Data movement reduction

- Technology contribution
  - Frequency

![Graph showing LATencies for different stages with AIM-SRAM and Sealer]
Throughput/Area, Energy & Power

- Lower latency, high parallelism -> higher throughput
- Least modification to SRAM arrays (< 1.55%) -> least area consumption
- Fewer operations -> lower energy, higher utilization -> higher power
Throughput/Area, Energy & Power

- Lower latency, high parallelism → higher throughput
- Least modification to SRAM arrays (< 1.55%) → least area consumption
- Fewer operations → lower energy, higher utilization → higher power

Sealer provides a high-performance and low-overhead on-chip encryption solution
Conclusion

- **Sealer** provides **low-latency, high-throughput, low-overhead, high-security** all by proposing an in-SRAM AES encryption solution.

- **Effective data organization** and **stage fusion** are proposed to efficiently map the algorithm to the **Sealer** architecture.

- **Sealer** can achieve **up to 323x** performance, **91x** throughput-per-area than state-of-the-art solutions with **< 1.55%** modification to conventional SRAM.
Q&A