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
- Bank
- Government
- Medical records
- Credit or debit cards
- 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

Demand for low-latency, high-throughput, low-overhead, on-chip AES

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

- **Core**
- **Cache**
- **LL $\$ Sealer**

---

**S-box**

- **Data**
- **Key**
- **SRAM**

---

**Substitution (S-box)**

- **SubBytes**
- **ShiftRows**
- **AddRoundKey**
- **Round Key**

---

**Matrix D1**

- 2 3 1 1
- 1 2 3 1
- 1 1 2 3
- 3 1 1 2

---

**Matrix D2**

- 9x Main Rounds
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

[1] Aga. Shaizeen, et al. "Compute caches." 2017 IEEE International Symposium on High Performance Computer Architecture (HPCA). IEEE, 2017.
Sealer: Effective Data Organization

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

Tradition design

Our design in 256x256 subarray

Our 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

Round Key

In-SRAM S-box

S-box
Data
Key
SRAM

Shift-enabled SA

BLB Vref BL
SA
MUX

Clk
Q V
EN
Si
Si-8
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

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]

---

[1] Shatzeen Aga et al. 2017. Compute Caches. In HPCA.
[2] Arun Subramaniyan et al. 2017. Cache Automaton. In MICRO.
[3] Design and implementation of low-area and low-power AES encryption hardware core. In DSD.
[4] 53Gbps native GF(24) 2 composite-field AES-encrypt/decrypt accelerator for content protection in 45nm high-performance microprocessors. In VLSIC.
[5] Securing emerging nonvolatile main memory with fast and energy-efficient AES in-memory implementation. TVLSI.
[6] DW-AES: a domain-wall nanowire-based AES for high throughput and energy-efficient data encryption in non-volatile memory. IEEE TIFS.
Latency

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

![Normalized Latency Chart](image)
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

![Diagram showing latency comparison between AIM-SRAM and Sealer]
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 -&gt; higher throughput
- Least modification to SRAM arrays (&lt; 1.55%) -&gt; least area consumption
- Fewer operations -&gt; lower energy, higher utilization -&gt; 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.
