In-DRAM True Random Number Generation Using Simultaneous Multiple-Row Activation An Experimental Study of Real DRAM Chips

İsmail Emir Yüksel

A. Olgun F. N. Bostancı O. Canpolat

G. F. Oliveira M. Sadrosadati A. G. Yağlıkçı O. Mutlu

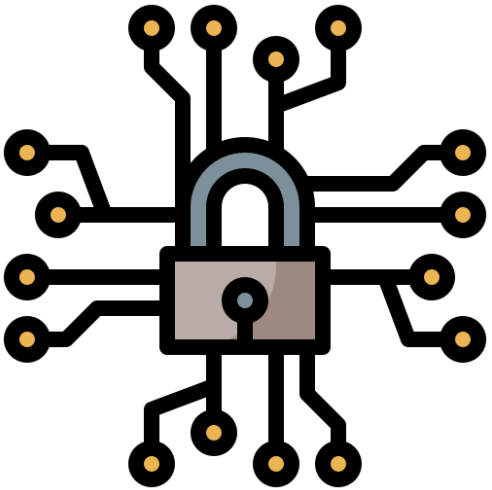
ETH zürich SAFARI CISPA

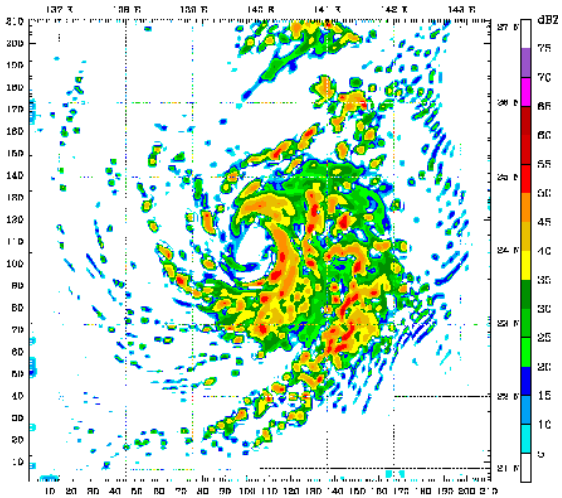
Outline

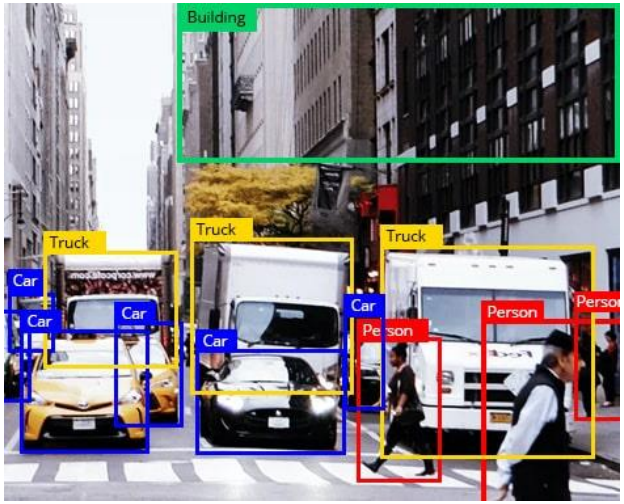
Outline

Use Cases of True Random Numbers

High-quality true random numbers
are critical to many applications







True random numbers can only be obtained
by sampling random physical processes

Unfortunately, not all computing systems are equipped with
true random number generator (TRNG) hardware

DRAM-Based TRNGs

DRAM chips are ubiquitous in modern computing platforms

DRAM-based TRNGs enable true random number generation within DRAM chips

Low-cost: No specialized circuitry for RNG

- Beneficial for constrained systems

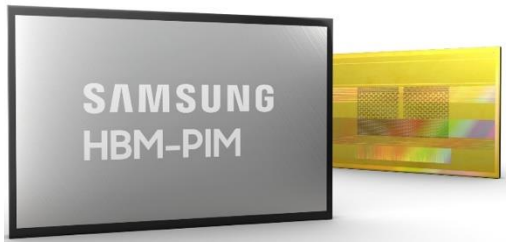
High-throughput: > Gb/s throughput

- Open application space that require high-throughput TRNG

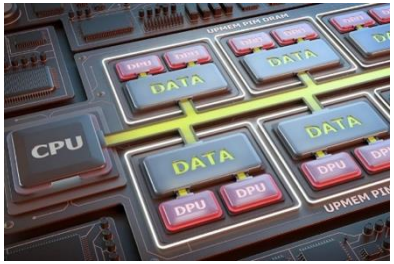
Synergy with Processing-in-Memory

Processing-in-Memory (PIM) Systems

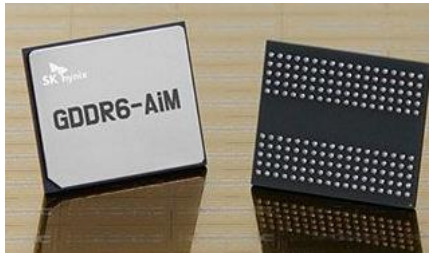
- Perform computation directly near/within a memory chip
- Improve system performance by avoiding off-chip data movement



[Samsung]



[UPMEM]

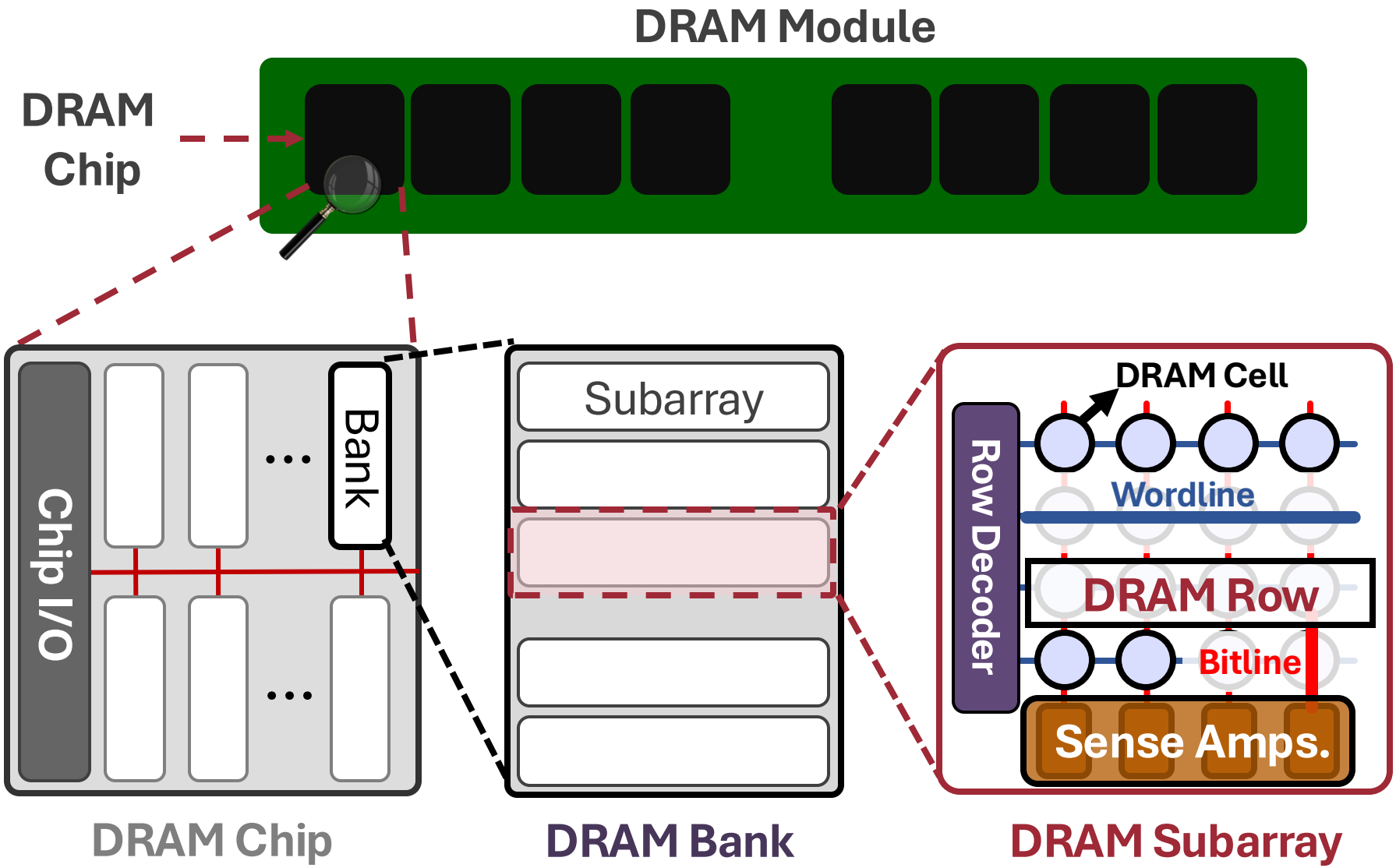


[SK hynix]

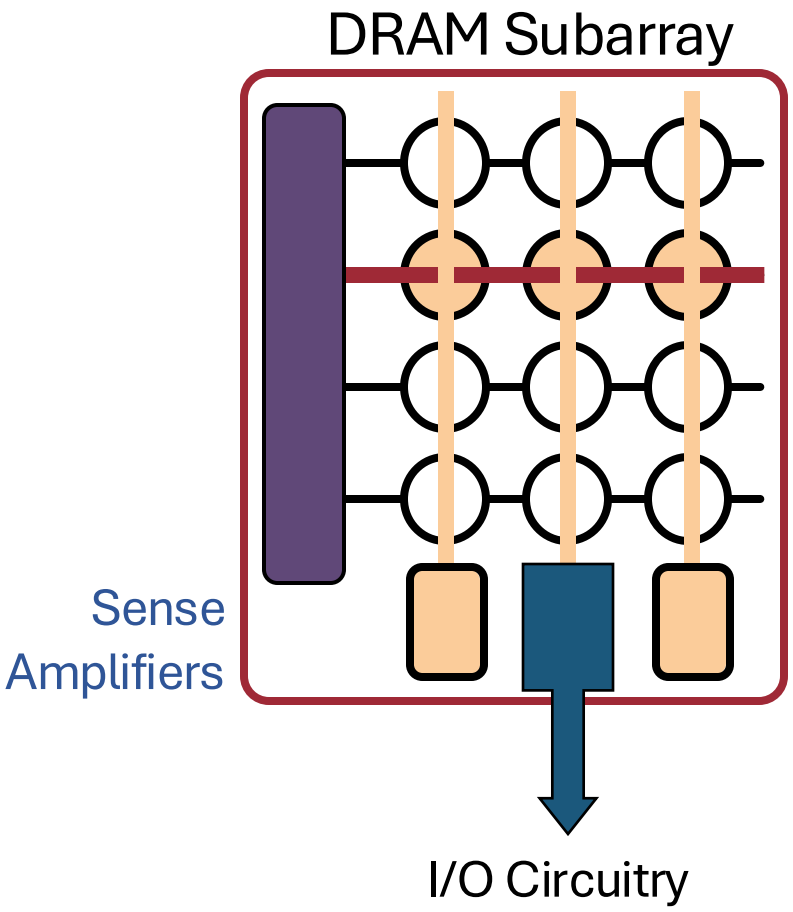
True random number generation within DRAM

- Enables PIM workloads to sample true random numbers directly within the memory chip
- Avoids inefficient communication to other possible off-chip TRNG sources, enhances security & privacy

DRAM Organization



DRAM Operation



1 **ACTIVATE (ACT):**
Fetch the row's content into the **sense amplifiers**

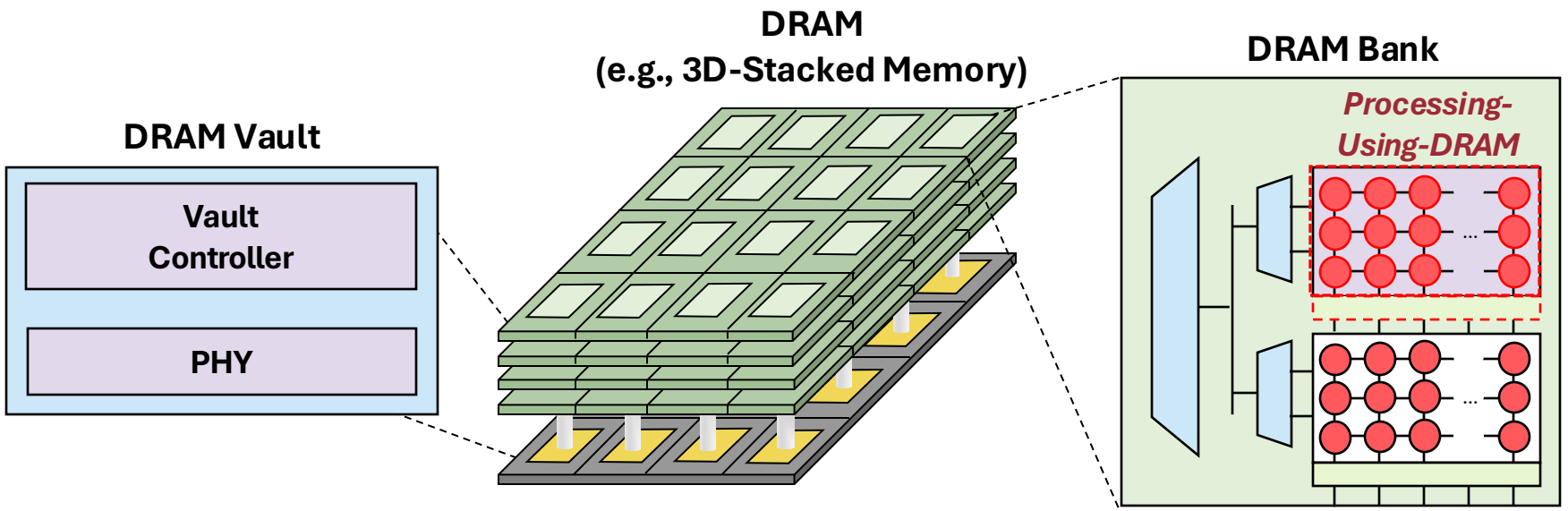
2 **Column Access (RD/WR):**
Read/Write the target column and drive to I/O

3 **PRECHARGE (PRE):**
Prepare the bank for a new ACTIVATE

Processing-using-DRAM (PuD)

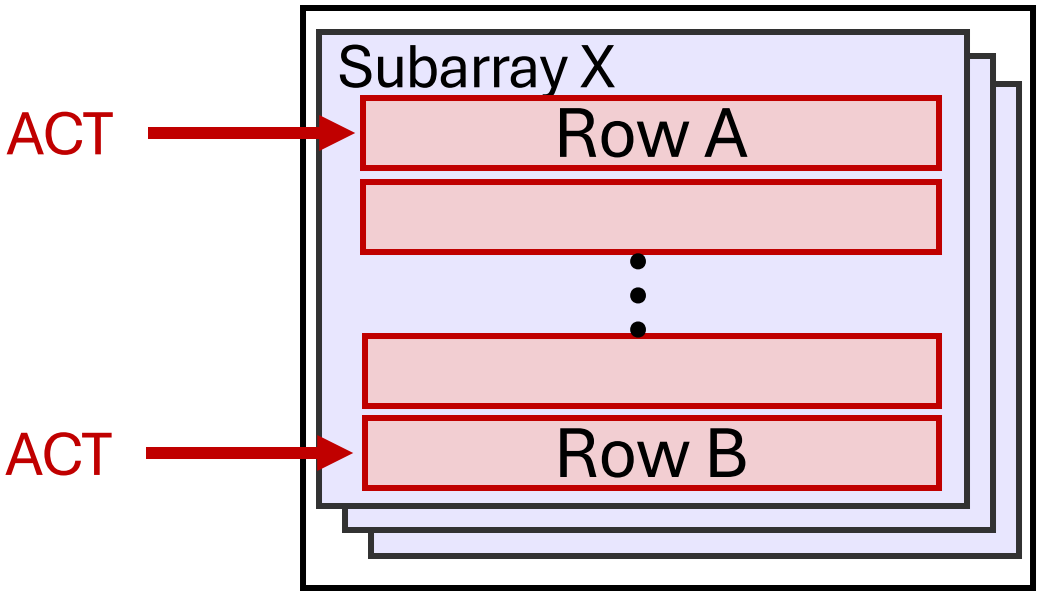
Processing-using-DRAM uses the analog operational principles of DRAM cells to perform computation

Simultaneous multiple-row activation (SiMRA) is a key technique to realize many PuD operations (e.g., AND and OR)



Simultaneous Multiple-Row Activation

Activating two rows in **quick succession** in **COTS** DRAM chips **simultaneously** activates **multiple rows** in a subarray



Our Goal

Experimentally understand
the **capability** of SiMRA **as a source of entropy**
in COTS DRAM chips

Experimentally analyze & characterize
such **capability**
under various **operating conditions**

Outline

Motivation & Goal

Generating Random Numbers via SiMRA

Testing Infrastructure

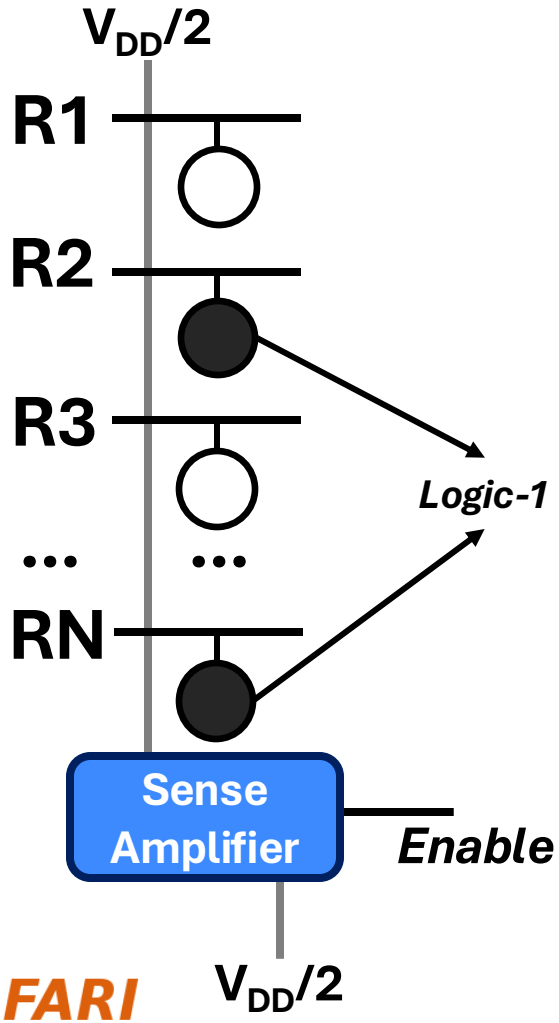
Real DRAM Chip Results

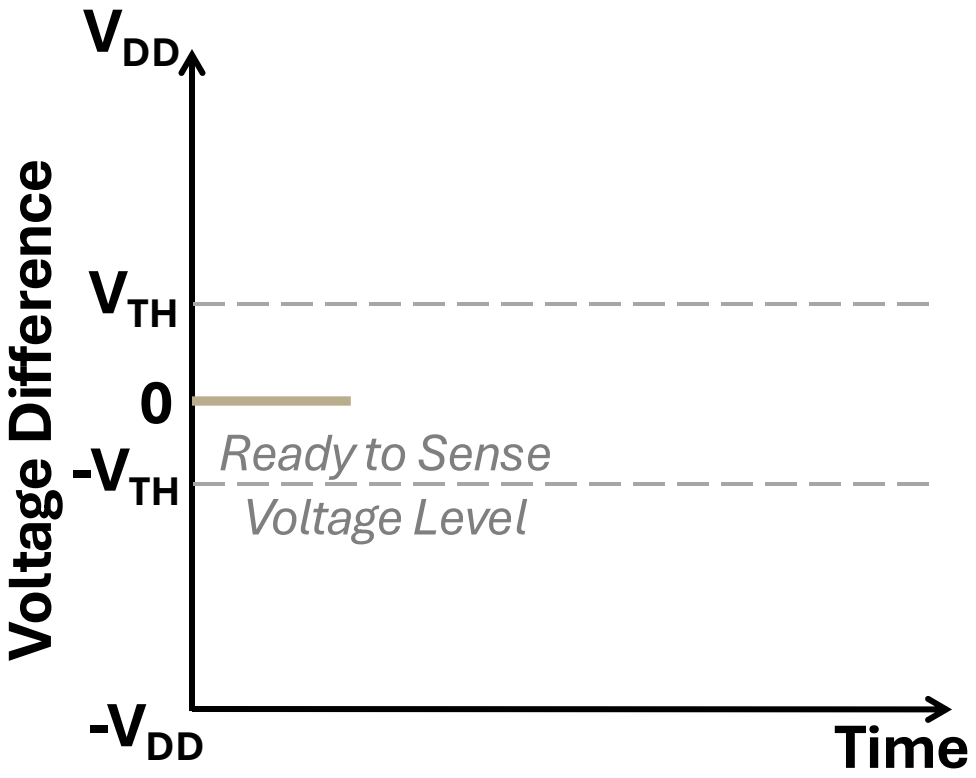
SiMRA-TRNG

Conclusion

Generating Random Values via SiMRA

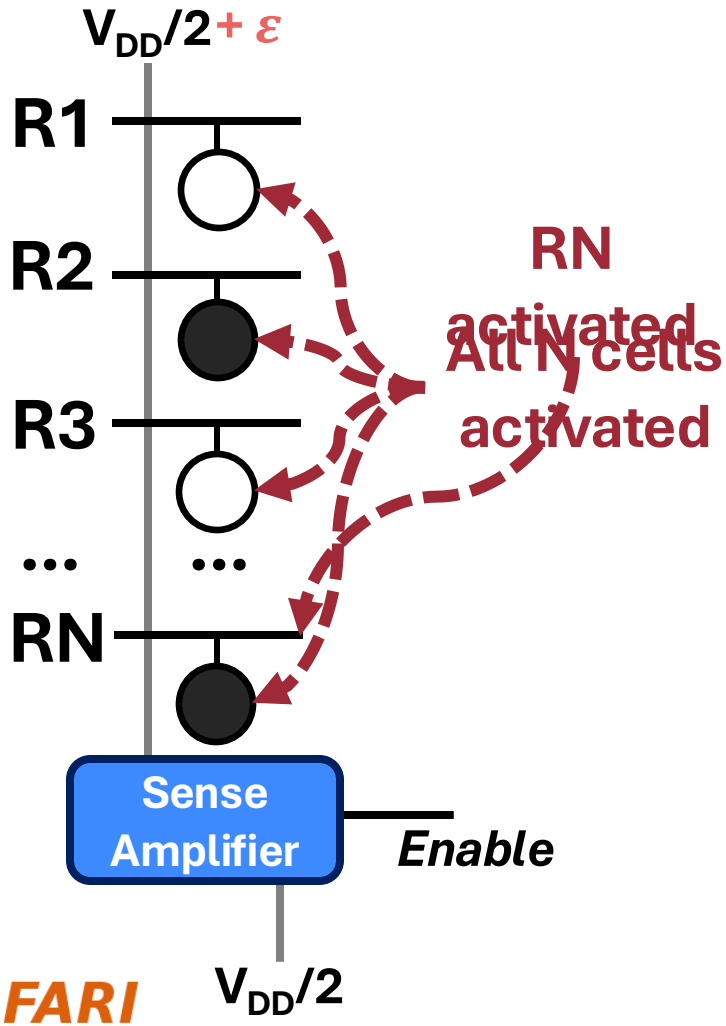
Simultaneously activating N rows with
half storing logic-1 and half storing logic-0

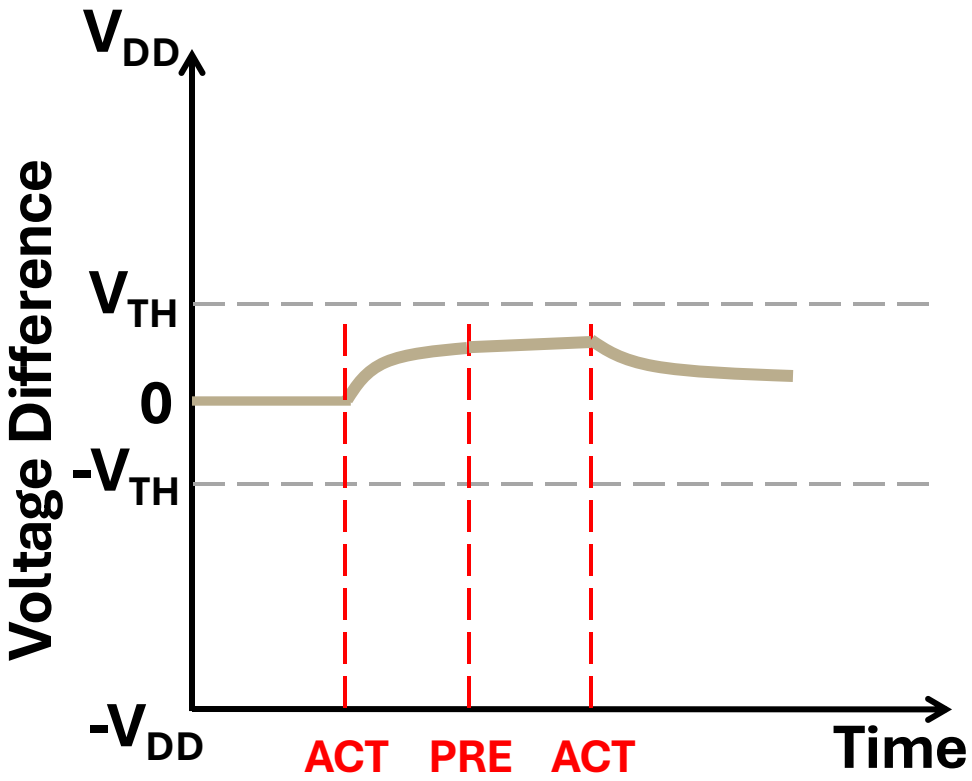




Generating Random Values via SiMRA

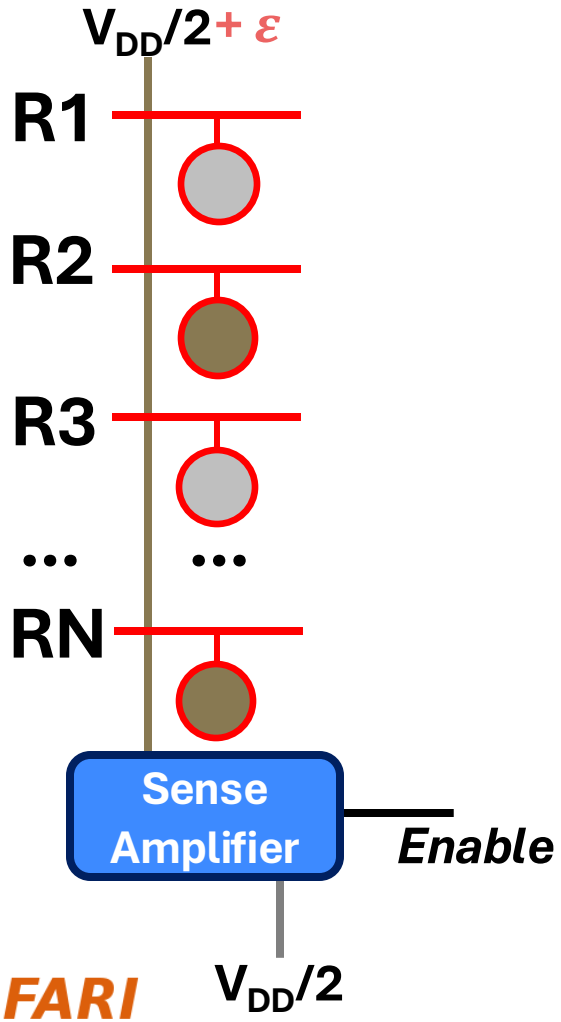
Simultaneously activating N rows with
half storing logic-1 and half storing logic-0

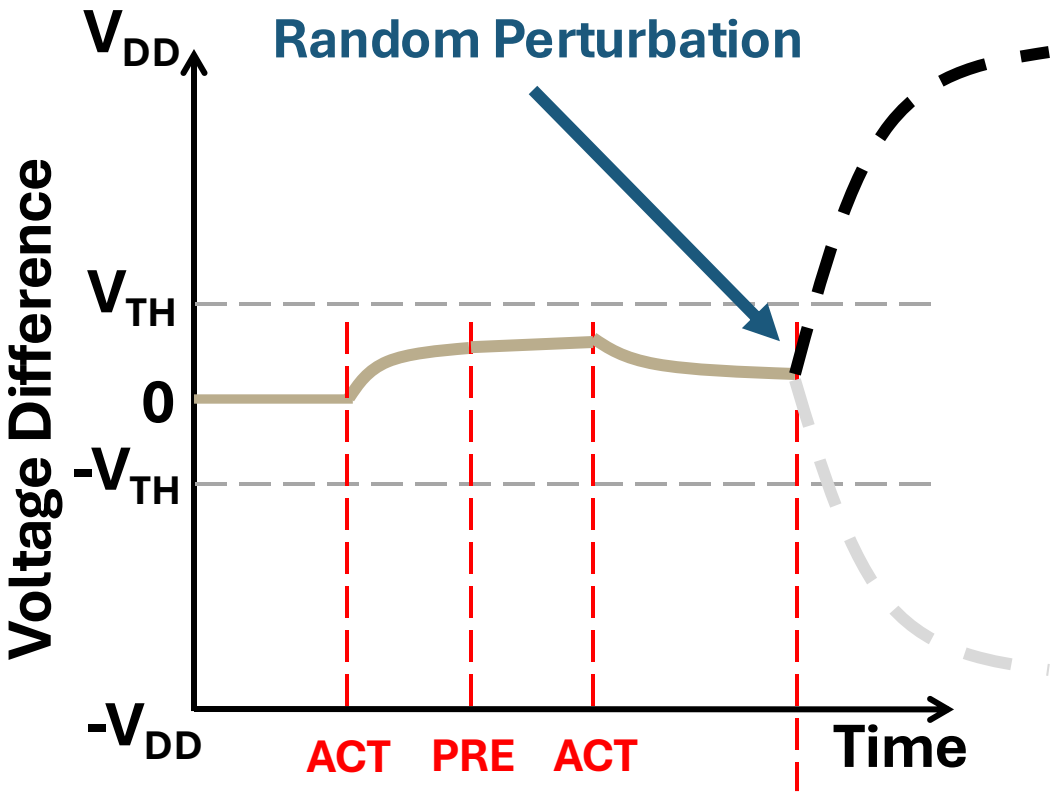




Generating Random Values via SiMRA

Simultaneously activating N rows with
half storing logic-1 and half storing logic-0





Outline

Motivation & Goal

Generating Random Numbers via SiMRA

Testing Infrastructure

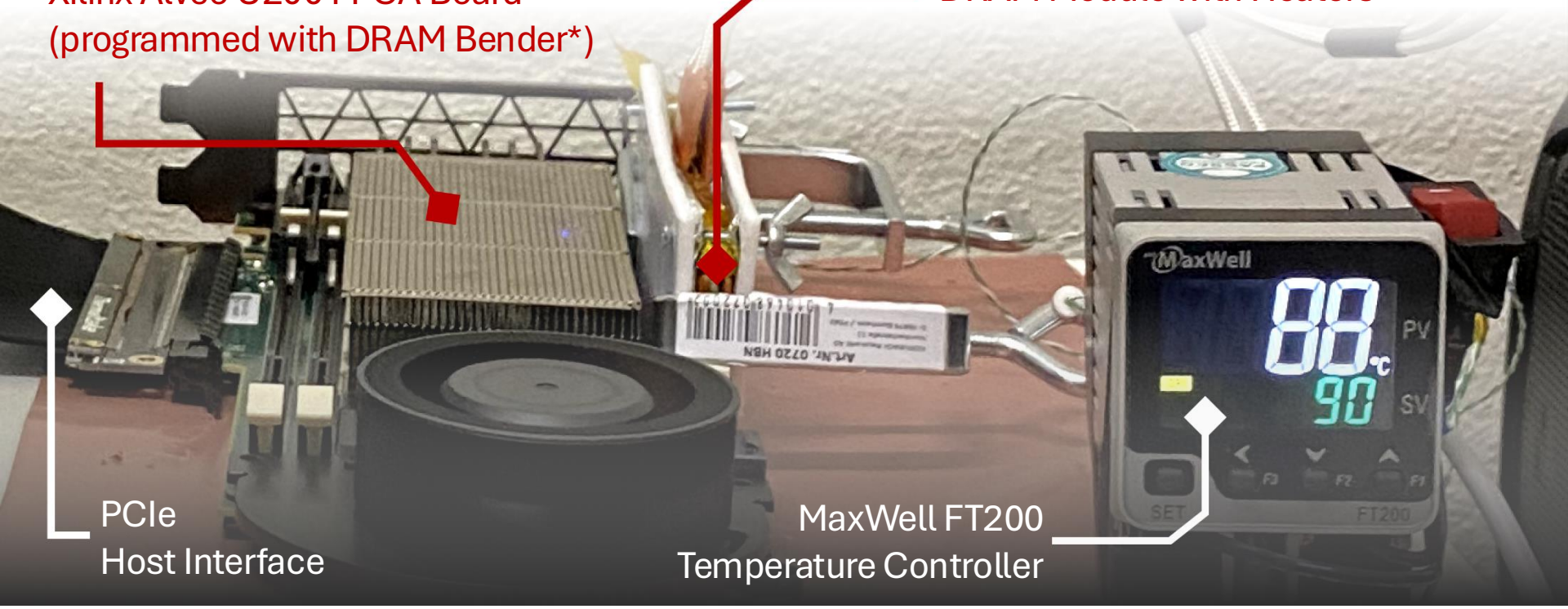
Real DRAM Chip Results

SiMRA-TRNG

Conclusion

DDR4 DRAM Testing Infrastructure

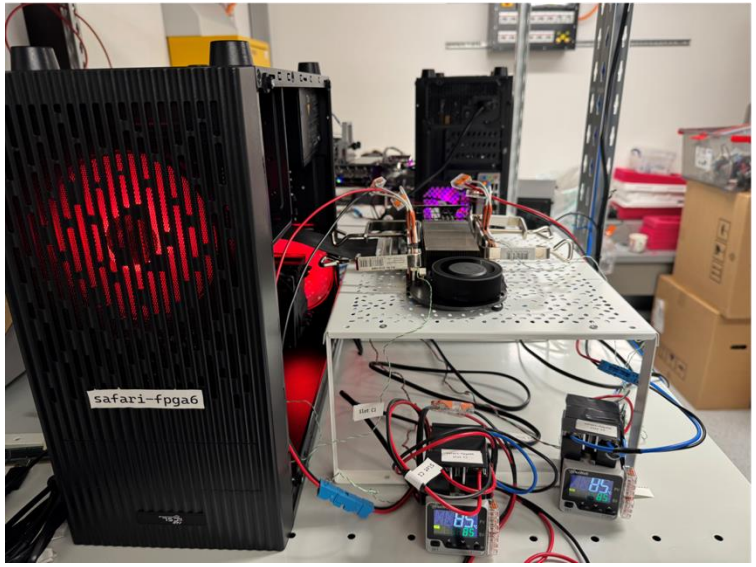
DRAM Bender on a Xilinx Virtex UltraScale+ XCU200



Fine-grained control over DRAM commands,
timing parameters (±1.5ns), and temperature (±0.5°C)

*Olgun et al., "DRAM Bender: An Extensible and Versatile FPGA-Based Infrastructure to Easily Test State-of-the-Art DRAM Chips," TCAD, 2023. [GitHub: https://github.com/CMU-SAFARI/DRAM-Bender]

Laboratory for Understanding Memory



Tested DRAM Chips

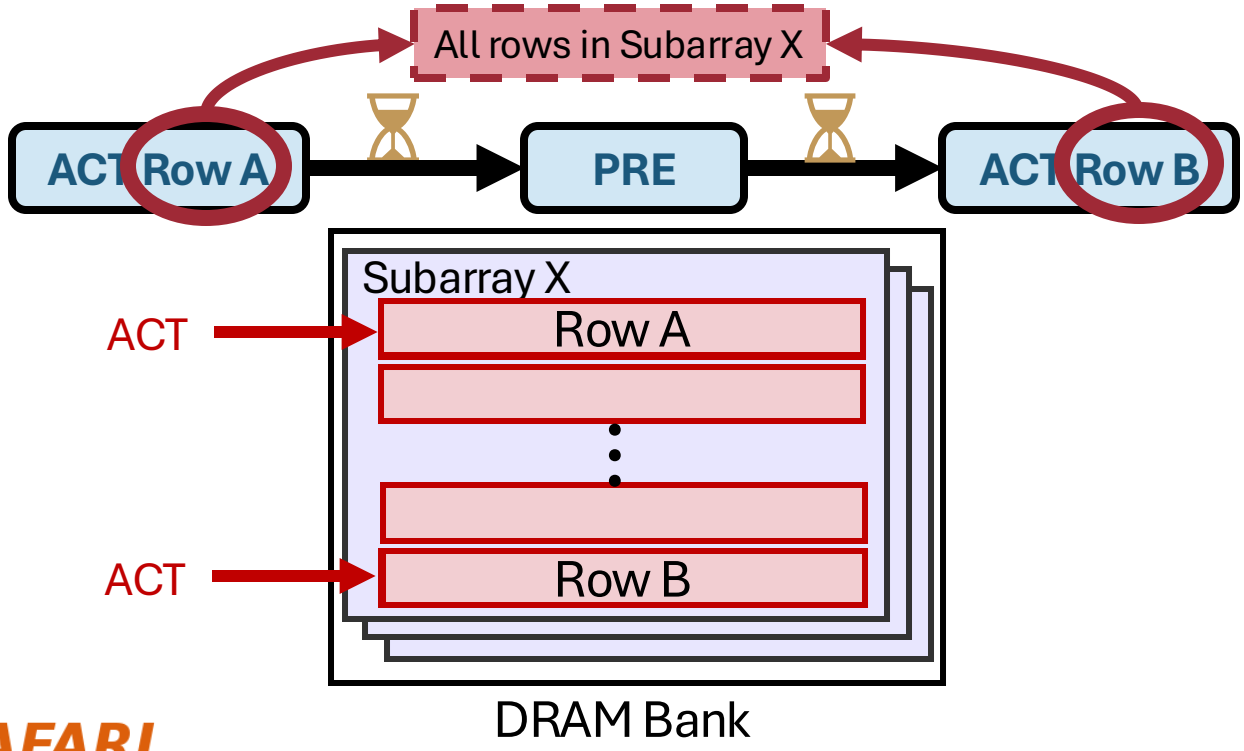
- 96 real DDR4 chips from SK Hynix
- Covering multiple die revision & chip density

Chip Mfr.	Module Mfr.	#Modules (#Chips)	Die Rev.	Chip Density	Chip Org.	Date[∇] year-week
SK Hynix	TimeTec	4 (32)	A	4Gb	x8	N/A
	TeamGroup	4 (32)	M	4Gb	x8	N/A
	SK Hynix	4 (32)	A	8Gb	x8	18-43

[∇] We report "N/A" if no date is marked on the label of a module.

Characterization Methodology

- Carefully sweep
  - Row addresses: Row A and Row B (>3M row pairs)
  - Timing parameters: Between ACT → PRE and PRE → ACT
  - Data Pattern: 0x00, 0xFF, and random
  - Temperature (°C): 50, 60, 70, and 80



Randomness Metric - I

- Measure randomness of bitstreams using **Shannon Entropy (SE)**

$$H(x) = -\sum_{i=1}^{2} p(x_i) \log_2 p(x_i)$$

Calculating probabilities:

→ Proportion of *logic-1* and *logic-0* values in the random bitstream

- Sample each bitline following SiMRA *1000 times* and calculate the bitline's Shannon Entropy

SE(*1111…111*) = 0

0 < **SE(*1001…010*)** < 1

Randomness Metric - II

- Calculate *cache block (512-bit) entropy*

$$\sum all\ bitline\ Shannon\ entropies\ in\ the\ cache\ block$$

- **Average cache block entropy**
  - *Average* entropy across all cache blocks in a module

Outline

Motivation & Goal

Generating Random Numbers via SiMRA

Testing Infrastructure

Real DRAM Chip Results

SiMRA-TRNG

Conclusion

Key Takeaways from Major Characterization Experiments

Takeaway 1

SiMRA's entropy generally **increases** as the number of simultaneously **activated rows increases**. This trend **varies across DRAM microarchitectures**

Takeaway 2

Data pattern significantly **affects** **SiMRA's** average cache block **entropy**

Takeaway 3

DRAM chip temperature significantly **affects** **SiMRA's** average cache block **entropy**

Key Takeaways from Major Characterization Experiments

Takeaway 1

SiMRA's entropy generally **increases** as the number of simultaneously **activated rows increases**. This trend **varies across DRAM microarchitectures**

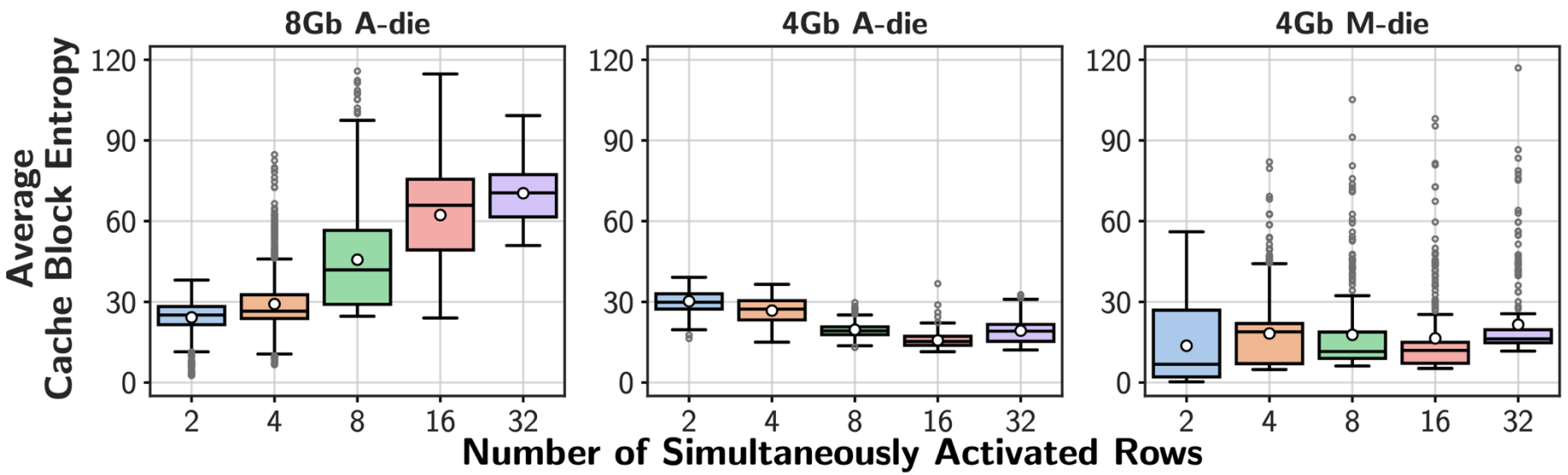
Takeaway 2

Data pattern significantly **affects** **SiMRA's** average cache block **entropy**

Takeaway 3

DRAM chip temperature significantly **affects** **SiMRA's** average cache block **entropy**

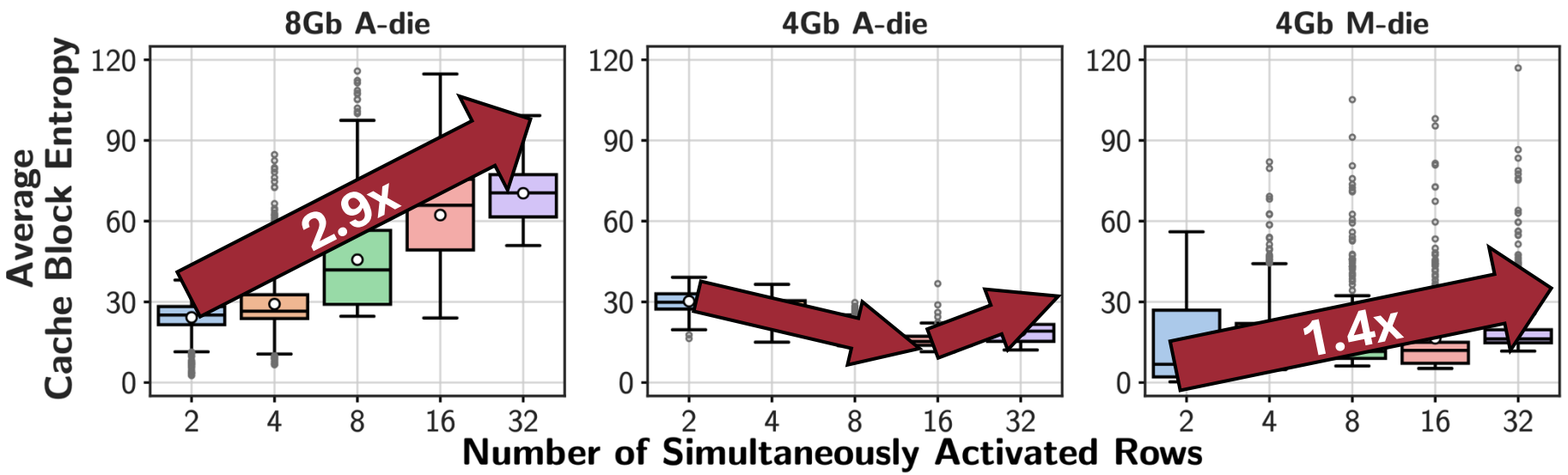
Average Entropy for All DRAM Chips



SiMRA generates **random values** with
all tested numbers of simultaneously activated rows

This trend is **consistent**
across all tested chip density & die revision pairs

Average Entropy for All DRAM Chips



In most of the tested chips, entropy increases as the number of simultaneously activated rows increases

However, not all chip density & die revision pairs have the same trend

Key Takeaways from Major Characterization Experiments

Takeaway 1

SiMRA's entropy generally **increases** as
the number of simultaneously **activated rows increases**.
This trend **varies across DRAM microarchitectures**

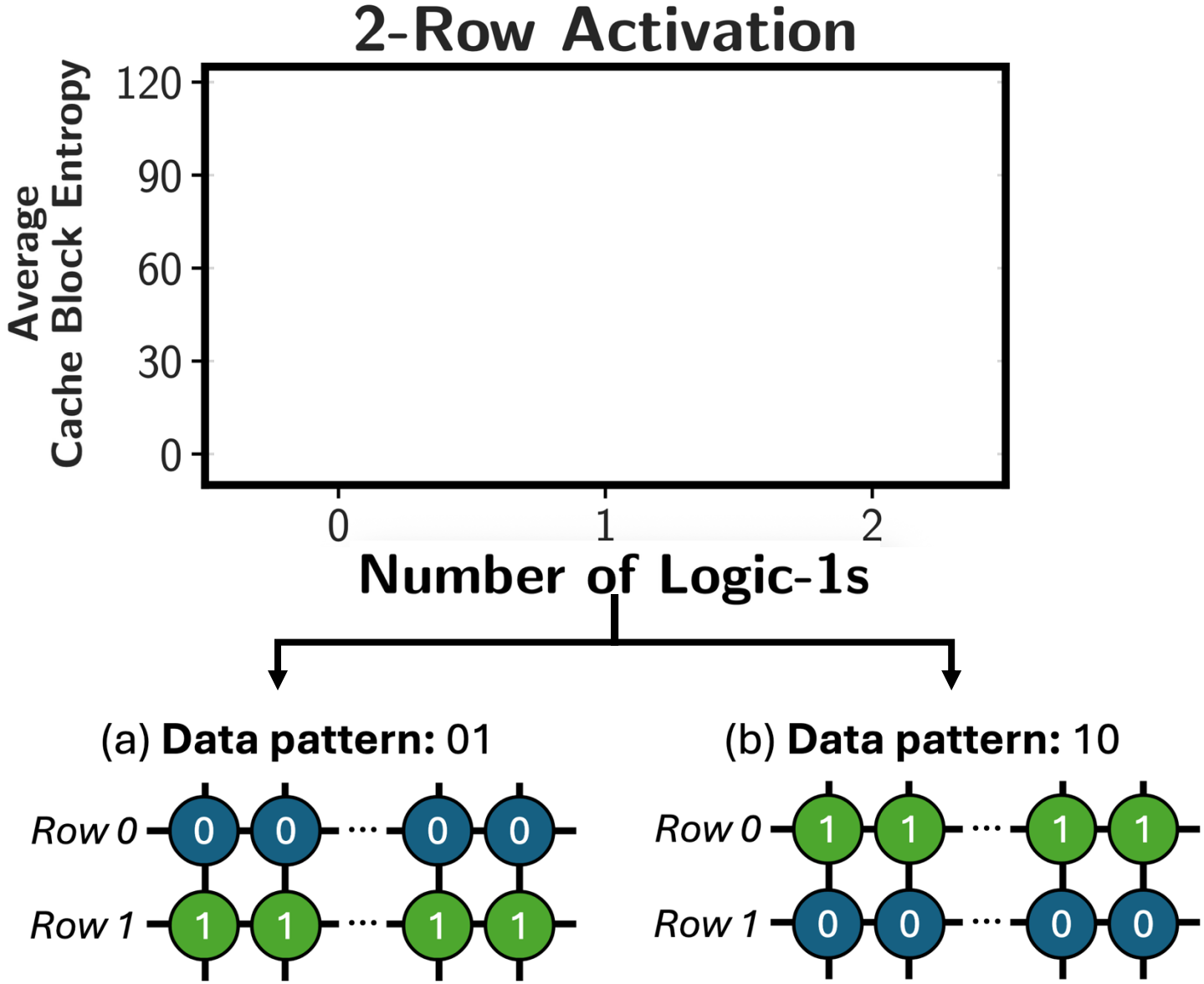
Takeaway 2

Data pattern significantly **affects**
SiMRA's average cache block **entropy**

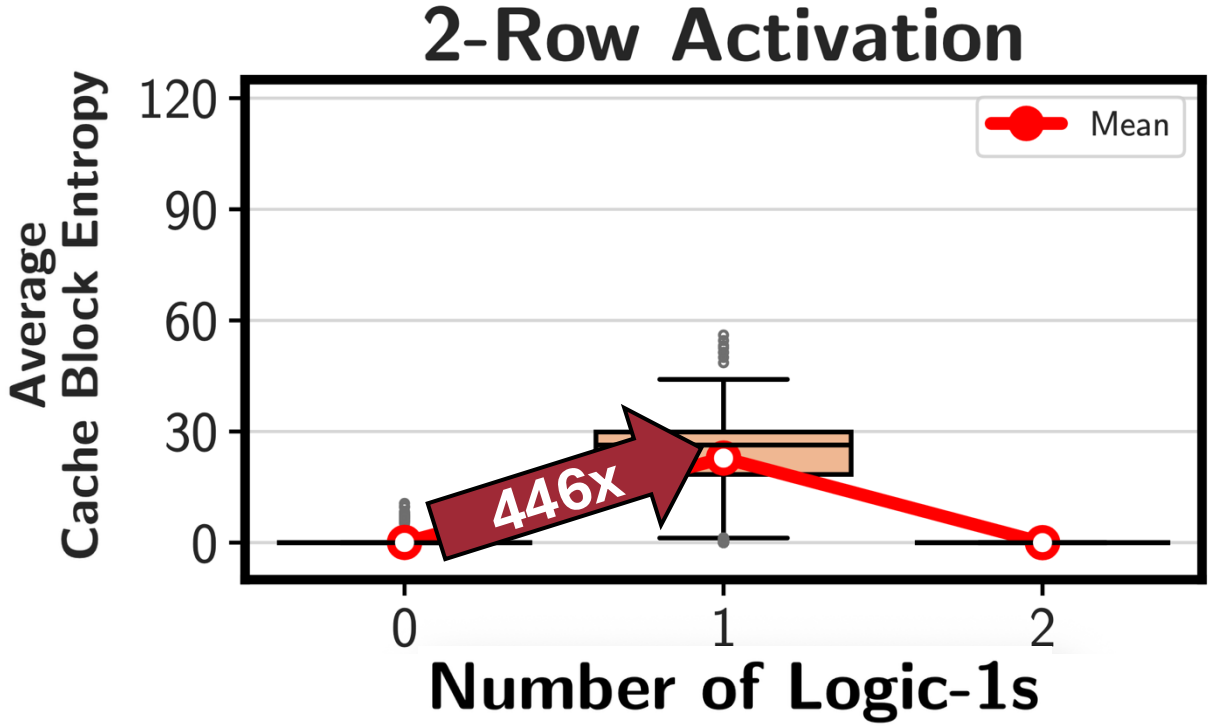
Takeaway 3

DRAM chip temperature significantly **affects**
SiMRA's average cache block **entropy**

Effect of Data Pattern

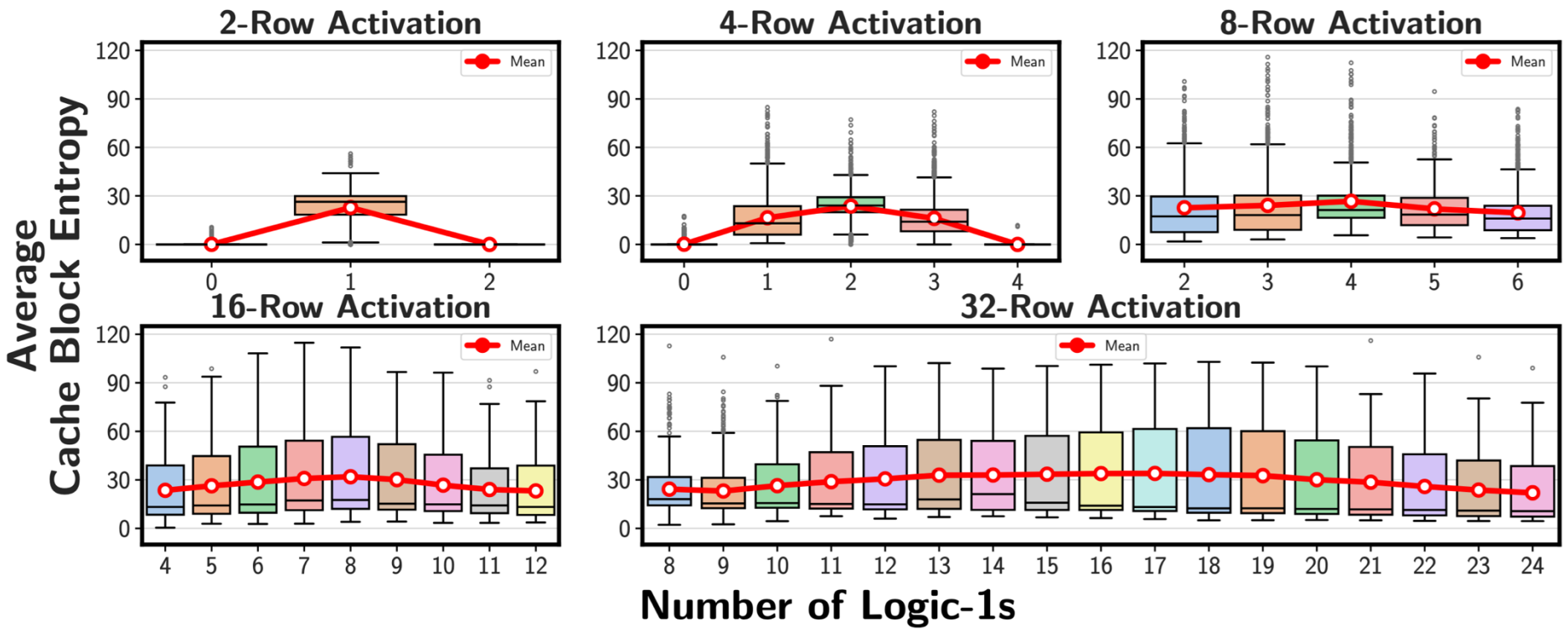


Effect of Data Pattern



Entropy significantly varies across tested number of logic-1s

Effect of Data Pattern



This observation is consistent for every number of simultaneous multiple row activations

Key Takeaways from Major Characterization Experiments

Takeaway 1

SiMRA's entropy generally **increases** as
the number of simultaneously **activated rows increases**.
This trend **varies across DRAM microarchitectures**

Takeaway 2

Data pattern significantly **affects**
SiMRA's average cache block **entropy**

Takeaway 3

DRAM chip temperature significantly **affects**
SiMRA's average cache block **entropy**

Outline

Motivation & Goal

Generating Random Numbers via SiMRA

Testing Infrastructure

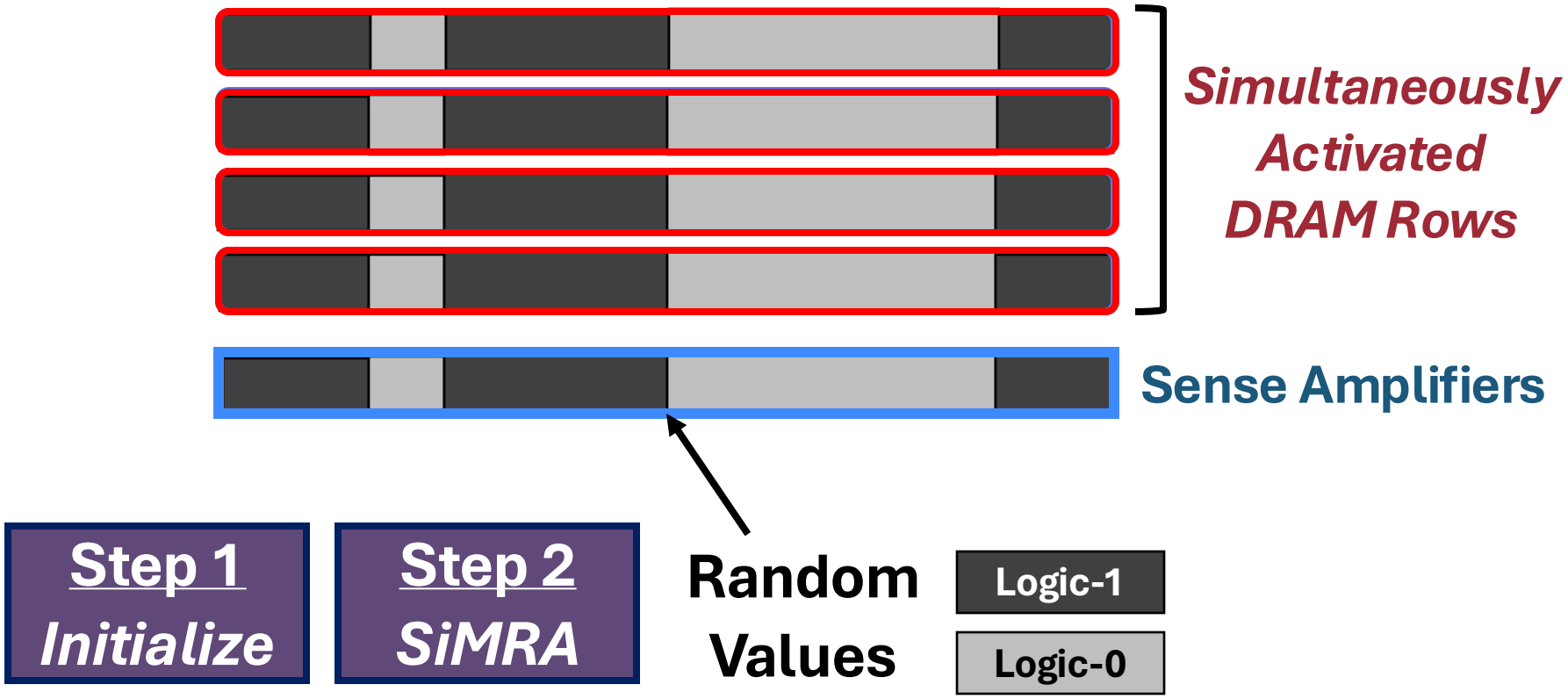
Real DRAM Chip Results

SiMRA-TRNG

Conclusion

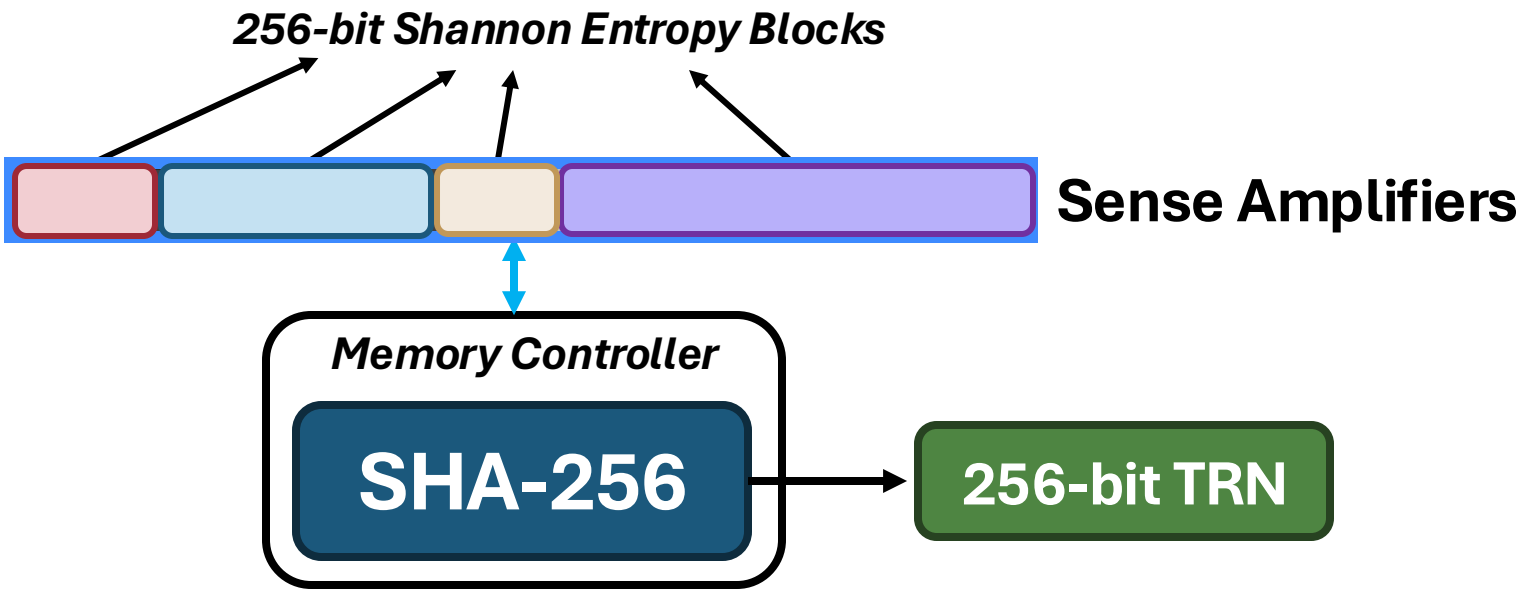
SiMRA-TRNG

Key Idea: Leverage random values on sense amplifiers generated by SiMRA operations as source of entropy



SiMRA-TRNG

Key Idea: Leverage random values on sense amplifiers generated by SiMRA operations as source of entropy



Step 1	Step 2	Step 3	Step 4
Initialize	*SiMRA*	*Read*	*Post-Process*

SiMRA-TRNG: Quality and Latency

SiMRA-TRNG's Quality

SiMRA-TRNG reliably produces high-quality true random bitstreams

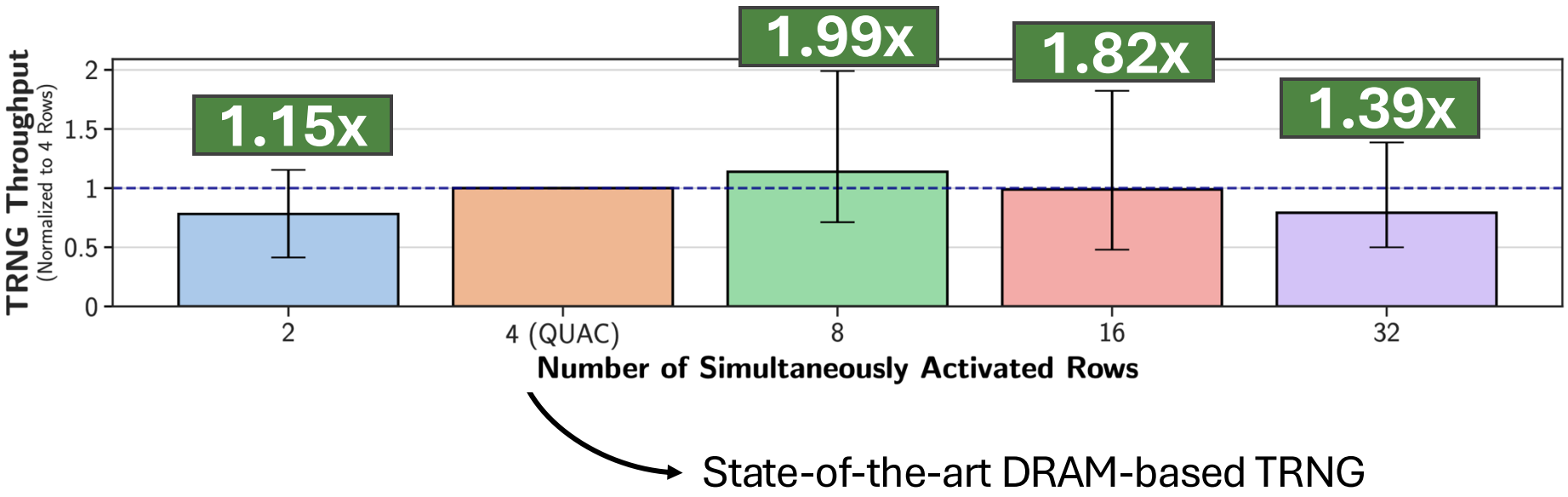
- Evaluated using **NIST Statistical Test Suite (NIST STS) tests**
  - SiMRA-TRNG **passes all** 15 NIST STS **tests for all N for SiMRA-N**

SiMRA-TRNG's Latency

{2, 4, 8, 16, 32}-row-based SiMRA-TRNG has 165ns, 255ns, 435ns, 795ns, and 1.51us of latency, respectively

- Measured latency of **each SiMRA-TRNG step**
  - Initialization, SiMRA, Read, Post-Process

SiMRA-TRNG: Throughput



SiMRA-based TRNG designs outperform the state-of-the-art DRAM-based TRNG

More in the Paper

- **More characterization results**
  - Effect of temperature
  - Spatial variation across DRAM subarrays
  - Many hypotheses based on the empirical observations
- **Details and analyses on SiMRA-TRNG**
  - Quality
  - Latency
  - Throughput
- **Discussion on the limitations of testing real DRAM chips**

Available on arXiv

In-DRAM True Random Number Generation Using Simultaneous Multiple-Row Activation: An Experimental Study of Real DRAM Chips

İsmail Emir Yüksel[§] Ataberk Olgun[§] F. Nisa Bostancı[§] Oğuzhan Canpolat[§]
Geraldo F. Oliveira[§] Mohammad Sadrosadati[§] A. Giray Yağlıkçı[§Γ] Onur Mutlu[§]

[§]ETH Zürich [Γ]CISPA

In this work, we experimentally demonstrate that it is possible to generate true random numbers at high throughput and low latency in commercial off-the-shelf (COTS) DRAM chips by leveraging simultaneous multiple-row activation (SiMRA) via an extensive characterization of 96 DDR4 DRAM chips. We rigorously analyze SiMRA's true random generation potential in terms of entropy, latency, and throughput for varying numbers of simultaneously activated DRAM rows (i.e., 2, 4, 8, 16, and 32), data patterns, temperature levels, and spatial variations. Among our 11 key experimental observations, we highlight four key results. First, we evaluate the quality of our TRNG

computing systems, ranging from microcontrollers to supercomputers, and systems that employ Processing-in-Memory (PIM) [29–46]. Several prior works [47–53] demonstrate that by leveraging the analog operational properties of DRAM, it is possible to generate true random numbers in commercial off-the-shelf (COTS) DRAM chips.

Recent works [35, 53–57] experimentally demonstrate a new phenomenon in COTS DRAM chips that could potentially be used as a source of entropy: *simultaneous multiple-row activation (SiMRA)*. These works [35, 53–57] carefully engineer a sequence of DRAM commands that allows the DRAM chip

https://arxiv.org/pdf/2510.20269

Outline

Conclusion

We demonstrate and extensively study the potential of using simultaneous multiple-row activation (SiMRA) to generate true random numbers

We characterize 96 off-the-shelf DDR4 DRAM chips

- **SiMRA** generates **random values** in **all tested DRAM chips**
- **Data pattern** significantly **affects** SiMRA's **entropy**
- **Temperature** has a **significant impact** on SiMRA's **entropy**

We develop a new DRAM-based TRNG: SiMRA-TRNG

- **Passes all** NIST STS **tests** and generates **true random numbers**
- **Outperforms** the **state-of-the-art DRAM-based TRNG**

We hope these empirical results demonstrate the promising potential of using DRAM as a TRNG substrate

In-DRAM True Random Number Generation Using Simultaneous Multiple-Row Activation An Experimental Study of Real DRAM Chips

İsmail Emir Yüksel

A. Olgun F. N. Bostancı O. Canpolat

G. F. Oliveira M. Sadrosadati A. G. Yağlıkçı O. Mutlu

ETH zürich SAFARI CISPA

In-DRAM True Random Number Generation Using Simultaneous Multiple-Row Activation An Experimental Study of Real DRAM Chips

Backup Slides

İsmail Emir Yüksel

A. Olgun F. N. Bostancı O. Canpolat

G. F. Oliveira M. Sadrosadati A. G. Yağlıkçı O. Mutlu

ETH zürich SAFARI CISPA

Limitations of Tested COTS DRAM Chips (I)

- **Some COTS DRAM chips do not support all in-DRAM operations**
  - We do not observe simultaneous multiple-row activation in tested 64 Samsung & Micron chips
  - Hypothesis
    - Internal DRAM circuitry ignores the PRE command or the second ACT command when the timing parameters are greatly violated

If such a limitation were not imposed, we believe these DRAM chips are also fundamentally capable of performing the operations we examine in this work

Limitations of Tested COTS DRAM Chips (II)

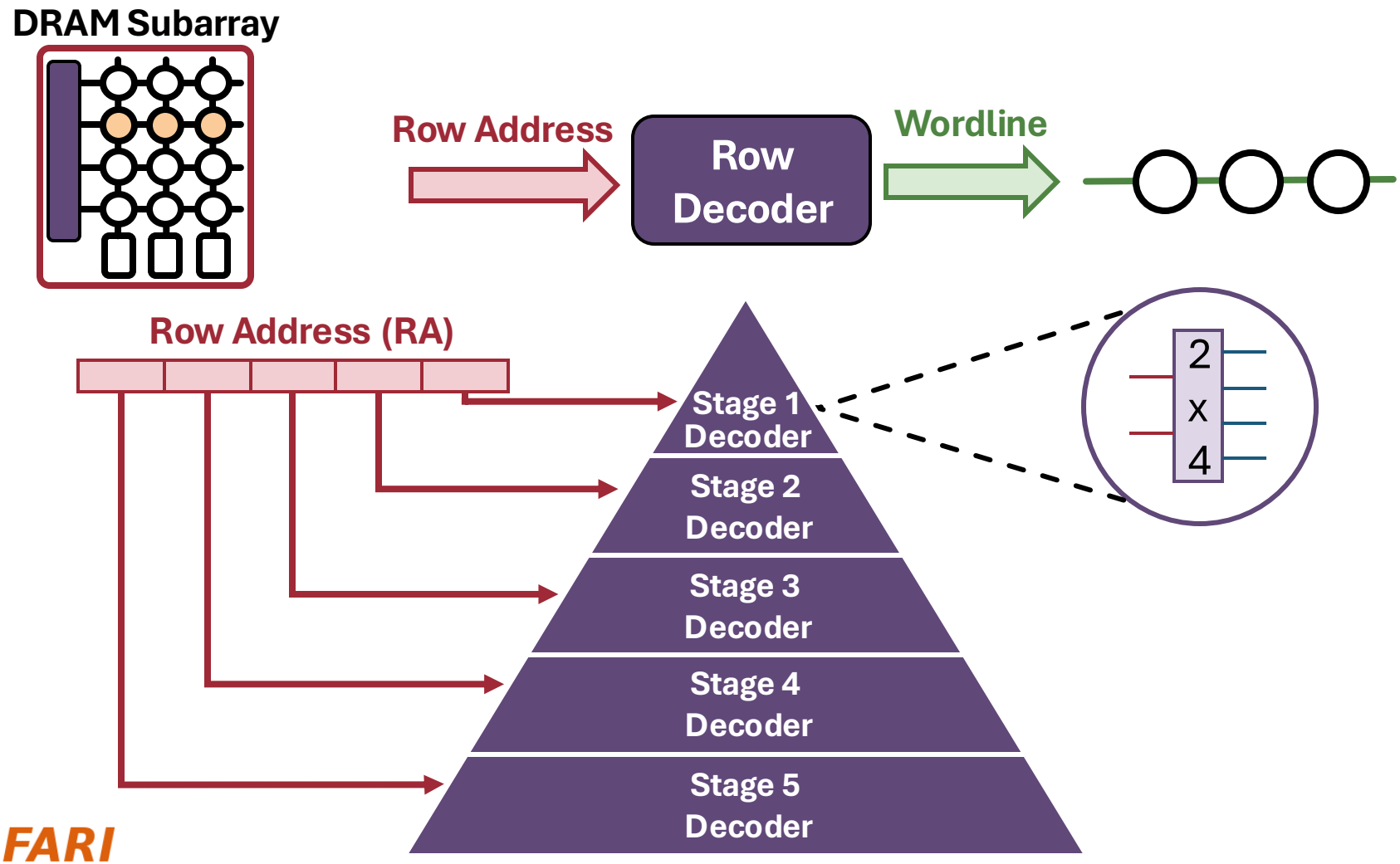
- **Tested COTS DRAM chips support only consecutive two row activation and simultaneous activation of 2, 4, 8, 16, and 32 rows**

  - Hypothesis
    - This is due to our current infrastructure limitations, where we can issue DRAM commands at intervals of only 1.5ns.

    - Having fine-grained control on timing would allow us to deassert/assert desired intermediate signals in the row decoder circuitry

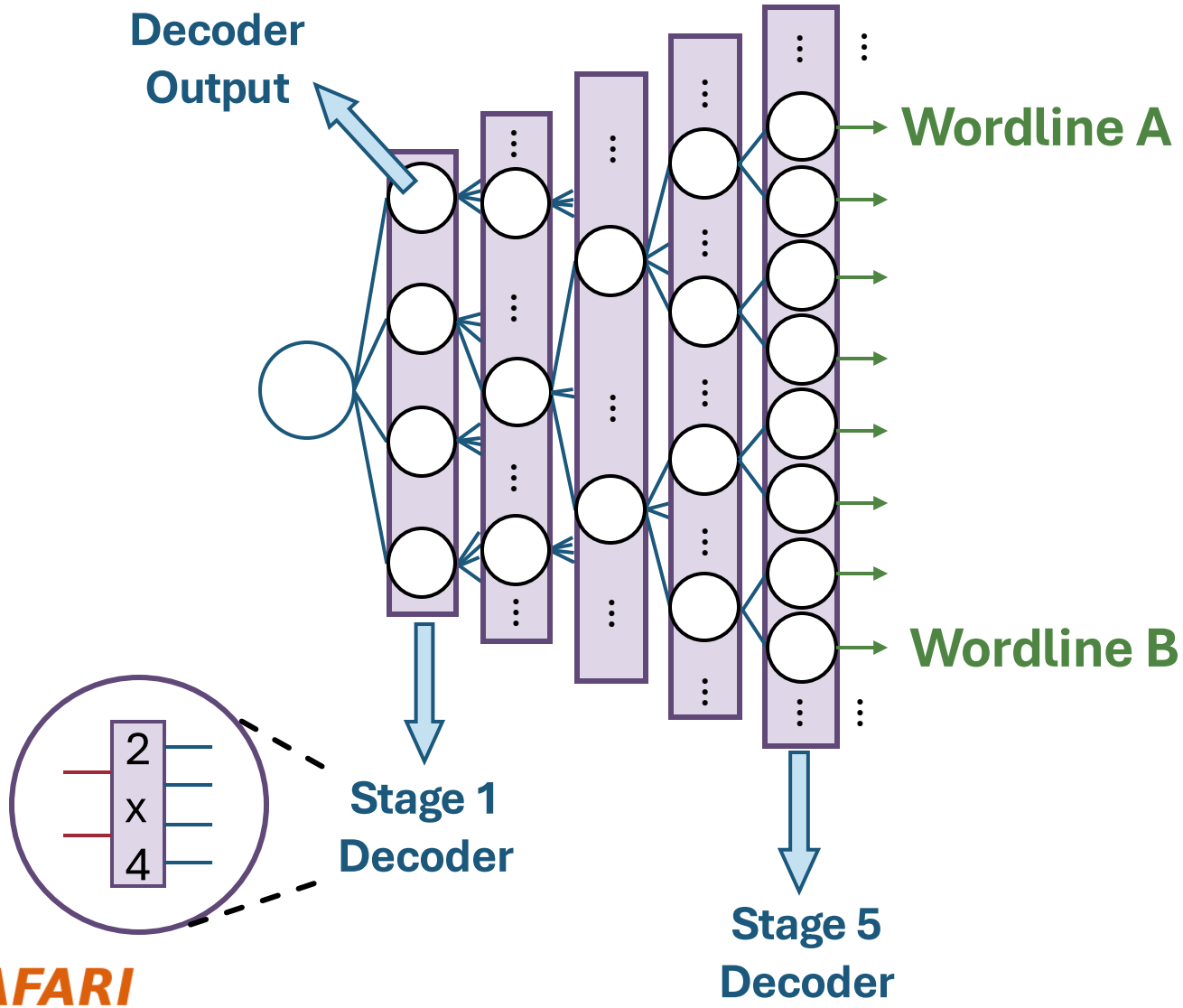
Hypothesis: Row Decoder Circuitry

Simultaneous many-row activation is possible
due to the **hierarchical DRAM row decoder design**

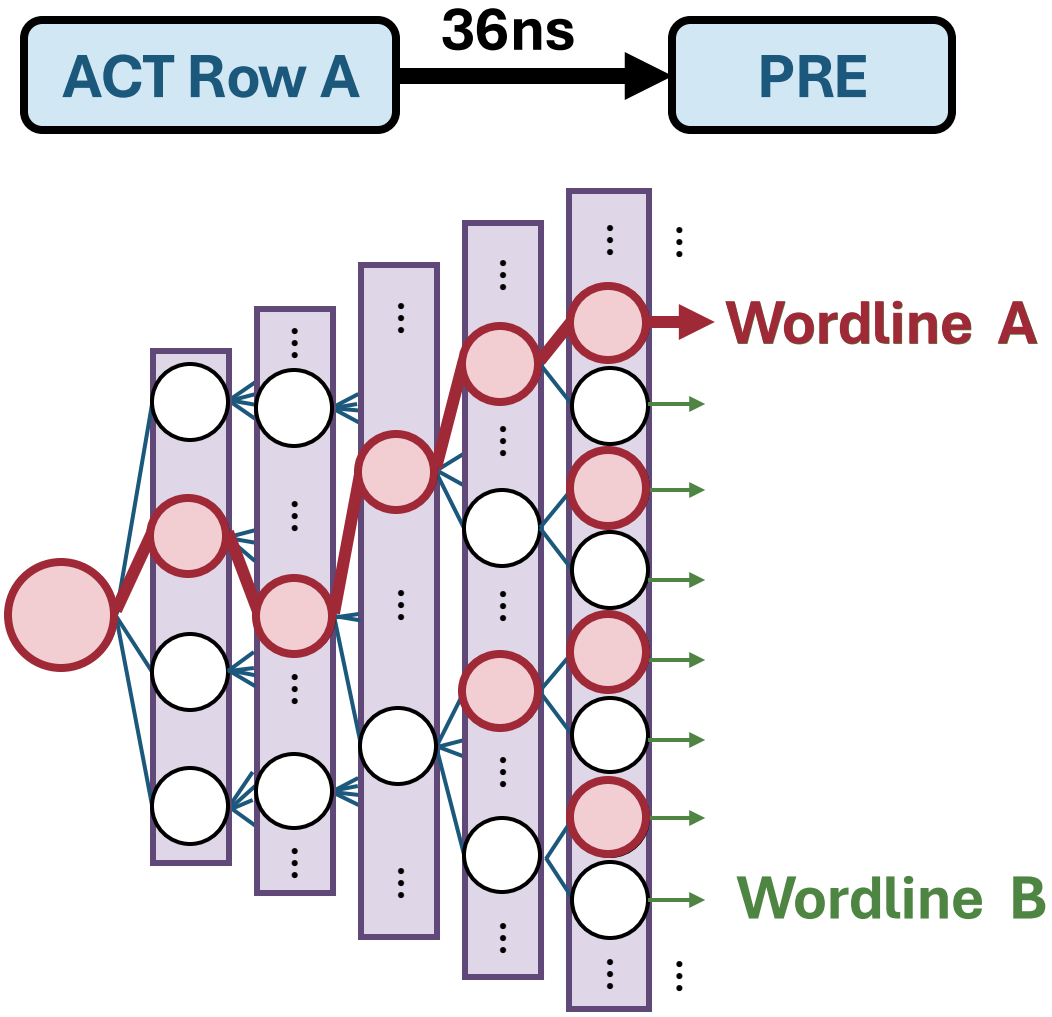


Row Decoder: A Tree Example

- We can visualize the hierarchical row decoder circuitry as a tree



Activating a Single Row



1. Decode Row A address

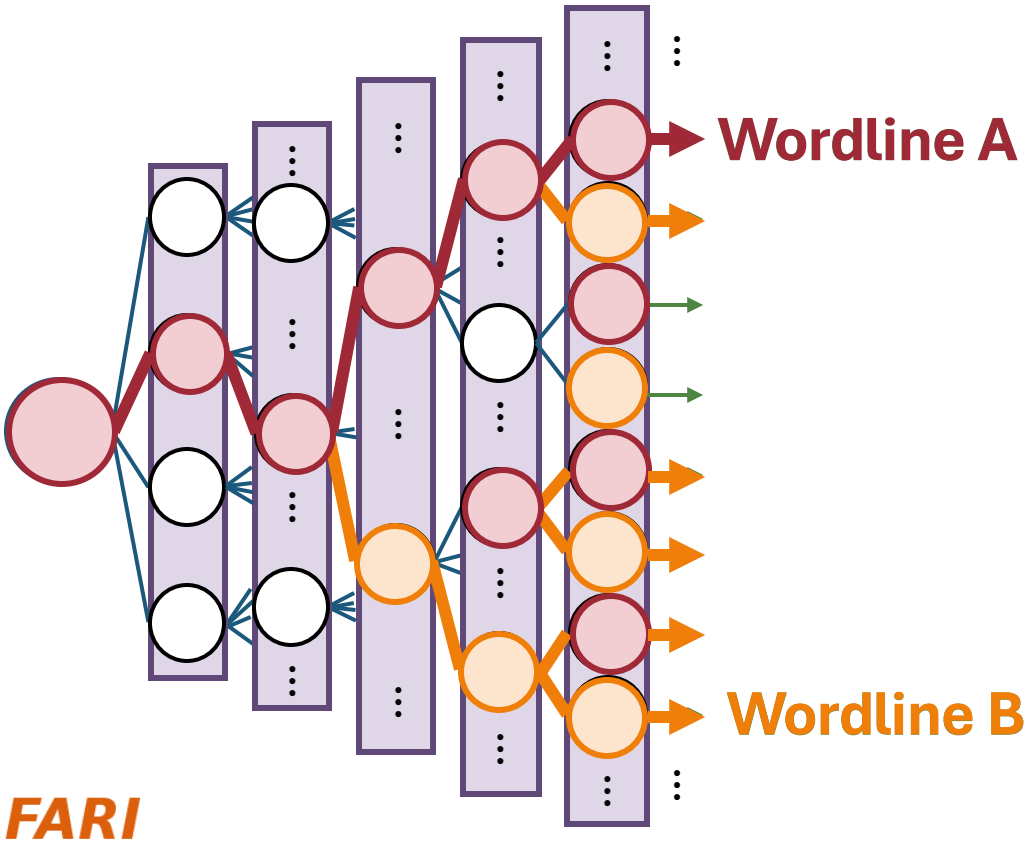
2. Assert corresponding signals

3. Follow asserted nodes path from root to leaf

4. Assert wordline

Activating Many Rows: A Walkthrough

Back-to-back ACT commands with violated timings **asserts many more signals** in the row decoder



1. Keep the previously asserted signals active

2. Decode Row B & assert corresponding signals

3. Follow newly asserted nodes' path

4. Simultaneously activate many rows

Hypothesis: Row Decoder Circuitry

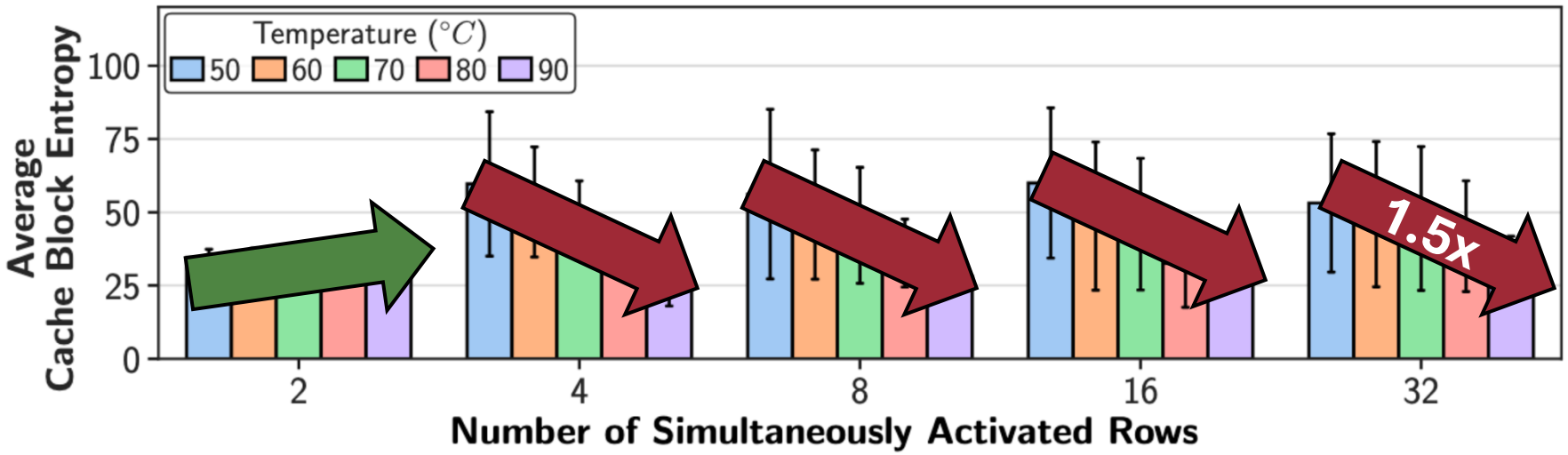
Simultaneous Many-Row Activation in Off-the-Shelf DRAM Chips: Experimental Characterization and Analysis

İsmail Emir Yüksel[1] Yahya Can Tuğrul[1,2] F. Nisa Bostancı[1] Geraldo F. Oliveira[1]
A. Giray Yağlıkçı[1] Ataberk Olgun[1] Melina Soysal[1] Haocong Luo[1]
Juan Gómez-Luna[1] Mohammad Sadrosadati[1] Onur Mutlu[1]
[1]*ETH Zürich* [2]*TOBB University of Economics and Technology*

(More discussions & hypotheses in the paper)

https://arxiv.org/pdf/2405.06081

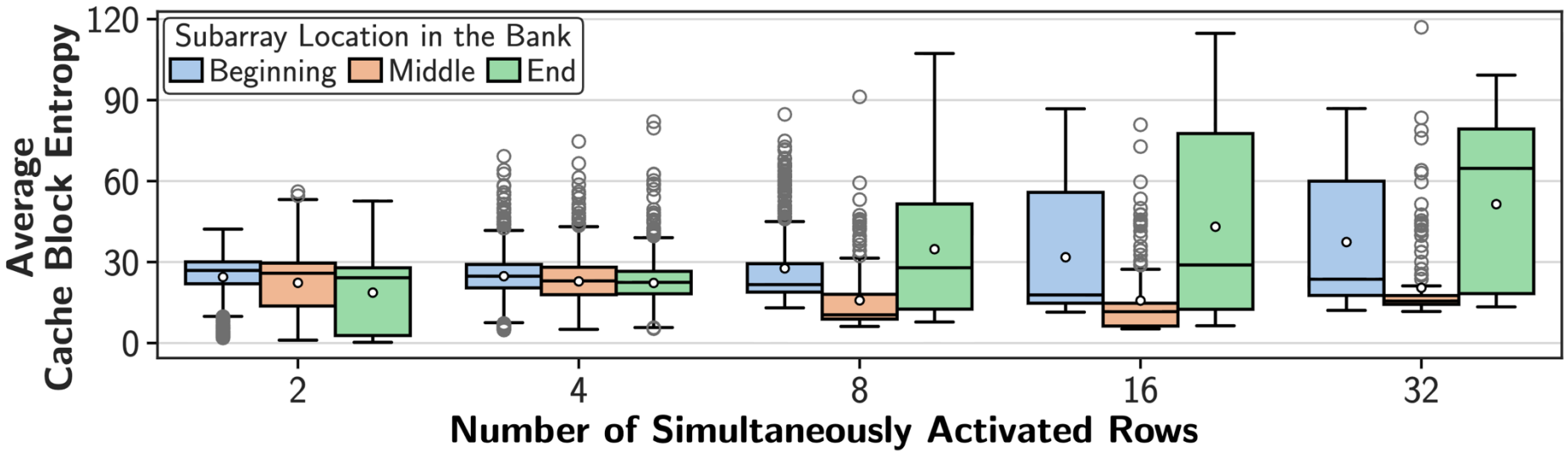
Effect of Temperature



As temperature increases SiMRA's entropy tends to decrease

This observation is consistent across most of the number of simultaneously activated rows (except 2-row activation)

Spatial Variation



SiMRA's entropy significantly varies based on the subarray's location in a bank

Each number of row activations has a different entropy variation trend

SiMRA-TRNG: Quality Evaluation

- **Methodology:** Use NIST Statistical Test Suite (NIST STS) tests (15 tests)
  - Extract a 1Gb bitstream by repeatedly performing SiMRA
  - Post process the bitstream with SHA-256

- **Results:**
  - **All** tested number of simultaneously activated rows (2-, 4-, 8-, 16-, and 32-row activation) **pass all** 15 NIST STS **tests**

SiMRA-TRNG **reliably** produces **high-quality true** random bitstreams

SiMRA-TRNG: Latency Evaluation

- Measured each step of SiMRA-TRNG design
  - **Step 1: Initialize**
    - Performing N RowClone operation for N-row activation
  - **Step 2: SiMRA**
    - Executing DRAM commands that activates N rows simultaneously (ACT – PRE – ACT)
  - **Step 3: Read**
    - Reading random values from sense amplifiers
  - **Step 4: Post-processing**
    - Post processing the random bitstream using SHA-256

Latencies of SiMRA-TRNG designs based on 2-, 4-, 8-, 16-, and 32-row activation are 165ns, 255ns, 435ns, 795ns, and 1.51us, respectively

Throughput & Latency Comparison

Proposal	Entropy Source	TRNG Throughput	256-bit TRNG Latency
QUAC-TRNG	Quadruple ACT	13.76 Gb/s	274 ns
Talukder+ [15]	Precharge Failure	0.68 - 6.13 Gb/s	249 ns - 201 ns
D-RaNGe [88]	Activation Failure	0.92 - 9.73 Gb/s	260 ns - 36 ns
D-PUF [150]	Retention Failure	0.20 Mb/s	40 s
DRNG [47]	DRAM Start-up	N/A	700 µs
Keller+ [81]	Retention Failure	0.025 Mb/s	40 s
Pyo+ [126]	DRAM Cmd Schedule	2.17 Mb/s	112.5 µs