Functionally-Complete Boolean Logic in Real DRAM Chips

Experimental Characterization and Analysis

Ismail Emir Yüksel

Yahya C. Tugrul Ataberk Olgun F. Nisa Bostancı

A. Giray Yaglıkçı Geraldo F. Oliveira Haocong Luo

Juan Gómez–Luna Mohammad Sadr Onur Mutlu

SAFARI

ETH zürich

Executive Summary

- **Motivation: Processing-using-DRAM** can alleviate the performance and energy bottlenecks caused by **data movement**
  - **Prior works** show that **existing DRAM chips** can perform **three-input majority** and **two-input AND and OR** operations
- **Problem: Proof-of-concept** demonstrations on **commercial off-the-shelf (COTS)** DRAM chips do not provide
  - **functionally-complete** operations (e.g., NAND or NOR)
  - **NOT operation**
  - **AND** and **OR** operations with **more than two inputs**
- **Experimental Study: 256 DDR4 chips** from **two major manufacturers**
- **Key Results:**
  - COTS DRAM chips can perform **NOT** and **{2, 4, 8, 16}-input AND, NAND, OR, and NOR** operations with **very high reliability (>94% success rate)**
  - **Data pattern** and **temperature** only slightly affect the reliability of these operations **(<1.98% decrease in success rate)**

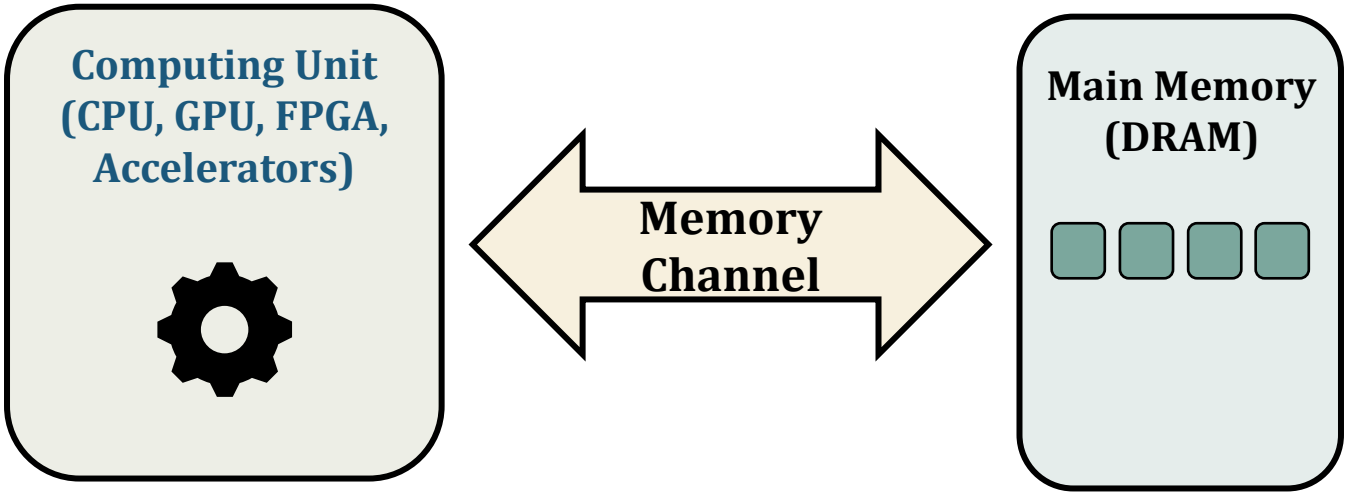
Outline

- Background
- Goal & Overview
- Experimental Methodology
- Multiple-Row Activation in Neighboring Subarrays
- NOT Operation
- AND, NAND, OR, and NOR Operations
- Conclusion

Outline

Data Movement Bottleneck

- Today's computing systems are processor centric
- All data is processed in the processor → at great system cost



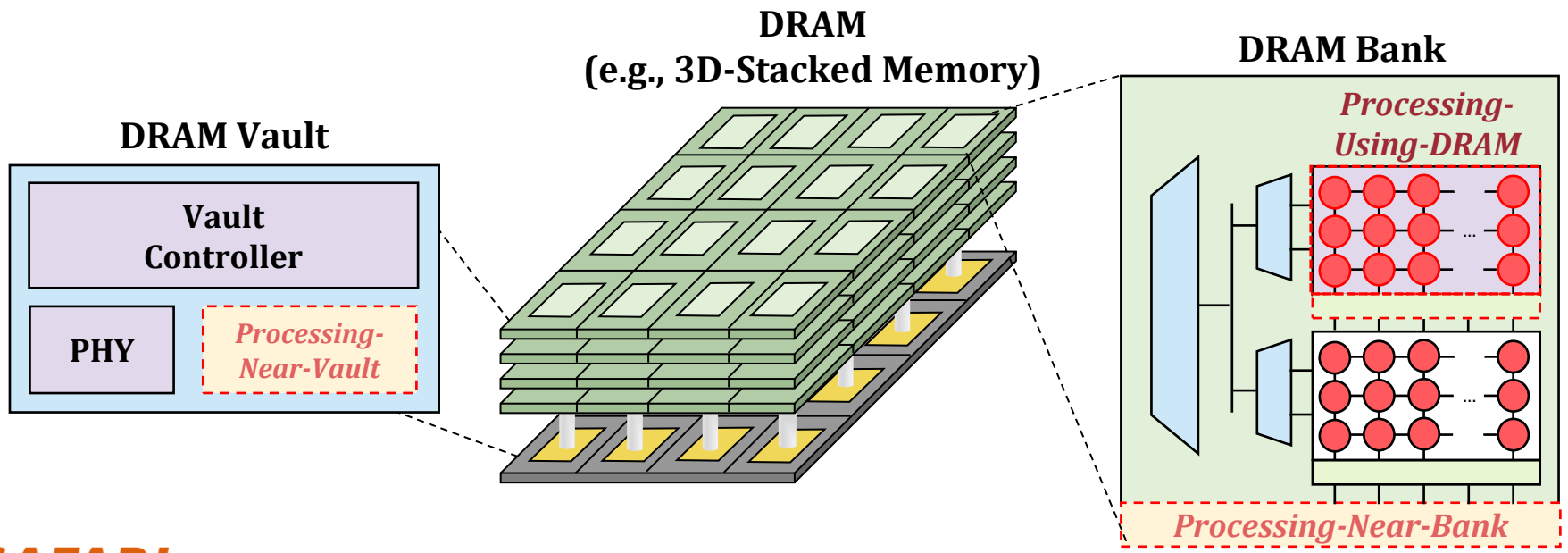
More than 60% of the total system energy is spent on data movement[1]

[1] A. Boroumand et al., "Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks," ASPLOS, 2018

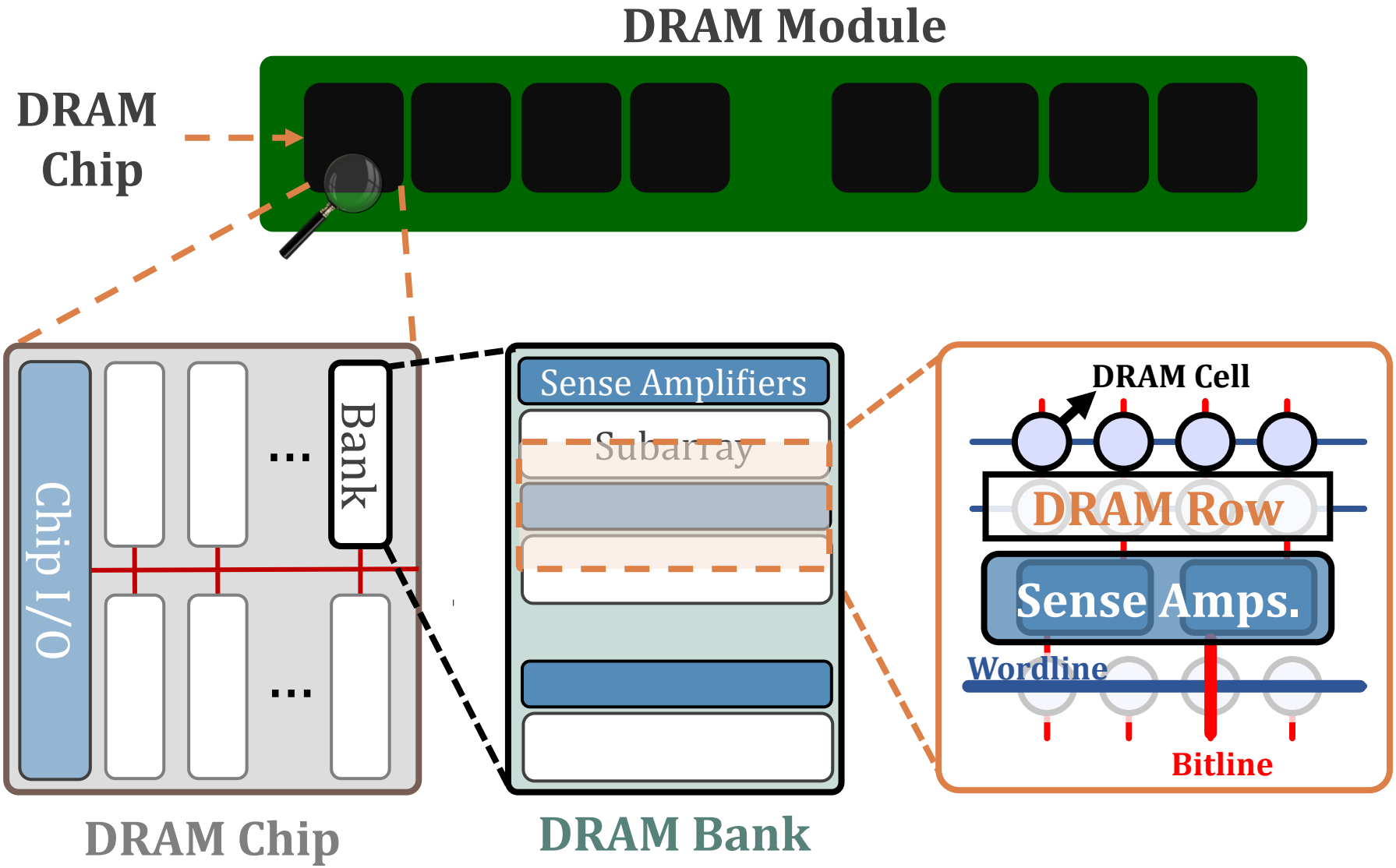
Processing-In-Memory (PIM)

Two main approaches for Processing-In-Memory:

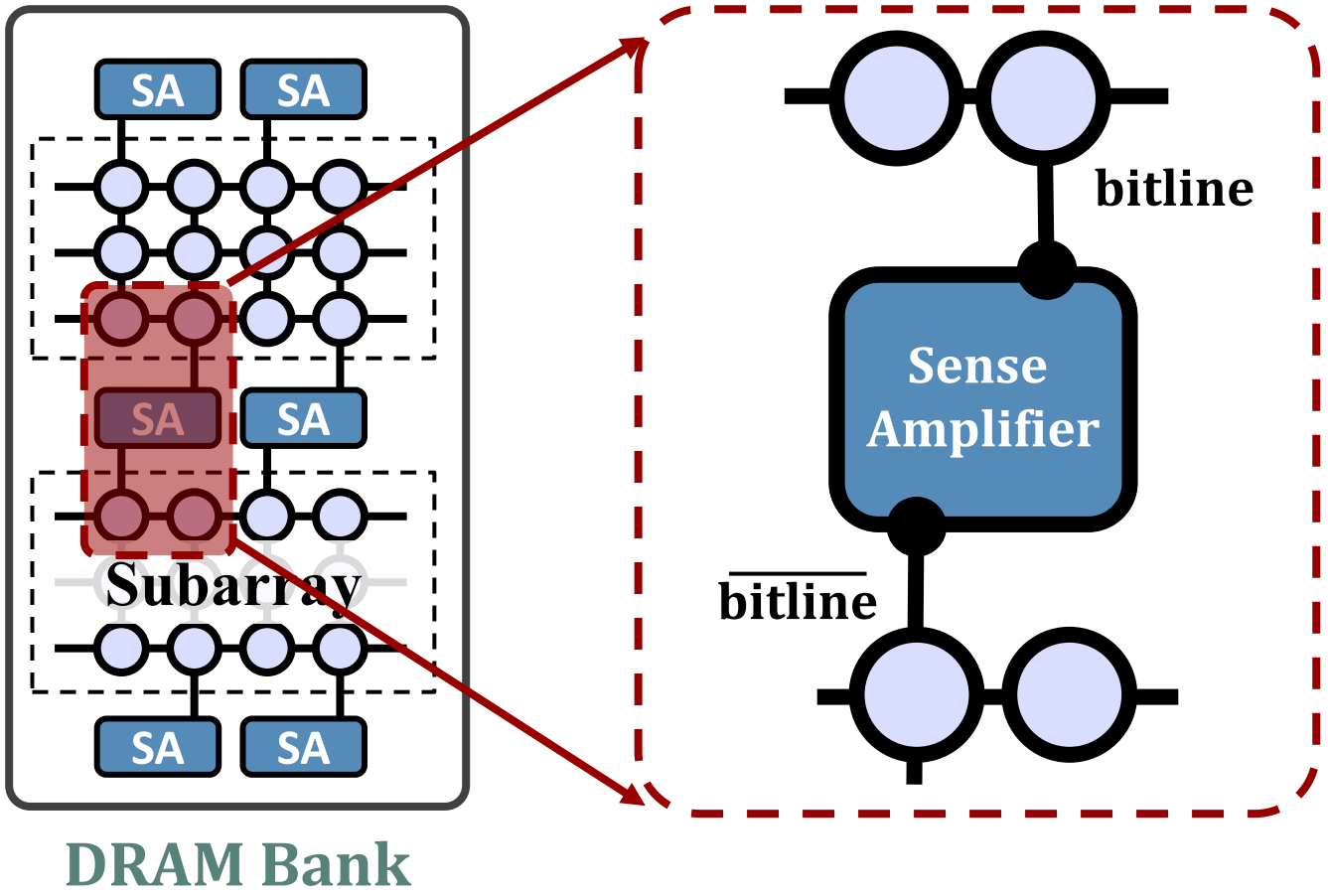
1. **Processing-Near-Memory**: PIM logic is added near the memory arrays or to the logic layer of 3D-stacked memory
2. **Processing-Using-Memory**: uses the analog operational principles of memory cells to perform computation



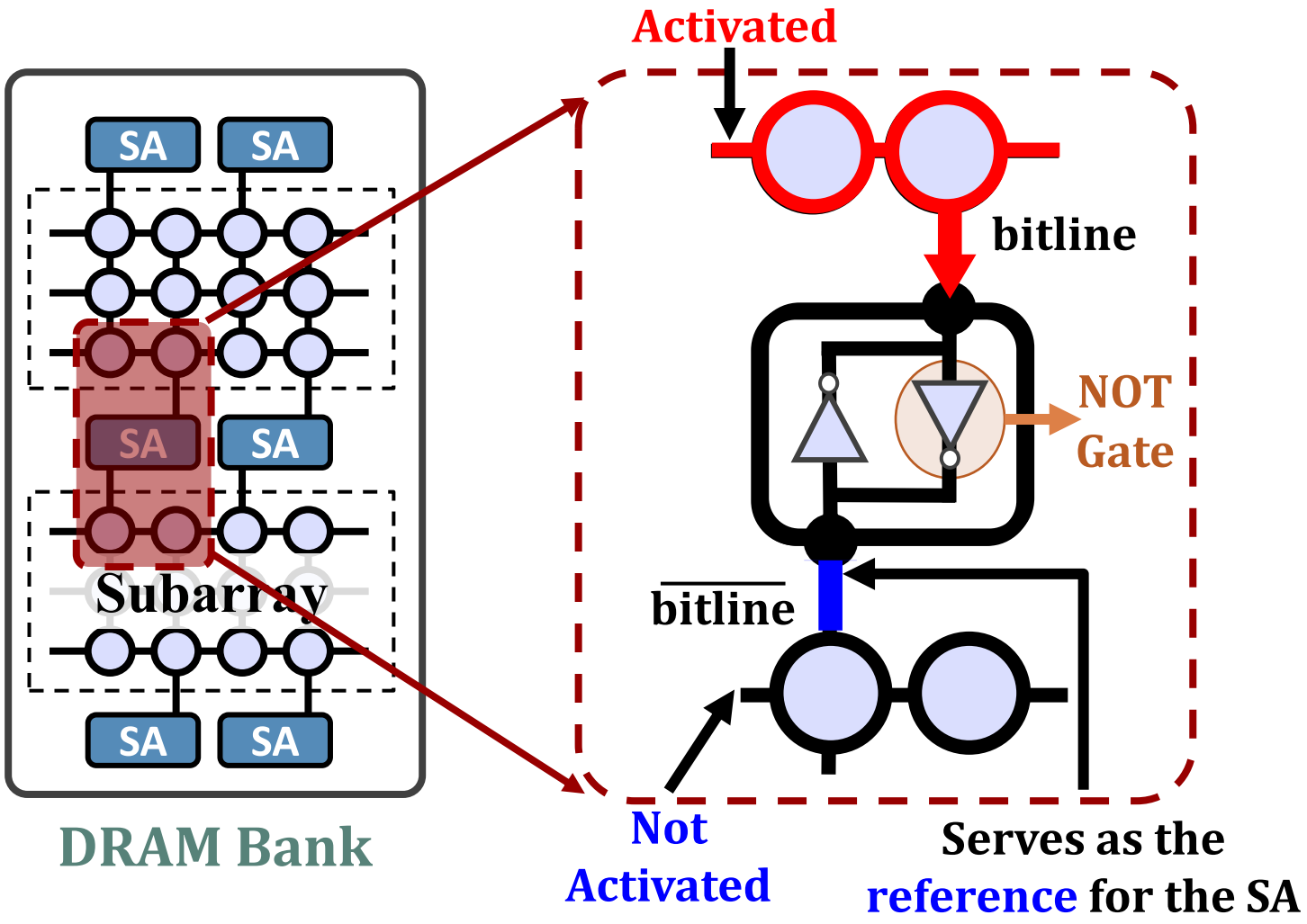
DRAM Organization



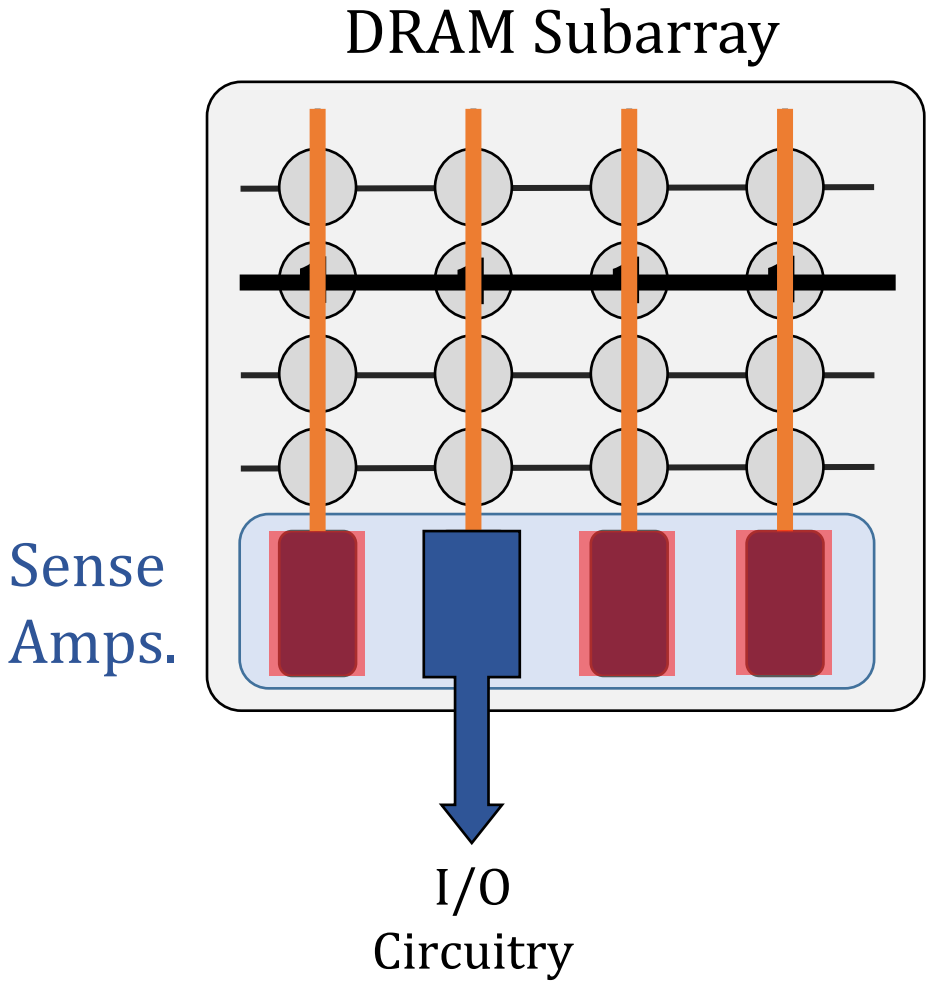
DRAM Open Bitline Architecture



DRAM Open Bitline Architecture



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 subarray for a new ACTIVATE

Outline

- Background
- Goal & Overview
- Experimental Methodology
- Multiple-Row Activation in Neighboring Subarrays
- NOT Operation
- AND, NAND, OR, and NOR Operations
- Conclusion

Our Goal

Understand the **capability** of COTS DRAM chips
beyond just storing data

Rigorously **characterize**
the reliability of this capability

COTS: Commercial Off-The-Shelf

The Capability of COTS DRAM Chips

We **demonstrate** that **COTS DRAM chips:**

1 Can **simultaneously activate** up to **48 rows** in **two neighboring subarrays**

2 Can perform **NOT operation** with up to **32 output operands**

3 Can perform up to **16-input AND, NAND, OR, and NOR** operations

Outline

Background

Goal & Overview

Experimental Methodology

Multiple-Row Activation in Neighboring Subarrays

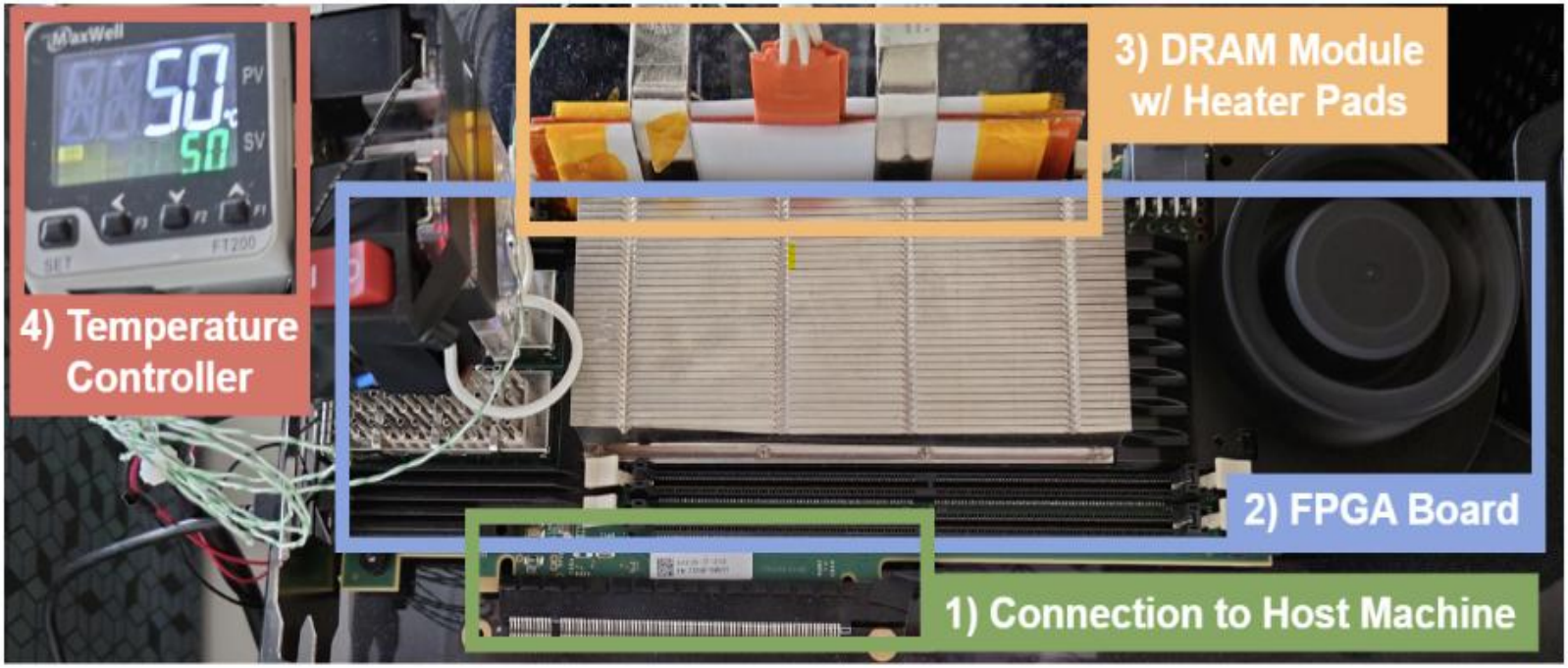
NOT Operation

AND, NAND, OR, and NOR Operations

Conclusion

DRAM Testing Infrastructure

- Developed from DRAM Bender [Olgun+, TCAD'23]*
- Fine-grained control over DRAM commands, timings, and temperature



*Olgun et al., "DRAM Bender: An Extensible and Versatile FPGA-based Infrastructure to Easily Test State-of-the-art DRAM Chips," TCAD, 2023.

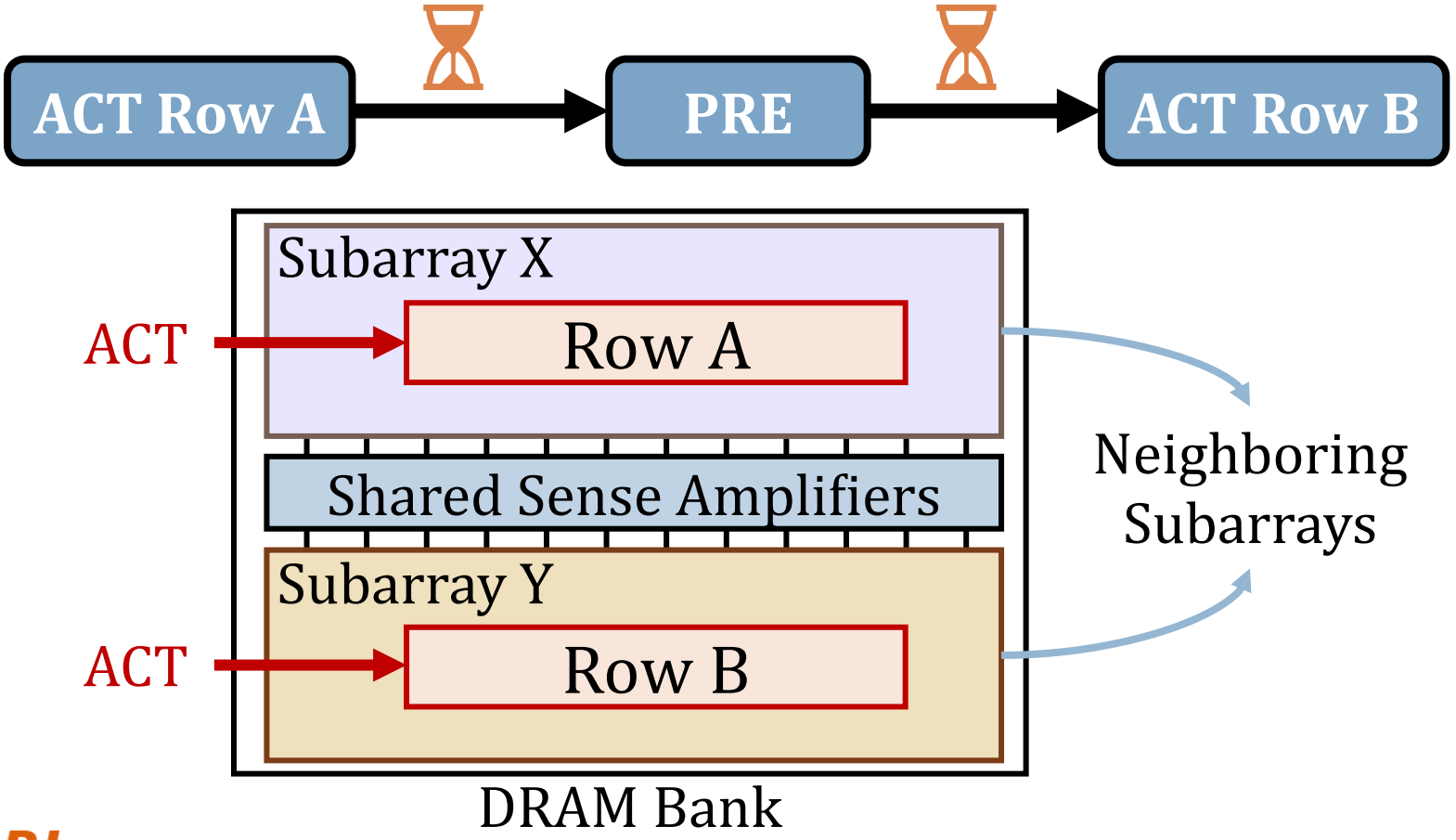
DRAM Chips Tested

- 256 DDR4 chips from two major DRAM manufacturers
- Covers different die revisions and chip densities

Chip Mfr.	#Modules (#Chips)	Die Rev.	Mfr. Date[a]	Chip Density	Chip Org.	Speed Rate
SK Hynix	9 (72)	M	N/A	4Gb	x8	2666MT/s
	5 (40)	A	N/A	4Gb	x8	2133MT/s
	1 (16)	A	N/A	8Gb	x8	2666MT/s
	1 (32)	A	18-14	4Gb	x4	2400MT/s
	1 (32)	A	16-49	8Gb	x4	2400MT/s
	1 (32)	M	16-22	8Gb	x4	2666MT/s
Samsung	1 (8)	F	21-02	4Gb	x8	2666MT/s
	2 (16)	D	21-10	8Gb	x8	2133MT/s
	1 (8)	A	22-12	8Gb	x8	3200MT/s

Testing Methodology

- Carefully sweep:
  - Row addresses: Row A and Row B
  - Timing parameters: Between ACT → PRE and PRE → ACT



Outline

Background

Goal & Overview

Experimental Methodology

Multiple-Row Activation in Neighboring Subarrays

NOT Operation

AND, NAND, OR, and NOR Operations

Conclusion

The Capability of COTS DRAM Chips

We **demonstrate** that **COTS DRAM chips:**

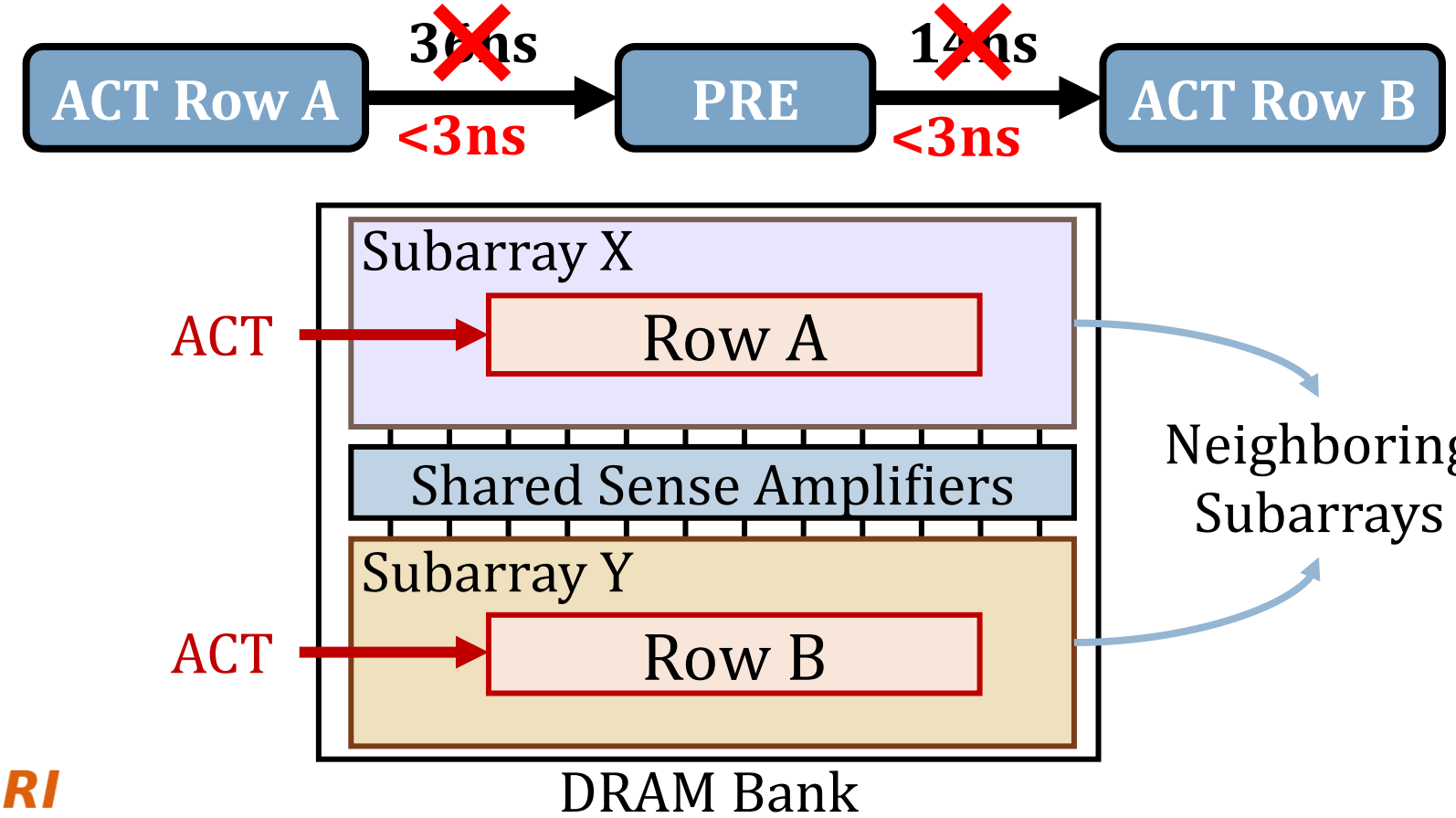
1 Can **simultaneously activate** up to **48 rows** in **two neighboring subarrays**

2 Can perform **NOT operation** with **up to 32** output operands

3 Can perform **up to 16-input AND, NAND, OR, and NOR** operations

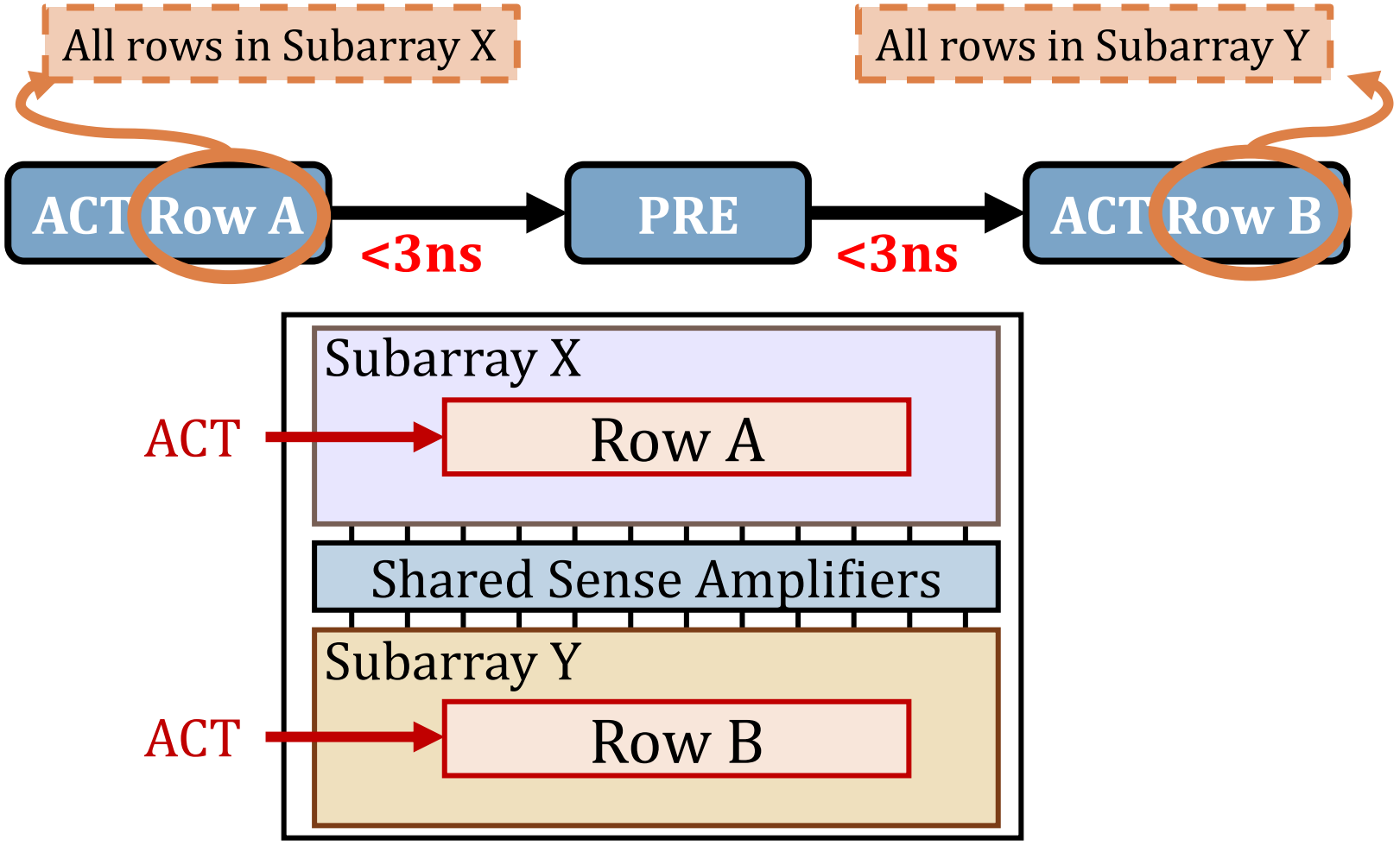
Key Observation

Activating two rows in **quick succession**
can **simultaneously** activate
multiple rows in neighboring subarrays



Characterization Methodology

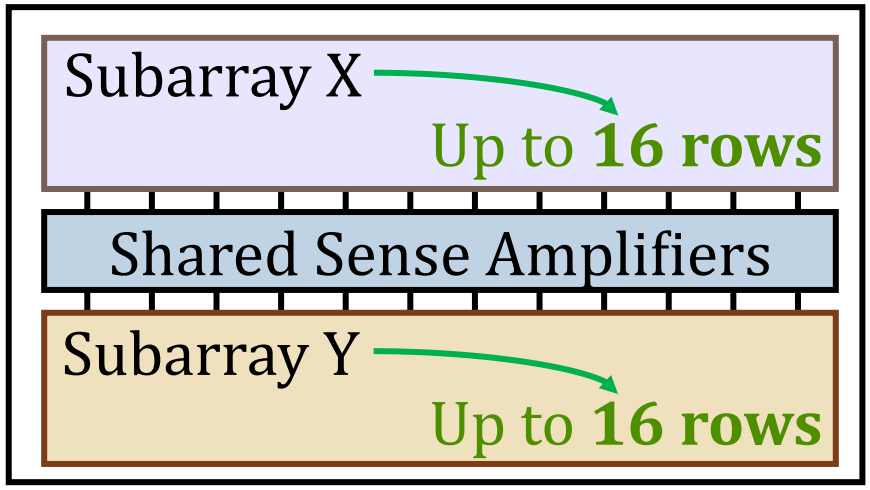
- To understand **which and how many** rows are simultaneously activated
  - **Sweep** Row A and Row B addresses



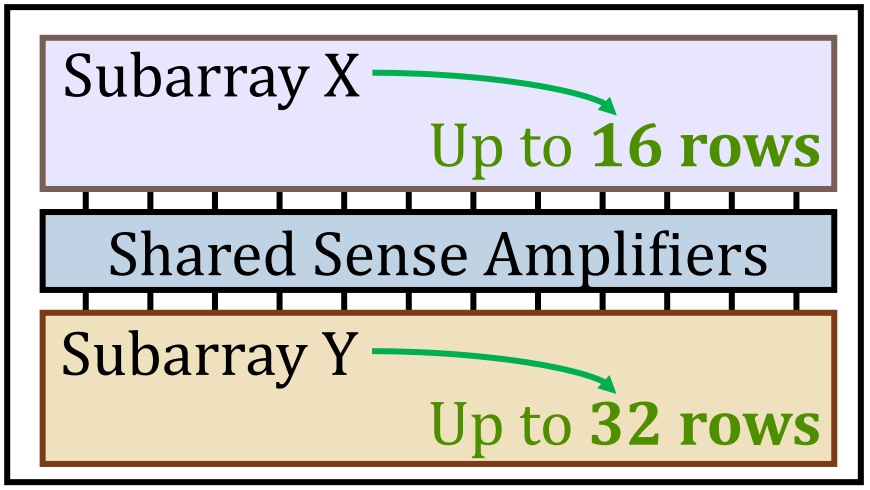
Key Results

COTS DRAM chips have **two distinct** sets of activation patterns in **neighboring subarrays** when two rows are activated with **violated timings**

Exactly the same number of rows in each subarray are activated	**Twice as many** rows in one subarray **compared to its neighbor subarray** are activated



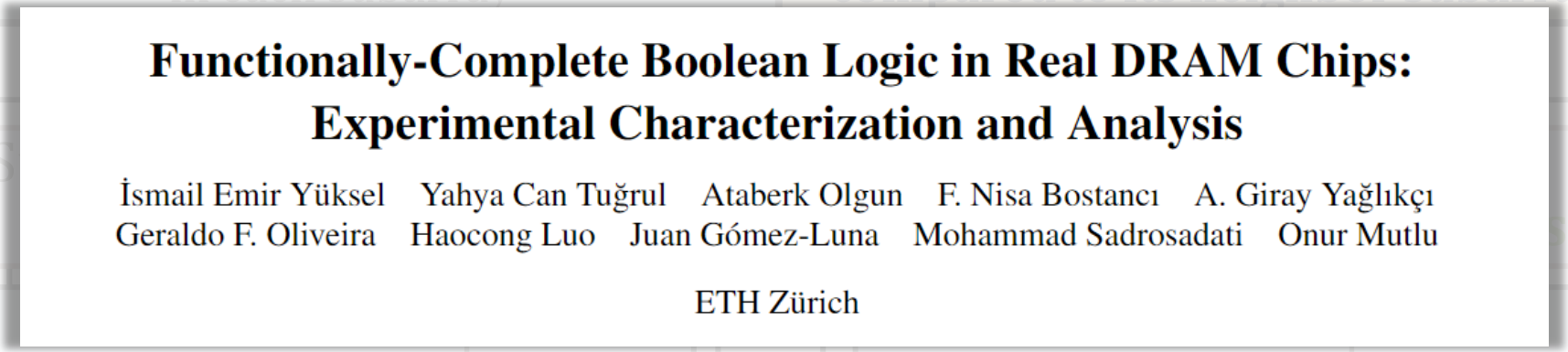
A total of **32 rows**



A total of **48 rows**

Key Takeaway

COTS DRAM chips can simultaneously activate up to 48 rows in two neighboring subarrays



Functionally-Complete Boolean Logic in Real DRAM Chips: Experimental Characterization and Analysis

İsmail Emir Yüksel Yahya Can Tuğrul Ataberk Olgun F. Nisa Bostancı A. Giray Yağlıkçı
Geraldo F. Oliveira Haocong Luo Juan Gómez-Luna Mohammad Sadrosadati Onur Mutlu

ETH Zürich

(More results in the paper)

https://arxiv.org/pdf/2402.18736.pdf

Outline

Background

Goal & Overview

Experimental Methodology

Multiple-Row Activation in Neighboring Subarrays

NOT Operation

AND, NAND, OR, and NOR Operations

Conclusion

The Capability of COTS DRAM Chips

We **demonstrate** that **COTS DRAM chips:**

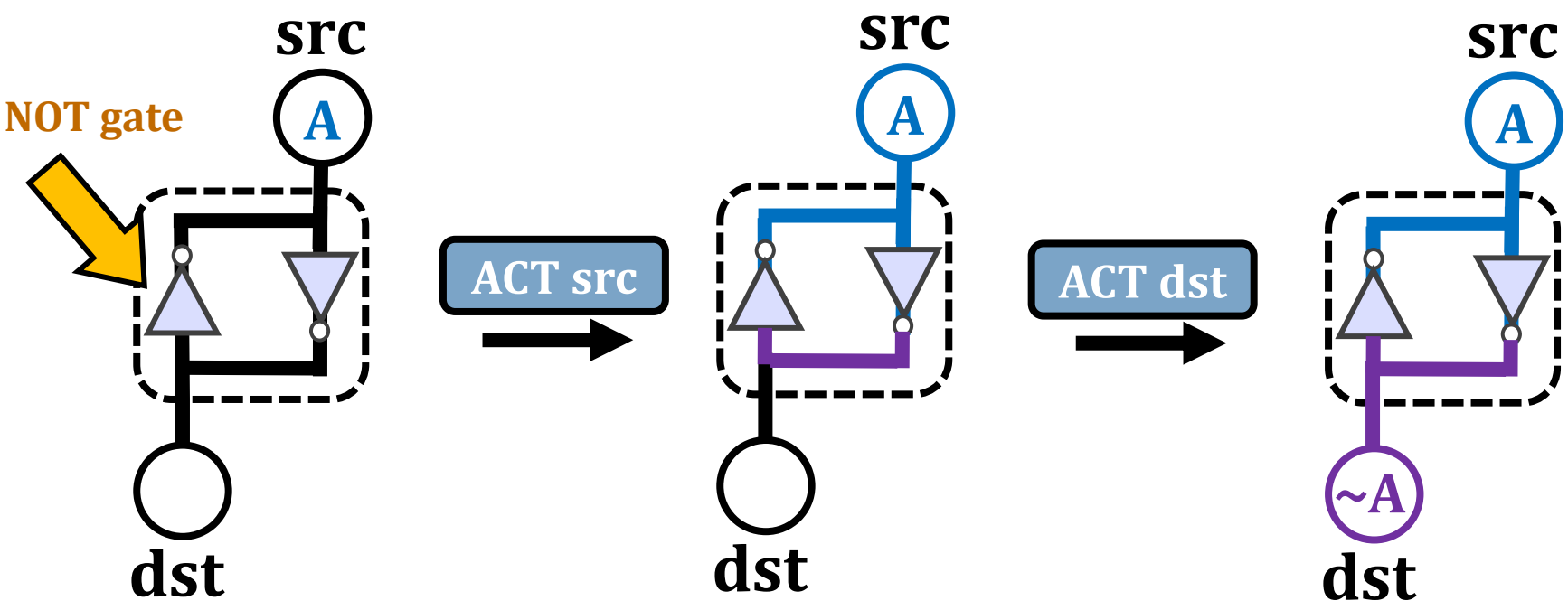
1 Can simultaneously activate up to 48 rows in two neighboring subarrays

2 Can perform **NOT operation** with **up to 32** output operands

3 Can perform up to 16-input AND, NAND, OR, and NOR operations

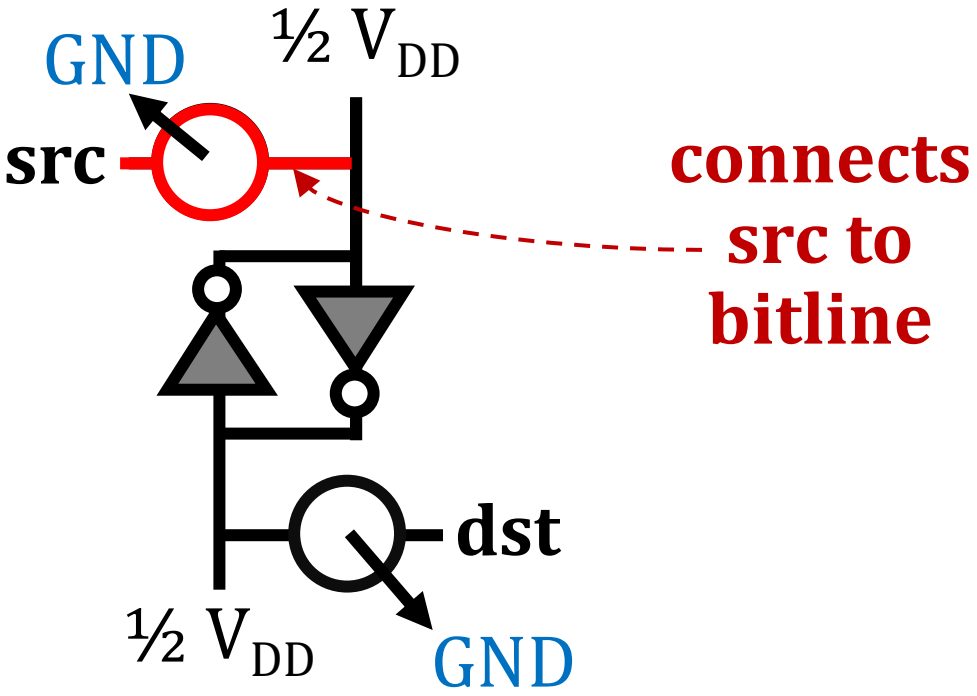
Key Idea

Connect rows in neighboring subarrays
through **a NOT gate** by simultaneously activating rows

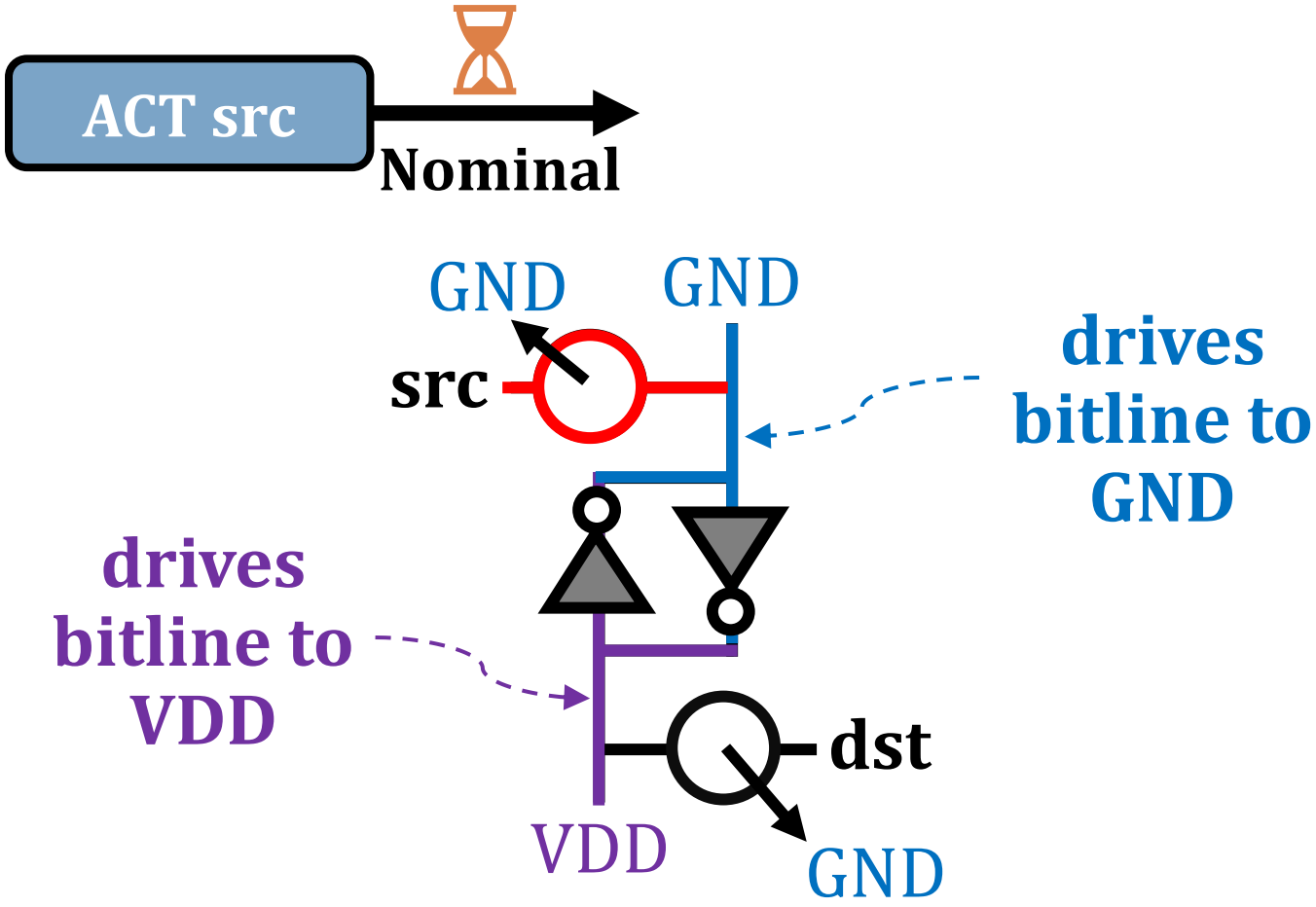


NOT Operation: A Walkthrough

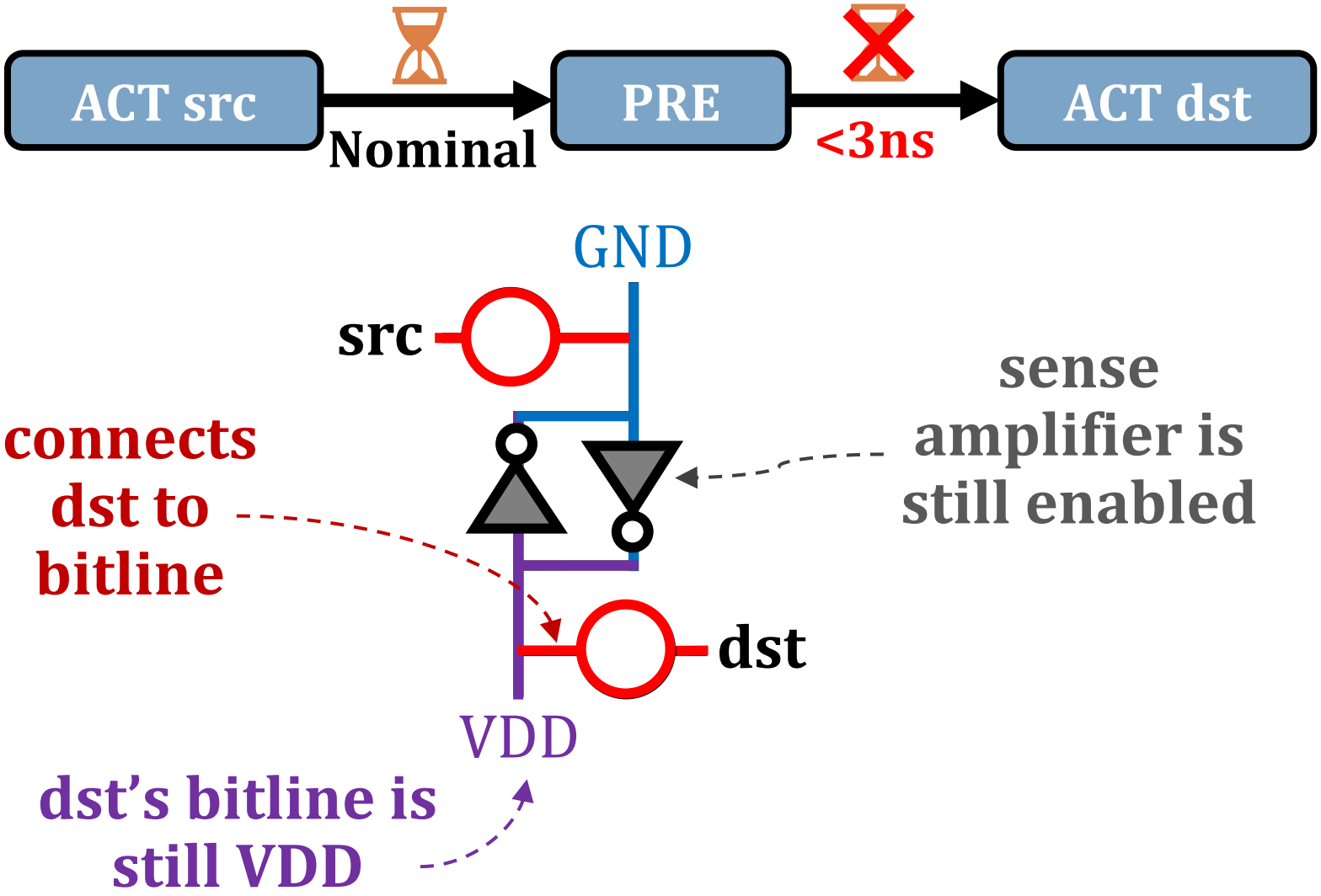
ACT src



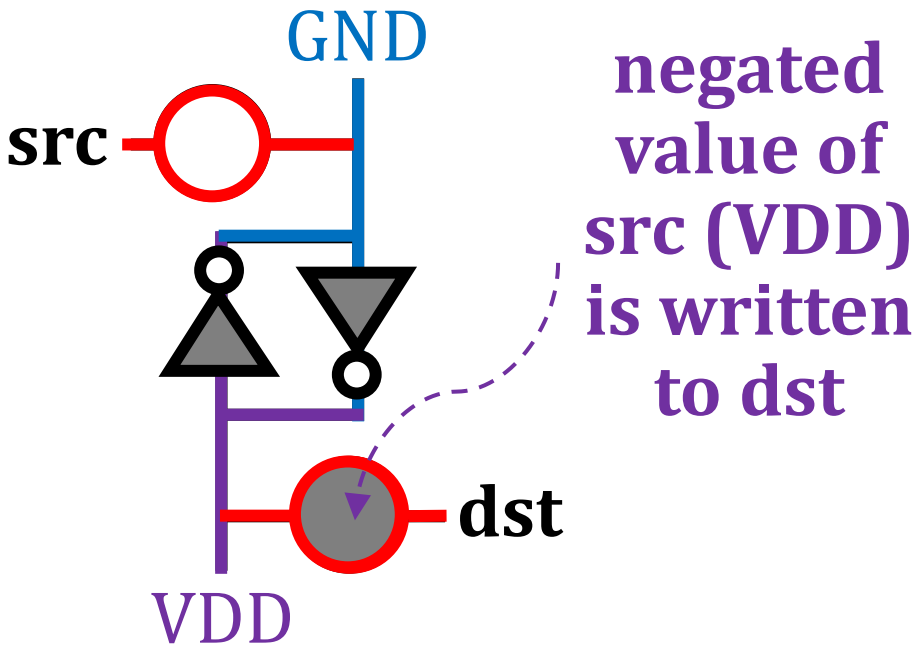
NOT Operation: A Walkthrough



NOT Operation: A Walkthrough

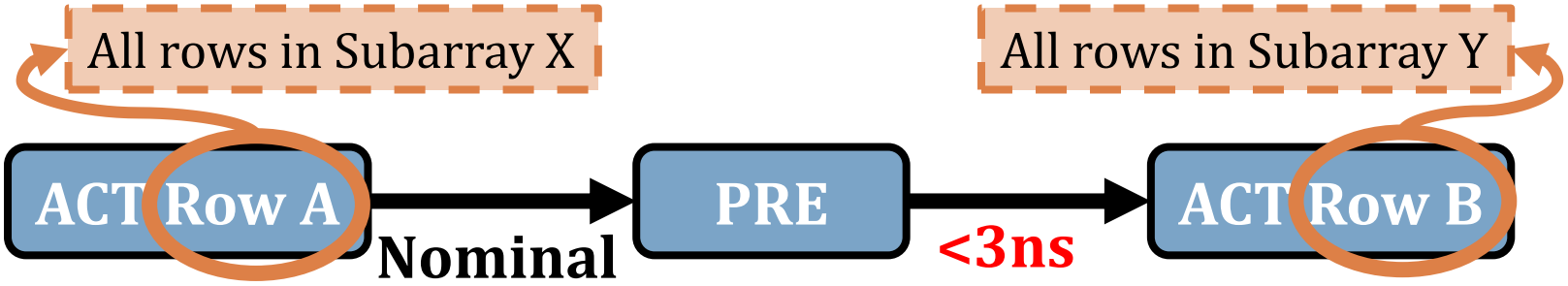


NOT Operation: A Walkthrough

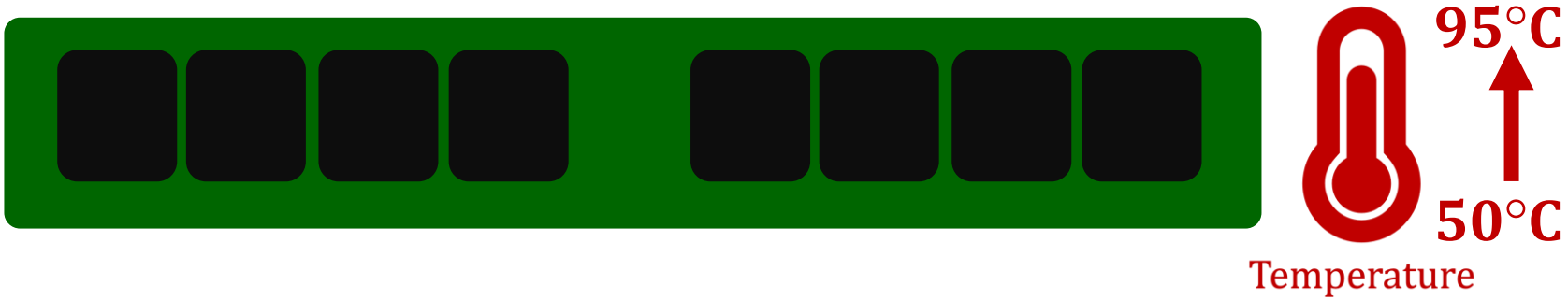


Characterization Methodology

- Sweep **Row A and Row B addresses**



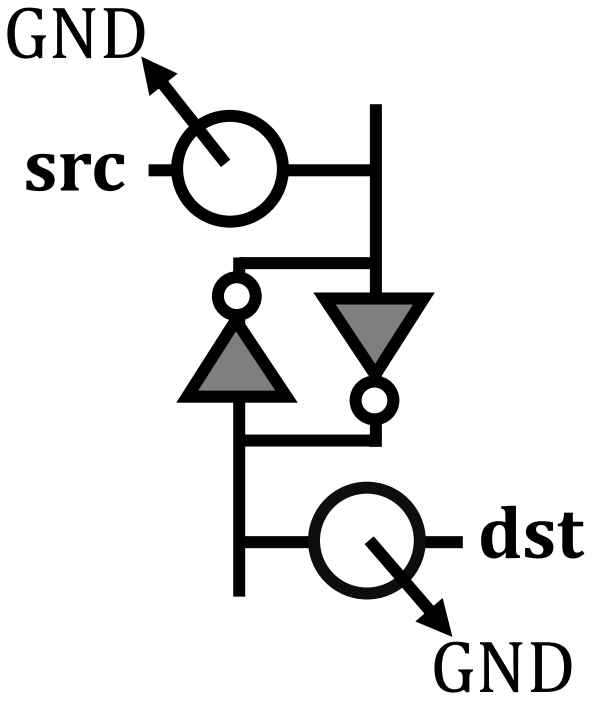
- Sweep **DRAM chip temperature**



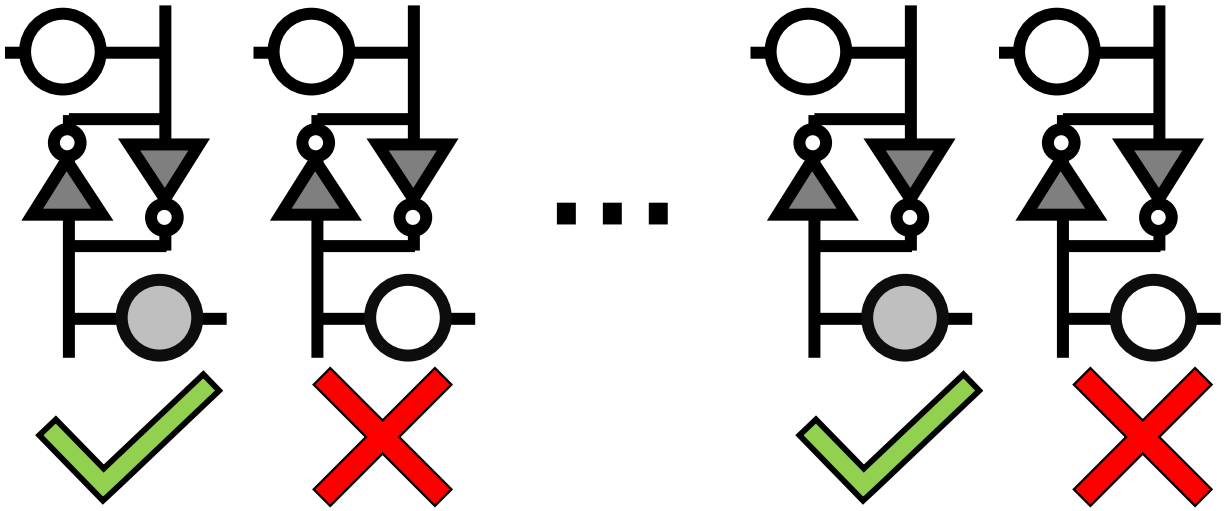
Reliability Metric

Success Rate (for a DRAM cell)

Percentage of trials where the **correct output** of a tested operation is stored in the cell



Total of 10000 trials



Success rate for this cell: 50%

Key Takeaways from In-DRAM NOT Operation

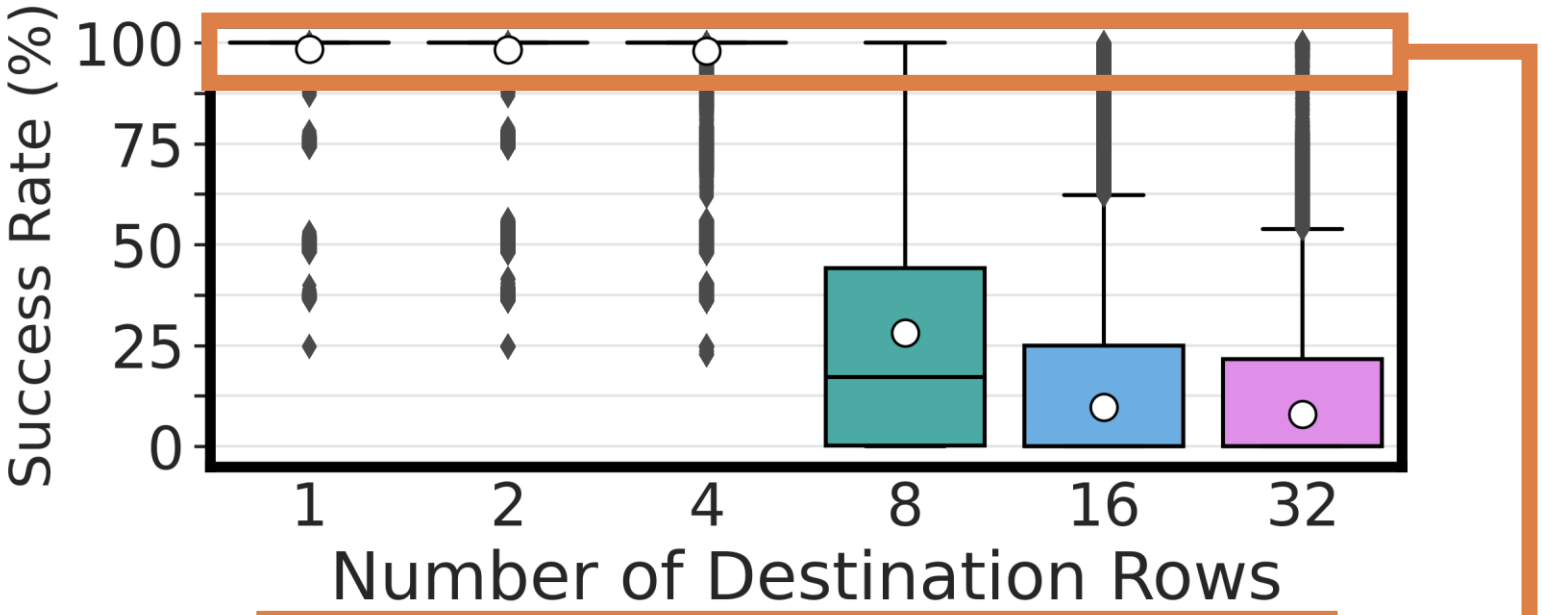
Key Takeaway 1

COTS DRAM chips can perform NOT operations with up to 32 destination rows

Key Takeaway 2

Temperature has a small effect on the reliability of NOT operations

Performing NOT in COTS DRAM Chips

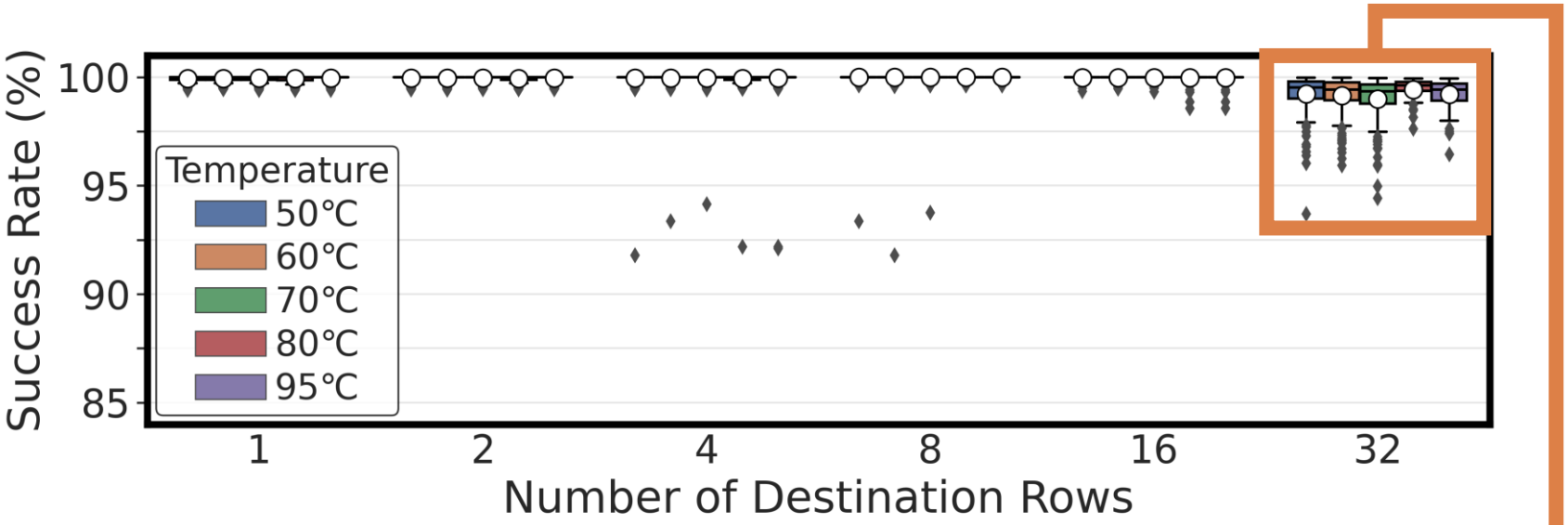


There is **at least one DRAM cell** that can perform **the NOT operation with a 100% success rate**

COTS DRAM chips can perform NOT operations with up to 32 destination rows

Impact of Temperature

- Used **destination cells** that can perform NOT operation with **>90% success rate at 50°C**



from 50°C to 95°C
only 0.2% variation in average success rate

Temperature has a small effect on the reliability of NOT operations

Outline

Background

Goal & Overview

Experimental Methodology

Multiple-Row Activation in Neighboring Subarrays

NOT Operation

AND, NAND, OR, and NOR Operations

Conclusion

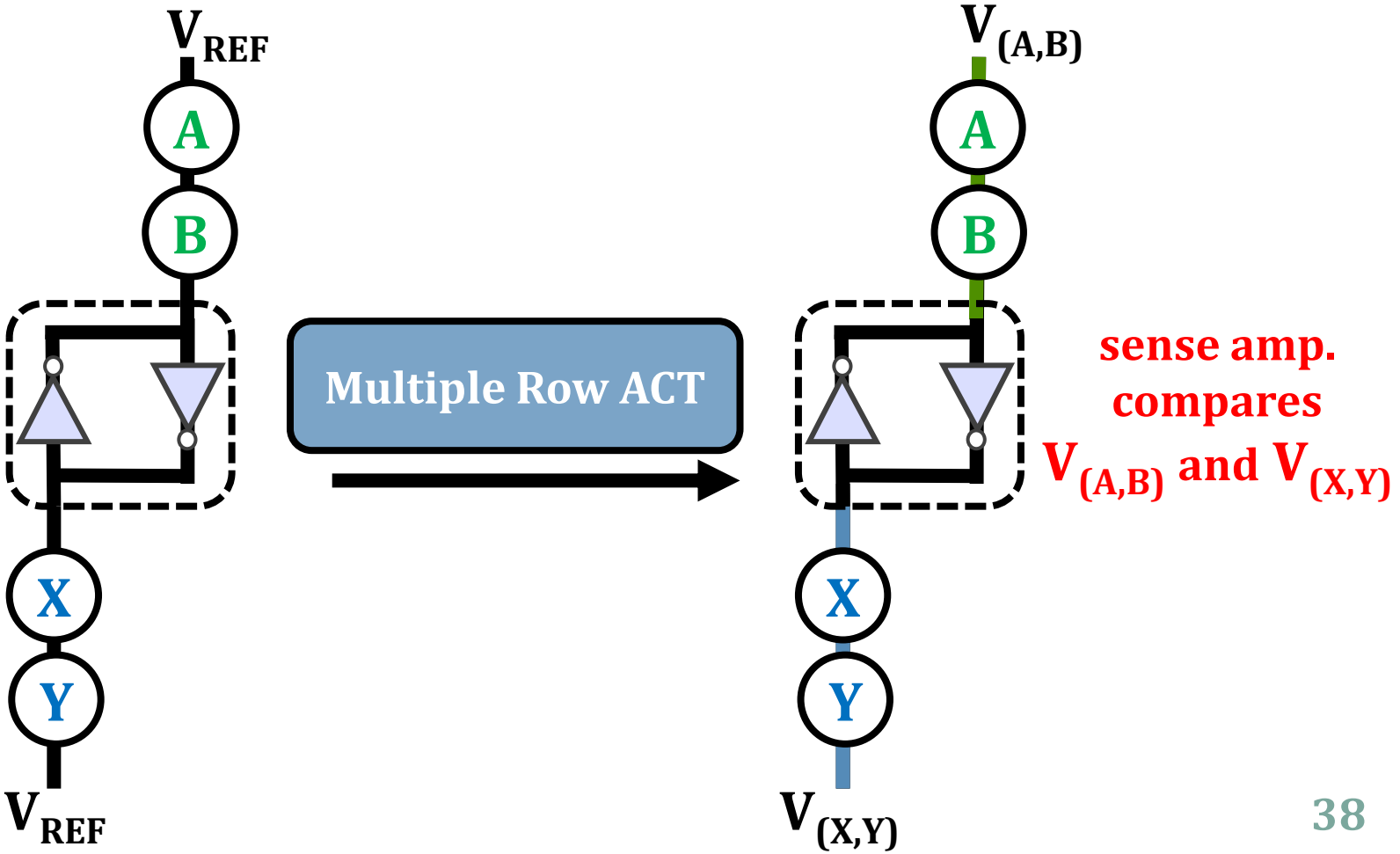
The Capability of COTS DRAM Chips

We **demonstrate** that **COTS DRAM chips:**

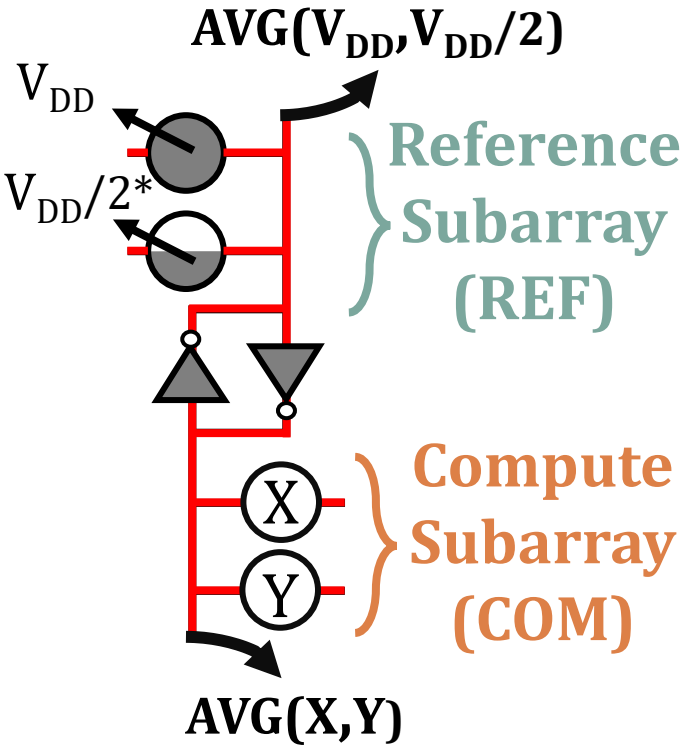
1. Can simultaneously activate up to 48 rows in two neighboring subarrays
2. Can perform NOT operation with up to 32 output operands
3. Can perform **up to 16-input AND, NAND, OR, and NOR** operations

Key Idea

Manipulate the bitline voltage to express
a wide variety of functions using
multiple-row activation in neighboring subarrays

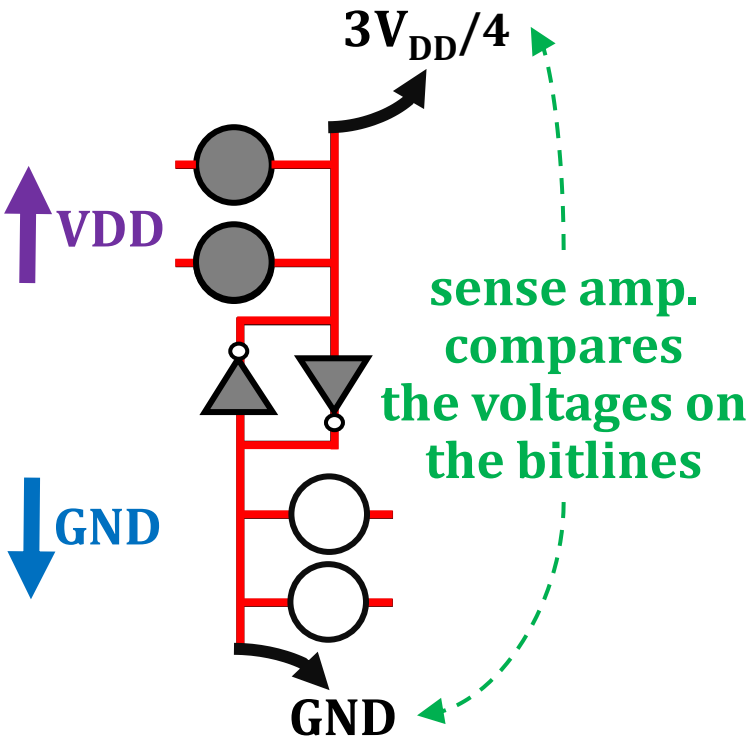


Two-Input AND and NAND Operations



*Gao et al., "FracDRAM: Fractional Values in Off-the-Shelf DRAM," in MICRO, 2022.

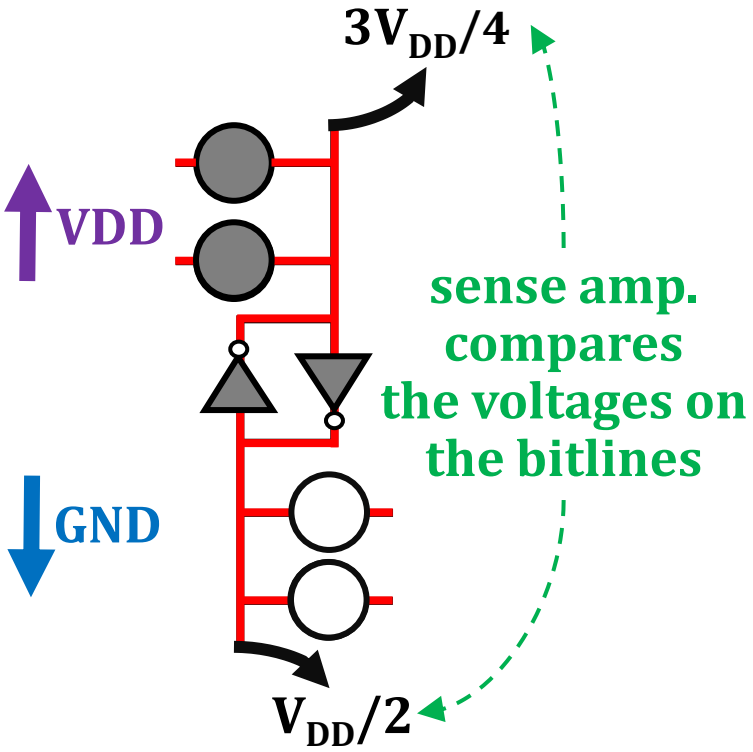
Two-Input AND and NAND Operations



V_{DD}=1 & GND = 0

X	Y	COM	REF
0	0	0	1

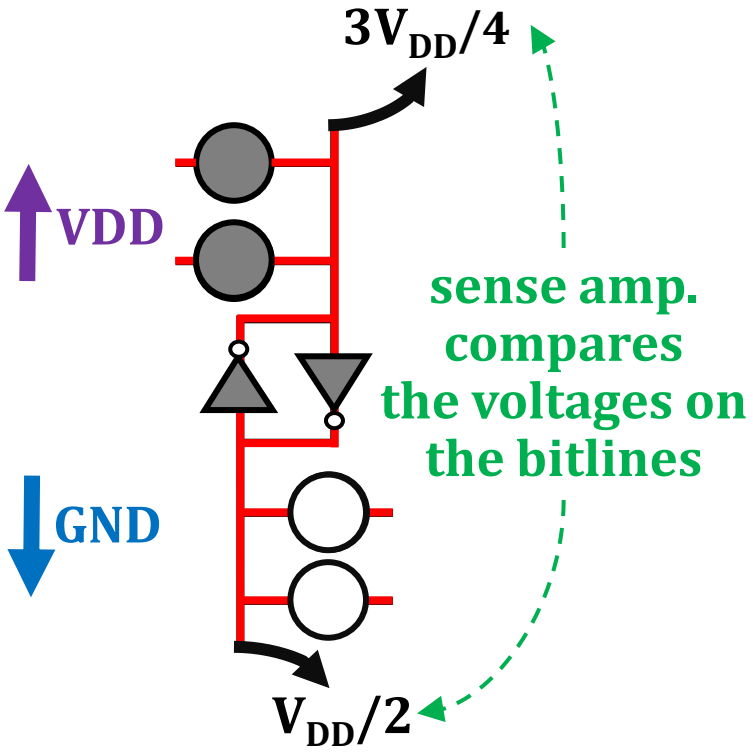
Two-Input AND and NAND Operations



V_{DD}=1 & GND = 0

X	Y	COM	REF
0	0	0	1
0	1	0	1

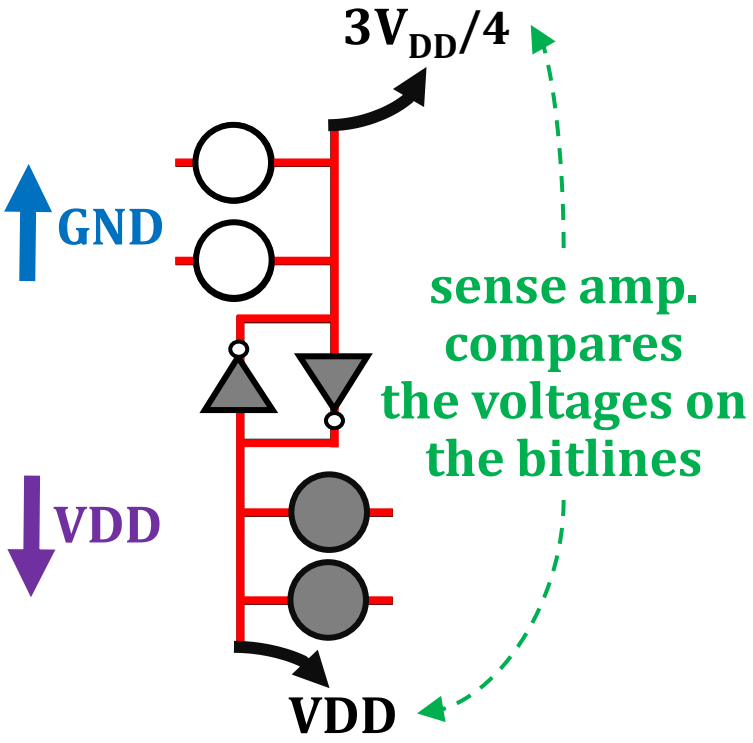
Two-Input AND and NAND Operations



V_{DD}=1 & GND = 0

X	Y	COM	REF
0	0	0	1
0	1	0	1
1	0	0	1

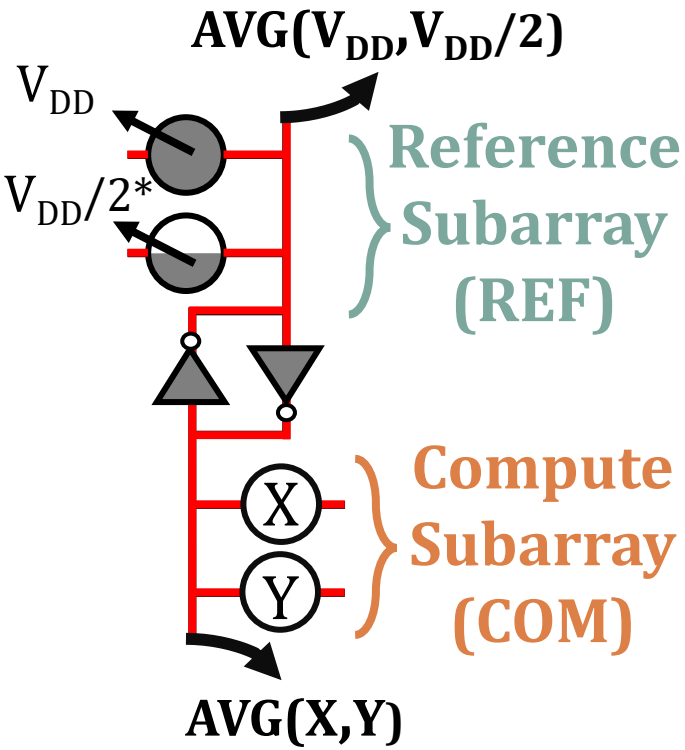
Two-Input AND and NAND Operations



V_{DD}=1 & GND = 0

X	Y	COM	REF
0	0	0	1
0	1	0	1
1	0	0	1
1	1	1	0

Two-Input AND and NAND Operations



V_{DD}=1 & GND = 0

X	Y	COM	REF
0	0	0	1
0	1	0	1
1	0	0	1
1	1	1	0
		AND	NAND

*Gao et al., "FracDRAM: Fractional Values in Off-the-Shelf DRAM," in MICRO, 2022.

Many-Input AND, NAND, OR, and NOR Operations

We can express AND, NAND, OR, and NOR operations by carefully manipulating the **reference voltage**

Functionally-Complete Boolean Logic in Real DRAM Chips: Experimental Characterization and Analysis

İsmail Emir Yüksel Yahya Can Tuğrul Ataberk Olgun F. Nisa Bostancı A. Giray Yağlıkçı
Geraldo F. Oliveira Haocong Luo Juan Gómez-Luna Mohammad Sadrosadati Onur Mutlu

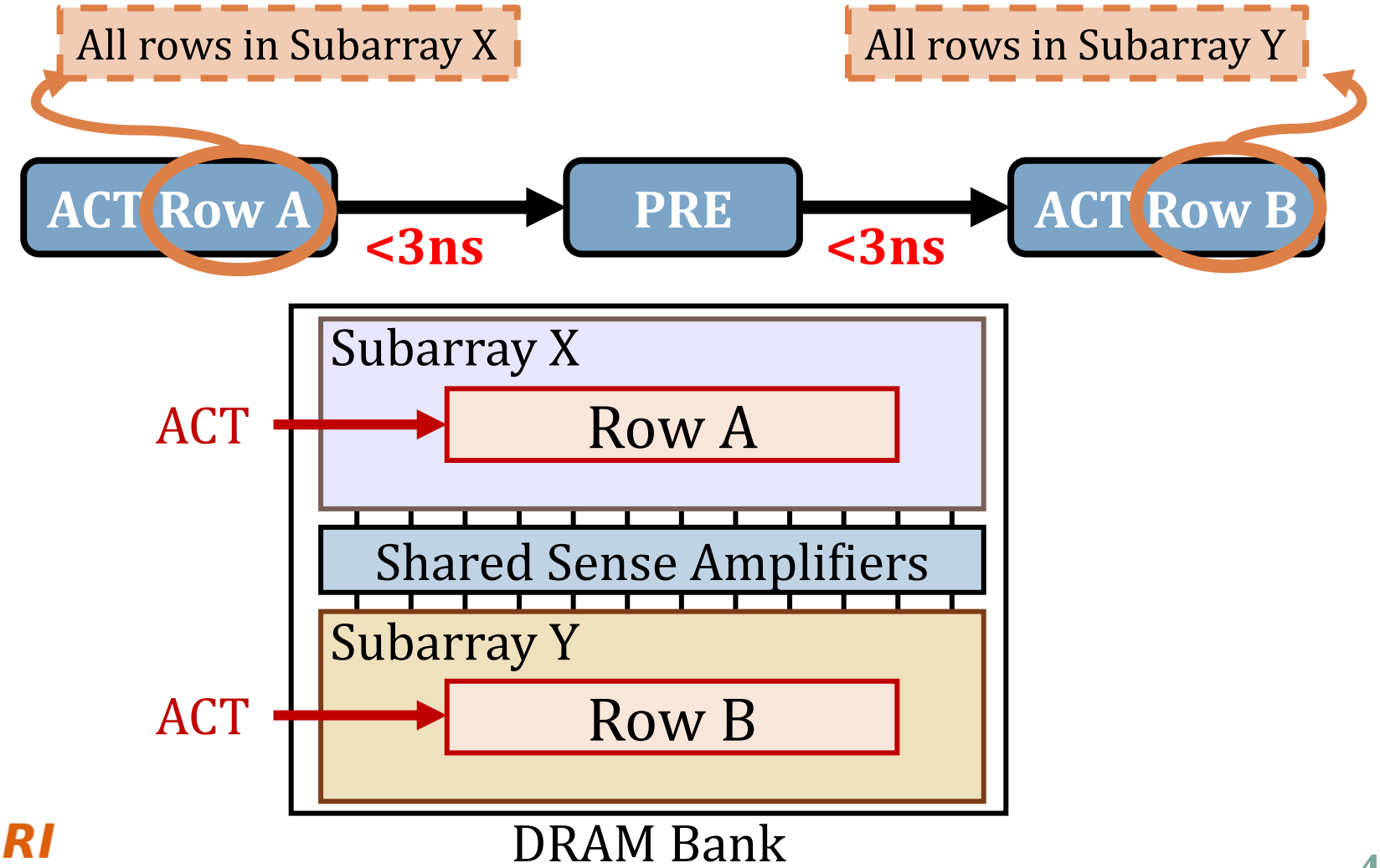
ETH Zürich

(More details in the paper)

https://arxiv.org/pdf/2402.18736.pdf

Characterization Methodology

- **Sweep** Row A and Row B addresses



Key Takeaways from In-DRAM Operations

Key Takeaway 1

COTS DRAM chips can perform {2, 4, 8, 16}-input AND, NAND, OR, and NOR operations

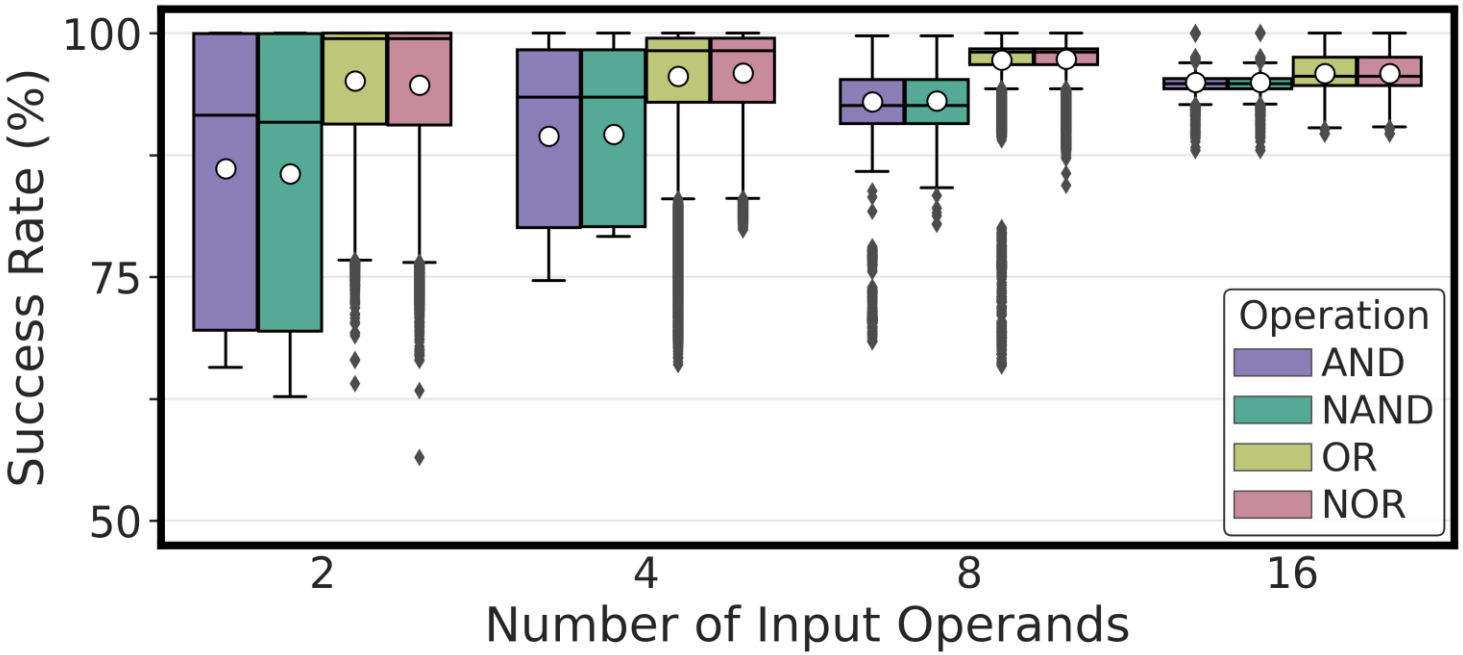
Key Takeaway 2

COTS DRAM chips can perform AND, NAND, OR, and NOR operations with very high reliability

Key Takeaway 3

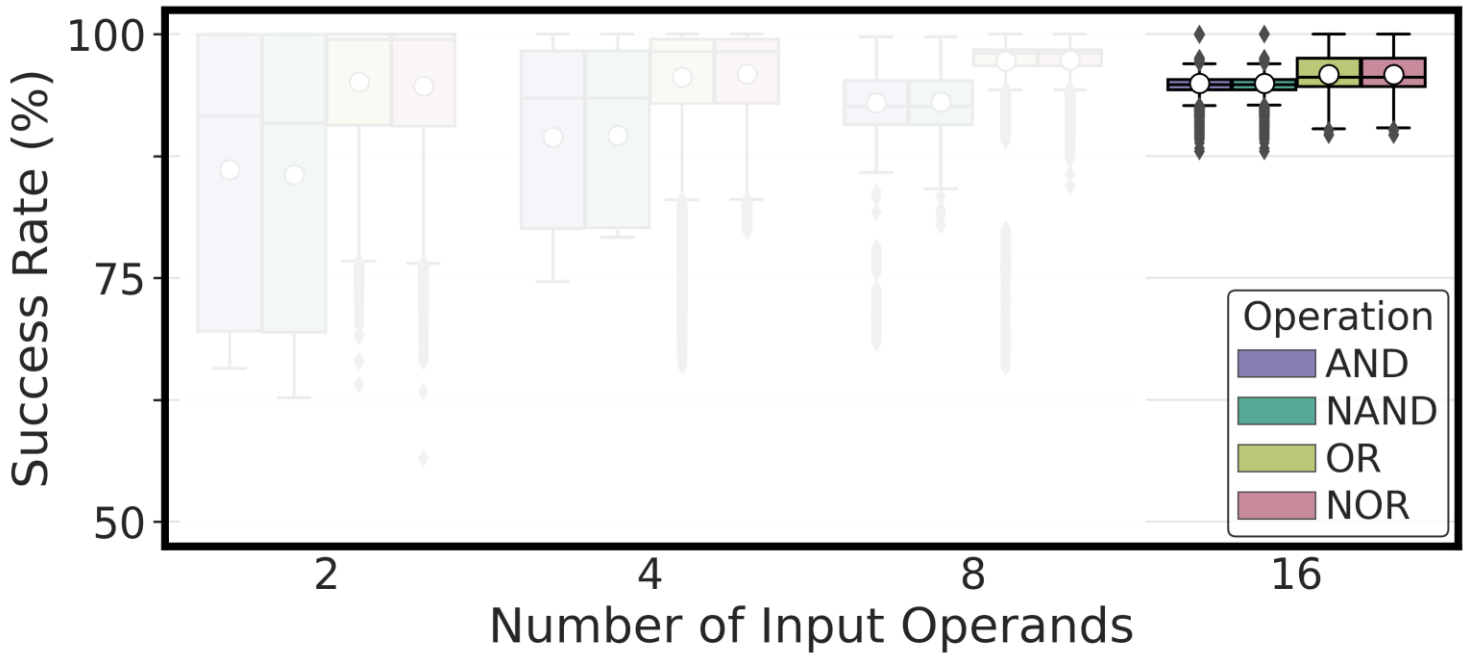
Data pattern slightly affects the reliability of AND, NAND, OR, and NOR operations

Performing AND, NAND, OR, and NOR



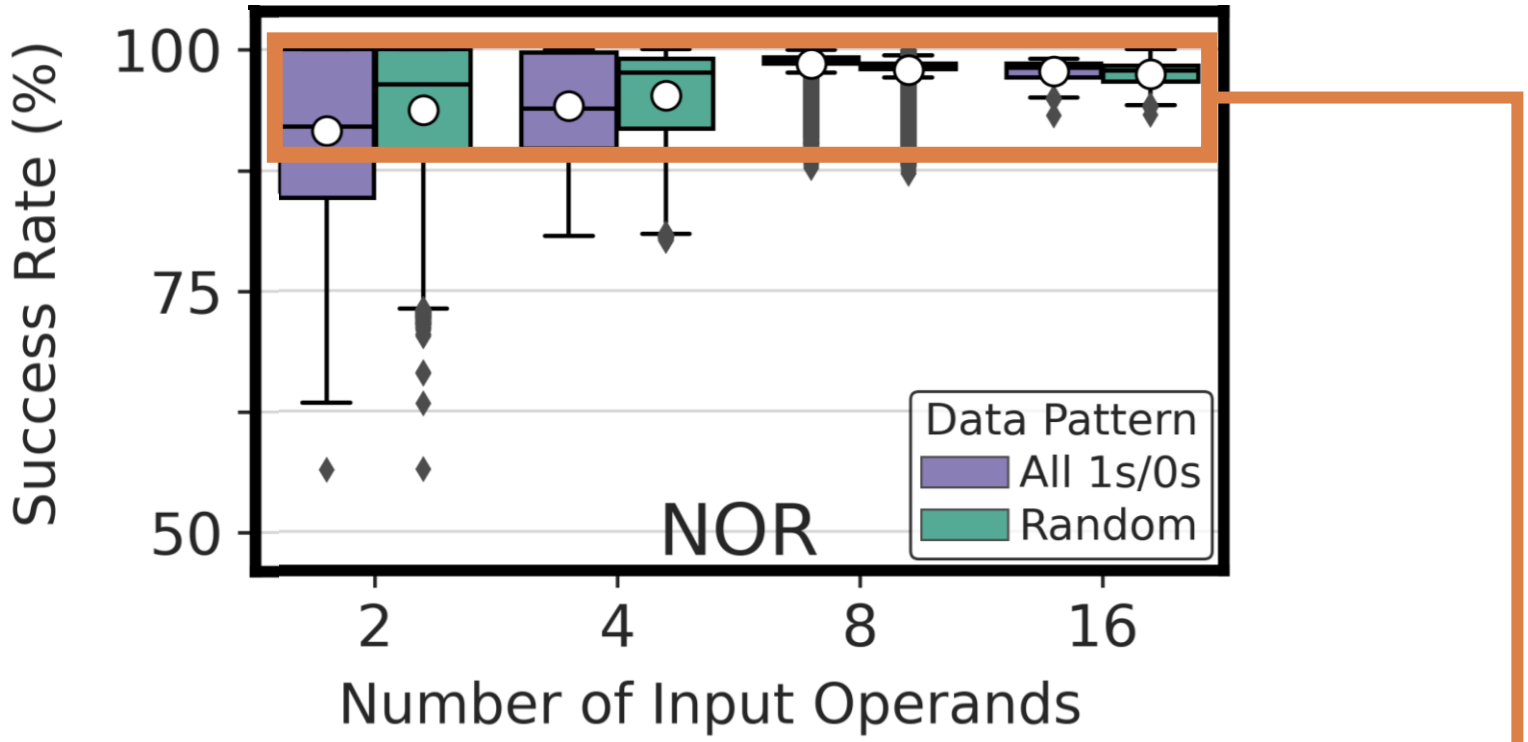
COTS DRAM chips can perform {2, 4, 8, 16}-input AND, NAND, OR, and NOR operations

Performing AND, NAND, OR, and NOR



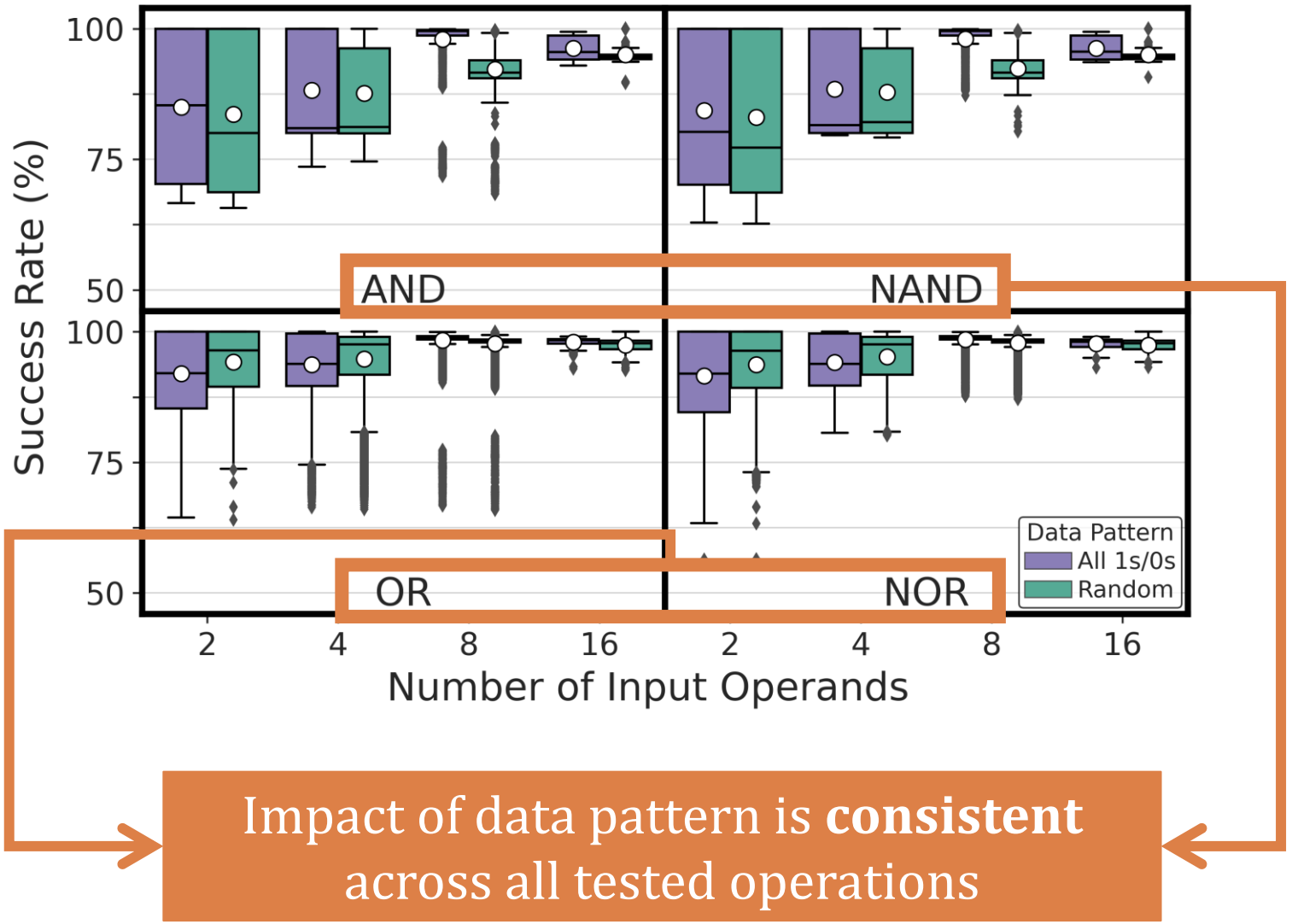
COTS DRAM chips can perform 16-input AND, NAND, OR, and NOR operations with very high success rate (>94%)

Impact of Data Pattern



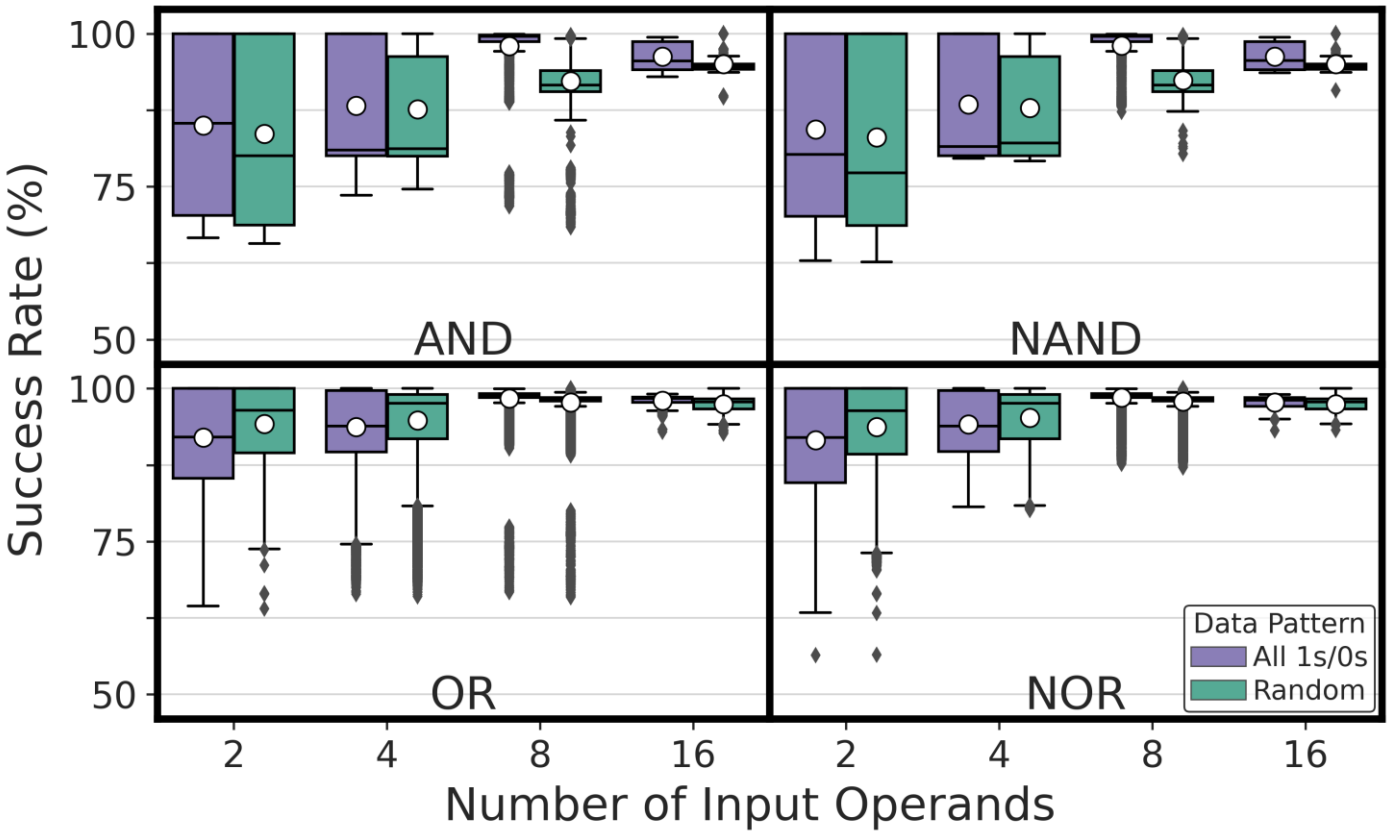
1.98% variation in **average success rate** across **all number of input operands**

Impact of Data Pattern



Impact of data pattern is **consistent** across all tested operations

Impact of Data Pattern



Data pattern slightly affects the reliability of AND, NAND, OR, and NOR operations

More in the Paper

- Detailed hypotheses & key ideas to perform
  - NOT operation
  - Many-input AND, NAND, OR, and NOR operations

- How the reliability of bitwise operations are affected by
  - The location of activated rows
  - Temperature (for AND, NAND, OR, and NOR)
  - DRAM speed rate
  - Chip density and die revision

- Discussion on the limitations of COTS DRAM chips

Available on arXiv

Functionally-Complete Boolean Logic in Real DRAM Chips: Experimental Characterization and Analysis

İsmail Emir Yüksel Yahya Can Tuğrul Ataberk Olgun F. Nisa Bostancı A. Giray Yağlıkçı
Geraldo F. Oliveira Haocong Luo Juan Gómez-Luna Mohammad Sadrosadati Onur Mutlu

ETH Zürich

Processing-using-DRAM (PuD) is an emerging paradigm that leverages the analog operational properties of DRAM circuitry to enable massively parallel in-DRAM computation. PuD has the potential to significantly reduce or eliminate costly data movement between processing elements and main memory. A common approach for PuD architectures is to make use of bulk bitwise computation (e.g., AND, OR, NOT). Prior works experimentally demonstrate three-input MAJ (i.e., MAJ3) and two-input AND and OR operations in commercial off-the-shelf (COTS) DRAM chips. Yet, demonstrations on COTS DRAM chips do not *provide a functionally complete set of operations (e.g., NAND or AND and NOT).*

We experimentally demonstrate that COTS DRAM chips are capable of performing 1) functionally-complete *Boolean operations: NOT, NAND, and NOR and 2) many-input (i.e., more than two-input) AND and OR operations. We present an extensive*

systems and applications [12, 13]. Processing-using-DRAM (PuD) [29–32] is a promising paradigm that can alleviate the data movement bottleneck. PuD uses the analog operational properties of the DRAM circuitry to enable massively parallel in-DRAM computation. Many prior works [29–53] demonstrate that PuD can greatly reduce or eliminate data movement.

A widely used approach for PuD is to perform bulk bitwise operations, i.e., bitwise operations on large bit vectors. To perform bulk bitwise operations using DRAM, prior works propose modifications to the DRAM circuitry [29–31, 33, 35, 36, 43, 44, 46, 48–58]. Recent works [38, 41, 42, 45] experimentally demonstrate the feasibility of executing data copy & initialization [42, 45], i.e., the RowClone operation [49], and a subset of bitwise operations, i.e., three-input bitwise majority (MAJ3) and two-input AND and OR operations in unmodified commercial off-the-shelf (COTS) DRAM chips by operating beyond

https://arxiv.org/pdf/2402.18736.pdf

Outline

Background

Goal & Overview

Experimental Methodology

Multiple-Row Activation in Neighboring Subarrays

NOT Operation

AND, NAND, OR, and NOR Operations

Conclusion

Conclusion

- We experimentally demonstrate that **commercial off-the-shelf (COTS)** DRAM chips can perform:
  - **Functionally-complete** Boolean operations: NOT, NAND, and NOR
  - **Up to 16-input** AND, NAND, OR, and NOR operations
- We characterize **the success rate** of these operations on **256 COTS DDR4 chips** from **two major manufacturers**
- We highlight **two key results:**
  - We can perform **NOT** and **{2, 4, 8, 16}-input AND, NAND, OR, and NOR** operations on COTS DRAM chips with **very high success rates (>94%)**
  - **Data pattern** and **temperature** only slightly affect the reliability of these operations

We believe these empirical results demonstrate the promising potential of using DRAM as a computation substrate

Functionally-Complete Boolean Logic in Real DRAM Chips

Experimental Characterization and Analysis

Ismail Emir Yüksel

Yahya C. Tugrul Ataberk Olgun F. Nisa Bostancı

A. Giray Yaglıkçı Geraldo F. Oliveira Haocong Luo

Juan Gómez–Luna Mohammad Sadr Onur Mutlu

SAFARI

ETH zürich

Simultaneous Many-Row Activation in Off-the-Shelf DRAM Chips

Experimental Characterization and Analysis

İsmail Emir Yüksel

Yahya C. Tuğrul F. Nisa Bostancı Geraldo F. Oliveira

A. Giray Yağlıkçı Ataberk Olgun Melina Soysal Haocong Luo

Juan Gómez–Luna Mohammad Sadr Onur Mutlu

SAFARI

ETH zürich

Executive Summary

Motivation:

- **Processing-Using-DRAM (PUD)** alleviates **data movement bottlenecks**
- Commercial off-the-shelf (COTS) DRAM chips can perform **three-input majority (MAJ3)** and **in-DRAM copy** operations

Goal: To experimentally analyze and understand

- The **computational capability** of COTS DRAM chips beyond that of prior works
- The **robustness** of such capability under various **operating conditions**

Experimental Study: 120 DDR4 chips from **two major manufacturers**

- COTS DRAM chips can perform **MAJ5, MAJ7, and MAJ9** operations and **copy** one DRAM row to **up to 31 different rows** at once
- Storing **multiple redundant copies** of MAJ's input operands (i.e., input replication) drastically increases **robustness** (>30% higher success rate)
- **Operating conditions** (temperature, voltage, and data pattern) **affect** the robustness of in-DRAM operations (by up to 11.52% success rate)

Leveraging Simultaneous Many-Row Activation

1 Perform **MAJX (where X>3)** operations

2 **Increase the robustness** of MAJX operations

3 Copy **one row's content** to **multiple rows**

Leveraging Simultaneous Many-Row Activation

1 Perform **MAJX (where X>3)** operations

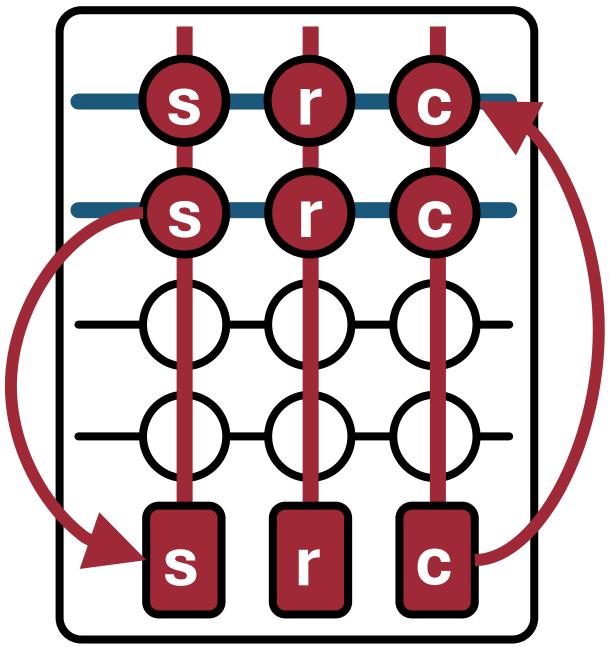
2 **Increase the robustness** of MAJX operations

3 Copy **one row's content** to **multiple rows**

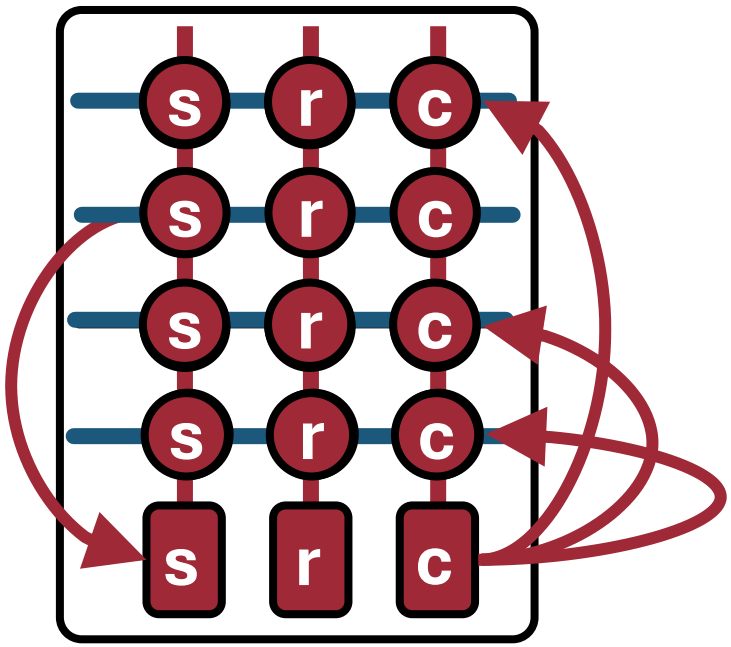
In-DRAM Multiple Row Copy (Multi-RowCopy)

Simultaneously activate many rows to
copy **one row's content** to **multiple destination rows**

RowClone

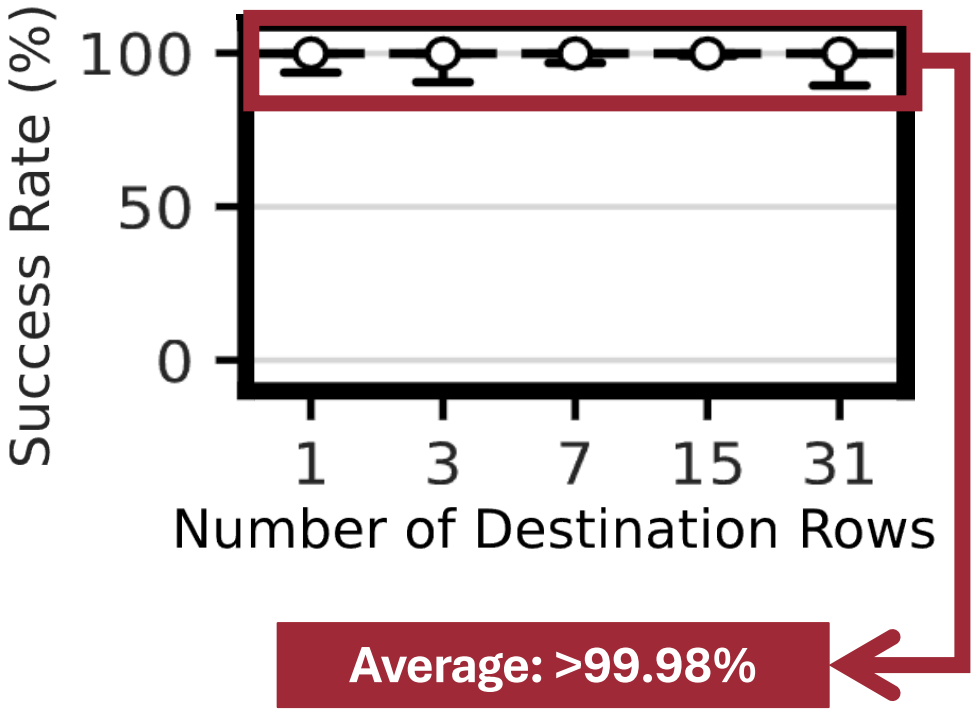


Multi-RowCopy



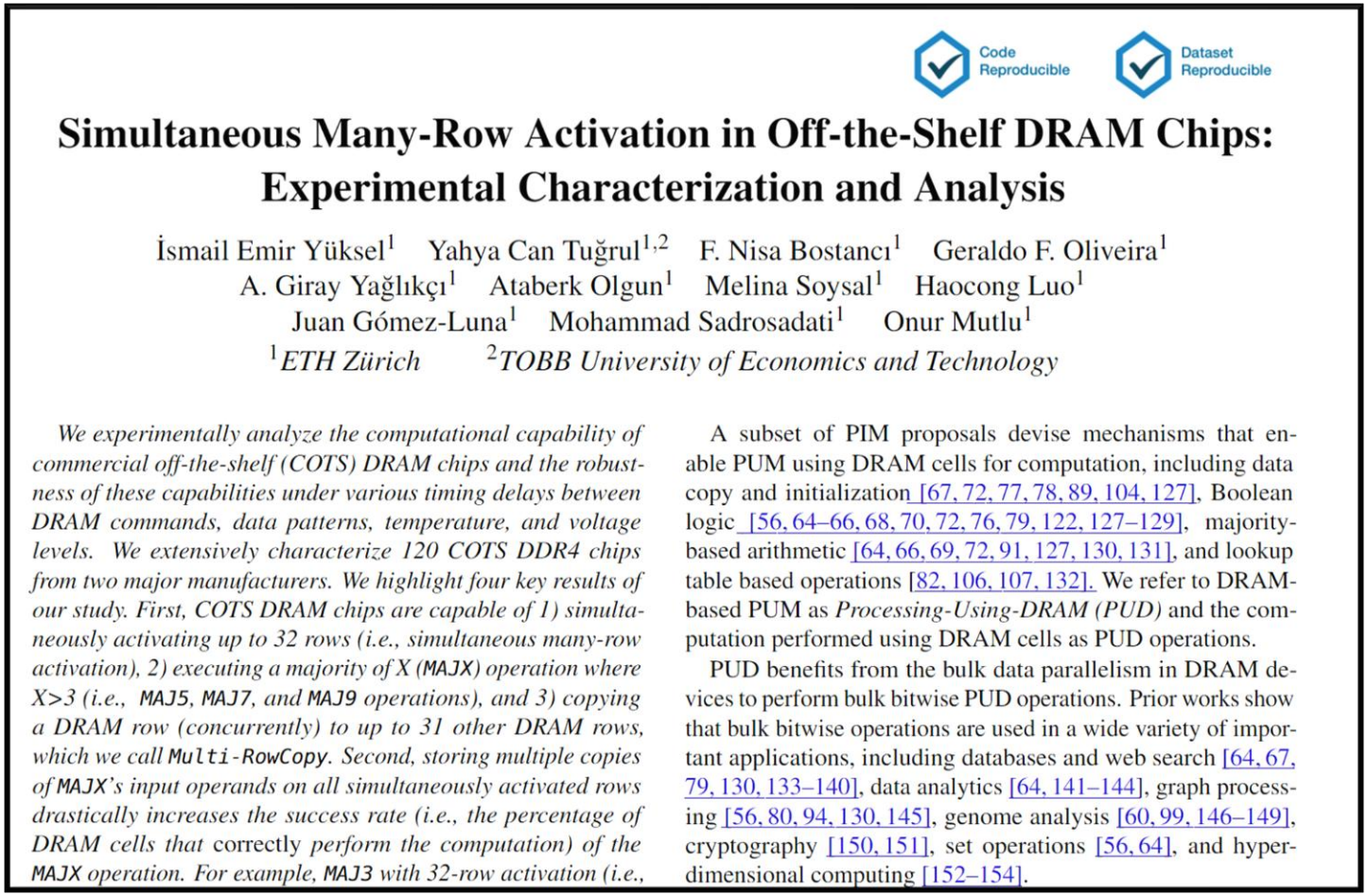
[Seshadri+ MICRO'13]

Robustness of Multi-RowCopy



COTS DRAM chips can copy one row's content to up to 31 rows with a very high success rate

Available on arXiv



Code Reproducible

Dataset Reproducible

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*

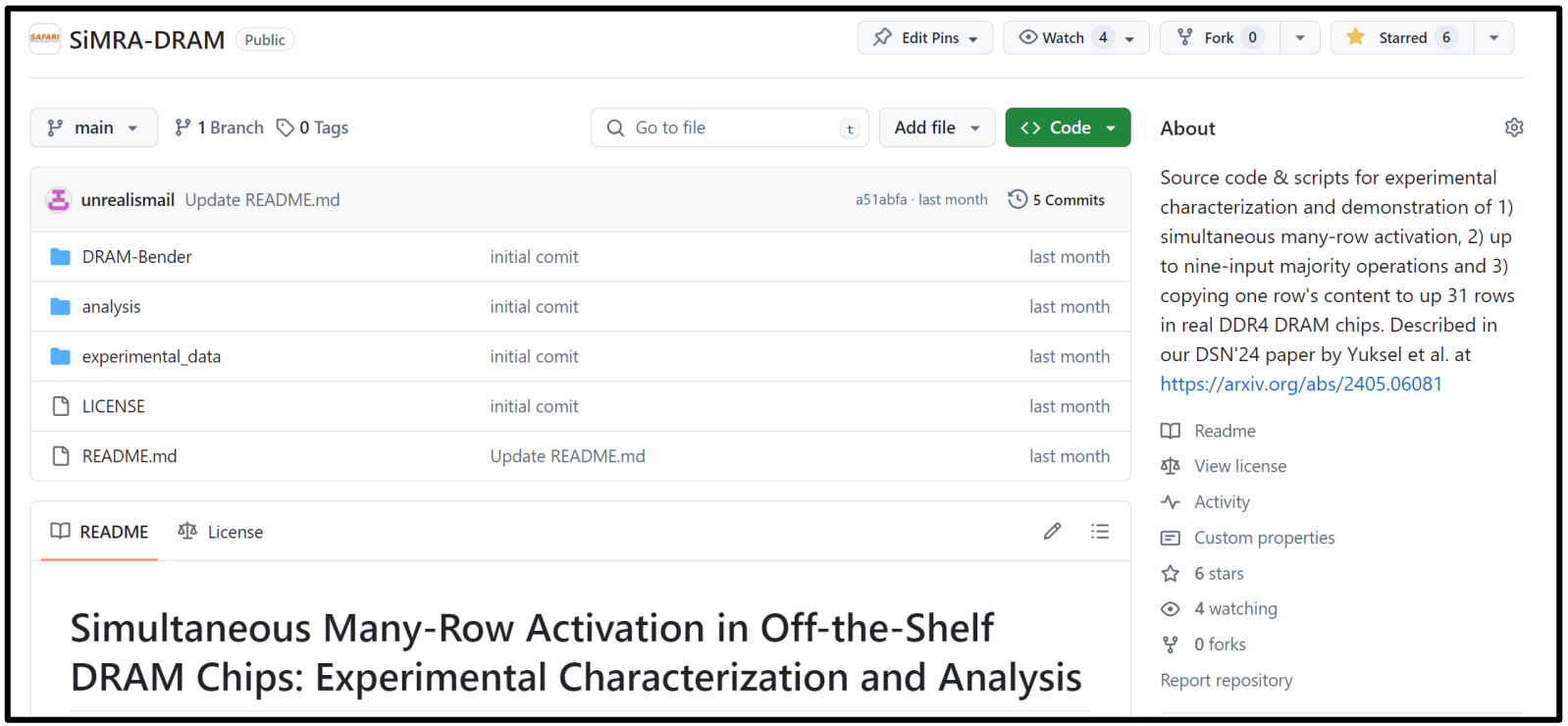
We experimentally analyze the computational capability of commercial off-the-shelf (COTS) DRAM chips and the robustness of these capabilities under various timing delays between DRAM commands, data patterns, temperature, and voltage levels. We extensively characterize 120 COTS DDR4 chips from two major manufacturers. We highlight four key results of our study. First, COTS DRAM chips are capable of 1) simultaneously activating up to 32 rows (i.e., simultaneous many-row activation), 2) executing a majority of X (`MAJX`) operation where X>3 (i.e., `MAJ5`, `MAJ7`, and `MAJ9` operations), and 3) copying a DRAM row (concurrently) to up to 31 other DRAM rows, which we call `Multi-RowCopy`. Second, storing multiple copies of `MAJX`'s input operands on all simultaneously activated rows drastically increases the success rate (i.e., the percentage of DRAM cells that correctly *perform the computation) of the `MAJX` operation. For example, `MAJ3` with 32-row activation (i.e.,*

A subset of PIM proposals devise mechanisms that enable PUM using DRAM cells for computation, including data copy and initialization [67, 72, 77, 78, 89, 104, 127], Boolean logic [56, 64–66, 68, 70, 72, 76, 79, 122, 127–129], majority-based arithmetic [64, 66, 69, 72, 91, 127, 130, 131], and lookup table based operations [82, 106, 107, 132]. We refer to DRAM-based PUM as *Processing-Using-DRAM (PUD)* and the computation performed using DRAM cells as PUD operations.

PUD benefits from the bulk data parallelism in DRAM devices to perform bulk bitwise PUD operations. Prior works show that bulk bitwise operations are used in a wide variety of important applications, including databases and web search [64, 67, 79, 130, 133–140], data analytics [64, 141–144], graph processing [56, 80, 94, 130, 145], genome analysis [60, 99, 146–149], cryptography [150, 151], set operations [56, 64], and hyperdimensional computing [152–154].

https://arxiv.org/pdf/2405.06081

Our Work is Open Source and Artifact Evaluated



https://github.com/CMU-SAFARI/SiMRA-DRAM

Simultaneous Many-Row Activation in Off-the-Shelf DRAM Chips

Experimental Characterization and Analysis

Paper

GitHub

İsmail Emir Yüksel

Yahya C. Tuğrul F. Nisa Bostancı Geraldo F. Oliveira

A. Giray Yağlıkçı Ataberk Olgun Melina Soysal Haocong Luo

Juan Gómez–Luna Mohammad Sadr Onur Mutlu

SAFARI

ETH zürich

Functionally-Complete Boolean Logic in Real DRAM Chips

Experimental Characterization and Analysis

Backup Slides

Ismail Emir Yüksel

Yahya C. Tugrul Ataberk Olgun F. Nisa Bostancı

A. Giray Yaglıkçı Geraldo F. Oliveira Haocong Luo

Juan Gómez–Luna Mohammad Sadr Onur Mutlu

SAFARI

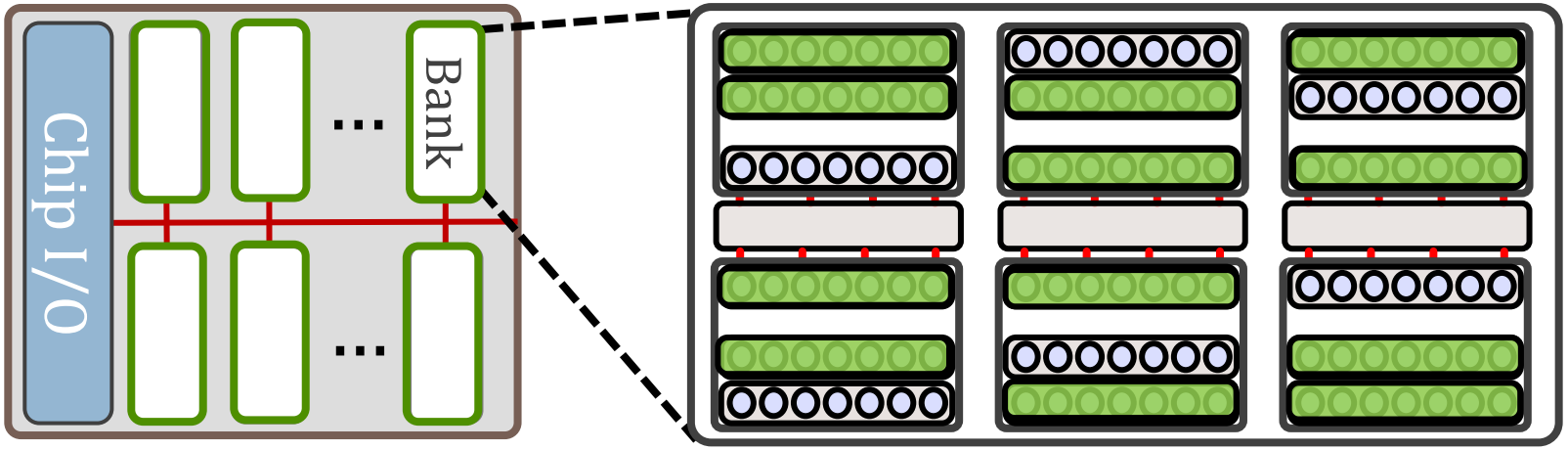
ETH zürich

Experimental Methodology

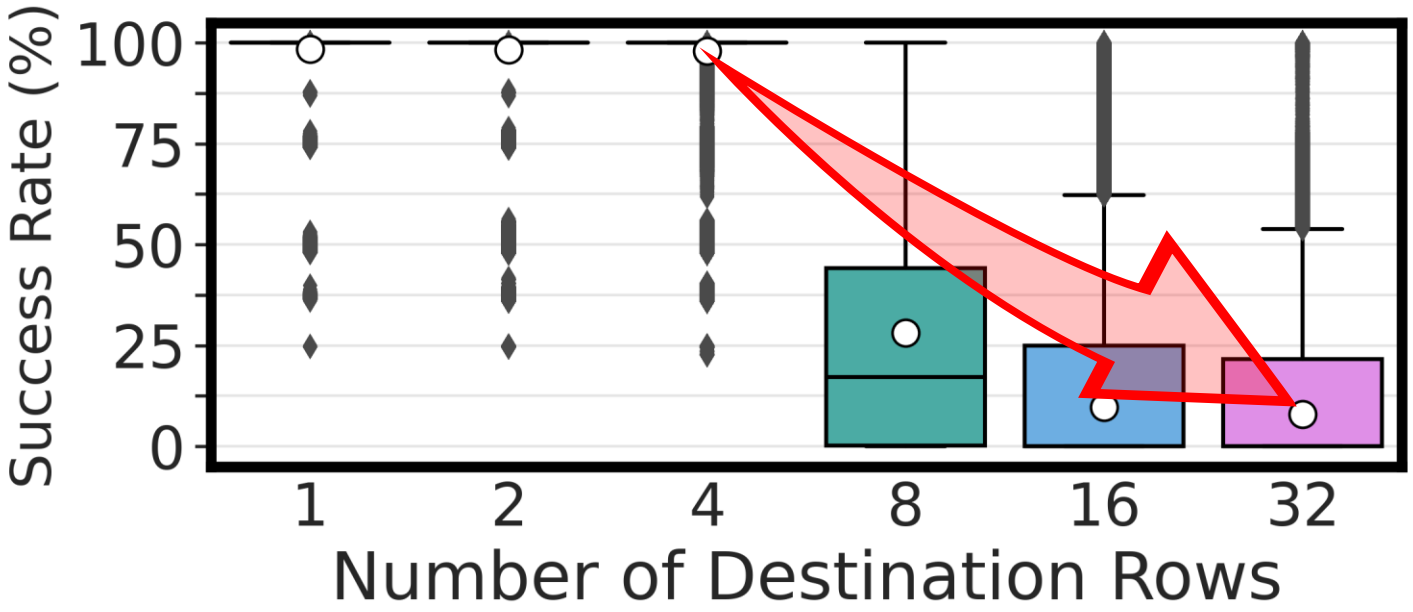
We test **all banks** in **each DRAM chip**

We test **three neighboring subarray pairs** in **each bank**

We test **all possible combinations of activated rows**



Performing NOT in COTS DRAM Chips



The average success of the NOT operation with
four destination row: 98.37%
thirty-two destination rows: 7.95%

As the number of destination rows increases, more DRAM cells produce incorrect results.

The Coverage of Multiple-Row Activation

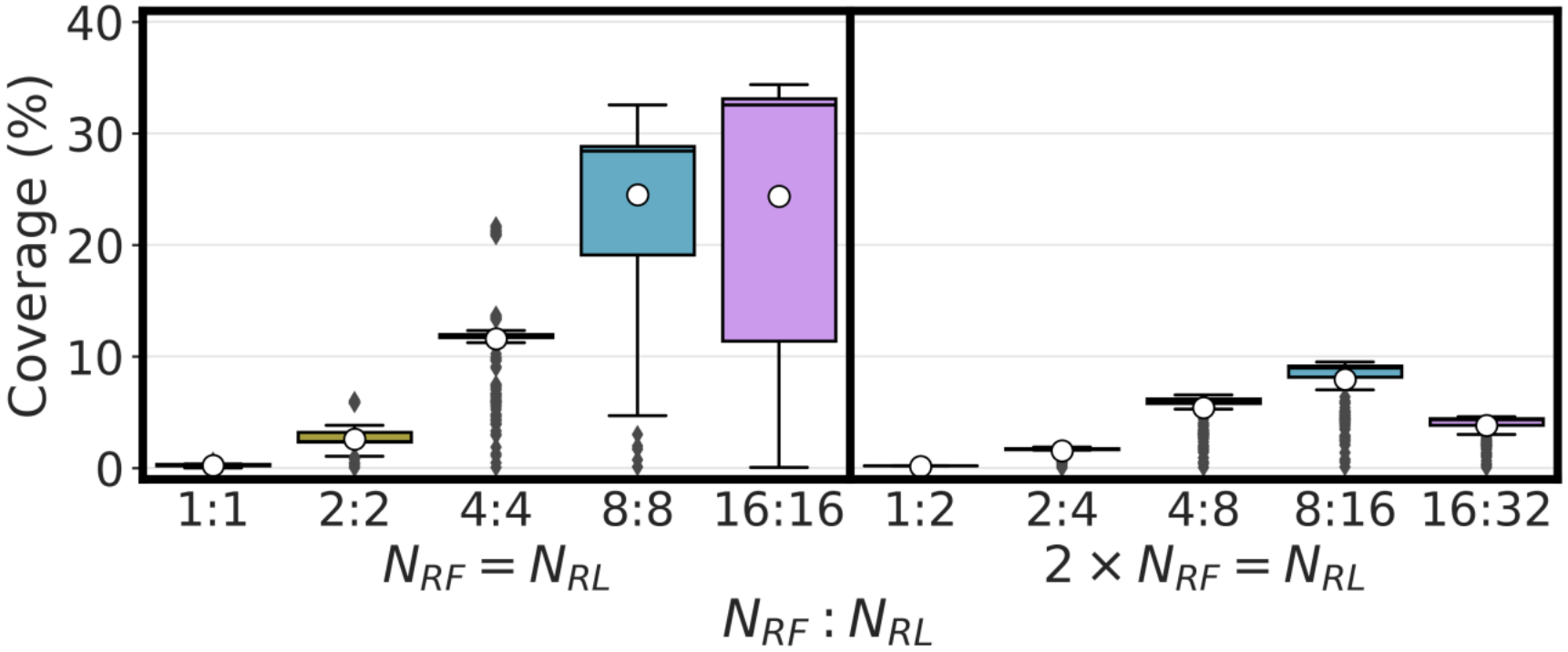


Figure 5: Coverage of each N_{RF}:N_{RL} activation type across tested R_F and R_L row pairs.

NOT vs. Activation Trend

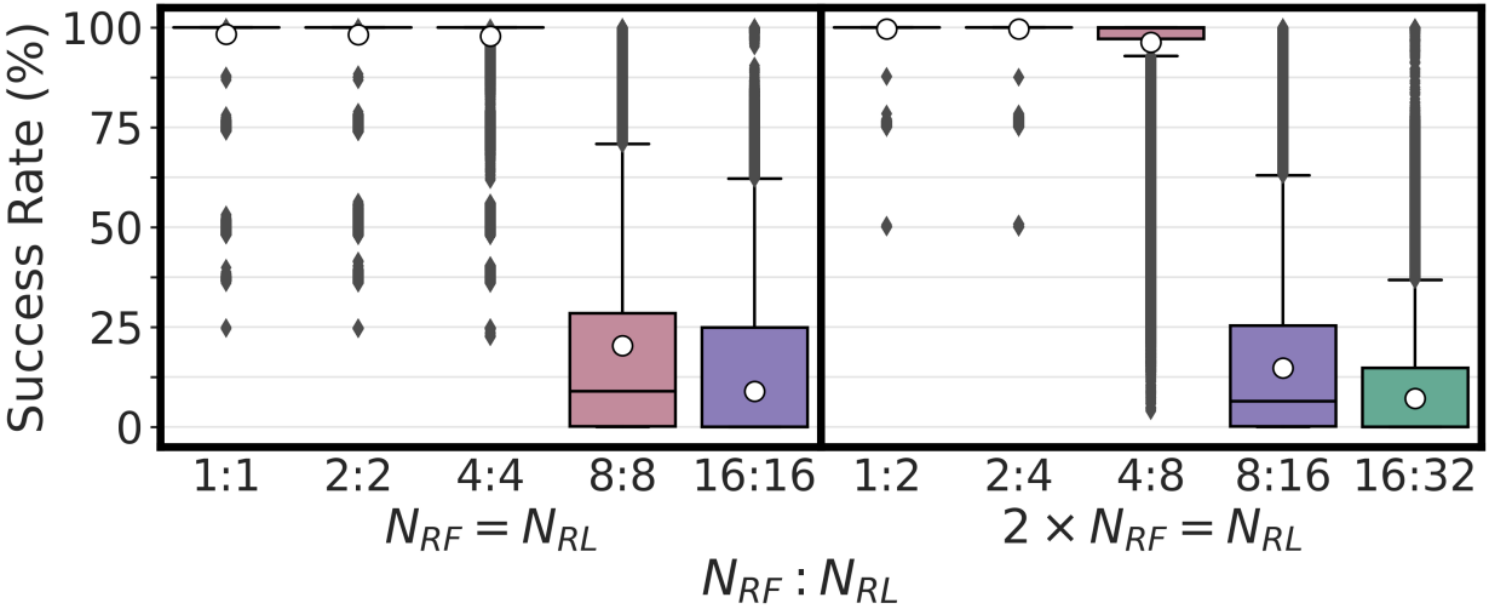
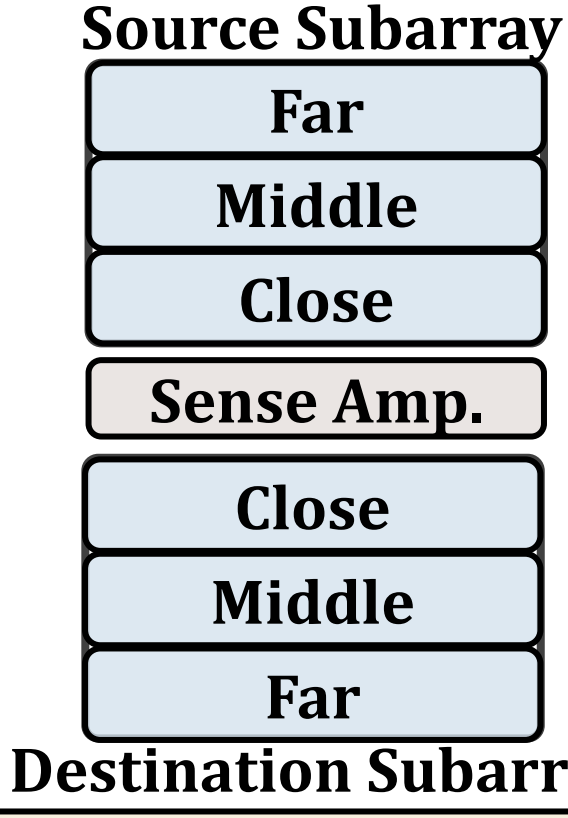
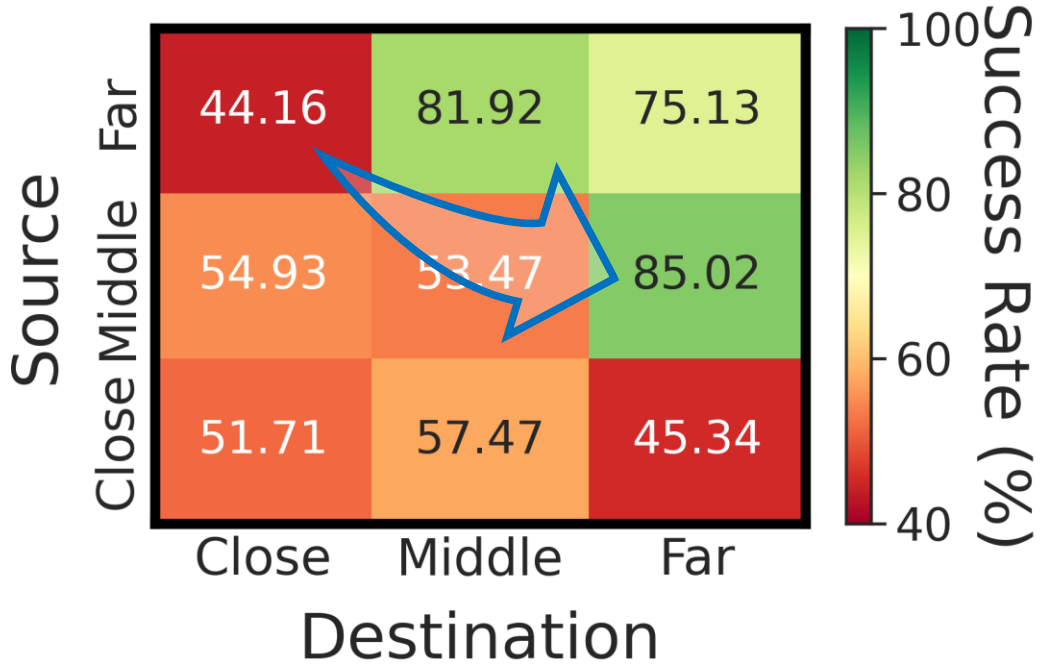


Figure 8: Success rate of the NOT operation vs. N_{RF}:N_{RL} activation type.

Impact of Location in NOT Op.

- Categorize the distance between activated rows (source and destination rows) and the sense amplifiers into three regions: Far, Middle, and Close





The distance between activated rows and the sense amplifiers significantly affects the reliability

The effect of DRAM Speed Rate on NOT

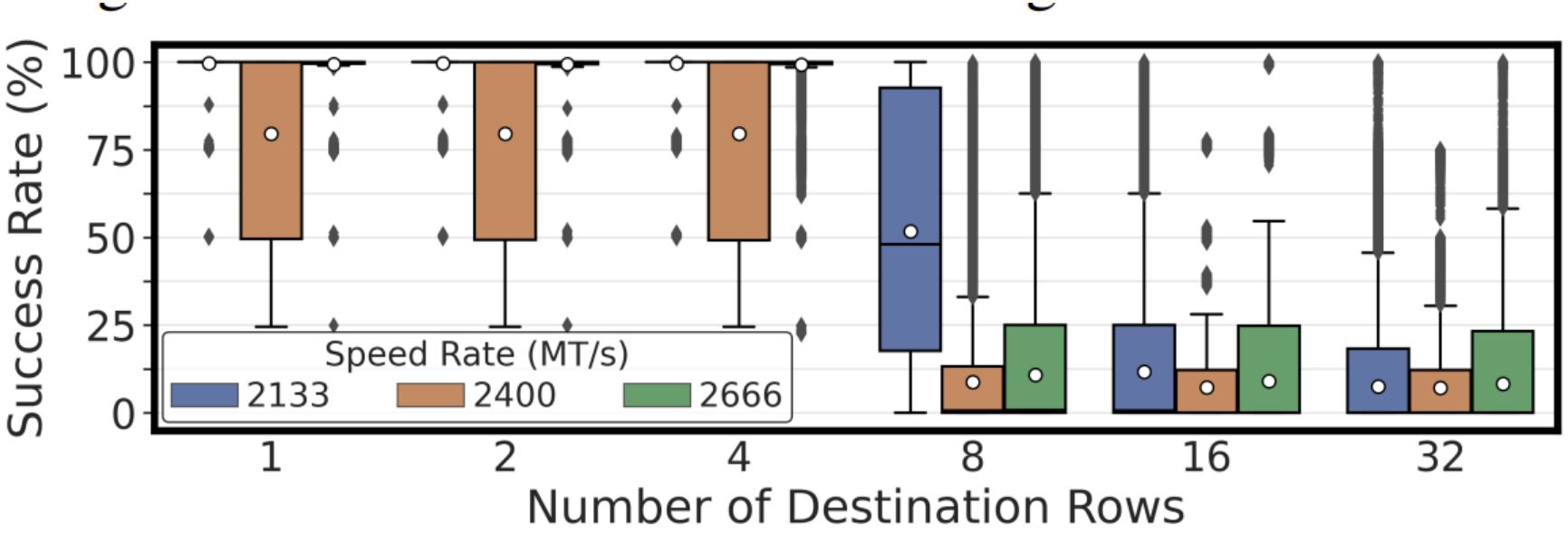


Figure 11: Success rate of the NOT operation for different DRAM speed rates.

Chip Density & Die Revision (NOT)

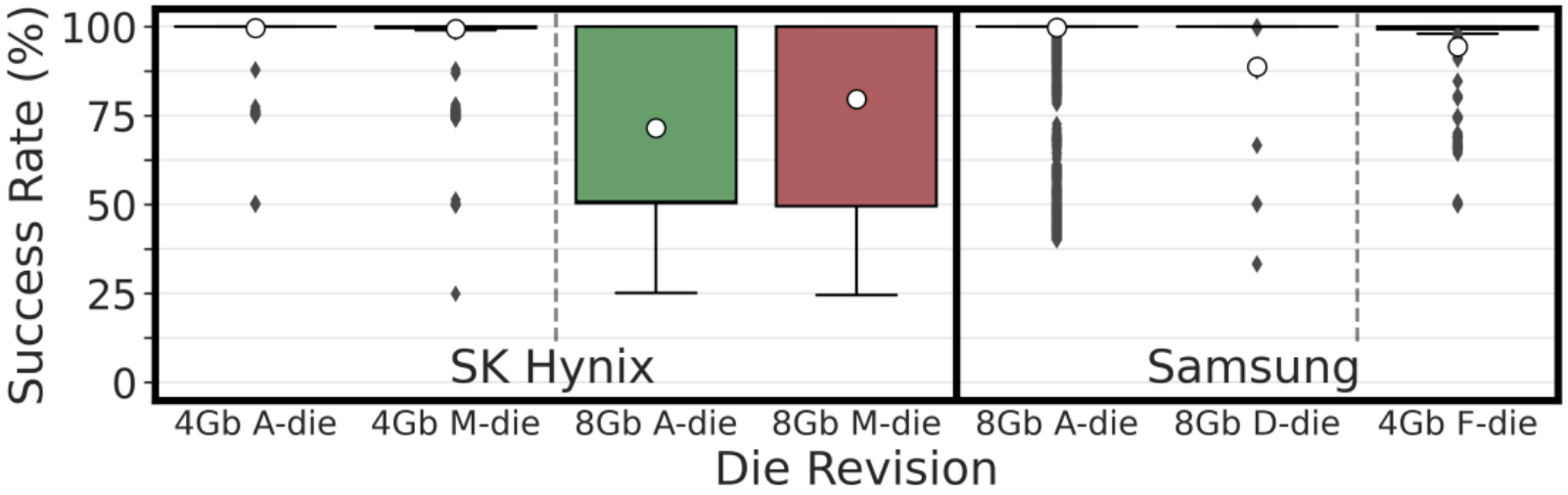
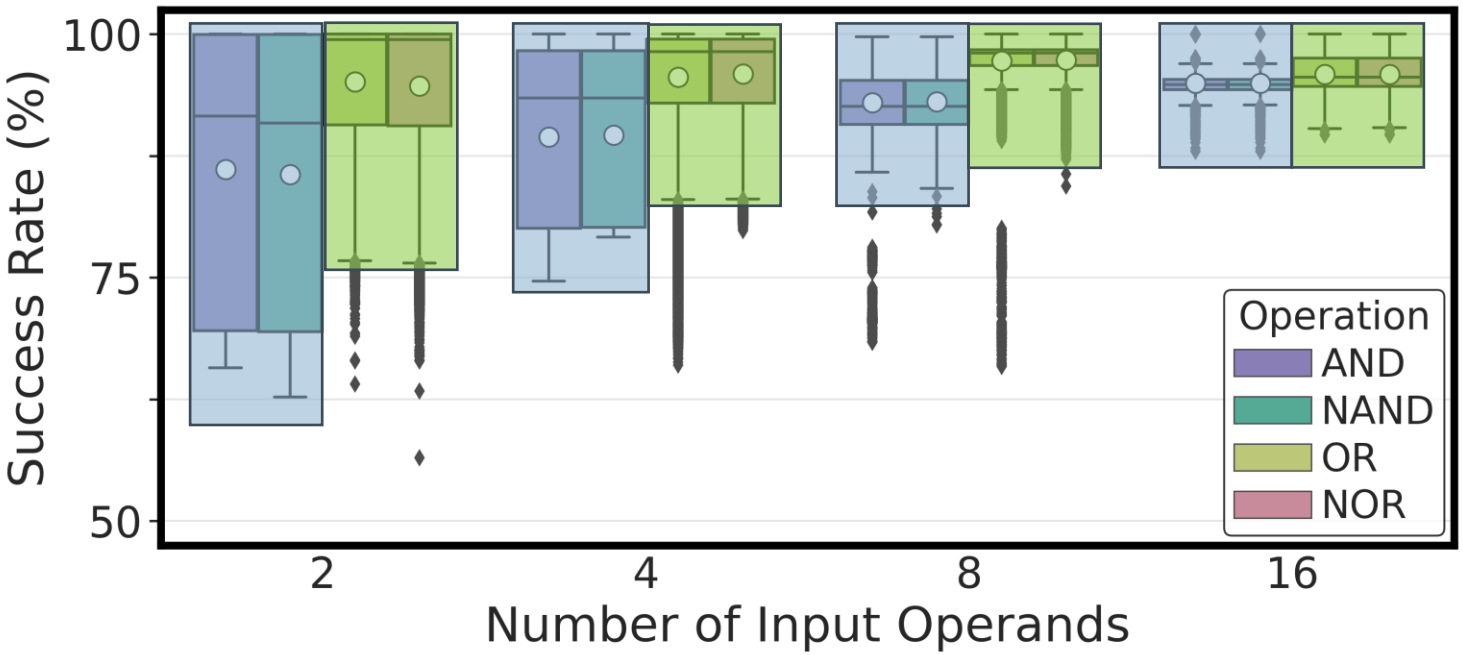


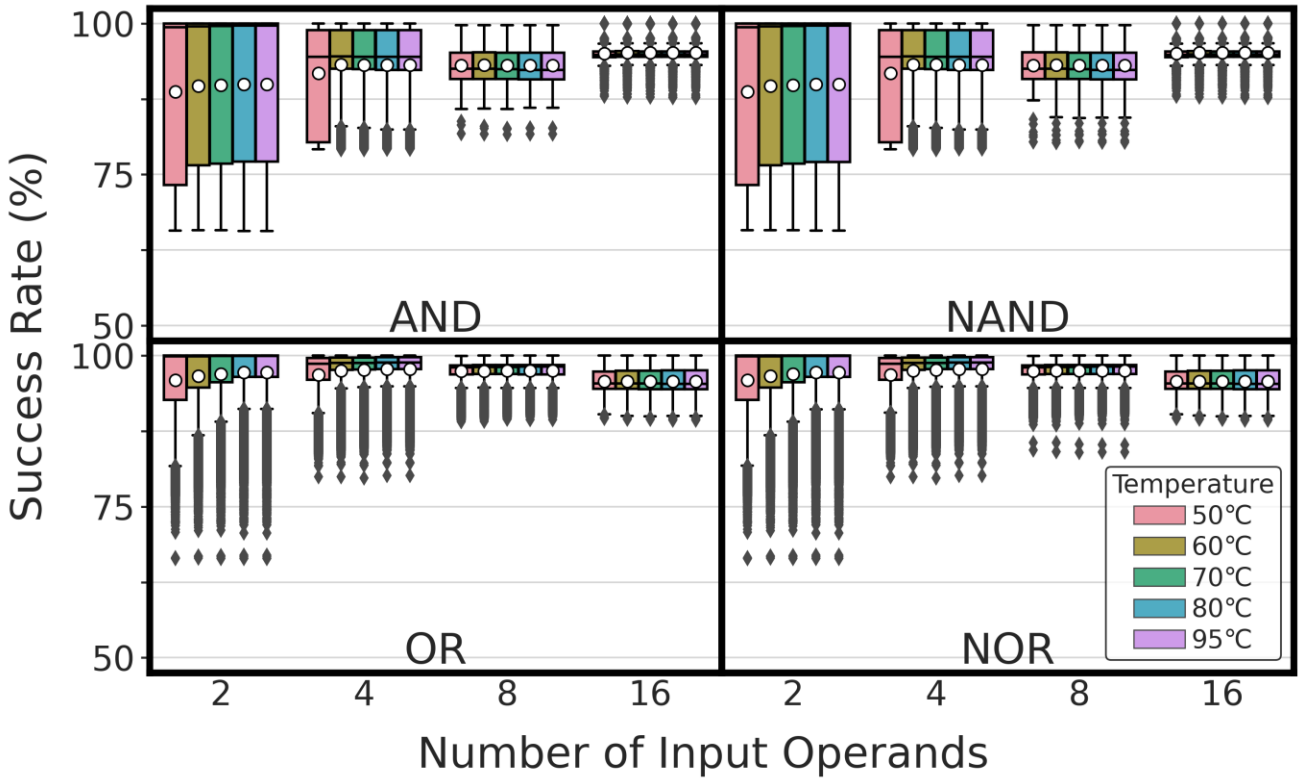
Figure 12: Success rate of the NOT operation for different chip density and die revision combinations for two major manufacturers.

Performing AND, NAND, OR, and NOR



The reliability distributions are very similar between 1) AND-NAND and 2) OR – NOR operations.

Impact of Temperature



Temperature has a small effect on the reliability of AND, NAND, OR, and NOR operations

Boolean Operations vs. Number of 1s

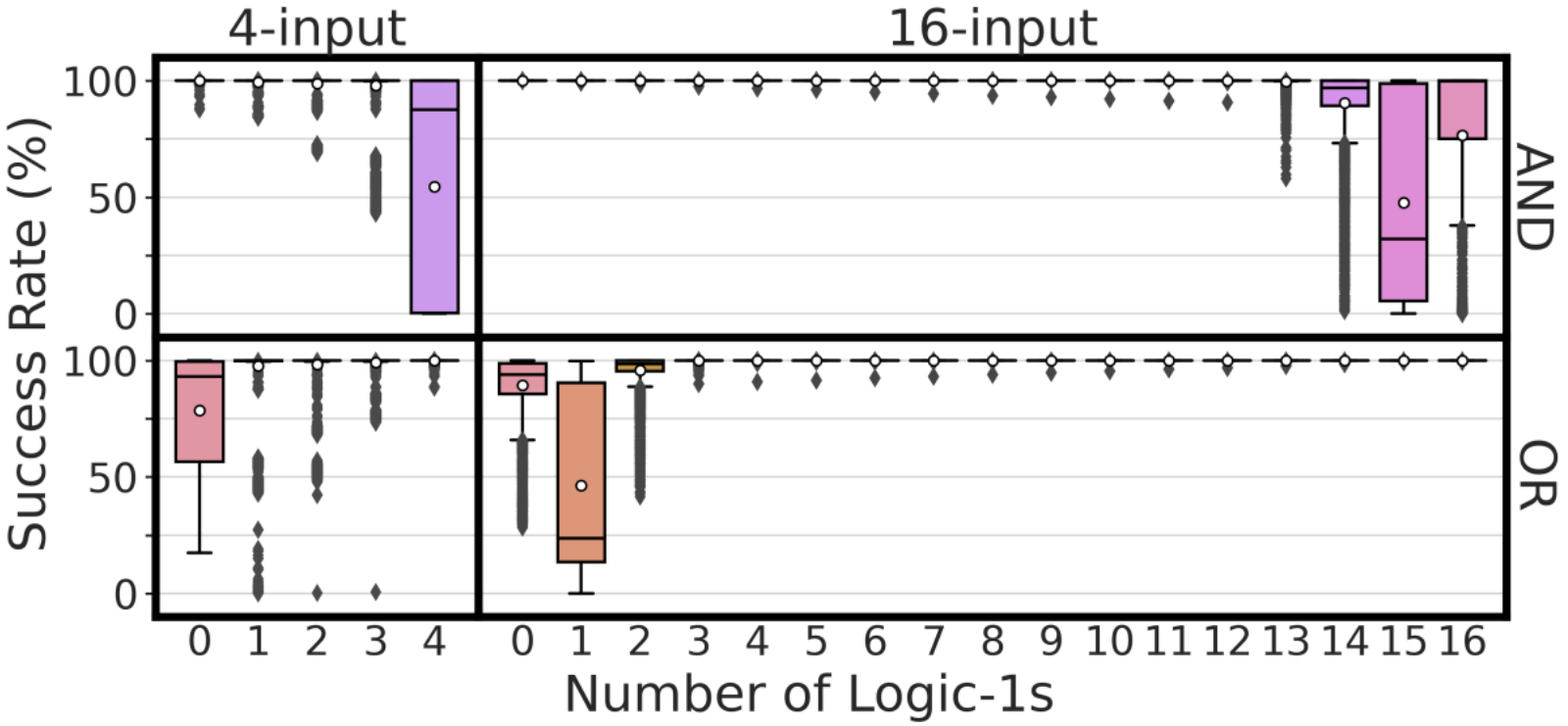
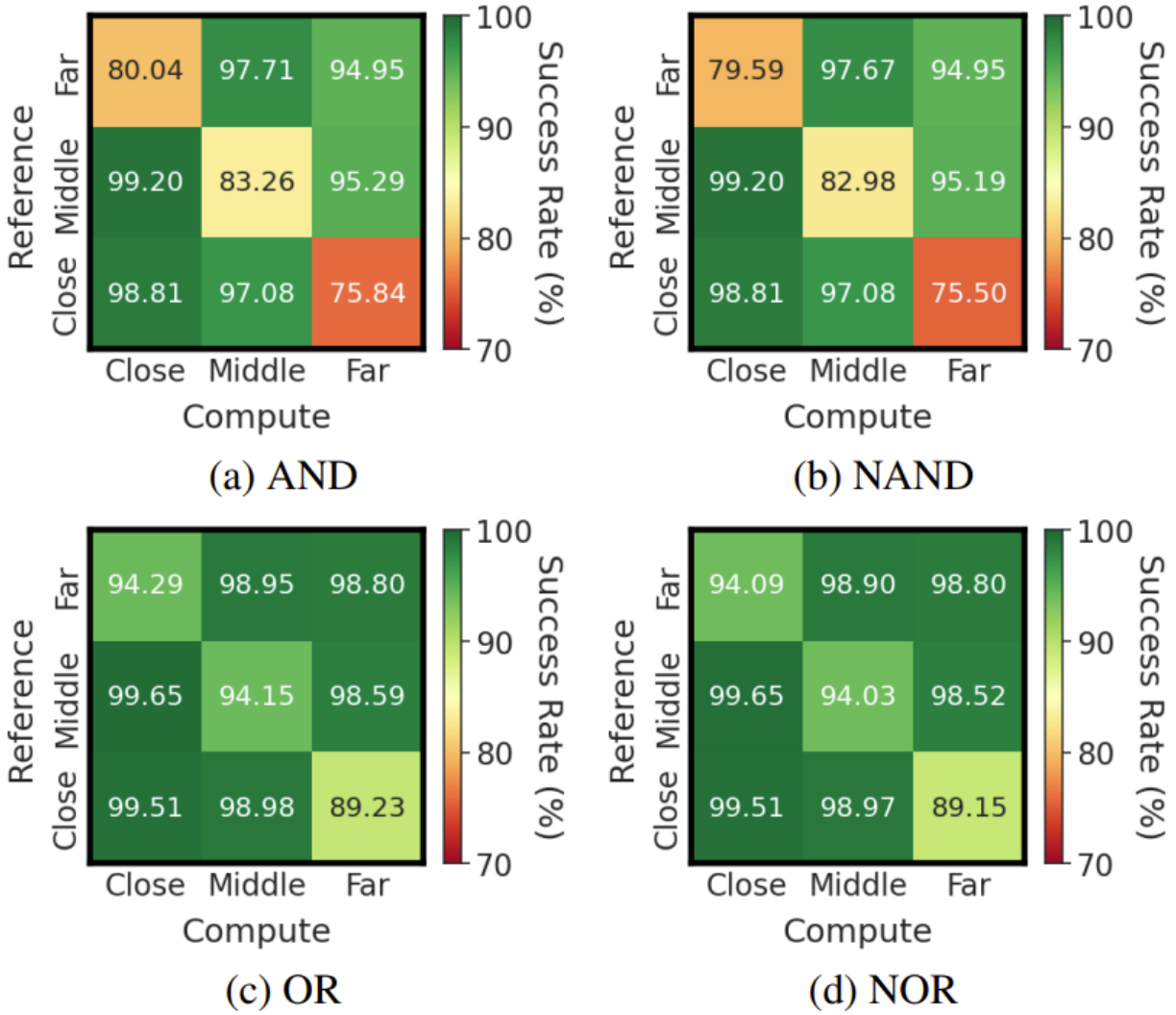


Figure 16: Success rates of AND and OR operations based on the number of logic-1s in the input operands.

The Effect of the Location



DRAM Speed Rate vs. Bitwise Ops.

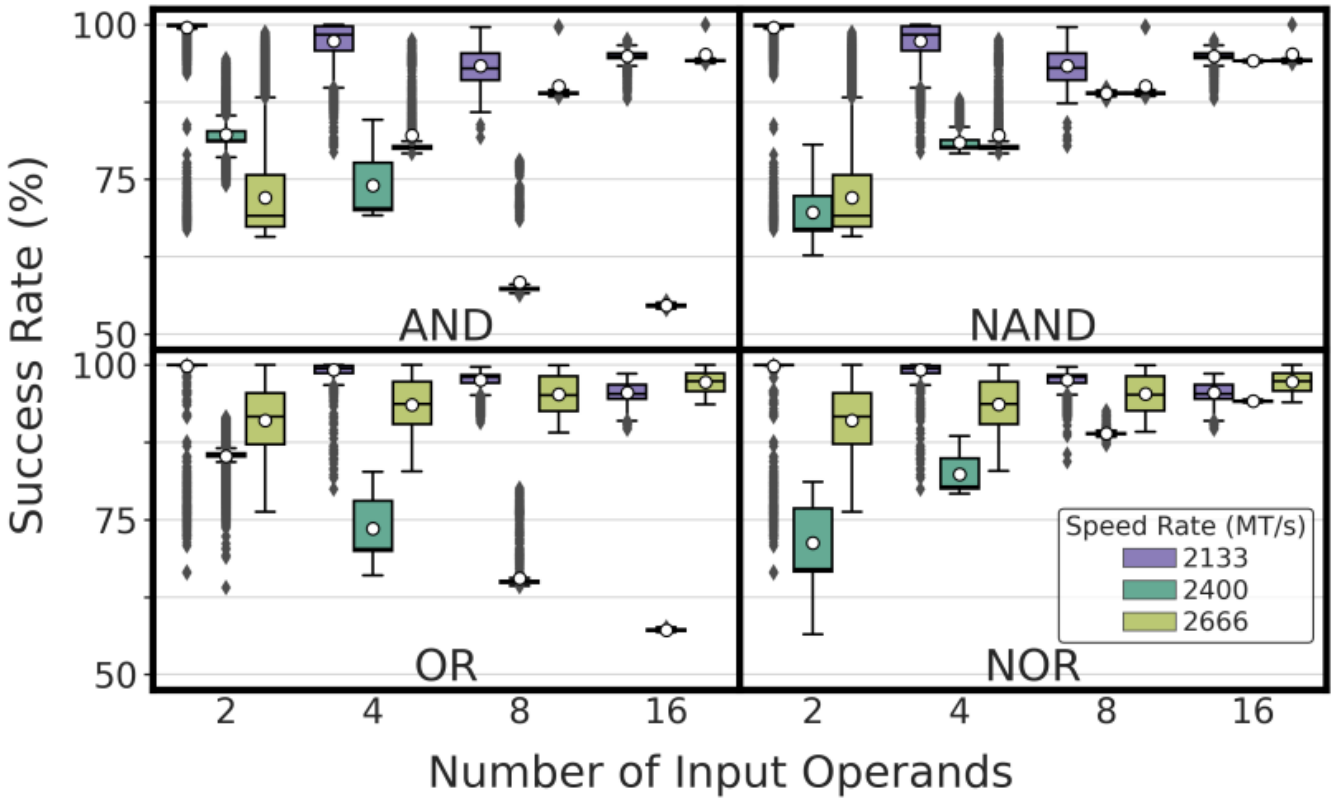
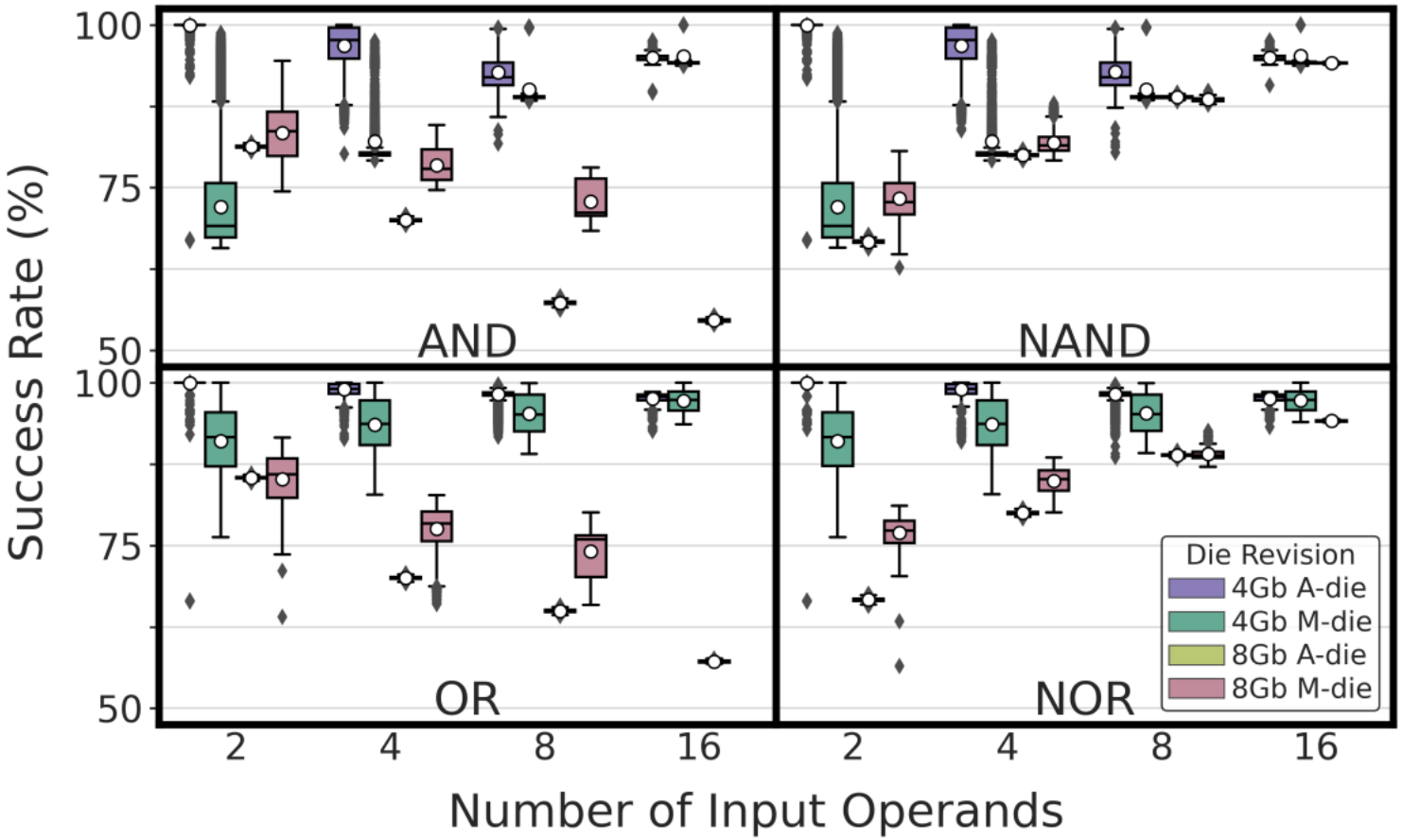
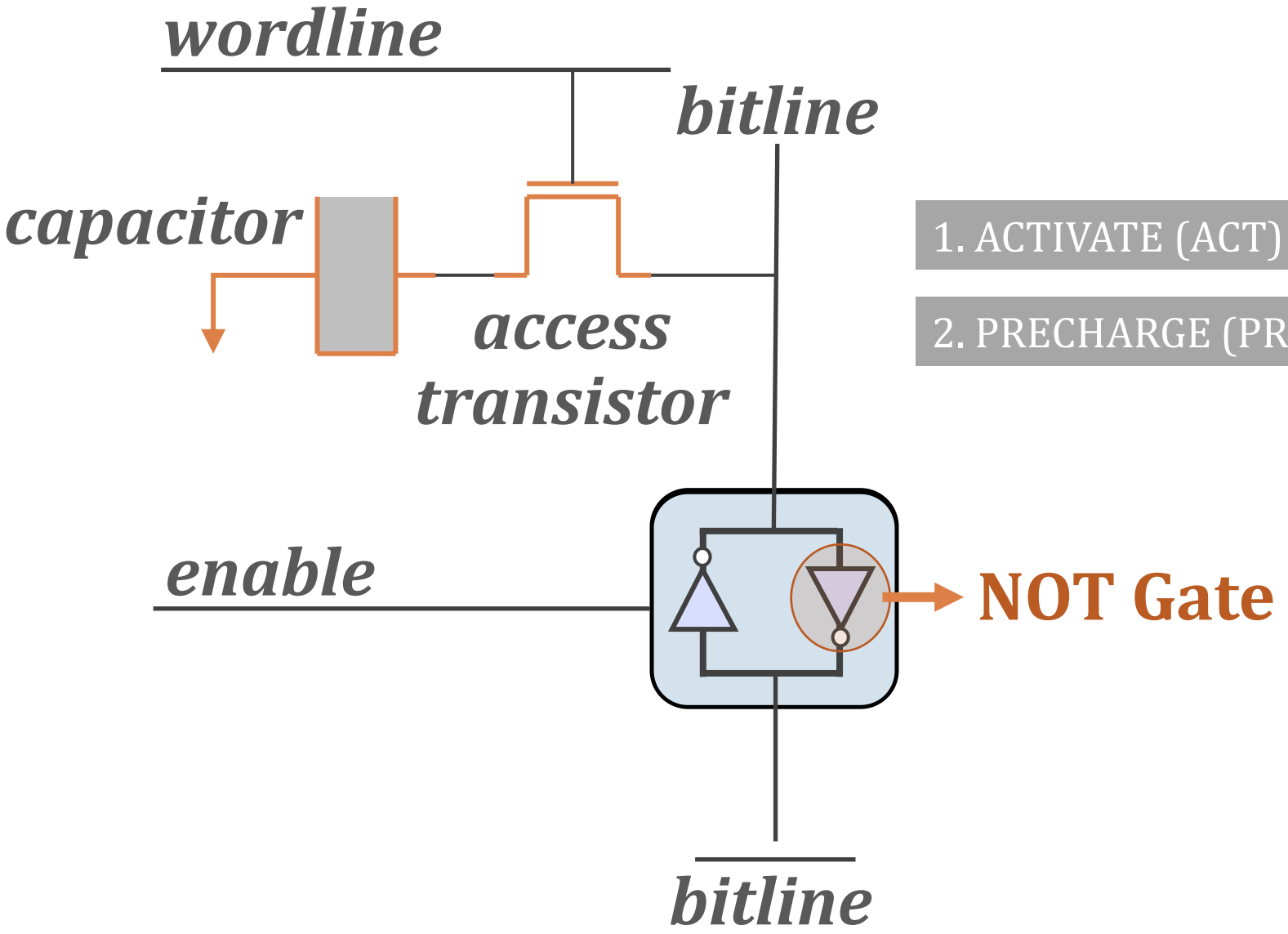


Figure 20: Success rates of AND, NAND, OR, and NOR operations for three DRAM speed rates.

Chip Density&Die Revision vs. Bitwise Ops.

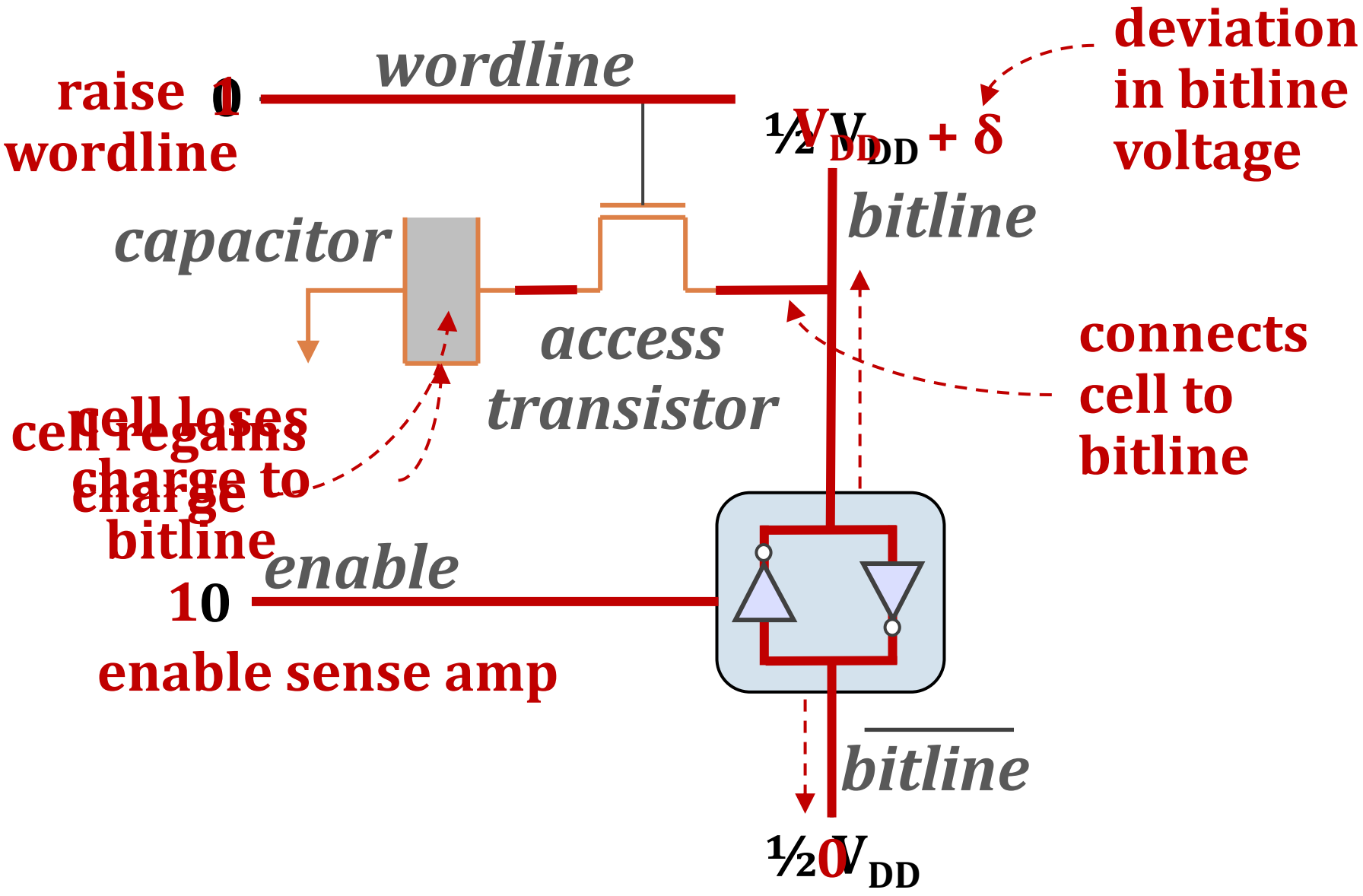


DRAM Cell Operation

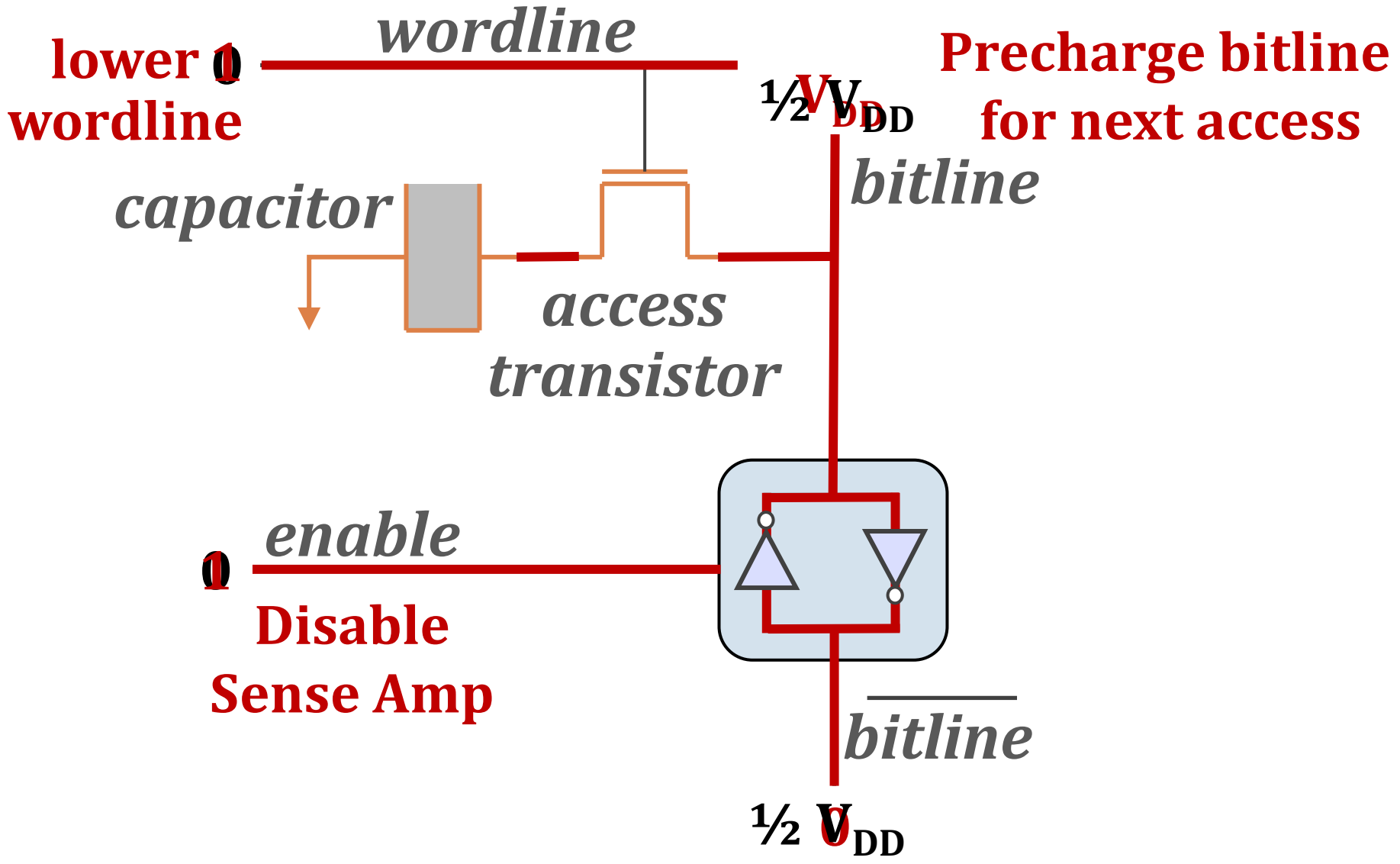


1. ACTIVATE (ACT)
2. PRECHARGE (PRE)

DRAM Cell Operation - ACTIVATE



DRAM Cell Operation - PRECHARGE



Simultaneous Many-Row Activation in Off-the-Shelf DRAM Chips

Experimental Characterization and Analysis

İsmail Emir Yüksel

Yahya C. Tuğrul F. Nisa Bostancı Geraldo F. Oliveira

A. Giray Yağlıkçı Ataberk Olgun Melina Soysal Haocong Luo

Juan Gómez–Luna Mohammad Sadr Onur Mutlu

SAFARI

ETH zürich

Executive Summary

Motivation:

- **Processing-Using-DRAM (PUD)** alleviates **data movement bottlenecks**
- Commercial off-the-shelf (COTS) DRAM chips can perform **three-input majority (MAJ3)** and **in-DRAM copy** operations

Goal: To experimentally analyze and understand

- The **computational capability** of COTS DRAM chips beyond that of prior works
- The **robustness** of such capability under various **operating conditions**

Experimental Study: 120 DDR4 chips from **two major manufacturers**

- COTS DRAM chips can perform **MAJ5, MAJ7, and MAJ9** operations and **copy** one DRAM row to **up to 31 different rows** at once
- Storing **multiple redundant copies** of MAJ's input operands (i.e., input replication) drastically increases **robustness** (>30% higher success rate)
- **Operating conditions** (temperature, voltage, and data pattern) **affect** the robustness of in-DRAM operations (by up to 11.52% success rate)

Outline

Motivation & Background

Goal

Experimental Methodology

Simultaneous Many-Row Activation

MAJX Operation

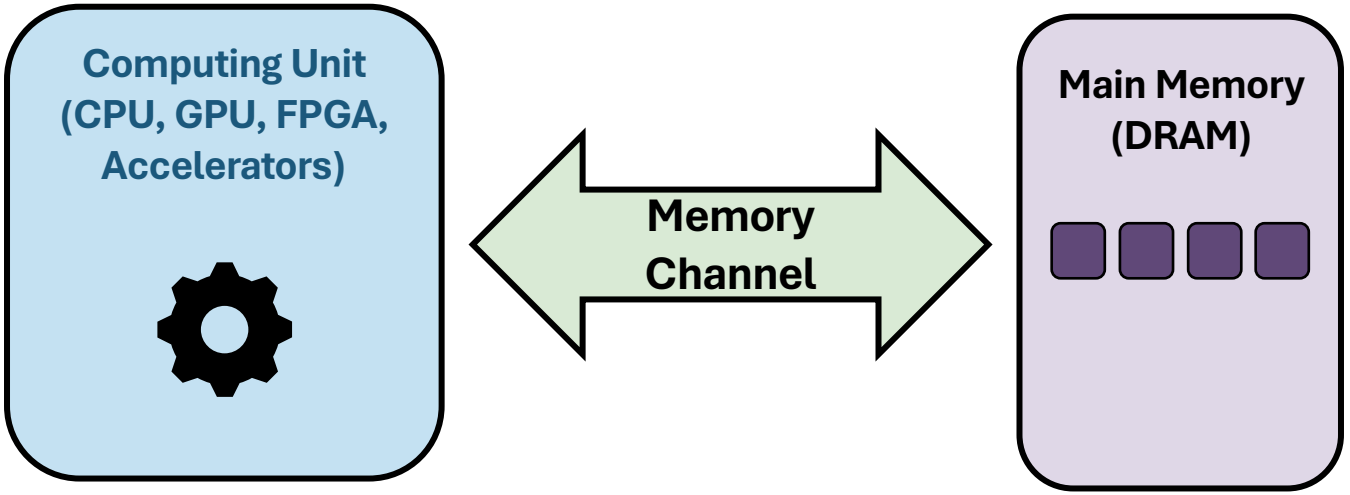
Multi-RowCopy Operation

Conclusion

Outline

Data Movement Bottleneck

- Today's computing systems are processor centric
- All data is processed in the processor → at great system cost



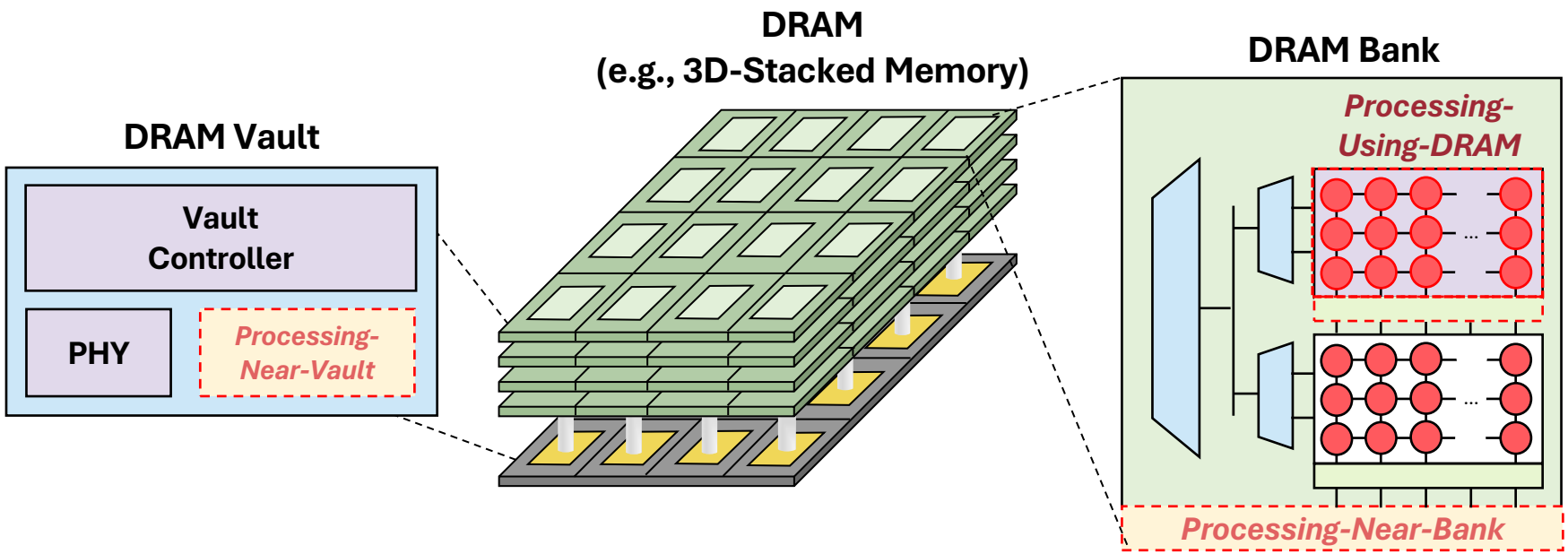
More than 60% of the total system energy is spent on data movement[1]

[1] A. Boroumand et al., "Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks," ASPLOS, 2018

Processing-In-Memory (PIM)

Two main approaches for Processing-In-Memory:

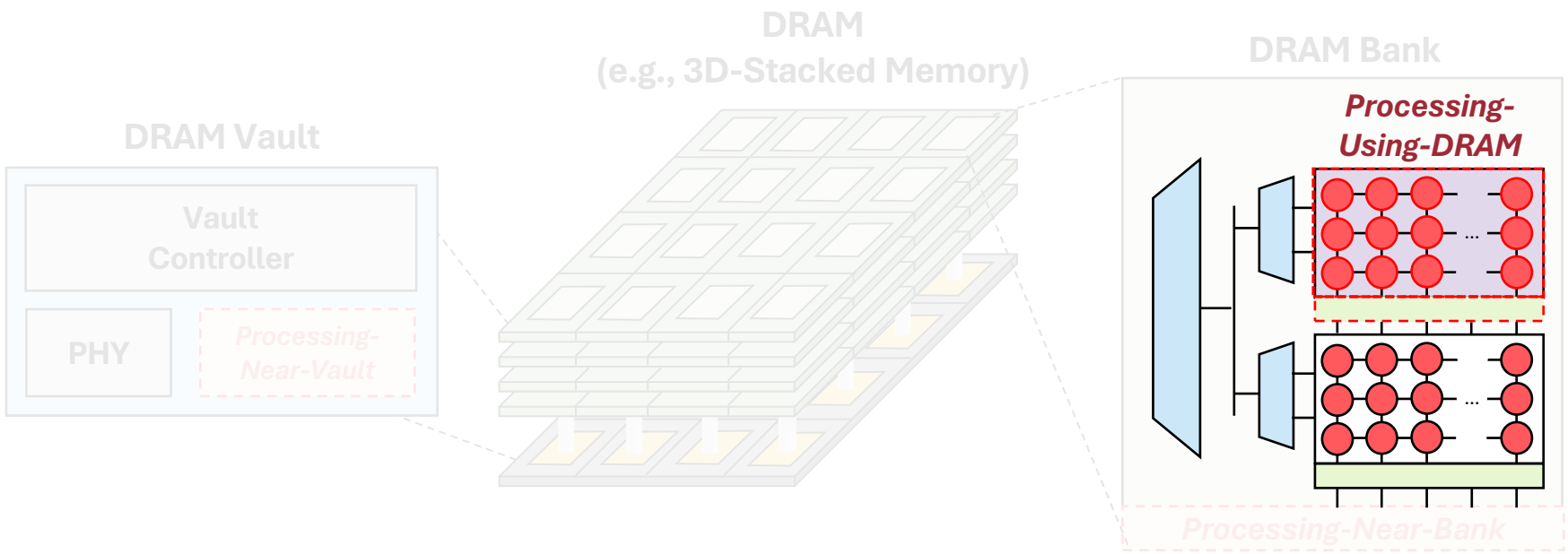
1. **Processing-Near-Memory**: PIM logic is added near the memory arrays or to the logic layer of 3D-stacked memory
2. **Processing-Using-Memory**: uses the analog operational principles of memory cells to perform computation



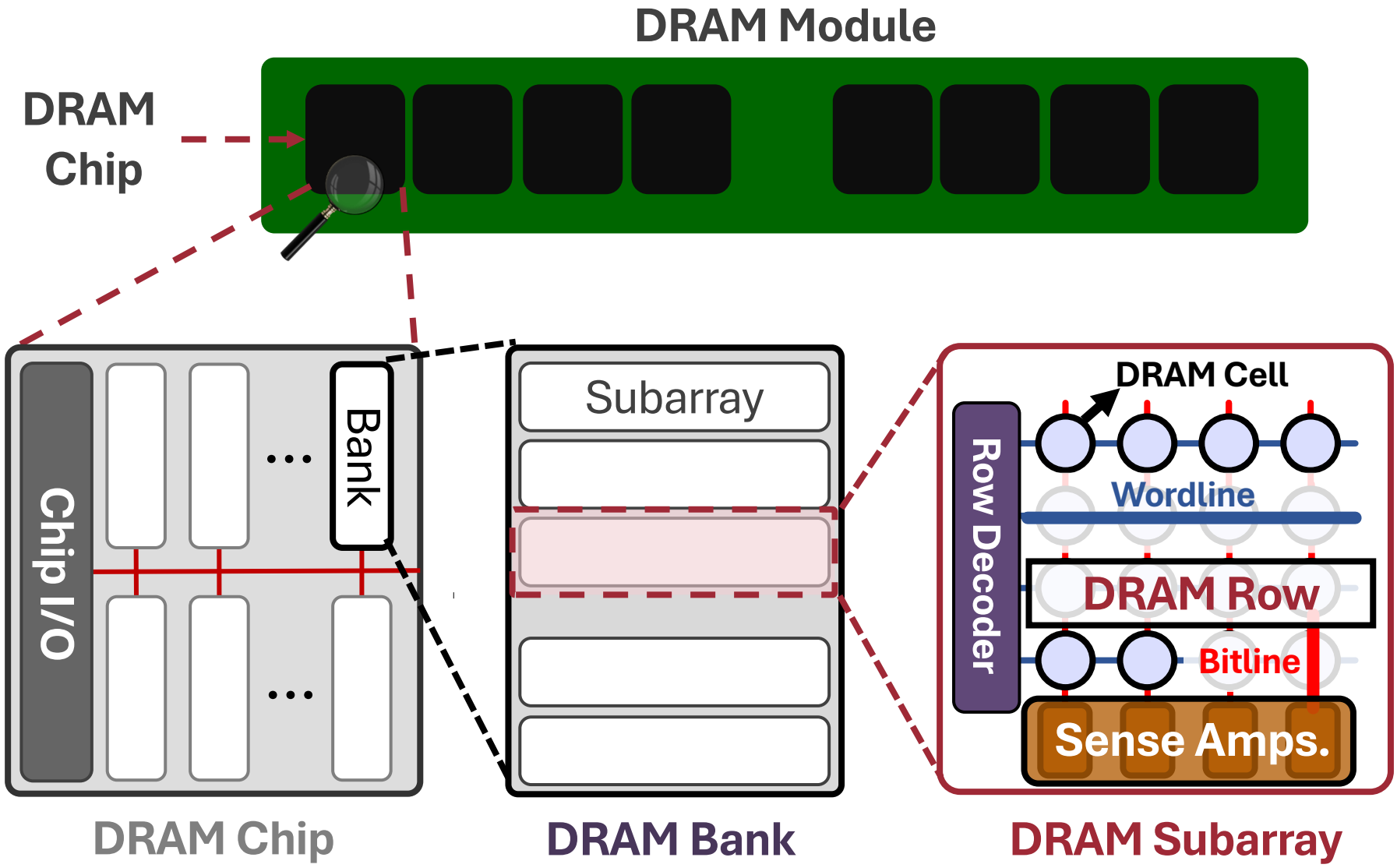
Processing-In-Memory (PIM)

Two main approaches for Processing-In-Memory:

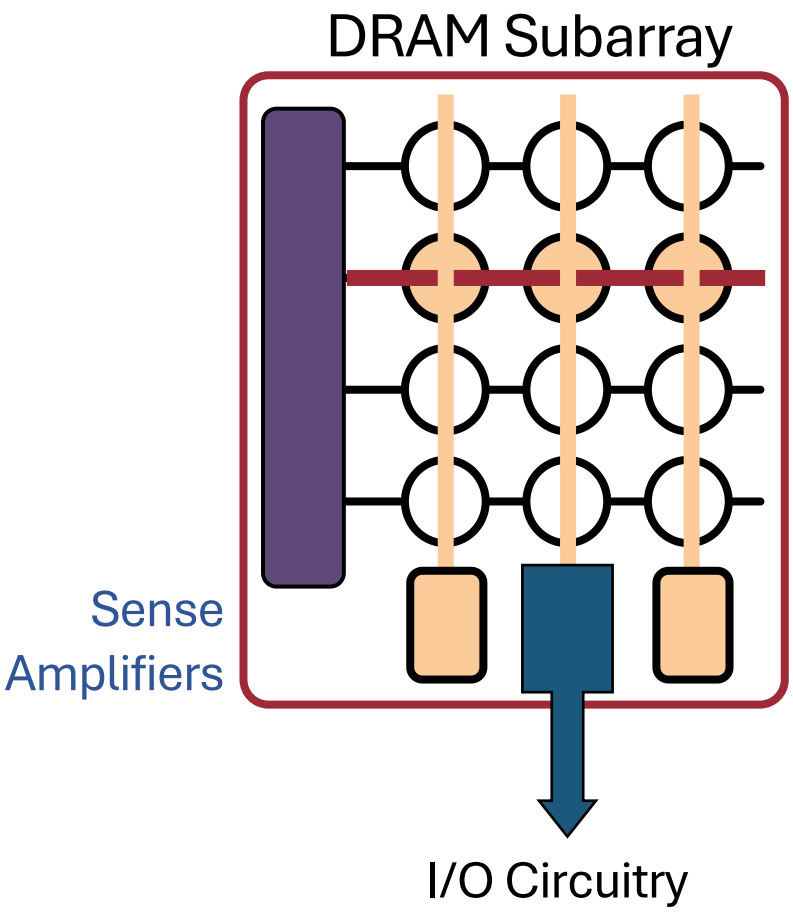
1. **Processing-Near-Memory**: PIM logic is added near the memory arrays or to the logic layer of 3D-stacked memory
2. **Processing-Using-Memory**: uses the analog operational principles of memory cells to perform computation



DRAM Organization



DRAM Operation



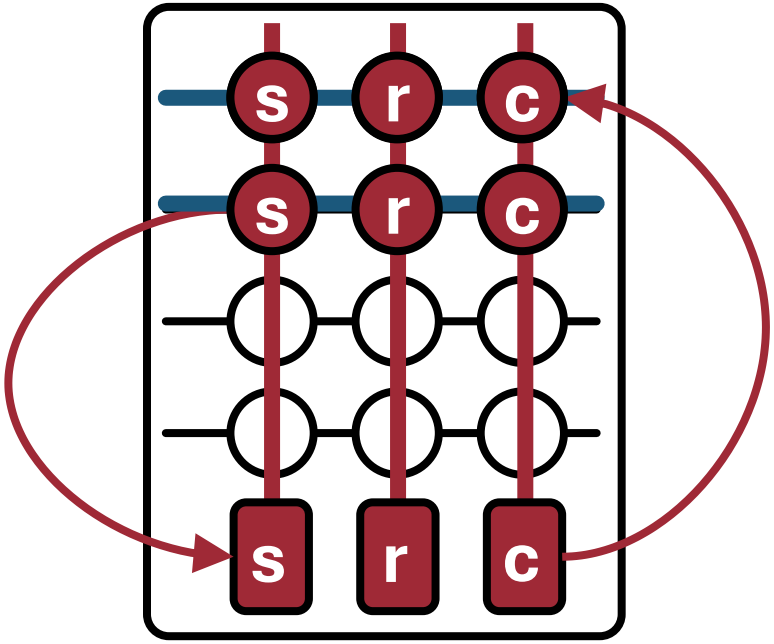
1 **ACTIVATE (ACT):**
Fetch the row's content into the **sense amplifiers**

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

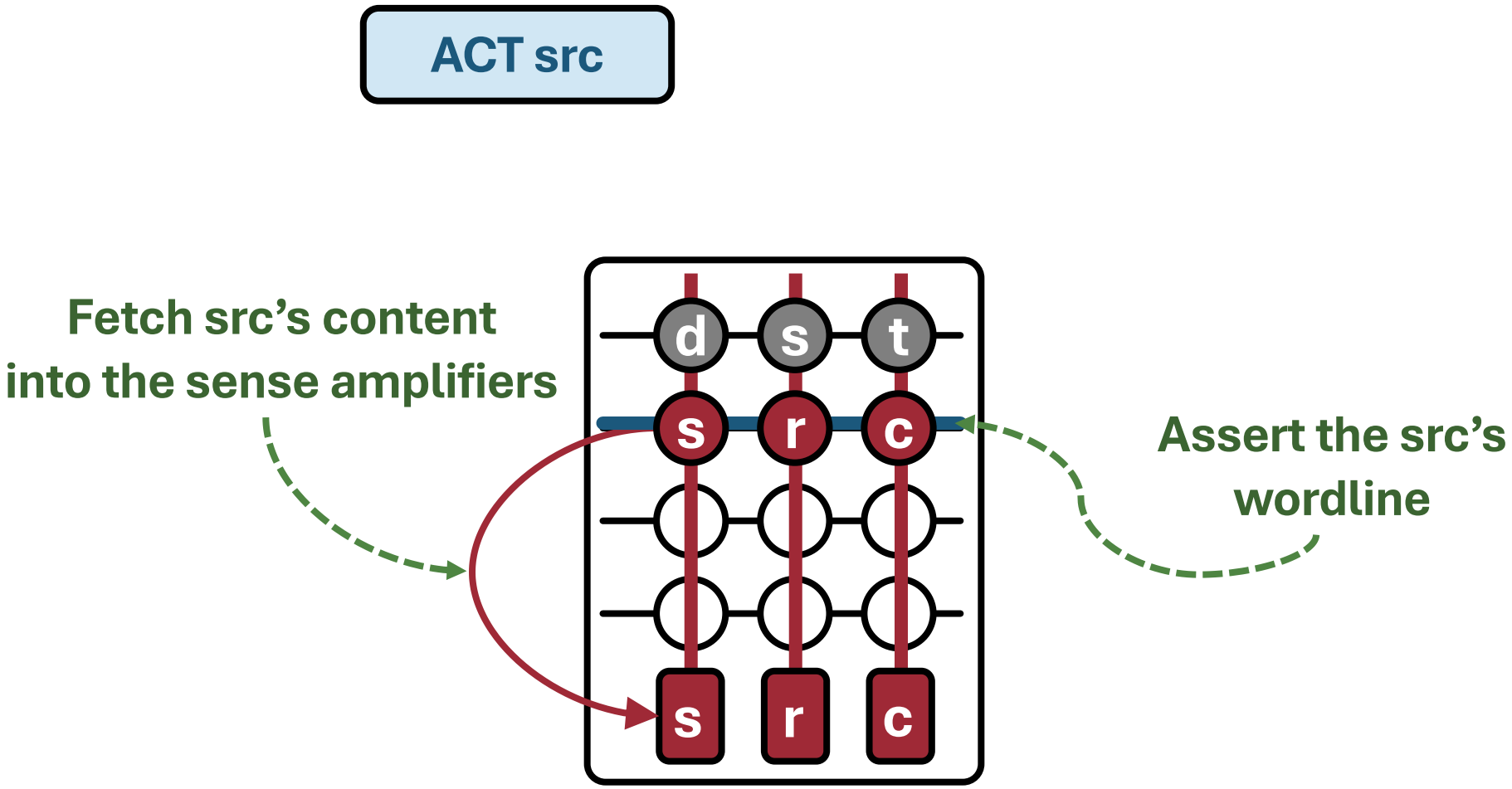
In-DRAM Row-Copy (RowClone)

Copying the source (src) row's content to the destination (dst) row



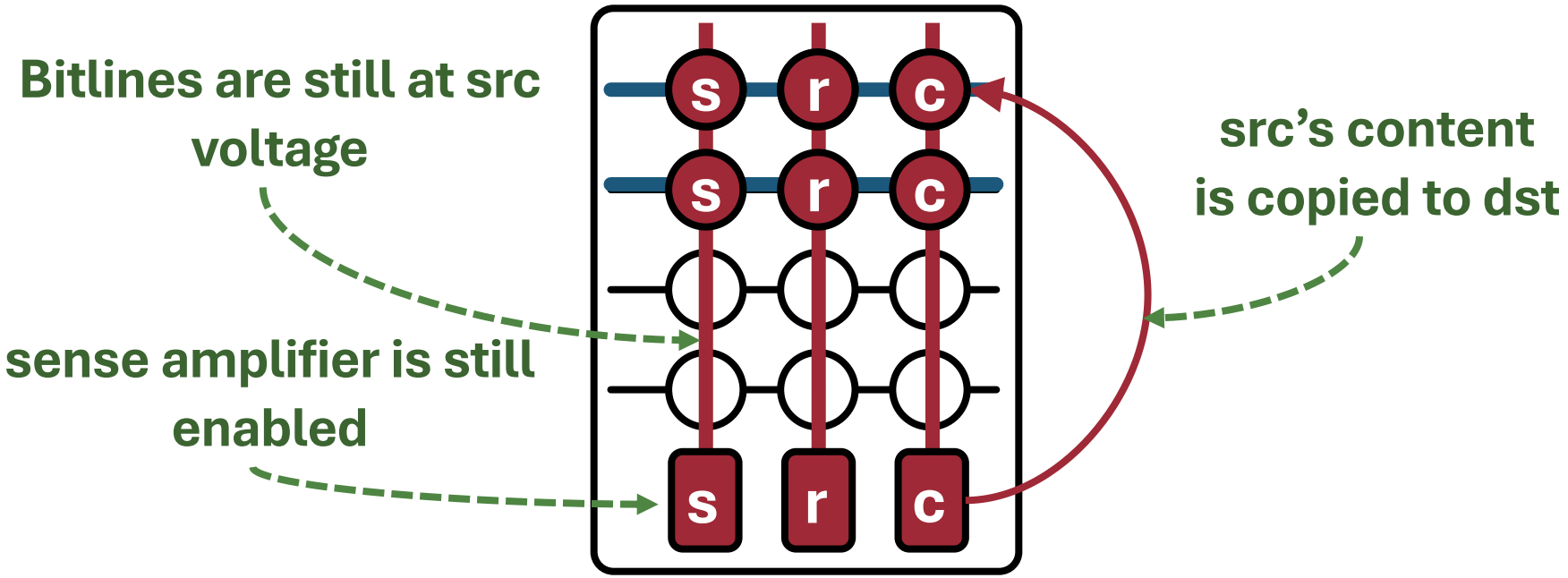
[Seshadri+ MICRO'13]

In-DRAM Row-Copy (RowClone)



[Seshadri+ MICRO'13]

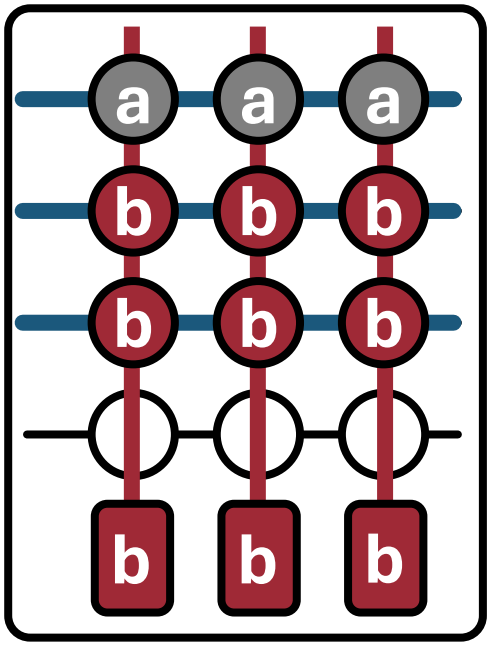
In-DRAM Row-Copy (RowClone)



[Seshadri+ MICRO'13]

In-DRAM Majority-of-Three (MAJ3)

Performing a MAJ3 operation using three rows as input operands

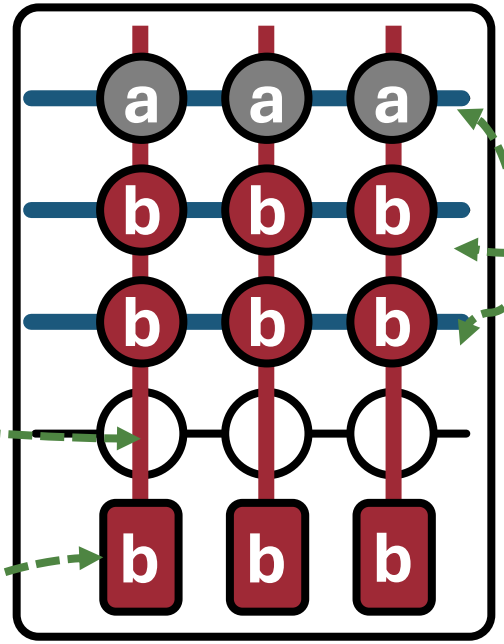


MAJ3(a, b, b) = b

[Seshadri+ MICRO'17]

In-DRAM Majority-of-Three (MAJ3)

Activate three rows simultaneously



Three rows perturb bitlines simultaneously

Assert three rows' wordlines

Sense amplifiers sample the output of majority

[Seshadri+ MICRO'17, Gao+ MICRO'19]

Outline

Motivation & Background

Goal

Experimental Methodology

Simultaneous Many-Row Activation

MAJX Operation

Multi-RowCopy Operation

Conclusion

Our Goal

Experimentally understand
the **computational capability** of COTS DRAM chips

Experimentally analyze
the **robustness** of such capability
under various **operating conditions**

Outline

Motivation & Background

Goal

Experimental Methodology

Simultaneous Many-Row Activation

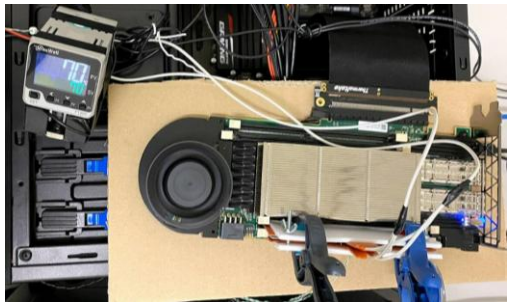
MAJX Operation

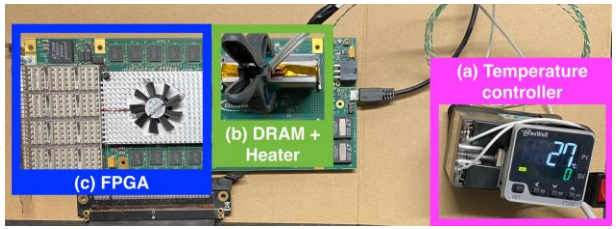
Multi-RowCopy Operation

Conclusion

DRAM Testing Infrastructure (I)

DRAM Bender DDR3/4 Testing Infrastructure





https://github.com/CMU-SAFARI/DRAM-Bender

CMU-SAFARI / **DRAM-Bender**

Code | Issues 1 | Pull requests 1

DRAM-Bender Public

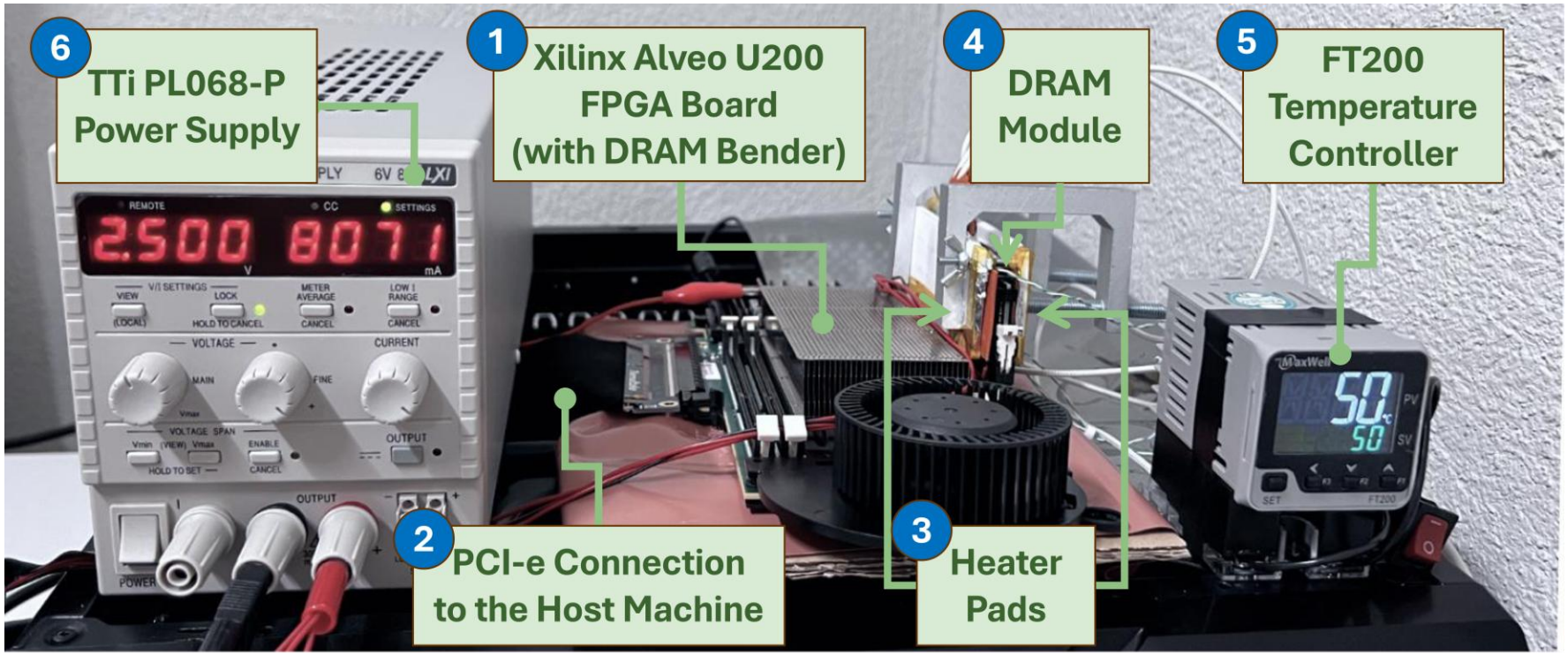
About

DRAM Bender is the first open source DRAM testing infrastructure that can be used to easily and comprehensively test state-of-the-art DDR4 modules of different form factors. Five prototypes are available on different FPGA boards.

Olgun et al., "DRAM Bender: An Extensible and Versatile FPGA-based Infrastructure to Easily Test State-of-the-art DRAM Chips," TCAD, 2023.

DRAM Testing Infrastructure (II)

Fine-grained control over DRAM commands, timings, temperature, and voltage



Olgun et al., "DRAM Bender: An Extensible and Versatile FPGA-based Infrastructure to Easily Test State-of-the-art DRAM Chips," TCAD, 2023.

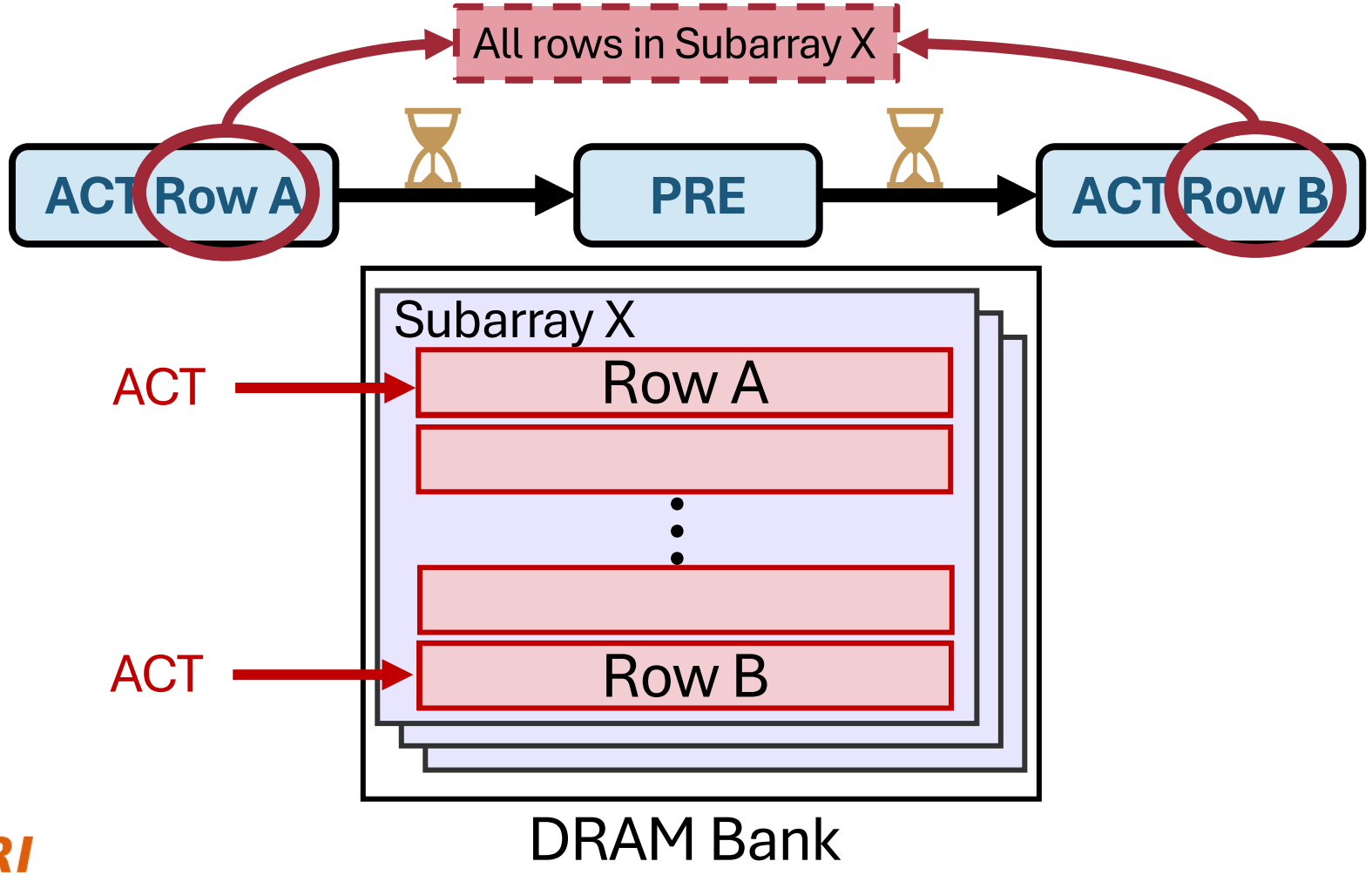
DRAM Chips Tested

- 120 DDR4 chips from two major DRAM manufacturers
- Covers different die revisions and chip densities

DRAM Mfr.	#Modules	#Chips	Die Rev.	Density	Org.	Subarray Size
SK Hynix	7	56	M	4Gb	x8	512 or 640
(Mfr. H)	5	40	A	4Gb	x8	512
Micron	4	16	E	16Gb	x16	1024
(Mfr. M)	2	8	B	16Gb	x16	1024

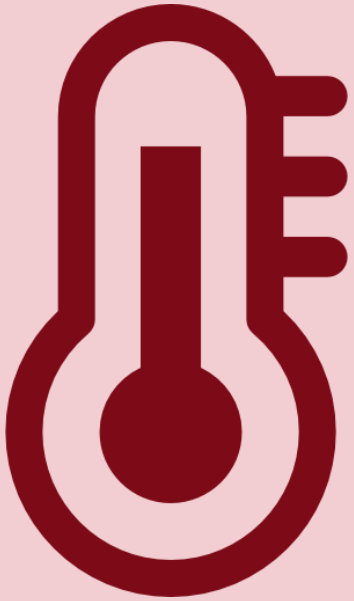
Testing Methodology (I)

- Carefully sweep
  - Row addresses: Row A and Row B (>3M row pairs)
  - Timing parameters: Between ACT → PRE and PRE → ACT



Testing Methodology (II)

Temperature



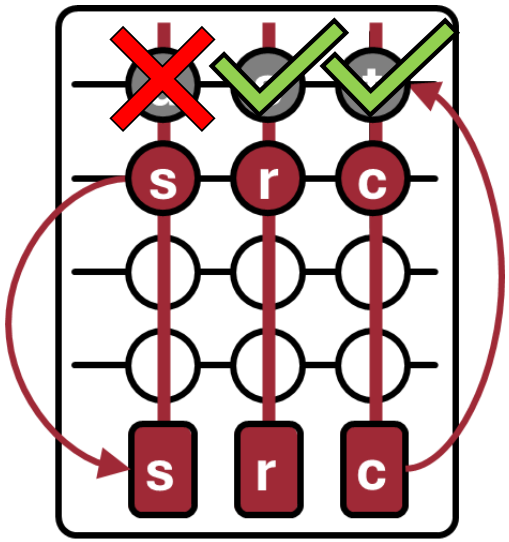
50°C → 90°C

Wordline Voltage

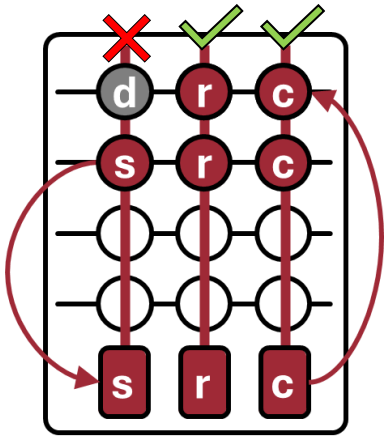
2.5V → 2.1V

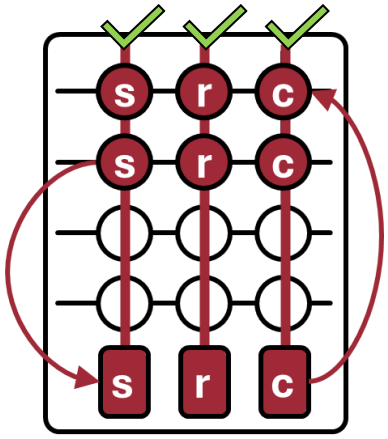
Robustness Metric: Success Rate

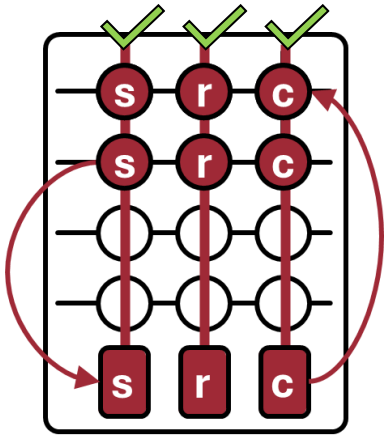
Percentage of DRAM cells that produce **correct output** of a tested operation in **all test trials**



Total of 10,000 trials





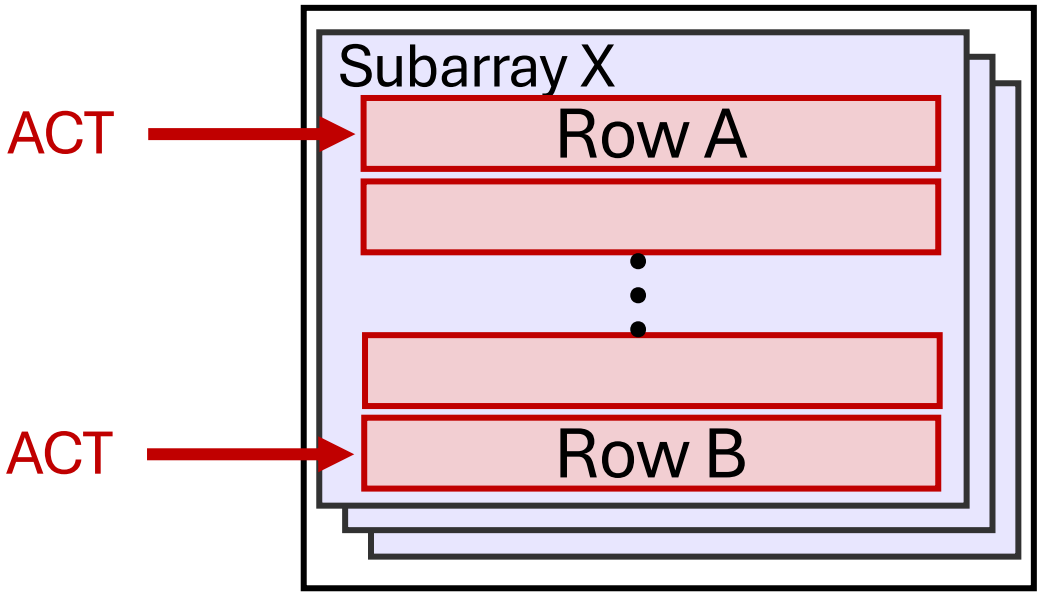


Success rate for this example: 66.67% (2/3)

Outline

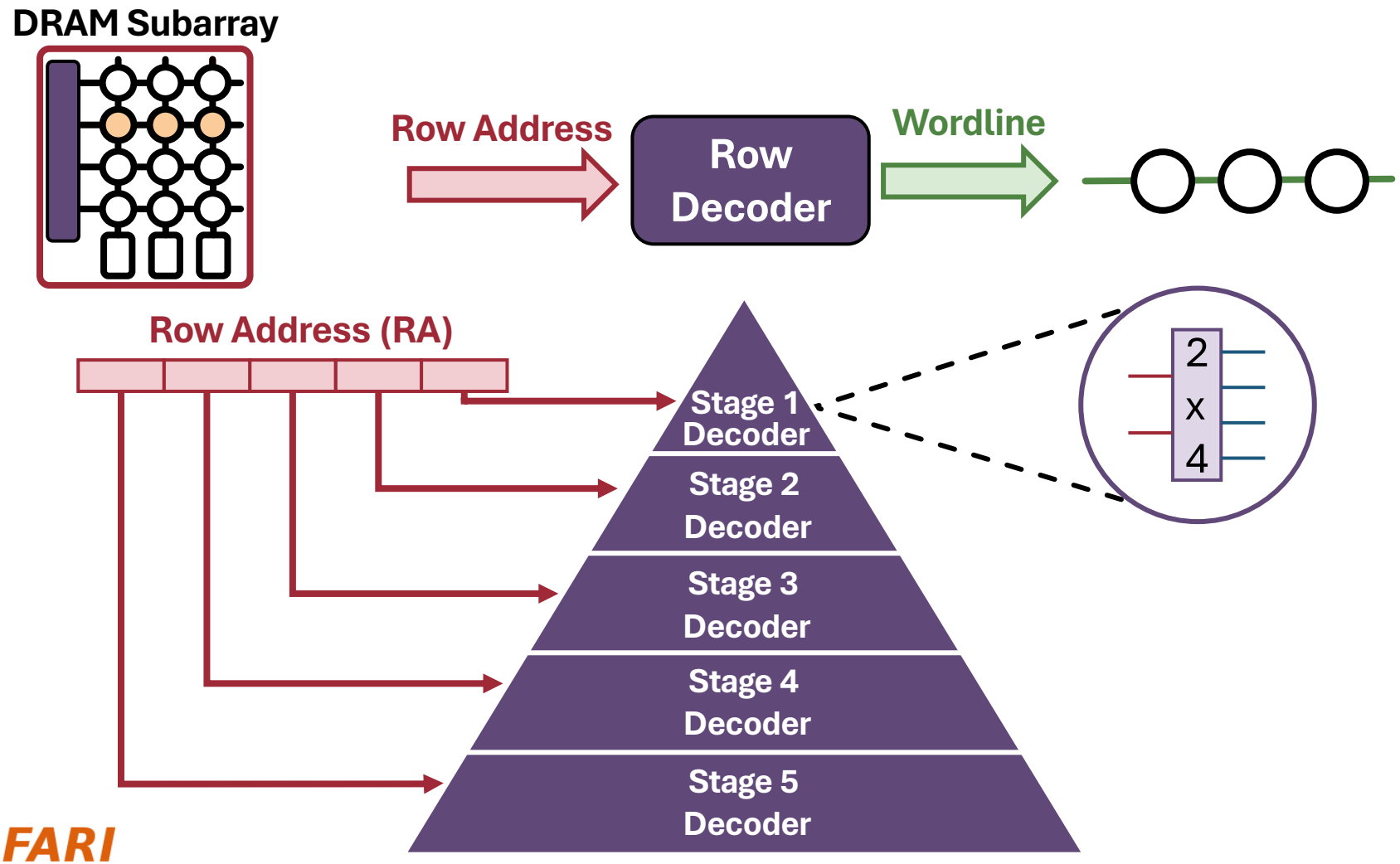
Key Observation

Activating two rows in **quick succession**
can **simultaneously** activate
many rows in a subarray



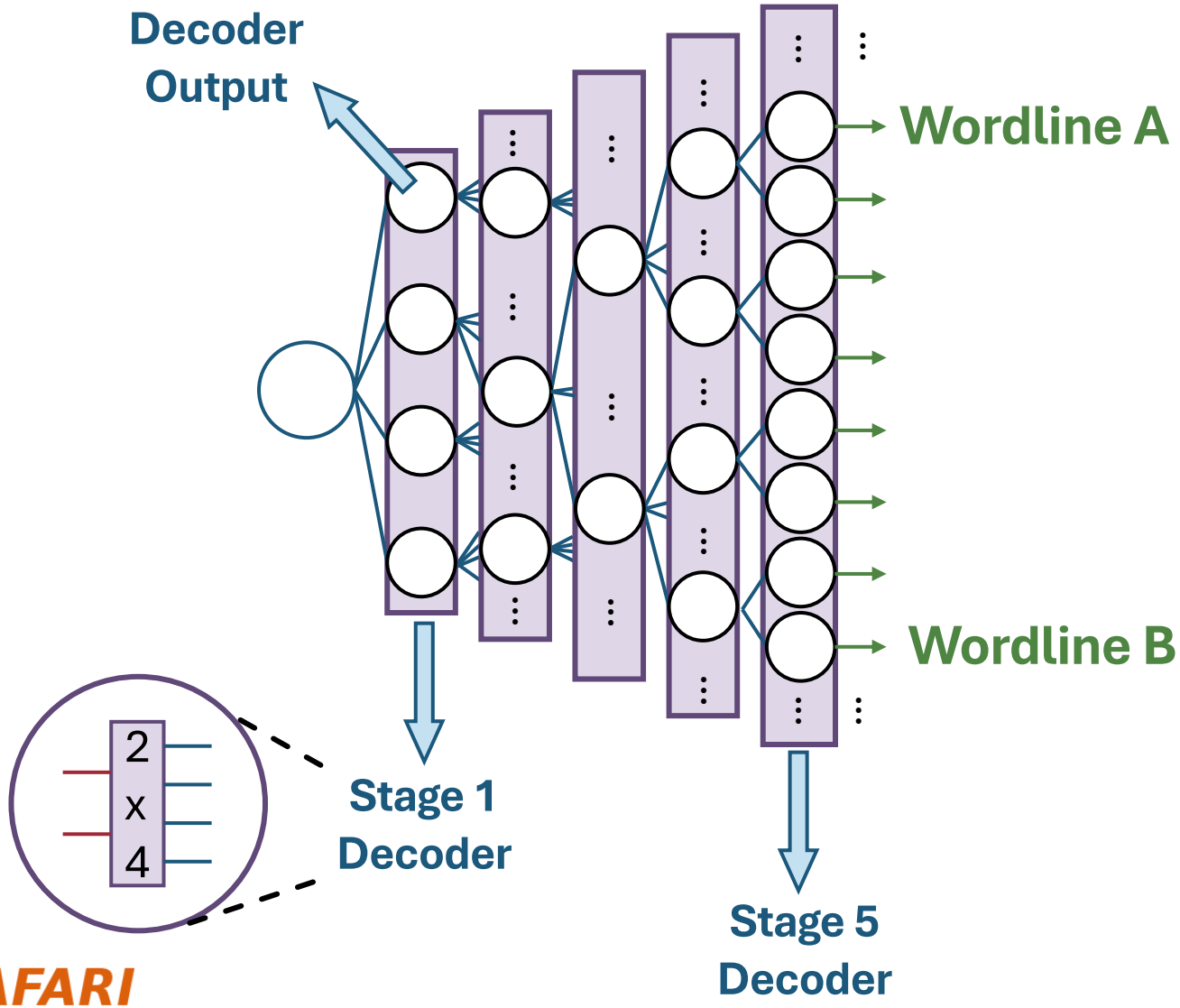
Hypothesis: Row Decoder Circuitry

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

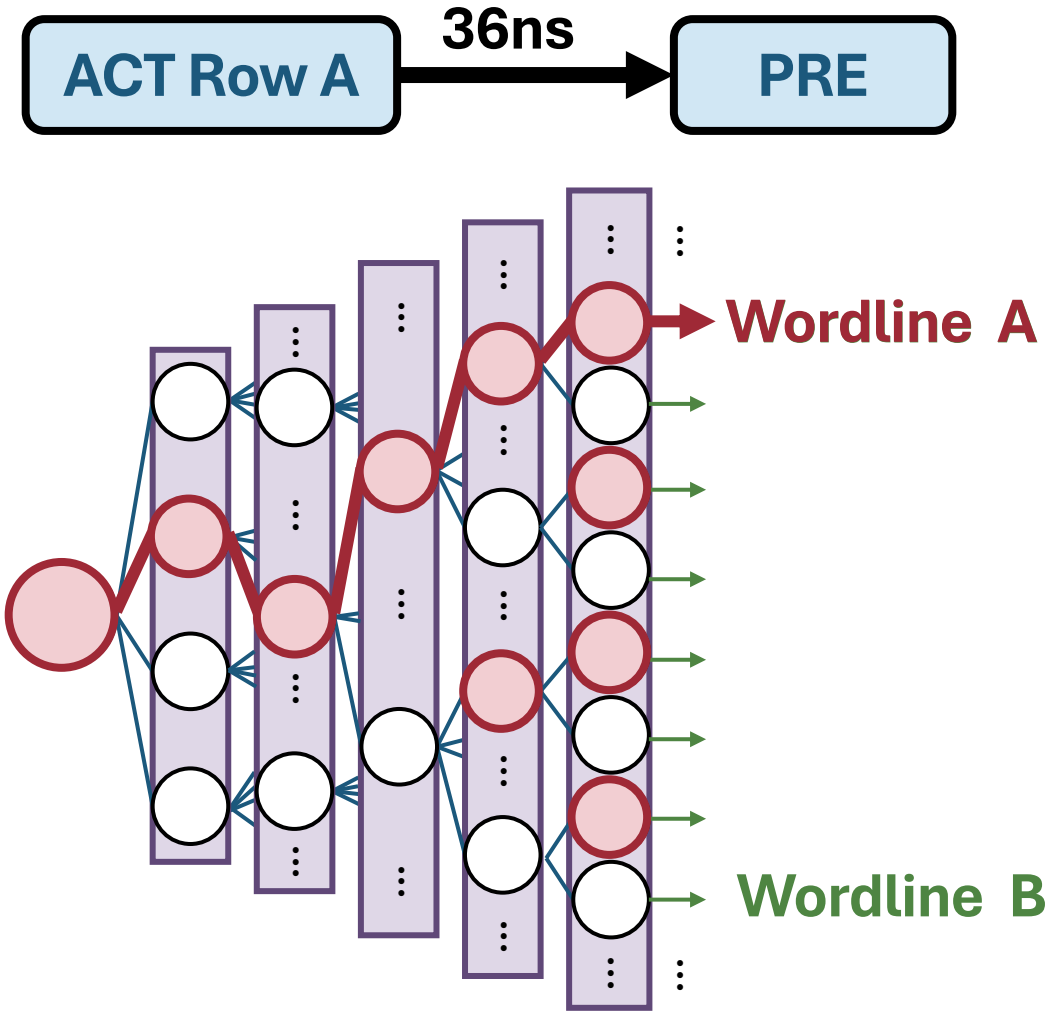


Row Decoder: A Tree Example

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



Activating a Single Row



1. Decode Row A address

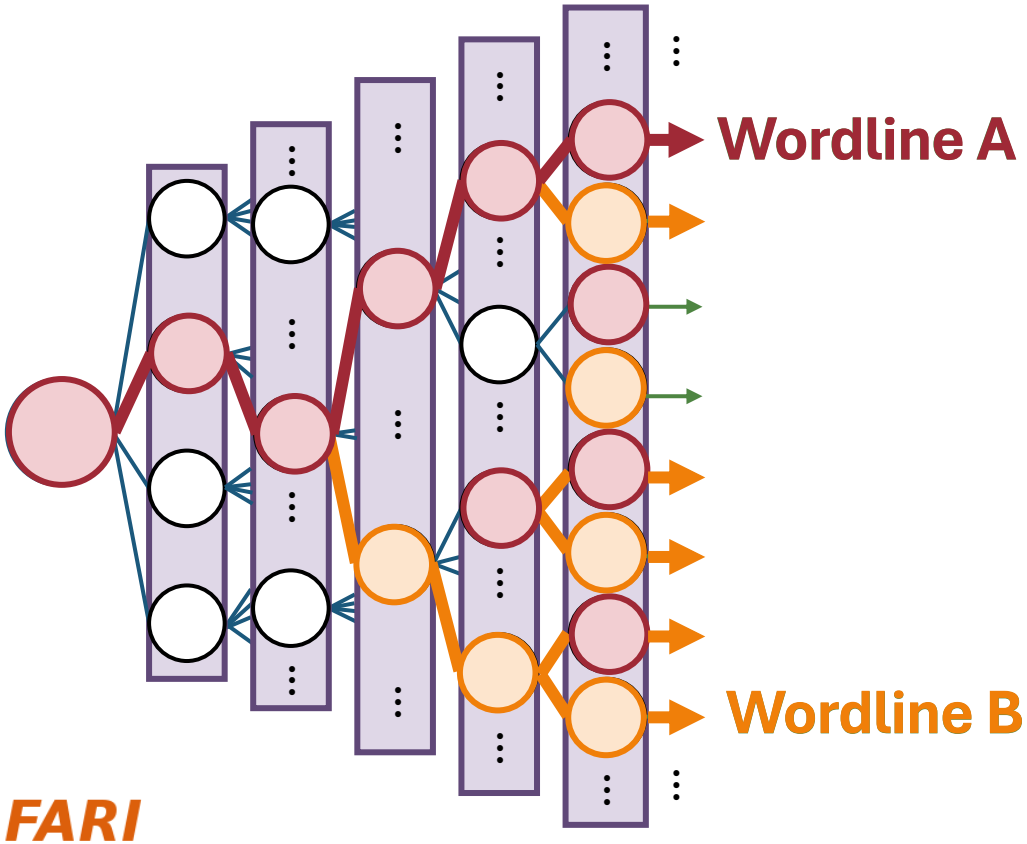
2. Assert corresponding signals

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

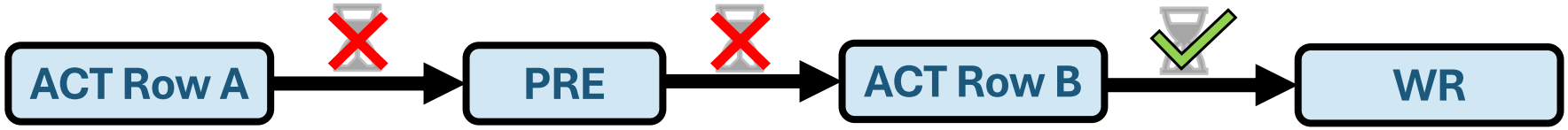
İsmail Emir Yüksel[1] Yahya Can Tuğrul[1,2] F. Nisa Bostancı[1] Geraldo F. Oliveira[1]
A. Giray Yağlıkçı[1] Ataberk Olgun[1] Melina Soysal[1] Haocong Luo[1]
Juan Gómez-Luna[1] Mohammad Sadrosadati[1] Onur Mutlu[1]
[1]*ETH Zürich* [2]*TOBB University of Economics and Technology*

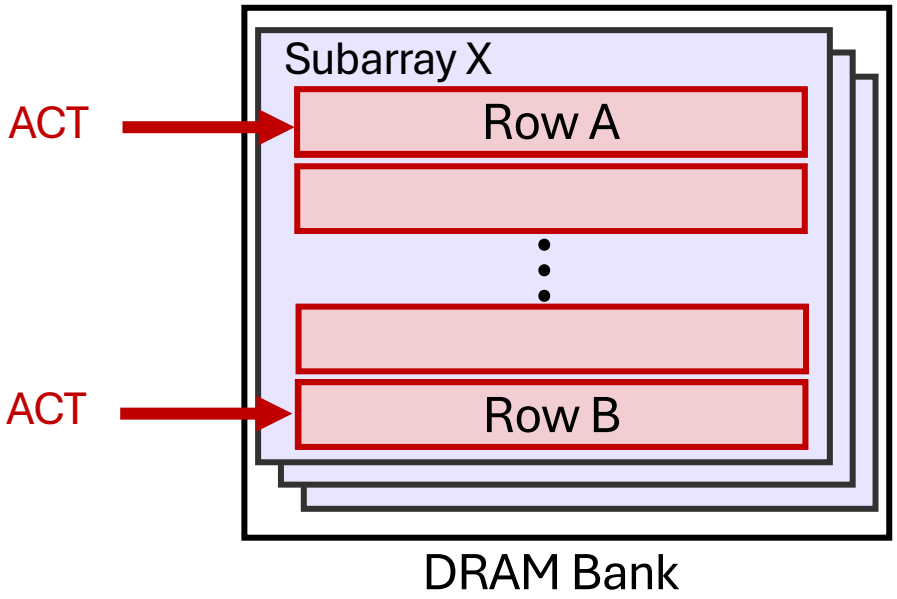
(More discussions & hypotheses in the paper)

https://arxiv.org/pdf/2405.06081

Characterization Methodology (I)

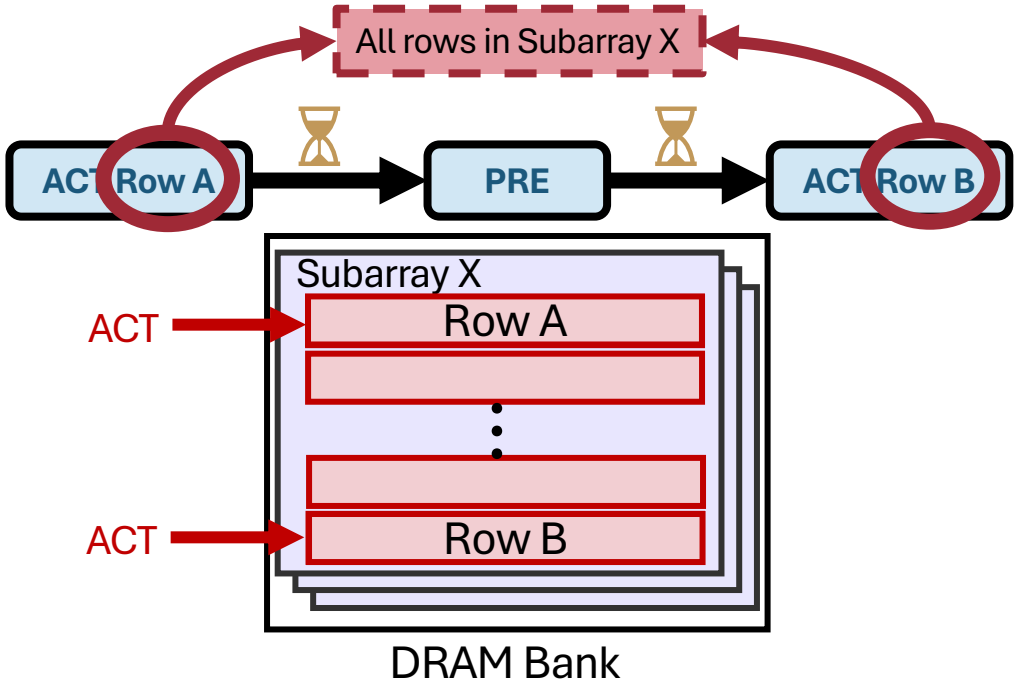
If rows are activated, WR command overwrites all of the activated rows' content

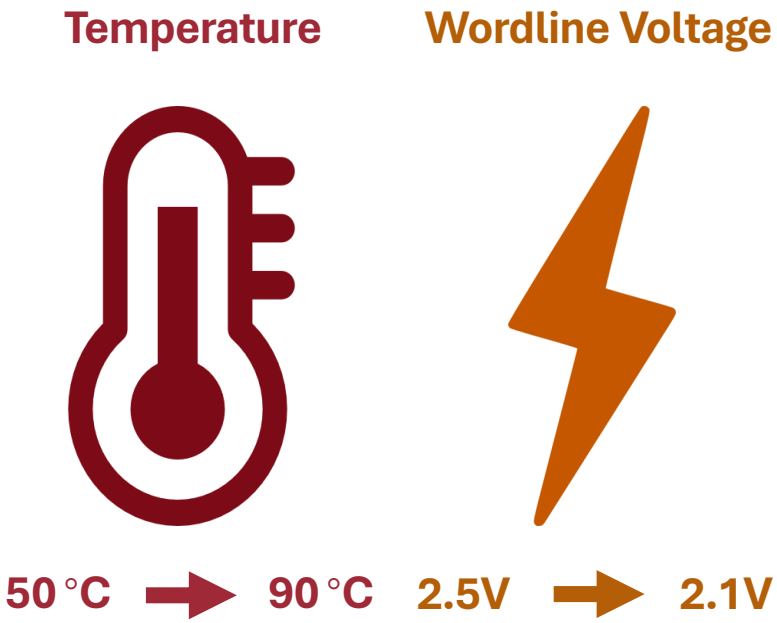




Characterization Methodology (II)

- Carefully sweep
  - Row addresses: Row A and Row B
  - Timing parameters: Between ACT → PRE and PRE → ACT
  - Temperature (°C): 50, 60, 70, 80, and 90
  - Wordline Voltage (V): 2.5, 2.4, 2.3, 2.2, and 2.1





Key Takeaways from Simultaneous Many-Row ACT

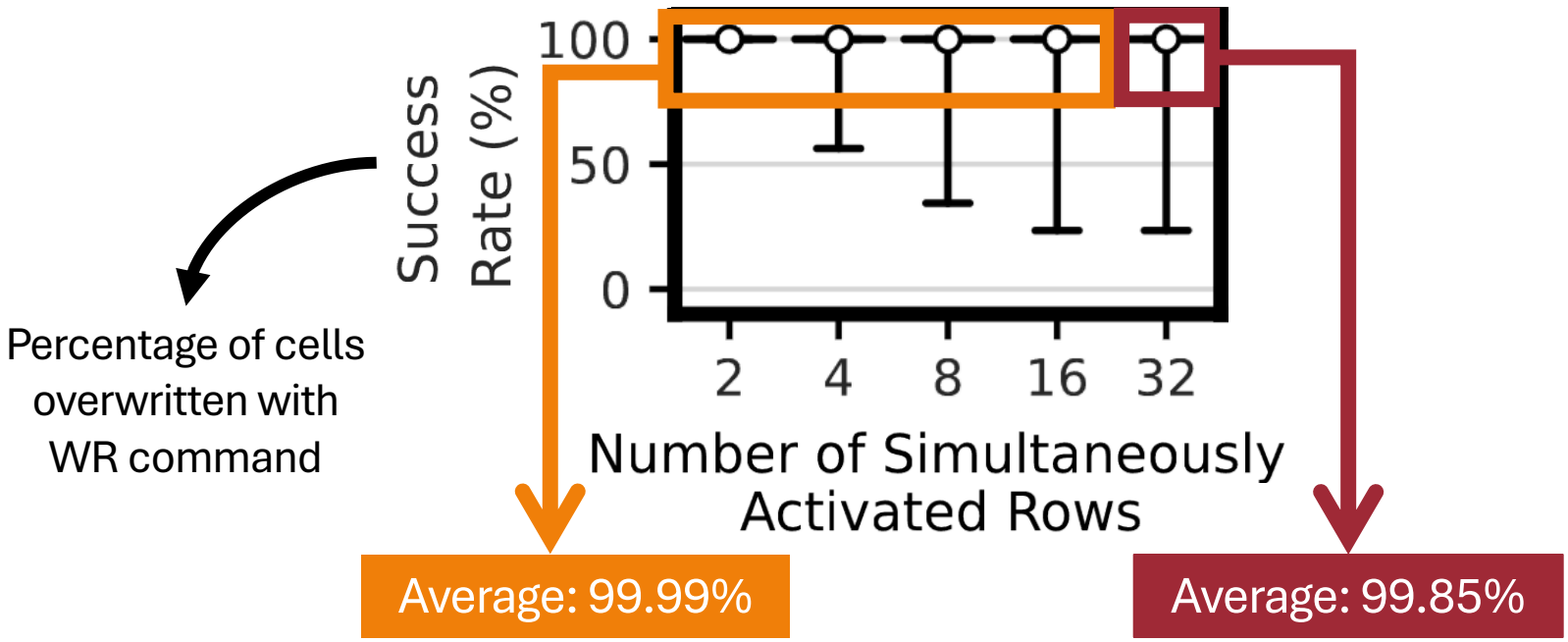
Key Takeaway 1

COTS DRAM chips are capable of simultaneously activating 2, 4, 8, 16, and 32 rows

Key Takeaway 2

Simultaneous many-row activation is highly resilient to temperature and wordline voltage changes

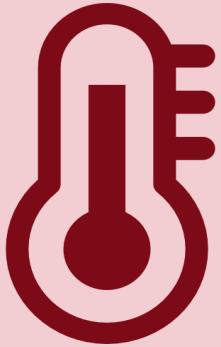
Robustness of Simultaneous Many-Row Activation



COTS DRAM chips can simultaneously activate 2, 4, 8, 16, and 32 rows in the same subarray

Also in the Paper: Impact of Temperature & Voltage

Temperature



50 °C → 90 °C

Increasing temperature up to 90°C has a small effect on the success rate of simultaneous many-row activation

Wordline Voltage

2.5V → 2.1V

Reducing the wordline voltage only slightly affects the success rate of simultaneous many-row activation

Leveraging Simultaneous Many-Row Activation

1 Perform **MAJX (where X>3)** operations

2 **Increase the robustness** of MAJX operations

3 Copy **one row's content** to **multiple rows**

Outline

Leveraging Simultaneous Many-Row Activation

1 Perform **MAJX (where X>3)** operations

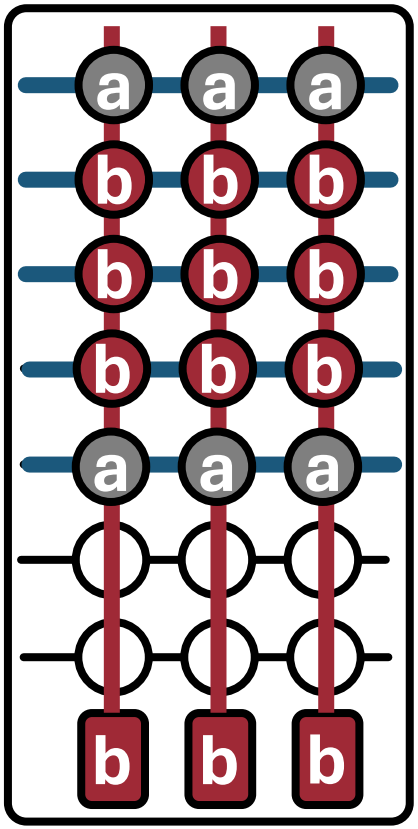
2 **Increase the robustness** of MAJX operations

3 Copy **one row's content** to **multiple rows**

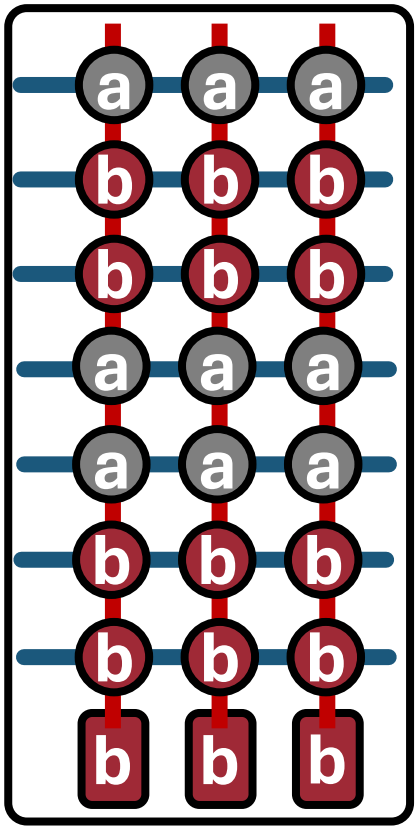
In-DRAM Majority-of-X (MAJX)

Simultaneously activate many rows to perform **MAJX (where X>3)** operations

MAJ5(a, b, b, b, a) = b

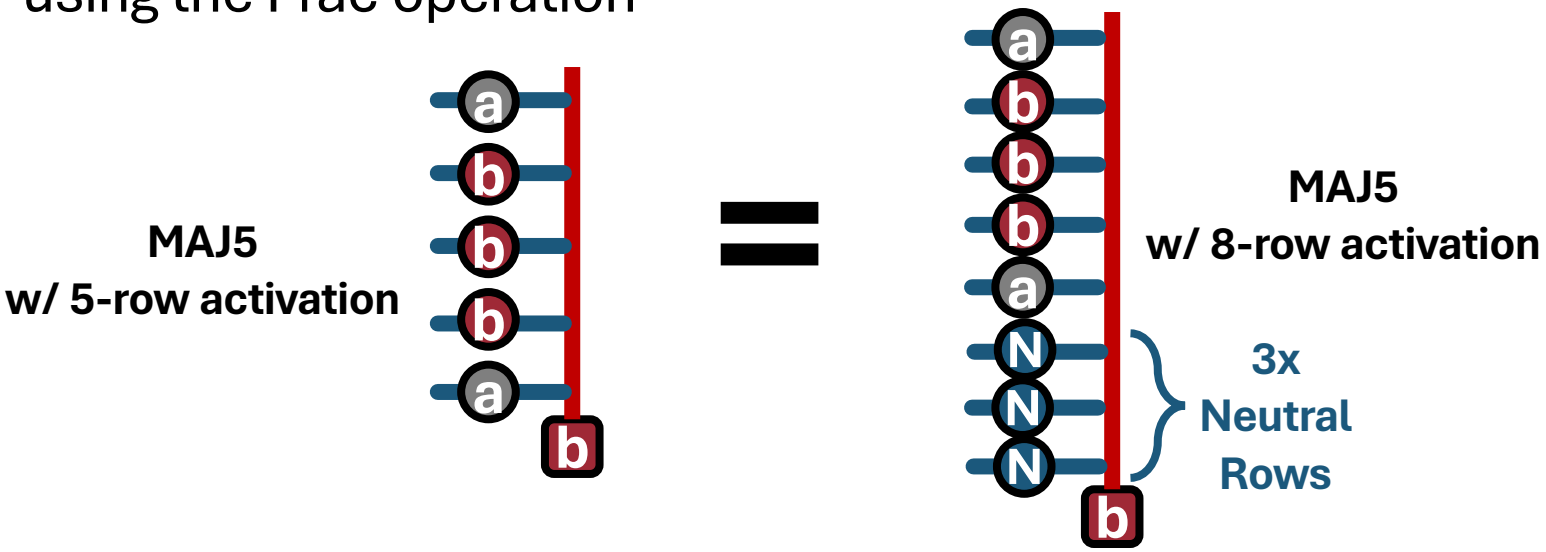


MAJ7(a, b, b, a, a, b, b) = b



MAJX in Real DRAM Chips

- For MAJX, we need to activate X rows simultaneously
- We can only simultaneously activate 2, 4, 8, 16, and 32 rows
- **Question**
  - How do we perform **MAJX** while simultaneously activating **more than X rows**?
- **Answer**
  - Making **some rows neutral** during the MAJX operation using the Frac operation*



*[Gao+MICRO'22]

Leveraging Simultaneous Many-Row Activation

1 Perform **MAJX (where X>3)** operations

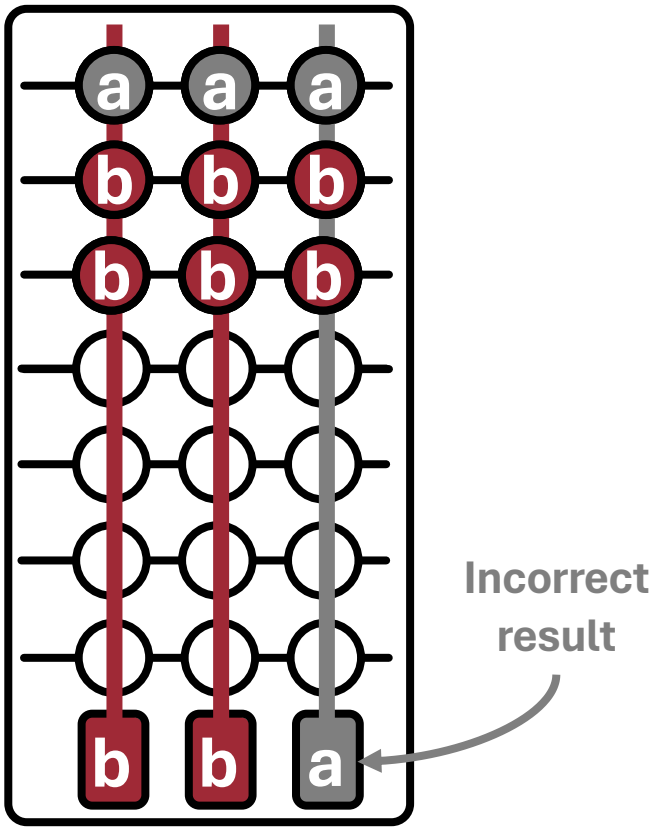
2 **Increase the robustness** of MAJX operations

3 Copy **one row's content** to **multiple rows**

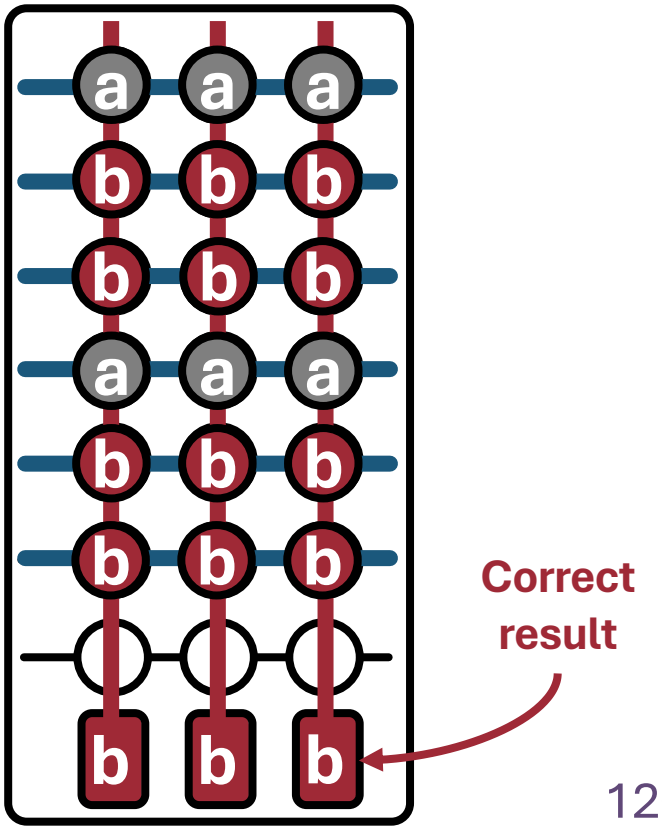
Improving the Robustness (Input Replication)

Storing **multiple copies** of MAJX input operands can **increase the robustness** of MAJX operations

MAJ3(a, b, b) = b

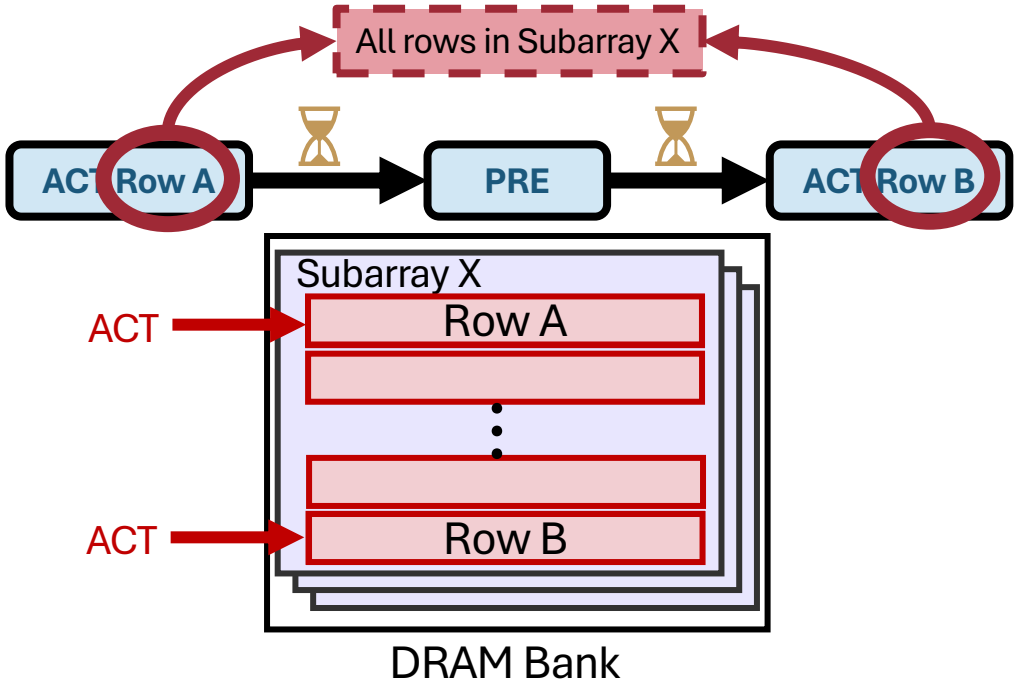


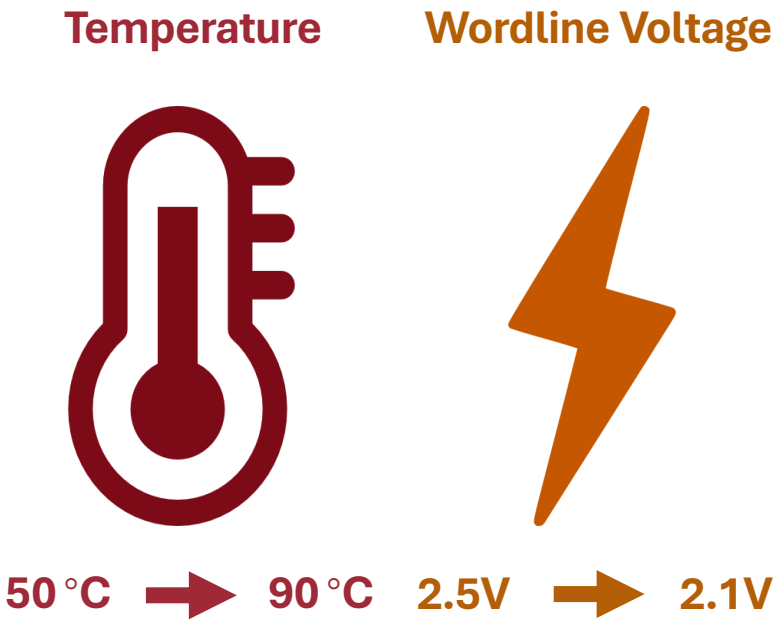
MAJ6(a, b, b, a, b, b) = b



Characterization Methodology

- Carefully sweep
  - Row addresses: Row A and Row B
  - Timing parameters: Between ACT → PRE and PRE → ACT
  - Temperature (°C): 50, 60, 70, 80, and 90
  - Wordline Voltage (V): 2.5, 2.4, 2.3, 2.2, and 2.1





Key Takeaways from MAJX Operation

Key Takeaway 1

COTS DRAM chips are capable of performing MAJ5, MAJ7, and MAJ9 operations

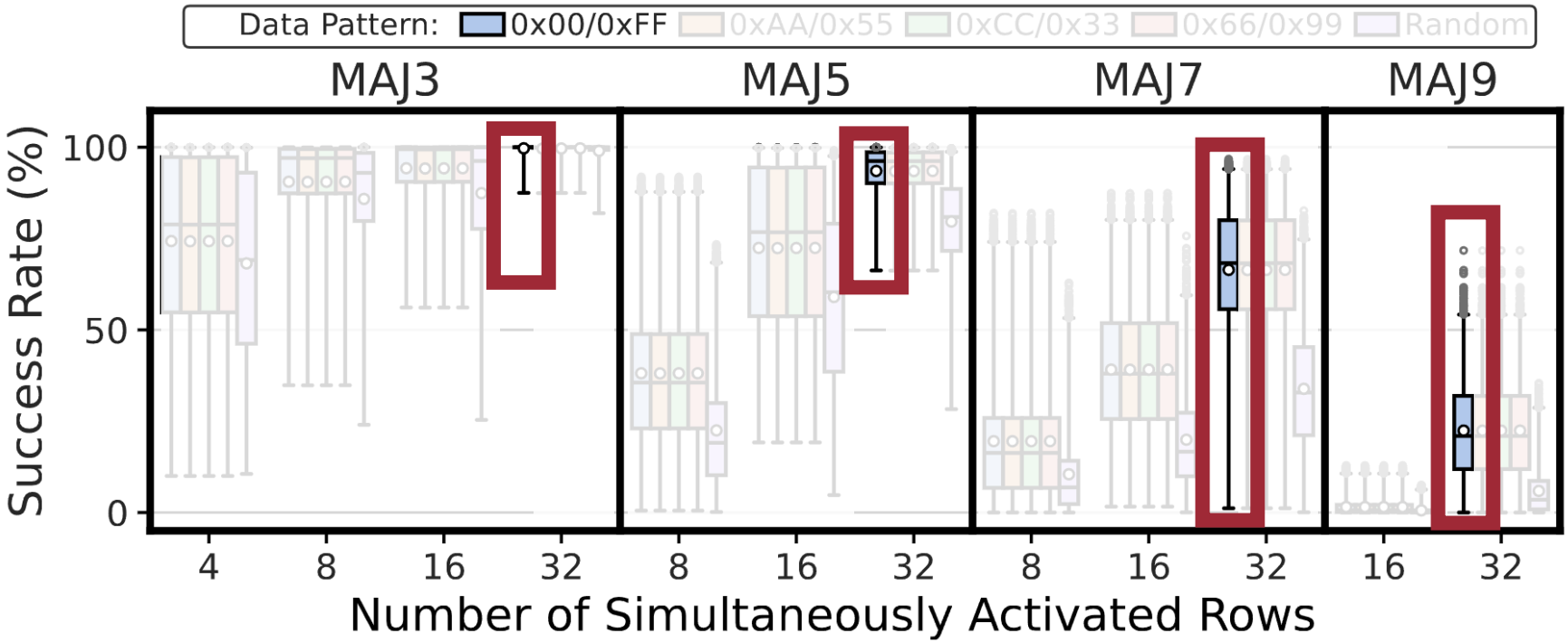
Key Takeaway 2

Storing multiple copies of MAJX's input operands significantly increases the MAJX's success rate

Key Takeaway 3

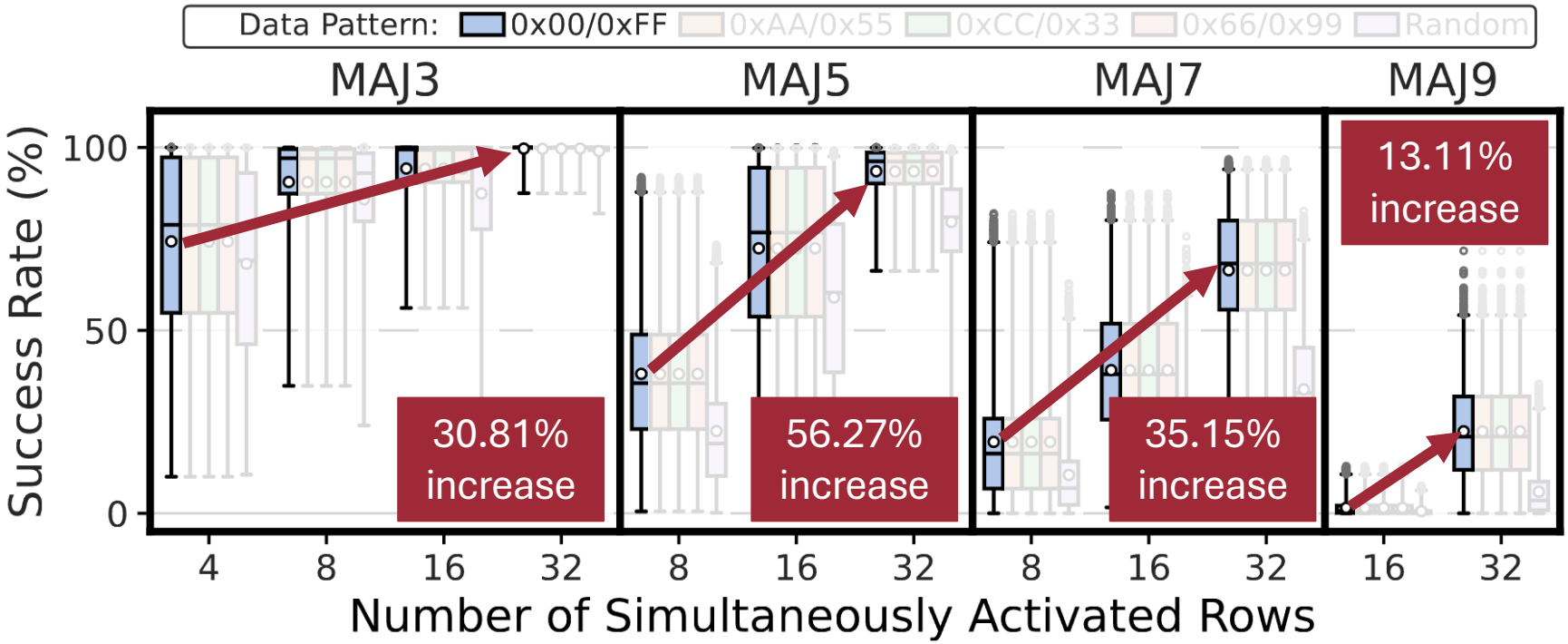
Voltage and temperature slightly affect the success rate, whereas data pattern affects significantly

Robustness of MAJX Operations



COTS DRAM chips are capable of performing MAJ5, MAJ7, and MAJ9 operations

Impact of Input Replication

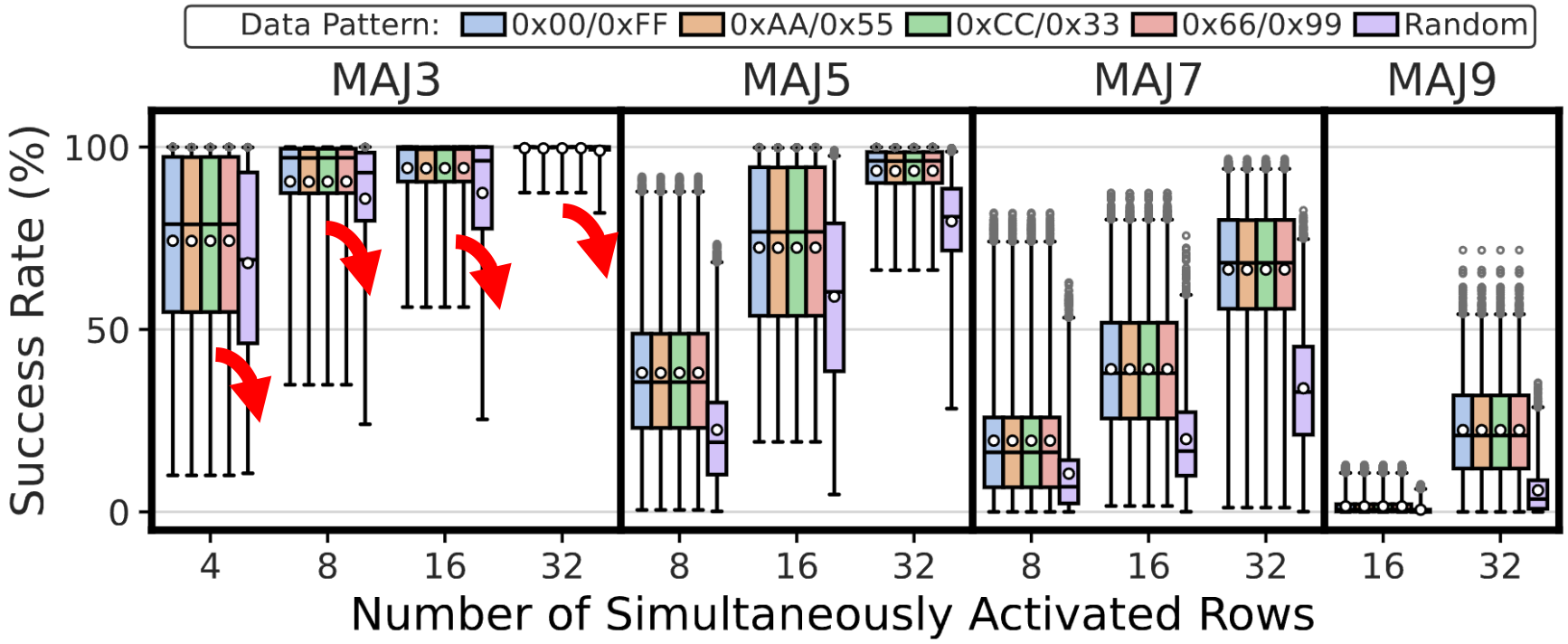


Similar trend in success rate increase in all tested MAJX operations

Final average success rate MAJ3 (MAJ5): 99.68% (93.49%)

Storing multiple copies of MAJ's input operands increases the success rate of MAJ3, MAJ5, MAJ7, and MAJ9 operations

Impact of Data Pattern

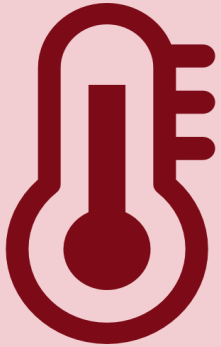


11.52% decrease in success rate on average (up to 32.56%) across all tested MAJX operations

Data pattern significantly affects the success rate of the MAJX operation

Also in the Paper: Impact of Temperature & Voltage

Temperature



50 °C → 90 °C

Temperature slightly affects the success rate of the MAJX operation

Wordline Voltage

2.5V → 2.1V

Wordline voltage has a small effect on the success rate of the MAJX operation

Outline

Leveraging Simultaneous Many-Row Activation

1 Perform **MAJX (where X>3)** operations

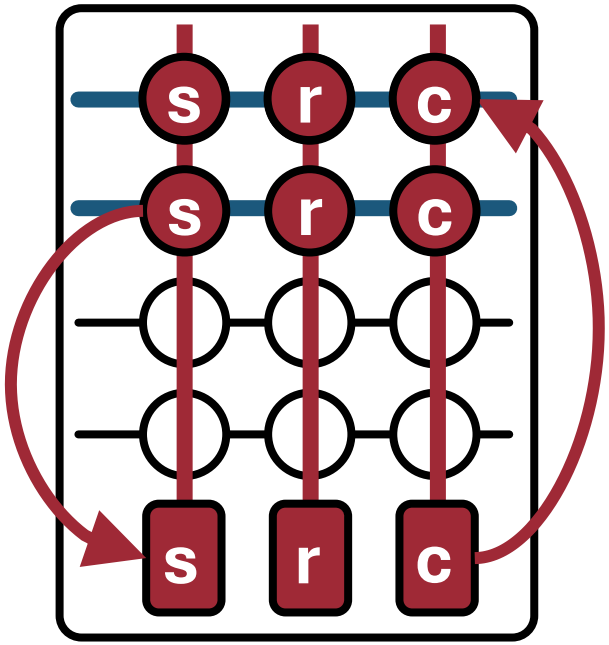
2 **Increase the robustness** of MAJX operations

3 Copy **one row's content** to **multiple rows**

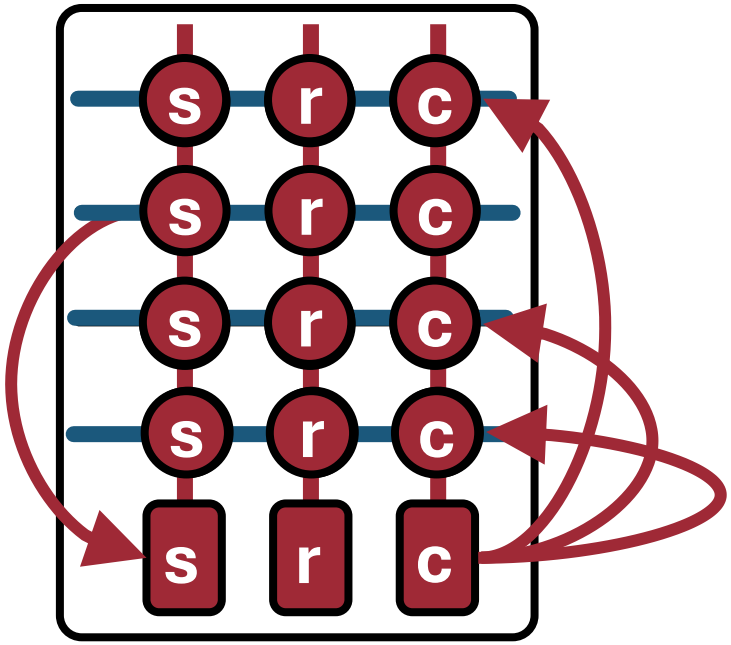
In-DRAM Multiple Row Copy (Multi-RowCopy)

Simultaneously activate many rows to
copy **one row's content** to **multiple destination rows**

RowClone



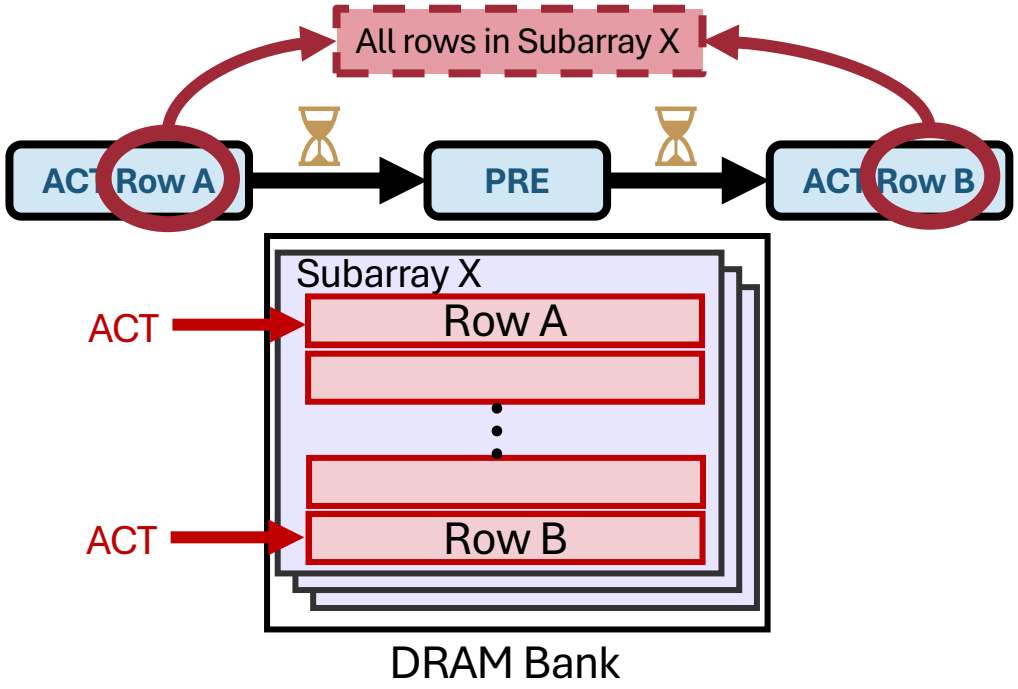
Multi-RowCopy

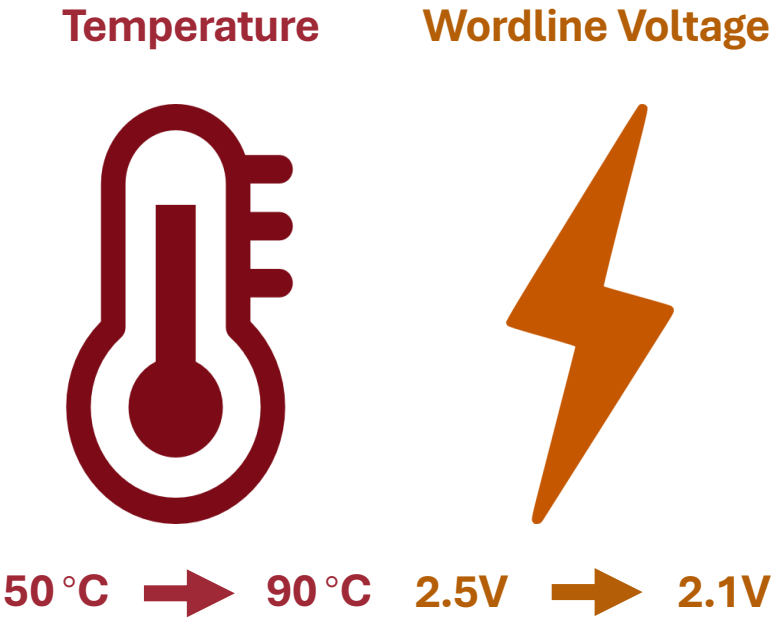


[Seshadri+ MICRO'13]

Characterization Methodology (II)

- Carefully sweep
  - Row addresses: Row A and Row B
  - Timing parameters: Between ACT → PRE and PRE → ACT
  - Temperature (°C): 50, 60, 70, 80, and 90
  - Wordline Voltage (V): 2.5, 2.4, 2.3, 2.2, and 2.1





Key Takeaways from Multi-RowCopy

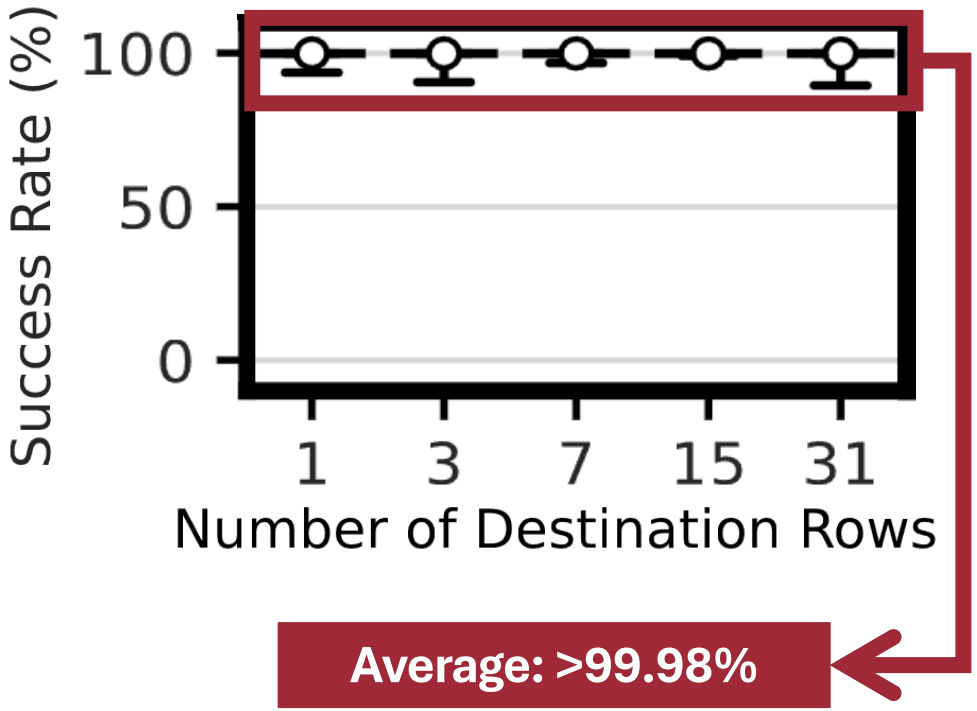
Key Takeaway 1

COTS DRAM chips are capable of copying one row's data to 1, 3, 7, 15, and 31 other rows at very high success rates

Key Takeaway 2

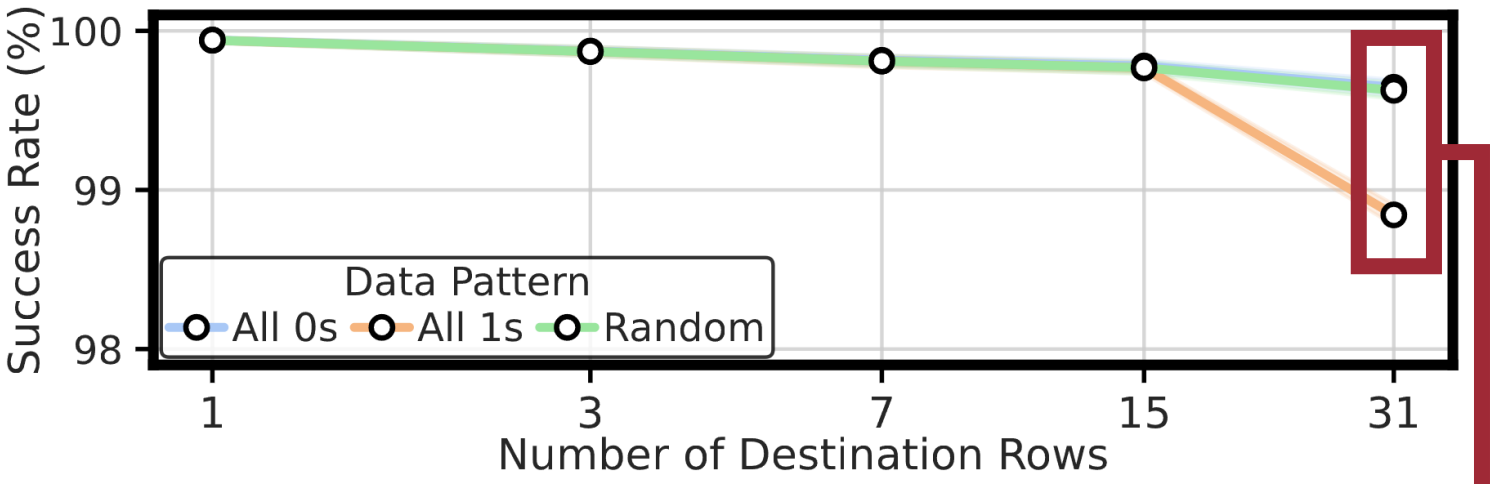
Multi-RowCopy in COTS DRAM chips is highly resilient to changes in data pattern, temperature, and wordline voltage

Robustness of Multi-RowCopy



COTS DRAM chips can copy one row's content to up to 31 rows with a very high success rate

Impact of Data Pattern

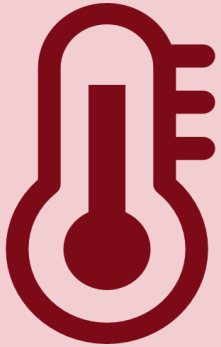


At most 0.79% decrease in average success rate

Data pattern has a small effect on the success rate of the Multi-RowCopy operation

Also in the Paper: Impact of Temperature & Voltage

Temperature



50 °C → 90 °C

Increasing temperature up to 90°C
has a very small effect on
the success rate of the Multi-RowCopy operation

Wordline Voltage

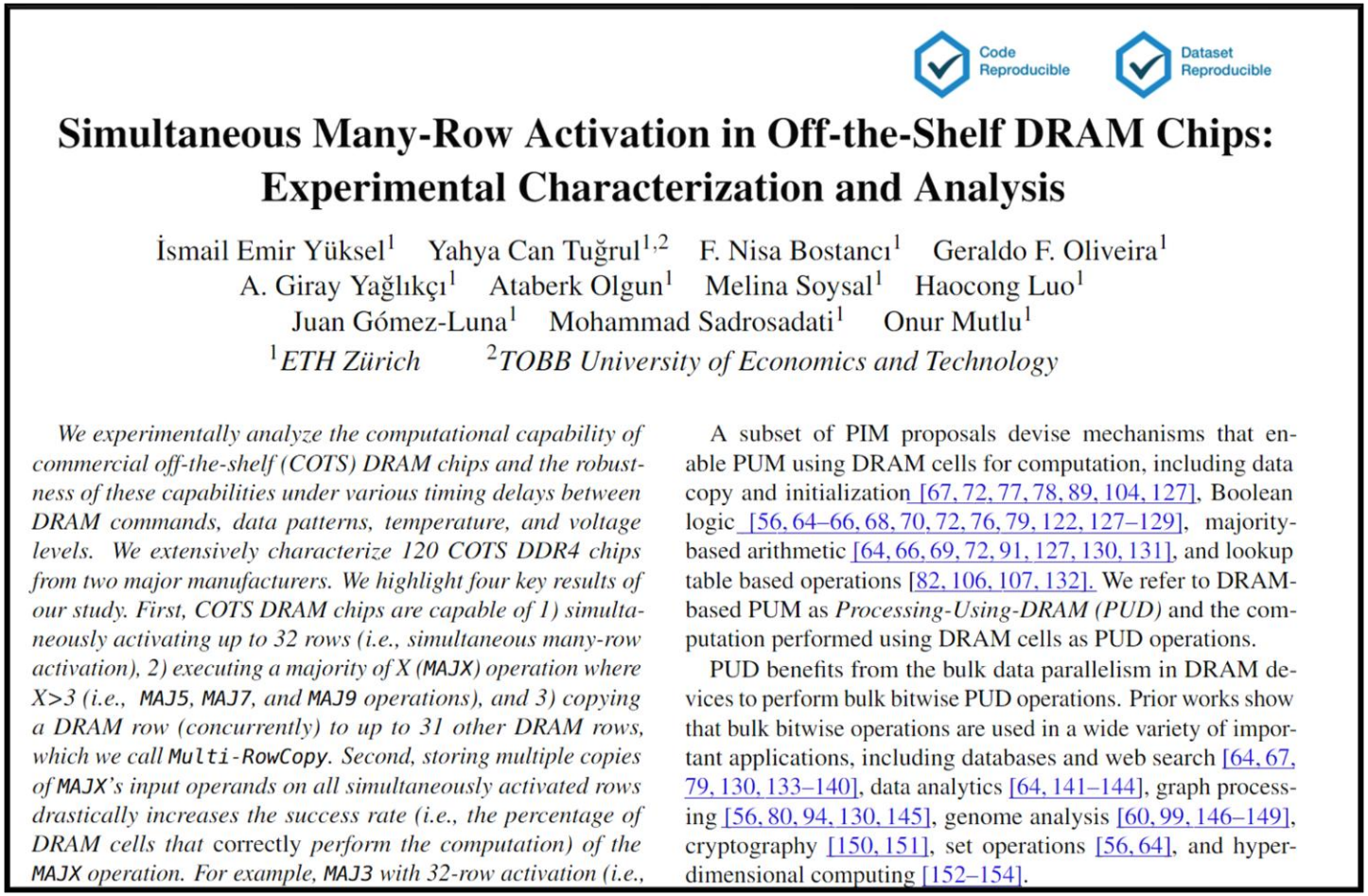
2.5V → 2.1V

Reducing the wordline voltage
only slightly affects
the success rate of the Multi-RowCopy operation

More in the Paper

- Detailed hypotheses and key ideas on
  - Hypothetical row decoder circuitry
  - Input Replication
- More characterization results
  - Power consumption of simultaneous many-row activation
  - Effect of timing delays between ACT-PRE and PRE-ACT commands
  - Effect of temperature and wordline voltage
- Circuit-level (SPICE) experiments for input replication
- Potential performance benefits of enabling new in-DRAM operations
  - Majority-based computation
  - Content destruction-based cold-boot attack prevention
- Discussions on the limitations of tested COTS DRAM chips

Available on arXiv



Code Reproducible

Dataset Reproducible

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*

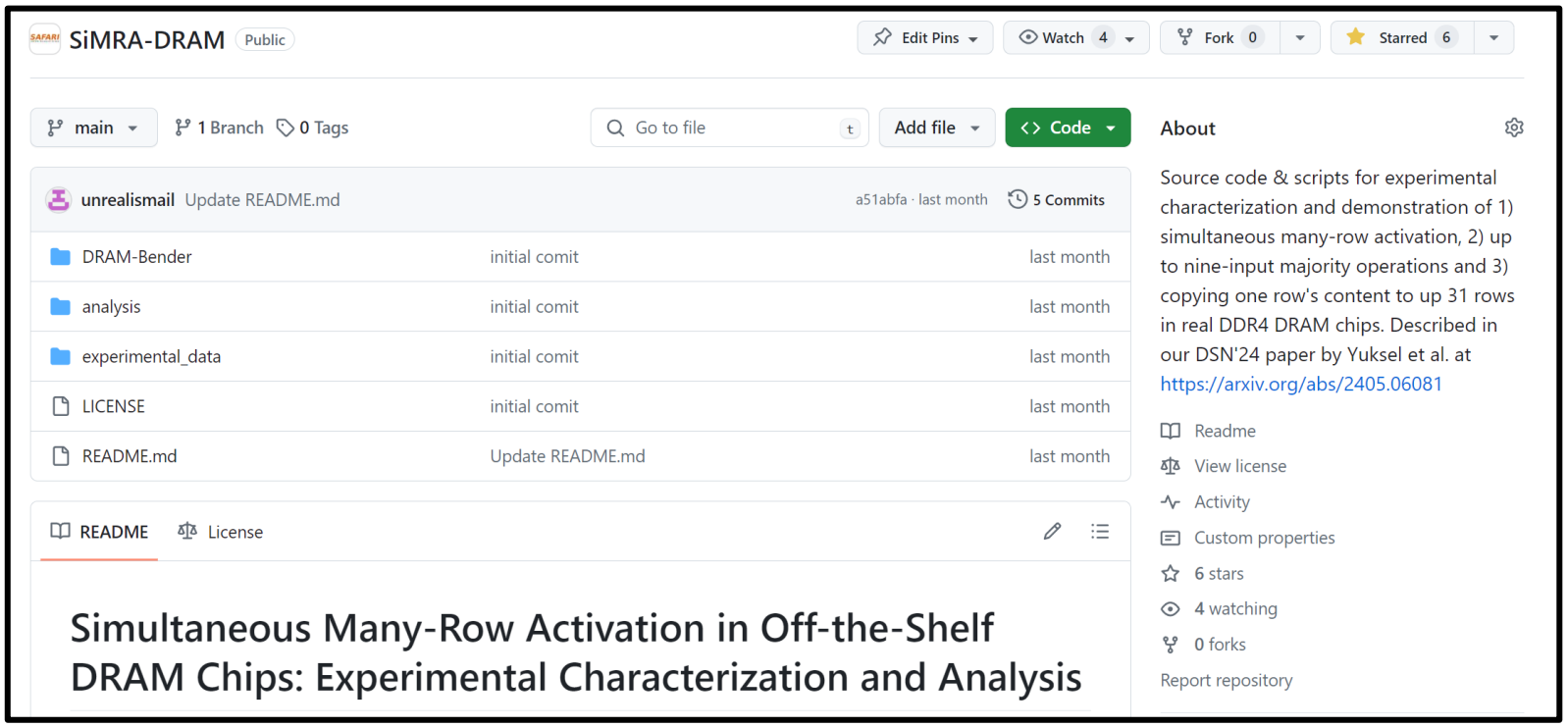
We experimentally analyze the computational capability of commercial off-the-shelf (COTS) DRAM chips and the robustness of these capabilities under various timing delays between DRAM commands, data patterns, temperature, and voltage levels. We extensively characterize 120 COTS DDR4 chips from two major manufacturers. We highlight four key results of our study. First, COTS DRAM chips are capable of 1) simultaneously activating up to 32 rows (i.e., simultaneous many-row activation), 2) executing a majority of X (`MAJX`) operation where X>3 (i.e., `MAJ5`, `MAJ7`, and `MAJ9` operations), and 3) copying a DRAM row (concurrently) to up to 31 other DRAM rows, which we call `Multi-RowCopy`. Second, storing multiple copies of `MAJX`'s input operands on all simultaneously activated rows drastically increases the success rate (i.e., the percentage of DRAM cells that correctly *perform the computation) of the `MAJX` operation. For example, `MAJ3` with 32-row activation (i.e.,*

A subset of PIM proposals devise mechanisms that enable PUM using DRAM cells for computation, including data copy and initialization [67, 72, 77, 78, 89, 104, 127], Boolean logic [56, 64–66, 68, 70, 72, 76, 79, 122, 127–129], majority-based arithmetic [64, 66, 69, 72, 91, 127, 130, 131], and lookup table based operations [82, 106, 107, 132]. We refer to DRAM-based PUM as *Processing-Using-DRAM (PUD)* and the computation performed using DRAM cells as PUD operations.

PUD benefits from the bulk data parallelism in DRAM devices to perform bulk bitwise PUD operations. Prior works show that bulk bitwise operations are used in a wide variety of important applications, including databases and web search [64, 67, 79, 130, 133–140], data analytics [64, 141–144], graph processing [56, 80, 94, 130, 145], genome analysis [60, 99, 146–149], cryptography [150, 151], set operations [56, 64], and hyperdimensional computing [152–154].

https://arxiv.org/pdf/2405.06081

Our Work is Open Source and Artifact Evaluated



https://github.com/CMU-SAFARI/SiMRA-DRAM

Outline

Conclusion

We experimentally demonstrate that COTS DRAM chips can

- **simultaneously activate** up to 32 DRAM rows
- perform **MAJ3, MAJ5, MAJ7, and MAJ9** operations
- copy one row's content to **up to 31 rows**

We characterize 120 DDR4 chips and highlight three key results

- **Storing multiple copies** of MAJX's input operands (i.e., input replication) drastically increases the success rate of MAJX operations
- **Voltage and temperature slightly** affect the success rate of **MAJX operation**, whereas **data pattern** affects **significantly**
- **Multi-RowCopy is highly resilient** to changes in data pattern, temperature, and wordline voltage

We believe these empirical results demonstrate the promising potential of using DRAM as a computation substrate

Simultaneous Many-Row Activation in Off-the-Shelf DRAM Chips

Experimental Characterization and Analysis

Paper

GitHub

İsmail Emir Yüksel

Yahya C. Tuğrul F. Nisa Bostancı Geraldo F. Oliveira

A. Giray Yağlıkçı Ataberk Olgun Melina Soysal Haocong Luo

Juan Gómez–Luna Mohammad Sadr Onur Mutlu

SAFARI

ETH zürich

Simultaneous Many-Row Activation in Off-the-Shelf DRAM Chips

Experimental Characterization and Analysis

Backup Slides

İsmail Emir Yüksel

Yahya C. Tuğrul F. Nisa Bostancı Geraldo F. Oliveira

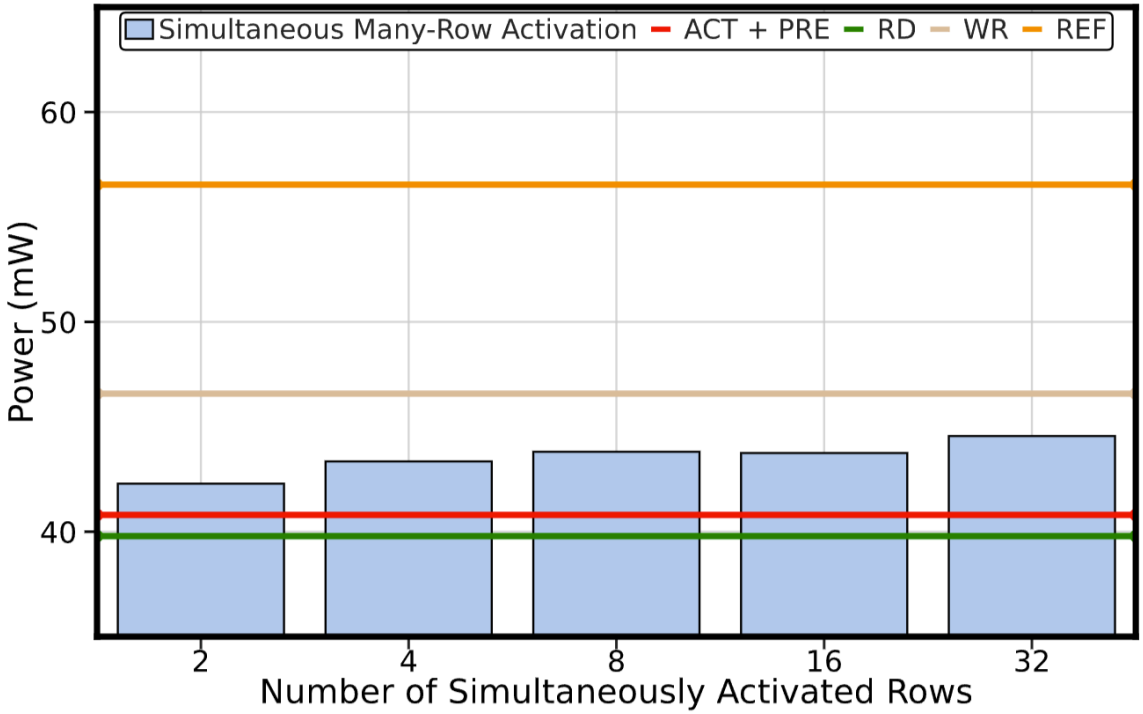
A. Giray Yağlıkçı Ataberk Olgun Melina Soysal Haocong Luo

Juan Gómez–Luna Mohammad Sadr Onur Mutlu

SAFARI

ETH zürich

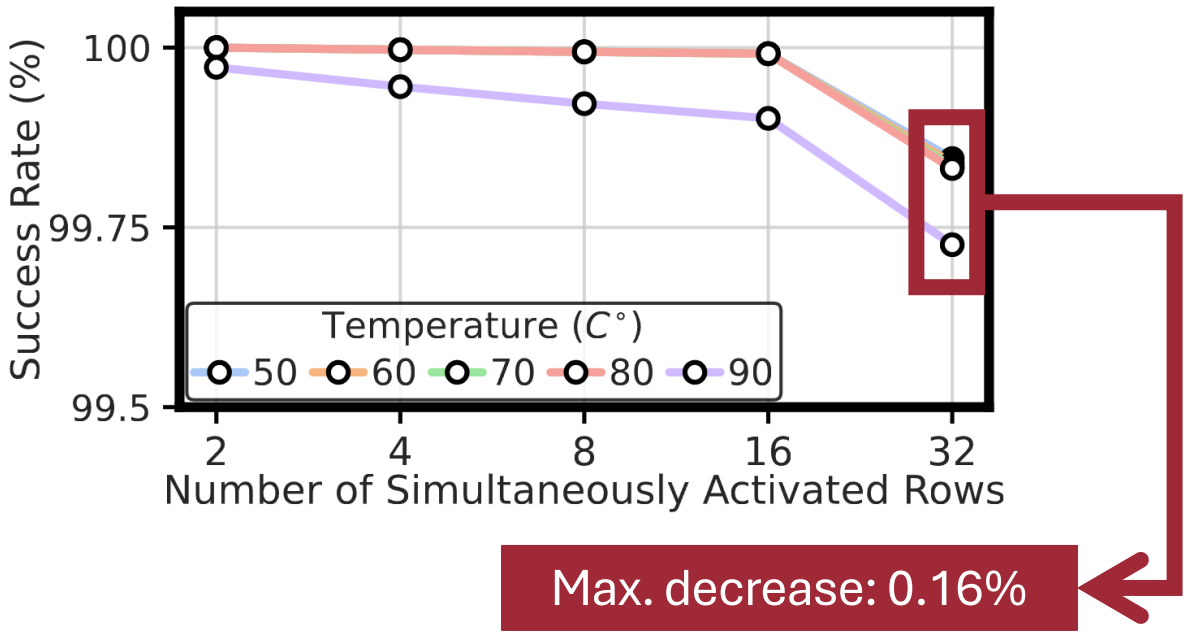
Power Consumption of Many-Row ACT



32-row activation consumes 21.19% less power than the most power-consuming single DRAM operation (i.e., REF)

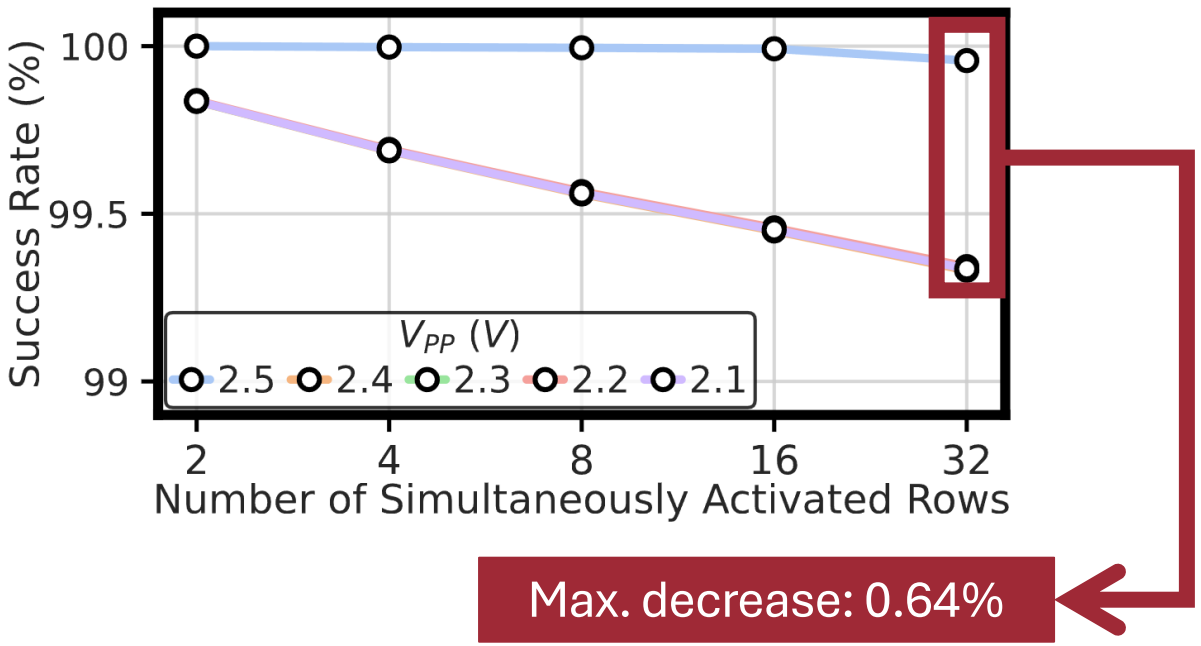
Simultaneous many-row activation power draw likely meets the power budget of DDR4 chips

Impact of Temperature in Many-Row ACT



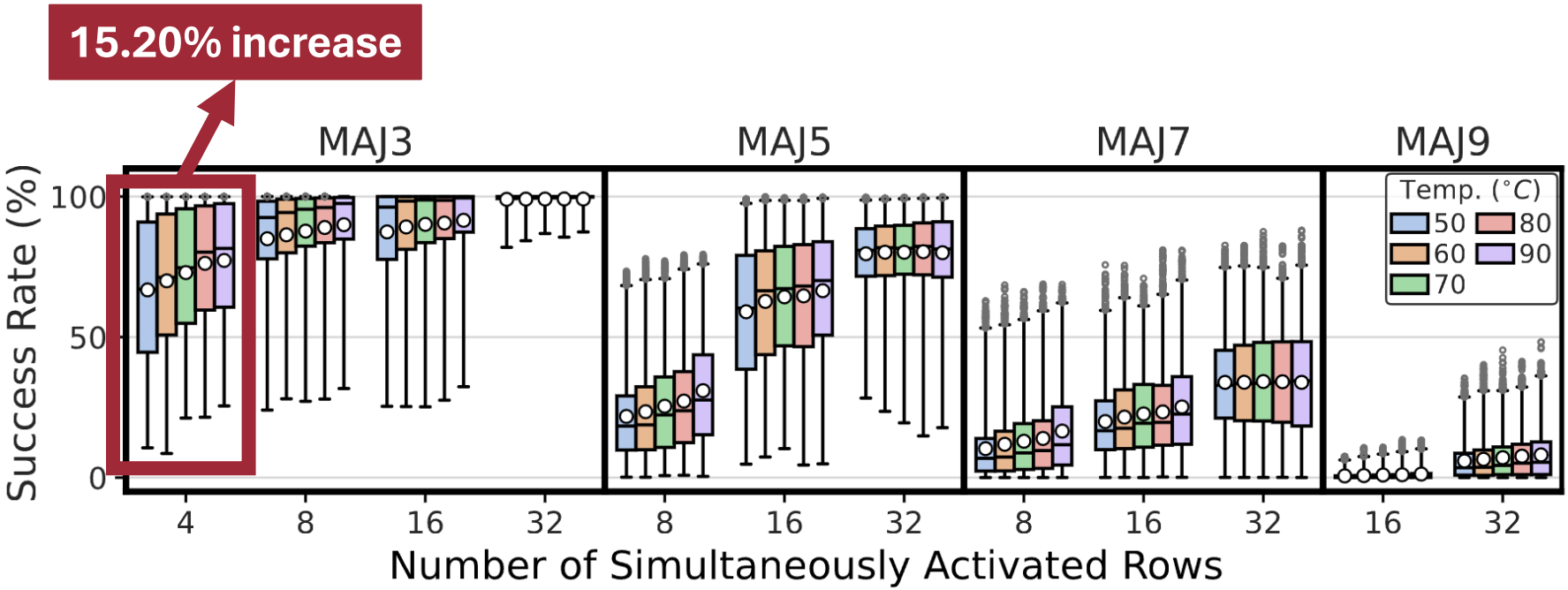
Increasing temperature up to 90°C has a small effect on the success rate

Impact of Voltage in Many-Row ACT



Reducing the wordline voltage only slightly affects the success rate

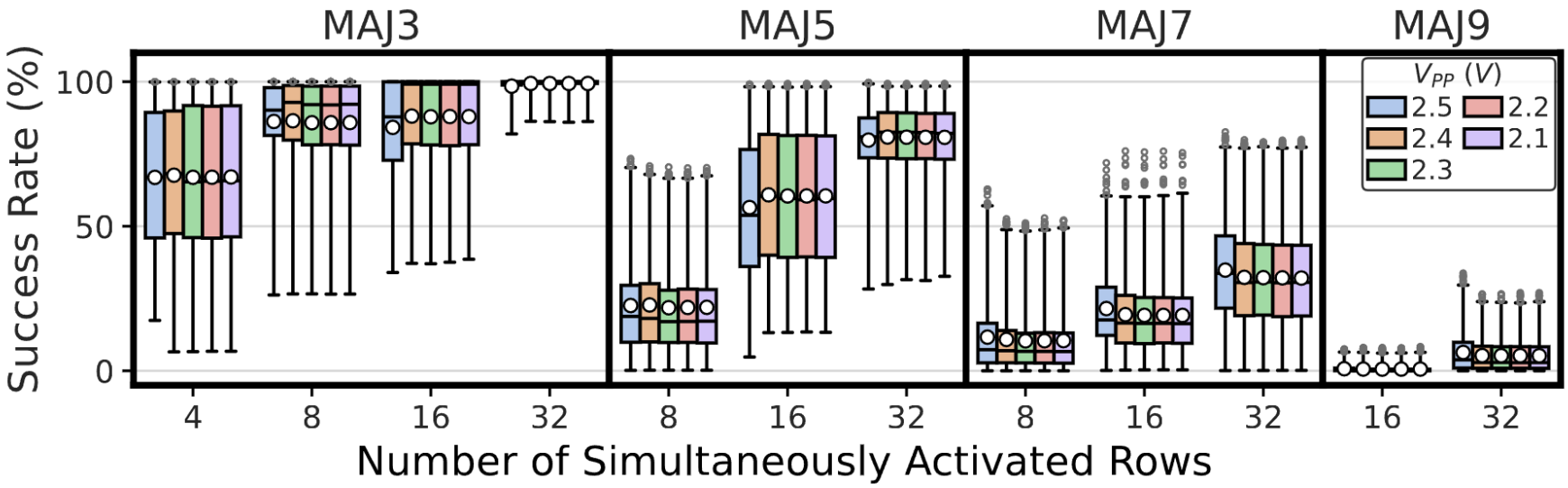
Impact of Temperature in MAJX



from 50 °C to 90 °C , the success rate varies by 4.25% on average across all the tested operations

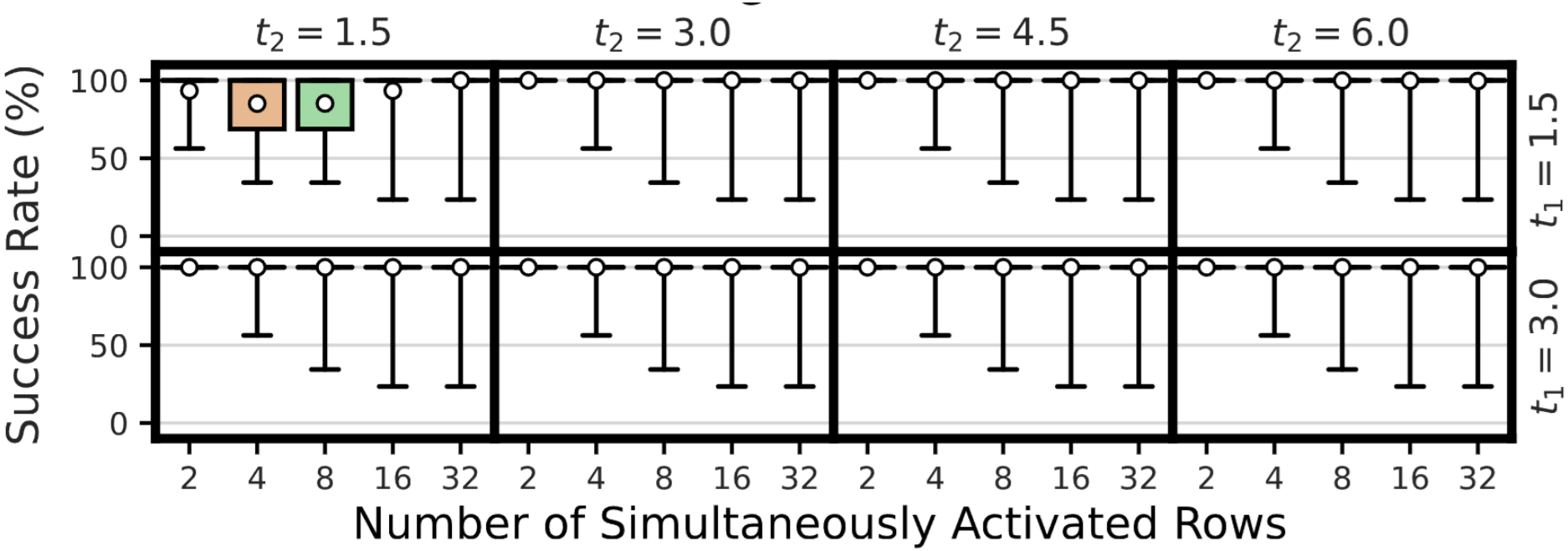
Temperature slightly affects the success rate of the MAJX operation

Impact of Voltage in MAJX Operations

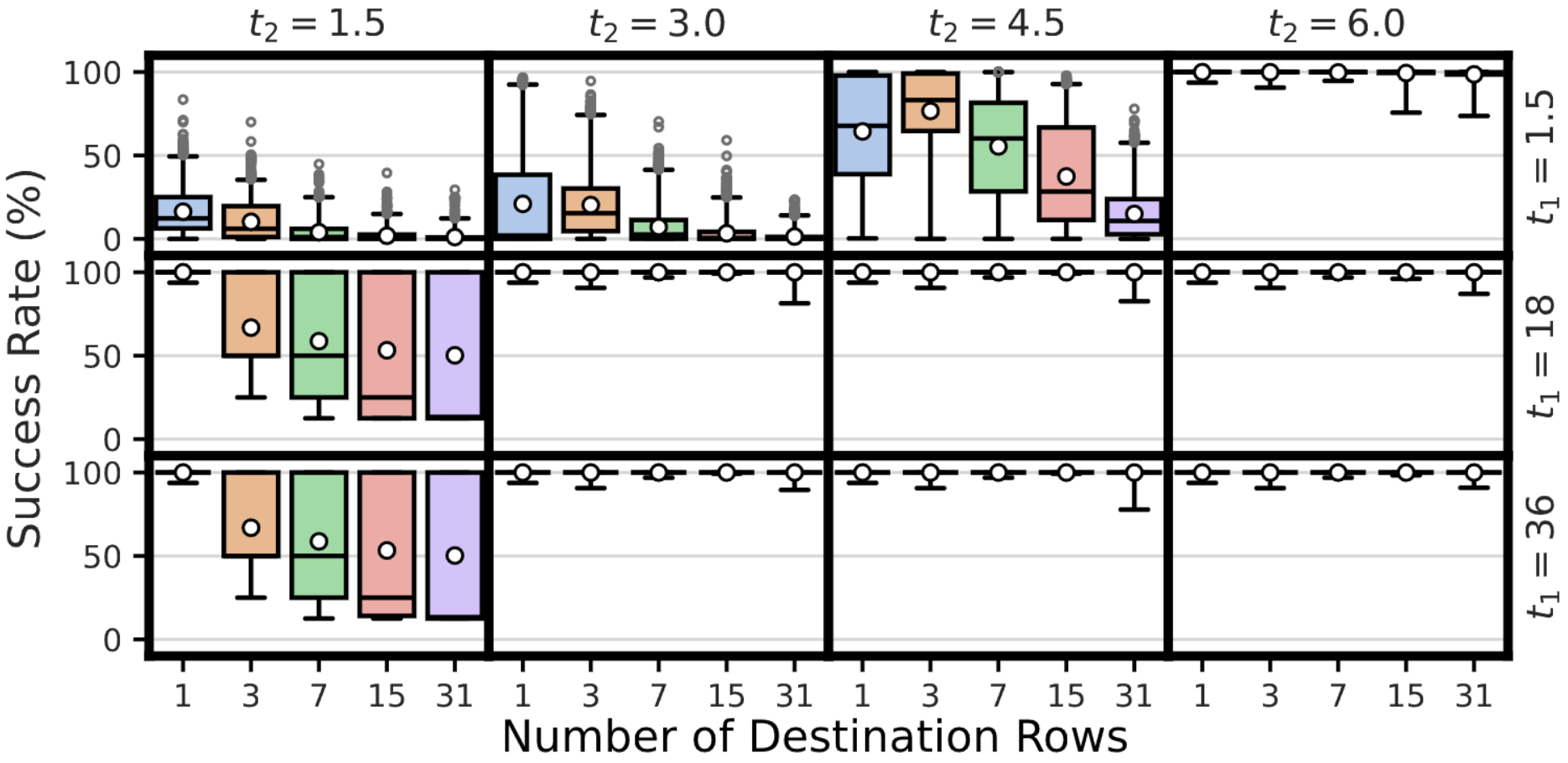


Wordline voltage slightly affects the success rate of the MAJX operation

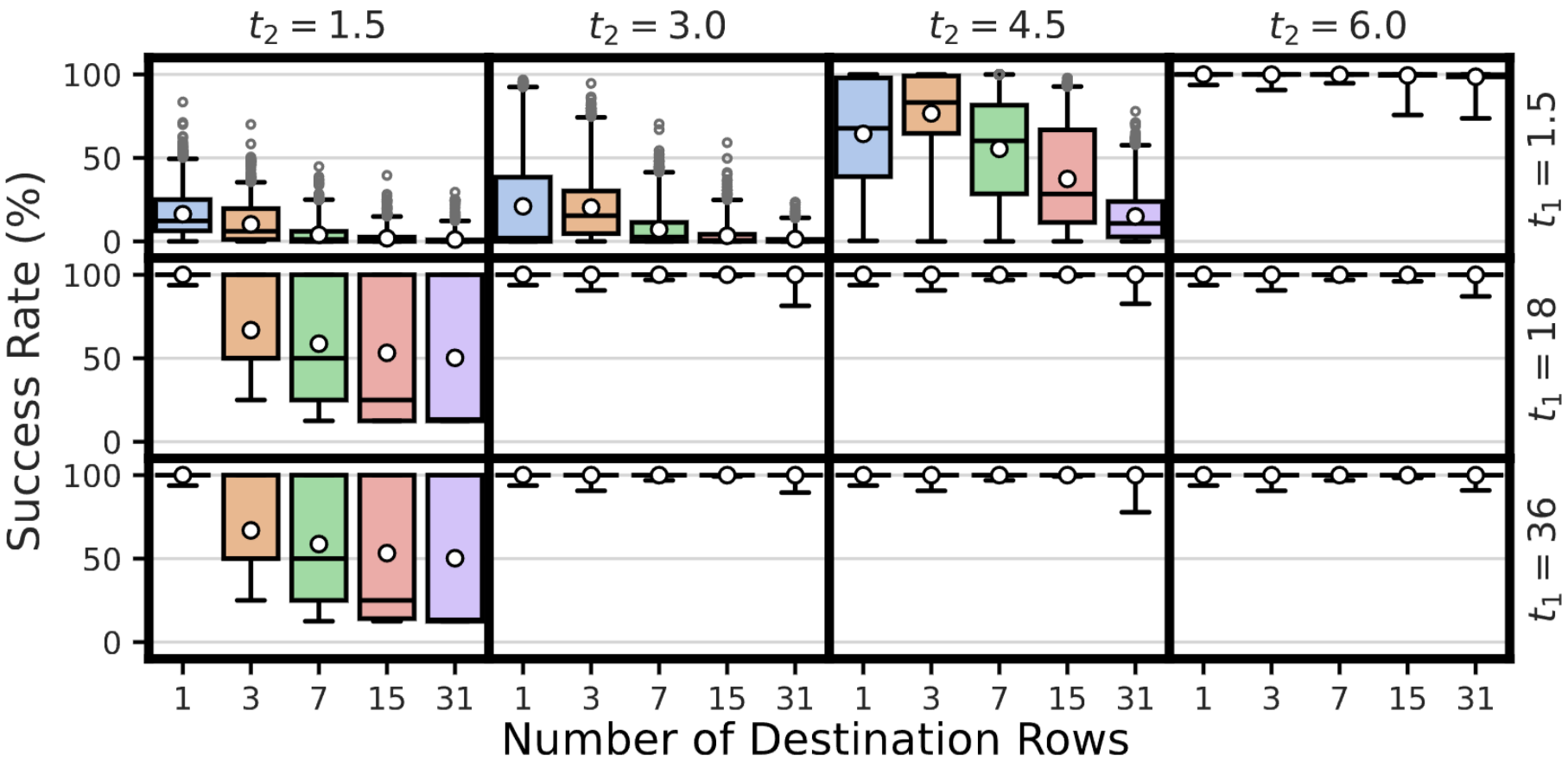
Impact of Timing Delays in Many-Row ACT



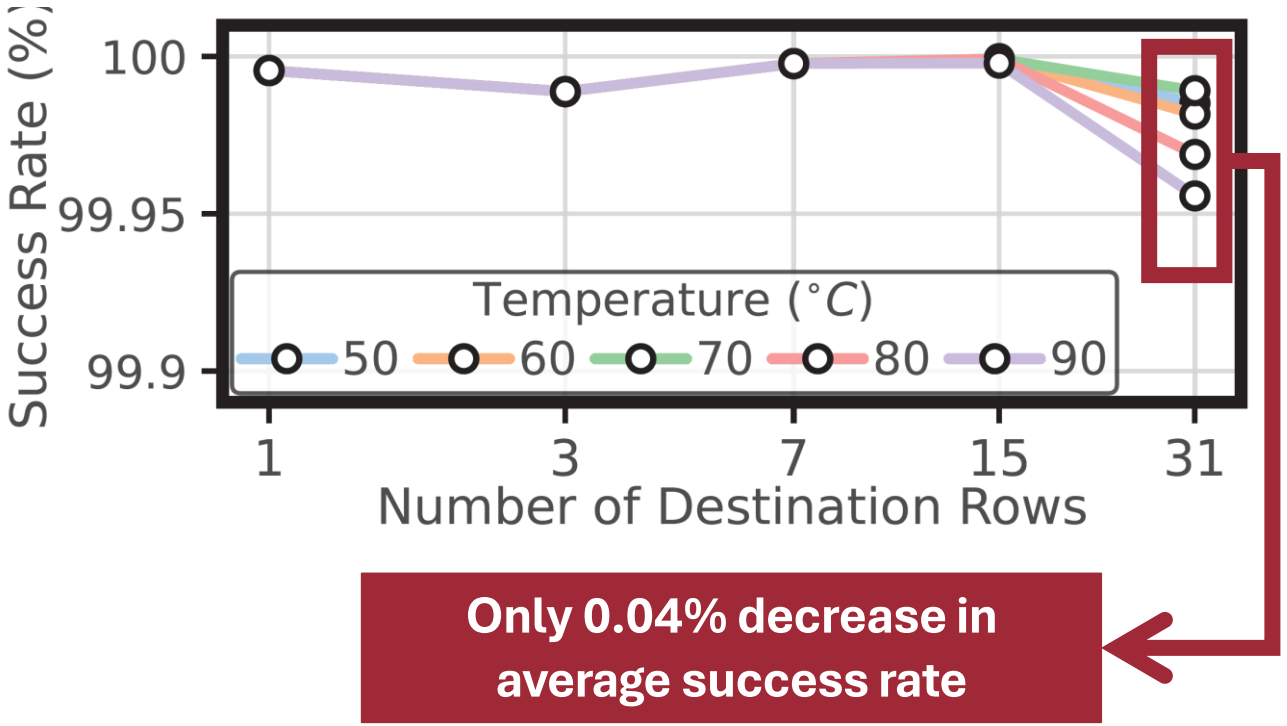
Impact of Timing Delays in MAJX



Impact of Timing Delays in Multi-RowCopy

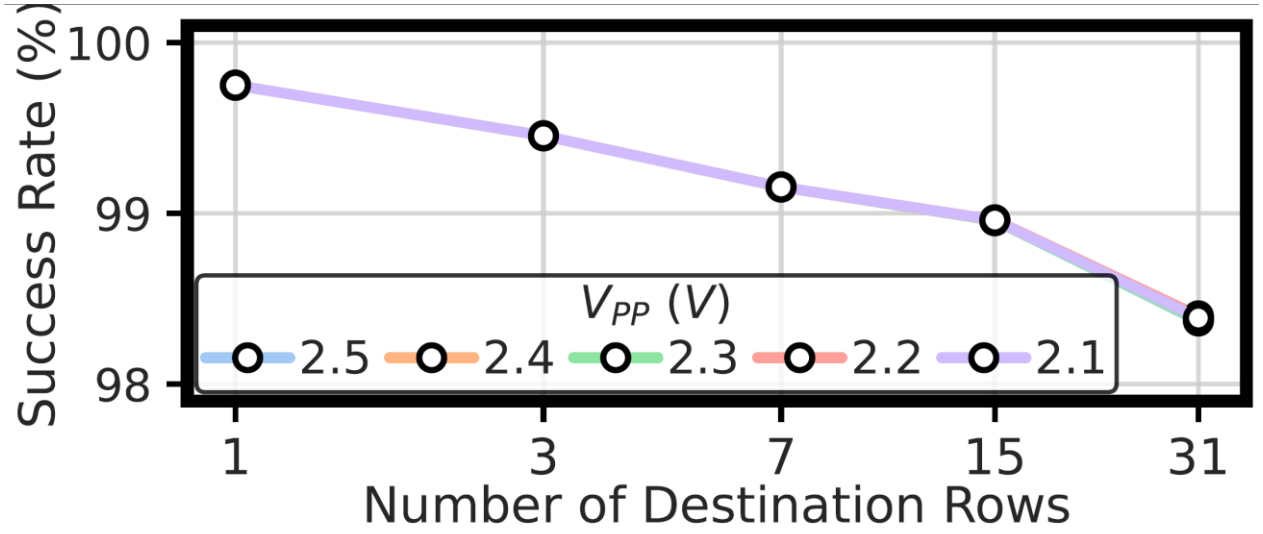


Impact of Temperature in Multi-RowCopy



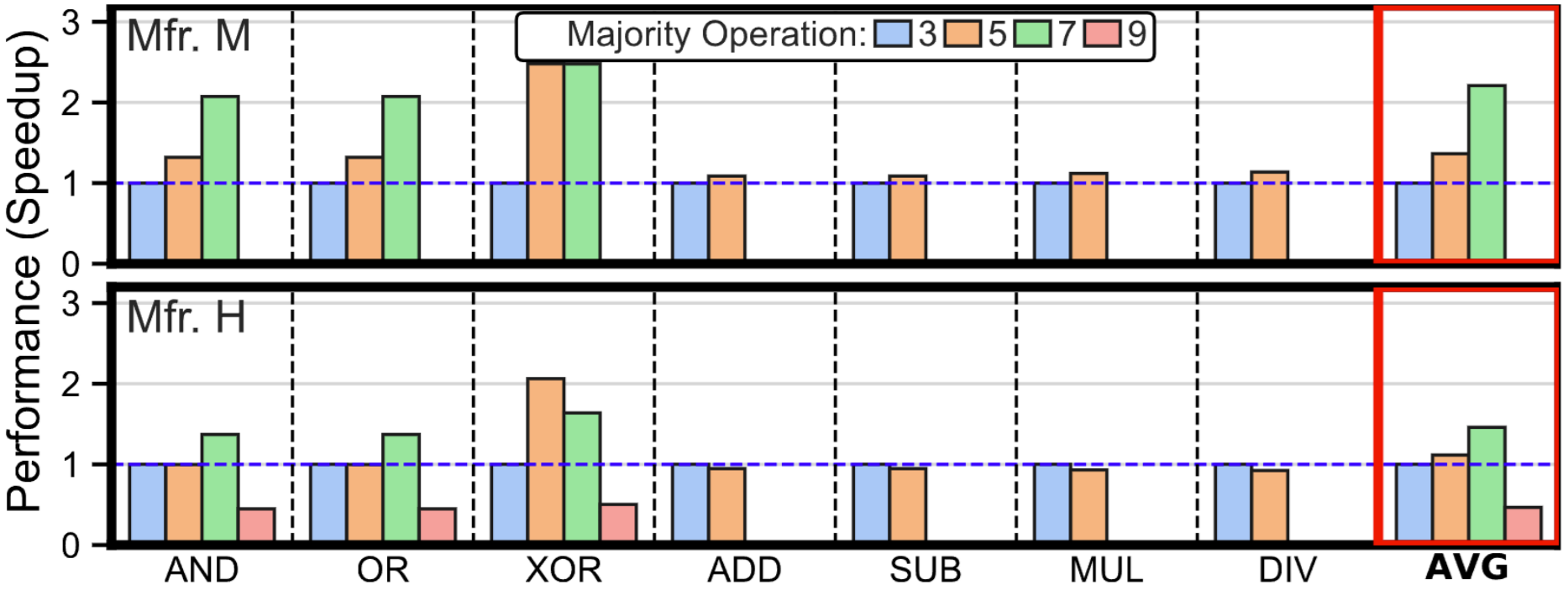
Increasing temperature up to 90°C has a very small effect on the success rate

Impact of Voltage in Multi-RowCopy



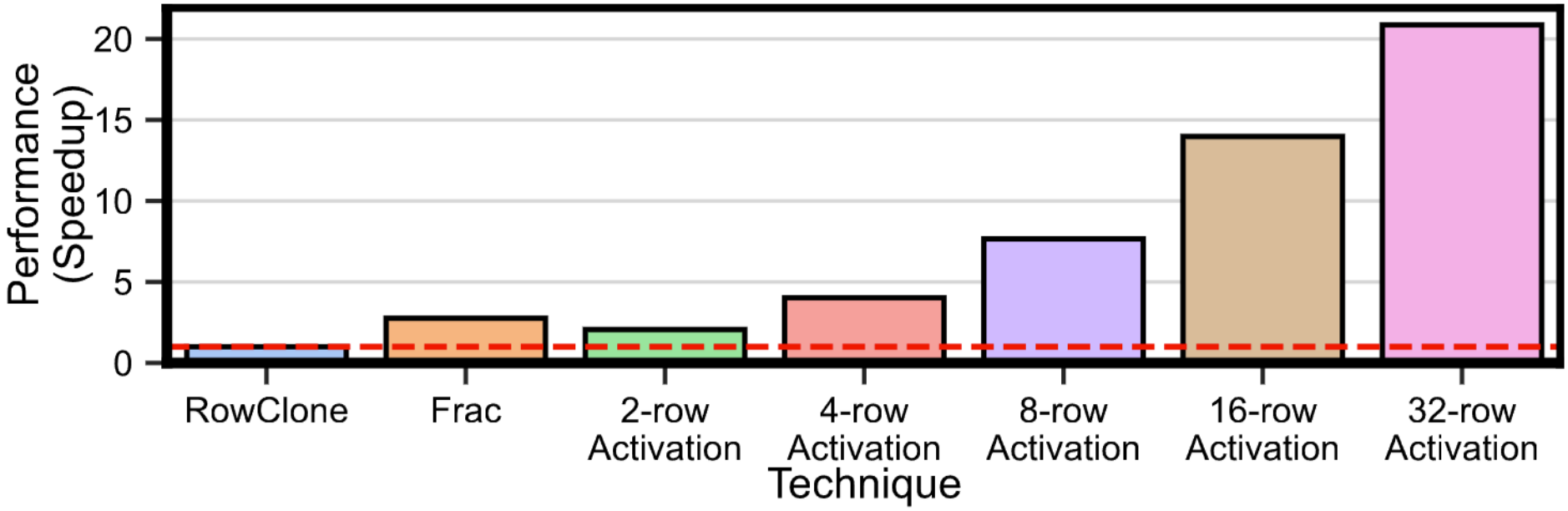
Reducing the wordline voltage only slightly affects the success rate

Majority-based Computation



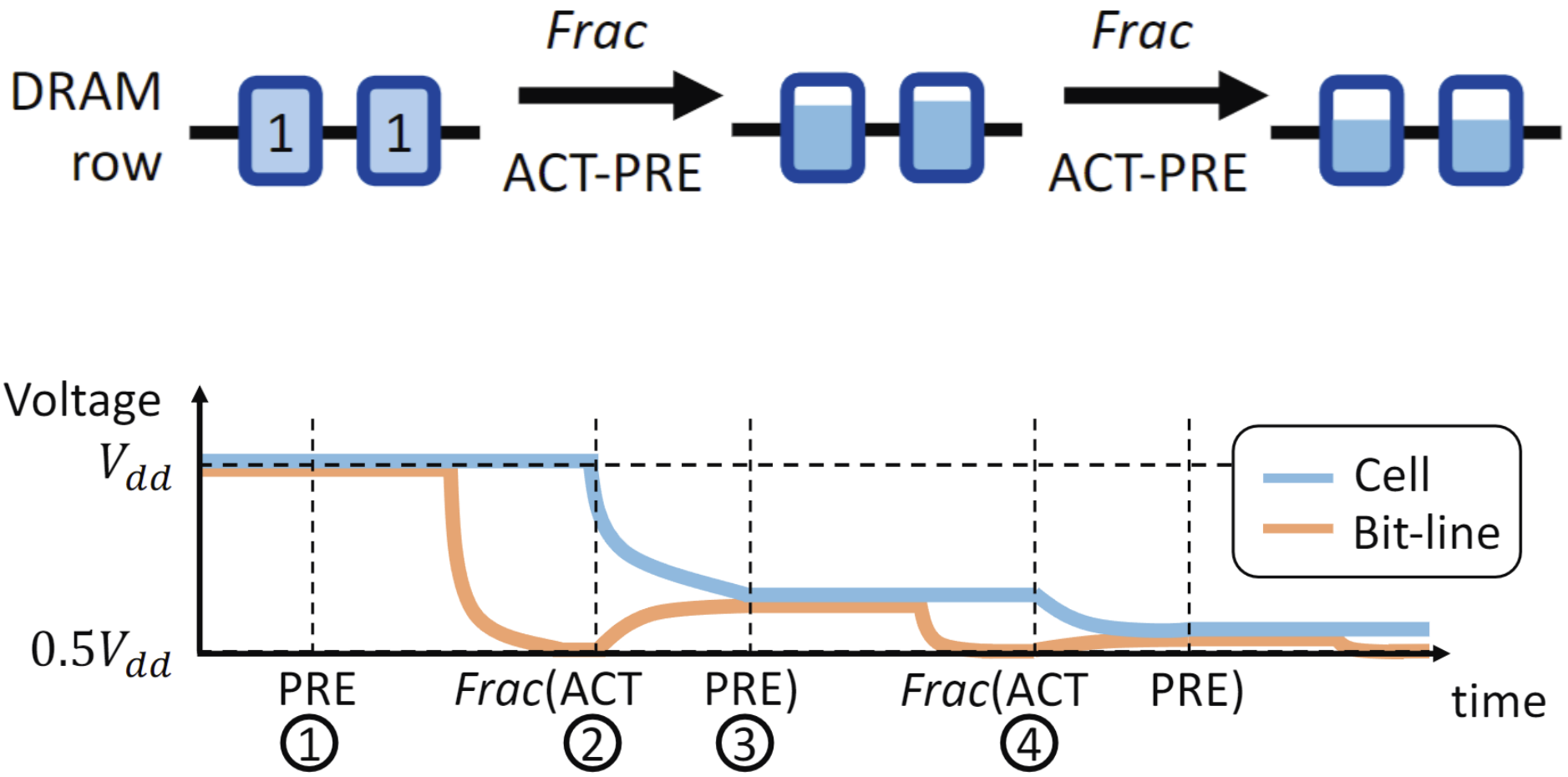
New MAJX operations provide 121.61% (46.54%) higher performance over using only MAJ3 in Mfr. M (Mfr. H) on average.

Cold Boot Attack Prevention

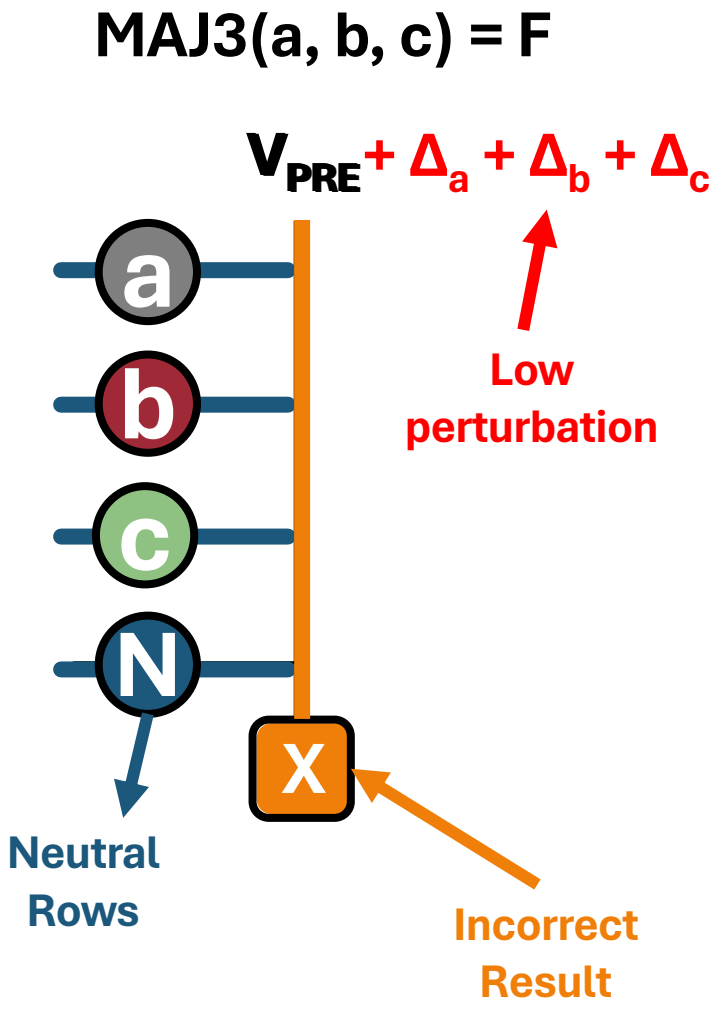


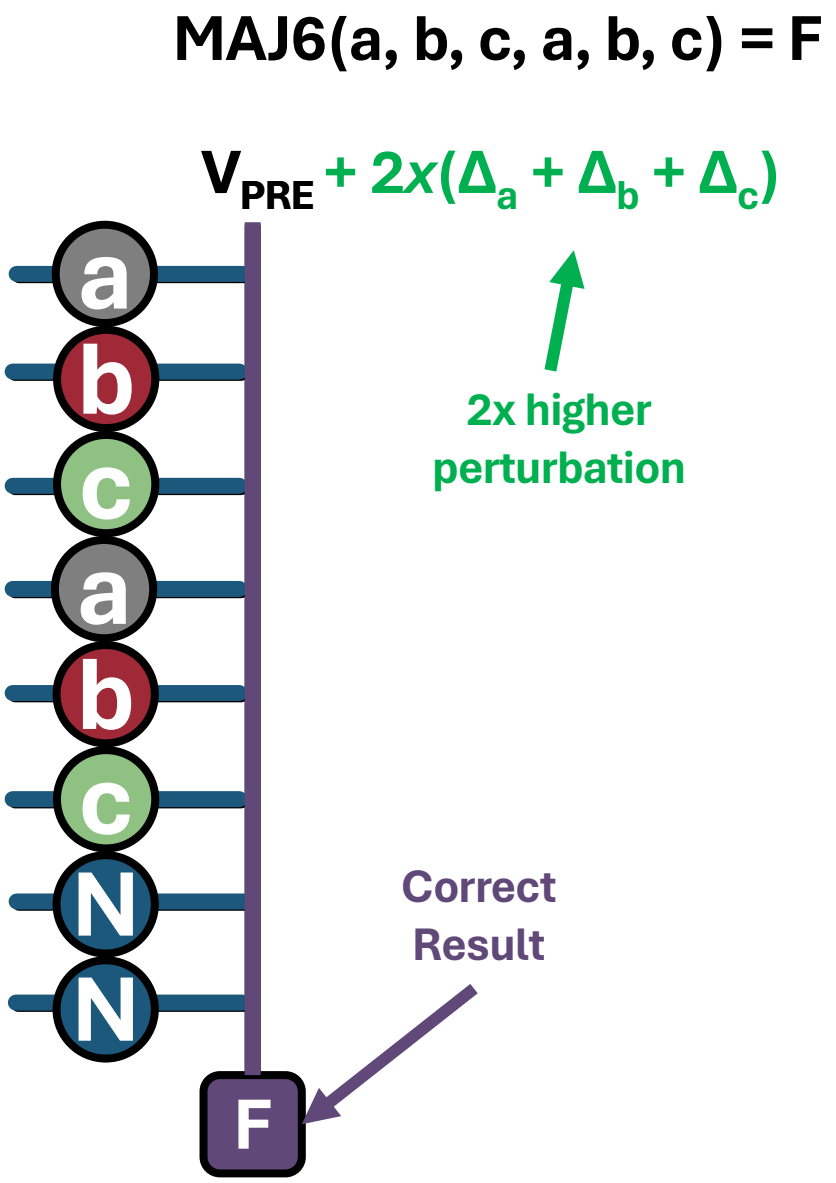
Multi-RowCopy-based content destruction outperforms both RowClone-based and Frac-based content destruction by up to 20.87× and 7.55×, respectively.

Frac Operation



Input Replication in Real Chips

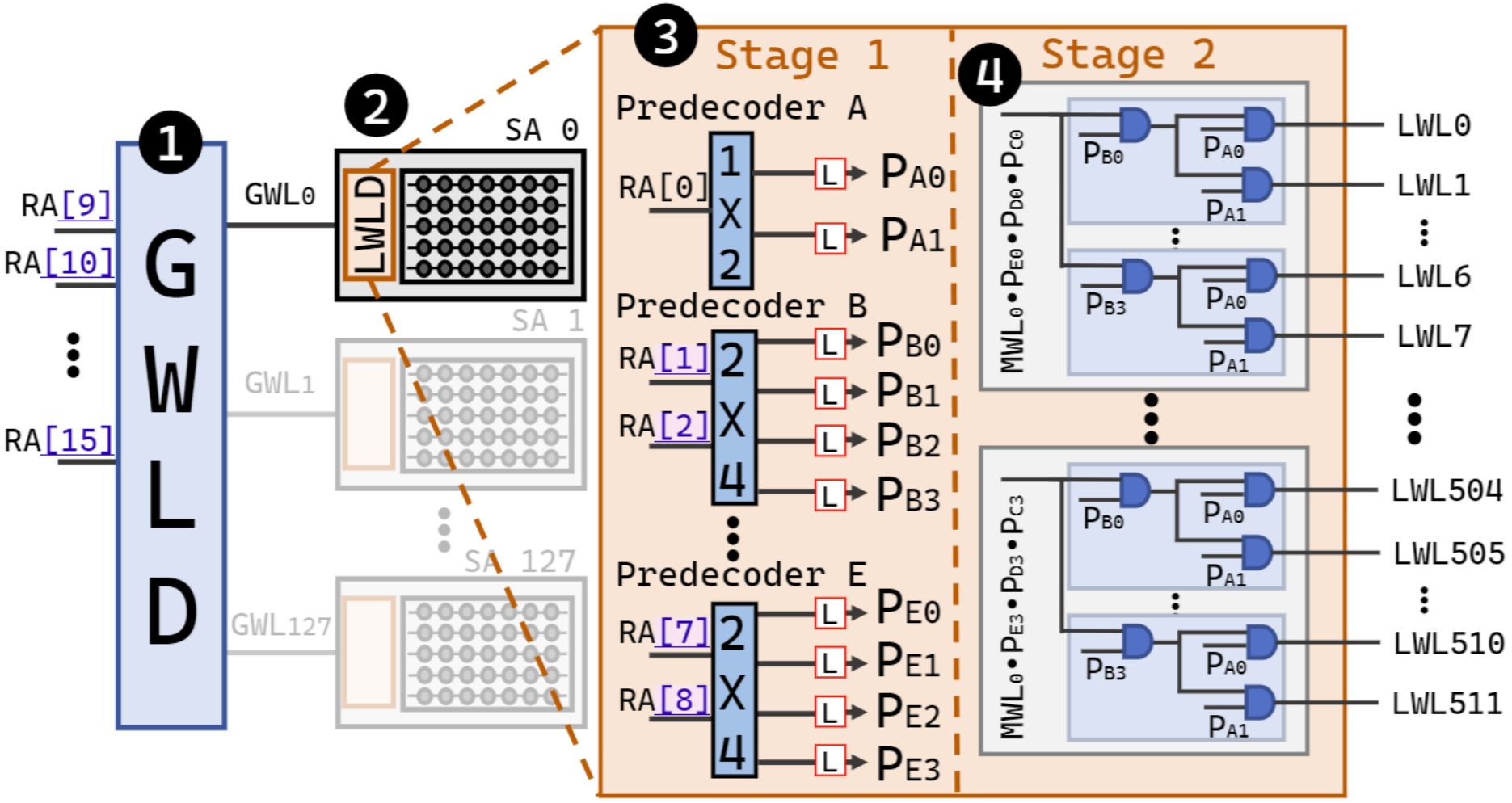




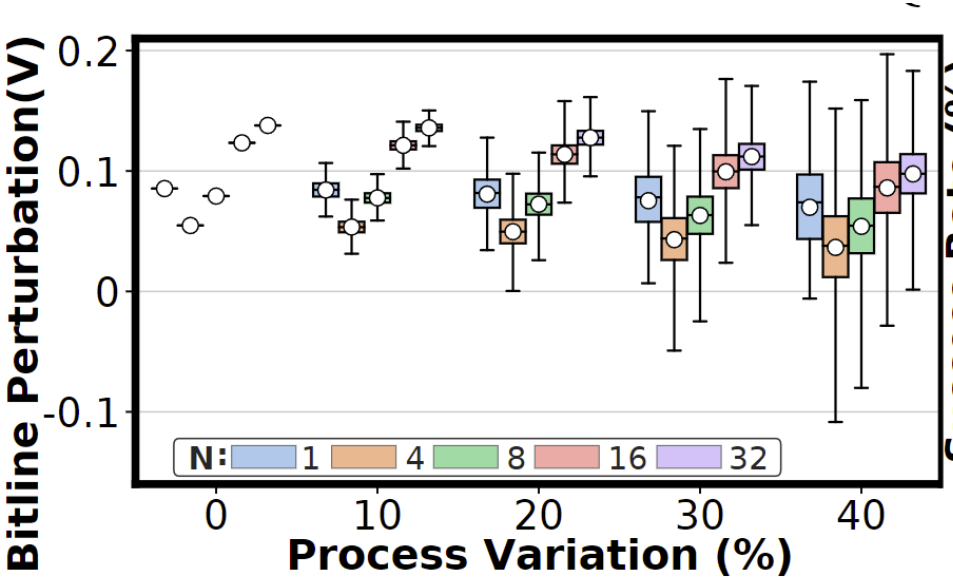
DRAM Chips Tested: Extended Table

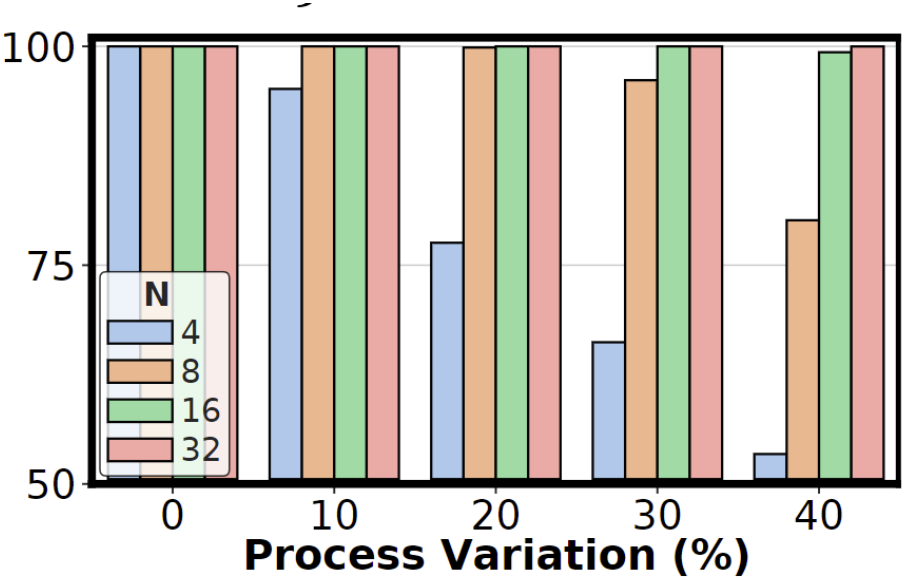
Module Vendor	Chip Vendor	Module Identifier Chip Identifier	#Modules (#Chips)	Freq (MT/s)	Mfr. Date ww-yy	Chip Den.	Die Rev.	Chip Org.	Subarray Size
TimeTec	SK Hynix	TLRD44G2666HC18F-SBK [240] H5AN4G8NMFR-TFC [241]	7 (56)	2666	Unknown	4Gb	M	×8	512 or 640
TeamGroup	SK Hynix	76TT21NUS1R8-4G [242] H5AN4G8NAFR-TFC [243]	5 (40)	2133	Unknown	4Gb	M	×8	512
Micron	Micron	MTA4ATF1G64HZ-3G2E1 [244] MT40A1G16KD-062E:E [245]	4 (16)	3200	46-20	16Gb	E	×16	1024
Micron	Micron	MTA4ATF1G64HZ-3G2B2 [246] MT40A1G16RC-062E:B [247]	2 (8)	2666	26-21	16Gb	B	×16	1024

Row Decoder Circuitry



Effect of Input Replication on the Bitline Deviation





Limitations of Tested COTS DRAM Chips (I)

- **Some COTS DRAM chips do not support all in-DRAM operations**
  - We do not observe simultaneous many-row activation in tested 64 Samsung 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

Limitations of Tested COTS DRAM Chips (III)

- **Performing in-DRAM operations potentially have an effect on transient errors in DRAM chips**
  - We perform each test (a single data point in the distribution) 10K times
  - We do not observe any errors in rows outside of the simultaneously activated row group

We believe that investigating all potential effects (e.g., on transient errors) requires a much more extensive exploration of various aspects

Open Research Questions

1. Is it possible to **robustly** activate **more than four** DRAM rows simultaneously?

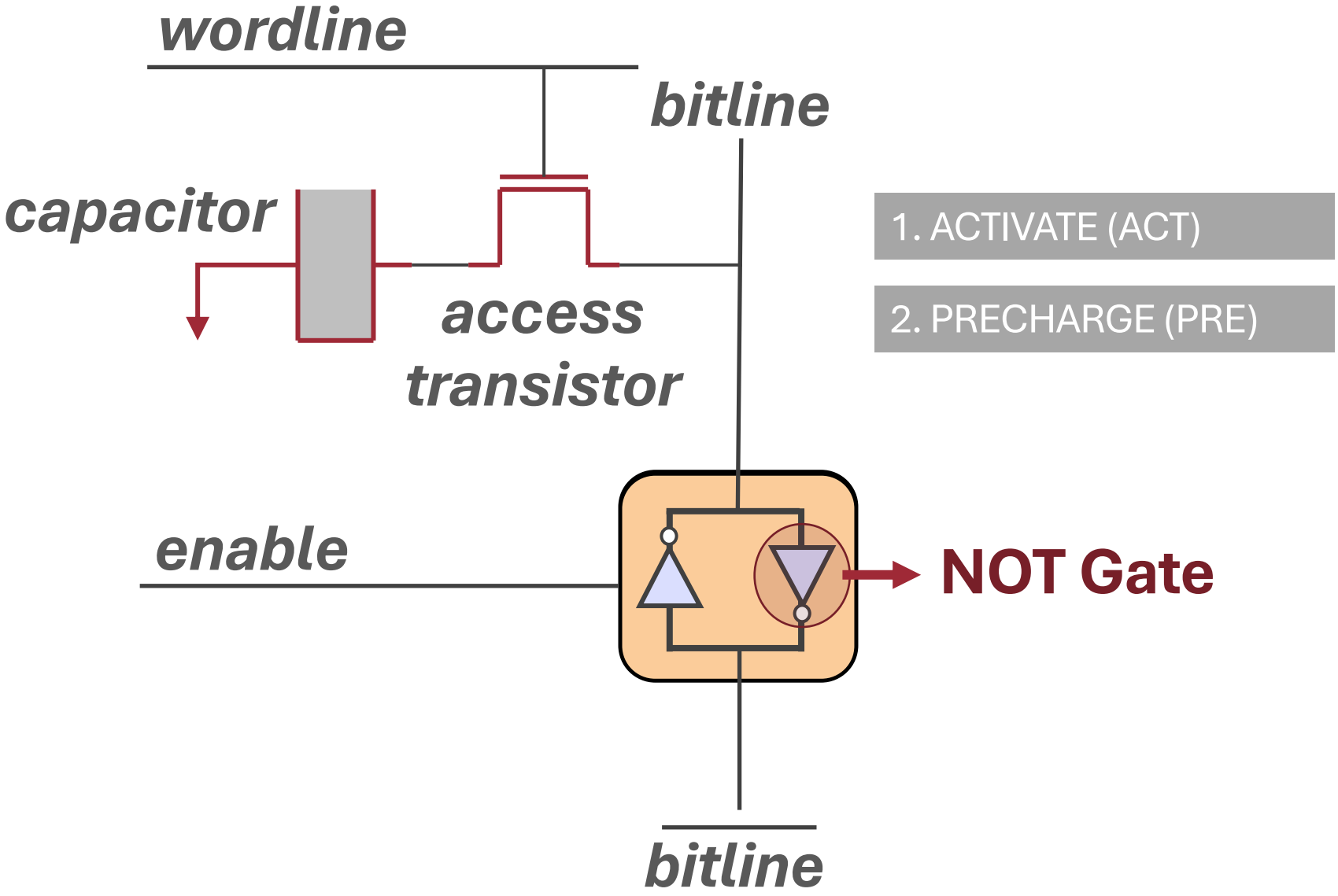
2. What **other PUD operations** can be realized in COTS DRAM chips?

3. How **robustly** can PUD operations be performed in COTS DRAM chips?

4. Can **the robustness** of PUD operations **be improved**?

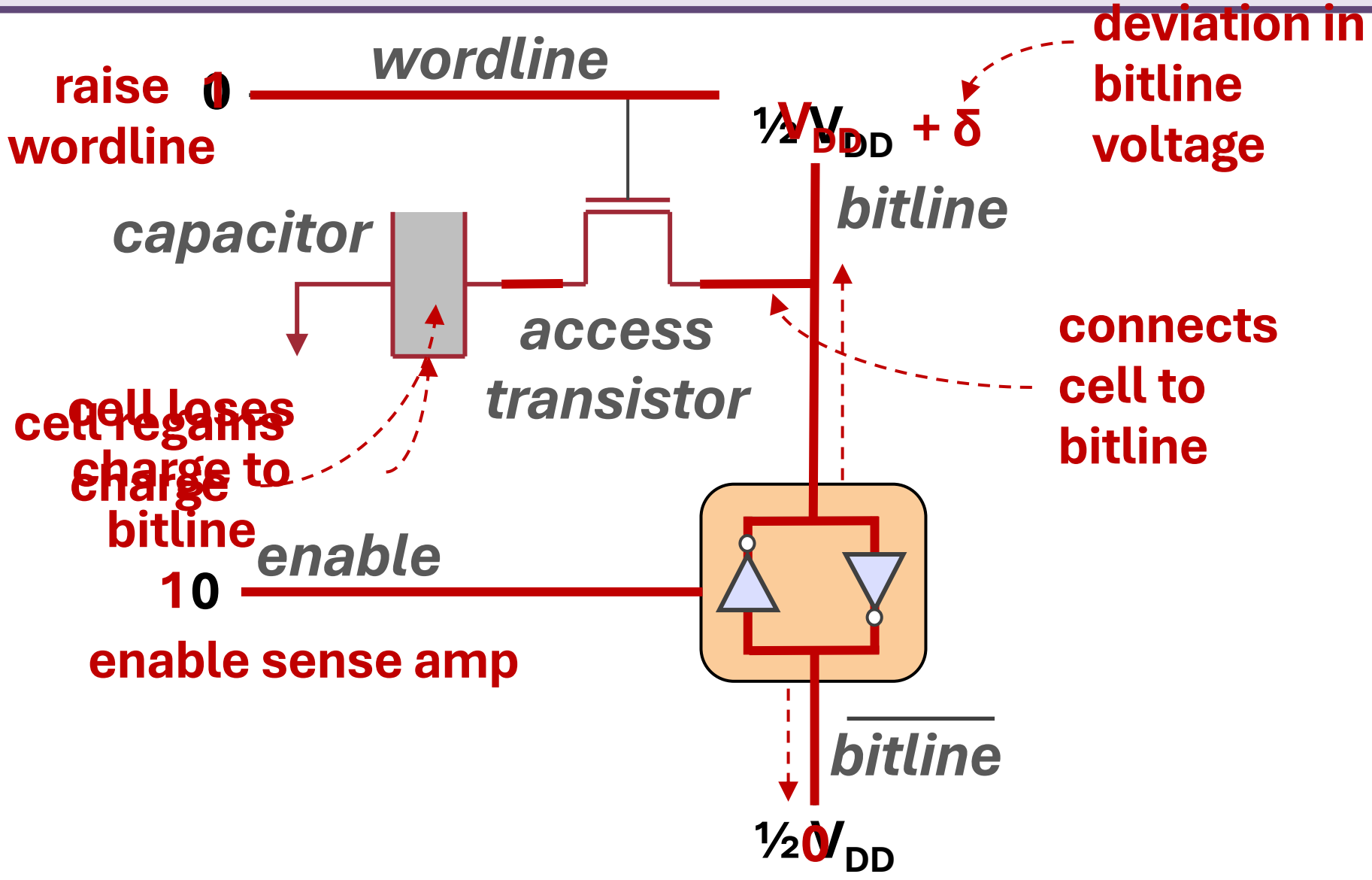
5. What are **the effects of operating conditions** on the robustness of PUD operations?

DRAM Cell Operation



1. ACTIVATE (ACT)
2. PRECHARGE (PRE)

DRAM Cell Operation - ACTIVATE



DRAM Cell Operation - PRECHARGE