Fundamentally Understanding DRAM Reliability & Enabling Fast and Secure Memory

Abdullah Giray Yaglikci

agyaglikci@gmail.com

https://agyaglikci.github.io

9 April 2024

Microsoft Swiss Joint Research Center

agyaglikci.github.io

safari.ethz.ch

SAFARI

ETH zürich

Outline

Lack of Memory Isolation

Infrastructure

Understanding Read Disturbance on Real DRAM Chips

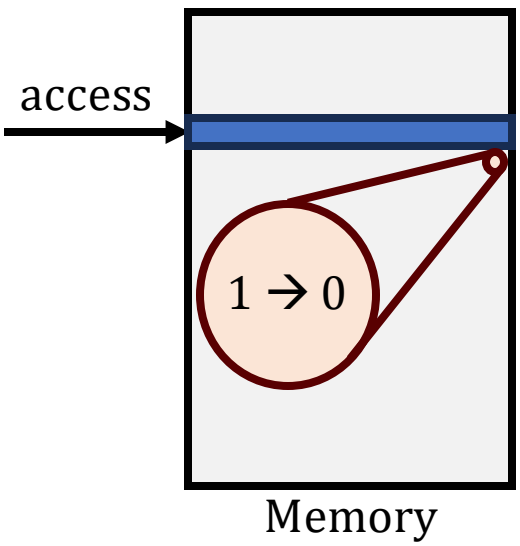
Efficient and Scalable DRAM Read Disturbance Solutions

Data Movement Bottleneck

Addressing Data Movement Bottleneck on Off-the-Shelf DRAM Chips

Current and Future Challenges

Lack of Memory Isolation



Data **loss** or **corruption**

Compromise application **correctness**

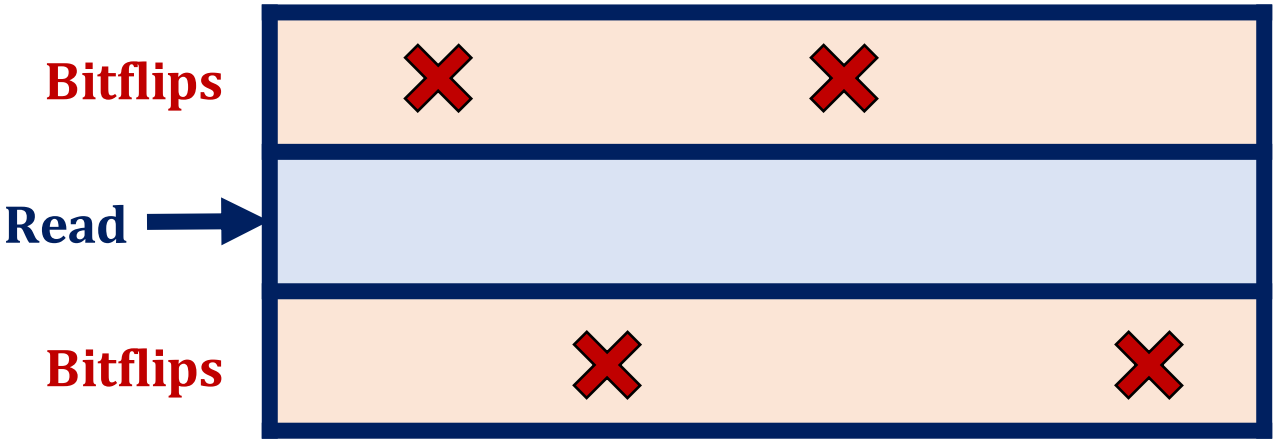
Leak private information

Take over a computer

An access to one memory address
should not have **unintended side effects**
on data stored in **other addresses**

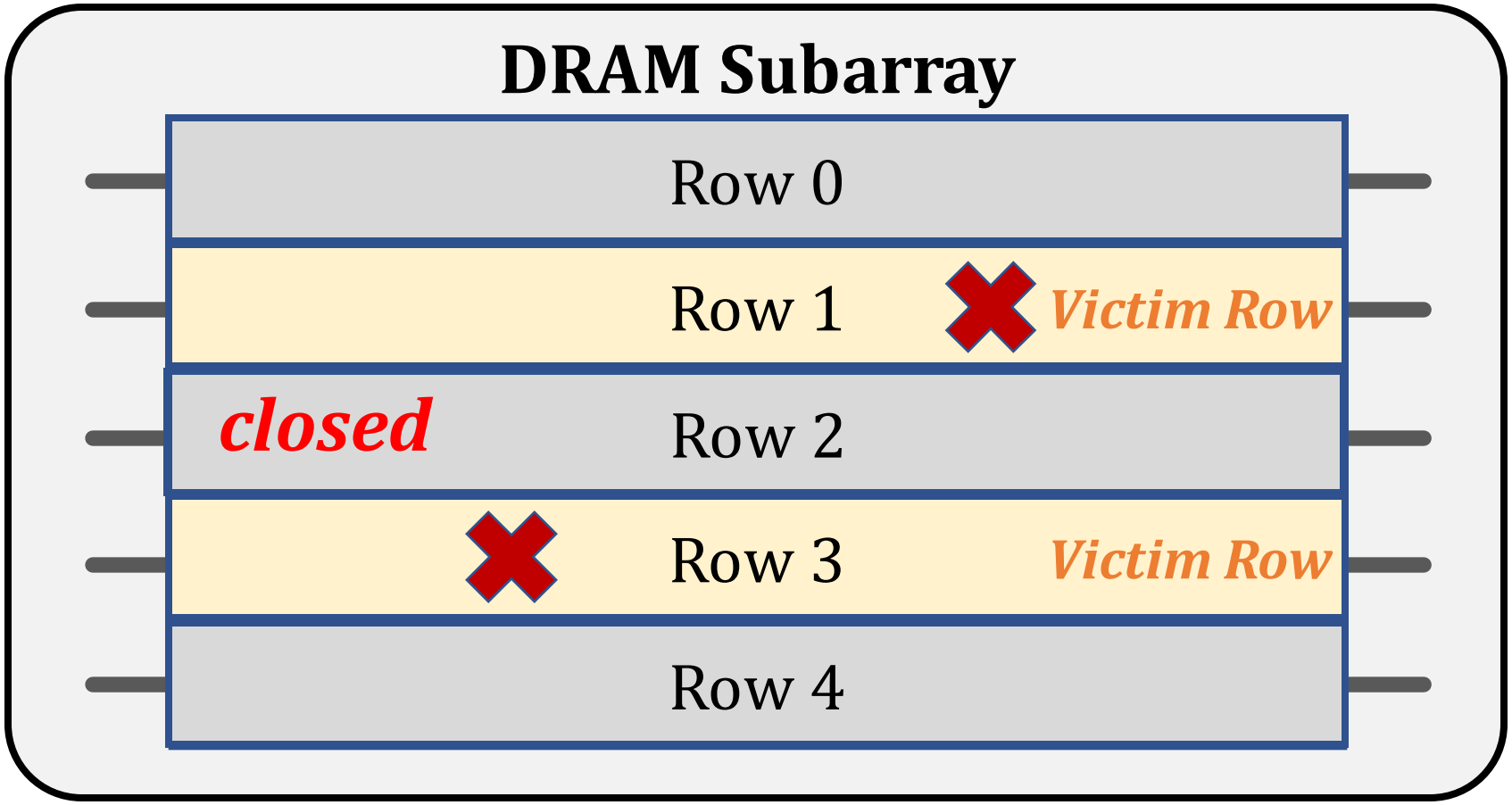
Memory isolation is **difficult in modern memory chips**

DRAM Read Disturbance



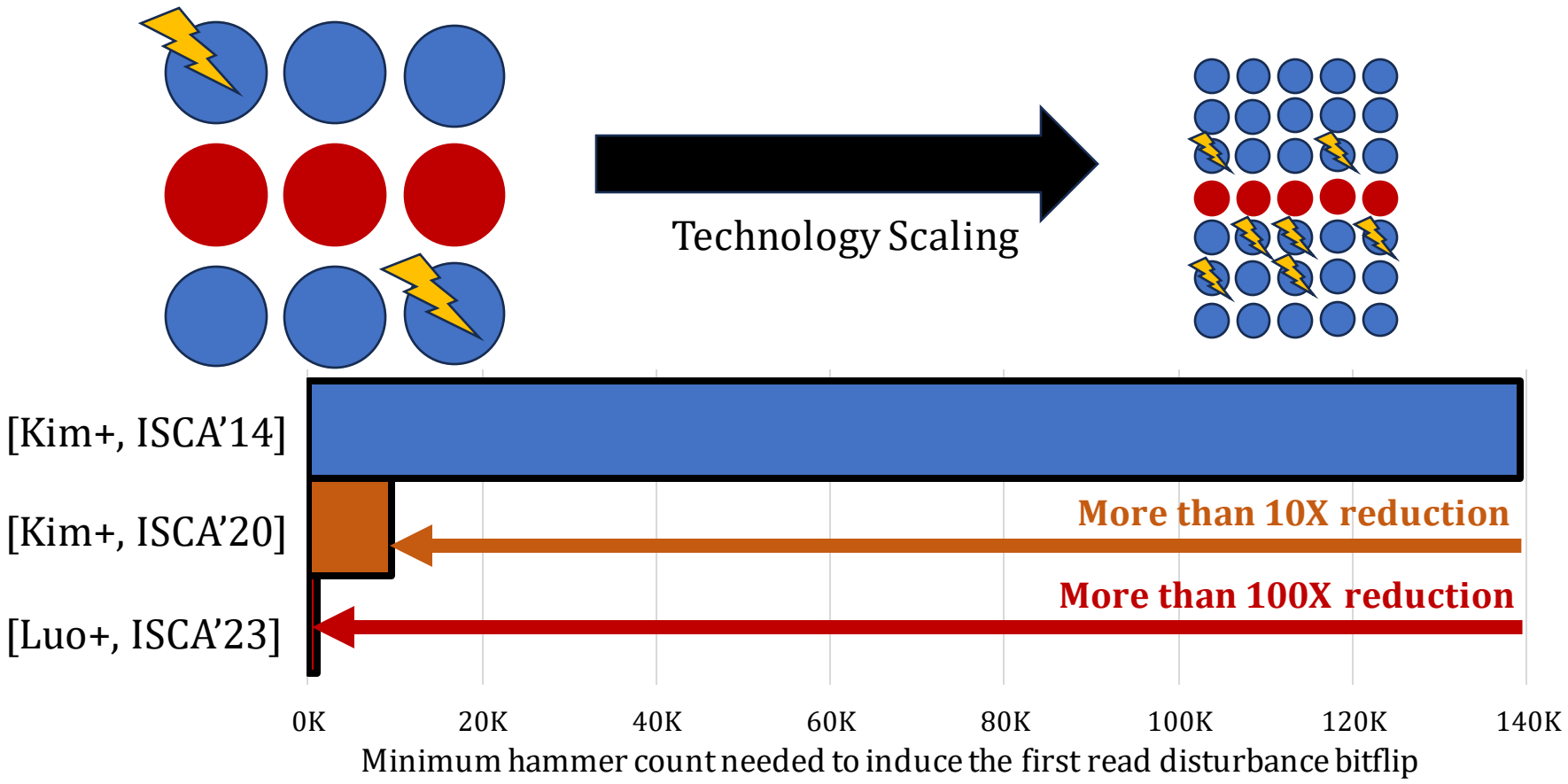
Reading from a memory location
disturbs data in **physically nearby** locations

The RowHammer Vulnerability [Kim+, ISCA'14]



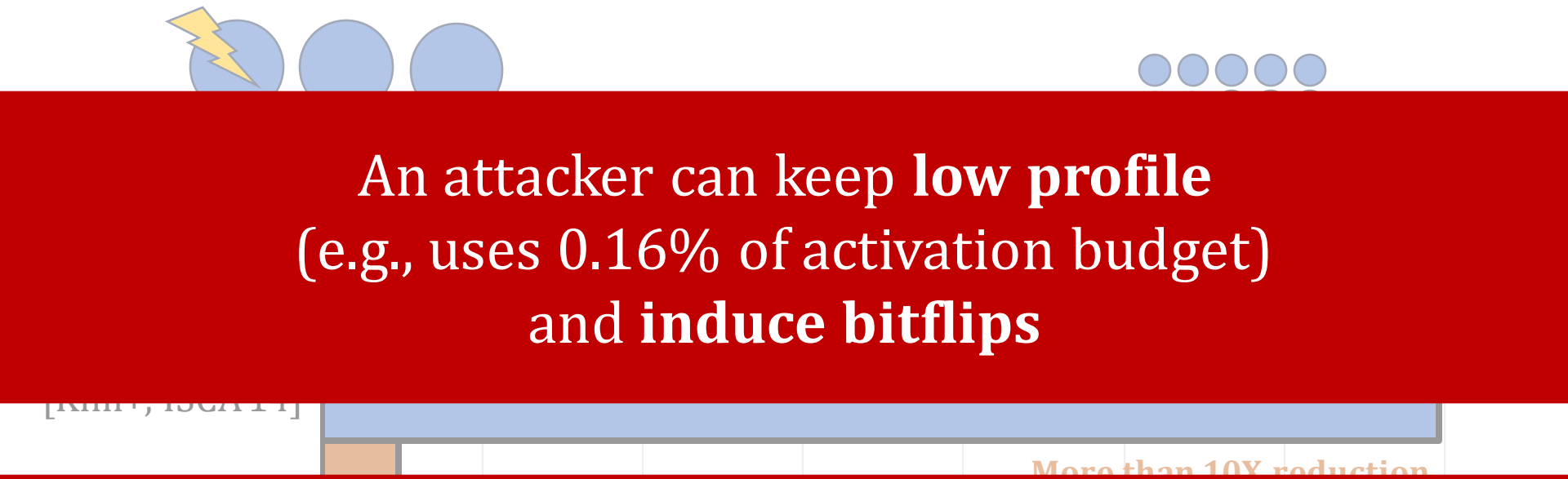
Repeatedly **opening** (activating) and **closing** (precharging) a DRAM row causes **RowHammer bitflips** in nearby cells and breaks **memory isolation**

Motivation



DRAM chips are **increasingly more vulnerable** to **read disturbance** with **technology scaling**

Motivation



An attacker can keep **low profile**
(e.g., uses 0.16% of activation budget)
and **induce bitflips**

Preventing bitflips requires
tracking many rows and **performing many refreshes**

DRAM chips are **increasingly more vulnerable**
to **read disturbance** with **technology scaling**

Read Disturbance is an Outstanding Problem

Increasing DRAM chip density
exacerbates DRAM read disturbance

Attackers can keep **low profile**
(using <0.16% of the row activation budget)

Efficient and scalable solutions are needed

A **deeper understanding** of
DRAM read disturbance is the key to enable
efficient and scalable solutions

Outline

Lack of Memory Isolation

Infrastructure

Understanding Read Disturbance on Real DRAM Chips

Efficient and Scalable DRAM Read Disturbance Solutions

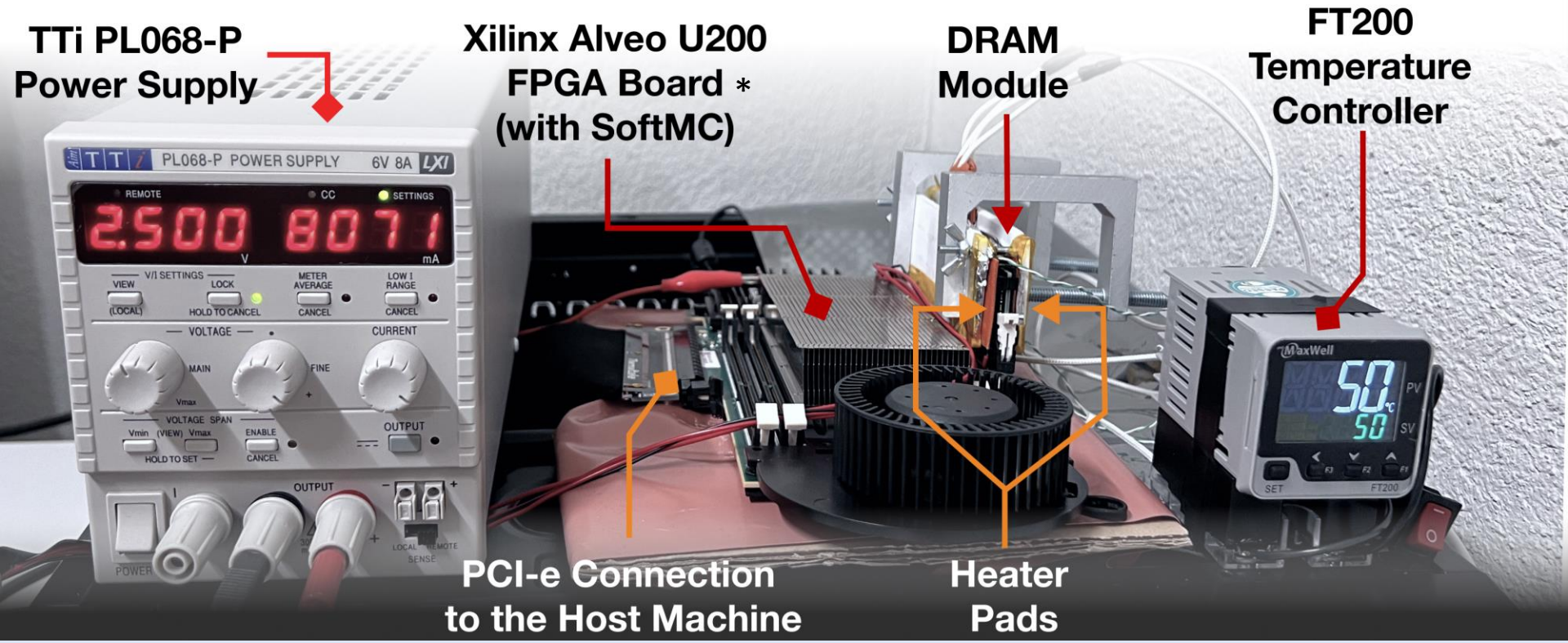
Data Movement Bottleneck

Addressing Data Movement Bottleneck on Off-the-Shelf DRAM Chips

Current and Future Challenges

DRAM Testing Infrastructure: DRAM Bender

DRAM Bender on a Xilinx Virtex UltraScale+ XCU200



Fine-grained control over **DRAM commands**, **timing parameters (±1.5ns)**, **temperature (±0.5°C)**, and **wordline voltage (±1mV)**

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

DRAM Bender

- Ataberk Olgun, Hasan Hassan, A Giray Yağlıkçı, Yahya Can Tuğrul, Lois Orosa, Haocong Luo, Minesh Patel, Oğuz Ergin, and Onur Mutlu,
  "DRAM Bender: An Extensible and Versatile FPGA-based Infrastructure to Easily Test State-of-the-art DRAM Chips"
  IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (***TCAD***), 2023.
  [Extended arXiv version]
  [DRAM Bender Source Code]
  [DRAM Bender Tutorial Video (43 minutes)]

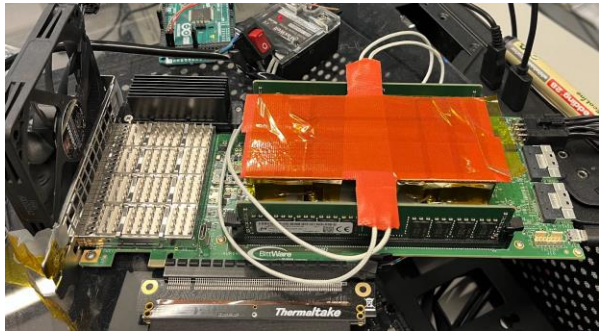
DRAM Bender: An Extensible and Versatile FPGA-based Infrastructure to Easily Test State-of-the-art DRAM Chips

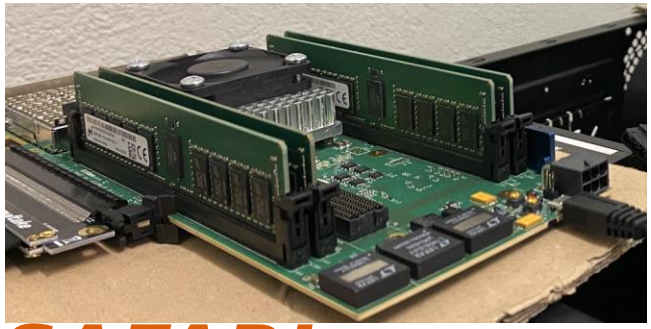
Ataberk Olgun[§] Hasan Hassan[§] A. Giray Yağlıkçı[§] Yahya Can Tuğrul[§†]
Lois Orosa[§⊙] Haocong Luo[§] Minesh Patel[§] Oğuz Ergin[†] Onur Mutlu[§]
[§]*ETH Zürich* [†]*TOBB ETÜ* [⊙]*Galician Supercomputing Center*

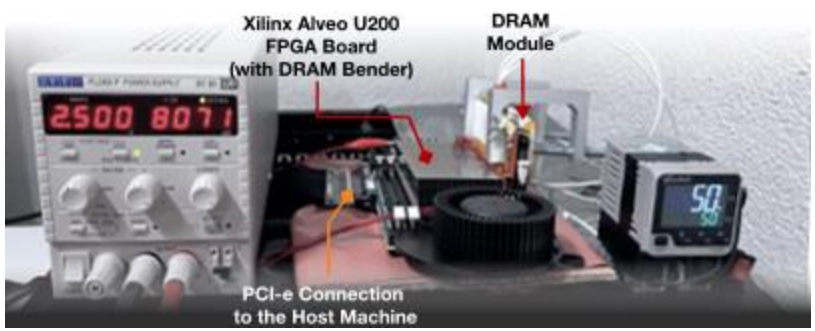
DRAM Bender: Prototypes

Testing Infrastructure	Protocol Support	FPGA Support
SoftMC [134]	DDR3	One Prototype
LiteX RowHammer Tester (LRT) [17]	DDR3/4, LPDDR4	Two Prototypes
DRAM Bender (this work)	DDR3/DDR4/HBM2	**Five Prototypes**

Five out of the box FPGA-based prototypes







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

Outline

Lack of Memory Isolation

Infrastructure

Understanding Read Disturbance on Real DRAM Chips

Efficient and Scalable DRAM Read Disturbance Solutions

Data Movement Bottleneck

Addressing Data Movement Bottleneck on Off-the-Shelf DRAM Chips

Current and Future Challenges

Outline

Lack of Memory Isolation

Infrastructure

Understanding Read Disturbance on Real DRAM Chips

- Temperature
- Memory Access Patterns
- Voltage
- RowHammer in HBM2
- Spatial Variation across DRAM Rows

Efficient and Scalable DRAM Read Disturbance Solutions

Data Movement Bottleneck

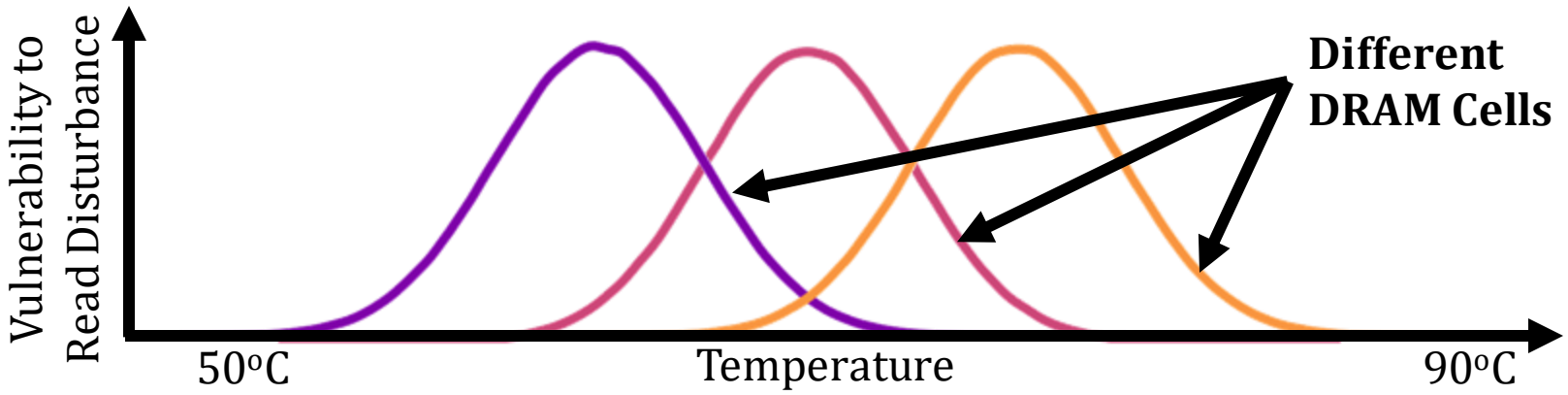
Addressing Data Movement Bottleneck on Off-the-Shelf DRAM Chips

Current and Future Challenges

A Deeper Look into RowHammer

- Lois Orosa*, **Abdullah Giray Yağlıkçı***, Haocong Luo, Ataberk Olgun, Jisung Park, Hasan Hassan, Minesh Patel, Jeremie S. Kim, and Onur Mutlu,
  "A Deeper Look into RowHammer's Sensitivities: Experimental Analysis of Real DRAM Chips and Implications on Future Attacks and Defenses"
  *Proceedings of the 54th International Symposium on Microarchitecture (**MICRO**)*, Virtual, October 2021.
  [Slides (pptx) (pdf)] [Talk Video (21 minutes)]
  [Short Talk Slides (pptx) (pdf)]
  [Lightning Talk Slides (pptx) (pdf)] [Lightning Talk Video (1.5 minutes)]
  [arXiv version]

Temperature



DRAM read disturbance is more effective **within a bounded temperature range**

Vulnerable temperature range varies **across cells**

A DRAM cell should be tested
at **each possible** operating temperature

Trap-Assisted Charge Leakage Model [Yang+, EDL 2019]

- Hammering a wordline **pulls and pushes electrons**
- Electrons **get trapped** and **exacerbate charge leakage**, leading to cause bitflips
- With **increasing temperature**, it becomes **less likely for an electron to get trapped**

A Follow-Up Work: SpyHammer [Orosa+, arXiv'22]

SpyHammer: Using RowHammer to Remotely Spy on Temperature

Lois Orosa[1,2] Ulrich Rührmair[3,4] A. Giray Yağlıkçı[1] Haocong Luo[1] Ataberk Olgun[1]
Patrick Jattke[1] Minesh Patel[1] Jeremie Kim[1] Kaveh Razavi[1] Onur Mutlu[1]
[1]*ETH Zürich* [2]*Galicia Supercomputing Center (CESGA)* [3]*LMU München* [4]*University of Connecticut*

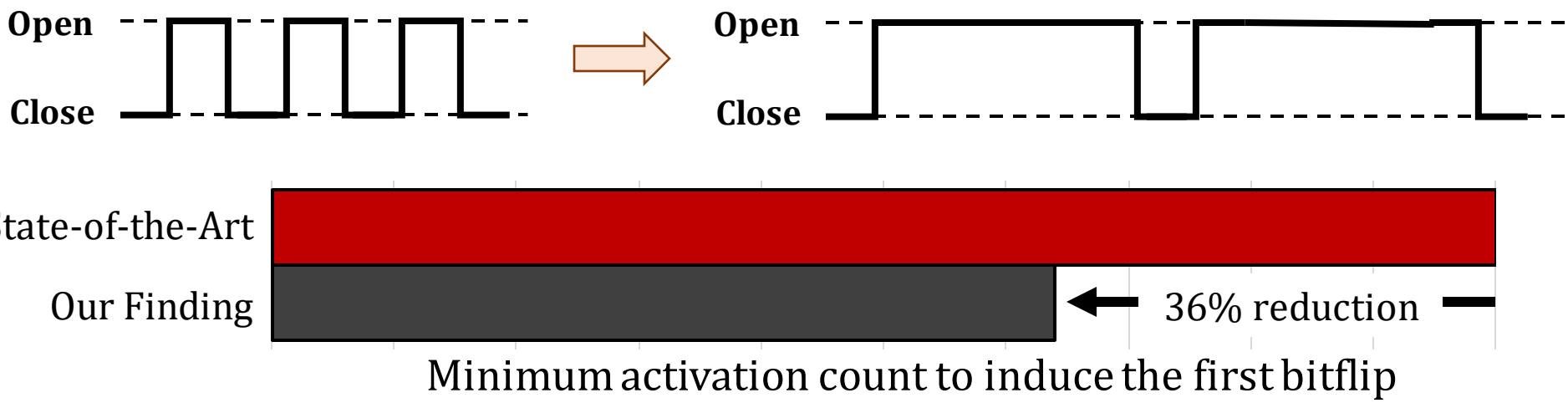
https://arxiv.org/pdf/2210.04084.pdf

Outline

- Lack of Memory Isolation
- Infrastructure
- Understanding Read Disturbance on Real DRAM Chips
  - Temperature
  - Memory Access Patterns
  - Voltage
  - RowHammer in HBM2
  - Spatial Variation across DRAM Rows
- Efficient and Scalable DRAM Read Disturbance Solutions
- Data Movement Bottleneck
- Addressing Data Movement Bottleneck on Off-the-Shelf DRAM Chips
- Current and Future Challenges

Key Findings: Memory Access Patterns

Read disturbance is **more effective**
if the **activated aggressor row** stays **active longer**



Fewer reads cause a **more significant** read disturbance
when the activated aggressor row stays **active longer**

Existing mitigations are **ineffective** without this insight

RowPress [Luo+, ISCA 2023]

- Haocong Luo, Ataberk Olgun, Giray Yaglikci, Yahya Can Tugrul, Steve Rhyner, M. Banu Cavlak, Joel Lindegger, Mohammad Sadrosadati, and Onur Mutlu,
  "RowPress: Amplifying Read Disturbance in Modern DRAM Chips"
  Proceedings of the 50th International Symposium on Computer Architecture (***ISCA***), Orlando, FL, USA, June 2023.
  [Slides (pptx) (pdf)] [Lightning Talk Slides (pptx) (pdf)] [Lightning Talk Video (3 min)]
  [RowPress Source Code and Datasets (Officially Artifact Evaluated with All Badges)]
  Best artifact award at ISCA 2023.

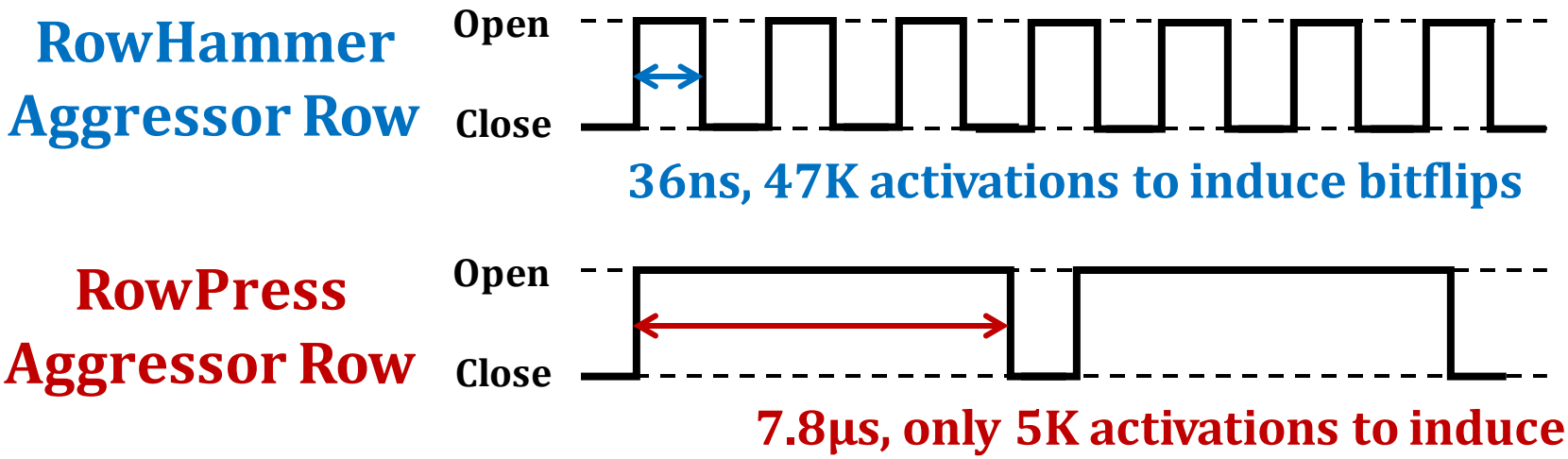
RowPress: Amplifying Read-Disturbance in Modern DRAM Chips

Haocong Luo Ataberk Olgun A. Giray Yağlıkçı Yahya Can Tuğrul Steve Rhyner
Meryem Banu Cavlak Joël Lindegger Mohammad Sadrosadati Onur Mutlu
ETH Zürich

RowPress vs. RowHammer

Instead of using a high activation count,

☞ increase the time that the aggressor row stays open



We observe bitflips even with **ONLY ONE activation** in extreme cases where the row stays open for 30ms

Defenses should perform preventive actions (e.g., refresh) at **much lower activation counts**

Outline

Lack of Memory Isolation

Infrastructure

Understanding Read Disturbance on Real DRAM Chips

- Temperature
- Memory Access Patterns
- Voltage
- RowHammer in HBM2
- Spatial Variation across DRAM Rows

Efficient and Scalable DRAM Read Disturbance Solutions

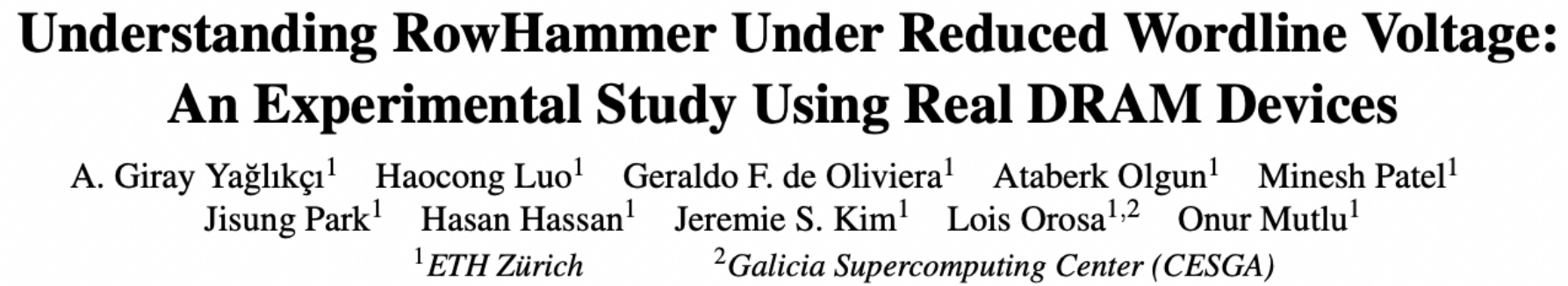
Data Movement Bottleneck

Addressing Data Movement Bottleneck on Off-the-Shelf DRAM Chips

Current and Future Challenges

RowHammer Under Reduced Voltage

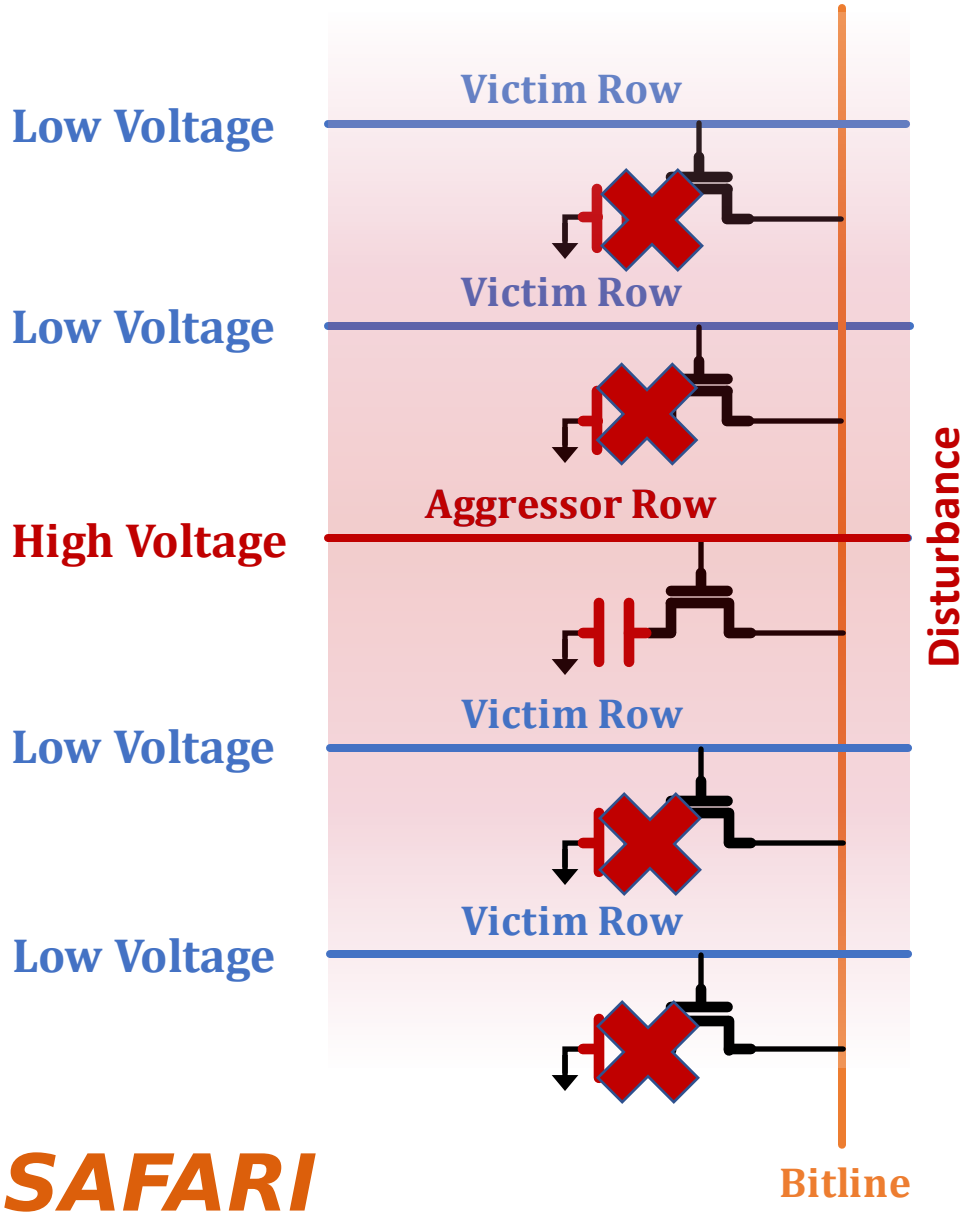
- A. Giray Yağlıkçı, Haocong Luo, Geraldo F. de Oliviera, Ataberk Olgun, Minesh Patel, Jisung Park, Hasan Hassan, Jeremie S. Kim, Lois Orosa, and Onur Mutlu, **"Understanding RowHammer Under Reduced Wordline Voltage: An Experimental Study Using Real DRAM Devices"**
  Proceedings of the 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (***DSN***), Baltimore, MD, USA, June 2022.
  [Slides (pptx) (pdf)]
  [Lightning Talk Slides (pptx) (pdf)]
  [arXiv version]
  [Talk Video (34 minutes, including Q&A)]
  [Lightning Talk Video (2 minutes)]

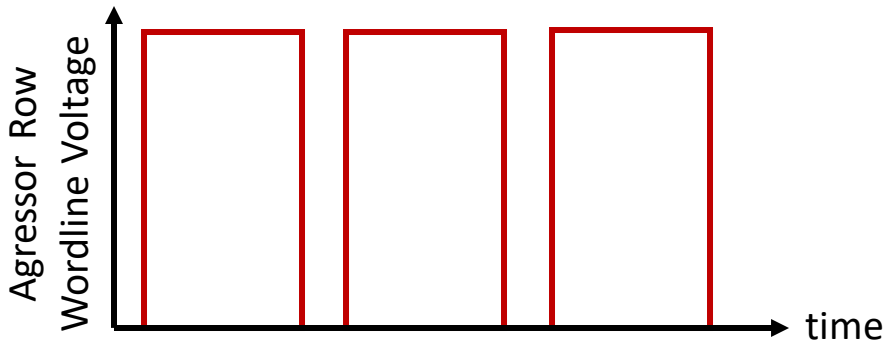


Understanding RowHammer Under Reduced Wordline Voltage: An Experimental Study Using Real DRAM Devices

A. Giray Yağlıkçı[1] Haocong Luo[1] Geraldo F. de Oliviera[1] Ataberk Olgun[1] Minesh Patel[1]
Jisung Park[1] Hasan Hassan[1] Jeremie S. Kim[1] Lois Orosa[1,2] Onur Mutlu[1]
[1]*ETH Zürich* [2]*Galicia Supercomputing Center (CESGA)*

A Closer Look into RowHammer

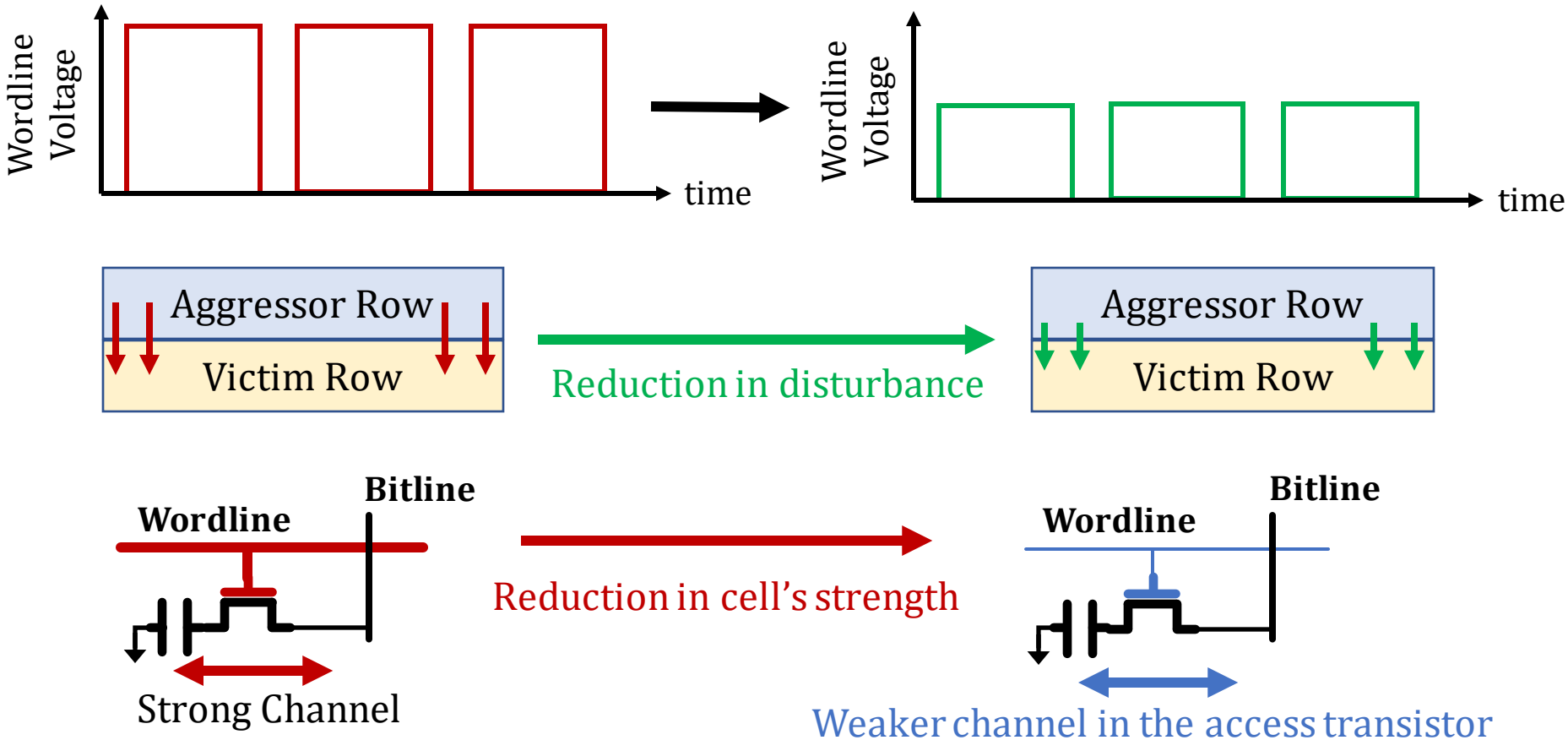




Repeatedly toggling wordline voltage is the *key* to inducing **RowHammer bit flips**

Effects of Reducing Wordline Voltage

Reducing **wordline voltage**
can **reduce RowHammer vulnerability**
without significantly affecting **reliable DRAM operation**



Outline

- Lack of Memory Isolation
- Infrastructure
- Understanding Read Disturbance on Real DRAM Chips
  - Temperature
  - Memory Access Patterns
  - Voltage
  - RowHammer in HBM2
  - Spatial Variation across DRAM Rows
- Efficient and Scalable DRAM Read Disturbance Solutions
- Data Movement Bottleneck
- Addressing Data Movement Bottleneck on Off-the-Shelf DRAM Chips
- Current and Future Challenges

RowHammer in HBM2

- First detailed experimental RowHammer characterization in a modern HBM2 DRAM chip
- Different channels in 3D-stacked HBM chips exhibit different RowHammer vulnerability
- DRAM rows near the end of a DRAM bank are more RowHammer resilient
- A modern HBM chip implements undisclosed on-DRAM-die RowHammer mitigation (e.g., similar to DDR4 chips)

Ataberk Olgun, Majd Osserian, A. Giray Yağlıkçı, Yahya Can Tugrul, Haocong Luo, Steve Rhyner, Behzad Salami, Juan Gomez-Luna, and Onur Mutlu, **"An Experimental Analysis of RowHammer in HBM2 DRAM Chips"** in *Proceedings of the 53nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks Disrupt Track (**DSN Disrupt**)*, Porto, Portugal, June 2023. [Slides (pptx) (pdf)] [Talk Video (24 minutes, including Q&A)]

An Experimental Analysis of RowHammer in HBM2 DRAM Chips

Ataberk Olgun[1] Majd Osseiran[1,2] A. Giray Yağlıkçı[1] Yahya Can Tuğrul[1]
Haocong Luo[1] Steve Rhyner[1] Behzad Salami[1] Juan Gomez Luna[1] Onur Mutlu[1]
[1]*SAFARI Research Group, ETH Zürich* [2]*American University of Beirut*

Outline

Lack of Memory Isolation

Infrastructure

Understanding Read Disturbance on Real DRAM Chips

- Temperature
- Memory Access Patterns
- Voltage
- RowHammer in HBM2
- Spatial Variation across DRAM Rows

Efficient and Scalable DRAM Read Disturbance Solutions

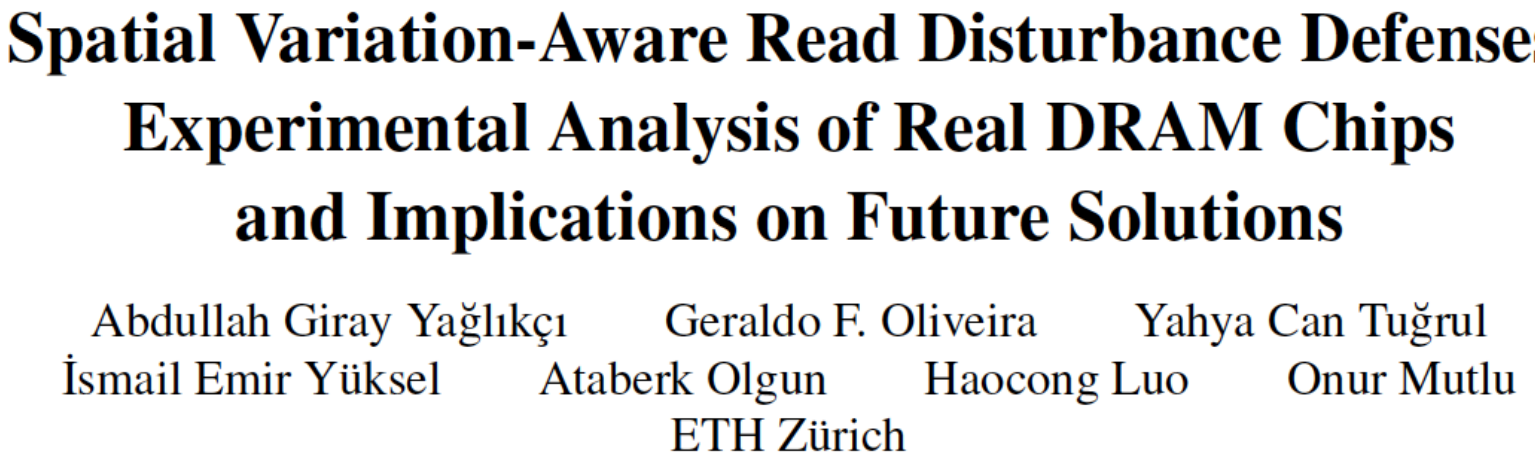
Data Movement Bottleneck

Addressing Data Movement Bottleneck on Off-the-Shelf DRAM Chips

Current and Future Challenges

Spatial Variation-Aware Read Disturbance Defenses

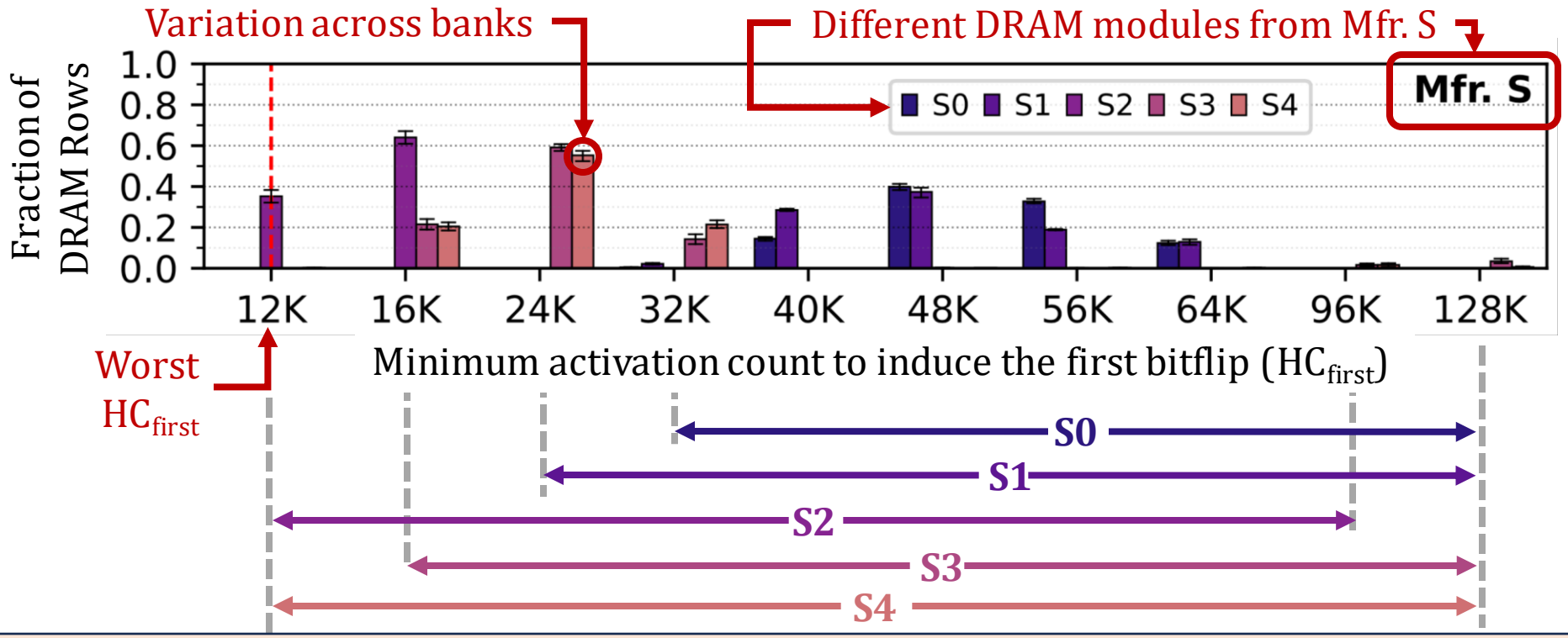
- A. Giray Yağlıkçı, Geraldo F. de Oliviera, Yahya Can Tuğrul, İsmail Emir Yüksel, Ataberk Olgun, Haocong Luo, and Onur Mutlu, **"Spatial Variation-Aware Read Disturbance Defenses: Experimental Analysis of Real DRAM Chips and Implications on Future Solutions,"** *Proceedings of the 30th Edition of The **International Symposium on High-Performance Computer Architecture** (HPCA), Edinburgh, Scotland, UK, March 2024.*



Spatial Variation-Aware Read Disturbance Defenses: Experimental Analysis of Real DRAM Chips and Implications on Future Solutions

Abdullah Giray Yağlıkçı Geraldo F. Oliveira Yahya Can Tuğrul
İsmail Emir Yüksel Ataberk Olgun Haocong Luo Onur Mutlu
ETH Zürich

Spatial Variation in the Minimum Hammer Count to Induce the First Bitflip across DRAM Rows



The minimum activation count to induce the first bitflip **significantly varies across rows** in a DRAM bank

Not all rows need **the same level** of protection

Outline

Lack of Memory Isolation

Infrastructure

Understanding Read Disturbance on Real DRAM Chips

Efficient and Scalable DRAM Read Disturbance Solutions

Data Movement Bottleneck

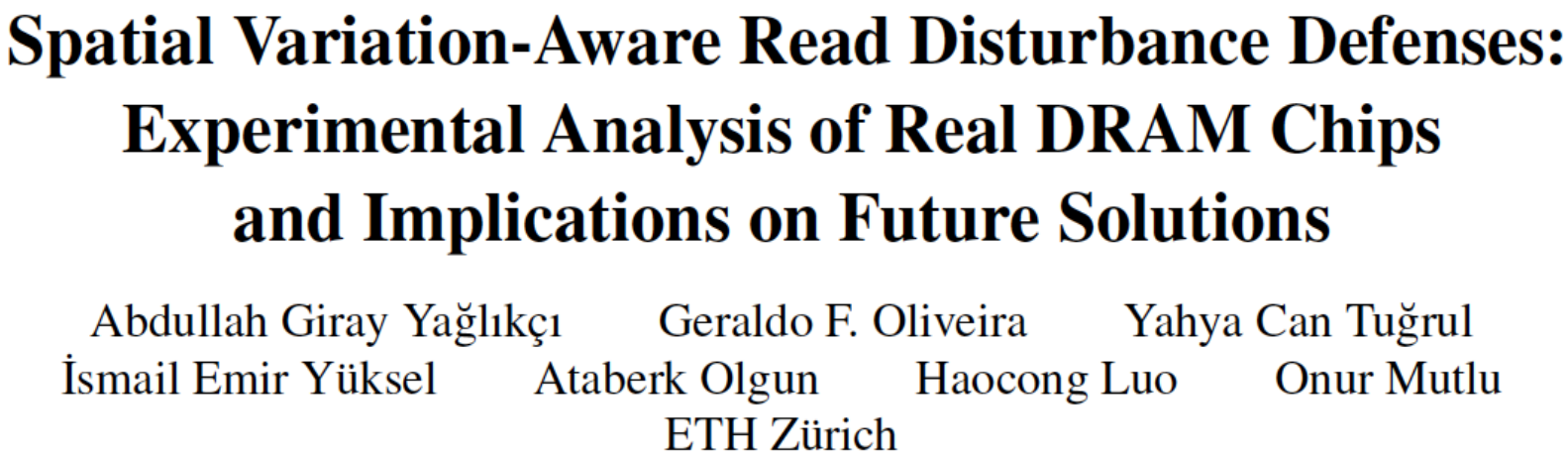
Addressing Data Movement Bottleneck on Off-the-Shelf DRAM Chips

Current and Future Challenges

Outline

Svärd: Spatial Variation-Aware Read Disturbance Defenses

- A. Giray Yağlıkçı, Geraldo F. de Oliviera, Yahya Can Tuğrul, İsmail Emir Yüksel, Ataberk Olgun, Haocong Luo, and Onur Mutlu, **"Spatial Variation-Aware Read Disturbance Defenses: Experimental Analysis of Real DRAM Chips and Implications on Future Solutions,"** *Proceedings of the 30th Edition of The **International Symposium on High-Performance Computer Architecture** (HPCA), Edinburgh, Scotland, UK, March 2024.*



Spatial Variation-Aware Read Disturbance Defenses: Experimental Analysis of Real DRAM Chips and Implications on Future Solutions

Abdullah Giray Yağlıkçı Geraldo F. Oliveira Yahya Can Tuğrul
İsmail Emir Yüksel Ataberk Olgun Haocong Luo Onur Mutlu
ETH Zürich

Svärd: Spatial Variation-Aware Read Disturbance Defenses

Goal:
Reduce the performance overhead of existing read disturbance solutions

Key Idea:
To leverage the **spatial variation of DRAM read disturbance across DRAM rows**

Svärd: Spatial Variation-Aware Read Disturbance Defenses
- **Dynamically tunes** a solution's aggressiveness (e.g., perform more/less refresh) to the **victim row's vulnerability** to DRAM read disturbance
- Implemented either in **the memory controller** or in **the DRAM chip**

Evaluation:
- Showcase on **five state-of-the-art read disturbance defenses**
- **Reduces** existing read disturbance solutions' **performance overheads**
- **Significantly improves** system performance (e.g., >4.8x)

Outline

- Lack of Memory Isolation
- Infrastructure
- Understanding Read Disturbance on Real DRAM Chips
- Efficient and Scalable DRAM Read Disturbance Solutions
  - Spatial Variation-Aware Read Disturbance Defenses
  - All-Bank Activation Counters for Scalable Mitigations
  - Count-Min-Sketch-Based Row Tracking
  - Hidden Row Activation for Reducing Refresh Latency
- Data Movement Bottleneck
- Addressing Data Movement Bottleneck on Off-the-Shelf DRAM Chips
- Current and Future Challenges

ABACuS: All-Bank Activation Counters

- Ataberk Olgun, Yahya Can Tugrul, Nisa Bostanci, Ismail Emir Yuksel, Haocong Luo, Steve Rhyner, Abdullah Giray Yaglikci, Geraldo F. Oliveira, and Onur Mutlu,
  "ABACuS: All-Bank Activation Counters for Scalable and Low Overhead RowHammer Mitigation"
  *To appear in Proceedings of the 33rd USENIX Security Symposium (**USENIX Security**)*, Philadelphia, PA, USA, August 2024.
  [arXiv version]
  [ABACuS Source Code]

ABACuS: All-Bank Activation Counters for Scalable and Low Overhead RowHammer Mitigation

Ataberk Olgun Yahya Can Tugrul Nisa Bostanci Ismail Emir Yuksel
Haocong Luo Steve Rhyner Abdullah Giray Yaglikci Geraldo F. Oliveira Onur Mutlu

ETH Zurich

ABACuS: All-Bank Activation Counters

Goal: Prevent RowHammer bitflips at low performance, energy, and area cost

Key Observation: Workloads tend to access the same row in all DRAM banks at around the same time

Key Idea: Use one hardware counter to keep track of activation counts of the same row across all banks

- Make high-performance, area-hungry counter-based mechanisms practical

Key Results:

Faster than the lowest-area-cost counter-based defense mechanism
Smaller than the lowest-performance-overhead counter-based defense mechanism

0.59% avg. performance overhead (single-core) at a RowHammer threshold (1K)

- Only 9.79 KiB on-chip storage per DRAM rank (0.02% of a Xeon processor)

1.52% avg. performance overhead (single-core) at an ultra-low threshold (125)

- 75.70 KiB on-chip storage per DRAM rank (0.11% of the Xeon processor)

https://github.com/CMU-SAFARI/ABACuS

SAFARI

Outline

Lack of Memory Isolation

Infrastructure

Understanding Read Disturbance on Real DRAM Chips

Efficient and Scalable DRAM Read Disturbance Solutions

- Spatial Variation-Aware Read Disturbance Defenses
- All-Bank Activation Counters for Scalable Mitigations
- Count-Min-Sketch-Based Row Tracking
- Hidden Row Activation for Reducing Refresh Latency

Data Movement Bottleneck

Addressing Data Movement Bottleneck on Off-the-Shelf DRAM Chips

Current and Future Challenges

CoMeT: Count-Min-Sketch-based Row Tracking

- Nisa Bostanci, Ismail Emir Yuksel, Ataberk Olgun, Konstantinos Kanellopoulos, Yahya Can Tugrul, A. Giray Yaglikci, Mohammad Sadrosadati, Onur Mutlu
  "CoMeT: Count-Min-Sketch-based Row Tracking to Mitigate RowHammer at Low Cost,"
  *in Proceedings the 30th International Symposium on High-Performance Computer Architecture (**HPCA**), Edinburgh, March 2024.*
  [arXiv version]
  [CoMeT Source Code]

CoMeT: Count-Min-Sketch-based Row Tracking to Mitigate RowHammer at Low Cost

F. Nisa Bostancı İsmail Emir Yüksel Ataberk Olgun Konstantinos Kanellopoulos
Yahya Can Tuğrul A. Giray Yağlıkçı Mohammad Sadrosadati Onur Mutlu
ETH Zürich

Executive Summary

Goal: Prevent RowHammer bitflips with low area, performance, and energy overheads in highly RowHammer-vulnerable DRAM-based systems

Key Idea: Use low-cost and scalable hash-based counters to accurately track DRAM rows

CoMeT:
- tracks most DRAM rows with scalable hash-based counters by employing the Count-Min-Sketch technique to achieve a low area cost
- tracks only a small set of DRAM rows that are activated many times with highly accurate per-DRAM-row activation counters to reduce performance penalties

Evaluation: CoMeT achieves a good trade-off between area, performance and energy costs
- incurs significantly less area overhead (74.2×) compared to the state-of-the-art technique
- outperforms the state-of-the-art technique (by up to 39.1%)

https://github.com/CMU-SAFARI/CoMeT

Outline

Lack of Memory Isolation

Infrastructure

Understanding Read Disturbance on Real DRAM Chips

Efficient and Scalable DRAM Read Disturbance Solutions

- Spatial Variation-Aware Read Disturbance Defenses
- All-Bank Activation Counters for Scalable Mitigations
- Count-Min-Sketch-Based Row Tracking
- Hidden Row Activation for Reducing Refresh Latency

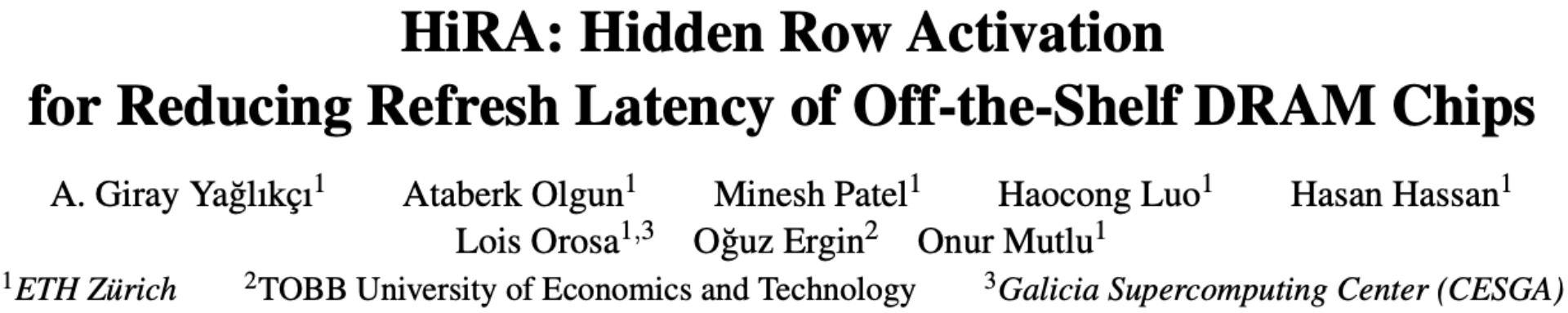
Data Movement Bottleneck

Addressing Data Movement Bottleneck on Off-the-Shelf DRAM Chips

Current and Future Challenges

HiRA: Hidden Row Activation

- **Abdullah Giray Yağlıkçı,** Ataberk Olgun, Minesh Patel, Haocong Luo, Hasan Hassan, Lois Orosa, Oguz Ergin, and Onur Mutlu, **"HiRA: Hidden Row Activation for Reducing Refresh Latency of Off-the-Shelf DRAM Chips,"** in ***MICRO*** 2022.
  [Slides (pptx) (pdf)]
  [Longer Lecture Slides (pptx) (pdf)]
  [Lecture Video (36 minutes)]
  [arXiv version]



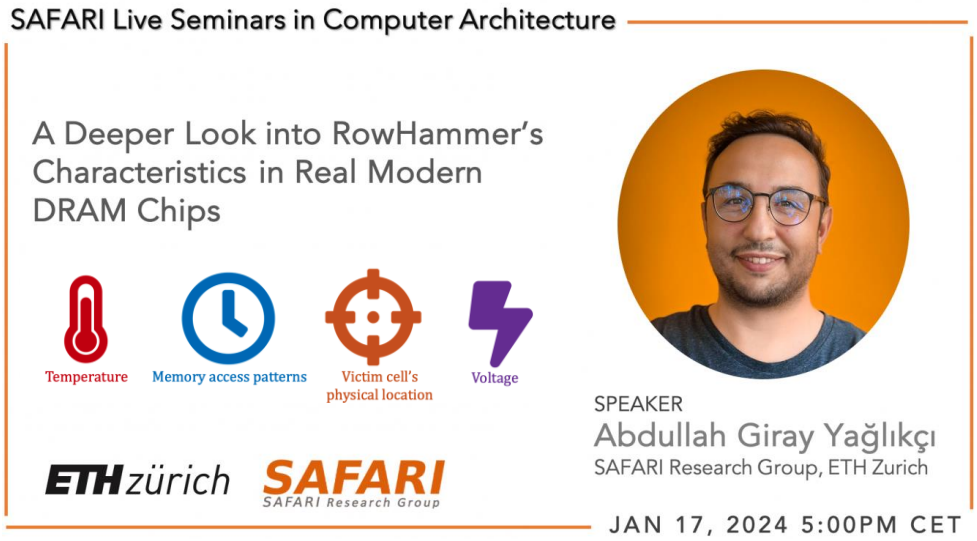
HiRA: Hidden Row Activation
for Reducing Refresh Latency of Off-the-Shelf DRAM Chips

A. Giray Yağlıkçı[1] Ataberk Olgun[1] Minesh Patel[1] Haocong Luo[1] Hasan Hassan[1]
Lois Orosa[1,3] Oğuz Ergin[2] Onur Mutlu[1]
[1]*ETH Zürich* [2]TOBB University of Economics and Technology [3]*Galicia Supercomputing Center (CESGA)*

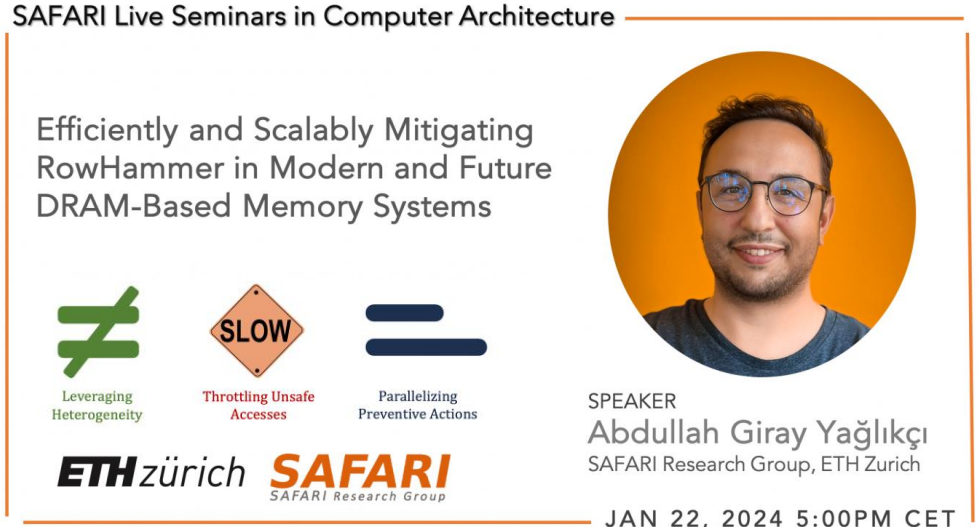
HiRA: Hidden Row Activation

- **Problem:** DRAM Refresh
  - is a **fundamental operation** to avoid bit flips due to **leakage** and **RowHammer**
  - incurs **increasingly large performance overhead** with DRAM chip **density scaling**
- **Goal:** Reduce the **performance overhead** of DRAM Refresh
- **Key Idea: Hide refresh latency** by **refreshing** a DRAM row *concurrently with* **activating** another row in a **different subarray** of the **same bank**
- **HiRA:** Hidden Row Activation – a new DRAM operation that
  - Issues **DRAM commands** in **quick succession** to concurrently open two rows in **different subarrays**
  - Works on **real off-the-shelf DRAM chips** by violating timing constraints
  - **Significantly reduces** (51.4%) the time spent for refresh operations
- **HiRA-MC:** HiRA Memory Controller – a new mechanism
  - **Leverages HiRA** to perform **refresh requests** *concurrently with* **DRAM accesses** and **other refresh requests**
  - **Significantly improves** system performance by **hiding refresh latency** for both **regular periodic** and **RowHammer-preventive** refreshes

More Details and Discussion on YouTube



https://www.youtube.com/live/CRtm1es4n3o?si=8N5zB6e_RUc5Ejl8



https://www.youtube.com/live/YQwRYWpCsk0?si=jXPueMHb5wgs69-q

Outline

Lack of Memory Isolation

Infrastructure

Understanding Read Disturbance on Real DRAM Chips

Efficient and Scalable DRAM Read Disturbance Solutions

Data Movement Bottleneck

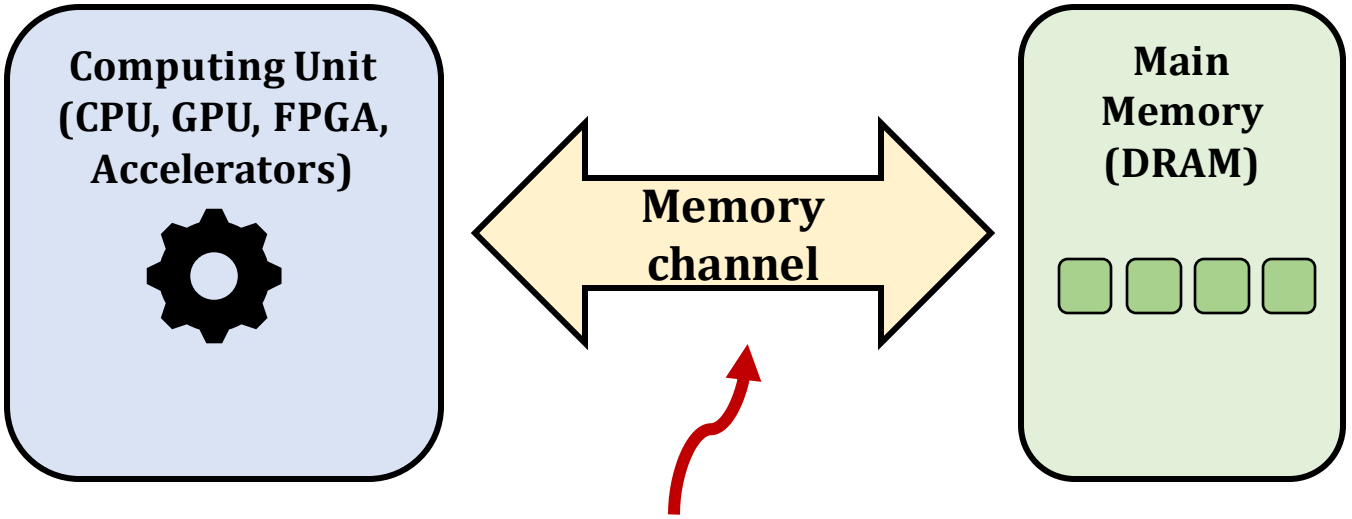
Addressing Data Movement Bottleneck on Off-the-Shelf DRAM Chips

Current and Future Challenges

Data Movement Bottleneck

- Data movement is a major bottleneck

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

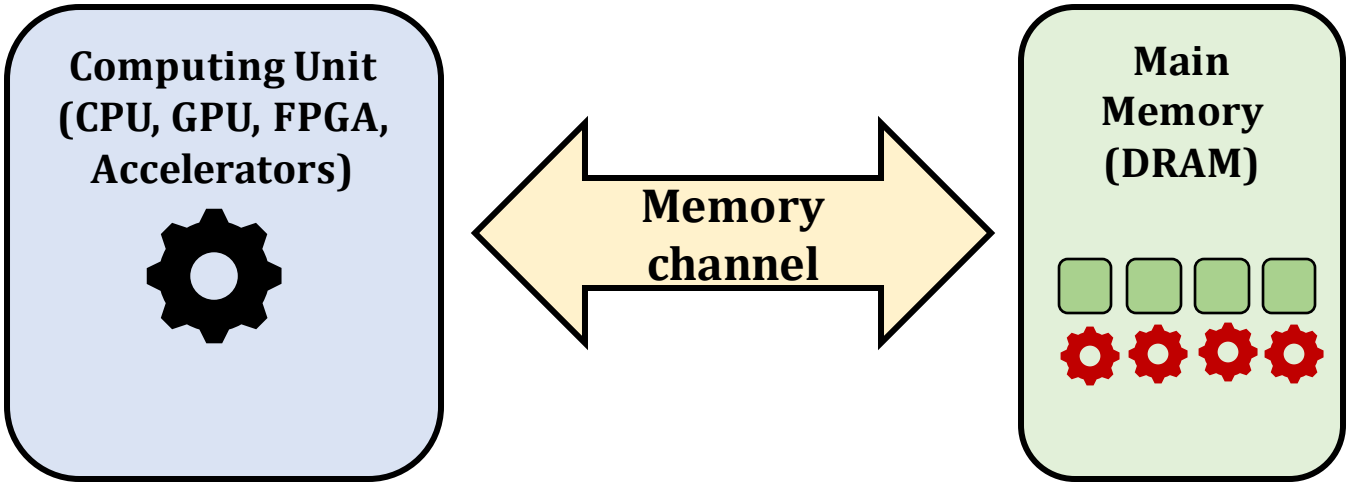


Bandwidth-limited and power-hungry memory channel

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

Processing-in-Memory (PIM)

- **Processing-in-Memory:** moves computation closer to where the data resides
  - Reduces/eliminates the need to move data between processor and DRAM



Processing-using-Memory (PuM)

- PuM: Exploits analog operation principles of the memory circuitry to perform computation
  - Leverages the large internal bandwidth and parallelism available inside the memory arrays
- A common approach for PuM architectures is to perform bulk bitwise operations
  - Simple logical operations (e.g., AND, OR, XOR)
  - More complex operations (e.g., addition, multiplication)

Limitations of *Processing using Memory (PuM)*

Existing PuM mechanisms are
not widely applicable

Support only a limited set of operations

Require significant changes
to the DRAM subarray

Goal:

Functionally-complete and reliable
PuM on off-the-shelf DRAM chips

Outline

Lack of Memory Isolation

Infrastructure

Understanding Read Disturbance on Real DRAM Chips

Efficient and Scalable DRAM Read Disturbance Solutions

Data Movement Bottleneck

Addressing Data Movement Bottleneck on Off-the-Shelf DRAM Chips

Current and Future Challenges

Outline

Lack of Memory Isolation

Infrastructure

Understanding Read Disturbance on Real DRAM Chips

Efficient and Scalable DRAM Read Disturbance Solutions

Data Movement Bottleneck

Addressing Data Movement Bottleneck on Off-the-Shelf DRAM Chips

- Functionally-Complete Boolean Logic in Real DRAM Chips
- Simultaneous Many-Row Activation in Off-the-Shelf DRAM Chips

Current and Future Challenges

Functionally-Complete Boolean Logic in Real DRAM Chips

- Ismail Emir Yuksel, Yahya Can Tugrul, Ataberk Olgun, Nisa Bostanci, A. Giray Yaglikci, Geraldo F. Oliveira, Haocong Luo, Juan Gomez Luna, Mohammad Sadrosadati, and Onur Mutlu,
  "Functionally-Complete Boolean Logic in Real DRAM Chips: Experimental Characterization and Analysis"
  Proceedings of the 30th International Symposium on High-Performance Computer Architecture ***(HPCA)***, *Edinburgh, March 2024.*
  [arXiv version]

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

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

Functionally-Complete Boolean Logic in Real DRAM Chips

- **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 (>94%) success rate**
  - **Data pattern** and **temperature** only slightly affect the reliability of these operations **(<1.98% decrease in success rate)**

Outline

Lack of Memory Isolation

Infrastructure

Understanding Read Disturbance on Real DRAM Chips

Efficient and Scalable DRAM Read Disturbance Solutions

Data Movement Bottleneck

Addressing Data Movement Bottleneck on Off-the-Shelf DRAM Chips

- Functionally-Complete Boolean Logic in Real DRAM Chips
- Simultaneous Many-Row Activation in Off-the-Shelf DRAM Chips

Current and Future Challenges

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

- Ismail Emir Yuksel, Yahya Can Tugrul, F. Nisa Bostanci, Abdullah Giray Yaglikci, Ataberk Olgun, Geraldo F. Oliveira, Melina Soysal, Haocong Luo, Juan Gomez Luna, Mohammad Sadrosadati, Onur Mutlu,
  "PULSAR: Simultaneous Many-Row Activation for Reliable and High-Performance Computing in Off-the-Shelf DRAM Chips"
  in Proceedings of the 54th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), Brisbane, Australia, June 24-27, 2024
  [arXiv version]

PULSAR: Simultaneous Many-Row Activation for Reliable and High-Performance Computing in Off-the-Shelf DRAM Chips

Ismail Emir Yuksel Yahya Can Tugrul F. Nisa Bostanci Abdullah Giray Yaglikci Ataberk Olgun
Geraldo F. Oliveira Melina Soysal Haocong Luo Juan Gomez Luna Mohammad Sadrosadati
Onur Mutlu

ETH Zurich

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

- **Goal:** Improve PuD **reliability** and **performance**
- **Research questions:**
  - Can off-the-shelf DRAM chips reliably activate many rows?
  - What other PuD operations can be realized?
- **Experimental Study: Extensive** characterization on **120 DDR4 chips** under varying **timing delays, data patterns, temperature,** and **voltages**
- **Key Results:** Off-the-shelf DRAM chips can
  - Activate up to **32 DRAM rows**
  - Execute **MAJ5**, **MAJ7**, and **MAJ9** operations
  - Enables writing data into DRAM rows in bulk

Significantly improves **PuD reliability** by **replicating input** across many rows

Provides the user with **MAJ5**, **MAJ7**, and **MAJ9** operations for better **performance**

Thanks for your support

DRAM Bender: An Extensible and Versatile FPGA-based Infrastructure to Easily Test State-of-the-art DRAM Chips

TCAD'23

RowPress: Amplifying Read-Disturbance in Modern DRAM Chips

ISCA'23

A Deeper Look into RowHammer's Sensitivities: Experimental Analysis of Real DRAM Chips and Implications on Future Attacks and Defenses

MICRO'21

HiRA: Hidden Row Activation for Reducing Refresh Latency of Off-the-Shelf DRAM

MICRO'22

Understanding RowHammer Under Reduced Wordline Voltage: An Experimental Study Using Real DRAM Devices

DSN'22

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

HPCA'24

Spatial Variation-Aware Read Disturbance Defenses: Experimental Analysis of Real DRAM Chips and Implications on Future Solutions

HPCA'24

CoMeT: Count-Min-Sketch-based Row Tracking to Mitigate RowHammer at Low Cost

HPCA'24

PULSAR: Simultaneous Many-Row Activation for Reliable and High-Performance Computing in Off-the-Shelf DRAM Chips

DSN'24

An Experimental Analysis of RowHammer in HBM2 DRAM Chips

DSN Disrupt'23 DSN'24

ABACuS: All-Bank Activation Counters for Scalable and Low Overhead RowHammer Mitigation

USENIX Security'24

Thanks for your support

DRAM Bender: An Extensible and Versatile FPGA-based Infrastructure to Easily Test State-of-the-art DRAM Chips

Thanks for your support!

MICRO'22

HPCA'24

Functionally-Complete Boolean Logic in Real DRAM Chips:

There still are outstanding research problems and more things to do

DSN Disrupt'23

DSN'24

Outline

Lack of Memory Isolation

Infrastructure

Understanding Read Disturbance on Real DRAM Chips

Efficient and Scalable DRAM Read Disturbance Solutions

Data Movement Bottleneck

Addressing Data Movement Bottleneck on Off-the-Shelf DRAM Chips

Current and Future Challenges

Current and Future Challenges

Reliability

Performance

Fairness

Energy Efficiency

While the memory systems

1. **scale-up and are shared** across many users (e.g., disaggregated memory systems)

2. **scale-down** in manufacturing technology node size

3. support **processing near/using memory**

Future Research for Better Memory Systems

Deeper Understanding of
Physics and Vulnerabilities

Flexible and Intelligent Memory
Chips, Interfaces, Controllers

Cross-Layer
Communication

Fundamentally Understanding DRAM Reliability & Enabling Fast and Secure Memory

Abdullah Giray Yaglikci

agyaglikci@gmail.com

https://agyaglikci.github.io

9 April 2024

Microsoft Swiss Joint Research Center

agyaglikci.github.io

safari.ethz.ch

SAFARI

ETH zürich

Fundamentally Understanding DRAM Reliability & Enabling Fast and Secure Memory

Future Research

Abdullah Giray Yaglikci

agyaglikci@gmail.com

https://agyaglikci.github.io

9 April 2024

Microsoft Swiss Joint Research Center

agyaglikci.github.io

safari.ethz.ch

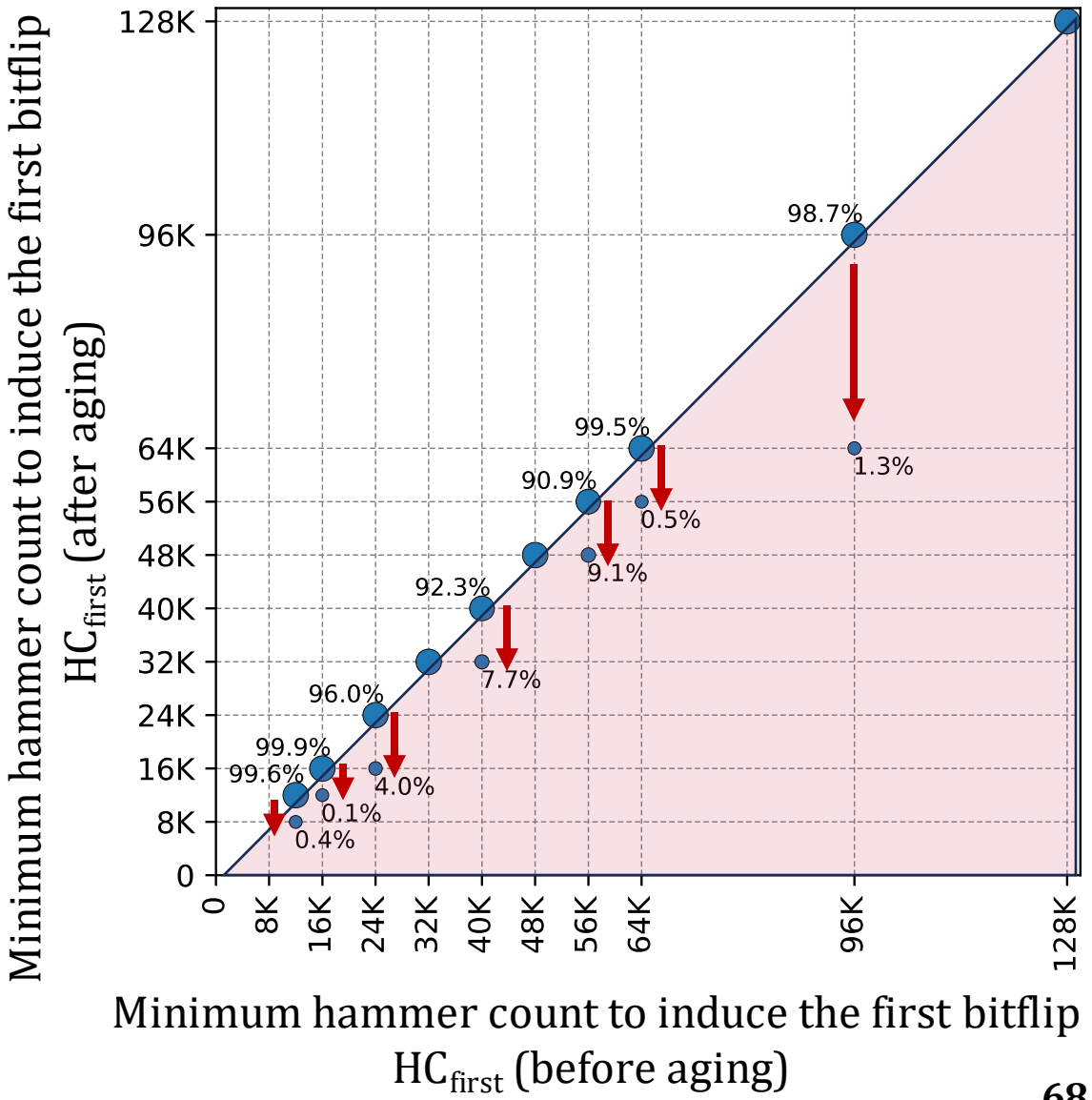
SAFARI

ETH zürich

Deeper Understanding of Physics and Vulnerabilities

- The effect of **aging**
  Preliminary data on aging via 68-day of continuous hammering

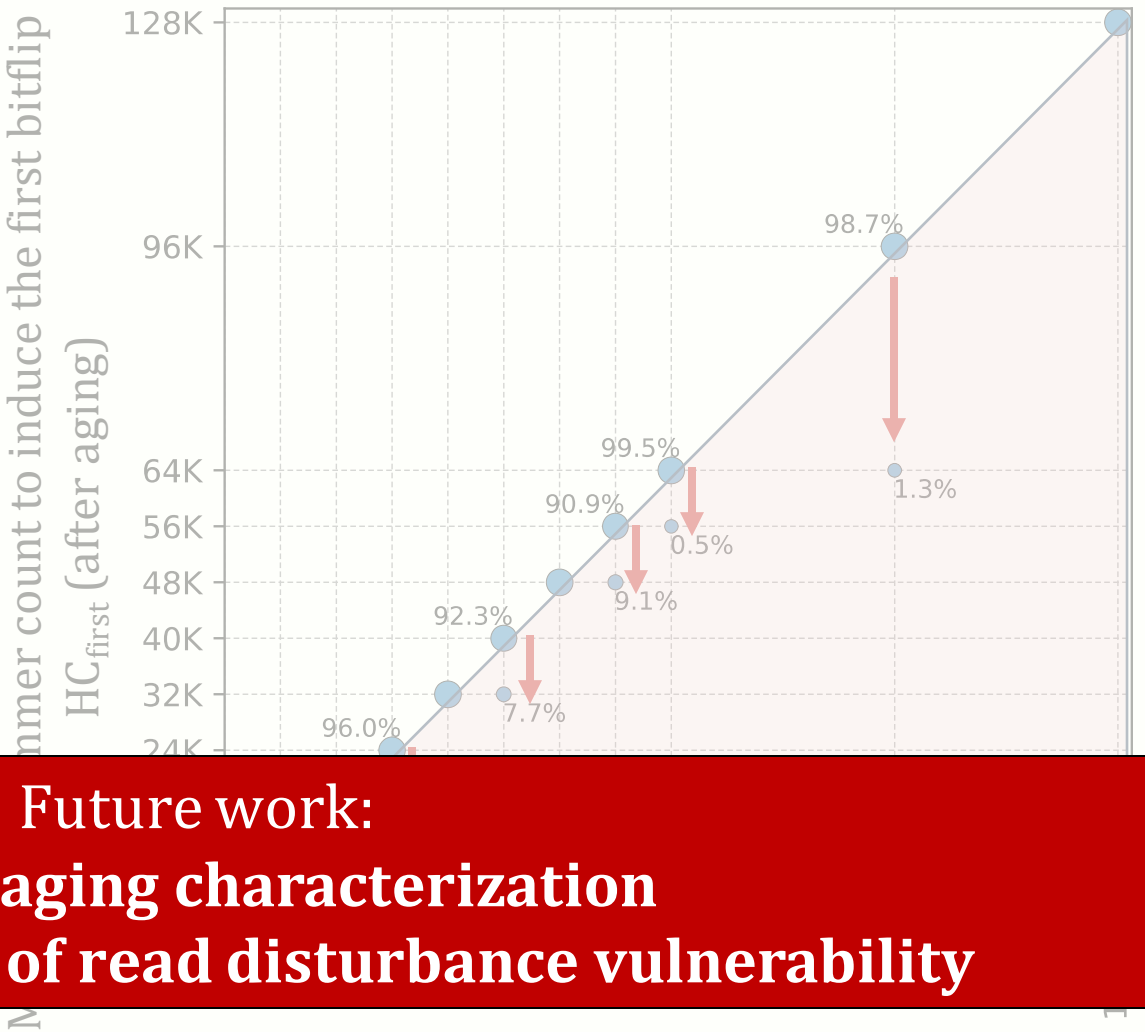
Aging can lead to read disturbance bitflips at **smaller** hammer counts



Deeper Understanding of Physics and Vulnerabilities

- The effect of **aging**
  Preliminary data on aging via 68-day of continuous hammering

Aging can lead to read disturbance bitflips at **smaller** hammer counts



Future work:
rigorous aging characterization
and **online profiling of read disturbance vulnerability**

Deeper Understanding of Physics and Vulnerabilities

- The effect of **aging**

- **Interactions** across different error mechanisms
  - RowHammer
  - RowPress
  - Data retention time errors
  - Variable retention time
  - ...

Deeper Understanding of Physics and Vulnerabilities

- The effect of **aging**
- **Interactions** across different error mechanisms
- What is **the worst-case**?
  - Temperature
  - Data pattern
  - Memory access pattern
  - Spatial variation
  - Voltage

What is **the worst-case** considering all **these sensitivities**?

What is **the minimum hammer count** to induce a read disturbance bitflip?

Deeper Understanding of Physics and Vulnerabilities

- The effect of **aging**
- **Interactions** across different error mechanisms
- What is **the worst-case**?

How reliable are our DRAM chips?

How reliable will our DRAM chips be tomorrow?

We **do not** know! This is an **open research problem**

Future Research for Better Memory Systems

Deeper Understanding of
Physics and Vulnerabilities

Flexible and Intelligent Memory
Chips, Interfaces, Controllers

Cross-Layer
Communication

Flexible and Intelligent Chips, Interfaces, Controllers

- **In-field patching** is necessary

DRAM read disturbance solution is here

Configuration

DRAM chip is here

Reconfiguration is needed

+ add more hardware counters
+ increase metadata cache capacity

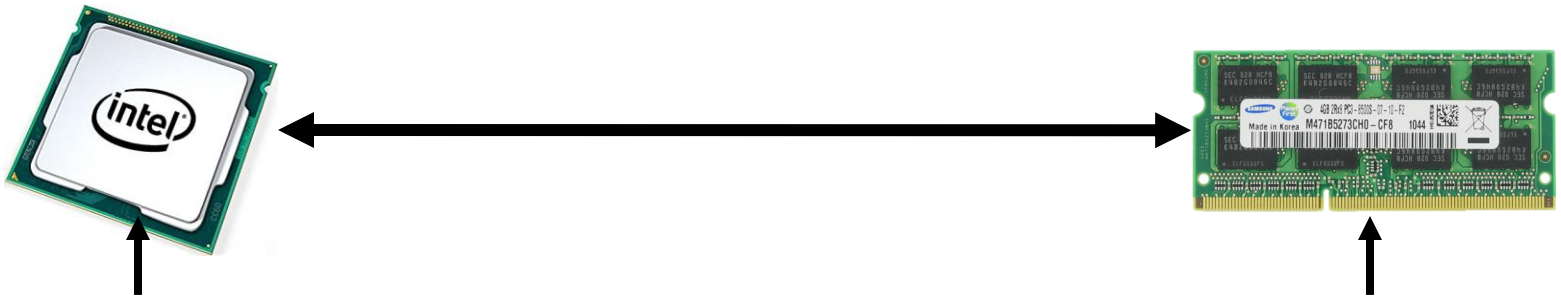
DRAM module can be replaced

Aging might change the chip's characteristics

Deployed solutions should be patchable in field

Flexible and Intelligent Chips, Interfaces, Controllers

- **In-field patching** is necessary
- Interfaces should be **more flexible**



Memory controller decides what should be done when

DRAM chip has read disturbance solution inside (tracking+prevention)

- The memory controller should provide the DRAM chip with **necessary time window** to perform **preventive actions (e.g., refreshing rows)**
- The memory controller **does not have** the tracking information
- Communicating is **not straightforward** due to strict communication protocol

A more flexible interface is necessary

Flexible and Intelligent Chips, Interfaces, Controllers

- **In-field patching** is necessary
- Interfaces should be **more flexible**
- Memory controllers should be **more intelligent** in detecting **malicious activity**

- DRAM chips become **more and more vulnerable** to RowHammer **and RowPress**
- **Key Insight:**
  - **A thousand activations** are enough **to induce bitflips**
  - Benign applications **perform as many activations**
- **Problem:** DRAM read disturbance solutions are getting **prohibitively expensive**
- **Research Question:** How to identify malicious threads/processes/users?
- More **intelligent detection** mechanisms are needed → AI can play an important role
- The **memory controller** observes all memory accesses → has the ground truth data

More intelligent memory controllers can help

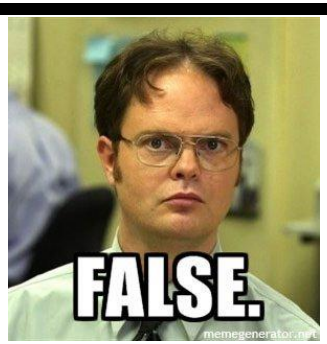
Future Research for Better Memory Systems

Deeper Understanding of
Physics and Vulnerabilities

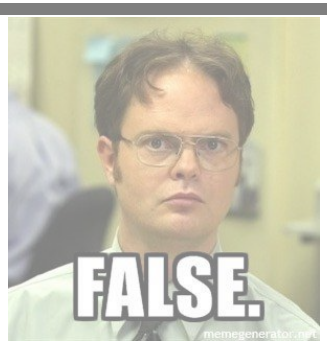
Flexible and Intelligent Memory
Chips, Interfaces, Controllers

Cross-Layer
Communication

Cross-Layer Communication

		Detection	Mitigation
Software	 • Memory allocations • Memory access patterns • Control flow patterns • Time / power measurements		++
uArch	 • Memory request scheduling • Speculative execution • Prefetching, branch prediction • Power management	+	+
Device	 • Bitflips occur • Memory isolation is broken	++	~

Cross-Layer Communication

		Detection	Mitigation
Software	• Memory allocations • Memory access patterns • Control flow patterns • Time / power measurements		++
Device	• Bitflips occur • Memory isolation is broken	++	~

Cross-layer communication is crucial going forward

Future Research for Better Memory Systems

Deeper Understanding of
Physics and Vulnerabilities

Flexible and Intelligent Memory
Chips, Interfaces, Controllers

Cross-Layer
Communication

Fundamentally Understanding DRAM Reliability & Enabling Fast and Secure Memory

Backup Slides

agyaglikci.github.io

Abdullah Giray Yaglikci

agyaglikci@gmail.com

https://agyaglikci.github.io

9 April 2024

Microsoft Swiss Joint Research Center

safari.ethz.ch

SAFARI

ETH zürich

My Dissertation Works

Deeper understanding of RowHammer	Efficiently and scalably mitigating RowHammer
Chapter 4 MICRO'21: Temperature	Chapter 7 HPCA'21: Throttling unsafe accesses
Chapter 4 MICRO'21: Memory access patterns; Victim cell's physical location	Chapter 6 HPCA'24: Leveraging heterogeneity
Chapter 5 DSN'22: Voltage	Chapter 8 MICRO'22: Parallelizing preventive actions

DRAM Testing Methodology

To characterize our DRAM chips at **worst-case** conditions:

1. Prevent sources of interference during core test loop

- **No DRAM refresh**: to avoid refreshing victim row
- **No DRAM calibration events**: to minimize variation in test timing
- **No RowHammer mitigation mechanisms**: to observe circuit-level effects
- Test for **less than a refresh window (32ms)** to avoid retention failures
- **Repeat tests** for ten times

2. Worst-case access sequence

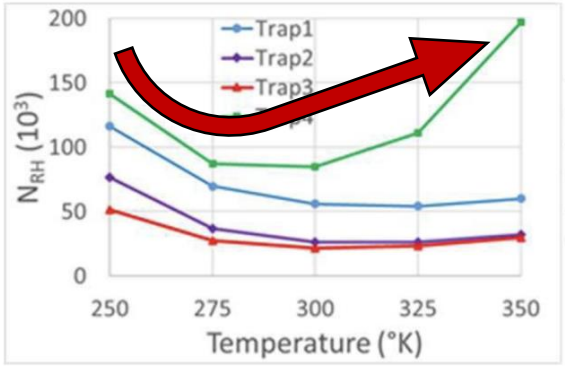
- We use **worst-case** access sequence based on prior works' observations
- For each row, **repeatedly access the two physically-adjacent rows as fast as possible**

Circuit-Level Justification

Trap-Assisted Charge Leakage Model

- Hammering a wordline **pulls and pushes electrons**
- Electrons **get trapped** and **exacerbate charge leakage**, leading to cause bitflips
- With **increasing temperature**, it becomes **less likely for an electron to get trapped**

3D TCAD Evaluation [Yang+, EDL'19]



	Row	Column	Energy
Trap1	1	4	0.50eV
Trap2	2	3	0.45eV
Trap3	3	3	0.45eV
Trap4	4	3	0.45eV

Fig. 6. Hammering threshold N_{RH} vs. temperature from 250 to 350°K for different traps. Location in row and column refers to matrix in Fig. 2b.

Until a temperature inflection point:

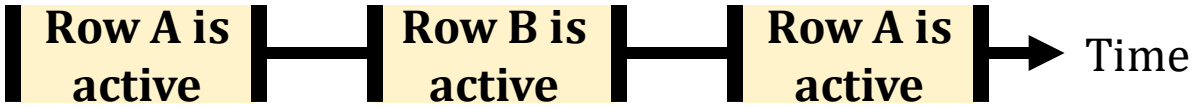
As temperature increases, **fewer activations** can cause bitflips

After the temperature inflection point:

As temperature increases, **more activations** are needed to cause bitflips

Example Attack Improvement: Bypassing Defenses with Aggressor Row Active Time

Activating aggressor rows **as frequently as possible**:



Keeping the aggressor rows **active for a longer time**:

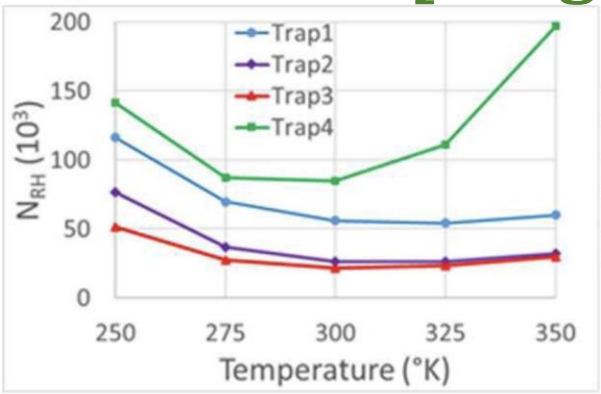
Reduces the minimum activation count to induce a bitflip **by 36%**

Bypasses defenses that do not account for this reduction

Circuit-Level Justification

We hypothesize that our observations are caused by the **non-monotonic behavior of charge trapping** characteristics of DRAM cells

3D TCAD model [Yang+, EDL'19]



	Row	Column	Energy
Trap1	1	4	0.50eV
Trap2	2	3	0.45eV
Trap3	3	3	0.45eV
Trap4	4	3	0.45eV

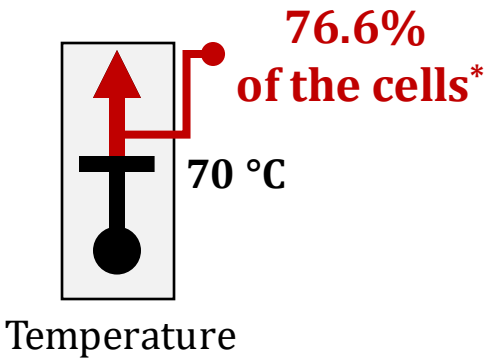
Fig. 6. Hammering threshold N_{RH} vs. temperature from 250 to 350°K for different traps. Location in row and column refers to matrix in Fig. 2b.

***HC_{first}* decreases as temperature increases**, until a temperature inflection point where ***HC_{first}* starts to increase as temperature increases**

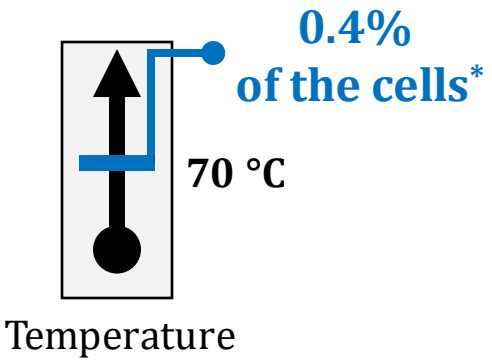
A **cell is more vulnerable** to RowHammer at **temperatures close to its temperature inflection point**

Example Attack Improvement: Temperature-Dependent Trigger

1. Identify **abnormal increase** in temperature to attack a data center **during its peak hours**



2. **Precisely measure** the temperature **to trigger an attack** exactly at the desired temperature

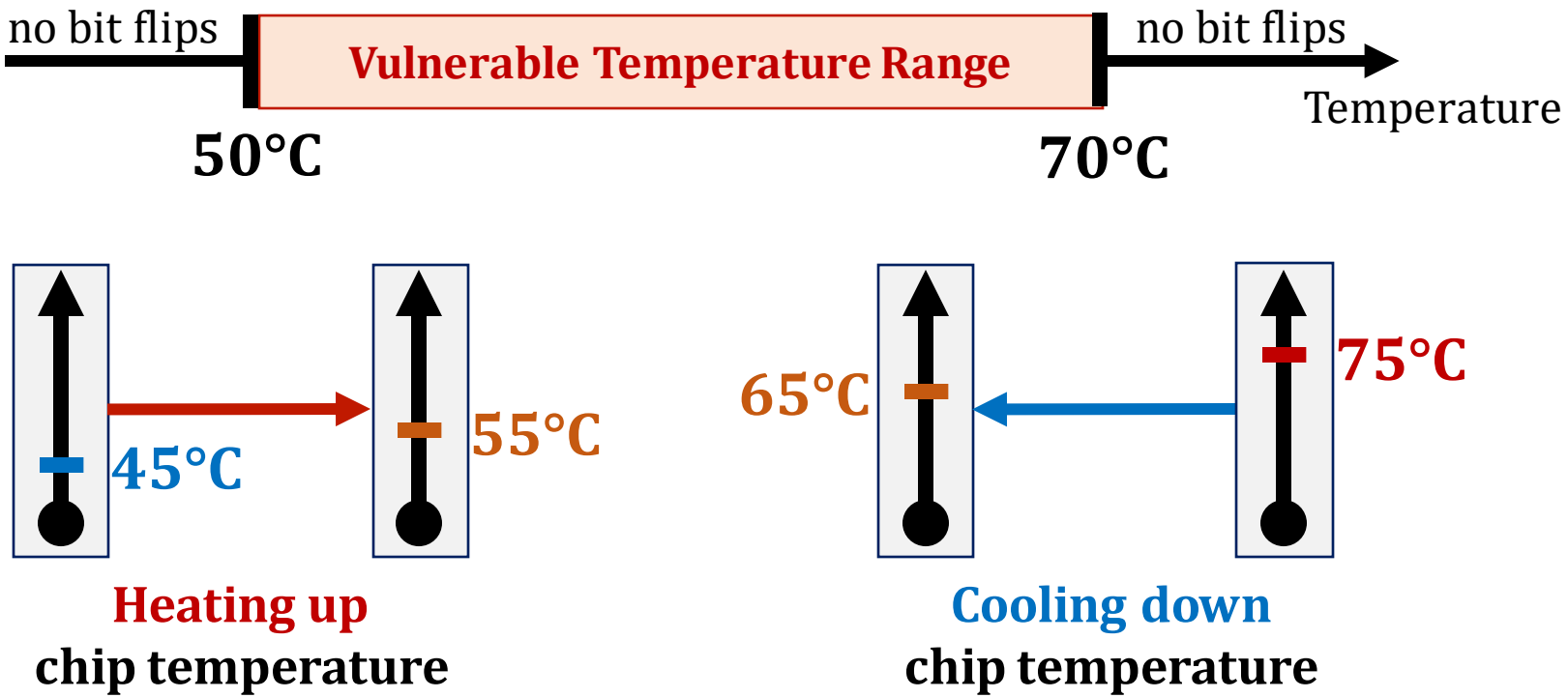


*Example fraction values from SK Hynix modules

Attack Improvement 1:
Making DRAM Cells More Vulnerable

An attacker can **manipulate temperature** to make the cells that store sensitive data **more vulnerable**

DRAM cells are vulnerable in a **bounded temperature range**



How Large is 1000 Activations?

- Bitflips occur **at ~1000 activations**
- Mitigation mechanisms trigger **preventive actions** (e.g., preventive refresh) **at ~500 activations**
- Is an **activation count of 500 common** or **rare**?
- Benign workloads activate **hundreds of rows** more than **512 times** in a refresh window

Memory intensive workloads
from SPEC'06/17, TPC, YCSB, and MediaBench

Benchmark	MPKI	# of Rows w/ ACT count >512
429.mcf	68.27	2564
470.lbm	28.09	664
519.lbm	24.37	2482
434.zeusmp	22.24	292
510.parest	17.79	94
437.leslie3d	15.82	7
483.xalancbmk	13.67	113
482.sphinx3	12.59	304
505.mcf	11.35	732
471.omnetpp	10.72	122
tpch2	9.09	88
520.omnetpp	9.00	32
tpch17	7.43	26

Sorted ↓

* Right-most column shows the total number of rows across all banks

How Large is 1000 Activations?

- Bitflips occur **at ~1000 activations**
- Mitigation mechanisms trigger **preventive actions** (e.g., preventive refresh) **at ~500 activations**

more than **512 times** in a refresh window

Memory intensive workloads
from SPEC'06/17, TPC, YCSB, and MediaBench

Benchmark	MPKI	# of Rows w/ ACT count >512
429.mcf	68.27	2564
470.lbm	28.09	664
519.lbm	24.37	2482
434.zeusmp	22.24	292
tpch2	9.09	88
520.omnetpp	9.00	32
tpch17	7.43	26

* Right-most column shows the total number of rows across all banks

Benign workloads **might not be so benign** even if they are **not very memory intensive**

Thesis Statement

Developing

a deeper understanding of DRAM read disturbance

and

revisiting memory controller designs

enable scientists and engineers to build

reliable, secure, and safe DRAM-based systems

My Dissertation Works

- A deeper look into DRAM read disturbance

Temperature

Memory Access Patterns

In-Chip Variations

Voltage

- Solutions to DRAM read disturbance

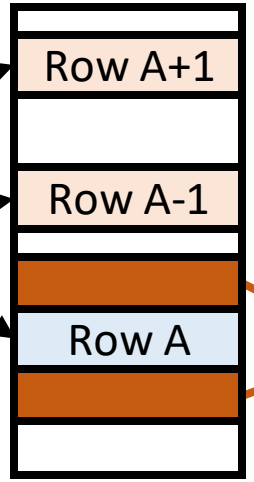
Throttling Unsafe Accesses

Parallelizing Preventive Measures

Leveraging Heterogeneity

Compatibility with Commodity DRAM Chips

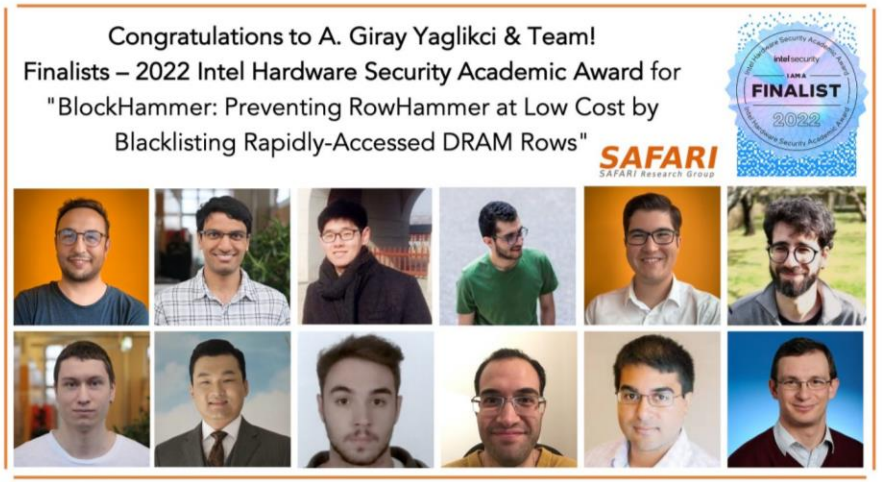
- A **RowHammer attack** hammers Row A
- Existing mechanisms **detect the attack**
- Refresh rows **A+1** and **A-1**
- Bit flips **still may occur** due to **unknown DRAM-internal row mapping**



Existing read disturbance mitigation mechanisms need to know proprietary DRAM-internal row address mapping

BlockHammer – HPCA 2021

- A. Giray Yaglikci, Minesh Patel, Jeremie S. Kim, Roknoddin Azizi, Ataberk Olgun, Lois Orosa, Hasan Hassan, Jisung Park, Konstantinos Kanellopoulos, Taha Shahroodi, Saugata Ghose, and Onur Mutlu,
  "BlockHammer: Preventing RowHammer at Low Cost by Blacklisting Rapidly-Accessed DRAM Rows"
  Proceedings of the 27th International Symposium on High-Performance Computer Architecture (***HPCA***),
  Virtual, February-March 2021.
  [Slides (pptx) (pdf)]
  [Short Talk Slides (pptx) (pdf)]
  [Intel Hardware Security Academic Awards Short Talk Slides (pptx) (pdf)]
  [Talk Video (22 minutes)]
  [Short Talk Video (7 minutes)]
  [Intel Hardware Security Academic Awards Short Talk Video (2 minutes)]
  [BlockHammer Source Code]
  Intel Hardware Security Academic Award Finalist (one of 4 finalists out of 34 nominations)

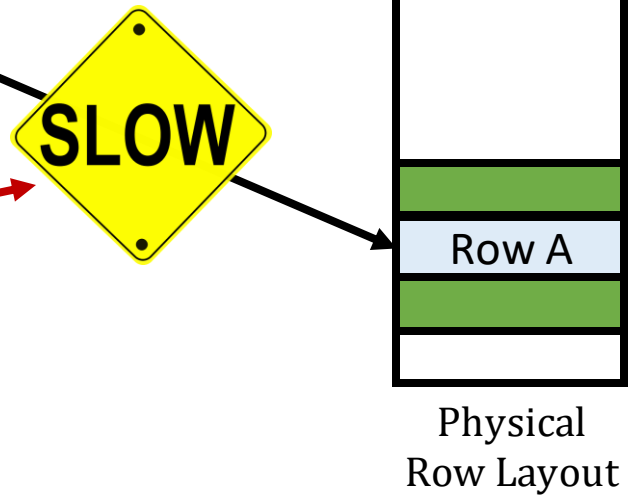


BlockHammer: Preventing RowHammer at Low Cost by Blacklisting Rapidly-Accessed DRAM Rows

A. Giray Yağlıkçı[1] Minesh Patel[1] Jeremie S. Kim[1] Roknoddin Azizi[1] Ataberk Olgun[1] Lois Orosa[1]
Hasan Hassan[1] Jisung Park[1] Konstantinos Kanellopoulos[1] Taha Shahroodi[1] Saugata Ghose[2] Onur Mutlu[1]
[1]*ETH Zürich* [2]*University of Illinois at Urbana–Champaign*

BlockHammer: Throttling Unsafe Accesses

- A RowHammer attack hammers Row A
- BlockHammer detects and **selectively throttles accesses** from within **the memory controller**
- Bit flips **do not** occur

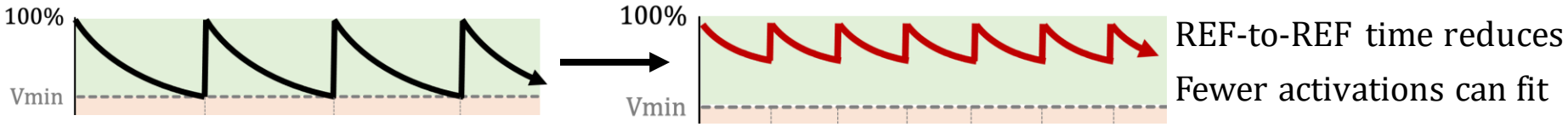


- BlockHammer can *optionally* **inform the system software** about the attack

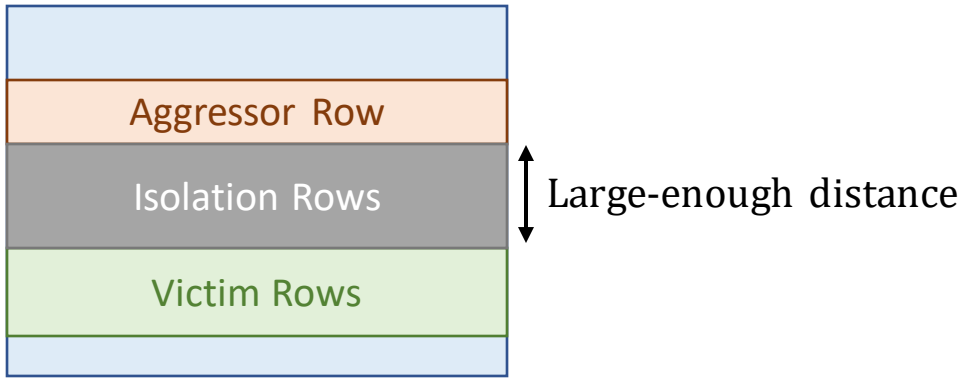
BlockHammer is compatible with commodity DRAM chips
No need for proprietary info of or modifications to DRAM chips

RowHammer Mitigation Approaches

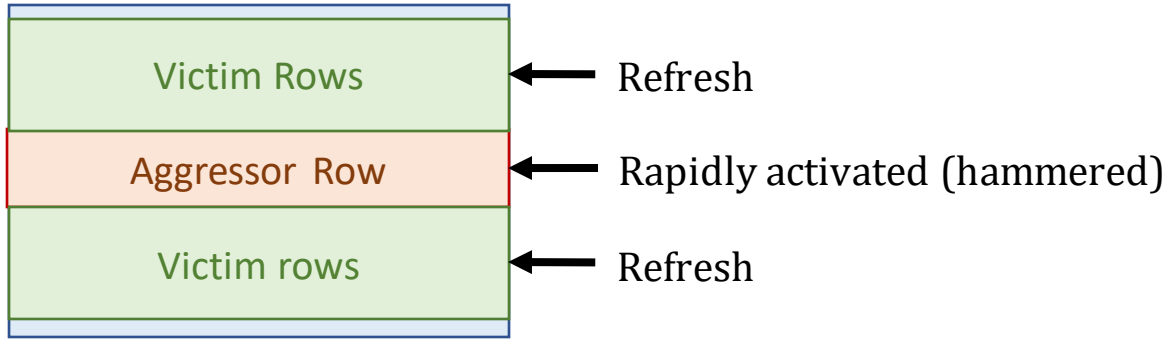
- Increased refresh rate



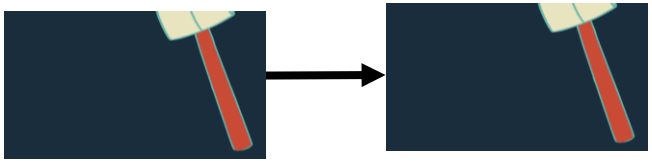
- Physical isolation



- Reactive refresh

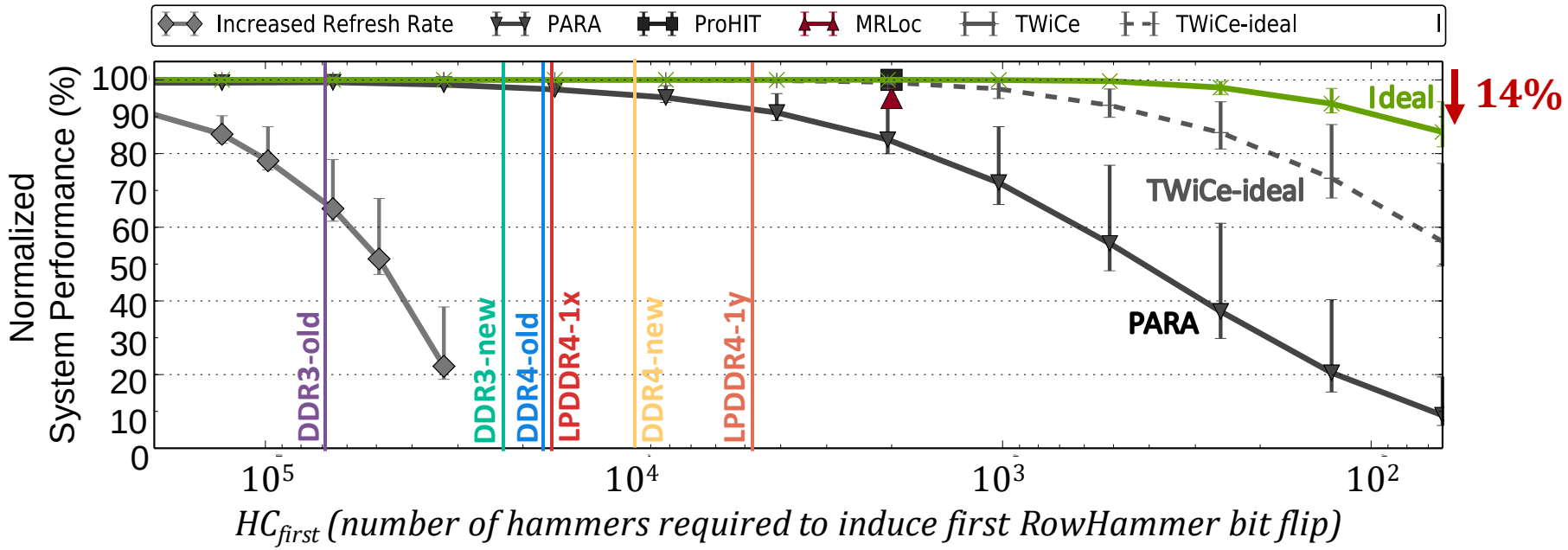


- Proactive throttling



Fewer activations can be performed

RowHammer Mitigation across Generations



J. S. Kim, M. Patel, A. G. Yaglikci, H. Hassan, R. Azizi, L. Orosa, and O. Mutlu, "Revisiting RowHammer: An Experimental Analysis of Modern Devices and Mitigation Techniques," in *ISCA*, 2020.

Fundamentally Understanding DRAM Reliability & Enabling Fast and Secure Memory

Backup Slides

agyaglikci.github.io

Abdullah Giray Yaglikci

agyaglikci@gmail.com

https://agyaglikci.github.io

9 April 2024

Microsoft Swiss Joint Research Center

safari.ethz.ch

SAFARI

ETH zürich