

Computer Architecture

Lecture 11a: Memory Latency, Energy, and Power

Prof. Onur Mutlu

ETH Zürich

Fall 2018

24 October 2018

Why the Long Memory Latency?

- Reason 1: Design of DRAM Micro-architecture
 - Goal: Maximize capacity/area, not minimize latency
- Reason 2: “One size fits all” approach to latency specification
 - Same latency parameters for all temperatures
 - Same latency parameters for all DRAM chips
 - Same latency parameters for all parts of a DRAM chip
 - Same latency parameters for all supply voltage levels
 - Same latency parameters for all application data
 - ...

Tackling the Fixed Latency Mindset

- Reliable operation latency is actually very heterogeneous
 - Across temperatures, chips, parts of a chip, voltage levels, ...
- Idea: Dynamically find out and use the lowest latency one can reliably access a memory location with
 - Adaptive-Latency DRAM [HPCA 2015]
 - Flexible-Latency DRAM [SIGMETRICS 2016]
 - Design-Induced Variation-Aware DRAM [SIGMETRICS 2017]
 - Voltron [SIGMETRICS 2017]
 - DRAM Latency PUF [HPCA 2018]
 - ...
- We would like to find sources of latency heterogeneity and exploit them to minimize latency

Analysis of Latency Variation in DRAM Chips

- Kevin Chang, Abhijith Kashyap, Hasan Hassan, Samira Khan, Kevin Hsieh, Donghyuk Lee, Saugata Ghose, Gennady Pekhimenko, Tianshi Li, and Onur Mutlu,

"Understanding Latency Variation in Modern DRAM Chips: Experimental Characterization, Analysis, and Optimization"

*Proceedings of the ACM International Conference on Measurement and Modeling of Computer Systems (**SIGMETRICS**), Antibes Juan-Les-Pins, France, June 2016.*

[[Slides \(pptx\)](#) ([pdf](#))]

[[Source Code](#)]

Understanding Latency Variation in Modern DRAM Chips: Experimental Characterization, Analysis, and Optimization

Kevin K. Chang¹

Abhijith Kashyap¹

Hasan Hassan^{1,2}

Saugata Ghose¹

Kevin Hsieh¹

Donghyuk Lee¹

Tianshi Li^{1,3}

Gennady Pekhimenko¹

Samira Khan⁴

Onur Mutlu^{5,1}

¹Carnegie Mellon University ²TOBB ETÜ ³Peking University ⁴University of Virginia ⁵ETH Zürich

Solar-DRAM: More on Spatial Variation

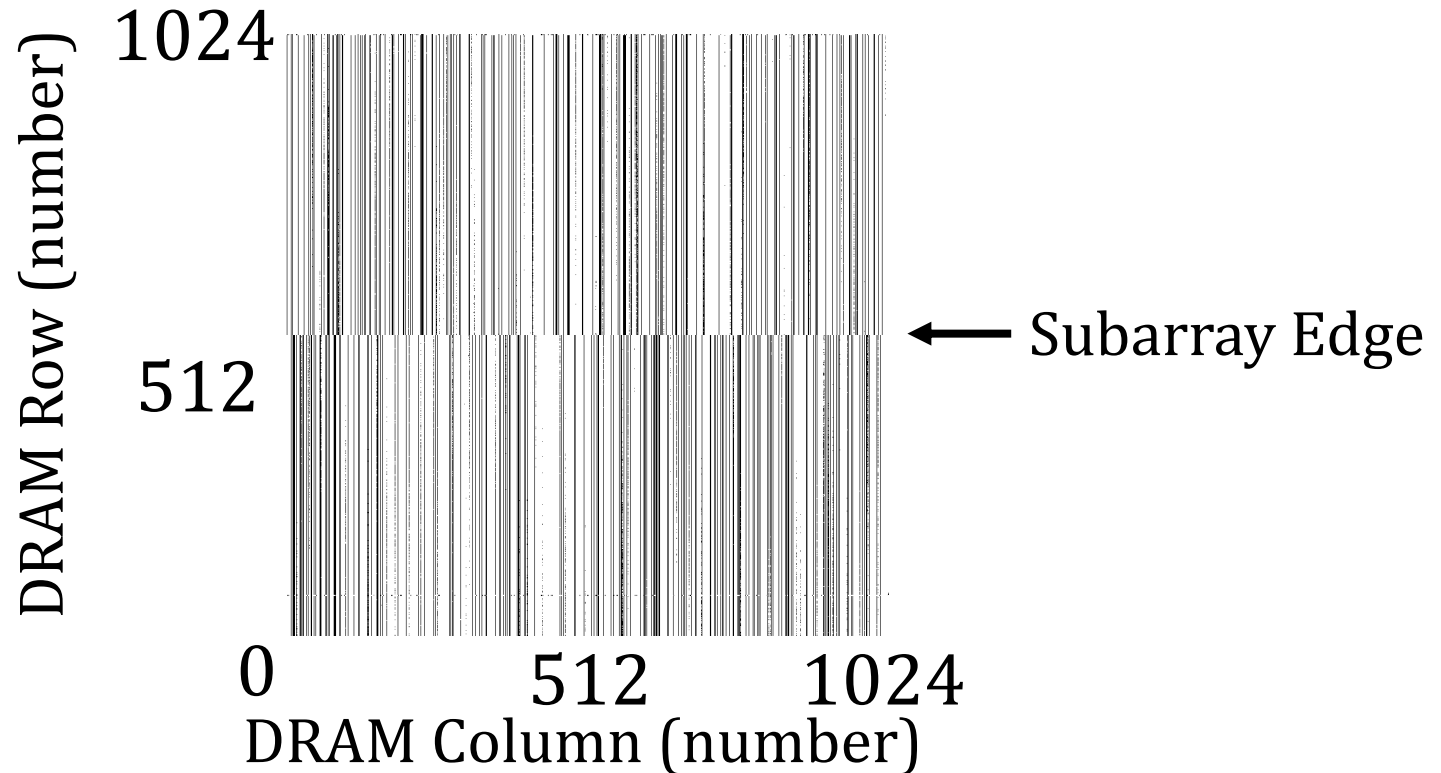
- Jeremie S. Kim, Minesh Patel, Hasan Hassan, and Onur Mutlu,
"Solar-DRAM: Reducing DRAM Access Latency by Exploiting the Variation in Local Bitlines"
Proceedings of the 36th IEEE International Conference on Computer Design (ICCD), Orlando, FL, USA, October 2018.

Solar-DRAM: Reducing DRAM Access Latency by Exploiting the Variation in Local Bitlines

Jeremie S. Kim^{‡§} Minesh Patel[§] Hasan Hassan[§] Onur Mutlu^{§‡}
 ‡Carnegie Mellon University §ETH Zürich

Spatial Distribution of Failures

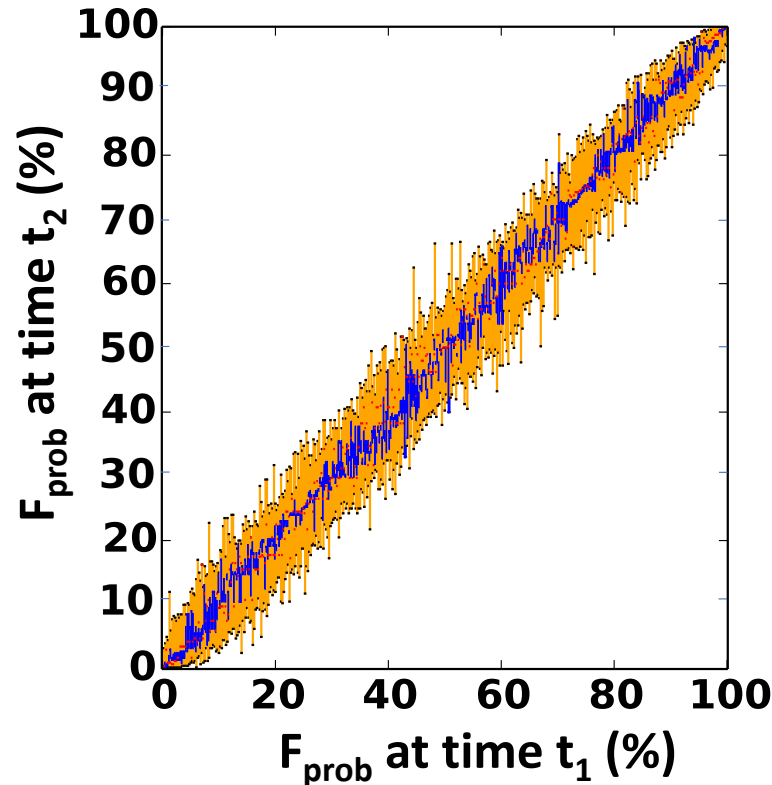
How are activation failures spatially distributed in DRAM?



Activation failures are **highly constrained**
to local bitlines

Short-term Variation

Does a bitline's probability of failure change over time?



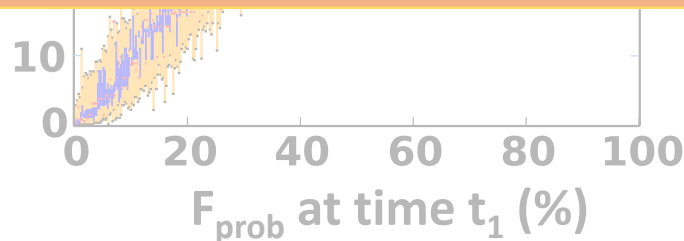
A **weak** bitline is likely to remain **weak** and
a **strong** bitline is likely to remain **strong** over time

Short-term Variation

Does a bitline's probability of failure change over time?



We can rely on a **static profile** of weak bitlines to determine whether an access will cause failures

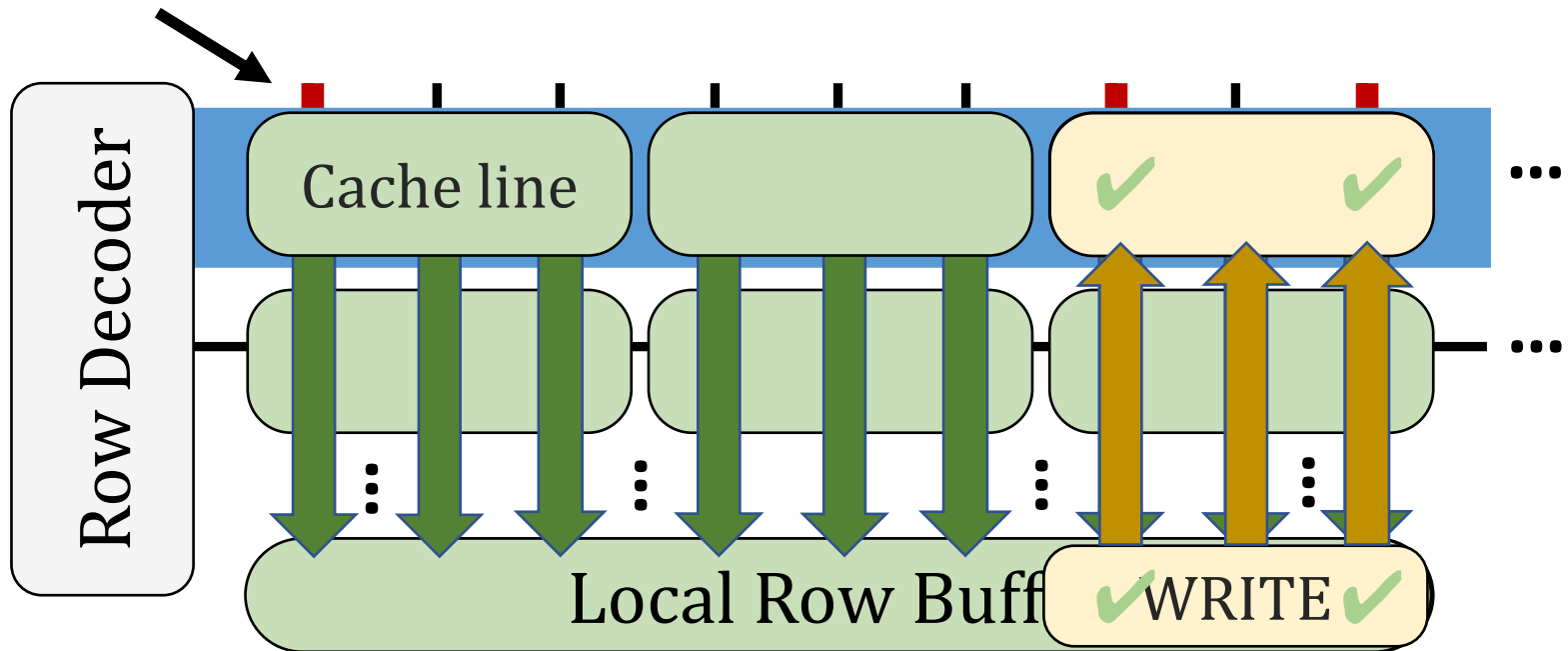


A **weak bitline** is likely to remain **weak** and a **strong bitline** is likely to remain **strong** over time

Write Operations

How are write operations affected by reduced t_{RCD} ?

Weak bitline



We can reliably issue write operations
with significantly reduced t_{RCD} (e.g., by 77%)

Solar-DRAM

Uses a **static profile of weak subarray columns**

- Identifies subarray columns as weak or strong
- Obtained in a one-time profiling step

Three Components

1. Variable-latency cache lines (VLC)
2. Reordered subarray columns (RSC)
3. Reduced latency for writes (RLW)

Solar-DRAM

Uses a **static profile of weak subarray columns**

- Identifies subarray columns as weak or strong
- Obtained in a one-time profiling step

Three Components

1. Variable-latency cache lines (VLC)
2. Reordered subarray columns (RSC)
3. Reduced latency for writes (RLW)

Solar-DRAM

Uses a **static profile of weak subarray columns**

- Identifies subarray columns as weak or strong
- Obtained in a one-time profiling step

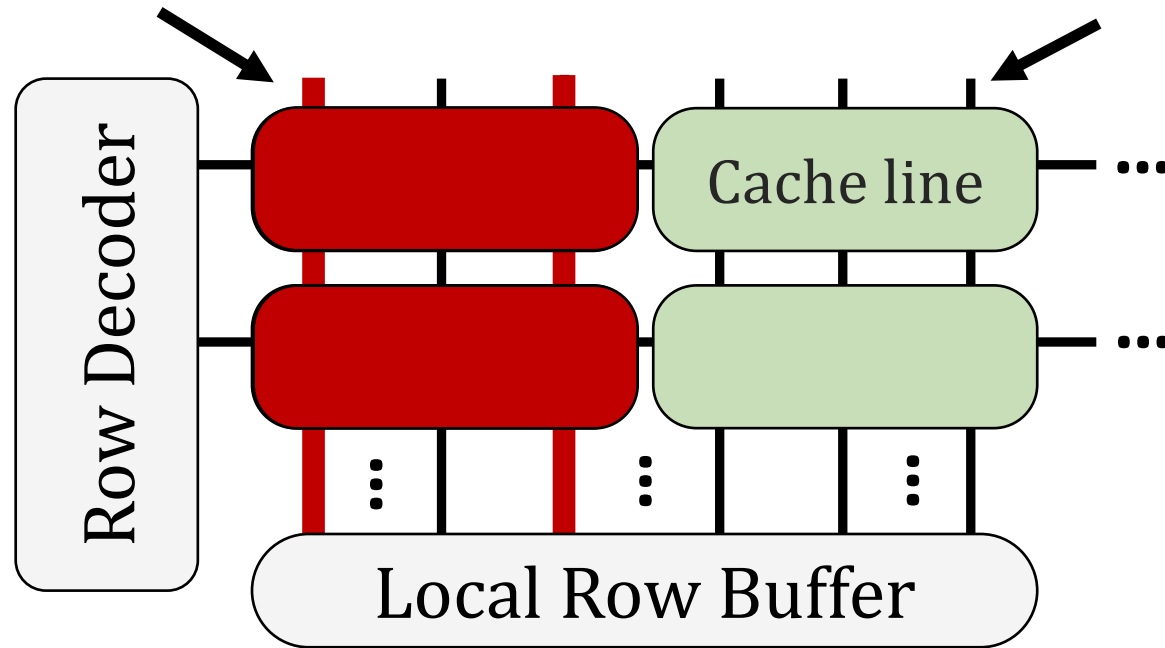
Three Components

1. Variable-latency cache lines (VLC)
2. Reordered subarray columns (RSC)
3. Reduced latency for writes (RLW)

Solar-DRAM: VLC (I)

Weak bitline

Strong bitline



Identify cache lines comprised of **strong bitlines**

Access such cache lines with a **reduced t_{RCD}**

Solar-DRAM

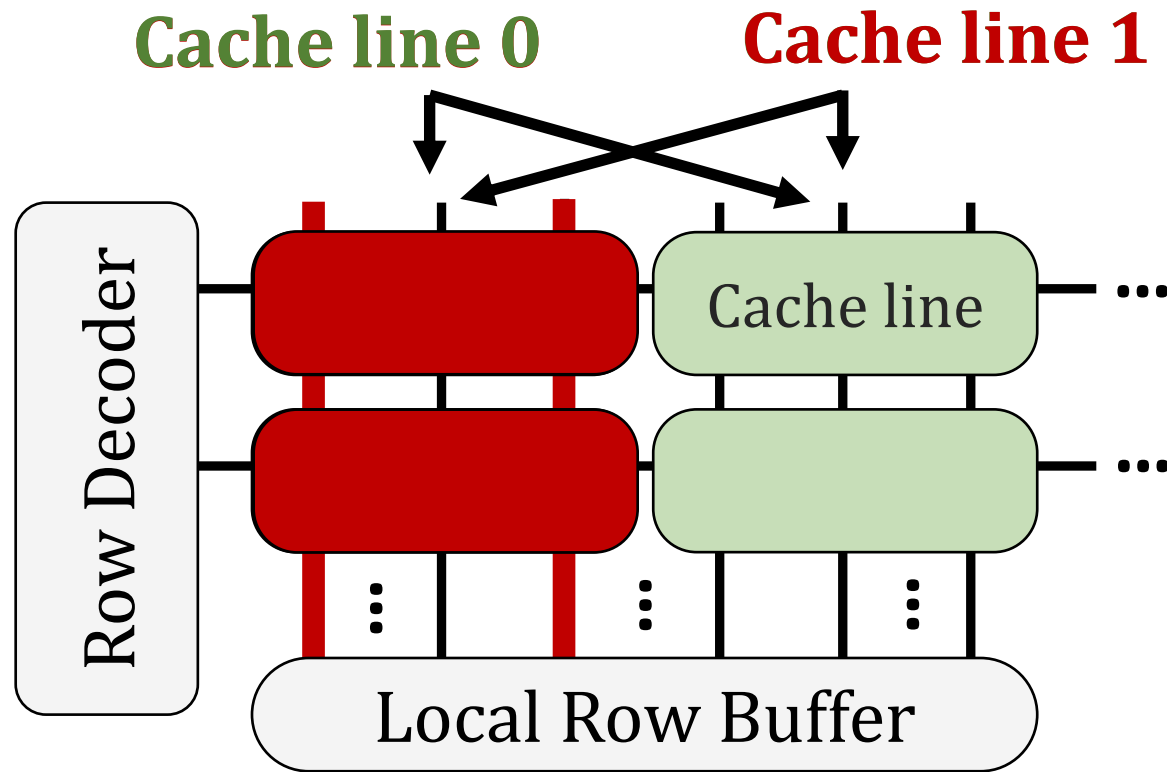
Uses a **static profile of weak subarray columns**

- Identifies subarray columns as weak or strong
- Obtained in a one-time profiling step

Three Components

1. Variable-latency cache lines (VLC)
2. Reordered subarray columns (RSC)
3. Reduced latency for writes (RLW)

Solar-DRAM: RSC (II)



Remap cache lines across DRAM at the memory controller level so cache line 0 will likely map to a **strong** cache line

Solar-DRAM

Uses a **static profile of weak subarray columns**

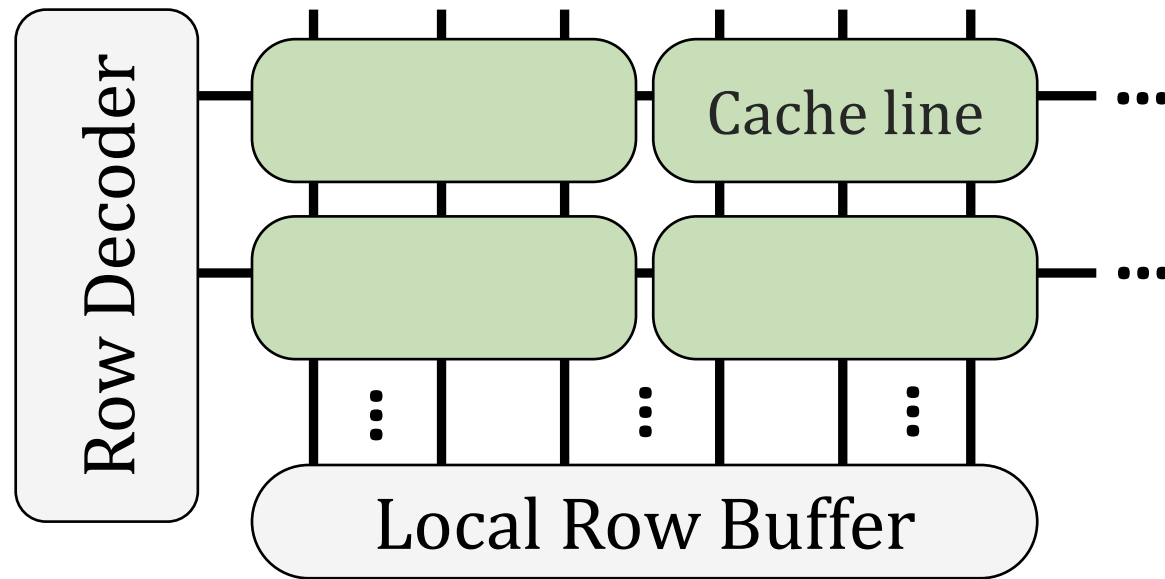
- Identifies subarray columns as weak or strong
- Obtained in a one-time profiling step

Three Components

1. Variable-latency cache lines (VLC)
2. Reordered subarray columns (RSC)
3. Reduced latency for writes (RLW)

Solar-DRAM: RLW (III)

All bitlines are strong when issuing writes



Write to all locations in DRAM with a significantly reduced t_{RCD} (e.g., by 77%)

More on Solar-DRAM

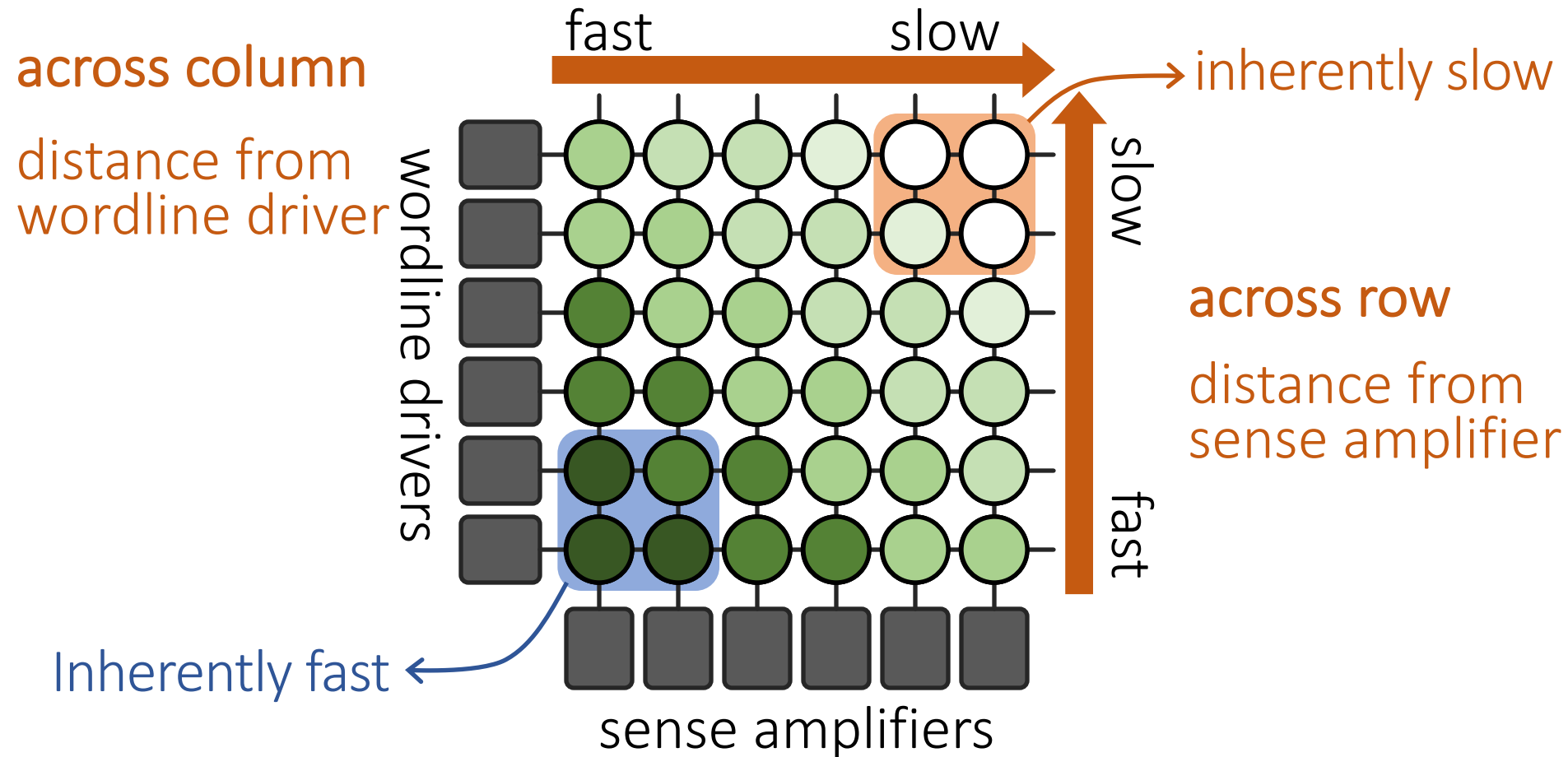
- Jeremie S. Kim, Minesh Patel, Hasan Hassan, and Onur Mutlu,
"Solar-DRAM: Reducing DRAM Access Latency by Exploiting the Variation in Local Bitlines"
Proceedings of the 36th IEEE International Conference on Computer Design (ICCD), Orlando, FL, USA, October 2018.

Solar-DRAM: Reducing DRAM Access Latency by Exploiting the Variation in Local Bitlines

Jeremie S. Kim^{‡§} Minesh Patel[§] Hasan Hassan[§] Onur Mutlu^{§‡}
 ‡Carnegie Mellon University §ETH Zürich

Why Is There Spatial Latency Variation Within a Chip?

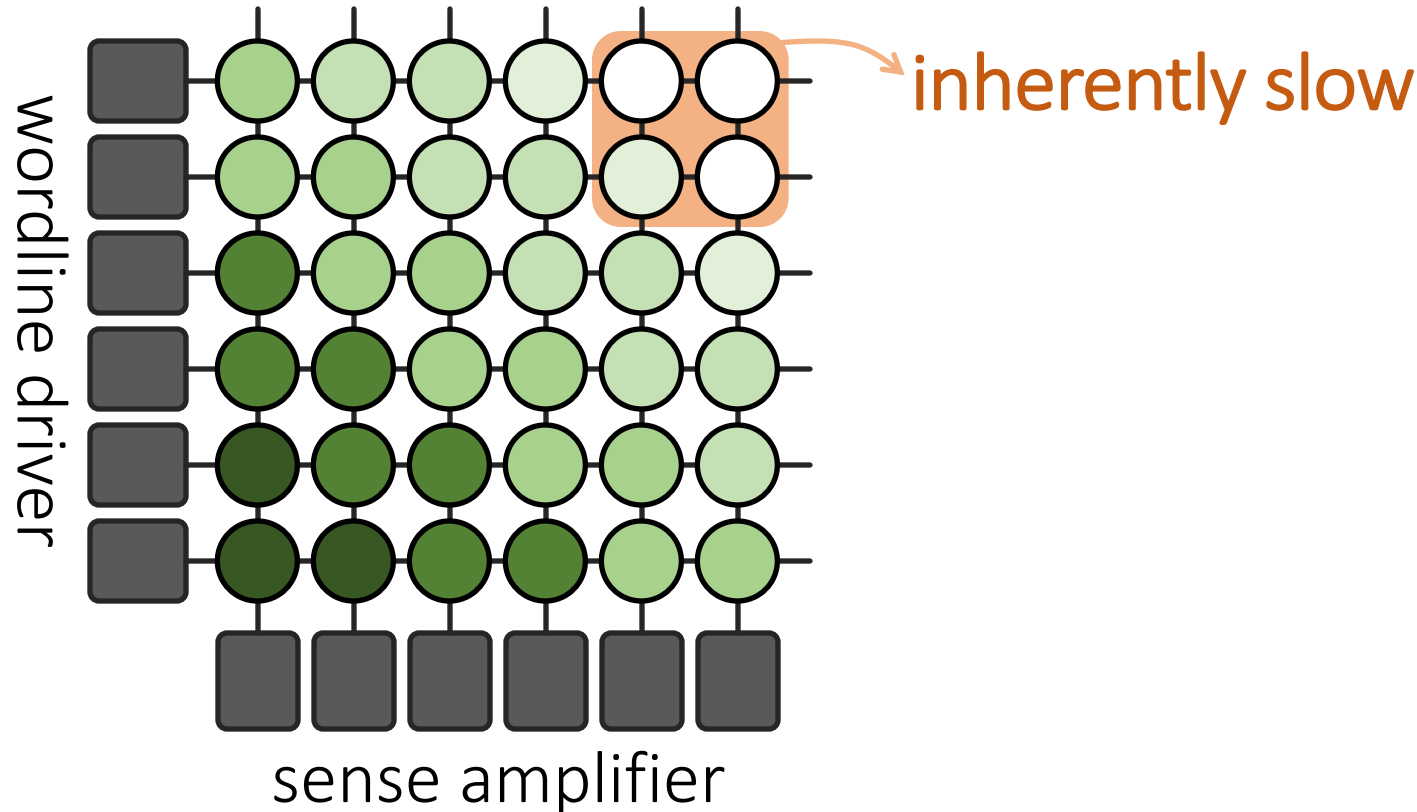
What Is Design-Induced Variation?



Systematic variation in cell access times
caused by the ***physical organization*** of DRAM

DIVA Online Profiling

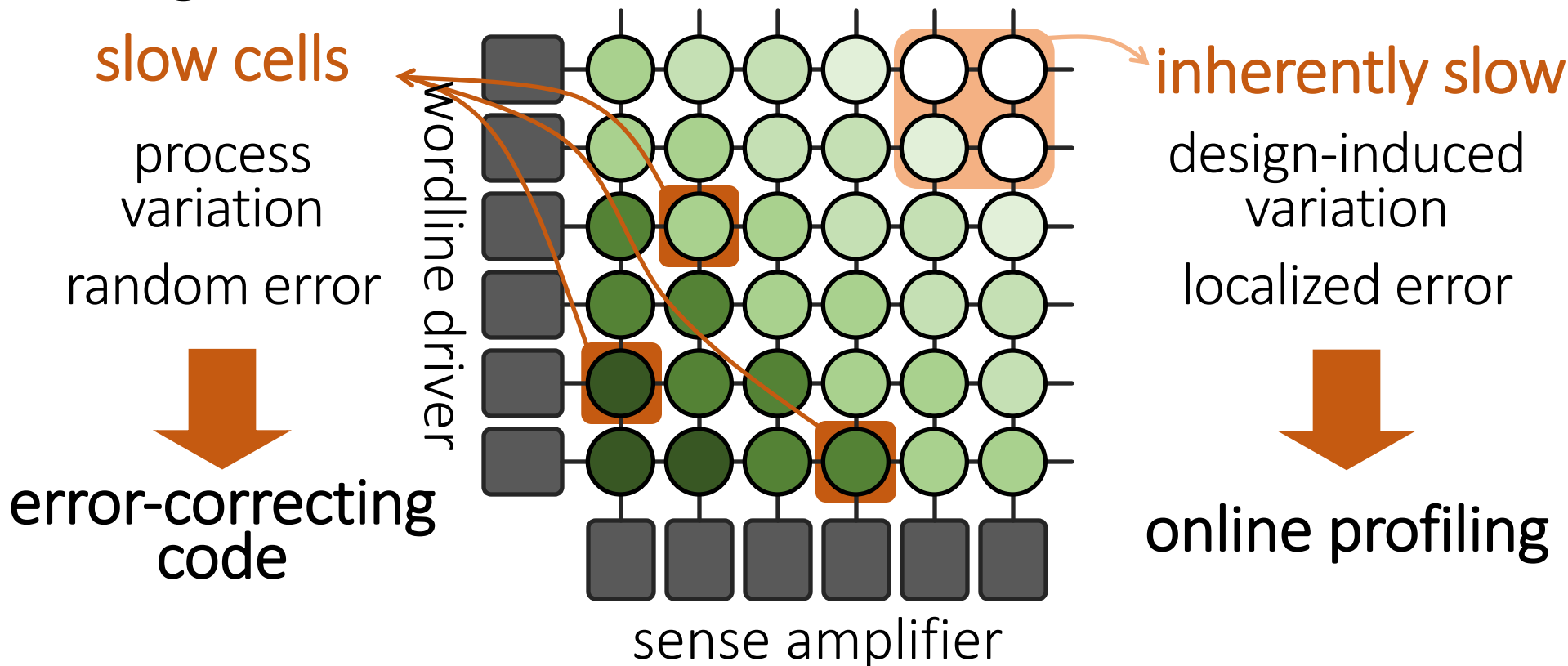
Design-Induced-Variation-Aware



Profile *only slow regions* to determine min. latency
→ *Dynamic* & *low cost* latency optimization

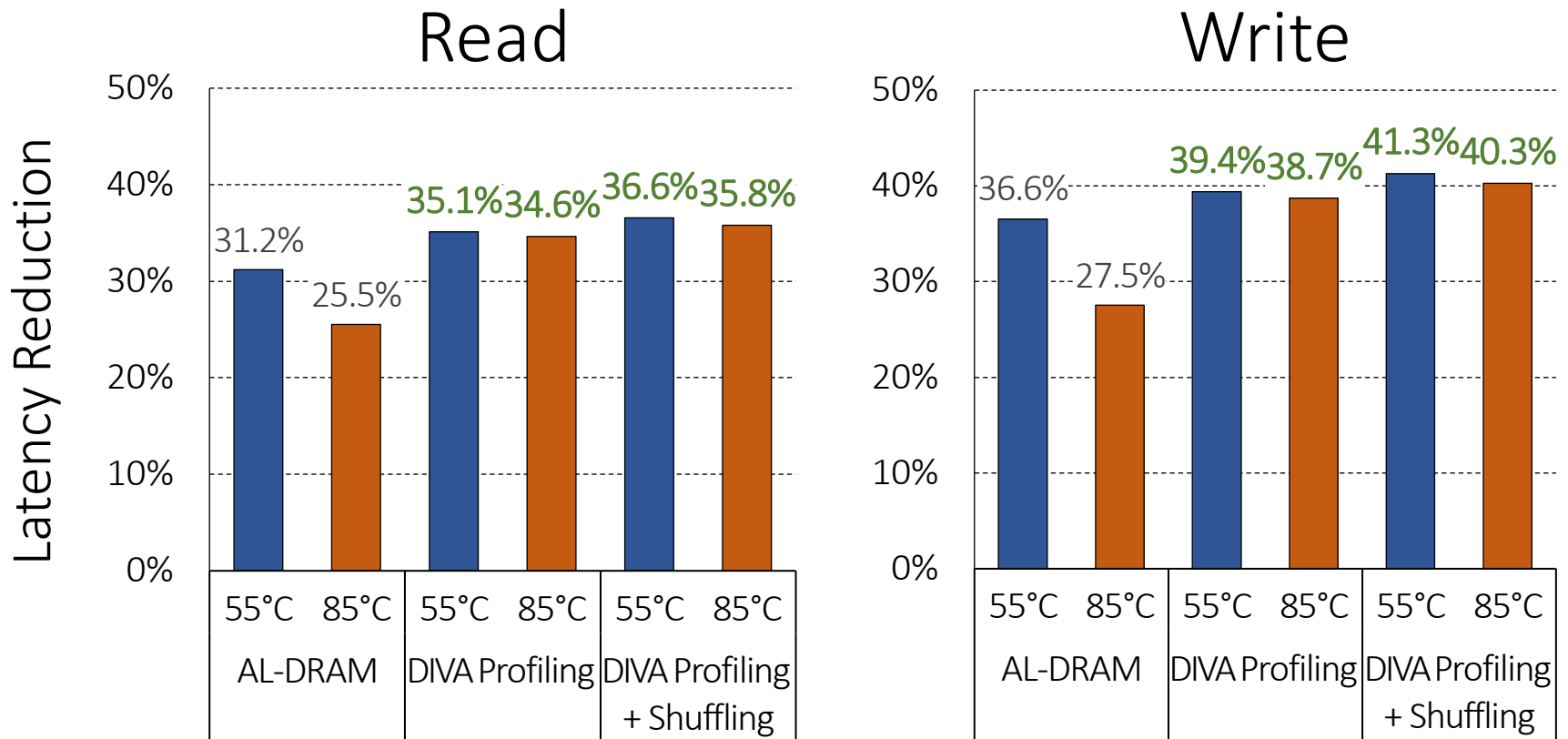
DIVA Online Profiling

Design-Induced-Variation-Aware



Combine **error-correcting codes** & **online profiling**
→ **Reliably** reduce DRAM latency

DIVA-DRAM Reduces Latency



DIVA-DRAM *reduces latency more aggressively*
and uses ECC to correct random slow cells

DIVA-DRAM: Advantages & Disadvantages

■ Advantages

- ++ Automatically finds the lowest reliable operating latency at system runtime (lower production-time testing cost)
- + Reduces latency more than prior methods (w/ ECC)
- + Reduces latency at high temperatures as well

■ Disadvantages

- Requires knowledge of inherently-slow regions
- Requires ECC (Error Correcting Codes)
- Imposes overhead during runtime profiling

Design-Induced Latency Variation in DRAM

- Donghyuk Lee, Samira Khan, Lavanya Subramanian, Saugata Ghose, Rachata Ausavarungnirun, Gennady Pekhimenko, Vivek Seshadri, and Onur Mutlu,
"Design-Induced Latency Variation in Modern DRAM Chips: Characterization, Analysis, and Latency Reduction Mechanisms"
*Proceedings of the ACM International Conference on Measurement and Modeling of Computer Systems (**SIGMETRICS**), Urbana-Champaign, IL, USA, June 2017.*

Design-Induced Latency Variation in Modern DRAM Chips: Characterization, Analysis, and Latency Reduction Mechanisms

Donghyuk Lee, NVIDIA and Carnegie Mellon University

Samira Khan, University of Virginia

Lavanya Subramanian, Saugata Ghose, Rachata Ausavarungnirun, Carnegie Mellon University

Gennady Pekhimenko, Vivek Seshadri, Microsoft Research

Onur Mutlu, ETH Zürich and Carnegie Mellon University

Understanding & Exploiting the Voltage-Latency-Reliability Relationship

High DRAM Power Consumption

- Problem: High DRAM (memory) power in today's systems



>40% in POWER7 (Ware+, HPCA'10)



>40% in GPU (Paul+, ISCA'15)

Low-Voltage Memory

- Existing DRAM designs to help reduce DRAM power by lowering supply voltage conservatively
 - $Power \propto Voltage^2$
- DDR3L (low-voltage) reduces voltage from 1.5V to 1.35V (-10%)
- LPDDR4 (low-power) employs low-power I/O interface with 1.2V (lower bandwidth)

Can we reduce DRAM power and energy by further reducing supply voltage?

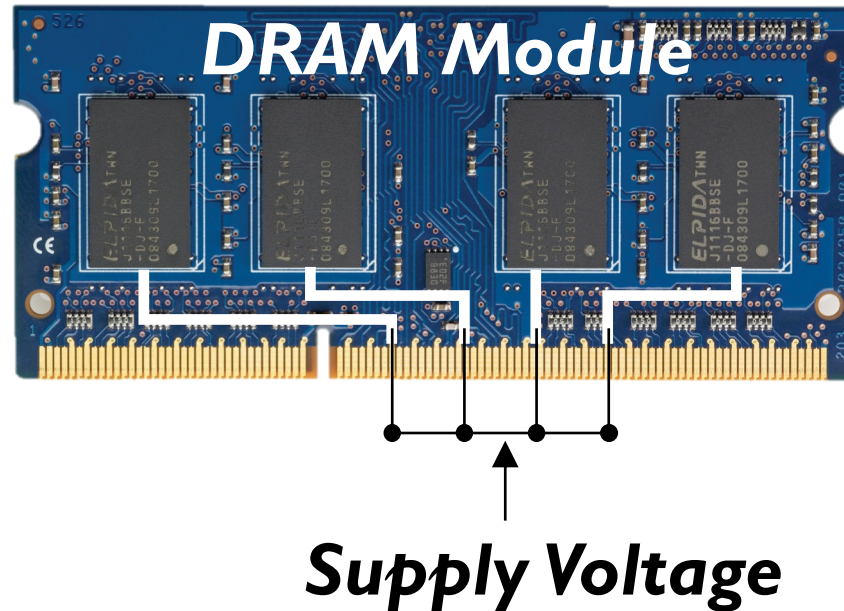
Goals

- 1 Understand and characterize the various characteristics of DRAM under **reduced voltage**
- 2 Develop a mechanism that reduces DRAM energy by **lowering voltage** while keeping performance loss within a target

Key Questions

- How does reducing voltage affect ***reliability*** (errors)?
- How does reducing voltage affect ***DRAM latency***?
- How do we design a new DRAM energy reduction mechanism?

Supply Voltage Control on DRAM



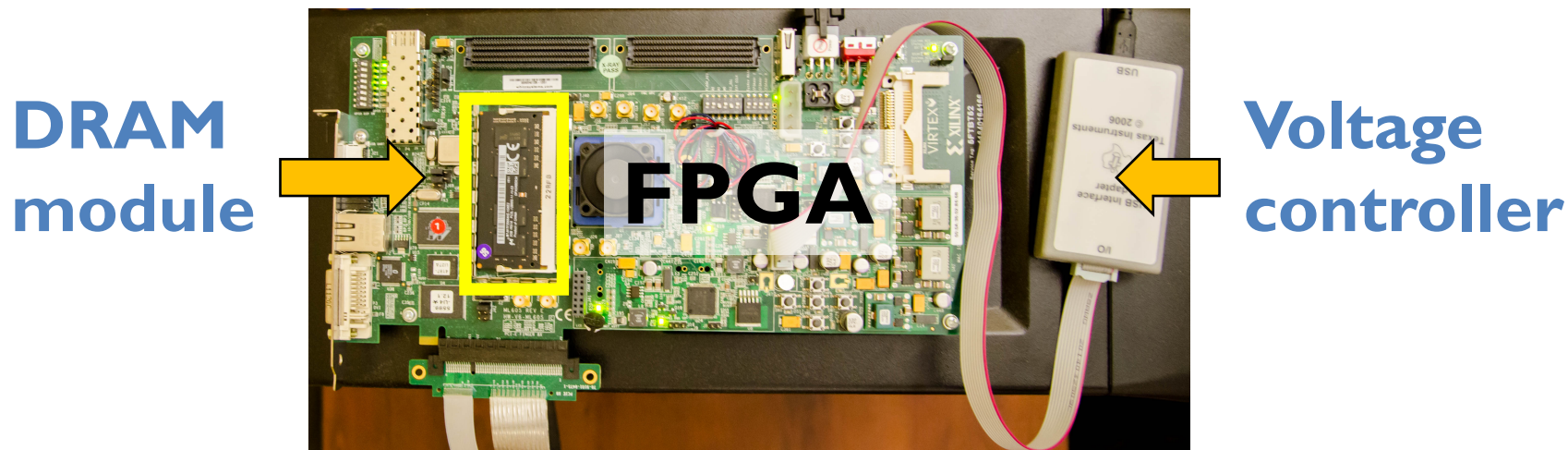
Adjust the *supply voltage* to every chip on the same module

Custom Testing Platform

SoftMC [Hassan+, HPCA'17]: FPGA testing platform to

- 1) Adjust supply voltage to DRAM modules
- 2) Schedule DRAM commands to DRAM modules

Existing systems: DRAM commands not exposed to users

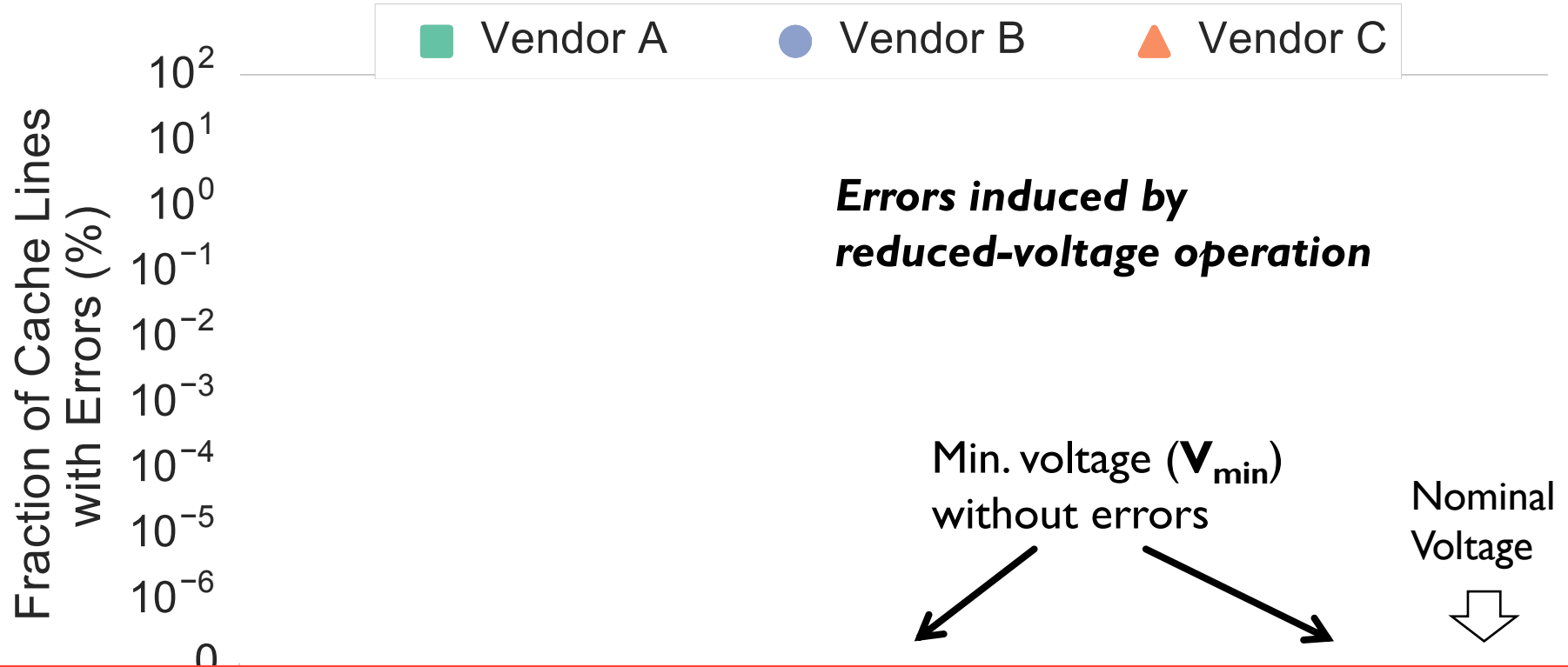


<https://github.com/CMU-SAFARI/DRAM-Voltage-Study>

Tested DRAM Modules

- **124 DDR3L** (low-voltage) DRAM chips
 - **31 SO-DIMMs**
 - **1.35V** (DDR3 uses 1.5V)
 - Density: 4Gb per chip
 - Three major vendors/manufacturers
 - Manufacturing dates: 2014-2016
- Iteratively read every bit in each 4Gb chip under a wide range of supply voltage levels: 1.35V to 1.0V (**-26%**)

Reliability Worsens with Lower Voltage



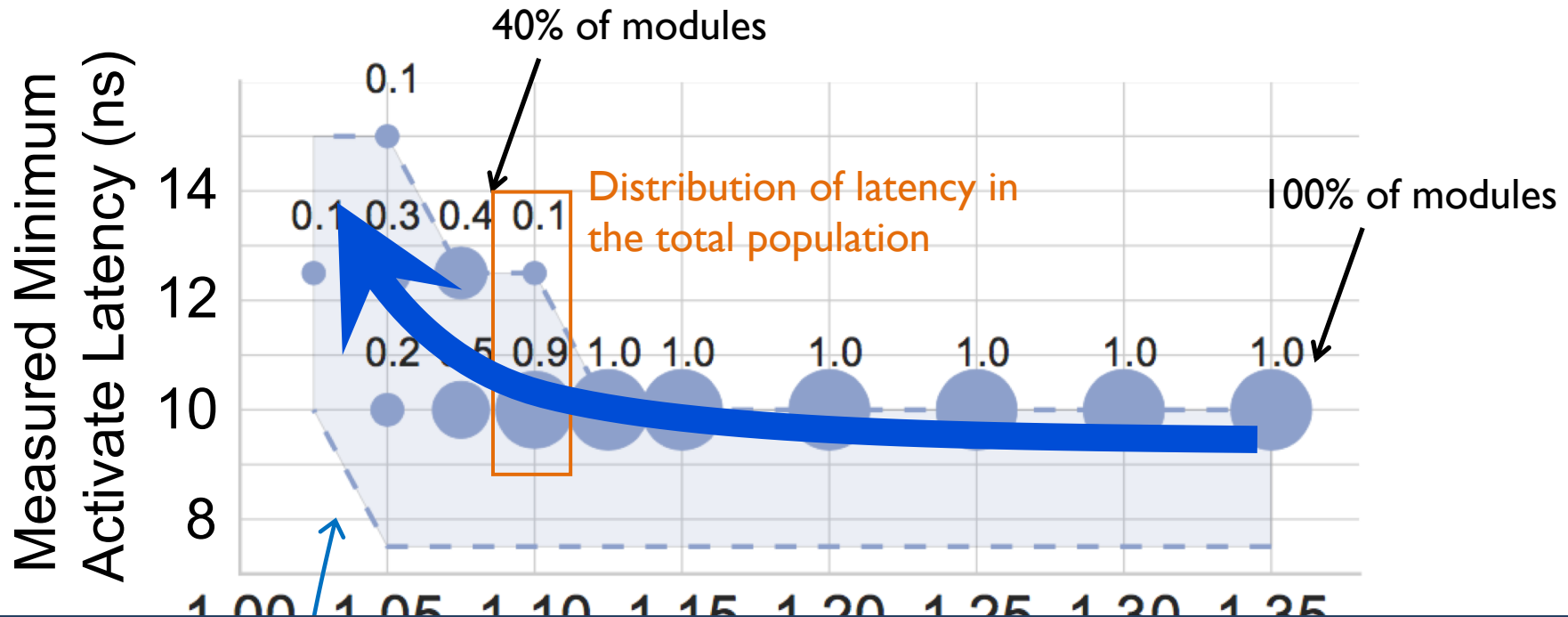
Reducing voltage below V_{\min} causes an increasing number of errors

Detailed circuit simulations (SPICE) of a DRAM cell array to model the behavior of DRAM operations



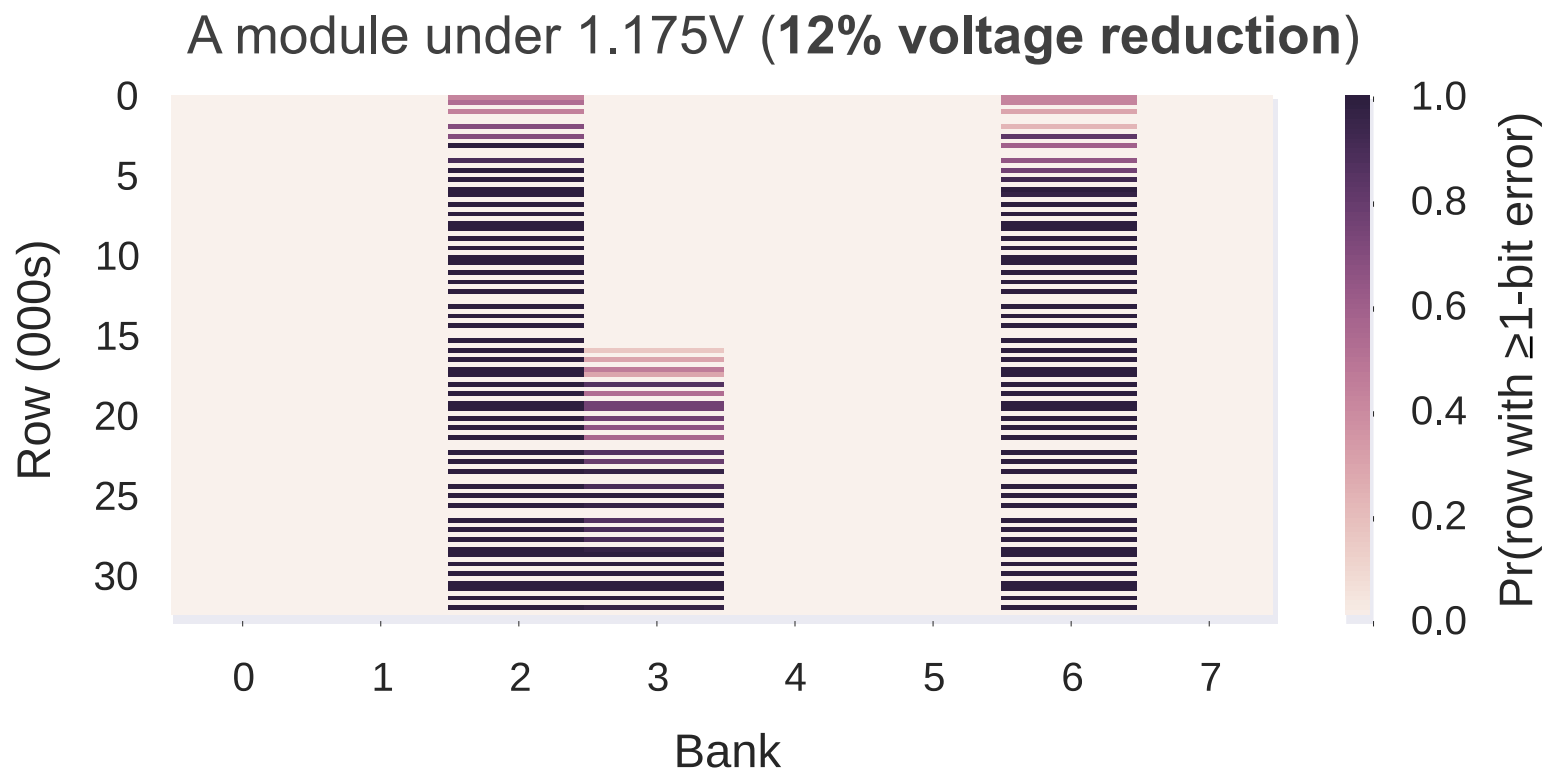
DIMMs Operating at Higher Latency

Measured minimum latency that *does not* cause errors in DRAM modules



DRAM requires longer latency to access data **without errors** at lower voltage

Spatial Locality of Errors



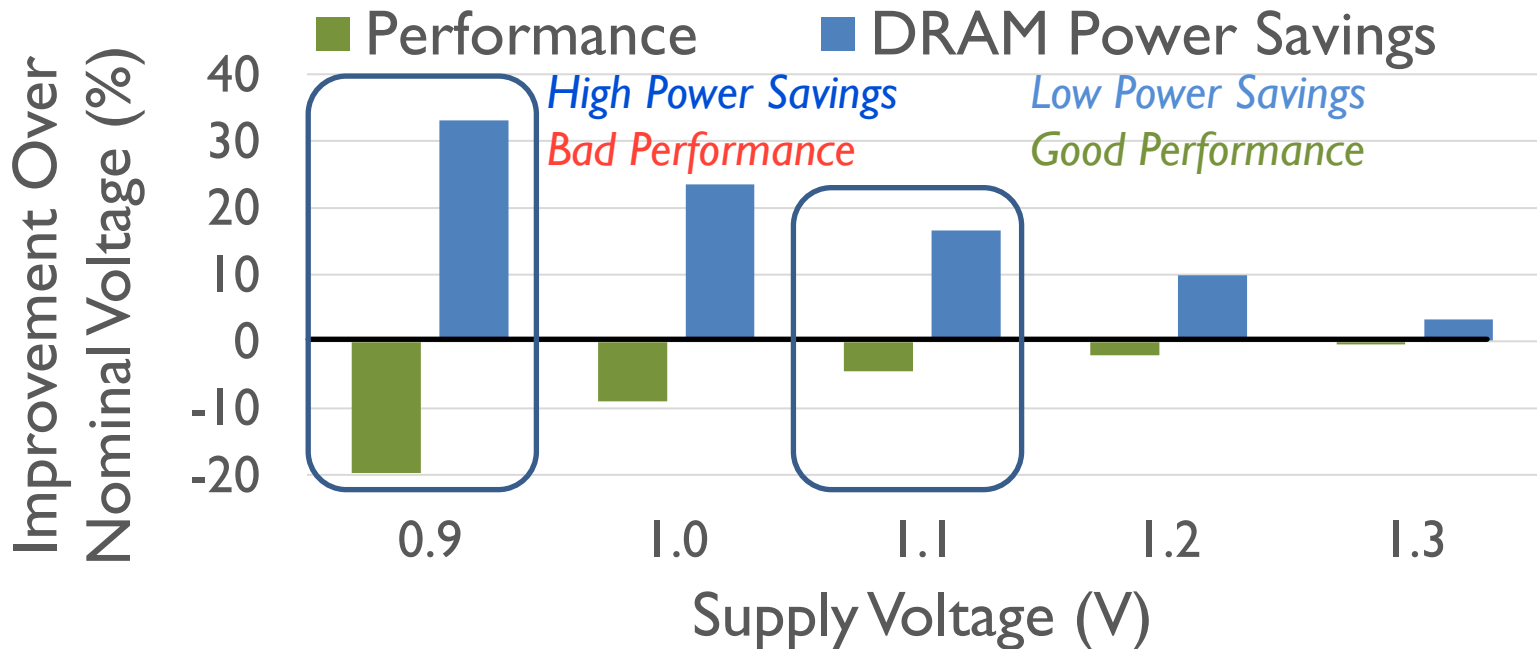
Errors concentrate in certain regions

Summary of Key Experimental Observations

- Voltage-induced errors increase as voltage reduces further below V_{\min}
- Errors exhibit spatial locality
- Increasing the latency of DRAM operations mitigates voltage-induced errors

DRAM Voltage Adjustment to Reduce Energy

- Goal: Exploit the trade-off between voltage and latency to reduce energy consumption
- Approach: Reduce DRAM voltage **reliably**
 - **Performance loss** due to increased latency at lower voltage

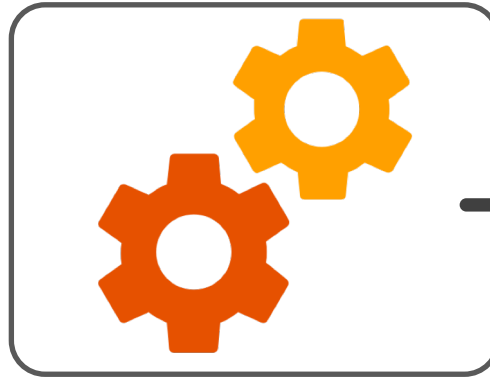


Voltron Overview

Voltron



User specifies the
performance loss target

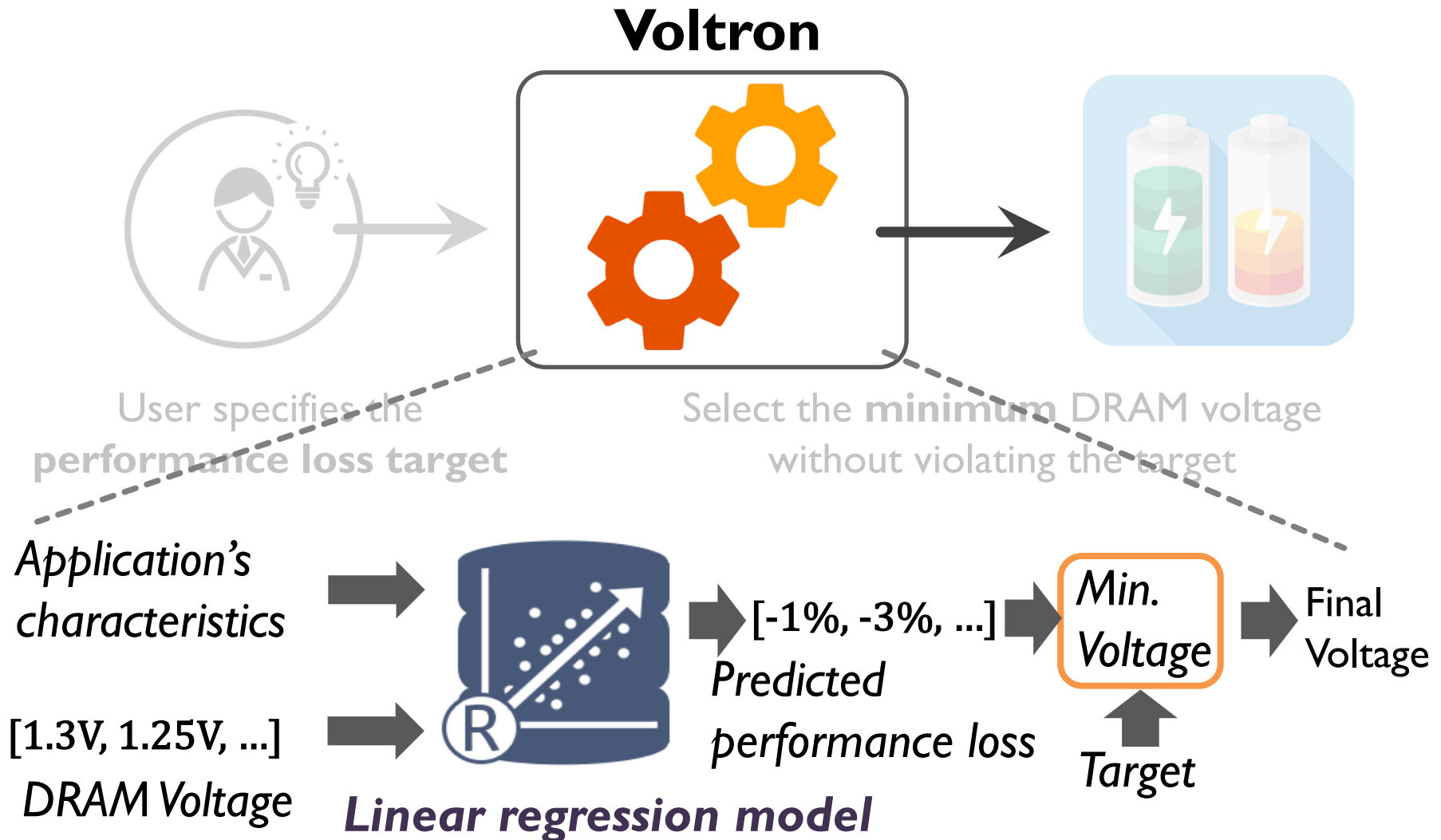


Select the **minimum** DRAM voltage
without violating the target



How do we predict performance loss due to increased latency under low DRAM voltage?

Linear Model to Predict Performance

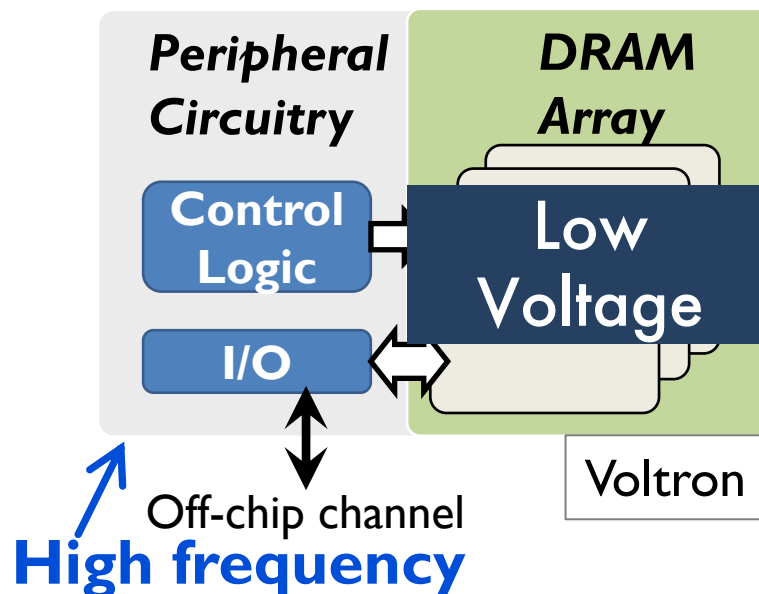
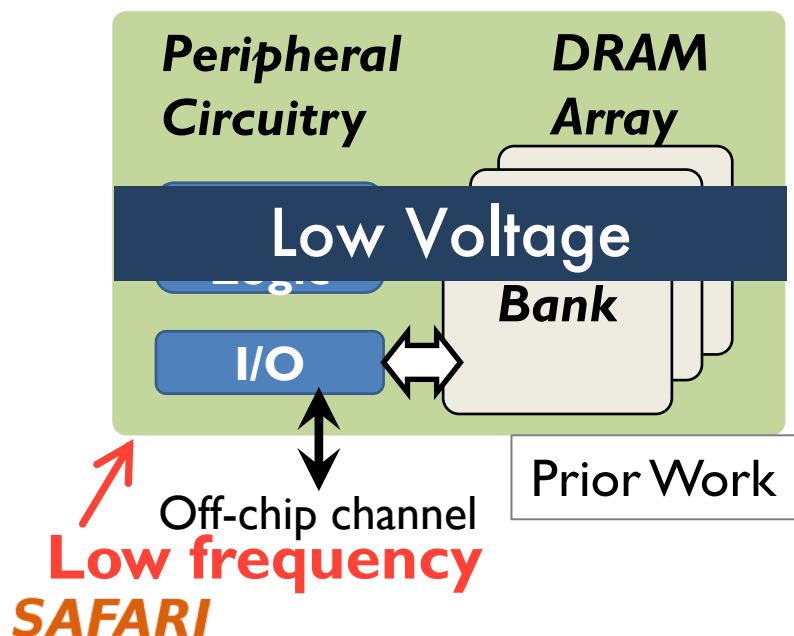


Regression Model to Predict Performance

- Application's characteristics for the model:
 - **Memory intensity**: Frequency of last-level cache misses
 - **Memory stall time**: Amount of time memory requests stall commit inside CPU
- Handling multiple applications:
 - Predict a performance loss for each application
 - Select the minimum voltage that satisfies the performance target for all applications

Comparison to Prior Work

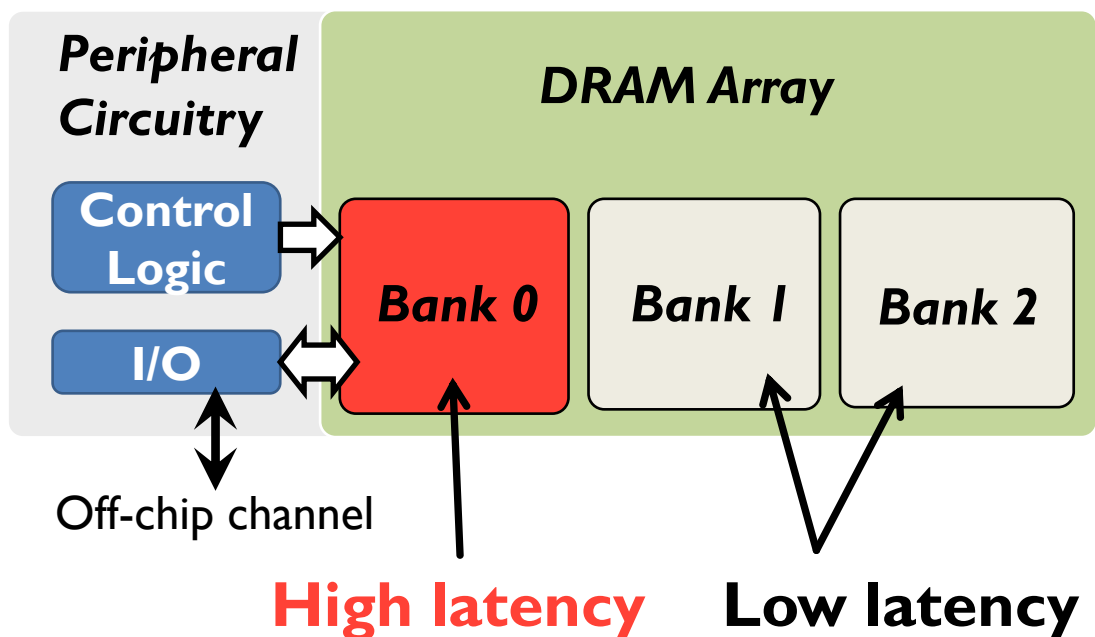
- Prior work: Dynamically scale *frequency and voltage* of the entire DRAM based on bandwidth demand [David+, ICAC'11]
 - Problem: Lowering voltage on the peripheral circuitry decreases channel frequency (memory data throughput)
- Voltron: Reduce voltage to only **DRAM array** without changing the voltage to peripheral circuitry



Exploiting Spatial Locality of Errors

Key idea: Increase the latency only for DRAM banks that observe errors under low voltage

- Benefit: Higher performance

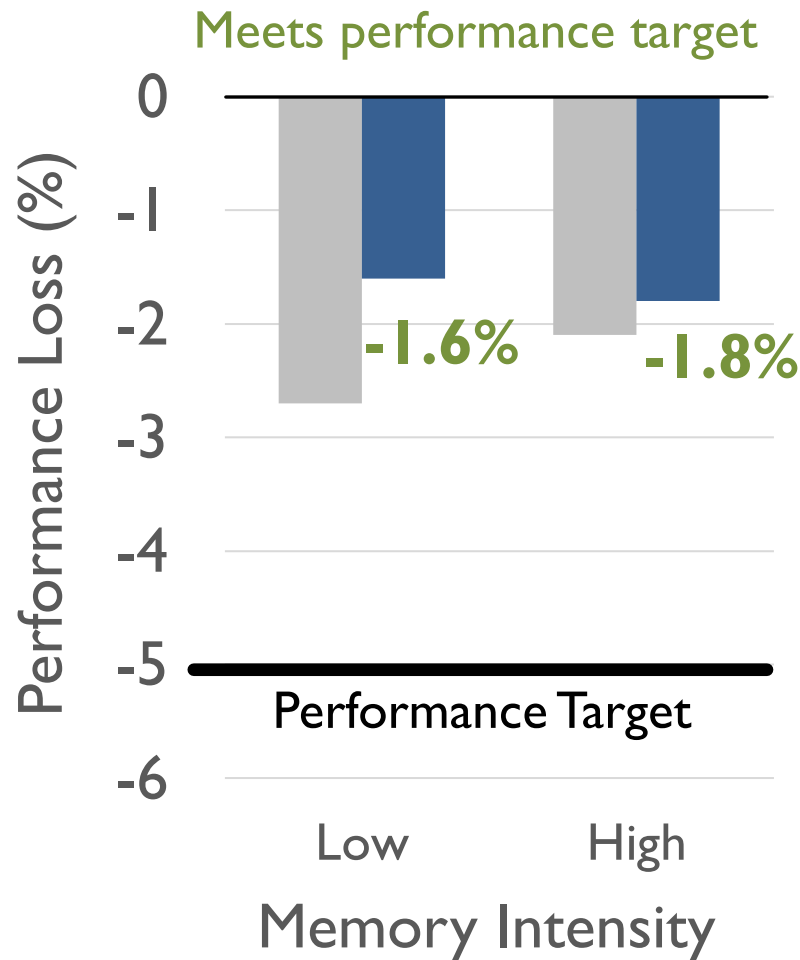
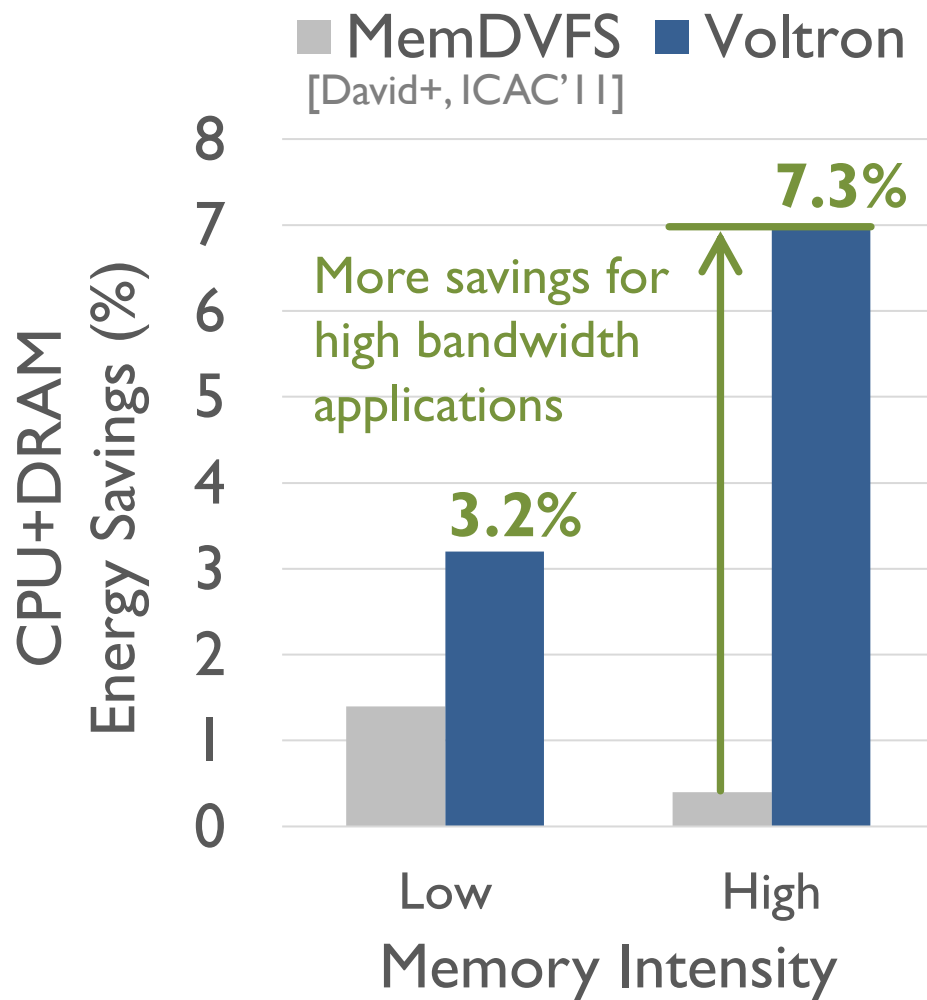


Voltron Evaluation Methodology

- **Cycle-level simulator:** Ramulator [CAL'15]
 - **McPAT** and **DRAMPower** for energy measurement

<https://github.com/CMU-SAFARI/ramulator>
- **4-core** system with DDR3L memory
- **Benchmarks:** SPEC2006, YCSB
- Comparison to prior work: **MemDVFS** [David+, ICAC'11]
 - Dynamic DRAM frequency and voltage scaling
 - Scaling based on the *memory bandwidth consumption*

Energy Savings with Bounded Performance



Voltron: Advantages & Disadvantages

■ Advantages

- + Can trade-off between voltage and latency to improve energy or performance
- + Can exploit the high voltage margin present in DRAM

■ Disadvantages

- Requires finding the reliable operating voltage for each chip → higher testing cost

Analysis of Latency-Voltage in DRAM Chips

- Kevin Chang, A. Giray Yaglikci, Saugata Ghose, Aditya Agrawal, Niladrish Chatterjee, Abhijith Kashyap, Donghyuk Lee, Mike O'Connor, Hasan Hassan, and Onur Mutlu,

"Understanding Reduced-Voltage Operation in Modern DRAM Devices: Experimental Characterization, Analysis, and Mechanisms"

*Proceedings of the ACM International Conference on Measurement and Modeling of Computer Systems (**SIGMETRICS**), Urbana-Champaign, IL, USA, June 2017.*

Understanding Reduced-Voltage Operation in Modern DRAM Chips: Characterization, Analysis, and Mechanisms

Kevin K. Chang[†] Abdullah Giray Yağlıkçı[†] Saugata Ghose[†] Aditya Agrawal[¶] Niladrish Chatterjee[¶]
Abhijith Kashyap[†] Donghyuk Lee[¶] Mike O'Connor^{¶,‡} Hasan Hassan[§] Onur Mutlu^{§,†}

[†]Carnegie Mellon University

[¶]NVIDIA

[‡]The University of Texas at Austin

[§]ETH Zürich

And, What If ...

- ... we can sacrifice reliability of some data to access it with even lower latency?

Reducing Refresh Latency

On Reducing Refresh Latency

- Anup Das, Hasan Hassan, and Onur Mutlu,
"VRL-DRAM: Improving DRAM Performance via Variable Refresh Latency"
*Proceedings of the 55th Design Automation Conference (**DAC**), San Francisco, CA, USA, June 2018.*

VRL-DRAM: Improving DRAM Performance via Variable Refresh Latency

Anup Das
Drexel University
Philadelphia, PA, USA
anup.das@drexel.edu

Hasan Hassan
ETH Zürich
Zürich, Switzerland
hhasan@ethz.ch

Onur Mutlu
ETH Zürich
Zürich, Switzerland
omutlu@gmail.com

The DRAM Latency PUF:

Quickly Evaluating Physical Unclonable Functions
by Exploiting the Latency-Reliability Tradeoff
in Modern Commodity DRAM Devices

Jeremie S. Kim Minesh Patel

Hasan Hassan Onur Mutlu



QR Code for the paper

https://people.inf.ethz.ch/omutlu/pub/dram-latency-puf_hpca18.pdf

HPCA 2018

SAFARI



ETH zürich

Carnegie Mellon

DRAM Latency PUFs

- Jeremie S. Kim, Minesh Patel, Hasan Hassan, and Onur Mutlu,
"The DRAM Latency PUF: Quickly Evaluating Physical Unclonable Functions by Exploiting the Latency-Reliability Tradeoff in Modern DRAM Devices"
Proceedings of the 24th International Symposium on High-Performance Computer Architecture (HPCA), Vienna, Austria, February 2018.
[[Lightning Talk Video](#)]
[[Slides \(pptx\)](#)] [[pdf](#)] [[Lightning Session Slides \(pptx\)](#)] [[pdf](#)]

The DRAM Latency PUF:

Quickly Evaluating Physical Unclonable Functions

by Exploiting the Latency-Reliability Tradeoff in Modern Commodity DRAM Devices

Jeremie S. Kim^{†§}

Minesh Patel[§]

Hasan Hassan[§]

Onur Mutlu^{§†}

[†]Carnegie Mellon University

[§]ETH Zürich

Reducing Memory Latency by Exploiting Memory Access Patterns

More on ChargeCache

- Hasan Hassan, Gennady Pekhimenko, Nandita Vijaykumar, Vivek Seshadri, Donghyuk Lee, Oguz Ergin, and Onur Mutlu,
"ChargeCache: Reducing DRAM Latency by Exploiting Row Access Locality"
Proceedings of the 22nd International Symposium on High-Performance Computer Architecture (HPCA), Barcelona, Spain, March 2016.
[[Slides \(pptx\)](#)] [[pdf](#)]
[[Source Code](#)]

ChargeCache: Reducing DRAM Latency by Exploiting Row Access Locality

Hasan Hassan^{†*}, Gennady Pekhimenko[†], Nandita Vijaykumar[†]
Vivek Seshadri[†], Donghyuk Lee[†], Oguz Ergin^{*}, Onur Mutlu[†]

A Very Recent Work

- Yaohua Wang, Arash Tavakkol, Lois Orosa, Saugata Ghose, Nika Mansouri Ghiasi, Minesh Patel, Jeremie S. Kim, Hasan Hassan, Mohammad Sadrosadati, and Onur Mutlu,
"Reducing DRAM Latency via Charge-Level-Aware Look-Ahead Partial Restoration"
Proceedings of the 51st International Symposium on Microarchitecture (MICRO), Fukuoka, Japan, October 2018.

Reducing DRAM Latency via Charge-Level-Aware Look-Ahead Partial Restoration

Yaohua Wang^{†§} Arash Tavakkol[†] Lois Orosa^{†*} Saugata Ghose[‡] Nika Mansouri Ghiasi[†]
Minesh Patel[†] Jeremie S. Kim^{‡†} Hasan Hassan[†] Mohammad Sadrosadati[†] Onur Mutlu^{‡†}

[†]*ETH Zürich*

[§]*National University of Defense Technology*

[‡]*Carnegie Mellon University*

^{*}*University of Campinas*

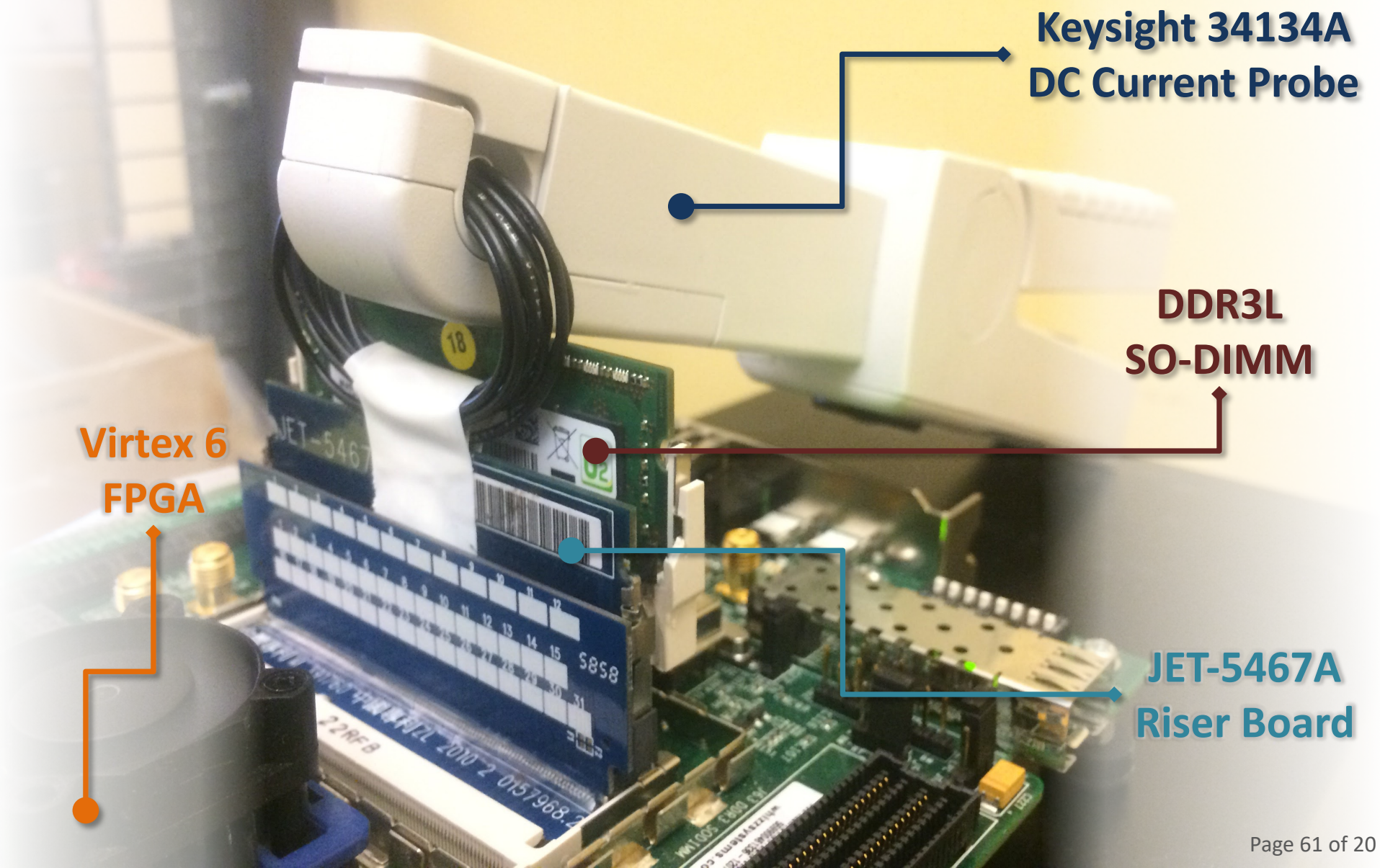
Summary: Low-Latency Memory

Summary: Tackling Long Memory Latency

- Reason 1: Design of DRAM Micro-architecture
 - Goal: Maximize capacity/area, not minimize latency
- Reason 2: “One size fits all” approach to latency specification
 - Same latency parameters for all temperatures
 - Same latency parameters for all DRAM chips
 - Same latency parameters for all parts of a DRAM chip
 - Same latency parameters for all supply voltage levels
 - Same latency parameters for all application data
 - Same latency parameters for all access patterns
 - ...

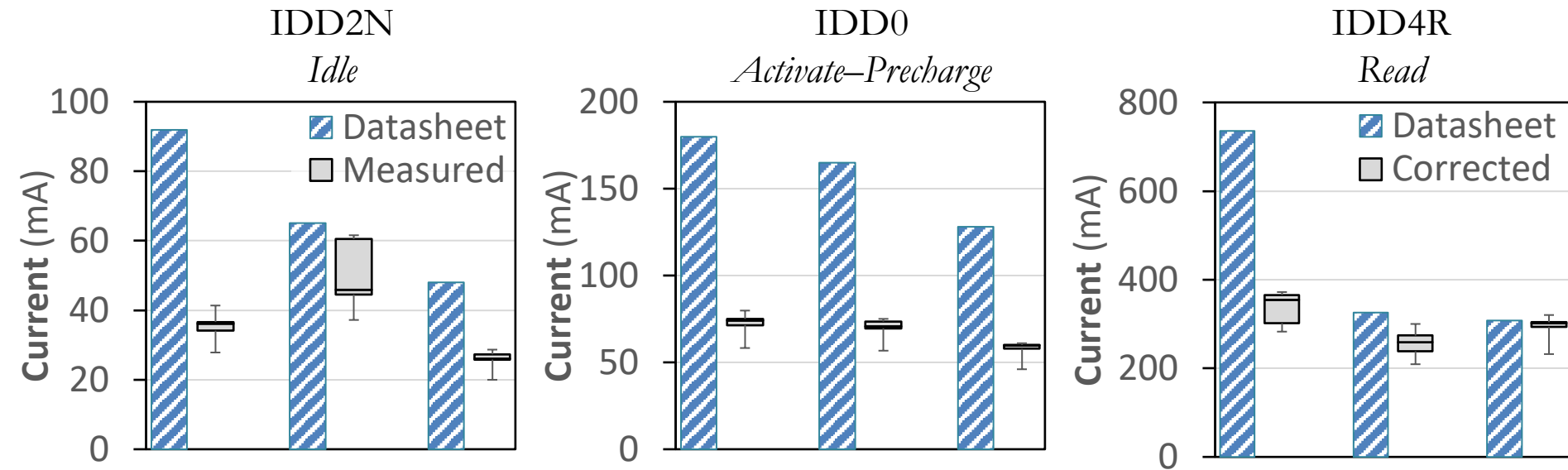
Fundamentally Low Latency Computing Architectures

On DRAM Power Consumption



- **SoftMC: an FPGA-based memory controller** [Hassan+ HPCA '17]
 - Modified to repeatedly loop commands
 - Open-source: <https://github.com/CMU-SAFARI/SoftMC>
- **Measure current consumed by a module during a SoftMC test**
- **Tested 50 DDR3L DRAM modules** (200 DRAM chips)
 - Supply voltage: 1.35 V
 - **Three major vendors: A, B, C**
 - Manufactured between 2014 and 2016
- **For each experimental test that we perform**
 - 10 runs of each test per module
 - At least 10 current samples per run

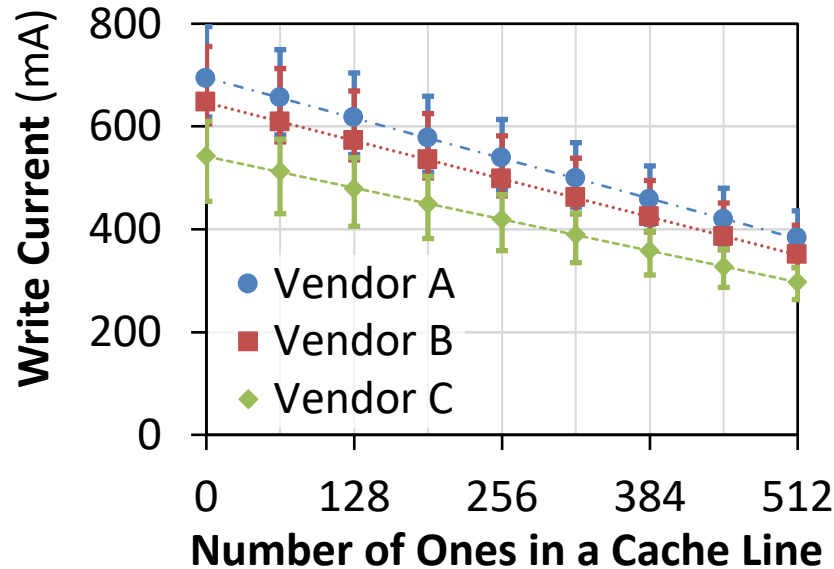
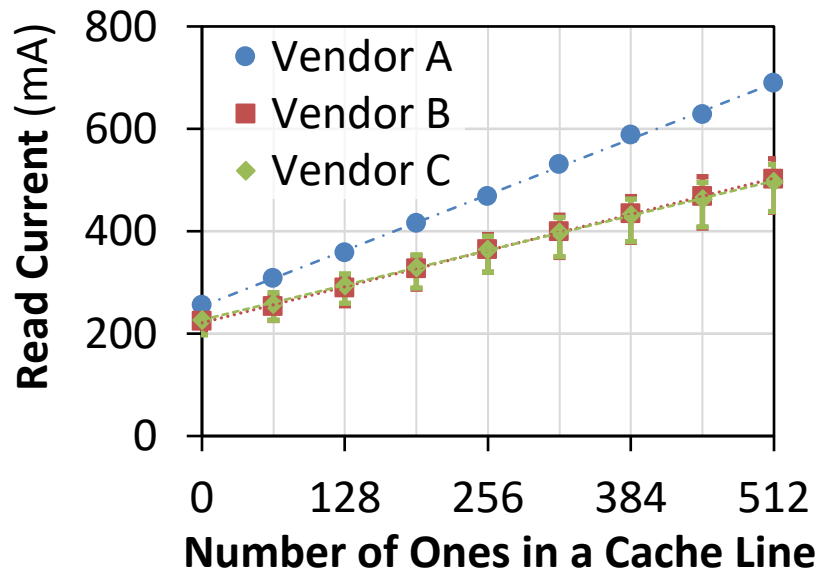
1. Real DRAM Power Varies Widely from IDD Values **SAFARI**



- Different vendors have very different margins (i.e., *guardbands*)
- Low variance among different modules from same vendor

Current consumed by real DRAM modules varies significantly for all IDD values that we measure

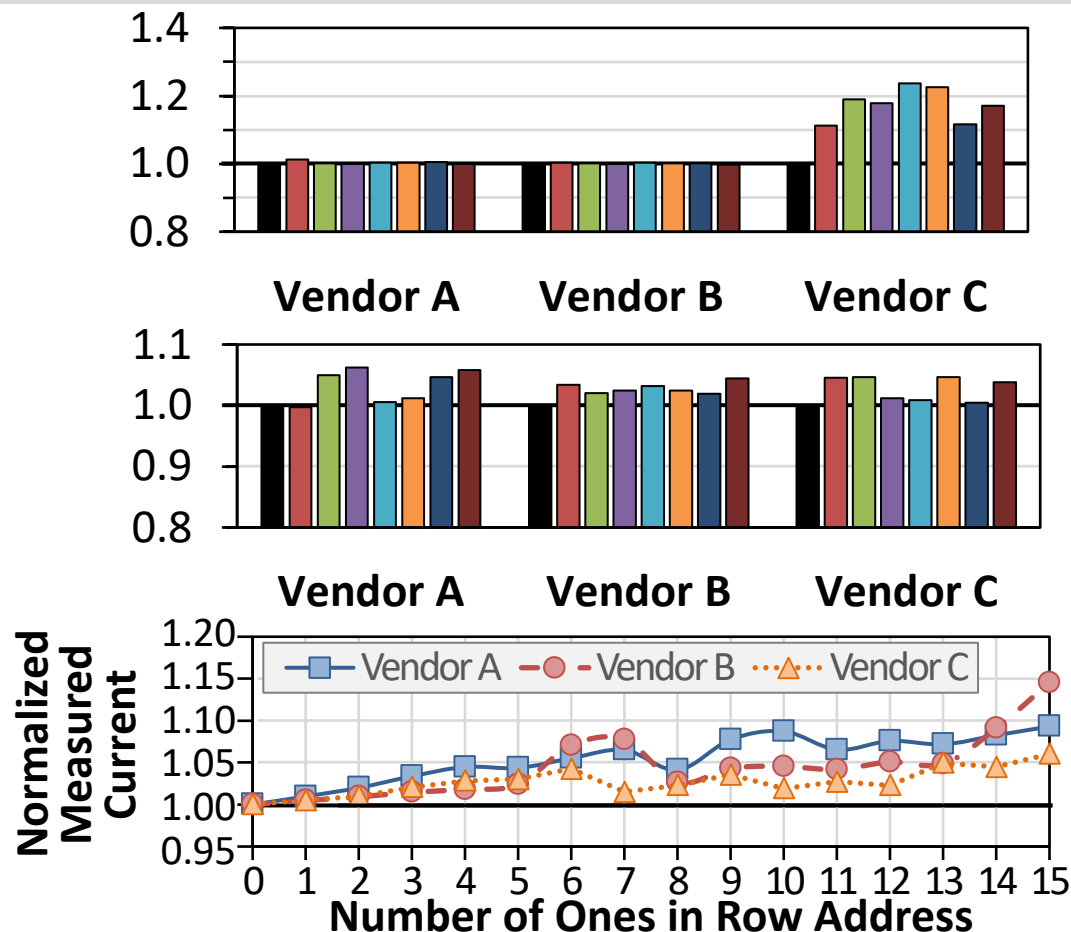
2. DRAM Power is Dependent on Data Values



- Some variation due to infrastructure – can be subtracted
- Without infrastructure variation: up to 230 mA of change
- Toggle affects power consumption, but < 0.15 mA per bit

DRAM power consumption depends *strongly* on the data value

3. Structural Variation Affects DRAM Power Usage **SAFARI**



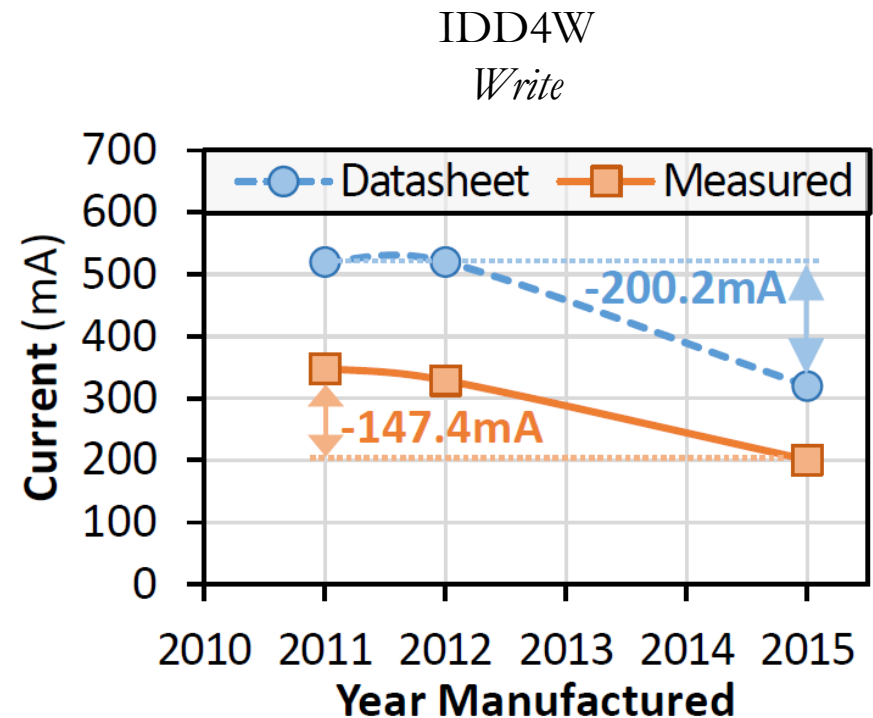
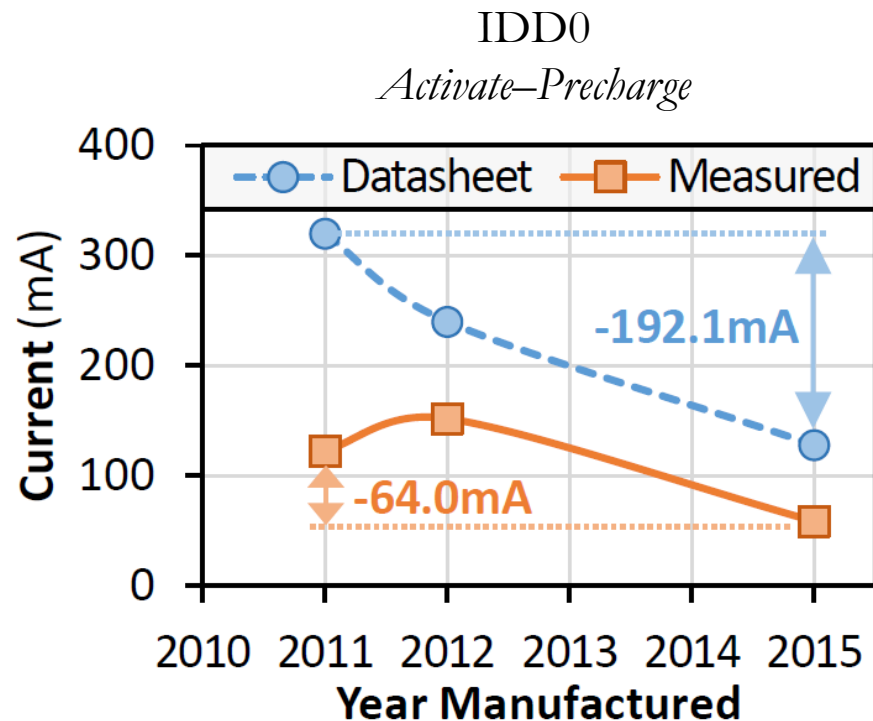
- Vendor C: variation in idle current across banks

- All vendors: variation in read current across banks

- All vendors: variation in activation based on

Significant structural variation:
DRAM power varies systematically by bank and row

4. Generational Savings Are Smaller Than Expected **SAFARI**



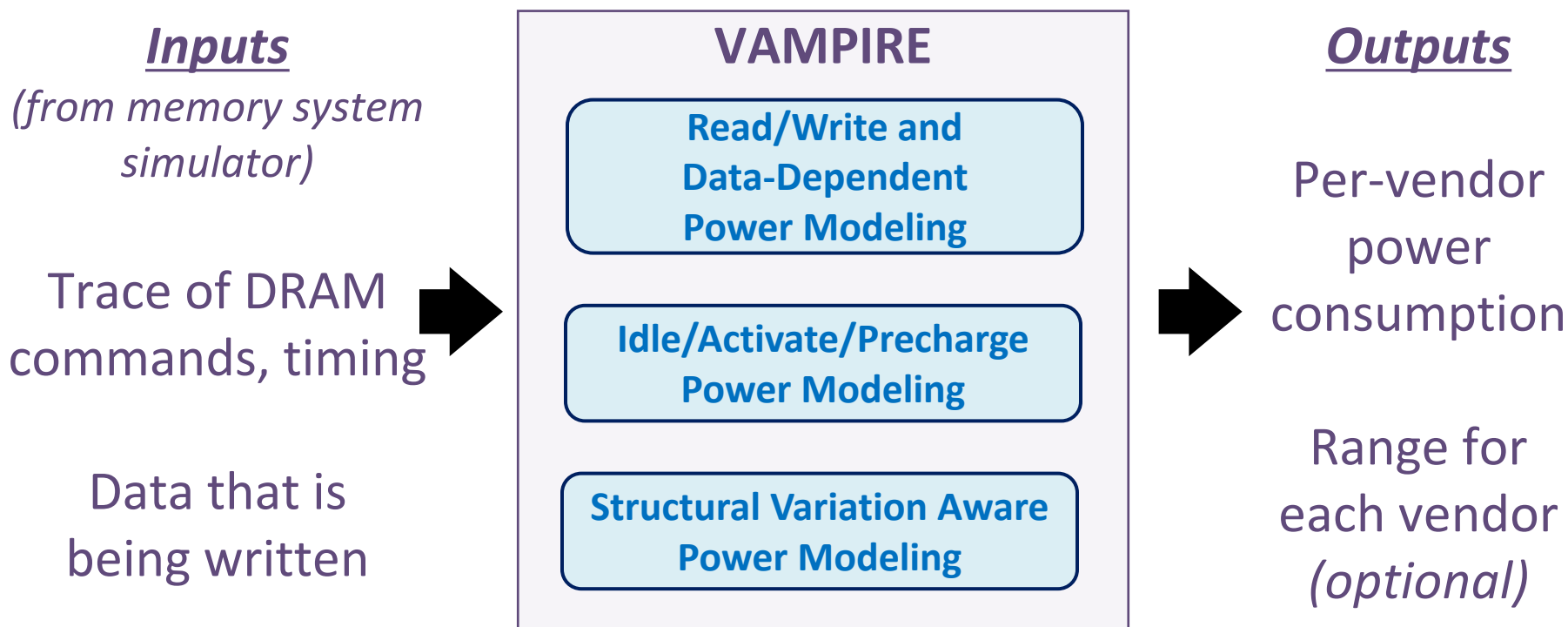
- Similar trends for idle and read currents

Actual power savings of newer DRAM is *much lower* than the savings indicated in the datasheets

1. Real DRAM modules often **consume less power** than vendor-provided IDD values state
2. DRAM power consumption is **dependent on the data value** that is read/written
3. Across banks and rows, **structural variation affects power consumption of DRAM**
4. **Newer DRAM modules save less power** than indicated in datasheets by vendors

Detailed observations and analyses in the paper

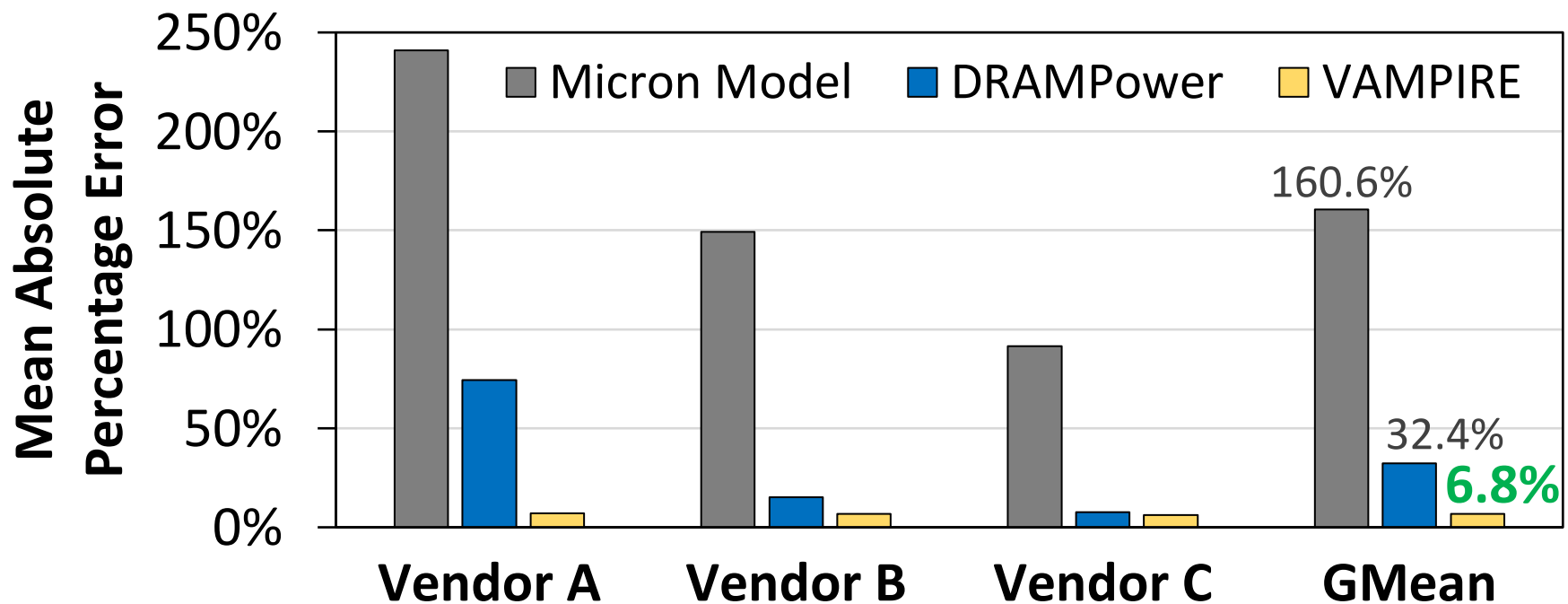
- **VAMPIRE: Variation-Aware model of Memory Power Informed by Real Experiments**



- VAMPIRE and raw characterization data will be open-source: <https://github.com/CMU-SAFARI/VAMPIRE>

VAMPIRE Has Lower Error Than Existing Models **SAFARI**

- Validated using new power measurements:

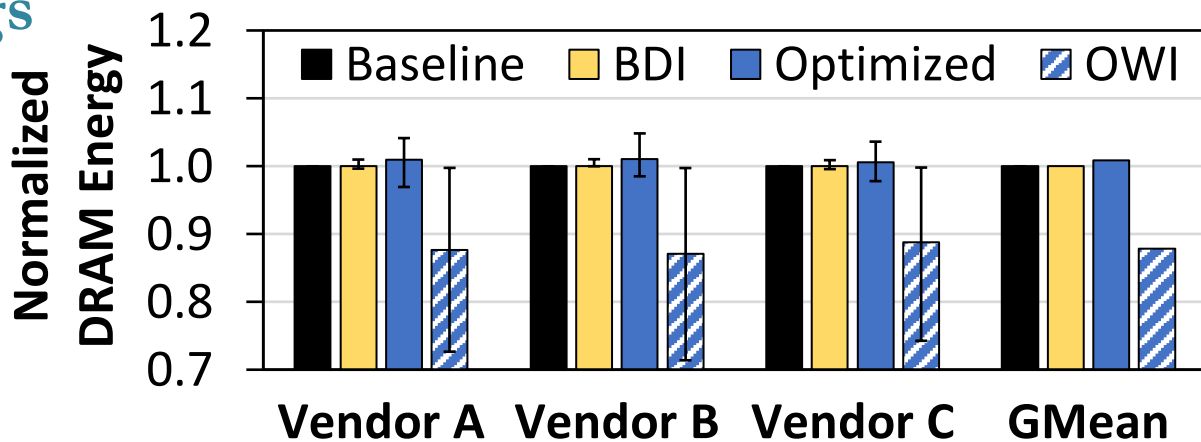


VAMPIRE has very low error for *all* vendors: 6.8%
Much more accurate than prior models

- Taking advantage of structural variation to perform **variation-aware physical page allocation** to reduce power
- Smarter DRAM **power-down scheduling**

- Reducing DRAM energy with **data-dependency-aware cache line encodings**

- 23 applications from the SPEC 2006 benchmark suite
- Traces collected using Pin and Ramulator



- We expect there to be many other new studies in the future

VAMPIRE DRAM Power Model

- Saugata Ghose, A. Giray Yaglikci, Raghav Gupta, Donghyuk Lee, Kais Kudrolli, William X. Liu, Hasan Hassan, Kevin K. Chang, Niladrish Chatterjee, Aditya Agrawal, Mike O'Connor, and Onur Mutlu,
"What Your DRAM Power Models Are Not Telling You: Lessons from a Detailed Experimental Study"

*Proceedings of the ACM International Conference on Measurement and Modeling of Computer Systems (**SIGMETRICS**), Irvine, CA, USA, June 2018.*

[[Abstract](#)]

What Your DRAM Power Models Are Not Telling You: Lessons from a Detailed Experimental Study

Saugata Ghose [†]	Abdullah Giray Yağlıkçı ^{‡†}	Raghav Gupta [†]	Donghyuk Lee [§]
Kais Kudrolli [†]	William X. Liu [†]	Hasan Hassan [‡]	Kevin K. Chang [†]
Niladrish Chatterjee [§]	Aditya Agrawal [§]	Mike O'Connor ^{§¶}	Onur Mutlu ^{‡†}

[†]Carnegie Mellon University

[‡]ETH Zürich

[§]NVIDIA

[¶]University of Texas at Austin

Fundamentally Low Latency Computing Architectures

Fundamentally Low Energy Computing Architectures

More Fundamentally Reducing Latency and Energy

Up Next: Processing In Memory

Computer Architecture

Lecture 11a: Memory Latency, Energy, and Power

Prof. Onur Mutlu

ETH Zürich

Fall 2018

24 October 2018