Understanding Reduced-Voltage Operation in Modern DRAM Devices

Experimental Characterization, Analysis, and Mechanisms

Kevin Chang[†]

A. Giray Yaglikci[†], Saugata Ghose[†], Aditya Agrawal[*], Niladrish Chatterjee[*], Abhijith Kashyap[†], Donghyuk Lee[*], Mike O'Connor[*], Hasan Hassan[‡], Onur Mutlu[†‡]

[†]Carnegie Mellon University

SAFARI

[*]NVIDIA

[‡]ETH Zürich

Executive Summary

- DRAM (memory) power is significant in today's systems
  - Existing low-voltage DRAM reduces voltage **conservatively**
- Goal: Understand and exploit the reliability and latency behavior of real DRAM chips under ***aggressive reduced-voltage operation***
- Key experimental observations:
  - Errors occur and increase with lower voltage
  - Errors exhibit spatial locality
  - Higher operation latency mitigates voltage-induced errors
- **Voltron**: A new DRAM energy reduction mechanism
  - Reduce DRAM voltage **without introducing errors**
  - Use a **regression model** to select voltage that does not degrade performance beyond a chosen target → 7.3% system energy reduction

Outline

- Executive Summary
- **Motivation**
- DRAM Background
- Characterization of DRAM
- Voltron: DRAM Energy Reduction Mechanism
- Conclusion

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?

Outline

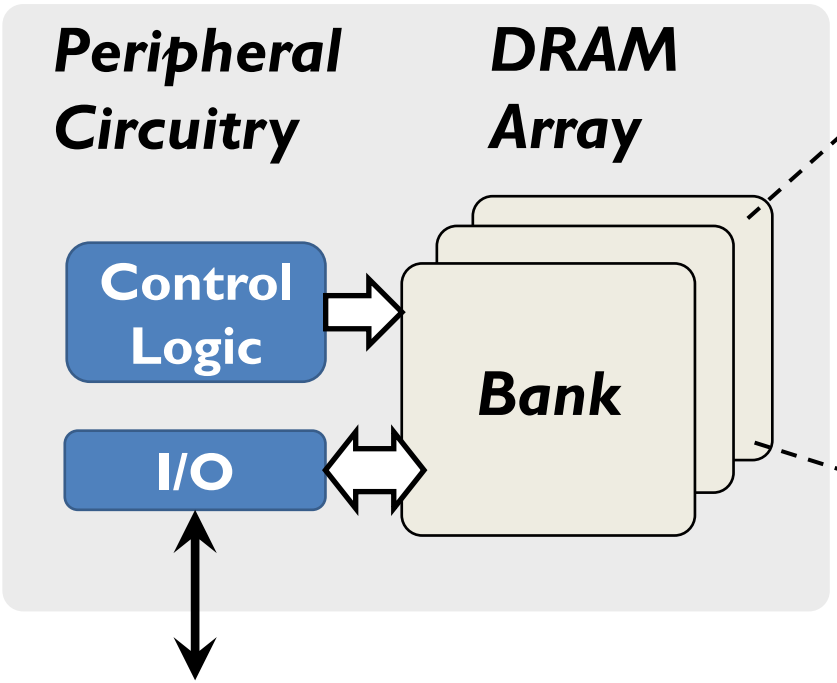
- Executive Summary
- Motivation
- **DRAM Background**
- Characterization of DRAM
- Voltron: DRAM Energy Reduction Mechanism
- Conclusion

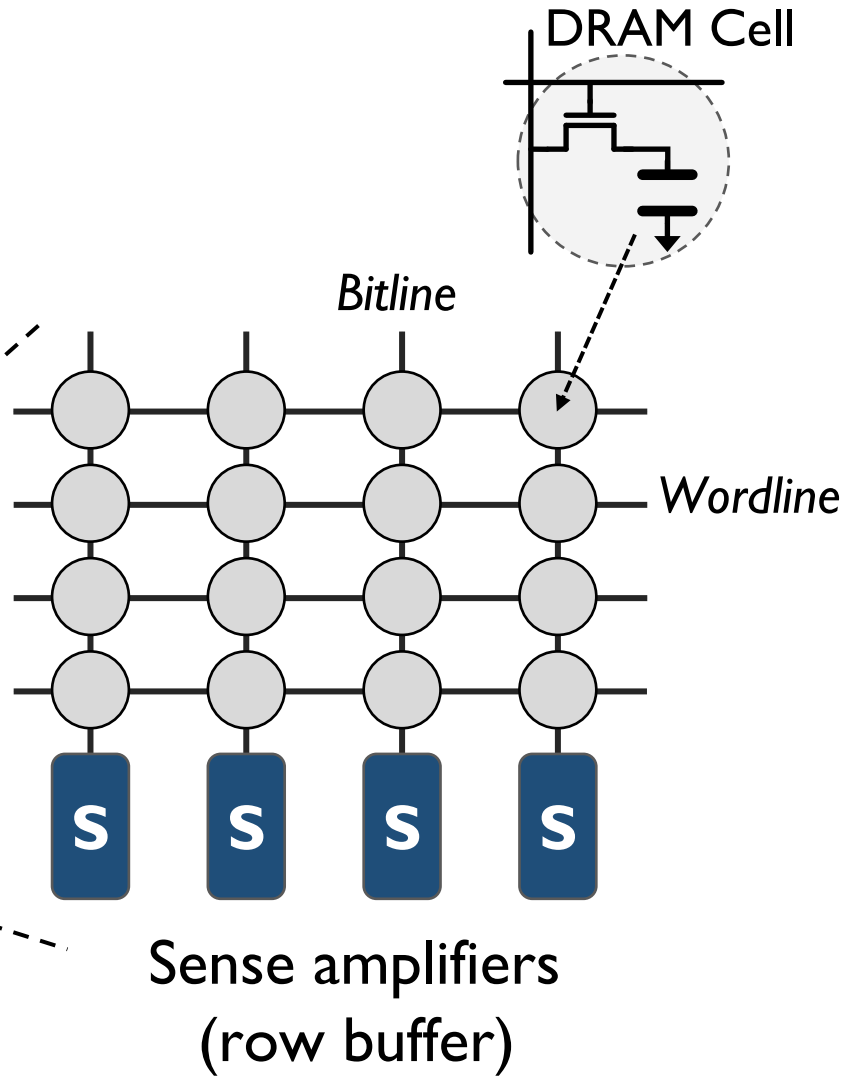
High-Level DRAM Organization

DRAM
Channel

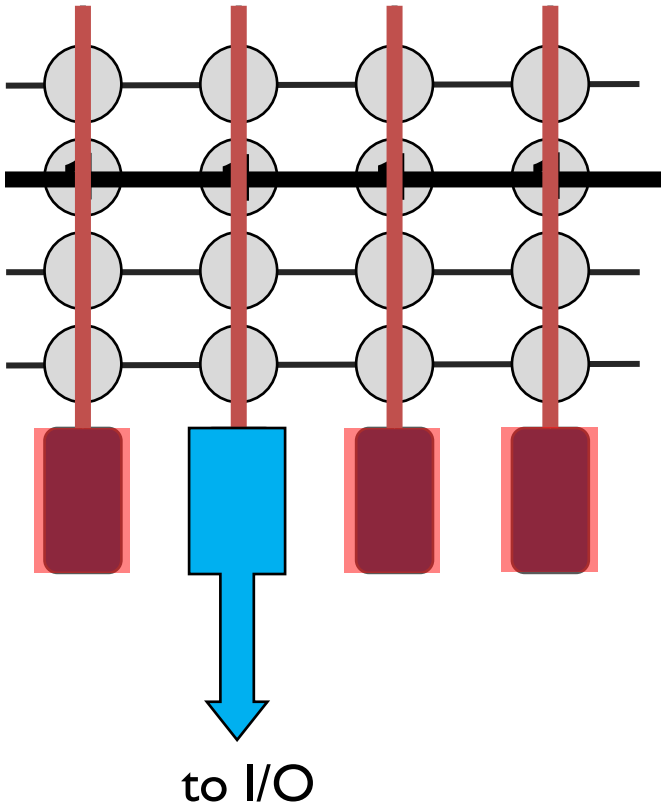
DRAM Module

DRAM Chip Internals





DRAM Operations

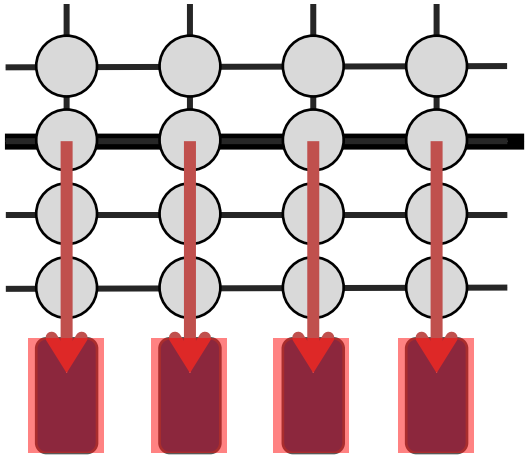


1 **ACTIVATE**: Store the row into the **row buffer**

2 **READ**: Select the target cache line and drive to CPU

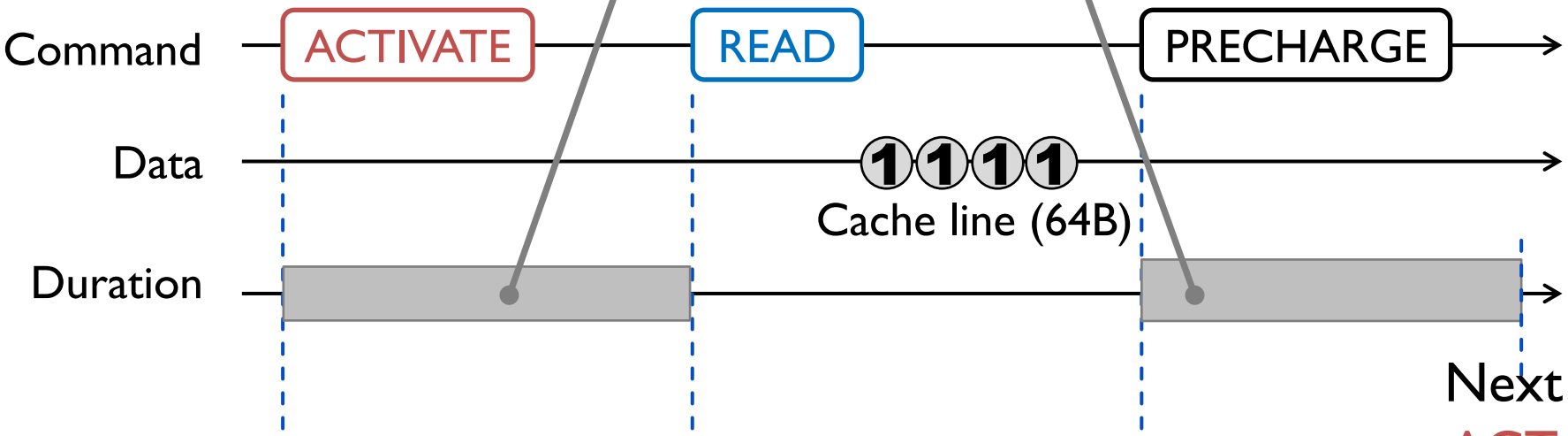
3 **PRECHARGE**: Prepare the array for a new ACTIVATE

DRAM Access Latency



1 **Activation latency**
(13ns / 50 cycles)

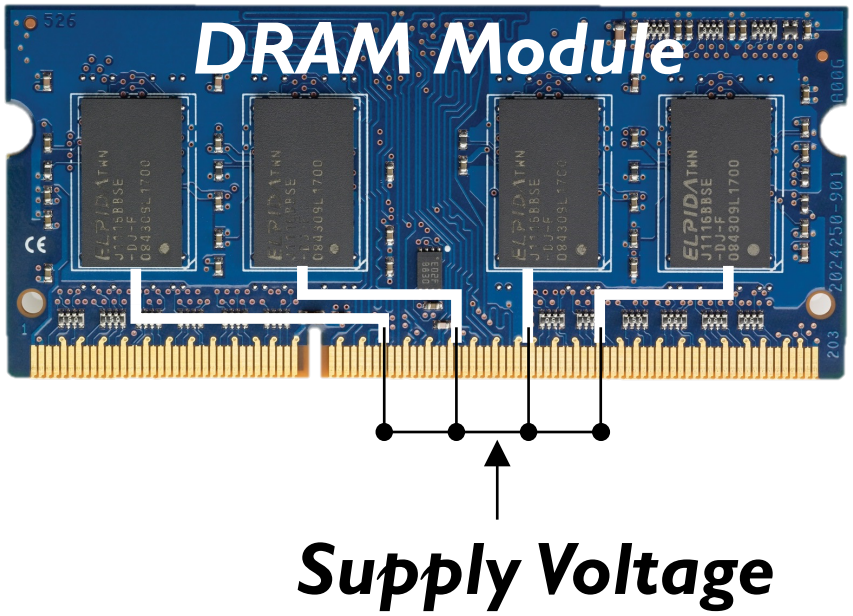
2 **Precharge latency**
(13ns / 50 cycles)



Outline

- Executive Summary
- Motivation
- DRAM Background
- **Characterization of DRAM**
  - Experimental methodology
  - Impact of voltage on reliability and latency
- Voltron: DRAM Energy Reduction Mechanism
- Conclusion

Supply Voltage Control on DRAM



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

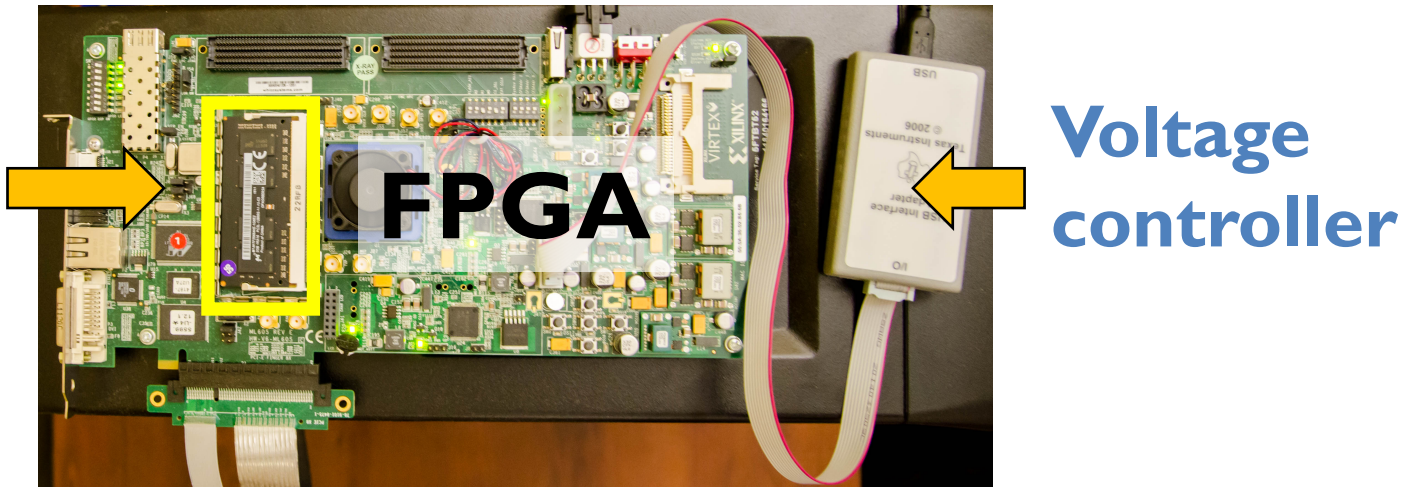
Custom Testing Platform

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

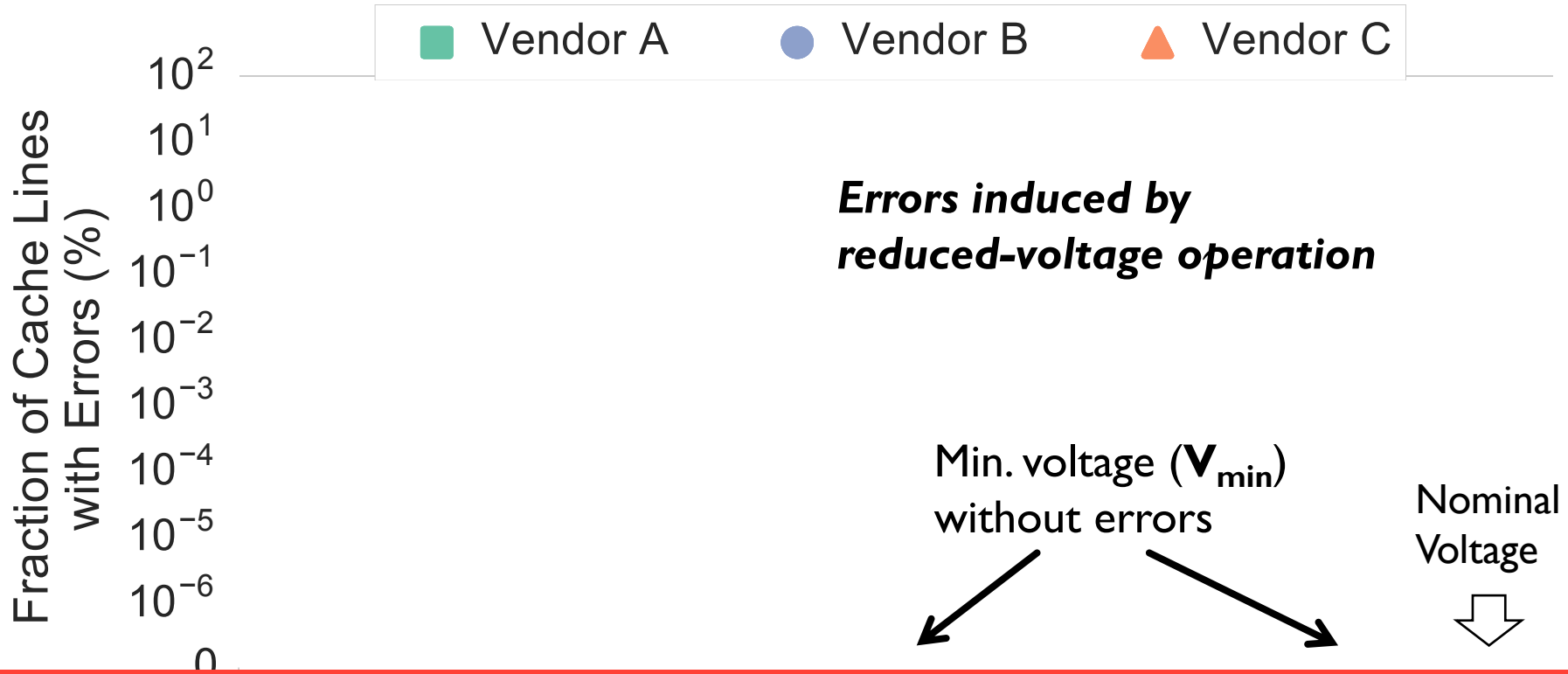
- **124 DDR3L** (low-voltage) DRAM chips
  - **31 SO-DIMMs**
  - **1.35V** (DDR3 uses 1.5V)
  - Density: 4Gb per chip
  - Three major vendors/manufacturers
  - Manufacturing dates: 2014-2016

- Iteratively read every bit in each 4Gb chip under a wide range of supply voltage levels: 1.35V to 1.0V (**-26%**)

Outline

- Executive Summary
- Motivation
- DRAM Background
- **Characterization of DRAM**
  - Experimental methodology
  - Impact of voltage on reliability and latency
- Voltron: DRAM Energy Reduction Mechanism
- Conclusion

Reliability Worsens with Lower Voltage

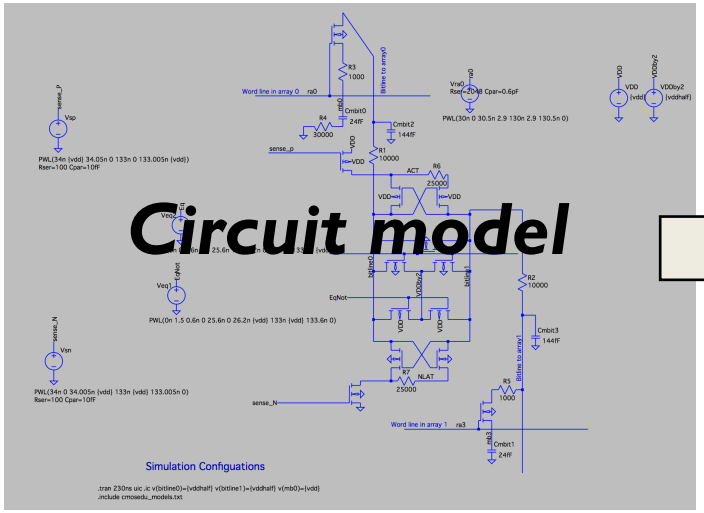


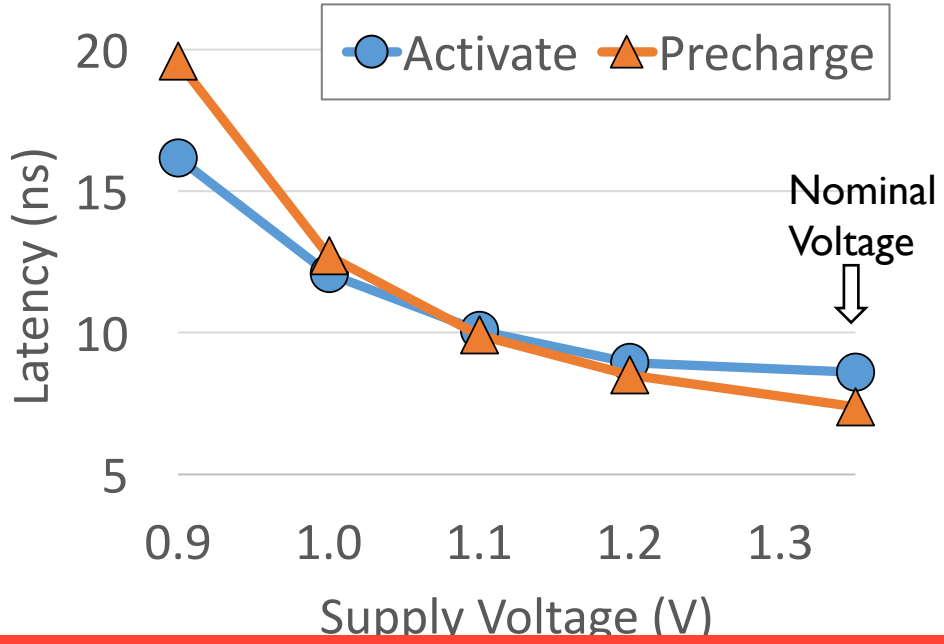
Reducing voltage below V_{min} causes an increasing number of errors

Source of Errors

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

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

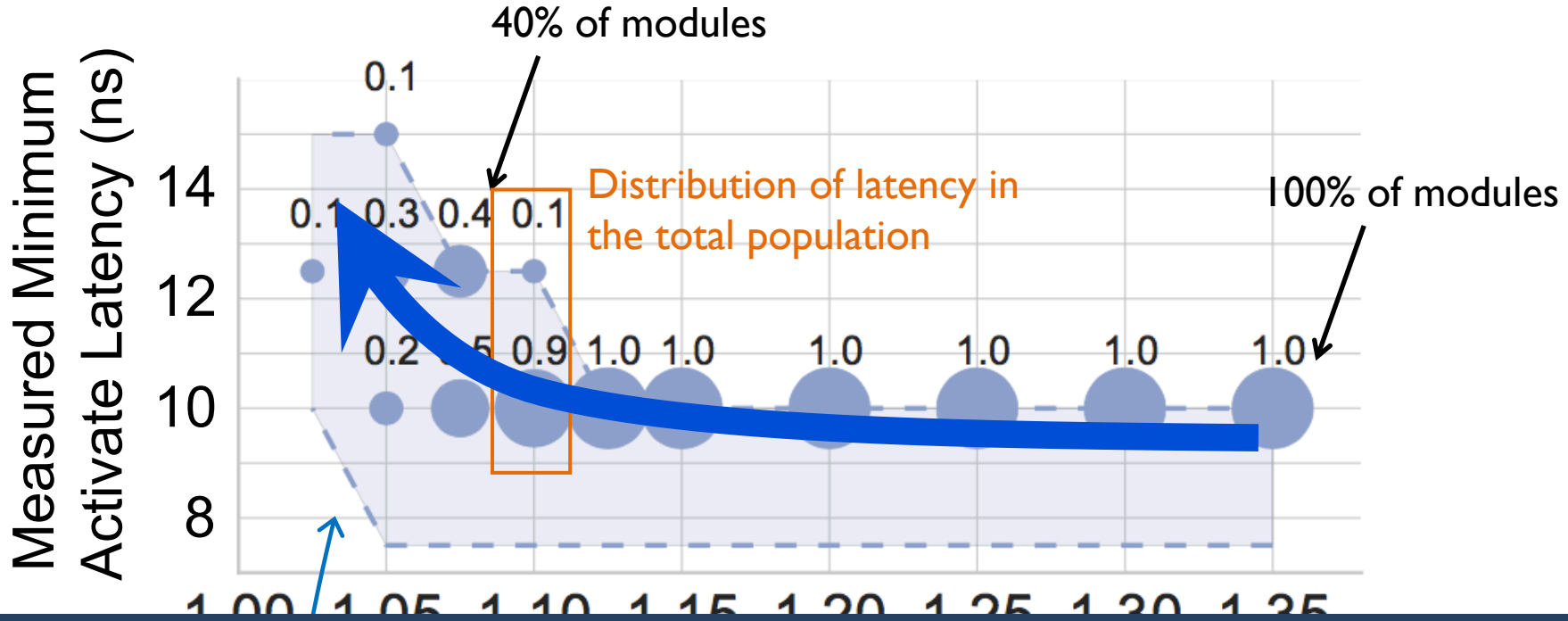




Reliable low-voltage operation requires higher latency

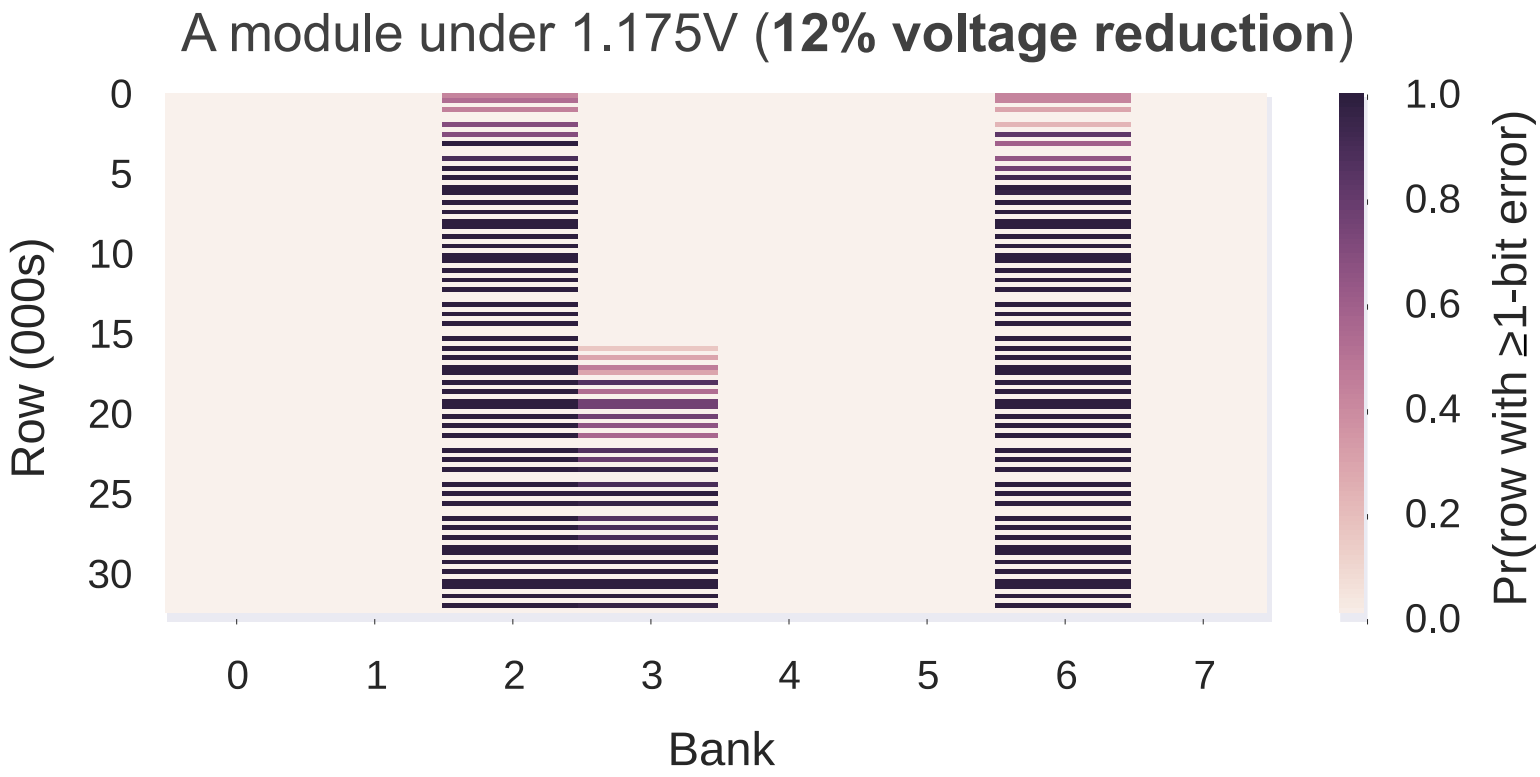
DIMMs Operating at Higher Latency

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

Other Results in the Paper

- Error-Correcting Codes (ECC)
  - ECC (SECDED) is **not** sufficient to mitigate the errors
- Effect of temperature
  - Higher temperature requires higher latency under some voltage levels
- Data retention time
  - Lower voltage does **not** require more frequent refreshes
- Effect of stored data pattern on error rate
  - Difference is **not** statistically significant to draw conclusion

Summary of Key Experimental Observations

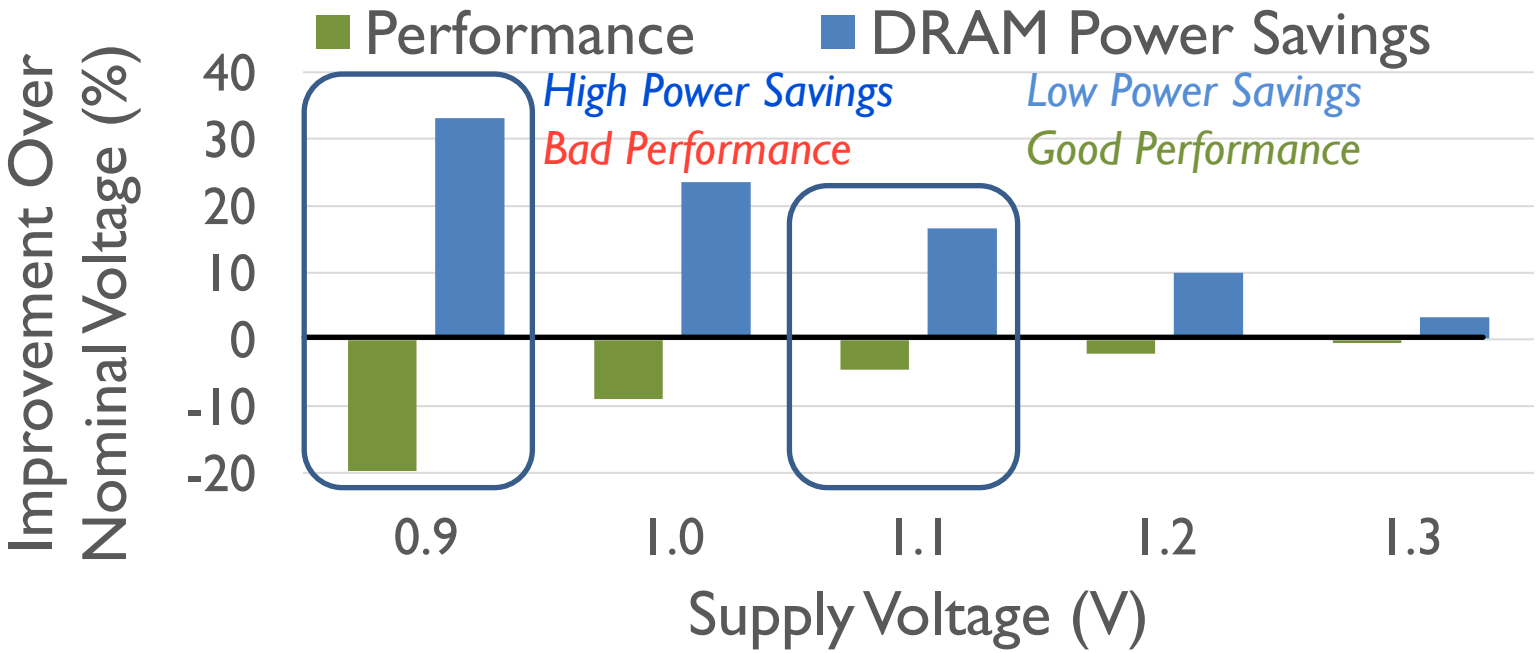
- Voltage-induced errors increase as voltage reduces further below V_{min}
- Errors exhibit spatial locality
- Increasing the latency of DRAM operations mitigates voltage-induced errors

Outline

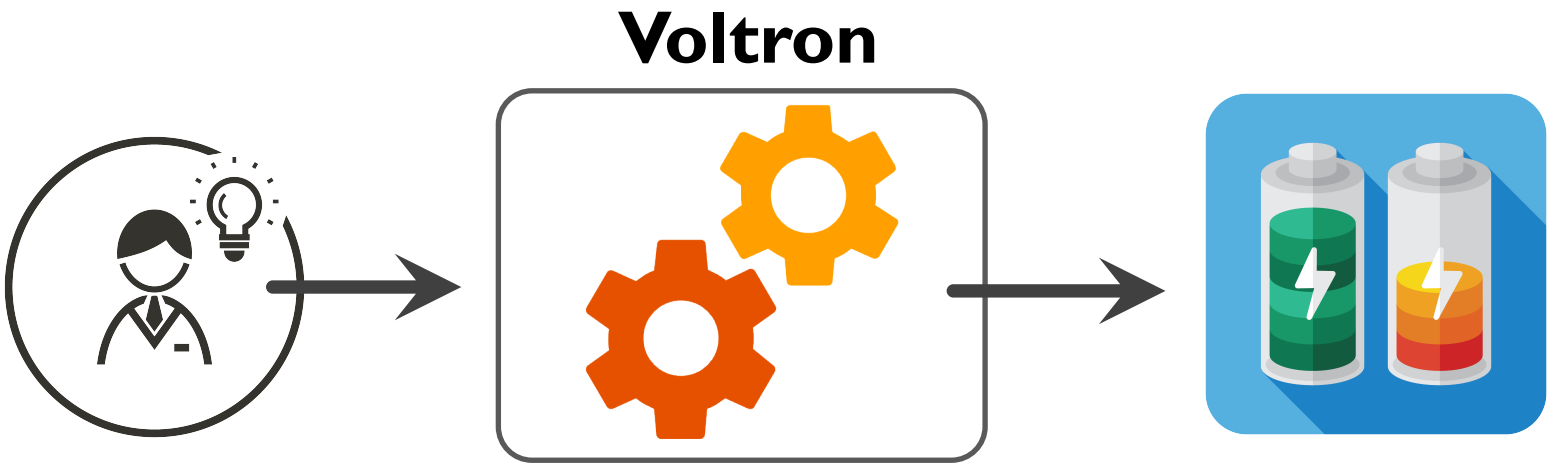
- Executive Summary
- Motivation
- DRAM Background
- Characterization of DRAM
  - Experimental methodology
  - Impact of voltage on reliability and latency
- **Voltron: DRAM Energy Reduction Mechanism**
- Conclusion

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

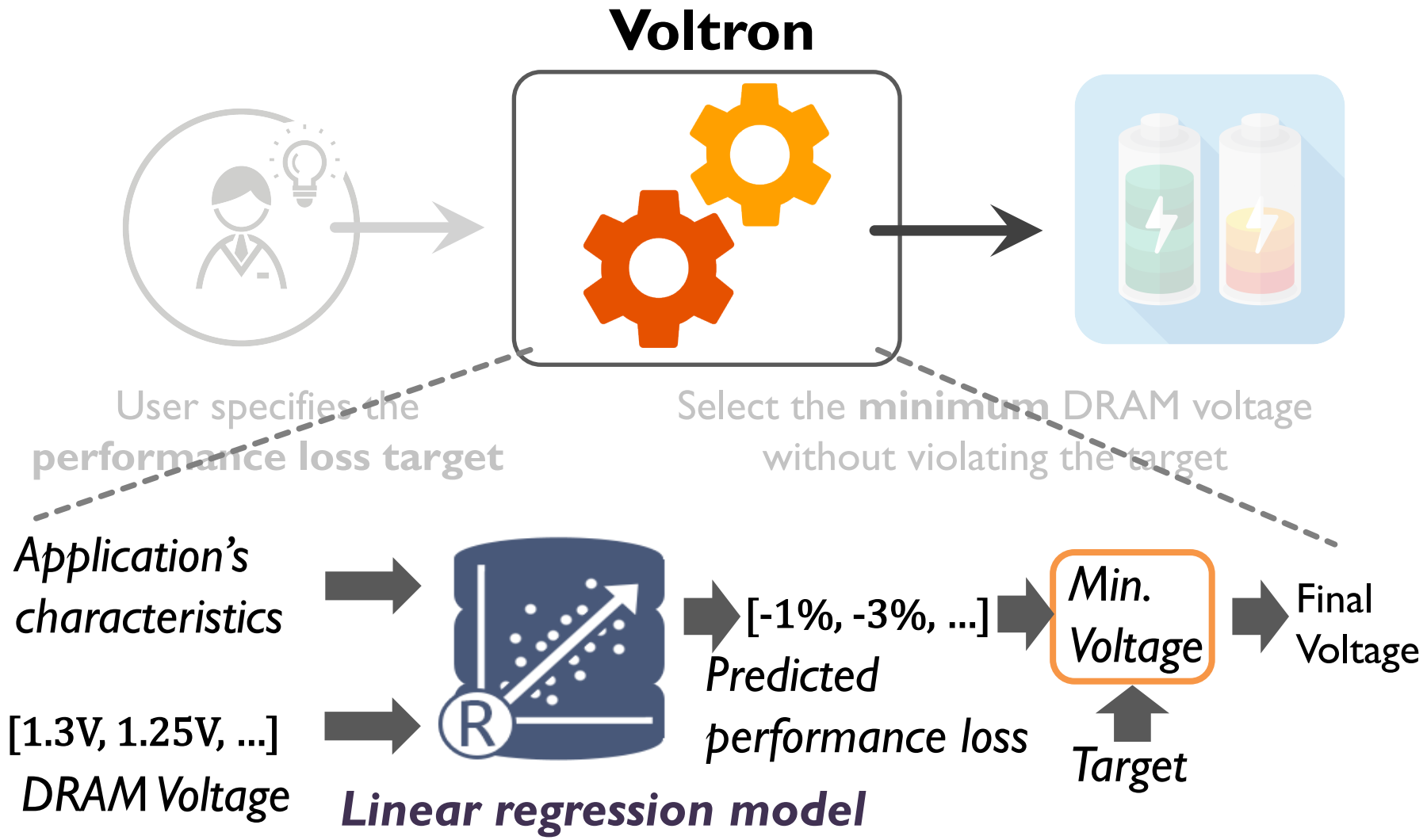


User specifies the **performance loss target**

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



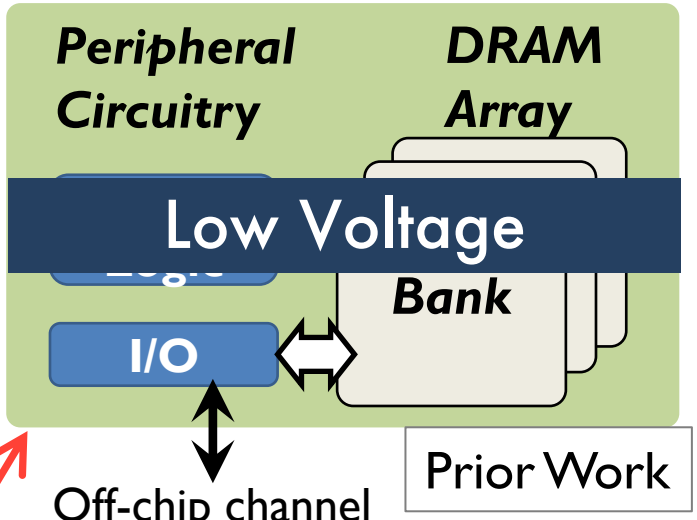
Linear Model to Predict Performance

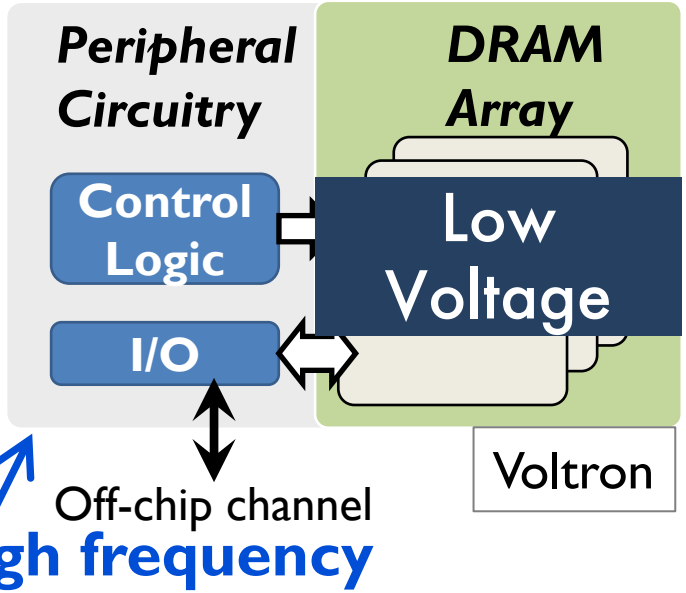
- Application's characteristics for the model:
  - ***Memory intensity***: Frequency of last-level cache misses
  - ***Memory stall time***: Amount of time memory requests stall commit inside CPU

- 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

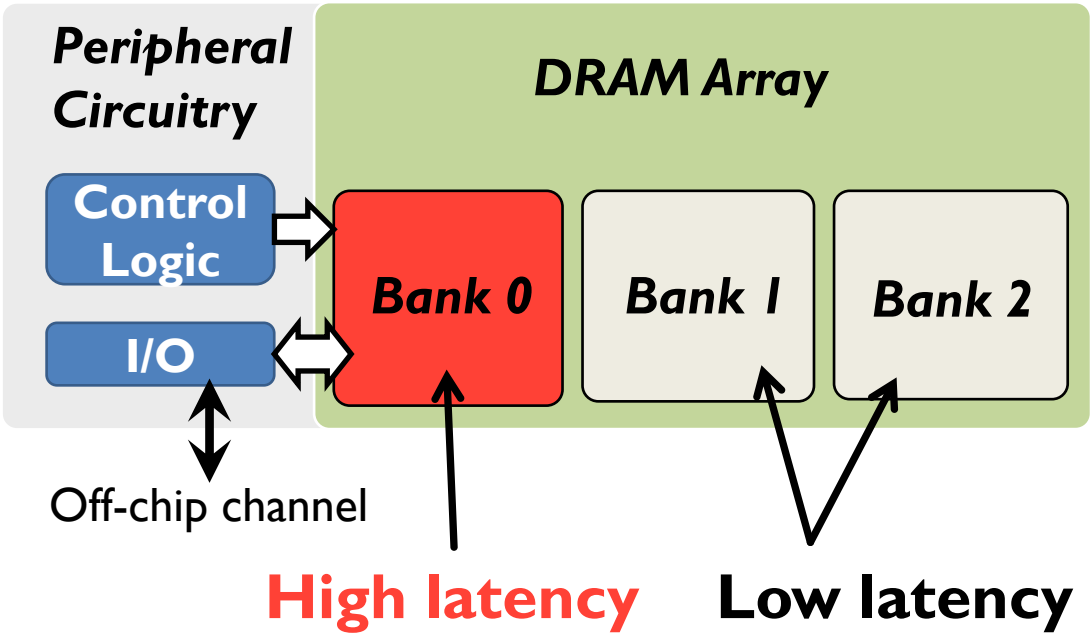




Exploiting Spatial Locality of Errors

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

- Benefit: Higher performance



Outline

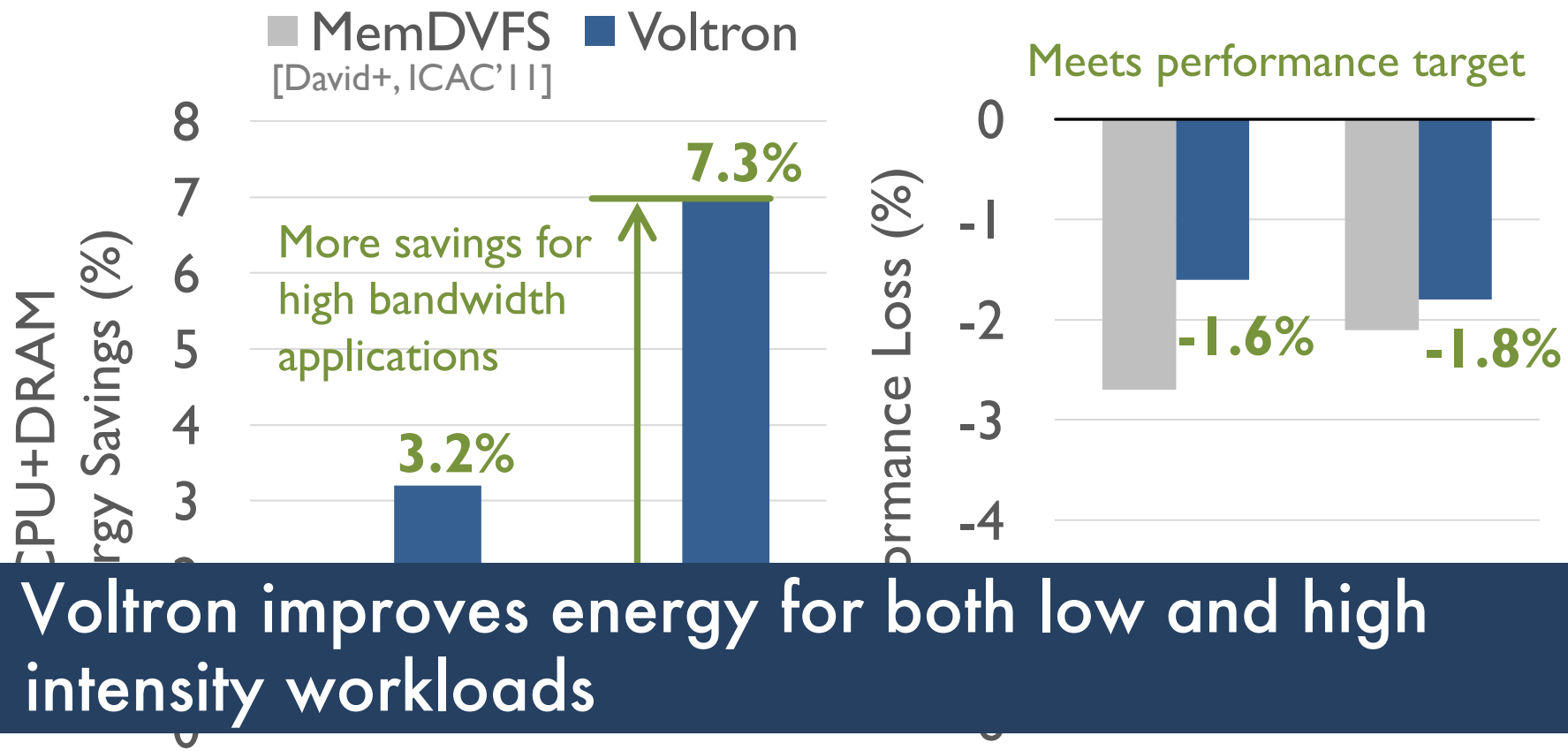
- Executive Summary
- Motivation
- DRAM Background
- Characterization of DRAM
  - Experimental methodology
  - Impact of voltage on reliability and latency
- **Voltron: DRAM Energy Reduction Mechanism**
  - Evaluation
- Conclusion

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



1. Voltron improves energy for both low and high intensity workloads

2. Voltron satisfies the performance loss target via a regression model

Outline

- Executive Summary
- Motivation
- DRAM Background
- Characterization of DRAM
  - Experimental methodology
  - Impact of voltage on reliability and latency
- Voltron: DRAM Energy Reduction Mechanism
- **Conclusion**

Conclusion

- DRAM (memory) power is significant in today's systems
  - Existing low-voltage DRAM reduces voltage **conservatively**
- Goal: Understand and exploit the reliability and latency behavior of real DRAM chips under ***aggressive reduced-voltage operation***
- Key experimental observations:
  - Errors occur and increase with lower voltage
  - Errors exhibit spatial locality
  - Higher operation latency mitigates voltage-induced errors
- **Voltron**: A new DRAM energy reduction mechanism
  - Reduce DRAM voltage **without introducing errors**
  - Use a **regression model** to select voltage that does not degrade performance beyond a chosen target → 7.3% system energy reduction

Understanding Reduced-Voltage Operation in Modern DRAM Devices

Experimental Characterization, Analysis, and Mechanisms

Kevin Chang[†]

A. Giray Yaglikci[†], Saugata Ghose[†], Aditya Agrawal[*], Niladrish Chatterjee[*], Abhijith Kashyap[†], Donghyuk Lee[*], Mike O'Connor[*], Hasan Hassan[‡], Onur Mutlu[†‡]

[†]Carnegie Mellon University

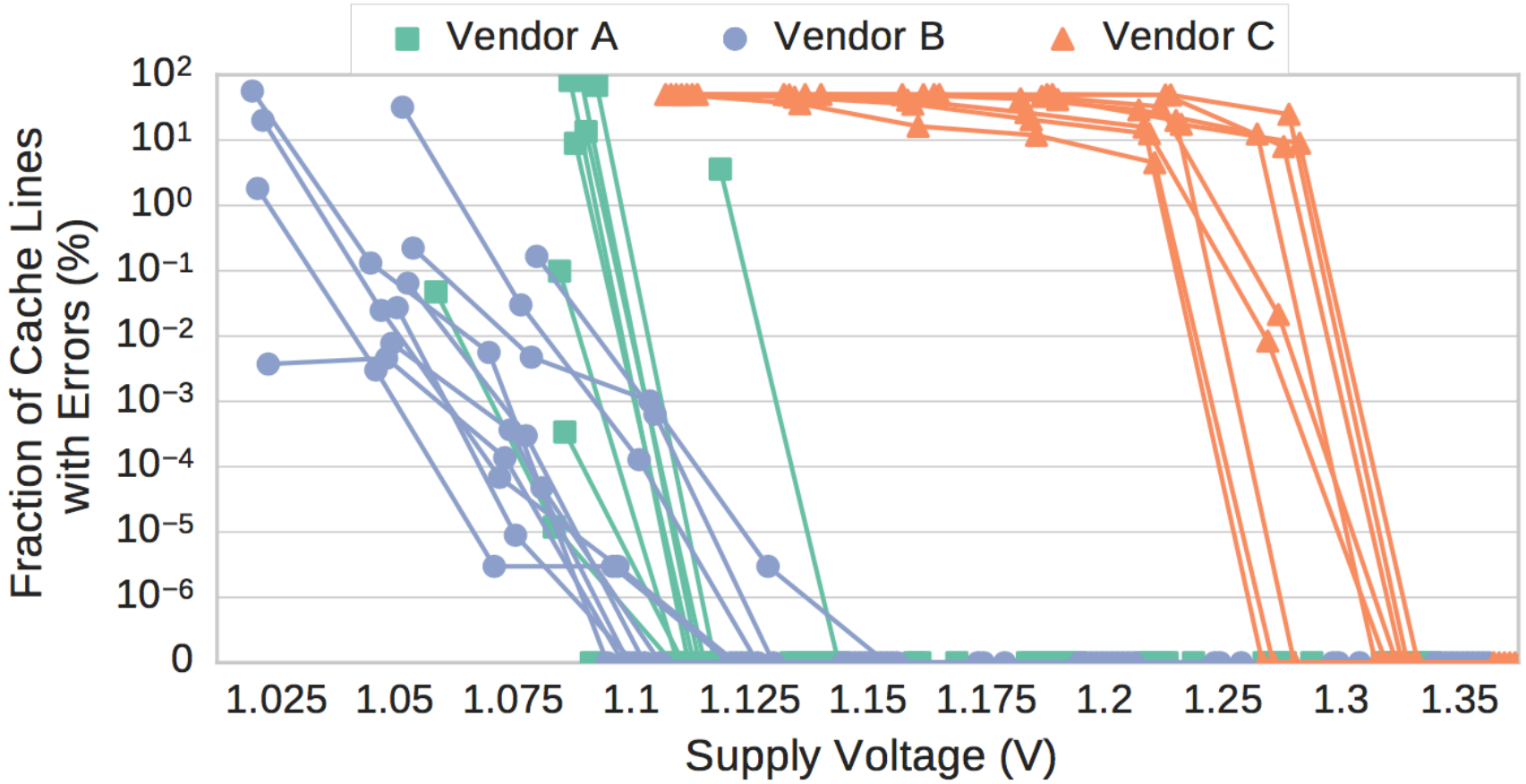
SAFARI

[*]NVIDIA

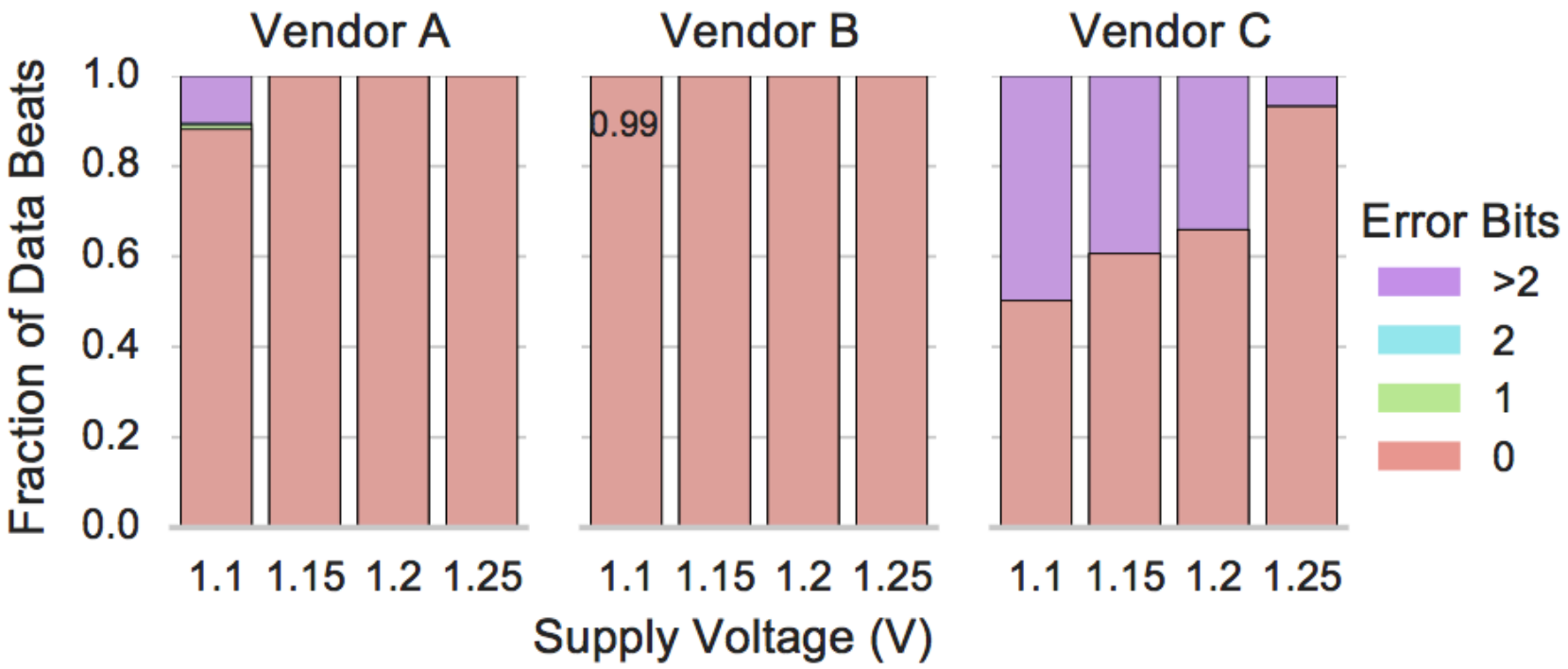
[‡]ETH Zürich

BACKUP

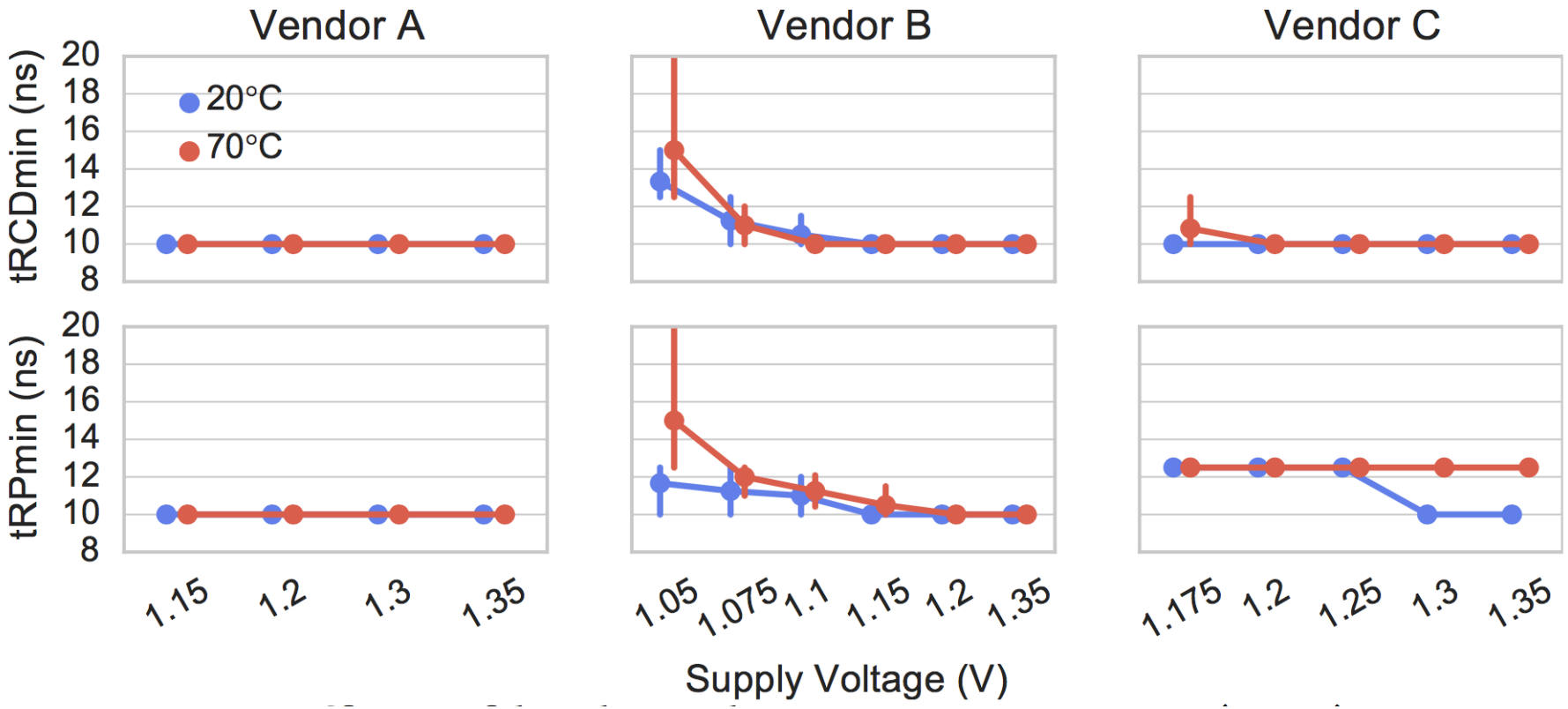
Errors Rates Across Modules



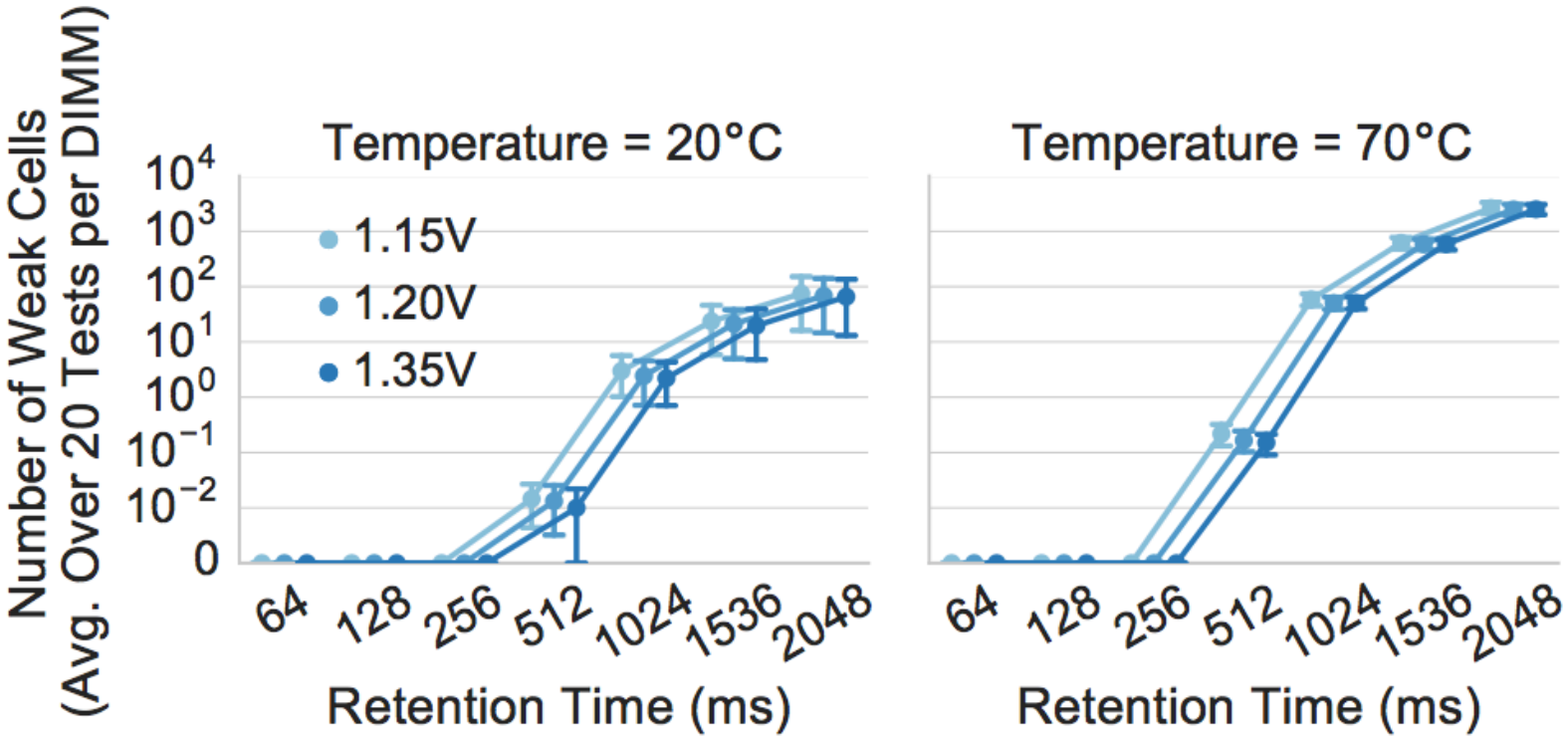
Error Density



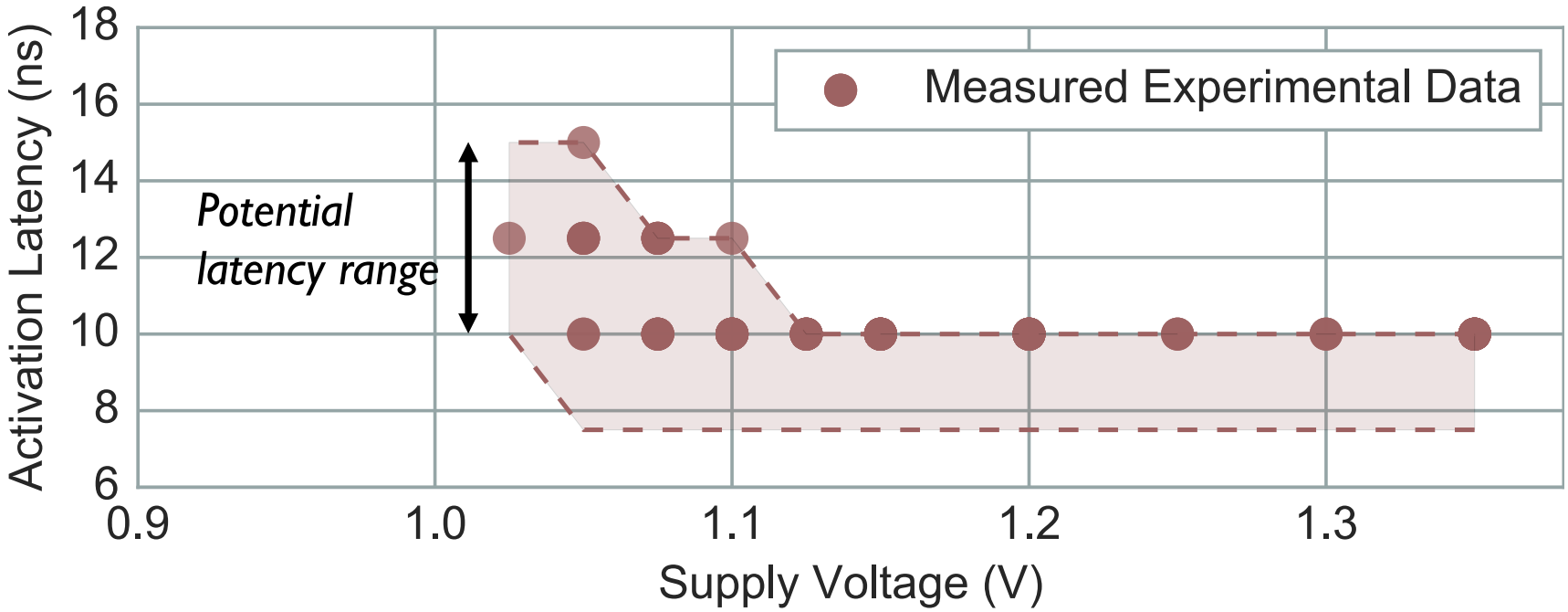
Temperature Impact



Impact on Retention Time



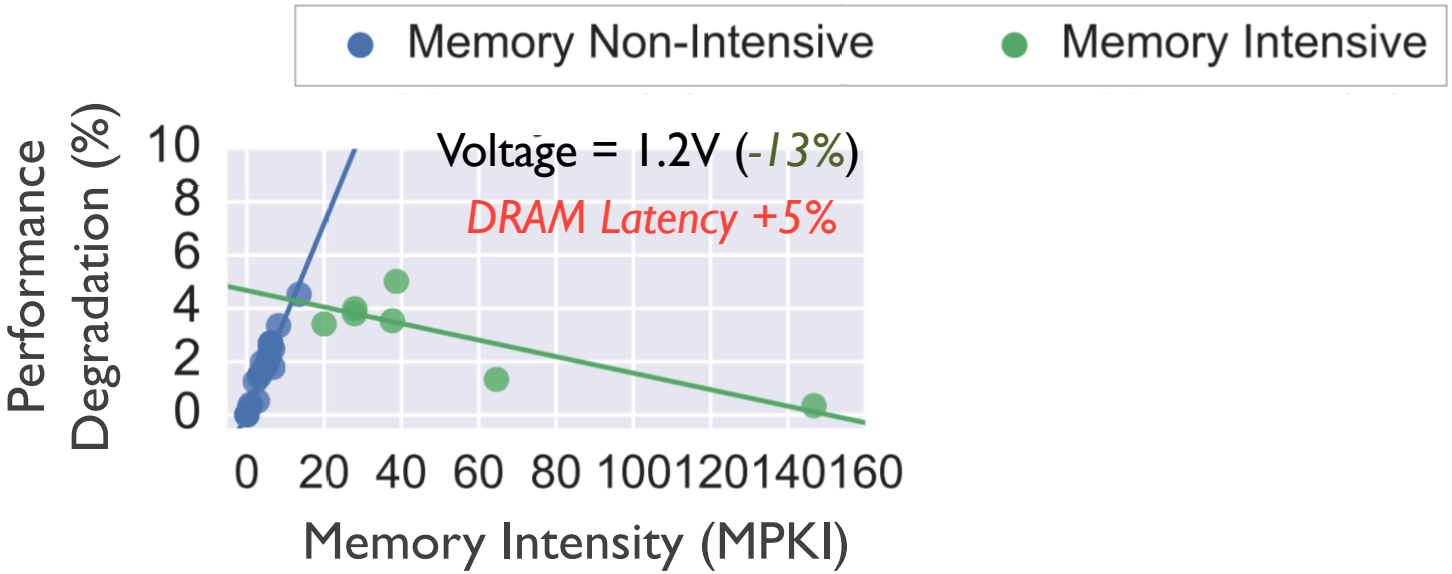
Derivation of More Precise Latency



DRAM circuit model validates our experimental results and provides more precise latency

Performance Loss Correlation

- Observation: Application's performance loss due to higher latency has a strong linear relationship with its memory intensity



MPKI = Last-level cache Misses Per Thousand Instruction

Performance-Aware Voltage Adjustment

- Build a performance (linear-regression) model to predict performance loss based on the selected voltage

$$PredictedLoss = \theta_0 + \theta_1 Latency + \theta_2 App.Itensity + \theta_3 App.StallTime$$

Latency due to voltage adjustment

The running application's characteristics

- $\boldsymbol{\theta}$s are trained through 151 application samples
- Use the model to select a minimum voltage that satisfies a performance loss target specified by the user

Linear Model Accuracy

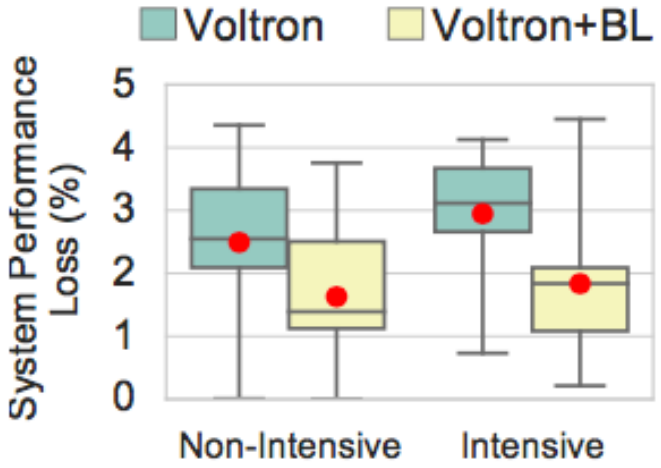
- R^2 = 0.75 / 0.9 for low and high intensity workloads
- RMSE = 2.8 / 2.5 for low and high intensity workloads

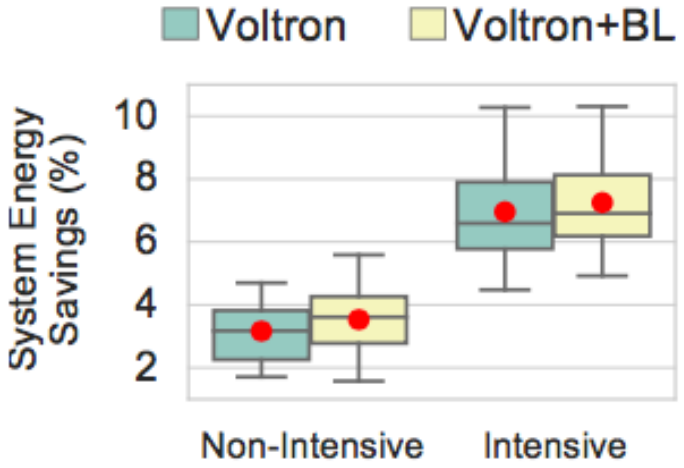
Dynamic Voltron

Algorithm 1 Array Voltage Selection

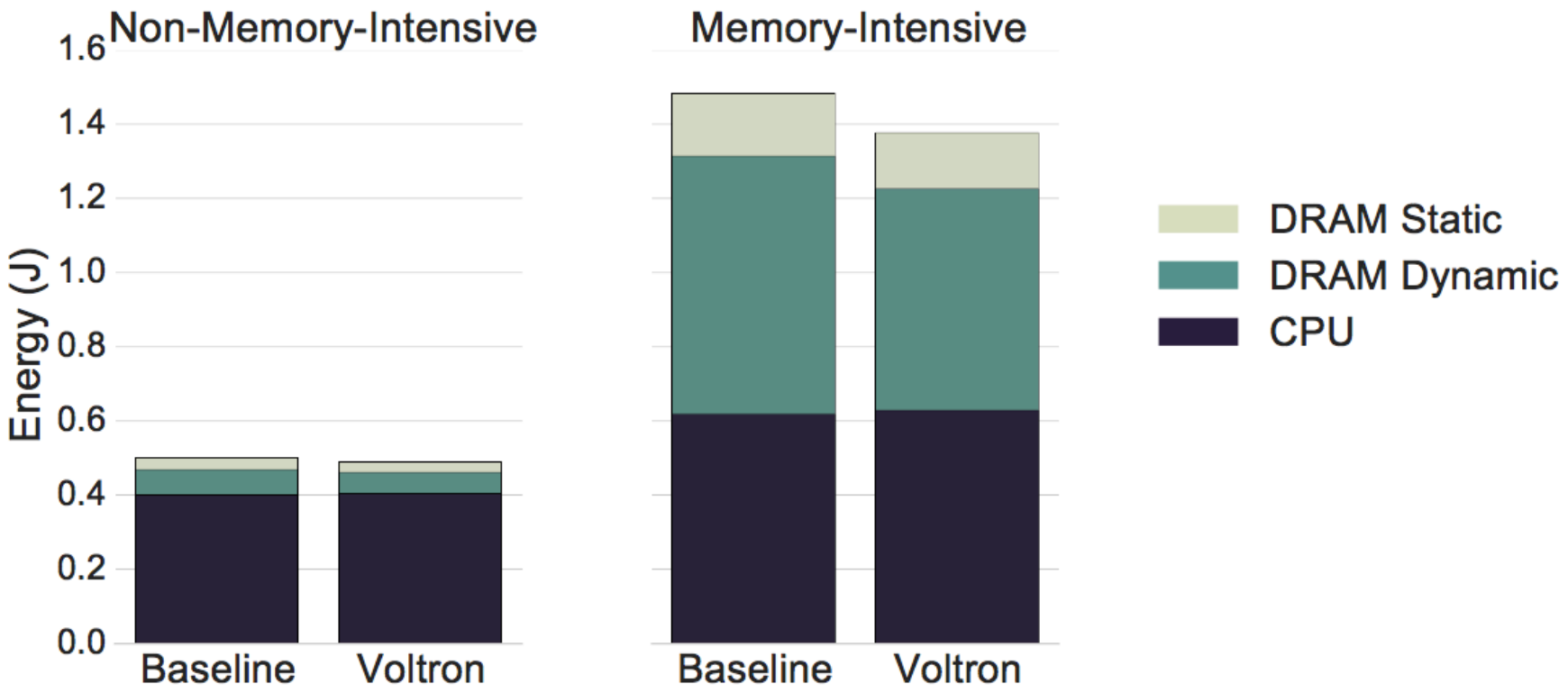
```
1  SelectArrayVoltage(target_loss)
2    for each interval                                   ▷ Enter at the end of an interval
3      profile = GetMemoryProfile()
4      NextV_array = 1.35
5      for V_array ← 0.9 to 1.3          ▷ Search for the smallest V_array that satisfies the performance loss target
6        predicted_loss = Predict(Latency(V_array), profile.MPKI, profile.StallTime)   ▷ Predict performance loss
7        if predicted_loss ≤ target_loss then              ▷ Compare the predicted loss to the target
8          NextV_array = V_array                ▷ Use the current V_array for the next interval
9          break
10     ApplyVoltage(NextV_array)                ▷ Apply the new V_array for the next interval
```

Effect of Exploiting Error Locality

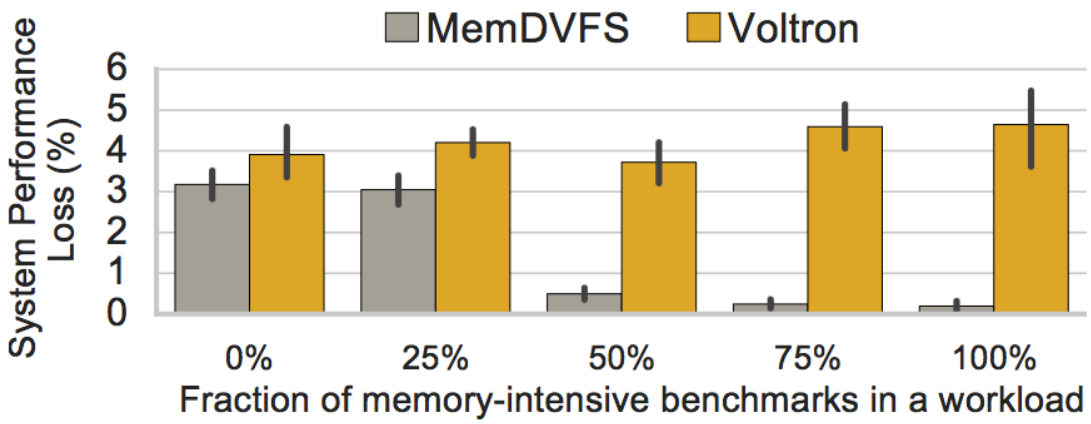


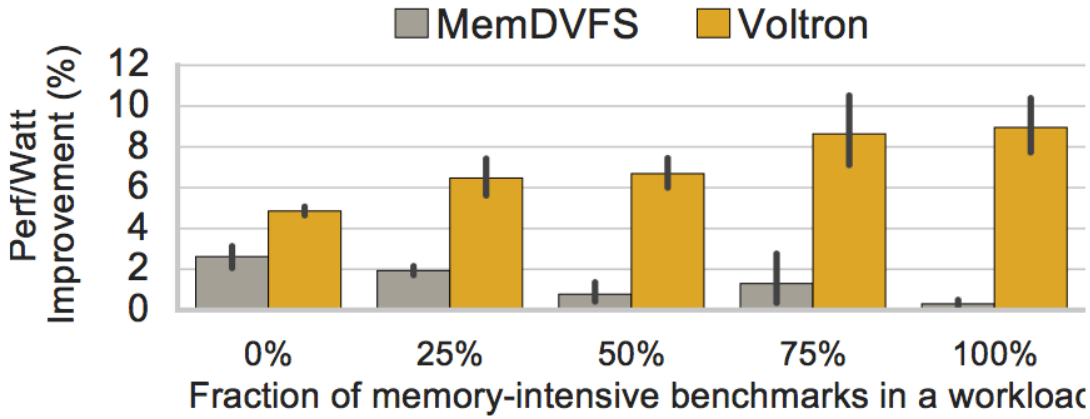


Energy Breakdown

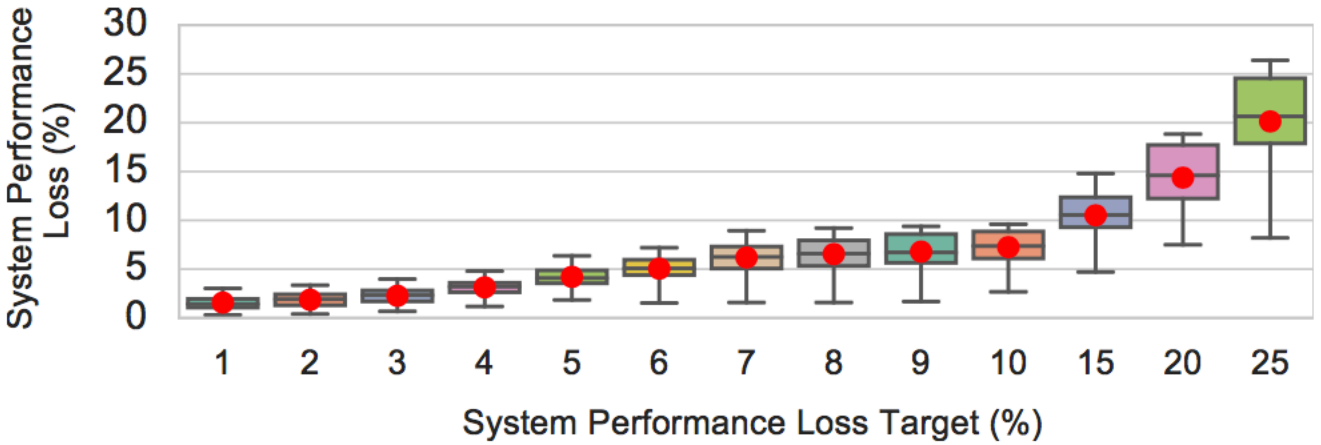


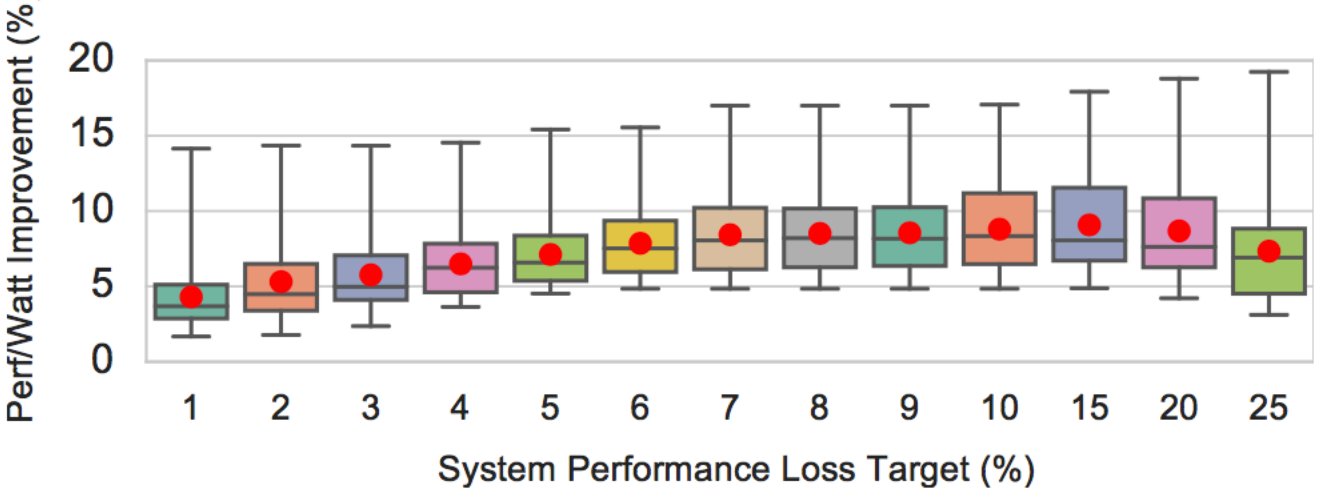
Heterogeneous Workloads





Performance Target Sweep





Sensitivity to Profile Interval Length