Understanding Latency Variation in Modern DRAM Chips

Experimental Characterization, Analysis, and Optimization

Kevin Chang

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

Carnegie Mellon University

TOBB ETÜ
TOBB Ekonomi ve Teknoloji Üniversitesi

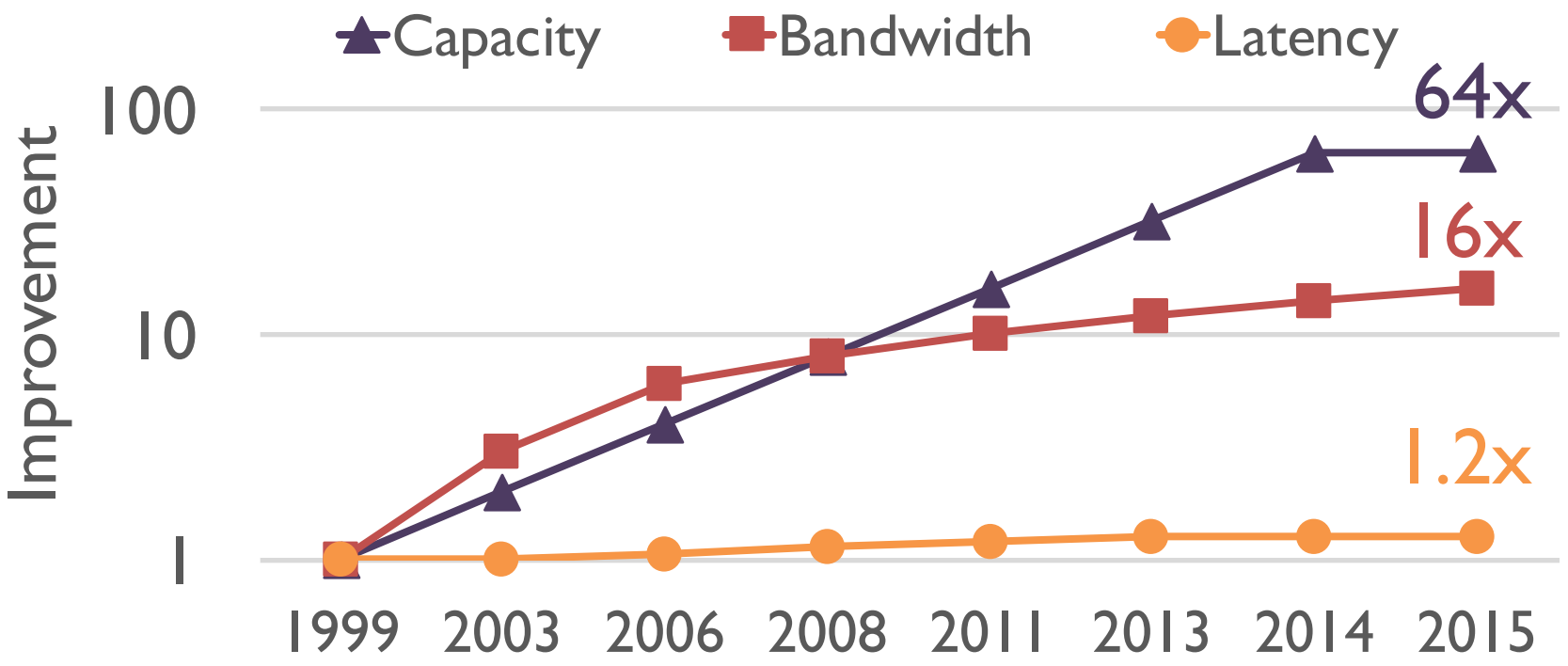
University of Virginia

Peking University

ETH zürich

v1.3

Main Memory Latency Lags Behind

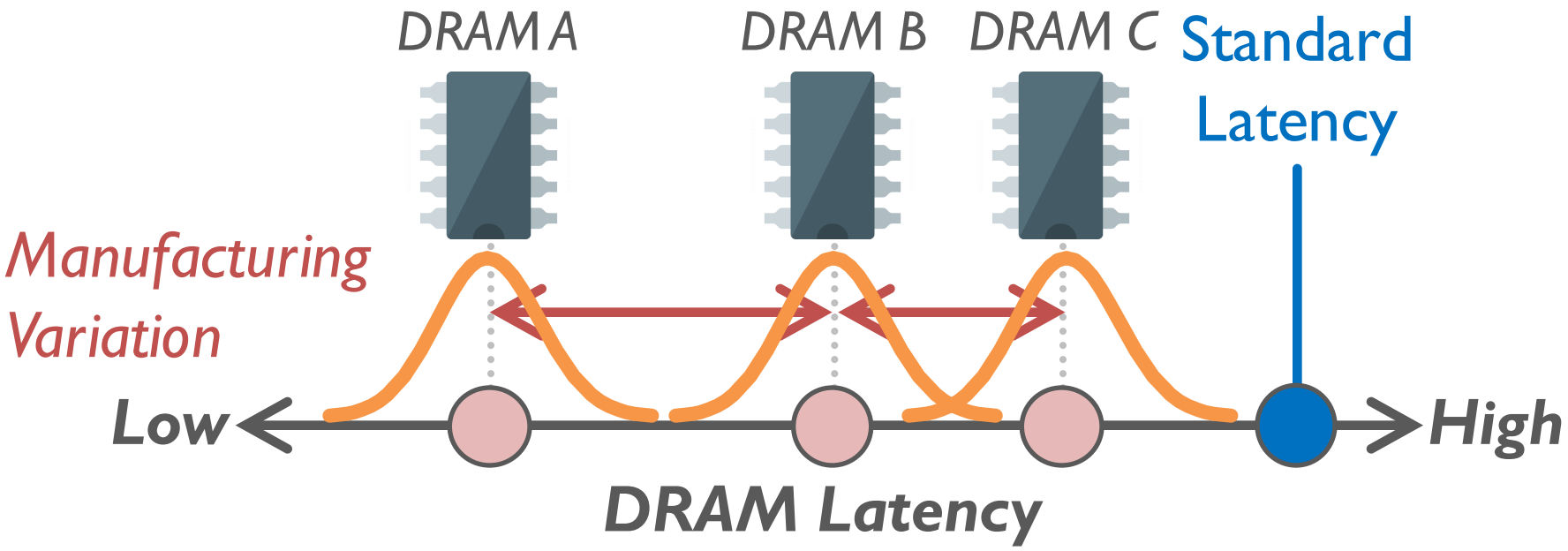


Long DRAM latency → performance bottleneck

In-memory DB, Spark, JVM, … [Clapp+ (**Intel**), IISWC'15]
Google warehouse-scale workloads [Kanev+ (**Google**), ISCA'15]

Why is Latency High?

- DRAM latency: Delay as specified in DRAM standards
  - Doesn't reflect true DRAM device latency
- Imperfect manufacturing process → latency variation
- **High standard latency** chosen to increase yield



Goals

1 Understand and characterize latency variation in modern DRAM chips

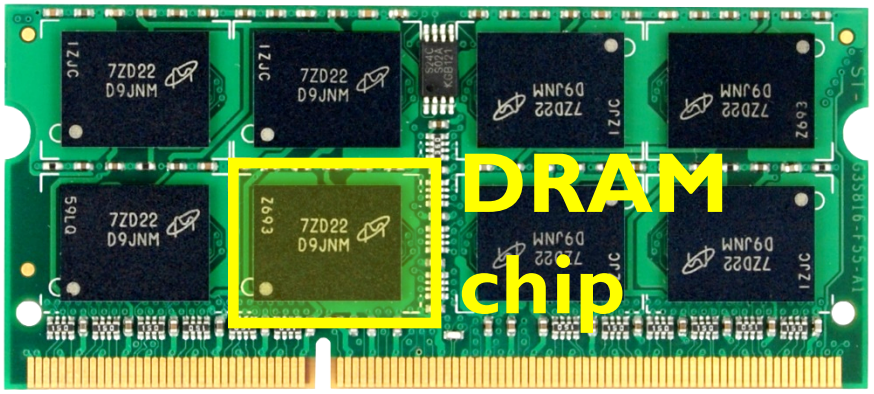
2 Develop a mechanism that exploits latency variation to reduce DRAM latency

Outline

- Motivation and Goals
- DRAM Background
- Experimental Methodology
- Characterization Results
- Mechanism: Flexible-Latency DRAM
- Conclusion

High-Level DRAM Organization

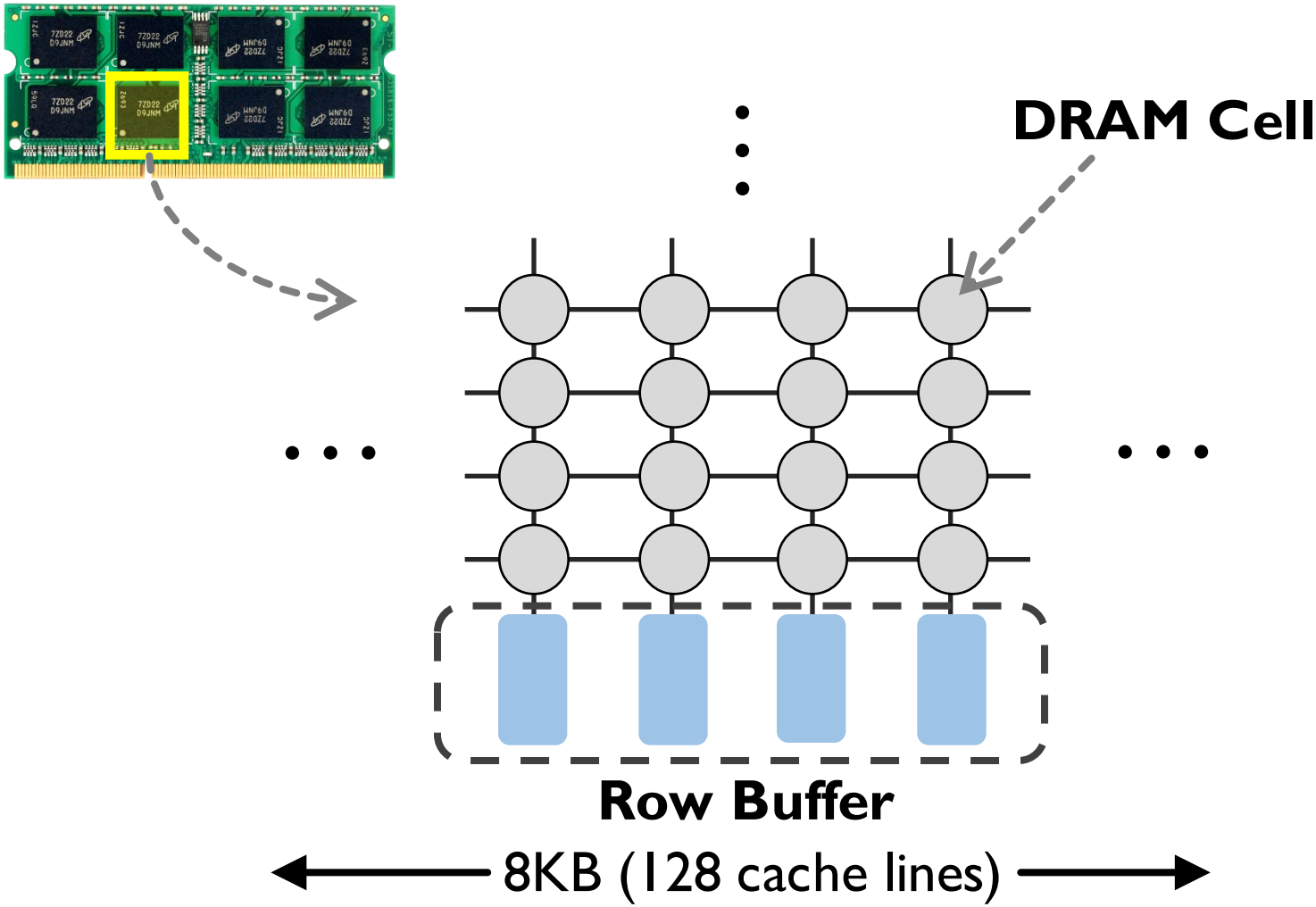
DRAM
Channel



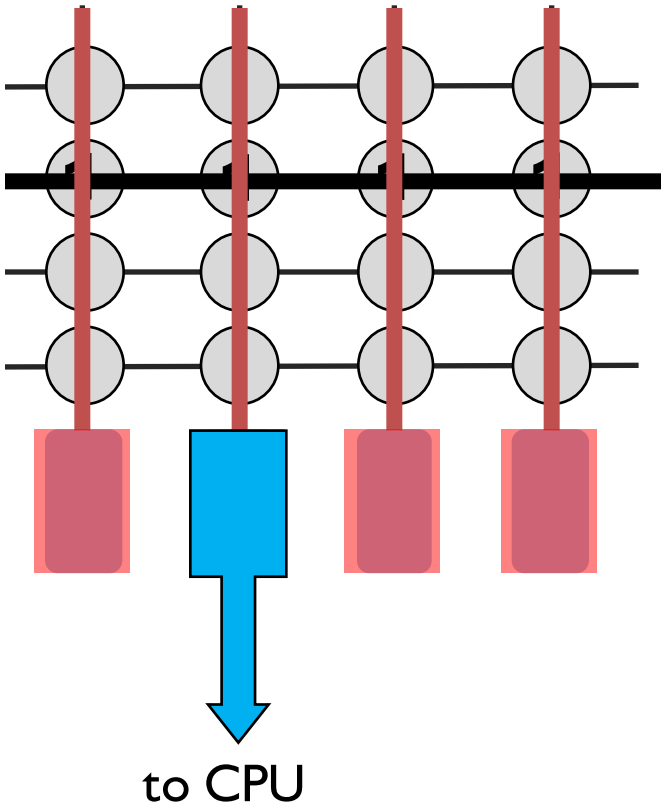
DIMM

(Dual in-line memory module)

DRAM Chip Internals



DRAM Operations

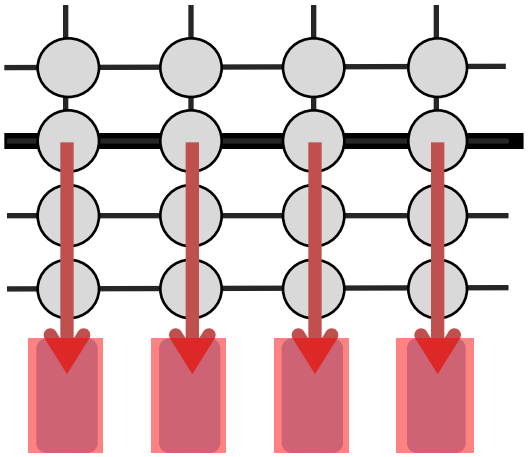


1 **ACTIVATE**: Store the row into the **row buffer**

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

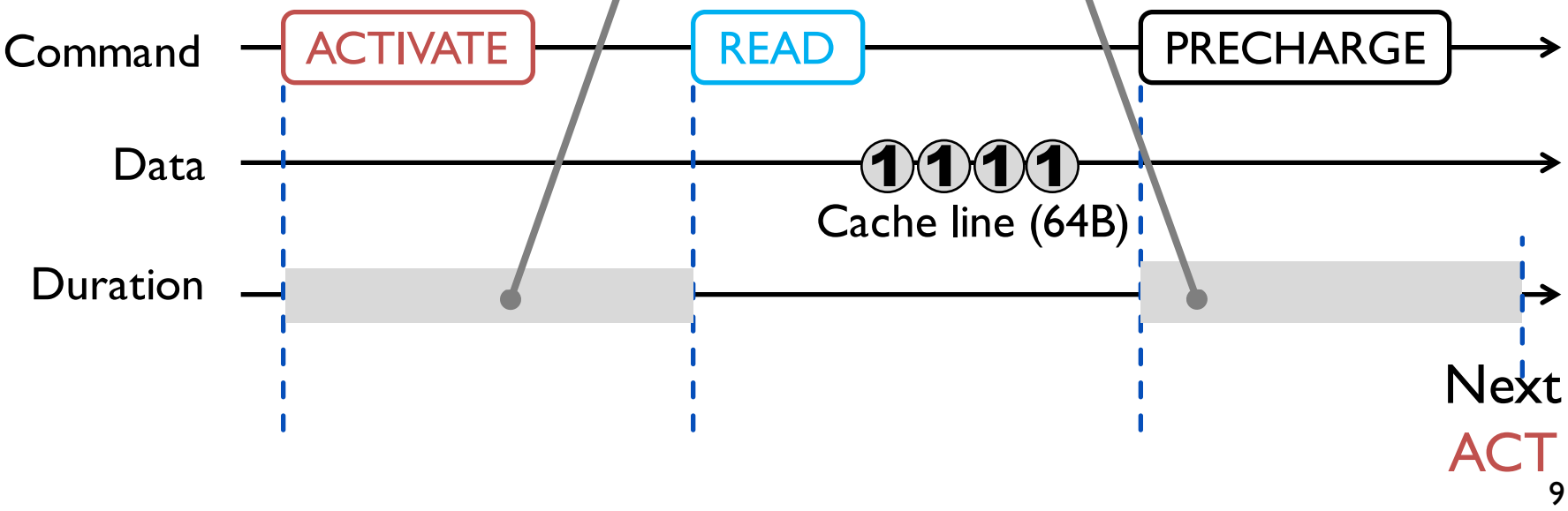
3 **PRECHARGE**: Prepare the array for a new ACTIVATE

DRAM Timing Parameters



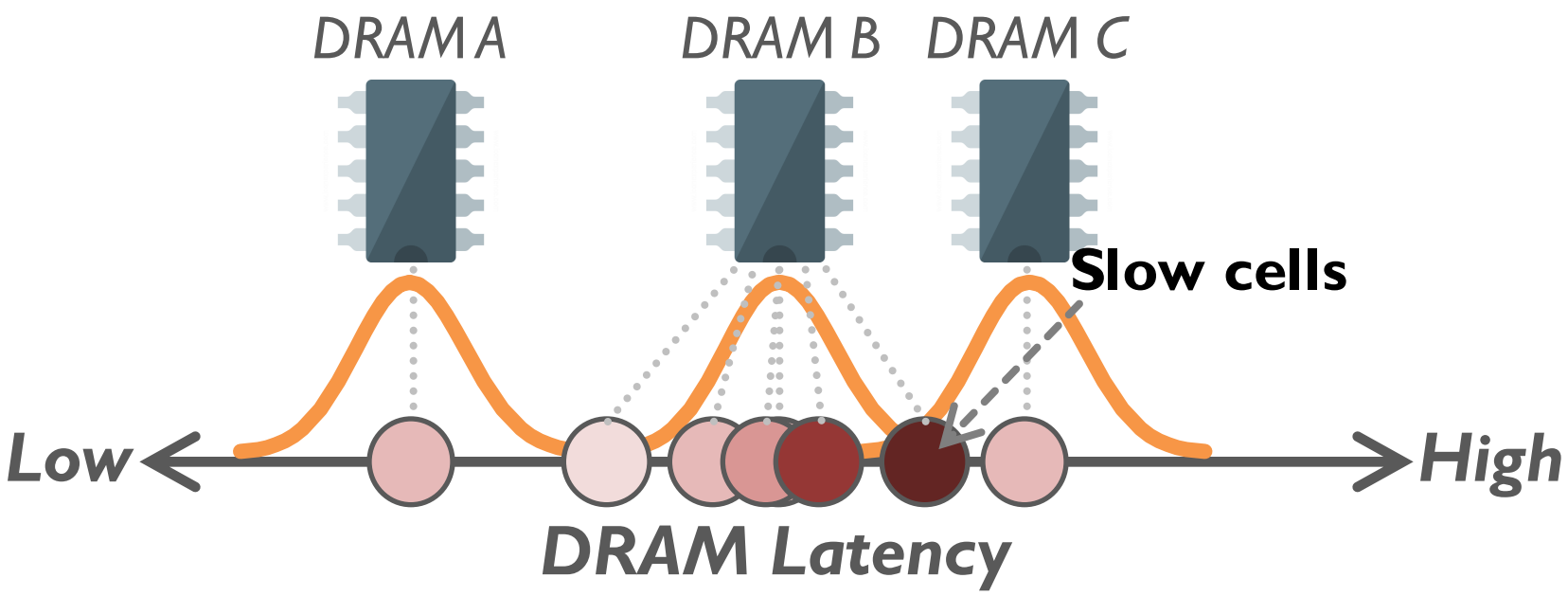
1 Activation latency: tRCD
(13ns / 50 cycles)

2 Precharge latency: tRP
(13ns / 50 cycles)



DRAM Latency Variation

Imperfect manufacturing process → latency variation



Experimental Questions

Imperfect manufacturing process → latency variation

Can we show latency variation in these parameters?

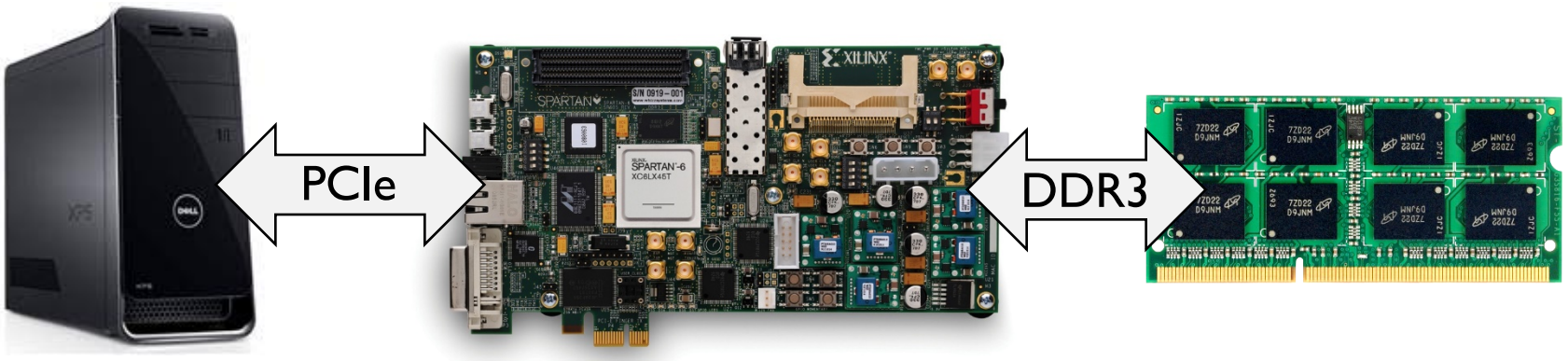
How large is latency variation in modern DRAM chips?

Can we identify the properties of slow cells with long latency?

Can we isolate slow cells to make DRAM faster?

Experimental Methodology

- Tool that enables us to freely issue DRAM commands
  - Existing systems: Commands are generated and controlled by HW
- Custom FPGA-based infrastructure



PC

C++ programs to specify commands

FPGA

Generate command sequence

DIMM

Experiments

- Swept each timing parameter to read data
  - Time step of 2.5ns (FPGA cycle time)

- Quantified *timing errors*: bit flips when using reduced latency

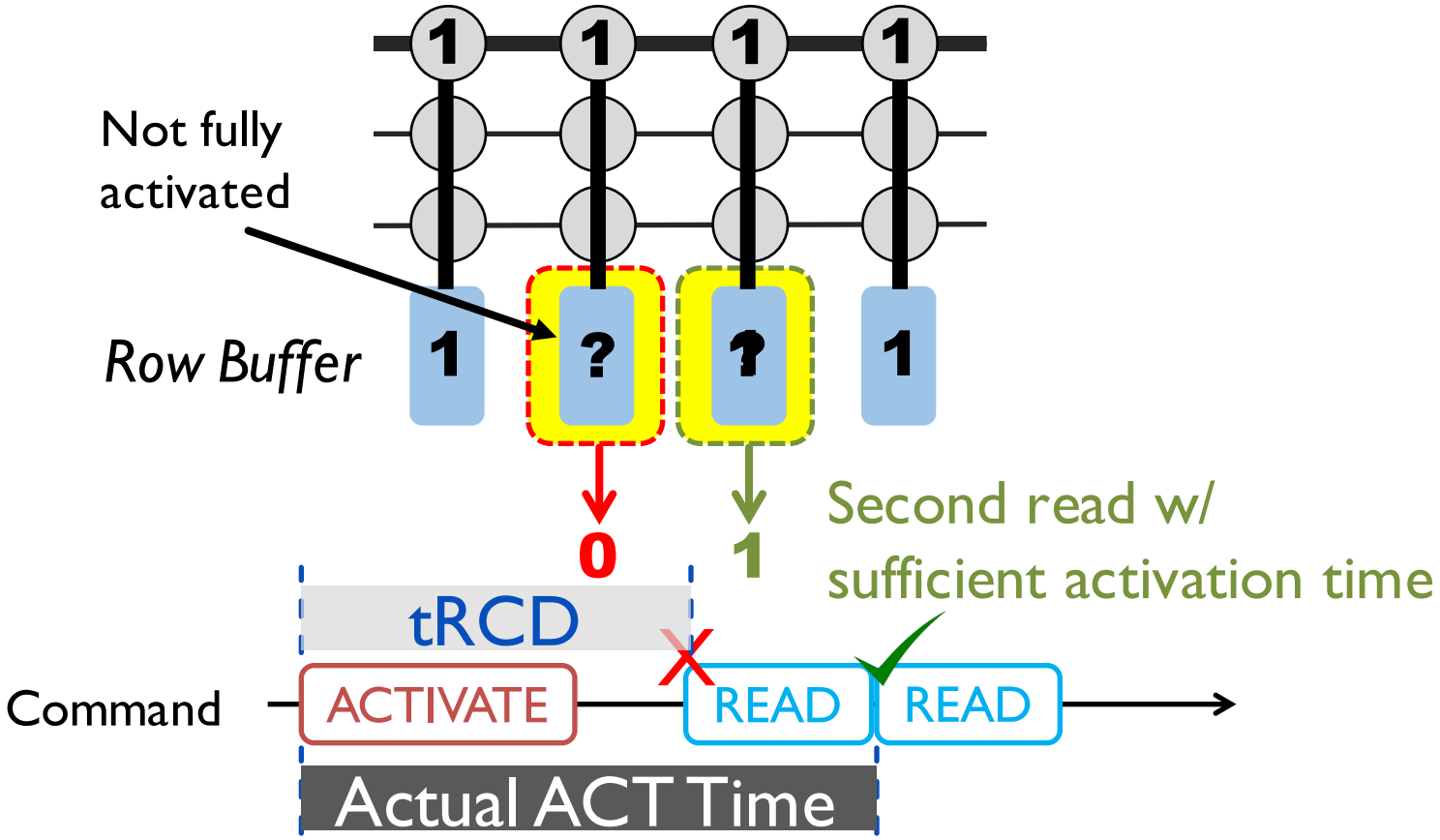
- Tested 240 DDR3 DRAM chips from three vendors
  - 30 DIMMs
  - Manufacturing dates: 2011 – 2013
  - Capacity: 1GB
  - Ambient temperature: 20°C

Outline

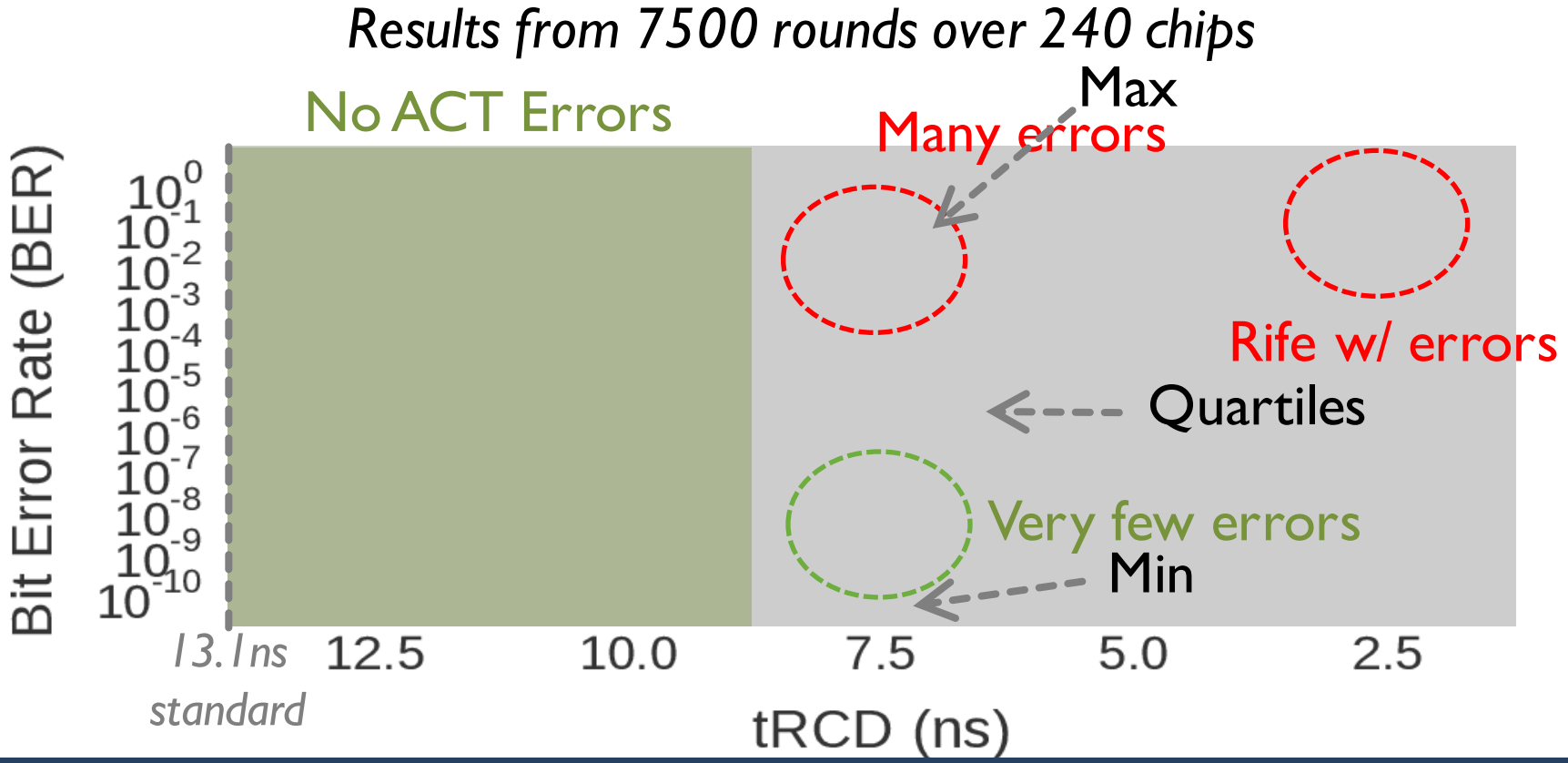
- Motivation and Goals
- DRAM Background
- Experimental Methodology
- Characterization Results
  - Activation latency
  - Precharge latency
- Mechanism: Flexible-Latency DRAM
- Conclusion

Activation Latency: Key Observation

Observation: ACT errors are isolated in the cells read in the first cache line

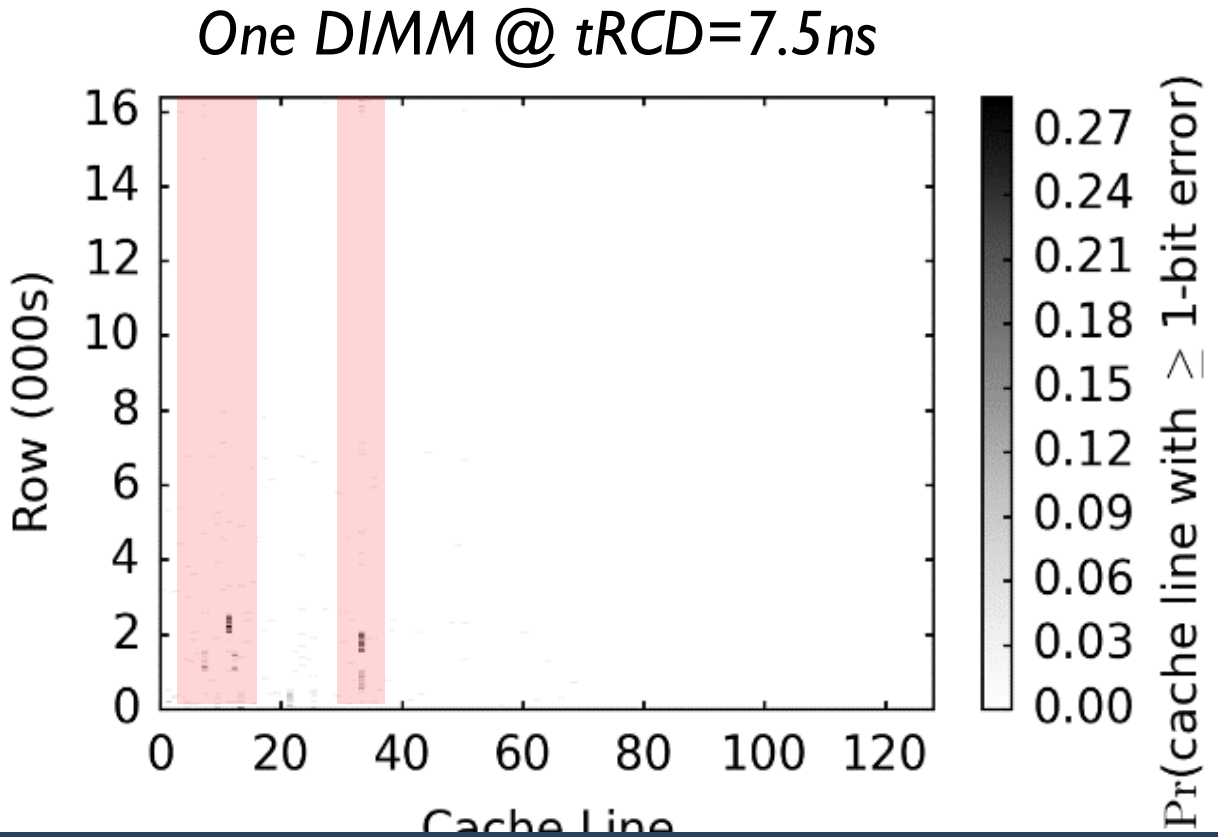


Variation in Activation Errors



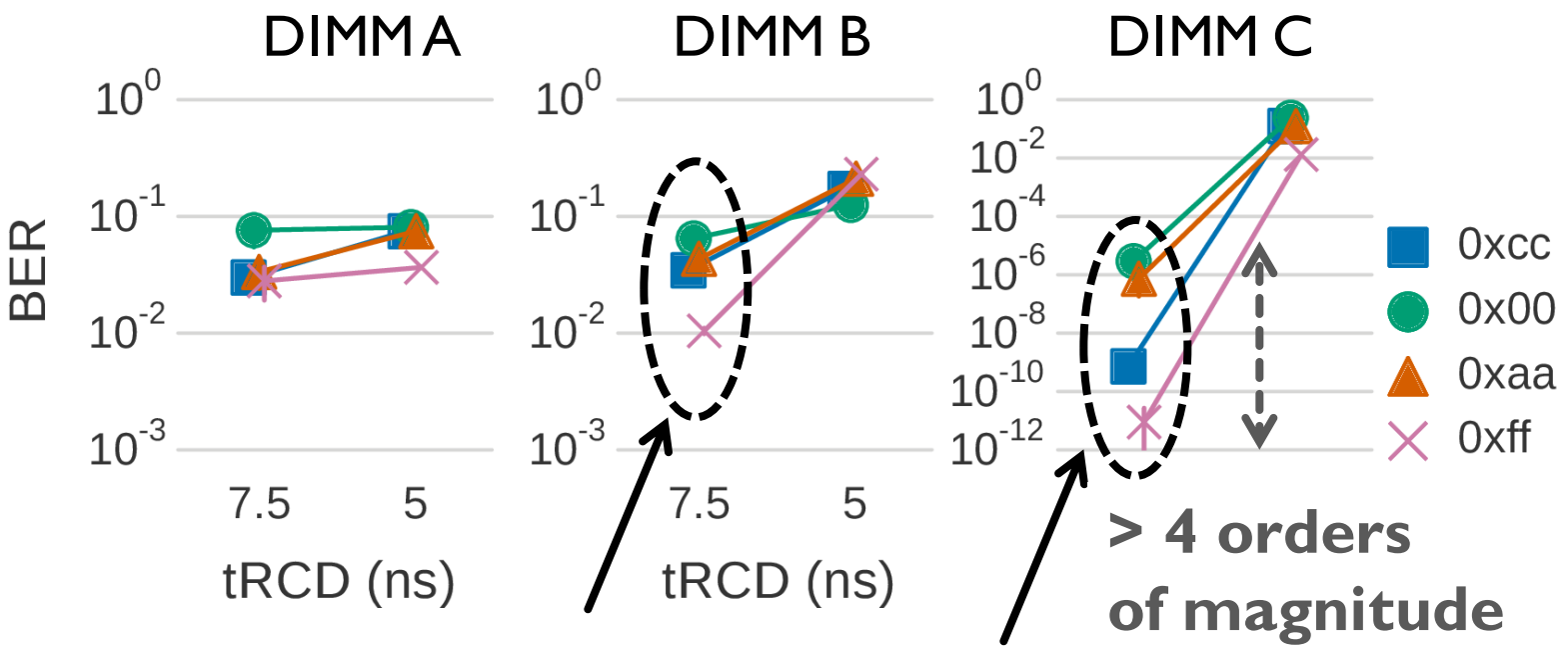
Modern DRAM chips exhibit significant variation in activation latency

Spatial Locality of Activation Errors



Activation errors are concentrated at certain columns of cells

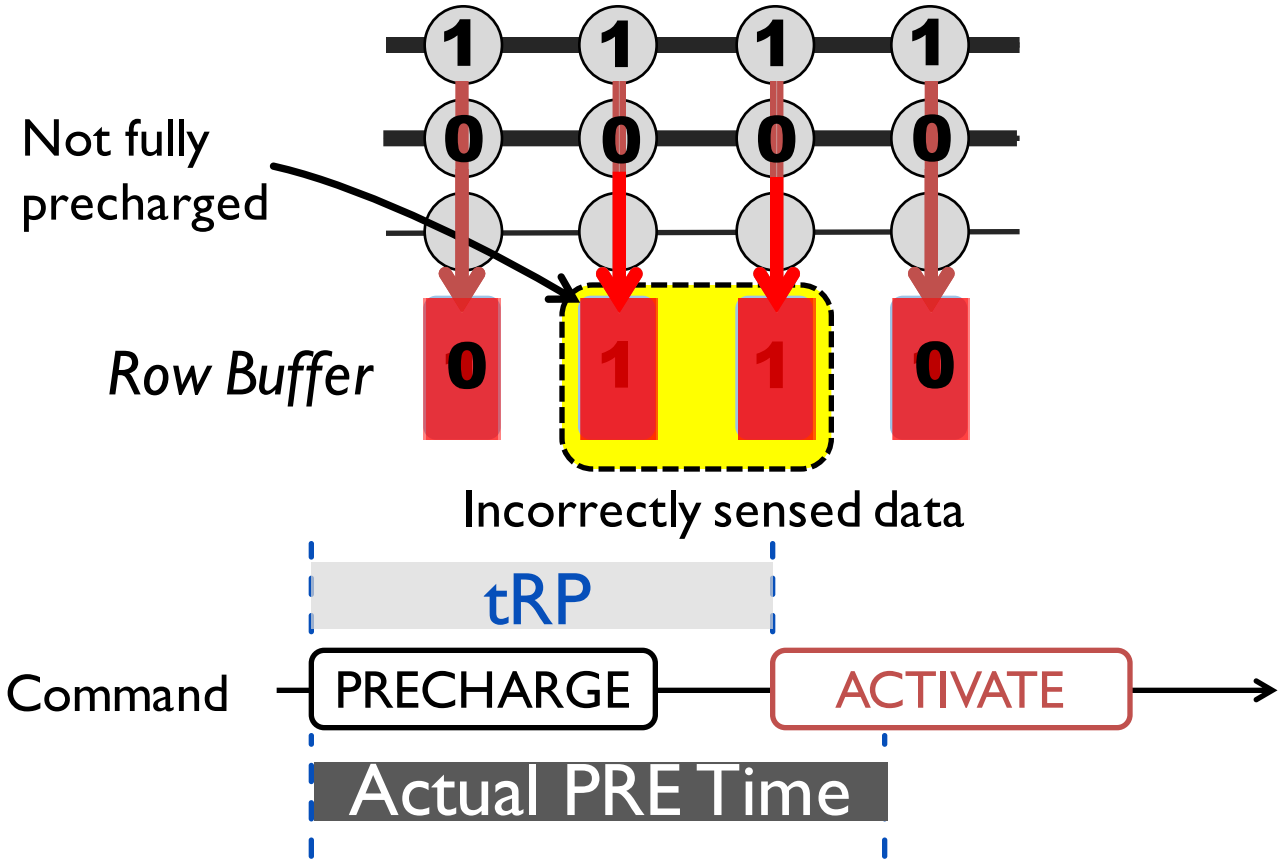
Strong Pattern Dependence



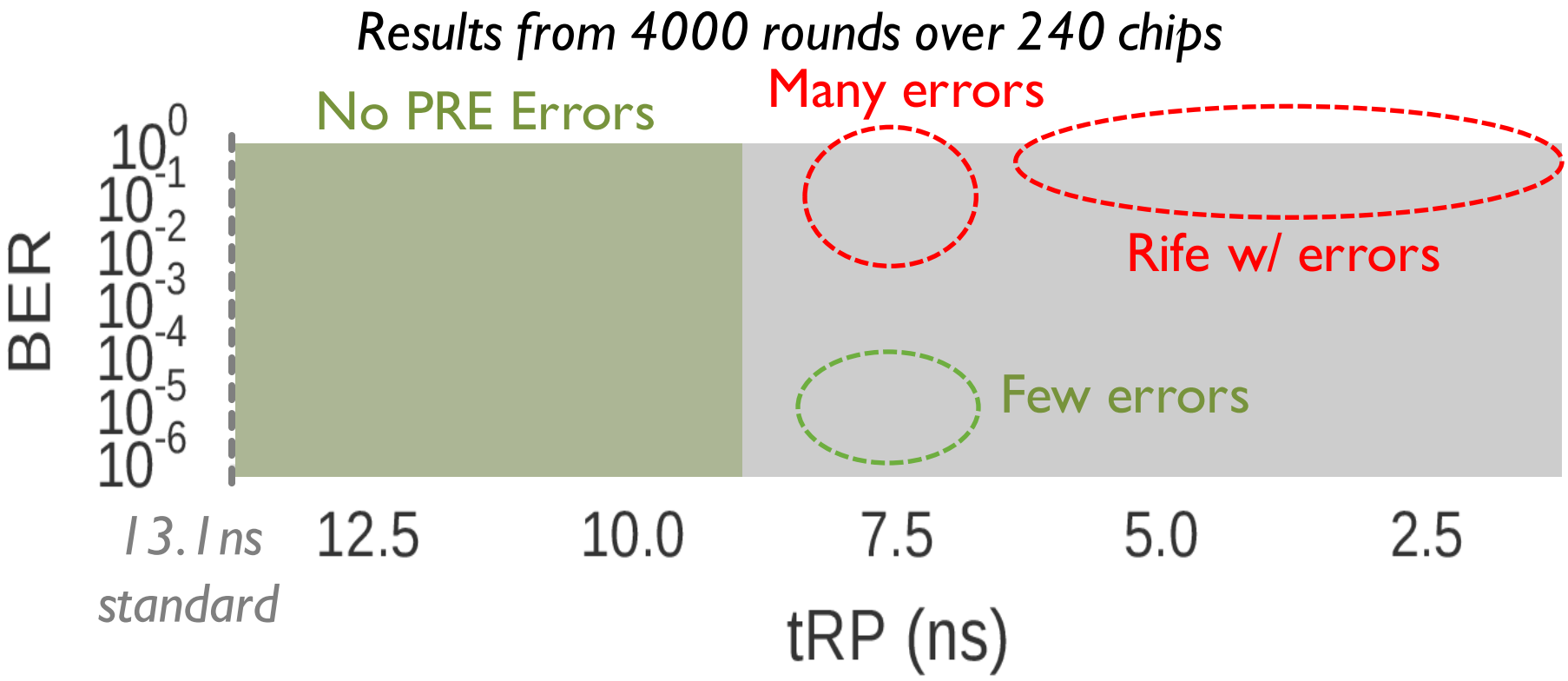
Activation errors have a strong dependence on the stored data patterns

Precharge Latency: Key Observation

Observation: PRE errors occur in multiple cache lines in the row activated after a precharge

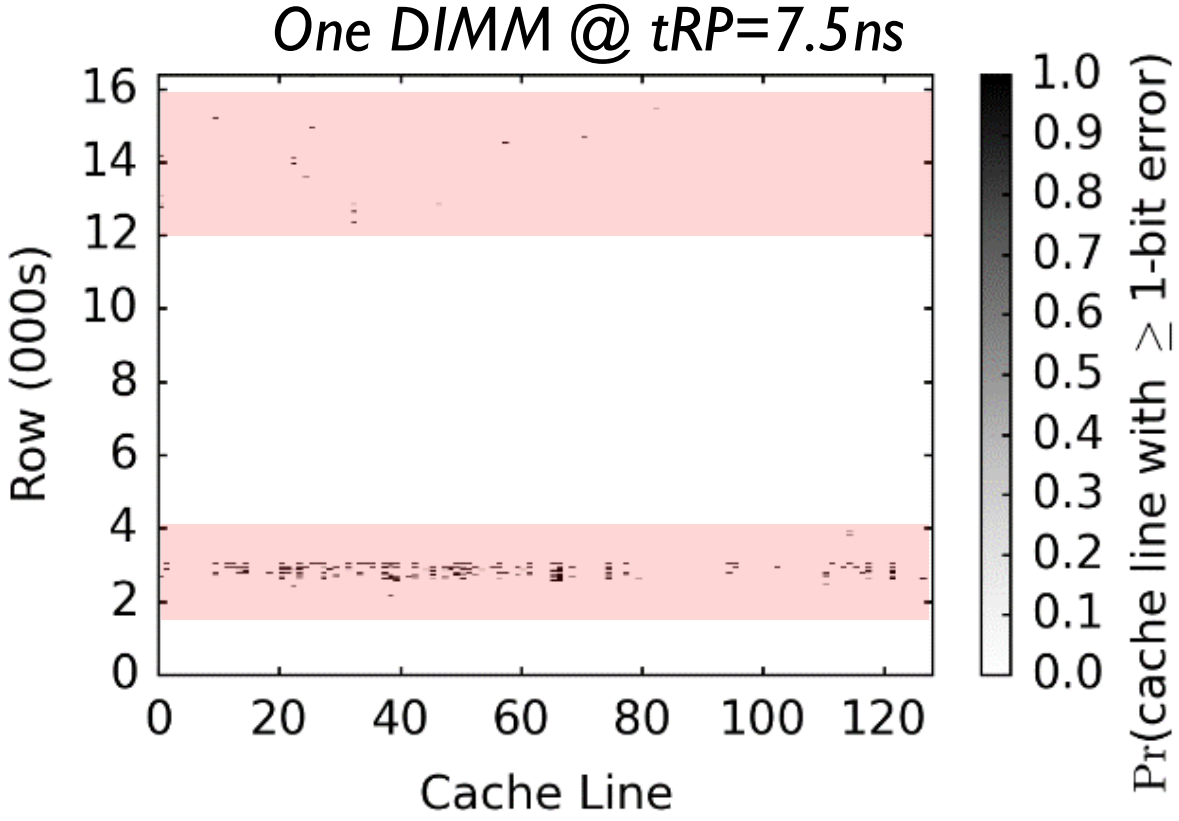


Variation in Precharge Errors



Modern DRAM chips exhibit significant variation in precharge latency

Spatial Locality of Precharge Errors



Precharge errors are concentrated at certain rows of cells

Outline

- Motivation and Goals
- DRAM Background
- Experimental Methodology
- Characterization Results
- Mechanism: Flexible-Latency DRAM
- Conclusion

Mechanism to Reduce DRAM Latency

- **Observations**
  - DRAM timing errors are concentrated on certain regions
  - All cells operate without errors at 10ns tRCD and tRP
- **Flexible-LatencY (FLY) DRAM**
  - A software-transparent design that reduces latency
- **Key idea**:
  1) Divide memory into regions of different latencies
  2) *Memory controller:* Use lower latency for regions without slow cells; higher latency for other regions

FLY-DRAM Evaluation Methodology

- **Cycle-level simulator**: Ramulator [CAL'15]

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

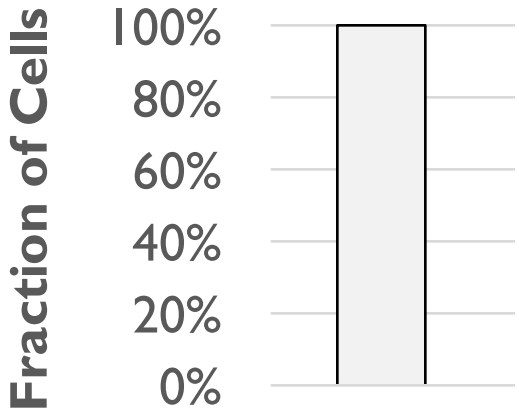
- **8-core** system with DDR3 memory

- **Benchmarks**: SPEC2006, TPC, STREAM, random

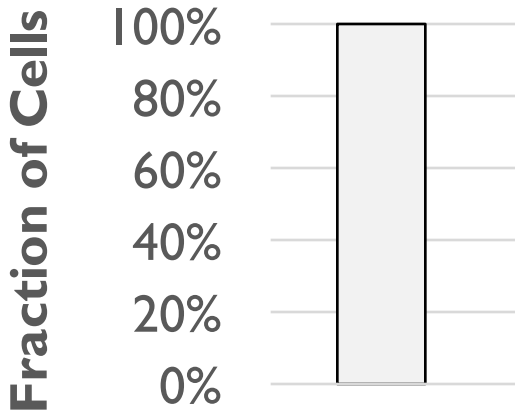
  – 40 8-core workloads

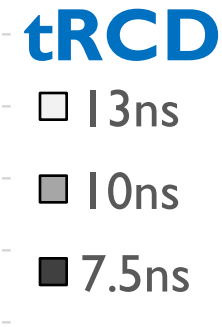
- **Performance metric**: Weighted Speedup (WS)

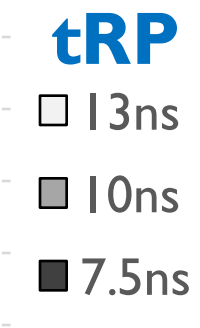
FLY-DRAM Configurations



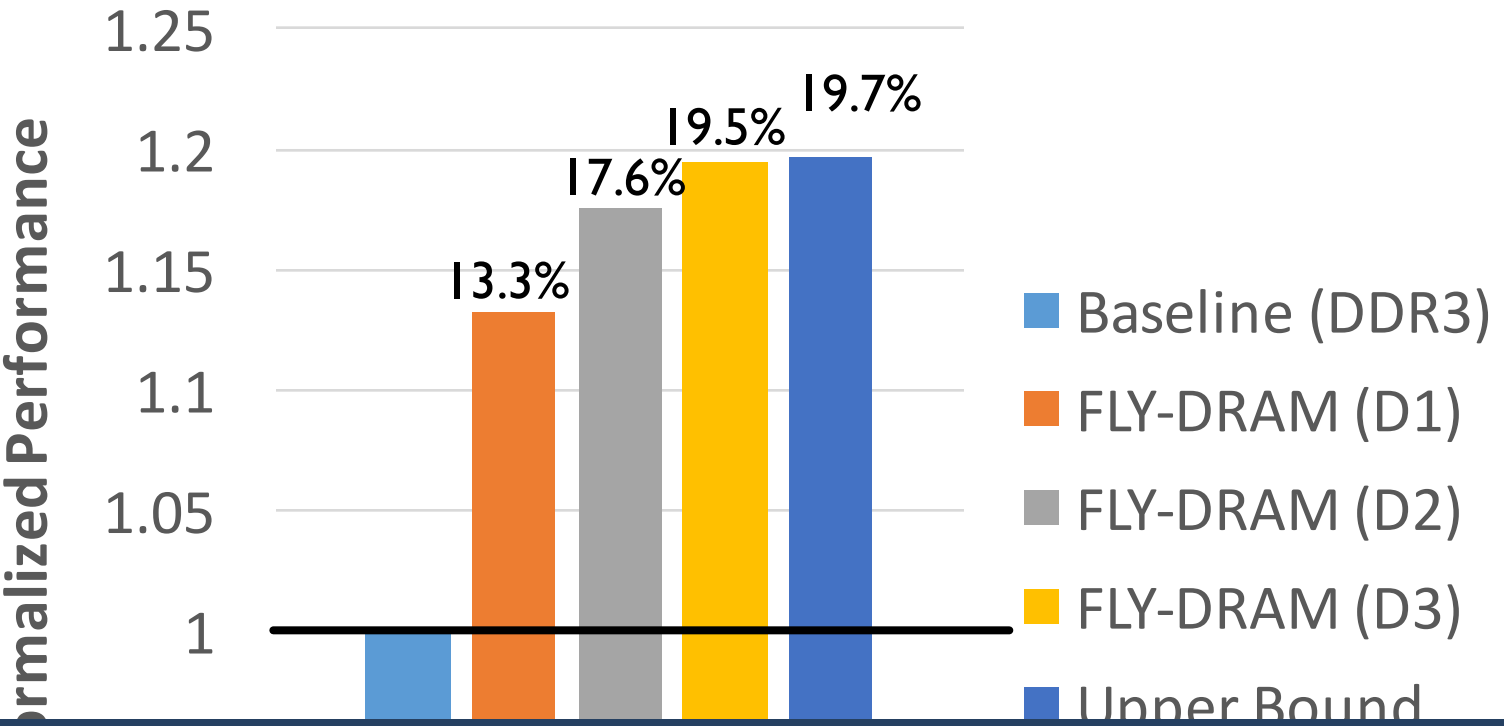
Baseline (DDR3)







Results



FLY-DRAM improves performance by exploiting latency variation in DRAM

Other Results in the Paper

- Error-correcting codes (ECC)
  - Effective at correcting activation errors
- Restoration latency
  - Significant margin to complete without errors
- Effect of temperature
  - Difference is not statistically significant to draw conclusion

Conclusion

- First to experimentally demonstrate and analyze latency variation behavior *within* real DRAM chips
- Show across 240 DRAM chips that:
  - All cells work below standard latency
  - Some regions of cells work even faster, but slow cells in other regions start to fail
  - Error rate is data-dependent
- FLY-DRAM reduces latency by using low latency for regions without slow cells and high latency for others
  - 13%/17%/19% speedup based on profiles of 3 real DIMMs

https://github.com/CMU-SAFARI/DRAM-Latency-Variation-Study

Understanding Latency Variation in Modern DRAM Chips

Experimental Characterization, Analysis, and Optimization

Kevin Chang

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

Carnegie Mellon University

TOBB ETÜ
TOBB Ekonomi ve Teknoloji Üniversitesi

University of Virginia

Peking University

ETH zürich

BACKUP SLIDES

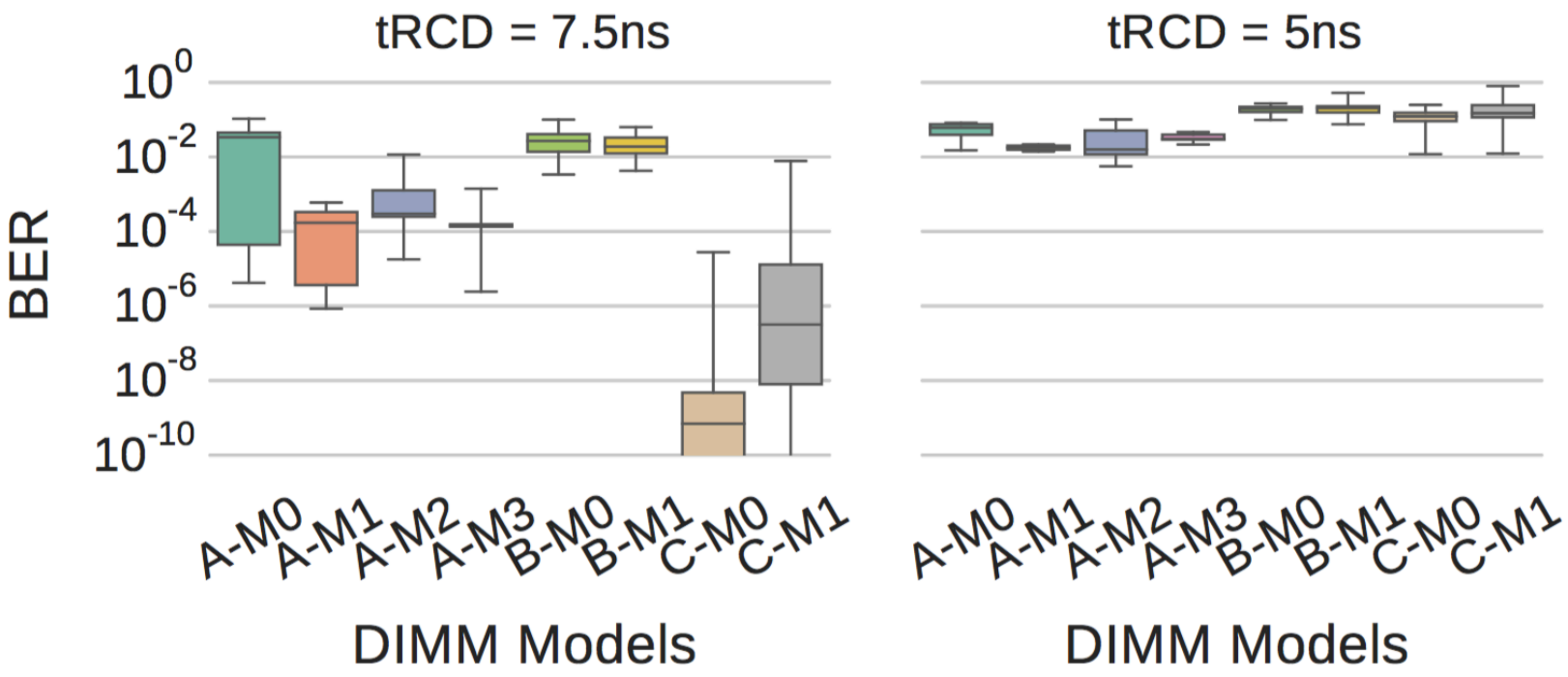
Infrastructure

DRAM DIMMs

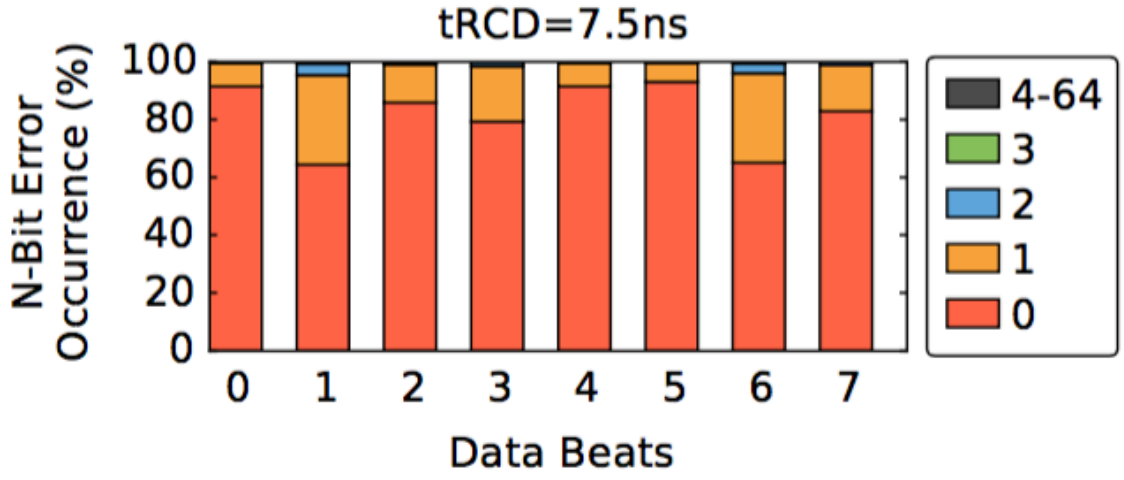
Vendor	DIMM Name	Model	Timing (ns) (tRCD/tRP/tRAS)	Assembly Year
A Total of 8 DIMMs	D_A^{0-1}	M0	13.125/13.125/35	2013
	D_A^{2-3}	M1	13.125/13.125/36	2012
	D_A^{4-5}	M2	13.125/13.125/35	2013
	D_A^{6-7}	M3	13.125/13.125/35	2013
B Total of 9 DIMMs	D_B^{0-5}	M0	13.125/13.125/35	2011-12
	D_B^{6-8}	M1	13.125/13.125/35	2012
C Total of 13 DIMMs	D_C^{0-5}	M0	13.125/13.125/34	2012
	D_C^{6-12}	M1	13.125/13.125/36	2011

Table 1: Properties of tested DIMMs.

Activation Latency Variation by DRAM Models

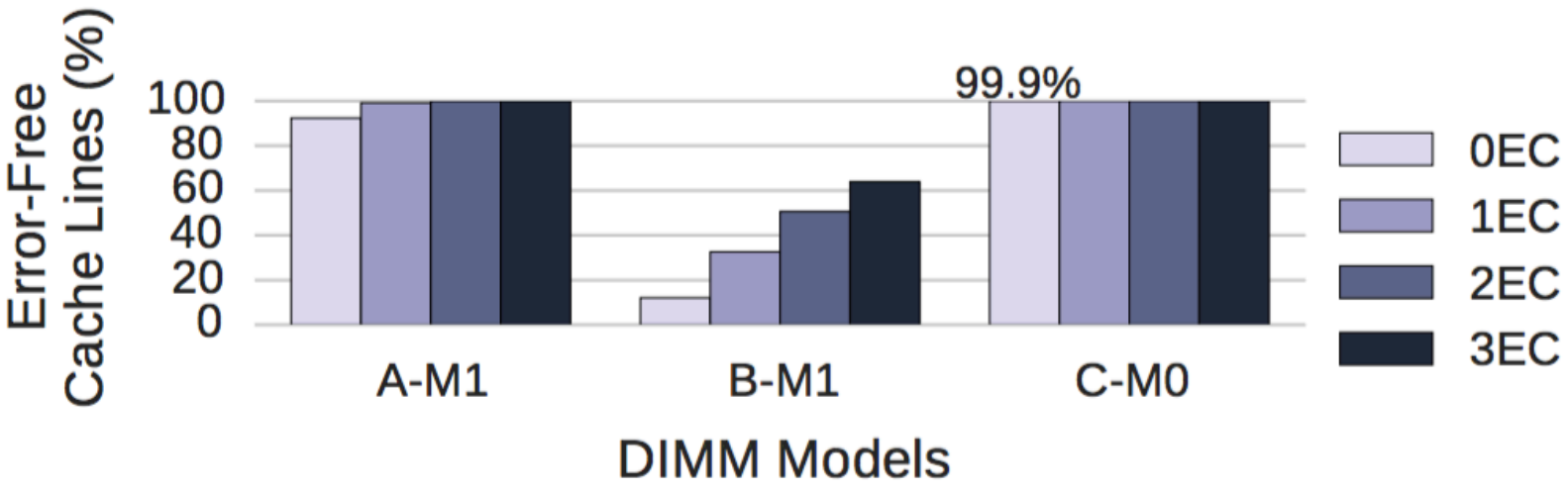


Activation Errors in Data Bursts

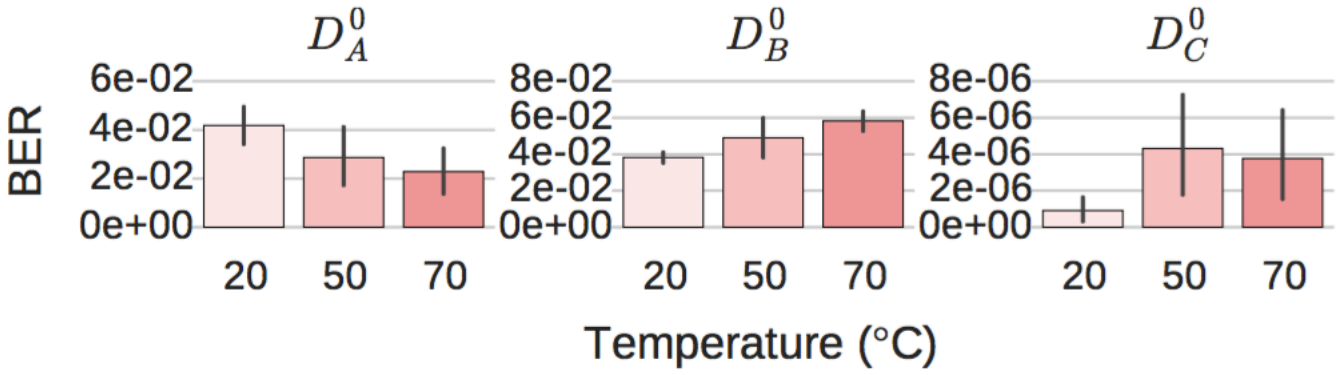


(a) A-M1

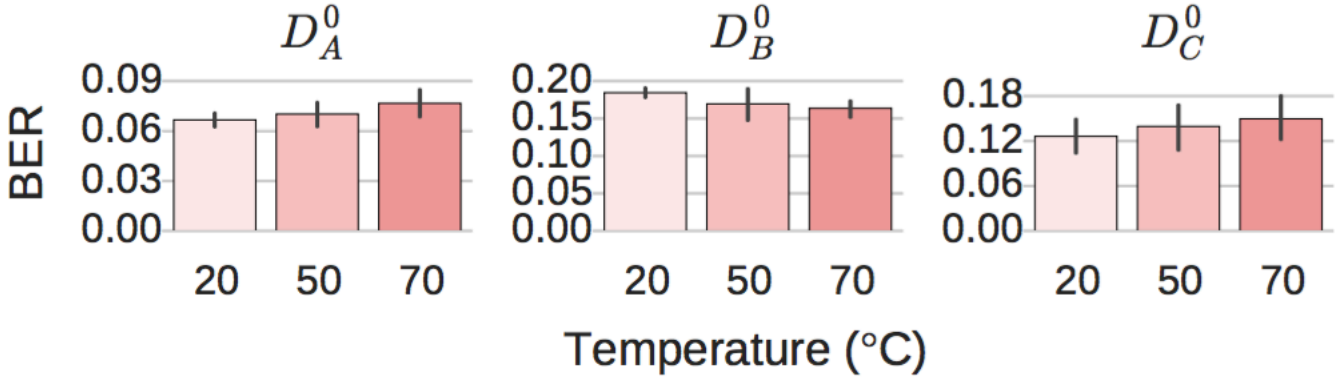
Effect of ECC on Activation Errors



Activation Errors by Temperature



(a) tRCD=7.5ns



(b) tRCD=5ns

Precharge Latency Variation by DRAM Models