Understanding and Improving Latency of DRAM-Based Memory Systems

Thesis Oral

Kevin Chang

Committee:

Prof. Onur Mutlu (Chair)

Prof. James Hoe

Prof. Kayvon Fatahalian

Prof. Stephen Keckler (NVIDIA, UT Austin)

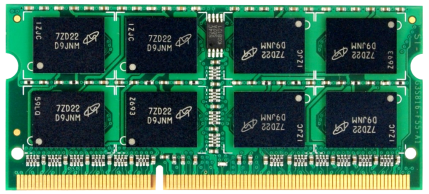
Prof. Moinuddin Qureshi (Georgia Tech.)

Carnegie Mellon University

The March For "Moore"

4B transistors

8Gb



4K transistors

1Kb

Intel 8080, 1974

Intel 1103, 1970

Processor

Main Memory

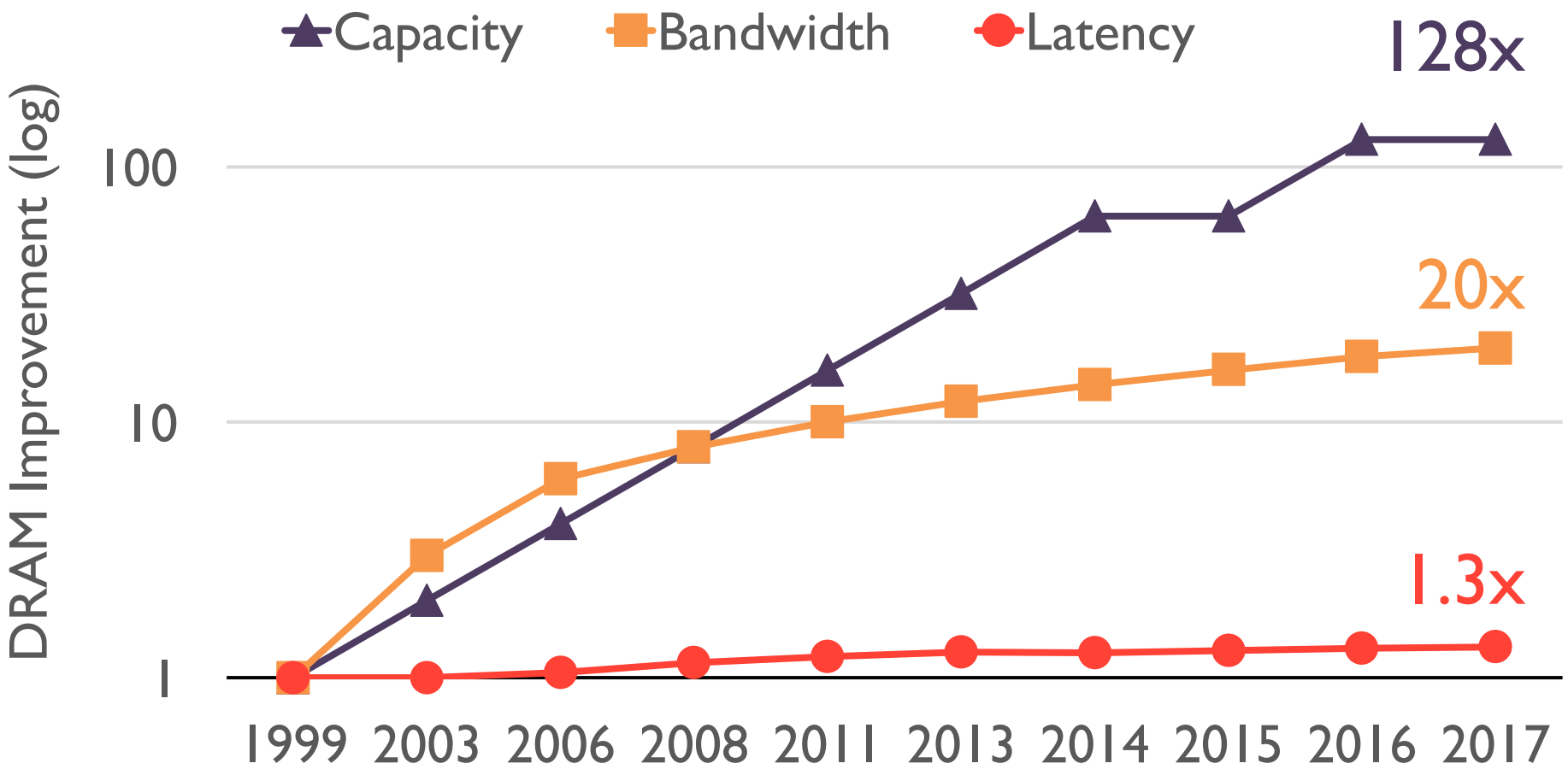
or

DRAM (Dynamic Random Access Memory)

PROBLEM

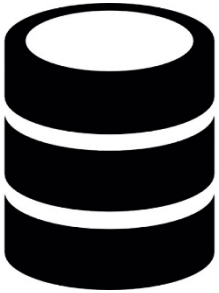
DRAM latency has been relatively stagnant

Main Memory Latency Lags Behind

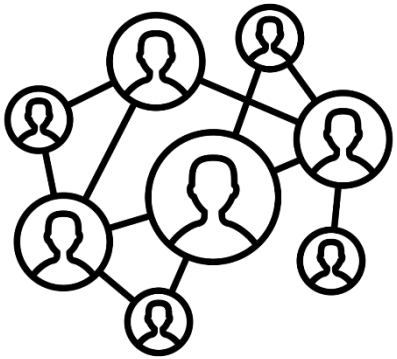


Memory latency remains almost constant

DRAM Latency Is Critical for Performance



In-memory Databases

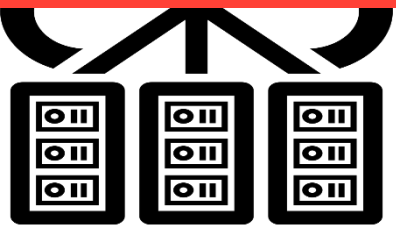


Graph/Tree Processing

Long memory latency → performance bottleneck

In-Memory Data Analytics

[Clapp+ (**Intel**), IISWC'15;
Awan+, BDCloud'15]



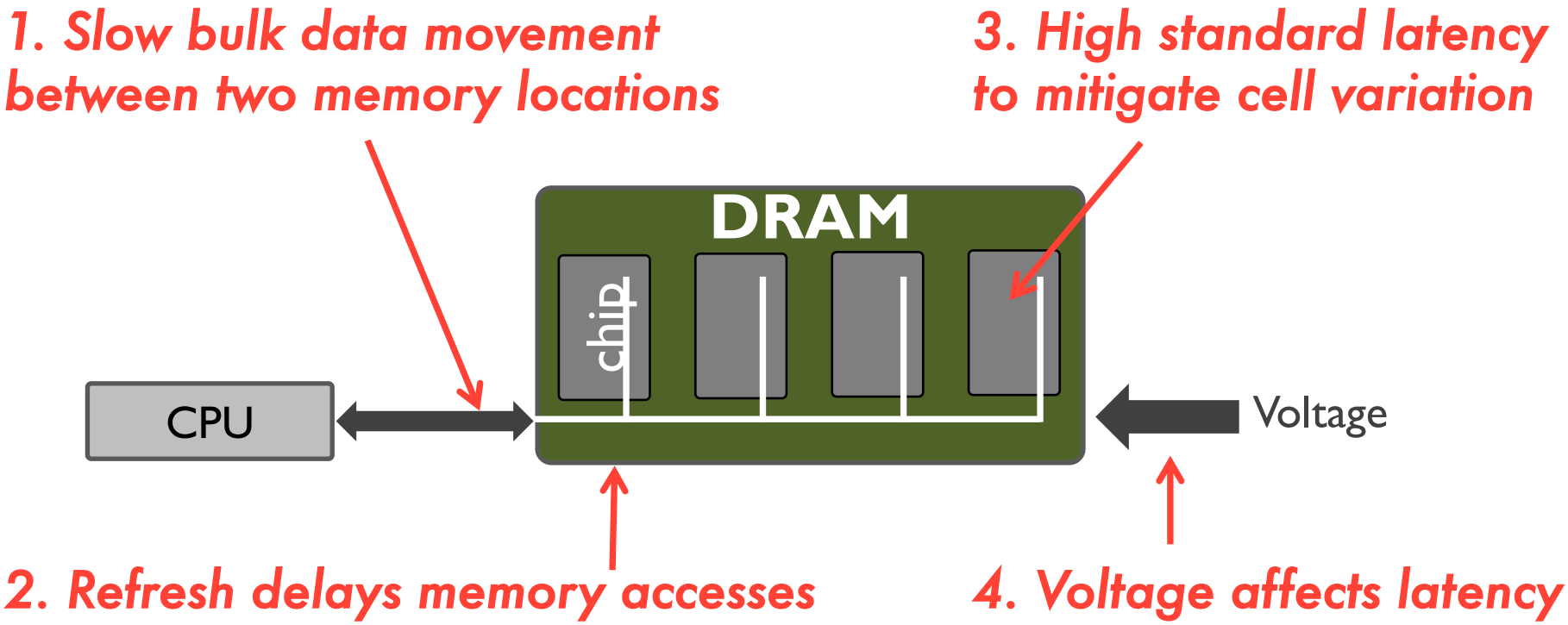
Datacenter workloads

[Kanev+ (**Google**), ISCA'15]

Goal

Improve latency of DRAM (main memory)

Different DRAM Latency Problems



Thesis Statement

Memory latency can be significantly reduced with a multitude of **low-cost architectural techniques** that aim to **reduce different causes of long latency**

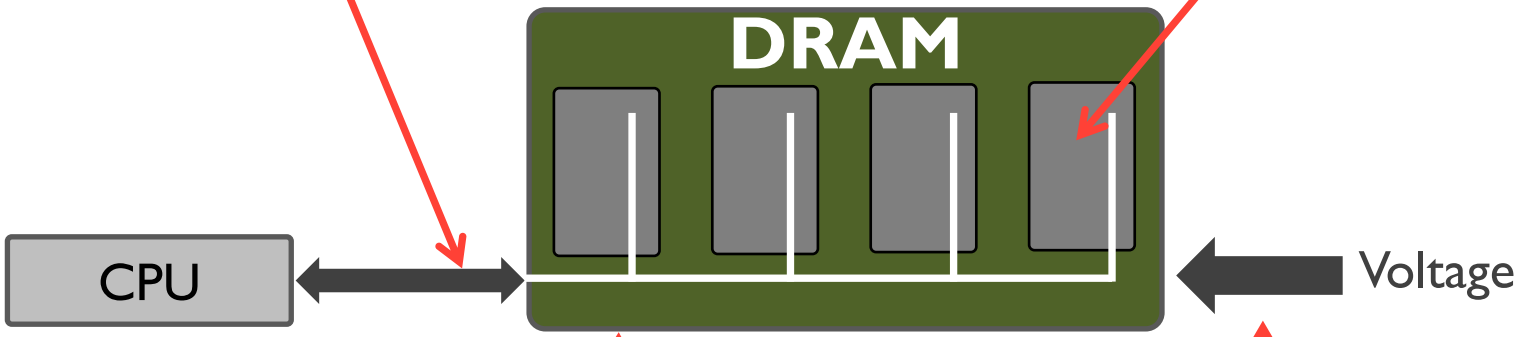
Contributions

Low-Cost Architectural Features in DRAM

Understanding and overcoming the latency limitation in DRAM

Low-Cost Inter-Linked Subarrays (LISA) [HPCA'16]

Understanding and Exploiting Latency Variation in DRAM (FLY-DRAM) [SIGMETRICS'16]



Mitigating Refresh Latency by Parallelizing Accesses with Refreshes (DSARP) [HPCA'14]

Understanding and Exploiting Latency-Voltage Trade-Off (Voltron) [SIGMETRICS'17]

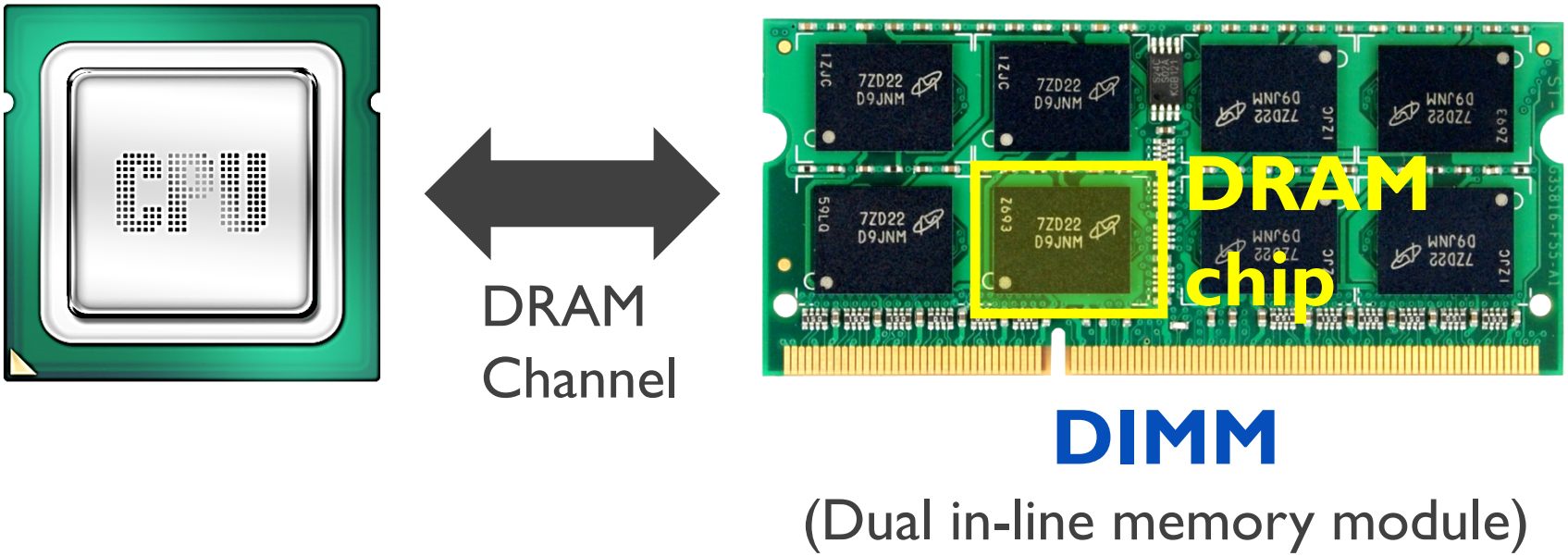
DRAM Background

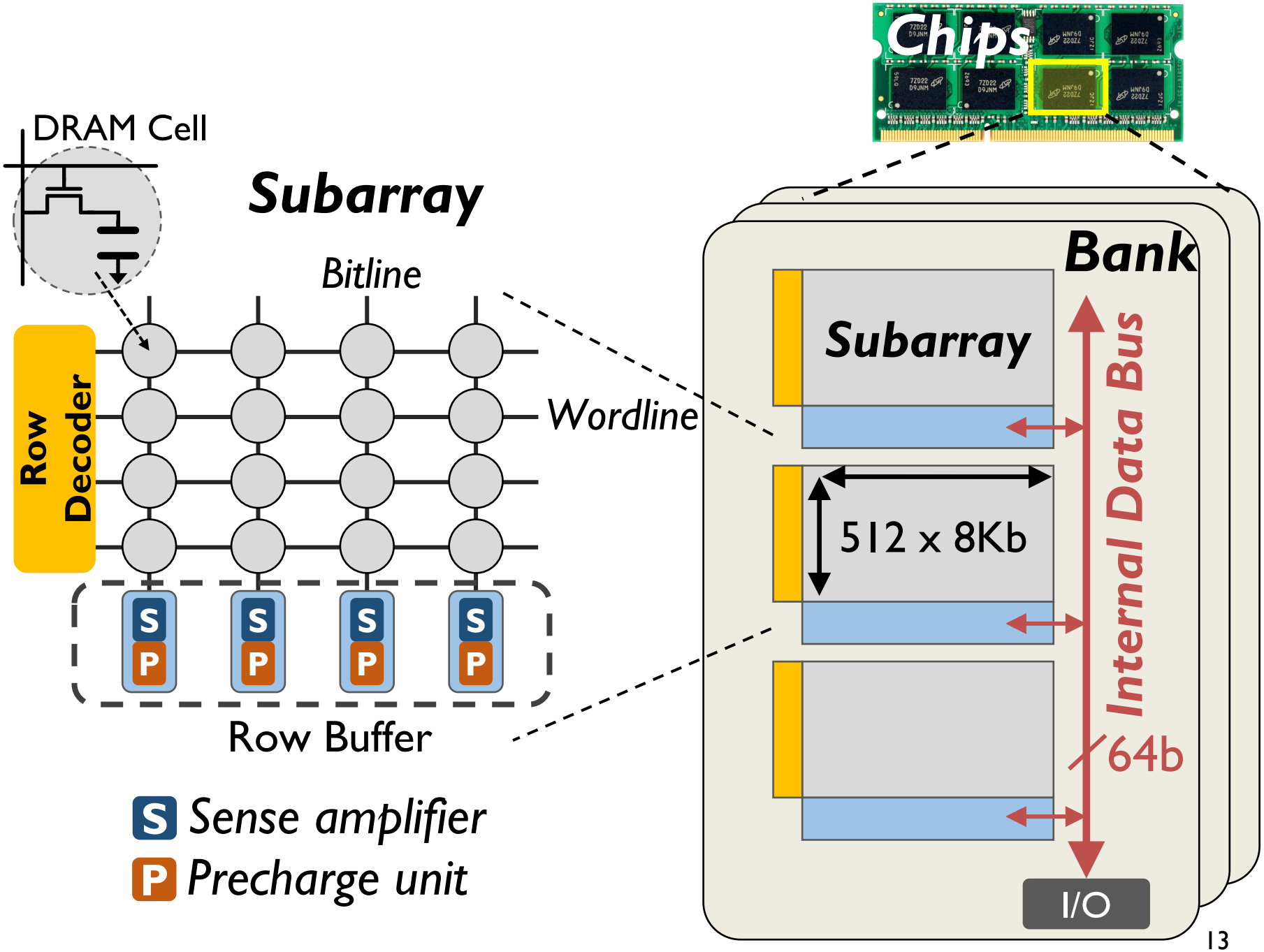
What's inside a DRAM chip?

How to access DRAM?

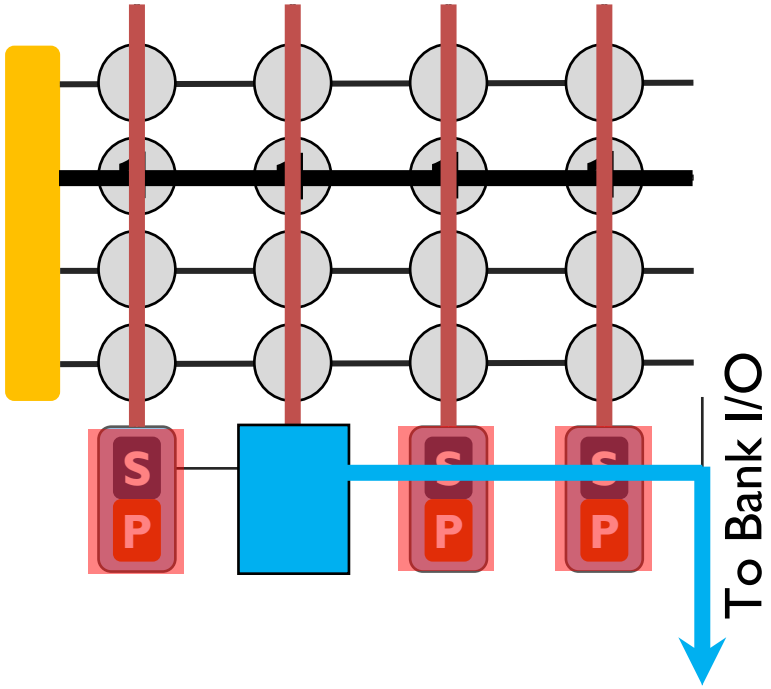
How long does accessing data take?

High-Level DRAM Organization




Chips
DRAM Cell
Subarray
Bitline
Row Decoder
Wordline
S
P
Row Buffer
S Sense amplifier
P Precharge unit
Bank
Subarray
512 x 8Kb
Internal Data Bus
64b
I/O

Reading Data From DRAM

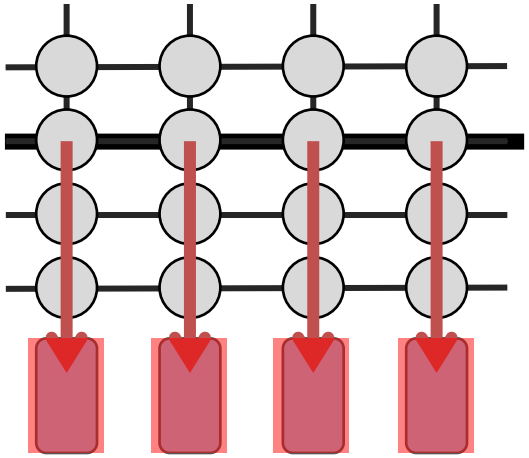


1 **ACTIVATE**: Store the row into the **row buffer**

2 **READ**: Select the target column and drive to CPU

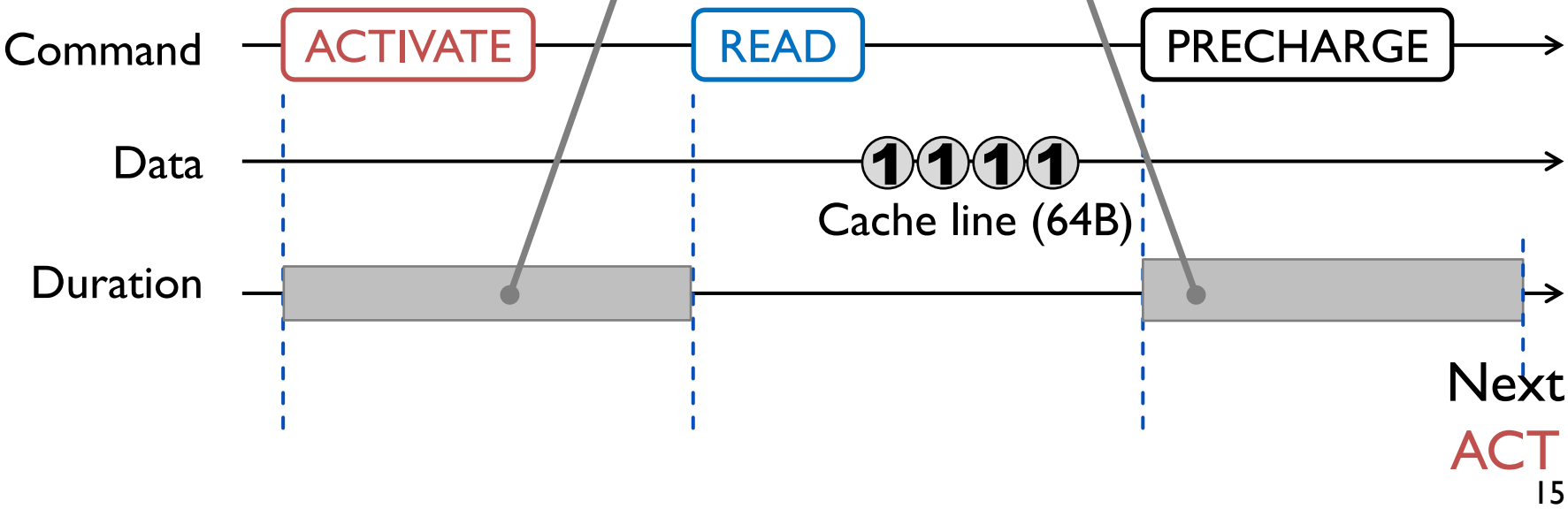
3 **PRECHARGE**: Reset the bitlines for a new ACTIVATE

DRAM Access Latency



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

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

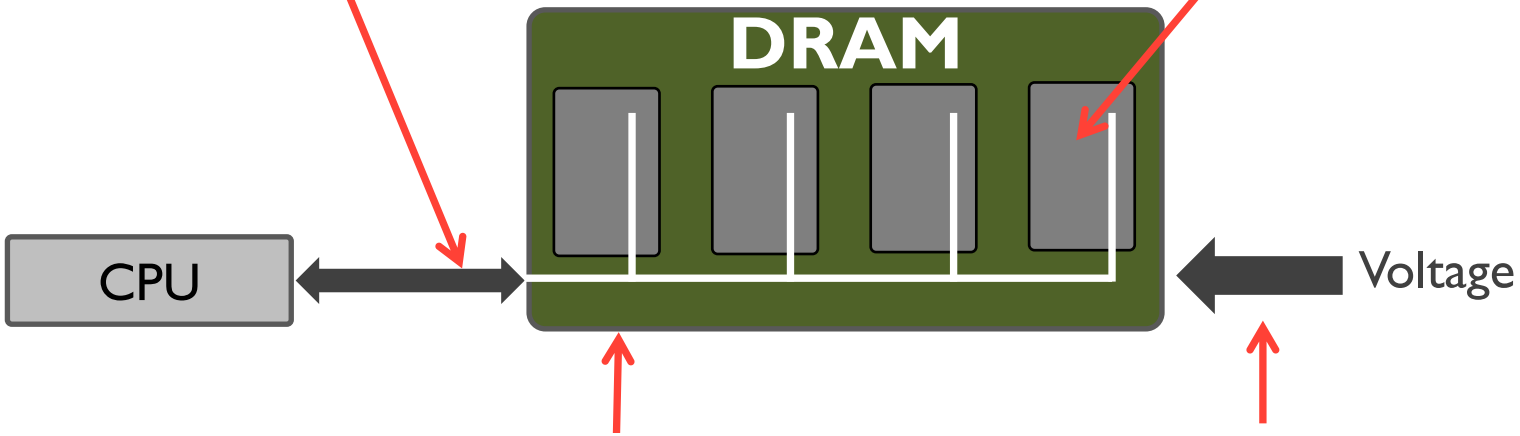


Low-Cost Architectural Features in DRAM

Understanding and overcoming the latency limitation in DRAM

Low-Cost Inter-Linked Subarrays (LISA) [HPCA'16]

Understanding and Exploiting Latency Variation in DRAM (FLY-DRAM) [SIGMETRICS'16]



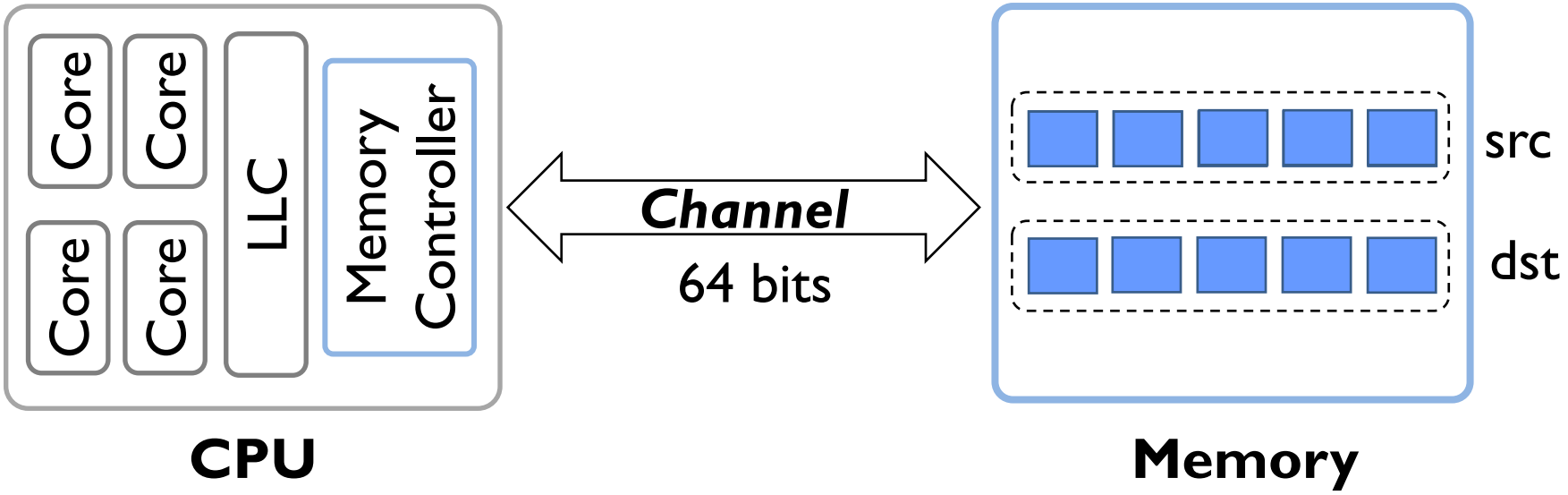
Mitigating Refresh Latency by Parallelizing Accesses with Refreshes (DSARP) [HPCA'14]

Understanding and Exploiting Latency-Voltage Trade-Off (Voltron) [SIGMETRICS'17]

Problem: Inefficient Bulk Data Movement

Bulk data movement is a key operation in many applications

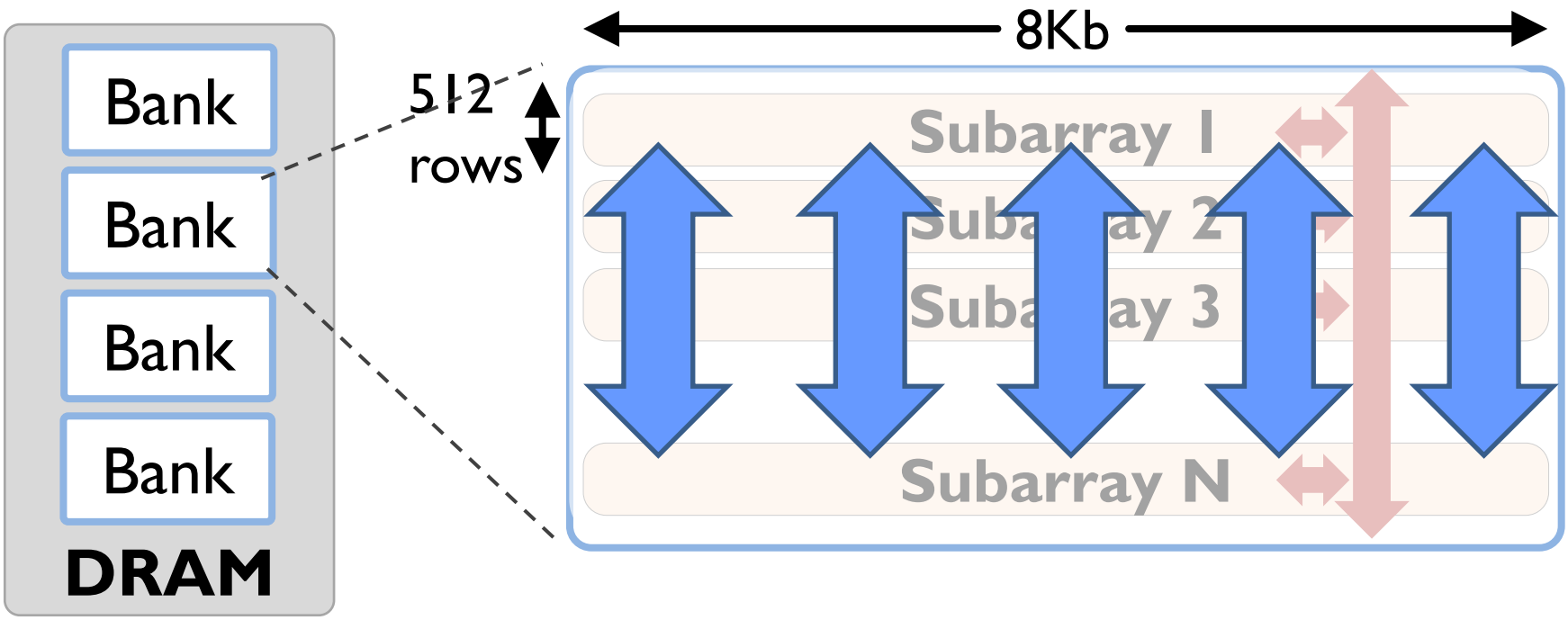
– *memmove* & *memcpy*: 5% cycles in Google's datacenter [Kanev+, ISCA'15]



Long latency and high energy

Move Data inside DRAM?

Moving Data Inside DRAM?

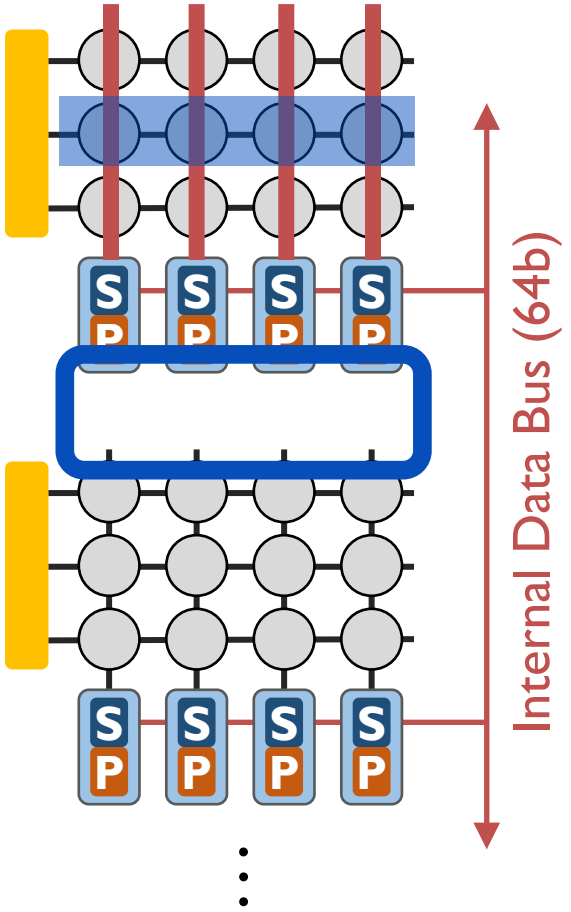


Goal: Provide a new substrate to enable wide connectivity between subarrays

Our proposal:

Low-Cost Inter-Linked SubArrays (LISA)

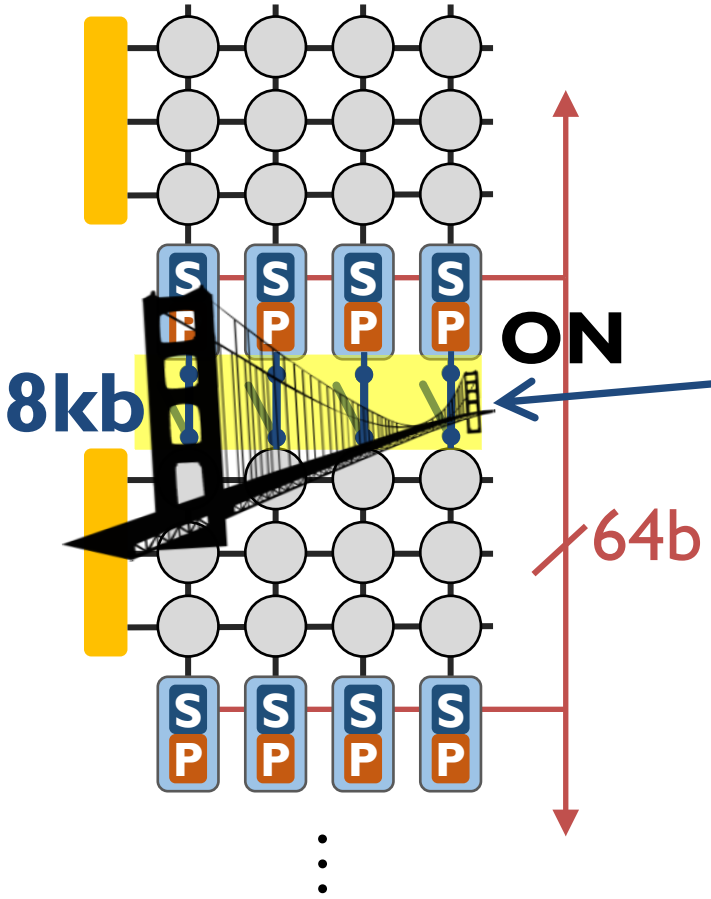
Observations



1 Bitlines serve as a **bus** that is as wide as a row

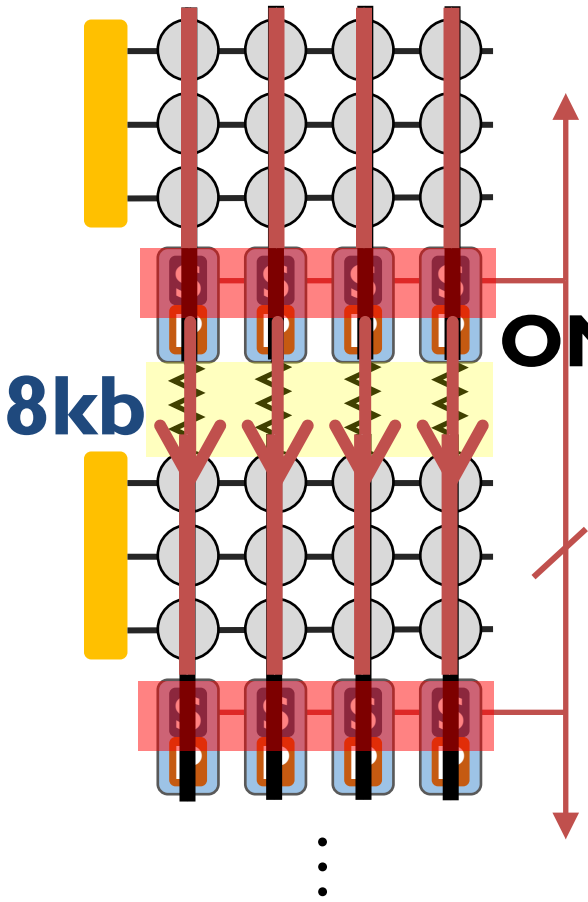
2 Bitlines between subarrays are close but disconnected

Low-Cost Interlinked Subarrays (LISA)



Interconnect bitlines of adjacent subarrays in a bank using **isolation transistors (links)**

Low-Cost Interlinked Subarrays (LISA)



Row Buffer Movement (RBM): Move a row of data in an activated row buffer to a precharged one

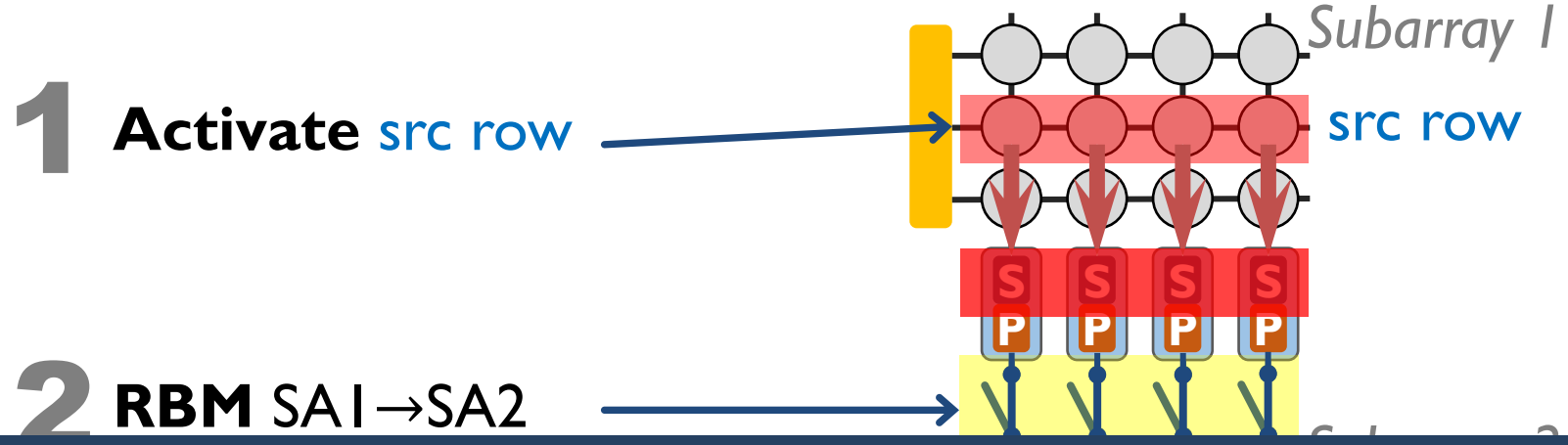
- **4KB data in 8ns**
  → **500 GB/s, 26x** bandwidth of a DDR4-2400 channel
- **0.8%** DRAM chip area overhead

Three New Applications of LISA to Reduce Latency

1 Fast bulk data copy

1. Rapid Inter-Subarray Copying (RISC)

- **Goal:** Efficiently copy a row across subarrays
- **Key idea:** Use *RBM* to form a new command sequence

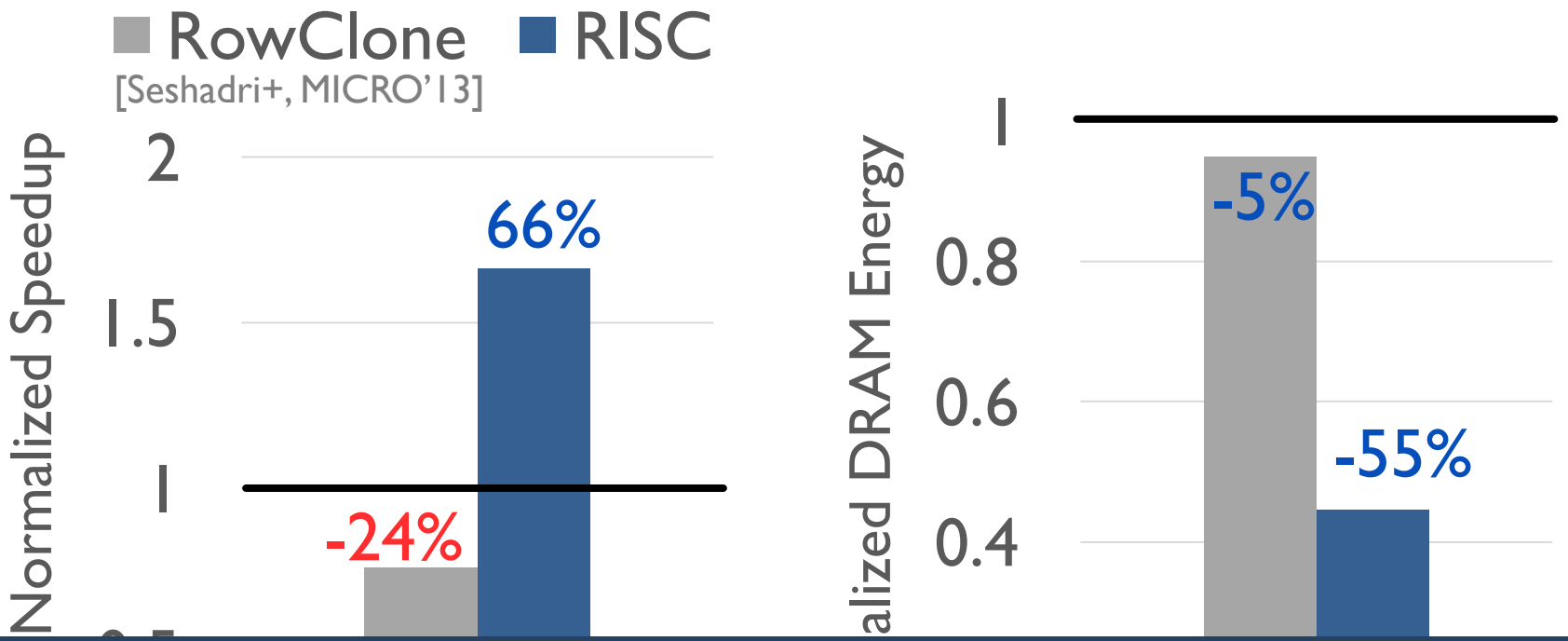


Reduces row-copy latency by 9x,
DRAM energy by 48x

Methodology

- Cycle-level simulator: Ramulator [Kim+, CAL'15]
- **Four out-of-order cores**
- **Two DDR3-1600 channels**
- Benchmarks: TPC, STREAM, SPEC2006, DynoGraph, random, bootup, forkbench, shell script

RISC Outperforms Prior Work



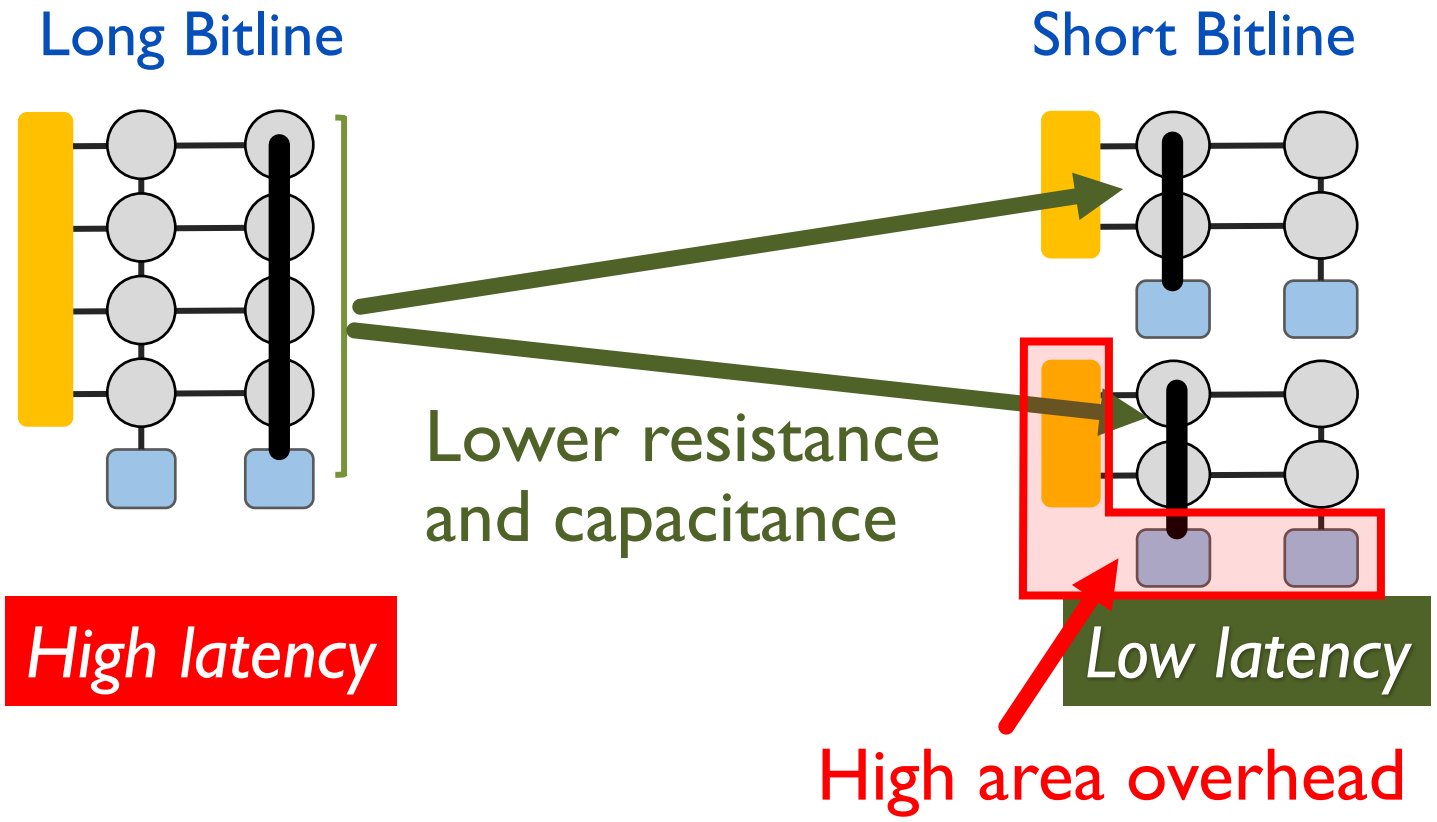
Rapid Inter-Subarray Copying (RISC) using LISA improves system performance

Three New Applications of LISA to Reduce Latency

1. Fast bulk data copy
2. In-DRAM caching

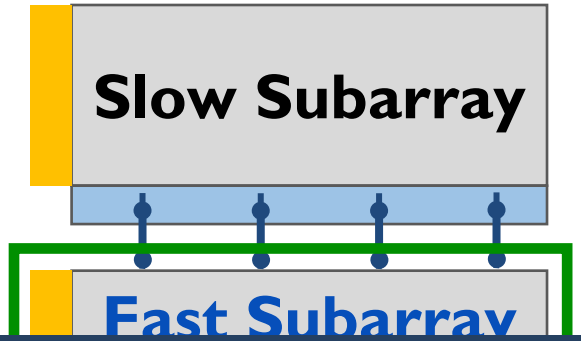
2. Variable Latency DRAM (VILLA)

- **Goal**: Reduce access latency with low area overhead
- **Motivation**: Trade-off between area and latency



2. Variable Latency DRAM (VILLA)

- **Key idea**: **Heterogeneous DRAM** design by adding a few fast subarrays as a low-cost cache in each bank
- **Benefits**: Reduce access latency for frequently-accessed data

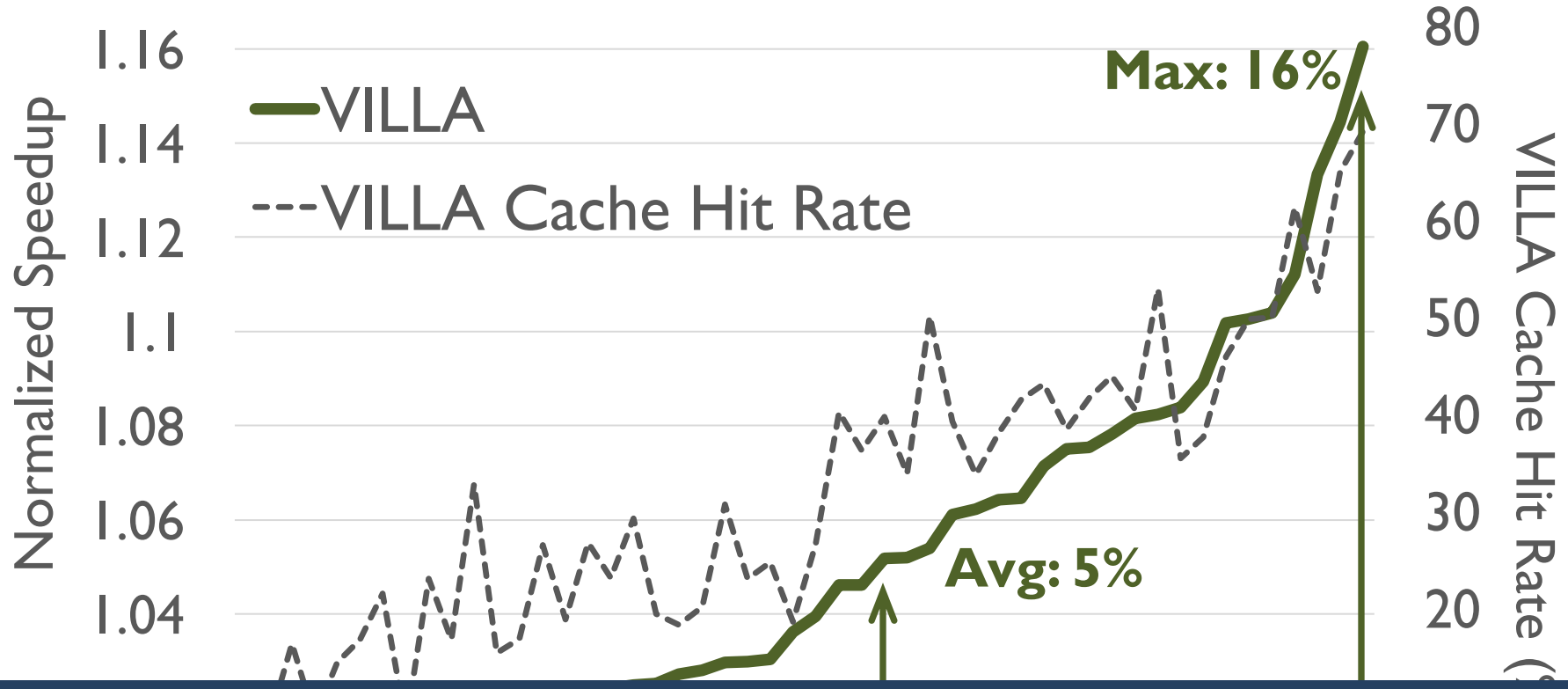


Challenge: How to move data efficiently from slow to fast subarrays?

LISA: Cache rows rapidly from slow

Reduces hot data access latency by 2.2x at only 1.6% area overhead

VILLA Improves System Performance by Caching Hot Data



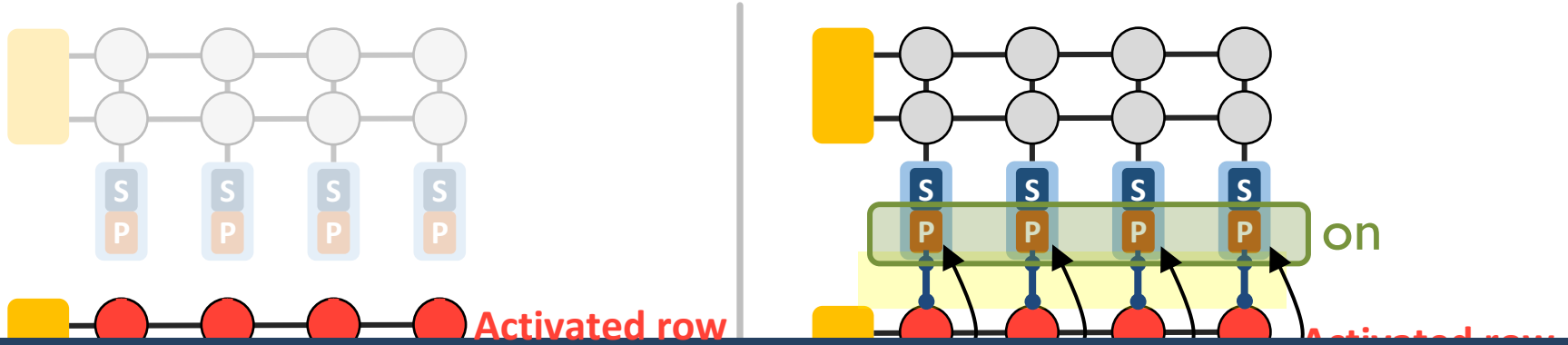
LISA enables an effective in-DRAM caching scheme

Three New Applications of LISA to Reduce Latency

1. Fast bulk data copy
2. In-DRAM caching
3. Fast precharge

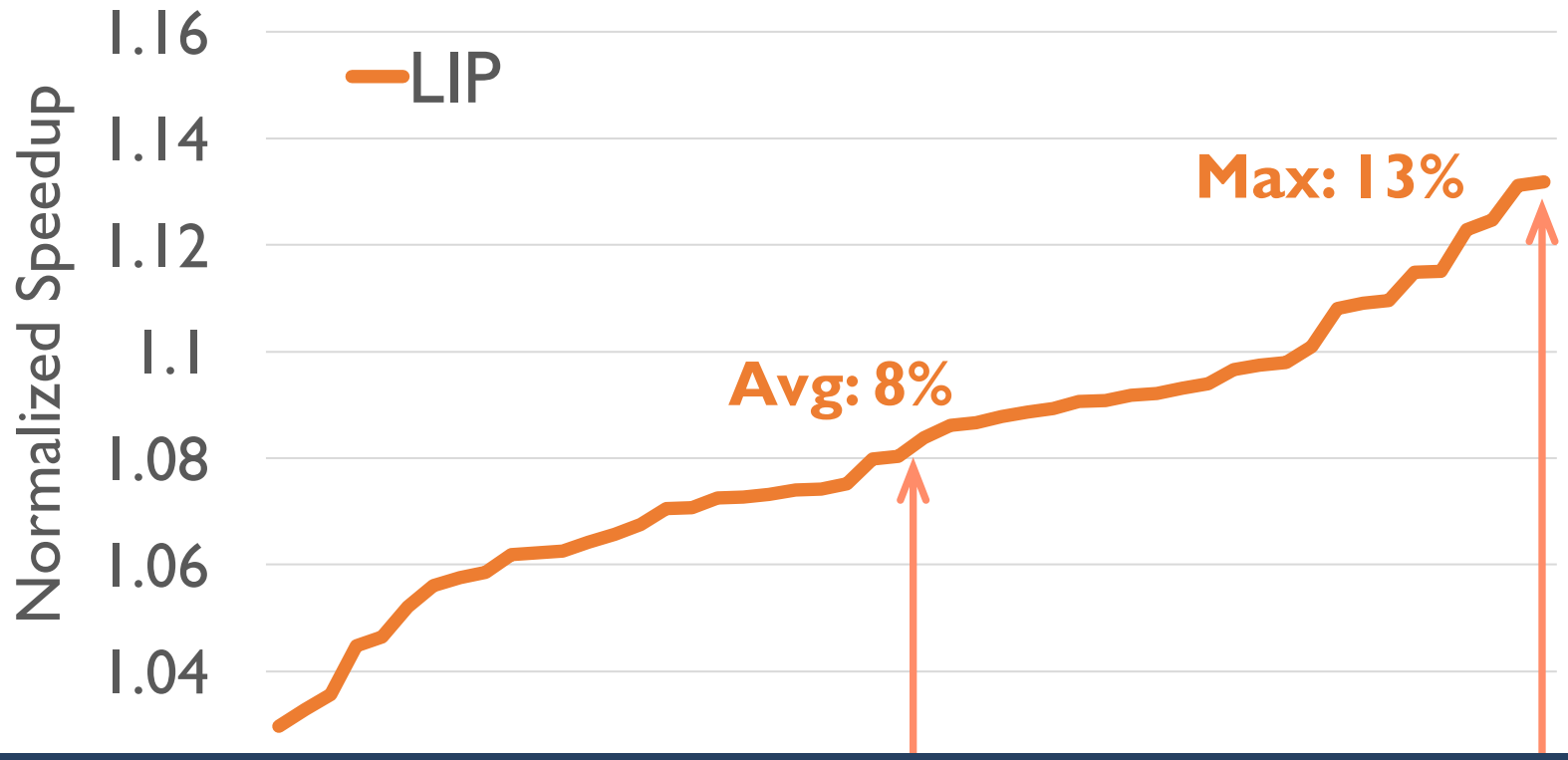
3. Linked Precharge (LIP)

- **Problem**: The precharge time is limited by the strength of one precharge unit
- **Linked Precharge (LIP)**: LISA precharges a subarray using multiple precharge units



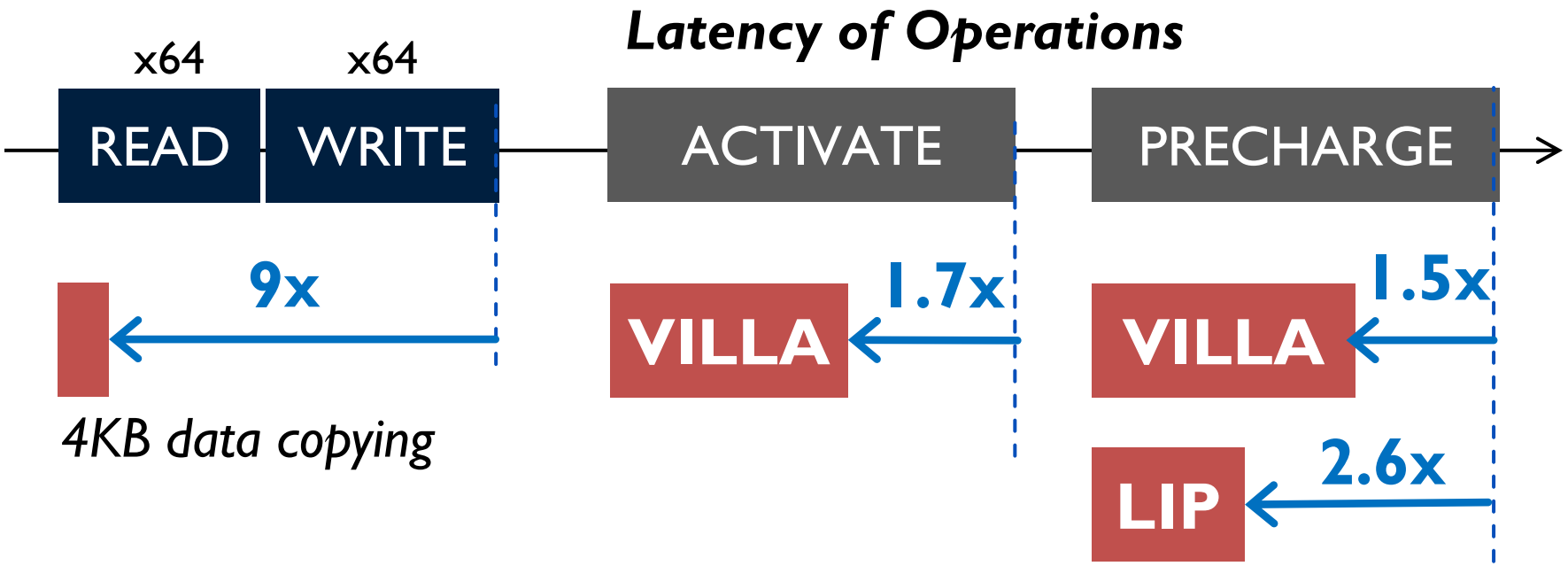
Reduces precharge latency by 2.6x

LIP Improves System Performance by Accelerating Precharge



LISA reduces precharge latency

Latency Reduction of LISA



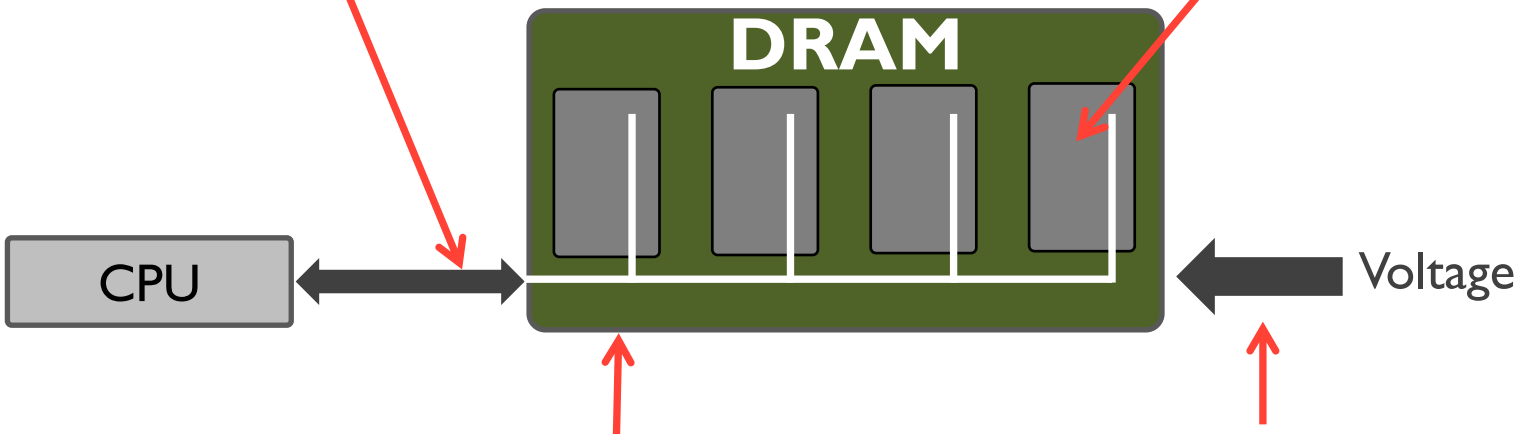
LISA is a versatile substrate that enables many new techniques

Low-Cost Architectural Features in DRAM

Understanding and overcoming the latency limitation in DRAM

Low-Cost Inter-Linked Subarrays (LISA) [HPCA'16]

Understanding and Exploiting Latency Variation in DRAM (FLY-DRAM) [SIGMETRICS'16]



Mitigating Refresh Latency by Parallelizing Accesses with Refreshes (DSARP) [HPCA'14]

Understanding and Exploiting Latency-Voltage Trade-Off (Voltron) [SIGMETRICS'17]

What Does DRAM Latency Mean to You?

- DRAM latency: Delay as specified in DRAM standards

Speed Grade	Data Rate (MT/s)	Target ^{t}RCD-^{t}RP-CL	^{t}RCD (ns)	^{t}RP (ns)	CL (ns)
-075E[1]	2666	18-18-18	13.5	13.5	13.5
-083E[2]	2400	16-16-16	13.32	13.32	13.32
-083[2]	2400	17-17-17	14.16	14.16	14.16
-093E	2133	15-15-15	14.06	14.06	14.06
-093	2133	16-16-16	15	15	15

- Memory controllers use these standardized latency to access DRAM

  "The purpose of this Standard is to define the minimum set of requirements for JEDEC compliant … SDRAM devices"
  (p.1) JEDEC DDRx standard

- Key question: How does reducing latency affect DRAM accesses?

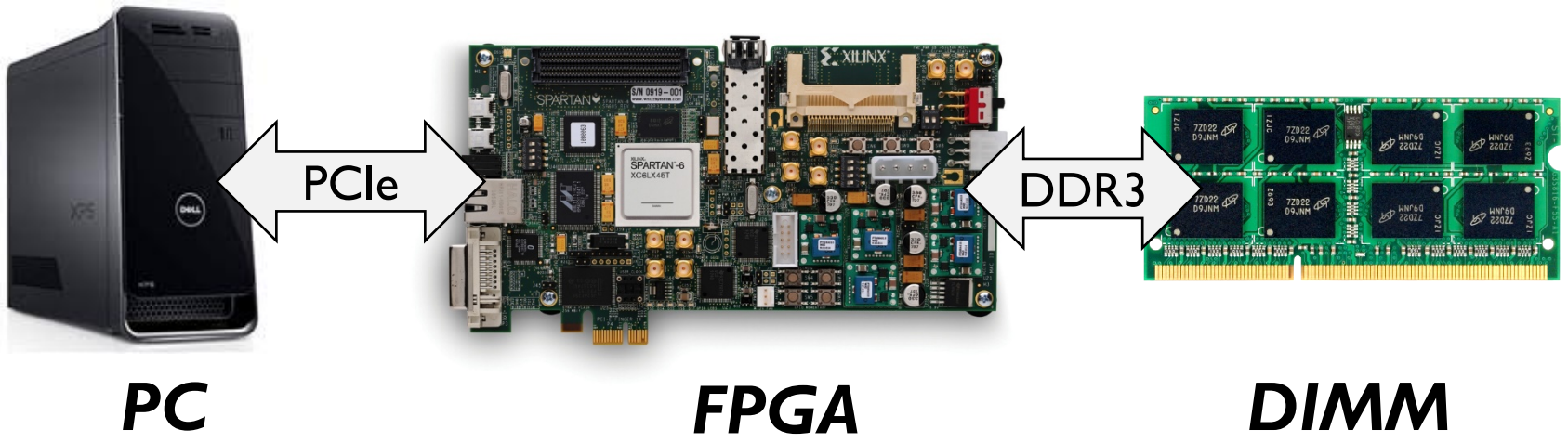
Goals

1 Understand and characterize reduced-latency behavior in modern DRAM chips

2 Develop a mechanism that exploits our observation to improve DRAM latency

Experimental Setup

- Custom FPGA-based infrastructure
  - Existing systems: Commands are generated and controlled by HW



Experiments

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

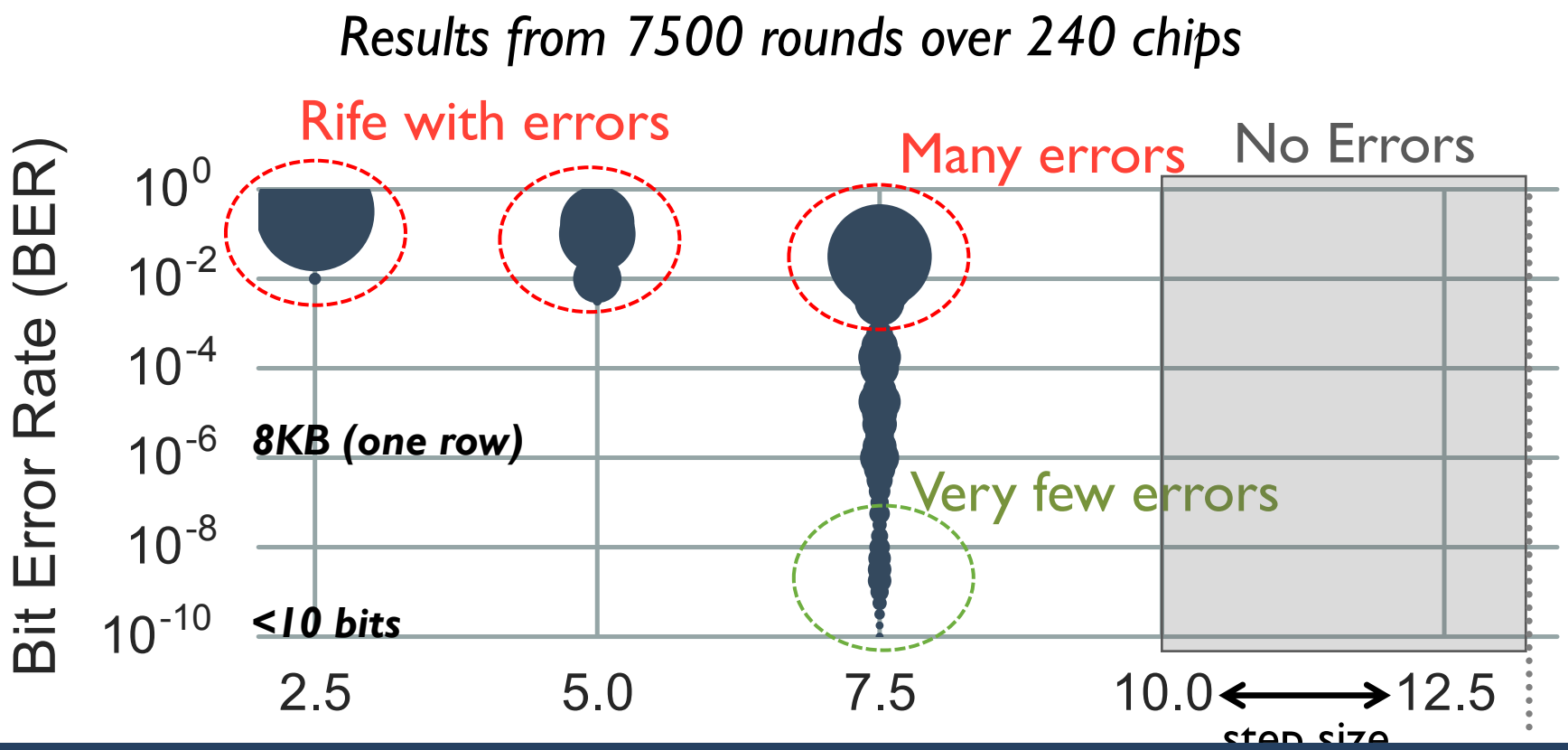
- Check the correctness of data read back from DRAM
  - Any errors (bit flips)?

- Tested 240 DDR3 DRAM chips from three vendors
  - 30 DIMMs
  - Capacity: 1GB

Experimental Results

Activation Latency

Variation in Activation Errors

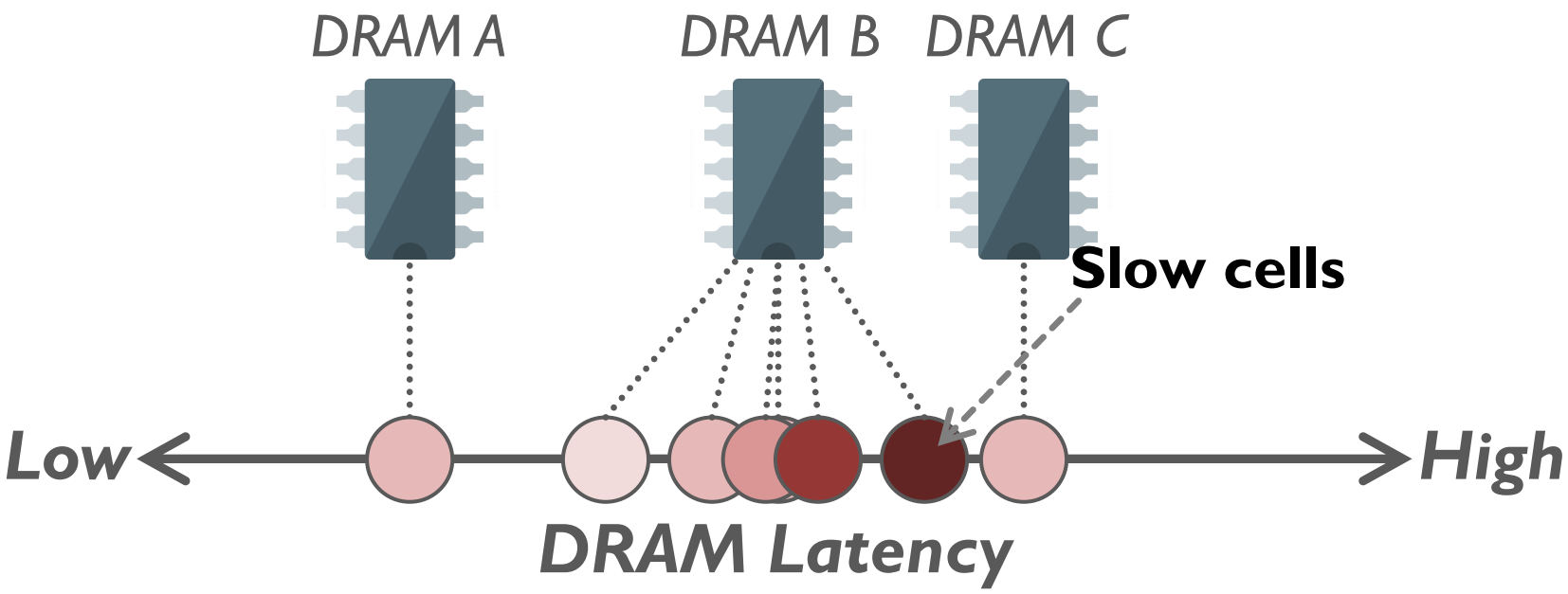


Modern DRAM chips exhibit significant variation in activation latency

DRAM Latency Variation

Imperfect manufacturing process →

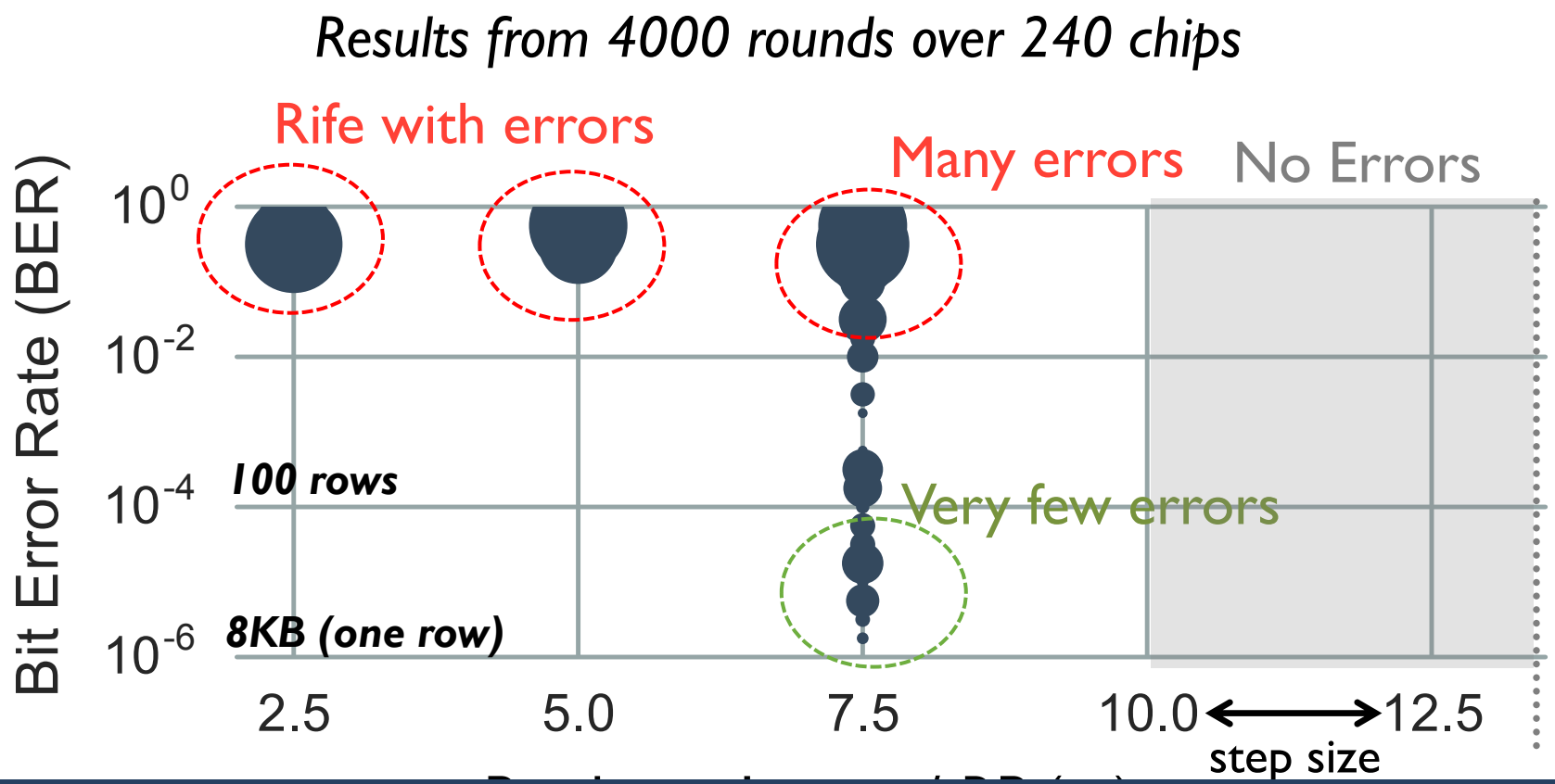
latency variation in timing parameters



Experimental Results

Precharge Latency

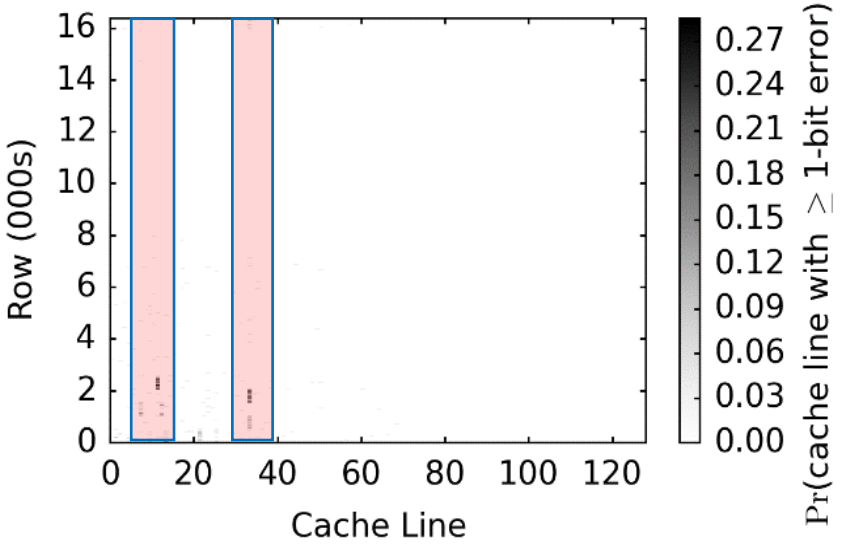
Variation in Precharge Errors



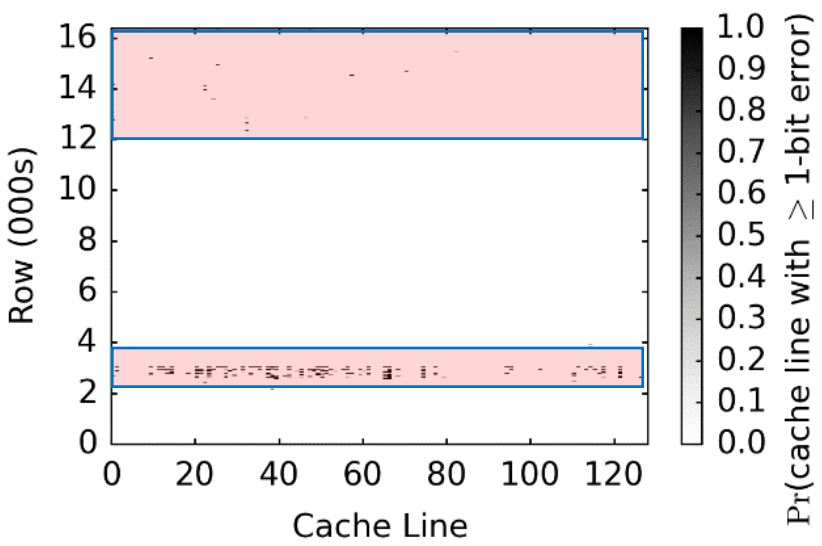
Modern DRAM chips exhibit significant variation in precharge latency

Spatial Locality of Slow Cells

One DIMM: tRCD=7.5ns



One DIMM: tRP=7.5ns



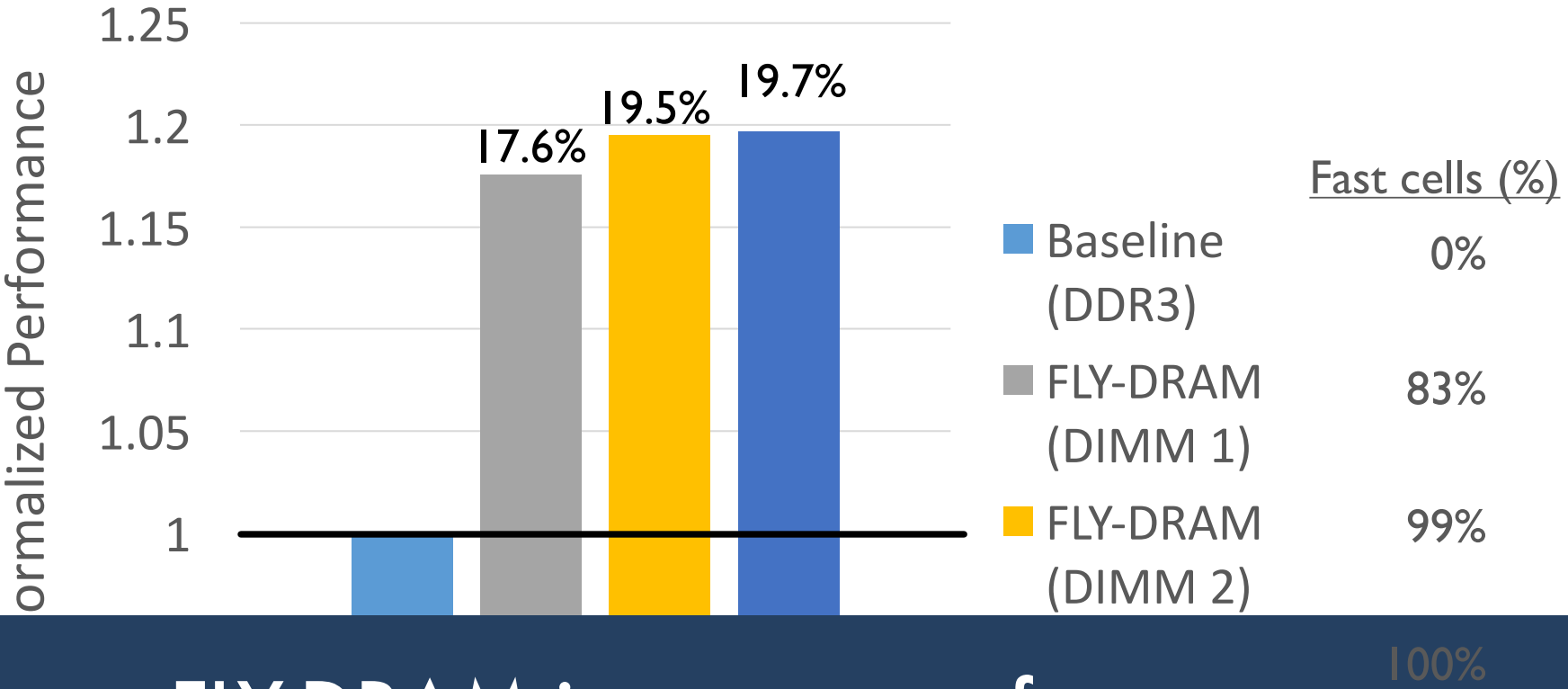
Slow cells are concentrated at certain regions

Mechanism:
Flexible-Latency (FLY) DRAM

Mechanism to Reduce DRAM Latency

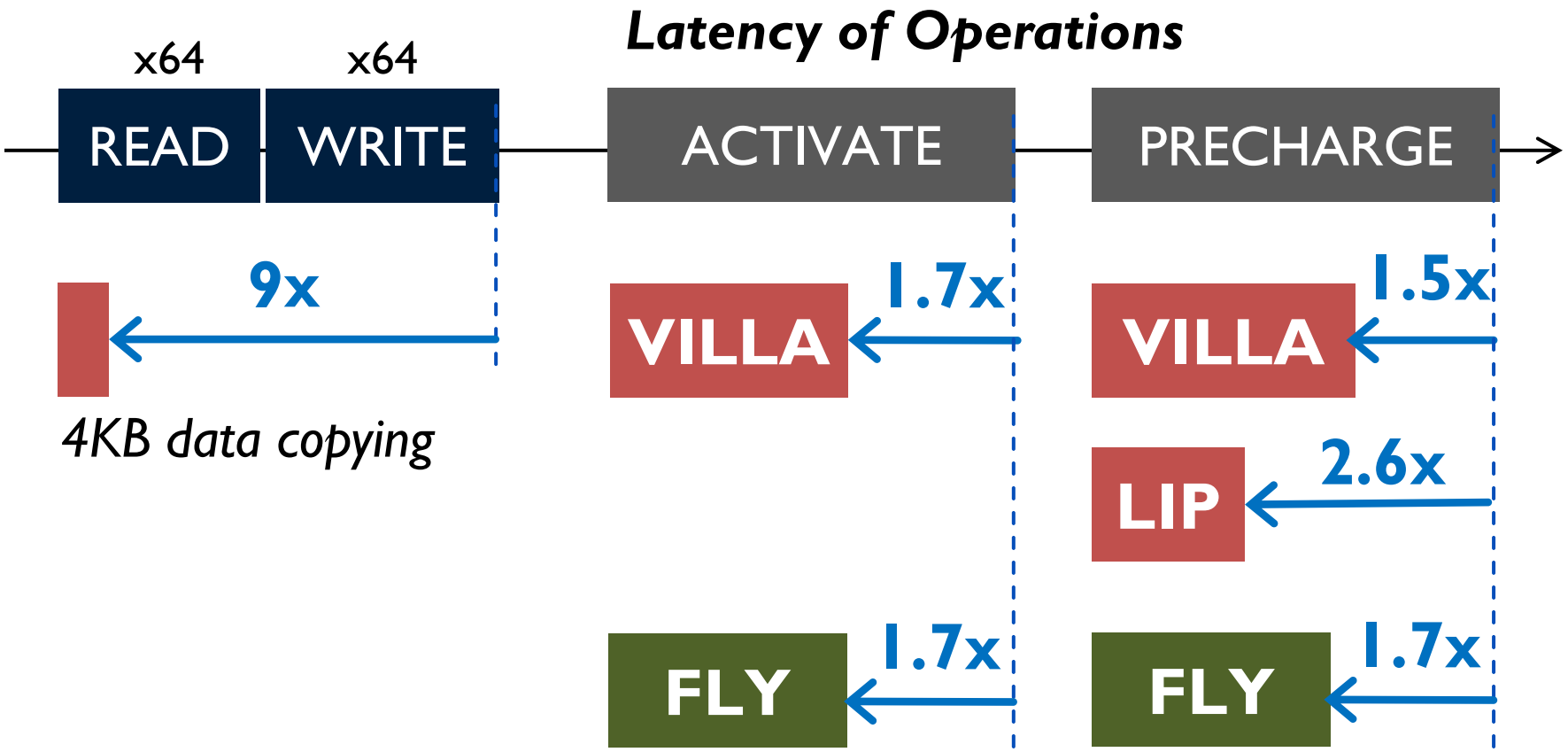
- **Observation:** DRAM timing errors (slow DRAM cells) are concentrated on certain regions
- **Flexible-LatencY (FLY) DRAM**
  - A memory controller design that reduces latency
- **Key idea**:
  1) Divide memory into regions of different latencies
  2) *Memory controller:* Use lower latency for regions without slow cells; higher latency for other regions
- Latency profile through DRAM vendors or online tests

Benefits of FLY-DRAM



FLY-DRAM improves performance by exploiting latency variation in DRAM

Latency Reduction of FLY-DRAM



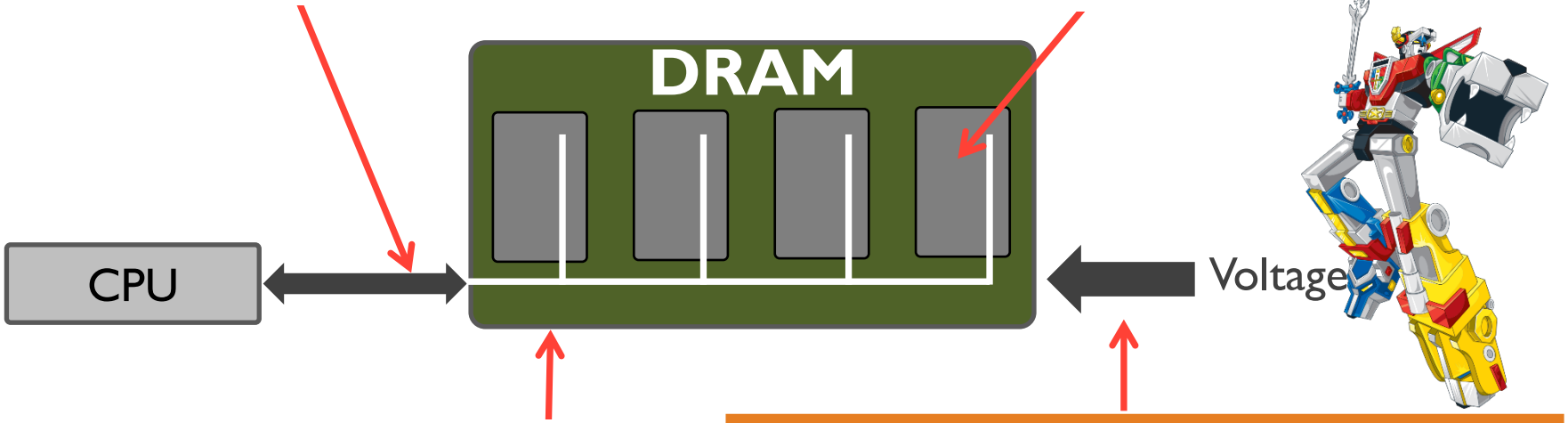
Experimental demonstration of latency variation enables techniques to reduce latency

Low-Cost Architectural Features in DRAM

Understanding and overcoming the latency limitation in DRAM

Low-Cost Inter-Linked Subarrays (LISA) [HPCA'16]

Understanding and Exploiting Latency Variation in DRAM (FLY-DRAM) [SIGMETRICS'16]



Mitigating Refresh Latency by Parallelizing Accesses with Refreshes (DSARP) [HPCA'14]

Understanding and Exploiting Latency-Voltage Trade-Off (Voltron) [SIGMETRICS'17]

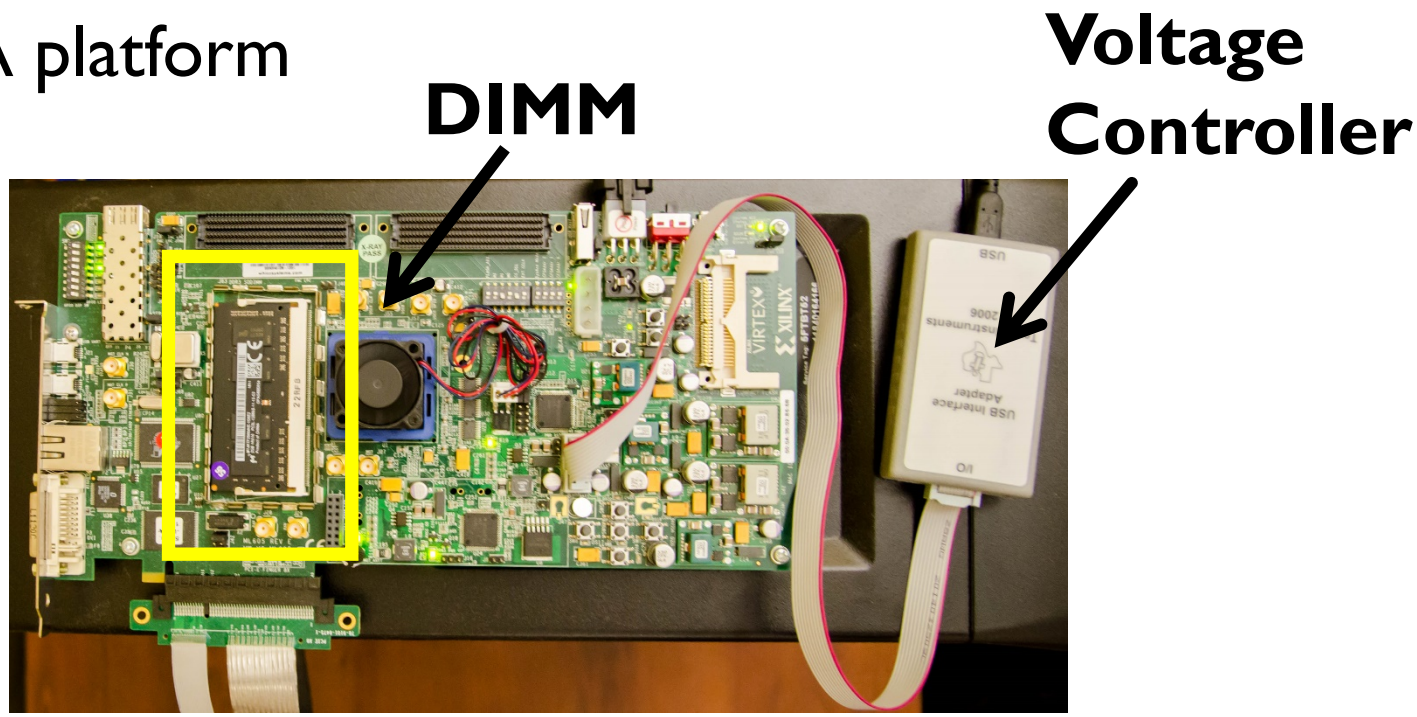
Motivation

- DRAM voltage is an important factor that affects: latency, power, and reliability

- Goal: Understand the relationship between latency and DRAM voltage and exploit this trade-off

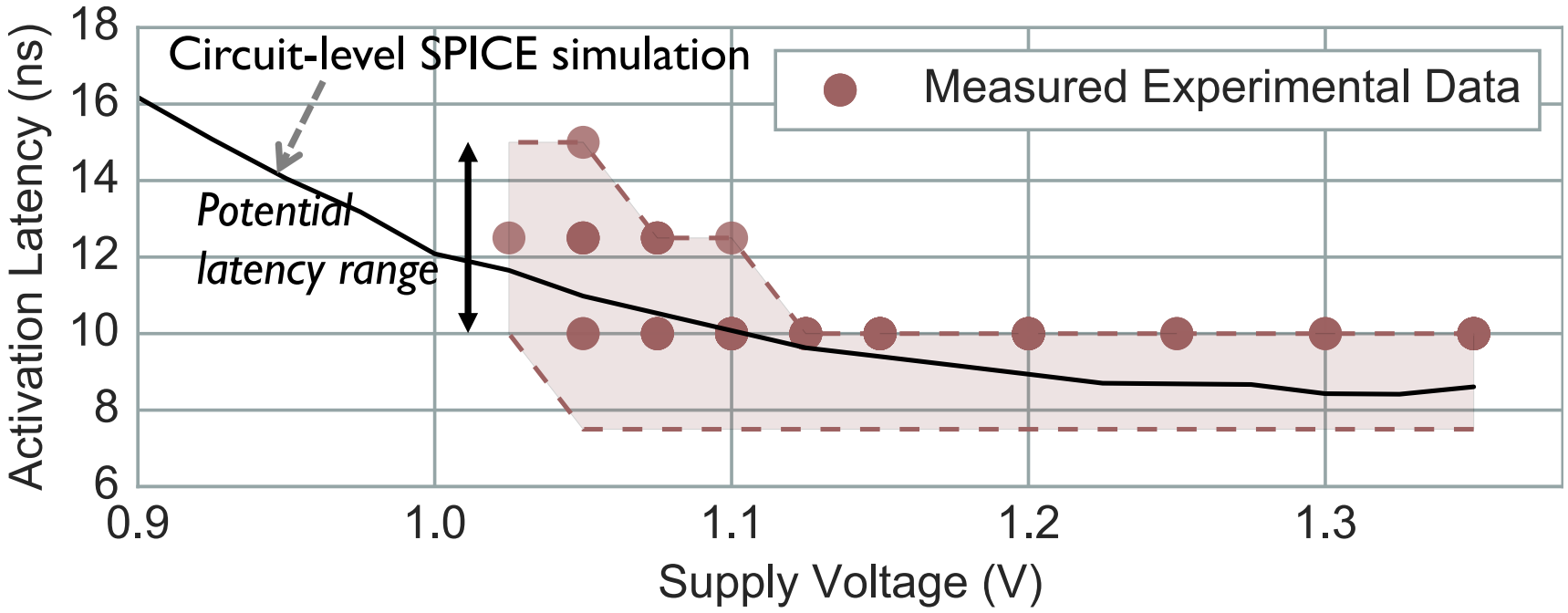
Methodology

- FPGA platform



- Tested 124 DDR3L DRAM chips (31 DIMMs)

Key Result: Voltage vs. Latency



Trade-off between access latency and voltage

Goal and Key Observation

- Goal: Exploit the trade-off between voltage and latency to reduce energy consumption
- Approach: Reduce voltage
  - Performance loss due to increased latency
  - Energy: Function of time (performance) and power (voltage)
- Observation: Application's performance loss due to higher latency has a strong linear relationship with its memory intensity

Mechanism: Voltron

- Build a performance (linear) model to predict performance loss based on the selected voltage value
- Use the model to select a minimum voltage that satisfies a performance loss target specified by the user
- Results: Reduces system energy by 7.3% with a small performance loss of 1.8%

Reducing Latency by Exploiting Voltage-Latency Trade-Off

- Voltron exploits the latency-voltage trade-off to improve energy efficiency

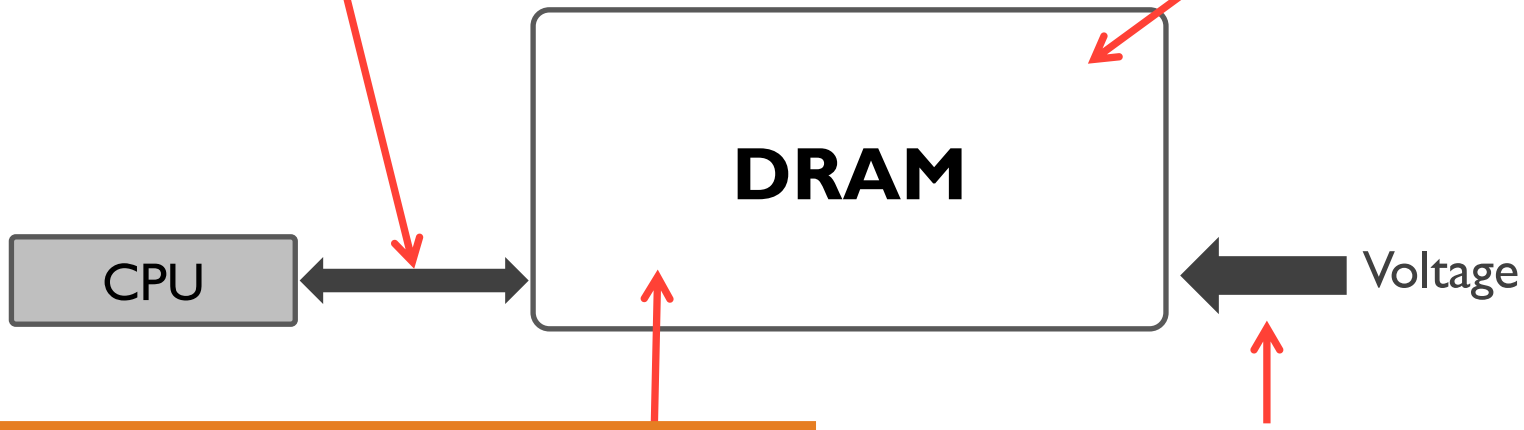
- Another perspective: Increase voltage to reduce latency

Low-Cost Architectural Features in DRAM

Understanding and overcoming the latency limitation in DRAM

Low-Cost Inter-Linked Subarrays (LISA) [HPCA'16]

Understanding and Exploiting Latency Variation in DRAM (FLY-DRAM) [SIGMETRICS'16]



Mitigating Refresh Latency by Parallelizing Accesses with Refreshes (DSARP) [HPCA'14]

Understanding and Exploiting Latency-Voltage Trade-Off (Voltron) [SIGMETRICS'17]

Summary of DSARP

- Problem: Refreshing DRAM blocks memory accesses
  - Prolongs latency of memory requests
- Goal: Reduce refresh-induced latency on demand requests
- Key observation: Some subarrays and I/O remain completely **idle** during refresh
- Dynamic Subarray Access-Refresh Parallelization (**DSARP**):
  - DRAM modification to enable idle DRAM subarrays to serve accesses during refresh
  - **0.7%** DRAM area overhead
- 20.2% system performance improvement for 8-core systems using 32Gb DRAM

Prior Work on Low-Latency DRAM

- Uniform short-bitlines DRAM: *FCRAM*, *RLDRAM*
  - Large area overhead (30% - 80%)
- Heterogeneous bitline design
  - TL-DRAM: Intra-subarray [Lee+, HPCA'13]
    Requires two fast rows to cache one slow row
  - CHARM: Inter-bank [Son+, ISCA'13]
    High movement cost between slow and fast banks
- SRAM cache in DRAM [Hidaka+, IEEE Micro'90]
  - Large area overhead (38% for 64KB) and complex control
- Our work:
  - Low cost
  - Detailed experimental understanding via characterization of commodity chips

CONCLUSION

Conclusion

- Memory latency has remained mostly constant over the past decade
  - System performance bottleneck for modern applications

- Simple and low-cost architectural mechanisms
  - New DRAM substrate for fast inter-subarray data movement
  - Refresh architecture to mitigate refresh interference

- Understanding latency behavior in commodity DRAM
  - Experimental characterization of:
    1) Latency variation inside DRAM
    2) Relationship between latency and DRAM voltage

Thesis Statement

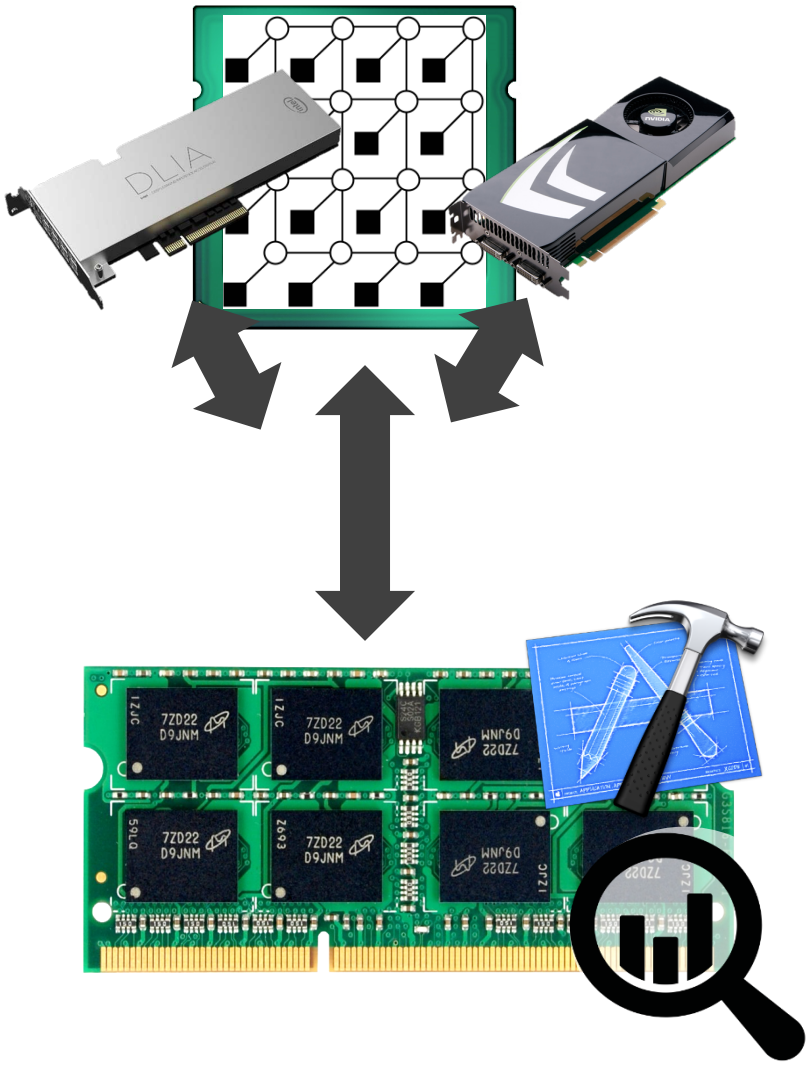
Memory latency can be significantly reduced with a multitude of **low-cost architectural techniques** that aim to **reduce different causes of long latency**

Future Research Direction

- Latency characterization and optimization for other memory technologies
  - eDRAM
  - Non-volatile memory: PCM, STT-RAM, etc.

- Understanding other aspects of DRAM
  - Variation in power/energy consumption
  - Security/reliability

Other Areas Investigated



Energy Efficient Networks-On-Chip
[NOCS'12, SBACPAD'12, SBACPAD'14]

Memory Schedulers for Heterogeneous Systems
[ISCA'12, TACO'16]

Low-latency DRAM Architecture
[HPCA'15]

DRAM Testing Platform
[HPCA'17]

Acknowledgements

- Onur Mutlu
- James Hoe, Kayv[illegible] Fatahalian, Moinuddin [illegible]reshi, and Steve Keckler
- Safari group: [illegible]chata Aus[illegible]ngnirun, Amirali Bo[illegible]oumand, [illegible]ris Fallin, Saugata Ghose, Hasan H[illegible]an, Kevin [illegible]engjan[illegible] Kashyap, [illegible]mira Khan, Yoongu Kim, Donghyuk [illegible]e, Yang Li[illegible] [illegible]Meza, Ger[illegible]dy Pekhimenko, Vivek Seshadri, [illegible]anya Subramanian, Nandita Vijaykumar, Hanbin [illegible]on, Hongyi Xin
- Georgia Tec[illegible] collabora[illegible] [illegible]Nair, Jaewoong [illegible]m
- CALCM grou[illegible]
- Friends
- Family — parents, [illegible]er, and girlfriend
- Intern mentors and indus[illegible]ators:

Sponsors

- Intel and SRC for my fellowship
- NSF and DOE
- Facebook, Google, Intel, NVIDIA, VMware, Samsung

Thesis Related Publications

- **Improving DRAM Performance by Parallelizing Refreshes with Accesses**
  Kevin Chang, Donghyuk Lee, Zeshan Chishti, Alaa Alameldeen, Chris Wilkerson, Yoongu Kim, Onur Mutlu
  HPCA 2014
- **Low-Cost Inter-Linked Subarrays (LISA): Enabling Fast Inter-Subarray Data Movement in DRAM**
  Kevin Chang, Prashant J. Nair, Donghyuk Lee, Saugata Ghose, Moinuddin K. Qureshi, and Onur Mutlu
  HPCA 2016
- **Understanding Latency Variation in Modern DRAM Chips: Experimental Characterization, Analysis, and Optimization**
  Kevin Chang, Abhijith Kashyap, Hasan Hassan Samira Khan, Kevin Hsieh, Donghyuk Lee, Saugata Ghose, Gennady Pekhimenko, Tianshi Li, Onur Mutlu
  SIGMETRICS 2016
- **Understanding Reduced-Voltage Operation in Modern DRAM Devices: Experimental Characterization, Analysis, and Mechanisms**
  Kevin Chang, Abdullah Giray Yağlikçi, Saugata Ghose, Aditya Agrawal, Niladrish Chatterjee, Abhijith Kashyap, Donghyuk Lee, Mike O'connor, Hasan Hassan, Onur Mutlu
  SIGMETRICS 2017

Understanding and Improving Latency of DRAM-Based Memory Systems

Thesis Oral

Kevin Chang

Committee:

Prof. Onur Mutlu (Chair)

Prof. James Hoe

Prof. Kayvon Fatahalian

Prof. Stephen Keckler (NVIDIA, UT Austin)

Prof. Moinuddin Qureshi (Georgia Tech.)

Carnegie Mellon University