

Computer Architecture

Lecture 6: Computation in Memory

Prof. Onur Mutlu

ETH Zürich

Fall 2020

8 October 2020

Sub-Agenda: In-Memory Computation

- Major Trends Affecting Main Memory
- **The Need for Intelligent Memory Controllers**
 - **Bottom Up: Push from Circuits and Devices**
 - **Top Down: Pull from Systems and Applications**
- Processing in Memory: Two Directions
 - Minimally Changing Memory Chips
 - Exploiting 3D-Stacked Memory
- How to Enable Adoption of Processing in Memory
- Conclusion

Three Key Systems Trends

1. Data access is a major bottleneck

- ▣ Applications are increasingly data hungry

2. Energy consumption is a key limiter

3. Data movement energy dominates compute

- ▣ Especially true for off-chip to on-chip movement

Observation and Opportunity

- High latency and high energy caused by data movement
 - ❑ Long, energy-hungry interconnects
 - ❑ Energy-hungry electrical interfaces
 - ❑ Movement of large amounts of data
- Opportunity: Minimize data movement by performing computation directly (near) where the data resides
 - ❑ Processing in memory (PIM)
 - ❑ In-memory computation/processing
 - ❑ Near-data processing (NDP)
 - ❑ General concept applicable to any data storage & movement unit (caches, SSDs, main memory, network, controllers)

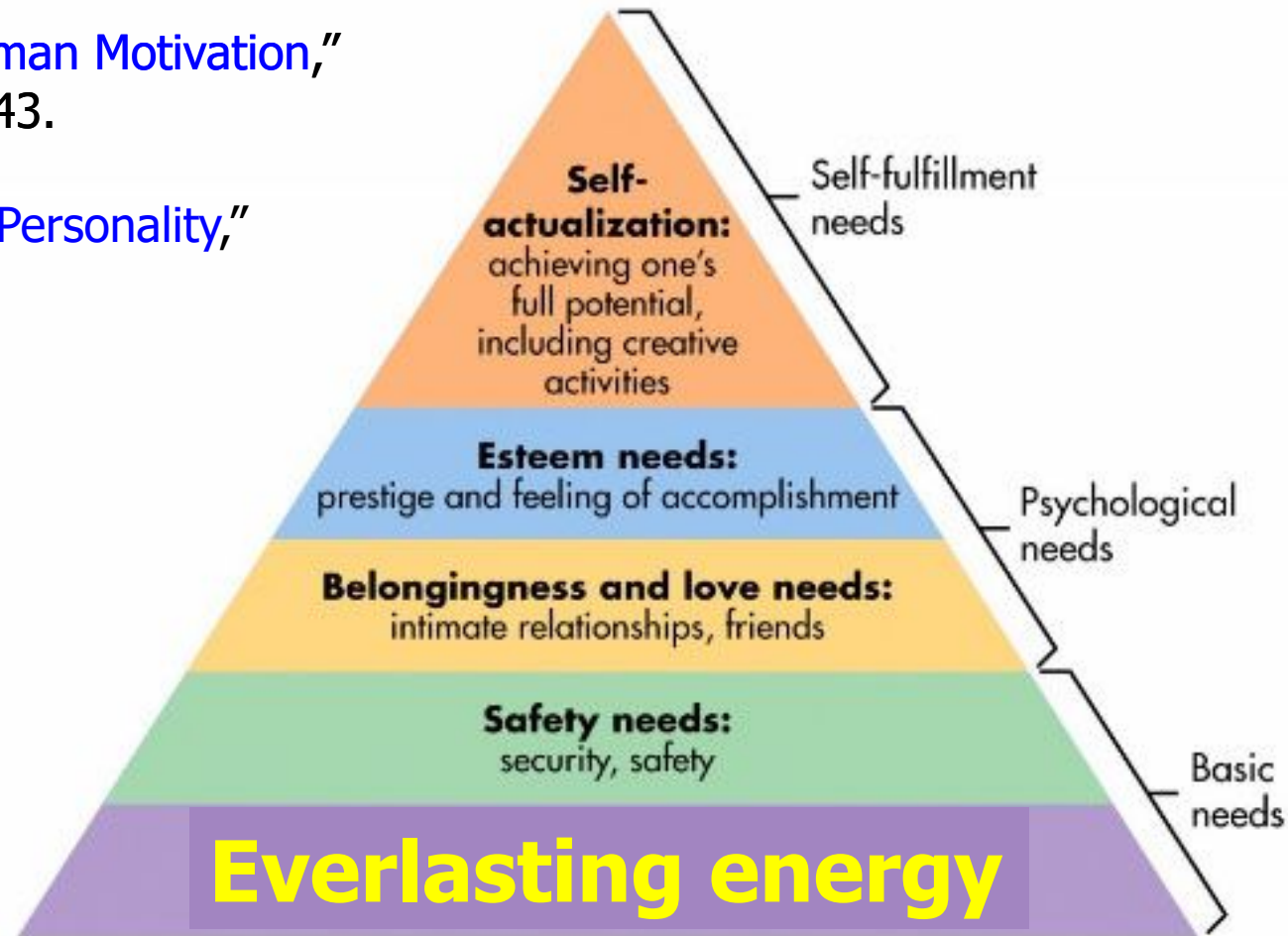
Four Key Issues in Future Platforms

- Fundamentally **Secure/Reliable/Safe** Architectures
- Fundamentally **Energy-Efficient** Architectures
 - **Memory-centric** (Data-centric) Architectures
- Fundamentally **Low-Latency** Architectures
- Architectures for **Genomics, Medicine, Health**

Maslow's (Human) Hierarchy of Needs, Revisited

Maslow, "A Theory of Human Motivation,"
Psychological Review, 1943.

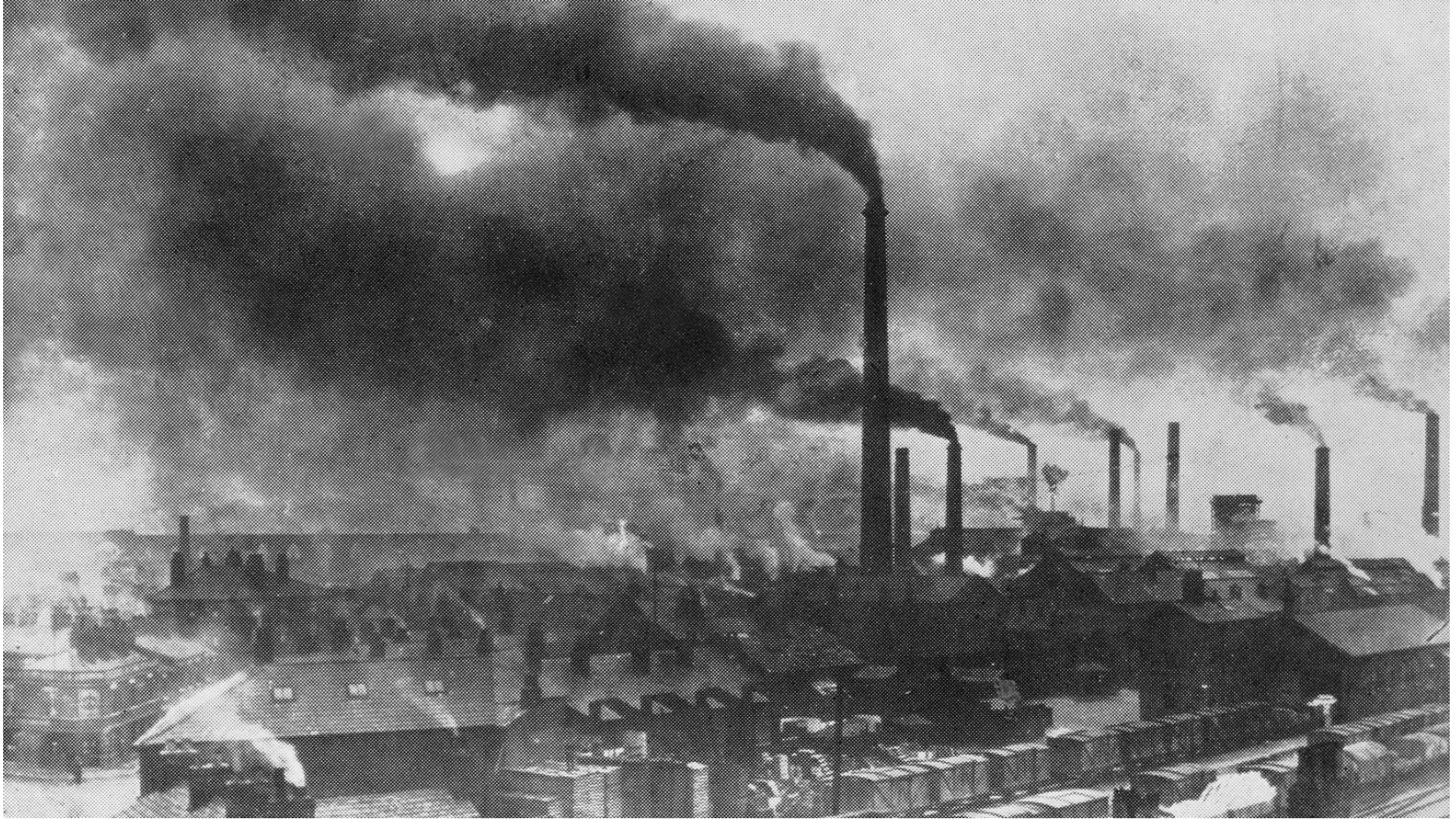
Maslow, "Motivation and Personality,"
Book, 1954-1970.



Do We Want This?



Or This?



High Performance,
Energy Efficient,
Sustainable

The Problem

Data access is the major performance and energy bottleneck

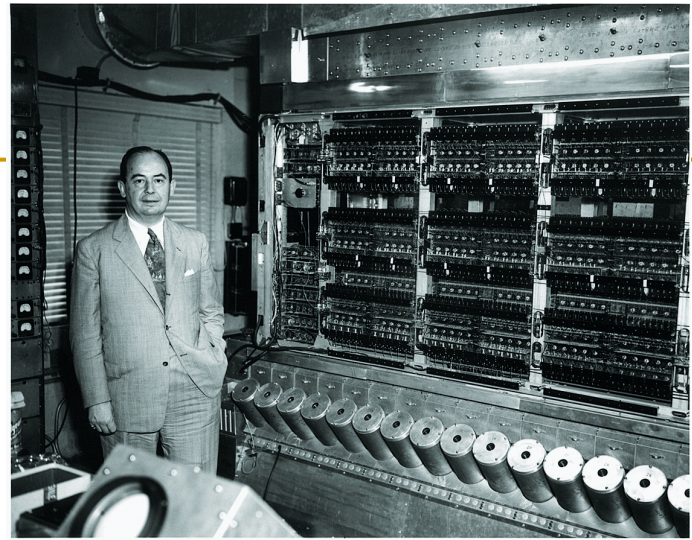
Our current
design principles
cause great energy waste
(and great performance loss)

The Problem

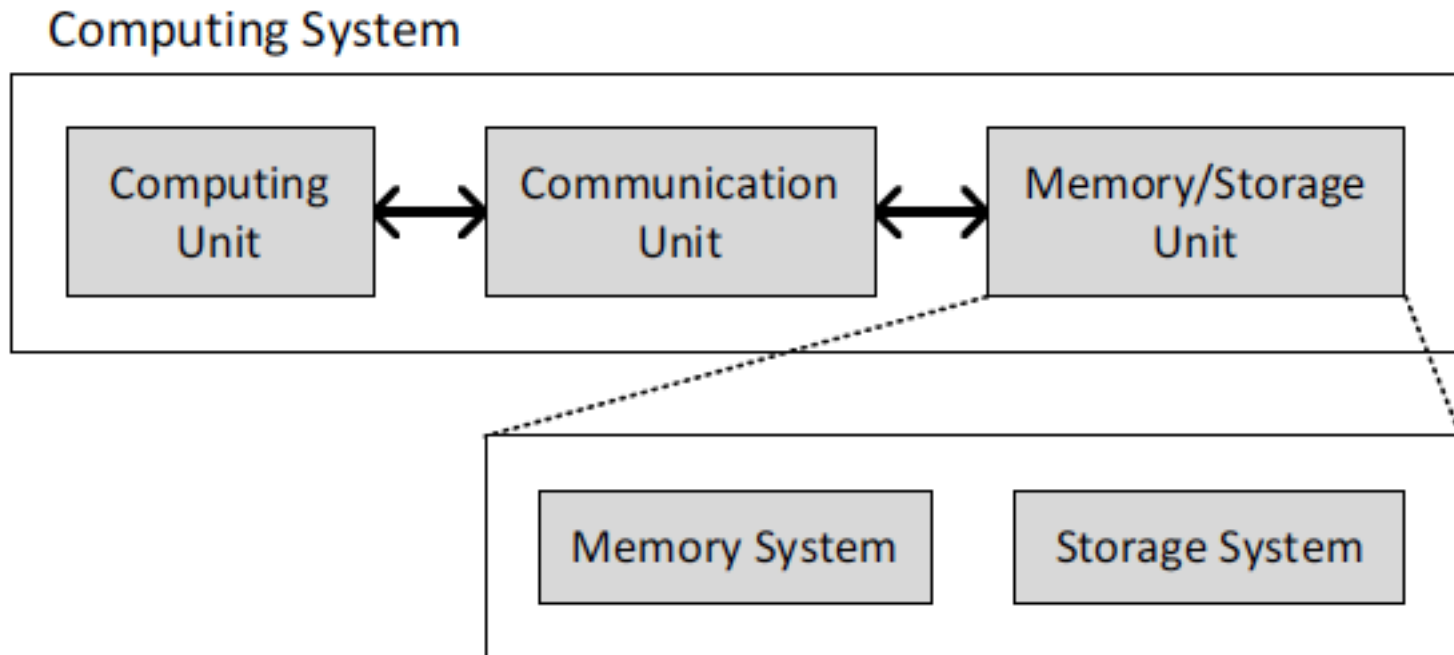
Processing of data
is performed
far away from the data

A Computing System

- Three key components
- Computation
- Communication
- Storage/memory

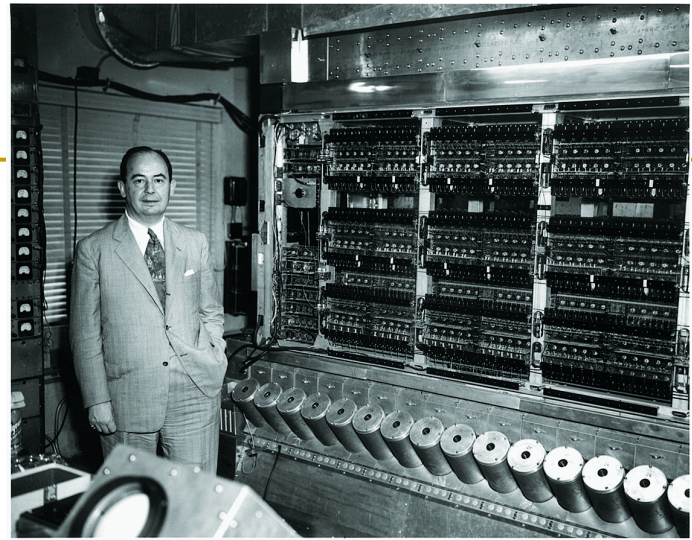


Burks, Goldstein, von Neumann, "Preliminary discussion of the logical design of an electronic computing instrument," 1946.



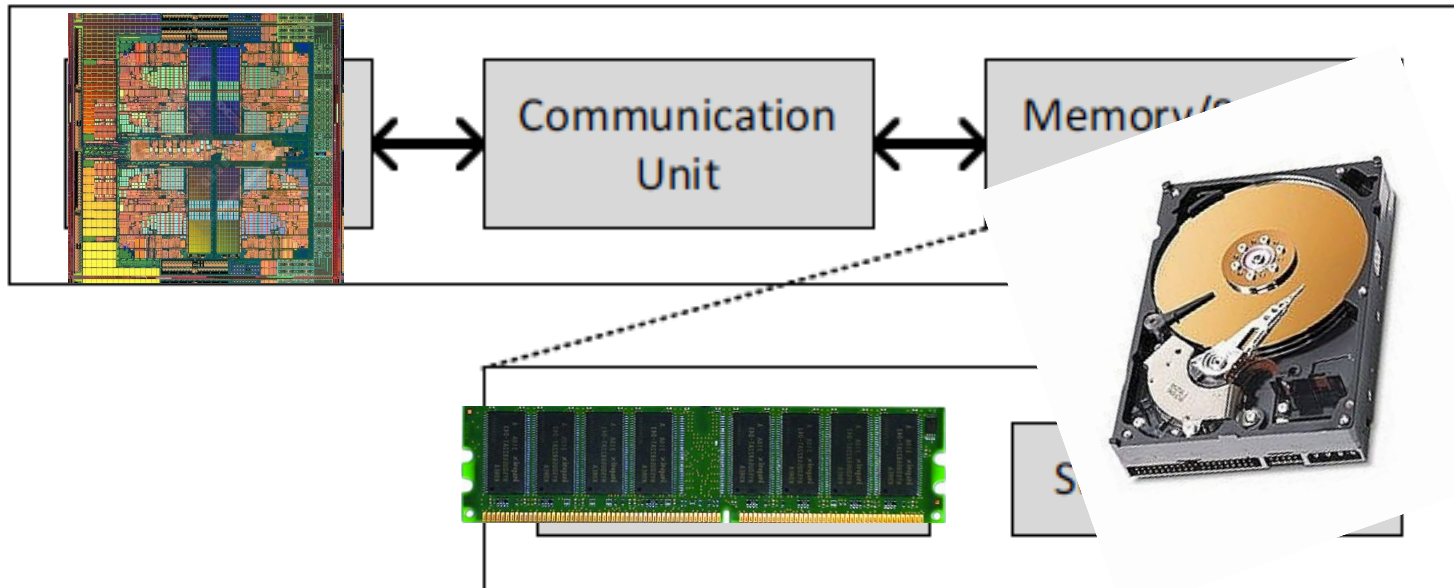
A Computing System

- Three key components
- Computation
- Communication
- Storage/memory



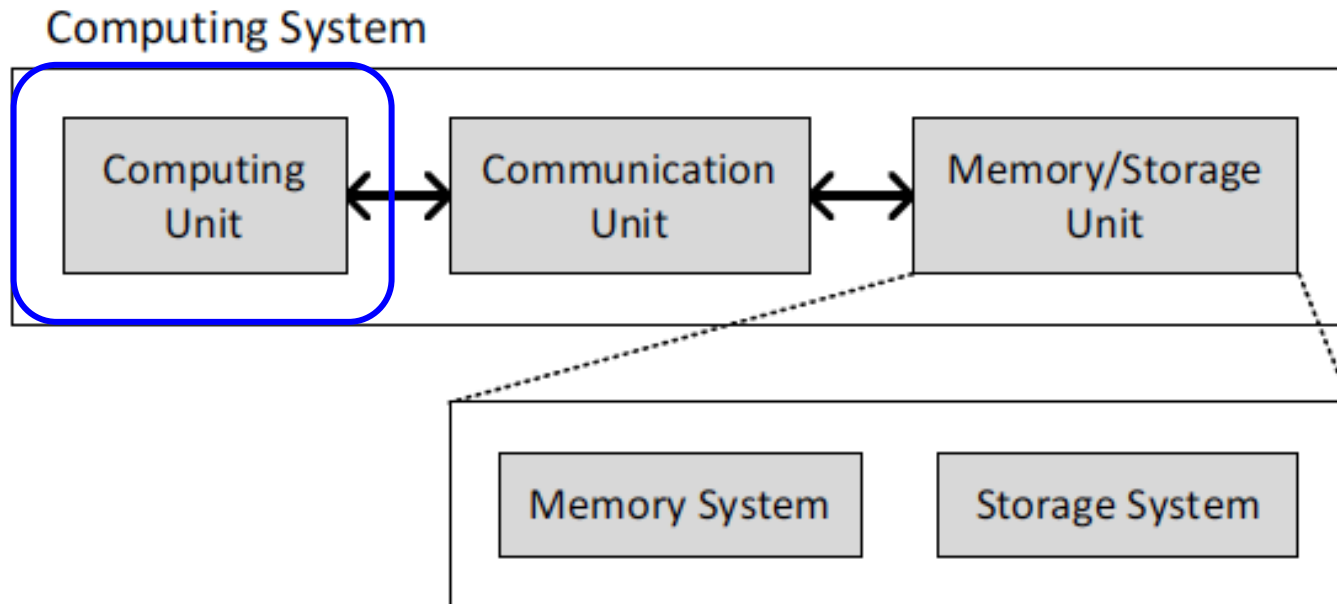
Burks, Goldstein, von Neumann, "Preliminary discussion of the logical design of an electronic computing instrument," 1946.

Computing System



Today's Computing Systems

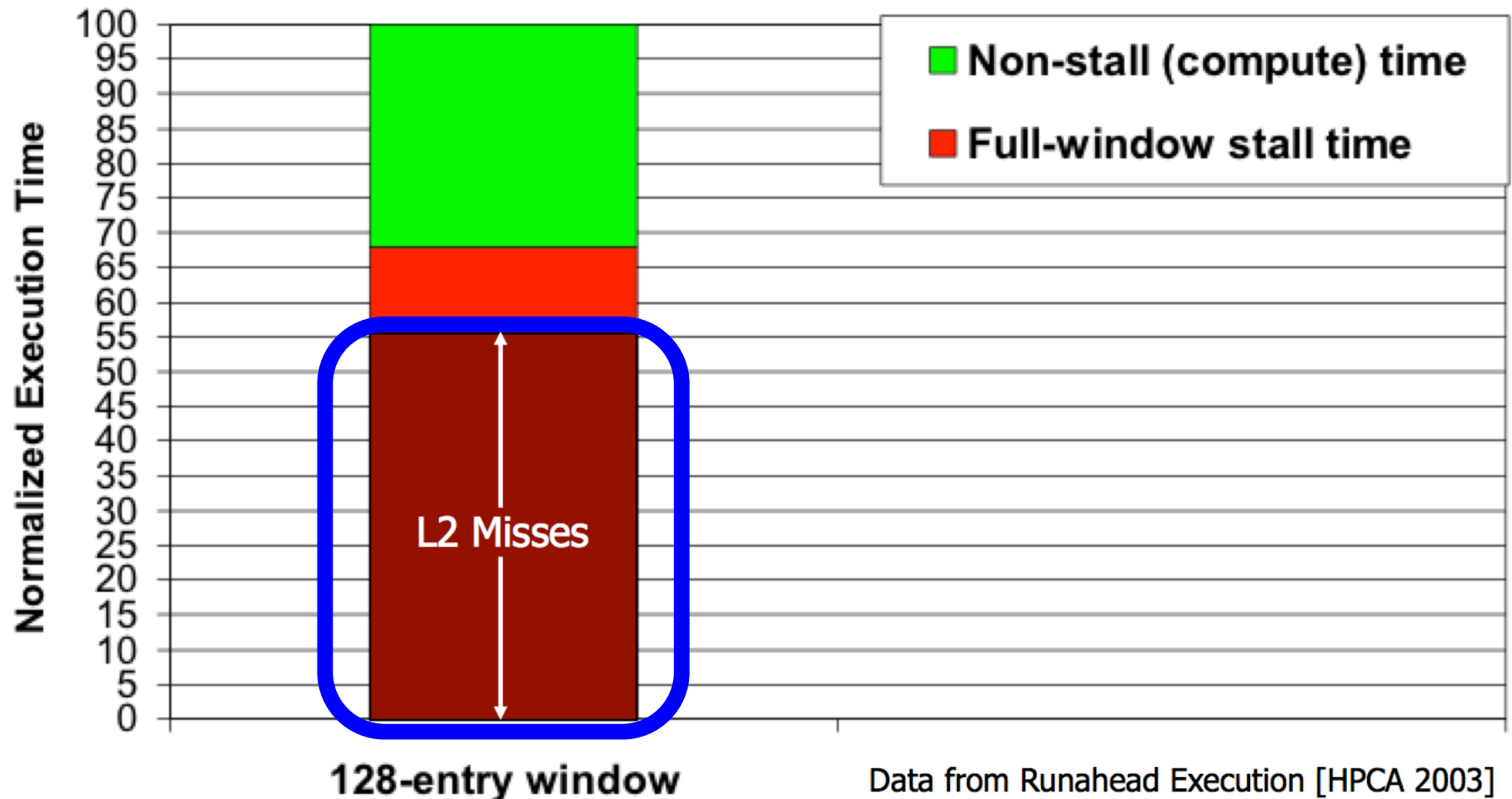
- Are overwhelmingly processor centric
- All data processed in the processor → at great system cost
- Processor is heavily optimized and is considered the master
- Data storage units are dumb and are largely unoptimized (except for some that are on the processor die)



Yet ...

I expect that over the coming decade memory subsystem design will be the *only* important design issue for microprocessors.

- **“It’s the Memory, Stupid!”** (Richard Sites, MPR, 1996)



The Performance Perspective

- Onur Mutlu, Jared Stark, Chris Wilkerson, and Yale N. Patt,
"Runahead Execution: An Alternative to Very Large Instruction Windows for Out-of-order Processors"
Proceedings of the 9th International Symposium on High-Performance Computer Architecture (HPCA), Anaheim, CA, February 2003. [Slides \(pdf\)](#)
One of the 15 computer architecture papers of 2003 selected as Top Picks by IEEE Micro.

Runahead Execution: An Alternative to Very Large Instruction Windows for Out-of-order Processors

Onur Mutlu § Jared Stark † Chris Wilkerson ‡ Yale N. Patt §

§ECE Department
The University of Texas at Austin
{onur,patt}@ece.utexas.edu

†Microprocessor Research
Intel Labs
jared.w.stark@intel.com

‡Desktop Platforms Group
Intel Corporation
chris.wilkerson@intel.com

The Memory Bottleneck

- Onur Mutlu, Jared Stark, Chris Wilkerson, and Yale N. Patt,
"Runahead Execution: An Effective Alternative to Large Instruction Windows"
IEEE Micro, Special Issue: Micro's Top Picks from Microarchitecture Conferences (MICRO TOP PICKS), Vol. 23, No. 6, pages 20-25, November/December 2003.

RUNAHEAD EXECUTION: AN EFFECTIVE ALTERNATIVE TO LARGE INSTRUCTION WINDOWS

It's the Memory, Stupid!

RICHARD SITES

It's the Memory, Stupid!

When we started the Alpha architecture design in 1988, we estimated a 25-year lifetime and a relatively modest 32% per year compounded performance improvement of implementations over that lifetime (1,000× total). We guesstimated about 10× would come from CPU clock improvement, 10× from multiple instruction issue, and 10× from multiple processors.

5, 1996  MICROPROCESSOR REPORT

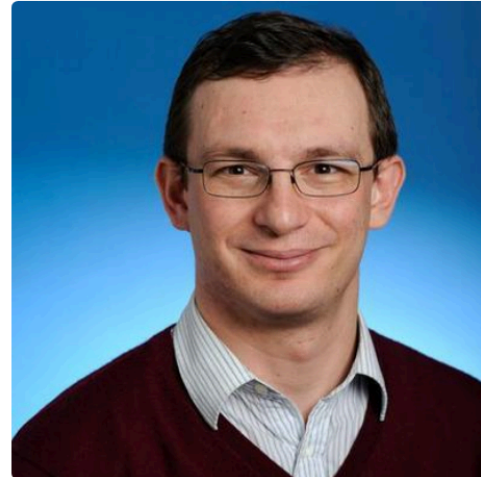
An Informal Interview on Memory

- Madeleine Gray and Onur Mutlu,
"It's the memory, stupid': A conversation with Onur Mutlu"
HiPEAC info 55, HiPEAC Newsletter, October 2018.
[Shorter Version in Newsletter]
[Longer Online Version with References]

'It's the memory, stupid': A conversation with Onur Mutlu

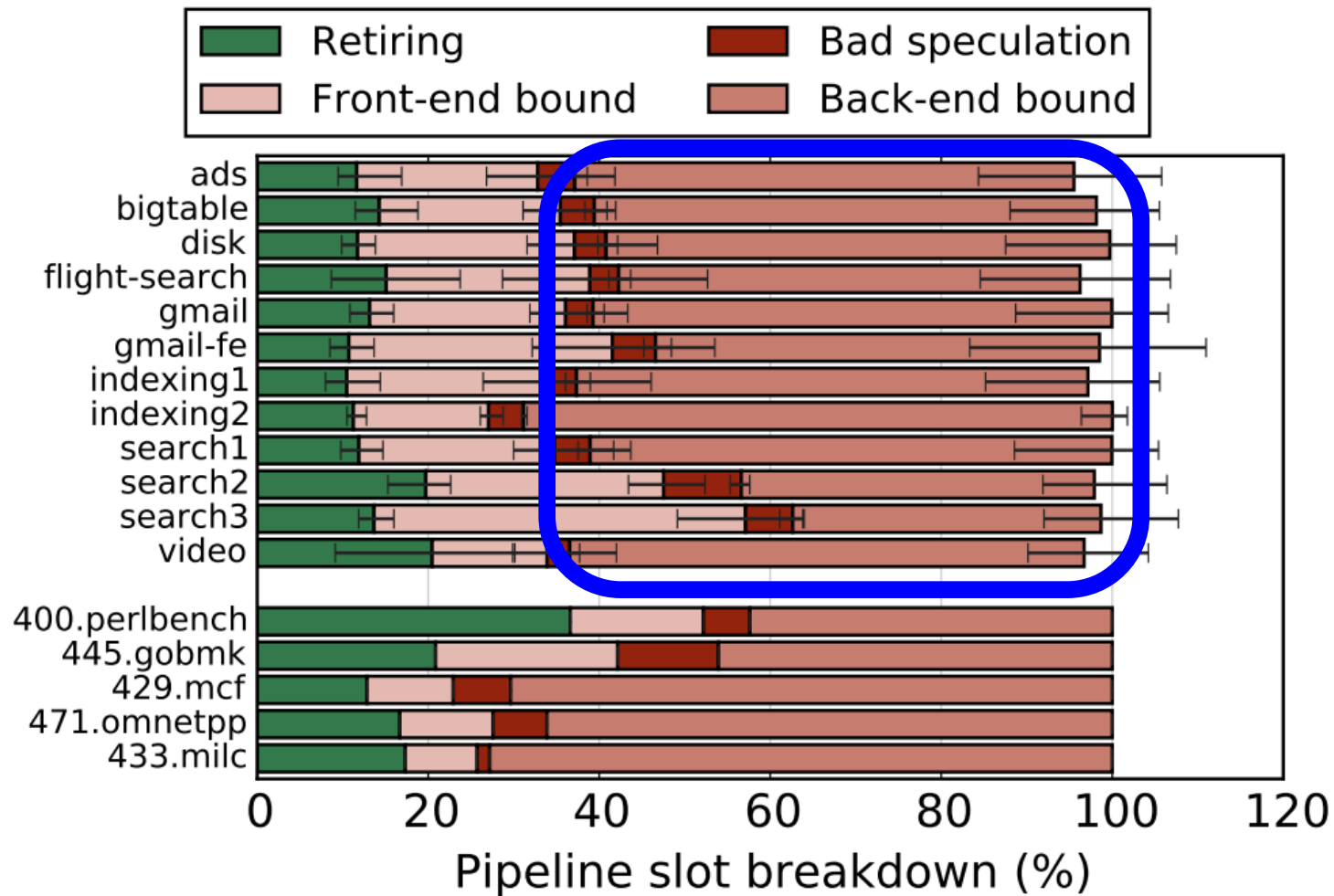
'We're beyond computation; we know how to do computation really well, we can optimize it, we can build all sorts of accelerators ... but the memory – how to feed the data, how to get the data into the accelerators – is a huge problem.'

This was how ETH Zürich and Carnegie Mellon Professor Onur Mutlu opened his course on memory systems and memory-centric computing systems at HiPEAC's summer school, ACACES18. A prolific publisher – he recently bagged the top spot on the International Symposium on Computer Architecture (ISCA) hall of fame – Onur is passionate about computation and communication that are efficient and secure by design. In advance of our Computing Systems Week focusing on data centres, storage, and networking, which takes place next week in Heraklion, HiPEAC picked his brains on all things data-based.



The Performance Perspective (Today)

- All of Google's Data Center Workloads (2015):



The Performance Perspective (Today)

- All of Google's Data Center Workloads (2015):

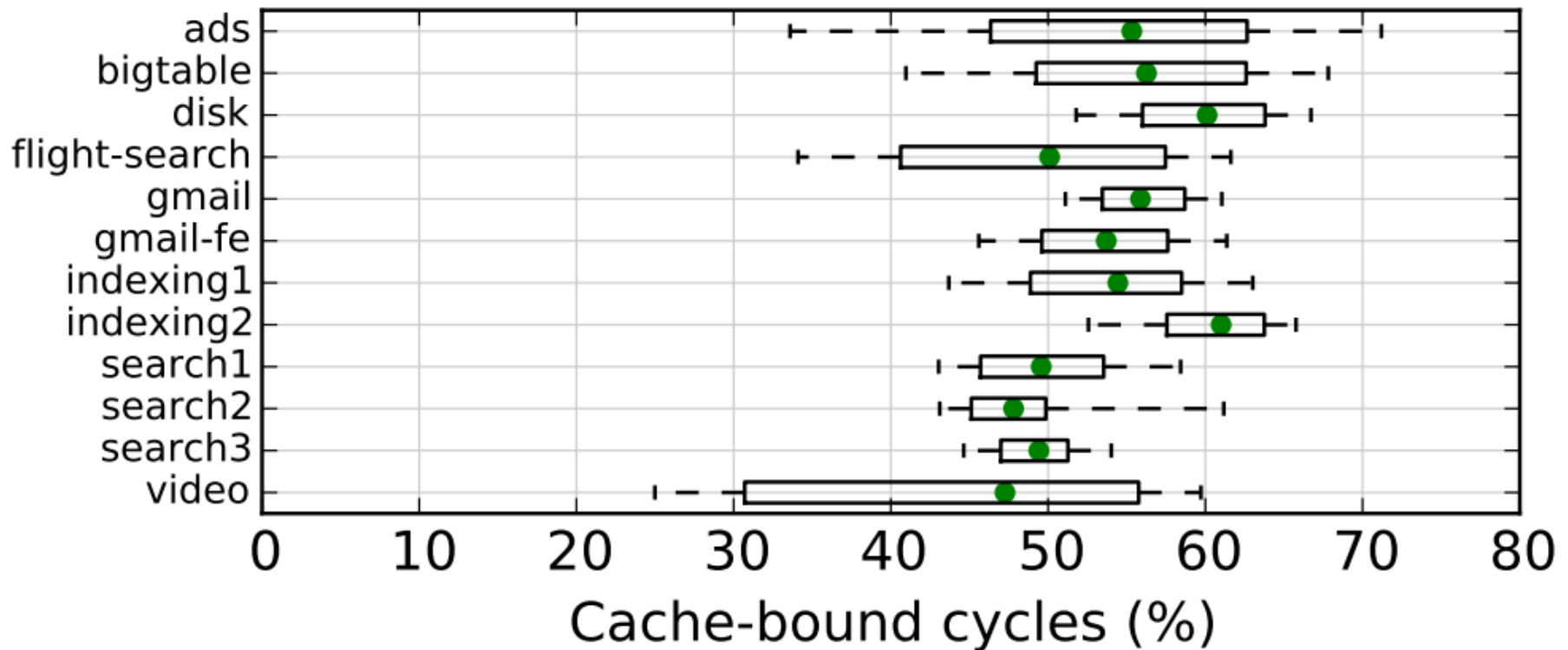


Figure 11: Half of cycles are spent stalled on caches.

Perils of Processor-Centric Design

■ Grossly-imbalanced systems

- ❑ Processing done only in **one place**
- ❑ Everything else just stores and moves data: **data moves a lot**
 - Energy inefficient
 - Low performance
 - Complex

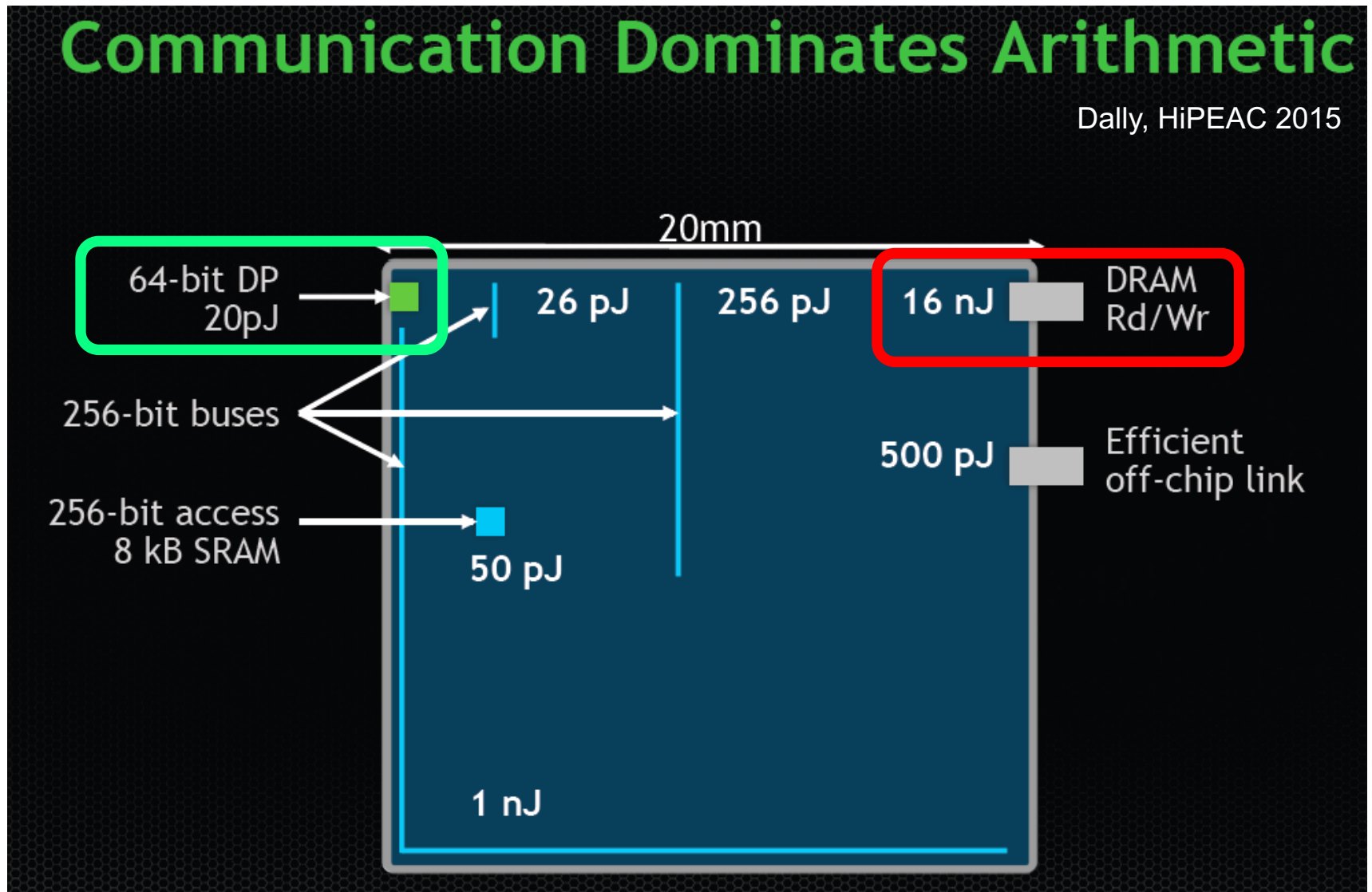
■ Overly complex and bloated processor (and accelerators)

- ❑ To tolerate data access from memory
- ❑ Complex hierarchies and mechanisms
 - Energy inefficient
 - Low performance
 - Complex

The Energy Perspective

Communication Dominates Arithmetic

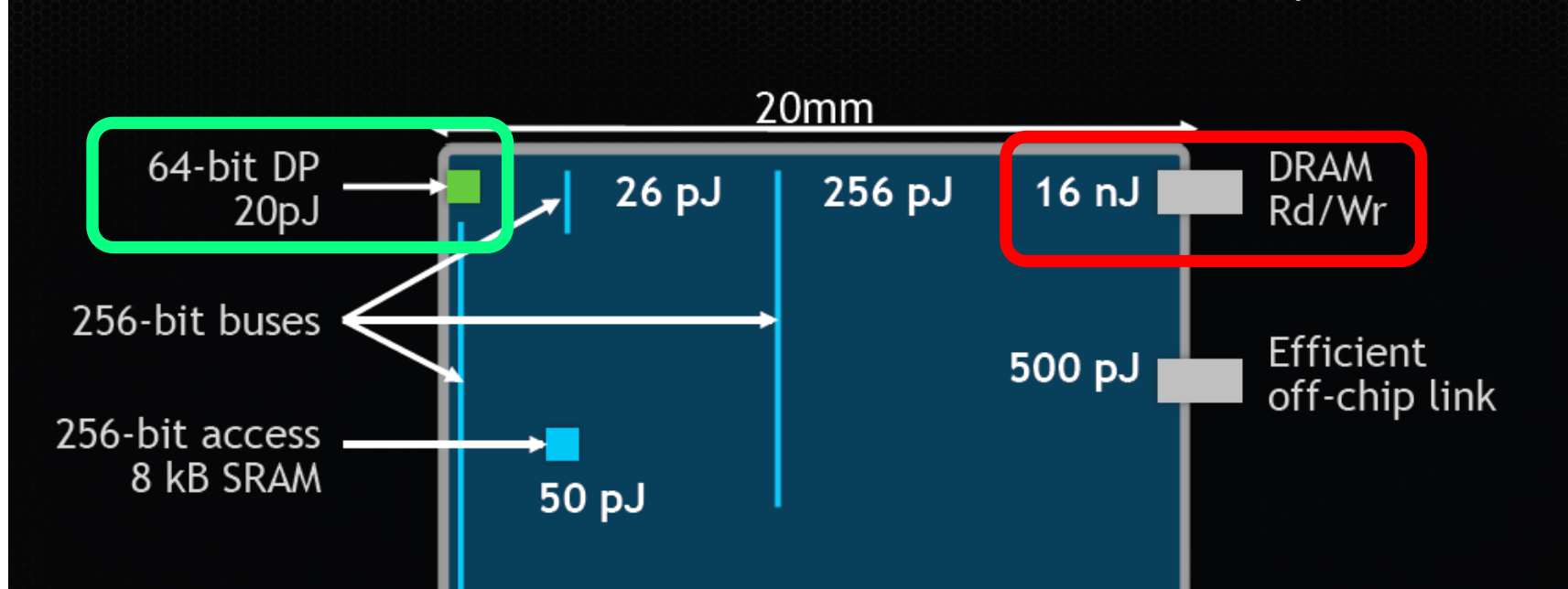
Dally, HiPEAC 2015



Data Movement vs. Computation Energy

Communication Dominates Arithmetic

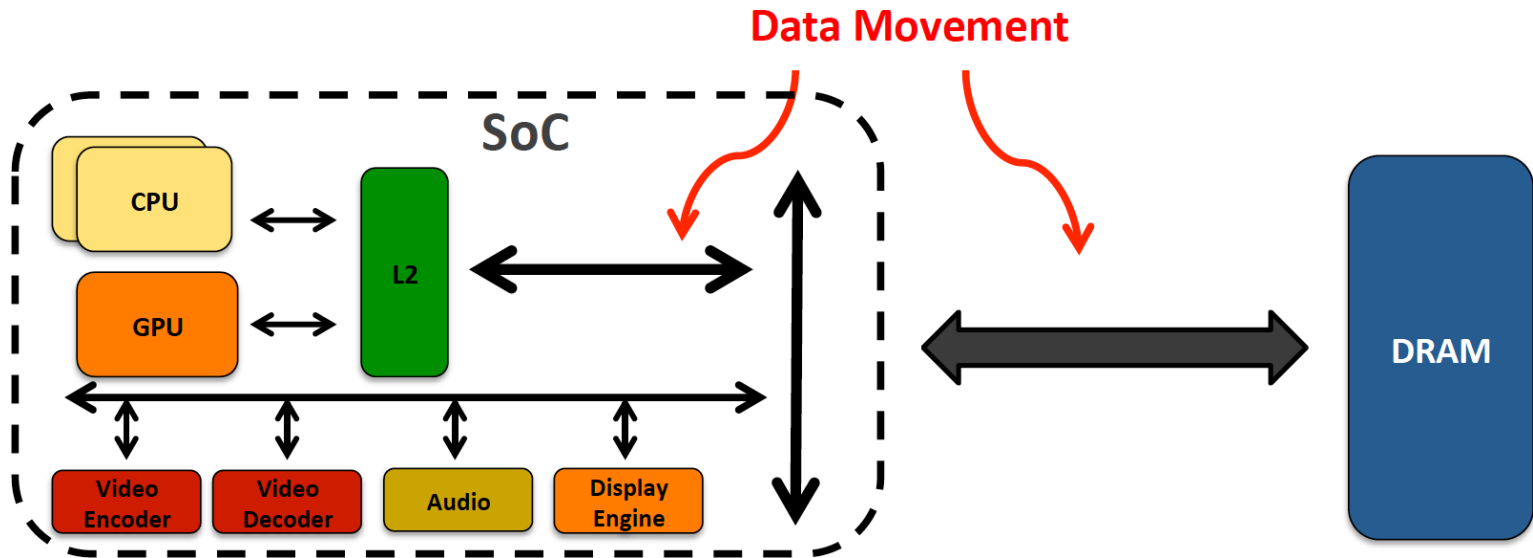
Dally, HiPEAC 2015



A memory access consumes $\sim 100\text{-}1000\times$ the energy of a complex addition

Data Movement vs. Computation Energy

- **Data movement** is a major system energy bottleneck
 - ❑ Comprises 41% of mobile system energy during web browsing [2]
 - ❑ Costs ~ 115 times as much energy as an ADD operation [1, 2]



[1]: Reducing data Movement Energy via Online Data Clustering and Encoding (MICRO'16)

[2]: Quantifying the energy cost of data movement for emerging smart phone workloads on mobile platforms (IISWC'14)

Energy Waste in Mobile Devices

- Amirali Boroumand, Saugata Ghose, Youngsok Kim, Rachata Ausavarungnirun, Eric Shiu, Rahul Thakur, Daehyun Kim, Aki Kuusela, Allan Knies, Parthasarathy Ranganathan, and Onur Mutlu, **"Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks"** *Proceedings of the 23rd International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, Williamsburg, VA, USA, March 2018.

**62.7% of the total system energy
is spent on data movement**

Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks

Amirali Boroumand¹

Saugata Ghose¹

Youngsok Kim²

Rachata Ausavarungnirun¹

Eric Shiu³

Rahul Thakur³

Daehyun Kim^{4,3}

Aki Kuusela³

Allan Knies³

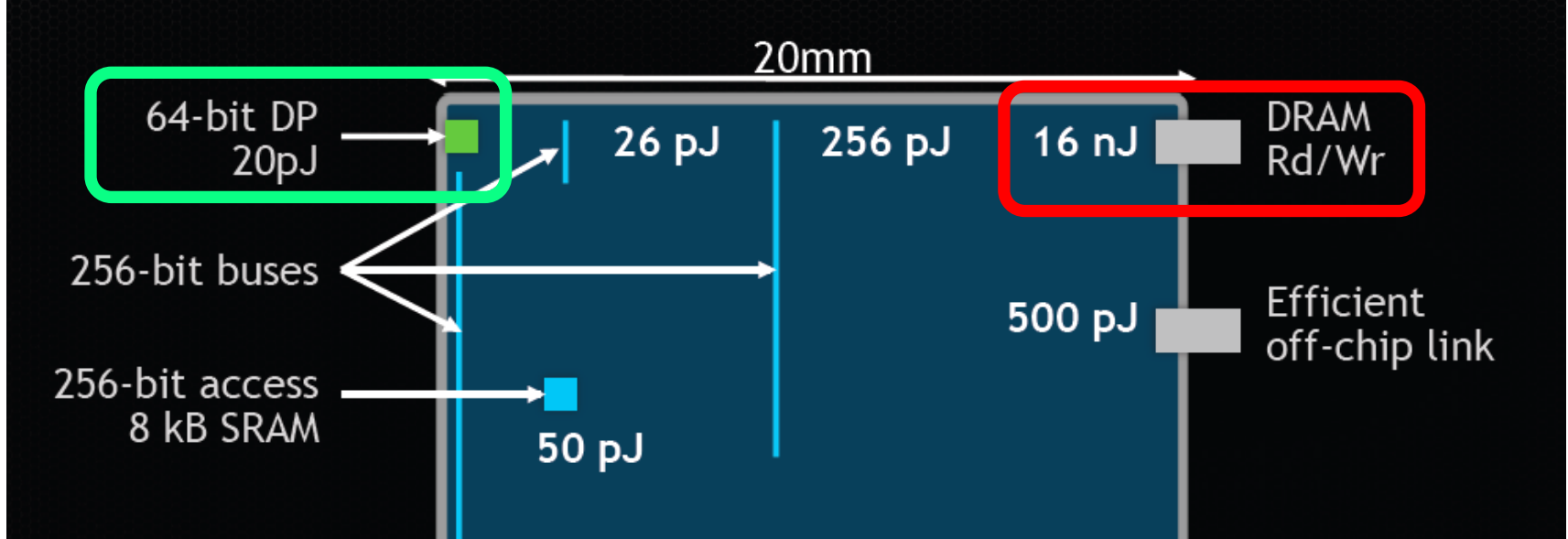
Parthasarathy Ranganathan³

Onur Mutlu^{5,1}

We Do Not Want to Move Data!

Communication Dominates Arithmetic

Dally, HiPEAC 2015

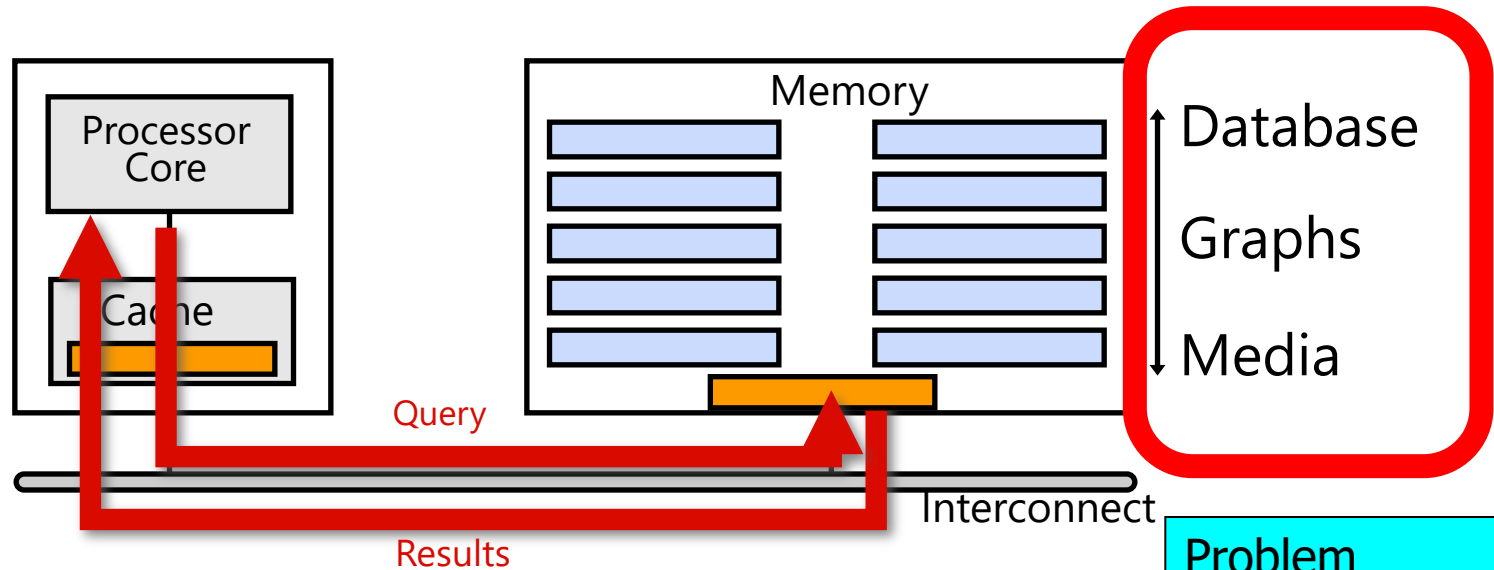


A memory access consumes $\sim 100\text{-}1000\times$ the energy of a complex addition

We Need A Paradigm Shift To ...

- Enable computation with minimal data movement
- Compute where it makes sense (where data resides)
- Make computing architectures more data-centric

Goal: Processing Inside Memory



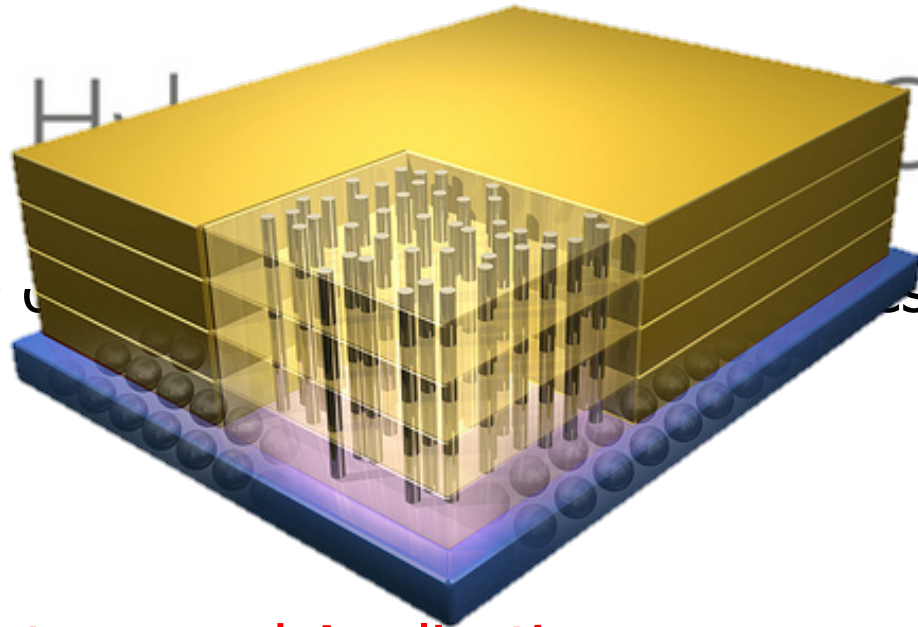
- Many questions ... How do we design the:
 - ❑ compute-capable memory & controllers?
 - ❑ processor chip and in-memory units?
 - ❑ software and hardware interfaces?
 - ❑ system software, compilers, languages?
 - ❑ algorithms and theoretical foundations?

Problem
Algorithm
Program/Language
System Software
SW/HW Interface
Micro-architecture
Logic
Devices
Electrons

Why In-Memory Computation Today?



→ Industry C

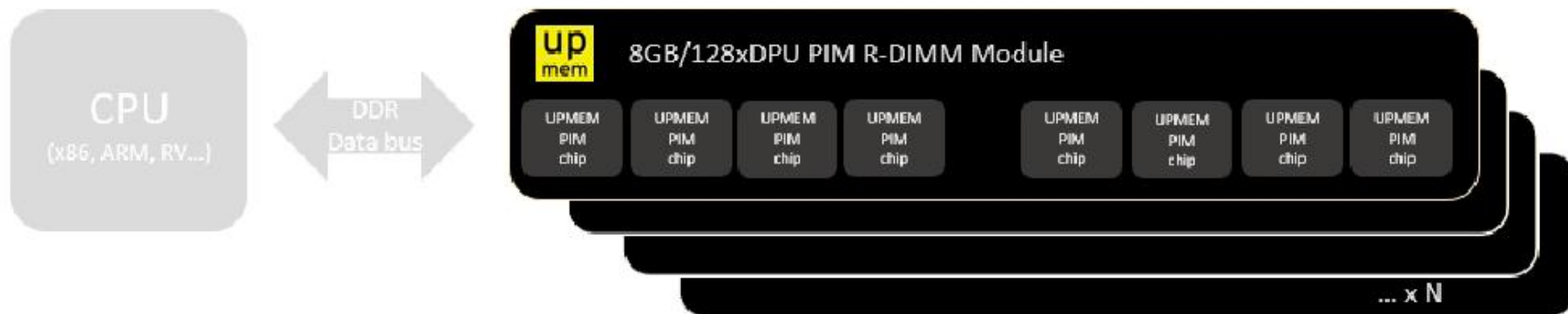
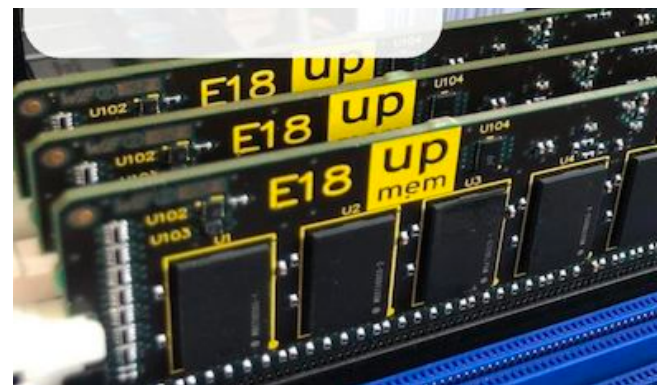


■ Pull from Systems and Applications

- ❑ Data access is a major system and application bottleneck
- ❑ Systems are energy limited
- ❑ Data movement much more energy-hungry than computation

UPMEM Processing-in-DRAM Engine (2019)

- **Processing in DRAM Engine**
- Includes **standard DIMM modules**, with a **large number of DPU processors** combined with DRAM chips.
- Replaces **standard DIMMs**
 - DDR4 R-DIMM modules
 - 8GB+128 DPUs (16 PIM chips)
 - Standard 2x-nm DRAM process
 - **Large amounts of** compute & memory bandwidth



We Need to Think Differently
from the Past Approaches

Sub-Agenda: In-Memory Computation

- Major Trends Affecting Main Memory
- The Need for Intelligent Memory Controllers
 - Bottom Up: Push from Circuits and Devices
 - Top Down: Pull from Systems and Applications
- Processing in Memory: Two Directions
 - Minimally Changing Memory Chips
 - Exploiting 3D-Stacked Memory
- How to Enable Adoption of Processing in Memory
- Conclusion

Processing in Memory: Two Approaches

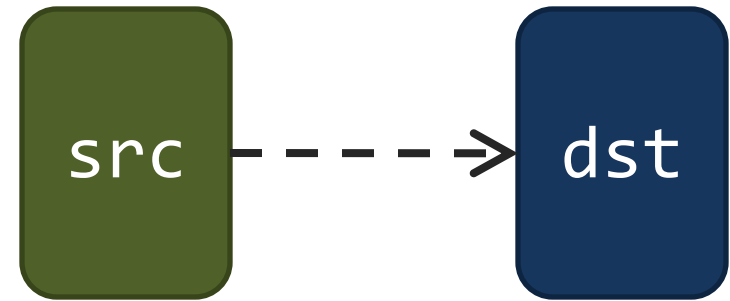
1. Minimally changing memory chips
2. Exploiting 3D-stacked memory

Approach 1: Minimally Changing DRAM

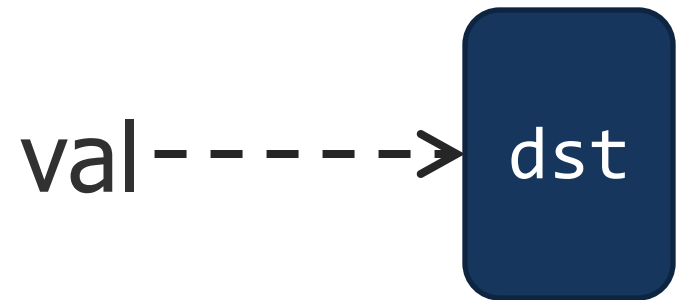
- DRAM has great capability to perform **bulk data movement and computation** internally with small changes
 - Can **exploit internal connectivity** to move data
 - Can **exploit analog computation capability**
 - ...
- Examples: RowClone, In-DRAM AND/OR, Gather/Scatter DRAM
 - RowClone: Fast and Efficient In-DRAM Copy and Initialization of Bulk Data (Seshadri et al., MICRO 2013)
 - Fast Bulk Bitwise AND and OR in DRAM (Seshadri et al., IEEE CAL 2015)
 - Gather-Scatter DRAM: In-DRAM Address Translation to Improve the Spatial Locality of Non-unit Strided Accesses (Seshadri et al., MICRO 2015)
 - "Ambit: In-Memory Accelerator for Bulk Bitwise Operations Using Commodity DRAM Technology" (Seshadri et al., MICRO 2017)

Starting Simple: Data Copy and Initialization

**Bulk Data
Copy**



**Bulk Data
Initialization**



Bulk Data Copy and Initialization

The Impact of Architectural Trends on Operating System Performance

Mendel Rosenblum, Edouard Bugnion, Stephen Alan Herrod,
Emmett Witchel, and Anoop Gupta

Hardware Support for Bulk Data Movement in Server Platforms

Li Zhao[†], Ravi Iyer[‡], Srihari Makineni[‡], Laxmi Bhuyan[†] and Don Newell[‡]

[†]Department of Computer Science and Engineering, University of California, Riverside, CA 92521
Email: {zhao, bhuyan}@cs.ucr.edu

[‡]Communications Technology Lab, Intel Corp.

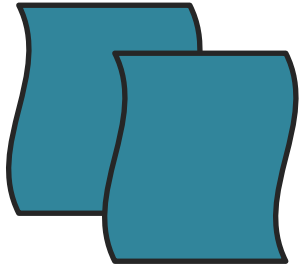
Architecture Support for Improving Bulk Memory Copying and Initialization Performance

Xiaowei Jiang, Yan Solihin
Dept. of Electrical and Computer Engineering
North Carolina State University
Raleigh, USA

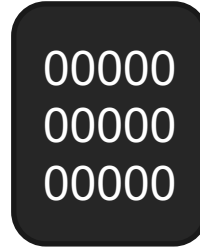
Li Zhao, Ravishankar Iyer
Intel Labs
Intel Corporation
Hillsboro, USA

Starting Simple: Data Copy and Initialization

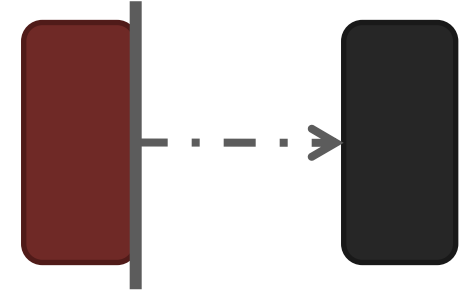
memmove & memcpy: 5% cycles in Google's datacenter [Kanev+ ISCA'15]



Forking



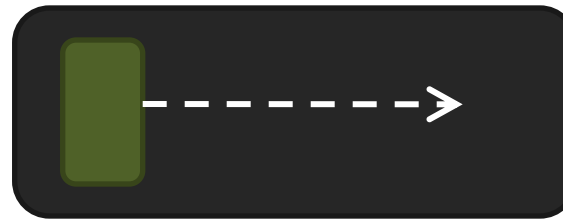
**Zero initialization
(e.g., security)**



Checkpointing



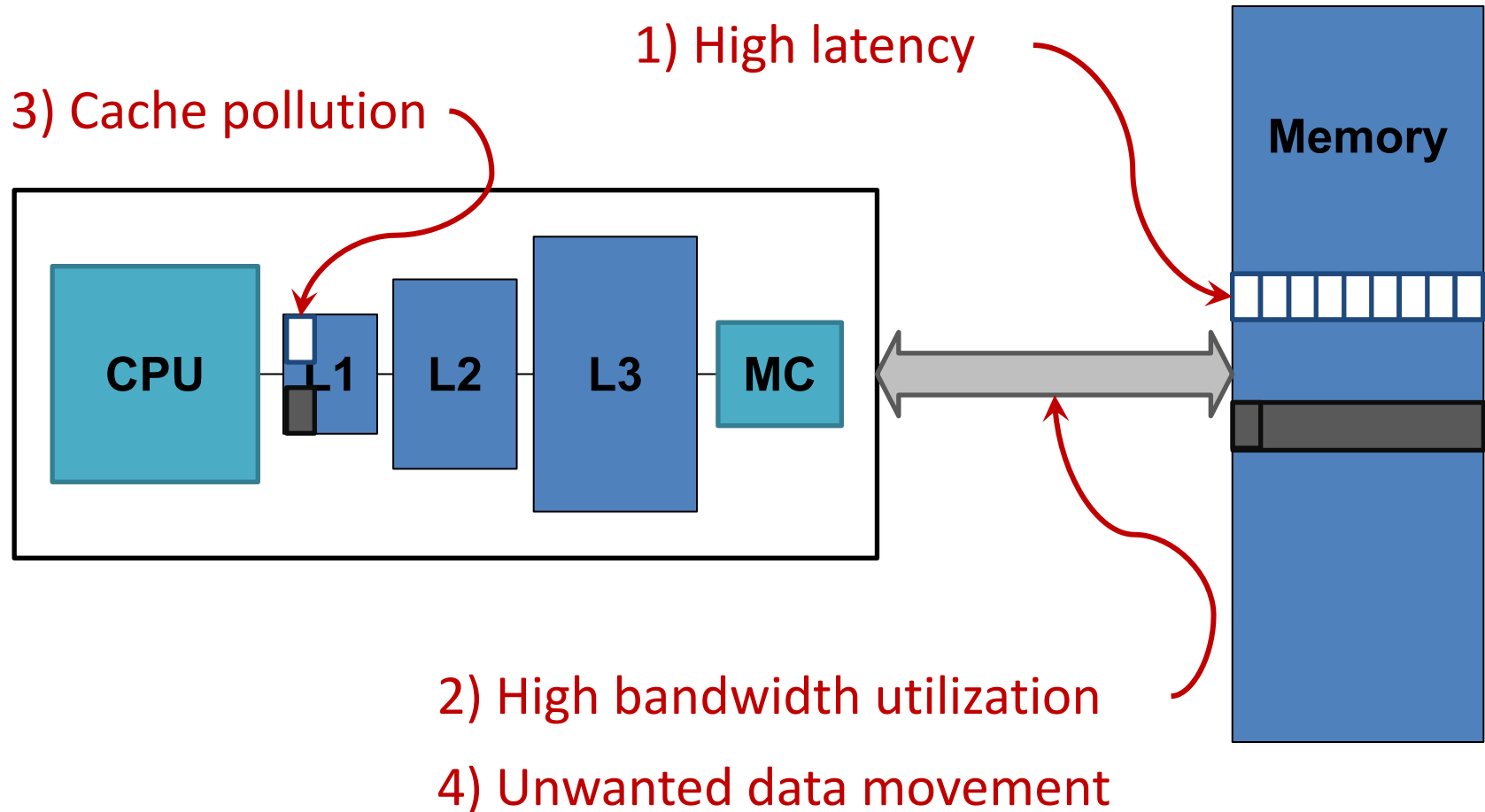
**VM Cloning
Deduplication**



Page Migration

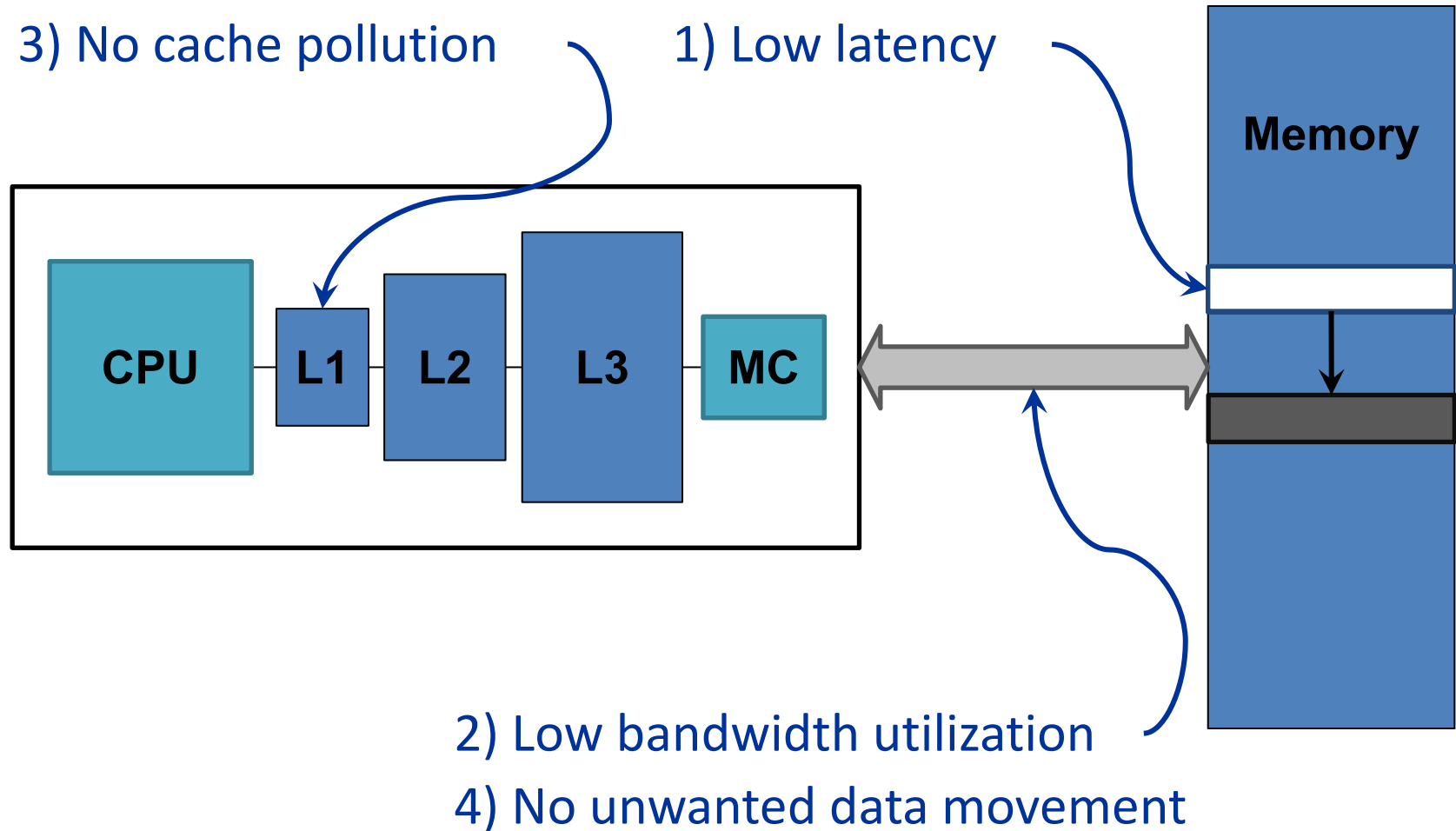
...
Many more

Today's Systems: Bulk Data Copy



1046ns, 3.6uJ (for 4KB page copy via DMA)

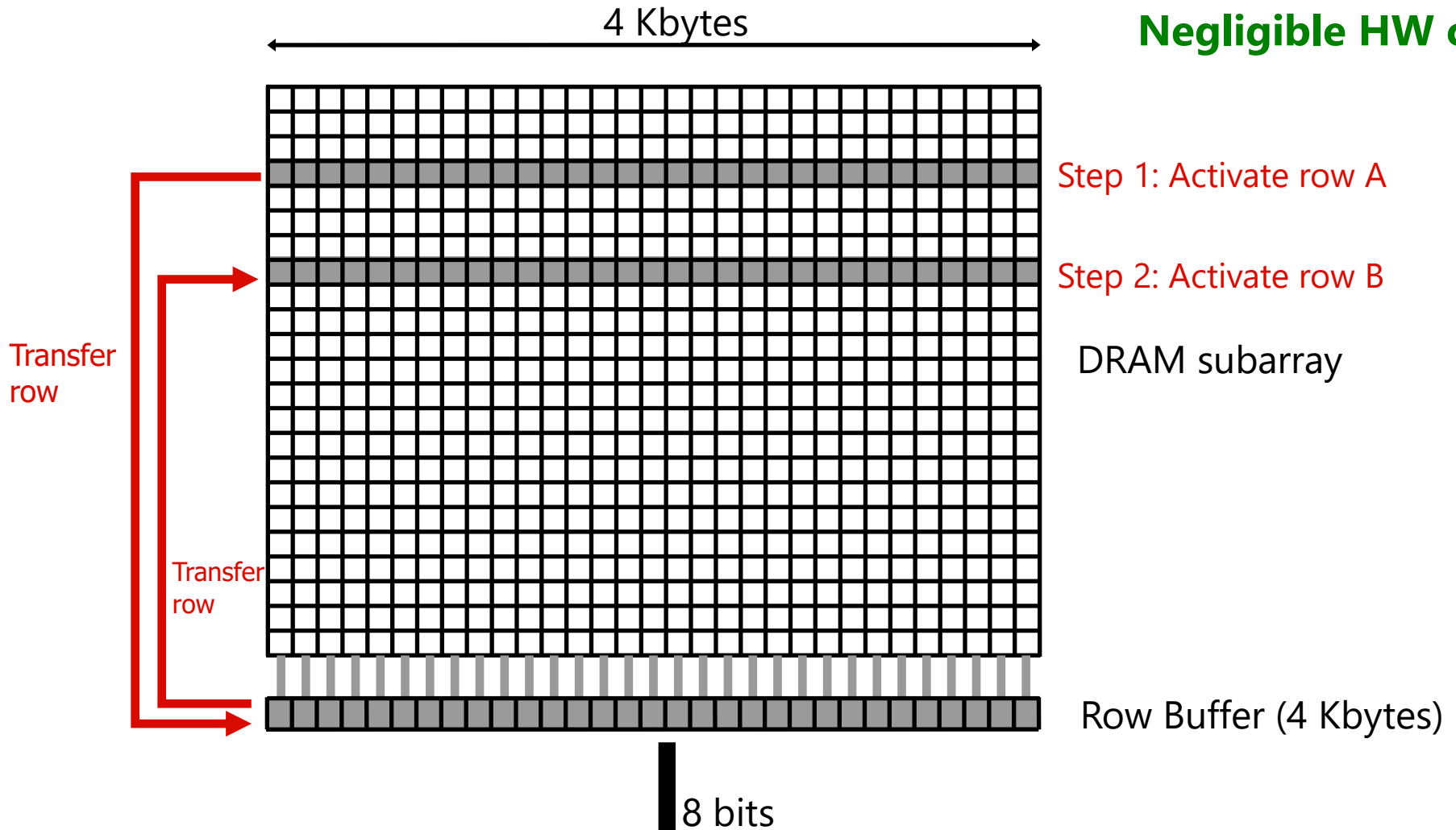
Future Systems: In-Memory Copy



1046ns, 3.6uJ → 90ns, 0.04uJ

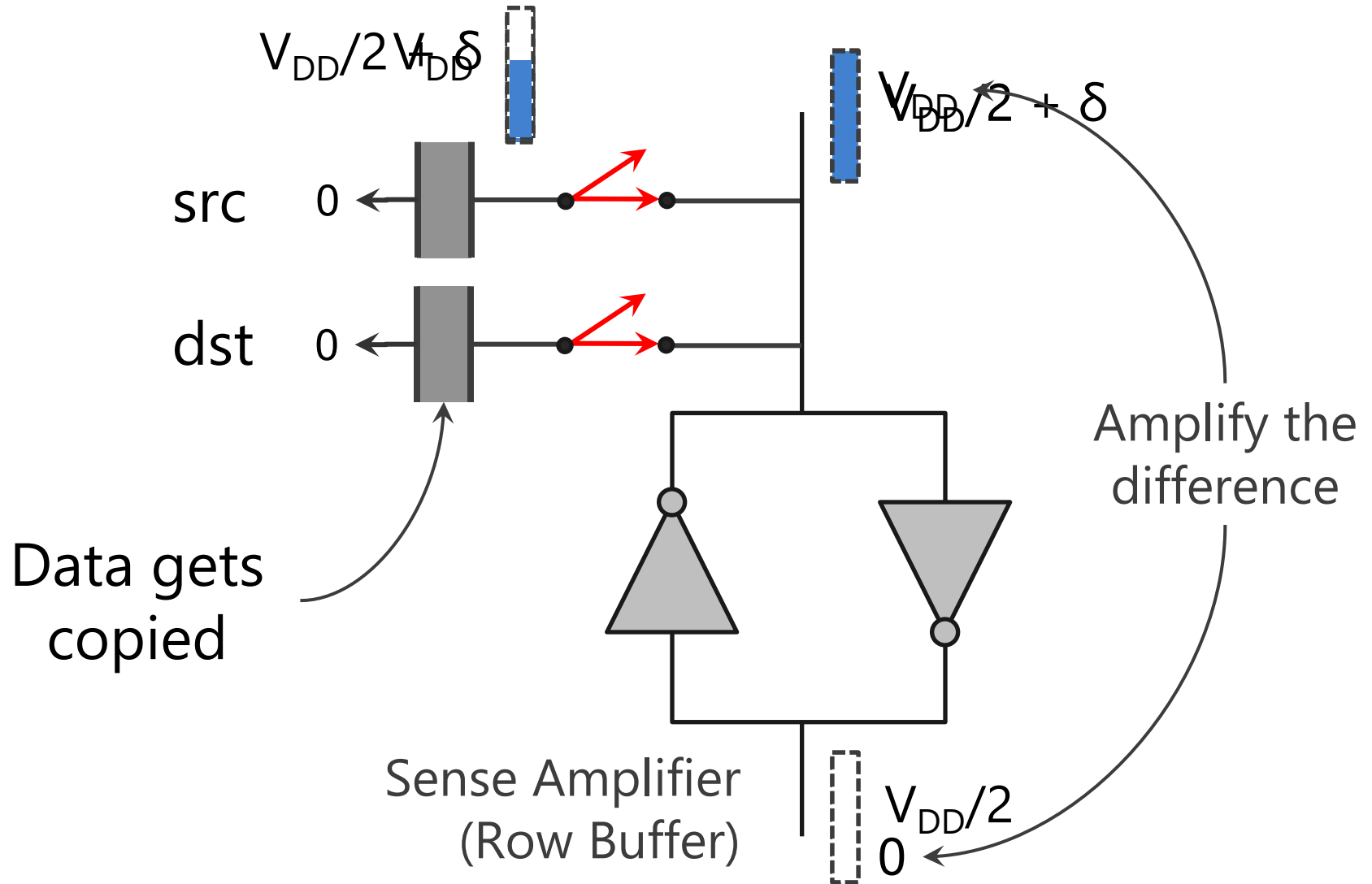
RowClone: In-DRAM Row Copy

Idea: Two consecutive ACTivates
Negligible HW cost

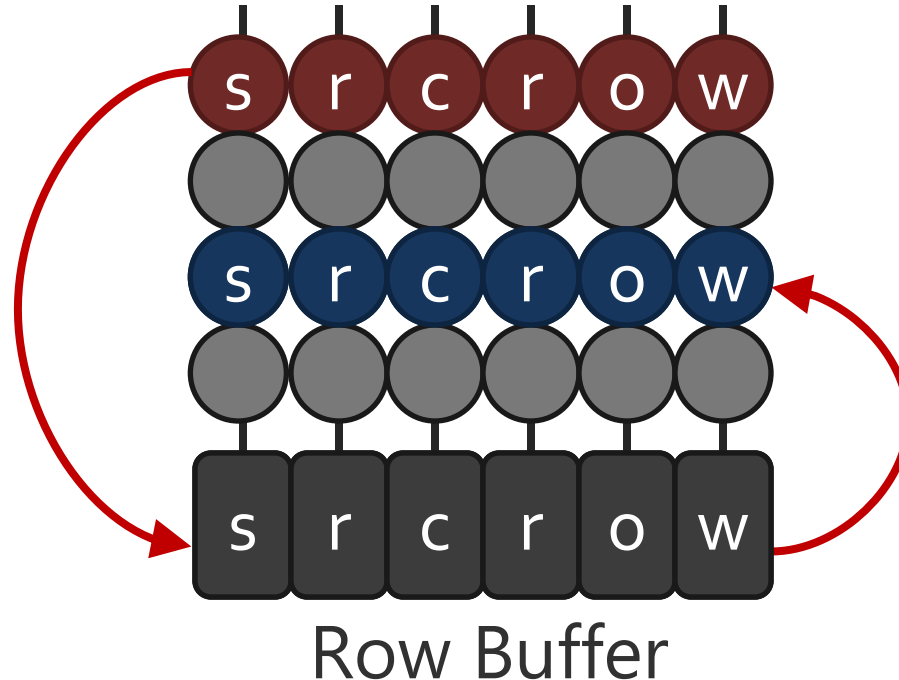


11.6X latency reduction, **74X** energy reduction

RowClone: Intra-Subarray

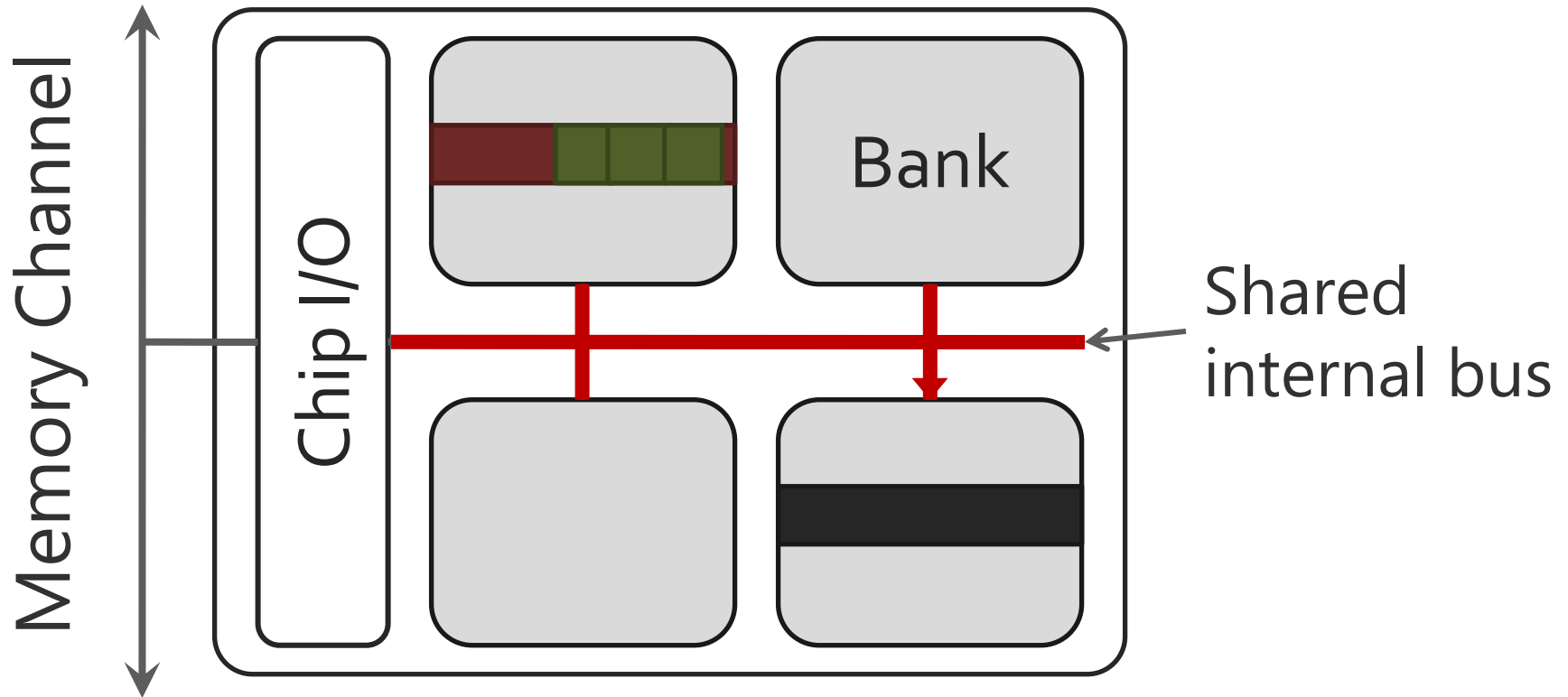


RowClone: Intra-Subarray (II)



1. **Activate** src row (copy data from src to row buffer)
2. **Activate** dst row (disconnect src from row buffer, connect dst – copy data from row buffer to dst)

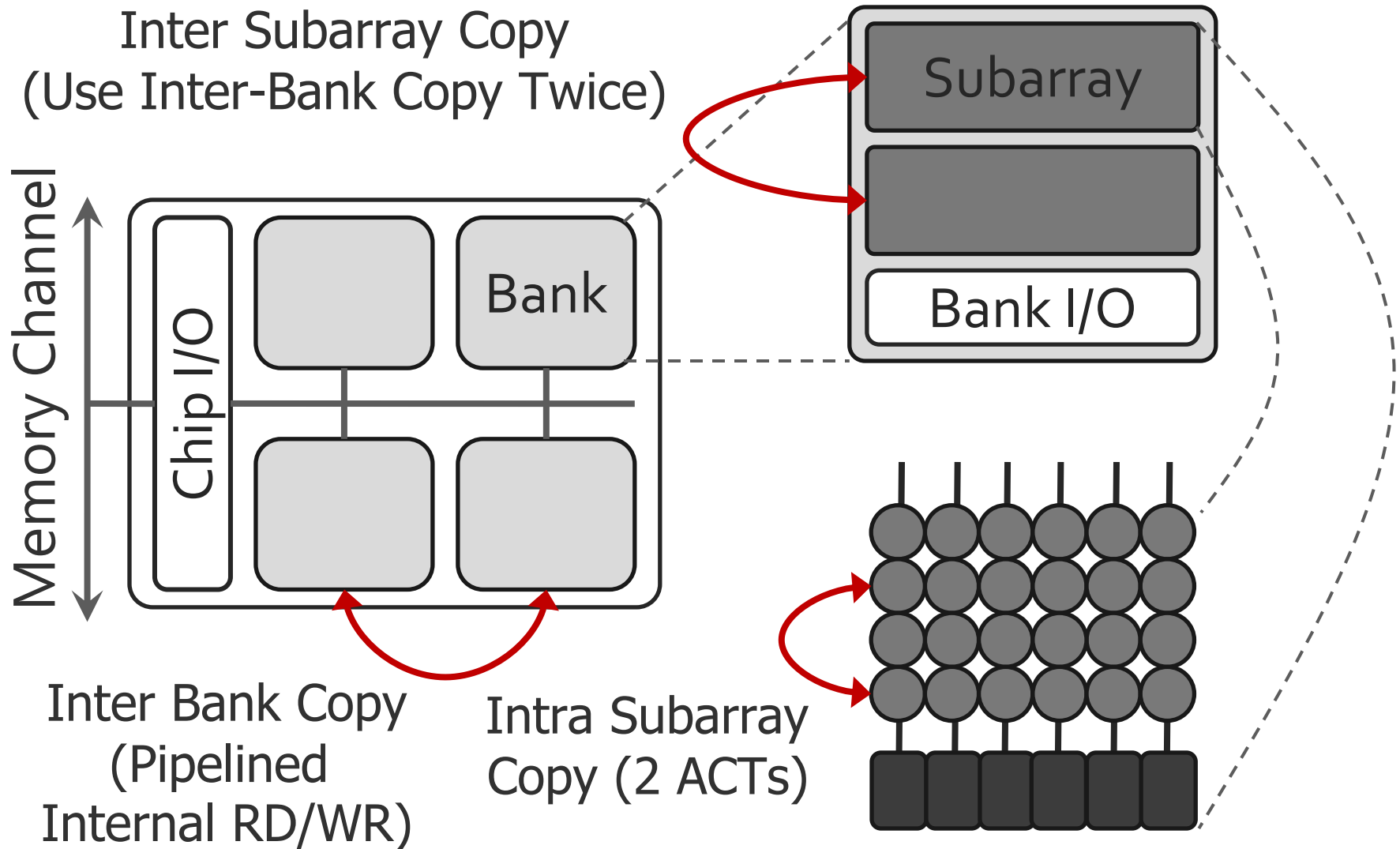
RowClone: Inter-Bank



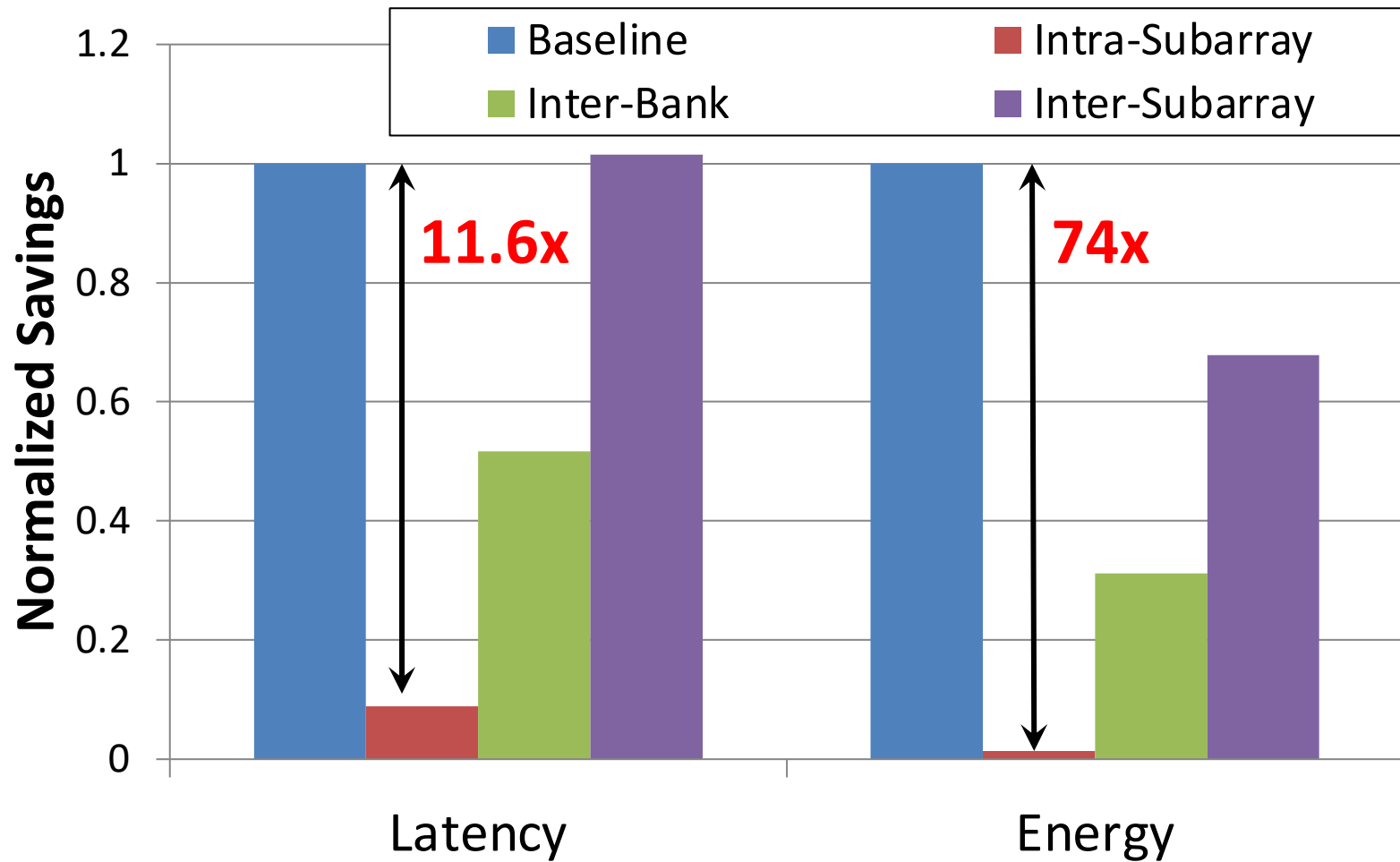
Overlap the latency of the read and the write
1.9X latency reduction, **3.2X** energy reduction

Generalized RowClone

0.01% area cost

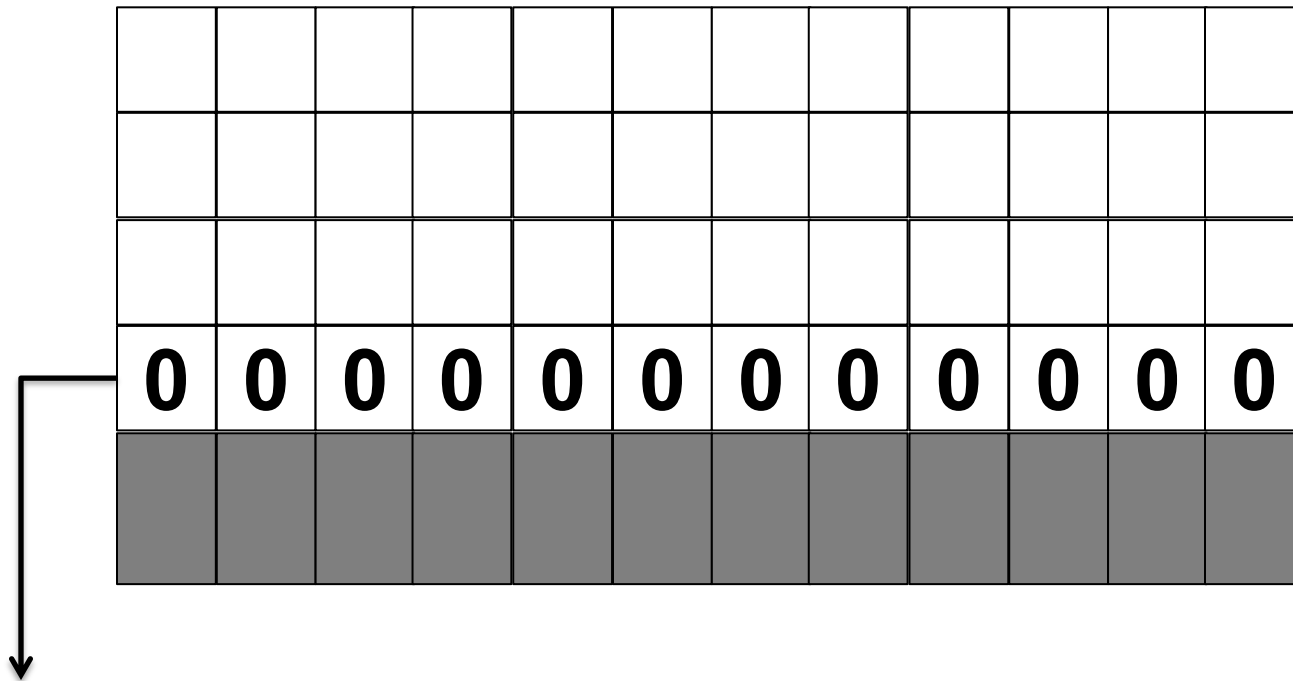


RowClone: Latency and Energy Savings



Seshadri et al., "RowClone: Fast and Efficient In-DRAM Copy and Initialization of Bulk Data," MICRO 2013.

RowClone: Fast Row Initialization



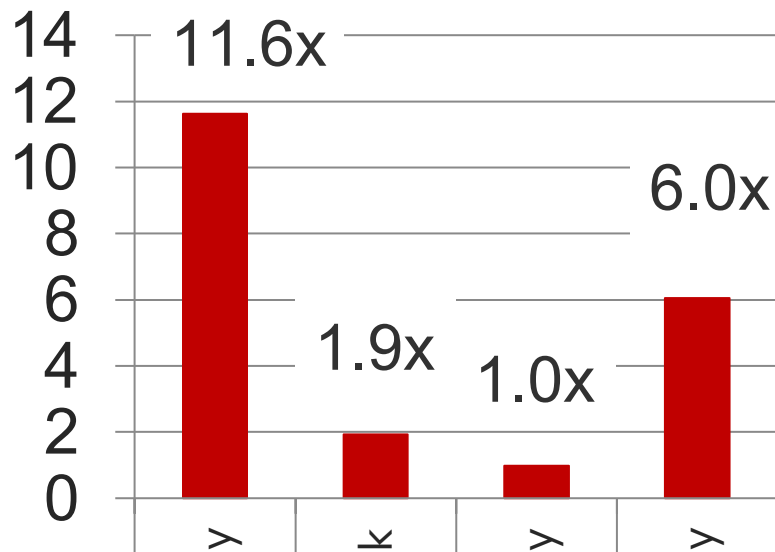
Fix a row at Zero
(0.5% loss in capacity)

RowClone: Bulk Initialization

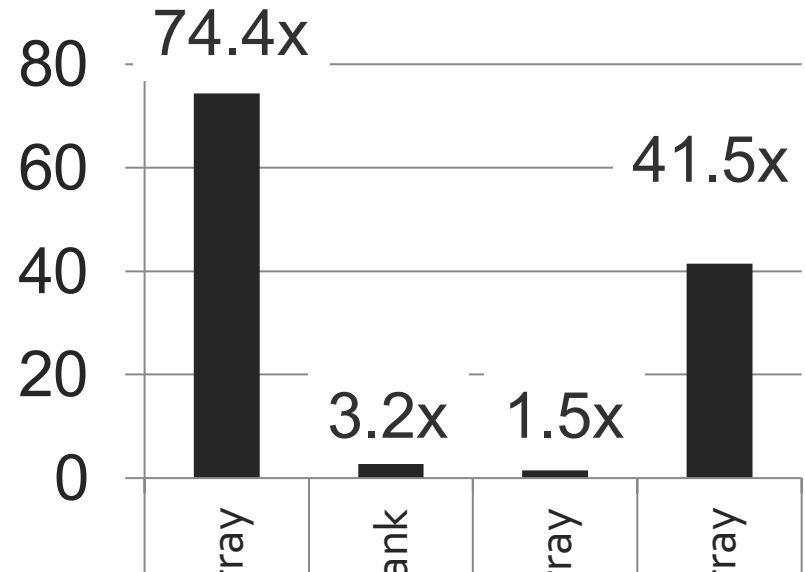
- Initialization with arbitrary data
 - Initialize one row
 - Copy the data to other rows
- Zero initialization (most common)
 - Reserve a row in each subarray (always zero)
 - Copy data from reserved row (FPM mode)
 - **6.0X** lower latency, **41.5X** lower DRAM energy
 - 0.2% loss in capacity

RowClone: Latency & Energy Benefits

Latency Reduction

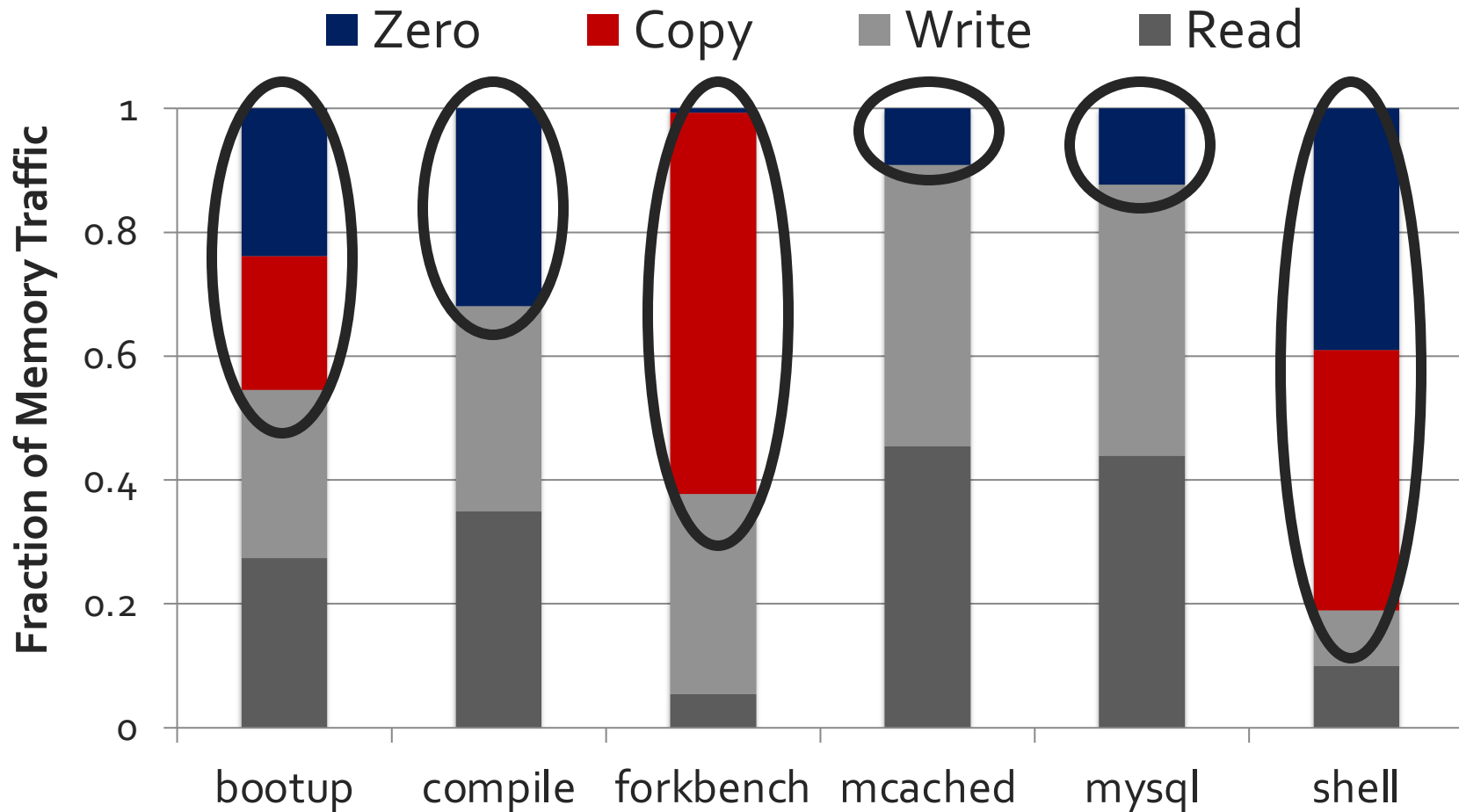


Energy Reduction

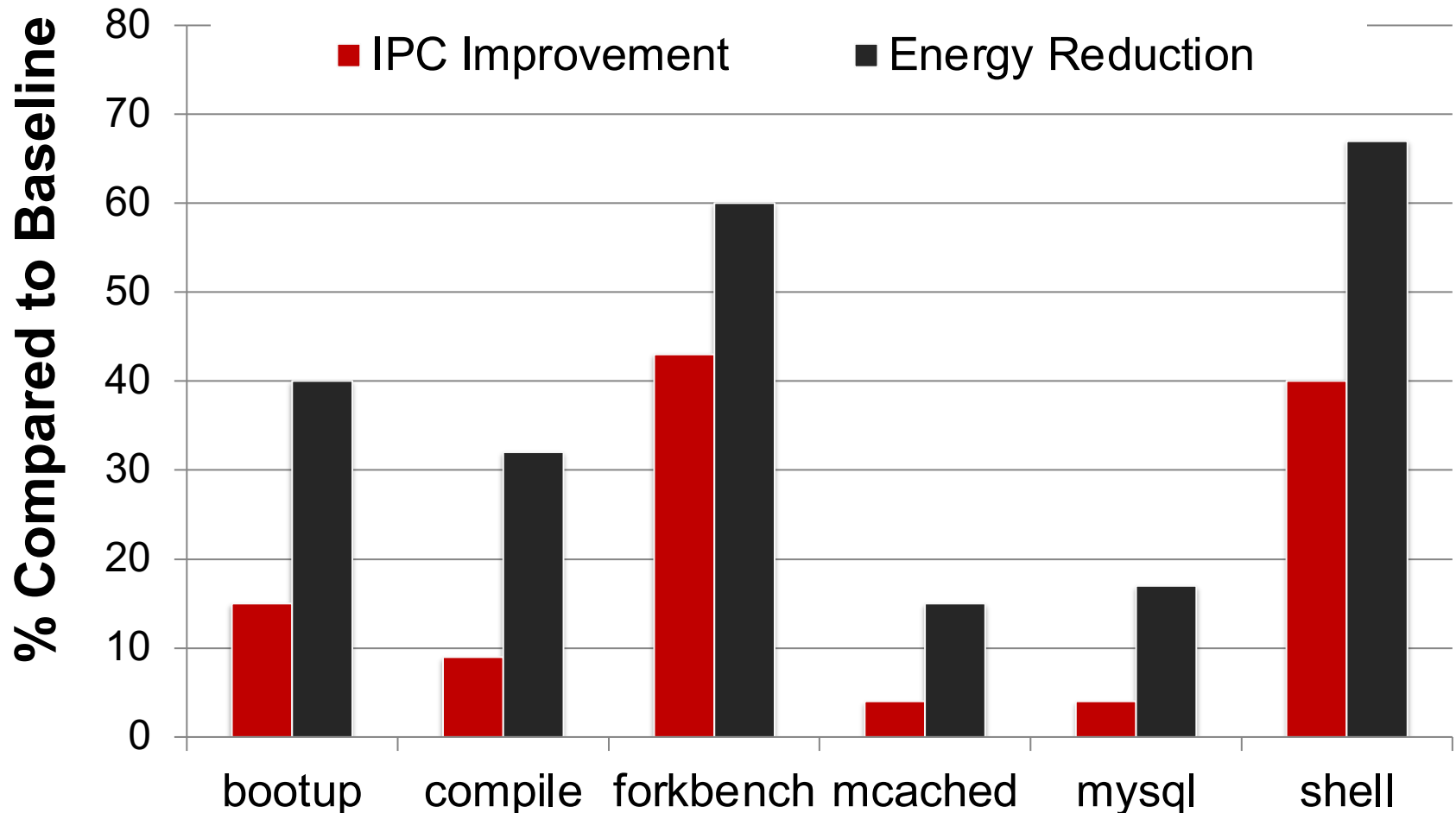


Very low cost: 0.01% increase in die area

Copy and Initialization in Workloads



RowClone: Application Performance



End-to-End System Design

Application

How to communicate occurrences of bulk copy/initialization across layers?

Operating System

How to ensure cache coherence?

ISA

Microarchitecture

How to maximize latency and energy savings?

DRAM (RowClone)

How to handle data reuse?

More on RowClone

- Vivek Seshadri, Yoongu Kim, Chris Fallin, Donghyuk Lee, Rachata Ausavarungnirun, Gennady Pekhimenko, Yixin Luo, Onur Mutlu, Michael A. Kozuch, Phillip B. Gibbons, and Todd C. Mowry,
"RowClone: Fast and Energy-Efficient In-DRAM Bulk Data Copy and Initialization"
Proceedings of the 46th International Symposium on Microarchitecture (MICRO), Davis, CA, December 2013. [[Slides \(pptx\)](#)] [[pdf](#)] [[Lightning Session Slides \(pptx\)](#)] [[pdf](#)] [[Poster \(pptx\)](#)] [[pdf](#)]

RowClone: Fast and Energy-Efficient In-DRAM Bulk Data Copy and Initialization

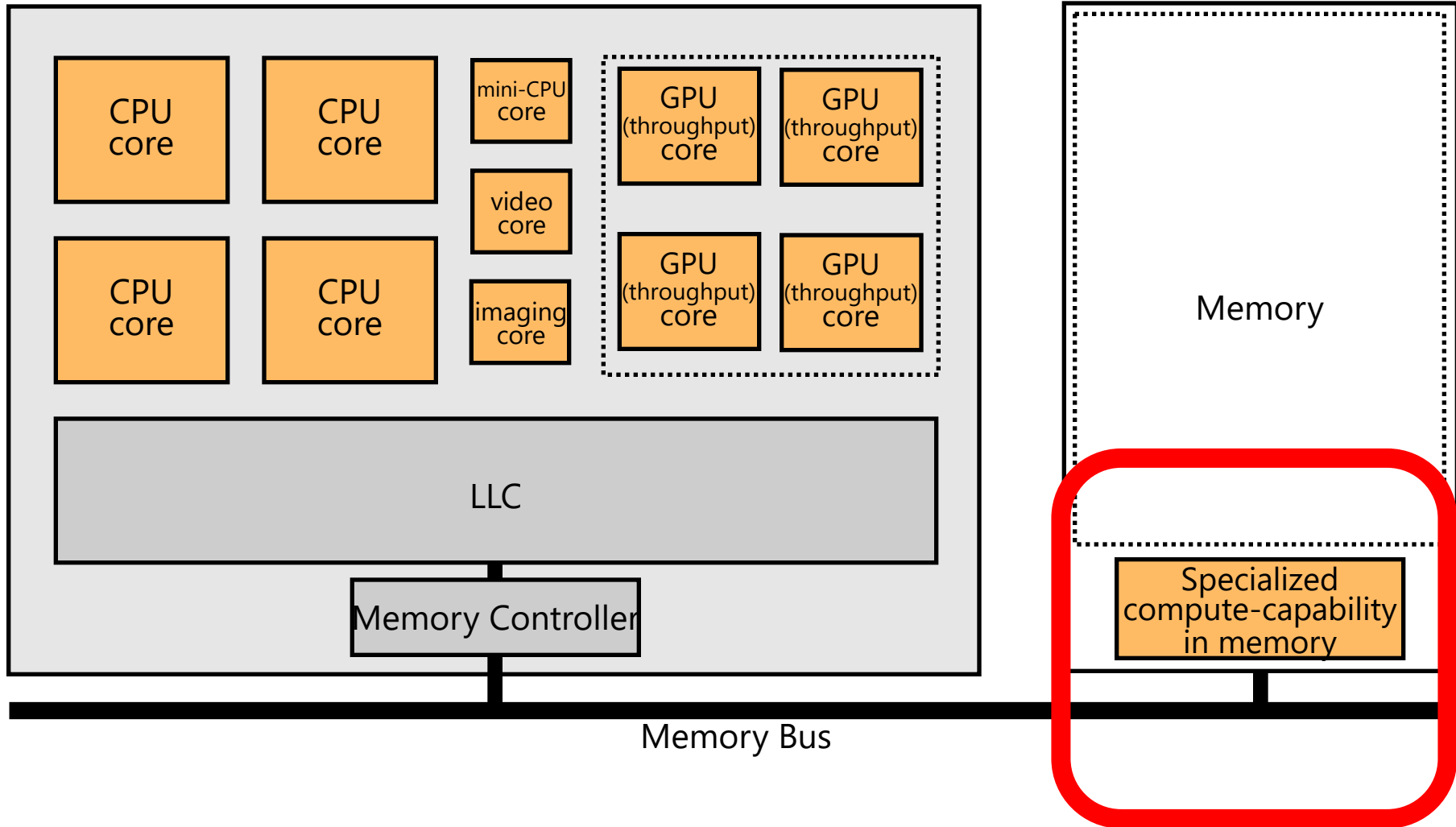
Vivek Seshadri vseshadr@cs.cmu.edu	Yoongu Kim yoongukim@cmu.edu	Chris Fallin* cfallin@c1f.net	Donghyuk Lee donghyuk1@cmu.edu
---------------------------------------	---------------------------------	----------------------------------	-----------------------------------

Rachata Ausavarungnirun rachata@cmu.edu	Gennady Pekhimenko gpekhime@cs.cmu.edu	Yixin Luo yixinluo@andrew.cmu.edu
--	---	--------------------------------------

Onur Mutlu onur@cmu.edu	Phillip B. Gibbons† phillip.b.gibbons@intel.com	Michael A. Kozuch† michael.a.kozuch@intel.com	Todd C. Mowry tcm@cs.cmu.edu
----------------------------	--	--	---------------------------------

Carnegie Mellon University †Intel Pittsburgh

Memory as an Accelerator



Memory similar to a "conventional" accelerator

RowClone Strengths

Strengths of the Paper

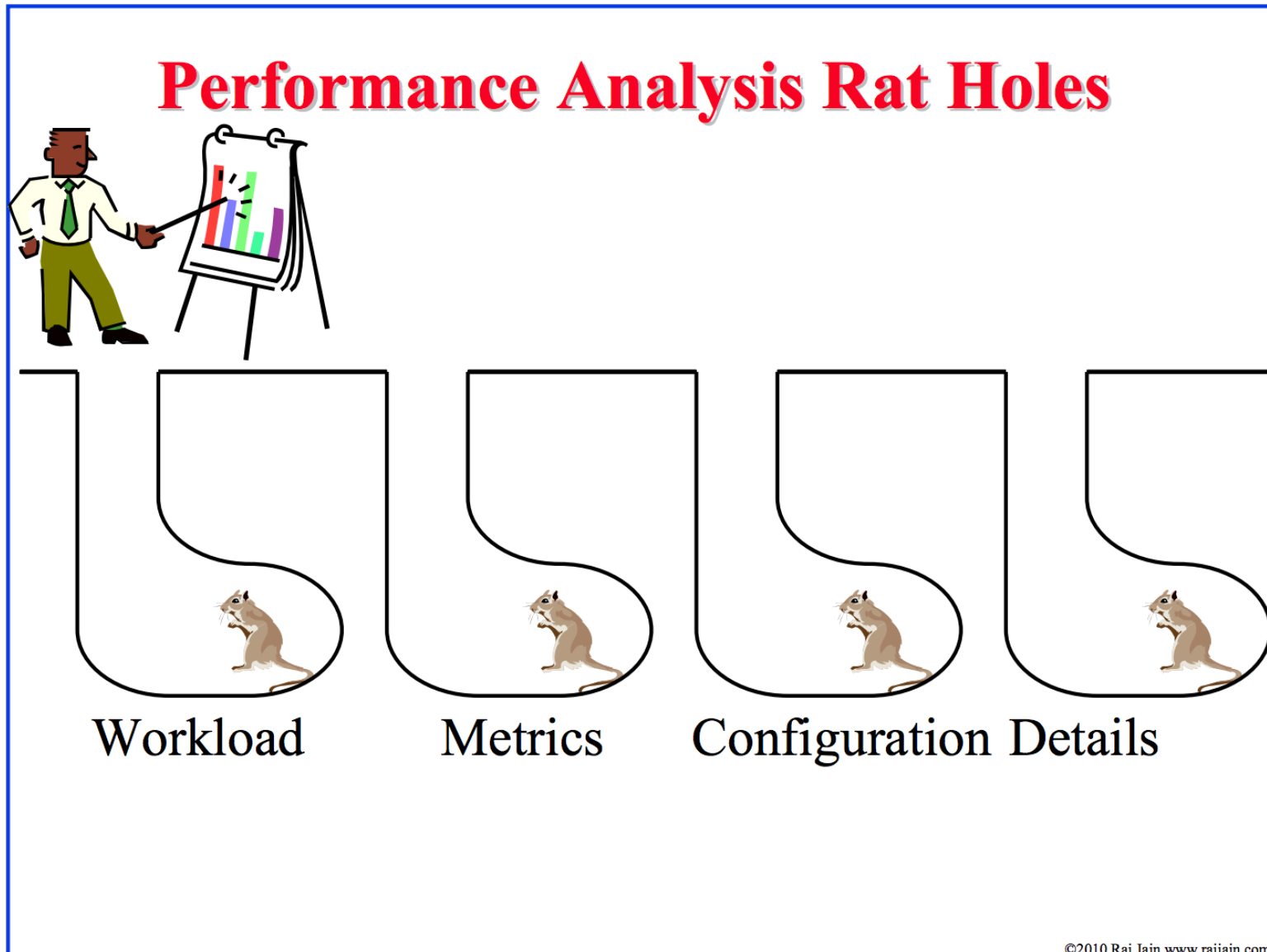
- Simple, novel mechanism to solve an important problem
 - Effective and low hardware overhead
 - Intuitive idea!
 - Greatly improves performance and efficiency (assuming data is mapped nicely)
 - Seems like a clear win for data initialization (without mapping requirements)
 - Makes software designer's life easier
 - If copies are 10x-100x cheaper, how to design software?
 - Paper tackles many low-level and system-level issues
 - Well-written, insightful paper
-

RowClone Weaknesses

Weaknesses

- Requires data to be mapped in the same subarray to deliver the largest benefits
 - Helps less if data movement is not within a subarray
 - Does not help if data movement is across DRAM channels
- Inter-subarray copy is very inefficient
- Causes many changes in the system stack
 - End-to-end design spans applications to circuits
 - Software-hardware cooperative solution might not always be easy to adopt
- Cache coherence and data reuse cause real overheads
- Evaluation is done solely in simulation
- Evaluation does not consider multi-chip systems
- Are these the best workloads to evaluate?

Recall: Try to Avoid Rat Holes



Improvements on RowClone

Extensions and Follow-Up Work

- Can this be improved to do faster inter-subarray copy?
 - Yes, [see the LISA paper \[Chang et al., HPCA 2016\]](#)
- Can we enable data movement at smaller granularities within a bank?
 - Yes, [see the FIGARO paper \[Wang et al., MICRO 2020\]](#)
- Can this be improved to do better inter-bank copy?
 - Yes, [see the Network-on-Memory paper \[CAL 2020\]](#)
- Can similar ideas and DRAM properties be used to perform computation on data?
 - Yes, [see the Ambit paper \[Seshadri et al., MICRO 2017\]](#)

LISA: Fast Inter-Subarray Data Movement

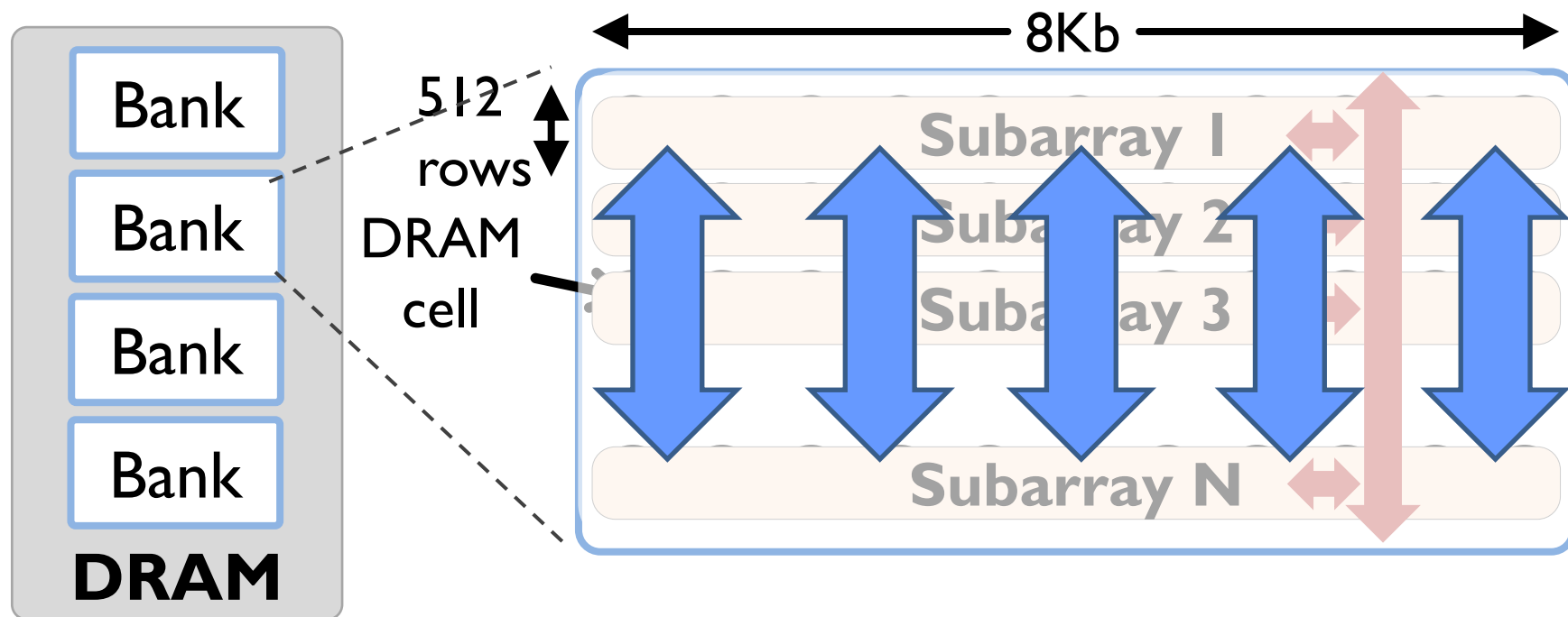
- Kevin K. Chang, Prashant J. Nair, Saugata Ghose, Donghyuk Lee, Moinuddin K. Qureshi, and Onur Mutlu,
"Low-Cost Inter-Linked Subarrays (LISA): Enabling Fast Inter-Subarray Data Movement in DRAM"
Proceedings of the 22nd International Symposium on High-Performance Computer Architecture (HPCA), Barcelona, Spain, March 2016.
[[Slides \(pptx\)](#)] [[pdf](#)]
[[Source Code](#)]

Low-Cost Inter-Linked Subarrays (LISA): Enabling Fast Inter-Subarray Data Movement in DRAM

Kevin K. Chang[†], Prashant J. Nair^{*}, Donghyuk Lee[†], Saugata Ghose[†], Moinuddin K. Qureshi^{*}, and Onur Mutlu[†]

[†]Carnegie Mellon University ^{*}Georgia Institute of Technology

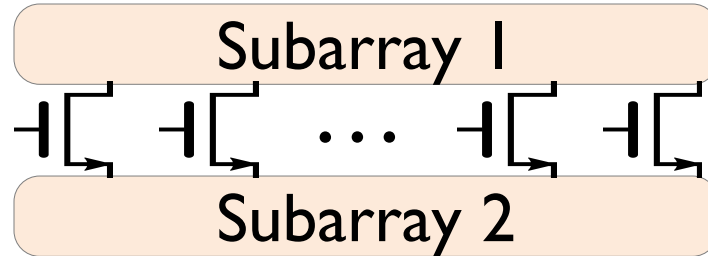
Moving Data Inside DRAM?



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

Key Idea and Applications

- **Low-cost Inter-linked subarrays (LISA)**
 - Fast bulk data movement between subarrays
 - **Wide datapath via isolation transistors**: 0.8% DRAM chip area



- LISA is a **versatile substrate** → new applications
 - Fast bulk data copy**: Copy latency 1.363ms→0.148ms (9.2x)
→ 66% speedup, -55% DRAM energy
 - In-DRAM caching**: Hot data access latency 48.7ns→21.5ns (2.2x)
→ 5% speedup
 - Fast precharge**: Precharge latency 13.1ns→5.0ns (2.6x)
→ 8% speedup

More on LISA

- Kevin K. Chang, Prashant J. Nair, Saugata Ghose, Donghyuk Lee, Moinuddin K. Qureshi, and Onur Mutlu,
"Low-Cost Inter-Linked Subarrays (LISA): Enabling Fast Inter-Subarray Data Movement in DRAM"
Proceedings of the 22nd International Symposium on High-Performance Computer Architecture (HPCA), Barcelona, Spain, March 2016.
[[Slides \(pptx\)](#)] [[pdf](#)]
[[Source Code](#)]

Low-Cost Inter-Linked Subarrays (LISA): Enabling Fast Inter-Subarray Data Movement in DRAM

Kevin K. Chang[†], Prashant J. Nair^{*}, Donghyuk Lee[†], Saugata Ghose[†], Moinuddin K. Qureshi^{*}, and Onur Mutlu[†]

[†]Carnegie Mellon University ^{*}Georgia Institute of Technology

FIGARO: Fine-Grained In-DRAM Copy

- Yaohua Wang, Lois Orosa, Xiangjun Peng, Yang Guo, Saugata Ghose, Minesh Patel, Jeremie S. Kim, Juan Gómez Luna, Mohammad Sadrosadati, Nika Mansouri Ghiasi, and Onur Mutlu,
"FIGARO: Improving System Performance via Fine-Grained In-DRAM Data Relocation and Caching"
Proceedings of the 53rd International Symposium on Microarchitecture (MICRO), Virtual, October 2020.

FIGARO: Improving System Performance via Fine-Grained In-DRAM Data Relocation and Caching

Yaohua Wang^{*} Lois Orosa[†] Xiangjun Peng^{⊙*} Yang Guo^{*} Saugata Ghose^{◇‡} Minesh Patel[†]
Jeremie S. Kim[†] Juan Gómez Luna[†] Mohammad Sadrosadati[§] Nika Mansouri Ghiasi[†] Onur Mutlu^{†‡}

^{*}National University of Defense Technology [†]ETH Zürich [⊙]Chinese University of Hong Kong

[◇]University of Illinois at Urbana–Champaign [‡]Carnegie Mellon University [§]Institute of Research in Fundamental Sciences

Network-On-Memory: Fast Inter-Bank Copy

- Seyyed Hossein SeyyedAghaei Rezaei, Mehdi Modarressi, Rachata Ausavarungnirun, Mohammad Sadrosadati, Onur Mutlu, and Masoud Daneshtalab,
"NoM: Network-on-Memory for Inter-Bank Data Transfer in Highly-Banked Memories"
IEEE Computer Architecture Letters (**CAL**), to appear in 2020.

NOm: NETWORK-ON-MEMORY FOR INTER-BANK DATA TRANSFER IN HIGHLY-BANKED MEMORIES

Seyyed Hossein SeyyedAghaei Rezaei¹
Mohammad Sadrosadati³

Mehdi Modarressi^{1,3}
Onur Mutlu⁴

Rachata Ausavarungnirun²
Masoud Daneshtalab⁵

¹University of Tehran

²King Mongkut's University of Technology North Bangkok

³Institute for Research in Fundamental Sciences

⁴ETH Zürich

⁵Mälardalens University

In-DRAM Bulk Bitwise AND/OR

- Vivek Seshadri, Kevin Hsieh, Amirali Boroumand, Donghyuk Lee, Michael A. Kozuch, Onur Mutlu, Phillip B. Gibbons, and Todd C. Mowry,
"Fast Bulk Bitwise AND and OR in DRAM"
IEEE Computer Architecture Letters (***CAL***), April 2015.

Fast Bulk Bitwise AND and OR in DRAM

Vivek Seshadri*, Kevin Hsieh*, Amirali Boroumand*, Donghyuk Lee*,
Michael A. Kozuch†, Onur Mutlu*, Phillip B. Gibbons†, Todd C. Mowry*

*Carnegie Mellon University

†Intel Pittsburgh

Ambit: Bulk-Bitwise in-DRAM Computation

- Vivek Seshadri, Donghyuk Lee, Thomas Mullins, Hasan Hassan, Amirali Boroumand, Jeremie Kim, Michael A. Kozuch, Onur Mutlu, Phillip B. Gibbons, and Todd C. Mowry,
"Ambit: In-Memory Accelerator for Bulk Bitwise Operations Using Commodity DRAM Technology"
Proceedings of the 50th International Symposium on Microarchitecture (MICRO), Boston, MA, USA, October 2017.
[\[Slides \(pptx\) \(pdf\)\]](#) [\[Lightning Session Slides \(pptx\) \(pdf\)\]](#) [\[Poster \(pptx\) \(pdf\)\]](#)

■

Ambit: In-Memory Accelerator for Bulk Bitwise Operations Using Commodity DRAM Technology

Vivek Seshadri^{1,5} Donghyuk Lee^{2,5} Thomas Mullins^{3,5} Hasan Hassan⁴ Amirali Boroumand⁵
Jeremie Kim^{4,5} Michael A. Kozuch³ Onur Mutlu^{4,5} Phillip B. Gibbons⁵ Todd C. Mowry⁵

¹Microsoft Research India ²NVIDIA Research ³Intel ⁴ETH Zürich ⁵Carnegie Mellon University

In-DRAM Bulk Bitwise Execution Paradigm

- Vivek Seshadri and Onur Mutlu,
"In-DRAM Bulk Bitwise Execution Engine"
Invited Book Chapter in Advances in Computers, to appear
in 2020.
[[Preliminary arXiv version](#)]

In-DRAM Bulk Bitwise Execution Engine

Vivek Seshadri
Microsoft Research India
visesha@microsoft.com

Onur Mutlu
ETH Zürich
onur.mutlu@inf.ethz.ch

Extensions and Follow-Up Work (II)

- Can this idea be evaluated on a real system? How?
 - Yes, [see the Compute DRAM paper \[MICRO 2019\]](#)
- Can similar ideas be used in other types of memories? Phase Change Memory? RRAM? STT-MRAM?
 - Yes, [see the Pinatubo paper \[DAC 2016\]](#)
- Can we have more efficient solutions to
 - Cache coherence (minimize overhead)
 - Data reuse after copy and initialization

Pinatubo: PCM RowClone and Bitwise Ops

Pinatubo: A Processing-in-Memory Architecture for Bulk Bitwise Operations in Emerging Non-volatile Memories

Shuangchen Li^{1*}, Cong Xu², Qiaosha Zou^{1,5}, Jishen Zhao³, Yu Lu⁴, and Yuan Xie¹

University of California, Santa Barbara¹, Hewlett Packard Labs²

University of California, Santa Cruz³, Qualcomm Inc.⁴, Huawei Technologies Inc.⁵
{shuangchenli, yuanxie}@ece.ucsb.edu¹

RowClone Demonstration in Real DRAM Chips

ComputeDRAM: In-Memory Compute Using Off-the-Shelf DRAMs

Fei Gao

feig@princeton.edu

Department of Electrical Engineering
Princeton University

Georgios Tziantzioulis

georgios.tziantzioulis@princeton.edu

Department of Electrical Engineering
Princeton University

David Wentzlaff

wentzlaf@princeton.edu

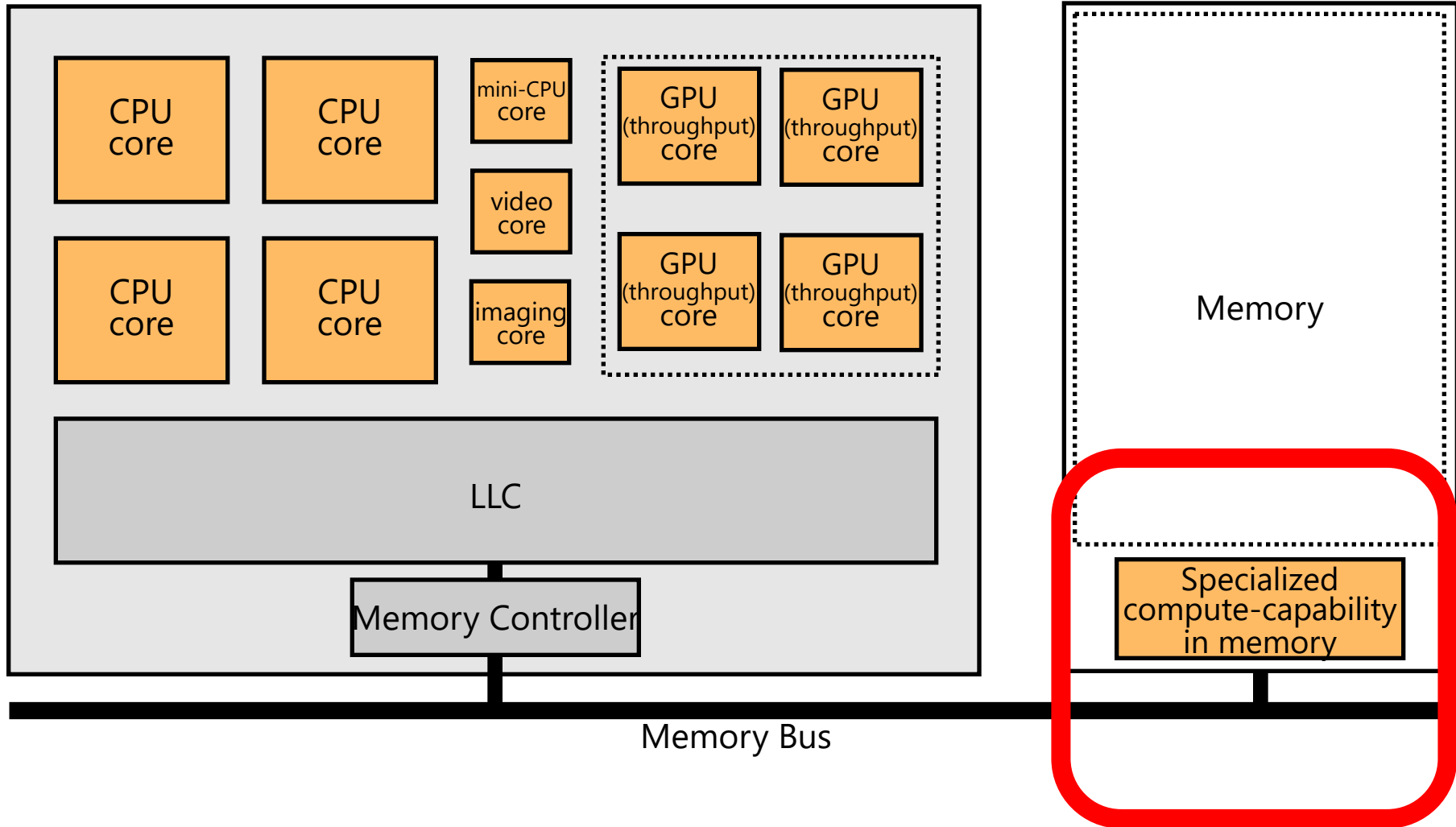
Department of Electrical Engineering
Princeton University

Takeaways

Key Takeaways

- A novel method to accelerate data copy and initialization
- Simple and effective
- Hardware/software cooperative
- Good potential for work building on it to extend it
 - To different granularities
 - To make things more efficient and effective
 - Many works have already built on the paper (see LISA, FIGARO, Ambit, ComputeDRAM, and other works in Google Scholar)
- Easy to read and understand paper

Memory as an Accelerator



Memory similar to a "conventional" accelerator

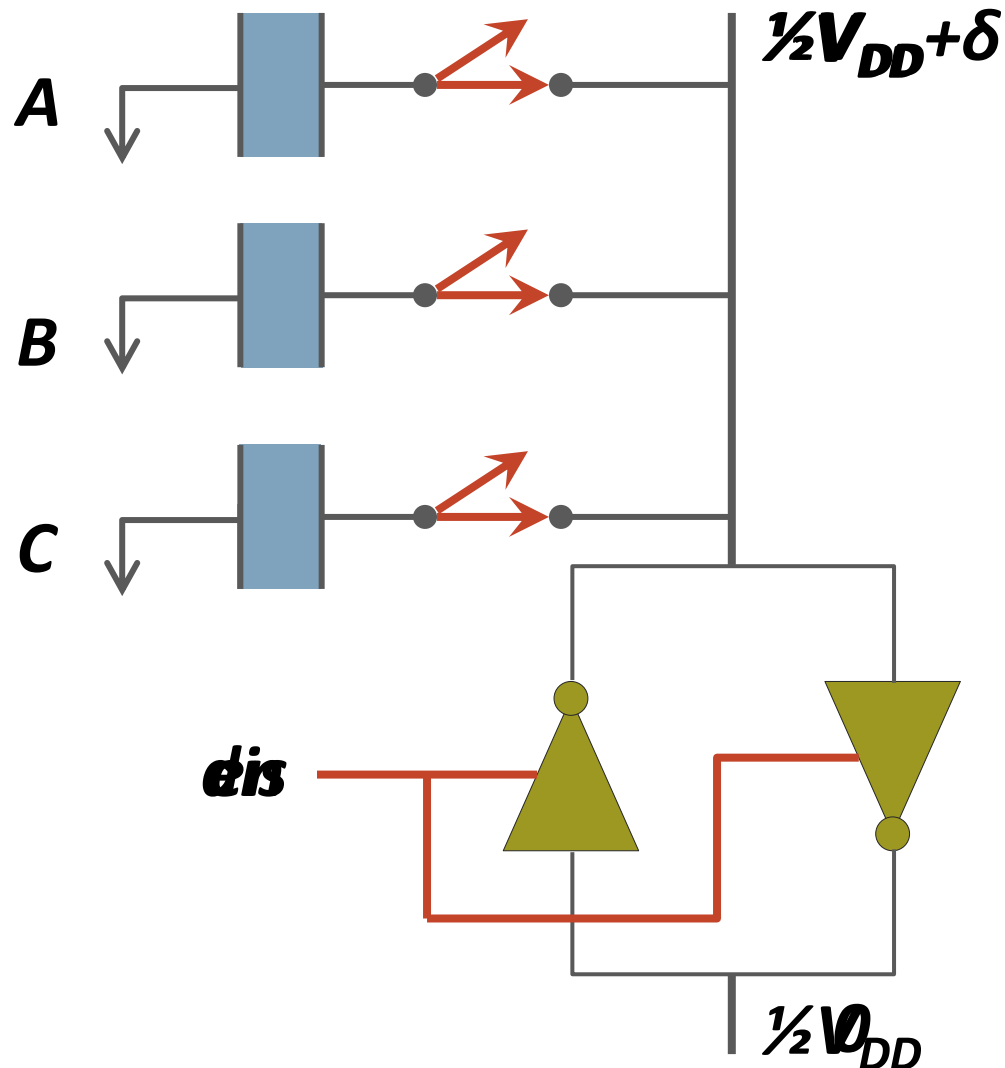
In-DRAM

Bulk Bitwise Operations

In-Memory Bulk Bitwise Operations

- We can support in-DRAM COPY, ZERO, AND, OR, NOT, MAJ
- At low cost
- Using analog computation capability of DRAM
 - Idea: activating multiple rows performs computation
- 30-60X performance and energy improvement
 - Seshadri+, "Ambit: In-Memory Accelerator for Bulk Bitwise Operations Using Commodity DRAM Technology," MICRO 2017.
- New memory technologies enable even more opportunities
 - Memristors, resistive RAM, phase change mem, STT-MRAM, ...
 - Can operate on data with minimal movement

In-DRAM AND/OR: Triple Row Activation



Final State
 $AB + BC + AC$

**$C(A + B) +$
 $\sim C(AB)$**

In-DRAM Bulk Bitwise AND/OR Operation

- **BULKAND A, B → C**
 - Semantics: Perform a bitwise AND of two rows A and B and store the result in row C
 - R0 – reserved zero row, R1 – reserved one row
 - D1, D2, D3 – Designated rows for triple activation
-
1. RowClone A into D1
 2. RowClone B into D2
 3. RowClone R0 into D3
 4. ACTIVATE D1,D2,D3
 5. RowClone Result into C

More on In-DRAM Bulk AND/OR

- Vivek Seshadri, Kevin Hsieh, Amirali Boroumand, Donghyuk Lee, Michael A. Kozuch, Onur Mutlu, Phillip B. Gibbons, and Todd C. Mowry,
"Fast Bulk Bitwise AND and OR in DRAM"
IEEE Computer Architecture Letters (***CAL***), April 2015.

Fast Bulk Bitwise AND and OR in DRAM

Vivek Seshadri*, Kevin Hsieh*, Amirali Boroumand*, Donghyuk Lee*,
Michael A. Kozuch†, Onur Mutlu*, Phillip B. Gibbons†, Todd C. Mowry*

*Carnegie Mellon University

†Intel Pittsburgh

In-DRAM NOT: Dual Contact Cell

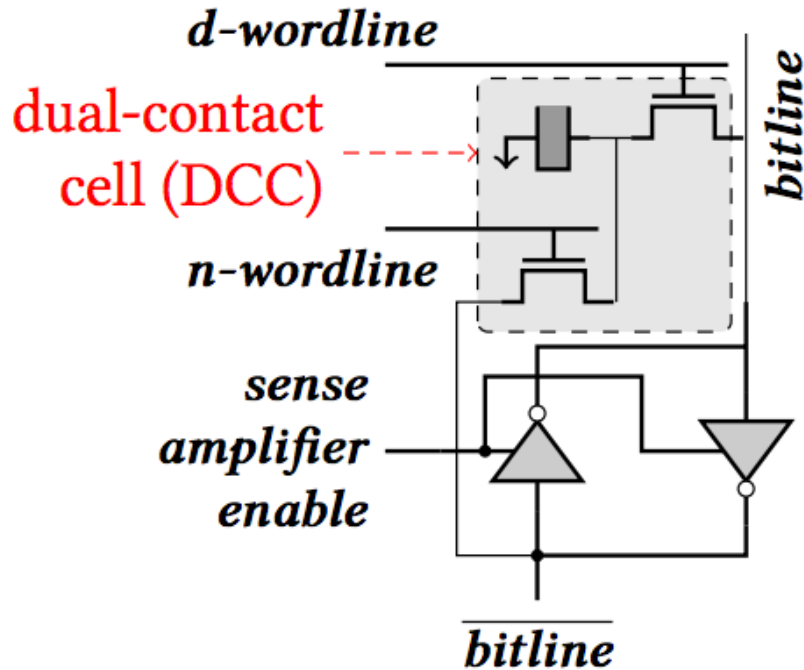


Figure 5: A dual-contact cell connected to both ends of a sense amplifier

Idea:
Feed the
negated value
in the sense amplifier
into a special row

In-DRAM NOT Operation

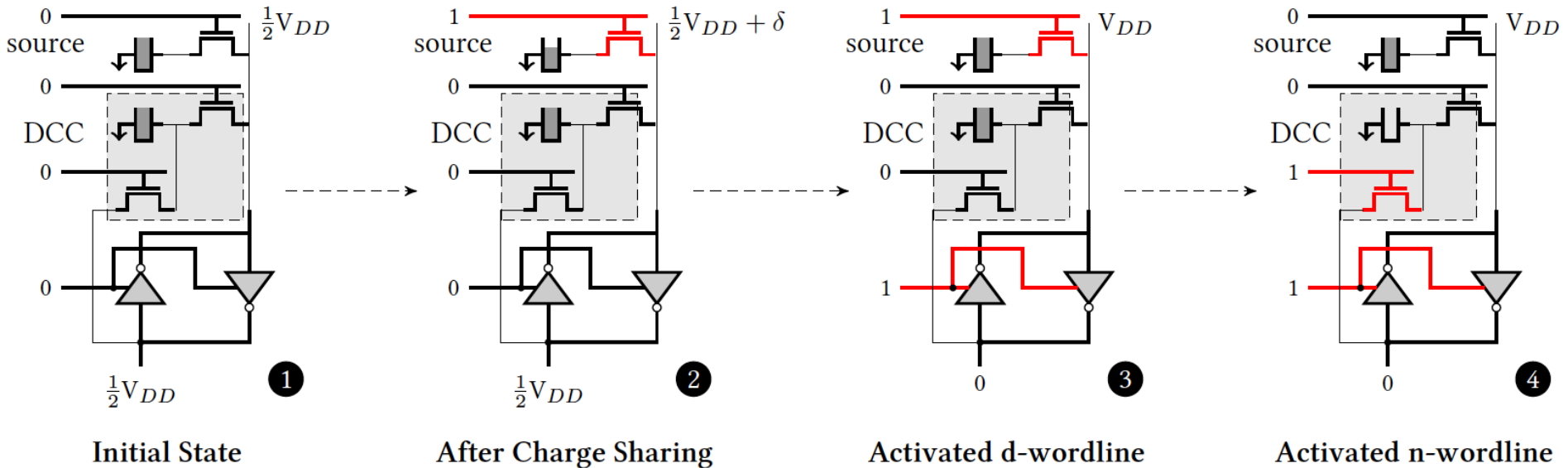


Figure 5: Bitwise NOT using a dual contact capacitor

Seshadri+, "Ambit: In-Memory Accelerator for Bulk Bitwise Operations using Commodity DRAM Technology," MICRO 2017.

Performance: In-DRAM Bitwise Operations

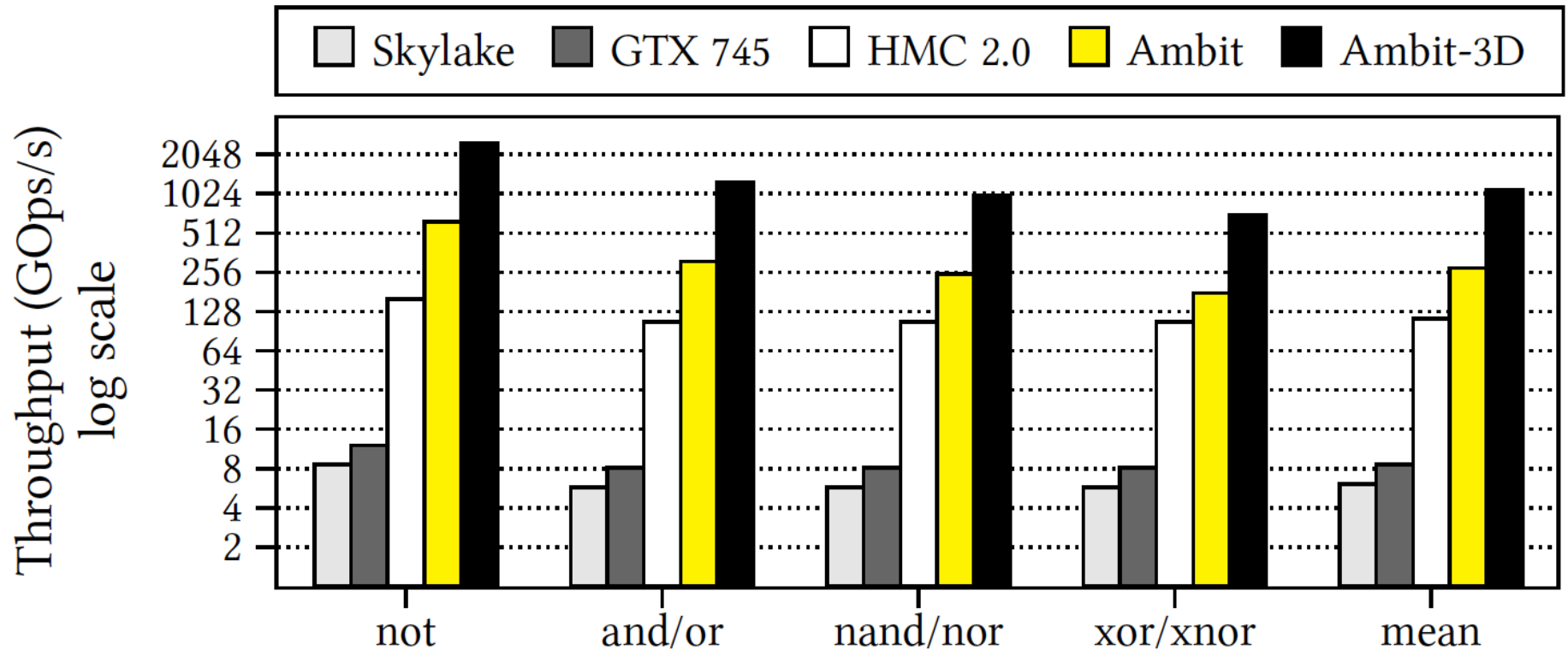


Figure 9: Throughput of bitwise operations on various systems.

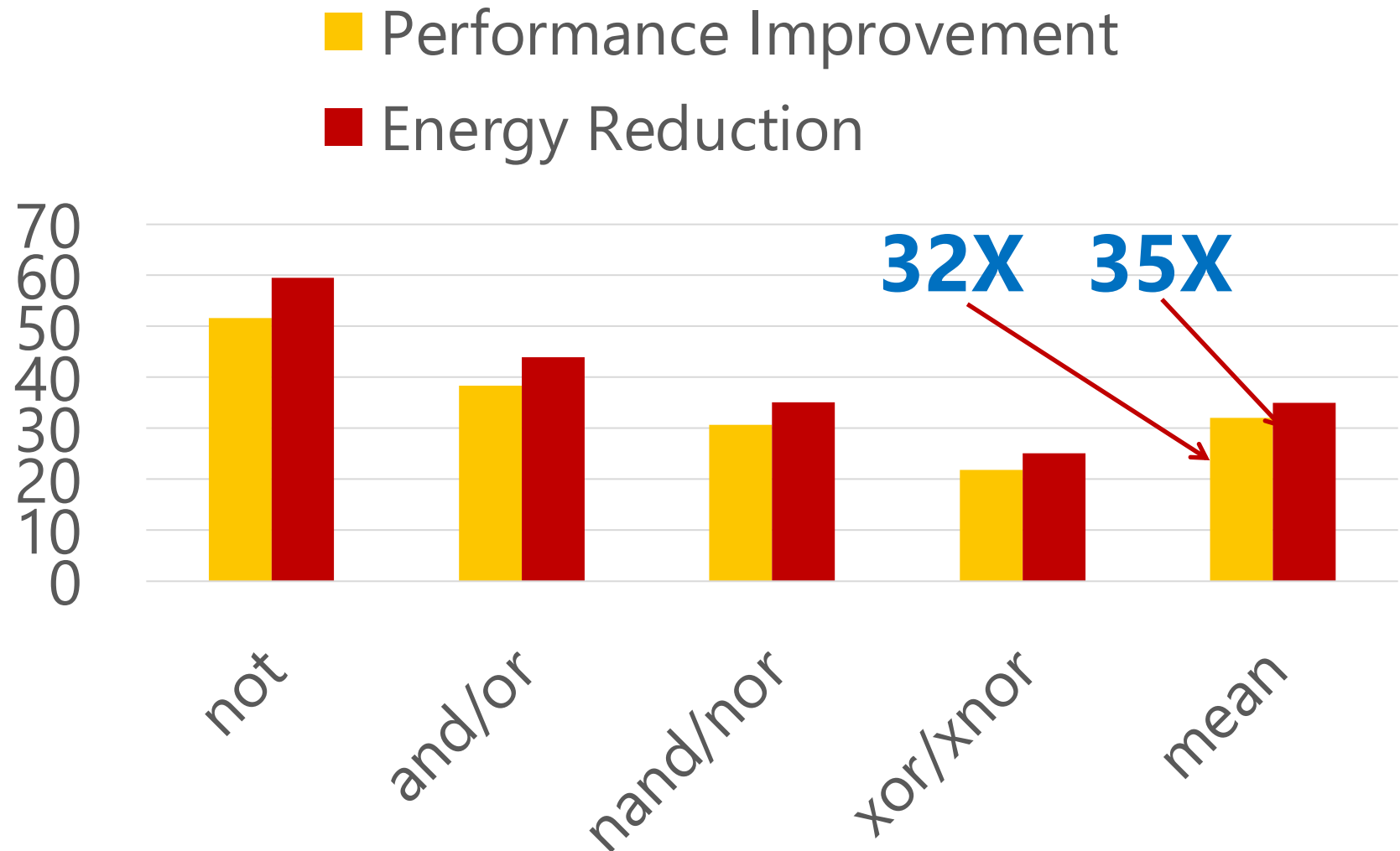
Energy of In-DRAM Bitwise Operations

	Design	not	and/or	nand/nor	xor/xnor
DRAM & Channel Energy (nJ/KB)	DDR3	93.7	137.9	137.9	137.9
	Ambit	1.6	3.2	4.0	5.5
	(↓)	59.5X	43.9X	35.1X	25.1X

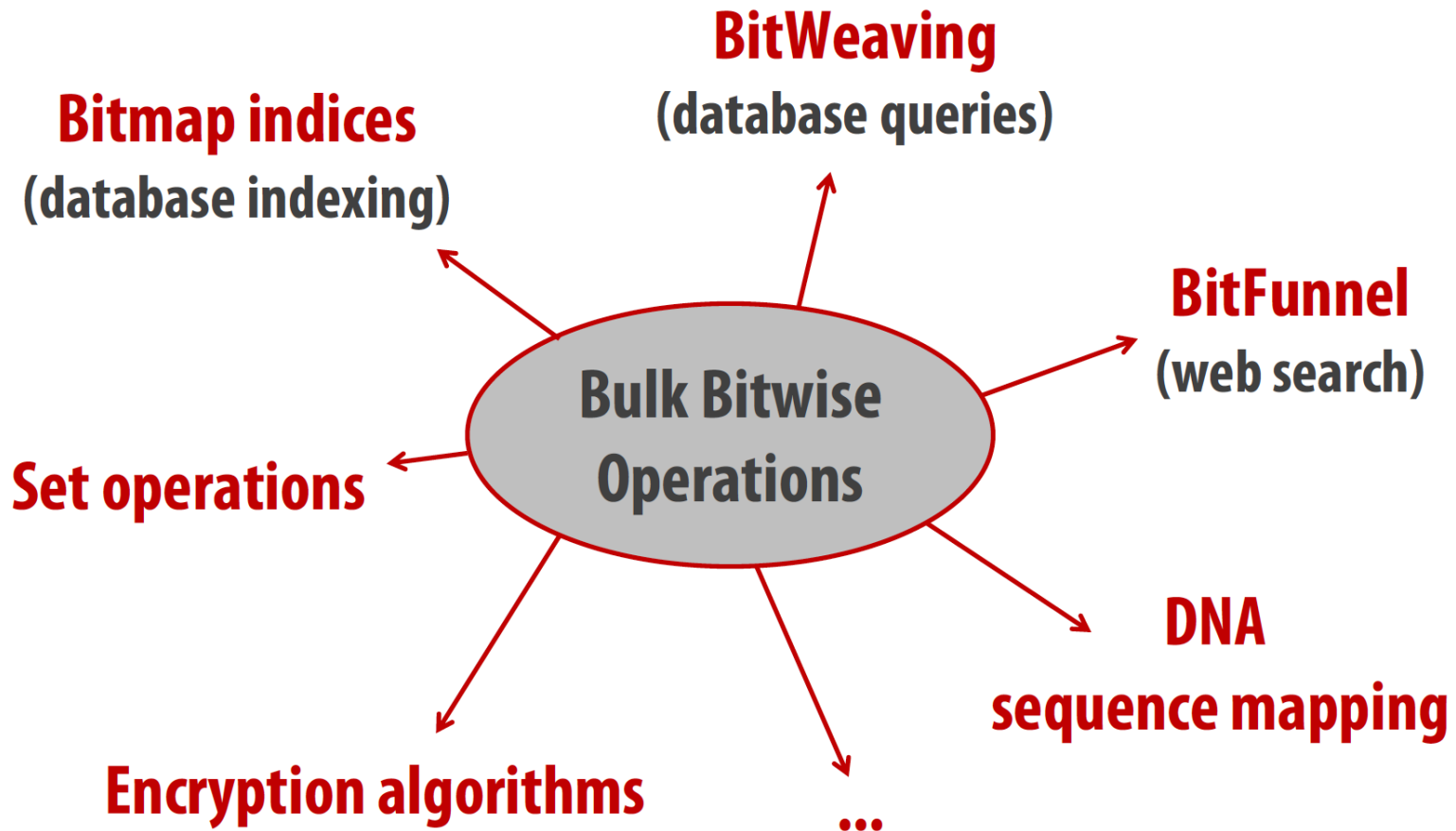
Table 3: Energy of bitwise operations. (↓) indicates energy reduction of Ambit over the traditional DDR3-based design.

Seshadri+, "Ambit: In-Memory Accelerator for Bulk Bitwise Operations using Commodity DRAM Technology," MICRO 2017.

Ambit vs. DDR3: Performance and Energy



Bulk Bitwise Operations in Workloads



Example Data Structure: Bitmap Index

- Alternative to B-tree and its variants
- Efficient for performing *range queries* and *joins*
- **Many bitwise operations to perform a query**

age < 18 18 < age < 25 25 < age < 60 age > 60

Bitmap 1

Bitmap 2

Bitmap 3

Bitmap 4

Performance: Bitmap Index on Ambit

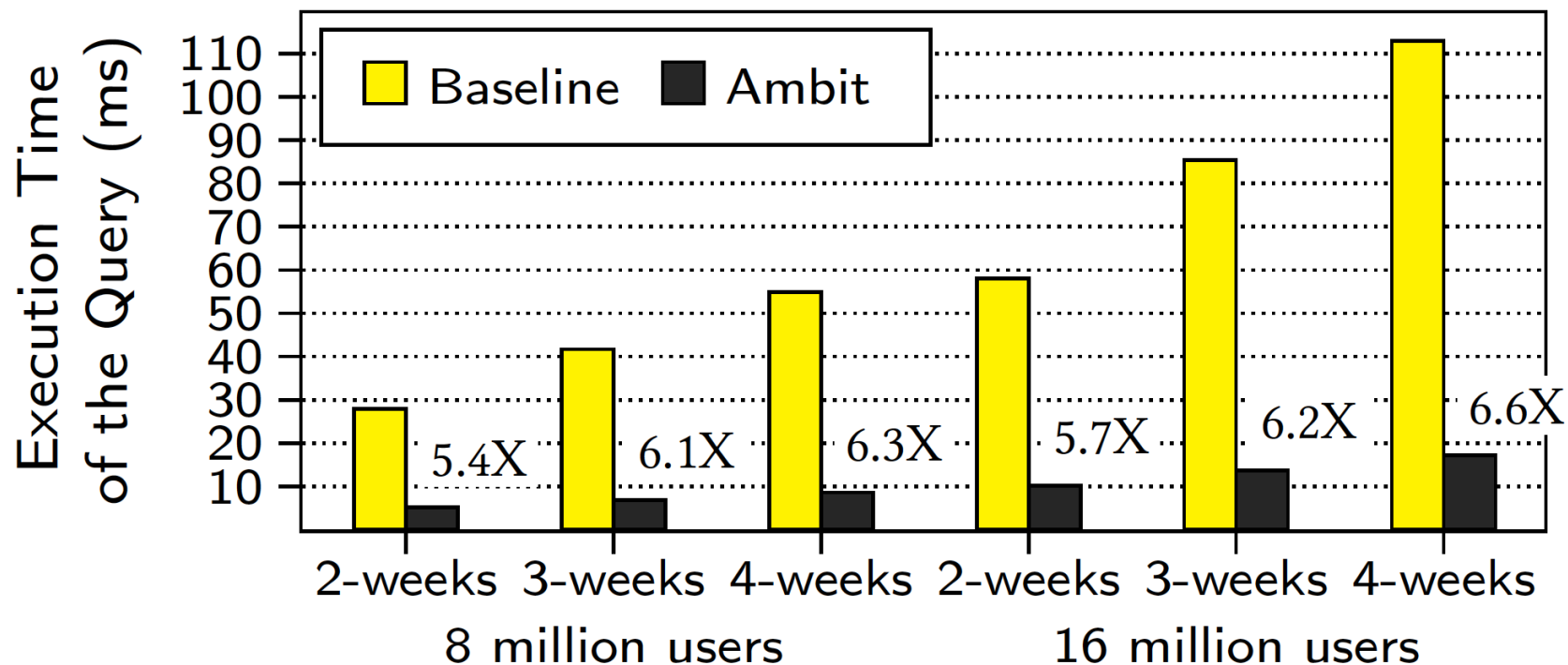


Figure 10: Bitmap index performance. The value above each bar indicates the reduction in execution time due to Ambit.

>5.4-6.6X Performance Improvement

Performance: BitWeaving on Ambit

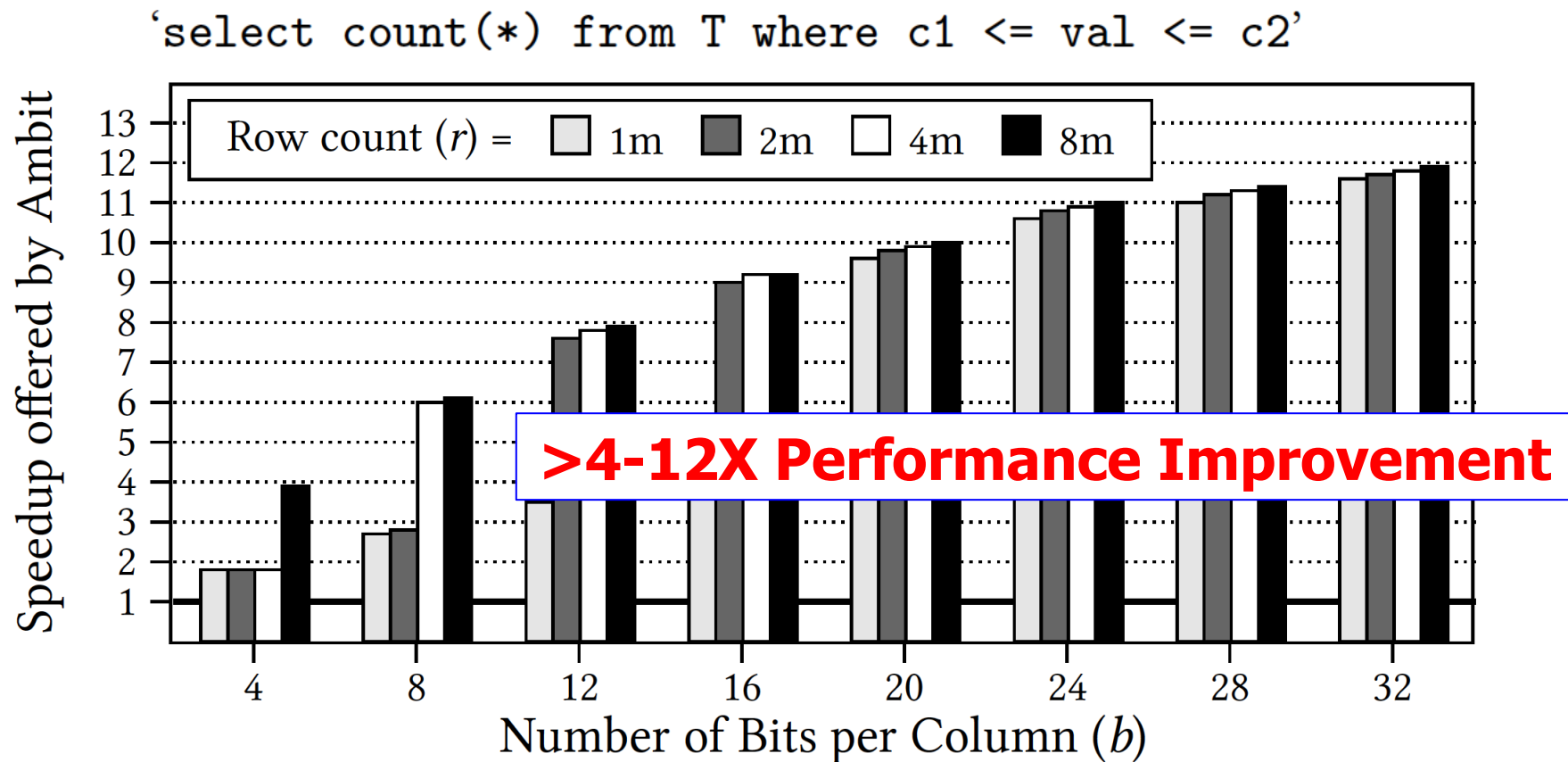


Figure 11: Speedup offered by Ambit over baseline CPU with SIMD for BitWeaving

Seshadri+, "Ambit: In-Memory Accelerator for Bulk Bitwise Operations using Commodity DRAM Technology," MICRO 2017.

More on In-DRAM Bulk AND/OR

- Vivek Seshadri, Kevin Hsieh, Amirali Boroumand, Donghyuk Lee, Michael A. Kozuch, Onur Mutlu, Phillip B. Gibbons, and Todd C. Mowry,
"Fast Bulk Bitwise AND and OR in DRAM"
IEEE Computer Architecture Letters (***CAL***), April 2015.

Fast Bulk Bitwise AND and OR in DRAM

Vivek Seshadri*, Kevin Hsieh*, Amirali Boroumand*, Donghyuk Lee*,
Michael A. Kozuch†, Onur Mutlu*, Phillip B. Gibbons†, Todd C. Mowry*

*Carnegie Mellon University

†Intel Pittsburgh

More on In-DRAM Bitwise Operations

- Vivek Seshadri et al., “**Ambit: In-Memory Accelerator for Bulk Bitwise Operations Using Commodity DRAM Technology**,” MICRO 2017.

Ambit: In-Memory Accelerator for Bulk Bitwise Operations
Using Commodity DRAM Technology

Vivek Seshadri^{1,5} Donghyuk Lee^{2,5} Thomas Mullins^{3,5} Hasan Hassan⁴ Amirali Boroumand⁵
Jeremie Kim^{4,5} Michael A. Kozuch³ Onur Mutlu^{4,5} Phillip B. Gibbons⁵ Todd C. Mowry⁵

¹Microsoft Research India ²NVIDIA Research ³Intel ⁴ETH Zürich ⁵Carnegie Mellon University

More on In-DRAM Bulk Bitwise Execution

- Vivek Seshadri and Onur Mutlu,
"In-DRAM Bulk Bitwise Execution Engine"
Invited Book Chapter in Advances in Computers, to appear
in 2020.
[[Preliminary arXiv version](#)]

In-DRAM Bulk Bitwise Execution Engine

Vivek Seshadri
Microsoft Research India
`visesha@microsoft.com`

Onur Mutlu
ETH Zürich
`onur.mutlu@inf.ethz.ch`

Challenge: Intelligent Memory Device

Does **memory**
have to be
dumb?

Computing Architectures with Minimal Data Movement

Historical Perspective & A Detour on the Review Process

Ambit and RowClone
Sound Great!
No?

Some History: RowClone

RowClone: Historical Perspective

- This work is perhaps the first example of “minimally changing DRAM chips” to perform data movement and computation
 - Surprising that it was done as late as 2013!
- It led to a body of work on in-DRAM (and in-NVM) computation with “hopefully small” changes
- Work building on RowClone still continues
- Initially, it was dismissed by some reviewers
 - Rejected from ISCA 2013 conference

One Review (ISCA 2013 Submission)

PAPER STRENGTHS

The paper includes a well written background on DRAM organization/operation. The proposed technique is simple and elegant; it nicely exploits key circuit-level characteristics of DRAM designs and minimizes the changes necessary to commodity DRAM chips.

PAPER WEAKNESSES

I am concerned on the applicability of the technique and found the evaluation to be unconvincing in terms of motivating the work as well as quantifying the potential benefit. Details on how to efficiently manage the coherence between the cache hierarchy and DRAM to enable the proposed technique are glossed over, but in my opinion are critical to the narrative.

Another Review and Rebuttal

DETAILED COMMENTS

The paper proposes a simple and not new idea, block copy in a DRAM, and the creates a complete

Reviewer B mentions that our idea is "not new". An explicit reference by the reviewer would be helpful here. While the reviewer may be referring to one of the patents that we cite in our paper (citations 2, 6, 25, 26, 27 in the paper), these patents are at a superficial level and do *not* provide a concrete mechanism. In contrast, we propose three concrete mechanisms and provide details on the most important architectural and microarchitectural modifications required at the DRAM chip, the memory controller, and the CPU to enable a system that supports the mechanisms. We also analyze their latency, hardware overhead, power, and performance in detail. We are not aware of any prior work that achieves this.

ISCA 2013 Submission

ISCA40

Paper #295

onur@cmu.edu [Profile](#) | [Help](#) | [Sign out](#)



Main



[Edit](#)

#268 Papers #353

(All)

Search

#295 RowClone: Fast and Efficient In-DRAM Copy and Initialization of Bulk Data



COMMENT

NOTIFICATION

If selected, you will receive email when updated comments are available for this paper.

+ OTHER CONFLICTS

Rejected



1014kB

Thursday 22 Nov 2012 12:11:45am EST

| 0fd459a9adc6194cda028a394d2e4d929f662f32

You are an **author** of this paper.

– ABSTRACT

Many programs initialize or copy large amounts of memory data. Initialization and copying are

+ AUTHORS

V. Seshadri, Y. Kim, D. Lee, C. Fallin, R. Ausavarungnirun, G. Pekhimenko, Y. Luo, O. Mutlu,

[Review #295A](#)

[Review #295B](#)

[Review #295C](#)

OveMer Nov WriQua RevConAnd

3

4

5

3

4

3

4

3

3

4

4

3

Yet Later... in ISCA 2015...

Profiling a warehouse-scale computer

Svilen Kanev[†]
Harvard University

Juan Pablo Darago[†]
Universidad de Buenos Aires

Kim Hazelwood[†]
Yahoo Labs

Parthasarathy Ranganathan
Google

Tipp Moseley
Google

Gu-Yeon Wei
Harvard University

David Brooks
Harvard University

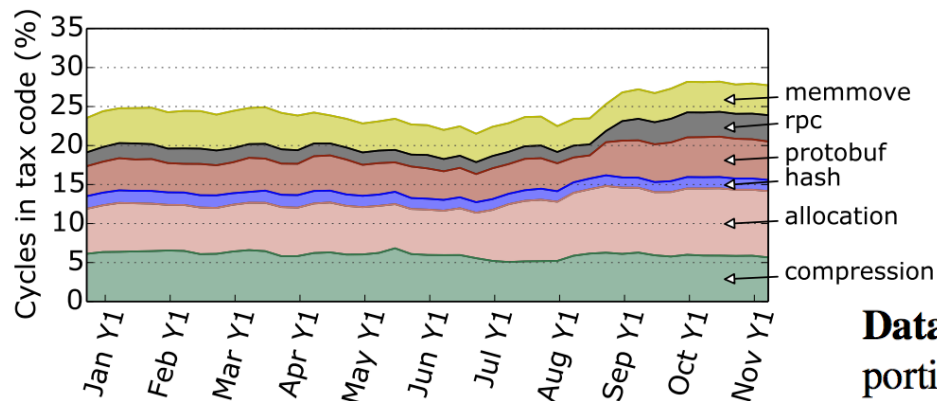


Figure 4: 22-27% of WSC cycles are spent in different components of “datacenter tax”.

we see common building blocks once we aggregate sampled profile data across many applications running in a datacenter. In this section, we quantify the performance impact of the **datacenter tax**, and argue that its components are prime candidates for hardware acceleration in future datacenter SoCs.

Data movement In fact, RPCs are by far not the only code portions that do data movement. We also tracked all calls to the `memcpy()` and `memmove()` library functions to estimate the amount of time spent on explicit data movement (i.e., exposed through a simple API). This is a conservative estimate because it does not track inlined or explicit copies. Just the variants of these two library functions represent 4-5% of datacenter cycles.

Recent work in performing data movement in DRAM [45] could optimize away this piece of tax.

MICRO 2013 Submission

#206 RowClone: Fast and Energy-Efficient In-DRAM Bulk Data Copy and Initialization

ATION
ceive

e for

Accepted



1947kB

Friday 31 May 2013 1:48:46pm PDT |
fd8423acdd9a222280302355899340083e5a40b1

You are an **author** of this paper.

+ **ABSTRACT**

Bulk data copy and initialization operations are frequently triggered by several system level operations in modern systems. Despite the fact that these operations do not require [\[more\]](#)

+ **AUTHORS**

V. Seshadri, Y. Kim, C. Fallin, D. Lee, R. Ausavarungrun, G. Pekhimenko, Y. Luo, O. Mutlu, P. Gibbons, M. Kozuch, T. Mowry
[\[details\]](#)

+ **TOPICS**

	OveMer	Nov	WriQ	Qua	RevExp
Review #206A	5	4	4	4	
Review #206B	4	2	4	4	
Review #206C	3	4	4	4	
Review #206D	3	3	4	3	
Review #206E	4	3	5	3	

More History: Ambit

Ambit

- First work on performing bulk bitwise operations in DRAM
 - By exploiting analog computation capability of bitlines
 - Extends and completes our IEEE CAL 2015 paper
- Disruptive -- spans algorithms to circuits/devices
 - Requires hardware/software cooperation for adoption
- Led to a large amount of work in similar approaches in DRAM and NVM
 - The work continues to build
- Initially, it was dismissed by many reviewers
 - Rejected from 4 conferences!

ISCA 2016: Rejected

Buddy RAM: Fast and Efficient Bulk Bitwise Operations Using DRAM

Rejected



2006kB

23 Nov 2015 11:30:23pm EST ·

7f7234da178e644380275ce12a4f539ef45c4418

You are an **author** of this paper.

► Abstract

Many data structures (e.g., database bitmap indices) rely on fast bitwise operations on large bit vectors to achieve high performance. Unfortunately, the throughput of such bulk [\[more\]](#)

► Authors

V. Seshadri, D. Lee, T. Mullins, A. Boroumand, J. Kim, M. Kozuch, O. Mutlu, P. Gibbons, T. Mowry [\[details\]](#)

► Topics and Options

	RelISC	OveMerPos	RevConAnd	Nov	WriQua
Review #171A	3	4	4	2	3
Review #171B	2	4	3	3	4
Review #171C	3	4	4	2	3
Review #171D	3	5	2	2	3
Review #171E	2	3	2	3	3

MICRO 2016: Rejected



Submission (1662kB) 10 Apr 2016 9:32:31pm EDT ·
e518c6a8916109492574858db80a6184fe61ca0c

► **Abstract**

Certain widely-used data structures
(e.g., bitmap indices) rely on

[\[more\]](#)

▼ **Authors**

Vivek Seshadri (CMU)
<vseshadr@cs.cmu.edu>
Donghyuk Lee (NVIDIA Research)
<donghyuk1@cmu.edu>
Thomas Mullins (Intel)
<thomas.p.mullins@intel.com>
Amirali Boroumand (CMU)
Jeremie Kim (CMU)
Michael A. Kozuch (Intel)
<michael.a.kozuch@intel.com>
Onur Mutlu (CMU/ETH)
<omutlu@gmail.com>
Phillip B. Gibbons (CMU)
<gibbons@cs.cmu.edu>
Todd C. Mowry (CMU) <tc>

► **Topics**

Rejected · You are an **author** of this paper.

	Pos	Reb	Ove	OveMer	RevExp	Nov	WriQua
Review #249A	2	2	4	3	3		
Review #249B	4	4	3	3	5		
Review #249C	2	3	4	2	3		
Review #249D	5	5	2	3	3		
Review #249E	5	5	2	2	3		
Review #249F	3	3	3	3	4		

HPCA 2017: Rejected

1)~significantly improves the performance of queries in applications that use bitmap indices for fast analytics, and
2)~makes bit vectors more attractive than red-black trees to represent sets. We believe Buddy can trigger programmers to redesign applications to use bitwise operations with the goal of achieving high performance and efficiency.

Rejected · You are an **author** of this paper.

	OveMer	RevExp	WriQua	ExpMet	Nov
Review #119A	1	2	3	2	2
Review #119B	4	1	4	4	3
Review #119C	4	4	4	4	4
Review #119D	3	1	4	4	3
Review #119E	3	2	5	4	4

1 Comment: [Response \(V. Seshadri\)](#)

ISCA 2017: Rejected

Rejected



Submission ⌚ 19 Nov 2016 12:03:02am EST ·

📌 3eea263e35e53552851cab5225162776f809eaa

► Abstract

Bitwise operations are an important component of

► Authors

V. Seshadri, D. Lee, T. Mullins, H. Hassan, A. Boroumand, J. Kim, M. Kozuch, O. Mutlu, P. Gibbons, T. Mowry [\[details\]](#)

[\[more\]](#)

► Topics and Options

	PosRebOve	OveMer	Nov	WriQua	RevExp
Review #162A	1	2	2	4	5
Review #162B	2	2	3	3	3
Review #162C	4	4	3	4	4
Review #162D	3	3	3	4	4
Review #162E	4	4	3	4	3

Ambit Sounds Good, No?

Review from ISCA 2016

Paper summary

The paper proposes to extend DRAM to include bulk, bit-wise logical operations directly between rows within the DRAM.

Strengths

- Very clever/novel idea.
 - Great potential speedup and efficiency gains.
-

Weaknesses

- Probably won't ever be built. Not practical to assume DRAM manufacturers will change DRAM in this way.

Very Interesting and Novel, BUT ...

Comments for the authors

I found this idea very interesting and novel. In particular, while there have been many works proposing moving computation closer to memory, I'm not aware of any work which proposes to leverage the DRAM rows themselves to implement the computation. The benefits to this approach are large in that no actual logic is used to implement the logical functions. Further the operation occurs in parallel across the whole row, a huge degree of data parallelism.

... This Will Never Get Implemented

- The biggest problem with the work is that it underestimates the difficulty in modifying DRAM process for benefit in only a subset of applications which do bulk bitwise operations. In particular, I find it hard to believe that the commodity DRAM industry will incorporate this into their standard DRAM process. DRAM process is, at this point, a highly optimized, extremely tuned endeavor. Adding this kind of functionality will have a big impact on DRAM cost. The performance benefit on the subset of applications isn't enough to justify the higher costs this will incur and this will never get implemented.

Another Review

Another Review from ISCA 2016

Strengths

The proposed mechanisms effectively exploit the operation of the DRAM to perform efficient bitwise operations across entire rows of the DRAM.

Weaknesses

This requires a modification to the DRAM that will only help this type of bitwise operation. It seems unlikely that something like that will be adopted.

... This Will Never Get Implemented

Comments for the authors

This paper shows that DRAM could be modified to support bitwise operations directly within the DRAM itself. The performance advantages are compelling for situations in which bulk bitwise operations matter.

However, I am not really convinced that any DRAM manufacturer would really consider modifying the DRAM in this way. It benefits one specific type of operation, and while that is important for some applications, it is not really a general-purpose operation. It is not like the STL library would be changed to use this for its implementation of sets.

Yet Another Review

Yet Another Review from ISCA 2016

Weaknesses

The core novelty of Buddy RAM is almost all circuits-related (by exploiting sense amps). I do not find architectural innovation even though the circuits technique benefits architecturally by mitigating memory bandwidth and relieving cache resources within a subarray. The only related part is the new ISA support for bitwise operations at DRAM side and its induced issue on cache coherence.

This paper suits better to be peer-reviewed and published in a circuit conference or with a fabricated chip in ISSCC.

A Review from HPCA 2017: REJECT

#119 - HPCA23

Review #119A

Paper summary

Paper proposes DRAM technology changes (inverts, etc) to implement bit-wise operations directly on DRAM rows.

Overall merit

1. Reject

Reviewer expertise

2. I have passing familiarity with this area

Experimental methodology

2. Poor

Strengths

Seems like a new idea. Processor-in-Memory (PIM) ideas have resurged.

Weaknesses



* Impractical. Too many implications on ISA, DRAM design, and coherence protocols.

* Unlikely to benefit real-world computations.

* Evaluation did not consider full-program performance.

Comments for author

I am skeptical this would benefit real-world computations. I've never seen real-world program profiles with hot functions or instructions that are bit-wise operations.

On the other hand, I *have* seen system profiles that show non-trivial time zeroing pages. Suggest re-tooling your work to support page zeroing and evaluating that with a full-system simulation. Take a look at when/why the Linux kernel zeroes pages. You might be surprised at the possible impact.

A Review from ISCA 2017

#162 - ISCA 2017

Review #162A Updated 28 Jan 2017 5:16:50am EST

 [Plain text](#)

Post rebuttal overall merit

1. Reject

Overall merit

2. Weak reject

Novelty

2. Incremental improvement

Writing quality

4. Well-written

Reviewer expertise

5. This is my area

Paper summary

This paper proposes in-DRAM bit-wise operations by activating more than one word lines (and cells connected to the wordiness). Basically, it's a charge-based computation where the difference in charge stored cells connected to the same bit line is used for the logic operation.

Strengths

- conceptually a very interesting proposal (but practically not sure).
- consider various aspects including the interaction between

processors and RAM (although there isn't any new contribution and rather use the same proposal as prior work).

Weaknesses

- negative impact on the regularity of DRAM array design (and associated overhead evaluation seems to be very weak.
- significantly increase the testing cost

Comments to authors

This is an interesting proposal and well presented paper. However, I have some concerns regarding the evaluation (especially related to circuit level issues).

Especially, I feel that the variation related modeling and evaluation are weak as there are multiple sources of variations such as access transistors and sense-amp mismatches, minor defects in either access transistors and/or capacitor that can manifest in this particular proposed operation scenarios. That is, the authors oversimplify the variation modeling, which I believe failed to convince me this will work in practice. Also, the area overhead analysis sounds hand-waivy. I totally understand the difficulty of DRAM overhead analysis but also we must pursue more precise ways of evaluating the area impact as DRAM is very cost-sensitive.

Another Review from ISCA 2017

Review #162B Updated 1 Feb 2017 6:50:31pm EST

 [Plain text](#)

Post rebuttal overall merit

2. Weak reject

Overall merit

2. Weak reject

Novelty

3. New contribution

Writing quality

3. Adequate

Reviewer expertise

3. I know the material, but am not an expert

Paper summary

This paper proposes performing bulk bit-wise operations at DRAM. They leverage analog operation of DRAM, and add some extra

#162 - ISCA 2017

circuits to do bit-wise operations at row granularity.

Strengths

The idea of handling bit-wise operations in memory is interesting.

Weaknesses

Not motivated well.

Not convinced the possible gains worth all the complexity.

Not convinced if the proposal is applicable in real world applications that do bit-wise operations on different data granularity.

Comments to authors

* The paper lacks motivation. The authors talk about how common bit-wise operations are. However, they do not provide any stat on how often these operations are being used, and more importantly, on what data granularity.

* Although bit-wise operations are common in some applications, they are not necessarily done at large granularity. For example, many applications do bit-wise operations at small 64-byte (or even smaller) entities. For such cases, this paper requires copying two whole rows to some temporary rows, and doing the operation on those rows. Please explain how you handle such cases, and what the benefits would be.

* What happens if the user does bit-wise operation on two 8-byte data, and want to store it in a third block?

* What happens if both operands are located in one row?

* The main issue with this work is that it requires flushing blocks out of caches to do the bit-wise operations. Imagine you have blocks A and B in the cache, as discussed in section 6.2.3., the proposal would flush them out of caches (not sure how?), writes

ISCA 2017 Summary

@A1 6 Mar 2017

This paper was discussed both online and at the PC meeting. Reviewers were uniformly positive about the novelty of the proposed Buddy-RAM design. However, reviewers were also concerned about the feasibility of the design. During the post-rebuttal and PC discussion, the main concerns raised were (1) the impact of process variation on the design's functional correctness;

#162 - ISCA 2017

(2) the potential reliability issues that arise due to the lack of ECC/CRC mechanisms; and (3) the impact on DRAM testing cost.

Specifically on point (1), some reviewers raised concerns about the limitations of the simulations performed to address variability: "Monte-Carlo cannot capture tail distribution of cell failures. Also Monte-Carlo cannot capture random correlated WID process variation issues (only some random uncorrelated variations)."

Given these concerns, the PC ultimately decided to reject the paper. We hope that this feedback is useful in preparing a future version of the paper.

The Reviewer Accountability Problem

Acknowledgments

We thank the reviewers of ISCA 2016/2017, MICRO 2016/2017, and HPCA 2017 for their valuable comments. We

MICRO 2017: Accepted

Accepted



Submission (1837kB)

4 Apr 2017 11:33:57pm EDT ·

7420f9f02c549bccca0dc6216a5e9887dffe0d422



Revision (1852kB)

14 Jun 2017 4:16am EDT ·

22f0d123a22cf04960928e0ac43d972b5a33a848

▼ Abstract

Many important applications trigger bitwise operations on large bit vectors (bulk bitwise operations). In fact, recent

► Authors

V. Seshadri, D. Lee, T. Mullins, H. Hassan, A. Boroumand, J. Kim, M. Kozuch, O. Mutlu, P. Gibbons, T. Mowry [\[details\]](#)

	PosResOve	OveMer	RevExp	Nov	PotImp	WriQua	ImpRev
Review #347A	4	3	5	4	3	4	2
Review #347B	3	4	5	4	3	4	3
Review #347C		2	5	3	2	4	4
Review #347D	3	4	4	4	4	4	2
Review #347E	3	3	4	3	3	3	4
Review #347F	3	2	4	3	3	4	1

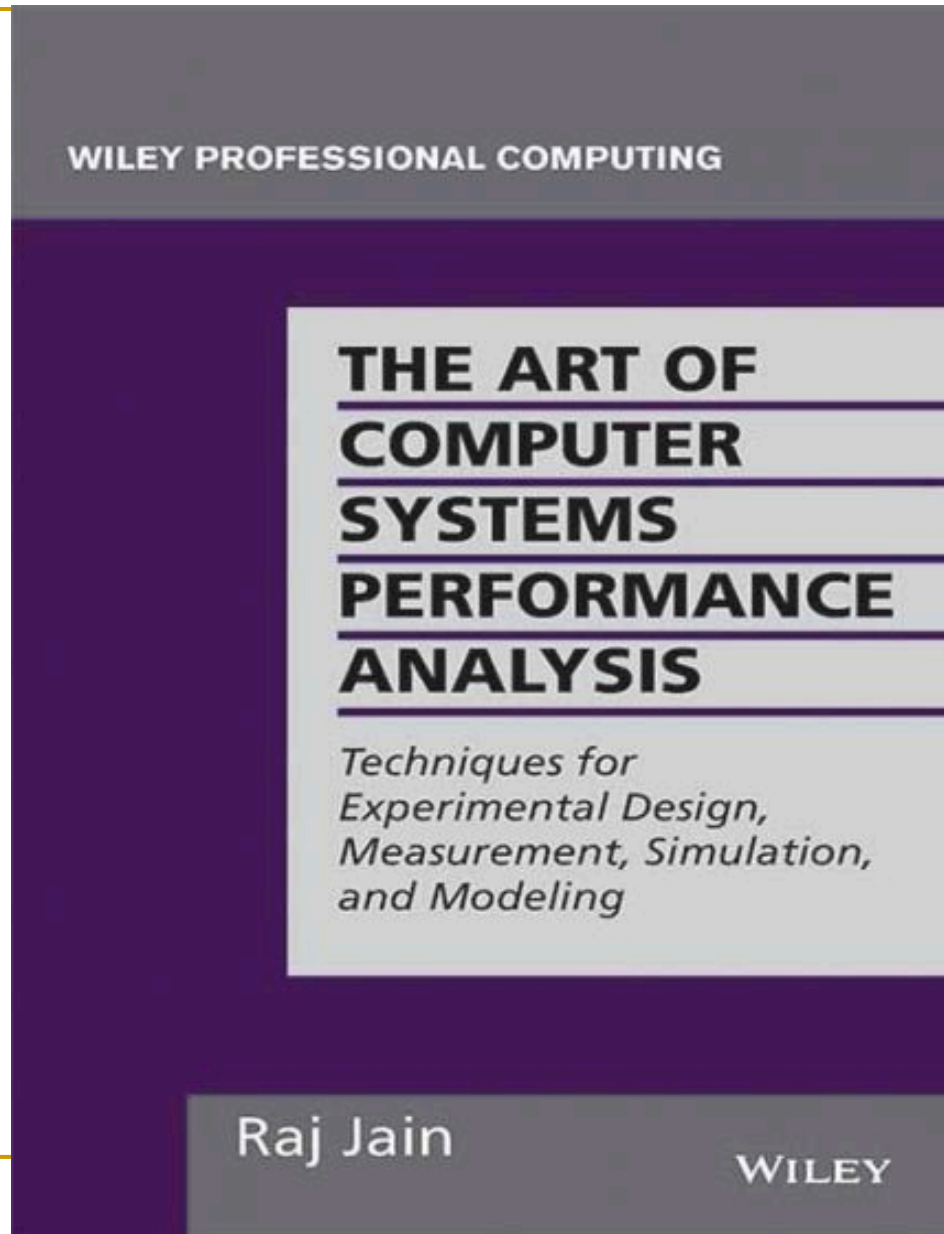
1 Comment: [Rebuttal Response \(V. Seshadri\)](#)

We Have a Mindset Issue...

- There are many other similar examples from reviews...
 - For many other papers...
- And, we are not even talking about JEDEC yet...
- How do we fix the mindset problem?
- By doing more research, education, implementation in alternative processing paradigms

We need to work on enabling the better future...

Aside: A Recommended Book



Raj Jain, "[The Art of Computer Systems Performance Analysis](#)," Wiley, 1991.

10.8 DECISION MAKER'S GAMES

Even if the performance analysis is correctly done and presented, it may not be enough to persuade your audience—the decision makers—to follow your recommendations. The list shown in Box 10.2 is a compilation of reasons for rejection heard at various performance analysis presentations. You can use the list by presenting it immediately and pointing out that the reason for rejection is not new and that the analysis deserves more consideration. Also, the list is helpful in getting the competing proposals rejected!

There is no clear end of an analysis. Any analysis can be rejected simply on the grounds that the problem needs more analysis. This is the first reason listed in Box 10.2. The second most common reason for rejection of an analysis and for endless debate is the workload. Since workloads are always based on the past measurements, their applicability to the current or future environment can always be questioned. Actually workload is one of the four areas of discussion that lead a performance presentation into an endless debate. These “rat holes” and their relative sizes in terms of time consumed are shown in Figure 10.26. Presenting this cartoon at the beginning of a presentation helps to avoid these areas.

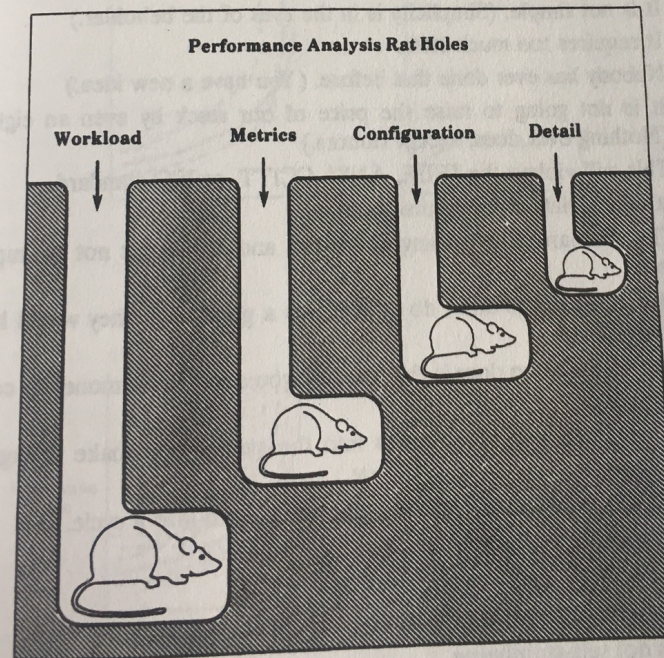


FIGURE 10.26 Four issues in performance presentations that commonly lead to endless discussion.

Raj Jain, "The Art of Computer Systems Performance Analysis," Wiley, 1991.

Box 10.2 Reasons for Not Accepting the Results of an Analysis

1. This needs more analysis.
2. You need a better understanding of the workload.
3. It improves performance only for long I/O's, packets, jobs, and files, and most of the I/O's, packets, jobs, and files are short.
4. It improves performance only for short I/O's, packets, jobs, and files, but who cares for the performance of short I/O's, packets, jobs, and files; its the long ones that impact the system.
5. It needs too much memory/CPU/bandwidth and memory/CPU/bandwidth isn't free.
6. It only saves us memory/CPU/bandwidth and memory/CPU/bandwidth is cheap.
7. There is no point in making the networks (similarly, CPUs/disks/...) faster; our CPUs/disks (any component other than the one being discussed) aren't fast enough to use them.
8. It improves the performance by a factor of x , but it doesn't really matter at the user level because everything else is so slow.
9. It is going to increase the complexity and cost.
10. Let us keep it simple stupid (and your idea is not stupid).
11. It is not simple. (Simplicity is in the eyes of the beholder.)
12. It requires too much state.
13. Nobody has ever done that before. (You have a new idea.)
14. It is not going to raise the price of our stock by even an eighth. (Nothing ever does, except rumors.)
15. This will violate the IEEE, ANSI, CCITT, or ISO standard.
16. It may violate some future standard.
17. The standard says nothing about this and so it must not be important.
18. Our competitors don't do it. If it was a good idea, they would have done it.
19. Our competition does it this way and you don't make money by copying others.
20. It will introduce randomness into the system and make debugging difficult.
21. It is too deterministic; it may lead the system into a cycle.
22. It's not interoperable.
23. This impacts hardware.
24. That's beyond today's technology.
25. It is not self-stabilizing.
26. Why change—it's working OK.

Raj Jain, "The Art of Computer Systems Performance Analysis," Wiley, 1991.

We Need to Fix the Reviewer Accountability Problem

Main Memory Needs Intelligent Controllers

Research Community
Needs

Accountable Reviewers

Suggestions to Reviewers

- Be fair; you do not know it all
- Be open-minded; you do not know it all
- Be accepting of diverse research methods: there is no single way of doing research
- Be constructive, not destructive
- Do not have double standards...

Do not block or delay scientific progress for non-reasons

RowClone & Bitwise Ops in Real DRAM Chips

ComputeDRAM: In-Memory Compute Using Off-the-Shelf DRAMs

Fei Gao

feig@princeton.edu

Department of Electrical Engineering
Princeton University

Georgios Tziantzioulis

georgios.tziantzioulis@princeton.edu

Department of Electrical Engineering
Princeton University

David Wentzlaff

wentzlaf@princeton.edu

Department of Electrical Engineering
Princeton University

RowClone & Bitwise Ops in Real DRAM Chips

MICRO-52, October 12–16, 2019, Columbus, OH, USA

Gao et al.

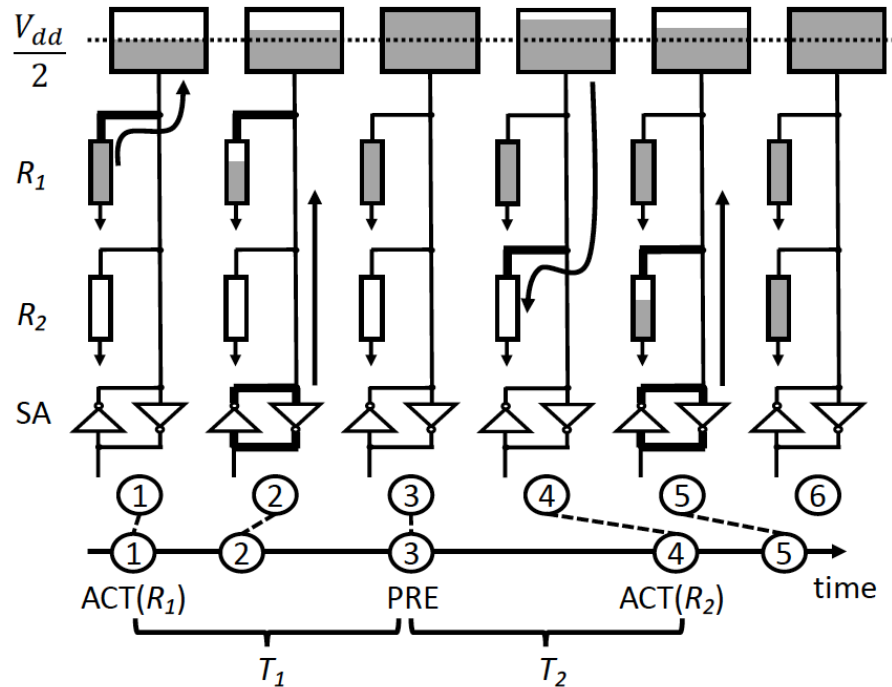


Figure 4: Timeline for a single bit of a column in a row copy operation. The data in R_1 is loaded to the bit-line, and overwrites R_2 .

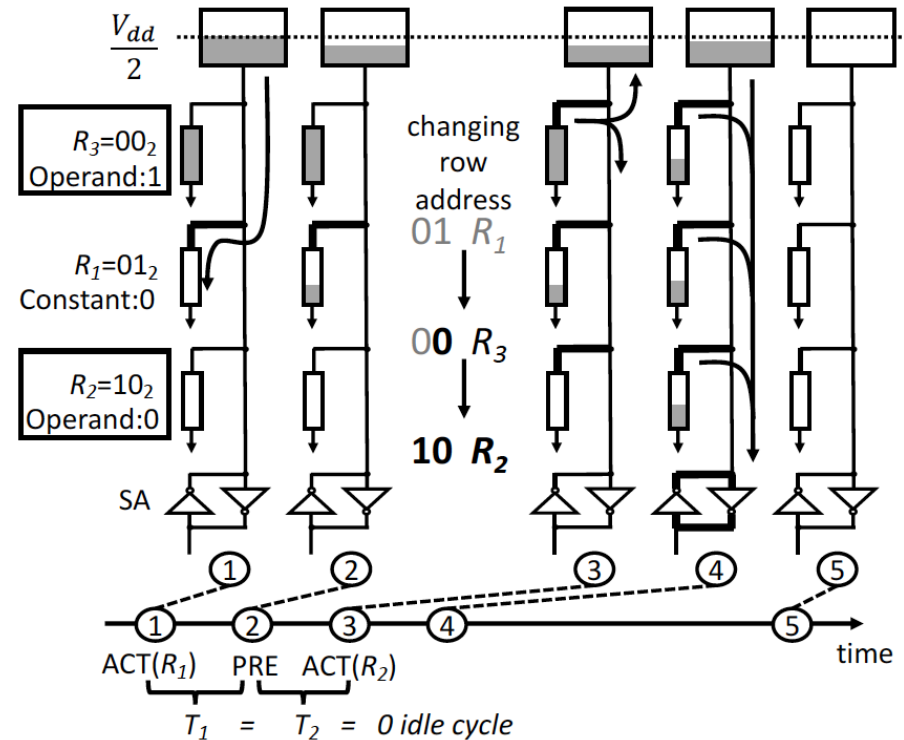
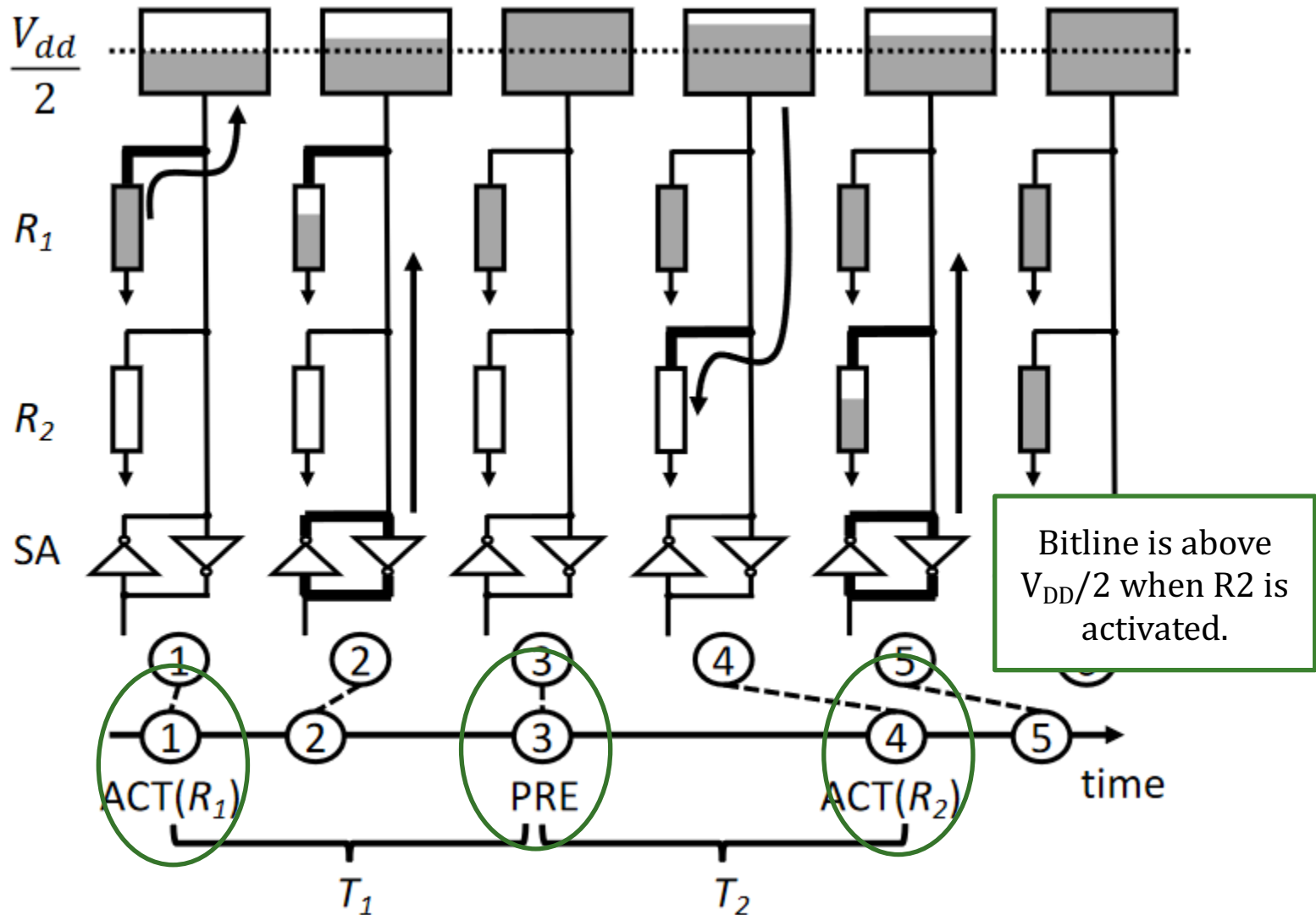
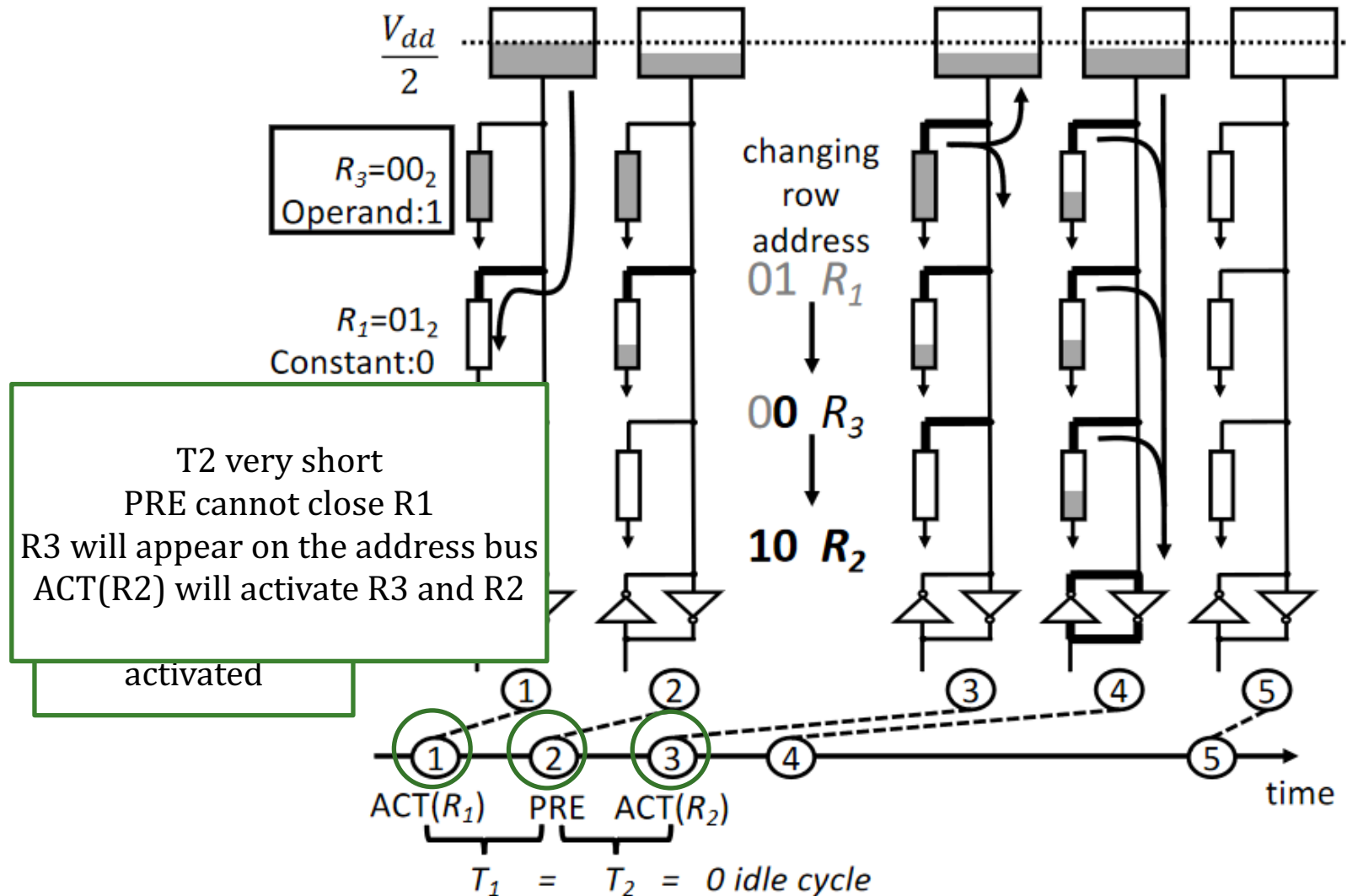


Figure 5: Logical AND in ComputeDRAM. R_1 is loaded with constant zero, and R_2 and R_3 store operands (0 and 1). The result ($0 = 1 \wedge 0$) is finally set in all three rows.

Row Copy in ComputeDRAM



Bitwise AND in ComputeDRAM



Experimental Methodology

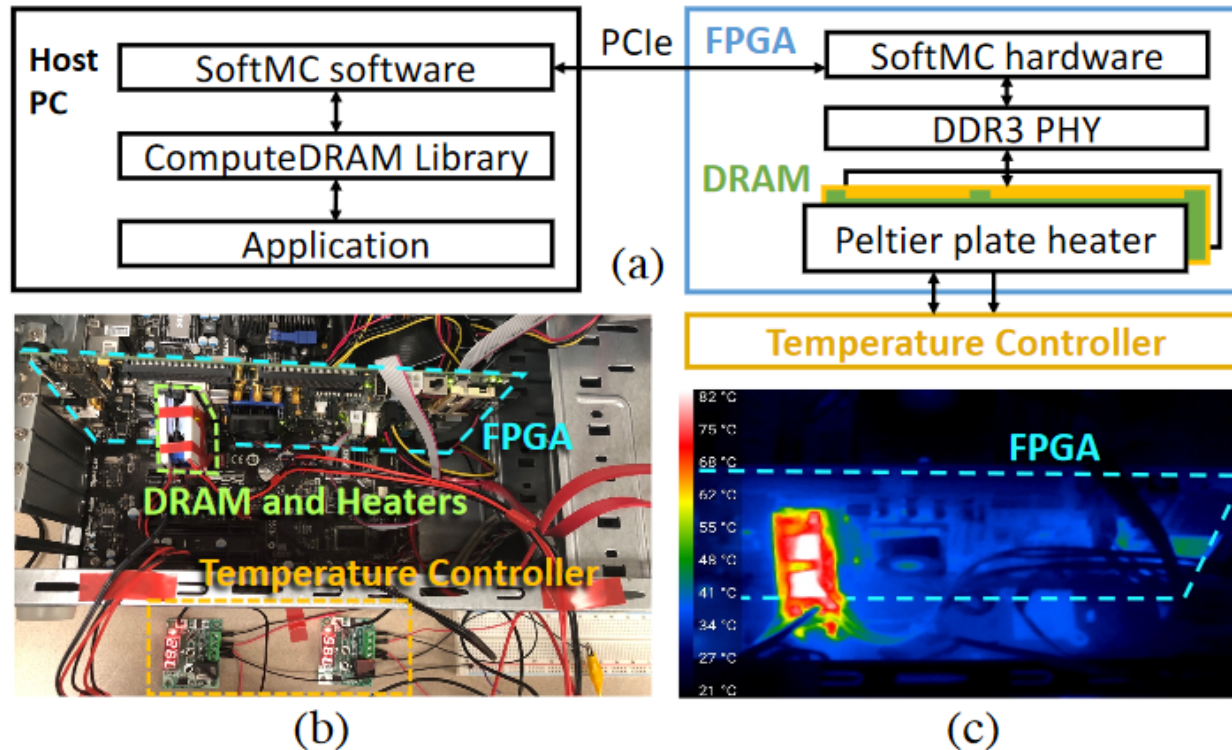


Figure 9: (a) Schematic diagram of our testing framework. (b) Picture of our testbed. (c) Thermal picture when the DRAM is heated to 80 °C.

Experimental Methodology

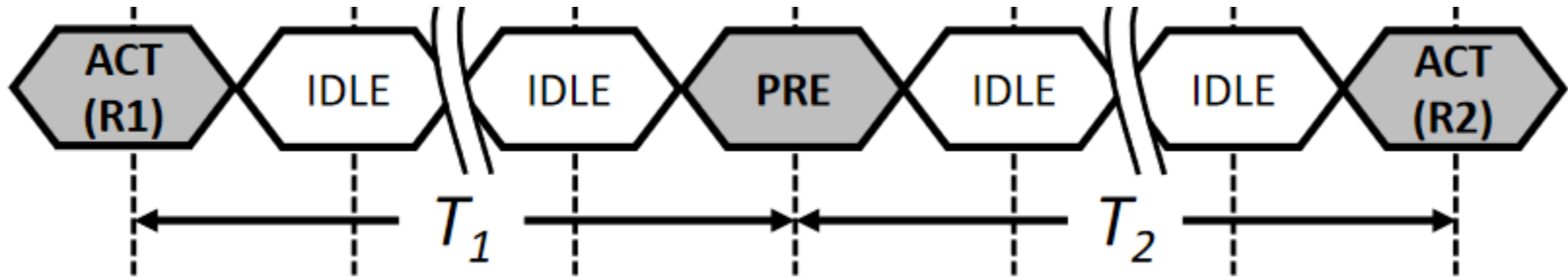
Table 1: Evaluated DRAM modules

Group ID: Vendor_Size_Freq(MHz)	Part Num	# Modules
SKhynix_2G_1333	HMT325S6BFR8C-H9	6
SKhynix_4G_1333		2
SKhynix_4G_1333		2
SKhynix_4G_1333		4
SKhynix_4G_1333		2
Samsung_4G_1333		2
Samsung_4G_1333		2
Micron_2G_1333		2
Micron_2G_1333		2
Elpida_2G_1333	EBJ21UE8BDS0-DJ-F	2
Nanya_4G_1333	NT4GC64B8HG0NS-CG	2
TimeTec_4G_1333	78AP10NUS2R2-4G	2
Corsair_4G_1333	CMSA8GX3M2A1333C9	2

**32 DDR3 Modules
~256 DRAM Chips**

Proof of Concept

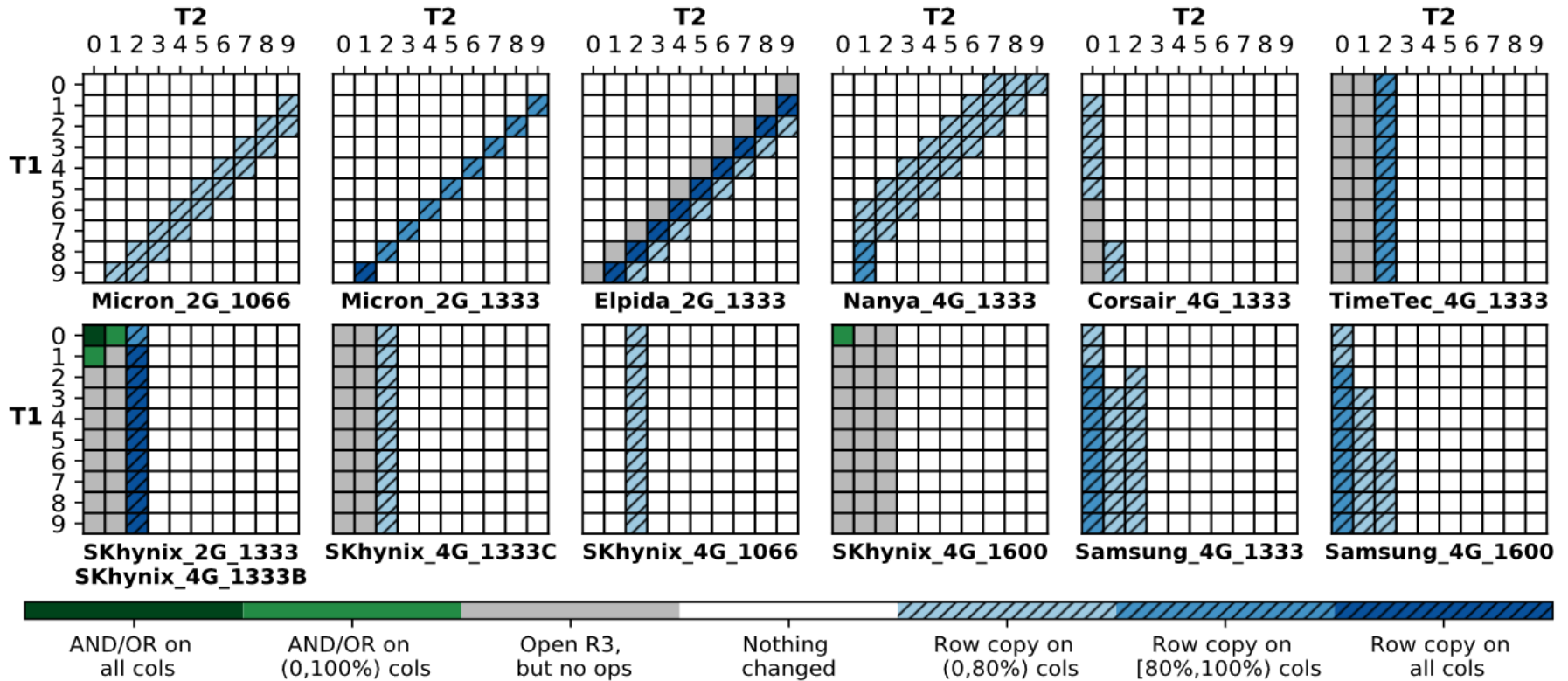
- How they test these memory modules:
 - Vary T_1 and T_2 , observe what happens.



SoftMC Experiment

1. Select a random subarray
2. Fill subarray with random data
3. Issue ACT-PRE-ACTs with given T_1 & T_2
4. Read out subarray
5. Find out how many columns in a row support either operation
 - Row-wise success ratio

Proof of Concept



- Each grid represents the success ratio of operations for a specific DDR3 module.

Pinatubo: RowClone and Bitwise Ops in PCM

Pinatubo: A Processing-in-Memory Architecture for Bulk Bitwise Operations in Emerging Non-volatile Memories

Shuangchen Li^{1*}, Cong Xu², Qiaosha Zou^{1,5}, Jishen Zhao³, Yu Lu⁴, and Yuan Xie¹

University of California, Santa Barbara¹, Hewlett Packard Labs²

University of California, Santa Cruz³, Qualcomm Inc.⁴, Huawei Technologies Inc.⁵
{shuangchenli, yuanxie}@ece.ucsb.edu¹

Pinatubo: RowClone and Bitwise Ops in PCM

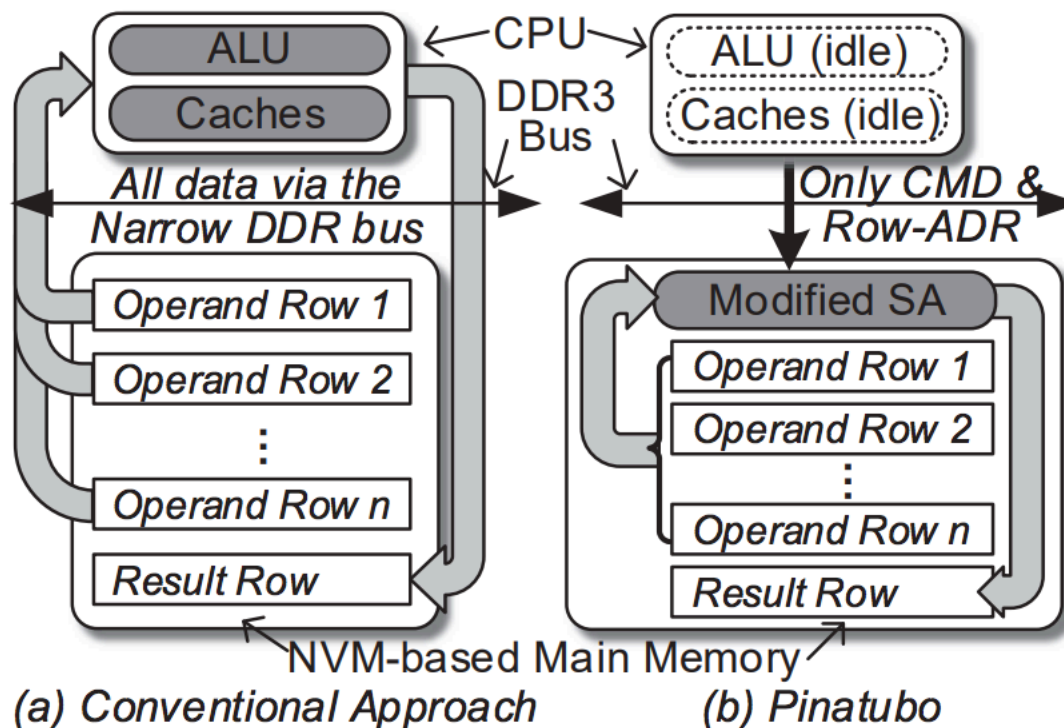


Figure 2: Overview: (a) Computing-centric approach, moving tons of data to CPU and write back. (b) The proposed Pinatubo architecture, performs n -row bitwise operations inside NVM in one step.

Other Examples of “Why Change? It’s Working OK!”

Mindset Issues Are Everywhere

- “Why Change? It’s Working OK!” mindset limits progress
- There are many such examples in real life
- Examples of Bandwidth Waste in Real Life
- Examples of Latency and Queueing Delays in Real Life
- Example of Where to Build a Bridge over a River

Suggestion to Researchers: Principle: Passion

Follow Your Passion
**(Do not get derailed
by naysayers)**

Suggestion to Researchers: Principle: Resilience

Be Resilient

Principle: Learning and Scholarship

Focus on
learning and scholarship

Principle: Learning and Scholarship

The quality of your work
defines your impact

An Interview on Research and Education

- Computing Research and Education (@ ISCA 2019)
 - https://www.youtube.com/watch?v=8ffSEKZhmvo&list=PL5Q2soXY2Zi_4oP9LdL3cc8G6NIjD2Ydz

- Maurice Wilkes Award Speech (10 minutes)
 - https://www.youtube.com/watch?v=tcQ3zZ3JpuA&list=PL5Q2soXY2Zi8D_5MGV6EnXEJHnV2YFBJI&index=15

More Thoughts and Suggestions

- Onur Mutlu,
"Some Reflections (on DRAM)"
*Award Speech for ACM SIGARCH Maurice Wilkes Award, at the **ISCA** Awards Ceremony, Phoenix, AZ, USA, 25 June 2019.*
[[Slides \(pptx\)](#)] [[pdf](#)]
[[Video of Award Acceptance Speech \(Youtube; 10 minutes\)](#)] [[Youku; 13 minutes](#)]
[[Video of Interview after Award Acceptance \(Youtube; 1 hour 6 minutes\)](#)] [[Youku; 1 hour 6 minutes](#)]
[[News Article on "ACM SIGARCH Maurice Wilkes Award goes to Prof. Onur Mutlu"](#)]

- Onur Mutlu,
"How to Build an Impactful Research Group"
*57th Design Automation Conference Early Career Workshop (**DAC**), Virtual, 19 July 2020.*
[[Slides \(pptx\)](#)] [[pdf](#)]

Computer Architecture

Lecture 6: Computation in Memory

Prof. Onur Mutlu

ETH Zürich

Fall 2020

8 October 2020