Four Key Current Directions

- Fundamentally **Secure/Reliable/Safe** Architectures

- Fundamentally **Energy-Efficient** Architectures
  - Memory-centric (Data-centric) Architectures

- Fundamentally **Low-Latency and Predictable** Architectures

- Architectures for **AI/ML, Genomics, Medicine, Health, ...**
Solving the Hardest Problems

Source: https://farm9.staticflickr.com/8571/16376102935_8628150df8_o.jpg
Solving the Hardest Problems
Solving the Hardest Problems

Source: Jane Ades, NHGRI
Solving the Hardest Problems
Solving the Hardest Problems

Source: http://spectrum.ieee.org/image/MjYzMzAyMg.jpeg
Maslow’s Hierarchy of Needs, A Third Time


Source: https://www.simplypsychology.org/maslow.html
Challenge and Opportunity for Future

Fundamentally Low-Latency Computing Architectures
More Motivation for Low Latency

- Watch Satya’s (CMU) keynote talk at SYSTOR 2020
  - Edge Computing: A New Disruptive Force
More Motivation for Low Latency

- Watch Satya’s (CMU) keynote talk at SYSTOR 2020
- Edge Computing: A New Disruptive Force

**Value Proposition of Edge Computing**

What is the edge doing for you?

1. **Edge analytics in IoT**
   - "Scalable real-time sensor analytics"

2. **Highly responsive cloud-like services**
   - "New applications and microservices"

3. **Exposure firewall in the IoT**
   - "Crossing the IoT Chasm"

4. **Mask disruption of cloud services**
   - "Disconnected operation for cloud services"

5. **Honor data export restrictions**
   - "In-country storage and processing"

**Value Propositions**

- **Bandwidth**
  - (peak and average)

- **Latency**
  - (mean and tail)

- **Privacy**
  - (control of sensor data)

- **Availability**
  - (UPS for cloud)

- **Provenance**
  - (bring processing to the data)
More Motivation for Low Latency

- Watch Satya’s (CMU) keynote talk at SYSTOR 2020
- Edge Computing: A New Disruptive Force

Human Cognition is Amazing

Fast, accurate and robust

- face detection under hostile conditions < 700 ms
  (low lighting, distorted optics)
- face recognition 370 ms – 620 ms
- is this sound from a human? 4 ms
- VR head tracking < 16 ms

To be “superhuman” we need to beat these speeds
Leave time for additional software processing (e.g. database lookup) to add value to user
Future of Genome Sequencing & Analysis

MinION from ONT

SmidgION from ONT

Recall Our Dream (from 2007)

- An embedded device that can perform comprehensive genome analysis in real time (within a minute)

- Still a long ways to go
  - Energy efficiency
  - Performance (latency)
  - Security
  - **Huge memory bottleneck**
Data-Centric (Memory-Centric) Architectures
Data-Centric Architectures: Properties

- **Process data where it resides** *(where it makes sense)*
  - Processing in and near memory structures

- **Low-latency & low-energy data access**
  - Low latency memory
  - Low energy memory

- **Low-cost data storage & processing**
  - High capacity memory at low cost: hybrid memory, compression

- **Intelligent data management**
  - Intelligent controllers handling robustness, security, cost, scaling
Low-Latency & Low-Energy Data Access
Memory Latency: Fundamental Tradeoffs
Review: Memory Latency Lags Behind

Memory latency remains almost constant
A Closer Look …

![Graph showing DRAM latency trends over time](image)

**Figure 1:** DRAM latency trends over time [20, 21, 23, 51].

Chang+, "**Understanding Latency Variation in Modern DRAM Chips: Experimental Characterization, Analysis, and Optimization**", SIGMETRICS 2016.
DRAM Latency Is Critical for Performance

**In-memory Databases**
[Mao+, EuroSys’12; Clapp+ (Intel), IISWC’15]

**Graph/Tree Processing**
[Xu+, IISWC’12; Umuroglu+, FPL’15]

**In-Memory Data Analytics**
[Clapp+ (Intel), IISWC’15; Awan+, BDCloud’15]

**Datacenter Workloads**
[Kanev+ (Google), ISCA’15]
DRAM Latency Is Critical for Performance

Long memory latency $\rightarrow$ performance bottleneck

**In-memory Databases**

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

**Graph/Tree Processing**

[Kanev+ (Google), ISCA’15]

**In-Memory Data Analytics**

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

**Datacenter Workloads**

[Kanev+ (Google), ISCA’15]
New DRAM Types Increase Latency!


[Abstract]

---

Demystifying Complex Workload–DRAM Interactions: An Experimental Study

Saugata Ghose† Tianshi Li† Nastaran Hajinazar‡†
Damla Senol Cali† Onur Mutlu§†

†Carnegie Mellon University ‡Simon Fraser University §ETH Zürich
Why Study Workload–DRAM Interactions?

- Manufacturers are developing many new types of DRAM
  - DRAM limits performance, energy improvements: new types may overcome some limitations
  - Memory systems now serve a very diverse set of applications: can no longer take a one-size-fits-all approach

- So which DRAM type works best with which application?
  - Difficult to understand intuitively due to the complexity of the interaction
  - Can’t be tested methodically on real systems: new type needs a new CPU

- We perform a wide-ranging experimental study to uncover the combined behavior of workloads and DRAM types
  - 115 prevalent/emerging applications and multiprogrammed workloads
  - 9 modern DRAM types: DDR3, DDR4, GDDR5, HBM, HMC, LPDDR3, LPDDR4, Wide I/O, Wide I/O 2
## Modern DRAM Types: Comparison to DDR3

<table>
<thead>
<tr>
<th>DRAM Type</th>
<th>Banks per Rank</th>
<th>Bank Groups</th>
<th>3D-Stacked</th>
<th>Low-Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDR3</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DDR4</td>
<td>16</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>GDDR5</td>
<td>16</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>HBM High-Bandwidth Memory</td>
<td>16</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>HMC Hybrid Memory Cube</td>
<td>256</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Wide I/O</td>
<td>4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Wide I/O 2</td>
<td>8</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>LPDDR3</td>
<td>8</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>LPDDR4</td>
<td>16</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

- **Bank groups**
  - Bank Group
  - Bank
  - Bank
  - memory channel
- **3D-stacked DRAM**
  - high bandwidth with Through-Silicon Vias (TSVs)
  - Memory Layers
  - dedicated Logic Layer

- **increased latency**
- **increased area/power**
- **narrower rows, higher latency**
4. Need for Lower Access Latency: Performance

- New DRAM types often increase access latency in order to provide more banks, higher throughput
- Many applications can’t make up for the increased latency
  - Especially true of common OS routines (e.g., file I/O, process forking)

A variety of desktop/scientific, server/cloud, GPGPU applications

Several applications don’t benefit from more parallelism
Key Takeaways

1. DRAM latency remains a critical bottleneck for many applications

2. Bank parallelism is not fully utilized by a wide variety of our applications

3. Spatial locality continues to provide significant performance benefits if it is exploited by the memory subsystem

4. For some classes of applications, low-power memory can provide energy savings without sacrificing significant performance
Conclusion

- Manufacturers are developing many new types of DRAM
  - DRAM limits performance, energy improvements: new types may overcome some limitations
  - Memory systems now serve a very diverse set of applications: can no longer take a one-size-fits-all approach
  - Difficult to intuitively determine which DRAM–workload pair works best

- We perform a wide-ranging experimental study to uncover the combined behavior of workloads, DRAM types
  - 115 prevalent/emerging applications and multiprogrammed workloads
  - 9 modern DRAM types

- 12 key observations on DRAM–workload behavior

Open-source tools: https://github.com/CMU-SAFARI/ramulator
Full paper: https://arxiv.org/pdf/1902.07609
New DRAM Types Increase Latency!

- Saugata Ghose, Tianshi Li, Nastaran Hajinazar, Damla Senol Cali, and Onur Mutlu, "Demystifying Workload–DRAM Interactions: An Experimental Study"


[Preliminary arXiv Version]

[Abstract]

[Slides (pptx) (pdf)]

[MemBen Benchmark Suite]

[Source Code for GPGPUSim-Ramulator]

[Source Code for Ramulator modeling Hybrid Memory Cube (HMC)]
The Memory Latency Problem

- High memory latency is a significant limiter of system performance and energy-efficiency.

- It is becoming increasingly so with higher memory contention in multi-core and heterogeneous architectures:
  - Exacerbating the bandwidth need
  - Exacerbating the QoS problem

- It increases processor design complexity due to the mechanisms incorporated to tolerate memory latency.
Retrospective: Conventional Latency Tolerance Techniques

- Caching [initially by Wilkes, 1965]
  - Widely used, simple, effective, but inefficient, passive
  - Not all applications/phases exhibit temporal or spatial locality

- Prefetching [initially in IBM 360/91, 1967]
  - Works well for regular memory access patterns
  - Prefetching irregular access patterns is difficult, inaccurate, and hardware-intensive

- Multithreading [initially in CDC 6600, 1964]
  - Works well if there are multiple threads
  - Improving single thread performance using multithreading hardware is an ongoing research effort

- Out-of-order execution [initially by Tomasulo, 1967]
  - Tolerates cache misses that cannot be prefetched
  - Requires extensive hardware resources for tolerating long latencies
Retrospective: Conventional Latency Tolerance Techniques

- **Caching** [initially by Wilkes, 1965]
  - Widely used, simple, effective, but inefficient, passive
  - Not all applications/phases exhibit temporal or spatial locality

- **Prefetching** [initially in IBM 360/91, 1967]
  - Works well for regular memory access patterns
  - Prefetching irregular access patterns is difficult, inaccurate, and hardware-intensive

- **Multithreading** [initially in CDC 6600, 1964]
  - Works well if there are multiple threads
  - Improving single thread performance using multithreading hardware is an ongoing research effort

- **Out-of-order execution** [initially by Tomasulo, 1967]
  - Tolerates cache misses that cannot be prefetched
  - Requires extensive hardware resources for tolerating long latencies

None of These Fundamentally Reduce Memory Latency

ongoing research effort
Runahead Execution
Perfect Caches:
Load 1 Hit  Load 2 Hit

Small OoO Instruction Window:
Load 1 Miss  Load 2 Miss

Runahead:
Load 1 Miss  Load 2 Miss  Load 1 Hit  Load 2 Hit

Saved Cycles
Effect of Runahead in Sun ROCK

- Shailender Chaudhry talk, Aug 2008.

![Graph showing normalized IPC vs L2 cache size with notes on runahead benefits.](image-url)
More on Runahead Execution

- Onur Mutlu, Jared Stark, Chris Wilkerson, and Yale N. Patt, "Runahead Execution: An Alternative to Very Large Instruction Windows for Out-of-order Processors"


One of the 15 computer architecture papers of 2003 selected as Top Picks by IEEE Micro.
More on Runahead Execution (Short)

Runahead Readings

- Required

- Recommended
More on Runahead Execution (I)

Computer Architecture - Lecture 19a: Execution-Based Prefetching (ETH Zürich, Fall 2020)

395 views • Nov 29, 2020

https://www.youtube.com/watch?v=zPewo6laJ_8&list=PL5Q2soXY2Zl9xidyIgxBxUz7xRPS-wisBN&index=34
More on Runahead Execution (II)

Runahead Execution in NVIDIA Denver

Reducing the effects of long cache-miss penalties has been a major focus of the microarchitecture, using techniques like prefetching and run-ahead. An aggressive hardware prefetcher implementation detects L2 cache requests and tracks up to 32 streams, each with complex stride patterns.

Run-ahead uses the idle time that a CPU spends waiting on a long latency operation to discover cache and DTLB misses further down the instruction stream and generates prefetch requests for these misses. These prefetch requests warm up the data cache and DTLB well before the actual execution of the instructions that require the data. Run-ahead complements the hardware prefetcher because it's better at prefetching non-stripped streams, and it trains the hardware prefetcher faster than normal execution to yield a combined benefit of 13 percent on SPECInt2000 and up to 60 percent on SPECfp2000.


Gwnnap, "NVIDIA's First CPU is a Winner," MPR 2014.

Onur Mutlu - Runahead Execution: A Short Retrospective (HPCA Test of Time Award Talk @ HPCA 2021)

https://www.youtube.com/watch?v=KFCOecRQTlc
Retrospective: Conventional Latency Tolerance Techniques

- Caching [initially by Wilkes, 1965]
  - Widely used, simple, effective, but inefficient, passive
  - Not all applications/phases exhibit temporal or spatial locality

- Prefetching [initially in IBM 360/91, 1967]
  - Works well for regular memory access patterns
  - Prefetching irregular access patterns is difficult, inaccurate, and hardware-intensive

- Multithreading [initially in CDC 6600, 1964]
  - Works well if there are multiple threads
  - Improving single thread performance using multithreading hardware is an ongoing research effort

- Out-of-order execution [initially by Tomasulo, 1967]
  - Tolerates cache misses that cannot be prefetched
  - Requires extensive hardware resources for tolerating long latencies

None of These Fundamentally Reduce Memory Latency
Two Major Sources of Latency Inefficiency

- Modern DRAM is not designed for low latency
  - Main focus is cost-per-bit (capacity)

- Modern DRAM latency is determined by worst case conditions and worst case devices
  - Much of memory latency is unnecessary

Our Goal: Reduce Memory Latency at the Source of the Problem
Truly Reducing Memory Latency
What Causes the Long Memory Latency?
Why the Long Memory Latency?

- **Reason 1: Design of DRAM Micro-architecture**
  - Goal: Maximize capacity/area, not minimize latency

- **Reason 2: “One size fits all” approach to latency specification**
  - Same latency parameters for all temperatures
  - Same latency parameters for all DRAM chips
  - Same latency parameters for all parts of a DRAM chip
  - Same latency parameters for all supply voltage levels
  - Same latency parameters for all application data
  - ...
Brief Review:  
Inside A DRAM Chip
DRAM Module and Chip
Goals

• Cost
• Latency
• Bandwidth
• Parallelism
• Power
• Energy
• Reliability
• ...

...
DRAM Chip
Sense Amplifier

enable

top

bottom

Inverter
Sense Amplifier – Two Stable States

Logical “1”

Logical “0”
Sense Amplifier Operation

\[ V_T > V_B \]
DRAM Cell – Capacitor

Empty State
Logical “0”

Fully Charged State
Logical “1”

1 Small – Cannot drive circuits
2 Reading destroys the state
Capacitor to Sense Amplifier
DRAM Cell Operation
DRAM Subarray – Building Block for DRAM Chip

- Row Decoder
- Cell Array
- Array of Sense Amplifiers (Row Buffer) 8Kb
- Cell Array
DRAM Bank

Row Decoder

Cell Array

Array of Sense Amplifiers (8Kb)

Cell Array

Address

Array of Sense Amplifiers

Cell Array

Bank I/O (64b)

Data
DRAM Chip

Shared internal bus

Memory channel - 8bits
DRAM Operation

1. ACTIVATE Row
2. READ/WRITE Column
3. PRECHARGE

Row Address

Column Address

Bank I/O

Row Decoder

Cell Array

Array of Sense Amplifiers

Data
More on DRAM Operation: Section 2

  [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
Why the Long Memory Latency?

- **Reason 1: Design of DRAM Micro-architecture**
  - Goal: Maximize capacity/area, not minimize latency

- **Reason 2: “One size fits all” approach to latency specification**
  - Same latency parameters for all temperatures
  - Same latency parameters for all DRAM chips
  - Same latency parameters for all parts of a DRAM chip
  - Same latency parameters for all supply voltage levels
  - Same latency parameters for all application data
  - ...

SAFARI
Tiered Latency DRAM
What Causes the Long Latency?

**DRAM Chip**

\[
\text{DRAM Latency} = \text{Subarray Latency} + \text{I/O Latency}
\]

*Dominant* Subarray 

---

63
Why is the Subarray So Slow?

- **Long bitline**
  - Amortizes sense amplifier cost → Small area
  - Large bitline capacitance → High latency & power
Trade-Off: Area (Die Size) vs. Latency

Long Bitline

Short Bitline

Faster

Smaller

Trade-Off: Area vs. Latency
Trade-Off: Area (Die Size) vs. Latency

- **Fancy DRAM Short Bitline**
- **Commodity DRAM Long Bitline**

Normalized DRAM Area vs. Latency (ns)

- Cheaper
- Faster

GOAL: 512 cells/bitline
Approximating the Best of Both Worlds

Long Bitline

Small Area

High Latency

Our Proposal

Add Isolation Transistors

Short Bitline

Large Area

Low Latency

Need Isolation

Fast

67
Approximating the Best of Both Worlds

- Long Bitline Tiered-Latency DRAM
- Short Bitline

Small Area
- Small Area
- Large Area
- High Latency

Low Latency

Small area using long bitline

Low Latency
Commodity DRAM vs. TL-DRAM [HPCA 2013]

- DRAM Latency ($t_{RC}$)
- DRAM Power

- DRAM Area Overhead
  
  ~3%: mainly due to the isolation transistors
Trade-Off: Area (Die-Area) vs. Latency

Normalized DRAM Area vs. Latency (ns)

- Cheaper
- Near Segment
- Far Segment

Goal: Cheaper and Faster

Cells/bitline: 32, 64, 128, 256, 512
Leveraging Tiered-Latency DRAM

• TL-DRAM is a *substrate* that can be leveraged by the hardware and/or software

• Many potential uses

1. Use near segment as hardware-managed *inclusive* cache to far segment
2. Use near segment as hardware-managed *exclusive* cache to far segment
3. Profile-based page mapping by operating system
4. Simply replace DRAM with TL-DRAM

Near Segment as Hardware-Managed Cache

• **Challenge 1:** How to efficiently migrate a row between segments?

• **Challenge 2:** How to efficiently manage the cache?
Inter-Segment Migration

- **Goal:** Migrate source row into destination row
- **Naïve way:** Memory controller reads the source row byte by byte and writes to destination row byte by byte → High latency
Inter-Segment Migration

• Our way:
  – Source and destination cells *share bitlines*
  – Transfer data from source to destination across *shared bitlines* concurrently

![Diagram of Inter-Segment Migration with Far and Near Segments, Isolation Transistors, and Sense Amplifiers]
Inter-Segment Migration

• Our way:
  – Source and destination cells share bitlines
  – Transfer data from source to destination using shared bitlines concurrently

Migration is overlapped with source row access
Additional ~4ns over row access latency

Step 1: Activate source row
Step 2: Activate destination row to connect cell and bitline
Near Segment as Hardware-Managed Cache

- **Challenge 1:** How to efficiently migrate a row between segments?
- **Challenge 2:** How to efficiently manage the cache?
Using near segment as a cache improves performance and reduces power consumption

By adjusting the near segment length, we can trade off cache capacity for cache latency.
More on TL-DRAM

- Donghyuk Lee, Yoongu Kim, Vivek Seshadri, Jamie Liu, Lavanya Subramanian, and Onur Mutlu,
  "Tiered-Latency DRAM: A Low Latency and Low Cost DRAM Architecture"
  Proceedings of the 19th International Symposium on High-Performance Computer Architecture (HPCA), Shenzhen, China, February 2013. Slides (pptx)

Tiered-Latency DRAM: A Low Latency and Low Cost DRAM Architecture
Donghyuk Lee  Yoongu Kim  Vivek Seshadri  Jamie Liu  Lavanya Subramanian  Onur Mutlu
Carnegie Mellon University
LISA: Low-Cost Inter-Linked Subarrays [HPCA 2016]
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
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
New DRAM Command to Use LISA

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

RBM transfers an entire row b/w subarrays
RBM Analysis

• The range of RBM depends on the DRAM design
  – Multiple RBMs to move data across > 3 subarrays

• Validated with SPICE using worst-case cells
  – NCSU FreePDK 45nm library

• 4KB data in 8ns (w/ 60% guardband)
  → 500 GB/s, 26x bandwidth of a DDR4-2400 channel

• 0.8% DRAM chip area overhead [O+ ISCA’14]
1. Rapid Inter-Subarray Copying (RISC)

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

1. **Activate** src row

2. **RBM** SA1 $\rightarrow$ SA2

Reduces row-copy latency by 9.2x, DRAM energy by 48.1x
2. Variable Latency DRAM (VILLA)

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

Long Bitline (DDRx)

Short Bitline (RLDRAM)

Shorter bitlines $\rightarrow$ faster activate and precharge time

High area overhead: $>40\%$
2. Variable Latency DRAM (VILLA)

- **Key idea**: Reduce access latency of hot data via a heterogeneous DRAM design [Lee+ HPCA’13, Son+ ISCA’13]
- **VILLA**: Add fast subarrays as a cache in each bank

**Challenge**: VILLA cache requires frequent movement of data rows

Reduces hot data access latency by 2.2x at only 1.6% area overhead
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 (43% guardband)
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"
[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
CROW: The Copy Row Substrate
[ISCA 2019]
Challenges of DRAM Scaling

1. access latency
2. refresh overhead
3. exposure to vulnerabilities
Conventional DRAM

DRAM Subarray

row decoder

sense amplifier
Copy Row DRAM (CROW)

Row copy

Multiple row activation

DRAM Subarray

regular rows

copy rows

sense amplifier
Use Cases of CROW

- **CROW-cache**
  - reduces access latency

- **CROW-ref**
  - reduces DRAM refresh overhead

- A mechanism for protecting against RowHammer

Use Cases of CROW

- **CROW-cache**
  - reduces access latency

- **CROW-ref**
  - reduces DRAM refresh overhead

- A mechanism for protecting against RowHammer
Key Results

**CROW-cache + CROW-ref**
- 20% speedup
- 22% less DRAM energy

**Hardware Overhead**
- 0.5% DRAM chip area
- 1.6% DRAM capacity
- 11.3 KiB memory controller storage
More on CROW

- Hasan Hassan, Minesh Patel, Jeremie S. Kim, A. Giray Yaglikci, Nandita Vijaykumar, Nika Mansouighiasi, Saugata Ghose, and Onur Mutlu,

"CROW: A Low-Cost Substrate for Improving DRAM Performance, Energy Efficiency, and Reliability"


[Slides (pptx) (pdf)]
[Lightning Talk Slides (pptx) (pdf)]
[Poster (pptx) (pdf)]
[Lightning Talk Video (3 minutes)]
[Full Talk Video (16 minutes)]
[Source Code for CROW (Ramulator and Circuit Modeling)]

CROW: A Low-Cost Substrate for Improving DRAM Performance, Energy Efficiency, and Reliability

Hasan Hassan† Minesh Patel† Jeremie S. Kim†§ A. Giray Yaglikci†
Nandita Vijaykumar†§ Nika Mansouri Ghiasi† Saugata Ghose§ Onur Mutlu†§

†ETH Zürich §Carnegie Mellon University
CLR-DRAM: Capacity-Latency Reconfigurability

- Haocong Luo, Taha Shahroodi, Hasan Hassan, Minesh Patel, A. Giray Yaglikci, Lois Orosa, Jisung Park, and Onur Mutlu,

"CLR-DRAM: A Low-Cost DRAM Architecture Enabling Dynamic Capacity-Latency Trade-Off"


[Slides (pptx) (pdf)]
[Lightning Talk Slides (pptx) (pdf)]
[Talk Video (20 minutes)]
[Lightning Talk Video (3 minutes)]
CLR-DRAM: Capacity-Latency Reconfigurable DRAM [ISCA 2020]
CLR-DRAM: A Low-Cost DRAM Architecture Enabling Dynamic Capacity-Latency Trade-off

Haocong Luo  Taha Shahroodi  Hasan Hassan  Minesh Patel
A. Giray Yaglıkçı  Lois Orosa  Jisung Park  Onur Mutlu
Motivation & Goal

- Workloads and systems have **varying** main memory capacity and latency demands.
- Existing commodity DRAM makes **static** capacity-latency trade-off at **design time**.
- Systems miss opportunities to improve performance by adapting to changes in main memory capacity and latency demands.

**Goal**: Design a low-cost DRAM architecture that can be **dynamically** configured to have high capacity or low latency at a fine granularity (i.e., at the granularity of a row).
CLR-DRAM (Capacity-Latency-Reconfigurable DRAM)

- **CLR-DRAM (Capacity-Latency-Reconfigurable DRAM):**
  - A low cost DRAM architecture that enables a single DRAM row to *dynamically* switch between max-capacity mode or high-performance mode.

- **Key Idea:**
  *Dynamically* configure the connections between DRAM cells and sense amplifiers in the density-optimized open-bitline architecture.
CLR-DRAM (Capacity-Latency-Reconfigurable DRAM)

- Max-capacity mode

- High-performance mode

The same storage capacity as the conventional open-bitline architecture

Reduced latency and refresh overhead via coupled cell/SA operation

mimics the cell-to-SA connections as in the open-bitline architecture

coupled cells

coupled sense amplifiers

SA1

SA2

Type 1

Type 2

Type 1

Type 2

Type 1

Type 2

Type 1

Type 2

Type 1

Type 2

Type 1

Type 2

SA

SA
Key Results

- **DRAM Latency Reduction:**
  - Activation latency (tRCD) by **60.1%**
  - Restoration latency (tRAS) by **64.2%**
  - Precharge latency (tRP) by **46.4%**
  - Write-recovery latency (tWR) by **35.2%**

- **System-level Benefits:**
  - Performance improvement: **18.6%**
  - DRAM energy reduction: **29.7%**
  - DRAM refresh energy reduction: **66.1%**

We hope that CLR-DRAM can be exploited to develop more flexible systems that can adapt to the diverse and changing DRAM capacity and latency demands of workloads.
More on CLR-DRAM

  
  [Slides (pptx) (pdf)]
  [Lightning Talk Slides (pptx) (pdf)]
  [Talk Video (20 minutes)]
  [Lightning Talk Video (3 minutes)]
SALP: Reducing DRAM Bank Conflict Impact

Kim, Seshadri, Lee, Liu, Mutlu
*A Case for Exploiting Subarray-Level Parallelism (SALP) in DRAM*
ISCA 2012.
SALP: Problem, Goal, Observations

- Problem: Bank conflicts are costly for performance and energy
  - serialized requests, wasted energy (thrashing of row buffer, busy wait)
- Goal: Reduce bank conflicts without adding more banks (low cost)
- Observation 1: A DRAM bank is divided into subarrays and each subarray has its own local row buffer
Observation 2: Subarrays are mostly independent
- Except when sharing global structures to reduce cost

Key Idea of SALP: Minimally reduce sharing of global structures
- Reduce the sharing of ...
  - Global decoder → Enables almost parallel access to subarrays
  - Global row buffer → Utilizes multiple local row buffers
SALP: Reduce Sharing of Global Decoder

Instead of a global latch, have *per-subarray latches*
SALP: Reduce Sharing of Global Row-Buffer

Selectively connect local row-buffers to global row-buffer using a **Designated** single-bit latch.
SALP: Baseline Bank Organization

Global Decoder

Latch

Global bitlines

Local row-buffer

Global row-buffer

Local row-buffer
Overhead of SALP in DRAM chip: 0.15%

1. Global latch → per-subarray local latches
2. Designated bit latches and wire to selectively enable a subarray
SALP: Results

- Wide variety of systems with different #channels, banks, ranks, subarrays
- Server, streaming, random-access, SPEC workloads
- **Dynamic DRAM energy reduction:** 19%
  - DRAM row hit rate improvement: 13%
- **System performance improvement:** 17%
  - Within 3% of ideal (all independent banks)
- **DRAM die area overhead:** 0.15%
  - vs. 36% overhead of independent banks
More on SALP

- Yoongu Kim, Vivek Seshadri, Donghyuk Lee, Jamie Liu, and Onur Mutlu,
  "A Case for Exploiting Subarray-Level Parallelism (SALP) in DRAM"
Proceedings of the 39th International Symposium on Computer Architecture (ISCA), Portland, OR, June 2012. Slides (pptx)
More on SALP

DRAM Process Scaling Challenges

- Refresh
  - Difficult to build high-aspect ratio cell capacitors decreasing cell capacitance

Co-Architecting Controllers and DRAM to Enhance DRAM Process Scaling

Uksong Kang, Hak-soo Yu, Churoo Park, *Hongzhong Zheng, **John Halbert, **Kuljit Bains, SeongJin Jang, and Joo Sun Choi

Samsung Electronics, Hwasung, Korea / *Samsung Electronics, San Jose / **Intel
More on SALP

Sub-array Level Parallelism with tWR Relaxation

- **tWR relaxation**
  - Relaxing tWR results in DRAM yield improvement but can degrade performance requiring new compensating features
  - By increasing tWR 5X (from 15ns to 75ns), fail bit counts are expected to reduce by 1 to 2 orders of magnitudes

- **Sub-array level parallelism (SALP)**
  - Allows a page in another sub-array in the same bank to be opened in parallel with the currently activated sub-array
  - Results in performance gain by increasing the row access parallelism within a bank
  ⇒ Used to compensate for the performance loss caused by tWR relaxation

More on SALP

Performance Impact of SALP and tWR relaxation

- Performance simulations run for various workloads when tWR is relaxed by 2X and 3X, and when SALP is applied with 2 sub-banks

- Results show that performance is reduced by ~5% and ~2% in average if tWR is relaxed by 3X and 2X, respectively

- Results also show that performance is compensated, and even improved to up to ~3% in average when SALP is applied, even with tWR relaxed by 3X
Why the Long Memory Latency?

- **Reason 1:** Design of DRAM Micro-architecture
  - Goal: Maximize capacity/area, not minimize latency

- **Reason 2:** “One size fits all” approach to latency specification
  - Same latency parameters for all **temperatures**
  - Same latency parameters for all **DRAM chips**
  - Same latency parameters for all **parts of a DRAM chip**
  - Same latency parameters for all **supply voltage levels**
  - Same latency parameters for all **application data**
  - ...
Tackling the Fixed Latency Mindset

- Reliable operation latency is actually very heterogeneous
  - Across temperatures, chips, parts of a chip, voltage levels, ...

- Idea: Dynamically find out and use the lowest latency one can reliably access a memory location with
  - Adaptive-Latency DRAM [HPCA 2015]
  - Flexible-Latency DRAM [SIGMETRICS 2016]
  - Design-Induced Variation-Aware DRAM [SIGMETRICS 2017]
  - Voltron [SIGMETRICS 2017]
  - DRAM Latency PUF [HPCA 2018]
  - Solar DRAM [ICCD 2018]
  - DRAM Latency True Random Number Generator [HPCA 2019]
  - ...

- We would like to find sources of latency heterogeneity and exploit them to minimize latency (or create other benefits)
Latency Variation in Memory Chips

Heterogeneous manufacturing & operating conditions → latency variation in timing parameters

DRAM A  DRAM B  DRAM C

Low  →  High

DRAM Latency

Slow cells
Why is Latency High?

• **DRAM latency**: Delay as specified in DRAM standards
  – Doesn’t reflect true DRAM device latency
• Imperfect manufacturing process → latency variation
• **High standard latency** chosen to increase yield
What Causes the Long Memory Latency?

- **Conservative timing margins!**

- DRAM timing parameters are set to cover the worst case

- **Worst-case temperatures**
  - 85 degrees vs. common-case
  - to enable a wide range of operating conditions

- **Worst-case devices**
  - DRAM cell with smallest charge across any acceptable device
  - to tolerate process variation at acceptable yield

- **This leads to large timing margins for the common case**
Understanding and Exploiting Variation in DRAM Latency
DRAM Stores Data as Charge

Three steps of charge movement

1. Sensing
2. Restore
3. Precharge
Why does DRAM need the extra timing margin?
Two Reasons for Timing Margin

1. Process Variation
   - DRAM cells are not equal
   - Leads to extra timing margin for a cell that can store a large amount of charge

2. Temperature Dependence
DRAM Cells are Not Equal

**Ideal**

- Same Size
- Same Charge
- Same Latency

**Real**

- Different Size
- Different Charge
- Different Latency

Large variation in cell size
Large variation in charge
Large variation in access latency

Smallest Cell
Largest Cell
Process Variation

Small cell can store small charge

- Small cell capacitance
- High contact resistance
- Slow access transistor

High access latency
Two Reasons for Timing Margin

1. *Process Variation*
   - DRAM cells are not equal
   - Leads to *extra timing margin* for a cell that can store a large amount of charge

2. *Temperature Dependence*
   - DRAM leaks more charge at higher temperature
   - Leads to *extra timing margin* for cells that operate at low temperature
Cells store small charge at high temperature and large charge at low temperature → Large variation in access latency
DRAM Timing Parameters

• **DRAM timing parameters are dictated by the worst-case**
  
  – The smallest cell with the smallest charge in all DRAM products
  
  – Operating at the highest temperature

• **Large timing margin for the common-case**
Idea: Optimize DRAM timing for the common case
- Current temperature
- Current DRAM module

Why would this reduce latency?

- A DRAM cell can store much more charge in the common case (low temperature, strong cell) than in the worst case
- More charge in a DRAM cell
  - Faster sensing, charge restoration, precharging
  - Faster access (read, write, refresh, ...)

Extra Charge $\rightarrow$ Reduced Latency

1. Sensing
   Sense **cells with extra charge** faster
   $\rightarrow$ **Lower sensing latency**

2. Restore
   No need to fully restore **cells with extra charge**
   $\rightarrow$ **Lower restoration latency**

3. Precharge
   No need to fully precharge bitlines for **cells with extra charge**
   $\rightarrow$ **Lower precharge latency**
DRAM Characterization Infrastructure

Temperature Controller

FPGAs

Heater

FPGAs

PC

DRAM Characterization Infrastructure


- **Flexible**
- **Easy to Use (C++ API)**
- **Open-source**

  [github.com/CMU-SAFARI/SoftMC](https://github.com/CMU-SAFARI/SoftMC)
SoftMC: Open Source DRAM Infrastructure

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

SoftMC: A Flexible and Practical Open-Source Infrastructure for Enabling Experimental DRAM Studies

Hasan Hassan¹,²,³  Nandita Vijaykumar³  Samira Khan⁴,³  Saugata Ghose³  Kevin Chang³  Gennady Pekhimenko⁵,³  Donghyuk Lee⁶,³  Oguz Ergin²  Onur Mutlu¹,³

¹ETH Zürich  ²TOBB University of Economics & Technology  ³Carnegie Mellon University  ⁴University of Virginia  ⁵Microsoft Research  ⁶NVIDIA Research
Observation 1. Faster Sensing

Typical DIMM at Low Temperature

More Charge
Strong Charge Flow
Faster Sensing

Timing (tRCD)
17% ↓
No Errors

115 DIMM Characterization

Typical DIMM at Low Temperature
⇒ More charge ⇒ Faster sensing
Observation 2. Reducing Restore Time

Typical DIMM at Low Temperature

- Less Leakage ➔
- Extra Charge
- No Need to Fully Restore Charge

115 DIMM Characterization

Read ($t_{RAS}$)
- 37% ↓

Write ($t_{WR}$)
- 54% ↓
- No Errors

Typical DIMM at lower temperature

➔ More charge ➔ Restore time reduction
AL-DRAM

• **Key idea**
  – Optimize DRAM timing parameters online

• **Two components**
  – DRAM manufacturer provides multiple sets of **reliable DRAM timing parameters** at different temperatures for each DIMM
  – System monitors **DRAM temperature** & uses appropriate DRAM timing parameters

---

DRAM Temperature

- **DRAM temperature measurement**
  - Server cluster: Operates at under 34°C
  - Desktop: Operates at under 50°C
  - **DRAM standard optimized for 85 °C**

DRAM operates at low temperatures in the common-case

- **Previous works – Maintain low DRAM temperature**
  - David+ ICAC 2011
  - Liu+ ISCA 2007
  - Zhu+ Itherm 2008
Latency Reduction Summary of 115 DIMMs

- **Latency reduction for read & write (55°C)**
  - Read Latency: **32.7%**
  - Write Latency: **55.1%**

- **Latency reduction for each timing parameter (55°C)**
  - Sensing: **17.3%**
  - Restore: **37.3%** (read), **54.8%** (write)
  - Precharge: **35.2%**

AL-DRAM: Real System Evaluation

- **System**
  - **CPU**: AMD 4386 (8 Cores, 3.1GHz, 8MB LLC)

---

<table>
<thead>
<tr>
<th>Bits</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>31:30</td>
<td>Reserved.</td>
</tr>
</tbody>
</table>

**D18F2x200_dct[0]_mp[1:0] DDR3 DRAM Timing 0**

Reset: 0F05_0505h. See 2.9.3 [DCT Configuration Registers].

<table>
<thead>
<tr>
<th>Bits</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>29:24</td>
<td><strong>Tras: row active strobe.</strong> Read-write. BIOS: See 2.9.7.5 [SPD ROM-Based Configuration]. Specifies the minimum time in memory clock cycles from an activate command to a precharge command, both to the same chip select bank.</td>
</tr>
<tr>
<td>07h-00h</td>
<td>Reserved</td>
</tr>
<tr>
<td>2Ah-08h</td>
<td>&lt;Tras&gt; clocks</td>
</tr>
<tr>
<td>3Fh-2Bh</td>
<td>Reserved</td>
</tr>
<tr>
<td>23:21</td>
<td>Reserved.</td>
</tr>
</tbody>
</table>

**20:16** **Trp: row precharge time.** Read-write. BIOS: See 2.9.7.5 [SPD ROM-Based Configuration]. Specifies the minimum time in memory clock cycles from a precharge command to an activate command or auto refresh command, both to the same bank.
AL-DRAM improves performance on a real system
AL-DRAM: Multi-Core Evaluation

Performance Improvement

| Workload  | Single Core | Multi Core | Improvement
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>soplex</td>
<td><img src="soplex" alt="Graph Bar" /></td>
<td><img src="multi-core-soplex" alt="Graph Bar" /></td>
<td>14.0%</td>
</tr>
<tr>
<td>mcf</td>
<td><img src="mcf" alt="Graph Bar" /></td>
<td><img src="multi-core-mcf" alt="Graph Bar" /></td>
<td>10.4%</td>
</tr>
<tr>
<td>milc</td>
<td><img src="milc" alt="Graph Bar" /></td>
<td><img src="multi-core-milc" alt="Graph Bar" /></td>
<td>2.9%</td>
</tr>
<tr>
<td>libq</td>
<td><img src="libq" alt="Graph Bar" /></td>
<td><img src="multi-core-libq" alt="Graph Bar" /></td>
<td></td>
</tr>
<tr>
<td>lbm</td>
<td><img src="lbm" alt="Graph Bar" /></td>
<td><img src="multi-core-lbm" alt="Graph Bar" /></td>
<td></td>
</tr>
<tr>
<td>gems</td>
<td><img src="gems" alt="Graph Bar" /></td>
<td><img src="multi-core-gems" alt="Graph Bar" /></td>
<td></td>
</tr>
<tr>
<td>copy</td>
<td><img src="copy" alt="Graph Bar" /></td>
<td><img src="multi-core-copy" alt="Graph Bar" /></td>
<td></td>
</tr>
<tr>
<td>s.cluster</td>
<td><img src="s.cluster" alt="Graph Bar" /></td>
<td><img src="multi-core-s.cluster" alt="Graph Bar" /></td>
<td></td>
</tr>
<tr>
<td>gups</td>
<td><img src="gups" alt="Graph Bar" /></td>
<td><img src="multi-core-gups" alt="Graph Bar" /></td>
<td></td>
</tr>
</tbody>
</table>

Average Improvement

- Non-intensive workloads: 14.0%
- Intensive workloads: 10.4%
- All workloads: 2.9%

AL-DRAM provides higher performance for multi-programmed & multi-threaded workloads.
Reducing Latency Also Reduces Energy

- AL-DRAM reduces DRAM power consumption by 5.8%
- Major reason: reduction in row activation time
AL-DRAM: Advantages & Disadvantages

■ Advantages
  + Simple mechanism to reduce latency
  + Significant system performance and energy benefits
    + Benefits higher at low temperature
  + Low cost, low complexity

■ Disadvantages
  - Need to determine reliable operating latencies for different temperatures and different DIMMs → higher testing cost
    (might not be that difficult for low temperatures)
More on AL-DRAM

- Donghyuk Lee, Yoongu Kim, Gennady Pekhimenko, Samira Khan, Vivek Seshadri, Kevin Chang, and Onur Mutlu, "Adaptive-Latency DRAM: Optimizing DRAM Timing for the Common-Case"

[Slides (pptx) (pdf)] [Full data sets]
Different Types of Latency Variation

- AL-DRAM exploits latency variation
  - Across time (different temperatures)
  - Across chips

- Is there also latency variation within a chip?
  - Across different parts of a chip
We Stopped Here in Lecture. Remaining Slides are For Future.
Why the Long Memory Latency?

- **Reason 1: Design of DRAM Micro-architecture**
  - Goal: Maximize capacity/area, not minimize latency

- **Reason 2: “One size fits all” approach to latency specification**
  - Same latency parameters for all temperatures
  - Same latency parameters for all DRAM chips
  - **Same latency parameters for all parts of a DRAM chip**
  - Same latency parameters for all supply voltage levels
  - Same latency parameters for all application data
  - ...

SAFARI
Variation in Activation Errors

Results from 7500 rounds over 240 chips

Modern DRAM chips exhibit significant variation in activation latency
Spatial Locality of Activation Errors

Activation errors are concentrated at certain columns of cells
Mechanism to Reduce DRAM Latency

- **Observation:** DRAM timing errors (slow DRAM cells) are concentrated on certain regions

- **Flexible-Latency (FLY) DRAM**
  - A software-transparent design that reduces latency

- **Key idea:**
  1) Divide memory into regions of different latencies
  2) *Memory controller:* Use lower latency for regions without slow cells; higher latency for other regions

Chang+, “*Understanding Latency Variation in Modern DRAM Chips: Experimental Characterization, Analysis, and Optimization*,” SIGMETRICS 2016.
FLY-DRAM Configurations

Results

FLY-DRAM improves performance by exploiting spatial latency variation in DRAM

FLY-DRAM: Advantages & Disadvantages

- **Advantages**
  + Reduces latency significantly
    + Exploits significant within-chip latency variation

- **Disadvantages**
  - Need to determine reliable operating latencies for different parts of a chip → higher testing cost
  - More complicated controller
Analysis of Latency Variation in DRAM Chips

Kevin Chang, Abhijith Kashyap, Hasan Hassan, Samira Khan, Kevin Hsieh, Donghyuk Lee, Saugata Ghose, Gennady Pekhimenko, Tianshi Li, and Onur Mutlu,
"Understanding Latency Variation in Modern DRAM Chips: Experimental Characterization, Analysis, and Optimization"
[Slides (pptx) (pdf)]
[Source Code]
Putting It All Together: Solar-DRAM
Solar-DRAM: Putting It Together

  [Slides (pptx) (pdf)]
  [Talk Video (16 minutes)]

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

Jeremie S. Kim‡§ Minesh Patel§ Hasan Hassan§ Onur Mutlu§†
‡Carnegie Mellon University §ETH Zürich

SAFARI
More on Solar DRAM

Spatial Distribution of Failures

How are activation failures spatially distributed in DRAM?

- Subarray Edge

Activation failures are highly constrained to local bitlines (i.e., subarrays)

Solar-DRAM: Reducing DRAM Access Latency by Exploiting the Variation in Local Bitlines - ICCD 2018

https://www.youtube.com/watch?v=WPMdIxmKrU
Why Is There Spatial Latency Variation Within a Chip?
What Is Design-Induced Variation?

Systematic variation in cell access times caused by the physical organization of DRAM
DIVA Online Profiling

Design-Induced-Variation-Aware

Profile only slow regions to determine min. latency

⇒ Dynamic & low cost latency optimization
DIVA Online Profiling

Design-Induced-Variation-Aware

slow cells
process variation
random error

inherently slow
design-induced variation
localized error

error-correcting code

online profiling

Combine error-correcting codes & online profiling → Reliably reduce DRAM latency
DIVA-DRAM reduces latency more aggressively and uses ECC to correct random slow cells.
DIVA-DRAM: Advantages & Disadvantages

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

- **Disadvantages**
  - Requires knowledge of inherently-slow regions
  - Requires ECC (Error Correcting Codes)
  - Imposes overhead during runtime profiling
Design-Induced Latency Variation in Modern DRAM Chips: Characterization, Analysis, and Latency Reduction Mechanisms

Donghyuk Lee, NVIDIA and Carnegie Mellon University
Samira Khan, University of Virginia
Lavanya Subramanian, Saugata Ghose, Rachata Ausavarungrunrun, Carnegie Mellon University
Gennady Pekhimenko, Vivek Seshadri, Microsoft Research
Onur Mutlu, ETH Zürich and Carnegie Mellon University
Why the Long Memory Latency?

- **Reason 1: Design of DRAM Micro-architecture**
  - Goal: Maximize capacity/area, not minimize latency

- **Reason 2: “One size fits all” approach to latency specification**
  - Same latency parameters for all temperatures
  - Same latency parameters for all DRAM chips
  - Same latency parameters for all parts of a DRAM chip
  - Same latency parameters for all supply voltage levels
  - Same latency parameters for all application data
  - ...

SAFARI 169
Another Example: EDEN for DNNs

- Deep Neural Network evaluation is very DRAM-intensive (especially for large networks)

1. Some data and layers in DNNs are very tolerant to errors

2. Reduce DRAM latency and voltage on such data and layers

3. While still achieving a user-specified DNN accuracy target by making training DRAM-error-aware

Data-aware management of DRAM latency and voltage for Deep Neural Network Inference
Example DNN Data Type to DRAM Mapping

Mapping example of ResNet-50:

Map more error-tolerant DNN layers to DRAM partitions with lower voltage/latency

4 DRAM partitions with different error rates
EDEN: Overview

**Key idea:** Enable **accurate, efficient** DNN inference using approximate DRAM

EDEN is an **iterative** process that has **3 key steps**
CPU: DRAM Energy Evaluation

Average 21% DRAM energy reduction maintaining accuracy within 1% of original
Average 8% system speedup
Some workloads achieve 17% speedup

EDEN achieves close to the ideal speedup possible via tRCD scaling
GPU, Eyeriss, and TPU: Energy Evaluation

- **GPU**: average 37% energy reduction
- **Eyeriss**: average 31% energy reduction
- **TPU**: average 32% energy reduction
EDEN: Data-Aware Efficient DNN Inference

- Skanda Koppula, Lois Orosa, A. Giray Yağlıkçı, Roknoddin Azizi, Taha Shahroodi, Konstantinos Kanellopoulos, and Onur Mutlu,

"EDEN: Enabling Energy-Efficient, High-Performance Deep Neural Network Inference Using Approximate DRAM"

Proceedings of the 52nd International Symposium on Microarchitecture (MICRO), Columbus, OH, USA, October 2019.

[Lightning Talk Slides (pptx) (pdf)]
[Lightning Talk Video (90 seconds)]
More on EDEN

Key idea: Enabling accurate, efficient DNN inference using approximate DRAM

EDEN is an iterative process that has 3 key steps:
Recall: Exploiting Memory Error Tolerance with Hybrid Memory Systems

On Microsoft’s Web Search workload
Reduces server hardware cost by 4.7 %
Achieves single server availability target of 99.90 %

Heterogeneous-Reliability Memory [DSN 2014]
Recall: Heterogeneous-Reliability Memory

Step 1: Characterize and classify application memory error tolerance

Step 2: Map application data to the HRM system enabled by SW/HW cooperative solutions
More on Heterogeneous-Reliability Memory

- Yixin Luo, Sriram Govindan, Bikash Sharma, Mark Santaniello, Justin Meza, Aman Kansal, Jie Liu, Badriddine Khessib, Kushagra Vaid, and Onur Mutlu,

"Characterizing Application Memory Error Vulnerability to Optimize Data Center Cost via Heterogeneous-Reliability Memory"

Proceedings of the 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), Atlanta, GA, June 2014. [Summary] [Slides (pptx) (pdf)] [Coverage on ZDNet]
Why the Long Memory Latency?

- **Reason 1: Design of DRAM Micro-architecture**
  - Goal: Maximize capacity/area, not minimize latency

- **Reason 2: “One size fits all” approach to latency specification**
  - Same latency parameters for all temperatures
  - Same latency parameters for all DRAM chips
  - Same latency parameters for all parts of a DRAM chip
  - **Same latency parameters for all supply voltage levels**
  - Same latency parameters for all application data
  - ...

SAFARI
Understanding & Exploiting the Voltage-Latency-Reliability Relationship
Analysis of Latency-Voltage in DRAM Chips

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

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

More on Voltron

https://www.youtube.com/watch?v=F17sytMs80o&list=PL5Q2soXY2Zi-DyoI3HbqcdtUm9YWRR_z&index=17
Reducing Memory Latency to Support Security Primitives
Using Memory for Security

- **Generating True Random Numbers (using DRAM)**
  - Kim et al., HPCA 2019
  - Olgun et al., ISCA 2021

- **Evaluating Physically Unclonable Functions (using DRAM)**
  - Kim et al., HPCA 2018

- **Quickly Destroying In-Memory Data (using DRAM)**
  - Orosa et al., arxiv 2019 + ISCA 2021
DRAM Latency PUFs


[Lightning Talk Video]
[Slides (pptx) (pdf)] [Lightning Session Slides (pptx) (pdf)]
[Full Talk Lecture Video (28 minutes)]
More on DRAM Latency PUFs

- A cell’s latency failure probability is inherently related to random process variation from manufacturing
- We can provide repeatable and unique device signatures using latency error patterns

High % chance to fail with reduced $t_{RCD}$
DRAM Latency True Random Number Generator

- Jeremie S. Kim, Minesh Patel, Hasan Hassan, Lois Orosa, and Onur Mutlu,
  "D-RaNGe: Using Commodity DRAM Devices to Generate True Random Numbers with Low Latency and High Throughput"


[Slides (pptx) (pdf)]
[Full Talk Video (21 minutes)]
[Full Talk Lecture Video (27 minutes)]

*Top Picks Honorable Mention by IEEE Micro.*
More on DRAM Latency TRNGs

D-RaNGe: Extracting Random Values

- Identify all DRAM cells that fail randomly when accessed with a reduced $t_{RCD}$ (RNG Cell)
- When accessing an RNG Cell with a reduced $t_{RCD}$, the values read will be truly random values

RNG Cell

0110100110011101000110101
In-DRAM True Random Number Generation

- Ataberk Olgun, Minesh Patel, A. Giray Yaglikci, Haocong Luo, Jeremie S. Kim, F. Nisa Bostanci, Nandita Vijaykumar, Oguz Ergin, and Onur Mutlu,
  "QUAC-TRNG: High-Throughput True Random Number Generation Using Quadruple Row Activation in Commodity DRAM Chips"
  [Slides (pptx) (pdf)]
  [Short Talk Slides (pptx) (pdf)]
  [Talk Video (25 minutes)]
  [SAFARI Live Seminar Video (1 hr 26 mins)]

**QUAC-TRNG: High-Throughput True Random Number Generation Using Quadruple Row Activation in Commodity DRAM Chips**

- Ataberk Olgun§†
- Minesh Patel§
- A. Giray Yağlıkçı§
- Haocong Luo§
- Jeremie S. Kim§
- F. Nisa Bostancı§†
- Nandita Vijaykumar§○
- Oğuz Ergin†
- Onur Mutlu§

§ETH Zürich  †TOBB University of Economics and Technology ○University of Toronto
More on QUAC-TRNG

Real Chip Characterization

Experimentally study QUAC and QUAC-TRNG using 136 real DDR4 chips from SK Hynix

DDR4 SoftMC → DRAM Testing Infrastructure

SAFARI Live Seminar: High-Throughput TRNG Using Quadruple Row Activation in Commodity DRAM Chips

https://www.youtube.com/watch?v=snvF3g3GfkI&list=PL5Q2soXY2Zi_tOTAYm--dYByNPL7JhwR9&index=6
Reducing Refresh Latency
Reducing Refresh Latency

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

VRL-DRAM: Improving DRAM Performance via Variable Refresh Latency

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

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

Onur Mutlu
ETH Zürich
Zürich, Switzerland
omutlu@gmail.com
Reducing Memory Latency by Exploiting Memory Access Patterns
ChargeCache: Exploiting Access Patterns

- Hasan Hassan, Gennady Pekhimenko, Nandita Vijaykumar, Vivek Seshadri, Donghyuk Lee, Oguz Ergin, and Onur Mutlu,
  "ChargeCache: Reducing DRAM Latency by Exploiting Row Access Locality"


[Slides (pptx) (pdf)]
[Source Code]
More on ChargeCache

Observation 1
A highly-charged DRAM row can be accessed with low latency
- tRCD: 44%
- tRAS: 37%

How does a row become highly-charged?

https://www.youtube.com/watch?v=snvF3q3GfkI&list=PL5Q2soXY2Zi_tOTAYm--dYByNPL7JhwR9&index=6
Partial Restoration of Cell Charge

- Yaohua Wang, Arash Tavakkol, Lois Orosa, Saugata Ghose, Nika Mansouri Ghiasi, Minesh Patel, Jeremie S. Kim, Hasan Hassan, Mohammad Sadrosadati, and Onur Mutlu,

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

Parallelizing Refreshes and Accesses

- Kevin Chang, Donghyuk Lee, Zeshan Chishti, Alaa Alameldeen, Chris Wilkerson, Yoongu Kim, and Onur Mutlu,
  "Improving DRAM Performance by Parallelizing Refreshes with Accesses"
  [Summary] [Slides (pptx) (pdf)]

Reducing Performance Impact of DRAM Refresh by Parallelizing Refreshes with Accesses

Kevin Kai-Wei Chang   Donghyuk Lee   Zeshan Chishti†
Alaa R. Alameldeen†   Chris Wilkerson†   Yoongu Kim   Onur Mutlu
Carnegie Mellon University   †Intel Labs
On DRAM Power Consumption
Saugata Ghose, A. Giray Yaglikci, Raghav Gupta, Donghyuk Lee, Kais Kudrolli, William X. Liu, Hasan Hassan, Kevin K. Chang, Niladrish Chatterjee, Aditya Agrawal, Mike O'Connor, and Onur Mutlu,

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


Abstract

VAMPIRE DRAM Power Model

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

Saugata Ghose† Abdullah Giray Yağlıkçi‡† Raghav Gupta† Donghyuk Lee‡
Kais Kudrolli† William X. Liu† Hasan Hassan‡ Kevin K. Chang†
Niladrish Chatterjee§ Aditya Agrawal§ Mike O’Connor§‖ Onur Mutlu‡†

†Carnegie Mellon University ‡ETH Zürich §NVIDIA ‖University of Texas at Austin
Summary: Low-Latency Memory
Challenge and Opportunity for Future

Fundamentally Low Latency Computing Architectures
Reason 1: Design of DRAM Micro-architecture
- Goal: Maximize capacity/area, not minimize latency

Reason 2: “One size fits all” approach to latency specification
- Same latency parameters for all temperatures
- Same latency parameters for all DRAM chips (e.g., rows)
- Same latency parameters for all parts of a DRAM chip
- Same latency parameters for all supply voltage levels
- Same latency parameters for all application data
- ...
Takeaway I

We Can Reduce Memory Latency with Change of Mindset
Main Memory Needs Intelligent Controllers to Reduce Latency
Some Solution Principles (More Compact)

- Data-centric design
- All components intelligent
- Better cross-layer communication, better interfaces
- Better-than-worst-case design
- Heterogeneity
- Flexibility, adaptability
- Open minds
Four Key Current Directions

- Fundamentally **Secure/Reliable/Safe** Architectures

- Fundamentally **Energy-Efficient** Architectures
  - Memory-centric (Data-centric) Architectures

- Fundamentally **Low-Latency and Predictable** Architectures

- Architectures for **AI/ML, Genomics, Medicine, Health, ...**
Solar-DRAM
Solar-DRAM: Putting It Together

- Jeremie S. Kim, Minesh Patel, Hasan Hassan, and Onur Mutlu, "Solar-DRAM: Reducing DRAM Access Latency by Exploiting the Variation in Local Bitlines"
  [Slides (pptx) (pdf)]
  [Talk Video (16 minutes)]

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

Jeremie S. Kim†§, Minesh Patel§, Hasan Hassan§, Onur Mutlu§†
†Carnegie Mellon University §ETH Zürich
Activation failures are **highly constrained** to local bitlines.
Short-term Variation

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

A weak bitline is likely to remain weak and a strong bitline is likely to remain strong over time.
Short-term Variation

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

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

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

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

Weak bitline

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

Uses a **static profile of weak subarray columns**
- Identifies subarray columns as weak or strong
- Obtained in a one-time profiling step

**Three Components**

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

Uses a static profile of weak subarray columns
  • Identifies subarray columns as weak or strong
  • Obtained in a one-time profiling step

Three Components

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

Identify cache lines comprised of strong bitlines
Access such cache lines with a reduced $t_{RCD}$
Solar-DRAM

Uses a \textit{static profile of weak subarray columns}
- Identifies subarray columns as weak or strong
- Obtained in a one-time profiling step

\textbf{Three Components}

1. Variable-latency cache lines (VLC)
2. Reordered subarray columns (RSC)
3. Reduced latency for writes (RLW)
Remap cache lines across DRAM at the memory controller level so cache line 0 will likely map to a strong cache line
Solar-DRAM

Uses a static profile of weak subarray columns
  • Identifies subarray columns as weak or strong
  • Obtained in a one-time profiling step

Three Components

1. Variable-latency cache lines (VLC)
2. Reordered subarray columns (RSC)
3. Reduced latency for writes (RLW)
Solar-DRAM: Putting It Together


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

Jeremie S. Kim†§, Minesh Patel§, Hasan Hassan§, Onur Mutlu§†
†Carnegie Mellon University §ETH Zürich

SAFARI 222
More on Solar DRAM

Spatial Distribution of Failures
How are activation failures spatially distributed in DRAM?

Activation failures are highly constrained to local bitlines (i.e., subarrays)

Solar-DRAM: Reducing DRAM Access Latency by Exploiting the Variation in Local Bitlines - ICCD 2018
101 views - Oct 23, 2018

https://www.youtube.com/watch?v=WPmDlIx1mKrU
Understanding & Exploiting the Voltage-Latency-Reliability Relationship
Analysis of Latency-Voltage in DRAM Chips

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

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


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

Kevin K. Chang† Abdullah Giray Yaglıkçı† Saugata Ghose† Aditya Agrawal¶ Niladrish Chatterjee¶
Abhijith Kashyap† Donghyuk Lee¶ Mike O’Connor¶,‡ Hasan Hassan§ Onur Mutlu§,†

†Carnegie Mellon University ‡NVIDIA ‡The University of Texas at Austin §ETH Zürich
High DRAM Power Consumption

• **Problem:** High DRAM (memory) power in today’s systems

>40% in POWER7 \((\text{Ware}+, \text{HPCA’10})\)

>40% in GPU \((\text{Paul}+, \text{ISCA’15})\)
Low-Voltage Memory

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

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

1. Understand and characterize the various characteristics of DRAM under reduced voltage.

2. Develop a mechanism that reduces DRAM energy by lowering voltage while keeping performance loss within a target.
Key Questions

• How does reducing voltage affect reliability (errors)?

• How does reducing voltage affect DRAM latency?

• How do we design a new DRAM energy reduction mechanism?
Supply Voltage Control on DRAM

Adjust the supply voltage to every chip on the same module.
Custom Testing Platform

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

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

Existing systems: DRAM commands not exposed to users

[Image of custom testing platform with labels for DRAM module, FPGA, and Voltage controller]

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

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

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

Errors induced by reduced-voltage operation

Reducing voltage below $V_{\text{min}}$ causes an increasing number of errors
Source of Errors

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

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

Reliable low-voltage operation requires higher latency
DIMMs Operating at Higher Latency

Measured minimum latency that does not cause errors in DRAM modules

DRAM requires longer latency to access data without errors at lower voltage
Spatial Locality of Errors

A module under 1.175V (12% voltage reduction)

Errors concentrate in certain regions
Summary of Key Experimental Observations

• Voltage-induced errors increase as voltage reduces further below $V_{\text{min}}$

• Errors exhibit spatial locality

• Increasing the latency of DRAM operations mitigates voltage-induced errors
**DRAM Voltage Adjustment to Reduce Energy**

- **Goal**: Exploit the trade-off between voltage and latency to reduce energy consumption

- **Approach**: Reduce DRAM voltage *reliably*
  - Performance loss due to increased latency at lower voltage

![Bar chart showing performance and power savings over different supply voltages.](chart)

- **Improvement Over Nominal Voltage (%)**
- **Supply Voltage (V)**
- **Performance**
- **DRAM Power Savings**
  - High Power Savings
  - Low Power Savings
  - Bad Performance
  - Good Performance

*SAFARI*
Voltron Overview

User specifies the performance loss target

Voltron

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

How do we predict performance loss due to increased latency under low DRAM voltage?
**Linear Model to Predict Performance**

Voltron

User specifies the performance loss target

Select the minimum DRAM voltage without violating the target

Application’s characteristics

[1.3V, 1.25V, ...]

DRAM Voltage

Linear regression model

Predicted performance loss

[-1%, -3%, ...]

Min. Voltage

Target

Final Voltage
Regression Model to Predict Performance

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

• **Handling multiple applications:**
  – Predict a performance loss for each application
  – Select the minimum voltage that satisfies the performance target for all applications
Comparison to Prior Work

• Prior work: Dynamically scale *frequency and voltage* of the entire DRAM based on bandwidth demand [David+, ICAC’11]
  – Problem: Lowering voltage on the peripheral circuitry decreases channel frequency (memory data throughput)

• Voltron: Reduce voltage to only **DRAM array** without changing the voltage to peripheral circuitry
Exploiting Spatial Locality of Errors

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

- **Benefit**: Higher performance
Voltron Evaluation Methodology

- **Cycle-level simulator**: Ramulator [CAL’15]
  - McPAT and DRAMPower for energy measurement
    https://github.com/CMU-SAFARI/ramulator

- **4-core** system with DDR3L memory

- **Benchmarks**: SPEC2006, YCSB

- **Comparison to prior work**: MemDVFS [David+, ICAC’11]
  - Dynamic DRAM frequency and voltage scaling
  - Scaling based on the *memory bandwidth consumption*
Energy Savings with Bounded Performance

More savings for high bandwidth applications

MemDVFS  Voltron

Meets performance target

Performance Target

-1.6%  -1.8%

3.2%  7.3%

Low  High

Memory Intensity

CPU+DRAM Energy Savings (%)

[David+, ICAC’11]
Voltron: Advantages & Disadvantages

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

- **Disadvantages**
  - Requires finding the reliable operating voltage for each chip → higher testing cost
Analysis of Latency-Voltage in DRAM Chips

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

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

Reducing Memory Latency to Support Security Primitives
Using Memory for Security

- **Generating True Random Numbers (using DRAM)**
  - Kim et al., HPCA 2019
  - Olgun et al., ISCA 2021

- **Evaluating Physically Unclonable Functions (using DRAM)**
  - Kim et al., HPCA 2018

- **Quickly Destroying In-Memory Data (using DRAM)**
  - Orosa et al., arxiv 2019 + ISCA 2021
D-RaNGe: Using Commodity DRAM Devices to Generate True Random Numbers with Low Latency and High Throughput

Jeremie S. Kim  Minesh Patel
Hasan Hassan  Lois Orosa  Onur Mutlu

SAFARI

ETH Zürich  Carnegie Mellon

HPCA 2019
D-RaNGe Executive Summary

• **Motivation**: High-throughput true random numbers enable system security and various randomized algorithms.
  - Many systems (e.g., IoT, mobile, embedded) do not have dedicated **True Random Number Generator (TRNG)** hardware but have DRAM devices.

• **Problem**: Current DRAM-based TRNGs either
  1. do not sample a fundamentally non-deterministic entropy source
  2. are too slow for continuous high-throughput operation

• **Goal**: A novel and effective TRNG that uses existing commodity DRAM to provide random values with 1) **high-throughput**, 2) **low latency** and 3) no adverse effect on concurrently running applications.

• **D-RaNGe**: Reduce DRAM access latency **below reliable values** and exploit DRAM cells’ failure probabilities to generate random values.

• **Evaluation**:
  1. Experimentally characterize **282 real LPDDR4 DRAM devices**
  2. **D-RaNGe (717.4 Mb/s)** has significantly higher throughput (**211x**)
  3. **D-RaNGe (100ns)** has significantly lower latency (**180x**)
DRAM Latency Characterization of 282 LPDDR4 DRAM Devices

• Latency failures come from accessing DRAM with reduced timing parameters.

• Key Observations:
  1. A cell’s latency failure probability is determined by random process variation
  2. Some cells fail randomly
Process variation during manufacturing results in cells having unique behavior.
Weaker cells have a higher probability to fail.
D-RaNGe Key Idea

**High % chance to fail with reduced $t_{RCD}$**

**Low % chance to fail with reduced $t_{RCD}$**

Fails randomly with reduced $t_{RCD}$
D-RaNGe Key Idea

We refer to cells that fail randomly when accessed with a reduced $t_{\text{RCD}}$ as RNG cells.
Our D-RaNGe Evaluation

• We generate random values by repeatedly accessing RNG cells and aggregating the data read

• The random data satisfies the NIST statistical test suite for randomness

• The D-RaNGe generates random numbers
  - Throughput: 717.4 Mb/s
  - Latency: 64 bits in <1us
  - Power: 4.4 nJ/bit
D-RaNGe: Using Commodity DRAM Devices to Generate True Random Numbers with Low Latency and High Throughput

Jeremie S. Kim  Minesh Patel
Hasan Hassan  Lois Orosa  Onur Mutlu

SAFARI  HPCA 2019

ETH Zürich  Carnegie Mellon
More on D-RaNGe

Jeremie S. Kim, Minesh Patel, Hasan Hassan, Lois Orosa, and Onur Mutlu, "D-RaNGe: Using Commodity DRAM Devices to Generate True Random Numbers with Low Latency and High Throughput"


[Slides (pptx) (pdf)]
[Full Talk Video (21 minutes)]
[Full Talk Lecture Video (27 minutes)]

Top Picks Honorable Mention by IEEE Micro.

D-RaNGe: Using Commodity DRAM Devices to Generate True Random Numbers with Low Latency and High Throughput

Jeremie S. Kim†§ Minesh Patel§ Hasan Hassan§ Lois Orosa§ Onur Mutlu§†
†Carnegie Mellon University §ETH Zürich
More on DRAM Latency TRNGs

D-RaNGe: Extracting Random Values

Identify all DRAM cells that fail randomly when accessed with a reduced $t_{RCD}$ (RNG Cell).

- When accessing an RNG Cell with a reduced $t_{RCD}$, the values read will be truly random values.

RNG Cell

SAFARI

https://www.youtube.com/watch?v=Y3hPv1I5f8Y&list=PL5Q2soXY2Zj-DyoI3HbgcdtUm9YWRR_z&index=16
Ataberk Olgun, Minesh Patel, A. Giray Yaglikci, Haocong Luo, Jeremie S. Kim, F. Nisa Bostanci, Nandita Vijaykumar, Oguz Ergin, and Onur Mutlu,
"QUAC-TRNG: High-Throughput True Random Number Generation Using Quadruple Row Activation in Commodity DRAM Chips"
[Slides (pptx) (pdf)]
[Short Talk Slides (pptx) (pdf)]
[Talk Video (25 minutes)]
[SAFARI Live Seminar Video (1 hr 26 mins)]

QUAC-TRNG: High-Throughput True Random Number Generation Using Quadruple Row Activation in Commodity DRAM Chips

Ataberk Olgun§†, Minesh Patel§, A. Giray Yağlıkçı§, Haocong Luo§, Jeremie S. Kim§, F. Nisa Bostanci§†, Nandita Vijaykumar§○, Öğuz Ergin†, Onur Mutlu§

§ETH Zürich †TOBB University of Economics and Technology ○University of Toronto
More on QUAC-TRNG

Real Chip Characterization

Experimentally study QUAC and QUAC-TRNG using 136 real DDR4 chips from SK Hynix

DDR4 SoftMC $\rightarrow$ DRAM Testing Infrastructure

SAFARI Live Seminar: High-Throughput TRNG Using Quadruple Row Activation in Commodity DRAM Chips

713 views • Streamed live on Sep 15, 2021

https://www.youtube.com/watch?v=snvF3g3GfkI&list=PL5Q2soXY2Zi_tOTAYm--dYByN7JhwR9&index=6
DRAM Latency PUFs

Jeremie S. Kim, Minesh Patel, Hasan Hassan, and Onur Mutlu, "The DRAM Latency PUF: Quickly Evaluating Physical Unclonable Functions by Exploiting the Latency-Reliability Tradeoff in Modern DRAM Devices"
[Lightning Talk Video]
[Slides (pptx) (pdf)] [Lightning Session Slides (pptx) (pdf)]
[Full Talk Lecture Video] (28 minutes)
The DRAM Latency PUF:
Quickly Evaluating Physical Unclonable Functions by Exploiting the Latency-Reliability Tradeoff in Modern Commodity DRAM Devices

Jeremie S. Kim     Minesh Patel
Hasan Hassan       Onur Mutlu

SAFARI

Systems@ETH Zürich

ETH Zürich

Carnegie Mellon
DL-PUF: Executive Summary

**Motivation:**
- We can authenticate a system via unique signatures if we can evaluate a Physical Unclonable Function (PUF) on it
- Signatures (PUF response) reflect inherent properties of a device
- DRAM is a promising substrate for PUFs because it is widely used

**Problem:** Current DRAM PUFs are 1) very slow, 2) require a DRAM reboot, or 3) require additional custom hardware

**Goal:** To develop a novel and effective PUF for existing commodity DRAM devices with low-latency evaluation time and low system interference across all operating temperatures

**DRAM Latency PUF:** Reduce DRAM access latency below reliable values and exploit the resulting error patterns as unique identifiers

**Evaluation:**
1. Experimentally characterize 223 real LPDDR4 DRAM devices
2. DRAM latency PUF (88.2 ms) achieves a speedup of 102x/860x at 70°C/55°C over prior DRAM PUF evaluation mechanisms
Motivation

We want a way to ensure that a system’s components are not **compromised**

- **Physical Unclonable Function (PUF)**: a function we **evaluate** on a device to **generate** a **signature unique** to the device
- We refer to the unique signature as a **PUF response**
- Often used in a **Challenge-Response Protocol (CRP)**
Motivation

1. We want a **runtime-accessible** PUF
   - Should be evaluated **quickly** with **minimal** impact on concurrent applications
   - Can protect against **attacks that swap system components with malicious parts**

2. DRAM is a **promising substrate** for evaluating PUFs because it is **ubiquitous** in modern systems
   - Unfortunately, current DRAM PUFs are **slow** and get **exponentially slower** at lower temperatures
DRAM Latency Characterization of 223 LPDDR4 DRAM Devices

- Latency failures come from accessing DRAM with reduced timing parameters.

- Key Observations:
  1. A cell’s latency failure probability is determined by random process variation
  2. Latency failure patterns are repeatable and unique to a device
DRAM Latency PUF Key Idea

- A cell’s latency failure probability is inherently related to random process variation from manufacturing.

- We can provide repeatable and unique device signatures using latency error patterns.

High % chance to fail with reduced $t_{RCD}$

Low % chance to fail with reduced $t_{RCD}$
DRAM Latency PUF Key Idea

- A cell’s latency failure probability is inherently related to random process variation from manufacturing.
- We can provide repeatable and unique device signatures using latency error patterns.

The **key idea** is to compose a PUF response using the DRAM cells that fail with **high probability**.
The DRAM Latency PUF Evaluation

• We generate PUF responses using *latency errors* in a region of DRAM

• The latency error patterns *satisfy PUF requirements*

• The DRAM Latency PUF *generates PUF responses in 88.2ms*
Results – PUF Evaluation Latency

DRAM latency PUF is

1. Fast and constant latency (88.2ms)
Results – PUF Evaluation Latency

- **8KiB memory segment**: 869.8x faster than the previous segment.
- **64KiB memory segment**: 108.9x faster than the previous segment.

**DRAM latency PUF is**

1. Fast and constant latency (88.2ms)
Results – PUF Evaluation Latency

DRAM latency PUF is
1. Fast and constant latency (88.2ms)
Results – PUF Evaluation Latency

<table>
<thead>
<tr>
<th>Temperature (°C)</th>
<th>Evaluation Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>56</td>
<td>10^4</td>
</tr>
<tr>
<td>58</td>
<td>10^3</td>
</tr>
<tr>
<td>60</td>
<td>10^2</td>
</tr>
<tr>
<td>62</td>
<td>10^1</td>
</tr>
<tr>
<td>64</td>
<td>10^0</td>
</tr>
<tr>
<td>66</td>
<td>10^-1</td>
</tr>
<tr>
<td>68</td>
<td>10^-2</td>
</tr>
<tr>
<td>70</td>
<td>10^-3</td>
</tr>
</tbody>
</table>

**DRAM latency PUF is**

1. Fast and constant latency (88.2ms)
2. On average, 102x/860x faster than the previous DRAM PUF with the same DRAM capacity overhead (64KiB)
Other Results in the Paper

• How the **DRAM latency PUF** meets the basic requirements for an effective PUF

• A detailed analysis on:
  - Devices of **the three major DRAM manufacturers**
  - The **evaluation time** of a PUF

• Further discussion on:
  - **Optimizing** retention PUFs
  - **System interference** of DRAM retention and latency PUFs
  - Algorithm to **quickly and reliably** evaluate DRAM latency PUF
  - **Design considerations** for a DRAM latency PUF
  - The DRAM Latency PUF overhead analysis
The DRAM Latency PUF:
Quickly Evaluating Physical Unclonable Functions by Exploiting the Latency-Reliability Tradeoff in Modern Commodity DRAM Devices

Jeremie S. Kim    Minesh Patel
Hasan Hassan      Onur Mutlu

QR Code for the paper

HPCA 2018

ETH Zürich

Carnegie Mellon
More on DRAM Latency PUFs

- Jeremie S. Kim, Minesh Patel, Hasan Hassan, and Onur Mutlu, "The DRAM Latency PUF: Quickly Evaluating Physical Unclonable Functions by Exploiting the Latency-Reliability Tradeoff in Modern DRAM Devices"


[Lightning Talk Video] [Slides (pptx) (pdf)] [Lightning Session Slides (pptx) (pdf)]
More on DRAM Latency PUFs

- A cell’s latency failure probability is inherently related to random process variation from manufacturing.
- We can provide repeatable and unique device signatures using latency error patterns.
- High % chance to fail with reduced $t_{\text{RCQ}}$.
Reducing Memory Latency by Exploiting Memory Access Patterns
ChargeCache: Executive Summary

- **Goal**: Reduce average DRAM access latency with no modification to the existing DRAM chips

- **Observations**:
  1) A highly-charged DRAM row can be accessed with low latency
  2) A row’s charge is restored when the row is accessed
  3) A recently-accessed row is likely to be accessed again: Row Level Temporal Locality (RLTL)

- **Key Idea**: Track recently-accessed DRAM rows and use lower timing parameters if such rows are accessed again

- **ChargeCache**:
  - Low cost & no modifications to the DRAM
  - Higher performance (8.6-10.6% on average for 8-core)
  - Lower DRAM energy (7.9% on average)
DRAM Charge over Time

- Cell
- Sense Amplifier

Graph showing charge over time with stages:
- Ready to Access
- Ready to Precharge
- Data 0
- Data 1

Events:
- tRCD
- tRAS

SAFARI
Accessing Highly-charged Rows

- Ready to Access
- Ready to Precharge

Cell
Sense-Amplifier

Data 1
Data 0

Sensing
Restore
Precharge

time

charge

ACT
R/W
PRE

tRCD
trRAS
Observation 1

A highly-charged DRAM row can be accessed with low latency

- tRCD: 44%
- tRAS: 37%

How does a row become highly-charged?
How Does a Row Become Highly-Charged?

DRAM cells lose charge over time

Two ways of restoring a row’s charge:

- Refresh Operation
- Access
Observation 2

A row’s charge is restored when the row is accessed

How likely is a recently-accessed row to be accessed again?
Row Level Temporal Locality (RLTL)

A **recently-accessed** DRAM row is likely to be accessed again.

- **t-RLTL**: Fraction of rows that are accessed within time *t* after their previous access.
Key Idea

Track **recently-accessed** DRAM rows and use **lower timing parameters** if such rows are accessed again.
ChargeCache Overview

ChargeCache

Memory Controller

Requests: A D A

ChargeCache Hits: Use Default Timings

DRAM

A
B
C
D
E
F
Area and Power Overhead

• Modeled with CACTI

• Area
  – ~5KB for 128-entry ChargeCache
  – 0.24% of a 4MB Last Level Cache (LLC) area

• Power Consumption
  – 0.15 mW on average (static + dynamic)
  – 0.23% of the 4MB LLC power consumption
Methodology

• Simulator
  – DRAM Simulator (Ramulator [Kim+, CAL’15])
    https://github.com/CMU-SAFARI/ramulator

• Workloads
  – 22 single-core workloads
    • SPEC CPU2006, TPC, STREAM
  – 20 multi-programmed 8-core workloads
    • By randomly choosing from single-core workloads
  – Execute at least 1 billion representative instructions per core (Pinpoints)

• System Parameters
  – 1/8 core system with 4MB LLC
  – Default tRCD/tRAS of 11/28 cycles
Single-core Performance

ChargeCache improves single-core performance
Eight-core Performance

- NUAT: 2.5%
- ChargeCache: 9%
- ChargeCache + NUAT
- LL-DRAM (Upperbound): 13%

ChargeCache significantly improves multi-core performance.
ChargeCache reduces DRAM energy
More on ChargeCache

- Hasan Hassan, Gennady Pekhimenko, Nandita Vijaykumar, Vivek Seshadri, Donghyuk Lee, Oguz Ergin, and Onur Mutlu,

"ChargeCache: Reducing DRAM Latency by Exploiting Row Access Locality"

[Slides (pptx) (pdf)]
[Source Code]
More on ChargeCache

Observation 1

A highly-charged DRAM row can be accessed with low latency

- tRCD: 44%
- tRAS: 37%

How does a row become highly-charged?

https://www.youtube.com/watch?v=snvF3q3GfkI&list=PL5Q2soXY2Zi_tOTAYm--dYByNPL7JhwR9&index=6
Partial Restoration of Cell Charge

- Yaohua Wang, Arash Tavakkol, Lois Orosa, Saugata Ghose, Nika Mansouri Ghiasi, Minesh Patel, Jeremie S. Kim, Hasan Hassan, Mohammad Sadrosadati, and Onur Mutlu,

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

On DRAM Power Consumption
VAMPIRE DRAM Power Model

- Saugata Ghose, A. Giray Yaglikci, Raghav Gupta, Donghyuk Lee, Kais Kudrolli, William X. Liu, Hasan Hassan, Kevin K. Chang, Niladrish Chatterjee, Aditya Agrawal, Mike O'Connor, and Onur Mutlu,

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


[Abstract]
[POMACS Journal Version (same content, different format)]
[Slides (pptx) (pdf)]
[VAMPIRE DRAM Power Model]
Power Measurement Platform

- Keysight 34134A DC Current Probe
- DDR3L SO-DIMM
- Virtex 6 FPGA
- JET-5467A Riser Board

Page 301 of 20
Power Measurement Methodology

- **SoftMC: an FPGA-based memory controller** [Hassan+ HPCA ’17]
  - Modified to repeatedly loop commands
  - Open-source: [https://github.com/CMU-SAFARI/SoftMC](https://github.com/CMU-SAFARI/SoftMC)

- Measure current consumed by a module during a SoftMC test

- **Tested 50 DDR3L DRAM modules** (200 DRAM chips)
  - Supply voltage: 1.35 V
  - **Three major vendors:** A, B, C
  - Manufactured between 2014 and 2016

- **For each experimental test that we perform**
  - 10 runs of each test per module
  - At least 10 current samples per run
1. Real DRAM Power Varies Widely from IDD Values

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

Current consumed by real DRAM modules varies significantly for all IDD values that we measure.
2. DRAM Power is Dependent on Data Values

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

DRAM power consumption depends *strongly* on the data value
### 3. Structural Variation Affects DRAM Power Usage

**Vendor C:** variation in idle current across banks

**All vendors:** variation in read current across banks

**All vendors:** variation in activation based on row address

**Significant structural variation:** DRAM power varies systematically by bank and row
4. Generational Savings Are Smaller Than Expected

- Similar trends for idle and read currents

Actual power savings of newer DRAM is much lower than the savings indicated in the datasheets
Summary of New Observations on DRAM Power

1. Real DRAM modules often consume less power than vendor-provided IDD values state

2. DRAM power consumption is dependent on the data value that is read/written

3. Across banks and rows, structural variation affects power consumption of DRAM

4. Newer DRAM modules save less power than indicated in datasheets by vendors

*Detailed observations and analyses in the paper*
A New Variation-Aware DRAM Power Model

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

**Inputs**
(from memory system simulator)
- Trace of DRAM commands, timing
- Data that is being written

**VAMPIRE**
- Read/Write and Data-Dependent Power Modeling
- Idle/Activate/Precharge Power Modeling
- Structural Variation Aware Power Modeling

**Outputs**
- Per-vendor power consumption
- Range for each vendor (optional)

- **VAMPIRE** and raw characterization data are open-source: https://github.com/CMU-SAFARI/VAMPIRE
VAMPIRE has very low error for all vendors: 6.8%

Much more accurate than prior models.
VAMPIRE Enables Several New Studies

- Taking advantage of structural variation to perform variation-aware physical page allocation to reduce power

- Smarter DRAM power-down scheduling

- Reducing DRAM energy with data-dependency-aware cache line encodings
  - 23 applications from the SPEC 2006 benchmark suite
  - Traces collected using Pin and Ramulator

- We expect there to be many other new studies in the future
VAMPIRE DRAM Power Model

- Saugata Ghose, A. Giray Yaglikci, Raghav Gupta, Donghyuk Lee, Kais Kudrolli, William X. Liu, Hasan Hassan, Kevin K. Chang, Niladrish Chatterjee, Aditya Agrawal, Mike O'Connor, and Onur Mutlu,

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


[Abstract]
[POMACS Journal Version (same content, different format)]
[Slides (pptx) (pdf)]
[VAMPIRE DRAM Power Model]