Understanding Reduced-Voltage Operation in Modern DRAM Devices

Experimental Characterization, Analysis, and Mechanisms

Kevin Chang†

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

†Carnegie Mellon University  SAFARI  *NVIDIA  ‡ETH Zürich
Executive Summary

- DRAM (memory) power is significant in today’s systems
  - Existing low-voltage DRAM reduces voltage conservatively

- Goal: Understand and exploit the reliability and latency behavior of real DRAM chips under aggressive reduced-voltage operation

- Key experimental observations:
  - Errors occur and increase with lower voltage
  - Errors exhibit spatial locality
  - Higher operation latency mitigates voltage-induced errors

- Voltron: A new DRAM energy reduction mechanism
  - Reduce DRAM voltage without introducing errors
  - Use a regression model to select voltage that does not degrade performance beyond a chosen target → 7.3% system energy reduction
Outline

• Executive Summary
• **Motivation**
• DRAM Background
• Characterization of DRAM
• Voltron: DRAM Energy Reduction Mechanism
• Conclusion
High DRAM Power Consumption

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

>40% in POWER7 \((\text{Ware+}, \text{HPCA’10})\)  >40% in GPU \((\text{Paul+}, \text{ISCA’15})\)
Low-Voltage Memory

• Existing DRAM designs to help reduce DRAM power by lowering supply voltage conservatively
  – \( \text{Power} \propto \text{Voltage}^2 \)
• 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?
Outline

• Executive Summary
• Motivation
• **DRAM Background**
  • Characterization of DRAM
  • Voltron: DRAM Energy Reduction Mechanism
• Conclusion
High-Level DRAM Organization

CPU

DRAM Channel

DRAM Module

DRAM chips
DRAM Chip Internals

Peripheral Circuitry
- Control Logic
- I/O

DRAM Array
- Bank

Off-chip channel

Bitline
Wordline
Sense amplifiers (row buffer)

DRAM Cell

SAFARI
**DRAM Operations**

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

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

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

to I/O
## DRAM Access Latency

### 1. Activation latency
(13ns / 50 cycles)

### 2. Precharge latency
(13ns / 50 cycles)

**Command**
- ACTIVATE
- READ
- PRECHARGE

**Data**
- Cache line (64B)

**Duration**
- Next ACT

**Diagram**: Illustration of DRAM access latency with commands (ACTIVATE, READ, PRECHARGE) and data (cache line 64B) showing activation and precharge latencies.

**SAFARI**
Outline

• Executive Summary
• Motivation
• DRAM Background
• Characterization of DRAM
  – Experimental methodology
  – Impact of voltage on reliability and latency
• Voltron: DRAM Energy Reduction Mechanism
• Conclusion
Supply Voltage Control on DRAM

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

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

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

Existing systems: DRAM commands not exposed to users

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

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

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

• Executive Summary
• Motivation
• DRAM Background
• Characterization of DRAM
  – Experimental methodology
  – Impact of voltage on reliability and latency
• Voltron: DRAM Energy Reduction Mechanism
• Conclusion
Reliability Worsens with Lower Voltage

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

Errors induced by reduced-voltage operation
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
Other Results in the Paper

• **Error-Correcting Codes (ECC)**
  – ECC (SECDED) is **not** sufficient to mitigate the errors

• **Effect of temperature**
  – Higher temperature requires higher latency under some voltage levels

• **Data retention time**
  – Lower voltage does **not** require more frequent refreshes

• **Effect of stored data pattern on error rate**
  – Difference is **not** statistically significant to draw conclusion
Summary of Key Experimental Observations

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

- Errors exhibit **spatial locality**.

- **Increasing the latency** of DRAM operations mitigates voltage-induced errors.
Outline

• Executive Summary
• Motivation
• DRAM Background
• Characterization of DRAM
  – Experimental methodology
  – Impact of voltage on reliability and latency
• Voltron: DRAM Energy Reduction Mechanism
• Conclusion
DRAM Voltage Adjustment to Reduce Energy

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

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

![Graph showing the relationship between supply voltage, DRAM power savings, and performance improvement over nominal voltage. The graph illustrates different regions for high and low power savings, with corresponding performance levels.]
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

Final Voltage

Target
Linear Model to Predict Performance

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

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

- **Prior work**: Dynamically scale *frequency and voltage* of the entire DRAM based on bandwidth demand [David+, ICAC’11]
  - **Problem**: Lowering voltage on the peripheral circuitry decreases channel frequency (memory data throughput)
- **Voltron**: Reduce voltage to only DRAM array without changing the voltage to peripheral circuitry
Exploiting Spatial Locality of Errors

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

- **Benefit:** Higher performance
Outline

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

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

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

- **Benchmarks**: SPEC2006, YCSB

- **Comparison to prior work**: MemDVFS [David+, ICAC’11]
  - Dynamic DRAM frequency and voltage scaling
  - Scaling based on the memory bandwidth consumption
1. Voltron improves energy for both low and high intensity workloads

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

Energy Savings with Bounded Performance

- More savings for high bandwidth applications
- Meets performance target

MemDVFS [David+, ICAC’11]
Outline

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

• DRAM (memory) power is significant in today’s systems
  – Existing low-voltage DRAM reduces voltage conservatively

• Goal: Understand and exploit the reliability and latency behavior of real DRAM chips under aggressive reduced-voltage operation

• Key experimental observations:
  – Errors occur and increase with lower voltage
  – Errors exhibit spatial locality
  – Higher operation latency mitigates voltage-induced errors

• Voltron: A new DRAM energy reduction mechanism
  – Reduce DRAM voltage without introducing errors
  – Use a regression model to select voltage that does not degrade performance beyond a chosen target \( \rightarrow \) 7.3% system energy reduction
Understanding Reduced-Voltage Operation in Modern DRAM Devices

Experimental Characterization, Analysis, and Mechanisms

Kevin Chang†

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

† Carnegie Mellon University

SAFARI

‡ ETH Zürich
Errors Rates Across Modules

![Graph showing errors rates across modules with different supply voltages for vendors A, B, and C.](image)

- **Vendor A**
- **Vendor B**
- **Vendor C**

The graph plots the fraction of cache lines with errors (%) on the y-axis against supply voltage (V) on the x-axis, demonstrating the performance of different vendors under varying supply voltages.
Error Density

Vendor A

Vendor B

Vendor C

Fraction of Data Beats

Supply Voltage (V)

Error Bits
- >2
- 2
- 1
- 0

SAFARI
Temperature Impact

Vendor A

Vendor B

Vendor C

Supply Voltage (V)
Impact on Retention Time

Temperature = 20°C
- 1.15V
- 1.20V
- 1.35V

Temperature = 70°C
Derivation of More Precise Latency

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

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

![Graph showing the relationship between memory intensity and performance degradation](image)

- **Voltage = 1.2V (-13%)**
  - **DRAM Latency +5%**

**MPKI = Last-level cache Misses Per Thousand Instruction**
Performance-Aware Voltage Adjustment

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

\[ \text{PredictedLoss} = \theta_0 + \theta_1 \text{Latency} + \theta_2 \text{App.Intensity} + \theta_3 \text{App.StallTime} \]

- Latency due to voltage adjustment
- The running application’s characteristics

• \( \theta \)s are trained through 151 application samples

• Use the model to select a minimum voltage that satisfies a performance loss target specified by the user
Linear Model Accuracy

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

Algorithm 1 Array Voltage Selection

1. SELECT_ARRAY_VOLTAGE\(target\_loss\)
2. \textbf{for each} interval
3. \hspace{1cm} profile = GetMemoryProfile()
4. \hspace{1cm} \texttt{NextVarray} = 1.35
5. \hspace{1cm} \textbf{for} \hspace{1cm} V_{array} \leftarrow 0.9 \hspace{0.5cm} \textbf{to} \hspace{0.5cm} 1.3
6. \hspace{1.5cm} \texttt{predicted\_loss} = \texttt{Predict}\left(\texttt{Latency(V_{array}), profile.MPKI, profile.StallTime}\right)
7. \hspace{1.5cm} \textbf{if} \hspace{0.5cm} \texttt{predicted\_loss} \leq \texttt{target\_loss} \hspace{0.5cm} \textbf{then}
8. \hspace{2cm} \texttt{NextVarray} = \texttt{V_{array}}
9. \hspace{2cm} \texttt{break}
10. \texttt{ApplyVoltage(NextVarray)}

\hspace{1cm} \textgreater{} Enter at the end of an interval
\hspace{1cm} \textgreater{} Search for the smallest $V_{array}$ that satisfies the performance loss target
\hspace{1cm} \textgreater{} Predict performance loss
\hspace{1cm} \textgreater{} Compare the predicted loss to the target
\hspace{1cm} \textgreater{} Use the current $V_{array}$ for the next interval
\hspace{1cm} \textgreater{} Apply the new $V_{array}$ for the next interval
Effect of Exploiting Error Locality

![Graph showing system performance loss and energy savings for Voltron and Voltron+BL in non-intensive and intensive scenarios.](image-url)
Energy Breakdown

Non-Memory-Intensive

Memory-Intensive

[Chart showing energy breakdown for Baseline and Voltron in both Non-Memory-Intensive and Memory-Intensive scenarios, with categories for DRAM Static, DRAM Dynamic, and CPU.]
Heterogeneous Workloads

![Graph showing system performance loss and Perf/Watt Improvement](image)

- **System Performance Loss (%)**: Comparison between MemDVFS and Voltron across different fractions of memory-intensive benchmarks (0%, 25%, 50%, 75%, 100%).
- **Perf/Watt Improvement (%)**: Similar comparison for Perf/Watt Improvement across the same fractions of benchmarks.
Performance Target Sweep

System Performance Loss (%) vs. System Performance Loss Target (%)

Perf/Watt Improvement (%) vs. System Performance Loss Target (%)
Sensitivity to Profile Interval Length

[Graph showing the relationship between Profile Interval Length (in million cycles) and Perf/Watt Improvement (%).]