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

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

June 21, 2018
Main memory in computers consists of DRAM modules.

DRAM consumes up to half of total system power.

State-of-the-art DRAM power models are not adequate:
- Based on IDD values: standard current measurements provided by vendors
- Often have a high mean absolute percentage error: 32% for DRAMPower, 161% for Micron power model

Our goal:
Measure and analyze the power used by real DRAM, and build an accurate DRAM power model.
Outline

Background: DRAM Organization & Operation

Characterization Methodology

New Findings on DRAM Power Consumption

VAMPIRE: A Variation-Aware DRAM Power Model

Conclusion
Simplified DRAM Organization and Operation

- Fundamental DRAM commands: activate, read, write, precharge
- Fundamental DRAM commands: activate, read, write, precharge
- One row of DRAM: 8 kB
- One cache line of data: 64 B
Fundamental DRAM commands: activate, read, write, precharge

One row of DRAM: 8 kB

One cache line of data: 64 B
Power Measurement Platform

Keysight 34134A DC Current Probe

DDR3L SO-DIMM

Virtex 6 FPGA

JET-5467A Riser Board
Methodology Details

- **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
Outline

Background: DRAM Organization & Operation

Characterization Methodology

New Findings on DRAM Power Consumption

VAMPIRE: A Variation-Aware DRAM Power Model

Conclusion
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, but not on bit toggling**
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.
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.
Outline

Background: DRAM Organization & Operation

Characterization Methodology

New Findings on DRAM Power Consumption

VAMPIRE: A Variation-Aware DRAM Power Model

Conclusion
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 will be open-source: [https://github.com/CMU-SAFARI/VAMPIRE](https://github.com/CMU-SAFARI/VAMPIRE) (August 2018)
VAMPIRE Has Lower Error Than Existing Models

- Validated using new power measurements: details in the paper

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
Outline

Background: DRAM Organization & Operation

Characterization Methodology

New Findings on DRAM Power Consumption

VAMPIRE: A Variation-Aware DRAM Power Model

Conclusion
### Conclusion

- **DRAM consumes up to half of total system power:** need to develop new low-power solutions

- **State-of-the-art DRAM power models are based only on IDD values, and have a high error**

- **We make four new observations on DRAM power consumption using 50 real DRAM modules from three major vendors**
  - Real DRAM modules often consume less power than IDD values state
  - Power consumption is dependent on the data value being read/written
  - Across banks and rows, structural variation affects power consumption
  - Newer DRAM modules save less power than indicated in datasheets

- **VAMPIRE:** a new DRAM power model built on our observations
  - Mean absolute percentage error of only 6.8%
  - Case study: dependency-aware data encoding reduces DRAM power by 12%

More information: [https://github.com/CMU-SAFARI/VAMPIRE](https://github.com/CMU-SAFARI/VAMPIRE)
What Your DRAM Power Models Are Not Telling You:
Lessons from a Detailed Experimental Study

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

More information: https://github.com/CMU-SAFARI/VAMPIRE
Backup Slides
More Information in the Paper…

- Full characterization analysis
- Application-level comparison to existing power models
- Case study: dependency-aware data encoding

Paper available at https://github.com/CMU-SAFARI/VAMPIRE
Today’s Models Leave a Lot to Be Desired

- Most models reliant on JEDEC-based IDD values
  - Micron power calculator
  - DRAMPower
  - gem5/GPGPU-Sim

- Some rely on circuit-level models
  - Vogelsang model for memory scaling
  - CACTI

- None are all that accurate
  - One value for each DRAM
  - Does not capture any inherent variation (e.g., data, structure)
How Do We Measure Current?

Foundation of Current Power Models

Just how bad are current models?

JEDEC defines a set of IDD values

<table>
<thead>
<tr>
<th>IDD</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDD0</td>
<td>Activation and Precharge</td>
</tr>
<tr>
<td>IDD1</td>
<td>Activation – 1 Column Read – Precharge</td>
</tr>
<tr>
<td>IDD2N</td>
<td>Precharge Standby (all banks are precharged/closed)</td>
</tr>
<tr>
<td></td>
<td>✔ clk enabled</td>
</tr>
<tr>
<td>IDD3N</td>
<td>Active Standby (all banks are active/opened)</td>
</tr>
<tr>
<td></td>
<td>✔ clk enabled</td>
</tr>
<tr>
<td>IDD2P</td>
<td>Precharge Power-Down (all banks are precharged/closed)</td>
</tr>
<tr>
<td></td>
<td>✘ clk disabled</td>
</tr>
<tr>
<td>IDD3P</td>
<td>Active Power-Down (all banks are active/opened)</td>
</tr>
<tr>
<td></td>
<td>✘ clk disabled</td>
</tr>
<tr>
<td>IDD4R/W</td>
<td>Burst mode Read/Write</td>
</tr>
<tr>
<td>IDD5B</td>
<td>Burst mode Refresh</td>
</tr>
<tr>
<td>IDD7</td>
<td>Activate – Column Read w/ Auto Precharge</td>
</tr>
</tbody>
</table>
JEDEC defined IDD measurement loops cover:

- Average power consumption of all banks
  » missing variation across banks
- Average power consumption of only two rows: 00 and F0
  » missing variation across rows in a subarray
  » missing variation across subarrays
- Average power consumption of only two data patterns: 00 and 33
  » missing effect of number of ones/zeroes in data
  » missing effect of toggling bits
IDD0: Activation and Precharge Energy

![Diagram showing ACT, tRAS, PRE, tRP, and ACT with time on the x-axis and current on the y-axis for Vendor A, Vendor B, and Vendor C.](Image)

- **normalized measured current**
  - **Datasheet**
  - **Measured**

- **current (mA)**
  - 0
  - 50
  - 100
  - 150
  - 200

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

- **normalized measured current**
  - A
  - B
  - C

- **time**
  - 0xF0
  - 0x00

- **DRAM Array**
- **Row Buffer**
IDD1: Activation, Read, and Precharge Energy

![Diagram showing DRAM Array, Row Buffer, ACT, tRCD, RD, PRE, tRP, ACT, tRAS, time with vendors A, B, C data points]

- **Vendor A**
  - Datasheet: 350 mA
  - Measured: 300 mA

- **Vendor B**
  - Datasheet: 250 mA
  - Measured: 200 mA

- **Vendor C**
  - Datasheet: 150 mA
  - Measured: 100 mA

**Normalized Measured Current**

- **Vendor A**
  - Normalized: 0.70

- **Vendor B**
  - Normalized: 0.60

- **Vendor C**
  - Normalized: 0.40
IDD2N: Precharged Standby

(1) Precharge All Banks (Close Row Buffers)

(2) Wait

Vendor A Vendor B Vendor C

Datasheet Measured

Normalized Measured Current

Current (mA)
IDD3N: Active Standby

Bank 0
DRAM Array
Row Buffer

Bank 1
DRAM Array
Row Buffer

Bank 7
DRAM Array
Row Buffer

(1) Activate All Banks (Open Row Buffers)

(2) Wait

---

Vendor A Vendor B Vendor C

Row Buffer
DRAM Array
Bank 0 Bank 1 Bank 7

Datasheet Measured

<table>
<thead>
<tr>
<th>Vendor</th>
<th>Measured Current (mA)</th>
<th>Norm Current</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
</tbody>
</table>

---

Normalized Measured Current

(Area A)
IDD2P: Precharged Power Down

(1) Precharge All Banks (Close Row Buffers)

Bank 0
DRAM Array

Bank 1
DRAM Array

Bank 7
DRAM Array

Row Buffer Row Buffer Row Buffer

(2) Wait

CLK is Disabled

Normalized Measured Current

Datasheet

Normalized Measured Current

Vendor A Vendor B Vendor C

Row Buffer

Row Buffer

Row Buffer

80
60
40
20
0

Current (mA)

Vendor A
Vendor B
Vendor C

80
60
40
20
0

Normalized Measured Current

0.00
0.20
0.40
0.60
0.80
1.00
1.20

A
B
C
### IDD4R: Burst Read Current

1. **Activate All Banks (Open Row Buffers)**
2. **Read one column at a time**
3. **Interleave across banks after each read**

![Diagram](image)

**Bar Graph**

- **X-axis (Vendor)**: Vendor A, Vendor B, Vendor C
- **Y-axis (Current)**: 0 to 800 (mA)

- **Datasheet**
- **Measured**
- **Corrected**
IDD4W: Burst Write Current

1. Activate All Banks (Open Row Buffers)
2. Write one column at a time
3. Interleave across banks after each read

<table>
<thead>
<tr>
<th>Bank 0</th>
<th>Bank 1</th>
<th>Bank 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM Array</td>
<td>DRAM Array</td>
<td>DRAM Array</td>
</tr>
<tr>
<td>Row Buffer</td>
<td>Row Buffer</td>
<td>Row Buffer</td>
</tr>
<tr>
<td>0x00 0x33</td>
<td>0x00 0x33</td>
<td>0x00 0x33</td>
</tr>
</tbody>
</table>

---

**Graphs:**

- **Current (mA):**
  - **Vendor A:** 400 mA
  - **Vendor B:** 200 mA
  - **Vendor C:** 200 mA

- **Normalized Measured Current:**
  - **Vendor A:** 0.60
  - **Vendor B:** 0.55
  - **Vendor C:** 0.50

---

**Legend:**
- □ Datasheet
- ■ Measured
IDD5B: Refresh in Burst Mode

Burst Mode:

REF \rightarrow t_{RFC} \rightarrow REF \rightarrow t_{RFC} \rightarrow REF

t_{REFI} (64ms)

t_{RFC}

Datasheet

Measured

Normalized Measured Current

Current (mA)

Vendor A  Vendor B  Vendor C

0 200 400 600 800 1000

A  B  C
IDD7: Read, Auto-Precharge

- ACT-RDA
  - $t_{RRD}$
  - ACT-RDA
  - $t_{RRD}$
  - ACT-RDA
  - $t_{RRD}$

(time)

---

### Normalized Measured Current

<table>
<thead>
<tr>
<th>Vendor</th>
<th>Datasheet</th>
<th>Measured</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.65</td>
<td>0.55</td>
</tr>
<tr>
<td>B</td>
<td>0.50</td>
<td>0.45</td>
</tr>
<tr>
<td>C</td>
<td>0.40</td>
<td>0.35</td>
</tr>
</tbody>
</table>

---

### Current (mA)

- Vendor A: 600 mA
- Vendor B: 500 mA
- Vendor C: 400 mA

---

Box plots showing the distribution of measured current for vendors A, B, and C.
Impact of Bit Toggling on DRAM Power

Bank 0 Row Buffer
0000 1010 1111 ...

Bank 1 Row Buffer
1011 0010 1011 ...

Bank 7 Row Buffer
1011 0010 1011 ...

Column Select

Bank Select
global bitlines

peripheral bus to I/O drivers

Column Interleaving

Bank and Column Interleaving

Toggle Sensitivity (mA/bit)

Vendor A
Vendor B
Vendor C
## Data Dependency Model

### Read

<table>
<thead>
<tr>
<th>Vendor</th>
<th>F (mA)</th>
<th>G (mA)</th>
<th>H (mA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vendor A</td>
<td>246.44</td>
<td>0.433</td>
<td>0.0515</td>
</tr>
<tr>
<td>Vendor B</td>
<td>217.42</td>
<td>0.157</td>
<td>0.0947</td>
</tr>
<tr>
<td>Vendor C</td>
<td>234.42</td>
<td>0.154</td>
<td>0.0856</td>
</tr>
</tbody>
</table>

### Write

<table>
<thead>
<tr>
<th>Vendor</th>
<th>F (mA)</th>
<th>G (mA)</th>
<th>H (mA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vendor A</td>
<td>531.18</td>
<td>-0.246</td>
<td>0.0461</td>
</tr>
<tr>
<td>Vendor B</td>
<td>466.84</td>
<td>-0.215</td>
<td>0.0166</td>
</tr>
<tr>
<td>Vendor C</td>
<td>368.29</td>
<td>-0.116</td>
<td>0.0229</td>
</tr>
</tbody>
</table>

\[ y = F + Gn + Ht \]

- **Additional current per logic-1**
- **Additional current per bit toggle**

---

![](image1.png)  

![](image2.png)

![](image3.png)

![](image4.png)
## Models

### No Interleaving (Same Bank & Column)

<table>
<thead>
<tr>
<th>Vendor</th>
<th>Read</th>
<th>Write</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F$</td>
<td>$G$</td>
</tr>
<tr>
<td>A</td>
<td>250.88</td>
<td>0.449</td>
</tr>
<tr>
<td>B</td>
<td>226.69</td>
<td>0.164</td>
</tr>
<tr>
<td>C</td>
<td>222.11</td>
<td>0.134</td>
</tr>
</tbody>
</table>

### Bank Interleaving Only

<table>
<thead>
<tr>
<th>Vendor</th>
<th>Read</th>
<th>Write</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F$</td>
<td>$G$</td>
</tr>
<tr>
<td>A</td>
<td>287.24</td>
<td>0.244</td>
</tr>
<tr>
<td>B</td>
<td>228.14</td>
<td>0.159</td>
</tr>
<tr>
<td>C</td>
<td>289.99</td>
<td>0.034</td>
</tr>
</tbody>
</table>

### Column Interleaving Only

<table>
<thead>
<tr>
<th>Vendor</th>
<th>Read</th>
<th>Write</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F$</td>
<td>$G$</td>
</tr>
<tr>
<td>A</td>
<td>277.13</td>
<td>0.267</td>
</tr>
<tr>
<td>B</td>
<td>223.61</td>
<td>0.152</td>
</tr>
<tr>
<td>C</td>
<td>266.51</td>
<td>0.099</td>
</tr>
</tbody>
</table>

[https://github.com/CMU-SAFARI/VAMPIRE](https://github.com/CMU-SAFARI/VAMPIRE)
Normalized **Active Standby** Energy across Banks

Normalized **Read** Burst Energy across Banks

Normalized **Write Burst** Energy across Banks
Evaluated System Configuration

Processor  x86-64 ISA, one core 3.2 GHz, 128-entry instruction window
Cache   L1: 64 kB, 4-way associative; L2: 2 MB, 16-way associative
Memory  64/64-entry read/write request queues,
Controller  FR-FCFS [119, 149]
DRAM    DDR3L-800 [57], 1 channel, 1 rank/8 banks per channel

- Application traces collected using Pin
- DRAM command timings generated using Ramulator: https://github.com/CMU-SAFARI/ramulator
Trends Across Generations

- Basically, if you’re building a system, you aren’t getting the kinds of savings you were promised.
Data Encoding

- Baseline: No coding
- BDI: Base Delta Immediate
- Optimized: Minimize the number of ones
- OWI: Minimize ones for reads, maximize ones for writes

![Normalized Energy Graph]

12.5% Energy Reduction
Validating Our DRAM Power Models

- New tests run on 22 of our DDR3L DRAM SO-DIMMs
- Validation command sequence
  - Activate
  - $n$ reads
    - Sweep $n$ from 0 to 764
    - All reads contain data value 0xAA
    - All reads to Bank 0, Row 128
    - Column interleaved
  - Precharge
- Error metric: mean absolute percentage error (MAPE)

- Best prior model (DRAMPower): 32.4% MAPE
- VAMPIRE: 6.8% MAPE