QUAC-TRNG

High-Throughput True Random Number Generation Using Quadruple Row Activation in Real DRAM Chips

Ataberk Olgun
Minesh Patel  A. Giray Yağlıkçı  Haocong Luo
Jeremie S. Kim  F. Nisa Bostancı  Nandita Vijaykumar
Oğuz Ergin  Onur Mutlu
Executive Summary

- **Motivation**: DRAM-based true random number generators (TRNGs) provide **true random numbers at low cost** on a **wide range** of computing systems.

- **Problem**: Prior DRAM-based TRNGs are slow:
  1. Based on fundamentally slow processes → **high latency**
  2. Cannot effectively harness entropy from DRAM rows → **low throughput**

- **Goal**: Develop a **high-throughput** and **low-latency** TRNG that uses **commodity DRAM** devices.

- **Key Observation**: Carefully engineered sequence of DRAM commands can activate **four DRAM rows** → QUadruple ACtivation (QUAC)

- **Key Idea**: Use QUAC to activate DRAM rows that are initialized with **conflicting data** (e.g., two ‘1’s and two ‘0’s) to generate random values.

- **QUAC-TRNG**: DRAM-based TRNG that generates true random numbers at **high-throughput** and **low-latency** by repeatedly performing QUAC operations.

- **Results**: We evaluate QUAC-TRNG using **136** real DDR4 chips
  1. **5.4 Gb/s** maximum (**3.4 Gb/s** average) TRNG throughput per DRAM channel
  2. Outperforms existing DRAM-based TRNGs by **15.08x** (base), and **1.41x** (enhanced)
  3. Low TRNG latency: **256-bit RN in 274 ns**
  4. Passes **all 15** NIST randomness tests
Outline

True Random Numbers in DRAM

DRAM Organization and Operation

QUadruplet ACtivation (QUAC)

QUAC-TRNG

Evaluation
Use Cases of True Random Numbers

High-quality true random numbers are critical to many applications

True random numbers can only be obtained by sampling random physical processes

Unfortunately, not all computing systems are equipped with TRNG hardware (e.g., dedicated circuitry)
DRAM-Based TRNGs

**DRAM chips** are ubiquitous in modern computing platforms

DRAM-based TRNGs enable true random number generation within DRAM chips

**Low-cost:** No specialized circuitry for RNG
- Beneficial for constrained systems

**High-throughput:** > Gb/s throughput
- Open application space that require high-throughput TRNG
Synergy with Processing-in-Memory

Processing-in-Memory (PIM) Systems

• Perform computation directly within a memory chip
• Improve system performance by avoiding off-chip data movement

True random number generation within DRAM

• Enables PIM workloads to sample true random numbers directly within the memory chip
• Avoids inefficient communication to other possible off-chip TRNG sources, enhances security & privacy
Outline

True Random Numbers in DRAM

DRAM Organization and Operation

QUadruple ACtivation (QUAC)

QUAC-TRNG

Evaluation
DRAM Organization
Accessing a DRAM Cell

- wordline
- capacitor
- access transistor
- enable
- Sense Amp
- bitline
Accessing a DRAM Cell

1. Enable wordline

2. Connects cell to bitline

3. Cell loses charge to bitline

4. Deviation in bitline voltage

5. Enable sense amp

6. Cell loses charge to bitline

- Capacitor
- Access transistor
- Sense Amp

$V_{DD}$ $V_{DD} + \delta$

$\frac{1}{2} V_{DD}$
DRAM Operation

DRAM Command Sequence

$t_{RAS}$ (Activation Latency)
$t_{RP}$ (Precharge Latency)

[Kim+ HPCA’19]
Outline

True Random Numbers in DRAM

DRAM Organization and Operation

QUadruple ACtivation (QUAC)

QUAC-TRNG

Evaluation
Quadruple Activation (QUAC)

New Observation

Carefully-engineered DRAM commands can activate four rows in real DRAM chips

Activate four rows with two ACT commands
Quadruple Activation (QUAC)

**Characteristic 1**

Activates a set of four DRAM rows whose addresses differ only in their two LSBs.
Quadruple Activation (QUAC)

Characteristic 2

First and second ACT's addresses must have their two LSBs inverted

- 2 011
- 1 010
- 2 001
- 1 000

ACT 000 PRE ACT 011 ACT 010 PRE ACT 001
QUAC on Real DRAM Chips

Valid QUAC behavior on 136 DDR4 chips
Why Does QUAC Work?

Hypothetical circuit to explain QUAC

Hierarchical Wordlines

• High density and performance requirements

• Hierarchical organization of DRAM wordlines enable high-density and low-latency DRAM operation
Hierarchical Wordlines

A master wordline drives multiple local wordlines
Hierarchical Wordlines

A master wordline drives multiple local wordlines

Select signals for rows 0-3
Hypothetical Row Decoder

Predecode the least significant two bits

Drive control signals

Address [0]

Latch

\( \overline{A0} \)

\( \overline{A1} \)

Address [1]

\( A0 \)

\( A1 \)

\( \overline{A0} \)

\( \overline{A1} \)

\( A0 \)

\( A1 \)

\( A0 \)

\( A1 \)
Hypothetical Row Decoder

Command

ACT R0

First ACT command drives a single wordline
Hypothetical Row Decoder

Command | ACT R0 \( \xrightarrow{\text{Violate Timing}} \) PRE

\[
\begin{align*}
0 & \quad 0 \\
L & \quad \overline{A0} \\
L & \quad A0 \\
L & \quad A0 \\
L & \quad A1 \\
\end{align*}
\]

PRE command **cannot** disable latches
Hypothetical Row Decoder

Command | ACT R0 | Violate Timing | PRE | Violate Timing | ACT R3

0

Second ACT drives the **remaining three** wordlines
Hypothetical Row Decoder

Command

ACT R0 \(\overset{\text{Violate Timing}}{\Rightarrow}\) PRE \(\overset{\text{Violate Timing}}{\Rightarrow}\) ACT R3

1 \(\overset{\text{L}}\rightarrow\) \(\overline{A_0}\)

1 \(\overset{\text{L}}\rightarrow\) \(A_0\)

1 \(\overset{\text{L}}\rightarrow\) \(\overline{A_1}\)

1 \(\overset{\text{L}}\rightarrow\) \(A_1\)

\(\overline{A_0} \quad \overline{A_1}\)

\(A_0 \quad A_1\)

\(\overline{A_0} \quad A_1\)

\(A_0 \quad A_1\)

S0

S1

S2

S3

All four wordlines are enabled

Quadruple Activation
Outline

True Random Numbers in DRAM

DRAM Organization and Operation

QUadruplet ACtivation (QUAC)

QUAC-TRNG

Evaluation
Generating Random Values via QUAC

Voltage Difference

\[ V_{dd} \]
\[ V_{TH} \]
\[ 0 \]
\[ -V_{TH} \]
\[ -V_{dd} \]

Time

Ready to Sense Voltage Level

Logic-1

Enable

Sense Amplifier

R0

R1

R2

R3
Generating Random Values via QUAC

Voltage Difference

\[ V_{TH} \]

\[ 0 \]

\[ -V_{TH} \]

\[ -V_{dd} \]

\[ V_{dd}^- \]

Enable  

Sense Amplifier

\[ V_{DD}/2 \]

\[ R0 \]

\[ R1 \]

\[ R2 \]

\[ R3 \]

\[ V_{DD}/2 + \varepsilon \]

Time

ACT R0  PRE  ACT R3
Generating Random Values via QUAC

- Voltage difference
- Random perturbation
- Enable Sense Amplifiers
- Voltage Difference
- Enable
- R0
- R1
- R2
- R3
- \( V_{DD}/2 + \varepsilon \)
- \( V_{DD}/2 \)
**QUAC-TRNG**

**Key Idea:** Leverage random values on sense amplifiers generated by QUAC operations as source of entropy

![Diagram](image)

**Step 1** Initialize

**Step 2** QUAC

Random Values

Logic-1

Logic-0

DRAM Segment

Sense Amplifiers
**QUAC-TRNG**

**Key Idea:** Leverage random values on sense amplifiers generated by QUAC operations as source of entropy

---

**Step 1:** Initialize

**Step 2:** QUAC

**Step 3:** Read

**Step 4:** Post-process

![Diagram of QUAC-TRNG workflow](image-url)
Key Idea: Leverage random values on sense amplifiers generated by QUAC operations as source of entropy

Generates a 256-bit random number for every 256-bit Shannon Entropy block
Outline

True Random Numbers in DRAM

DRAM Organization and Operation

QUadruple ACTivation (QUAC)

QUAC-TRNG

Evaluation
Real Chip Characterization

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

**DDR4 SoftMC → DRAM Testing Infrastructure**

[b. FPGA Board]

[a. DRAM Module]

c. PCIe Host Interface

d. Temperature Controller

[SAFARI] [kasirga] [Hassan+ HPCA’17] https://github.com/CMU-SAFARI/SoftMC
Real Chip Characterization

Measure randomness of bitstreams using **Shannon Entropy**

\[
H(x) = - \sum_{i=1}^{2} p(x_i) \log_2 p(x_i)
\]

**Calculating probabilities:**
Proportion of *logic-1* and *logic-0* values in the random bitstream

Sample each bitline following QUAC *1000 times* and calculate the bitline’s Shannon Entropy

\[
SE(1111...111) = 0
\]

\[
0 < SE(1001...010) < 1
\]
Real Chip Characterization

At 50°C and nominal voltage:

Repeatedly perform QUAC 1000 times and measure the Shannon Entropy of each bitline in 8K DRAM Segments (32K DRAM Rows), using all 16 different four-bit data patterns.

Data pattern: 1111 (four ones)

- 1 R3
- 1 R2
- 1 R1
- 1 R0

Data pattern: 1000

- 0 R3
- 0 R2
- 0 R1
- 0 R0
Calculate cache block entropy (CBE)

\[ \sum \text{all bitline entropies in the cache block} \]

Metrics based on CBE:

1. **Average CBE**: Average entropy across all cache blocks in a module

2. **Maximum CBE**: Maximum of the cache block entropies in a module
Data Pattern Dependence

Entropy varies with data pattern

Highest average entropy with pattern “0111”
Spatial Distribution

Segment entropy = \sum \text{all bitline entropies in the segment}

Segment entropy behavior is different for different modules
Spatial Distribution

Segment entropy = \( \sum \text{all bitline entropies in the segment} \)

Entropy significantly increases towards the end of the DRAM bank.
Takeaways: QUAC Entropy

We observe that entropy resulting from QUAC operations changes according to the

- data pattern used in initialization
- physical location of DRAM segments

attributed to:

- systematic manufacturing process variation
- design-induced variation
Two experiments:

1. Collect bitstreams by repeatedly sampling bitlines after QUAC operations
   • 1 Mb bitstreams
   • Post-processing: Von Neumann Corrector

2. Collect bitstreams using QUAC-TRNG
   • 1 Gb bitstreams
   • Post-processing: SHA-256
QUAC-TRNG’s Quality

Two experiments:

1. Collect bitstreams by repeatedly sampling bitlines after QUAC operations
   - 1 Mb bitstreams
   - Post-processing: Von Neumann Corrector

QUAC and QUAC-TRNG bitstreams pass all 15 NIST randomness tests
   - Post-processing: SHA-256
QUAC-TRNG Throughput Estimation

Estimate QUAC-TRNG’s throughput according to:

\[
\frac{256 \times SIB}{L \times 10^{-9}} \text{ bps}
\]

**SIB**: # of SHA Input Blocks in the highest-entropy segment

**L**: Latency of one QUAC operation in nanoseconds
QUAC-TRNG Throughput Estimation

Estimate QUAC-TRNG’s throughput according to:

\[
(256 \times \text{SIB}) / (L \times 10^{-9}) \text{ bps}
\]

**SIB:** # of SHA Input Blocks in the highest-entropy segment

**L:** Latency of one QUAC operation in nanoseconds

256-bit Entropy Blocks

\[
\sum \text{cache block entropy} = 256
\]
QUAC-TRNG Throughput Estimation

Estimate QUAC-TRNG’s throughput according to:

\[(256 \times SIB)/(L \times 10^{-9})\] bps

**SIB**: # of SHA Input Blocks in the highest-entropy segment

**L**: Latency of one QUAC operation in nanoseconds
QUAC-TRNG Throughput Estimation

Estimate QUAC-TRNG’s throughput according to:

\[
\frac{(256 \times SIB)}{(L \times 10^{-9})} \text{ bps}
\]

**SIB**: # of SHA Input Blocks in the highest-entropy segment

**L**: Latency of one QUAC operation in nanoseconds
QUAC-TRNG Configurations

1. **One Bank**
   - Use a single DRAM bank

2. **BGP**
   - Bank Group-Level Parallelism
   - Use four banks from different bank groups

3. **RC + BGP**
   - RowClone + BGP
   - Use in-DRAM copy to initialize DRAM rows and use four banks from different bank groups

---

**DRAM Subarray**

- **Source Row**
- **Destination Row**

**Activate Source**

- **Precharge**

**Activate Destination**

---

[Seshadri+ MICRO’13] [Gao+ MICRO’19]
QUAC-TRNG Throughput

Achieves 3.44 Gb/s throughput per DRAM channel on average across all modules

In-DRAM initialization greatly improves throughput
QUAC-TRNG vs State-Of-The-Art

High-throughput DRAM-based TRNGs:
- **D-RaNGe**: Activation latency failures
- Talukder et. al: Precharge latency failures

Calculate throughput by tightly scheduling the DDR4 commands required to induce failures

Evaluate two versions of these past two works:
- **Base**: As proposed
- **Enhanced (Fair)**: Throughput-optimized (SHA-256)

Assume four-channel DDR4 memory

[Kim+, HPCA’19] [Talukder+, IEEE Access 2019]
QUAC-TRNG vs State-Of-The-Art

Outperforms best prior DRAM-based TRNG
(i) “base” by **15.08x** at 2.4 GT/s
(ii) “enhanced” by **2.03x** at 12 GT/s
More in the Paper

• NIST randomness tests results

• Throughput & latency comparison against four other DRAM-based TRNGs

• System Integration
  - How QUAC-TRNG can be implemented in real systems
  - System performance study
    • QUAC-TRNG’s throughput with concurrently running applications
  - Area overhead: 0.04% of a contemporary CPU (7 nm)
  - Memory overhead: 0.002% of an 8 GiB DRAM module

• Sensitivity Analysis
  - Effect of temperature on QUAC’s entropy
    • Entropy changes with temperature
  - Time dependence study
    • Entropy remains stable for at least up to a month
Executive Summary

- **Motivation**: DRAM-based true random number generators (TRNGs) provide true random numbers at low cost on a wide range of computing systems.

- **Problem**: Prior DRAM-based TRNGs are slow:
  1. Based on fundamentally slow processes → high latency
  2. Cannot effectively harness entropy from DRAM rows → low throughput

- **Goal**: Develop a high-throughput and low-latency TRNG that uses commodity DRAM devices

- **Key Observation**: Carefully engineered sequence of DRAM commands can activate four DRAM rows → QUadruple ACtivation (QUAC)

- **Key Idea**: Use QUAC to activate DRAM rows that are initialized with conflicting data (e.g., two ‘1’s and two ‘0’s) to generate random values

- **QUAC-TRNG**: DRAM-based TRNG that generates true random numbers at high-throughput and low-latency by repeatedly performing QUAC operations

- **Results**: We evaluate QUAC-TRNG using 136 real DDR4 chips
  1. 5.4 Gb/s maximum (3.4 Gb/s average) TRNG throughput per DRAM channel
  2. QUAC-TRNG has low TRNG latency: 256-bit RN in 274 ns
  3. Outperforms existing DRAM-based TRNGs by 15.08x (base), and 1.41x (enhanced)
  4. QUAC-TRNG passes all 15 NIST randomness tests
QUAC-TRNG

High-Throughput True Random Number Generation Using Quadruple Row Activation in Real DRAM Chips

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

SAFARI    kasırga

ETH Zurich    TOBB ETÜ University of Economics & Technology    UNIVERSITY OF TORONTO
Spatial Distribution

Cache block entropy is the highest around the middle of the DRAM segment.
## Table 1: NIST STS Randomness Test Results

<table>
<thead>
<tr>
<th>NIST STS Test</th>
<th>VNC* (p-value)</th>
<th>SHA-256 (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>monobit</td>
<td>0.430</td>
<td>0.500</td>
</tr>
<tr>
<td>frequency_within_block</td>
<td>0.408</td>
<td>0.528</td>
</tr>
<tr>
<td>runs</td>
<td>0.335</td>
<td>0.558</td>
</tr>
<tr>
<td>longest_run_ones_in_a_block</td>
<td>0.564</td>
<td>0.533</td>
</tr>
<tr>
<td>binary_matrix_rank</td>
<td>0.554</td>
<td>0.548</td>
</tr>
<tr>
<td>dft</td>
<td>0.538</td>
<td>0.364</td>
</tr>
<tr>
<td>non_overlapping_template_matching</td>
<td>&gt;0.999</td>
<td>0.488</td>
</tr>
<tr>
<td>overlapping_template_matching</td>
<td>0.513</td>
<td>0.410</td>
</tr>
<tr>
<td>maurers_universal</td>
<td>0.493</td>
<td>0.387</td>
</tr>
<tr>
<td>linear_complexity</td>
<td>0.483</td>
<td>0.559</td>
</tr>
<tr>
<td>serial</td>
<td>0.355</td>
<td>0.510</td>
</tr>
<tr>
<td>approximate_entropy</td>
<td>0.448</td>
<td>0.539</td>
</tr>
<tr>
<td>cumulative_sums</td>
<td>0.356</td>
<td>0.381</td>
</tr>
<tr>
<td>random_excursion</td>
<td>0.164</td>
<td>0.466</td>
</tr>
<tr>
<td>random_excursion_variant</td>
<td>0.116</td>
<td>0.510</td>
</tr>
</tbody>
</table>

*VNC: Von Neumann Corrector
System Performance Study

The maximum throughput QUAC-TRNG provides without reducing the total off-chip memory bandwidth

Ramulator: 3.2 GHz core, four-channel DDR4 memory

QUAC-TRNG achieves 74.13% of the empirical average throughput
## Table 2: Summary of prior DRAM-TRNGs vs QUAC-TRNG

<table>
<thead>
<tr>
<th>Proposal</th>
<th>Entropy Source</th>
<th>TRNG Throughput</th>
<th>256-bit TRNG Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td>QUAC-TRNG</td>
<td>Quadruple ACT</td>
<td>13.76 Gb/s</td>
<td>274 ns</td>
</tr>
<tr>
<td>Talukder+ [15]</td>
<td>Precharge Failure</td>
<td>0.68 - 6.13 Gb/s</td>
<td>249 ns - 201 ns</td>
</tr>
<tr>
<td>D-RaNGe [88]</td>
<td>Activation Failure</td>
<td>0.92 - 9.73 Gb/s</td>
<td>260 ns - 36 ns</td>
</tr>
<tr>
<td>D-PUF [150]</td>
<td>Retention Failure</td>
<td>0.20 Mb/s</td>
<td>40 s</td>
</tr>
<tr>
<td>DRNG [47]</td>
<td>DRAM Start-up</td>
<td>N/A</td>
<td>700 μs</td>
</tr>
<tr>
<td>Keller+ [81]</td>
<td>Retention Failure</td>
<td>0.025 Mb/s</td>
<td>40 s</td>
</tr>
<tr>
<td>Pyo+ [126]</td>
<td>DRAM Cmd Schedule</td>
<td>2.17 Mb/s</td>
<td>112.5 μs</td>
</tr>
</tbody>
</table>
Temperature Dependence

Figure 14: Maximum and average segment entropy at different temperatures.
## DDR4 Modules

<table>
<thead>
<tr>
<th>Module</th>
<th>Module Identifier</th>
<th>Chip Identifier</th>
<th>Freq. (MT/s)</th>
<th>Organization</th>
<th>Segment Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Size (GB)</td>
<td>Chips</td>
</tr>
<tr>
<td>M1</td>
<td>Unknown</td>
<td>H5AN4G8NAFR-TFC</td>
<td>2133</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>M2</td>
<td>Unknown</td>
<td>Unknown</td>
<td>2133</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>M3</td>
<td>Unknown</td>
<td>H5AN4G8NAFR-TFC</td>
<td>2133</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>M4</td>
<td>76TT21NUS1R8-4G</td>
<td>H5AN4G8NAFR-TFC</td>
<td>2133</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>M5</td>
<td>Unknown</td>
<td>T4D5128HT-21</td>
<td>2133</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>M6</td>
<td>TLRD44G2666HC18F-SBK</td>
<td>H5AN4G8NMFR-VKC</td>
<td>2666</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>M7</td>
<td>TLRD44G2666HC18F-SBK</td>
<td>H5AN4G8NMFR-VKC</td>
<td>2666</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>M8</td>
<td>TLRD44G2666HC18F-SBK</td>
<td>H5AN4G8NMFR-VKC</td>
<td>2666</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>M9</td>
<td>TLRD44G2666HC18F-SBK</td>
<td>H5AN4G8NMFR-VKC</td>
<td>2666</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>M10</td>
<td>TLRD44G2666HC18F-SBK</td>
<td>H5AN4G8NMFR-VKC</td>
<td>2666</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>M11</td>
<td>TLRD44G2666HC18F-SBK</td>
<td>H5AN4G8NMFR-VKC</td>
<td>2666</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>M12</td>
<td>TLRD44G2666HC18F-SBK</td>
<td>H5AN4G8NMFR-VKC</td>
<td>2666</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>M13</td>
<td>KSM32RD8/16HDR</td>
<td>H5AN4G8NAFA-UHC</td>
<td>2400</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>M14</td>
<td>F4-2400C17S-8GNT</td>
<td>H5AN4G8NMFR-UHC</td>
<td>2400</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>M15</td>
<td>F4-2400C17S-8GNT</td>
<td>H5AN4G8NMFR-UHC</td>
<td>3200</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>M16</td>
<td>KSM32RD8/16HDR</td>
<td>H5AN8G8NDJR-XNC</td>
<td>3200</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>M17</td>
<td>KSM32RD8/16HDR</td>
<td>H5AN8G8NDJR-XNC</td>
<td>3200</td>
<td>16</td>
<td>8</td>
</tr>
</tbody>
</table>

†The maximum possible entropy in a DRAM segment is 64K (65,536) bits.