Rubix: Reducing the Overhead of Secure Rowhammer Mitigations via Randomized Line-to-Row Mapping

Anish Saxena
anish.saxena@cc.gatech.edu
Georgia Institute of Technology
Atlanta, USA

Saurav Mathur
smathur44@gatech.edu
Georgia Institute of Technology
Atlanta, USA

Moinuddin Qureshi
moin@gatech.edu
Georgia Institute of Technology
Atlanta, USA

Abstract

Modern systems mitigate Rowhammer using victim refresh, which refreshes neighbours of an aggressor row when it encounters a specified number of activations. Unfortunately, complex attack patterns like Half-Double break victim-refresh, rendering current systems vulnerable. Instead, recently proposed secure Rowhammer mitigations perform mitigative action on the aggressor rather than the victims. Such schemes employ mitigative actions such as row-migration or access-control and include AQUA, SRS, and Blockhammer. While these schemes incur only modest slow-downs at Rowhammer thresholds of few thousand, they incur prohibitive slow-downs (15%-600%) for lower thresholds that are likely in the near future. The goal of our paper is to make secure Rowhammer mitigations practical at such low thresholds.

Our paper provides the key insights that benign application encounter thousands of hot rows (receiving more activations than the threshold) due to the memory mapping, which places spatially proximate lines in the same row to maximize row-buffer hitrate. Unfortunately, this causes row to receive activations for many frequently used lines. We propose Rubix, which breaks the spatial correlation in the line-to-row mapping by using an encrypted address to access the memory, reducing the likelihood of hot rows by 2 to 3 orders of magnitude. To aid row-buffer hits, Rubix randomizes a group of 1-4 lines. We also propose Rubix-D, which dynamically changes the line-to-row mapping. Rubix-D minimizes hot-rows and makes it much harder for an adversary to learn the spatial neighbourhood of a row. Rubix reduces the slow-down of AQUA (from 15% to 1%), SRS (from 60% to 2%), and Blockhammer (from 600% to 3%) while incurring a storage of less than 1 Kilobyte.

CCS Concepts: • Security and privacy → Systems security; Hardware security implementation.

Keywords: DRAM, Rowhammer, Memory Mapping

ACM Reference Format:

1 Introduction

Rowhammer is a data-disturbance error where frequently activating a row induces bit flips in nearby rows [24]. Rowhammer is a severe security threat and has been used to leak confidential data and escalate privilege [3, 6–9, 13, 27, 27, 44, 48]. Rowhammer worsens with higher memory density. The number of activations required to induce bit-flips, termed as the Rowhammer Threshold ($T_{RH}$), has plummeted from 139K (DDR3) in 2014 to just 4.8K in 2020 (LPDDR4), as shown in Figure 1 (a). The threshold is expected to reduce even further, and if the current trend continues (30X reduction in 6 years), we can expect $T_{RH}$ of about 100 over the next decade. Solutions that protect against Rowhammer must be viable not just at the current threshold, but also at future thresholds.

Rowhammer defenses typically incorporate a tracking mechanism to count row activations and a mitigative action to perform when the activation count reaches the threshold. The most popular form of mitigative action is victim refresh, which simply refreshes the nearby victim rows when the aggressor row reaches the specified threshold of activations. Victim refresh has been deployed in commercial systems (e.g. DDR4, DDR5) in the form of Target Row Refresh (TRR) [7]. However, the drastic reduction of $T_{RH}$ poses two problems. First, due to severe area limitation in DRAM (9% area required for per-row tracking [11]), TRR is unable to identify all aggressors, even in DDR5 [11]. In fact, two recent whitepapers from JEDEC[15, 16] mention that “in-DRAM mitigations cannot eliminate all forms of Rowhammer attacks”. Second, even if tracking is perfect, the act of victim-refresh can itself be used to induce bit-flips. As shown in Figure 1 (b), Half-Double [2] leverages victim-refresh to cause bit-flips at a distance-of-2 from the aggressor row, thereby breaking all defenses relying on victim-refresh. Thus, current systems remain vulnerable to Rowhammer. In this paper, we focus on mitigations resilient to complex patterns.
Recent studies [43, 53, 54] propose such resilient mitigations that are aggressor-focused instead of being victim-focused (such as victim-refresh). AQUA [43] and SRS [53] migrate the aggressor row to another row, breaking the spatial correlation between the aggressor and victim. Blockhammer [54] limits the number of activations to any row to less than $T_{RH}$, preventing large number of activations to a row (critical for complex attacks). While these schemes invoke high-overhead mitigating actions that take several microseconds or more, at current $T_{RH}$ only a small number of rows require any mitigation, and these schemes incur a modest slowdown. Unfortunately, as $T_{RH}$ reduces, many more rows reach the threshold, requiring mitigations, which causes significantly higher overheads. Figure 1 (c) shows the average slowdown of AQUA, SRS, and Blockhammer as the threshold reduces from 1K to 128. At a threshold of 128, AQUA suffers a slowdown of 15%, SRS of 60%, and Blockhammer of 600%, rendering them impractical at lower thresholds.

The goal of our paper is to make secure Rowhammer mitigations practical even at a low threshold (128), as such thresholds can occur if the trend holds for the next decade.

High slowdown occurs due to the dramatic increase in number of rows that receive more than $T_{RH}$ activations in 64ms, which we define as hot rows. In our evaluations, we observe only about 200 hot-rows, on average, with 512 or more activations, but 9500 hot-rows with 64 or more activations (45X more). Reducing the number of hot-rows would reduce the slowdown stemming from secure mitigations.

We make the key observation that hot rows are primarily caused by the memory mapping function, which determines the set of lines co-residing within the same row. The memory-mapping in modern processors places lines with spatial proximity in the same row to maximize row-buffer hits. For example, Intel Coffee Lake [49] mapping places the entire 4KB page within the same row and Intel Skylake [49] round-robin the lines of each 4KB page between rows of two banks. Thus, 32-64 lines of each 4KB page co-reside within the same row, and if the page is heavily accessed, these lines would contribute to the aggregate activation count of the row. While each line incurs only a few activations, the sum of activations due to all the lines makes the row a hot-row.

Typical workloads access only a small fraction of the memory within 64ms. In our evaluations, less than 5% of rows are touched within 64ms. Thus, spreading activations from hot-rows to the entire memory would greatly reduce hot-rows. With this insight, we propose Rubix, a memory mapping that breaks spatial correlation of lines to rows by using an encrypted address to access memory. We present two flavors of Rubix: Static (Rubix-S) and Dynamic (Rubix-D).

Rubix-S uses the low-latency programmable bit-width K-Cipher [26] for address-space randomization, which is kept in the memory controller. On a memory access, the memory controller encrypts the line-address, accessing the memory with the encrypted line address. Encryption randomizes the line-to-row mapping, so the lines co-resident in the same row have no spatial correlation. This significantly reduces the likelihood of heavily accessed lines getting placed in the same row, virtually eliminating all the hot-rows. As a result, mitigations are invoked much less, reducing slowdown. To preserve row buffer hit rate, Rubix-S encrypts a gang of 1-4 contiguous lines, as line-level address encryption would result in virtually zero row buffer hits. Our evaluations show that at $T_{RH}$ of 128, Rubix-S reduces the slowdown of AQUA (from 15% to 1%), SRS (from 60% to 3%), and Blockhammer (from 600% to 3%) while requiring just 16 bytes of storage, thereby making it practical to deploy secure mitigations.

With Rubix-S, the group of lines that co-reside in the row are randomized, however, this group remains unchanged throughout the system uptime. To this end, Rubix-D provides dynamic randomization of line-to-row mapping without needing a programmable cipher, by using an xor operation with a randomly generated key. The mapping changes gradually from the current-key to the next-key. Rubix-D remaps vertically (gangs in the same position of different rows) instead of horizontally (gangs within the row), which not only reduces hot-rows, but also makes it much harder for an adversary to determine set of spatially contiguous rows, a critical step in launching a targeted Rowhammer attack. Our evaluations show that at $T_{RH}$ of 128, Rubix-D reduces the slowdown of AQUA (from 15% to 1.5%), SRS (from 60% to 2%), and Blockhammer (from 600% to 3%) while incurring a storage overhead of less than 1 KB.
2 Background and Motivation

2.1 Threat Model

We assume an unprivileged attacker that can run code on the system vulnerable to Rowhammer. The attacker can run a process under user privilege and exploit Rowhammer to flip bits in critical data structures (such as page-table) or in the data of another program. We assume the Rowhammer bit-flip occurs at the victim location when any row in memory incurs more activations than $T_{RH}$ within the refresh interval of 64ms. Thus, the attack is successful if no mitigation is issued when a row has encountered more than $T_{RH}$ activations.

2.2 Background on DRAM

Modern DRAM-based memory is organized into several banks, each of which is a two-dimensional array of DRAM cells, organized as rows and columns. Each bank caches the most recently opened row in a row buffer. Data is accessed by bringing it into the row buffer. To access data in another row, the bank clears the row buffer, followed by activation of the given row. DRAM cells leak charge and require periodic refresh operations (at 64ms). $T_{RC}$ determines the time between consecutive activations for a given bank and is about 45ns.

2.3 Memory Mapping

The memory-mapping function routes a given line address to a particular bank and row, determining the set of lines that co-reside in a row [10, 49]. It also affects row-buffer hit-rate and performance. Memory systems place spatially proximate lines in the same row and we consider two mappings used in Intel systems. While not exhaustive, we note that most deployed mappings use similar xor-based hashing [49].

Coffee Lake Mapping: This mapping places consecutive 128 lines within the same row buffer. So, two consecutive 4KB pages would be resident in the same row. It uses a xor-based hashed mapping for bank selection.

Skylake Mapping: This mapping alternatively places a pair of lines between two banks (selected using xor). So, for a 4KB page, lines 0,1,4,5 ... 60, 61 reside in a row of one bank, and lines 2,3,6,7 ... 62,63 are in row of the other bank. This mapping causes 32 lines from a 4KB page to reside in a row, with contents of four consecutive pages in the same row.

2.4 Rowhammer

Rowhammer is a data-disturbance error [17, 24] where activating a row frequently induces bit-flips in nearby rows. The Rowhammer Threshold ($T_{RH}$) denotes minimum activations required on a row to induce bit-flips with any access pattern. Rowhammer is a severe security threat as the attacker can flip bits in the page table and take over the system. When Rowhammer was characterized in 2014, $T_{RH}$ was 139K with single-sided attack, whereas it reduced by 30x to 4.8K [19] in 2020 (with double-sided attack). As established by a decade of threshold characterization [20], ultra-low thresholds will be reached by the next decade and as memory gets denser, more nearby rows experience aggressor activations [28].

We can avoid ultra-low thresholds if DRAM organization changes fundamentally or DRAM vendors mitigate Rowhammer. Unfortunately, neither option has materialized, as stated by JEDEC [15, 16] and recent industry papers [11, 23]. Moreover, as systems remain deployed for several years, Rowhammer defenses must scale to future $T_{RH}$.

Hardware-based defenses for Rowhammer have two parts: activation-tracker and mitigating-action. Several studies [4, 22, 29, 36, 37, 46, 47, 56] have looked at storage-efficient trackers. Comparatively, mitigating action is less well studied. Most modern systems simply rely on victim-refresh, which is vulnerable to address-correlation attacks [25]. Thus, modern systems continue to be vulnerable to Rowhammer attacks.

2.5 Secure Rowhammer Mitigation

Performing mitigative action on aggressor row, instead of the victim, prevents address-correlation attacks. Figure 2 shows three such recently proposed aggressor-focused mitigation.

Blockhammer [54] controls the access rate of frequently accessed rows, such that no row incurs more than $T_{RH}/2$ activations within 64ms, by delaying accesses for an appropriate time. As the adversary cannot perform an overwhelmingly large number of activations (required for Half-Double) on a single row, it prevents complex attacks.

AQUA [43] migrates the aggressor row when it receives $T_{RH}/2$ activations (halving of threshold due to tracker reset), to a quarantine-region in memory. AQUA breaks the spatial connection between aggressor and victim, limiting the time for an attacker to craft complex attacks.

Secure Row-Swap (SRS) [53] swaps the aggressor row, once it has received $T_{RH}/3$ activations (reduction due to birthday-paradox attacks), with another randomly selected row in memory. Like AQUA, SRS breaks the spatial connection between the aggressor and the victim.
2.6 Scalability Problem of Secure Mitigations

Secure Rowhammer mitigations (such as Blockhammer, AQUA, and SRS) incur significantly more overhead than victim-refresh. While performing two victim-refresh activation takes less than 100 nanoseconds, these schemes incur significantly more latency. For example, row migration in AQUA and SRS ties up the memory bus for several microseconds, during which the channel cannot service any requests. The problem is even worse for Blockhammer, as rate control can delay accesses by several tens to hundreds of microseconds.

These schemes are designed for the current thresholds of few thousands, where very few rows reach the threshold in benign workloads and require mitigation. However, at lower thresholds, many more rows reach the threshold, which causes more frequent high-overhead mitigations, causing drastic slowdown. Figure 3 shows the performance of AQUA, SRS, and Blockhammer as thresholds ranging from 1K to 128, for the Coffee Lake and Skylake memory mappings, normalized to Coffee Lake-based baseline.

At the threshold of 1K, AQUA and SRS have negligible slowdown, whereas Blockhammer suffers 10% (Coffee Lake) to 25% (Skylake). However, at $T_{RH}$ of 128, all schemes incur significant overheads. AQUA and SRS incur 15% and 60% slowdown, respectively, and Blockhammer has 500% to 600% slowdown (note that normalized IPC of 0.2 implies 5x slowdown), making secure mitigations impractical for adoption.

3.2 Workloads

We evaluate with 18 SPEC2017 [1] rate workloads and 16 mixed workloads (each with four random SPEC2017 workloads). We fast-forward 25 billion instructions and simulate for 250 million instructions. Table 2 shows the Misses Per 1K Instructions (MPKI) and the average number of unique rows touched, and "hot-rows" with 64 or more activations (ACT-64+) and with 512 or more activations (ACT-512+).
Figure 4: Illustration: Understanding the impact of memory-mapping in generating hot-rows (a) System configuration (b) Workloads (c) Number of hot-rows for 4MB footprint (and 4KB rows). Under baseline mapping, both stride-64 and random have 1K hot rows (100%), however, with an encrypted address virtually all the hot-rows are eliminated.

Table 2: Workloads Characteristics: MPKI, Unique Rows Touched (within 64ms), and Hot-Rows (within 64ms).

<table>
<thead>
<tr>
<th>Workload</th>
<th>MPKI (LLC)</th>
<th>Unique Rows Activated</th>
<th>Total number of Hot-Rows</th>
</tr>
</thead>
<tbody>
<tr>
<td>blender</td>
<td>12.78</td>
<td>8.8K</td>
<td>347K, 2.9K</td>
</tr>
<tr>
<td>ibm</td>
<td>20.87</td>
<td>29.4K</td>
<td>70.3K</td>
</tr>
<tr>
<td>gcc</td>
<td>6.12</td>
<td>10.4K</td>
<td>21.8K, 384</td>
</tr>
<tr>
<td>cactuBSSN</td>
<td>2.57</td>
<td>5.2K</td>
<td>12.2K</td>
</tr>
<tr>
<td>mcf</td>
<td>5.81</td>
<td>4.9K</td>
<td>10.5K, 425</td>
</tr>
<tr>
<td>roms</td>
<td>3.33</td>
<td>27.9K</td>
<td>6.6K</td>
</tr>
<tr>
<td>perbench</td>
<td>0.71</td>
<td>11.4K</td>
<td>1.7K</td>
</tr>
<tr>
<td>xx</td>
<td>0.40</td>
<td>10.8K</td>
<td>496</td>
</tr>
<tr>
<td>nab</td>
<td>0.53</td>
<td>4.4K</td>
<td>189</td>
</tr>
<tr>
<td>namd</td>
<td>0.37</td>
<td>3.4K</td>
<td>105</td>
</tr>
<tr>
<td>imagick</td>
<td>0.13</td>
<td>1.1K</td>
<td>89</td>
</tr>
<tr>
<td>bwaves</td>
<td>0.21</td>
<td>1.7K</td>
<td>20</td>
</tr>
<tr>
<td>wrf</td>
<td>0.02</td>
<td>70Z</td>
<td>20</td>
</tr>
<tr>
<td>exchange2</td>
<td>0.01</td>
<td>122</td>
<td>14</td>
</tr>
<tr>
<td>deepsjeng</td>
<td>0.25</td>
<td>68.1K</td>
<td>12</td>
</tr>
<tr>
<td>povray</td>
<td>0.01</td>
<td>390</td>
<td>8</td>
</tr>
<tr>
<td>parest</td>
<td>0.10</td>
<td>2.4K</td>
<td>3</td>
</tr>
<tr>
<td>leela</td>
<td>0.02</td>
<td>879</td>
<td>0</td>
</tr>
<tr>
<td>Average</td>
<td>3.01</td>
<td>10.7K</td>
<td>9528, 206</td>
</tr>
</tbody>
</table>

4. A Case for Randomized Memory

The reason secure Rowhammer mitigations incur significant overheads at low thresholds is because more rows reach the threshold number of activations (we refer to such rows as hot-rows). We identify the root cause of hot-rows to be the memory mapping function that determines the line-to-row mapping. In this section, we first present this insight, then our workload characterization, then our solution Rubix, and results for slowdown and mitigation.

4.1 Dependence of "Hot Rows" on Mapping

We illustrate the dependence of hot-rows on line-to-row mapping using a simple model, as shown in Figure 4 (a). The processor accesses a memory system containing one bank. The memory system is 4GB and contains 1 million rows of 4KB each. We use sequential mapping that places the 4KB page within the same row.

We consider three kernels as shown in Figure 4 (b): stream, stride-64, and random, with a 4MB footprint. Each kernel incurs 1 million memory accesses within 64ms. We deem a row to be a hot-row if it has at least 64 activations. We analyze the number of hot-rows for these kernels.

For the stream kernel, the first access causes an activation, and subsequent 63 accesses get a row-buffer hit. Therefore, a million memory accesses cause a total of 15.6K activations, which get spread over the 1K rows, with a uniform activation rate of about 16 activations per row, with no hot-rows. The stride-64 kernel has a stride of 64 lines and each access goes to a different page. When all pages are exhausted, the stride continues with the next line on the page. As each memory access causes an activation, this kernel incurs 1 million activations, spread equally over 1K pages, and each row gets 1K activations. Thus, all the 1K rows are hot-rows. The random kernel accesses a random line in memory. The likelihood of a row buffer hit is negligibly small, so the 1 million accesses cause 1 million activations, spread over 1K rows. The average activations per row are 1000 (standard deviation of 32), with more than 99% of the rows having more than 900 activations. Thus, we deem all the 1K rows to be hot-rows. The results are summarized in Figure 4 (c).

The conventional mapping of placing sequential lines in the same row buffer causes hot-rows for both the stride pattern and the random pattern. We have 64 lines that cause activation of the same row in memory, thus compounding the total number of activations incurred by the given row.

Consider an alternative mapping that uses an encrypted line-address to access the memory system. There are 64K lines in a 4MB footprint. These 64K lines would be spread over the 1 million rows in memory. We estimate (using binomial distribution) that 61.5K rows have exactly 1 line from the kernel mapped to them, 1.9K rows with 2 lines, and 40 rows with 3 lines (no row with 4 or more lines). For both stream and stride, each line gets accessed 16 times. So, we have 61.5K rows with 16 activations, 1.9K rows with 32 activations, and 40 rows with 48 activations. Thus, no row is deemed a hot-row. For random, we estimate the expected number of hot-rows to be 0.4, so less than 1 row will be deemed a hot-row. Thus, randomizing the line-to-row mapping eliminates the hot rows of all three kernels.
4.2 Characterizing Lines in Hot-Rows

For our baseline system, we examine how many lines (out of the 128 lines) of the row contribute to making the row a hot-row. For each row that reaches 64 activations, we measure the number of lines in the row that encountered at least 1 activation. Table 3 shows the percentage of hot-rows that had 1-8 lines, 8-16 lines, 32-64 lines, and 64-128 lines (and the average) contributing to the row activation counts.

Table 3. Number of lines that add to activation counts of hot-rows (data for workloads with 100+ hot-rows).

<table>
<thead>
<tr>
<th>Workload</th>
<th>Number of Activating Lines in a Hot-Row</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1-32</td>
</tr>
<tr>
<td>blender</td>
<td>2%</td>
</tr>
<tr>
<td>ibm</td>
<td>0</td>
</tr>
<tr>
<td>gcc</td>
<td>1%</td>
</tr>
<tr>
<td>cactuBSSN</td>
<td>0</td>
</tr>
<tr>
<td>mcf</td>
<td>0</td>
</tr>
<tr>
<td>roms</td>
<td>3%</td>
</tr>
<tr>
<td>perlbench</td>
<td>7.3%</td>
</tr>
<tr>
<td>xz</td>
<td>0</td>
</tr>
<tr>
<td>nab</td>
<td>0</td>
</tr>
<tr>
<td>namd</td>
<td>0</td>
</tr>
<tr>
<td>Average</td>
<td>2%</td>
</tr>
</tbody>
</table>

We observe that for 98% of hot-row activations come from at-least 32 lines in the row. On average, 56 out of 128 lines incur at least one-activations within the hot-row. This validates our hypothesis that hot-rows occur because many lines of the row contribute to the activation counts. Thus, the line-to-row mapping which decides which set of lines co-reside within the same row is the main reason for the occurrence of hot-rows.

4.3 Rubix: Randomized Line-to-Row Mapping

Rubix breaks the spatial correlation of lines to row by using an encrypted address to access memory. Figure 5 shows an overview of the static version of Rubix, called Rubix-S. Consider the access pattern where requests for four consecutive lines A, B, C, D are set to memory. In conventional mapping, these four lines will co-reside within the same row. However, with encryption, these lines get scattered to different rows.

Rubix-S uses K-Cipher [26], a low-latency programmable bit-width cipher, for address-space randomization. K-Cipher is kept in the memory controller and incurs a latency of 3 cycles (with 10nm process technology [26]). On a memory access, it encrypts the line-address which is used to access the memory. As we have 16GB memory, we use a 28-bit cipher. Encryption randomizes the line-to-row mapping, breaking the spatial correlation between lines co-residing in the row.

The exact line-to-row mapping depends on the 96-bit key of the K-Cipher. The key is set to a random value (based on PRNG) at boot time. As each system will have a different key, the memory mapping for each system will be different.

4.4 Recouping Row-Buffer Hits via Gangs

While line-address encryption virtually eliminates hot-rows, it degrades the row-buffer hit-rate to approximately zero. Rubix minimizes hot-rows while still retaining some row-buffer hits, by encrypting a gang of 2-4 contiguous lines. Figure 6 shows Rubix-S with gang-level randomization.

Instead of encrypting the entire n-bit line-address, Rubix-S skips the k least significant bits of the line-address and only encrypts the gang-address, which is the remaining (n-k) bits. The encrypted gang-address is concatenated with the unmodified k-bits, and this line-address is used to access the memory. Thus, lines within a gang that co-reside in a row provide temporal locality to aid row-buffer hits. For example, in Figure 6, lines 1 and 2 co-reside in the same row. Note that with k-bits, we would have a gang-size of 2^k lines and if k is set to zero, this design degenerates into Rubix-S with line-address encryption. We denote Rubix-S with a gang-size of X lines as Rubix-S (GSX). The size of the cipher is adjusted per gang size, so Rubix-S (GS4) uses a 26-bit K-cipher.

4.5 Results: Impact on Mitigations

We observe that Rubix-S (GS1, line-level) eliminates all hot-rows for our workloads. Figure 7 shows the number of hot-rows for the baseline system with Coffee Lake mapping, Skylake mapping, and Rubix-S (GS4). Rubix-S eliminates hot-rows for all but six workloads. On average, Coffee Lake...
Figure 7. Number of hot-rows (activations of 64 or more) with Intel mappings and Rubix-S with Gang-Size of 4 (GS4). Mean implies arithmetic mean. While baselines have more than 7K hot rows on average, Rubix-S (GS4) reduces it by 220x to 33.

and Skylake mappings have 7.6K and 7.2K hot-rows respectively, whereas Rubix-S (GS4) reduces it by 220x to only 33. Line-to-row mapping primarily determines hot-rows, and our design significantly reduces hot-rows. Mitigations are invoked much less, greatly reducing performance overheads.

4.6 Results: Impact on Performance

Figure 8 shows performance of AQUA, SRS, and Blockhammer with Intel Coffee Lake, Skylake, and Rubix-S mappings. Performance is normalized to an unprotected Coffee Lake baseline. Intel mappings incur unacceptable slowdown with secure mitigations. We compare Rubix-S with Coffee Lake mapping which performs slightly better than Skylake. Coffee Lake incurs a significant average slowdown of 15% for AQUA, while Rubix-S reduces it to a negligible 1% (for gang-size 4). SRS and BlockHammer are impractical with baseline policies, incurring 60% and 600% average slowdown, respectively. Rubix-S not only enables SRS and BlockHammer with a negligible average slowdown of 3.1% (GS 4) and 2.8% (GS 1), respectively, it retains application-level performance with
a worst-case slowdown of 42% for lbm with SRS and just 11% for BlockHammer, 28X and 350X improvement.

Overall, Rubix-S makes secure mitigations viable at ultralow TRH of 128 with just 2-3% overhead. While we do not change access scheduling and DRAM page policies, fine-tuning them would likely reduce the overheads even further.

4.7 Sensitivity: Varying Gang-Size

Gang-size (GS) balances row-buffer hits and reduction in hot-rows. With larger GS, row-buffer hit rate increases along with hot-rows and mitigation overheads. Figure 9 shows the performance of secure mitigations with Rubix-S as GS is varied from 1 to 4. Due to high mitigation overhead, BlockHammer works best with GS1 which eliminates hot-rows. AQUA has lower overhead mitigation, so GS4 works best which retains row-buffer hits. For SRS, GS2 offers the best balance between row-buffer hits and minimizing hot-rows. Thus, the best GS size depends on the scheme and the mitigation overhead and Rubix-S provides the flexible trade-off of minimizing hot-rows while retaining row-buffer locality.

4.8 Results: Impact on Row-Buffer Hits

A key effect of small gang-size is decreased row-buffer hit rate. The baseline Coffee Lake and Skylake policies provide an average row-buffer hit rate of 55% and 63%, respectively. Rubix-S shows a gradual increase in row-buffer hit-rate from 0 with GS1, to 19% at GS2 to 31% at GS4, with up-to 2.7X more activations for GS1. Thus, GS2 and GS4 recoup some of the row-buffer hits. The overall system performance depends not only on row-buffer hits but also on mitigation overheads.

4.9 Results: Storage and Power Overheads

Rubix requires negligible power for the K-Cipher and address mapping logic. We use Micron’s power calculator [34] to compute DRAM power, the primary overhead due to lower row-buffer hit rate. Rubix-S increases the DRAM power by 120mW at a gang-size of 4 (4.3% increase), and by 300mW at gang-size of 1 (10.6% increase), due to a lower row-buffer hit rate than baseline that result in additional activations. The power consumption of Rubix-S with secure mitigations remains within 10% of the baseline, because of virtually eliminating mitigations, unlike existing memory mappings, which incur prohibitive energy overheads.

4.10 Security Analysis of Rubix-S

The security of Rubix-S stems from the security of the underlying mitigation schemes (SRS, AQUA, Blockhammer). The security guarantees of these schemes are not dependent on using a specific memory-mapping. Rubix-S remains secure because we simply change the memory mapping.

4.10.1 Defining TRH. We define TRH as the minimum number of activations to at least one row within 64ms which causes a bit flip via any attack pattern (single-sided, double-sided, Half-Double[25], or a future attack pattern). So to ensure security of our solution, our only assumption is:

A successful Rowhammer attack requires activating at least one row more than TRH times within a refresh period.

4.10.2 Security of SRS, AQUA, and BlockHammer. SRS and AQUA rely on row migration to guarantee that now row receives more than TRH activations within a 64ms window. SRS does so by randomization, guaranteeing that even under continuous attacks for several years, the likelihood of randomly finding migrated rows is negligibly small. With AQUA, a row that receives TRH activations is moved to a quarantine region, and by design, it guarantees that no physical row will ever receive more than TRH activations. BlockHammer controls the activation rates to a physical row such that no row ever receives more than TRH activations. These schemes rely on accurate tracking of row counts, and we use Misra-Gries tracker for SRS and AQUA, and one-counter-per-row for BlockHammer, which provide guaranteed tracking. The security guarantees of SRS, AQUA, Blockhammer are applicable for all access patterns (including Half Double) and all possible memory mapping (the mapping of lines to rows).

4.10.3 Proving Security of Rubix-S Using Lemmas.

Rubix-S reduces performance overheads while retaining the security guarantees of the underlying SRS, AQUA, and Blockhammer schemes. The underlying schemes (SRS, AQUA, and BlockHammer) are secure against all access patterns (including Half Double), and these guarantees work for any memory mapping. Using the randomized memory mapping of Rubix-S retains these guarantees, as we show using lemmas.

Lemma-1: The security guarantee of SRS, AQUA, and Blockhammer is not dependent on memory mapping, so these designs are secure for all memory mappings.

Lemma-2: Rubix-S is a memory mapping which randomizes the line-to-row mapping.

From Lemma-1 and Lemma-2, it follows that secure mitigations continue to be secure with Rubix-S. For example, Half Double requires that an aggressor row be activated about 100x more times than TRH. As no row is activated TRH times in secure mitigations, their security is unaffected by Rubix-S.
We perform remapping every 100 accesses. Figure 10 (a) shows the mapping at the start of the epoch with all lines using the currKey whereas by the end of the epoch, all lines use the nextKey. We perform remapping every 100 accesses. Figure 10 (a) shows the mapping at the start of the epoch with all lines located at their original address xor-ed with the currKey (010). After 100 accesses, the first remapping is invoked, so the physical location 000 (pointed by the Ptr) is swapped with the destination 110 (Ptr xor-ed with the nextKey). Ptr is incremented to 001.

The next three remappings also result in swaps (Figure 10 (b), (c), and (d)) and the pointer is incremented. For the next four remapping episodes, swapping is skipped as the Ptr points to an already remapped line. After 8 episodes, all lines use the mapping with nextKey (Figure 10(h)). Next, the currKey becomes currKey xor-ed with nextKey, and the nextKey is initialized to a new value obtained using a hardware-based PRNG. The Ptr is reset to 000, indicating a new epoch.

We translate line-address to physical-address in two steps:

1. Translate line-address \( L \) to \( L' = (L \text{ xor currKey}) \).
2. Perform two checks: First, is \( L' < \text{Ptr} \)? and Second, is \( (L' \text{ xor nextKey}) < \text{Ptr} \) ?. If either is yes, \( L' = (L' \text{ xor nextKey}) \).

The memory access is routed to location \( L' \). The simple xor and check operations are performed within one cycle. Thus, xor-based dynamic remapping randomizes line-to-row addresses with negligible SRAM (three registers – currKey, nextKey, and Ptr) and latency (one cycle). For properties and proof of xor-based randomization, please refer to [45].

5.2 Pitfall of Xor at Randomizing Line-to-Row

While xor-based mapping dynamically randomizes memory addresses, we cannot directly apply it in our context, due to the linear mapping of xor. For example, if there are 128 lines co-residing in a row, then after an xor with a random key, these 128 lines still co-reside in one row (at another location). As all the top (n-7) bits of the lines that get mapped to the same row are identical, an xor with the (n-7) bits in the key results in the same remapped value. Reordering of lines within the destination row, unfortunately, does not reduce the likelihood of it becoming a hot-row. Instead, our proposal Rubix-D reorganizes the xor-based mapping to dynamically randomize the group of lines that co-reside in a row.

5.3 Overview of Rubix-D

Figure 11 shows an overview of Rubix-D. We randomize gangs vertically (across rows but for same gang-in-row). For \( G \) gangs in a row, we provision \( G \) sets of remapping circuits (currKey, nextKey, and Ptr). As each gang in the row uses a different key, gangs co-residing in the same row in the baseline are scattered to different rows in memory, breaking the spatial correlation between gang mapping to a row.

In Figure 11, the memory has 4 gangs in a row (colored red, yellow, etc.). The same-colored gangs across all rows form a vertical-group (v-group). Each v-group is allocated a pair of keys (curr and next) and a pointer. The line-address is split into three parts: the least significant \( k \) bits identify the line-in-gang, next \( p \) bits identify the gang-in-row, and remaining \( n - p - k \) bits identify the row-address.
Rubix-D keeps the k+p bits of the line address unchanged, randomizing only the bits for (global) row address. The p bits identify the v-group and its keys and pointer translate the row-address to the remapped-row-address. The remapped-row-address is concatenated with the k+p bits to form the remapped-line-address, which is used to access the memory. With a 28-bit line address, Rubix-D with gang-size of 4 uses 2 bits to identify line-in-gang, the next 5 bits for gang-in-row, and remaining 21 bits for global row address. With less than 8 bytes for each pair of keys and ptr, we need total SRAM of 512 bytes (for 32 v-groups).

5.4 Remapping Rate and Remapping Period

Remapping Rate (RR) determines the frequency of remapping. We set RR to occur with 1% probability on each activation (thereby avoiding ACT counters for v-groups). V-gangs with more activations are remapped more frequently. During remap, the gangs pointed by the Ptr of the v-group are swapped with their destination (based on nextKey). At GS4, the memory controller streams 4 lines from source and destination rows and swaps them (open-row-X, read-DataX, open-row-Y, read-DataY, write-DataX-to-Y, open-row-X, write-DataY-to-X). Swapping incurs 3 ACTs, 8 CAS reads and 8 CAS writes, consuming bandwidth and energy. As half of the remap operations are skipped (Figure 10 (e)-(h)), at an RR of 1%, the average overhead is low at 1.5% extra activations.

Remapping Period (RP) is the time to remap the v-group. With RR=1% and two million rows in memory, a v-group has a remap-period of about 200 million activations. We can reduce the remapping-period by dividing such that every Nth row of the v-group forms a v-segment. Each v-segment has its own set of keys and pointer. With N=32, the remapping-period of the v-segment is 6.25 million activations; however, this requires 16 KB SRAM overhead for metadata.

5.5 Security Analysis of Rubix-D

Even though Rubix-D remaps dynamically, it is not a standalone mitigation for Rowhammer, as an adversary can use Flush+Reload [55] to cause bit-flips. Thus, Rubix-D must always be used with a Rowhammer mitigation scheme. Rubix-D’s security stems from the underlying mitigation (AQUA/SRS/Blockhammer). As the security of these schemes is not dependent on line-to-row mapping, Rubix-D retains their security (please see Section 4.10). Thus, per Lemma-1 and the fact that Rubix-D is simply a memory mapping, the overall design (with AQUA,SRS, Blockhammer) of Rubix-D is secure against all access patterns, including Half-Double.

5.6 Impact of Rubix-D on Future Attacks

Complex attacks, such as Half-Double and BLASTER [28], attack multiple rows and identifying spatially contiguous rows is critical for success [5]. Once inferred, the mapping remains constant in Rubix-S until system reboot, whereas with Rubix-D the neighbor information gets changed within a few seconds due to remapping. Thus, with Rubix-D, not only do we get security for known attacks, it makes orchestrating future complex pattern attacks much harder.

5.7 Results: Storage and Power Overheads

Rubix-D needs 8-byte metadata (currKey, nextKey, Ptr) for each v-group, so 512 bytes for gang-size of 4 lines. For segmented Rubix-D, the storage overhead is proportional to the number of segments (e.g., 16KB SRAM for 32 segments). DRAM power, computed using Micron’s power calculator [34], increases by 130mW at GS4 (4.2% more than baseline), 180mW at GS2 (5.8% increase), and 320mW at GS1 (10.9% increase). Note that with baseline mappings, secure mitigations would incur significant energy overheads.

5.8 Results: Impact on Mitigations

Rubix-D reduces hot-rows within 64ms as shown in Figure 12, which plots hot-rows for conventional policies, Rubix-S, and Rubix-D (as GS is varied). The baseline policies each have more than 7K hot-rows. Rubix with GS1 eliminates hot-rows which GS2 incurs a negligible number of hot-rows, which increase to few tens with GS1. The reduction in hot-rows makes secure mitigations viable at TRH of 128.

5.9 Results: Impact on Performance

We evaluate Rubix-D with Remapping-Rate of 1% without any segments as they do not impact performance (they affect the Remapping-Period and storage overheads). Figure 13 shows Hot-rows in baseline and Rubix (atleast 100x less).
shows the performance of Rubix-D compared to Intel mappings, normalized to an unprotected Coffee Lake baseline. Rubix-D incurs low overhead of just 1-3% on average at $T_{RH}$ of 128. AQUA, SRS, and BlockHammer perform best at different gang-sizes. AQUA almost achieves these benefits from row buffer locality at GS4. SRS operates at a lower threshold of $\frac{L}{3}$ and launches more mitigations, performing best at GS2 with negligible hot rows. BlockHammer has high mitigation overhead and works best minimal hot-rows at GS1. Rubix-D incurs worst-case slowdown of just 10%, compared to more than 100X in baseline (for BlockHammer). The remapping of Rubix-D also avoids getting stuck with an accidentally bad mapping, as the mapping gets changed over program execution.

### 5.10 Sensitivity: Mapping Overhead of Rubix

Table 4 shows the isolated slowdown of Rubix mappings without any mitigative action. Rubix incurs low overhead of 1%-3% due to lower row-buffer hit rate than baseline mapping. Rubix-D overheads are slightly higher than Rubix-S due to extra activations required for dynamic remapping. As randomization incurs a small performance cost while minimizing episodes of hot rows, it avoids expensive mitigations.

<table>
<thead>
<tr>
<th>Slowdown of Rubix Mapping</th>
<th>Rubix-S</th>
<th>Rubix-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gang of 4 lines (GS4)</td>
<td>1%</td>
<td>1.3%</td>
</tr>
<tr>
<td>Gang of 2 lines (GS2)</td>
<td>1.6%</td>
<td>1.9%</td>
</tr>
<tr>
<td>Gang of 1 line (GS1)</td>
<td>2.6%</td>
<td>2.7%</td>
</tr>
</tbody>
</table>

#### 5.11 Sensitivity: Higher Rowhammer Thresholds

Figure 14 shows the slowdown of secure mitigations with Rubix-S and Rubix-D at higher $T_{RH}$. Randomizing the line-to-row mapping practically eliminates hot-rows at higher thresholds, even at gang-size of 4 employed by Rubix at $T_{RH}$ of 1K for all secure mitigations. Consequently, the performance overhead is negligible 1.1% to 1.4% at $T_{RH}$ of 1K.
We evaluate Rubix and the baseline mappings with memory-intensive STREAM workloads [33] using 1 GiB arrays (LLC MPKI of more than 50). Figure 15 shows the performance of Rubix normalized to unprotected CoffeeLake and Skylake mappings. Rubix eliminates hot-rows in all STREAM workloads. On average, Rubix incurs 2% to 5% slowdown compared to CoffeeLake mapping (5% to 8% slowdown compared to Skylake mapping) due to lower row buffer hit rate (Rubix-D incurs more slowdown due to dynamic remapping). Overall, Rubix is low-cost even with memory-bound workloads.

In this section, we describe alternative designs that can reduce hot-rows without relying on a cipher or remapping.

6.1 Randomizing Line-to-Row without Cipher
Rubix-S breaks line-to-row spatial proximity via randomization. An alternative strategy to reduce hot rows is to use the most significant bits of the memory address for the gang-in-row, which stripes gangs co-resident in a row. For example, 16GB memory and 32 gangs-per-row stripes the gangs in the same row by 512MB. As lines that are much further away (512MB) from each other are unlikely to be accessed within a short time of each other, this mapping also reduces the line-to-row correlation without relying on a cipher. We also evaluated such a large-stride design and found that it has overheads similar to Rubix-S (1.8% to 3.8% slowdown with secure mitigations compared to unprotected CoffeeLake mapping). However, unlike Rubix-S, gang-level striding would not be robust against all access patterns, such as patterns with large strides, whereas, cipher-based randomization provides a principled solution for all patterns.

6.2 Randomizing Line-to-Row with Keyed XOR
Rubix-D assigns each gang-in-row to a separate remapping circuit to XOR-hash with its randomly generated key. If dynamic remapping is skipped, Rubix-D still retains static randomization while avoiding the performance and energy overheads of swapping gangs. In our evaluations, Rubix-D without dynamic remapping incurs an average slowdown of just 0.9%-2.6% with secure mitigations. The randomized mapping remains unchanged until system is rebooted (like Rubix-S). Note that static randomization virtually eliminates all hot-rows, and the additional benefit of dynamic randomization is to make targeted Rowhammer attacks difficult.
7 Related Works

7.1 Mapping of Memory Systems

Minimalist Open-Page (MOP) [18] balances performance and fairness by placing only four lines of a 4KB page in the same row. Unfortunately, as MOP round-robins across all banks, spatially proximate lines from consecutive pages co-reside in the same row, maintaining spatial correlation. We find hot-rows with MOP are similar to our baseline mapping. Figure 17 shows the normalized performance of secure mitigations with MOP, Rubix-S, and Rubix-D. We observe that MOP still suffers significant slowdowns, whereas Rubix virtually eliminates the hot-rows and the associated slowdown. Rather than hand-crafting a mapping, our work uses encryption for breaking the spatial correlation of lines.

![Figure 17. Performance of AQUA, SRS, and Blockhammer on MOP and Rubix. MOP suffers large slowdowns.](image)

7.2 Randomization in Memory Systems

Randomization is a popular technique to improve the reliability and security of memory systems. For example, Start-Gap [40] and Security-Refresh [45] randomize mapping in non-volatile memories for wear-leveing. Cache randomization [30, 38, 39, 41, 50, 51] techniques randomize the line-to-set mapping to mitigate conflict-based cache attacks.

7.3 In-DRAM Rowhammer Mitigations

DRAM modules contain Target Row Refresh (TRR), which tracks aggressors and refreshes victims. Recent attacks [7, 14], break TRR by exploiting its insufficient tracking capability to capture all possible aggressor rows. Samsung’s DSAC [11] and SK Hynix’s PAT [23] improve TRR for DDR5, but due to severe area limitation in DRAM, still allow aggressors to escape detection. DSAC has an escape probability of 13.9% between two mitigations and PAT fails 6.9% of the time (compared to DDR4-TRR). Two recent whitepapers from JEDEC[15, 16] mention that the deployed “in-DRAM mitigations cannot eliminate all forms of Rowhammer attacks”.

Even if all aggressors are tracked accurately, victim-refresh is still not secure as it preserves spatial proximity between aggressor and victims, enabling attacks such as Half-Double. Note that increasing the victim refreshed to two on each side does not solve Half-Double, as rows distance-of-three away can now incur bit flips. Instead, our solution Rubix makes secure Rowhammer mitigations, which are resilient to complex attacks, practical at ultra-low thresholds, as shown in Table 5. Rubix is a memory mapping and is compatible with any tracking and mitigation mechanism.

![Table 5. Comparison of Rowhammer Mitigations](image)

Rubix can also greatly reduce the overheads of existing mitigations, which rely on victim refresh, by eliminating the root cause of overheads – hot-rows, thereby requiring much reduced number of mitigative actions. Moreover, all our evaluated secure mitigations (AQUA, SRS, and BlockHammer) work with commodity DRAM and DDR protocol, while in-DRAM mechanisms like Rega [32] typically require changes to DDR protocol and DRAM architecture (and significant energy overheads to mitigate low thresholds). Thus, such solutions are orthogonal to our work.

7.4 Randomization to Mitigate Rowhammer

Recent row migration proposals [42, 43, 52, 53] mitigate Rowhammer by moving an aggressor row to another row in memory. However, such row-to-row randomization does not change the set of lines that co-reside in the row. Likewise, randomized DRAM address remapping [21] retains the set of lines coresident in the same memory row. Thus, unlike our solution, these schemes do not reduce the hot-rows.

8 Conclusion

Rowhammer gets worse as thresholds drop and attacks develop complex patterns that defeat the commonly used victim-refresh. Mitigations resilient to complex attacks, like AQUA, SRS, and Blockhammer, suffer from drastic slowdown at low thresholds due to many hot-rows. We identify the line-to-row mapping as the root cause of hot-rows, as it places spatially correlated lines in same row. Our proposal, Rubix, breaks this spatial correlation by randomizing the line-to-row mapping, reducing the number of hot rows by more than 100x. Rubix reduces overheads of the prior schemes by 10-100x, making them viable for practical adoption.

Acknowledgments

We thank Salman Qazi (Google) for feedback on an earlier draft of our paper. We also thank our shepherd, Prof. Alaa Alameldeen, and the anonymous reviewers of MICRO-2023 and ASPLOS-2024 for their comments and feedback. This work was supported in part by a gift from Intel.
References


