How to Kill the Second Bird with One ECC: The Pursuit of Row Hammer Resilient DRAM

Michael Jaemin Kim[†], Minbok Wi[†], Jaehyun Park[†], Seoyoung Ko[†], Jaeyoung Choi[†], Hwayong Nam[†], Nam Sung Kim[‡], Jung Ho Ahn[†], Eojin Lee[§] Seoul National University[†], University of Illinois Urbana Champaign[‡], Inha University[§] Presenter: Michael Jaemin Kim (michael604@scale.snu.ac.kr)







Overview

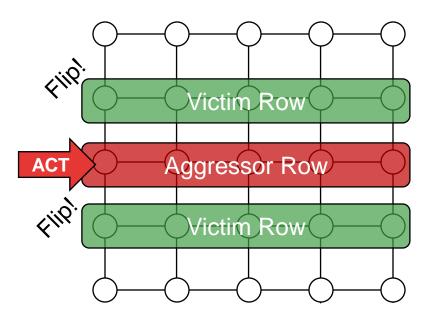
- Problem 1: Scalability problem of existing Row Hammer solutions.
- Problem 2: Limited success of DRAM ECC against Row Hammer.

- Goal 1: Exploit the ECC against Row Hammer.
- Goal 2: Cooperate with the probabilistic Row Hammer solutions.

- Technique 1: Row address scramble for Chipkill against Row Hammer.
- Technique 2: Victim diagnosis using On-die ECC error profile against Row Hammer.

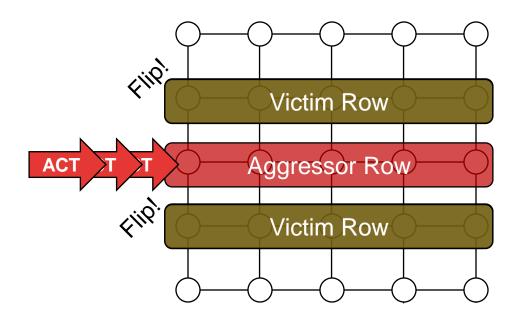
Row Hammer

- Row Hammer?
 - Row Hammer Threshold (T_{RH})

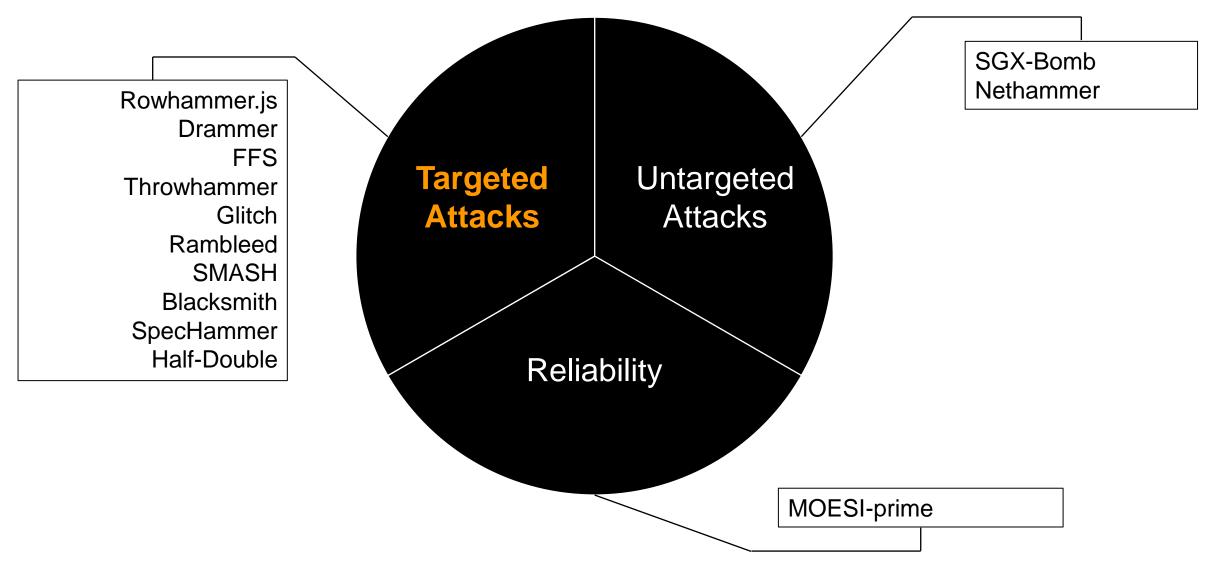


Row Hammer

- Row Hammer?
 - Row Hammer Threshold (T_{RH})

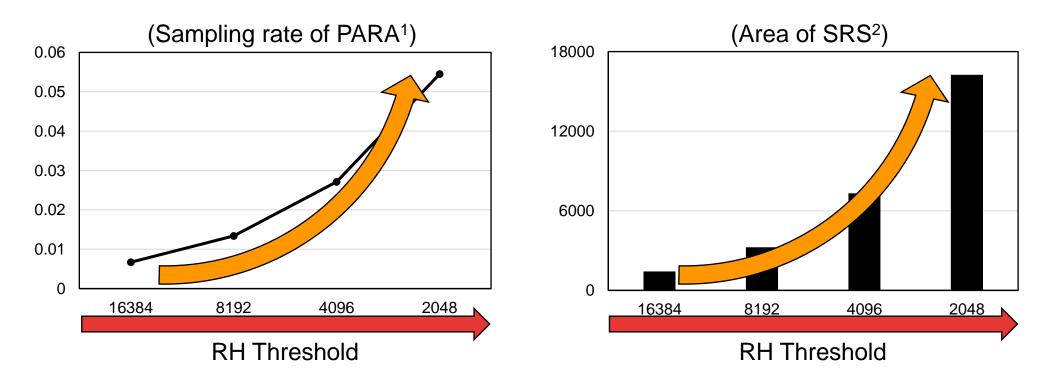


Row Hammer Attacks



Row Hammer Mitigation

- Scalability problem following lowering Row Hammer Threshold
 - Performance
 - Area



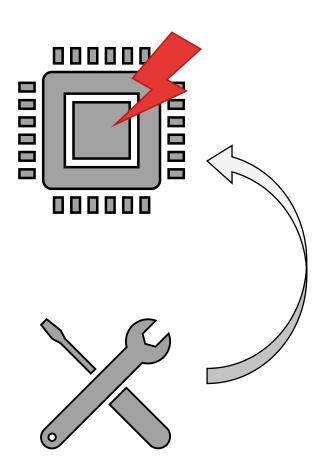
¹ Yoongu Kim, Ross Daly, Jeremie Kim, Chris Fallin, Ji Hye Lee, Donghyuk Lee, Chris Wilkerson, Konrad Lai, and Onur Mutlu. 2014. Flipping Bits in Memory Without Accessing Them: An Experimental Study of DRAM Disturbance Errors.

² Gururaj Saileshwar, Bolin Wang, Moinuddin Qureshi, and Prashant J. Nair. 2022. Randomized Row-Swap: Mitigating Row Hammer by Breaking Spatial Correlation between Aggressor and Victim Rows.

DRAM ECC

- DRAM errors
 - Single cell faults
 - Process scaling faults, cosmic rays, etc.
 - Large granularity row/column hard faults

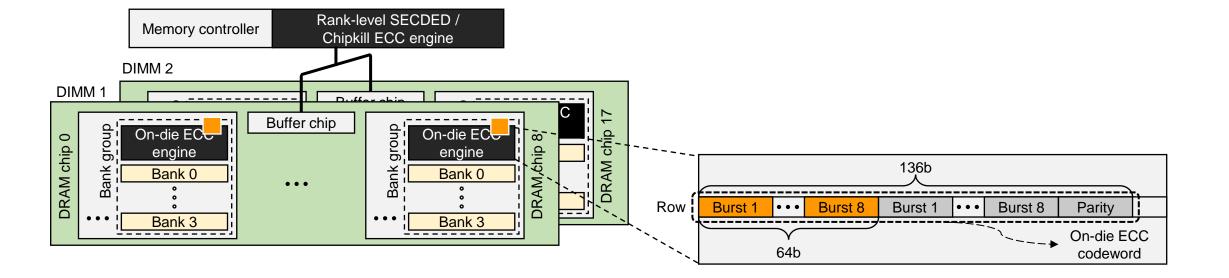
- Error correcting codes (ECC)
 - Redundancy / parity bits
 - Encoding
 - Correct or detect errors
 - Decoding



DRAM ECC

- Example
 - Two DIMMs per memory channel,
 - One rank per DIMM
 - ×8 chips with a burst length of eight

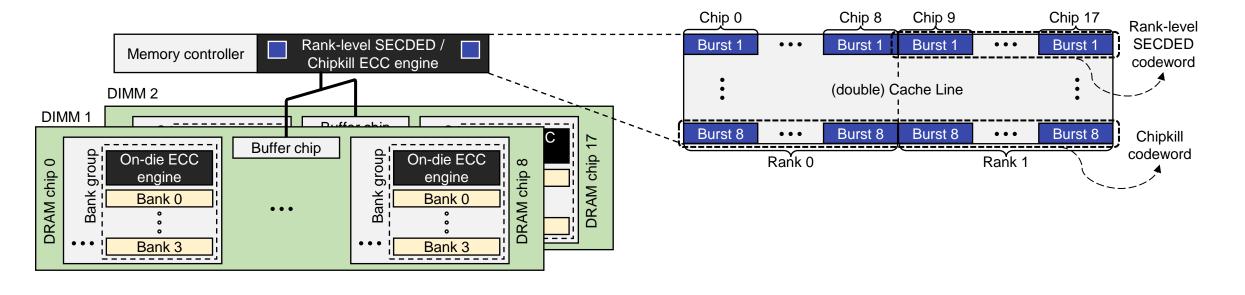
- Example
 - On-die ECC (OECC): Single Error Correction
 - Memory-Controller ECC: Chipkill ECC



DRAM ECC

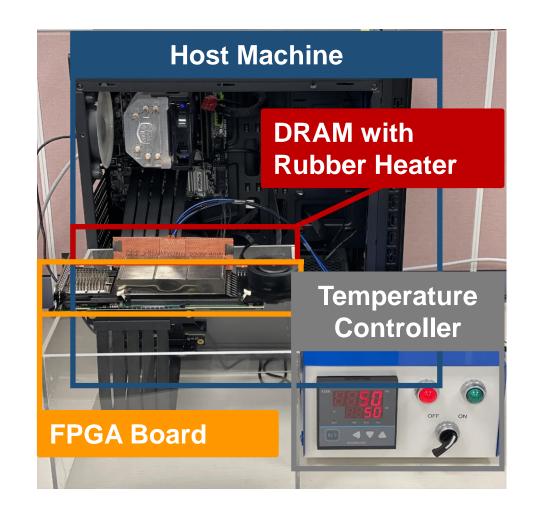
- Example
 - Two DIMMs per memory channel,
 - One rank per DIMM
 - ×8 chips with a burst length of eight

- Example
 - On-die ECC (OECC): Single Error Correction
 - Memory-Controller ECC: Chipkill ECC



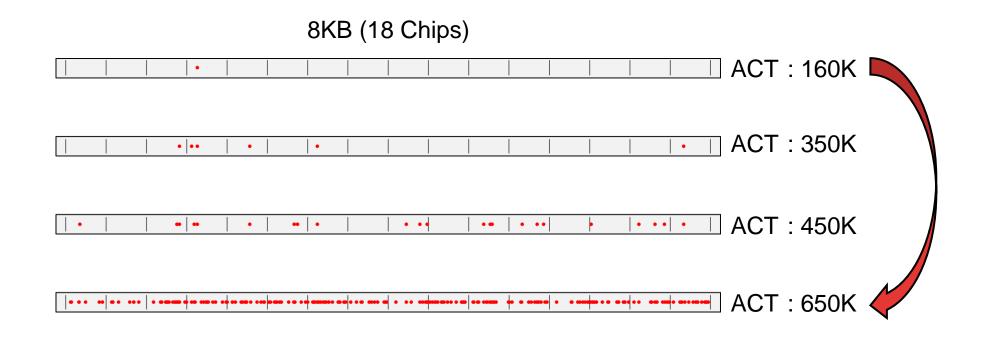
DRAM ECC against Row Hammer Error

- Row Hammer experiment
 - SoftMC-based¹ FPGA environment
 - Xilinx Alveo U280
 - DDR4 RDIMM
 - 2 vendors
 - 14 DIMMS from 2016 to 2021
 - Total of 208 chips
 - No OECC
 - 8 intra-column data patterns
 - No cross-column patterns
 - Total of 16K rows, 1K rows per each bank
 - 50°C



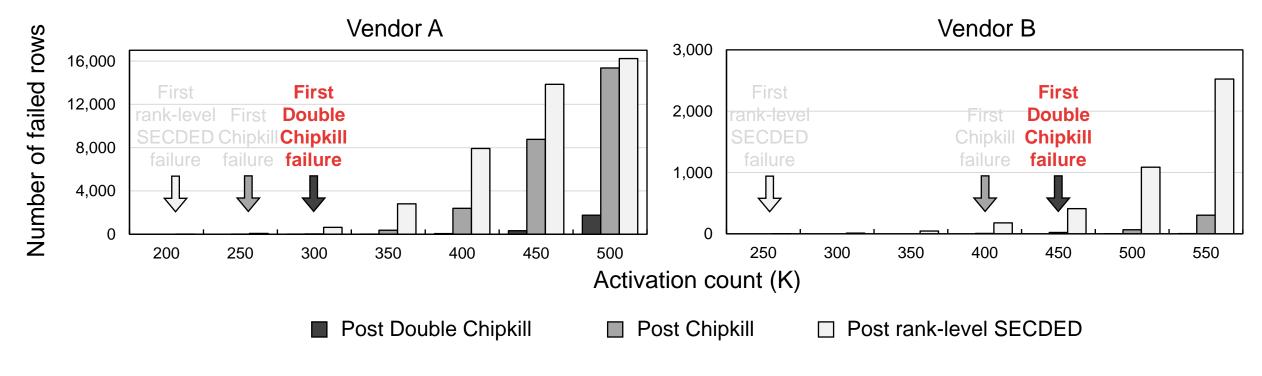
DRAM ECC again Row Hammer?

- Errors occur in a **bursty** way, as ACT count increases.
 - Example: row #10142



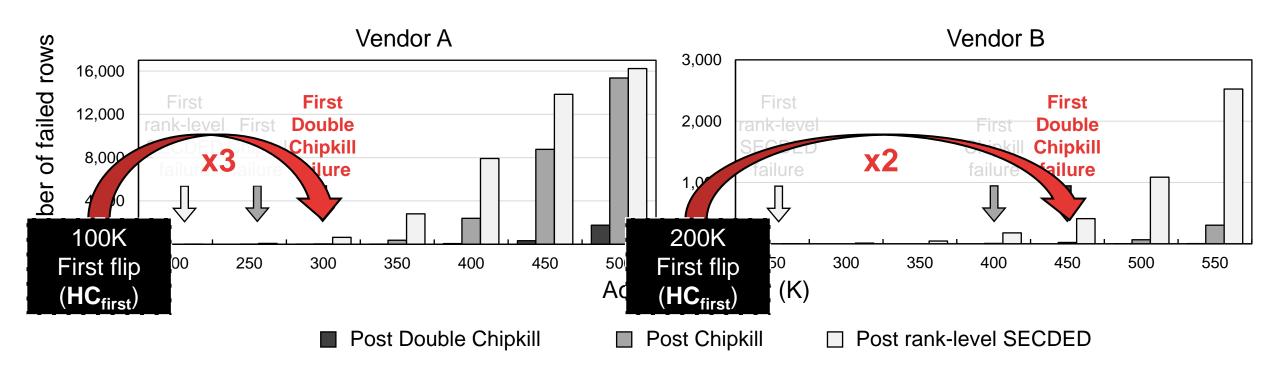
DRAM ECC again Row Hammer?

- ECC cannot tolerate victim row concentrated bursty error.
 - Double Chipkill-level ECC provides only a x3 increase in tolerable hammer count.



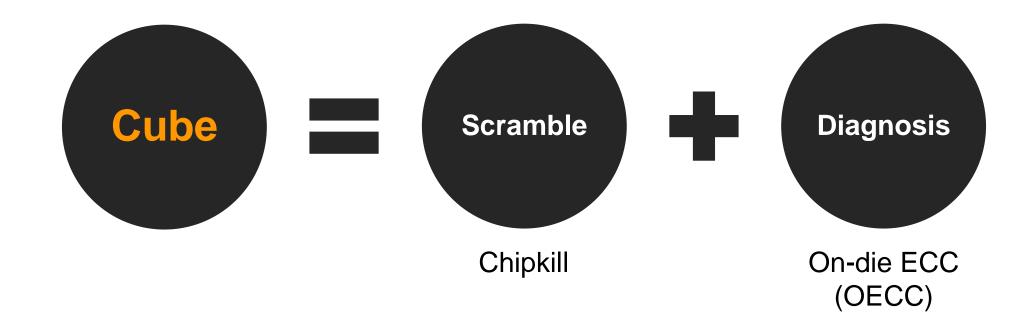
DRAM ECC again Row Hammer?

- ECC cannot tolerate victim row concentrated bursty error.
 - Double Chipkill-level ECC provides only a x3 increase in tolerable hammer count.

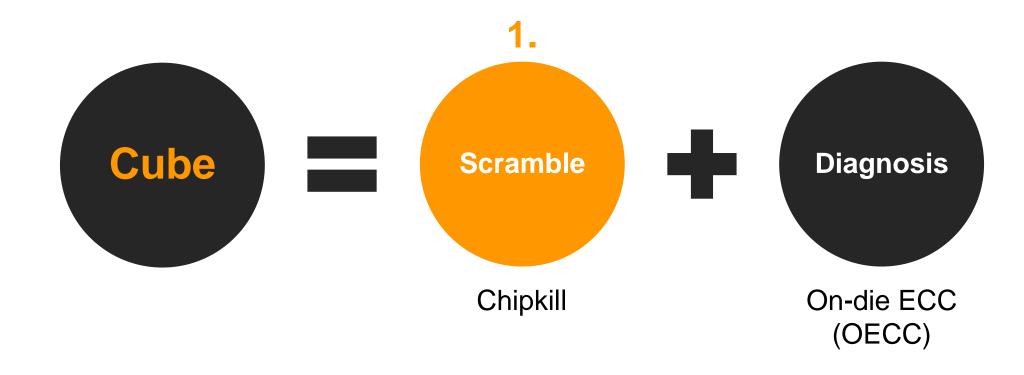


- Goal of Cube:
 - Better utilize the correction & detection capabilities of the existing ECC against Row Hammer.
 - Reduce the overall cost by **cooperating** with the probabilistic Row Hammer schemes.

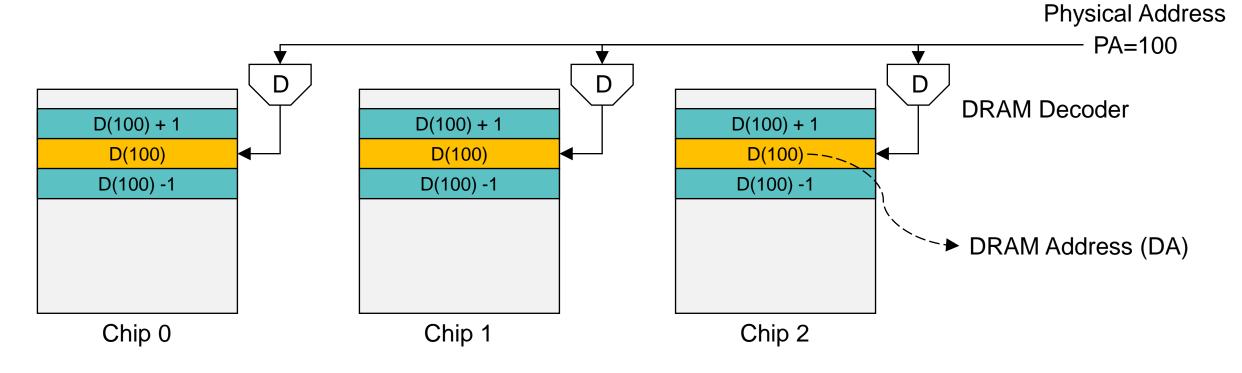
- Goal of Cube:
 - Better utilize the correction & detection capabilities of the existing ECC against Row Hammer.
 - Reduce the overall cost by cooperating with the probabilistic Row Hammer schemes.



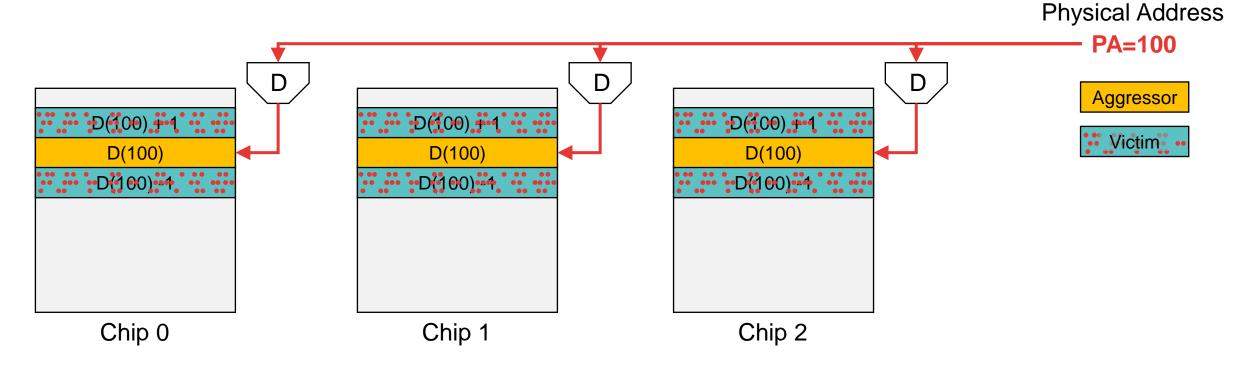
- Goal of Cube:
 - Better utilize the correction & detection capabilities of the existing ECC against Row Hammer.
 - Reduce the overall cost by cooperating with the probabilistic Row Hammer schemes.



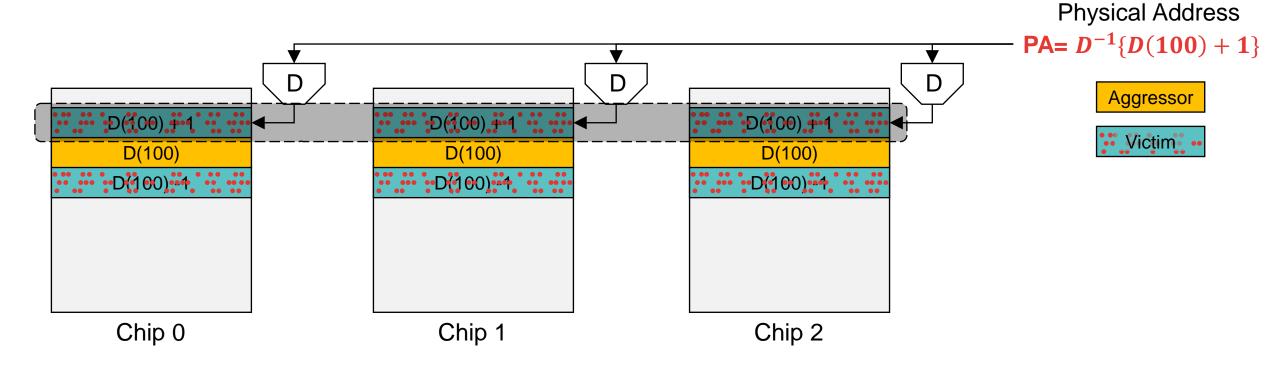
- Conventional addressing mechanism
 - Victims are concentrated to two Chipkill codewords



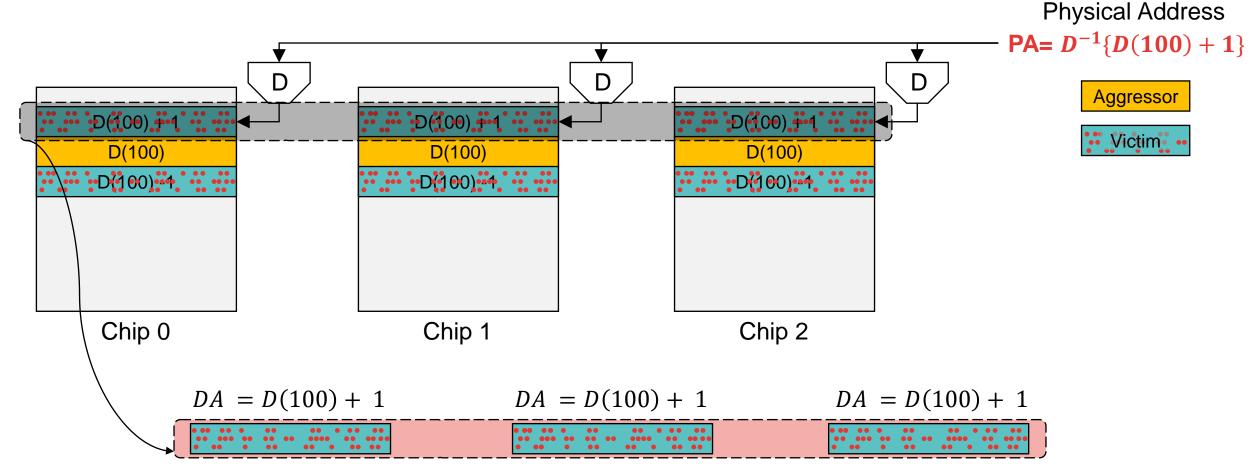
- Conventional addressing mechanism
 - Victims are concentrated to two Chipkill codewords



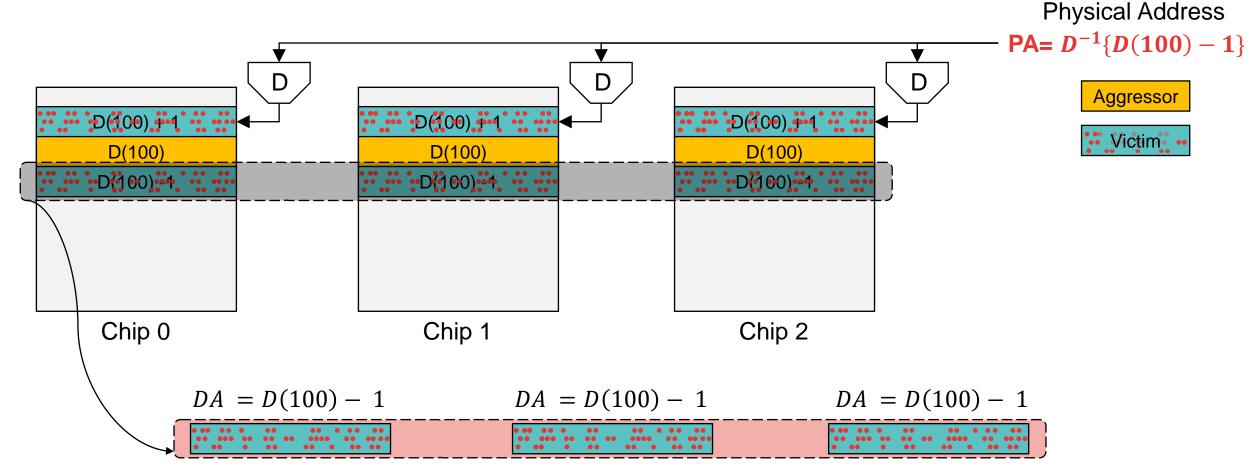
- Conventional addressing mechanism
 - Victims are concentrated to two Chipkill codewords



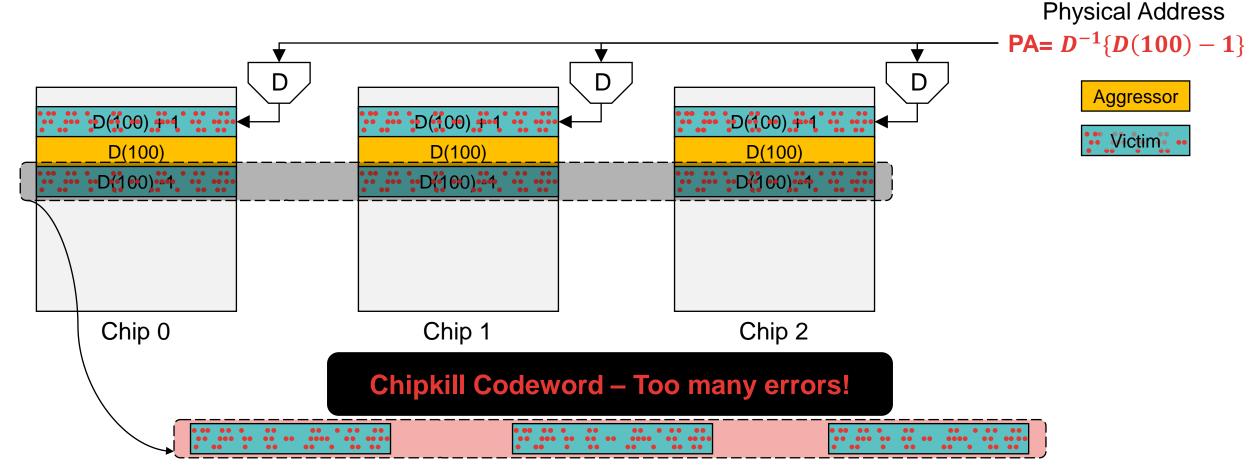
- Conventional addressing mechanism
 - Victims are concentrated to two Chipkill codewords



- Conventional addressing mechanism
 - Victims are concentrated to two Chipkill codewords

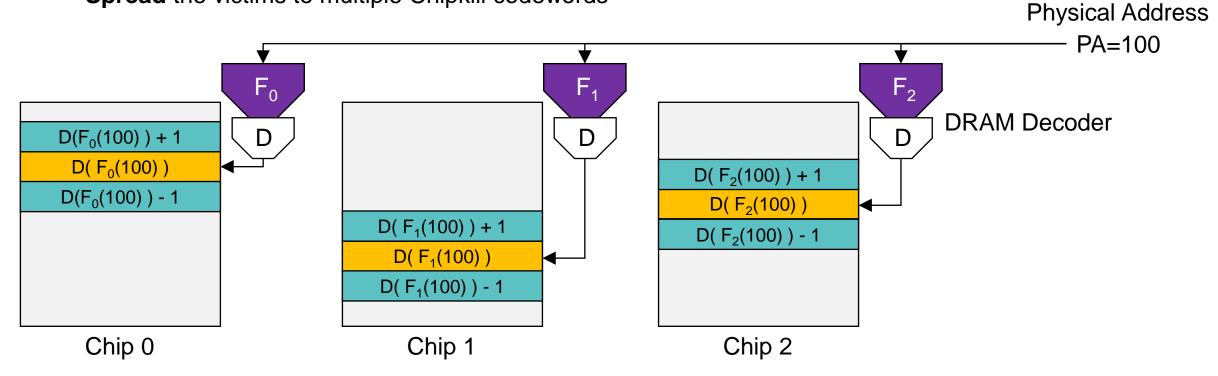


- Conventional addressing mechanism
 - Victims are concentrated to two Chipkill codewords

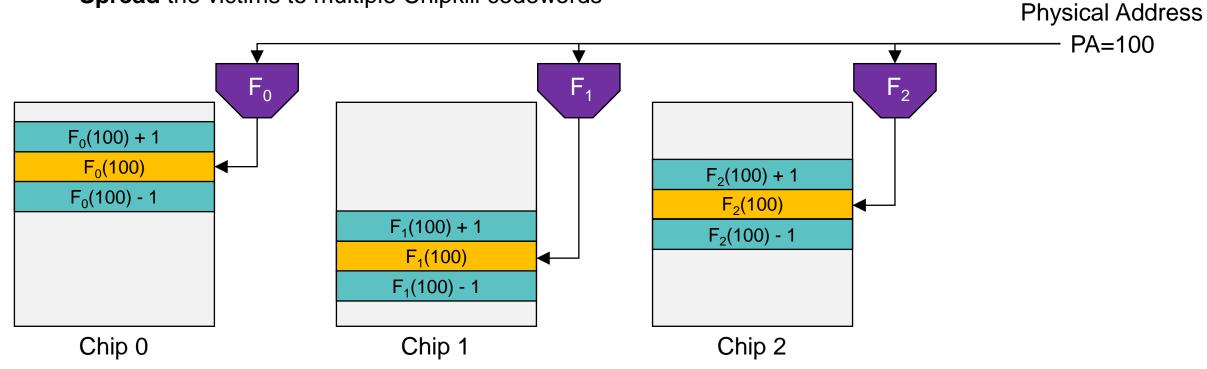


- Make every chip's victim unique
 - Spread the victims to multiple Chipkill codewords

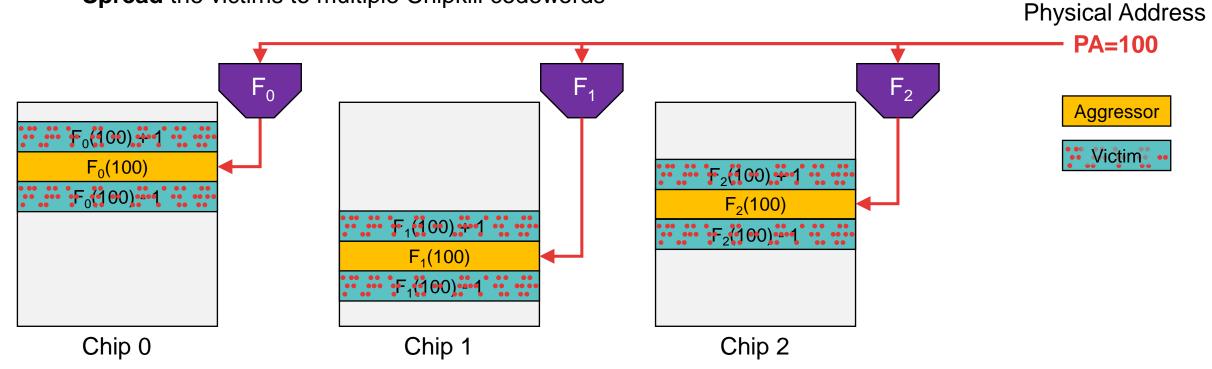
- Make every chip's victim unique
 - Spread the victims to multiple Chipkill codewords



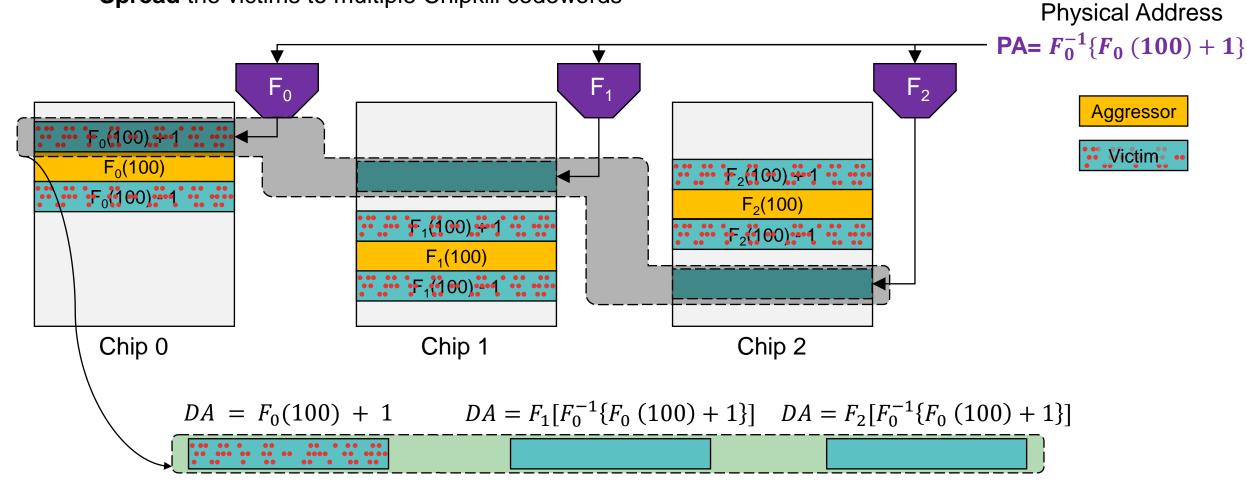
- Make every chip's victim unique
 - Spread the victims to multiple Chipkill codewords



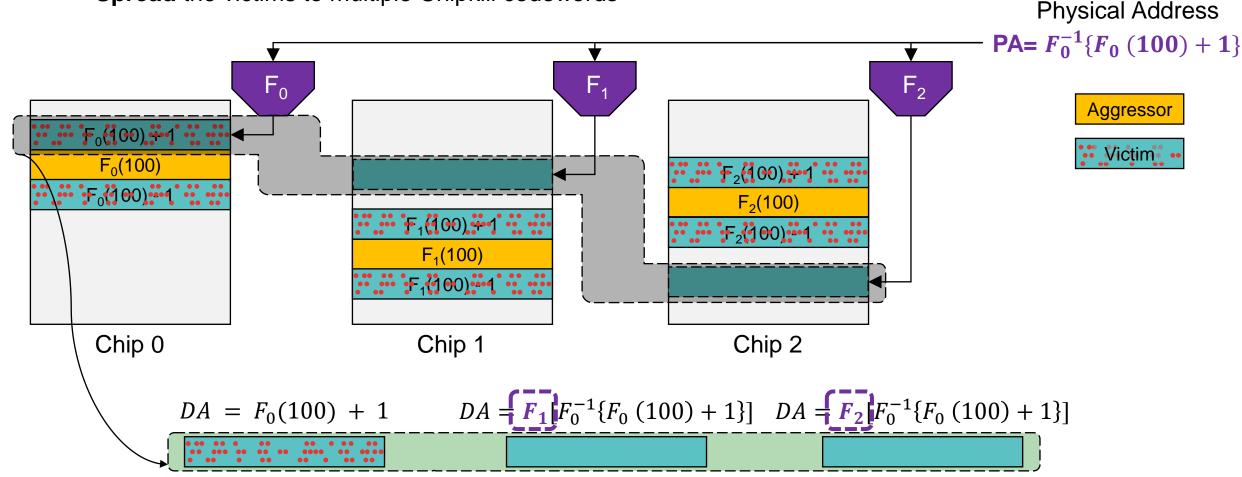
- Make every chip's victim unique
 - Spread the victims to multiple Chipkill codewords



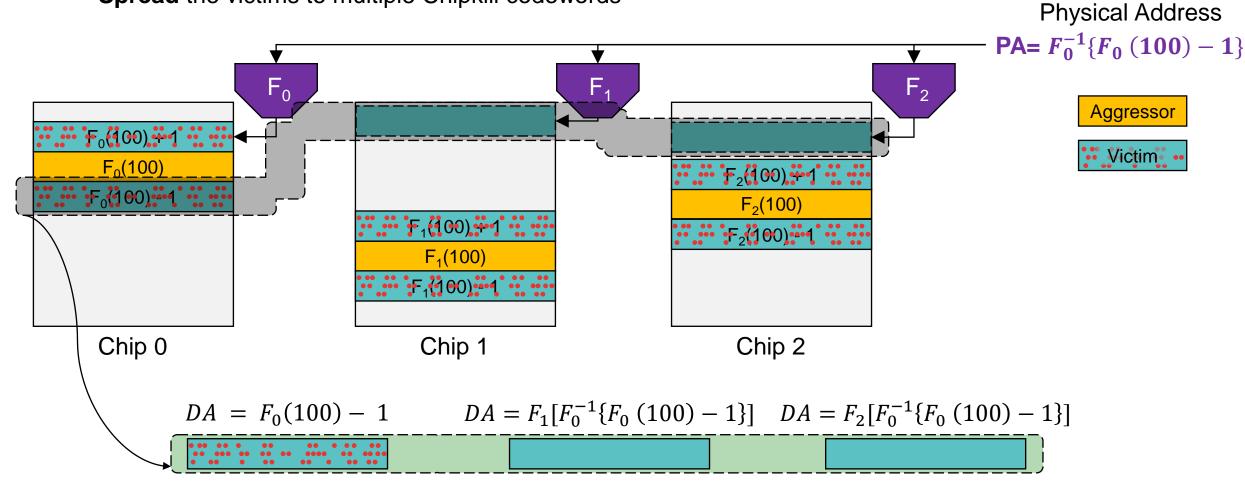
- Make every chip's victim unique
 - Spread the victims to multiple Chipkill codewords



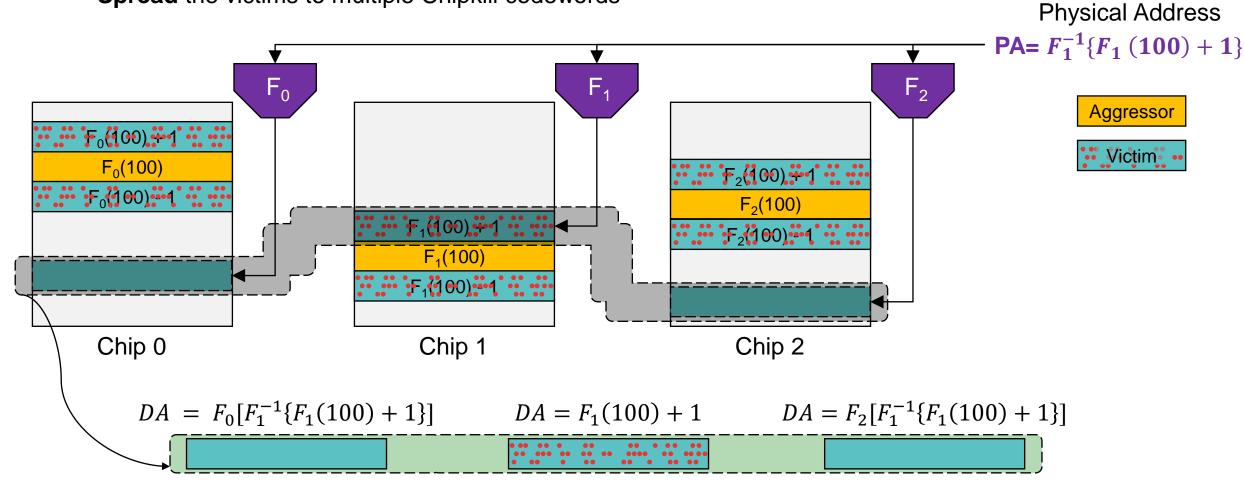
- Make every chip's victim unique
 - **Spread** the victims to multiple Chipkill codewords



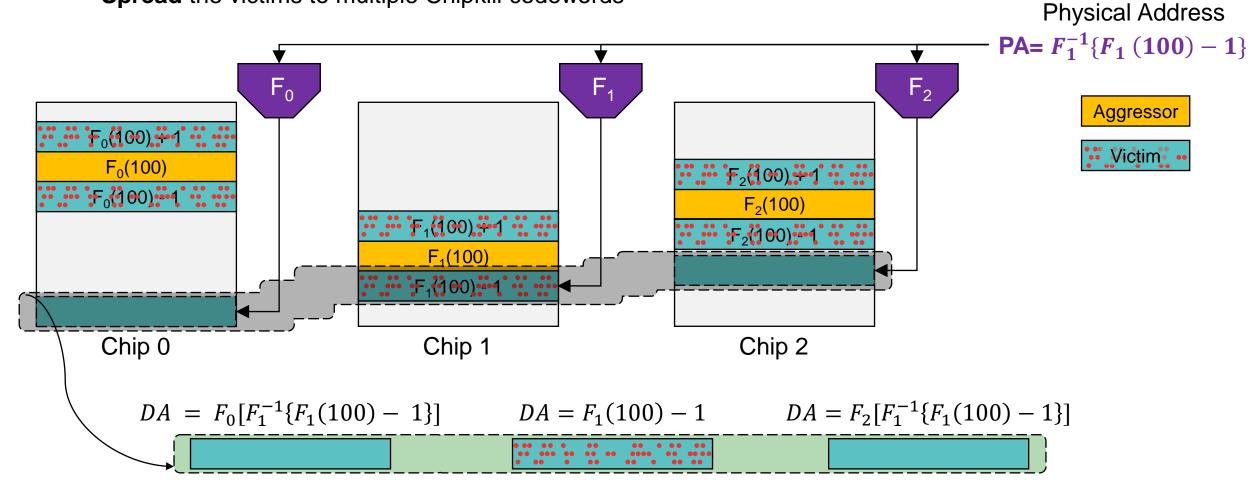
- Make every chip's victim unique
 - Spread the victims to multiple Chipkill codewords



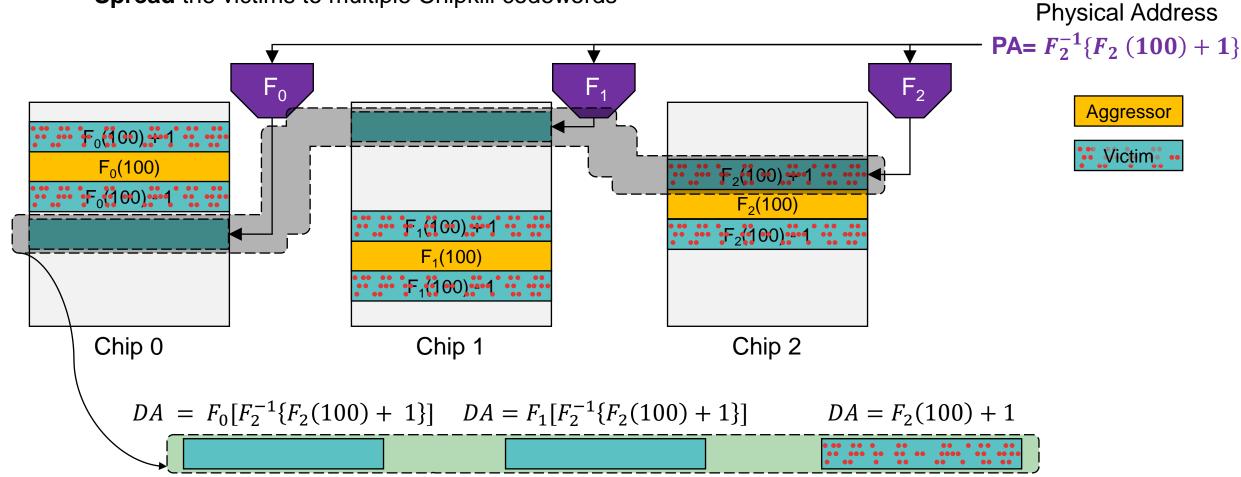
- Make every chip's victim unique
 - Spread the victims to multiple Chipkill codewords



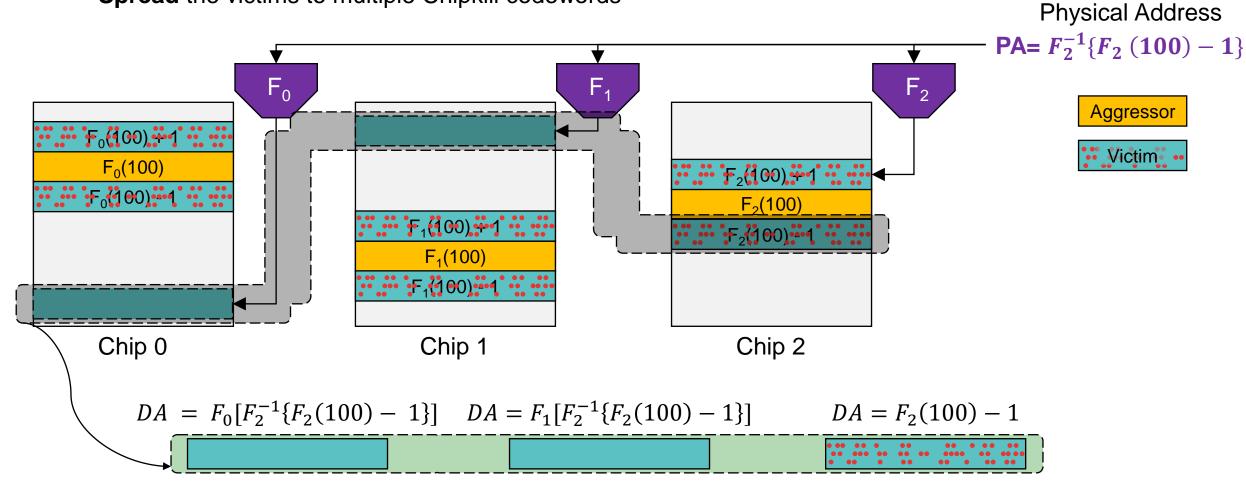
- Make every chip's victim unique
 - Spread the victims to multiple Chipkill codewords



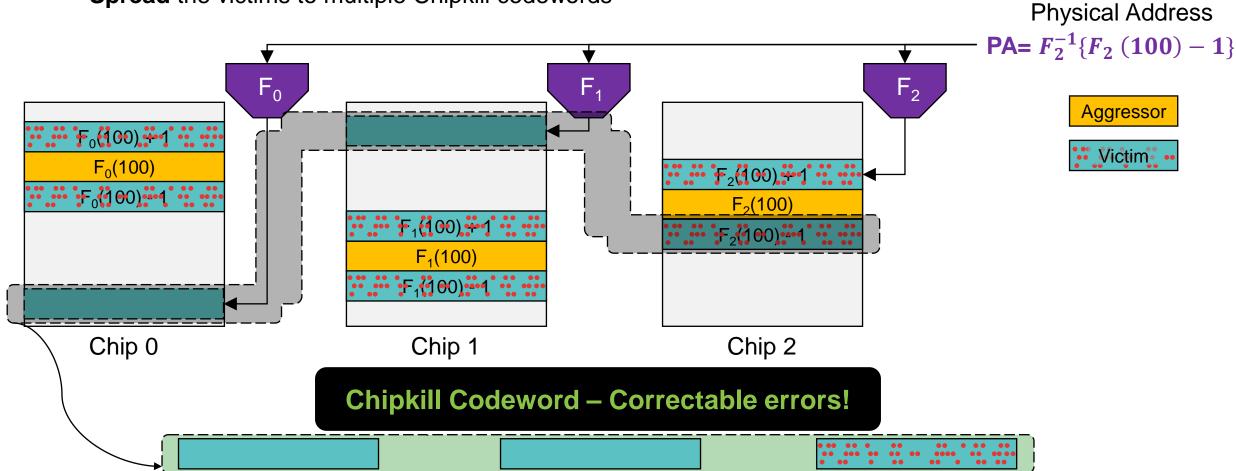
- Make every chip's victim unique
 - Spread the victims to multiple Chipkill codewords



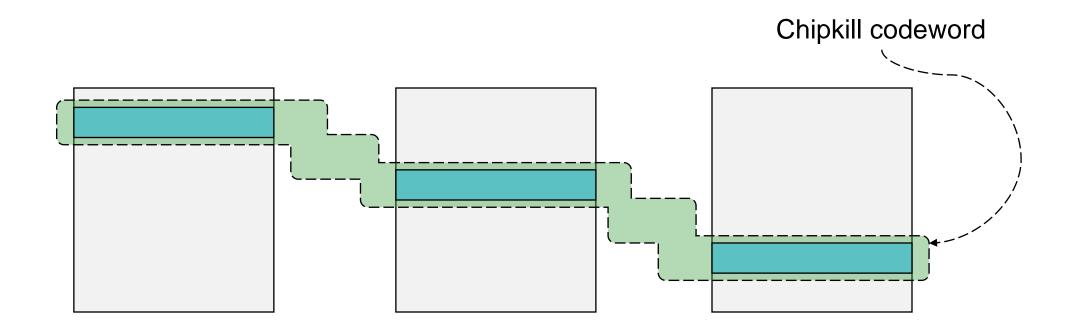
- Make every chip's victim unique
 - Spread the victims to multiple Chipkill codewords



- Make every chip's victim unique
 - Spread the victims to multiple Chipkill codewords

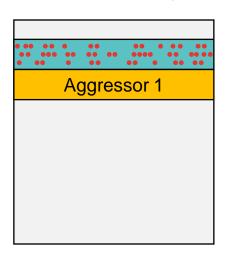


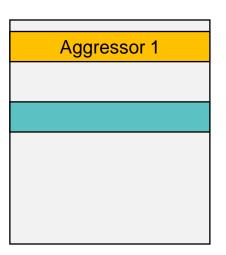
- Force the attacker to succeed in not only one, not only two, but multiple Row Hammer attacks!
 - Theoretically, two successful Row Hammer attack can be enough, but with low probability.

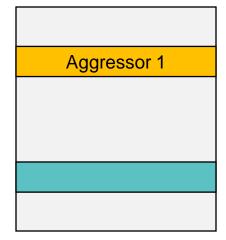


- Force the attacker to succeed in not only one, not only two, but multiple Row Hammer attacks!
 - Theoretically, two successful Row Hammer attack can be enough, but with low probability.

1. (Theoretical) Two successful Row Hammer attack

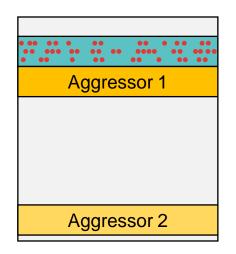




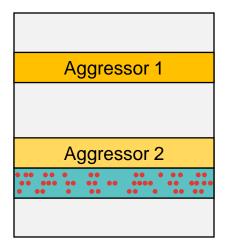


- Force the attacker to succeed in not only one, not only two, but multiple Row Hammer attacks!
 - Theoretically, two successful Row Hammer attack can be enough, but with low probability.

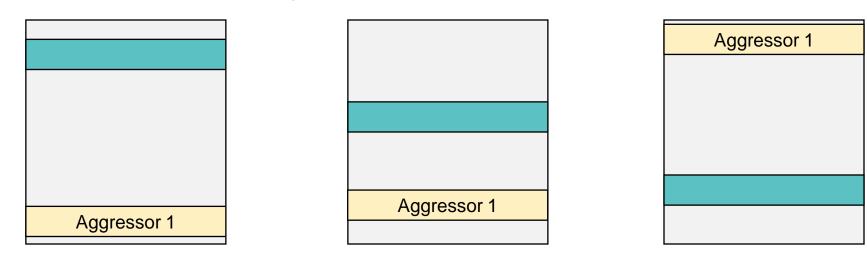
1. (Theoretical) Two successful Row Hammer attack



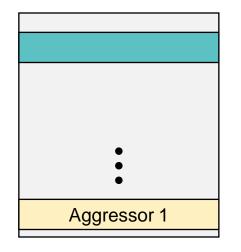


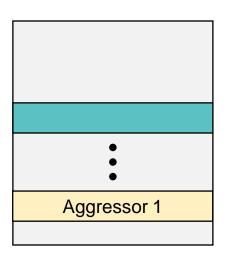


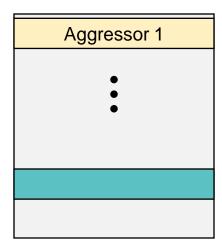
- Force the attacker to succeed in not only one, not only two, but multiple Row Hammer attacks!
 - Theoretically, two successful Row Hammer attack can be enough, but with low probability.
 - Without knowing the mapping function F, multiple Row Hammer attacks are likely needed.
 - "Birthday collision"



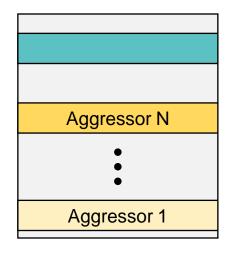
- Force the attacker to succeed in not only one, not only two, but multiple Row Hammer attacks!
 - Theoretically, two successful Row Hammer attack can be enough, but with low probability.
 - Without knowing the mapping function F, multiple Row Hammer attacks are likely needed.
 - "Birthday collision"

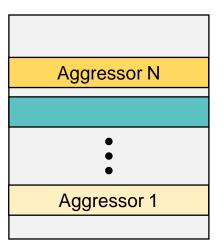


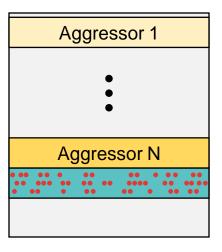




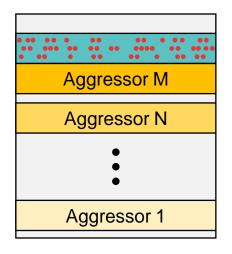
- Force the attacker to succeed in not only one, not only two, but multiple Row Hammer attacks!
 - Theoretically, two successful Row Hammer attack can be enough, but with low probability.
 - Without knowing the mapping function F, multiple Row Hammer attacks are likely needed.
 - "Birthday collision"

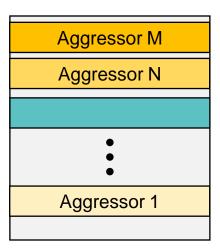


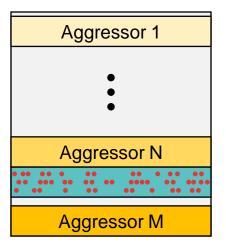




- Force the attacker to succeed in not only one, not only two, but multiple Row Hammer attacks!
 - Theoretically, two successful Row Hammer attack can be enough, but with low probability.
 - Without knowing the mapping function F, multiple Row Hammer attacks are likely needed.
 - "Birthday collision"







- Scramble function (F) requirements
 - R-i) For any given aggressor, victims must not collide.
 - *R-ii)* Must be fast, as it lies on the DRAM access critical path.
 - R-iii) No aliasing.
 - R-iv) Mapping function must be hidden from the attacker.

- Scramble function (F) requirements
 - R-i) For any given aggressor, victims must not collide.
 - All victim PAs must not collide within the same victim set.
 - *R-ii)* Must be fast, as it lies on the DRAM access critical path.
 - No AES or Fisher-Yates shuffle.
 - R-iii) No aliasing.
 - No hashing.
 - R-iv) Mapping function must be hidden from the attacker.

- Scramble function (F) requirements
 - R-i) For any given aggressor, victims must not collide.
 - All victim PAs must not collide within the same victim set.
 - R-ii) Must be fast, as it lies on the DRAM access critical path.
 - No AES or Fisher-Yates shuffle.
 - R-iii) No aliasing.
 - No hashing.
 - R-iv) Mapping function must be hidden from the attacker.

Vanilla Scramble Function for chip i, $F_i(PA) := (PA \times c_i) \mod N_{row}$

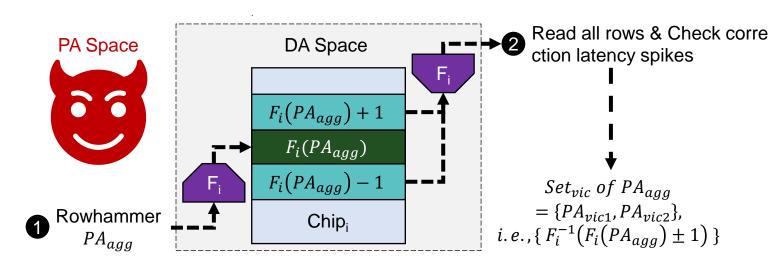
$$R-i) \Rightarrow \{c_i \neq c_j \text{ and } c_i \neq (N_{row} - c_j), 0 \leq i, j < 18\}$$

 $R-iii) \Rightarrow c_i \text{ is odd}$

- Scramble function (F) requirements
 - R-i) For any given aggressor, victims must not collide.
 - R-ii) Must be fast, as it lies on the DRAM access critical path.
 - R-iii) No aliasing.
 - R-iv) Mapping function must be hidden from the attacker.
 - If not, two Row Hammer attacks are enough.
 - Possibility of side-channel from Chipkill correction latency¹.
 - A threat for static scheme (e.g., PARA)
 - Not an issue for dynamic scheme (e.g., SHADOW, SRS)

¹ Lucian Cojocar, Kaveh Razavi, Cristiano Giuffrida, and Herbert Bos. 2019. Exploiting Correcting Codes: On the Effectiveness of ECC Memory Against Rowhammer Attacks.

- Scramble function (F) requirements
 - R-i) For any given aggressor, victims must not collide.
 - *R-ii*) Must be fast, as it lies on the DRAM access critical path.
 - R-iii) No aliasing.
 - R-iv) Mapping function must be hidden from the attacker.
 - If not, two Row Hammer attacks are enough.
 - Possibility of side-channel from Chipkill correction latency¹.



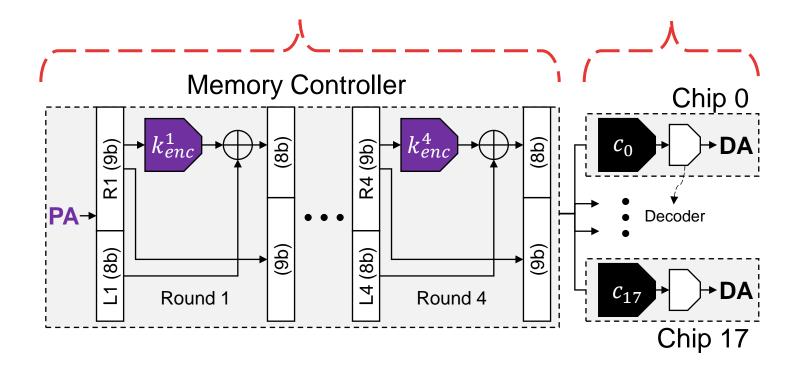
¹ Lucian Cojocar, Kaveh Razavi, Cristiano Giuffrida, and Herbert Bos. 2019. Exploiting Correcting Codes: On the Effectiveness of ECC Memory Against Rowhammer Attacks.

- Scramble function (F) requirements
 - R-i) For any given aggressor, victims must not collide.
 - R-ii) Must be fast, as it lies on the DRAM access critical path.
 - R-iii) No aliasing.
 - R-iv) Mapping function must be hidden from the attacker.
 - If not, two Row Hammer attacks are enough.
 - Possibility of side-channel from Chipkill correction latency¹.

(Optional) Encrypted Scramble Function
$$F_i(PA) := (Enc(PA) \times c_i) \mod N_{row}$$

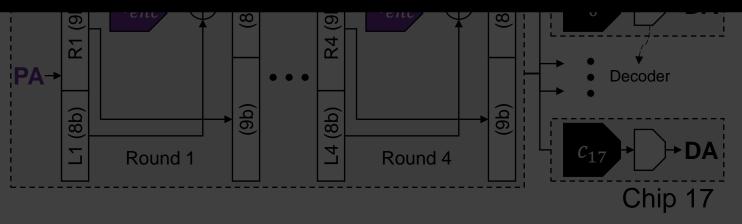
¹ Lucian Cojocar, Kaveh Razavi, Cristiano Giuffrida, and Herbert Bos. 2019. Exploiting Correcting Codes: On the Effectiveness of ECC Memory Against Rowhammer Attacks.

- Scramble function (F) implementation
 - Global encryption scheme (LLBC¹)
 - Chip-wise fixed multiplication unit



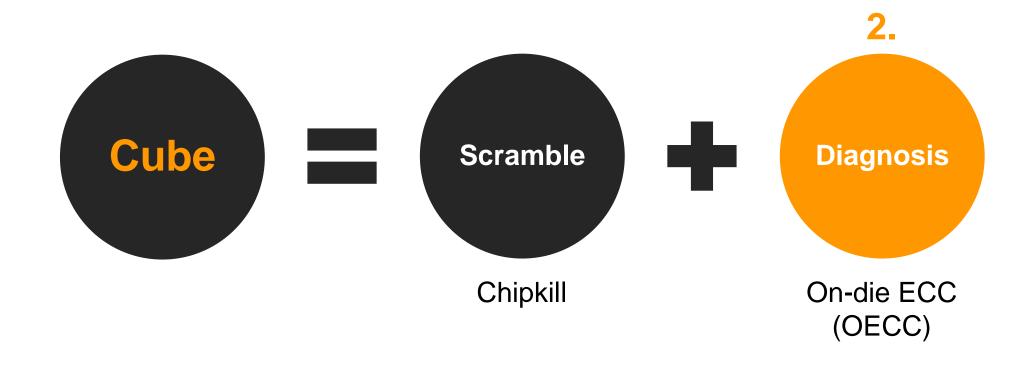
- Scramble function (F) implementation
 - Chip-wise fixed multiplication unit
 - Global encryption scheme (LLBC¹)

Please refer to the full paper for more detailed discussion on security and implementation.



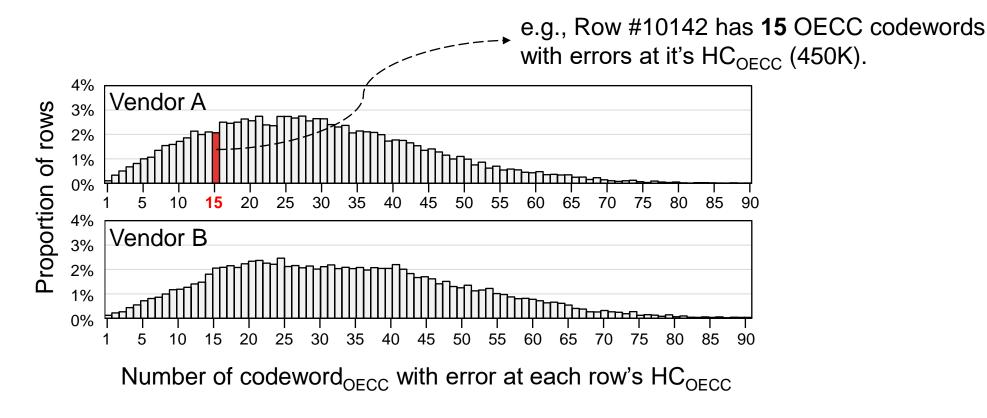
Cube

- Goal of Cube:
 - Better utilize the correction & detection capabilities of the existing ECC against Row Hammer.
 - Reduce the overall cost by cooperating with the probabilistic Row Hammer schemes.

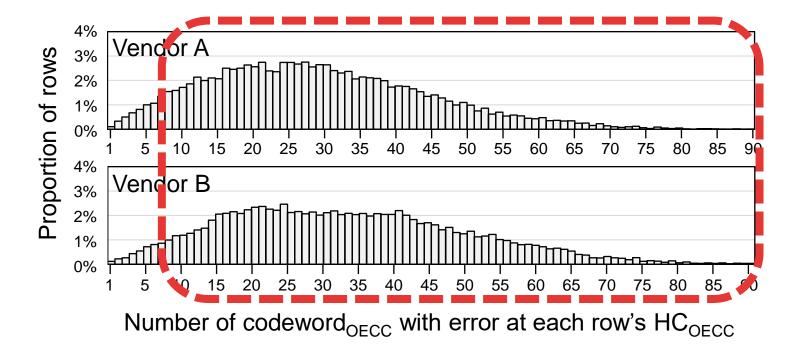


Row with a lot of OECC errors can indicate that it has been Row Hammer victimized.

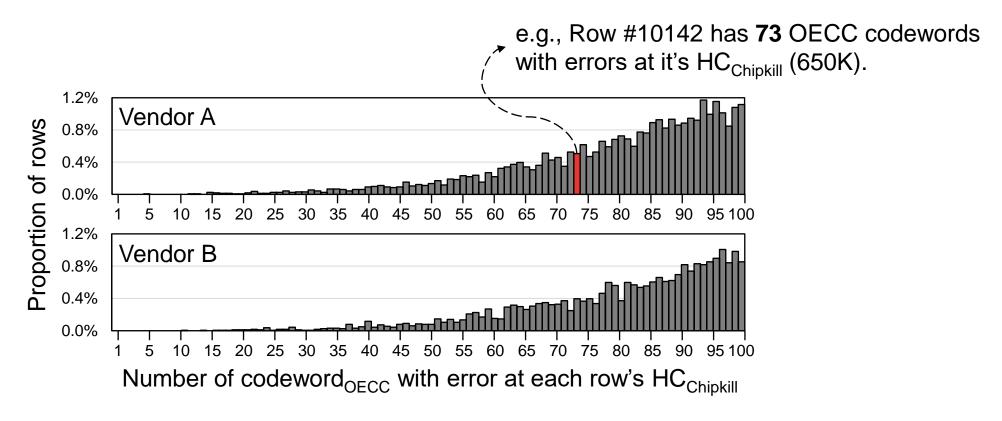
- Row with a lot of OECC errors can indicate that it has been Row Hammer victimized.
 - At the first OECC uncorrectable errors (HC_{OECC}), many OECC correctable errors are likely to have occurred before.



- Row with a lot of OECC errors can indicate that it has been Row Hammer victimized.
 - At the first OECC uncorrectable errors (HC_{OECC}), many OECC correctable errors are likely to have occurred before.



- Row with a lot of OECC errors can indicate that it has been Row Hammer victimized.
 - At the first Chipkill uncorrectable errors (HC_{Chipkill}), the phenomenon is more pronounced.

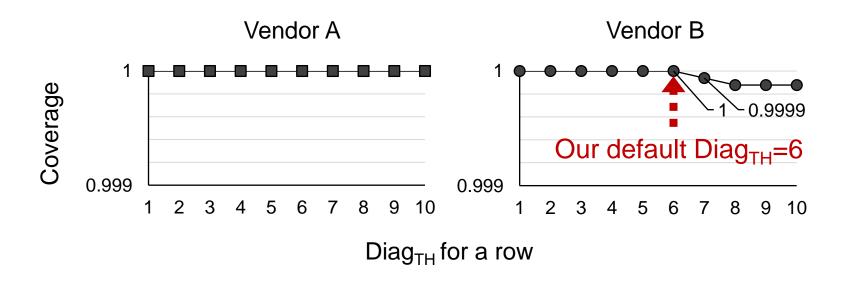


- Cube diagnoses a row as Row Hammer victim, when it has more than Diag_{TH} number of OECC codeword with error.
 - When diagnosed, executes sequence of Chipkill correction.

- Cube diagnoses a row as Row Hammer victim, when it has more than Diag_{TH} number of OECC codeword with error.
 - When diagnosed, executes sequence of Chipkill correction.

Diag_{TH}

- Too low Diag_{TH} (e.g., 1) would incur too much false positive from non-Row Hammer errors.
- Too high Diag_{TH} (e.g., 100) would incur too much **false negative** and lose actual victims.

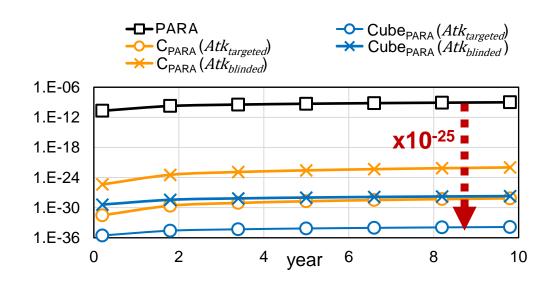


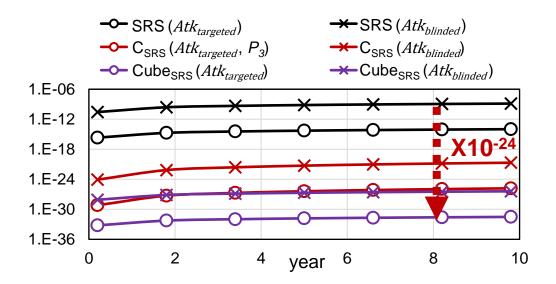
- OECC error profile can be collected from patrol scrubbing (e.g., ECS mode mode adopted for DDR5).
- We adopt a more optimized version of OECC scrubbing with a few additional registers.
 - Bank-group level parallelism
 - Scrub address register (17b)
 - 70B buffering register
 - Scrub all rows every 10 minutes
 - 1.37% reduced tREFI

Evaluation and Result

Cooperation with the Probabilistic Row Hammer Solutions

- Can be co-run with most probabilistic Row Hammer protection solutions.
 - Lowers the probability of Row Hammer attack success.
 - ii. Reduces the overhead of probabilistic Row Hammer solutions, when the target probability is fixed





Evaluation Methodology

- Pin-based cycle accurate simulator (McSimA+)
- SPEC CPU2017
 - Sampled using Simpoint
- Evaluate prior probabilistic schemes
 - PARA1
 - SHADOW²
 - SRS³
 - Did not considered the LLC pinning case, which is rather a deterministic solution.

Core Configurations (16 cores)					
Core	3.6 GHz 4-way O3 cores				
LLC	16 MB (shared)				
Memory System Configurations					
Module	DDR5-4800, 32Gb chips				
Channel	2 channels				
Configuration	1 rank; 32 banks per rank				
Scheduling	BLISS [102]				
Page-Policy	Minimalist-open [44]				

Parameters for architectural simulation

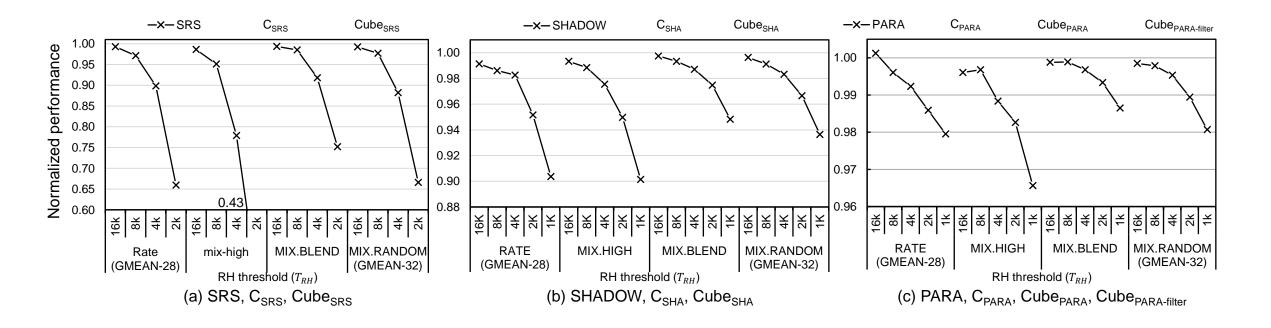
- Targets RH protection guarantee of 10⁻¹⁰ per rank per year.
- Conservatively only consider Atk_{Blinded}.
 - Cube is stronger against Atk_{Targeted} and most attacks belong to the latter.

¹ Yoongu Kim, Ross Daly, Jeremie Kim, Chris Fallin, Ji Hye Lee, Donghyuk Lee, Chris Wilkerson, Konrad Lai, and Onur Mutlu. 2014. Flipping Bits in Memory Without Accessing Them: An Experimental Study of DRAM Disturbance Errors.

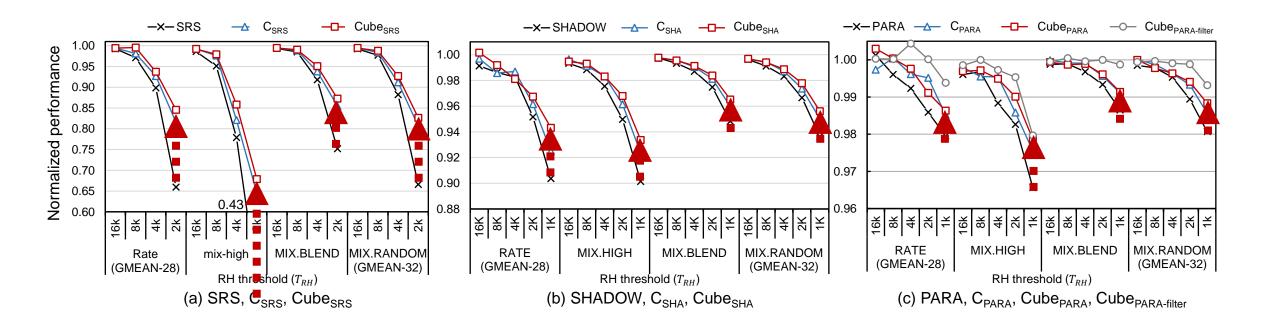
² Minbok Wi, Jaehyun Park, Seoyoung Ko, Michael Jaemin Kim, Nam Sung Kim, Eojin Lee, and Jung Ho Ahn. 2023. SHADOW: preventing Row Hammer in DRAM with Intra-Subarray Row Shuffling.

³ Gururaj Saileshwar, Bolin Wang, Moinuddin Qureshi, and Prashant J. Nair. 2022. Randomized Row-Swap: Mitigating Row Hammer by Breaking Spatial Correlation between Aggressor and Victim Rows.

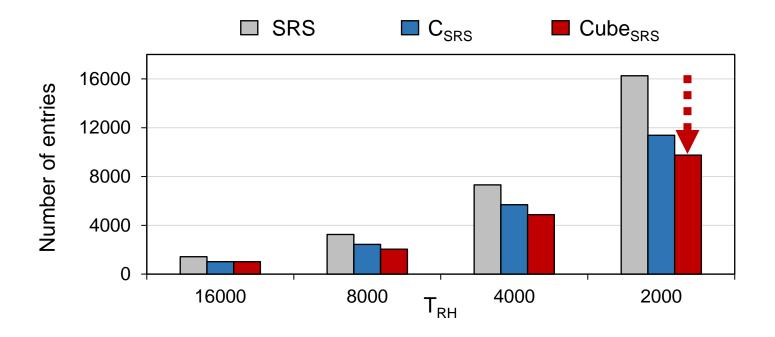
- Improves the performance degradation of all evaluated probabilistic Row Hammer protection schemes.
 - Up to 24.3% reduction.



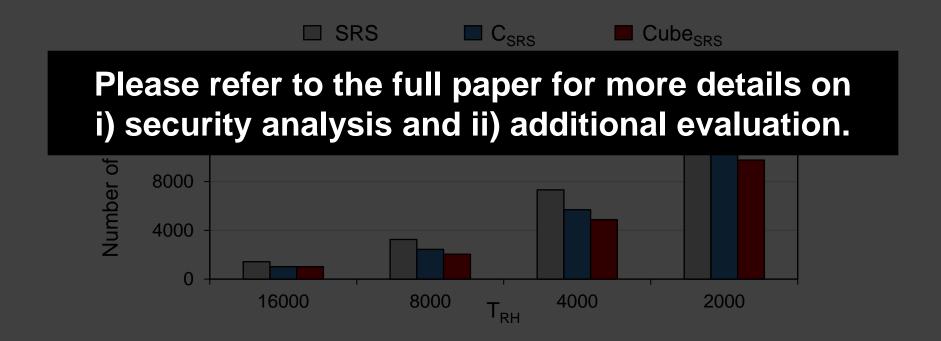
- Improves the performance degradation of all evaluated probabilistic Row Hammer protection schemes.
 - Up to 24.3% reduction.



- Improves the area overhead of probabilistic scheme that requires counter table structure (SRS).
 - Up to 39.9% reduction.



- Improves the area overhead of probabilistic scheme that requires counter table structure (SRS).
 - Up to 39.9% reduction.



Summary

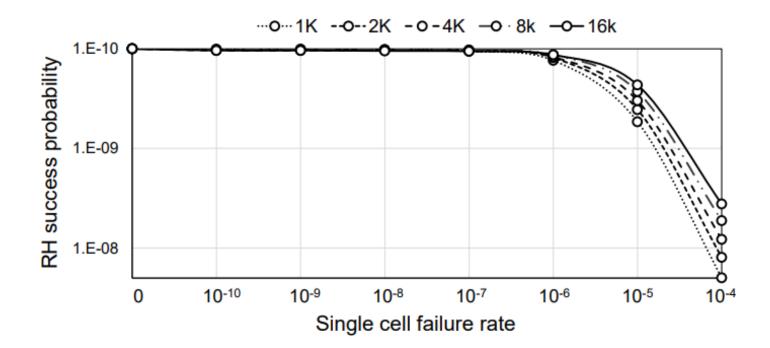
Cube

- 1. Row address **scramble**
 - Spreads the Row Hammer victims across multiple Chipkill codewords
- 2. Victim diagnosis
 - Identify the Row Hammer victims and execute Chipkill correction
- 3. Cooperation with most existing probabilistic Row Hammer solutions
 - i. Improves the **RH protection guarantee** of up to **10**⁻²⁵
 - ii. Improves the **performance** degradation by up to **24.3%** or **area** overhead by up to **39.9%**

Thank you!

Backup: Non-Row Hammer Errors

- 1) Large granularity faults
 - Assumed to be handled by the higher-level RAS features, thus unlikely to coexist with the Row Hammer victim simultaneously.
- 2) Single cell faults



Backup: Configurations

• Fixed target RH protection guarantee of 10⁻¹⁰ per rank per year

Table 5: $Cube_{SRS}$ and SRS k configurations, $Cube_{SHA}$ and SHADOW RAAIMT configurations

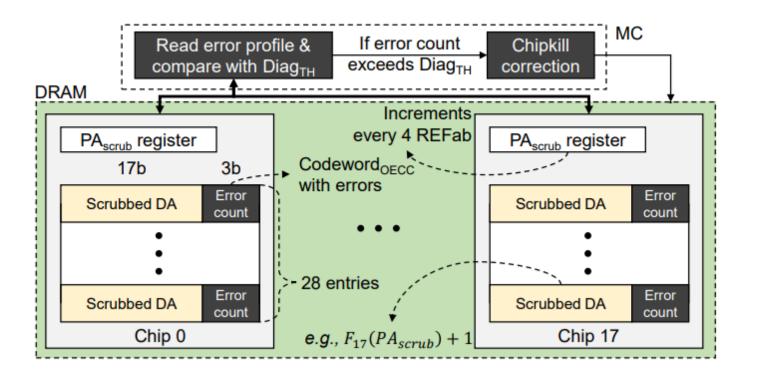
T_{RH}	SRS	C _{SRS}	Cube _{SRS}	SHADOW	C _{SHA}	Cube _{SHA}
1K	-	-	-	19	25	29
2K	10	7	6	39	51	60
4K	9	7	6	79	106	123
8K	8	6	5	161	218	255
16K	7	5	5	332	451	530

Table 6: Cube_{PARA} and PARA p configurations

T_{RH}	PARA	C _{PARA}	Cube _{PARA}	Cube _{PARA-filter}
2K	0.0545	0.0388	0.0334	0.0554 & 400
4K	0.0271	0.0192	0.0164	0.0206 & 400
8K	0.0134	0.0094	0.0081	0.0092 & 500
16K	0.0067	0.0046	0.0039	0.0041 & 300

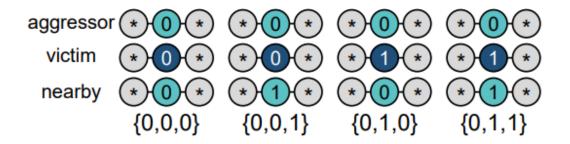
Backup: OECC Scrubbing Methodology

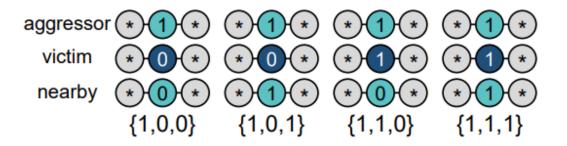
- Must be scrubbed at the "victim set" granularity.
 - For a given specific aggressor (PA), scan it's victims (DA).



Backup: Experimented Data Pattern

- Only considered intra-column data patterns.
 - Did not consider the cross-column data patterns, referring to prior studies.
 - Lucian Cojocar, Kaveh Razavi, Cristiano Giuffrida, and Herbert Bos. 2019. Exploiting Correcting Codes: On the Effectiveness of ECC Memory Against Rowhammer Attacks.
 - Sangwoo Ji, Youngjoo Ko, Saeyoung Oh, and Jong Kim. 2019. Pinpoint Rowhammer: Suppressing Unwanted Bit Flips on Rowhammer Attacks.

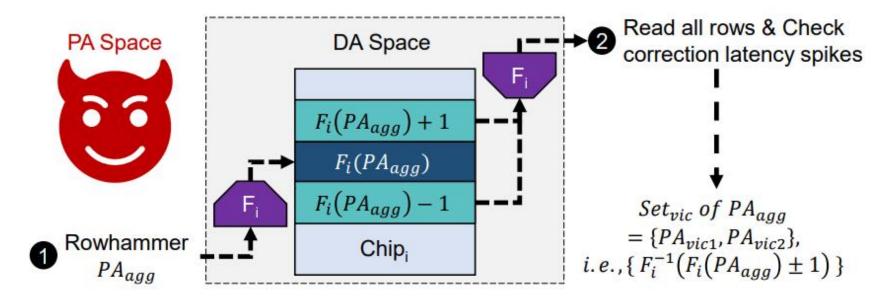




Backup: Implications of Row Scramble

- Targeted RH attacks require templating phase.
- Tailored data pattern attack requires profiling phase.
- Row scramble can actually prevent the attacker from executing such stages, by greatly limiting the possible knowledge gained by the attacker.
 - Only can get side-channel information about the aggressor's PA, and the possible victims' PA, even
 if you succeed in RH attack once, and be able to scan the whole DRAM space.
 - You cannot observe the actual bitflip locations, because they are going to be fixed.

Backup: Row Hammer Side-channel



Vanilla Function

Two victim PAs =
$$c_i^{-1}$$
(PA_{agg} × $c_i \pm 1$)

$$\Rightarrow c_i^{-1}$$
(PA_{agg} × $c_i + 1$) - c_i^{-1} (PA_{agg} × $c_i - 1$) = $2c_i^{-1}$

Encrypted Function

Enc. Scramble Function $F_i(PA) := (Enc(PA) \times c_i) \mod N_{row}$, Two victim PAs : $Dec\{c_i^{-1}(Enc(PA_{agg}) \times c_i \pm 1)\}$

Backup: LLBC Shortcut Attacks

```
For Y = (Y_0 || Y_1 || \cdots Y_n),

Y_i = F_i(X, K) = \mathcal{A}_i(X) \oplus \mathcal{B}_i(K)

= a_1 x_1 \oplus a_2 x_2 \oplus \cdots \oplus a_n x_n \oplus b_1 k_1 \oplus b_2 k_2 \oplus \cdots \oplus b_{mn} k_{mn}

if \mathcal{A}_i(X_1) \oplus \mathcal{B}_i(K) = Y_i = \mathcal{A}_i(X_2) \oplus \mathcal{B}_i(K), \forall i \mid 1 \leq i \leq \lfloor \log_2(L) \rfloor

\Rightarrow \mathcal{A}_i(X_1) = \mathcal{A}_i(X_2), \forall i \mid 1 \leq i \leq \lfloor \log_2(L) \rfloor, independent of K
```

Backup: LLBC Shortcut Attacks

- Redefine the shortcut attack as follows:
 - 1) Using the FPGA environment that bypasses in-DRAM TRR, find the set of of aggressor PAs (e.g., PA_{agg1} and PA_{agg2}) that hammers the same victim PA (e.g., PA_{vic}) on different chips (e.g., chip 1 and 2).

$$PA_{vic1} = Dec[c_{1}^{-1} \cdot \{DA_{vic1}^{[1]}\}, K] \text{ for } PA_{agg1} \text{ on chip } 1$$

$$= Dec[c_{2}^{-1} \cdot \{DA_{vic1}^{[2]}\}, K] \text{ for } PA_{agg2} \text{ on chip } 2$$
where,
$$DA_{vic1}^{[1]} = [c_{1} \cdot \{Enc(PA_{agg1}, K)\} + 1] \mod(2^{17})$$

$$= [c_{1} \cdot \{F_{1}(PA_{agg1}, K)\| \cdot \cdot \cdot \|F_{17}(PA_{agg1}, K)\} + 1] \mod(2^{17})$$

$$= [c_{1} \cdot \{A_{1}(PA_{agg1}) \oplus B_{1}(K)\| \cdot \cdot \cdot \} + 1] \mod(2^{17})$$

$$DA_{vic1}^{[2]} = [c_{2} \cdot \{Enc(PA_{agg2}, K)\} + 1] \mod(2^{17})$$

$$= [c_{2} \cdot \{F_{1}(PA_{agg2}, K)\| \cdot \cdot \cdot \|F_{17}(PA_{agg2}, K)\} + 1] \mod(2^{17})$$

$$= [c_{2} \cdot \{A_{1}(PA_{agg2}) \oplus B_{1}(K)\| \cdot \cdot \cdot \} + 1] \mod(2^{17})$$

- 2) The attacker can try to reuse the discovered set of aggressor PAs on a server environment with a new LLBC key (e.g., K').
- Multiplication and Addition (from victim)