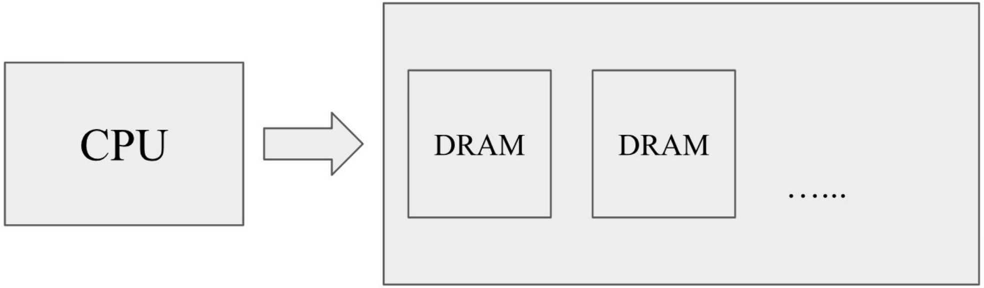
Designing a Secure Hybrid DRAM+NVM Memory Module

Wang Xu and Israel Koren
University of Massachusetts Amherst

Introduction



DRAM Memory Module

Traditionally DRAM as Main Memory

DRAM Limitation:

Easy loss of the stored charges when it comes to 10nm and beyond

NVM (such as PCM) as Main Memory

NVM could be scaled down to nanoscale

However,

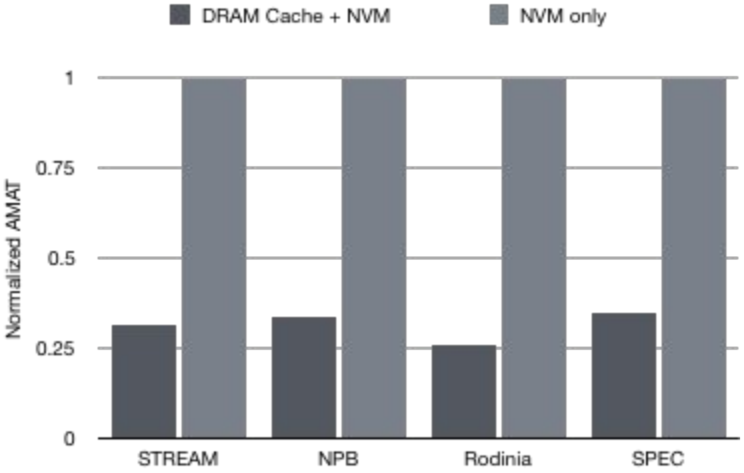
#1. Slower Operations (read: 4x slower than DRAM; write: 10x) [1][2]

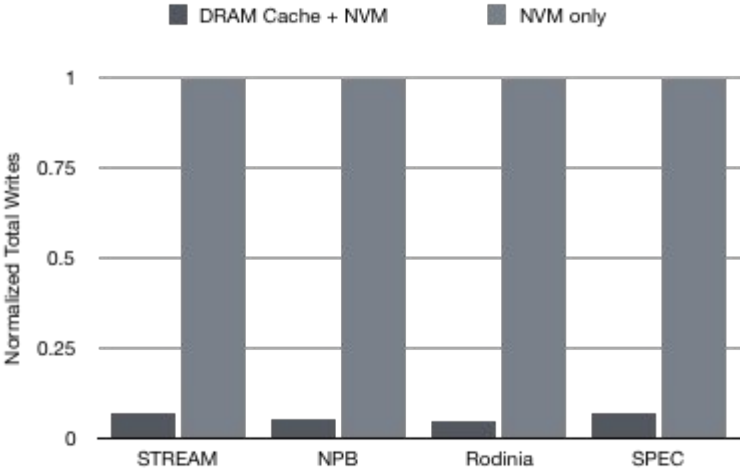
#2. Poor Write Endurance (10^8 vs 10^{16} on DRAM) [1]

#3. Data Security (Non-volatility) [3]

Solving NVM Challenges

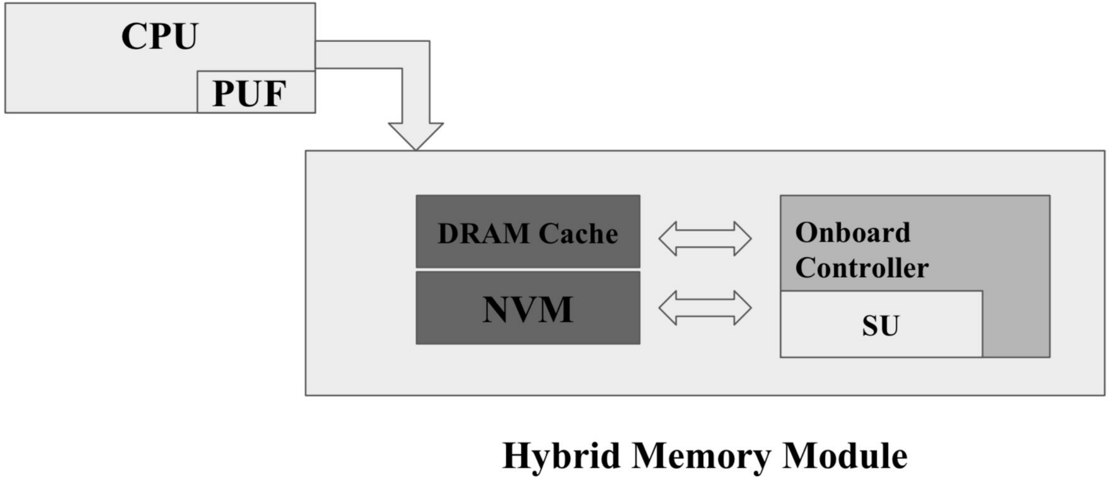
1.DRAM Cache to solve slower operation problem and poor write endurance problem





2.A hardware security support to solve the data security problem

Our Proposal



In this work, we design a secure hybrid memory module.

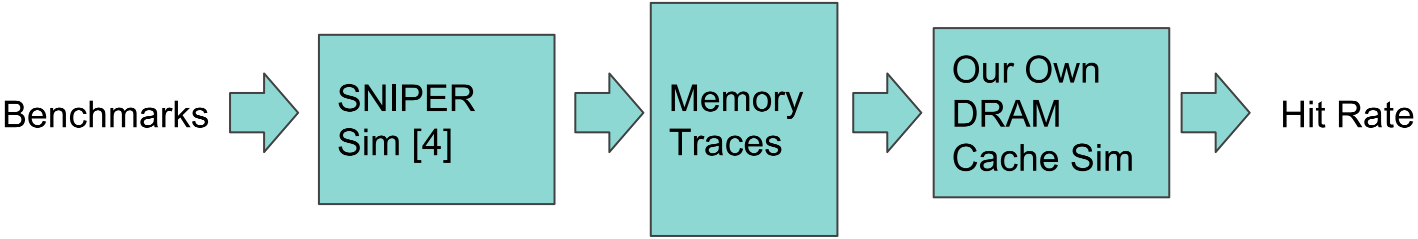
It includes NVM (PCM) as main memory, a DRAM cache and a security unit.

The security unit includes a AES-GCM [4] engine and a NVM vault (STT-RAM) for tags and counter values

It follows NVDIMM-N standard: A backup power source is built to transfer DRAM blocks into NVM when the system fails.

Methodology for DRAM Cache Design

#1. DRAM Cache Configuration: Evaluate Hit Rate for different combinations of DRAM cache parameters (associativity, cache line size and total size) and choose the one that is most beneficial



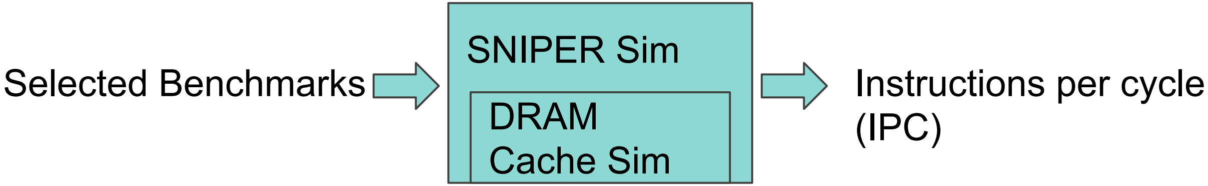
Benchmarks: SPEC CPU06 with single thread [5]

NPB [6], STREAM [7], Rodinia [8] with 4 threads

CPU Frequency	2.67GHz
L1, L2, L3 Cache	32KB, 256KB, 8MB
DRAM Row Size	2KB[9]
DRAM tCL-tRCD-tRP	13.75ns-13.75ns-13.75ns[9]
NVM Main Memory Size	8GB
NVM tCL-tRCD-tRP	13.75ns-55ns-150ns[2]
NVM Row Size	2KB[2]
DDR Frequency/period	1600MHz/0.625ns
AES Latency(128 bit block)-GHASH Latency(1KB)	10 memory cycles - 69 memory cycles [4]

Methodology on Performance Evaluation

#2 Performance Evaluation



Selected Benchmarks (memory bound benchmarks)

SPEC CPU06: gobmk, sjeng, gcc, wrf, zeusmp, GemsFDTD, lbm, mcf, soplex (single thread)

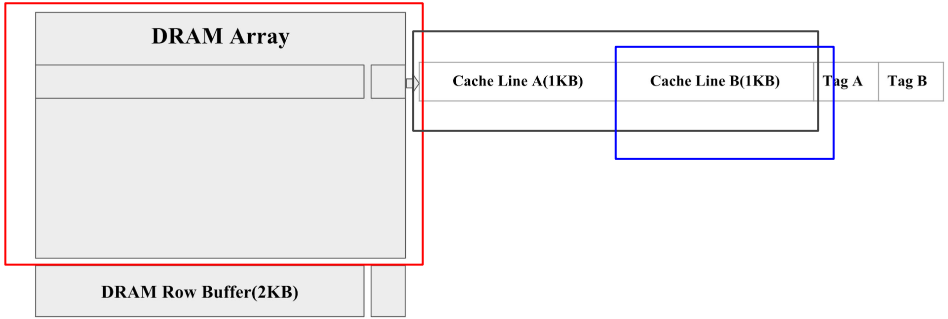
STREAM: STREAM (4 threads)

Rodinia: backprop, nw, hotspot, bfs, cfd (4 threads)

NPB: is, cg, ua, mg, lu (4 threads)

DRAM Cache Parameters

Three parameters should be determined to design a DRAM cache: Associativity, Cache Line Size and Total Size



We suggest 2-way associativity, 1KB cache line and 256MB total size

Associativity

For SPEC and STREAM, multiway cache doesn't provide better hit rate. For NPB and Rodinia, the difference of hit rates between 1-way to 32-way is less than 5%

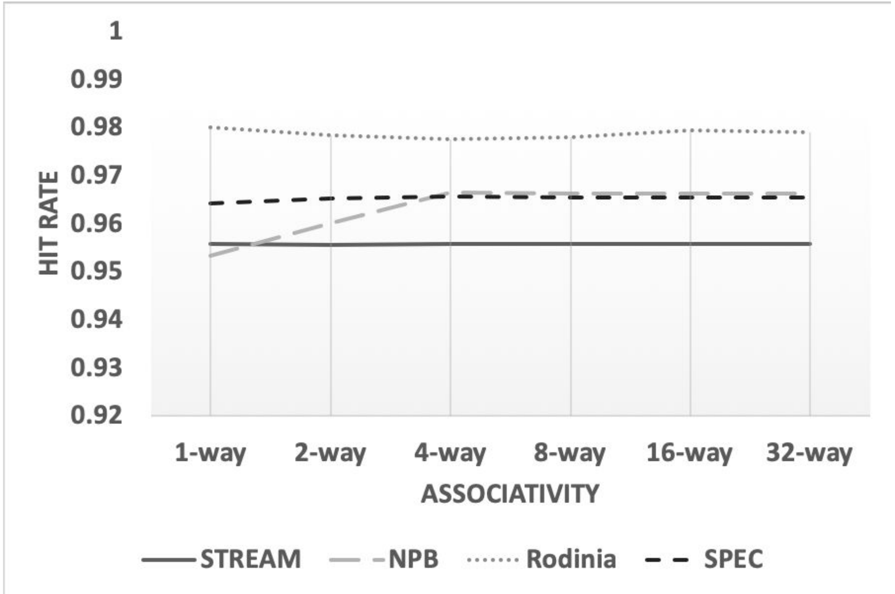
We suggest 2-way as associativity:

#1. 2-way is more effective against write attack than 1-way

#2. More than 2-way is more complex to design for LRU replacement policy

Cache Total Size: 256MB
Cache Line Size: 1024B

Average Hit Rate



Cache Line Size

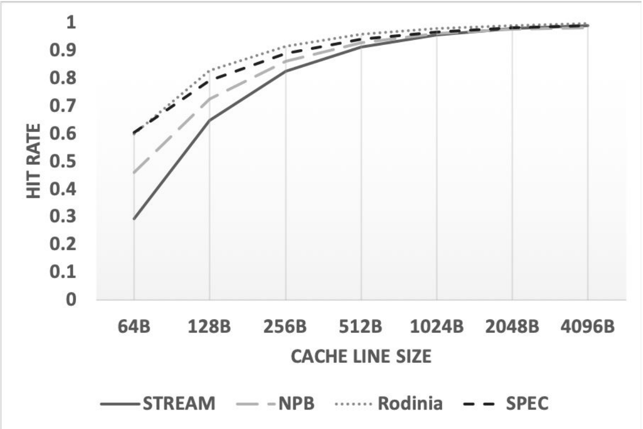
For all benchmarks, the hit rate increases until it reaches 1KB

However, larger cache line needs more time to transfer.

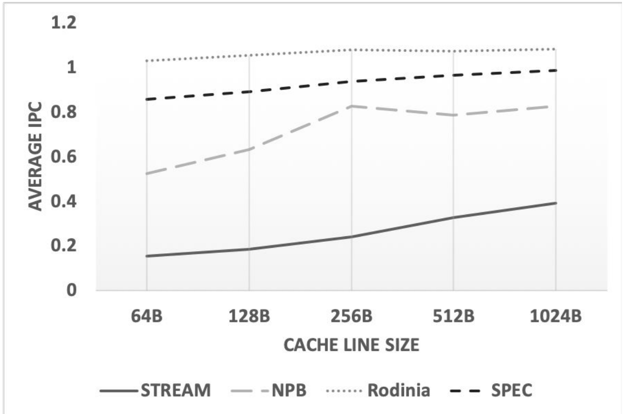
We also evaluate the system performance for different cache line sizes. 1KB does provide better IPC

Cache Total Size: 256MB
Associativity: 2-way

Average Hit Rate



Average IPC



Total Size

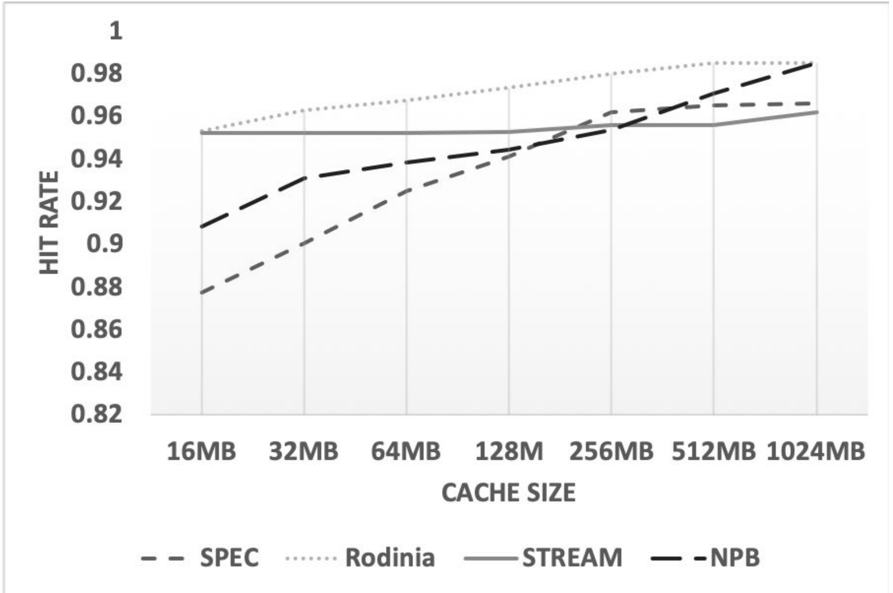
For STREAM, larger size has no benefit. For SPEC and Rodinia, 256MB is sufficient. For NPB, larger size could always provide better hit rate.

We suggest 256MB as the total size.

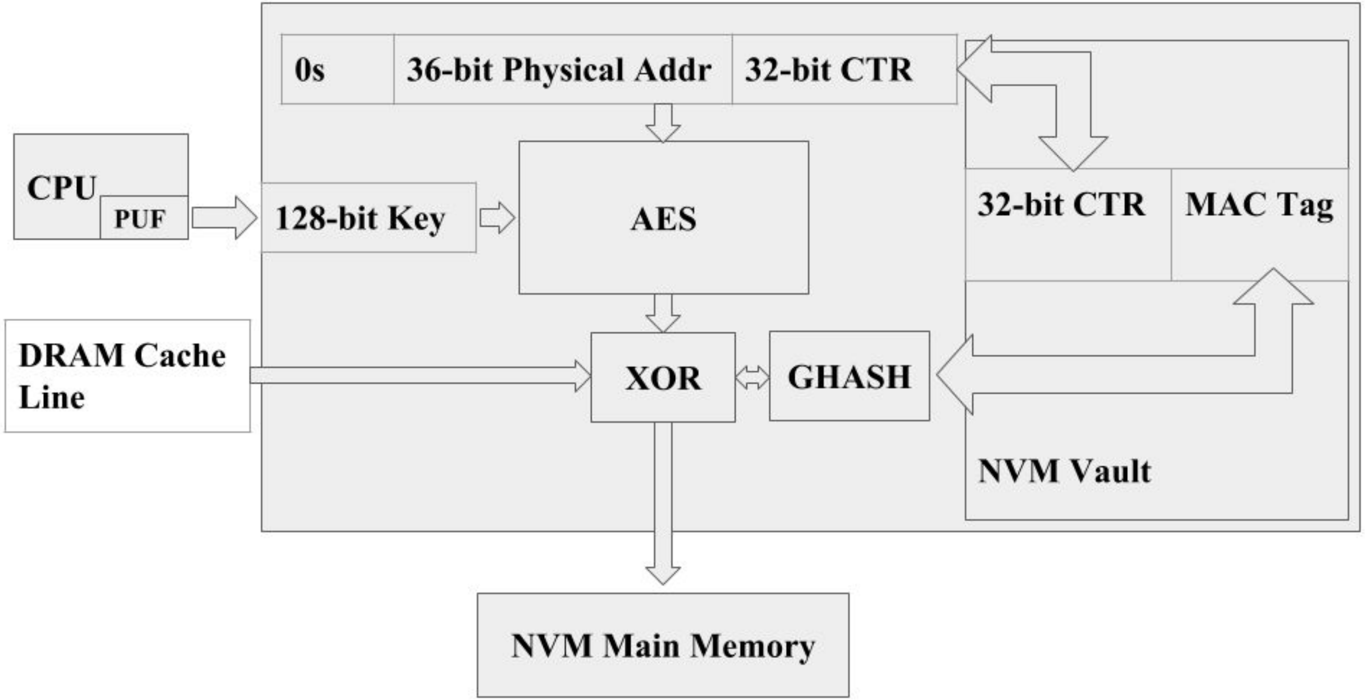
Some previous research also chose 256MB as the DRAM cache size [9][10][15]

Cache Associativity: 2-way
Cache Line Size: 1024B

Average Hit Rate



Overview of Security Unit



Putting It All Together

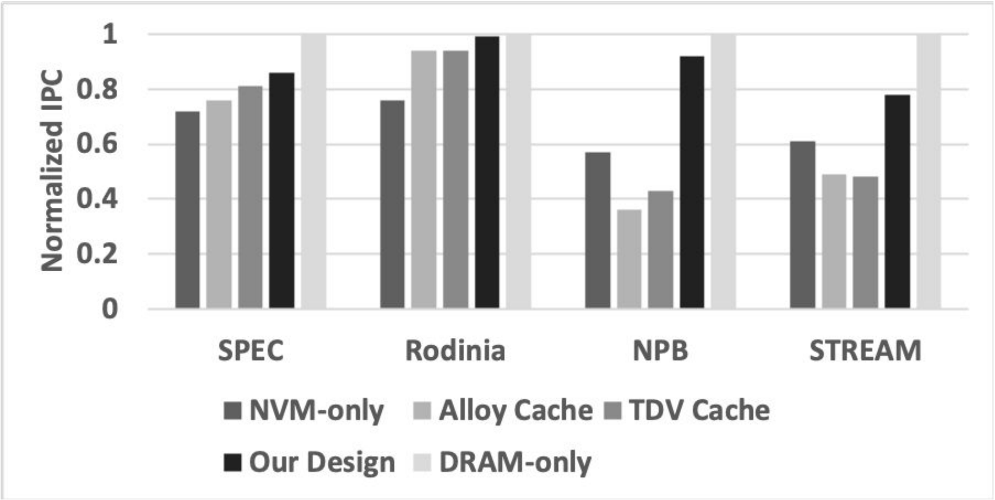
SPEC CPU06: gobmk, sjeng, gcc, wrf, zeusmp,
GemsFDTD, lbm, mcf, soplex (single thread)
STREAM: STREAM (4 threads)
Rodinia: backprop, nw, hotspot, bfs, cfd (4 threads)
NPB: is, cg, ua, mg, lu (4 threads)

Compared with NVM-only and DRAM-only

Overall, it improves the performance by 32% compared to NVM-only Memory Module, and only is 6.8% slower than DRAM-only Memory Module

Our DRAM Cache design is also faster than Alloy memory system [9] and TDV memory system [10] by 16.9% and 13.5%, respectively

Normalized IPC



Reference

[1]S. Yu and P. Chen, “Emerging Memory Technologies: Recent Trends and Prospects”, IEEE Solid-State Circuits Magazine, vol. 8, no. 2, pp. 43-56, 2016

[2]Benjamin C. Lee et al, “Architecting Phase Change Memory as a Scalable Dram Alternative”, ISCA ‘09, Proceedings of the 36th Symposium on Computer Architecture, pp.2-13, 2009

[3]Sparsh Mittal and Ahmed Izzat Alsalibi, “A Survey of Techniques for Improving Security of Non-Volatile Memories”, Journal of Hardware and System Security, vol. 2, issue 2, pp. 179-200, 2018

[4]Karim M.Abdellatif, Roselyne Chotin-Avot and Habib Mehrez, “AES-GCM and AEGIS: Efficient and High Speed Hardware Implementations”, Journal of Signal Processing Systems, v.88, Issue 1, pp. 1-12, 2017

[5]https://www.spec.org/cpu/

[6]https://www.nas.nasa.gov/publications/npb.html

[7]John D.McCalpin, “Memory Bandwidth and Machine Balance in Current High Performance Computers”, IEEE Computer Society Technical Committee on Computer Architecture Newsletter, 1995

[8]Shuai Che, Michael Boyer, Jiayuan Meng, David Tarjan, Jeremy W.Sheaffer, Sang-Ha Lee and Kevin Skadron, “Rodinia: A Benchmark Suit for Heterogeneous Computing”, IISWC ’09, Proceedings of the 2009 IEEE International Symposium on Workload Characterization, pp. 44-54, 2009

[9]M. K. Qureshi and G. H. Loh, “Fundamental Latency Trade-off in Architecting DRAM Caches: Outperforming Impractical SRAM-Tags with a Simple and Practical Design”, MICRO ‘12, Proceedings of 45th Annual IEEE/ACM International Symposium on Microarchitecture, pp. 235-246, 2012

[10]T. Lu, Y. Liu, H. Pan and M. Chen, “TDV Cache: Organizing Off-Chip DRAM Cache of NVMM from a Fusion Perspective”, ICCD ‘17, Proceedings of IEEE International Conference on Computer Design, pp. 65-72, 2017

[11]Djordje Jevdjic, Stavros Volos, and Babak Falsafi, “Die-Stacked DRAM Caches for Servers: Hit Ratio, Latency, or Bandwidth? Have It All with Footprint Cache”, ISCA ’13, Proceedings of the 40th Annual International Symposium on Computer Architecture, pp.404-415, 2013

[12]Fangyong Hou and Hongjun He, “Ultra Simple Way to Encrypt Non-Volatile Main Memory”, Security and Communication Networks, vol. 8, issue 7, pp. 1155-1168, 2015

[13]Jingfei Kong and Huiyang Zhou, “Improving Privacy and Lifetime of PCM-Based Main Memory”, DSN ‘10, Proceedings of IEEE/IFIP International Conference on Dependable Systems and Networks, pp. 333-342, 2010

[14]S. Chhabra and Y. Solihin, “i-NVM: A Secure Non-Volatile Main Memory System with Incremental Encryption”, ISCA ‘11, Proceedings of the 38th Symposium on Computer Architecture, pp.177-188, 2011

[15]Xiaobing Lee, Florian Longnos, Shaojie Chen and Wei Yang, “DDR4 DIMM Devices with Hybrid DRAM & NVM for Big Dara Performances at Low Cost”, 2016

Appendix:AES-GCM Parameters

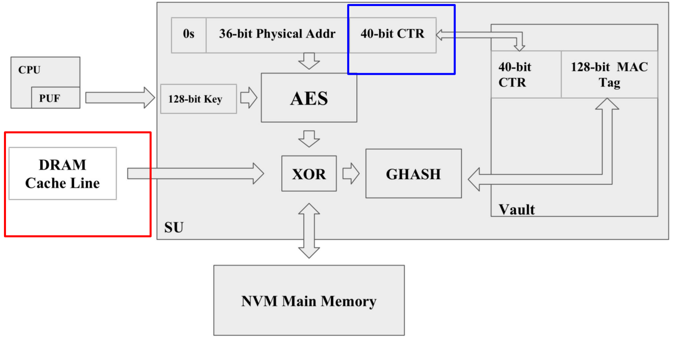
Encryption Granularity

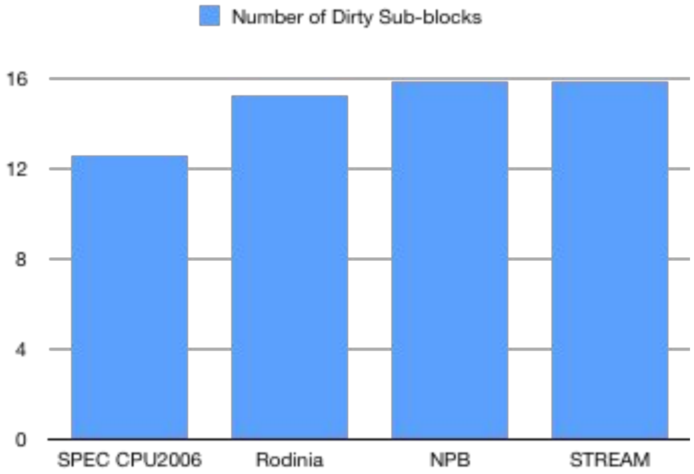
We divide 1024B cache line into 64B subblocks and see how many dirty subblocks are there when a DRAM cache line is written back. It turns out most of cases, more than 12 subblocks are dirty when a 1024B cache line written back.

Counter Value

When a dram cache line is written back, its counter value is incremented by 1

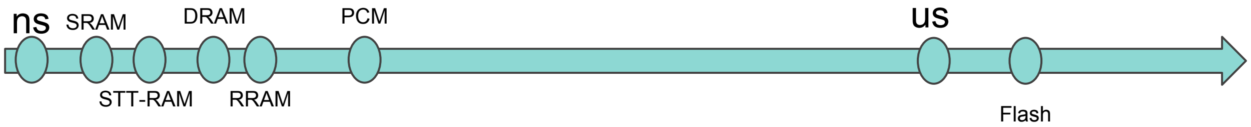
A NVM cell could be written 10^8 and 27-bit counter value is enough..





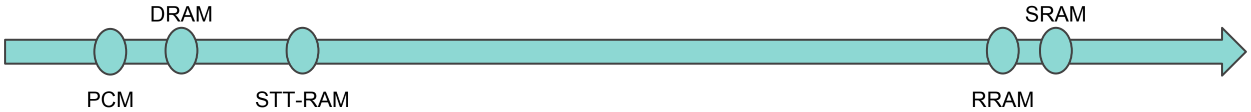
Appendix: RAM Technology Comparison

Latency

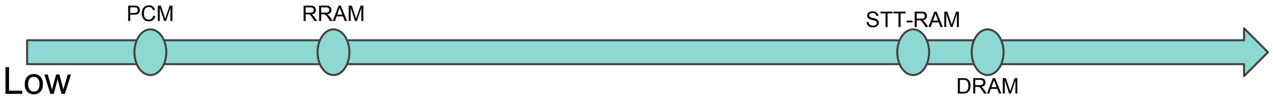


PCM:
Phase Change Memory
STT-RAM:
Spin-Transfer Torque
RRAM
Resistive RAM

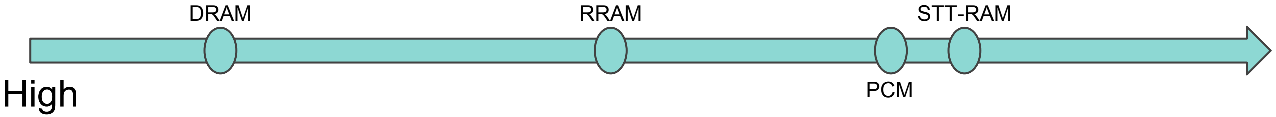
Cost



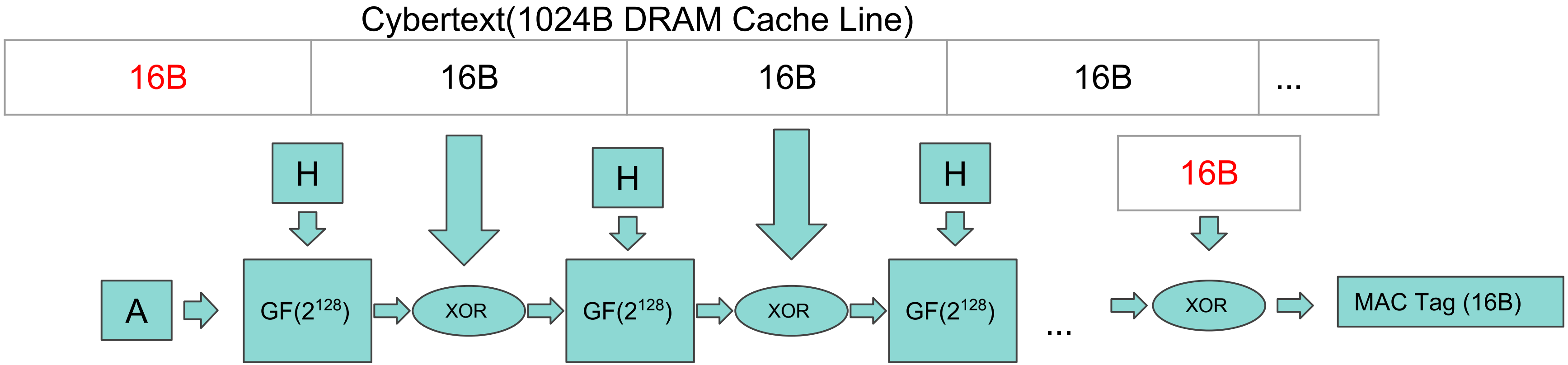
Write Endurance



Density



Appendix: GHASH



A : Associated Data(User Password)

H : AES(0000…)$_{PUF}$

GF(2^{128}) : Galois Field Multiplication with $P(x) = x^{128}+x^7+x^2+1$