CMPE 110: Computer Architecture

Week 8
Cache / Main Memory

Jishen Zhao (http://users.soe.ucsc.edu/~jzhao/)

[Adapted in part from Jose Renau, Mary Jane Irwin, Joe Devietti, Onur Mutlu, and others]

Reminder

- Quiz 3 will be posted today
  - Due at 11:59pm on Nov. 9.

Review

- Handling cache hit and miss
  - Read hit
  - Write hit: write-back cache vs. write-through cache
  - Cache miss handling: cache controller
- Cache performance, and AMAT
  - $t_{avg} = t_{hit} + \%_{miss} * t_{miss}$
  - AMAT = t_{avg}(L1)
  - Calculate CPI
- Types of caches

Types of cache

Directly vs. Associative Mapped Caching

- Direct mapped caching allows any given main memory block to be mapped into exactly one unique cache location.

- Set-associative mapped cache allows any given main memory block to be mapped into two or more cache locations.

- Fully-associative mapped caching allows any given main memory block to be mapped into any cache location.

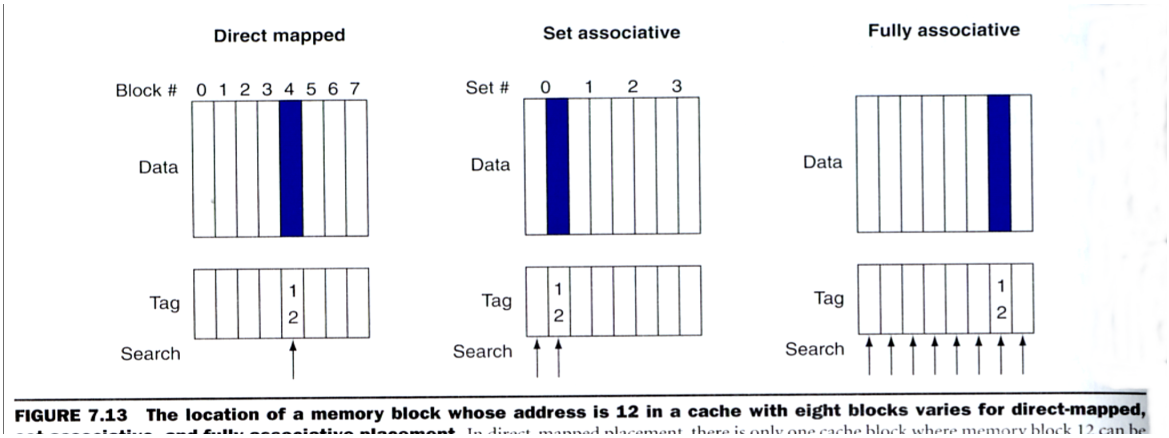


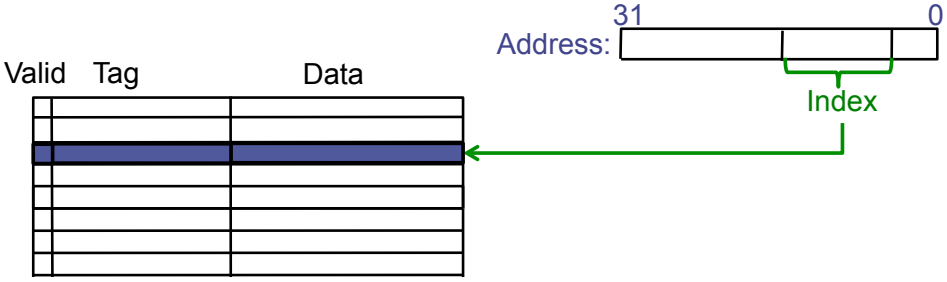
FIGURE 7.13 The location of a memory block whose address is 12 in a cache with eight blocks varies for direct-mapped,

Directly vs. Associative Mapped Caching

- Direct mapped caching allows any given main memory block to be mapped into exactly one unique cache location.

- Set-associative mapped cache allows any given main memory block to be mapped into two or more cache locations.

- Fully-associative mapped caching allows any given main memory block to be mapped into any cache location.

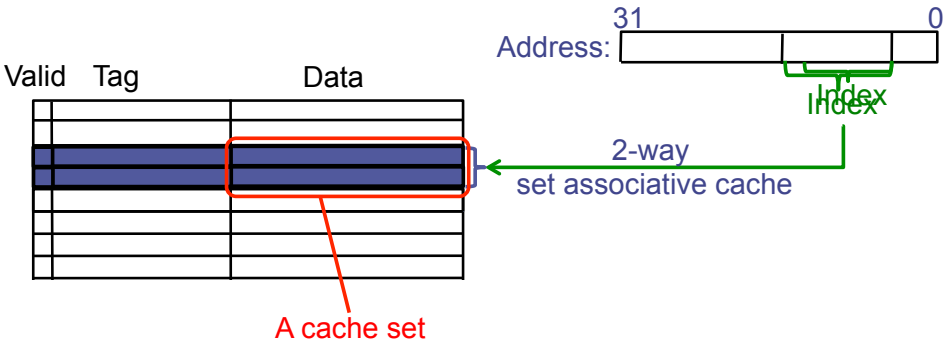


Directly vs. Associative Mapped Caching

- Direct mapped caching allows any given main memory block to be mapped into exactly one unique cache location.

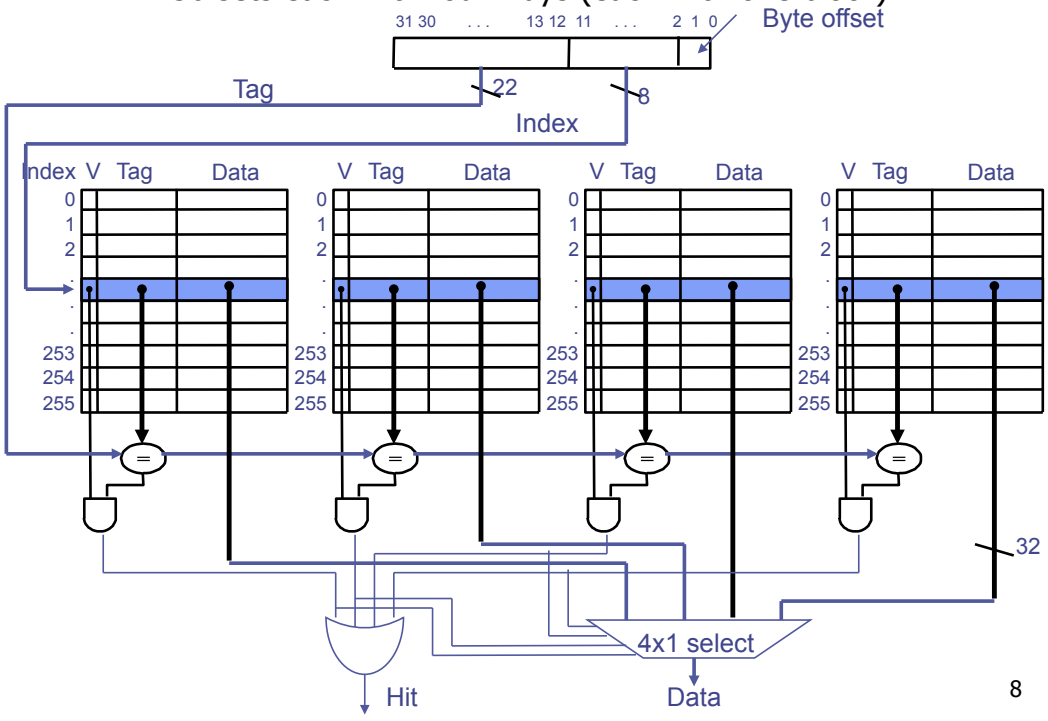
- Set-associative mapped cache allows any given main memory block to be mapped into two or more cache locations.

- Fully-associative mapped caching allows any given main memory block to be mapped into any cache location.



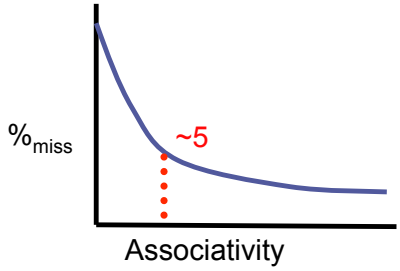
Example: 4-Way Set Associative Cache

- 2^8 = 256 sets each with four ways (each with one block)



Associativity and Performance

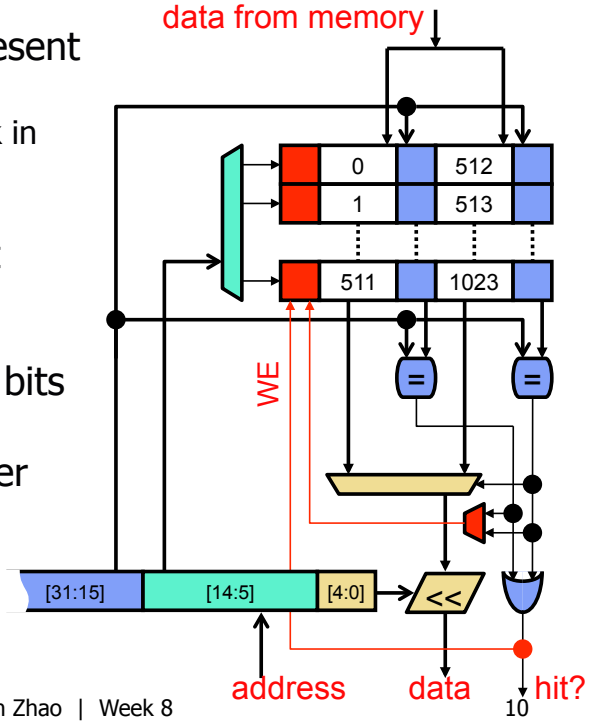
- Higher associative caches
  - + Have better (lower) $\%_{miss}$
    - Diminishing returns
  - – However t_{hit} increases
    - The more associative, the slower
  - What about t_{avg}?



- Block-size and number of sets should be powers of two
  - Makes indexing easier (just rip bits out of the address)
- 3-way set-associativity? No problem

Miss Handling & Replacement Policies

- Set-associative caches present a new design choice
  - On cache miss, which block in set to replace (kick out)?
- Add **LRU** field to each set
  - "Least recently used"
- Each access updates LRU bits
- Pseudo-LRU used for larger associativity caches

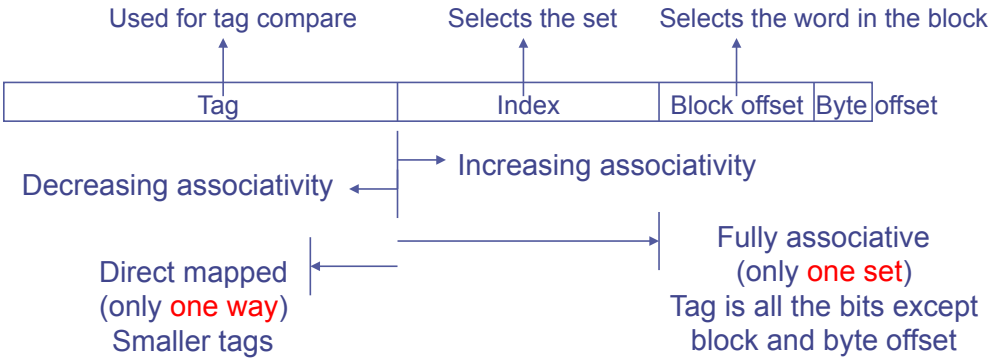


Replacement Policies

- Set-associative caches present a new design choice
  - On cache miss, which block in set to replace (kick out)?
- Some options
  - **Random**
  - **FIFO (first-in first-out)**
  - **LRU (least recently used)**
    - Fits with temporal locality, LRU = least likely to be used in future
  - **NMRU (not most recently used)**
    - An easier to implement approximation of LRU
    - Same as LRU for 2-way set-associative caches
  - **Belady's**: replace block that will be used furthest in future
    - Unachievable optimum

Range of Set Associative Caches

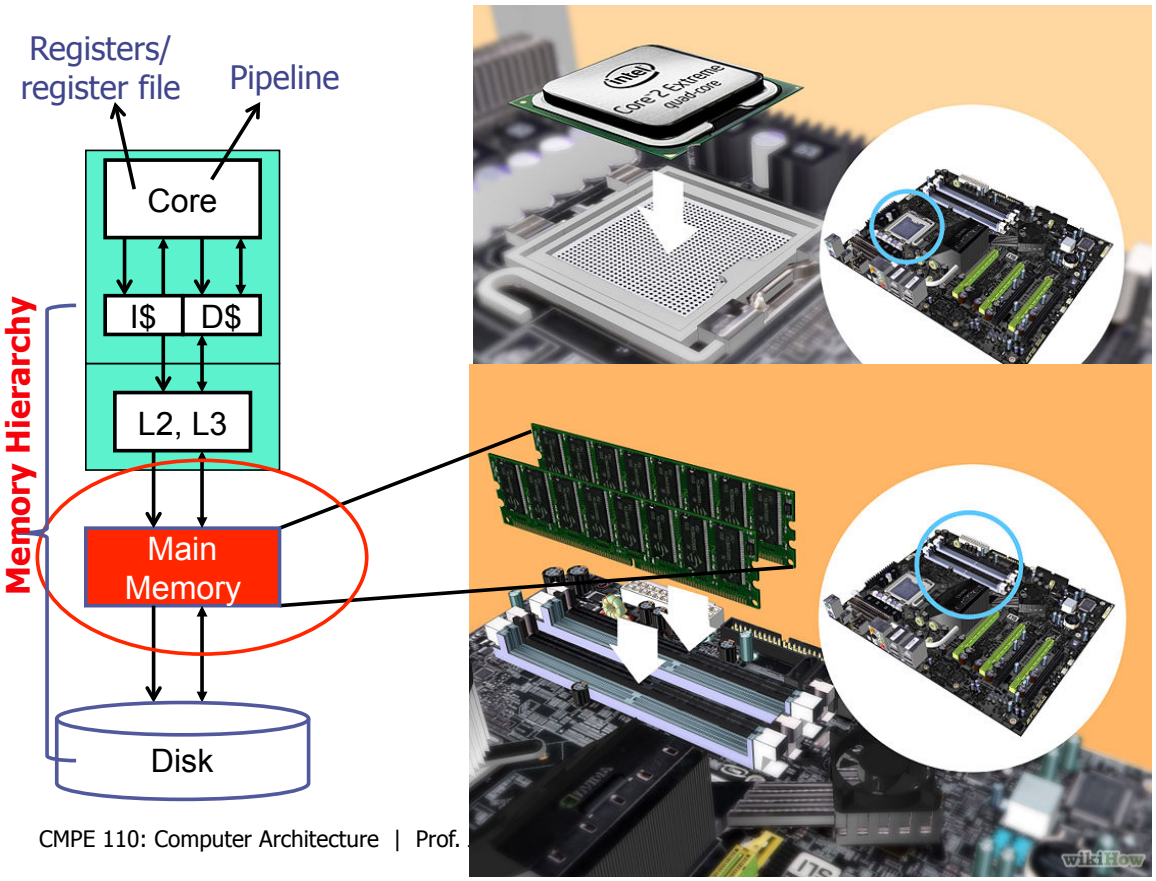
- For a fixed size cache, each increase by a factor of two in associativity doubles the number of blocks per set (i.e., the number or ways) and halves the number of sets – decreases the size of the index by 1 bit and increases the size of the tag by 1 bit



Classifying Misses: 3C Model

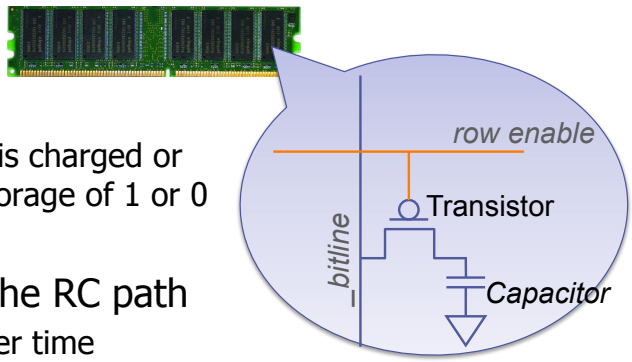
- Divide cache misses into three categories
  - **Compulsory (cold)**: never seen this address before
    - **Would miss even in infinite cache**
  - **Capacity**: miss caused because cache is too small
    - **Would miss even in fully associative cache**
    - Identify? Consecutive accesses to block separated by access to at least N other distinct blocks (N is number of blocks in cache)
  - **Conflict**: miss caused because cache associativity is too low
    - Identify? **All other misses**

Main Memory



Review: Main Memory technology -- DRAM

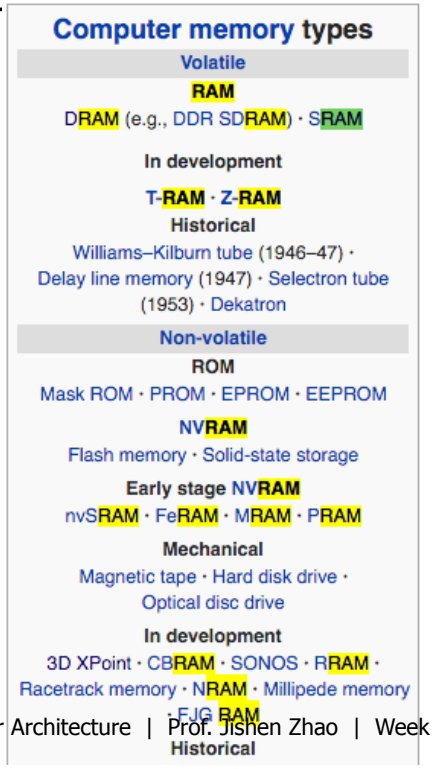
- Dynamic random access memory
- Capacitor charge state indicates stored value
  - 1T1C
  - 1 access transistor
  - 1 capacitor
    - Whether the capacitor is charged or discharged indicates storage of 1 or 0



- Capacitor leaks through the RC path
  - DRAM cell loses charge over time
  - DRAM cell needs to be refreshed

Other memory technologies (not required materials)

Other memory technologies



Source: Wikipedia

Memory technologies

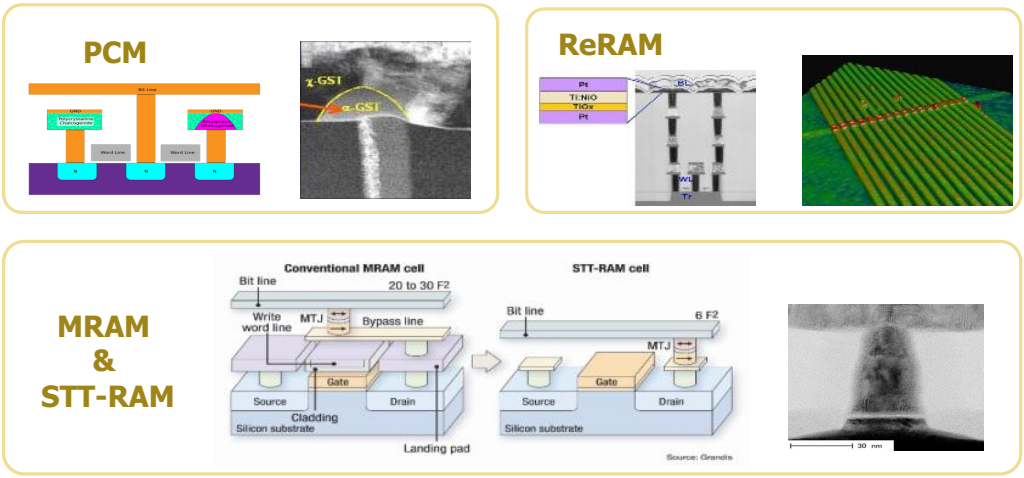
RAM: random access memory – Accesses (reads/writes) in almost the same amount of time irrespective of the physical location of data inside the memory

- DRAM (dynamic – need refresh)
- SRAM (static – does not need refresh)
- Nonvolatile RAM (NVRAM) (does not need refresh)
  - Spin-transfer torque magnetic RAM (STT-MRAM, a type of MRAM)
  - Phase-change memory (PCRAM or PRAM or PCM)
  - Resistive RAM (ReRAM)

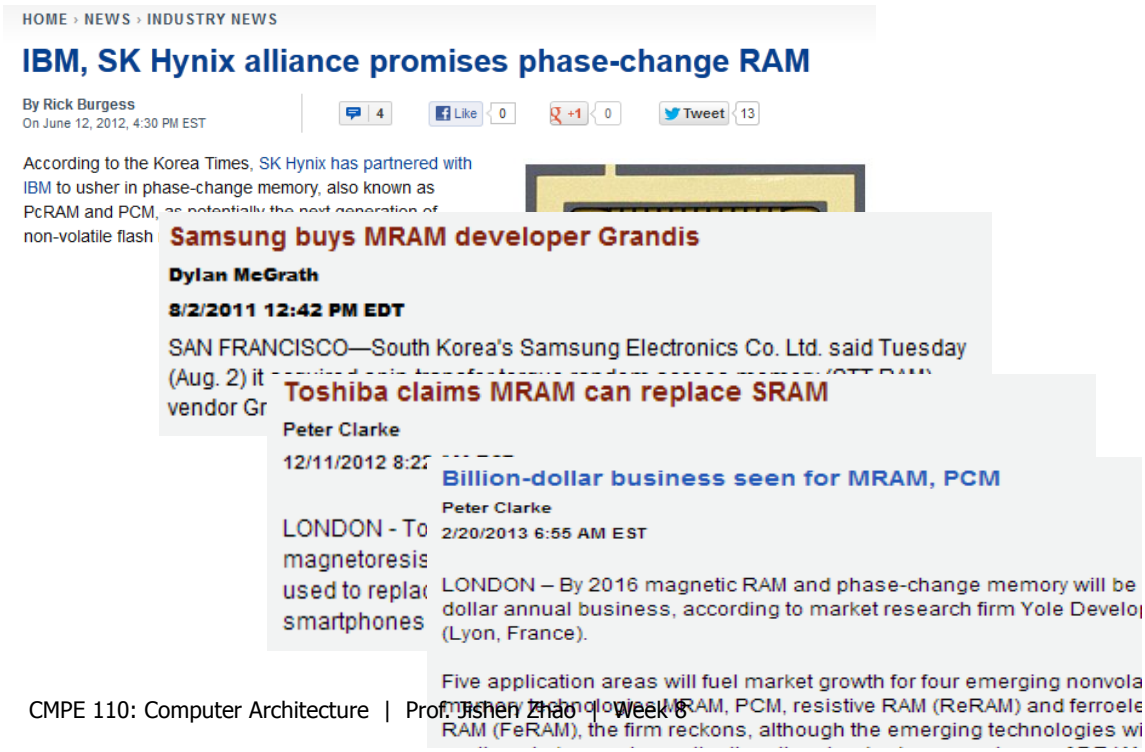
- Flash – NAND flash and NOR flash – flash is not RAM
  - Flash is often called nonvolatile memory (NVM), but actually not a "memory"
- ROM (read-only memory) – not RAM

Nonvolatile RAMs (NVRAMs)

- Various types of NVRAMs with various physical cell devices



NVRAMs are hot in research -- HW



HOME › NEWS › INDUSTRY NEWS

IBM, SK Hynix alliance promises phase-change RAM

By Rick Burgess
On June 12, 2012, 4:30 PM EST

4 Like 0 +1 0 Tweet 13

According to the Korea Times, SK Hynix has partnered with IBM to usher in phase-change memory, also known as PcRAM and PCM, as potentially the next generation of non-volatile flash

Samsung buys MRAM developer Grandis

Dylan McGrath

8/2/2011 12:42 PM EDT

SAN FRANCISCO—South Korea's Samsung Electronics Co. Ltd. said Tuesday (Aug. 2) it vendor Gr

Toshiba claims MRAM can replace SRAM

Peter Clarke

12/11/2012 8:27

LONDON - To magnetoresis used to repla smartphones

Billion-dollar business seen for MRAM, PCM

Peter Clarke

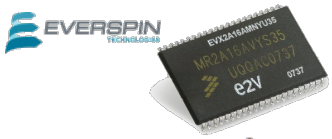
2/20/2013 6:55 AM EST

LONDON – By 2016 magnetic RAM and phase-change memory will be dollar annual business, according to market research firm Yole Develo (Lyon, France).

Five application areas will fuel market growth for four emerging nonvola memory technologies: MRAM, PCM, resistive RAM (ReRAM) and ferroele RAM (FeRAM), the firm reckons, although the emerging technologies wi

NVRAM vendors

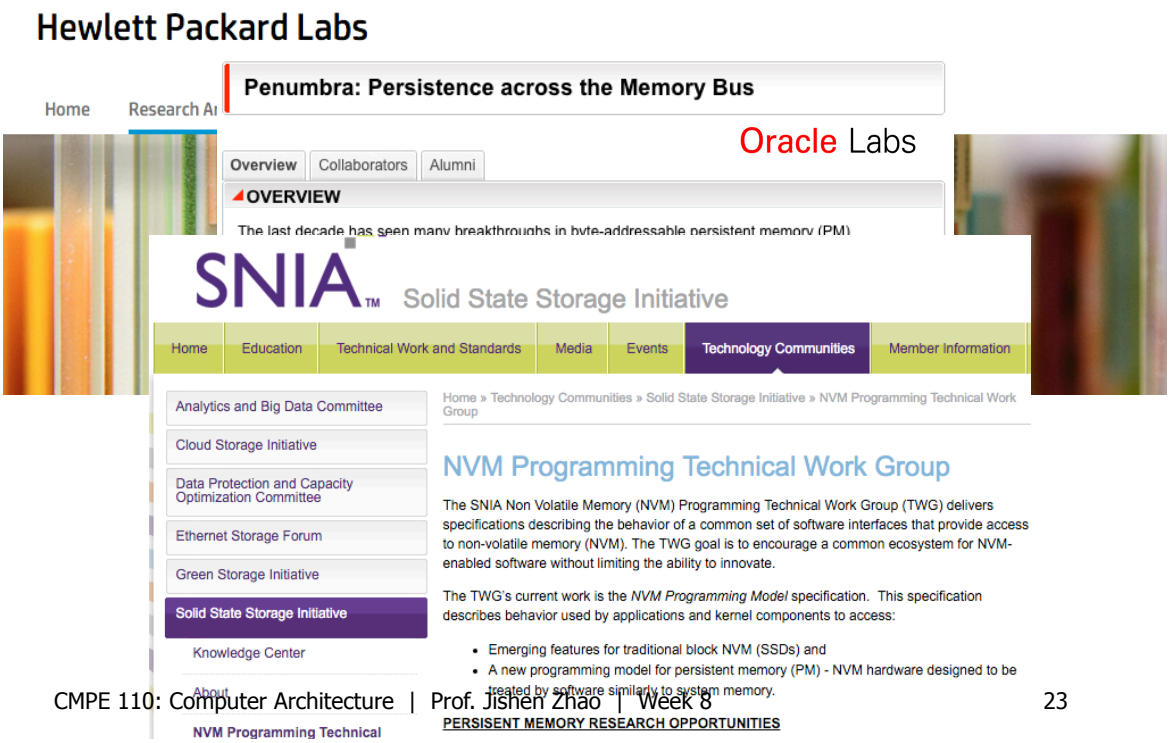
- Magnetic RAM (MRAM)
  - EverSpin (130nm, up to 16Mb)
- Spin-Transfer-Torque RAM (STTRAM)
  - Grandis (54nm, acquired by Samsung)
- Phase-Change RAM (PCRAM)
  - Samsung (20nm, diode, up to 8Gb)
- Resistive RAM (ReRAM)
  - Panasonic (180nm process, 4-layer xpoint)
  - Unity Semi (64MB, acquired by Rambus)



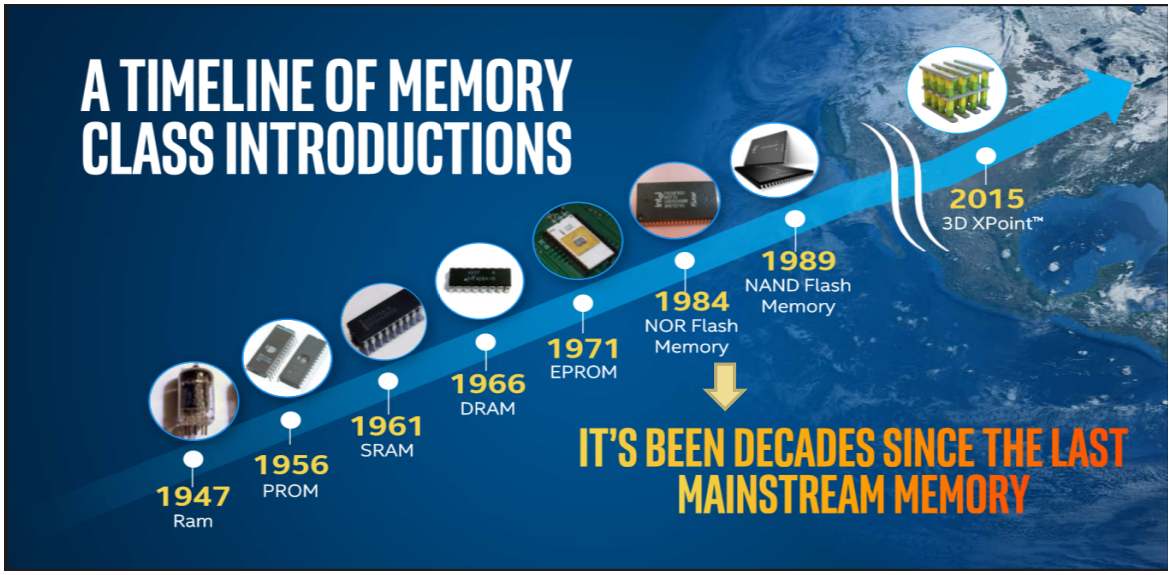
GRANDIS
RENESAS
TOSHIBA

hynix
hp
Panasonic
ideas for life
Rambus

NVRAMs are hot in research – SW & sys



NVRAM is coming to the market



NVRAM is coming to the market

Intel and Micron Produce Breakthrough Memory Technology

Posted by IntelPR in Intel Newsroom on Jul 28, 2015 9:00:28 AM

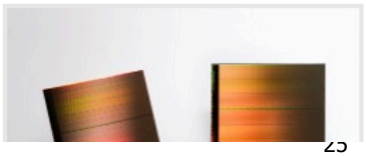
Tweet 1,208 Like 5.5k +1 514

New Class of Memory Unleashes the Performance of PCs, Data Centers and More

NEWS HIGHLIGHTS

- Intel and Micron begin production on new class of non-volatile memory, creating the first new memory category in more than 25 years.
- New 3D XPoint™ technology brings non-volatile memory speeds up to 1,000 times faster[1] than NAND, the most popular non-volatile memory in the marketplace today.
- The companies invented unique material compounds and a cross point architecture for a memory technology that is 10 times denser than conventional memory[2].
- New technology makes new innovations possible in applications ranging from machine learning to real-time tracking of diseases and immersive 8K gaming.

SANTA CLARA, Calif., and BOISE, Idaho, July 28, 2015 – Intel Corporation and Micron Technology, Inc. today unveiled 3D XPoint™ technology, a non-volatile memory that has the potential to revolutionize any device, application or service that benefits from fast access to large sets of data. Now in production, 3D XPoint technology is a major breakthrough in memory process technology and the first new memory category since the introduction of NAND flash in 1989.



NVRAM is coming to the market – STT-MRAM

April 18, 2016

Everspin Aims MRAM at SSD Storage Tiers

by Gary Hilson, EE Times, ***Designlines***

TORONTO – MRAM pioneer Everspin Technologies Inc. is continuing its efforts to expand the applications for its non-volatile memory by displacing DRAM as a persistent memory in enterprise storage applications.

In a telephone interview with EE Times, Everspin CEO Phill LoPresti said storage and servers OEMs have been evaluating its ST-MRAM products, and the company will deliver further density increases for its MRAM-based storage class memory (SCM) and expects to sample a 1Gb product based on its proprietary perpendicular magnetic tunnel junction (pMTJ) ST-MRAM later this year.

Read the entire interview from EE Times >

March 1, 2016

Everspin Produces 1Mb MRAM with Quad SPI Interface

Storage Newsletter March 1, 2016 - Everspin Technologies, Inc., has announced production availability of the fastest non-volatile memory with a Quad SPI interface, the MR10Q010, 1Mb QSPI MRAM.

This high performance device can read and write data at 104MHz without the write delays encountered in other non-volatile technology such as NOR Flash

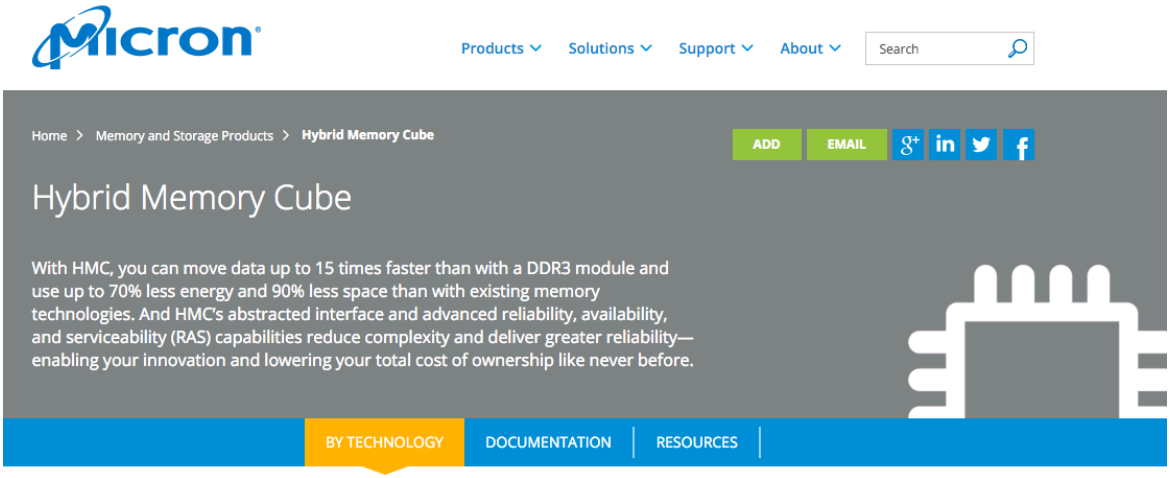
Other new memory tech research

Active Memory Cube

Ravi Nair

IBM T. J. Watson Research Center

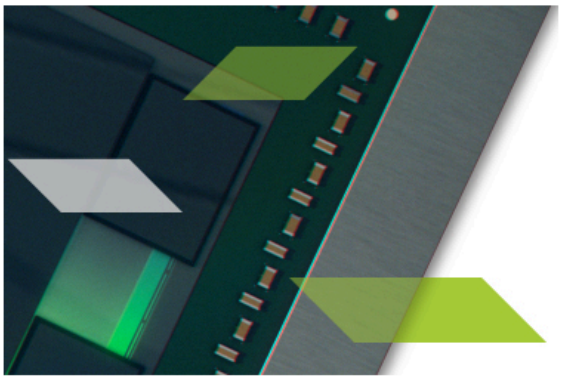
December 14, 2014



By Technology

AMD > Innovations > AMD Technologies + Software > High Bandwidth Memory

High Bandwidth Memory | Reinventing Memory Technology



Another first for AMD with High Bandwidth Memory (HBM)

The latest invention in AMD's long history of pioneering open standard technologies that push the boundaries of what is possible

Project Zero

News and updates from the Project Zero team at Google

Monday, March 9, 2015

Exploiting the DRAM rowhammer bug to gain kernel privileges

Posted by Mark Seaborn, sandbox builder and breaker, with contributions by Thomas Dullien, reverse engineer

[This guest post continues Project Zero's practice of promoting excellence in security research on the Project Zero blog]

Overview

"Rowhammer" is a problem with some recent DRAM devices in which repeatedly accessing a row of memory can cause bit flips in adjacent rows. We tested a selection of laptops and found that a subset of them exhibited the problem. We built two working privilege escalation exploits that use this effect. One exploit uses rowhammer-induced bit flips to gain kernel privileges on x86-64 Linux when run as an unprivileged userland process. When run on a machine vulnerable to the rowhammer problem, the process was able to induce bit flips in page table entries (PTEs). It was able to use this to gain write access to its own page table, and hence gain read-write access to all of physical memory.

We don't know for sure how many machines are vulnerable to this attack, or how many existing vulnerable machines are fixable. Our exploit uses the x86 `CLFLUSH` instruction to generate many accesses to the underlying DRAM, but other techniques might work on non-x86 systems too.