CMPE 110: Computer Architecture

Week 8
Main Memory

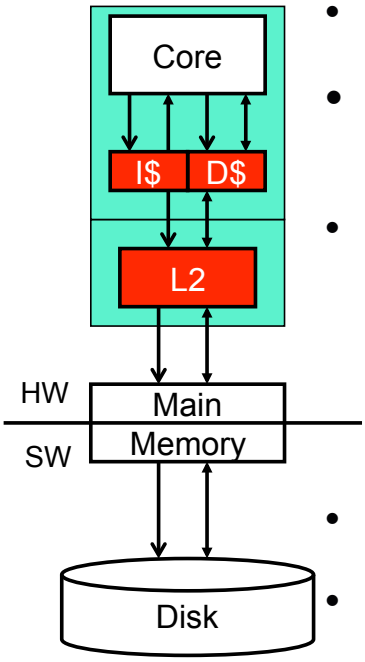
Jishen Zhao (http://users.soe.ucsc.edu/~jzhao/)

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

Reminder

- Quiz 3 due at 11:59pm on Nov. 9
  - Both questions shown on a single page
- No class on Friday Nov. 11 (holiday)

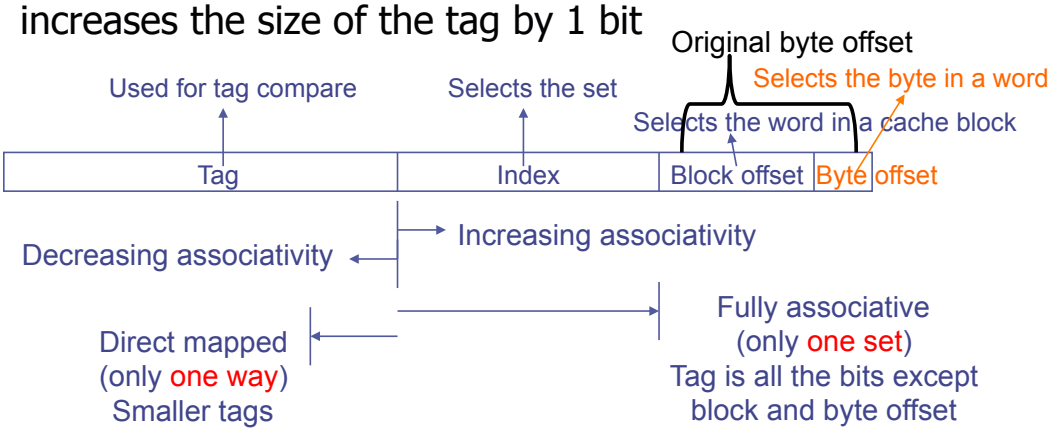
Review: Cache



- **Memory hierarchy**
  - Smaller, faster, expensive → bigger, slower, cheaper
- "Cache": hardware managed
  - Hardware automatically retrieves missing data
  - Built from fast on-chip SRAM
- Cache ABCs (**associativity, block size, capacity**)
  - How to access cache
  - Directly-mapped, set-associative, fully-associative caches
  - Handling cache hits and cache misses
  - Types of cache misses: 3C model
  - Cache replacement policy
- **Handling write hits**
  - Write-back vs. write-through
- **Average access time** of a memory component
  - $t_{avg} = t_{hit} + (\%_{miss} * t_{miss})$

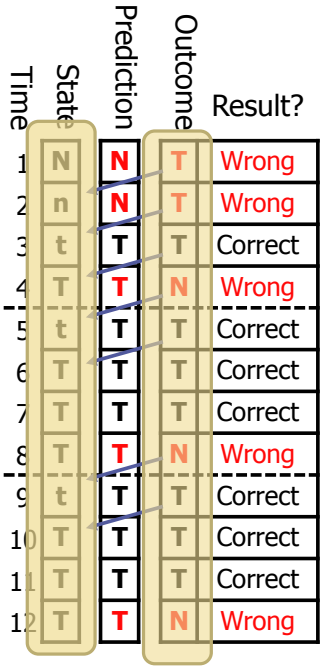
Review: Range of Set Associative Caches

- 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

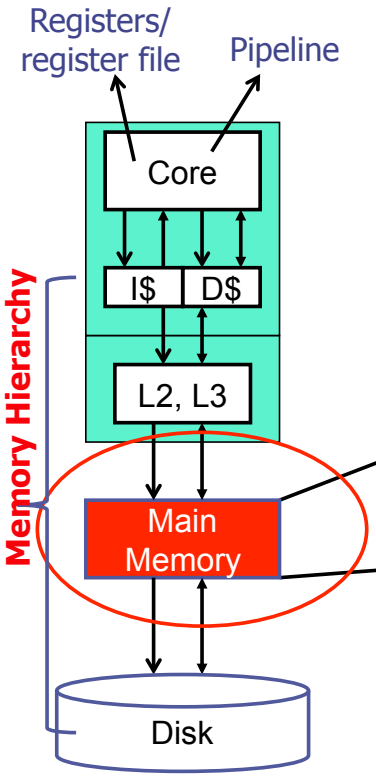


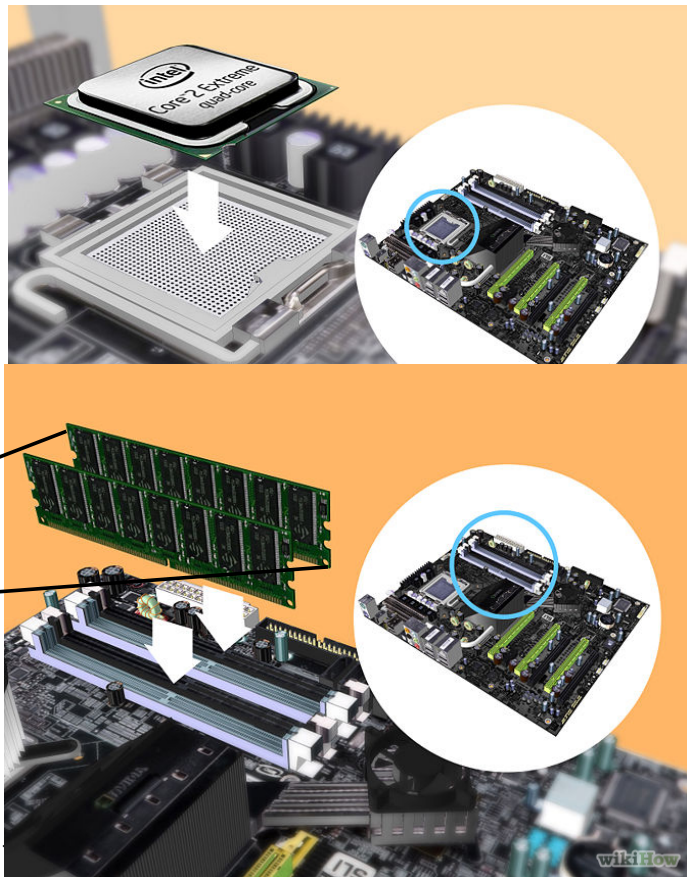
Review: Two-bit history-based branch prediction

- **Two-bit saturating counters (2bc)** [Smith 1981]
  - Replace each single-bit prediction
    - (0,1,2,3) = (N,n,t,T)
  - Adds "hysteresis"
    - Force predictor to mis-predict twice before "changing its mind"
  - One misprediction each loop execution (rather than two)
    - \+ Fixes this pathology (which is not contrived, by the way)
    - Can we do even better?



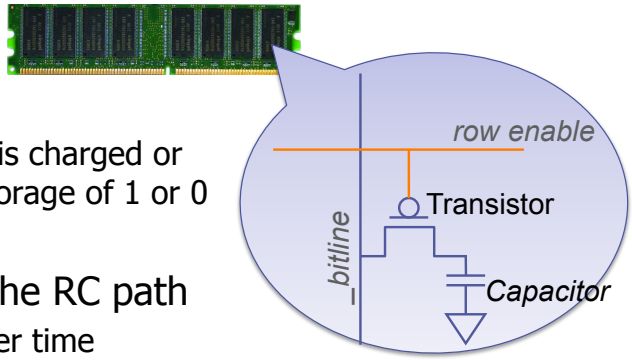
Main Memory





Review: Main Memory technology -- DRAM

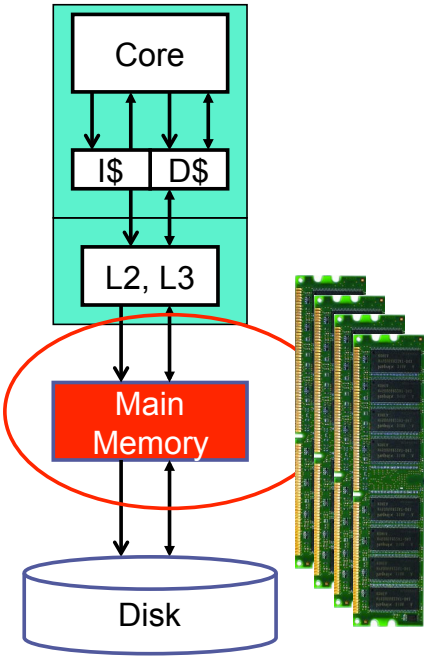
- Dynamic random access memory
- Capacitor charge state indicates stored value
  - 1T1C
  - 1 access transistor
  - 1 capacitor
    - Whether the capacitor is charged or discharged indicates storage of 1 or 0



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

Today: Main memory organization

- Channel
- DIMM
- Rank
- Chip
- Bank
- Row/column



Optional reading

Memory Systems

1st Edition

Cache, DRAM, Disk

Authors: Bruce Jacob, Spencer Ng, David Wang

Hardcover ISBN: 9780123797513
eBook ISBN: 9780080553849

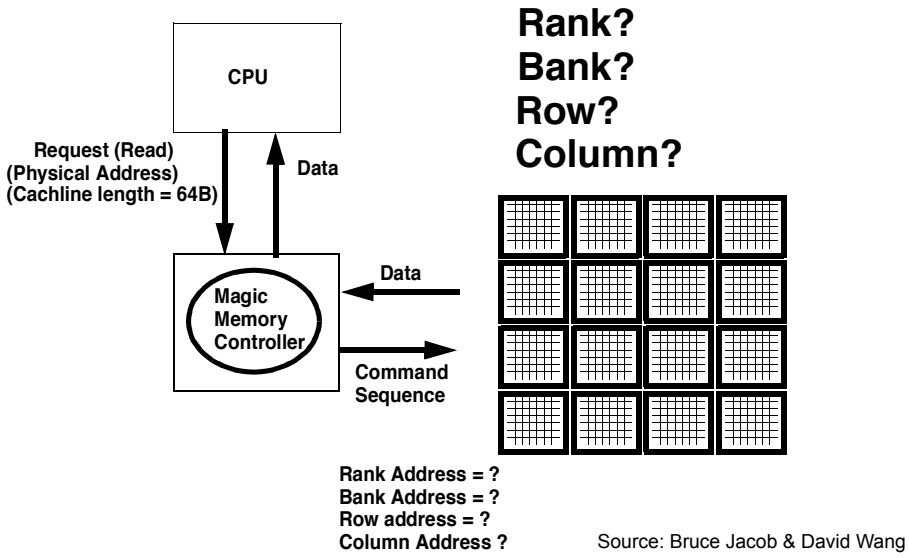
View on ScienceDirect

Imprint: Morgan Kaufmann
Published Date: 17th September 2007

Page Count: 900

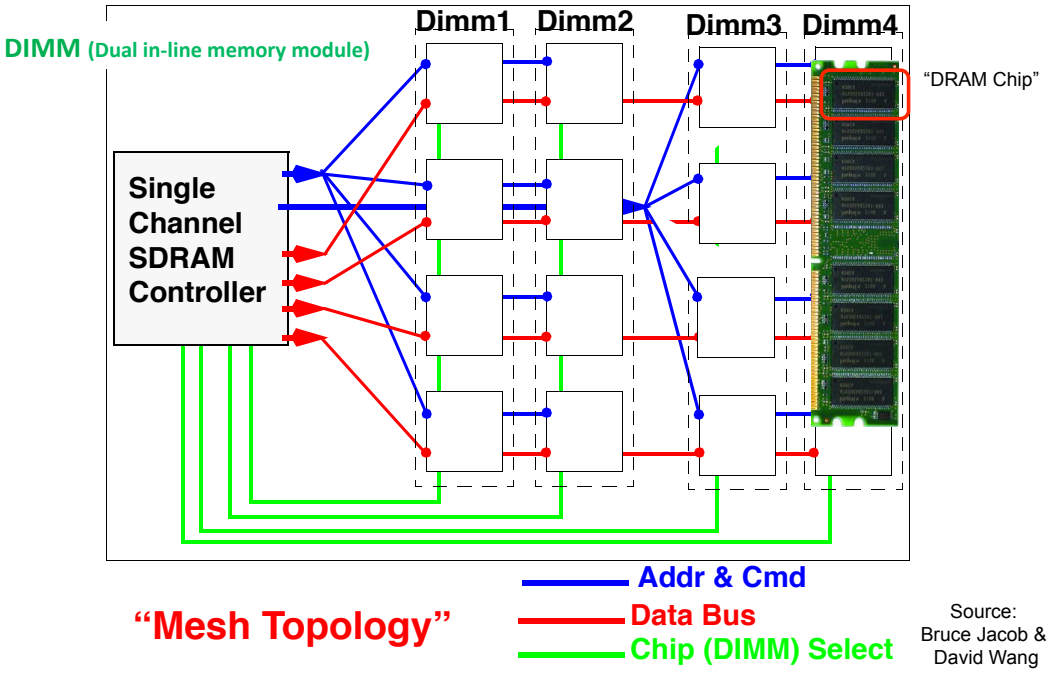
Memory system organization

Where is the data?



Source: Bruce Jacob & David Wang

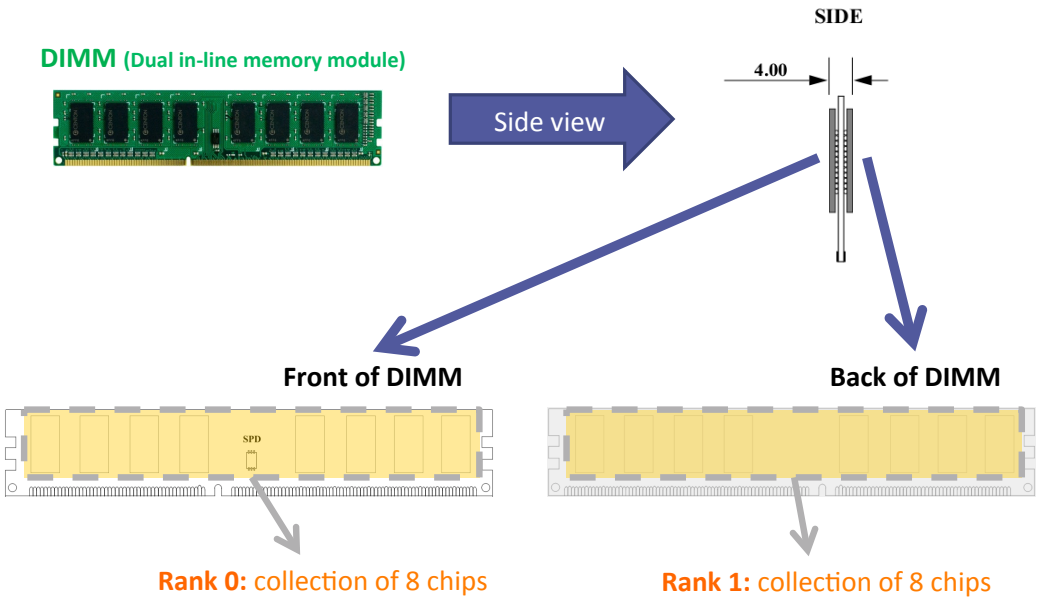
Memory system organization



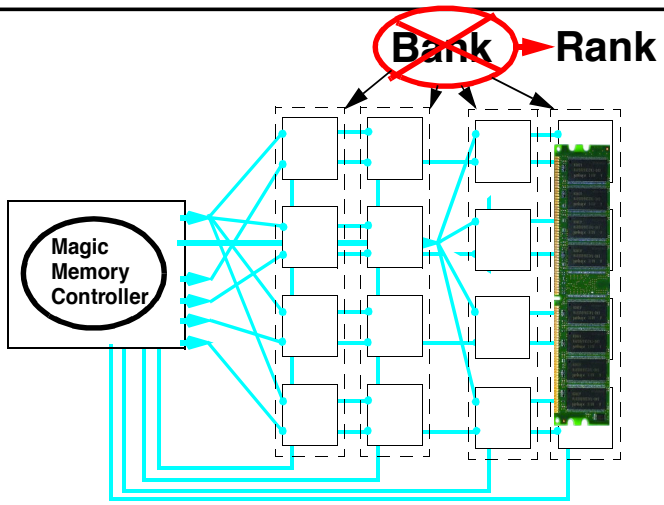
"Mesh Topology"

Source:
Bruce Jacob &
David Wang

Breaking down a DIMM

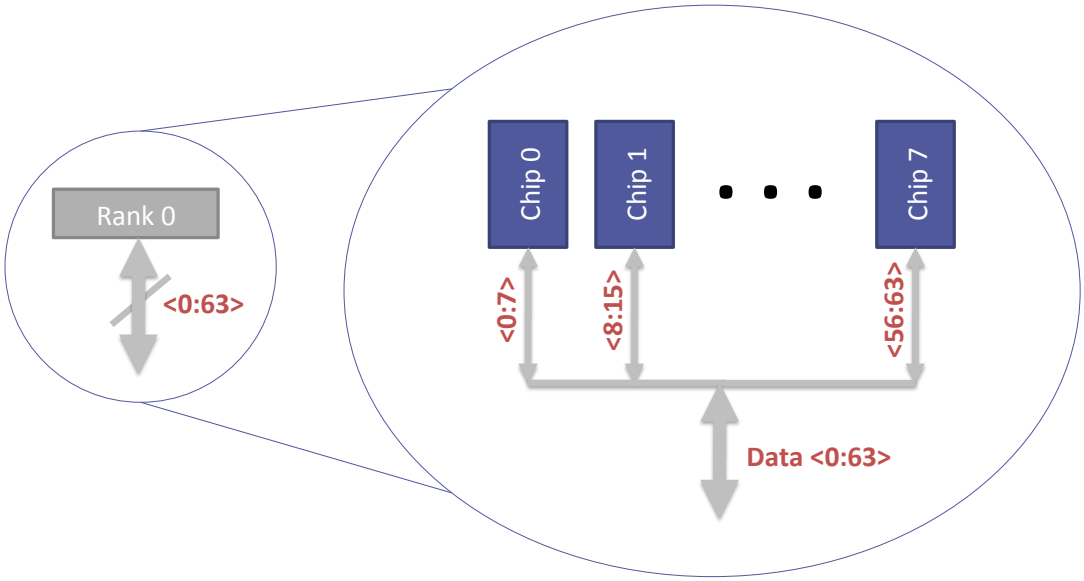


Rank

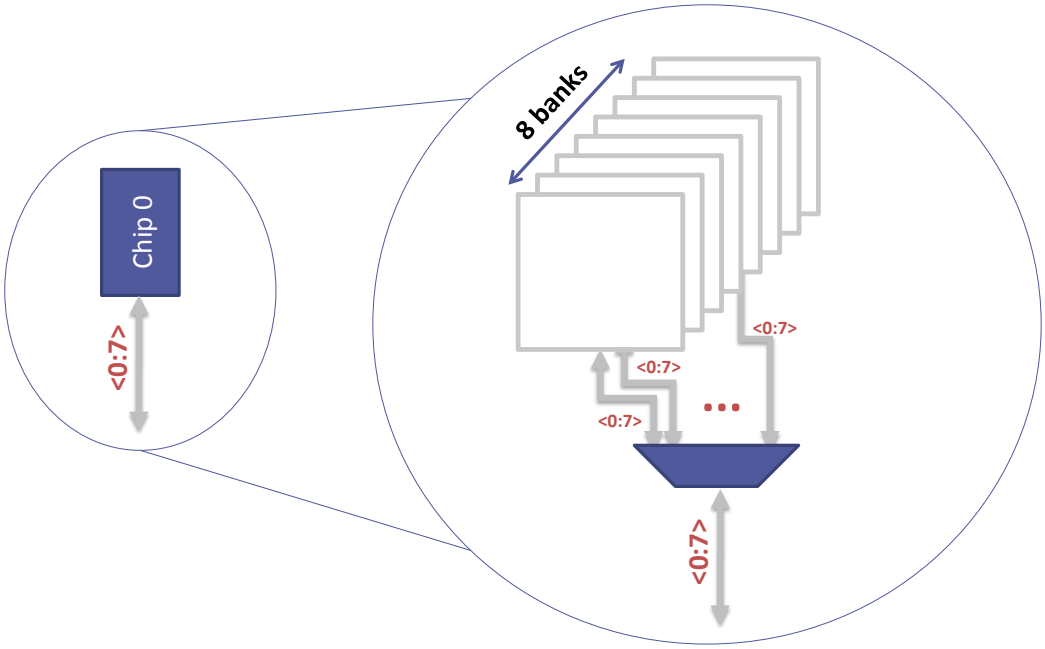


- A set of chips that shares command and control signals
- Data pins for each DRAM are separate in a rank
  - Data in different DRAM chips of the same rank: accessed in parallel

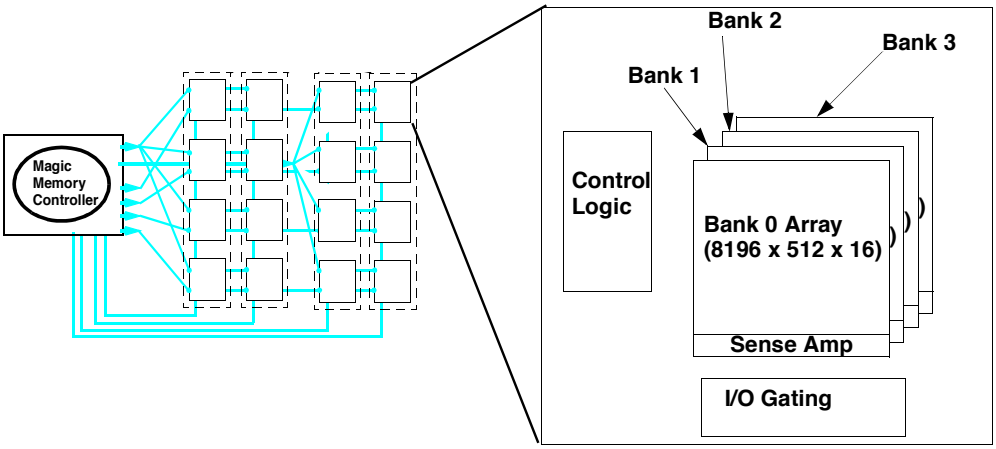
Breaking down a Rank: DRAM chips



Breaking down a Chip: Banks

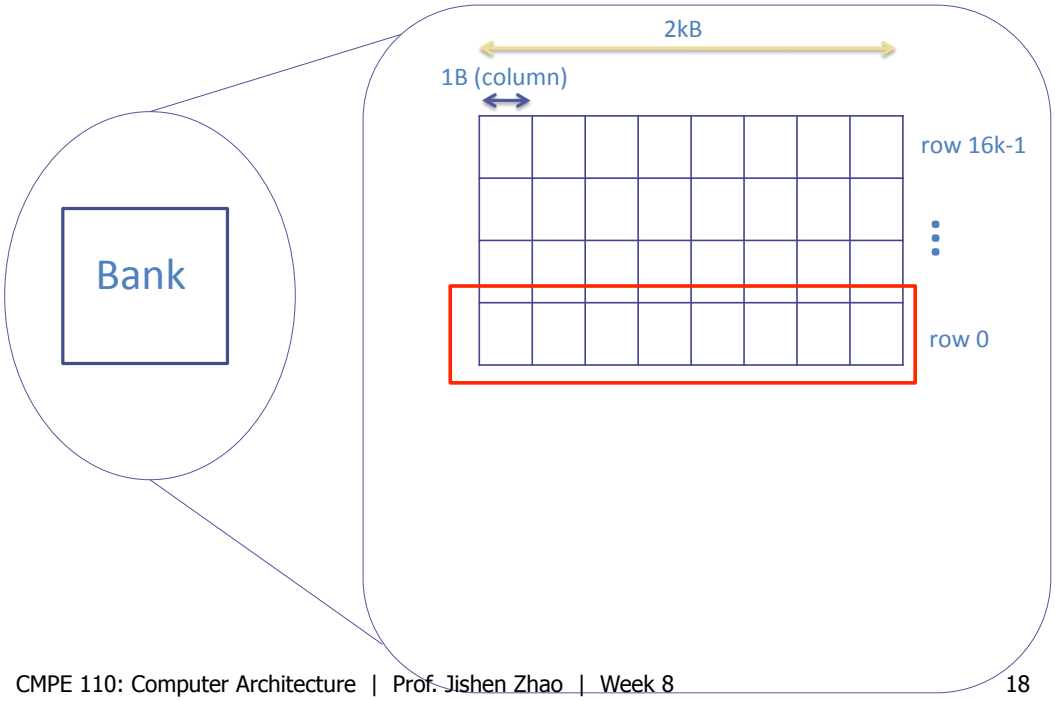


Bank



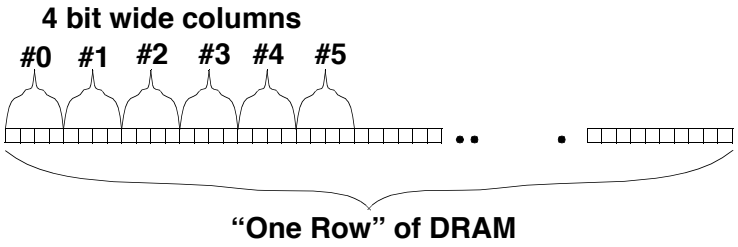
- Banks: independent memory arrays inside of a DRAM Chip
- DDRx DRAM system: 4~16 banks

Breaking down a Bank: Rows and columns

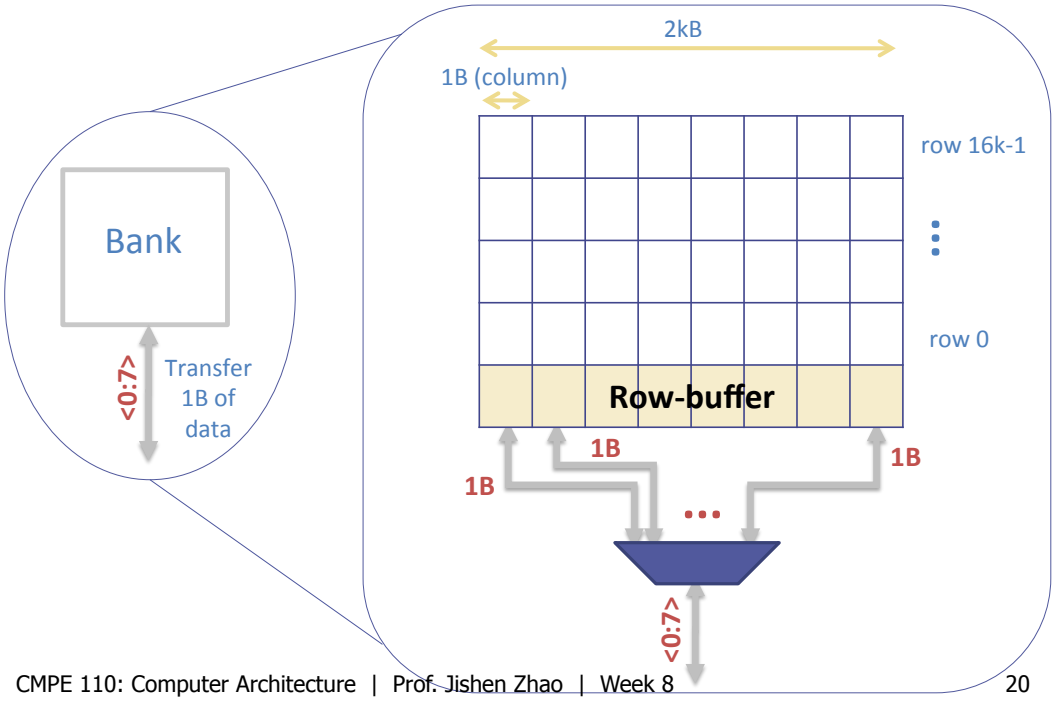


Rows and columns

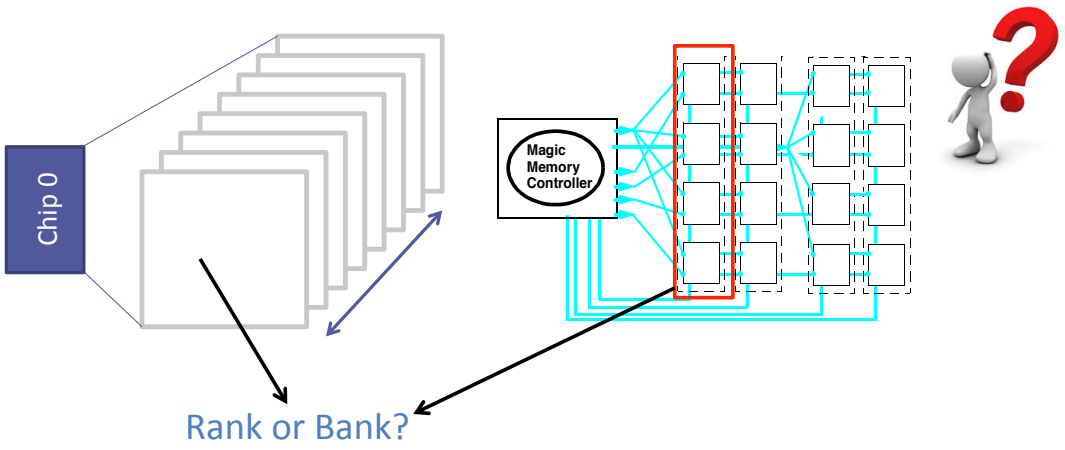
- Column: Smallest addressable quantity of DRAM chip
  - 1 column: Column size == DRAM chip data bus width (4, 8,16, 32)
    - (i.e., 4 bits, 8 bits, 16 bits, 32 bits)
  - Get "n" columns per access. n = (1, 2, 4, 8)
    - "Burst length" == n



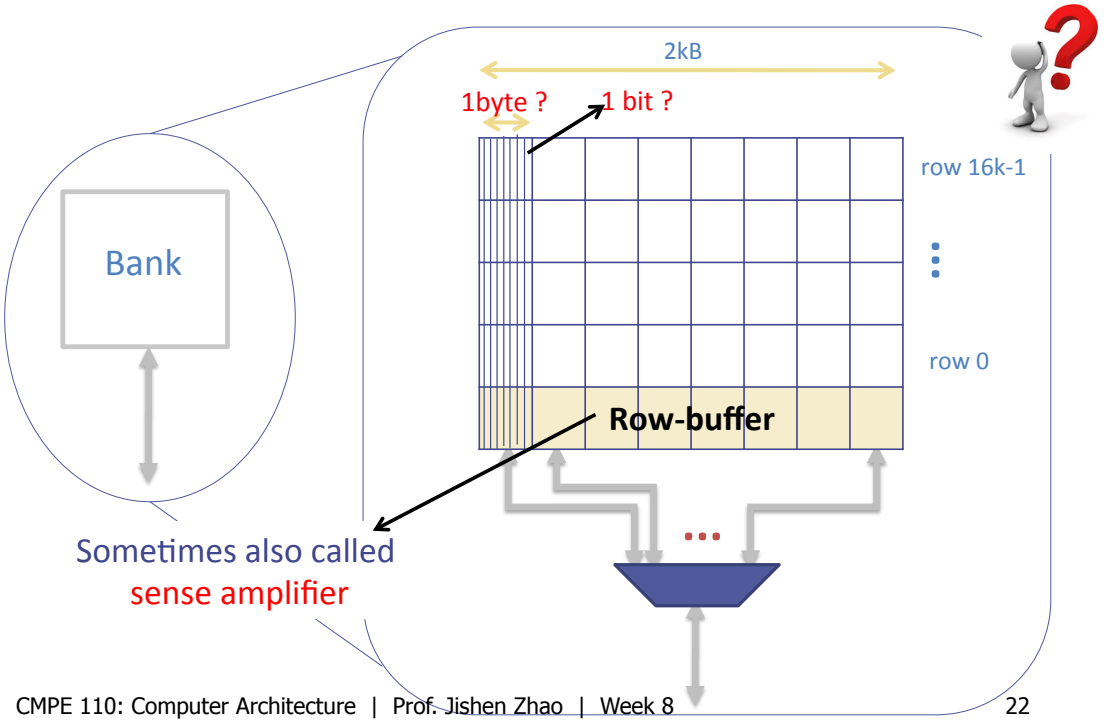
Rows and columns



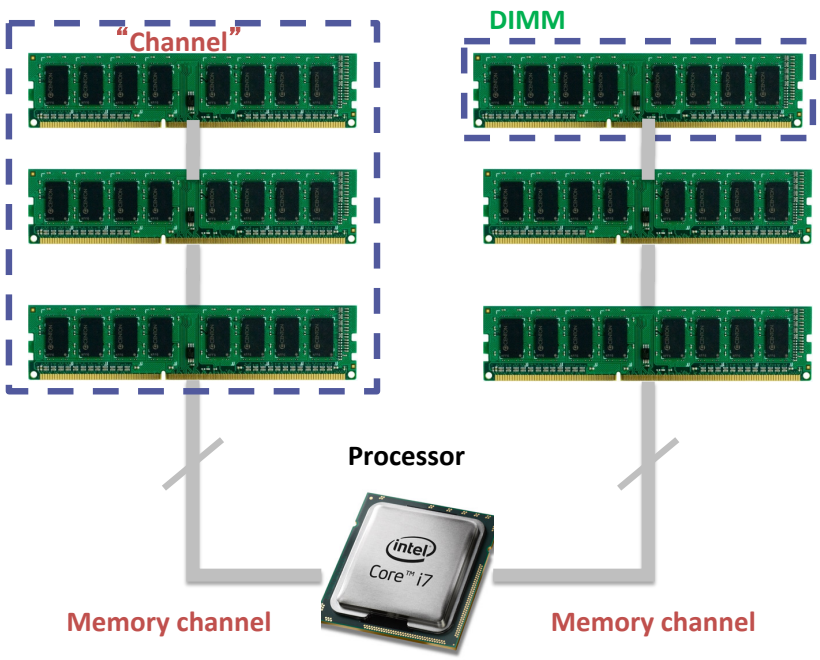
Test yourself: Rank and bank



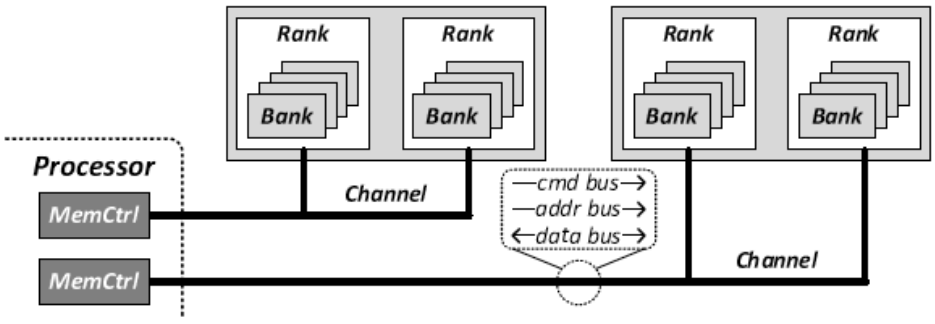
Test yourself: row and column



Channel

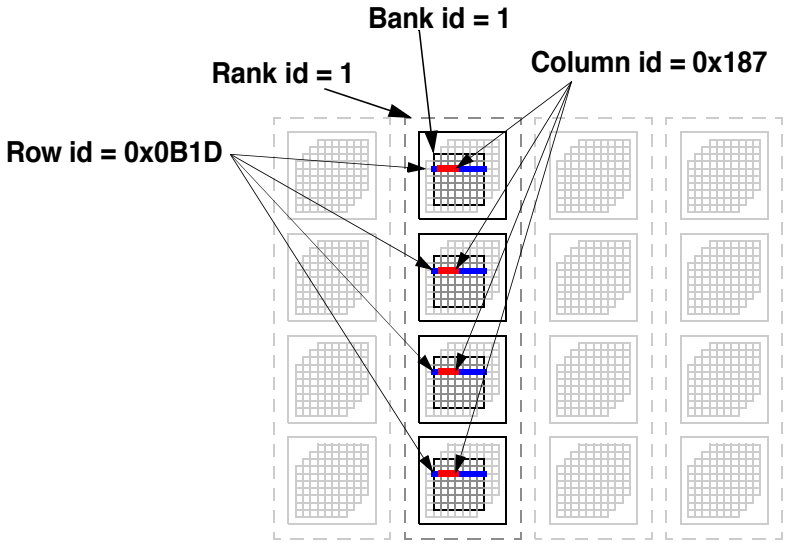


Generalized Memory Structure



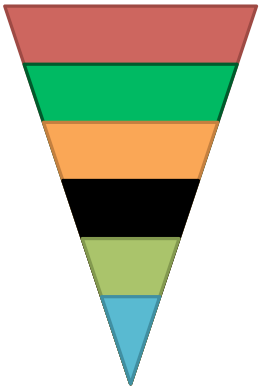
Source: Kim+, "A Case for Exploiting Subarray-Level Parallelism in DRAM," ISCA 2012.

Where's the data (later)?



What we learned: main memory

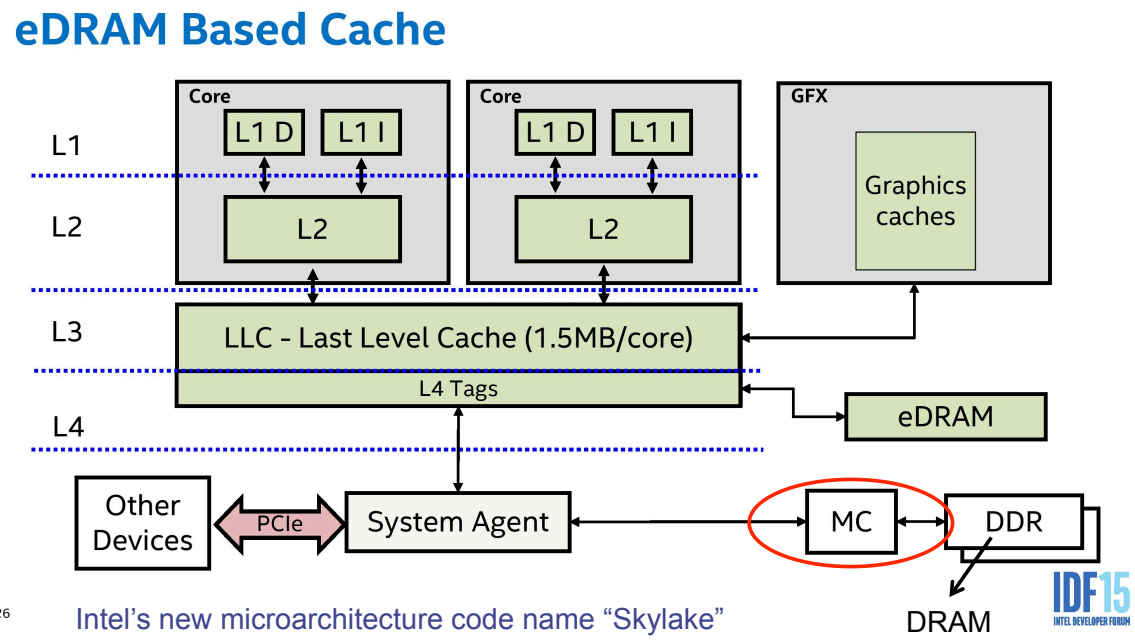
- Channel
- DIMM
- Rank
- Chip
- Bank
- Row/Column



Memory controller (not required)

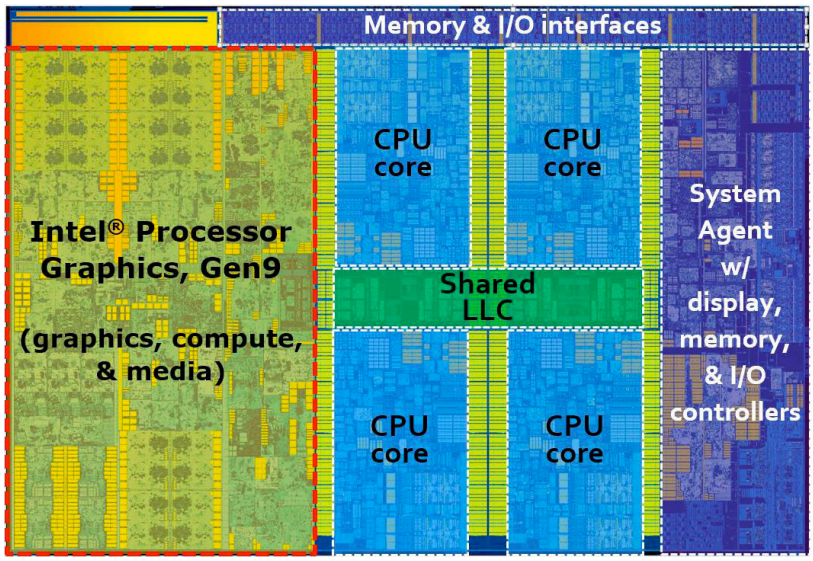
- Purpose and functions
  - Ensure correct operation of DRAM (e.g., refresh)
- Service DRAM requests while obeying timing constraints of DRAM chips
  - Constraints: resource conflicts (bank, bus, channel), minimum write-to-read delays
  - Translate requests to DRAM command sequences (e.g., activate, read, write, precharge)
- Buffer and schedule requests to improve performance
  - Reordering read/write requests and row-buffer management
- Manage power consumption and thermals in DRAM
  - Turn on/off DRAM chips, manage power modes

Memory controller (MC)



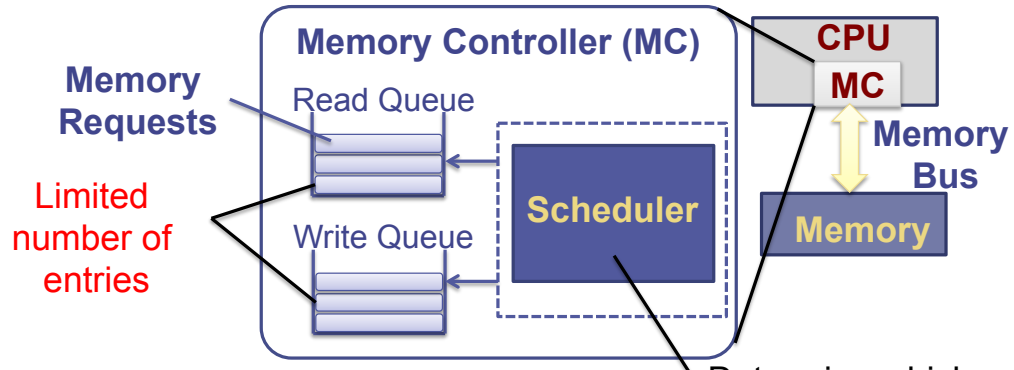
26 Intel's new microarchitecture code name "Skylake"

MC in the latest processors



Intel Core i7 (Skylake, 2015, 14nm)

Memory controller (not required)



If one of queues overflows, need to **drain** it by stalling current memory service

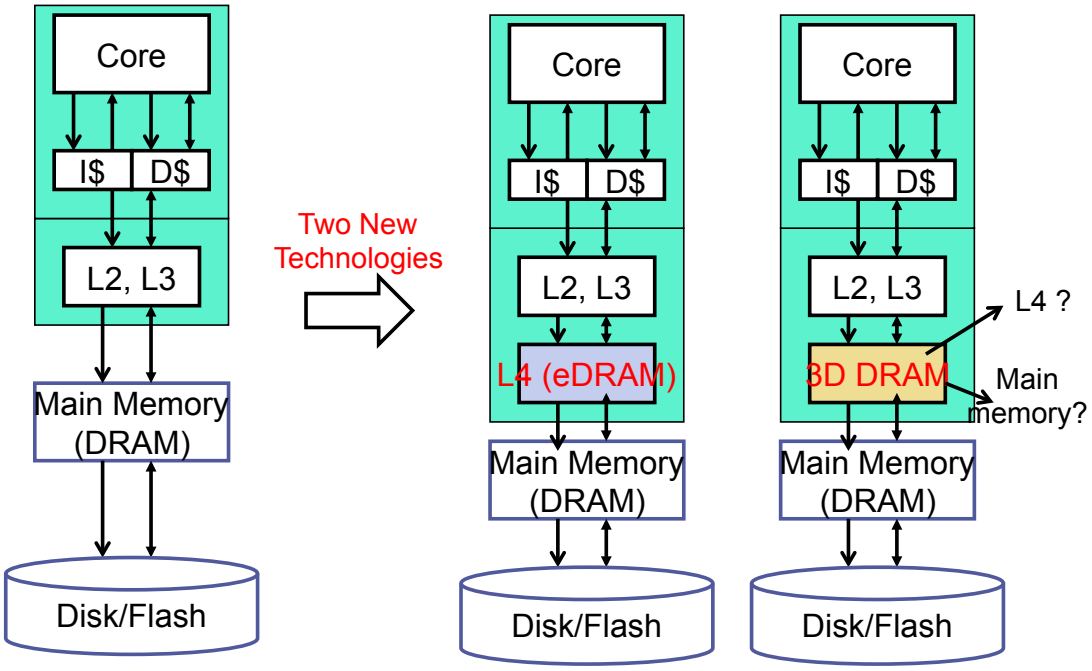
- Interrupting the on-going memory service
  - Bus turnaround overhead

Determine which requests can be sent on the memory bus to be serviced

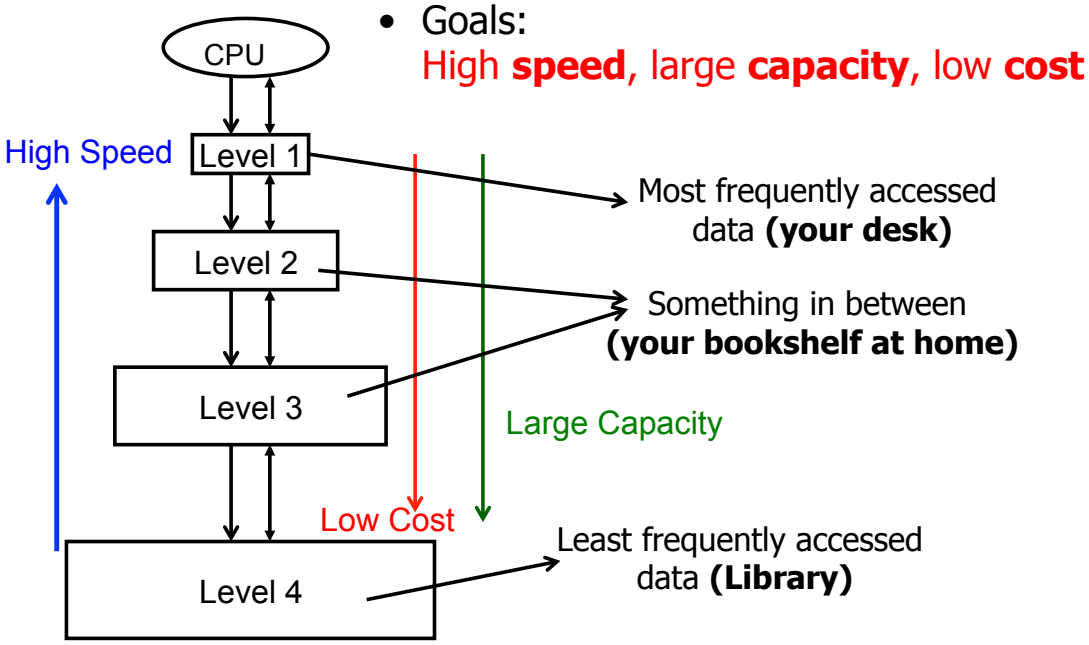
Another set of new memory technologies

On-chip Memory

Integrate DRAM on the processor chip



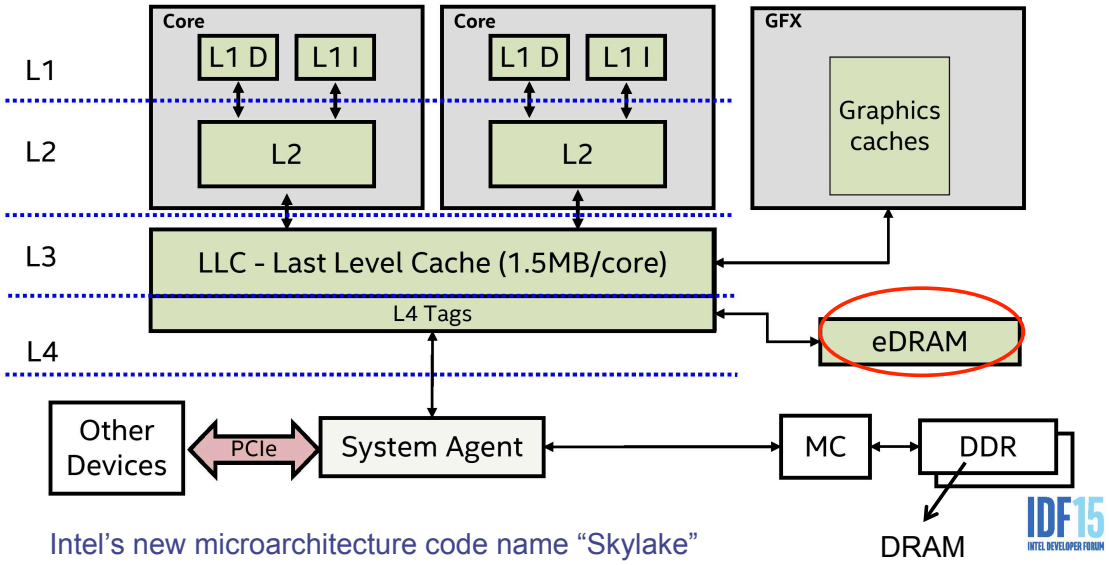
Recall: Basic idea of memory hierarchy!



- Goals:
  High **speed**, large **capacity**, low **cost**

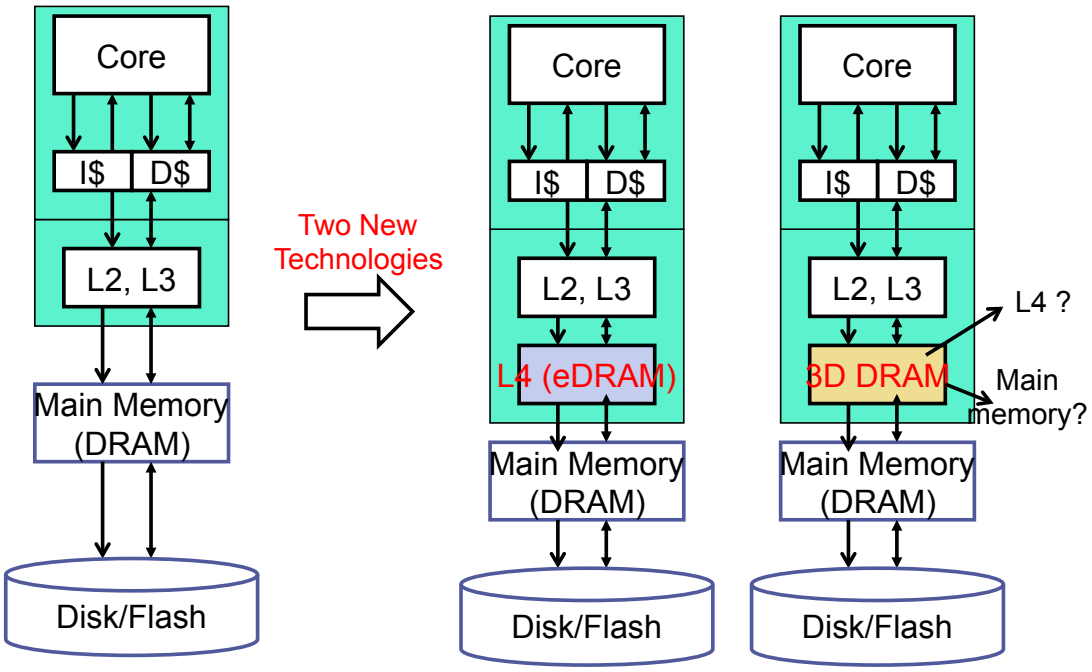
eDRAM (embedded DRAM)

eDRAM Based Cache



26 Intel's new microarchitecture code name "Skylake"

Integrate DRAM on the processor chip



From 2D to 3D architectures