Processors and Memory

System Processors

CPU Architecture

Memory Systems

Memory and Caching

Main memory

- DDR3 SDRAM – (2007) faster than DDR3 and more power efficient, speeds from 800 -2133 MTS

- DDR4 SDRAM – from late 2014, Lower voltage (1.2 vs 1.5), speeds up to 4000 MTS

- DDR5 SDRAM

What is all this useful for?

- Matching CPU speed to memory speed

- Program optimization

RAM speed and PCnnnn ratings

- Sometimes memory is described as PCnnnn, eg
  - PC3200
  - PC19200

- This is the maximum speed in Mbytes/sec that data can be transferred between the CPU and the memory.

- However, it will only reach this maximum with the correct CPU clock speed

Examples

- A CPU with a Front Side Bus running at 400 MHz can transfer data at a maximum rate of 400 x 106 x 8 = 3200Mbs
- A CPU with a Front Side Bus running at 2400 MHz can transfer data at a maximum rate of 400 x 106 x 8 = 19200Mb

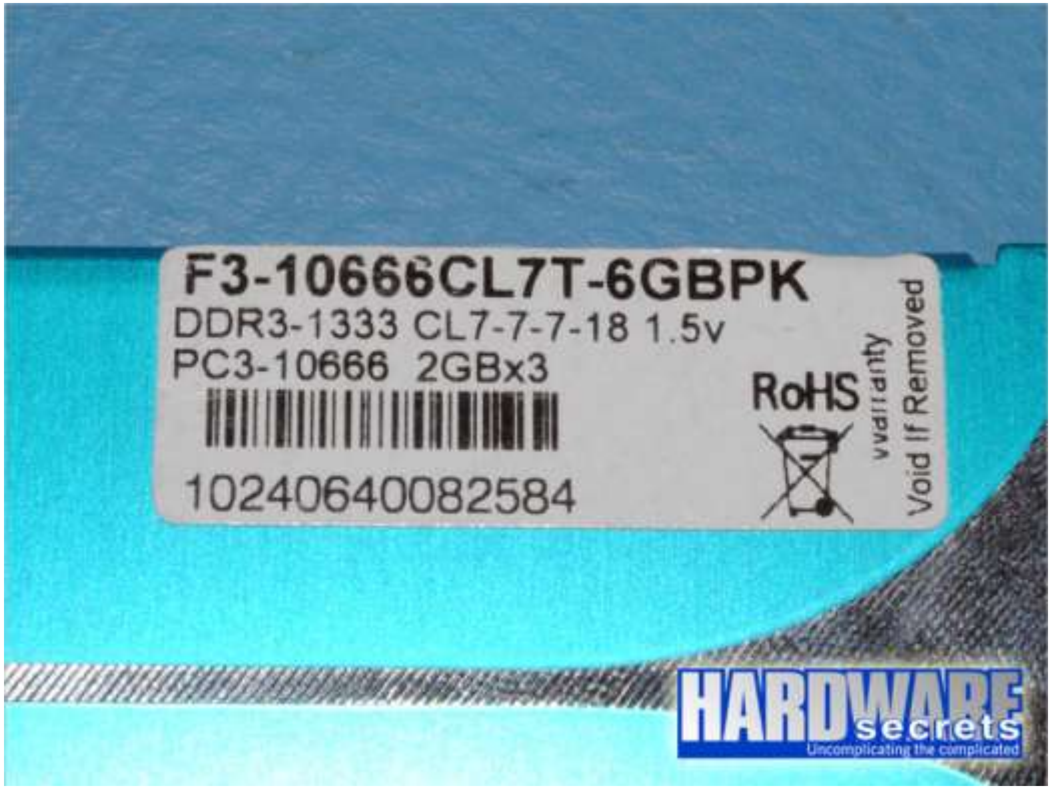
- Put another way, PC3200 memory will work correctly at FSB speeds of up to 400 MHz, and PC19200 will work correctly for FSB speeds up to 2400 MHz
- PC19200 memory would work with a 400MHz FSB, but only running at 3200MHz (excluding other compatibility issues)
- The memory will generally run at 8 times the speed of the FSB, providing it doesn’t exceed its maximum speed

Memory Latency

- Maximum Burst Speed is the rate data can be transferred to the CPU, but doesn't take account of latency.

- DDR1 SDRAM running at 100MHz (cycle time 1/108 = 10nS), had typical timing of 2-3-3, meaning the CAS delay was 20nS.

- What about this one



Other bus structures

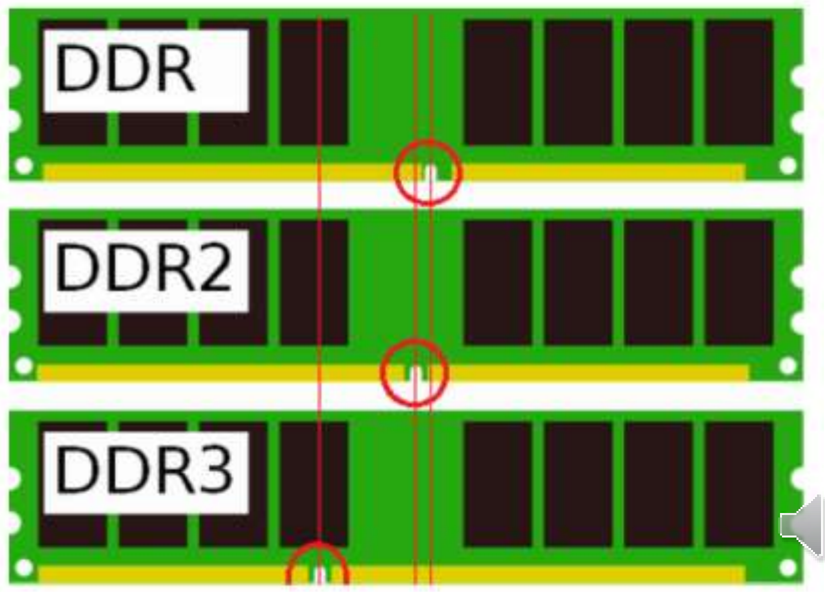
- Some newer CPUs have dual data buses which potentially double the maximum data transfer rate

Memory Chips

- Memory chips are arranged on circuit boards in DIMM (Dual In-Line Memory Module) format
  - edge card connectors on both sides of board

- each SDRAM type has different electrical properties are are incompatible

- keyed' so that they will only fit in the socket corresponding to that specific memory technology.



Memory Metrics

Memory metrics:

- Bandwidth
  - capacity: how much data can be transferred in a given time
  - peak bandwidth is theoretical efficiency

- Latency
  - delay: time from request for data to when data is available
  - chip memory organized as 2D matrix – takes time to activate a row, select a column, switch between rows and columns, etc.
-

Main memory

- Example peak bandwidth and speeds:

Technology	Speed Rating	Peak Bandwidth
SDRAM	PC133	1.06 GB/sec
DDR SDRAM	DDR-400	3.2 GB/sec
DDR2 SDRAM	DDR2-800	6.4 GB/sec
DDR3 SDRAM	DDR3-1600	12.8 GB/sec
DDR4 SDRAM	DDR4-2400	24.8 GB/sec

Non-volatile memory

- Computers require some information to remain in memory: the BIOS (Basic Input/output System)
  - enables computer to access components of hardware
  - needs to be in memory when computer boots
  - stored in non-volatile memory

- Flash EEPROM
  - Flash Electrically Erasable Programmable Read Only Memory

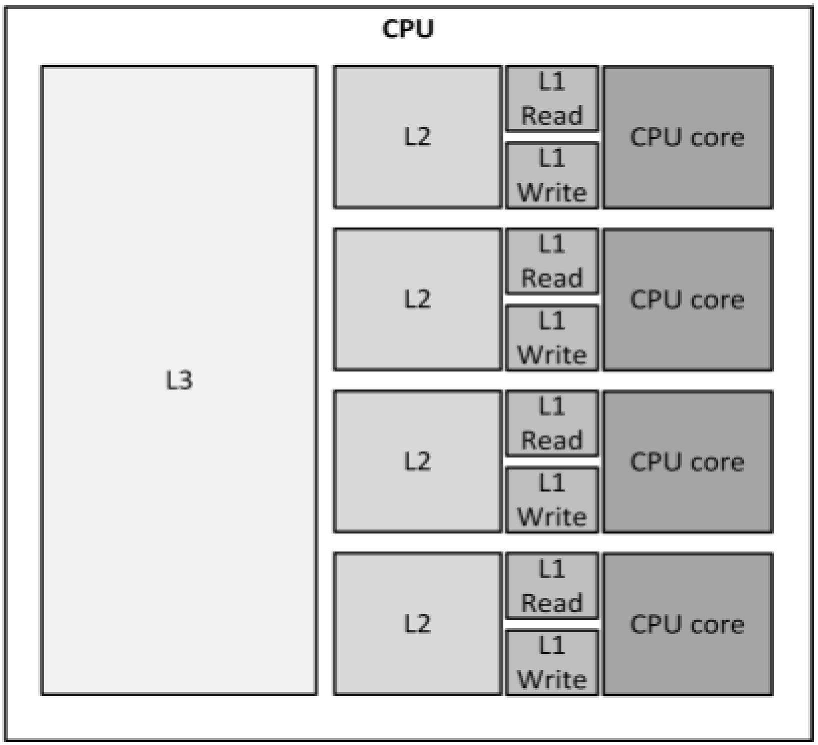
Cache

- Main memory is slow
- If CPU has to wait for a memory access, many CPU cycles are wasted

- Cache
  - static memory between CPU and main memory
  - holds recently accessed data
  - holds data predicted to be needed soon (prediction might not be accurate)
  - multiple levels of cache

CPU Cache

- *Level 1 cache*
  - small (e.g. 64 KB)
  - on CPU for ultra fast access
  - runs at or near CPU speeds so little CPU delay to access
  - Write-through is possible
- *Level 2 cache*
  - larger than level 1 (e.g. 256 KB)
  - slower than level 1 cache
  - on CPU
- *Level 3 cache*
  - larger than level 2 (e.g. 8 MB)
  - slower than level 2 cache
  - either on CPU or sometimes on a support chip



CPU Cache

- Cache is tiny compared to main memory
  - yet 90% hit rates are not uncommon on modern systems
- Principle of Locality
  - Locality of execution
  - Locality of data access

Memory Type	Size	Proportion of Total
Main Memory	8 GB	~99.9%
Level 3 Cache	8 MB	~0.1%
Level 2 Cache	256 KB	~0.00003%
Level 1 Cache	64 KB	~0.000008%

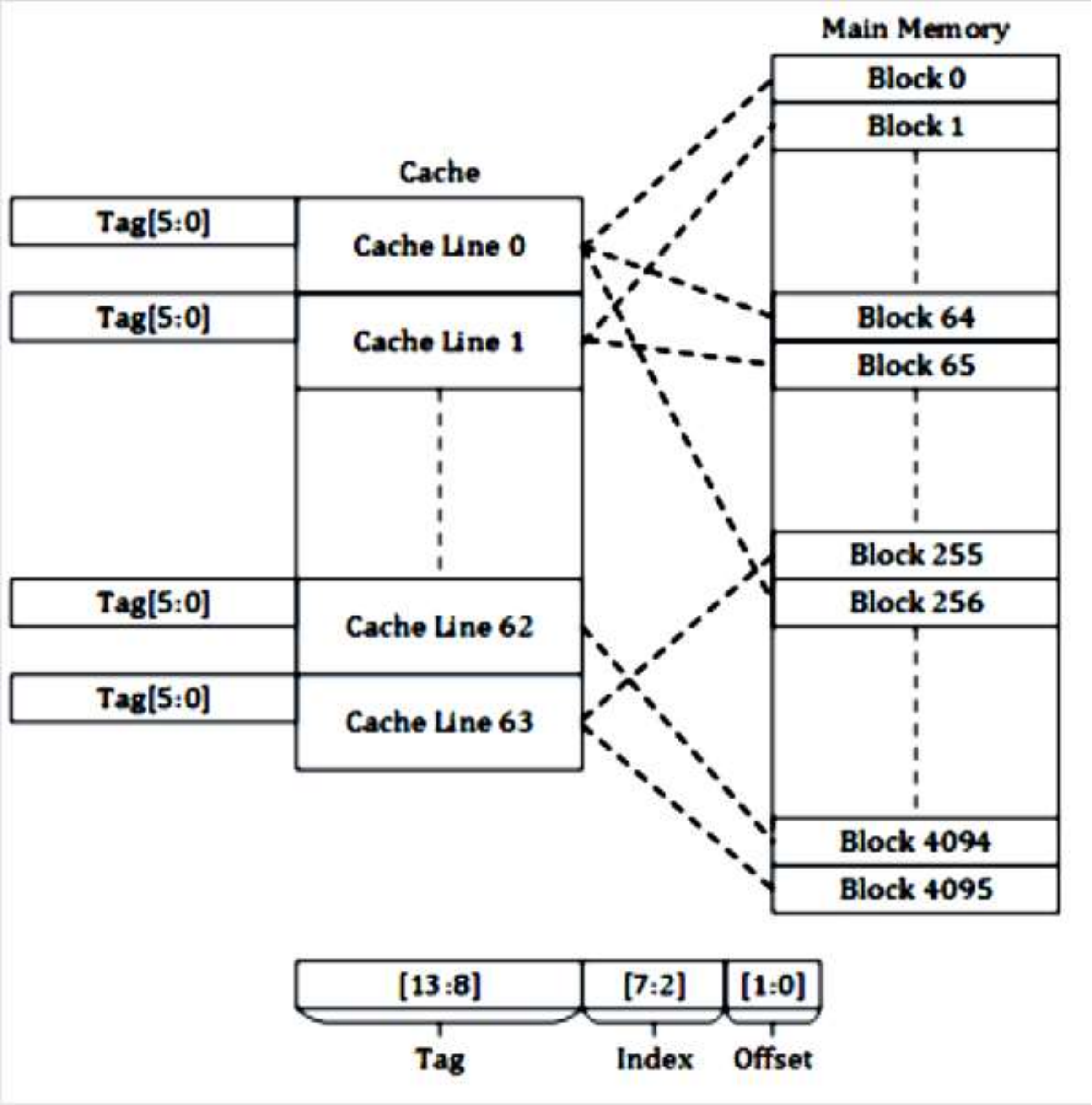
Cache for different purposes

- Types of cache

– Instruction Cache	– locality of execution
– Data Cache	– locality of access
– Translation look aside cache	– locality of indexed access

- The fastest type of cache is content addressable
  - This is usually in the form of tables containing
    - The starting address of a notional block of memory
    - The block contents itself
  - These are mapped into a virtual memory address space
    - In operating systems they are called pages
    - In CPUs they are the L1,2,3 caches
    - In networking, they are called proxy caches

Memory Cache



Memory Cache

- When a memory access occurs the processor first checks the cache to see if cache contains the memory address

- If yes, it is loaded directly (Cache hit)
- If no, the data must be requested from memory, and is loaded into the cache
  - This also means a new space has to be found in the cache
  - How this is done is know a the Cache replacement policy
  - Eg, least recently used (study more in OS course)

Cache Write Policies

- If data is written to the cache, it needs to be written back to main memory at some point, unless it is not changed
- When Cache memory address is written to, a status bit is set, marking the location as dirty, meaning it must be written back
- Simplest approach is write through, where every cache write is written back to main memory, however this can be slow
- Alternatively, cache can be write back, whereby data is periodically written back to memory, or perhaps queued and written back asynchronously when the memory bus is free

Cache Communication

- It is common to have dedicated L1 and L2 cache for each core in a multicore CPU

- If one core loads a memory location that is currently held in the cache of another core, communication needs to occur to ensure that values are maintained correctly

Summary

- Memory
  - Technologies
  - Timing
  - Evolution of SDRAM
- Main Memory
- Memory Cache