Solutions to Week 6 Memory

Question 1

What is the difference between memory bandwidth, and memory latency? Can you think of any kinds of workloads with which one quality may be more important than the other?

Answer:
Memory **bandwidth** is a measure of how much data can be transferred in a given time. The faster that memory can deliver data (the more data can be transferred in a given time), the faster the computer will perform.
Memory **latency** refers to how long it takes from a request for data being issued, and that data being retrieved from memory. Memory latency is the response time for any type of memory.

Workload Examples:

Desktop versus Server Workloads

Consider the workload of a desktop where memory latency does not have a significant effect contrasted to the workload of server systems where memory latency is more critical.
Streaming Videos

Memory bandwidth would be supremely important in cases where very large volumes of sequential data is important — like streaming video. Being able to move very large volumes of data is important. Because it's sequential the cache algorithms should be easily able to make sure that "all" data the CPU requires is available in the cache, in which case latency would be minimal and of little consequence.
Non-Sequential Data

Memory latency becomes crucially important when dealing with non-sequential data, because then the volume of data is not so important. Being able to get data in a non-predictable sequence means that latency is very important because the hit rate on the cache would be severely degraded. Complex mathematical calculations (with “unpredictability”) would be an example.

Question 2

Compare and contrast and find a hardware example of each:

1. PROM
2. EPROM
3. EEPROM

Answer: They All are ROM- Read Only Memory.

ROM — Once data has been written onto a ROM chip, it cannot be removed and can only be read. Unlike main memory (RAM), ROM retains its contents even when the computer is turned off, e.g, a small core program called the BIOS is stored on the motherboard ROM ROM is referred to as being nonvolatile.

PROM — Programmable ROM is a computer memory chip that can be programmed once after it has been created. Once the PROM has been programmed, the information written is permanent and cannot be erased or deleted. PROM was first developed by Wen Tsing Chow in 1956 and a good example of a PROM is a computer BIOS in early computers. Today, PROM in computers has been replaced by EEPROM.

EPROM - EPROM is Erasable Programmable ROM where ultra-violet light is used to erase memory. Classical example of EPROMs in micro-controllers include some versions of the Intel 8048, the Freescale 68HC11, and the "C" versions of the PIC micro-controller.

EEPROM (Electrically Erasable Programmable ROM) — EEPROM is a type of non-volatile memory used in computers and other electronic devices to store small amounts of data that must be saved when power is removed. Unlike bytes in most other kinds of non-volatile memory, individual bytes in a traditional EEPROM can be independently read, erased, and rewritten. Note that there is now Flash EEPROM and in modern systems BIOS is flash EEPROM — also used in modern music players, smartphones, and tablets.

Question 3

Describe what cache is, and explain how it is used by the system to manage data.

Answer:

A CPU cache is a cache used by the CPU of a computer to reduce the average time to access data from the main memory. The cache is a smaller, faster memory which stores copies of the data from frequently used main memory locations.

Cache Levels	Size	Access Time	Location	Proportion of Total
L1	Smallest (2KB — 64KB)	Ultra fast access	A static memory integrated with processor core.	~ 0.000008%
L2	Slightly larger than L1 (256KB — 2MB)	Slower than L1	On the motherboard on earlier computers, it is now found on the processor core - integrated or on-die cache.	~ 0.00003%
L3	Larger than L2 (1MB — 32MB)	Slower than L2	Motherboard-based caches.	~ 0.1%
Main Memory	Gigabytes in size.			~ 99.9%

- Static memory between CPU and main memory
- Holds recently accessed data
- Holds data predicted to be needed soon (prediction might not be accurate)

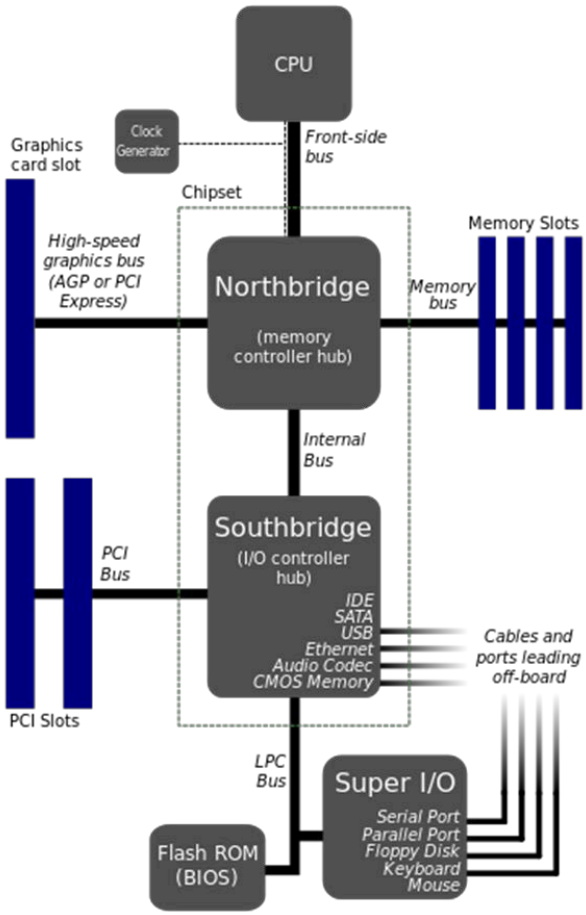
- Goal is to have effective memory access time be close to the access time of the fastest memory
- Multiple levels of cache

Multi-level caches generally operate by checking the fastest, level 1 (L1) cache first, when a miss occurs in L1, L2 is examined, and only if a miss occurs there is main memory referenced.

Question 4

Why is it preferable to balance the peak bandwidths of the memory technology and the CPU's front side bus FSB (or connection to the memory)?

Answer:



In memory technology, the peak bandwidth is the theoretical efficiency.
The FSB connects the computer's processor to the system memory (RAM) and other components on the motherboard. These components include the system chipset, AGP card, PCI devices, and other peripherals.
Most computers' processors run faster than their system buses, so the FSB speed is typically a ratio of the processor speed.
For example, a Pentium 4 processor that runs at 2.4 GHz may have an FSB speed of only 400 MHz. The CPU to FSB ratio would be 6:1. A Power Mac G5, however, with a 2.0 GHz processor, has a 1.0 GHz FSB. Therefore, its CPU to FSB ratio is 2:1.
The smaller the ratio, the more efficiently the processor can work. Therefore, faster FSB speeds lead to faster overall performance.
If the bandwidth for the memory is slower than the FSB, it will bottleneck the performance. (The peak bandwidth can only be as fast as the slowest component.) If the RAM is slower, the CPU will be "waiting around" for the memory to catch up.
The system can only transfer data to/from memory as fast as the infrastructure (FSB) allows. Installing memory that is faster than the FSB will just make your expensive, fast memory run slower.

Question 5

In the above figure, you can see a PC3-10666 memory module, which uses DDR3-1333 memory chips.

1. What is the maximum theoretical transfer rate? What is the real clock rate?
2. What are the causes of latency in DDR type RAM?

Answer: They are typically of the form: **DDRx-yyyy CLa-b-c-d**
The first figure is the memory speed. **DDR3-1333** refers to DDR3 memory that is running at an effective clock speed of 1333 MHz (1333 million cycles per second). 1333 indicates the maximum clock speed that the memory chips support.
It is important to note that this is not the real clock rate of the memory.
The real clock rate of the DDR, DDR2, and DDR3 memories is **half** of the labeled transfer rate, since double data rate (DDR) type RAM operates using two transfers per clock cycle. Therefore DDR400 memories work at 200 MHz, DDR2-800 memories work at 400 MHz, and DDR3-1333 memories work at 666 MHz.
Read more at http://www.hardwaresecrets.com/understanding-ram-timings/

NOTE latency measured in clock cycles instead of real time.

Similar to an Excel spreadsheet, memory is also organised into a grid of rows and columns. To activate a row of RAM, we have to send the memory controller the address of the row we're interested in, and similarly, to activate a column, we have to send it the address of the column we're interested in. The memory module in the above figure has 7-7-7-18 timing.
Four more data points follow:

- The first figure, often prefixed with CL, stands for "CAS latency". This is the time it takes, in clock ticks, for the memory module to start finding the data requested of it. In this instance, it will take 7 clock ticks.
- The second figure is the "RAS to CAS delay". This is the time taken between accessing the row of the memory matrix, and the column. For this memory module, it takes 7 clock ticks.
- The third figure is the "RAS precharge". This is how long it takes to go from one row in the memory matrix, to another. This memory module takes 7 clock ticks to perform this function.

- The fourth figure is "Active to precharge delay". This is the time the memory controller has to wait from one memory access instruction to another. This memory module takes 18 clock ticks to perform this function.

These timings or delays occur in a particular order:

- When a Row of memory is activated to be read by the memory controller, there is a delay before the data on that Row is ready to be accessed, this is known as tRCD (RAS to CAS, or Row Address Strobe to Column Access Strobe delay).
- Once the contents of the row have been activated, a read command is sent, again by the memory controller, and the delay before it starts actually reading is the CAS (Column Access Strobe) latency.
- When reading is complete, the Row of data must be de-activated, which requires another delay, known as tRP (RAS Precharge), before another Row can be activated.
- The final value is tRAS, which occurs whenever the controller has to address different rows in a RAM chip. Once a row is activated, it cannot be de-activated until the delay of tRAS is over.

Question 6

At a basic level, latency refers to the time delay between when a command is entered and executed. **Latency is specified in clock cycles instead of real time.**

With this in mind, there are two variables that determine a module's latency:

- The total number of clock cycles the data must go through (measured in CAS Latency or CL)
- The duration of each clock cycle (measured in nanoseconds)

Combining these two variables gives us the latency equation:

true latency (ns) = clock cycle time (ns) × number of clock cycles (CL)

The **clock cycle time** is the **inverse** of clock speed (**keep in mind that you need to use the real clock rate.** (The real clock rate of the DDR, DDR2, and DDR3 memories is **half** of the labeled transfer rate, since double data rate type RAM operates using two transfers per clock cycle.)

Compare the true latency of the following memory modules, and discuss which is more important: speed or latency?

- DDR3-1333 CL9
- DDR4-1866 CL13
- DDR4-2133 CL15
- DDR4-2400 CL17
- DDR4-2666 CL18

Answer:

Technology	Module Speed (MHz)	Real Clock Rate	Clock Cycle time (ns)	CAS Latency (CL)	True Latency (ns)
DDR3	1333	666.66	1.50	9	13.5
DDR4	1866	933	1.07	13	13.93
DDR4	2133	1066	0.94	15	14.06
DDR4	2400	1200	0.83	17	14.17
DDR4	2666	1333	0.75	18	13.50

(Note: the clock cycle time of a DDR3-1333 memory running at 1333 MHz (666.66 MHz clock) would be 1/666.66 = 1.50 ns.)

Based on in-depth engineering analysis and extensive testing in the Crucial Performance Lab, the answer to this classic question is **speed**. In general, as speeds have increased, true latencies have remained approximately the same, meaning faster speeds enable you to achieve a higher level of performance.

True latencies haven't necessarily increased, just CAS latencies. And CL ratings are an inaccurate, and often misleading, indicator of true latency (and memory) performance.

Optimise your system by installing as much memory as possible, using the latest memory technology, and choosing modules with as much speed as is cost-effective and/or relevant for the applications you're using.

Question 7

What are the potential issues with trying to run memory faster than rated (either in terms of bus speed, or latency timings)?
Can these issues also occur when trying to run other core components faster than originally intended?

Answer:
Overclocking is the process of forcing a computer or component to operate faster than the manufactured clock frequency. Computer components that may be overclocked include processors (CPU), video cards, motherboard chipsets, and RAM.
A processor's speed = its **multiplier** × the FSB speed (MHz/GHz)
For example, an Intel processor with a multiplier of **16** working with a FSB speed of **200MHz** would run at **3.2GHz**.

There are **two** ways a processor can be made to run faster: **increasing the multiplier** OR **increasing FSB speed**.

- Many modern processors have 'multiplier locks' which prevent users from changing (increasing) the internal multiplier settings partially or completely.

- **Increasing FSB speed tends to be the most common and effective method of overclocking.**

Can Overclocking Damage Computer Hardware?
Yes, but it's typically unlikely.
Generally speaking, when computer hardware is pushed beyond its limits, it will lock up, crash or show other obvious errors long before it gets to the point where the processor or memory might be permanently damaged.
How to Overclock Your Video Card and Boost Your Gaming Performance:
http://lifehacker.com/how-to-overclock-your-video-card-and-boost-your-gaming- 30799346

Question 8

PassMark Software has delved into the thousands of benchmark results that PerformanceTest users have posted to its web site and produced charts to help compare the relative performance of different Memory from major manufacturers such as G- Skill, Corsair, Mushkin, Kingston, Patriot, Crucial and others. Higher quality Memory improves overall system performance for many computing activities such as PC gaming, video editing, software development and normal everyday activity.

Take a look at the "Memory Benchmarked & Graphically Compared" at http://www.memorybenchmark.net/.

Discuss how it can be used to choose the right RAM for your PC.

Answer:

Three main benchmarks in PassMark:
Bandwidth
- write transfer rate
- read transfer rate

Latency

What benchmark is important for the following use
- PC gaming (needs high latency benchmark)
- Video editing (needs high read/ write rate benchmark)
- Software development (needs high read/write rate benchmark)