System Architecture

Todd Davies

February 5, 2015

Introduction

The basic architecture of computer systems has been covered in first year course units which detailed both the instruction set architecture and the micro- architecture (hardware structure) of simple processors. Although these principle underlie the vast majority of modern computers, there are a wide range of both hardware and software techniques which are employed to increase the performance, reliability and flexibility of systems.

Aims

The aims of this course are to introduce the most important system architecture approaches. To give a wider understanding of how real systems operate and, from that understanding, the ability to optimise their use.

The syllabus includes:

- The motivation behind advanced architectural techniques.
- Caching
- The need to overcome latency. Caching as a principle, examples of caching in practice. Processor cache structure and operation.
- Pipelining
- Principles of pipelining. Implementation of a processor pipeline and its properties. Pipelining requirements and limitations. Additional support for pipelining.
- Multi-Threading
- Basic multi-threading principles. Processor support for multi-threading. Simultaneous multi-threading.
- Multi-Core
- Motivation for multi-core. Possible multi-core structures. Cache coherence.
- File System Support
- Implementation of file systems. RAID
- Virtual Machines
- Motivation for Virtual Machines. Language Virtual Machines. System Virtual Machines. Virtual Machine implementation. Binary Translation

Contents

1 Introduction

Performance is always an attribute in high demand in computer systems. Even though processors have become so much more powerful over the last half century, there's still loads of stuff that we cannot do with current technology, such as synthesising HD video in realtime, or computing realistic game physics.

Since 2004/5, companies haven't been able to increase the speed of microprocessors at such a rapid rate due to physical limits, such as power dissipation and device variability. Our devices are still getting faster, but now architecture and the design of systems play a larger role in making stuff run faster. An example of this include making computation more parallel.

2 Caches

Not all technology has improved at the same relative speed. CPU's have become over three orders of magnitude faster over the past thirty years, but memory has increased by only one order of magnitude. This is problematic, since it means that we need to reconcile this gap in order to achieve efficient computation.

Processor caching is used to let the processor do useful computation while it's also waiting on the memory. Modern processors couldn't perform anywhere near how fast they do now without equally modern caching techniques, since the imbalance between the CPU and main memory is so high.

Caches (in general) provide a limited, but very fast local space for the CPU to use. They are used in lots of places all over computer science, including web browsers, mobile phone UI's etc. Likewise, a processor cache is a temporary store for frequently used memory locations.

The principle of locality is what makes caches work for processors, which is that the CPU will only use a small subset of memory over a short period of time. If this subset of memory can be loaded into the cache, then the computation can be sped up significantly.

Every 'cache miss' takes *at least* sixty times longer to execute than a 'cache hit' will (that's assuming there are no page faults etc). Circuit capacitance is the thing that makes electronic devices slow, and larger components have a larger capacitance, henceforth large memories are slow. Dynamic memories (DRAM) store data using capacitance, and are therefore slower than static memories (SRAM) that work using bistable circuits.

Even the wires between the processor and the memory have a significant capacitance. Driving signals between chips needs specialised high power interface circuits. An ideal situation would be to have everything on a single chip, however current manufacturing limitations prevent this; maybe one day we will be able to do this.

2.1 Why are caches expensive?

L1, L2 and (usually) L3 caches are SRAM instead of DRAM (which is what main memory is made from).

SRAM needs six transistors per bit, DRAM needs one.

SRAM is henceforth physically larger, taking up more space on the chip, which is expensive, since real estate costs money.

2.2 L1 Cache

The L1 cache is the first level of caching between the processor and the main memory. The L1 cache is around 32kb, which is very small in comparison to the size of the main memory, but this is driven out of necessity, since the cache needs to be small to be fast. The cache must be

able to hold any arbitrary location in the main memory (since we don't know in advance what the CPU will want), and henceforth requires specialised structures to implement this.

2.3 Types of cache

The cpu will give the cache a full address of a word in memory that it wants to read. The cache will contain a small selection of values from memory that it has locally, but will ask the main memory for values that it does not have. This is called a cache miss and is expensive in comparison to a cache hit.

2.3.1 Fully associative

A **Fully Associative** cache is one where the cache is small (around 32,000 values), but stores both addresses and their corresponding data. The hardware compares the input address with all of the stored addresses (it does this in parallel). If the address is found, then the value is returned with no need to ask the RAM (cache hit), if the value isn't found, then a cache miss occurs, and the request must go to the main memory.

Caches rely on locality in order to function effectively. There are two types of locality; temporal locality, which is the principle that if you use an address once, you may use it again soon (e.g. loops), and spatial locality, where if you use an address once, you are also likely to use addresses nearby (e.g. arrays).

Spatial locality is exploited better by having a bigger data area in the cache (returning say 512 bits for every address instead of just one word)

The cache hit rate is the ratio of cache hits to misses. We need a hit rate of 98% to hide the speed of memory from the CPU. Instruction hit rates are usually better than data hit rates (although they are in the same cache remember). The reason for this is that instructions are accessed in a more regular pattern, incrementing by one word every time, or looping around etc (have higher locality).

When we do a cache miss (read), we should add the missed value to the cache. In order to do this, we need a cache replacement policy to find room in the cache to put the new value:

- LRU - slow, good for hit rates
- Round Robin - not as good, easier to implement
- Random - Easy to implement, works better than expected.

Memory writes are more complicated than reads. If we've already got the value in the cache, then we change the value in the cache. We can use three write strategies for cache writes (on hits):

- Write through (slow)
- Write through + buffer (faster, slow when heavily used)
- Copy back on cache replacement.

On misses:

- We can find a location in the cache, and write to that, then rely on copy back later, or write back straight away.
- we can skip the cache, and write directly to RAM. Subsequent read will add to the cache if necessary (good if you're initialising datastructures with zeroes).

Fastest one is write allocate or copy back. Main memory and the cache aren't coherent, which can be a problem for stuff like multiprocessors, autonomous IO devices etc. This may need cleaning up later.

Each cache line is at least half address and half data, but often, we store more data per address, so will have 64 bytes of data per 32 bit address.

A fully associative cache is ideal, but this is expensive (in terms of silicon and power).

2.3.2 Directly mapped

We can use standard RAM to create a directly mapped cache, which mimics the functionality of an ideal cache. Usually, this uses static RAM, which is more expensive than dynamic RAM, but is faster. The address is divided into two parts,

2.3.3 Set associative

Set associative caches are a compromise. They comprise of a number of directly mapped caches operating in parallel. If one matches, we have a hit and select the appropriate data. This is good because we can have more flexible cache replacement strategies. In a 4 way, we could choose any one of the four caches for example. The hit rate of set associative caches improves with the number of caches, but increasing the number increase the cost.

3 Practical caches

3.1 Cache control bits

When the system is started, the cache is empty. We need a bit for each cache entry to indicate that the data is meaningful (i.e. it isn't just an unitialised zero or something). We also need a dirty bit if we're using the 'write back' caching strategy (see above), rather than the 'write through' strategy.

3.2 Exploiting spatial locality

In order to exploit spatial locality, we need to have a wider 'cache line', where each entry will give you more data than just one word. Each entry tag could correspond to two, four, eight etc words. Spatial locality says that if we get one byte, we'll probably want one from close by too.

The lowest bits are used to select the word in the cache line. Most cache lines are 16 or 32 bytes, which is 4 or 8 32bit words. The data is transferred from RAM in bursts equal to the width of the line size, using specialised memory access modes.

The line size is important, since we want to have a line size of multiple words to exploit spatial locality, but if the line is too big, then parts of it will never be used. The number of cache misses decreases as you increase the cache line size, until one point, where the line size will be too long to use all the words, and then the number of misses will increase.

3.3 Separate instruction and data caches

Since instructions and data have different access patterns in memory (but they are stored in the same memory), we could use different caches for each type of word, so that the different caches can use different strategies to minimise misses according to their different access patterns.

3.4 Multi level caches

As chips get bigger, in theory, we should build bigger caches to perform better. However, big caches are slow, and the L1 cache needs to run at processor speed. We can instead put another cache between the RAM and the L1 cache, and keep the L1 cache the same size.

The L2 cache is typically sixteen times bigger than the L1 cache, but also four times slower. It's still ten times faster than RAM though. The L1D and L1I caches both share the L2 cache.

If a chip has an L3 cache, then it is usually quite large (maybe around 8Mb), but its performance is only about twice as good as that of RAM.

3.5 Cache misses

There are three types of cache misses, called the three C's

Compulsory misses
When we first start the computer, the cache is empty, so until the cache is populated, we're going to have a lot of misses.

Capacity misses
Since the cache is limited in size, we can't contain all of the pages for a program, so some misses will occur.

Conflict misses
In a direct mapped or set associative cache, there is competition between memory locations for places in the cache. If the cache was fully associative, then misses due to this wouldn't occur.

3.6 More cache performance

In order to fill a cache from empty, it takes $\frac{\text{Cache size}}{\text{Line size}}$ memory accesses. If we multiply this by the time it takes for a single memory access (say 10μs), then we can work out how long it will take to fill the cache (assuming each access is to a unique memory address). We can derive how many CPU cycles this takes from this.

3.7 Cache consistency

We need to make sure that the values stored in the CPU cache are consistent with those in main memory. There are situations when they can disagree, for example if IO reads or writes directly to memory (perhaps using DMA), then that value could be different from whatever is in the cache.

Solutions:

Non-cacheable
One solution is to make areas of memory that IO can access non-cachable, or clear the cache before and after the IO takes place.

IO use data cache
Another is to have the IO go directly through the CPU's L1d (data) cache before accessing memory, but this is slow.

Snoop on IO activity
We could have hardware logic that will look at the reads and writes to memory from IO and make sure the cache is consistent with memory for those addresses.

3.8 Virtual Addresses

Since the CPU deals with virtual addresses when accessing memory, and uses a Translation Lookaside Buffer to derive the correct physical address. However, which address does the cache store? Does it sit before the TLB, or after it between the CPU and memory?

If we make addresses go through the TLB before they reach the cache, then this is slow, since they must pass through extra logic etc before hitting the cache. However, if we make the cache store virtual addresses, and have the TLB sit inbetween the cache and memory, this makes snooping hard to implement along with more functional difficulties.

The answer is to have the TLB operate in parallel to the cache. Since address translation only affects the high order bits of the cache (the low order bits are the offset which remains the same). The cache index is selected from the low order offset bits, and only the tag is changed by address translation.

4 Pipelines

5 Multi-Threading

6 Multi-Core

7 Vitalisation

8 Permanent Storage