Lecture 8: Faster Memory

Hossam A. H. Fahmy

Cairo University, Faculty of Engineering

Overview

1. Performance
   - Access time and bandwidth
   - Wide memories
   - Interleaved memories
2. CPU time
   - Solved examples
3. Points to remember

Access time and bandwidth

Let us look at a large system: many processors each with its own cache accessing the memory system.

The access time of the memory is the period from the time the address is given to the memory chip till the data is ready to be sent to the requester.

The bandwidth is the number of bytes (or requests) served per unit time.

The bandwidth is not simply the inverse of the access time. *Why? How can you increase the bandwidth given a fixed access time?*

Higher bandwidth: wide memories

Example 1 In a system, the access time of main memory is 10 clock cycles. The cache block size is 16 bytes. The address is transfered on the bus in one clock cycle and any result from the memory is transfered back in one cycle. What is the bandwidth for a narrow bus and memory (4 bytes) and that of a wide bus and memory (16 bytes)?
Solution: We may assume that the address of the whole block is sent only once and that the memory results are not overlapped with the memory access.

$$\begin{aligned} T_n &= 1 + 4 \times (10 + 1) = 45 \text{ cycles.} \\ BW_n &= 16/45 \approx 0.3556 \text{ bytes/cycle} \\ T_w &= 1 + 10 + 1 = 12 \text{ cycles.} \\ BW_w &= 16/12 \approx 1.3333 \text{ bytes/cycle} \end{aligned}$$

Is it worth the price?

Higher bandwidth: interleaved memories

What about multiple memory banks but a single narrow bus?

Example 2 Now, if we use a simple interleaving scheme where four banks are used and each clock cycle one of them starts to access its data, what is the bandwidth?
Solution: In this case, at time 0 the address is sent then at time 1 the first bank starts and its data is ready at time 11. The second bank starts at time 2 and its data is ready at time 12 and so on.

$$\begin{aligned} T_i &= 1 + 10 + 4 \times 1 = 15 \text{ cycles.} \\ BW_i &= 16/15 \approx 1.0667 \text{ bytes/cycle} \end{aligned}$$

In word interleaved systems, the bank number is

$$\text{(word address)mod(\# banks)}.$$

How much time are we loosing on misses?

- Each instruction accesses the memory for its own fetch.
- It may also access the memory for data.

$$\Rightarrow \text{memory accesses per instruction} \geq 1.$$

In addition to AMAT, we calculate the total CPU time for the program.

$$\begin{aligned}
\text{CPU time} &= (\text{CPU cycles} + \text{Stall cycles}) \times \text{Clock cycle} \\
&= (\text{CPU cycles} + \text{Read Stalls} + \text{Write Stalls}) \times \text{Clock cycle} \\
\text{CPU time} &\approx IC \times (CPI + \frac{Mem\ access}{Inst.} \times \text{miss rate} \times \text{miss penalty}) \\
&\quad \times \text{Clock cycle}
\end{aligned}$$

Simple example

Example 3 A system with CPI of 1 has load/store frequency of 25% with cache miss rate 5% and miss penalty 10 cycles. A suggested change to the cache reduces the miss rate to 2% but increases the size of the clock cycle by 20%. Should you use this change?

Solution: We should calculate the CPU time for both cases to decide

$$\begin{aligned} T_1 &= IC(1 + (1 + \frac{25}{100}) \times 0.05 \times 10) \times cycle \\ T_2 &= IC(1 + (1 + \frac{25}{100}) \times 0.02 \times 10) \times 1.2cycle \\ \frac{T_1}{T_2} &= \frac{1 + 1.25 \times 0.5}{(1 + 1.25 \times 0.2) \times 1.2} = \frac{1.625}{1.5} \end{aligned}$$

Yes, this change is beneficial.

Is 'faster' really that much better?

Example 4 The gcc compiler runs on a system with a miss rate of 5% for instructions and 10% for data. The perfect CPI is 1 cycle, the miss penalty is 12 cycles, and the load/store frequency is 33%. Study the effect of changing to 1) a faster architecture with $CPI = 0.5$ instead of 1 and 2) a faster system with three times the clock frequency.

Solution: First let us calculate the original CPU time

$$\begin{aligned} T_{orig} &= IC(1 + 0.05 \times 12 + \frac{1}{3} \times 0.10 \times 12) \times cycle \\ &= IC(1 + 1) \times cycle \end{aligned}$$

Is 'faster' really that much better?

The changed CPU times are

$$\begin{aligned} T_{CPI} &= IC(0.5 + 0.05 \times 12 + \frac{1}{3} \times 0.10 \times 12) \times cycle \\ &= IC(0.5 + 1) \times cycle \\ T_{freq} &= IC(1 + 0.05 \times 36 + \frac{1}{3} \times 0.10 \times 36) \times \frac{1}{3} cycle \\ &= IC(1 + 3) \times \frac{1}{3} cycle \end{aligned}$$

The smaller CPI gives only $1.5/2 = 3/4$ reduction while the higher frequency gives only $(4/3)/2 = 2/3$.
Remember that marketing is different from engineering!

Final notes

- The hit time is also important.
- Spatial locality leads us to blocks with multiple words.
- A larger block size may increase the miss penalty. We must balance different factors.
- Wide and interleaved memories boost the system performance.
- The total time taken by a program is a very important measure.