

CSSE7013 - Advanced Computer Architecture Tutorial 1

Learning Objectives

This tutorial aims to help you to understand practical application of quantitative techniques. Some examples may be clearer later in the course.

QUESTION 1—Quantitative Techniques

Use the following data for the original machine, a load-store architecture, for all parts of the question where calculations are required. In a load-store machine, all memory data references are either copies from memory to a register (loads), or copies from a register to memory (store). All ALU operations are on 2 source registers and 1 destination register. The timings assume no cache misses, and are averages based on the possibility that the pipeline may stall (for reasons we'll see in later chapters). CPI of less than 1 (as here) is usually given as IPC as explained in the notes.

instruction categories	dynamic frequency %	clock cycles (no stalls)
ALU	43	0.25
loads	21	0.5
stores	12	0.25
branches	24	1

- A modification to the design is proposed whereby 90% of branches can be predicted accurately, resulting in a 50% reduction in branch CPI when a branch is predicted correctly. However, when a branch is predicted incorrectly, CPI for a branch is increased by 50%. The modification increases clock cycle time (slows down clock speed) by 15%.
 - What is the CPI of the original machine with the given instruction mix?
 - What is the CPI of the modified machine with the given instruction mix?
 - Which is faster, and by how much?
 - Now assume the instruction miss rate is 0.8% for the original and 1% for the modified architecture. In both cases, the data miss rate is 1% and a miss costs 200 extra cycles. Redo the CPI calculations for both machines; which is faster—and by how much?
- Recent designs rely on a very fast on-chip interface to minimize the cost of misses from the L1 cache to the L2 cache. A consequence of this design is that the L2 cache size can be limited, as opposed to earlier designs with off-chip L2 caches.
 - Consider a 2-level cache, in which the penalty for references to DRAM is constant at 200 cycles, but the penalty for misses to L2 varies. In the 2 design variations, the following figures apply:
 - fraction of references which miss to L2 in both designs: 1%
 - fraction of references which miss from L2 in small, slower L2 design: 20% (relative)
 - penalty for misses to small, fast L2: 10 cycles
 - fraction of references which miss from L2 in bigger, slower L2 design: 10% (relative)
 - penalty for misses to bigger, slower L2: 20 cycles
 - In the light of (i), comment on the trend towards faster but smaller on-chip caches.

QUESTION 2—Amdahl's Law

- You are travelling to Sydney by car. Halfway there, you realize you have been going too slowly: in fact, you are taking twice as long as you should. Explain what if anything you can do to get to Sydney on time. Think about how Amdahl's Law applies.
- It is proposed that a new fast train be introduced, linking Brisbane and Gold Coast, over a distance of 50km, with 2 stops on the way. Assume the train averages 150km/h while moving, but needs 10min at each stop to load and unload. What is the speedup versus a train which averages 75km/h but which does not stop on the way?

QUESTION 3—Performance Measurement

- To evaluate new computer architectures, you can use back of the envelope calculations (e.g., question 1), running real-world applications on a simulation of a new design or running traces of memory references from applications from an existing machine through a memory simulator. Trace file entries are *type_of_operation memory_address* where type of operation is usually one of *fetch*, *read* or *write*. Explain why any of the performance measures do or don't apply in these situations:
 - You are experimenting with variations on the pipeline timing but otherwise the design is very similar.
 - The pipeline design is just like an existing machine, but you are experimenting with variations in DRAM speed.
 - You are planning significant additions to the instruction set with 2 possibilities:
 - the instructions can be generated by an experimental compiler
 - the new instructions are too different from any existing ISA to create an experimental compiler easily
- Another way to measure where time is being spent in a program is to use a *profiler*. A profiler uses various techniques like interrupting the program at random intervals, or taking measurements at specific control points (e.g., procedure call, on any change of control flow). What kinds of architecture measurements do you think profiling could apply to?

QUESTION 4—More Exercise

- Work through the examples in Chapter 1.
- Work through questions at the end of Chapter 1. Warning: the exercises in the book can sometimes be hard e.g. because details are left out, requiring you to make assumptions.