Chapter I
Quantitative analysis of computer performance

1

Quantitative analysis of computer performance

Objectives

The exercises in this chapter will demonstrate how to analyze quantitatively the performance of a computer, as well as how to apply Amdahl's law.

Exercise 1.

A computer uses a motherboard with two processor sockets. The system can work with one or two processors. The goal is to evaluate the performance of the computer using one or two processors, each one with two cores.

An estimation of the performance is obtained by running one or several instances of a benchmark in four different scenarios:

A) Only one processor is installed in the computer and only one instance of the benchmark is run. In this scenario, the average response time is 15 ms.

B) Only one processor is installed and two instances of the benchmark are run simultaneously, one on each core of the processor. In this scenario the average response time is 17 ms.

C) Two processors are installed and only one instance of the benchmark is executed. In this scenario the average response time is 16 ms.

D) Two processors are installed and four instances of the benchmark are run simultaneously, one on each processor core. In this scenario the average response time is 20 ms.

Each experiment is repeated several times and the value obtained is the average response time.

❑ **1.1** What do you think is the reason for the increment in the average response time when several instances of the benchmark are run simultaneously in several cores?

The use of several processor cores allows the system to have several execution units, but the rest of the components of the computer, such as memory and I/O devices, are not replicated. Thus, all the instances of the benchmark compete for these shared resources.

❑ **1.2** Think of a user that never launches tasks simultaneously. Would this user get a better performance with the monoprocessor or with the multiprocessor system? Justify quantitatively the answer.

They get better performance with a monoprocessor system since the response time executing a task (15 ms) is less than the response time in the multiprocessor system. The throughput in the monoprocessor system is $1/15 = 0.067$ tasks/ms = 67 tasks/s, whereas it is $1/16 = 0.063$ tasks/ms = 63 task/s in the multiprocessor system.

Taking the above results into account it could be concluded that using two processors the performance is worse than using only one processor. However, this conclusion is reached because the performance metric used is only valid for comparing the elapsed time of a task when it is the only task being executed in the system. When more tasks are run simultaneously in the computer, this metric is not valid.

❑ **1.3** If the user needs to run four tasks simultaneously, which system provides a better performance, monoprocessor or multiprocessor? Justify quantitatively your answer.

In this case the throughput must be analyzed:
Monoprocessor: $2/17 = 0.12$ tasks/ms = 120 tasks/s
Multiprocessor: $4/20 = 0.2$ tasks/ms = 200 tasks/s.
Thus, the multiprocessor system provides a better performance.

Exercise 2.

Improving the performance of a computer requires being aware of Amdahl's law:

> *The maximum improvement to an overall system when only part of the system is enhanced is limited by the fraction of the computation time that can take advantage of the enhancement.*

To illustrate this concept, think of a program in which 25% of the response time is used by input/output operations whereas the rest is used by the CPU. In this scenario two improvements are set out:

A) Replace the hard disk by another one twice as fast.

B) Replace the CPU by another one twice as fast.

❑ **2.1** Which is the speedup obtained taking each improvement into account? You must use the mathematical expression of the speedup deduced from the Amdahl's law.

$A_{(A)} = 1/[(1 - 0.25) + (0.25/2)] = 1.14$
$A_{(B)} = 1/[(1 - 0.75) + (0.75/2)] = 1.6$

❑ **2.2** In order to better understand the concept of speedup, you must compute now the speedup obtained by using the (A) improvement without using the theoretical expressions provided by the enhanced response time and the speedup. To do so, you can assume that the response time before the improvement is T, and try to divide it by the response time after the improvement.

`Initial time =` T
`Enhancement time =` $0,75 \times T + (0.25 \times T)/2$
`Division:` $A_{(A)} = 1/(0.75 + 0.125) = 1.14$

Exercise 3.

A program is run in 100 seconds; 35 seconds is CPU time and the rest is I/O time. The CPU time is reduced one third each year in the next 5 years; the I/O time, however, does not decrease.

❑ **3.1** Initially, which percentage of the response time of the program does the I/O time represent?

$65/100 \times 100 = 65\%$

❑ **3.2** Which will be the response time of the program after 5 years? Round the result to an integer.

$T = T_{CPU} + T_{I/O} = (35 \times (2/3)^5) + 65 \approx 70\,\mathrm{s}$

❑ **3.3** After 5 years, which percentage of the response time of the program does the I/O time represent?

$65/70 \times 100 \approx 92.86\%$

❑ **3.4** Which will be the speedup factor of the CPU after 5 years?

$35/(70 - 65) \approx 7$

❑ **3.5** Which will be the speedup factor of the computer after 5 years if the aforementioned program is used as a benchmark?

$100/70 \approx 1.43$

❑ **3.6** The results provided by the two previous answers should be different. How do you explain this difference?

```
The I/O time does not decrease and the percentage of I/O time
in the program increases.
```

Exercise 4.

The SPEC organization has decided to develop a new benchmark suite, called SPEC GPU. This new suite consists of two programs: `x264` and `ffmpeg`. These programs are in charge of measuring the performance of the GPU based on the elapsed time while coding several video sequences. Next, the average response time of both programs in the reference machine, S_{ref}, and in the system under inspection, SUT, are shown.

Benchmark	Response time SUT (s)	Response time S_{ref}	Speedup
`x264`	125	540	4.32
`ffmpeg`	165	775	4.7

❑ **4.1** Complete the above table computing the speedup of each benchmark in the SUT with regard to the S_{ref}.

If you have to choose an statistical value which includes the two speedup values as a general index representing the average performance of the SUT with regard to the S_{ref},

❑ **4.2** which one of the following would you choose?

1. Arithmetic mean
2. Geometric mean
3. Moving average
4. Standard deviation
5. Variance

```
2
```

❑ **4.3** Why?

```
The geometric mean is the only correct mean when averaging nor-
malized results.
The geometric mean consistently maintains the performance rela-
tionships regardless of the computer that is used as the basis
for normalization [Computer Organization and Architecture, W.
Stallings, 8th ed. p.74].
```

Exercise 5.

A speedup factor of 2 is desired in the response time of a program. 40% of response time of this program is CPU time and 60% is I/O time.

Several improvements can be made in the system to achieve this goal. The cost of each improvement is also shown.

I/O		CPU	
Speedup	Cost (Euro)	Speedup	Cost (Euro)
2	100	2	50
3	200	5	100
6	1000	20	1000

❑ **5.1** What is the required speedup in the CPU to achieve the required speedup for the program?

It is not possible to achieve the speedup goal for the program improving only the CPU, whatever the improvement of the CPU is.
$A = 1/[(1 - Fraction_{enhanced}) + (Fraction_{enhanced}/Speedup_{enhanced})]$
$2 = 1/[(1 - 0.4) + (0.4/X)] \Rightarrow X = -4$ (This result is not coherent)

❑ **5.2** What is the required speedup in the I/O to achieve the required speedup for the program?

$A = 1/[(1 - Fraction_{enhanced}) + (Fraction_{enhanced}/Speedup_{enhanced})]$
$2 = 1/[(1 - 0.6) + (0.6/X)] \Rightarrow X = 6$

❑ **5.3** If both components are improved simultaneously, with which option can the goal be achieved at a minimum cost?

The goal is to reduce the CPU and the I/O time 50% at least. Although several configurations would achieve this goal, the one with the minimum cost is that where the CPU and the I/O are improved using a speedup of 2 (CPU time is reduced down to 20% whereas I/O is reduced down to 30%). The cost of this configuration is 150 Euro.

Exercise 6.

Several benchmarks were run in two computers, A and B, as well as in the system of reference, S_{ref}.

Benchmark	Response time S_{ref} (s)	Response time A (s)	Ratio S_{ref}/A	Response time B (s)	Ratio S_{ref}/B
`wupwise`	1420	53	26.79	54.1	26.24
`swim`	2400	86	27.90	90	26.67
`mgrid`	1650	74	22.29	69	23.91
`applu`	2530	96	26.35	81	31.23
`mesa`	1120	42	26.67	41	27.31

❑ **6.1** Fill the table with the improvement ratios of each computer compared to the system of reference.

If you were requested to describe the performance of each computer based on the above benchmarks:

❑ **6.2** What is the speedup ratio of each computer?

The geometric mean of the speedup ratios of each bechmark for each computer is computed:
$A = 25.92$
$B = 26.97$

❑ **6.3** Which computer is the best? And in what proportion?

Computer B. $B/A \Rightarrow 26.97/25.92 = 1.04$

Exercise 7.

A program is run in Linux and its response time measured with the `time` command is shown below:

```
real   0m0.50s
user   0m0.38s
sys    0m0.02s
```

❑ **7.1** What is the percentage of CPU time?

$(0.38 + 0.02)/0.50 \times 100 = 80\%$

The clock frequency of the computer where the program was run is $2\,\text{GHz}$.

❑ **7.2** How many clock cycles does the CPU require to complete the processing part of this program?

$CPUtime/T = CPUtime \times f = 0.4 \times 2GHz = 800 \times 10^6$ clock cycles

It is also known that the execution of the program requires 500 millions of instructions.

❑ **7.3** What is the number of clock cycles per instruction or CPI (in average)?

CPI = 800×10^6 clock cycles / 500×10^6 instructions = 1.6

There are three options to reduce the response time of the program. Each option would be implemented independently. The options are the following:

1. Using a new compiler with code optimization. The number of instructions to be executed is reduced down to 400 millions.
2. Replacing the CPU by a new one with the same architecture but with a clock frequency of $2.5\,\text{GHz}$.
3. Replacing the CPU by a new one with the same clock frequency but with a new architecture where the average CPI is 1.4.

❑ **7.4** Which option is more effective? Justify your answer computing the CPU time for each one.

$CPUtime =$ `# of instructions` $\times CPI \times T$
`Option 1)` $400 \times 10^6 \times 1.6 \times [1/(2 \times 10^9)] = 0,32s$
`Option 2)` $500 \times 10^6 \times 1.6 \times [1/(2.5 \times 10^9)] = 0,32s$
`Option 3)` $500 \times 10^6 \times 1.4 \times [1/(2 \times 10^9)] = 0,35s$
`Options 1 and 2 are the best.`

Exercise 8.

Which of the following are TRUE statements? You may answer NONE if you think all of them are false.

A) Let T_A and T_B be the response times of a task in two computers A and B, respectively; then, A provides a speedup of T_A/T_B in relation to B.

B) Generally, when the number of cores of a processor is incremented, the throughput is also incremented.

C) Performance is only an important factor when talking about CPU, since the performance of the rest of the components of a computer barely influences its global performance.

D) The response time of a task can vary from execution to execution in the same computer.

E) Usually, the standard deviation of the response time of a task is zero.

F) Applying Amdahl's argument to program development, a good programming strategy is optimizing the code from the beginning without having a clear idea of the response time of each section of the program.

`B and D`

Exercise 9.

Which of the following are TRUE statements? You may answer NONE if you think all of them are false.

A) It can be deduced from Amdahl's law that a speedup value of 2 in a processor can never provide a response time in a task lower than half its original response time.

B) Given a computer with a larger MIPS value than another one, the performance of the former is better.

C) Let A and B be two processors with the same core; if processor A runs at a higher frequency than B, it can be said that the MIPS value of A is larger than the MIPS value of B and, furthermore, that the performance of A is better.

D) The response time of a program depends only on the CPU time.

E) Let A and B be two processors implementing the same instruction set; if processor A runs at a higher frequency than B, this does not imply that processor B has a worse performance the processor A since B can provide a lower CPI value than A.

F) The CPU time of a program shows the time that the CPU is running the program, without taking into account the waiting intervals.

`A, C, E and F`

Exercise 10.

Which of the following are TRUE statements? You may answer NONE if you think all of them are false.

A) The benchmarks used to measure the performance of a computer should be representative of the workload of the computer.

B) While comparing the performance of two computers using benchmarks, it is not possible that one benchmark indicates a better performance on one computer and another benchmark indicates a better performance on the other computer.

C) Real workload-based benchmarks have the drawback that they are difficult to reproduce in the same test conditions.

D) There are no benchmarks to measure the performance of different parts of a computer, such as the hard disks or the GPU.

E) The test conditions are really difficult to reproduce when working with synthetic workload-based benchmarks, but the results are significant.

F) Analytic workload-based benchmarks use mathematical models of the devices under test as well as of the workload.

```
A, C and F
```