CH3、Performance

計算機效能的評估

重點一：效能的定義

反應時間(response time、又稱執行時間 execution time)和生產量(throughput)是常見的評估標準。通常個人電腦會以『執行時間』做為效能的定義；多人伺服器則以『生產量』為評估標準

$$\text{Performance}_X = \frac{1}{\text{Execution Time}_X}$$

$$\text{Performance}_X > \text{Performance}_Y \rightarrow \frac{1}{\text{Execution Time}_X} > \frac{1}{\text{Execution Time}_Y}$$

$$\rightarrow \text{Execution Time}_Y > \text{Execution Time}_X$$

而不同電腦的速度會以 n 倍來代表，X 的速度是 Y 的 n 倍

$$\frac{\text{Performance}_X}{\text{Performance}_Y} = n$$

例(20)：In data center, what is the appropriate definition of good performance? It (1) gets a program done first (2) completes the most jobs during a certain period (3) has faster response time.

(2)

例(5)：In evaluating the performance of a computer, two performance metrics response time and throughput are usually used. Explain the meanings of these two terms.

Response time (execution time): the total time required for the computer to complete a task.
Throughput (bandwidth): it is the number of tasks completed per time unit.

練習：若 machine A 執行一程式需要 10 second，而 machine B 則花了 15 second.請問 A 比 B 快多少？

A 比 B 快 1.5 倍

$$\frac{Performance_A}{Performance_B} = \frac{Execution\ \ time_B}{Execution\ \ time_A} = \frac{15}{10} = 1.5$$

例(9)：If computer A runs a program in 15 seconds and computer B runs the same program in 25 seconds, how much faster is A than B?

Computer A is 25/15=1.67 times faster than computer B.

Execution Time

<table>
<tr><td rowspan="3"><i>執行一個程式所需時間 (wall-clock time、response time、elapsed time)</i></td><td colspan="2"><i>程式執行及等待IO時間</i></td></tr>
<tr><td rowspan="2"><i>程式使用CPU時間</i></td><td><i>CPU執行作業系統時間 (System CPU Time，不討論)</i></td></tr>
<tr><td><i>CPU執行使用者程式時間 (User CPU Time)</i></td></tr>
</table>

時脈(clock)

訊號到達的時間可能會因為電路的關係，而不同時到達加法器，故需要有一個時脈做為類似同步啟動計算的機制，有些以上緣觸發、有些以下緣觸發。

單一指令所需之時脈週期數(Cycles Per Instruction, CPI)

指令個數(Instruction Count, IC)

*CPU time = Clock Cycle * Clock Cycle Time*

= Clock Cycle / Clock Rate

*= IC * CPI * Clock Cycle Time*

*= IC * CPI / Clock Rate*

*= IC / (IPC * Clock Rate)*

(註)Peak performance = Clock Rate / smallest CPI

<table>
<tr><td colspan="3"><i>Execute Time</i></td></tr>
<tr><td colspan="2"><i>CPU Clock=</i> $\sum CPI*C_i$</td><td rowspan="2"><i>Clock Cycle Time=</i>$\frac{1}{Clock\ \ Rate}$</td></tr>
<tr><td rowspan="2"><i>IC</i></td><td>$CPI=\sum CPI * Frequence_i$</td></tr>
<tr><td colspan="2"><i>Instruction Time =</i> $\frac{1}{Instrstuction\ rate} = \frac{1}{MIPS*10^6}$</td></tr>
</table>

練習：假設我們有兩部使用相指令集架構*(相容，IC 一樣)*的計算機，對某個程式來說，Machine A 的 clock cycle time 為 250ps，CPI 為 2.0，Machine B 的 clock cycle time 為 500ps，CPI 為 1.2。請問哪一部機器較快？快多少？

令 IC=I

*CPU clock cycle$_A$ = I*2.0, CPU clock cycle$_B$=I*1.2*

*CPU time$_A$= CPU clcock cycle$_A$ * cycle time$_A$ = I*2.0*250 ps = 500I ps*

*CPU time$_B$= CPU clcock cycle$_A$ * cycle time$_A$ = I*1.2*500 ps = 600I ps*

由此可知，600I/500I = 1.2；Machine A 比 Machine B 快 1.2 倍

例(36)：Suppose we have two implementations of the same instruction set architecture. Computer A has a clock cycle time of 500 ps, and computer B has a clock rate of 2.5 GHz. Consider a program having 100 instructions.

1. Suppose computer A has a clock cycles per instruction (CPI) 2.3 for the program. Find the CPU time (in ns) for the computer A.
2. Suppose the CPU time of the computer B is 800 ns for the same program. Compute the CPI of computer B for the program.

*1. CPU time for computer A = 1000*2.3*500 ps = 1150 ns*

*2. 800 ns = 100 * CPI$_B$ *0.4 ns => CPI$_B$ =2*

例(52)：A program contains the following instruction mix:

60% load/store instructions with execution time of 1.2 us each.
10% ALU instructions with execution time of 0.8 us each.
30% branch instructions with execution time of 1.0 us each.

1. If the clock period is 0.2 us, calculate the average CPI (clock cycles per instruction) for the program.
2. What is the average MIPS rate of the program?

*1. CPI = 0.6 * (1.2/0.2) + 0.1 * (0.8/0.2) + 0.3 * (1.0/0.2) = 5.5*
2. MIPS = 1 / $(2.0*10^{-6})$ / $(5.5*10^{6})$ = 0.91

例(45)：We have the following statistics for two processors M1 and M2 (they have the same classes of instructions):

M1:200 MHz

1. Calculate the average CPI for the two instruction class
2. Calculate the MIPS (Million Instructions Per Second) for them.
3. Which machine is faster? How much faster?

*1. Average CPI for M1 = 5*0.25 + 2*0.4 + 3*0.35 = 3.1*
*Average CPI for M2 = 3*0.4 + 3*0.35 + 4*0.25 = 3.25*
2. MIPS for M1 = $200*10^{6}$ / $3.1*10^{6}$ = 64.52, MIPS for M2 = $250*10^{6}$ / $3.25*10^{6}$ = 76.92
*3. M2 is faster than M1 by 3.1*5ns / 3.25*4ns = 1.2 times*

練習：某一程式在 Machine A 上執行需 10 秒，Machine A 的 clock rate 為 2 GHz。我們現在要試著幫助一電腦設計人員設計另一 Machine B，預期只要 6 秒來完成同一程式的執行。該設計人員已確定大富增加 clock rate 是可行的，但是這種頻率的增加會影響處理器其它方面的設計，使得 Machine B 需要 A 的 1.2 倍 clock cycles 來完成相同程式的執行。我們得建議此設計人員採用的 clock rate 為何？

	A	*B*
Clock Rate	*2G*	*?*
Execution Time	*10*	*6*
CPU clock	*X (20G)*	*1.2x (24G)*

⇨ *Ans = 4G*

練習：編譯程式設計人員想要在兩套程式碼做選擇，他們提供如下的資料以供參考：

指令類別	CPI
A	1
B	2
C	3

現有一高皆語言之敘述，編譯器設計人員考慮以下兩種有不同指令數的程式碼

程式碼	各類別指令數		
	A	B	C
1	2	1	2
2	4	1	1

哪一種程式碼的指令數較多？何者較快？又兩者之 CPI 值為何？

程式碼 1 共執行了 2+1+2=5 指令。程式碼 2 共執行了 4+1+1=6 指令，所以程式碼 1 執行指令數較少。

*CPU clock cycle1 = 2*1 + 1*2 + 2*3 = 10 cycles*

*CPU clock cycle2 = 4*1 + 1*2 + 1*3 = 9 cycles*

所以程式碼 2 較快

CPI1 = CPU clock cycle1 / Instruction count1 = 10/5 = 2

CPI2 = CPU clock cycle2 / Instruction count2 = 9/6 = 1.5

例(23)：Consider two different implementations of the same instruction set architecture. There are four classes of instructions A, B, C, and D. The clock rate and CPI (cycles per instruction) of each implementation are given in the following table. Given a program with 108 instructions divided into classes as follows: 20% class A, 30% class B, 40% class C, and 10% class D.

1. Find the average CPI for P1
2. Find the average CPI for P2
3. Find the execution time for P1
4. Find the execution time for P2

	Clock Rage	CPI class A	CPI class B	CPI class C	CPI class D
P1	2.6 GHz	1	2	1	4
P2	1.5 GHz	2	3	1	2

1. *Average CPI for P1 = 1*0.2 + 2*0.3 + 1*0.4 + 4*0.1 = 1.6*
2. *Average CPI for P2 = 2*0.2 + 3*0.3 + 1*0.4 + 2*0.1 = 1.9*
3. *Execution time for P1 = (10^8*1.6) / 2.6G = 0.0615 sec*
4. *Execution time for P2 = (10^8*1.9) / 1.5G = 0.1267 sec*

練習：一個以 Java 寫成的應用程式在桌上型電腦執行需花 15 秒。一個新釋出的 Java 編譯器產生的指令數只有原先編譯器的 0.6 倍。但它會使 CPI 變成原先的 1.1 倍。我們可以預測新編譯器產生的程式碼可以跑多快？

1. 15*0.6 / 1.1 = 8.2 sec
2. 15 * 0.6 * 1.1 = 9.9 sec
3. 15 * 1.1 / 0.6 = 27.5 sec

2

練習：Consider three different processors P1, P2, and P3 executing the same instruction set with the clock rates and CPIs given in the following table.

Processor	Clock Rate	CPI
P1	2 GHz	1.5
P2	1.5 GHz	1.0
P3	3 GHz	2.5

1. Which processor has the highest performance?
2. If the processors each execute a program in 10 seconds, find the number of cycles and the number of instructions.
3. We are trying to reduce the time by 30% but this leads to an increase of 20% in the CPI. What clock rate should we have to get this time reduction?

1. *P2 has the highest performance*
   Instructions time of P1 = 1.5/2G = 0.75 ns
   Instructions time of P2 = 1.0/1.5G = 0.67 ns
   Instructions time of P3 = 2.5/3G = 0.83 ns

2. *Number of cycles = Execute time * Clock rate*
   *Cycles of P1 = 10*2G = 20G*
   *Cycles of P2 = 10*1.5G = 15G*
   *Cycles of P3 = 10*3G = 30G*
   Number of instructions = Number of cycles/CPI
   Instruction of P1 = 20G/1.5 = 13.33G
   Instruction of P2 = 15G/1 = 15G
   Instruction of P3 = 30G/2.5 = 5.14G
3. *Clock rate = (Number of Instruction * CPI) / Execution time*
   *Clock rate(P1) = 13.33G * 1.5 * 1.2 / 7 = 3.42 GHz*
   *Clock rate(P2) = 15G * 1 * 1.2 / 7 = 2.57 GHz*
   *Clock rate(P3) = 12G * 2.5 * 1.2 / 7 = 5.14 GHz*

軟硬體如何影響效能

	IC	CPI	Clock Rate	
演算法	O	O		*主要為軟體部分，故不影響 CR*
程式語言	O	O		*主要為軟體部分，故不影響 CR*
編譯器	O	O		*主要為軟體部分，故不影響 CR*
ISA	O	O	O	*ISA 介於軟硬體之間，故都影響*
電腦組織		O	O	*硬體不影響 IC*
VLSI 技術			O	

例(39)：The law of performance indicates that CPU time can be shown as a product of 3 terms. What are those three terms? Explain what factors may have impact on the three terms respectively.

Term	*Impact Factors*
IC	*演算法、程式語言、編譯器、ISA*
CPI	*演算法、程式語言、編譯器、ISA、電腦組織*
Clock cycle time	*ISA、電腦組織、VLSI*

練習：For problems below, use the information in the following table.

Processor	Clock Rate	IC	Time
P1	2 GHz	$20*10^9$	7 s
P2	1.5 GHz	$30*10^9$	10 s
P3	3 GHz	$90*10^9$	9 s

1. Find the IPC (instructions per cycle) for each processor.
2. Find the clock rate for P2 that reduces its execution time to that of P1.
3. Find the number of instructions for P2 that reduces its execution time to that of P3.

1. *IPC = 1/CPI = IC/(Execution Time * Clock Rate)*
   IPC(P1) = 1.42, IPC(P2) = 2, IPC(P3) = 3.33
2. *$Time_{new}/Time_{old}$ = 7/10 = 0.7*
   So Clock $Rate_{new}$ = Clock Rate/0.7 = 1.5 GHz/0.7 = 2.14 GHz
3. *$Time_{new}/Time_{old}$ = 9/10 = 0.9*
   So $IC_{newnew} = IC_{old}*0.9 = 30*10^9*0.9 = 27*10^9$ GHz

練習：Assume that multiply instructions take 12 cycles and account for 15% of the instructions in a typical program, and the other 85% of the instructions require an average of 4 cycles for each instruction.

1. What percentage of time does the CPU spend doing multiplication?
2. Your hardware engineering team has indicated that it would be possible to reduce the number of cycles required for multiplication to 8, but this will require a 20% increase in the cycle time. Nothing else will be affected by the change. Should they proceed with the modification?

1. (0.15*12) / (0.15*12+0.85*4) = 1.8/5.2 = 34.6%
2. New CPI = 0.15*8 + 0.85*4 = 4.6
   If the original cycle time is T, the instruction time for original machine is 5.2T and the instruction time for modified machine= 4.6*1.2T = 5.52T. So, the modification should not be made.

例(第四章 16)：同上

重點三：使用 MIPS 做為效能評估標準的謬誤

MIPS(million instructions per second, native MIPS)

$$\text{MIPS} = \frac{\text{IC}}{\text{Execution Time} * 10^6} = \frac{\text{IC}}{\frac{\text{IC} * \text{CPI}}{\text{Clock rate}} * 10^6} = \frac{\text{Clock Rate}}{\text{CPI} * 10^6}$$

以 MIPS 做為評估標準時需注意：

1. MIPS 並沒有把每一個指令的能力考慮進來(不絕對公平)
2. 同一電腦上的不同程式，MIPS 可能不同
3. MIPS 甚至可能會與效能呈反比(見下例)

例(41)：Assume that a processor is a load-store RISC CPU, running with 600 MHz. The instruction mix and clock cycles for a program as follows:

Instruction type	Frequency	Clock cycles
A	205%	2
B	10%	2
C	15%	3
D	30%	4
E	20%	1

1. Find the CPI.
2. Find the MIPS.

1. CPI = 0.25*2 + 0.1*2 + 0.15*3 + 0.3*4 + 0.2*1 = 2.55
2. MIPS = (600*10^6) / (2.55*10^6) = 235.29

例(28)：The performance of a 100MHz microprocessor P is measured by executing 10^7 instructions of Benchmark code, which is found to take 0.25s. What are the values of CPI and MIPS for this performance experiment?

CPI = Clock cycles/Instruction count = 100*10^6*0.25 / 10^7 = 2.5

MIPS = Instruction count / Execution time*10^6 = 107 / 0.25*10^6 = 40

例(32)：Does a computer with higher MIPS (millions of instructions per second) always have a faster response time than a computer with lower MIPS when executing the same program? Why?

1. MIPS 表示指令的執行率，但並沒有把每一指令的能力考慮進來。
2. 即使在同一台電腦上的不同程式，其 MIPS 亦可能不同
3. MIPS 甚至可能會與效能呈現反比

練習：某一電腦有三類指令 A、B、和 C，其 CPI 分別為 1、2、3。假設有兩個編譯器對同一程式進行編譯說產生程式碼，我們得到以下的資料：

Code from	Instruction counts (in billions)		
	A	B	C
Compiler 1	5	1	1
Compiler 2	10	1	1

假設電腦的時脈頻率是 4 GHz。若用 MIPS 當衡量標準，哪一個編譯器產生的程式碼將會執行得比較快？若用執行時間當衡量標準，又是哪一個比較快？

	Compiler 1	*Compiler 2*
CPU clocks	*1*5+2*1+3*1=10G*	*1*10+2*1+3*1=15G*
Clock Rate	*4G*	*4G*
Execution time	*10G/4G=2.5*	*15G/4G=3.75*
IC	*5+1+1=7G*	*10+1+1=12G*
MIPS	$7G / 2.5*10^6 = 2800$	$12G / 3.75*10^6 = 32000$

練習：Consider the following performance measurements for a program:

Measurement	Computer A	Computer B
IC	10 billion	8 billion
Clock Rate	4 GHz	4 GHz
CPI	1.0	1.1

1. Which computer has the higher MIPS rating?
2. Which computer is faster?

1. *Computer A*
   $MIPS_A = 4G / 1*10^6 = 4000, MIPS_B = 4G / 1.1*10^6 = 3636$
2. *Computer B*
   $ExTime_A = 10*10^9*1 / 4*10^9 = 2.5, ExTime_B = 8*10^9*1.1 / 4*10^9 = 2.2$

例(29)：The following results are measured for the same source code that is performed in two different processors:

Measurement	A	B
IC	5 billion	4 billion
Clock Rate	500 MHz	500 MHz
CPI	1	1.2

1. Which processor has the higher MIPS (million instructions per second) rating?
2. Which processor is faster?

1. *Computer A*
   $MIPS_A = 500*10^6 / 1*10^6 = 500, MIPS_B = 500*10^6/ 1.2*10^6 = 416.7$
2. *Computer B*
   $ExTime_A = 5*10^9*1 / 500*10^6 = 10, ExTime_B = 4*10^9*1.2 / 500*10^6 = 9.6$

例(27)：Compiling a same C program on two different computers, A and B, as follows, which of the following is correct?

	A	B
IC	10 million	Unknown
Clock Rate	4 GHz	5 GHz
CPI	1.0	1.2

1. The instruction count on B must be greater than 9.6 billion.
2. The program will spend less CPU time on B.
3. Computer B must have higher MIPS rating.
4. If the program runs faster (i.e. takes shorter CPU time) on A, the instruction count on B must be greater than 10 billion.
5. Computer A must consume less power than B.

3

註(3) ：$MIPS_A = 4G/1M = 4000 < MIPS_B = 5G/1.2M = 4166$

註(4) ：$CPU\_Time_A = 10*10^9*1 / 4*10^9 = 2.5 < CPU\_Time_B = IC_B*1.2 / 5*10^9 => IC_B > 10.42$ billions

註(5) ：$power = f * c * v^2$

練習 ：For the following set of variables, identify all of the subsets that can be used to calculate execution time. Each subset should be minimal; that is, it should not contain any variable that is not needed. {CPI, clock rate, cycle time, MIPS number of instructions in program, number of cycles in program}

<table>
<tr><td colspan="3">Execute Time</td></tr>
<tr><td colspan="2">CPU Clock= $\sum CPI*Ci$ =C</td><td rowspan="2">Clock Cycle Time=$\frac{1}{Clock\ \ Rate}$</td></tr>
<tr><td rowspan="2">IC = I</td><td>CPI= $\sum CPI * Frequence_i$</td></tr>
<tr><td colspan="2">Instruction Time = $\frac{1}{Instrstuction\ rate} = \frac{1}{MIPS*10^6}$</td></tr>
</table>

I = number of instructions in program ；C = number of cycles in program.

{Clock Rate, C} 、{Cycle Time, C} 、{MIPS, I} 、{CPI, C, MIPS} 、{CPI, I, Clock Rate} 、{CPI, I, Cycle Time}

練習 ：Consider two different implementations, P1 and P2, of the same instruction set. There are five classes of instructions (A, B, C, D, and E) in the instruction set. P1 has a clock rate of 4 GHz. P2 has a clock rate of 6 GHz. The average number of cycles for each instruction class for P1 and P2 is as follows:

Class	CPI on P1	CPI on P2
A	1	2
B	2	2
C	3	2
D	4	4
E	3	4

1. Assume that peak performance is defined as the fastest rate that a computer can execute any instruction sequence. What re the peak performances of P1 and P2 expressed in instructions per second?
2. If the number of instructions executed in a certain program is divided equally among the classes of instructions except for class A, which occurs twice as often as each of the others, how much faster is P2 than P1?

1. *For Machine M1:*
   *CPI = 0.6*1 + 0.3*2 + 0.1*4 = 1.6*
   For Machine M2:
   *CPI = 0.6*2 + 0.3*3 + 0.1*4 = 2.5*
2. *For Machine M1:*
   Average MIPS rating = $(80*10^6) / (1.6*10^6) = 50.0$
   For Machine M2:
   Average MIPS rating = $(100*10^6) / (2.5*10^6) = 40.0$

3. *Machine M2 has a smaller MIPS rating. If we change the CPI of instruction class A for Machine M2 to 1, we will have a better MIPS rating than M1:*
   *CPI = 0.61*1 + 0.3*3 + 0.1*4 = 1.9*
   Average MIPS rating = $(100*10^6) / (1.9*10^6) = 52.6$

例(33)：Your company uses a benchmark C to evaluate the performance of a computer A used in your company. But the computer A can only execute integer instructions, and it uses a sequence of integer instructions to emulate a single floating-point instruction. The computer A is rated at 200 MIPS on the benchmark C. Now, your boss would like to attach a floating-point coprocessor B to the computer A such that the floating-point instructions can be executed by the coprocessor for performance improvement. Note that, however, the combination of computer A and the coprocessor B is rated only at 60 MIPS on the same benchmark C. The following symbols are used in this problem:
I: the number of integer instructions executed on the benchmark C.
F: the number of floating-point instructions executed on the benchmark C.
N: the number of integer instructions to emulate a floating-point instruction.
Y: time to execute the benchmark C on the computer A alone.
Z: time to execute the benchmark C on the combination of computer A and the coprocessor B.

1. Write an equation for the MIPS rating of computer A using the symbols above.
2. Given I = 10^{*6}, F = $5*10^5$, N=30, find Y and Z.
3. Do you agree with your boss from the performance point of view? Please state the reasons to justify your answer.

1. $MIPS_A = (I+F*N) / (Y*10^6)$
2. $MIPS_A = (I+F*N) / (Y*10^6) \Rightarrow Y = (I+F*N) / (MIPS_A*10^6) = (5*10^6 + 5*10^5*30) / (MIPS_A *10^6) = (5*10^6 + 5*10^5*30) / (200*10^6) = 100$ *ms*
   $MIPS_{A+B} = (I+F) / (Z*10^6) \Rightarrow Z = (I+F) / (MIPS_{A+B} *10^6) = (5*10^6 + 5*10^5) / (60*10^6) = 91.67$ *ms*
3. *Yes. Although the MIPS of the processor/coprocessor combination seems to be lower than that of the processor alone, that is not the case. This is clearly seen from the execution times since it only takes 91.67 ms to execute the program with the coprocessor present as opposed to the 100 ms seconds without it.*

例(24)：We are interested in two implementations of a machine, one with and one without special floating-point hardware.
Consider a program P, with the following mix of operations:

Floating-point multiply	10%
Floating-point add	15%
Floating-point divide	5%
Integer instructions	70%

Machine MFP (Machine with Floating Point) has floating-point hardware and can therefore implement the floating point operations directly. It requires the following number of clock cycles for each instruction class:

Floating-point multiply	6
Floating-point add	4
Floating-point divide	20
Integer instructions	2

Machine MNFP (Machine with No Floating Point) has no floating-point hardware and so must emulate the floating-point operations using integer instructions. The integer instructions all take 2 clock cycles. The number of integer instructions needed to implement each of the floating-point operations is as follows:

Floating-point multiply	30
Floating-point add	20
Floating-point divide	50

Both machines have a clock rate of 100 MHz. Please find the native MIPS (Million Instructions per Second) ratings for both machines.

MIPS = Clock Rate / (CPI*10^6)
CPI for MFP = 0.1-*6 + 0.15*4 +0.05*20 + 0.7*2 = 3.6
The CPI for MNFP is simply 2.
So MIPS for MFP = 1000/CPI = 278 and MIPS for MNFP = 1000/2 = 500

重點四、AMDAHL'S 阿姆達爾定律

$$\text{Speedup} = \frac{1}{\frac{\text{F(受影響時間比例)}}{\text{S(改善倍率)}} + (1 - \text{F})}$$

例(4)：

1. Describe the definition of Amdahl's law.
2. Suppose we enhance a machine making all floating-point instructions run 10 times faster. If the execution time of some benchmark before the floating-point enhancement is 80 seconds, what will the speedup be if three-forth of the 80 seconds are spent executing floating-point instructions?

1. Amdahl's law: A rule stating that the performance enhancement possible with a given improvement is limited by the amount that the improved feature is used.
2. Speedup = 80 /(60/10 +20) = 3.08

練習：假設一程式在某機器上的執行時間為 100 秒鐘，而乘法運算就佔了 80 秒。如果想要讓此程式變成現在的 4 倍快，那麼該讓乘法運算部份改善多少才能達到這個要求呢？

100/4 = (100-80) + 80/x => x=16

例(26)：True or false: Suppose a program runs in 60 seconds on a machine, with multiplication responsible for 40 seconds of the time. According to Amdahl's law, we can simply improve the speed of multiplication to have the program run at 3 times faster.

False. Suppose x is the improvement rate, then 60/3 = 20 + 40/x => x 無限大

練習：假設我們改善一部機器，使得所有浮點指令執行的速度是原來的 5 倍快。讓我們來觀察當引入較快的浮點硬體時，其效能提昇(speedup)之情形。如果某個評效程式在改良浮點運算硬體之前的執行時間是 10 秒，若其中的 5 秒是花在浮點數運算指令上，則改善後的 speedup 值是多少？

Speedup = 10 / (10-5 + 5/5) = 5/3

例(2)：Consider a computer running a program with CPU times shown in the following table:

FP instruction	INT instruction	L/S instruction	Branch instruction
400 s	300 s	250 s	50 s

1. If the time for FP instructions is reduced by 40%, the total execution time of the program reduced in percentage is 20%.
2. If the time for L/S instructions is reduced by 80%, the speedup of the program is 1.25.
3. If the speed of the INT instructions is improved to 3 times faster, this program may run 1.25 times faster.
4. This program may run 1.8 times faster by improving the speed of FP instructions.

2、3

*註(1)：(400*0.6 + 300 + 25 + 50) // 1000 = 0.84；The Extime is reduced 16%.*

*註(2)：speedup = 100 / (400 + 300 + 250*0.2 + 50) = 1.25*

註(3)：speedup = 1 / (0.3/3 + 1-0.3) = 1.25

註(4)：speedup = 1 / (0.4/無限 + 1-0.4) = 1.67 < 1.8

練習：我們希望尋找新的效能評估程式來展示如上題中所敘述的新型浮點運算硬體，並期望整體效能提昇值是 3。目前一個正在考慮中的效能評估程式，在舊的浮點運算硬體上執行需 100 秒。為達到整體 speedup 值為 3，請問原花在執行浮點運算指令的時間應為多少？

3 = 100 / (100-x + x/5) => x = 250/3

例(48)：Amdahl's law is often written in terms of overall speedup as a function of two variables: the size of the enhancement (or amount of improvement) and the fraction of the original execution time that the enhanced feature is being used. Let t_{before}, t_{after}, $t_{affected}$, $t_{unaffected}$, $a_{improvement}$, and f denote execution time before improvement, execution time after improvement, execution time affected by improvement, execution time unaffected by improvement, amount of improvement, and the fraction affected, respectively. Derive this form of the speedup equation.

$Speedup = t_{before}/t_{after} = t_{before} / (t_{affected}/a_{improvement} + t_{unaffected}) = 1 / (f/a_{improvement} + 1-f)$

例(18)：Fill in O(True) or X(False)：Processor Performance

1. Program execution time increases when the clock rate increases.
2. Program execution time increases when the CPI increases.
3. Program execution time increases when the instruction count (IC) increases.
4. A speedup of 40 on 40% of the program will result in an overall speedup of at least 2.
5. A speedup of 20 on 50% of the program will result in an overall speedup of at least 2.
6. A speedup of 10 on 60% of the program will result in an overall speedup of at least 2.

1	2	3	4	5	6
X	O	O	X	X	O

註(4)：Speedup = 1 / (0.4/40 + 0.6) = 1.64

註(5)：Speedup = 1 / (0.5/20 + 0.5) = 1.9

註(6)：Speedup = 1 / (0.6/10 + 0.4) = 2.17

練習：對某一部機器而言，以下哪一種改變較有效率：

1. 將求浮點數平方根運算加速 10 倍，而浮點數平方根運算佔總執行時間的 20%
2. 加速其他浮點數運算 2 倍，其他浮點數運算佔總執行時間的 50%

假設兩種改變的成本是一種的而且是互斥的。

假設第一種及第二種改變的加速分別為 s1 及 s2

S1 = 1 / (0.2/10 + 1-0.2) = 1.22；s2 = 1 / (0.5/2 + 1-0.5) = 1.33。因此第二種改變較有效率

例(38)：You are going to enhance a computer, and there are two possible improvements: either make multiply instructions run four times faster than before, or make memory access instructions run two times faster than before. You repeatedly run a program that takes 100 seconds to execute. Of this time, 20% is used for multiplication, 50% for memory access instructions, and 30% for other tasks. Calculate the speedup:

1. Speedup if we improve only multiplication
2. Speedup if we only improve memory access
3. Speedup if both improvements are made

1. *Speedup= 1 / (0.2/4 + 0.8) = 1.18*
2. *Speedup = 1/(0.5/2 +0.5) = 1.33*
3. *Speedup = 1 / (0.2/4 + 0.5/2 + 0.3) = 1.67*

練習：Suppose a program segment consists of a purely sequential part which takes 25 cycles to execute, and an iterated loop which takes 100 cycles per iteration. Assume the loop iterations are independent, and cannot be further parallelized. If the loop is to be executed 100 times, what is the maximum speedup possible using an infinite number of processors (compared to a single processor)?

*Execution time before improvement = 100*100 + 25 = 10025 cycles*
Execution time after improvement = 100 + 25 = 125 cycles
Speedup = 10025/125 = 80.2

例(16)：

1. Please describes the Amdahl's law, and uses the Amdahl's law to explain why the multicore processor cannot get the similar improvement factor as the core numbers.
2. Suppose a program segment consists of a purely sequential part which takes 20 cycles to execute, and an iterated loop which takes 200 cycles per iteration. Assume the loop iterations are independent, and cannot be further parallelized. If the loop is to be executed 200 times, what is the maximum speedup possible using an infinite number of processors (compared to a single processor)?

1. *Amdahl's law makes it possible to determine the theoretical maximum speed increase in a system that can be obtained by using multicore processor. Since only some portions of the system can be run in parallel, a limit will eventually be reached where no more cores can be run alongside one another. That is, the speed-up of a program using multicore processor in parallel is limited by the time needed for sequential portions of the program to execute.*
2. *Execution time before improvement = 200*200 + 20 = 40020 cycles*
   Execution time after improvement = 200 + 20 = 220 cycles
   Speedup = 40020/220 = 181.91

例(42)：We make an enhancement to a computer that improves some mode of execution by a factor of 10. Enhanced mode is used 80% of the time, measured as a percentage of the execution time when the enhanced mode is in use.

1. What is the speedup we have obtained from fast mode?
2. What percentage of the original execution time has been converted to fast mode.

Hint: The Amdahl's law depends on the fraction of the original, unenhanced execution time that could make use of enhanced mode. Thus, we cannot directly use this 80% measurement to compute speedup with Amdahl's law.

1. *Speedup = $Time_{unenhanced}$/$Time_{enhanced}$*
   The unenhanced time is the sum of the time that does not benefit from the 10 times faster speedup, plus the time that does benefit, but before its reduction by the factor of 10. Thus, $Time_{unenhanced} = 0.2 * Time_{enhanced} + 10*0.8*Time_{enhanced} = 8.2Time_{enhanced}$
   Substituting into the equation for speedup gives us:
   Speedup = $Time_{unenhanced}/Time_{enhanced} = 8.2$
2. *Using Amdahl's law, the given value of 10 for the enhancement factor, and the value for Speedup from Part(a), we have: 8.2 = 1 / (1-f + f/10) => f = 0.9756. Solving shows the enhancement can be applied 97.56% of the original time.*

練習：Assume that a design team is considering enhancing a machine by adding MMX (multimedia extension instruction) hardware to a processor. When a computation is run in MMX mode on the MMX hardware, it is 10 times faster than the normal mode of execution. Call the percentage of time that could be spent using the MMX mode the percentage of media enhancement.

1. What percentage of media enhancement is needed to achieve an overall speedup of 2?
2. What percentage of the run-time is spent in MMX mode if a speedup of 2 is achieved? (Hint: You need to calculate the new overall time.)
3. What percentage of the media enhancement is needed to achieve 1/2 the maximum speedup attainable from using the MMX mode?

1. *Let x be the percent of media enhancement needed for achieving an overall speedup of 2. Then, Speedup = 1 / (x/10 + 1-x) = 2 => x = 55%*
2. *The percentage of the run-time that is spent in MMX mode = 0.55/10 / (0.55/10 + 1-0.55) = 10.89%*
3. *The maximum speedup using MMX mode occurs when the whole program can run in media enhancement mode and is 10. One-half of this is 5.*

重點五：效能總評

1. 算數平均
2. 加權算術平均(Weighted arithmetic mean, WAM)
3. SPECratio(愈大愈好)：程式執行時間/機器執行時間(正規化)
4. 幾何平均：優點是無關執行時間與哪台機器；缺點是無法預測時間

	電腦 A	電腦 B	對 A 正規化		對 B 正規化	
			A	B	A	B
Program 1	1	10	1	0.1	10	1
Program 2	1000	100	1	10	0.1	1
AM	500.5	55	1	5.05	5.05	1
GM	31.6	31.6	1	1	1	1

例(11)：Table 1 shows the execution times of three programs on two different computers. Please compare the performance between computers A and B.

	Computer A	Computer B
Program 1	100 second	200 second
Program 2	200 second	150 second
Program 3	150 second	100 second

AM for computer A = (100+200+150)/3 =150 sec
AM for computer B = (200+150+100)/3 = 150 sec
So, computer A is as fast as computer B.

練習：The following table shows results for SPEC2006 benchmark programs running on an AMD Barcelona.

		IC*109	Execution time (sec)	Reference time (sec)
a.	peri	2118	500	9770
b.	mcf	336	1200	9120

1. Find the CPI if the clock cycle time is 0.333 ns.
2. Find the SPEC ration.
3. For these two benchmarks, find the geometric mean.

Clock rate = 1/cycle time = 3 GHz

	1	2	3
a.	$(3G*500)/21118*10^9=0.7$	9770/500=19.54	$(19.54*7.6)^{1/2} = 12.19$
b.	$(3G*1200)/336*10^9=10.7$	9120/1200=7.6	

例(25)：Suppose the execution times for 3 programs on3 machines are given below:

	Program X	Program Y	Program Z
Machine A	50	30	10
Machine B	10	30	20
Machine C	20	15	10

1. Using Program X, Y and Z as benchmarks, please calculate the unweighted arithmetic means of the execution times for the 3 machines.
2. Using Program X, Y and Z as benchmarks, please calculate the geometric means of the execution times for the 3 machines, with Machine A as the reference machine (as in the SPEC benchmarks).
3. Which machine has the overall highest performance? Please justify your answer in detail.

	Program X		Program Y		Program Z		1	2
	ExTime	SPEC ratio	ExTime	SPEC ratio	ExTime	SPEC ratio	AM	GM
M_A	50	1	30	1	10	1	30	1
M_B	10	5	30	1	2	0.5	20	Sqrt3(2.5)
M_C	20	2.5	15	2	10	1	15	Sqrt3(5)

3. Machine C has the highest performance since its AM is the smallest among the three machines.

練習：The following table shows data for benchmarks.

	Name	CPI	Clock Rate	SPECratio
a.	sjeng	0.96	4 GHz	14.5
b.	omnetpp	2.94	4 GHz	9.1

1. Find the increase in CPU time if the number of instruction of the benchmark is increased by 10% without affecting the CPI.
2. Find the increase in CPU time if the number of instruction of the benchmark is increased by 10% and the CPI is increased by 5%.
3. Find the change in the SPECratio for the change described in 2.

1. If CPI and clock rate do not change, the CPU time increase is equal to the increase in the number of instructions, that is, 10%

*2. CPU time (after) = (1.1*Instruction * 1.05*CPI) / Clock rate*
*CPU time(after)/CPU time(before) = 1.1*1.05=1.155. Thus, CPU time is increased by 15.5%.*

3. The SPECratio is decreased by 14%.
SPEC(after)/SPEC(before) = Time(before)/Time(after) = 1/1.155 = 0.86

練習：You wonder how the performance of the three computers in the following table would compare using other means to normalize performance. Which computer is fastest by the geometric mean?

Program	Floating-point operations	Execution time in seconds		
		Computer A	Computer B	Computer C
1	1000 0000	1	10	20
2	1 0000 0000	1000	100	20

Computer C is faster
*GM(A) =sqrt(1*1000) = 32, GM(B) =sqrt(10*100) = 32, GM(C) =sqrt(20*20) = 20*

重點六：效能評估程式(benchmark)

SPEC(System Performance Evaluation Corporation)
最新版本的 SPEC 處理器評效程式是 SPEC CPU2000，其中包括了 12 整數程式(CINT 2000)與 14 浮點數程式(CFP 2000)；第一個專門測網頁伺服器效能的是 SPECweb99

例(22)：SPEC is an organization about 1.benchmark 2.embedded computers 3.IC technology.
3

例(19)：(Circle all correct answers) Consider the performance of SPECweb99 listed in the following table. Which of the following statements are correct?

1. Clock rate determines the web service performance.
2. The system with 2 1 GHz processors, 7 disk, and 3 network connections is likely better than the system with 2 1.4 GHz processor, 2 disks, and 1 network connection for this benchmark.
3. Throughput is a more suitable metric than response time in measuring the performance of SPECweb99.
4. Since this is a web service, increasing the number of networks always improve the response time.

Processor	# of disk drives	# of CPU	# of networks	Clock Rate(GHz)	Result
P3	3	2	1	1.4	1810
P3	8	2	4	1.13	3435
P3 xeon	5	4	4	0.7	4200
P4 xeon	10	2	4	2.2	4615

2、3

練習：Consider the SPEC benchmark. Name two factors that influence the resulting performance on any particular architecture.

1. *The compiler (flags) used to compile the benchmark*
2. *The input data that is given to the benchmark while measuring performance*

例(51)：A certain machine with a 10 ns ($10*10^{-9}$s) clock period can perform jumps (1 cycle), branches (3 cycles), arithmetic instructions (2 cycles), multiply instructions (5 cycles), and memory instructions (4 cycles). A certain program has 10% jumps, 10% branches, 50% arithmetic, 10% multiply, and 20% memory instructions. Answer the following question. Show your derivation in sufficient detail.

1. What is the CPI of this program on this machine?
2. If the program executes 10^9 instructions, what is its execution time?
3. A 5-cycle multiply-add instruction is implemented that combines an arithmetic and a multiply

instruction. 50% of the multiplies can be turned into multiply-adds. What is the new CPI?

4. Following 3 above, if the clock period remains the same, what is the program's new execution time.

1. *1*0.1 + 3*0.1 + 2*0.5 + 5*0.1 + 4*0.2 = 2.7*
2. *Execution time = 10^9 *2.7*10ns = 27s*
3. *CPI = (1*0.1 + 3*0.1 + 2**0.45 + 5*0.05 + 4*0.2 + 5*0.05) / 0.95 = 2.74*
4. *Execution time = 10^9*0.95*2.74*10 ns = 26.03 s*

例(7)：A processor and two options for improving its hardware and compiler design are described as follows:

The base machine, M_{base}:

M_{base} has a clock rate of 200 MHz and the following measures:

Instruction class	CPI	Frequency
A	1	50%
B	2	25%
C	4	25%

The machine with improved hardware, M_{hw}:

M_{hw} has a clock rate of 500 MHz and the following measures:

Instruction class	CPI	Frequency
A	2	50%
B	4	25%
C	6	25%

The combination of the improved compiler and the base machine, M_{comp}:

The instruction improvements from this enhanced compiler are as follows:

Instruction class	Percentage of instruction executed vs M_{base}
A	80%
B	80%
C	40%

Calculate the CPI (clock cycles per instruction) for each machine and the speedups of M_{hw} and M_{comp} with respect to M_{base}.

Machine	CPI (cycle/instruction)	Speedup
M_{base}		1
M_{hw}		
M_{comp}		

Machine	*CPI (cycle/instruction)*	*Speedup*
M_{base}	*2*	*1*
M_{hw}	*3.5*	*1.43*
M_{comp}	*2.57*	*1.11*

*CPI_{Mbase} = 0.5*1 + 0.25*2 + 0.25*4 = 2*

*CPI_{Mhw} = 0.5*2 + 0.25*4 + 0.25*6 = 3.5*

*0.5*0.8 + 0.25*0.8 + 0.25*0.4 =0.7*

*CPI_{Mcomp} = (0.5*0.8*1 + 0.25*0.8*2 + 0.25*0.4*4) / 0.7 = 1.714*

*ExeTime of M_{base}/M_{hw} = IC*2/200*10^6 / IC*3.5/500*10^6 = 1.43*

*ExeTime of M_{base}/M_{Comp} = IC*2/200*10^6 / 0.7IC*1.714/200*10^6 = 1.67*

例(15)：You are considering adding a register-memory addressing mode to the MIPS instruction set. With this new addressing mode, there would be 5% increase in cycle time and the CPI of the ALU instruction type (including both the register-register and register-memory types) becomes 2.

1. What is the average CPI of the old implementation?

2. What type of instructions could be reduced?
3. If half of the instructions that you answer in question 2 can be reduced with the new addressing mode, what is the average CPI of the new implementation? Will you decide to add the new addressing mode? Please explain your answer.

OP	Freq	CPI
ALU	40%	1
Load	20%	3
Store	20%	3
Branch	20%	2

1. *CPI_{old} = 1*0.4 + 3*0.2 + 3*0.2 + 2*0.2 = 2*
2. *Load*
3. *Suppose the IC is IC and the cycle time for the old implementation is T, then the execution time for the old implementation is IC*2*T*
   *CPI_{new} = (2*0.4 + 3*0.1 + 3*0.2 + 2*0.2) / 0.9 = 2.33*
   *The execution time for the new implementation is 0.9*IC*2.33*1.05T = IC*2.2*T*
   So, to add the new addressing mode is not a good choice.

例(13)：The following table shows the number of floating-point operations executed in three different programs and the runtime for those programs on three different computers:

Program	FP operations	Execution in seconds		
		Computer A	Computer B	Computer C
P1	$10*10^9$	1	2	5
P2	$40*10^9$	10	10	10
P3	$80*10^9$	100	25	8

Suppose that the total number of floating-point operations executed in the workload is equally divided among the three programs. That is , program 1 runs 8 times for every time program 3 runs, and program 2 runs twice for every time program 3 runs. Find which computer is fastest for this workload and by what factor.

*Computer A = 8*1 + 2*20 + 1*100 = 148s*
*Computer B = 8*2 + 2*10 + 1*25 = 61s*
*Computer C = 8*5 + 2*10 + 1*8 = 68s*
So with this workload, computer B is faster by a factor of 68/61 = 1.14 with respect to computer C and a factor of 148/61=2.42 with respect to computer A