Capítulo 2

MIDIENDO EL DESEMPEÑO

Toda decisión que tome un arquitecto de computadoras tiene impacto en dos rubros: el costo y el desempeño del equipo. Así que, como en el resto de los ámbitos que componen nuestra experiencia en el mundo, la mejor decisión es aquella con la mejor relación costo-beneficio. Un primer requisito para tomar una buena decisión es, entonces, poseer una clara idea de cómo las diferentes alternativas disponibles tienen impacto en el desempeño y esto significa *cuantificarlo*.

Existen dos tipos diferentes de medidas de desempeño:

- Aquellas directamente relacionadas con el tiempo que tarda un sistema en terminar una tarea determinada. Esto es, el *tiempo de respuesta* del sistema y, claro está, un desempeño es tanto mejor cuanto menor sea el tiempo de respuesta, asi que es una de las medidas que suelen clasificarse como *mejor cuanto menor* o LIB por la siglas de *Lower Is Better*.

- Las relacionadas directamente con el número de labores realizadas por unidad de tiempo. Es decir el *rendimiento* (*throughput*) del sistema. En este caso un desempeño mejor se asocia con un número mayor de tareas realizadas, así que es una medida del tipo *mejor cuanto mayor* o HIB por las siglas de *Higher is Better*.

En general los usuarios de un sistema de cómputo están más interesados en conocer el desempeño del sistema en términos del primer grupo de medidas y los administradores o proveedores de servicios prefieren

conocer una medida del segundo tipo. Por supuesto en general si se tiene una medida del primert tipo es posible transformarla en una del segundo usando su inverso multiplicativo y viceversa.

Prueba de desempeño

Necesariamente para medir el desempeño de un sistema de cómputo se requiere de un instrumento de medida y dado que nuestras computadoras lo son porque ejecutan programas, se requiere entonces de uno o varios programas que pongan a trabajar al sistema y medir los tiempos de ejecución. A la evaluación del desempeño de un sistema se le deonomina *benchmark* en inglés, en español se suele traducir como *prueba de desempeño*, *análisis comparativo* o símplemente *comparativa*, lo que es bastante adecuado dado que, en efecto, medir es comparar. Por supuesto nos interesa poder comparar el desempeño de un sistema con el de otro. Algo como "El rendimiento del sistema A es 2.5 veces mayor que el del sistma B", formalmente:

$$\frac{\mathrm{R}end(A)}{\mathrm{R}end(B)} = 2.5$$

o equivalentemente, si el tiempo de respuesta es representado por T:

$$\frac{T(B)}{T(A)} = 2.5$$

Comparar el desempeño de dos sistemas es una labor delicada. Observemos, por ejemplo, los datos de tiempo de ejecución de dos sistemas mostrados en la tabla 2.1. Se puede decir, sin faltar a la verdad, que el sistema B es mejor que A ejecutando los programas 1 y 2. Se puede decir, también, que A es mejor ejecutando los programas 3, 4 y 5. ¿Cómo zanjar el conflicto?

Síntesis de desempeño

Lo primero que se requiere es obtener una única medida de desempeño. Un sólo número que refleje la evaluación hecha. Una buena opción podría ser considerar el tiempo total de ejecución, listado en el último renglón de la tabla. Considerándolo como una medida de síntesis del desempeño, podríamos decir que tanto A como B son igualmente buenos. En general el tiempo total de ejecución puede ser una manera correcta de sitetizar el desempeño, pero quisieramos tener un parámetro estadísticamente más significativo. Que no dependa ni de cuantos programas se ejecutan ni de cuantas veces se ejecutan, que nos proporcione una estimación del desempeño del equipo en su uso cotidiano. Lo que queremos es, de hecho, lo que los estadísticos denominan una *medida de tendencia central*: un número alrededor del cual, en general, se encontrará las más de las veces el desempeño de la máquina.

Programa	Computadora A	Computadora B
Prog1	20	1
Prog2	15	2
Prog3	10	15
Prog4	10	27
Prog5	10	20
Total	**65**	**65**

Tabla 2.1. Tiempos de ejecución en milisegundos de cinco programas para dos sistemas de cómputo diferentes.

Medidas de tendencia central

Existen varias medidas de tendencia central usadas en estadística. La *moda* es una de ellas, definida como el valor de una muestra de datos que más se repite, puede ni siquiera existir. La *mediana*, definida como el punto medio entre los valores extremos de la muestra, tiene la desventaja de dejarse llevar fácilmente por valores excesivamente altos o excesivamente bajos en la muestra. Necesitamos medidas de tendencia central mucho más robustas. Por fortuna tenemos aún mucho de donde escoger. Si $D = \{\delta_1, \delta_2, \ldots, \delta_n\}$ es una muestra de n datos, se definen:

Media aritmética

$$\mathrm{A}(D) = \frac{1}{n} \sum_{i=1}^{n} \delta_i \tag{2.1}$$

Media armónica

$$\mathrm{H}(D) = \frac{n}{\sum_{i=1}^{n} \frac{1}{\delta_i}} \tag{2.2}$$

Media geométrica

$$\mathrm{G}(D) = \sqrt[n]{\prod_{i=1}^{n} \delta_i} \tag{2.3}$$

La media aritmética es lo que normalmente llamamos el *promedio* de la muestra de datos y es, con mucho, la medida de tendencia central más usada en la vida cotidiana, además de ser la más simple de calcular. Sin embargo puede no ser la mejor en toda circunstancia. Es bien conocida, por ejemplo, su indeseable tendencia a dejarse llevar por valores excesivamente grandes respecto a la meyoría de los datos (*outliers*) de una muestra.

Velocidad (tareas/min)
70
30
40
60

Tabla 2.2. Las cuatro velocidades a las que puede operar un sistema de cómputo. La media aritmética es 50 tareas/minuto, la media armónica es 44.8 tareas/minuto.

Relación entre las medias

Dados un conjunto de datos, la media aritmética siempre será la mayor de las tres, la armónica la menor y la geométrica estará siempre entre ambas. La media armónica, por cierto, es el inverso de la media aritmética de los inversos de los elementos de la muestra, por lo que también se puede sesgar indeseablemente ante valores relativamente alejados del resto de la muestra. La media geométrica, por su parte, tiene un significado interesante: es el tamaño que debería tener el lado de un hipercubo para tener un volumen igual al hiperprisma cuyos lados son los elementos de la muestra. Sí, por ejemplo, tenemos $\{3, 5, 4\}$ como muestra y consideramos estos como las medidas en centímetros del ancho, alto y largo de un prisma cuadrangular, el volúmen del prisma sería 60 cm^3. La media geométrica de la muestra es 3.915, si suponemos que esto es el largo en centímetros del lado de un cubo, enonces el volúmen del cubo es igual al del prisma previo.

Diferentes medias y diferentes usos

Cada media es útil por sí misma como medida de tendencia central en ciertas circunstancias. No todas son utilizables a ultranza. Supongamos, por ejemplo, que tenemos un sistema que puede realizar cierto tipo de tareas en cuatro velocidades diferentes como se muestra en la tabla 2.2. La media aritmética de las velocidades es de 50 tareas/min, mientras la media armónica es de 44.8 tareas/min.

Supongamos ahora que ponemos a funcionar este sistema durante dos minutos en cada uno de los regímenes de velocidad, haciendo entonces un total de 8 minutos. En la tabla 2.3 se muestran los resultados. El número total de tareas ejecutadas es 400. Si usamos la media armónica (44.8 tareas/min) como medida de tendencia central y decimos que ese fué la tasa a la que se ejecutaron tareas durante 8 minutos, entonces obtenemos: $44.8 \times 8 = 358.4$ tareas ejecutadas, lo que discrepa de las 400 que obtenemos sumándolas. Si en cambio usamos la media aritmética (50 tareas/min) como la tasa de ejecución durante los 8

Vel (tareas/min)	Trabajo (min)	#tareas
70	2	140
30	2	60
40	2	80
60	2	120
Total	8	400

Tabla 2.3. Resultado de ejecutar tareas durante dos minutos en cada régimen en el sistema de ejemplo. La suma del total de tareas ejecutadas es 400.

Vel (tareas/min)	Tareas	Tiempo (min)
70	14	0.2
30	14	0.47
40	14	0.35
60	14	0.23
Total	56	1.25

Tabla 2.4. Resultado de ejecutar 14 tareas en cada régimen en el sistema de ejemplo. La suma del total de tiempo es 1.25 minutos.

minutos obtenemos: $50 \times 8 = 400$ tareas, lo que es apegado a la realidad.

Si, en cambio, la carga de trabajo fija es el número de tareas realizadas en cada régimen, como se muestra en la tabla 2.4, el tiempo total de ejecución (suma de la tercera columna) es de 1.25 minutos. Si consideramos la media aritmética como la tasa a la que se ejecutan las 56 tareas totales, entonces el tiempo estimado sería: $56/50 = 1.12$ minutos, muy distinto de los 1.25 minutos. Por el contrario, si consideramos la media armónica, entonces obtenemos: $56/44.8 = 1.25$ minutos, tal como debe ser.

Así que cuando la cantidad que se desea medir es directamente proporcional a la métrica de las muestras se debe usar la media aritmética. Por el contrario, si la relación es inversamente proporcional, se debe usar la media armónica.

Producción (ton)	Factor de crecimiento
100	
130	1.3000
180	1.3846
240	1.3333
305	1.2708
345	1.1311
M. arit.	1.284
M. geom.	1.281
M. armo.	1.278

Tabla 2.5. Producción anual, en toneladas, de una empresa productora de cereales durante los últimos seis años.

La media geométrica tiene también su propio ámbito de uso adecuado. Supongamos que la tabla 2.5 muestra la producción anual de una empresa agrícola productora de cereales.

Media geométrica y tasas

Si consideramos la media aritmética del factor de crecimiento como la tasa promedio del mismo, obtendríamos: $100 \times 1.284^5 = 348.98$ toneladas de producción en el último año, mayor que la producción real. Si usaramos la media armónica obtendriamos $100 \times 1.278^5 = 340.91$ lo que es inferior a la producción real. Pero si usamos la media geométrica tendríamos: $100 \times 1.281^5 = 345$ lo que realmente concide con la producción del último año. En este caso la media geométrica es la medida de tendencia central que se debe usar. La métrica que usamos es, realmente, una razón sin unidades, comparamos una magnitud con otra de la misma métrica. La media geométrica es lo adecuado para obetener medidas de tendencia central en razones de proporción o *tasas* (lo que suele llamarse *ratio* en inglés). Más adelante veremos una cualidad adicional muy útil de esta media.

Media aritmética ponderada

En el conjunto de programas que se utilizan en una prueba de desempeño habrá algunos más representativos que otros. Para juzgar correctamente habría que asignar a cada programa un cierto peso que fuera indicativo de qué tan estadísticamente representativo es. Podríamos pensar entonces en usar una media aritmética en la que se asocie un peso diferente a cada uno de los programas en función de su representatividad. La medida en cuestión es la *media aritmética ponderada*.

Programa	A	B	C
Prog1	5	13	35
Prog2	730	250	45
Prog3	1200	230	50
Total	**1935**	**493**	**130**

Tabla 2.6. Tiempos de ejecución (ms) para tres programas en tres diferentes máquinas.

Dada una muestra $D = \{\delta_1, \delta_2, \ldots, \delta_n\}$ de n datos y un conjunto de *pesos* $W = \{w_1, w_2, \ldots, w_n\}$, tales que $\sum_{i=1}^{n} w_i = 1$, la media aritmética ponderada se define como:

$$\mathrm{P}(D, W) = \sum_{i=1}^{n} w_i \, \delta_i \tag{2.4}$$

Así la media aritmética convencional resulta ser un caso particular de la ponderada en el que los pesos de los programas son todos iguales a $1/n$.

Normalización de tiempos

Tendríamos ahora la dificultad de elegir adecuadamente los pesos de los programas. Otra opción es fijar una máquina como patrón de medida. Decidir que hay una cierta computadora p que posee una arquitectura suficientemente general que le concede un comportamiento representativo y referir los tiempos de ejecución a esa máquina. Los tiempos a considerar deben ser entonces normalizados, divididos por los tiempos en la computadora de referencia. Así, el valor a considerar para el i-ésimo programa sería:

$$\hat{t}_i = \frac{t_{i,c}}{t_{i,p}} \tag{2.5}$$

En este contexto la medida de tendencia central adecuada sería la media geométrica, dado que los valores de la muestra son cocientes de tiempos, razones. Utilizar la media aritmética, por ejemplo, puede entregar resultados inconsistentes, en función de la máquina usada como referencia.

En la tabla 2.6 se muestran los tiempos de ejecución en milisegundos para tres diferentes programas en tres diferentes computadoras: A, B y C. En las tablas 2.7, 2.8 y 2.9 se muestran los tiempos normalizados usando a A, B y C como máquina de referencia, respectivamente.

Programa	A	B	C
Prog1	1	2.6	7
Prog2	1	0.34	0.06
Prog3	1	0.19	0.04
Total	**3**	**3.13**	**7.1**
M. arit.	**1**	**1.04**	**2.37**
M. geom.	**1**	**0.55**	**0.26**

Tabla 2.7. Desempeño de los tres sistemas de ejemplo, usando los tiempos de A para normalizar.

Programa	A	B	C
Prog1	0.38	1	2.69
Prog2	2.92	1	0.18
Prog3	5.22	1	0.22
Total	**8.52**	**3**	**3.09**
M. arit.	**2.84**	**1**	**1.03**
M. geom.	**1.8**	**1**	**0.47**

Tabla 2.8. Desempeño de los tres sistemas de ejemplo, usando los tiempos de B para normalizar.

Programa	A	B	C
Prog1	0.14	0.37	1
Prog2	16.22	5.56	1
Prog3	24	4.6	1
Total	**40.37**	**10.53**	**3**
M. arit.	**13.46**	**3.51**	**1**
M. geom.	**3.82**	**2.12**	**1**

Tabla 2.9. Desempeño de los tres sistemas de ejemplo, usando los tiempos de C para normalizar.

Media geométrica y máquina de referencia

Como puede observar el lector, si la medida de desempeño usada es el tiempo normalizado total o su media aritmética, entonces la computadora ganadora es justo la que se elije como referencia. La única medida consistente es la obtenida usando la media geométrica. En ese caso la máquina ganadora es siempre la C, lo que es congruente con el tiempo total de ejecución de la prueba de la tabla 2.6. Este comportamiento resulta del hecho de que la media geométrica de los cocientes de dos muestras, es igual al cociente de las medias geométricas de las muestras, es decir:

$$G\left(\frac{X}{Y}\right) = \frac{G(X)}{G(Y)}$$

La desventaja de usar la media geométrica de tiempos normalizados es que perdemos la noción intuitiva que deseabamos en principio: poder decir, "el sistema X tiene un desempeño 3.5 veces mejor que el sistema Y". Como lo que se maneja son razones de tiempos y se pasan por la media geométrica ya no hay una noción intuitiva de nada. El hecho de que la media aritmética "se deje llevar" por valores grandes o pequeños, significa, desde el punto de vista positivo, que se toma en cuenta la información que se recibe, si deseamos que nuestra medida no se sesgue tan fácilmente es porque, implícitamente, algo de la información se pierde. Todo tiene un costo.

Las ventajas, sin embargo son grandes: no se sesga ante valores extrañamente grandes o pequeños; se puede usar un conjunto de valores normalizados sin que importe la elección de los valores de referencia. Es por esto que la media geométrica fue la elegida para sintetizar las pruebas de desempeño más científicas que poseemos y que nos ocuparán en la siguiente sección.

2.1 PRUEBAS DE DESEMPEÑO

Carga de trabajo

Al conjunto de programas usados para evaluar dicho desempeño se le llama *carga de trabajo* o *workload* en inglés y, evidentemente, es fundamental la elección de esta carga de trabajo para dar validez a la prueba de desempeño. A fin de cuentas queremos una medida de desempeño para poder predecir cómo se comportará un sistema cuando sea usado en el mundo real, así que la carga de trabajo usada en la prueba debe ser diseñada de tal forma que permita estimar, tan exactamente como sea posible, el desempeño del equipo cuando esté en operación.

Existen varias opciones para diseñar la carga de trabajo:

1. Programas reales. Ejemplo: correr un conjunto de programas populares en dos computadoras diferentes, ejecutando las mismas tareas y comparar los tiempos de ejecución.

2. Kernels. Son trozos de programas reales que aislan el desempeño de ciertas características individuales. Por ejemplo, el famoso *Linpack* que prueba el desempeño en tareas relacionadas con álgebra de matrices o *Livermore loops* para medir la eficiencia en la ejecución de ciclos. Los kernels no son útiles por sí mismos.

3. Benchmarks de juguete. De 10 a 100 lineas de código que produce resultados conocidos. Por ejemplo la tradicional criba de Eratóstenes o Quicksort. Era usual hace algunas décadas reportar el desempeño de un sistema diciendo cuanto tiempo tardaba obteniendo los primeros n números primos o cuanto tardaba ordenando una secuencia de enteros.

4. Bechmarks sintéticos. No son programas útiles por si mismos, en general es un programa diseñado para tratar de simular la frecuencia promedio de las instrucciones que se ejecutan. Algunos fabricantes de computadoras comenzaron a hacer modificaciones a sus compiladores de tal forma que al compilar un bechmark de este tipo se hicieran optimizaciones no estándares para obtener mejores resultados en el benchmark falsificando los resultados[1]. Los más famosos son *Whetstone* y *Dhrystone*.

SPEC

Otra opción es hacer algo que mezcle varias de estas opciones en aras de construir una prueba que sea a la vez, apegada a la realidad, representativa y general. Esa es la opción que eligió SPEC (siglas de *Standard Performance Evaluation Corporation*) en su prueba comparativa que es, con mucho, la más científica y reproducible con la que contamos. La prueba general de SPEC evalúa el desempeño en dos rubros: aritmética entera y de punto flotante (conocidas como CINT y CFP, respectivamente), usando dos métricas diferentes: una basada en el tiempo de respuesta y la otra en el rendimiento (denominada con el sufijo *rate*). El conjunto de prueba de SPEC está constituido por varios programas de uso común como el compilador de C de GNU, algunos programas de compresión de datos como `gzip` o `bzip2` en la parte de aritmética entera o algunos programas de modelación científica en la parte de punto flotante. Todos ellos se ejecutan con una entrada determinada, usando banderas específicas para compilarlos.

[1]De hecho es posible optimizar tirando a la basura el 25% del código de Dhrystone, uno de los benchmarks más populares.

Los benchmarks de SPEC.

Las baterías de prueba (*benchmark suites*) son colecciones de programas o fragmentos de ellos, al estilo de los kernels. La intención es proporcionar resultados reproducibles y que realmente den una idea del desempeño del sistema en condiciones reales respecto a otros sistemas. Los programas elegidos deben ser entonces, estadísticamente significativos y las medidas obtenidas, más que dar un resultado que sea, *per se*, indicativo, proporcionan un criterio para comparar el sistema evaluado con otros. Las baterías más famosos y usuales son las de SPEC (*Systems Performance Evaluation Corporation*) una organización internacional no lucrativa entre cuyos miembros se cuentan las principales compañias fabricantes de hardware y software, universidades, centros de investigación y compañias consultoras.

SPEC CPU

SPEC ha elabrado diferentes baterías de pruebas para evaluar diferentes aspectos de los sistemas. Algunas resultan entonces mucho más útiles que otras en cierto contexto. Hay una batería para evaluar servidores web, otra para estaciones de trabajo de capacidades gráficas, otra para servidores de correo electrónico y otra, para evaluar el desempeño del sistema en general. Esta última es la prueba de mayor utilidad general y es, por tanto, la más usual. Se encuentra dividida en dos partes: la que evalúa el desempeño del sistema ejecutando tareas que involucran artitmética entera (las más usuales), llamada CINT2006 en su versión mas reciente (2011) y la que evalúa el desempeño del sistema utilizando aritmética de punto flotante, CFP2006. Cada una de estas partes evalúa el desempeño usando los dos tipos de métricas mencionados: tiempo de respuesta (SPECint2006 y SPECfp2006), llamada *speed* en la documentación de SPEC y rendimiento (SPECint_rate2006 y SPECfp_rate2006), llamada *throughput* por SPEC.

Desempeño base y máximo

Cualquier persona u organización puede ejecutar las baterías de SPEC y reportar los resultados obtenidos. Pero para evitar que, atendiendo a intereses comerciales, las compañías modifiquen la batería o manipulen el ambiente en el que se ejecuta la prueba para obtener resultados mejores, SPEC fija todos los parámetros de control de la prueba, por ejemplo las banderas del compilador y los datos de entrada a los programas de la prueba. Cualquier reporte de desempeño que se diga acorde con SPEC está obligado a contener los resultados atendiendo a las restricciones de SPEC, a lo que se le llama reporte de desempeño *base*. Adicionalmente, si la entidad que hace el reporte lo desea, puede relajar las restricciones y reportar lo que en terminología de SPEC se denomina máximo desempeño *peak performance*. La intención es dar a los fabricantes la libertad de decir qué tan bien se desempeñan sus sistemas pero sin pretender engañar al público. El reporte base, tanto

en punto flotante como en aritmética entera, es lo que el sistema en condiciones estándar puede lograr y el público podría, si quisiera, reproducir el experimento y obtener los mismos números. Es un punto de partida confiable. El reporte de desempeño máximo puede o no, ser tomado en cuenta por el público.

2.2 LEY DE AMDAHL

Toda decisión en el ámbito del diseño de computadoras (como en la mayoría de los ámbitos de la vida en general), está regida por una relación costo–beneficio. Así que necesariamente se debe poder evaluar, en particular, el beneficio que se obtendría de adoptar una decisión particular que mejora, de alguna manera, el estado actual de las cosas. En 1967 Gene Amdahl, entonces arquitecto de computadoras en IBM, formuló lo que hoy conocemos como la *ley de Amdahl* con el propósito de evaluar la mejora en el desempeño en computadoras de procesamiento paralelo. La regla encontrada por Amdahl es, sin embargo, aplicable de manera general para cuantificar el beneficio obtenido al introducir una cierta mejora en un sistema preexistente.

Beneficio de una mejora

Para evaluar el beneficio de introducir una mejora en un proceso es buena idea comparar, por ejemplo, el tiempo que tomaba la realización del proceso sin la mejora con el tiempo luego de introducirla. Conviene, entonces usar una expresión del tipo:

$$\frac{Tiempo_{sin}}{Tiempo_{con}} \tag{2.6}$$

Ganancia bruta y ganancia neta

Son más bien raras las situaciones en las que, en un proceso constituido de varias etapas, es posible introducir una mejora que tenga impacto directo en todas y cada una de ellas. Generalmente la mejora incide sobre una de ellas. Claro que esto tiene impacto sobre el desempeño total a la largo del proceso, pero sólo indirectamente. Convendría entonces utilizar una evaluación como la anterior, en diferentes niveles: para cuantificar el impacto directo de la mejora en la etapa para la que fue diseñada y después para evaluar su impacto indirecto a nivel global sobre todo el proceso. Llamaremos a estas *ganancia bruta* y *ganancia neta*, respectivamente.[2] Aclaremos mediante un ejemplo.

Imaginemos que debemos hacer, tan rápido como sea posible, un recorrido a campo traviesa que comprende dos diferentes tipos de terreno:

[2]Se ha utilizado el término *ganancia* para denotar lo que en la literatura en inglés suele llamarse *speedup*. Por una parte la traducción literal es "aceleración" lo que no resulta estrictamente correcto y además hace pensar que la ley de Amdahl es aplicable sólo a mejoras en el tiempo.

una parte en terreno irregular, pedregoso o cubierto de hierba entre el bosque y otra sobre nieve no muy firme. Si hacemos todo el recorrido usando unas botas de *trekking* nos desempeñaremos razonablemente bien en la parte boscosa del recorrido, pero tendremos problemas para andar sobre nieve blanda cuando se nos undan los pies y nuestro avance sea muy lento. Una mejora posible es llevar un par de raquetas de nieve o *snowshoes* que distribuyen el peso en una superficie mayor e impiden el hundimiento de los pies. Digamos que, sobre la nieve, cuando usamos las raquetas podemos ir al doble de velocidad que si no las usamos, es decir, la ganancia (g), *específicamente en el tramo en el que pueden usarse las raquetas*, es de 2. Formalmente diríamos:

$$g = \frac{Tiempo_{sin}}{Tiempo_{con}} = 2$$

Esta es una evaluación de la ganancia bruta obtenida al usar la mejora. Porque no es cierto que la mejora pueda usarse durante toda la ejecución de la tarea encomendada. Supongamos que, de los 30 km del recorrido, sólo 4 km son sobre nieve. Es decir el 13.33% del recorrido es sobre nieve y el restante 86.66% sobre terreno "normal". Podríamos decir, entonces, que la fracción F del recorrido en que es posible usar las raquetas es: 0.1333, mientras que no es posible usarlas en la fracción: $1 - F = 0.8666$.

Si suponemos que, usando sólo botas de *trekking*, el recorrido completo lo hacemos en un tiempo $T_0 = 12$ horas, podríamos decir que, usando las raquetas en el tramo en que es posible usarlas, el tiempo que usariamos en el recorrido completo debe ser:

$$T_m = T_0 \left(0.8666 + \frac{0.1333}{2}\right) = 0.9333\, T_0 \approx 11.2$$

lo que podemos leer "nos tardamos el tiempo usual cuando no podemos usar la mejora y la mitad cuando sí podemos usarla".

Ahora podemos hacer un cociente análogo al expresado en 2.6 pero entre el los tiempos totales y no sólo en la fracción en que se puede usar la mejora. Esta es, entonces, una evaluación cuantitativa de la ganancia neta:

$$G = \frac{T_0}{T_m} = \frac{12}{11.2} = 1.07$$

lo que significa que, en general, el uso de raquetas de nieve en el trayecto mejora el tiempo de recorrido en un 7%.

Generalizando lo hecho en el ejemplo. Si g es la ganancia bruta de una mejora introducida en una fracción F de un proceso P, T_0 el

tiempo total de ejecución de P sin usar la mejora, T_m el tiempo total de ejecución de P usando la mejora en la parte en que es posible; entonces:

$$T_m = T_0 \left[(1 - F) + \frac{F}{g} \right]$$

de donde, la *ganancia neta* obtenida por la introducción de la mejora es:

$$G = \frac{T_0}{T_m} = \frac{1}{(1 - F) + \frac{F}{g}} \tag{2.7}$$

Hacer mejor lo más común

Esta es una medida HIB y, dado que $T_m \leq T_0$, es una fracción impropia: la mejora es tanto más insignificante cuanto más cercano a uno es el cociente de 2.7. La regla coloquial que podemos extraer de la ley de Amdahl es: *hay que hacer mejor lo más común.* La ecuación 2.7 justamente nos permite cuantificar que tan mejor es lo "mejor" (factor g) y que tan común es lo "común" (factor F).

2.3 COSTO–BENEFICIO

En general, como hemos dicho, las decisiones que un arquitecto de hardware toma, están regidas por dos factores fundamentales: el desempeño del equipo y el costo de lo necesario para obtenerlo. Es conveniente, por cierto, aclarar que por "costo" no nos referimos ahora sólo al costo cuantificado en términos monetarios. A fin de cuentas se traducirá justamente en eso, pero no es ni necesario, ni recomendable, usar sólo el costo monetario como medida de lo necesario para obtener un cierto desempeño.

En este rubro es útil usar medidas como por ejemplo:

- (Desempeño SPEC) / precio.
- (Instrucciones / segundo) / (área en el chip).
- (Transacciones / segundo) / precio.
- (Frecuencia de reloj) / temperatura.
- (Tareas ejecutadas / segundo) / (Watt).

2.4 MALAS MÉTRICAS

Son populares algunas métricas para evaluar el desempeño y que, en general, no proporcionan una idea clara del comportamiento del sistema

en su uso cotidiano. Parecen prácticas e intuitivas, pero al usarlas arbitrariamente se corre el riesgo de terminar con una idea completamente equivocada de la realidad.

Un ejemplo clásico de esto es lo que suele llamarse MIPS (Millones de Instrucciones Por Segundo).

$$MIPS = \frac{\text{Número de instrucciones}}{\text{Tiempo de ejecución} \times 10^6} = \frac{F}{R \times 10^6}$$

donde R y F tienen el significado que les fue atribuido en la ecuación fundamental de desempeño (1.5), es decir ciclos por instrucción (*clocks per instruction*) y frecuencia de operación, respectivamente. Por ejemplo, en una máquina de 700 MHz cuya tasa de ejecución sea de 6 ciclos por instrucción:

$$MIPS = \frac{700,000,000}{6,000,000} = 116.6$$

La ventaja de usar esta medida de desempeño es que es fácil de entender, sobre todo por personas no muy avezadas en el área, es intuitiva.

Las desventajas son que:

- Depende del conjunto de instrucciones de la máquina. Una sola instrucción en una arquitectura particular, puede equivaler a toda una rutina en otra.
- No todas las instrucciones tardan lo mismo, así que dependiendo de las instrucciones que utilice un programa la medida en MIPS varía. Entonces ¿cuales instrucciones del conjunto total del procesador se deben usar? ¿las más usuales? ¿todas? ¿las más económicas en tiempo?

Incluso puede ocurrir que una máquina A con mejor desempeño real que otra B resulte con una medida en MIPS menor. Por ejemplo, si A tiene unidad de punto flotante (*floating point unit* o FPU) y tarda 8 ns. en una multiplicación mientras B tiene que emular la multiplicación a través de una rutina de 200 instrucciones tardándose 16 ns. resultará que $MIPS(B) > MIPS(A)$.

También ocurre que no se considera la "calidad" de las operaciones realizadas. 100 sumas de punto flotante tardan menos que 100 divisiones enteras, pero siguen siendo 100 instrucciones.

Aún el tiempo de ejecución está sujeto a diversas interpretaciones. Supongamos que se ejecuta un programa en una computadora, el tiempo de ejecución puede ser considerado como:

- el tiempo que tarda en completarse el programa incluyendo entrada/salida, accesos a memoria, etc.
- el tiempo efectivo en el que el procesador está ejecutando las instrucciones del programa.

En el primer caso estamos hablando del tiempo que le tomó al sistema completo (incluyendo dispositivos de e/s, sistema operativo, etc.) la ejecución del programa. Estamos midiendo el desempeño del sistema. En el segundo caso estamos midiendo el desempeño de la unidad central de proceso (CPU) únicamente.

La popularidad del sistema operativo Linux puso en uso una unidad similar llamada *BogoMIPS*. Linus Torvalds mismo, quien introdujo la medida en la versión 0.99.11 del kernel de Linux en 1993, dice hablerla nombrado BogoMIPS en referencia a la palabra "bogus", algo que es falso. En efecto, la medida en BogoMIPS de una computadora es sólo un truco técnico para poder sincronizar adecuadamente el hardware con el software, para determinar tiempos de espera adecuados para varificar el estado de agunos dispositivos, no para evaluar el desempeño del sistema.

En algunos lugares se utiliza también la frecuencia de operación del reloj del sistema como medida de desempeño. Claro, esto sólo es válido en las mismas circunstancias que los MIPS, cuando se comparan procesadores de la misma arquitectura de conjunto de instrucciones. Sólo si las instrucciones ejecutadas son exactamente las mismas, es valido usar el ritmo de trabajo como medida de la rapidez.

En los sistemas de alto desempeño, orientados a cómputo científico suele usarse una medida similar llamada MFLOPS (megaflops), la cantidad de millones de instrucciones de punto flotante ejecutadas por segundo. En este caso particular la medida no es del todo injusta: se comparan máquinas orientadas a las mismas tareas, cuantificando cuantas instrucciones, de un cierto tipo específico, se ejecutan por unidad de tiempo.