# Formulario de ISE

Ismael Sallami Moreno

Universidad de Granada
Doble Grado en Ingeniería Informática y ADE

## Introducción a la Ingeniería de Servidores

### Ganancia en velocidad de la máquina A respecto de B

$$S_B(A) = \frac{v_A}{v_B} = \frac{t_B}{t_A} \qquad \triangle v_{A,B}(\%) = \frac{v_A - v_B}{v_A} \times 100 = (S_A(B) - 1) \times 100$$

### Coste y relación prestaciones/coste

$$\frac{\frac{Prestaciones_A}{Coste_A}}{\frac{Prestaciones_B}{Coste_B}} = \frac{\frac{v_A}{Coste_A}}{\frac{v_B}{coste_B}} \quad \text{con } v_a \text{ (análogo para } v_B) \rightarrow v_A = \frac{1}{t_A} \quad \text{“mayor = mejor”}$$

### Ley de Amdahl

$$T_m = (1 - f) \times T_0 + \frac{f \times T_0}{k}$$

$$S \equiv S_{original}(mejorado) = \frac{v_m}{v_0} = \frac{t_0}{t_m} = \frac{T_0}{(1 - f) \times T_0 + \frac{f \times T_0}{k}}$$

$$\text{Ley de Amdahl} \rightarrow S = \frac{1}{1 - f + \frac{f}{k}}$$

Siendo:

- k: veces que se mejora.
- f: fracción donde se aplica la mejora.

Puede darse el caso de que tengamos varias mejoras:

$$S = \frac{1}{(1 - \sum_{i=1}^{n} f_i) + \sum_{i=1}^{n} \frac{f_i}{k_i}}$$

# Monitorización

$$Sobrecarga_{Recurso}(\%) = \frac{\text{Uso del recurso por parte del monitor}}{\text{Capacidad total del recurso ó periodo de activación}} \times 100$$

# Análisis Comparativo de Rendimiento

$$T_{ejec} = NI \times CPI \times T_{reloj} = \frac{NI \times CPI}{f_{reloj}}$$

$$MIPS = \frac{NI}{T_{ejec} \times 10^6} = \frac{f_{reloj}}{CPI \times 10^6}$$

$$MFLOPS = \frac{\text{Operaciones en coma flotante realizadas}}{T_{ejec} \times 10^6}$$

$$\text{índice SPEC} = \sqrt[n]{\frac{t_1^{REF}}{t_1} \times \frac{t_2^{REF}}{t_2} \times \ldots \times \frac{t_n^{REF}}{t_n}} = \sqrt[n]{\prod_{i=1}^{n} \frac{t_i^{REF}}{t_i}}$$

### Media Aritmética

$$\text{Media Aritmética} = \bar{t} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

### Media Aritmética Ponderada

$$\overline{t_W} = \sum_{k=1}^{n} w_K \times t_K \quad \text{donde} \quad \sum_{k=1}^{n} w_k = 1$$

$$w_K \equiv \frac{C}{t_K^{REF}} \Rightarrow C = \frac{1}{\sum_{k=1}^{n} \frac{1}{t_K^{REF}}} \quad \text{Siendo C una constante de normalización}$$

### Media Geométrica

$$\text{Media Geométrica} = \sqrt[n]{x_1 \times x_2 \times \ldots \times x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

### Propiedad del índice SPEC y comparación entre máquinas

**Propiedad:** Cuando las medidas son ganancias en velocidad (*speedups*) respecto a una máquina de referencia, el índice SPEC mantiene el mismo orden en las comparaciones independientemente de la máquina de referencia elegida (siempre que sea la misma en todos los casos).

$$SPEC(M) = \sqrt[n]{\frac{t_1^{REF}}{t_1^M} \times \frac{t_2^{REF}}{t_2^M} \times \cdots \times \frac{t_n^{REF}}{t_n^M}} = \sqrt[n]{\frac{t_1^{REF} \times t_2^{REF} \times \cdots \times t_n^{REF}}{t_1^M \times t_2^M \times \cdots \times t_n^M}}$$

**Comparación entre dos máquinas (MA y MB):**

$$\frac{SPEC(MA)}{SPEC(MB)} = \sqrt[n]{\frac{t_1^{MB} \times t_2^{MB} \times \cdots \times t_n^{MB}}{t_1^{MA} \times t_2^{MA} \times \cdots \times t_n^{MA}}}$$

**Orden de SPEC y medias geométricas:**

$$SPEC(MA) > SPEC(MB) \iff$$
$$\sqrt[n]{t_1^{MA} \times t_2^{MA} \times \cdots \times t_n^{MA}} < \sqrt[n]{t_1^{MB} \times t_2^{MB} \times \cdots \times t_n^{MB}}$$

*Es decir, la máquina con mayor SPEC es la que tiene menor media geométrica de los tiempos de ejecución.*

### Probabilidad: t Student

$$t_{exp} = \frac{\overline{d} - \overline{d}_{real}}{s/\sqrt{n}} \qquad \overline{d} = \frac{\sum_{i=1}^{n} d_i}{n} \qquad s = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \overline{d})^2}{n-1}}$$

$$\text{Error estándar} = \frac{s}{\sqrt{n}}$$

### P-value

Cuando el p-value $< \alpha$ (nivel de significación), se rechaza la hipótesis nula.
Cuando el p-value $> \alpha$ (nivel de significación), no se rechaza la hipótesis nula.
Donde la hipótesis nula es:

$$H_0 : \overline{d} = \overline{d}_{real} \quad \text{(no hay diferencia significativa entre las medias)}$$

( A y B rendimientos equivalentes)

**Intervalos de confianza**

Si nuestro intervalo no contiene el 0, rechazamos la hipótesis nula de que ambas máquinas tienen el mismo rendimiento al % del intervalo de confianza.

# Optimización del Rendimiento

*Todas las variables operacionales deducidas que se usan en este apartado son valores medios.* Además, suponemos que se tiene K estaciones de servicio.

- W: waiting time, tiempo de espera en la cola.
- S: service time, tiempo de servicio.
- R: response time, tiempo de respuesta.

$$R = W + S$$

**Variables y leyes operacionales:**

- $N_0$: número de trabajos en el servidor.
- $N_z$: número de clientes en reflexión (esperando a que los clientes vuelvan a introducirlos en el servidor).
- $T$: duración del periodo de media para el que se extrae el modelo.
- $A_i$: número de trabajos solicitados a la estación (**arrivals**).
- $B_i$: tiempo que el dispositivo ha estado en uso (**busy**).
- $C_i$: número de trabajos completados en el periodo (**completed**).
- $S_i$: tiempo medio de servicio (**service**). Se mide en $\frac{segundos}{trabajo}$ o bien en segundos.
- $W_i$: tiempo medio de espera en la cola (**waiting**). Se mide en segundos [/trabajo].
- $R_i$: tiempo medio de respuesta (**response**). Se mide en segundos [/trabajo].

$$S_i = \frac{B_i}{C_i} \qquad\qquad R_i = W_i + S_i$$

- $\lambda_i$: tasa media de llegada (**arrival rate**). Unidades $\frac{trabajos}{segundos}$.
- $X_i$: Productividad media (**throughput**). Unidades $\frac{trabajos}{segundos}$.

- $U_i$: Utilización media (**utilization**). Unidades %, pero no suele tener. Valor máx = $U_{i,max} = 1 \rightarrow 100\,\%$

$$U_i = \frac{B_i}{T} \qquad \lambda_i = \frac{A_i}{T} \qquad S_i = \frac{B_i}{C_i} \qquad X_i = \frac{C_i}{T}$$

**Haciendo referencia al número de trabajos en la estación de servicio:**

- $N_i$: Número de trabajos en la estación de servicio.
- $Q_i$: Número medio de trabajos en la cola de espera.
- $U_i$: Número medio de trabajos siendo servidos por el dispositivo.

$U_i = N_i - Q_i$ Coincide númericamente con la Utilización Media

**Variables operacionales de un servidor:**

- Básicas:
  - $A_0$: número de trabajos solicitados al servidor.
  - $C_0$: número de trabajos completados en el servidor.
- Deducidas:
  - $\lambda_0$: tasa media de llegada al servidor.
  - $X_0$: Productividad media del servidor.
  - $N_0$: Número medio de trabajos en el servidor.
  - $R_0$: Tiempo medio de respuesta del servidor.

$$\lambda_0 = \frac{A_0}{T} \qquad X_i = \frac{C_0}{T}$$

**Razón de visita y demanda de servicio:**

- Razón media de visita al servidor: $V_i$ (**visit ratio**): Proporción entre el número de trabajos completados por el servidor y el número de trabajos completados por la estación de servicio i-ésima.
- Demanda de servicio: $D_i$ (**service demand**): Cantidad de tiempo que, por término medio, el dispositivo de la estación de servicio i-ésima le ha dedicado a cada trabajo que abandona el servidor.

$$V_i = \frac{C_i}{C_0} \qquad D_i = \frac{B_i}{C_0} = V_i \times S_i$$

**Ley de Utilización**

$$\forall i = 1, \ldots, K \quad U_i = X_i \times S_i \overset{\text{equilibrio de flujo}}{=} \lambda_i \times S_i$$

**Ley del fujo forzado**

$$\forall i = 1, \ldots, K \quad X_i = X_0 \times V_i \overset{\text{equilibrio de flujo}}{=} \lambda_0 \times V_i = \lambda_i$$

**Relación Utilización-demanda de servicio**

$$\forall i = 1, \ldots, K \quad U_i = X_0 \times D_i \overset{\text{equilibrio de flujo}}{=} \lambda_0 \times D_i$$

**Ley de Little**

- Aplicada a un servidor:

$$N_0 = \lambda_0 \times R_0 = X_0 \times R_0$$

- Aplicada a toda una estación de servicio:

$$N_i = \lambda_i \times R_i = X_i \times R_i$$

- Aplicada a una cola de una estación de servicio:

$$Q_i = \lambda_i \times W_i = X_i \times W_i$$

**Ley general del tiempo de respuesta**

$$R_0 = \sum_{i=1}^{K} V_i \times R_i$$

**Ley del tiempo de respuesta interactivo**

$$R_0 = \frac{N_T}{X_0} - Z$$

- $Z$: tiempo de reflexión, tiempo que requiere el cliente antes de volver a lanzar una petición al servidor tras la respuesta de este.

**Identificación del cuello de botella**

- *b (bottleneck): índice del dispositivo cuello de botella*

$$D_b = \max_{i=1,\ldots,K} D_i = V_b \times S_b$$
$$U_b = \max_{i=1,\ldots,K} U_i = X_0 \times D_b$$

**Saturación del servidor**

- El saturación, el cuello de botella está al máximo de su productividad.

$$1 = U_b = X_b \times S_b \Rightarrow X_b = \frac{1}{S_b}$$

**Límites optimistas: redes abiertas**

$$R_0 \underset{\text{optimista} = \min}{\Rightarrow} R_0^{min} = \sum_{i=1}^{K} V_i \times S_i = \sum_{i=1}^{K} D_i \equiv D$$

$$\text{Si } U_b = 1 \Rightarrow X_0^{max} = \frac{1}{D_b}$$

Cuando $\lambda_0 \leq X_0^{max}$ estamos en equilibrio de flujo

### Límites optimistas: redes cerradas

- Valores de carga altos

$$\text{Cuando esta cerca de la saturación: Si } U_b = 1 \Rightarrow X_0^{max} = \frac{1}{D_b}$$

$$\text{Valor optimista de respuesta medio: } R_0 = \left(\frac{N_T}{X_0^{max}}\right) - Z = D_b \times N_T - Z$$

- Valores de carga bajos

$$\text{Valor optimista de respuesta medio: } R_0^{min} = \sum_{i=1}^{K} V_i \times S_i = \sum_{i=1}^{K} D_i \equiv D$$

$$\text{Valor optimista de productividad media: } X_0 = \frac{N_T}{R_0^{min} + Z} = \frac{N_T}{D + Z}$$

### Punto teórico de saturación

$$D = D_b \times N_T^* - Z \Rightarrow N_T^* = \frac{D + Z}{D_b}$$