Evaluación y Explotación de Sistemas Informáticos

Introducción a la evaluación de rendimiento

- Técnicas de evaluación del rendimiento
- Sistemas de Referencia
- Medidas de Rendimiento
- Rendimiento en Entornos Web
- Métricas de Rendimiento

Técnicas de Evaluación del Rendimiento

- Medición
  - Índices
- Simulación
  - Carga de trabajo
- Modelado Analítico
  - Análisis Operacional
  - Teoría de Colas

Técnicas de Evaluación del Rendimiento

	Técnica de Solución		
Características	Modelado Analítico	Simulación	Medición
Flexibilidad	Alto	Alto	Bajo
Costo	Bajo	Medio	Alto
Credibilidad	Bajo	Medio	Alto
Exactitud	Bajo	Medio	Alto

Una comparación de las técnicas de análisis de rendimiento

Sistemas de Referencia - Clásico

- Sistemas de procesamiento en Lotes o Batch
- Sistemas interactivos o por demanda
- Sistemas transaccionales

Sistemas de Referencia

- Sistemas centralizados (batch e interactivos)
- Sistemas de Bases de Datos
- Sistemas en Red
- Sistemas Multiprocesador

Esquema clásico de evaluación de prestaciones

- Definir objetivos
- Especificar el conjunto de datos del sistema y los datos de la carga de trabajo
- Definir los índices de rendimiento y variables a ser evaluadas
- Repetir hasta que se alcancen los objetivos definidos
  - Medición del sistema y la carga de trabajo
  - Recolección de datos e interpretación de resultados
  - Si no se alcanzan los objetivos: *modificar el sistema*
- Planear verificaciones periódicas

Índices de rendimiento: Internos

- Factor de Utilización
  - Utilización de CPU
- Solapamiento de actividades
  - Solapamiento de componentes
- Overhead
- Factor de carga de multiprogramación
- Factor de ganancia de multiprogramación
- *Paging Rate*, Tasa de error de paginado, frecuencia de fallo de página
- Frecuencia de swapping
- Tiempo de reacción

Índices de rendimiento: Externos

- Tiempo de espera
- Tiempo de respuesta
- *Throughput*, *Productividad*, Capacidad de tratamiento/procesamiento de datos
- Capacidad

- Disponibilidad
- Confiabilidad
- Seguridad

- Facilidad de mantenimiento
- Tolerancia a fallos

Volumen de negocios en e-commerce

Type of Business	1997	2001 (forecast)
Business to Business	8.000	183.000
Travel	0.654	7.400
Financial Services	1.200	5.000
PC Hardware & Software	0.863	3.800
Entertainment	0.298	2.700
Ticket Event Sales	0.079	2.000
Books & Music	0.156	1.100
Apparel & Footware	0.092	0.514
Total	11.342	205.514

(Business Week, June 22, 1998; numbers in $US billion)

AMAZON.com (2005) : $8.7 billion

Volumen de negocios en e-commerce

- El comercio electrónico en España alcanzó los 3.740 millones de euros en 2007 *
  - Un 52% más que en 2006
- El comercio electrónico en España alcanzó los 5.751,7 millones de euros en 2009 **
  - Un 11% más que en 2008
- El 53,7%% de las compras correspondieron a portales españoles: 3088.7 millones de euros.
- El último trimestre del 2009 el comercio online español llegó a mover 1.574,2 millones de euros, y se incrementaron las operaciones de compra-venta en un 28,3%
- Por sectores (2009):
  - Turismo: 514,6 millones, 32%
  - Marketing directo: 154,9 millones, 9,8%
  - Juegos de azar y apuestas: 95,8 millones, 6%.

* Fuente: Comisión del Mercado de las Telecomunicaciones, 2008
** Fuente: Comisión del Mercado de las Telecomunicaciones, 2010

Rendimiento en entornos Web

- Evolución muy rápida
- Estándares creados por la industria
  - Aceptados/Rechazados en algunos casos por el mercado
- Gran cantidad de usuarios
- Problemas de tráfico
  - Congestionamiento
- Tiempo de respuesta : Crucial
- Factibilidad comercial
- ... pero no solo cuenta el nivel de respuesta
  - USABILIDAD

Rendimiento en entornos Web (cont.)

- Objetivo:
  - Proporcionar la calidad de servicio que requiere el usuario
- Problemas
  - Dimensionamiento adecuado de la infraestructura.
  - Seguimiento de la intensidad de la carga de trabajo
  - Detección de cuellos de botella
  - Predicción de las capacidades futuras
  - Determinar la forma de actualización más rentable para:
    - Solventar los problemas de rendimiento
    - Dar respuesta a los incrementos de carga de trabajo

Rendimiento en entornos Web (cont.)

- Características
  - Diferencias importantes entre las cargas medias y los picos de carga
  - Migración de aplicaciones legadas a entornos Web
  - Generación dinámica y compleja de páginas web
  - Contenidos personalizados
  - Integración con bases de datos y sistemas de planificación y seguimiento
  - Requisitos de calidad del servicio (QoS) de alto rendimiento y disponibilidad
  - Desarrollo de aplicaciones críticas
  - Caso particular de aplicaciones Cliente/Servidor

Concepto de planificación de capacidad

- Es el proceso de predecir los niveles de carga en los que el sistema se saturará y la determinación de la solución más eficaz en coste que permita retrasar la saturación del sistema lo más posible.
- La predicción ha de considerar la evolución de la carga de trabajo y los niveles de servicio deseados.
- No hay que esperar a que los problemas sucedan, hay que anticiparse a ellos.
- La calidad de servicio de un sistema Web o de comercio electrónico tiene una correlación elevada con el coste de la infraestructura necesaria para proporcionar el servicio.
- Es necesaria la planificación ya que la resolución de los problemas de rendimiento no es instantánea.
- La mejora del rendimiento no siempre tiene que ser alcanzada mediante la adquisición de nuevos componentes.

Entornos Web: Niveles de Servicio
Service-Level Agreements (SLA)

- Principalmente orientados al usuario
- Límite superior del tiempo de respuesta
  - El tiempo de respuesta a la transacción debe ser menor a 6 segundos.
- Productividad mínima por el servidor Web.
  - Procesar al menos 100 peticiones web por segundo.
- Disponibilidad mínima del sitio web.
  - El sitio ha de estar operativo al 99.9% del tiempo
- Porcentaje de transacciones que han de tener un tiempo de respuesta menor o igual que un cierto valor.
  - El 95% de las transacciones han de tener un tiempo de respuesta inferior a 2 segundos.

Capacidad Adecuada

- Para definir la capacidad adecuada de un sistema basado en Web se toma en cuenta:
  - Compromisos de Nivel de Servicio (Service Level Agreements - SLA)
  - Estándares y Tecnologías Adoptadas
  - Restricciones de Costo
- El objetivo es proporcionar una buena calidad de servicio (capacidad adecuada para proporcionarla)
- Se requiere de una metodología de *planeación de capacidad*

Métricas de Rendimiento

- Características *básicas* de un sistema computacional que se miden usualmente:
  - Una *cuenta* del número de veces que un evento ocurre
  - La *duración* de algún intervalo de tiempo
  - El *tamaño* de algún parametro
  - Por ejemplo:
  - El número de veces que un procesador inicia un requerimiento de entrada salida.
  - Cuanto dura la petición en ser completada.
  - La cantidad (número de bits) transferidos.
- A partir de estos valores podemos derivar el *valor* actual que deseamos usar para describir el rendimiento del sistema evaluado. Este valor es llamado como *métrica de rendimiento*.
- Podemos usar estos valores directamente o …

Métricas de Rendimiento

- Los valores básicos pueden normalizarse.
- Los contadores pueden relacionarse con el tiempo del intervalo de conteo para obtener una métrica de velocidad (operaciones por segundo)
- Las métricas normalizadas se denominan *tasas/índices de medida* o *productividad* (*rate metric* ó *throughput*)

Características de una buena métrica de rendimiento

- Linealidad
- Confiabilidad
- Repetible
- Facilidad de Medida
- Consistencia
  - MIPS y MFLOPS no son consistentes entre si
- Independencia

Métricas de Rendimiento de Procesador y Sistemas (*Centradas en el procesador*)

- Índice del reloj, ciclos del reloj (clock rate)
- MIPS
- MFLOPS
- SPEC
  - Standard Performance Evaluation Corporation (grupo de fabricantes)
  - Consta de los siguientes pasos
    - Medir el tiempo requerido para ejecutar cada programa en conjunto en el sistema evaluado
    - Dividir el tiempo medido por cada programa en el primer paso por el tiempo requerido para ejecutar el programa en una máquina estándar (Para normalizar el tiempo de ejecución)
    - Promediar estos valores usando la media geométrica para producir un valor simple (la métrica de desempeño).
  - Desventaja: No es líneal

Métricas de Rendimiento de Procesador y Sistemas (*Centradas en el procesador*)

- QUIPS (Quality Improvement per Second)
  - Benchmark Hint
  - Mide la calidad de la mejora en lugar de la cantidad
- Tiempo de ejecución
  - Distorsión en sistemas multitarea y por influencia del S.O.

Otras métricas de rendimiento: Variables internas

- Factor de Utilización
- Solapamiento de actividades
- Overhead
- Factor de carga de multiprogramación
- Factor de ganancia de multiprogramación
- *Paging Rate*, Tasa de error de paginado, Frecuencia de fallo de página
- Frecuencia de swapping
- Tiempo de reacción

Otras métricas de rendimiento: Variables Externas

- Tiempo de espera
- Tiempo de respuesta
- *Throughput*, *Productividad*, Capacidad de tratamiento/procesamiento de datos
- Capacidad

Otras Métricas de rendimiento: Otras magnitudes

- Disponibilidad
- Confiabilidad
- Seguridad
- Facilidad de mantenimiento
- Tolerancia a fallos

Turnaround Time

- Definido como el intervalo de tiempo entre el instante en el que un programa es enviado a un sistema de procesamiento batch y el instante en el que finaliza la ejecución.
- Proporciona información acerca de la eficiencia del procesamiento.
- Si el tiempo de turnaround de un programa es

  $$T = P - R$$

  donde R es el momento en el que comienza la lectura de las instrucciones del programa y P en el que finaliza la impresión de los resultados.
- El tiempo medio de turnaround (T_m) para n programas es:

  $$T_m = \frac{1}{n}\sum_{i=1}^{n} T_i = \frac{1}{n}\sum_{i=1}^{n} (P_i - R_i)$$

  Esto puede conducir a conclusiones imprecisas acerca de la eficiencia del procesamiento si n es pequeño.

Tunaround Time (cont.)

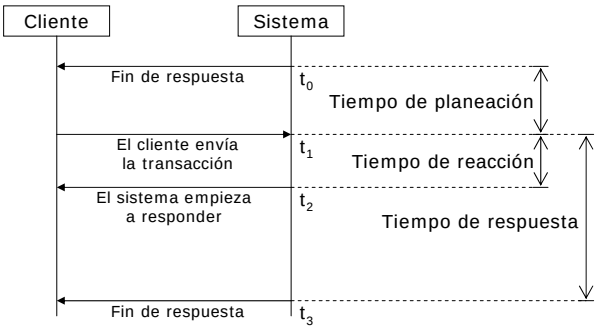
- El tiempo de tournaround ponderado T_w se define como el ratio entre el tiempo de turnaround T y el tiempo de procesamiento del programa T_p.

  $$T_w = T/T_p$$
- El tiempo de turnaround ponderado medio T_{wm} se define como la media aritmética de la sumatoria tiempos de turnaround ponderados.

  $$T_{wm} = \frac{1}{n}\sum_{i=1}^{n} T_{wi}$$
- Ambas métricas se ven afectadas por las políticas de gestión de recursos implementadas en el sistema y por las características de la carga de trabajo.

Definición de Tiempos de: Planeación, Reacción y Respuesta

- Tiempo de planeación, (*Think time*), el usuario/cliente elabora el requerimiento.
- Tiempo de reacción (*Reaction time*), el usuario/cliente envia el requerimiento, el sistema empieza a responder.
- Tiempo de respuesta (*Response time*), el sistema responde al requerimiento del usuario.



- Por simplicidad no se considera el retardo que puede presentarse en la comunicación entre Sistema y Cliente.

Tiempo de respuesta

- El tiempo de respuesta es fuertemente dependiente del tipo de comando que ejecuta el sistema.
  - Comandos ligeros: precisan de menos de un quantum de tiempo de la CPU. (comandos de edición, petición de información, etc.)
  - Comandos pesados: precisan de más de un quantum de tiempo de CPU para su ejecución. (compilación, ejecución, clasificación, etc.)
- El tiempo medio de respuesta R_m no proporciona la información completa acerca del rendimiento de un sistema interactivo, hay que tener en cuenta su variabilidad.
- Puede modelarse a través de la distribución Gamma

Distribución Gamma

- Es la distribución más general para modelar el tiempo de respuesta.
- Los parámetros k y θ determinan la forma y la escala de la distribución respectivamente.
- Cuando $\alpha=1$ y $\beta=1/\lambda$ entonces se reduce a la distribución exponencial.

$$f(x,k,\theta) = x^{k-1}\frac{\theta^{-k}e^{-x/\theta}}{\Gamma(k)} \; ; x>0$$

Reemplazando $t = x,\ \alpha = k$ y $\beta = 1/\theta$

$$g(t,\alpha,\beta) = t^{\alpha-1}\frac{\beta^{\alpha}e^{-t\beta}}{\Gamma(\alpha)} \; ; \; t>0$$

Productividad (Throughput)

- Cantidad de trabajo útil ejecutado por unidad de tiempo en un entorno de carga determinado.
- Tasa (peticiones por unidad de tiempo) a la que el sistema sirve las peticiones.
- Tipos de productividad
  - Sistema batch - trabajos por segundo
  - Sistema Interactivo – peticiones por segundo
  - CPU's – MIPS o MFLOPS
  - Redes – paquetes por segundo o bits por segundo
  - Sistema transaccional – transacciones por segundo.

Productividad (*Throughput*)

- Una definición general del *throughput* es

  $$X = N_p / T_{tot}$$

  donde Np es el número de programas procesados en en el intervalo de medida $Ttot$. X proporciona un índice de la velocidad de ejecución para el conjunto de Np programas (carga de trabajo).

- Factores que influyen en el *throughput*
  - Las características de la carga bajo las cuales se evalúa
  - La velocidad de los componentes hardware y software
  - El grado de multiprogramación permitido por el hardware
  - La configuración del sistema
  - El algoritmo de reserva de recursos utilizado.

Otras magnitudes

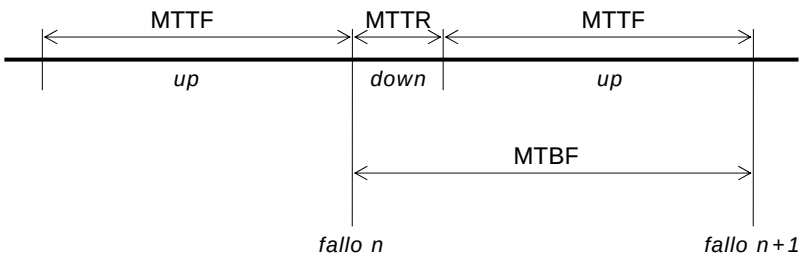
- Eficiencia
  - La proporción entre la *capacidad utilizable* y la *capacidad nominal*
- Utilización de un recurso
  - Es la fracción de tiempo que el recurso está ocupado sirviendo peticiones.
  - Es la proporción entre el tiempo ocupado y el tiempo total transcurrido en un periodo dado.

    Tiempo_ocupado / Tiempo_total
  - Tiempo en espera (*idle time*), periodo en que el recurso no es usado
- Confiabilidad
  - Se mide como la probabilidad de errores o el tiempo medio entre errores
    - MTTF (*mean time to failure*) Tiempo medio para un fallo
    - MTBF (*mean time between failures*) Tiempo medio entre fallos
    - MTTT (*mean time to transition*) Tiempo medio para la transición
    - MTTR – (*mean time to recover*) Tiempo medio para recuperarse

      MTBF = MTTF + MTTR
- Disponibilidad
  - Es la proporción del tiempo total durante el que el sistema está a disposición de los usuarios.

    Disp= MTTF / (MTTF+MTTR) = MTTF / MTBF

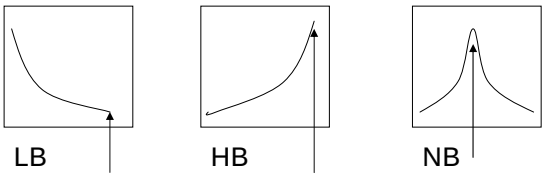
Otras magnitudes: MTTF, MTTR, MTTB



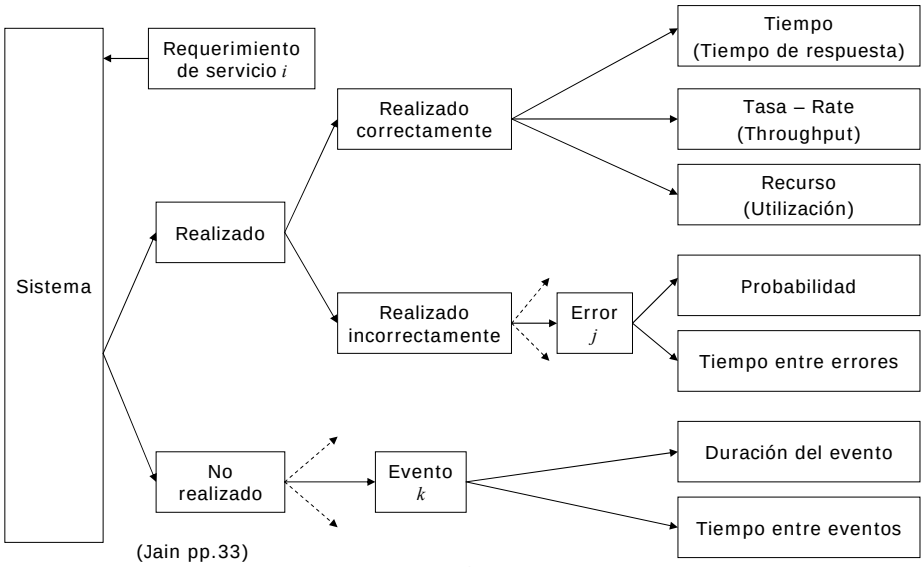
[Menascé & Almeida 2002, pp. 420]

Categorías de métricas de evaluación de rendimiento

- Las métricas de rendimiento pueden categorizarse en tres clases:
  - LB (lower is better) el valor más bajo es el mejor. Por ejemplo el tiempo de respuesta.
  - HB (higher is better) el valor más alto es el mejor. Por ejemplo el throughput (productividad)
  - NB (nominal is best) el valor nominal es el mejor como por ejemplo la utilización.



Selección de métricas de rendimiento



(Jain pp.33)