

Taller de Administración de Servidores Linux CC5308

Clase 33: 02 de junio de 2011

Desempeño y Aprovisionamiento

Luis León Cárdenas Graide
lcardena [at] nic . cl

Copyright © 2011
Creative Commons 3.0-cl by-nc-sa

Benchmarking

- Niveles de Carga: Cuantificar (Monitoreo)
- Stress: Diseñar Tests
- Cuellos de Botella: Identificarlos
- Saturación controlada
 - ¿Saturación del saturador?
- Análisis de fallas
 - Pre-saturación, Durante, Después
- Resiliencia del sistema
 - Recuperación post-falla: Técnica y Procesos
 - Volver a estresar: ¿mantiene calidad?

Recursos Agotables

- Dinero, Tiempo (dinero), H/H (dinero y tiempo)
- Hardware
 - CPU: % utilización
 - RAM: % libre, velocidad, trashing
 - Disco: % libre, throughput, latencia, fragmentación
 - Red: ancho de banda, latencia, pérdida
- Software
 - Descriptores: Procesos, File Descriptors, Sockets, Sistema de Archivos
 - Contadores y Acumuladores
- Eventuales fallas anteriores a agotamiento total

¿Qué es "carga"?

- Contención: Requerimientos listos para ser procesados a la espera de liberación de recursos
- Ej: "Puntos de Carga"
 - #Procesos en cola Ready del SO esperando por CPU
- Cola de Requerimientos
 - $\Delta \text{Contención} / \Delta T = \text{Tasa Ingreso} - \text{Tasa Despacho}$
- Tasa de Ingreso también tiene límites
 - Requerimientos 1 a 1
 - Concurrencia
- Latencia = $T(\text{despacho}) - T(\text{encolamiento})$

¿Qué Requerimientos?

- Operaciones Técnicas: Hardware
 - Red: bits por segundo (bps), paquetes por segundo
 - Disco: I/O, operaciones de read o write por segundo
- Transacciones: Base de datos
 - select/insert/update/delete por segundo (Locks)
- Servicios: a Clientes y entre otros Servicios
 - Conexiones: handshakes por segundo
 - Complejidad dependiendo de capa
 - Clientes: light vs heavy users
- Procesos de Negocio
 - Ventas, Cobros, Atenciones, Pagos

Contención

- Comportamiento del sistema según variaciones
 - Ingreso de Requerimientos
 - Stock de Requerimientos pendientes
 - Despacho de Requerimientos
 - Tiempo total de Ejecución
 - Fresh start vs Long running
 - Carga y recarga
 - Efectos de caché y flush
 - Inconsistencias por saturación anterior
- Efectos feedback
 - Latencia
 - Pérdida: Retry, Inconsistencia

Test de Stress

- Tasa Encolamiento
 - Constante: Instantánea, Promedio
 - Creciente: Lineal, Exponencial
 - Peaks, Mesetas, Valles, Intermittencia
- Contención Objetivo: uso intensivo de un recurso
 - Operaciones livianas vs pesadas
 - Distintos impactos sobre distintos recursos
- Estresores: Despacho posible vs Encolamiento
 - Simple vs Múltiple
 - Serial vs Concurrente
- Output: Límite de saturación del sistema

Proyección de Requerimientos

- Reactivo vs Proactivo: SLA y BCP
- Upgrade: Proyecto
 - Tradeoff Plazos y Costos
 - Definir inicio, término y presupuesto
 - Medición, Adquisición, Instalación
 - Ineficiencia y Costo de Oportunidad
 - Subdimensionamiento: merma utilidades
 - Ventas perdidas, Costos hundidos
 - Sobredimensionamiento: capital ocioso
 - Retorno económico considerando costos
 - Precio hardware, Horas hombre, Consumo energético, Disipación, dependencias físicas
- Objetivos en Throughput / Plazo / Costo: ¡Medir!

Actividad Personal

- ¿Cuál es la capacidad de tu red?
- Ancho de banda, Paquetes por segundo, Latencia, Jitter
- ¿Contra quién? ¿De qué forma? ¿Cuál es el Cuello de Botella? ¿Seguro?
- Herramientas: hping, netcat
- Distintas capas: Enlace, Red, Transporte, Aplicación