Tema 3

Remember…
queueing theory

El escenario

- Server and queue
- Parámetros
  - Proceso de llegadas: tiempo entre llegadas, llegadas acumuladas
  - Proceso de servicio: tiempo de servicio

- Tiempo de servicio, numero de clientes en el sistema
- Factor de utilizacion

Modelos

- Formula de Little
- Procesos de Markov
- Sistemas de colas en equilibrio

Sistemas de colas

- M/M/1
  - M/D/1 M/G/1
- M/M/n
- M/M/1/n
- M/M/n/m
- ...

Sistema M/M/1

- Resumiendo:
  - Número medio de clientes en el sistema (cola+servidor)

$$N = \frac{\rho}{1-\rho}$$

  - Número medio de cliente en la cola

$$N_q = \frac{\rho^2}{1-\rho}$$

  - Número medio de clientes en el servidor

$$N_s = \rho$$

  - Tiempo medio de espera en el sistema (cola+servidor)

$$S = \frac{1/\mu}{1-\rho}$$

  - Tiempo medio de espera en la cola

$$W = \frac{\rho}{\mu(1-\rho)}$$

  - Tiempo medio de espera en el servidor es simplemente 1/μ

Ejemplo: paquetes

- Un router recibe paquetes desde diferentes redes
- Los paquetes tienen un tamaño medio de 400Bytes
- Llegando por 10 puertos de 1Gbps, por cada puerto llegan unos 50Mbps

- Es esto mucho tráfico?
- Cuantos paquetes habrá en media en la cola del puerto de 1Gbps de salida?
- Cuanto tiempo extra tardan en salir los paquetes por culpa del exceso de trafico? Cuanto mas que si la red estuviera descargada?
- Si la cola tiene espacio para 10 paquetes cual es la probabilidad de descarte de paquetes por esta causa

Ejemplo: servidores

- Un servidor web sirve las peticiones de forma iterativa con un solo hilo (no puede atender a dos a la vez)
- El tiempo principal que tarda en atender una petición viene de una consulta a la base de datos que tiene un tiempo medio de 30ms
- Si tenemos en media 10peticiones al servidor por segundo…
- Cual es el tiempo medio que ve el cliente para atender su operación?

- Cual es la probabilidad de que vea mas de 100ms?
- Si las peticiones que no se atienden inmediatamente no se pueden almacenar cual es la probabilidad de que una petición no sea atendida
- Suponiendo que la base de datos puede atender mas peticiones simultáneas y seguir tardando 30ms…
- Como cambiarían los resultados si el servidor tuviera N hilos que atiendan peticiones en paralelo