SISTEMAS M/M/2 CON SERVIDORES DE DISTINTAS VELOCIDADES.

INTRODUCCION:

Supongamos que en una M/M/1 esta en uso (o en funciones) un servidor que pareciera estar resultando insuficiente. Se observa que las colas se alargan y que los tiempos de permanencia en el sistema de los clientes parecen haber aumentado considerablemente.

Eso puede deberse a una cuestión circunstancial y ser una situación transitoria o puede ser que definitivamente la tasa de arribos haya aumentado "para siempre", es decir que las solicitudes hayan aumentado de manera permanente. Para determinar si es algo transitorio o si llego para quedarse hay que probarlo matemáticamente.

Para eso se considera una muestra suficientemente grande, de un tiempo considerado adecuado, por ejemplo 1 mes o lo que el analista a cargo determine, y se calcula el NUMERO MEDIO DE CLIENTES EN EL SISTEMA (**N**) aplicando el teorema de LITTLE.

Luego se considera:

Si .... $N < \lambda/(\mu_1 - \lambda)$ entonces el servidor todavía es adecuado, NO MODIFICAR

Si .... $N > \lambda/(\mu_1 - \lambda)$ entonces hay que aumentar la tasa de servicio.

Si .... $N = \lambda/(\mu_1 - \lambda)$ entones es indistinto, peroes evidente que la tasa de arribos esta aumentando y seguramente en forma inmediata la inecuación se cumpla por ">"; el sentido común indica que también hay que aumentar la tasa de servicio.

Nunca hay que olvidarse que la TEORIA DE COLAS ES UNA PARTE DE INVESTIGACION OPERATIVA Y QUE **LA INVESTIGACION OPERATIVA SE HA DEFINIDO COMO SENTIDO COMUN CUANTIFICADO.**

Asumiendo que se realiza el análisis precedente y se determina que hay que realizar algún cambio porque el servidor resulta insuficiente, hay tres opciones:

1.- Cambiar el servidor que se esta utilizando por otro mas rápido (nunca falla)

2.- Agregar otro servidor de igual velocidad o mas rápido (también sirve siempre)

3.- Supuesto que ninguna de las dos anteriores fuese posible, por ejemplo por razones de costos,y solo se dispusiese de un servidor mas lento (el que se dispone sin tener que ir a comprarlo), NO SIEMPRE CONVIENE AGREGARLO CONFIGURANDO UNA M/M/2. En algunos casos puede mejorar, aunque sea de manera temporaria, la situación y en otros casos no solo no mejora sino que empeora.

Entre el caso 1 y 2, si se logra la misma tasa de servicio del sistema, desde el punto de vista de la velocidad y sin tomar en cuenta que en el caso "2" si un servidor se cae le

queda uno funcionando, si el ρ de la M/M/1 es suficientemente alto como para mantener a los dos servidores de la M/M/2 ocupados "casi" todo el tiempo, entonces es indistinto porque cualquiera de los sistemas trabaja a tasa 2μ. Si, en cambio, el ρ de la M/M/1 no suficiente para mantener los dos servidores de la M/M/2 ocupados todo el tiempo (o casi todo el tiempo) puede ocurrir que muchas veces quede un servidor ocioso, y en ese supuesto la velocidad de procesamiento de todo el sistema cae a la tasa μ , es decir, en los periodos que trabaja uno solo de los dos servidores de la M/M/2, el sistema cae a la velocidad de un solo servidor, o sea, trabajaría a la mitad de la velocidad del sistema, entonces si se diese ese supuesto, el caso 1 es mas conveniente porque siempre procesa a tasa $\mu_S = 2\mu$ en cambio en la M/M/2, cuando un solo servidor queda trabajando, el sistema, durante esos periodos, procesa a la mitad de la velocidad, y si esa situación se repite con mucha frecuencia, la velocidad promedio de todo el sistema baja y el servidor agregado, en vez de mejorar la situación, la empeora.

Para el caso de no poder implementar alguna de las soluciones de los casos 1 y 2, queda por ver si es posible configurar una M/M/2 con un servidor mas lento que el que esta en uso en la M/M/1 (la alternativa "3"). Para eso hay que considerar el ρ_c (rho critico), que es el valor de ρ de la M/M/1 a partir del cual conviene agregar el servidor lento. El análisis que se hace para determinar si conviene o si no conviene agregar el servidor lento configurando una M/M/2 es el siguiente:

Si $\rho_{M/M/1} > \rho_c$ entonces conviene agregar el servidor lento

Si $\rho_{M/M/1} < \rho_c$ NO conviene agregar el servidor lento

Si $\rho_{M/M/1} = \rho_c$ es una situación de indiferencia, es decir, no mejora ni empeora.

El único caso en que conviene agregarlo, hasta que se pueda conseguir el servidor que realmente se necesita, es el primero, en el cual la desigualdad se cumple por ">".

CONFIGURACIONES:

Para el caso de tener que configurar una M/M/2 agregando un servidor mas lento que el de la M/M/1, existen dos configuraciones posibles: SIN SELECCIÓN Y CON SELECCIÓN DE SERVIDOR.

1er. CASO: sin selección de servidor:

Su dinámica es la siguiente: Si llega un cliente y los dos servidores están ocupados, el cliente espera en cola. Si al producirse la llegada de un cliente un servidor esta ocupado y el otro desocupado, entonces el cliente va al servidor que esta ocioso.

Cuando llega un cliente y los dos servidores están desocupados, el cliente tiene la misma probabilidad de ir a cualquiera de los dos servidores. Con el sistema en régimen, en un numero suficiente grande de ocurrencias, la mitad de los clientes que llegan en esa situación va a ir al servidor rápido y la otra mitad va a ir al servidor lento.

(Aclaracion: es en un numero suficientemente grande de ocurrencias porque en un numero grande de repeticiones se cumple, de manera muy aproximada, la verdadera probabilidad del suceso – definición frecuencial de probabilidades-)

2do. CASO: Con selección de servidor

Cuando llega un cliente al sistema y los dos servidores están ocupados, el cliente se pone en cola. Si llega un cliente y un servidor esta ocupado y el otro esta libre, el cliente siempre va al que esta libre, sin importar si es el rápido o el lento. Hasta ahí los dos casos son iguales. La diferencia se produce si un cliente llega y los dos servidores están libres; en ese caso, el cliente siempre va al servidor rápido, nunca al lento. De alguna manera se lo direcciona, ya sea porque hay un algoritmo que lo hace, o alguien que le indica adonde ir, o por lo que sea, el cliente siempre va al servidor rápido.

DIAGRAMAS DE ESTADO

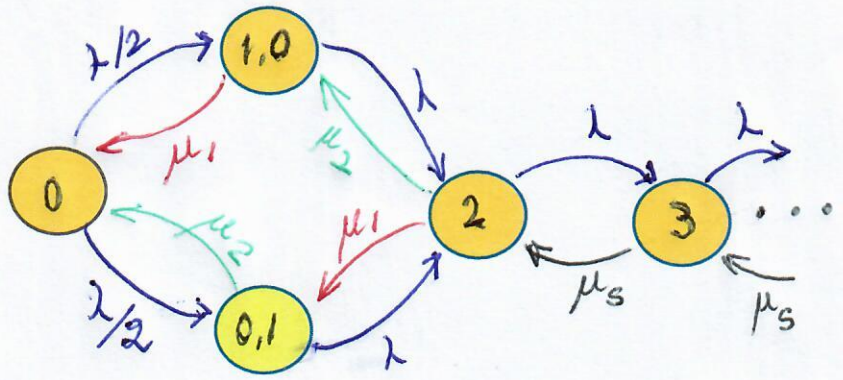
1er. Caso: SIN SELECCIÓN DE SERVIDOR

Referencias: λ = tasa de arribos, $\mu 1$ = tasa del servidor rápido; $\mu 2$ = tasa del servidor lento

$\mu_s = \mu_1 + \mu_2$ ESTADOS : 1,0 servidor rápido ocupado, lento ocioso

0,1 servidor rápido ocioso, lento ocupado

Cuando esta en estado "0" y llega un cliente, tiene la misma probabilidad de ir al servidor rápido que de ir al servidor lento, entonces la mitad de la tasa de arribos aporta al servidor rápido y la otra mitad aporta al lento .

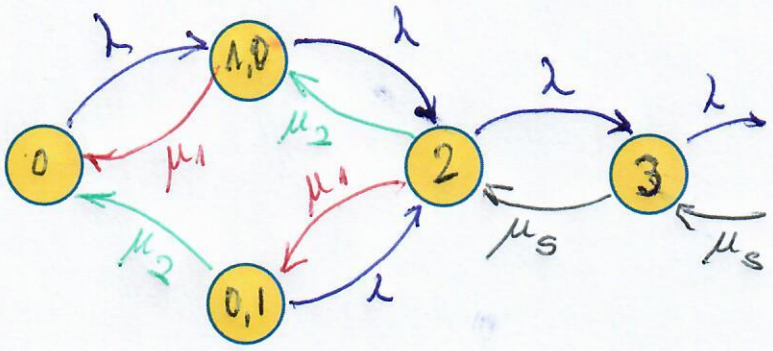


Cuando esta en estado "1,0" para pasar al estado "2" tiene que llegar un cliente e ir al servidor que lento, que es el que esta libre y eso ocurre con tasa λ. Si esta en estado "0,1" y arriba un cliente, este va a ir al servidor rápido que es el que esta ocioso y esto

ocurre con tasa λ . Si esta en estado "2" y arriba un cliente, se pone en cola, y esto ocurre con tasa λ y asi siguiendo.... LAS SALIDAS: Si el sistema esta en estado 1,0 y sale un cliente, este esta saliendo del servidor rapido y ocurre con la tasa del rapido, μ_1. Si el sistema esta en estado 0,1 y sale un cliente, esta saliendo del servidor lento y sale con la tasa del lento, μ_2. Cuando el sistema esta en estado 2 tiene dos opciones: vaciarse el lento y pasar al estado 1.0 con tasa μ_2 o vaciarse el rápido y eso ocurre con la tasa del servidor rápido, μ_1. Si el sistema esta en estado 3 para pasar al estado 2 puede salir el cliente que esta en el servidor rápido o el que esta en el lento, y eso ocurre con la tasa de servicio del sistema, porque las tasas se suman. Del estado 4 al 3 tambien es con la tasa de salida del sistema, y asi siguiendo para todos los estados superiores.

2do. Caso: CON SELECCIÓN DE SERVDOR

Las referencias son las mismas del diagrama anterior.



La diferencia se produce cuando el sistema esta en estado "0". Los dos servidores están libres y llega un cliente, el cliente que esta arribando siempre va al servidor rápido. El sistema pasa del estado "0" al 1,0 siempre y lo hace con la tasa λ, NUNCA pasa del estado "0" al 0,1. El resto de los pasajes son iguales que en el caso sin selección. La única diferencia esta en la llegada de un cliente cuando esta en estado "0". Observese que al estado 0,1 se puede llegar por la salida de un cliente cuando el sistema esta en estado 2.

CALCULOS EN LOS SISTEMAS M/M/2 CON SERVIDORES DE DISTINTAS VELOCIDADES

Cabe diferenciar los dos casos. Notese que el ρ_c se calcula con formulas diferentes para cada caso. Con respecto al calculo de π_0 y N, las formulas son las mismas para los dos casos pero los coeficientes designados con las letras a y a´, tienen formulas diferentes. Las diferencias se producen porque el caso con selección de servidor es un

poquito mas rápido que el caso sin selección de servidor porque cuando los dos servidores están libres, en el caso con selección si llega un cliente siempre va al servidor rápido, en el caso sin selección, si los dos servidores están desocupados, la mitad de las veces que llega un cliente va al servidor lento y eso le disminuye un poco la velocidad.

Para el caso sin selección de servidor:

$\rho_c = 1 - \left(r(1+r)/(1+r^2)\right)$ donde $r = \mu_2/\mu_1$

$\pi_0 = (1-\rho)/(1-\rho+\lambda/a)$ donde $a = (2\mu_1 * \mu_2)/(\mu_1 + \mu_2)$

$N = \lambda / \{(1-\rho)[\lambda+(1-\rho)*a]\}$

Para el caso con selección de servidor:

$\rho_c^2(1+r^2) - \rho_c(2+r^2) - (2r-1)(1+r) = 0$ **resolviendo la cuadrática se obt. ρ_c**

$\pi_0 = (1-\rho)/(1-\rho+\lambda/a')$ donde $a' = (2\lambda+\mu)(\mu_1 * \mu_2)/\mu(\lambda+\mu_2)$

$N = \lambda / \{(1-\rho)[\lambda+(1-\rho)*a']\}$