TEORÍA DE COLA

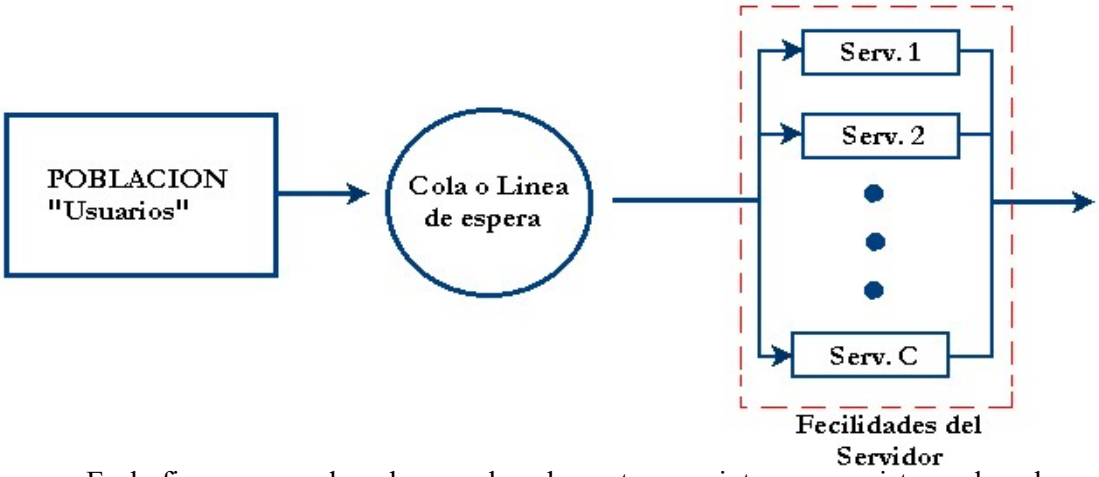
5.1 INTRODUCCIÓN

La teoría de redes de cola de espera (teoría de cola) es la fundamentación matemática para la mayor cantidad de modelos analíticos de sistemas de computación.

Las relaciones establecidas por la teoría de cola son relaciones entre cantidades abstractas que no pueden ser observadas directamente. Para hallar las soluciones a estas relaciones se las simplifica.

5.2 ELEMENTOS DE UN SISTEMA DE COLAS ABIERTAS

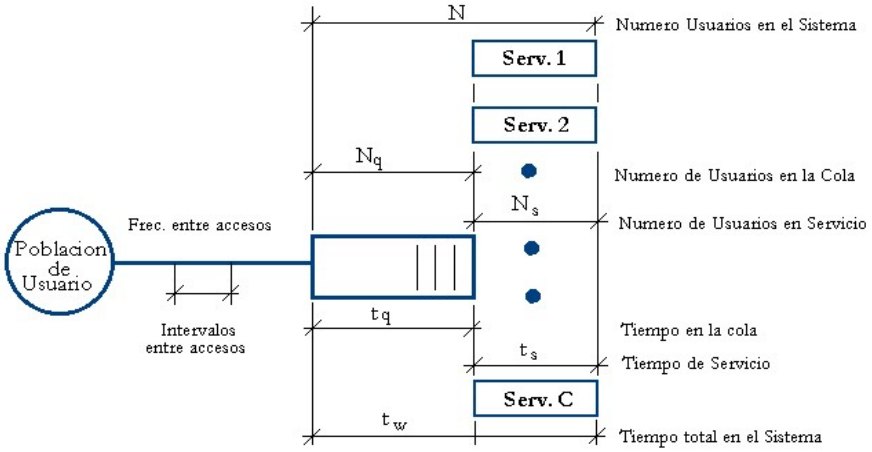
Una cola es una línea de espera y la teoría de cola es el estudio del fenómeno de la espera en la cola.



En la figura se pueden observar los elementos que integran un sistema de cola. “Usuarios” significa una entidad que desea algún tipo de servicio de un conjunto de servicios que ofrece el sistema.

Estos servicios pueden ser: establecer una determinada comunicación (sistema telefónico), procesar una pregunta, atender algún pedido de entrada/salida.

Si al ingresar un usuario al sistema todos los servidores están ocupados, aquel tendrá que formar una cola a la espera de la disponibilidad de alguno de ellos que pueda atenderlo.



La figura muestra un conjunto de variables aleatorias utilizada para el estudio de sistemas de cola. Se usa $t(_Q)$ representar el tiempo que un usuario arbitrario necesita estar en la cola a la espera de un servidor disponible (tiempo de espera). Y t(s) tiempo requerido para que el servidor provea el servicio (tiempo de atención o de servicio). Finalmente se tiene t(w) que es el tiempo total en el sistema de cola, obviamente t(w) será:

$$\mathbf{t(w) = t(q) + t(s)}$$

INFORME DE LA CONFERENCIA DE ESTANDARIZACIÓN DE COLAS POR ORSA, AIIE CORS Y TIMS

NOTACIONES DE MAGNITUDES DE LA TEORÍA DE COLA

☒ **C(c, u):** es la formula C de Erlang o la probabilidad de que todos los usuarios estén ocupados en un sistema de cola de espera M/M/c.

☒ **E[s]:** es el tiempo promedio de servicio para un usuario.

☒ **E[τ]:** es el intervalo promedio de tiempo entre arribos $E[\tau] = {}^{1}/_{\lambda}$, con λ velocidad o frecuencia promedio de arribos.

☒ **L:** es el número promedio de usuarios en el sistema cuando el sistema esta en estado estable.

☒ **Lq:** es el número promedio de usuarios en la cola (no incluye el número de usuarios que se hallan recibiendo atención) en un estado estable del sistema.

☒ **λ:** frecuencia o velocidad de arribo al sistema de cola $\lambda = {}^{1}/_{E[\tau]}$.

☒ **λ^T:** es la frecuencia de atención media en tareas o interacciones por unidad de tiempo.

☒ **μ:** es la frecuencia promedio de servicio por servidor $\mu = {}^{1}/_{E[s]}$.

☒ **N:** es una variable aleatoria que describe el número de usuarios en el sistema de cola de espera cuando el sistema esta estable.

☒ **Nq:** es una variable aleatoria que describe el número de usuarios en la cola durante el estado estable del sistema.

☒ **Ns:** es una variable aleatoria que describe el número que están recibiendo servicios durante un estado estable del sistema.

☒ **t(q), t(s) y t(w):** son variables aleatorias que describen el tiempo de un usuario en la cola, en servicio o en el sistema respectivamente.

☒ **ρ:** es el factor de utilización del servidor. Así ρ = intensidad de trafico/c = λ E[s]/c = (λ/μ) / c.

☒ **τ:** es una variable aleatoria que describe el intervalo entre arribos.

☒ **W:** es el tiempo promedio esperado en el sistema en estado estable. W = Wq + E[s].

☒ **Wq:** es el tiempo promedio esperado en la cola (tiempo de espera), excluye el tiempo de servicio, para el sistema en estado estable.

5.3 NOCIONES DE UN SISTEMA DE COLA

FUENTES: esto es el conjunto de usuarios que pueden solicitar el servicio del sistema, puede ser una fuente finita o infinita. Si se tuviera un sistema de fuente finita la longitud de la cola estaría determinada por el número de usuarios en el sistema, ya que estos afectan la frecuencia de arribo a ella. En cambio, en un sistema con fuente infinita, la cola es ilimitada, por lo tanto la frecuencia de arribos no esta afectada por el número de usuarios presentes en el sistema. Si la fuente fuese limitada pero con gran número de usuarios potenciales se supone una población infinita con el objeto de simplificar la tarea matemática.

PROCESO DE ARRIBO: se supone que los usuarios entran en el sistema en los tiempos: $t_0 < t_1 < t_2 < ..... < t_n$. Las variables aleatorias $\tau_k = t_k - t$ (k-1), (donde $k \geq 1$) se llaman intervalos entre arribos. Se supone que la τ_k forma una secuencia de variables aleatorias distribuidas independiente e idénticamente, y se usa el símbolo τ para un intervalo arbitrario entre arribos. El patrón más común de arribos en la terminología de la teoría de cola es el patrón de arribo aleatorio o proceso de arribo de Poisson. Esto significa que la distribución de intervalos entre arribos es exponencial, esto es:

Para τ $\qquad$ **$P[\tau \leq t] = 1 - e^{-\lambda t}$**

Para cada intervalo entre arribos, y la probabilidad de n arribos en algún intervalo de tiempo t es:

$e^{-\lambda t} \ (\lambda t)^n / n!$

Donde n = 0, 1, 2, ... Aquí λ es la frecuencia promedio de arribos, y el número de arribos por unidad de tiempo tiene una distribución de Poisson.

DISTRIBUCIÓN DEL TIEMPO DE SERVICIO: sea s(k) el tiempo requerido para la atención del k-esimo usuario. Supongamos los s(k) como una variable aleatoria independiente e idénticamente distribuidos. Por lo tanto, se puede considerar un tiempo de servicio arbitrario s. Se supone también una función de distribución común $Ws(t) = P[s \leq t]$ del tiempo de servicio para todos los usuarios. La función de distribución del tiempo de servicio en la teoría de cola es exponencial, lo cual define el denominado servicio (o atención) aleatoria. El símbolo μ se reserva para la frecuencia promedio de servicio y la función de reserva para la frecuencia promedio de servicio y la función de distribución para el servicio aleatorio será:

$$W_s(t) = 1 - e^{-\mu t}$$

Donde $t \geq 0$. Otras distribuciones de tiempos de servicios comunes son: las Enlang-k, hiper-exponencial y la constante.

Un parámetro estadístico que es útil como una medida de la característica de la distribución de probabilidades para el intervalo de arribo y para el tiempo2 de servicio es

el coeficiente de variación al cuadrado – $C^2(x)$ – el que se defina para una variable aleatoria x como:

$$C^2_{(x)} = \frac{VAR[x]}{E^2_{[x]}}$$

si x es constante: $C^2(x) = 0$;
si x tiene una distribución Erlang-k: $C^2(x) = {}^1/_k$;
si x tiene una distribución exponencial: $C^2(x) = 1$ y
si x tiene una distribución: $C^2(x) \geq 0$;

En base a estos datos se puede concluir que:

- **Para $C^2(\tau)$ aproximadamente igual a cero, el proceso de arribo tiene un patrón regular.**

- **Para $C^2(\tau)$ aproximadamente igual a uno, el proceso de arribo es aproximadamente aleatorio.**

- **Para $C^2(\tau)$ mayor que uno, los arribos tienden a agruparse.**

Similar conclusión puede ser tomada con respecto a la distribución del tiempo de servicio, donde pequeños valores de $C^2(s)$ corresponden a tiempos de servicios aproximadamente constantes y valores altos de $C^2(s)$ a mayores variaciones en los tiempos de servicio.

CAPACIDAD MÁXIMA DEL SISITEMA DE COLA: en algunos sistemas de cola la capacidad de la cola se supone infinita. Esto es cualquier arribo de un usuario se permite. Otros sistemas, llamados sistemas de perdida tienen una capacidad de linea de ampers igual cero. En ellos si un usuario arriba cuando las facilidades de servicio están cubiertas no podrá esperar. Sin embargo, esto tienen una capacidad de cola positiva pero no infinita.

NÚMERO DE SERVIDORES: el sistema de cola más simple es el de servidor único, el cual puede atender un usuario solamente por vez. Un sistema multiservidor tiene servidores idénticos y puede servir hasta n usuarios simultáneamente.

DISCIPLINA DE COLA: la disciplina de cola, a veces llamada disciplina de servidor es la regla para seleccionar el próximo usuario a recibir servicio. La disciplina más común de cola es la FIFO (primero en arribar, primero en atender). También existen disciplinas como LIFO (último en arribar, primero en atender) o disciplinas donde se consideran algunas prioridades de los usuarios.

Existe una notación, llamada ***Notación kendall***, que se ha desarrollado para especificar un sistema de cola y que tiene la forma A/B/c/k/n/Z. En ella A indica la distribución del intervalo entre arribos, B la distribución del tiempo de servicio, c el número de servidores, k la capacidad del sistema, n el número de usuarios y Z la

disciplina de él. Una notación más frecuente es A/B/c, se la utiliza cuando la cola de espera no tiene limites, la fuente de usuarios es infinita y la disciplina es FIFO.

Para A y B se utilizan los siguientes símbolos:

- **GI:** intervalo entre arribos general e independiente.
- **G:** tiempo de servicio general.
- **Ek:** distribución de tiempo de servicio o de intervalo entre arribos Erlang-k.
- **M:** distribución de tiempo de servicio o de intervalo entre arribos exponencial.
- **D:** distribución determinística o constante de tiempo de servicio o de intervalos entre arribos.
- **Hk:** distribución hiper-exponencial en k estados de tiempo de servicio o de intervalo entre arribos.

INTENSIDAD DE TRAFICO: la intensidad de tráfico u es la relación entre el tiempo de servicio promedio E[s] y el intervalo promedio entre arribos E[τ] esta relación es un parámetro importante de un sistema de cola y se define como:

$$u = \frac{E[s]}{E[\tau]} = \lambda \cdot E[s]$$

La intensidad de tráfico u determina el número mínimo de servidores que se requieren para cumplir con el flujo entrante de usuarios. De este mono, por ejemplo, si E[s] es 15 segundos y E[τ] es 10 segundos, luego, u = 1.5, esto indica que al menos se requerirán dos servidores. La unidad de intensidad de tráfico es el Erlang.

FACTOR DE UTILIZACIÓN DEL SERVICIO: Otro parámetro importante es la intensidad de tráfico por servidor (u/c), llamado factor de utilización del servicio - ρ - (cuando el tráfico se divide igualmente entre los servidores). Es la probabilidad de que algún servidor este ocupado. Para sistemas de un único servidor coincide con la intensidad de tráfico.

PROBABILIDAD DE n USUARIOS ESTEN EN EL SISTEMA EN EL TIEMPO t: Esta probabilidad, pn(t), depende no solo del tiempo t, si no también de las condiciones iniciales del sistema de cola, esto es, de número de usuarios presentes cuando se inicia el servicio y de las distribuciones y parámetros mencionados. Para los sistemas más útiles de cola, como t crece, pn(t) se aproxima al valor del estado estable, el es independiente de t y de las condiciones iniciales. El sistema se dice entonces que esta en un estado estable. Consideraremos solamente soluciones en estados estables. Las soluciones a problemas de cola que dependan del tiempo o soluciones transitorias son más complejas de hallar.

La teoría de cola provee mediciones estadísticas de la performance de los sistemas de cola. Algunas mediciones estadísticas incluyen Wq, W, L y Lq.

Las siguientes formulas llamadas leyes de Littlo, son bastante útiles relacionándolas con cuatro medidas de performance primarias:

$$\mathbf{Lq = \lambda\ Wq}$$
$$\mathbf{L = \lambda\ W}$$

Otra medición de performance útil es el valor del 90 percentil del tiempo en el sistema, w(90), el cual define la cantidad de tiempo necesaria tal que el 90% de todos los que arribaron a la cola (o sea al sistema) no esperan más que ese tiempo. Expresado simbólicamente, $\pi_{w(90)}$, se define por la ecuación:

$$\mathbf{P[w \le \pi_{w(90)}] = 0.9}$$

Se define de forma similar el 90 percentil del tiempo en la cola.

5.4 MODELO DE COLA ABIERTA. – APLICACIONES -

El sistema de cola M/M/c es un modelo de cola simple, abierta, que puede ser usado para modelar, algún sistema de computo. Se considera abierto, pues los usuarios entran al sistema desde el exterior, reciben el servicio y parten. Los sistemas cerrados, en los cuales el usuario nunca deja el sistema, se consideran luego. Las ecuaciones del sistema de cola M/M/c se muestran a continuación:

$$u = \frac{\lambda}{\mu} = \lambda\ E[s] \quad \textit{intencidad de Trafico}$$

$$\rho = \frac{u}{c} \quad \textit{foactor de utilizacion del servidor}$$

Probabilidad de que todos los c servidores estén ocupados de tal forma de que un nuevo usuario deba esperar servicio –C(c, u)-. Se lo puede calcular mediante la formula C de Erlang.

$$C_{(c,\,u)} = \frac{u^c/c!}{\frac{\mu}{c!} + (1-\rho)\sum_{n=0}^{c-1}\frac{\mu^n}{n!}}$$

$$Wq = \frac{C_{(c,\,u)}\ E[s]}{c\,(1-\rho)} \quad \textit{tiempo medio de cola}$$

$$W = Wq + E[s] \quad \textit{tiempo medio en el sistema}$$

$$\pi_q(90) = \frac{E[s]}{c\,(1-\rho)} \ln(10\,C_1) \quad \textit{tiempo del 90 percentil en la cola}$$

Cuando c = 1 las formulas se simplifican resultando:

$$C_{(c,\,u)} = \rho = \lambda \; E[s]$$

$$Wq = \frac{\rho \; E[s]}{1 - \rho}$$

$$W = \frac{E[s]}{1 - \rho}$$

$$\pi q\,(90) = W \;\; \ln\,(10\,\rho)$$

- Ejemplo N°1 -

Una oficina de una firma consultora tiene un terminal conectado a un sistema central de computo 8 horas al día. Ingenieros que trabajan fuera de la ciudad se dirigen a la oficina para hacer cálculos en el terminal. Sus patrones de arribo son aleatorio (Poisson) con un promedio de 10 personas por día usando el terminal. La distribución de tiempo de espera por ingeniero en el terminal es aproximadamente exponencial con un valor promedio de media hora. El terminal se utiliza entonces 5/8 del tiempo total (10 ing. X 0,5 horas = 5 horas sobre las 8 horas disponibles). Es necesario para la empresa mejorar el tiempo de espera de los ingenieros para manejar el terminal cuando el que se dispone se usa solamente 5/8 del tiempo en promedio. Como ayuda la teoría de cola a este problema?

- SOLUCIÓN -

El sistema de cola M/M/c es un modelo razonable para este sistema con ρ = 5/8 (hay verdaderamente un número finito de ingenieros pero la presunción de un número infinito parece razonable). Luego, usando las ecuaciones anteriores para el modelo M/M/1 (con c = 1) podemos calcular las medidas de la performance externa.

- Tiempo promedio que el ingeniero espera en la cola:

$$Wq = \frac{\rho \; E[s]}{1 - \rho} \qquad Wq = 50m.$$

$$Wq = 30m.\,(5/8)\,(3/8)$$

- Tiempo promedio que un ingeniero esta en el sistema (espera en cola más el tiempo de servicio):

$$W = \frac{E[s]}{1 - \rho} \qquad W = 80m.$$

$$W = 30m./(3/8)$$

- 90 percentil del tiempo de espera:

$$\pi_q(90) = W \; \ln(10\,\rho) \qquad \pi_q(90) = 147m.$$

$$\pi_q(90) = 80m. \; \ln 6$$

Como λ = 10 ing. / 8horas diarias resulta: $1/48$ing./min, usando las leyes de Little se calcula:

- El número promedio de ingenieros en la cola:

$$Lq = \lambda \; Wq \;\Rightarrow\; Lq = 1{,}0417 \text{ ing.}$$

- El número promedio de ingenieros en la oficina usando el terminal:

$$L = \lambda \; W \;\Rightarrow\; L = 1{,}667 \text{ ing.}$$

Estos valores estadísticos demuestran que apenas más de un ingeniero (en promedio) esta esperando para usar el terminal.

- Ejemplo N°2 -

El ejemplo anterior nos muestra una situación en la cual un terminal remoto se utiliza solamente un 62.5% (5/8) del tiempo que se dispone, en el tiempo promedio que un ingeniero debe esperar es de 50 minutos (con un 10% de ellos con una espera de cola superior a 146.6 minutos. Considerar un replanteo del sistema con el objeto de tener un tiempo medio en la cola que no debe exceder de 10 minutos y un 10% de los usuarios solamente esperar no más de 15 minutos.

Podrá aplicarse como opciones un único sistema con varios servidores (M/M/c) o varios sistemas con servidor único (M/M/1).

- SOLUCIÓN 1-

✔ El primer caso puede plantearse con dos servidores (M/M/2), se tendrá entonces:

- Factor de utilización del servidor:

$$\rho = \frac{u}{c} \Rightarrow \rho = 0{,}625 / 2 \Rightarrow \boxed{\rho = 0{,}3125}$$

- Probabilidad de que ambos servidores estén ocupados:

$$C_{(c,\,u)} = \frac{u^c/c!}{\frac{\mu}{c!} + (1-\rho)\sum_{n=o}^{c-1}\frac{\mu^n}{n!}} \Rightarrow \boxed{C_{(2,\,0{,}625)} = 0{,}1488}$$

- Tiempo medio en la cola:

$$Wq = \frac{C_{(c,\,u)}\ E[s]}{c\,(1-\rho)} \Rightarrow \boxed{Wq = 3{,}247 \text{ min}}$$

- Tiempo de cola del 90 percentil:

$$\pi q\,(90) = \frac{E[s]}{c\,(1-\rho)} \ln\left(10\,C_{(c,\,u)}\right) \Rightarrow \boxed{\pi q\,(90) = 8{,}67 \text{ min}}$$

Luego, un terminal más en la oficina satisface las necesidades.

- SOLUCIÓN 2-

✔ Se prueba la otra solución, instalar dos sistemas M/M/1:

- Factor de utilización del servidor:

$$\boxed{\rho = 0{,}3125}$$

- Tiempo medio en la cola:

$$Wq = \frac{\rho\ E[s]}{1-\rho} \Rightarrow \boxed{Wq = 13{,}63 \text{ min}}$$

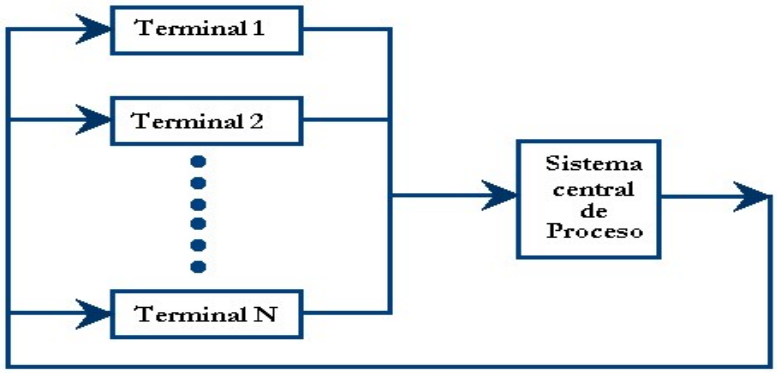
- Tiempo de cola del 90 percentil:

$$\pi q\,(90) = W\ \ln(10\,\rho) \Rightarrow \boxed{\pi q\,(90) = 49{,}72 \text{ min}}$$

Luego dos terminales en dos sistemas no son suficientes para satisfacer las necesidades.

5.5 MODELO DE COLA EN SISTEMAS INTERACTIVOS CON POBLACIÓN FINITA.

Se considerará ahora un modelo de cola más realista; modelo de cola cerrado (ningún usuario ni entra ni sale) y con una población finita de usuarios.



Un sistema de procesamiento central consiste de una cola para la unidad central de proceso (CPU), algunos terminales de entrada/salida y un sistema asociado de cola para cada uno de estos dispositivos. Los usuarios interactuan con la CPU a través de los n terminales. Cada usuario se supone en uno de los siguientes estados en cada instante:

(1) "Pensando" en el terminal ("tiempo de pensamiento", t)
(2) Esperando algún tipo de atención.
(3) Recibiendo el servicio.

El tiempo de pensamiento incluye todo el paso de tiempo entre el cumplimiento del servicio por parte de la CPU (para una dada interacción) y una nueva solicitud de servicio de la CPU por parte del usuario. Un usuario en el terminal no puede solicitar servicio hasta que el servicio pedido anteriormente no haya sido satisfecho. En la figura anterior el usuario puede representarse como una señal que circula, en el sistema y que en algún instante está en el terminal, en la CPU o en la cola de espera.

La estructura de cola particular que se estudia esta determinada por el modelo que se selecciona para el sistema de procesamiento central. A continuación se verán las ecuaciones que describen el modelo en condiciones generales:

$$W = \frac{N\,E[s]}{(1 - p_0)} - E[t] \quad \textit{tiempo de respuesta medio}$$

$$\lambda_T = \frac{N}{W + E[t]} \quad \textit{interaccion por unidad de tiempo}$$

Donde:

E[s]: tiempo medio de servicio de la CPU por interacción.
p_0: probabilidad de que el sistema este libre.
E[t]: "tiempo de pensamiento" medio (tiempo entre la finalización de un servicio y el próximo pedido).

El sistema de procesamiento central tendrá un modelo de tiempo compartido si el tiempo de servicio de la CPU es continuo y se usa la disciplina de cola por muestreo por medio de la CPU. Esta disciplina supone que la potencia de procesamiento de la CPU se distribuye igualmente entre todos los pedidos. Sin embargo, si hay n interacciones pendientes para el servicio de la CPU cada una de ellas se sirven instantáneamente por el sistema con una velocidad (frecuencia) de u/n usuarios por unidad de tiempo, donde u es la velocidad de procesamiento de la CPU. En la práctica da buenos resultados para tiempos de servicios cuantificado en cuantos finitos. Si existe un único CPU usando la disciplina de cola por muestreo, el p_0 se calcula de la siguiente forma:

$$p_0 = \left[\sum_{n=0}^{N} \frac{N!}{(N-n)!} \left(\frac{E[s]}{E[t]} \right)^n \right]^{-1}$$

También se puede calcular:

$\rho = 1 - p_0$ *el factor de ubicación de la CPU*

$\lambda_T = \frac{\rho}{E[s]}$ *la interacción por unidad de tiempo*

- Ejemplo N°3-

Un sistema interactivo de computo, por muestreo finito, tiene 20 terminales activos, un tiempo de pensamiento medio de 3 seg., una frecuencia de servicio de CPU de 500000 instrucciones por segundo (0,5 MIPS) y un requerimiento de interacción promedio de 100000 instrucciones. Se desea encontrar el tiempo de respuesta medio (W), las interacciones promedio (λ_T), la utilización de la CPU y el número de interacciones promedio pendientes en el sistema. Analizar que sucede si se agregan 10 terminales.

- SOLUCIÓN 1-

Cada interacción necesita la ejecución de un promedio de 100000 instrucciones de CPU, luego; E[s] = 100000/500000 seg. = 0.2seg. (tiempo promedio por interacción). La probabilidad de que el CPU este libre es:

$$\mathbf{p_0 = 0{,}0456}$$

Luego, W = 1,91 seg. Es el tiempo de respuesta medio.

El factor de utilización de la CPU es:

$$\mathbf{\rho = 1 - p_0 = 1 - 0{,}0456}$$

resultando:

$$\mathbf{\rho = 0.9544}$$
$$\mathbf{\lambda_T = 4{,}772}$$

Según las leyes de Little el número de interacciones pendientes en la CPU será:

$$\mathbf{L = \lambda W \quad \Rightarrow \quad L = 5{,}68\ int.}$$

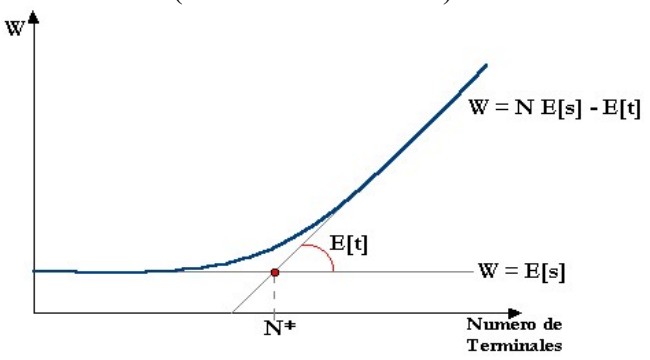
Si el número de terminales se lleva a 30 resulta:

$$\mathbf{p_0 = 0{,}00022}$$
$$\mathbf{\rho = 0{,}99978}$$
$$\mathbf{W = 3{,}00\ seg.}$$
$$\mathbf{\lambda_T = 5\ ^{int.}/_{seg.}}$$
$$\mathbf{L = 15\ int.}$$

Se ve que un aumento en el 50% en los terminales incrementa la cantidad de interacciones solo en un 4.78% a costa de un incremento en el tiempo de respuesta del 150%.

Esto ilustra el concepto de saturación de un sistema.

Considerando la figura, en ella se grafica el tiempo de respuesta medio W en función de N (número de terminales).



Para N =1 no hay cola (solamente se tiene un terminal) por lo tanto el tiempo de respuesta será E[s].

Para pocos terminales N la interferencia entre usuarios es mínima, esto quiere decir que cuando uno de los usuarios busca una interacción con la CPU los demás generalmente están en el modo de pensamiento, por lo tanto, ocasionalmente se formara una cola de espera. Es posible entonces considerar la curva estratégica a W = E[s] cuando se tiene pocos terminales.

Cuando el número de terminales tiende a infinito, la probabilidad de tener la CPU desocupada (p_0) tiende a cero, consecuentemente la curva será asintótica a la recta N x E[s] – E[t] con N tendiendo a infinito.

La intersección de las dos asíntotas se produce en N* tal como:

$$N^* = (E[s] + E[t]) / E[s]$$

Este valor de N se llama punto de saturación del sistema. Se concluye que si cada interacción requiere exactamente E[s] unidades de tiempo de servicio de CPU y exactamente E[t] "tiempo de pensamiento", N* es el número máximo de terminales que pueden ser incluidos de tal forma de no causar ninguna interferencia mutua.
En el ejemplo anterior:

$$N^* = 3 . 2 / 0.2 = 16$$

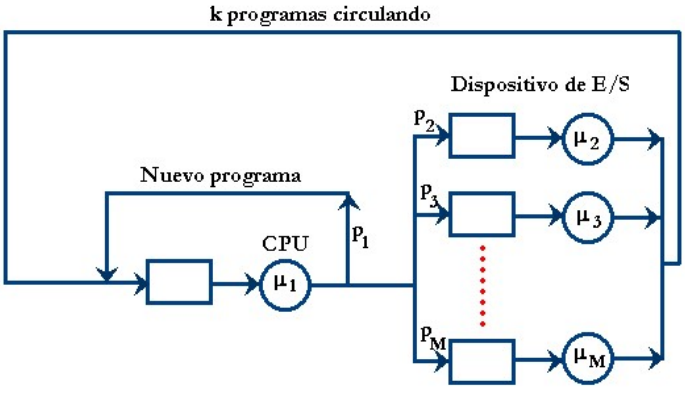
Además, el incremento de W después de un número N* de terminales es aproximadamente E[s] por el número de terminales incrementados. En el ejemplo: 10 x 0.2 = 2 seg. (realmente se tiene un tiempo de respuesta de 1.8 seg.).

.....**no**.....

5.6 MODELO DE MULTIPROGRAMACIÓN PARA EL SERVIDOR CENTRAL

Para representar el Sistema Central de Procesamiento se han formulado una serie de modelos. Uno de ellos de gran existo practico, es el *modelo de servidor central*. También se ha usado el modelo de sistema de multiprogramación. Este último es más complejo que los anteriores.

El modelo de multiprogramación es un modelo (ver figura) cerrado, pues contiene un número fijo de programas k, los cuales pueden ser pensados como atendidos siguiendo un ciclo interminable. Sin embargo, en cada momento hay un programa realizando el ciclo de la CPU. Una vez que este cumplió su ciclo uno nuevo entra en el sistema central.



Hay M-1 terminales de E/S, cada uno de los cuales tiene su propia cola y cada uno un tiempo de atención distribuido exponencialmente con una frecuencia promedio de atención u_i (con i = 2, 3, ..., M). La CPU se supone también con un u_1. Una vez que finaliza la ejecución de la CPU la tarea tendrá una probabilidad p1 de retornar a la CPU para completar la ejecución y probabilidades pi de atender un terminal E/S i.

Luego de completar la atención del terminal la tarea retorna a la cola de la CPU para otro ciclo. Si consideramos k = (k1, k2, ..., k_M) como representación de un estado interno del sistema, donde ki es el número de tareas (marcadores) en la i'esima cola (de espera o en servicio). Buzen muestra que la probabilidad P(k_1, ..., k_M) que el sistema se encuentre en el estado k está dada por:

$$P(k_1, \ldots\ldots, k_M) = \frac{1}{G(K)} \prod_{i=2}^{M} \left(\frac{\mu_i p_i}{\mu_i}\right)^{k_i}$$

para algun $(k_1, \ldots\ldots, k_M)$ **tales que:**

$$\sum_{i=1}^{M} k_i = K$$

Donde G(k) se define como que hace que las probabilidades sumen 1. En la figura los rectángulos representan las colas y los círculos las facilidades de servicio. Existen técnicas para calcular los G(k) con k = 0, 1, 2, ..., k. Estas técnicas constituyen el *ALGORITNO de BUZEN.*

Luego: las utilizaciones del servidor están dadas por:

$$\rho_i = \begin{cases} \dfrac{G(\mu - 1)}{G(K)} & i = 1 \\ \dfrac{\mu_i \rho_1 p_i}{\mu_i} & i = 2, \ldots\ldots M \end{cases}$$

El acceso promedio, λ_T, esta dado por:

$$\lambda_T = \mu_1 \rho_1 p_1$$

Si el modelo de servidor central es el modelo del sistema central de proceso del sistema interactivo, el tiempo de respuesta promedio W, se calcula por:

$$W = \frac{N}{\lambda_T} - E[t] = \frac{N}{\mu_1 \rho_1 p_1} - E[t]$$