CAPITULO 16

PLANIFICACION DE LA CARGA

16.1 - **Introducción**

Los objetivos de la Planificación de la Carga son :
- La ejecución de la mayor cantidad de "trabajos" en el menor tiempo posible.
- La no saturación de los recursos.

16.2. - **SISTEMAS BATCH**.

En los sistemas batch es donde existe una real planificación de la carga, ya que frente a un conjunto de trabajos posibles a ser ejecutados, es posible realizar la selección de los mismos de acuerdo a algún criterio.
A continuación detallaremos los más usuales.

16.2.1. - **Tiempo de llegada o planificación secuencial**.

A medida que llegan los trabajos van ingresando al sistema hasta la saturación de algún recurso (por ejemplo, que no haya más particiones disponibles).

16.2.2. - **El más corto primero**.

Ante la posibilidad de seleccionar entre varios trabajos, se seleccionan los más cortos, para que **menos** programas terminen más tarde.
Esta selección es sólo posible frente a un conocimiento previo del tiempo que necesitará cada trabajo.

16.2.3. - **Demoras**

Para poder determinar entre dos planificaciones distintas cuál es la mejor se introduce en elemento de medida que definiremos así:

Demora Ponderada = Demora Absoluta / Duración

definiendo Demora Absoluta el tiempo desde que un trabajo ingresa y espera ser ejecutado hasta que termina su ejecución y Duración al tiempo de ejecución de un trabajo.

Si tomamos el siguiente ejemplo:

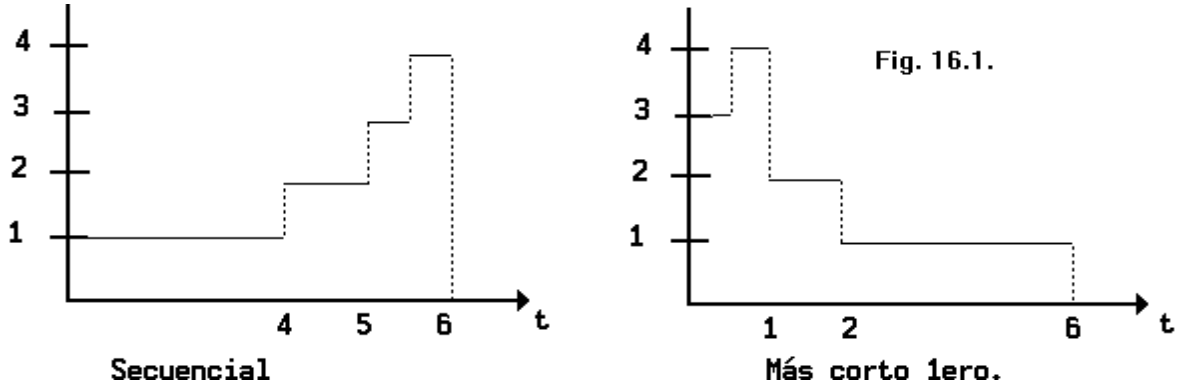


Fig. 16.1.

Trabajo	Duración
1	4
2	1
3	0,5
4	0,5

y los ejecutamos en monoprogramación (Fig. 16.1) en los dos sistemas de Planificación recién vistos resulta que en ambas planificaciones se termina todo en el mismo tiempo, pero será alguna más eficiente que la otra ?.

Apliquemos el elemento de medida.

Secuencial

Trab	Duración	Dem. Absoluta	Dem. Ponderada
1	4	4	1
2	1	5	5

3	0,5	5,5	11
4	0,5	6	12
Promedios		5.125	7,25

El más corto primero

Trab	Duración	Dem. Absoluta	Dem. Ponderada
3	0,5	0,5	1
4	0,5	1	2
2	1	2	2
1	4	6	1.5
Promedios		2,38	1,625

Si observamos la Demora Ponderada en ambos casos vemos que es mucho menor en el más corto primero. Lo que indica que tendremos más usuarios "satisfechos" antes, si bien en ambos casos el tiempo total de ejecución es el mismo.

La Demora Ponderada es un buen índice de medición, ya que independiza a los trabajos de los tiempos de su propia duración

16.2.4. - **Planificación con conocimiento futuro**.

Se asemeja al más corto primero.

Dado un trabajo largo por ingresar, pero conociendo que en poco tiempo ingresarán otros cortos, es posible que convenga esperar la ejecución de esos trabajos cortos, según se ve en el ejemplo:

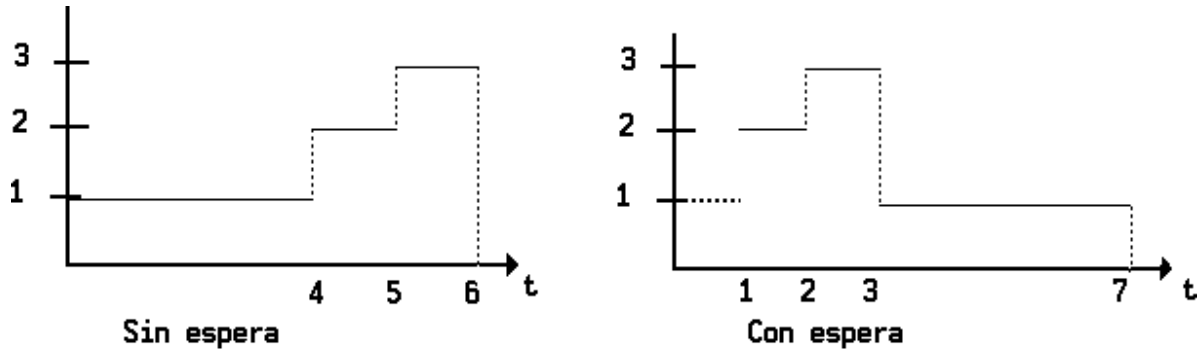


Fig. 16.2.

Sin espera

Trab	Duración	Dem. Absoluta	Dem. Ponderada
1	4	4	1
2	1	5	5
3	1	6	6
Promedios		5	4

Con espera

Trab	Duración	Dem. Absoluta	Dem. Ponderada
1	4	7	1.75
2	1	2	2
3	1	3	3
Promedios		4	2.25

16.2.5. - **Planificación por Mejor Aprovechamiento de los Recursos**.

Veamos el siguiente ejemplo (100K de memoria, 4 unidades de cinta y multiprogramación).

Nota: Consideraremos que la existencia de multiprogramación no afecta la duración de los trabajos, esto no es lógicamente así pero lo utilizaremos para simplificar los cálculos.

Trab	Duración	Memoria	Cintas
1	4	50 K	1
2	1	50 K	2
3	0,5	50 K	2
4	0,5	50 K	3

Planificación Simple

Trab	Duración	Dem. Absoluta	Dem. Ponderada
1	4	4	1
2	1	1	1

3	0,5	1,5	3
4	0,5	2	4
Promedios		2.13	2.25

Más corto primero

Trab	Duración	Dem. Absoluta	Dem. Ponderada
1	4	5	1,25
2	1	2	2
3	0,5	0,5	1
4	0,5	1	2
Promedios		2,375	1,5625

Mejor uso de Cintas

Trab	Duración	Dem. Absoluta	Dem. Ponderada
1	4	4,5	1,125
2	1	1	1
3	0,5	0,5	1
4	0,5	1,5	3
Promedios		1,875	1,53

Fig. 16.3.

El mismo criterio puede aplicarse para cualquier otro recurso o combinación de ellos, como memoria, dispositivos, uso de procesador, etc.

16.2.6. - **Planificación por Agotamiento de Recursos**.

Esta planificación permite la entrada de trabajos hasta que un recurso o grupo de recursos se saturan, por lo tanto puede ser por:

- cantidad de memoria,
- que la paginación no exceda un determinado valor,
- hasta el 100% de uso del procesador (en este caso es necesario verificar que el procesador sea en realidad utilizado por procesos usuario y no por el sistema operativo, en general si éste utiliza más del 5% del tiempo de procesador existe algún otro recurso utilizado en exceso, por ejemplo alta paginación).
- que la operaciones de E/S no superen un determinado nivel (por ejemplo, en general una utilización del canal mayor al 30% es excesiva).

16.2.7. - **Planificación por prioridades**.

Esta planificación provoca que se rompan todos los esquemas anteriores de mejor uso de recursos. Se utiliza cuando un determinado trabajo debe ser ejecutado, no importando lo que se está ejecutando en ese momento.

Es el caso, no frecuente, en que se convive con procesos Industriales de Control, los cuales deben ser ejecutados bajo pena de destrucción física o la necesidad de obtener resultados determinados para una hora determinada del día para tomar decisiones financieras, etc.

16.2.8. - **Planificación Algorítmica**.

Se elige un algoritmo que represente mejor los puntos débiles del sistema y con él se seleccionan los trabajos.

Un ejemplo de algoritmo es el de Balance entre uso de procesador y operaciones de Entrada/Salida.

Se solicitan los tiempos estimados de procesador (TCPU) y cantidad de operaciones estimadas de E/S (OPES) con los cuales se realizan las siguientes operaciones:

TOPES = OPES * (T.Posicionamiento + T.Latencia + T.Transferencia)
Duración Teórica = TOPES + TCPU

y se obtiene un coeficiente :

Coef = TCPU / Duración Teórica que es menor o igual a 1

Por ejemplo, el coeficiente en $TCPU_1 / (TCPU_1 + TOPES_1) + TCPU_2 / (TCPU_2 + TOPES_2)$ será = 1 cuando

$TOPES_1 = TCPU_2$

Y $TOPES_2 = TCPU_1$

Esto implica que las ráfagas de CPU de un proceso coinciden con las ráfagas de E/S del otro proceso.
En el caso de tres procesos la fórmula sería igual a 1 cuando:

$TOPES_1 = TCPU_2 + TCPU_3$

$TOPES_2 = TCPU_{1+} TCPU_3$

$TOPES_3 = TCPU_{1+} TCPU_2$

Luego se permitirá entrar la cantidad de trabajos que cumplan :

$$\sum \text{Coef} \leq 1$$

16.2.9. - **Planificación por Balance**.

Se busca cargar una mezcla de trabajos de mucha E/S con otros de alto uso de procesador.
Esto generalmente se utiliza con una planificación similar a la de prioridades en la cual se ejecutan durante el día trabajos cortos y de poco uso de procesador y se deja para la noche los de alto uso de procesador y poca E/S.

16.3. - **SISTEMAS INTERACTIVOS**.

Aquí no es posible establecer una planificación de antemano, solo se puede intentar que los tiempos de respuesta a los usuarios sean razonables.
Luego su objetivo es mantener satisfechos al usuario que está delante de un puesto de trabajo, y generalmente se utiliza el método de darle prioridad al usuario altamente interactivo.

16.3.1. - **Planificación por Contención**.

Significa admitir hasta **n** usuarios, el usuario **n+1** es rechazado. Si existiese alguno de mayor prioridad que desea entrar, se debería forzar a alguno de baja prioridad.
También en este caso puede utilizarse alguna de las variantes de agotamiento de recursos.

16.3.2. - **Planificación Ponderada**.

Cada tipo de usuario tiene asociada una carga, por ejemplo :

Común interactivo	1
Batch	1,5
Priorizado	0,5

Se determina cuantas unidades es capaz de soportar el sistema, por ejemplo 50, y pasadas estas unidades no se aceptan más usuarios.
Los usuarios son agrupados y cada grupo tiene un máximo de unidades, cuando es excedido ese máximo no se aceptan más usuarios del grupo.

16.3.3. - **Planificación Algorítmica**.

Esta se basa en que la utilización de algún recurso debe ser, en general, menor a un valor tope.

Utilización (recursos) = Tiempo Uso Recurso / Tiempo de Observación
< Tope >

Contención (recurso) = Procesos en espera / Tiempo de Observación
< Tope > (ideal 0)

Servicio (t. de respuesta prom.) = (Tiempo Servicio + Tiempo Cola) / Tiempo Observación
< Tope >

Proporción = (CPU (S.O.) + CPU.(Usuarios)) / CPU (Usuarios)
(lo más cercano posible a 1)

Para saber si el sistema anda bien, o se puede permitir el ingreso de un nuevo usuario, se debe verificar no exceder los valores máximos indicados (Topes) que correspondan a un sistema en particular.

TEORIA DE COLAS

16.4. **Modelización estocástica de los instantes de llegada y duración de trabajos**

16.4.1. **Introducción**

Consideraremos el caso en que **no** tenemos un sistema batch. Para eso vamos a estudiar la situación general en la que los trabajos llegan en instantes aleatorios para ser atendidos y cada uno necesita un tiempo de atención que es aleatorio y que no conocemos con anticipación. No podremos hacer una planificación de la carga pero, bajo ciertos supuestos bastante razonables, vamos a obtener conclusiones útiles sobre el comportamiento del sistema. Los resultados que vamos a ver son parte de una rama de la teoría de probabilidades llamada *teoría de colas*.

16.4.2. **Descripción del modelo y notación**

Los trabajos llegan en instantes aleatorios para ser procesados. A pesar de que no sabemos en qué instante va a llegar un trabajo, existe un promedio o **tasa de arribos** por unidad de tiempo, que denotamos con la letra $\boldsymbol{\lambda}$. Se supone que si observamos el sistema durante un tiempo no muy corto, podemos contar el número de arribos y dividiéndolo por el tiempo transcurrido vamos a obtener un buen estimador de λ.

Por ejemplo: Observamos una ventanilla y su cola durante 3 horas. En ese período llegaron 45 personas para hacer un trámite allí. Cuánto vale el estimador de λ?. En qué unidades?. Por ejemplo, en este caso λ sería 45 personas/hora o 0,25 personas/minuto.

c será el número de puntos de atención (**despachadores**) para los trabajos. Es el número de procesadores (o de cajeros cuando se modeliza la atención a clientes en un banco, por ejemplo) que están disponibles para ocuparse con las tareas que van llegando. Cuando llega un trabajo, si hay un despachador libre, es atendido por éste hasta su finalización. Si no, espera en una cola. Cuando se libera algún despachador, si hay cola, alguno de la cola va inmediatamente a ser atendido.

$\boldsymbol{\mu}$ denotará la **tasa de atención**, común a los *c* despachadores. Es el número de trabajos que un despachador es capaz de atender por unidad de tiempo. Lo podemos estimar contando la cantidad de atenciones por parte de un despachador en un período no muy corto y dividiendo por el tiempo transcurrido menos el tiempo en que el despachador estuvo ocioso. Si hay varios despachadores se realiza esa estimación para cada uno y luego se promedia (recordar que suponemos que todos atienden con la misma tasa).

Supondremos que $\boldsymbol{\lambda < c\ \mu}$ porque de no ser así, estaríamos incluyendo casos en los que se incrementa indefinidamente la cantidad de trabajos por atender sin posibilidad de retorno.
Por ejemplo en el caso de caso $\lambda = 15$ personas/hora y teniendo 4 despachadores $\boldsymbol{\mu}$ podría ser 4, es decir si cada despachador atiende a 4 personas/hora podemos satisfacer la necesidad del sistema ya que

$$15 < 4 * 4 = 16$$

Diremos que el sistema se encuentra en estado *i* (con $0 \leq i < \infty$) si el número de trabajos en el sistema, contando los que se están atendiendo y también los que esperan, es *i*.

$\boldsymbol{p_i}$ será la **probabilidad de que en un instante cualquiera el sistema se encuentre en estado *i***. Una interpretación práctica de esta magnitud esta dada por la proporción de tiempo (a largo plazo) en que el sistema tiene *i* trabajos.

Por ejemplo: Si observo al sistema durante 3 horas y veo que la suma de los períodos en que hubo 3 personas en el sistema es de 25 minutos entonces P_3 será aproximadamente 25/180 = 0,138.

Sabemos que la $\sum p_i = 1$ con i de 0 a ∞ es igual a 1 por ser probabilidades, entonces podrían tenerse observando el sistema digamos por un lapso de 5 horas, probabilidades del estilo de :

$P_0 = 20 / 300$
$P_1 = 30 / 300$
$P_2 = 50 / 300$
$P_3 = 0$
$P_4 = 0$
$P_5 = 200 / 300$
$P_6 = p_7 = p_8 = 0$

De todas maneras aspiramos a *calcular* las p_i a partir de λ, μ, c y los supuestos del modelo. Así podremos, por ejemplo, ver cómo se modifica la proporción de tiempo en que el sistema está ocioso (p_0) cuando duplicamos el número de despachadores, o qué efecto tiene una modificación en la velocidad de atención (μ) en la probabilidad de que haya cola ($\Sigma_{i>c}\, p_i$).

16.4.2.1. **Más supuestos sobre el modelo. Ecuaciones de balance.**

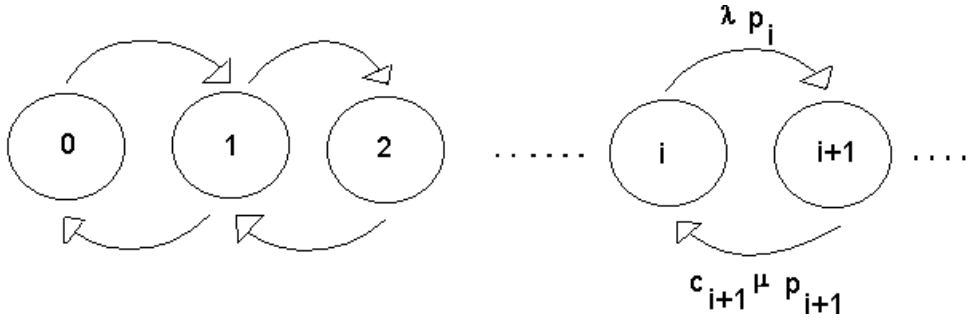
Los tiempos entre llegadas y los tiempos de atención son independientes entre si, en el sentido de que, por ejemplo, un tiempo de atención no da información sobre lo que durará la atención siguiente. Sí sabemos, como se dijo, que se verifican las frecuencias medias λ y μ. Los tiempos también son independientes del estado en que se encuentra el sistema. Esto implica, entre otras cosas, que con cola FIFO o LIFO obtendremos los mismos resultados sobre las p_i porque da lo mismo cuál trabajo tomemos de la cola para dárselo a un despachador. Existen variantes de este modelo que no suponen independencia.

Finalmente suponemos que para cada estado *i* se cumple la *ecuación de balance*

$$p_i\, \lambda = p_{i+1}\, c_{i+1}\, \mu$$

donde $c_k = k$ si $k < c$, y $c_k = c$ si $k \geq c$.
c_i es la cantidad de despachadores ocupados cuando el sistema está en el estado i.

El primer miembro de la igualdad es la tasa de pasaje del estado i al estado i+1, en tanto que el segundo miembro de la igualdad es la tasa de pasaje del estado i+1 al estado i.



Estas ecuaciones pueden deducirse de supuestos más elementales, dentro de la teoría de procesos de Markov. Estos consisten en considerar que los tiempos entre arribos y de atención siguen distribuciones exponenciales. Se puede ver además que la distribución exponencial describe muy bien el tipo de aleatoriedad que estamos manejando. Aquí vamos a prescindir de esas deducciones de las ecuaciones de balance viendo que ellas mismas constituyen suposiciones razonables.

En efecto, dado que p_i es la proporción de tiempo en que hay *i* trabajos en el sistema y λ es la cantidad de arribos por unidad de tiempo, tenemos que $p_i\, \lambda$ representa la tasa de salida del estado *i* hacia el estado *i+1*. Por otro lado, como c_{i+1} es la cantidad de despachadores atendiendo cuando hay *i+1* trabajos en el sistema, $p_{i+1}\, c_{i+1}\, \mu$ representa la tasa de salida del estado *i+1* hacia el estado *i*. Las ecuaciones de balance dicen entonces que a largo plazo (digamos en tiempo infinito) se iguala la cantidad de pasajes de *i* a *i+1* con los pasajes de *i+1* a *i*. Esto se debe a que si se produce un pasaje entre estos dos estados en un sentido, el siguiente pasaje que involucre estos dos estados deberá ser en el sentido opuesto. Así que en tiempo infinito los pasajes en un sentido se balancean con los del sentido opuesto, a no ser que a partir de un instante deje de haber pasajes entre estos dos estados (en cuyo caso podría haber hasta ese momento un pasaje más en un sentido que en otro). Vamos a descartar esto último notando que puede ocurrir solamente en dos casos razonablemente imposibles para nuestro modelo:

Primero: El número de trabajos se agrande irreversiblemente y entonces un paso de *i* a *i+1* nunca es compensado por el opuesto.

Segundo: Este sería el caso en que el número de trabajos en el sistema queda, a partir de un instante, siempre más chico que *i* o siempre más grande que *i+1* pero sin aumentar irreversiblemente en tamaño. Entonces deberá existir algún estado distinto de *0* por el que se pasa infinitas veces pero tal que a partir de un cierto instante, siempre que se sale de él, es en un único sentido.

Ambos casos pueden suponerse imposibles en nuestro modelo. Así completamos una justificación heurística de las ecuaciones de balance.

La notación de Kendall (muy difundida) llama *M/M/c* a un sistema como el descripto (M de Markoviano). Esta notación se refiere a que los arribos y las atenciones siguen la distribución exponencial, lo que hace markovianos a los procesos, y a que hay *c* despachadores. Podemos tener $c = \infty$, en el caso en que consideramos que siempre va a haber suficientes despachadores como para atender todos los trabajos que lleguen, sin que se forme cola. Se ha usado este último modelo para representar un disco que por tener un gran número de brazos móviles, va a poder atender en el acto toda solicitud de disco que se presente.

16.4.3. **Cálculos para un sistema *M/M/1***

En este caso tenemos c = 1, es decir $c_0 = 0$, $c_1 = 1$, $c_2 = 1$, …etc., siempre hay un despachador atendiendo excepto en el caso en que no hay trabajos para atender. Luego en virtud de las ecuaciones de balance:

$p_0 \lambda = \mu p_1$

$p_1 = p_0 (\lambda / \mu)$

$p_2 = p_1 (\lambda / \mu) = p_0 (\lambda / \mu)^2$

y en general, para $i = 1, 2, \ldots$:

$p_i = p_0 (\lambda / \mu)^i$

Como las p_i son probabilidades (o fracciones de un tiempo total), deben sumar 1. Así:

$1 = \Sigma_{i \geq 0} p_i = \Sigma_{i \geq 0} p_0 (\lambda / \mu)^i = p_0 (1- (\lambda / \mu)^{-1}$ (sumando la serie geométrica, que es posible ya que $\lambda / \mu < 1$ por que $\boldsymbol{\lambda < c \mu}$)

Recordemos que $\Sigma_{i \geq 0} q^i = 1 / (1 – q)$ *si* $q < 1$

Luego, despejando de aquí p_0 tenemos

$p_0 = 1- \lambda / \mu$, y reemplazando más arriba,

$p_i = (1- (\lambda / \mu)) (\lambda / \mu)^i$.

Por ejemplo, calcular la probabilidad de que haya cola aquí sería $\Sigma_{i > C} p_i = 1 – p_0 – p1$

Llamemos $N_s = N_c + N_a$ a la cantidad (aleatoria) de trabajos que tiene en un instante el sistema. N_a son los que están siendo atendidos y N_c los que están en la cola. La esperanza (o valor promedio) del número de trabajos en el sistema, $E(N_s)$, será la suma de todos los posibles números de trabajos en el sistema, cada uno multiplicado por la probabilidad (o fracción de tiempo) que le corresponde:

$\Sigma_{i \geq 0} i p_i = \Sigma_{i \geq 0} i (1- (\lambda / \mu)) (\lambda / \mu)^i = (1- (\lambda / \mu)) (\lambda / \mu) \Sigma_{i \geq 0} i (\lambda / \mu)^{i-1}$

$= (1- (\lambda / \mu)) (\lambda / \mu) (1- (\lambda / \mu))^{-2}$ (sumando la derivada de la serie geométrica)

Entonces, $E(N_s) = (\lambda / \mu) (1 – (\lambda / \mu))^{-1} = ((\mu / \lambda) – 1)^{-1} = \lambda / (\mu – \lambda)$

$E(N_c)$ será una suma análoga a la anterior pero a *i* trabajos atendiéndose le corresponden *i -1* trabajos en la cola si $i > 0$, y *0* trabajos si $i = 0$. Luego:

$E(N_c) = 0 . p_0 + \Sigma_{i \geq 1} (i -1) p_i = \Sigma_{i \geq 1} i p_i - \Sigma_{i \geq 1} p_i = \Sigma_{i \geq 0} i p_i - (1 - p_0) = E(N_s) – (1 - p_0) = E(N_s) + p_0 - 1$.

Reemplazando tenemos una expresión alternativa:

$E(N_c) = (\lambda / \mu) (1 - \lambda / \mu)^{-1} + 1 - \lambda / \mu - 1 = (\lambda / \mu)^2 (1 – (\lambda / \mu))^{-1}$.

Para calcular $E(N_a)$ tenemos en cuenta que hay 0 trabajos atendiéndose si hay 0 trabajos en el sistema, y hay 1 atendiéndose si hay 1 o más en el sistema. Luego:

$E(N_a) = 0 . p_0 + \Sigma_{i \geq 1} 1. p_i = 1 - p_0$

Notar la aditividad de la esperanza: $E(N_c+N_a) = E(N_c) + E(N_a)$, una propiedad que se cumple en general para la suma de dos variables aleatorias.

16.4.4. ***M/M/c* y *M/M/*∞**

Los razonamientos y los cálculos son análogos, salvo que aparecen otros tipos de series para sumar en lugar de la geométrica. Con $c = \infty$, la del desarrollo de la función exponencial y con $1 \leq c < \infty$, aparecen series geométricas salvo los primeros términos y las expresiones son un poco más complicadas, pero manejables.

Veamos por ejemplo el cálculo de las p_i en el caso $c = 3$:

$p_1 = p_0 (\lambda / \mu)$, como con $c = 1$, pero la segunda ecuación de balance es $p_1 \lambda = p_2 2 \mu$. Luego:

$p_2 = p_1 \lambda / (2 \mu) = p_0 (1/2) (\lambda / \mu)^2$,

y como la tercera ecuación de balance es $p_2 \lambda = p_3 3 \mu$, tenemos

$p_3 = p_2 \lambda/(3 \mu) = p_0 (1/6)(\lambda / \mu)^3$.

En este punto observemos que con $c = \infty$, obtenemos en general $p_i = p_0 (1 / i!)(\lambda / \mu)^i$,
y usando que $\Sigma_{i \geq 0} \; p_i = 1$, junto con la igualdad $e^x = \Sigma_{i \geq 0} x^i / i !$, podemos hallar las p_i.

Volviendo al caso $c = 3$, tenemos como cuarta ecuación de balance $p_3 \lambda = p_4 3 \mu$, porque no hay más que tres despachadores. En general, para $i \geq 4$, $p_{i-1} \lambda = p_i 3 \mu$, y por lo tanto

$p_i = p_{i-1} \lambda / (3 \mu) = p_0 (1/6)(1/3)^{i-3}(\lambda / \mu)^3$. Luego

$1 = \Sigma_{i \geq 0} \; p_i = p_0 [1+ \lambda/ \mu + (1/2) \Sigma_{i \geq 2} (1/3)^{i-2}(\lambda/\mu)^i]$

$= p_0 [1+ \lambda/ \mu + (1/2) (\lambda/\mu)^2 \Sigma_{i \geq 2} (1/3)^{i-2}(\lambda/\mu)^{i-2}]$

$= p_0 [1+ \lambda/ \mu + (1/2) (\lambda/\mu)^2 \Sigma_{i \geq 0} (1/3)^{i}(\lambda/\mu)^{i}]$

$= p_0 [1+ \lambda/ \mu + (1/2) (\lambda/\mu)^2 (1-\lambda/(3 \mu))^{-1}]$.

De aquí se despeja p_0 (obteniendo que $p_0 = e^{(-\lambda / \mu)}$) y con las ecuaciones de más arriba, el resto de las p_i.

16.4.5. - **<u>Tiempos esperados</u>**

<u>Tiempo esperado de atención de 1 trabajo en el caso M/M/1.</u>

Como μ es la cantidad de trabajos atendidos en una unidad de tiempo :

μ	trabajos en	1	unidad de tiempo
1	trabajo en	$1 / \mu$	unidad de tiempo

Ahora, para deducir heurísticamente la esperanza del tiempo que demora un trabajo en la cola y la esperanza del tiempo que demora un trabajo en pasar por todo el sistema, nos ponemos en la posición de un trabajo que llega al mismo

16.4.5.1. - <u>Tiempo esperado de espera en la cola</u> (si la cola es FIFO) - T_c

$E(T_c)$ = "el tiempo para un trabajo multiplicado por la cantidad esperada de trabajos que se tienen que procesar antes de mi llegada al despachador"

$= 1 / \mu (E (N_s)) = (1 / \mu) \; \lambda / (\mu - \lambda)$

$E (N_s)$ es la esperanza de una cantidad de usuarios en el sistema.

16.4.5.2. - Tiempo esperado total en el sistema

Partiendo de que $E(T_s)$ es la esperanza de tiempo de permanencia en el sistema

$E(T_s) = 1 / \mu (E (N_s) + 1)$ "el tiempo para un trabajo multiplicado por la cantidad de trabajos que están antes que yo, más yo mismo (todos estos son atendidos en un tiempo promedio de $1 / \mu$ cada uno)"

$= (1 / \mu) (\lambda / (\mu - \lambda) + 1)$

$= (1 / \mu) \mu / (\mu - \lambda)$

$= 1 / (\mu - \lambda)$

16.4.5.3. - Tiempo esperado de atención

Sabemos que $E(T_a) = E(T_s) - E(T_c)$ o equivalentemente $E(T_s) = E(T_c) + E(T_a)$ que viene a ser el tiempo de atención a todos los trabajos incluyéndome menos el tiempo de atención a mí mismo.

Remplazando ahora tenemos

$E(T_a) = 1 / \mu (E(N_S) + 1) - 1 / \mu (E(N_S))$

$= 1 / \mu E(N_S) + 1 / \mu - 1 / \mu (E(N_S))$

$E(T_a) = 1 / \mu$

Con estos resultados y basándonos en la definición de demora ponderada de un trabajo como cociente entre la demora absoluta y la duración, podemos calcular un cociente entre esperanzas para obtener una magnitud análoga a la demora ponderada pero para un sistema probabilístico M/M/1.

Demora ponderada en el sistema M/M/1 = Demora absoluta / duración

$= E(T_s) / E(T_a)$

$= (1 / (\mu - \lambda)) / (1 / \mu)$

$= \mu / (\mu - \lambda)$

$= 1 / (1 - \lambda / \mu)$

Ejercicios resueltos

1) *Proponer una manera de estimar μ que involucre los c despachadores.*

Se observa el sistema durante un tiempo. Para cada despachador k ($1 \leq k \leq c$) se mide el tiempo que no estuvo ocioso, t_k, y el número de personas que atendió, n_k.

n_k / t_k será un estimador de μ que luego se promedia para todos los despachadores.

Es decir: *Estimación de* $\mu = (\Sigma n_k / t_k) / c$.

2) *En un sistema M/M/1, la tasa de arribos es de 2.1 por segundo y la de atención de 3.2. Calcular la probabilidad de que el despachador no se encuentre ocioso y la longitud esperada de la cola.*

Probabilidad de que el despachador no esté ocioso = $1 - p_0 = \lambda / \mu = 2.1/3.2 = 0.65625$.

Longitud esperada de la cola = $E(N_c) = (\lambda / \mu)(1 - \lambda/\mu)^{-1} - \lambda / \mu = 0.65625 (.34375)^{-1} - 0.65625$
$= 1.91 - 0.66 = 1.25,$ aproximadamente.

3) *Para un sistema M/M/∞, obtener la probabilidad de que el sistema esté ocioso (p_0) en función de la tasa de arribos (λ) y la de atención (μ).*

Como en el apunte, partiendo de las ecuaciones de balance llegamos a que

$p_i = p_0 (1/i!)(\lambda/\mu)^i$. Luego

$1 = \Sigma_{i \geq 0} \; p_i = p_0 \Sigma_{i \geq 0} (1/i!)(\lambda/\mu)^i = p_0 e^{\lambda/\mu}$. Entonces:

$p_0 = e^{-\lambda/\mu}$.

4) *Para un sistema M/M/1, con tasa de atención 3 veces mayor que la tasa de arribo, hallar la probabilidad de que haya cola.*

Probabilidad de que haya cola $= \Sigma_{i>1} \; p_i = 1 - p_0 - p_1$
$= 1 - (1 - \lambda/\mu) - (1 - \lambda/\mu)(\lambda/\mu) = 1/3 - (2/3)(1/3) = 1/3 - 2/9 = 1/9$.

Algunos libros que incluyen el tema de teoría probabilística de colas

Introducción a los Sistemas Operativos, H. M. Deitel, 2da. Edición, Addison-Wesley.
Sistemas Operativos, S. E. Madnick y J. J. Donovan, Editorial Diana.
Real-Time Systems Design and Analysis. P. Laplante, IEEE Press.