SISTEMAS DE COLAS CON DISTRIBUCIONES DE TIEMPO DE SERVICIO DERIVADAS DE LA DISTRIBUCIÓN EXPONENCIAL

MIGUEL MIRANDA
Facultad de Ingeniería. Universidad de Buenos Aires. ARGENTINA[1]
miguelmiranda@netscape.net

Fecha Recepción: Diciembre 2014 - Fecha Aceptación: Abril 2015

RESUMEN

Si bien la suposición de que el tiempo de prestación del servicio en una disposición de colas tiene una distribución exponencial es muy apropiada para describir una considerable cantidad de situaciones reales, hay sin embargo muchos sistemas, especialmente aquellos en donde el servicio proporcionado comprende la ejecución de varias actividades (o fases), por lo que resulta inadecuado considerar dicha hipótesis. El objeto del presente trabajo es presentar un enfoque analítico de modelos matemáticos para describir distintas situaciones prácticas de servicios multitareas, muy comunes en la fabricación de productos, analizar la distribución de probabilidades resultante y formular expresiones cuantitativas de las variables de eficiencia tanto para el sistema global como para cada una de las correspondientes fases.

PALABRAS CLAVES: Sistemas de colas - MG1 - Distribución general de tiempos de servicio.

ABSTRACT

Although assuming that the server service time in a queue facility has exponential distribution is very suitable to describe a number of situations, there are however many systems, especially those where the service rendered by the channels includes the carrying out of various activities (or phases), for which it is inappropriate to consider said hypothesis. The object of this paper is presenting an analytical approach to mathematical models to describe different practical situations of multitasking services, very common in the manufacture of products, determining the resulting probability distribution, and formulating quantitative expressions of the efficiency variables, both for the system as a whole and for each of the corresponding phases.

KEYWORDS: Queueing systems - MG1 - General service time distribution.

[1] También Facultad de Ingeniería – Universidad Austral – Argentina; Facultad de Ingeniería – Universidad Católica Argentina – Argentina.

1. INTRODUCCIÓN

Cuando en un proceso Poisson la variable es el tiempo (o, de manera más general, el continuo) que debe transcurrir para que se verifiquen una cantidad determinada de eventos, la distribución se llama Gamma. Un caso especial de distribución Gamma es la denominada distribución Erlang-K, en donde la cantidad de "eventos que se deben verificar es un número entero "K".

El caso particular de distribución Erlang-1, es decir cuando la variable es el tiempo que transcurre hasta que se verifique un evento, se denomina exponencial. Una característica muy interesante de la distribución exponencial es la falta de memoria, también llamada propiedad de Markov. Esto significa que la probabilidad de que se verifique el próximo evento, independientemente del momento de la observación, no depende del tiempo transcurrido desde que se produjo el último evento. La distribución exponencial es la única distribución continua que exhibe esta propiedad de falta de memoria. El coeficiente de variación (relación entre desvío estándar y media) de la exponencial es igual a 1.

En los sistemas de colas, la distribución exponencial es, sin duda alguna, la más habitual, ya que su aleatoriedad presenta características únicas que la hacen muy adecuada para describir procesos de arribos y de servicios en infinidad de casos prácticos.

Sin embargo, hay muchas situaciones en donde una variable aleatoria exponencial no describe correctamente el intervalo de tiempo que se estudia. En efecto, cuando el trabajo que se brinda en los canales a los clientes consiste en la ejecución de varias tareas de tiempos exponenciales, la distribución del tiempo de servicio resultante será típicamente no exponencial. Las tareas que constituyen el servicio pueden desarrollarse en forma secuencial o en paralelo. Las distribuciones que se derivan como combinación de variables de distribución exponencial en serie se denominan distribuciones con pendiente (que tienen un coeficiente de variación menor a uno), mientras que las que surgen como combinación de variables en paralelo se las llama planas (o de alta varianza, con coeficiente de variación mayor a uno). En muchos casos es posible describir el tiempo de realización del servicio como una combinación de distribuciones con pendiente y planas. En este trabajo se desarrollarán ejemplos de distribuciones con coeficiente de variación menor a uno.

Otras distribuciones de tiempo, no derivadas de la exponencial, también de interés en algunos sistemas reales de colas, pero que no serán objeto de este trabajo, son las derivadas de la distribución beta, tales como la uniforme, la triangular y la doble triangular (esta última utilizada en el sistema de administración de proyectos PERT). Asimismo, hay procesos determinísticos, en donde los tiempos están perfectamente determinados; a la distribución correspondiente (si bien no se trata de una distribución) se la denomina 'degenerada' (o determinística).

Para un sistema de un solo canal de capacidad infinita al cual arriban clientes no impacientes conforme a un proceso Poisson a una tasa media λ, en estado de régimen permanente, en donde la duración del servicio es una

variable aleatoria con cualquier tipo de distribución, de media T_S = 1/μ, cuya notación Kendall es MG1, se verifica que la longitud promedio de la cola está dada por:

$$L = \rho + \frac{\rho^2 + \lambda^2 \cdot \sigma_s^2}{2 \cdot (1 - \rho)} \qquad [1]$$

en donde σ_S es el desvío estándar de la distribución de tiempo de servicio y ρ el factor de tránsito (ρ = λ/μ) que debe ser menor a 1 para que el sistema alcance el estado de equilibrio, de modo tal que la cola no aumente indefinidamente. La expresión [1], conocida como fórmula de Pollaczek-Khintchine[2] (P-K), es muy flexible en términos de aplicación a diferentes distribuciones estadísticas de la duración del servicio, pero no permite obtener los valores de las variables de eficiencia para cada fase del proceso de atención.

2. DESARROLLO DE MODELOS MATEMÁTICOS

Se presentarán aquí los casos de sistemas de colas más habituales de tiempos no exponenciales resultantes de combinaciones de tareas de tiempos exponenciales. En cada caso se hará un comentario general de la distribución estadística pertinente, formulando un ejemplo de aplicación, se desarrollará el modelo matemático para determinar los valores tanto de las variables de eficiencia como de las probabilidades de estado y, finalmente, se realizará un análisis comparativo de los resultados obtenidos contra los que resultarían de aplicar incorrectas hipótesis de trabajo.

En los modelos desarrollados se utilizará básicamente la nomenclatura propuesta en el texto "Teoría de Colas" (Miranda, 2013) y se asumirá que el sistema está en condiciones de régimen permanente, que no hay restricciones de capacidad, que la población de clientes es infinita, que se dispone de un solo canal con modalidad de atención FIFO, y que los clientes pertenecientes a una población homogénea y no impaciente arriban al sistema según un proceso Poisson.

2.1 Modelos con distribución de tiempo de servicio Hipoexponencial

La distribución hipoexponencial, un caso generalizado de la distribución Erlang-K, describe el tiempo (o longitud de un parámetro t) que transcurre hasta que se completan K tareas con distribuciones de tiempo exponencial de distinto parámetro. Llamando T_i = 1/μ_i al tiempo promedio de cada fase, la media y el desvío estándar de esta distribución son, respectivamente:

$$T = \sum_{1}^{K} T_i = \sum_{1}^{K} \frac{1}{\mu_i}$$

[2] La fórmula fue publicada en 1930 por el ingeniero austro-francés Félix Pollaczek y reformulada en términos estadísticos por el matemático ruso Alexandre Khintchine.

$$\sigma = \sqrt{\sum_{1}^{K} T_i^2} = \sqrt{\sum_{1}^{K} \frac{1}{\mu_i^2}}$$

Un ejemplo de distribución hipoexponencial puede ser el proceso de fabricación de una pieza que requiere tres tareas secuenciales diferentes (por ejemplo, posicionamiento, estampado y extracción), en donde las duraciones de cada operación son variables aleatorias distribuidas exponencialmente. En la FIGURA 1 se grafican los tiempos de cada tarea t_1, t_2 y t_3 (exponenciales) para este ejemplo y los tiempos totales de fabricación $t = t_1 + t_2 + t_3$ (hipoexponencial):

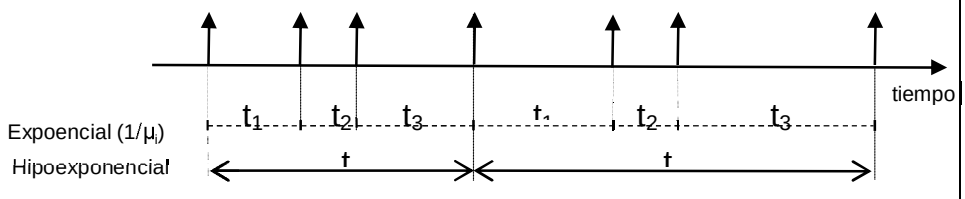

FIGURA 1

Conforme a la expresión anterior, la media y el desvío estándar de la variable "tiempo total de fabricación" serán:

$$T_S = \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\mu_3}$$

$$\sigma = \sqrt{\frac{1}{\mu_1^2} + \frac{1}{\mu_2^2} + \frac{1}{\mu_3^2}}$$

Se formulará ahora un modelo matemático para describir un sistema cuyo servicio consiste en la realización de una secuencia de K tareas con tiempos de ejecución diferentes. Se asume que el canal no se desocupa para atender a un nuevo cliente hasta tanto se termine la k-ésima actividad que está ejecutando sobre el cliente actual. Llamaremos:

T_{Si}: Tiempo promedio de servicio de la tarea i

μ_i: Tasa de atención promedio de la tarea i $\left(T_{S_i} = 1/\mu_i\right)$

ρ_i: factor de tráfico de la fase: Es la relación entre el flujo de clientes que pasa por el canal y la velocidad promedio de atención de la operación i $\left(\rho_i = \lambda/\mu_i\right)$.

H_i: Número promedio de clientes que se encuentra recibiendo la tarea i. Dado que el sistema es monocanal, este valor es también el porcentaje del tiempo que el canal está realizando la tarea i (porcentaje de ocupación en fase i).

Utilizaremos el símbolo h_K para indicar que la distribución es hipoexponencial, de modo que la notación Kendall correspondiente a este caso será: Mh_K1. En la Figura 2 se muestra una representación esquemática del sistema, con el cliente recibiendo la segunda tarea.

La duración total del servicio para cada cliente es la suma de los tiempos que se insume en cada tarea, y en consecuencia tendrá distribución hipoexponencial con media:

$$T_S = \sum_{i=1}^{K} T_{S_i} \qquad [2]$$

La tasa promedio del servicio completo es la inversa de dicho valor:

$$\mu = \frac{1}{T_S} \qquad [3]$$

El porcentaje de atención del canal será igual a la suma de los porcentajes de ocupación de las fases:

$$H = \sum_{i} H_i \qquad [4]$$

Haciendo el balance en régimen permanente de flujo ingreso-egreso del sistema y de cada fase, se tiene que:

$$\lambda = \mu \cdot H = \mu_i \cdot H_i \qquad [5]$$

de manera que

$$H_i = \frac{\lambda}{\mu_i} = \rho_i \qquad [6]$$

y también:

$$\mu = \mu_i \cdot \frac{H_i}{H} \qquad [7]$$

En definitiva, el porcentaje de ocupación de cada fase i es igual al factor de tráfico para esa fase i, y el porcentaje de ocupación global del canal igual al factor de tráfico:

$$H = \frac{\lambda}{\mu} = \rho = \sum_{1}^{K} H_i = \sum_{1}^{K} \frac{\lambda}{\mu_i} = \sum_{1}^{K} \rho_i \qquad [8]$$

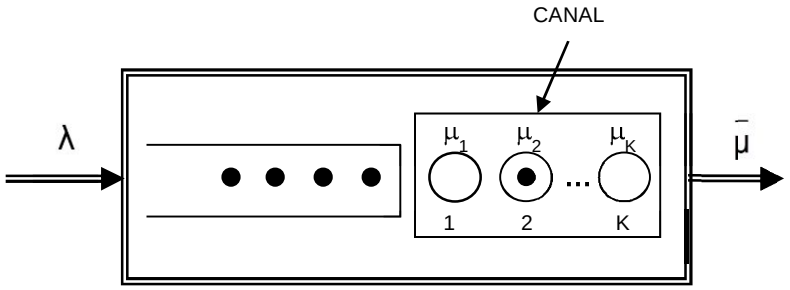

FIGURA 2

Para determinar las variables características del sistema se tendrá en cuenta la propiedad PASTA (*Poisson Arrivals See Time Averages*) que tienen los arribos poissonianos y la ley de Little (Little, 1961 y 2011). El tiempo promedio que tendrá que esperar un cliente que llega al sistema, cuando se está desarrollando la tarea i, hasta que pueda comenzar su servicio es igual a la suma de:

- el tiempo probable que deberá esperar hasta que termine el servicio que se estaba brindando (es decir, desde la fase i hasta la fase K), que es igual a la probabilidad de que el sistema se encuentre en la tarea i en el instante de arribo multiplicado por la suma de tiempos promedios de ejecución desde la operación i a la K:

$$\sum_{i=1}^{K}(H_i \cdot \sum_{i}^{K} T_{S_i})$$

- y del tiempo que tendrá que esperar por los clientes que están en cola:

$$T_S \cdot L_C$$

En consecuencia, el tiempo promedio de espera es:

$$W_C = \sum_{i=1}^{K}(H_i \cdot \sum_{i}^{K} T_{S_i}) + T_S \cdot L_C$$

Teniendo en cuenta Little ($L_C = \lambda \cdot W_C$):

$$W_C = \sum_{i=1}^{K}(H_i \cdot \sum_{i}^{K} T_{S_i}) + T_S \cdot \lambda \cdot W_C \qquad [9]$$

y que, por [8] y [3], $H = T_S \cdot \lambda$:

$$W_C \cdot (1-H) = \sum_{i=1}^{K}(H_i \cdot \sum_{i}^{K} T_{S_i})$$

Luego, tendremos:

$$W_C = \frac{\sum_{i=1}^{K}(H_i \cdot \sum_{i}^{K} T_{S_i})}{(1-H)} \qquad [10]$$

El tiempo de permanencia en el sistema será:

$$W = \frac{\sum_{i=1}^{K}(H_i \cdot \sum_{i}^{K} T_{S_i})}{(1-H)} + T_S \qquad [11]$$

Volviendo a considerar la ley de Little y las expresiones [6] y [8], nos queda expresada la longitud (cantidad de clientes) promedio del sistema:

$$L = \frac{\sum_{i=1}^{K}(H_i \cdot \sum_{i}^{K} H_i)}{(1-H)} + H \qquad [12]$$

Finalmente, teniendo en cuenta que la longitud promedio del sistema es igual a la suma de la cantidad promedio de clientes esperando recibir el servicio y de la cantidad promedio de clientes recibiendo el servicio, tendremos que:

$$L_C = \frac{\sum_{i=1}^{K}(H_i \cdot \sum_{i}^{K} H_i)}{(1-H)} \qquad [13]$$

Para determinar las probabilidades de estado en régimen permanente se analizará la cadena markoviana, representada en la FIGURA 3. Un estado, excepto cuando el sistema está vacío, queda definido por dos dimensiones: (n,i), en donde n = 1, 2, … es la cantidad de clientes que se encuentra en el sistema e i = 1, 2, …, K es el número de la fase operativa del canal.

La probabilidad de que el sistema se encuentre en el estado 0 es:

$$p(0) = 1 - H$$

Balanceando el nodo 0:

$$p(0) \cdot \lambda = p(1,K) \cdot \mu_K$$

se determina la probabilidad de que el sistema se encuentre con un cliente en la última etapa de ejecución:

$$p(1,K) = p(0) \cdot \frac{\lambda}{\mu_K} = p(0) \cdot \rho_K \qquad [14]$$

Balanceando ahora los nodos en donde hay un solo cliente en el resto de las etapas:

$$p(1,i)\cdot(\lambda+\mu_i)=p(1,i-1)\cdot\mu_{i-1}$$

de donde:

$$\boxed{p(1,i-1)=p(1,i)\cdot\frac{\lambda+\mu_i}{\mu_{i-1}}} \qquad [15]$$

La expresión [15] permite determinar todas las probabilidades para un cliente en el sistema (n=1) desde la etapa K hasta la etapa1.

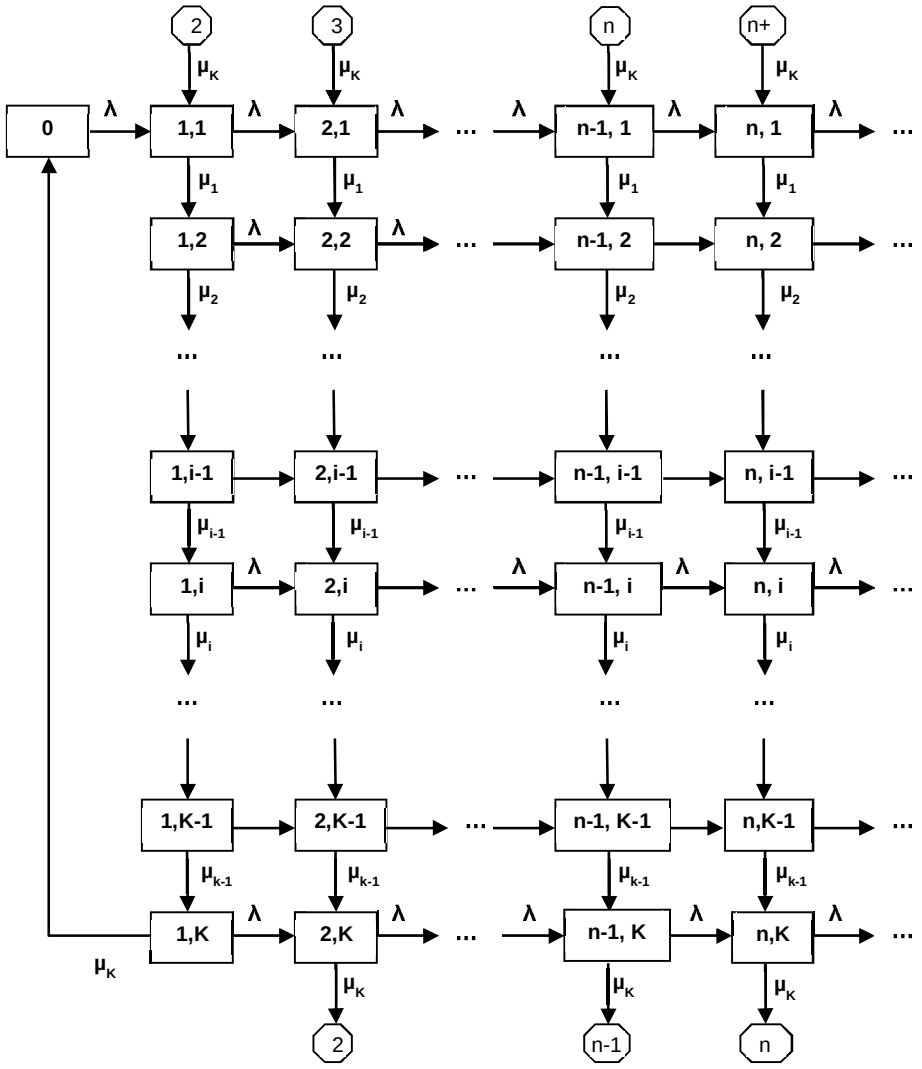

FIGURA 3

Haciendo ahora el balance en el nodo 1,1:

$$\boxed{p(1,1)\cdot(\lambda+\mu_1)=p(0)\cdot\lambda+p(2,K)\cdot\mu_K} \qquad [16]$$

se determina la probabilidad de que haya dos cliente en la última etapa de ejecución (K):

$$p(2,K)=\frac{p(1,1)\cdot(\lambda+\mu_1)-p(0)\cdot\lambda}{\mu_K}$$

Así, en general, para la última fase:

$$\boxed{p(n+1,K)=\frac{p(n,1)\cdot(\lambda+\mu_1)-p(n-1)\cdot\lambda}{\mu_K}} \qquad [17]$$

Para una cantidad de clientes genérica n, en una fase i, el balance es:

$$\boxed{p(n,i)=\frac{p(n-1,i)\cdot\lambda+p(n,i-1)\cdot\mu_{i-1}}{\lambda+\mu_i}} \qquad [18]$$

En definitiva las ecuaciones de estado son [14], [15] para n = 1 e i > 2, [16], [17] para n≥1 y [18] para n > 1 y 1 < i < K.

Por supuesto, que la probabilidad de que el sistema se encuentre con una cantidad n de clientes es:

$$p(n)=\sum_{i=1}^{k}p(n,i) \qquad [19]$$

En la TABLA 1 se muestra la diferencia entre los valores de las variables características que surgen de aplicar este modelo para K = 3 con parámetros T_{S_1} = 0,10, T_{S_2} = 0,14, T_{S_3} = 0,12, y λ = 0,9, contra un sistema MM1 con igual tasa de arribos y asumiendo que la media es igual a la suma de los tiempos medios de cada una de las actividades (T_S = 0,36).

2.1 Modelos con distribución de tiempo Erlang-K

Un caso especial de distribución hipoexponencial es la Erlang-K, que describe el tiempo (o longitud de un parámetro t) que transcurre hasta que suceden K eventos en un proceso de tipo Poisson de media λ. La media y el desvío estándar de esta distribución son:

$$T=\sum_{1}^{K}T_i=\sum_{1}^{K}\frac{1}{\mu_i}=\frac{K}{\mu_i}$$

$$\sigma=\sqrt{\sum_{1}^{K}T_i^2}=\sqrt{\sum_{1}^{K}\frac{1}{\mu_i^2}}=\frac{\sqrt{K}}{\mu_i}$$

La distribución erlangiana se obtiene combinando K distribuciones exponenciales en serie, todas ellas con idéntico parámetro μ. Un ejemplo de aplicación en un sistema de colas puede ser el proceso de fabricación de una pieza que requiere tres perforaciones similares en forma secuencial, en donde las duraciones de cada perforación tienen distribución exponencial e igual media.

Para formular el modelo matemático correspondiente se propone indicar el modo de atención de los procesos Erlang con el símbolo E_K, de modo que la notación Kendall será: $M/E_K/1$.

Dado que el tiempo de atención total en el canal es:

$$T_S = \sum_{i=1}^{K} \frac{1}{\mu_i} = \frac{K}{\mu_i} \qquad [20]$$

y además, la inversa de la tasa total de atención

$$T_S = \frac{1}{\mu} \qquad [21]$$

tendremos que:

$$\mu = \frac{\mu_i}{K} \qquad [22]$$

Luego, las expresiones del factor de tráfico de fase y del porcentaje de ocupación total del canal son:

$$\rho_i = \frac{\lambda}{\mu_i}$$

De igual manera que en el caso anterior, tendremos que

$$H_i = \frac{\lambda}{\mu_i} = \rho_i \qquad [23]$$

El porcentaje de atención del canal (o número promedio de clientes atendiéndose) será igual a la suma de los porcentajes de ocupación de las fases:

$$H = \sum_i H_i = K \cdot H_i$$

Es decir,

$$H_i = \frac{H}{K} \qquad [24]$$

Reemplazando estas variables en las expresiones [10] a [13] se resuelve el presente modelo. En la TABLA 2 se puede observar la diferencia entre los valores de las variables características que surgen de aplicar las expresiones anteriores con K = 3, para los parámetros $T_{S_1} = T_{S_2} = T_{S_3} = 0{,}12$, y

λ = 0,9, contra un sistema MM1 con parámetro: T_S = 0,36 e igual tasa de arribos.

2.2 Modelos con distribución de tiempo de máxima duración

Suponiendo que t_1, t_2, …, t_K son variables independientes y exponencialmente distribuidas, con parámetros μ_1, μ_2, …, μ_K, la variable aleatoria $t_{MÁX} = MAX(t_1, t_2, ..., t_K)$ que determina el valor máximo de ellas, tiene una distribución que llamaremos MAXX-K, cuya media es:

$$T_{MÁX} = T_1 + \frac{T_2^2}{T_1 + T_2} + \frac{T_3^2}{T_1 + T_2 + T_3} + ... + \frac{T_K^2}{T_1 + T_2 + ... + T_K}$$

siendo $T_1 \geq T_2 \geq \ldots T_{K-1} \geq T_K$.

Es decir:

$$T_{MÁX} = \sum_{i=1}^{K} \frac{T_i^2}{\sum_{j=1}^{i} T_j} \qquad [25]$$

Para el caso particular de que las variables estén idénticamente distribuidas (esto es, igual parámetro μ), la expresión anterior queda:

$$T_{MÁX} = T + \frac{T^2}{2 \cdot T} + \frac{T^2}{3 \cdot T} + ... + \frac{T^2}{K \cdot T} = \sum_{i=1}^{K} \frac{T}{i} \qquad [25A]$$

Estas expresiones han sido debidamente validadas por el autor con técnicas de simulación.

En sistemas de colas en los cuales los canales deben realizar varias tareas independientes en forma simultánea sobre el cliente, el tiempo total del servicio queda determinado por el de la tarea que haya durado más tiempo. Un ejemplo de ello puede ser un centro de carga de camiones de una empresa con una sola plataforma, en donde se posiciona un camión para que se desarrollen las actividades de carga de mercadería y de carga de combustible, ambas con tiempos aleatorios exponencialmente distribuidos. La plataforma se liberará recién cuando hayan finalizado las dos actividades; es decir el tiempo de ocupación del canal por parte de un camión es una variable aleatoria con distribución MAXX-2.

Se formulará a continuación un modelo matemático con idénticos supuestos que en el punto 2.1, excepto en el hecho de que la atención al cliente consiste en la ejecución de una cantidad K de tareas diferentes en el canal, no vinculadas entre sí, que comienzan simultáneamente y que se desarrollan en paralelo. El canal no se desocupa para atender a otro cliente hasta que terminen todos los trabajos.

Para la notación Kendall se propone utilizar el símbolo Mx_K con relación a la distribución correspondiente al tiempo máximo de los tiempos de las K

tareas, de modo que para este sistema será: $M/Mx_K/1$. En la FIGURA 4 se muestra una representación esquemática del sistema, estando el cliente del canal recibiendo la primera y la tercera tarea (habiéndose finalizado ya el resto).

El tiempo medio del servicio, por [25], será:

$$T_S = T_{S_1} + \frac{T_{S_2}^{\ 2}}{T_{S_1} + T_{S_2}} + \frac{T_{S_3}^{\ 2}}{T_{S_1} + T_{S_2} + T_{S_3}} + \ldots + \frac{T_{S_K}^{\ 2}}{\sum_{i=1}^{K} T_{Si}} \qquad [26]$$

en donde $T_{S_1} \geq T_{S_2} \geq \ldots \geq T_{S_{K-1}} \geq T_{S_K}$.

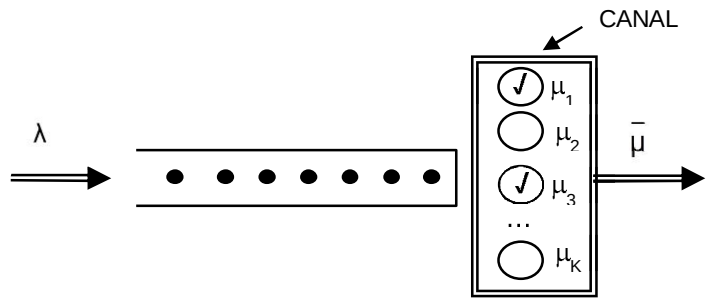

FIGURA 4

Luego, la velocidad promedio de atención del canal es:

$$\mu = \frac{1}{T_{MAX_K}} = \sum_{i=1}^{K} \frac{\sum_{j=1}^{i} T_{S_j}}{T_{S_i}^{\ 2}} \qquad [27]$$

El porcentaje de tiempo que el canal se encuentra realizando una tarea i cualquiera solamente (es decir, cuando ya han terminado el resto de las actividades) es:

$$H_i = \frac{\lambda}{\mu_i}$$

Del mismo modo, la probabilidad de que el canal esté ejecutando al mismo tiempo solamente las tareas i y j será:

$$H_{ij} = \frac{\lambda}{\mu_i + \mu_j}$$

Así, la probabilidad de que el canal esté ejecutando simultáneamente las K tareas (esto es, sin que haya terminado ninguna de ellas) es:

$$H_{ij\ldots K} = \frac{\lambda}{\mu_i + \mu_j + \ldots + \mu_K}$$

Por otra parte, para el sistema en estado de régimen

$$H = \frac{\lambda}{\mu}$$

Y, teniendo en cuenta [27]:

$$H = \sum_{i-1}^{K} \frac{\lambda \cdot T_{S_i}^{\ 2}}{\sum_{j-1}^{i} T_{S_j}} = \sum_{i-1}^{K} \frac{H_i \cdot T_{S_i}}{\sum_{j-1}^{i} T_{S_j}} \qquad [28]$$

Se procederá ahora a determinar cuantitativamente las variables características para un sistema que ejecuta dos tareas A y B en paralelo al cliente en el canal, utilizando el mismo enfoque que en los casos anteriores. Para ello, se definirán los estados de la siguiente forma:

(n,AB): n clientes en el sistema, ambas tareas ejecutándose,

(n,A): n clientes en el sistema, la tarea B finalizada, ejecutándose la tarea A.

(n,B): n clientes en el sistema, la tarea A finalizada, ejecutándose la tarea B.

El tiempo promedio que permanecerá un cliente que arriba al sistema para comenzar su servicio es igual a la suma de:

- el tiempo que deberá esperar para que termine el servicio actual que se está brindando, que es igual a T_{SMAX} si el sistema se encuentra en el estado AB, T_{SA} si el sistema se encuentra en el estado A, y T_{SB} si se encuentra en el estado B y, es decir: $H_{AB} \cdot T_{S_{MAX}} + H_A \cdot T_{S_A} + H_B \cdot T_{S_B}$
- el tiempo que deberá esperar por los clientes que están esperando en cola:
  $L_C \cdot T_{S_{MAX}}$
- el tiempo de su propio servicio: $T_{S_{MAX}}$

Luego:

$$W = H_{AB} \cdot T_{S_{MAX}} + H_A \cdot T_{S_A} + H_B \cdot T_{S_B} + L_C \cdot T_{S_{MAX}} + T_{S_{MAX}}$$

Considerando que $L = L_C + H$, teniendo en cuenta la ecuación de Little y operando, tendremos que el tiempo promedio de permanencia y el número promedio de clientes en el sistema es:

$$W = \frac{H_A \cdot T_{S_A} + H_B \cdot T_{S_B} + (1 - H + H_{AB}) \cdot T_{S_{MAX}}}{(1 - H)} \qquad [29]$$

Y, nuevamente, por Little:

$$L = \frac{H_A \cdot T_{S_A} + H_B \cdot T_{S_B} + (1 - H + H_{AB}) \cdot T_{S_{MAX}}}{(1 - H)} \cdot \lambda \qquad [30]$$

$$L_C = L - H \qquad [31]$$

$$W_C = \frac{L_C}{\lambda} = W - T_{S_{MAX}} \qquad [32]$$

Los porcentajes de tiempo que el canal está ejecutando ambas tareas simultáneamente solamente la tarea A, solamente la tarea B son, respectivamente:

$$H_A = \frac{\lambda}{\mu_A} \quad ; \quad H_B = \frac{\lambda}{\mu_B} \quad ; \quad H_{AB} = \frac{\lambda}{\mu_A + \mu_B}$$

En el caso de la expresión [28], para este modelo, H representa la cantidad de trabajos en ejecución.

Para la determinación de las probabilidades de estado, hay que plantear la cadena markoviana para la cantidad de tareas K del problema. Supongamos un sistema en donde el canal efectúa dos tareas A y B, siendo $T_A \geq T_B$. La cadena markoviana de un modelo general correspondiente a un estado genérico "n" se muestra en la Figura 5.

Las ecuaciones de estado generales son:

$$p(n,AB)\cdot(\lambda+\mu_A+\mu_B) = p(n-1,AB)\cdot\lambda + p(n+1)\cdot(\mu_B+\mu_A)$$

[33]

$$p(n,A)\cdot(\mu_A+\lambda) = p(n,AB)\cdot\mu_B + p(n-1,A)\cdot\lambda \qquad [34]$$

$$p(n,B)\cdot(\mu_B+\lambda) = p(n,AB)\cdot\mu_A + p(n-1,B)\cdot\lambda \qquad [35]$$

las que conjuntamente con la condición de colectividad exhaustiva:

$$\sum_{1}^{\infty}\left[p(n,AB)+p(n,A)+p(n,B)\right] = 1 \qquad [36]$$

permiten determinar las probabilidades de estado.

Para el caso especial en donde los tiempos de procesamiento son los mismos para todas las operaciones (T_{Si}), la expresión [26] del tiempo promedio del máximo de las K tareas queda:

$$T_S = T_{S_1} + \frac{T_{S_2}}{2} + \frac{T_{S_3}}{3} + \ldots + \frac{T_{S_K}}{K} = \sum_{i=1}^{K}\frac{T_{Si}}{i} \qquad [37]$$

Luego, la velocidad promedio de atención del canal es:

$$\mu = \frac{1}{T_S} = \sum_{i=1}^{K}\frac{i}{T_{Si}} \qquad [38]$$

En la TABLA 3 se puede observar la diferencia entre los valores de las variables características que surgen de aplicar este modelo con K = 2, para los parámetros T_{S_1} =0,12, T_{S_2} = 0,10, y λ = 2, contra un sistema MM1 con igual tasa de arribos y parámetro T_S = 0,165454 (es decir el promedio del tiempo máximo de las exponenciales).

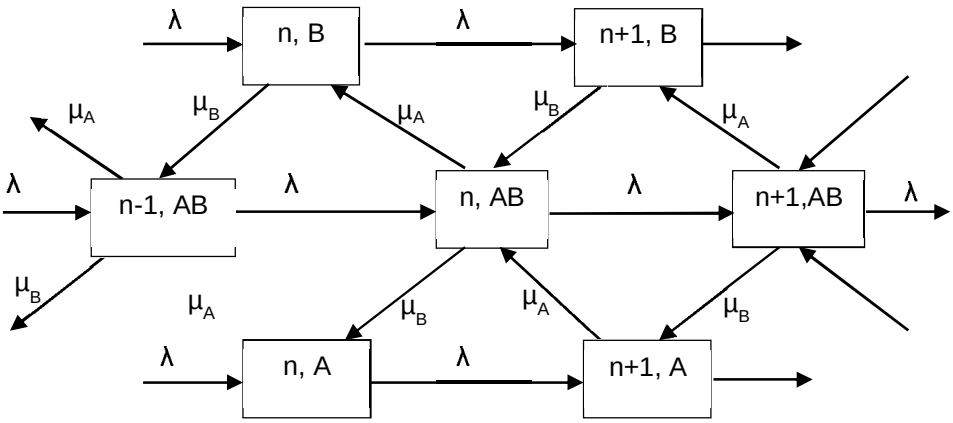

FIGURA 5

2.3 Modelos con distribución de tiempo de mínima duración

Suponiendo nuevamente que t_1, t_2, …, t_K son variables independientes y exponencialmente distribuidas, con parámetros μ_1, μ_2, …, μ_K, la variable aleatoria $t_{MIN} = \min(t_1, t_2, ..., t_K)$ que determina el valor mínimo de ellas, tiene una distribución que llamaremos MINX-K, cuya media es:

$$T_{MIN_K} = \frac{1}{\sum_{i=1}^{K} \mu_i} \qquad [39]$$

y el desvío estándar:

$$\sigma_{T_{MIN}} = \frac{1}{\sum_{i=1}^{K} \mu_i} \qquad [40]$$

Esta distribución es exponencial, tal como se observa, dado que el coeficiente de variación (es decir la relación entre el desvío estándar y la media) es igual a 1. La probabilidad de que t_i asuma el mínimo valor es:

$$p\{t_i = \text{mín}(t_1, t_2, ..., t_K)\} = \frac{\mu_i}{\sum_{1}^{K} \mu_n} \qquad [41]$$

En el caso particular de que todas las variables exponenciales estén idénticamente distribuidas (con parámetro λ), las expresiones de la media y el desvío estándar quedan:

$$T_{MIN} = \frac{1}{K \cdot \mu} \qquad [39A]$$

$$\sigma_{T_{MIN}} = \frac{1}{K \cdot \mu} \qquad [40A]$$

Sistemas de colas en donde se deben llevar a cabo diversos trabajos independientes en forma simultánea, y que el tiempo del servicio quede determinado por el mínimo de ellos no son muy frecuentes en la práctica. Un ejemplo sería la verificación de anomalías de máquinas que se realiza por dos métodos diferentes llevados a cabo en forma simultánea. El tiempo de la verificación será el tiempo del método que terminó primero.

Otro ejemplo puede ser el de un sistema de Programación Matemática que ejecuta modelos que se van recibiendo en un procesador central para ser resueltos. El sistema comprende tres optimizadores lineales que utilizan algoritmos diferentes (por ejemplo, Simplex, Karmarkar y un método híbrido), y da por resuelto el problema lineal (y en consecuencia, concluido el trabajo) cuando uno de los tres optimizadores termine la resolución.

Se formulará ahora un modelo matemático con idénticos supuestos que en el punto anterior, excepto en el hecho de que la atención al cliente finaliza cuando termina una de las K tareas que se desarrollan en forma simultánea, con el resto de las hipótesis iguales a las del modelo desarrollado en el punto 2.3. Se propone indicar en la notación Kendall a esta distribución con el símbolo Mn_K, de modo que para este modelo será: $M/Mn_K/1$. La representación gráfica del sistema de la FIGURA 6 muestra que, cuando el canal está activo, se ejecutan todas las tareas.

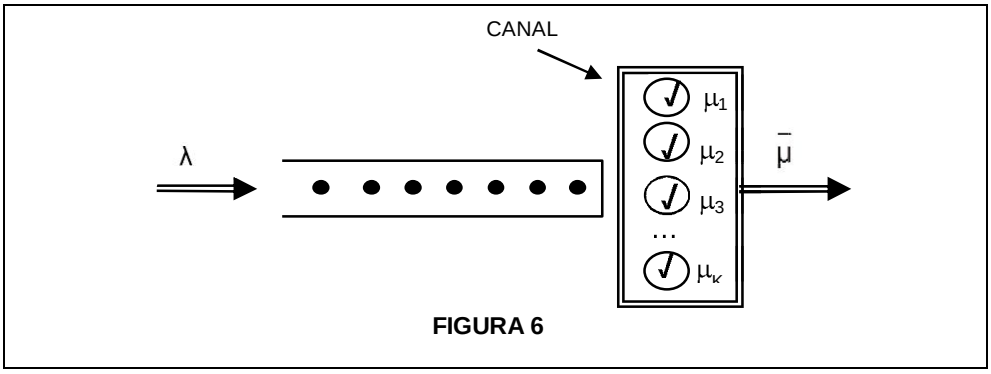

FIGURA 6

Siendo:

T_{S_i}: Tiempo promedio de servicio de la tarea i

μ_i: Tasa de atención promedio de la tarea i $\left(T_{S_i} = 1/\mu_i\right)$

tendremos, por [39]:

$$T_S = \frac{1}{\sum_{i=1}^{K} \mu_i} \qquad [42]$$

En consecuencia, para resolver este modelo basta con calcular la velocidad promedio de atención:

$$\mu = \frac{1}{T_S} = \sum_{i=1}^{K} \mu_i \qquad [43]$$

y reemplazar en las expresiones del modelo básico:

$$L = \frac{\lambda}{\mu - \lambda} \qquad [45]$$

$$L_C = \frac{\lambda^2}{(\mu - \lambda) \cdot \mu} \qquad [46]$$

$$W = \frac{1}{\mu - \lambda} \qquad [47]$$

$$W_C = \frac{\lambda}{\mu \cdot (\mu - \lambda)} \qquad [48]$$

La cadena markoviana correspondiente a un estado genérico “n” puede observarse en la FIGURA 7.

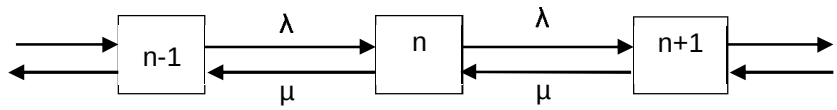

FIGURA 7

La ecuación general de estado de régimen permanente, para n > 0, es:

$$p(n) \cdot (\lambda + \mu) = p(n-1) \cdot \lambda + p(n+1) \cdot \mu \qquad [49]$$

que se puede expresar en forma simplificada de la siguiente forma:

$$p(1) = p(0) \cdot H \quad \text{para } n = 0 \qquad [50]$$

$$p(n) = p(0) \cdot H^n \quad \text{para } n \geq 1 \qquad [51]$$

El número promedio de trabajos que se están realizando, en promedio, en el canal es K·H.

Siendo:

$$H_i = \frac{\lambda}{\mu_i} \qquad [52]$$

el porcentaje de tiempo que estaría ocupado el canal si se realizara solamente la tarea i, la probabilidad de que la tarea "i" sea la de menor tiempo es:

$$P(T_{S_i} \rightarrow MIN) = \frac{\mu_i}{\sum_{j=1}^{K} \mu_j} = \frac{H}{H_i} \qquad [53]$$

Por supuesto que en estado de régimen se verifica:

$$\lambda = \mu_i \cdot H_i = \mu \cdot H = \lambda \qquad [54]$$

En la TABLA 4 se compara el sistema $MMn_2 1$ para los parámetros T_{S_1} = 0,12, T_{S_2} = 0,10, y λ = 2 con un sistema MM1 de parámetros T_S=0,054545, es decir el promedio del mínimo entre T_{S_1} y T_{S_2}, e igual tasa de arribos.

3. CONSIDERACIONES FINALES Y CONCLUSIONES

Una extensión natural al enfoque aquí propuesto es el estudio de problemas de colas cuyo tiempo de atención tenga distribución hiperexponencial, tal como el que se verifica en sistemas multiclases (Miranda, 2010). Estos sistemas son aquellos en los cuales la población de clientes no es homogénea, pudiendo clasificarse en categorías de clientes que requieren un tipo de atención diferente.

El presente trabajo apunta a problemas de distribuciones generales en los procesos de atención, pero pueden abordarse también modelos de distribuciones generales para arribos de clientes (GM1), o para ambos procesos (GG1), tal como los estudiados por Ward Whitt quien presentó una fórmula para efectuar un cálculo aproximado del tiempo medio en cola (Whitt, 1991).

El campo de aplicación de los modelos con distribuciones no exponenciales es muy amplio y el enfoque presentado puede utilizarse también en la formulación analítica de modelos para describir procesos de atención masiva con arribos individuales, procesos de arribos masivos con atención individual, procesos de arribos y atención en masa, sistemas con prioridades, sistemas con canales en serie que otorgan el mismo servicio (tales como peajes de doble cabina), sistemas con interrupción de servicios, etc., mediante combinación de distribuciones con pendiente y planas.

Cabe mencionar que el método aquí descripto es aplicable a problemas con procesos de tiempo de tareas markovianos, dado que se basa en la característica de falta de memoria de la distribución exponencial. No es utilizable, en cambio, en problemas con otras formas de distribuciones de tiempo, tales como las tipo Beta.

La suposición de que el servicio de una distribución generada a partir de exponenciales es una distribución exponencial es improcedente, con excepción de la MINX-K. Los resultados comparativos aquí obtenidos así lo demuestran: si bien el factor de utilización del canal (H) y la consecuente probabilidad de que el

sistema esté vacío son coincidentes con los de la distribución generada, los valores de las variables características presentan diferencias considerables (por ejemplo, entre un 9 y un 15%, para el caso del número promedio de clientes).

El método constituye un enfoque alternativo a Pollaczek-Khintchine, presentando las ventajas de poder determinar las variables de eficiencia sin la necesidad de conocer el desvío estándar de la distribución del tiempo de atención (no siempre fácilmente determinable), y de poder desagregar las variables de eficiencia para cada fase del proceso. Por otra parte, este planteamiento comprende la determinación de la distribución de las probabilidades de estado en régimen permanente.

La distribución MAXX-K está muy poco divulgada en la literatura científica. La expresión original [25] aquí presentada para calcular su media resulta de gran utilidad para la utilización del presente método en sistemas cuyos tiempos de servicios respondan a esta distribución.

4. TABLAS

TABLA 1: Comparación entre sistemas $Mh_3$1 y MM1 de igual H

	L	L_C	W	Wc	p(0)	p(1)	p(2)
$MH_3$1	0,42800	0,10400	0,47556	0,11556	0,6760	0,2433	0,06247
MM1	0,47929	0,15529	0,53254	0,17250	0,6760	0,2190	0,07096

TABLA 2: Comparación entre sistemas $ME_3$1 y MM1 de igual H

	L	L_C	W	Wc	p(0)	p(1)	p(2)
$ME_3$1	0,42753	0,10353	0,47503	0,11503	0,6760	0,2435	0,06237
MM1	0,47929	0,15529	0,53254	0,17250	0,6760	0,2190	0,07096

TABLA 3: Comparación entre sistemas $MMx_2$1 y MM1de igual H

	L	L_C	W	Wc	p(0)	p(1)	p(2)
$MMx_2$1	0,45900	0,12808	0,22950	0,06404	0,66901	0,23736	0,06823
MM1	0,49456	0,16366	0,24728	0,08183	0,66901	0,22141	0,07327

TABLA 4: Comparación entre sistemas $MMn_2$1 y MM1de igual H

	L	L_C	W	Wc	p(0)	p(1)	p(2)
$MMn_2$1	0,12245	0,01336	0,06122	0,00668	0,89091	0,09719	0,01060
MM1	0,12245	0,01336	0,06122	0,00668	0,89091	0,09719	0,01060

5. AGRADECIMIENTO

A la ing. Ayelén Barreto por su valiosa contribución en el presente trabajo.

6. REFERENCIAS BIBLIOGRÁFICAS

LITLLE, J. D. C. (1961): "A PROOF FOR THE QUEUING FORMULA: L = λ W", Operations Research, Vol. 9, No.3, pgs.383,387.

LITLLE, J. D. C. (2011): "LITTLE'S LAW AS VIEWED ON ITS 50TH ANNIVERSARY", Operations Research, Vol. 59, No. 3, Mayo–Junio 2011, pgs. 536,549.

MIRANDA, M. (2013): "TEORÍA DE COLAS". 2ª edición. Editorial: EDUCA.

MIRANDA, M. (2010): "SISTEMAS DE COLAS MULTICLASES SIN PRIORIDAD". Anales XXIII ENDIO- XXI EPIO/ENDIO, II ERABIO, pgs. 556, 558.

MIRANDA, M. (2012): "SISTEMAS DE COLAS CON INTERRUPCIÓN DE SERVICIOS", Anales XXV ENDIO- XXIII EPIO/ENDIO, pgs. 389, 407.

MIRANDA, M. (2014): "SISTEMAS DE COLAS CON INTERMISIÓN DE SERVICIOS Y CLIENTES SIN TOLERANCIA". Revista Investigación Operativa de la Escuela de Perfeccionamiento en Investigación Operativa, Nro. 35, Mayo 2014, año XXII.

WHIT, W., MELAMED B. (1990): "ON ARRIVALS THAT SEE TIME AVERAGES". *Operations Research*, vol. 38, No. 1, 1990, pgs. 156,172.