Modelos de Decisión en Ambientes Inciertos

(Apunte de Clases para el Curso Investigación Operativa IN44A)

Departamento de Ingeniería Industrial - Universidad de Chile

René A. Caldentey Susana V. Mondschein [1]

Enero, 1999

[1]La presente es una versión preliminar de este apunte docente, el cual se encuentra en construcción. Los autores agradecen los comentarios y correcciones de eventuales errores que aún permanezcan en el texto, los cuales pueden ser comunicados a smondsch@dii.uchile.cl, rcaldent@mit.edu o hawad@dii.uchile.cl

Capítulo 8

Teoría de Espera

8.1 Introdrucción

El origen de la *Teoría de Espera* o *Teoría de Colas* se encuentra en el año 1909 en una publicación de A. Erlang sobre congestión en el tráfico de llamadas telefónicas. Posteriormente Kendall durante los años 1951-1953 formula en términos más formales la teoría de espera bajo un enfoque de procesos estocásticos. Su trabajo tuvo amplia aceptación y generó en los años posteriores un importante desarrolló a nivel teórico y práctico. Una cola o línea

de espera se forma cuando un conjunto de entidades (personas, productos, documentos,etc) demandan un cierto servicio en un momento dado del tiempo excediendo la capacidad para prestarlo en ese instante. Por ejemplo, consideremos un banco al cual llegan personas a realizar diversos trámites. Si en un momento dado todos los cajeros del banco están ocupados, las personas que sigan llegando formarán una cola en espera de ser atendidas. Los elementos básicos que caracterizan un sistema de espera son los siguientes:

1. Proceso de LLegada: El proceso de llegada de las entidades al sistema representa la forma en que las llegadas ocurren. Usualmente se caracteriza por el tiempo entre llegadas sucesivas, el cual puede ser determinístico en cuyo caso es constante, o bien estocástico en cuyo caso se representa mediante una distribución de probabilidades. El proceso de llegadas también define si las llegadas son individuales o en grupos (batch), en este último caso se especifica la forma en que se constituye el tamaño del batch.

2. Proceso de Atención: El proceso de atención representa la forma en que el servicio es entregado. Lo usual es caracterizarlo mediante el tiempo necesario para completar el servicio. Este tiempo puede ser determinístico, es decir cada entidad demora lo mismo en ser atendida, o bien estocástico en cuyo caso es necesario especificar una distribución de probabilidades. Además debe definir si la atención es individual o en grupo y en este último caso la forma en que se selecciona el tamaño del batch.

3. Número de Servidores: Un sistema puede tener un sólo servidor o varios en paralelo. Un sistema con varios servidores puede tener una cola común, o bien tener para cada

servidor una cola. Si una entidad llega y encuentra más de un servidor desocupado escogerá en forma aleatoria uno para su atención.

4. Capacidad del Sistema: Un sistema de atención puede tener una capacidad infinita, es decir que el tamaño de la cola puede crecer indefinidamente, o bien tener una capacidad finita en cuyo caso el número de entidades en el sistema (o cola) está acotado. Si un sistema tiene capacidad finita y en un momento dado se alcanza, entonces las entidades que sigan llegando no podrán ingresar al sistema y lo abandonarán.

5. Disciplina de Atención La disciplina de atención indica la forma en que se seleccionan las personas de la cola para ser atendidas. Lo usual es que se use un enfoque FIFO, es decir el primero en la cola es el primero en ser atendido. También se pueden usar enfoques LIFO, random o de prioridad.

La formulación matemática de una línea de espera requiere que cada uno de los 5 elementos anterior sean perfectamente conocidos. Mediante dicha formulación se persigue en general responder preguntas relevantes relacionadas con la operación de estos sistemas, como por ejemplo:

- ¿Cuál es el número de entidades en la cola en un instante cualquiera?
- ¿Cuál es el valor esperado del tiempo que una entidad permanece en el sistema?
- ¿Qué fracción del tiempo permanece desocupado el servidor?
- ¿Cuál es el número mínimo de servidores necesarios para que al menos el 95% de las entidades permanezca no más de 12 minutos en el sistema?

Por otro lado, el estudio de sistemas de espera puede realizarse desde dos perspectivas describiendo el comportamiento en estado transiente o en estado estacionario. Un sistema se encuentra en estado transiente si las condiciones iniciales bajo las cuales comenzó su evolución afectan su estado actual, por lo general describir analíticamente el estado transiente de un sistema es difícil y es preferible muchas veces utilizar técnicas de simulación. Un sistema ha alcanzado el estado estacionario si las condiciones de borde iniciales no afecta su estado actual, esta condición se presenta en sistemas que llevan evolucionando mucho tiempo. En este capítulo nos centraremos principalmente en el estudio de sistema de espera en estado estacionario, sin embargo, explicaremos como es posible determinar la conducta transiente en algunos casos especiales. Antes de entrar de lleno en el estudio de los sistemas de espera, se describirá primero la notación introducida por Kendall para clasificarlos y en segundo lugar se deducirán algunas propiedades generales que se satisfacen en estado estacionario.

8.2 Notación

Como se vio anteriormente los sistemas de espera están compuestos por 5 elementos básicos que son (i): Un proceso de llegada de las entidades al sistema, (ii): Un proceso de atención de

las entidades, (iii): Un número de servidores, (iv): Una capacidad para el sistema, (v): Una disciplina de atención. La notación que introdujo Kendall (1951) tiene por objeto simplificar la forma de especificar los 4 primeros elementos señalados. Para ello hace uso del siguiente esquema: $A/B/C/D/E$ en donde los símbolos A, B, C, D, E representan:

1. **A**: Distribución del tiempo entre llegadas sucesivas.
2. **B**: Distribución del tiempo de atención.
3. **C**: Número de servidores en paralelo.
4. **D**: Capacidad máxima del sistema.

El símbolo **E** corresponde a un elemento que no habíamos discutido: indica el tamaño de la población que da origen a las llegadas. Si el tamaño de la población es finito, la distribución del tiempo entre llegadas se verá afectada por el número de entidades que haya en el sistema (mientras más entidades en el sistema hay menos "afuera", i.e. menos llegadas potenciales).

Las cantidades C y D se representan númericamente por su valor en cambio A y B corresponden a distribuciones de probabilidad, los símbolos usados en los casos más comunes son:

- M: para la distribución exponencial (satisface la propiedad Markoviana).
- E_k: para la distribución Erlang-k.
- D: en el caso determinístico.
- G: para una distribución general.

Así por ejemplo la notación $M/M/1$ senála un sistema cuyo proceso de llegada es Poisson (tiempo entre llegadas exponencial), cuyo tiempo de atención es exponencial y que tiene un sólo servidor, si se omite el cuarto símbolo D se entiende que el sistema tiene capacidad infinita ($M/M/1 \equiv M/M/1/\infty$). Un sistema $G/D/k/K$ tiene un proceso de llegada arbitrario, con tiempo de atención determinístico, con k servidores y con capacidad K. En algunos casos se agrega un quinto símbolo que representa el tamaño de la fuente de donde provienen las entidades. Por ejemplo, consideremos un sistema que represente el taller mecánico de una empresa de transporte que dispone de 20 camiones, entonces el tamaño de la fuente de entidades es 20 y por ejemplo el sistema se podría representar por $M/M/3/5/20$.

8.3 Conducta Transiente y Estacionaria

Designemos por $N(t)$ el número de entidades en el sistema (en la cola más las que se están atendiendo) en el instante t y sea

$$p_n(t) = Prob(N(t) = n) \qquad n = 0, 1, 2, 3, ...$$

la distribución de probabilidades de $N(.)$. Determinar el comportamiento en estado transiente del sistema corresponde a encontrar $p_n(t) \quad \forall n, t$ lo que en general puede llegar a ser muy difícil. En muchas aplicaciones prácticas sin embargo, se necesita conocer la conducta de equilibrio del sistema, es decir, cuando el sistema lleva operando un tiempo suficientemente largo y las condiciones iniciales ya no influyen en la evolución del sistema. En otras palabras lo que se busca es determinar:

$$p_n = \lim_{t \to \infty} p_n(t) \qquad n = 0, 1, 2, ...$$

que representa en el largo plazo la fracción del tiempo que el sistema a contenido n entidades. No siempre el límite anterior existe y es necesario determinar bajo que condiciones es posible encontrar p_n. Estas condiciones se discutirán más adelante. Si el límite anterior existe $\forall n$ se dice que el sistema alcanza un estado estacionario y p_n se conoce como la probabilidad estacionaria de encontrar n entidades en el sistema.

8.3.1 Fórmulas de Conservación, Fórmula de Little

Existen algunas relaciones en la teoría de colas que se satisfacen bajo condiciones bastantes generales, las cuales se basan principalmente en principios de *Conservación* en estado estacionario. Una de las más importantes es

$$L = \lambda \cdot W$$

donde λ es la tasa promedio de llegada de entidades al sistema, L es el número promedio de entidades en el sistema y W es el tiempo promedio de permanencia de una entidad en el sistema en estado estacionario. En forma equivalente denotando el número promedio de entidades en la cola y el tiempo promedio de permanencia de una entidad en la cola en estado estacionario por L_Q y W_Q respectivamente se tiene que

$$L_Q = \lambda \cdot W_Q$$

Esta relación se conocía desde hace ya mucho tiempo, sin embargo fue recién en 1961 que Little dio una prueba formal de ella y es por ello que se conocen como *Fórmula de Little*. Una forma intuitiva para justificar la fórmula de Little es observando que para un sistema cualquiera en estado estacionario la tasa de entrada de las entidades al sistema debe ser igual a la tasa de salida (conservación del flujo de entidades a través del sistema). De no ser así, o bien el sistema se estaría llenando (tasa de entrada mayor que tasa de salida), o bien el sistema se estaría vaciando (tasa de entrada menor que tasa de salida). En cualquiera de los dos casos no existiría estado estacionario que es condición necesaria para la fórmula de Little. La tasa de entrada al sistema es simplemente λ. Para determinar la tasa de salida basta observar lo siguiente. Si una entidad llega al sistema en estado estacionario encontrará, en promedio, que junto con ella hay L entidades en el sistema, además ella dejará el sistema, en promedio, en W unidades de tiempo por lo tanto encuentra que existe un flujo de salida de L entidades en W unidades de tiempo es decir una tasa de $\frac{L}{W}$ entidades por unidad de tiempo. Finalmente, igualando λ con $\frac{L}{W}$ se obtiene el resultado. Una demostración simple

de la fórmula de Little es la dada por Eilon (1961) (ver apéndice VII). La importancia de la fórmula de Little radica principalmente en lo general que es. Para cualquier sistema que alcanza un comportamiento estacionario se cumple $L = \lambda \cdot W$ independiente del número de servidores, de la capacidad del sistema, de los tiempos de atención, etc. Algunas relaciones adicionales que se pueden deducir de la fórmula de Little son las siguientes:

1. $L = L_Q + L_S$
2. $W = W_Q + W_S$
3. $L_Q = \lambda \cdot W_Q$
4. $L_S = \lambda \cdot W_S$

en donde L_S y W_S son el número promedio de entidades siendo atendidas y el tiempo promedio de atención de una entidad respectivamente. Por otro lado, el principio de conservación del flujo de clientes en estado estacionario permite determinar una condición necesaria para la existencia de estado estacionario. Tomemos un ejemplo, consideremos un sistema $G/G/c$ en donde la tasa media de llegada es λ, cada servidor tiene una tasa promedio de atención μ y existe una cola única. Supongamos que existe estado estacionario y sea $\{p_n\}$ el conjunto de probabilidades estacionarias del sistema. De esta forma, la tasa promedio de salida del sistema se calcula como

$$\sum_{k=0}^{\infty} min(k, c) \cdot \mu \cdot p_k \tag{8.1}$$

Ahora bien, igualando la tasa de entrada con la tasa de salida se tiene

$$\begin{aligned} \lambda &= \sum_{k=0}^{\infty} min(k, c) \cdot \mu \cdot p_k \\ \frac{\lambda}{c \cdot \mu} &= \sum_{k=0}^{\infty} min(\frac{k}{c}, 1) \cdot p_k \\ &< \sum_{k=0}^{\infty} p_k = 1 \end{aligned} \tag{8.2}$$

Luego para que exista estado estacionario es necesario que $\frac{\lambda}{c \cdot \mu} < 1$. El resultado es intuitivo si se piensa que la tasa máxima de atención del sistema se alcanza cuando están todos los servidores ocupados y en este caso en promedio vale $c \cdot \mu$. Por lo tanto, para que el sistema no se congestione la tasa de entrada tiene que ser menor que la mayor tasa de salida, de donde se obtiene el resultado anterior. El término $\frac{\lambda}{c \cdot \mu}$ se conoce como intensidad de tráfico y se suele denotar por ρ. Es fácil ver que el número promedio de servidores ocupados en estado estacionario es $c \cdot \rho$. En el caso particular que $c = 1$ igualar la tasa de entrada con la tasa de salida equivale a

$$\lambda = \mu \cdot (1 - p_0) \tag{8.3}$$

es decir, $p_0 = \frac{\lambda}{\mu} = \rho$ que corresponde a la probabilidad estacionaria de encontrar al sistema vacío o equivalentemente a la fracción del tiempo que el servidor está desocupado (tiempo oscioso).

8.4 Teoría de Espera en modelos exponenciales de Nacimiento y Muerte

Muchos problemas de líneas de espera son susceptibles de modelarse como procesos de nacimiento y muerte. Es decir, que el sistema de evolucionar lo hace a estados vecinos. Denotando el estado k como aquel en el cual el sistema tiene k entidades entonces para que el sistema evolucione sólo a estados vecinos en necesario que las llegadas y las atenciones sean individuales. En esta sección nos preocuparemos de estudiar sistemas de espera con la característica anterior y que además presenta un comportamiento exponencial tanto en la llegada como en la atención, es decir, sistemas del tipo $M/M/...$, además el estudio se centrará en el estado estacionario de estos sistemas. Como se vio en el capítulo anterior, un

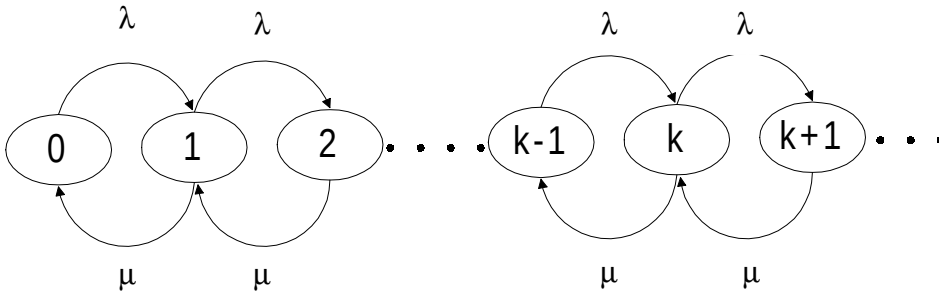
proceso de nacimiento y muerte con tiempos de permanencia en cada estado exponenciales es un tipo especial de cadena de Markov en tiempo continuo y por tanto se cumple, en estado estacionario, que *la tasa a la cual el sistema entra en un estado k cualquiera es igual a la tasa a la cual lo abandona.* A continuación veremos como este resultado es suficiente para determinar el comportamiento en estado estacionario de los sistema de interés, entre los que se encuentran $M/M/1$, $M/M/c$, $M/M/c/k$, $M/M/\infty$, entre otros.

8.4.1 Sistema $M/M/1$

Consideremos un sistema para el cual las entidades llegan de acuerdo a un proceso Poisson de tasa λ, existe un único servidor que atiende a un entidad a la vez y se demora un tiempo aleatorio exponencialmente distribuido con tasa μ. Supongamos además que el sistema tiene capacidad infinita y que la disciplina de atención es FIFO. El número de entidades en este sistema se puede modelar como una cadena de markov en tiempo continuo, más específicamente como un proceso de nacimiento y muerte, con el grafo representante que se muestra en la Figura 8.1:

Si aplicamos el principio de igualdad de tasas en los distintos estados del sistema se tiene que:

- **Estado 0**: $\lambda p_0 = \mu p_1$
- **Estado 1**: $\lambda p_1 + \mu p_1 = \lambda p_0 + \mu p_2$
- ....



- **Estado** $k > 1$: $\lambda p_k + \mu p_k = \lambda p_{k-1} + \mu p_{k+1}$

El sistema anterior más la condición de normalización $\sum_k p_k = 1$ determinan completamente la solución $\{p_k\}$. Para resolver el sistema anterior basta observar que:

$$\begin{aligned} \lambda p_k - \mu p_{k+1} &= \lambda p_{k-1} - \mu p_k \\ &= \lambda p_{k-2} - \mu p_{k-1} \\ &= ..... \\ &= \lambda p_0 - \mu p_1 = 0 \end{aligned} \tag{8.4}$$

luego

$$p_{k+1} = (\lambda \; over \mu) p_k = (\frac{\lambda}{\mu})^2 p_{k-1} = ... = (\frac{\lambda}{\mu})^{n+1} p_0 \tag{8.5}$$

usando $\sum_{k=0}^{\infty} p_k = 1$ se tiene que

$$p_0 \cdot \sum_{k=0}^{\infty} (\frac{\lambda}{\mu})^k p_k = 1 \tag{8.6}$$

la ecuación anterior admite solución si sólo si $\rho = \frac{\lambda}{\mu} < 1$, en dicho caso se tiene que

$$p_0 = 1 - \frac{\lambda}{\mu} = 1 - \rho \tag{8.7}$$

y en general

$$p_k = (1 - \rho) \cdot \rho^k \tag{8.8}$$

es decir tiene una distribución geométrica. Anteriormente vimos que la condición $\lambda < \mu$ era necesaria para la existencia de probabilidades estacionarias en un sistema $G/G/1$. Ahora bien, del desarrollo anterior vemos que para el caso $M/M/1$ la condición es además suficiente.

Medidas de Efectividad

El estudio práctico de un sistema de espera no se limita a conocer cual es la distribución de probabilidades estacionarias, en lo que realmente se está interezado es en determinar las

medidas de efectividad del sistema como por ejemplo número promedio de entidades, tiempo promedio de permanencia en el sistema, fracción del tiempo que el servidor está ocupado, etc. Estas medidas de efectividad reflejan el funcionamiento del sistema y permiten apoyar decisiones sobre el manejo de los sistemas de espera.Por ejemplo si al momento de diseñar una cola se considera un sólo servidor y se calcula que en estas condiciones el 98% del tiempo el servidor estará ocupado entonces parece razonable pensar en aumentar el número de servidores a utilizar, la conclusión sería distintas si se hubiese obtenido una fracción de ocupación del 40%. Sea N la variable aleatoria número de entidades en el sistema y W la

variable aleatoria tiempo de permanecia en el sistema. En forma análoga se pueden definir N_Q y W_Q para la cola. El número promedio de entidades en el sistema se calcula como [1]

$$L = E(N) = \sum_{k=0}^{\infty} kp_k = \sum_{k=1}^{\infty} k(1-\rho)\rho^n = \frac{\rho}{1-\rho} \tag{8.9}$$

Utilizando la formula de Little se tiene que

$$W = \frac{L}{\lambda} = \frac{\rho}{\lambda \cdot (1-\rho)} = \frac{1}{\mu - \lambda} \tag{8.10}$$

Además, es directo que $W_S = \frac{1}{\mu}$ y por tanto es posible calcular $W_Q = W - W_S$, es decir

$$W_Q = \frac{1}{\mu - \lambda} - \frac{1}{\mu} = \frac{\lambda}{\mu \cdot (\mu - \lambda)} = \rho \cdot W \tag{8.11}$$

Finalmente conocido W_Q, L_Q se obtiene usando Little

$$L_Q = \lambda \cdot W_Q = \frac{\lambda^2}{\mu \cdot (\mu - \lambda)} = \rho \cdot L \tag{8.12}$$

Las medidas de efectividad anteriormente calculadas representan los valores medios del número de entidades en el sistema o en la cola (L, L_Q) y también del tiempo de permanecia en el sistema o en la cola (W, W_Q). Si bien esta información es de por sí es muy útil para el estudio de un sistema, puede ser insuficiente por no entregar un medida de la variabilidad. Es necesario conocer también como es la varianza de N y W por ejemplo para tener una visión más completa del funcionamiento del sistema. Varianza de N

La varianza del número de entidades en el sistema $var(n)$ se puede calcular fácilmente usando la relación:

$$var(N) = E(N^2) - [E(N)]^2 \tag{8.13}$$

Dejamos propuesto al lector chequear que:

$$E(N^2) = \frac{\rho^2 + \rho}{(1-\rho)^2} \tag{8.14}$$

y que por tanto $var(N) = \frac{\rho}{(1-\rho)^2}$. Se puede ver tanto de $E(N)$ como de $var(N)$ que el comportamiento del sistema se hace más inestable cuando $\rho \to 1$. En estos casos el número

[1] El resultado se obtiene facilmente utilizando la relación $\sum_n n\rho^n = \rho \cdot \frac{d}{d\rho}(\sum_n \rho^n)$.

promedio de entidades en el sistema es muy grande al igual que la varianza lo que implica que el número de entidades observadas en el sistema en un instante cualquiera tiene una probabilidad alta de ser muy diferente del valor medio. Varianza de W
Para poder calcular $var(W)$ es necesario conocer cuál es la distribución de probabilidades de esta variable. Para ello observemos los siguiente, supongamos que una entidad E llega al sistema y encuentra n entidades en su interior, entonces el tiempo que deberá permanecer antes de salir W_n puede escribirse como:

$$W_n = v_1 + v_2 + ... + v_{n+1} \tag{8.15}$$

en donde v_1' representa el tiempo residual de atención de la entidad que está siendo atendida cuando llega E, v_2,...,v_n representan los tiempos de atención de las $n-1$ entidades en la cola al momento de llegar E y v_{n+1} es el tiempo de atención de E. Como el proceso de atención es exponencial entonces v_2,...,v_{n+1} son claramente variables aleatorias independientes exponencialmente distribuidas, más aún la falta de memoria de la distribución exponencial permite asegurar que v_1 también está exponencialmente distribuida. Por lo tanto, W_n es la suma de $n+1$ variables aleatorias exponecial independientes de tasa μ por lo que W_n tiene una distribución gamma de parámetros μ y $n+1$ y su función de densidad viene dada por:

$$f_{W_n}(w) = \frac{\mu^{n+1} \cdot w^n \cdot e^{-\mu \cdot w}}{\Gamma(n+1)} \tag{8.16}$$

Sea por otro lado $f_W(w)$ la función de densidad del tiempo de permanencia en el sistema. Entonces la probabilidad que una entidad permanezca en el sistema un tiempo comprendido entre $[w, w+dw]$ viene dada por:

$$Prob(w \leq W \leq w+dw) = f_W(w) \cdot dw \tag{8.17}$$

Además la probabilidad anterior puede reescribirse condicionándola al número de entidades que nuestra entidad de prueba E detecta al llegar como:

$$\begin{aligned} Prob(w \leq W \leq w+dw) &= \sum_{n=0}^{\infty} Prob(w \leq W_n \leq w+dw) \cdot p_n \\ &= \sum_{n=0}^{\infty} \frac{\mu^{n+1} \cdot w^n \cdot e^{-\mu \cdot w}}{\Gamma(n+1)} \cdot (1-\rho)\rho^n \cdot dw \\ &= \mu(1-\rho)e^{-\mu \cdot w} \sum_{n=0}^{\infty} \frac{(\mu w \rho)^n}{\Gamma(n+1)} dw \\ &= \mu(1-\rho)e^{-\mu(1-\rho)w} dw \end{aligned} \tag{8.18}$$

Luego igualando 8.17 con 8.18 se tiene que

$$f_W(w) = \mu(1-\rho)e^{-\mu(1-\rho)w} \tag{8.19}$$

Por lo tanto W sigue una distribución exponencial de tasa $\mu(1-\rho)$. Luego la varianza de W viene dada por $var(W) = \frac{1}{(\mu(1-\rho))^2}$ y nuevamente los problemas de variabilidad para W se alcanzan cuando $\rho \to 1$.

8.4.2 Sistema M/M/c

Consideremos un sistema de espera cuyo proceso de llegada es poissoniano de tasa λ y que dispone de c servidores en paralelo, teniendo cada uno un tiempo de atención exponencialmente distribuido con tasa μ.

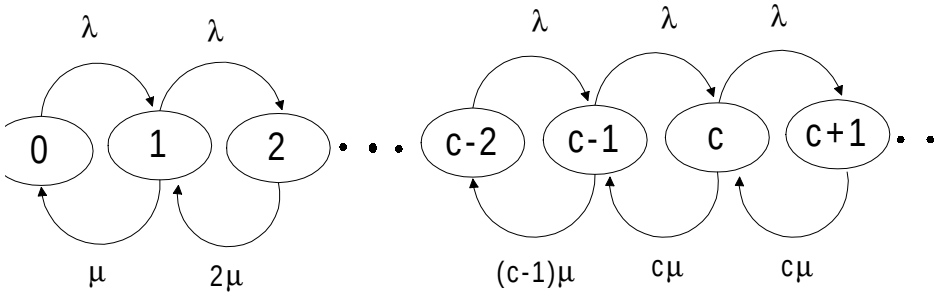
Probabilidades Estacionarias

El sistema $M/M/c$ corresponde a un sistema de nacimiento y muerte para el cual las tasa de transición entre estados no son siempre iguales. En efecto si existen n entidades en el sistema y $n < c$ entonces sólo n servidores estarán ocupados y el tiempo entre dos salidas consecutivas del sistema está exponencialmente distribuidas con tasa $n \cdot \mu$. Si se tiene en cambio que $n \geq c$ entonces todos los servidores estarán ocupados y el tiempo entre salidas consecutivas del sistema está exponencialmente distribuido con tasa $c \cdot \mu$. Por lo tanto, el sistema $M/M/c$ es un proceso de nacimiento y muerte con tasa de nacimiento constante $\lambda_n = \lambda$ y con tasa de muerte μ_n dependiente del estado n del sistema con:

$$\mu_n = n \cdot \mu \qquad n = 0, 1, 2, ..., c-1$$

$$\mu_n = c \cdot \mu \qquad n = c, c+1, ...$$

El grafo representante se muestra en la Figura 8.2:



Suponiendo que el estado estacionario existe entonces la probabilidad estacionaria de que el sistema se encuentre en el estado n (p_n) viene dada por:

$$p_n = \frac{\Pi_{i=0}^{n-1} \lambda_i}{\Pi_{i=1}^{n} \mu_i} p_0 \tag{8.20}$$

es decir,

$$\begin{aligned} p_n &= \frac{\lambda^n}{n! \cdot \mu^n} p_0 \quad n = 0, 1, 2, ..., c-1 \\ p_n &= \frac{\lambda^n}{c! \cdot c^{n-c} \cdot \mu^n} p_0 \quad n = c, c+1, ... \end{aligned} \tag{8.21}$$

Utilizando el resultado anterior y el hecho que $\sum_{n=0}^{\infty} p_n = 1$ se obtiene

$$p_0 = \left[\sum_{n=0}^{c-1} \frac{\lambda^n}{n! \cdot \mu} + \frac{\lambda^c}{c! \cdot \mu^c \cdot (1 - \frac{\lambda}{c\mu})} \right]^{-1} \tag{8.22}$$

El resultado anterior asume que se satisface la condición $\lambda < c\mu$ que es condición necesaria y suficiente para la existencia del estado estacionario para un sistema $M/M/c$. Una pregunta interesante en este modelo es la que tiene relación con la probabilidad que una entidad que llega al sistema tenga que esperar para ser atendidad o equivalentemente que encuentre a los c servidores ocupados.

$$C \equiv Prob(N \geq c) = \sum_{n=c}^{\infty} p_n = \frac{p_c}{1 - \rho}$$

donde $\rho = \frac{\lambda}{c\mu}$. La expresión anterior se conoce como la fórmula C de Erlang y aparece tabulada para valores diferentes de c y $\frac{\lambda}{\mu}$.

Medidas de Efectividad

Número Esperado de Servidores Ocupados

Una medida de efectividad para estudiar el dimensionamiento en términos del número de servidores a utilizar es el número promedio de servidores que están ocupados en un momento dado. Mientras más cercano es este valor al número de servidores totales mayor será la tasa de ocupación de estos. A partir de las relaciones de Little se tiene que $L = \lambda W$ y $L_Q = \lambda W_Q$ y por tanto $L - L_Q = \lambda(W - W_Q)$. El término $L - L_Q$ representa el número promedio de entidades que están siendo atendidas en un momento dado del tiempo y es exactamente igual al número de servidores ocupados en ese momento. Por otro lado, $W - W_Q$ representa el tiempo de atención de una entidad el cual viene dado por $\frac{1}{\mu}$. Por lo tanto, el número promedio de servidores ocupados es igual a $\frac{\lambda}{\mu} = c\rho$. Número Esperado de entidades en el sistema

El número esperado de entidades en el sistema se puede calcular como la suma del número esperado de entidades siendo atendidas más el número esperado de entidades en la cola. Del punto anterior el número esperado de entidades siendo atendidas es $c\rho$. Por otro lado, el número esperado de entidades en la cola viene dado por:

$$L_Q = \sum_{n=c}^{\infty} (n - c) \cdot p_c$$

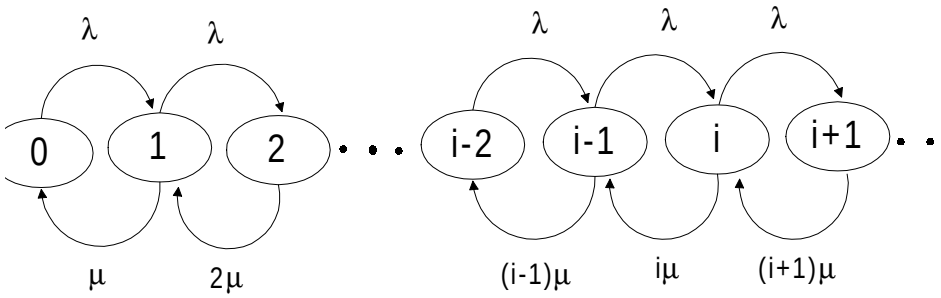
$$
\begin{aligned}
&= \sum_{n=c}^{\infty}(n-c)\frac{\lambda^n}{\mu^n \cdot c! \cdot c^{n-c}}p_0 \\
&= \frac{\rho}{1-\rho}Prob(N \geq c) \qquad (8.23)
\end{aligned}
$$

Por lo tanto, el número esperado de entidades en el sistema es igual a $L = c\rho + \frac{\rho C}{1-\rho}$ con $C = Prob(N \geq c)$. A partir de la fórmula de Little se pueden deducir facilmente los valores de W y W_Q.

$$
\begin{aligned}
W_Q &= \frac{L_Q}{\lambda} = \frac{C}{c \cdot \mu \cdot (1-\rho)} \\
W &= \frac{L}{\lambda} = \frac{1}{\mu} + W_Q \qquad (8.24)
\end{aligned}
$$

8.4.3 Sistema M/M/∞

Consideremos un sistema con llegadas poissonianas, tiempos de atención exponenciales y que tenga un número ilimitado de servidores. Este tipo de modelos se ajusta bien a situaciones de autoservicio, es decir, en donde cada entidad que llega al sistema se proporciona el servicio. Este sistema puede modelarse como un proceso de nacimiento y muerte con el grafo representante que se muestra en la Figura 8.3:



Las probabilidades estacionarias vienen dadas en este caso por:

$$p_n = \frac{\lambda^n}{\mu^n \cdot n!}p_0 \qquad (8.25)$$

Luego imponiendo que $\sum_{n=0}^{\infty} p_n = 1$ se tiene que:

$$p_n = \frac{\lambda^n}{\mu^n \cdot n!}e^{\frac{-\lambda}{\mu}} \qquad \forall n = 0, 1, 2, ... \qquad (8.26)$$

Es decir, el número de personas en el sistema tiene una distribución Poisson de media $L = \frac{\lambda}{\mu}$. El tiempo promedio de permanencia de una entidad en este sistema es claramente $\frac{1}{\mu}$, ya al exsistir capacidad ilimitada en la atención no se forma cola y el tiempo en el sistema es igual al tiempo de atención cuya media es $\frac{1}{\mu}$.

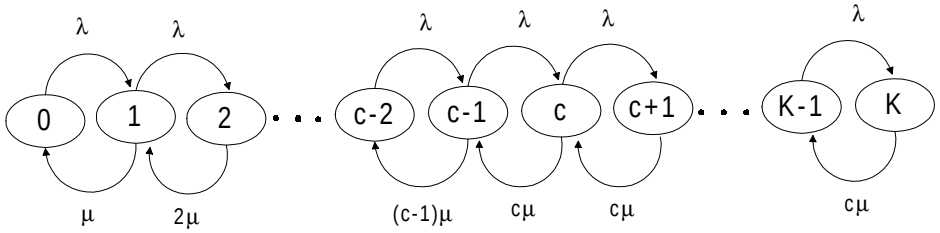
8.4.4 Sistema M/M/c/K

Consideremos un sistema de espera cuyo proceso de llegada es Poisson de tasa λ y que dispone de c servidores en paralelo, teniendo cada uno un tiempo de atención exponencialmente distribuido con tasa μ. El sistema tiene capacidad finita, no pudiendo haber más de K entidades simultáneamente en el sistema (i.e. sólo hay $K - c$ lugares de espera). Podemos asumir $c \leq K$ pues si $c > K$ el sistema se comporta igual que un $M/M/K/K$.

Probabilidades Estacionarias

El sistema $M/M/c/K$ corresponde a un sistema de nacimiento y muerte en el cual las tasas de transición entre estados son iguales a las del sistema $M/M/c$ para los estados $0, 1, \ldots, K$ y la tasa de transición a estados con índice mayor que K es nula. Si existen n entidades en el sistema y $n < c$ entonces sólo n servidores estarán ocupados y el tiempo entre dos salidas consecutivas del sistema (muertes) está exponencialmente distribuidas con tasa $\mu_n = n \cdot \mu$. Si se tiene en cambio que $n \geq c$ entonces todos los servidores estarán ocupados y el tiempo entre salidas consecutivas del sistema está exponencialmente distribuido con tasa $\mu_n = c \cdot \mu$. Por otro lado, si hay $n < K$ entidades en el sistema la tasa de nacimiento es $\lambda_n = \lambda$, mientras que si hay $n = K$ entidades la tasa de nacimiento es $\lambda_n = 0$. El grafo representante se muestra en la Figura 8.4:

Figura 8.4: Grafo representante de una cola M/M/C/K



Como vemos, la evolución del número de entidades en el sistema puede ser representado mediante una cadena de Markov en tiempo continuo irreducible y *finita*, de modo que con seguridad existe una ley de probabilidades estacionarias, no importa cuál sea la relación entre λ, μ, cyK (asumiendo $\lambda > 0, \mu > 0, 0 < c \leq K < \infty$).

La probabilidad estacionaria de tener n entidades en el sistema, p_n viene dada por:

$$p_n = \Pi_{i=0}^{n-1} \frac{\lambda_i}{\mu_{i+1}} p_0 \tag{8.27}$$

es decir,

$$p_n = \begin{cases} \frac{\lambda^n}{n! \cdot \mu^n} p_0 & n < c \\ \\ \frac{\lambda^n}{c! \cdot c^{n-c} \cdot \mu^n} p_0 & c \leq n \leq K \end{cases} \tag{8.28}$$

Utilizando el resultado anterior y el hecho que $\sum_{n=0}^{\infty} p_n = 1$ se obtiene

$$p_0 = \left[\sum_{n=0}^{c-1} \frac{\lambda^n}{n! \cdot \mu} + \frac{\lambda^c \cdot c^c}{c! \cdot \mu^c} \cdot \frac{1 - \left(\frac{\lambda}{c\mu}\right)^{K+1}}{1 - \frac{\lambda}{c\mu}} \right]^{-1} \tag{8.29}$$

Medidas de Efectividad

Entre las medidas de efectividad para este sistema resulta de especial interés la tasa de pérdida de entidades en el largo plazo. Una entidad que llega cuando el sistema está lleno no puede entrar, y se retira sin haber recibido servicio. Diremos que esa entidad se ha perdido. Dado que el proceso de llegadas es Poisson, la probabilidad que una entidad que llega encuentre el sistema lleno es igual a p_K, la probabilidad estacionaria que el sistema esté lleno (recordar "PASTA"). Así, en el largo plazo una fracción p_K de las entidades encuentra el sistema lleno, y se pierden, de modo que la tasa de pérdida de entidades es igual a λp_K [entidades/u. de tiempo].

Podemos calcular el número medio de entidades en el sistema como

$$L = \sum_{i=0}^{K} i \cdot p_i$$

. Una vez conocido L podemos calcular el tiempo medio que pasa una entidad en el sistema, W, a partir de la Fórmula de Little. Sin embargo se debe tener cuidado al aplicar aquí la fórmula de Little: hay que tomar en cuenta que al sistema sólo entran las entidades que no lo encuentran lleno, de manera que la tasa efectiva de entrada al sistema es $\lambda_{ef} = \lambda(1 - p_K)$ [entidades/u. de tiempo]. De esa forma se tiene

$$W = \frac{L}{\lambda_{ef}} = \frac{L}{\lambda(1 - p_K)}$$

.

8.4.5 Sistema $M/M/1/\infty/N$

Consideramos ahora un sistema en que las llegadas provienen de una población finita, de N entidades. Una entidad que está fuera del sistema llegará a él en un tiempo exponencialmente distribuido de media $1/\lambda$ [u. de tiempo], independiente de las demás. Los tiempos de atención son variables aleatorias i.i.d con distribución exponencial con media $1/\mu$ [u. de tiempo].

Probabilidades Estacionarias

Este sistema es susceptible de ser modelado como un proceso de nacimiento y muerte. Cuando hay i entidades en el sistema hay $N - i$ fuera de él. El tiempo que transcurre hasta la llegada de la próxima entidad es el mínimo de los tiempos de llegada de cada una de las $N - i$ entidades que están fuera del sistema, y como todos ellos son exponenciales de tasa λ, la tasa de nacimientos es $(N - i)\lambda$, para $0 \leq i < N$. Cuando hay N entidades en el sistema la tasa de nacimientos es nula (no hay ninguna entidad fuera que pueda llegar al sistema). La tasa de muerte es constante e igual a μ mientras haya un número positivo de entidades en el sistema. El grafo representante se muestra en la Figura 8.5.