2 Teoría de colas o líneas de espera

El tráfico en redes se puede modelar con la ayuda de la teoría de colas, es por ello que es importante estudiarlas y comprenderlas. Existen varias definiciones sobre la teoría de colas, una de ellas y de suma importancia es la que menciona Jaime Enrique Varela en el libro Introducción a la Investigación de Operaciones, ya que indica que la teoría de colas se ocupa del análisis matemático de los fenómenos de las líneas de espera o colas. Además, menciona que las colas se presentan con frecuencia cuando se solicita un servicio por parte de una serie de clientes y tanto el servicio como los clientes son de tipo probabilístico.

La teoría de colas es únicamente un modelo del comportamiento del tráfico que se ve todos los días, como lo puede ser un semáforo, la espera en un banco, la fila para conseguir el ticket para un concierto, así como el tráfico que se presenta en el envío

de paquetes en redes, siendo este último caso el que se va a analizar. La teoría de colas presenta un panorama del comportamiento de la cola a través del tiempo y el entorno de la misma.

Existen varios tipos de colas que se mencionarán a lo largo del capítulo, sin embargo se hará hincapié en tres casos especiales que son el cimiento del modelo, los modelos a estudiar serán el M/M/1, M/M/1/K y M/M/C que se describirán posteriormente.

2.1 Conceptos básicos del modelo de colas

Un ejemplo de una cola es: cuando se va a comprar un boleto para viajar, si existen pocas personas para ser atendidas, será una cola pequeña; sin embargo, si hay un gran número de personas esperando ser atendidas será una cola muy grande. Ahora bien, el número de servidores dependerá de cuantas personas están atendiendo y el cliente será la persona que quiere comprar el boleto, el número de servidores podrá ser de 1 hasta infinito. A continuación se muestra el ejemplo de una cola con un único servidor.

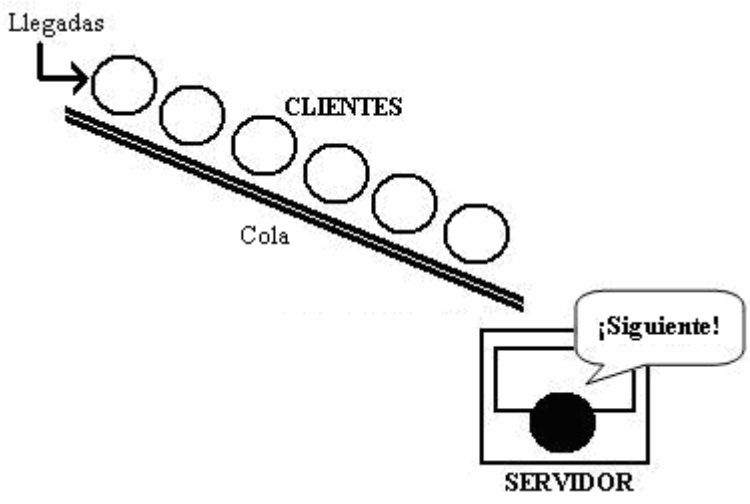

Figura 2-1 Modelo de una sola cola con un único servidor.

Ahora se muestra el mismo ejemplo pero con más servidores.

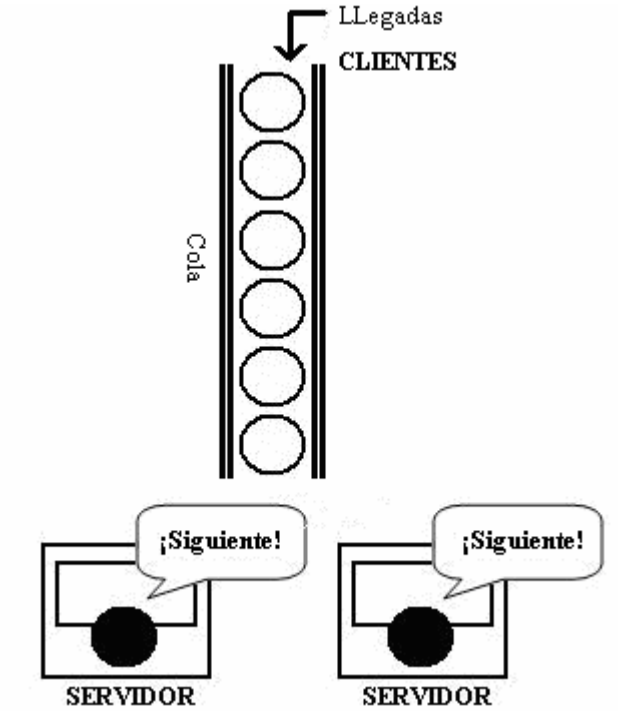

Figura 2-2 Modelo de una sola cola con dos servidores.

Un sistema de colas se especifica por seis características principales [8].

1. El tipo de distribución de entradas o llegadas (tiempo entre llegadas)
2. El tipo de distribución de salidas o retiros (tiempo de servicio)

3. Los canales de servicio
4. La disciplina del servicio
5. El número máximo de clientes permitidos en el sistema
6. La fuente o población

Una vez mencionadas las características de las colas, es importante comentar cada una de ellas. Para empezar, las distribuciones de entrada y salida, también conocidas como distribuciones de llegada y retiro, determinan los modelos por los cuales los clientes entran y salen. En la característica 1 y 2, como puede observarse, se le hace referencia a lo que es el tiempo entre llegadas y el tiempo de servicio, éstos también son conocidos como patrones.

El patrón de llegadas de los clientes generalmente está especificado por el tiempo entre llegadas, que es el tiempo entre las llegadas de los clientes sucesivos a la instalación que ofrece el servicio [9]. En esta parte es importante indicar que a veces los clientes prefieren no esperar en la cola para recibir el servicio y es cuando se presentan dos casos, los cuales son el rechazo y el abandono, el primero ocurre cuando el cliente observa una cola demasiado grande o larga y prefiere no ingresar a ella, el segundo caso se presenta cuando un usuario se encuentra en la cola pero prefiere dejarla.

Generalmente el patrón de servicio está especificado por el tiempo de servicio, que es el tiempo que le toma a un servidor atender a un cliente [9]. En esta parte es importante determinar si un servidor atiende por completo a un cliente o si el cliente

requiere una secuencia de servidores [9]. Para esta parte del trabajo se considerará siempre que un solo servidor está atendido a un solo usuario.

El canal de servicio es el proceso o sistema que está efectuando el servicio para el cliente [10]. De manera complementaria, el canal de servicio puede ser un canal en serie, paralelo o mixto, es decir una combinación de ambas. La diferencia entre el canal en serie y el paralelo es el número de clientes que pueden ser atendidos de manera simultánea. Así pues, se pueden atender varios clientes al mismo tiempo en un canal paralelo, sin embargo en un canal en serie los clientes tendrán que pasar por todos lo canales hasta obtener el servicio.

La disciplina de servicio es una regla para seleccionar clientes de la línea de espera al inicio del servidor [8]. Una de las disciplinas más utilizadas es la denominada “First In First Out”, FIFO, en la cual los primeros que llegan serán los primeros en salir; otra disciplina es la denominada “Last In First Out”, LIFO, en la cual los últimos en llegar serán los primeros en salir. Existen otras disciplinas denominadas *al azar* y *de prioridad*, sin embargo para este trabajo se utilizará únicamente la disciplina de servicio FIFO.

El parámetro mencionado anteriormente como el número máximo de clientes permitidos, es el cupo de clientes permitidos en una cola dependiendo de las características que presenta el sistema; es decir, de acuerdo a las características del sistema se podrá tener una cola infinita o finita. Si una cola es infinita no hay problema en que lleguen mil clientes ya que los mil clientes podrán ser atendidos; mientras que en una cola finita hay un cupo máximo o límite y cuando la cola se

encuentre llena los demás clientes serán rechazados. Este caso en específico se le conoce como caso de frustración.

Finalmente, la fuente (o población) representa un factor importante en el análisis de teoría de colas ya que el modelo de llegadas depende de la fuente de donde provienen los clientes. La fuente que genera las llamadas puede ser finita o infinita. Existe una fuente finita cuando una llegada afecta la tasa de llegadas de futuros clientes potenciales [8]. Así pues, la cola se puede ver de la siguiente manera:

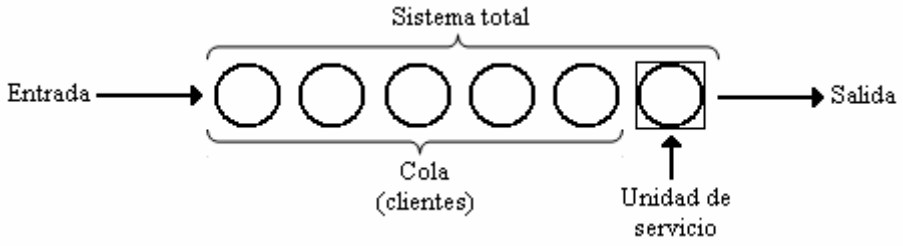

Figura 2-3 Elementos principales de un sistema de colas [10].

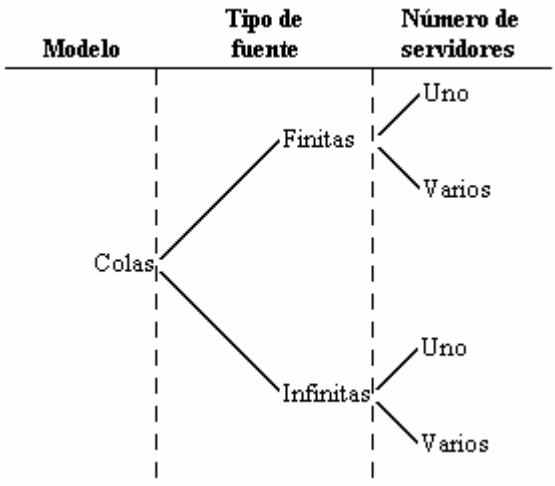

Figura 2-4 Clasificación del modelo de colas.

2.2 Notación de Kendall

En el año de 1953 el matemático David G. Kendall, originario de Inglaterra, implementó la notación de colas, la cual es utilizada para identificar las características de una línea de espera por medio de iniciales. En el sub-capítulo anterior se describieron las características de las colas, en este se aclarará cada inicial.

Un sistema podrá ser notado de la siguiente manera, A/B/X/Y/Z/V, donde [10]:

- A es el modelo de llegadas, valores posibles:
    - M=tiempos entre llegadas exponenciales
    - D=tiempos entre llegadas deterministas
    - G=tiempos entre llegadas generales (cualquier distribución)
- B es el modelo de servicio , puede tomar los mismos valores que A
- X es el número de servidores
- Y es la capacidad del sistema (número máximo de clientes en el sistema), se puede omitir si es infinita.
- Z es la disciplina, se puede omitir si es FIFO
- V es el número de estados de servicio, se puede omitir si es 1

2.3 Tipos de sistemas

Un sistema de líneas de espera es un conjunto de clientes, un conjunto de servidores, y un orden en el cual los clientes llegan y son atendidos. Un sistema de líneas de espera es un proceso de nacimiento-muerte con una población formada por clientes en espera del servicio o que están en servicio; una muerte ocurre cuando un cliente

abandona la instalación. El estado del sistema es el número de clientes en la instalación [9].

En la figura 2-5 se muestran los tipos de sistemas existentes, donde se describe para cada caso qué tipo de sistema es. Es importante mencionar nuevamente que se estudiarán los sistemas M/M/1, M/M/1/K y M/M/C, los cuales se pueden observar en los dos primeros casos; sin embargo, los sistemas más complejos se pueden resolver teniendo como base éstos, pero en muchos casos no es posible resolverlos analizándolos matemáticamente y se analizan por medio de su comportamiento.

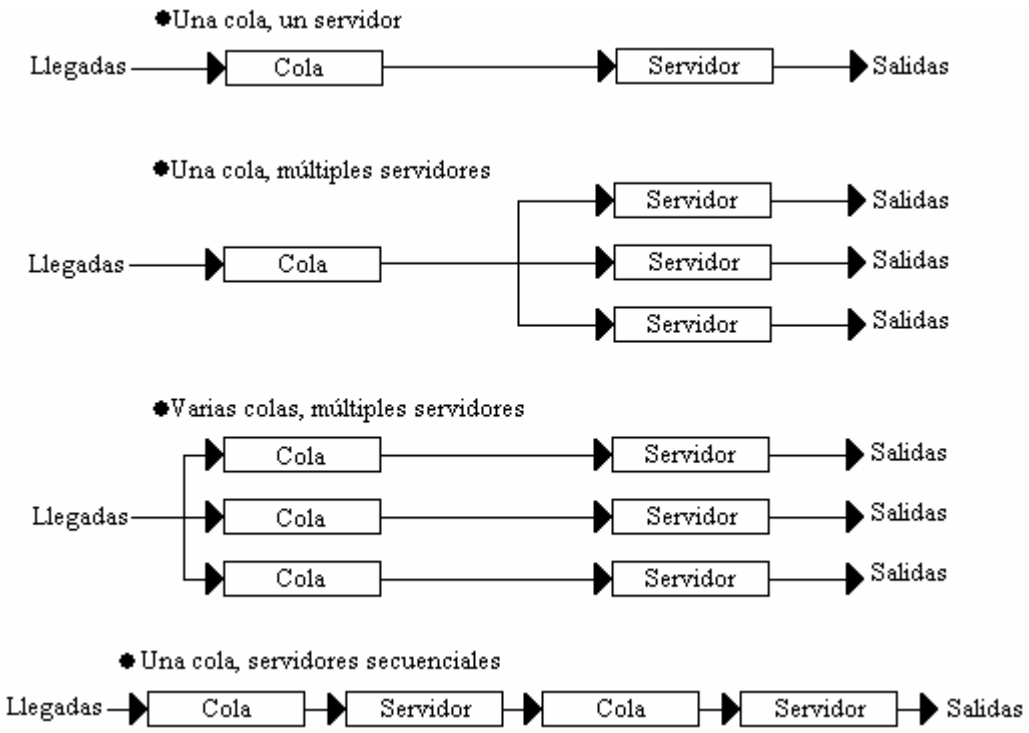

Figura 2-5 Tipos de sistemas [12].

Es importante señalar en este punto, que para nuestro caso utilizaremos la siguiente analogía en la teoría de colas: los paquetes serán los clientes, el servidor podrá ser un “router” o conmutador y la cola será el “buffer” de los servidores.

2.3.1 Sistemas M/M/1

Con respecto a la notación de Kendall, para este sistema se tienen las siguientes características:

A) Se tiene un sistema de llegadas que se producen según un proceso de Poisson de razón λ, donde los tiempos entre llegadas estarán distribuidos exponencialmente $Exp(\lambda)$

- Donde λ es el número medio de llegadas por unidad de tiempo

B) Los tiempos entre servicios son distribuidos de manera exponencial, $Exp(\mu)$

- Donde μ es el número medio de paquetes que el servidor es capaz de atender por unidad de tiempo

X) Se posee un único servidor en el sistema

Y) La capacidad del sistema es infinita, la cual se puede omitir

Z) La disciplina del sistema será FIFO, la cual se puede omitir

V) Se tiene un estado de servicio igual a uno, es decir una sola cola, el cual se puede omitir también

Es decir, el sistema es el siguiente: M/M/1/∞/FIFO/1, pero se abrevia como M/M/1. A continuación se irá analizando el sistema exclusivamente en su condición de no saturación, es decir como un estado estable, ya que si el sistema llega a saturarse el número de paquetes en la cola crecerá indefinidamente, esto quiere decir que el sistema tendrá una tasa mayor de la que el servidor puede manejar.

Para este tipo de sistema, se define la intensidad de tráfico, también conocida como factor de utilización, como:

$$\rho = \frac{\lambda}{\mu} \tag{2.1}$$

donde:

ρ = Intensidad de tráfico en el sistema,

λ = Número medio de llegadas por unidad de tiempo, y

μ = Número medio de paquetes que el servidor es capaz de atender por unidad de tiempo.

Por lo tanto, para que el sistema sea estable, se tiene la siguiente condición de no saturación:

$$\rho < 1 \tag{2.2}$$

donde el parámetro ρ se le domina también como carga o flujo. Este parámetro mide la relación entre la media de los paquetes por unidad de tiempo y la capacidad de atenderlos por unidad de tiempo. Si se cumple la condición de no saturación, las probabilidades del estado estable existen y están dadas por:

$$\rho_n = \rho^n (1 - \rho) \tag{2.3}$$

donde:

ρ_n = Probabilidad de que haya n paquetes en el sistema.

Esta fórmula indica la probabilidad de que haya n paquetes en el sistema, dependiendo del tipo de red que se tenga, ya que el tamaño del paquete es diferente en cada red. En los sistemas de líneas de espera, las medidas de rendimiento, también conocidas como medidas de efectividad, son las medidas de mayor interés. En ellas el tiempo medio en que un paquete permanece en el sistema se le denomina como W. Si hacemos un supuesto de que llega un paquete a una cola y hay j paquetes antes que éste, se tendrá la siguiente fórmula:

$$W = \sum_{j=0}^{\infty} (j+1)\frac{1}{\mu} P_j = \sum_{j=0}^{\infty} j\frac{1}{\mu} P_j + \sum_{j=0}^{\infty} \frac{1}{\mu} p_j \quad (2.4)$$

donde:

W = Tiempo medio que un paquete permanece en el sistema,

j = Número de paquetes que se encuentran en la cola antes del paquete actual, y

P_j = Probabilidad de que haya j paquetes en el sistema.

Al evaluarse y simplificar la ecuación resultante, se obtendrá la siguiente fórmula de W:

$$W = \frac{L}{\mu} + \frac{1}{\mu} = \frac{1}{\mu - \lambda} \quad (2.5)$$

donde:

L =Número medio de paquetes en el sistema.

Una vez que se obtuvo W, es posible obtener el tiempo medio de espera en la cola, denominado como W_q, el cual se calcula en base a W, al tiempo medio que un paquete permanece en la cola se le resta el tiempo medio de servicio, siendo este $\frac{1}{\mu}$.

Por lo tanto:

$$W_q = W - \frac{1}{\mu} \quad (2.6)$$

donde:

W_q = Tiempo medio de espera en la cola.

Una consideración especial de la fórmula para este tipo de cola es la siguiente:

$$W_q = \frac{\rho}{\mu - \lambda} \quad (2.7)$$

Una de las últimas medidas de rendimiento que son importantes es el número medio de trabajos en la cola, conocido como L_q, el cual es calculado al restarle al número medio de paquetes en el sistema, la carga de tráfico que existe en el sistema, así como se muestra a continuación:

$$L_q = L - (1 - \rho_0) = L - \rho = \frac{\rho}{1-\rho} - \rho \tag{2.8}$$

donde:

L_q = Número medio de paquetes en la cola, y

ρ_0 = Probabilidad de que no existan paquetes en el sistema.

La ecuación (2.8) puede simplificarse de la siguiente manera:

$$L_q = \frac{\rho^2}{1-\rho} \tag{2.9}$$

2.3.2 Sistemas M/M/1/K

Con respecto a la notación de Kendall, para este sistema se tienen las siguientes características:

A) Se tiene un sistema de llegadas que se producen según un proceso de Poisson de razón λ, donde los tiempos entre llegadas estarán distribuidos exponencialmente $Exp(\lambda)$

- Donde λ es el número medio de llegadas por unidad de tiempo

B) Los tiempos entre servicios son distribuidos de manera exponencial, $Exp(\mu)$

- Donde μ es el número medio de paquetes que el servidor es capaz de atender por unidad de tiempo

X) Se posee un único servidor en el sistema

Y) La capacidad del sistema es finita, ésta se expresa por la constante K

Z) La disciplina del sistema será FIFO, la cual se puede omitir

V) Se tiene un estado de servicio igual a uno, es decir una sola cola, el cual se puede omitir también

En este sistema debe de considerarse que se está limitando el número de paquetes que van a poder entrar a la cola, es decir si la cola estuviera llena los paquetes que llegaran después serían rechazados. La ventaja que tiene este tipo de sistemas es que no se necesita utilizar una condición de no saturación debido a que la capacidad es limitada y por ello se encuentra siempre en un estado estable, sin importar cual sea el valor de ρ, siendo ρ igual a la ecuación (2.1). Las probabilidades en este sistema están dados por:

$$Pn = \left\{ \begin{array}{l} \dfrac{\rho^n (1-\rho)}{1-\rho^{K+1}} \rightarrow cuando(\rho \neq 1) \\ \dfrac{1}{K+1} \rightarrow cuando(\rho = 1) \end{array} \right\} \tag{2.10}$$

donde:

P_n = Probabilidad de que haya n paquetes en el sistema,

ρ = Intensidad de tráfico en el sistema, y

K = Número de paquetes que caben en el sistema.

En este caso, la ρ determina cómo varían las probabilidades, ya que si $\rho < 1$ los estados más probables son aquellos donde la oferta de servicio supera a la demanda, en cambio cuando se tiene $\rho > 1$ la oferta de servicio no es suficiente para el servicio que se está solicitando, por último se tiene el caso equilibrado donde $\rho = 1$.

En este sistema, como en el de M/M/1, se tienen medidas de rendimiento que son de gran interés, una de ellas es el número medio de paquetes en el sistema, L, para el cual las condiciones de fórmula son las siguientes:

$$L = \begin{cases} \dfrac{\rho}{1-\rho} - \dfrac{(K+1)\rho^{K+1}}{1-\rho^{K+1}} \rightarrow cuando(\rho \neq 1) \\ \dfrac{K}{2} \rightarrow cuando(\rho = 1) \end{cases} \qquad (2.11)$$

donde:

L = Número medio de paquetes en el sistema.

Otra medida de rendimiento de gran utilidad es la tasa efectiva de llegadas, representada como λ_{ef}. Este parámetro es el número medio de clientes que son admitidos al sistema. Es por ello que la tasa efectiva de llegadas será siempre menor al número medio de llegadas, ambas por unidad de tiempo. Se tendrá la siguiente fórmula:

$$\lambda_{ef} = \lambda(1 - P_K) \qquad (2.12)$$

donde:

λ_{ef} = Tasa efectiva de llegadas,

λ = Número medio de llegadas por unidad de tiempo, y

P_K = Probabilidad de que haya K paquetes en el sistema.

Finalmente, considerando la tasa efectiva de llegadas, se obtendrán las siguientes fórmulas de W, L y L_q. Resolviéndose W_q con la ecuación (2.7)

$$W = W_q + \frac{1}{\mu} \qquad (2.13)$$

$$L = \lambda_{ef} W \qquad (2.14)$$

$$L_q = \lambda_{ef} W_q \qquad (2.15)$$

donde:

W = Tiempo medio que un paquete permanece en el sistema,

W_q = Tiempo medio de espera en la cola,

μ = Número medio de paquetes que el servidor es capaz de atender por unidad de tiempo, y

L_q = Número medio de paquetes en la cola.

2.3.3 Sistemas M/M/c

Con respecto a la notación de Kendall, para este sistema se tienen las siguientes características:

A) Se tiene un sistema de llegadas que se producen según un proceso de Poisson de razón λ, donde los tiempos entre llegadas estarán distribuidos exponencialmente $Exp(\lambda)$

- Donde λ es el número medio de llegadas por unidad de tiempo

B) Los tiempos entre servicios son distribuidos de manera exponencial, $Exp(\mu)$

- Donde μ es el número medio de paquetes que el servidor es capaz de atender por unidad de tiempo

X) El número de servidores en el sistema de denotará con la constante c

Y) La capacidad del sistema es infinita, la cual se puede omitir

Z) La disciplina del sistema será FIFO, la cual se puede omitir

V) Se tiene un estado de servicio igual a uno, es decir una sola cola, el cual se puede omitir también

Este sistema al igual que el sistema M/M/1 presenta una capacidad del sistema infinita por lo cual se establece una condición de no saturación para alcanzar el estado estable, ya que de esta manera se cuida que el número de paquetes no crezca indefinidamente. Para este software sólo se ocuparán colas que no se saturan, por lo que la condición será la siguiente:

$$\rho < 1 \tag{2.16}$$

donde se tiene que ρ se calcula así:

$$\rho = \frac{\lambda}{c\mu} \tag{2.17}$$

donde:

ρ = Intensidad de tráfico en el sistema,

λ = Número medio de llegadas por unidad de tiempo,

c = Número de servidores en el sistema, y

μ = Número medio de paquetes que el servidor es capaz de atender por unidad de tiempo.

Para un estado estable, es decir no saturado, se tienen las siguientes probabilidades:

$$\rho_0 = \left(\frac{c^c \rho^c}{c!(1-\rho)} + \sum_{n=0}^{c=1} \frac{(cp)^n}{n!} \right)^{-1} \tag{2.18}$$

$$P_n = \begin{Bmatrix} \frac{(c\rho)^n}{n!}\rho_0 \rightarrow n = 0,1,...,c \\ \frac{c^c \rho^n}{c!}\rho_0 \rightarrow otro\_caso \end{Bmatrix} \tag{2.19}$$

donde:

ρ_0 = Probabilidad de que no existan paquetes en el sistema,

n = Paquetes en el sistema, y

P_n = Probabilidad de que haya n paquetes en el sistema.

En cuanto a las medidas de rendimiento para este sistema, se tiene que el número medio de clientes en la cola está dado por:

$$L_q = \frac{c^c \rho^{c+1} \rho_0}{c!(1-\rho)^2} \tag{2.20}$$

donde:

L_q = Número medio de paquetes en la cola.

Otras medidas como lo es W puede obtenerse mediante la ecuación (2.5), mientras que W_q puede calcularse con la ecuación (2.6). Otros razonamientos como L y L_q podrán obtenerse con las siguientes fórmulas:

$$L = \lambda W \tag{2.21}$$

$$L_q = \lambda W_q \tag{2.22}$$

donde:

L = Número medio de paquetes en el sistema,

W = Tiempo medio que un paquete permanece en el sistema, y

W_q = Tiempo medio de espera en la cola.

Una última medida de rendimiento lo es el número medio de servidores ocupados, caracterizada por la constante S, se calculará mediante la siguiente fórmula:

$$S\mu = \lambda \Rightarrow S = \frac{\lambda}{\mu} = c\rho \tag{2.23}$$

donde:

S = Número medio de servidores ocupados.