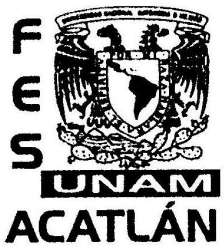



# MODELOS EXPONENCIALES DE COLAS APLICADOS A LA EFICIENCIA DE UN CENTRO DE ATENCIÓN TELEFÓNICA: UN CASO PRÁCTICO

**T E S I N A**

PARA OBTENER EL TÍTULO DE:

**LICENCIADO EN ACTUARÍA**

P R E S E N T A :

**JOSE ALEJANDRO LÓPEZ GONZÁLEZ**

ASESOR ACTUARIO. MAHIL HERRERA MALDONADO

MÉXICO, D.F. JULIO DE 2009

A mis padres, por el apoyo y cariño incondicional que siempre me han brindado.

A mi hermano, por estar siempre presente.

A México, por poner el conocimiento al alcance de todos sus ciudadanos.

A la Universidad Nacional Autónoma de México, por haberme formado desde la infancia.

# Agradecimientos

Agradezco al Actuario Mahil Herrera la experiencia, disposición y paciencia que mostró en el asesoramiento de este proyecto. Asimismo, expreso mi gratitud a todas las personas que de diferentes maneras contribuyeron para que el trabajo llegara a su término.

La única característica que de verdad nos distingue
del resto de los animales es la capacidad de prever.

BERTRAND RUSELL.

No puede existir un lenguaje más universal y simple,
más carente de errores y oscuridades, y por lo tanto
más apto para expresar las relaciones invariables de
las cosas naturales [...] [*Las matemáticas*]
parecen constituir una facultad de la mente humana destinada
a compensar la brevedad de la vida y la imperfección de los sentidos.

JOSEPH FOURIER,
*Théorie analytique de la chaleur.*
Discurso preliminar (1822)

# Índice

# Objetivo

Mostrar que los modelos exponenciales de colas son una buena herramienta para lograr los niveles de eficiencia establecidos por un centro de atención telefónica.

# Introducción

Las colas, o líneas de espera, son fenómenos que es posible observar con gran frecuencia en distintos ámbitos de la realidad. Hay muchos fenómenos que pueden ser descritos como un conjunto de clientes esperando para recibir algún servicio. Por ejemplo el número de llamadas que llegan a una central telefónica, en un determinado tiempo, el número de personas que pueden ser atendidas en una caja de banco o en una bomba de gasolina por unidad de tiempo, así como el número de personas que llegan a cualquier tipo de servicio y tienen que esperar para ser atendidos. En todos estos casos hay tres cosas que nos interesan: Cuántas personas llegan por unidad de tiempo, es decir, el flujo de llegada, cuánto tiempo tienen que esperar para ser atendidos, esto es el tiempo de espera en cola y cuánto tiempo tardarán siendo atendidos, es decir, el tiempo del servicio. Estos tres parámetros serán fundamentales para poder entender cómo se comportan las líneas de espera y poder modelar este comportamiento para tratar de predecir el tiempo de espera dependiendo de la cantidad de servidores en paralelo ó poder predecir el tiempo de espera en función de la cantidad de servidores.

La teoría de colas es la modelación, mediante herramientas matemáticas, de líneas de espera o colas. Desarrolladas inicialmente por el matemático danés A.K.(Agner Krarup) Erlang, quien en 1909 publicó su documento fundamental en la congestión del tráfico telefónico. Más tarde otros matemáticos como David G. Kendall, quien en 1953 introdujo la notación A/B/C, que es hoy en día universalmente utilizada para describir el tipo de colas que se pretende modelar, han seguido con el desarrollo de esta teoría.

En el presente trabajo haremos uso de este conocimiento, principalmente mediante una modelación como procesos estocásticos exponenciales de las llegadas, tiempos de espera, y tiempos de servicio, para tratar de optimizar los recursos y predecir los tiempos en servicio, de espera en cola y de estancia en el sistema de cada cliente.

El objetivo del desarrollo de estos modelos es lograr un uso eficiente del recurso principal en cualquier centro de atención telefónica, que es el personal, aunque esto implicará también hacer uso eficiente de otros recursos como infraestructura y los enlaces telefónicos, ofreciendo un nivel de servicio óptimo. Es decir, lograr un balance entre el uso eficiente de los recursos con que se cuenta sin descuidar brindar una atención suficientemente buena a los clientes.

En el presente trabajo realizamos el desarrollo de modelos de colas exponenciales que con mayor frecuencia se presentan en la realidad para después hacer una aplicación práctica de estos modelos en un caso real de flujo de llamadas en un centro telefónico.

En el capítulo 1 se tratan de la definición de los objetivos dentro de un centro telefónico, los

cuales son relativos al costo y nivel de servicio, intentando lograr el mejor nivel de servicio al menor costo posible.

En el capítulo 2 nos ocuparemos de los conceptos generales que son las definiciones de las partes de un sistema de colas así como el proceso de encolamiento y la notación que será utilizada para describir los sistemas de colas que pretendemos modelar. Se deducen algunas propiedades generales de las colas y se establecen algunas propiedades importantes de la distribución *Poisson.*

En el capítulo 3, partiendo del supuesto de estabilidad en el sistema, lo que nos permite utilizar las ecuaciones de balance, desarrollaremos los distintos modelos de colas, así como todas las fórmulas que posteriormente utilizaremos en la aplicación.

En el capítulo 4, hacemos una aplicación práctica de algunos de los resultados que hemos desarrollado en los capítulos 2 y 3 procurando lograr los objetivos definidos en el capítulo 2. Veremos qué tan bien se ajusta la realidad a las predicciones hechas con nuestros modelos. Además haremos un contraste con simulaciones, para ver qué tanto nuestros modelos se ajustan con la realidad y con la simulación de colas.

Por último llegamos a las conclusiones, en donde establecemos los beneficios que acarrea la utilización de estos modelos. Así mismo, se señalan algunas consideraciones al usarlos.

# 1. Objetivos de la Administración de Centros Telefónicos

## 1.1. Costo y Nivel de Servicio

El objetivo principal de un centro telefónico es brindar el servicio que esperan los clientes que se comunican telefónicamente para recibir algún tipo de servicio. El grado de satisfacción de los clientes depende de distintos aspectos relacionados como la calidad de la respuesta, el tiempo de espera del cliente, amabilidad del agente, etc. Algunos de estos aspectos son incuantificables.

El administrador de un centro telefónico tiene que procurar el mejor nivel de servicio posible mediante el uso inteligente de los recursos con que cuenta. Recursos como presupuesto, número de posiciones, infraestructura para recibir las llamadas telefónicas y la fuerza laboral. Es claro que mientras mayor sea la disponibilidad de recursos, mejor será el nivel de servicio que es posible ofrecer a los clientes. Sin embargo, tomando en cuenta que el principal recurso de un centro de atención telefónica es el agente o representante, la infraestructura y los procesos deben establecerse de manera tal, que sea posible maximizar el efectivo y eficiente empleo de la fuerza de trabajo.

El balance costo-servicio se centra, principalmente, en el manejo cuantitativo del Centro Telefónico. En general, cuando se incrementa el costo, entonces el nivel de servicio (NS) se incrementa también.

## 1.2. Costo y Productividad

El costo más importante en un centro telefónico son los salarios de los agentes, es por eso que los agentes deben trabajar de la manera más eficiente y efectiva posible. El indicador más importante de efectividad de desempeño es la Solución en el Primer Intento (FTR First-Time-Resolution o también llamada First-Call-Resolution). Otro importante indicador es el Tiempo Promedio de Espera (AHT Average Holding Time), que es el tiempo que, en promedio, tiene que esperar un cliente antes de ser atendido. Así, si el SPI(FTR) es alto, no habrá necesidad de que el cliente tenga que hacer otra llamada para ser atendido. Es así que mantener el SPI (FTR) lo más alto posible así como el TPE (AHT) lo más bajo posible son los objetivos principales en cualquier centro telefónico.

El principal indicador de eficiencia es la productividad en algún periodo de tiempo determinado. Usualmente la productividad está dada como el *porcentaje del tiempo que el agente está trabajando, del total de su tiempo de trabajo*:

$$\text{Productividad} = \frac{\text{Tiempo total de Trabajo}}{\text{Tiempo Total Disponible}}$$

El tiempo total de trabajo se define como la suma del tiempo en llamada y el tiempo en espera de llamada (wrap-up). El tiempo total disponible es el resto del tiempo que el operador tiene que estar en el centro telefónico. Aunque en ocasiones para el tiempo total se establecen distintos criterios por ejemplo para incluir, o no, en el tiempo total los tiempos de descanso o capacitación.

Por ejemplo. Suponiendo que un operador telefónico tiene que trabajar 36 horas cada semana, de las cuales 6 horas constituyen la suma de sus periodos habituales de descanso y ha estado recibiendo o esperando llamadas durante 1632 minutos. Entonces, si no contamos los descansos como tiempo sin trabajar, su productividad ha sido de $1632/2160 = 75{,}56\,\%$. Ahora, si consideramos los descansos como tiempo sin trabajar tendríamos: $(1632 + 360)/2160 = 92\,\%$.

## 1.3. Medidas de Nivel de Servicio

Aunque existen diversos factores que pueden provocar la ocurrencia de abandonos de llamadas como, por ejemplo, un trato poco amable por parte de algún agente o la falta de una señal clara en la llamada telefónica, el análisis cuantitativo se enfoca principalmente en los tiempos de espera como causa de los posibles abandonos. Una manera común de definir el Nivel de Servicio (SL Service Level) es considerando el total de llamadas que son contestadas antes de un tiempo fijo de espera considerado como *aceptable*. A este tiempo *aceptable* de espera se le denomina *Tiempo Aceptable de Espera (AWT Acceptable Answering Time)*. Así el número de llamadas contestadas en un tiempo menor o igual al TAE (AWT) es denominado *Factor de Servicio Telefónico (TSF Telephone Service Factor)*

La interpretación del FST (TSF) es la siguiente: Supongamos que un centro telefónico ha encontrado que 10 segundos de espera es un tiempo razonable para evitar la pérdida de clientes. Así que define su TAE(AWT) = 10 segundos y considera que un FST(TSF) de 80/20 (80 de cada 100 llamadas por debajo del Tiempo Aceptable De Espera AWT) es un Nivel de Servicio (SL) aceptable. Así que 1 de cada 5 clientes que llamen recibirá un mal servicio. Si vuelve a llamar, tiene nuevamente una probabilidad de 20 % de obtener un mal servicio y así sucesivamente. Con un TSF definido de esta manera, la probabilidad de que un cliente tenga 3 llamadas consecutivas consideradas como mal servicio es menor a 1 % lo cual puede ser considerado razonable o no, dependiendo de la oferta del mismo servicio por otros centros telefónicos.

El objetivo principal del centro telefónico es evitar, en la medida de lo posible, los abandonos después del Tiempo de Espera establecido (AWT), es así que es posible definir el nivel servicio como:

$$\text{SL} = \frac{\text{Número de llamadas contestadas antes de TAE(AWT)}}{\text{Llamadas Contestadas Totales}}$$

Otra posibilidad es contarlas como la proporción de llamadas que alcanzaron en nivel de servicio (SL) propuesto.

Así que el nivel de servicio puede medirse en una proporción entre 0 y 1.

Es así que es necesario utilizar un modelo que nos permita optimizar tanto el costo como el nivel de servicio.

# 2. Composición y conceptos generales de los sistemas de colas

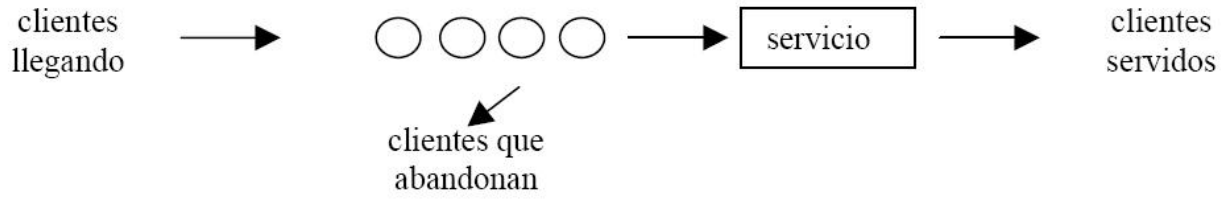



Figura 1: Sistema de Colas típico

## 2.1. Introducción

La teoría de colas data de 1909 cuando el matemático Agner Krarup Erlang, (1878-1929) publicó su documento fundamental en la congestión de tráfico telefónico. Años después Kendall (1951,1953) fue el pionero en el desarrollo de la teoría de colas desde la perspectiva de los procesos estocásticos. La teoría de colas es el estudio matemático de las *líneas de espera* o *colas*. Una cola se forma siempre que la demanda de algún servicio excede la capacidad de proveer el servicio en ese punto del tiempo, es decir, cuando no es posible proveer al cliente con el servicio de forma inmediata. Un sistema de colas se conforma por clientes o unidades que necesitan algún tipo de servicio que llegan a una instancia donde el servicio es provisto, y se integran a una cola si el servicio no está disponible inmediatamente y eventualmente se retiran después de haber recibido el servicio; también hay casos en los que los clientes dejan el sistema sin haber recibido el servicio o después de haber esperado cierto tiempo.

Los términos clientes y servidor son genéricos. Los clientes son aquellos que necesitan algún tipo de servicio, y llegan a la instancia en donde dicho servicio es brindado. La entidad que realiza el servicio a los clientes se llama servidor o canal, por ejemplo clientes de un banco o llamadas llegando a un conmutador telefónico. Si al momento de llegar, el cliente encuentra al servidor ocupado, se forma o se incorpora a una cola.

## 2.2. Elementos del Sistema de Colas

- El patrón de llegada de clientes.
- El patrón de servicio.
- El número de servidores o canales de servicio.

- La capacidad del sistema; y
- La disciplina de la cola

A continuación describimos cada uno de estos elementos.

### 2.2.1. El patrón de Llegada de Clientes

El patrón de llegada es la forma en la que la llegada de clientes ocurre. Está especificado por el tiempo entre llegadas consecutivas. Para esta medida, usualmente se considera el tiempo promedio entre llegadas o su recíproco, el número promedio de llegadas por unidad de tiempo. El tiempo entre llegadas puede ser determinístico, o estocástico. Si es determinístico, entonces el número de llegadas será el mismo en todos los intervalos de tiempo, pero si es estocástico su distribución de probabilidad debe ser especificada. La hipótesis más comúnmente utilizada es que las llegadas ocurren de acuerdo con una distribución Poisson. El patrón de llegada también especifica si las llegadas ocurren individualmente o en grupos, si ocurren en grupos, la forma en que cada uno de estos grupos son constituidos es especificada también por el patrón de entrada. En ocasiones las llegadas pueden no incorporarse a la cola debido a su longitud de la misma o debido a que son excluidos de esa opción debido al espacio de espera que no puede tener más clientes esperando de lo que su limitada capacidad permite. Por último, el patrón de entrada toma en cuenta si las llegadas provienen de una fuente infinita o en ocasiones de una fuente finita.

### 2.2.2. El Patrón de Servicio

El patrón de servicio es la forma en que el servicio es brindado. Es especificado por el tiempo que toma completar un servicio. El tiempo de servicio puede ser constante (determinístico) o estocástico. Si es estocástico, la especificación del patrón considera la distribución del tiempo de servicio de una unidad o cliente. La medida típicamente considerada es el tiempo promedio requerido para servir a una unidad o el número de unidades servidas por unidad de tiempo. En ocasiones el servicio puede brindarse en bloques o grupos, como en el caso de un elevador, en vez del servicio a una sola unidad a la vez. En este caso la forma en la que forman los grupos para el servicio debe ser especificada también.

### 2.2.3. Número de Servidores

Un sistema puede tener un solo servidor o un cierto número de servidores en paralelo. Una unidad que encuentra a todos los servidores ocupados tiene que incorporarse a la cola, de lo contrario tendrá que acudir con alguno de ellos para que le brinde el servicio. A menos que se

especifique lo contrario, vamos a suponer al referirnos a sistemas con servidores múltiples que el primer cliente que llega es el primero que recibe el servicio del servidor que quedó libre primero.

### 2.2.4. La Capacidad Del Sistema

Un sistema puede tener una capacidad infinita, es decir, la cola frente a los servidores puede crecer ilimitadamente; así que puede haber limitación de espacio, por lo que cuando el espacio se llena a su capacidad máxima, una llegada no podrá incorporarse al sistema y se perderá del sistema. El sistema se llama *sistema de retraso (delay sistem)* o *sistema de pérdida (loss sistem)*, según la capacidad sea infinita o finita. Si es finito, debe especificarse el número de lugares disponibles para la cola además de los clientes en servicio.

### 2.2.5. Disciplina de la Cola

La disciplina de la cola se refiere a la forma en que los clientes son elegidos para brindarles el servicio. La disciplina de la cola más usual es aquella en la que el primer cliente en llegar, es el primero en ser atendido, FCFS o FIFO (First Come First Served First In First Out), aunque pueden definirse otros tipos de disciplinas como el último en llegar, primero en ser atendido, servicio en orden aleatorio, o de acuerdo a algún tipo de prioridad establecida.

Si las llegadas ocurren en grupos y el servicio les es ofrecido individualmente, entonces la forma en la que los clientes llegan al grupo y son ordenados para el servicio, debe ser también especificada.

Tres importantes medidas son la tasa promedio de llegadas (denotada por $\lambda$), la tasa promedio de servicio (denotado por $\mu$), y el número ($c \geq 1$) de servidores en paralelo con una única cola. La cantidad

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

en el caso del sistema de un sólo servidor y $\rho = \lambda/c\mu$ en el caso del sistema de $c$ servidores se llama *intensidad de tráfico* o *carga* del sistema. El resultado de estas ecuaciones nos da unidades de medida llamadas Erlangs, en memoria a E.K. Erlang.

## 2.3. El Proceso de Encolamiento

El análisis de sistemas de colas con tiempos entre llegadas y de servicio determinísticos no presenta mucha dificultad, además considerando la aplicación práctica para la que es necesario el presente desarrollo teórico, este trabajo está enfocado en los modelos o sistemas en los que ambos, tiempo entre llegadas y de servicio, son estocásticos. Este análisis tiene que involucrar

la descripción estocástica del sistema y las medidas de comportamiento relacionadas que a continuación mencionaremos.

1. $N_t$ Distribución del número $N_t$ de unidades en el sistema al tiempo $t$ (número en la cola más el número que están siendo servidos, si es que los hay).
2. $W_n$ Distribución del tiempo de espera en la cola (en el sistema) para la enésima llegada. El tiempo que una llegada tiene que esperar en la cola (manteniéndose en el sistema).
3. $W_t$ Distribución del tiempo virtual de espera $W_t$ el tiempo que una llegada tiene que esperar habiendo llegado al tiempo $t$.
4. Distribución del periodo de ocupación que es la cantidad de tiempo en la cual el servidor se mantiene ocupado. El periodo de ocupación es el intervalo desde el momento de la llegada de una unidad a un sistema vacío hasta el momento en que el canal queda libre por primera vez.

## 2.4. Notación A/B/X/Y/Z

La notación generalmente utilizada para describir los modelos de colas es la introducida por Kendall(1951). Consiste en la especificación de las tres características básicas: la entrada, el tiempo de servicio, y el número de servidores en paralelo. Los símbolos utilizados para denotar estas características son los siguientes.

- A: Distribución de tiempo entre llegadas consecutivas.
- B: Distribución de tiempo de servicio.
- X: Número de canales de servicio.
- Y: Capacidad del sistema.
- Z: Disciplina de la cola.

- $M$: Para tiempo entre llegadas exponencial (Entrada Poisson) o tiempo de servicio exponencial.
- $E_k$: Para la distribución Erlang-k
- $H$ Para la distribución Hiperexponencial.
- $D$ Para tiempo entre llegadas o de servicio determinístico (constante).

- $G$ Para una distribución General (arbitraria).

Así, por ejemplo, la notación $M/G/1$ denota una cola o modelo con entrada Poisson, distribución de tiempo de servicio general y un solo servidor. Dos descriptores más pueden agregarse, cuando es necesario; el cuarto denota la capacidad del sistema y el quinto denota el tamaño (finito) de la fuente de la cual provienen las llegadas. Así el modelo $M/G/1$ significa lo mismo que el modelo $M/G/1/\infty$. El modelo $G/G/c/K/N$ se refiere a un modelo con $c$ servidores con distribución de tiempo entre llegadas y servicio arbitraria o general, espacio antes de ser atendido limitado a K (incluyendo aquellos siendo servidos si es que los hay) y $N$ siendo el tamaño de la fuente de la que provienen las llegadas.

Así podemos definir distintos tipos de modelos de colas según el tipo de sistema con el cual estemos trabajando.

## 2.5. Estado Estable y Estado Transitorio

Denotaremos $N_t$ al número de clientes en sistema (clientes en la cola más clientes siendo servidos) al tiempo $t$ medidos a partir de un instante fijo ($t = 0$) y su distribución de probabilidad como

$$p_n(t) = Pr\{N_t = n\}, \qquad n = 0, 1, 2, \ldots$$

Entonces

$$p_i(0) = 1, (p_j(0) = 0, j \neq i)$$

implica que el número de clientes en el momento inicial era $i$ (donde $i$ puede ser $0, 1, 2, \ldots$). Para una descripción completa del comportamiento estocástico de procesos de medida de colas $\{N_t, t \geq 0\}$, necesitamos encontrar una solución dependiente del tiempo $p_n(t), n \geq 0$. Es difícil obtener estas soluciones. Y de hallarlas, es muy complicado manejarlas. De cualquier modo, para muchas situaciones prácticas, es necesario tener un comportamiento de equilibrio, es decir, un comportamiento en el que el sistema alcance un estado de equilibrio después de estar operando por tiempo suficiente. En otras palabras, nos interesa el comportamiento de $p_n(t)$ cuando $t \to \infty$. Denotemos

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

siempre que el límite exista. Así, $p_n$ es la probabilidad límite de que haya $n$ clientes en el sistema. Siempre que este límite exista, se dice que el sistema alcanza un estado estable o de equilibrio y $p_n$ es independiente del tiempo. Usualmente resulta que $p_n$ es igual a la proporción de tiempo que el sistema contiene exactamente $n$ clientes. En particular, $p_0$ denota la proporción de tiempo que el sistema está vacío. Se sigue que

$$\sum_{n=0}^{\infty} p_n = 1;$$

esta se llama condición de normalización. Consideremos otras *probabilidades límites* $\{a_n, n \geq 0\}$ y $\{d_n, n \geq 0\}$ definidas como sigue

$a_n$=probabilidad de que las llegadas (clientes que llegan) encuentren $n$ en el sistema cuando llegan.
$d_n$=probabilidad de que las salidas (clientes que salen) encuentren $n$ en el sistema cuando salen.

Resulta que $a_n$ es la proporción a largo plazo de clientes que, al llegar, encuentran $n$ en el sistema, y $d_n$ es la proporción a largo plazo de clientes que, al partir, encuentran $n$ en el sistema. Las tres cantidades $p_n, a_n$, y $d_n$ no siempre son iguales.

**Teorema 2.1.** *En cualquier sistema en el que las llegadas ocurren uno a uno y ha alcanzado el estado de equilibrio,*

$$a_n = d_n \qquad \textit{para todo } n \geq 0$$

*Demostración.* Consideremos una llegada que encontrará $n$ en el sistema; entonces el número en el sistema se incrementará en 1 e irá de $n$ a $n+1$. Nuevamente, una salida dejará $n$ en el sistema, implicando que el número en el sistema decrece en 1, de $n+1$ a $n$. En cualquier intervalo de tiempo $T$, el número de transiciones $A$ de $n$ a $n+1$ y el número de transiciones $B$ de $n+1$ a $n$ diferirá cuando más en 1; en otras palabras, ya sea que $A = B$ o $A \approx B = 1$. Entonces para $T$ grande, las tasas de transición $A/T$ y $B/T$ serán iguales. Así, en promedio, las llegadas y salidas siempre encuentran el mismo número de clientes, lo que significa que $a_n = d_n$ siempre para toda $n \geq 0$. □

## 2.6. Algunas Relaciones Generales de Teoría de Colas (Fórmula de Little)

Existen algunos resultados y relaciones en teoría de colas que se cumplen bajo condiciones bastante generales. Aunque las pruebas matemáticas de tales relaciones son algo complicadas. A continuación mencionamos algunas de estas relaciones que aplican para sistemas en *estado estable*.La más importante es:

$$L = \lambda W \tag{2.1}$$

donde $\lambda$ es la tasa media de llegada, $L$ es el número promedio de unidades en el sistema y $W$ es el tiempo esperado, o promedio, en el sistema en estado estable. Denotemos el número esperado en la cola y el tiempo esperado en la cola en estado estable por $L_Q$ y $W_Q$, respectivamente, Estos están relacionados por una ecuación similar.

$$L_Q = \lambda W_q \tag{2.2}$$

Puede encontrarse una prueba rigurosa de esta relación [6], razón por la cual esta relación es conocida como Fórmula de Little.

Este resultado, de gran generalidad, aplica independientemente del tipo de distribuciones del tiempo entre llegadas y de servicio. Aplica bajo condiciones muy generales para cualquier sistema siempre y cuando el sistema esté en estado estable.

Así que la ecuación $L = \lambda W$ es válida en general y relaciona el tiempo promedio de espera de los clientes $W$ con el número promedio de clientes en cola $L$, dada una tasa de llegada $\lambda$.

## 2.7. El papel de la distribución Exponencial

Procesos de llegadas en diversas situaciones pueden ser modelados como procesos Poisson. Así que la suposición de que el tiempo entre llegadas T, tiene una distribución exponencial, tiene algunas implicaciones importantes para los modelos de teoría de colas, razón por la cual, resulta útil revisar algunas propiedades de la distribución exponencial que mencionaremos a continuación.

### 2.7.1. Propiedad 1

$f_T(t)$ es una función estrictamente decreciente de $t$ $(t \geq 0)$.

Una consecuencia de la propiedad 1 es que

$$P\{0 \leq T \leq h\} > P\{t \leq T \leq t + h\}$$

para cualesquiera valores estrictamente positivos de $h$ y $t$.

### 2.7.2. Propiedad 2

Carencia de memoria.

$$P\{T \geq t + h \mid T \geq h\} = P\{T \geq t\} \qquad h, t > 0$$

Esto significa que la distribución del tiempo restante hasta que ocurre el incidente (llegada o compleción del servicio) es siempre la misma, sin importar cuánto tiempo $h$ ya ha transcurrido.

Es decir, el proceso carece de memoria. Este interesante resultado ocurre con la distribución exponencial debido a que:

$$\begin{aligned} P\{T > t+h \mid T > h\} &= \frac{P\{T > h, T > t+h\}}{p\{T > h\}} \\ &= \frac{P\{T > t+h\}}{p\{T > h\}} \\ &= \frac{exp^{-\lambda\{t+h\}}}{exp^{-\lambda h}} \\ &= exp^{-\lambda t} \\ &= P\{T \geq t\} \end{aligned}$$

Para los tiempos entre llegadas, este resultado describe la situación en que el tiempo hasta la siguiente llegada no es influido en forma alguna por el hecho de cuándo ocurrió el evento anterior. Para los tiempos de servicio describe el hecho de que el tiempo restante hasta la siguiente compleción de servicio no tiene que ver con hace cuánto se completó el último servicio.

### 2.7.3. Propiedad 3

El mínimo de un conjunto de variables aleatorias exponenciales tiene una distribución exponencial.

Sean $T_1, T_2, \ldots, T_n$ variables aleatorias exponenciales con parámetros $\lambda_1, \lambda_2, \ldots, \lambda_n$, respectivamente. Además, sea $U$ la variable aleatoria que toma el valor igual al mínimo de los valores que en realidad son tomados por $T_1, T_2, \ldots, T_n$. Es decir

$$U = min\{T_1, T_2, \ldots, T_n\}.$$

Por lo tanto, si $T_j$ representa el tiempo hasta que ocurre una clase particular de incidente, entonces $U$ es el tiempo hasta que ocurre el primero de los $n$ incidentes diferentes. Ahora, para cualquier $t \geq 0$

$$\begin{aligned} P\{U > t\} &= P\{T_1 > t, T_2 > t, \ldots, T_n > t\} \\ &= P\{T_1 > t\}\{T_2 > t\} \ldots \{T_3 > t\} \\ &= exp^{\lambda_1 t} exp^{\lambda_2 t} \ldots exp^{\lambda_n t} \\ &= exp\{-\textstyle\sum_{i=1}^{n} \lambda_i t\} \end{aligned}$$

de modo que, en efecto, $U$ tiene una distribución exponencial con parámetro

$$\lambda = \sum_{i=1}^{n} \lambda_i$$

Como consecuencia de esta propiedad, supongamos que se tienen $n$ tipos diferentes de clientes. Los tiempos entre llegadas para cada tipo $i$ tienen una distribución exponencial con parámetro $\lambda_i$ $(i = 1, 2, \ldots, n)$. Por la propiedad 2, el tiempo restante desde cualquier instante especificado hasta la siguiente llegada de un cliente del tipo $i$ tendría esta misma distribución. Por lo tanto, sea $T_i$ este tiempo restante, medido desde el instante en que llega un cliente de cualquier tipo. Entonces, de acuerdo con la propiedad 3, los tiempos entre llegadas para el sistema de colas como un todo, tiene una distribución exponencial con parámetro $\lambda$ definido por la ecuación anterior. Así que es posible ignorar la distinción entre los clientes y todavía tener tiempos exponenciales para el modelo de colas.

Otra implicación, para los tiempos de servicio, es la siguiente. Consideremos la situación en que todos los servidores tienen la misma distribución exponencial de tiempo de servicio, con parámetro $\mu$. Sea $n$ el número de servidores que están proporcionando servicio actualmente y sea $T_i$ el tiempo de servicio restante para el servidor $i$, el cual también tiene una distribución exponencial con parámetro $\lambda_i = \mu$. Entonces U, el tiempo hasta la compleción del siguiente servicio por parte de cualquiera de estos servidores, tiene una distribución exponencial con parámetro $n\mu$. Así que el sistema de colas estará desempeñándose como un sistema de un solo servidor en el que los tiempos de servicio tienen una distribución exponencial con parámetro $n\mu$

### 2.7.4. Propiedad 4

Relación con la distribución de Poisson.
Sea $N_t$ el número de ocurrencias consecutivas (por ejemplo llegadas o compleciones de servicio) hasta el instante $t$, en donde el instante 0 es aquel en el que se inicia el conteo, entonces

$$P\{N_t = n\} = \frac{(\lambda t)^n e^{-\lambda t}}{n!}$$

es decir, $N_t$ tiene una distribución de Poisson con parámetro $\lambda t$. Así que la media de $N_t$ es

$$E\{N_t\} = e^{\lambda t}$$

de modo que el número esperado de incidentes por unidad de tiempo es $\lambda$. Es así que $\lambda$ es la *tasa media* a la cual ocurren los incidentes. Al contar los incidentes de manera continua, tenemos un *proceso de conteo* $\{X_t \mid t \in \mathbb{R}^+\}$, que es un *proceso Poisson* con parámetro $\lambda$.

Esta propiedad es útil tanto para modelar el número de compleciones de servicio en un periodo de tiempo [0,t) así como para modelar el comportamiento de las llegadas en el mismo tipo de periodo.

### 2.7.5. Propiedad 5

Para t $\geq 0$ $P\{T \leq t+h \mid T \geq t\} \approx \alpha h$, para $h$ pequeño. Así que para una ind exponencial con parámetro $\alpha$, la propiedad 2 implica que

$$\begin{aligned} P\{T < t+h \mid T > t\} &= P\{T \leq h\} \\ &= exp^{-\alpha h} \end{aligned}$$

para cualesquiera cantidades positivas $t$ y $h$. Por lo tanto como la expresión, como serie infinita, de $e^x$, para cualquier exponente $x$

$$e^x = 1 + x + \sum_{n=2}^{\infty} \frac{x^n}{n!}$$

concluimos que

$$P\{T \leq t+h \mid T > t\} = 1 - 1 + \alpha h - \sum_{n=2}^{\infty} \frac{(-\alpha h)^n}{n!} \approx \alpha h$$

para $h$ pequeño.

Debido a que los términos en la suma se vuelven relativamente despreciables para valores $\alpha h$ suficientemente pequeños. Debemos notar, además, que el valor de $t$ no afecta esta probabilidad en absoluto.

Como hemos mencionado, $T$ podría representar tiempos entre llegadas así como tiempos de servicio en los modelos de colas. Por lo tanto esta propiedad nos permite conocer la probabilidad de que la llegada o la compleción de servicio ocurra en el siguiente intervalo de tiempo pequeño $\alpha t$. ( También es posible hacer exacto el análisis basado en la aproximación tomando los límites apropiados conforme $h \to 0$,). La propiedad también indica que esta probabilidad es esencialmente proporcional a $h$, para valores pequeños de $h$.

## 2.8. Una característica del proceso de llegada Poisson: PASTA

Los procesos Poisson tienen una propiedad única. Para llegadas Poisson

$$a_n = p_n \qquad \text{para toda } n \geq 0 \tag{2.3}$$

Así que si $N_t(t \geq 0)$ denota el estado del sistema (número de ocurrencias o llegadas), si $t_0$ es un instante arbitrario, y las llegadas ocurren de acuerdo con un proceso Poisson, entonces la distribución de la variable aleatoria $N_{t0}$ es independiente de la ocurrencia, o no, de una llegada en $t_0$.

Esta propiedad se llama *PASTA* (Poisson Arrivals See Time Averages). En una cola con llegadas Poisson, la proporción límite de llegadas que encuentran el sistema en un estado $n$ es igual a la proporción límite de tiempo que el sistema está en estado $n$. Así que el término PASTA se refiere a la igualdad entre estas dos proporciones.

# 3. Desarrollo de los modelos de colas mediante sistemas de nacimiento y muerte.

## 3.1. Introducción

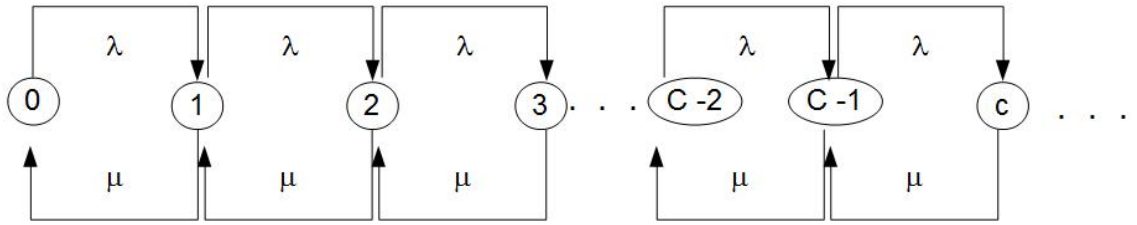



Figura 2: Diagrama Sistema Nacimiento y Muerte

Muchos modelos de colas simples, pero interesantes pueden ser modelados a través de procesos de nacimiento y muerte . En este tipo de procesos las transiciones ocurren desde cualquier estado a alguno de sus estados inmediato anterior o posterior. Esto es, con una llegada, hay una transición del estado $i(\geq 0)$ al estado $i+1$ y al completar algún servicio habrá una transición de $j$ a $(j-1)(j \geq 0)$. A continuación desarrollaremos los resultados para el modelo que nos será de utilidad en el caso real que analizaremos más adelante.

## 3.2. El proceso de Nacimiento y Muerte.

La mayor parte de los modelos de colas utilizan la hipótesis de que las entradas (clientes que llegan) y las salidas (clientes que salen) ocurren de acuerdo con un *proceso de nacimiento y muerte*. Este importante proceso estocástico tiene aplicación en diversas áreas. Sin embargo, en el ámbito de la teoría de colas, un *nacimiento* se refiere a la llegada de un nuevo cliente al sistema de colas , y la *muerte* se refiere a la salida de un cliente servido. El estado del sistema, en el instante $t(t \geq 0)$ está dado por $N_t$. En consecuencia, el proceso de nacimiento y muerte describe cómo cambia $N_t$ a medida que se incrementa $t$. De manera más precisa, las hipótesis de nacimiento y muerte son las siguientes:

**Hipótesis 1** Dado $N_t = n$, la distribución actual de probabilidad del tiempo restante hasta el siguiente nacimiento (llegada) es exponencial con parámetro $\lambda_n(n = 0, 1, 2, \ldots)$.

**Hipótesis 2** Dado $N_t = n$, la distribución actual de probabilidad del tiempo restante hasta la siguiente muerte o compleción del servicio es exponencial con parámetro $\mu_n(n = 1, 2, \ldots)$.

**Hipótesis 3** Sólo puede ocurrir un nacimiento o una muerte en un instante.

El que los nacimientos y muertes se distribuyan exponencialmente implica que $\lambda_n$ y $\mu_n$ son tasas medias. Estas hipótesis son ilustradas en la figura 2 Las flechas en este diagrama muestran las únicas transiciones posibles en cualquier estado del sistema (como lo especifica la hipótessis 3) y la entrada para cada flecha de la tasa media para esa transición (como lo especifican las hipótesis 1 y 2), cuando se encuentra en el estado de la base de la flecha.

El análisis de proceso de nacimiento y muerte es muy difícil cuando el sistema se encuentra en una condición *transitoria*, además es relativamente sencillo modelar el proceso de nacimiento y muerte considerando que el proceso es estacionario para conocer la distribución de $N_t$, razón por la cual suele establecerse esta hipótesis para desarrollar los modelos de colas. Lo anterior puede llevarse a cabo directamente a partir del diagrama de tasas, como describiremos a continuación.

Considérese cualquier estado particular del sistema $n(n = 0, 1, 2, \ldots)$. Supóngase que se estuviera empezando a contar el número de veces en que el proceso entra a este estado y el número de veces en que sale de él. Como los dos tipos de incidentes (entrada y salida) deben alterarse, estos dos números siempre deben ser iguales o diferir tan sólo en 1. Esta diferencia posible de 1 al final causaría únicamente una diferencia despreciable en las tasas promedio (número total de ocurrencias por unidad de tiempo) a las que han ocurrido estos dos tipos de incidentes ( es decir, $1/t \to 0$, a medida que $t \to \infty$). Por lo tanto, a largo plazo, estas dos tasas deben ser iguales. Esto conduce al siguiente principio clave:

**Definición 3.1.** *Ecuación de Balance*

*Principio de la Tasa de Entrada = Tasa de Salida.*

*Para cualquier estado del sistema, $n(n = 0, 1, 2, \ldots)$, la tasa media o número esperado de ocurrencias por unidad de tiempo a la que los incidentes de entrada ocurren deben ser igual a la tasa media a la cual ocurren los incidentes de salida.*

Después de construir las ecuaciones de balance para todos los estados, en términos de las probabilidades $P_n$ desconocidas, entonces puede resolverse este sistema de ecuaciones para hallar estas probabilidades.

Consideremos el estado 0. El proceso entra en este estado únicamente desde el estado 1. Por lo tanto, la probabilidad de estado estacionario de encontrarse en el estado 1 ($P_1$) representa la proporción de veces que le sería posible al proceso entrar al estado 0. Dado que el proceso se encuentra en el estado 1, la tasa media de entrada al estado 0 es $\mu_1$. Es decir, por cada unidad acumulada de tiempo que el proceso pase en el estado 1, el número esperado de veces en que saldría del estado 1 para entrar al estado 0 es $\mu_1$.Desde cualquier otro estado, la tasa media de entrada al estado 0 es 0. Por lo tanto, la probabilidad de que el proceso salga de su estado actual para entrar al estado 0 es

$$\mu_1 P_1 + 0(1 - P) = \mu_1 P_1$$

Utilizando el mismo razonamiento, la tasa media de ocurrencia de los incidentes de salida debe ser $\lambda_0 P_0$, de modo que la ecuación de balance para el estado 0 es

$$\mu_1 P_1 = \lambda_0 P_0$$

Para todos los demás estados existen dos transiciones posibles, tanto hacia adentro como hacia afuera del estado. Por lo tanto cada miembro de las ecuaciones de balance para estos estados representa la suma de las tasas medias para las dos transiciones que intervienen. En caso contrario, el razonamiento es precisamente el mismo que para el estado 0. Estas ecuaciones de balance se resumen en la tabla a continuación.

$$\begin{array}{ccccc}
0 & \mu_1 P_1 & = & \lambda_0 P_0 \\
1 & \lambda_0 P_0 + \mu_2 P_2 & = & (\lambda_1 + \mu_1) P_1 \\
2 & \lambda_1 P_1 + \mu_3 P_3 & = & (\lambda_2 + \mu_2) P_2 \\
\vdots & \vdots & & \\
n-1 & \lambda_{n-2} P_{n-1} + \mu_n P_n & = & (\lambda_{n-1} + \mu_{n-1}) P_{n-1} \\
n & \lambda_{n-1} P_{n-1} + \mu_{n+1} P_{n+1} & = & (\lambda_n + \mu_n) P_n \\
\vdots & \vdots & &
\end{array}$$

Notemos que la primera ecuación de balance contiene dos variables para las cuáles debe ser resuelta($P_0$ y $P_1$); las dos primeras ecuaciones contienen tres variables ($P_0, P_1, P_2$), y así sucesivamente, de modo que siempre existe una variable extra. Por lo tanto el proceso de solución de estas ecuaciones es resolver en términos de una de las variables, siendo $P_0$ la más conveniente. De donde, se usa la primera ecuación con el fin de resolver para $P_1$ en términos de $P_0$, entonces se usa este resultado en la segunda ecuación con el fin de resolver para $P_2$ en términos de $P_0$, etc. Al final, puede utilizarse la propiedad de que la suma de todas las probabilidades es igual a 1 para evaluar $P_0$.

Aplicando este procedimiento llegamos a los resultados siguientes

$$
\begin{array}{cccccccc}
0 & P_1 & = & \dfrac{\lambda_0}{\mu_1}P_0 & & & & \\
1 & P_2 & = & \dfrac{\lambda_1}{\mu_2}P_1 + \dfrac{1}{\mu_2}(\mu_1 P_1 - \lambda_0 P_0) & = & \dfrac{\lambda_1}{\mu_2}P_1 & = & \dfrac{\lambda_1\lambda_0}{\mu_2\mu_1}P_0 \\
2 & P_3 & = & \dfrac{\lambda_2}{\mu_3}P_2 + \dfrac{1}{\mu_3}(\mu_2 P_2 - \lambda_1 P_1) & = & \dfrac{\lambda_2}{\mu_3}P_2 & = & \dfrac{\lambda_2\lambda_1\lambda_0}{\mu_3\mu_2\mu_1}P_0 \\
\vdots & & \vdots & & & & & \\
n-1 & P_n & = & \dfrac{\lambda_{n-1}}{\mu_n}P_{n-1} + \dfrac{1}{\mu_n}(\mu_{n-1}P_{n-1} - \lambda_{n-2}P_{n-2}) & = & \dfrac{\lambda_{n-1}}{\mu_n}P_{n-1} & = & \dfrac{\lambda_{n-1}\lambda_{n-2}\cdots\lambda_0}{\mu_n\mu_{n-1}\cdots\mu_1}P_0 \\
n & P_{n+1} & = & \dfrac{\lambda_n}{\mu_{n+1}}P_n + \dfrac{1}{\mu_{n+1}}(\mu_n P_n - \lambda_{n-1}P_{n-1}) & = & \dfrac{\lambda_n}{\mu_{n+1}}P_n & = & \dfrac{\lambda_n\lambda_{n-1}\cdots\lambda_0}{\mu_{n+1}\mu_n\cdots\mu_1}P_0 \\
\vdots & & \vdots & & & & &
\end{array}
$$

Sea

$$S_n = \frac{\lambda_{n-1}\lambda_{n-2}\cdots\lambda_0}{\mu_n\mu_{n-1}\cdots\mu_1}. \quad \text{para } n = 1, 2, \ldots$$

Así, las probabilidades de estado estacionario son

$$P_n = S_n P_0 \qquad \text{para } n = 1, 2, \ldots \tag{3.1}$$

Como

$$\sum_{n=0}^{\infty} P_n = 1$$

Entonces

$$\left[1 + \sum_{n=1}^{\infty} S_n\right] P_0 = 1,$$

$$P_0 = \frac{1}{1 + \sum_{n=1}^{\infty} S_n} \tag{3.2}$$

Dada esta información

$$L = \sum_{n=0}^{\infty} nP_n$$

También, puesto que el número de servidores $s$ representa el número de clientes que pueden ser servidos, y por lo tanto eliminados de la cola , simultáneamente,

$$L_q = \sum_{n=s}^{\infty} (n-s)P_n$$

Además las relaciones 2.1 y 2.2 conducen a

$$W = \frac{L}{\bar{\lambda}}, W_q = \frac{L_q}{\bar{\lambda}}$$

en donde $\lambda$ es la tasa promedio de llegadas a largo plazo. Puesto que $\lambda_n$ es la tasa media de llegadas mientras el sistema se encuentra en el estado $n(n = 0, 1, 2, \ldots)$ y $P_n$ es la proporción del tiempo que el sistema se halla en este estado,

$$\bar{\lambda} = \sum_{n=0}^{\infty} \lambda_n P_n$$

Los resultados anteriores se han obtenido bajo la suposición de que los parámetros $\lambda_n$ y $\mu_n$ tienen valores tales que, en realidad, el proceso puede alcanzar una condición de estado estacionario. Siempre se cumple esta hipótesis si $\lambda_n = 0$ a partir algún valor de $n$, de modo que solo son posibles un númerofinito de estados (aquellos menores que esta n). También siempre se cumple cuando se definen $\lambda$ y $\mu$, con $\rho = \lambda/s\mu < 1$. No se cumple si $\sum_{n=1}^{\infty} S_n = \infty$.

## 3.3. El modelo M/M/c

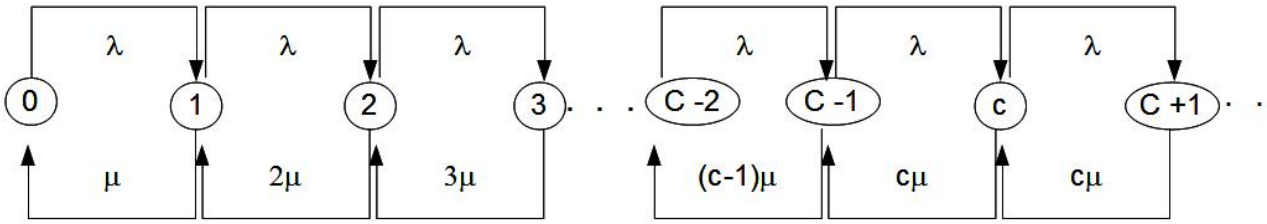



Figura 3: Diagrama Sistema M/M/c

### 3.3.1. Distribución de estado estable

. Consideremos una cola con entrada Poisson, con parámetro $\lambda$ y $c(1 \leq c \leq \infty)$ canales de servicio en paralelo cuya distribución de tiempo servicio es I.I.D. exponencial con parámetro $\mu$. Si hay $n$ clientes en el sistema, y $n$ es menor que $c$, entonces, $n$ canales están ocupados y el intervalo entre dos compleciones consecutivas de servicio es también exponencial con parámetro $c\mu$ (Propiedad 3 de la sección 2.7). Si hay $n \geq c$ clientes en el sistema, entonces todos los $c$ canales están ocupados y el intervalo entre dos servicios completados consecutivamente es exponencial con parámetro $c\mu$. Así tenemos un modelo de nacimiento y muerte con parámetro de entrada (nacimiento) $\lambda$ y parámetros de salida (muerte):

$$\mu_n = \begin{cases} n\mu & n = 0, 1, 2, \ldots, c \\ c\mu & n = c+1, c+2, \ldots \end{cases}$$

Sea

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

Asumimos que el proceso es estable y que el sistema ha alcanzado la estabilidad. Remplazando los valores de $\lambda_n$ y $\mu_n$ en las ecuaciones 3.2 y 3.1 , tenemos, para $n = 1, 2, \ldots, c$,

$$\begin{aligned} p_n &= \frac{\lambda\lambda\lambda}{(\mu)(2\mu)\cdots(n\mu)} p_0 = \frac{(\lambda/\mu)^n}{n!} p_0 \\ &= \frac{\lambda}{n\mu} p_{n-1} \end{aligned} \tag{3.3}$$

y para $n = c, c+1, c+2, \ldots$

$$\begin{aligned} p_n &= \frac{\lambda\lambda\lambda\cdots\lambda}{[(\mu)(2\mu)\cdots(c\mu)][(c\mu)(c\mu)\cdots(c\mu)]}p_0 \\ &= \frac{\lambda^n}{c!\mu^c c^{n-c}\mu^{n-c}}p_0 \\ &= \frac{(\lambda/\mu)^n}{c!c^{n-c}}p_0 \\ &= \frac{\lambda}{c\mu}p_{n-1} = \rho^{n-c}p_c. \end{aligned} \tag{3.4}$$

La condición $\sum_{n=0}^{\infty} p_n = 1$ implica

$$p_0^{-1} = 1 + \sum_{n=1}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \sum_{n=c}^{\infty}\frac{\left(\frac{\lambda}{\mu}\right)^n}{c!c^{n-c}} = \sum_{n=0}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \frac{1}{c!c^{-c}}\sum_{n=c}^{\infty}\left(\frac{\lambda}{c\mu}\right)^n.$$

De donde, para que exista una solución de estado estable, la serie

$\sum_{n=c}^{\infty}\left(\frac{\lambda}{c\mu}\right)^n$ debe ser convergente, lo cual ocurrirá si, y sólo si, $\rho < 1$. Así que, para $\rho < 1$

$$p_0 = \left[\sum_{n=0}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!(1-\lambda/c\mu)}\right]^{-1} \tag{3.5}$$

y la distribución de estado estable está dada por las ecuaciones 3.3, 3.4 , con $p_0$ dado por 3.5. Las probabilidades de estado estable $p_n$ satisfacen las relaciones de recurrencia:

$$p_n = \begin{cases} \frac{\lambda}{n\mu}p_{n-1} = \frac{c}{n}\rho p_{n-1} & n = 1, 2, \ldots, c-1 \\ \frac{\lambda}{c\mu}p_{n-1} = \rho p_{n-1} & n = c, c+1, \ldots \end{cases}$$

Así, para $n < c$, $p_n/p_{n-1} > 1$ si $(n/c) < \rho < 1$; en este caso $p_n$ es monótona creciente en $n$ hasta que $n$ excede $c\rho$ y entonces es monótona decreciente hasta $n = c$ Para $n > c, p_n$ es monótona decreciente en $n$.

Además, para $n$ finita, $\{p_n\}$ se distribuye como una Poisson para $n \leq c$ y como una distribución geométrica para $n > c$ [3].

### 3.3.2. Probabilidad de Espera en Cola Fórmula C de Erlang

La probabilidad de que una llegada tenga que esperar, es decir, que encuentre a todos los servidores ocupados, puede ser evaluada utilizando la siguiente ecuación llamada fórmula C de Erlang (o segunda fórmula de Erlang).

$$C = C\left(c, \frac{\lambda}{\mu}\right) = Pr\{N \geq c\} = \sum_{n=c}^{\infty} p_n$$

$$= \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!(1-\rho)} p_0 = \frac{p_c}{1-\rho} \tag{3.6}$$

Es conocida como *Fórmula C de Erlang.*

## 3.4. Número de servidores en espera u ocupados

El número promedio de servidores ocupados $E(B)$ está dado por:

$$\begin{aligned}
E(B) &= \sum_{n=0}^{c-1} np_n + \sum_{n=c}^{\infty} cp_n \\
&= \left[ \sum_{n=0}^{c-1} \frac{n\left(\frac{\lambda}{\mu}\right)^n}{n!} + \frac{c\left(\frac{\lambda}{\mu}\right)^c}{c!(1-\rho)} \right] p_0 \\
&= \frac{\lambda}{\mu} \left[ \sum_{n=1}^{c-1} \frac{\left(\frac{\lambda}{\mu}\right)^{n-1}}{(n-1)!} + \frac{\left(\frac{\lambda}{\mu}\right)^{c-1}}{(c-1)!(1-\rho)} \right] p_0 \\
&= \frac{\lambda}{\mu} \left[ \sum_{m=0}^{c-2} \frac{\left(\frac{\lambda}{\mu}\right)^{m}}{m!} + \frac{\{(1-\rho)+\rho\}\left(\frac{\lambda}{\mu}\right)^{c-1}}{(c-1)!(1-\rho)} \right] p_0 \\
&= \frac{\lambda}{\mu} \left[ \sum_{m=0}^{c-2} \frac{\left(\frac{\lambda}{\mu}\right)^{m}}{m!} + \frac{\left(\frac{\lambda}{\mu}\right)^{c-1}}{(c-1)!} + \frac{\rho\left(\frac{\lambda}{\mu}\right)^{c-1}}{(c-1)!(1-\rho)} \right] p_0 \\
&= \frac{\lambda}{\mu} \left[ \sum_{m=0}^{c-1} \frac{\left(\frac{\lambda}{\mu}\right)^{m}}{m!} + \frac{\left(\frac{\lambda}{\mu}\right)^{c}}{c!(1-\rho)} \right] p_0 \\
&= \frac{\lambda}{\mu} p_0^{-1} p_0 = \frac{\lambda}{\mu} = c\rho.
\end{aligned}$$

$$= \frac{\lambda}{\mu} p_0^{-1} p_0 = \frac{\lambda}{\mu} = c\rho. \tag{3.7}$$

Así que el número esperado de servidores en espera $E(I)$ está dado por

$$\begin{aligned} E(I) &= E(c-B) = E(c) - E(B) \\ &= c - c\rho = c(1-\rho) \end{aligned} \tag{3.8}$$

### 3.4.1. Número Esperado de Clientes en el Sistema

$$E(N) = E(B) + E(Q)$$

donde $E(Q)$ es el número esperado de clientes en cola. Tenemos

$$\begin{aligned} E(Q) &= \sum_{n=c}^{\infty}(n-c)p_n \\ &= \sum_{n=c}^{\infty}(n-c)\frac{\left(\frac{\lambda}{\mu}\right)^n}{c!c^{n-c}}p_0 \\ &= \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}\sum_{n=c}^{\infty}(n-c)\left(\frac{\lambda}{c\mu}\right)^{n-c}p_0 \\ &= \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}\frac{\lambda}{c\mu}\sum_{m=0}^{\infty}m\left(\frac{\lambda}{c\mu}\right)^{m-1}p_0 \\ &= \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}\frac{\lambda}{c\mu}p_0\frac{1}{\left(1-\frac{\lambda}{c\mu}\right)^2} \\ &= \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}p_0\frac{\rho}{(1-\rho)^2} \\ &= \frac{\rho p_c}{(1-\rho)^2} = \frac{\rho}{(1-\rho)}Pr\{N \geq c\} \end{aligned} \tag{3.9}$$

Así,

$$\begin{aligned} E(N) &= E(B) + E(Q) \\ &= c\rho + \rho\frac{p_c}{(1-\rho)^2} \\ &= c\rho + \frac{\rho C}{1-\rho} \end{aligned} \tag{3.10}$$

donde $C = Pr\{N \geq c\}$.

Utilizando la fórmula de Little (Sección 2.6), es posible encontrar $E(W_Q)$, tiempo esperado en la cola, y $E(W)$, tiempo esperado en el sistema o tiempo de respuesta. Tenemos

$$E(W_Q) = \frac{E(Q)}{\lambda} = \frac{p_c}{c\mu(1-\rho)^2} = \frac{1}{c\mu(1-\rho)}Pr\{N \geq c\} \tag{3.11}$$

$$E(W) = \frac{E(N)}{\lambda} = \frac{1}{\mu} + \frac{p_c}{c\mu(1-\rho)^2} \tag{3.12}$$

Usando la fórmula de Little, podemos ver además que los resultados 3.7 y 3.8 son válidos para cualquier sistema de colas general $G/G/c$ con $\rho = \dfrac{\lambda}{c\mu} < 1$. Esto lo observamos obteniendo el resultado como sigue.

$$\begin{aligned} L &= \lambda W \\ L_Q &= \lambda W_Q \end{aligned}$$

Restando, tenemos

$$L - L_Q = \lambda(W - W_Q).$$

En donde el lado izquierdo de la igualdad es el número promedio de canales de servicio o el número promedio de canales ocupados $E(B)$, y $(W - W_Q)$ es el tiempo promedio en servicio que equivale a $1/\mu$. Así

$$E(B) = \frac{\lambda}{\mu} = c\rho.$$

que es el mismo resultado de la ecuación 3.7 al que habíamos llegado anteriormente.

## 3.5. Distribuciones de Tiempo de Espera

Debemos considerar dos tipos de tiempo de espera:(1) Tiempo de espera $W_q$ en cola o *tiempo de espera* y (2) tiempo de espera $W$ en el sistema, el cual incluye tiempo de espera, en caso de haberlo, en cola más tiempo del servicio, que es el tiempo total utilizado en el sistema por alguna unidad de prueba (también llamado tiempo de respuesta). Para hallar esta distribución, debe tomarse en cuenta la disciplina de la cola. Asumiremos que es FCFS, primero en llegar, primero en ser atendido.

Sean $w_q(x)$ y $w(x)$ las FDP (Función de Densidad de Probabilidad) de los tiempos de espera $W_q$ y $W_s$ en la cola y en el sistema, respectivamente. de la unidad de prueba, y sean $w_q^*(s)$ y $w^*(s)$ sus Tiempos Restantes (TR). Luego, sean $w_q^*(s \mid n)$ y $w^*(s \mid n)$ los Tiempos Restantes de las PDF de las distribuciones condicionales de los tiempos de espera respectivos dado que la unidad de prueba encuentra, a su llegada, $n$ en el sistema. Obtenemos $w^*(s)$ basados en el número de unidades que la unidad de prueba encuentra a su llegada. Si la unidad de prueba encuentra, al llegar, $n < c$ unidades, no tiene que esperar para ser atendida y su tiempo de espera en el sistema es igual al tiempo de servicio, esto es,

$$w^*(s \mid n) = \frac{\mu}{s+\mu}, \qquad \text{para } n < c. \tag{3.13}$$

Si encuentra $n \geq c$ unidades en el sistema, tendrá que esperar en la cola hasta que se complete el servicio de $(n-c+1)$ unidades, todos los $c$ canales de servicio ocupados tienen entonces tasa de servicio $c\mu$. Tomando en consideración su propio tiempo de servicio, tendrá que esperar en el sistema hasta que sean completados $(n-c+1)$ servicios a una tasa $c\mu$ y su propio servicio a tasa $\mu$, esto es

$$w^*(s \mid n) = \left(\frac{c\mu}{s+c\mu}\right)^{n-c+1}\left(\frac{\mu}{s+\mu}\right), n \geq c \tag{3.14}$$

Como para este proceso Poisson $a_n = p_n$ (propiedad 2.3 PASTA), tenemos

$$w^*(s) = \sum_{n=0}^{c-1} w^*(s \mid n)p_n + \sum_{n=c}^{\infty} w^*(s \mid n)p_n$$

$$= \frac{\mu}{s+\mu}\left[\sum_{n=0}^{c-1} p_n + \sum_{n=c}^{\infty} p_n \left(\frac{c\mu}{s+c\mu}\right)^{n-c+1}\right] \tag{3.15}$$

Remplazando los valores de $p_n$, tenemos

$$w^*(s) = \frac{\mu}{s+\mu}\left[\sum_{n=0}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \sum_{n=c}^{\infty}\frac{\left(\frac{\lambda}{\mu}\right)^n}{c!c^{n-c}}\left(\frac{c\mu}{s+c\mu}\right)^{n-c+1}\right]p_0$$

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

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

$$= \frac{\mu}{s+\mu}\left[\sum_{n=0}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}\frac{c\mu}{s+c\mu-\lambda}\right]p_0 \tag{3.16}$$

Obtenemos $w(x)$ invirtiendo $w^*(s)$. De la ecuación anterior, tenemos

$$w^*(s) = \sum_{n=0}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}p_0\frac{\mu}{s+\mu} + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}p_0\left(\frac{\mu}{s+\mu}\right)\frac{c\mu}{s+c\mu-\lambda}$$

$$= \sum_{n=0}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}p_0\frac{\mu}{s+\mu} + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}p_0\frac{c\mu^2}{(c-1)\mu-\lambda}\left[\frac{1}{s+\mu} - \frac{1}{s+c\mu-\lambda}\right]$$

Revirtiendo la transformación, tenemos

$$w(x) = \sum_{n=0}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}p_0\mu e^{-\mu x} + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}p_0\frac{c\mu^2}{(c-1)\mu-\lambda}[e^{-\mu x} - e^{-(1-\rho)c\mu x}]$$

Para $c = 1$,

$$\begin{aligned} w(x) &= \mu p_0 e^{-\mu x} + \frac{\left(\frac{\lambda}{\mu}\right) p_0 \mu^2}{-\lambda} [e^{-\mu x} - e^{-(1-\rho)\mu x}] \\ &= \mu(1-\rho) e^{-(1-\rho)\mu x} \end{aligned}$$

Como $w^*(s) = w_q^*(s)[\mu/(s+\mu)]$, tenemos de 3.15

$$\begin{aligned} w_q^*(s) &= \sum_{n=0}^{c-1} p_n + \sum_{n=c}^{\infty} p_n \left(\frac{c\mu}{s+c\mu}\right)^{n-c+1} \\ &= \left[\sum_{n=0}^{c-1} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!} \frac{c\mu}{s+c\mu-\lambda}\right] p_0 \end{aligned} \tag{3.17}$$

Invirtiendo, la ecuación anterior obtenemos la FDP

$$w_q(x) = \left(\sum_{n=0}^{c-1} p_n\right) \delta(x) + \sum_{n=c}^{\infty} p_n \frac{c\mu(c\mu x)^{n-c} e^{-c\mu x}}{(n-c)!}, \qquad x \geq 0 \tag{3.18}$$

donde $\delta$ es la función delta Dirac (o unidad de impulso). Poniendo en las expresiones para $p_n$, y simplificando, tenemos

$$w_q(x) = \left(1 - \frac{p_c}{1-\rho}\right) \delta(x) + c\mu p_c e^{-c\mu(1-\rho)x} \tag{3.19}$$

### 3.5.1. Función Complementaria de Distribución

$$\begin{aligned} Pr\{W_q > t\} &= \int_t^\infty w_q(x)dx \\ &= \frac{\left(\frac{\lambda}{\mu}\right)^c}{(c-1)!}\mu\frac{e^{-c(\mu-\lambda)t}}{c(\mu-\lambda)}p_0 \\ &= C\left(c, \frac{\lambda}{\mu}\right)e^{-(1-\rho)c\mu t} \end{aligned} \tag{3.20}$$

Donde $C\left(c, \frac{\lambda}{\mu}\right) = Pr\{W_q \geq 0\}$ (pérdida de Erlang) es la probabilidad de encontrar todas las posiciones ocupadas dada por 3.6.

Usando el resultado anterior, podemos ver que

$$\begin{aligned} Pr\{W_q > t \mid W_q > 0\} &= \frac{Pr\{W_q > t\}}{Pr\{W_q > 0\}} \\ &= e^{-(1-\rho)c\mu t} \end{aligned} \tag{3.21}$$

La distribución del tiempo condicional de espera en cola, dado que la unidad de prueba tenga que esperar, es exponencial con media:

$$E\{W_q \mid W_q > 0\} = \frac{1}{(1-\rho)c\mu} \tag{3.22}$$

### 3.5.2. Tiempo Esperado en el Sistema

Tenemos

$$E(W_s) = -\frac{d}{ds}w^*(s) \Big|_{s=0}$$

$$= \frac{1}{\mu}\left[\sum_{n=0}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}p_0 + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}\frac{c\mu}{c\mu-\lambda}p_0\right] + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}\frac{c\mu}{(c\mu-\lambda)^2}p_0$$

$$= \frac{1}{\mu}\left[\sum_{n=0}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}\frac{1}{1-\rho}\right]p_0 + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!(c\mu)}\frac{1}{(1-\rho)^2}p_0$$

Así

$$E(W_s) = \frac{1}{\mu} + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!c\mu}\frac{1}{(1-\rho)^2}p_0 = \frac{1}{\mu} + \frac{p_c}{c\mu(1-\rho)^2}. \tag{3.23}$$

$$= \frac{1}{\mu} + \frac{C}{c\mu(1-\rho)}. \tag{3.24}$$

Se sigue que

$$E(W_q) = \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!c\mu}\frac{1}{(1-\rho)^2}p_0 = \frac{p_c}{c\mu}\frac{1}{(1-\rho)^2}. \tag{3.25}$$

Poniendo $c = 1$, podemos encontrar el resultado que corresponde a un sistema M/M/1, en donde los clientes son atendidos por un servidor único.

### 3.5.3. El Proceso de Salida

Ahora consideraremos el proceso de salida de un sistema de colas M/M/c.

**Teorema 3.1.** *En un sistema de colas M/M/c en estado estable, con tasas de llegada y servicio $\lambda$ y $\mu$, respectivamente, los tiempos entre salidas son independientes e identicamente distribuidos con distribución exponencial con media $1/\lambda$, es decir, el proceso de salida es Poisson con parámetro $\lambda$.*

## 3.6. El Modelo de pérdida $M/M/c/c$

Consideremos un modelo con $c$ servidores con entrada Poisson y tiempo de servicio exponencialmente distribuidos, tales que cuando todos los $c$ canales están ocupados la siguiente llegada abandona el sistema sin esperar por el servicio. Este sistema se llama *Sistema de Pérdida de c canales*.

Este es un modelo de colas de nacimiento y muerte con

$$\begin{aligned} &\lambda_n = \lambda, \quad \mu_n = n\mu, \quad n = 0, 1, 2, \ldots, c-1 \\ &\lambda_n = 0, \quad \mu_n = c\mu, \quad n \geq c \end{aligned} \tag{3.26}$$

Usando los resultados 3.1 y 3.2 obtenemos

$$\begin{aligned} p_n &= \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} p_0, \quad n = 1, \ldots, c \\ &= 0, \qquad\qquad n \geq c \end{aligned}$$

y

$$p_0 = \left[ \sum_{k=0}^{c} \frac{\left(\frac{\lambda}{\mu}\right)^k}{k!} \right]^{-1}$$

Así

$$p_n = \frac{\left(\frac{\lambda}{\mu}\right)^n / n!}{\sum_{k=0}^{c} \left(\frac{\lambda}{\mu}\right)^k / k!} \qquad n = 0, 1, 2, \ldots, c. \tag{3.27}$$

La ecuación anterior es conocida como primera fórmula de Erlang (o fórmula de retraso). Al llegar una unidad al sitema, ésta se pierde, si al llegar encuentra ocupados todos los canales. La probabilidad de este evento es:

$$p_c = \frac{\left(\frac{\lambda}{\mu}\right)^c \Big/ c!}{\sum_{k=0}^{c} \left(\frac{\lambda}{\mu}\right)^k \Big/ k!}$$

La ecuación anterior es conocida como fórmula de pérdida (de bloqueo o saturación) o fórmula B, denotada por $B\left(c, \frac{\lambda}{\mu}\right)$.

### 3.6.1. Número Esperado de Canales Ocupados

Sea $B$ la variable aleatoria que denota el número de canales ocupados. Tenemos

$$\begin{aligned} E\{B\} &= \sum_{n=1}^{c} n p_n = \sum_{n=1}^{c} \frac{n\left(\frac{\lambda}{\mu}\right)^n}{n!} p_o \\ &= \left(\frac{\lambda}{\mu}\right) p_0 \sum_{n=1}^{c} \frac{\left(\frac{\lambda}{\mu}\right)^{n-1}}{(n-1)!} = \left(\frac{\lambda}{\mu}\right) p_0 \left[\sum_{n=0}^{c} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} - \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}\right] \\ &= \frac{\lambda}{\mu}[1 - p_c] = \frac{\lambda}{\mu}\left[1 - B\left(c, \frac{\lambda}{\mu}\right)\right]. \end{aligned} \tag{3.28}$$

Ahora, sea $I$ la variable aleatoria que denota el número de canales libres, entonces

$$\begin{aligned} E\{I\} &= E\{c - B\} = c - E(B) \\ &= c - \frac{\lambda}{\mu}(1 - p_c) \\ &= c - \frac{\lambda}{\mu}\left[1 - B\left(c, \frac{\lambda}{\mu}\right)\right]. \end{aligned} \tag{3.29}$$

### 3.6.2. Probabilidad de Espera en Cola

Sea $X_i$ la variable indicadora para el el i-ésimo canal elegido aleatoriamente; $X_i = 1$ o 0 según si el i-ésimo canal esté ocupado o libre. Sea $P_c\{A\}$ la probabilidad de ocurrencia de un evento A en uns sistema $M/M/c/c$ en equilibrio. Entonces

(i)

$$P_c\{X_1 = 1, \ldots, X_k = 1\} = \frac{B\left(c, \frac{\lambda}{\mu}\right)}{B\left(c-k, \frac{\lambda}{\mu}\right)}, \qquad 1 \leq k \leq c$$

(ii)

$$P_c\{X_1 = 1\} = \frac{\left(\frac{\lambda}{\mu}\right)\left[1 - B\left(c, \frac{\lambda}{\mu}\right)\right]}{c}, \qquad \text{y}$$

(iii)

$$P_c\{X_{k+1} = 1 \mid X_1 = 1, X_2 = 1, \ldots, X_k = 1\} = P_{c-k}\{X_1 = 1\}$$

*Demostración i)*

Considerando el número de canales ocupados, tenemos

$$P_c\{X_1 = 1, \ldots, X_k = 1\} = \sum_{j=k}^{c} Pr\{X_1 = 1, \ldots, X_k = 1 \mid n = j\} p_j$$

$$= \sum_{j=k}^{c} \frac{\binom{j}{k}}{\binom{c}{k}} p_0 \frac{\left(\frac{\lambda}{\mu}\right)^j}{j!} = \frac{(c-k)!}{c!} p_0 \sum_{j=k}^{c} \frac{\left(\frac{\lambda}{\mu}\right)^j}{(j-k)!}$$

$$= \frac{(c-k)!}{c!} \left(\frac{\lambda}{\mu}\right)^{k-c} \sum_{j=k}^{c} \left[\left(\frac{\lambda}{\mu}\right)^c p_0\right] \frac{\left(\frac{\lambda}{\mu}\right)^{j-k}}{(j-k)!}$$

$$= \left[\frac{\left(\frac{\lambda}{\mu}\right)^c}{c!} p_0\right] \left[\frac{(c-k)!}{\left(\frac{\lambda}{\mu}\right)^{c-k}} \sum_{r=0}^{c-k} \frac{\left(\frac{\lambda}{\mu}\right)^r}{r!}\right]$$

$$= \frac{B\left(c, \frac{\lambda}{\mu}\right)}{B\left(c-k, \frac{\lambda}{\mu}\right)}$$

*Demostración ii)*

$$P_c\{X_1 = 1\} = \sum_{j=1}^{c} P_c\{X_1 = 1 \mid B = j\} Pr\{B = j\}$$

(B= Número de canales ocupados)

$$= \sum_{j=1}^{c} \frac{j}{c} p_j = \frac{1}{c} \sum_{j=1}^{c} j p_j$$

$$= \frac{1}{c} E(B)$$

$$= \frac{\lambda/\mu}{c}\left[1 - B\left(c, \frac{\lambda}{\mu}\right)\right] \quad \text{De la Ec.} \quad 3.28 \tag{3.30}$$

$$= \frac{\lambda/\mu}{c}\left[1 - B\left(c, \frac{\lambda}{\mu}\right)\right] \tag{3.31}$$

$$= \frac{\lambda/\mu}{c}\left[1 - B\left(c, \frac{\lambda}{\mu}\right)\right] \quad \text{De la Ec.} \quad 3.28 \tag{3.32}$$

*Demostración iii)* Una vez más

$$P_c\{X_{k+1} = 1 \mid X_1 = 1, X_2 = 1, \ldots, X_k = 1\}$$

$$= \frac{P_c\{X_1 = 1, X_2 = 1, \ldots, X_k = 1, X_{k+1} = 1\}}{P_c\{X_1 = 1, X_2 = 1, \ldots, X_k = 1\}}$$

$$= \frac{B\left(c, \frac{\lambda}{\mu}\right)}{B\left(c-k-1, \frac{\lambda}{\mu}\right)} \frac{\left(c-k, \frac{\lambda}{\mu}\right)}{B\left(c, \frac{\lambda}{\mu}\right)} = \frac{B\left(c-k, \frac{\lambda}{\mu}\right)}{B\left(c-k-1, \frac{\lambda}{\mu}\right)}$$

$$= \frac{\left[\left(\frac{\lambda}{\mu}\right)^{c-k}(c-k)!\right]\left[\sum_{r=0}^{c-k}\left(\frac{\lambda}{\mu}\right)^{r}/r!\right]^{-1}}{\left[\left(\frac{\lambda}{\mu}\right)^{c-k-1}/(c-k-1)!\right]\left[\sum_{r=0}^{c-k-1}\left(\frac{\lambda}{\mu}\right)^{r}/r!\right]^{-1}}$$

$$= \frac{\frac{\lambda}{\mu}}{(c-k)}\left[1 - \frac{\left(\frac{\lambda}{\mu}\right)^{c-k}/(c-k)!}{\sum_{r=0}^{c-k}\left(\frac{\lambda}{\mu}\right)^{r}/r!}\right]$$

$$= \frac{\frac{\lambda}{\mu}}{(c-k)}\left[1 - B\left(c-k, \frac{\lambda}{\mu}\right)\right]$$

$$= P_{c-k}\{X_1 = 1\}.$$

El resultado (*iii*) implica que, en un sistema $M/M/c/c$, en equilibrio, si $(k+1)$ canales son escogidos aleatoriamente, sin remplazo, y los primeros $k$ canales están ocupados, entonces la probabilidad condicional de que el $(k+1)$ ésimo canal esté ocupado es igual a la probabilidad a priori de que un canal escogido aleatoreamente en un sistema de colas $M/M/c-k/(c-k)$ esté ocupado.

# 4. Aplicación del Modelo. Central Telefónica

## 4.1. Introducción

Llegamos aquí al punto en que aplicaremos toda la teoría desarrollada en los capítulos anteriores. La aplicación será en una pequeña central telefónica, para definir el número óptimo de operadores necesarios, según los periodos de tráfico, que se presentan a lo largo de la jornada laboral. Seguiremos los lineamientos expuestos en el capítulo 2, para definir la eficiencia deseada, de acuerdo con la cual estamos dispuestos a cubrir el costo de operación para poder alcanzarla, luego haremos el análisis, usando el modelo de colas exponencial, expuesto en los capítulos 2 y 3, para determinar el número de agentes que es necesario tener en el sistema para lograr la eficiencia, o calidad del servicio, deseada.

## 4.2. Objetivos de Costo y Eficiencia

Debido a que la empresa telefónica tiene como objetivo brindar prioridad al óptimo nivel de servicio, la eficiencia tendría como prioridad brindar un muy alto nivel de servicio, o llamadas no perdidas, razón por la cual, lo primero que debemos hacer es establecer los objetivos de nivel de servicio, una vez hecho esto, será posible saber cuál será el grado de ocupación necesario para mantener dicho nivel de servicio.

La compañía ha establecido los siguientes objetivos de nivel de servicio:

AWT (Tiempo Aceptable de Espera)= 5 segundos.
Es decir, se espera que el tiempo máximo de espera en cola, para un cliente que encuentre todos los canales ocupados, sea de 5 segundos.

TSF (Factor de Servicio Telefónico)= 99/1.
Esto significa que espera tener solamente uno por ciento de las llamadas por encima de esos cinco segundos de espera.

De acuerdo con las anteriores definiciones, se espera tener un nivel de servicio muy elevado (SL), que es el objetivo principal de la empresa. Aunque, mantener ese nivel de servicio eleva el costo, la empresa está dispuesta a pagarlo, ya que considera más elevado el costo de perder al cliente en el primer intento. Sin embargo, usaremos los modelos de colas para minimizar ese costo. Más adelante veremos cuál es el nivel de servicio que es posible alcanzar con el AWT y TSF que han sido definidos y el costo que debe ser asumido para lograrlo.

## 4.3. Elementos del Sistema

Como hemos mencionado en el capítulo 2, la primera parte consiste en definir los componentes del modelo de colas.

### 4.3.1. Patrón de Llegada de Clientes

En el caso de la compañía se tiene que hay dos periodos durante el día con distintos patrones de entrada. El primero de ellos, que comprende de las 09:30 horas, a las 19:30 horas, y el segundo, que comprende de las 19:30 horas a las 22:00 horas. De modo que se tienen dos variables $\lambda$ para cada uno de estos intervalos, es decir

$$\lambda = \begin{cases} 67 & \text{si } 9{,}5 \leq t \leq 19{,}5 \\ 100 & \text{si } 19{,}5 \leq t \leq 22 \end{cases}$$

De manera que, para hacer el análisis, dividiremos la jornada laboral en dos partes para estudiar por separado cada caso de forma independiente.

Por otra parte, se sabe que los clientes llegarán individualmente.

### 4.3.2. Patrón de Servicio

Después de medir el tiempo de servicio, se sabe que, en promedio cada servicio es completado en, aproximadamente, 80 segundos. Debido a que utilizaremos el modelo exponencial para la explicación de nuestro sistema, consideraremos que este tiempo de servicio se distribuye como una variable aleatoria exponencial. Es así que supondremos que el tiempo de servicio es una variable aleatoria exponencial con media $1/\mu = 80$ segundos.

### 4.3.3. Número de Servidores

Esta es la incógnita fundamental que intentaremos responder con el modelo de colas.
Como un solo servidor, podría atender un máximo de 22.5 llamadas consecutivas ininterrumpidamente, por intervalo de tiempo,

$$(\text{segundos})\ \frac{1800}{80} = 22{,}5$$

Lo redondearemos a 23 para un manejo más fácil.

De acuerdo con lo estudiado en la sección 3.2, teniendo solo servidor, la carga del sistema se desbordaría y la cola crecería infinitamente, incluso para el valor más pequño $\lambda = 67$.

$$\rho = \frac{\lambda}{\mu} = \frac{67}{23} = 2{,}91 > 1$$

Por esto, para que se cumpla el supuesto de estabilidad, será necesario que el sistema tenga al menos tres servidores, ya que con $c \geq 3$

$$\rho = \frac{\rho}{c\mu} = \frac{67}{69} = {,}97 < 1$$

Con tres servidores se reducirá el costo al mínimo, pero el objetivo de nivel de servicio estará muy lejos de los objetivos planteados, es así que debemos encontrar el número $c$ de agentes mínimo necesario, reduciendo así el costo, que permitirá a la empresa alcanzar los objetivos que ha establecido.

### 4.3.4. Capacidad del Sistema

El sistema tiene una capacidad finita de 90 llamadas en cola, correspondientes a tres enlaces telefónicos que puede soportar hasta 90 llamadas esperando, así que tenemos un sistema con capacidad finita de hasta 90 clientes en la cola, sin embargo, debido a que haremos la probabilidad de espera en cola muy pequeña, y a que podrían agregarse líneas de entrada, en caso necesario, consideraremos que se trata de un sistema con capacidad infinita.

### 4.3.5. Disciplina de la Cola

La disciplina de la cola es FCFS (Primero en Llegar Primero en Servicio).

### 4.3.6. Intervalo de Tiempo

Para analizar los flujos de llegada, así como los tiempos de servicio, es necesario dividir el tiempo en periodos pequeños, de forma que sea posible detectar las fluctuaciones y actuar en consecuencia. Es así que en los centros telefónicos se divide el tiempo en intervalos de 15 o 30 minutos. En nuestro caso de estudio, los tiempos están divididos en medias horas, ya que es una manera eficiente de manejar al personal dentro del centro telefónico, además de que es una de las más usuales en cualquier centro de atención telefónica. Así que el intervalo de tiempo que usaremos como unidad será media hora.

Hemos definido así el sistema con sus componentes, y las tres variables principales son las siguientes:

Una tasa de llegada con distribución exponencial y parámetro $\lambda$.
Una tasa de servicio con distribución exponencial y parámetro $\mu$
La carga del sistema

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

## 4.4. Modelo M/M/c para $\lambda = 67$

Utilizaremos el modelo $M/M/c$, para definir el número de posiciones necesario para cubrir la demanda eficientemente. Primeramente, haremos el análisis con tres operadores, con la intención de demostrar que así habría una alta eficiencia, pero un bajo nivel de servicio (SL), para luego buscar el número más adecuado de agentes. En ambos casos se observarán los distintos resultados obtenidos, así como la gráfica de distribución de probabilidad de las llegadas y haremos una simulación para ambos escenarios, que debe ser congruente con el resultado de nuestro modelo de colas al colocar cierto número de agentes atendiendo en paralelo.

Utilizaremos una hoja de cálculo en la que se han introducido los resultados del capítulo 3, para el modelo M/M/c.

### 4.4.1. Modelo M/M/3

En primera instancia, analizaremos el resultado de poner tres agentes sirviendo en paralelo.

Los parámetros que hemos introducido en el modelos son

$$\lambda = 67$$

$$\frac{1}{\mu} = 80 \text{ segundos} \Rightarrow \mu = 23$$

$$AWT = \ 5 \text{ segundos.}$$

Con 3 agentes, observamos que la carga del sistema $\rho$ es de 97 % lo cual representa una alta productividad con pocos recursos, ya que los agentes estarían ocupados el 97 % del tiempo, o tendremos, en promedio, al 97 % de los agentes ocupados. Esto significa que prácticamente todo el tiempo los 3 agentes estarían recibiendo llamadas, sin poder descansar un solo segundo. Sin embargo, a mayor eficiencia, menor nivel de servicio, por lo cual es de esperarse que el nivel de servicio obtenido con este número de operadores sea muy bajo.

Esto último se deduce de que la probabilidad de tener que esperar en la cola $C$ es de 95 %, lo que implica que 95 % de los clientes tendrán que esperar. Además el promedio de tiempo en cola $E(W_Q)$, es de 851 segundos, o sea, alrededor de 14 minutos, lo cual está muy por encima de los 5 segundos establecidos como Tiempo Aceptable de Espera (AWT). Así que casi todos los clientes tendrían que esperar, en promedio, 14 minutos para ser atendidos. Esto último hace que el tiempo promedio en el sistema para los clientes $E(W)$, contando el tiempo en cola más el tiempo de servicio, sea de 929 segundos, casi 16 minutos. Finalmente, la probabilidad de esperar en cola más tiempo del AWT establecido es de $Pr\{W_q \geq t\}$ es de 94 %, cuando el objetivo establecido es que sea menor o igual a 1 %.

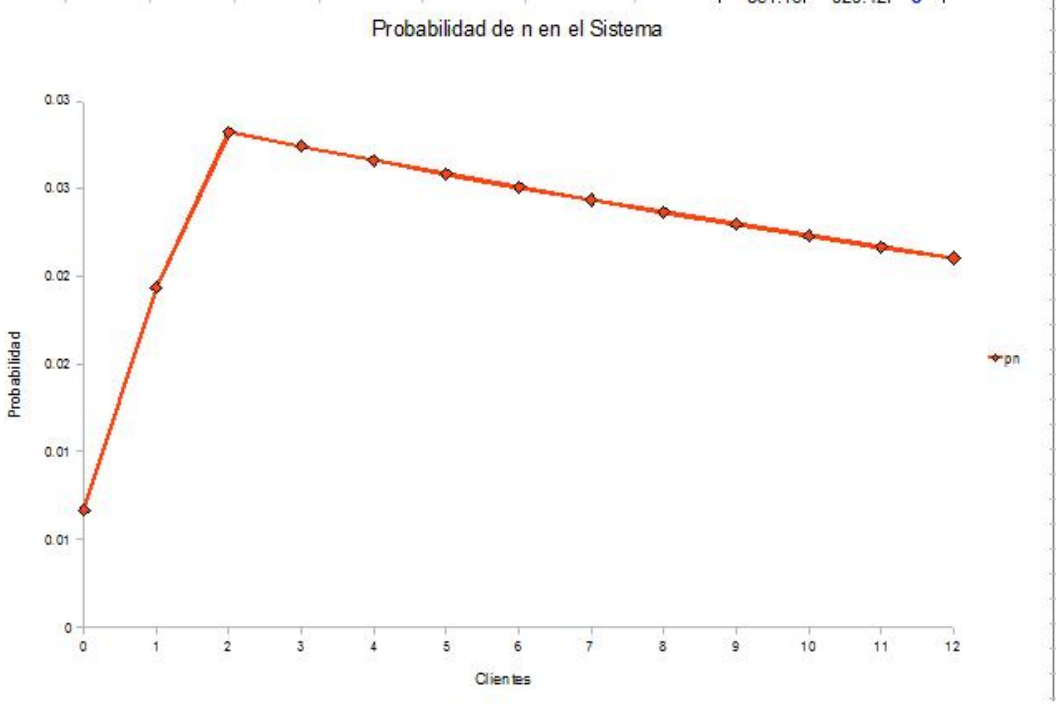
Todo lo anterior, nos muestra cómo poniendo el mínimo número de agentes necesarios para evitar que el sistema se desborde, implica brindar el peor nivel de servicio posible, que en algún tipo de situación podría ser suficiente, sin embargo, en el caso de la compañía de nuestro estudio, significa un nivel de servicio (SL) que está muy lejos de lo que la compañia pretende brindar a sus clientes. Es comprensible que los centros telefónicos establezcan estos objetivos ya que, en casos como el anterior, puede resultar más cara la pérdida de clientes por el mal servicio, que el ahorro que haremos de algunas posiciones.

Observando la distribución de probabilidad del número de clientes en el sistema, es posible apreciar que, debido a que los clientes frecuentemente tendrán que esperar en cola y, el tiempo que deberán estar en la cola será considerable, provoca que haya más clientes en el sistema, con muchos de ellos esperando en la cola.

| $p_0$ | $p_c$ | $\lambda$ | $\mu$ | c | $\rho$ | $1-\rho$ | C | Agentes Ocupados E(B) | Agentes Libres E(I) | E(Q) Clientes | E(N) Clientes | $E(W_Q)$ Tiempo | E(W) Tiempo | t | $p_0$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.006654 | 0.027413 | 67.00 | 23.00 | 3.00 | 0.971014 | 0.02899 | 0.9457 | 2.91 | 0.09 | 31.68 | 34.5951 | 0.47287 | 0.52 | 0.003 | 0.9405 |

| $1/\mu$ | Segs | Segs | AWT Segs |
|---|---|---|---|
| 80 | 851.16 | 929.42 | 5 |

| n | | $p_n$ | $p_0$ | |
|---|---|---|---|---|
| 0 | 1 | 0.01 | 0.0069863 | 0.03 |
| 1 | 2.91 | 0.02 | 0.0068469 | 0.09 |
| 2 | 4.24 | 0.03 | **0.0066536** | 0.15 |
| 3 | 4.12 | **0.03** | 0.0064761 | 0.15 |
| 4 | 3 | 0.03 | 0.0063527 | 0.12 |
| 5 | 1.75 | 0.03 | 0.0062829 | 0.1 |
| 6 | 0.85 | 0.03 | 0.0062496 | 0.08 |
| 7 | 0.35 | 0.02 | 0.0062358 | 0.06 |
| 8 | 0.13 | 0.02 | 0.0062308 | 0.05 |
| 9 | 0.04 | 0.02 | 0.0062292 | 0.04 |
| 10 | 0.01 | 0.02 | 0.0062287 | 0.03 |
| 11 | 0 | 0.02 | 0.0062286 | 0.02 |
| 12 | 0 | 0.02 | 0.0062286 | 0.02 |
| 13 | 0 | 0.02 | 0.0062286 | 0.01 |
| 14 | 0 | 0.02 | 0.0062286 | 0.01 |
| 15 | 0 | 0.02 | 0.0062286 | |
| 16 | 0 | 0.02 | 0.0062286 | |
| 17 | 0 | 0.02 | 0.0062286 | |
| 18 | 0 | 0.02 | 0.0062286 | |
| 19 | 0 | 0.02 | 0.0062286 | |
| 20 | 0 | 0.02 | 0.0062286 | |
| 21 | 0 | 0.02 | 0.0062286 | |
| 22 | 0 | 0.02 | 0.0062286 | |
| 23 | 0 | 0.02 | 0.0062286 | |



### 4.4.2. Simulación del sistema $M/M/3$

Finalmente, hacemos una simulación de 800 llegadas al sistema con los parámetros que definimos en la sección 4.4.1. Obteniendo la tabla de resultados 4, en donde observamos que, en la simulación, la carga del sistema es de 99 % mientras, valo muy cercano al 97 % obtenido con el modelo, el tiempo promedio en el sistema es de 939 segundos (15.7 minutos), contra 929 segundos (15.48 minutos), que esperábamos. Además, en la simulación, el tiempo promedio de espera en cola fue de 864 segundos, habíamos pronosticado 851 segundos con nuestro modelo. El tiempo promedio en servicio en la simulación fue de 75 segundos, lo cual se aproxima a los 80 segundos que planteamos en el modelo de colas. Por otra parte, el número promedio de clientes en servicio es de 2.97 , valor muy próximo 2.91 que esperábamos en el modelo. Lo mismo ocurre con de clientes en el sistema obtenido en la simulación, 38, contra 35 que obtuvimos como resultado, mediante la teoría de colas. Y finalmente, en la simulación hubo 35 clientes en cola, cuando esperábamos 32. Es así que observamos que los resultados de la simulación concuerdan significativamente con lo que habíamos previsto con nuestro modelo, y que en una experiencia real habría generado un sistema absolutamente lleno, con grandes colas, y los operadores trabajando como máquinas lo más rápidamente posible, cosa que no puede ocurrir realmente.

Ahora, graficando la simulación que acabamos de realizar, observamos claramente que prácticamente la totalidad del tiempo habrá clientes en cola, lo que contribuye a que el tiempo en el sistema se alargue mucho más de lo que dura el servicio, ya que el tiempo de servicio es de 75 segundos, pero el promedio de espera en cola es de más de 14 minutos. Es así el número de personas en cola llegó a ser de hasta 15 personas, mientras los tres agentes están ocupados. La simulación nos confirma lo que habíamos ya descubierto, con nuestro modelo, que con cuatro agentes el nivel de servicio sería intolerablemente malo, haciendo esperar a los clientes con demasiada frecuencia, teniendo que esperar durante más tiempo del que tomará su servicio.

### 4.4.3. Modelo M/M/8

Usando la hoja de cálculo, y probando los resultados al aumentar el número de operadores, observamos que con ocho operadores es posible encontrar los resultados teóricos que pretendemos, es decir, un TEF (AWT) de 99/1.Los parámetros que introducimos son los mismos

$$\lambda = 67$$
$$\frac{1}{\mu} = 80 \text{ segundos} \Rightarrow \mu = 23$$
$$AWT = \ 5 \text{ segundos.}$$

Debemos ir aumentando la cantidad de agentes hasta que la probabilidad de espera por encima

| | A | B | C | D | G | H | I | J | K | N | T | U | Z | AA | AB | AC | AF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | **Queueing Simulaton** | | | | | | | | | | | | S | S | S | A | |
| 2 | Next Event-Dynamic | | | | | | | Event | Event Time | Event Name | Sistem a | Cola | S1 Cal. | S2 Cal. | S3 Cal. | Arrive Cal. | Tiempo (Min) |
| 3 | Queue Station | SimQ3b | **Segs.** | Number | TBA | TFS | | 0 | 0 | 0 | 0 | 0 | 10000 | 10000 | 10000 | 0.001 | 0 |
| 4 | Arrival Rate | 67 | | 1 | 0.001 | 0.041 | | 1 | 0.001 | A | 1 | 0 | 0.042 | 10000 | 10000 | 0.002 | 0 |
| 5 | Service Rate | 23 | | 2 | 6E-04 | 0.043 | | 2 | 0.002 | A | 2 | 0 | 0.042 | 0.045 | 10000 | 0.015 | 0 |
| 6 | Number of Servers | 3 | | 3 | 0.013 | 0.012 | | 3 | 0.015 | A | 3 | 0 | 0.042 | 0.045 | 0.028 | 0.039 | 0 |
| 7 | Max. Number in System | *** | | 4 | 0.024 | 0.137 | | 4 | 0.028 | S | 2 | 0 | 0.042 | 0.045 | 1E+05 | 0.039 | 1 |
| 8 | Type | M/M/3 | | 5 | 8E-04 | 0.029 | | 5 | 0.039 | A | 3 | 0 | 0.042 | 0.045 | 0.176 | 0.04 | 1 |
| 9 | Arrival Seed | *** | | 6 | 0.005 | 0.001 | | 6 | 0.04 | A | 4 | 1 | 0.042 | 0.045 | 0.176 | 0.045 | 1 |
| 10 | Service Seed | *** | | 7 | 0.005 | 0.012 | | 7 | 0.042 | S | 3 | 0 | 0.071 | 0.045 | 0.176 | 0.045 | 1 |
| 11 | Number in Simulation | 800 | | 8 | 0.005 | 0.015 | | 8 | 0.045 | A | 4 | 1 | 0.071 | 0.045 | 0.176 | 0.049 | 1 |
| 12 | Start Data Time | 0 | | 9 | 0.01 | 0.007 | | 9 | 0.045 | S | 3 | 0 | 0.071 | 0.046 | 0.176 | 0.049 | 1 |
| 13 | Stop Data Time | 11.134202 | | 10 | 0.001 | 0.007 | | 10 | 0.046 | S | 2 | 0 | 0.071 | 1E+05 | 0.176 | 0.049 | 1 |
| 14 | Mean Number at Station | 37.437828 | | 11 | 0.016 | 0.021 | | 11 | 0.049 | A | 3 | 0 | 0.071 | 0.061 | 0.176 | 0.055 | 1 |
| 15 | Mean Time at Station | 0.5217026 | 939 | 12 | 0.019 | 0.022 | | 12 | 0.055 | A | 4 | 1 | 0.071 | 0.061 | 0.176 | 0.064 | 2 |
| 16 | Mean Number in Queue | 34.460156 | | 13 | 0.005 | 0.081 | | 13 | 0.061 | S | 3 | 0 | 0.071 | 0.076 | 0.176 | 0.064 | 2 |
| 17 | Mean Time in Queue | 0.4802082 | 864 | 14 | 0.028 | 0.005 | | 14 | 0.064 | A | 4 | 1 | 0.071 | 0.076 | 0.176 | 0.066 | 2 |
| 18 | Mean Number in Service | 2.9776728 | | 15 | 0.005 | 4E-04 | | 15 | 0.066 | A | 5 | 2 | 0.071 | 0.076 | 0.176 | 0.082 | 2 |
| 19 | Mean Time in Service | 0.0414944 | 75 | 16 | 0.007 | 0.043 | | 16 | 0.071 | S | 4 | 1 | 0.078 | 0.076 | 0.176 | 0.082 | 2 |
| 20 | Arrival Rate | 71.760866 | | 17 | 0.007 | 0.211 | | 17 | 0.076 | S | 3 | 0 | 0.078 | 0.083 | 0.176 | 0.082 | 2 |
| 21 | Throughput Rate | 71.760866 | | 18 | 0.016 | 0.047 | | 18 | 0.078 | S | 2 | 0 | 1E+05 | 0.083 | 0.176 | 0.082 | 2 |
| 22 | Efficiency | 0.9925576 | | 19 | 0.004 | 0.053 | | 19 | 0.082 | A | 3 | 0 | 0.104 | 0.083 | 0.176 | 0.101 | 2 |
| 23 | Probability Balk | 0 | | 20 | 0.033 | 0.13 | | 20 | 0.083 | S | 2 | 0 | 0.104 | 1E+05 | 0.176 | 0.101 | 2 |
| 24 | | | | 21 | 0.012 | 0.013 | | 21 | 0.101 | A | 3 | 0 | 0.104 | 0.124 | 0.176 | 0.106 | 3 |
| 25 | | | | 22 | 0.017 | 0.007 | | 22 | 0.104 | S | 2 | 0 | 1E+05 | 0.124 | 0.176 | 0.106 | 3 |
| 26 | | | | 23 | 0.003 | 0.085 | | 23 | 0.106 | A | 3 | 0 | 0.187 | 0.124 | 0.176 | 0.134 | 3 |
| 27 | | | | 24 | 0.038 | 0.034 | | 24 | 0.124 | S | 2 | 0 | 0.187 | 1E+05 | 0.176 | 0.134 | 4 |
| 28 | | | | 25 | 0.003 | 0.066 | | 25 | 0.134 | A | 3 | 0 | 0.187 | 0.139 | 0.176 | 0.139 | 4 |
| 29 | | | | 26 | 0.011 | 0.002 | | 26 | 0.139 | A | 4 | 1 | 0.187 | 0.139 | 0.176 | 0.145 | 4 |
| 30 | | | | 27 | 0.002 | 0.002 | | 27 | 0.139 | S | 3 | 0 | 0.187 | 0.14 | 0.176 | 0.145 | 4 |
| 31 | | | | 28 | 0.005 | 0.066 | | 28 | 0.14 | S | 2 | 0 | 0.187 | 1E+05 | 0.176 | 0.145 | 4 |
| 32 | | | | 29 | 0.013 | 0.018 | | 29 | 0.145 | A | 3 | 0 | 0.187 | 0.188 | 0.176 | 0.152 | 4 |
| 33 | | | | 30 | 0.065 | 0.019 | | 30 | 0.152 | A | 4 | 1 | 0.187 | 0.188 | 0.176 | 0.168 | 5 |
| 34 | | | | 31 | 0.098 | 0.005 | | 31 | 0.168 | A | 5 | 2 | 0.187 | 0.188 | 0.176 | 0.172 | 5 |
| 35 | | | | 32 | 0.015 | 0.013 | | 32 | 0.172 | A | 6 | 3 | 0.187 | 0.188 | 0.176 | 0.206 | 5 |

Figura 4: Simulación 3 Agentes

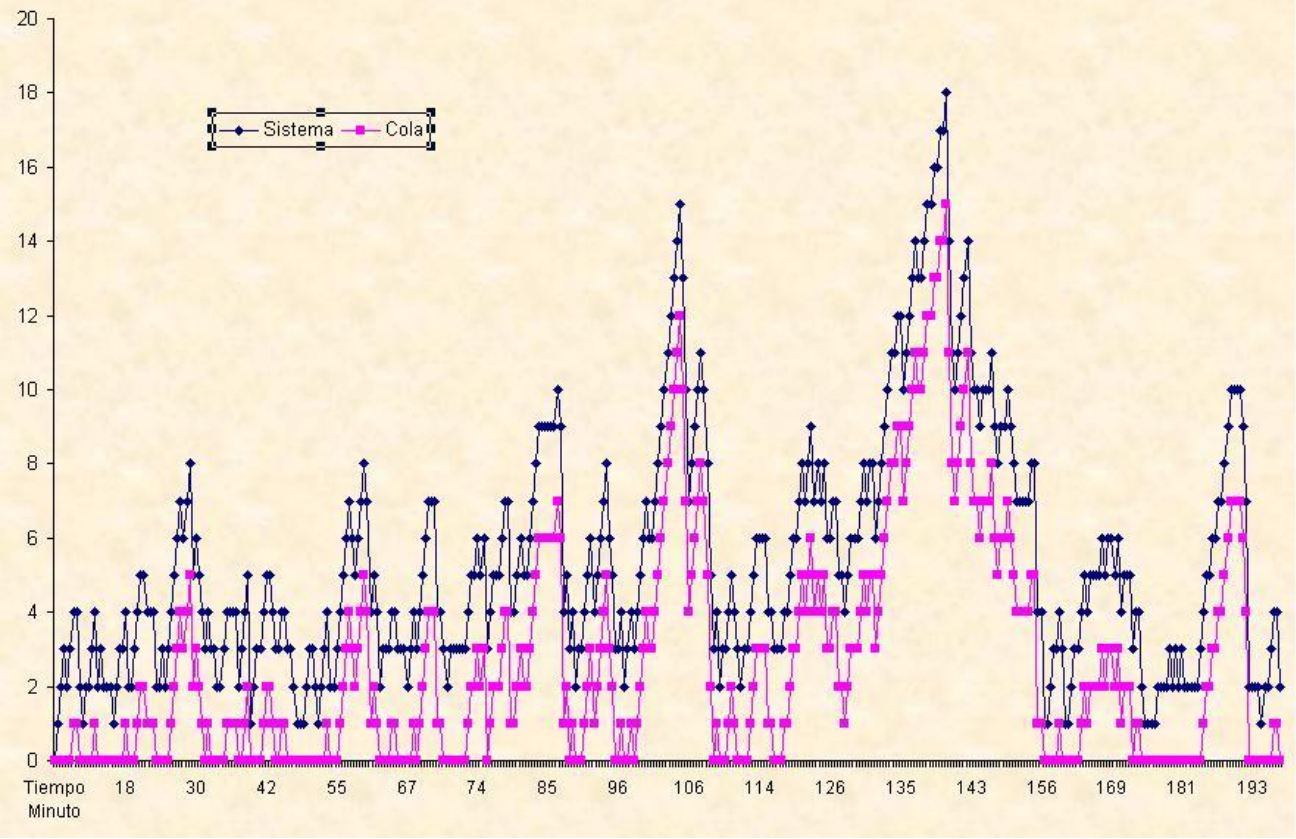



Figura 5: Simulación 800 llegadas 3 Canales

del AWT sea menor a 1 %. Mediante este procedimiento es que hemos observado que el número mínimo de agentes necesarios para cumplir esta condición, bajo los supuestos establecidos, es de 8.

Ahora, con 8 agentes, observamos que la carga del sistema $\rho$ es de 36 % lo cual significa que los agentes estarán atendiendo llamadas un tercio del tiempo, o que, en promedio, estarán ocupados un tercio de los 8 agentes. Así que los operadores estarán desocupados, en promedio, $1 - \rho =$ 64 % del tiempo, sin embargo, es el precio que la compañía está dispuesta a pagar para brindar un excelente nivel de servicio.

Con los 8 agentes el número promedio de agentes ocupados $B$ es el mismo 2.9, mientras que el de agentes desocupados $E(I)$ es de 5.09. El número promedio de clientes en el sistema $E(N)$ sería de 2.91. Ahora, con 8 agentes el número promedio de clientes en cola es de casi cero y el tiempo promedio en cola $E(W_Q)$sería de sólo .17 segundos.

La probabilidad de encontrar a todos los agentes ocupados, y tener que esperar en cola, $C$ es apenas un poco más de 1 %.Por último, la probabilidad de esperar en cola más tiempo del AWT establecido es de $Pr\{W_q \geq t\}$ es de 0,80 %, cantidad que está por debajo del 1 % establecido.

Todo lo anterior concuerda también con la gráfica de la distribución de probabilidad de clientes en el sistema, en la que observamos que con ocho agentes, quedará suficientemente bien cubierta la demanda de servicio.

De acuerdo con esto, será necesario tener 8 posiciones de forma constante para poder lograr el nivel de servicio establecido por la compañía.

| | $P_0$ | $P_c$ | $\lambda$ | $\mu$ | c | $\rho$ | $1-\rho$ | C Pr{W$q$>0} | Agentes Ocupados E(B) | Agentes Libres E(I) | E(Q) Clientes | E(N) Clientes | E($W_q$) Tiempo | E(W) Tiempo | $t$ | TSF Pr{W$q$>$t$} |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.054264 | 0.006979 | 67.00 | 23.00 | 8.00 | 0.364130 | 0.63587 | 0.0110 | 2.91 | 5.09 | 0.01 | 2.9193 | 0.00009 | 0.04 | 0.003 | 0.0079 |

| 1/$\mu$ |
|---|
| 80 |

| Segs | Segs | AWT Segs |
|---|---|---|
| 0.17 | 78.43 | 5 |

| n | $P_n$ | $P_0$ | |
|---|---|---|---|
| 0 | 1 | 0.8317742 | 0.03 |
| 1 | 2.91 | 0.2429961 | 0.09 |
| 2 | 4.24 | 0.1196429 | 0.15 |
| 3 | 4.12 | 0.0801402 | 0.15 |
| 4 | 3 | 0.0646057 | 0.12 |
| 5 | 1.75 | 0.0580499 | 0.1 |
| 6 | 0.85 | 0.0553243 | 0.08 |
| 7 | 0.35 | 0.0542640 | 0.06 |
| 8 | 0.13 | 0.0538879 | 0.05 |
| 9 | 0.04 | 0.0537673 | 0.04 |
| 10 | 0.01 | 0.0537323 | 0.03 |
| 11 | 0 | 0.0537230 | 0.02 |
| 12 | 0 | 0.0537207 | 0.02 |
| 13 | 0 | 0.0537202 | 0.01 |
| 14 | 0 | 0.0537201 | 0.01 |
| 15 | 0 | 0.0537201 | |
| 16 | 0 | 0.0537201 | |
| 17 | 0 | 0.0537201 | |
| 18 | 0 | 0.0537201 | |
| 19 | 0 | 0.0537201 | |
| 20 | 0 | 0.0537201 | |
| 21 | 0 | 0.0537201 | |
| 22 | 0 | 0.0537201 | |
| 23 | 0 | 0.0537201 | |

**Queueing Simulaton**
Next Event-Dynamic

| | |
|---|---|
| Queue Station | Sim8Oper |
| Arrival Rate | 67 |
| Service Rate | 23 |
| Number of Servers | 8 |
| Max. Number in System | *** |
| Type | M/M/8 |
| Arrival Seed | *** |
| Service Seed | *** |
| Number in Simulation | 800 |
| Start Data Time | 0 |
| Stop Data Time | 11.948481 |
| Mean Number at Station | 2.9974454 |
| Mean Time at Station | 0.0448247 |
| Mean Number in Queue | 0.0313718 |
| Mean Time in Queue | 0.0004691 |
| Mean Number in Service | 2.9660736 |
| Mean Time in Service | 0.0443555 |
| Arrival Rate | 66.870425 |
| Throughput Rate | 66.870425 |
| Efficiency | 0.3707592 |
| Probability Balk | 0 |

80.684423
0.8444582
79.839965

| Number | TBA | TFS |
|---|---|---|
| 1 | 0.005 | 0.061 |
| 2 | 0.008 | 0.024 |
| 3 | 0.01 | 0.082 |
| 4 | 0.02 | 0.018 |
| 5 | 0.024 | 0.072 |
| 6 | 0.034 | 0.014 |
| 7 | 0.015 | 0.053 |
| 8 | 0.042 | 0.137 |
| 9 | 0.008 | 0.027 |
| 10 | 0.013 | 0.013 |
| 11 | 0.004 | 0.028 |
| 12 | 0.005 | 0.04 |
| 13 | 0.009 | 0.002 |
| 14 | 0.002 | 0.016 |
| 15 | 0.003 | 0.036 |
| 16 | 0.006 | 0.02 |
| 17 | 0.008 | 0.045 |
| 18 | 0.01 | 8E-04 |
| 19 | 0.004 | 0.013 |
| 20 | 0.008 | 0.01 |
| 21 | 0.001 | 0.032 |
| 22 | 0.046 | 0.017 |
| 23 | 0.011 | 0.001 |
| 24 | 0.026 | 0.142 |
| 25 | 0.035 | 0.002 |
| 26 | 0.003 | 0.045 |
| 27 | 0.02 | 0.088 |
| 28 | 0.012 | 0.047 |

| Event | Event Time | Event Name | Sistema | Cola | S S1 Cal. | S S2 Cal. | S S3 Cal. | S S4 Cal. | S S5 Cal. | S S6 Cal. | S S7 Cal. | S S8 Cal. | A Arrive Cal. | Tiempo (Min) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 0.005 | 0 |
| 1 | 0.005 | A | 1 | 0 | 0.066 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 0.013 | 0 |
| 2 | 0.013 | A | 2 | 0 | 0.066 | 0.038 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 0.024 | 0 |
| 3 | 0.024 | A | 3 | 0 | 0.066 | 0.038 | 0.106 | 10000 | 10000 | 10000 | 10000 | 10000 | 0.044 | 0 |
| 4 | 0.038 | S | 2 | 0 | 0.066 | 1E+05 | 0.106 | 10000 | 10000 | 10000 | 10000 | 10000 | 0.044 | 0 |
| 5 | 0.044 | A | 3 | 0 | 0.066 | 0.062 | 0.106 | 10000 | 10000 | 10000 | 10000 | 10000 | 0.068 | 0 |
| 6 | 0.062 | S | 2 | 0 | 0.066 | 1E+05 | 0.106 | 10000 | 10000 | 10000 | 10000 | 10000 | 0.068 | 1 |
| 7 | 0.066 | S | 1 | 0 | 1E+05 | 1E+05 | 0.106 | 10000 | 10000 | 10000 | 10000 | 10000 | 0.068 | 1 |
| 8 | 0.068 | A | 2 | 0 | 0.139 | 1E+05 | 0.106 | 10000 | 10000 | 10000 | 10000 | 10000 | 0.101 | 1 |
| 9 | 0.101 | A | 3 | 0 | 0.139 | 0.115 | 0.106 | 10000 | 10000 | 10000 | 10000 | 10000 | 0.116 | 1 |
| 10 | 0.106 | S | 2 | 0 | 0.139 | 0.115 | 1E+05 | 10000 | 10000 | 10000 | 10000 | 10000 | 0.116 | 1 |
| 11 | 0.115 | S | 1 | 0 | 0.139 | 1E+05 | 1E+05 | 10000 | 10000 | 10000 | 10000 | 10000 | 0.116 | 1 |
| 12 | 0.116 | A | 2 | 0 | 0.139 | 0.169 | 1E+05 | 10000 | 10000 | 10000 | 10000 | 10000 | 0.158 | 1 |
| 13 | 0.139 | S | 1 | 0 | 1E+05 | 0.169 | 1E+05 | 10000 | 10000 | 10000 | 10000 | 10000 | 0.158 | 1 |
| 14 | 0.158 | A | 2 | 0 | 0.295 | 0.169 | 1E+05 | 10000 | 10000 | 10000 | 10000 | 10000 | 0.166 | 2 |
| 15 | 0.166 | A | 3 | 0 | 0.295 | 0.169 | 0.193 | 10000 | 10000 | 10000 | 10000 | 10000 | 0.179 | 2 |
| 16 | 0.169 | S | 2 | 0 | 0.295 | 1E+05 | 0.193 | 10000 | 10000 | 10000 | 10000 | 10000 | 0.179 | 2 |
| 17 | 0.179 | A | 3 | 0 | 0.295 | 0.192 | 0.193 | 10000 | 10000 | 10000 | 10000 | 10000 | 0.183 | 2 |
| 18 | 0.183 | A | 4 | 0 | 0.295 | 0.192 | 0.193 | 0.211 | 10000 | 10000 | 10000 | 10000 | 0.188 | 2 |
| 19 | 0.188 | A | 5 | 0 | 0.295 | 0.192 | 0.193 | 0.211 | 0.228 | 10000 | 10000 | 10000 | 0.197 | 2 |
| 20 | 0.192 | S | 4 | 0 | 0.295 | 1E+05 | 0.193 | 0.211 | 0.228 | 10000 | 10000 | 10000 | 0.197 | 2 |
| 21 | 0.193 | S | 3 | 0 | 0.295 | 1E+05 | 1E+05 | 0.211 | 0.228 | 10000 | 10000 | 10000 | 0.197 | 2 |
| 22 | 0.197 | A | 4 | 0 | 0.295 | 0.198 | 1E+05 | 0.211 | 0.228 | 10000 | 10000 | 10000 | 0.199 | 2 |
| 23 | 0.198 | S | 3 | 0 | 0.295 | 1E+05 | 1E+05 | 0.211 | 0.228 | 10000 | 10000 | 10000 | 0.199 | 2 |
| 24 | 0.199 | A | 4 | 0 | 0.295 | 0.215 | 1E+05 | 0.211 | 0.228 | 10000 | 10000 | 10000 | 0.202 | 2 |
| 25 | 0.202 | A | 5 | 0 | 0.295 | 0.215 | 0.238 | 0.211 | 0.228 | 10000 | 10000 | 10000 | 0.207 | 2 |
| 26 | 0.207 | A | 6 | 0 | 0.295 | 0.215 | 0.238 | 0.211 | 0.228 | 0.227 | 10000 | 10000 | 0.215 | 2 |
| 27 | 0.211 | S | 5 | 0 | 0.295 | 0.215 | 0.238 | 1E+05 | 0.228 | 0.227 | 10000 | 10000 | 0.215 | 2 |
| 28 | 0.215 | S | 4 | 0 | 0.295 | 1E+05 | 0.238 | 1E+05 | 0.228 | 0.227 | 10000 | 10000 | 0.215 | 2 |

Figura 6: Simulación 8 Agentes

### 4.4.4. Simulación del sistema $M/M/8$

Nuevamente hacemos una simulación de 800 llegadas al sistema con los parámetros que definimos en la sección 4.4.1. Obteniendo la tabla de resultados 6, en donde observamos que, en la simulación, la carga del sistema es de 37 %, casi igual al 38 % obtenido antes, el tiempo promedio en el sistema es de 81 segundos, cuando en nuestro modeloesperabamos 79 segundos. Además, en la simulación, el tiempo promedio de espera en cola fue de 0,84 segundos, habíamos pronosticado 0,17 segundos con nuestro modelo. El tiempo promedio en servicio en la simulación fue de 79,83 segundos, lo cual concuerda con los 80 segundos que planteamos en el modelo de colas. Por otra parte, el número promedio de clientes en servicio es de 2,96 , valor muy próximo al 2,91 que esperábamos en el modelo. Lo mismo ocurre con el número de clientes en el sistema obtenido en la simulación 2,99, contra 2,92 que obtuvimos como resultado, mediante la teoría de colas.El tiempo en el sistema es de 80,5 segundos, casi igual a los 78,6 que teníamos. Y finalmente, en la simulación el tiempo promedio en cola es de 0,84 segundos, contra 0,17 segundos obtenidos anteriormente. Observamos que los resultados de la simulación concuerdan significativamente con lo que habíamos previsto con nuestro modelo.

Ahora, graficando la simulación que acabamos de realizar, observamos claramente que hay muy pocos periodos con clientes en cola, llegando a ser éstas de hasta 7 clientes en espera, sin embargo, la cantidad de clientes que esperan por encima de los cinco segundos establecidos no es más del 1 %, lo cual está dentro de los parámetros del Nivel de Servicio (SL) establecido. Habrá que ver qué tan cercanos son los resultados que hemos obtenido, teóricamente y a través de simulaciones, al comportamiento real de las llamadas telefónicas durante una jornada en el centro telefónico.

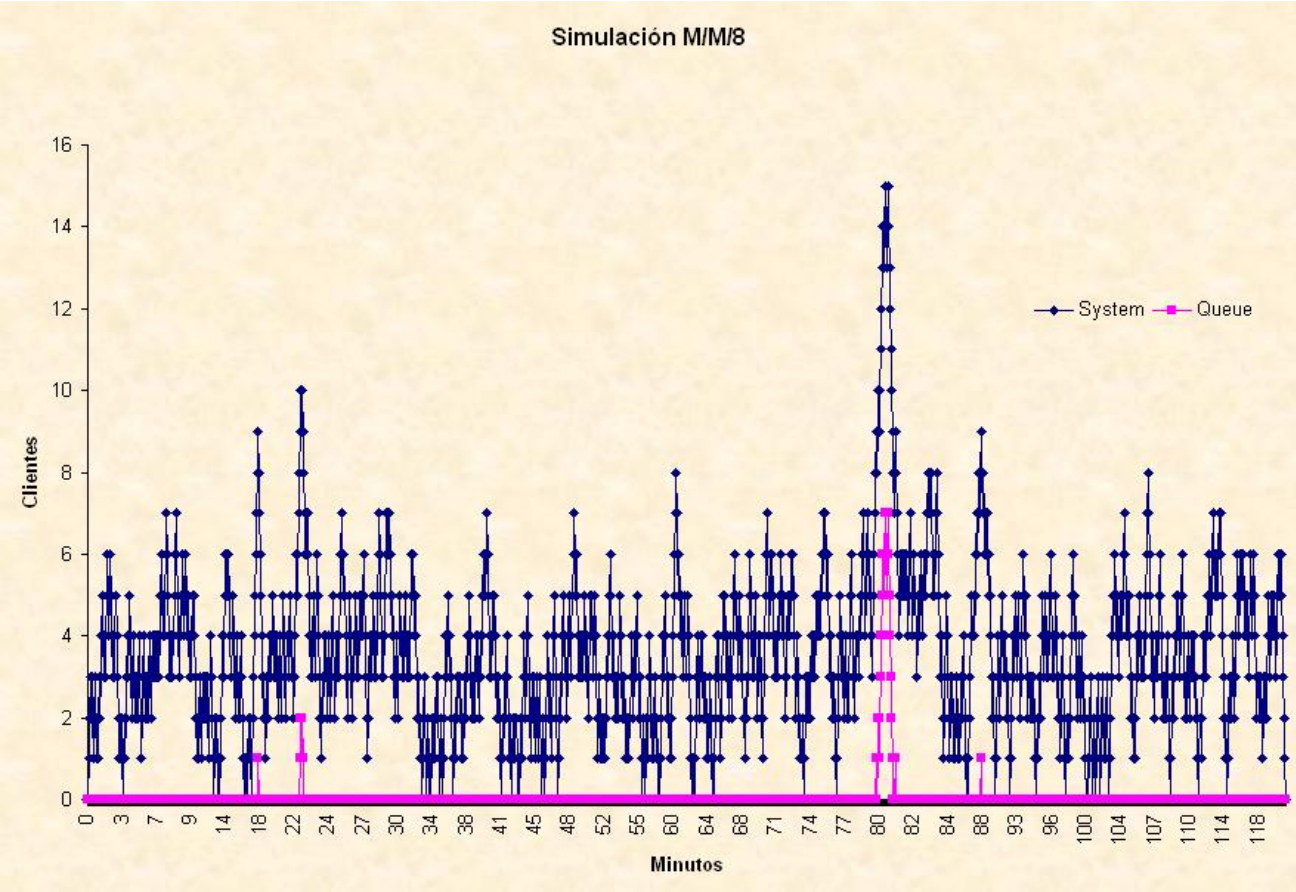



Figura 7: Simulación 800 llegadas 8 Canales

## 4.5. Consideraciones Finales

Hemos llegado a la conclusión de que con 8 agentes atendiendo de manera constante el tráfico de llamadas serán suficientes para lograr los objetivos de la compañía.

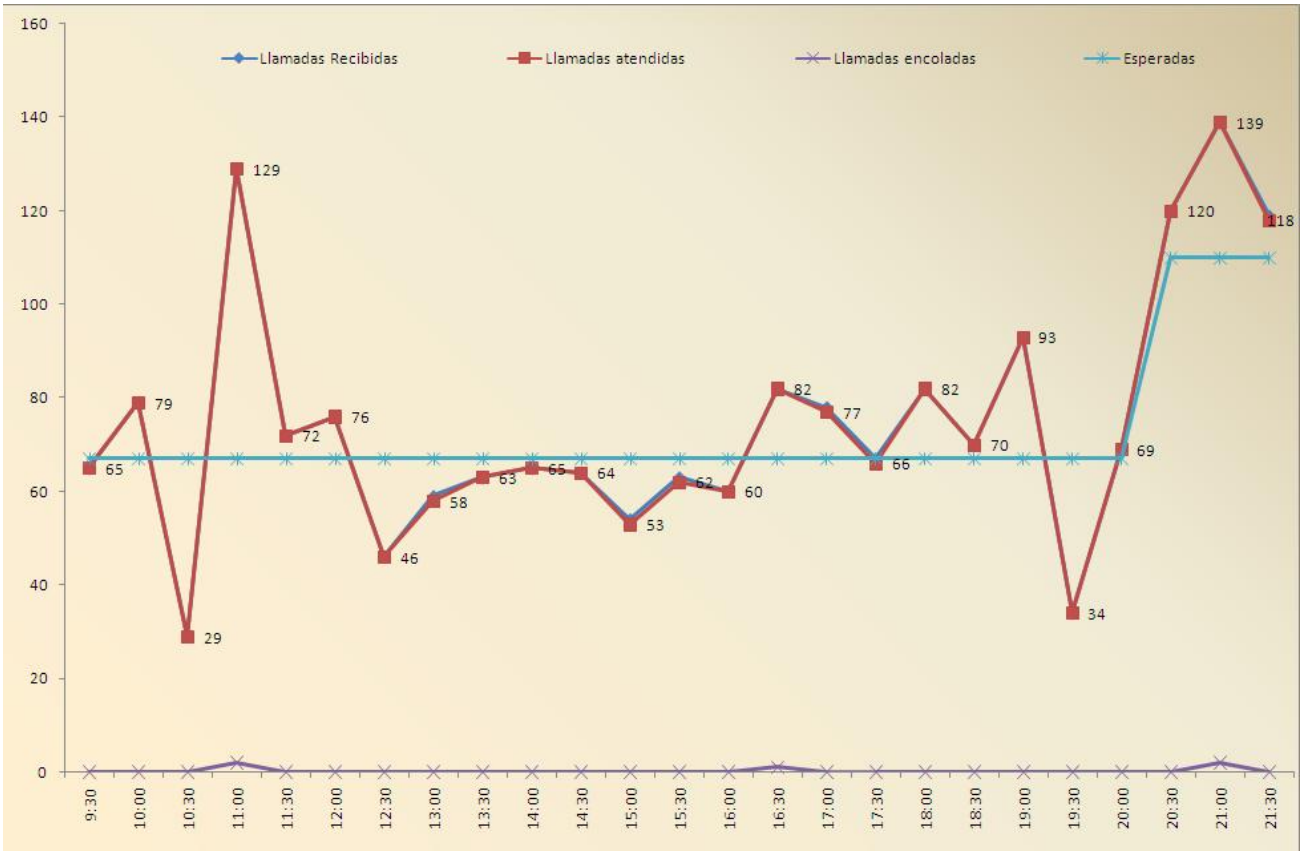



Figura 8: Tráfico Durante la Jornada

## 4.6. Contraste con el tráfico Real para el caso M/M/8

Vamos a contrastar la información real en el centro telefónico con lo que hemos pronosticado con nuestro modelo exponencial de colas y mediante la simulación.

Abajo tenemos (tabla 9) la información, en medias horas del tráfico telefónico durante la jornada. Así como una gráfica del flujo de llamadas durante la jornada (Figura 8). Es posible observar, claramente, que el flujo se distribuye de dos maneras distintas, una hasta las 19:30 horas y otra hasta el final de la jornada. Los distintos fluos se notan an la cantidad de llamadas recibidas, así como en la duración, o tiempo de servicio (AHT IN) de las mismas. Así que el procedimiento de separar el tráfico de la jornada en dos porciones para analizarlas de manera separada ha sido correcto.

Observemos la primera columna: *Llamadas recibidas*. En promedio llegaron 70 llamadas por intervalo de media hora, muy cerca de los 67 que esperábamos, esto es muy bueno, ya que es uno de los parámetros fundamentales para que el modelo y la realidad sean lo más parecidos posible.

| Hora | Llamadas Recibidas | Llamadas atendidas | % Abandono | Abandono Temprano < 5 Segs | Llamadas encoladas | SL | Llamadas Atendidas < 20 Segs | Llamadas Atendidas > 20 segs | AHT IN [segundos] | Ejecutivos Conectados | % Ocupación |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 9:30 - 10:00 | 65 | 65 | 0% | 0 | 0 | 100% | 65 | 0 | 74 | 8 | 33% |
| 10:00 - 10:30 | 79 | 79 | 0% | 0 | 0 | 100% | 79 | 0 | 70 | 8 | 38% |
| 10:30 - 11:00 | 29 | 29 | 0% | 0 | 0 | 100% | 29 | 0 | 70 | 8 | 14% |
| 11:00 - 11:30 | 129 | 129 | 0% | 0 | 2 | 100% | 129 | 0 | 80 | 8 | 72% |
| 11:30 - 12:00 | 72 | 72 | 0% | 0 | 0 | 100% | 72 | 0 | 79 | 8 | 40% |
| 12:00 - 12:30 | 76 | 76 | 0% | 0 | 0 | 100% | 76 | 0 | 71 | 8 | 37% |
| 12:30 - 13:00 | 46 | 46 | 0% | 0 | 0 | 100% | 46 | 0 | 77 | 8 | 25% |
| 13:00 - 13:30 | 59 | 58 | 2% | 1 | 0 | 98% | 59 | 0 | 73 | 8 | 29% |
| 13:30 - 14:00 | 63 | 63 | 0% | 0 | 0 | 100% | 63 | 0 | 69 | 8 | 30% |
| 14:00 - 14:30 | 65 | 65 | 0% | 0 | 0 | 100% | 65 | 0 | 74 | 8 | 33% |
| 14:30 - 15:00 | 64 | 64 | 0% | 0 | 0 | 100% | 64 | 0 | 80 | 8 | 36% |
| 15:00 - 15:30 | 54 | 53 | 2% | 1 | 0 | 98% | 54 | 0 | 82 | 8 | 30% |
| 15:30 - 16:00 | 63 | 62 | 2% | 1 | 0 | 98% | 63 | 0 | 68 | 8 | 29% |
| 16:00 - 16:30 | 60 | 60 | 0% | 0 | 0 | 100% | 60 | 0 | 71 | 8 | 30% |
| 16:30 - 17:00 | 82 | 82 | 0% | 0 | 1 | 100% | 82 | 0 | 74 | 8 | 42% |
| 17:00 - 17:30 | 78 | 77 | 1% | 1 | 0 | 99% | 78 | 0 | 69 | 8 | 37% |
| 17:30 - 18:00 | 67 | 66 | 1% | 1 | 0 | 99% | 67 | 0 | 77 | 8 | 35% |
| 18:00 - 18:30 | 82 | 82 | 0% | 0 | 0 | 100% | 82 | 0 | 73 | 8 | 42% |
| 18:30 - 19:00 | 70 | 70 | 0% | 0 | 0 | 100% | 70 | 0 | 71 | 8 | 35% |
| 19:00 - 19:30 | 93 | 93 | 0% | 0 | 0 | 100% | 93 | 0 | 75 | 8 | 48% |
| **19:30 - 20:00** | **34** | **34** | **0%** | **0** | **0** | **100%** | 34 | **0** | **104** | **14** | **14%** |
| **20:00 - 20:30** | **69** | **69** | **0%** | **0** | **0** | **100%** | 69 | **0** | **107** | **14** | **29%** |
| **20:30 - 21:00** | **120** | **120** | **0%** | **0** | **0** | **100%** | 120 | **0** | **117** | **14** | **56%** |
| **21:00 - 21:30** | **139** | **139** | **0%** | **1** | **2** | **100%** | 139 | **0** | **132** | **14** | **73%** |
| **21:30 - 22:00** | **119** | **118** | **1%** | **0** | **0** | **99%** | 119 | **0** | **116** | **14** | **54%** |
| Total del día | 1,877 | 1,871 | 0% | 6 | 5 | 100% | 1,877 | 0 | 82 | 230 | 38% |

36%

45%

Figura 9: Tabla de Resultados

Ahora revisemos la duración de las llamadas o tiempo de servicio, este es el otro parámetro fundamental de nuestro modelo, y es por eso que es el segundo valor que revisamos. Este valor se encuentra en la columna *AHT IN [segundos]* (Average Handling Time o Tiempo Promedio de Atención), en donde durante el día, las llamadas tardaron, en promedio, 73 segundos, esto está por debajo de los 80 segundos que pronosticamos, ésto puede generar que nos hayamos excedido un poco en la cantidad de agentes colocados. Sin embargo, al introducir una duración de llamada de 73 segundos en el modelo, resulta que el número de agentes necesarios sigue siendo de 8, lo cual indica que no habrá gran diferencia entre lo que calculamos anteriormente y la realidad.

El nivel de ocupación promedio es del 36 %, valor que coincide con el valor de carga del sistema $\rho$ que obtuvimos en nuestro modelo.

Finalmente, revisaremos el número de llamadas encoladas, que es de solamente 5 llamadas, de donde tenemos que

$$5/1817 = 0{,}28\,\%$$

Este número está cerca del valor obtenido $C$=1.10 % de proporción de llamadas en espera.

Aunque el ACD (Automatic Call Distributor) no nos permite saber el tiempo en cola, al ser el total de llamadas menora al 1 % hemos logrado el objetivo de nivel de servicio (SL) que pretendía

no permitir que más del 1 % de las llamadas tuvieran que esperar más de 5 segundos.

Hemos así colocado el mínimo número posible de agentes para lograr el objetivo de servicio, durante el primer periodo de tiempo en el que el flujo de llamadas es de 67 llamadas cada media hora.

## 4.7. Modelo M/M/c para $\lambda = 100$

Debido a que al final del día, tanto el flujo de llegadas como la duración de las mismas, será diferente al del resto del día, consideraremos que, tanto las llegadas como la duración, tienen otra distribución, razón por la cual es necesario hacer el mismo análisis, en forma independiente para esta porción de la jornada.

Nuevamente haremos lo que ya hicimos en las sección 4.4.4, para definir el número de posiciones necesario para cubrir la demanda eficientemente.

### 4.7.1. Modelo M/M/14

Utilizando la experiencia de jornadas anteriores, sabemos que el flujo de llegadas en la segunda parte de la jornada es aproximadamente de 100 llamadas por periodo de tiempo de media hora. Igualmente, sabemos que el tiempo de servicio será aproximadamente de 120 segundos, así que el número máximo de servicios que podría atender un solo agente es de 15.

$$(\text{segundos})\ \frac{1800}{120} = 15$$

Nuevamente, usando los resultados de los capítulos anteriores en la hoja de cálculo, junto con los nuevos parámetros de llegada y tiempo de servicio, observamos, al ir aumentando la cantidad de agentes hasta que la probabilidad de espera por encima del AWT sea menor a 1 %, que catorce es el número mínimo de operadores para que la probabilidad de espera, mayor a cinco segundos, sea menor a 1 %. Los parámetros introducidos son:

$$\lambda = 100$$
$$\frac{1}{\mu} = 120 \text{ segundos} \Rightarrow \mu = 15$$
$$AWT = \ 5 \text{ segundos.}$$

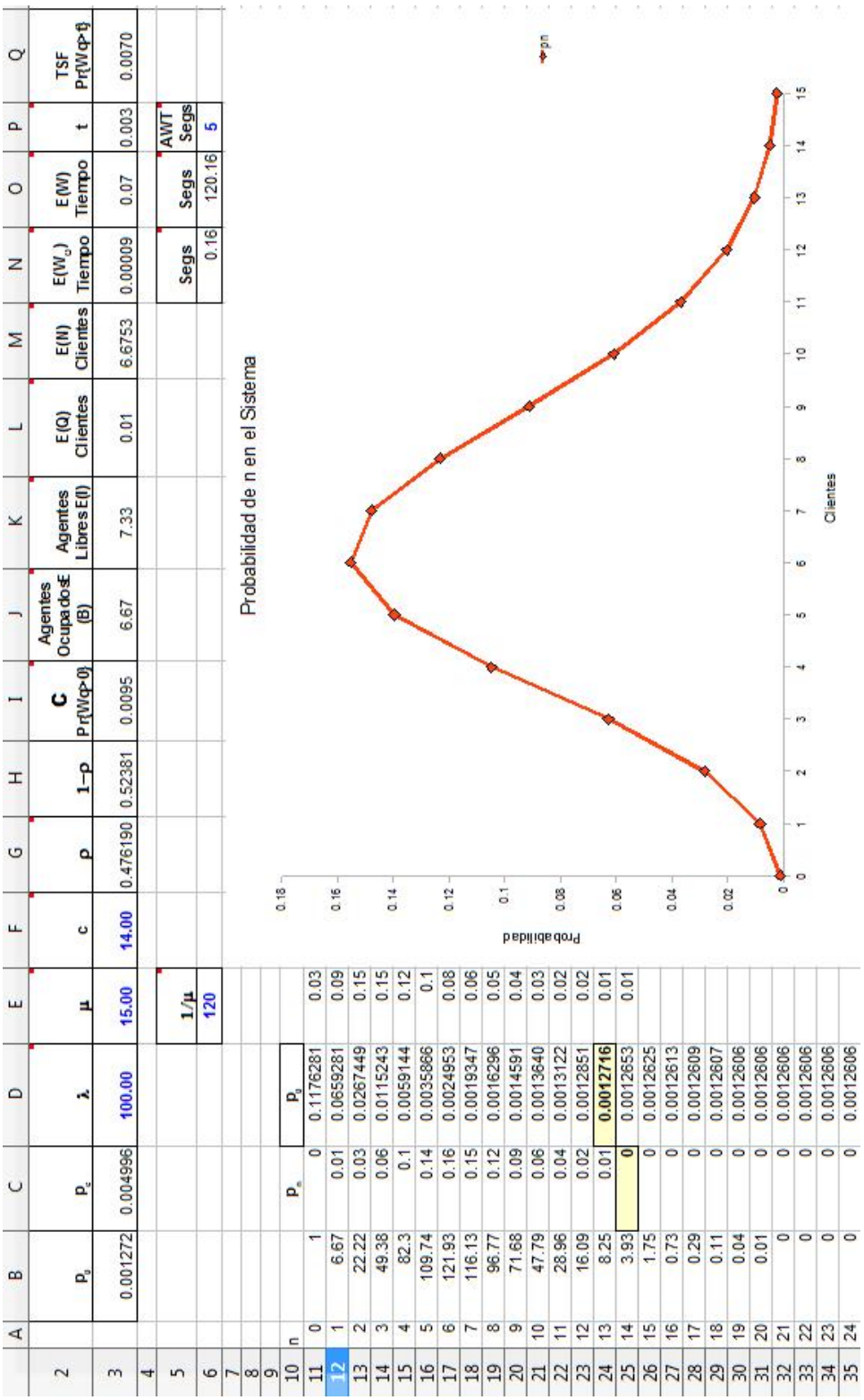
Ahora, con 14 agentes, podemos ver que la carga del sistema $\rho$ es de 48 % lo cual significa que los agentes estarán atendiendo llamadas 48 % del tiempo, o que, en promedio, estarán ocupados el 48 %, es decir, siete de los 14 agentes. Así que los operadores estarán desocupados, en promedio, 52 % del tiempo, en esta ocasión el porcentaje de inactividad es menor que el que resultó para el modelo de nueve canales.

Con los 14 agentes, el número promedio de agentes ocupados $E(B)$ es el mismo 6,67, mientras que el de agentes desocupados $E(I)$ es de 7,33. El número promedio de clientes en el sistema $E(N)$ sería de 6,7 . Además con los 14 agentes, el número promedio de clientes en cola $E(Q)$ es

de casi cero. Asimismo, el tiempo promedio de espera en cola $E(W_Q)$es de 0,16 segundos, y el tiempo en el sistema $E(W)$ es de 120,16, prácticamente la duración del tiempo en servicio.

La probabilidad $C$ de encontrar a todos los agentes ocupados , y tener que esperar en cola $Pr\{W_q \geq 0\}$, es de 1 %, el promedio de tiempo en cola es de 0,16 segundos. Por último, la probabilidad de esperar en cola más tiempo del AWT establecido es de $Pr\{W_q \geq t\}$ es de 0,70 %, cantidad que está por debajo del 1 % establecido.

De acuerdo con esto, será necesario tener 14 posiciones de forma constante para poder lograr el nivel de servicio establecido por la compañía.




| | $P_0$ | $P_n$ | λ | μ | c | ρ | 1−ρ | C Pr[Wq>0] | Agentes Ocupados E (B) | Agentes Libres E(I) | E(Q) Clientes | E(N) Clientes | $E(W_q)$ Tiempo | E(W) Tiempo | t | TSF Pr[Wq>t] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.001272 | 0.004996 | 100.00 | 15.00 | 14.00 | 0.476190 | 0.52381 | 0.0095 | 6.67 | 7.33 | 0.01 | 6.6753 | 0.00009 | 0.07 | 0.003 | 0.0070 |

| 1/μ |
|---|
| 120 |

| Segs | Segs | AWT Segs |
|---|---|---|
| 0.16 | 120.16 | 5 |

| n | $P_n$ | $P_0$ |
|---|---|---|
| 0 | 1 | 0.1176281 |
| 1 | 6.67 | 0.0659281 |
| 2 | 22.22 | 0.0267449 |
| 3 | 49.38 | 0.0115243 |
| 4 | 82.3 | 0.0059144 |
| 5 | 109.74 | 0.0035886 |
| 6 | 121.93 | 0.0024953 |
| 7 | 116.13 | 0.0019347 |
| 8 | 96.77 | 0.0016296 |
| 9 | 71.68 | 0.0014591 |
| 10 | 47.79 | 0.0013640 |
| 11 | 28.96 | 0.0013122 |
| 12 | 16.09 | 0.0012851 |
| 13 | 8.25 | **0.0012716** |
| 14 | 3.93 | 0.0012653 |
| 15 | 1.75 | 0.0012625 |
| 16 | 0.73 | 0.0012613 |
| 17 | 0.29 | 0.0012609 |
| 18 | 0.11 | 0.0012607 |
| 19 | 0.04 | 0.0012606 |
| 20 | 0.01 | 0.0012606 |
| 21 | 0 | 0.0012606 |
| 22 | 0 | 0.0012606 |
| 23 | 0 | 0.0012606 |
| 24 | 0 | 0.0012606 |

La gráfica nos muestra lo que ya hemos mencionado, con la distribución centrada entre el 6 y

el 8 y casi la totalidad de la distribución de probabilidad de llegada de clientes cubierta con 14 agentes.

### 4.7.2. Simulación del sistema $M/M/14$

Una vez más, hacemos una simulación de 800 llegadas al sistema con los parámetros que definimos en la sección 4.7.1. Obteniendo la tabla de resultados correspondientes a la figura 10, en donde observamos que, en la simulación, el tiempo promedio en el sistema fue de 118 segundos, cuando en nuestro modelo esperábamos 120 segundos. Además, en la simulación, el tiempo promedio de espera en cola $E(W_Q)$fue de 0.14 segundos, habíamos pronosticado 0.16 segundos con nuestro modelo. El tiempo promedio en servicio $E(W)$ en la simulación fue de 118 segundos, lo cual concuerda con los 120 segundos que planteamos en el modelo de colas. Por otra parte, el número promedio de clientes en servicio $E(N) - E(Q)$ es de 6,45 , valor muy próximo al 6,67 que esperábamos en el modelo. Lo mismo ocurre con de clientes en el sistema $E(N)$obtenido en la simulación, 6,47, contra 6,68 que obtuvimos como resultado, mediante la teoría de colas. Y finalmente, en la simulación hubo 0.008 clientes en cola, cuando esperábamos 0.01. Los anteriores resultados de la simulación concuerdan, una vez más, significativamente con lo que habíamos previsto con nuestro modelo.

Ahora, observando la gráfica de la simulación (figura 11), existen solamente dos colas de dimensión dos, que, casi seguramente, no sobrepasarán el tiempo de espera de cinco segundos, lo cual mantendrá el objetivo de espera por encima de los cinco segundos dentro del objetivo establecido. Más adelante veremos, una vez más, que tan cercanos son los resultados de nuestro modelo con el tráfico real durante la jornada de trabajo en el centro telefónico.

## 4.8. Consideraciones Finales

Hemos llegado a la conclusión de que con 14 agentes atendiendo de manera constante el tráfico de llamadas serán suficientes para lograr los objetivos establecidos en el segundo periodo de análisis .

## 4.9. Contraste con el tráfico Real para el caso M/M/14

En esta ocasión analizaremos el tráfico de llamadas en la segunda parte de la jornada, partiendo de las 19:30 horas y hasta el final de la misma. Como podemos obsevar, de nuevo en la gráfica de la figura 8. Observamos que hay una mayor variabilidad en la cantidad de llamadas durante este periodo. Usando nuevamente la tabla de la figura 9, observamos lo siguiente:

La primera columna *Llamadas recibidas* en promedio recibimos 96 llamadas, por cada periodo de

| Queueing Simulaton | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Next Event-Dynamic | | | | | | Event | Event Time | Time Interv | Event Index | Event Name | Next Server | Server 1 | Server 2 | Server 3 | Server 4 | Server 5 | Server 6 | Server 7 | Server 8 | Server 9 |
| Queue Station | SimQ2 | | Number | TBA | TFS | 0 | 0 | 0.0022 | Start | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Arrival Rate | 100 | | 1 | 0.0022 | 0.0242 | 1 | 0.0022 | 0.0242 | 15 | A | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Service Rate | 15 | | 2 | 0.0284 | 0.0367 | 2 | 0.0264 | 0.0042 | 1 | S | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Number of Servers | 14 | | 3 | 0.0086 | 0.0972 | 3 | 0.0307 | 0.0086 | 15 | A | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Max. Number in System | *** | | 4 | 0.0049 | 0.1703 | 4 | 0.0393 | 0.0049 | 15 | A | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Type | M/M/14 | | 5 | 0.0049 | 0.0775 | 5 | 0.0442 | 0.0049 | 15 | A | 3 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Arrival Seed | *** | | 6 | 0.0062 | 0.0391 | 6 | 0.0491 | 0.0062 | 15 | A | 4 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| Service Seed | *** | | 7 | 0.0008 | 0.0394 | 7 | 0.0552 | 0.0008 | 15 | A | 5 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| Number in Simulation | 800 | | 8 | 0.0056 | 0.0158 | 8 | 0.056 | 0.0056 | 15 | A | 6 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| Start Data Time | 0 | | 9 | 0.0007 | 0.0278 | 9 | 0.0616 | 0.0007 | 15 | A | 7 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| Stop Data Time | 8.0822838 | | 10 | 0.0063 | 0.0402 | 10 | 0.0623 | 0.0051 | 15 | A | 8 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| Mean Number at Station | 6.46568744 | | 11 | 0.0022 | 0.0474 | 11 | 0.0674 | 0.0013 | 1 | S | 9 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| Mean Time at Station | 0.06540366 | 117.73 | 12 | 0.0032 | 0.0728 | 12 | 0.0686 | 0.0022 | 15 | A | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| Mean Number in Queue | 0.00773025 | | 13 | 0.0178 | 0.0535 | 13 | 0.0708 | 0.0032 | 15 | A | 9 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Mean Time in Queue | 7.8195E-05 | 0.14 | 14 | 0.0042 | 0.2269 | 14 | 0.0741 | 0.0033 | 15 | A | 10 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Mean Number in Service | 6.45795719 | | 15 | 0.0176 | 0.0114 | 15 | 0.0774 | 0.0127 | 7 | S | 11 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| Mean Time in Service | 0.06532546 | 117.59 | 16 | 0.0115 | 0.0081 | 16 | 0.0901 | 0.0019 | 8 | S | 7 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 |
| Arrival Rate | 98.8581965 | | 17 | 0.0259 | 0.03 | 17 | 0.0919 | 0.0024 | 15 | A | 7 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| Throughput Rate | 98.8581965 | | 18 | 0.0015 | 0.1182 | 18 | 0.0943 | 0.0011 | 5 | S | 8 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 |
| Efficiency | 0.46128266 | | 19 | 0.0024 | 0.0044 | 19 | 0.0954 | 0.0008 | 6 | S | 5 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 |
| Probability Balk | 0 | | 20 | 0.0007 | 0.0717 | 20 | 0.0962 | 0.0126 | 15 | A | 5 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 |
| | | | 21 | 0.0033 | 0.0069 | 21 | 0.1088 | 0.005 | 1 | S | 6 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 |
| | | | 22 | 0.0165 | 0.0076 | 22 | 0.1138 | 0.0045 | 15 | A | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 |
| | | | 23 | 0.0145 | 0.1673 | 23 | 0.1182 | 0.007 | 9 | S | 6 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 |

Figura 10: Simulación 14 Agentes

tiempo, número que está por muy próximo, aunque un poco por debajo de los 100 que habíamos establecido. Esto ayudará mucho a que el modelo ajuste bien a lo observado en la jornada real. Ahora, el otro valor importante, que queremos saber es el que está en la columna *AHT IN [segundos]* y corresponde al tiempo de servicio es de 115 segundos, valor que también está muy cerca a los 120 establecidos previamente. Dado que los dos valores se encuentran muy cerca de los que establecimos en el modelo, esperamos que no haya grandes diferencias con los resultados reales.

Ahora, en esta parte del día, el nivel de ocupación promedio fue de 45 %, valor muy próximo al 48 % obtenido con el modelo lo cual indica que el modelo funciona bien, cuando los parámetros introducidos son cercanos a los reales.

Por último, revisamos el número de llamadas encoladas que ha sido de 2 llamadas, de donde:

$$2/480 = ,42\,\%$$

que es una proporción que se encuentra por debajo del 1 del valor $C$, que habíamos pronosticado alrededor del 1 % y, en consecuencia, el porcentae de llamadas por debao del Tiempo Aceptable de Respuesta AWT, se encuentra también por debajo de 1 %.

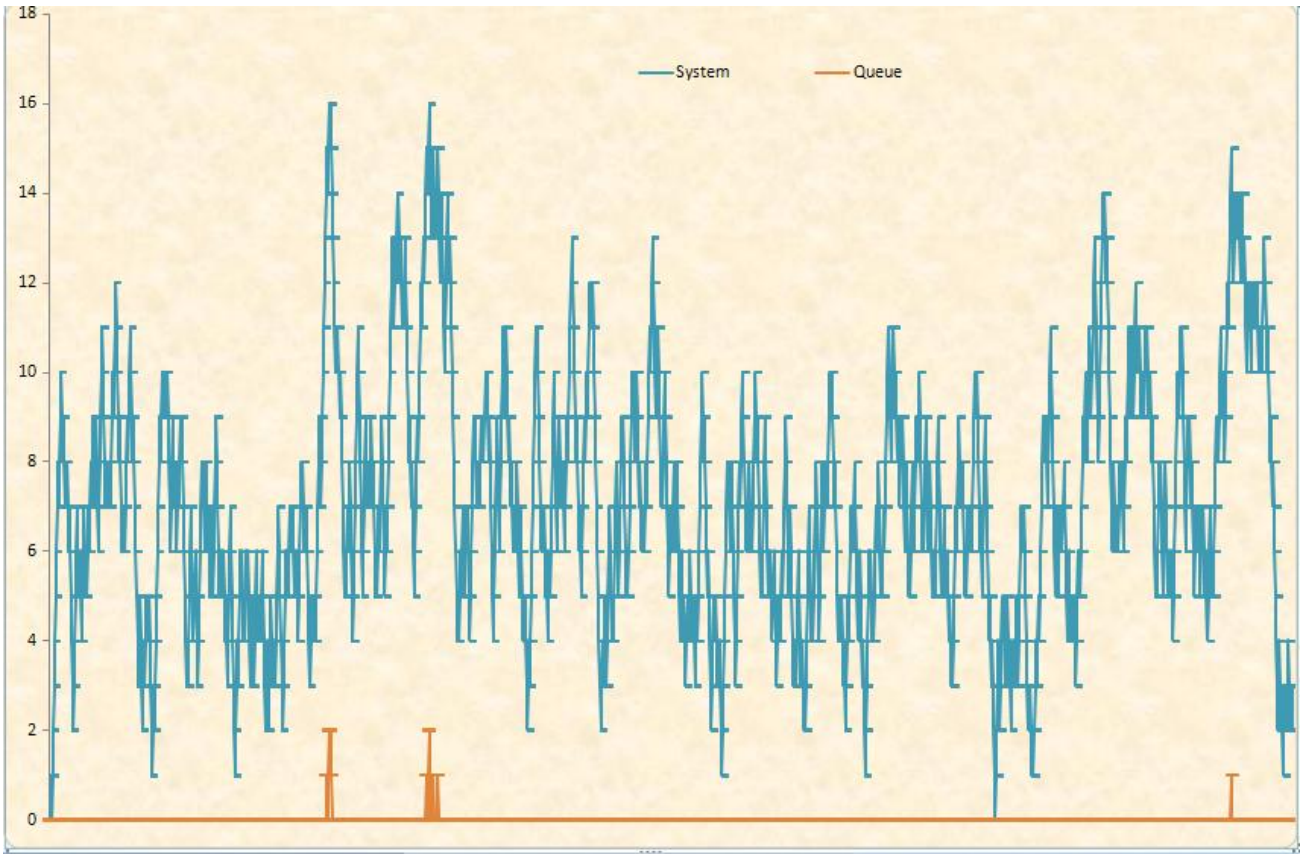



Figura 11: Simulación 800 llegadas 14 Canales

# 5. Conclusiones

De este trabajo se desprenden los siguientes puntos importantes:

Los modelos estocásticos de colas pueden ser una herramienta poderosa y muy útil para modelar líneas de espera en centros de atención telefónica.

Sin embargo, es importante que la estimación de los parámetros fundamentales en estos modelos, que son el tiempo de servicio y la tasa de entrada al sistema, sean estimados de manera precisa, ya que de estar estos valores lejos de la realidad, las predicciones que hagamos con nuestro modelo estarán lejos del comportamiento real de las líneas de espera. Pequeños cambios no muy significativos en el tiempo de servicio o de la tasa de llegada pueden producir un gran cambio en el número de servidores necesarios. Usualmente resulta más fácil obtener el tiempo aproximado de servicio mediante experimentación de algunas llamadas ficticias midiendo el tiempo de atención. Resulta más difícil conocer el número de llamadas, en este caso, es necesario tener experiencia previa de algún tipo, por ejemplo de la cantidad de llamadas en campañas similares hechas anteriormente.

Utilizando estas herramientas, es posible imaginar distintos escenarios con posibles flujos de llegada o tiempos de servicio y saber cómo se estarían comportando las líneas de espera, de acuerdo con los valores introducidos. Esto resulta de gran utilidad para las organizaciones que pueden evitar desastres en los que el flujo de clientes se desborde, o grandes gastos teniendo muchos más servidores de los necesarios.

Al utilizar adecuadamente los modelos de colas, es posible lograr una eficiencia máxima con el número mínimo de servidores, alcanzando los niveles de servicio establecidos. Esto beneficia tanto a la institución, que no gasta recursos innecesariamente, a los trabajadores que no están demasiado ocupados u ociosos y a los clientes, que reciben un buen nivel de servicio.

En el caso aplicado, que hemos realizado en este trabajo, el lograr tener la cantidad mínima de agentes para cubrir la demanda de servicio, al nivel establecido ha ayudado a reducir considerablemente los costos de operación, lo cual genera mayor utilidad para la compañía, así como disminuir el tiempo inactivo u ocioso de los operadores, que se desesperan cuando están demasiado tiempo sin recibir llamada.

# A. Formulario

$L$: Número promedio de clientes en el sistema.
$L_q$: Número promedio de clientes en cola.
$W$: Tiempo de espera promedio en el sistema.
$W_q$: Tiempo promedio en cola

$$L = \sum_{n=0}^{\infty} n p_n \quad L_q = \sum_{n=0}^{\infty} (n-s) p_n$$

$$W = \frac{L}{\bar{\lambda}} \qquad W_q = \frac{L_q}{\bar{\lambda}}$$

Modelo $M/M/c$

Distribución de Estado Estable

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

$$p_n = \begin{cases} \dfrac{\lambda}{n\mu} p_{n-1} = \dfrac{c}{n} \rho p_{n-1} (\text{F. Rec}) & n = 1, \ldots, c \\ \dfrac{\lambda}{c\mu} p_{n-1} = \rho^{n-c} p_c = \rho p_{n-1} (\text{F. Rec}) & n = c, c+1, \ldots \end{cases}$$

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

Probabilidad de Espera $Pr\{N \geq c\}$

$$C = C\left(c, \frac{\lambda}{\mu}\right) = \frac{p_c}{1-\rho}$$

Número de servidores en espera u Ocupados

$$E(B) = c\rho \quad E(I) = c(1-\rho)$$

**Número Esperado de Clientes en el sistema**

$$E(N) = E(B) + E(Q); \qquad E(N) = c\rho + \frac{\rho C}{1-\rho}$$

$$E(W_Q) = \frac{1}{c\mu(1-\rho)} Pr\{N \geq c\}; \quad E(Q) = \frac{\rho}{(1-\rho)} Pr\{N \geq c\}$$

$$E(W) = \frac{E(N)}{\lambda} = \frac{1}{\mu} + \frac{p_c}{c\mu(1-\rho)^2}$$

Distribución de tiempo de espera

$$Pr\{W_q > t \mid W_q > 0\} = e^{(1-\rho)c\mu t}$$

$$Pr\{W_q > t\} = C\left(c, \frac{\lambda}{\mu}\right) e^{(1-\rho)c\mu t}$$

**Tiempo de Espera en el Sistema**

$$E(W_s) = \frac{1}{\mu} + \frac{C}{c\mu(1-\rho)}$$

$$E(W_q) = \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!c\mu} \frac{1}{(1-\rho)^2} p_0 = \frac{p_c}{c\mu} \frac{1}{(1-\rho)^2}$$

**El Proceso de Salida**

**Teorema (Burkle)** En un sistema de colas M/M/c en estado estable, con tasas de llegada y servicio $\lambda$ y $\mu$, respectivamente, los tiempos entre salidas son independientes e identicamente distribuidos con distribución exponencial con media $1/\lambda$, es decir, el proceso de salida es Poisson con parámetro $\lambda$.

El Modelo $M/M/c/c$

$$\lambda_n = \lambda, \quad \mu_n = n\mu, \quad n = 0, 1, 2, \ldots, c-1$$
$$\lambda_n = 0, \quad \mu_n = c\mu, \quad n \geq c$$

$$p_n = \frac{\left(\frac{\lambda}{\mu}\right)^n / n!}{\sum_{k=0}^{c} \left(\frac{\lambda}{\mu}\right)^k / k!} \qquad n = 0, 1, 2, \ldots, c.$$

Formula de Pérdida

$$B\left(c, \frac{\lambda}{\mu}\right) = p_c = \frac{\left(\frac{\lambda}{\mu}\right)^c \Big/ c!}{\sum_{k=0}^{c} \left(\frac{\lambda}{\mu}\right)^k \Big/ k!}$$

Número Esperado de Canales Ocupados y Canales Libres

$$E\{B\} = \frac{\lambda}{\mu}[1 - p_c] = \left(\frac{\lambda}{\mu}\right)\left[1 - B\left(c, \frac{\lambda}{\mu}\right)\right]$$
$$E\{I\} = c - \frac{\lambda}{\mu}\left[1 - B\left(c, \frac{\lambda}{\mu}\right)\right]$$

Probabilidad de Espera en Cola Sea $X_i$ la variable indicadora para el el i-ésimo canal elegido aleatoriamente; $X_i = 1$ o 0 según si el i-ésimo canal esté ocupado o libre. Sea $P_c\{A\}$ la probabilidad de ocurrencia de un evento A en uns sistema $M/M/c/c$ en equilibrio. Entonces

i

$$P_c\{X_1 = 1, \ldots, X_k = 1\} = \frac{B\left(c, \frac{\lambda}{\mu}\right)}{B\left(c - k, \frac{\lambda}{\mu}\right)}, \qquad 1 \leq k \leq c$$

ii

$$P_c\{X_1 = 1\} = \frac{\left(\frac{\lambda}{\mu}\right)\left[1 - B\left(c, \frac{\lambda}{\mu}\right)\right]}{c}, \qquad \text{y}$$

iii

$$P_c\{X_{k+1} = 1 \mid X_1 = 1, X_2 = 1, \ldots, X_k = 1\} = P_{c-k}\{X_1 = 1\}$$

# B. Glosario

- *AHT* Average Handling Time. Es el tiempo en segundos que el agente estuvo en conversación (Tiempo en ACD/NACD) más el tiempo que tardó de captura (After Call Work) más el tiempo que mantuvo en Hold la llamada.
- *ACD* Automatic Call Distribution. Llamadas que entran del conmutador a un grupo determinado.
- *AHT* Es el tiempo en segundos que el agente estuvo en conversación (Tiempo en ACD/NACD) mas el tiempo que tardó de captura (After Call Work) mas el tiempo que mantuvo en Hold la llamada.
- *ASA* Average Speed of Answer. Tiempo promedio de espera, antes de ser atendido por un agente.
- *AWT* Acceptable Waiting Time. Es el tiempo que se establece como máximo para que los clientes que llaman esperen en cola, antes de ser atendidos, usualmente es de 20 segundos.
- *Calls* Llamadas totales. Total de llamadas recibidas y no realizadas.
- *Centro Telefónico*. Es una colección de recursos (Típicamente agentes y equipo CTI), capaz de brindar servicios telefónicos.
- *CTI* Computer-Telephony Integration. Integración de procesos que permite la comunicación e integración de equipo telefónico y sistemas de cómputo.
- *DID* Direct Inward Dialing o Direct Incoming Dialing. Es la facilidad de comunicarse del exterior a una extensión de un conmutador digital, sin la intervención de una operadora, dando el efecto de una línea directa.
- *FCFS* First Come First Served. Se refiere a la forma en que las llamadas encoladas son atendidas, en orden de llegada, es la forma más usual en que se da prioridad a las llamadas entrantes al conmutador de un call center, aunque no siempre es así como son atendidas las llamadas.
- *Grupo ID* Identificador del grupo de ACD. Es el grupo dado de alta en el conmutador para su uso en la realización y recepción de llamadas.
- *HACD* Network Automatic Call Distribuitor. Es el tiempo, en segundos, que el agente estuvo en conversación con el cliente.
- *OUT* Fuera, no disponible. Es el tiempo en segundos que el agente se mantiene fuera y no está en espera de llamada o disponible para realizarla.
- *WRAP (ACW)* After Call Work. Es el tiempo en segundos que el agente emplea para corrección y captura de datos y/ o retroalimentación de la llamada con el supervisor.

- *RDY IDLE* Disponible. Es el tiempo que el agente se mantiene en espera de llamada o disponible para realizarla.
- *SL* Service Level. Nivel de Servicio. Término que se refiere, en ocasiones, a todos los aspectos del servicio (Tiempo de espera, abandonos, etc.) o en ocasiones solamente al TSF.
- *Tiempo de Espera* Tiempo que transcurre entre el momento en que el cliente se incorpora a la cola y el momento en que el agente es conectado a la llamada.
- *TSF* Telephone Service Factor. Factor de Servicio Telefónico. A menudo llamado SL, es el porcentaje de llamadas contestadas antes del AWT.
- *Wrap Up Time* Tiempo después de terminada la llamada, que el agente utiliza en esa misma llamada. Usualmente consiste en el tiempo en que el agente anota alguna información relativa a la llamada que ha terminado completando alguna información en el ordenador.

# Referencias


[1] Koole Ger, 2006. Call Center Mathematics, A cientific Method forunderstanding and improving call centers.

[2] Donald E. Knuth, 1989. Typesetting Concrete Mathematis.

[3] Herrera Mahil, Introducción a los procesos estocásticos.

[4] Medhi, Jyotiprasad, Stochastic Models in Queueing Theory.Academic Press, 1991.

[5] Taha, A. Hambdy, 2004. Investigación de Operaciones.

[6] Little, J.D.C. A proof for the queueing Formula: $L = lambdaW$.Operations Research 1961.

[7] Winston, Wayne L. Investigacion de operaciones : aplicaciones y algoritmos. Editorial Iberoamericana, 1994.

[8] Cooper, Robert B. Queueing Theory. North Holland, 1981.

[9] Góngora Amaro, Marco Antonio. Simulación de un centro de atención telefónica a través de la teoría de colas. Tesis. U.N.A.M.

[10] Caballero Ruiz, osé Victor, Teoría de colas y análisis de su aplicación a un problema real. Tesis U.N.A.M., 1999.

[11] Aguilar Urbano, Fernando. Elaboración de una hoja de cálculo para el análisis y planeación de sistemas de filas utilizando modelos de colas. Tesis U.N.A.M., 2005.

[12] Arriaga Sánchez, Martha. Desarrollo de un sistema para la simulación de líneas de espera en un centro de atención telefónica. Tesis U.N.A.M., 2005.

[13] José Pedro García Sabater. Resumen traducido de parte del libro "Fundamentals of Queueing Theory"por Donald Gross y Carl Harris.