UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE INGENIERÍA

INTRODUCCIÓN A LA TEORÍA DE COLAS

DIVISIÓN DE INGENIERÍA MECÁNICA E INDUSTRIAL

Idalia Flores De La Mota

CONTENIDO

1

2

3

4

Idalia Flores De La Mota

INTRODUCCIÓN A LA TEORÍA DE COLAS

DIMEI
División de Ingeniería Mecánica e Industrial

Flores De La Mota, Idalia
Introducción a la Teoría de Colas
Universidad Nacional Autónoma de México,
Facultad de Ingeniería, 2023, 152 p.

Introducción a la Teoría de Colas

Primera edición electrónica
de un ejemplar (3 MB) Formato PDF
Publicado en línea en agosto de 2023

FACULTAD DE INGENIERÍA
http://www.ingenieria.unam.mx/

Hecho en México.

Unidad de Apoyo Editorial

Cuidado de la edición: María Cuairán Ruidíaz
Diseño y formación editorial: Nismet Díaz Ferro
Fotografía de portada: rawpixel.com, Freepik

CONTENIDO

3 MODELOS DE COLAS GENERALES Y REDES DE COLAS

4 TOMA DE DECISIONES Y SIMULACIÓN

PRÓLOGO

La teoría de colas o teoría de la espera es una estructura matemática de suma importancia en nuestro medio; ya que no sólo trata de resolver problemas de Investigación de Operaciones, también es una herramienta importante en algunas las áreas de la ingeniería. Muchas veces los textos que tratan sobre esta disciplina son muy teóricos y se limitan a resolver unos cuantos ejemplos abundando en la demostración de teoremas, lo cual no es nuestro propósito. Más bien, se tiene como objetivo mostrar los modelos de colas con ejemplos sencillos y la teoría que sustenta dichos modelos.

El objetivo central de estos apuntes consiste en reconocer todas las fuentes de aleatoriedad, modelarlas e inferir conclusiones sobre la eficiencia del sistema y propuestas de mejora.

Estos apuntes se desarrollaron como un apoyo didáctico para la clase y como un complemento de la bibliografía sugerida para el curso, en este sentido es importante mencionar que no pretenden sustituir a los libros de teoría de colas, son simplemente un complemento para la materia. No se presenta ningún complemento computacional ya que este va cambiando y desarrollándose muy rápidamente.

Los presentes apuntes van dirigidos a estudiantes de licenciatura y posgrado que requieren de un cierto grado de conceptualización y una buena dosis de ejemplos y aplicaciones. También al final de

cada capítulo se presentan notas históricas que buscan dar al lector una idea más completa del tema, viéndolas desde una perspectiva histórica, inmersas en la realidad de la época.

Parafraseando a Isaac Asimov: "La ciencia gana realidad cuando es visualizada no como una abstracción, sino como la suma concreta de los científicos, pasados y presentes, vivos y muertos. No hay estamentos en ciencia, ni una observación, ni un pensamiento que exista por sí mismo. Cada uno de ellos es producto de un duro esfuerzo de algún hombre, y a menos que conozcáis al hombre y el mundo con el cual trabajaba, los supuestos que él aceptaba como verdades, los conceptos que él consideraba insostenibles, no podréis comprender sus afirmaciones, u observaciones, o pensamiento".

Quiero agradecer el apoyo en revisión y correcciones a la Maestra María Cuairán Ruidíaz, de la Unidad de Apoyo Editorial de la Facultad de Ingeniería de la UNAM, por el respaldo brindado y por la colaboración para la realización de la edición; y a la LDG Nismet Díaz Ferro por la dedicación y disposición para el diseño y formación de la obra. Asimismo, a la Lic. Patricia García Naranjo titular de la UAE por su apoyo en la publicación de este texto.

Agradezco a Alejandro Felipe Zárate Pérez y José Antonio Sánchez Calderón, por su apoyo en la elaboración de este material.

1 CONCEPTOS DE TEORÍA DE COLAS

1.1 INTRODUCCIÓN

En la actualidad y sobre todo en los servicios, existe una serie de situaciones en donde "alguien" o "algo" requiere de un determinado tipo de servicio, para lo cual, en ocasiones tienen que hacer una cola o esperar en línea ante "alguien" o "algo", que es quien proporciona dicho servicio. En este sentido la teoría de colas ofrece elementos importantes de decisión en la solución a problemas como: disminuir o eliminar la congestión de un determinado servicio, minimizar pérdidas de tiempo debidas a operaciones deficientes, minimizar excesos de capacidad, etc.

La naturaleza de esta situación de espera puede analizarse matemáticamente, si se conocen las leyes que gobiernan las llegadas para un servicio, el orden en que son atendidos, y los tiempos para dar el servicio.

Existe una fuente, de la cual provienen las unidades que requieren un determinado servicio, mismas que llegan al sistema de espera formando una cola. El mecanismo de servicio, a través de estaciones de servicio selecciona en el tiempo a una de las unidades en la cola para prestar el servicio solicitado. Una vez terminado éste, la unidad deja el sistema de espera, como se observa en la figura 1.1

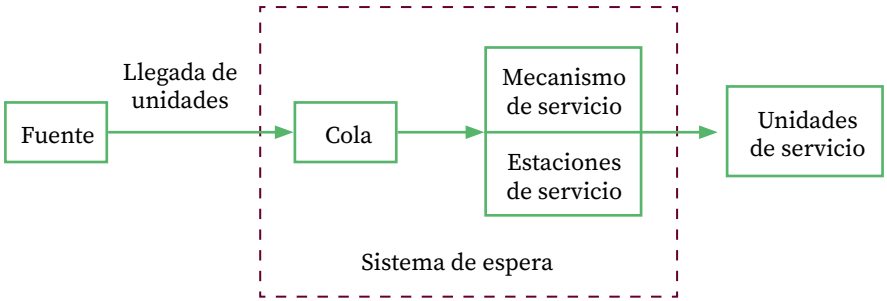


Figura 1.1 Proceso básico de una cola o línea de espera

La gran diversidad de situaciones en cada una de las fases del proceso descritas origina una multiplicidad de situaciones de espera, como las que se presentan en la tabla 1.1.

Tabla 1.1 Procesos de servicio

Tipos de entradas	Naturaleza del servicio	Prestadores de servicio
Clientes Barcos Maquinaria Pacientes	Venta de artículos Carga y descarga Reparación Consulta	Dependientes Muelles Mecánicos Médicos

De igual forma, se pueden señalar como ejemplos de situaciones de espera los siguientes:

Tabla 1.2 Situaciones de espera

Servicio	Líneas de espera	Estación de servicio
Consultorio Peluquería Gasolinera Banco	1 1 2 muchas	1 muchas 1 muchas

Las características más frecuentes del proceso de espera se mencionan a continuación:

1.2 CARACTERÍSTICAS DEL PROCESO DE ESPERA

1 Tipo de llegadas

Una característica de la fuente es el tamaño del número total de unidades que solicitan servicio puede ser finito o infinito. Otra característica es la forma en que las unidades llegan al sistema de espera, se pueden distinguir los siguientes casos:

Llegadas de unidades al sistema con intervalos iguales de tiempo.

Llegadas de unidades al sistema con intervalos desiguales de tiempo, pero perfectamente conocidos.

Llegadas de unidades al sistema con intervalos desiguales de tiempo, cuyas probabilidades son conocidas (intervalos aleatorios).

Llegadas de unidades al sistema con intervalos desiguales de tiempo, con probabilidades desconocidas y en cuyo caso no puede ser estimado.

Los casos más típicos de distribución de llegadas con su correspondiente distribución de tiempo entre llegadas, que se presentan en la tabla 1.3.

Tabla 1.3 Distribución de llegadas y tiempo entre llegadas

Distribución de llegadas	Distribución de tiempo entre llegadas
Llegadas que ocurren a intervalos iguales de tiempo	Constante
Llegadas poissonianas	Exponencial

Tabla 1.3 Distribución de llegadas y tiempo entre llegadas (continuación)

Tipo poissoniano es intermedia entre las dos anteriores	Erlang (dist. Gama) donde: sí $k = 1$ exponencial sí $k = \alpha$ constante
Caso general	Distribución general

E Ejemplo 1.1

Llegadas	3:20	3:25	3:48	3:60	**Poisson**
Tiempo entre llegadas	5	23	12		**Exponencial**

En la práctica existen tipos de llegadas más complejos: llegadas programadas a ciertos instantes de tiempo, pero sujetas a variación; llegadas en grupos, colas cíclicas en las cuales un pequeño número de unidades recicla, etc.

2 Disciplina de la línea de espera

Existen diferentes disciplinas en la línea de espera, siendo el caso más sencillo y el que normalmente se considera en modelos de espera: primero en llegar es a quien se le da primero servicio (FIFO). Otros tipos de disciplina se establecen de acuerdo con un cierto orden preferencial para otorgar servicio.

Las diferentes disciplinas de servicio son:

- *FIFO* (primero en llegar, primero en ser atendido), la más común.
- *LIFO* (último en llegar, primero en ser atendido).
- *RS* (servicio en orden aleatorio).

También es posible que los clientes que lleguen a una instalación sean colocados en líneas de *espera con prioridad,* para que aquellas personas con mayor prioridad reciban preferencia para ser atendidos en primer término.

3 Mecanismo de servicio

El mecanismo de servicio puede constar de una o más unidades de servicio y cada una contiene una o más estaciones de servicio.

Igual que las llegadas de unidades al sistema, el tiempo de servicio (comprendido desde que la unidad entra al servicio hasta que termina) sigue la misma clasificación que en el punto 1.

Tabla 1.4 Distribución de salidas y del tiempo de servicio

Distribución de salidas	Distribución del tiempo de servicio (intervalos de tiempo entre servicios consecutivos)
Salidas que ocurren a intervalos iguales de tiempo	Constante
Salidas poissonianas con parámetro	Exponencial $\mu\ exp(-\mu t)$
Tipo poissoniano	Erlang $f_k(t) = t_{k-1}(\mu k)\, k\ exp(-\mu kt)/(k-1)!$ *sí* $k=1$ *exponencial* *sí* $k=\alpha$ *constante*
Caso general	Distribución general

Otros factores que pueden influir en el servicio son:

Servicio por grupo, colas en serie en las cuales la salida de una unidad de una estación de servicio es la entrada en la siguiente estación, etc.

Los elementos básicos de un modelo de espera dependen de los siguientes factores:

1. Distribución de llegadas (llegadas individuales o en grupo).
2. Distribución del tiempo de servicio (servicio individual o masivo).
3. Diseño de la instalación de servicio (estaciones en serie, en paralelo o en red).
4. Disciplina de servicio (FIFO, LIFO, otras).
5. Tamaño de la cola (finito o infinito).
6. Fuente de llegadas (finita o infinta).
7. Conducta humana (cambios, elusión y renuncia).

En general, las soluciones analíticas son difíciles de obtener para problemas muy complejos, debiendo usar la simulación.

1.3 Notación de Kendall

Un código que describe el proceso de llegada. Los códigos usados son:

M para "Markoviano"(la tasa de llegadas sigue una distribución de Poisson), lo que significa una distribución exponencial para los tiempos entre llegadas.

D para unos tiempos entre llegadas deterministas, es decir, no siguen un proceso probabilista a la hora de su determinación.

G para una "distribución general" de los tiempos entre llegadas, o del régimen de llegadas. Mismo código para tiempos de servicio.

Una forma de describir un proceso de espera en forma sencilla es a través de una serie de símbolos y diagonales tales como $A/B/X/Y$,

$Z/$, notación de Kendall (1953), en donde A indica la distribución de tiempo entre llegadas, B la distribución de probabilidades de tiempo entre servicios, X el número de estaciones de servicio paralelas, Y la restricción en la capacidad del sistema y Z la disciplina de espera, como se muestra en la tabla 1.5.

Tabla 1.5 Notación de Kendall para diferentes distribuciones

Característica	Símbolo	Explicación
Distribución de tiempo entre llegadas (A)	M	Exponencial o Poisson
	D	Determinística
	E_k	Tipo Erlang b ($b = 1, 2, \cdots$)
	GI	General (independiente)
Distribución de tiempo de servicio (B)	M	Exponencial
	D	Determinística
	E_k	Erlang tipo b
	G	General
Número de estaciones de servicio paralelas (X)	$1, 2, \cdots, X$	
Restricción en la capacidad del sistema (Y)	$1, 2, \cdots, Y$	
Disciplina de espera (Z)	FIFO	Primero en entrar, primero en salir
	LIFO	Último en entrar, primero en salir
	SIRO	Servicio en forma aleatoria
	PRI	Prioridad
	GD	Disciplina general

Normalmente se utilizan los primeros tres símbolos, omitiéndose los símbolos Y y Z. Así $M/D/2$ representa un sistema de espera con tiempo entre llegadas exponencial, servicio determinístico, dos estaciones de servicio, ningún límite en la capacidad de servicio y como disciplina primero en llegar primero en ser atendido.

/M/M/2 Llegadas poissonianas servicio poissoniano 2 estaciones de servicio.

/M/G/8 Llegadas poissonianas servicio con distribución general de tiempo entre servicios 8 estaciones de servicio.

En la descripción de los modelos de espera se usa la siguiente nomenclatura:

E_n = Estado en el cual hay n unidades en el sistema.

λ = Relación media de llegadas = número esperado de llegadas por unidad de tiempo.

μ = Relación media de servicios = número esperado de unidades servidas por unidad de tiempo (de una estación de servicio).

$1/\lambda$ = Tiempo esperado entre llegadas.

$1/\mu$ = Tiempo esperado entre salidas tiempo medio esperado de servicio.

s = Número de estaciones de servicio.

λn = Relación media de llegadas de nuevas unidades cuando hay n unidades en el sistema.

n = Relación media de servicios cuando hay n unidades en el sistema.

$pj(t)$ = Probabilidad de que existan j unidades en el sistema en el instante t.

$t = \lambda / s\mu$ Porcentaje de utilización del servicio = intensidad de tráfico = probabilidad de que el servicio esté ocupado.

1.4 Régimen permanente o estado estacionario

Por régimen permanente o estado estacionario, se entiende la zona de respuesta del sistema en la que, tras haber transcurrido tiempo

suficiente, todos los parámetros del sistema se han estabilizado y permanecen a un valor constante. En este sentido se definen las siguientes:

L = Número esperado de unidades en el sistema
L_q = Número esperado de unidades en la cola
L_s = Número esperado de unidades siendo servidas o en servicio
W_q = Tiempo esperado de permanencia en la cola
W_s = Tiempo esperado de permanencia en el servicio $= 1/\mu$
W = Tiempo esperado de permanencia en el sistema
W = $W_q + W_s$

Modelo generalizado de cola de Poisson.
El desarrollo del modelo generalizado se basa en el comportamiento a largo plazo, o de estado estable de la cola que se alcanza después de que el sistema ha estado funcionando durante un tiempo suficientemente largo.

Este modelo supone que las frecuencias tanto de las llegadas como de salidas dependen del estado, lo que significa que dependen de la cantidad de clientes en la instalación de servicio.

Características:
- Combinación de llegadas y salidas/Poisson
- Comportamiento a largo plazo o estado estable.
- Las frecuencias tanto de llegada como de salida dependen del estado.

Ejemplos de modelos de Poisson.
1. En la caseta de cobro de una autopista, los empleados tienden a acelerar el cobro en las horas pico.

2. En la caseta de cobro de una autopista, los empleados tienden a acelerar el cobro en las horas pico.

Ley de Little[1]

Un importante resultado matemático es el demostrado por John Little en 1961, el cual relaciona las siguientes variables:

L : Número promedio de clientes en un sistema
W : Tiempo promedio de espera en un sistema
λ : Número promedio de clientes que llegan al sistema por unidad de tiempo

Luego la ley de Little establece que el número promedio de clientes en un sistema (L) es igual a la tasa promedio de llegada de los clientes al sistema (λ) por el tiempo promedio que un cliente está en el sistema (W).

$$L = \lambda W$$

La fórmula es válida para sistemas y para subsistemas, es decir:

$$L_q = \lambda\ W_q$$

Donde L_q es el número promedio de clientes que esperan en la fila y W_q el tiempo promedio que un cliente espera en la fila. Adicionalmente μ representa el nivel del servicio o capacidad del sistema.

1 https://www.gestiondeoperaciones.net/lineas-de-espera/que-es-la-ley-de-little-y-su-aplicacion-en-el-analisis-de-lineas-de-espera/

Ejemplo 1.2

Un pequeño banco está considerando abrir un servicio para que los clientes paguen desde su automóvil. Se estima que los clientes llegarán a una tasa promedio de 15 por hora. El cajero que trabajará en la ventanilla puede atender a los clientes a un ritmo promedio de uno cada tres minutos. Suponiendo que el patrón de llegadas es Poisson y el patrón de servicios es Exponencial, encuentre:

La utilización promedio del cajero:

$$\rho = \frac{\lambda}{\mu} = \frac{15}{20} = 75\,\%$$

El número promedio de clientes en la línea de espera es:

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{15^2}{20(20-15)} = 2.25 \text{ clientes}$$

El número promedio de clientes en el sistema:

$$L_s = \frac{\lambda}{\mu - \lambda} = \frac{15}{20-15} = 3 \text{ clientes}$$

El tiempo promedio de la espera en la fila:

$$W_q = \frac{L_q}{\lambda} = \frac{2.25}{15} = 0.15 \left[\text{horas}\right] \text{ o } 9 \left[\text{min}\right]$$

El tiempo promedio de espera en el sistema:

$$W_s = \frac{L_s}{\lambda} = \frac{3}{15} = 0.2 \left[\text{horas}\right] \text{ o } 12 \left[\text{min}\right]$$

En el capítulo 2 veremos con más detalle estas fórmulas y su uso en diferentes sistemas de colas.

1.5 Distribuciones de probabilidad

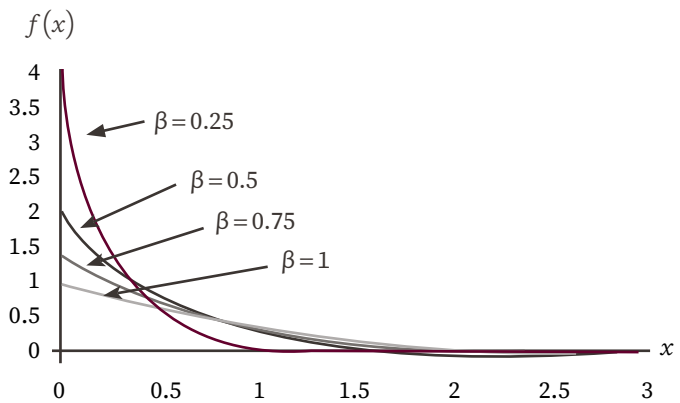
Como se ha visto en las secciones anteriores existen distribuciones de probabilidad asociadas a las llegadas al sistema y los tiempos de servicio. Las distribuciones más usadas en la teoría de colas son Poisson, Exponencial y Erlang, aunque también existe el caso de distribuciones generales, que no se ajustan a ninguna de estas y que pueden ser construidas al hacer el modelo, esto se debe a la naturaleza de los datos del sistema.

En esta sección se hace un breve resumen de las características de dichas funciones.

Función de distribución exponencial

Esta distribución se emplea para modelar el tiempo entre llegadas. También se emplea para modelar tiempos de servicio que son muy variables, por ejemplo, la duración de una llamada telefónica. Esta distribución se relaciona con la de Poisson, dado que si una tasa de llegadas (llegadas por unidad de tiempo = λ) sigue una distribución de Poisson, el tiempo entre llegadas sigue una distribución exponencial de parámetro $\beta = 1/\lambda$.

Exponencial	*Expo* (β)
Posible interés	Tiempo entre llegadas de clientes cuando la frecuencia media de llegadas es constante.
Densidad de probabilidad	$f(x)=\begin{cases}\frac{1}{\beta}e^{\frac{-x}{\beta}} & x\geq 0\\ 0 & x<0\end{cases}$
Distribución acumulativa	$F(x)=\begin{cases}1-e^{\frac{-x}{\beta}} & x\geq 0\\ 0 & x<0\end{cases}$
Media	β
Varianza	β^2



Función de distribución de Poisson

La frecuencia de aparición de eventos en un proceso de llegadas puede formalizarse al especificar el tiempo entre dos llegadas sucesivas, o especificando el número de eventos de llegada por intervalo.

» ***Tiempo entre 2 eventos de llegada sucesivos:*** en general, el tiempo entre dos eventos independientes de llegada consecutivos suele responder a una distribución exponencial.

» ***Número de eventos de llegada por intervalo:*** en lugar de describir el tiempo entre eventos de llegada, se describe el número de eventos en un intervalo de tiempo constante. Nótese, por ejemplo, que no es posible describir, mediante una distribución exponencial, la llegada de material a una unidad de producción cuando ésta es transportada en *pallets* con un número de piezas variables, ya que el tiempo entre la llegada de una pieza y la siguiente es 0. La distribución de Poisson es una de las más utilizadas para describir este tipo de comportamiento. Esta distribución fue desarrollada originalmente para modelar las llamadas telefónicas a una central. Otros fenómenos que pueden ser modelados son:

1. El número de entidades temporales que llegan por unidad de tiempo.
2. El número total de defectos en una pieza.
3. El número de veces que un recurso es interrumpido por unidad de tiempo.

Poisson	*Poisson* (β)
Función de probabilidad	$f(x)=\begin{cases}\dfrac{e^{-\lambda}\lambda^{x}}{x!} & x\in\{0,1,\cdots\}\\ 0 & resto\end{cases}$
Distribución acumulativa	$F(x)=\begin{cases}0 & x<0\\ e^{-\lambda}\sum_{i=0}^{[x]}\dfrac{\lambda i}{i!} & 0\le x\end{cases}$
Media	λ
Varianza	λ

Función de distribución gamma (Erlang es una versión particular)

En general, el tiempo que una unidad de producción requiere para realizar una operación repetitiva de procesamiento de materia prima, o bien el tiempo consumido en una actividad repetitiva de transporte de material entre dos estaciones de trabajo, suele seguir un valor constante con pequeñas variaciones provocadas por ciertos aspectos físicos. Éstos podrían ser modelados de forma determinante, pero con el objetivo de simplificar la tarea, se suelen describir como el resultado de una actividad aleatoria mediante modelos estadísticos.

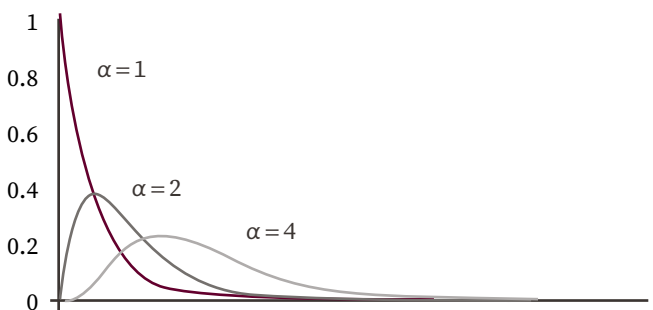
De acuerdo con los parámetros de la función de distribución de probabilidad (fdp) gamma, ésta presenta una gráfica muy similar a la de la fdp normal, pero con una cierta asimetría, que responde a la presencia de datos con valores superiores al valor promedio. Esta asimetría permite modelar secuencias de actividades (por ejemplo, unidades de procesamiento o de transporte) que se realizan en paralelo, tales que cada una de ellas responde a una fdp Normal, pero el tiempo consumido en la secuencia de actividades presenta una asimetría sesgada hacia los valores superiores a la media.

La distribución gamma, cuando α es un entero positivo se conoce con el nombre de Erlang. Existe una asociación entre los modelos de probabilidad de Poisson y de Erlang. Si el número de eventos aleatorios independientes que ocurren en un lapso específico es una variable aleatoria de Poisson con frecuencia constante de ocurrencia igual a $1/\theta$, entonces, para una α dada, el tiempo de espera hasta que ocurre el a-ésimo evento de Poisson sigue una distribución de Erlang.

Cuando $\alpha = 1$, la distribución de Erlang se reduce a una distribución exponencial negativa. Observe que la variable aleatoria de una dis-

tribución exponencial negativa puede pensarse como el lapso que transcurre hasta el primer evento de Poisson. De acuerdo con esto, la variable aleatoria de Erlang es la suma de variables aleatorias independientes distribuidas exponencialmente.

Gamma	*Gamma* (α, β)
Densidad de probabilidad	$f(x)=\begin{cases}\dfrac{\beta^{-\alpha}\,x^{\alpha-1}\,e^{\frac{-x}{\beta}}}{\Gamma(\alpha)} & x\geq 0\\ 0 & x<0\end{cases}$ $\Gamma(\alpha)$ es la función gamma $\Gamma(\alpha)=\int t^{\alpha-1}\,e^{-t}\,dt$ Si α es entero positivo $\Gamma(\alpha)=(\alpha-1)!$
Distribución acumulativa	$F(x)=\begin{cases}1-e^{\frac{-x}{\beta}}\sum_{j=0}^{\alpha-1}\dfrac{(x/\beta)^j}{j!} & x\geq 0\\ 0 & x<0\end{cases}$ Si α es entero positivo; en caso contrario no hay fórmula cerrada
Media	$\alpha\beta$
Varianza	$\alpha\beta^2$



Función de distribución gamma $(\beta=1)$

1.6 Resumen

La **teoría de colas** es el estudio matemático de las colas o líneas de espera dentro de un sistema. Esta teoría estudia factores como el tiempo de espera medio en las colas o la capacidad de trabajo del sistema sin que llegue a colapsar. Dentro de las matemáticas, la teoría de colas se engloba en la investigación de operaciones y es un complemento muy importante a la teoría de sistemas y la teoría de control.

Se trata así de una teoría que encuentra aplicación en una amplia variedad de situaciones como negocios, comercio, industria, ingenierías, transporte y logística o telecomunicaciones.

Elementos de las colas:

Proceso básico de colas: Los clientes que requieren un servicio se generan en una fase de entrada. Estos clientes entran al sistema y se unen a una cola. En determinado momento se selecciona un miembro de la cola, para proporcionarle el servicio, mediante alguna regla conocida como disciplina de servicio. Luego, se lleva a cabo el servicio requerido por el cliente en un mecanismo de servicio, después de lo cual el cliente sale del sistema de colas.

Fuente de entrada o población potencial: Una característica de la fuente de entrada es su tamaño. El tamaño es el número total de clientes que pueden requerir servicio en determinado momento. Puede suponerse que el tamaño es infinito o finito.

Cliente: Es todo individuo de la población potencial que solicita servicio como por ejemplo una lista de trabajo esperando para imprimirse.

Capacidad de la cola: Es el máximo número de clientes que pueden estar haciendo cola (antes de comenzar a ser servidos). De nuevo, puede suponerse finita o infinita.

Disciplina de la cola: La disciplina de la cola se refiere al orden en el que se seleccionan sus miembros para recibir el servicio. Por ejemplo, puede ser:

FIFO (first in first out) primero en entrar, primero en salir, según la cual se atiende primero al cliente que antes haya llegado.

LIFO (last in first out) también conocida como pila que consiste en atender primero al cliente que ha llegado el último.

RSS (random selection of service) que selecciona los clientes de manera aleatoria, de acuerdo con algún procedimiento de prioridad o a algún otro orden.

Processor Sharing – sirve a los clientes igualmente. La capacidad de la red se comparte entre los clientes y todos experimentan con eficacia el mismo retraso.

Mecanismo de servicio: El mecanismo de servicio consiste en una o más instalaciones de servicio, cada una de ellas con uno o más canales paralelos de servicio, llamados servidores.

Redes de colas: Sistema donde existen varias colas y los trabajos fluyen de una a otra. Por ejemplo: las redes de comunicaciones o los sistemas operativos multitarea.

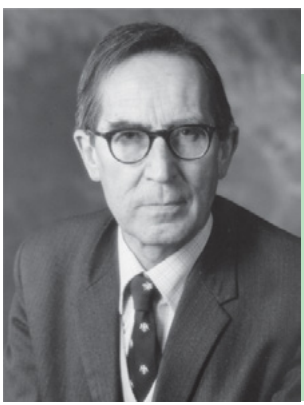
El proceso de servicio: Define cómo son atendidos los clientes.

Por régimen permanente o **estado estacionario**, se entiende la zona de respuesta del sistema en la que, tras haber transcurrido tiempo suficiente, todos los parámetros del sistema se han estabilizado y permanecen a un valor constante

La Ley de Little establece que el número promedio de clientes en un sistema (L) es igual a la tasa promedio de llegada de los clientes al sistema (λ) por el tiempo promedio que un cliente está en el sistema (W).

1.7 Notas históricas

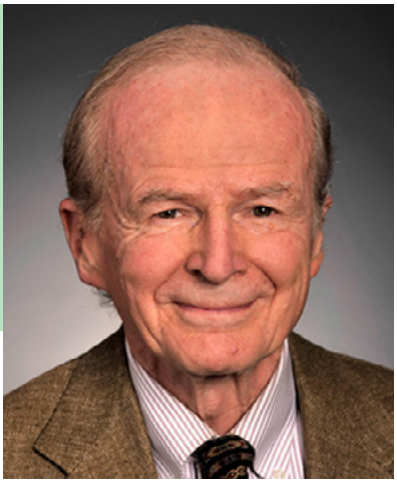
El origen de la teoría de colas está en el esfuerzo de Agner Kraup Erlang (Dinamarca, 1878 - 1929) para analizar en 1909 la congestión de tráfico telefónico con el objetivo de cumplir la demanda incierta de servicios en el sistema telefónico de Copenhague. Sus investigaciones acabaron en una nueva teoría denominada teoría de colas o de líneas de espera. Esta teoría es ahora una herramienta de valor en negocios debido a que un gran número de problemas pueden caracterizarse, como problemas de congestión llegada - salida.



David George Kendall (15 de enero de 1918 - 23 de octubre de 2007) [3] fue un matemático y estadístico inglés, conocido por su trabajo sobre probabilidad, análisis estadístico de formas, líneas ley y teoría de colas. Pasó la mayor parte de su vida académica en la Universidad de Oxford (1946-1962) y la Universidad de Cambridge (1962-1985). Trabajó con MS Bartlett durante la Segunda Guerra Mundial.

En 1962 fue nombrado primer profesor de Estadística Matemática en el Laboratorio de Estadística de la Universidad de Cambridge; en cuyo cargo permaneció hasta su jubilación en 1985. Fue elegido para una beca de profesor en Churchill College, y fue fideicomisario fundador de Rollo Davidson Trust. En 1986, la Universidad de Bath le otorgó un título honorífico (Doctor en Ciencias).

Kendall era un experto en análisis de probabilidad y datos, y fue pionero en el análisis estadístico de formas, incluido el estudio de líneas ley. Definió la notación de Kendall para la teoría de las colas. La Royal Statistical Society le otorgó la Medalla Guy en Plata en 1955, seguida en 1981 por la Medalla Guy en Oro. En 1980, la London Mathematical Society otorgó a Kendall su premio Senior Whitehead y en 1989 su medalla De Morgan. Fue elegido miembro de la Royal Society en 1964. Kendall también jugó un papel clave en la fundación de la Bernoulli Society en 1975, y fue su presidente inicial. Tomado de: https://en.wikipedia.org/wiki/David_George_Kendall



John Dutton Conant Little, conocido por el desarrollo en el campo de la investigación operativa de la llamada ley de Little, nació en Boston en 1928. Obtuvo su título de grado en física en el Instituto Tecnológico de Massachusetts en 1948, tras lo cual trabajó en General Electric (1948-50). Su tesis doctoral, *Use of Storage Water in a Hydroelectric System*, que empleó programación dinámica y fue dirigida por Philip M. Morse, fue la primera tesis sobre investigación operativa en Estados Unidos (1955). Tras ello, fue profesor en la Universidad Case de la Reserva Occidental entre 1957 y 1962, antes de regresar al MIT, donde es profesor desde 1962. En 1988, fue profesor visitante en INSEAD.

En un artículo de 1954, se asumió que la ley de Little era cierta y se usó sin pruebas. La forma $L = \lambda W$ fue publicada por primera vez por Philip M. Morse, donde desafió a los lectores a encontrar una situación en la que la relación no se sostuviera. Little publicó en 1961 su

prueba demostrando que no existía tal situación. La demostración de Little fue seguida por una versión más simple de Jewell y otra por Eilon. Shaler Stidham publicó una prueba diferente y más intuitiva en 1972. (https://en.wikipedia.org/wiki/Little %27s_law)

Telefonía. Es interesante, y de particular importancia, señalar que el origen de la teoría de colas se encuentra en los problemas de congestión de redes telefónicas, campo en el cual se siguen presentando problemas en todo el mundo.

Una breve descripción de este problema es la siguiente: Se observan inmediatamente dos situaciones básicas. Cuando todos los aparatos de un grupo están ocupados, se dice que el grupo está bloqueado, se pierde o se demora. El primer caso es característico de un sistema de pérdidas y el segundo de un sistema de espera, siendo posible los casos mixtos. (Tomado de: Elementos de la teoría de colas, Thomas L. Saaty. Ed Aguilar 1967).

1.8 Ejercicios propuestos

1. Identifique al cliente y al servidor en cada uno de los casos siguientes:

- **a)** Aviones que llegan a un aeropuerto.
- **b)** Base de taxis donde éstos esperan a que lleguen pasajeros.
- **c)** Verificación de las herramientas en un almacén de un taller de maquinado.
- **d)** Cartas procesadas en una oficina de correos.
- **e)** Inscripción a las clases en una universidad.
- **f)** Juicios en la corte.
- **g)** Funcionamiento de las cajas de un supermercado.
- **h)** Funcionamiento de un estacionamiento.

2. Para cada uno de los casos del problema anterior, identifique lo siguiente:

- **a)** Naturaleza de la fuente (finita o infinita).
- **b)** Naturaleza de los clientes que llegan (individualmente o en grupo).
- **c)** Clase de tiempo entre llegadas (probabilísticas o determinísticas).
- **d)** Definición y clase de tiempo de servicio.
- **e)** Capacidad de la cola (finita o infinita).
- **f)** Disciplina de la cola.

3. Estudie el sistema siguiente e identifique las situaciones su procesamiento. Al recibirlas, el supervisor decide si el trabajo es normal o urgente. Algunas órdenes requieren usar una o varias máquinas idénticas. Las demás órdenes se procesan en una línea de producción en dos etapas y hay dos de ellas disponibles. En cada uno de los grupos, se asigna una instalación para manejar los trabajos urgentes.

Los trabajos que llegan a una instalación se procesan en el orden de llegada. Los trabajos terminados se embarcan al llegar, en una zona de embarque con capacidad limitada.

Las herramientas afiliadas para las distintas máquinas son suministradas en un almacén central de herramientas. Cuando una máquina se descompone se manda a un repartidor, de un grupo de servicio para arreglarla.

Las máquinas que trabajan en pedidos urgentes siempre reciben prioridad tanto en la adquisición de herramientas nuevas del almacén como en recibir reparaciones.

4. ¿Cierto o falso?

a) Un cliente impaciente que espera puede optar por desistir (irse).

b) Si se prevé un largo tiempo de espera, un cliente que llega puede optar por rehusar.

c) El cambio de una línea de espera a otra se hace para reducir el tiempo de espera.

5. En cada uno de los casos del problema 1, comente sobre la posibilidad de que los clientes opten por a) b) o c) del ejercicio 4.

2 MODELOS DE COLAS DE POISSON

2.1 Introducción

En este capítulo se consideran los modelos de colas de Poisson usando la notación de Kendall para posteriormente en el capítulo 3 incluir modelos más generales, comenzando con un breve análisis del modelo generalizado de cola de Poisson.

Modelo generalizado de cola de Poisson

El desarrollo del modelo generalizado se basa en el comportamiento a largo plazo, o de estado estable de la cola que se alcanza después de que el sistema ha estado funcionando durante un tiempo suficientemente largo.

Este modelo supone que las frecuencias tanto de las llegadas como de salidas dependen del estado, lo que significa que dependen de la cantidad de clientes en la instalación de servicio.

Definición 2.1

El modelo generalizado define a p_n como función de λ_n y μ_n.
Después se usan esas probabilidades para determinar las medidas de funcionamiento del sistema como la longitud promedio de la

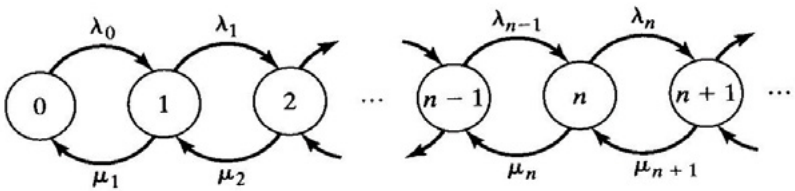
cola, el tiempo promedio de espera y la utilización promedio de la instalación.

Hipótesis 2.1

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

El estado 0 solo puede cambiar al estado 1 cuando hay una llegada con la frecuencia λ_0. Observe que μ_0 no está definida porque no puede haber salidas del sistema si está vacío. Ver figura 2.1.

Figura 2.1 Cambio de estados y las frecuencias asociadas



Hipótesis 2.2

Dado $E_n(t) = n$, la distribución actual de probabilidad del tiempo restante hasta la siguiente muerte (complexión del servicio) es exponencial con parámetro $\mu_n (n = 1, 2, \cdots)$.

Hipótesis 2.3

Solo puede ocurrir un nacimiento o una muerte en un instante. Las expresiones λ_n y μ_n son tasas medias.

Ecuación de balance

Para cualquier estado del sistema *n*, la tasa media (número esperado de ocurrencias por unidad de tiempo) a la que los incidentes de entrada ocurren debe ser igual a la tasa media de los incidentes de salida.

Ecuaciones de estado

Con el fin de considerar una ecuación de balance valore el estado 0. El proceso entra a 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 está en estado 1, la tasa media de entrar al estado 0 es μ_1, desde cualquier otro estado esta tasa media es 0.

Por lo tanto, la tasa media global a la cual el proceso sale de su estado actual para entrar al estado 0 (tasa media de ocurrencias de entrada) es:

$$\mu_1 p_1 + 0(1-p) = \mu_1 p_1$$

Con el mismo razonamiento se tiene que la tasa media de ocurrencia de incidentes de salida es $\lambda_0 q_0$ entonces la ecuación de balance para el estado 0 es:

$$\mu_1 p_1 = \lambda_0 p_0$$

En la tabla 2.1 se muestran las ecuaciones de balance del proceso de nacimiento y muerte.

Tabla 2.1 Ecuaciones de balance

Estado	Tasa de entrada = tasa de salida
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	$\lambda_1 p_1 + \mu_3 p_3 = (\lambda_2 + \mu_2) p_2$
..	...
$n-1$	$\lambda_{n-2} p_{n-2} + \mu_n p_n = (\lambda_{n-1} + \mu_{n-1}) p_{n-1}$
n	$\lambda_{n-1} p_{n-1} + \mu_{n+1} p_{n+1} = (\lambda_n + \mu_n) p_n$

Nótese que la primera ecuación de balance contiene dos variables para las cuales se debe resolver (q_0 y q_1); las dos primeras ecuaciones contienen 3 variables (q_0, q_1 y q_2) y así sucesivamente, de donde se encuentran los valores para las q's en forma recursiva como se ve en la figura.

Tabla 2.2 Ecuaciones de balance para las p

Estado	Ecuaciones
0	$p_1 = \frac{\lambda_0}{\mu_1} p_0$
1	$p_2 = \frac{(\lambda_1 + \mu_1) p_1}{\mu_2} - \frac{\lambda_0 p_0}{\mu_2} = \frac{\lambda_1}{\mu_2} p_1 \frac{1}{\mu_2} \{(\mu_1 p_1 - \lambda_0 q_0) = 0\} = \frac{\lambda_1}{\mu_2} p_1 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} p_0$
...	...
$n-1$	$p_n = \frac{\lambda_{n-1}}{\mu_{n+1}} p_{n-1} + \frac{1}{\mu_{n+1}} (\mu_{n-1} p_{n-1} - \lambda_{n-2} p_{n-2}) = \frac{\lambda_{n-1}}{\mu_{n+1}} p_{n-1} = \frac{\lambda_{n-1} \lambda_{n-2} \cdots \lambda_0}{\mu_n \mu_{n-1} \cdots \mu_1} p_0$
n	$p_{n+1} = \frac{\lambda_n}{\mu_{n+1}} p_n + \frac{1}{\mu_{n+1}} (\mu_n p_n - \lambda_{n-1} p_{n-1}) = \frac{\lambda_n}{\mu_{n+1}} p_n = \frac{\lambda_n \lambda_{n-1} \cdots \lambda_0}{\mu_{n+1} \mu_n \cdots \mu_1} p_0$

De esta manera se llega a las fórmulas de los modelos de Poisson.

Ejemplo 2.1

Una dulcería opera con tres cajas. El gerente usa el siguiente programa para determinar la cantidad de cajas en operación, en función de la cantidad de clientes en la tienda:

Tabla 2.3 Cajas en operación

# clientes en la tienda	# cajas en función
1 a 3	1
4 a 6	2
Más de 6	3

Los clientes llegan a las cajas siguiendo una distribución Poisson, con una frecuencia media de 10 por hora. El tiempo promedio de atención a un cliente es exponencial con 12 minutos de promedio.

Calcular la probabilidad p de estado estable de que haya n clientes en las cajas.

Solución

De la información del problema se tiene:

$$\lambda n = \lambda = 10 \qquad \textit{clientes por hora}, \qquad n = 0, 1, \cdots$$

$$\mu_n = \begin{cases} 60/12 = 5 & \textit{clientes por hora}, \quad n = 1, 2, 3 \\ 2 \times 5 = 10 & \textit{clientes por hora}, \quad n = 4, 5, 6 \\ 3 \times 5 = 15 & \textit{clientes por hora}, \quad n = 7, 8, \cdots \end{cases}$$

$$p_1 = (10/5)\, p_0 = 2\, p_0$$

$$p_2 = (10/5)^2\, p_0 = 4\, p_0$$

$$p_3 = (10/5)^3\, p_0 = 8\, p_0$$

$$p_4 = (10/5)^3\, (10/10)\, p_0 = 8\, p_0$$

$$p_5 = (10/5)^3\, (10/10)^2\, p_0 = 8\, p_0$$

$$p_6 = (10/5)^3\, (10/10)^3\, p_0 = 8\, p_0$$

$$p_n = (10/5)^3\, (10/10)^3\, (10/15)^{n-6}\, p_0 = 8(2/3)^{n-6}\, p_0,\ n = 7,\ 8, \cdots$$

El valor de p_0 se determina con la ecuación

$$p_0 + p_0 \left\{ 2 + 4 + 8 + 8 + 8 + 8 + 8(2/3) + 8(2/3)^2 + 8(2/3)^3 + \cdots \right\} = 1$$

lo que es igual a:

$$p_0 \left\{ 31 + 8\left(1 + (2/3) + (2/3)^2 + \cdots\right) \right\} = 1$$

Se aplica la fórmula de la suma de la serie geométrica:

$$\sum_{i=0}^{\infty} x^i = \frac{1}{1-x}, \quad |x| < 1$$

Para obtener:

$$p_0 \left\{ 31 + 8 \left(\frac{1}{1-(2/3)} \right) \right\} = 1$$

En consecuencia: $p_0 = 1/55$.

Conocida p_0 ya se pueden conocer cada una de las probabilidades del problema, por ejemplo, la probabilidad de que haya entre 1 y 3 clientes en el sistema, es decir:

$$p_1 + p_2 + p_3 = (2 + 4 + 8)(1/55) \approx 0.255$$

Se puede usar p_n para determinar medidas de eficiencia, por ejemplo:

Cantidad esperada de cajas vacías

$$= 3p_0 + 2(p_1 + p_2 + p_3) + 1(p_4 + p_5 + p_6) + 0(p_7 + p_8)$$
$$= 1 \text{ caja}$$

2.2 Una cola-un sevidor-población infinita (/M/M/1: FIFO/∞/∞)

En este modelo se consideran clientes que llegan a un sistema para pedir un servicio. Si el canal de servicio está vacío, la unidad entra y recibe servicio. Si hay ya uno o más clientes en el canal, la disciplina de la cola es "primero en llegar-primero en ser servido". Existe un solo servidor.

Se supone una población infinita y una línea de espera de capacidad ilimitada. Es importante mencionar que en los modelos aquí expuestos asumen que el proceso de llegadas es Poisson.

Es necesario que el factor de utilización del sistema $\rho = \lambda/\mu < 1$, debido a que cuando la tasa de arribos λ es mayor o igual a la tasa de servicios μ, el largo de la cola crece sin límite.

Se tiene entonces para λ y μ lo siguiente:

$$\begin{cases} \lambda_n = \lambda \\ \mu_n = \mu \end{cases} \qquad n = 0, 1, 2, \cdots$$

También $\lambda_{ef} = \lambda$ y $\lambda_{perdida} = 0$. Ya que todos los clientes que llegan pueden entrar al sistema.

Como $L_s(t)$ es el número de clientes en el sistema en el tiempo t, tanto en servicio como en la cola, el incremento $L_s(t)$ en una unidad corresponde al arribo de un cliente, p_n es la probabilidad de que existan n clientes en el sistema en determinado momento.

Así para obtener los resultados de cómo se comporta el sistema se tienen que resolver, sustituyendo los valores respectivos, las siguientes fórmulas.

La probabilidad de que el sistema este vacío al llegar es:

$$p_o = 1 - \rho$$

La probabilidad de que existan clientes en el sistema al llegar es:

$$p_n = (1 - \rho)\rho^n$$

El número esperado de clientes en el sistema L_s es:

$$L_s = \frac{\lambda}{\mu - \lambda}$$

El número esperado de clientes en la cola L_q es:

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)}$$

La cantidad esperada de servidores ocupados $\bar{c}$

$$\bar{c} = L_s - L_q = \rho$$

El tiempo promedio de espera de un cliente en el sistema W_s está dado por:

$$W_s = \frac{L_s}{\lambda} = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu - \lambda}$$

El tiempo promedio de espera de un cliente en la cola W_q es:

Ejemplo 2.2

Una nueva compañía de juegos multimedia cuenta con un centro de asistencia técnica por teléfono. Un técnico toma las llamadas, el tiempo que requiere para responder a las preguntas del cliente tiene una distribución exponencial con media de 5 minutos. Las llamadas llegan según un proceso Poisson con tasa media de 9 por hora.

Este ejemplo se refiere a un modelo de un sistema una cola —un servidor— población infinita. Se tiene presente que no existe prioridad en el servicio y se considera que el mecanismo del servicio es "primero en llegar primero en ser atendido".

Primero hay que especificar los valores de λ y μ:

$\lambda = 9$ llamadas por hora
$\mu = 12$ llamadas por hora

Solución

Para encontrar L_s, el número esperado de llamadas

$$L_s = \frac{\lambda}{\mu - \lambda} = \frac{9}{12-9} = 3 \; llamadas$$

El largo promedio de la cola, L_q, está dado por

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{9^2}{12(12-9)} = \frac{81}{12(3)} = 2.25$$

Esto es, aproximadamente

$$L_q = 2 \; llamadas$$

El tiempo promedio de espera en el sistema, W_s, se obtiene de

$$W_s = \frac{1}{\mu - \lambda} = \frac{1}{12-9} = \frac{1}{3} = 0.333 \text{ de hora}$$

Es decir 20 minutos.

El tiempo promedio de espera en la cola se encuentra utilizando

$$W_q = W - \frac{1}{\mu} = 0.333 - \frac{1}{12} = 0.333 - 0.083 = 0.25 \text{ de hora}$$

Esto es aproximadamente 15 minutos.

Al llamar e ingresar a este sistema en promedio se encuentran 3 personas en él, se espera en una cola virtual de 2 personas durante 15 minutos y se recibe el servicio en 5 minutos más.

Se puede conocer más información al respecto como la probabilidad de que nadie espere en el sistema o que espere solo 1 por ejemplo.

La probabilidad de que nadie esté en el sistema es

$$p_0 = 1 - \rho$$

$$p_0 = 1 - \frac{\lambda}{\mu}$$

$$p_0 = 1 - \frac{9}{12}$$

$$p_0 = 1 - 0.75$$

$$p_0 = 0.25$$

La probabilidad de que solo 1 cliente esté en el sistema es

$$p_n = (1-\rho)\rho^n$$
$$p_1 = (1-0.75)(0.75)$$
$$p_1 = (0.25)(0.75)$$
$$p_1 = 0.1875$$

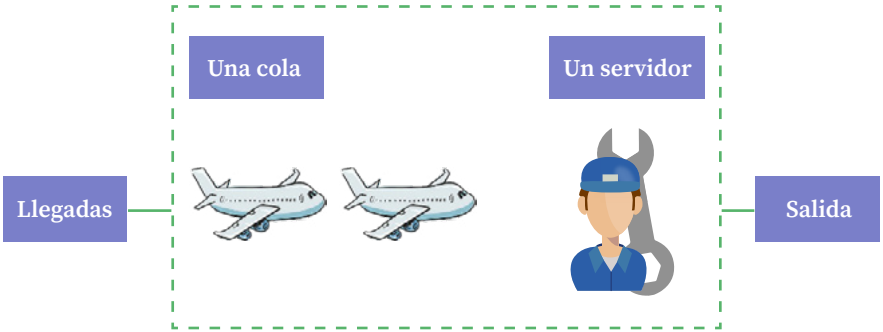
La información que se obtiene se puede manejar según las necesidades personales.

2.3 Una cola-un servidor-población finita (/M/M/1: FIFO/N/∞)

Este es un modelo donde se tiene un número de clientes en el sistema no mayor a un número especificado por N y donde la longitud máxima de la cola es igual a $N-1$. A cualquier cliente que llega cuando la cola está "llena", se le evita la entrada al sistema y, por tanto, sale.

Por ejemplo, cuando la flota de aviones que requieren de revisión para un buen funcionamiento llega a la sala de espera (hangar) y el servidor (grupo de mecánicos) realiza su labor (servicio) empezando por el primero que llega, así después de ser revisado el avión sale del sistema. Como el lugar de revisión tiene un área determinada y los aviones ocupan un espacio grande, tienen que limitar a cierto número de clientes.

Otro ejemplo se presenta al asistir al médico donde es necesario esperar a ser atendidos por un doctor que solo consulta a un determinado número de pacientes.



La tasa media de entrada al sistema se vuelve cero cuando la cola está llena. Entonces es necesario para considerar una cola finita que los parámetros λ_n se modifiquen a

$$\lambda_n = \begin{cases} \lambda & \text{para} \quad n = 0, 1, \cdots, N-1 \\ 0 & \text{para} \quad n \geq N \end{cases}$$

La probabilidad de que el sistema esté vacío al llegar un cliente es:

$$p_0 = \begin{cases} \dfrac{1-\rho}{1-\rho^{N+1}} & \rho \neq 1 \\ \dfrac{1}{N+1} & \rho = 1 \end{cases}$$

La probabilidad de que existan n clientes en el sistema al llegar es:

$$p_n = \begin{cases} \rho^N p_0 & \rho \neq 1 \\ \dfrac{1}{N+1} & \rho = 1 \end{cases}$$

El número esperado de clientes en el sistema L_s es:

$$L_s = E\{n\} = \sum_{n=0}^{N} n p_n = \frac{\rho}{1-\rho} - \frac{(N+1)\rho^{N+1}}{1-\rho^{N+1}}$$

El número esperado de clientes en la cola L_q es:

$$L_q = L_s - \frac{\rho(1-\rho^N)}{1-\rho^{N+1}}$$

Por otra parte, se calcula la λ_{ef} que es la frecuencia de llegada efectiva al sistema, que es igual a λ cuando todos los clientes que llegan pueden entrar al sistema, en caso contrario si hay límite de espacio

para atender a los clientes entonces $\lambda_{ef} < \lambda$, y por lo tanto es la que se usará para calcular W_q.

El tiempo promedio de espera de un cliente en la cola W_q es:

$$W_q = \frac{L_q}{\mu(1-P_0)} = \frac{L_q}{\lambda_{ef}}$$

El tiempo promedio de espera de un cliente en el sistema W es:

$$W_s = W_q + \frac{1}{\mu} = \frac{L_s}{\lambda(1-p_N)}$$

También se demuestra que:

$$\lambda_{ef} = \mu(L_s - L_q) = \lambda(1-p_N)$$

Ejemplo 2.3

Una tienda de servicio por correo tiene una sola línea telefónica, atendida por una operadora que tiene instrucciones de mantener en espera a un máximo de tres clientes en línea, mientras toma sus órdenes. Las llamadas llegan según una distribución de Poisson cada 5 minutos. El tiempo necesario para tomar cada orden es exponencial con un promedio de 6 minutos.

a) ¿Cuánto tiempo espera un cliente en promedio antes de ser atendido por la operadora?

b) ¿Opina usted que el tiempo de espera obtenido en a) es razonable para una tienda de este tipo?

c) Suponiendo que la tienda continuará usando solo una línea telefónica, ¿qué sugeriría para reducir el tiempo de espera en la línea?

Solución

a) Para calcular el tiempo de espera de un cliente W_q primero se calculan los valores de pn para $n=0, 1, 2, 3, 4$. Se considera que puede haber 4 clientes en el sistema, 3 en cola y uno siendo atendido. Sabemos también lo siguiente:

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

Además, se sabe que:

$$\mu = \frac{60}{6} = 10 \qquad \lambda = \frac{60}{5} = 12 \qquad \rho = \frac{\lambda}{\mu} = 1.2$$

Cuando el número de clientes permitido en el sistema es limitado como en este caso $N=4$. Se tiene que p_0 es igual a:

$$p_0 = \begin{cases} \dfrac{1-\rho}{1-\rho^{N+1}} & \rho \neq 1 \\ \dfrac{1}{N+1} & \rho = 1 \end{cases}$$

Por lo tanto, sustituyendo se tiene:

$$p_0 = \frac{1-1.2}{1-(1.2)^5} = 0.13438$$

Para p_n se tiene la formula siguiente:

$$p_n = \begin{cases} \rho^N p_0 & \rho \neq 1 \\ \dfrac{1}{N+1} & \rho = 1 \end{cases}$$

Por lo tanto:

$$p_1 = 0.13438\ (1.2) = 0.16126$$
$$p_2 = 0.13438\ (1.2)^2 = 0.19361$$
$$p_3 = 0.13438\ (1.2)^3 = 0.23221$$
$$p_4 = 0.13438\ (1.2)^4 = 0.27865$$

La cantidad esperada de clientes en el sistema L_s se calcula como sigue:

$$L_s = E\{n\} = \sum_{n=0}^{N} n p_n = 2.359$$

El número esperado de clientes en la cola L_q se calcula con la formula siguiente:

$$L_q = L_s - \frac{\lambda(1 - p_N)}{\mu}$$

Sustituyendo se obtiene $L_q = 1.49338$

Por otra parte, se calcula la λ_{ef} que es la frecuencia de llegada efectiva al sistema, que es igual a λ cuando todos los clientes que llegan pueden entrar al sistema, en caso contrario si hay límite de espacio para atender a los clientes entonces $\lambda_{ef} < \lambda$, y por lo tanto es la que se usará para calcular W_q.

$$\lambda_{ef} = \lambda(1 - p_N) = 12\ (1 - p_4) = 12(1 - 0.27865) = 8.6575$$

Finalmente se tiene que el tiempo de espera en cola para un cliente está dado por la fórmula

$$W_q = \frac{L_q}{\lambda_{ef}} = \frac{1.49338}{8.6575} = 0.1725$$

En términos de una hora es $(0.17258)\ 60 \approx 10$ minutos.

b) El tiempo de espera es excesivo

c) La única opción es hacer más rápido el tiempo de entrega de una orden. Si se reduce por ejemplo el tiempo de servicio a 4 min entonces se tiene

$$\mu = \frac{60}{4} = 12 \qquad \rho = \frac{\lambda}{\mu} = 1 \qquad p_0 = 0.2 \qquad p_N = 0.2$$

$$\lambda_{ef} = \lambda(1 - p_N) = 12(0.8) = 9.6$$

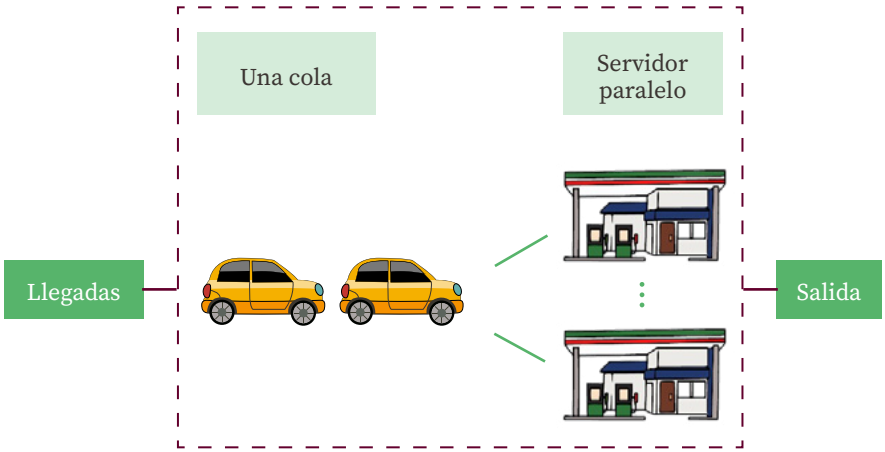
$$L_s = E\{n\} = \sum_{n=0}^{N} n p_n = 0(0.2) + 1(0.2) + 2(0.2) + 3(0.2) + 4(0.2) = 2$$

$$L_q = L_s - \frac{\lambda(1 - p_N)}{\mu} = 2 - \frac{12(1 - 0.2)}{12} = 1.2$$

$$W_q = \frac{L_q}{\lambda_{ef}} = \frac{1.2}{9.6} = 0.125 = 7.5 \text{ min.}$$

2.4 Una cola – c servidores en paralelo – Población infinita (M/M/C):(FIFO/∞/∞)

Este modelo presenta servidores múltiples y una cola como sucede por ejemplo en la mayoría de los bancos. Se llega al sistema y si todos los servidores están ocupados el cliente se forma en la única fila que hay y espera a que sean atendidos los que están adelante, para así recibir el servicio y después abandonar el sistema. Otro ejemplo se presenta cuando al llegar a una gasolinera se dirige hacia una bomba.



Este modelo con c servidores independientes con idéntico tiempo de servicio exponencial, tiene la particularidad de que si hay más de c clientes en el sistema, todos los c servidores están ocupados y cada uno atiende con una tasa media μ: la tasa media de salida del sistema es $c\mu$. Cuando hay menos de c clientes en el sistema, $n < c$, solamente n de los c servidores están ocupados y el sistema tiene una tasa media de servicio de $n\mu$

Así, la tasa de servicio es

$$\mu_n = \begin{cases} n\mu & (1 \le n < c) \\ c\mu & (n \ge c) \end{cases}$$

Y la tasa de llegada es

$$\lambda_n = \lambda$$

Recordando que el proceso de llegadas se asume Poisson

$$\frac{\rho}{c} < 1$$

O bien

$$(\lambda/c\mu \leq 1)$$

La probabilidad de que existan n clientes en el sistema al llegar es:

$$p_n = \begin{cases} \dfrac{\lambda^n}{\mu(2\mu)(3\mu)\cdots(n\mu)}\, p_0 = \dfrac{\lambda^n}{n!\,\mu^n}\, p_0 = \dfrac{p^n}{n!}\, p_0 & n < c \\ \dfrac{\lambda^n}{\prod_{i=1}^{c} i\mu (c\mu)^{n-c}}\, p_0 = \dfrac{\rho^n}{c!\, c^{n-c}}\, p_0 & n \geq c \end{cases}$$

Si $\rho = \lambda/\mu$ y además $\rho/c < 1$ el valor de p_0 se determina con:

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

La probabilidad de que el sistema esté vacío al llegar un cliente es:

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

$$\rho/c < 1$$

El número esperado de clientes en el sistema L_s es:

$$L_s = \frac{\lambda}{\mu} + \left[\frac{(\lambda/\mu)^C \lambda\mu}{(C-1)!\,(c\mu-\lambda)^2}\right] p_0$$

El número esperado de clientes en la cola L_q es:

$$L_q = \left[\frac{(\lambda/\mu)^C \lambda\mu}{(C-1)!\,(c\mu-\lambda)^2}\right] p_0 = \sum_{n=c}^{\infty}(n-c)p_n = \frac{\rho^{c+1}}{(c-1)!\,(c-\rho)^2}\, p_0$$

Como $\lambda_{ef} = \lambda$ entonces $L_s = L_q + \rho$.

El tiempo promedio de espera de un cliente en la cola W_q es:

$$W_q = \left[\frac{(\lambda/\mu)^C \mu}{(C-1)!\,(c\mu - \lambda)^2} \right] p_0 = \frac{L_q}{\lambda}$$

El tiempo promedio de espera de un cliente en el sistema W_s es:

$$W_s = \frac{1}{\mu} + \left[\frac{(\lambda/\mu)^C \mu}{(C-1)!\,(c\mu - \lambda)^2} \right] p_0 = W_q + \frac{1}{\mu}$$

Ejemplo 2.4

Un centro de verificación vehicular tiene 3 casetas de chequeo de automóviles, el tiempo que dura la revisión tiene una distribución exponencial con una media de 4 minutos. Los carros llegan de acuerdo con un proceso de Poisson con una media de 30 autos por hora, bajo la disciplina primero en llegar, primero en ser atendido. Encuentre la probabilidad de que el sistema esté vacío cuando se llega, la cantidad de clientes en cola y el tiempo promedio de espera en cola.

Solución

Los valores para λ y μ son:

$\lambda = 30$ autos por hora
$\mu = 15$ autos por hora
$\rho = \lambda/\mu = 30/15 = 2$ $\qquad\qquad$ $\rho/c = 2/3 < 1$

Primero se calcula p_0, la probabilidad de que el sistema esté vacío

$$p_0 = \left\{ \sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c}{c!\left(1-(\rho/c)\right)} \right\}^{-1} = \left(\sum_{n=0}^{2} \frac{1}{n!} 2^n + \frac{1}{3!}\left(2^3\right) \frac{45}{45-30} \right)^{-1} =$$

$$= \left(1+2+2+\frac{2^3(45)}{3!(15)} \right)^{-1} = 1/9$$

La cantidad de clientes en la cola:

$$L_q = \sum_{n=c}^{\infty} (n-c) p_n = \frac{\rho^{c+1}}{(c-1)!(c-\rho)^2}\, p_0 = \left[\frac{2^3(30)(15)}{2!(15)^2} \right] \frac{1}{9} = \frac{8}{9} \approx 1 \text{ carro}$$

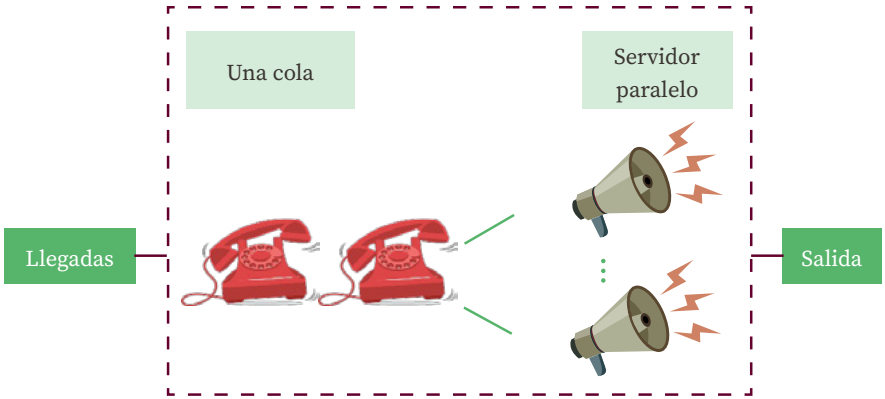
Tiempo promedio de espera en la cola

$$W_q = \frac{L_q}{\lambda} = \frac{0.889}{30} = 0.0296$$

Esto es 1 minuto con 46 segundos.

2.5 Una cola – c servidores en paralelo – Población finita (/M/M/C): (FIFO/N/∞, C ≤ N)

Considerando un modelo de servidores en paralelo en el que en el sistema tiene una capacidad limitada, los modelos que lo representan pueden ser los verificentros que admiten sólo un número determinado de carros, los salones de belleza que atienden a un número finito de clientes, aparatos eléctricos que requieren de reparación y son arreglados por el personal de una empresa determinada. En estos ejemplos el tamaño máximo de la cola es igual a N-c



Se considera un modelo de servidores en paralelo en el que en el sistema tiene una capacidad limitada, con c servidores.

Si $$\lambda_n \begin{cases} \lambda & (0 \le n < K) \\ 0 & (n \ge K) \end{cases}$$

Y $$\mu_n = \begin{cases} n\mu & n < c \\ c\mu & n \ge c \end{cases}$$

La probabilidad de que el sistema esté vacío al llegar es:

$$p_0 = \begin{cases} \left[\sum_{n=0}^{C-1} \frac{1}{n!} \left(\frac{\lambda}{\mu} \right)^n + \frac{(\lambda/\mu)^c}{c!} \frac{1-(\lambda/c\mu)^{K-c+1}}{1-\lambda/c\mu} \right]^{-1} & \lambda/c\mu \neq 1 \\ \left[\sum_{n=0}^{C-1} \frac{1}{n!} \left(\frac{\lambda}{\mu} \right)^n + \frac{(\lambda/\mu)^c}{c!} (K-c+1) \right]^{-1} & \lambda/c\mu = 1 \end{cases}$$

La probabilidad de que al llegar existan n clientes en el sistema es:

$$p_n = \begin{cases} \frac{\rho^n}{n!} p_0 & (0 \le n < c) \\ \frac{\rho^n}{c^{n-c}\, c!} p_0 & (c \le n \le N) \end{cases}$$

El número esperado de clientes en la cola L_q es:

$$L_q = \frac{p_0 (c\rho)^c \rho_c}{c!(1-\rho_c)^2}\left[1-\rho_c^{N-c+1}-(1-\rho_c)(N-c+1)\,\rho_c^{N-c}\right]$$

Por lo tanto, se tiene:

$$L_q = \begin{cases} p_0 \dfrac{\rho^{c+1}}{(c-1)!(c-\rho)^2}\left\{1-\left(\dfrac{\rho}{c}\right)^{N-c}-(N-c)\left(\dfrac{\rho}{c}\right)^{N-c}\left(1-\dfrac{\rho}{c}\right)\right\} & \rho/c \neq 1 \\ p_0 \dfrac{\rho^c\,(N-c)(N-c+1)}{2c!} & \rho/c = 1 \end{cases}$$

El número esperado de clientes en el sistema L_s es:

$$L_s = L_q + c - p_0 \sum_{n=0}^{C-1} \frac{(c-n)(\rho_c c)^n}{n!} = L_Q + (c-\bar{c}) = L_q + \frac{\lambda_{ef}}{\mu}$$

Para determinar W_s se calcula λ_{ef} como sigue:

$$\lambda_{perdido} = \lambda\, p_N$$
$$\lambda_{ef} = \lambda - \lambda_{perdido} = (1-p_N)\,\lambda$$

El tiempo promedio de espera de un cliente en el sistema W_s es:

$$W_s = \frac{L}{\lambda_{ef}}$$

El tiempo promedio de espera de un cliente en la cola es:

$$W_q = \frac{L_q}{\lambda_{ef}}$$

Ejemplo 2.5

En un lote de estacionamiento existen 10 espacios solamente. Los automóviles llegan según una distribución de Poisson con media de 10 por hora. El tiempo de estacionamiento está exponencialmente distribuido con media de 10 minutos. Determine lo siguiente:

a) Número esperado de espacios de estacionamiento vacíos.
b) Probabilidad de que un automóvil que llegue no encontrará un espacio para estacionarse.
c) Tasa efectiva de llegadas del sistema.

Solución

Este es un modelo con varios servidores del tipo (M/M/C) donde $C=10$. Como se tienen varios servidores en paralelo la frecuencia de llegadas λ y la rapidez de servicio es μ por servidor. Como no hay un límite de cantidad en el sistema $\lambda_{ef}=\lambda$, al usar c servidores en el sistema se tiene un aumento en la tasa de servicio proporcional a c.

En nuestro ejemplo se tiene:

$$\lambda_n=\lambda=10$$

$$\mu_n=\begin{cases} n\mu & n<c \\ c\mu & n\geq c \end{cases}$$

$$\mu=60/10=6$$

$$p=\frac{\lambda}{\mu}=\frac{10}{6}=1.667$$

Para calcular las medidas de desempeño se tiene:

$$p_n = \begin{cases} \dfrac{\lambda^n}{\mu(2\mu)(3\mu)\cdots(n\mu)}\, p_0 = \dfrac{\lambda^n}{n!\,\mu^n}\, p_0 = \dfrac{p^n}{n!}\, p_0 & n < c \\ \dfrac{\lambda^n}{\prod_{i=1}^{c} i\mu(c\mu)^{n-c}}\, p_0 = \dfrac{\rho^n}{c!\,c^{n-c}}\, p_0 & n \geq c \end{cases}$$

Si $\rho = \lambda/\mu$ y además $\rho/c < 1$ el valor de p_0 se determina con:

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

Esto da como resultado

$$p_0 = \left\{ \sum_{n=0}^{\infty} \frac{\rho^n}{n!} + \frac{\rho^c}{c!} \sum_{n=c}^{\infty} \left(\frac{\rho}{c!} \right)^{n-c} \right\}^{-1} \qquad \rho/c < 1$$

Sustituyendo se tiene que $p_0 = 0.188888$.

De aquí se tienen los valores para $n = 1, \cdots, 9$ de p_n

n	p_n	n	p_n	n	p_n
1	0.31479	4	0.06072	7	0.00134
2	0.26233	5	0.02024	8	0.00028
3	0.14574	6	0.00562	9	0.00005

Entonces se tiene para L_q y L_s las siguientes formulas:

$$L_q = \sum_{n=c}^{\infty} (n-c)\, p_n = \frac{\rho^{c+1}}{(c-1)!\,(c-\rho)^2}\, p_0$$

Como $\lambda_{ef} = \lambda$ entonces $L_s = L_q + \rho$.

Se tiene que $L_q = 0$ por lo cual $L_s = 0 + \rho = 1.667$

a) Número esperado de lugares para estacionar
$= 10 - L_s = 10 - 1.67 = 8.33$

b) $P\{10 \text{ carros en el estacionamiento.}\} = P_{10} < 0.00001$

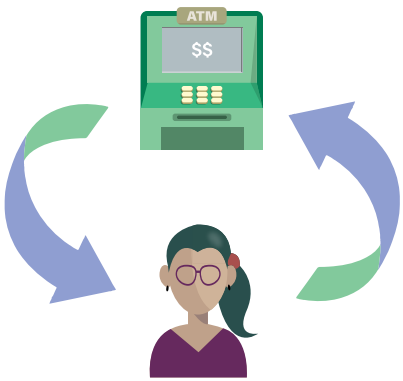
c) $\lambda_{efectiva} = \lambda(1 - P_{10}) \approx \lambda = 10$

2.6 Modelo de autoservicio - servicio infinito: (/M/M/∞)(FIFO/∞/∞)

En este modelo, el número de servidores es ilimitado porque el cliente mismo es también el servidor. Esto es común en los establecimientos de autoservicio, por ejemplo, una prueba escrita para tramitar la licencia de conducir. Se debe tener cuidado con situaciones como los bancos con servicio de 24 horas o ATM, ya que en estos casos el servidor es el cajero automático, aunque sea el cliente quien opera el equipo.

En términos del modelo generalizado se tiene:

$\lambda_n = \lambda$ para $n = 0, 1, 2, \cdots$
$\mu_n = n\mu$ para $n = 0, 1, 2, \cdots$
$\rho = \lambda / n\mu$



Las medidas de desempeño son entonces:

$$p_n = \frac{\lambda^n}{n!\,\mu^n} p_0 = \frac{p^n}{n!} p_0 \qquad n = 0, 1, 2, \cdots$$

Como $\sum_{n=0}^{\infty} p_n = 1$, se tiene entonces:

$$p_0 = \frac{1}{1 + \rho + \frac{\rho^2}{2!} + \cdots} = \frac{1}{e^{\rho}} = e^{-\rho}$$

Se tiene que

$$L_s = \rho$$
$$W_s = 1/\mu$$
$$L_q = W_q = 0$$

Note que $W_q = 0$ ya que cada cliente se atiende a sí mismo.

Ejemplo 2.6

En una instalación de autoservicio las llegadas ocurren según una distribución de Poisson con media de 50 por hora. Los tiempos de servicio por cliente están exponencialmente distribuidos con media de 5 minutos.

a) Encuentre el número esperado de clientes en servicio.
b) ¿Cuál es el porcentaje de tiempo en el que la instalación está inactiva?

Solución

$\lambda_n = \lambda$ para $n = 0, 1, 2, \ldots$.
$\mu_n = n\mu$ para $n = 0, 1, 2, \ldots$.
$\rho = \lambda / n\mu$

Como

$$p_0 = \frac{1}{1 + \rho + \frac{\rho^2}{2!} + \cdots} = \frac{1}{e^{\rho}} = e^{-\rho}$$

Para p_n se tiene: $p_n = \frac{e^{-\rho}\rho^n}{n!}$ $\quad n = 0, 1, 2, \cdots$

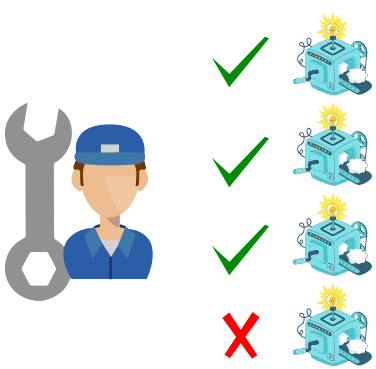
Que es una Poisson con media $L_s = \rho$. Y como es de esperar $L_q = w_q = 0$

a) $L_s = \rho = 4.17$

b) $p_0 = e^{-\rho} = 0.0155$

2.7 Modelo de servicio a máquinas (/M/M/R): (FIFO/*K*/*K*) *R*<*K*

Este es un modelo (M/M/R) donde R ≤ K, también conocido como un modelo de servicio a máquinas. El entorno de este modelo es un taller con *K* máquinas. Cuando se descompone una máquina se llama a un mecánico para hacer la reparación. La frecuencia es de λ descomposturas por máquina por unidad de tiempo, y un mecánico las repara a una tasa de μ máquinas por unidad de tiempo. Todas las descomposturas y servicios siguen una distribución Poisson.



Este modelo tiene una fuente finita de clientes. Si se considera por ejemplo el caso de que todas las máquinas del taller estén descompuestas, no se pueden generar más solicitudes de servicio. Como la frecuencia de descomposturas por máquina es λ, la frecuencia de descomposturas en todo el taller es proporcional a la cantidad

de máquinas que están funcionando. Si se tienen n máquinas en el sistema quiere decir que n máquinas están descompuestas ya que el sistema no es el taller. Entonces la frecuencia de descomposturas en todo el taller es:

$$\lambda_n = (k-n)\lambda \qquad \text{para} \qquad 0 \le n \le K$$

En general se puede expresar como:

$$\lambda_n = \begin{cases} (K-n)\lambda & 0 \le n < K \\ 0 & n \ge K \end{cases}$$

$$\mu_n = \begin{cases} n\mu & 0 \le n < R \\ R\mu & R \le n \le K \end{cases}$$

Para el modelo generalizado se tiene

$$p_n = \begin{cases} \binom{K}{n} \rho^n p_0 & 0 \le n \le R \\ \binom{k}{n} \dfrac{n!\,\rho^n}{R!\,R^{n-R}}\, p_0 & R \le n \le K \end{cases}$$

$$p_0 = \left\{ \sum_{n=0}^{R} \binom{K}{n} \rho^n + \sum_{n=R+1}^{K} \binom{k}{n} \frac{n!\,\rho^n}{R!\,R^{n-R}} \right\}^{-1}$$

Las otras medidas son:

$$L_q = \sum_{n=R+1}^{K} (n-R) p_n$$

$$L_s = L_q + (R - \overline{R}) = L_q + \frac{\lambda_{ef}}{\mu}$$

Donde $\overline{R} = \sum_{n=0}^{R} (R-n) p_n$ es el número esperado de técnicos.

Y $\lambda_{ef} = \mu(R - \overline{R}) = \lambda(K - L_s)$

Esta expresión de λ_{ef} se obtiene de la siguiente manera: Como la tasa de llegadas dadas n máquinas en el sistema es $\lambda(K-n)$ donde λ es la tasa de descompostura de la máquina en condiciones de estado estable se tiene

$$\lambda_{ef} = E\{\lambda(K-n)\} = \lambda(K - L_s)$$

Para $R = 1$ es decir cuando se tiene solo un técnico es:

$$L_q = K - \left(1 + \frac{1}{\rho}\right)(1 - p_0)$$

$$L_s = K - \frac{1 - p_0}{\rho}$$

E **Ejemplo 2.7**

Diez máquinas están siendo atendidas por una sola grúa. Cuando una máquina termina su carga se pide a la grúa que descargue la máquina y la provea de una nueva carga tomada de un área de funcionamiento adyacente. El tiempo de maniobra por carga se supone exponencial con media de 30 minutos. El tiempo desde el momento en que la grúa pone a trabajar una máquina hasta que le trae una nueva carga, también es exponencial con media de 10 minutos.

a) Encuentre el porcentaje de tiempo que la grúa está ociosa.
b) ¿Cuál es el número estimado de máquinas que esperan servicio de la grúa?

Solución

Tenemos que $\lambda = \frac{60}{30} = 2$ $\mu = \frac{60}{10} = 6$ $\rho = \frac{1}{3}$

$$p_0 = \left\{ \sum_{n=0}^{1} \binom{10}{n} \rho^n + \sum_{n=2}^{10} \binom{10}{n} \frac{n!\, p^n}{1!\, 1^{n-1}} \right\}^{-1}$$

Recordemos la fórmula del número de combinaciones de n elementos tomados de m en m.

$$C_n^m = \frac{m!}{n!(m-n)!}$$

a) Entonces $p_0 = 0.00081$

b) $L_q = 10 - \left(1 + \frac{1}{(1/3)}\right)(1 - 0.00081) = 10 - (3.999) = 6.003$

2.8 RESUMEN

Modelo generalizado de cola de Poisson: Este modelo supone que las frecuencias tanto de las llegadas como de salidas dependen del estado, lo que significa que dependen de la cantidad de clientes en la instalación de servicio.

Ecuación de Balance: Para cualquier estado del sistema n, la tasa media (número esperado de ocurrencias por unidad de tiempo) a la que los incidentes de entrada ocurren debe ser igual a la tasa media de los incidentes de salida.

Una cola - un sevidor - población infinita (/M/M/1: FIFO/∞/∞). En este modelo se consideran clientes que llegan a un sistema para pedir un servicio. Si el canal de servicio está vacío, la unidad entra y recibe servicio. Si hay ya uno o más clientes en el canal, la disciplina de la cola es "primero en llegar-primero en ser servido". Existe un solo servidor.

Una cola- un servidor - población finita (/M/M/1: FIFO/N/∞). Este es un modelo donde se tiene un número de clientes en el sistema no mayor a un número especificado por N y donde la longitud máxima de la cola es igual a $N-1$. A cualquier cliente que llega cuando la cola está "llena", se le evita la entrada al sistema y, por tanto, sale.

Una cola - c servidores en paralelo - población infinita (M/(M/c)) : (FIFO /(∞/∞)). Este modelo presenta c servidores múltiples y una cola como sucede por ejemplo en la mayoría de los bancos. Se llega al sistema y si todos los servidores están ocupados el cliente se forma en la única fila que hay y espera a que sean atendidos los que están adelante, para así recibir el servicio y después abandonar el sistema.

Una cola - c servidores en paralelo - población finita (/M/M/C): (FIFO /N/∞, c ≤ N). Considerando un modelo de servidores en paralelo en el que en el sistema tiene una capacidad limitada, los modelos que lo representan pueden ser los verificentros que admiten solo un número determinado de carros.

Modelo de autoservicio - servicio infinito: (/M/M/∞) (FIFO/∞/∞). En este modelo, el número de servidores es ilimitado porque el cliente mismo es también el servidor.

Modelo de servicio a máquinas (/M/M/R): (FIFO/K/K) R<K. Este es un modelo (M/M/R) donde $R \leq K$, también conocido como un modelo de servicio a máquinas. El entorno de este modelo es un taller con

K máquinas. Cuando se descompone una máquina se llama a un mecánico para hacer la reparación.

2.9 NOTAS HISTÓRICAS

Siméon Denis Poisson (1781-1840) matemático, astrónomo y físico francés. Fue alumno de Lagrange y Laplace en l'École Polytechnique, donde comenzó su actividad docente como ayudante de Fourier.

Miembro de la Academia de Ciencias, presidente del Bureau des Longitudes y profesor de Mecánica de la Facultad de Ciencias. Para Poisson "la vida es trabajo". De su esfuerzo continuado a lo largo de su vida surgieron más de trescientas obras que recogen importantes aportaciones a la física (elasticidad, magnetismo, calor, capilaridad, mecánica celeste) y a la matemática (teoría de números, probabilidad, series de Fourier).

Su nombre está asociado a un buen número de conceptos relacionados con estas ciencias: ecuación de Poisson, coeficiente de Poisson, ley de Poisson, paréntesis de Poisson, distribución de Poisson, integral de Poisson.

Poisson dedicó su vida a la investigación y enseñanza de las matemáticas. De su mano surgieron numerosas memorias (sus biógrafos las cifran entre 300 y 400) con aportaciones originales en muchos campos. Y una serie de tratados con los que pretendió formar una gran obra de física matemática que no llegó a concluir.

Tomado de: https://virtual.uptc.edu.co/ova/estadistica/docs/autores/pag/mat/poisson-1.asp.htm

2.10 EJERCICIOS PROPUESTOS

1. Se sabe que el tiempo entre fallas de un refrigerador Kencore es exponencial, con una media de 9000 horas (más o menos 1 año de funcionamiento), y la empresa otorga una garantía de 1 año con el refrigerador. ¿Cuáles son las probabilidades de que la garantía cubra una reparación por descompostura?

2. La U. de G. administra dos líneas de autobuses en el campus: roja y verde. La línea roja da servicio al campus norte, y la verde al sur; y hay una estación de trasbordo que enlaza a las dos líneas. Los autobuses verdes llegan al azar (tiempo exponencial entre llegadas) a la estación de transferencia cada 10 minutos. Los rojos también llegan al azar, cada 7 minutos en promedio.

a) ¿Cuál es la distribución de probabilidades de tiempo de espera para que un alumno que llega en la línea roja se suba a la línea verde?

b) ¿Cuál es la distribución de probabilidades de tiempo de espera para que un alumno que llega en la línea verde se suba a la línea roja?

3. Una tienda de servicio por correo tiene una sola línea telefónica, atendida por una operadora que tiene instrucciones de mantener en espera a un máximo de tres clientes en línea, mientras toma sus órdenes. Las llamadas llegan según una distribución de Poisson cada 5 minutos. El tiempo necesario para tomar cada orden es exponencial con un promedio de 6 minutos.

a) ¿Cuánto tiempo en promedio espera un cliente antes de ser atendido por la operadora?

b) ¿Opina usted que el tiempo de espera obtenido en a) es razonable para una tienda de este tipo?

c) Suponiendo que la tienda continuará usando solo una línea telefónica, ¿qué sugeriría para reducir el tiempo de espera en la línea?

4. En un lote de estacionamiento existen 10 espacios solamente. Los automóviles llegan según una distribución de Poisson con media de 10 por hora. El tiempo de estacionamiento está exponencialmente distribuido con media de 10 minutos. Determine lo siguiente:

a) Número esperado de espacios de estacionamiento vacíos.

b) Probabilidad de que un automóvil que llegue no encontrará un espacio para estacionarse.

c) Tasa efectiva de llegadas del sistema.

5. En una instalación de autoservicio las llegadas ocurren según una distribución de Poisson con media de 50 por hora. Los tiempos de servicio por cliente están exponencialmente distribuidos con media de 5 minutos.

a) Encuentre el número esperado de clientes en servicio.

b) ¿Cuál es el porcentaje de tiempo que la instalación está inactiva?

6. Diez máquinas están siendo atendidas por una sola grúa. Cuando una máquina termina su carga se pide a la grúa que descargue la máquina y la provea de una nueva carga tomada de un área de funcionamiento adyacente. El tiempo de maniobra por carga se supone exponencial con media de 30 minutos. El tiempo desde el momento en que la grúa pone a trabajar una máquina hasta que le trae una nueva carga, también es exponencial con media de 10 minutos.

a) Encuentre el porcentaje de tiempo que la grúa está ociosa.

b) ¿Cuál es el número estimado de máquinas que esperan servicio de la grúa?

7. Lavado automático para automóviles funciona solo con un lugar. Los autos llegan siguiendo una distribución de Poisson, con 4 autos por hora, que pueden esperar en el estacionamiento de la instalación, si el lugar de lavado está ocupado. El tiempo para lavar y limpiar un automóvil es exponencial, con media de 10 minutos. Los automóviles que no se pueden estacionar en la instalación lo pueden hacer enfrente del establecimiento, lo que implica que no hay límite del tamaño del sistema.

a) El gerente quiere determinar el tamaño del estacionamiento.

b) Suponemos ahora que el tiempo de servicio se cambia a constante igual a 10 minutos, ¿cómo afecta ese nuevo sistema al funcionamiento de la instalación?

3 MODELOS DE COLAS GENERALES Y REDES DE COLAS

En este capítulo se consideran modelos de colas generales donde pueden presentarse variaciones de modelos de Poisson o de modelos que no se comportan de esta manera, con la finalidad de ampliar el panorama de los modelos y su uso al tomar decisiones. También se presenta el tema de redes de colas, sus métricas y aplicaciones.

3.1 MODELO DE COLAS QUE NO OBEDECE UNA DISTRIBUCIÓN POISSON[2].

Los modelos de colas que donde los procesos de llegada y/o salida no siguen una distribución de Poisson arrojan resultados complejos y poco manejables en cuyo caso es preferible usar la simulación. Sin embargo hay modelos de colas que sin seguir una distribución Poisson tienen resultados analíticos, concretamente el modelo *(M/G/1):(DG/∞/∞)*, donde se tiene un tiempo de servicio general con media $E\{t\}$ y varianza $var\{t\}$. Desafortunadamente, el análisis de esta situación es un tanto restringido ya que no se proporciona una expresión analítica manejable para las probabilidades p_n. En vez de esto los resultados del modelo sólo proporcionan las medidas básicas de desempeño L_s, L_q, W_s y W_q.[3]

2 Taha, Introducción a la Investigación de Operaciones, 5ª ed. Alfaomega, 1995.
3 Taha, Introducción a la Investigación de Operaciones, 2004.

Sea λ la tasa de llegadas a una instalación con un servidor, dadas la media $E\{t\}$ y la varianza $var\{t\}$ de la distribución del tiempo de servicio, se demuestra usando cadenas de Markov que:

$$L_s = \lambda E\{t\} + \frac{\lambda^2\left(E^2\{t\} + var\{t\}\right)}{2\left(1-\lambda E\{t\}\right)} \qquad (M/G/1)$$

Donde $\lambda E\{t\} < 1$. Esta fórmula se denomina la fórmula de **Pollaczek-Khintchine** (P-K).[4]

De esta fórmula se obtienen las siguientes medidas de desempeño:

$$W_s = \frac{L_s}{\lambda}$$

$$L_q = Ls - \lambda E\{t\}$$

$$W_q = \frac{L_q}{\lambda}$$

La tasa de servicio está dada por $\mu = 1/E\{t\}$. En este modelo $\lambda_{ef} = \lambda$. Para cuando el tiempo de servicio es aproximadamente constante, $\text{var}\{t\} = 0$. Y la fórmula P-K se reduce a:

$$L_s = \rho + \frac{\rho^2}{2\left(1-\rho\right)} \qquad (M/D/1)$$

4 Para una explicación más amplia de la fórmula se puede consultar https://www.netlab.tkk.fi/opetus/s383143/kalvot/E_mg1jono.pdf La fórmula se publicó por primera vez por Felix Pollaczek en 1930. y fue readaptada en términos probabilísticos por Aleksandr Khinchin dos años después. En teoría de riesgo, la fórmula puede usarse para calcular la probabilidad de ruina final (probabilidad de que una compañía de seguros quiebre). https://es.wikipedia.org/wiki/F %C3 %B3rmula_Pollaczek-Khintchine.

Donde $\rho = \lambda/\mu$ y μ es la tasa constante de servicio.

Si el tiempo de servicio es tipo Erlang con parámetros m y μ, con $E\{t\} = 1/\mu$ y $var\{t\} = 1/m\mu^2$, con la fórmula P-K se tiene:

$$L_s = \rho + \frac{1+m}{2m}\left(\frac{\rho}{1-\rho}\right) \qquad (M/E_m/1)$$

Ejemplo 3.1

Considere un establecimiento de lavado de autos que se realiza con máquinas automáticas por lo que el tiempo de servicio se considera constante para todos los autos. El ciclo de la máquina lavadora tarda exactamente 10 minutos, para analizar la situación se consideran llegadas Poisson con una media de 4 por hora.

Solución

Para analizar la situación se observa que $\lambda = 4$ por hora. Por otra parte, como el tiempo de servicio es constante, tenemos $E\{t\} = 1/6$ hora y $var\{t\} = 0$. Por lo tanto:

$$L_s = 4\left(\frac{1}{6}\right) + \frac{4^2\left[(1/6)^2 + 0\right]}{2\left(1-(4/6)\right)} = 1.333 \text{ automóviles}$$

$$L_q = 1.333 - (4/6) = 0.667 \text{ automóviles}$$

$$W_s = \frac{1.333}{4} = 0.333 \text{ hora}$$

$$Wq = \frac{0.667}{4} = 0.167 \text{ hora}$$

3.2 Lineas de espera con prioridad de servicio

Los modelos de colas con prioridad suponen varias líneas en paralelo incluyendo clientes que pertenecen a cierto orden de prioridad. Si la instalación tiene m filas, suponemos que la fila 1 tiene la más alta prioridad de servicio y la línea de espera m incluye a clientes con la más baja prioridad. Las tasas de llegada y servicio pueden variar para las diferentes filas de prioridad. Sin embargo, supondremos que los clientes formados en cada línea de espera son atendidos bajo la disciplina FIFO.

El servicio de prioridad puede seguir dos reglas:

1. Regla de prioridad, donde el servicio de un cliente de menor prioridad puede ser interrumpido para atender a un cliente que llegue con más prioridad.
2. Regla de no prioridad, donde un cliente una vez que está siendo atendido, saldrá del establecimiento una vez que termine su servicio, independientemente de la prioridad del cliente que llegue.

Comenzaremos con dos modelos de no prioridad que se aplican a servidores únicos y múltiples. El modelo de servidor único supone llegadas de Poisson y distribuciones de servicio arbitrarias. En el caso de los servidores múltiples las llegadas y salidas siguen la distribución de Poisson. Se usa la terminología NPRP para denotar la disciplina de no prioridad; M_i y G_i representan distribuciones de Poisson y arbitraria.

Modelo (Mi/Gi/1): (NPRP/∞/∞)

Sea $F_i(t)$ la función de distribución acumulada del tiempo de servicio arbitraria para la *i-ésima* línea de espera ($i = 1, 2, \cdots, m$) y sea Ei{t}

y vari{t} la media y la varianza, respectivamente. Sea λ_i la tasa de llegada en la *i-ésima* línea de espera por el tiempo unitario.

Se definen $L_q^{(k)}$, $L_s^{(k)}$, $W_q^{(k)}$ y $W_s^{(k)}$ en la forma habitual salvo que ahora representan las medidas de la *k-ésima* línea de espera. Por lo tanto, los resultados de esta situación están dados como sigue:

$$W_q^{(k)} = \frac{\sum_{i=1}^{m} \lambda_i \left(E_i^2\{t\} + var_i\{t\}\right)}{2\left(1 - S_{k-1}\right)\left(1 - S_k\right)}$$

$$L_q^{(k)} = \lambda_k \, W_q^{(k)}$$

$$W_s^{(k)} = W_q^{(k)} + E_k\{t\}$$

$$L_s^{(k)} = L_q^{(k)} + \rho_k$$

Donde

$$\rho_k = \lambda_k \, E_k\{t\}$$

$$S_k = \sum_{i=1}^{k} \rho_i < 1 \qquad K = 1, 2, \cdots, m$$

$$S_0 \equiv 0$$

Observamos que W_q, que es el tiempo de espera estimado en la línea de espera para cualquier cliente sin importar su prioridad, está dado por:

$$W_q = \sum_{k=1}^{m} \frac{\lambda_k}{\lambda} \, W_q^{(k)}$$

Donde $\lambda = \sum_{i=1}^{m} \lambda_i$ y $\frac{\lambda_k}{\lambda}$ es el peso relativo de $W_q^{(k)}$

Un resultado similar se aplica a W_s.

Ejemplo 3.2

A un taller de producción llegan trabajos en tres categorías: orden urgente, orden regular o normal y orden de baja prioridad. Aunque los trabajos urgentes son procesados antes que cualquier otro trabajo y los trabajos normales o regulares tienen preferencia sobre las órdenes de baja prioridad, cuando cualquier trabajo ha empezado, deberá terminarse antes de que se inicie uno nuevo. Las llegadas de órdenes de trabajo de las tres categorías son de Poisson con medias 4, 3 y 1 por día. Las tasas de servicio respectivas son constantes e iguales a 10, 9 y 5 por día.

Solución

En esta situación de espera, tenemos tres líneas de espera de no prioridad. Supongamos que las líneas se representan por 1, 2 y 3 de acuerdo con la categoría de los trabajos, por lo tanto, se tiene:

$$\rho_1 = \lambda_1 E\{t_1\} = 4\,(1/10) = 0.4$$

$$\rho_2 = 3\,(1/9) = 0.333$$

$$\rho_3 = 1\,(1/5) = 0.2$$

Se tiene asimismo que:

$$S_1 = \rho_1 = 0.4$$

$$S_2 = \rho_1 + \rho_2 = 0.733$$

$$S_3 = \rho_1 + \rho_2 + \rho_3 = 0.933$$

Como $S_3 < 1$, el sistema puede alcanzar condiciones de estado estable. Por lo tanto, se puede determinar el tiempo de espera estimado en cada línea de espera de la siguiente manera.

$$\sum_{i=1}^{m} \lambda_i \left(E_i^2\{t\} + var_i\{t\} \right) = 4\left[(1/10)^2 + 0\right] + 3\left[(1/9)^2 + 0\right] + 1\left[(1/5)^2 + 0\right] = 0.117$$

De aquí se tiene:

$$W_q^2 = \frac{0.117}{2(1-0.4)(1-0.733)} = 0.365 \text{ días} \cong 8.77 \text{ horas}$$

$$W_q^3 = \frac{0.117}{2(1-0.733)(1-0.933)} = 3.27 \text{ días} \cong 78.5 \text{ horas}$$

$$W_q^1 = \frac{0.117}{2(1-0)(1-0.4)} = 0.0975 \text{ días} \cong 2.34 \text{ horas}$$

El tiempo de espera general estimado para cualquier cliente sin importar la prioridad está dado por

$$W_q = \frac{4(2.34) + 3(8.77) + 1(78.5)}{4+3+1} = 14.27$$

También se puede obtener el número promedio de trabajos en espera de ser procesados en cada línea de espera de prioridad.

$$L_q^1 = 4(0.0975) = 0.39 \text{ trabajos}$$

$$L_q^2 = 3(0.365) = 1.095 \text{ trabajos}$$

$$L_q^3 = 1(3.27) = 3.27 \text{ trabajos}$$

Modelo (M1/M/c) : (NPRP/∞/∞)

Este modelo supone que todos los clientes tienen la misma distribución del tiempo de servicio, independientemente de sus prioridades, y que los c canales tienen una distribución de servicio exponencial idéntica con tasa de servicio μ. Las llegadas en la *k-ésima* línea de espera con prioridad ocurren según una distribución Poisson con una tasa de llegadas λ_k, $k=1, 2, \cdots, m$. Se puede demostrar para la *k-ésima* línea de espera que:

$$W_q^{(k)} = \frac{E\{\xi_0\}}{(1-S_{k-1})(1-S_k)}, \; k=1, 2, \cdots, m$$

Donde $S \equiv 0$ y

$$S_k = \sum_{i=1}^{k} \frac{\lambda_i}{c\mu} < 1, \; \forall k$$

E **Ejemplo 3.3**

Para ilustrar los cálculos que se realizan en el modelo, suponemos que se tienen tres líneas de espera con prioridad y tasas de llegadas $\lambda_1 = 2$, $\lambda_2 = 5$ *y* $\lambda_3 = 10$ por día. Existen dos servidores y la tasa de servicio es de 10 por día. Las llegadas y salidas siguen distribuciones Poisson.

$$S_1 = \frac{\lambda_1}{c\mu} = \frac{2}{2(10)} = 0.1$$

$$S_2 = S_1 + \frac{\lambda_2}{c\mu} = 0.1 + \frac{5}{2(10)} = 0.35$$

$$S_3 = S_2 + \frac{\lambda_3}{c\mu} = 0.35 + \frac{10}{2(10)} = 0.85$$

Como todas las $S_i < 1$, se puede alcanzar el estado estable. Ahora, por definición,

$$\rho = \frac{\lambda_1 + \lambda_2 + \lambda_3}{\mu} = \frac{17}{10} = 1.7$$

Por lo tanto:

$$E\{\xi_0\} = \frac{1}{(10)\,2\{(1.7)^{-2}(2-1.7)(1!)(1+1.7)+1\}} = 0.039$$

En consecuencia

$$W_q^1 = \frac{0.039}{(1-0.1)} = 0.0433$$

$$W_q^2 = \frac{0.039}{(1-0.1)(1-0.35)} = 0.0665$$

$$W_q^3 = \frac{0.039}{(1-0.35)(1-0.85)} = 0.4$$

El tiempo de espera en la fila para cualquier cliente está dado por

$$W_q = \frac{\lambda_1}{\lambda} W_q^{(1)} + W_q = \frac{\lambda_2}{\lambda} W_q^{(2)} + W_q = \frac{\lambda_3}{\lambda} W_q^{(3)} + \frac{2}{17}(0.0433) +$$

$$+ \frac{5}{17}(0.0665) + \frac{10}{17}(0.4) = 0.26$$

Por último, la espera estimada en la línea para todo el sistema está dada por:

$$L_q = \lambda W_q = 17(0.26) = 4.42$$

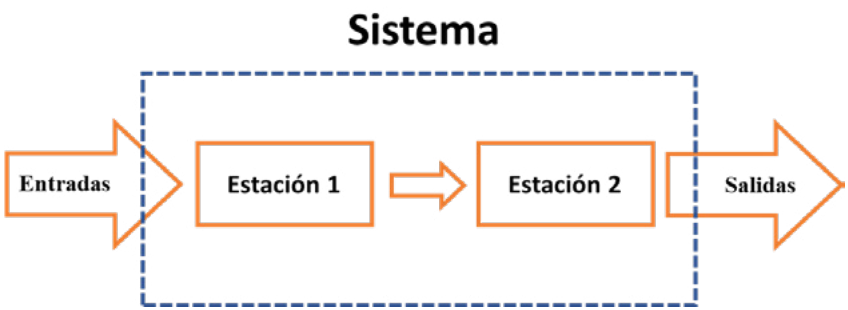
3.3 Líneas de espera sucesivas o en serie

En esta sección se consideran líneas de espera de Poisson con estaciones de servicio dispuestas en serie, de manera que el cliente debe pasar por todas las estaciones antes de completar su servicio. Primero se presenta un caso sencillo de dos estaciones en serie donde no se admiten líneas de espera, y después se presenta un resultado importante para la línea de espera en serie de Poisson, sin límite de espera.

Modelo en serie de dos estaciones con capacidad de líneas de espera cero

Como un ejemplo de análisis de filas en serie, se considera un sistema de colas de un canal simplificado como se muestra en la figura 3.1.

Figura 3.1 Modelo de cola en serie con dos servidores



Un cliente que llega para ser atendido debe pasar por la estación 1 y la estación 2. Los tiempos de servicio en cada estación están exponencialmente distribuidos con una tasa de servicio μ. Las llegadas ocurren según una distribución de Poisson con tasa λ. No se permite ninguna cola enfrente de las estaciones 1 o 2.

La construcción del modelo requiere primero identificar los estados del sistema en cualquier punto del tiempo, lo que se logra como sigue. Cada estación puede estar libre u ocupada. La estación 1 se dice que está bloqueada si el cliente en esta estación completa su servicio antes de que la estación 2 llegue a estar libre. Los símbolos 0, 1 y b representan los estados libre, ocupado y bloqueado, respectivamente.

Sean i y j los estados de las estaciones 1 y 2. Entonces los estados del sistema se pueden representar como:

$$\{(i, j)\} = \{(0, 0), (1, 0), (0, 1), (1,1), (b, 1)\}.$$

Se define $p_{ij}(t)$ como la probabilidad de que el sistema se halle en el estado (i, j) en el tiempo t. Las probabilidades de transición entre los tiempos t y $t+h$ (donde h es un incremento pequeño positivo en el tiempo), se resumen en la tabla siguiente, donde los cuadros vacíos indican que las transiciones entre los estados indicados en t y $t+h$ son imposibles ($=0$).

Tabla 3.1. Estados del sistema

		Estados en $(t+h)$				
		(0, 0)	(0, 1)	(1, 0)	(1, 1)	(b, 1)
Estados en t	(0, 0)	$1-\lambda h$		λh		
	(0, 1)	$\mu h(1-\lambda h)$	$1-\mu h-\lambda h$		$\lambda h(1-\mu h)$	
	(1, 0)		$\mu h\,(1-\lambda h)$	1-μh		
	(1, 1)		$\mu h\,(1-\lambda h)$	μh	$(1-\mu h)\,(1-\mu h)$	μh
	(b, 1)		$\mu h\,(1-\lambda h)$			$(1-\mu h)$

Por esto (sin tomar en cuenta los términos en h2), se pueden escribir las siguientes ecuaciones:

$$
\begin{aligned}
p_{00}(t+h) &= p_{00}(t)\,(1-\lambda h) + p_{01}(t)(\mu h) \\
p_{01}(t+h) &= p_{01}(t)\,(1-\mu h-\lambda h) + p_{10}(t)(\mu h) + p_{b1}(t)(\mu h) \\
p_{10}(t+h) &= p_{00}(t)\,(\lambda h) + p_{10}(t)\,(1-\mu h) + p_{11}(t)(\mu h) \\
p_{11}(t+h) &= p_{01}(t)\,(\lambda h) + p_{11}(t)\,(1-2\mu h) \\
p_{b1}(t+h) &= p_{11}(t)\,(\mu h) + p_{b1}(t)\,(1-\mu h)
\end{aligned}
$$

Reordenando los términos y tomando los límites apropiados, las ecuaciones de estado estable son:

$$
\begin{aligned}
p_{01} - \rho p_{00} &= 0 \\
p_{10} + pb1 - (1+\rho)\,p_{01} &= 0 \\
\rho p_{00} + p_{11} - p_{10} &= 0 \\
\rho p_{01} - 2p_{11} &= 0 \\
p_{11} - p_{b1} &= 0
\end{aligned}
$$

Una de estas ecuaciones es redundante, por lo tanto, agregando la condición

$$p_{00} + p_{11} + p_{10} + p_{01} + p_{b1} = 1$$

La solución para p_{ij} es:

$$p_{00} = \frac{2}{A} \qquad p_{01} = \frac{2\rho}{A}$$

$$p_{10} = \frac{\rho^2 + 2\rho}{A} \qquad p_{11} = p_{b1} = \frac{\rho^2}{A}$$

Donde A es igual a:

$$A = 3\rho^2 + 4\rho + 2$$

El número esperado de clientes en el sistema se puede obtener como:

$$L_s = 0\,p_{00} + 1(p_{01} + p_{10}) + 2(p_{11} + p_{b1}) = \frac{5\rho^2 + 4\rho}{A}$$

Ejemplo 3.4

Una línea de subensamblaje de dos estaciones es operada por un sistema de bandas de transporte. El tamaño del producto ensamblado no permite que se almacene más de una unidad en cada estación. El producto llega a la línea de subensamblaje de otra instalación de producción, según una distribución de Poisson con la media de 10 por hora. Los tiempos de ensamblaje en las estaciones 1 y 2 son exponenciales con media de 5 minutos cada uno. Todos los artículos que llegan y que no pueden entrar directamente a la línea de ensamblaje son dirigidos a otras líneas de subensamblaje.

Solución

Como $\lambda = 10$ por hora y $\mu = 60/5 = 12$ por hora, tenemos $\rho = \lambda/\mu = 10/12 = 0.833$. Podemos determinar todas las probabilidades observando que:

$$A = 3\,(0.833)^2 + 4(0.833) + 2 = 7.417$$

Por lo tanto,

$$p_{00} = \frac{2}{7.417} = 0.2697 \qquad p_{01} = \frac{2(0.833)}{7.417} = 0.2247$$

$$p_{10} = \frac{(0.833)^2 + 2(0.833)}{7.417} = 0.3183 \qquad p_{11} = p_{b1} = \frac{(0.833)^2}{7.417} = 0.917$$

La probabilidad de que un artículo que llega entrará en la estación 1 es

$$p_{00} + p_{01} = 0.2697 + 0.2247 = 0.4944$$

De tal manera que la tasa efectiva de llegadas es

$$\lambda_{ef} = 0.4944(10) = 4.944 \text{ trabajos por hora.}$$

Ya que

$$L_s = \frac{5(0.833)^2 + 4(0.833)}{7.417} = 0.917$$

Se deduce que el tiempo de espera en el sistema es:

$$W_s = \frac{L_s}{\lambda_{ef}} = \frac{0.917}{4.944} = 0.185 \text{ por hora}$$

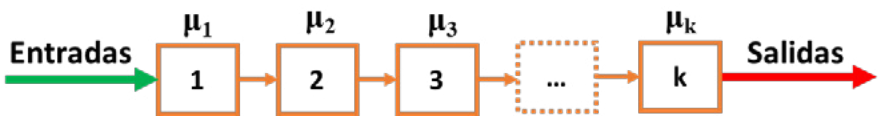
Se observa que W_s representa el tiempo de servicio estimado por artículo ya que no se permiten líneas de espera. Notamos que un artículo puede ser atendido en un tiempo promedio de $5 + 5 = 10$ minutos, o sea, 0.167 por hora siempre que no esté bloqueada la estación 1. Por lo tanto, la diferencia entre W_s $(= 0.185)$ y 0.167 puede considerarse en realidad como el tiempo promedio que espera un artículo

en virtud del bloqueo de la estación 1, es decir $0.185 - 0.167 = 0.018$ por hora o 1.08 minutos.

Modelo en serie de k estaciones con capacidad de líneas de espera infinita

Considere un sistema con k estaciones en serie, como se ilustra en la figura 3.2, y suponga que las llegadas a la estación 1 son generadas por una población infinita, de acuerdo con una distribución de Poisson con tasa media de llegada λ. Las unidades atendidas pasarán sucesivamente de una estación a la siguiente hasta que se descarguen de la estación k. La distribución de tiempo de servicio en cada estación i es exponencial con tasa media μ_i, $i = 1, 2, \cdots k$. Además, no hay límite de líneas de espera en cualquier estación.

Figura 3.2 Modelo en serie de k estaciones o servidores



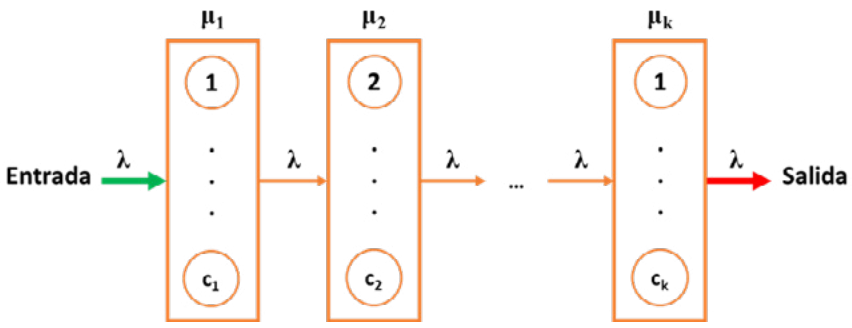
En estas condiciones puede comprobarse que, para toda i, la salida de la estación i, o entrada a la estación $i+1$ es Poisson con una media λ, y que cada estación puede tratarse independientemente como (M/M/1). Esto significa que para la *i-ésima* estación, las probabilidades de estado estable p_{ni} están dadas por:

$$p_{ni} = (1 - \rho_i)\rho_i^{ni}, \qquad n_i = 0, 1, 2, \cdots$$

Para $i = 1, 2, \cdots, k$, donde n_i es el número en el sistema que solo consta de la estación i. Los resultados de estado estable existirán únicamente si $\rho_i = \lambda / \mu_i < 1$.

El mismo resultado puede extenderse al caso donde la estación i incluye c_i servidores en paralelo, cada uno con la misma tasa de servicio exponencial μi por unidad de tiempo, como se muestra en la figura 3.3.

Figura 3.3 Servidores en paralelo en estaciones en serie



En este caso cada estación se puede tratar de manera independiente como $(M/M_i/c_i)$ con tasa media de llegadas λ. Y nuevamente los resultados de estado estable presentados previamente prevalecerán únicamente si $\lambda < c_i\,\mu_i$, para $i = 1, 2, \cdots, k$.

Ejemplo 3.5

En una línea de producción con cinco estaciones en serie, llegan trabajos a la estación 1 con una distribución Poisson con tasa media $\lambda = 20$ por hora. El tiempo de producción en cada estación es exponencial con media de 2 minutos. La salida de la estación i se utiliza como entrada de la estación $i+1$. La parte de artículos en buenas condiciones que se producen en la estación i es α_i de la entrada total a la misma estación. La parte restante $(1 - \alpha_i)$ son artículos defectuosos y se deben desechar.

Suponemos que nos interesa el espacio de almacenamiento entre estaciones sucesivas que darán cabida a todos los artículos que lleguen el $100\beta\,\%$ del tiempo. La probabilidad pni de ni artículos en la estación i está dada por:

$$p_{ni} = (1-\rho_i)\,\rho_i^{ni}$$

Donde $\rho_i = \lambda_i/\mu_i$. Por lo tanto, el requisito de almacenamiento se cumple si el espacio de almacenamiento de la estación i da cabida a $N_i - 1$ artículos, donde N_i se determina a partir de:

$$\sum_{ni=1}^{N_i} p_{ni} = \sum_{ni=1}^{N_i} (1-\rho_i)\,\rho_i^{ni} \geq \beta, \qquad i = 1, 2, \cdots, 5$$

Al simplificar se obtiene:

$$N_i \geq \frac{\ln(1-\beta)}{\ln \rho_i} - 1, \qquad i = 1, 2, \cdots, 5$$

Solución

Se supone que $\alpha_i = 0.9$; por lo tanto

$$\lambda_1 = \lambda = 20$$
$$\lambda_2 = \alpha_1\,\lambda_1 = 20\,\alpha_1 = 18$$
$$\lambda_3 = \alpha_2\,\alpha_1\,\lambda_1 = 20\,\alpha_1\,\alpha_2 = 16.2$$
$$\lambda_4 = 20\,\alpha_1\,\alpha_2\,\alpha_3 = 14.58$$
$$\lambda_5 = 20\,\alpha_1\,\alpha_2\,\alpha_3\,\alpha_4 = 13.12$$

Como $\mu_i = \mu = 30$ por hora, se tiene

$\rho_1 = 0.67$ $\rho_2 = 0.6$ $\rho_3 = 0.54$ $\rho_4 = 0.486$ $\rho_5 = 0.437$

Si deseamos establecer almacenamiento para todos los artículos que llegan el 99 % del tiempo es decir $\beta = 0.99$, los límites sobre N_i se pueden determinar como

$$N_1 \geq 10.499\,(\approx 11), \qquad N_2 \geq 8, \qquad N_3 \geq 6.47\,(\approx 7),$$

$$N_4 \geq 5.38\,(\approx 6), \qquad N_5 \geq 4.57\,(\approx 5)$$

3.4 REDES DE COLAS[5]

Una red de colas es un conjunto de nodos interconectados por medio de arcos, cada uno de estos nodos está formado por un sistema de colas con unos o más servidores.

Estas colas están conectadas con líneas que operan de forma asíncrona y concurrente, es decir, no hay sincronismo entre entradas y salidas, y actúan simultáneamente.

Las colas pueden estar conectadas entre ellas en serie o en paralelo, donde el tráfico saliente de una cola es el tráfico entrante de la siguiente. También pueden aparecer bifurcaciones y fusiones de tráfico donde se divide el flujo de tráfico o se unen diversos flujos de tráfico.

Algunos ejemplos de redes de colas son redes de computadoras, líneas de producción en una fábrica, tráfico de vehículos en una ciudad. Un ejemplo más detallado es el siguiente:

5 https://estadistica.net/IO/7-7-TEORIA-COLAS.pdf

Asesoría empresarial como red de colas: Los clientes llegan y esperan a ser atendidos por el servicio de recepción, desde allí son derivados al servicio solicitado (contable, fiscal, etc.), allí esperan la cola correspondiente y una vez que son atendidos, tienen que hacer cola en un servicio de gestión de cobros. Para decidir a qué cola se dirige un cliente que acaba de salir de una cola hay dos tipos de criterios:

La teoría de redes de colas contempla dos modelos:

a) Redes cerradas: No entran nuevos clientes y los existentes nunca salen, esto es, el número de clientes es constante en el tiempo, como puede ser la reparación de máquinas.
b) Redes abiertas: Los clientes pueden entrar y salir del sistema.

Es decir, cada flujo entra en el sistema por un punto en un momento dado y, después de pasar por unas o más colas, sale del sistema.

Considerando el número de unidades constante, pueden ser:

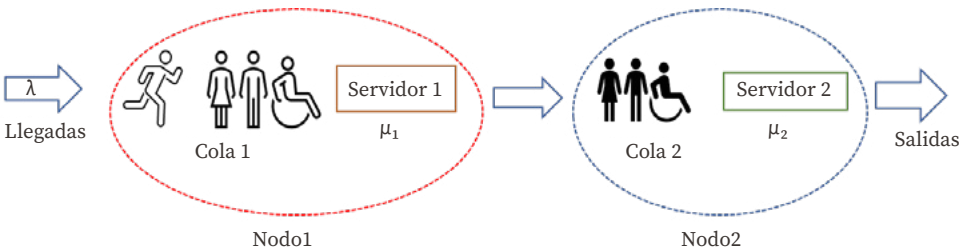
» Acíclicas: Un cliente nunca puede volver a la misma cola.
» Cíclicas: Cuando hay bucles en la red.

3.4.1 Sistema de colas en serie

En un sistema de colas en serie un cliente debe visitar diversos servidores antes de completar el servicio requerido. Se utiliza para casos en los que el cliente llega de acuerdo con un proceso de Poisson y el tiempo de atención se distribuye exponencialmente en cada estación. Esto ya lo vimos en la sección anterior de este libro, pero aquí se van a considerar diferentes tasas de servicio μ.

Se considera un ejemplo en el que los clientes llegan según un proceso de Poisson de parámetro λ, y pasan sucesivamente por dos colas en serie, respectivamente, con tasas de servicio μ_1 y μ_2 respectivamente.

Figura 3.4 Sistema de colas en serie



El número de clientes de cada uno de los servidores es independiente del otro.

- Los tiempos de espera de un cliente en cada cola no son independientes.
- Los tiempos totales de espera (cola + servicio) son independientes.

El estado del sistema es un par (n, m) con n clientes en el nodo 1 y m clientes en el nodo 2.

Las ecuaciones del balance o de equilibrio (tasa de entrada debe de ser igual a la de salida), para $n>0$, $m>0$, son:

Tabla 3.2
Ecuaciones de Balance

Estado	Tasa de entrada = Tasa de salida
(0,0)	$\mu_2 p_0,\ 1 = \lambda p_{0,0}$
$(n,0)$	$\lambda p_{n-1,0} + \mu_2 p_{n,1} = (\lambda + \mu_1) p_{n,0}$
$(0, m)$	$\mu_1 p_{1,m-1} + \mu_2 p_{0,m+1} = (\lambda + \mu_2) p_{0,m}$
(n, m)	$\lambda p_{n-1,m} + \mu_1 p_{n+1,m-1} + \mu_2 p_{n,m+1} = (\lambda + \mu_1 + \mu_2) p_{n,m}$

Con

$$\sum_{n,m} p_{n,m} = 1$$

Donde

$p_{n,0}$ = probabilidad de n clientes en el nodo 1

$p_{0,m}$ = probabilidad de m clientes en el nodo 2.

El nodo 1 es un modelo de cola M/M/1 y por el teorema de Burke, el nodo 2 también es un modelo de cola M/M/1. Por lo tanto,

$$p_{n,0} = \left(\frac{\lambda}{\mu_1}\right)^n \left(1 - \frac{\lambda}{\mu_1}\right)$$

$$p_{0,m} = \left(\frac{\lambda}{\mu_2}\right)^m \left(1 - \frac{\lambda}{\mu_2}\right)$$

Si los clientes en los nodos 1 y 2 son variables aleatorias independientes se verifica que: $p_{n,m} = p_{n,0} \cdot p_{0,m}$, propiedad que cumplen las ecuaciones de equilibrio.

En consecuencia, $p_{n,m} = (p_{n,0}).(p_{0,m})$, es la solución estacionaria y el número de clientes en el nodo 1 es independiente del número de clientes en el nodo 2, lo que no implica que los tiempos de espera de un cliente en las dos colas sean independientes. Sin embargo, los tiempos totales de espera (cola más servicio) son independientes.

Teorema de Burke

Una propiedad interesante de las colas M/M/1 que simplifica enormemente su combinación dentro de una red es el hecho de que la salida de una cola M/M/1 con una tasa de llegada λ es un proceso de Poisson de tasa de llegada λ.

En cualquier instante de tiempo t, el número de unidades que hay en el sistema es independiente de las salidas que ha habido antes de este instante. Se puede decir que el sistema es reversible.

Según el teorema de Burke, para un sistema de colas M/M/s/ ∞ si la capacidad de las colas es infinita, se puede estudiar cada una de ellas por separado.

Por lo tanto, la serie estará formada por k colas independientes.

La probabilidad de que en un instante haya n_1 unidades en la cola 1, n_2 unidades en la cola 2 ⋯ y n_k unidades en la cola k es:

$$p(n)=\sum_{i=1}^{k} p_i(n_i)$$

El teorema de Burke dice:

Para un sistema de colas con entrada de Poisson, una sola línea de espera sin deserciones, con los tiempos de servicio exponenciales e idénticamente distribuidos independientes (negativos), la distribución de equilibrio del número de terminaciones de servicios en un intervalo de tiempo arbitrario se demuestra que es la misma que la distribución de entrada, para cualquier número de servidores.[6]

El número medio de clientes en la red en serie o secuencial:

$$L_{red}=\sum_{n,m}(n+m)\,p_{n,m}=\sum_{n} n\,p_{n,0}+\sum_{m} m\,p_{0,m}=$$

$$=\frac{\lambda}{\mu_1-\lambda}+\frac{\lambda}{\mu_2-\lambda}=\sum_{i=1}^{2}\frac{\lambda}{\mu_i-\lambda}$$

6 The Output of a Queuing System Operations. Research. Vol 4, No 6. 1956

El tiempo medio de un cliente en la red (desde que entra hasta que sale):

$$W_{red} = \frac{L_{red}}{\lambda_{red}}$$

El tiempo medio de un cliente en la cola es:

$$W_q = W_{red} - \left(\frac{1}{\mu_1} + \frac{1}{\mu_2} \right)$$

Para ilustrar todos estos conceptos se resuelve un sistema de colas en el siguiente ejemplo:

Ejemplo 3.6

Un restaurante de autoservicio dispone de tres empleados, un mesero sirve el primer plato, el segundo mesero sirve el segundo plato y el tercero se encarga de la caja.

El primer mesero dispone de suficiente espacio para atender a clientes sin limitación, mientras que los otros dos empleados tienen un espacio limitado a dos personas como máximo. El autoservicio, modelado como red, muestra que la tasa media de llegada a la hora de la comida es de 54 clientes/hora, el primer mesero tiene un tiempo medio de servicio de un minuto y el segundo de treinta segundos. Se quiere encontrar:

a) El valor máximo del tiempo de servicio del tercer empleado para que su trabajo no interrumpa al de sus compañeros.
b) La longitud de las colas que forman el sistema.
c) El tiempo medio que un cliente pasa en el autoservicio desde que llega hasta que sale dispuesto para comer.

d) Número promedio de clientes en la cola.

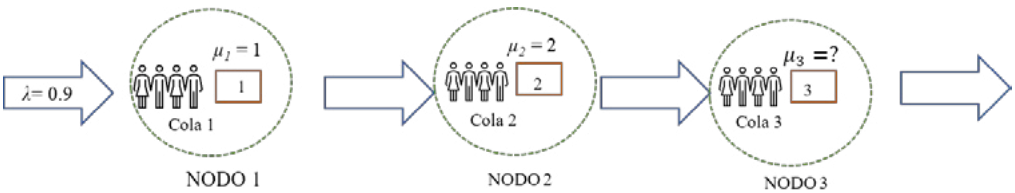
e) La longitud de las colas que se forman en el sistema para cada nodo.

f) El tiempo promedio de estancia en el sistema para cada nodo.

Solución

Es un modelo de red de colas en serie, con tres nodos (subsistemas), cada uno es un modelo de cola M/M/1.

Figura 3.5 El sistema de 3 colas en serie



$\lambda = 54$ clientes/ hora $= 54/60 = 0.9$ clientes/minuto

$1/\mu_1 = 1$ minuto, $\rightarrow \mu_1 = 1$ minuto

$1/\mu_2 = 30$ segundos $= 30/60 = 0.5$ minutos $\rightarrow \mu_2 = 2$ minutos

El factor de utilización o intensidad de tráfico $\rho = \lambda/\mu < 1$ para que la red no se sature y el estado sea estacionario.

$$\rho_1 = \frac{\lambda}{\mu_1} = \frac{0.9}{1} = 0.9 \qquad \rho_2 = \frac{\lambda}{\mu_2} = \frac{0.9}{2} = 0.45$$

a) Número promedio de clientes en la cola:

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{\rho^2}{(1-\rho)}$$

Número máximo de clientes en la cola para el nodo 3:

$$L_{q3} = \frac{\rho_3^2}{(1-\rho_3)} = 2$$

Se resuelve la ecuación cuadrática para encontrar el valor de $\rho 3$

$$\rho_3^2 + 2\rho_3 - 2 = 0 \rightarrow \rho_3 = 0.732 \text{ y } \rho_3 = -2.732$$

Por lo tanto, la intensidad de tráfico del nodo 3 es la siguiente:

$$\rho_3 = \frac{0.9}{\mu_3} \rightarrow \mu_3 = \frac{0.9}{0.732} = 1.2295 \text{ minutos}$$

b) La longitud de las colas que se forman en el sistema para cada nodo:

$$L_{q1} = \frac{\rho_1^2}{(1-\rho_1)} = \frac{0.9^2}{(1-0.9)} = 8.1 \text{ clientes}$$

$$L_{q2} = \frac{\rho_2^2}{(1-\rho_2)} = \frac{0.45^2}{(1-0.45)} = 0.37 \text{ clientes}$$

$$L_{q3} = \frac{\rho_3^2}{(1-\rho_3)} = 2 \text{ clientes}$$

c) El tiempo promedio de estancia en el sistema para cada nodo:

$$W_{si} = \frac{1}{\mu_i - \lambda}$$

$$W_{s1} = \frac{1}{1-0.9} = 10 \text{ minutos} \qquad W_{s2} = \frac{1}{2-0.9} = 0.9091 \text{ minutos}$$

$$W_{s3} = \frac{1}{1.223-0.9} = 3.035 \text{ minutos}$$

$$W_{si} = \sum_{i=1}^{3} \frac{1}{\mu_i - \lambda} = 10 + 0.9091 + 3.035 = 13.9441 \text{ minutos}$$

$1/\mu_2 = 30$ segundos $= 30/60 = 0.5$ minutos $\rightarrow \mu_2 = 2$ minutos

El factor de utilización o intensidad de tráfico $\rho = \lambda/\mu < 1$ para que la red no se sature y el estado sea estacionario.

$$\rho_1 = \frac{\lambda}{\mu_1} = \frac{0.9}{1} = 0.9 \qquad \rho_2 = \frac{\lambda}{\mu_2} = \frac{0.9}{2} = 0.45$$

d) Número promedio de clientes en la cola:

$$L_q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{\rho^2}{(1-\rho)}$$

Número máximo de clientes en la cola para el nodo 3:

$$L_{q3} = \frac{\rho_3^2}{(1-\rho_3)} = 2$$

Se resuelve la ecuación cuadrática para encontrar el valor de ρ_3

$$\rho_3^2 + 2\rho_3 - 2 = 0 \rightarrow \rho_3 = 0.732 \text{ y } \rho_3 = -2.732$$

Por lo tanto, la intensidad de tráfico del nodo 3 es la siguiente:

$$\rho_3 = \frac{0.9}{\mu_3} \rightarrow \mu_3 = \frac{0.9}{0.732} = 1.2295 \text{ minutos}$$

e) La longitud de las colas que se forman en el sistema para cada nodo:

$$L_{q1} = \frac{\rho_1^2}{(1-\rho_1)} \rightarrow \mu_3 = \frac{0.9^2}{(1-0.9)} = 8.1 \text{ clientes}$$

$$L_{q2} = \frac{\rho_2^2}{(1-\rho_2)} = \frac{0.45^2}{(1-0.45)} = 0.37 \text{ clientes}$$

$$L_{q3} = \frac{\rho_3^2}{(1-\rho_3)} = 2 \text{ clientes}$$

f) El tiempo promedio de estancia en el sistema para cada nodo:

$$W_{si} = \frac{1}{\mu_i - \lambda}$$

$$W_{s1} = \frac{1}{1-0.9} = 10 \text{ minutos} \qquad W_{s2} = \frac{1}{2-0.9} = 0.9091 \text{ minutos}$$

$$W_{s3} = \frac{1}{1.223-0.9} = 3.035 \text{ minutos}$$

$$W_{si} = \sum_{i=1}^{3} \frac{1}{\mu_i - \lambda} = 10 + 0.9091 + 3.035 = 13.9441 \text{ minutos}$$

3.5 REDES DE JACKSON[7]

Una red de Jackson consta de J nodos (o estaciones), cada uno con uno o varios servidores. Los tiempos de procesamiento de los trabajos en cada nodo son independientes e idénticamente distribuidos, siguiendo una distribución exponencial con media unitaria.

El costo del servicio, es decir, el costo por el cual se agota el trabajo, en cada nodo puede ser dependiente del nodo y dependiente del estado. Específicamente, siempre que haya trabajos x_i en el nodo i, la tasa de procesamiento es $\mu_i(x_i)$.

Los trabajos viajan entre los nodos siguiendo una matriz de enrutamiento $P := (p_{ij})$, donde $i, j = 1, \cdots, J$, y p_{ij} es la probabilidad de que un trabajo que salga del nodo i irá al nodo j.

En una red de Jackson todos los trabajos en cada nodo pertenecen a una sola "clase": todos los trabajos siguen la misma distribución del tiempo de servicio y el mismo mecanismo de enrutamiento; en este sentido, la red Jackson es un modelo de una sola clase. En consecuencia, no existe una noción de prioridad en el desempeño de los trabajos: en cada nodo, todos los trabajos se atienden por orden de llegada.

De acuerdo con las especificaciones de la matriz de enrutamiento, existen 3 variaciones diferentes para las redes de Jackson: red abierta, cerrada y semiabierta. Para efectos de este capítulo solamente se presenta un ejemplo de una red de Jackson abierta y de qué tratan las cerradas y semiabiertas, ya que el objetivo del libro es solamente introducir al lector a la teoría de colas.

7 H. Chen et al., Fundamentals of Queueing Networks © Springer Science+Business Media New York 2001

3.5.1 Redes de Jackson abiertas[8]

Son redes con k nodos que contemplan la posibilidad de entrada de clientes desde el exterior.

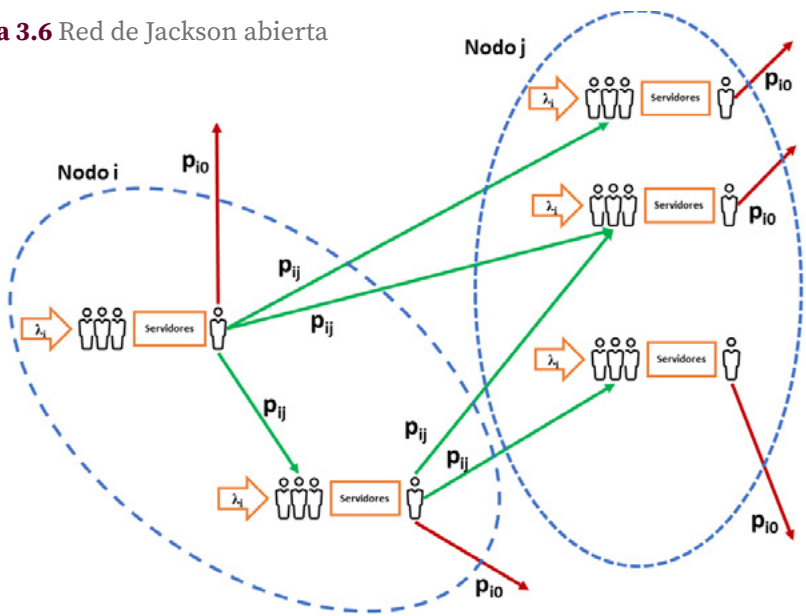
Las redes abiertas verifican tres propiedades:

a) La llegada de clientes al nodo i desde fuera del sistema sigue un proceso de Poisson de parámetro o tasa λ_i. También pueden llegar clientes al nodo i desde otros nodos de dentro de la red.

b) Cada nodo i consiste en si servidores, cada uno con tiempo de servicio exponencial de parámetro μ_i

c) El cliente una vez servido en el nodo i pasa (instantáneamente) al nodo j con $j = 1, 2, \cdots, k$ con probabilidad p_{ij}, o abandona la red con probabilidad p_{i0}

Figura 3.6 Red de Jackson abierta



8 https://estadistica.net/IO/7-7-TEORIA-COLAS.pdf

Dado que el flujo total de entrada a un nodo i ($i = 1, 2, \cdots, k$) debe ser igual al flujo total de salida del nodo, se obtienen las ecuaciones de equilibrio:

$$\Lambda_i = \lambda_i + \sum_{j=1}^{k} \Lambda_j \rho_{ji}$$

$$\begin{pmatrix} \textit{llegadas al nodo } i \\ \textit{fuera y dentro del sistema} \end{pmatrix} = \begin{pmatrix} \textit{llegadas al nodo } i \\ \textit{desde fuera del sistema} \end{pmatrix} + \begin{pmatrix} \textit{llegadas al nodo } i \\ \textit{desde dentro del sistema} \end{pmatrix}$$

Las ecuaciones de los Λ_i son sencillas:
$\Lambda_i \equiv$ Tasa de llegadas al nodo i desde fuera y dentro del sistema
$\lambda_i \equiv$ Tasa de llegadas al nodo i desde fuera del sistema
$\Lambda_j p_{ji} \equiv$ Tasa de llegadas al nodo i que salen del nodo j

Las k ecuaciones anteriores forman un sistema lineal con solución única, que se resuelve para hallar las tasas de llegada a cada nodo Λ_i.

En forma matricial se tiene:

$$\left(\vec{\Lambda} = \vec{\lambda} + \vec{\Lambda} p\right): \begin{pmatrix} \Lambda_1 \\ \Lambda_2 \\ \cdots \\ \Lambda_k \end{pmatrix} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \cdots \\ \lambda_k \end{pmatrix} + \begin{pmatrix} p_{11} & p_{12} & p_{1k} \\ p_{11} & p_{12} & p_{1k} \\ \cdots & \cdots & \cdots \\ p_{11} & p_{12} & p_{1k} \end{pmatrix} \begin{pmatrix} \Lambda_1 \\ \Lambda_2 \\ \cdots \\ \Lambda_k \end{pmatrix}$$

La solución $\left(\vec{\Lambda} = \vec{\lambda} + \vec{\Lambda} p\right)$ proporciona las tasas totales de llegada a cada subsistema (venga de fuera o de otro nodo).

El teorema de Jackson indica que las redes con realimentación son tales que los nodos se comportan como si fueran alimentados totalmente por llegadas de Poisson, aunque en realidad, no sea así.

Las probabilidades estacionarias en cada nodo son las de un modelo M/M/s, incluso aunque el modelo no sea un modelo M/M/s. Los estados n_i de los nodos individuales son variables aleatorias independientes.

Para que ninguna de las colas del sistema se sature, es preciso que se cumpla la condición:

$$\rho_i = \frac{\Lambda_i}{s_i\,\mu_i}\,;\ \rho_i < 1\ \forall_i = 1,\,2,\cdots,\,k$$

Que es la condición de no saturación del modelo M/M/s, aplicada a cada uno de los nodos por separado.

La probabilidad de que en el estado estacionario haya n_1 clientes en el nodo 1, n_2 clientes en el nodo 2, y así sucesivamente.

$$P_{n1n2\cdots nr} = \prod_{i=1}^{k} \frac{{p_i}^{ni}}{a_i(n_i)}\,P_{0i}\,; \qquad p_i = \frac{\Lambda_i}{\mu_i}\,; \qquad a(n_i) = \begin{cases} n_i! & n_i < s_i \\ {s_i}^{(n_i - s_i)}\,s_i & n_i \geq s_i \end{cases}$$

Concretamente si $s_i = 1$ para toda $i = 1,\,2,\cdots,\,k$

Las medidas de desempeño para cada nodo se calculan según las ecuaciones del modelo M/M/s, teniendo las siguientes consideraciones: En una red Jackson abierta que cumple la condición de no saturación, en estado estacionario, la distribución del número de clientes en cada nodo es:

$$P(n) = \prod_{i=1}^{k} P_i(n_i) \qquad \forall\, n_1,\cdots,n_k \geq 0$$

$P_i(n_i) \equiv$ Probabilidad de que haya n_i clientes en el nodo i

$\lambda_{red} = \sum_{i=1}^{k} \equiv$ Número de llegadas que entran en la red por unidad de tiempo desde fuera del sistema.

$\Lambda_{red} \equiv$ Tasa global de salidas del sistema, número promedio de clientes que salen del sistema por unidad de tiempo, que coincide con el número de clientes que entran desde dentro del sistema: $\lambda_{red} = \sum_{i=1}^{k} \Lambda_i$

$L_{red} \equiv$ Número medio de clientes en el sistema (cola + servicio), suma del número medio de clientes en cada uno de los nodos: $L_{red} = \sum_{i=1}^{k} L_{si}$

El hecho de que los nodos se comporten como si fueran modelo M/M/s podría interpretarse como el poder usar las distribuciones de los tiempos de espera de estos modelos. Sin embargo, esto no es necesariamente cierto en las redes de Jackson, en donde se permite la retroalimentación.

$W_{red} \equiv$ Tiempo medio en el sistema, tiempo medio que un cliente pasa desde que entra en la red hasta que sale de ella: $W_{red} = \dfrac{L_{red}}{\Lambda_{red}}$

$V_i \equiv$ Número medio que un cliente visita el nodo i, número medio de veces que un cliente visita el nodo i desde que entra en la red hasta que sale: $V_i = \dfrac{\Lambda_i}{\Lambda_{red}} \quad \forall_i = 1,2,\cdots,k$

Los supuestos considerados son:

- Capacidad infinita en los nodos.
- Efecto bloqueo: Si un cliente ha finalizado su servicio en el nodo i y se dirige a un nodo j que está al máximo de su capacidad. El sistema se bloquea con tres posibilidades:

a) Las llegadas al nodo i se rechazan.
b) El cliente debe ir inmediatamente a otro nodo.
c) El cliente debe abandonar el sistema.

Medidas de desempeño de nodos para un modelo M/M/1

Factor de saturación del nodo i :

$$\rho_i = \frac{\Lambda_i}{\mu_i} < 1 \qquad i = 1,2,\cdots$$

Número medio de clientes en cola (nodo i):

$$L_{qi} = \frac{\Lambda_i^2}{\mu_i(\mu_i - \Lambda_i)}$$

Número medio de clientes en el sistema (nodo i):

$$L_{si} = \frac{\rho_i}{1-\rho_i} = \frac{\Lambda_i}{\mu_i - \Lambda_i}$$

Tiempo medio de espera en cola del nodo i:

$$W_{qi} = \frac{L_{qi}}{\Lambda_i} = \frac{\Lambda_i}{\mu_i(\mu_i - \Lambda_i)}$$

Tiempo medio de espera en cada subsistema para el nodo i:

$$W_{si} = W_{qi} + \frac{1}{\mu_i} \; ; \; W_{si} = \frac{L_{si}}{\Lambda_i} = \frac{1}{\mu_i - \Lambda_i}$$

Medidas de desempeño de nodos para un modelo M/M/s

Factor de saturación del nodo i:

$$\rho_i = \frac{\Lambda_i}{s_i\,\mu_i} \begin{cases} \lambda_i \equiv \text{tasa de llegadas de procesos al nodo } i \\ \Lambda i \equiv \text{tasa de procesos que salen del nodo } i \\ \quad (\text{tasas totales llegadas}) \end{cases}$$

Utilización promedio del nodo i: $u_{si} = \Lambda_i / \mu_i$

La probabilidad que ningún cliente se encuentre en el sistema de cola nodo i:

$$P_{0i} = \frac{1}{\sum_{n=0}^{si-1} \frac{\left(\frac{\Lambda_i}{\mu_i}\right)^n}{n!} + \frac{1}{s_i!}\left(\frac{\Lambda_i}{\mu_i}\right)^{si} \frac{1}{1-\rho_i}}$$

Número medio de clientes en la cola del nodo i:

$$L_{si} = L_{qi} + \frac{\Lambda_i}{\mu_i}\ ; \quad L_{si} = \Lambda_i\, W_{si}$$

Tasa total de llegadas desde el exterior: $\lambda_{red} = \sum_{i=1}^{k} \lambda_i$

Número medio de clientes en la red: $\lambda_{red} = \sum_{i=1}^{s} L_i$

Tasa global de salidas del sistema: $\Lambda_{red} = \sum_{i=1}^{k} \Lambda_i$

Tiempo promedio en la red: $W_{red} = \frac{L_{red}}{\Lambda_{red}}$

Número medio de clientes que visitan un nodo: $V_i = \frac{\Lambda_i}{\Lambda_{red}} \quad \forall\, i = 1, 2, \cdots, k$

RED CÍCLICA Y ACÍCLICA: Una red es acíclica si no contiene ciclos (lazos), en caso contrario es cíclica. En una red acíclica cada cliente tiene que visitar cada nodo a lo sumo una vez, es decir: $V_i \leq 1$ para toda $i = 1, 2, \cdots, k$.

E **Ejemplo 3.7**

Los servidores de dos terminales del aeropuerto de Madrid siguen una disciplina FIFO y un proceso de Poisson, reciben respectivamente 20 y 30 procesos de usuarios por minuto. El servidor de la primera terminal tiene capacidad para atender una media de cien procesos por minuto, mientras que cualquiera de los dos procesa-

dores del servidor de la segunda terminal puede atender a veinticinco procesos, con tiempo de procesado exponenciales.

Cuando un proceso está a punto de finalizar en el servidor de la segunda terminal crea un nuevo proceso hijo en el servidor de la primera terminal el 25 % de los casos, en otro caso termina totalmente su ejecución.

Por otra parte, los procesos que se encuentran a punto de finalizar en el servidor de la primera terminal crean un nuevo proceso en su servidor el 20 % de los casos, en caso contrario cuando terminan su ejecución envían otro proceso al servidor de la segunda terminal un 10 % de las veces.

Se necesita conocer:

a) El número medio de procesos en cada servidor.
b) Número medio que un proceso visita cada nodo.
c) Tiempo medio que tarda un proceso en la red.

Solución

a) Es una red de Jackson cíclica abierta con $K=2$ nodos.

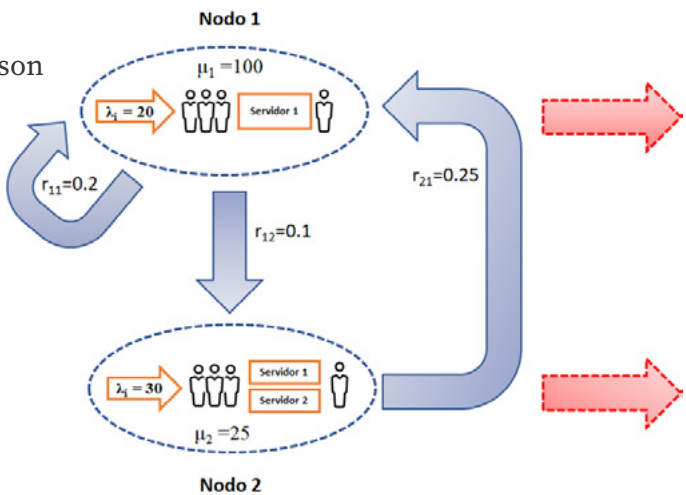


Figura 3.7 Red de Jackson abierta con $k=2$ nodos

Nodo 1 con un servidor $s_1 = 1$
Nodo 2 con dos servidores $s_2 = 2$
Tasas de llegada y servicio (procesos/ minuto) desde fuera del sistema son:

$$\lambda_1 = 20 \qquad \lambda_2 = 30 \qquad \mu_1 = 100 \qquad \mu_2 = 25$$

Ecuaciones de tráfico o ecuaciones de equilibrio:

$$\Lambda_i \equiv \lambda_i + \sum_{j=1}^{2} \Lambda_j \, p_{ji}$$

Λ_i ≡ Tasa de llegadas de procesos al nodo i desde fuera y dentro del sistema
λ_i ≡ Tasa de llegadas de procesos al nodo i desde fuera del sistema
Λ_j ≡ Tasa de procesos que salen del nodo j
$\Lambda_j \, p_{ji}$ ≡ Tasa de procesos que llegan al nodo i desde el nodo j

En forma matricial

$$\vec{\Lambda} = \vec{\lambda} + \vec{\Lambda} p$$

Explícitamente:

$$\begin{pmatrix} \Lambda_1 \\ \Lambda_2 \end{pmatrix} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} + \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} \begin{pmatrix} \Lambda_1 \\ \Lambda_2 \end{pmatrix}$$

Las probabilidades de transición de unos estados a otros se reflejan en la matriz:

$$p_{ij} = \begin{pmatrix} \square & T1 & T2 \\ T1 & 0.2 & 0.1 \\ T2 & 0.25 & 0 \end{pmatrix}$$

Considerando que $p_{ij}^{t} = p_{ji}$ y sustituyendo:

$$\begin{pmatrix} \Lambda_1 \\ \Lambda_2 \end{pmatrix} = \begin{pmatrix} 20 \\ 30 \end{pmatrix} + \begin{pmatrix} 0.2 & 0.25 \\ 0.1 & 0 \end{pmatrix} \begin{pmatrix} \Lambda_1 \\ \Lambda_2 \end{pmatrix}$$

Por lo tanto, se tiene un sistema de dos ecuaciones con dos incógnitas:

$$\Lambda_1 = 20 + 0.2\,\Lambda_1 + 0.25\,\Lambda_2$$
$$\Lambda_2 = 30 + 0.1\,\Lambda_1$$

Resolviendo el sistema:

$$\Lambda_1 = 35.484 \qquad \Lambda_2 = 33.548$$

Y $\lambda_{red} = \sum_{i=1}^{2} \lambda_i = 20 + 30 = 50$

En cada nodo el flujo que entra debe ser igual al flujo que sale. La tasa global de salidas del sistema coincide con el número de procesos que entran en el sistema:

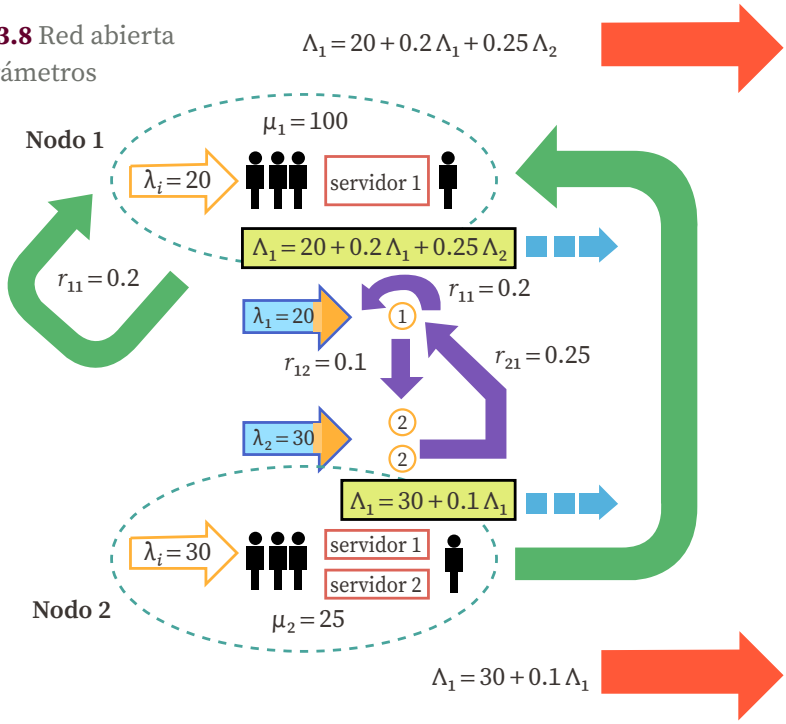
$$\Lambda_{red} = \sum_{i=1}^{2} \Lambda_i = 35.484 + 33.548 = 69.032$$

La condición de no saturación aplicada a cada uno de los nodos por separado es:

$$\rho_i = \frac{\Lambda_i}{s_i\,\mu_i}$$

$$\rho_i < 1 \qquad \forall\, i = 1, 2, \cdots \begin{cases} \lambda_i \equiv \text{Tasa de llegadas de procesos al nodo } i \\ \Lambda i \equiv \text{Tasa total de procesos que llegan al nodo } i \end{cases}$$

Figura 3.8 Red abierta con parámetros



$$\text{Nodo 1:}\quad \rho_1 = \frac{35.484}{100.1} = 0.35484 < 1 \qquad s_1 = 1 \text{ servidor}$$

$$\text{Nodo 2:}\quad \rho_2 = \frac{333.548}{25.21} = 0.67096 < 1 \qquad s_2 = 2 \text{ servidores}$$

Por lo tanto, ambos servidores son estacionarios.

» La terminal 1 es una cola tipo M/M/1

El número medio de procesos en el sistema (cola + servicio):

$$L_{si} = \frac{\rho_i}{1-\rho_i} = \frac{\Lambda_i}{\mu_i\,\Lambda_i}$$

$$L_{s1} = \frac{\rho_1}{1-\rho_1} = \frac{0.35484}{1-0.35484} = 0.55$$

Tiempo promedio de estancia en el sistema (cola + servicio):

$$W_{si} = \frac{L_{si}}{\Lambda_i} = \frac{1}{\mu_i - \Lambda_i}$$

$$W_{s1} = \frac{L_{s1}}{\Lambda_1} = \frac{0.55}{35.484} = 0.0155 \text{ minutos} = 0.93 \text{ segundos}$$

» La terminal 2 es una cola tipo M/M/s

La probabilidad de que ningún proceso se encuentre en el sistema de cola es:

$$P_{02} = \frac{1}{\sum_{n=0}^{1} \frac{\left(\frac{\Lambda_2}{\mu_2}\right)^n}{n!} + \frac{1}{s_{2!}} \left(\frac{\Lambda_2}{\mu_2}\right)^{s2} \frac{1}{1-\rho_2}} = \frac{1}{1 + \frac{33.548}{25} + \left(\frac{1}{2}\right)\left(\frac{33.548}{25}\right)^2 \left(\frac{1}{1-0.67096}\right)} = 0.1969167$$

El número medio de procesos en la cola de la terminal

$$L_{q2} = \frac{1}{s_{2!}} \left(\frac{\Lambda_2}{\mu_2}\right)^{s_2} \frac{\rho_2}{(1-\rho_2)^2} P_{02} = \frac{1}{2!} \left(\frac{33.548}{25}\right)^2 \frac{0.67096}{(1-0.67096)^2} 0.1969167 = 1.0988$$

a) Número medio de procesos en el sistema (cola + servicio):

$$L_{si} = L_{qi} + \frac{\Lambda_i}{\mu_i};$$

En consecuencia, $L_{s2} = L_{q2} + \frac{\Lambda_2}{\mu_2} = 1.0988 + \frac{33.548}{25} = 2.44072$

Número medio de procesos en la red:

$$L_{red} = \sum_{i=1}^{k} L_s = \sum_{i=1}^{2} L_i = 0.55 + 2.44072 = 2.9907$$

b) Número medio que un proceso visita cada nodo, desde que entra a la red hasta que sale:

$$V_i = \frac{\Lambda_i}{\Lambda_{red}} \qquad \forall\, i = 1, 2, \cdots, k$$

$$V_1 = \frac{\Lambda_1}{\Lambda_{red}} = \frac{35.484}{69.032} = 0.514 \text{ veces/min}$$

$$V_2 = \frac{\Lambda_2}{\Lambda_{red}} = \frac{33.548}{69.032} = 0.486 \text{ veces/min}$$

c) Tiempo medio de un proceso en la red (desde que entra hasta que sale):

$$W_{red} = \frac{L_{red}}{\Lambda_{red}} = \frac{2.9907}{69.032} = 0.04332 \text{ minutos} = 3 \text{ segundos}$$

3.5.2 Redes de Jackson cerradas[9]

En muchas aplicaciones, el número total de trabajos en la red se mantiene a un nivel constante, digamos N. Una vez que un trabajo completa todos sus requisitos de procesamiento y abandona la red, se libera inmediatamente un nuevo trabajo en la red (por supuesto, aquí está implícita la suposición de que existe una fuente infinita de nuevos trabajos, listos para ser lanzados a la red en cualquier momento). Conceptualmente, este tipo de operación también puede verse como si tuviera un número fijo de trabajos circulando

9 H. Chen *et al.*, Fundamentals of Queueing Networks © Springer Science+Business Media New York 2001

en la red, sin que ningún trabajo salga de la red y ningún trabajo externo ingrese a la red; y en este sentido, la red está "cerrada".

Ejemplos de este modo de operación incluyen sistemas de producción que mantienen un nivel constante de WIP (trabajo en proceso) o siguen una regla de control de inventario base (también conocida como política de reabastecimiento uno por uno). Además, es bastante común en muchos sistemas de fabricación automatizados que cada trabajo que se procesa debe montarse en un palet especialmente diseñado a lo largo de su circulación en el sistema; una vez que se completa un trabajo, se puede liberar otro trabajo nuevo en el sistema para ocupar el palet. Por lo tanto, el número total de palets determina naturalmente el nivel de WIP. En un sistema de computación multinivel, la constante N corresponde al nivel de computación.

En las redes de comunicación es bastante común utilizar un número fijo de "tokens" o fichas para lograr el control del flujo.

En una red cerrada no entran ni salen clientes, el número de clientes es constante en el tiempo.

- No es necesario que los buffers de espera sean infinitos solo que tengan capacidad suficiente para mantener (N– 1) clientes para que no haya bloqueo.
- El cliente al finalizar el proceso en el nodo i pasa al nodo j con probabilidad p_{ij}
- Todos los tiempos de servicio son exponenciales negativos μ_i y los clientes se procesan según el orden de llegada a un nodo.
- Cada nodo i es una cola M/M/s_i

3.5.3 Redes de Jackson semiabiertas[10]

El modelo semiabierto unifica las características de las redes tanto abiertas como cerradas.

Específicamente, la red está abierta, siguiendo las descripciones de la red en la sección 3.5.1, con la excepción de que el número total de trabajos en la red está limitado a K trabajos en cualquier momento. Nos referimos a K como el "límite de búfer". Cuando se alcance este límite, las llegadas externas se bloquean y se pierden. Dado que las llegadas externas siguen un proceso de Poisson, debido a la propiedad sin memoria de los tiempos entre llegadas, este mecanismo de bloqueo es equivalente a detener el proceso de llegada tan pronto como el límite del búfer se alcanza. El proceso de llegada se reanudará la próxima vez que un trabajo salga de la red, lo que reduce el número total de trabajos en la red a K–l.

Resulta que este modelo semiabierto se puede reducir a una red cerrada con una constante de K trabajos y $J+1$ nodos. El nodo adicional, indexado como nodo 0, nuevamente representa el mundo externo, con enrutamiento desde y hacia el nodo 0 siguiendo las probabilidades P_{0j} y P_{i0}, al igual que en la red abierta.

Además, el nodo 0 se considera una función de "servicio", con tasa de servicio $\mu_0(n) = \alpha$ para todo $n \geq 1$, y $\mu_0(0) = 0$ (con la definición de que α es la tasa del proceso de llegada de Poisson externo). De esta manera, el nodo 0 efectivamente genera el proceso de llegadas de la red semiabierta original. En particular, $x_0 = 0$ significa que hay K trabajos en los otros J nodos (que constituyen la red original), y

10 H. Chen et al., Fundamentals of Queueing Networks © Springer Science+Business Media New York 2001

por lo tanto $\mu_0(0) = 0$ captura correctamente el bloqueo de llegadas externas cuando el búfer está lleno.

3.6 RESUMEN

Modelos de colas que no obedecen a una distribución Poisson. Son modelos de colas que sin seguir una distribución Poisson tienen resultados analíticos, concretamente el modelo (M/G/1):(DG/∞/∞), donde se tiene un tiempo de servicio general con media $E\{t\}$ y varianza $var\{t\}$.

Sea λ la tasa de llegadas a una instalación con un servidor, dadas la media $E\{t\}$ y la varianza $var\{t\}$ de la distribución del tiempo de servicio, y se demuestra usando cadenas de Markov el uso de la fórmula que se denomina la fórmula de **Pollaczek-Khintchine.**

Modelos de colas con prioridad de servicio. Los modelos de colas con prioridad suponen varias líneas en paralelo incluyendo clientes que pertenecen a cierto orden de prioridad. Si la instalación tiene m filas, suponemos que la fila 1 tiene la más alta prioridad de servicio y la línea de espera m incluye a clientes con la más baja prioridad.

Modelos de colas sucesivas o en serie. Se consideran líneas de espera de Poisson con estaciones de servicio dispuestas en serie, de manera que el cliente debe pasar por todas las estaciones antes de completar su servicio.

3.7 Notas históricas

Colas generales. William (Vilim) Feller, cuyo nombre original era Vilibald Srećko Feller (7 de julio de 1906 – 14 de enero de 1970), fue un matemático estadounidense de origen croata conocido por sus contribuciones a la teoría de la probabilidad. Numerosos resultados en teorías de la probabilidad están asociados a él, como los procesos de Feller, el test de explosión de Feller, el movimiento de Feller-Brown y el teorema de Lindberg-Feller. Sus libros han sido fundamentales para la popularización de la teoría de la probabilidad. Realizó contribuciones importantes en la teoría de la renovación, los teoremas tauberianos, paseos aleatorios, procesos de difusión y la ley del logaritmo iterado. Inició la publicación de la revista Mathematical Reviews. https://es.wikipedia.org/wiki/William_Feller

Su libro *Introducción a la teoría de probabilidades y sus aplicaciones*, que apareció en 1950 estudia problemas de congestión telefónica.

Colas con prioridad. En 1954, Alan Cobham publicó el primer trabajo sobre prioridades no absolutas, para ingreso de Poisson, tiempo de duración exponencial y uno o múltiples canales, donde plantea lo siguiente: Hay varias situaciones que ocurren comúnmente en las que la posición de una unidad o miembro de una línea de espera está determinada por una prioridad asignada a la unidad en lugar de su hora de llegada a la línea. Un ejemplo es la línea formada por mensajes en espera de transmisión a través de un canal de comunicación abarrotado en el que los mensajes urgentes pueden tener prioridad sobre los rutina-

rios (Tomado de: *Elementos de la teoría de colas,* Thomas L. Saaty, Ed. Aguilar, 1967).

Con el paso del tiempo, una unidad dada puede avanzar en la línea debido al servicio de unidades en el frente de la línea o puede retroceder debido a la llegada de unidades que tienen prioridades más altas. Aunque no proporciona una descripción completa de este proceso, el tiempo promedio transcurrido entre la llegada a la línea de una unidad de una prioridad determinada y su admisión a la instalación para su servicio es útil para evaluar el procedimiento mediante el cual se realizan las asignaciones de prioridad. Las expresiones para esta cantidad se derivan para dos casos: el sistema de un solo canal en el que los tiempos de servicio de la unidad se distribuyen arbitrariamente y el sistema de múltiples canales en el que los tiempos de servicio se distribuyen exponencialmente. En ambos casos se asume que las llegadas ocurren al azar. https://econpapers.repec.org/article/inmoropre/v_3a2_3ay_3a1954_3ai_3a1_3ap_3a70-76.htm.

James R. Jackson (1924-2011)

Colas en serie. En 1954, G.G. O´Brien estudió el problema de dos colas (dos fases en equilibrio, con ingreso de Poisson y el tiempo de duración exponencial) en serie, dando una expresión de la distribución de la longitud y de la esperanza del tiempo de espera. En el mismo año, James R. Jackson consideró el problema con ingreso limitado e ilimitado para dos y tres fases, dando la distribución de la longitud de la cola (Tomado de: *Elementos de la teoría de colas,* Thomas L. Saaty, Ed. Aguilar, 1967).

En la teoría de colas, una disciplina dentro de la teoría matemática de la probabilidad, una red de Jackson (a veces red Jacksoniana) es una clase de red de colas donde la distribución de equilibrio es particularmente simple de calcular ya que la red tiene una solución en forma de producto. Fue el primer desarrollo significativo en la teoría de redes de colas, y la generalización y aplicación de las ideas del teorema para buscar soluciones similares en forma de producto en otras redes ha sido objeto de mucha investigación, incluyendo ideas utilizadas en el desarrollo de Internet.

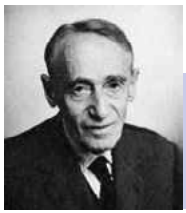
Las redes fueron identificadas por primera vez por James R. Jackson y su artículo fue reimpreso en los 'Diez títulos más influyentes de los primeros cincuenta años de las ciencias de la gestión' de la revista *Management Science*.

Jackson se inspiró en el trabajo de Burke y Reich, aunque Jean Walrand señala que "los resultados en forma de producto ... [son] un resultado mucho menos inmediato del teorema de salida de lo que el propio Jackson parecía creer en su artículo fundamental".

RRP Jackson encontró una solución anterior en forma de producto para colas en tándem (una cadena finita de colas donde cada cliente debe visitar cada cola en orden) y redes cíclicas (un bucle de colas donde cada cliente debe visitar cada cola en orden).

Una red Jackson consta de varios nodos, donde cada nodo representa una cola en la que la tasa de servicio puede ser dependiente del nodo (diferentes nodos tienen diferentes tasas de servicio) y dependiente del estado (las tasas de servicio cambian según la longitud de la cola). Los trabajos viajan entre los nodos siguiendo una matriz de enrutamiento fija. Todos los trabajos de cada nodo pertenecen a una única "clase" y los trabajos siguen la misma dis-

tribución del tiempo de servicio y el mismo mecanismo de enrutamiento. En consecuencia, no hay noción de prioridad en el servicio a los puestos de trabajo: todos los puestos de trabajo en cada nodo se sirven en un primer llegado, primer servido base.



Las redes de Jackson, donde una población finita de trabajos viaja alrededor de una red cerrada, también tienen una solución en forma de producto descrita por el teorema de Gordon-Newell.
https://en.wikipedia.org/wiki/Jackson_network.

Colas que no son de Poisson. En una cola que no es de Poisson, o el ingreso o el servicio, o ambos no son de Poisson, Khintchine dio una elegante deducción del número esperado de unidades y del valor esperado del tiempo de espera, para una cola ordenada en equilibrio, con llegadas Poisson y tiempo de duración arbitrario, que es la fórmula de Pllaczek-Khintchine.

Félix Pollaczek (1 de diciembre de 1892 en Viena - 29 de abril de 1981 en Boulogne-Billancourt) fue un ingeniero y matemático austriaco - francés, conocido por sus numerosas contribuciones a la teoría de números, análisis matemático, física matemática y teoría de probabilidades. Es más conocido por la fórmula de Pollaczek-Khinchine en la teoría de las colas (1930) y los polinomios de Pollaczek.

La fórmula de Pollaczek–Khinchine establece una relación entre la longitud de la cola y las transformadas de Laplace del tiempo de servicio para una M/G/1 (donde las llegadas siguen una distribución de Poisson y los tiempos de servicio una distribución general). Este término también se usa para referir a las relaciones entre la lon-

gitud media de cola y el tiempo medio de espera/servicio en dicho modelo.

La fórmula se publicó por primera vez por Felix Pollaczek en 1930 y fue readaptada en términos probabilísticos por Aleksandr Khinchin dos años después. En teoría de riesgo, la fórmula puede usarse para calcular la probabilidad de ruina final (probabilidad de que una compañía de seguros quiebre).

Aleksandr Yakovlevich Jinchin (en ruso: Алекса́ндр Я́ковлевич Хи́нчин, en francés: Alexandre Khintchine; 19 de julio de 1894 - 18 de noviembre de 1959) fue un matemático soviético, uno de los más importantes contribuyentes a la escuela soviética de la teoría de la probabilidad. https://en.m.wikipedia.org/wiki/Félix_Pollaczek

3.8 Ejercicios propuestos

1. En el ejemplo 3.1 suponga que la distribución del tiempo de servicio está dada de la siguiente manera:

a) Uniforme de t = de 5 a 15 minutos
b) Normal con media de 9 minutos y varianza de 4 minutos
c) Discreta con valores de 5, 10 y 15 minutos y probabilidades de ¼, ½ y ¼ respectivamente

2. Una línea de producción consta de dos estaciones. El producto debe pasar por las dos estaciones en serie. El tiempo que pasa el producto en la primera estación es constante e igual a 30 minutos. La segunda estación hace un ajuste (y cambios menores) y, por lo tanto, su tiempo dependerá de la condición del artículo cuando se reciba de la estación 1. Se calcula que el tiempo en

la estación 2 es uniforme entre 5 y 10 minutos. Los artículos se reciben en la estación 1 con un flujo de Poisson a la tasa de uno cada 40 minutos. Debido al tamaño de los artículos, no puede entrar una nueva unidad a la línea de producción sino hasta que salga de la estación 2 el que se encuentra ya en la instalación. Determine el número esperado de artículos en espera enfrente de la estación 1.

3. En una instalación de servicio la atención se ofrece en tres etapas consecutivas. El tiempo de servicio en cada etapa es exponencial con media de 10 minutos. Un nuevo cliente debe esperar hasta que el que está en servicio pasa por la etapa 3. Los clientes llegan a la etapa 1 de acuerdo con una distribución de Poisson con tasa media de uno por hora. Determine el número estimado de clientes en espera en la etapa 1.

4. Las órdenes de trabajo que llegan a una instalación de producción se dividen en tres grupos. El grupo 1 tendrá la más alta prioridad de procesamiento; el grupo 3 se procesará solo si no hay órdenes en espera de los grupos 1 y 2. Se supone que un trabajo una vez admitido en la instalación deberá completarse antes de que se inicie uno nuevo. Las solicitudes de los grupos 1, 2 y 3 ocurren según distribuciones de Poisson con medias 4, 3 y 2 por día, respectivamente. Los tiempos de servicio de los tres grupos son constantes con tasa de 10, 9 y 10 por día, respectivamente. Determine lo siguiente:

a) El tiempo de espera estimado en el sistema para cada una de las tres colas.
b) El número estimado de trabajos en espera en cada uno de los tres grupos.
c) El número estimado en espera en el sistema.

5. Una empresa de ITV (inspección y mantenimiento de vehículos) en una localidad dispone de una superficie que consta de tres partes: Una caseta donde los clientes entregan la documentación del vehículo y realizan el pago de tasas, con un espacio físico para un máximo de quince vehículos. Una nave formada por dos circuitos (equipamiento y personal técnico) para revisar los vehículos, con una tasa de servicio medio de 45 clientes/hora. Una oficina con dos puestos donde los conductores recogen la documentación y la ficha de la inspección técnica.

Acude a la nave una media de 57 clientes/hora, un mayor número de vehículos colapsaría el trabajo de la caseta, cuyo empleado atiende a un ritmo medio de 1 cliente/minuto; mientras que un oficinista tarda una media de 2 minutos/cliente.

Las llegadas siguen una Poisson y el tiempo de servicio exponencialmente. Se pide:

a) Longitud media de la cola de vehículos que habiendo pagado las tasas se encuentran esperando a la entrada de la nave.
b) Tiempo medio que un cliente pasa en la oficina.
c) Tiempo medio que un cliente se encuentra en la ITV
d) Para agilizar el proceso la empresa estudia la posibilidad de ampliar el número de servidores en la caseta o en la oficina. Suponiendo que el costo de ampliación en uno u otro lugar fuera equivalente, ¿qué criterio sería más acertado para que el tiempo de servicio del sistema fuera menor?

4 TOMA DE DECISIONES Y SIMULACIÓN

En este capítulo se presenta dos temas importantes en la teoría de colas que complementan la teoría vista en los capítulos anteriores, como son la elección de un modelo apropiado con los costos asociados y el uso de simulación.

Los modelos que se vieron en los capítulos 2 y 3 proporcionan resultados que describen el comportamiento de un número de situaciones que se presentan en las líneas de colas, sin embargo, no es inherente a la teoría de colas construir un modelo de decisiones o llevarlas a un proceso de simulación por lo que en este capítulo se abordan brevemente estos temas para complementar el análisis de colas.

4.1 ELECCIÓN DEL MODELO APROPIADO[11]

La elección del tipo de modelo por utilizar y las distribuciones apropiadas de probabilidad de las llegadas, servicios y salidas de un sistema de colas, son fundamentales en el análisis del sistema. Su determinación se realiza, cuando el sistema está funcionando, con base en un muestreo de las llegadas y las salidas a partir del cual se pueden obtener histogramas de frecuencia, ajustando posteriormente una distribución de probabilidad a estos y probando la bon-

11 *Apuntes de Teoría de la espera*. Antonio García Arana, DEPFI, UNAM, 1986.

dad de ajuste con una prueba χ^2, una Kolmogórov-Smirnov o una Anderson-Darling.

En el caso en que se esté diseñando un nuevo sistema de colas, analizando la naturaleza de la fuente de entrada y el mecanismo de servicio, es posible estimar las futuras distribuciones de entradas y salidas. Así, si las llegadas son aleatorias y futuras llegadas son independientes de llegadas pasadas entonces se puede pensar que siguen un proceso de Poisson.

Normalmente es difícil encontrar que la distribución de tiempos de servicio es exponencial debido a que dicha distribución tiene ciertas propiedades que la convierten en inapropiada para muchas situaciones de servicio. Una de estas propiedades es que la distribución exponencial tiene una función de densidad decreciente.

Lo anterior implica que el tiempo más probable de servicio es muy cercano a cero ($f(t)$ es máximo para $t=0$), contrario a la media $1/\mu$.

Una gran proporción de los problemas de colas comprenden la toma de una o más de las siguientes decisiones:

1. Número de servidores en un medio de servicio
2. Eficiencia de los servidores
3. Número de medios de servicio

Una mejor aproximación en muchos casos es la distribución de Erlang con un valor k muy grande, posiblemente $k=\infty$, de donde resulta un tiempo de servicio constante. Se pueden también tener casos donde la naturaleza del servicio es la misma pero el tipo y cantidad de servicio pueden diferir mucho, pudiendo tener la distribución de servicios en una forma acampanada.

Estimación de Parámetros

Si μ y λ son constantes fijas, λ se puede estimar contando el número de llegadas N en un intervalo de tiempo fijo t o viceversa, de la siguiente manera:

$$\hat{\lambda} = \frac{N}{t} \text{ estimador insesgado de } \lambda$$

De igual manera se puede estimar μ, con M como el número de unidades atendidas y B la suma de los tiempos ocupados de las unidades que dan servicio en un tiempo específico.

$$\hat{\mu} = \frac{M}{B} \text{ estimador insesgado de } \mu$$

Normalmente es preferible fijar el número de unidades solicitando servicio o en servicio en vez de fijar un tiempo de muestreo.

En el caso de varias estaciones de servicio puede ser conveniente trabajar con la suma de "m" tiempos de servicio para solo una de dichas estaciones y tomar el valor de μ resultante para el análisis posterior.

Si las entradas o salidas son poissonianas los estimadores de λ y μ son también de máxima verosimilitud.

4.2 Toma de decisiones

No es posible establecer una sola metodología para la toma de decisiones en problemas de espera, debido a la gran variedad de problemas que se presentan.

En general, los problemas de decisión que se presentan son el resultado de las combinaciones de las siguientes decisiones:

1. Número de unidades o estaciones de servicio en un mecanismo de servicio *S*
2. Eficiencia de las unidades o estaciones de servicio μ
3. Número de mecanismos de servicio λ

Las principales variables de decisión son S, μ y λ, respectivamente.

En la mayoría de los problemas interviene la determinación del nivel adecuado de servicio en un sistema de colas.

Esencialmente se cuenta con dos consideraciones para la toma de decisiones: Costo del servicio y el costo del tiempo de espera por servicio para un determinado nivel de servicio. Un correcto equilibrio entre las dos consideraciones conducirá a la mejor alternativa, aunque la última decisión dependerá de factores tales como: factores sicológicos por largas esperas (hospitales), costos muy elevados por espera (barcos), lo que se busca disminuir al mínimo el tiempo de espera, el impacto en las unidades que vienen para recibir servicio y tienen que esperar mucho tiempo por este, entre otros.

Normalmente el costo por tiempo de espera es difícil de estimar, por lo que un posible criterio sería fijar el tiempo de espera deseable.

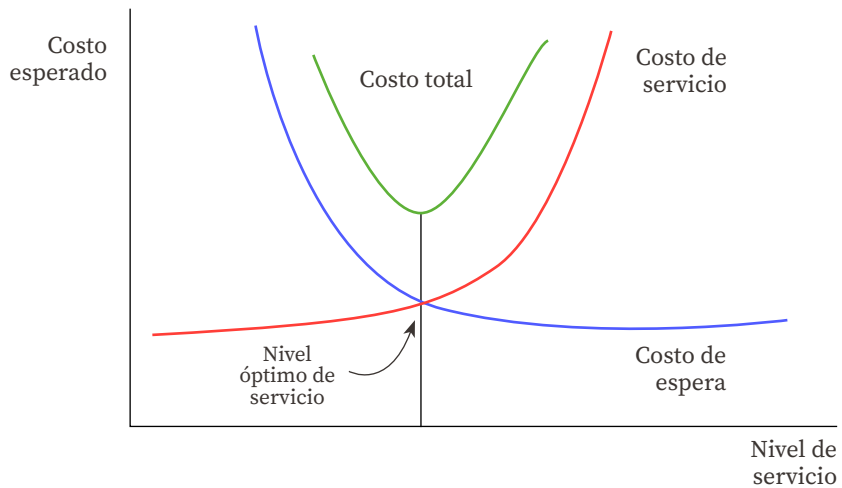
En la evaluación de los costos de tiempo de espera se podría pensar en el caso de los empleados, en el costo por unidad del empleado (sueldo) y el valor de la producción que se está dejando de obtener. En el caso de los clientes, el costo del tiempo de espera sería el costo de perder ventas ya sea de inmediato (el cliente que se desespera y se va)

o futuro (no volver a comprar en esa empresa por el tiempo exagerado de espera), aunque en este caso es difícil estimar el costo respectivo.

> *El objetivo principal es determinar el nivel de servicio que minimiza el costo total esperado de servicio y de espera por el servicio.*

Podemos resumir esto en la figura siguiente:

Figura 4.1 Costos asociados a la decisión de espera



4.3 Clasificación de los modelos de costos

Los modelos de costos como se muestra en la figura 4.1, básicamente equilibran los dos tipos de costo en conflicto:

1. Costo de servicio
2. Costo de espera en ofrecer el servicio (demora)

El primer costo refleja el punto de vista del servidor, en tanto que el segundo representa el del cliente. En este sentido abordamos primero la determinación de la tasa óptima de servicio en una instalación con un servidor, en el segundo caso la determinación óptima de servidores en paralelo, para ofrecer el servicio.

4.3.1 Tasa óptima de servicio (se desconoce μ)

Este modelo trata de un solo servidor donde se conoce la tasa λ de llegadas. Se desea determinar la tasa óptima de servicio μ con base en un modelo de costo apropiado. Sean:

CEO(μ) = costo estimado de operar la instalación por unidad de tiempo dada μ

CEE(μ) = costo estimado de espera por unidad de tiempo.

Se busca determinar el valor de μ que minimiza la suma de estos dos costos.

Las formas específicas para CEO y CEE como funciones de μ dependen del caso en estudio, por ejemplo, estas funciones pueden ser lineales, no lineales, continuas o discretas, dependiendo de las características de μ.

Ejemplo 4.1[12]

Una compañía impresora desea adquirir una copiadora comercial de alta velocidad para cubrir la creciente demanda en su servicio de copiado. La tabla siguiente presenta las especificaciones de diferentes modelos:

12 Taha. *Investigación de Operaciones*, Ed. Alfaomega, 2004.

Tabla 4.1 Tipo de copiadora, costos y velocidad

Tipo de copiadora	Costo de operación (h.)	Rapidez (hojas/min.)
1	15	30
2	20	36
3	24	50
4	27	66

Los pedidos llegan a la compañía de acuerdo con una distribución de Poisson, a razón de 4 cada 24 horas. La cantidad de cada pedido es aleatoria, pero se estima que en promedio es de 10 000 copias. Los contratos con los clientes estipulan una multa por entrega tardía de $ 80 por día y por pedido.

Solución

Usando un tamaño promedio por pedido de 10 000 copias, las tasas de servicio de las diferentes copiadoras se calculan de la siguiente manera, usaremos la copiadora 1 como ejemplo:

Para la copiadora 1

$$\text{tiempo promedio por pedido} = \frac{\text{pedido promedio} \times \text{día}}{\text{rapidez (hojas} \times \text{min.)}} \left(\frac{1}{60 \text{ min.}} \right) =$$

$$= \text{cantidad de horas por pedido}$$

$$\text{tiempo promedio por pedido} = \frac{10\,000}{36} \left(\frac{1}{60} \right) = 5.56 \text{ horas}$$

$$\text{tasa de servicio estimada } \hat{\mu}_1 = \frac{1 \text{ día (24 hrs.)}}{\text{tiempo promedio} \times \text{pedido}} =$$

$$= \text{cantidad de pedidos por día}$$

$$\text{tasa de servicio estimada } \hat{\mu}_1 = \frac{24}{5.56} = 4.32 \text{ pedidos} \times \text{día}$$

De la misma manera se calculan las tasas para las otras tres copiadoras y se muestran en la tabla 4.2.

Tabla 4.2. Tasas de servicio estimadas por tipo de copiadora

Tipo de copiadora	Tasa de servicio estimada μi (pedidos por día)
1	4.32
2	5.18
3	7.20
4	9.50

Un modelo de costo apropiado para la situación reconoce que μ se encuentra en cuatro valores discretos, correspondientes a los cuatro tipos de copiadoras. Esto nos indica que la tasa óptima de servicio se puede obtener comparando los costos totales correspondientes.

La determinación del costo total asociado con cada tipo de copiadora se calcula de la siguiente manera.

Tomamos un día (24 horas) como unidad de tiempo, entonces el costo de operar la instalación por día está dado por:

$$CEO_i = 24\, C_i \qquad \text{donde} \qquad i = 1, 2, 3, 4$$

C_i = costo de operación por hora de la copiadora i.

Por otra parte, el costo de espera por día considera la multa estipulada en los contratos de $80 por día por entrega tardía, este elemento de costo se expresa como:

$$CEE_i = 80\, L_{si}$$

L_{si} = número promedio de trabajos no terminados de la copiadora i.

La función de costo total queda entonces de la siguiente manera:

$$CET_i = 24\,C_i + 80\,L_{si}$$

Los valores de L_{si} corresponde al número esperado en el sistema para el tipo i, si se usan las fórmulas del modelo M/M/1, que recapitulando se tiene:

$$L_s = E\{n\} = \frac{\rho}{1-\rho}$$

Con $\rho = \lambda/\mu$, y se nos dice que $\lambda = 4$ la tasa de pedidos. En la tabla 4.3 se resumen los datos para cada copiadora.

Tabla 4.3 Cálculo de número promedio de trabajos no terminados de la copiadora *i*

Tipo de copiadora	λ_i	μ_i	L_{si}
1	4	4.32	13.50
2	4	5.18	4.39
3	4	7.20	2.25
4	4	9.50	1.73

Con esta parte de los cálculos se procede al cálculo general de CETi como se muestra en la tabla 4.4.

Tabla 4.4 Cálculo total de costos estimados por copiadora *i*

Tipo de copiadora	CEO_i	CEE_i	CET_i
1	360	1080.00	1440.00
2	480	351.20	831.20
3	**576**	**180.00**	**756.00**
4	648	138.16	786.16

De acuerdo con los cálculos la copiadora 3 tiene el menor costo total por día.

4.3.2 Número óptimo de servidores (se desconoce C)

El modelo anterior puede ampliarse para determinar el número óptimo de servidores en paralelo en una instalación. Si c es el número de servidores en paralelo, el problema se reduce en determinar el valor de c que minimiza

$$CET(c) = CEO(c) + CEE(c)$$

El valor óptimo de c debe satisfacer las siguientes condiciones necesarias:

$$CET(c-1) \geq CET(c) \qquad \text{y} \qquad CET(c+1) \geq CET(c)$$

Consideramos las funciones de costo asociadas al costo total de la siguiente manera:

$$CEO(c) = C_1(c)$$
$$CEE(c) = C_2\, L_s(c)$$

Donde
C_1 = costo por servidor adicional por unidad de tiempo
C_2 = costo por tiempo unitario de espera por cliente
$L_s(c)$ = número esperado de clientes en el sistema, dado c

La situación corresponde a un modelo M/M/C en el que se busca determinar el valor óptimo de C. Considerando el modelo de costos se tiene:

$$CET(c) = C_{1c} + C_2 L_s(c)$$

Ejemplo 4.2

En un almacén de herramientas, las solicitudes de intercambio de herramienta ocurren de acuerdo con una distribución Poisson con media de 17.5 solicitudes por hora. Cada empleado de la instalación puede manejar un promedio de 10 solicitudes por hora. El costo de incluir un nuevo empleado al almacén se estima en $ 12 por hora. El costo de la producción perdida por máquina en espera por hora se estima en $ 50 la hora. ¿Cuántos empleados se deben contratar para minimizar los costos?

Solución

Observe que $L_s(c)$ es una función del número de empleados s en paralelo. Se tiene además de acuerdo con la fórmula de costos lo siguiente:

$$CET(c) = 12c + 50\ L_s(c)$$

Se observa también que $L_s(1) = \infty$, ya que $\lambda > \mu$, donde $\lambda = 17.5$; $\mu = 10$ y $\rho = 1.75$

Para que el modelo alcance el estado estable se requiere que, $c > \lambda/\mu$, es decir $c > 2$. Se hace uso de la fórmula (del capítulo 2):

$$L_s = \frac{\lambda}{\mu} + \left[\frac{(\lambda/\mu)^c\,\lambda\mu}{(C-1)!\,(c\mu-\lambda)^2}\right] p_0$$

La probabilidad de que el sistema esté vacío al llegar un cliente es:

$$p_0 = \left\{\sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c}{c!\,(1-(\rho/c))}\right\}^{-1} \qquad \rho/c < 1$$

Los cálculos se efectúan de la siguiente manera, se hará para $c=2$:

$$p_0=\left[1+1.75+\frac{(1.75)^2}{2!}\left(\frac{1}{1-(1.75/2)}\right)\right]^{-1}=\left[1+1.75+1.531(8)\right]^{-1}=\frac{1}{15}$$

$$L_s(2)=\frac{17.5}{10}+\left[\frac{\left(\frac{17.5}{10}\right)^2 17.5(10)}{1!\left[2(10)-17.5\right]^2}\right]p_0=1.75+\left(\frac{535.93}{6.25}\right)p_0=$$

$$=1.75+85.75/15=7.467$$

Por otro lado, se tiene:

$$CET(2)=12(2)+50(7.467)=24+373.35=397.35$$

Se hacen los cálculos para $c=2,\cdots,6$ y se resumen en la tabla siguiente:

Tabla 4.5 Cálculo de número de empleados y costos

C servidores	L_s(c) solicitudes	CET (c) ($)
2	7.467	397.35
3	2.217	146.85
4	**1.842**	**140.10**
5	1.769	148.45
6	1.754	159.70

La cantidad que minimiza el costo y que satisface la demanda es de 4 servidores.

4.3.3 Tipo y cantidad de equipo para proporcionar un servicio (se desconocen S y μ)

En este modelo desconocemos el número de servidores, así como la tasa de servicio, se tiene entonces lo siguiente:

$F(\mu)$ = Costo marginal de una estación de servicio por unidad de tiempo para una relación de servicio μ

A = Conjunto de valores factibles de μ

Dados λ, $CEE(\mu)$, $F(\mu)$ y A encontrar μ y S. Por lo tanto, el objetivo es determinar el valor de μ que minimiza la suma de estos dos costos.

$$\text{Min } E(C) = S\,F(\mu) + CEE(\mu)$$

Si F(μ) es igual a CEO(μ) se tiene el caso del primer modelo. Así para un valor fijo de μ se puede utilizar el primer modelo; corriendo el modelo para todos los valores posibles de μ se escogería el de mínimo valor.

Ejemplo 2.3

Una organización desea ampliar las instalaciones de un puerto construyendo muelles dedicados a atender una flotilla de barcos que transportan mineral ferroso. Debe decidirse el número de muelles y el tipo de instalación en cada una, de tal manera que se minimicen los costos totales de descarga.

Se puede construir un máximo de 3 muelles y se requiere que todos los muelles nuevos tengan el mismo tipo de instalación de descarga. La información referente a los tipos de instalación disponibles se muestra en la siguiente tabla:

Tabla 4.6 Tipo de instalación con costos fijos y de operación y capacidades

Tipo de instalación	Costos diarios fijos	Costos diarios de operación	Capacidad de descarga diaria Tons.
A	840	840	3600
B	1350	1350	5800
C	1500	1600	6400

Los costos fijos incluyen conceptos tales como amortización de la instalación, mantenimiento general entre otros, siendo aplicables independientemente de que se utilice o no el equipo. Los costos de operación solo se presentan cuando se usan las instalaciones.

Los barcos transportan 8000 toneladas de mineral ferroso y arriban al puerto en forma poissoniana con una relación media de llegadas de 5 barcos por semana. Los tiempos de servicio de cada tipo de instalación se consideran exponenciales con una relación media de servicio definida por la capacidad de descarga correspondiente.

El tiempo medio de permanencia en el sistema representa un costo de $ 2000 por barco por día, esto es el tiempo de espera + el tiempo de descarga. ¿Qué tipo de instalación debería elegirse y cuántos muelles deberían construirse?

Solución

Existen 9 políticas posibles (tipo de instalación *A*, *B* o *C*) y número de lugares (1, 2, 3).

La combinación (*A*, 1) se puede descartar de forma inmediata, ya que la tasa media de llegadas es de $\lambda = 5$ por semana $= 5/7 = 0.714$ barcos por día, mientras que:
$\mu_A = 3600/8000 = 0.450$ barcos por día, lo que implica que $\lambda > \mu$

De la misma manera se calculan:

$\mu_B = 5800/8000 = 0.725$ barcos por día

$\mu_C = 6400/8000 = 0.800$ barcos por día

Calcular el costo esperado total por día para cada política incluye:

1. Costo fijo de las instalaciones
2. Costo de operación de las instalaciones
3. Costo de tiempo de espera del barco

Entonces el costo total está dado como sigue:

Costo total = costo fijo + costo de operación + costo de permanencia

En la tabla siguiente se muestran los costos fijos por día considerando número de lugares para los tres tipos de instalación, *A*, *B* y *C*.

Tabla 4.7 Costos fijos por día: tipo de instalación y número de lugares

Tipo de instalación	1	2	3
A	-	1680	2520
B	1350	2700	4050
C	1500	3000	4500

Ahora calculamos los factores de utilización del sistema: $\rho = \frac{\lambda}{s\mu} < 1$. Sabemos que la combinación (*A*, 1) no es factible, y como ejemplo calculamos para (*A*, 2):

$$\rho = \frac{\lambda}{s\mu} = \frac{0.714}{2(0.450)} = 0.793$$

Haciendo lo mismo para las demás combinaciones, se pueden ver resumidas en la tabla 4.8.

Tabla 4.8 Factores de utilización del sistema ρ

Tipo de instalación	1	2	3
A	-	0.793	0.529
B	0.985	0.493	0.328
C	0.892	0.448	0.297

Ahora procedemos a calcular los costos diarios de operación que consiste en multiplicar el número de lugares por el factor de utilización correspondiente por los costos diarios de operación dados en la tabla 4.6. Por ejemplo, para la opción $(A, 2)$ el cálculo es: $2\,(0.793)\,(840) = 1332$. En la tabla 4.9 se muestran los costos de operación para las 9 combinaciones.

Tabla 4.9 Costos diarios de operación

Tipo de instalación	1	2	3
A	-	1332	1333
B	1330	1331	1328
C	1427	1434	1426

Para calcular el costo de espera del barco usamos la fórmula para el tiempo medio de permanencia del sistema que involucra a p_0 que es la probabilidad de que el sistema esté vacío al llegar un cliente y cuya expresión es igual a:

$$p_0 = \left\{ \sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c}{c!\left(1-(\rho/c)\right)} \right\}^{-1} \qquad \rho/c < 1$$

Nuevamente, los cálculos para las nueve opciones de que el sistema esté vacío se resumen en la tabla 4.10

Tabla 4.10 Probabilidad de que el sistema esté vacío

Tipo de instalación	1	2	3
A	-	0.115	0.190
B	0.015	0.340	0.369
C	0.108	0.383	0.406

Para calcular el tiempo medio en el sistema nuevamente usamos la fórmula dada en el capítulo 2 para el modelo M/M/C. y está dada de la siguiente manera:

$$W_s = \frac{1}{\mu} + \left[\frac{(\lambda/\mu)^c \, \mu}{(C-1)! \, (c\mu - \lambda)^2} \right] p_0 = W_q + \frac{1}{\mu}$$

Con base en esto y la información que se tiene se presentan los cálculos en la tabla 4.11, nuevamente hacemos el cálculo para la combinación $(A, 2)$

Se tienen los datos siguientes: $p_0 = 0.115$, $\rho = 0.793$, $c = 2$, $\lambda = 0.714$, $\mu_A = 0.450$, por lo tanto

$$W_s = \frac{1}{0.450} + \left[\frac{(1.59)^2 \, (0.450)}{0.0346} \right] 0.115 = 5.96$$

Tabla 4.11 Tiempos medios de permanencia en el sistema

Tipo de instalación	1	2	3
A	-	5.96	2.48
B	90.90	1.82	1.43
C	11.60	1.56	1.29

Con toda esta información procedemos a calcular el costo de permanencia por día de los barcos por ejemplo para $(A; 2)$ se tiene $\lambda = 0.714$

por el costo de $2000 que es el costo del tiempo medio de permanencia, por el costo del tiempo medio de permanencia de esa opción como se muestra en la tabla 4.11 que es de 5.96, entonces se tiene

Costo de permanencia para $(A, 2) = 0.714\,(2000)\,(5.96) = 8511$.

Los costos de permanencia se muestran en la siguiente tabla 4.12.

Tabla 4.12 Costos de permanencia por día

Tipo de instalación	1	2	3
A	-	8511	3541
B	129 805	2599	2042
C	16 565	2228	1842

Finalmente se calculan los costos totales de descarga por día que es la suma de los costos dados en las tablas 4.7, 4.9, y 4.12 que son:

Costos totales de descarga = costos fijos por día +
costos diarios de operación + costos de permanencia por día

La tabla 4.13 resume estos cálculos:

Tabla 4.13 Costos totales de descarga por día

Tipo de instalación	1	2	3
A	-	11 523	7 393
B	132 485	6 629	7 422
C	19 492	6 655	7 769

La opción más económica es la construcción de dos muelles de tipo B.

4.4 Teoría de colas y uso de la simulación

Todos los modelos de colas vistos en los capítulos anteriores y en este son susceptibles de ser simulados, existe "software" que usa modelos de colas para simular un proceso. Sin embargo, es muy importante construir el modelo conceptual de la simulación antes de seleccionar algún "software" de simulación y no ser un simple usuario, para que de esta manera se entienda claramente lo que se requiere para efectuar de manera exitosa una simulación.

Como el objetivo de este escrito no es la simulación, nos limitaremos a presentar algunos conceptos y ejemplos que están relacionados con el uso de la teoría de colas.

Una simulación por computadora es un intento de modelar un proceso de un sistema real o hipotético por medio de un programa de computadora con el objetivo de observar, analizar y mejorar su comportamiento. En términos más prácticos, la simulación puede ser utilizada para pronosticar el comportamiento futuro de un sistema y determinar qué se puede hacer para influir en tal comportamiento. A fin de analizar, estudiar y mejorar algún sistema utilizando las técnicas de simulación digital, es necesario primero desarrollar un modelo conceptual que describa la dinámica de interés y, después, codificarlo en un simulador con el fin de analizar los resultados.[13]

Históricamente, la simulación es muy antigua ya que es inherente al proceso de aprendizaje de los seres humanos, como se observa en los juegos de los niños; los cuales se pueden considerar como una simulación del mundo real. Por otra parte, la simulación digital es reciente ya que para ser capaces de entender la realidad y toda

13 Idalia Flores De La Mota et al. *Robust Modelling and Simulation.* Ed. Springer, 2017.

la complejidad que un sistema puede implicar, ha sido necesario construir objetos artificiales y experimentar dinámicamente con ellos antes de interactuar con el sistema real. La simulación digital puede ser vista como el equivalente electrónico de este tipo de experimentación.[14]

En la mayoría de los casos, la simulación se usa cuando las alternativas matemáticas son pobres, es decir, es el "último recurso", algo así como: "cuando todo falle, use la simulación".

En realidad, si la solución analítica es relativamente sencilla, siempre será preferible a la simulación, ya que se considera el modelo general. Sin embargo, el problema es que existen muchos sistemas que no generan problemas sencillos de resolver, en este caso se recurre a la simulación. Por ejemplo, se tienen las colas o líneas de espera que involucran procesos aleatorios distribuidos en una serie de componentes del sistema: los modelos de inventarios, de recursos compartidos, de pronósticos de series de tiempo, de comportamientos económicos, de esquemas de producción, de movimiento de vehículos, de dinámicas de cruceros viales, etc.

Otra ventaja de la simulación es que se puede experimentar sin exponer a la organización a los perjuicios de errores en el mundo real. Por ejemplo, algunos bancos han estudiado el cambio de su sistema de filas múltiples a fila única empleando modelos de simulación y sin necesidad de experimentar con los clientes. La experimentación directa con los clientes podría tener consecuencias desagradables si no funciona como se espera.

14 Idalia Flores De La Mota et al. *Robust Modelling and Simulation*. Ed. Springer, 2017.

Por otro lado, es más sencillo controlar condiciones experimentales en un modelo de simulación que en un sistema real. Podemos pensar en un modelo de un crucero vial, donde puedan analizarse diferentes sincronizaciones de semáforos sin afectar a los elementos reales, lo cual tendría un costo excesivo que podría llegar hasta lo invaluable de una vida humana.

En un modelo de simulación es posible comprimir largos periodos de tiempo y analizar el comportamiento en forma inmediata. Podemos visualizar, por ejemplo, cómo será la población dentro de 30 años y si los servicios de transporte serán suficientes para satisfacer las necesidades de movilidad.

Por supuesto, hay casos en los que el sistema que se quiere analizar ni siquiera existe, de modo que, definitivamente, lo ideal será usar la simulación o algún método de tipo cualitativo. La simulación no remplaza 'per se' a otras formas de experimentación ni al juicio subjetivo, pero es una solución alternativa conveniente cuando el modelo es de complejidad elevada o el número de variables es muy grande. En cualquier caso, la experiencia y la intuición, así como el profundo conocimiento de los fenómenos, deberán ser ingredientes constantes para el éxito de los modelos de simulación.[15]

Ciclo de vida de un proyecto de simulación

Existe un consenso entre la gente involucrada en el desarrollo y mantenimiento de modelos de simulación en que los modelos simples son preferibles a los complejos. A pesar de ello, en muchos proyectos los modelos generalmente son grandes y complejos. Es nece-

15 Idalia Flores De La Mota et al. *Robust Modelling and Simulation.* Ed. Springer, 2017.

sario enfatizar que el exceso de complejidad en los modelos no solo tiene impacto en el rendimiento computacional, sino que también afecta en otros aspectos, como el tiempo necesario para el desarrollo del modelo, su mantenimiento, verificación y validación.

Aunque parece un concepto muy intuitivo, no existe una definición o medida de complejidad aceptada como estándar por los expertos del área. Algunos autores relacionan la complejidad del modelo, por ejemplo, con el "nivel de detalle" y, otros, con "la generalización del sistema". Algunas de las ventajas de trabajar con modelos simples son:

- Más fáciles de implementar, validar y analizar.
- Es más sencillo, en cierto modo menos "doloroso", descartar un modelo simple que tiene, por ejemplo, un error de diseño que un modelo complejo en donde se ha invertido un número considerable de horas de personal experto.
- Es más fácil adaptar un modelo simple que un modelo complejo si las condiciones o hipótesis de operación en el sistema real cambian.
- El tiempo de duración del ciclo de vida total del proyecto es generalmente menor.

Un proyecto de simulación es de naturaleza dinámica. Los resultados que se obtienen mientras se desarrollan exponen nuevos problemas, así como limitaciones inherentes al sistema estudiado. Esto puede obligar a reconsiderar la orientación inicial del proyecto. Además, la motivación del cliente también puede cambiar a lo largo del proyecto, como consecuencia de los resultados obtenidos o por factores externos al mismo. A fin de tener éxito en tal ambiente dinámico, es necesario utilizar una metodología correcta.

La tabla siguiente muestra las fases de un proyecto de simulación. Aunque pueda parecer que el desarrollo de un proceso de simulación es secuencial, en realidad no es así. Por ejemplo, si el modelo de simulación obtenido no aprueba la fase de validación (fase 5), es posible que sea necesario modificar el modelo conceptual, así como el modelo de simulación.[16]

Fase		Descripción
1	Formulación del problema	Definición del problema y ajuste de los objetivos.
2	Diseño del modelo conceptual	Especificación de los elementos del sistema y sus interacciones considerando los objetivos del problema.
3	Recolección de datos	Identificación, recolección y análisis de los datos necesarios para el estudio.
4	Construcción del modelo	Construcción del modelo de simulación basado en el modelo conceptual y en los datos recopilados.
5	Verificación y validación	Verificación de que el comportamiento del modelo concuerda con el modelo conceptual y los datos recopilados, comprobando que el modelo de simulación representa al sistema real.
6	Experimentación y análisis	Análisis de los resultados de la simulación con el propósito de detectar problemas en el sistema real y recomendar mejoras.
7	Documentación	Proveer de documentación sobre el estudio llevado a cabo.
8	Implantación	Poner en práctica las decisiones tomadas con el apoyo del estudio de simulación.

16 Idalia Flores De La Mota et al. *Robust Modelling and Simulation*, Ed. Springer, 2017.

Modelos continuos versus modelos discretos

Los *modelos continuos* se caracterizan por presentar la evolución de las variables de interés de manera continua. En general, se usan ecuaciones diferenciales ordinarias para modelar la evolución de una variable con respecto al tiempo, o ecuaciones diferenciales parciales para modelar también una evolución de la variable, pero con respecto al espacio.

De forma análoga a la definición de modelos continuos, los *modelos discretos* se caracterizan por representar la evolución de las variables de interés de manera discreta.

Es importante tener en cuenta que, a partir de la clasificación anterior de modelos, es posible describir un sistema continuo por medio de un modelo discreto y viceversa. La decisión de utilizar un modelo continuo o uno discreto depende de los objetivos particulares para cada estudio, y no tanto de las características del modelo. Así, por ejemplo, es posible encontrar modelos de flujo de automóviles en una autopista, donde ha sido elegida una formulación continua debido a que los objetivos de estudio están centrados, por ejemplo, en evaluar la evolución del tráfico con la presencia de un accidente, donde el movimiento de un automóvil en particular carece de importancia.

Los modelos de *eventos discretos* son dinámicos, estocásticos y discretos, en donde las variables de estado cambian su valor en instantes no periódicos de tiempo, sin estar dirigidos por un reloj. Estos instantes corresponden con la ocurrencia de un evento. Así, un evento se define como la acción instantánea que puede cambiar el estado de un modelo.[17]

17 Idalia Flores De La Mota et al. *Robust Modelling and Simulation*, Ed. Springer, 2017.

Ejemplo 4.4

El tiempo entre llegadas de clientes a Barbería Pérez tiene distribución exponencial con una media de 15 minutos. En el local hay un solo peluquero, y tarda de 10 a 15 minutos, con una distribución uniforme, para terminar un corte de pelo. A los clientes se les atiende con el sistema PEPS o FIFO (primero en llegar, primero en salir). El objetivo de la simulación es calcular las siguientes medidas de desempeño:

- La utilización promedio del local
- La cantidad promedio de clientes en espera
- El tiempo de espera promedio de un cliente en la cola

La lógica del modelo de simulación se puede escribir en términos de las acciones asociadas con sus eventos de llegada y de salida.

Solución

Evento de llegada

1. Genere y guarde cronológicamente la hora de llegada del siguiente cliente (= hora de simulación actual + tiempo entre llegada).
2. Si la instalación (el peluquero) está inactiva.
3. Inicie el servicio y declare ocupada la instalación. Actualice las estadísticas de utilización de la instalación.
4. Genere y guarde cronológicamente la hora de salida del cliente (= hora de simulación actual + tiempo de servicio).
5. Si la instalación está ocupada, ponga al cliente en la línea de espera y actualice las estadísticas de la cola.

Evento de salida

1. Si la cola está vacía, declare inactiva a la instalación. Actualice las estadísticas de utilización de la instalación.
2. Si la cola no está vacía.

Seleccione un cliente de la cola y colóquelo en la instalación. Actualice las estadísticas de la cola y de utilización de la instalación.

Genere y guarde cronológicamente la hora de salida del cliente (= hora de simulación actual + tiempo de servicio).

Según los datos del programa, el tiempo entre llegadas tiene distribución exponencial con una media de 15 minutos, y el tiempo de servicio tiene distribución uniforme entre 10 y 15 minutos. Si p y q representan muestras aleatorias de tiempo entre llegadas y de servicio, entonces, como se obtiene:

$$p = -15 \ln(R) \text{ minutos}, \qquad 0 \le R \le 1$$
$$q = 10 + 5R \text{ minutos}, \qquad 0 \le R \le 1$$

Para fines de este ejemplo usaremos R de la siguiente tabla de números aleatorios, comenzando con la columna 1. También usaremos el símbolo T para representar la simulación de la hora indicada en el reloj. Además, supondremos que el primer cliente llega cuando $T = 0$, y que la instalación comienza vacía.

Como los cálculos de simulación suelen ser voluminosos y tediosos, la simulación se limitará a las primeras 5 llegadas.

Tabla de números aleatorios

0.0589	0.3529	0.5869	0.3455	0.7900	0.6307
0.6733	0.3646	0.1281	0.4871	0.7698	0.2346
0.4799	0.7676	0.2867	0.8111	0.2871	0.4220
0.9486	0.8931	0.8216	0.8912	0.9534	0.6991
0.6139	0.3919	0.8261	0.4291	0.1394	0.9745
0.5933	0.7876	0.3866	0.2302	0.9025	0.3428
0.9341	0.5199	0.7125	0.5954	0.1605	0.6037
0.1782	0.6358	0.2108	0.5423	0.3567	0.2569
0.3473	0.7472	0.3575	0.4208	0.3070	0.0546
0.5644	0.8954	0.2926	0.6975	0.5513	0.0305

Llegada del cliente 1 cuando $T = 0$

Generar la llegada del cliente 2 a los

$$T = 0 + p_1 = 0 + \left[-15 \ln \left(0.0589 \right) \right] = 42.48 \text{ minutos.}$$

Como la instalación está inactiva cuando $T = 0$, el cliente 1 inicia el servicio de inmediato. La hora de salida se calcula como sigue:

$$T = 0 + q_1 = 0 + \left(10 + 5 * 0.6733 \right) = 13.37 \text{ minutos}$$

La lista *cronológica* de los eventos futuros es entonces:

Hora T	Evento
13.37	Salida del cliente 1
42.48	Llegada del cliente 2

Salida del cliente 1 cuando $T = 13.37$

Como la instalación está vacía, se declara inactiva. Al mismo tiempo se anota que la instalación ha estado ocupada entre T = 0 y T = 13.37 minutos. La lista actualizada de eventos futuros es

Hora T	Evento
42.48	Llegada del cliente 2

Llegada del cliente 2 cuando $T = 42.48$

El cliente 3 llegará a los

$$T = 42.48 + \left[-15 \ln \left(0.4799 \right) \right] = 53.49 \text{ minutos.}$$

Como la instalación está inactiva, el cliente 2 inicia el servicio y la instalación se declara ocupada. La hora de salida es

$$\mathrm{T} = 42.48 + (10 + 5 * 0.9486) = 57.22 \text{ minutos.}$$

La lista de eventos futuros
está actualizada

Llegada del cliente 3 cuando $T = 53.49$

El cliente 4 llegará a los

$$T = 53.49 + \left[-15 \ln \left(0.6139 \right) \right] = 60.81 \text{ minutos.}$$

Como en ese momento la instalación está ocupada (hasta que $T = 57.22$), el cliente 3 se forma en la cola cuando $T = 53.49$. La lista actualizada de eventos futuros es:

Hora T	Evento
57.22	Salida del cliente 2
60.81	Llegada del cliente 4

Salida del cliente 2 cuando $T = 57.22$

El cliente 3 sale de la cola e inicia su servicio. Su tiempo de espera fue

$$W_3 = 57.22 - 53.49 = 3.73 \text{ minutos}$$

La hora de salida es

$$T = 57.22 + (10 + 5 * 0.5933) = 70.19 \text{ minutos}$$

La lista actualizada de los eventos futuros es:

Hora T	Evento
60.81	Llegada del cliente 4
70.19	Salida del cliente 3

Llegada del cliente 4 cuando $T = 60.81$

El cliente 5 llegará a los

$$T = 60.81 + [-15 \ln(0.9341)] = 61.83 \text{ minutos}$$

Como la instalación está ocupada hasta que T = 70.19, el cliente 4 se pone en la cola. La lista actualizada de los eventos futuros es:

Hora T	Evento
61.83	Llegada del cliente 5
70.19	Salida del cliente 3

Llegada del cliente 5 cuando $T = 61.83$

La simulación solo se limitará a 5 llegadas, por lo que no se genera la llegada del cliente 6. La instalación sigue ocupada y en consecuencia el cliente se forma en la cola cuando T = 61.83. La lista actualizada de eventos es:

Hora T	Evento
70.19	Salida del cliente 3

Salida del cliente 3 cuando $T = 70.19$

El cliente 4 sale de la cola para iniciar su servicio. Su tiempo de espera fue

$$W_4 = 70.19 - 60.81 = 9.38 \text{ minutos}$$

La hora de salida es

$$T = 70.19 + \left[10 + 5 * 0.1782\right] = 81.08 \text{ minutos}$$

La lista actualizada de eventos futuros es:

Hora T	Evento
81.08	Salida del cliente 4

Salida del cliente 4 cuando $T = 81.08$

El cliente 5 sale de la cola para iniciar su servicio. Su tiempo de espera fue

$$W_5 = 81.08 - 61.83 = 19.25 \text{ minutos}$$

La hora de salida es

$$T = 81.08 + \left(10 + 5 * 0.3473\right) = 92.82 \text{ minutos.}$$

La lista actualizada de los eventos futuros es:

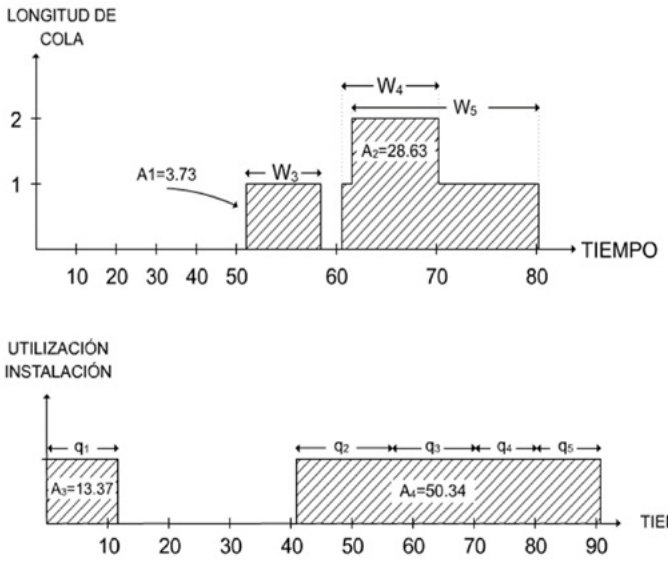
Hora T	Evento
92.82	Salida del cliente 5

Salida del cliente 5 cuando $T = 92.82$

No hay más clientes en el sistema (cola e instalación) y termina la simulación.

La figura siguiente es un resumen de los cambios en la longitud de la cola y de la utilización de la instalación, en función de la hora o tiempo de simulación.

Figura 4.2: a) Longitud de la cola y b) Utilización de la instalación en el tiempo de la simulación



La longitud de la cola y la utilización de la instalación se llaman variables **basadas en tiempo**, porque su variación es función del tiempo. En consecuencia, sus valores promedio se calculan como sigue:

$$(\text{Valor medio de una variable basada en tiempo}) = \frac{\text{Área bajo la curva}}{\text{período simulado}}$$

Al aplicar esta fórmula a los datos de la figura anterior se obtiene

$$(\text{Longitud promedio de la cola}) = \frac{A_1 + A_2}{92.82} = \frac{32.36}{92.82} = 0.349 \text{ cliente}$$

$$(\text{Utilización promedio de la instalación}) = \frac{A_3 + A_4}{92.82} = \frac{63.71}{92.82} = 0.686 \text{ peluquero}$$

El tiempo promedio de espera en la cola es una variable **basada en observación**, cuyo valor se calcula como sigue:

$$\text{Valor promedio de una variable basada en observación} = \frac{\text{Suma de las observaciones}}{\text{Cantidad de observaciones}}$$

Al examinar la figura 4.2 a) se ve que el área bajo la curva de longitud de cola en realidad es igual a la suma del tiempo de espera de los tres clientes que formaron la cola, es decir:

$$W_1 + W_2 + W_3 + W_4 + W_5 = 0 + 0 + 3.73 + 9.38 + 19.25 = 32.36 \text{ min.}$$

El tiempo promedio de espera de todos los clientes se calcula entonces como:

$$\overline{W}_q = \frac{32.36}{5} = 6.47 \text{ minutos}$$

Como podemos ver, este modelo no se ajusta a los vistos en el capítulo 2, pero como estamos considerando un tiempo de servicio uniforme tendríamos que usar el modelo de la sección 3.1 (M/G/1), capítulo 3.

c) L_s = Número esperado de unidades siendo servidas o en servicio.

$$L_s = \lambda E\{t\} + \frac{\lambda^2\left(E^2\{t\} + var\{t\}\right)}{2\left(1 - \lambda E\{t\}\right)} \qquad \text{(M/G/1)}$$

d) W_s = Tiempo esperado de permanencia en el servicio

$$W_s = \frac{L_s}{\lambda}$$

e) L_q = Número esperado de unidades en la cola

$$L_q = L_s - \lambda E\{t\}$$

f) W_q = Tiempo esperado de permanencia en la cola

$$W_q = \frac{L_q}{\lambda}$$

En este caso se tiene que $\lambda = 1/15$ que es la media de llegadas. $E\{t\} = 25/2$ la media de la distribución uniforme y $\text{var}\{t\} = 25/3$ la varianza de la distribución uniforme.

Por lo tanto, se tiene:

a) L_s= 2.93 este cálculo no se hizo en la simulación.
b) $W_s = 45.41$ si este tiempo se divide entre 5 clientes da 8.8 minutos contra 6.47 que se obtuvo de la simulación.
c) $L_q = 2.1$ este número no es similar a 0.349 como se obtuvo con la simulación.
d) $W_q \approx 31.5$ min que se aproxima a 32.36 min.

¿Simulación o teoría de colas?

De todo lo anterior podemos observar que las fórmulas para los modelos de colas corresponden a un modelo probabilístico pero estático, y las que corresponden a la simulación manual, a un modelo dinámico, ya que se van haciendo los cálculos con base en una tabla de números aleatorios, o más precisamente *pseudoaleatorios*, lo que se puede considerar una ventaja, sin embargo, la simulación manual es muy corta y discreta.

En general la ventaja de los resultados teóricos de las colas es que son exactos, es decir, no están sujetos a variación estadística (aunque en algunos casos en los que el análisis numérico puede estar involucrado para encontrar la solución, podría haber un error de redondeo). Por otro lado, los resultados de la simulación no son exactos y tienen una incertidumbre estadística asociada, que debemos reconocer, medir y abordar correctamente.

Pero la teoría de las colas tiene otras deficiencias propias, principalmente centradas en los supuestos que debemos hacer para derivar fórmulas como las vistas en el capítulo 2 y en este. En muchas situaciones del mundo real, esas suposiciones probablemente serán simplemente incorrectas y es difícil decir qué impacto podría tener en la corrección de los resultados y, por lo tanto, en la validez del modelo.

Los resultados teóricos de las colas no están disponibles universalmente para todas las distribuciones de tiempo entre llegadas y tiempo de servicio (recuerde la conveniencia matemática de la distribución exponencial sin memoria), aunque los métodos de aproximación pueden ser bastante precisos.

La simulación, por otro lado, puede manejar más fácilmente marcos de tiempo a corto plazo (o de horizonte finito). De hecho, el estado estacionario es más difícil para la simulación, ya que tenemos que correr durante mucho tiempo y también nos preocupan los efectos de sesgo de las condiciones iniciales que no son característicos del estado estacionario. Y al simular, podemos usar las distribuciones entre llegadas y tiempo de servicio que parezcan apropiadas para ajustarse al sistema real que estamos estudiando, sin embargo, como ya se mencionó existe la incertidumbre estadística por lo que

los modelos deben analizarse con técnicas estadísticas adecuadas para poder sacar conclusiones justificadas y precisas.[18]

En general las soluciones analíticas son difíciles de obtener para problemas complejos, por lo que se hace preferible el uso de la simulación.

18 https://textbook.simio.com/SASMAA/ch-queueing.html#problems

4.5 Resumen

Elección del modelo apropiado. La aplicación de la teoría de colas en la práctica implica dos aspectos importantes:

1. Selección del modelo matemático adecuado que representará al sistema real.
2. Implementación de un modelo de decisión basado en las medidas de desempeño del sistema con el fin de diseñar la instalación de servicio.

Modelos de decisión. La selección del modelo adecuado solo puede darnos medidas de desempeño que describen el comportamiento del sistema investigado, el paso siguiente es idear modelos de decisión que se puedan usar en la optimización del diseño de los sistemas de líneas de espera.

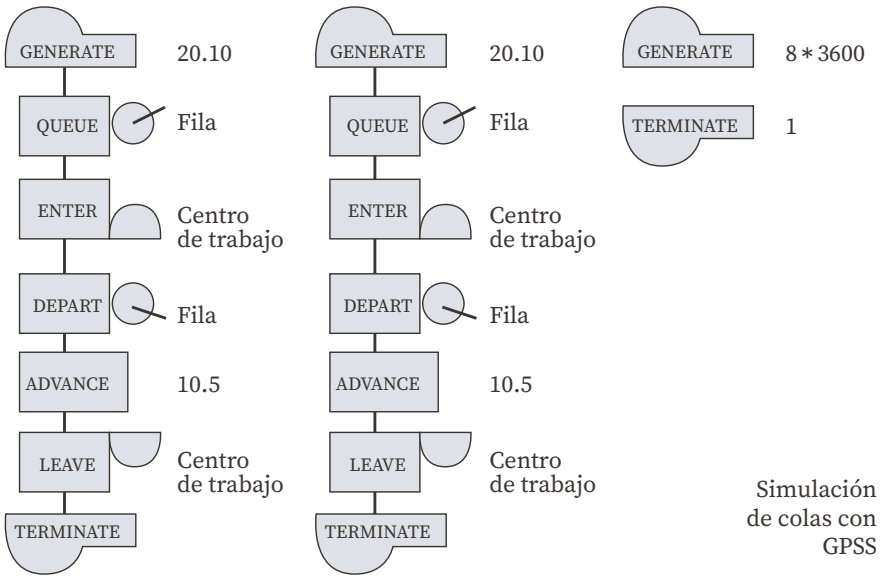
En general, un modelo de costos en líneas de espera busca equilibrar los costos de espera contra los costos de incrementar el nivel de servicio.

Teoría de colas y uso de la simulación. Las fórmulas para los modelos de colas corresponden a un modelo probabilístico pero estático, y las que corresponden a la simulación manual, a un modelo dinámico, ya que se van haciendo los cálculos con base en una tabla de números pseudoaleatorios.

4.6. Notas Históricas

El modelado de colas es muy importante para optimizar la longitud de la cola, el tiempo del cliente y para un mejor uso de los recur-

sos. Sin embargo, la simulación tiene en cuenta la aleatoriedad y la interdependencia que caracterizan el comportamiento del entorno empresarial de la vida real.



Simulación de colas con GPSS

Mediante la simulación, la aleatoriedad se puede incluir a través de distribuciones de probabilidad debidamente identificadas tomadas directamente de los datos del estudio. Sin embargo, a medida que aumenta la complejidad del análisis, también aumenta la necesidad de emplear herramientas informáticas. Si bien las hojas de cálculo pueden realizar muchos cálculos para determinar el estado operativo de sistemas simples, utilizan promedios para representar horarios, tiempos de actividad y disponibilidad de recursos. Esto no refleja con precisión la aleatoriedad y la interdependencia presentes en la realidad con los recursos y otros elementos del sistema.

En 1960, en la etapa inicial del desarrollo de las redes de telecomunicaciones con paquetes de conmutación, se aplicó la teoría de las colas a modelos de tales redes (Kleinrock L. 1964). Se han estudiado

varios modelos de colas para aplicaciones comerciales (Iliadis, I. y Denzel W.E. 1993). Las redes de colas se generan como una notación de modelado potente y se pueden aplicar a muchos dominios diferentes, incluidas las redes informáticas, el análisis de la cadena de suministro, los sistemas de software, el tráfico en las calles y otros (Serazzi G. 2008). Alan Weiss (Weiss A. 1995) proporciona una descripción general de la teoría de las grandes desviaciones y sus aplicaciones a los sistemas de telecomunicaciones. La simulación de colas para un hospital se optimiza mediante la simulación (Hang J. 1998).

4.7 EJERCICIOS PROPUESTOS

1. Considere el modelo M/M/1 con tasas de llegadas y servicio λ y μ respectivamente. Defina:
C1 = Costo por incremento unitario de μ por unidad de tiempo
C2 = Costo de espera por tiempo unitario de espera por cliente
   a) Desarrolle un modelo general de costo para el problema.
   b) Determine la μ óptima que minimiza la función de costos en a).

2. Aplique el resultado del problema 1 a la siguiente situación: los pedidos llegan a un taller de maquinaria de acuerdo con una distribución de Poisson a razón de 10 por día. Una máquina automática representa el cuello de botella en el taller. Se estima que un incremento unitario en la tasa de producción de la máquina costará al taller $ 1000 por semana. Los pedidos que no se puedan entregar a tiempo normalmente ocasionan pérdidas para el negocio, con un estimado de $ 2000 por pedido por semana. Determine la rapidez óptima de la máquina en unidades de la tasa de producción.

3. Una compañía vende dos modelos de restaurantes en franquicia. El modelo A tiene una capacidad de 80 comensales, mientras que el modelo *B* puede dar cabida a 100. El costo mensual de operación del modelo *A* es de $10 000 y el de *B* de $12 000 dólares. Un prospecto de inversionista desea abrir un restaurante en su ciudad. Él estima que sus clientes llegarán según una distribución de Poisson a una tasa de 30 por hora. El modelo A ofrecerá servicio a la tasa de 20 clientes por hora y el modelo *B* servirá a 35 comensales por hora. Cuando el restaurante esté lleno a toda su capacidad (en ambos modelos), los nuevos clientes que lleguen se van sin esperar el servicio. La pérdida por cliente por día se estima en cerca de $8. Una demora en la atención a los clientes que esperan dentro del restaurante (en ambos modelos) se calcula que tiene un costo para el dueño de $0.40 por comensal por hora, debido a la pérdida de la buena voluntad del cliente. ¿Qué modelo debe elegir el inversionista? Suponga que el restaurante estará abierto por 10 horas diarias.

4. SIMULACIÓN DE UNA TIENDA

Los clientes llegan a una tienda con un tiempo medio entre llegadas de 10 segundos, las llegadas son poissonianas. Los clientes compran de uno a cuatro artículos, con los siguientes porcentajes:

Tabla 1. Porcentajes de compra

Artículo	Porcentaje
1	50 %
2	20 %
3	20 %
4	10 %

Al llegar a la caja, si está desocupada pasan a pagar, si no, observan la cola, si en la cola hay 3 clientes o más, dejan la compra y se van, si por el contrario hay menos de tres clientes, se forman

y esperan su turno para pagar. El cajero tarda en despachar a sus clientes: 10 segundos en saludarlo más 10 segundos por cada artículo que compra. Al iniciar la simulación la tienda está vacía.

El objetivo de la simulación es calcular las siguientes medidas de desempeño:

a) Clientes atendidos
b) Artículos que no se compraron
c) El tiempo de espera promedio de un cliente en la cola
d) El tiempo promedio de utilización de la caja

5. Para el problema del barbero de este capítulo, ahora se tienen los siguientes datos: el tiempo entre llegadas sigue una distribución uniforme de 18±6 y que el tiempo que toman los cortes de pelo siguen también una distribución uniforme de 16±4. Con los datos que se dan a continuación, el diagrama de flujo y el formato anexo calcule los siguientes indicadores de desempeño:

a) Utilización promedio del servicio = suma de los tiempos de corte / longitud de la simulación.
b) Ocio promedio del servicio = suma del tiempo de ocio / longitud de la simulación.
c) Tiempo promedio de corte por cliente = suma de tiempo de corte / cortes hechos.
d) Tiempo promedio de espera por cliente = suma de tiempo de espera / personas que entraron a la fila.

Se pide simular la llegada de 16 clientes

BIBLIOGRAFÍA

Bath, N. (2008) *An Introduction to Queueing Theory*, Birkhauser Boston.

Flores De La Mota, I. Guasch, A., Mújica Mota, M., Piera, M. (2017). *Robust Modelling and Simulation.* DOI 10.1007/978-3-319-53321-6 Springer.

García Arana, A. (1987). *Apuntes de Teoría de la Espera*, DEPFI, UNAM.

Chen, H., Yao, D.,(2001). *Fundamentals of Queueing Networks*. Springer Science Business Media New York.

Taha, H. (2004). *Introducción a la Investigación de Operaciones*, 7ta. Ed. Pearson Prentice Hall.

Citas

https://www.gestiondeoperaciones.net/lineas-de-espera/que-es-la-ley-de-little-y-su-aplicacion-en-el-analisis-de-lineas-de-espera/

https://www.netlab.tkk.fi/opetus/s383143/kalvot/E_mg1jono.pdf

https://es.wikipedia.org/wiki/F%C3%B3rmula_Pollaczek-Khintchine.

https://estadistica.net/IO/7-7-TEORIA-COLAS.pdf

UNIDAD DE APOYO EDITORIAL

Introducción a la Teoría de colas
Se publicó la primera edición electrónica de un ejemplar
(3 MB) en formato PDF en agosto de 2023,
en el repositorio de la Facultad de Ingeniería, UNAM,
Ciudad Universitaria, Ciudad de México. C.P. 04510

El diseño estuvo a cargo de la Unidad de Apoyo Editorial
de la Facultad de Ingeniería. Las familias tipográficas
utilizadas fueron Source Serif Pro y Sienna Math Pro para
texto y fórmulas matemáticas, y Brevia para títulos.