Aportes básicos de la **TEORÍA DE COLAS** a la **administración** y la **gerencia**

Alberto **Cadevilla**, Amaire **Mora**

UNELLEZ
La Universidad que Siembra

Ediciones de la Universidad Ezequiel Zamora
Colección: **Docencia Universitaria**

albertocadevillasoto@gmail.com
alcaso24@gmail.com
cadevilla@unellez.edu.ve

Alberto José Cadevilla Soto

Profesor - Investigador. Politólogo Cum Laude (Universidad de los Andes [ULA], Mérida -Venezuela). Magister Scientiarum en Administración Mención Gerencia General (Universidad Nacional Experimental de los Llanos Occidentales «Ezequiel Zamora» [UNELLEZ]). Doctor en Ciencias Administrativas y Gerenciales (Facultad de Ciencias Económicas y Sociales [FaCES], Universidad de Carabobo [UC]). Profesor Agregado a Dedicación Exclusiva en la Categoría de Agregado, Programa de Ciencias Sociales y Económicas (PCSE) del Vicerrectorado de Producción Agrícola (UNELLEZ - VPA), Guanare - Venezuela. Facilitador en las Maestrías en Administración Mención Gerencia General, Gerencia y Planificación Institucional y Gerencia Pública del Programa de Estudios Avanzados de UNELLEZ - VPA. Profesor Invitado de la Maestría en Administración de Empresas (UC - Guanare). Investigador del Centro de Investigación y Desarrollo de la Pequeña y Mediana Empresa y la Microempresa del Estado Carabobo (CIDPyMESMicro) de FaCES - UC. Miembro de la Red de Investigadores Latinoamericanos en Administración y Gerencia (RILAyG). Miembro del Comité Editorial de la Revista Equidad del PCSE - UNELLEZ - VPA. Autor de artículos científicos, otras publicaciones y ponencias a nivel nacional e internacional en el campo de la Ciencia Política, la Administración y la Gerencia tanto Pública como Privada.

amairemora@gmail.com
amaire@unellez.edu.ve

Amaire Josefina Mora Guerrero

Profesora - Investigadora. Politóloga Cum Laude (Universidad de los Andes [ULA], Mérida - Venezuela). Magister Scientiarum en Administración Mención Gerencia General (Universidad Nacional Experimental de los Llanos Occidentales «Ezequiel Zamora» [UNELLEZ]). Doctora en Ciencias Administrativas y Gerenciales (Facultad de Ciencias Económicas y Sociales [FaCES], Universidad de Carabobo [UC]). Profesora Agregada a Dedicación Exclusiva en la Categoría de Agregada, Programa de Ciencias Sociales y Económicas (PCSE) del Vicerrectorado de Producción Agrícola (UNELLEZ - VPA), Guanare - Venezuela. Facilitadora en las Maestrías en Administración Mención Gerencia General, Gerencia y Planificación Institucional y Gerencia Pública del Programa de Estudios Avanzados de UNELLEZ - VPA. Profesora Invitada de la Maestría en Administración de Empresas (UC - Guanare). Investigadora del Centro de Investigación y Desarrollo de la Pequeña y Mediana Empresa y la Microempresa del Estado Carabobo (CIDPyMESMicro), FaCES - UC. Miembro de la Red de Investigadores Latinoamericanos en Administración y Gerencia (RILAyG). Miembro del Comité Editorial de la Revista Equidad del PCSE - UNELLEZ - VPA. Autora de artículos científicos, otras publicaciones y ponencias a nivel nacional e internacional en el campo de la Ciencia Política, la Administración y la Gerencia tanto Pública como Privada.

AUTORIDADES UNIVERSITARIAS:

Adán Chávez Frías
Rector

Joneidi Carolina Rivas
Secretaría General

Antonio José Albarrán
Vicerrector de Servicios

Gyzel Guillén
Vicerrectora de Planificación
y Desarrollo Social

Luis Eduardo Rosales
Vicerrector de Producción Agrícola

Hayden Pirela
Vicerrectora de Infraestructura
y Procesos Industriales

Marielida Rodríguez
Vicerrectora de Planificación
y Desarrollo Regional

Dalia González
Gerente de la Fundación Editorial
Universidad Ezequiel Zamora

Aportes básicos de la teoría de colas a la administración y la gerencia

Primera edición, 2024

Gustavo Quintana
Diseño de cubierta

Depósito Legal: BA2023000009
ISBN: 978-980-248-312-9

UNELLEZ
Universidad Nacional Experimental
de los Llanos Occidentales «Ezequiel Zamora»
La Universidad que Siembra

feduez
Fundación Editorial
Universidad Ezequiel Zamora

APORTES BÁSICOS DE LA TEORÍA DE COLAS A LA ADMINISTRACIÓN Y LA GERENCIA

ÍNDICE DE FIGURAS

INTRODUCCIÓN

La espera para la adquisición de un bien, la obtención de un servicio, satisfacer una necesidad, cumplir con un requerimiento, es una situación recurrente en la vida moderna; incluso, el acceso a la información, la llegada de correos electrónicos, la ejecución de un comando de una aplicación conlleva un tiempo de espera para su ejecución.

De ahí la importancia de la teoría de colas para la toma de decisiones en la administración y la gerencia, en tanto mecanismo de orientación y guía con respecto a los parámetros de acción para asegurar un manejo y gestión eficaz, eficiente y efectivo de los recursos de la organización, bien sea pública, bien sea privada.

En ese sentido, las hileras uno en fondo, o las filas, o las colas, o las líneas de espera, nombre con el que se suele designar la espera y, además, recogen muchas de las situaciones de la vida diaria que reflejan la espera, entre ellas cabe mencionar: los compradores hacer colas en los supermercados y abastos para pagar los productos que han escogido, los usuarios hacen filas frente a las taquillas de los servicios públicos –agua, electricidad, aseo urbano, gas, por indicar algunos– para pagar su consumo mensual, los consumidores desarrollan líneas de espera ante las cajas de pago de las tiendas para poder adquirir sus productos.

Estas situaciones se refieren al ámbito de las personas, pero también en el ámbito de las organizaciones se desarrollan líneas de espera, cuando las piezas esperan en las líneas de ensamblaje, o cuando los productos terminados hacen fila para ser embalados, o cuando se establecen colas para la atención de demandas de producción.

Es por ello que las empresas buscan que los gerentes de planta, de producción, de comercialización y, en general, cualquier gerente que tenga a su cargo el desarrollo del proceso productivo, o parte de este, implementen los sistemas o modelos que

ayuden a gestionar los tiempos de espera. Es común buscar disminuir, minimizar o reducir la espera en las líneas de montaje, de no hacerlo la espiral de costos tiende a incrementarse.

En razón de lo cual, es preciso encontrar la tasa óptima de servicio que haga factible reducir los costos de espera junto a los costos de servicio, haciendo que el costo total no impacte negativamente en los balances financieros de las organizaciones.

Todo esto viene a demostrar la importancia de estudiar y analizar las colas, o las líneas de espera, para poder comprender su naturaleza, características y complejidades, así como la forma de manejarlas, superarlas y facilitar su resolución.

El abordaje de la teoría de colas –líneas de espera, filas, hileras uno en fondo, entre otras denominaciones– se hace a partir de algunas generalidades, para luego definirla y establecer sus objetivos; así, lograr determinar los elementos básicos de los modelos de colas, sus costos y las medidas de rendimiento que se emplean para evaluar los sistemas de líneas de espera.

Con ello se hace un análisis económico de estos sistemas y se estructuran las características de los modelos de colas de acuerdo con criterios físicos y de funcionalidad. Posteriormente, se habla de los parámetros de los sistemas de colas, que son la antesala a la operacionalización de los modelos de líneas de espera a través de los modelos de llegada y de tiempo de servicio.

A continuación, se señala la relación entre las funciones de las distribuciones de Poisson y Exponencial, lo cual permite hablar de lo modelos de nacimiento puro y muerte pura. Se hace mención a la distribución de Erlang, las colas especializadas de Poisson y la Notación de Kendall y, por supuesto, se alude al modelo generalizado de Poisson. Lo cual posibilita la mención de las medidas de desempeño de estado estable y de las líneas de espera especializadas de Poisson.

En este punto se describen los modelos:

M/M/1, (M/M/1):(DG/N/∞),

(M/M/c): (DG/∞/∞),

(M/M/c):(DG/N/∞) $c \leq N$,

Modelos de autoservicio,

Modelos de servicio de máquina.

Asimismo, también se alude a los modelos que obedecen a la distribución de Poisson, a los de líneas de espera con prioridad en el servicio, a los de no prioridad con un servidor, a los de no prioridad con varios servidores, a modelos en serie de dos estaciones con capacidad de líneas de espera cero y en serie de k estaciones con capacidad de líneas espera infinita.

Por último, se hace referencia a la selección del modelo apropiado de líneas de espera, a los modelos de decisión en línea de espera, a los modelos de costo tanto con tasa óptima de servicio como número óptimo de servidores y se cierra con el modelo de nivel de aceptación.

Es preciso tener presente que el propósito de este libro es generar una teorización básica sobre la teoría de colas desde la perspectiva de la administración como ciencia y de la gerencia como praxis que enriquezca el conocimiento necesario para facilitar el proceso de toma de decisiones que asegure la gestión eficiente de los procesos productivos, administrativos, comerciales y humanos.

La idea fuerza detrás de texto es comprender, entender y aprehender la dinámica operacional de las líneas de espera y a partir de ahí generar una explicación que facilite la comprensión del proceso.

Al mismo tiempo, allane la formación gerencial de administradores y gerentes en aspectos matemáticos, administrativos y gerenciales de amplia influencia en la superación de los «cuellos de botella» de las colas con la mayor eficacia, eficiencia y, por ende, efectividad.

El desarrollo de la teorización propuesta en este libro se realiza en tres etapas, a saber:

Primera etapa: se describen los aportes de la teoría de colas, para ello se hace una indagación y a partir de ahí se genera una categorización de los aspectos más resaltantes y que, a su vez, permitan una develación de los elementos que conforman tales categorías.

Segunda etapa: se comprenden los aportes de la teoría de colas a la administración y la gerencia; el desarrollo y fundamentación de las categorías develadas posibilita darles contenido en el marco de su categorización en la administración y la gerencia, como esos factores que se convierten en variables decisivas en el proceso de toma de decisiones gerenciales.

Tercera etapa: se explican los aportes básicos de la teoría de colas a la administración y la gerencia. Develados y comprendidos los aportes es momento de explicarlos, de internalizarlos, de ejemplificarlos; de manera de hacerlos factibles y palpables dentro de las ciencias administrativas y gerenciales.

Este texto, ***los aportes básicos de la teoría de colas a la administración y la gerencia***, busca profundizar en cómo una teoría matemática de modelos cuantitativos de múltiples variantes puede ser aplicada a heterogéneas situaciones de gestión y administración de recursos en distintos ámbitos, departamentos o áreas funcionales –compras, cuentas, *marketing*, producción, atención al público, inventarios, entre muchas otras– de una empresa –pública, privada o mixta– de naturaleza extractiva, industrial, comercial, de servicios y de cualquier otra naturaleza económica.

Todas las organizaciones humanas son susceptibles de sufrir el fenómeno de las colas o de las líneas de esperas y, en razón de ello, todas son proclives a la administración y gestión de dicho fenómeno para asegurar la rentabilidad, productividad, competitividad, sustentabilidad y operatividad de la organización.

Las categorías que se emplean para agrupar y codificar la información están relacionadas con: eficiencia del servicio, satisfacción del consumidor o cliente, asertividad y oportunidad de la toma de decisiones, *engagement* o compromiso de los empleados, fidelización de los usuarios, sostenibilidad de la organización.

COLAS, FILAS O LÍNEAS DE ESPERA

Las colas, filas o líneas de espera son parte de la vida diaria; todas las personas esperan en filas para comprar un boleto para el cine, hacer un depósito en el banco, pagar en el supermercado, enviar un paquete por correo, simplemente repostar de combustible el tanque del automóvil, entrar un correo electrónico en la bandeja de entrada, la salida de una comunicación de un servidor electrónico, entre otras situaciones. De ahí, la necesidad de estudiarlas y analizarlas, para poder comprender su naturaleza, características y complejidades, así como la forma de manejarlas, superarlas y facilitar su resolución.

Teoría de Colas

En tal sentido, la teoría de colas es el estudio de la espera en sus diversas modalidades y manifestaciones. Para Vega (2004),

> Las 'colas' son un aspecto de la vida moderna que nos encontramos continuamente en nuestras actividades diarias. En el contador de un supermercado, accediendo a Internet,... el fenómeno de las colas surge cuando unos recursos compartidos necesitan ser accedidos para dar servicio a un elevado número de trabajos o clientes. (p.12).

Se usan los modelos de colas para representar los tipos de sistemas de líneas de espera, que deben ser entendidos como sistemas que involucran algún tipo de filas que surgen en la práctica. Las fórmulas para cada modelo indican cuál debe ser el desempeño del sistema correspondiente y señalan la cantidad promedio de espera que ocurrirá, en una gama amplia de circunstancias. De ahí que, los modelos de líneas de espera que se han desarrollado han resultado muy eficientes en la gestión de los sistemas de colas, apuntando: a mejorar la forma de realizar los procesos productivos, de llevar el inventario y de realizar todas las actividades involucradas con la gestión del inventario mismo.

- A la eficiencia en la satisfacción del servicio requerido por el usuario o cliente.

- A la eficacia en lograr esa satisfacción del servicio en el menor tiempo de espera posible.

- A la efectividad, equilibrando ambos factores en el mayor número de usuarios probables para salvaguardar la reputación, el mercadeo e imagen de la organización por la gestión de sus procesos y actividades.

- Por consiguiente, a la salud económico-financiera de la organización.

El impacto de las colas o de las líneas de espera resulta fundamental para la estabilidad, sostenibilidad y rentabilidad de las empresas, organizaciones o instituciones, muchas veces llega a ser catastrófico en términos económicos, financieros, de *marketing* y de reputación. Esto se debe a que, por un lado, mientras mayor sea la espera de un usuario o cliente para la satisfacción de la necesidad requerida, mayor será su tiempo invertido en esa espera interminable; lo que es la mismo, la pérdida de un tiempo que pudo haber invertido en la consecución de otro objetivo o satisfactor. Lo que termina afectando la efectividad de la empresa en términos de prestigio o reputación y a la larga económicos por la ineficiencia del proceso.

Pero, más importante aún, mientras mayor sea el tiempo que la empresa tarde en satisfacer un requerimiento de un usuario, más es el tiempo que se invierte en buscar la forma de gestionar la satisfacción que se le ha requerido, se incrementa la inoperancia de los procesos administrativos y gerenciales. Además, ocurre una concentración y redundancia de actividades para tratar de superar el «cuello de botella» o la falta de fluidez del servicio. Así mismo, se empieza a desarrollar la frustración de los operarios o empleados encargados de brindar el servicio a los clientes o usuario; lo cual engendra cuadros de estrés y tensión. Todo conjugado

genera costos económicos que se van incrementando y verificando en el corto, mediano y largo plazo.

En virtud de ello, la teoría de colas –que también pueden ser denominadas filas, series, líneas– es el estudio matemático de las líneas de espera; pues la formación de filas es un fenómeno común que ocurre siempre que la demanda efectiva de un servicio excede a la oferta efectiva. Las empresas toman decisiones con relación a la cantidad de servicios y bienes de que disponen o pueden ofrecer en un momento determinado, procurando la mayor escala de satisfacción viable. No obstante, es imposible fijar con exactitud el tiempo de llegada de los clientes que solicitarán o usarán el bien o el servicio; así como tampoco, es factible precisar cuánto tiempo se empleará para facilitar el servicio o cumplir con la entrega del bien. Convirtiéndose tales situaciones en un problema, típicamente administrativo y gerencial de muy difícil respuesta para un decisor, sin las adecuadas herramientas formativas y las competencias matemáticas requeridas.

Necesidad de Atender las Líneas de Espera

Por lo anteriormente señalado, tomar una decisión con respecto a tales situaciones genera disyuntivas y conflictos que hay que solucionar con la mucha o poca información disponible y formación que se posea, aquí es donde la teoría de colas juega un papel preponderante. Toda organización o empresa no puede estar preparada para prestar todos los servicios que le requieran sus usuarios, ni facilitar todos los bienes que le exigen los consumidores; ello sería muy costoso en cuanto al mantenimiento de recursos ociosos. Sin embargo, no contar con una eficiente prestación de servicio o de rapidez en la facilitación del bien requerido, lo que se conoce como capacidad de servicio, puede ocasionar filas excesivamente largas en algunos periodos de tiempo. Tal situación, de hecho, es costosa para la empresa ya que trae pérdida de prestigio y de clientes; mientras que los clientes al esperar en filas pagan un coste en incomodidad, cansancio, hastío y, sobre todo, en tiempo, como se ha descrito previamente.

Estos costos pueden generar varios hechos:

- Reducción del número de clientes.

- Incremento del ausentismo laboral.

- Bajas médicas por estrés laboral.

- Alta rotación de trabajadores.

- Disminución del atractivo de la empresa tanto para trabajadores como para clientes o usuarios.

De acuerdo con Render, Stair y Hanna (2018: 593), es necesario tener en cuenta que "El estudio de líneas de espera, llamado teoría de colas, es una de las técnicas de análisis cuantitativo más antiguas y que se utilizan con mayor frecuencia.". Debido a que, las situaciones de espera en la atención del servicio, en la línea de producción, en la respuesta a un pedido siempre ha estado presente en la administración y gerencia como factor decisivo en la gestión de los recursos. Asimismo, agregan Render, Stair y Hanna (2018: 300), "Los tres componentes básicos de un proceso de colas son las llegadas, las instalaciones de servicio y la línea de espera real.".

Es por ello que se busca responder a la pregunta: ¿cuáles son los aportes básicos de la teoría de colas a la administración y la gerencia? En ese sentido, se pretende explicar las contribuciones básicas de la teoría de colas a las ciencias administrativas y a la praxis gerencial. En ese sentido, se debe decir que este libro parte de la idea de abordar elementos básicos de primer orden de dicha teoría y sus implicaciones aludiendo, en todo momento y de manera transversal, a todo tipo de organizaciones, empresas productivas y de servicio, tanto de naturaleza pública como de naturaleza privada, siempre desde la perspectiva administrativo-gerencial.

Tomando en consideración el hecho que las personas nunca terminan de acostumbrarse a esperas largas e incómodas mientras hacen numerosas y extensas filas. Adicionalmente, valorando y reflexionando sobre las actuales tendencias organizacionales y de los sistemas humanos, es significativo poder gestionar de manera exitosa la reducción de los tiempos de espera y de esa manera contribuir, en alguna medida, a la generación de satisfactores en la población que contribuyan a reducir el estrés tanto laboral como social, a su vez, reducir los niveles de tensión laboral.

A lo que es necesario agregar aumentar el número de clientes o usuario debido a la naciente reputación por presentar menores tiempos de espera en la prestación del servicio o satisfactor. Adicionalmente, permite reducir el ausentismo laboral, pues el ambiente se torna agradable y el clima laboral se torna motivador e inspirador, favoreciendo el espíritu colaborativo y de servicio. Lo que viene a contribuir a una menor rotación de los trabajadores, a la par que la empresa, organización o institución se va convirtiendo de a poco en atractiva tanto para los trabajadores como para los clientes o usuarios.

En general, uno de sus factores clave de éxito –de las organizaciones, empresas e instituciones–, es el aprovechamiento del tiempo y los recursos en procura del continuo mejoramiento y de la optimización de procesos y productos para satisfacer las necesidades de los consumidores y usuarios, la continua demanda de estos últimos en procura de una satisfacción rápida, cuando no inmediata; las filas generan molestia, siendo un factor de frustración que atenta contra el desarrollo sostenible, el bien común y la felicidad social.

En ese sentido, sin lugar a dudas, tener que esperar no es sólo una molestia personal, el valioso e irrecuperable tiempo que la población de un país pierde en colas o filas, es un factor importante que redunda tanto en la calidad de vida como en la productividad y competitividad de organizaciones y empresas tanto públicas como privadas a nivel local, regional y nacional; así como de la economía en su conjunto.

De ahí la importancia de subsanar las situaciones que contribuyen a su generación y tienden a la inoperancia de los procesos productivos, comerciales, comunicacionales y de variada índole.

Es por ello que en este libro se parte de la necesidad de describir los aportes básicos de la teoría de colas para llegar a comprender los aportes básicos de la teoría de colas a la administración y la gerencia y, de esa forma, poder explicar los aportes básicos de la teoría de colas a la administración y la gerencia. Se busca hacer una revisión amplia y profunda de la literatura existente y a partir de ahí lograr categorizar tales aportes, brindando un contexto que permita su comprensión desde la administración y la gerencia y, a su vez, sirva en la orientación y guía del decisor, gerente o administrador para lograr la efectividad del manejo de los tiempos de espera.

Se procura ampliar la comprensión de la teoría de colas y su incidencia en las organizaciones, desde la perspectiva de la administración y la gerencia. Las decisiones sobre las líneas de espera, en cuanto a su manejo y gestión, producen grandes debates en las empresas en particular y en las organizaciones en general, pues la relación costo/beneficio que suele establecerse en torno a las filas, es de marcado influjo sobre los rendimientos, la productividad y la competitividad empresarial. En tal sentido, es necesario, a partir de las revisiones y teorizaciones, generar un enfoque y conceptualización que amplíe el marco de acción del tomador de decisiones dentro de una administración y una gerencia eficaz, eficiente y efectiva.

El interés general es ampliar la comprensión sobre los aportes a la teoría de cola o líneas de espera a la administración y gerencia a objeto de mejorar la eficiencia de los procesos, de garantizar tomas de decisiones contextualizadas y acordes a las necesidades de satisfacción de usuarios y/o clientes. Se insiste en que el propósito fundamental es ampliar los márgenes de certidumbre de los tomadores de decisiones en los diferentes niveles o ámbitos organizacionales. Por ello, se trata de demostrar la incidencia e influjo de las filas, o colas, o líneas de espera, en la dinámica

organizacional, los costos añadidos que se generan asociados a los inconvenientes que engendran este tipo de situaciones que entorpecen la buena marcha productiva, el desempeño eficiente de las personas y el cumplimiento de los fines organizacionales.

En virtud de esto, la reflexión crítica sobre los contenidos revisados, el análisis de la documentación recabada, la valoración del soporte teórico recaudado posibilita la construcción de una teorización propia, con una perspectiva particular, que apunta al desarrollo de nuevos aportes básicos de la teoría de colas a la administración y la gerencia. Ello es posible, en gran parte, a la necesidad de ampliar las perspectivas y opciones de acción, de gestión, de decisión de los administradores, de los gerentes y de los tomadores de decisiones, a la hora de enfrentar esta problemática tan compleja y con profundas repercusiones en la rentabilidad, estabilidad y sostenibilidad de las organizaciones. Más cuando las sociedades humanas se complejizan cada vez más y son más las situaciones que generan espera, contratiempos, demoras en su arribo o entrada, bien a las líneas de producción, bien a las líneas de respuesta, bien a la atención del servicio.

GENERALIDADES SOBRE LA TEORÍA DE COLAS

Las colas o líneas de espera se han convertido en un fenómeno cotidiano de la vida moderna, es común encontrarlas en casi todas las actividades diarias de los seres humanos, ejemplo de ello vendrían a ser: los clientes que esperan ser atendidos por la cajera de un supermercado, los automóviles que esperan avanzar en un semáforo con luz roja, los pacientes que esperan ser atendidos en un centro de salud y asistencia social, los aviones que están por despegar o por aterrizar en las pistas de los aeropuertos, las maquinarias que deben ser reparadas por un técnico, las cartas que deben ser escritas por una secretaria o un asistente administrativo; así como los programas y aplicaciones que deben ser ejecutadas por un procesador o un dispositivo electrónico.

Aspectos comunes de las Líneas de Espera

El elemento común de todas las situaciones presentadas es la espera, por ello, se piensa en la necesidad de ofrecer y prestar servicios sin la condición de esperar. No obstante, explica Serra (2002: 96), esto es inevitable pues "las colas surge cuando unos recursos compartidos necesitan ser accedidos para dar servicio a un elevado número de trabajos o clientes". Si se considera que uno de los mayores dilemas de la economía es «satisfacer necesidades infinitas con recursos escasos», es inevitable que siempre vaya a presentarse la espera, por varias razones: una, los primeros en llegar accederán primero a los recursos; dos, la llegada de demandantes es continua lo que incrementa la presión sobre los recursos; entre muchas otras razones.

El origen de la teoría de colas se puede precisar en al año 1909, con el esfuerzo del danés Agner Krarup Erlang para analizar la congestión de tráfico telefónico con el objetivo de cumplir la demanda incierta de servicios en el sistema telefónico de Copenhague, Suecia. Sus investigaciones acabaron en una nueva teoría llamada teoría de colas o de líneas de espera; esta se ha convertido en una valiosa herramienta

en el área de los negocios, debido a que muchos de los problemas que se presentan en estos, los negocios, pueden ser caracterizados a partir de dicha teoría, es decir, como problemas de congestión llegada - partida.

Definiciones de la Teoría de Colas

El estudio de las colas proporciona tanto una base teórica del tipo de servicio que se puede esperar de un determinado recurso –tecnológico, financiero, administrativo, gerencial, comercial, o de cualquiera otra naturaleza–, como la forma en que dicho recurso puede ser diseñado para proporcionar un determinado grado de servicio a sus clientes. De ahí que Cao (2002) las define como el estudio matemático del comportamiento de líneas de espera; se presentan cuando «clientes» llegan a un «lugar» demandando un servicio a un «servidor», el cual tiene cierta capacidad de atención. Si el servidor no está disponible inmediatamente y el cliente decide esperar, entonces se forma en la línea de espera.

Cao (2002) delimita a los clientes con base a ejemplos: gente esperando, líneas telefónicas desocupadas, máquinas que esperan ser reparadas.

Adicionalmente, explica el lugar o las instalaciones de servicio en función de los ejemplos dados, en este caso son las líneas telefónicas y los talleres de reparación.

En el mismo orden, las llegadas son entendidas por Cao (2002) como el número de clientes que llegan a las instalaciones de servicio para solicitarlo, adquirirlo, usarlo o servirse de dicho servicio.

En tanto la tasa de servicio designa la capacidad de servicio que tiene la empresa, organización o institución en cuestión.

Siguiendo con los ejemplos dados, para Cao (2002), un sistema telefónico entre dos ciudades puede manejar hasta 90 llamadas por minuto, una instalación de

reparación o taller puede reparar en promedio una máquina cada 8 horas. En la figura 1 se establece lo que se considera o se define como una cola clásica o típica.

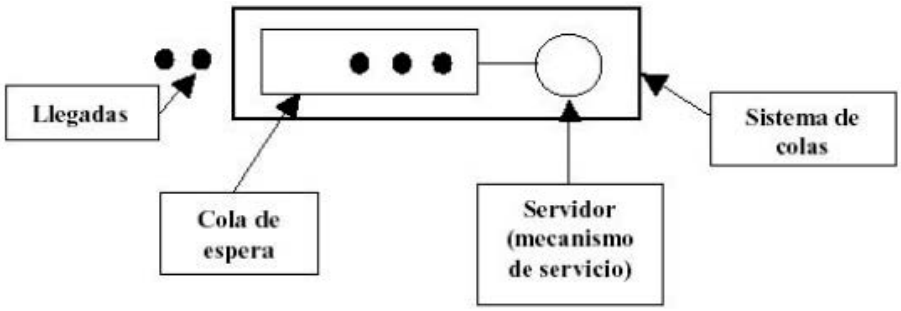


Figura 1. Definición de la teoría de colas
Fuente: Serra (2002: 122).

Para Moskowitz y Wright (1993) una cola es una línea de espera y la teoría de colas es una colección de modelos matemáticos que describen sistemas de línea de espera particulares o sistemas de colas. Los modelos sirven para encontrar un equilibrio entre costes del sistema y los tiempos promedio de la línea de espera para un sistema dado. El problema es determinar qué capacidad o tasa de servicio proporciona el balance correcto de espera tanto para los clientes como para la propia organización, donde los tiempos de espera no se traduzcan en un incremento continuo de los costos financieros –presentes y futuros–, económicos, de reputación, humanos, de efectividad administrativo-operacional, de efectividad gerencial-organizacional, de salud –física, psicológica, emocional de empleados y gerentes–, temporales, sostenibilidad de la organización, entre otros tipos de costos.

Establecer todo lo anterior no es sencillo, ya que un cliente no llega en un horario fijo o en horarios preestablecidos o predeterminado; es decir, no se sabe con exactitud en qué momento llega o llegará un cliente, o en todo caso cuándo llegarán –lapso temporal– varios clientes o los clientes en general, o si llegarán de uno en uno o varios en un solo momento, o si llegarán todos los clientes al mismo tiempo. También es necesario acotar que el tiempo de servicio no tiene un horario fijo o determinado con anticipación, pues varía de acuerdo con las necesidades y expectativas de los

clientes. Es por ello que surgen los dilemas de decisión, tal como se presentan en la figura 2.

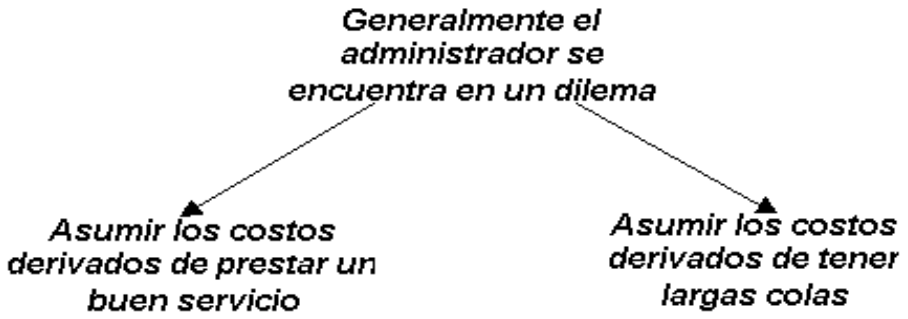


Figura 2: Dilema de la teoría de colas
Fuente: Adaptación de Moskowitz y Wright (1993).

En tal sentido, los sistemas de colas son modelos de sistemas que proporcionan servicio, por eso como modelos pueden representar cualquier sistema en donde los trabajos, usuarios, demandantes o clientes llegan buscando un servicio de algún tipo y salen después que dicho servicio haya sido atendido, o la atención haya sido prestada, o la necesidad haya sido satisfecha. Se pueden modelar los sistemas de éste tipo tanto como colas sencillas o como un sistema de colas interconectadas formando una red de colas.

El modelo de colas sencillo puede usarse para representar una situación típica en la cual los clientes llegan, esperan si los servidores están ocupados, son servidos por un servidor disponible y se marchan cuando se obtiene el servicio requerido. Tal como se representa en la figura 3. En dicha figura se representa una situación muy corriente para la formación y manejo de las hileras de uno en fondo; ejemplo de este tipo de colas son las líneas de espera de los bancos o de atención al público en los centros de gestión de los servicios públicos. Los usuarios o demandantes arriban al lugar, esperan si los despachadores, operarios, máquinas están ocupados, cuando les corresponde en la línea son atendidos por un servidor que haya quedado disponible: satisfecha la demanda se marchan del lugar o culminan su operación, dando cabida al siguiente de la fila, como se muestra en las figuras 4 y 5.

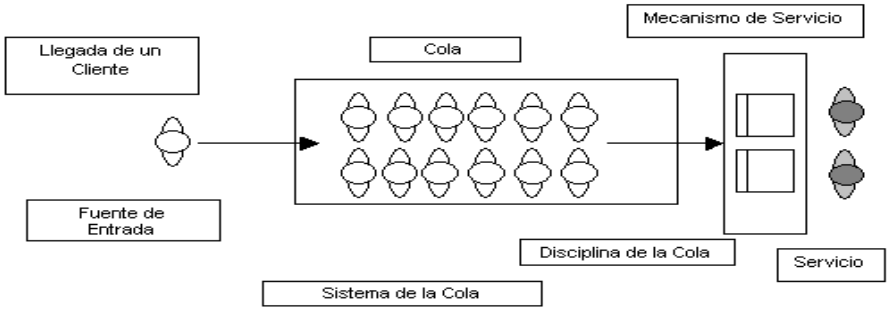


Figura 3. Modelo de colas sencillo

Fuente: Moskowitz y Wright (1993).

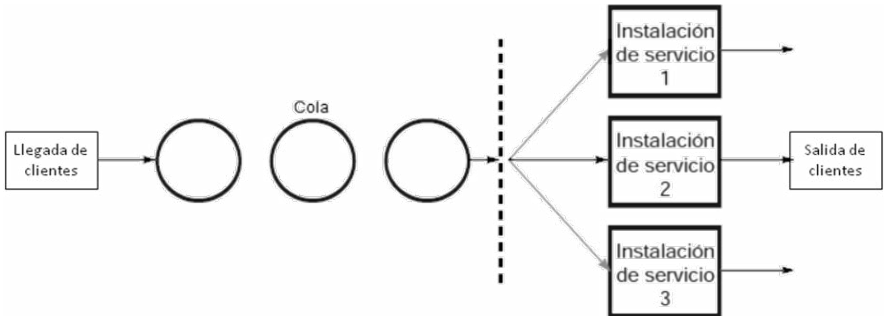


Figura 4. Modelo de colas multicanal una sola fase

Fuente: Render, Stair y Hanna (2018: 504).

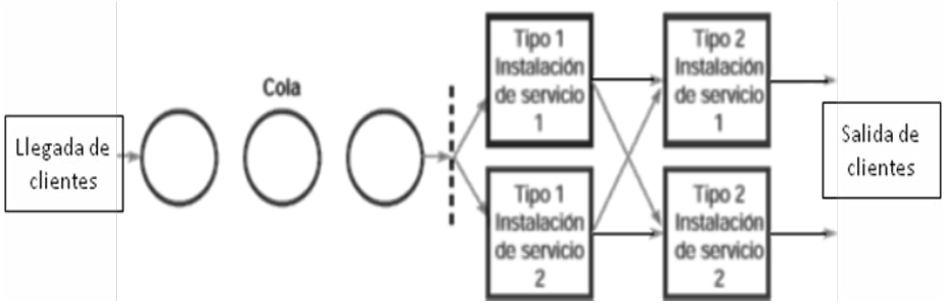


Figura 5. Modelo de colas multicanal con multifase

Fuente: Render, Stair y Hanna (2018: 504).

Objetivos de la Teoría de Colas

Para Moskowitz y Wright (1993) los objetivos de la teoría de colas consisten en:

- Identificar el nivel óptimo de capacidad del sistema que minimiza su coste global.

- Evaluar el impacto que las posibles alternativas de modificación de la capacidad del sistema tendrían en su coste total.

- Establecer un balance equilibrado, es decir, óptimo entre las consideraciones cuantitativas de costes y las cualitativas de servicio.

- Hay que prestar atención al tiempo de permanencia en el sistema o en la cola: la paciencia de los clientes depende del tipo de servicio específico considerado y eso puede hacer que un cliente abandone el sistema.

En ese sentido, Serra (2002: 124) explica que todos los modelos o sistemas que devienen de la teoría de colas o de líneas de espera tienen esencialmente dos propósitos: uno, "minimización del tiempo de espera" tanto para la organización como para los usuarios o clientes; debido a que ello implica un ahorro de recursos. Y, dos, "minimización de los costes totales de funcionamiento del sistema", lo cual redunda en la efectividad de funcionamiento y operatividad de la organización.

Empero se debe considerar que ambos propósitos son contradictorios, conflictivos y paradójicos. Para Serra (2002: 124) la complicación nace porque "para reducir el tiempo de espera se necesitan poner más recursos en el sistema, con el consiguiente aumento de los costes de producción"; en otras palabras, las incertidumbres asociadas a la inversión recurrente se hacen presentes, además de la incidencia que pueden tener la eficiencia o ineficiencia del servicio prestado. Adicionalmente, enfatiza Serra (2002: 124), "En muchos casos el tiempo de espera es difícil de determinar, sobre todo cuando se trata de un sistema en donde seres humanos están implicados"; por la volubilidad de las personas en sus procesos de toma de decisiones.

Elementos de los Modelos de Colas

Rodríguez y Gámez (2002) revelan que se pueden utilizar sistemas de colas para modelar procesos en los cuales los clientes van llegando, esperan su turno para receptar el servicio, reciben el servicio y luego se marchan. Es por ello que afirman que los sistemas de colas pueden definirse mediante cinco componentes:

- La función de densidad de probabilidad del tiempo entre llegadas.
- La función de densidad de probabilidad del tiempo de servicio.
- El número de servidores.
- La disciplina de ordenamiento en las colas.
- El tamaño máximo de las colas.

En razón de tales componentes se hace imprescindible precisar que, arguyen Rodríguez y Gámez (2002: 194), "La densidad de probabilidad del tiempo entre llegadas describe el intervalo de tiempo entre llegadas consecutivas." En otras palabras, mediante un proceso de observación se anota o registra la llegada de cada usuario, demandante o cliente a la organización, empresa o institución, entre cada anotación se debe indicar la hora precisa de arribo para de esa manera poder establecer la ocurrencia de arribo o llegada y, a partir de ahí, establecer la intermitencia de llegada de cada usuario o cliente.

Es necesario subrayar que lo mismo debe hacerse con la salida de los clientes, usuarios o demandantes de los servicios de la organización, para medir la eficacia, eficiencia y efectividad del servicio o atención prestada o del bien producido. Tal como se muestra en la figura 6.

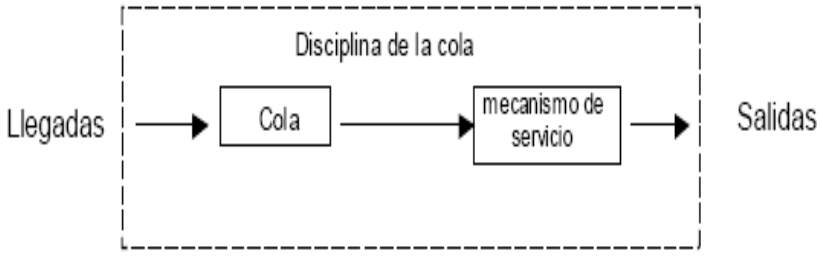


Figura 6. Esquema de un sistema de colas
Fuente: Rodríguez y Gámez (2002: 196).

La densidad de probabilidad del tiempo entre llegadas debe ser entendida como el intervalo de tiempo entre llegadas consecutivas y puede ser representado así:

una persona observa la llegada de los clientes, a cada llegada de un cliente se registra el tiempo transcurrido desde que ocurrió la llegada del anterior cliente. Luego, de un buen tiempo de estar registrando se procede a clasificar y agrupar los datos obtenidos;

Así, la densidad de probabilidad de estas muestras caracteriza al proceso de llegadas, tal como se visualiza en la figura 7.

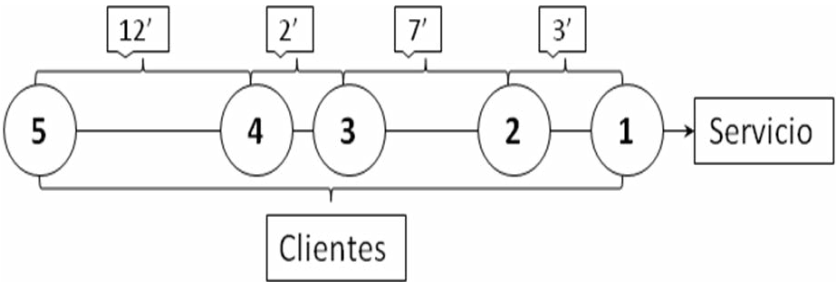


Figura 7. Esquema de densidad de tiempo de llegada
Fuente: Elaboración propia.

La función de densidad de probabilidad del tiempo empleado en prestar servicio.

Cada cliente requiere cierta cantidad de tiempo, el que precise el servidor para realizar el servicio que este cliente demanda.

Así, el tiempo de servicio requerido por cada cliente es tiempo de trabajo activo para el servidor y varía entre un cliente y otro, las necesidades de las personas son diferentes y dependen de sus circunstancias y contextos. Para ejemplificar lo expresado se propone la siguiente situación: en la caja de un supermercado un cliente puede presentar un carro lleno de artículos y el próximo cliente puede traer únicamente una lata de refresco, los tiempos de servicio varían notablemente entre un cliente y otro en función del número de artículos. Se detalla en la figura 8.

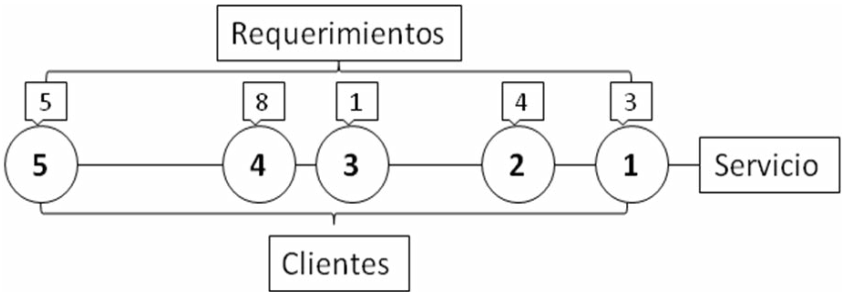


Figura 8. Esquema de densidad de tiempo de servicio
Fuente: Elaboración propia

El número de servidores se explica a través de los ejemplos siguientes: muchos bancos, por ejemplo, tienen una sola cola larga para todos sus clientes y cada vez que uno de los cajeros se libera el cliente que se encuentra primero en la cola se dirige a la caja que ha quedado libre; a este sistema se le denomina sistema de cola multiservidor. Se refleja en las figura 9.

En otros bancos cada cajero tiene su propia cola; en este caso se tiene un conjunto de colas independientes de un solo servidor. Se muestra en la figura 10.

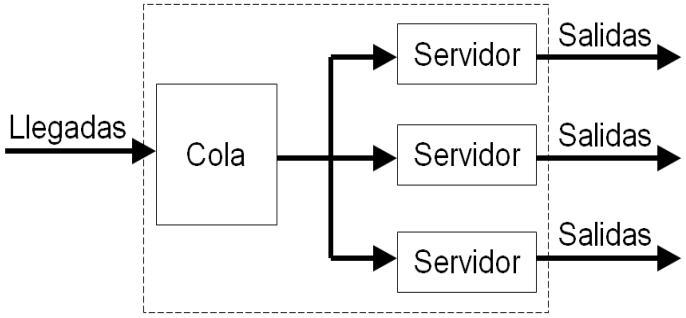


Figura 9. Esquema de cola multiservidor

Fuente: Caba, Chamorro y Fontalvo (2011:109).

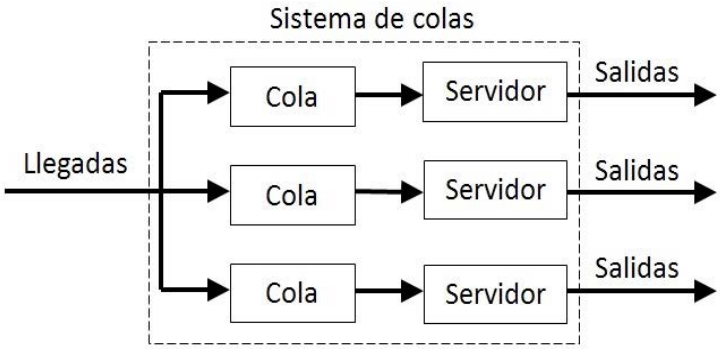


Figura 10. Esquema de colas monoservidor

Fuente: Caba, Chamorro y Fontalvo (2011: 109).

La disciplina de una cola, viene a describir el orden según el cual los clientes van siendo atendidos. Existen varias modalidades, las cuales dependen del tipo de bien o de servicio que facilite la organización en cuestión. Así se tiene que:

- En los supermercados utilizan el método de servir primero al cliente que ha llegado antes a la caja o servicio de cobro.

- En las salas de emergencia de los hospitales o centro de atención de salud de emergencia se emplea el criterio de atender primero al que esté más grave, o donde la vida del paciente se ve comprometida en lo inmediato.

- En una fotocopiadora se puede atender primero al que tenga menor trabajo, esto es: entra primero el que tenga que hacer menos fotocopias.

Al respecto Caba, Chamorro y Fontalvo (2011) resaltan que

> Es cierto que todas las colas tienen límites en el tamaño, pero este límite no las desanima o evita las llegadas, [aunque] puede ignorarse. La disciplina de la cola es primero en llegar primero en ser servido sin prioridades especiales. También se supone que las llegadas no pueden cambiar lugares en la línea o dejar la cola antes de ser servidas. (p.111-112).

Render, Stair y Hanna (2018: 502) desarrollan la explicación al establecer que la disciplina de la cola se relaciona con la "regla con la cual los clientes que están en la línea van a recibir el servicio." Así, se cuenta como varias disciplinas, a saber:

Primeras entradas, primeras salidas (PEPS), también llamada FIFO (*first in first out*, por sus siglas en inglés), que adicionalmente puede ser denominada FCFS (*first come first served*, por sus siglas del inglés y traducen: primero en llegar, primero en ser servido); lo cual deviene del tipo de organización en donde se organiza la cola.

Últimas entradas, primeras salidas (UEPS), conocida como LIFO (*last in first out*, por sus siglas en inglés, que traducen último en entrar primero en salir), también conocida como LCFS (*last come first served*, por sus siglas del inglés, que traducen último en llegar, primero en ser atendido). Esta disciplina procura atender primero al cliente que ha llegado el último; lo cual es consecuencia del tipo de organización que permite el surgimiento de la cola.

Selección aleatoria de servicio (SAS), que además recibe el nombre de (RSS) (*random selection of service*, por sus siglas en inglés). A esta disciplina también se le llama SIRO (*service in random order, por sus siglas del inglés,* que significan servicio en orden aleatorio). Esta disciplina de cola implica la selección del o de los clientes que son y serán atendidos de forma aleatoria; esta también es producto del tipo de organización por la cual se origina la cola. En la figura 11 se muestran en conjunto los diferentes tipos de disciplina en las colas.

Se puede presentar la situación que los clientes sean colocados en líneas de espera con prioridad, se les otorga la categoría de preferentes, priorizados o prioritarios y reciben atención preferencial. Esto no es un factor limitante para que les sea aplicada una disciplina de servicio en particular por el número de usuarios que lleguen para recibir el servicio/bien.

Otro elemento que cabe mencionar, refiere Taha (2017), es el comportamiento de los usuarios, clientes o demandantes en las colas; debido a que,

> Los clientes pueden cambiarse de una cola más larga a una más corta para reducir el tiempo de espera, pueden desistir del todo de hacer cola debido a la larga tardanza anticipada, o salirse de una cola porque han estado esperando demasiado. (p.595).

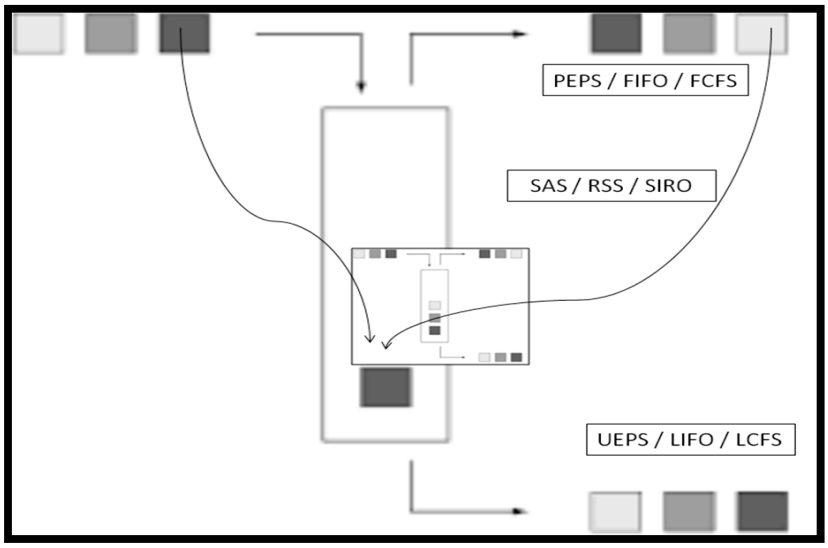


Figura 11. Disciplina en las colas
Render, Stair y Hanna (2018: 502).

Mecanismo de servicio, es entendido, aseguran Hieller y Lieberman (2021: 710), como “El tiempo que transcurre desde el inicio del servicio para un cliente hasta su terminación en una estación”; se refleja en la figura 12.

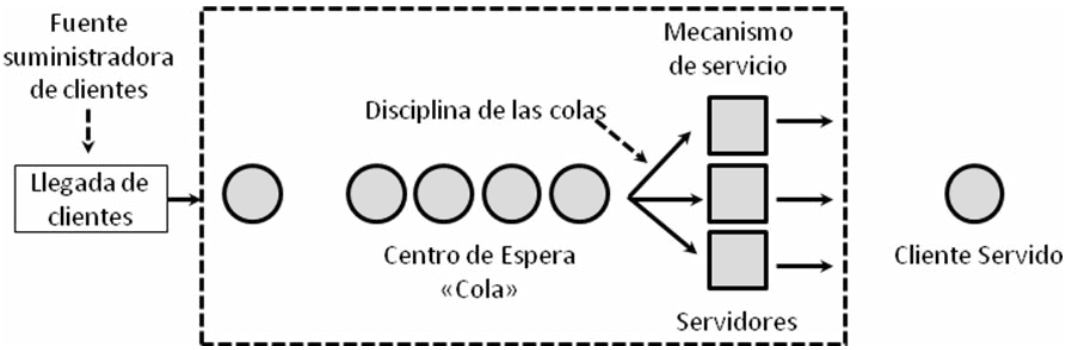


Figura 12. Mecanismo de servicio de las colas

Fuente: Adaptación de Villalobos (2014: párr. 3).

Capacidad de la cola, según Santiago (2017: párr. 25), "Es el máximo número de clientes que pueden estar haciendo" fila. Ello se debe, fundamentalmente, a que los espacios, instalaciones, unidades estructurales o los lugares que se tengan previstos para que alberguen las colas que se desarrollen tienen una capacidad de alojamiento finita o limitada. Esto influye decisivamente en el tamaño o la capacidad de las colas, pudiéndose tomar varias acciones: una, limitar la entrada al recinto; dos, establecer cuotas de recepción; tres, priorizar la admisión para la recepción de clientes o usuarios. De ahí que se tiende a ver las colas en el mundo real como finitas y lleva a Santiago (2017: párr.26) a indicar que "Es el máximo número de clientes que pueden estar haciendo cola (antes de comenzar a ser servidos)." Situación frecuente, común e, incluso, obligatoria en la atención al público en recintos cerrados durante la pandemia de Covid-19. Se reseña en la figura 13.

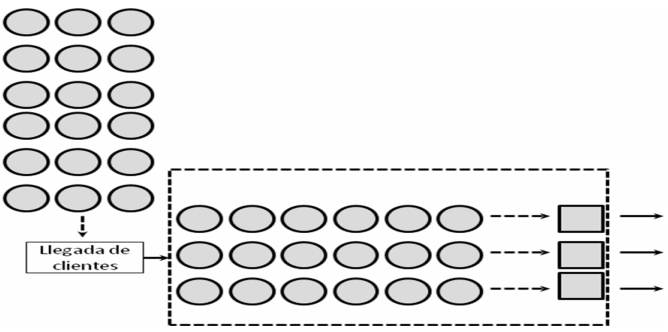


Figura 13. Capacidad de las colas

Fuente: Elaboración propia.

Otro factor de los modelos de espera es el diseño de la instalación y la ejecución del servicio, al respecto De La Fuente y Pino (2007) señalan que el lugar donde se presta el servicio o donde se genera la espera puede implicar más de un servidor, por ello es posible atender tantos clientes en forma simultánea o paralela como número de servidores existan, la mejor forma de entenderlo es a través de los cajeros de los bancos. En estas circunstancias todos los servidores ofrecen el mismo servicio y por esto se habla de servidores paralelos.

Existen otras situaciones, aseguran De La Fuente y Pino (2007), donde el lugar puede implicar un número de estaciones en serie por las que pueden o deben pasar los clientes antes que se complete el servicio. La manera más gráfica de entenderlo sería la producción en línea de un automóvil o el procesamiento de un producto en un conjunto de máquinas. De esto se derivan bien sea las líneas de espera en serie o las líneas de espera sucesivas, bien las líneas de espera en paralelo, bien las líneas de espera en red. Su graficación se expone en la figura 14.

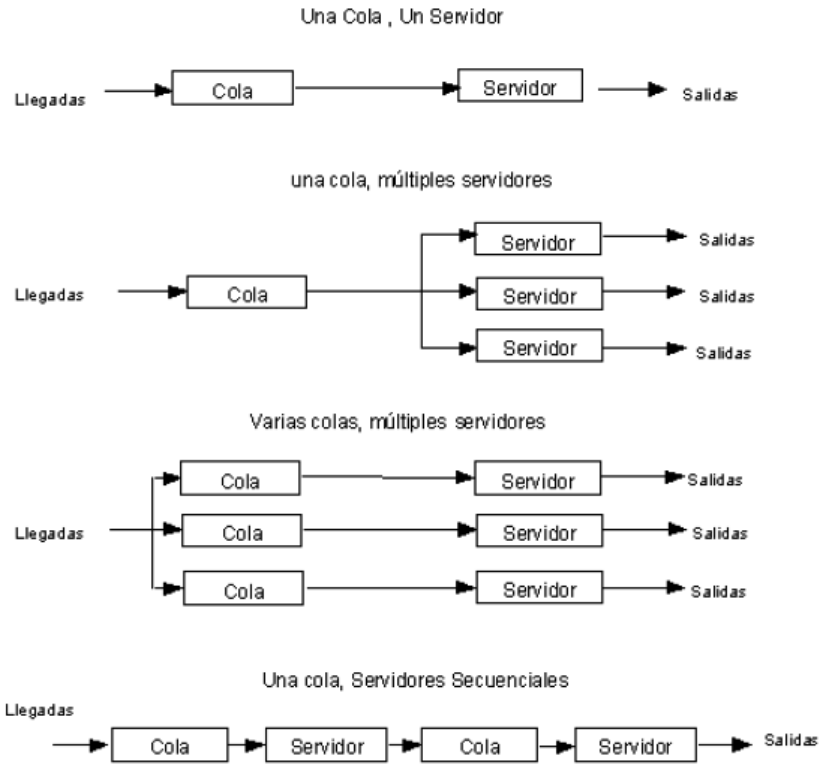


Figura 14. Tipos de colas con número de estaciones

Fuente: De La Fuente y Pino (2007).

De acuerdo con el Fuente: (2007) otro factor de los modelos de espera es la fuente de llamadas, esta puede suscitar un número finito o infinito de clientes.

Se habla de número finito de clientes cuando una llegada afecta la tasa de llegada de nuevos clientes, así por ejemplo, en un supermercado con M cajeras, la fuente de llamadas antes que cualquier cajera se levante a tomar su hora de almuerzo implica M clientes en potencia.

Cuando una cajera se toma su hora de almuerzo se convierte en un cliente y, por ende, no puede generar nuevas llamadas hasta que no se reincorpore a su puesto de trabajo.

No obstante, insiste Equipo Vértice (2007), se debe diferenciar entre la situación de la oficina y otros donde la causa para generar llamadas está limitada, aunque puede generar un número infinito de llegadas.

Tal es el caso de un departamento de publicaciones cualquiera, el número de usuarios es finito, aunque cada usuario puede generar un número ilimitado de llegadas; puesto que, en términos generales un usuario o cliente no necesita esperar a que se concluya el material entregado con anterioridad antes de generar otro nuevo.

Sarabia (1996) apunta que cuando la cola se compone de objetos inanimados que esperan algún tipo de procesamiento, el problema es básicamente económico, cuánto equipo hay que comprar y otras preguntas similares; en cambio, cuando la cola está formada por personas que esperan un servicio, el problema tiene aspectos psicológicos además de los económicos, que son bastante difíciles de cuantificar.

Asimismo, se habla que para las líneas de espera que están formadas por seres humanos las rigen las Leyes de Harper:

Primera Ley: *No importa en qué cola se sitúe: La otra siempre avanzará más rápido*.

Segunda Ley: *Y si se cambia de cola, aquélla en la que estaba al principio empezará a ir más deprisa*.

Por todo esto, los modelos de teorías de cola que se diseñen deben tomar en cuenta el efecto que produce la conducta humana en el comportamiento general bien del sistema, bien del modelo. Debido a que, se pueden presentar los siguientes casos:

- un cliente puede cambiarse de una línea de espera a otra pensando que reducirá su tiempo de espera,

- otros clientes pueden evitar formar una línea de espera al presumir anticipadamente una demora apreciable en la línea de espera,

- otros usuarios o clientes, sencillamente, pueden renunciar o dejar de esperar luego de estar algún tiempo en la fila pensando que ha pasado mucho tiempo o la espera es aún mayor.

Todas las cualidades o características humanas sólo pueden ser consideradas en la medida que sean tanto cuantificables como comunes.

En resumen, como bien lo indica Winston (2005), se puede decir que un factor determinante en el análisis de problemas de colas lo constituye la característica de las llegadas de los elementos al sistema de colas. Así, existen cuatro características que determinan el tipo de llegada al sistema:

- Estructura de la llegada: controlable o incontrolable.

- Tamaño de las unidades de llegada, de uno en uno, en lote.

- Patrón de distribución, tiempo entre llegadas constante o de acuerdo a alguna distribución estadística, como la de Poisson, la exponencial o la de Erlang.

- Nivel de paciencia, si la llegada permanece en la cola o se va.

Todos estos elementos se presentan en la figura 15.

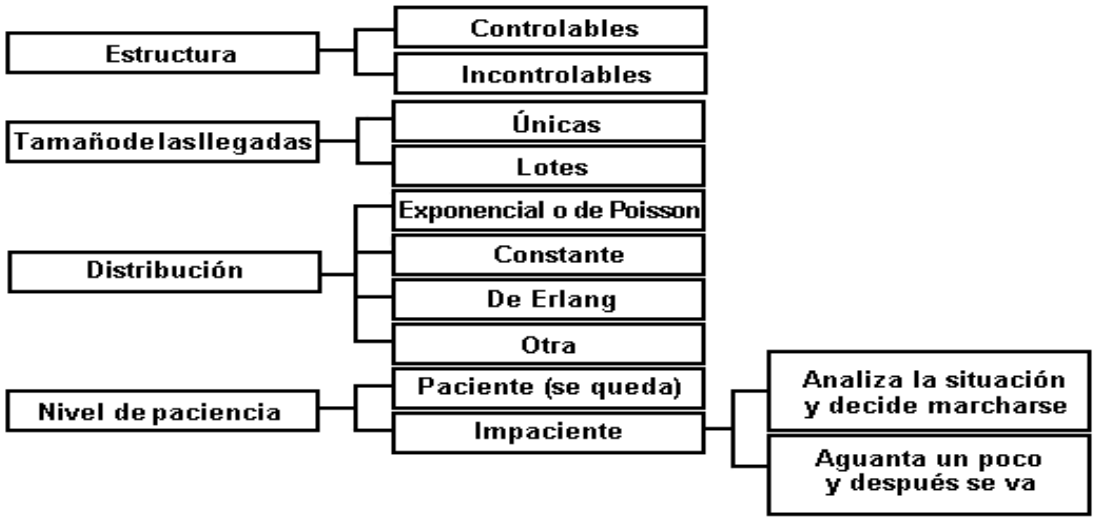


Figura 15. Características de las llegadas de los sistemas de colas

Fuente: Winston (2005: 1063).

COSTOS EN LOS SISTEMAS DE COLAS

Medida del Tiempo en los Sistemas de Colas

Los sistemas de colas puede apreciarse a partir de sus elementos más significativos, saber: la cola y la instalación de servicio. En razón de ello, Carro (2014: 357), teniendo presente que las colas se inician con las llegadas de usuarios, clientes o demandantes, donde "Las llegadas son las unidades que entran en el sistema para recibir el servicio. Siempre se unen primero a la cola". Luego de la llegada de los clientes, estos se organizan según la disciplina de cola pautada –PEPS, UEPS, SAS–. Por último, cumplido el servicio, satisfecho la necesidad, llenada la expectativa, cubierta la demanda se ha completado el servicio y ahora las llegadas se transforman en salidas y, de ese modo, continua el proceso con el resto de los demandantes en la línea de espera. A ellas se alude en la figura 16.

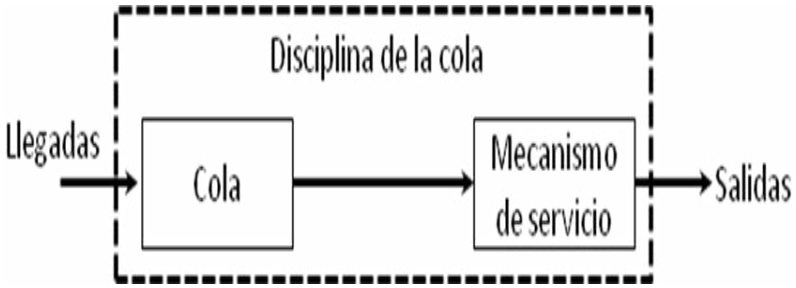


Figura 16. Elementos de los sistemas de colas
Fuente: Carro y González (2015: 357).

Así pues, asegura Carro (2014), al estudiar la operación de una instalación de servicio en condiciones aleatorias se busca asegurar que este cumpla con algunas características que midan el desempeño de dicho sistema. En tal sentido, las medidas de desempeño son varias, a saber:

- Tiempo de espera del cliente, es el tiempo que se calcula esperará un cliente antes de ser atendido, el cual está estrechamente relacionada con la percepción del cliente.

- Porcentaje de tiempo que no se utiliza en la instalación de servicio, se evalúa el grado de uso de la instalación.

De ese modo, cuanto mayor sea el tiempo de espera del cliente tanto menor es el porcentaje de tiempo que se mantendría ociosa la instalación y al contrario.

$$> \text{Tiempo de espera} \quad < \text{Tiempo de ocio}$$

$$< \text{Tiempo de espera} \quad > \text{Tiempo de ocio}$$

Explica Taha (2017) se presume que no se pierde tiempo entre el momento en que un cliente ya atendido sale de la instalación y la admisión de un nuevo cliente de la línea de espera. Así mismo, siguiendo a Taha (2017), en los modelos de espera la interacción entre el cliente y el servidor sólo es de interés si se relaciona con el periodo que necesita el cliente para completar su servicio; por lo tanto, de las llegadas de los clientes interesan los intervalos de tiempo que separan llegadas sucesivas y del servicio es importante el tiempo se servicio por cliente. De ahí, que en los modelos de espera, las llegadas y los tiempos de servicio se sintetizan en distribuciones de probabilidades; estas distribuciones pueden representar situaciones donde llegan clientes y son atendidos individualmente –caso de los bancos y de los supermercados– o donde llegan clientes y son atendidos en grupos –caso de los restaurantes–.

Tipos de Costos en las Líneas de Espera

Para Anderson, Sweeney, Williams, Camm y Kipp (2011), se tiene que:

1. Costo de Espera. Debido a que esperar significa desperdicio de algún recurso activo que bien se podría aprovechar en otra cosa. El coste medio de una cola por unidad de tiempo está dado por:

$$\mathbf{C \times L}$$

Donde:

C = Costo de espera por cliente y unidad de tiempo

L = Número promedio de clientes en cola.

2. Costo de Servicio. Está asociado a la compra de las instalaciones de servicio, así como los gastos de ponerlas en uso como pueden ser los gastos de mantenimiento y personal.

3. Sistema de costo mínimo o costos totales del sistema de servicio.

Aquí hay que tomar en cuenta que tasas bajas de servicio normalmente darán lugar a largas colas y costos de espera muy altos.

Conforme aumenta el servicio disminuyen los costos de espera, pero aumenta el costo de servicio; entonces, el propósito es encontrar el balance adecuado para que el costo total sea el mínimo.

Se obtiene de la suma de los dos costos anteriores –de espera y de servicio–. Esta sumatoria da como resultado una función de costos totales del sistema en función de la capacidad que tendrá una forma similar a la que se hace referencia en la figura 17.

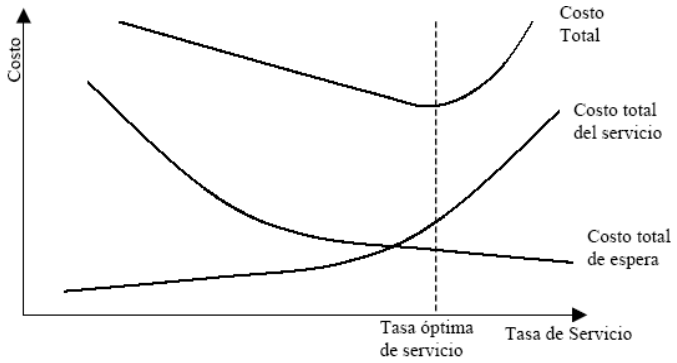


Figura 17. Costos de los sistemas de colas

Fuente: Anderson, Sweeney, Williams, Camm y Kipp (2011: 673)

Medidas de Rendimiento para Evaluar un Sistema de Colas

De acuerdo con Leandro (2004) el objetivo último de la teoría de colas consiste en responder cuestiones administrativas pertenecientes al diseño y a la operación de un sistema de colas. El gerente de un banco puede querer decidir si programa tres o cuatro cajeros durante la hora de almuerzo; en una estructura de producción, el administrador puede desear evaluar el impacto de la compra de una nueva máquina que pueda procesar los productos con más rapidez.

Cualquier sistema de colas pasa por dos fases básicas. Por ejemplo, cuando el banco abre en la mañana no hay nadie en el sistema, de modo que el primer cliente es atendido de forma inmediata. Conforme van llegando más clientes, lentamente se va formando la cola y la cantidad de tiempo que tienen que esperar se empieza a incrementar. A medida que avanza el día, el sistema llega a una condición en la que el efecto de la falta inicial de clientes ha sido eliminado y el tiempo de espera de cada cliente ha alcanzado niveles bastante estables.

La fase inicial que conserva los efectos de las condiciones iniciales se conoce como fase transitoria; después que los efectos de las condiciones son eliminados el sistema entra en un estado estable. Por ello, se plantean medidas de rendimiento que buscan evaluar el nivel de respuesta del sistema en estado estable. Para diseñar y poner en operación un sistema de colas, por lo general, los administradores se preocupan por el nivel de servicio que recibe un cliente, así como por el uso apropiado de las instalaciones de servicio de la empresa.

Algunas de las medidas que se utilizan para evaluar el rendimiento surgen de hacerse las siguientes preguntas:

Preguntas relacionadas con el tiempo, centradas en el cliente, como:

a. ¿Cuál es el tiempo promedio que un cliente recién llegado tiene que esperar en la fila antes de ser atendido?

La medida de rendimiento asociada es el tiempo promedio de espera, representado con W_q.

b. ¿Cuál es el tiempo que un cliente invierte en el sistema entero, incluyendo el tiempo de espera y el de servicio?

La medida de rendimiento asociada es el tiempo promedio en el sistema, denotado con W.

Preguntas cuantitativas relacionadas al número de cliente, como:

a. En promedio ¿cuántos clientes están esperando en la cola para ser atendidos?

La medida de rendimiento asociada es la longitud media de la cola, representada con L_q.

b. ¿Cuál es el número promedio de clientes en el sistema?

La medida de rendimiento asociada es el número medio en el sistema, representado con L.

Preguntas probabilísticas que implican tanto a los clientes como a los servidores, por ejemplo:

a. ¿Cuál es la probabilidad que un cliente tenga que esperar a ser atendido?

La medida de rendimiento asociada es la probabilidad de bloqueo, que se representa por p_w.

b. En cualquier tiempo particular, ¿cuál es la probabilidad que un servidor esté ocupado?

La medida de rendimiento asociada es la utilización, denotada con U, esta medida indica también la fracción de tiempo que un servidor está ocupado.

c. ¿Cuál es la probabilidad que existan n clientes en el sistema?

La medida de rendimiento asociada se obtiene calculando la probabilidad P_o que no haya clientes en el sistema, la probabilidad P_i que haya un cliente en el sistema y así sucesivamente.

Esto tiene como resultado la distribución de probabilidad de estado, representada por P_n, n = 0,1...

d. Si el espacio de espera es finito, ¿Cuál es la probabilidad que la cola esté llena y un cliente que llega no sea atendido?

La medida de rendimiento asociada es la probabilidad de negación del servicio, representada por P_d.

Preguntas relacionadas con los costos, como:

a. ¿Cuál es el costo por unidad de tiempo por operar el sistema?

b. ¿Cuántas estaciones de trabajo se necesitan para lograr mayor efectividad en los costos?

Siguiendo a Leandro (2004), el cálculo específico de estas medidas de rendimiento depende de la clase de sistema de colas que se esté analizando. Algunas de estas medidas están relacionadas entre sí. Conocer el valor de una medida permita encontrar el valor de una medida relacionada. El cálculo de muchas de las medidas de rendimiento depende de los procesos de llegadas y de servicio del sistema de colas

en específico. Incluso sin conocer la distribución especifica, las relaciones entre algunas de las medidas de rendimiento pueden obtenerse para ciertos sistemas de colas, únicamente mediante el uso de los siguientes parámetros de los procesos de llegada y de servicio.

λ = número promedio de llegadas por unidad de tiempo.

μ = número promedio de clientes atendidos por unidad de tiempo en una sección

Se supone una población de clientes infinita y una cantidad limitada de espacio de espera en la fila, el tiempo total que un cliente invierte en el sistema es la cantidad de tiempo invertido en la fila más el tiempo durante el cual es atendido:

Tiempo promedio en el sistema	=	tiempo promedio de espera	+	tiempo promedio de servicio

El tiempo promedio en el sistema y el tiempo promedio de espera están representados por las cantidades W y W_q, respectivamente, el tiempo promedio de servicio puede expresarse en términos de parámetros de &. Por ejemplo, si & es cuatro clientes por hora, en promedio, cada cliente requiere un cuarto de hora para ser atendido.

En general, el tiempo de servicio es 1/&, lo cual conduce a la siguiente relación:

$$W = Wq + 1/\&.$$

Considerando la relación entre el número promedio de clientes en el sistema y el tiempo promedio que cada cliente pasa en el sistema, un cliente acaba de llegar y se espera que permanezca en el sistema un promedio de media hora, durante esta media hora otros clientes siguen llegando a una tasa X, doce por hora. Cuando el cliente en cuestión abandona el sistema, después de media hora, deja tras de sí un promedio de

(1/2)*12 = 6 clientes nuevos; es decir, en promedio, existen seis clientes en el sistema en cualquier tiempo dado.

En términos de & y de las medidas de rendimiento, entonces:

Número promedio de clientes en el sistema	=	número promedio de llegadas por unidad de tiempo	*	tiempo promedio en el sistema

De modo que: $L = \lambda * W$

Utilizando una lógica parecida se obtiene la relación entre el número promedio de clientes que esperan en la cola y el tiempo promedio de espera en la fila, de manera que:

$$Lq = \lambda * W_q$$

Análisis Económico de los Sistemas de Colas

Las realidades y análisis presentados muestran las ventajas de tener más de un servidor, a saber: la reducción del tiempo de espera y el número de clientes que esperan para ser atendidos. Claramente, mientras más servidores se tengan, mejor será el servicio a los clientes; sin embargo, cada servidor implica costos de operación.

¿De qué manera se equilibra el nivel de servicio y el nivel de costo? La mejor manera es ubicar el modelo de cola más adecuado para las actividades que se llevan a cabo dentro de las instalaciones estudiadas.

Características de los Modelos de Cola

Físicas

- Servidor: elemento que presta el servicio solicitado por los clientes.

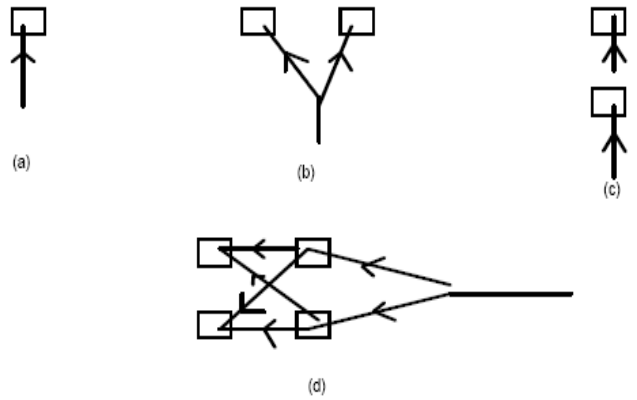
- Cola: elementos esperando recibir servicio.

- Sistema: incluye cola, servidor y el elemento que está siendo servido.

- Cadena: número de líneas de cola del sistema, los sistemas de colas son mono o multicadenas; en los casos más simples el número de cadenas es el número de servidores en paralelo.

- Número de fases: número de servicios diferentes que hay que esperar antes de completar el servicio total; en los casos más simples es el número de servidores en serie.

Se tratan de reproducir todos estos elementos en la figura 18.



a. Una cadena y una sola fase.
b. Multicadena y una sola fase.
c. Una cadena y multifase.
d. Multicadena y multifase.

Figura 18. Modelos de los sistemas de colas con cadenas
Fuente: Fonollosa, Sallán y Suñe (2005).

Fonollosa, Sallán y Suñe (2005) ofrecen las anteriores características en su obra, además para aclarar tales conceptos exponen los siguientes ejemplos:

Cadena (a) de la figura 19 corresponde a la taquilla de un cine.

Cadena (b) de la figura 19 concerniría a los cajeros en un banco.

Cadena (c) de la figura 19 sería una línea de montaje con distintos elementos que hay que fabricar.

Cadena (d) de la figura 19 vendría a ser una situación de automóviles esperando paso en distintos semáforos.

Funcionalidad

Con la funcionalidad Fonollosa, Sallán y Suñe (2005) hacen referencia a los elementos que afectan el funcionamiento o dinámica de los sistemas de líneas de espera, tal es el caso de: la distribución del intervalo de tiempo entre llegadas, la distribución de los tiempos de servicio, o tiempos empleados por el servidor para prestar los servicios requeridos por cada uno de los clientes, las distintas formas en que se reorganizan las colas en el supuesto que haya varias cadenas o varias fases y la disciplina de la cola que es la forma en que los clientes que están esperando acceden al servidor.

Usualmente se considera que el primero que ha llegado es el primero al que se le presta servicio; no obstante, en algunas circunstancias esto no es así. Otras disciplinas de colas pueden ser aleatorias, como la forma en que suben al tren los viajeros que esperan en una estación, en este caso, el orden de entrada depende de lo cerca que haya quedado la puerta de cada viajero.

También puede haber algunas prioridades en determinados servicios, se ha de considerar si existe abandono de la cola, es decir, elementos, clientes, demandantes o usuarios que al ver una cola demasiado larga no se deciden a esperar, o elementos, clientes, demandantes o usuarios que habiendo esperado un cierto tiempo no desean esperar más y abandonan la cola.

El sistema se dice que tiene una capacidad limitada si sólo admite, como máximo, un cierto número de elementos.

Parámetros de los Sistemas de Colas

De acuerdo con Rodríguez y Gámez (2002) si la cola es de comportamiento aleatorio no se puede saber exactamente la situación que se tendrá en cada momento; de ahí, que para describir su comportamiento se emplean promedios y probabilidades. Entre los parámetros más usuales se consideran los siguientes:

- Probabilidad que no haya elementos en la cola.
- Probabilidad que haya un cierto número de unidades en el sistema.
- Probabilidad que un elemento que llega tenga que esperar para recibir servicio.
- Número promedio de elementos en cola.
- Número promedio de elementos presentes en el sistema.
- Tiempo medio que ha de esperar cada elemento que accede a la cola.
-Tiempo promedio que un elemento pasa en el sistema.

Modelos de Llegadas y de Tiempo de Servicio

Para Zaragoza (2004) los clientes pueden acceder al sistema: bien de forma determinada, porque se sabe exactamente cuándo van a venir cada uno de ellos (por ejemplo, a intervalos de tiempo de 3 segundos). Bien puede ocurrir que los intervalos de llegada sigan una variable aleatoria, es decir, que aunque no se sepa en qué momento van a llegar cada uno de los elementos, si se conoce la distribución de probabilidad de los intervalos de tiempo entre llegadas consecutivas. En el primer caso se habla de distribución de llegada determinista, en el segundo caso se dice que los tiempos de llegada siguen una distribución aleatoria.

La distribución que se usa más frecuentemente para modelar los intervalos de tiempos entre dos llegadas consecutivas es la distribución exponencial.

Se presume que en un instante sólo puede haber una llegada, se anota para t_i –la hora a la que llega el cliente *i*–, por $T_i = t_i+1 - t_i$ el tiempo transcurrido entre dos llegadas consecutivas.

Se supone que los valores de T_i son independientes, que T_i es una variable continua y que el estado es estacionario; es decir, se admite la hipótesis que la distribución que modela la cola, –a saber la probabilidad que haya un cierto número de elementos en la cola– es la misma a todas las horas del día.

Normalmente esto no es estrictamente cierto, pero puede cumplirse aproximadamente considerando ciertos intervalos de horarios cada día.

Si se admite el modelo exponencial para la distribución de la variable aleatoria *T*

$$f(t) = \lambda \exp(-\lambda t); \quad \lambda > 0, \quad t > 0$$

La probabilidad que una llegada ocurra en un tiempo $t < c$ unidades después que la anterior es:

$$P(t < c) = \int_0^c f(t)dt = \int_0^c \lambda \exp(-\lambda t)dt$$

Se puede comprobar que la media de esta distribución es 1/λ, y la varianza $\frac{1}{\lambda^2}$. El parámetro λ hay que interpretarlo como el número promedio de elementos que llegan al sistema por unidad de tiempo.

Ejemplo tomado de Rodríguez y Gámez (2002), donde el número promedio de llegadas por hora al consultorio de un hospital es de 60 pacientes. Si acaba de llegar

un paciente, ¿cuál es la probabilidad que el siguiente venga dentro del siguiente minuto? ¿Y de que tarde más de 4 minutos?

$$Se\ toma\ \ para\ \ \lambda = \frac{60\,pacientes}{hora} = \frac{60\,pacientes}{60\,\mathrm{min}utos} = 1\frac{paciente}{\mathrm{min}utos}$$

$$P(t<1) = \int_0^1 e^{-t}dt = -e^{-t}\Big|_0^1 = 1 - e^{-1} = 0{,}632$$

$$P(t>4) = \int_4^\infty \exp(-t)dt = 0 - (-e^{-4}) = 0{,}0183$$

Es importante resaltar que la distribución exponencial cumple la siguiente relación:

$$P(t \le h) = P(t \le c + \frac{h}{t} \ge c)$$

Esta propiedad significa que en todo momento la probabilidad que el siguiente elemento venga en un intervalo de *h* segundos no depende del momento concreto *c*, sino exclusivamente del intervalo de tiempo *h* considerado. Esta probabilidad no cambia con el tiempo y es independiente de lo que haya pasado ante; por eso esta propiedad se suele enunciar diciendo que la función exponencial carece de memoria. Lo cual quiere decir que la distribución no guarda información sobre lo que ha pasado antes de *c* y, por tanto, no se necesita tener información del pasado para predecir el futuro.

Para ilustrarlo se toma un ejemplo de Taha (2017): una máquina en servicio tiene una unidad de reserva para sustituirla de inmediato cuando falle, el tiempo a la falla (tiempo entre fallas) de la máquina (o de su unidad de reserva) es exponencial y sucede cada 40 minutos en promedio. El operador de la máquina dice que ésta tiene la costumbre de descomponerse cada noche a eso de las 8:30pm. Al analizar lo que dice el operador se tiene que:

La tasa promedio de fallas de la máquina es $\lambda = \frac{60}{40} = 1{,}5$ fallas por hora; así la distribución exponencial del tiempo a la falla es

$$f(t) = 1{,}5e^{-1{,}5t}, \quad t > 0$$

En cuanto a lo que dice el operador, se sabe que no es correcto, porque se opone al hecho que el tiempo entre fallas es exponencial y, en consecuencia, es totalmente aleatorio. La probabilidad que una falla suceda a las 8:30pm no se puede usar para respaldar ni refutar esa afirmación, porque el valor de esa probabilidad depende de la hora del día (en relación con las 8:30pm) con la que se calcule. Por ejemplo, si ahora son las 8:20pm la probabilidad que lo que dice el operador sea cierta esta noche es

$$p\left\{t < \frac{10}{60}\right\} = 1 - e^{-1{,}5\left(\frac{10}{60}\right)} = 0{,}22$$

Baja, este valor indica que no se puede analizar la afirmación del operador con base en estimaciones de probabilidades y confiar en las características de la distribución exponencial (aleatoriedad total) para refutar la afirmación.

Relación entre las Funciones de las Distribución de Poisson y Exponencial

Acá se sigue a Taha (2017), se considera la situación de espera en la cual el número de llegadas y salidas durante un intervalo de tiempo es controlado por las siguientes condiciones:

Condición1: la probabilidad que un evento –llegada o salida– ocurra entre los tiempo t y $t + h$ depende únicamente de la longitud de h, lo que significa que la probabilidad no depende ni del número de eventos que ocurren hasta el tiempo t ni del valor específico del periodo $(0, t)$.

Matemáticamente se dice que la función de probabilidad tiene incrementos independientes estacionarios.

Condición 2: la probabilidad que ocurra un evento durante un intervalo de tiempo muy pequeño h es positiva, pero menor que 1.

Condición 3: cuando mucho puede ocurrir un evento durante un intervalo de tiempo muy pequeño h.

Las tres condiciones dadas describen un proceso donde el conteo de eventos durante un intervalo de tiempo dado, sigue la distribución de Poisson y que, equivalentemente, el intervalo de tiempo entre eventos sucesivos es exponencial. En tal caso se dice que las condiciones representan un proceso de Poisson.

$p_n(t)$ = probabilidad que ocurra n eventos durante el tiempo t.

Entonces, por la condición de 1, la probabilidad que no ocurra ningún evento durante el tiempo $t + h$ es

$$p_0(t+h) = p_0(t)p_0(h)$$

Para $h > 0$ y suficientemente pequeña, la condición 2 indica que $0 < p_0(h) < 1$. Bajo estas condiciones la ecuación anterior tiene la siguiente solución

$$p_0(t) = e^{-\alpha t}, \quad t \geq 0$$

Donde α es: para un proceso descrito por $p_n(t)$, el intervalo de tiempo entre eventos sucesivos es exponencial; usando la relación conocida entre las distribuciones exponencial y de Poisson se puede concluir que $p_n(t)$ debe ser tipo Poisson. Sea

$f(t)$ = función densidad de probabilidad (fdp) del intervalo de tiempo t entre la ocurrencia de eventos sucesivos, $t \geq 0$. *Se supone que* T es el

intervalo de tiempo desde la ocurrencia del último evento; entonces, es válido el siguiente enunciado probabilístico:

P {el tiempo entre eventos excede a T} = P {no ocurren eventos durante T}

En términos matemáticos esto se expresa como

$$\int_{T}^{\infty} f(t)dt = p_0(T)$$

Sustituyendo el valor de p_oT, se obtiene que:

$$\int_{T}^{\infty} f(t)dt = e^{-\alpha T}, \quad T > 0$$

O bien

$$\int_{0}^{T} f(t)dt = 1 - e^{-\alpha T}, \quad T > 0$$

Derivando ambos miembros de la ecuación respecto a T, se obtiene:

$$f(t) = \alpha e^{-\alpha t}, \quad t \geq 0 \quad \text{(exponencial)}$$

Una distribución exponencial con media E{t} = 1/α unidades de tiempo. Como *f(t)* es una distribución exponencial, la teoría de probabilidades indica que $p_n(t)$ debe ser una distribución de Poisson.

$$p_n(t) = \frac{(\alpha t)^n e^{-\alpha t}}{n!}, \quad n = 0, 1, 2, \ldots \text{(Poisson)}$$

El valor medio de n durante un periodo dado t es $E\{n \mid t\} = \alpha t$ eventos. Esto significa que α representa la tasa a la que ocurren los eventos.

La conclusión es que si el intervalo de tiempo entre eventos es exponencial con media1/α la cantidad de llagadas durante un periodo *t* específico tiene distribución de Poisson con media αt.

El proceso de Poisson es un proceso completamente aleatorio porque tiene la propiedad que el intervalo de tiempo que permanece hasta la ocurrencia del próximo evento es totalmente independiente del intervalo de tiempo que ha transcurrido desde la ocurrencia del último evento. Esta probabilidad equivale a demostrar el siguiente enunciado de probabilidad:

$$P\{t > T + S | t > S\} = P\{t > T\}$$

Donde *S* es el intervalo de tiempo desde la ocurrencia del último evento, como *t* es exponencial se tiene que:

$$P\left\{t > T + S | t > S\right\} = \frac{P\{t > T + S | t > S\}}{P\{t > S\}} = \frac{P\{t > T + S\}}{P\{t > S\}}$$

$$= \frac{e^{-\alpha(T+S)}}{e^{-\alpha S}} = e^{-\alpha T}$$

$$= P\{t > T\}$$

Es la propiedad de olvido o falta de memora de la distribución exponencial y es la base para demostrar que la distribución de Poisson es totalmente aleatoria. Otra característica que distingue a la de Poisson es que es la única distribución cuya media y varianza son iguales.

Para explicar lo dicho hasta ahora se tomará un ejemplo del mismo autor, Taha (2017). Una máquina de servicio automático tiene siempre una unidad de reserva para su reemplazo inmediato en caso de falla. El tiempo para que ocurra la falla de la máquina o de su unidad de reserva es exponencial con media de 10 horas. Así, las

fallas ocurren a razón de 0,1 eventos por hora. La distribución exponencial del tiempo para que ocurra la falla está dada por:

$$f(t) = 0{,}1e^{-0{,}1t}, \quad t > 0$$

En tanto que la distribución de Poisson del número de fallas en un periodo T está dada por:

$$p_n(T) = \frac{(0{,}1T)^n e^{-0{,}1T}}{n!}, \; n = 0, 1, 2, \ldots$$

Interesa calcular la probabilidad que ocurra una falla en un intervalo de 5 horas. Esta probabilidad está dada por:

$$P\{t < 5\} = \int_0^5 f(t)dt$$

Alternativamente, la probabilidad que una falla ocurra después de 6 horas, a partir de ahora, considerando que la última falla tuvo lugar 3 horas antes, utilizando la propiedad de olvido de la distribución exponencial , sigue que:

$$P\{t > 9 \mid t > 3\} = P\{t > 6\} = e^{-0{,}1*6} = 0{,}549$$

La relación entre las distribuciones de Poisson y exponencial se demuestra calculando la probabilidad que ninguna falla tendrá lugar durante un periodo de 1 día (24 horas), o sea

$$p_0(24) = \frac{(0{,}1*24)^0 e^{-0{,}1*24}}{0!} = e^{-2{,}4} = 0{,}091$$

Se observa que $p_0\,(24)$ es equivalente a tener un tiempo entre fallas de 24 horas por lo menos, o sea:

$$P\left\{t > 24\right\} = \int_{24}^{\infty} 0{,}1e^{-0{,}1t}\,dt = e^{-2{,}4}$$

Modelo de Nacimiento Puro

Para explicar este modelo se acude a Cao (2002). En este modelo los clientes llegan y nunca parten, siendo un proceso completamente aleatorio; siendo el mejor ejemplo una oficina de registro civil cualquiera. Se supone que λ es el proceso de nacimiento puro de tener n arribos o llegadas durante el periodo de tiempo:

$$p_n(t) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}, \quad n = 0, 1, 2, \ldots \text{ (nacimiento puro)}$$

Donde λ es la tasa de llegadas por unidad de tiempo con el número esperado de llegadas durante t igual a λt.

Para explicarlo mejor se expone una adaptación de un ejemplo del mismo autor. Se supone que los nacimientos en Canaguá, Mérida, Venezuela están separados en el tiempo, de acuerdo con una distribución exponencial, presentándose un nacimiento cada 7 minutos en promedio. Como el tiempo promedio entre arribos es de 7 minutos, la tasa de nacimientos en la ciudad se calcula como

$$\lambda = \frac{24 * 60}{7} = 205{,}7 nacimientos / día$$

El número de nacimientos en la ciudad por año está dado por:

□ = (20577*365) = 75.080 nacimientos/año

La probabilidad de ningún nacimiento en cualquier día es:

$$P_0(1) = \frac{(205{,}7/1)^0 e^{-205{,}7*1}}{0!} \cong 0$$

Se supone que interesa emitir 45 actas de nacimiento al final de un período de 3 horas, si se pudieron emitir 35 actas en las primeras dos horas; se observa que debido a que los nacimientos ocurren según un proceso de Poisson, la probabilidad requerida se reduce a tener 45 - 35 = 10 nacimientos en hora (= 3 –2). Dado $\lambda = 60 / 7 = 8,57$ nacimientos por hora, se obtiene:

$$P_{10}(1) = \frac{(8,57 * 1)^{10} e^{-8,57*1}}{10!} = 0,11172$$

Modelo de Muerte Pura

En este modelo los clientes se retiran de un abasto inicial, para entender el mismo se seguirá al manual de dirección de operaciones del Equipo Vértice (2007). Se considera la situación de almacenar *N* unidades de un artículo al inicio de semana para satisfacer la demanda de los clientes en la misma. Se presume que la demanda se presenta a una tasa de μ unidades por día y que el proceso de demanda es completamente aleatorio, la probabilidad asociada de tener *n* artículos en almacén después de un tiempo *t*, la da la siguiente distribución truncada de Poisson:

$$p_n(t) = \frac{(\mu t)^{N-n} e^{-,\mu t}}{(N-n)!}, \quad n = 1, 2, \ldots N$$

$$p_0(t) = 1 \sum_{n=1}^{N} p_n(t)$$

Un ejemplo dado en el Manual del Equipo Vértice (2007) es el siguiente: La sección de florería en un supermercado tiene 18 docenas de calas al iniciar cada semana, en promedio el florista vende 3 docenas por día (una docena cada vez), pero la demanda sigue en realidad una distribución de Poisson; siempre que la existencia llega a 5 docenas, o menos, se coloca un pedido nuevo de 18 docenas para entregar al principio de la semana que entra. Por la naturaleza de la mercancía todas las calas que quedan al final de la semana se desechan; determinar lo siguiente: la probabilidad

de colocar un pedido en cualquier día de la semana y la cantidad promedio de docenas de calas que se desechan al final de la semana.

Como las compras se hacen con una frecuencia de μ =3 docenas diarias, la probabilidad de colocar un pedido al final del día *t* es

$$p_{n\leq 5}(t) = p_0(t) + p_1(t) + ... + p_5(t)$$

$$pn \leq 5(t) = po(t) + \sum_{n=1}^{5} \frac{(3t)^{18-n} e^{-3t}}{(18-n)!}, t = 1,2,...,7$$

t(días)	1	2	3	4	5	6	7
μ*t*	3	6	9	12	15	18	21
$p_{n\leq 5}(t)$	0,0000	0,0088	0,1242	0,4240	0,7324	0,9083	0,9755

La cantidad promedio de docena de calas desechadas al final de la semana *(t = 7),* se calcula como sigue:

$$E\{n \mid t = 7\} = \sum_{n=0}^{18} np_n(7) = 0,664 docena$$

Distribución de Erlang

Para explicarla se sigue a Winston (2005), quien dice que a veces se modelan los intervalos de llegadas con una distribución de Erlang, la función de densidad de esta distribución viene dada por dos parámetros:

$$f(t) = \frac{R(Rt)^{h-1}\exp(-Rt)}{(k-1)!}, \quad t \geq 0, \text{ donde } E(t) = \frac{k}{R} \text{ y } Var(t) = \frac{k}{R^2}$$

Si tomamos R = Kλ, tenemos esta otra expresión para la función de densidad:

$$f(t) = \frac{k\lambda(k\lambda t)^{k-1}\exp(-k\lambda t)}{(k-1)!},\ t \geq 0,\ \text{siendo en este caso}\ E(t) = \frac{1}{\lambda}\ \text{y}\ Var(t) = \frac{k}{(k\lambda)^2} = \frac{1}{k\lambda^2}$$

Si $k=1$, la distribución es una exponencial de parámetro λ. La representación gráfica de la función puede tomar muy diversas formas para los distintos valores de los dos parámetros, por lo que es adaptable a distintas situaciones reales.

Puede demostrarse que la distribución de Erlang es la distribución de la suma de k variables exponenciales independientes del parámetro λ; por tanto, cuando los intervalos entre llegadas consecutivas se modelan con una función exponencial de parámetro λ, el intervalo entre k llegadas consecutivas sigue una distribución Erlang de parámetros k y λ. Se visualiza en la figura 19.

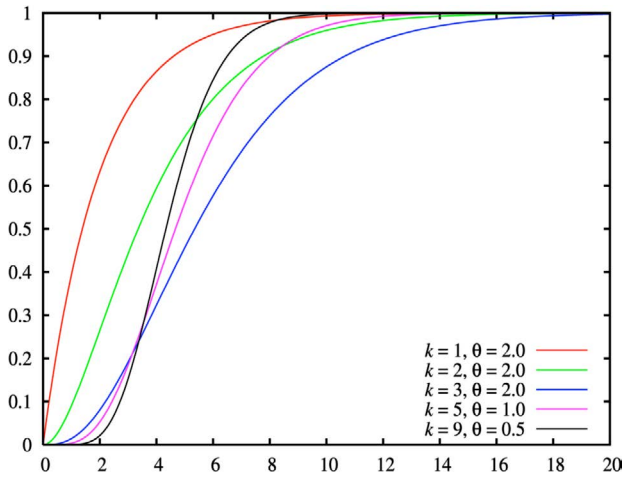


Figura 19. Función de distribución de probabilidades de Erlang

Fuente: Wiston (2005: 1059).

Colas Especializadas de Poisson y Notación de Kendall

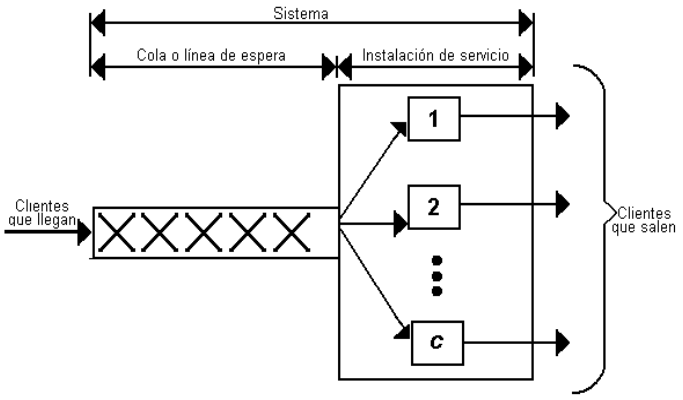


Figura 20. Modelo de colas en paralelo
Fuente: De La Fuente y Pino (2007: 97).

Para explicitar este punto se utiliza a De La Fuente y Pino (2007) y se reseña en la figura 20. Ellos puntualizan que las colas especializadas de Poisson se refieren a líneas de espera donde los clientes son atendidos por *c* servidores en paralelo, de manera que se pueda dar servicio a *c* clientes al mismo tiempo. Todos los servidores ofrecen servicios iguales, el número de clientes en el sistema en cualquier punto se define como el que incluye a aquellas personas que están en línea de espera y en servicio.

La anotación que se utiliza y resume las características del modelo de cola en paralelo es la de Kendall, aunque él solo creo la primera parte *(a/b/c)* en 1953; Lee en 1966 le agregó los símbolos *d* y *e* y, finalmente, Taha en 1968 le agregó el símbolo *f*; quedando definitivamente como sigue:

$$(a/b/c):(d/e/f)$$

Donde los símbolos representan:

a ≡ distribución de llegadas

b ≡ distribución de salidas o de tiempo de servicio

c ≡ número de servidores en paralelo

d ≡ disciplina de servicio

e ≡ número máximo admitido en el sistema

f ≡ tamaño de la fuente de llamadas

En la figura 21 se expone la notación de Kendall.

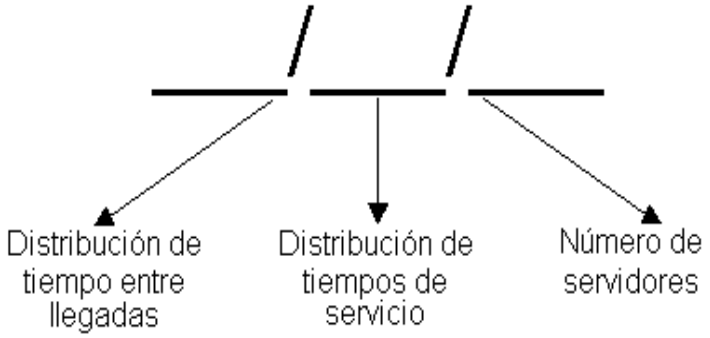


Figura 21. Notación de Kendall
Fuente: De La Fuente y Pino (2007: 99).

Los símbolos de llegada *(a)* y salida *(b)* pueden ser sustituidos por los códigos siguientes:

M ≡ distribución de llegadas o salidas de Poisson (o de Markov); o intervalos entre llegadas independientes e idénticamente distribuidos que se rigen por la distribución exponencial.

D ≡ tiempo entre llegadas o de servicio constante o determinista.

E_k ≡ distribución de Erlang o θ de la distribución de tiempo entre llegadas o de servicios con parámetro *k*.

GI ≡ distribución de llegadas general independiente, o tiempo entre llegadas.

G ≡ distribución de salidas general, o tiempo de servicio.

Si un sistema de cola se representa con el esquema M/M/1/fcfs/∞/∞ significa: que los intervalos entre llegadas consecutivas y los tiempos empleados en prestar el servicio demandado se distribuyen con distribuciones exponenciales; que hay un solo servidor; que la disciplina de cola consiste en atender primero al que haya llegado antes al sistema; que el sistema puede recibir un número ilimitado de individuos; y, que el número de clientes potenciales es infinito (muy grande).

Modelo Generalizado de Poisson

Para entender el modelo generalizado de Poisson se debe saber que significa tasas de llegada y servicio dependientes del estado, de la cantidad de clientes en la instalación de servicios; para ello se seguirá a Fonollosa, Sallán y Suñe (2005). Según los autores la mejor forma de entenderlo es a través de la ejemplificación; así pues, considérese un taller de maquinaria con un total de N máquinas, la tasa de falla de las máquinas es una función del número de máquinas en condiciones de trabajo. Esto es, si λ es la tasa de falla por máquina, la tasa de falla en todo el taller donde $(n \leq N)$ máquinas están en estado operativo, es $n\,\lambda$. De manera similar, si un sistema tiene c servidores en paralelo y μ es la tasa de servicio por servidor, entonces, dado que n es el número de clientes en el sistema, tanto en espera como en servicio, la tasa de salidas del sistema entero es n_μ si $n < c$ y c_μ si $n \geq c$.

El ejemplo demuestra cómo el modelo generalizado de líneas de espera, las tasas de llegadas y salidas pueden ser funciones del estado del sistema representado por el número de clientes n; por ello, se usa la notación λ_n y μ_n para definir las tasas de llegadas y salidas como una función de n.

El objetivo es deducir una expresión para p_n, que es la probabilidad de estado estable de n clientes en el sistema, como una función de λ_n y μ_n.

La deducción de una expresión para p_n se logra empleando el diagrama de tasa de transición.

Dado que hay *n* clientes en el sistema en cualquier tiempo *t*, el número de clientes en el sistema al final de un intervalo de tiempo *h* suficientemente pequeño, serán *n – 1 o n + 1*, dependiendo de si una salida o llegada tiene lugar durante el intervalo *h*, la probabilidad que ocurra más de un evento durante *h* tiende a 0 cuando $h \rightarrow 0$.

Se dice que en un proceso de Poisson el estado *n* sólo se puede comunicar con los estados *n – 1 y n + 1*

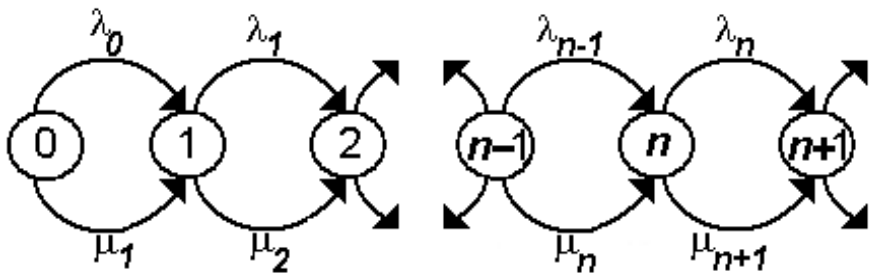


Figura 22. Tasas de transición
Fuente: Fonollosa, Sallán y Suñe (2005: 14).

Las flechas en la figura 23 representan la transición entre los estados *n, n-1* y *n+1*, los valores asociados con cada flecha representan las tasas de transición de los estados. Por ejemplo, tasa de transición del estado *n-1* al estado *n* es la tasa de llegadas λ_{n-1}, donde λ_{n-1} es una función del estado de origen, *n-1*; la tasa de transición de estado *n* al estado *n-1* es la tasa de salidas μ_n que también es una función del estado de origen *n*, y así sucesivamente. Bajo condiciones de estado estable, las tasas esperadas de flujo entrante y saliente del estado *n* deben ser iguales, como el estado *n* se comunica sólo con los estados *n-1* y *n+1*, las tasas de transición desde todos los otros estados (0, 1, 2,…, n – 2, n + 2, n + 3,…) deben ser cero.

$$\left[\begin{array}{c}\textbf{Tasas esperada de}\\ \textbf{flujo desde el estado } \boldsymbol{n}\end{array}\right] = 0(p_{o+} + \ldots + p_{n-2}) + \lambda_{n-1}p_{n-1} + \mu_{n+1}p_{n+1} + 0(p_{n+2} +$$

$$= \lambda_{n-1}p_{n-1} + \mu_{n+1}p_{n+1}$$

Similarmente $\left[\begin{array}{c}\textbf{Tasas esperada de flujo}\\ \textbf{desde el estado } \boldsymbol{n}\end{array}\right] = (\lambda_n + \mu_n)p_n$

Se igualan las dos tasas y se obtiene la ecuación de equilibrio:

$$\lambda_{n-1}p_{n-1} + \mu_{n+1}p_{n+1} = (\lambda_n + \mu_n)p_n, \quad n = 1, 2, \ldots$$

Esta ecuación es válida sólo para $n > 0$. Las ecuaciones de equilibrio se resuelven en forma recursiva comenzando con p_1 y procediendo por inducción para determinar p_n. De la ecuación de equilibrio para $n = o$, se obtiene:

$$p1 = \frac{\lambda_0}{\mu_1} p_0$$

Para $n = 1$ se tiene $\lambda_0 p_0 + \mu_2 p_2 = (\lambda_1 + \mu_1)p_1$, sustituyendo $p_1 = (\lambda_0 / \mu_1)p_0$ y simplificando $p_2 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} p_0$. En general, se puede demostrar por inducción que:

$$p_n = \frac{\lambda_{n-1}\lambda_{n-2}\ldots\lambda_0}{\mu_n \mu_{n-1}\ldots\mu_1} p_{0,} \quad n = 1, 2, \ldots$$

El valor de p_0 *se determina con:* $\sum_{n=0}^{\infty} p_n = 1$

Con un ejemplo de Fonollosa, Sallán y Suñe (2005) se puede esclarecer aún más lo dicho hasta ahora. En supermercados MAS se opera con tres cajas, el gerente usa el siguiente programa para determinar la cantidad de cajeros en operación en función de la cantidad de clientes en la tienda:

Cantidad de clientes en la tienda	Cantidad de cajeros funcionando
1 a 3	1
4 a 6	2
Más de 6	3

Los clientes llegan a las cajas siguiendo una distribución de 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 que haya *n* clientes en las cajas.

$$p_1 = \left(\frac{10}{5}\right)p_0 = 2p_0, \quad p_2 = \left(\frac{10}{5}\right)^2 p_0 = 4p_0, \quad p_3 = \left(\frac{10}{5}\right)^3 p_0 = 8p_0,$$

$$p_4 = \left(\frac{10}{5}\right)^3\left(\frac{10}{10}\right)p_0 = 8p_0, \quad p_5 = \left(\frac{10}{5}\right)^3\left(\frac{10}{10}\right)^2 p_0 = 8p_0, \; p_6 = \left(\frac{10}{5}\right)^3\left(\frac{10}{10}\right)^3 p_0 = 8p_0,$$

$$p_n = \left(\frac{10}{5}\right)^3\left(\frac{10}{10}\right)^3\left(\frac{10}{15}\right)^{n-6} p_0 = 8\left(\frac{2}{3}\right)^{n-6} p_0, \; n = 7, 8, \ldots$$

El valor de p_o se determina así:

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

$$p_0\left\{31 + 8\left(\frac{1}{1-\frac{2}{3}}\right)\right\} = 1 \quad p_0 = \frac{1}{55}$$

Conocido p_0 se puede determinar cualquiera de las probabilidades. La probabilidad que sólo hay una caja abierta se calcula como la de que haya entre 1 y 3 clientes en el sistema:

$$p_1 + p_2 + p_3 = (2 + 4 + 8)\left(\frac{1}{55}\right) \approx 0.255$$

Se puede usar p_n para determinar medidas de funcionamiento, o de eficiencia, para el caso de supermercado MAS

$$\begin{bmatrix}\text{Cantidad esperada}\\ \text{de cajas vacías}\end{bmatrix} = 3p_0 + 2(p_1 + p_2 + p_3) + 1(p_4 + p_5 + p_6) + 0(p_7 + p_8 + \ldots) = 1caja$$

Medidas de Desempeño de Estado Estable

Las medidas de desempeño del estado estable como lo señalan Marrero, Asencio, Abreu, Orozco y Granela (2006), se pueden usar para analizar la operación de las líneas de espera con el fin de hacer recomendaciones sobre el diseño del sistema. Como medidas de desempeño se pueden mencionar:

L_s = número esperado de clientes en el sistema.

L_q = número esperado de clientes en la fila.

W_s = tiempo estimado de espera en el sistema.

W_q = tiempo estimado de espera en la fila.

Se está estudiando un servicio con c servidores en paralelo, de la definición de p_n se obtiene:

$$L_s = \sum_{n=0}^{\infty} np_n \qquad L_q = \sum_{n=c+1}^{\infty} (n-c)p_n$$

Se puede decir que existe una íntima relación entre L_s y W_s, así como entre L_q y W_q de manera que cualquier medida se determina automáticamente a partir de la otra.

Siendo λ_{ef} la tasa promedio efectiva de llegadas (independientemente del número en el sistema *n*), entonces:

$$L_s = \lambda_{ef}\ W_s \qquad\qquad L_q = \lambda_{ef}\ W_q$$

El valor de λ_{ef} se determina a partir de λ_n dependiente del estado y las probabilidades de p_n $\lambda_{ef} = \sum_{n=0}^{\infty} \lambda_n p_n$

También existe una relación directa entre W_s y W_q, pues por definición:

$$\begin{pmatrix}\text{Tiempo de}\\ \text{espera estimado}\\ \text{en el sistema}\end{pmatrix} = \begin{pmatrix}\text{Tiempo de}\\ \text{espera estimado}\\ \text{en la fila}\end{pmatrix} + \begin{pmatrix}\text{Tiempo}\\ \text{estimado de}\\ \text{servicio}\end{pmatrix}$$

Dado que μ es la tasa de servicio por servidor activo, el tiempo estimado de servicio es $1/\mu$, se obtiene:

$$W_s = W_q + 1/\lambda$$

Se multiplica ambos miembros de la ecuación por λ_{ef} $L_s = L_q + \dfrac{\lambda_{ef}}{\mu}$

La utilización estimada de servicio se define como una función del número promedio de servidores activos, como la diferencia entre L_s y L_q debe ser igual al número estimado de servidores ocupados:

$$\begin{pmatrix}\text{Número estimado de}\\ \text{servidores activos}\end{pmatrix} = \bar{c} = L_s - L_q = \frac{\lambda_{ef}}{\mu}$$

Porcentaje de utilización de un servicio con *c* servidores en paralelo $= \dfrac{\bar{c}}{c} * 100.$

Un ejemplo de lo dicho hasta ahora es el siguiente, tomado de los mismos autores. Se considera la situación de una línea de espera con un solo servidor donde las tasas de llegadas y salidas son constantes y están dadas por $\lambda_n = 3$ llegadas/hora y $\mu_n = 8$ salidas/hora para toda $n \geq 0$. Se usan las probabilidades p_n para calcular las medidas de desempeño, se observa que $\lambda_n = \lambda = 3$ llegadas/hora para toda $n \geq 0$; así la tasa de promedio de llegadas se calcula como $\lambda_{ef} = 3\ (p_0 + p_1 + p_2 + \ldots) = 3$ llegadas/hora

$$L_s = \sum_{n=0}^{\infty} np_n = 0*0*0.625 + 1*0.234 + 2*0.088 + 3*0.033 + 4*0.012 + 5*0.005 + 6*0.002$$
$$+ 7*0.001 = 0.6 cliente$$

Usando $L_s = \square_{ef}\ W_s$ se obtiene el tiempo de espera en el sistema como:

$W_s = \frac{L_s}{\lambda_{ef}} = \frac{0,6}{3} = 0,2 hora$ De aquí se obtiene el tiempo estimado de espera en la fila:

$W_q = W_s - \frac{1}{\mu} = 0,2 - \frac{1}{8} = 0,075 hora$ El número esperado en la fila se calcula:

$L_q = \lambda_{ef} W_q = 3*0,075 = 0,225 cliente$ Como el establecimiento tiene un solo servidor $c = 1$, el porcentaje de utilización se calcula así,

$$= \frac{\bar{c}}{c} * 100$$

$$= \frac{L_s - L_q}{1} * 100$$

$$= \frac{0,6 - 0,225}{1} * 100 = 37,5\%$$

Líneas de Espera Especializadas de Poisson

Cada modelo de los que se estudia seguidamente se describen en términos de la notación extendida por Kendall, como la deducción de p_n es completamente independiente de la disciplina de la línea de espera, es apropiado usar el símbolo DG (disciplina general) en la notación de Kendall.

Estudio de una Cola M/M/1

Este punto será analizado en base a los aporte de Rodríguez y Gámez (2002). Ellos establecen que se denomina estado del sistema en *t* al número de elementos presentes en el instante de tiempo *t*; para $t = 0$, el estado del sistema sería el número de elementos que están en el sistema inicialmente. Se presume que el sistema ha llegado al estado estacionario, que se caracteriza porque la probabilidad de cada estado no varía con el tiempo; P_j es la probabilidad que el sistema esté en el estado *j*, puede interpretarse como la fracción de tiempo en que hay *j* elementos en el sistema. Se muestra en la figura 23.

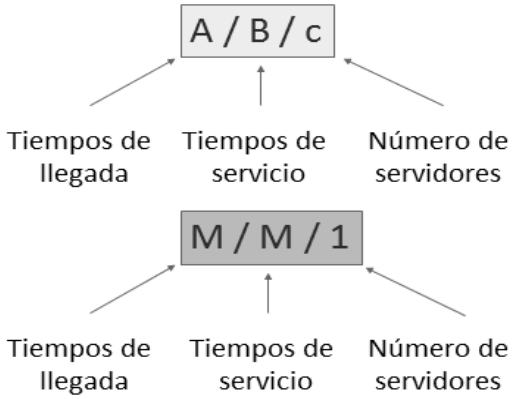


Figura 23. Cola M/M/1

Fuente: Gestión de Operaciones (2015: párr.3).

Un sistema de COLAS M/M/1/fcfs/∞/∞ sigue las leyes siguientes:

1. La probabilidad de un llegada entre t y $t + \Delta t$, puede darse por $\lambda\Delta t + o(\Delta t)$. Una llegada incrementa el estado del sistema en 1.

2. La probabilidad de una salida entre t y $t + \Delta t$ (siempre que haya algún elemento recibiendo servicio en el instante t) puede darse por $\mu\Delta t + o(\Delta t)$. Una salida disminuye en 1 el estado del sistema.

3. Las llegadas y salidas son sucesos independientes.

4. El estado estacionario se alcanza si $\lambda < \mu$, siendo λ y μ las tasas de llegada y servicio (número de llegadas o servicios por unidad de tiempo).

5. Dos o más sucesos (llegadas o salidas) no pueden ocurrir simultáneamente (esto es una forma de decir que la probabilidad de ocurrencia de más de un suceso en el tiempo Δt es un infinitésimo de orden superior a Δt).

Para calcular la probabilidad que el sistema esté en el estado j en el instante $t + \Delta t$, a partir de su estado en el tiempo t. Para el caso de $P_0 (t+\Delta t)$ = Probabilidad que no haya nadie en el sistema en el instante $t + \Delta t$; se dará esta circunstancia en uno de los supuestos siguientes:

1. No había nadie en el sistema en el instante t y no ha venido nadie en este intervalo, la probabilidad que ocurra este supuesto es: $P_0 (t) (1 - \lambda\Delta t + o(\Delta t))$.

2. Había 1 elemento en el instante t, no ha venido nadie en ese intervalo y se ha ido el que estaba, la probabilidad es: $P_1 (t) (\mu\Delta t + o(\Delta t))(1 - \lambda\Delta t + o(\Delta t))$.

3. Los casos restantes requieren que al menos dos sucesos (entradas o salidas) ocurran en el intervalo de tiempo Δt, según la propiedad 5 esta probabilidad es de orden superior a Δt. Por lo tanto,

$$P_0 (t+\Delta t) = P_0 (t)(1-\lambda\Delta t+ o(\Delta t))+ P_1 (t)(\mu\Delta t + o(\Delta t)(1 -\lambda\Delta t + o(\Delta t)) + o(\Delta t) \Rightarrow$$

$$P_0\,(t + \Delta t) = P_0(t)(1 - \lambda\Delta t + o(\Delta t)) + P_1(t)\,(\mu\Delta t + o(\Delta t)) + o(\Delta t) \Rightarrow$$

$$\frac{P_0(t+\Delta t) - P_0(t)}{\Delta t} = -\lambda P_0(t) + \mu P_1(t) + \frac{o(\Delta t)}{\Delta t}$$

Si $\Delta t \to 0$ se obtiene en el primer miembro la derivada de $P_0(t)$, si se considera que se está estado estacionario P_0 es constante y su derivada es 0; el último sumando del segundo término también tiende a 0 puesto que el orden del numerador es mayor que el del denominador (el numerador tiende a cero más rápidamente que el denominador). Por tanto ha de cumplirse cuando $\Delta t \to 0$:

$0 = -\lambda P_0 + \mu P_1$, de donde se deduce que $P_1 = \frac{\lambda}{\mu} P_0 = \rho P_0$

En el caso general, tras agrupar todos los infinitésimos de orden superior a Δt se tiene:

$$P_j(t)(t + \Delta t) = P_{j-1}(t)(\lambda\Delta t))(1-\mu\Delta t) + P_j(t)(1-\lambda\Delta t)(1-\mu\Delta t) + P_{j+1}(t)(\mu\Delta t)(1-\lambda\Delta t) + o(\Delta t)$$

Se procede de forma análoga para obtener la expresión de P_1:

$P_j(\lambda + \mu) = \lambda P_j - 1 + \mu P_j + 1$ En concreto si $j = 1$ $P_1(\lambda + \mu) = \lambda P_0 + \mu P_2$

Sustituyendo $P_1 = \frac{\lambda P_0}{\mu}$ se obtiene $\frac{\lambda P_0}{\mu}(\lambda + \mu) = \lambda P_0 + \mu P_2$ y despejando P_2

$P_2 = \frac{\lambda^2 P_0}{\mu^2} = \rho^2 P_0$ Por inducción, se obtiene para el estado j: $P_j = \frac{\lambda^j P_0}{\mu^j}$ Como los posibles estados del sistema son: 0, 1, 2, 3,...

$$\sum_{i=0}^{\infty} P_i = 1 = P_0 + \rho P_0 + \rho^2 P_0 + \ldots = P_0(1 + \rho + \rho^2 + \ldots) = P_0\left(\frac{1}{1-\rho}\right)$$

La serie representa la suma de los términos de una progresión geométrica de razón ρ, si el estado es estacionario $\rho = \frac{\lambda}{\mu} < 1$ y en este caso $1 + \rho + \rho^2 + ... = \frac{1}{1-\rho}$; luego $P_0\left(\frac{1}{1-\rho}\right) = 1$, de donde se deduce que $P_0=1-\rho$, por tanto $P_j=(1-\rho)\rho^j$

Para calcular L, número medio de elementos en el sistema, se usa el concepto de esperanza matemática

$$L = E(j) = \sum_{j=0}^{\infty} jP_j = \sum_{j=0}^{\infty} j(1-\rho)\rho^j = (1-\rho)\sum_{1}^{\infty} j\rho^j = (1-\rho)^S, \quad \text{siendo} \quad S = \sum_{1}^{\infty} i\rho^j \text{ y } j \text{ la}$$

variable aleatoria que denota el número de elementos en el sistema.

Para hallar S se parte de la igualdad $S - \rho S = \frac{\rho}{1-\rho}$, de donde se deduce $S = \frac{\rho}{(1-\rho)^2}$. Se sustituye esta expresión en la de L y se obtiene para el número medio de elementos en el sistema $L = \frac{\rho}{1-\rho}$.

Para hallar el número de elementos medios en cola hay que hallar la media de j, que es la variable aleatoria que denota el número de elementos en la cola, su promedio es:

$$L_q = 0(P_0 + P_1) + 1P_2 + 2P_3 + 3P_4 + ... = \sum_{j=1}^{\infty} (j+1)P_j = \sum_{j=1}^{\infty} (j-1)(1-\rho)\rho^j =$$

$$= (1-\rho)\rho\sum_{1}^{\infty} (j-1)\rho^{j-1} = (1-\rho)\rho S = \frac{\rho^2}{1-\rho}$$

El número de elementos recibiendo servicio, es la diferencia entre los que están en el sistema y los que están en la cola, su valor medio, L_s, es:

$$L_s = L - L_q = \frac{\rho}{1-\rho} - \frac{\rho^2}{1-\rho} = \frac{\rho(1-\rho)}{1-\rho} = \rho$$

Este valor, ρ, también puede interpretarse como la fracción de tiempo en que el servidor está ocupado. Un ejemplo de Taha (2017), puede servir para ilustrar lo hasta ahora escrito sobre el punto. Lavado automático para carros funciona sólo con un lugar, los carros llegan siguiendo una distribución de Poisson, con 4 autos por hora, que pueden esperar en el estacionamiento del autolavado si el lugar de lavado está ocupado. El tiempo para lavar y limpiar un carro es exponencial, con 10 minutos de promedio, los coches que no se pueden estacionar en la instalación pueden esperar en la quebrada junto al autolavado; quiere decir que para todo fin práctico no hay límite del tamaño del sistema. El gerente del mismo desea determinar el tamaño del estacionamiento.

Para este caso $\lambda = 4$ autos por hora y $\mu = 60/10 = 6$ coches por hora, $\rho = \lambda / \mu < 1$, el sistema puede funcionar en condiciones de estado estable. La cantidad promedio de carros en la cola, L_q, es 1,33; pero esta no debe ser la única base para determinar el número de puestos de estacionamiento, porque el diseño debe contener la longitud máxima posible de la cola. Puede ser factible diseñar el estacionamiento de modo que un carro que llegue encuentre lugar al menos el 90% de las veces. Sea K la cantidad de puestos de estacionamiento, tener K puestos equivale a tener $K + 1$ lugares en el sistema (en cola y en servicio); un coche que llega encontrará un puesto el 90% de las veces si hay cuando mucho K autos en el sistema; equivale al siguiente enunciado de probabilidades: $p_0 + p_1 + \ldots p_k \geq 0,9$. La cantidad K de espacios se puede determinar usando la definición de p_n $(1-\rho)(1+\rho+\rho^2+\ldots\rho^k) \geq 0,9$.

La suma de la serie geométrica es igual a $\frac{1-\rho^{k-1}}{1-\rho}$, la condición se reduce a $(1-\rho^{k-1})$ $\geq 0,9$. Al simplificar la desigualdad se obtiene $\rho^{k+1} \leq 0,1$. Se sacan logaritmos de ambos lados para obtener:

$$K \geq \frac{\ln(0,1)}{\ln\left(\frac{4}{6}\right)} - 1 = 4,679 \approx 5$$

En consecuencia, $K \geq 5$ puestos de estacionamiento.

Modelo (M/M/1):(DG/N/∞)

Para explicar este modelo se utilizan las explicaciones dadas por Anderson, Sweeney, Williams, Camm y Kipp (2011). Señalan que en este modelo el número máximo de clientes permitidos en el sistema es *N* (longitud máxima de la línea de espera es = *N*-1); significa cuando haya *N* clientes en el sistema, se impiden todas las nuevas llegadas o no se les permite unirse a la línea de espera. Haciendo □ = □ □ se obtiene:

$$P_0 = \begin{cases} (1-\square)\ (1-\square^{N+1}); \ \square \ \square\ 1. \\ 1 / N+1; \ \square = 1. \end{cases}$$

Las formulas para $p_{n:}$

$$P_n = \begin{cases} [\ (1-\square)\ (1-\square^{N+1})]\square^{n}; \ \square \ \square\ 1 \qquad n = 0,\ 1,2,... \\ 1 / (N+1); \ \square = 1 \end{cases}$$

Para este modelo no es necesario que □>1, el número de unidades en el sistema está controlado por la longitud de la línea de espera (= N-1). Usando el valor de p_n, se encuentra que el número esperado de unidades en el sistema se calcula:

$$L_s = \begin{cases} \{\square[1-(N+1)\square\square^{N} + N\square^{N+1}]\}\ \ [(1-\square)(1-\square^{N+1})]; \quad \square \ \square\ 1 \\ N / 2; \ \square = 1 \end{cases}$$

Las medidas L_q, W_s y W_q se pueden calcular a partir de L_s, una vez que se determina la tasa efectiva de llegadas $\square_{ef}$: $\square_{e,\,f} = \square(1-p_N)$. Usando L_s y $\square_{ef}$ se obtienen las formulas para calcular, L_q, W_q y W_s:

$L_q = L_s-(\square_{e\ f}\square) = L_s-[(1-p_N)]\square$ p_N = Probabilidad que una unidad no se una al sistema.

$W_q = L_q/\square_{e\ f} = L_s/[\square(1-p_N)]$

$W_s = W_q+1/\square = L_s/[\square(1-p_N)]$

El ejemplo utilizado para aclara el modelo se toma del mismo autor, Anderson, Sweeney, Williams, Camm y Kipp (2011). Así pues, se considera una heladería, la misma tiene 4 mesas, si están llenas, las personas que lleguen irán a otras; se observa que = 4 + 1 = 5. Un dato que puede interesar al dueño es saber cuántos clientes se pierden debido al limitado número de mesas, equivale a determinar el valor de λ_{pN}, o λ-$\lambda_{ef.}$ λ-λ_{ef} = 4 – 3.8075 – 0,1925 clientes/hora, o bien con base en una jornada de ocho horas la heladería pierde 2 (≈8*0,1925) clientes por jornada en promedio, (2/4*8)*100 = 6,25% de todos los comensales de helados que llegan por jornada. Una decisión respecto a aumentar el número de mesas a más de 4 debe basarse en el valor de clientes perdidos. El tiempo total esperado desde que un cliente pide su helado, se sienta en la mesa y se le sirve el helado es W_s = 0.3736 horas, cercano a 22 minutos.

Modelo (M/M/c):(DG/□□)

El autor que sirve de fundamento al análisis de este modelo es Sarabia (1996), quien señala que en este modelo los clientes llegan con una tasa constante □ y un máximo de c unidades puede ser atendidos simultáneamente. La tasa de servicio por servidor activo es constante e igual a □ y $\square_{ef}$ = □. El efecto de usar c servidores paralelos es acelerar la tasa de servicio, permite servicios simultáneos, si el número de clientes en el sistema, n, ≥ c, la tasa combinadas de salidas de la instalación es c□□; si n < c, la tasa de servicio es igual a n□□. Así, se tiene:

$\lambda_n = \lambda$ $\quad n \geq 0$ $\qquad\qquad$ $\lambda_n = \begin{cases} n\lambda, & n < c \\ c\lambda, & n \geq c \end{cases}$

Si $\rho = \lambda / \mu$; el valor de p_n y p_0 se calcula de la siguiente forma:

$$P_n = \begin{cases} (\rho^n n!)p_0 & 0 \leq n \leq c \\ (\rho^n c^{n-c} c!)p_0 & n > c \end{cases}$$

$$P_0 = \sum_{n=0}^{c-1} (\rho^n n!) + \rho^c [c!(1-\rho/c)]\}^{-1}$$

Los valores de las medidas de desempeño se obtienen así:

$L_q = [\rho^{c+1} / (c-1)!(c-\rho)^2]p_0 = [c\rho/(c-\rho)^2]p_c$

$L_s = L_q + \rho$

$W_q = L_q/\lambda$

$W_s = W_q + 1/\mu$

Las operaciones asociadas con este modelo pueden ser tediosas, Morse da dos aproximaciones útiles para p_o y L_q, para ρ mucho menor que 1,

$$P_0 \approx 1-\rho \quad y \quad L_q \approx \rho^{c+1} c^2$$

Y para ρ/c muy próxima a 1,

$$P_0 \approx [(c-\rho)(c-1)]!c^c \quad y \quad L_q \approx \rho/c-\rho$$

Un ejemplo del mismo autor es el que sigue, hay dos (2) empresas de taxi que dan servicio a una población, cada una posee 2 taxis y comparten partes iguales del

mercado, llegan ocho (8) llamadas por hora a la oficina de cada empresa; el tiempo promedio en el viaje es doce (12) minutos. Las llamadas llegan siguiendo una distribución de Poisson y el tiempo de viaje es exponencial, un inversionista compró las dos (2) empresas y busca consolidarlas en una (1) sola oficina para dar mejor servicio a los clientes. Los taxis son los servidores y el viaje es el servicio.

Modelo (M/M/2):(DG/N/∞/∞) λ = 8 llamadas por hora μ = 60/12 = 5 viajes por taxi por hora; al consolidarlas (M/M/4):(DG/N/∞/∞) λ = 2*8 = 16 llamadas por hora y μ = 5 viajes por taxi por hora.

Una medida adecuada para comparar los 2 modelos es el tiempo promedio que espera un cliente para un viaje, W_q.

Los resultados indican que el tiempo de espera para un viaje es 0,356 hora (≈ 21 minutos) para el caso de dos empresas y 0,149 (≈ 9 minutos) para el caso consolidado. Es una reducción de más del 50% y una evidencia que da garantía de la consolidación de las dos (2) empresas en una (1).

Modelo (M/M/c):(DG/N/□), c □ N

Esta modelo impone un límite N sobre la capacidad del sistema (tamaño máximo de la línea de espera = N-c). Así lo señala Handy Taha en su libro *Investigación de operaciones de 1995.* En términos del modelo generalizado, λ_n y μ_n para el modelo actual están dadas por:

$$\lambda_n = \lambda, \quad \begin{cases} 0 \leq n < N \\ 0, \quad n \geq N \end{cases}$$

$$\mu_n = n\mu, \quad \begin{cases} 0 \leq n \leq c \\ c\mu, \quad c\mu \leq n \leq N \end{cases}$$

Se sustituye por λ_n y μ_n en la expresión general de p_n y observando que $\rho = \lambda/\mu$; se obtiene:

$$P_n = \begin{cases} (\lambda^n n!)^* p_0, & 0 \le n \le c \\ [\lambda^n (c! c^{n-c})] p_0, & c \le n \le N \end{cases}$$

Donde:

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

La diferencia entre p_n en este modelo y el anterior ocurre en la expresión de p_0 y el factor de uso λc no necesita ser menor que 1. Las medidas de desempeño se calculan de la siguiente forma:

$$Lq = \begin{cases} p_0[\lambda^{c+1}(c-1)!(c-\lambda)^2]\{1-(\lambda c)^{N-c}-(N-c)(\lambda c)^{N-c}[1-(\lambda c)]\}, & \lambda c \neq 1 \\ p_0[\lambda^c (N-c)(N-c+1)2c! & \lambda c = 1 \end{cases}$$

$L_s = L_q + (c + \bar{c}) = L_q + \lambda_{ef}$

Donde:

$\bar{c}$ = número estimado de servidores inactivos $= \sum_{n=0}^{c} (c-n)p_n$

$\lambda_{ef} = \lambda(1-p_n) = \lambda(c-\bar{c})$

$\lambda_{e,f}$ = La tasa efectiva de llegada.

Modelo de Autoservicio (M/M/□):(DG/□□)

Los autores que orienta la explicación de este modelo son Marrero, Asencio, Abreu, Orozco y Granela (2006). En este modelo el número de servidores es ilimitado, porque el cliente mismo es también servidor; caso de los establecimientos de autoservicio; en términos del modelo generalizado se tiene:

$\square_n = \square$ para toda n $\square$ 0

$\square_n = n\square$ para toda n $\square$ 0

La sustitución directa en la expresión de p_n produce:

$P_n = (\square^n \ n!\square^n)p_0 = (\square^n n!)p_{0,}$ ya que $\sum_{n=0}^{\infty} p_n = 1$ se deduce que:

$P_0 = 1 / 1+ \square + (\square^2 2!) + \square = 1e^{\square} = e^{-\square}$

Como resultado:

$P_n = (e^{-\square}\square^n)n!$, n = 0,1,2,□, p_n sigue una distribución de Poisson con media $E\{n\} = \square$. Las medidas de desempeño se obtienen por:

$L_s = E\{n\} = \square$

$W_s = 1/\square$

$L_q = W_q = 0$

$W_q = 0$ cada cliente se atiende, razón por que W_s es igual al tiempo de servicio medio 1/□.

Los cálculos muestran que cuando □ se hace chica, □ es mucho mayor que □, (M/M/□) es una aproximación bastante exacta de (M/M/c) aun para c tan chica como 10.

Modelo de Servicio de Máquinas (M/M/R):(DG/K/K), R > k

Según Rincón (2001) en este modelo se asume que se dispone de R técnicos en reparaciones para dar servicio a k máquinas; como una maquina descompuesta no puede generar nuevas llamadas mientras está en servicio. El modelo es ejemplo de una fuente de llamadas finita. En la figura 24 se representa.

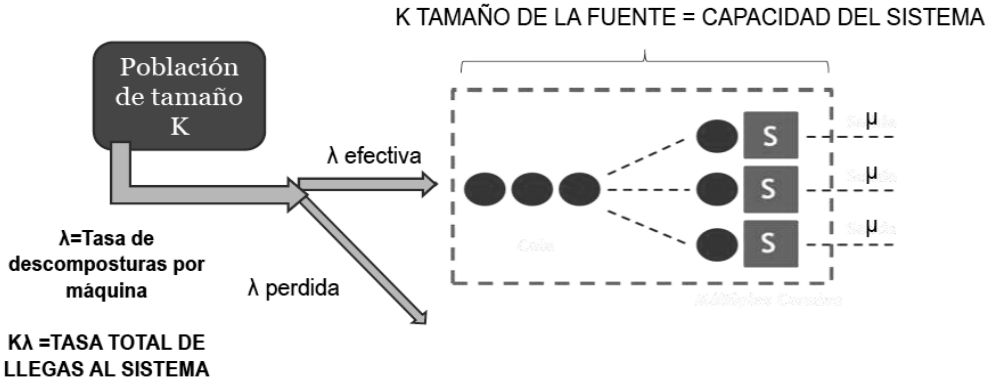


Figura 24. Modelo de servicio de máquinas (M/M/R):(DG/K/K), R > k l
Fuente: Campos (2016: 10).

Modelos en Líneas de Espera que No Obedecen la Distribución de Poisson

De acuerdo a De La Fuente y Pino (2007), los modelos de líneas de espera donde los procesos de llegada y/o salida no siguen las hipótesis de Poisson, conducen a resultados complejos y poco manejables; puesto que, el tiempo de servicio se describe por medio de una distribución general de probabilidades con media E{t} y varianza var{t}. El análisis no provee una expresión analítica manejable para las

probabilidades p_n, los resultados sólo aportan las medidas básicas de desempeño, L_s, L_q, W_s, y W_q.

Para □ igual a la tasa de llegadas a una instalación con un solo servidor y dadas E{t} y var{t} como la media y la varianza de la distribución del tiempo de servicio, se puede demostrar usando el análisis de cadena de Markov que:

L_s = □ E{t}+[□^{2}(E^2{t}+var{t})]2 (1-□ E{t}), (M/G/1):(DG/□□) □ E{t}< 1.

$W_s = L_s/$□

$L_q = L_s$-□ E{t}

$W_q = L_q/$□

La tasa de servicio está dada por □ = 1/E{t} y □$_{ef}$ = □ para este modelo. Para el caso cuando el tiempo de servicio es constante, var{t}= 0 la fórmula de L_s anterior se reduce a: L_s = □+□^{2}[2(1□□)], (M/D/1):(DG/□□) □ =□□ y □ es la tasa constante de servicio. Las otras medidas de desempeño permanecen iguales. Si el tiempo de servicio es tipo Erlang con parámetros m y □, con E{t}= 1□ y var{t}=1/(m□^{2}), la fórmula de L_s sería: □+(1+m)/(2m)[□(1-□)], (M/Em/1):(DG/□□).

La ejemplificación del modelo vendría dada por una lavandería, el lavado lo realizan las lavadoras automáticas, el tiempo de servicio es el mismo y constante para todos los clientes; el ciclo de la lavadora es de 10 minutos exactos.

λ = 4 hora, el tiempo de servicio es constante,

E{t} = 10/30 = 1/6 hora y var {t} = 0.

$$L_s = 4(1/6) + \frac{4_2(1/6)^2 + 0}{2(1-4/6)} = 1{,}333\ ciclos\ de\ lavado$$

$$L_q = 1{,}333 - (4/6) = 0{,}667\ ciclos\ de\ lavado$$

$$W_s = \frac{1{,}333}{4} = 0{,}333 hora$$

$$W_q = \frac{0{,}667}{4} = 0{,}167 hora$$

Modelos de Líneas de Espera con Prioridades en el Servicio

Según Cao (2002) en los modelos de espera con prioridad se presume que se forman varias líneas de espera en paralelo, incluyendo los clientes que pertenezcan a cierto orden de prioridad. Si la instalación tiene m líneas de espera, la línea de espera 1 tiene más alta prioridad de servicio y la línea de espera m incluye a clientes con más baja prioridad; las tasas de llegadas y servicio pueden variar para las diferentes filas de prioridad. La figura 25 es un ejemplo de ello.

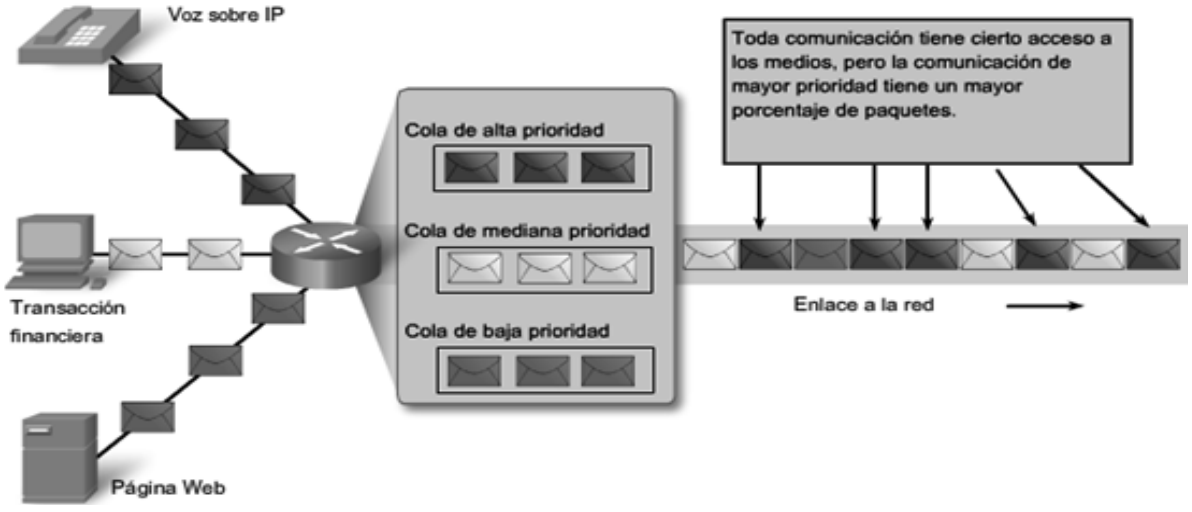


Gráfico 25. Uso de colas para priorizar la comunicación

Fuente: Ingeniería de Sistemas (2016: párr.13).

El servicio de prioridad puede seguir una de estas dos reglas:

1. Regla de prioridad, el servicio de un cliente de más baja prioridad puede ser interrumpido para favorecer a un cliente que llegue con más alta prioridad.

2. Regla de no prioridad, un cliente, una vez que está siendo atendido, saldrá del establecimiento sólo después que acabe de ser atendido, sin importar la prioridad del cliente que llegue.

Los modelos que se rigen por la regla de no prioridad se aplican a servidores únicos y múltiples.

En el caso de servidor único, supone llegadas de Poisson y distribuciones de servicio arbitrarias.

En el caso de servidores múltiples, las llegadas y salidas siguen la distribución de Poisson.

El símbolo NPRP se utiliza con la notación de Kendall para representar la disciplina de no prioridad, M_i y G_j representan distribuciones de Poisson y arbitraria.

Modelo de No Prioridad con Un Servidor (M/G/1):(NPRP/□□)

Sea F_i (t) la FDA de la distribución de tiempo de servicio arbitraria para la i-ésima línea de espera (i = 1,2,… m) y sea E{t} y var{t} la media y la varianza, sea λ_i la tasa de llegada en la i-ésima fila por el tiempo unitario. Los valores de las medidas de desempeño se logran con las siguientes expresiones:

$$W_q^{(k)} = \frac{\sum_{i=1}^{m} \lambda_i (E_i^2\{t\} + \mathrm{var}_i\{t\})}{2(1 - S_{k-1})(1 - S_k)}$$

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

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

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

Donde:

$\square_k = \square_k E\{t\}$ $\quad S_k = \sum_{i=1}^{k} < 1$, k = 1,2,3 …, m $\quad S_0 = 0$

Se definen $W_q^{(k)}\ L_s^{(k)}\ L_q^{(k)}\ W_s^{(k)}$ como las medidas de desempeño para la k-ésima línea de espera, el tiempo de espera estimado en la línea de espera para cualquier cliente sin importar su prioridad está dado por: $Wq = \sum_{k=1}^{m} (\square_k \square) W_q^{(k)}$

Donde: $\square = \sum_{i=1}^{m} \square_i$ y $\square_k \square$ es el peso relativo de $W_q^{(k)}$, resultado similar se aplica a W_s.

A una panadería llegan trabajos en tres categorías: urgencia, regular y baja prioridad; los trabajos urgentes son procesados antes que cualquiera, los trabajos regulares tienen preferencia sobre los de baja prioridad, cualquier trabajo una vez empezado debe terminarse antes de iniciar otro. La llegada de orden 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, 3 y 5 por días. Existen tres líneas de espera de no prioridad, se tiene:

$\rho_1 = 4(1/10) = 0,4$ $\quad \rho_2 = 3(1/9) = 0,333$ $\quad \rho_3 = 1(1/5) = 0,2$

$S_1 = \rho_1 = 0,4$ $\quad S_2 = \rho_1 + \rho_2 = 0,733$ $\quad S_3 = \rho_1 + \rho_2 + \rho_3 = 0,933$

Como $S_3 > 1$ el sistema puede alcanzar condiciones de estado estable. Se calcula el tiempo de espera estimado en cada línea:

$$4[(1/10)^2+0] + 3\,[(1/9)^2 + 0] + 1\,[(1/5)^2 + 0] = 0{,}117$$

$$W_q^1 = \frac{0{,}117}{2(1-0)(1-0{,}4)} = 0{,}0975 día \cong 2{,}34 horas$$

$$W_q^2 = \frac{0{,}117}{2(1-0{,}4)(1-0{,}733)} = 0{,}365 día \cong 8{,}77 horas$$

$$W_q^3 = \frac{0{,}117}{2(1-0{,}733)(1-0{,}933)} = 3{,}27 día \cong 78{,}5 horas$$

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

$$W_q = \frac{4*2{,}34 + 3*8{,}77 + 1*78{,}5}{4+3+1} = 14{,}27 horas$$

Modelo de No Prioridad con Varios Servidores (M/M/c):(NPRP/□□)

La comprensión de este modelo está guiado por las notas de Zaragoza (2004), quien supone que todos los clientes tienen la misma distribución del tiempo de servicio, los c canales tienen una distribución de servicio exponencial idéntica con tasa de servicio □. Las llegadas de K-ésima línea de espera con prioridad ocurren según una distribución de Poisson con una tasa de llegada $\square_K$, k = 1,2,…, m; para la k-ésima línea de espera:

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

Donde: $S_0 \equiv 0$

$$Sk = \sum_{i=1}^{k} \frac{\lambda_i}{c\mu} < 1 \quad \text{para toda k}$$

$$E\{\xi_0\} = \frac{1}{\mu\left[c\rho^{-c}(c-\rho)(c-1)!\sum_{n=0}^{c-1}\frac{\rho n}{n!}+1\right]}, \quad \rho = \frac{\lambda}{\mu}$$

La espera numérica estimada en la línea de espera para todo el sistema viene dada por: $L_q = \square W_q$.

Un ejemplo que ilustra es el siguiente, hay tres líneas de espera con prioridad y tasa de llegada $\lambda_1 = 2$, $\lambda_2 = 5$ y $\lambda_3 = 10$ por día; existen dos servidores y la tasa de servicio es de 10 por día. Las llegadas y salidas siguen distribución de Poisson.

$$S_1 = 0,1 \qquad S_2 = 0,35 \qquad S_3 = 0,85$$

Como todas las expresiones $S < 1$, se alcanza el estado estable.

$$\rho = 1,7 \qquad E\{\xi_0\} = 0,039$$

$$W_q^1 = 0,0433 \qquad W_q^2 = 0,0665 \qquad W_q^3 = 0,4$$

El tiempo de espera en la fila para cualquier cliente por $W_q = 0,26$.

La espera numérica estimada en la línea de espera para todo el sistema está dada por: $L_q = 4,42$.

Líneas de Espera Sucesivas o En Serie

Modelo En Serie de Dos Estaciones con Capacidad de Líneas de Espera Cero

De acuerdo a Render, Stair y Hanna (2018), este modelo es de líneas de espera de Poisson con estaciones de servicio dispuestas en serie, de manera que todo cliente debe pasar por las dos estaciones antes de completar el servicio. Los tiempos de servicio en cada estación están exponencialmente distribuidos con la misma tasa de

servicio □, las llegadas ocurren según una distribución de Poisson con tasa □ y no se permiten COLAS en ninguna de las dos estaciones. Se detalla en la figura 26.

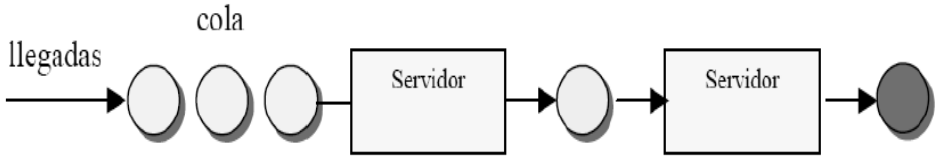


Figura 26. Modelo en serie de dos estaciones
Fuente: Jaime (2013: 13).

Es preciso para plantear este modelo identificar los estados del sistema; los símbolos 0, 1, b, representan los estados libres, ocupado y bloqueado respectivamente. Sean i y j los estados en la estación 1 y 2. Entonces los estados del sistema se pueden representar: {(i, j)} = {(0, 0); (1, 0); (0, 1); (1, 1); (b, 1)}

Definiendo $p_{ij}(t)$ 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 se resumen en la siguiente tabla:

Estados t	(0, 0)	(0, 1)	(1, 0)	(1, 1)	(b, ¡)
(0, 0)	1- □h		□h		
(0, 1)	□h(1-□h)	1-□h-□h		□h(1-□h)	
(1, 0)		□h(1-□h)	1 - □h		
(1, 1)		□h(1-□h)	□h	(1-□h)(1-□h)	□h
(b, 1)		□h(1-□h)			1- □h

Los cuadros vacíos indican que las transiciones entre los estados indicados t y t+h son imposibles (=0).

Por esto se pueden plantear las siguientes ecuaciones:

$p_{00}(t+h) = p_{00}(t)(1-□h) + p_{01}(t)(□h)$

$p_{01}(t+h) = p_{01}(t)(1-□h-□h) + p_{10}(t)(□h) + p_{b,1}(t)(□h)$

$p_{10}(t+h) = p_{00}(t)(\square h) + p_{10}(t)(1-\square h)\ p_{11}(t)(\square h)$

$p_{11}(t+h) = p_{01}(t)(\square h) + p_{11}(t)(1-2\square h)$

$p_{b1}(t+h) = p_{11}(t)(\square h) + p_{b1}(t)(1-\square h)$

Las ecuaciones de estado estable son:

$p_{01} - \square p_{00} = 0$

$p_{10} + p_{b1} - (1 + \rho)p_{01} = 0$

$\square p_{00} + p_{11} - p_{10} = 0$

$\square p_{01} + 2p_{11} = 0$

$p_{11} - p_{b1} = 0$

Se analizan las ecuaciones y se agrega la condición:

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

Se llega a la conclusión que la solución para p_{ij} es:

$$p_{00} = 2/A$$

$$p_{01} = (2\square)\square$$

$$p_{10} = (\square^2 + 2\square)\square$$

$$p_{11} = p_{b1} = \square^2\square$$

Donde $A = 3\square^2 + 4\square + 2$.

El número esperado en el sistema puede obtenerse como:

$$L_s = 0p_{00} + 1(p_{01} + p_{10}) + 2(p_{11} + p_{b1}) = (5\square^2 + 4\square)4.$$

Modelo en Serie de K Estaciones con Capacidad de Líneas de Espera Infinita

En consonancia con el Manual de operaciones del Equipo Vértice (2007), este modelo considera un sistema con k estaciones en serie, se presume que las llegadas a la estación provienen de una población infinita de acuerdo con una distribución de Poisson con tasa media de llegada □. Las unidades atendidas pasarán sucesivamente de una estación a la siguiente hasta que salgan por la estación k. Se puede apreciar en la figura 27.

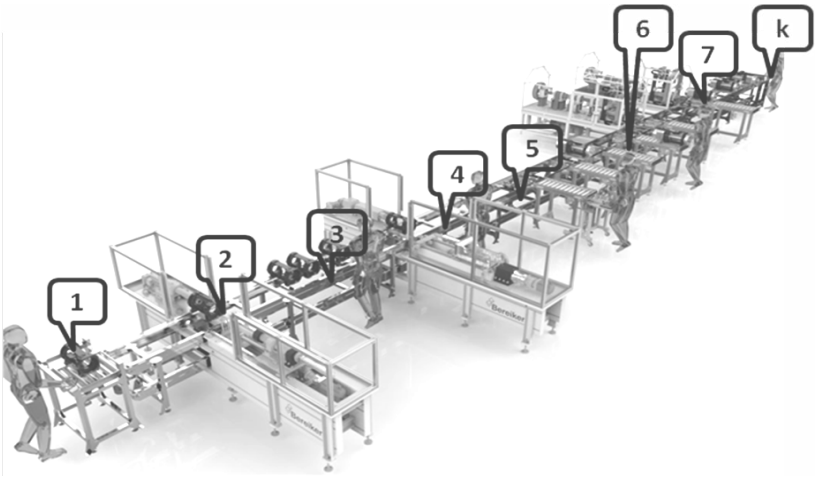


Figura 27. Línea de ensamblaje
Fuente: Adaptación de Bereiker (2018: párr.1).

La distribución del tiempo de servicio en cada estación i es exponencial con tasa media □$_i$ para i = 1,2,..., k. En estas condiciones puede probarse que para i la salida de la estación i sigue una distribución de Poisson con tasa media □ y que cada estación puede tratarse de forma independiente como un modelo (M/M/1):(DG/□□);

significa que para la i-ésima estación las probabilidades de estado estable $p_{n,\,i}$ están dadas por:

$$p_{n,\,i} = (1-\rho_i)\,\rho_i^{ni} \quad n_i = 0, 1, 2 \ldots \text{ para } i = 1, 2, 3,\ldots,k$$

Donde n_i es el número en el sistema que solo consta de la estación i. Los resultados del estado estable solo existirán si: $\rho_i = \lambda/c_i\mu_i < 1$. 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 μ_i por unidad de tiempo. En este caso cada estación puede ser tratada como un modelo (M/M/c):(DG/∞/∞) con tasa media de llegadas λ. Los resultados de estado estable prevalecerán si $\lambda = c_i\mu_i$ para i = 1, 2,..., k. Así se indica en la figura 28.

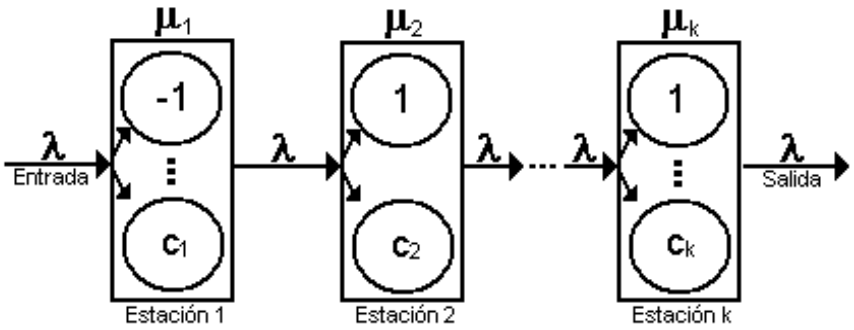


Figura 28. Modelo en serie k de estaciones con capacidad de colas infinita
Fuente: Equipo Vértice (2007: 32).

Selección del Modelo Apropiado de Líneas de Espera

Para desarrollar estos puntos se tomarán las ideas de Taha (2017). En tal sentido, la aplicación de la teoría de espera en la práctica implica dos aspectos:

1. Selección del modelo matemático que representará al sistema real en forma apropiada con el fin de determinar las medidas de desempeño del sistema.

2. Implantación de un modelo de decisión basado en las medidas de desempeño del sistema a objeto de diseñar las instalaciones o la disposición del servicio.

La escogencia de un modelo específico para analizar una línea de espera está signado por las distribuciones de los tiempos de llegada y de servicio; en la práctica la determinación de estas distribuciones implica observar las líneas de espera durante su operación y registrar las observaciones. Esta recolección de información debe hacerse de acuerdo con:

1. ¿Cuándo observar el sistema?

2. ¿Cómo registrar los datos?

En la mayoría de las líneas de espera existen los periodos ocupados, durante los cuales la tasa de llegadas al sistema crece en comparación con otros periodos del día; ejemplo, son las horas picos en las autopistas. En estas situaciones se debe recopilar datos, pues el sistema debe ser diseñado para soportar el nivel de demanda a que puede estar sometido en cualquier momento del día o de la jornada.

La recolección de datos se puede realizar bien midiendo el tiempo entre las llegadas o las salidas sucesivas para establecer los tiempos entre arribos o entre salidas; bien contando el número de llegadas o de salidas ocurridas durante una unidad de tiempo. El primer método proporciona las distribuciones de tiempos entre arribos o servicios y el segundo arroja las distribuciones del número de llegadas o de salidas. Estos métodos son la forma escogida para poder explicar los procesos de entrada y salida en los modelos de líneas de espera. La figura 29 da una representación de la curva que surge ante dichos fenómenos.

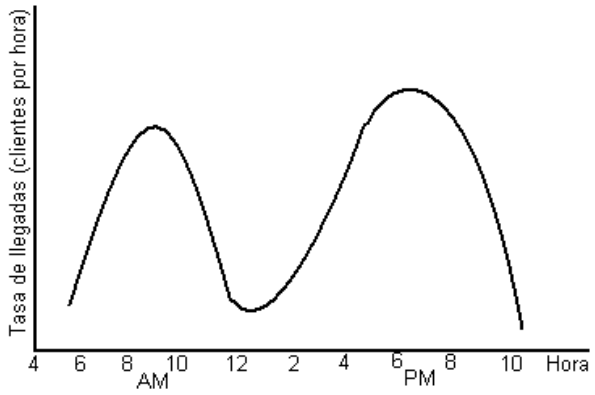


Figura 29. Variación característica en la tasa de llegada

Fuente: Taha (2012: 680)

Modelos de Decisión en Líneas de Espera

De acuerdo con Quesada y Vergara (2006) se deben perfilar modelos de decisión que se puedan usar en la optimización del diseño de los sistemas de línea de espera; a fin de minimizar los costos totales asociados con la operación de líneas de espera. Debido a que los modelos de costos en líneas de espera buscan equilibrar los costos de espera contra los costos de incrementar el nivel de servicio; puesto que, conforme crece el nivel de servicio, los costos de este también crecen y disminuye el tiempo de espera del cliente; así pues, el nivel de servicio óptimo se presenta cuando la suma de los dos costos es mínimo. En la figura 30 se presenta la situación ideal.

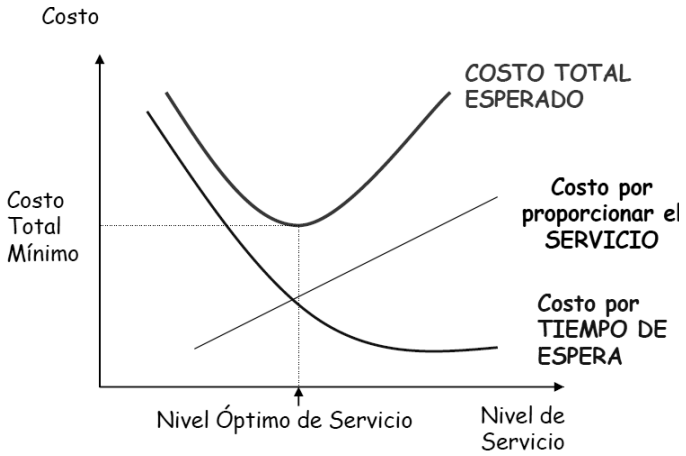


Figura 30. Equilibrio entre costos de espera y costos de servicio

Fuente: García y Oscar (2018: 37).

Algunos tipos de situaciones de líneas de espera pueden impedir el uso de modelos de decisión de costos; pues el costo de esperar es difícil de valorar, así como el grado de participación humana en la operación de las instalaciones. Los sistemas humanos son los más complejos y pueden ser vistos de dos formas: unas, situaciones donde los intereses del cliente y del servidor son mutuos; otras, sistemas donde los intereses pueden entrar en conflicto. Cuando los intereses del cliente y del servidor concilian se puede generar situación de niveles de productividad aceptable; pero cuando son conflictivos los intereses, los tiempos de espera dependen de la óptica de los clientes y del grado de satisfacción que le arguya a lo que va a obtener tras la espera.

Modelos de Costos

Los modelos de costos, como bien lo señalan Gross y Harris (2001) tratan de equilibrar los costos en conflicto, a saber: costos de ofrecer el servicio y costos que resultan de la demora en el ofrecimiento del servicio.

Tasa Óptima de Servicio

Siguiendo con Gross y Harris (2001), este modelo de costo trata la situación de un solo servidor donde se conoce la tasa λ de llegadas y se busca determinar la tasa μ óptima. 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 establecer el valor μ que minimiza la suma de esos dos costos; las formas de CEO y CEE como funciones de μ dependen del sistema en estudio. Un ejemplo ilustrativo es el siguiente, una compañía de fotocopiado presenta los siguientes datos:

Tipo de copiadora	Costo de operación (por hora)	Rapidez (hojas por minuto)
1	Bs. 15	30
2	Bs. 20	36
3	Bs. 24	50
4	Bs. 27	66

Los pedidos llegan a la empresa según una distribución de Poisson, a razón de 4 cada 24 horas; la cantidad de cada pedido es aleatorio, se estima que el promedio es de 10.000 copias, los contratos con los clientes establecen una multa por entrega tardía de Bs. 80 por día y por pedido. Empleando un tamaño promedio por pedido de 10.000 copias, las tasas de servicio de las diferentes copiadoras son:

Tipo de copiadora	Tasa de servicio, μ (pedidos por día)
1	4,32
2	5,18
3	7,20
4	9,50

El cálculo de los valores μ en esta tabla es:

Tiempo por pedido promedio $= \frac{10.000}{30} * \frac{1}{60} = 5{,}56 horas$.

Tasa de servicio correspondiente $= \frac{24}{5{,}56} = 4{,}32 \, pedidos / día$.

Un modelo de costo apropiado reconoce que μ se encuentra en cuatro valores discretos, correspondientes a los cuatro tipos de copiadoras, 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 hace tomando un día como unidad de tiempo, el costo de operar la instalación por día está dado por:

$$CEO_i = 24C_i \quad i = 1, 2, 3, 4$$

Se debe recordar el costo por multa $CEE_i = 80L_{si}$.

Entonces, $CET_i = 24C_i + 80L_{si}$. Se aplican las fórmulas del (M/M/1): (DG/∞/∞)

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

Costos por día:

Copiadora	CEO_i	CEE_i	CET_i
1	Bs. 360	Bs. 1.080,0	Bs. 1.440,0
2	Bs. 480	Bs. 351,20	Bs. 831,20
3	Bs. 576	Bs. 180,00	Bs. 756,00
4	Bs. 648	Bs. 138,16	Bs. 786,16

La copiadora 3 tiene el menor costo total por día.

Número Óptimo de Servidores

Explica Taha (2012), si c es el número de servidores en paralelo, el problema es determinar el valor de c que minimiza. Así, el valor óptimo de c debe satisfacer las siguientes condiciones:

$$CET(c-1) \geq CET(c) \qquad y \qquad CET(c+1) \geq CET(c)$$

Las funciones de costo son:

$CEO(c) = C_1c$

$CEE(c) = C_2L_s(c)$

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.

$$L_s(c) - L_s(c+1) \leq \frac{C_1}{C_2} \leq L_s(c-1) - L_s(c)$$

Un ejemplo del mismo autor, Taha (2017), es una instalación de almacenamiento de herramientas, las solicitudes de intercambio ocurren según una distribución de 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 a la instalación se estima en Bs. 6 por hora. El costo de la producción perdida por máquina en espera por hora se estima en Bs. 30 la hora. A ¿cuántos empleados debe contratar la instalación?

c	$L_s(c)$	$L_s(c-1) - L_s(c)$	
1	∞		
2	7,467	∞	
3	2,217	5,25	
4	1,842	0,375	$\leftarrow C_1 / C_2 = 0{,}2$
5	1,469	0,073	
6	1,754	0,015	
7	1,75	0,004	

Ya que $C_1 / C_2 = 6 / 30 = 0{,}2$ se tiene:

$$L_s(4) - L_s(5) = 0{,}073 < 0{,}2 < 0{,}375 = L_s(3) - L_s(4).$$

El óptimo es C = 4 empleados.

Modelo de Nivel de Aceptación

Revelan Anderson, Sweeney, Williams, Camm y Kipp (2011), este modelo reconoce la dificultad de estimar los parámetros de costos y emplea las características de operación del sistema al decidir sobre los valores óptimos de los parámetros de diseño; lo óptimo está referido a satisfacer ciertos niveles de aceptación establecidos por el decidor, definidos como límites superiores sobre los calores de las medidas conflictivas que desea balancear el decidor.

En el modelo de servidores múltiples donde se requiere determinar el valor óptimo del número c de servidores, las dos medidas en conflicto pueden tomarse:

Tiempo promedio de espera en el sistema W_s.

1. Porcentaje X de tiempo inactivo de los servidores

Estas dos medidas reflejan las aceptaciones del cliente y del servidor, sean α y β los niveles de aceptación, límites superiores para W_s y X, el método de nivel de aceptación puede expresarse así:

$$W_s \leq \alpha \qquad y \qquad X \leq \beta$$

La expresión para W_s, deviene de (m/M/c):(DG/∞/∞). La expresión de X está dada por:

$$X = \frac{100}{c}\sum_{n=0}^{c}(c+n)p_n = 100\left(1-\frac{\rho}{c}\right)$$

La solución a un problema planteado se puede obtener más rápidamente graficando W_s y X en función de c, localizando α y β se puede determinar inmediatamente un intervalo aceptable de c que satisfaga ambas restricciones. No obstante, si estas dos restricciones no se satisfacen simultáneamente, es necesario relajar una o ambas

restricciones antes de tomar una decisión. Su demostración gráfica se presenta en la figura 31.

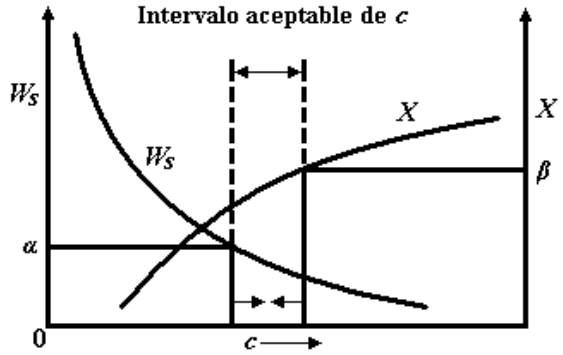


Figura 31. Intervalo aceptable de *c*

Fuente: Taha (2017: 687).

APORTES BÁSICOS DE LA TEORÍA DE COLAS A LA ADMINISTRACIÓN Y LA GERENCIA

Al describir los aportes básicos de la teoría de colas se entiende la necesidad de establecer mecanismos de manejo y gestión de las líneas de espera, que generan retrasos en la satisfacción de las necesidades de los clientes, así como cumplir con los requerimientos de los demandantes y atender a los usuarios. Estas situaciones van generando costos económicos, financieros, humanos, reputacionales, que terminan afectando la rentabilidad, la estabilidad y la sostenibilidad de las organizaciones en el mediano y largo plazo. Adicionalmente, las colas son situaciones que, dependiendo del área o ámbito donde ocurran, pueden distorsionar, entorpecer e, incluso, detener el proceso productivo; pues, la espera por componentes, partes o recursos para la continuidad de la marcha productiva ralentiza y puede llegar a detener el proceso operativo-productivo de la empresa, generando pérdidas tangibles e intangibles, que pueden llegar a ser incalculables por su impacto presente y futuro.

En síntesis:

Las colas generan retrasos en la satisfacción de las necesidades de los clientes.

La espera impide cumplir con los requerimientos de los demandantes y atender a los usuarios.

Las filas van generando costos económicos, financieros, humanos, reputacionales, afectando la estabilidad y sostenibilidad de las organizaciones en el mediano y largo plazo.

Balance entre Rapidez, Solución y Rentabilidad

En ese sentido, comprender los aportes básicos de la teoría de colas a la administración y la gerencia pasa por asumir la importancia de evitar los «cuellos de

botella», que son las colas o líneas de espera interminables, o en todo caso, reducir al mínimo su existencia. El éxito de una empresa depende del servicio que brinda al cliente o usuario, de ahí que sea muy importante considerar el tiempo que se demora la persona en realizar una operación dentro de la misma y, en tal sentido, este debe ser el más corto posible, sin que ello implique no realizar a cabalidad la actividad o sin que este no le engendre satisfacción al cliente. Pero, al mismo tiempo, debe implicar rapidez y efectividad para la organización de modo de garantizar rentabilidad y continuidad de uso. Por eso, es imprescindible que los modelos de colas ayuden a encontrar el balance entre rapidez, solución y rentabilidad.

Así pues, explicar los aportes básicos de la teoría de colas a la administración y la gerencia da la posibilidad de establecer que los modelos de líneas de espera son muy útiles para determinar cómo operar un sistema de colas de la manera más efectiva, eficaz y eficiente posible. Además, estos modelos permiten determinar si proporcionar demasiada capacidad de servicios para operar el sistema implica costos excesivos; pero, sin generar problemas. Pues, no contar con suficiente capacidad de servicio puede incrementar la espera, tiempo perdido del cliente, pérdida de operatividad de la organización, así como el prestigio de la empresa y a los clientes mismos. Es por eso que se puede afirmar que los modelos permiten encontrar un balance adecuado entre el costo de servicio y la cantidad de espera que debe presentar una organización o una empresa para garantizar rentabilidad y eficiencia.

Siempre que existe más de un usuario de un recurso limitado se puede formar una cola o línea de espera. Cuando la cola se compone de objetos inanimados (materiales) que esperan algún tipo de procesamiento, el problema es económico, cuánto equipo es necesario para poder producir u operar satisfactoriamente. Cuando la cola está formada por personas que esperan un servicio, el problema tiene además de aspectos económicos, psicológicos, que responden a estados subjetivos del ser y en términos matemáticos, suelen estar asociados más con valores discordantes que con elementos de cuantificación de interés para el sistema.

Aportes de los Modelos Matemáticos de la Teoría de Colas en la Administración y la Gerencia

Es esencial acotar que los resultados del modelo (M/M/□):(DG/□□) se pueden utilizar para determinar aproximadamente los del modelo (MMc):(DG/□□) cuando c crece lo suficiente. La ventaja que ofrece es que las operaciones son más sencillas en el modelo (M/M/□):(DG/□□). Debido a que, un tiempo de servicio constante indica mayor certeza en la operación de la empresa que se está analizando, sobre todo si se reduce el tiempo de espera calculado.

La teoría de colas es una gran ayuda para la toma de decisiones administrativas, el tener muchos clientes esperando en cola algún servicio, por largo tiempo, tiende a hacer que esos clientes se pierdan o se decepcionen de la organización. De ahí, que al plantear los modelos de colas se puede analizar de forma precisa el rendimiento del sistema. La teoría de las colas en si no resuelve directamente el problema de los tiempos de espera que se pueden generar en una organización cualquiera; pero contribuye con la información vital que se requiere para tomar las decisiones concernientes, prediciendo algunas características sobre la línea de espera, a saber:

- probabilidad que se formen y

- tiempo de espera promedio, por mencionar sólo dos.

Si se emplea el concepto de clientes internos en la empresa, asociándolo a la teoría de las colas, existe un acercamiento al modelo de organización empresarial o filosofía justo a tiempo, en la que se trata de minimizar el costo asociado a la ociosidad de recursos en la cadena productiva.

De ese modo, la teoría de colas es una técnica analítica de investigación de operaciones, pero se ha limitado a la formulación de una teoría matemática descriptiva. Es preciso que se convierta en elemento decisivo en los procesos de

toma de decisiones administrativas-gerenciales óptimas. No basta con conseguir información sobre el comportamiento del sistema de colas, sino hacer que esa información recabada y analizada sea una herramienta gerencial-decisional.

Existen casos en donde la población potencial del sistema es finita, la cola de la disciplina no es Primeras entradas, primeras salidas (PEPS) o FIFO, la tasa de servicio depende de las personas en la cola y las distribuciones de las llegadas no son de Poisson; en estos casos los modelos tienden a ser incompatibles. De ahí, la necesidad de ampliar los parámetros de estudio y evaluación que contemplen elementos de tipo subjetivo y cualitativo para ampliar los horizonte de decisión de los gerentes.

Allí es donde son fundamentales las categorías aludidas porque orientan al gerente y facilitan la toma de decisiones. Reconocer que la eficiencia del servicio, la satisfacción del cliente, la asertividad de la toma de decisiones, el compromiso de los empleados, la fidelización de los usuarios y la sostenibilidad de la organización no son parámetros contrarios, sino concomitantes que deben perseguirse en conjunto para asegurar la rentabilidad de la empresa, lo cual resulta en un hecho fundamental. Además, facilita que la organización pueda gestionar rápidamente los cuellos de botellas y lograr subsanar problemas subsecuentes o asociados a las esperas largas y recurrentes que pueden surgir tanto en los empleados como en los clientes, pasando por la propia organización en su conjunto.

En síntesis, se puede decir que al describir los aportes básicos de la teoría de colas se entiende la necesidad de establecer mecanismos de gestión de las líneas de espera, que generan retrasos en la satisfacción de las necesidades de los clientes, cumplir con los requerimientos de los demandantes y atender a los usuarios. Estas situaciones van generando costos económicos, financieros, humanos, reputacionales que terminan afectando la estabilidad y sostenibilidad de las organizaciones en el mediano y largo plazo.

Adicionalmente, las colas son situaciones que, dependiendo del área o ámbito donde ocurran, pueden distorsionar, entorpecer e, incluso, detener el proceso productivo; pues, la espera por componentes, partes o recursos para la continuidad de la marcha productiva detiene y ralentiza el proceso generando pérdidas tangibles e intangibles, que pueden llegar a ser incalculables por su impacto presente y futuro.

En ese sentido, comprender los aportes básicos de la teoría de colas a la administración y la gerencia pasa por asumir la importancia de evitar los cuellos de botella que son las colas o líneas de espera interminables, en todo caso, reducir al mínimo su existencia.

El éxito de una empresa depende del servicio que brinda al cliente, de ahí que sea muy importante el tiempo que se demora la persona en realizar una operación dentro de ésta y, en tal sentido, debe ser el más corto posible, sin que ello implique no realizar a cabalidad la actividad o sin que este no le engendre satisfacción al cliente. Pero, al mismo tiempo, debe implicar rapidez y efectividad para la organización de modo de garantizar rentabilidad y continuidad de uso. Por eso, es imprescindible que los modelos de colas ayuden a encontrar el balance entre rapidez, solución y rentabilidad.

Explicar los aportes básicos de la teoría de colas a la administración y la gerencia da la posibilidad de establecer que los modelos de líneas de espera son útiles para determinar cómo operar un sistema de colas de la manera más efectiva, eficaz y eficiente posible. Además, estos modelos permiten determinar si proporcionar demasiada capacidad de servicios para operar el sistema implica costos excesivos sin generar problemas; pues, no contar con suficiente capacidad de servicio puede incrementar la espera, el tiempo perdido del cliente, el prestigio de la empresa, a los clientes mismos.

Por eso se puede afirmar que los modelos permiten encontrar un balance adecuado entre el costo de servicio y la cantidad de espera que debe presentar una organización para garantizar rentabilidad y eficiencia.

La teoría de colas es una gran ayuda para la toma de decisiones administrativas, el tener muchos clientes esperando en cola algún servicio, por largo tiempo, tiende a hacer que esos clientes se pierdan o decepcionen de la organización. De ahí, que al plantear los modelos de colas se puede analizar de forma precisa el rendimiento del sistema. La teoría de las colas no resuelve el problema de los tiempos de espera que se pueden generar en una organización cualquiera; pero, contribuye con la información que se requiere para tomar las decisiones concernientes, prediciendo algunas características sobre la línea de espera: probabilidad que se formen y el tiempo de espera promedio, por mencionar sólo dos características.

Si se emplea el concepto de clientes internos en la organización de la empresa, asociándolo a la teoría de las colas, existe un acercamiento al modelo de organización empresarial o filosofía justo a tiempo, en la que se trata de minimizar el costo asociado a la ociosidad de recursos en la cadena productiva. La teoría de colas es preciso que se convierta en elemento decisivo en los procesos de toma de decisiones óptimas. No basta con conseguir información sobre el comportamiento del sistema de colas, sino hacer que esa información recabada y analizada sea una herramienta decisional.

Existen casos en donde la población potencial del sistema es finita, la cola de la disciplina no es PEPS o FIFO la tasa de servicio depende de las personas en la cola y las distribuciones de las llegadas no tienen una probabilidad de ocurrencia en un periodo establecido; en estos casos los modelos tienden a ser incompatibles. De ahí, la necesidad de ampliar los parámetros de estudio que contemplen elementos de tipo subjetivo y cualitativo, mejorando su utilidad para la toma de decisiones gerenciales.

Allí es donde son fundamentales las categorías: eficiencia del servicio, la satisfacción del cliente, la asertividad de la toma de decisiones, el compromiso de los empleados, la fidelización de los usuarios y la sostenibilidad de la organización; porque orientan al gerente y facilitan la toma de decisiones. Reconocer que son parámetros concomitantes que deben perseguirse en conjunto para asegurar la rentabilidad de la organización. Además, facilitan que la organización pueda gestionar rápidamente los cuellos de botellas y corregir problemas asociados a las esperas recurrentes, tanto en los empleados como en los clientes, pasando por la propia organización en su conjunto.

REFERENCIAS

Anderson, D., Sweeney, D., Williams, T., Camm, J. y Kipp, M. 2011. Métodos cuantitativos para los negocios. 11va. Ed. Cengage Learning. México.

Bereiker. 2018. Automatización del ensamblaje de 50 modelos diferentes de motores. [Documento en línea]. En: https://www.bereiker.com/es/linea-de-ensamblaje-de-motores/. [Consulta: Mayo, 20 de 2021].

Caba, N., Chamorro, O. y Fontalvo, T. 2011. Toma de decisiones a través de la investigación de operaciones. [Documento en línea]. En: https://1library.co/document/nq740ddq-villalobos-oswaldo-chamorro-altahona-tomas-jose-fontalvo-herrera.html. [Consulta: Abril, 13 de 2021].

Campos, L. 2016. Investigación de operaciones Teoría de colas. [Documento en línea]. En: https://www.slideshare.net/laura_campos07/teoria-de-colasmodelommc?from_action=save. [Consulta: Mayo, 19 de 2021].

Cao, R. 2002. Introducción a la simulación y a la teoría de colas. Netbiblo, S.L. La Coruña.

Carro, R. 2014. Investigación de operaciones en administración. Universidad de Mar de Plata. Argentina.

Carro, R. y González, D. 2015. Administración de las operaciones. Universidad Nacional de Mar del Plata – Facultad de Ciencias Económicas y Sociales. Mar del Plata.

De La Fuente, D. y Pino, R. 2007. Teoría de las líneas de espera. Modelos de cola. Universidad de Oviedo – Servicio de Publicaciones. Oviedo, España.

Equipo Vértice. 2007. Dirección de operaciones. Publicaciones Vértice. España.

Fonollosa, J., Sallán, J. y Suñe, A. 2005. Métodos cuantitativos de organización industrial II. 2da. Ed. Centro de Publicaciones de Campo Norte. Barcelona, España.

García, Ch. y Oscar, A. 2018. Teoría de redes. [Documento en líneas]. En: https://www.slideshare.net/MaberyRivera/teora-de-redes-teora-de-colas?from_ act ion=save. [Consulta: Mayo, 19 de 2021].

Gestión de Operaciones. 2015. Simulación de una Línea de Espera M/M/1 (Teoría de Colas) en Excel. [Documento en línea]. En: https://www.gestiondeoperaciones.net/lineas-de-espera/simulacion-de-una-linea-de-espera-mm1-teoria-de-colas-en-excel/. [Consulta: Mayo, 19 de 2021].

Gross, D. y Harris, C. 2001. Métodos cuantitativos de organización industrial. Departamento de organización de empresas. Colombia.

Hieller, F. y Lieberman, G. 2021. Introduction to Operations Research. 11va. Ed. Educación McGraw-Hill. Londres.

Ingeniería de Sistemas. 2016. Provisión de QoS - CCNA1 V5 - CISCO C1. [Documento en línea]. En: http://www.ingenieriasystems.com/2016/07/provision-de-qos-ccna1-v5-cisco-c1.html. [Consulta: Mayo 19 de 2021].

Jaime, J. 2013. Teoría de líneas de espera en el sector avícola para el diseño de muelles de despacho. [Documento en línea] En: https://repository.unimilitar.edu.co/bitstream/handle/10654/11013/L%EDneas%20de%20espera.pdf;jsessionid=D70660A589123E8C32F1A2600A3B832A?sequence=1. [Consulta: Mayo, 19 de 2021].

Leandro, G. 2004. Líneas de espera: teoría de colas. Curso métodos cuantitativos. [Documento en línea]. En: htpp://www.auladeeconomia.com. [Consulta: Abril, 15 de 2021].

Marrero, F., Asencio, J., Abreu, R., Orozco, R. y Granela, H. 2006. Herramientas para la toma de decisiones: la teoría de colas. Universidad Central de las Villas. Santa Clara, Cuba.

Moskowitz, Herbert y Gordon P. Wright. 1993. Investigación de Operaciones. Prentice-Hall. México.

Quesada, V. y Vergara, J. 2006. Análisis cuantitativo en WINQSE. Universidad de Cartagena. Cartagena.

Render, B., Stair, R. y Hanna, M. 2018. Métodos cuantitativos para los negocios. 13. Ed. Pearson / Prentice Hall. México.

Rincón, L. 2001. Investigación de operaciones para ingenierías y administración de empresas. FERIVA y Universidad Nacional de Colombia. Palmira, Colombia.

Rodríguez, R. y Gámez, A. 2002. Investigación operativa. Teoría, ejercicio y práctica con ordenadores. Universidad de Cádiz – Servicio de Publicaciones. Cádiz.

Santiago, H. 2017. Teoría de colas o de líneas de espera. [Artículo en línea]. En: https://www.emprendices.co/teoria-colas-lineas-espera/. [Consulta: Abril, 16 de 2021].

Sarabia, Á. 1996. La investigación operativa. Universidad Pontifica Comillas de Madrid. España.

Serra, D. 2002. Métodos cuantitativos para la toma de decisiones (con aplicaciones en el ámbito sanitario). Fundación BBV. Bilbao, España.

Taha, H. 1995. Investigación de operaciones. 5ta. Ed. Alfaomega. México.

Taha, H.2012. Investigación de operaciones. 9na. Ed. Pearson Educación. México.

Taha, H. 2017. Operations research: introduction. 10ma. Ed. Pearson Education Limited. Malaysia.

Vega, J. 2004. Diseño e implementación de una herramienta para la enseñanza y el aprendizaje de la teoría de cola.

Villalobos, J. 2014. Fenómenos de espera: experiencia. [Artículo en línea]. En: https://jimmysblogsp.wordpress.com/2014/10/22/fenomenos-de-espera-experiencia/. [Consulta: Abril, 13 de 2021].

Winston, W. 2005. Investigación de operaciones. Aplicaciones y algoritmos. Volumen II. 4ta. Ed. Internacional Thomson Editores S.A. México.

Zaragoza, A. 2004. Teoría de cola. [Documento en línea]. En: http://exa.unne.edu.ar/depar/areas/informatica/evalua/teoria_de_cola.pdf. [Consulta: Abril, 13 de 2021].

Aportes básicos de la **TEORÍA DE COLAS** a la **administración** y la **gerencia**

La relevancia y pertinencia de este libro nace de aprovechar los recursos para la mejora de los procesos organizacionales buscando ampliar la perspectiva teórica sobre la incidencia de las líneas de espera en las organizaciones desde la administración y la gerencia; también plantea ampliar los márgenes de certidumbres de los decisores en los diferentes ámbitos organizacionales.

La explicación se aborda en tres etapas: una, describir los aportes de la teoría de colas; dos, comprender tales aportes para la administración y la gerencia; tres, su interpretación. Adicionalmente, se establecen varias categorías: eficiencia del servicio, satisfacción del consumidor o cliente, asertividad y oportunidad de la toma de decisiones, engagement o compromiso de los empleados, fidelización de los usuarios, sostenibilidad de la organización.

Todo ello con miras a establecer mecanismos de manejo y gestión de las líneas de espera, que generan retrasos en la satisfacción de necesidades de los clientes e incremento de los costos de la organización.

ISBN: 978-980-248-312-9