Líneas de Espera

Líneas de espera (Colas)

Las líneas de espera, filas de espera o colas, son realidades cotidianas:

- Personas esperando para realizar sus transacciones ante una caja en un banco, estudiantes esperando por obtener copias en la fotocopiadora, vehículos esperando pagar ante una estación de peaje o continuar su camino, ante un semáforo en rojo, Maquinas dañadas a la espera de ser rehabilitadas.

Definiciones

- 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 buen compromiso entre costos del sistema y los tiempos promedio de la línea de espera para un sistema dado.

- Los sistemas de colas son modelos de sistemas que proporcionan servicio. Como modelo, pueden representar cualquier sistema en donde los trabajos o clientes llegan buscando un servicio de algún tipo y salen después de que dicho servicio haya sido atendido.

- La teoría de colas es el estudio matemático del comportamiento de líneas de espera. Esta se presenta, cuando los “clientes” llegan a un “lugar” demandando un servicio a un “servidor”, el cual tiene una cierta capacidad de atención. Si el servidor no esta disponible inmediatamente y el cliente decide esperar, entonces se forma la línea de espera.

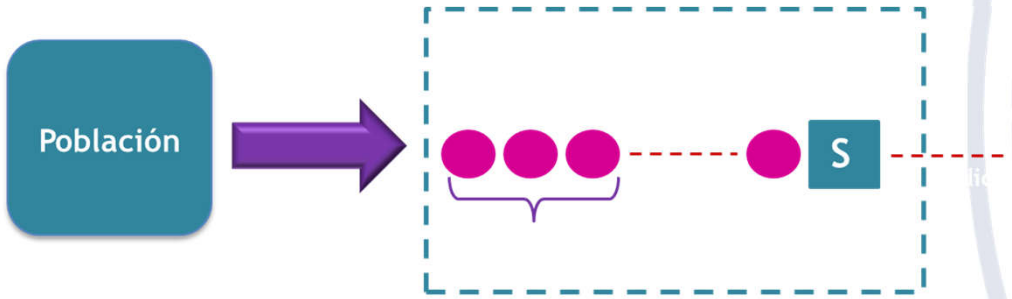
Objetivos del estudio de teoría de colas

- Identificar el nivel óptimo de capacidad del sistema que minimiza el costo global del mismo.
- Evaluar el impacto que las posibles alternativas de modificación de la capacidad del sistema tendrían en el coste total del mismo.
- Establecer un balance equilibrado (“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.

Tipos de Colas

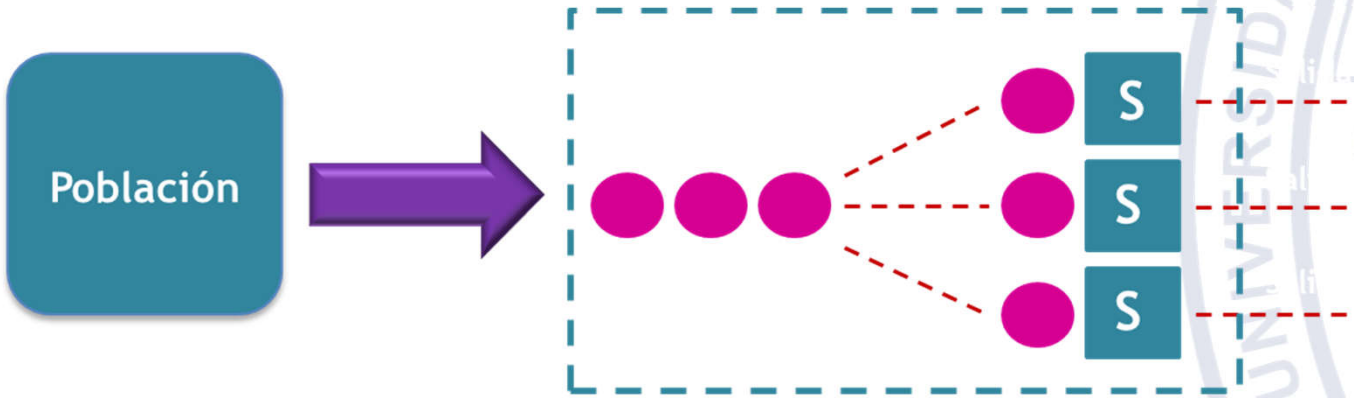
Según el tipo de sistema de colas, tenemos varios tipos de éstas, las cuales son:

- Una línea, un servidor : El primer sistema que se muestra se llama un sistema de un servidor y una cola o puede describir una consulta de un médico.



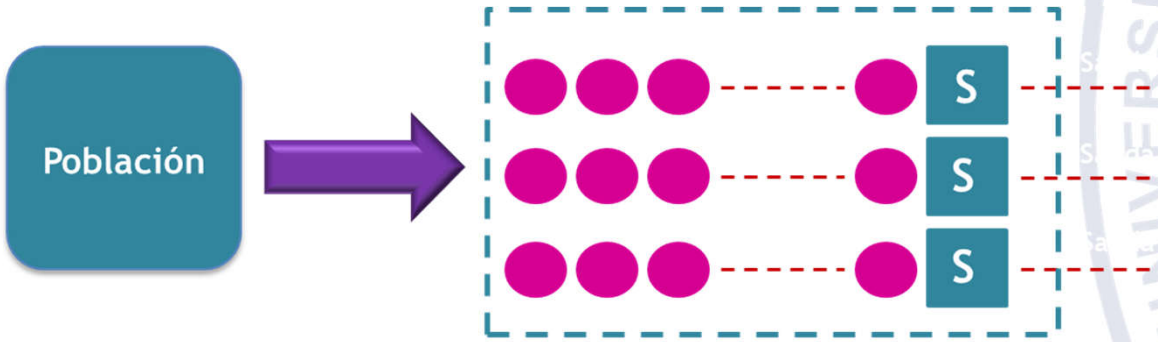
Tipos de colas

- Una línea, múltiples servidores: El segundo, una línea con múltiples servidores, es típico de una peluquería o una panadería en donde los clientes toman un número al entrar y se les sirve cuando les llega el turno.



Tipos de colas

- Varias líneas, múltiples servidores: cada servidor tiene una línea separada, es característico de los bancos y las tiendas de autoservicio. Para este tipo de servicio pueden separarse los servidores y tratarlos como sistemas independientes de un servidor y una cola.



Ejemplos

COLAS MÁS COMUNES		
SITIO	**ARRIBOS EN COLA**	**SERVICIO**
Supermercado	Compradores	Pago en cajas
Peaje	Vehículos	Pago de peaje
Consultorio	Pacientes	Consulta
Sistema de Cómputo	Programas a ser corridos	Proceso de datos
Compañía de teléfonos	Llamadas	Efectuar comunicación
Banco	Clientes	Depósitos y Cobros
Mantenimiento	Máquinas dañadas	Reparación
Muelle	Barcos	Carga y descarga

Definiciones

Fuente

- Recibe el nombre de fuente el dispositivo del que emanan las unidades que piden un servicio. Si el numero de unidades potenciales es finito, se dice que la fuente es finita; en caso contrario se dice que es infinita.

Tiempo entre llegadas

- Existen dos clases básicas de tiempo entre llegadas:
- Determinístico, en el cual clientes sucesivos llegan en un mismo intervalo de tiempo, fijo y conocido. Un ejemplo clásico es el de una línea de ensamble, en donde los artículos llegan a una estación en intervalos invariables de tiempo.

- Probabilístico, en el cual el tiempo entre llegadas sucesivas es incierto y variable. Los tiempos entre llegadas probabilísticos se describen mediante una distribución de probabilidad.

Definiciones

Capacidad de servicio

- Se llama capacidad del servicio al numero de clientes que pueden ser servidos simultáneamente. Si la capacidad es uno, se dice que hay un solo servidor (o que el sistema es mono canal) y si hay mas de un servidor, multicanal. El tiempo que el servidor necesita para atender la demanda de un cliente (tiempo de servicio) puede ser constante o aleatorio.

Disciplina de la cola

- En sistemas mono canal, el servidor suele seleccionar al cliente de acuerdo con uno de los siguientes criterios (prioridades):
    - El que llego antes. (PEPS)
    - El que llego el ultimo. (UEPS)
    - El que menos tiempo de servicio requiere.
    - El que más requiere. (Prioridad)

Terminología y Notación

- **Características operativas.-** Medidas de desempeño para una línea de espera que incluyen la probabilidad de que no haya unidades en el sistema, la cantidad promedio en la línea, el tiempo de espera promedio, etc.

- **Operación de estado estable.-** Operación normal de la línea de espera después de que ha pasado por un periodo inicial o transitorio. Las características operativas de las líneas de espera se calculan para condiciones de estado estable.

Terminología y Notación II

- **Tasa media de llegada.-** Cantidad promedio de clientes o unidades que llegan en un periodo dado.

- **Tasa media de servicio.-** Cantidad promedio de clientes o unidades que puede atender una instalación de servicio en un periodo dado.

- **Línea de espera de canales múltiples.-** Línea de espera con dos o más instalaciones de servicio paralelas.

Terminología y Notación III

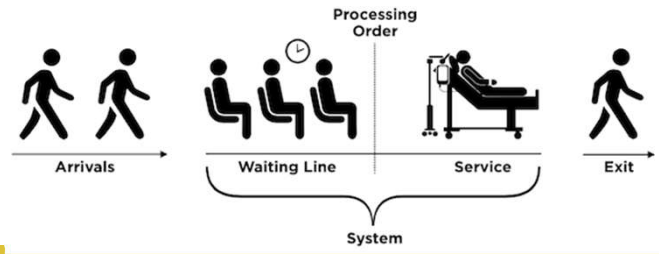
- **Bloqueado.-** Cuando las unidades que llegan no pueden entrar a la línea de espera debido a que el sistema esta lleno. Las unidades bloqueadas pueden ocurrir cuando no se permiten las líneas de espera o cuando las líneas de espera tienen una capacidad finita.

- **Población infinita.-** Población de clientes o unidades que pueden buscar servicio, no tiene un limite superior especificado.

- **Población finita.-** Población de clientes o unidades que pueden buscar servicio, tiene un valor fijo y finito.

Terminología y Notación IV

- El número de clientes en la cola es el número de clientes que esperan el servicio
- El número de clientes en el sistema es el número de clientes que esperan en la cola más el número de clientes que actualmente reciben el servicio
- La capacidad de la cola es el número máximo de clientes que pueden estar en la cola
- Generalmente se supone que la cola es infinita
- Aunque también la cola puede ser finita



Administración de las líneas de espera

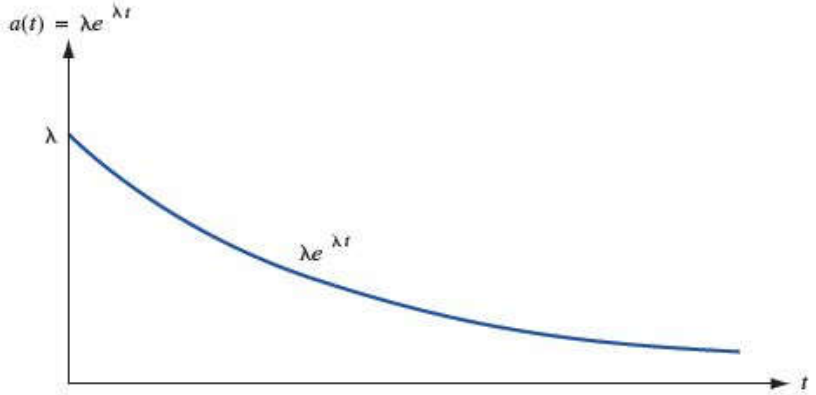
Los gerentes tienen un número de buenas razones para estar preocupados por las líneas de espera:

1. El costo de proporcionar espacio que espera.
2. Una pérdida posible de negocio debido a una porción de clientes que dejan la línea antes, o rechazan esperar en lo absoluto.
3. Una pérdida posible de buena voluntad.
4. Una reducción posible de satisfacción de cliente.
5. La congestión puede interrumpir otras operaciones de negocio y/o de clientes

Sistema de arribo –Dist. Exponencial

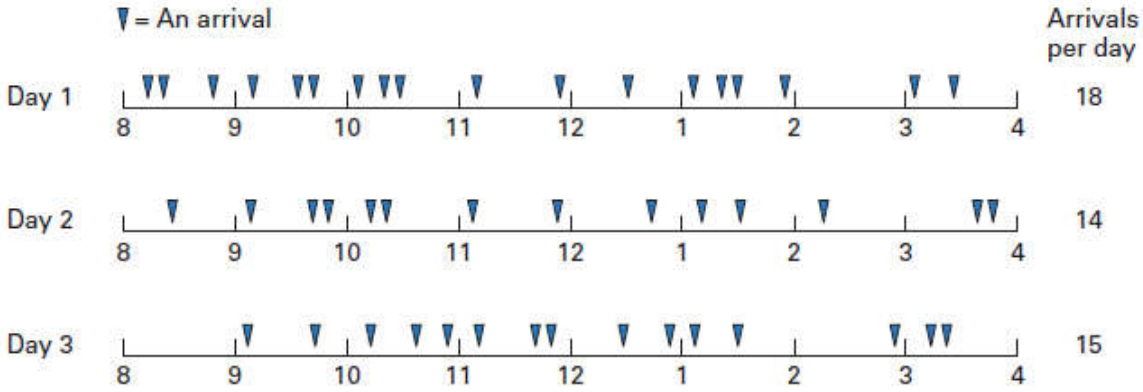
- La distribución exponencial supone una mayor probabilidad para tiempos entre llegadas pequeños,
- En general, se considera que las llegadas son aleatorias
- La última llegada no influye en la probabilidad de llegada de la siguiente,

$$a(t) = \lambda \cdot e^{\lambda t}$$



Sistema de arribo –Dist. Exponencial

- Es una distribución discreta empleada con mucha frecuencia para describir el patrón de las llegadas a un sistema de colas
- Para tasas medias de llegadas pequeñas es asimétrica y se hace más simétrica y se aproxima a la binomial para tasas de llegadas altas



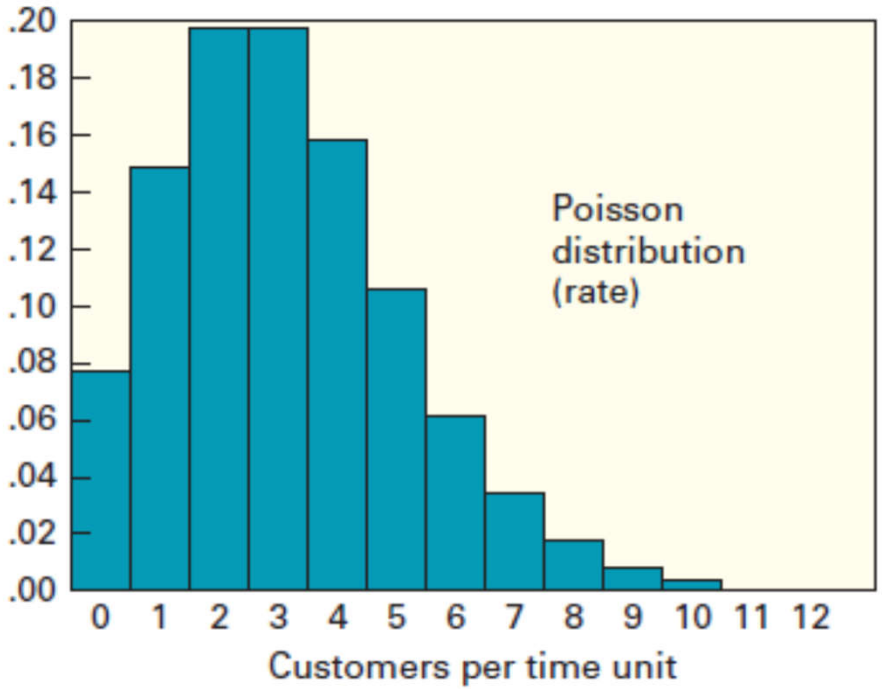
Sistema de arribo –Dist. Poisson

- Es una distribución discreta empleada con mucha frecuencia para describir el patrón de las llegadas a un sistema de colas
- Para tasas medias de llegadas pequeñas es asimétrica y se hace más simétrica y se aproxima a la binomial para tasas de llegadas altas

$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- Donde:
- P(k) : probabilidad de k llegadas por unidad de tiempo
- λ: tasa media de llegadas
- e = 2,7182818…

Sistema de arribo –Dist. Poisson



$$P_n = \frac{(\lambda T)^n}{n!} e^{-\lambda T}$$

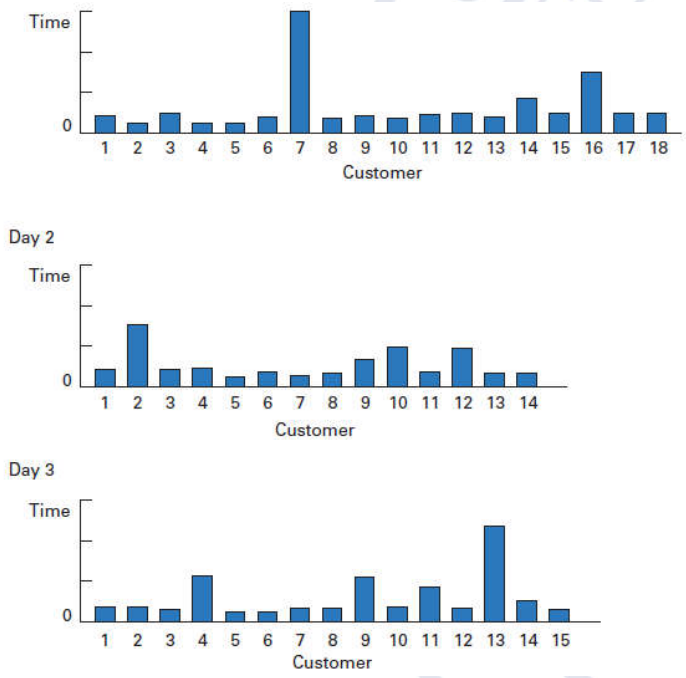
Sistema de servicio

Es necesario seleccionar una distribución de probabilidad para los tiempos de servicio,

Hay dos distribuciones que representarían puntos extremos:

- La distribución exponencial (σ=media)
- Tiempos de servicio constantes (σ=0)

$$P(t \leq T) = 1 - e^{-\mu T}$$



Sistema de arribo

- El tiempo que transcurre entre dos llegadas sucesivas en el sistema de colas se llama tiempo entre llegadas
- El tiempo entre llegadas tiende a ser muy variable
- El número esperado de llegadas por unidad de tiempo se llama tasa media de llegadas (λ)
- El tiempo esperado entre llegadas es $1/\lambda$
- Por ejemplo, si la tasa media de llegadas es $\lambda = 20$ clientes/hora
- Entonces el tiempo esperado entre llegadas es $1/\lambda = 1/20 = 0.05$ horas o 3 minutos

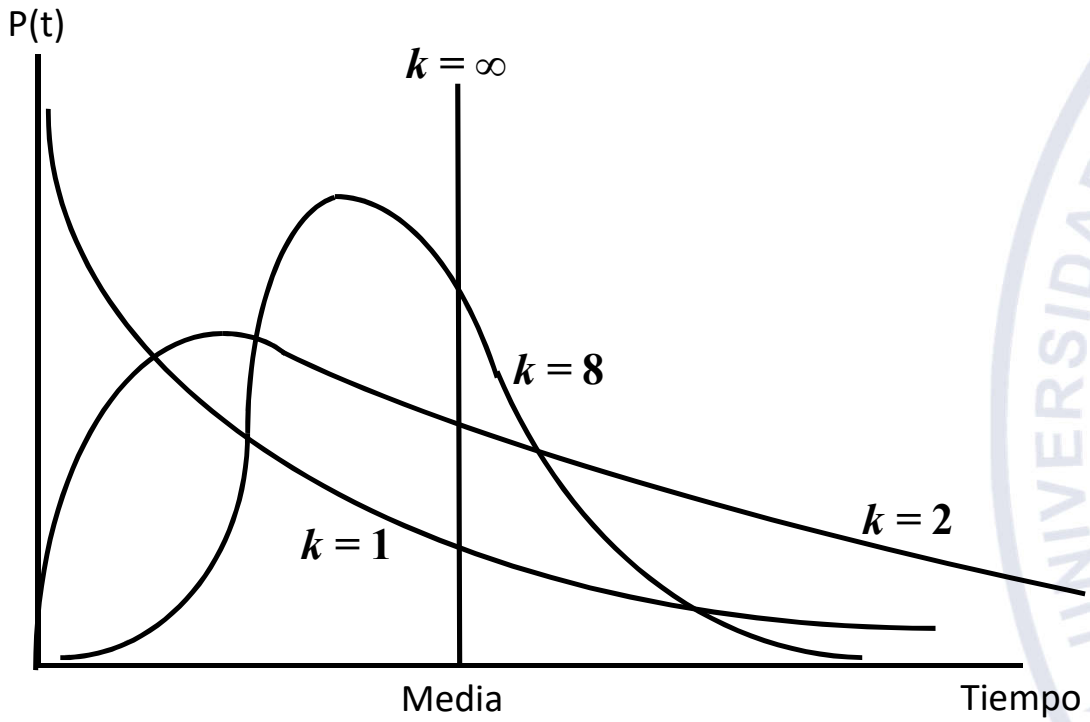
Sistema de servicio

- Una distribución intermedia es la distribución Erlang
- Esta distribución posee un parámetro de forma k que determina su desviación estándar:

$$\sigma = \frac{1}{\sqrt{k}} media$$

- Si k = 1, entonces la distribución Erlang es igual a la exponencial
- Si k = ∞, entonces la distribución Erlang es igual a la distribución degenerada con tiempos constantes
- La forma de la distribución Erlang varía de acuerdo con k

Sistema de servicio



Distribución Erlang

Distribución	Desviación estándar
Constante	0
Erlang, $k = 1$	*media*
Erlang, $k = 2$	
Erlang, $k = 4$	1/2 *media*
Erlang, $k = 8$	
Erlang, $k = 16$	1/4 *media*
Erlang, cualquier k	

Notación

- Reconociendo la diversidad de los sistemas de colas, Kendall (1953) propuso un sistema de notación para sistemas de servidores paralelos que ha sido adoptado universalmente.
- Una versión resumida de esta convención está basada en el formato A/B/c/N/K. Estas letras representan las siguientes características del sistema:
- A = Distribución de tiempo entre arribos.
- B = Distribución del tiempo de servicio.
- Los siguientes son símbolos comunes para A y B:
    - M = exponencial o Markov (1)
    - D = constante o determinística

Notación

- E_k = Erlang de orden k
- PH = Tipo fase
- H = Hiperexponencial
- G = Arbitrario o general
- GI = General independiente

- c = número de servidores paralelos
- N = Capacidad del sistema
- K = Tamaño de la población.
- Nota(1): A causa de las suposiciones de distribución exponencial en los procesos de arribo, estos modelos son llamados MARKOVIANOS

Notación-Resumen

Table 1.1 Queueing Notation $A/B/X/Y/Z$

Characteristic	Symbol	Explanation
Interarrival-time distribution (A)	M	Exponential
	D	Deterministic
	E_k	Erlang type $k (k = 1, 2, \ldots)$
Service-time distribution (B)	H_k	Mixture of k exponentials
	PH	Phase type
	G	General
# of parallel servers (X)	$1, 2, \ldots, \infty$	
Max. system capacity (Y)	$1, 2, \ldots, \infty$	
Queue discipline (Z)	FCFS	First come, first served
	LCFS	Last come, first served
	RSS	Random selection for service
	PR	Priority
	GD	General discipline

Notación

- Por ejemplo: **M/M/1/∞/∞** significa ***un solo servidor***, capacidad de ***cola ilimitada*** y ***población infinita*** de arribos potenciales. Los tiempos entre arribos y los tiempos de servicio son **distribuídos exponencialmente**.
- Cuando **N** y **K** son *infinitos*, pueden ser *descartados* de la notación. M/M/1/∞/∞ es reducido a **M/M/1**.

Modelo :M/M/1

Asumimos que existen las siguientes condiciones:

- Los clientes son servidos con una política PEPS y cada arribo espera a ser servido sin importar la longitud de la línea o cola.
- Los arribos son independientes de arribos anteriores, pero el promedio de arribos, no cambia con el tiempo.
- Los arribos son descritos mediante la distribución de probabilidad de Poisson y proceden de una población muy grande o infinita.
- Los tiempos de servicio varían de cliente a cliente y son independientes entre sí, pero su rata promedio es conocida.

Fórmulas para modelo M/M/1

λ = Número promedio de arribos por período de tiempo

μ = Número promedio de gente o cosas servidos por período de tiempo

n = número de unidades en el sistema

L_S = Número promedio de unidades (clientes) en el sistema $L_S = \dfrac{\lambda}{\mu - \lambda}$

ρ = Factor de utilización del sistema = $\dfrac{\lambda}{\mu}$

W_S = Tiempo promedio que una unidad permanece en el sistema =
(tiempo de espera + tiempo de servicio)

$W_S = \dfrac{1}{\mu - \lambda}$

Fórmulas para modelo M/M/1

L_q = Número promedio de unidades en la cola = $\frac{\lambda^2}{\mu(\mu-\lambda)} = \rho * L_S$

W_q = Tiempo promedio que una unidad espera en la cola = $\frac{\lambda}{\mu(\mu-\lambda)} = \rho * W_S$

P_n = Probabilidad de que "n" clientes estén en el sistema =

$$P_n = \left(1-\frac{\lambda}{\mu}\right) * \left(\frac{\lambda}{\mu}\right)^n = (1-\rho) * \rho^n$$

P_o = Probabilidad de cero unidades en el sistema (la unidad de servicio está vacía) =

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

$P_{n\rangle k}$ = Probabilidad de que más de "k" unidades estén en el sistema =

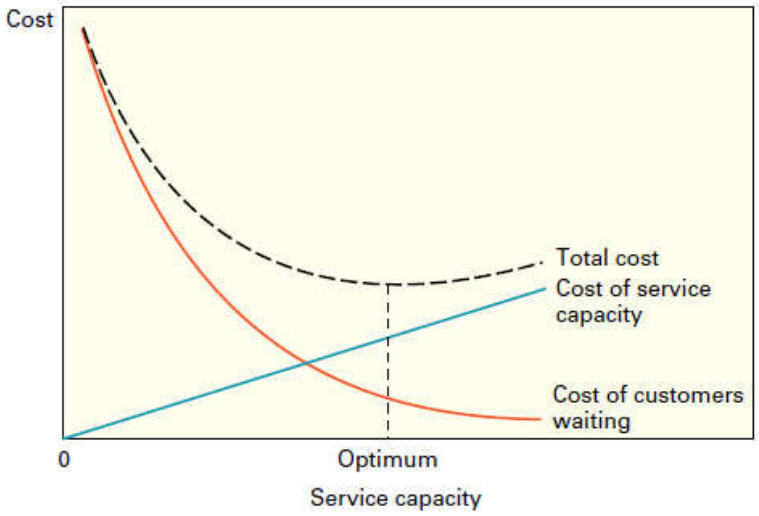
$$P_{n\rangle k} = \left(\frac{\lambda}{\mu}\right)^{k+1}$$

$$P(W_q > t) = \rho \cdot e^{-\mu(1-\rho)\cdot t}$$

$$P(W_s > t) = e^{-\mu(1-\rho)\cdot t}$$

Administración de las líneas de espera

CT= Costo de espera del cliente+ Costo de capacidad



Una de las decisiones habituales en el uso de este modelo puede serlo el de definir la cantidad de servidores necesarios. Por ejemplo

- La cantidad de ascensores en un edificio,
- La cantidad de escritorios para un equipo de trabajo.

La decisión se deberá basar en una relación entre dos costos básicos: el costo de proveer servidores adicionales versus el costo de demorar o no prestar el servicio.

Análisis de Costos

Los costos a los que nos acabamos de referir deben estar presentados por unidad de tiempo, a los efectos de realizar cálculos comparables. Si por ejemplo, el costo de un servidor consiste en el salario que debe pagarse a quien lo atiende, deberá anualizarse, para incluir aguinaldo, vacaciones, etc., y luego convertirlo en la misma unidad de tiempo que se use para determinar el tiempo de servicio o de espera.

Si se define:
C_d = Costo de demora por cliente por unidad de tiempo
C_s = Costo por unidad de tiempo para agregar otro servidor
L = Número promedio en el sistema

El costo total por unidad de tiempo para una estación con c servidores es:
$L\,C_d + c\,C_s$

A medida que c aumenta, la capacidad adicional incrementará la velocidad del servicio y *L* irá disminuyendo. Por consiguiente, una información útil que debe brindar el sistema es el número de servidores que minimice el costo total.