

Sistemas de colas: anatomía, desempeño y simulación

Juliana Quintana-Rojas

Departamento de Ingeniería Industrial

Universidad de Los Andes

Anatomía de un Sistema de colas

Def. Flujo de clientes que llegan a una instalación de servicio, esperan en una fila si los servidores están ocupados, reciben la atención requerida según una regla establecida y luego se retiran

01

**Población de
llamadas**

(Calling Population)

02

**Capacidad del
sistema**

(System Capacity)

03

**Proceso de
llegada**

(Arrival Process)

04

**Disciplina de
cola**

(Queue discipline)

05

**Mecanismo de
servicio**

(Service mechanism)

Población de Llamadas

(Calling Population)

Def. Representa el conjunto de clientes potenciales que pueden solicitar servicio.

Infinita

Ingreso de un cliente al sistema no tiene un impacto apreciable sobre la tasa de llegada de los clientes restantes

Finita

clientes potenciales es limitado. Cuando uno de ellos entra al sistema de colas, el tamaño de la población externa disminuye, reduciendo temporalmente la tasa de llegadas futuras

Ejemplo

Los clientes que entran a un supermercado o a un banco

Ejemplo

Grupo de K máquinas que fallan de vez en cuando y deben ser reparadas por un único operario

Capacidad del sistema

(System capacity)

Def. Establece el límite máximo de clientes que el sistema puede albergar simultáneamente. Incluye tanto a los clientes en fila como a aquellos bajo servicio

Infinita

No hay restricciones físicas sobre el tamaño de la fila; cualquier cliente que arribe puede esperar

Ejemplo

Preventa telefónica masiva o una fila larga en un parque de diversiones

Finita

El espacio físico o lógico de espera es limitado. Si un cliente llega y encuentra el sistema lleno, se retira inmediatamente del sistema y su venta o servicio se pierde

Ejemplo

Una bahía con espacio para máximo 10 autos

Proceso de Llegada

(Arrival process)

Def. Forma en la que los clientes aparecen en el sistema a lo largo del tiempo, comúnmente caracterizado por la distribución de probabilidad de los tiempos entre llegadas sucesivas

- Las llegadas pueden ser aleatorias, programadas o constantes
- El modelo más representativo en ingeniería es el Proceso de Poisson, donde los tiempos entre llegadas siguen una distribución exponencial con una media de $1/\lambda$, siendo λ la tasa promedio de arribos por unidad de tiempo.

Ejemplo - aleatorias

Llegada la fila de una atracción en un parque de diversiones

Ejemplo - programadas

Citas médicas programadas

Ejemplo - constantes

Llegada de tren cada 5 minutos

Disciplina de cola (Queue discipline)

Def. Es la regla que utiliza el servidor para decidir quién será el próximo cliente atendido cuando se libera un recurso

FIFO - First In First Out / FCFS

Primero en llegar, primero en ser servido

Fila de pago de un supermercado, los clientes se sirven en el orden que llegan

LIFO - Last In First Out

El último cliente en llegar es el primero en recibir atención

Orden de lavado de una torre de platos, el primero en ser lavado es el que más reciente se pone

Prioridad (Priority)

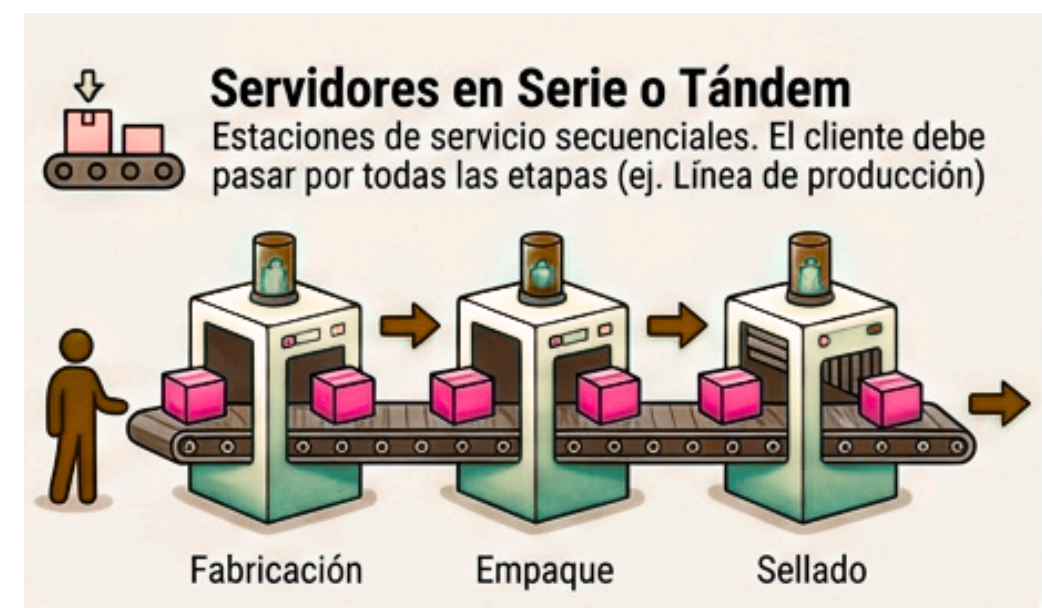
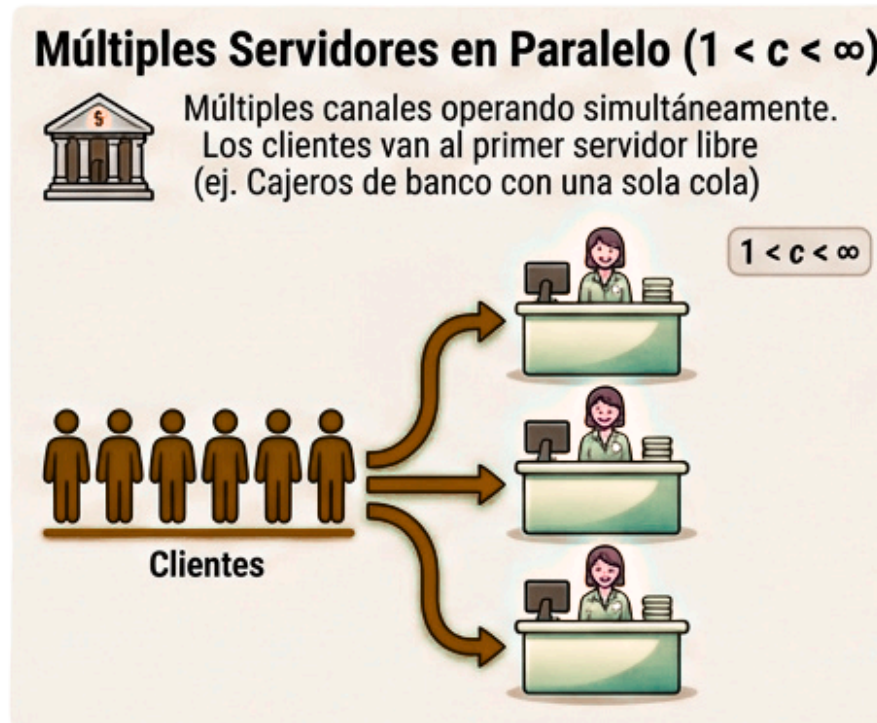
Se atiende según la urgencia, tipo de cliente (ej. preferencial) o requerimiento de servicio

Cuarto de emergencias de un hospital, el paciente más crítico es el primero en ser atendido

Mecanismo de servicio

(Service mechanism)

Def. Número de servidores disponibles en paralelo (c o s) y la duración del tiempo de servicio de cada cliente (S_n)



Medidas de desempeño en colas

Tiempo promedio en el sistema (W)

$$\hat{W} = \frac{1}{N} \sum_{i=1}^N W_i$$

Número promedio entidades en el sistema (L)

$$\hat{L} = \frac{1}{T} \int_0^T L(t) dt$$

Tiempo promedio en cola (W_q)

$$\hat{W}_q = \frac{1}{N} \sum_{i=1}^N W_{qi}$$

Número promedio entidades en cola (L_q)

$$\hat{L}_q = \frac{1}{T} \int_0^T L_q(t) dt$$

Tiempo promedio en servicio (W_s)

$$\hat{W}_s = \frac{1}{N} \sum_{i=1}^N W_{si}$$

Número promedio entidades en servicio (L_s)

$$\hat{L}_s = \frac{1}{T} \int_0^T L_s(t) dt$$

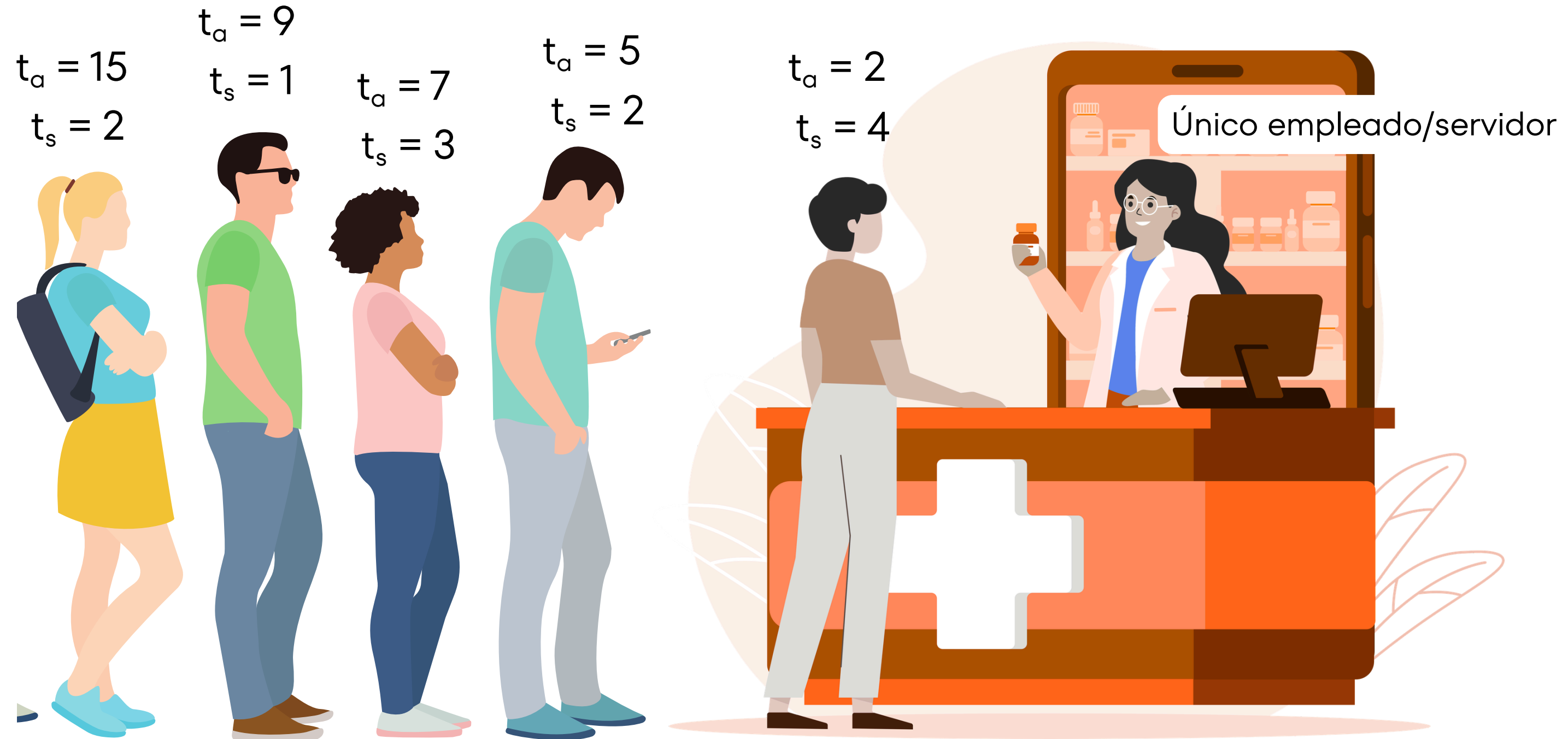
Ley de little (Eq. Conservación)

Def. Establecida por **John D. C. Little en 1961**, es el resultado más general y útil de la teoría de colas. Postula que el número promedio de clientes en un sistema estable es igual a la tasa promedio de arribos multiplicada por el tiempo promedio de permanencia en el sistema.

#Promedio Clientes
 en el sistema = Tasa promedio
 arribos en sistema $\times$ Tiempo promedio
 en sistema.

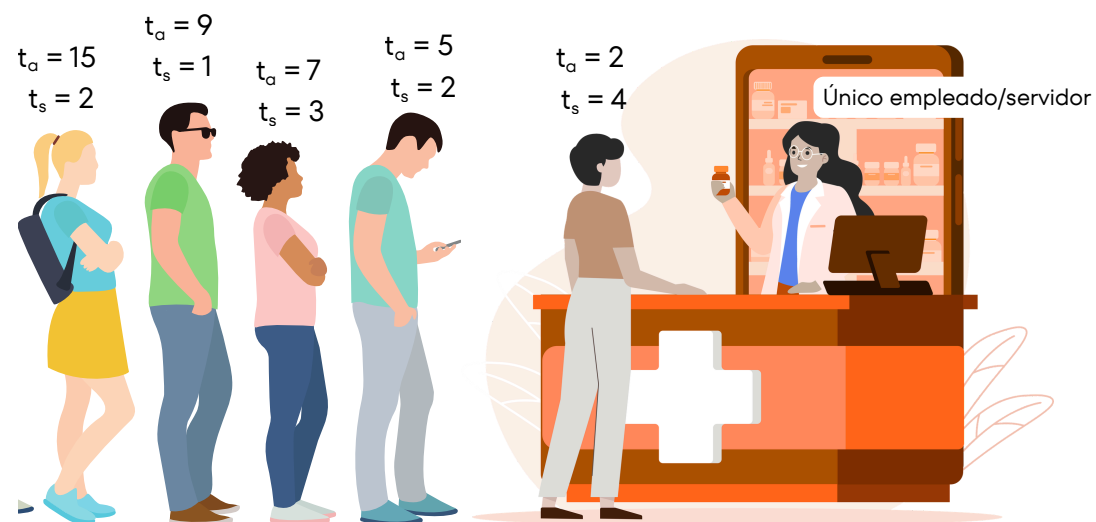
$\hat{L} = \hat{\lambda} \hat{W}$ sistema	$\hat{L}_q = \hat{\lambda} \hat{W}_q$ En cola	$\hat{L}_s = \hat{\lambda} \hat{W}_s$ en servicio
--	--	--

Ejercicio



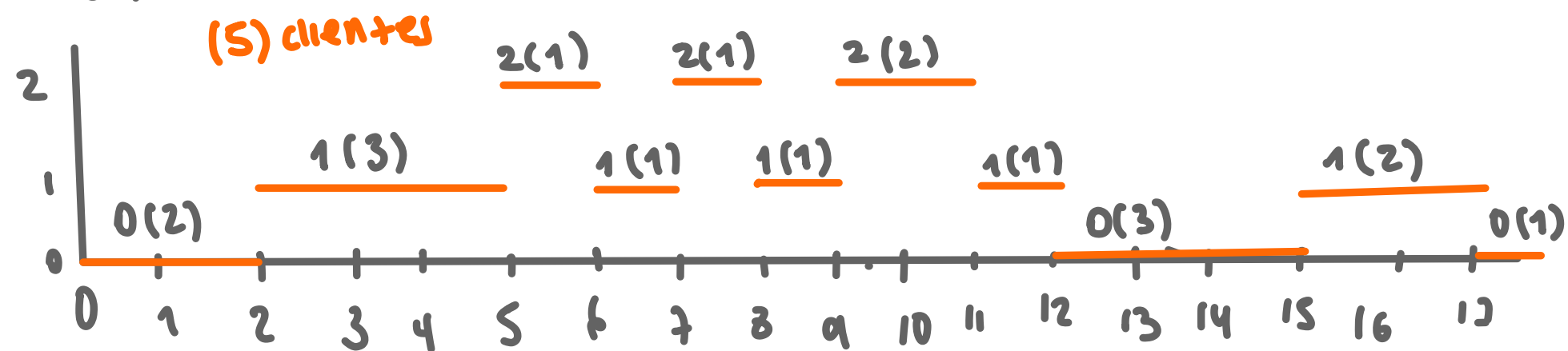
Ejercicio

Max(Finc., Llegada)

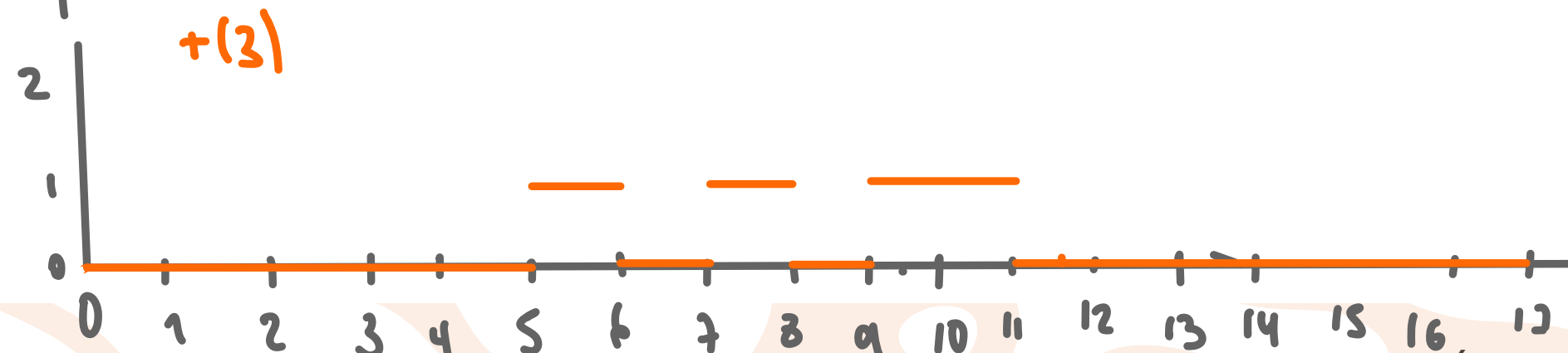


cliente	Llega a la	Inicio	servicio	Fin
C1	2	2	4	<u>6</u>
C2	<u>5</u>	6	2	8
C3	7	8	3	11
C4	9	11	1	12
C5	15	15	2	17

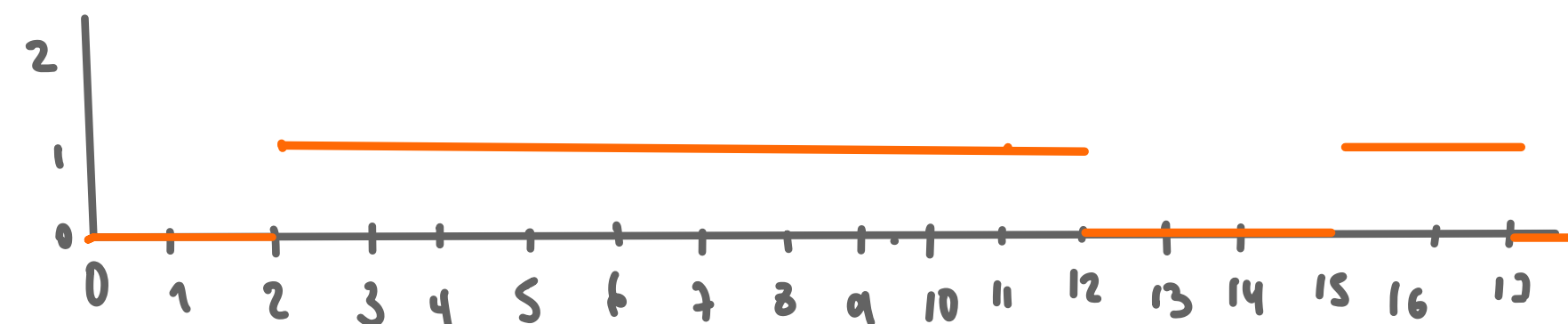
$L(t) \rightarrow$ # clientes en la farmacia



$L_q(t) \rightarrow$ # clientes en cola



$L_s(t) \rightarrow$ # clientes servicio



Ejercicio

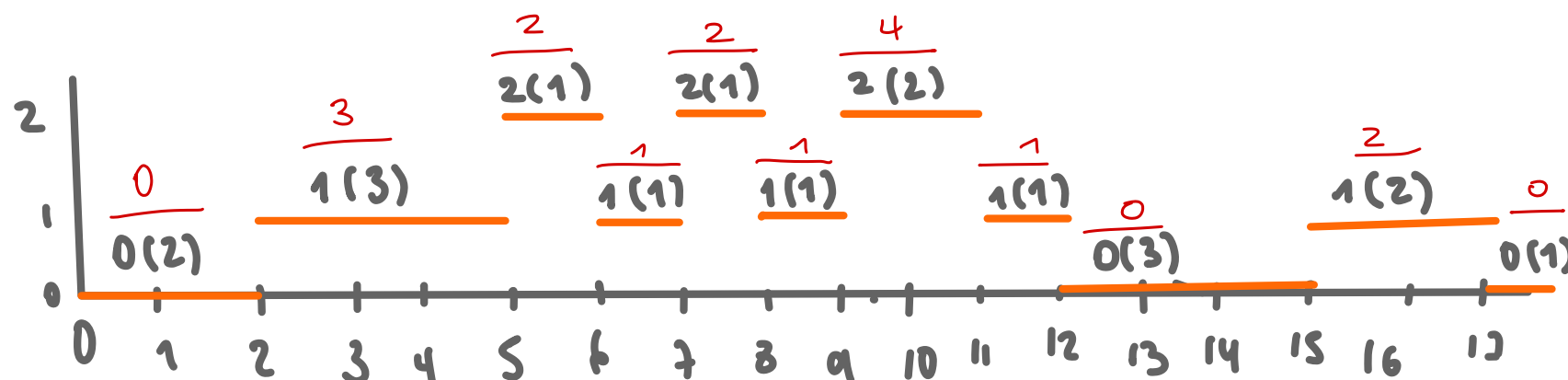
Calculo métrica

$T = 17$ (Tiempo maximo)

$$\hat{L} = \frac{1}{T} \int_0^T L(t) dt$$

$$\hat{L} = \frac{1}{17} \int_0^{17} L(t) dt$$

Área bajo la curva



$$\hat{L} = \frac{0 + 3 + 2 + 1 + 2 + 1 + 4 + 1 + 0 + 2 + 0}{17} = 16/17 = 0.94 \text{ clientes/min}$$

$$\hat{W} = \frac{1}{N} \sum_{i=1}^N W_i \rightarrow W_i = T_{\text{salida}} - T_{\text{ llegada}}$$

Cliente	Llega da	Inicio	Servicio	Fin	W _i
C ₁	2	2	4	6	4
C ₂	<u>5</u>	6	2	8	3
C ₃	7	8	3	11	4
C ₄	9	11	1	12	3
C ₅	15	15	2	17	2

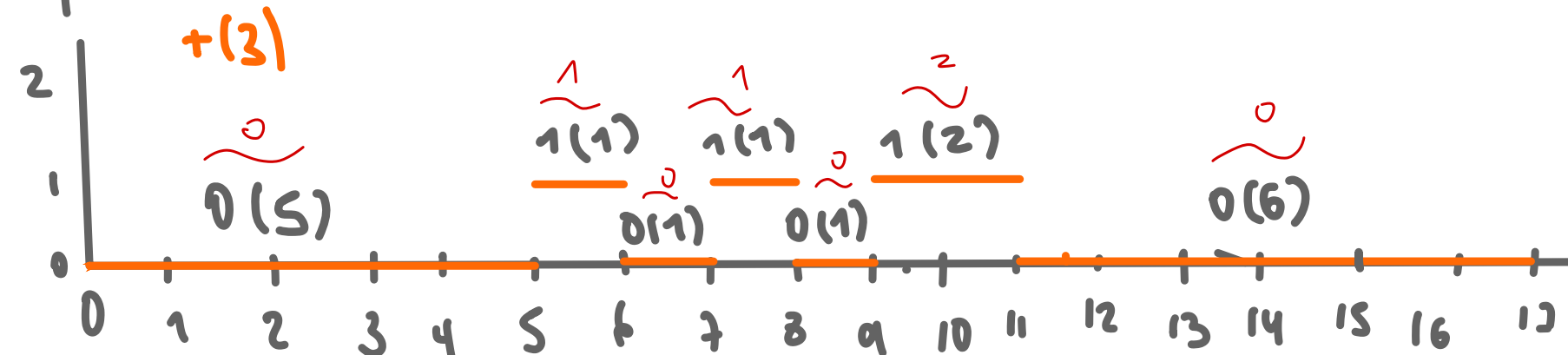
$$\hat{W} = \frac{1}{5} \cdot (4 + 3 + 4 + 3 + 2) = 16/5 = 3.2 \text{ min}$$

Ejercicio

Calculo métrica

$T = 17$ (Tiempo maximo)

$L_q(t) \rightarrow \#$ clientes en cola



$$\hat{L}_q = \frac{1}{T} \int_0^T L_q(t) dt$$

$$\hat{L}_q = \frac{1}{17} \int_0^{17} L_q(t) dt$$

Área bajo la curva

$$\hat{L}_q = \frac{0+1+0+1+0+2+0}{17} = 4/17 = 0.24 \text{ clientes/min}$$

$$\hat{W}_q = \frac{1}{N} \sum_{i=1}^N W_{qi} \rightarrow W_{qi} = \text{inicio servicio} - \text{llegada}$$

$$\hat{W}_q = \frac{1}{5} (0+1+1+2+0) = 4/5 = 0.8 \text{ min}$$

cliente	Llega aq	Inicio	servicio	Fin	W_i	W_{qi}
C ₁	2	2	4	<u>6</u>	4	0
C ₂	<u>5</u>	6	2	8	3	1
C ₃	7	8	3	11	4	1
C ₄	9	11	1	12	3	2
C ₅	15	15	2	17	2	0

Ejercicio

Calculo métrica

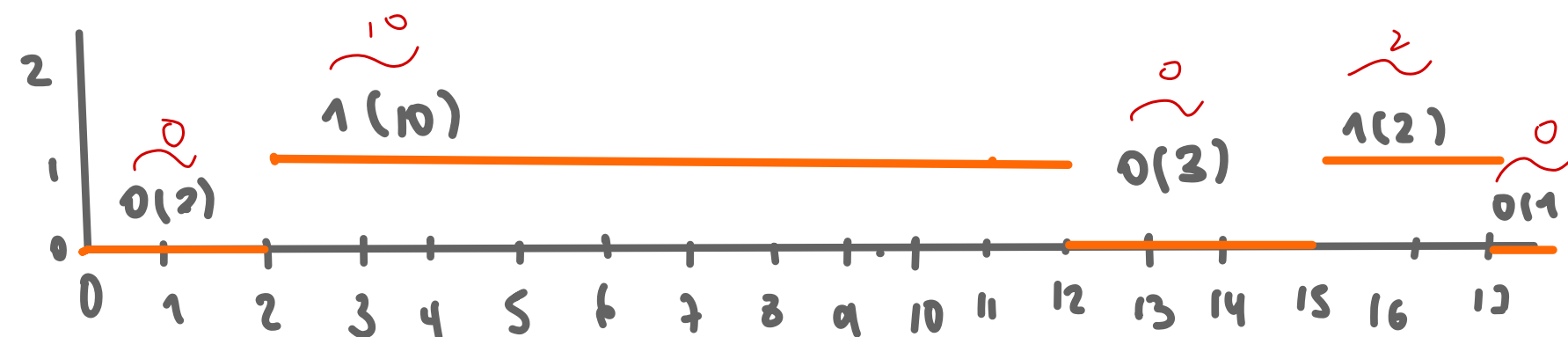
$T = 17$ (Tiempo maximo)

$$\hat{L}_s = \frac{1}{T} \int_0^T L_s(t) dt$$

$$\hat{L}_s = \frac{1}{17} \int_0^{17} L_s(t) dt$$

Área bajo la curva

$L_s(t) \rightarrow \# \text{ clientes servicio}$



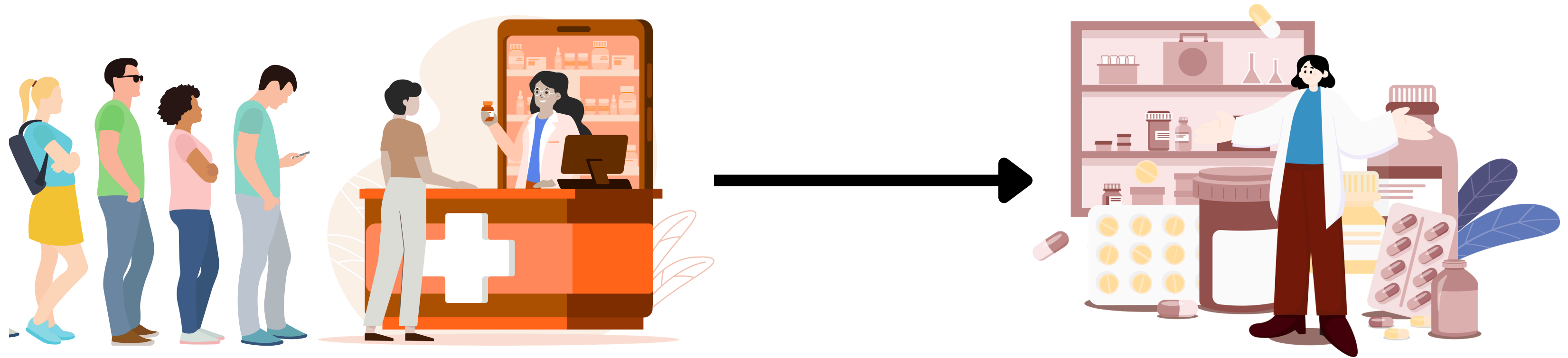
$$\hat{L}_s = \frac{0 + 10 + 0 + 2 + 0}{17} = 12/17 = 0.706 \text{ clientes servicio/min}$$

$$\hat{W}_s = \frac{1}{N} \sum_{i=1}^N W_i \rightarrow W_s = \text{salida} - \text{Inicio servicio}$$

$$\hat{W}_s = \frac{1}{5} (4 + 2 + 3 + 1 + 2) = 12/5 = 2.4 \text{ min}$$

cliente	Llega aq	Inicio	servicio	Fin	W_i	W_{qi}	W_s
C1	2	2	4	<u>6</u>	4	0	4
C2	<u>5</u>	6	2	8	3	1	2
C3	7	8	3	11	4	1	3
C4	9	11	1	12	3	2	1
C5	15	15	2	17	2	0	2

Ejercicio

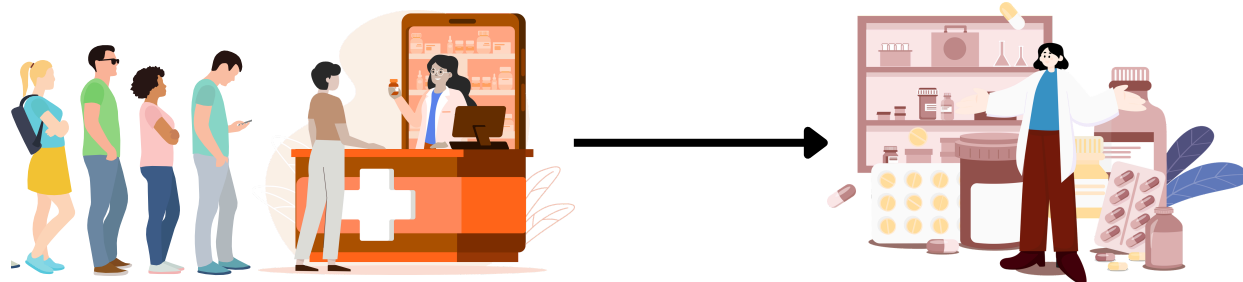


La misma farmacia, pero ahora con dos estaciones: mostrador principal (cobro) y una estación de empaque especial para medicamentos refrigerados.

- **Tasa total de arribos al sistema:** $\lambda = 12$ clientes/hora
- **Tiempo promedio en la estación principal:** 3 min
- El 40% de los clientes pasa además por la estación de empaque
- En promedio hay 0.9 clientes en todo el sistema en estado estable

- (1) • #clientes estación principal
- (2) • #clientes estación empaque.
- (3) • W_s , estación empaque
- (4) • Tiempo total prom sistema

Ejercicio



$$\lambda = 12 \text{ clientes / hora} \rightarrow \frac{12 \text{ c}}{\text{h}} \cdot \frac{1 \text{ h}}{60 \text{ m}} = 0.2 \text{ c/h.}$$

$$T_{\text{promedio}} S_1 = 3 \text{ min}$$

40% estación empaque

0.9 clientes en todo el sistema estado estable

ley Little

$$L = \lambda W \Rightarrow 0.9 \text{ clientes} = 0.2 \frac{\text{clientes}}{\text{hora}} \cdot W$$

$$W = \frac{0.9}{0.2} = 4.5 \text{ minutos} \rightarrow \text{Tiempo total sistema}$$

(2)

$$L_{S2} \rightarrow L = L_{S1} + L_{S2} \rightarrow L_{S2} = L - L_{S1} = 0.9 - 0.6 = 0.3 \text{ clientes} \rightarrow \text{estación empaque}$$

$$\lambda_2 \rightarrow \text{tasa arribos estación 2 : } \lambda_2 = 0.4 \cdot \lambda_1 = 0.4 (0.2) = 0.08$$

(3)

$$W_{S2} = L_{S2} / \lambda_2$$

$$W_{S2} = 0.3 / 0.08 = 3.75$$

↳ tiempo estación empaque

Notación de Kendall

Def. Propuesta por David G. Kendall en 1953, es una nomenclatura compacta de formato **A/B/C/D/E/F** utilizada para describir rápidamente las características estructurales de un sistema de colas en paralelo

- A: proceso de arribos
- B: proceso de servicio
- C: número de servidores
- D: disciplina de cola
- E: capacidad en el sistema
- F: tamaño de la población de usuarios

A/B/C/D/E/F

Notación de Kendall

Proceso Arribos

Número de servidores

A/B/C/D/E/F

Proceso servicio

1 : uno

K : servidores

∞ servidores

G : general

Gi : general con arribos independientes

M Markoviano. tiempo entre arribos exponenciales
Procesos poisson

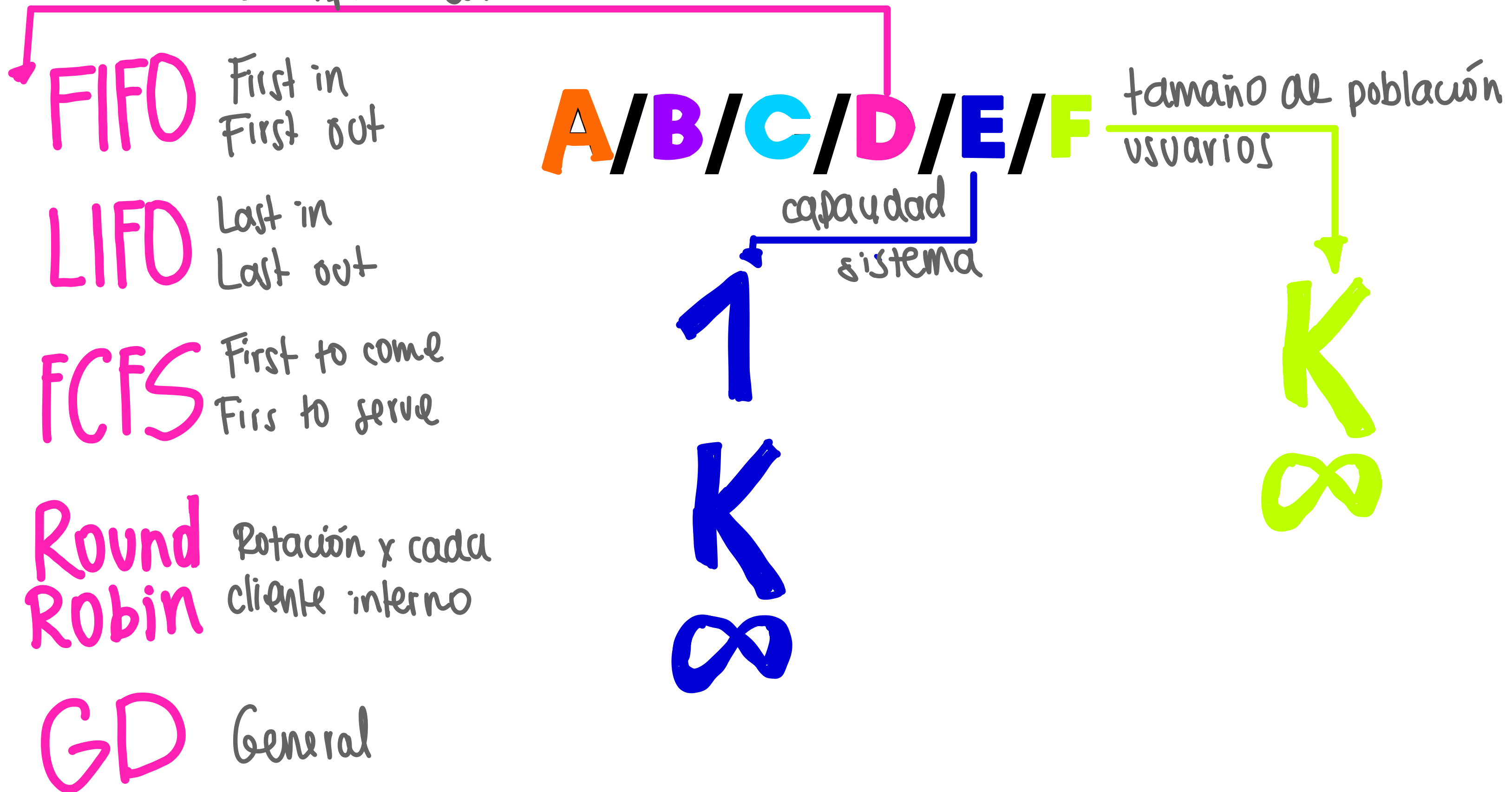
Gamma / Log Normal / etc.

G : general

M Markoviano. Tiempo servicio exponenciales

Gamma / Log Normal / etc.

Notación de Kendall



Ej. Notación de Kendall

Un cajero automático único; llegadas Poisson, retiro exponencial; la fila física solo tiene espacio para 6 personas en total (incluida la que se atiende); población de usuarios ilimitada; FIFO.

$$M/M/1/FCFS/6/\infty$$

$$M/M/1/FIFO/6/\infty$$

Un restaurante buffet de autoservicio: nunca se forma fila porque siempre hay espacio para atender a todos a la vez; el tiempo que cada quien se sirve es variable, sin distribución conocida; llegadas Poisson.

$$M/G/\infty/GD/\infty/\infty$$

Una fábrica con 5 máquinas idénticas que fallan (tiempo hasta la falla exponencial) y un solo técnico que las repara (reparación exponencial); no hay clientes "externos", el universo entero son esas 5 máquinas.

$$M/M/1/FCFS/5/5$$

$$M/M/1/FIFO/5/5$$

Una clínica veterinaria con 3 doctores idénticos; llegadas Poisson, consultas exponenciales; sala de espera sin límite; FIFO.

$$M/M/3/FIFO/\infty/\infty$$

Medidas de desempeño largo plazo

Tiempo promedio en el sistema (W)

$$W = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N W_i$$

Tiempo promedio en cola (W_q)

$$W_q = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N W_{qi}$$

Tiempo promedio en servicio (W_s)

$$W_s = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N W_{si}$$

Número promedio entidades en el sistema (L)

$$L = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T L(t) dt$$

Número promedio entidades en cola (L_q)

$$L_q = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T L_q(t) dt$$

Número promedio entidades en servicio (L_s)

$$L_s = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T L_s(t) dt$$

Ley little sistema (L)

$$L = \lambda W$$

Ley little colas (L_q)

$$L_q = \lambda W_q$$

Ley little servicio (L_s)

$$L_s = \lambda W_s$$

Ej. Mecánico

Un mecánico atiende máquinas de producción que fallan de acuerdo a un proceso de Poisson con una tasa promedio de 1.5 máquinas por hora. Los tiempos de reparación del mecánico siguen una distribución exponencial con un tiempo medio de reparación de 30 minutos.

¿Cuál es la utilización?

$$\rho = \frac{\lambda}{\mu} = \frac{1.5}{2} = 0.75$$

→ 75% servidor es utilizado | $\rho < 1$

$\mu \rightarrow$ tasa servicio

$$15 = 30 \text{ min} = 0.5 \text{ hora} \rightarrow \mu = 1/0.5$$

$$\mu = 2 \text{ máquinas/minutos}$$

El estado es en estado estable entonces existen métricas al largo plazo

promedio máquinas rotas

$$L = \frac{\rho}{1 - \rho} = \frac{0.75}{1 - 0.75} = 3 \text{ máquinas}$$

Ej. Mecánico

Un mecánico atiende máquinas de producción que fallan de acuerdo a un proceso de Poisson con una tasa promedio de 1.5 máquinas por hora. Los tiempos de reparación del mecánico siguen una distribución exponencial con un tiempo medio de reparación de 30 minutos.

* Tiempo promedio máquina fuera servicio

$$L = \lambda \cdot W \Rightarrow W = L / \lambda = 3 / 1.5 = 2 \text{ horas}$$

* Tiempo promedio máquina espera en fila antes de ser reparada (tiempo cola)

$$W_q = W - W_s = 2.0 - 0.5 = 1.5 \text{ horas}$$

* # promedio máquinas fila esperando reparación (L_q)

$$L_q = \lambda \cdot W_q = 1.5 \text{ ma/h} \cdot 1.5 \text{ horas} = 2.25 \text{ máquina}$$