

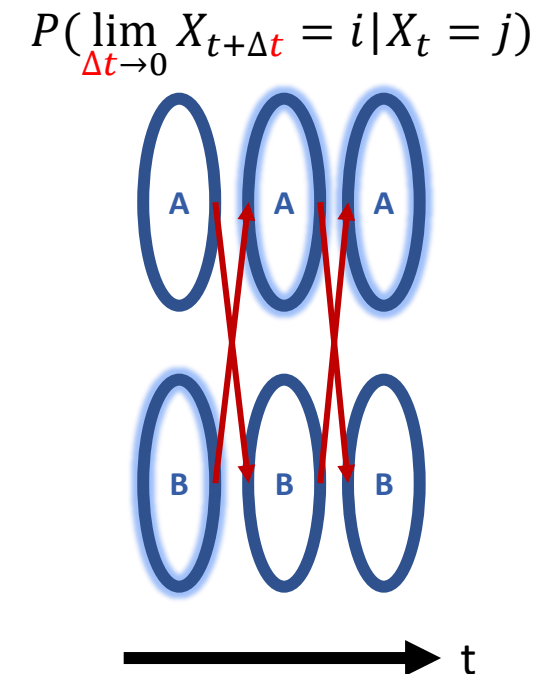
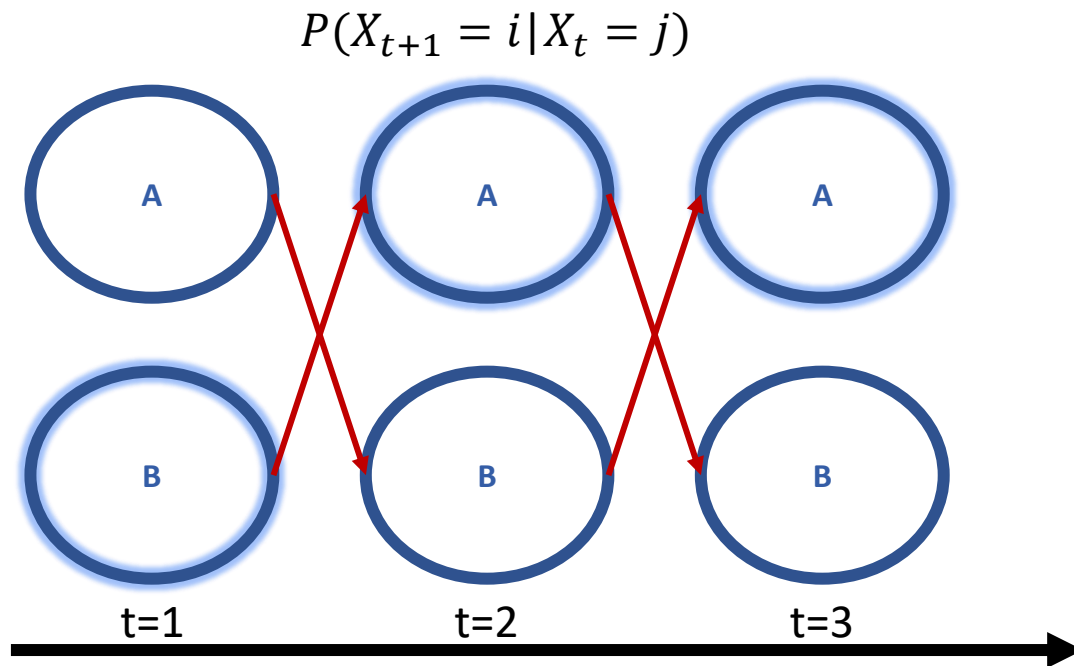


Cadenas de Markov Infinitas y Filas de Espera

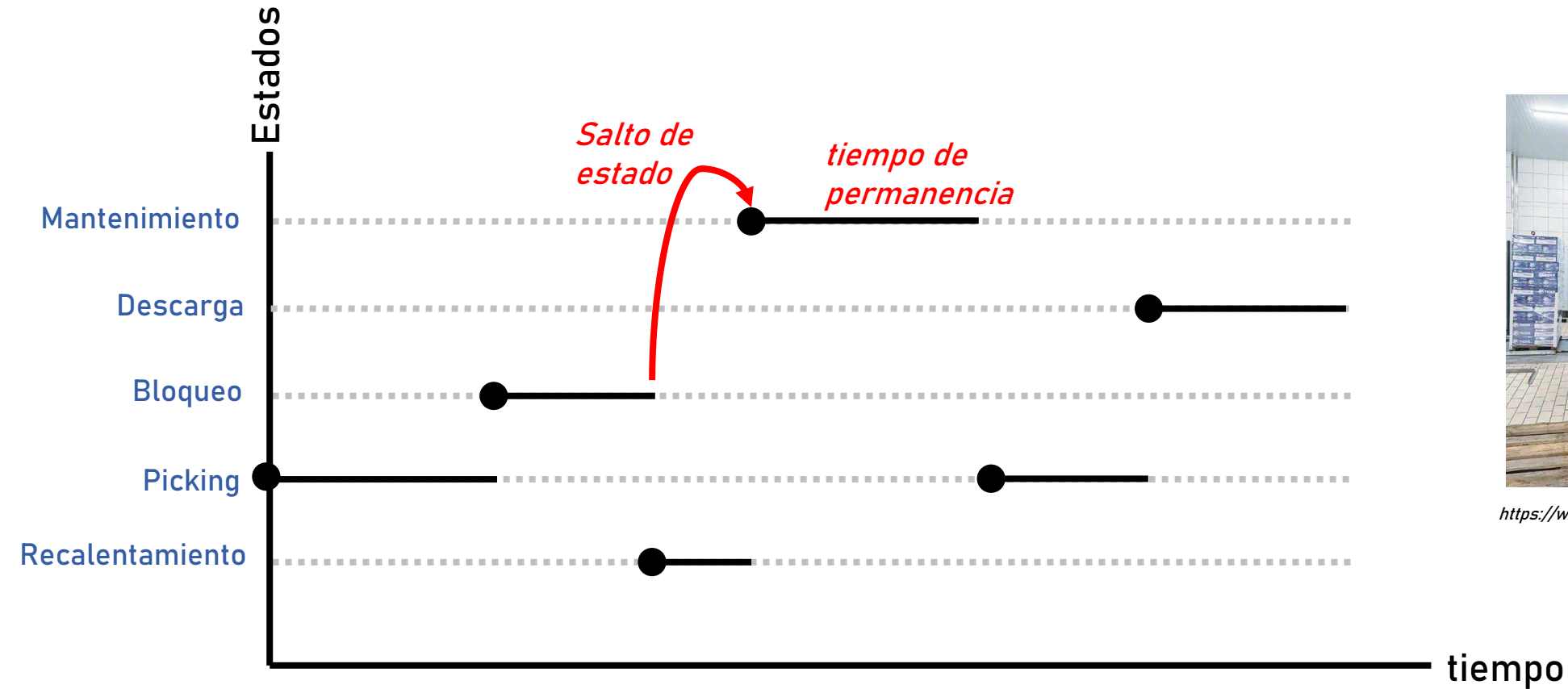
Rodrigo Maranzana

Repaso: Cadenas de Markov de tiempo continuo

Si intentamos achicar el paso del parámetro t , en la probabilidad de transición:

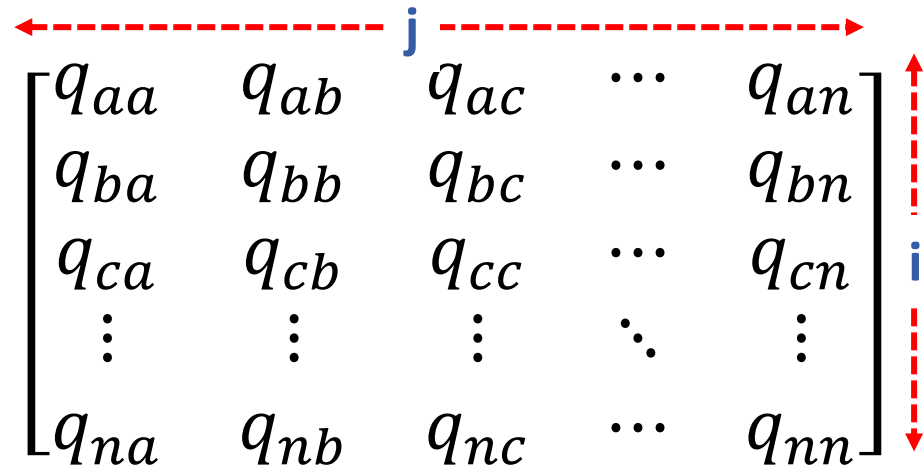


Repaso: Cadenas de Markov de tiempo continuo



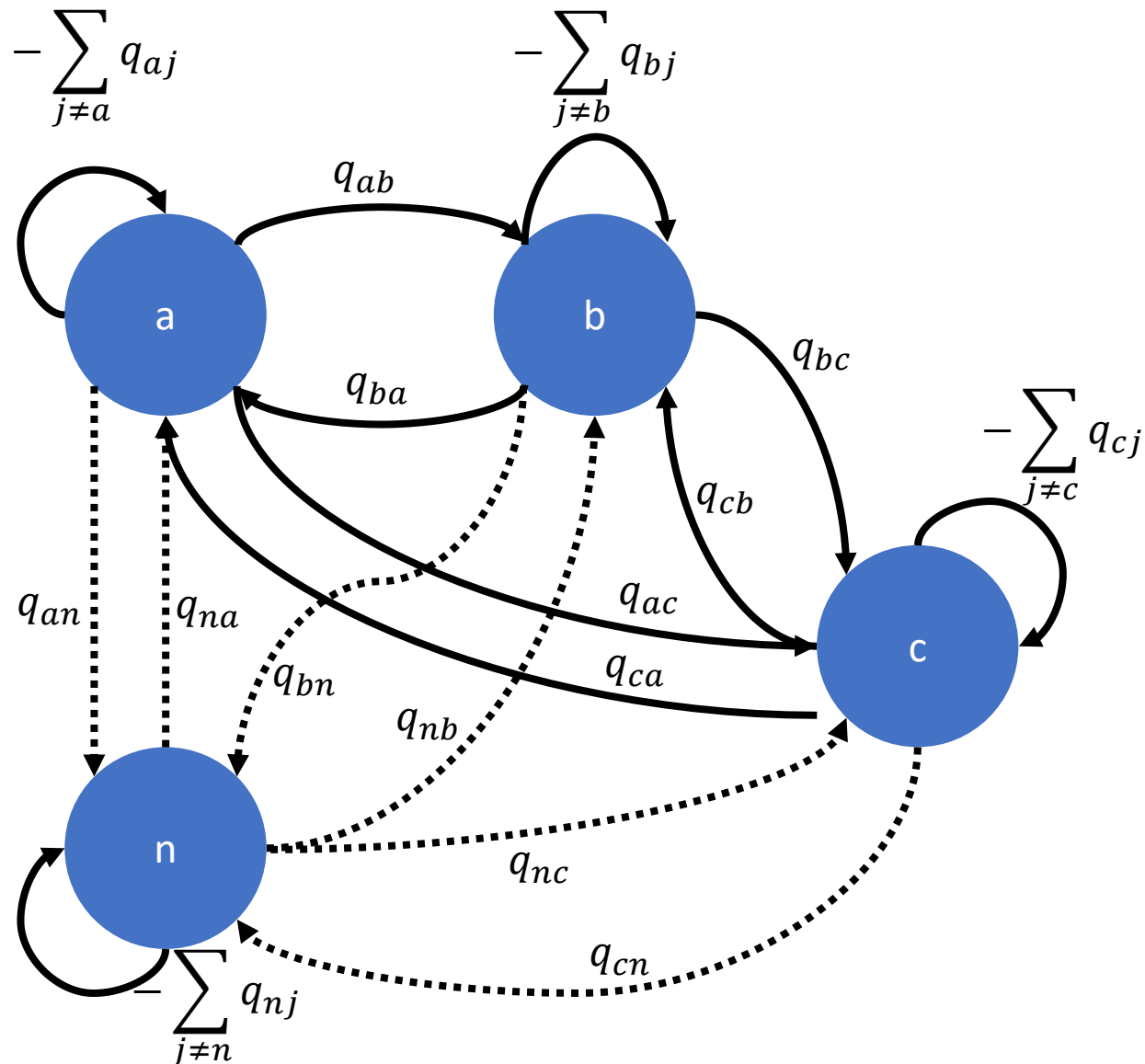
<https://www.mecalux.com.ar/blog/robot-de-picking>

Repaso: matriz generadora infinitesimal

$$Q = \begin{bmatrix} q_{aa} & q_{ab} & q_{ac} & \cdots & q_{an} \\ q_{ba} & q_{bb} & q_{bc} & \cdots & q_{bn} \\ q_{ca} & q_{cb} & q_{cc} & \cdots & q_{cn} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q_{na} & q_{nb} & q_{nc} & \cdots & q_{nn} \end{bmatrix}$$


Las tasas q_{ij} de cada componentes son escalares, representan la tasa de transición o de saltos entre estados.

Repaso: grafo y matriz generadora



$$Q = \begin{bmatrix} -\sum_{j \neq a} q_{aj} & q_{ab} & q_{ac} & \cdots & q_{an} \\ q_{ba} & -\sum_{j \neq b} q_{bj} & q_{bc} & \cdots & q_{bn} \\ q_{ca} & q_{cb} & -\sum_{j \neq c} q_{cj} & \cdots & q_{cn} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q_{na} & q_{nb} & q_{nc} & \cdots & -\sum_{j \neq n} q_{nj} \end{bmatrix}$$

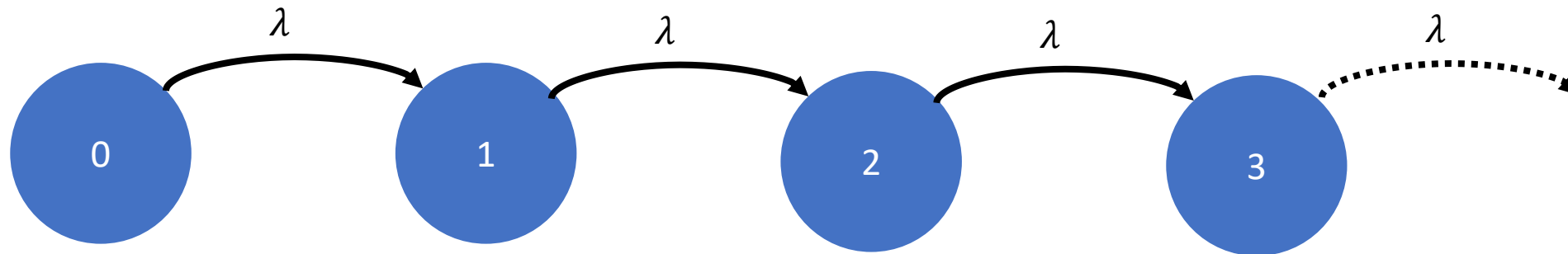
Repaso: Cadenas de Markov de tiempo continuo

Clasificación por estados:

- Estado finito
- Estado infinito

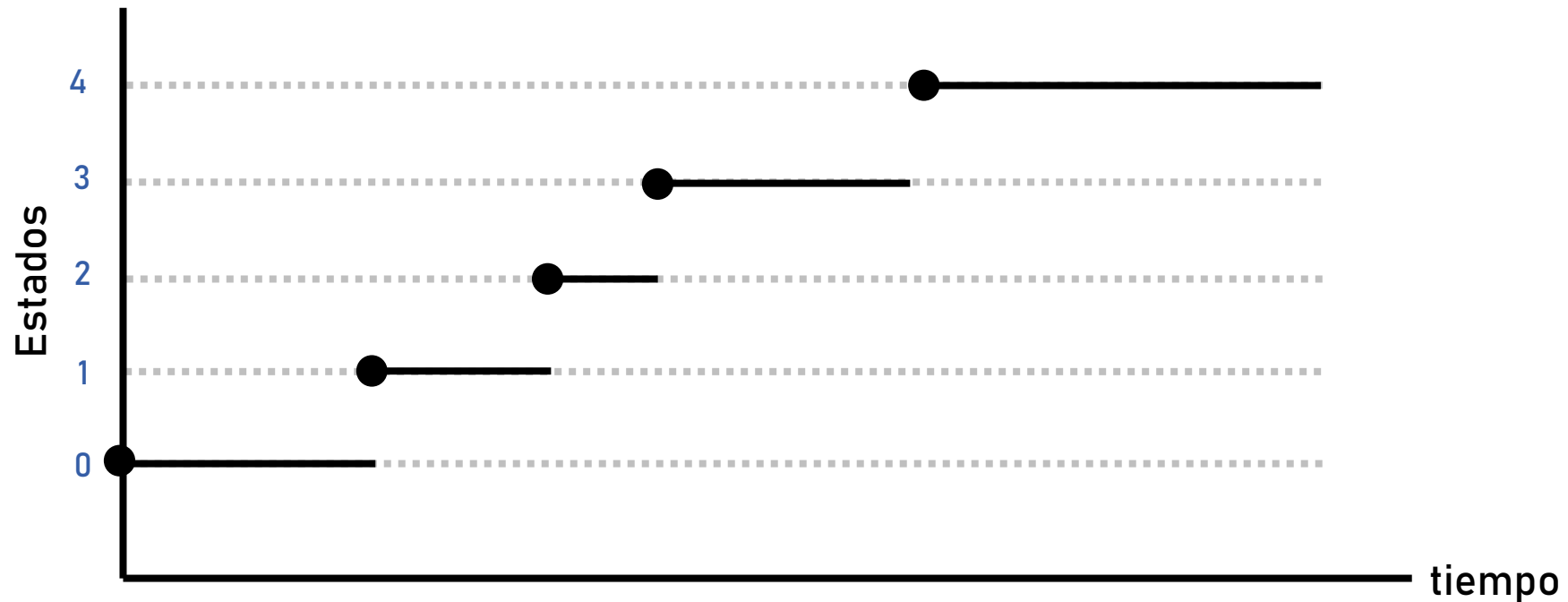
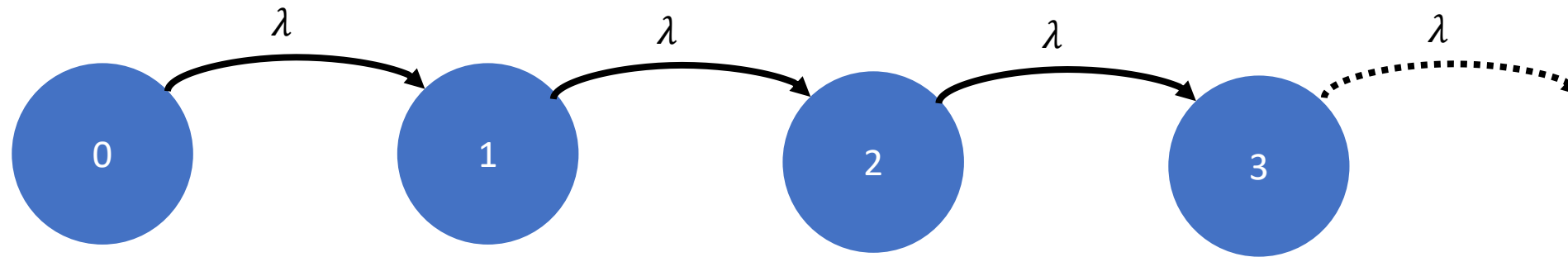
Proceso de Poisson (proceso de nacimiento)

- Es una cadena de markov de estado infinito.
- Las tasas de transición (λ) caracterizan la **distribución de Poisson** subyacente.
- Proceso de conteo, en donde el **tiempo entre eventos** sigue una distribución exponencial.
- Se genera un **proceso estocástico monótono creciente**.
- Los estados se definen por la cantidad de eventos generados.



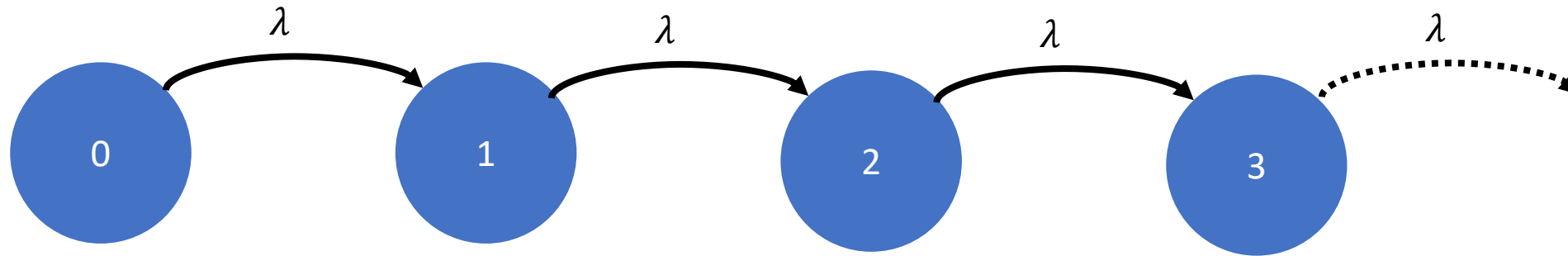
Proceso de Poisson (proceso de nacimiento)

Proceso de conteo:



Proceso de Poisson (proceso de nacimiento)

Matriz generadora:

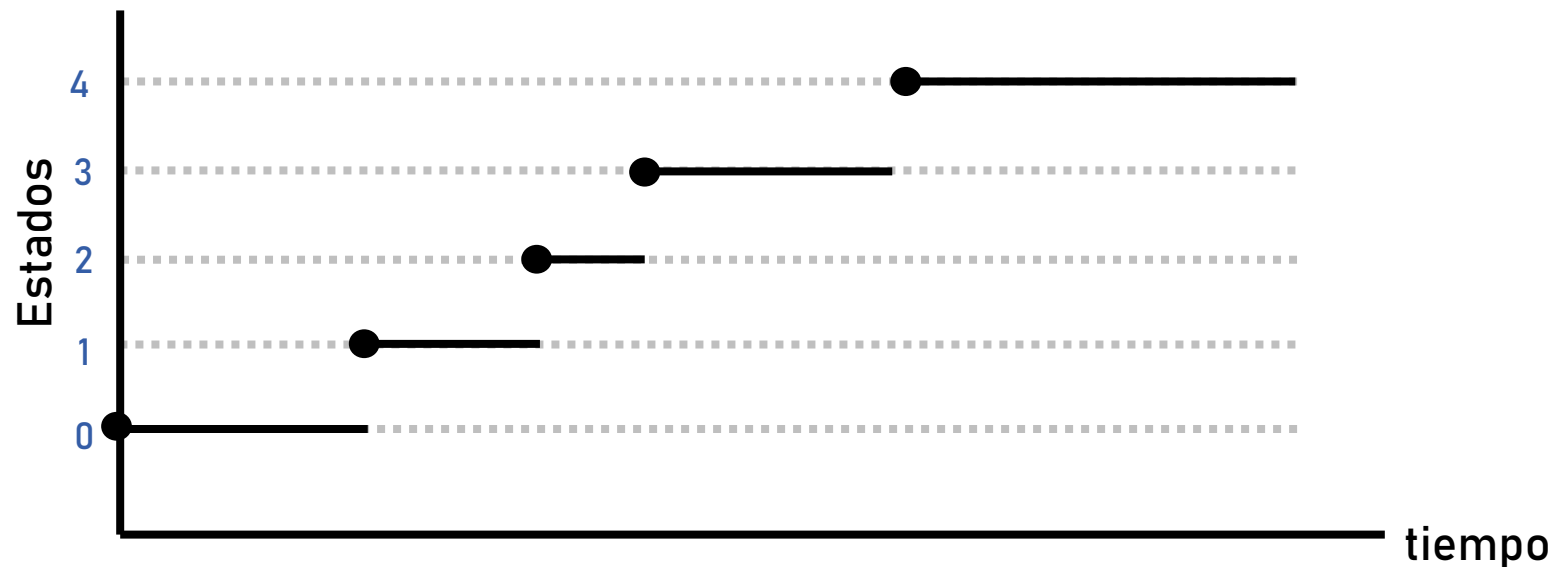


$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & \cdots \\ 0 & -\lambda & \lambda & \cdots \\ 0 & 0 & -\lambda & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Proceso de Poisson (proceso de nacimiento)

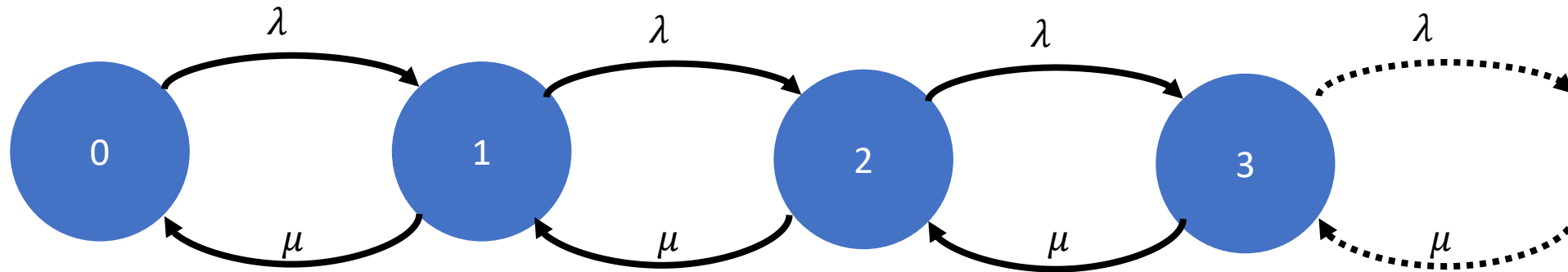
Ejemplos de aplicaciones:

- Conteo de **fallas** en un proceso/máquina.
- Conteo de **llamados** a un servicio.
- Finanzas: **riesgo de default**. Pricing de credit default swaps. La tasa de eventos se conoce como “hazard rate” y se estima la probabilidad de sobrevivir o de defaultear.



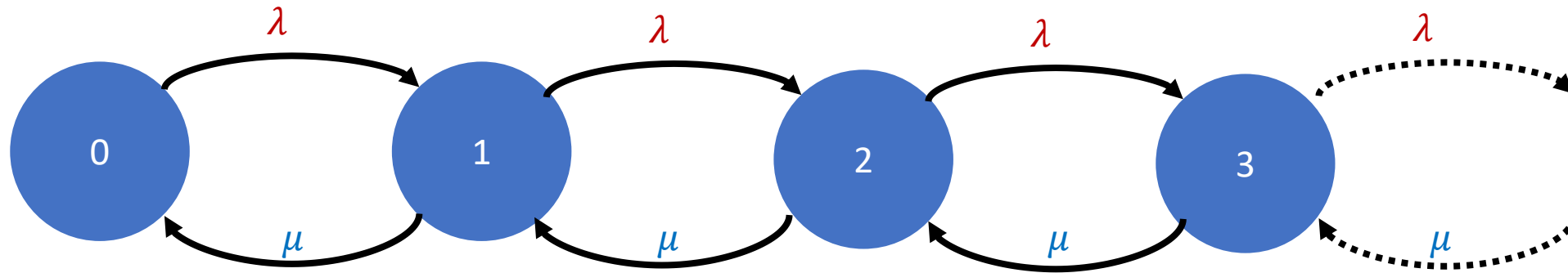
Proceso de Nacimiento y Muerte

- Es una cadena de markov de estado infinito.
- Los estados se definen por la cantidad de eventos generados.
 - Los nacimientos incrementan el conteo.
 - Las muertes restan al conteo.



Proceso de Nacimiento y Muerte

Matriz generadora:



$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \cdots \\ \mu & -\lambda - \mu & \lambda & 0 & \cdots \\ 0 & \mu & -\lambda - \mu & \lambda & \cdots \\ 0 & 0 & \mu & -\lambda - \mu & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

$$q_{ij} = \begin{cases} \lambda & \text{Si } j = i + 1 \\ \mu & \text{Si } j = i - 1 \\ 0 & \text{Si } |i - j| > 1 \\ -\sum_{j \neq i} q_{ij} & \text{Si } i = j \end{cases}$$

Estado estacionario

$$[p_0 \quad p_1 \quad p_2 \quad p_3 \quad \dots] \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \dots & 1 \\ \mu & -\lambda - \mu & \lambda & 0 & \dots & 1 \\ 0 & \mu & -\lambda - \mu & \lambda & \dots & 1 \\ 0 & 0 & \mu & -\lambda - \mu & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \dots \\ 1 \end{bmatrix}$$

Armando sistema de ecuaciones:

$$\begin{aligned} -p_0\lambda + p_1\mu &= 0 \\ p_0\lambda - p_1(\lambda + \mu) + p_2\mu &= 0 \\ p_1\lambda - p_2(\lambda + \mu) + p_3\mu &= 0 \\ &\dots \\ p_{n-2}\lambda - p_{n-1}(\lambda + \mu) + p_n\mu &= 0 \\ &\dots \\ \sum_n p_n &= 1 \end{aligned}$$

Estado estacionario

Si despejamos cada ecuación:

$$-p_0\lambda + p_1\mu = 0$$

$$p_0\lambda - p_1(\lambda + \mu) + p_2\mu = 0$$

$$p_1\lambda - p_2(\lambda + \mu) + p_3\mu = 0$$

...

$$p_{n-2}\lambda - p_{n-1}(\lambda + \mu) + p_n\mu = 0$$

...

$$\sum_n p_n = 1$$



$$p_1 = p_0 \frac{\lambda}{\mu}$$

$$p_2 = p_1 \frac{\lambda}{\mu}$$

...

$$p_n = p_{n-1} \frac{\lambda}{\mu}$$



$$p_2 = \left(p_0 \frac{\lambda}{\mu} \right) \frac{\lambda}{\mu}$$

...

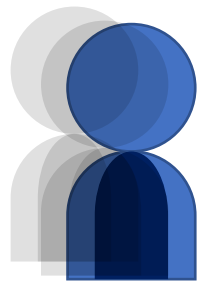
$$p_n = p_0 \left(\frac{\lambda}{\mu} \right)^n$$

Estado estacionario

Esta expresión nos permite calcular la probabilidad del sistema, de estar en un estado determinado “n”, conociendo la probabilidad de estar vacío.

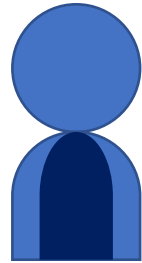
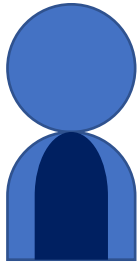
$$p_n = p_0 \left(\frac{\lambda}{\mu} \right)^n$$

Introducción a Filas de Espera



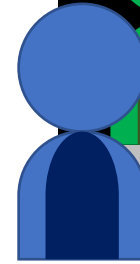
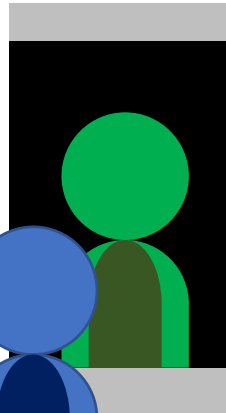
Llegada

Agente

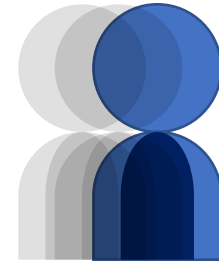


Fila de espera

Servidor



En servicio

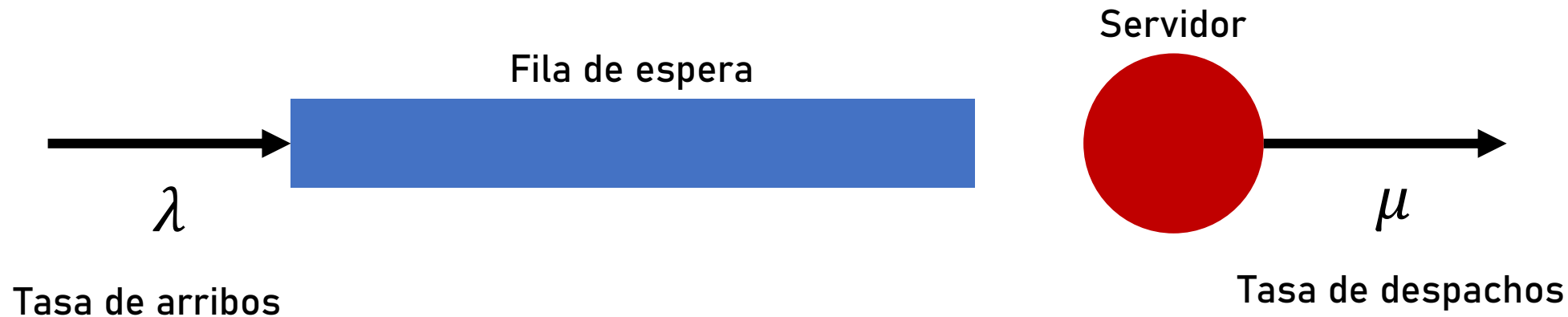


Agente servido

Representación de filas de espera

Arribo de agentes

Despacho de agentes



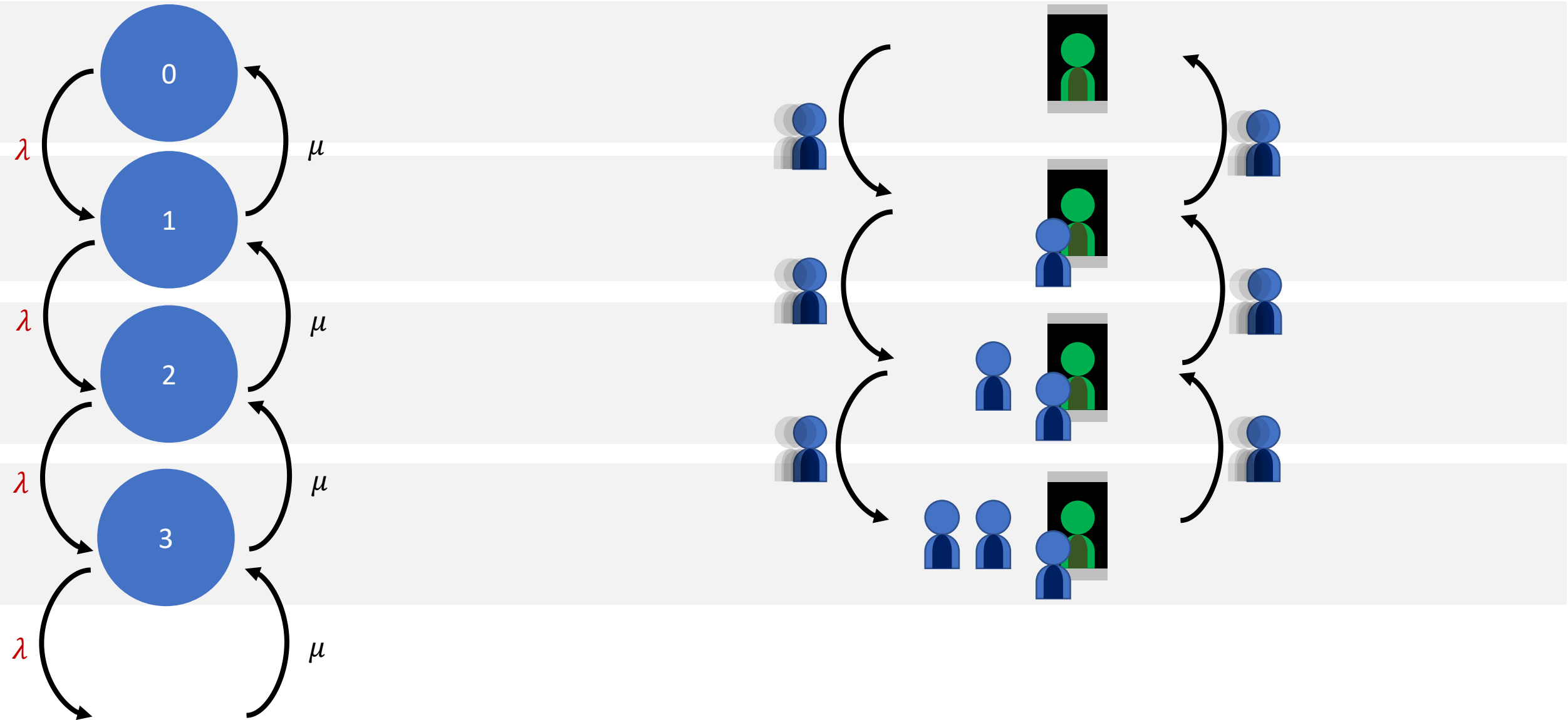
Representación de filas de espera

Podemos representar las filas como un **proceso de markov de nacimiento y muerte**.

- Las tasas de nacimiento son arribos de agentes: λ
- Las tasas de muerte son despachos de agentes: μ

Los estados del sistema es la **cantidad de agentes en el sistema**.

Representación de filas de espera



Configuraciones de filas de espera: servidores

Un solo servidor:

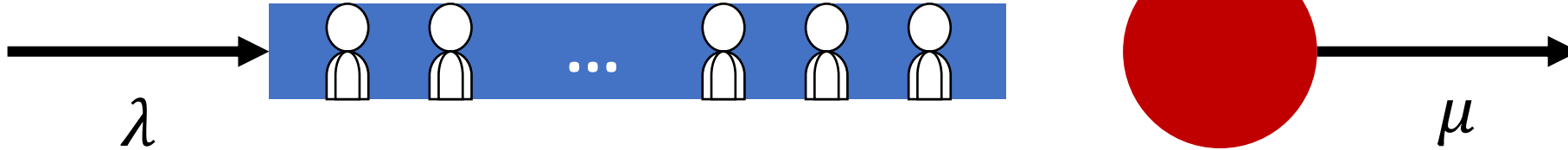


Múltiples servidores:

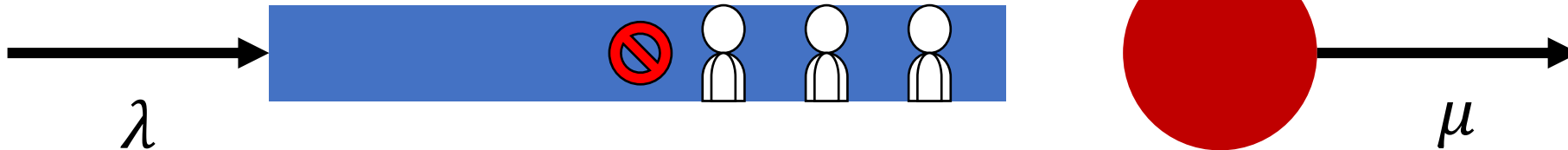


Configuraciones de filas de espera: capacidad

Capacidad de fila infinita

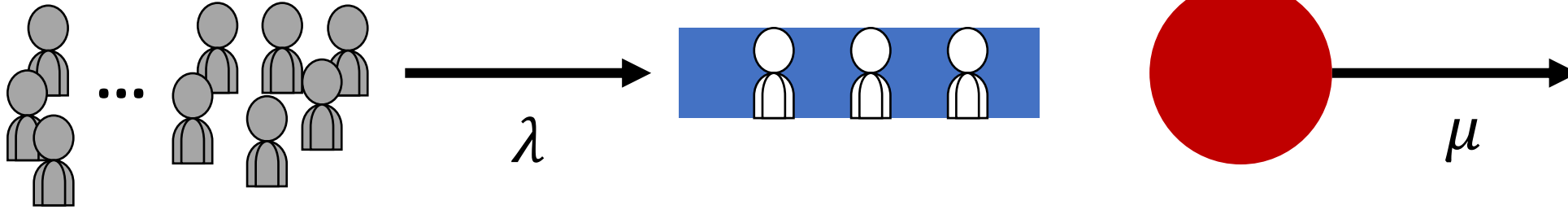


Capacidad de fila finita

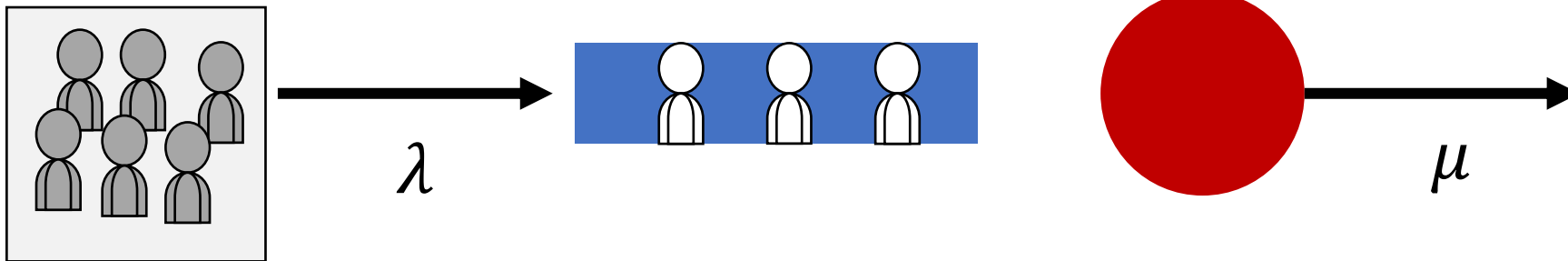


Configuraciones de filas de espera: fuente

Fuente infinita



Fuente finita



Notación de Kendall

Permite codificar la configuración de una fila.

$1/2/3/4/5/6$

Notación de Kendall

1/2/3/4/5/6

Naturaleza del proceso de arribo, ej:

- M: Los tiempos de diferencia de arribo independientes e idénticamente distribuidos (iid) siguiendo una distribución exponencial.
- D: Los tiempos de diferencia de arribo son iid y deterministas
- E_k: Los tiempos de diferencia de arribo son iid con distribución de Erlang con parámetro k.
- GI : Son iid y gobernados por una distribución general.

Notación de Kendall

1/**2**/3/4/5/6

Naturaleza del servicio:

- M: Los tiempos de servicio son independientes e idénticamente distribuidos (iid) siguiendo una distribución exponencial.
- D: Los tiempos de servicio son iid y deterministas
- E_k: Los tiempos de servicio son iid con distribución de Erlang con parámetro k.
- GI : Los tiempos de servicio son iid y gobernados por una distribución general.

Notación de Kendall

1/2/3/4/5/6

Número de servidores en paralelo.

Notación de Kendall

1/2/3/4/5/6

Disciplina de la fila:

- FCFS: First come, first served
- LCFS: Last come, first served
- SIRO: Served in random order
- SPT: Shortest processing time first
- PR: Service according to priority

Caso: política de requests con filas en Facebook

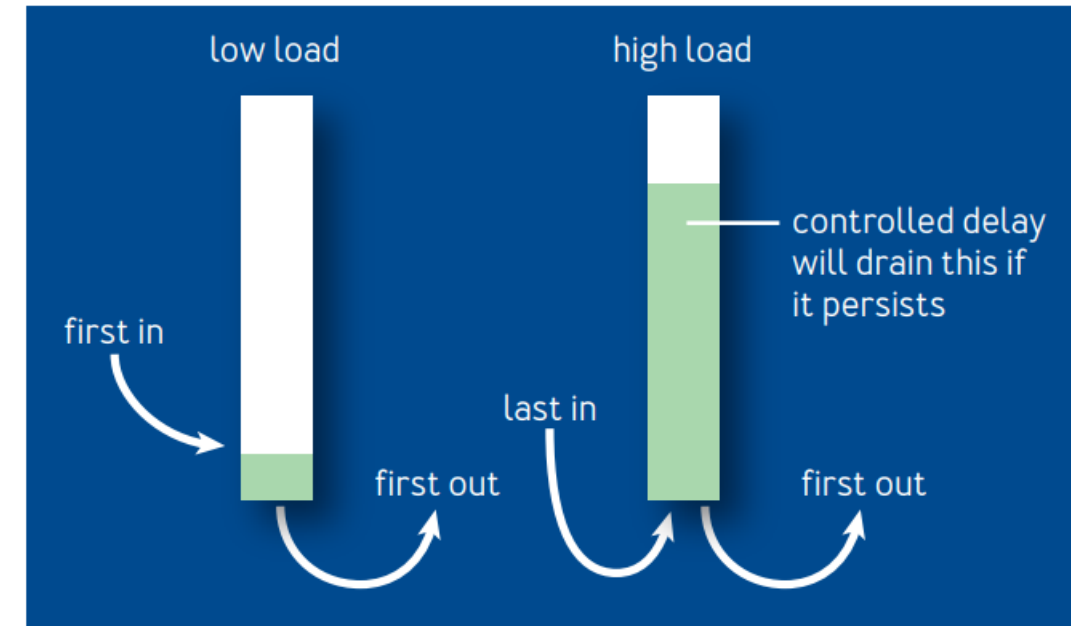
Política de fila adaptativa:

FIFO: en condiciones normales.

LIFO: alta demanda.

- Un usuario que ya esperó mucho, es
- más difícil de satisfacer.
- Probablemente ya haya abandonado.

FIGURE 2: LIFO (LEFT) AND ADAPTIVE LIFO WITH CODEL (RIGHT)



Ben Maurer, Facebook (2015), "Fail at scale, Reliability in the face of rapid change"
(<https://dl.acm.org/doi/pdf/10.1145/2838344.2839461>)

Notación de Kendall

1/2/3/4/5/6

Cantidad de clientes que puede tener el sistema (esperando + servicio)

1/2/3/4/5/6

Tamaño de la fuente donde se extraen los clientes.

Notación de Kendall

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

$M/M/3/FCFS/25/\infty$

Notación de Kendall

M/M/1/FCFS/ ∞ / ∞



- Arribos $\sim \text{Exp}(\lambda)$
- Servicio $\sim \text{Exp}(\mu)$
- 1 servidor
- Primero llegado primero servido (FCFS)
- Capacidad infinita del sistema
- Fuente infinita

Se suele abreviar a:
M/M/1

M/M/3/FCFS/25/ ∞



- Arribos $\sim \text{Exp}(\lambda)$
- Servicio $\sim \text{Exp}(\mu)$
- 3 servidores
- Primero llegado primero servido (FCFS)
- Capacidad de 25 personas
- Fuente infinita

Se suele abreviar a:
M/M/3/25

Métricas: factor de tráfico

Es la relación entre la tasa de arribos y despachos.
Si “M” es la cantidad de servidores.

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

Casos:

$\rho \geq 1$ sistema inestable.

$\rho < 1$ sistema estable.

Métricas y parámetros: factor de tráfico

Cantidad de clientes promedio:

- En la fila: L_q [unidades o agentes]
- En el sistema: L_s o L [unidades o agentes]

Tiempo de espera promedio:

- En la fila: W_q [unidad de tiempo]
- En el sistema: W_s o W [unidad de tiempo]

Probabilidad de estado (que hayan “i” agentes):

$$P(X = i)$$

Caso M/M/1

Factor de tráfico:

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

Probabilidad de sistema ocioso:

$$P_0 = 1 - \rho = 1 - \frac{\lambda}{\mu}$$



Probabilidad de sistema con “n” agentes:

$$P_n = P_0 \rho^n$$

Métricas y parámetros: factor de tráfico

Cantidad de clientes promedio

- En el sistema:

$$L = \lambda W$$

(Ley de Little)

$$L = L_q + \rho$$

$$L = \frac{\rho}{1 - \rho}$$

- En la fila:

$$L_q = \lambda W_q$$

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

Tiempo de espera promedio

- En el sistema:

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

$$W = W_q + \frac{1}{\mu}$$

- En la fila:

$$W_q = W - \frac{1}{\mu}$$

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)}$$

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

Ejemplo inicial

A una unidad de empaquetado llegan 35 unidades por hora. La media de servicio de cada unidad es de 1 minuto.

La distribución de tiempo entre arribos es exponencial y la capacidad de la fila infinita.

- 1- Escribir la notación de Kendall y representar el sistema.
- 2- Factor de tráfico.
- 3- Probabilidad de sistema ocioso.
- 4- Número promedio de unidades en la fila.
- 5- Tiempo promedio de unidades esperando.
- 5- Clientes atendidos por hora.

1) Notación de Kendall y representación

M/M/1/FCFS/ ∞ / ∞ (M/M/1)



- Arribos $\sim \text{Exp}(\lambda)$
- Servicio $\sim \text{Exp}(\mu)$
- 1 servidor
- Primero llegado primero servido (FCFS)
- Capacidad infinita del sistema
- Fuente infinita

$$\lambda = 35 \text{ u/h}$$

$$\mu = 60 \text{ u/h}$$

2) Factor de tráfico

$$\rho = \frac{\lambda}{\mu} = \frac{35}{60} = 0.58$$

3) Probabilidad de sistema ocioso

$$P_0 = 1 - \rho = 1 - \frac{\lambda}{\mu}$$

$$P_0 = 1 - \frac{35}{60} = 0.41\hat{6}$$

La probabilidad de encontrar el sistema ocioso es de **41.66%**

4) Número promedio de unidades en la fila

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{35^2}{60(60 - 35)} = 0.8166 \text{ unidades}$$

5) Tiempo promedio de unidades esperando

$$W_q = \frac{L_q}{\lambda} = \frac{0.8166}{35} = 0.0233 \text{ horas} = 1.39 \text{ min}$$

6) Clientes atendidos por hora

a) calculamos la cantidad promedio en el servidor:

$$L_s - L_q = \rho$$

b) multiplicamos por la capacidad operativa del servidor:

$$\rho\mu = 0.58 * 60 = \text{35 clientes/hora}$$

6) Clientes atendidos por hora

Otra forma de entenderlo:

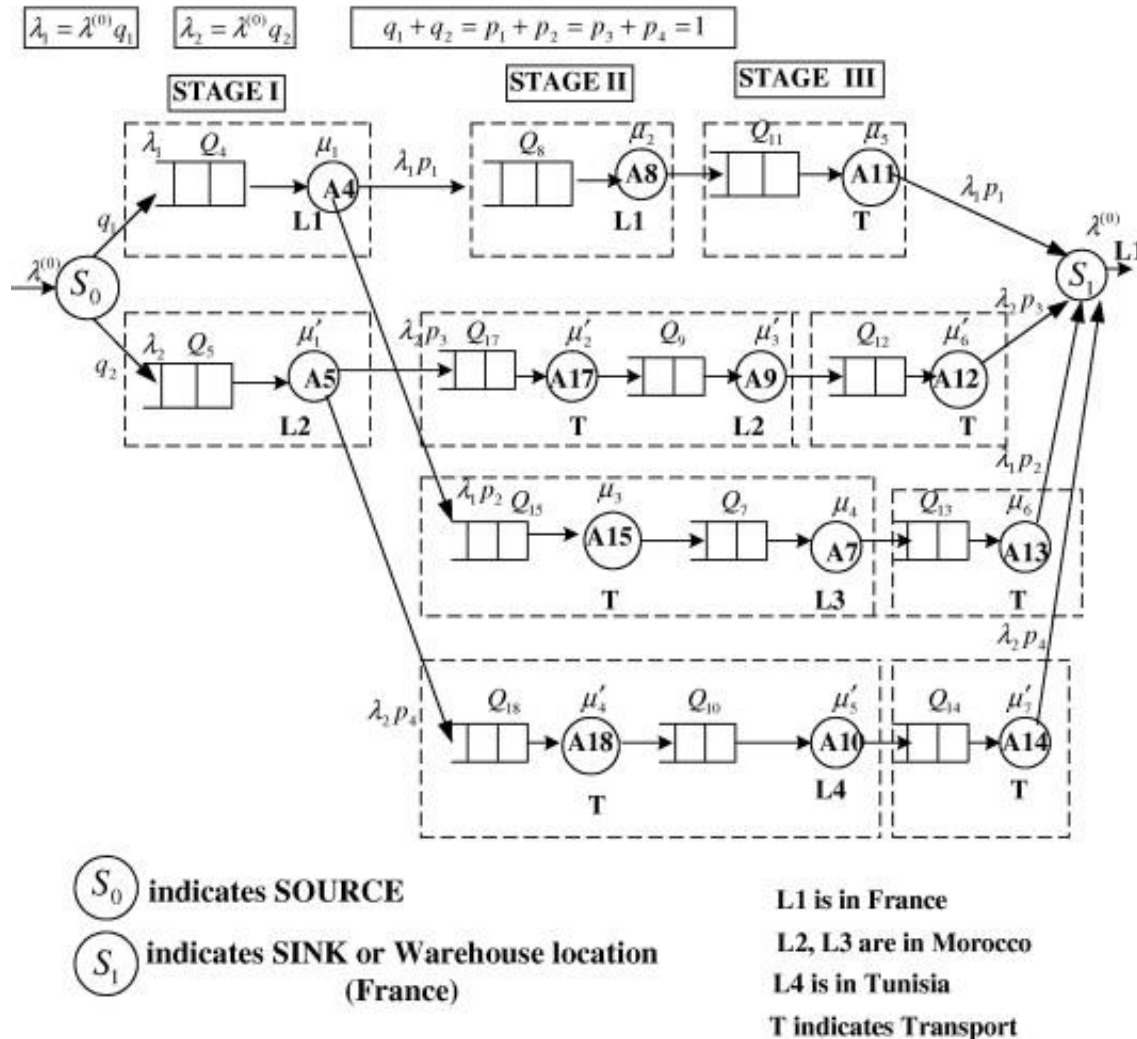
μ es la capacidad operativa total.

Dado que $\mu > \lambda$

Los clientes atendidos por hora son $\lambda = 35 \text{ clientes/hora}$

Caso: Supply Chain Modelling

Fig. 1: Queuing formulation of the network of processes



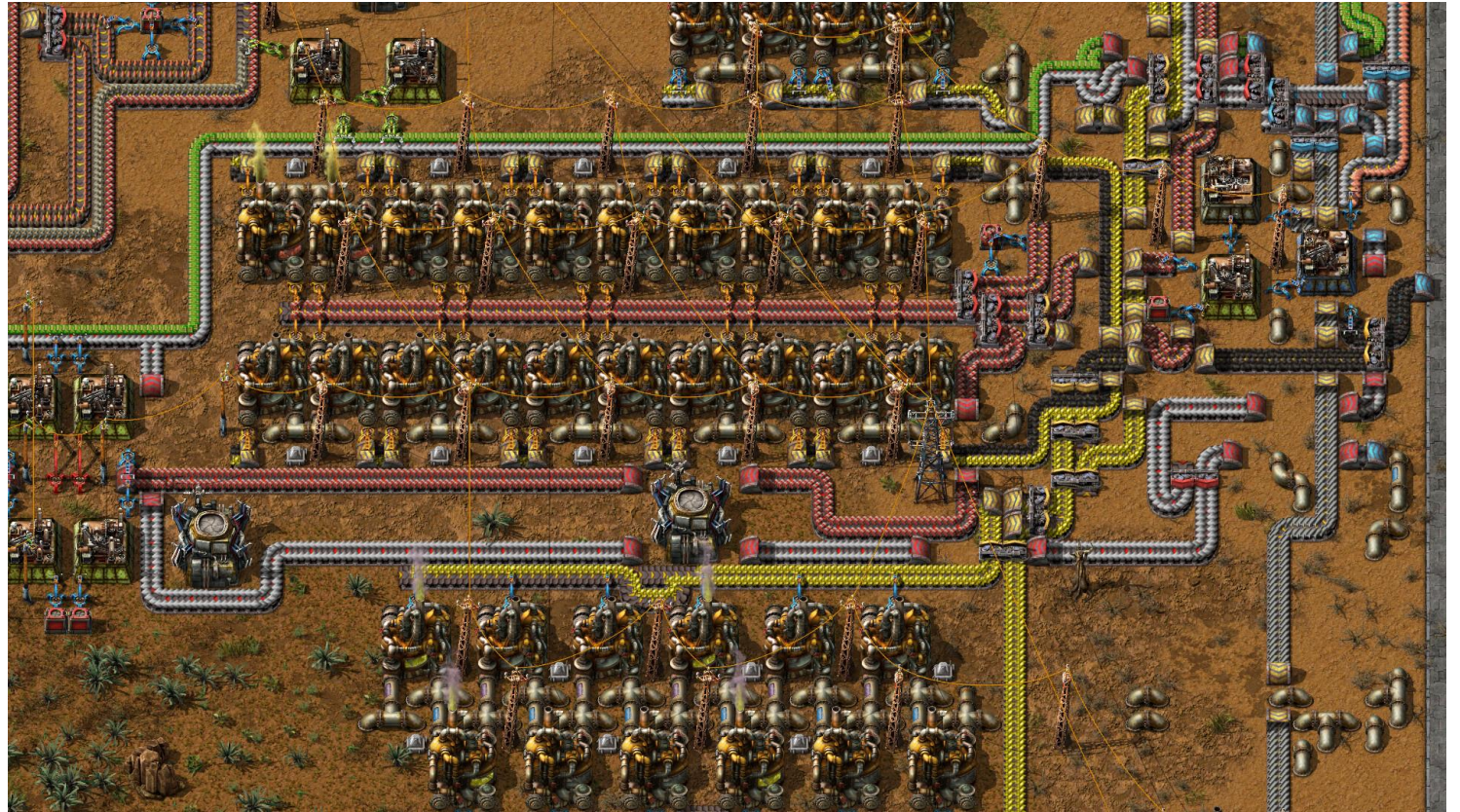
- Modelización con red de filas de espera.
- Casos límite.
- Casos equivalentes.
- Medición de performance.

Fuente: Bahskar (2010) "Modeling a supply chain using a network of queues"
 (<https://www.sciencedirect.com/science/article/pii/S0307904X09003382#bib29>)

Caso: Factorio (2020)



- Asignación de recursos.
- Supply Chain.
- Optimización de redes logísticas.
- Balanceo de línea.



Fuente: <https://store.steampowered.com/app/427520/Factorio/?l=spanish>

Little's Law in Factorio: <https://johanneshoff.com/little-factorio/>