MODELOS MATEMÁTICOS PARA ANALIZAR TIEMPOS DE ESPERA: MINIMIZANDO EL TIEMPO DE RESPUESTA EN CENTROS DE DATOS

Daniela Hurtado Lange

Estudiante de doctorado en Investigación Operativa

Georgia Institute of Technology

9 de enero 2020

¿QUIÉN SOY Y POR QUÉ POSTULAR?

EN MI CV…

PUC Chile

- Ingeniería Industrial Matemática
- Magíster en Ingeniería Industrial
  - Profesor guía: Pedro Gazmuri
- Tutora CARA
- Tutora Talento & Inclusión

Georgia Tech

- Alumna de doctorado en Investigación Operativa
- Tutora de investigación de pregrado (SURE)

TRABAJO EN CONJUNTO CON

Siva Theja Maguluri

Profesor Asistente

Georgia Institute of Technology

MOTIVACIÓN

Minimizar tiempo de espera

HOJA DE RUTA

- Función Generadora de Momentos (FGM)
- Cadenas de Markov en Tiempo Discreto (CMTD)
- Sistemas de espera
- Análisis asintótico: Alto tráfico

- Estado del arte
- Método de la Función Generadora de Momentos
- Limitaciones

FUNCIÓN GENERADORA DE MOMENTOS (FGM)

- Transformada de Laplace de una v.a.
- **Definición**: FGM de la v.a. X es $f(\theta) = E\left[e^{\theta X}\right]$, donde $\theta \in \mathbb{R}$ es un parámetro
- La FGM determina completamente la distribución de la v.a. X
- Propiedad:

$$\frac{d}{d\theta}E\left[e^{\theta X}\right] = E\left[Xe^{\theta X}\right] \quad \Rightarrow \quad \frac{d}{d\theta}E\left[e^{\theta X}\right]\Big|_{\theta=0} = E[X]$$

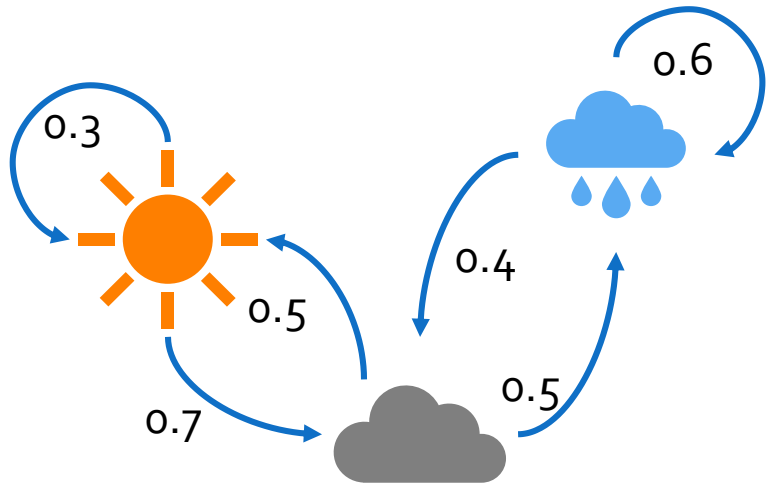
$$\frac{d^2}{d\theta^2}E\left[e^{\theta X}\right] = E\left[X^2e^{\theta X}\right] \quad \Rightarrow \quad \frac{d^2}{d\theta^2}E\left[e^{\theta X}\right]\Big|_{\theta=0} = E[X^2]$$

$$\vdots$$

$$\frac{d^m}{d\theta^m}E\left[e^{\theta X}\right] = E\left[X^me^{\theta X}\right] \quad \Rightarrow \quad \frac{d^m}{d\theta^m}E\left[e^{\theta X}\right]\Big|_{\theta=0} = E[X^m]$$

CADENAS DE MARKOV EN TIEMPO DISCRETO (CMTD)

- Proceso Estocástico
  - Tiempo discreto: Por etapas
  - Estacionario
  - Propiedad Markoviana:
    "El futuro depende del pasado solo a través del presente"
- Matriz de probabilidades de transición
  - Matriz estocástica
  - Evolución: $x(k+1) = x(k)P$
- Análisis de largo plazo: $x(k+1) = x(k)$
  - Positiva recurrente
  - Resolver $\pi = \pi P$



MAÑANA

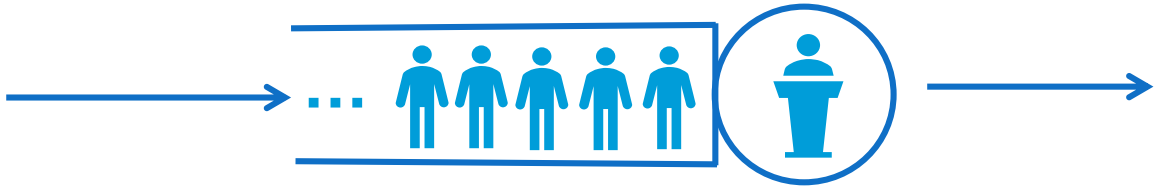
$$P = \begin{array}{c} \\ \text{HOY} \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc} \text{S} & \text{N} & \text{LL} \end{array} \\ \begin{bmatrix} 0.3 & 0.7 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 0.4 & 0.6 \end{bmatrix} \end{array} \begin{array}{l} \text{S} \\ \text{N} \\ \text{LL} \end{array}$$

$$\begin{array}{cccc} & \text{S} & \text{N} & \text{LL} \\ \pi = [& 0.24 & 0.34 & 0.42] \end{array}$$

En el **largo plazo**, la probabilidad de lluvia es 0.42

EJEMPLO DE CMTD: SISTEMA DE ESPERA

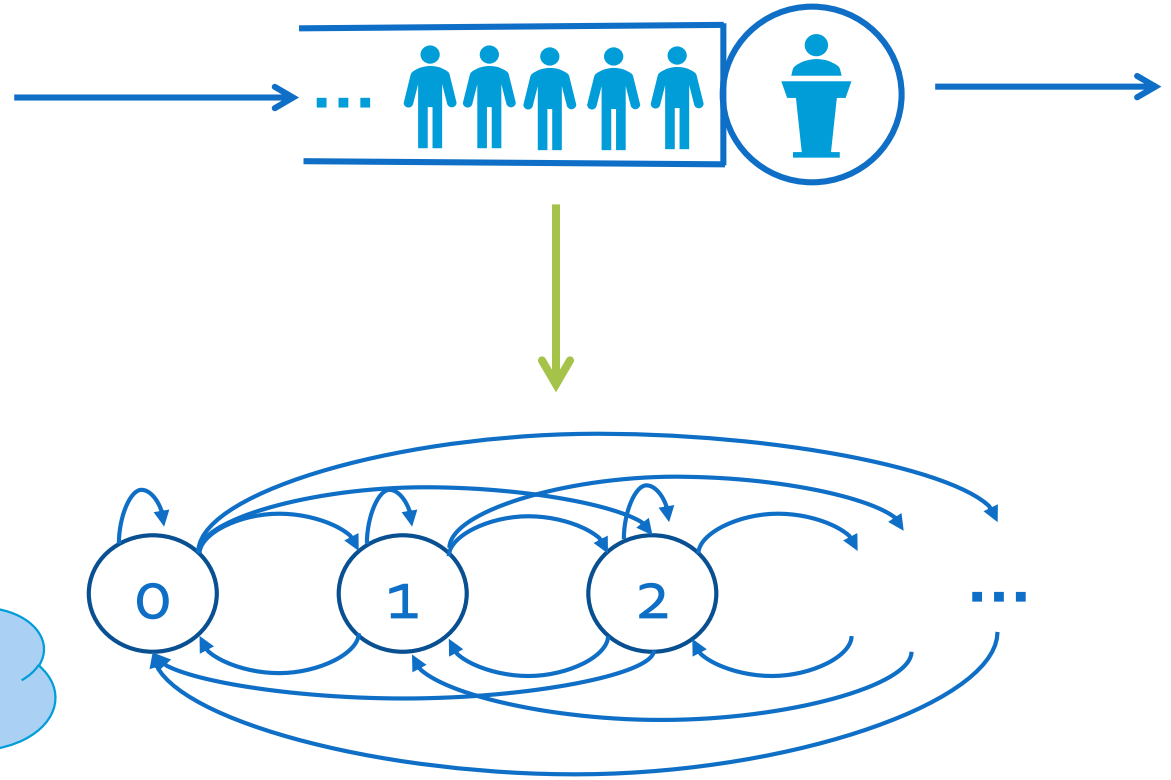
- Sistema de un solo servidor
- Número de llegadas en cada ventana de tiempo es aleatoria
- Si el servidor está ocupado, los trabajos esperan en una cola
- El servidor atiende un número aleatorio de clientes en cada ventana de tiempo
- Luego de ser atendidos, los trabajos dejan el sistema
- Hay infinito espacio en la cola



EJEMPLO DE CMTD: SISTEMA DE ESPERA (cont.)

MODELO

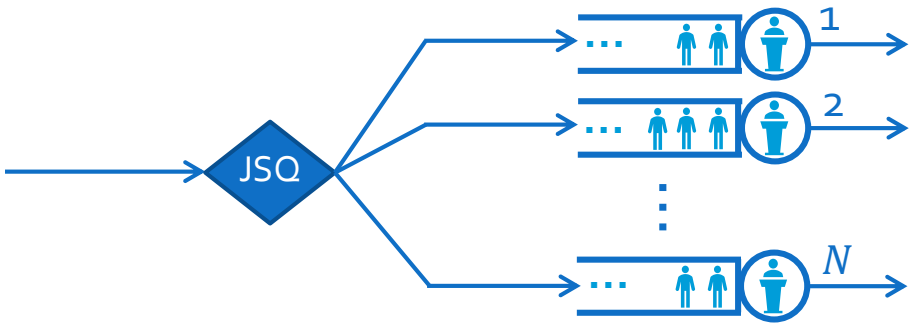
- Estado: #trabajos en sistema
- Transiciones
  - Depende de la distribución de llegadas
  - Depende de la distribución de servicio
- Infinitos estados

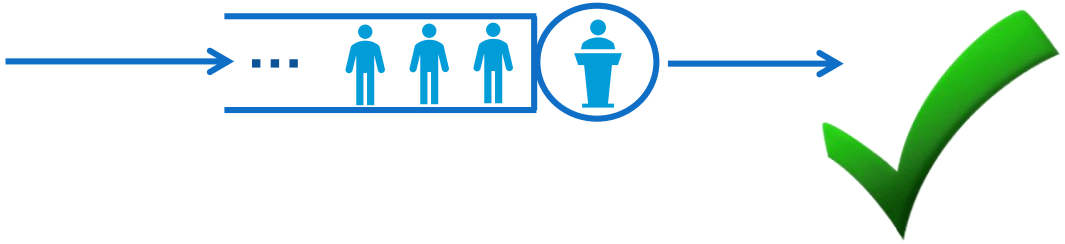


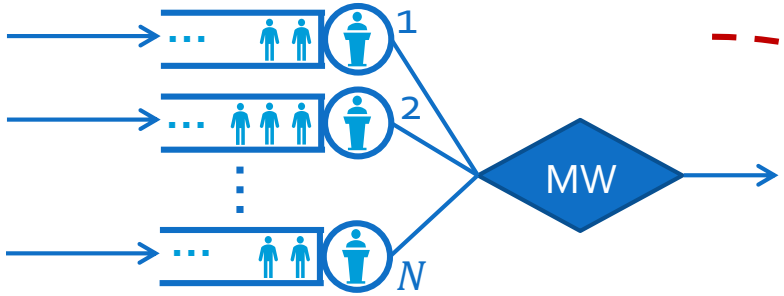
¿Matriz P?

¿Largo plazo?

ANÁLISIS ASINTÓTICO: TRÁFICO ALTO





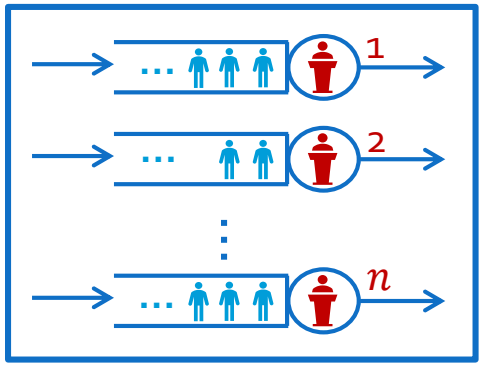


Colapso del espacio de estados

- Cargar sistema a capacidad máxima

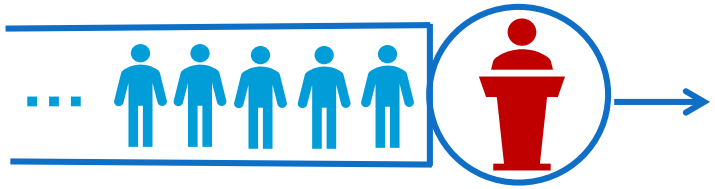
  Tasa de llegada ≈ Tasa de servicio

- "Peor de los casos"



$n < N$

o bien



HOJA DE RUTA

- Función Generadora de Momentos (FGM)
- Cadenas de Markov en Tiempo Discreto (CMTD)
- Sistemas de espera
- Análisis asintótico: Alto tráfico

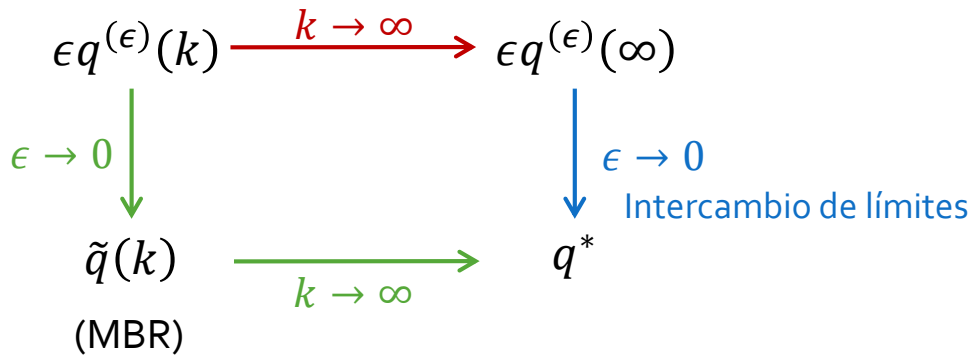
- Estado del arte
- Método de la Función Generadora de Momentos
- Limitaciones

LÍMITE DE TRÁFICO ALTO

- Cargar el sistema al máximo de su capacidad

LÍMITES DE DIFUSIÓN

- Escalar tiempo y espacio: $\epsilon = 1/\sqrt{n}$



MÉTODO *DRIFT* [Eryilmaz, Srikant 13]

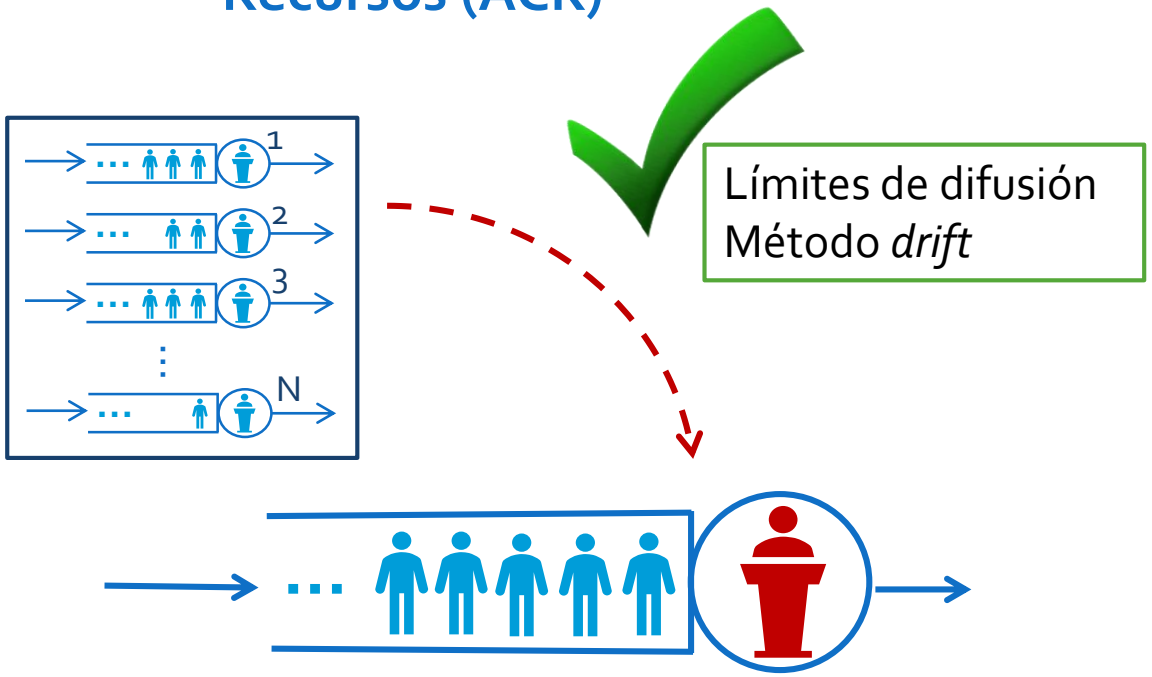
- *Drift* = Cambio en un intervalo de tiempo
- Igualar a cero el *drift* de una función de prueba
- $k \to \infty$:
$$E\left[V\left(q^{(\epsilon)}(k)\right)\right] = E\left[V\left(q^{(\epsilon)}(k+1)\right)\right]$$
- $\epsilon \to 0$: $E[V'(q^*)]$

- No se requiere intercambiar límites
- ¿Cómo escoger $V(\cdot)$?

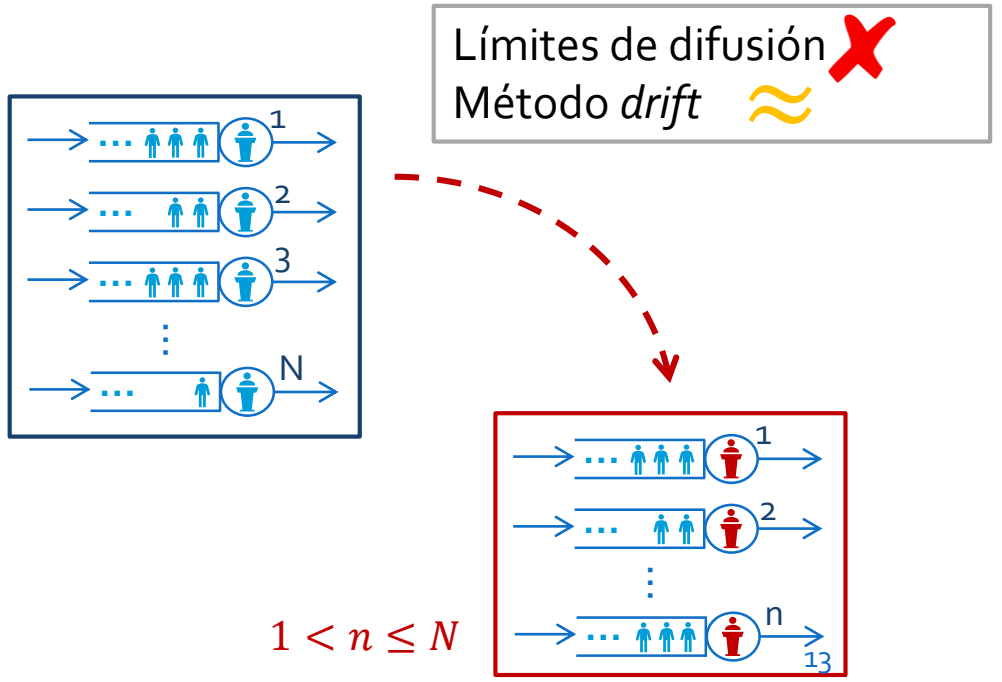
COLAPSO DEL ESPACIO DE ESTADOS (CEE)

- Sistema de espera se comporta como si tuviese una dimensión menor

Agrupamiento Completo de los Recursos (ACR)

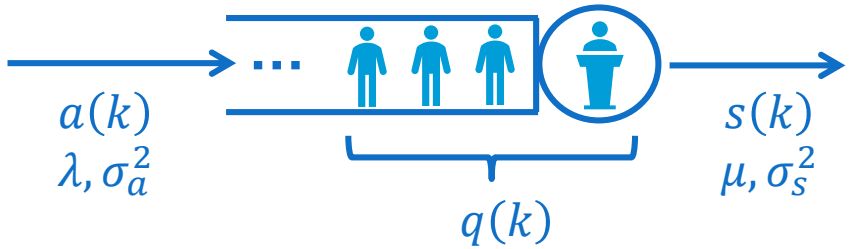


No ACR



MODELO DE COLA CON UN SOLO SERVIDOR

- Modelo en tiempo discreto
- $q(k)$: # trabajos en el sistema al comienzo de la ventana de tiempo k
- $a(k)$: # llegadas en la ventana de tiempo k
  - Media λ y varianza σ_a^2
- $s(k)$: # servicios **ofrecidos** en la ventana de tiempo k
  - Media μ y varianza σ_s^2
- Ambas son sequencias de v.a.i.i.d., independientes de la otra
- Estabilidad: $\lambda < \mu$



- Evolución de la cola

$$q(k+1) = [q(k) + a(k) - s(k)]^+$$
$$= q(k) + a(k) - s(k) + u(k)$$

Servicio no utilizado

$$q(k+1)u(k) = 0$$

ALTO TRÁFICO – MÉTODO *DRIFT* [Kingman, 1962]

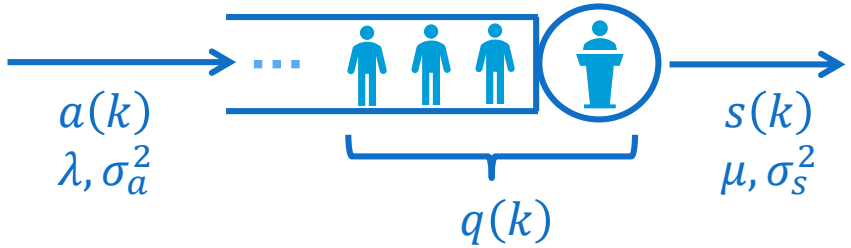
- Cargar sistema a su máxima capacidad
  - $\lambda = \mu - \epsilon$, con $\epsilon > 0$
  - Límite cuando $\epsilon \downarrow 0$

- En estado estacionario ($k \to \infty$):

$$E[q^2(k+1)] = E[q^2(k)]$$

- Por lo tanto,

$$\lim_{\epsilon \downarrow 0} E[\epsilon q] = \frac{\sigma_a^2 + \sigma_s^2}{2}$$

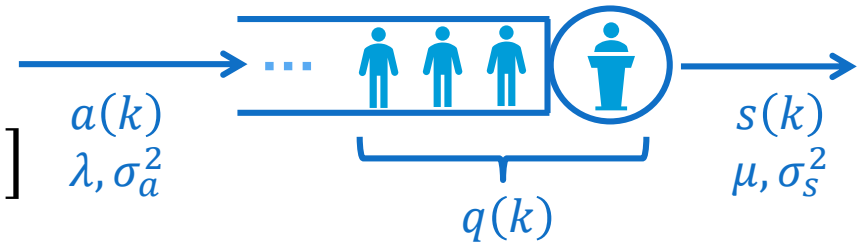


$$q(k+1) = [q(k) + a(k) - s(k)]^+$$
$$= q(k) + a(k) - s(k) + u(k)$$

$$q(k+1)u(k) = 0$$

MÉTODO *DRIFT* [Eryilmaz, Srikant 13]

$\epsilon = \mu - \lambda$



- En estado estacionario, $E\left[V\left(q(k+1)\right)\right] = E\left[V\left(q(k)\right)\right]$

$V(q)$:

q^2 $\quad$ q^3 $\quad$ … $\quad$ q^{m+1}

Obtenemos:

$E[\epsilon q]$ $\quad$ $E[\epsilon^2 q^2]$ $\quad$ … $\quad$ $E[\epsilon^m q^m]$

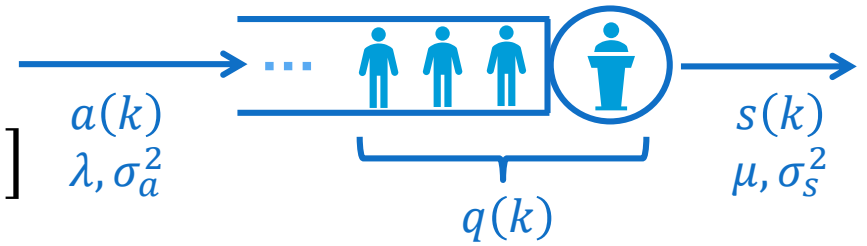
$\frac{\sigma_a^2 + \sigma_s^2}{2}$ $\quad$ $2\left(\frac{\sigma_a^2 + \sigma_s^2}{2}\right)^2$ $\quad$ $m!\left(\frac{\sigma_a^2 + \sigma_s^2}{2}\right)^m$

¿Podemos hacer este cálculo más eficiente?

$$\epsilon q \Rightarrow Expo\left(\frac{\sigma_a^2 + \sigma_s^2}{2}\right)$$

MÉTODO *DRIFT* [Eryilmaz, Srikant 13]

$\epsilon = \mu - \lambda$



- En estado estacionario, $E\left[V\left(q(k+1)\right)\right] = E\left[V\left(q(k)\right)\right]$

$V(q)$:
$$1 + \epsilon q + \frac{1}{2}\epsilon^2 q^2 + \frac{1}{3!}\epsilon^3 q^3 + \cdots + \frac{1}{(m+1)!}\epsilon^{m+1} q^{m+1} + \cdots = e^{\epsilon q}$$

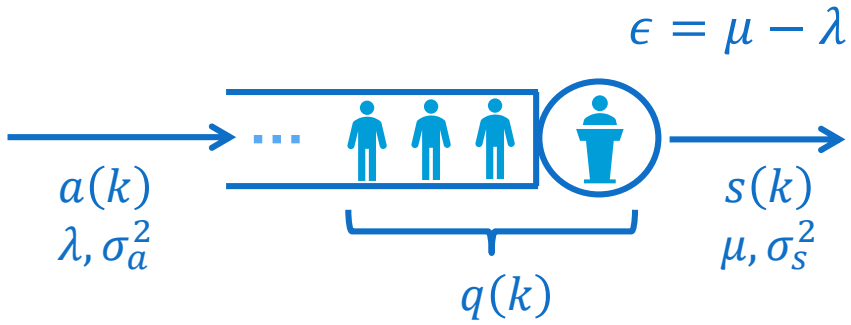
Obtenemos:
$$1 + E[\epsilon q] + \frac{1}{2}E[\epsilon^2 q^2] + \cdots + \frac{1}{m!}E[\epsilon^m q^m] + \cdots = E[e^{\epsilon q}]$$

$$\frac{\sigma_a^2 + \sigma_s^2}{2} \qquad 2\left(\frac{\sigma_a^2 + \sigma_s^2}{2}\right)^2 \qquad m!\left(\frac{\sigma_a^2 + \sigma_s^2}{2}\right)^m$$

$$\epsilon q \Rightarrow Expo\left(\frac{\sigma_a^2 + \sigma_s^2}{2}\right)$$

MÉTODO DE LA FGM [HL, Maguluri 19]

- En estado estacionario: $E\left[e^{\theta\epsilon q(k+1)}\right] = E\left[e^{\theta\epsilon q(k)}\right]$



- Lema clave:

$$\left(e^{\theta\epsilon q(k+1)} - 1\right)\left(e^{-\theta\epsilon u(k)} - 1\right) = 0$$

$$q(k+1) = q(k) + a(k) - s(k) + u(k)$$

- Reorganizando términos

$$E\left[e^{\theta\epsilon q}\right] = \frac{1 - E\left[e^{-\theta\epsilon u}\right]}{1 - E\left[e^{\theta\epsilon(a-s)}\right]}$$

$$q(k+1)u(k) = 0$$

$\epsilon \downarrow 0$

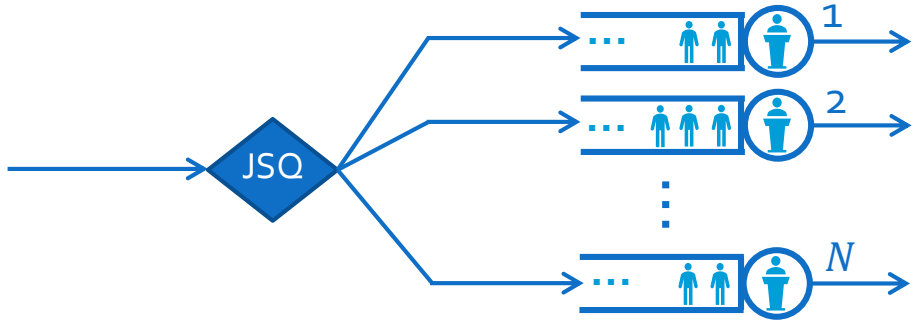
$$\frac{1}{1 - \theta\left(\frac{\sigma_a^2 + \sigma_s^2}{2}\right)}$$

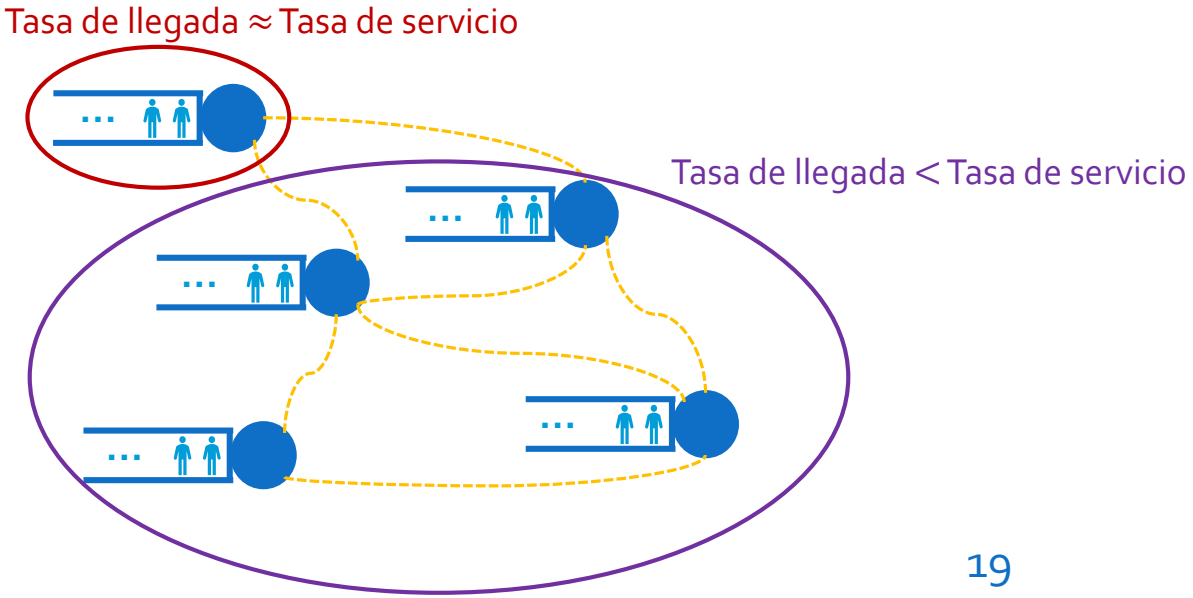
FGM de v.a. Exponencial con media $\frac{\sigma_a^2 + \sigma_s^2}{2}$

$$\epsilon q \Rightarrow Expo\left(\frac{\sigma_a^2 + \sigma_s^2}{2}\right)$$

¿OTROS SISTEMAS DE ESPERA?

- Colapso del Espacio de Estados (CEE):

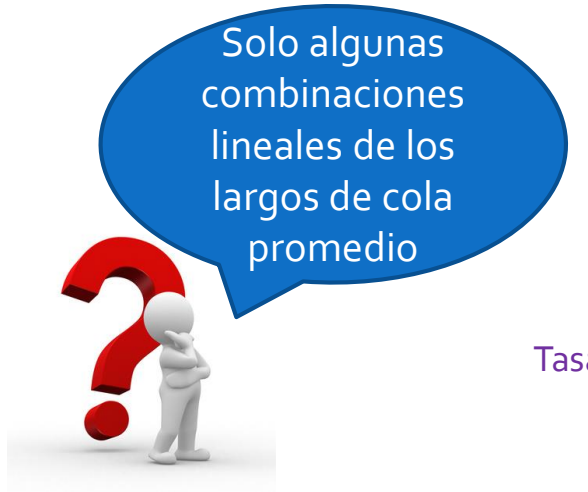


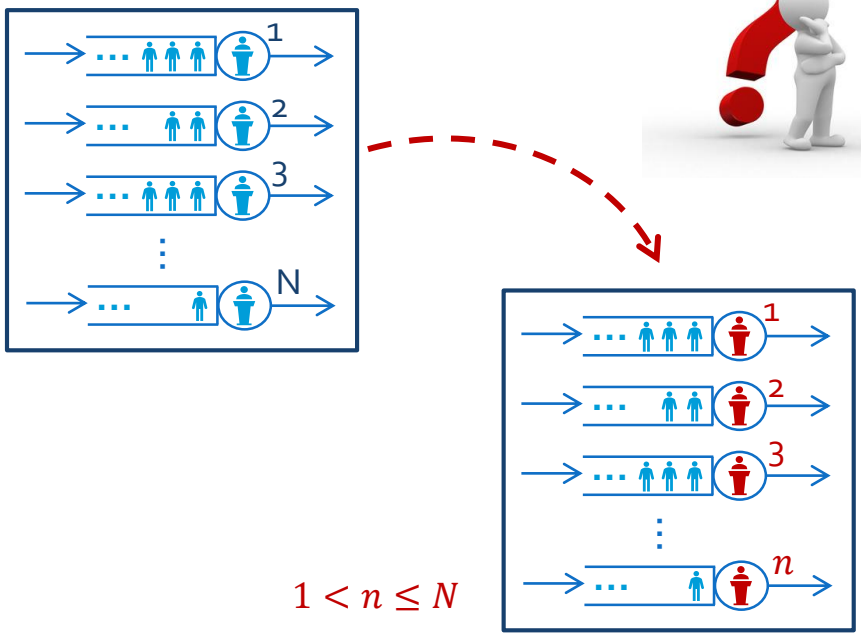


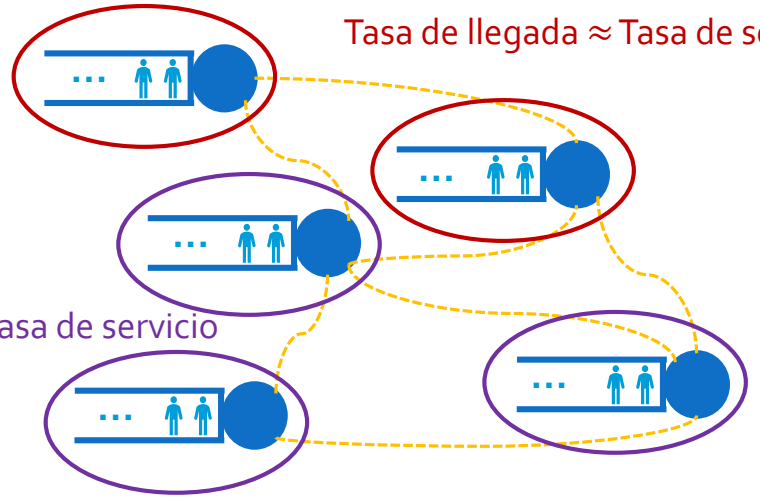
¿OTROS SISTEMAS DE ESPERA? (cont.)

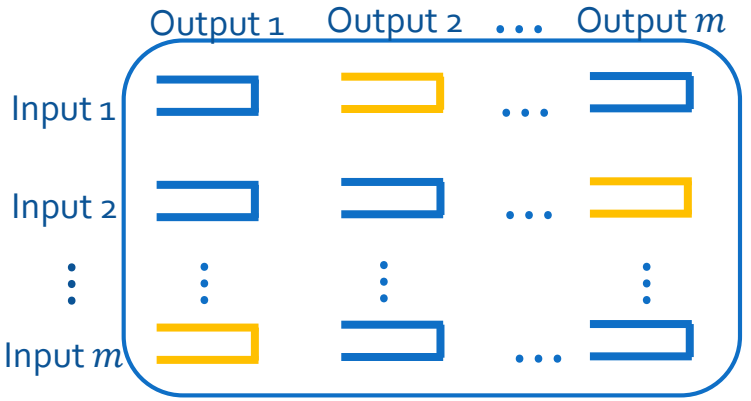
- Colapso del Espacio de Estados (CEE):

NO ACR









HOJA DE RUTA

Conceptos básicos

- Función Generadora de Momentos (FGM)
- Cadenas de Markov en Tiempo Discreto (CMTD)
- Sistemas de espera
- Análisis asintótico: Alto tráfico

Proyecto de investigación

- Estado del arte
- Método de la Función Generadora de Momentos
- Limitaciones

Conclusiones y trabajo futuro

CONCLUSIONES Y TRABAJO FUTURO

Agrupamiento Completo de los Recursos (ACR)

- Sistema de espera se comporta como una cola de un servidor
- Método FGM para obtener la distribución ϵq en límite de tráfico alto
- La clave está en el servicio no utilizado

$$q(k+1)u(k) = 0$$

$$\left(e^{\theta\epsilon q(k+1)} - 1\right)\left(e^{-\theta\epsilon u(k)} - 1\right) = 0$$

No ACR

- Solo podemos obtener **algunas** combinaciones lineales del **promedio** del largo de cola

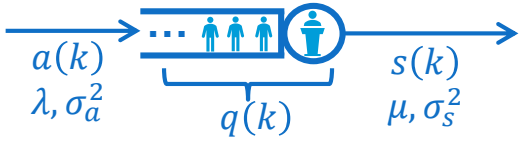
Trabajo futuro

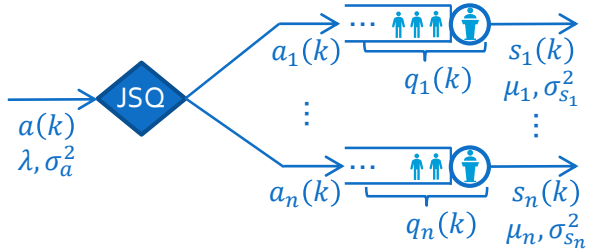
- Generalizar el método FGM a sistemas que no satisfacen ACR
- Optimizar problema de control
  - Ruteo
  - *Scheduling*

¡Gracias!

¿Preguntas?

MÉTODO FGM [Hurtado Lange, Maguluri 19]





$$\boldsymbol{q}_{\parallel} = \mathbf{1}\frac{\sum_i q_i}{n} \quad \& \quad \boldsymbol{q}_{\perp} = \boldsymbol{q} - \boldsymbol{q}_{\parallel}$$

- Igualar el *drift* de $V(q) = e^{\theta\epsilon q}$ a cero

$$\left(e^{\theta\epsilon q(k+1)} - 1\right)\left(e^{-\theta\epsilon u(k)} - 1\right) = 0$$

- Obtenemos

$$\lim_{\epsilon\to 0} E\left[e^{\theta\epsilon q}\right] = \frac{1}{1 - \theta\left(\frac{\sigma_a^2 + \sigma_s^2}{2}\right)}$$

$$\epsilon q \Rightarrow Expo\left(\frac{\sigma_a^2 + \sigma_s^2}{2}\right)$$

- Igualar el *drift* de $V(\boldsymbol{q}) = e^{\theta\epsilon \sum_i q_i}$ a cero

$$E\left[\left(e^{\theta\epsilon \sum_i q_i(k+1)} - 1\right)\left(e^{-\theta\epsilon \sum_i u_i(k)} - 1\right)\right] \text{ is } o(\epsilon^2)$$

- Obtenemos

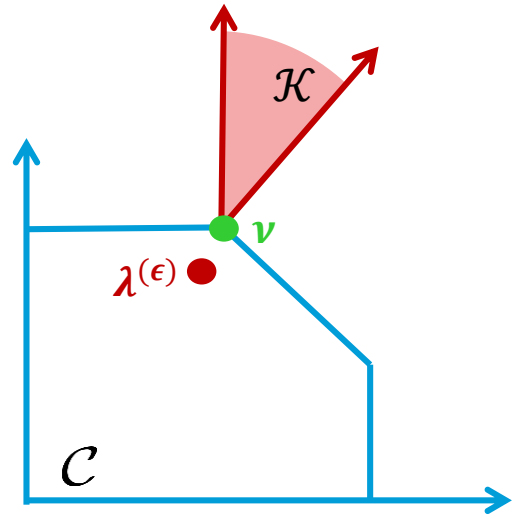
$$\lim_{\epsilon\to 0} E\left[e^{\theta\epsilon \sum_i q_i}\right] = \frac{1}{1 - \theta\left(\frac{\sigma_a^2 + \sum_i \sigma_{s_i}^2}{2}\right)}$$

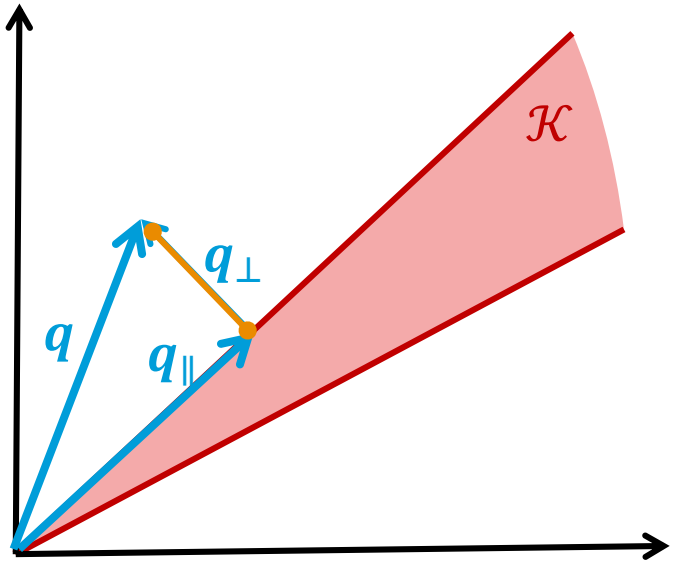
- Entonces, $\epsilon\boldsymbol{q}_{\parallel} \xrightarrow{d} \frac{\mathbf{1}}{n} Expo\left(\frac{\sigma_a^2 + \sum_i \sigma_{s_i}^2}{2}\right)$
- CEE implica $\epsilon\boldsymbol{q}_{\perp} \Rightarrow \mathbf{0}$

$$\epsilon\boldsymbol{q} \Rightarrow \frac{\mathbf{1}}{n} Expo\left(\frac{\sigma_a^2 + \sum_i \sigma_{s_i}^2}{2}\right)$$

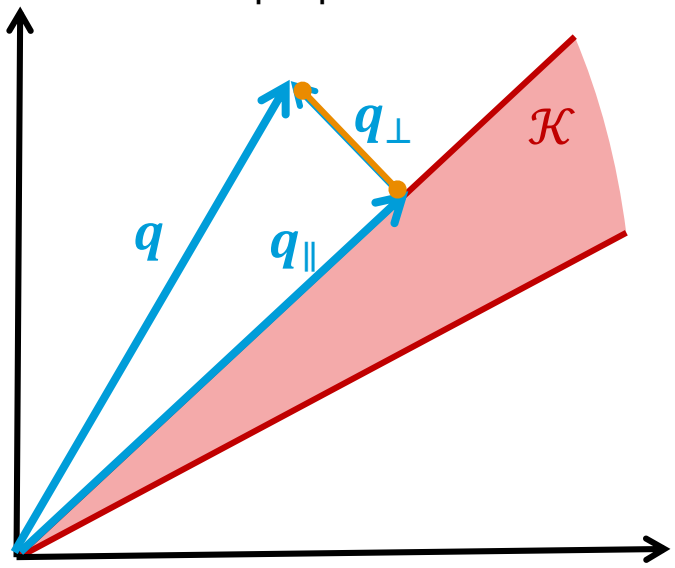
COLAPSO DEL ESPACIO DE ESTADOS

Proposición:
$\boldsymbol{q}_{\parallel}(k)$: proyección de $\boldsymbol{q}(k)$ en $\mathcal{K}$
$\boldsymbol{q}_{\perp}(k) \coloneqq \boldsymbol{q}(k) - \boldsymbol{q}_{\parallel}(k)$: error de approximar $\boldsymbol{q} \approx \boldsymbol{q}_{\parallel}$
Entonces, $E[\|\boldsymbol{q}_{\perp}\|^t] \leq T_t$ para todo $t = 1,2, \ldots$





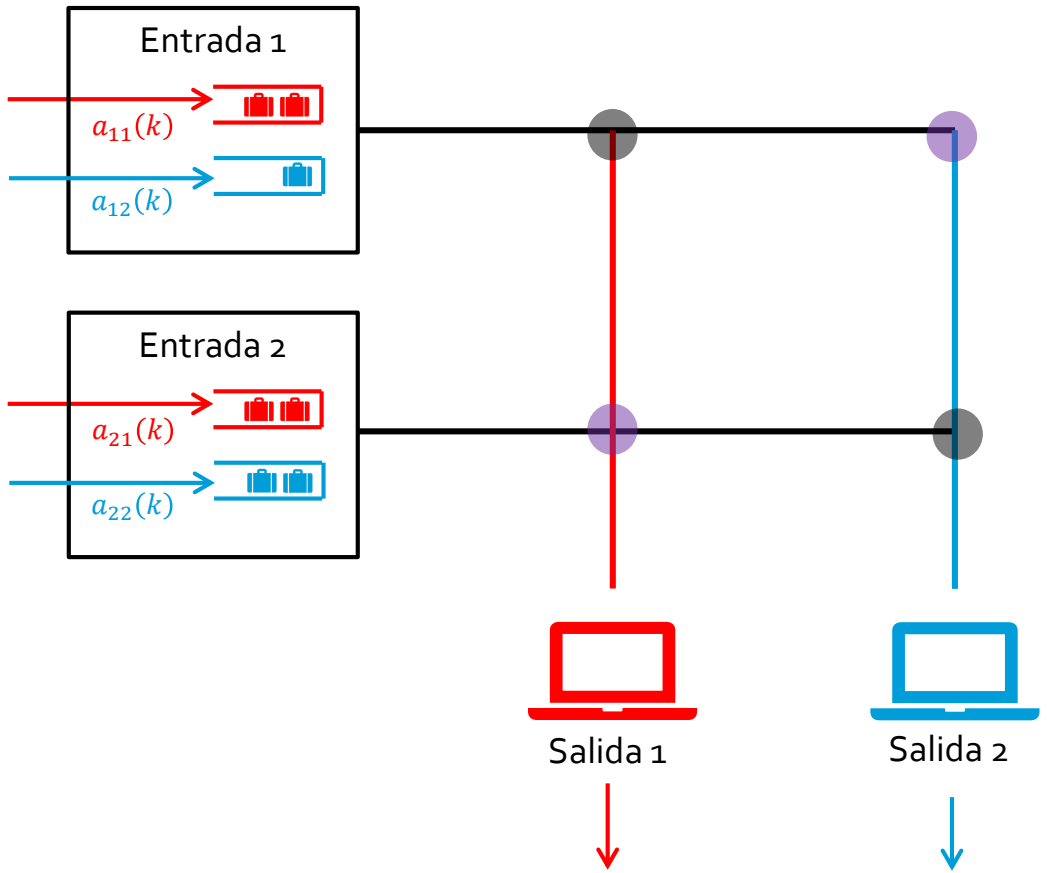
ϵ más pequeño:



$\boldsymbol{q} \approx \boldsymbol{q}_{\parallel}$ as $\epsilon \downarrow 0$

SWITCH CON COLAS EN LAS ENTRADAS

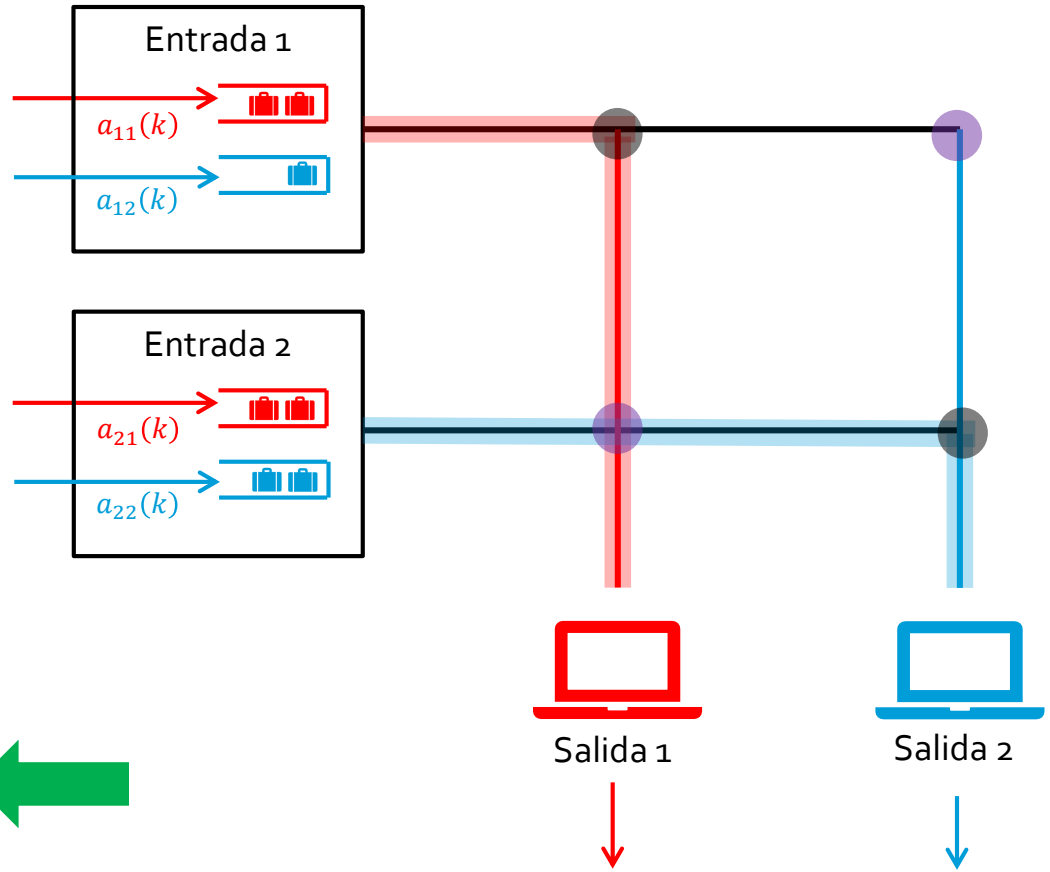
- Sistema de espera en tiempo discreto
- m entradas y m salidas
- Los trabajos llegan a cada entrada con una salida predeterminada
  - Hay una cola para cada par entrada/salida
- Todos los trabajos tienen tamaño 1
- Todas las entradas están conectadas a todas las salidas
- **Restricción**: Procesar a lo más un trabajo de cada entrada y uno en cada salida, en cada ventana de tiempo



¿Qué colas deberíamos servir?

ALGORITMO MAXWEIGHT

- MaxWeight = Escoger la conexión con el mayor peso
- Peso = Total de trabajos en línea en toda la conexión

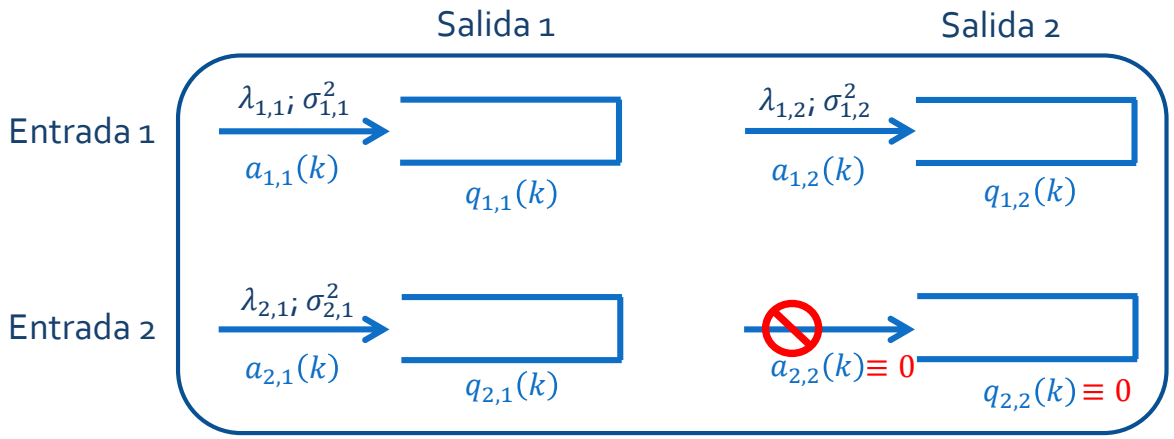


	Entrada 1	Entrada 2	Peso
Conexión 1	Salida 1	Salida 2	2+2 = 4
Conexión 2	Salida 2	Salida 1	1+2 = 3

NUEVA VISIÓN DEL MÉTODO *DRIFT*

- Sistema no-ACR más simple
  - *Switch* de 2x2 operando con MaxWeight
  - No hay llegadas a la cola (2,2)
  - Llegadas a las otras colas con media $\lambda_{i,j}$ y varianza $\sigma_{i,j}^2$
- Estabilidad:
  - Tasa de llegada < Tasa de servicio
  - $\lambda_{1,1} + \lambda_{1,2} < 1 \;\&\; \lambda_{1,1} + \lambda_{2,1} < 1$
- Tráfico alto: $\epsilon > 0$
  - Tasa de llegada ≈ Tasa de servicio
  - $\lambda_{1,1} = 1 - \lambda - \epsilon \;\&\; \lambda_{1,2} = \lambda_{2,1} = \lambda$

$\Rightarrow \lambda_{1,1} + \lambda_{1,2} = 1 - \epsilon \;\&\; \lambda_{1,1} + \lambda_{2,1} = 1 - \epsilon$



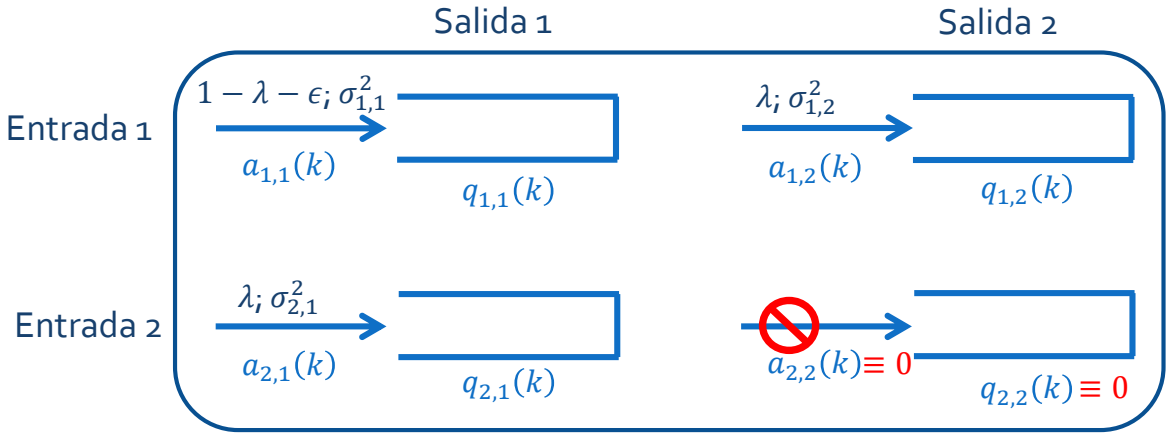
Dimensión 2
⇒No-ACR

Colapso del Espacio de Estados: [Maguluri et.al.,18]

$$\mathcal{K} = \left\{ \begin{bmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & 0 \end{bmatrix} : x_{1,1} = x_{1,2} + x_{2,1} \right\}$$

NUEVA VISIÓN DEL MÉTODO *DRIFT* (cont.)

- CEE:
  - $\boldsymbol{q}_{\|}$: Proyección de $\boldsymbol{q}$ en $\mathcal{K}$
  - $E[\|\boldsymbol{q}\|^2] \approx E\left[\left\|\boldsymbol{q}_{\|}\right\|^2\right]$
  - $q_{\|1,1} = q_{\|1,2} + q_{\|2,1}$



- Función de prueba cuadrática más general

$$V(\boldsymbol{q}) = \alpha_1 q^2_{\|1,2} + \alpha_2 q^2_{\|2,1} + \alpha_3 q_{\|1,2} q_{\|2,1}$$

$$V_1(\boldsymbol{q}) = q^2_{\|1,2}$$
$$V_2(\boldsymbol{q}) = q^2_{\|2,1}$$
$$V_3(\boldsymbol{q}) = q_{\|1,2} q_{\|2,1}$$

$$\mathcal{K} = \left\{ \begin{bmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & 0 \end{bmatrix} : x_{1,1} = x_{1,2} + x_{2,1} \right\}$$

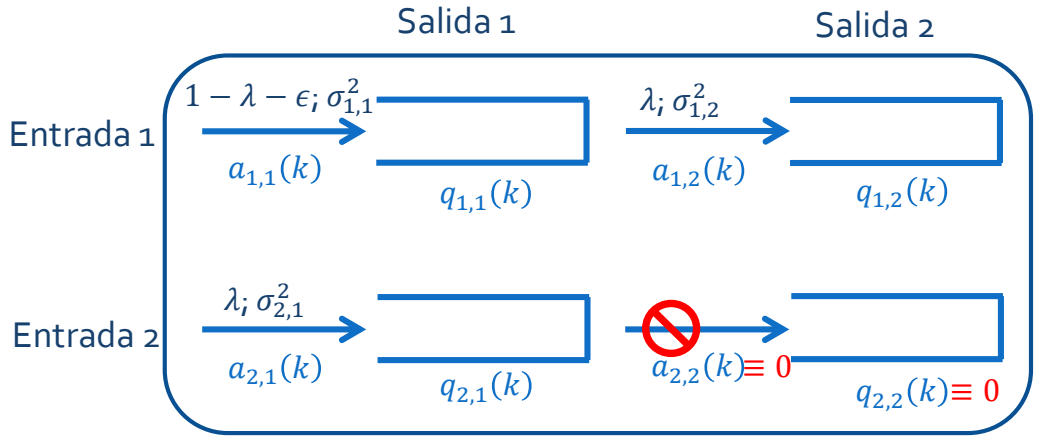
NUEVA VISIÓN DEL MÉTODO *DRIFT* (cont.)

- Igualar $E\left[V_i\left(\boldsymbol{q}(k+1)\right)\right] = E\left[V_i\left(\boldsymbol{q}(k)\right)\right]$ para $i = 1,2,3$:

(1) $$2\lim_{\epsilon\downarrow 0}\epsilon E\left[q_{1,2}\right] = \frac{\sigma_{1,1}^2 + 4\sigma_{1,2}^2 + \sigma_{2,1}^2}{3} - 2\lim_{\epsilon\downarrow 0} E\left[q_{1,2}^+ u_{2,1}\right]$$

(2) $$2\lim_{\epsilon\downarrow 0}\epsilon E\left[q_{2,1}\right] = \frac{\sigma_{1,1}^2 + \sigma_{1,2}^2 + 4\sigma_{2,1}^2}{3} - 2\lim_{\epsilon\downarrow 0} E\left[q_{2,1}^+ u_{1,2}\right]$$

(3) $$\lim_{\epsilon\downarrow 0}\epsilon E\left[q_{1,2} + q_{2,1}\right] = \frac{\sigma_{1,1}^2 - 2\sigma_{1,2}^2 - 2\sigma_{2,1}^2}{3} + 2\lim_{\epsilon\downarrow 0} E\left[q_{1,2}^+ u_{2,1} + q_{2,1}^+ u_{1,2}\right]$$



Teorema [Maguluri et.al. 18]:

$$\lim_{\epsilon\downarrow 0}\epsilon E\left[q_{1,1} + q_{1,2} + q_{2,1}\right] = \frac{2}{3}\left(\sigma_{1,1}^2 + \sigma_{1,2}^2 + \sigma_{2,1}^2\right)$$

Demostración: (1)+(2)+(3) y reorganizar términos

OTRAS COMBINACIONES LINEALES?

4 variables
3 ecuaciones!

Necesitamos
más ecuaciones!