Sistemas Informáticos II

Tema 2: Aspectos operacionales de los sistemas distribuidos: Rendimiento

Antonio E. Martínez, David Vallet, Iván Cantador, Irene Rodríguez, Daniel Hernández (daniel.hernandez@uam.es), Alvaro Ortigosa (alvaro.ortigosa@uam.es), Manuel Sánchez-Montañés (manuel.smontanes@uam.es)

Contenido

- Introducción a los aspectos operacionales de los sistemas distribuidos.
- Rendimiento.
- Teoría de colas.
  - Introducción.
  - Procesos aleatorios y cadenas de Markov.
  - Resolución de algunos modelos de colas: M/M/1, M/M/c, M/M/c/c, M/M/1/K, M/M/1/∞/M, M/M/c/∞/M , M/G/1.
  - Redes de colas.
- Bibliografía especial del tema.

Introducción

- Todo sistema informático tiene dos tipos de requerimientos:
  - **Requerimientos funcionales**:
    - Definen el sistema desde el punto de vista del problema que resuelven.
    - Representan, en general, el **comportamiento lógico** del sistema.
  - **Requerimientos no funcionales**:
    - Definen el **modo** en que opera el sistema.
    - Llamados también ***parámetros de calidad***.
    - Representan, en general, el **comportamiento físico** del sistema.
- En los sistemas distribuidos, los **requerimientos funcionales** reflejan su diseño de modo similar al caso de sistemas basados en un único ordenador.
  - Particularidades específicas que dependen del modelo de sistema elegido.
  - Similares a las que ocurren en el diseño de aplicaciones según distintos paradigmas.
- Los **requerimientos no funcionales**, por el contrario, producen un mayor impacto en el diseño global del sistema por la mayor complejidad física del mismo.

Requerimientos no funcionales

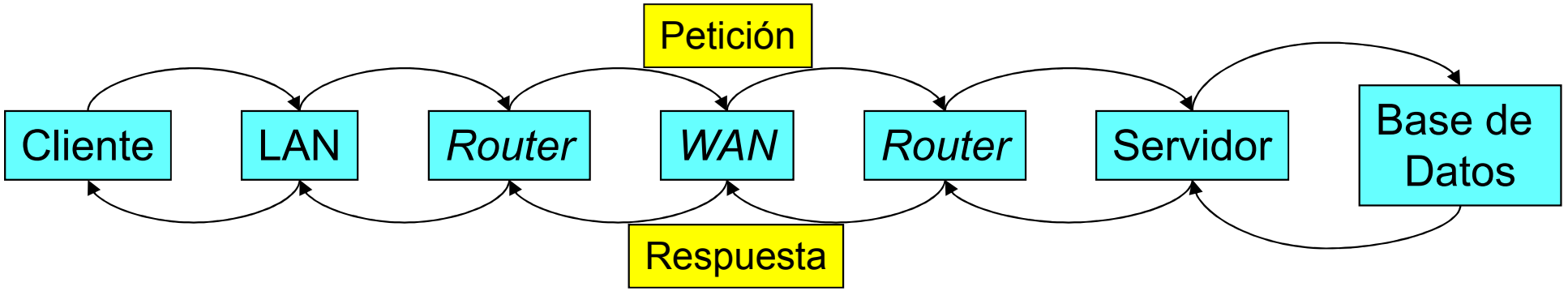
- Se deben considerar desde el punto de vista de arquitectura del sistema como aspectos fundamentales en la elaboración del mismo.
- Se pueden destacar los siguientes aspectos:
  - Rendimiento y capacidad.
  - Disponibilidad, fiabilidad y mantenibilidad.
  - Seguridad, integridad.
  - Gestionabilidad.
  - Usabilidad.
  - Escalabilidad.
- En todo sistema informático que preste un servicio a usuarios finales, los requerimientos no funcionales se plantean bajo la figura de los Acuerdos de Nivel de Servicio (*Service Level Agreements, SLA*).
  - Representan el acuerdo realizado entre el proveedor de un servicio y el usuario calificando los parámetros mínimos aceptables para el servicio.
  - Los sistemas se deben construir con todo lo necesario para cumplir sus SLAs, y no invertir más en intentar excederlos.

Rendimiento

- **Rendimiento *(performance)***: Es el atributo de un sistema informático que caracteriza la correcta disponibilidad temporal (oportunidad, puntualidad - *timeliness*) de los servicios que proporciona el sistema.
- **No es sinónimo de velocidad**. El rendimiento como parámetro de calidad incide más sobre la capacidad de predecir el comportamiento temporal de un sistema en el mayor número posible de situaciones.
- Los objetivos de rendimiento utilizados para especificar y validar un sistema son:
  - **Latencia**: Intervalo de tiempo en el cual se produce la respuesta a un evento.
    - Equivalente a tiempo de respuesta del sistema.
  - **Productividad *(throughput)***: Número de respuestas a eventos que se realizan por unidad de tiempo en un intervalo de observación.
  - **Capacidad**: Medida de la cantidad máxima de trabajo que un sistema o un componente de un sistema es capaz de realizar (productividad máxima).
    - Puede ir limitada por una condición de la latencia máxima permitida.
  - **Modos**: Los sistemas pueden tener distintos requerimientos de rendimiento en función de la fase de ejecución. Ejp: Modo interactivo – modo de proceso por lotes *(batch)*.

Cadena de procesamiento

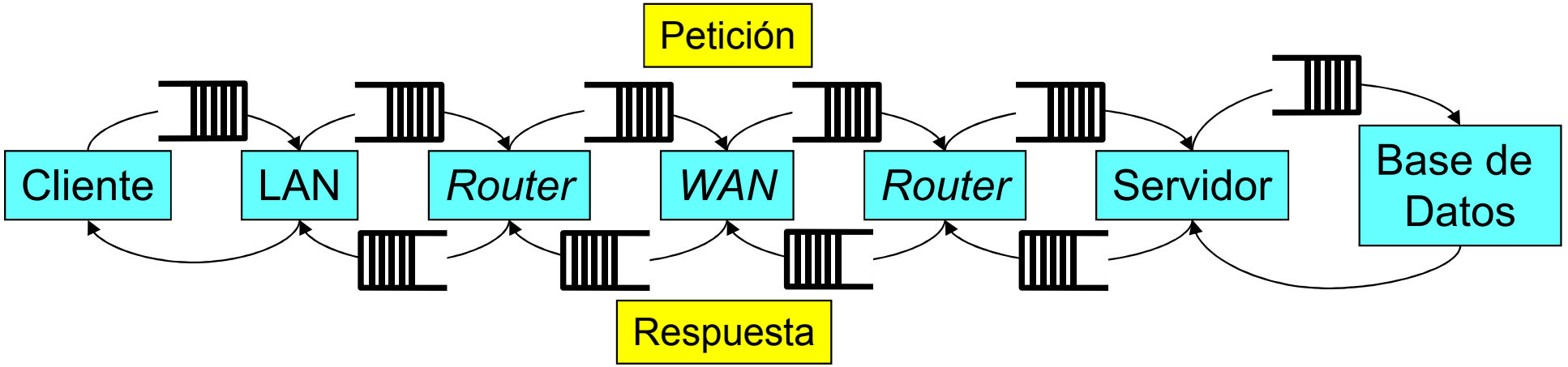
- La ejecución de una transacción en un sistema distribuido es el resultado del trabajo de múltiples elementos en secuencia.



- Estos elementos constituyen la **cadena de procesamiento**.
- Cada elemento en la cadena tiene una **capacidad de procesamiento propia**, que debe ser conocida.
- El rendimiento de un sistema distribuido es una característica extremo a extremo *(end-to-end)*. Se ve afectado por todos los componentes que intervienen en la cadena de procesamiento.
- **El rendimiento total del sistema siempre será peor que el rendimiento del elemento de menor capacidad de procesamiento.**

Recursos compartidos

- Algunos elementos de la cadena de procesamiento son de uso común para múltiples elementos.
- Cuando se desea utilizar uno de estos recursos, es posible que se encuentre ocupado en atender a otra petición. Se produce entonces:
  - Un **rechazo** de la petición. Es necesario reintentar para lograr su ejecución.
  - Una **espera** a que termine de procesar y quede libre para atender a una nueva petición. Es necesario esperar en una **cola**.
- En cualquiera de los casos se produce un retraso adicional en la cadena de procesamiento.



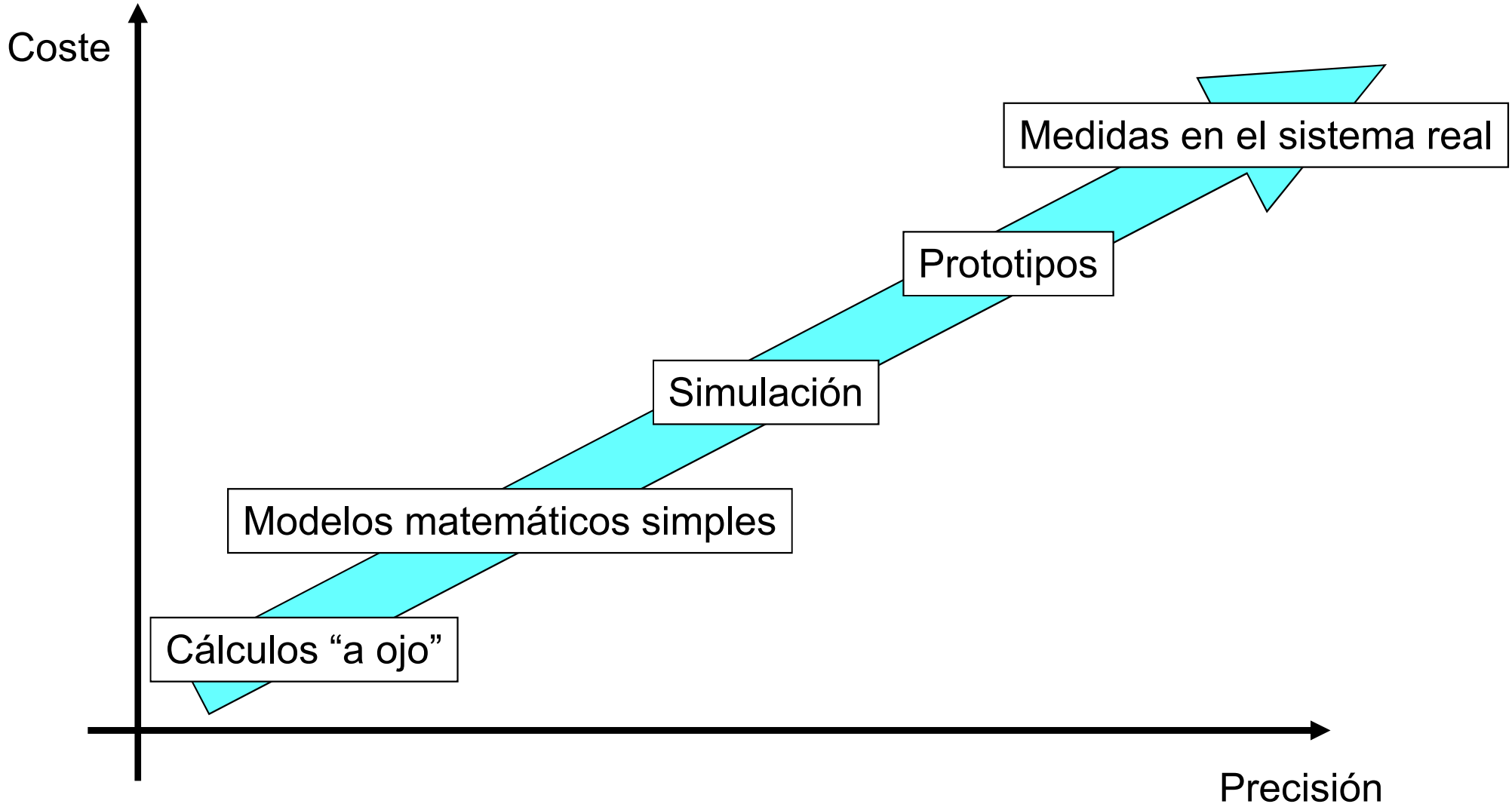
Tiempos de proceso y tiempos de espera

- La cadena de proceso presenta **dos tipos de tiempos de retardo**:
  - **Tiempo de proceso de cada elemento que la compone.**
    - Dependen de la capacidad del elemento.
    - Pueden venir dados por el fabricante.
    - Pueden ser medidos o estimados de forma no muy compleja.
    - Estudios teóricos específicos de cada tipo pueden ayudar a su evaluación
      - Ejp.: Arquitectura de ordenadores, arquitectura de redes de comunicaciones.
      - Acudir a lo visto en las asignaturas correspondientes.
  - **Tiempos de espera ante los recursos compartidos**.
    - Motivados por el acceso múltiple concurrente.
    - Herramienta matemática específica para evaluar sus efectos: **teoría de colas**.

Diseño orientado al rendimiento

1. Establecimiento de los requerimientos no funcionales relativos al rendimiento.
2. Definir la cadena o cadenas de procesamiento.
   - Pueden ser varias dependiendo de la complejidad del sistema o de algunas de sus funciones.
3. Cuantificar los elementos conocidos de la cadena de rendimiento.
   - Número de clientes, velocidad enlaces, capacidad procesadores…
4. Partir de unas especificaciones iniciales de capacidad para los elementos bajo diseño.
5. Realizar una **estimación de rendimiento** del conjunto.
6. Si los resultados de la estimación no dan resultados satisfactorios, alternativas:
   - Variar las especificaciones de capacidad de los elementos bajo diseño.
   - Variar la propia estructura de la cadena de procesamiento.
   - Volver a 3.
7. Si los resultados son satisfactorios, finaliza el diseño.
   - Establecer su validez conforme a la **precisión** de la estimación realizada.

Estimación del rendimiento (I)



Estimación del rendimiento (II)

- **Cálculos "a ojo" *(rules of thumb)***:
  - Estimación basada en conocimientos y experiencias anteriores.
  - Requiere gran experiencia del evaluador.
  - No todos los efectos relacionados con el rendimiento son intuitivos.
  - Poca precisión, bajo coste.
- **Modelos matemáticos simples**:
  - Simplificación del sistema para poder abordarlo con modelos teóricos.
  - Requiere conocimientos teóricos del evaluador.
  - Permite eliminar errores producidos por intuiciones erróneas.
  - Mayor precisión a mayor coste.
- **Simulación: Construcción de un modelo de la cadena de proceso del sistema**.
  - Mediante una herramienta de simulación discreta.
  - Ejecución del modelo en condiciones similares a las de trabajo.
  - Estudio de los resultados y análisis *what-if*.
  - *Garbage-In-Garbage-Out!*

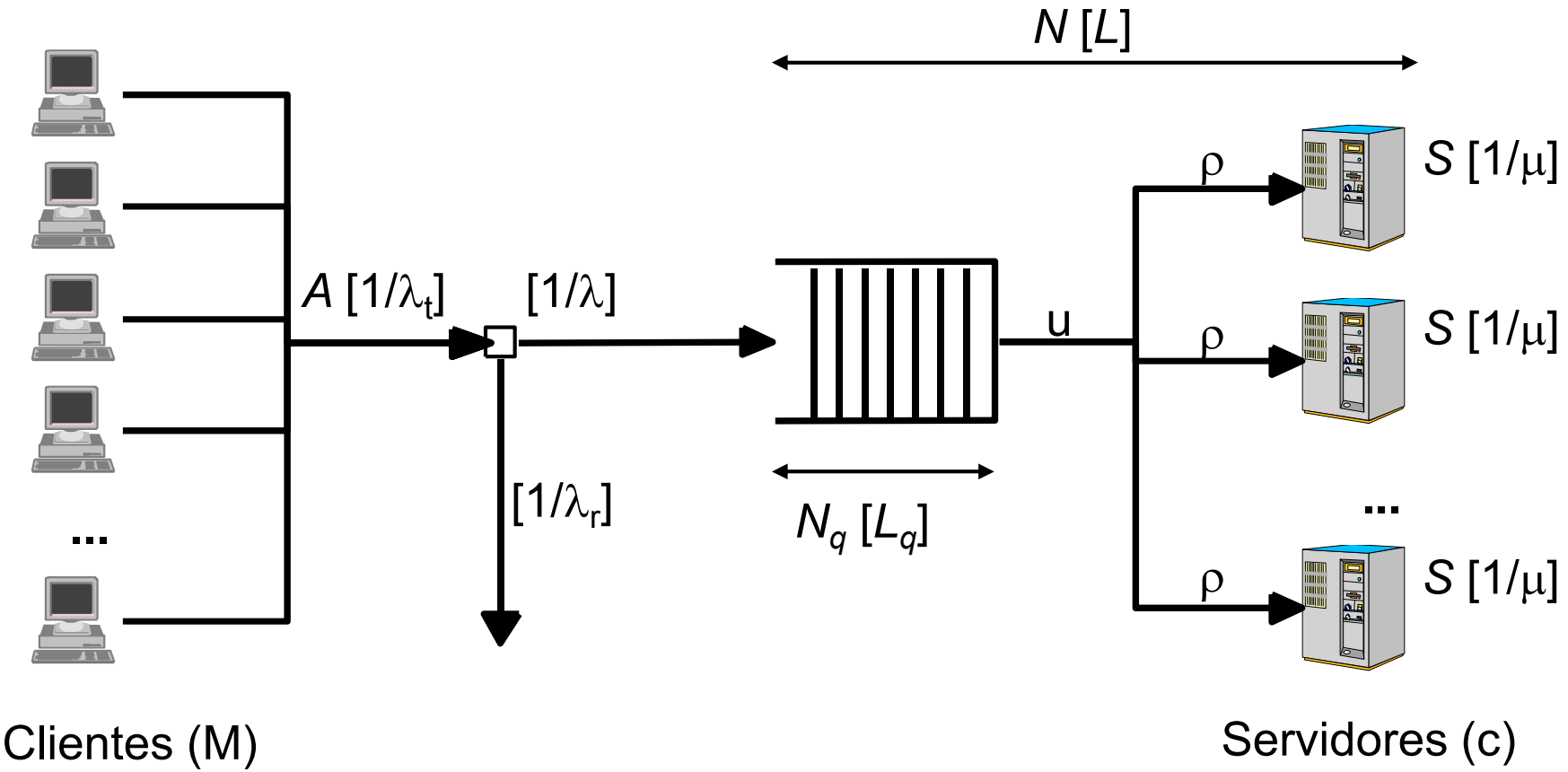
Estimación del rendimiento (III)

- **Prototipos: Versión a escala del sistema final**.
  - Permite detectar problemas ocultos en los análisis realizados.
  - Complejo obtener un prototipo válido para representar el sistema final:
    - Difícil reproducir el número de usuarios real del sistema.
    - Difícil obtener datos reales para su ejecución.
    - Difícil reproducir configuración real del sistema.
    - Es necesario crearlo en paralelo a la realización del producto.
  - Necesario realizar un escalado correcto para obtener valores del sistema final.
- **Medidas en el sistema real**.
  - No suelen ser posibles por estar en construcción. Mecanismo válido para estimar nuevas versiones de un producto existente.
  - Complejidad de la toma de datos sin interferir en su funcionamiento.

Teoría de colas

- La Teoría de Colas es la rama de la **teoría de la probabilidad** que estudia los retrasos producidos por compartición de recursos entre un determinado número de clientes, y establece **modelos** para predecir los retrasos.
- Generalmente hay tres tipos de resultados de interés:
  - Tiempo de espera de un cliente para utilizar un recurso.
  - Número de unidades en espera en un momento dado.
  - Una medida del tiempo en que el recurso está siendo utilizado.
- La naturaleza estocástica del problema hace que estas magnitudes sean casi siempre **variables aleatorias**.
- La teoría de colas proporciona los mecanismos para calcular sus funciones de distribución de probabilidad o, cuando menos, sus **valores esperados**.

Diagrama genérico de un sistema de colas



Definiciones y nomenclatura (I)

- **Cliente**: Solicitud de servicio recibida en un sistema.
  - La llegada de clientes al sistema es un proceso aleatorio.
  - El tiempo entre llegadas consecutivas es una variable aleatoria, *A.*
    - Se denomina *A(t)* a su función de distribución acumulada.
    - Su valor esperado o valor medio es E[*A*]=T_a.
    - El número medio de llegadas al sistema por unidad de tiempo será **$\lambda = 1 / T_a$** , que se denomina **tasa de llegadas**.
- **Servidor**: Elemento del sistema que atiende las solicitudes de servicio (clientes).
  - El tiempo de servicio de un servidor será una variable aleatoria, *S*.
    - Se denomina *S(t)* a su función de distribución acumulada.
    - Su valor esperado o valor medio es E[*S*]=T_s.
    - El número medio de clientes servidos por unidad de tiempo será **$\mu = 1 / T_s$** , que se denomina **tasa de servicio**.
- **Cola**: Elemento intermedio donde esperan los clientes a recibir el servicio.
- Un cliente que entra en el sistema estará siendo servido o esperando en cola.

Definiciones y nomenclatura (II)

- **u: Intensidad de tráfico**. Relación entre la tasa de llegadas y la tasa de servicio.

$$u = \frac{\lambda}{\mu} = \frac{T_s}{T_a} \qquad \textbf{(5.1)}$$

  Su unidad es el *Erlang.*
- **ρ: Factor de utilización del servidor**. Fracción de tiempo en que se encuentra ocupado un servidor.
  - Equivale a la probabilidad de que el servidor esté activo en un instante dado.
- La ocupación del sistema es un proceso aleatorio. El número de clientes en su estado estable es una variable aleatoria, *N*.
  - Su valor esperado, **E[*N*]=*L,*** es el **número medio de clientes en el sistema**.
  - $\boldsymbol{p_n}$ representa la probabilidad de que en el sistema haya *n* unidades.
- La ocupación de la cola del sistema es un proceso aleatorio. El número de clientes en su estado estable es una variable aleatoria, N_q.
  - Su valor esperado, **E[N_q]= L_q**, es el **número medio de clientes en cola**.

Definiciones y nomenclatura (III)

- El tiempo de estancia en el sistema es una variable aleatoria, *T*.
  - *W(t)* es su función de distribución acumulada.
  - Su valor esperado, **E[*T*]=*W***, es el **tiempo medio de estancia en el sistema**.
- El tiempo de espera en cola en el sistema es una variable aleatoria, T_q.
  - $W_q(t)$ es su función de distribución acumulada.
  - Su valor esperado, **E[T_q]= W_q**, es el **tiempo medio de espera en cola**.
- Entre ambos tiempos medios se verifica la siguiente relación:

$$W = W_q + T_s = W_q + 1/\mu \qquad \textbf{(5.2)}$$

- **Teorema de *Little***: Relaciona el número medio de clientes con el tiempo medio de estancia, tanto en el sistema como en cola:

$$L = \lambda W \qquad \textbf{(5.3)} \qquad\qquad L_q = \lambda W_q \qquad \textbf{(5.4)}$$

Por tanto, de (5.2), (5.3) y (5.4) se obtiene que:

$$L = L_q + \lambda T_s = L_q + \lambda/\mu \qquad \textbf{(5.5)}$$

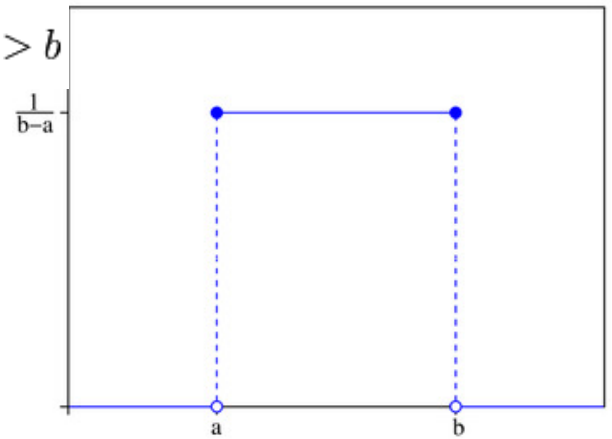
Modelos de sistemas de colas

- Los modelos de colas se caracterizan mediante la notación de Kendall *A/B/c/K/N/Z*, donde:
  - *A*: indica la distribución de probabilidades asumida para el tiempo entre llegadas.
  - B: indica la distribución de probabilidades asumida para el tiempo de servicio.
  - c: indica el número de servidores que contiene el sistema.
  - K: capacidad máxima de clientes que puede contener el sistema.
  - N: Número total de clientes del sistema.
  - Z: Es la disciplina de servicio de la cola.
- Para caracterizar las distribuciones se emplea la siguiente nomenclatura:
  - *M*: Distribución exponencial.
  - *G*: Distribución general (no específica).
  - *D*: Tiempo de servicio constante *(Deterministic)*.
  - H_n: Distribución hiperexponencial de orden n.
  - E_m: Distribución de Erlang-m
- Normalmente, K = ∞, N= ∞ y Z = FCFS (*first-come, first-served), y* se omiten.

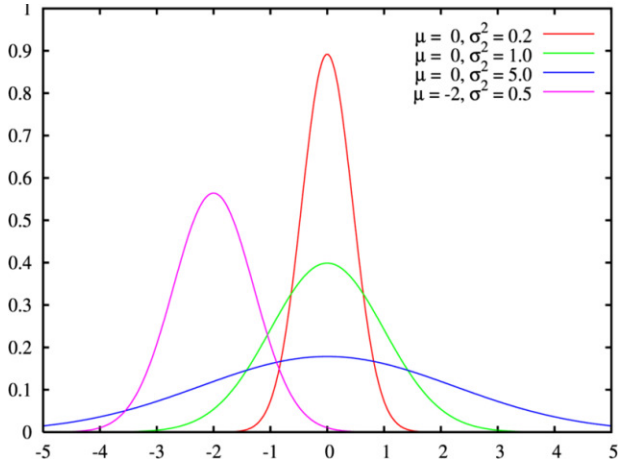
Distribuciones de probabilidad (I)

- **Función de densidad de probabilidad** f(x)
  - Ejemplos: distribuciones uniforme y normal

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{para } a \leq x \leq b \\ 0 & \text{para } x < a \text{ o } x > b \end{cases}$$



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad x \in \mathbb{R}$$



- **Función de distribución de probabilidad** F(x)

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$$

- **Valor esperado** E(x) y **Varianza** Var(x)

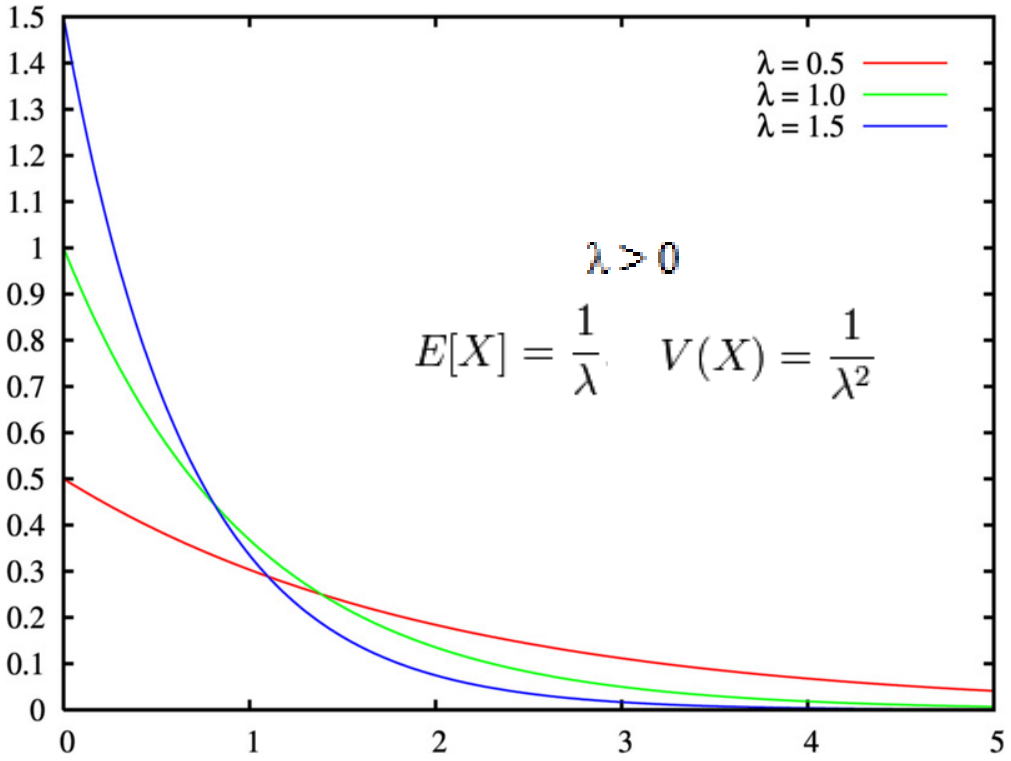
$$E[X] = \mu = \int_{-\infty}^{\infty} x f(x) dx \qquad \operatorname{Var}(X) = \mathrm{E}[(X-\mu)^2] = \int (x-\mu)^2 f(x)\,dx$$

Distribuciones de probabilidad (II)

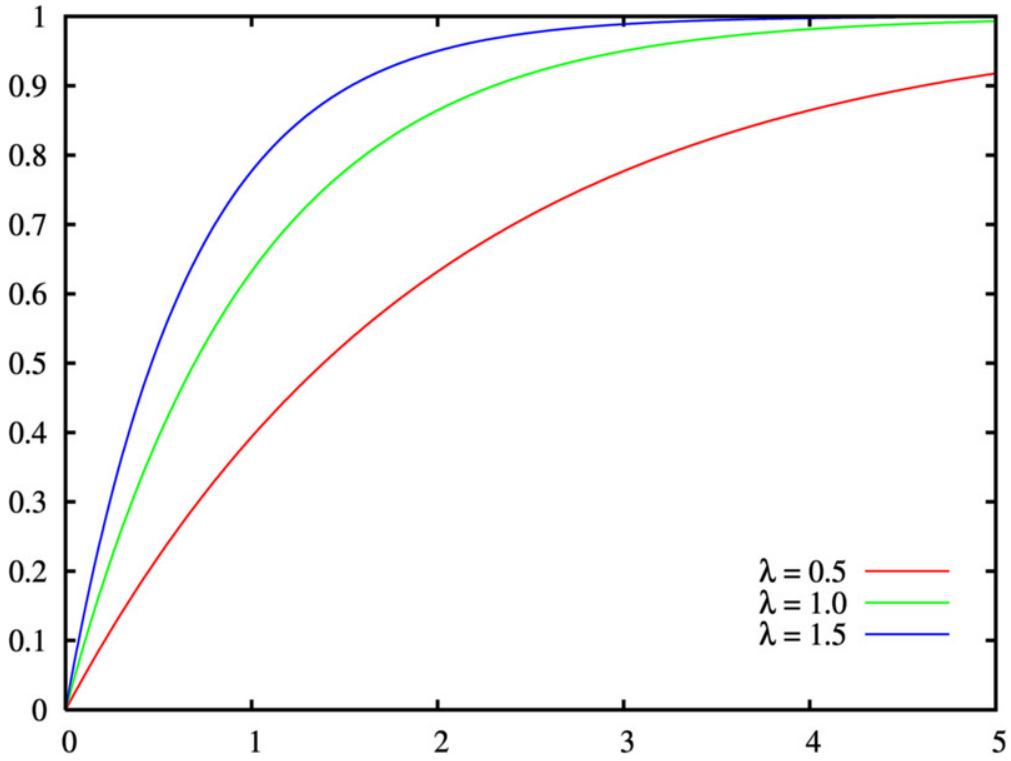
- **Distribución exponencial**

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{para } x \geq 0 \\ 0 & \text{de otro modo} \end{cases}$$

$$F(x) = P(X \leq x) = \begin{cases} 0 & \text{para } x < 0 \\ 1 - e^{-\lambda x} & \text{para } x \geq 0 \end{cases}$$



Función de densidad de probabilidad

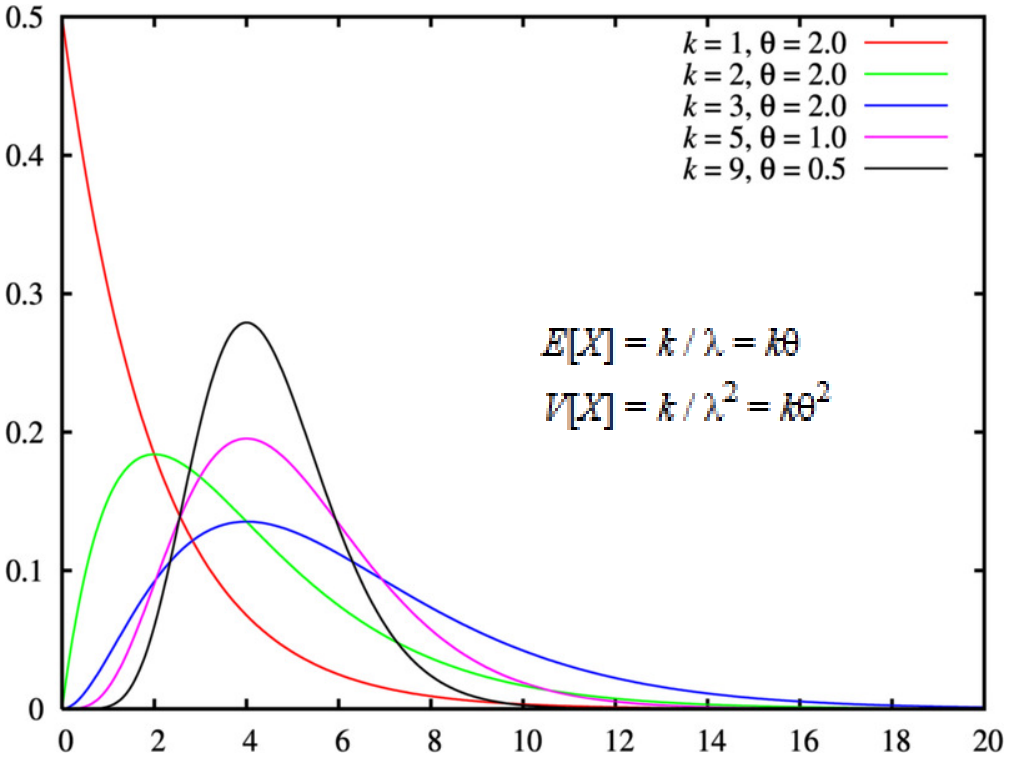


Función de distribución de probabilidad

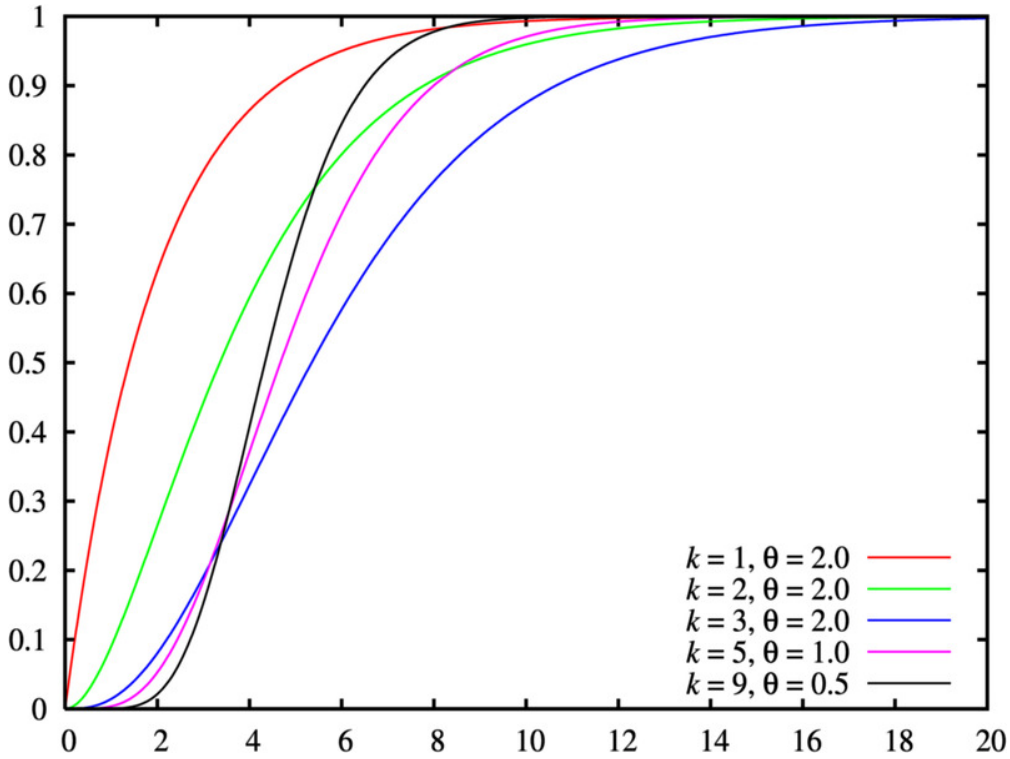
Distribuciones de probabilidad (III)

- **Distribución de Erlang**

$$f(x) = \lambda e^{-\lambda x} \frac{(\lambda x)^{k-1}}{(k-1)!} \quad \text{para } x, \lambda \geq 0$$



Función de densidad de probabilidad

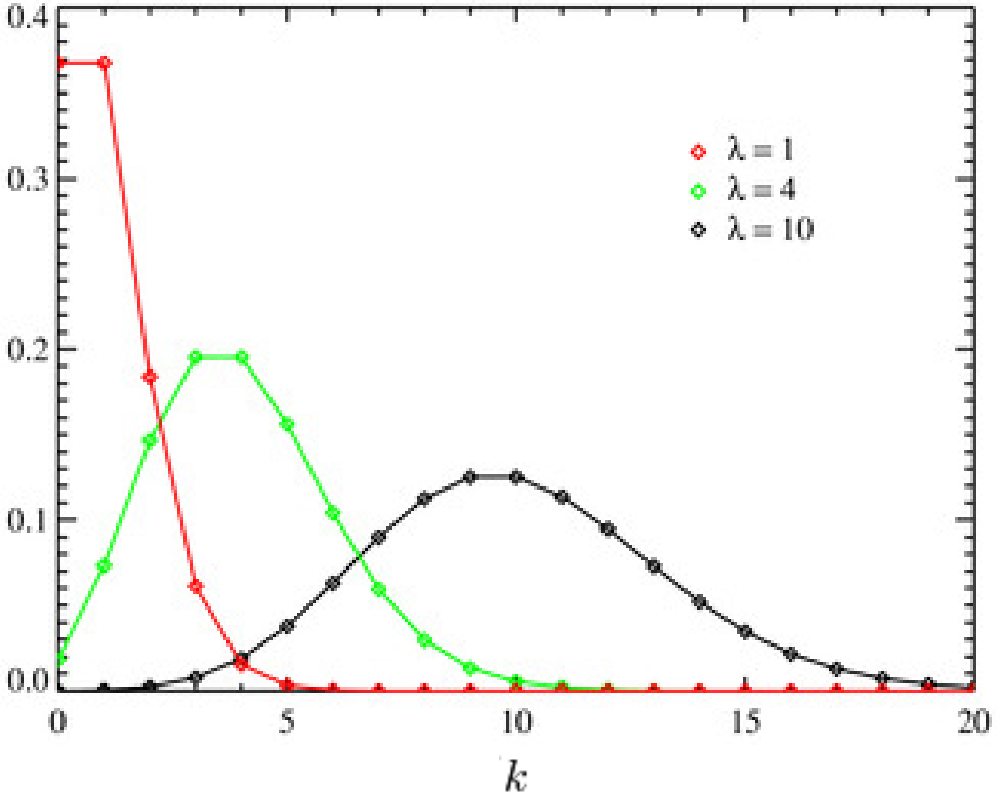


Función de distribución de probabilidad

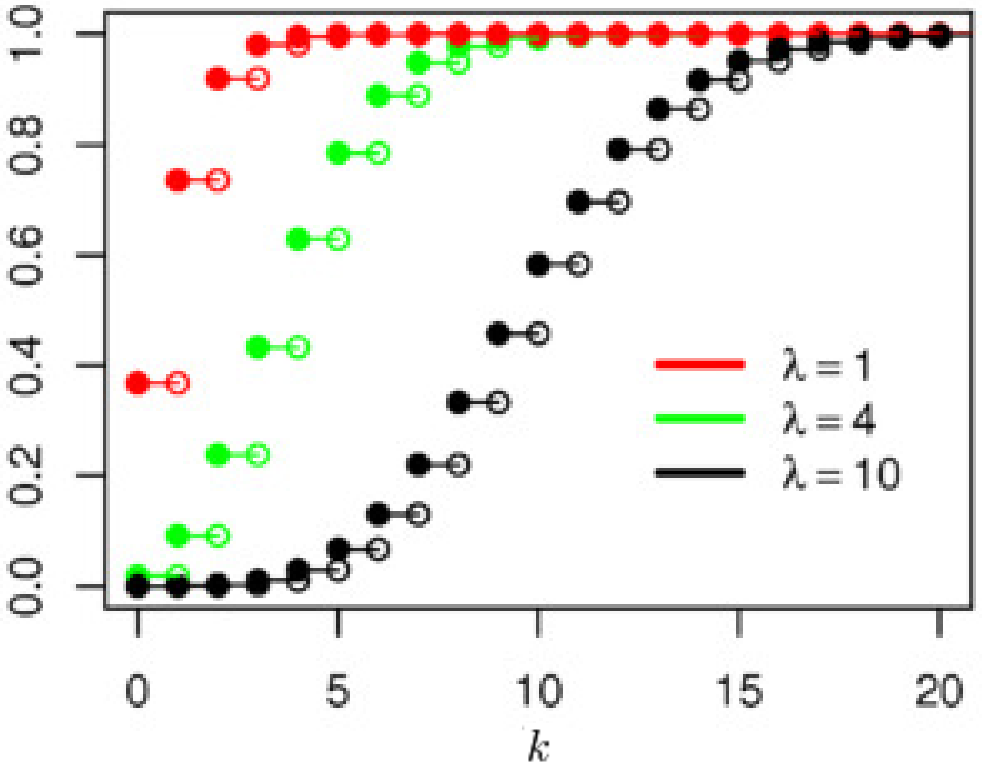
Distribuciones de probabilidad (IV)

- **Distribución de Poisson**

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



Función de masa



Función de distribución de probabilidad

Distribuciones de Probabilidad (V)

- **Mínimo de variables aleatorias exponenciales**: Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes distribuidas de forma exponencial con parámetros $\lambda_1, \lambda_2, \ldots, \lambda_n$. Entonces se cumple que

$$\min(X_1, X_2, \ldots, X_n) \sim \exp\left(\sum_{i=1}^{n} \lambda_i\right)$$

- **Falta de memoria**: Sea X una variable aleatoria continua no negativa y t > 0 y s > 0. Entonces se tiene que

$$P(X > t + s | X > s) = P(X > t) \Leftrightarrow X \sim \exp(\lambda)$$

- **Relación entre la distribución exponencial y la de Poisson:** Sea X una variable aleatoria que denota el número de ocurrencias en un intervalo [0,t]. Si X sigue una distribución de Poisson con parámetro λt (λ es el número medio de ocurrencias por unidad del intervalo), entonces se cumple que el tiempo entre ocurrencias está distribuido de forma exponencial con parámetro λ, y viceversa.

Procesos aleatorios o estocásticos

- Una **variable aleatoria** x es una función que asigna a cada resultado ζ de un experimento aleatorio S un número, $x(\zeta)$.
- Un **proceso aleatorio** $x(t)$, $t \in T$, es una función que asigna a cada resultado ζ de un experimento aleatorio S una función, $x(t, \zeta)$. Habitualmente, T representa el tiempo, y el proceso aleatorio es una familia de funciones temporales con parámetro ζ.
  - Si se fija ζ, $x(t)$ es una función del tiempo, llamada ***muestra* o *realización*** del proceso aleatorio.
  - Si se fija t y ζ es variable, $x(\zeta)$ es una variable aleatoria denominada el ***estado*** del proceso aleatorio en el instante t.
- Tipos de procesos aleatorios:
  - Parámetro discreto (*discrete-parameter* o *discrete-time*): T es un conjunto numerable.
  - Parámetro continuo (*continous-parameter* o *continous-time*): $T \subset \Re$
  - Estado discreto (*discrete-state*): El conjunto de estados es numerable, $\{e_1, e_2, e_3 \ldots e_j \ldots\}$
  - Estado continuo (*continous-state*): Estados toman valor en un conjunto continuo.

Cadenas de Markov (I)

- Un proceso aleatorio *x(t)* es un proceso de Markov si en cualquier instante t_n el estado del proceso $x(t_n)$ depende sólo del estado en el instante anterior, $x(t_{n-1})$.
  - El estado del proceso en el futuro es independiente de su pasado. El sistema carece de *memoria (memoryless)*.
- Un proceso de Markov de estado discreto se llama *Cadena de Markov*. Su comportamiento se caracteriza por:
  - Las probabilidades de que el proceso se encuentre en un estado e_i en un instante *t*.

$$p_i(t) = P\{x(t) = e_i\} \quad \textbf{(5.6)}$$

  - Las probabilidades de transición: Probabilidad de que el sistema pase al estado e_j en $t_0 + t$ si estaba en estado e_i en $t = t_0$.

$$p_{ij}(t) = P\{x(t_0 + t) = e_j \,|\, x(t_0) = e_i\} \quad \textbf{(5.7)}$$

    - Por tanto, la probabilidad de que no haya transición será:

$$p_{ii}(t) = 1 - \sum_{j \neq i} p_{ij}(t) \quad \textbf{(5.8)}$$

Cadenas de Markov (II)

- Si esta probabilidad no depende del instante t_0 sino solo del tiempo transcurrido t la cadena se dice que es ***homogénea en el tiempo***:

$$p_{ij}(t) = P\left\{ x(t+t_0) = e_j \middle| x(t_0) = e_i \right\} \qquad \forall t_0 > 0 \tag{5.9}$$

- Por el teorema de la probabilidad total,

$$p_j(t+t_0) = \sum_i P\left\{ x(t+t_0) = e_j \middle| x(t_0) = e_i \right\} \cdot P\left\{ x(t_0) = e_i \right\} = \sum_i p_i(t_0) p_{ij}(t) \tag{5.10}$$

- La probabilidad de estar en tiempo $t + t_0$ en el estado j es igual a la suma de la probabilidad de estar en cada estado i en tiempo t_0 y pasar a j en t segundos.

Cadenas de Markov (III)

- Nos centraremos en analizar el sistema cuando la cadena de Markov está en ***equilibrio o estado estable.*** Es decir, cuando después de un tiempo suficiente ($t \to \infty$), la probabilidad de encontrarse en el estado j no varía y es igual a p_j:

$$p_j = \lim_{t \to \infty} p_j(t)$$

- Asumiendo que la cadena es homogénea en el tiempo y haciendo $t_0 \to \infty$ (cadena en estado de equilibrio o estable) en la ecuación **5.10** se tiene:

$$p_j = \sum_i p_i \, p_{ij}(t) \qquad \textbf{(5.11)}$$

$$p_j(t + t_0) \xrightarrow{t_0 \to \infty} p_j \qquad p_i(t_0) \xrightarrow{t_0 \to \infty} p_i$$

$$p_j(t + t_0) = \sum_i P\{x(t + t_0) = e_j \mid x(t_0) = e_i\} = \sum_i p_i(t_0) \, p_{ij}(t)$$

Cadenas de Markov (IV)

- En toda cadena de Markov de parámetro continuo se verifica que el tiempo que se tarda en realizar un cambio de estado, τ , es una variable aleatoria con distribución exponencial. Según esto (para t pequeño):

$$p_{ij}(t) = P\{\tau \le t\} = 1 - e^{-\lambda_{ij} t} = \lambda_{ij} t + o(t) \qquad \textbf{(5.12)}$$

  o(t) representa términos de grado superior a 2 de t y λ_{ij} el número medio de transiciones por unidad de tiempo ***(tasa de transiciones)*** entre los estados e_i y e_j.

- Derivando (5.11) respecto a *t,* teniendo en cuenta la expresión (5.12), y considerando t muy pequeño se obtienen las **ecuaciones estacionarias de la Cadena de Markov**:

  En donde:

$$\sum_i p_i \lambda_{ij} = 0 \quad (\forall j) \qquad \textbf{(5.13)}$$

$$\lambda_{ij} = \frac{dp_{ij}(t)}{dt} \quad (i \neq j) \qquad \textbf{(5.14)}$$

$$\lambda_{jj} = \frac{dp_{jj}(t)}{dt} = -\sum_{i \neq j} \lambda_{ji} \qquad \textbf{(5.15)}$$

Cadenas de Markov (V)

- La expresión (5.13) representa un sistema con tantas ecuaciones y variables como estados. Desarrollando la ecuación para el estado *j* se tiene:

$$p_0\lambda_{0j} + p_1\lambda_{1j} + \ldots + p_{j-1}\lambda_{j-1j} + p_j\lambda_{jj} + p_{j+1}\lambda_{j+1j} + \ldots + p_n\lambda_{nj} + \ldots = 0 \quad \textbf{(5.13a)}$$

- Considerando (5.15), y pasando este término negativo al segundo miembro:

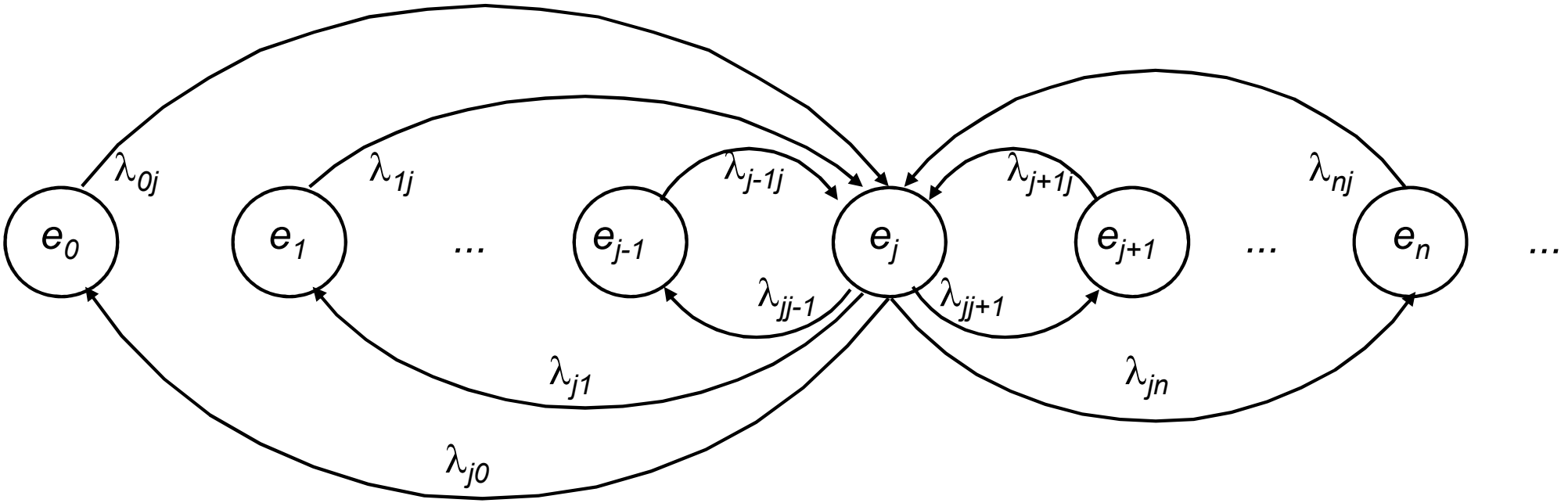
$$p_0\lambda_{0j} + p_1\lambda_{1j} + \ldots + p_{j-1}\lambda_{j-1j} + p_{j+1}\lambda_{j+1j} + \ldots p_n\lambda_{nj} = p_j\left(\lambda_{j0} + \lambda_{j1} + \ldots + \lambda_{jj-1} + \lambda_{jj+1} + \ldots + \lambda_{jn}\right) \quad \textbf{(5.13b)}$$

Por su forma, estas ecuaciones se conocen como **ecuaciones de balance global**.

- El primer término representa la suma de los productos de las tasas de transición desde todos los estados **hacia el estado *j*** por la probabilidad del estado origen de la transición.
- El segundo término representa la suma de los productos de las tasas de transición **desde el estado *j*** hacia todos los estados por la probabilidad del estado *j*.

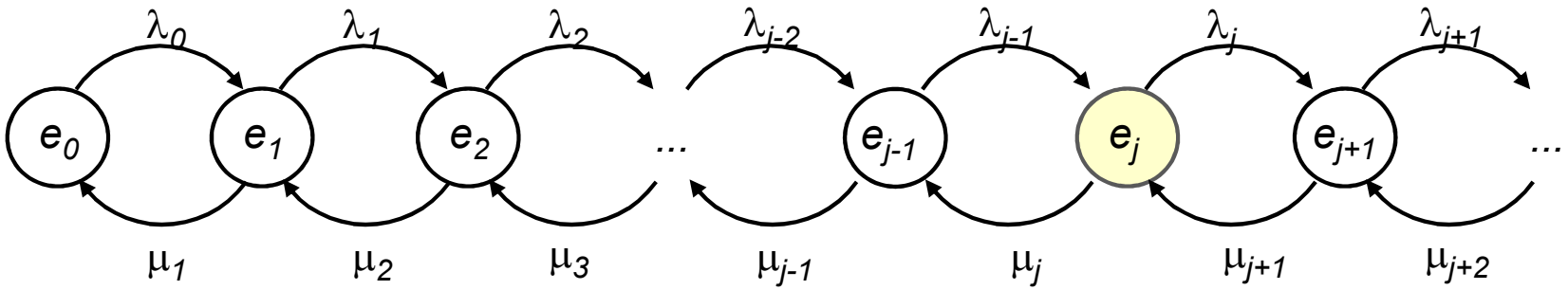
Cadenas de Markov (VI)

- La cadena de Markov de parámetro continuo se representa mediante un grafo en el que:
  - Los nodos representan cada estado posible de la cadena.
  - Los arcos representan las transiciones posibles de cada estado. Se numeran con la tasa de transiciones del estado origen al estado destino.
- Ejemplo: Para el estado *j* representado en (5.13a):



Proceso nacimiento-muerte (I)

- Ejemplo de proceso de Markov. Estado: número de individuos.
  - En un estado e_i, la tasa de nacimientos es λ_i y la tasa de defunciones, μ_i.
  - No se producen nacimientos ni muertes simultáneos.
  - El diagrama de transiciones de estados es de la siguiente forma:



  - En este modelo se verifica, por tanto:

$$\lambda_{jk}\begin{cases}\lambda_j & (k=j+1)\\ \mu_j & (k=j-1,\ j>0)\\ 0 & (k\neq j,\ j+1,\ j-1)\end{cases} \quad \textbf{(5.16)}$$

$$\lambda_{j,j+1}=\lambda_j \qquad \lambda_{j,j-1}=\mu_j \qquad \lambda_{jj}=-\left(\lambda_{j,j-1}+\lambda_{j,j+1}\right)=-\left(\lambda_j+\mu_j\right) \quad \textbf{(5.17)}$$

Proceso nacimiento-muerte (II)

- Eliminando los términos que en este caso son nulos en (**5.13**) se obtiene:

$$\begin{cases} \lambda_{j-1,j} p_{j-1} - \left(\lambda_{j,j-1} + \lambda_{j,j+1}\right) p_j + \lambda_{j+1,j} p_{j+1} = 0 \quad (j > 0) \\ -\lambda_{0,1} p_0 + \lambda_{1,0} p_1 = 0 \end{cases} \qquad \textbf{(5.18)}$$

- Introduciendo (5.16) y (5.17) en (5.18):

$$\begin{cases} \lambda_{j-1} p_{j-1} - \left(\lambda_j + \mu_j\right) p_j + \mu_{j+1} p_{j+1} = 0 \quad (j > 0) \\ -\lambda_0 p_0 + \mu_1 p_1 = 0 \end{cases} \qquad \textbf{(5.19)}$$

- Reescribiendo las ecuaciones se obtiene la siguiente identidad iterativa:

$$\mu_{j+1} p_{j+1} - \lambda_j p_j = \mu_j p_j - \lambda_{j-1} p_{j-1} = \ldots = \mu_1 p_1 - \lambda_0 p_0 = 0 \qquad \textbf{(5.19)}$$

Proceso nacimiento-muerte (III)

- Despejando y sustituyendo recursivamente se llega a la expresión:

$$p_{j+1} = \frac{\lambda_j}{\mu_{j+1}} p_j = \frac{\lambda_j \lambda_{j-1}}{\mu_{j+1}\mu_j} p_{j-1} = \ldots = \frac{\lambda_j \lambda_{j-1} \ldots \lambda_0}{\mu_{j+1}\mu_j \ldots \mu_1} p_0 \qquad \textbf{(5.20)}$$

$$p_n = \prod_{k=0}^{n-1} \frac{\lambda_k}{\mu_{k+1}} p_0 \quad (n > 0) \qquad \textbf{(5.21)}$$

- p_0 se calcula para que se cumpla el segundo axioma de la probabilidad:

$$\sum_{n=0}^{\infty} p_n = 1 \Rightarrow \left(1 + \sum_{n=1}^{\infty} \prod_{k=0}^{n-1} \frac{\lambda_k}{\mu_{k+1}} \right) p_0 = 1 \qquad \textbf{(5.22)}$$

- Por lo tanto, la solución estacionaria de esta cadena de Markov existe si la serie infinita de la expresión (5.22) converge.

Procesos de Poisson (I)

- Un proceso de Poisson es un proceso de Markov que sólo puede cambiar desde el estado e_i al estado e_{i+1} con una probabilidad que es independiente del estado en que se encuentre.
- Representan sucesos que ocurren en instantes aleatorios con las siguientes condiciones:
  - El número de ocurrencias es independiente del tiempo (no existen horas punta).
  - Una nueva ocurrencia del suceso es independiente de sucesos anteriores.
  - La probabilidad de dos ocurrencias simultáneas se puede considerar nula.
- En estas condiciones, si los sucesos ocurren a un ritmo de **λ s^{-1}**, la **probabilidad de *n* llegadas en *t* segundos** se expresa por:

$$p_n(t) = P\{x(t) = n\} = \frac{(\lambda t)^n}{n!} e^{-\lambda t};\ \lambda > 0,\ n = 0,1,2... \qquad \textbf{(5.23)}$$

Proceso de nacimiento puro con todas las tasas de nacimiento iguales a λ

Procesos de *Poisson* (II)

- El número medio de llegadas en un intervalo *t* y su varianza vienen dados por:

$$\mathrm{E}\left[x(t)\right] = \lambda t \quad \mathrm{Var}\left[x(t)\right] = \lambda t \tag{5.24}$$

- La probabilidad de que el tiempo entre llegadas sea menor que *t* sigue una **distribución exponencial de parámetro λ**, dada por la expresión:

$$A(t) = P\left\{A \le t\right\} = 1 - e^{-\lambda t} \tag{5.25}$$

- La función densidad de probabilidad del tiempo entre llegadas será:

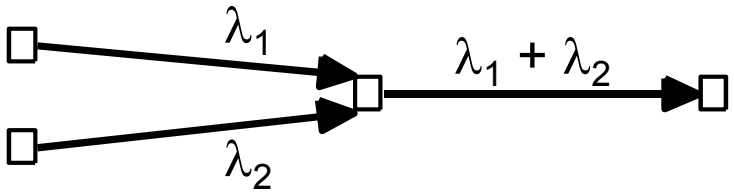
$$a(t) = \lambda e^{-\lambda t} \tag{5.26}$$

- El valor medio del tiempo entre llegadas es:

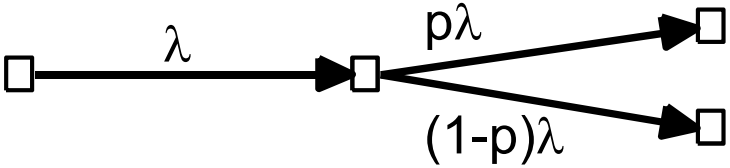
$$\mathrm{E}\left[A\right] = \frac{1}{\lambda} \quad \mathrm{Var}\left[A\right] = \frac{1}{\lambda^2} \tag{5.27}$$

Procesos de *Poisson* (III)

- Propiedades:
  - La suma de varios procesos de *Poisson* independientes es un proceso de *Poisson*, con tasa de ocurrencias igual a la suma de las de los procesos componentes



  - Los procesos resultantes de la partición de un proceso de *Poisson* asignando cada llegada a los procesos resultantes de modo aleatorio (independiente de las asignaciones previas) son procesos de *Poisson*.
    - La tasa de ocurrencias de cada uno es proporcional a la probabilidad de reparto. Por ejemplo, para una partición en dos procesos con probabilidad p y 1-p:



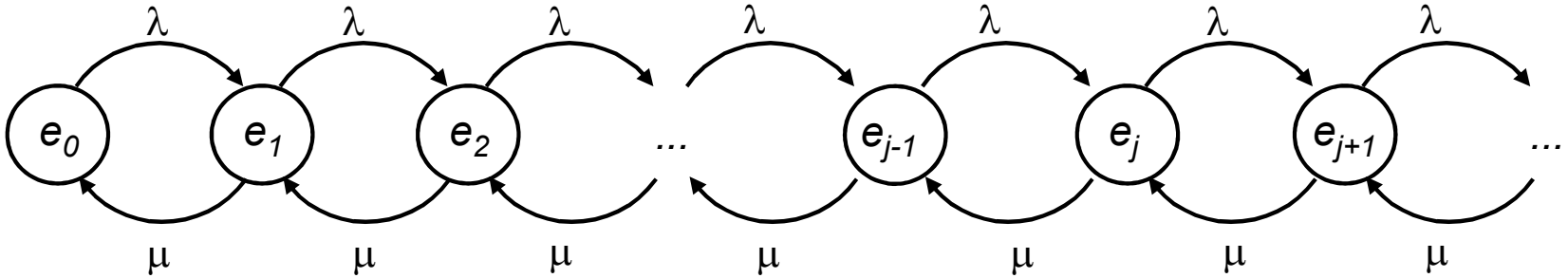
    - La partición con asignaciones fijas no es un proceso de *Poisson*.

Modelo *M/M/1* (I)

- Las llegadas al sistema siguen un proceso de Poisson con tasa de llegadas λ → tiempo entre llegadas sigue una distribución exponencial.
- El tiempo de servicio está distribuido de modo exponencial con media 1/μ.
- Hay un único servidor para atender las peticiones.
- En estas condiciones, el número de unidades en el sistema es una cadena de Markov, equivalente al proceso nacimiento – muerte, en la que:

$$\lambda_j = \lambda \quad \mu_j = \mu \quad (\forall j) \qquad \textbf{(5.28)}$$

- El diagrama de transiciones de estados será:



- Sustituyendo (5.28) en (5.21) se obtiene la distribución de probabilidad del número de unidades en el sistema:

$$p_n = p_0(\lambda/\mu)^n \qquad \textbf{(5.29)}$$

Modelo *M/M/1* (II)

- p_0 (probabilidad de 0 clientes, sistema inactivo) se puede calcular sabiendo que la suma de todas las probabilidades debe ser 1 (*segundo axioma probabilidad*).

$$\sum_{n=0}^{\infty} p_n = \sum_{n=0}^{\infty} p_0 \left(\lambda/\mu\right)^n = 1 \qquad \textbf{(5.30)}$$

Es una serie geométrica, que será convergente si su razón, λ/μ, es menor que 1:

$$\lambda/\mu < 1 \Rightarrow \sum_{n=0}^{\infty} p_0 \left(\lambda/\mu\right)^n = \frac{p_0}{1-\lambda/\mu} = 1 \Rightarrow p_0 = 1-\lambda/\mu \qquad \textbf{(5.31)}$$

El factor de utilización del servidor será:

$$\rho = 1 - p_0 = \lambda/\mu \qquad \textbf{(5.32)}$$

Sustituyendo (5.32) en (5.29) se obtiene:

$$p_n = \left(1-\rho\right)\left(\rho\right)^n \qquad \textbf{(5.33)}$$

Modelo *M/M/1* (III)

- El número medio de unidades en el sistema será el valor medio de (5.33):

$$L = \mathrm{E}[N] = \sum_{n=0}^{\infty} n p_n = \sum_{n=0}^{\infty} n(1-\rho)\rho^n = (1-\rho)\sum_{n=0}^{\infty} n\rho^n \qquad \textbf{(5.34)}$$

Considerando que:

$$\sum_{n=0}^{\infty} n\rho^n = \rho\sum_{n=1}^{\infty} n\rho^{n-1} = \rho\frac{\partial}{\partial\rho}\sum_{n=1}^{\infty}\rho^n = \rho\frac{\partial}{\partial\rho}\left(\frac{\rho}{1-\rho}\right) = \rho\frac{1}{(1-\rho)^2} \qquad \textbf{(5.35)}$$

Sustituyendo (5.35) en (5.34):

$$L = \rho(1-\rho)\frac{1}{(1-\rho)^2} = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda} \qquad \textbf{(5.36)}$$

Modelo *M/M/1* (IV)

- El tiempo medio de estancia en el sistema se calcula aplicando el Teorema de Little (5.3)

$$W = \frac{L}{\lambda} = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu - \lambda} \qquad \textbf{(5.37)}$$

- Para el tiempo medio de espera en cola, aplicando (5.2):

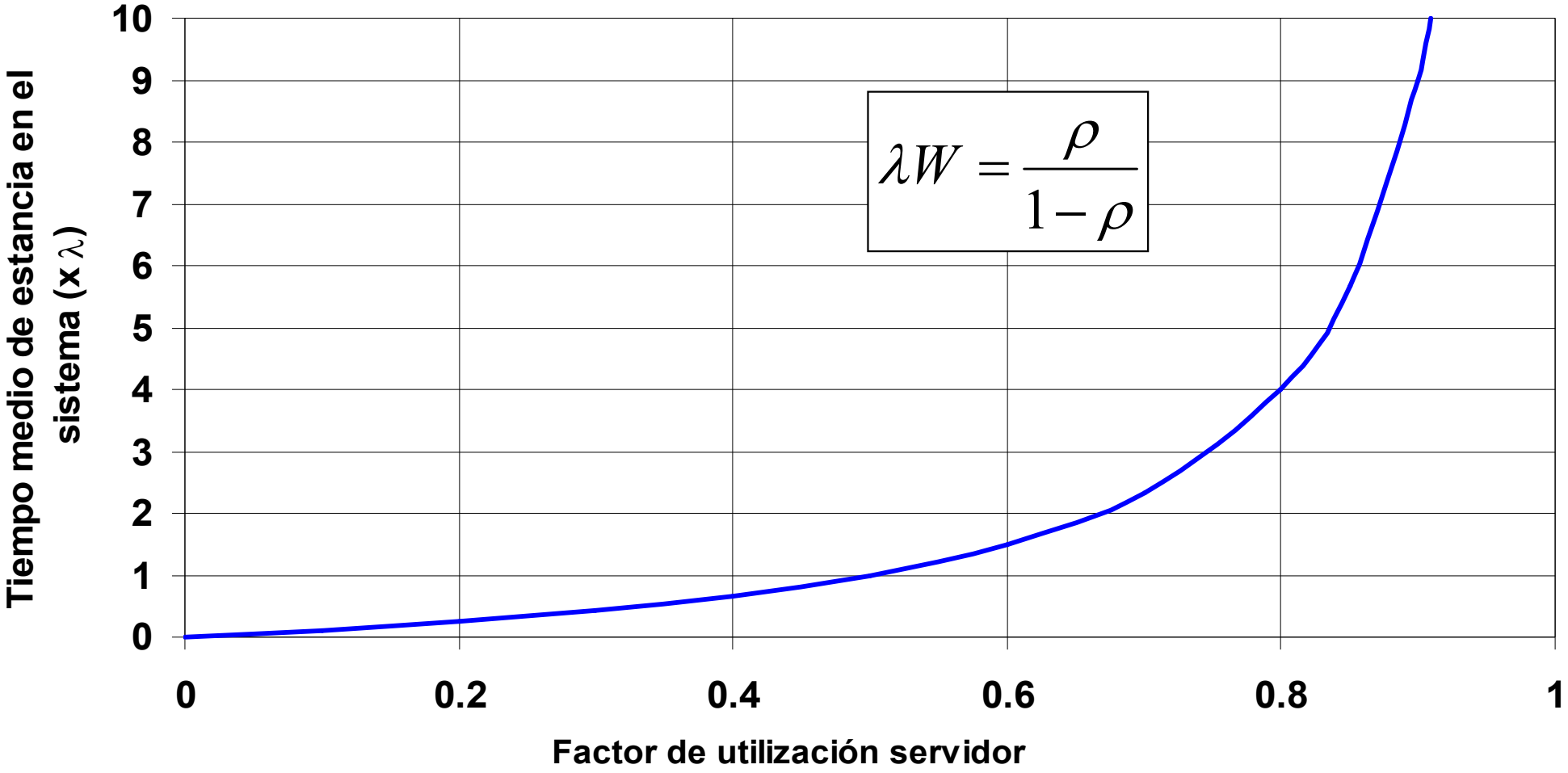
$$W_q = W - \frac{1}{\mu} = \frac{\rho}{\mu(1-\rho)} = \frac{\lambda}{\mu(\mu-\lambda)} \qquad \textbf{(5.38)}$$

- La ocupación media de la cola, aplicando el Teorema de Little (5.4):

$$L_q = \lambda W_q = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)} \qquad \textbf{(5.39)}$$

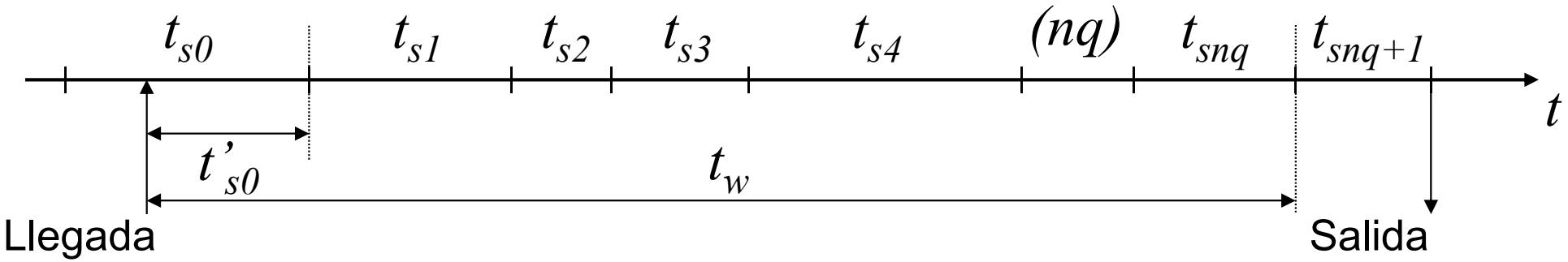
Modelo *M/M/1* (V)

Representación gráfica del tiempo medio de estancia en el sistema



Modelo *M/M/1* (VI)

- El tiempo de estancia en el sistema para un cliente se puede calcular considerando que es la suma de los siguientes tiempos:
  - El tiempo de servicio residual del cliente que está siendo atendido al producirse la llegada.
  - El tiempo de servicio de los *nq* clientes en cola al producirse la llegada.
  - El tiempo de servicio propio.



$$t_{Wn} = t'_{s0} + t_{s1} + t_{s2} + \ldots + t_{snq} + t_{snq+1} \qquad \textbf{(5.40)}$$

A partir de esta expresión se puede obtener la función de distribución para el tiempo de estancia en el sistema

$$W(t) = 1 - e^{-(\mu-\lambda)t} \qquad \textbf{(5.41)}$$

Modelo *M/M/1* (VII)

- A partir de esta expresión se pueden calcular los percentiles del tiempo de estancia en el sistema. Para un percentil p:

$$p = 1 - e^{-(\mu-\lambda)t} \Rightarrow t = \frac{1}{\mu-\lambda}\ln\frac{1}{1-p} = -W\ln(1-p) \qquad \textbf{(5.42)}$$

La siguiente tabla muestra algunos valores de los percentiles:

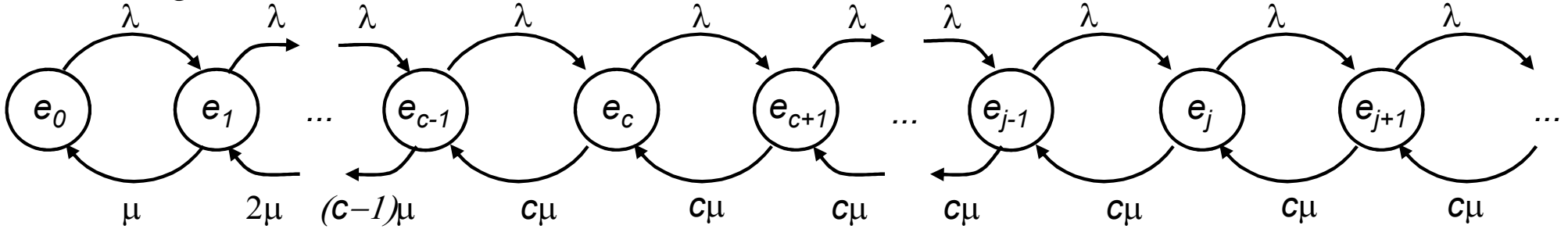
p	t
.63	W
.70	1.20W
.80	1.61W
.85	1.90W
.90	2.30W
.95	3.00W
.99	4.61W

Modelo *M/M/c* (I)

- Las llegadas al sistema siguen un proceso de Poisson con tasa de llegadas λ → tiempo entre llegadas sigue una distribución exponencial.
- El tiempo de servicio de cada servidor está distribuido de modo exponencial con media 1/μ.
- Hay *c* servidores para atender peticiones. El tráfico se reparte por igual entre todos los servidores.
- En estas condiciones, el número de unidades en el sistema es una cadena de Markov, equivalente al proceso nacimiento – muerte, en la que:

$$\lambda_j = \lambda \quad \forall j \quad \mu_j = \begin{cases} j\mu & (j < c) \\ c\mu & (j \geq c) \end{cases} \qquad \textbf{(5.43)}$$

- El diagrama de transiciones de estados será:



Modelo *M/M/c* (II)

- Sustituyendo (5.43) en (5.21) se obtiene la función de distribución de probabilidad del número de unidades en el sistema:

$$p_n = \begin{cases} p_0 \dfrac{(\lambda/\mu)^n}{n!} & (n < c) \\ p_0 \dfrac{c^c}{c!}\left(\dfrac{\lambda}{c\mu}\right)^n & (n \geq c) \end{cases} \qquad \textbf{(5.44)}$$

- p_0 se calcula aplicando el segundo axioma de la probabilidad:

$$\sum_{n=0}^{\infty} p_n = 1 \Rightarrow p_0 = \left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \sum_{n=c}^{\infty} \frac{c^c}{c!}\left(\frac{\lambda}{c\mu}\right)^n \right]^{-1} \qquad \textbf{(5.45)}$$

El segundo sumando es una serie geométrica de razón $\lambda/c\mu$, que converge si su razón es menor que 1.

Modelo *M/M/c* (III)

En este caso, por tanto

$$\frac{\lambda}{c\mu} < 1 \Rightarrow p_0 = \left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!(1-\rho)} \right]^{-1} \quad \textbf{(5.46)}$$

- El factor de utilización de cada servidor será una fracción 1/c del tráfico total:

$$\rho = u/c = \lambda/c\mu \quad \textbf{(5.47)}$$

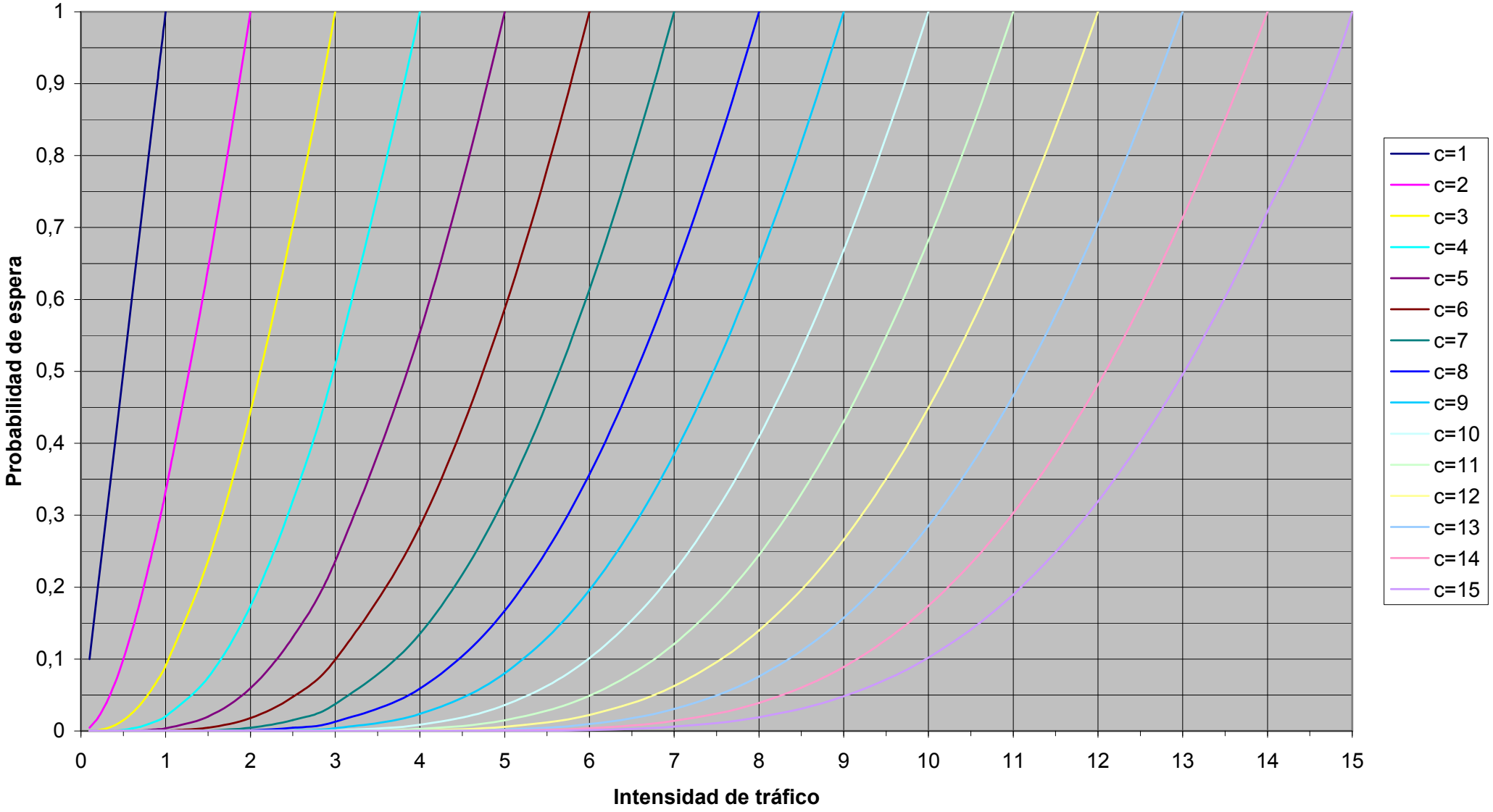
- La probabilidad de que al llegar un cliente tenga que esperar en cola es:

$$P_q = P\{N(t) \geq c\} = \sum_{n=c}^{\infty} p_n = \sum_{n=c}^{\infty} p_0 \frac{c^c}{c!} \left(\frac{\lambda}{c\mu} \right)^n = p_0 \frac{c^c}{c!} \left(\frac{\lambda}{c\mu} \right)^c \sum_{n=c}^{\infty} \left(\frac{\lambda}{c\mu} \right)^{n-c} =$$

$$= p_0 \frac{c^c}{c!} \left(\frac{\lambda}{c\mu} \right)^c \frac{1}{1-\lambda/c\mu} = \frac{p_c}{1-\rho} = E_c(c,u) \quad \textbf{(5. 48)}$$

conocida como *Fórmula C de Erlang* o *fórmula de Erlang de la llamada retardada.*

Representación gráfica de la función Erlang-C



Modelo *M/M/c* (IV)

- El número medio de clientes en cola será el valor medio de (5.44) para $n > c$:

$$L_q = \mathrm{E}\left[N_q\right] = \sum_{n=c}^{\infty}(n-c)p_n = \sum_{n=c}^{\infty}(n-c)p_0\frac{c^c}{c!}\left(\frac{\lambda}{c\mu}\right)^n =$$

$$= p_0\frac{c^c}{c!}\left(\frac{\lambda}{c\mu}\right)^c\sum_{n=c}^{\infty}(n-c)\left(\frac{\lambda}{c\mu}\right)^{n-c} = p_c\sum_{m=0}^{\infty}m\rho^m \qquad \textbf{(5.49)}$$

Sustituyendo en (5.49) los valores obtenidos en (5.35) y (5.48), sucesivamente, se obtiene:

$$L_q = p_c\frac{\rho}{(1-\rho)^2} = P_q\frac{\rho}{1-\rho} \qquad \textbf{(5.50)}$$

Modelo *M/M/c* (V)

- El tiempo medio de espera en cola, aplicando la forma (5.4) del Teorema de Little a (5.50), y considerando (5.47), viene dado por la relación:

$$W_q = \frac{L_q}{\lambda} = P_q \frac{\rho}{\lambda(1-\rho)} = \frac{P_q}{c\mu - \lambda} \qquad \textbf{(5. 51)}$$

- Y para el tiempo medio de estancia en el sistema, por (5.2):

$$W = W_q + T_s = \frac{P_q}{c\mu - \lambda} + \frac{1}{\mu} \qquad \textbf{(5.52)}$$

- El número medio de clientes en el sistema se obtiene aplicando la forma (5.3) del Teorema de Little a (5.52):

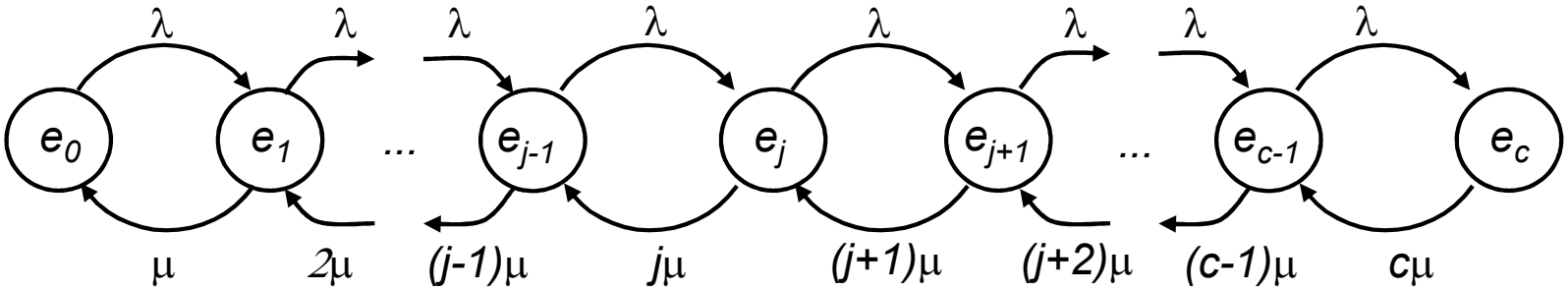
$$L = \lambda W = \frac{\lambda P_q}{c\mu - \lambda} + \frac{\lambda}{\mu} = \frac{P_q \rho}{1-\rho} + c\rho = L_q + c\rho \qquad \textbf{(5.53)}$$

Modelo *M/M/c/c* (I)

- Las condiciones son iguales que en el caso M/M/c, pero no hay colas de espera: Cualquier cliente adicional (> *c*) se rechaza.
- En estas condiciones, el número de unidades en el sistema es una cadena de Markov, equivalente al proceso nacimiento – muerte, en la que:

$$\lambda_j = \begin{cases} \lambda & (j < c) \\ 0 & (j \geq c) \end{cases} \quad \mu_j = \begin{cases} j\mu & (j \leq c) \\ 0 & (j > c) \end{cases} \qquad \textbf{(5.54)}$$

- El diagrama de transiciones de estados será:



Modelo *M/M/c/c* (II)

- Sustituyendo (5.54) en (5.21) se obtiene la función de distribución de probabilidad del **número de unidades en el sistema**:

$$p_n = p_0 \left(\frac{\lambda}{\mu} \right)^n \frac{1}{n!} \quad (0 \le n \le c) \tag{5.55}$$

- p_0 se calcula aplicando el segundo axioma de la probabilidad, y se obtiene:

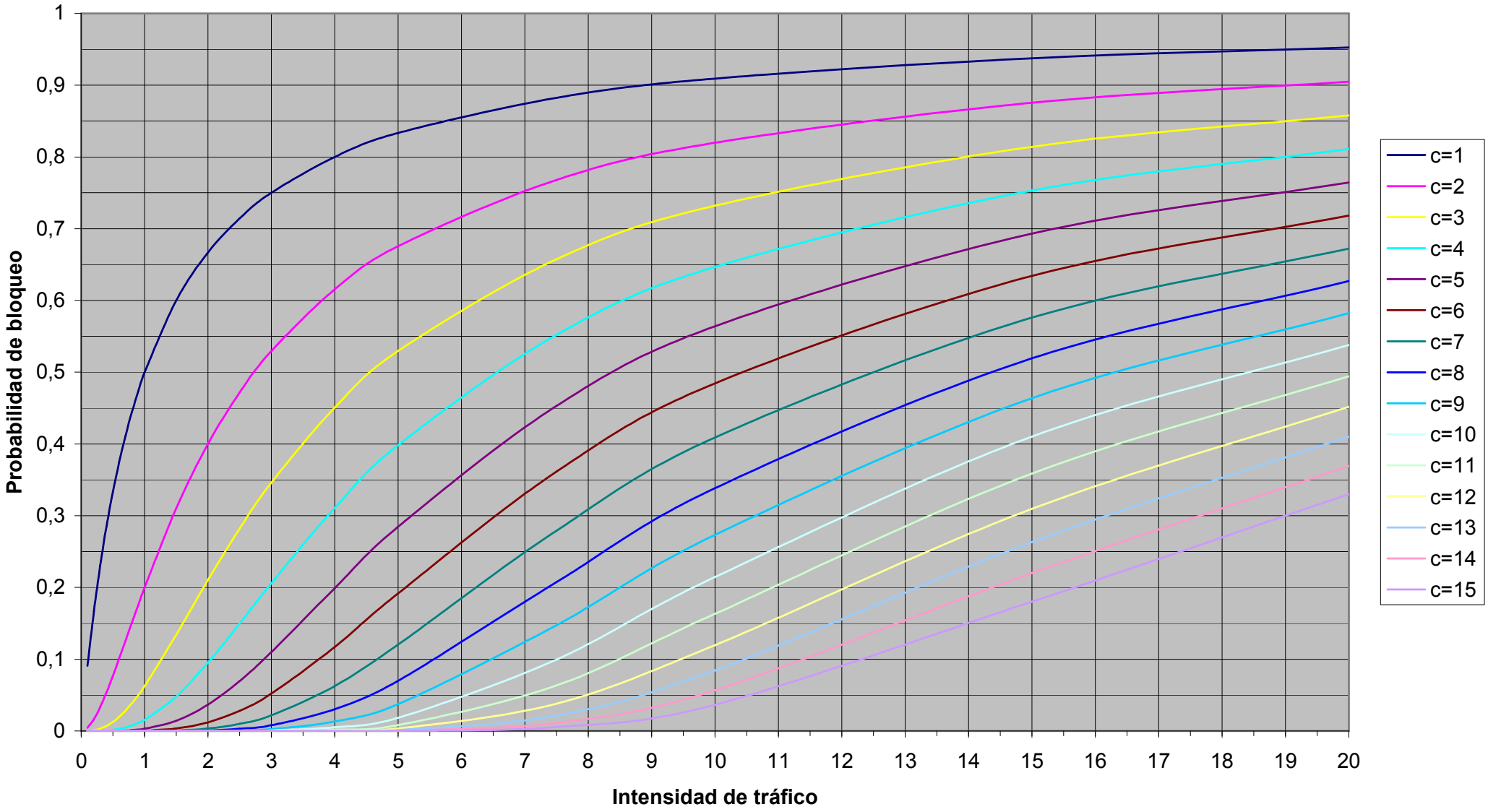
$$p_0 = \left[\sum_{n=0}^{c} \left(\frac{\lambda}{\mu} \right)^n \frac{1}{n!} \right]^{-1} \tag{5.56}$$

- La probabilidad de perder una petición será la probabilidad de que haya *c* unidades en el sistema:

$$p_c = \frac{(\lambda/\mu)^c / c!}{\sum_{i=0}^{c} \left[(\lambda/\mu)^i / i! \right]} = E_B(c,u) \tag{5.57}$$

conocida como *fórmula B de Erlang* o *fórmula de pérdidas de Erlang.*

Representación gráfica de la función Erlang-B



Modelo *M/M/c/c* (III)

- La tasa de llegadas efectiva al sistema será:

$$\lambda' = \lambda\left(1 - p_c\right) \tag{5.58}$$

- El factor de utilización de cada servidor viene dado por:

$$\rho = \frac{\lambda'}{c\mu} = \frac{\lambda}{c\mu}\left(1 - p_c\right) \tag{5.59}$$

- En este modelo, al no haber cola de espera, W_q=0 y L_q=0, con lo que por (5.2) y (5.3):

$$W = W_q + \frac{1}{\mu} = \frac{1}{\mu} \tag{5.60}$$

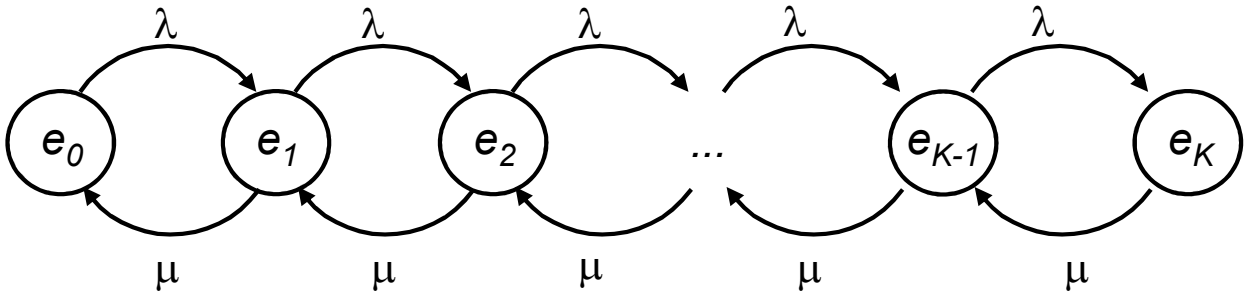
$$L = \lambda' W = \frac{\lambda'}{\mu} = \frac{\lambda}{\mu}\left(1 - p_c\right) = c\rho \tag{5.61}$$

Modelo *M/M/1/K* (I)

- Las condiciones son iguales que en el caso M/M/1, pero el número de unidades en el sistema está limitado a *K* (cola finita).
- En estas condiciones, el número de unidades en el sistema es una cadena de Markov, equivalente al proceso nacimiento – muerte, en la que:

$$\lambda_j = \begin{cases} \lambda & (j < K) \\ 0 & (j \geq K) \end{cases} \quad \mu_j = \begin{cases} \mu & (j \leq K) \\ 0 & (j > K) \end{cases} \qquad \textbf{(5.62)}$$

- El diagrama de transiciones de estados será:



- Sustituyendo (5.62) en (5.21) se obtiene la función de distribución de probabilidad del número de unidades en el sistema:

$$p_n = p_0 \left(\frac{\lambda}{\mu} \right)^n \quad (0 \leq n \leq K) \qquad \textbf{(5.63)}$$

Modelo *M/M/1/K* (II)

- p_0 se calcula a partir del segundo axioma de la probabilidad, y se obtiene:

$$p_0 = \left[\sum_{n=0}^{K}\left(\frac{\lambda}{\mu}\right)^n\right]^{-1} = \begin{cases} \dfrac{1-\lambda/\mu}{1-(\lambda/\mu)^{K+1}} & (\lambda \neq \mu) \\ \dfrac{1}{K+1} & (\lambda = \mu) \end{cases} \quad \textbf{(5.64)}$$

- La tasa de llegadas efectiva, λ' viene dada por la expresión:

$$\lambda' = \lambda(1-p_K) = \begin{cases} \lambda\dfrac{1-(\lambda/\mu)^{K}}{1-(\lambda/\mu)^{K+1}} & (\lambda \neq \mu) \\ \lambda\dfrac{K}{K+1} & (\lambda = \mu) \end{cases} \quad \textbf{(5.65)}$$

Modelo *M/M/1/K* (III)

- Por haber un único servidor, la probabilidad de que el sistema esté activo dará el factor de utilización del servidor, que será:

$$\rho = 1 - p_0 = \begin{cases} \dfrac{\lambda}{\mu}\left[\dfrac{1-(\lambda/\mu)^K}{1-(\lambda/\mu)^{K+1}}\right] & (\lambda \neq \mu) \\ \dfrac{K}{K+1} & (\lambda = \mu) \end{cases} \quad \textbf{(5.66)}$$

  Y en cualquiera de los dos casos se puede verificar que es igual a la intensidad de tráfico **efectiva** de entrada al servidor:

$$\rho = \frac{\lambda'}{\mu} = \frac{\lambda}{\mu}(1 - p_K) \quad \textbf{(5.67)}$$

- El número medio de unidades en el sistema será el valor esperado de (5.63).
  - Para λ≠μ:

$$L = \mathrm{E}[N] = \sum_{n=0}^{K} n p_n = \sum_{n=0}^{K} n p_0 \left(\frac{\lambda}{\mu}\right)^n = p_0 \sum_{n=0}^{K} n \left(\frac{\lambda}{\mu}\right)^n \quad \textbf{(5.68)}$$

Modelo *M/M/1/K* (IV)

Análogamente a (5.35), el sumatorio se puede calcular del siguiente modo.

Haciendo u=λ/μ :

$$\sum_{n=0}^{K} nu^n = u\sum_{n=1}^{K} nu^{n-1} = u\frac{\partial}{\partial u}\sum_{n=1}^{K} u^n = u\frac{\partial}{\partial u}\left(\frac{u-u^{K+1}}{1-u}\right) =$$

$$= u\frac{(1-u)[1-(K+1)u^K]+(u-u^{K+1})}{(1-u)^2} = u\frac{1-(K+1)u^k+Ku^{K+1}}{(1-u)^2} \qquad \textbf{(5.69)}$$

Deshaciendo el cambio de variable y sustituyendo se obtiene:

$$L = \frac{\lambda/\mu}{1-\lambda/\mu}\left[\frac{1-(K+1)(\lambda/\mu)^k+K(\lambda/\mu)^{K+1}}{1-(\lambda/\mu)^{K+1}}\right] \quad (\lambda \neq \mu) \qquad \textbf{(5.70)}$$

Modelo *M/M/1/K* (V)

– Y en el caso en que λ=μ:

$$L = \sum_{n=0}^{K} n \frac{1}{K+1} = \frac{1}{K+1} \sum_{n=0}^{K} n = \frac{1}{K+1} \frac{K+1}{2} K = \frac{K}{2} \quad (\lambda = \mu) \qquad (5.71)$$

- Y para el resto de los valores medios, aplicando (5.3), (5.2) y (5.5), respectivamente:

$$W = \frac{L}{\lambda'} = \frac{L}{\lambda(1 - p_K)} \qquad (5.72)$$

OJO!

$$W_q = W - \frac{1}{\mu} \qquad (5.73)$$

$$L_q = L - \frac{\lambda'}{\mu} = L - \frac{\lambda}{\mu}(1 - p_K) = L - \rho \qquad (5.74)$$

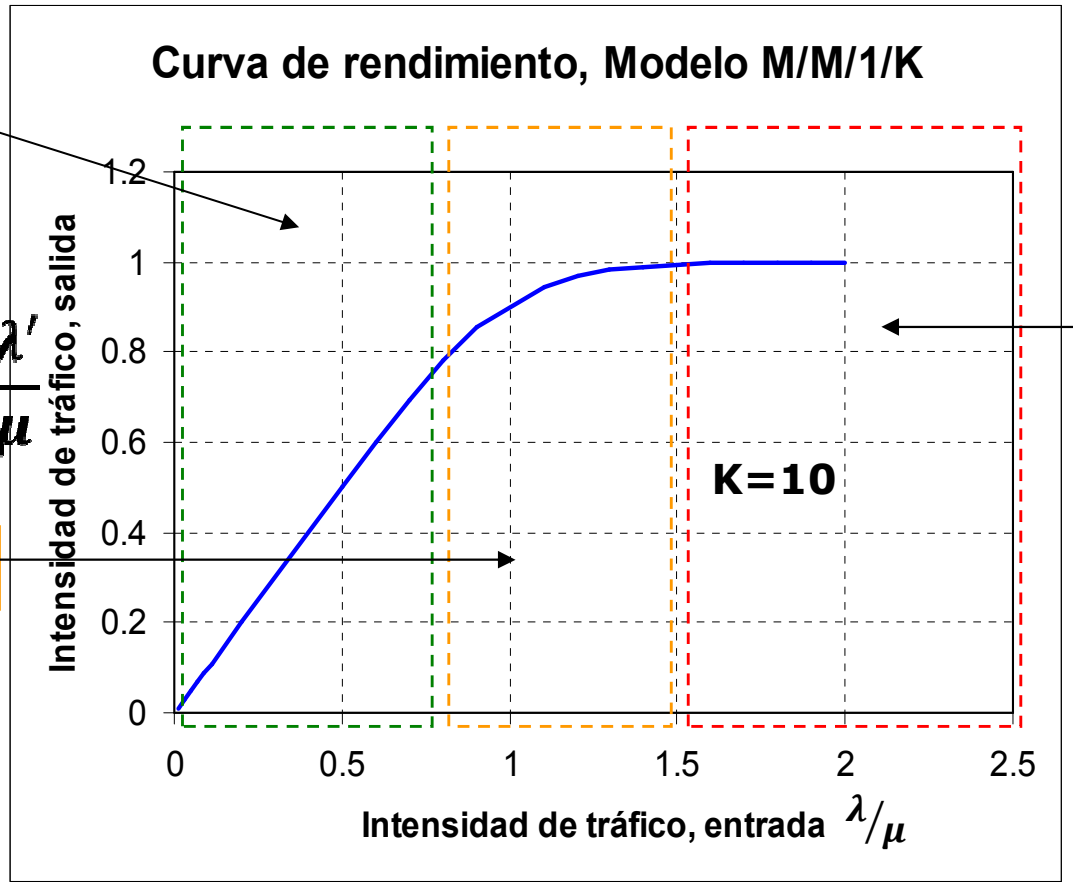
Modelo *M/M/1/K* (VI)

- La **curva de rendimiento del sistema** es la representación del tráfico efectivo que soporta el servidor (λ') en función del tráfico total que recibe a la entrada (λ) .
  - La curva tiene la siguiente forma:

1. Zona lineal, λ' = λ

$K \to \infty$

Cola grande comparada con la tasa de entrada



2. Zona de transición

$\rho = \frac{\lambda'}{\mu} = 1$

3. Zona de saturación, λ' = μ

$\lambda \to \infty$

Cola pequeña comparada con la tasa de entrada

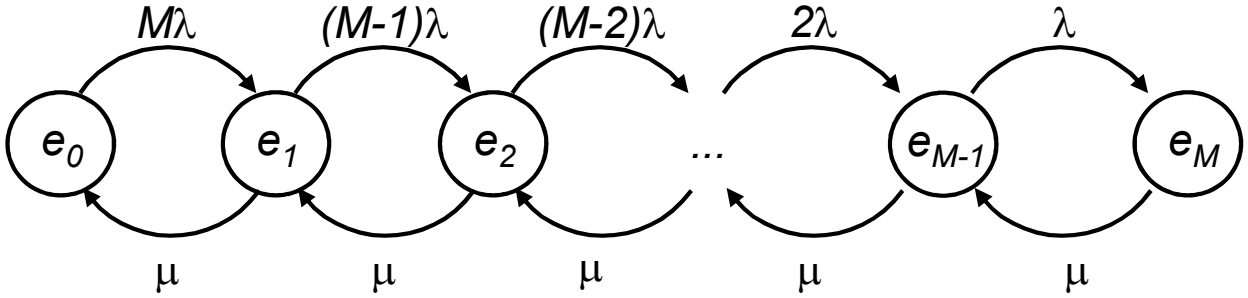
Se presenta normalizada con respecto a la tasa de servicio, μ, con lo que las magnitudes de los ejes son intensidades de tráfico (λ/ μ)

Modelo *M/M/1/∞/M* (I)

- Las condiciones son iguales que en el caso M/M/1, pero en este sistema hay un número finito de clientes *(M)*, que una vez servidos vuelven a poder hacer peticiones.
- La tasa de llegadas varía con el número de unidades en el sistema. Si cada cliente se encuentra operativo un tiempo T_c, genera $\lambda = 1 / T_c$ peticiones por unidad de tiempo.
- En estas condiciones, el número de unidades en el sistema es una cadena de Markov, equivalente al proceso nacimiento – muerte, en la que:

$$\lambda_n = \begin{cases} (M-n)\lambda & (0 \le n < M) \\ 0 & (n \ge M) \end{cases} \qquad \mu_n = \begin{cases} \mu & (0 \le n \le M) \\ 0 & (n > M) \end{cases} \tag{5.75}$$

- El diagrama de transiciones de estados será:



Modelo *M/M/1/∞/M* (II)

- Sustituyendo (5.75) en (5.21) se obtiene la función de distribución de probabilidad del número de unidades en el sistema:

$$p_n = p_0 \binom{M}{n} n! \left(\frac{\lambda}{\mu}\right)^n = p_0 \frac{M!}{(M-n)!} \left(\frac{\lambda}{\mu}\right)^n \qquad \textbf{(5.76)}$$

- p_0 se calcula para cumplir el segundo axioma de la probabilidad, y se obtiene:

$$p_0 = \left[\sum_{n=0}^{M} \frac{M!}{(M-n)!} \left(\frac{\lambda}{\mu}\right)^n \right]^{-1} \qquad \textbf{(5.77)}$$

- El factor de utilización del servidor, como sólo hay un servidor:

$$\rho = 1 - p_0 \qquad \textbf{(5.78)}$$

Modelo *M/M/1/∞/M* (III)

- Y por la misma razón, la tasa de llegadas **efectiva** vendrá dada por la expresión:

$$\lambda' = \mu\rho = \mu\left(1 - p_0\right) \qquad \textbf{(5.79)}$$

- Por otro lado, la tasa de llegadas **efectiva** será el valor esperado de las tasas de llegada λ_n:

$$\lambda' = \sum_{n=0}^{M} \lambda_n p_n = \sum_{n=0}^{M} \left(M - n\right)\lambda p_n = \lambda\left(M\sum_{n=0}^{M} p_n - \sum_{n=0}^{M} n p_n \right) = \lambda\left(M - L\right) \qquad \textbf{(5.80)}$$

- Despejando L y teniendo en cuenta (5.79) se obtiene:

$$L = M - \frac{\lambda'}{\lambda} = M - \frac{\mu}{\lambda}\rho \qquad \textbf{(5.81)}$$

que es el mismo valor al que se llegaría calculando E[*N*].

Modelo *M/M/1/∞/M* (IV)

- El resto de los valores medios se calculan por el teorema de Little (5.3) y por (5.2):

$$W = \frac{1}{\lambda'}\left[M - \frac{\mu}{\lambda}\rho\right] = \frac{M}{\mu\rho} - \frac{1}{\lambda} = \frac{MT_s}{\rho} - T_c \qquad \textbf{(5.82)}$$

$\boldsymbol{\lambda' = \mu\rho}$

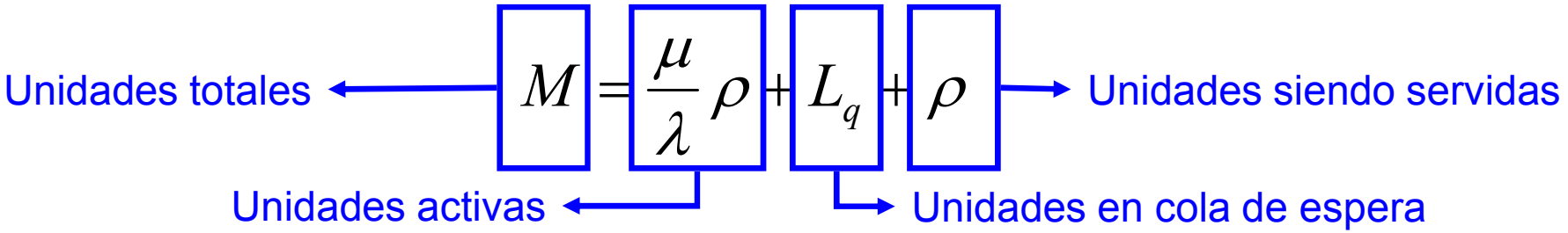
$$W_q = W - \frac{1}{\mu} = \frac{MT_s}{\rho} - T_c - T_s \qquad \textbf{(5.83)}$$

$$L_q = \lambda' W_q = M - \frac{\mu}{\lambda}\rho - \rho \qquad \textbf{(5.84)}$$

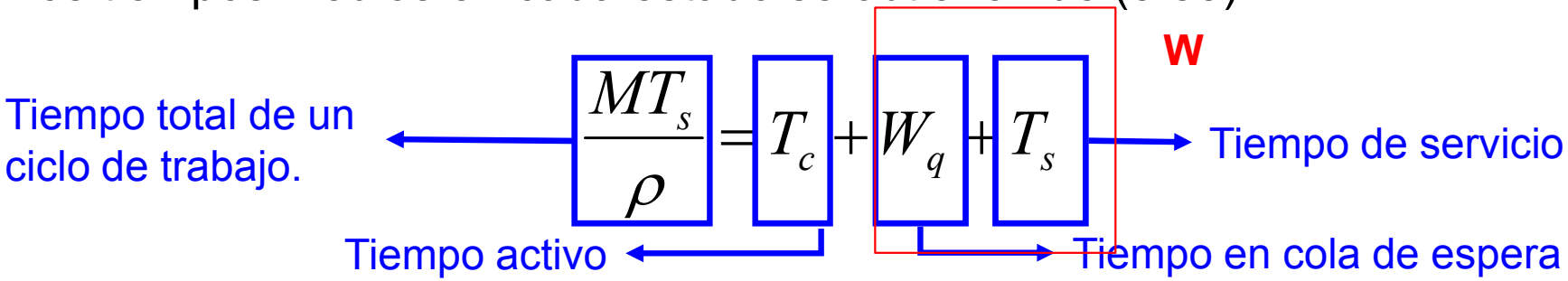
- Un cliente en el sistema puede encontrarse en 3 estados posibles:
  (a) Activo.
  (b) En espera de servicio.
  (c) En servicio.

Modelo *M/M/1/∞/M* (V)

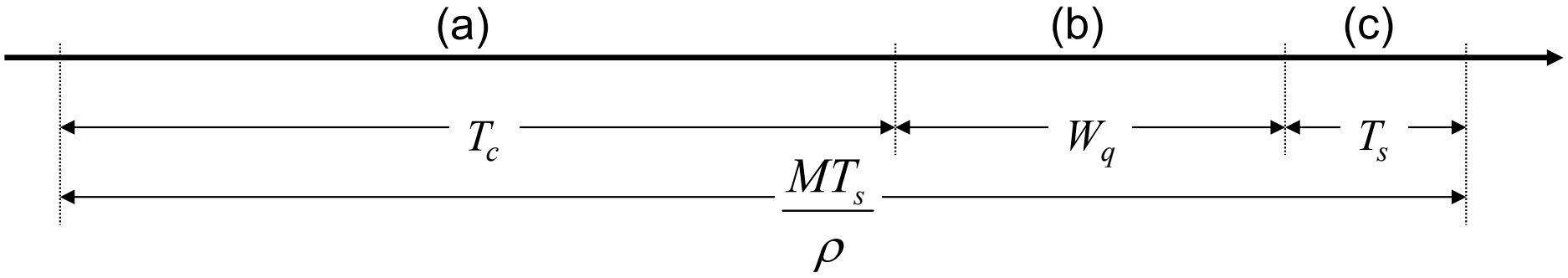
- La expresión (5.84) representa los valores medios de unidades en cada estado:



- Los tiempos medios en cada estado se obtienen de (5.83):



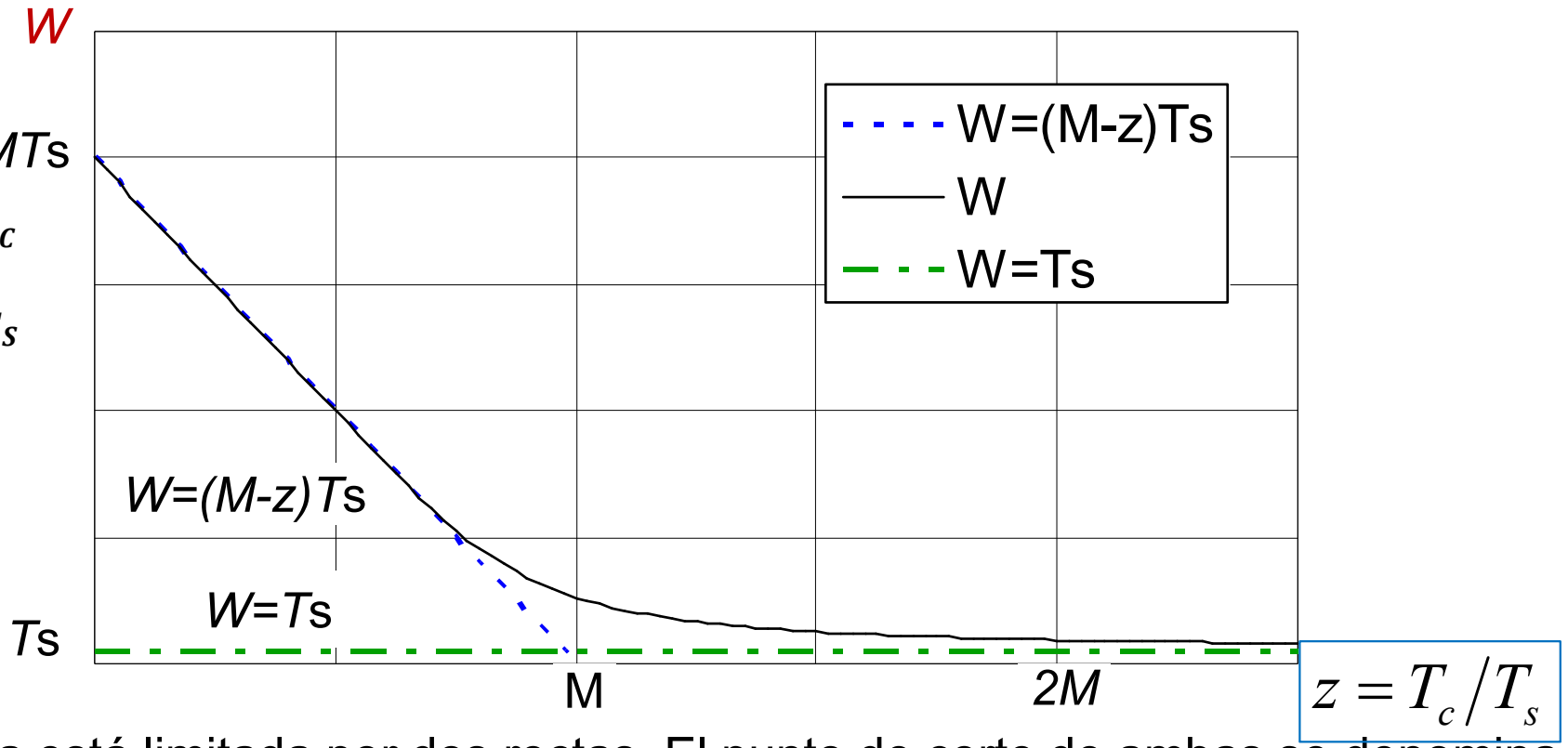
- El ciclo de funcionamiento de un cliente del sistema es:



Modelo *M/M/1/∞/M* (VI)

- Representación gráfica del **tiempo medio de estancia en el sistema** para M fijo y **T_c variable**:

$$W = \frac{MT_S}{\rho} - T_C = \left(\frac{M}{\rho} - \frac{T_C}{T_S}\right)T_S = \left(\frac{M}{\rho} - z\right)T_S$$



- La curva está limitada por dos rectas. El punto de corte de ambas se denomina ***Punto de Saturación***, y se encuentra en el valor de T_c/T_s:

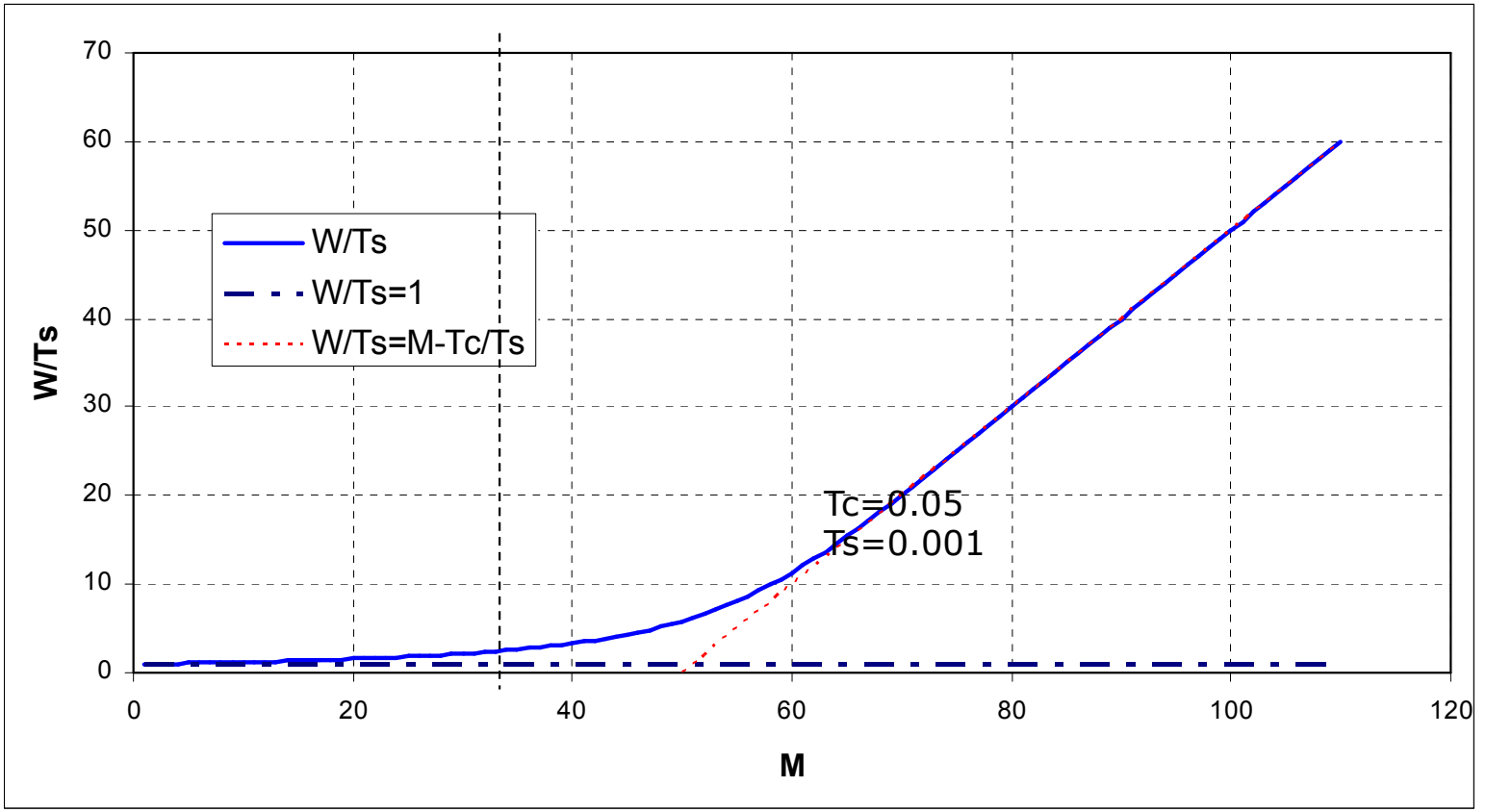
$$z_{sat} = T_c/T_s = M - 1 \qquad \textbf{(5.85)}$$

Modelo *M/M/1/∞/M* (VII)

- Representación gráfica del tiempo medio de estancia en el sistema para T_c fijo y M variable, normalizada respecto a T_s :

$$W = \frac{MT_S}{\rho} - T_C$$

$$\frac{W}{T_S} = \left(\frac{M}{\rho} - \frac{T_C}{T_S}\right)$$



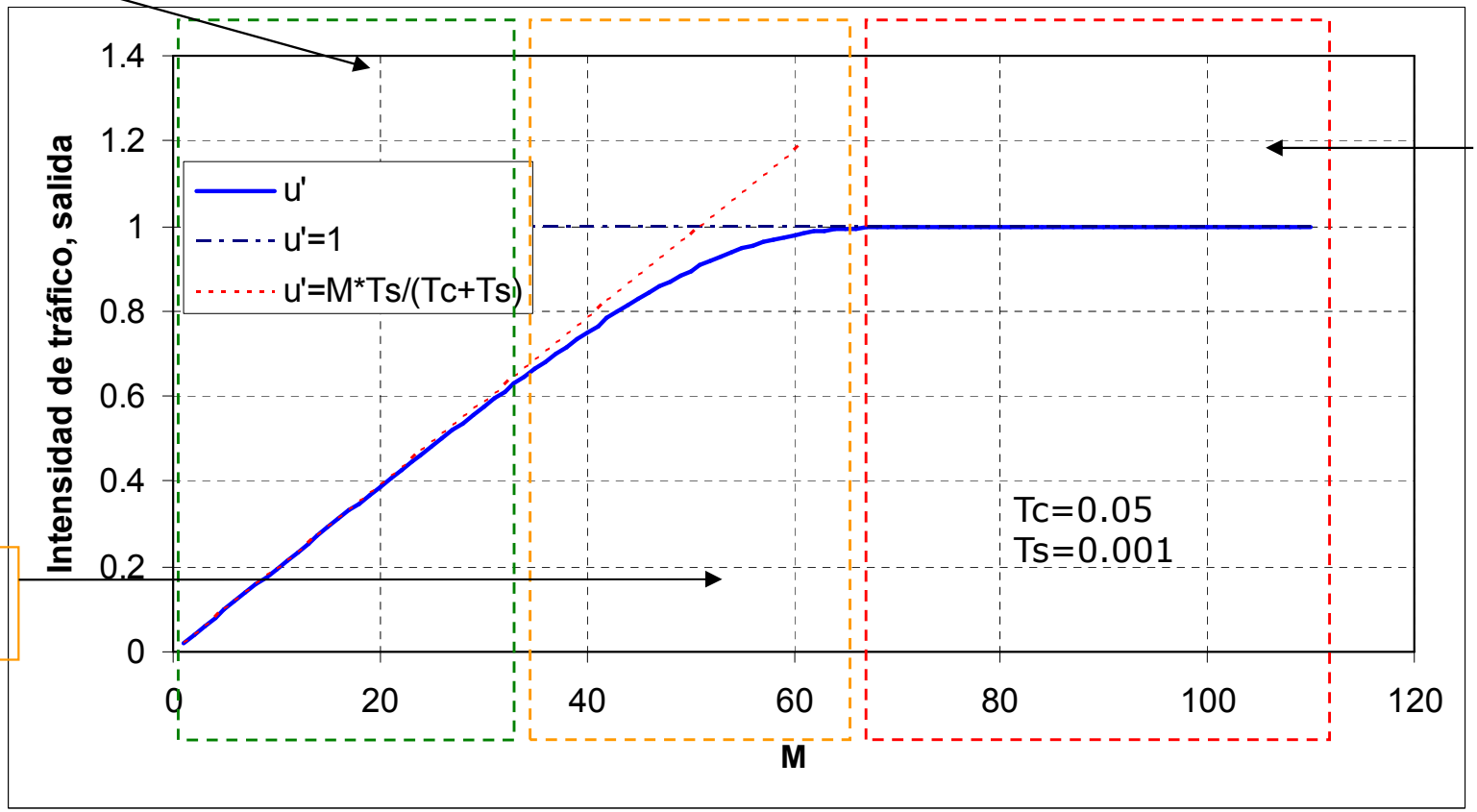
- Igual que en el caso anterior, tenemos dos rectas de límite que se cortan en:

$$M_{sat} = T_c / T_s + 1$$

Modelo *M/M/1/∞/M* (VIII)

- Representación gráfica de la curva de rendimiento, normalizada respecto a μ:

1. Zona lineal, λ' = M/(Tc+Ts)



3. Zona de saturación, λ' = μ

$$\frac{\lambda'}{\mu} = \lambda' T_s$$

2. Zona de transición

- Igual que en el caso anterior, tenemos dos rectas de límite que se cortan en:

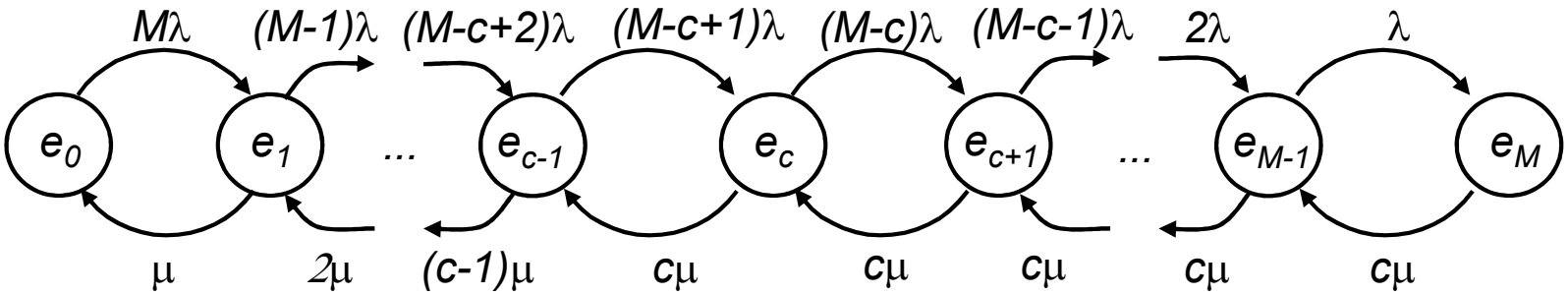
$$M_{sat} = T_c / T_s + 1$$

Modelo *M/M/c/∞/M* (I)

- Condiciones iguales que en el caso anterior, pero hay *c* servidores, y *c* < *M*.
- En estas condiciones, el número de unidades en el sistema es una cadena de Markov, equivalente al proceso nacimiento – muerte, en la que:

$$\lambda_n = \begin{cases} (M-n)\lambda & (0 \le n < M) \\ 0 & (n \ge M) \end{cases} \qquad \mu_n = \begin{cases} n\mu & (0 \le n < c) \\ c\mu & (c \le n \le M) \\ 0 & (n > M) \end{cases} \qquad \textbf{(5.86)}$$

- El diagrama de transiciones de estados será:



Modelo *M/M/c/∞/M* (II)

- Sustituyendo (5.86) en (5.21) se obtiene la función de distribución de probabilidad del número de unidades en el sistema:

$$p_n = \begin{cases} p_0 \binom{M}{n} \left(\frac{\lambda}{\mu}\right)^n & (0 \le n < c) \\ p_0 \binom{M}{n} \frac{n!}{c^{n-c} c!} \left(\frac{\lambda}{\mu}\right)^n & (c \le n < M) \end{cases} \qquad \textbf{(5.87)}$$

- p_0 se calcula para cumplir el segundo axioma de la probabilidad.

$$p_0 = \left[\sum_{n=0}^{c-1} \binom{M}{n} \left(\frac{\lambda}{\mu}\right)^n + \sum_{n=c}^{M} \binom{M}{n} \frac{n!}{c^{n-c} c!} \left(\frac{\lambda}{\mu}\right)^n \right]^{-1} \qquad \textbf{(5.87a)}$$

Modelo *M/M/c/∞/M* (III)

- El factor de ocupación se puede calcular a partir de la probabilidad de que un servidor esté libre. Si hay n servidores ocupados en el sistema (n<c), la probabilidad de que un servidor esté libre será

$$p\left\{\text{servidor libre}\middle|n\right\} = \frac{c-n}{c}$$

- Y por el teorema de la probabilidad total:

$$p\left\{\text{servidor libre}\right\} = \sum_{n=0}^{c-1} p_n \cdot p\left\{\text{servidor libre}\middle|n\right\} = \sum_{n=0}^{c-1} p_n \frac{c-n}{c}$$

- El factor de ocupación es la probabilidad complementaria:

$$\rho = 1 - \sum_{n=0}^{c-1} p_n \frac{c-n}{c} \qquad \textbf{(5.87b)}$$

Modelo *M/M/c/∞/M* (IV)

- El mecanismo de cálculo es el mismo que en el modelo con un único servidor, teniendo en cuenta que aquí:

$$\lambda' = c\mu\rho \qquad \textbf{(5.88)}$$

- El cálculo de la tasa de llegadas efectiva como valor medio de su función de distribución es idéntico al realizado en (5.80). Igualando (5.80) con (5.88) se obtiene el valor del número medio de unidades en el sistema:

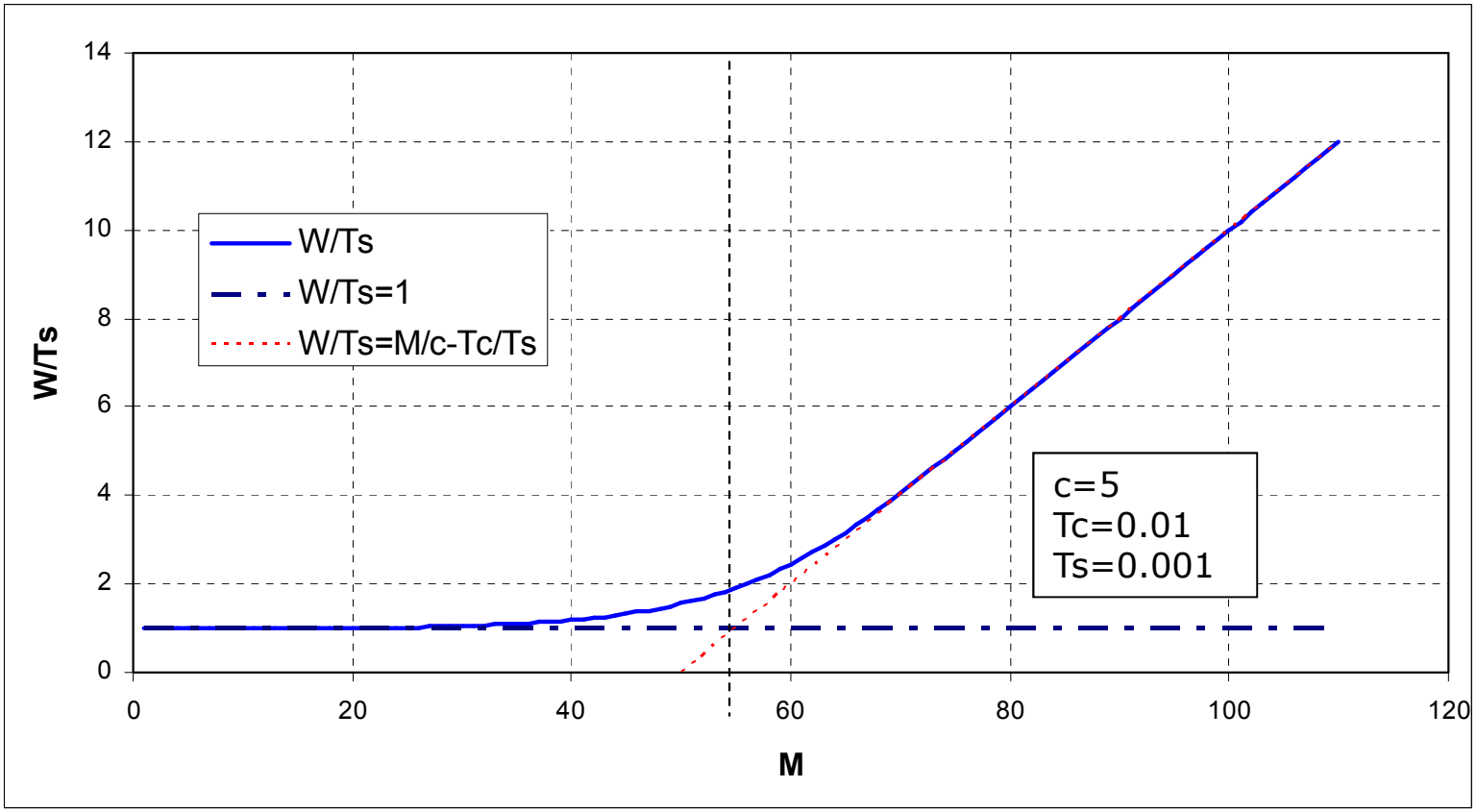
$$L = M - \frac{\lambda'}{\lambda} = M - \frac{c\mu}{\lambda}\rho \qquad \textbf{(5.88a)}$$

- Aplicando el teorema de Little:

$$W = \frac{MT_s}{c\rho} - T_c \qquad \textbf{(5.88b)}$$

Modelo *M/M/c/∞/M* (V)

- Representación gráfica del tiempo medio de estancia en el sistema para T_c fijo y *M* variable, normalizada respecto a T_s :



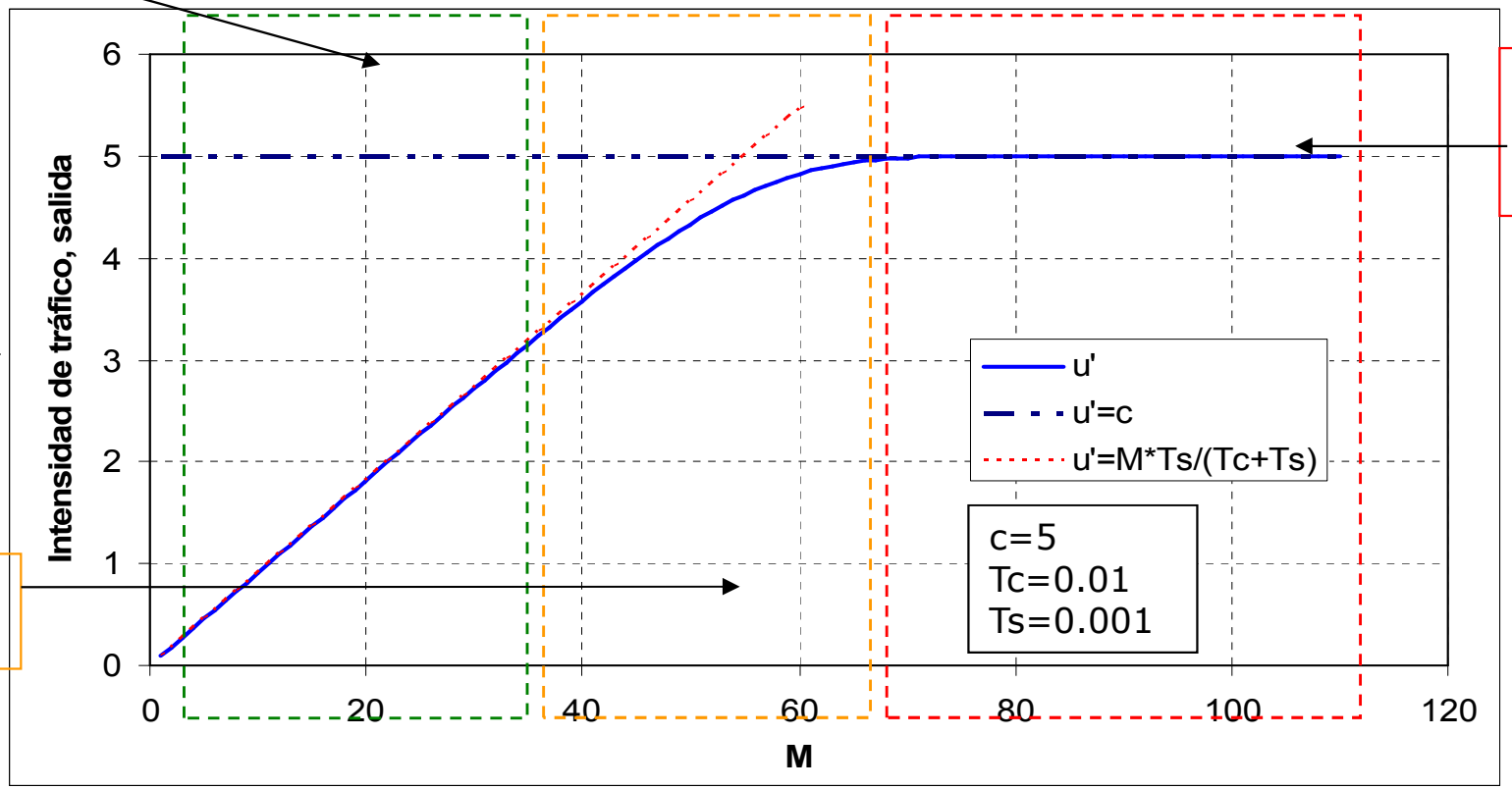
- Tenemos dos rectas de límite que se cortan en:

$$M_{sat} = c\left(T_c / T_s + 1\right)$$

Modelo *M/M/c/∞/M* (VI)

- Representación gráfica de la curva de rendimiento, normalizada respecto a μ:

1. Zona lineal, λ' = M/(Tc+Ts)



3. Zona de saturación, λ' = cμ

$$\frac{\lambda'}{\mu} = \lambda' T_s$$

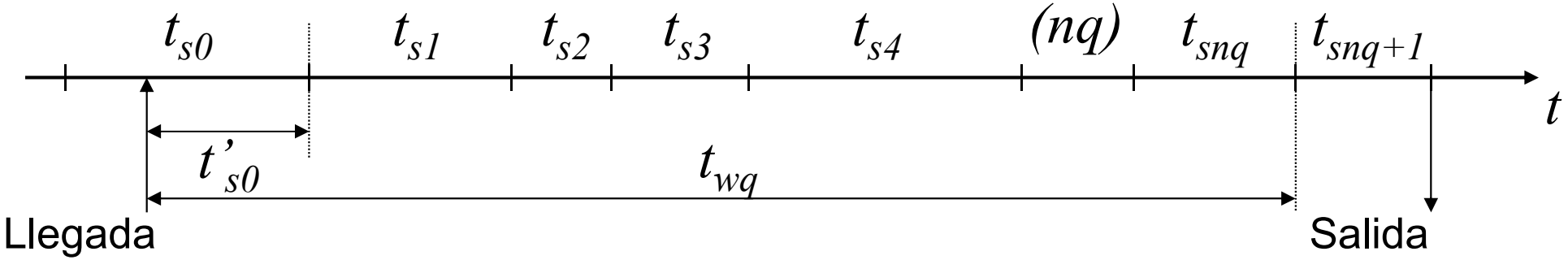
2. Zona de transición

- Igual que en el caso anterior, tenemos dos rectas de límite que se cortan en:

$$M_{sat} = c\left(T_c / T_s + 1\right)$$

Modelo *M/G/1* (I)

- El tiempo de servicio tiene una distribución aleatoria cualquiera S, de la que se conocen sus momentos E[S] y E[S^2].
- Para calcular el tiempo medio de estancia en el sistema, recordando (5.40)



$$t_{Wn} = t'_{s0} + t_{s1} + t_{s2} + \ldots + t_{snq} + t_{snq+1}$$

- Para calcular el valor esperado del tiempo medio en cola, considerando que todos los tiempos son independientes:

$$W_q = \mathrm{E}\left[t'_{s0} + t_{s1} + t_{s2} + \ldots + t_{snq}\right] = \mathrm{E}\left[t'_{s0}\right] + \mathrm{E}\left[N_q\right]\mathrm{E}\left[S\right] \qquad \textbf{(5.89)}$$

- Por el teorema de Little (5.4):

$$W_q = \mathrm{E}\left[t'_{s0}\right] + \lambda W_q \,\mathrm{E}\left[S\right] \qquad \textbf{(5.90)}$$

Modelo *M/G/1* (II)

- El valor esperado del tiempo residual se puede demostrar que viene dado por la expresión:

$$\mathrm{E}\left[t'_{s0}\right] = \frac{\lambda\,\mathrm{E}\left[S^2\right]}{2} \qquad \textbf{(5.91)}$$

- Sustituyendo en (5.90), y recordando que ρ=λ/μ= λE[S]:

$$W_q = \frac{\lambda\,\mathrm{E}\left[S^2\right]}{2} + W_q\rho \Rightarrow W_q = \frac{\lambda\,\mathrm{E}\left[S^2\right]}{2\left(1-\rho\right)} \qquad \textbf{(5.92)}$$

- El resto de los valores medios se obtienen a partir del teorema de Little y (5.2):

$$W = W_q + T_s = \frac{\lambda\,\mathrm{E}\left[S^2\right]}{2\left(1-\rho\right)} + \mathrm{E}\left[S\right] \qquad \textbf{(5.93)}$$

$$L_q = \lambda W_q = \frac{\lambda^2\,\mathrm{E}\left[S^2\right]}{2\left(1-\rho\right)} \qquad \textbf{(5.94)}$$

$$L = \lambda W = \frac{\lambda^2\,\mathrm{E}\left[S^2\right]}{2\left(1-\rho\right)} + \rho \qquad \textbf{(5.95)}$$

Modelo *M/G/1* (III)

- Verificando los resultados para una distribución exponencial, *(M/M/1)*:

$$E[S]=\frac{1}{\mu} \quad E[S^2]=\frac{2}{\mu^2} \quad W_q=\frac{\lambda}{2(1-\rho)}\frac{2}{\mu^2}=\frac{\rho}{\mu(1-\rho)} \qquad \textbf{(5.96)}$$

  que es el valor obtenido en (5.38).

- Para una distribución constante, *M/D/1*:

$$E[S]=\frac{1}{\mu} \quad E[S^2]=\frac{1}{\mu^2} \quad W_q=\frac{\lambda}{2(1-\rho)}\frac{1}{\mu^2}=\frac{\rho}{2\mu(1-\rho)} \qquad \textbf{(5.97)}$$

  Son los valores mínimos para cualquier cola M/G/1 con los mismos valores de λ y μ. El valor obtenido para W es la mitad que el caso M/M/1.

- En caso de no conocer de modo analítico la función de distribución de probabilidad de *S*, se pueden emplear estimadores de E[S] y E[S^2] a partir de muestras del tiempo de servicio:

$$\{t_{s1}, t_{s2}, \ldots t_{sn}\} \qquad \overline{S}=\frac{\sum_{i=0}^{n} t_i}{n} \quad \textbf{(5.98)} \qquad \overline{S^2}=\frac{\sum_{i=0}^{n} t_i^{\,2}}{n}=\sigma_s^{\,2}+\overline{S}^2 \quad \textbf{(5.99)}$$

Aproximación a distribuciones exponenciales

- Se define el *Coeficiente Cuadrático de Variación* de una variable aleatoria X como:

$$C^2 = \frac{\mathrm{Var}[X]}{\mathrm{E}[X]^2} \qquad \textbf{(5.100)}$$

  es decir, el cociente entre la varianza y el cuadrado del valor medio.
- Si X tiene distribución exponencial, $\mathbf{C^2 = 1}$.
- Si se desconoce la distribución del tiempo de servicio se pueden estimar la media y la varianza, calcular C^2 y analizar según sus valores el tipo de proceso:
  - $0 < C^2 < 0{,}7$: Tendencia uniforme.
    - Se puede modelar mediante una distribución de tipo **Erlang-m** ($C^2 = 1/m$).
  - $0{,}7 < C^2 < 1{,}3$: Comportamiento **Poissoniano**.
  - $1{,}3 < C^2$: Tendencia al agrupamiento.:
    - Se puede modelar mediante una distribución de tipo **hiperexponencial**.

Redes de colas

- Representan el flujo de peticiones de clientes a través de varios centros de servicio.

- Redes de colas **sin retroalimentación** → Teorema de **Burke**
- Redes de colas **con retroalimentación** → Teorema de **Jackson**

    - **Redes de colas abiertas**: Ningún cliente permanece en la red de colas indefinidamente.

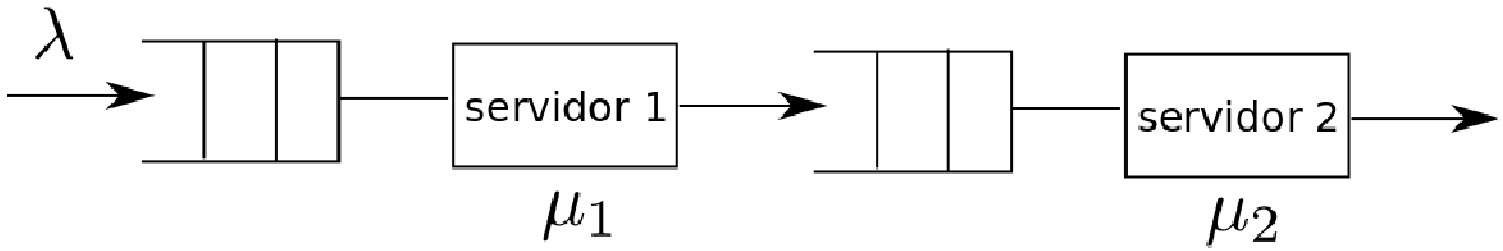
    - Redes de colas cerradas: un número fijo de clientes circulan indefinidamente por el sistema de colas.

Redes de colas: Teorema de Burke (I)

- Las salidas de algunas colas convierten en las llegadas de otras colas, pero **no hay realimentación** de una cola consigo misma.

TEOREMA DE BURKE

1. La salida **en estado estacionario** de un sistema M/M/c, con parámetro de entrada λ, es también un proceso de Poisson de parámetro λ.
2. Para cada t, el número de clientes a tiempo t es independiente de la secuencia de tiempos de partida anteriores a t.



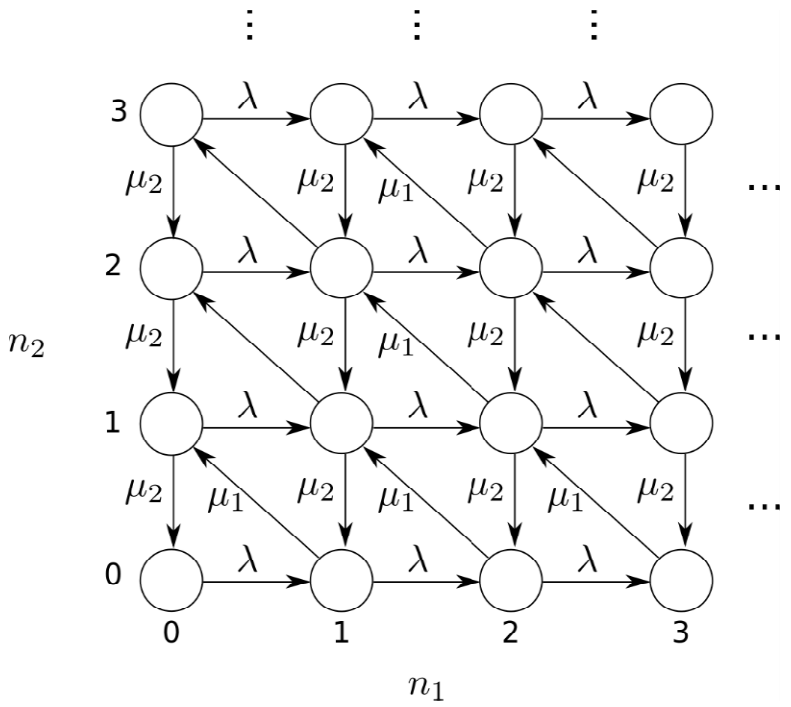
El número de clientes en el sistema 1 en t, $N_1(t)$, y el número de clientes en el sistema 2 en el mismo instante de tiempo t, $N_2(t)$, son **variables aleatorias independientes**. Por tanto:

$$P[N_1(t) = n, N_2(t) = m] = \mathrm{P}[N_1(t) = n]\, P[N_2(t) = m]$$

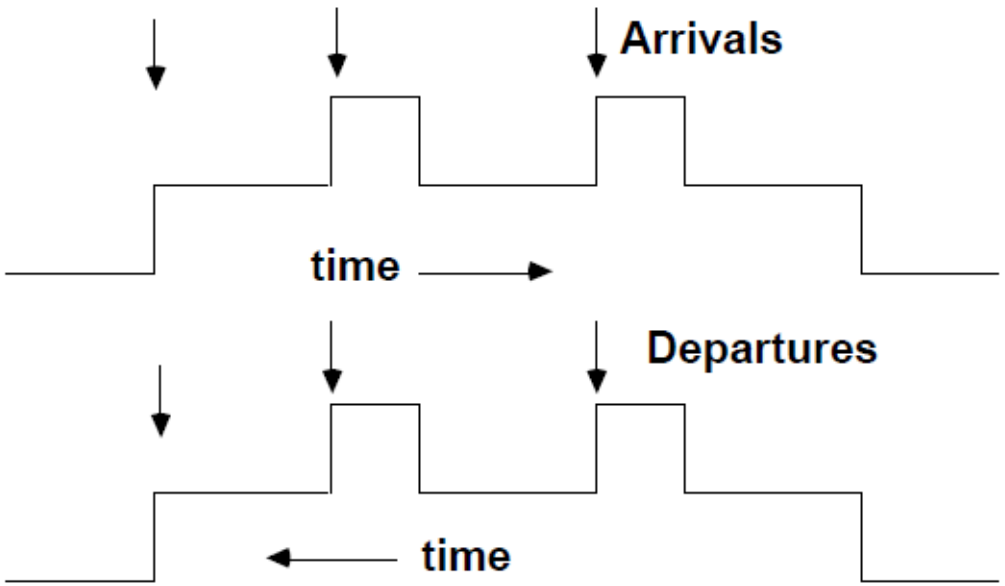
Redes de colas: Teorema de Burke (II)

- El estado del sistema viene determinado por el número de clientes en cada una de las colas en tiempo t. Por ejemplo para dos colas: $(N_1(t), N_2(t))$.

- Este vector de estados forma un proceso de Markov.



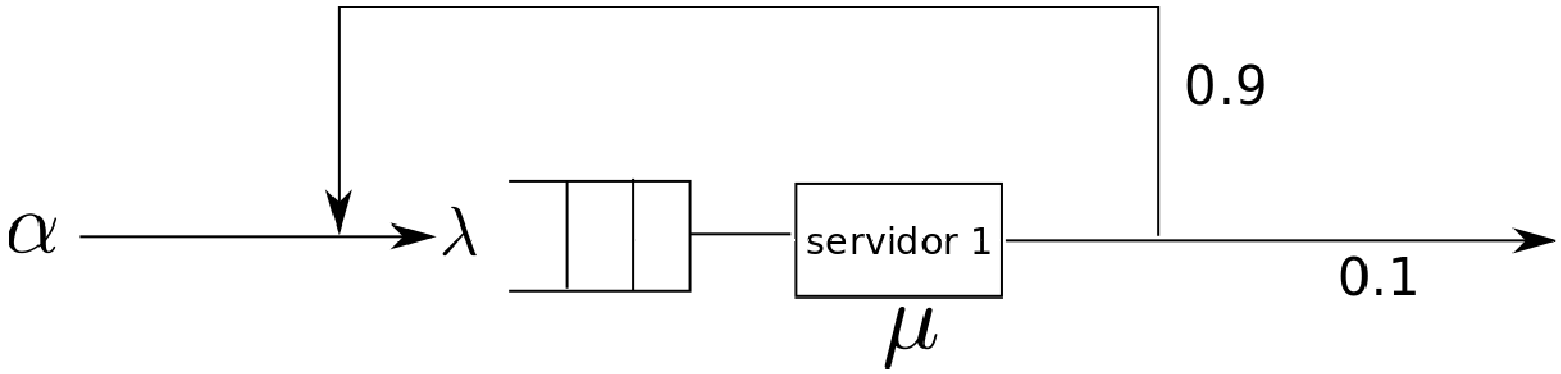
Redes de colas: Teorema de Burke (III)



1. Los procesos de nacimiento-muerte en estado estacionario son **reversibles en el tiempo**: las funciones de distribución *forward* y *backward* son indistinguibles → la salida del M/M/c tiene que seguir la misma distribución que la entrada.
2. Fijado un instante de tiempo t, las salidas antes de t en *forward* son llegadas después de t en backward. Estas llegadas son Poisson y son independientes de N(t) → N(t) y los instantes de partida son independientes.

Redes de colas: Teorema de Jackson (I)

- En muchas redes de colas, un cliente puede visitar una cola más de una vez → **hay retroalimentación**.
- En estos casos el teorema de Burke **NO** se puede aplicar.
- El teorema de Jackson extiende el Teorema de Burke para aquellos casos donde puede haber retroalimentación.
- Si un cliente puede visitar la cola más de una vez para la misma petición, entonces el proceso de llegada a la cola **NO** sigue una distribución de Poisson.



Redes de colas: Teorema de Jackson (II)

- **Redes de colas abiertas**: Conjunto de K colas las cuales:
  - Los clientes llegan del exterior en procesos de Poisson independientes con tasa α_j.
  - Cada cola j tiene c_j servidores. El tiempo de servicio de cada servidor está distribuido exponencialmente con media μ_j.
  - Tras el proceso en el sistema j, el cliente pasa a la cola i con una probabilidad p_{ji}, o sale del sistema con probabilidad $1-\sum_{i=1}^{K} p_{ji}$

- La **tasa total de llegadas** a cada servidor es:

$$\lambda_j = \alpha_j + \sum_{i=1}^{K} \lambda_i p_{ij} \qquad \textbf{(5.101)}$$

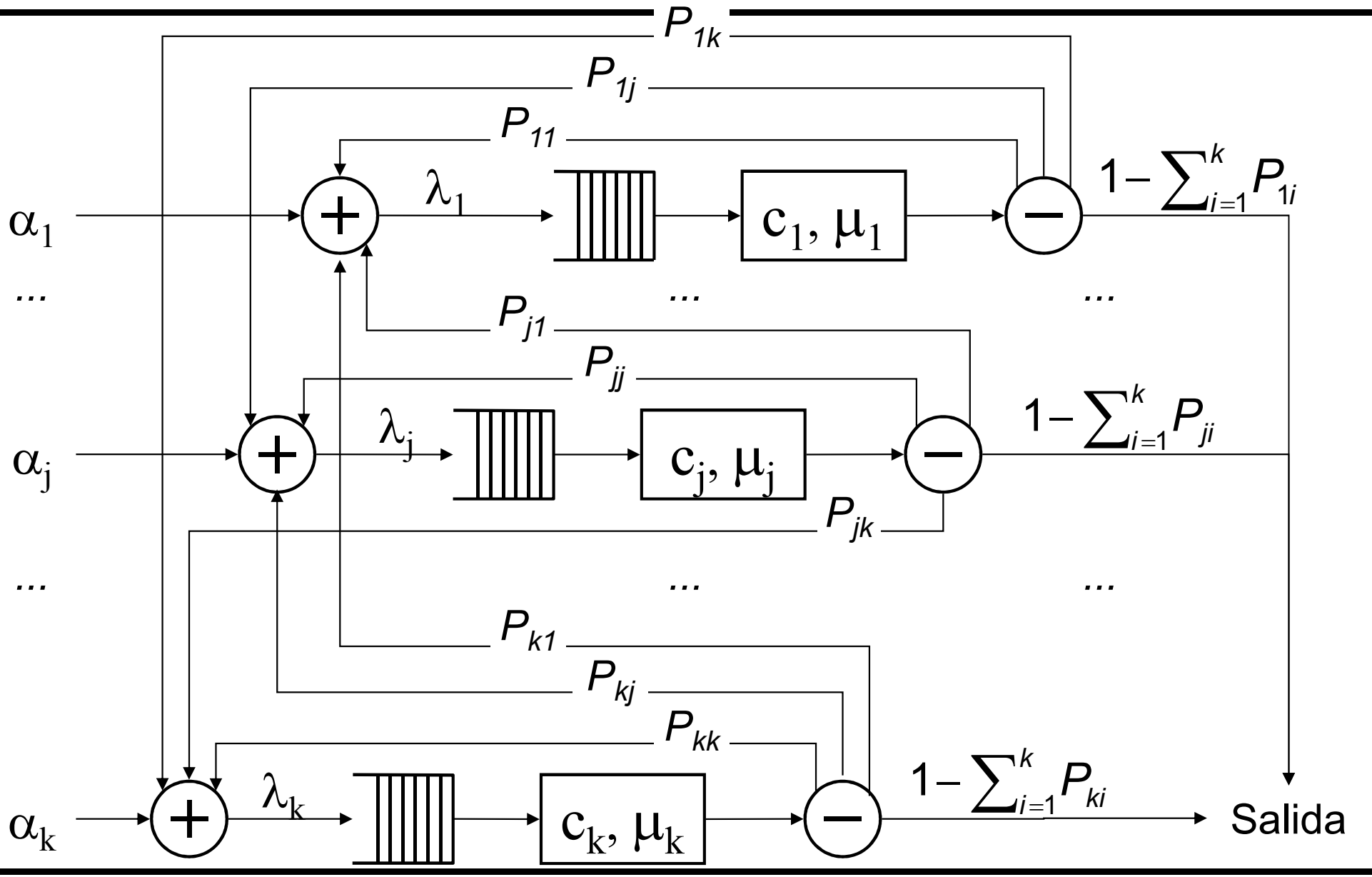
- Tiene solución única si no hay clientes que permanecen indefinidamente en el sistema. El vector del **número total de clientes en el sistema**:

$$N = \left(N_1, N_2, ..., N_K\right) \qquad \textbf{(5.102)}$$

  es un proceso de Markov.

- El **Teorema de Jackson** nos da la distribución estacionaria para N.

Redes de colas: Teorema de Jackson (III)



Redes de colas: Teorema de Jackson (IV)

TEOREMA DE JACKSON

En cualquier red de colas abiertas como la descrita, si se verifica que

$$\lambda_j < c_j \mu_j \quad \forall j \tag{5.103}$$

entonces para cualquier posible estado

$$\overline{n} = \left(n_1, n_2, ..., n_K \right) \tag{5.104}$$

la función de distribución de probabilidad del sistema viene dada por:

$$P\left[\overline{N} = \overline{n} \right] = P\left[N_1 = n_1 \right] P\left[N_2 = n_2 \right] ... P\left[N_K = n_K \right] \tag{5.105}$$

donde

$$P\left[N_j = n_j \right] \tag{5.106}$$

es la distribución del número de unidades en el sistema de un sistema *M/M/c_j* con tasa de llegadas λ_j y tasa de servicio μ_j.

Redes de colas (IV)

- El número total de unidades en el sistema es una variable aleatoria dada por la expresión:

$$N_T = \sum_{i=1}^{K} N_i \qquad \textbf{(5.107)}$$

y, por tanto, su valor esperado será:

$$L_T = \mathrm{E}\left[N_T\right] = \sum_{i=1}^{K} \mathrm{E}\left[N_i\right] = \sum_{i=1}^{K} L_i \qquad \textbf{(5.108)}$$

de este modo, aplicando el teorema de Little, obtenemos la siguiente expresión para el tiempo total de estancia en la red:

$$W_T = \frac{L_T}{\lambda_T} = \frac{\sum_{i=1}^{K} L_i}{\sum_{i=1}^{K} \alpha_i} \qquad \textbf{(5.109)}$$

Igualdades útiles (I)

- Suma finita S de los primeros N términos de una progresión geométrica de razón $r \neq 1$ y término inicial x_0:

$$S = \frac{x_0(1 - r^N)}{(1 - r)}$$

  Para $r = 1$, $S = Nx_0$

- Suma infinita S de los términos de una progresión geométrica de razón *r* y término inicial x_0:

$$S = \frac{x_0}{(1 - r)}$$

  Se tiene convergencia de la suma infinita S cuando $|r| < 1$.

- Dada una variable aleatoria X, cálculo de $\mathbb{E}[X^2]$ conocidas la media $\mathbb{E}[X]$ y la varianza $V[X]$:

$$\mathbb{E}[X^2] = V[X] + (\mathbb{E}[X])^2$$

$$V[X] = \mathbb{E}[(X - \mathbb{E}[X])^2 = \mathbb{E}[X^2] + (\mathbb{E}[X])^2 - 2\mathbb{E}\big[X\mathbb{E}[X]\big] = \mathbb{E}[X^2] + (\mathbb{E}[X])^2 - 2(\mathbb{E}[X])^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

Igualdades útiles (II)

- **Probabilidad condicional** del suceso A dado el suceso B:

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

- Por tanto: $P(A,B) = P(A|B)P(B) = P(B|A)P(A)$ (**regla del producto**).

- Por otro lado se cumple que $P(B) = \sum_{i=1}^{n} P(B,A_i)$, donde $A_1, A_2, \ldots, A_n$ son los posibles valores de la variable A (**regla de la suma**).

- **Teorema de la probabilidad total**: Sea B un suceso cualquiera del que se conocen las probabilidades condicionales $P(B|A_i)$ para todo $i = 1,2, \ldots, n$, entonces la probabilidad del suceso B viene dada por la expresión:

$$P(B) = \sum_{i=1}^{n} P(B,A_i) = \sum_{i=1}^{n} P(B|A_i)\,P(A_i)$$

Bibliografía especial del tema

- GROSS, D. y HARRIS, C.M., *Fundamentals of Queuing Theory*, Wiley, 1998. 3ª Ed.
- KLEINROCK, L., *Queuing Systems: Theory,* New York, Wiley, 1975.
- **LEON-GARCÍA, A, *Probability and Random Processes for Electrical Engineering*, Reading (MA), Addison-Wesley, 1994.**
- **PAPOULIS, A. y PILLAI, U., *Probability, Random Variables and Stocastic Processes,* McGraw-Hill, 2002. 4ª ed.**
- PAZOS, J.J., SUAREZ, A. y DÍAZ, R.P, *Teoría de Colas y Simulación de Eventos Discretos*, Prentice Hall, 2003.