Gestión y Planificación de Redes y Servicios
Área de Ingeniería Telemática

Herramientas para planificación y dimensionamiento

Area de Ingeniería Telemática
http://www.tlm.unavarra.es

Grado en Ingeniería en Tecnologías de Telecomunicación, 4º upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa

Gestión y Planificación de Redes y Servicios
Área de Ingeniería Telemática

Introducción al análisis y predicción de rendimiento

Introducción

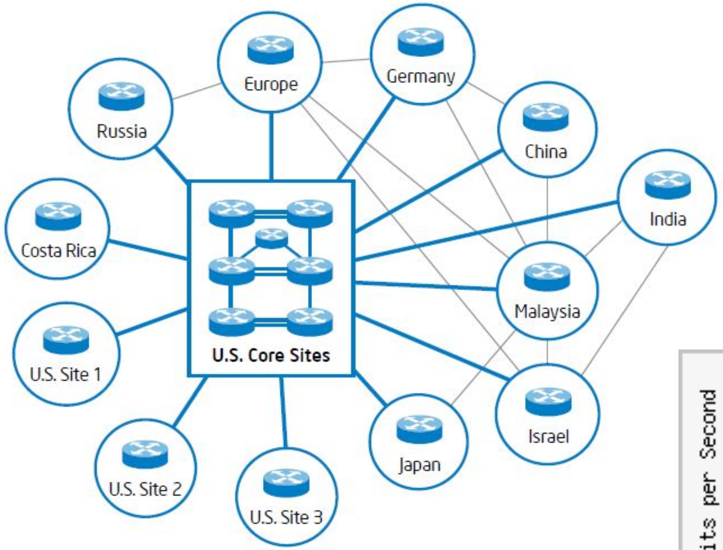
- Hasta ahora hemos visto cómo gestionar el equipamiento
- Cómo monitorizarlo, ver qué está sucediendo y por qué
- Sin embargo estas mediciones no nos dicen cómo se comportará el sistema en el futuro
- O cómo reaccionará ante cambios

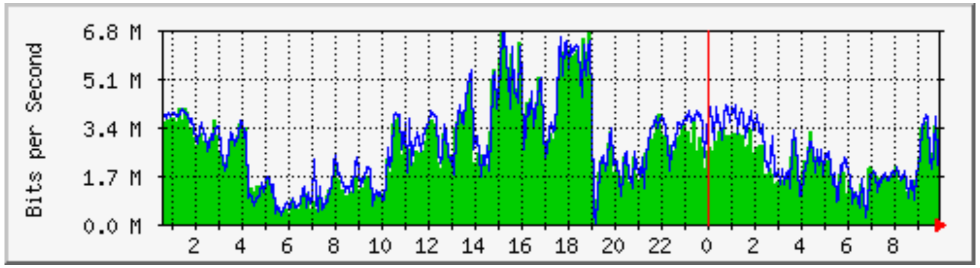
Ejemplo

- Queremos implementar una solución de VoIP
- Tenemos medidas pasivas de la utilización de los enlaces en un camino
- Tenemos medidas activas del OWD en ese trayecto
- ¿Qué OWD habrá cuando añadamos el tráfico de voz de alta prioridad?
- ¿Y si aumenta el tráfico best-effort?

Ejemplo (2)

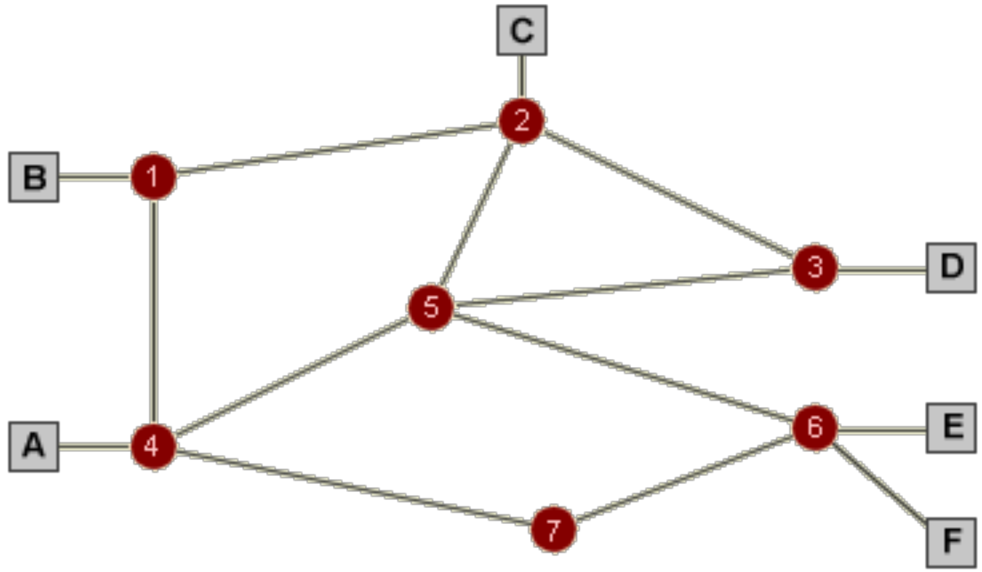
- Tenemos una red en producción
- Monitorizamos la utilización de los enlaces
- Monitorizamos las pérdidas (descartes en conmutadores o medidas activas)
- ¿Cuánto puede aumentar el tráfico tal que las pérdidas se mantengan en un porcentaje "aceptable"?
- ¿Cómo reacciona la red ante cambios en la matriz de tráfico?





Ejemplo (3)

- Se planea una red de conmutación de circuitos
- Se conoce la demanda de los usuarios
- Eso quiere decir no solo matrices de tráfico sino cuándo se generan los circuitos y cuánto duran
- ¿Probabilidad de bloqueo?
- ¿Cómo depende de cómo se haga el encaminamiento?

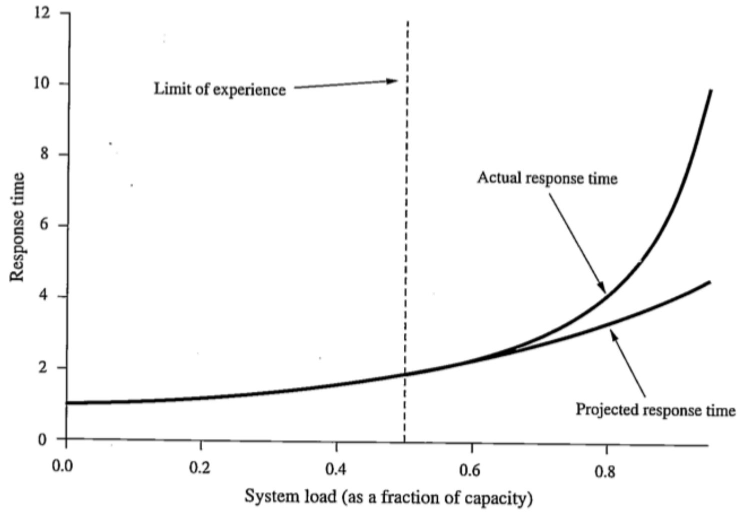


Escenario y aproximación

- Necesitamos calcular o predecir el comportamiento (rendimiento, *performance*)
- Nos basamos en la carga actual o la estimada para el nuevo escenario
- Opciones:
  - Medir en el nuevo escenario
    - Podemos crear una maqueta del nuevo escenario
    - Es complicado que el tráfico sea realista
    - Realista si incluimos a los usuarios
    - Pero entonces si el funcionamiento no les satisface estamos reaccionando ante el problema
  - (...)

Escenario y aproximación

- Necesitamos calcular o predecir el comportamiento (rendimiento, *performance*)
- Nos basamos en la carga actual o la estimada para el nuevo escenario
- Opciones:
  - Medir en el nuevo escenario
  - Hacer una predicción reescalando los valores actuales
    - Ejemplo: “Se duplica el tráfico, entonces se duplicarán las pérdidas”
    - El problema es que no suele ser tan simple cómo reaccionan los sistemas ante cambios
  - (...)



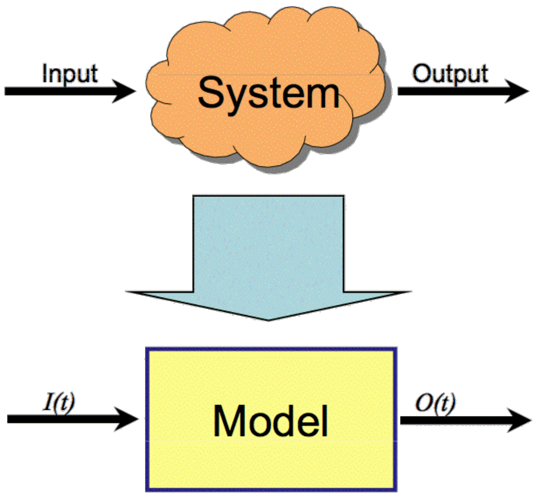
W.Stallings, “High-speed networks and Internets. Performance and quality of service”

Escenario y aproximación

- Necesitamos calcular o predecir el comportamiento (rendimiento, *performance*)
- Nos basamos en la carga actual o la estimada para el nuevo escenario
- Opciones:
  - Medir en el nuevo escenario
  - Hacer una predicción reescalando los valores actuales
  - Desarrollar un modelo del sistema (...)

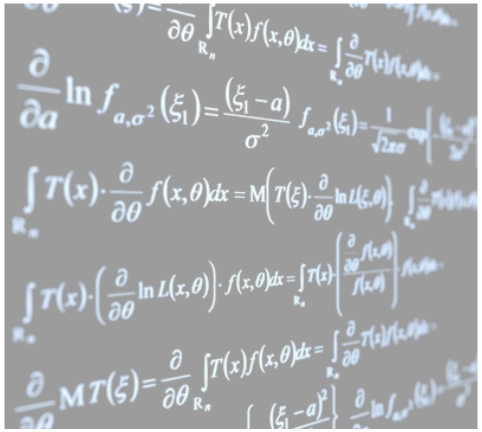
Modelo del sistema

- Representación de un sistema para estudiarlo
- Simplifica el sistema
- Considera solo los aspectos que afectan al problema en estudio
- Debe ser lo suficientemente detallado para poderse obtener conclusiones que apliquen al sistema real
- Vamos a ver algunos modelos
- Y la información que podemos extraer de ellos



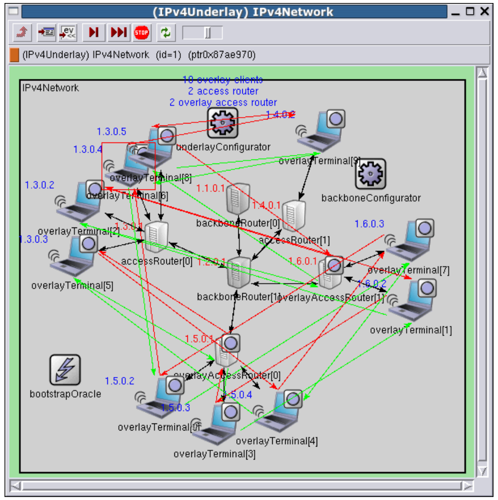
Escenario y aproximación

- Necesitamos calcular o predecir el comportamiento (rendimiento, *performance*)
- Nos basamos en la carga actual o la estimada para el nuevo escenario
- Opciones:
  - Medir en el nuevo escenario
  - Hacer una predicción reescalando los valores actuales
  - <u>Desarrollar un modelo analítico</u>
    - Ecuaciones que nos permitan calcular los parámetros deseados
    - Requieren modelos simplificados para el tráfico y el sistema
    - Si simplificamos demasiado pueden no ser útiles los resultados
  - (...)



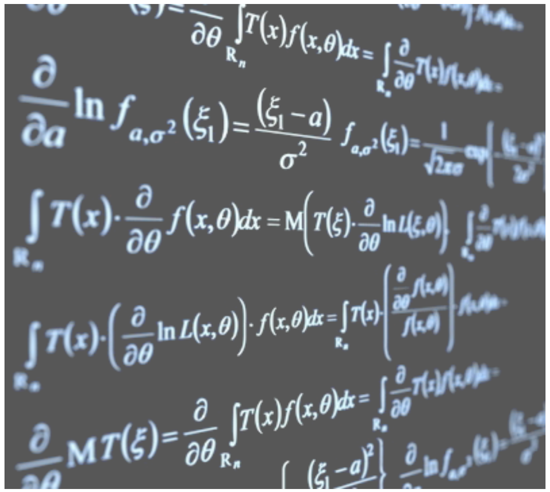
Escenario y aproximación

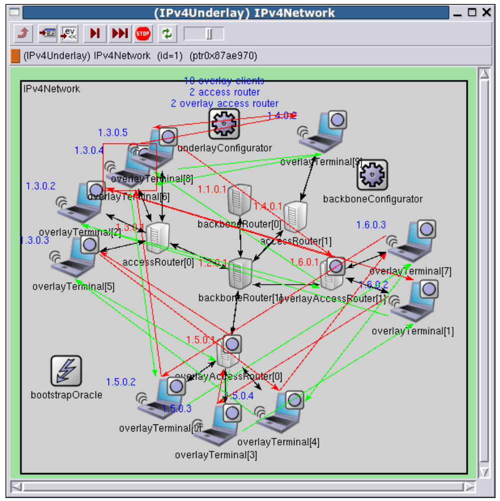
- Necesitamos calcular o predecir el comportamiento (rendimiento, *performance*)
- Nos basamos en la carga actual o la estimada para el nuevo escenario
- Opciones:
  - Medir en el nuevo escenario
  - Hacer una predicción reescalando los valores actuales
  - Desarrollar un modelo analítico
  - <u>Programar y ejecutar un modelo de simulación</u>
    - Un software simula el comportamiento del sistema con esos parámetros y medimos ahí
    - Seguimos necesitando buenos modelos
    - Es costoso en tiempo de desarrollo y de ejecución



En este tema

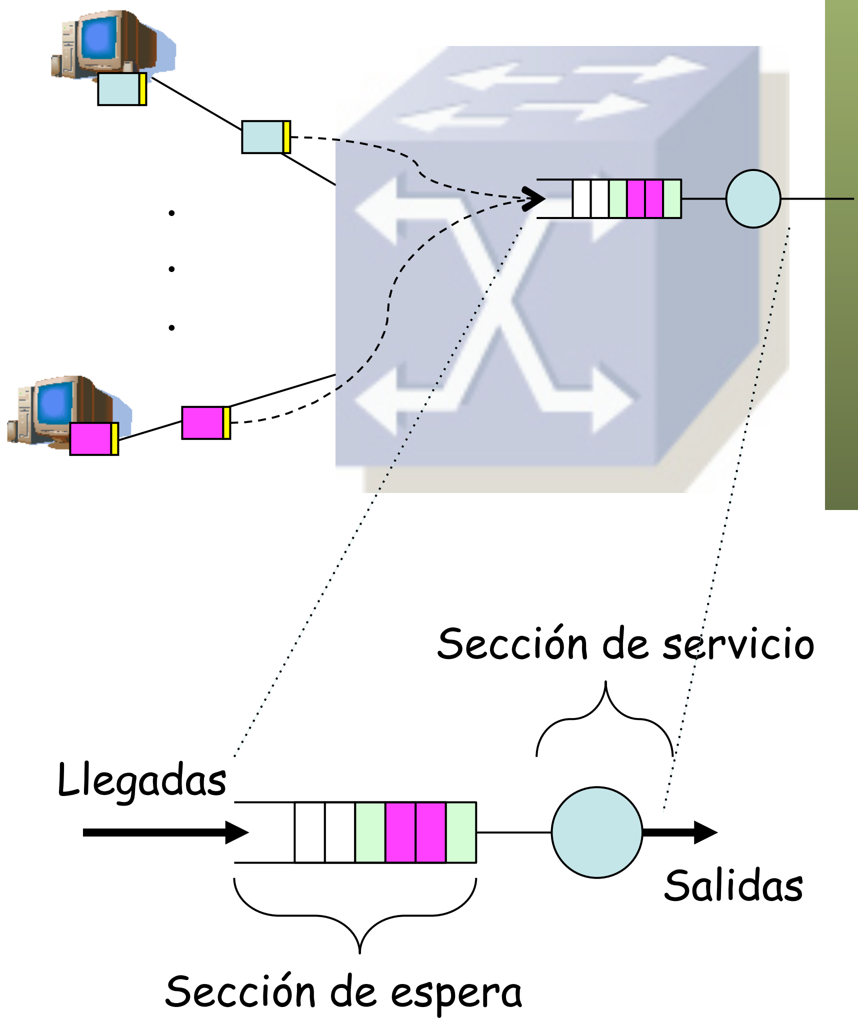
- Trataremos las dos últimas aproximaciones
  - Modelado analítico: teoría de colas
    - Ya vimos algo en ARSS: procesos de Poisson, Erlang-B
    - Fue aplicado en redes de conmutación de circuitos
    - Repasaremos y ampliaremos para conmutación de paquetes
  - Simulación: simulación de eventos discretos
    - Conceptos básicos en el desarrollo de simuladores
    - Aplicación en comparación con la teoría de colas





Ejemplo de pregunta

- ¿Cuál es el retardo que sufren los paquetes que atraviesan un conmutador?
- Simplificaciones:
- Tiempo de procesado despreciable
- La velocidad del enlace (out) es fija y conocida (no modulación variable)
- Si el enlace de salida está libre se puede transmitir (control de acceso al medio nulo): retardo de transmisión = tamaño/velocidad
- Si el interfaz de salida está ocupado se queda el paquete en espera (cola)
- Cola FIFO de gran tamaño
- Sabemos cuándo llegan los paquetes y de qué tamaño son
- Queremos conocer el retardo medio que sufren los paquetes
- O el valor de retardo tal que solo x% de los paquetes lo exceden



Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

Gestión y Planificación de Redes y Servicios
Área de Ingeniería Telemática

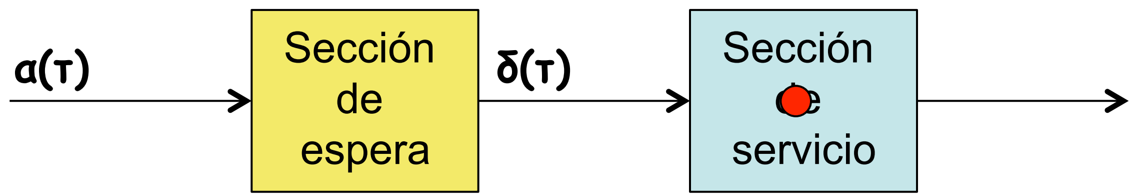
Terminología en el comportamiento de una cola

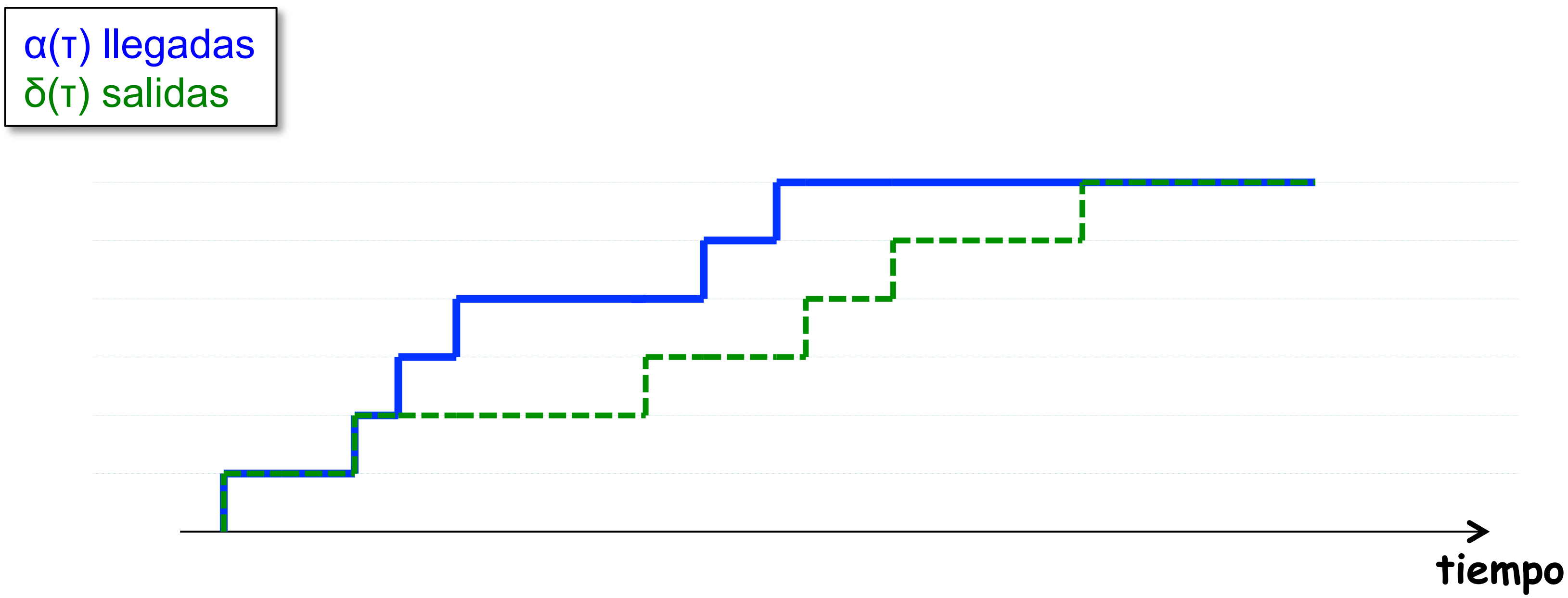
Funcionamiento del sistema

- Recordemos el funcionamiento del sistema con cola
- Y la implicación que tiene sobre las medias de:
  - Tiempo en el sistema
  - Tasa media de llegadas
  - Tiempo medio de servicio

Sistema con cola

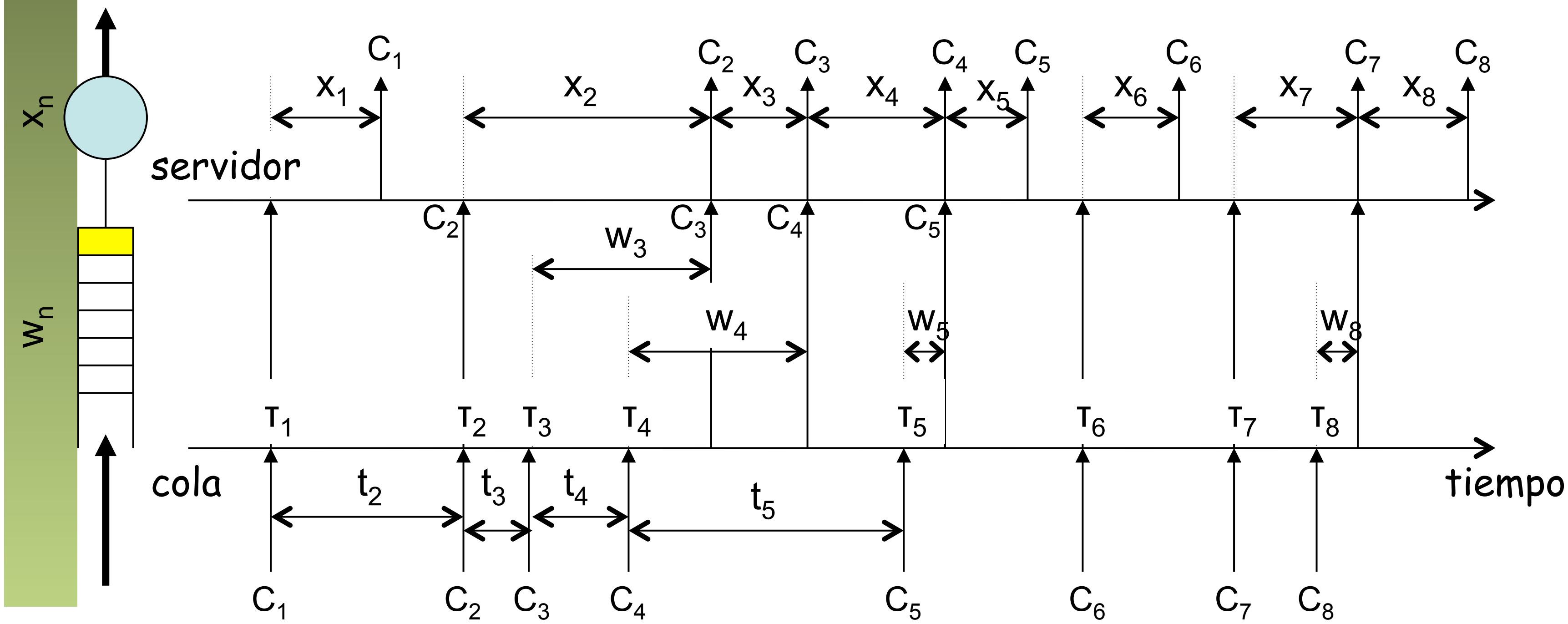
- A medida que los clientes terminan en la sección de servicio otros nuevos de la de espera pasan inmediatamente a ella (. . .)





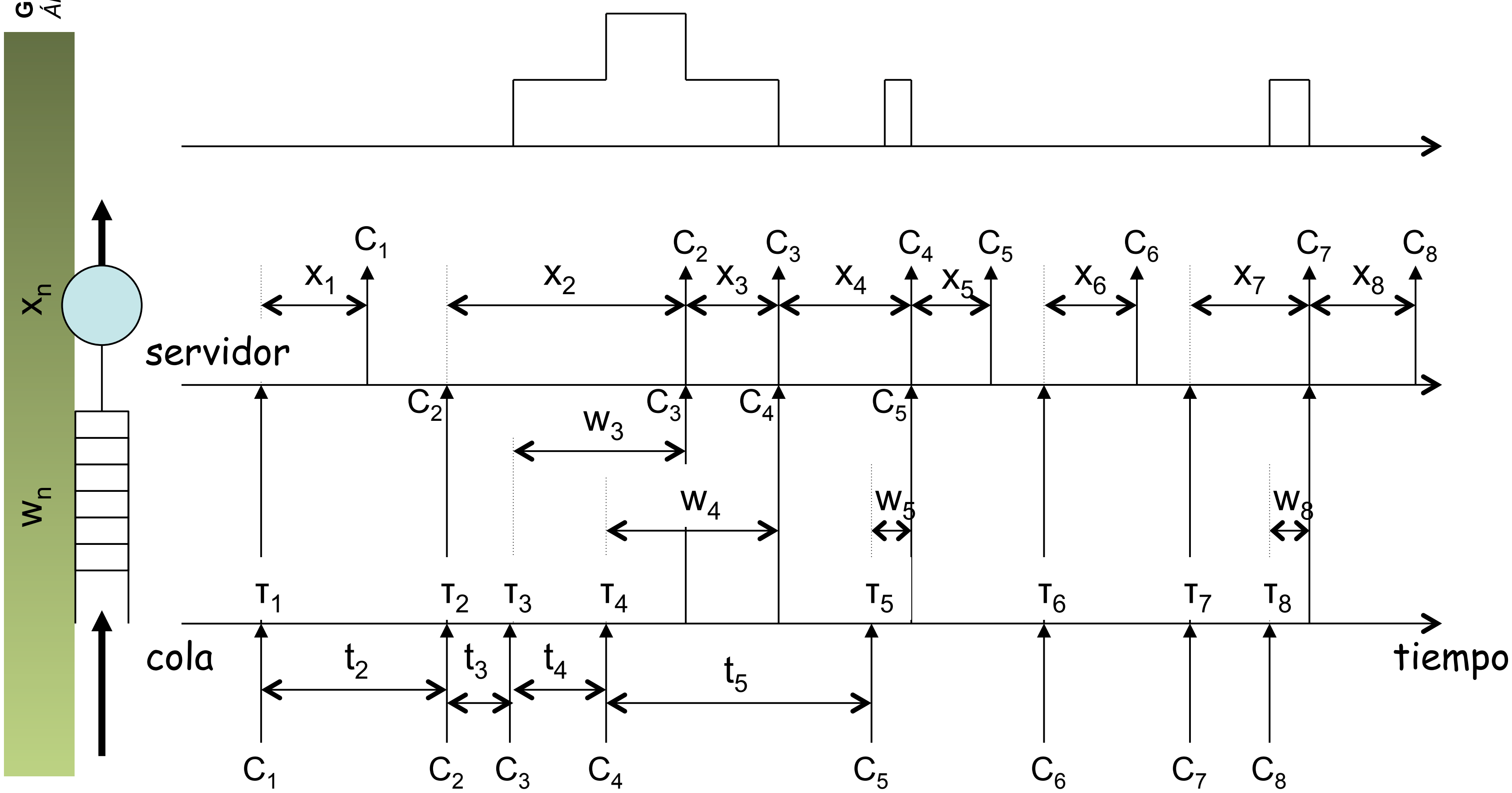
Llegadas y cola

- C_n : cliente n-ésimo del sistema
- τ_n : instante de llegada del cliente n-ésimo
- t_n : tiempo entre las llegadas del cliente n y n-1 = $\tau_n - \tau_{n-1}$
- w_n : Tiempo de espera en cola del cliente C_n , con media $E[w_n] = W$
- x_n : Tiempo de servicio del cliente C_n , con media $E[x_n] = 1/\mu$
- s_n : Tiempo en el sistema del Cliente C_n , $s_n = w_n + x_n$



Llegadas y cola

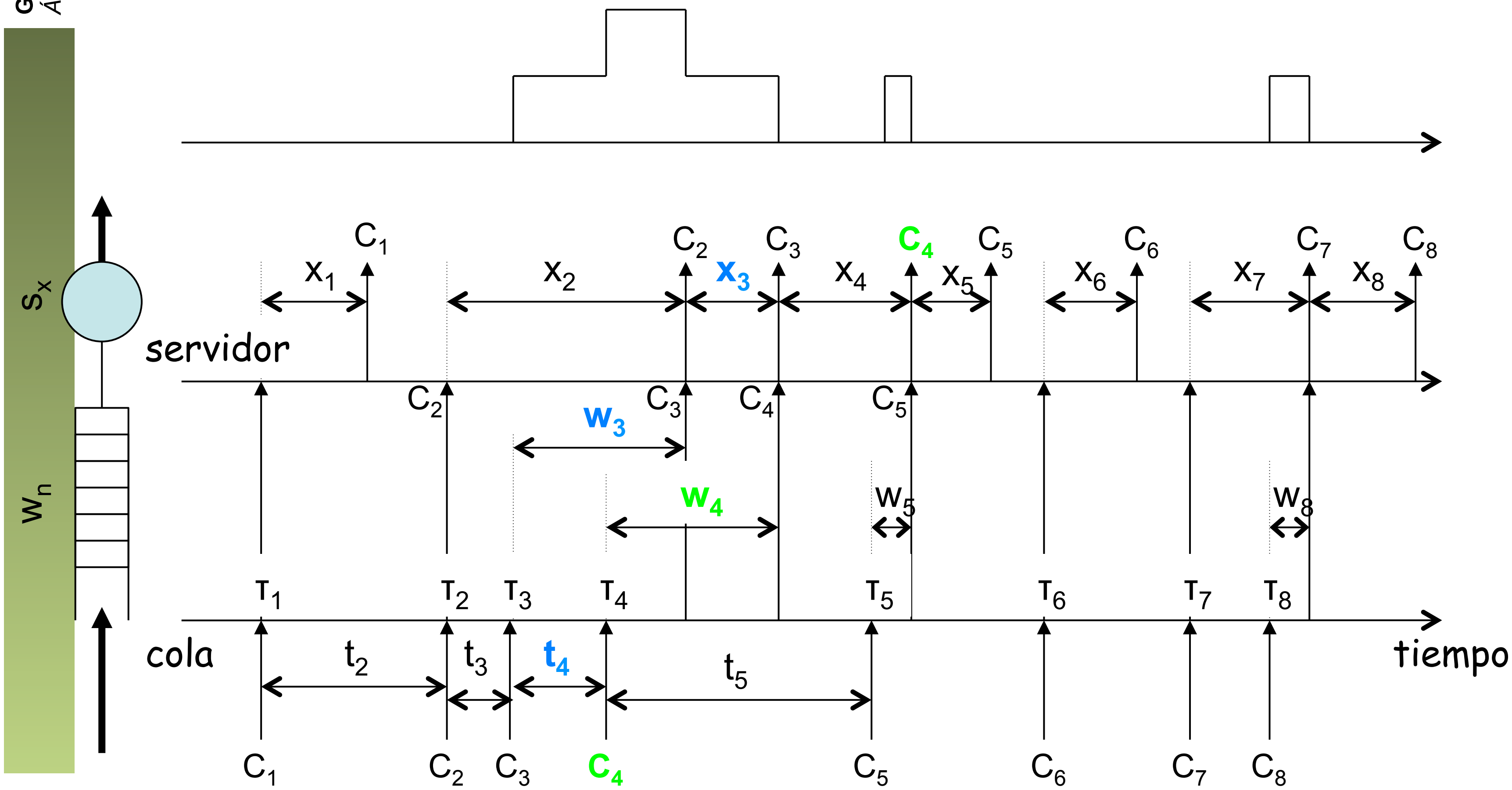
- Clientes en cola



Llegadas y cola

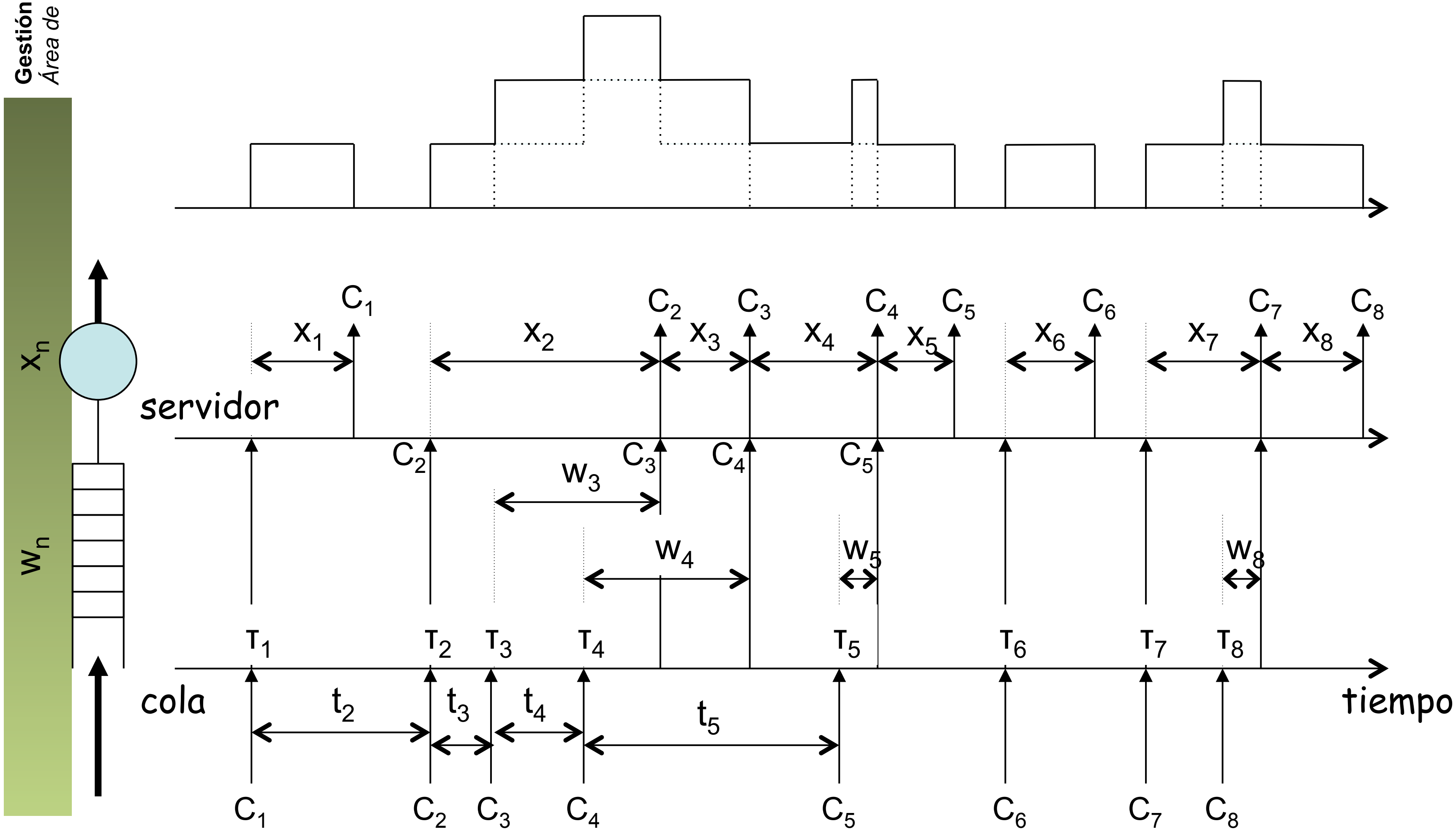
- Clientes en cola
- Tiempo de espera en cola del cliente C_{n+1}

$$w_{n+1} = \begin{cases} w_n + x_n - t_{n+1} & \text{if } w_n + x_n - t_{n+1} \geq 0, \\ 0 & \text{if } w_n + x_n - t_{n+1} < 0. \end{cases}$$



Llegadas y cola

- Clientes en cola
- Clientes en el sistema (cola+servidor)



Fórmula de Little

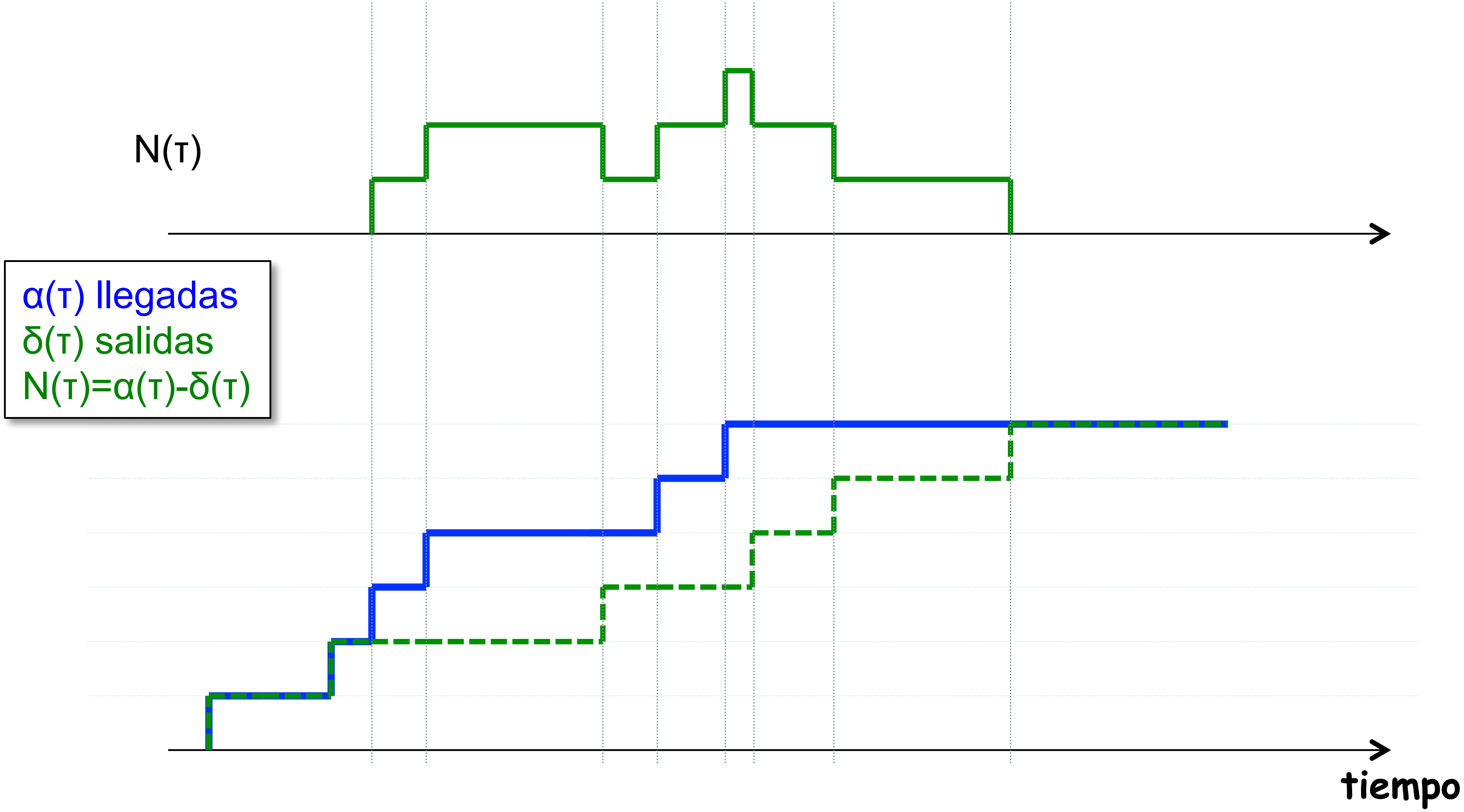
Fórmula de Little

- Es el resultado más general conocido
- Intuitivamente:
  - Tomamos una persona que llega a una cola
  - Vemos el tiempo que tarda el llegar a la cabeza de la cola (tiempo de espera)
  - ¿En ese tiempo cuántos clientes más llegan?

$$\overline{N} = \lambda \overline{W}$$

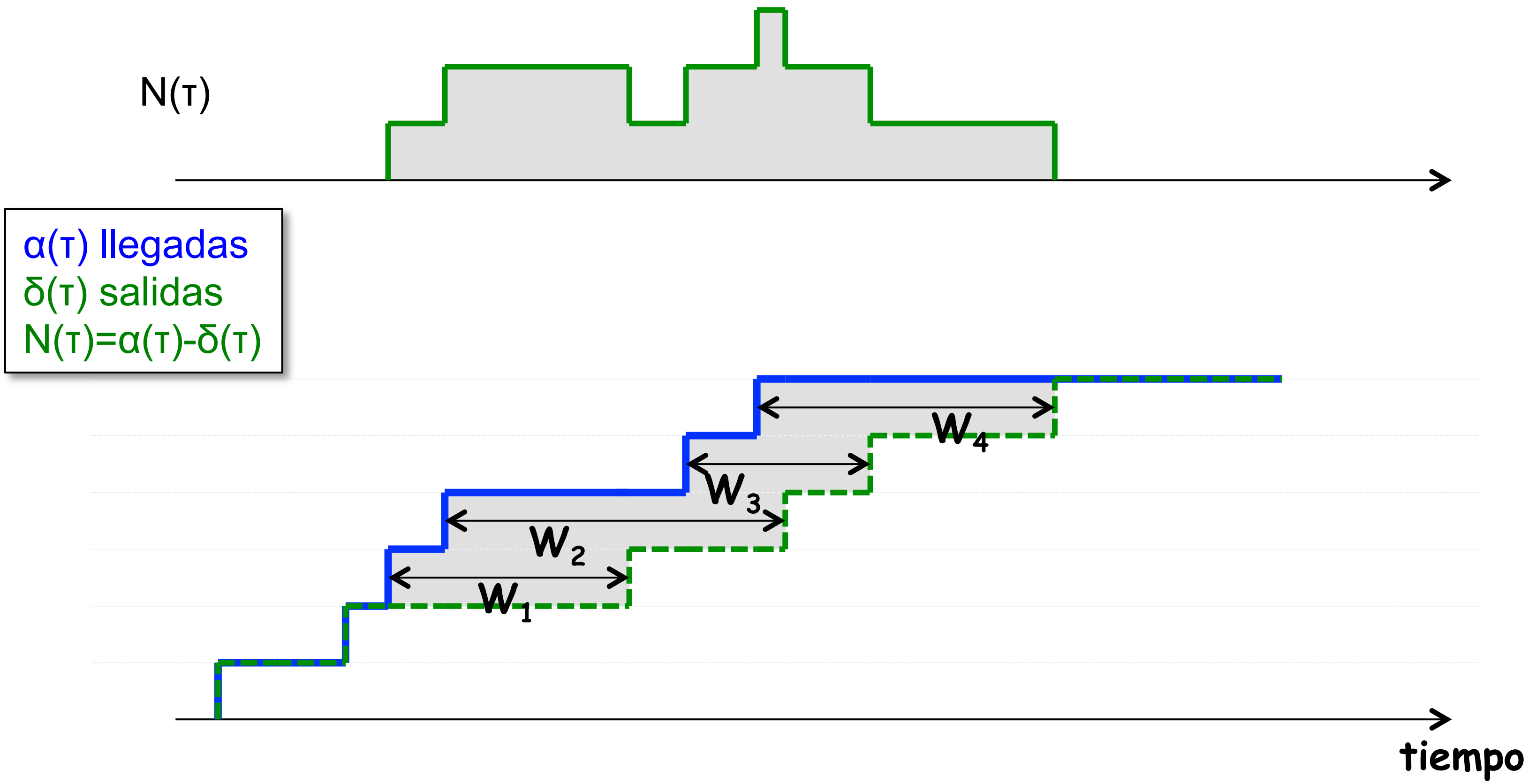
Fórmula de Little

- Volvemos al ejemplo anterior
- Suponemos que es conservativo: todos los clientes que llegan son atendidos
- $N(\tau) = \alpha(\tau) - \delta(\tau)$: número de usuarios en espera en el instante τ



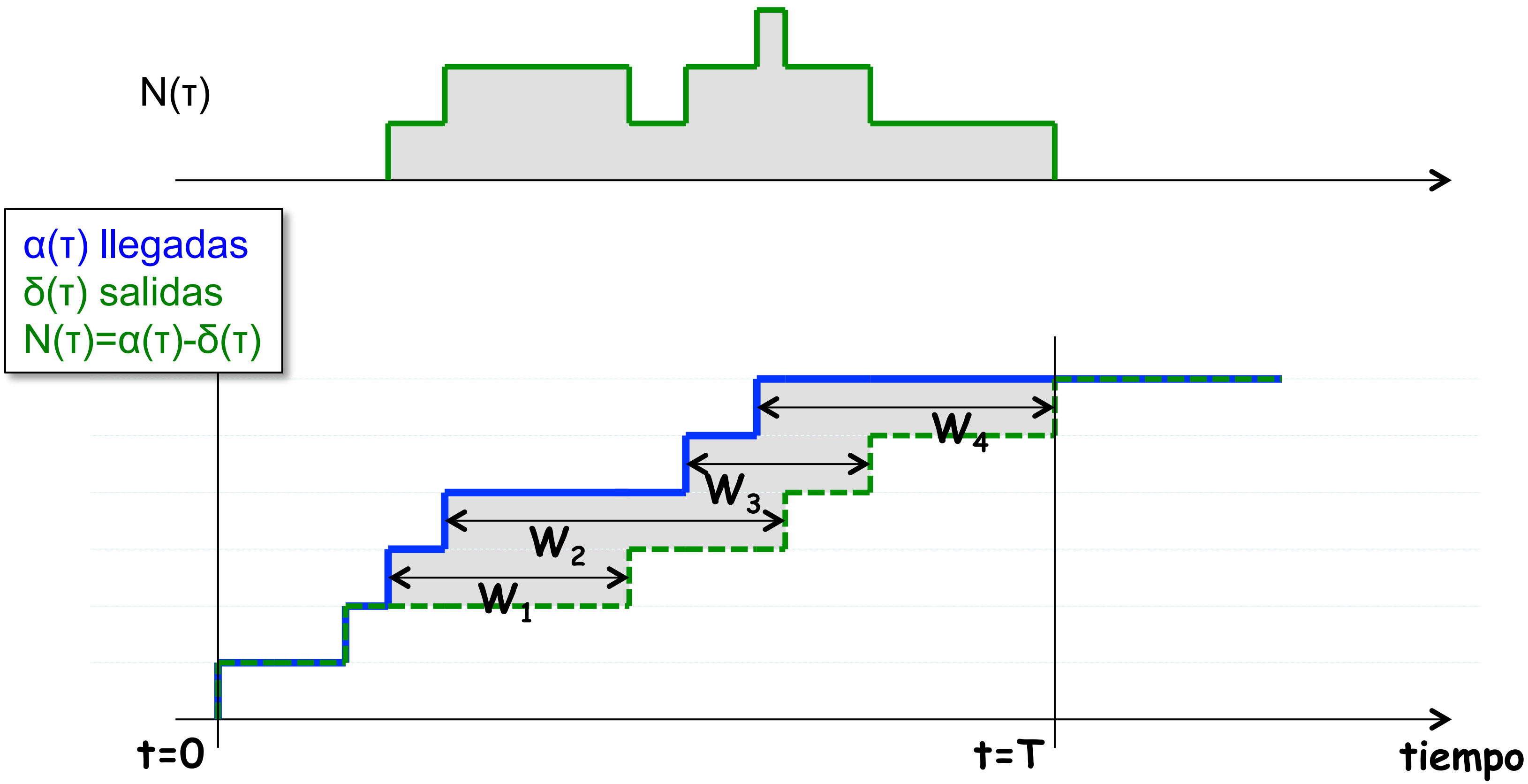
Fórmula de Little

- W_i son tiempos durante los cuales algún cliente estuvo esperando
- En el dibujo los W_i suponiendo FCFS



Fórmula de Little

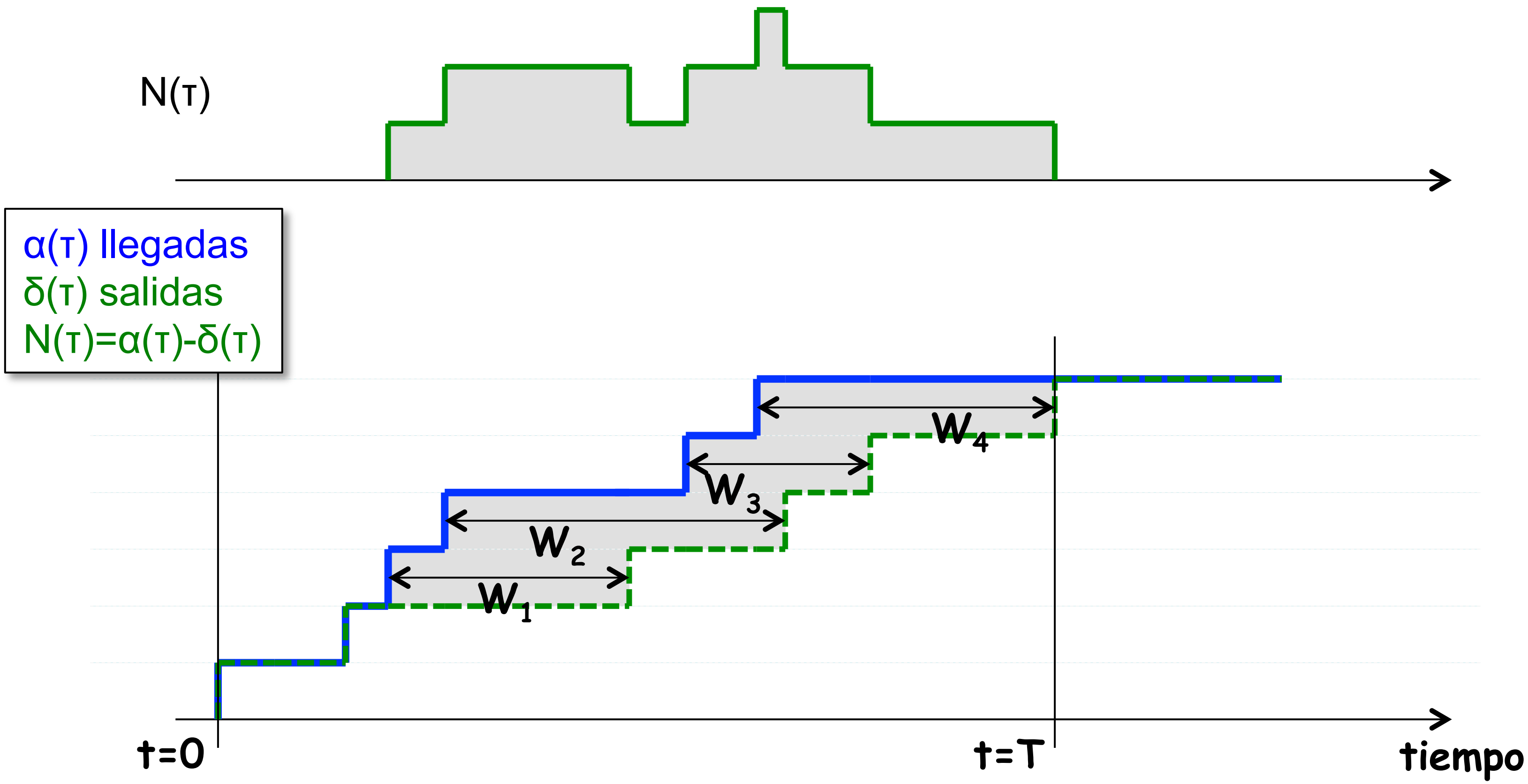
- W_i son tiempos durante los cuales algún cliente estuvo esperando
- Consideramos dos instantes en los que $\alpha(\tau)=\delta(\tau)$
- Por ejemplo $\tau=0$ y $\tau=T$



Fórmula de Little

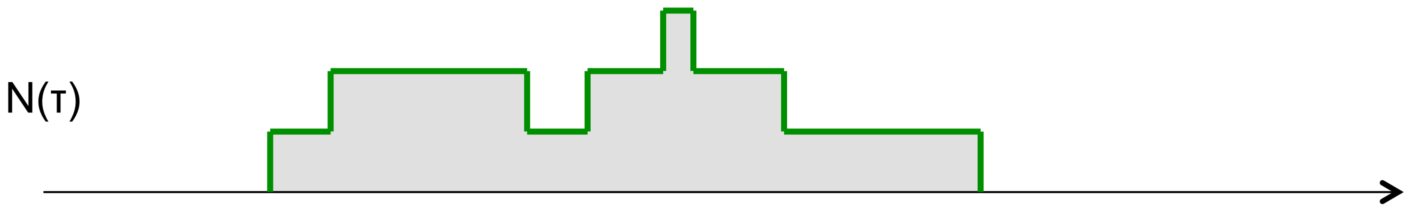
- El número de llegadas en ese intervalo es: n(T) = α(T) – α(0)
- El número *medio* de llegadas por unidad de tiempo en él es:

$$\lambda(T) = \frac{n(T)}{T}$$

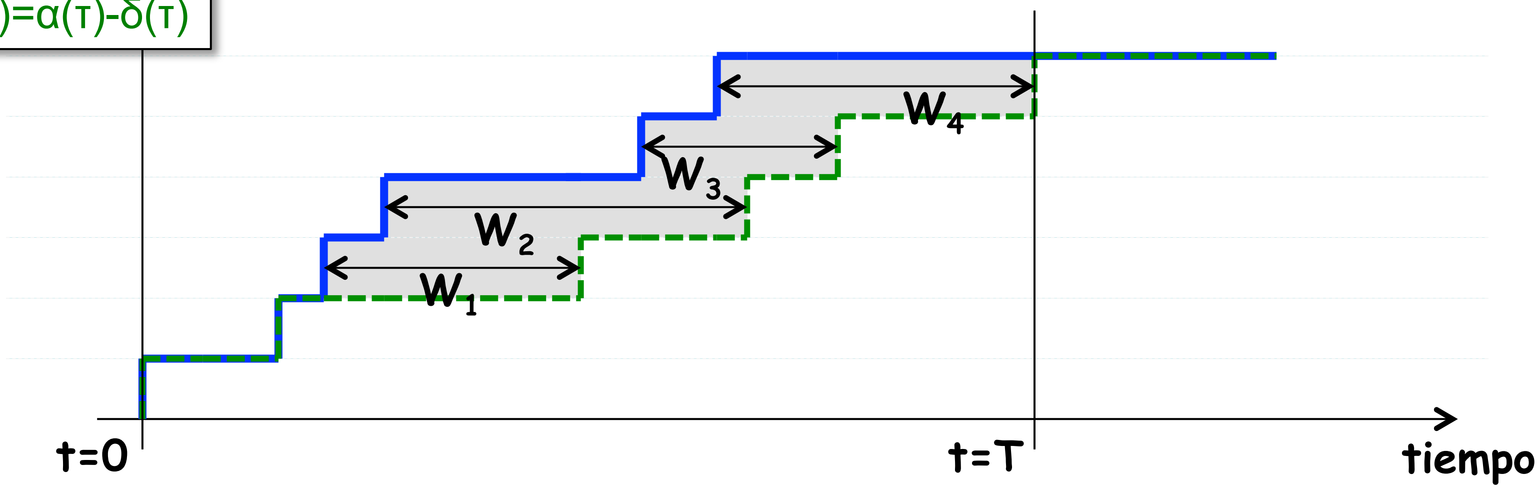


Fórmula de Little

- El área sombreada es: $\int_0^T N(t)\,dt = \sum_{j=1}^{n(T)} W_j$
- El tiempo medio de espera en ese intervalo es: $\overline{W}(T) = \frac{\sum_{j=1}^{n(T)} W_j}{n(T)}$
- Y el número medio de usuarios en él es: $\overline{N}(T) = \frac{\int_0^T N(t)\,dt}{T}$



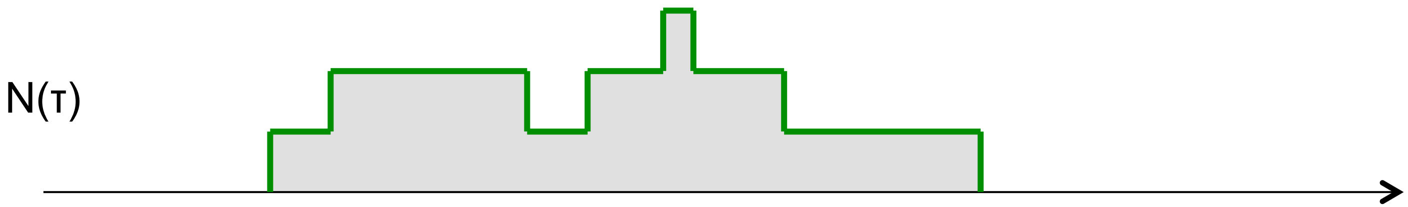
α(τ) llegadas
δ(τ) salidas
N(τ)=α(τ)-δ(τ)

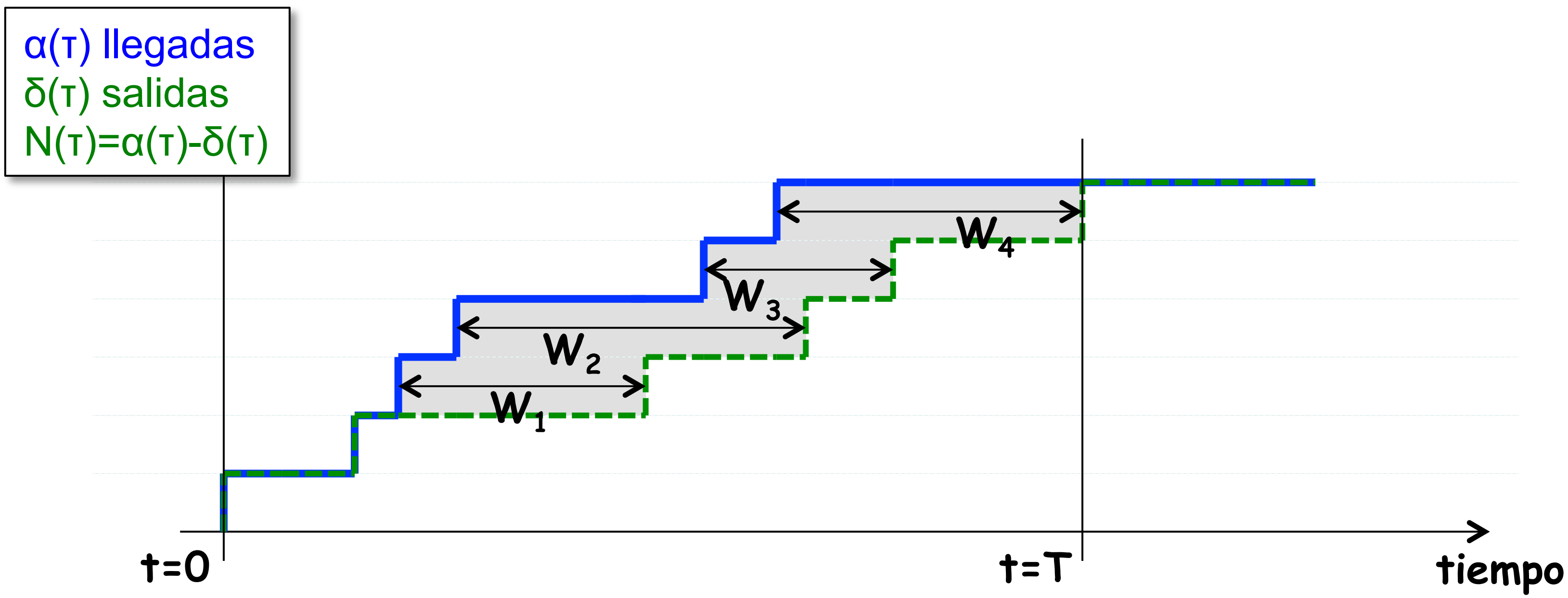


Fórmula de Little

$$\lambda(T) = \frac{n(T)}{T} \qquad \int_0^T N(t)\,dt = \sum_{j=1}^{n(T)} W_j \qquad \overline{N}(T) = \frac{\int_0^T N(t)\,dt}{T} \qquad \overline{W}(T) = \frac{\sum_{j=1}^{n(T)} W_j}{n(T)}$$

$$\overline{N}(T) =$$

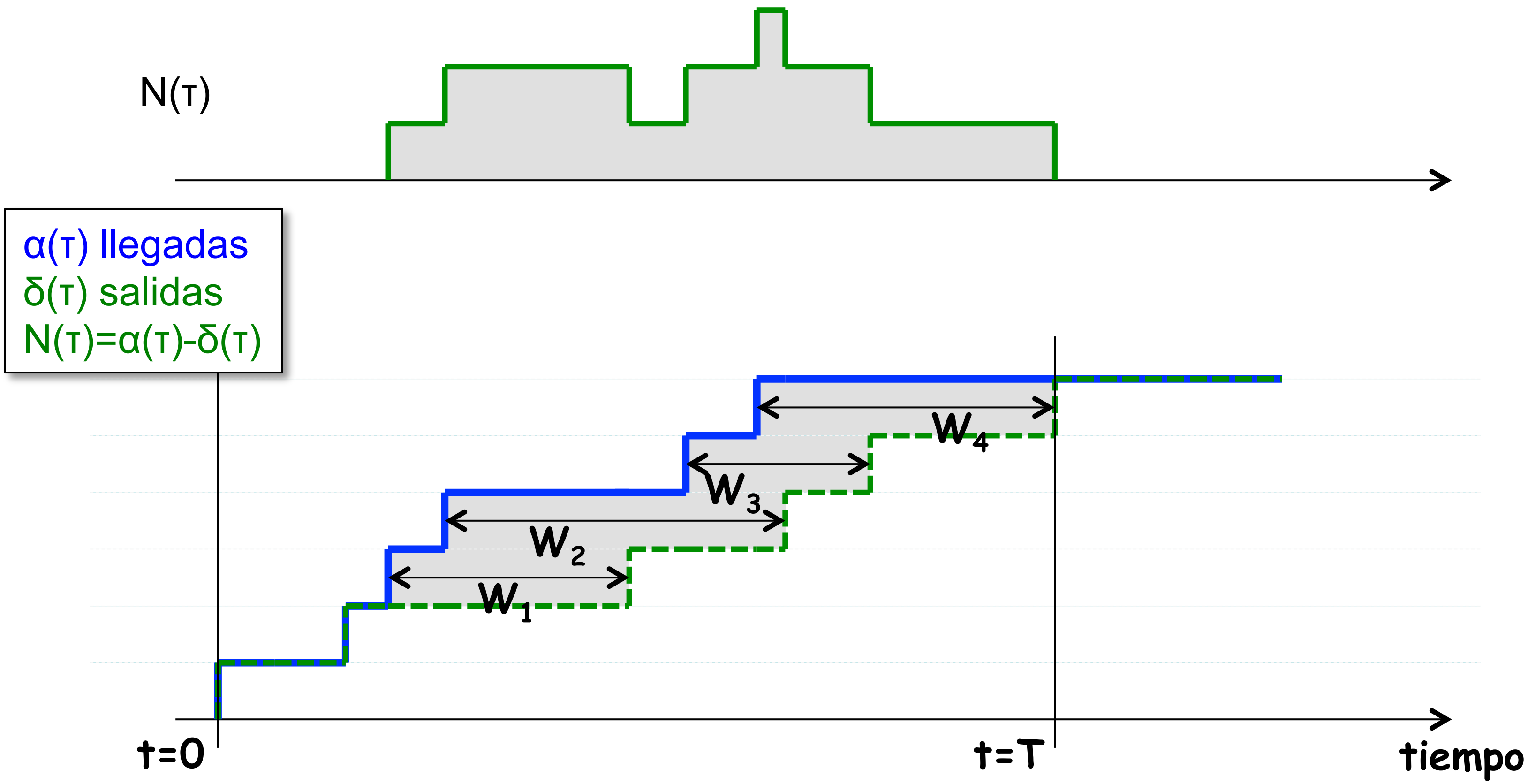




Fórmula de Little

$$\lambda(T) = \frac{n(T)}{T} \qquad \int_0^T N(t)\,dt = \sum_{j=1}^{n(T)} W_j \qquad \overline{N}(T) = \frac{\int_0^T N(t)\,dt}{T} \qquad \overline{W}(T) = \frac{\sum_{j=1}^{n(T)} W_j}{n(T)}$$

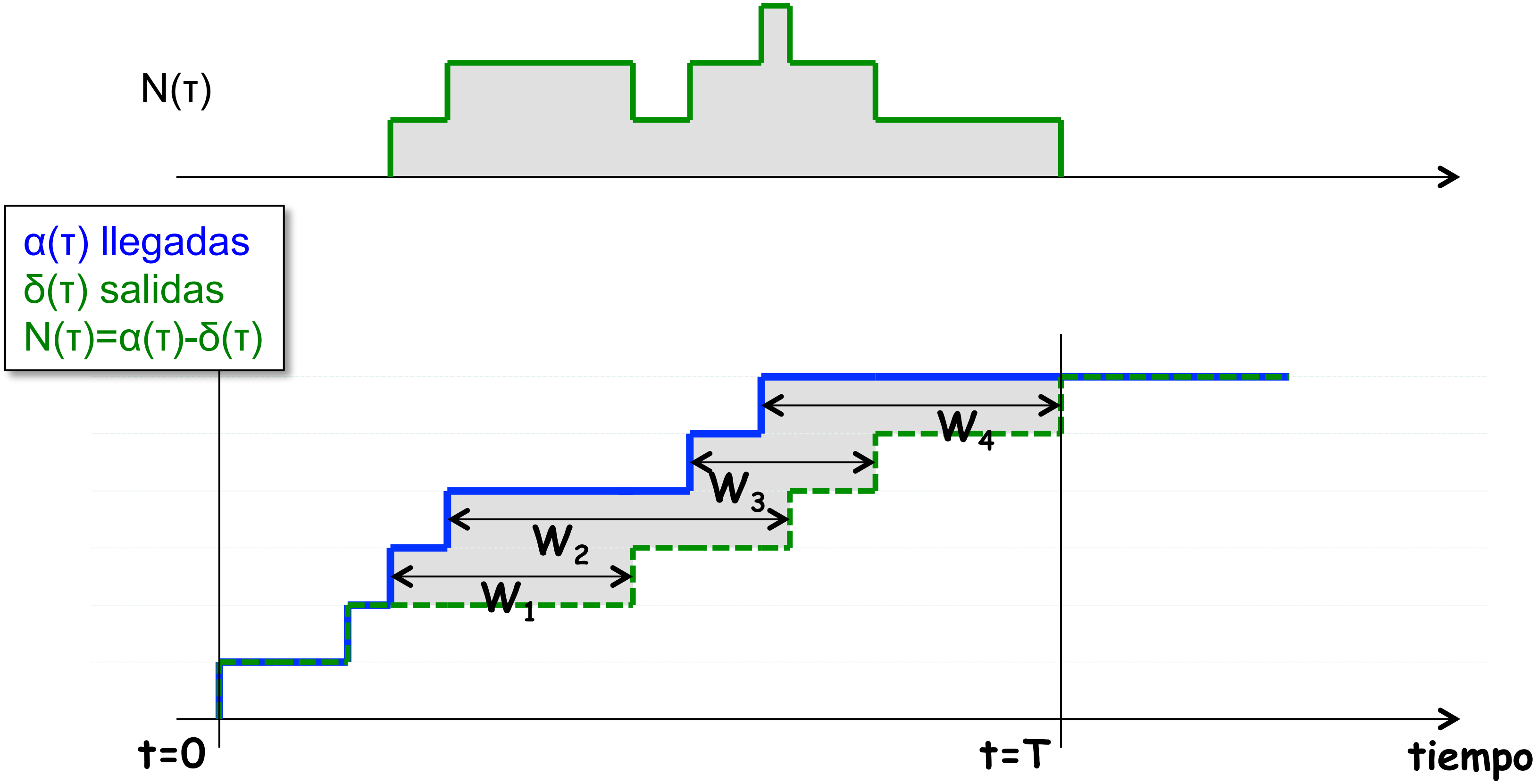
$$\overline{N}(T) = \frac{\int_0^T N(t)\,dt}{T} = \frac{\sum_{j=1}^{n(T)} W_j}{T} = \frac{n(T)\overline{W}(T)}{T} = \lambda(T)\overline{W}(T)$$



Fórmula de Little

$$\lambda(T) = \frac{n(T)}{T} \qquad \int_0^T N(t)\,dt = \sum_{j=1}^{n(T)} W_j \qquad \overline{N}(T) = \frac{\int_0^T N(t)\,dt}{T} \qquad \overline{W}(T) = \frac{\sum_{j=1}^{n(T)} W_j}{n(T)}$$

$$\overline{N}(T) = \lambda(T)\overline{W}(T) \xrightarrow[T \to \infty]{} \boxed{\overline{N} = \lambda \overline{W}}$$

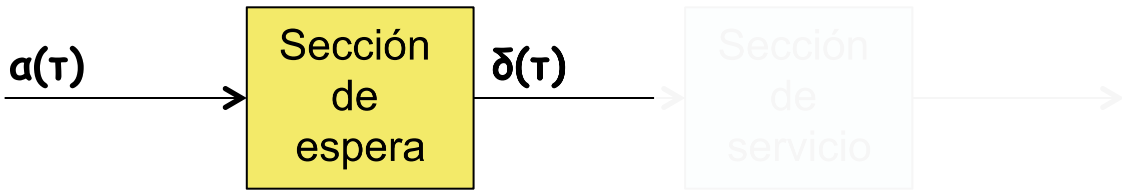


Fórmula de Little

- Número medio de usuarios en el sistema = tasa media de llegadas multiplicada por el tiempo medio de espera

$$\overline{N} = \lambda \overline{W}$$

- Demostrado para FIFO pero válido para cualquier política de servicio
- El "sistema" puede englobar cualquier número de elementos
- Podría ser por ejemplo solamente la sección de espera (…)

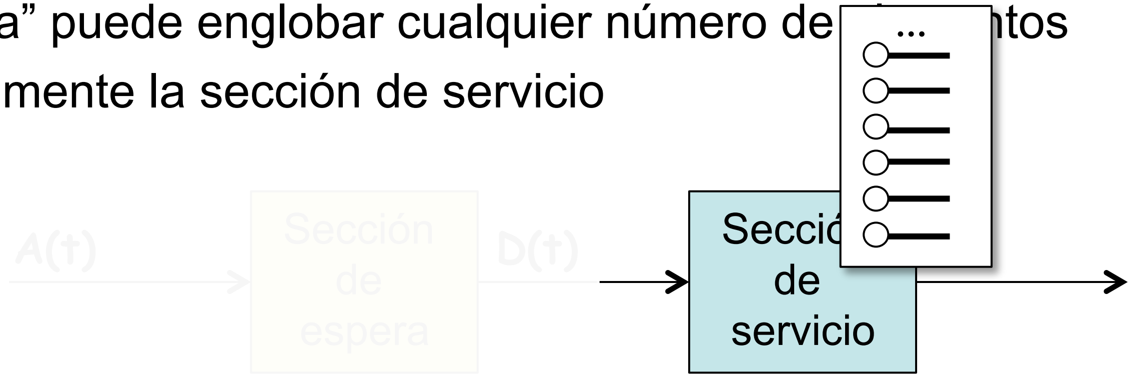


Fórmula de Little

- Número medio de usuarios en el sistema = tasa media de llegadas multiplicada por el tiempo medio de espera

$$\overline{N} = \lambda \overline{W}$$

- Demostrado para FIFO pero válido para cualquier política de servicio
- El “sistema” puede englobar cualquier número de [illegible]ntos
- O ser solamente la sección de servicio



- Supongamos que la sección de servicio es un número “infinito” de servidores y no hay cola
- La fórmula de Little nos dice que el número medio de servidores en uso (número medio de clientes en el sistema) es igual a la tasa media de llegadas multiplicada por el tiempo medio de servicio
- Es decir, igual a la intensidad de tráfico media $I = \lambda x$

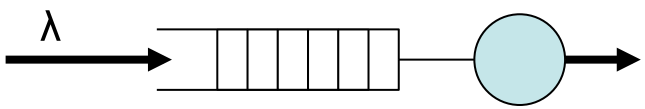
Factor de utilización

Factor de utilización

- Supongamos un sistema con un solo servidor
- Definimos el factor de utilización como el cociente entre la velocidad de llegada de "trabajo" y la máxima capacidad de llevarlo a cabo

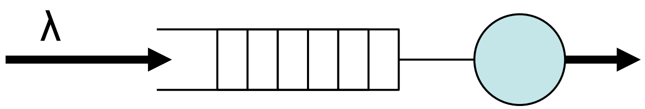
- Un cliente trae en media x unidades de trabajo
- En media llegan λ clientes por unidad de tiempo
- Luego en media llegan λx unidades de trabajo por unidad de tiempo

- Un servidor puede cursar un máximo de 1 unidad de trabajo por unidad de tiempo
- El factor de utilización entonces es simplemente $\rho = \lambda x$
- (...)



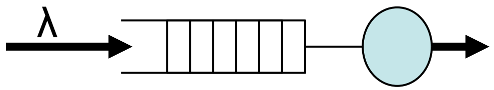
Factor de utilización

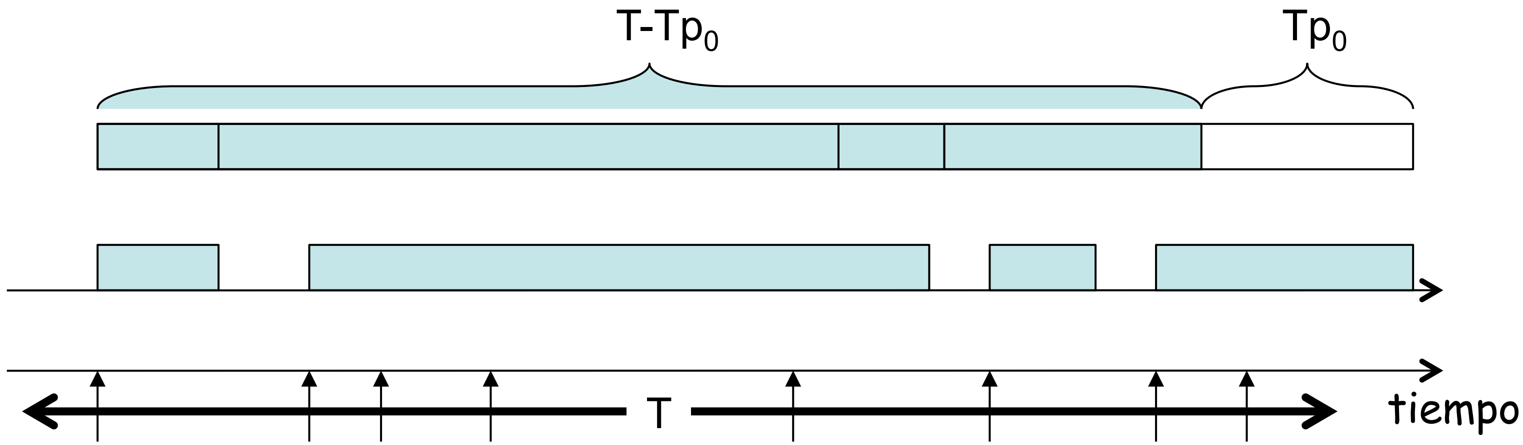
- Un cliente trae en media x unidades de trabajo
- En media llegan λ clientes por unidad de tiempo
- Ejemplo:
  - Las “unidades de trabajo” son bits y las de tiempo son segundos
  - Es decir en media llegan λ clientes (paquetes) por segundo y son de x bits
  - En media llegan λx bits/s
  - El servidor es capaz de servir C bits/s
  - El factor de utilización sigue siendo el cociente $\rho = \lambda x/C$ que no es más que el cociente de tasas en bits/s



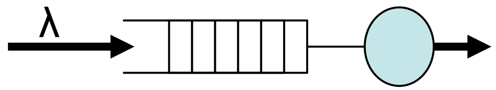
Factor de utilización

- Supongamos un sistema con un solo servidor
- *El factor de utilización es $\rho = \lambda x$*
- Sea T un intervalo de tiempo grande
- En él tendremos aproximadamente λT llegadas
- Sea p_0 la probabilidad de que el servidor esté desocupado en un instante al azar
- Durante el intervalo T el servidor estará desocupado durante Tp_0
- Y estará ocupado durante $T-Tp_0$
- Si el tiempo medio de servicio de un cliente es x
- El número de clientes servidos en el intervalo T será $(T-Tp_0)/x$
- (...)





Factor de utilización



- Supongamos un sistema con un solo servidor
- *El factor de utilización es $\rho = \lambda x$*
- Sea T un intervalo de tiempo grande
- En él tendremos aproximadamente λT llegadas
- Sea p_0 la probabilidad de que el servidor esté desocupado en un instante al azar
- Durante el intervalo T el servidor estará desocupado durante Tp_0
- Y estará ocupado durante $T-Tp_0$
- Si el tiempo medio de servicio de un cliente es x
- El número de clientes servidos en el intervalo T será $(T-Tp_0)/x$
- Si el sistema es estable, en un intervalo grande el número de llegadas debe ser aproximadamente igual al de salidas:

$$\lambda T \approx (T-Tp_0)/x \text{ , es decir } \lambda x \approx (T-Tp_0)/T = 1-p_0$$

- Es decir, para $T \rightarrow \infty$ tenemos que: $\rho = 1 - p_0$
- Es decir, el factor de utilización es la fracción del tiempo que el servidor está ocupado