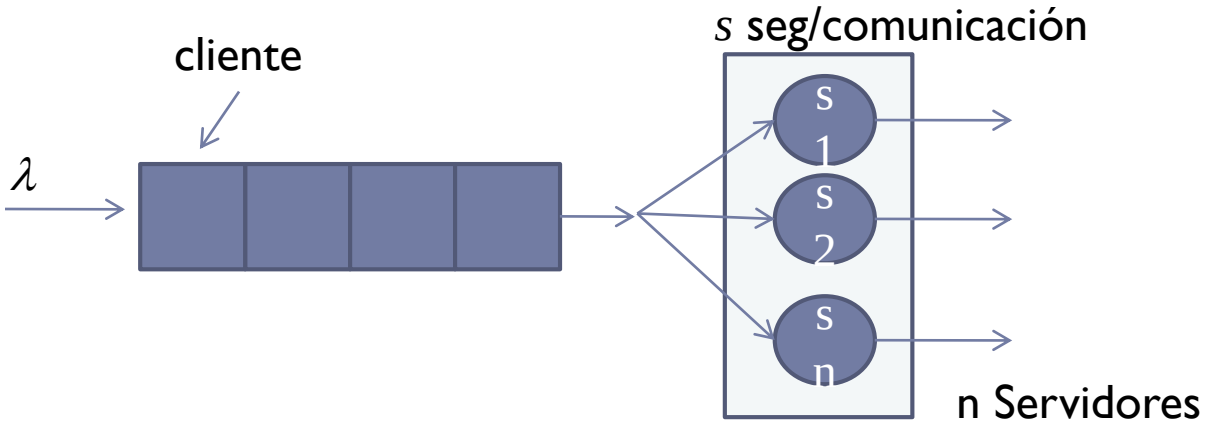
Sistemas de Espera

Jhon Jairo Padilla A., PhD.

Introducción

- En los sistemas de espera se forma una cola cuando los clientes que llegan no pueden ser atendidos
- Pueden haber varios servidores o un solo servidor
- La cola puede ser atendida de diferentes maneras: FIFO (FCFS), LIFO (LCFS), PQ (Priority Queueing), etc.



Notación de Kendall

- Kendall (1951) introdujo una notación para los modelos de colas:

$$A/B/n$$

- Donde:
  - A: Distribución del Proceso de llegadas
  - B: Distribución del tiempo de servicio
  - n: Número de servidores

Notación para las distribuciones en la notación de Kendall

M	$\sim$	Markov. Exponential time intervals (Poisson arrival process, exponentially distributed service times).
D	$\sim$	Deterministic. Constant time intervals.
E_k	$\sim$	Erlang-k distributed time intervals ($\mathrm{E}_1 = \mathrm{M}$).
H_n	$\sim$	Hyper-exponential of order n distributed time intervals.
Cox	$\sim$	Cox-distributed time intervals.
Ph	$\sim$	Phase-type distributed time intervals.
GI	$\sim$	General Independent time intervals, renewal arrival process.
G	$\sim$	General. Arbitrary distribution of time intervals (may include correlation).

Ejemplos: Notación de Kendall

- M/M/n:
  - M: Proceso de llegadas de Poisson
  - M: Tiempos de servicio distribuidos exponencialmente
  - n: n servidores
- GI/G/1: Sistema de espera general con un solo servidor

Notación de Kendall ampliada

- La notación completa debería introducir otras descripciones:

$$A/B/n/K/S/X$$

- Donde:
  - K: Capacidad total del sistema (ó número de posiciones de espera)
  - S: Tamaño de la población (número de clientes). Puede ser infinito
  - X: Disciplina de cola

Disciplinas de colas

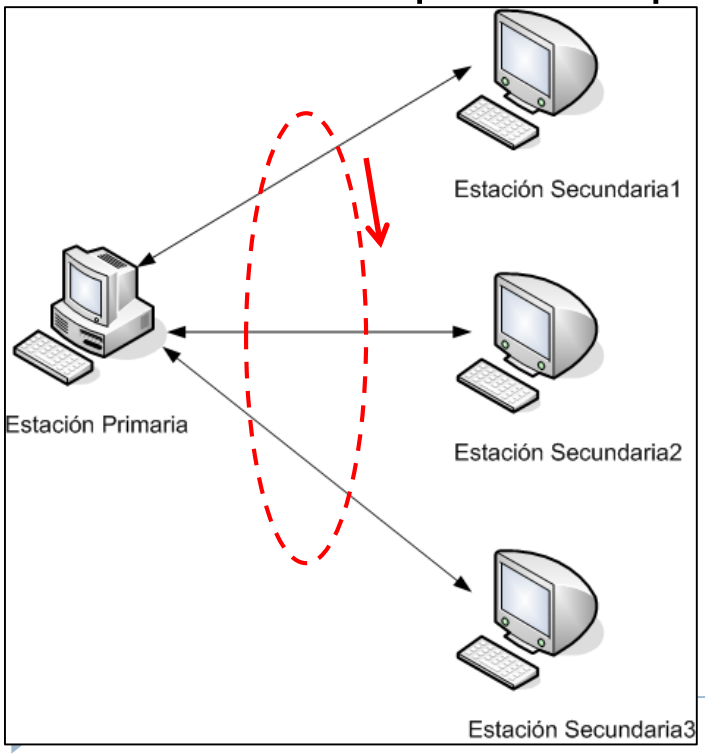
- Los clientes que esperan en cola pueden ser seleccionados para ser servidos de acuerdo a diferentes principios:
  - FCFS: First Come- First Served
    - Es llamada también cola ordenada.
    - También se conoce como FIFO: First In- First Out
  - LCFS: Last Come- First Served
    - Opera como una pila (Stack).
    - Se usa en almacenamientos.
    - También se conoce como LIFO: Last In- First Out
  - SIRO: Service in Random Order
    - Todos los clientes que esperan en cola tienen la misma probabilidad de ser elegidos para ser servidos.
    - También es llamada RS (Random Selection)

Disciplinas de colas

- Otras disciplinas toman en cuenta el tiempo de servicio como criterio:
  - SJF: Shortest Job First
    - También conocida como SJN (Shortest Job Next) ó SPF (Shortest Processing Time First).
    - Esta disciplina asume que conoce el tiempo de servicio con anticipación y minimiza el tiempo de espera total para todos los clientes.

Disciplinas de colas

- Otras disciplinas importantes:
  - RR: Round Robin
    - Se asigna a los clientes un tiempo de servicio fijo (ranura de tiempo). Si el servicio no se completa durante este intervalo, el cliente regresa a la cola, que es de tipo FCFS.

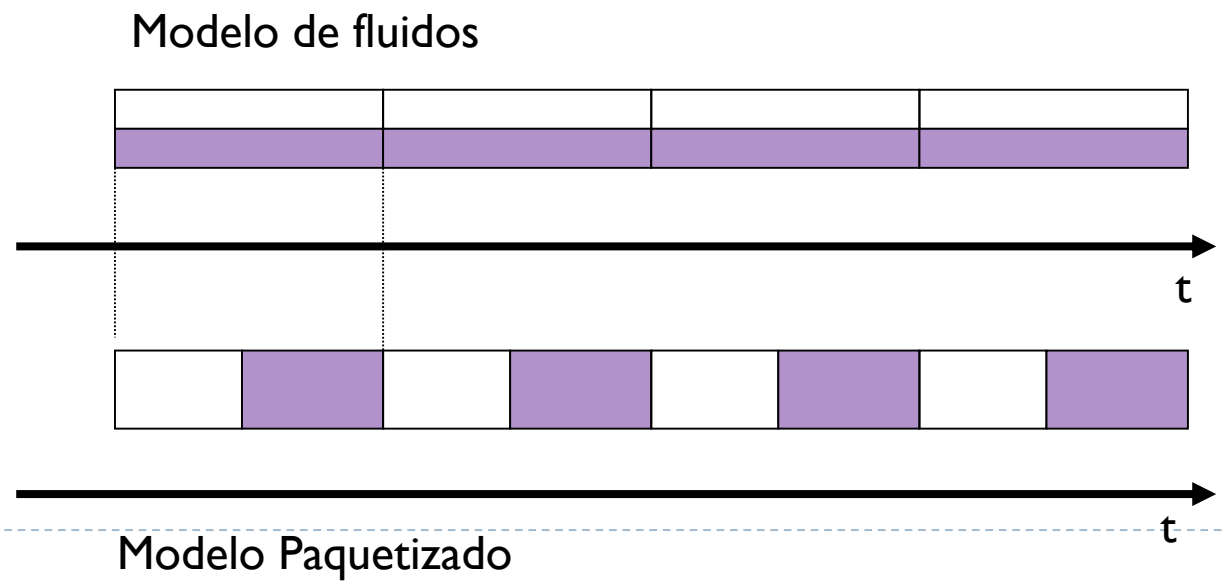


Disciplinas de colas

- PS: Procesor Sharing
  - Todos los clientes comparten la capacidad del servicio igualmente
- FB: Foreground-Background
  - Intenta implementar SJF sin conocer el tiempo de servicio. El servidor ofrecerá servicio al cliente que haya recibido la menor cantidad de servicio. Cuando todos los clientes han obtenido la misma cantidad de servicio, FB se comporta idéntico a PS.

Disciplinas de colas

- GPS: Generalized Processor Sharing
  - Es una variante de PS en que a los usuarios comparten la capacidad del sistema de forma que se les asigna una parte proporcional de la capacidad máxima según un peso pre-establecido con anterioridad.
  - Se conoce también como WFQ (Weighted Fair Queueing)

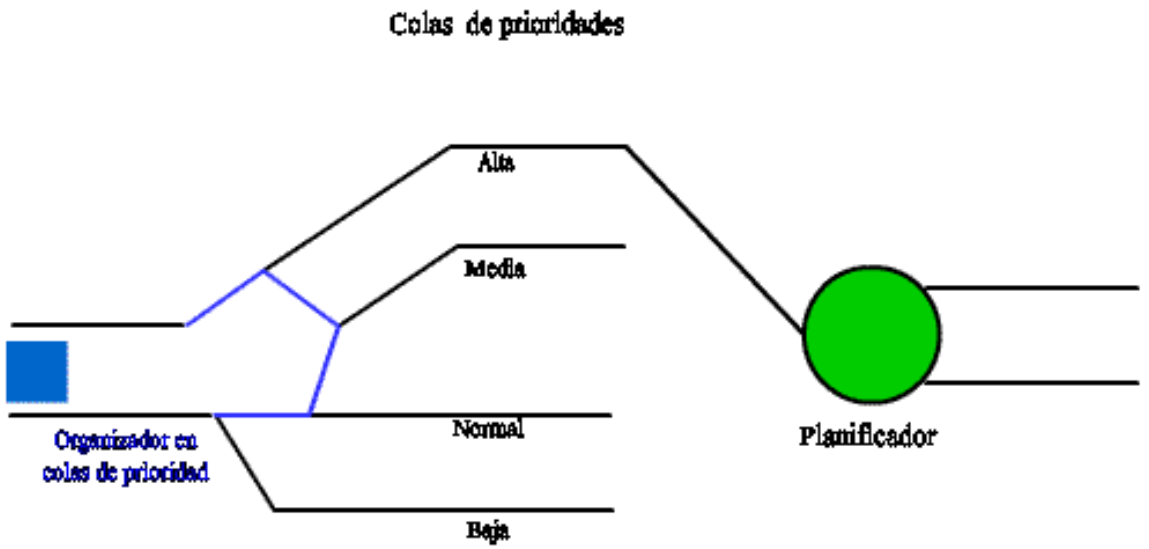


Disciplinas de clientes con prioridad

- En la vida real, los clientes suelen dividirse en N clases con diferente prioridad
- Un cliente perteneciente a la clase p tiene una más alta prioridad que un cliente perteneciente a la clase p+1.
- Hay dos tipos de prioridad:
  - Con derecho preferente (preemptive):
    - Un nuevo cliente con mayor prioridad puede interrumpir el servicio de un cliente de menor prioridad que está siendo atendido
  - Sin derecho preferente (Non-preemptive)
    - Un nuevo cliente con mayor prioridad debe esperar a que se atienda al cliente de menor prioridad que está siendo atendido. Incluso debe esperar a que se atienda a otros de su misma prioridad que llegaron antes.

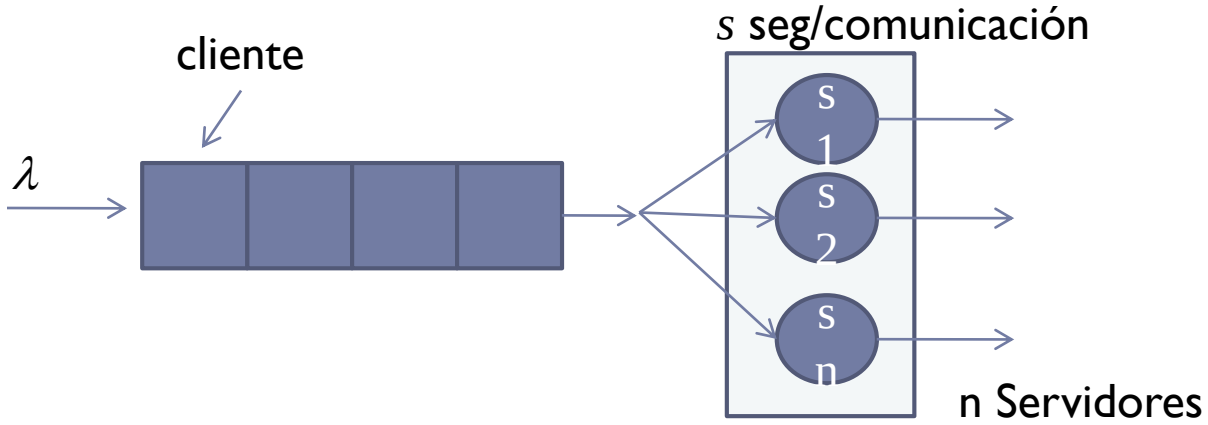
Disciplinas de colas

- PQ: Priority Queueing
  - Los clientes son clasificados por prioridades y son atendidos según el que tenga la prioridad más alta.



Sistemas de espera de Erlang

- Ahora se considerará el tráfico a un sistema con n servidores, con accesibilidad completa y un número infinito de posiciones de espera.
- Cuando los n servidores están ocupados, un cliente que llega se ubica en una cola y espera hasta que el servidor esté disponible (ocioso).
- Cuando un servidor está ocioso, es porque no hay clientes en cola (accesibilidad completa)

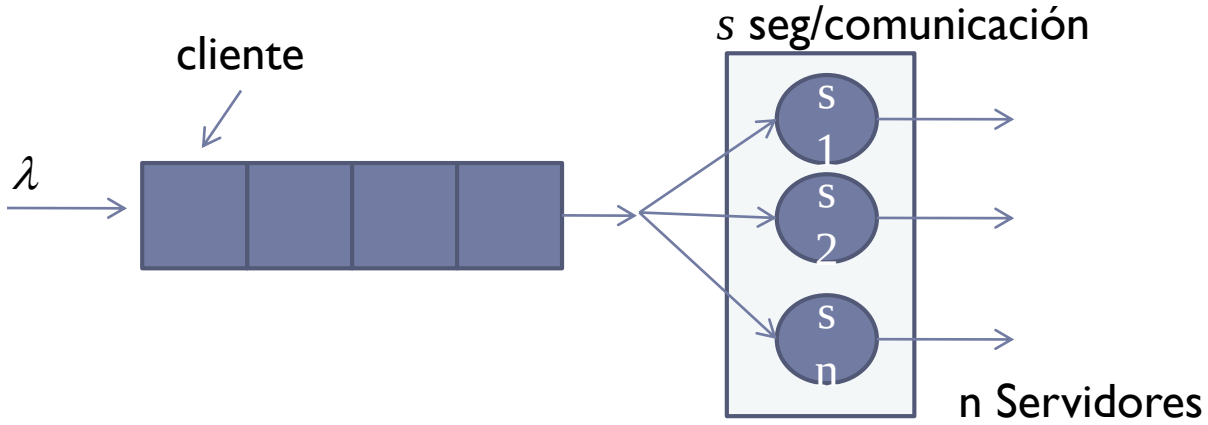


Tipos de sistemas considerados

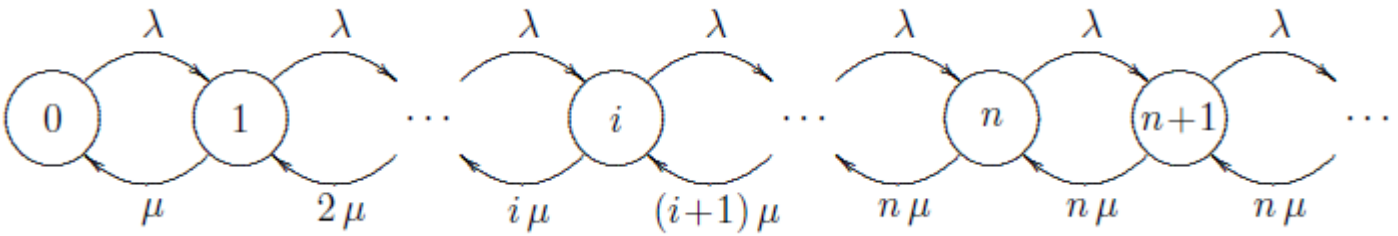
- Sistema de espera de Erlang:
  - Se tiene un proceso de llegadas de Poisson (infinito número de fuentes y tiempos de servicio exponencialmente distribuidos)
  - Se conoce también como un sistema M/M/n según la notación de Kendall
- Modelo de Colas cerradas (Reparación de máquina de Palm):
  - Se tiene un número limitado de fuentes y tiempos de servicio exponencialmente distribuidos
  - Este modelo ha sido aplicado ampliamente para dimensionamiento de computadores, terminales, sistemas de manufactura flexible.

Sistema de Espera de Erlang (M/M/n)

- Características:
  - Proceso de llegadas de Poisson (M)
  - Tiempos de Servicio exponenciales (M)
  - n servidores
  - Número infinito de posiciones de espera en cola
- El estado del sistema se define como el número total de clientes en el sistema (tanto siendo servidos como en cola de espera)



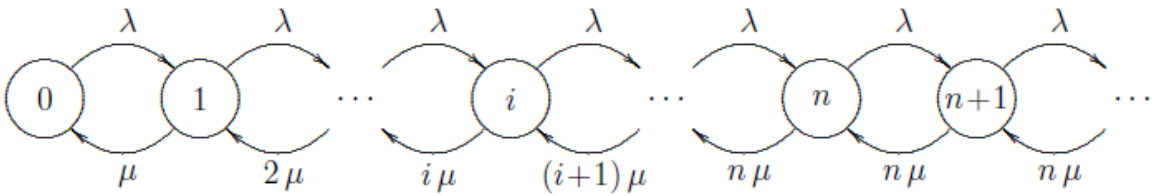
Sistema de Espera de Erlang



- Con base en el diagrama de transición de estados, se pueden obtener las probabilidades de estado estable.

Sistemas de Espera de Erlang

- Asumiendo equilibrio estadístico, las ecuaciones de corte serán:



Los n servidores están ocupados

$$
\begin{aligned}
\lambda \cdot p(0) &= \mu \cdot p(1)\,, \\
\lambda \cdot p(1) &= 2\,\mu \cdot p(2)\,, \\
&\vdots \quad \vdots \quad \vdots \\
\lambda \cdot p(i) &= (i+1)\,\mu \cdot p(i+1)\,, \\
&\vdots \quad \vdots \quad \vdots \\
\lambda \cdot p(n-1) &= n\,\mu \cdot p(n)\,, \\
\lambda \cdot p(n) &= n\,\mu \cdot p(n+1)\,, \\
&\vdots \quad \vdots \quad \vdots \\
\lambda \cdot p(n+j) &= n\,\mu \cdot p(n+j+1)\,.
\end{aligned}
$$

Sistemas de Espera de Erlang

- Como el tráfico ofrecido es $A=\lambda/\mu$, tendremos

$$p(i) = \begin{cases} p(0) \cdot \dfrac{A^i}{i!}\,, & 0 \le i \le n\,, \\ p(n) \cdot \left(\dfrac{A}{n}\right)^{i-n} = p(0) \cdot \dfrac{A^i}{n! \cdot n^{i-n}}\,, & i \ge n\,. \end{cases}$$

- Utilizando la normalización de las probabilidades, obtenemos p(0):

$$1 = \sum_{i=0}^{\infty} p(i)\,,$$

Progresión geométrica

$$1 = p(0) \cdot \left\{ 1 + \frac{A}{1} + \frac{A^2}{2!} + \cdots + \frac{A^n}{n!} \left(1 + \frac{A}{n} + \frac{A^2}{n^2} + \ldots \right) \right\}$$

Sistemas de Espera de Erlang

- La expresión de normalización sólo se cumple si:

$$A < n$$

- Por tanto, el equilibrio estadístico se obtiene para esta condición. De otra forma, la cola continuará creciendo hasta infinito.
- Así, obtenemos:

$$p(0) = \frac{1}{\sum_{i=0}^{n-1} \frac{A^i}{i!} + \frac{A^n}{n!} \frac{n}{n-A}}, \qquad A < n\,.$$

Características de Tráfico de los sistemas de espera

- Ahora se estudiarán la capacidad y el rendimiento del sistema.
- Para ello se estudiarán las probabilidades de estado estable.

La fórmula de Erlang-C

- Propiedad PASTA: Poisson Arrivals See Time Average
  - Cuando el proceso de llegadas es de Poisson, la probabilidad de que un cliente que llega tenga que esperar en cola es igual a la proporción del tiempo que los servidores están ocupados.
- Tiempo de espera: W
- Esta es la fórmula C de Erlang (1917). Esta se denota como $E_n(A)=E_{2,n}(A)$. La segunda forma se refiere al nombre alternativo de segunda fórmula de Erlang.

$$
\begin{aligned}
E_{2,n}(A) &= p\{\mathcal{W} > 0\} \\
&= \frac{\sum_{i=n}^{\infty} \lambda\, p(i)}{\sum_{i=0}^{\infty} \lambda\, p(i)} = \sum_{i=n}^{\infty} p(i) \\
&= p(n) \cdot \frac{n}{n - A}\,.
\end{aligned}
$$

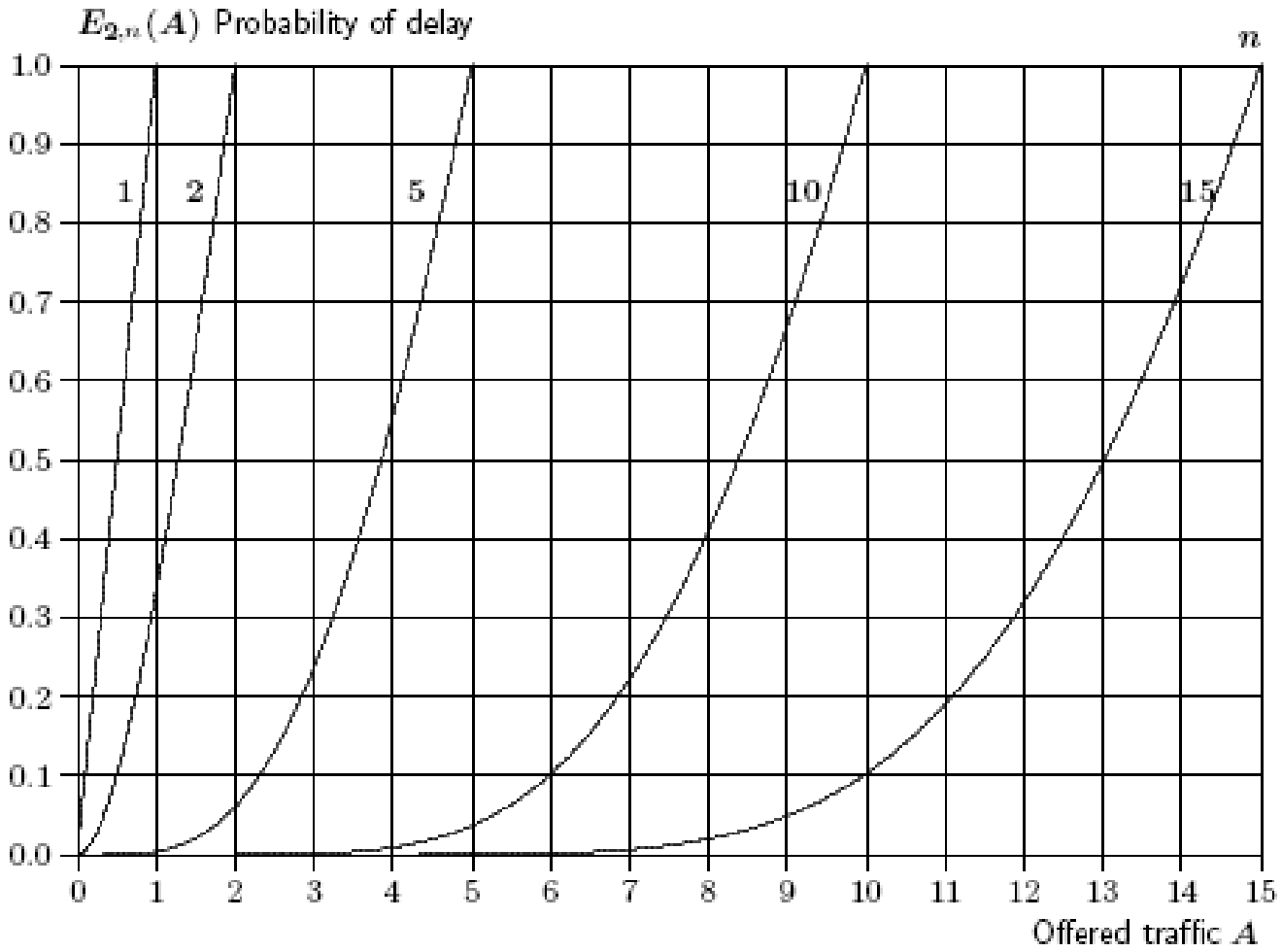
La fórmula de Erlang-C

- Reemplazando p(n), la fórmula de Erlang-C puede escribirse entonces como:

$$E_{2,n}(A) = \frac{\dfrac{A^n}{n!}\dfrac{n}{n-A}}{1 + \dfrac{A}{1} + \dfrac{A^2}{2!} + \cdots + \dfrac{A^{n-1}}{(n-1)!} + \dfrac{A^n}{n!}\dfrac{n}{n-A}}, \quad A < n\,.$$

- La probabilidad de retardo depende sólo de A=λ/μ y del número de servidores (n).
- A la fórmula de Erlang C se le conoce también como la segunda fórmula de Erlang o Fórmula de Erlang para sistemas de tiempo de espera.
- Otras notaciones son: $E_{2,n}(A) = D = D_n(A) = p\{W > 0\}$

Curva de Erlang-C para diferente número de servidores



Otros parámetros importantes

- La probabilidad de que un cliente sea atendido inmediatamente, sin tener que esperar es:

$$S_n = 1 - E_{2,n}(A)$$

- El tráfico servido Y es igual al tráfico ofrecido A, ya que los clientes no son rechazados y el proceso de llegadas es de tipo Poisson:

$$\begin{aligned} Y &= \sum_{i=1}^{n} i\, p(i) + \sum_{i=n+1}^{\infty} n\, p(i) \\ &= \sum_{i=1}^{n} \frac{\lambda}{\mu} p(i-1) + \sum_{i=n+1}^{\infty} \frac{\lambda}{\mu} p(i-1) \\ &= \frac{\lambda}{\mu} = A\,, \end{aligned}$$

Otros parámetros importantes

- La probabilidad de que un cliente tenga que hacer cola es la probabilidad de que la longitud de la cola (L) sea mayor que cero:

$$p\{\mathcal{L} > 0\} = \sum_{i=n+1}^{\infty} p(i) = \frac{\frac{A}{n}}{1 - \frac{A}{n}} \cdot p(n)\,,$$

$$p\{\mathcal{L} > 0\} = \frac{A}{n - A}\, p(n) = \frac{A}{n}\, E_{2,n}(A)\,.$$

Relación entre las fórmulas B y C de Erlang

- Las fórmulas B y C de Erlang están relacionadas así:

$$E_{2,n}(A) = \frac{n \cdot E_{1,n}(A)}{n - A\left(1 - E_{1,n}(A)\right)}$$

$$= \frac{E_{1,n}(A)}{1 - A\left\{1 - E_{1,n}(A)\right\}/n}, \qquad A < n\,.$$

- Nótese que:

$$E_{2,n}(A) > E_{1,n}(A)$$

Longitud media de la cola

- Se puede distinguir entre dos casos:
  - Longitud media de la cola en cualquier instante de tiempo
  - Longitud media de la cola cuando hay clientes en cola

Longitud media de la cola en cualquier instante de tiempo

- Es llamada también Longitud de cola virtual
- Es la cola experimentada por un cliente arbitrario que llega a un sistema con la propiedad PASTA
- Para cualquier instante de tiempo: $L_n = E\{\mathcal{L}\}$

$$
\begin{aligned}
L_n &= 0 \cdot \sum_{i=0}^{n} p(i) + \sum_{i=n+1}^{\infty} (i-n)\, p(i) \\
&= \sum_{i=n+1}^{\infty} (i-n)\, p(n) \left(\frac{A}{n}\right)^{i-n} \\
&= p(n) \cdot \sum_{i=1}^{\infty} i \left(\frac{A}{n}\right)^{i} \\
&= p(n) \cdot \frac{A}{n} \sum_{i=1}^{\infty} \frac{\partial}{\partial (A/n)} \left\{ \left(\frac{A}{n}\right)^{i} \right\}.
\end{aligned}
$$

- La serie es convergente ya que (A/n)<1 y el operador diferencial puede salir de la sumatoria. Luego,

$$
\begin{aligned}
L_n &= p(n)\,\frac{A}{n}\,\frac{\partial}{\partial(A/n)}\left\{\frac{A/n}{1-(A/n)}\right\} = p(n)\cdot\frac{A/n}{\{1-(A/n)\}^2} \\
&= p(n)\cdot\frac{n}{n-A}\cdot\frac{A}{n-A}\,, \\
L_n &= E_{2,n}(A)\cdot\frac{A}{n-A}\,.
\end{aligned}
$$

Longitud media de la cola cuando hay clientes en cola

- Es una probabilidad condicional: es el valor medio de la proporción de clientes en cola dado que ya hay n clientes siendo atendidos

$$L_{nq} = \frac{\sum_{i=n+1}^{\infty} (i-n)\, p(i)}{\sum_{i=n+1}^{\infty} p(i)}$$

- O también

$$L_{nq} = \frac{L_n}{p\{\mathcal{L} > 0\}},$$

$$= \frac{p(n) \cdot \dfrac{A/n}{(1-A/n)^2}}{p(n)\,\dfrac{A}{n-A}}$$

$$= \frac{n}{n-A}$$

Tiempos de espera medios

- Ahora se estudiarán dos tiempos de gran interés:
  - Tiempo medio de espera para todos los clientes (W):
    - Es un Indicador del nivel de servicio del sistema completo
  - Tiempo medio de espera para los clientes que experimentan espera (w):
    - Es un indicador para los clientes que sufren retardo por tener que esperar.
- Los tiempos medios son iguales a las llamadas medias debido a la propiedad PASTA

Tiempo medio de espera para todos los clientes

- Teorema de Little

$$L_n = \lambda W_n \,.$$

- Donde $L_n = L_n(A)$ y $W_n = W_n(A)$.
- Despejando W_n, y reemplazando L_n:

$$W_n = \frac{L_n}{\lambda} = \frac{1}{\lambda} \cdot E_{2,n}(A) \cdot \frac{A}{n - A} \,.$$

- Como $A = \lambda s$, donde s es el tiempo medio de servicio, obtenemos

$$W_n = E_{2,n}(A) \cdot \frac{s}{n - A} \,.$$

Tiempo medio de espera para clientes que esperan

- Es el tiempo de espera para los clientes que son puestos en cola.
- Puede calcularse como el tiempo de servicio (s) multiplicado por la tasa media de clientes que son dejados en espera (1/(n-A)):

$$w_n = \frac{s}{n - A}.$$

- Por tanto, la relación entre los tiempos de espera vistos es:

$$W_n = w_n \cdot E_{2,n}(A)$$

Ejemplo de Dimensionamiento: Sistemas Trunking

- Son sistemas de radiocomunicaciones de voz
- Se asigna un canal sólo cuando hay demanda.
- Cada usuario utiliza el canal únicamente durante el tiempo de conversación.
- Cuando concluye la conversación se libera el canal para que pueda ser asignado a otro usuario.
- Un usuario que llama puede ser puesto a esperar si no hay recursos disponibles
- Por tanto, es un sistema de espera.

Ejemplo de Dimensionamiento: Sistemas Trunking

- *El problema directo:*

Consiste en calcular el número de radiocanales necesarios (N) para dar servicio a M móviles con un GoS determinado.

- *Datos*:

GoS, Wo (t espera), H (duración media de las llamadas en la BH), M (Número de terminales), L (Número de llamadas/terminal en la BH)

- *Solución:*

1. Se calcula el tráfico ofrecido: A= MLH/3600
2. Se obtiene N a prueba y error de la expresión:

$$GoS = 100C(N, A)e^{\frac{-(N-A)W_o}{H}}$$

Ejemplo:

- Se desea determinar el número de radiocanales para un sistema Trunking con los siguientes datos:
  - Wo=H=20 segundos
  - GoS=5%
  - M=1000 terminales
  - L=1 (Número llamadas por terminal en la BH)

Ejemplo de Dimensionamiento: Sistemas Trunking

- **El problema Inverso:**

  Cuántos terminales adicionales podría admitir nuestra red con los N radiocanales?

- *Datos:*

N (Número de canales), GoS, L (número de llamadas/terminal en la BH), H (duración media de las llamadas en la BH)

- *Solución:*

1. A partir de la expresión del GoS se obtiene a prueba y error el valor de A.

$$GoS = 100C(N, A)e^{\frac{-(N-A)W_o}{H}}$$

2. Se obtiene a=LH/3600
3. Se obtiene M= Int (A/a)

Ejemplo:

- Cuando una empresa crece, se habilitan nuevos terminales.
- Grado de saturación:
  - Si se tolera un incremento del GoS hasta un valor del GoS de saturación, podemos determinar hasta cuántos terminales podrá crecer la red.
- Supóngase que en el ejemplo anterior se admite un GoS de saturación del 30%
- Datos:
  - Wo=H=20 segundos
  - N=8 radiocanales
  - L=1 (Número llamadas por terminal en la BH)
- Calcular el número de terminales que se soportan con este GoS.

Caso especial: M/M/1

- Cuando sólo se tiene un servidor, se obtiene el modelo M/M/1
- Este caso es muy utilizado en diferentes aplicaciones.
- Para este caso se obtienen los siguientes resultados:
  - Probabilidad del estado i:

$$p(i) = (1 - A) \cdot A^i, \qquad i = 0, 1, 2, \ldots,$$

  - Probabilidad de que haya retardo:

$$E_{2,1}(A) = A\,.$$

  - Longitud media de la cola:

$$L_1 = \frac{A^2}{1-A},$$

  - Tiempo medio de espera:

$$W_1 = \frac{A\,s}{1-A}.$$

Comportamiento de la cola M/M/1:

- De las expresiones para la cola M/M/1, podemos concluir que:
  - Cuando se **incrementa el tráfico ofrecido** por los usuarios, se **incrementa la longitud de la cola**, independientemente de si el incremento del tráfico es debido a que se incrementa el número de usuarios (λ) o que se incrementa el tiempo de servicio (s).
  - El **tiempo medio de espera se incrementa** en una segunda potencia de s, pero solamente por la primera potencia de λ. Por tanto, un incremento de carga (A) debido a más clientes afecta menos que un incremento de carga (A) debido al aumento del tiempo de servicio.
  - Es importante que los tiempos de servicio de un sistema no se incrementen durante la sobrecarga.

Tiempo en el sistema para una cola M/M/1

- Tiempo en el sistema= Tiempo de espera + Tiempo de servicio
- También es llamado tiempo de respuesta
- Se distribuye exponencialmente con intensidad (μ-λ):

$$F(t) = 1 - \mathrm{e}^{-(\mu-\lambda)t}, \quad \mu > \lambda, \quad t \geq 0.$$

- El tiempo medio en el sistema se calcula entonces como:

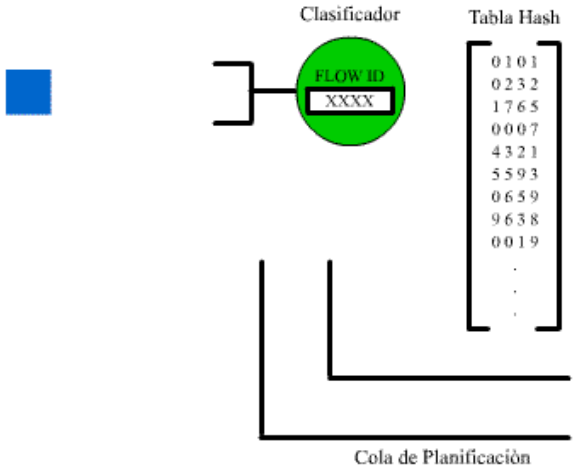
$$m = W_1 + s = \frac{A\,s}{1-A} + s,$$

$$m = \frac{s}{1-A} = \frac{1}{\mu-\lambda},$$

- Donde μ=1/s es la

Ejemplo: Clasificador de paquetes

- Un clasificador examina la cabecera de los paquetes para determinar a qué sesión pertenecen.
- Según la sesión, se le dará un tratamiento diferente en el enlace de salida.
- Para el modelo se suponen llegadas de paquetes con un proceso de Poisson (Tráfico elástico).



Ejemplo: Retardo de los paquetes en un clasificador de paquetes

Modelo Simulado

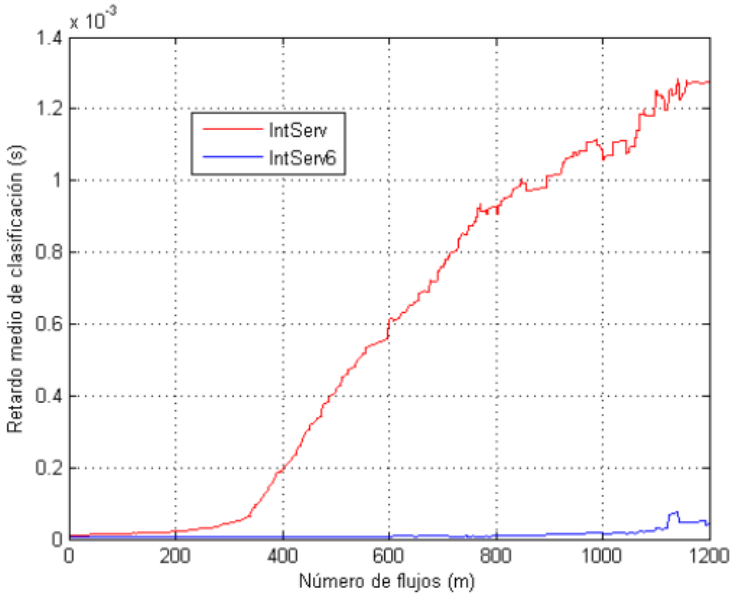
Comportamiento similar

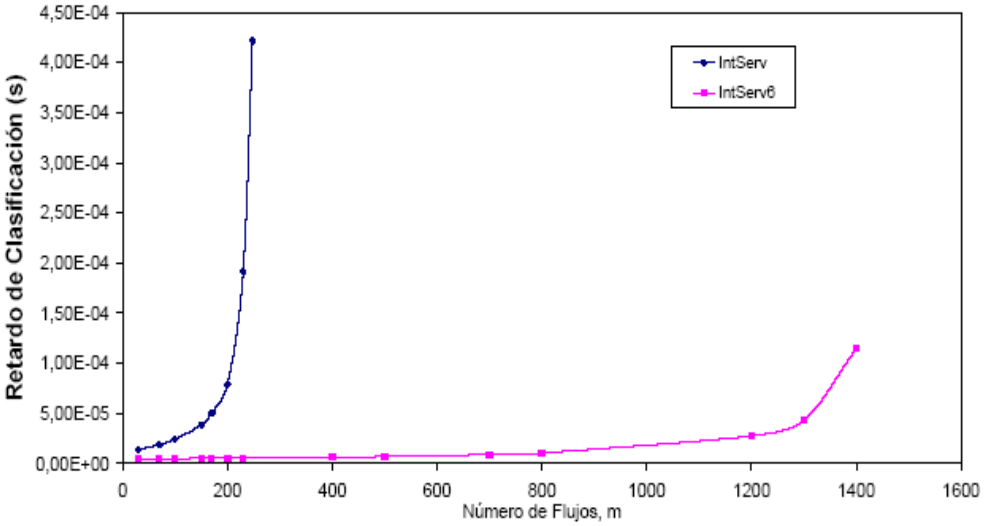
Modelo Teórico: M/M/1

$$d_c = \frac{1}{\frac{1}{\bar{x}_c} - \sum_{j=1}^{m} \lambda_j}$$

x_c : tiempo medio de servicio

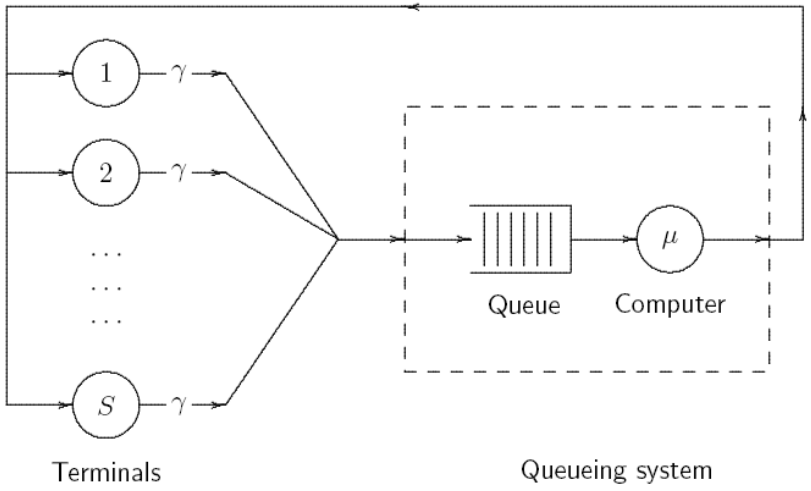
λ_j: tasa media de llegadas por sesión





Red de Colas Cerrada (Modelo de reparación de máquinas)

- Este modelo es un sistema de colas cerrada o cíclica
- El número de usuarios es limitado
- Se conoce como Caso Engset para sistemas de pérdidas
- Estudiado por Gnedenko (1933), C. Palm (1947) y Feller (1950)
- El caso para el que se estudió fue para cuando se tiene un sistema en que se reparan máquinas por uno o varios reparadores (servidores) y luego ellas operan nuevamente.
- El problema se enfoca en ajustar el número de servidores según el número de máquinas para que el costo total sea minimizado.
- Cuando una máquina pasa a ser reparada, esta debe reemplazarse por otra mientras tanto.



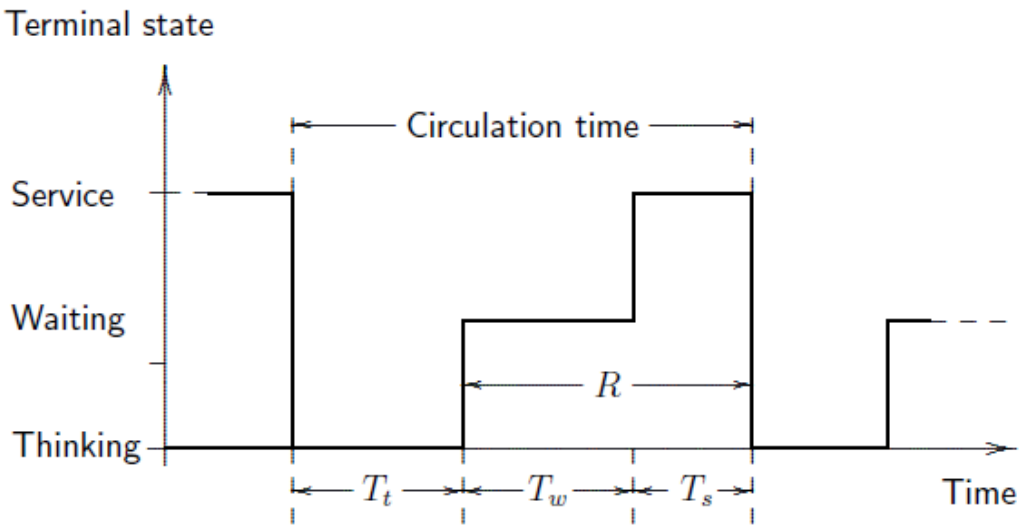
Sistema de colas cerrado

- Este modelo ha sido aplicado ampliamente para solucionar problemas de ingeniería de tráfico en sistemas computarizados.
- Notación Kendall: M/M/n/S/S
  - S: Número de clientes
  - n: Número de servidores
- Ejemplos:
  - En la Web, las máquinas corresponden a los clientes, mientras que los clientes corresponden a los servidores
  - En un sistema de terminales de computador las máquinas corresponden a los terminales y los servidores son los computadores que gestionan los terminales
  - En un sistema computador las máquinas podrían corresponder a un sistema de almacenamiento en disco y los servidores corresponden a los canales de Entrada/Salida.

Modelo de terminales

- Se aplica para sistemas que hacen multiplexación por división de tiempo (TDM)
- En un sistema TDM los usuarios sienten que son los únicos que utilizan al servidor.
- Un terminal individual cambia entre dos estados:
  - El usuario está pensando (trabajando)
  - El usuario está esperando por una respuesta del servidor

- T_t: Tiempo en que el terminal está pensando. (m_t: media de la v.a. T_t)
- R: Tiempo de respuesta. El terminal está esperando.
- $R = T_w + T_s$
- T_w: Tiempo de espera en cola (media: m_w)
- T_s: Tiempo de servicio (media: m_s)
- $T_t + R$ es el tiempo de circulación



Modelo de terminales

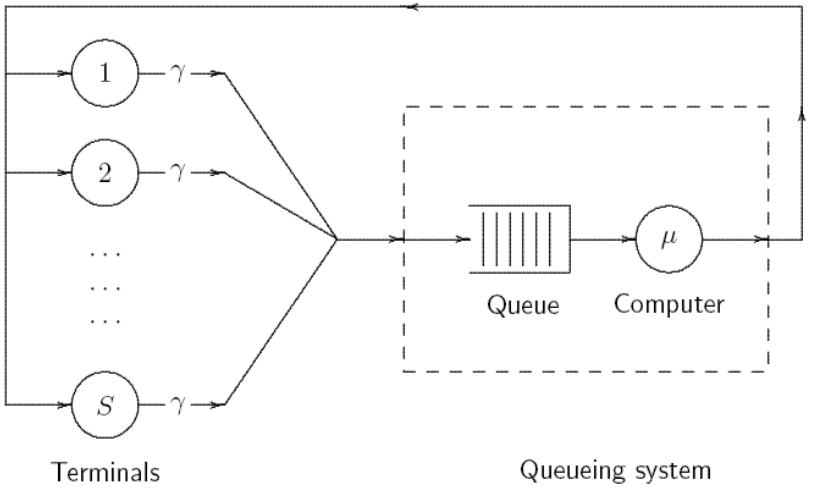
- Se consideran S terminales conectados a un computador
- Se asume que los tiempos de "pensar" de cada terminal son largos.
- Se asumen entonces llegadas de Poisson exponencialmente distribuidas con intensidad:

$$\gamma = 1/m_t$$

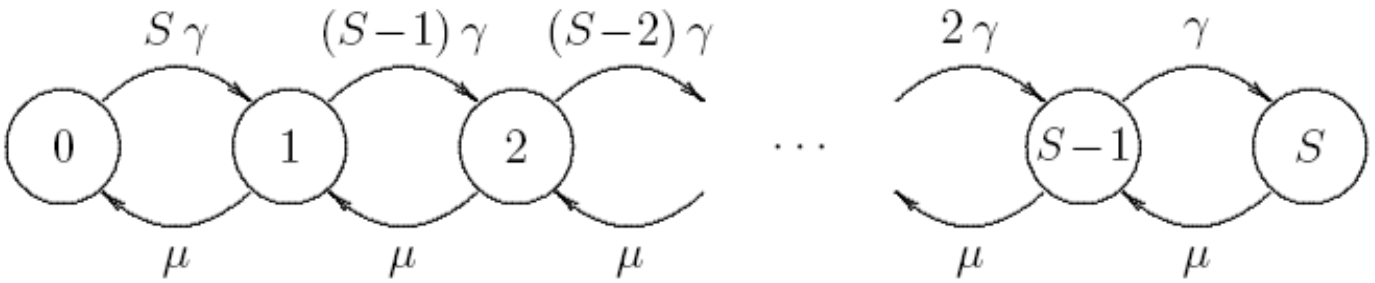
- El tiempo de servicio (ejecución) en el computador se asume también exponencialmente distribuido con tasa:

$$\mu = 1/m_s$$

- Cuando un terminal no puede ser atendido pero tiene una solicitud, debe esperar en cola.



Modelo de terminales



- Para el diagrama de estados se asume el estado i que representa el número de terminales que comprende los terminales en cola y los terminales siendo atendidos.
- Computador ocioso: i=0
- Computador trabajando: i>0, (i-1) terminales esperando.
- Se asume equilibrio estadístico.
- La intensidad de llegadas a la cola se decrementa en la medida que la longitud de la cola se incrementa (si todos los terminales están en cola, la intensidad de llegadas es cero).

Obtención de las probabilidades de estado

- Siguiendo el procedimiento descrito anteriormente, se halla p(0):
  - Haciendo

$$\varrho = \mu/\gamma \quad \text{:Tráfico ofrecido o tasa de servicio}$$

$$p(S-i) = \frac{\varrho^i}{i!}\, p(S)$$

$$= \frac{\dfrac{\varrho^i}{i!}}{\displaystyle\sum_{j=0}^{S} \frac{\varrho^j}{j!}}, \quad i = 0, 1, \ldots, S\,,$$

$$p(0) = E_{1,S}\,(\varrho)\ .$$

  - Que es la distribución truncada de Poisson
  - p(0) es la probabilidad de i=S (todos los clientes están en cola o siendo atendidos)

Ejemplo:

- Considere un sistema computador con 6 discos conectados a un mismo canal de E/S.
- El tiempo de búsqueda (posicionamiento del brazo) medio es 3ms y el tiempo medio de ubicación de un archivo es 1ms, correspondiendo a un tiempo de rotación de 2ms.
- El tiempo de lectura de un archivo es exponencialmente distribuido con media 0,8ms.
- El canal está ocupado sólo durante la lectura.
- Se quiere encontrar la máxima capacidad del sistema (en solicitudes por segundo)

- Solución:
  - Tiempo de "pensar": 4ms
  - Tiempo de servicio: 0,8ms
  - Tasa de servicio=1/0,8=5
- El número medio de terminales que son servidos por el sistema (pueden ser atendidos) es:

$$1 - p(0) = 1 - E_{1,6}(5) = 0.8082$$

- Por tanto, la tasa máxima de

$$\gamma_{max} = 0.8082/0.0008 = 1010$$

dada en solicitudes por segundo

Otros parámetros del modelo

- Número de terminales siendo atendidos:

$$n_s = 1 - E_{1,S}(\varrho)$$

- Número medio de terminales "pensando":

$$n_t = \frac{\mu}{\gamma}\{1 - E_{1,S}(\varrho)\} = \varrho\{1 - E_{1,S}(\varrho)\}$$

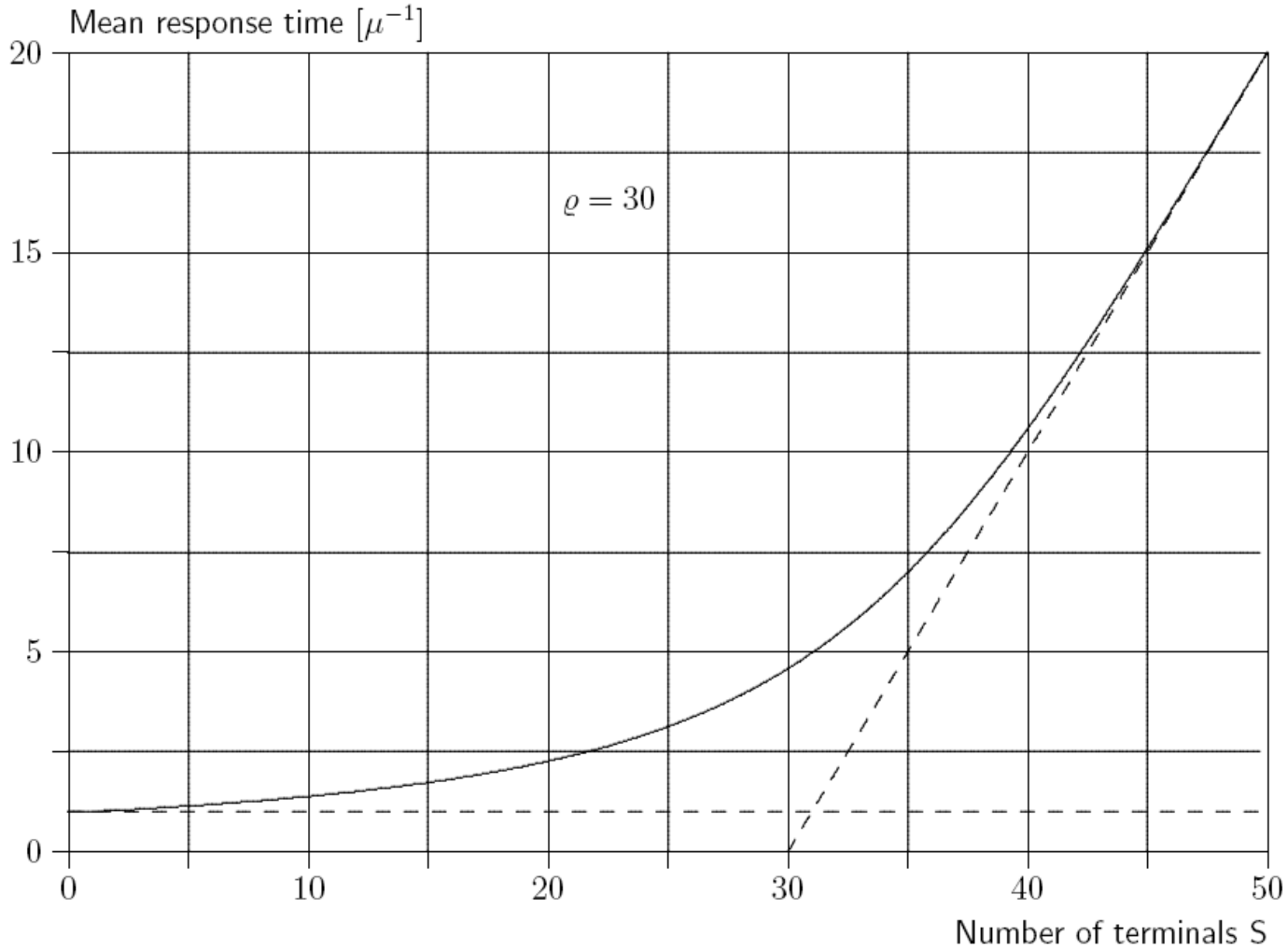
- Número medio de terminales esperando:

$$\begin{aligned} n_w &= S - n_s - n_t = S - \{1 - E_{1,S}(\varrho)\} - \varrho \cdot \{1 - E_{1,S}(\varrho)\} \\ &= S - \{1 - E_{1,S}(\varrho)\}\{1 + \varrho\}\,. \end{aligned}$$

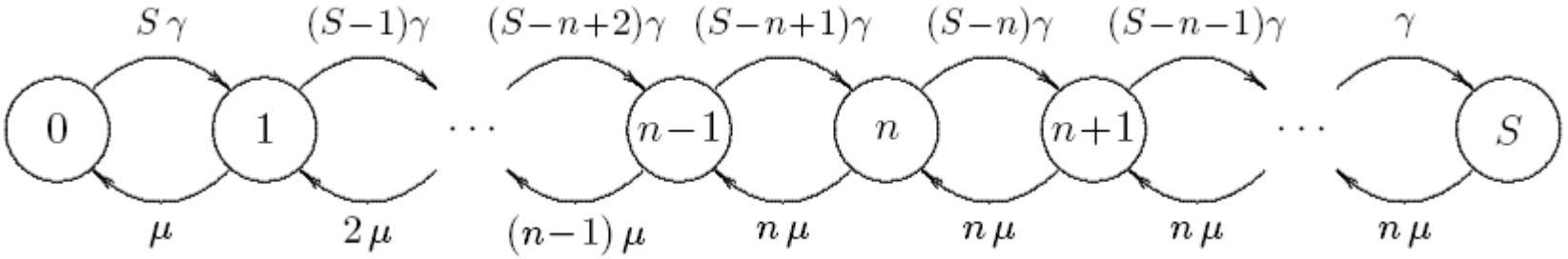
- Tiempo de respuesta medio:

$$m_r = \frac{S}{1 - E_{1,S}(\varrho)} \cdot m_s - m_t$$

Variación del tiempo de respuesta medio vs. Número de terminales



Modelo de cola cerrada con N servidores



$$p(i) = \binom{S}{i} \left(\frac{\gamma}{\mu}\right)^i p(0)\,,$$

$$p(i) = \frac{(S-n)!}{(S-i)!} \left(\frac{\gamma}{n\mu}\right)^{i-n} \cdot p(n)\,,$$

$p_s = p\,\{\textit{the terminal is served by a computer}\},$

$p_w = p\,\{\textit{the terminal is waiting for service}\},$

$p_t = p\,\{\textit{the terminal is thinking}\}.$

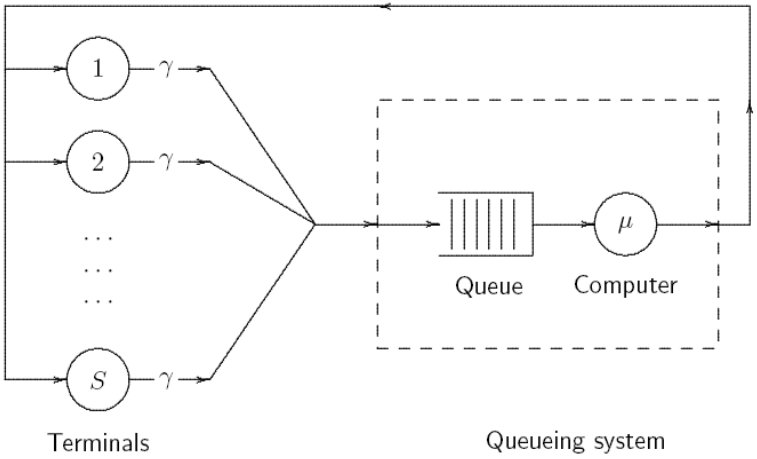
$$p_s = \frac{1}{S}\left\{\sum_{i=0}^{n} i \cdot p(i) + \sum_{i=n+1}^{S} n \cdot p(i)\right\}$$

$$p_t = p_s \cdot \frac{\mu}{\gamma}\,,$$

$$p_w = 1 - p_s - p_t\,.$$

Caso: Modelamiento de un canal de Internet en una red académica con una CQN

- El tráfico es de tipo http (tráfico elástico y por tanto, llegadas de Poisson)
- Se considera que el tráfico mínimo por usuario es de 50Kbps (b).
- Hay 235 usuarios (N)
- El tiempo medio en el estado "pensar" es de 0,135 seg=1/μ_1.
- Tamaño medio de los archivos: 1467.21bits (f)



- El ancho de banda por usuario es h
- El Número de usuarios es N
- B es el ancho de banda en el enlace
- Se cumple la relación:

$$B = N * h$$

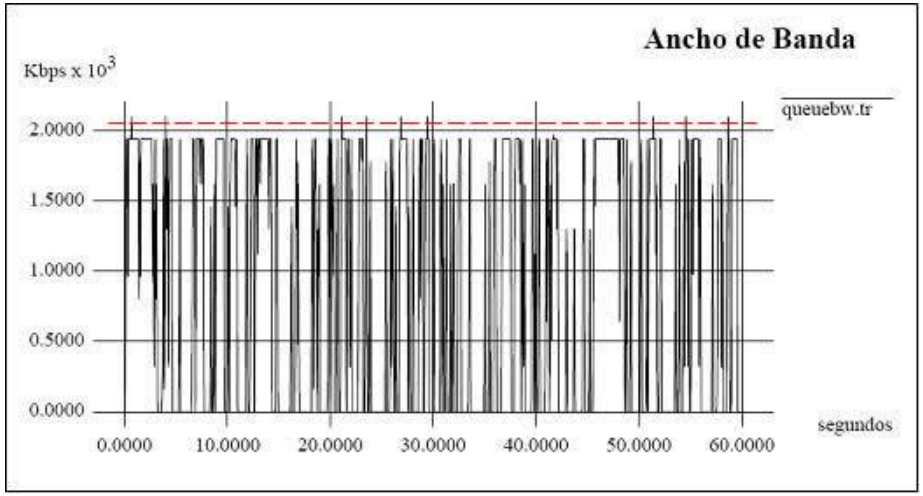
- Donde h:

$$h = \left(\frac{1}{b} + \frac{1}{\mu_1 f} \right)^{-1}$$

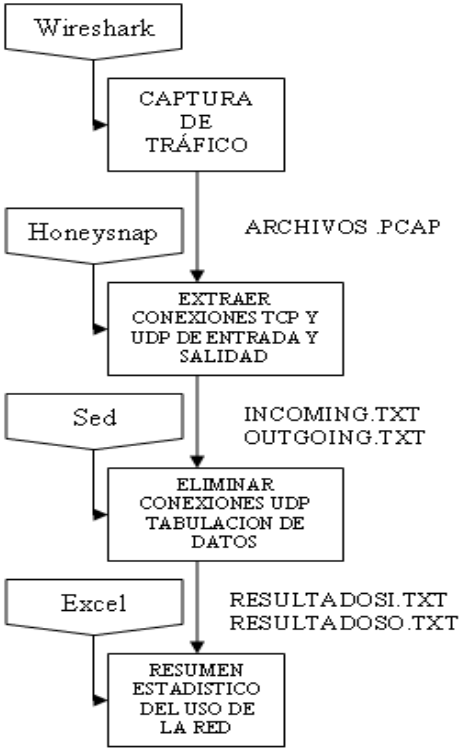
- *h=8900,78 bps*
- *B=2091684,95 bps*

Validación

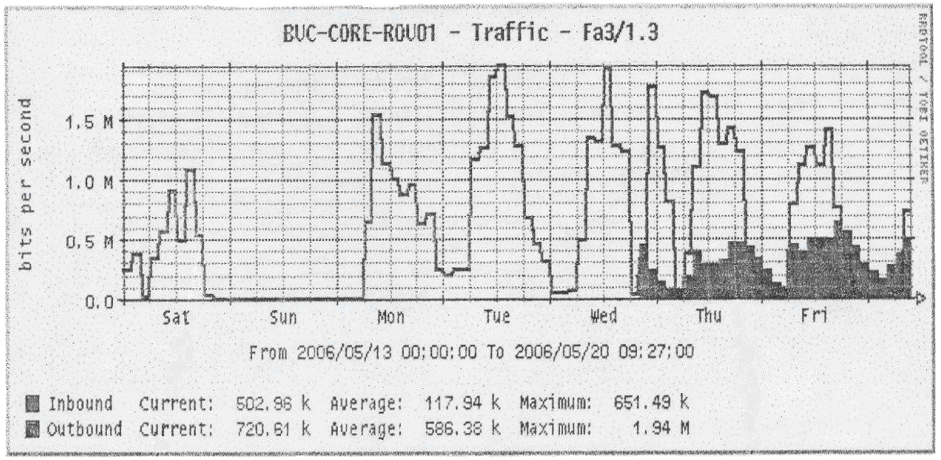
- Tráfico simulado con las características obtenidas



- Proceso de medición y análisis del tráfico:



- Tráfico medido:



Colas M/G/1

- Llegadas de Poisson
- Un servidor
- Distribución general para el tiempo de servicio, por lo que puede aplicarse a gran cantidad de situaciones.

La fórmula de Pollaczek-Khintchine para colas M/G/1

- W: Tiempo medio para todos los clientes
- s: Tiempo medio de servicio
- A: Tráfico ofrecido
- ε: factor de forma

$$W = \frac{A \cdot s}{2(1-A)} \cdot \varepsilon \,,$$

$$W = \frac{V}{1-A} \,,$$

$$V = A \cdot \frac{s}{2} \cdot \varepsilon = \frac{\lambda}{2} \cdot m_2 \,.$$

La fórmula de Pollaczek-Khintchine para colas M/G/1

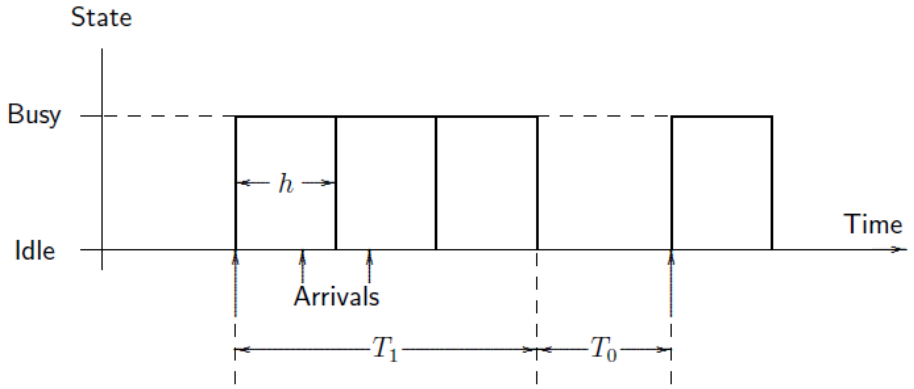
- Entre más regular (menor varianza y por tanto, el factor de forma) es el proceso de servicio, más pequeño será el tiempo medio de espera
- En sistemas telefónicos reales el factor de forma estará entre 4 a 6, en tráfico de datos estará entre 10 y 100.
- Un caso especial de la fórmula ya lo estudiamos para el caso M/M/1.

Período de ocupación para M/G/1

- El período de ocupación de un sistema es el intervalo de tiempo desde el instante en que todos los servidores están ocupados, hasta que un servidor está ocioso nuevamente.
- Consideraciones:
  - Cuando el sistema queda vacío, ha perdido su memoria debido al proceso de llegadas de Poisson. Estos instantes son puntos de regeneración (equilibrio)
  - El siguiente evento ocurrirá de acuerdo a un proceso de Poisson con intensidad λ

Período de ocupación para M/G/1

- Sólo se requiere considerar un ciclo desde el instante en que el servidor cambia de estado de Ocioso a Ocupado hasta el momento en que cambia nuevamente el estado de ocioso a ocupado.
- Este ciclo incluye un tiempo de ocupación (T_1) y un tiempo de ociosidad (T_0)



Período de ocupación para M/G/1

- La proporción de tiempo que el sistema está ocupado con respecto al ciclo completo (ocupado+ocioso) es:

$$\frac{m_{T_1}}{m_{T_0+T_1}} = \frac{m_{T_1}}{m_{T_0} + m_{T_1}} = A = \lambda \cdot s\,.$$

- Y m_{To}es el tiempo medio en estado ocioso, que es el inverso de la tasa de llegadas de Poisson:

$$m_{T_0} = 1/\lambda$$

- Por lo que se obtiene la media del período de ocupación como:

$$m_{T_1} = \frac{s}{1-A}$$

Momentos de la distribución del tiempo de espera en M/G/1

- Se supone una disciplina de servicio FCFS (por eso el subíndice F)
- Denótese el i-ésimo momento de la distribución del tiempo de servicio por m_i.
- Se puede calcular el k-ésimo momento con la fórmula recursiva siguiente, donde se escoge $m_1=s=1$

$$m_{k,F} = \frac{A}{1-A} \sum_{j=1}^{k} \binom{k}{j} \cdot \frac{m_{j+1}}{j+1} \cdot m_{k-j,F}\,, \quad m_{0,F} = 1\,.$$

Longitud de Cola Limitada M/G/1/k

- En sistemas reales, la longitud de la cola, por ejemplo el tamaño de un buffer, es finita.
- Los clientes que llegan son bloqueados cuando el buffer está lleno.
- En Internet, se aplica esta estrategia en Routers y es llamada Estrategia Drop Tail.

Longitud de Cola Limitada M/G/1/k

- Existe una relación simple entre las probabilidades de estado *p(i)* (*i=0,1,2,…*) de un sistema infinito M/G/1, y las probabilidades de estado $p_k(i)$ (*i=0,1,2,….*) de un sistema M/G/1/k, donde el número total de posiciones para clientes es k, incluyendo el cliente que está siendo servido (Keilson, 1966)

$$p_k(i) = \frac{p(i)}{(1 - A \cdot Q_k)}, \qquad i = 0, 1, \ldots, k-1,$$

$$p_k(k) = \frac{(1 - A) \cdot Q_k}{(1 - A \cdot Q_k)},$$

- Donde A<1 es el Tráfico Ofrecido, y $Q_k = \sum_{j=k}^{\infty} p(j)$

Longitud de Cola Limitada M/G/1/k

- Existen algoritmos para calcular *p(i)* para distribuciones de tiempo arbitrarias (M/G/1), basados en análisis de Cadenas de Markov embebidos (Kendall, 1953).
- Nótese que p(i) sólo existe para A<1, pero para buffers finitos también obtenemos el equilibrio estadístico para A>1. En este ultimo caso no se puede usar la aproximación descrita aquí.

Sistemas de Colas con tiempos de Servicio Constantes

- Ahora se considerarán sistemas con tiempos de servicio constantes y con disciplina de colas FCFS
- El primer artículo sobre el tema fue publicado por Erlang en 1909. Estudió sistemas con llegadas de Poisson y servicios constantes.
- Evolución histórica:
  - Erlang-1909: 1 servidor (errores para más de un servidor)
  - Erlang-1917: 1, 2 y 3 servidores (sin prueba)
  - Erlang-1920: número arbitrario de servidores (soluciones explícitas para n=1,2,3)

Evolución histórica

- Erlang derivó la distribución del tiempo de espera, pero no consideró las probabilidades de estado.
- 1928- Fry desarrolló ecuaciones para las probabilidades de estado con base en el trabajo de Erlang.
- 1932-Crommelin (Ingeniero de teléfonos Británico) presentó una solución general para M/D/n. Generalizó las ecuaciones de estado para un n arbitrario y derivó la distribución del tiempo de espera. LIlamada ahora la distrib ucion de Crommelin.
- Pollaczek (1930-34) presentó una solución muy general independiente del tiempo para distribuciones de tiempos de servicio arbitrarias. Pudo hallar con esto las soluciones específicas para los casos exponencial y constante de los tiempos de servicio.
- Khintchine (1932) también obtuvo la distribución para el tiempo de espera.

Utilidad de este modelo

- Se puede usar en situaciones donde haya tiempos de servicio constante y colas FCFS.
- Ejemplos:
  - Un Router ATM con servicio Best-Effort: El tiempo de procesamiento de cada celda es constante
  - Un proceso de un Router que tenga un número de operaciones constante independientemente de la longitud de los paquetes. (P.ej, marcar un paquete con un valor en el campo DSCP de IP)
  - Un algoritmo de Leaky bucket para marcado de paquetes no cumplientes con tamaño de paquetes constante (caso ATM)

Probabilidades de estado para M/D/1

- En general, la probabilidad de que haya i clientes en el sistema es:

$$p(i) = (1 - A) \cdot \sum_{j=1}^{i} (-1)^{i-j} \cdot e^{jA} \cdot \left\{ \frac{(jA)^{i-j}}{(i-j)!} + \frac{(jA)^{i-j-1}}{(i-j-1)!} \right\}, \quad i = 2, 3, \ldots$$

- En este caso se conoce p(0). Si no se conociera, tocaría hallarla de la expresión de normalización de las probabilidades (suma total =1)

Tiempo de espera medio y período de ocupación

- Los tiempos medios se derivan de la fórmula de Pollaczek-Khintchine para M/G/1, con tiempo medio de servicio h (constante)
- Tiempo medio de espera para todos los clientes (W):

$$W = \frac{A \cdot h}{2(1-A)}$$

- Tíempo medio de espera para los clientes que esperan (w):

$$w = \frac{h}{2(1-A)}$$

- Tiempo medio de ocupación (contenido de M/G/1):

$$m_{T_1} = \frac{h}{1-A}$$

Distribución del número de clientes que llegan durante el período de ocupación

- Está dado por una distribución de Bórel:

$$B(i) = \frac{(i\,A)^{i-1}}{i!}\,\mathrm{e}^{-i\,A}, \qquad i = 1, 2, \ldots$$

Probabilidad acumulada para el tiempo de espera (W)

- Para tiempos de espera pequeños (Erlang-1909)

$$p\{W \leq t\} = (1-\lambda) \cdot \sum_{j=0}^{T} \frac{\{\lambda(j-t)\}^j}{j!} \cdot \mathrm{e}^{-\lambda(j-t)}$$

- Para tiempos de espera largos:

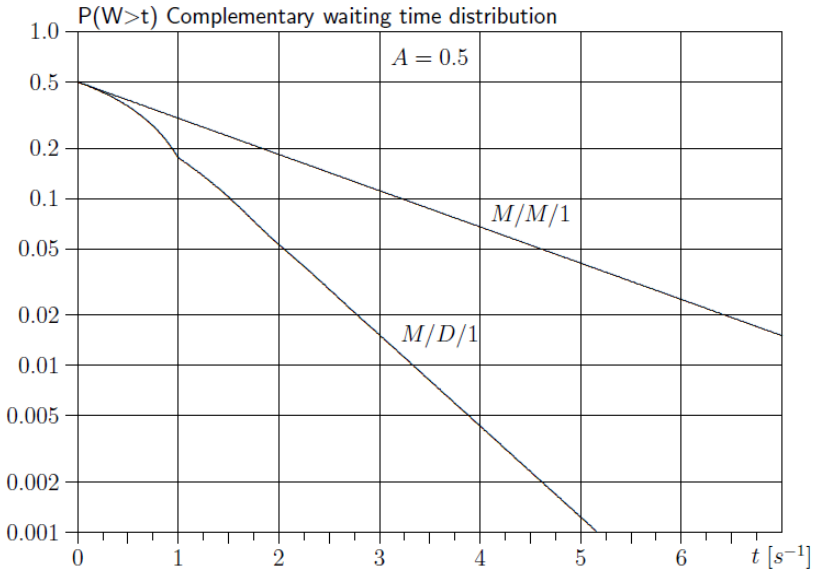
$$p\{W \leq T+\tau\} = \mathrm{e}^{\lambda\tau} \sum_{j=0}^{T} \frac{(-\lambda\tau)^j}{j!} \cdot p\{W \leq T-j\}$$

- Donde $p\{W \leq T-j\}$ está dada por (Iversen, 1982):

$$p\{W \leq t\} = p(0) + p(1) + \cdots + p(t)\,.$$

Comportamiento del tiempo medio de espera

- Distribución complementaria del tiempo de espera para colas M/M/1 y M/D/1 con disciplina de servicio FCFS
- Unidad de tiempo: Tiempo medio de servicio.



Sistemas de longitud finita de colas: M/D/1/k

- Las probabilidades de estado de un buffer finito pk(i) pueden ser obtenidas para cualquier tráfico ofrecido en la siguiente manera:
- En un sistema con un servidor y k-1 posiciones en cola, tenemos k estados incluyendo el cliente que está siendo servido.
- Las ecuaciones de balance de Fry para las probabilidades de estado pk(i), i_0,1,2,…,k-2 darán k-1 ecuaciones para los estados pk(0) a pk(k-1)
- También se usa la expresión de normalización (sumatoria total probabilidades es 1) $A = 1 - p_k(0) + A \cdot p_k(k)$
- Además, se usa la expresión

  (tráfico ofrecido=tráfico transportado+tráfico rechazado)

Ejemplo

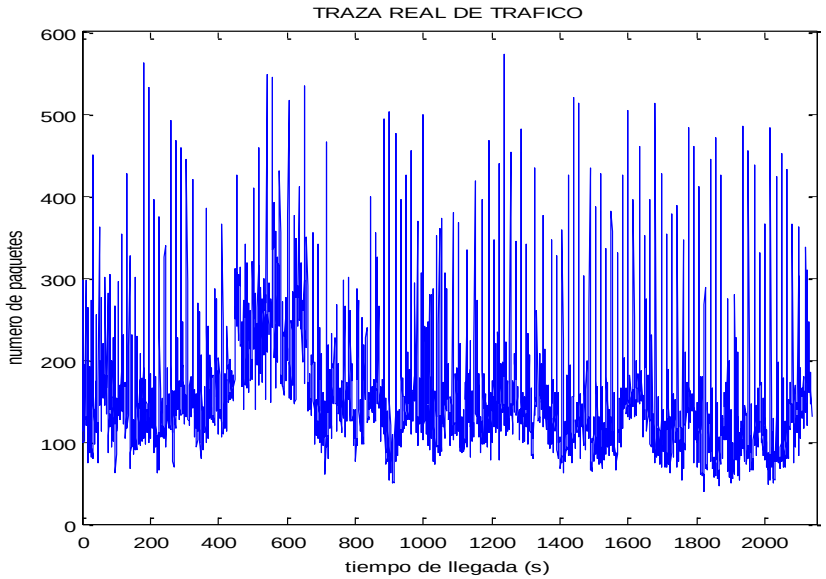
- Leaky Bucket en Sistemas ATM:
  - Es un sistema de colas con tiempo de servicio constante (tamaño de celda) y buffer finito.
  - Si el proceso de llegadas es un proceso de Poisson, entonces tendremos un sistema M/D/1/k
  - El tamaño del goteo corresponde a la intensidad de llegadas media aceptable a largo plazo, donde el tamaño del bucket describe el exceso (ráfaga) permitida.
  - El mecanismo copera como un sistema de colas virtual, donde las celdas son aceptadas inmediatamente o son rechazadas de acuerdo a los valores de un contador.

Caso: Modelamiento de un canal de Internet en una red académica con una cola MMPP/D/1

Metodología utilizada

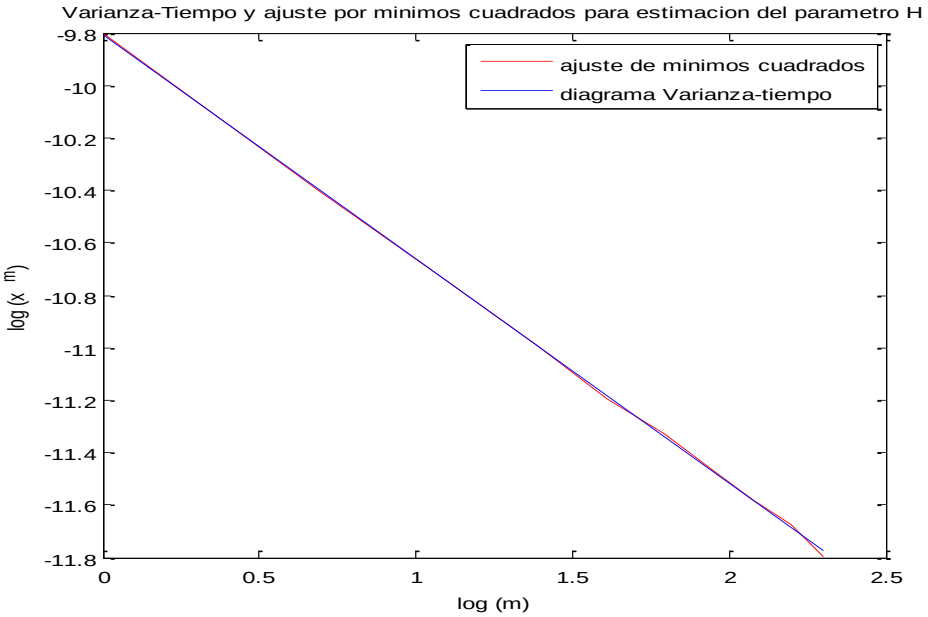
1. Muestreo y Procesamiento de las trazas de Tráfico
2. Caracterización del tráfico original y cálculo del parámetro Hurst
3. Generación de la traza de Tráfico mediante el modelo MMPP
4. Validación de la traza de Tráfico mediante Cuantiles
5. Evaluación del desempeño del modelo MMPP/D/1 y análisis de capacidad de canal

Muestreo y procesamiento



Cálculo del parámetro Hurst

Traza varianza-tiempo



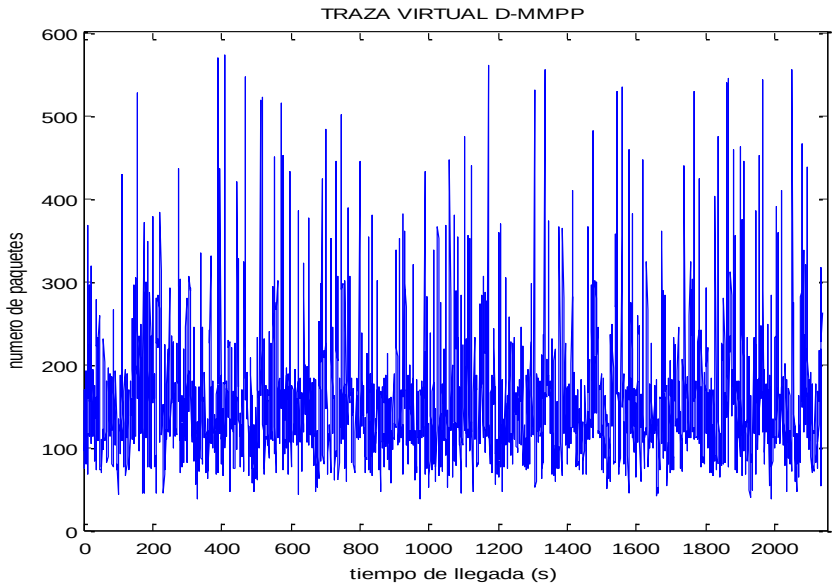
$$\log(X^m) = -\beta \log(m)$$

Donde, X^m **corresponde al proceso agregado.**
m son las divisiones del proceso X (niveles de agregación).
β es la pendiente de la recta y equivale a 2 (1-H)

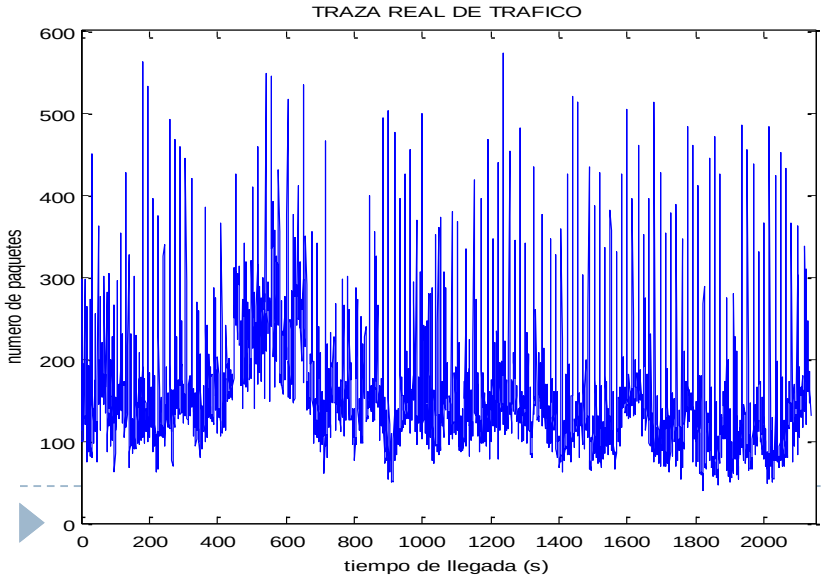
Generación traza tráfico MMPP

Generación de fuente en paquetes por segundo.

Tráfico Virtual



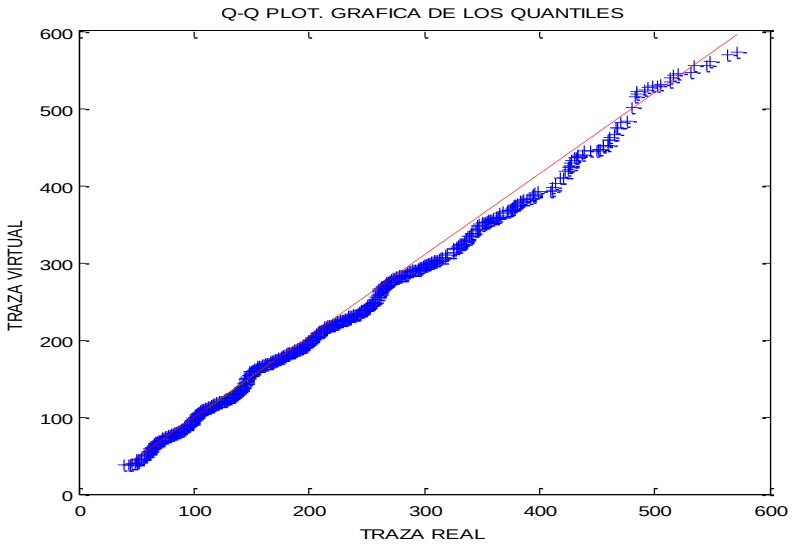
Tráfico Real



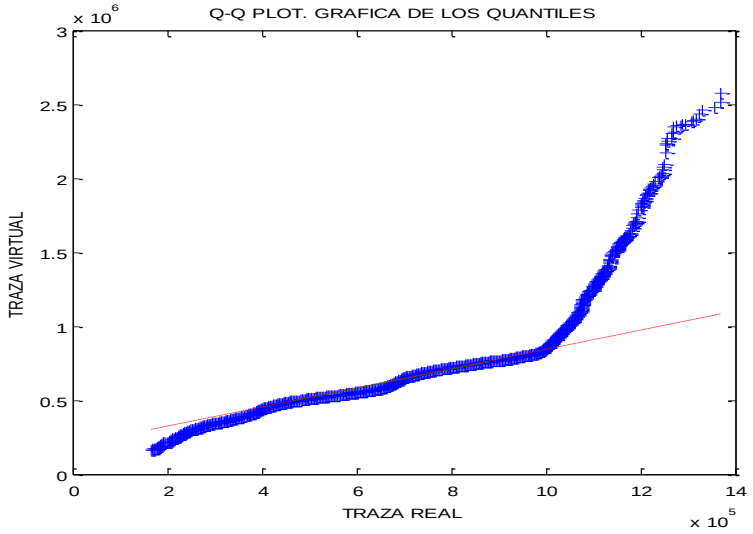
PRUEBA DE BONDAD : QUANTILES

Quantiles para la traza de tráfico, para el análisis en tramas por unidad de tiempo y para bits por unidad de tiempo.

Tráfico en (tramas por unidad de tiempo)

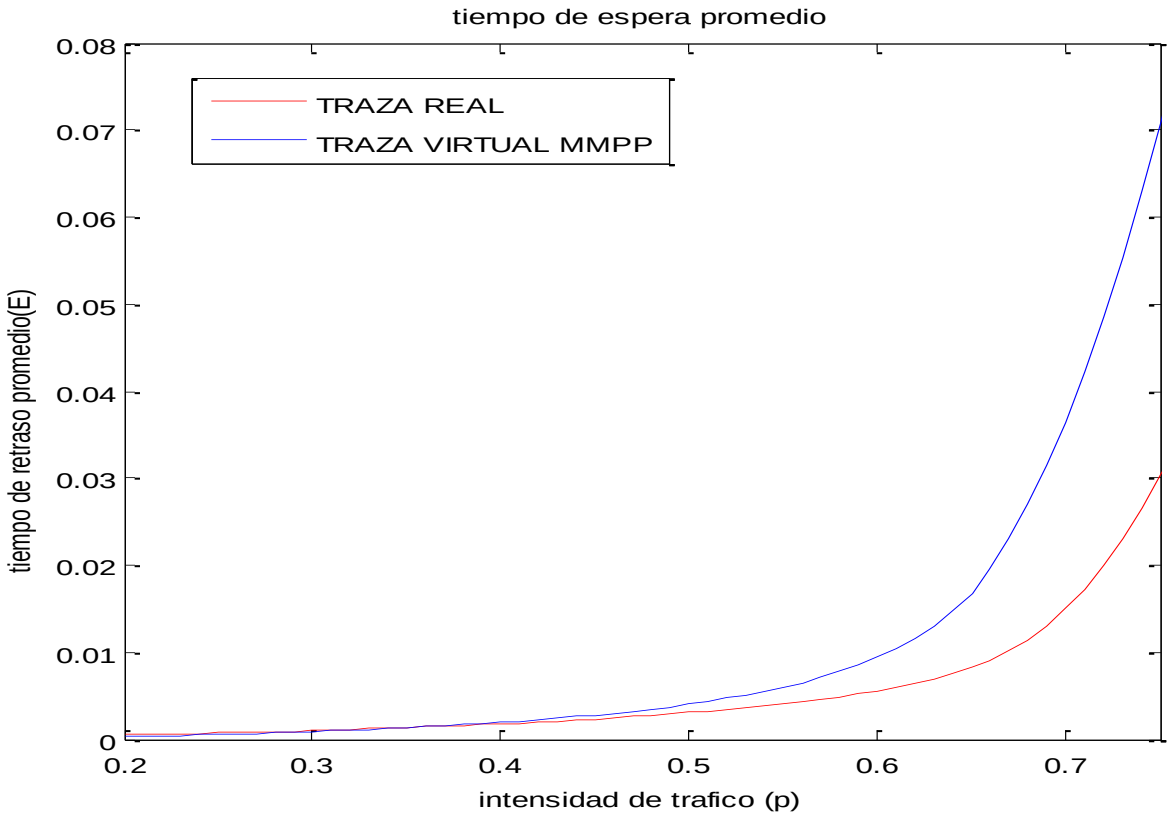


Tráfico en (bits por unidad de tiempo)



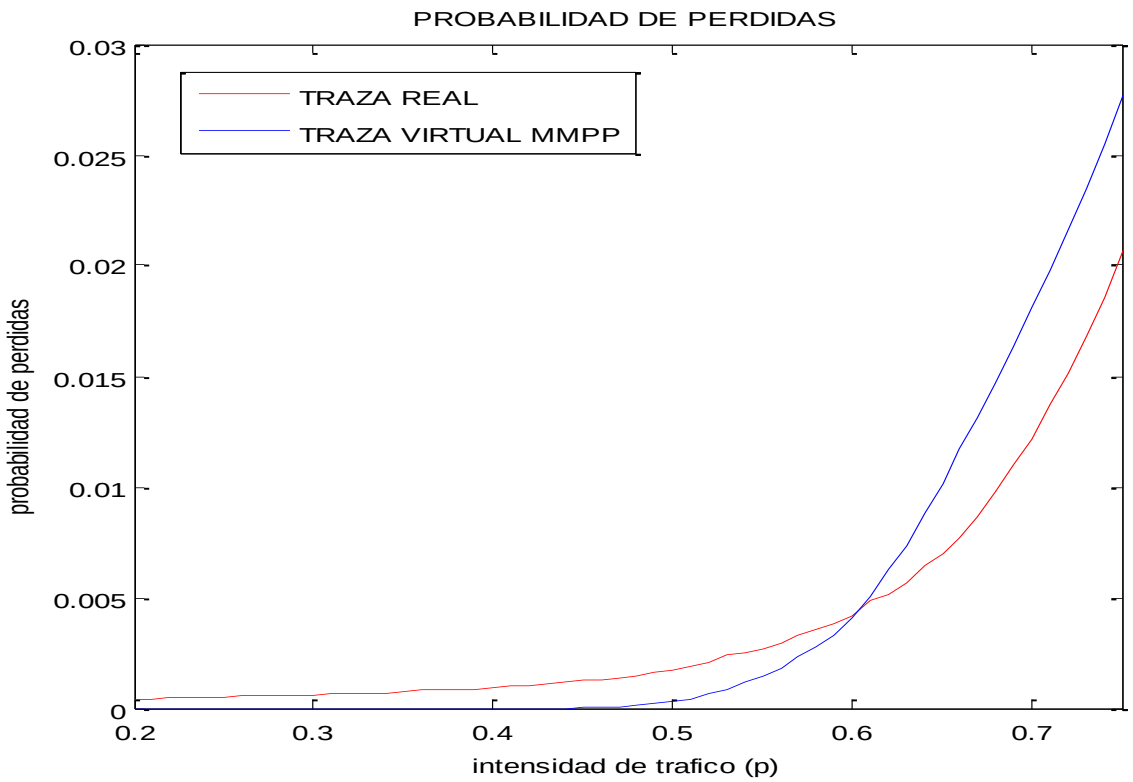
Validación modelo MMPP/D/1

Tiempo de espera promedio, Búfer infinito

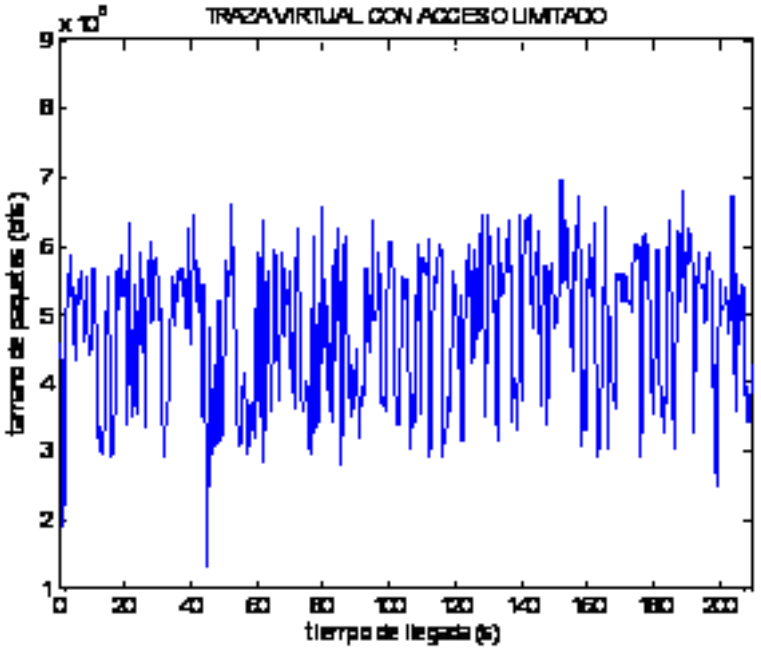


Análisis del sistema de cola MMPP/D/1

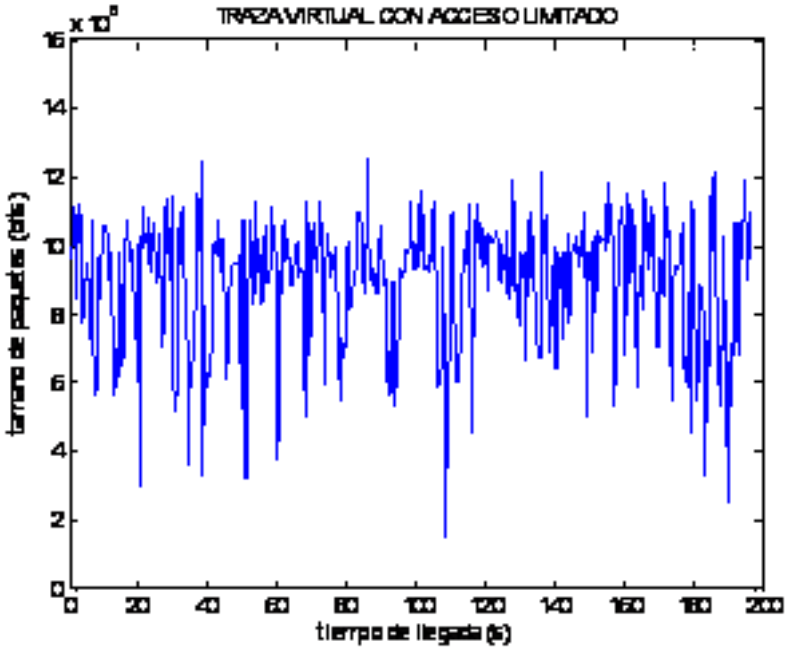
Probabilidad de pérdidas, búfer=10



Respuesta del enlace cuando se limita su capacidad



Tráfico virtual cuando se limita la capacidad del canal a 550Kbits/Δt para una captura simultanea



Tráfico virtual cuando se limita la capacidad del canal a 1Mbits/Δt para una captura individual