Tema: Modelos de Colas
(Modelos de Líneas de Espera)

Mg. Ariel González (Departamento de Computación – Fac. Cs. Exactas – UNRC). Mail: agonzalez(AT)dc.exa.unrc.edu.ar)

Propósito

- La simulación es muy usada en el análisis de los sistemas de colas.
- Los modelos de Colas, proveen al analista una herramienta poderosa para diseñar y evaluar la perfomance de los sistemas de colas (ya sea con soluciones analíticas o mediante simulaciones).
- Teoría de Colas y Simulación: predecir medidas de perfomance en función de los parámetros de entrada.
- *Parametros ?*
- Medidas típicas de perfomance de los sistemas:
  - Utiización del Servidor, longitud de las líneas de espera, y retardo de los clientes.
- Para sistemas relativamente simples → calc. análiticos.
- Para modelos de sistemas complejos → simulación es requerida.

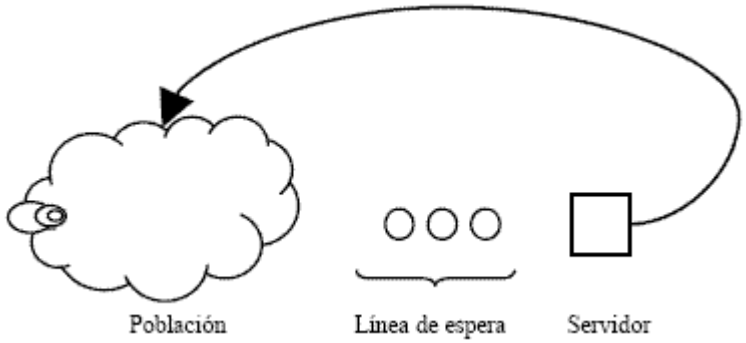
Esquema de los temas

- Discusión de algunos modelos mas conocidos:
  - Caraterísticas,
  - Significados y las relaciones de las medidas de rendimiento importante ,
  - Estimación de las medidas promedio de perfomance.
  - Efecto al variar los parámetros de entrada,
  - Soluciones matemáticas de algunos modelos de colas básicos.

Características

- Elementos claves:
  - Clientes: refiere a cualquier entidad que arriba al sistema y requiere un servicio, ej. gente, maquinas, emails.
  - Servidor: refiere a cualquier recurso que provee el servicio solicitado, ej., personas que reparan, máquinas lavarropas, pistas de aterrizaje en el aeropuerto.

Un tipitco modelo simple de Colas:



Población (*calling population*)

- La Población de potenciales clientes puede ser asumida como finita o infinita.
  - Modelos de población finita: si la tasa de arribos depende del número de clientes que están siendo servidos y esperando, ej. modelo de un avión de la empresa, si se repara, la tasa de llegada de reparación se convierte en cero.
  - Modelos de población infinita: si la tasa de llegada no se ve afectada por el número de clientes que se sirven y esperan, ej., los sistemas con gran población de clientes potenciales. Ej. autos que pasan frente a la universidad.
  - Diferencias principales:
    - Población infinita: tasa constante y aleatoria
    - Población finita: dependerá del número de clientes que se sirven y esperan.

Capacidad del Sistema

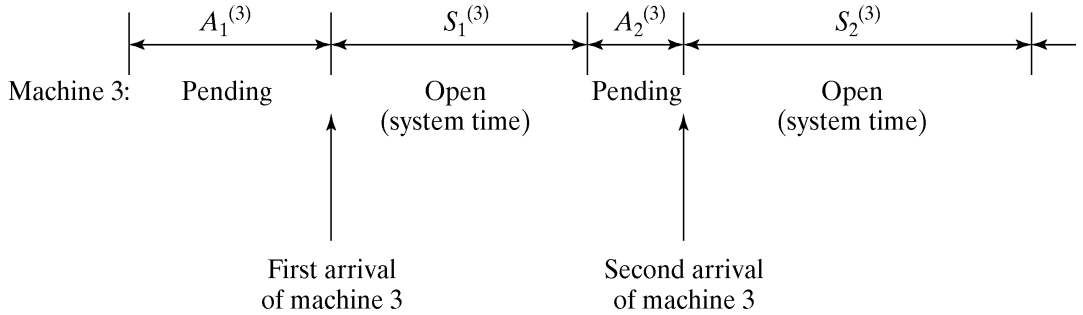
- Límite sobre el número de clientes en la cola o el sistema.
  - Capacidad limitada, ej., un lavadero de autos con capacidad para 10 autos en cola de espera.
    - Distinguir entre tasa de arribos y tasa de arribos efectiva.
  - Capacidad ilimitada, ej., número de gente que puede hacer cola para sacar entrada en la cancha de futbol.

Proceso de Arribo

- Para modelos de población infinita:
  - En termino de los tiempos de interarribo.
  - Arribos aleatorios: interarribos caracterizados por una distribución.
  - Arribos planificados: interarribos pueden ser constantes o constantes mas o menos una pequeño tiempo aleatorio.
    - Ej., parada de un colectivo.
  - Al menos un clientes es asumido a estar siempre presente. El servidor nunca está ocioso
    - Ej., suficiente materia prima para una máquina.

Proceso de Arribo

- Para Modelos de población finita:
  - Un cliente es “pendiente” cuando el cliente se encuentra fuera del sistema de colas,
    - por ejemplo, el problema de reparación de máquinas: una máquina está "pendiente" cuando esté en funcionamiento, se convierte en "*no pendiente*" en el instante en que demanda el servicio técnico de reparaciones.
  - "*Tiempo de ejecución*“ (*Runtime*) de un cliente es la longitud del tiempo desde la salida del sistema de espera, hasta el próximo arribo de ese cliente al sistema de cola,
    - por ejemplo, en el problema de reparación de la máquina, cual es el *runtime* de una máquina?



Comportamiento y Disciplina de la Cola

- Comportamiento de la Cola: acciones de un cliente mientras espera por un servicio, por ejemplo:
  - Negarse
  - Abandonar
  - Moverse a otra cola.
- Disciplina de una cola::
  - First-in-first-out (FIFO)
  - Last-in-first-out (LIFO)
  - Service in random order (SIRO)
  - Shortest processing time first (SPT)
  - Service according to priority (PR).

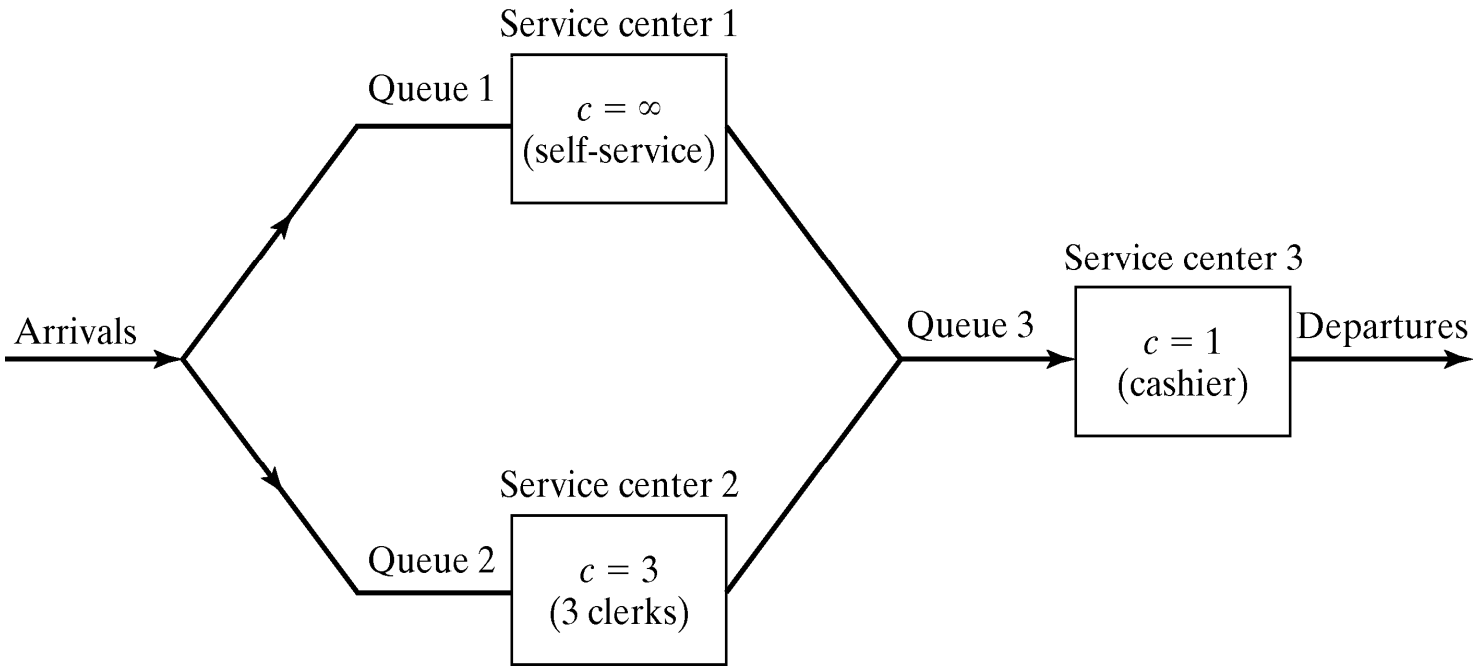
Tiempos de Servicios y Mecanismos de Servicios

- Los tiempos de servicios son denotados por S_1, S_2, S_3.
  - Puden ser Constantes o Aleatorios.
  - En algunos casos depende de la hora del día, o de la longitud de la cola.

- Un sistema de colas consiste de un núm. de Centros de Servicios y colas, interconectadas.
  - Cada Centro de servicio consiste de algún número de servidores, *c*, trabajando en paralelo. Un cliente al llegar a la salida de la cola toma el primer servidor disponible.

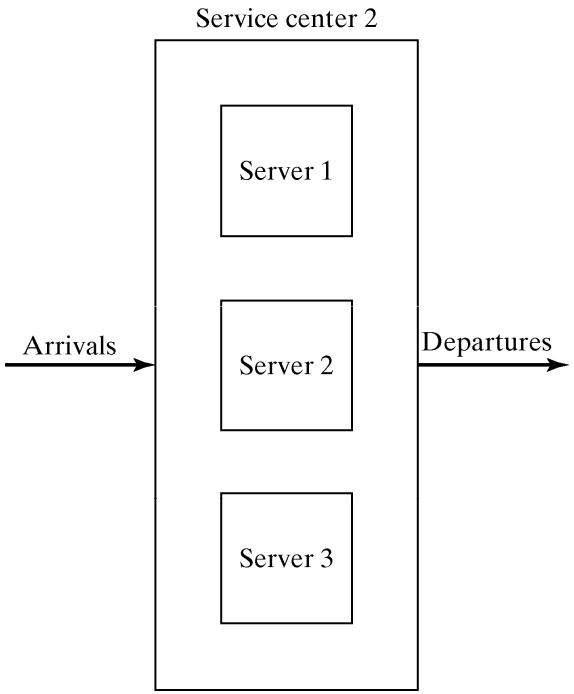
Tiempos de Servicios y Mecanismos de Servicios

- Ejemplo: considerar un supermercado, donde los clientes pueden:
  - Autoservirse antes de pasar por el cajero:



Tiempos de Servicios y Mecanismos de Servicios

- Esperar por uno de los 3 empleados:



- Servicio por Lotes (un servidor sirve varios clientes simultaneamente, o un cliente requiere de varios servidores simultaneamente).

Notación

- Un sistema de notación para sistemas de servidores paralelos: *A/B/c/N/K*
  - *A* representa la distribucion del tiempo de interarribo,
  - *B* representa la distribucion del tiempo de servicio,
  - *c* representa el numero de servidores paralelos,
  - *N* representa la capacidad del sistema,
  - *K* representa el tamaño de la población.

- Ejemplo: M/D/5/∞/ ∞
  - Interarrivo exponencial,
  - Tiempo de servicio constante,
  - 5 servidores
  - Y capacidad y tamaño de la población infinita.

Notación

- Medidas de perfomance de los sistemas de colas:

P_n	Probabilidad de estado estacionario de tener n clientes en el sistema
$P_n(t)$	Probabilidad de tener n clientes en el sistema en tiempo t
λ	Tasa de llegadas
λ_e	Tasa efectiva de llegadas
μ	Tasa de servicio de un servidor
ρ	Utilización del servidor
A_n	Tiempo entre llegadas del cliente $n-1$ y el cliente n
S_n	Tiempo de servicio del n-ésimo cliente arribado
W_n	Tiempo total en el sistema del n-ésimo cliente arribado
W_n^Q	Tiempo total esperando en la cola del n-ésimo cliente arribado
$L(t)$	El número de clientes en el sistema en tiempo t
$L_Q(t)$	El número de clientes en la cola en tiempo t
L	Número medio de clientes en el sistema en el tiempo a largo plazo
L_Q	Número medio de clientes en la cola en el tiempo a largo plazo
w	Tiempo medio pasado en el sistema por cliente a largo plazo
w_Q	Tiempo medio pasado en la cola por cliente a largo plazo

Número medio de clientes en el Sistema(*L)*

- Considere un sistema de colas sobre un período de tiempo *T*,
  - T_i denota el tiempo total durante *[0, T]* en el que el sistema contenía exactamente los *i* clientes, el número medio ponderado en el tiempo en el sistema se define por :

$$\hat{L} = \frac{1}{T}\sum_{i=0}^{\infty} iT_i = \sum_{i=0}^{\infty} i\left(\frac{T_i}{T}\right)$$

  - Considerando el area bajo la función *L(t)*, entonces,

$$\hat{L} = \frac{1}{T}\sum_{i=0}^{\infty} iT_i = \frac{1}{T}\int L(t)dt$$

  - Para sistemas estables:

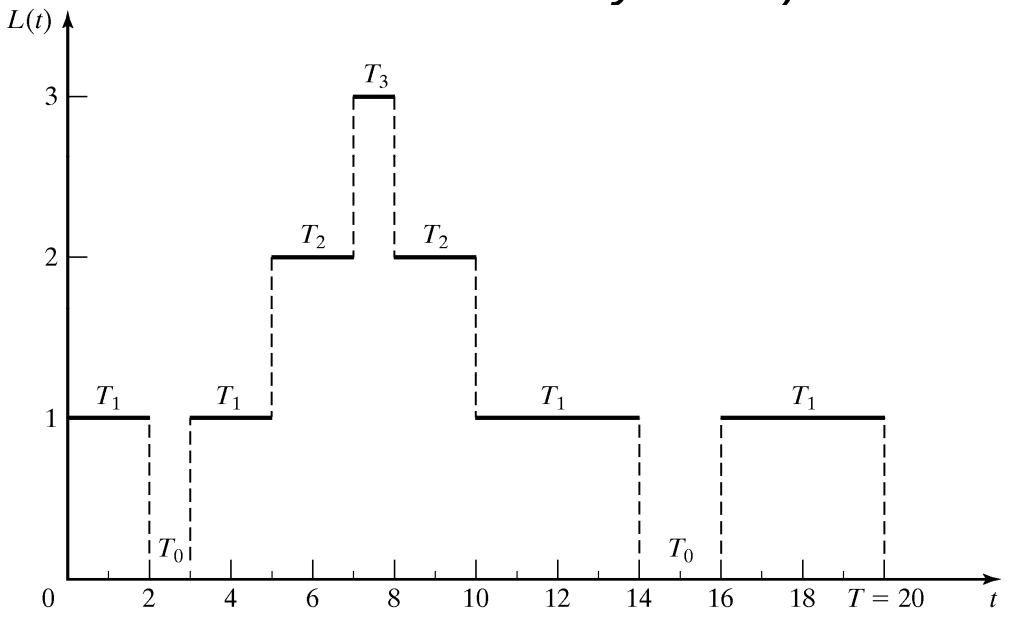
$$\hat{L} = \frac{1}{T}\int L(t)dt \to L \quad \text{as} \quad T \to \infty$$

Número medio de clientes en la cola(L_Q)

- Número promedio ponderado en el tiempo de clientes en la cola es:

$$\hat{L}_Q = \frac{1}{T}\sum_{i=0}^{\infty} iT_i^Q = \frac{1}{T}\int L_Q(t)dt \to L_Q \quad \text{as} \quad T \to \infty$$

- *Ejemplo G/G/1/N/K*: considere los resultados de un sistema de colas con *N > 4 y K > 4*) dados en la siguiente fig.:



$$\hat{L} = ? \text{ clientes}$$

$$L_Q(t) = \begin{cases} 0, & \text{if } L(t) = 0 \\ L(t) - 1, & \text{if } L(t) \geq 1 \end{cases}$$

$$\hat{L}_Q = 0{,}3 \text{ clientes}$$

Tiempo medio de permanencia por cliente en el sistema (*w*)

- También llamado tiempo medio en el sistema:

$$\hat{w} = \frac{1}{N}\sum_{i=1}^{N} W_i$$

donde $W_1, W_2, \ldots, W_N$ son los tiempos individuales que cada cliente pasa en el sistema durante $[0,T]$.

  - Para sistemas estables: $\hat{w} \to w \quad \text{as} \quad N \to \infty$
  - Si el sistema bajo consideracion es la cola (*delay*):

$$\hat{w}_Q = \frac{1}{N}\sum_{i=1}^{N} W_i^Q \to w_Q \quad \text{as} \quad N \to \infty$$

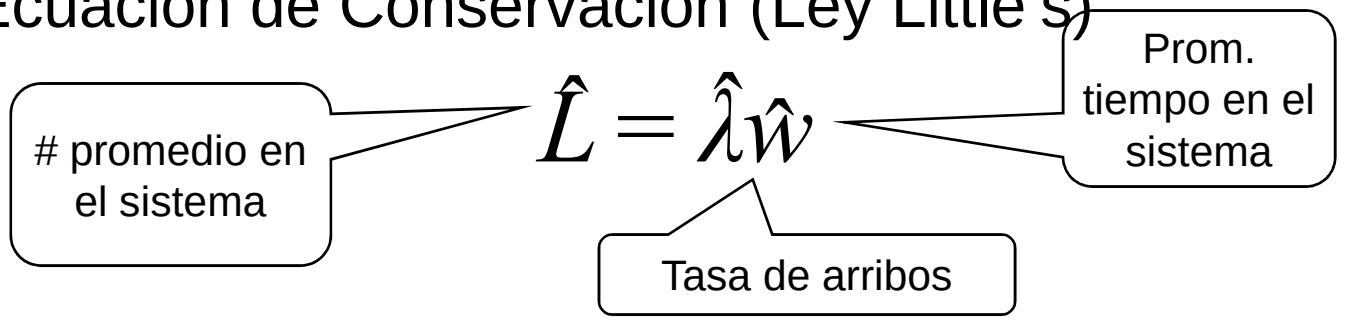
  - *Continuando con el ejemplo G/G/1/N/K*:

$$W = \frac{W_1 + W_2 + \ldots + W_5}{5} = \frac{2+5+5+7+4}{5} = 4.6 \text{ unidades de tiempo ¿}$$

$$\hat{w}_Q = ? \text{ unidadesde tiempo}$$

Ecuación de Conservación

- Ecuación de Conservación (Ley Little's)



$$L = \lambda w \quad \text{cuando} \quad T \rightarrow \infty \quad \text{y} \quad N \rightarrow \infty$$

  - Es válido para casi todos los sistemas o subsistemas de cola (independientemente del número de servidores, la disciplina de cola, u otras circunstancias especiales).
  - *Ejemplo G/G/1/N/K*: en promedio hay 1 arribo cada 4 unidades de tiempo y cada arribo gasta *4.6* unidades de tiempo en el sistema en promedio. En un punto arbitrario en el tiempo, hay *(1/4)(4.6) = 1.15* clientes presentes en promedio.
  - Demostración (pag. 239).

Utilización del Servidor

- La porción de tiempo que el servidor está ocupado.
  - $\hat{\rho}$, es definido sobre un intervalo de tiempo *[0,T]*.
  - ρ *,* es la utilización a largo plazo.
  - Para sistema estables a largo plazo:

$$\hat{\rho} \to \rho \text{ as } T \to \infty$$

- Continuación del ejemplo.

Utilización del Servidor

- Para colas *G/G/1/∞/∞*:
  - Considere: tasa de arribos de λ clientes por unidad de tiempo, y tiempo de servicio promedio *E(S) = 1/*μ unidades de tiempo, ent.
  - El servidor solo es un subsistema que puede ser consideado como un sistema de colas. ***COMO ??***
  - La ley de Little`s, *L* = λw, podría ser aplicada.
  - Para un sistema estable, la tasa de arribos promedio al servidor λ_s, debe ser idéntica a λ.
  - EL número promedio de clientes en el servidor es:

$$\hat{L}_s = \frac{1}{T}\int \left(L(t) - L_Q(t)\right)dt = \frac{T - T_0}{T}$$

Utilización del Servidor

- En general, para un sistema de colas de un servidor:

$$\hat{L}_s = \hat{\rho} \rightarrow L_s = \rho \quad \text{cuando} \quad T \rightarrow \infty$$

$$\text{y} \qquad \rho = \lambda E(s) = \frac{\lambda}{\mu}$$

  - Para un sistema de 1-servidor estable: $\rho = \frac{\lambda}{\mu} < 1$
  - Para un sistema inestable ($\lambda > \mu$), la utilización del servidor a largo plazo es *1*.

Utilización del Servidor

- Para colas *G/G/c/∞/∞* :
  - Con *c* servidores idénticos en paralelo.
  - Para sistemas estables, el número promedio de servidores ocupados, L_s, es: $L_s = \lambda E(s) = \lambda/\mu$.
  - La utilización del servidor es:

$$\rho = \frac{L_s}{c} = \frac{\lambda}{c\mu}\ , \quad \text{donde} \quad \lambda < c\mu \quad \text{para sistemas estables}$$

  - Cuanto vale la tasa de servicio máxima?

  - Ejemplo: “Renovación de Licencias”. (pag. 243)

Utilización del Servidor y Perfomance del Sistema

La perfomance del sistema puede variar ampliamente para un valor dado de utilización ρ.

- En una cola *D/D/1* donde $E(A) = 1/\lambda$ y $E(S) = 1/\mu$, donde:

$$L = \rho = \lambda/\mu, \quad w = E(S) = 1/\mu, \quad L_Q = W_Q = 0.$$

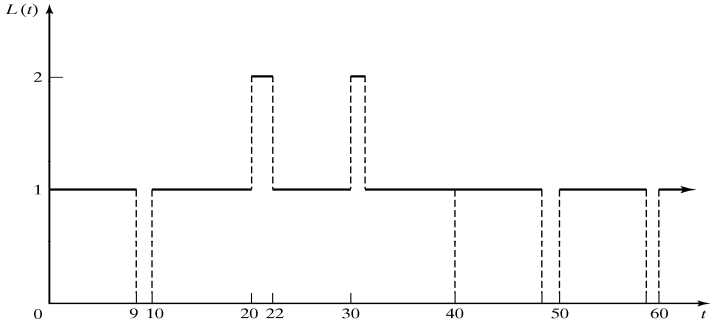
  - Variando λ y μ, la utilización del servidor toma valores entre *0* y *1*.
  - Sin embargo, no hay esperas.
- En general, la **variabilidad** de los interarribos y tiempos de servicio causan fluctuaciones en la longitud de las colas.

Utilización del Servidor y Perfomance del Sistema

- Ejemplo: Un médico planifica la atención de los pacientes cada 10 min. y pasa S_i min. con el $i^{e\text{-}simo}$ paciente:

$$S_i = \begin{Bmatrix} 9 \text{ minutes with probability } 0.9 \\ 12 \text{ minutes with probability } 0.1 \end{Bmatrix}$$

  - Los arribos son deterministicos, $A_1 = A_2 = \ldots = \lambda^{-1} = 10$.
  - Los servicios son estocásticos, $E(S_i) = 9.3$ min y $V(S_0) = 0.81$ min^2.
  - La utilización del medico es $\rho = \lambda/\mu = 0.93 < 1$.
  - Considerar que el sistema es simulado con tiempos de servicio: $S_1 = 9$, $S_2 = 12$, $S_3 = 9$, $S_4 = 9$, $S_5 = 9$, ….



  - La ocurrencia de un tiempo de servicio relativamente largo ($S_2 = 12$) causa una línea de espera temporal.

Problemas de Costos en los sistemas de colas

- Los costos pueden ser asociados con varios aspectos de las líneas de espera o servidores:
  - El sistema podría incurrir en un costo por cada cliente en la cola, a una tasa de *$10* por hora por cliente.
    - El costo promedio por cliente:

$$\sum_{j=1}^{N} \frac{\$10 * W_j^Q}{N} = \$10 * \hat{w}_Q$$

    - Si $\hat{\lambda}$ clientes arriban por hora (en promedio), el costo promedio por hora es:

$$\left(\hat{\lambda}\frac{\text{clientes}}{\text{hora}}\right)\left(\frac{\$10 * \hat{w}_Q}{\text{cliente}}\right) = \$10 * \hat{\lambda}\hat{w}_Q = \$10 * \hat{L}_Q \text{ / hora}$$

  - Los servidores pueden imponer costos sobre el sistema, si un grupo de *c* servidores paralelos ($1 \le c \le \infty$) tienen utilización *r*, cada servidor impone un costo de *$5* por hora mientras está ocupado.
    - EL costo total del Servidor es: $\$5 * c\rho$.
  - Si los costos estan asociados al tiempo ocioso, como es el costo total?

Comportamiento de estado-estable de los Modelos de Markov de población infinita

- Modelos Markovianos: arribos exponenciales (tasa de arribo media = λ).
- Los tiempos de servicio puede ser exponencialmente distribuidos (*M*) o arbitrarios (*G*).
- Disciplina de la cola: FIFO.
- Un sistema de colas se encuentra en “equilibrio estadístico” si la probabilidad de que el sistema esté en un estado determinado no depende del tiempo: $P(L(t) = n) = P_n(t) = P_n$.

- Los modelos matemáticos pueden ser usados para obtener resultados aproximados aun cuando los supuestos del modelo no se aplican estrictamente (como una guía aproximada).
- La Simulación puede ser usada para un análisis mas refinado (es una representación mas fiel para sistemas complejos).

Comportamiento de estado-estable de los Modelos de Markov de población infinita

- Para los modelos simples vistos, el parametro de estado estacionario *L*, puede ser computado como :

$$L = \sum_{n=0}^{\infty} nP_n$$

  - Si es conocido *L*, aplicando la ley de Little's a el sistema completo y a la cola sola:

$$w = \frac{L}{\lambda}, \qquad w_Q = w - \frac{1}{\mu}$$

$$L_Q = \lambda w_Q$$

- Ejemplo *G/G/c/∞/∞* : para alcanzar un equilibrio estadistico (estado estable), es condición necesaria y suficiente que $\lambda/(c\mu) < 1$.

Colas M/G/1

- Colas de un servidor con arribos Poisson y capacidad ilimitada.
- Suponiendo que los tiempos de servicios tienen media $1/\mu$ y varianza σ^2 y $\rho = \lambda/\mu < 1$, los parámetros de estado estable de una cola *M/G/1* son:

$$\rho = \lambda/\mu, \qquad P_0 = 1-\rho$$

$$L = \rho + \frac{\rho^2\left(1+\sigma^2\mu^2\right)}{2\left(1-\rho\right)}, \qquad L_Q = \frac{\rho^2\left(1+\sigma^2\mu^2\right)}{2\left(1-\rho\right)}$$

$$w = \frac{1}{\mu} + \frac{\lambda\left(1/\mu^2+\sigma^2\right)}{2\left(1-\rho\right)}, \qquad w_Q = \frac{\lambda\left(1/\mu^2+\sigma^2\right)}{2\left(1-\rho\right)}$$

Colas M/G/1

- No hay una expresión simple para las probabilidades P_0, P_1, …
- *Notar que* $L - L_Q = \rho$, el numero promedio ponderado en el tiempo de clientes que estan siendo servidos.
- La longitud promedio de la cola, L_Q, puede ser reescribirse como:

$$L_Q = \frac{\rho^2}{2(1-\rho)} + \frac{\lambda^2 \sigma^2}{2(1-\rho)}$$

  - Si λ y μ son constantes, L_Q depende solo de la variabilidad, σ^2, de los tiempos de servicios.

- *No se debe confundir "estado-estable" con baja variabilidad o líneas de espera cortas.*

Colas M/G/1

- Ejemplo: Dos trabajadores compiten por un trabajo, el primero dice ser mas rápido en promedio que el segundo, pero el segundo dice ser mas consistente (pag. 249),
  - Arribos Poisson con tasa $\lambda = 2$ por hora ($1/30$ por min.).
  - El primero: $1/\mu = 24$ min. y $\sigma^2 = 20^2 = 400$ min^2:

$$L_Q = \frac{(1/30)^2[24^2 + 400]}{2(1-4/5)} = 2.711 \text{ clientes}$$

    - La porción de arribos que encuentran al servidor desocupado es:

    $P_0 = 1-\rho = 1/5 = 20\%.$
  - El segundo: $1/\mu = 25$ min. y $\sigma^2 = 2^2 = 4$ min^2:

$$L_Q = \frac{(1/30)^2[25^2 + 4]}{2(1-5/6)} = 2.097 \text{ clientes}$$

    - La porción de arribos que encuentran al servidor desocupado es :

    $P_0 = 1-\rho = 1/6 = 16.7\%.$

  - ***Conclusion?***

Colas M/M/1

- Suponemos que los tiempos de servicio en una cola *M/G/1* son exponencialmente distribuidos con media $1/\mu$, entonces la varianza es $\sigma^2 = 1/\mu^2$.
  - *Los modelos de colas M/M/1* son muy usados cuando los tiempos de servicios tienen una desviación estandard aproximadamente igual a su media.
  - Los parametros de estado-estable son:

$$\rho = \lambda/\mu, \qquad P_n = (1-\rho)\rho^n$$

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

$$w = \frac{1}{\mu-\lambda} = \frac{1}{\mu(1-\rho)}, \qquad w_Q = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{\rho}{\mu(1-\rho)}$$

Colas M/M/1

- Ejemplo: suponga una cola *M/M/1* con una tasa de servicio $\mu=10$ clientes por hora.
  - Observar como *L* y *w* se incrementan en función de los incrementos de la tasa de arribos, λ, que aumenta desde *5* a *8.64* por incrementos del *20*%:

λ	*5,0*	*6,0*	*7,2*	*8,64*	*10,0*
ρ	*0,500*	*0,600*	*0,720*	*0,864*	*1,000*
L	*1,00*	*1,50*	*2,57*	*6,35*	∞
w	*0,20*	*0,25*	*0,36*	*0,73*	∞

  - Si $\lambda/\mu \geq 1$, las líneas de espera tienden a crecer continuamente.
  - Los incrementos de *w* y de *L* son altamente no lineales como una función de ρ.
  - Ejemplo de la peluquería (pag. 250).

El Efecto de la Utilización y Variabilidad del Servicio

- En la mayoria de los sistemas de colas, si las líneas de espera son demasiado largas, como pueden ser reducidas?
- Una medida de la variabilidad de una variable Aleatoria *X* es el *coeficiente de variación* (cv):

$$(cv)^2 = \frac{V(X)}{[E(X)]^2}$$

  - Cuanto mayor es su valor, mas variable es la distribución relativa a su valor esperado.
  - Cuanto vale para tiempos de servicios deterministicos?
  - Para tiempos de servicio exponenciales con tasa de servicio μ , el tiempo de servicio medio es E(X)=1/μ , y la varianza es 1/μ^2 , así *cv*=1.
  - Reescribiendo los calculos de las colas *M/G/1*

El Efecto de la Utilización y Variabilidad del Servicio

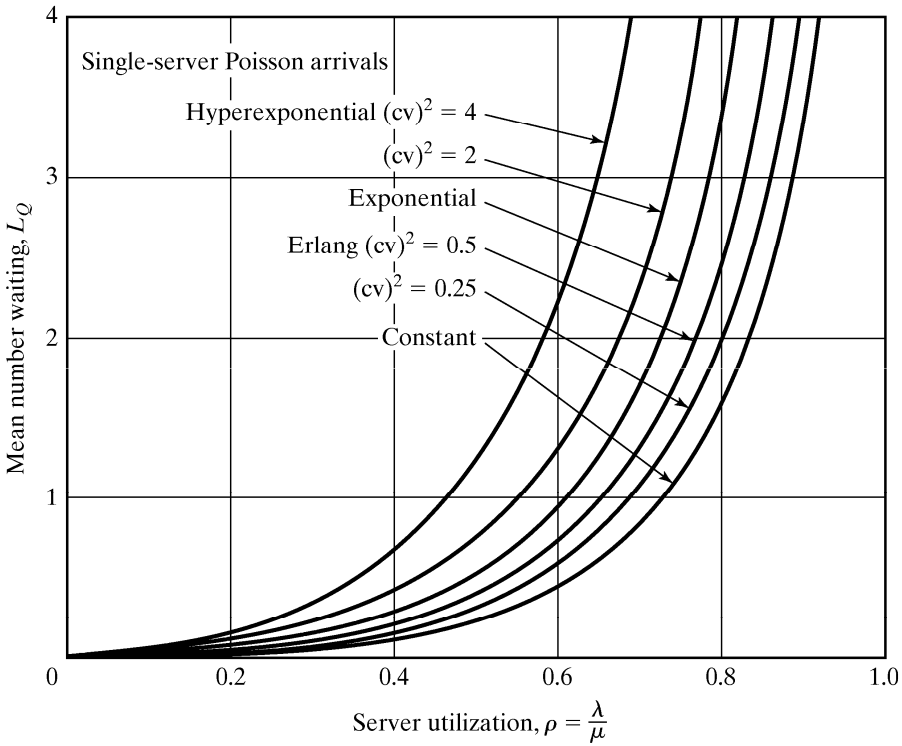
- Considerar L_Q para una cola *M/G/1*:

$$L_Q = \frac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)}$$

$$\left(\frac{\rho^2}{1-\rho}\right)\left(\frac{1+(cv)^2}{2}\right)$$

L_Q para colas *M/M/1*

"*Factor de Corrección*": corrige la fórmula de una cola *M/M/1* para distr. con tiempos de servicio NO exponencial



- El factor de corrección es solo aplicado a L_Q y w_Q

Colas Multiserver

- *Una cola M/M/c/∞/∞* : *c* canales (servidores) operando en paralelo.
  - Cada canal tiene una distr. de tiempo de servicio exponencial, son independientes e idénticos con media $1/\mu$.
  - Para alcanzar un estado-estable, la carga ofrecida (λ/μ) debe satisfacer $\lambda/\mu < c$, donde $\lambda/(c\mu) = \rho$ es la utilización del Servidor.

Colas Multiserver

- Estimaciones de estado-estable para *M/M/c/∞/∞* :

ρ	$\dfrac{\lambda}{c\mu}$
P_0	$\left\{\left[\sum_{n=0}^{c-1}\dfrac{(\lambda/\mu)^n}{n!}\right]+\left[\left(\dfrac{\lambda}{\mu}\right)^c\left(\dfrac{1}{c!}\right)\left(\dfrac{c\mu}{c\mu-\lambda}\right)\right]\right\}^{-1}$ $=\left\{\left[\sum_{n=0}^{c-1}\dfrac{(c\rho)^n}{n!}\right]+\left[(c\rho)^c\left(\dfrac{1}{c!}\right)\dfrac{1}{1-\rho}\right]\right\}^{-1}$
$P(L(\infty)\geq c)$	$\dfrac{(\lambda/\mu)^c P_0}{c!(1-\lambda/c\mu)}=\dfrac{(c\rho)^c P_0}{c!(1-\rho)}$
L	$c\rho+\dfrac{(c\rho)^{c+1}P_0}{c(c!)(1-\rho)^2}=c\rho+\dfrac{\rho P(L(\infty)\geq c)}{1-\rho}$
w	$\dfrac{L}{\lambda}$
w_Q	$w-\dfrac{1}{\mu}$
L_Q	$\lambda w_Q=\dfrac{(c\rho)^{c+1}P_0}{c(c!)(1-\rho)^2}=\dfrac{\rho P(L(\infty)\geq c)}{1-\rho}$
$L-L_Q$	$\dfrac{\lambda}{\mu}=c\rho$

Colas Multiserver

- Otros modelos de colas multiserver:
  - *M/G/c/∞*: Los parámetros pueden ser aproximados desde los modelos *M/M/c/∞/∞*. *Como ?*
  - *M/G/∞:* número de servidores o canales infinitos, ej., los clientes son su propio sistema, la capacidad del servicio excede ampliamente la demanda del servicio.
    - *ejemplo (pag. 257)*
    - *Estimaciones a largo*

    Plazo:

P_0	$e^{-\lambda/\mu}$
w	$\frac{1}{\mu}$
w_Q	0
L	$\frac{\lambda}{\mu}$
L_Q	0
P_n	$\frac{e^{-\lambda/\mu}(\lambda/\mu)^n}{n!}$, n=0,1,...

  - *M/M/C/N/∞*: capacidad total del sistema es *N (≥c)* clientes.

Colas Multiserver

- Parametros de estado-estable para colas *M/M/C/N/∞* , donde*:*
  N =capacidad del sistema, a=λ/μ, p= λ/cμ.

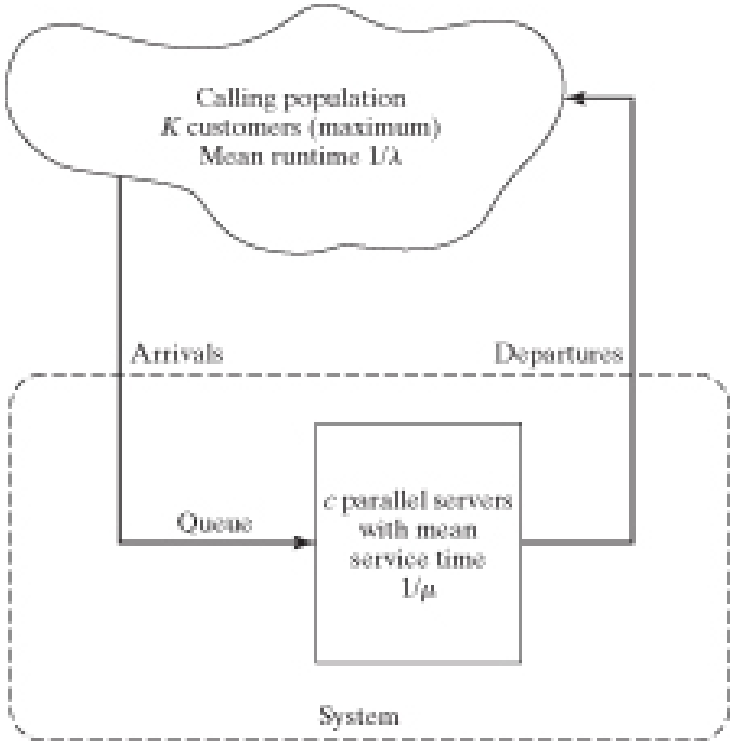
P_0	$\left[1+\sum_{n=1}^{c}\frac{a^n}{n!}+\frac{a^c}{c!}\sum_{n=c+1}^{N}\rho^{n-c}\right]^{-1}$
P_n	$\frac{a^N}{c!c^{N-c}}P_0$
L_Q	$\frac{P_0a^c\rho}{c!(1-\rho)^2}[1-\rho^{N-c}-(N-c)\rho^{N-c}(1-\rho)]$
λ_e	$\lambda(1-P_N)$
w_Q	$\frac{L_Q}{\lambda_e}$
w	$w_Q+\frac{1}{\mu}$
L	$\lambda_e w$

- Analizar λ_e.
- Que pasa con la utilización del servidor: aumenta o disminuye respecto al mismo problema sin restricción de capacidad?

Comportamiento de Estado-Estable de modelos con Población finita.

- Cuando la población demandante es pequeña, la presencia de uno o mas clientes dentro del sistema tiene un fuerte efecto sobre la distribución de los futuros arribos.
- Considere un modelo con población finita formada con *K* clientes (*M/M/c/K/K*):

  • El tiempo entre el final de un servicio y el próximo requerimiento del servicio es exponencialmente distribuído (Media = $1/\lambda$).
  • Tiempos de Servicios son también exponencialmente distribuídos.
  • *c* servidores en paralelo y sistema con capacidad *K*.



Comportamiento de Estado-Estable de modelos con Población finita.

- Algunos parámetros de estado-estable son:

$$P_0 = \left\{ \sum_{n=0}^{c-1} \binom{K}{n} \left(\frac{\lambda}{\mu}\right)^n + \sum_{n=c}^{K} \frac{K!}{(K-n)!c!c^{n-c}} \left(\frac{\lambda}{\mu}\right)^n \right\}^{-1}$$

$$P_n = \begin{cases} \binom{K}{n} \left(\frac{\lambda}{\mu}\right)^n P_0, & n = 0,1,\ldots,c-1 \\ \frac{K!}{(K-n)!c!c^{n-c}} \left(\frac{\lambda}{\mu}\right)^n, & n = c,c+1,\ldots K \end{cases}$$

$$L = \sum_{n=0}^{K} nP_n, \qquad w = L/\lambda_e, \qquad \rho = \lambda_e/c\mu$$

donde λ_e es la tasa de arribo efectiva a largo plazo de clientes que entran a la cola (o que entran/salen del servicio)

$$\lambda_e = \sum_{n=0}^{K} (K-n)\lambda P_n$$

Comportamiento de Estado-Estable de modelos con Población finita.

- Ejemplo: 2 trabajadores son responsables de *10* máquinas fresadoras.
  - Las máquinas trabajan sobre un promedio de 20 min., luego requieren de un período de servicio que lleva en promedio unos 5 minutos, ambos tiempos son exponencialmente distr.: $\lambda = 1/20$ y $\mu = 1/5$.
  - Todas las medidas de perfomance dependen de P_0:

$$P_0 = \left\{ \sum_{n=0}^{2-1} \binom{10}{n} \left(\frac{5}{20}\right)^n + \sum_{n=2}^{10} \frac{10!}{(10-n)!\,2!\,2^{n-2}} \left(\frac{5}{20}\right)^n \right\}^{-1} = 0.065$$

    - Luego se pueden obtener las P_n.
    - El número esperado de máquinas en el sistema es:

$$L = \sum_{n=0}^{10} nP_n = 3.17 \text{ machines}$$

    - *Cual es el cantidad promedio de máquinas trabajando?*

$$K - L = 10 - 3.17 = 6.83 \text{ machines}$$

    - Que pasaría si el número de servidores es incrementado o decrementado?

Redes de Colas

- Muchos sistemas son naturalmente modelados como redes de simples colas: los clientes salen de una cola y pasan a otra.
- Los siguientes resultados asumen un sistema estable con población demandante infinita y sin límite de capacidad de sistema:
  - Suponiendo que ningún cliente es creado ni destruído en las colas ➔
    - La tasa de salida de una cola es la misma que la de arribo de la cola destino (a largo plazo).
  - Si los clientes arriban a la cola *i* con tasa λ_i, y una fracción $0 \leq p_{ij} \leq 1$ de ellos son ruteados a la cola *j*, entonces la tasa de arribos desde la cola *i* a la cola *j* es $\lambda_i p_{ij}$ (a largo plazo).

Redes de Colas

- La tasa de arribo total a la cola *j*:

$$\lambda_j = a_j + \sum_{\text{all } i} \lambda_i p_{ij}$$

Tasa de Arribo desde fuera de la red

Suma de las tasa de arribo desde otras colas de la red

- Si la cola *j* tiene $c_j < \infty$ servidores paralelos trabajando con una tasa μ_j, entones la utilización a largo plazo de cada servidor es:
  - $\rho_j = \lambda_j/(c\mu_j)$ (se debe satisfacer $\rho_j < 1$ para que todas las colas sean estables).

- Si los arribos desde fuera de la Red siguen un proceso Poisson con tasa a_j para cada cola *j*, y si hay c_j servidores idénticos con tiempos de servicios exponencialmente distribuídos con media $1/\mu_j$, entonces, en estado-estable, la cola *j* se comporta como una cola $M/M/c_j$ con tasa de arribo:

$$\lambda_j = a_j + \sum_{\text{all } i} \lambda_i p_{ij}$$

Redes de Colas

- Ejemplo del Supermercado:
  - Clientes arriban con tasa de *80* por hora y el *40%* elige autoservirse:
    - La tasa de arribo al centro de servicio 1 is $\lambda_1 = 80(0.4) = 32$ por hora
    - La tasa de arribo al centro de servicio 2 is $\lambda_2 = 80(0.6) = 48$ por hora.
  - $c_2 = 3$ empleados y $\mu_2 = 20$ clientes por hora.
  - La utilización a largo plazo de los empleados es:

    $$\rho_2 = 48/(3*20) = 0.8$$

  - Todos los clientes van a abonar al centro de servicio 3. La tasa de arribo allí es:
    - $\lambda_3 = \lambda_1 + \lambda_2 = 80$ por hora.
    - Si $\mu_3 = 90$ por hora, entonces la utilización del cajero es:

      $$\rho_3 = 80/90 = 0.89$$

Modelación Preliminar

- Es muy util en muchos casos:
  - En algunos casos, los resultados de un analisis preliminar son tan convincentes que no habría una gran necesidad de una simulación mas detallada.
  - Provee al analista un mejor entendimiento del sistema a modelar.
  - Las medidas de perfomance obtenidas son utilizadas para *verificar* las salidas de una simulación.

  - Ejemplo: *Area de Otorgamiento de Licencias para conducir*.

Resumen

- Se introdujeron los conceptos básicos de Modelos de Colas.
- Se observó como la simulación y los análisis matemáticos son complementarios.
- Las medidas de perfomance mas usadas: L, L_Q, w, w_Q, ρ, y λ_e.
- Al simular cualquier sistema que evoluciona con el tiempo, el analista debe decidir si se debe estudiar el comportamiento transitorio o comportamiento en estado estacionario.
  - Se vieron algunas fórmulas para estimar los parámetros de comportamiento de estado-estable de los sistemas de colas mas comunes.
- Simples modelos pueden ser resueltos matematicamente, y podrían ser usados como estimaciones preliminares de una medida de perfomance.

Referencias

- COOPER, R.B.[1990], *Introduction to Queueing Theory*. 3d ed., George Washington University.
- GROSS, D., AND C. HARRIS [1997], *Fundamentals of Queueing Theory, 3d ed, Wiley, new York.*
- NELSON, B.L. [1995], *Stochastic Modeling: Analysis & Simulation*, Dover Publications, Mineola, NY.
- **Jerry Banks, John S. Carson, II, Barry L. Nelson, *Discrete-Event System Simulation*, David M. Nicol. Quinta Edición. ISBN-10: 0136062121. Publisher: Prentice Hall.**