Capítulo 5

Teoría de Colas

5.1. Estructura de los Sistemas de Colas

En un sistema de colas se distinguen básicamente los siguientes componentes:

- Población
- Patrón de Llegada
- Largo de la Cola
- Discilpina de la Cola
- Patrón de Servicio
- Salida

Las estructuras de clases más comunes son las siguientes:

1. Un Canal / Una Fase

Figura 5.1: Un Canal / Una Fase

2. Un Canal / Múltiples Fases

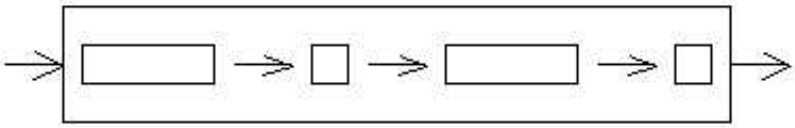

Figura 5.2: Un Canal / Múltiples Fases

3. Multicanal / Una Fase

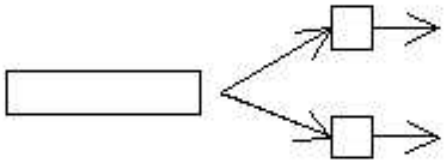

Figura 5.3: Multicanal / Una Fase

4. Multicanal / Múltiples Fases

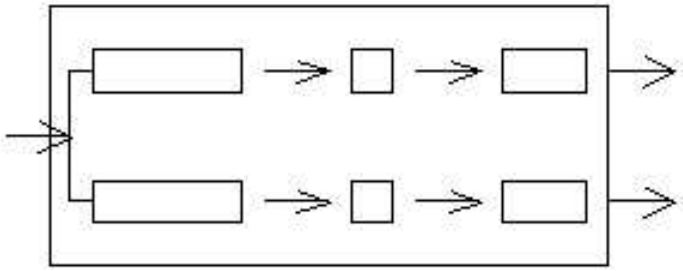

Figura 5.4: Multicanal / Múltiples Fases

5.2. Clasificación de los Sistemas de Colas

La nomenclatura de un sistema de colas tiene la siguiente forma:

(Tasa de Llegada / Tasa de Servicio / Número de Servidores / Tamaño de la Población / Largo de la Cola)

Por ejemplo, el caso más común es el siguiente:

$(M/M/1/I/I)$

que indica una Tasa de Llegada Poisson, una Tasa de Servicio exponencial, 1 Servidor y Tamaño de la Población y Largo de Cola infinitos.

Abreviaciones con supuesto FIFO:

M	Llegada Poisson, Servicio Exponencial
D	Determinística constante, llegada o servicio
K	Distribución *Erlang* del tiempo entre llegadas o servicio
GI	Cualquier distribución con tiempos entre llegadas independientes
G	Cualquier distribución del tiempo de servicio
S	Cualquier número de servidores
I	Infinito largo cola / población
F	Finito largo cola / población

Los casos utilizados más comúnmente son:

1. $(M/M/1/I/I)$
2. $(M/M/S/I/I)$
3. $(M/M/1/I/F)$
4. $(M/M/S/F/I)$

5.3. Proceso de Entrada

Supuestos:

- Proceso de Nacimiento Puro: Los clientes llegan y no abandonan
- Si las llegadas siguen un proceso de Poisson, esto implica que:

1. La probabilidad de una ocurrencia entre t, y $t+h$ sólo depende del ancho del intervalo h.
2. Con h muy pequeño a lo más puede ocurrir una llegada en el intervalo $(t, t+h)$.

- Disciplina FIFO
- Número de Ocurrencias se distribuye Poisson parámetro λt, con λ igual a la tasa de llegada por unidad de tiempo.
- Tiempo entre llegadas se distribuye exponencial con parámetro λ.
- Tiempo hasta la n-ésima llegada se distribuye Γ con parámetros $(n, \frac{1}{\lambda})$.

5.3.1. Llegadas

La teoría de colas se sustenta en el supuesto de que las llegadas de clientes al sistema siguen un *Proceso Poisson*, esto significa que:

Si X_t = número de llegadas en un intervalo $t, X_t \sim \text{Poisson}(\lambda t)$, luego se trata de una variable aleatoria discreta con la siguiente probabilidad:

$$P(X_t = x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t}$$

luego,

$$E[X_t] = \lambda t$$

que representa el número promedio de ocurrencias en el intervalo t y $\lambda = \frac{E[X_t]}{t}$ es el número promedio de ocurrencias por unidad de tiempo.

5.3.2. Tiempo entre llegadas

Considerando que las llegadas siguen un proceso de Poisson, vamos a determinar cuál es la distribución del tiempo entre llegadas.

Sea T_1 el instante de tiempo de la primera llegada, y sea t un instante de tiempo menor. T_1 se puede interpretar como el tiempo necesario para que suceda una ocurrencua, luego

$$P(T_1 > t) = P(\text{Número de ocurrencias hasta } t \text{ sea } 0)$$

$$P(T_1 > t) = P(X_t = 0) = e^{-\lambda t}$$

luego,

$$P(T_1 \leq t) = 1 - e^{-\lambda t}$$

La que corresponde a la función acumulada de una distribución exponencial, luego T_1 se distribuye $exp(\lambda)$
Analizando ahora el tiempo necesario hasta la segunda ocurrencia. Sea Y_t el

número de ocurrencias en el intervalo $[T_1, T_2]$, con T un instante de tiempo dentro de ese intervalo. Luego,

$P(T > t) = P(\text{no hay ocurrencias entre } T_1 \text{ y } T_1 + t, \text{ es decir, durante el intervalo } t)$

$$p(T > t) = P(X_t = 0) = e^{-\lambda t}$$

Así $T \sim exp(\lambda)$.

5.3.3. Tiempo Acumulado

Idea: Si el Tiempo hasta la n-ésima ocurrencia es mayor que un tiempo t dado, implica que hasta t han ocurrido sólo $n-1$ ocurrencias.

$$P(T_n > t) = P(X_t \leq n-1)$$

$$P(T_n > t) = \sum_{i=0}^{n-1} \frac{(\lambda t)^i e^{\lambda t}}{i!}$$

Aproximando la sumatoria por la integral se obtiene que $T_n \sim \Gamma(n, \lambda^{-1})$. Recordemos que una distribución exponencial con parámetro λ es equivalente a una distribución Γ con parámetros $(1, \lambda^{-1})$

5.4. Proceso de Salida

Supuestos:

- Fallecimiento Puro: No pueden reingresar al sistema
- Tasa de salida = μ = Número de clientes servidos por unidad de tiempo

5.4.1. Tiempo de Servicio

Si Z es el tiempo de servicio, $Z \sim exp(\mu)$

5.4.2. Número de Unidades Servidas durante el tiempo t

Sea Y_t el número de unidades servidas durante t, luego $Y_t \sim Poisson(\mu t)$.

5.5. Estado Estacionario

En estado estacionario las medidas de interés de desempeño del sistema a calcular son:

- Utilización del servidor
- Probabilidad de cero clientes en el sistema

- Probabilidad de n clientes en el sistema
- Número promedio en el sistema
- Número promedio en la cola
- Tiempo promedio de espera en la cola
- Tiempo promedio en el sistema
- Probabilidad de que una unidad tenga que esperar por servicio

5.6. Sistemas con una sola Cola, Población Infinita: Estadísticas en Estado Estacionario

Restricción: $\rho \leq 1$

- $\rho = \frac{\lambda}{\mu}$
- $P(0) = 1 - \rho$
- $P(n) = \rho^n \cdot (1 - \rho)$
- $L_q = \frac{\rho^2}{1-\rho}$
- $L_s = \frac{\rho}{1-\rho} = L_q + \rho$
- $W_q = \frac{L_q}{\lambda}$
- $W_s = \frac{1}{\mu - \lambda}$

5.7. Sistemas con Múltiples Servidores

Restricción: $\mu_i = \mu, \forall i$

- $\rho = \frac{\lambda}{k \cdot \mu}$
- $\frac{1}{P(0)} = \sum_{n=0}^{k-2} \frac{(p \cdot k)^n}{n!} + \frac{(p \cdot k)^{k-1}}{(k-1!) \cdot (1-\rho)}$
- $$P(n) = \begin{cases} (\rho \cdot k)^{\frac{n}{n!}} \cdot P(0) & , \text{si } n \leq k \\ \rho^n \cdot k^{\frac{k}{k!}} \cdot P(0) & , \text{si } n \geq k, \end{cases}$$
- $L_q = \frac{\rho^{k+1} \cdot k^{k-1}}{(k-1)! \cdot (1-\rho)^2} \cdot P(0)$
- $L_s = L_q + \rho \cdot k$
- $W_q = \frac{L_q}{\lambda}$
- $W_s = W_q + \frac{1}{\mu} = \frac{L_s}{\lambda}$

5.8. Colas con Prioridades

Idea:

- Entre clases existen prioridades
- Dentro de cada clase se rige mediante FIFO

Los supuestos son los mismos que en el modelo anterior excepto que las unidades que llegan no son tratadas como iguales. Se dividirán las unidades en m clases de acuerdo a una regla de prioridad. Cualquier unidad que llegue y que pertenece a una clase con mayor prioridad precederá a la clase de menor prioridad. Las unidades que forman cada clase se rigen por orden de llegada (FIFO).
Un ejemplo de este tipo de colas se presenta en aquellos programas procesados por facilidades de procesamiento de datos o la clasificación de correspondencia en el correo.

5.8.1. Fórmulas Matemáticas

Sea

- m = número de clases con prioridad.
- λ_i, μ_i, σ_i, ρ_i, $L_q(i)$, $W_q(i)$, $L_s(i)$ y $W_s(i)$ son las características de la cola para las clases $i = 1 \dots m$.
- λ = número total de unidades que llegan por hora, día o semana.
- W_q = tiempo esperado de espera de un cliente *típico*.

$$S_j = \sum_{i=1}^{j} \rho_i, \text{ con} S_0 = 0$$

- W_s = tiempo esperado que permanece un cliente *típico* en el sistema.

Por lo tanto, en estado estacionario, las clases i tienen las siguientes prioridades:

1. $\rho_i = \frac{\lambda_i}{\mu_i}$
2. $$W_q(i) = \frac{\sum_{j=1}^{m} (\rho_j^2) \cdot (1 + \mu_j^2 \cdot \sigma_j^2)/\lambda_j}{2 \cdot (1 - S_{i-1}) \cdot (1 - S_i)}$$
3. $W_s(i) = W_q(i) + \frac{1}{\mu_i}$
4. $L_q(i) = \lambda_i \cdot W_q(i)$
5. $L_s(i) = \lambda_i \cdot W_s(i)$

Las características del sistema completo se obtienen a partir de las clases individuales:

1. $\lambda = \sum_{i=1}^{m} \lambda_i$
2. $L_q = \sum_{i=1}^{m} L_q(i)$
3. $L_s = \sum_{i=1}^{m} L_s(i)$
4. $W_q = \frac{L_q}{\lambda}$
5. $W_s = \frac{L_s}{\lambda}$

Ejemplo

- **Problema:**
  Suponga que para tomar un tren se venden boletos de dos clases. Se ha observado que para la primera clase $\lambda_1 = 20$ y $\lambda_2 = 60$ para los pasajeros de segunda clase cada hora.
  La ventanilla opera con $\mu_1 = 60$ para primera clase y $\mu_2 = 120$ para la segunda clase cada hora. En otras palabras, el tiempo promedio de servicio es $\frac{1}{60}$ [horas], es decir, 1 [min] para procesar a un pasajero de primera clase y sólo $\frac{1}{2}$ [min] para los pasajeros de segunda clase (esto es porque los pasajeros de segunda clase pagan con sencillo y generalmente el valor justo del pasaje).

  Se tiene además que:
  $\sigma_1 = 0{,}85[min] = 0{,}141[horas]$
  $\sigma_2 = 0{,}38[min] = 0{,}0063[horas]$

- **Solución:**

  Veamos las caracteríticas de operación de la ventanilla:
  $\rho_1 = \frac{20}{60} = 0{,}33$
  $S_1 = \rho_1 = 0{,}33$

$$\frac{(\rho_1)^2 \cdot (1 + (\mu_1^2) \cdot (\sigma_1^2))}{\lambda_1} = 0{,}0959$$

$$\begin{aligned} W_q(1) &= \frac{0{,}0956 + 0{,}0657}{2 \cdot (1 - 0{,}33)} \\ &= 0{,}121[horas] \\ &= 0{,}73[min] \end{aligned}$$

  $W_s(1) = 0{,}0121 + \frac{1}{60} = 0{,}0288[horas] = 1{,}73[min]$
  $L_q(1) = 20 \cdot (0{,}0121) = 0{,}24[pasajeros]$

$L_s(1) = 20 \cdot (0{,}0288) = 0{,}58[pasajeros]$

$\rho_2 = \frac{60}{120} = 0{,}5$
$S_2 = \rho_1 + \rho_2 = 0.833$

$$\frac{(\rho_2)^2 * (1 + (\mu_2^2) * (\sigma_2^2))}{\lambda_2} = 0{,}0657$$

$$\begin{aligned} W_q(2) &= \frac{0{,}00956}{2 \cdot (0{,}6667) \cdot (0{,}1667)} \\ &= 0{,}0725[horas] \\ &= 4{,}35[min] \end{aligned}$$

$W_s(2) = 0{,}0725 + \frac{1}{120} = 0{,}0808[horas] = 4{,}85[min]$
$L_q(2) = 60 \cdot (0{,}0725) = 4{,}35[pasajeros]$
$L_s(2) = 60 \cdot (0{,}0808) = 4{,}85[pasajeros]$

Por lo tanto,

$$\begin{aligned} L_q &= 0{,}24 + 4{,}35 \\ &= 4{,}59 \qquad \text{(pasajeros en espera)} \\ L_s &= 5{,}43 \qquad \text{(pasajeros en el sistema)} \\ \lambda &= 20 + 60 \\ &= 80 \qquad \text{(pasajeros llegan por hora)} \end{aligned}$$

Un pasajero típico espera

$$\begin{aligned} W_q &= \frac{L_q}{\lambda} = \frac{4{,}59}{80} \\ &= 0{,}0547[horas] \\ &= 3{,}44[min] \\ W_s &= \frac{L_s}{\lambda} = \frac{5{,}43}{80} \\ &= 0{,}0679[horas] \\ &= 4{,}07[min] \end{aligned}$$

5.9. Colas con Restricciones

Vamos a analizar dos modelos que utilizan los siguientes supuestos:

1. El proceso de llegada es Poisson
2. Tiempos de Servicios son Exponencial
3. Disciplina FIFO

El primer modelo asume que el largo de la cola es ∞.

5.9.1. Sistemas Poisson Exponencial con pocas llamadas (llegadas)

En algunos casos los clientes que llegan al sistema son pocos, con lo cual el sistema básico Poisson-exponencial no se puede aplicar. Especificamente, la probabilidad de que una unidad necesitará servicio no es constante, ésta ahora depende del número de unidades en el sistema. En otras palabras se cuenta con una población finita.

Ejemplo

- **Problema:Comportamiento de los pacientes de un Hospital.**
- **Solución:**
  Sea
  M = número de clientes en la población
  λ = tasa media de llegada de cada unidad individual
  k = número de canales.
  Luego

$$\frac{1}{P(0)} = \sum_{n=0}^{k-1} \left(\begin{array}{c} M \\ n \end{array} \right) \cdot R^n + \frac{M!}{k!} \cdot \sum_{n=k}^{M} \frac{R^n}{(M-n)! \cdot k^{n-k}}$$

  , donde $R = \frac{\lambda}{\mu}$, y $\left(\begin{array}{c} M \\ n \end{array} \right)$ corresponde al número de combinaciones posibles en que se pueden tener n clientes provenientes de un población de tamaño M.

  Como $P(0)$ es conocido

$$P(n) = \begin{cases} \left(\begin{array}{c} M \\ n \end{array} \right) \cdot R^n \cdot P(0) & \text{, cuando } 0 \leq n \leq k \\ \frac{M! \cdot R^n \cdot P(0)}{(M-n)! \cdot k! \cdot k^{n-k}} & \text{, cuando } k \leq n \leq M, \end{cases}$$

$$L_s = \sum_{n=1}^{M} n \cdot P(n)$$

Las unidades L_s que están en el sistema en ese momento no están en la población.
La tasa efectiva de llegada es $\lambda_e = \lambda \cdot (M - L_s)$.
Además

- $W_s = \frac{L_s}{\lambda_e}$
- $W_q = W_s - \frac{1}{\mu}$
- $L_q = \lambda_e \cdot W_q$

5.9.2. Propiedades de un sistema con un sólo canal

$k = 1$

1. $$\frac{1}{P(0)} = M! \cdot \sum_{n=0} M \frac{R^n}{(M-n)!}$$
2. $P(n) = \frac{M!}{(M-n)!} \cdot R^n \cdot P(0)$
3. $L_q = M - \frac{\lambda+\mu}{\lambda} \cdot (1 - P(0))$
4. $L_s = L_q + 1 - P(0)$
5. $\lambda_e = \lambda \cdot (M - L_s) = \mu \cdot (1 - P(0))$
6. $W_q = \frac{L_q}{\lambda_e}$
7. $W_s = \frac{L_s}{\lambda_e}$

- **Problema:**

Pedro utiliza en su planta maquinaria anticuada. Cuando las 5 máquinas operan correctamente los resultados son únicos y hermosos. Sin embargo, él sabe que en promedio cada máquina necesita ser reacondicionada por cada hora de su operación, y que el procso de ajuste requiere en promedio 20[min]. Quizás él necesite un ayudante para repararlas ¿Cúales con las características de operación para el grupo de 2 personas?

- **Solución:**

$M = 5$ máquinas
$k = 2$ reparadores
$\lambda = 1$ máquinas paradas por hora de operación
$\mu = 3$ reparaciones cada hora $\Rightarrow R = \frac{1}{3}$

$$\begin{aligned} \frac{1}{P(0)} &= 1 + \left(\begin{array}{c} 5 \\ 1 \end{array} \right) \cdot \frac{1}{3} + \frac{5!}{2!} \cdot \left\{ \frac{(\frac{1}{3})^2}{3! \cdot 2^0} + \frac{(\frac{1}{3})^3}{2! \cdot 2^1} + \frac{(\frac{1}{3})^4}{1! \cdot 2^2} + \frac{(\frac{1}{3})^5}{0! \cdot 2^3} \right\} \\ &= 4{,}5489 \end{aligned}$$

$P(0) = 0{,}2198$
$P(1) = \begin{pmatrix} 5 \\ 1 \end{pmatrix} \cdot \frac{1}{3} \cdot 0{,}2198 = 0{,}3664$
$P(2) = \begin{pmatrix} 5 \\ 1 \end{pmatrix} \cdot (\frac{1}{3})^2 \cdot 0{,}2198 = 0{,}2443$
$P(3) = 0{,}1212$
$P(4) = 0{,}0407$
$P(5) = 0{,}0068$

$L_s = 1 \cdot 0{,}3664 + 2 \cdot 0{,}2443 + 3 \cdot 0{,}1212 + 4 \cdot 0{,}407 + 5 \cdot 0{,}0068 = 1{,}42$

$\lambda_e = \frac{1}{5-1{,}42} = 3{,}58$
$W_s = \frac{1{,}42}{3{,}58} = 0{,}40$
$W_q = 0{,}40 - \frac{1}{3} = 0{,}06$
$L_q = 0{,}06 * 3{,}58 = 0{,}21$

Como $P(0) = 0{,}2198$ implica que existe un 21.98 % de probabilidad que Pedro y su ayudante esten sin trabajo que reparar. En una hora típica 1,42 máquinas están en el taller de reparación y cada máquina permanece en promedio 1.4 [horas] $\cong$ 24 [min].
Cuando una máquina necesita un ajuste, ésta debe esperar (antes de ser atendida) 0.06 [horas] = 3.6 [min]. La línea de espera tiene un promedio de 0.21 máquinas.

5.9.3. Sistemas Poisson Exponencial con un sólo canal con cola truncada

Pueden existir 2 razones para limitar el largo de la cola.

1. La cola se limita sola, llega un momento en que ninguna persona desea ponerse a una cola con un largo excesivo.
2. Los sistemas de servicio limitados fisicamente, por ejemplo, la sala de espera en un centro médico.

Propiedades en Estado Estacionario
Sea
M = número máximo de unidades llegando al sistema, por lo tanto, el largo máximo de la cola será $M - 1$
$R = \frac{\lambda}{\mu}$

1. $P(0) = \frac{1-R}{1-R^{M+1}}$
2. $P(n) = R^n \cdot P(0)$

3. $L_s = \frac{R}{1-R} - \frac{(M+1)\cdot(R^{M+1})}{1-R^{M+1}}$
4. $L_q = L_s + P(0) - 1$
5. $\lambda_e = \lambda \cdot (1 - P(M))$
6. $W_q = \frac{L_q}{\lambda_e}$
7. $W_s = W_q + \frac{1}{\mu}$

donde λ_e corresponde a la tasa de llegada efectiva.
Ya que a lo más pueden haber M unidades en el sistema, $P(M)$ es la probabilidad que el sistema esté lleno, es decir, una unidad llegando en ese estado no podrá ingresar al sistema.
Por lo tanto, $1 - P(M)$ = probabilidad de que una unidad pueda entrar al sistema.
Y como los clientes varían entre 0 y M, entonces

$$\sum_{n=0}^{M} P(n) = 1$$

$$\Rightarrow P(0) = \frac{1-R}{1-R^{M+1}}$$

En el caso en que $\lambda = \mu$ todos los estados son igualmente probables, en este caso:

$$P(0) = \frac{1}{M+1}$$

$$L_s = \frac{M}{2}$$

Ahora si λ excede a μ se tendrá que el sistema llegará a saturarse con $L_s \cong M$ y $P(M) \cong 1$.

5.10. Ejercicios Propuestos

1. El Problema de Rafael

   - I Parte Los clientes llegan al negocio de Rafael según una distribución Poisson. El almacén abre a las 8h00 y los Martes en la mañana entre las 8h00 y las 9h00 llegan en promedio 6 clientes al almacén. Rafael se juntó a ver un partido con unos amigos el día Lunes en la noche, por eso le gustaría dormir una media hora más el Martes en la mañana. El sabe que si abre muy tarde puede perder muchos clientes por dejar de ganar por la venta perdida. Para tomar una buen decisión quiere ver una forma de estimar el número de clientes que llegarán entre las 8h00 y las 8h30 el martes y así estimar la eventual pérdida.

- II Parte Rafael estimó que se demora 4 minutos en servir a un cliente en su negocio y los tiempos de servicio siguen una distribución exponencial. Rafael tiene una cita importante asi es que desea encontrar la probabilidad que se tomará menos de tres minutos en servir al siguiente cliente.

2. Problema de una Zapatería El dueño de una zapatería recibió un reclamo de un cliente por el servicio recibido. Se sabe que los clientes llegan en promedio uno cada 12 minutos según un proceso de Poisson. El vendedor estima que puede servir un cliente en un promedio de 8 minutos y que su tiempo de servicio sigue una distribución exponencial. El vendedor siempre termina de atender un cliente antes del siguiente. El dueño desea determinar las medidas de desempeño del sistema actual, para evaluar la veracidad de lo que dice su cliente.

3. Consultas al Atudante Un curso de investigación de mercado tiene una sesión de laboratorio donde los estudiantes diseñan cuestionarios. Los alumnos tienen un ayudante para sus dudas. Un estudiante con una pregunta *volverá después* si hay 3 antes en la cola de espera. Hay un promedio de cuatro estudiantes por hora que acuden a hacer alguna pregunta en un promedio de 12 [min]. Suponga que es un número ilimitado de estudiantes.

   a) ¿Cuátos estudiantes en promedio estará en la cola de espera?

   b) ¿Cuál es la tasa efectiva de llegada?

   c) ¿Cuánto tiempo está con la duda un típico estudiante?