III.5 Teoria de Colas

May 13, 2019

1 Teoría de Colas

Una **cola o línea de espera** está formada por peticiones de servicios presentadas por clientes a una estación o servidor. Si el servicio solicitado no está inmediatamente disponible, la petición de servicio entra en cola, abandonando el sistema una vez atendido o antes por diversas razones (no espera, cola de tamaño finito, etc). Mientras una petición está esperando, pueden llegar otras. Un sistema de colas se caracteriza por:

- Proceso de entrada
- Mecanismo de servicio
- Disciplina de la cola
- Capacidad del servicio

Mm1_queue.png

1.0.1 Proceso de entrada

Se refiere a la caracterización probabilística del proceso estocástico que modela la llegada de las peticiones de servicio a lo largo del tiempo. Se caracteriza con:

- Secuencia de **instantes de llegada**: $t_1, t_2 \dots t_n \quad t_1 < t_2 < \dots < t_n$
- Secuencia de **tiempos entre llegadas**: $\tau_1 = \tau_2 - \tau_1, \tau_2 = \tau_3 - \tau_2, \dots \tau_n = t_{n+1} - t_n$

Es habitual que los instantes de llegada se modelen con un **Proceso Puntual de Poisson**, de modo que el **número de llegadas** $N(t)$ hasta el instante t, considerando $N(0) = 0$, es un proceso estocástico que en cada instante t tiene una función de masa de probabilidad de Poisson (recordar lo visto en el Bloque II), con tasa de llegada λ peticiones por segundo:

$$P(N(t) = n) \equiv p_N(n; t) = e^{-(\lambda t)} \frac{(\lambda t)^n}{n!}$$

El valor medio y la varianza del proceso en cada instante t son:

$$E(N(t)) = Var(N(t)) = \lambda t$$

Por definición del proceso puntual de Poisson, en el que las llegadas son independientes, **la secuencia de tiempos entre llegadas también es independiente**, por lo que la caracterización de un tiempo marginal de llegadas permite obtener la conjunta multiplicando las marginales, todas idénticas. Por otro lado, como se vio en el Bloque II, cuando los instantes de llegadas se modelan con un Proceso Puntual de Poisson con tasa λ, el tiempo transcurrido entre dos llegadas sucesivas tiene una función de densidad exponencial con idéntica tasa:

$$f_\tau(\tau) = \lambda e^{-\lambda\tau} u(\tau)$$

Su media es $E(\tau) = \frac{1}{\lambda}$ y varianza $Var(\tau) = \frac{1}{\lambda^2}$. Recordemos que la distribución exponencial **no tiene memoria**, esto es, $f_\tau(\tau + \tau_0 | \tau > \tau_0) = f_\tau(\tau)$. Dicho de otra manera, si tras τ_0 segundos nos tenemos una llegada, la función de densidad de probabilidad se mantiene invariante.

Es fácil entender que, en estas condiciones, el número de llegadas acumuladas hasta un instante t es una **cadena de Markov**, pues sólo depende de las acumuladas hasta el instante de la llegada anterior y no de toda la historia pasada.

La caracterización probabilística del proceso de entrada puede **generalizarse**, por ejemplo, manteniendo la suposición de que los tiempos entre llegadas son independientes, pero que siguen una distribución distinta de la exponencial. En tal caso se pierde la propiedad *sin memoria* y el número de llegadas acumuladas ya no puede modelarse con una cadena de Markov.

1.0.2 Mecanismo de Servicio

Se refiere a la caracterización probabilística del proceso estocástico que modela el tiempo que el servidor tarda en atender cada petición peticiones. Se caracteriza con la **secuencia de tiempos de servcio**: $s_1, s_2, \ldots s_n$. Es lógico pensar que **la secuencia de tiempos de servicio es independiente**. Los modelos más sencillos y habituales son:

- Duración del tiempo de serticio constante: $s_n T_s$
- Exponencial con parámetro μ. Como sabemos, esta distribución no tiene memoria y, en este caso, el número de peticiones atendidas hasta el instante t se modela como una cadena de Markov.
- istribución Erlang, que generaliza la exponencial

Notación de los procesos de entrada y de servicio

- M: Poisson o exponencial)Markov o sin memoria)
- D: Determinista
- En: Erlang
- G: distribución arbitraria
- GI: distribución arbitraria, pero con secuencia de tiempos de espera o servicio independiente

1.0.3 Disciplina de la cola

Especifica la regla que gestiona la cola, la permanencia o salida de la misma y la preferencia para atender las peticiones.

- Permanencia en cola: **wait** (espera hasta ser servido), **balking** (negarse a hacer cola), **renage** (hacer cola solo un tiempo e irse si no es atendido), **jockey** (saltar entre colas)
- Orden en atención: La más habitual es **FIFO** (First In First Out), siendo otras **LIFO** (Last In First Out), **Processor Sharing** (compartición del sevidor por igual), **Priority**, **Shortest Job First**, etc

1.0.4 Capacidad del servicio

- Número de canales (servidores) que proporcionan servicio con idénticas o diferentes tasas
- Capacidad de la cola, que puede ser infinita o finita, en este caso con una longitud máxima que si se supera provoca la pérdida de las nuevas llegadas de peticiones

La **intensidad de tráfico**, ρ, es el ratio entre la tasa media de llegadas de peticiones, λ y de peticiones servidas μ: $\rho = \frac{\lambda}{\mu}$

1.0.5 Caracterización de la cola

Dadoslos paráteros anteriores, definitorios del sistema de colas, la caracterización procura obtener:

- **Número de peticiones en el sistema**, considerando tanto en espera en la cola como siendo atendidas por los servidores
- **Tiempos de espera de las peticiones**, tanto en la cola, como en el servidor, como total

La caracterización de sistemas de colas Markovianos hace uso de los denominados **procesos de nacimiento y muerte**. La llegada (sin salida del sistema) es un proceso puro de nacimiento, mientras que la salida (sin nuevas entradas) lo es de muerte. Cada uno tiene sus propias tasas que, además, en general pueden variar a lo largo del tiempo. Ambos procesos pueden equilibrarse, dando lugar a un funcionamiento estacionario del sistema, que es el que tien mayor interés práctico.

El **Teorema de Little** permite relacionar los parámetros principales en el estado estacionario. Sean λ la tasa media de llegada de peticiones, $L = E(N(t))$ el número medio de peticiones en el sistema (tanto en cola como en los servidores) y $W = W_q + W_s$ el tiempo medio de las peticiones en el sistema, sumando en cola (q) y en el servidor (s). Se cumple: $L = \lambda W$

Notación de Kendall Un sistema de colas se especifica mediante tres parámetros, $\alpha/\beta/N_1$ al que puede añadirse un cuarto N_2:

- α: Modelo probabilístico de las llegadas, conforme a la notación vista antes
- β: Modelos probabilístico de los tiempos de servicio, conforme a la misma notación
- N_1: Número de canales o servidores
- N_2: Si la cola no tiene capacidad infinita, capacidad máxima de la misma

Por ejemplo un sistema $M/M/1$ se refiere al caso más sencillo, de llegdas y servicios markovianos con un único servidor. Sin embargo, $M/M/s$ denotaría s servidores mientras que, por ejemplo, $M/M/1/q$ se refiere al sistema $M/M/1$ anterior, pero limitando la cola a q peticiones. Pueden plantearse otros modelos, por ejemplo, $GI/En/r/q$ es un sistema con una distribución genérica en los tiempos entre llegadas (aunque independientes), con una distribución Erlang en los tiempos de serivicio, con r servidores y una cola con espacio para q peticiones en espera.