Simulación de un sistema MM1

Delmonti Agustín
Universidad Tecnológica Nacional Rosario
Zeballos 1341, S2000
agus.delmonti@gmail.com

Trilla Melody
Universidad Tecnológica Nacional Rosario
Zeballos 1341, S2000
trillamelo@gmail.com

Junio 2020

Resumen

El siguiente trabajo se realiza un análisis de un modelo de cola simple M/M/1. Se desarrolla un modelo matemático analítico y otro modelo basado en la simulación por eventos.

1. Introducción

Una **línea de espera** o **cola** es un proceso en el que las personas, los materiales o la información deben esperar en un momento determinado para **obtener un servicio** debido a que casi siempre, la capacidad de servicio es menor que la capacidad demandada.[1] Las colas ocurren en todas partes: los clientes esperan para ser atendidos en supermercados o bancos, o para comprar entradas en teatros y cines; las personas forman largas filas de autos en la congestión del tráfico; los usuarios en las redes de información o servidores web esperan para obtener un servicio, e inclusive los sistemas operativos que tratan de administrar las ejecuciones de las tareas que van apareciendo de la forma más eficiente.

Las colas afectan la productividad, la calidad del servicio (como el servicio es percibido por el cliente), consumen recursos valiosos y tiempo. Por lo tanto, el estudio de colas resulta muy importante y práctico.

Se cuenta con un conjunto de modelos matemáticos que se enmarcan en ***la teoría de colas***, una disciplina dentro de la teoría matemática de la probabilidad que provee modelos con resolución analítica para problemas simples de colas de espera. Sin embargo la mayoría de los sistemas complejos del mundo real con elementos estocásticos no pueden ser descritos precisamente por un modelo matemático que pueda ser evaluado analíticamente. Así, una simulación es comúnmente el único tipo de investigación posible. Por tal razón, en el presente trabajo se desarrolla un modelo simple de colas el cual es resuelto con ambas metodologías descritas.

2. Marco teórico

2.1. Teoría de colas

La teoría de colas o el modelo de colas es un conjunto de conocimientos que se ocupa de la línea de espera que intenta estimar el comportamiento de colas en función de ciertos supuestos.

En la forma más simple, el modelo de colas supone que los clientes arriban al sistema según una distribución de llegada y existen un número finito de servidores que sirven a los clientes según una distribución de tiempo de servicio. Cuando un cliente llega y se encuentra con todos los servidores ocupados, se posiciona en una cola a la espera de ser atendido.

El resultado directo de la teoría de colas es la medición de la efectividad o característica operativa que mide el *rendimiento* de un sistema de colas en su ***estado estable***. El objetivo principal es predecir el comportamiento a largo plazo (o en estado estable) y determinar una capacidad de servicio apropiada que

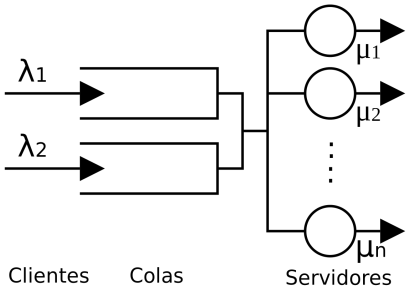


Figura 1: modelo conceptual de un sistema con 2 colas y n servidores.

garantice un equilibrio entre el factor cuantitativo (costos del sistema) y el factor cualitativo (satisfacción del cliente por el servicio). Por ejemplo, puede estimar el tiempo de espera o el retraso, la longitud de la cola, la probabilidad de que el servidor esté inactivo o la capacidad de trabajo del sistema sin que llegue a colapsar. A partir de esas mediciones de efectividad, se puede calcular la cantidad de servidores necesarios o estimar el tiempo de servicio requerido para satisfacer la demanda. También puede estimar el costo de las colas y experimentar con ciertas ideas para mejorar el sistema de colas.

2.1.1. Notación de Kendall

Cualquier modelo de colas se puede caracterizar por una cierta cantidad de parámetros del sistema. En la teoría de colas se utiliza una notación generalizada llamada notación de Kendall para indicar qué tipo de sistema se está modelizando.[2] Esta notación tiene la forma $A/S/c/K/N/D$.
Donde:

- A: arribo de los clientes.
- S: tiempos de servicio.
- c: cantidad de servidores.
- K: capacidad de la cola.
- D: disciplina de la cola.

La capacidad de la cola K denota el número máximo de clientes permitidos en la cola. Cuando el número está en este máximo, se rechazan nuevos arribos al sistema. La disciplina de cola puede ser $FIFO$, $LIFO$, por prioridad, circular o en tandas y determina en que orden los clientes de la cola son servidos. Cuando no se especifican los tres parámetros finales (por ejemplo, cola $M/M/1$), se supone $K = \infty$, $N = \infty$ y $D = FIFO$.

2.2. Estado estable

Cuando un sistema de colas apenas inicia su operación, el estado del sistema (número de clientes en el sistema), se encuentra bastante afectado por el estado inicial y el tiempo que ha pasado desde el inicio. Se dice entonces que el sistema se encuentra en una situación transitoria. Después de que ha pasado un tiempo suficiente, el estado del sistema se vuelve independiente del estado inicial y del tiempo transcurrido, y se puede decir que el sistema ha alcanzado su condición de estado estable. La teoría de colas tiende a dedicar su análisis a la condición de estado estable debido a que el caso transitorio es analíticamente más difícil. [3]

3. Sistema M/M/1

Considerando la notación de Kendall un sistema M/M/1 tiene las siguientes características:

- M: las llegadas que se producen según un proceso de Poisson a razón de λ arribos por unidad de tiempo, donde los tiempos entre llegadas están distribuidos exponencialmente $exp(\lambda^{-1})$.
- M: los tiempos entre servicios son distribuidos de manera exponencial, $exp(\mu^{-1})$ donde μ es el número medio de clientes que el servidor es capaz de atender por unidad de tiempo.
- 1: servidor único en el sistema.

3.1. Medidas de rendimiento.

Una vez en estado estacionario se pueden deducir las medidas de rendimiento del sistema que nos permitan evaluar la eficiencia y los límites del mismo.

Utilización del sistema El factor de utilización del sistema o intensidad de tráfico ρ es la proporción de tiempo en la que se espera encontrar a los servidores ocupados.

$$\rho = \frac{\lambda}{\mu} \tag{1}$$

Si $\rho < 1$ entonces el sistema se estabilizará. Mientras más cerca esté ρ a 1, más cargado estará el sistema, las colas serán más largas y los tiempos de espera serán mayores. Si $\rho > 1$ el número de clientes en el sistema se incrementará sin lmite: la tasa de servicio no satisface el ratio de llegada de los clientes al sistema. Se desprende además que la probabilidad de llegar al sistema y de hallar al servidor desocupado es $p_0 = 1 - \rho$.

Probabilidad de que haya n clientes en el sistema

$$P(X = n) = p_n = \rho^n(1 - \rho) \tag{2}$$

Luego

$$P(X \leq k) = \sum_{n=0}^{k-1} p_n = \sum_{n=0}^{k-1} \rho^n(1 - \rho) = 1 - \rho^k \tag{3}$$

Longitud media de la cola La longitud media de la cola es el valor esperado de clientes en cola

$$L_q = E(N_q) = \sum_{n=0}^{\infty}(n - 1)p_n = \frac{\rho^2}{1 - \rho} \tag{4}$$

Media de clientes en el sistema De la misma manera se puede calcular el valor esperado de clientes en el sistema

$$L = E(N) = \sum_{n=0}^{\infty} np_n = \frac{\rho}{1 - \rho} \tag{5}$$

Utilizando la expresión derivada en la Ecuación 4, se puede expresar la cantidad de clientes promedio en el sistema en función de la longitud promedio de la cola

$$L = L_q + \rho \tag{6}$$

Tiempo medio de un cliente en el sistema El tiempo medio que un cliente permanece en el sistema es el tiempo medio de respuesta W. Si suponemos que un cliente, al llegar al sistema, se encuentra con que hay n clientes antes que él, el tiempo medio que tardará en salir del sistema será $n+1$ veces el tiempo medio de servicio μ. Luego

$$\begin{aligned} W &= \sum_{n=0}^{\infty} \frac{1}{\mu}(n + 1)p_n \\ &= \frac{1}{\mu} \sum_{n=0}^{\infty} np_n + \frac{1}{\mu} \sum_{n=0}^{\infty} p_n \\ &= \frac{L}{\mu} + \frac{1}{\mu} \\ &= \frac{1}{\mu - \lambda} \end{aligned} \tag{7}$$

Tiempo medio de un cliente en la cola El tiempo medio de espera en la cola W_q se hallará restando a W el tiempo que tarda en ser servido el cliente.

$$W_q = W - \frac{1}{\mu} = \frac{\rho}{\mu - \lambda} \tag{8}$$

3.2. Simulaciones

Como es un sistema sencillo, ya vimos que podemos obtener un modelo analítico del sistema para comparar resultados. Sin embargo, para sistemas mucho más complejos puede que la única manera de conseguir resultados sea a través de simulaciones puesto que no se puede hallar solución analítica del modelo (muchas veces a causa de ecuaciones diferenciales complejas que modelizan el sistema dinámico). Se realiza un modelo de un sistema M/M/1 para registrar las medidas de rendimiento introducidas pero de forma empírica. La simulación termina cuando exactamente n clientes son servidos por el servidor, es decir cuando $n-1$ clientes han recibido servicio y el cliente n entra en el servidor. Para estimar los parámetros en el estado estable del sistema se utiliza el teorema central del limite de 30 corridas independientes de $n = 10k$ clientes servidos cada una.

Se puede observar en la Figura 2 la probabilidad de que se encuentren n clientes en la cola con diferentes

	L_q	ρ	W_q
		$\mu = 4$	
Analítico	0.0833	25,00 %	0.0833
Simulación	0.0846	25,03 %	0.0845
		$\mu = 2$	
Analítico	0.5	50,00 %	0.5
Simulación	0.4937	49,80 %	0.4941
		$\mu = 1{,}333$	
Analítico	2.25	75,00 %	2.25
Simulación	2.2544	74,96 %	2.2532
		$\mu = 1$	
Analítico	∞	100,0 %	∞
Simulación	79.5505	98,95 %	77.4634
		$\mu = 0{,}8$	
Analítico	∞	125,0 %	∞
Simulación	1557.776	99,949 %	1000.115

Cuadro 1: tabla comparativa de resultados de las simulaciones para un sistema M/M/1 con $\lambda = 1$ versus el valor teórico. Cada métrica es un promedio de 30 corridas.

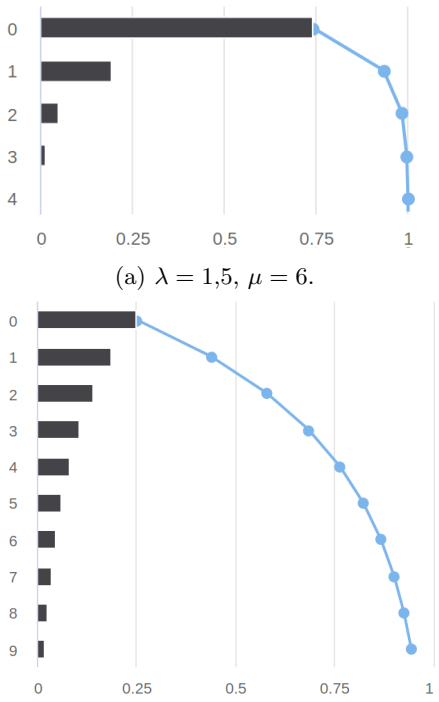
parámetros del sistema. En la Figura 2a se puede ver que un sistema con una utilización ρ del 25 % la cola es de tamaño mucho menor que uno con una utilización ρ del 75 % de la Figura 2b. La probabilidad de que un cliente llegue al sistema y encuentre la cola completamente vacía p_0 también disminuye drásticamente.

4. Sistemas M/M/1/K

Al igual que el modelo M/M/1, en este modelo, los tiempos entre llegadas y de servicios siguen distribuciones exponenciales con tasas λ y μ respectivamente, pero la cola tiene una capacidad máxima de clientes K.

4.1. Denegación de servicio

Como la cola admite a lo sumo K clientes, cuando ésta llega a su capacidad máxima y arriba un cliente nuevo, el sistema lo rechaza. A esto se lo denomina **denegación de servicio**.



(a) $\lambda = 1{,}5$, $\mu = 6$.

(b) $\lambda = 4{,}5$, $\mu = 6$

Figura 2: probabilidad de que haya n clientes en cola en modelo M/M/1.

Si la probabilidad de que haya a lo sumo n clientes en la cola es la probabilidad acumulada de la Ecuación 3, la probabilidad de denegarle el servicio a un cliente $K + 1$ esta dada por

$$P(X \geq K) = 1 - P(X \leq K) = \rho^K \tag{9}$$

En la Figura 3 se pueden observar las probabilidades discretas y acumuladas de que haya n clientes en cola. Si agregamos la restricción de un máximo de 4 clientes en cola, podemos observar en color rosa el margen de rechazo de clientes nuevos. La probabilidad de denegar el servicio es entonces lo acumulado en la sección rosa de la gráfica, es decir $P(X \geq 4) = 1 - F(4) = 0{,}1318$.

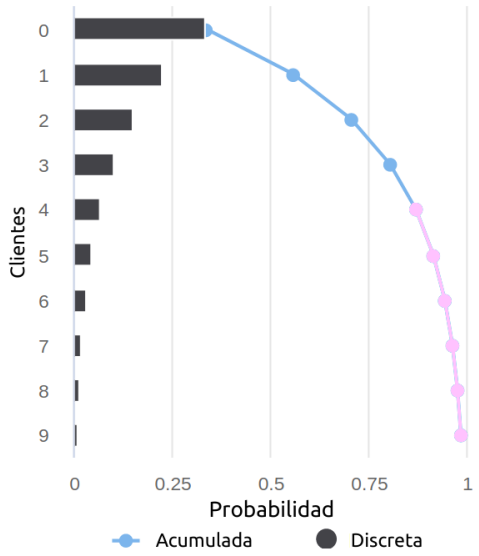


Figura 3: probabilidad de que haya n clientes en cola en modelo M/M/1/K con $\lambda = 4$, $\mu = 6$ y con tamaño de cola máxima $K = 4$.

5. Conclusión

Los sistemas de colas modelizan situaciones cotidianas de suma importancia. En sistemas de colas sencillos dichas relaciones se pueden encontrar analíticamente. En sistemas más complejos se pueden analizar mediante simulación, lo cual es más costoso.

Estos modelos nos permiten analizar casos específicos del sistema, inferir y tomar decisiones, que si fuesen a ser probadas en el modelo real podrían tener un costo elevado. Por ejemplo, si se detectara un incremento en la utilización del sistema o intensidad del tráfico se podrían probar cambios sin implementarlos en la vida real, como por ejemplo cambiar el servidor a uno con una tasa μ más alta, agregando más servidores en paralelo o inclusive implementando otra disciplina de cola. En definitiva, la simulación resulta una herramienta efectiva decisoria cuando no se puede hallar un modelo analítico del problema.

Referencias

[1] Kardi Teknomo. Queuing Theory. `http://people.revoledu.com/kardi/tutorial/Queuing`, 2014. Accessed: 2020-07-06.

[2] Wikipedia contributors. Kendall's notation — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Kendall%27s_notation&oldid=957278545`, 2020. [Online; accessed 6-July-2020].

[3] David de la Fuente García, Raúl Pino Diez. Teoría de líneas de espera - modelos de colas, 2001. [Online; accessed 7-July-2020].