Proyecto 1.
Simulación
Cola con n servidores en paralelo

Karen Dianelis Cantero Lopez C411
Loitzel Ernesto Morales Santiesteban C412

February 25, 2024

Contents

1 Introducción

En este proyecto pretendemos simular el comportamiento de una cola ante un servicio que posee n servidores en paralelo. Cuando un cliente llega, se une a la cola si todos los servidores están ocupados. Si algún servidor está libre, el cliente entrará al primer servidor (del 1 al n) que esté libre. Cuando un cliente completa el servicio con un servidor (no importa cuál), ese cliente luego abandona el sistema. El cliente que ha estado en la cola durante más tiempo (si hay clientes en la cola) entra en servicio.

1.1 Objetivos y Metas

El objetivo principal del proyecto es simular el sistema descrito anteriormente para comprender su comportamiento y desempeño bajo diferentes condiciones y distribuciones de arribos y servicio. Las metas específicas incluyen:

- Analizar el tiempo de espera promedio de los clientes en la cola para evaluar la eficiencia del sistema en la atención de los clientes.
- Determinar el tiempo de espera máximo experimentado por un cliente en la cola para identificar posibles casos extremos de espera.
- Calcular la cantidad total de clientes atendidos durante la simulación para entender la capacidad de procesamiento del sistema.
- Estimar la cantidad de clientes que no se llegan a atender debido a la falta de capacidad del sistema.
- Evaluar la probabilidad de que un cliente no pueda ser atentido inmediatamente debido a la falta de capacidad del sistema (probabilidad de que el sistema esté saturado) para comprender la eficiencia del sistema en la gestión de la demanda y la asignación de recursos.
- Estimar el tiempo promedio que un cliente en el sistema (en la cola y en el servidor)
- Calcular la probabilidad de que el sistema tenga 0 clientes para poder determinar la eficiencia y capacidad del sistema.

1.2 Variables que describen el problema

- Cantidad de servidores: Representa el número de servidores disponibles en el sistema para atender a los clientes.
- Tiempo de servicio: Indica la duración total del servicio.
- Distribución de llegada de clientes: Describe el patrón o la ley que sigue la llegada de nuevos clientes al sistema.
- Distribución de tiempo de servicio: Indica cómo se distribuyen los tiempos de servicio de los servidores al atender a los clientes.

2 Sistema Específico y Variables de Interés

2.1 Sistema Específico

Como ya se ha dicho en la introducción, el sistema que se va a simular es un sistema con n servidores en paralelo. Cada servidor puede seguir una distribución distinta del tiempo de atención. El tiempo de llegada de los clientes también puede seguir una distribución distinta.

Para facilitar el uso de nuestra implementación, se han dejado ya implementadas las distribuciones de poisson, exponencial, normal y uniforme.

2.2 Variables de Interés

Las variables de interés en este problema son:

- Tiempo de espera promedio de los clientes en la cola: Representa el tiempo promedio que un cliente pasa esperando en la cola antes de ser atendido.
- Promedio del tiempo de espera máximo experimentado por un cliente en la cola: Este valor ayuda a identificar casos extremos de espera.
- Promedio de la cantidad total de clientes atendidos durante la simulación: Representa el número total de clientes que han sido atendidos y han completado su servicio en el sistema durante la simulación.
- Promedio de la cantidad de clientes que no se llegan a atender debido a la falta de capacidad del sistema: Es una indicación de la capacidad del sistema.

- Probabilidad de que un cliente no pueda ser atendido inmediatamente debido a la falta de capacidad del sistema: Esta probabilidad es una medida de la eficacia del sistema en la gestión de la demanda y la asignación de recursos.

3 Detalles de Implementación

3.1 Lenguaje de Programación

Para la implementación de la simulación se utilizó el lenguaje de programación Python.

3.2 Pasos seguidos para la implementación

Los pasos seguidos para la implementación de la simulación fueron:

- Modelar distintos tipos de distribuciones.
- Crear las clases StatisticsHolder, StateVariables y Sim para mantener una estructura limpia en el código.
- Implementar el algoritmo de simulación teniendo en cuenta las diferentes distribuciones y almacenando los datos necesarios para el análisis.
- Realizar pruebas de los modelos implementados y comparar con los resultados matemáticos.

4 Resultados y Experimentos

4.1 Hallazgos de la simulación

Para poder comparar como afecta cada parámetro a los resultados de la simulación, simularemos 4 escenarios distintos, en cada uno variando solo una variable.

Para todas las simulaciones mantendremos constante la distribución Poisson como distribución de llegada a la cola (por lo tanto los tiempos entre las llegadas distribuyen de forma exponencial)y la distribucion Poisson como distribución de los tiempos de atención en los servidores ya que estos tiempos también son de forma exponencial.
λ va a representar el parámetro de la distribución de llegada de clientes, μ va a representar el parámetro de la distribución de los tiempos de atencián, y n representará la cantidad de servidores.

4.1.1 Simulación 1

Usaremos:

- $\lambda = 2$
- $\mu = 1$
- $n = 3$

Obtuvimos estos resultados:

variable	**media**	**varianza**	**desviación estándar**
probabilidad de encolarse	0.44	0.001	0.001
probabilidad de 0 clientes en el sistema	0.11	0.0001	0.0003
Longitud promedio de la cola	425	3268	1.8
Tiempo promedio en cola	0.43	0.009	0.003
Tiempo promedio en el sistema	1.43	0.01	0.003

4.1.2 Simulación 2

Ahora probaremos con las mismas variables, solo cambiando $\lambda = 4$

- $\lambda = 4$
- $\mu = 1$
- $n = 3$

variable	**media**	**varianza**	**desviación estándadr**
probabilidad de encolarse	0.99	0.000002	0.000049
probabilidad de 0 clientes en el sistema	0.0009	0.0000007	0.000014
Longitud promedio de la cola	1913	2043	1.42
Tiempo promedio en cola	59.85	54.42	0.34
Tiempo promedio en el sistema	60.59	54.29	0.23

4.1.3 Simulación 3

Ahora probaremos con las mismas variables, solo cambiando $\mu = 4$

- $\lambda = 2$
- $\mu = 4$
- $n = 3$

variable	**media**	**varianza**	**desviación estándar**
probabilidad de encolarse	0.015	0.000031	0.00017
probabilidad de 0 clientes en el sistema	0.60	0.00021	0.00046
Longitud promedio de la cola	14.68	30.40	0.17
Tiempo promedio en cola	0.0015	0.00000084	0.000029
Tiempo promedio en el sistema	0.2513	0.000073	0.00027

4.1.4 Simulación 4

Ahora probaremos con las mismas variables, solo cambiando $n = 8$
Usaremos:

- $\lambda = 2$
- $\mu = 1$
- $n = 8$

variable	**media**	**varianza**	**desviación estándar**
probabilidad de encolarse	0.001	0.0.0000028	0.000043
probabilidad de 0 clientes en el sistema	0.13	0.00027	0.0005
Longitud promedio de la cola	1.098	2.72	0.05
Tiempo promedio en cola	0.0002	0.00000018	0.000013
Tiempo promedio en el sistema	0.999	0.0010	0.001

4.2 Interpretación de los resultados

Tomaremos los resultados de la simulación 1 como base para comparar cómo el cambiar los parámetros afecta los resultados.

4.2.1 Variando el valor λ

Podemos apreciar que al aumentar el valor de λ y mantener todos los demás valores igual :

- La cantidad de personas encoladas aumentan, ya que la cantidad de personas en el sistema aumenta, debido a que lambda simboliza la cantidad de personas que llegan en un tiempo t.
- Los tiempos promedio que una persona permanece en la cola, y por ende en el sistema, aumentan.
- Al aumentar el flujo de personas, la probabilidad de encolarse aumenta, y la probabilidad de que el sistema esté vacío disminuye.

4.2.2 Variando el valor μ

Podemos apreciar que al aumentar el valor de μ y mantener todos los demás valores igual :

- La cantidad de personas encoladas disminuyen, ya que los tiempos de atención de los servidores disminuyen porque los servidores pueden soportar una carga mayor (esto es lo que μ representa en este caso)
- Los tiempos promedio que una persona permanece en la cola, y por ende en el sistema, disminuyen.
- Al disminuir el tiempo de atención, la probabilidad de encolarse disminuye, y la probabilidad de que el sistema esté vacío aumenta.

4.2.3 Variando el valor n

Podemos apreciar que al aumentar el valor de n y mantener todos los demás valores igual :

- La cantidad de personas encoladas disminuyen, ya que ahora existen más servidores disponibles para atender a los clientes.
- La probabilidad de encolarse también disminuye notablemente ya que existen más servidores disponibles.
- El tiempo promedio en la cola disminuye notablemente ya que existen más servidores disponibles.

- El tiempo promedio en el sistema no cambia notablemente con respecto a la Simulación 1 ya que no hemos variado el valor de μ que influye en el el tiempo de atención.

4.2.4 Alta varianza en resultados

Una alta varianza en los resultados nos dice que los valores varían notablemente entre las simulaciones.
Como puede ser observado, en todos los experimentos el valor de la varianza de la longitud promedio de la cola es mayor que la media de este mismo. Esto se debe principalmente a la capacidad del sistema. Si la capacidad del sistema (n) es insuficiente para manejar la demanda, se producirán congestiones y variaciones en la longitud de la cola. Esto es evidente al observar que la simulación con menor varianza es en la que se modifica el valor de n.

4.3 Necesidad de realizar el análisis estadístico a la simulación

El análisis estadístico de los resultados de una simulación es fundamental para comprender y extraer información relevante de los datos generados por la simulación.

Este análisis permite evaluar la precisión del modelo de simulación, y ayuda a identificar posibles problemas o limitaciones en el modelo. Además, proporciona una base para la toma de decisiones informadas y la identificación de áreas de mejora en el sistema simulado.

Nosotros hemos analizado la media, la varianza, la desviacón estándar y el percentil 95 de las variables que identificamos como de interés en la sección 2.2. Un ejemplo de este análisis se muestra más adelante en la sección 5.2.1

4.4 Análisis de parada de la simulación

Las colas se forman cuando hay un desequilibrio entre la demanda del servicio y la capacidad del sistema. Se debe encontrar un equilibrio entre proporcionar un buen servicio y minimizar el tiempo de espera. El análisis de parada ayuda a determinar cuándo agregar o reducir recursos (como servidores) para optimizar el sistema. A medida que se mejora el servicio disminuye el tiempo en la cola, pero aumentan los costos.

Para llevar a cabo un correcto análisis de parada, se deben generar simulaciones variando los parámetros λ, μ y n, observando cómo cambian las métricas y tratar de llegar a un balance.

Un índice a tomar en cuenta es el valor de $\rho = \lambda/c\mu$, ya que $\rho \lessdot 1$ indica que la cola puede alcanzar un estado estable y no crecer infinitamente.

En resúmen, el análisis de parada en simulaciones de colas es esencial para tomar decisiones informadas sobre la gestión de recursos y la eficiencia del sistema.

5 Modelo Matemático

5.1 Descripción del modelo de la simulación

En esta simulación, modelamos una cola M/M/c o de Erlang.
- M: Indica que las llegadas de clientes siguen una distribución de Poisson (M), lo que significa que los intervalos de tiempo entre las llegadas sucesivas de clientes siguen una distribución exponencial.
- M: También indica que los tiempos de servicio en el sistema siguen una distribución exponencial, lo que significa que la duración del servicio para cada cliente es independiente y sigue una distribución exponencial.
- c: Representa el número de servidores en el sistema. En el sistema M/M/c, hay múltiples servidores disponibles para atender a los clientes que llegan a la cola.
El tamaño de la cola es infinito, lo que significa que no hay límite para la cantidad de clientes que pueden esperar en la cola.

Usamos estas distribuciones ya que en la literatura se ha demostrado que son las más adecuadas para modelar sistemas de colas, y podremos usar los resultados de la teoría de cola para comparar con los resultados obtenidos en la simulación.

5.2 Comparación con la teoría de colas

La teoría de colas es un campo de estudio que se ocupa de la gestión de líneas de espera y la asignación de recursos. La teoría de colas proporciona herramientas y técnicas para analizar y comprender el comportamiento de los sistemas de colas y para evaluar su desempeño. Este modelo se compara con la teoría de colas para validar los resultados obtenidos. La teoría de colas proporciona fórmulas y ecuaciones para calcular y predecir el comportamiento de los sistemas de colas.

- $\rho = \lambda$ / (cμ) representa la utilización promedio del sistema, que es la fracción de tiempo que los servidores están ocupados. Si rho es mayor que 1, el sistema está sobrecargado y no puede manejar la tasa de llegada de clientes. Las métricas solo son confiables si $\rho < 1$.

- c: representa el número de servidores en el sistema.
- λ: representa la tasa de llegada promedio de los clientes al sistema siguiendo una distribución de Poisson.
- μ: representa la tasa de servicio promedio de los servidores en el sistema siguiendo una distribución exponencial.

Se utilizaron las siguientes fórmulas matemáticas para comparar con los resultados obtenidos en la simulación:

- Probabilidad de que el cliente tenga que esperar en la cola: La probabilidad de que un cliente tenga que esperar en la cola se puede calcular utilizando la fórmula de Erlang C, que se define como:

$$C(c,p) = \frac{1}{1 + (1-p)(\frac{c!}{(cp)^c})(\sum_{k=0}^{c-1} \frac{cp^k}{k!})}$$

- Cantidad promedio de clientes en el sistema: La cantidad promedio de clientes en el sistema se puede calcular utilizando la fórmula conocida como de Little:

$$L_s = \lambda W_s$$

- Probabilidad de que el sistema tenga 0 clientes: La probabilidad de que el sistema tenga 0 clientes se puede calcular utilizando la fórmula:

$$P_0 = \frac{1}{\sum_{m=0}^{c-1} \frac{(c\rho)^m}{m!} + \frac{(c\rho)^c}{c!(1-\rho)}}$$

- Cantidad promedio de clientes en la cola: La cantidad promedio de clientes en la cola se puede calcular utilizando la fórmula:

$$L_q = \frac{P_0(\frac{\lambda}{\mu})^c \rho}{c!(1-\rho)^2}$$

- Tiempo promedio que un cliente pasa en la cola: El tiempo promedio que un cliente pasa en la cola se puede calcular utilizando la fórmula:

$$W_q = \frac{L_q}{\lambda}$$

- Tiempo promedio que un cliente pasa en el sistema: El tiempo promedio que un cliente pasa en el sistema se puede calcular utilizando la fórmula:

$$W_s = W_q + \frac{1}{\mu}$$

5.2.1 Comparación

A continuación presentamos los resultados obtenidos de 3 simulaciones acompañados de los resultados que la teoría predice.

Valores	$\lambda = 2$, μ=1, c=8		$\lambda = 3$, μ=3, c=4		$\lambda = 1$, μ=2, c=4	
	Obtenido	Aprox.	Obtenido	Aprox.	Obtenido	Aprox.
Prob. de encolarse	0.0011	0.0011	0.02	0.02	0.0018	0.0018
Cant. de clientes	960.52	960.18	1438.47	483.26	480.911	240.12
Prob. 0 clientes	0.1366	0.1353	0.3676	0.3673	0.6081	0.6064
Longitud de cola	1.139	0.1833	29.101	3.2653	0.882	0.1237
Tiempo en cola	0.00019	0.00019	0.0022	0.0022	0.00027	0.00025
Tiempo en sistema	0.9986	1.0001	0.3357	0.3356	0.5001	0.5002

Ejemplo de análisis estadístico que realizamos a las variables:

Variable	Mean	Variance	Standard Error	95% Confidence Interval	Amplitude
Max Waiting Time	4.4402	1.667	0.040829	(4.360175425150171, 4.520224943070586)	0.16005
Probabilities of Waiting	0.441734	0.00243197	0.00155948	(0.4386773709563807, 0.44479051782573975)	0.00611315
Num Clients Served	961.181	890.369	0.943593	(959.3315568289387, 963.0304431710614)	3.69889
Num Clients	962.034	895.775	0.946454	(960.1789507351864, 963.8890492648136)	3.7101
Probability of 0 clients in the system	0.11194	0.000327493	0.00057227	(0.11081845355508466, 0.11306175041391789)	0.0022433
Avg Clients in Queue	425.918	3249.27	1.80257	(422.3849573961354, 429.4510426038646)	7.06609
Avg Time on Queue	0.437005	0.0198147	0.00445137	(0.4282802747544023, 0.4457296418131944 7)	0.0174494
Avg Time in System	1.4347	0.0257458	0.00507404	(1.4247524553685424, 1.4446426743985177)	0.0198902

Como se puede observar, la mayoría de los resultados son consistentes con los resultados teóricos. Esto indica que el modelo de simulación implementado es preciso y puede utilizarse para aproximar el comportamiento y el rendimiento del sistema M/M/c.

5.3 Supuestos y Restricciones

Los supuestos y restricciones del modelo son:

- Movimiento instantáneo entre servidores: Se asume que el movimiento de clientes entre servidores es instantáneo, lo que significa que un cliente puede pasar de un servidor a otro sin demora alguna.
- Distribución invariable de llegadas de clientes: Se supone que la distribución de llegadas de clientes no cambia con el tiempo. Esto implica que la tasa de llegada de clientes se mantiene constante durante toda la simulación.

- Permanencia de clientes no atendidos: Se establece que un cliente que llega al sistema permanece en el sistema hasta que es atendido por un servidor o hasta que concluye el tiempo de servicio sin ser atendido. No se permite que un cliente abandone el sistema sin ser atendido, a menos que expire el tiempo de servicio.