UNIVERSIDAD DE LA HABANA

Facultad de Matemática y Computación

Primer Proyecto de Simulación

Servidores Especializados vs Servidores Generalistas

Autor: Lidier Robaina Caraballo

Grupo: C-411

13 de abril de 2025

Índice general

Capítulo 1

Introducción

En el ámbito de la gestión de servicios, la elección entre estrategias especializadas o generalistas representa un dilema recurrente, donde la eficiencia operativa y la experiencia del usuario dependen de la configuración de los recursos disponibles. Este proyecto aborda dicho desafío mediante la simulación computacional basada en eventos discretos, un enfoque que permite modelar sistemas dinámicos bajo condiciones controladas.

El estudio se centra en comparar dos escenarios: uno con servidores dedicados a tareas específicas (estrategia especializada) y otro en el que los servidores son flexibles y pueden atender múltiples tipos de demandas (estrategia generalista). Aunque el análisis se contextualiza en un caso sencillo de una sucursal bancaria, los resultados se orientan a ofrecer conocimientos aplicables a sistemas de servicios más amplios y complejos.

1.1 Objetivos y metas

El proyecto tiene como objetivo principal analizar y comparar el desempeño de las dos configuraciones operativas mencionadas anteriormente. Para ello, se definen las siguientes metas:

1. Calcular la eficiencia global de cada estrategia, cuantificando métricas clave como: congestión del sistema, tiempo de espera del usuario, probabilidad de retrasos críticos y subutilización de recursos.

2. Evaluar trade-off entre la flexibilidad operativa (capacidad de atender múltiples tareas) y la velocidad de servicio, considerando posibles incrementos en los tiempos de atención al adoptar una estrategia generalista.

3. Proporcionar recomendaciones basadas en datos para la optimización de sistemas de servicios, extrapolables a contextos como logística, atención al cliente o salud.

1.2 Sistema específico a simular

> Una pequeña sucursal de un banco tiene dos empleados, uno para los pagos y otro para los cobros. Los clientes llegan a cada caja siguiendo una distribución de Poisson con una media de 20/hora (el total de llegada al banco es de 40/hora). El tiempo de servicio de cada empleado es una negativa exponencial de media 2 minutos. El encargado de la sección está pensando hacer un cambio en que los dos operarios puedan hacer tanto pagos como cobros para evitar situaciones en que una cola está llena y la otra parada. Sin embargo, se estima que cuando los empleados se encarguen de las dos cosas el tiempo de servicio aumentará a una media de 2,4 minutos. Compara el sistema que se emplea ahora con el propuesto, calculando el total de gente en el banco, el tiempo medio que pasaría un cliente en el banco hasta que es atendido, la probabilidad de que un cliente espere más de cinco minutos y el tiempo medio que están parados los empleados. [1]

De la definición del problema se obtienen las siguientes variables de interés:

- **L:** Media del total de clientes en el sistema en cada instante (congestión del sistema)
- $\mathbf{W_q}$**:** Media del tiempo durante el que un cliente permanece en la cola (tiempo de espera)
- $\mathbf{P(W_q > t_k)}$**:** Probabilidad de que el tiempo de espera sea superior a un tiempo determinado (retrasos críticos)
- $\mathbf{t_{free}}$**:** Media del tiempo durante el cual los servidores no tienen clientes (subutilización de recursos)

1.3 Variables que describen el problema

- λ_1**:** tasa de llegadas de clientes para pagos (distribución Poisson con media λ_1)
- λ_2**:** tasa de llegadas de clientes para cobros (distribución Poisson con media λ_2)
- $\mathbf{t_1}$**:** tiempo de atención en el servidor 1 (distribución exponencial con media t_1)
- $\mathbf{t_2}$**:** tiempo de atención en el servidor 2 (distribución exponencial con media t_2)
- **s:** estrategia especializada vs generalista (variable categórica)
- $\mathbf{t_k}$**:** tiempo de espera crítico

En caso de estrategia especializada, el servidor 1 atiende los pagos y el servidor 2 los cobros. En caso de estrategia generalista, $t_1 = t_2$.

[1] Problema 6.5 de [1]

Capítulo 2

Implementación

2.1 Detalles de implementación

1. **Gestión de Eventos:** Se emplea una cola de prioridad (módulo `heapq`) para manejar la lista cronológica de eventos. Cada evento contiene:
    - Marca temporal de ejecución
    - Tipo (llegada o salida)
    - Metadatos específicos (índice de servidor/cola)
2. **Estructuras de Datos:**
    - **Colas de espera**: Arreglos separados para cada servicio en modo especializado vs cola única compartida en modo generalista
    - **Estado de servidores**: Arreglo booleano que indica disponibilidad
    - **Contadores de clientes**: Registro separado por colas (especializado) o contador único (generalista)
3. **Mecánica de Simulación:**
    - **Llegadas**: Generadas mediante proceso Poisson usando `random.expovariate()`
    - **Tiempos de servicio**: Modelados con distribución exponencial negativa
    - **Asignación de servidores**: Política FIFO con prioridad a servidores disponibles
4. **Recolección de Métricas:**
    - *Área acumulativa* para cálculos promediados en el tiempo
    - Lista de tiempos de espera individuales
    - Registro de ocupación de servidores
    - Cálculo final mediante integración temporal (método de área bajo la curva)

2.2 Pasos de la simulación

El flujo de ejecución sigue esta secuencia lógica:

1. **Inicialización:**
   - Crear estructura de colas según estrategia
   - Programar primeros eventos de llegada usando tasas λ_1 y λ_2
   - Inicializar contadores y registros estadísticos
2. **Bucle Principal de Eventos:**

```
while time < sim_time:
    event = heappop(events)
    actualizar_estadisticas()
    procesar_evento(event)
```

3. **Procesamiento de Llegadas:**
   - Insertar cliente en la cola correspondiente
   - Si hay servidor disponible:
     - Iniciar servicio inmediato
     - Registrar tiempo de espera cero
     - Programar evento de salida
   - Generar próxima llegada según distribución Poisson
4. **Manejo de Salidas:**
   - Liberar servidor
   - Si existen clientes en cola:
     - Extraer siguiente cliente
     - Calcular tiempo de espera (current_time - arrival_time)
     - Programar nuevo evento de salida
5. **Actualización Estadística:**
   - Calcular tiempo transcurrido desde último evento
   - Acumular:
     - Clientes-tiempo en sistema
     - Tiempo ocupado de servidores
   - Mantener precisión temporal mediante integración continua

Capítulo 3

Resultados y experimentos

3.1 Variables de interés

Se ejecutaron 1000 simulaciones independientes para cada configuración del sistema. Para cada métrica, se comparan las distribuciones resultantes de ambas estrategias mediante el histograma normalizado y la distribución normal correspondiente a la media y desviación estándar de los datos.

3.1.1. Análisis exploratorio

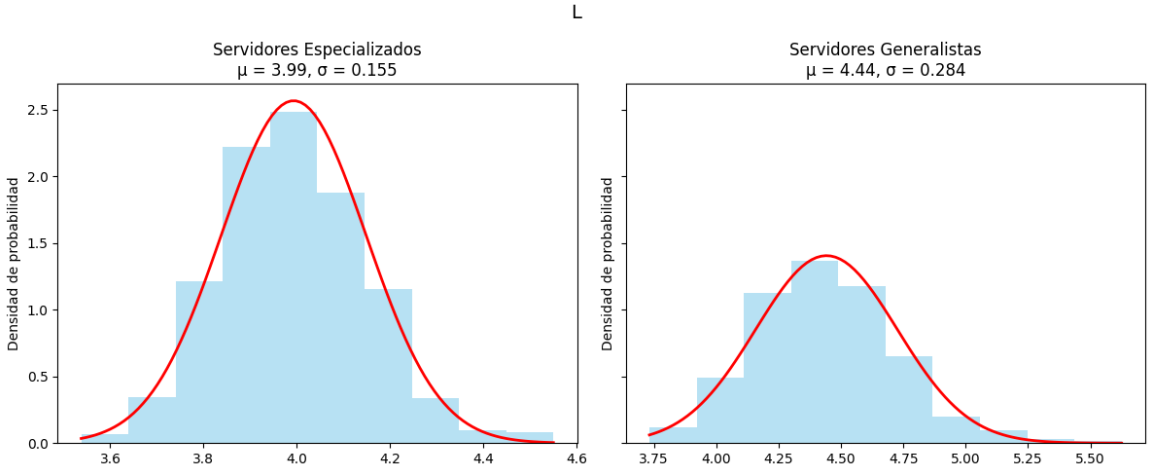

Figura 3.1: Histograma de la media de clientes en el sistema (L)

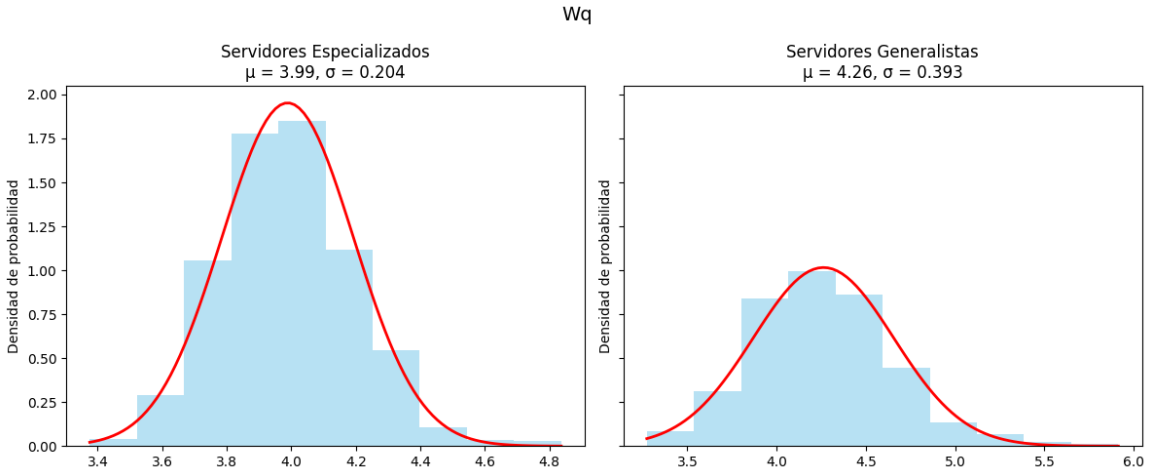

Figura 3.2: Histograma del tiempo (min) medio de espera (W_q)

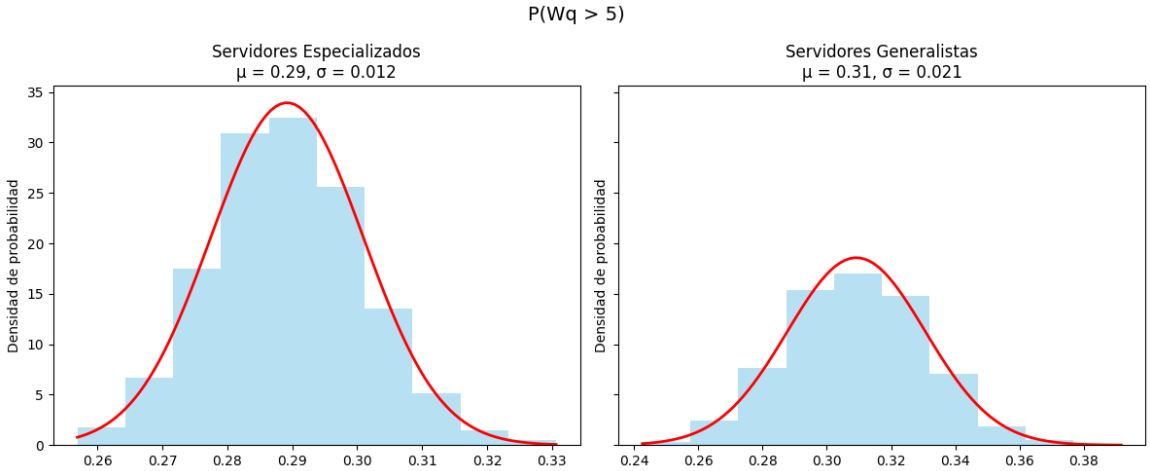

Figura 3.3: Histograma de la probabilidad de superar el umbral de 5 minutos de espera

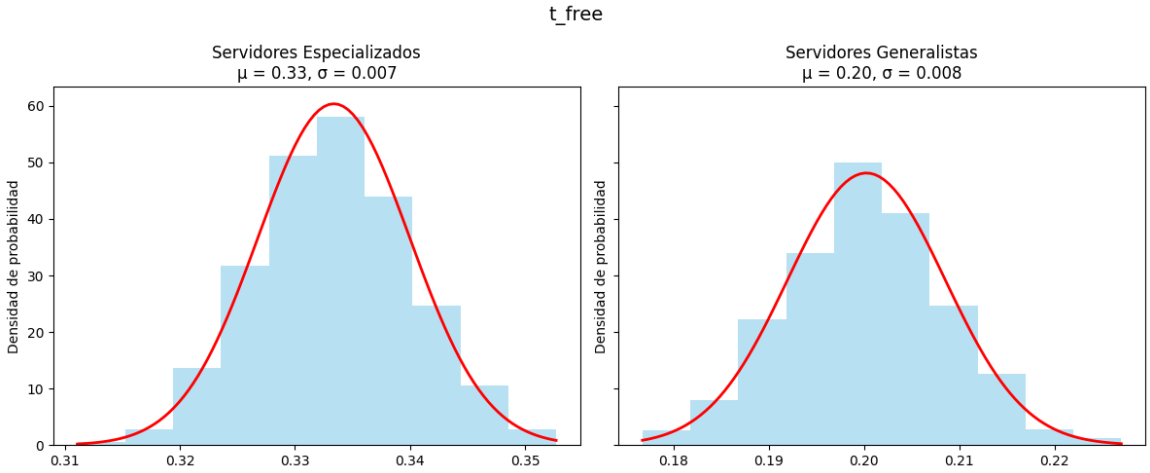

Figura 3.4: Histograma de la tasa de tiempo inactivo de los servidores

3.1.2. Interpretación de los resultados

- Al seguir una estrategia generalista, aunque se distribuye la carga entre dos servidores, el tiempo de servicio por cliente aumenta debido a la reducción de la especialización. Por tango, aumenta la congestión del sistema y el tiempo de espera de los usuario.
- La probabilidad de retrasos críticos no difiere significativamente, debido a que la distribución exponencial mitiga los efectos de las diferencias en las probabilidades de cola larga.
- En el sistema especializado, los servidores no pueden redistribuir la carga. Si una cola está vacía, su servidor permanece inactivo aunque haya demanda en la otra cola, por tanto tiene un peor uso de recursos.

3.1.3. Hipótesis extraídas

Hipótesis 1: Mayor congestión de clientes en sistema generalista

- **Hipótesis nula:** $H_0 : \mu_{\text{propuesto}} \leq \mu_{\text{actual}}$
- **Hipótesis alternativa:** $H_1 : \mu_{\text{propuesto}} > \mu_{\text{actual}}$

Hipótesis 2: Mayor tiempo de espera en sistema generalista

- **Hipótesis nula:** $H_0 : W_{\text{propuesto}} \leq W_{\text{actual}}$
- **Hipótesis alternativa:** $H_1 : W_{\text{propuesto}} > W_{\text{actual}}$

Hipótesis 3: Similar probabilidad de retrasos críticos

- **Hipótesis nula:** $H_0 : p_{\text{actual}} = p_{\text{propuesto}}$
- **Hipótesis alternativa:** $H_1 : p_{\text{actual}} \neq p_{\text{propuesto}}$

Hipótesis 4: Mayor inactividad en sistema especializado

- **Hipótesis nula:** $H_0 : I_{\text{actual}} \leq I_{\text{propuesto}}$
- **Hipótesis alternativa:** $H_1 : I_{\text{actual}} > I_{\text{propuesto}}$

3.1.4. Pruebas de Hipótesis

Se realizó un test t-student para cada hipótesis, en todos los casos se obtuvieron p-values muy cercanos a 0, así que en todos los casos se rechaza la hipótesis nula. Se puede asegurar que hay diferencias significativas en la media de cada variable para las dos estrategias.

Capítulo 4

Modelo matemático

4.1 Descripción del modelo

Consideramos dos sistemas de colas diferentes bajo teoría de colas markovianas, descritas detalladamente en [1]. Para ambos sistemas asumimos procesos Poisson de llegada con tasa λ y servicio con tasa μ. Definimos:

- $\rho = \frac{\lambda}{c\mu}$ (intensidad de tráfico, c cantidad de servidores en paralelo)
- L: Número promedio de clientes en el sistema
- L_q: Número promedio en cola
- W_q: Tiempo promedio en cola

4.1.1. Sistema Actual (Dos Colas M/M/1)

Para cada servidor simple independiente:

- $L = \frac{\lambda}{\mu - \lambda}$
- $W_q = \frac{\rho}{\mu - \lambda}$
- $P(W_q > t) = \rho e^{-(\mu-\lambda)t}$
- $P_{free} = 1 - \rho$

4.1.2. Sistema Propuesto (Una Cola M/M/2)

Para el sistema conjunto de los dos servidores en paralelo:

- $P_0 = \left[\sum_{k=0}^{1} \frac{(2\rho)^k}{k!} + \frac{(2\rho)^2}{2!(1-\rho)}\right]^{-1}$
- $L_q = \frac{(2\rho)^2 \rho}{2!(1-\rho)^2} P_0$

- $L = \frac{\lambda}{\mu} + L_q$
- $W_q = \frac{L_q}{\lambda}$
- $P(W_q > t) = \frac{2\rho e^{-2\mu(1-\rho)t}}{2-\rho} P_0$
- $P_{\text{free}} = 1 - \rho$

4.2 Supuestos y restricciones

- **Política FIFO:** Clientes son atendidos en orden de llegada
- **Independencia:** Llegadas y servicios son procesos independientes
- **Uniformidad:** Llegadas y servicios tienen la misma tasa para ambos sistemas y ambos servicios (valores de entrada del problema específico)
- **Espacio infinito:** Capacidad ilimitada en las colas
- **Homogeneidad:** Tasa de servicio constante en el tiempo

4.3 Comparación de resultados

Cuadro 4.1: Resultados Teóricos vs. Experimentales

Métrica	**Sistema actual** ($\lambda = 20$, $\mu = 30$)		**Sistema propuesto** ($\lambda = 40$, $\mu = 25$)	
	Teórico	Experimental	Teórico	Experimental
Clientes en el sistema	4.00	3.99	4.44	4.44
Tiempo en cola (min)	4.00	3.99	4.27	4.26
Prob. de retraso crítico	0.2895	0.29	0.3089	0.31
Tiempo inactivo (%)	33	33	20	20

Los resultados experimentales coinciden con las predicciones teóricas, confirmando la validez de los supuestos markovianos. Se concluye que ambos enfoques proveen una base sólida para la toma de decisiones operativas.

Capítulo 5

Conclusiones

1. Servidores especializados provocan bloqueo parcial del sistema debido a las colas separadas, lo que lleva a una subutilización de recursos.
2. Servidores generalistas son más lentos debido a la falta de especialización, provocando una peor experiencia de usuario.

Bibliografía

[1] SABATER, J. P. G. Aplicando Teoría de Colas en Dirección de Operaciones. [S.l.]: Grupo ROGLE, Departamento de Organización de Empresas, Universidad Politécnica de Valencia, 2015/2016.