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 (recursos dedicados a tareas específicas) y generalistas (recursos multifuncionales) constituye un dilema recurrente. La eficiencia del proceso y la experiencia del usuario están directamente vinculadas a cómo se estructuran los recursos disponibles [1] . Este proyecto aborda dicho desafío mediante la simulación basada en eventos discretos, una técnica que permite modelar sistemas dinámicos con precisión bajo condiciones controladas y replicables.

El estudio compara dos modelos de colas contrastantes:

a) Escenario especializado

- Servidores dedicados a un único tipo de tarea.
- *Configuración de colas*: Múltiples colas dedicadas (una por tipo de demanda).
- *Riesgo clave*: Costos asociados a desequilibrio de ocupación entre colas.

b) Escenario generalista

- Servidores flexibles que atienden múltiples demandas.
- *Configuración de colas*: Cola única atendida por todos los servidores.
- *Riesgo clave*: Incremento del tiempo de atención por cambios de contexto.

Aunque el análisis se desarrolla en un caso simplificado de una sucursal bancaria (procesos de retiros y depósitos), los hallazgos buscan extrapolarse a sistemas complejos como cadenas logísticas, centros de atención al cliente o redes hospitalarias.

1.1 Objetivos y metas

El objetivo principal es evaluar comparativamente el desempeño de ambas estrategias, con tres metas específicas:

1. Cuantificar la eficiencia global mediante métricas fundamentales.

2. Evaluar *trade-off* entre eficiencia y resiliencia.
3. Generar recomendaciones basadas en evidencia para optimizar sistemas de servicios.

1.2 Sistema específico a simular

El texto del problema fue extraído de [2] [1].

> 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.

Se tienen 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)
- $\mathbf{P(W_q > t_{max})}$**:** Probabilidad de que el tiempo de espera de un cliente sea superior a un tiempo determinado (retrasos críticos)
- $\mathbf{P_{idle}}$**:** Porcentaje del tiempo durante el cual los servidores no tienen clientes (capacidad ociosa)

1.3 Variables que describen el problema

- λ_1**:** tasa media de llegadas por hora de clientes para pagos (distribución de Poisson con parámetro λ_1)
- λ_2**:** tasa media de llegadas por hora de clientes para cobros (distribución de Poisson con parámetro λ_2)

[1] Problema Gasto de recursos por desequilibrio6.5: Sucursal Bancaria.

- **t_1:** media del tiempo (minutos) de atención en el servidor 1 (distribución exponencial con frecuencia $\mu_1 = \frac{60}{t_1}$ clientes atentidos por hora)
- **t_2:** media del tiempo (minutos) de atención en el servidor 2 (distribución exponencial con frecuencia $\mu_2 = \frac{60}{t_2}$ clientes atendidos por hora)
- **t_{max}:** tiempo (minutos) de espera crítico

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

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

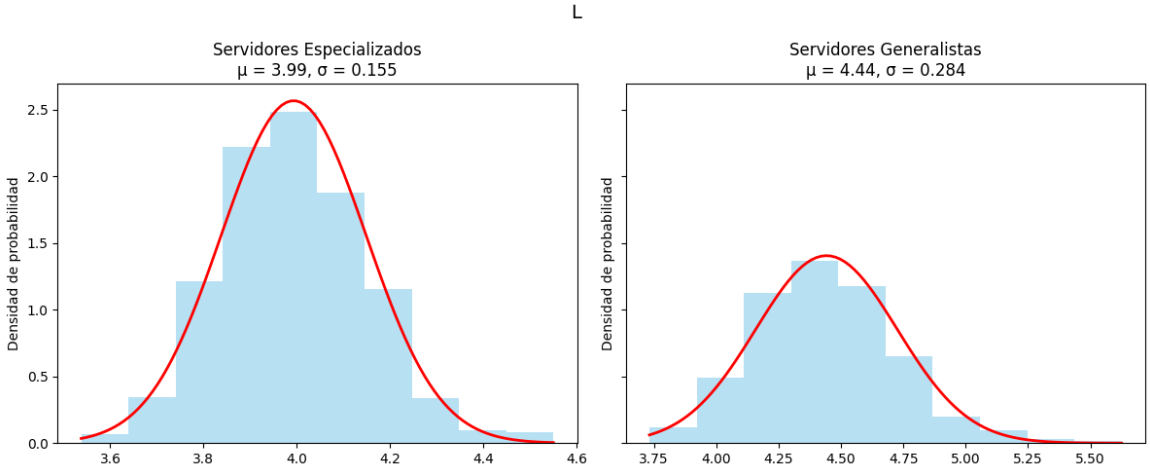

Figura 3.1: Histograma de la media de clientes en el sistema (L)

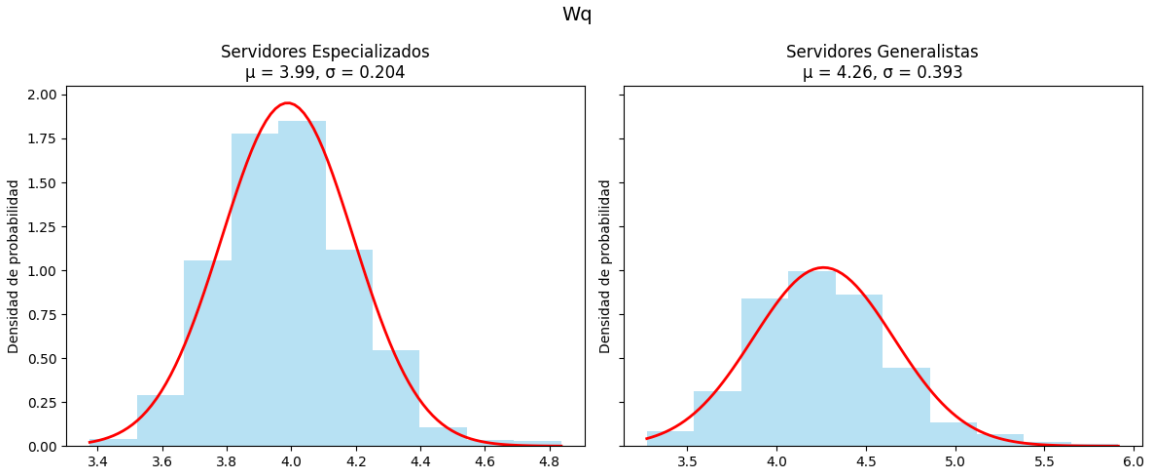

Figura 3.2: Histograma del tiempo (min) medio de espera (W_q)

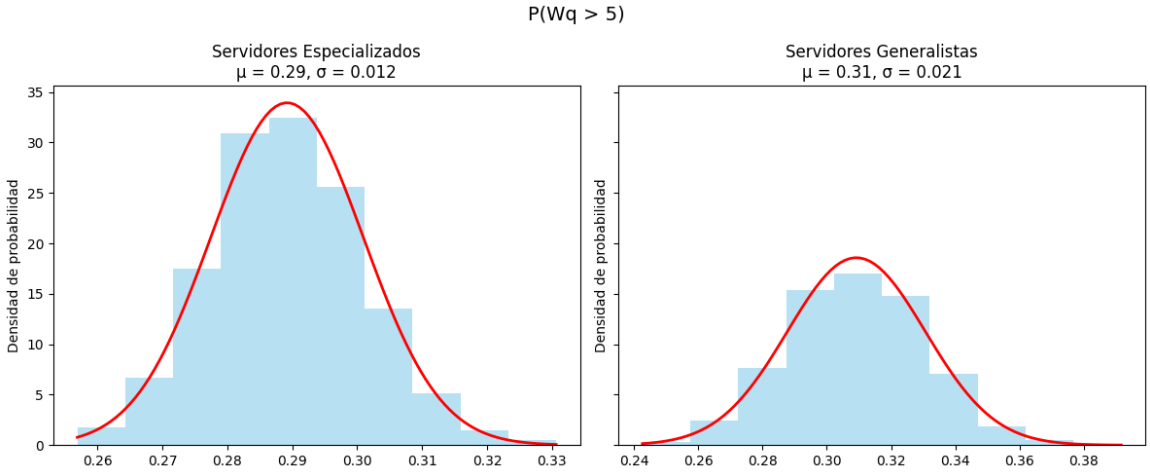

Figura 3.3: Histograma de la probabilidad de superar el umbral de 5 minutos de espera

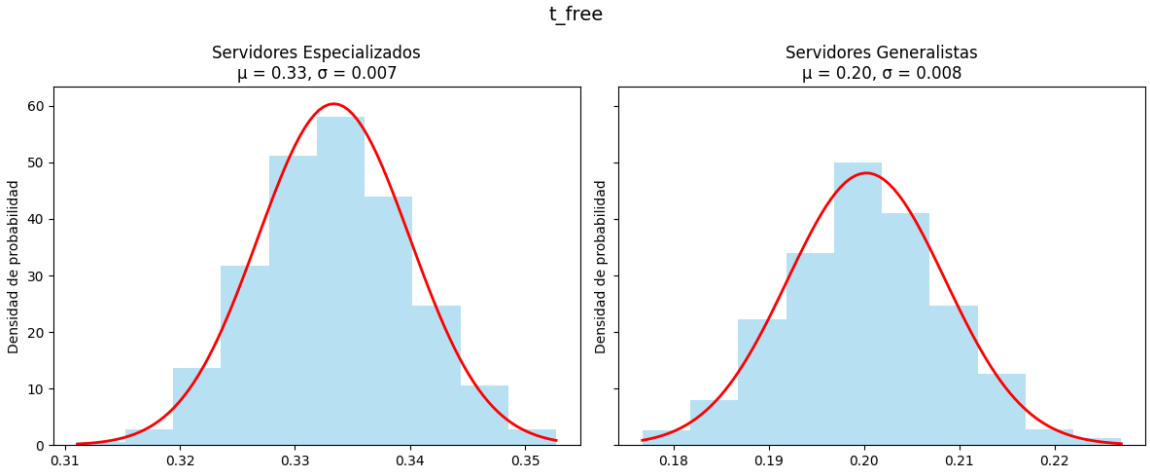

Figura 3.4: Histograma de la tasa de tiempo inactivo de los servidores

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

La teoría de colas analiza sistemas de espera mediante procesos estocásticos. En sistemas markovianos $M/M/c$, se asumen procesos Poisson de llegada con tasa λ y servicio con tasa μ, lo que permite modelar el sistema de nuestro problema como una cadena de Markov en tiempo continuo.

Definimos:

- c : Número de servidores en paralelo
- $\rho = \frac{\lambda}{c\mu}$: Factor de utilización de cada servidor
- $r = \frac{\lambda}{\mu}$: Número promedio de clientes que atienden en conjunto los servidores cada instante
- L: Número promedio de clientes en el sistema
- L_q: Longitud promedio de la cola
- W: Tiempo promedio en el sistema
- W_q: Tiempo promedio en cola
- P_0: Probabilidad de sistema vacío
- P_{idle}: Probabilidad de que un servidor no tenga clientes (capacidad ociosa)

Estas métricas clave se obtienen resolviendo ecuaciones de balance y aplicando la **Ley de Little**: $L = \lambda W$. En esta sección, las fórmulas que no están explícitamente justificadas fueron obtenidas de [2].

4.1.1. Sistema Especializado

Para cada servidor simple independiente ($\lambda = 20$, $\mu = 30$), aplicamos el modelo $M/M/1$.

- Clientes en sistema

  $$\rho = \frac{\lambda}{\mu} = \frac{2}{3}$$
  $$L = \frac{\rho}{1-\rho} = 2 \text{ por servidor} \Rightarrow 4 \text{ en total}$$

- Tiempo de espera

  $$W_q = \frac{\rho}{\mu - \lambda} = \frac{2}{30} \text{ horas} = 4 \text{ minutos}$$

- Probabilidad de retraso crítico

  En [3] se demuestra que el tiempo de espera en cola de un sistema $M/M/c$ se comporta según la siguiente distribución exponencial modificada:

  $$P(W_q \leq t) = 1 - P_w \; e^{-(c\mu - \lambda)t}$$

  donde P_w es la probabilidad de que el sistema esté lleno y el cliente tenga que esperar al llegar. En este caso $P_w = \rho$, por tanto:

  $$P(W_q > t) = 1 - P(W_q \leq t) = \rho e^{-(\mu - \lambda)t}$$
  $$P(W_q > 5) = P(W_q > \frac{1}{12} \text{ horas}) = \frac{2}{3} e^{-10\frac{1}{12}} \approx 0{,}2897$$

- Capacidad ociosa

  $$P_{idle} = 1 - \rho = \frac{1}{3}$$

4.1.2. Sistema Generalista

Para el sistema conjunto de los dos servidores en paralelo ($\lambda = 40$, $\mu = 25$, $c = 2$), aplicamos el modelo $M/M/2$.

- Clientes en sistema

  $$\rho = \frac{\lambda}{c\mu} = \frac{4}{5}$$
  $$r = \frac{\lambda}{\mu} = \frac{8}{5}$$
  $$P_0 = \left[\sum_{n=0}^{c-1} \frac{r^n}{n!} + \frac{r^c}{c!(1-\rho)} \right]^{-1} = \left[1 + r + \frac{r^4}{0{,}4} \right]^{-1} = \frac{1}{9}$$
  $$L_q = \frac{r^c \rho}{c!(1-\rho)^2} P_0 = \frac{2{,}56 \cdot 0{,}8}{2 \cdot 0{,}04 \cdot 9} = 2{,}8444 \quad \Rightarrow \quad L = L_q + r = 4{,}4444$$

- Tiempo de espera

  Aplicando Ley de Little:

  $W_q = \dfrac{L_q}{\lambda} = 0{,}07111 \text{ horas} \approx 4,2667 \text{ minutos}$

- Probabilidad de retraso crítico

  En este caso, $P_w = \frac{r^c P_0}{c!(1-\rho)}$, luego:

  $$P(W_q > t) = 1 - P(W_q \leq t) = \frac{r^c P_0}{c!(1-\rho)} e^{-(c\mu-\lambda)t}$$

  $$P(W_q > 5) = P(W_q > \frac{1}{12} \text{ horas}) = \frac{2{,}56 \cdot \frac{1}{9}}{2 \cdot 0{,}2} e^{-10\frac{1}{12}} \approx 0{,}3090$$

- Capacidad ociosa

  $P_{idle} = 1 - \rho = \dfrac{1}{5}$

4.2 Supuestos y restricciones

- **Proceso de nacimiento-muerte:** Las transiciones entre estados siguen una cadena de Markov con tasas λ (nacimientos) y μ (muertes). Los servidores se comportan de forma idéntica.
- **Equilibrio estable:** $\rho < 1$ garantiza estado estacionario ($\rho = 0{,}6667$ y $\rho = 0{,}8$ cumplen esto).
- **Política FIFO:** Clientes son atendidos en orden de llegada
- **Independencia:** Llegadas y servicios son procesos independientes
- **Memoria exponencial:** Los tiempos entre llegadas y servicios carecen de memoria, justificando el uso de distribuciones exponenciales.
- **Espacio infinito:** Capacidad ilimitada en las colas
- **Homogeneidad:** Tasas constantes en el tiempo

4.3 Comparación de resultados

Cuadro 4.1: Resultados Teóricos vs. Experimentales

Métrica	**Especializado**		**Generalista**	
	Teórico	Experimental	Teórico	Experimental
Clientes en el sistema	4	3.99	4.4444	4.44
Tiempo en cola (min)	4	3.99	4.2667	4.26
Prob. de retraso crítico	0.2897	0.29	0.3090	0.31
Tiempo inactivo (%)	33	33	20	20

La concordancia de resultados (error $< 1\,\%$) confirma los supuestos markovianos y la aplicabilidad del modelo. Se observa que ambos enfoques proveen una base sólida para la toma de decisiones operarivas.

Capítulo 5

Conclusiones

1. **Eficiencia cuantificada:**
    - El sistema especializado presenta mejor desempeño operativo inmediato gracias a su menor factor de utilización, pero con mayor capacidad ociosa.
    - El sistema generalista logra mayor eficiencia agregada al compartir recursos, pero con penalización en la congestión y los tiempos de espera.
2. **Trade-off entre eficiencia y resiliencia:**
    - El sistema especializado prioriza la rapidez en atención individual, pero su alta tasa de inactividad lo hace vulnerable a fluctuaciones en la demanda.
    - El sistema generalista, aunque presenta tiempos de espera mayores, garantiza mayor resiliencia al distribuir la carga entre servidores, reduciendo drásticamente la probabilidad de colapsos.
3. **Recomendaciones estratégicas:** La elección óptima depende del contexto operativo:
    - **Especializado** recomendable cuando:
      Los servicios tienen alta priorización de tiempos de respuesta (ej.: atención médica urgente).
      La congestión tiene consecuencias operativas severas.
    - **Generalista** preferible cuando:
      Los sistemas tienen demanda variable o alta incertidumbre.
      Los costes de capacidad ociosa superan los de congestión marginal.
    - **Solución híbrida** propuesta: Mantener especialización base + servidores generalistas flotantes para absorber picos de demanda.

Este análisis demuestra que no existe una solución universal óptima, sino configuraciones adaptativas donde la arquitectura de colas debe alinearse con los parámetros críticos del negocio y el perfil de riesgo organizacional.

Bibliografía

[1] Pinker, E. J., & Shumsky, R. A. (2000). The efficiency-quality trade-off of cross-trained workers. Manufacturing & Service Operations Management, 2(1), 32-48.

[2] Sabater, J. P., & ROGLE, G. (2015). Aplicando Teoría de Colas en dirección de operaciones.

[3] Gross, D., Shortle, J. F., Thompson, J. M., & Harris, C. M. (2011). Fundamentals of queueing theory (Vol. 627). John wiley & sons.