

Resumen Parcial 2

fedede

June 24, 2026

Contents

1	Intro	1
2	El Modelo M/M/2: Eficiencia y Estructura	2
2.1	Características y Beneficios	2
2.2	Parámetros Críticos y Estabilidad	3
3	Comparativa Estratégica: M/M/1 vs. M/M/2	3
4	Servidores Heterogéneos y Selección de Servidor	3
5	El Modelo M/G/1 y la Fórmula de Pollaczek-Khinchine	4
5.1	La Importancia de la Varianza	4
5.2	Aplicaciones de M/G/1	4
6	Sistemas con Prioridades (No Expulsivos)	4
6.1	Orden de Cálculo para el Tiempo de Espera (T_q)	4
7	Casos Prácticos y Modelización Numérica	5

1 Intro

El análisis de los sistemas de colas permite identificar cuellos de botella, optimizar recursos y mejorar la experiencia del usuario mediante la reducción de los tiempos de espera. Los hallazgos principales destacan que:

- Eficiencia del Sistema M/M/2

La transición de un sistema de servidor único (M/M/1) a uno doble con fila única reduce drásticamente el tiempo de permanencia, superando en eficiencia a la configuración de dos colas separadas.

- Impacto de la Variabilidad (M/G/1)

En sistemas con un solo servidor, la varianza del tiempo de servicio es un factor crítico; a mayor variabilidad, mayor es la longitud de la cola, incluso si la velocidad media se mantiene constante.

- Criterios de Intervención

La decisión de añadir servidores (especialmente si son de distinta velocidad) depende de un valor límite de utilización denominado “Rho Crítico” (ρ_c).

- Gestión de Prioridades

Los sistemas no expulsivos permiten dar atención preferencial a clases de clientes específicas sin interrumpir el servicio en curso.

2 El Modelo M/M/2: Eficiencia y Estructura

El modelo M/M/2 se define por tener arribos que siguen un proceso de Poisson, tiempos de servicio exponenciales y dos servidores que trabajan en paralelo atendiendo una fila única.

2.1 Características y Beneficios

- Detección de Cuellos de Botella: Permite identificar dónde se satura el proceso.
- Sensación de Equidad: Al utilizar una fila única, se evita la percepción de que otras filas avanzan más rápido, lo cual es común en sistemas de filas múltiples.
- Uso de Recursos: Maximiza la ocupación de los servidores, evitando que uno permanezca ocioso mientras el otro tiene clientes en espera.

2.2 Parámetros Críticos y Estabilidad

Para que el sistema sea estable, se debe cumplir la condición $\lambda < 2\mu$ (la tasa de llegadas debe ser menor a la capacidad total de servicio).

Característica	M/M/1	M/M/2
Número de servidores	1	2
Intensidad de tráfico (ρ)	$\frac{\lambda}{\mu}$	$\frac{\lambda}{2\mu}$
Tiempo en el sistema (W)	$\frac{1}{(\mu(1-\rho))}$	$\frac{1}{\mu(1-\rho^2)}$
Tiempo de espera	Mayor	Menor

3 Comparativa Estratégica: M/M/1 vs. M/M/2

La intervención en un sistema saturado es necesaria cuando las colas aumentan de forma sostenida o el servidor está cerca de su límite.

Diagnóstico de Intervención

- Sistema Aceptable: $E(n) < 3$ clientes.
- Necesita Intervención: $E(n) > 5$ clientes.

Caso de Ejemplo: Bajo una tasa de arribos $\lambda = 8$ y servicio $\mu = 10$:

1. En M/M/1, el tiempo medio en el sistema es de 30 minutos.
2. Al pasar a M/M/2, el tiempo medio se reduce a 7.1 minutos.

4 Servidores Heterogéneos y Selección de Servidor

Cuando se dispone de un servidor rápido (μ_1) y uno lento (μ_2), la decisión de añadir el segundo servidor depende del Rho Crítico (ρ_c).

- Sin selección de servidor: Los clientes entran al primer servidor disponible.
- Con selección de servidor: Existe una política para decidir a qué servidor enviar al cliente.
- Regla de Decisión: Si $\rho \geq \rho_c$, es conveniente agregar el segundo servidor. Si $\rho < \rho_c$, el sistema funciona mejor con uno solo.

5 El Modelo M/G/1 y la Fórmula de Pollaczek-Khinchine

A diferencia de los modelos anteriores, el M/G/1 permite que el tiempo de atención tenga cualquier distribución, requiriendo únicamente conocer su media ($E[S]$) y varianza ($\text{Var}[S]$).

5.1 La Importancia de la Varianza

La fórmula de Pollaczek-Khinchine para el número esperado de clientes en cola ($E[Lq]$) demuestra que la variabilidad es perjudicial: $E[Lq] = \frac{\rho^2 + \lambda^2 \cdot \text{Var}[S]}{2(1-\rho)}$

- M/D/1 (Servicio constante): Posee $\sigma = 0$, representando la referencia de mínima congestión.
- M/M/1 (Servicio exponencial): Representa la referencia de máxima congestión por su alta variabilidad inherente.

5.2 Aplicaciones de M/G/1

Este modelo es fundamental en entornos donde los tiempos de tarea varían según el caso, tales como:

- Hospitales: Consultas de duración variable.
- Redes / IT: Transmisión de paquetes de distinto tamaño.
- Manufactura: Piezas que requieren distintos tipos de trabajo.

6 Sistemas con Prioridades (No Expulsivos)

Este modelo se aplica cuando existen clientes de Clase 1 (prioritarios) y Clase 2. La prioridad es no expulsiva, lo que significa que no se interrumpe al cliente que ya está siendo atendido.

6.1 Orden de Cálculo para el Tiempo de Espera (T_q)

Un cliente de clase 1 debe esperar:

1. El residuo del tiempo del cliente que ya está en servicio (independientemente de su clase).

2. El tiempo de servicio de todos los clientes de clase 1 que ya estaban en la fila.
3. Posteriormente, recibe su propia atención.

7 Casos Prácticos y Modelización Numérica

Las fuentes detallan diversos escenarios de aplicación real:

- Fábrica de Periscopios: Un sistema de dos procesos consecutivos (corte y doblado) que requiere el cálculo de probabilidades de estado para determinar la saturación en cada etapa.
- Sector de Cajas de Supermercado: Operación mediante fila única que conduce a dos cajas registradoras idénticas. Se analizan métricas como la probabilidad de sistema vacío y el tiempo promedio en cola.
- Enrutamiento de Paquetes: Uso del Teorema de Little ($L = \lambda \cdot W$) para despejar la tasa de arribos y determinar el tiempo de permanencia en buffers de routers con alta capacidad de servicio.
- Sistemas de IA y APIs: Aplicación del modelo M/G/1 para gestionar solicitudes de backend donde la variabilidad del servicio es alta.