Ciência da Computação

Disciplina de Modelagem e Simulação

Teoria das Filas – Parte II

Aula 08

Professor: André Flores dos Santos

Santa Maria – RS
2025

Objetivos da Aula

- Compreender os conceitos fundamentais da Teoria das Filas
- Identificar sistemas que podem ser modelados como M/M/1
- Aplicar a notação de Kendall para classificar sistemas de filas
- Calcular as principais métricas de desempenho (L, Lq, W, Wq)
- Interpretar a utilização do sistema e sua relação com desempenho
- Resolver problemas práticos usando o modelo M/M/1
- Reconhecer limitações e quando usar outros modelos

Por que Estudar Filas?

Filas estão em todos os lugares na nossa vida digital e física!

Exemplos do Dia a Dia:

- Fila do banco ou supermercado
- Requisições em servidores web
- Processos aguardando CPU
- Chamadas em call center
- Pacotes de rede em roteadores
- Carros em semáforos

Impacto: Compreender filas ajuda a otimizar sistemas e reduzir tempos de espera

Em computação: desempenho de sistemas, dimensionamento de recursos, qualidade de serviço

Processo de Chegada e Atendimento

Processo de Chegada

Como e quando os clientes chegam ao sistema

- **Taxa de chegada (λ):** número médio de chegadas por unidade de tempo
- **Intervalos entre chegadas:** tempo entre uma chegada e a próxima
- **Padrão das chegadas:** regular, aleatório, ou seguindo alguma distribuição

Processo de Atendimento

Como os clientes são servidos pelo sistema

- **Taxa de atendimento (μ):** número médio de clientes atendidos por unidade de tempo
- **Tempo de serviço:** quanto tempo leva para atender cada cliente
- **Número de servidores:** quantos canais de atendimento existem

A relação entre λ e μ determina se o sistema será estável ou não

Disciplina da Fila e Capacidade

Disciplina da Fila

Regra que determina qual cliente será atendido primeiro

- **FIFO (First In, First Out):** primeiro a chegar, primeiro a ser atendido
- **LIFO (Last In, First Out):** último a chegar, primeiro a ser atendido
- **Prioridades:** clientes com maior prioridade são atendidos primeiro
- **Aleatória:** cliente é escolhido ao acaso da fila

Capacidade do Sistema

Limites físicos do sistema

- **Capacidade da fila:** quantos clientes podem aguardar
- **Capacidade total:** quantos clientes cabem no sistema (fila + atendimento)
- **População fonte:** quantos clientes podem potencialmente chegar

No modelo M/M/1: FIFO, capacidade infinita, população infinita

Notação de Kendall

Sistema padronizado para classificar modelos de fila (desenvolvido por D.G. Kendall, 1953)

A/B/c/K/N/D

Significado de cada elemento:

A = Processo de Chegada (M=Markoviano, D=Determinístico, G=Geral)

B = Tempo de Serviço (M=Markoviano, D=Determinístico, G=Geral)

c = Número de Servidores (1, 2, 3, ...)

K = Capacidade máxima do sistema (padrão: ∞)

N = Tamanho da população fonte (padrão: ∞)

D = Disciplina da fila (padrão: FIFO)

- Use **M** quando as chegadas são espontâneas, irregulares e a taxa é aproximadamente constante no período.
- Use **D** quando há agenda/ritmo fixo.
- Use **G** quando os dados mostram padrão diferente (ex.: rajadas, variação grande, horários muito distintos).

Formato simplificado: A/B/c quando K=∞, N=∞, D=FIFO

Exemplo: M/M/1 significa chegadas Markovianas, serviço Markoviano, 1 servidor

O que Significa M/M/1?

Vamos decifrar cada elemento da notação M/M/1:

Primeira letra M:

Processo de chegada Markoviano (sem memória)

Chegadas seguem processo de Poisson com taxa λ constante

Segunda letra M:

Tempo de serviço Markoviano (sem memória)

Tempos de serviço seguem distribuição exponencial com taxa μ

Número 1:

Sistema possui exatamente um servidor

Apenas um cliente pode ser atendido por vez

Premissas implícitas: capacidade infinita, população infinita, disciplina FIFO

M/M/1 = Sistema com chegadas Poisson, serviço exponencial, um servidor

Primeira e Segunda Letra M

O que significa 'Markoviano'?

Processo sem memória - o futuro depende apenas do estado presente

Primeira M - Chegadas (Processo de Poisson)

- Chegadas independentes e aleatórias
- Taxa de chegada λ constante ao longo do tempo
- Intervalos entre chegadas seguem distribuição exponencial
- Probabilidade de chegada não depende do histórico

Segunda M - Serviço (Distribuição Exponencial)

- Tempos de serviço independentes
- Taxa de atendimento μ constante
- Tempo restante de serviço não depende do tempo já decorrido
- Propriedade 'sem memória' simplifica a análise matemática

Na prática: assumimos aleatoriedade total em chegadas e atendimentos

Número 1 e Hipóteses do Modelo

O Número 1 - Um Servidor

Sistema possui exatamente um canal de atendimento

Apenas um cliente pode ser atendido simultaneamente

Hipóteses Fundamentais do M/M/1:

- Chegadas independentes seguindo processo de Poisson (taxa λ) - É um modelo para **chegadas "ao acaso"** com taxa média constante $\boldsymbol{\lambda}$ (chegadas por minuto, por exemplo)
- Tempos de serviço independentes e exponenciais (taxa μ)
- Capacidade infinita da fila (sem rejeição de clientes)
- População fonte infinita (sempre há clientes potenciais)
- Disciplina FIFO (primeiro a chegar, primeiro a ser atendido)
- Sistema em regime permanente (steady-state)
- **Condição de estabilidade: $\rho = \lambda/\mu < 1$**

Se $\rho \geq 1$, a fila cresce indefinidamente (sistema instável)

Quando Usar M/M/1?

Vantagens e Limitações

Quando usar o modelo M/M/1:

- Chegadas realmente aleatórias (sem padrões sazonais)
- Tempos de serviço muito variáveis
- Sistema com um ponto de atendimento
- Análise inicial ou aproximação rápida
- Sistema sem limitações de capacidade

Vantagens:

- Fórmulas analíticas simples e diretas
- Cálculos rápidos para análise inicial
- Base teórica sólida e bem estabelecida
- Fácil implementação em planilhas ou programas

Limitações:

- Pressupõe distribuições exponenciais (alta variabilidade)
- Não considera prioridades entre clientes
- Assume capacidade infinita (irrealista)
- Pode superestimar tempos de espera
- **Não adequado para sistemas com padrões regulares**

Exemplos Práticos de Aplicação

Sistemas de Computação:

- CPU processando requisições
- Servidor web atendendo requisições HTTP
- Sistema de arquivos (operações I/O)
- Banco de dados executando queries

Redes e Telecomunicações:

- Roteador processando pacotes
- Central telefônica atendendo chamadas
- Buffer de rede em switches
- Sistema de email processando mensagens

Outros Domínios:

- Caixa de banco ou supermercado
- Call center com atendentes
- Sistema de impressão em escritório
- Posto de gasolina (uma bomba)
- Máquina de café em empresa

Em todos os casos: chegadas aleatórias, um servidor, sem limitação de fila

O que Medir em uma Fila?

Para avaliar desempenho e otimizar sistemas, precisamos medir características quantitativas

Parâmetros Básicos do Sistema

- **Taxa de chegada (λ)** - quantos clientes chegam por unidade de tempo
- **Taxa de atendimento (μ)** - quantos clientes podem ser servidos por unidade de tempo
- **Utilização (ρ)** - fração do tempo que o servidor está ocupado

Métricas de Desempenho

- **Número médio de clientes no sistema (L)**
- **Número médio de clientes na fila (Lq)**
- **Tempo médio no sistema (W)**
- **Tempo médio na fila (Wq)**

Objetivo: entender como o sistema se comporta e identificar gargalos

Taxa de Chegada (λ) e Taxa de Atendimento (μ)

Taxa de Chegada (λ - lambda)

Número médio de clientes que chegam ao sistema por unidade de tempo

λ = 10 clientes/hora → em média, 10 clientes chegam a cada hora

λ = 5 processos/segundo → 5 processos chegam

por segundo à CPU

λ = 100 requisições/minuto → 100 requisições

HTTP por minuto

Unidade: clientes/tempo (hora, minuto, segundo, etc.)

Taxa de Atendimento (μ - mi)

Número médio de clientes que PODEM ser atendidos por unidade de tempo

μ = 15 clientes/hora → servidor consegue atender até 15 clientes/hora

μ = 8 processos/segundo → CPU consegue processar 8

processos/segundo

μ = 120 requisições/minuto → servidor web processa até

120 req/min

Unidade: clientes/tempo (mesma unidade que λ)

Para estabilidade: $\lambda < \mu$ (chegadas menores que capacidade de atendimento)

Utilização do Sistema (ρ)

Utilização (ρ - rho) representa a fração de tempo que o servidor está ocupado

$$\rho = \lambda/\mu$$

ρ = Taxa de Chegada / Taxa de Atendimento

Interpretação:

- $\rho = 0.5$ → servidor ocupado 50% do tempo (metade do tempo livre)
- $\rho = 0.8$ → servidor ocupado 80% do tempo (sistema sob pressão)
- $\rho = 0.95$ → servidor ocupado 95% do tempo (sistema crítico)

Impacto no Desempenho:

- $\rho < 0.7$ → Sistema confortável, baixos tempos de espera
- $0.7 \leq \rho < 0.9$ → Sistema sob pressão, tempos começam a crescer
- $\rho \geq 0.9$ → Sistema crítico, tempos de espera muito altos

FUNDAMENTAL: Para estabilidade, ρ DEVE ser < 1

Se $\rho \geq 1$: chegadas ≥ capacidade → fila cresce indefinidamente

Número Médio no Sistema e na Fila

Número Médio no Sistema (L)

Quantidade média de clientes no sistema (sendo atendidos + esperando)

$L = \rho/(1-\rho)$

Se $\rho = 0.8$, então $L = 0.8/(1-0.8) = 0.8/0.2 = 4$ clientes

Número Médio na Fila (Lq)

Quantidade média de clientes aguardando na fila (exclui quem está sendo atendido)

$Lq = \rho^2/(1-\rho)$

Se $\rho = 0.8$, então $Lq = 0.8^2/(1-0.8) = 0.64/0.2 = 3.2$ clientes

Relação: $L = Lq + \rho$ (sempre há 'ρ' clientes sendo atendidos em média)

Comportamento: quando 'ρ' aumenta, L e Lq crescem rapidamente

Quando $\rho \to 1$, tanto L quanto $Lq \to \infty$

Tempo Médio no Sistema e na Fila

Tempo Médio no Sistema (W)

Tempo total que um cliente passa no sistema (espera + atendimento)

$W = 1/(\mu-\lambda)$

Se $\lambda=10$ e $\mu=12.5$, então $W = 1/(12.5-10) = 1/2.5 = 0.4$ unidades de tempo

Tempo Médio na Fila (Wq)

Tempo que um cliente passa esperando antes de ser atendido

$Wq = \lambda/(\mu(\mu-\lambda))$

Com os mesmos valores: $Wq = 10/(12.5\times2.5) = 0.32$ unidades

Relação: $W = Wq + 1/\mu$ (tempo total = espera + atendimento)

Lei de Little

$L = \lambda \times W$

$Lq = \lambda \times Wq$

Conecta quantidades médias com tempos médios

Verificação: $L = 10 \times 0.4 = 4$ ✓ e $Lq = 10 \times 0.32 = 3.2$ ✓

Fórmulas Principais do M/M/1

Folha de Referência

Parâmetros Básicos:

λ = taxa de chegada

μ = taxa de atendimento $\rho = \lambda/\mu$ (utilização)

Tempos Médios: $W = 1/(\mu-\lambda)$ (no sistema)

$Wq = \lambda/(\mu(\mu-\lambda))$ (na fila)

Quantidades Médias:

$L = \rho/(1-\rho)$ (no sistema)

$Lq = \rho^2/(1-\rho)$ (na fila)

Relações Importantes:

$L = Lq + \rho$ (clientes sendo atendidos em média, ocupação média do servidor)

$W = Wq + 1/\mu$ (tempo total = espera + atendimento)

$L = \lambda \times W$ (Lei de Little)

$Lq = \lambda \times Wq$ (Lei de Little)

LEMBRE-SE: As fórmulas são válidas apenas se $\rho < 1$!

Exemplo 1: Caixa do Banco

Cenário:

Um banco possui um caixa eletrônico onde clientes chegam para realizar saques e depósitos. O gerente quer analisar o desempenho do sistema durante o horário de pico.

Dados Observados:

- Clientes chegam em média a cada 3 minutos
- Cada operação no caixa demora em média 2 minutos
- As chegadas parecem aleatórias (sem padrão específico)
- Os tempos de operação variam bastante entre clientes

Justificativa para usar M/M/1: chegadas aleatórias, tempos variáveis, um caixa

O que o gerente quer saber:

- Qual a utilização do caixa eletrônico?
- Quantos clientes estão no sistema em média?
- Quanto tempo um cliente demora para ser atendido?
- O sistema está funcionando adequadamente?

Na próxima tela: resolução passo a passo

Resolução do Exemplo 1

Passo 1: Determinar λ e μ

λ = 1 cliente a cada 3 min = 1/3 = 0.33 clientes/min

μ = 1 operação a cada 2 min = 1/2 = 0.50 operações/min

Passo 2: Calcular Utilização

$\rho = \lambda/\mu = 0.33/0.50 = 0.67 = 67\%$

O caixa está ocupado 67% do tempo (sistema confortável)

Passo 3: Calcular Métricas de Desempenho (No Sistema e Fila)

$L = \rho/(1-\rho) = 0.67/(1-0.67) = 0.67/0.33 = 2.0$ clientes

$W = 1/(\mu-\lambda) = 1/(0.50-0.33) = 1/0.17 = 6.0$ minutos

$Lq = \rho^2/(1-\rho) = 0.67^2/0.33 = 0.45/0.33 = 1.3$ clientes

$Wq = \lambda/(\mu(\mu-\lambda)) = 0.33/(0.50\times0.17) = 4.0$ minutos

Interpretação dos Resultados:

- Sistema em bom funcionamento ($\rho = 67\% < 70\%$)
- Em média 2 clientes no sistema, 1.3 esperando
- Cliente demora 6 min total (4 min espera + 2 min operação)
- Não há necessidade de outro caixa neste momento

Fórmulas:

Quantidades Médias de espera:

$L = \rho/(1-\rho)$ (no sistema)

$Lq = \rho^2/(1-\rho)$ (na fila)

Tempos Médios: $W = 1/(\mu-\lambda)$ (no sistema)

$Wq = \lambda/(\mu(\mu-\lambda))$ (na fila)

Exemplo 2: Sistema Computacional

Cenário:

Um servidor web recebe requisições HTTP. O administrador quer avaliar se o servidor está adequadamente dimensionado para a carga atual.

Dados do Sistema:

- Chegam 150 requisições/minuto (média)
- Servidor processa 200 requisições/minuto
- Chegadas são independentes (tráfego web)
- Tempos de processamento variam

Contexto: Requisições competem pelo mesmo recurso (CPU/memória).
Justificativa M/M/1: Chegadas independentes, tempos variáveis, um servidor.

Questões para o Administrador:

- Qual o percentual de utilização do servidor?
- Quantas requisições ficam em espera?
- Qual o tempo de resposta médio?
- O sistema suporta um aumento de 20% no tráfego?

Próximo slide: análise técnica completa

Resolução do Exemplo 2

Passo 1: Parâmetros do Sistema

$\lambda = 150$ req/min, $\mu = 200$ req/min

Passo 2: Análise de Utilização

$\rho = \lambda/\mu = 150/200 = 0.75 = 75\%$

Servidor operando a 75% da capacidade (zona de atenção)

Passo 3: Métricas de Desempenho

$L = \rho/(1-\rho) = 0.75/0.25 = 3.0$ requisições no sistema

$Lq = \rho^2/(1-\rho) = 0.5625/0.25 = 2.25$ requisições na fila

$W = 1/(\mu-\lambda) = 1/(200-150) = 0.02$ min $= 1.2$ segundos

$Wq = \lambda/(\mu(\mu-\lambda)) = 150/(200\times50) = 0.015$ min $= 0.9$ segundos

Passo 4: Análise de Capacidade (+20% tráfego, 100%=1, então 1+0,20)

Novo $\lambda = 150 \times 1.2 = 180$ req/min

Novo $\rho = 180/200 = 0.90 = 90\%$

Novo $W = 1/(200-180) = 0.05$ min $= 3.0$ segundos

Conclusões Técnicas:

- Sistema atual: desempenho adequado (1.2s tempo de resposta)
- Com +20% tráfego: sistema crítico (90% utilização, 3s para responder)
- Recomendação: upgrade do servidor ou load balancing antes do crescimento
- Monitoramento contínuo necessário na zona de 75-80% utilização

Fórmulas:

Quantidades Médias de espera:

$L = \rho/(1-\rho)$ (no sistema)

$Lq = \rho^2/(1-\rho)$ (na fila)

Tempos Médios: $W = 1/(\mu-\lambda)$ (no sistema)

$Wq = \lambda/(\mu(\mu-\lambda))$ (na fila)

Resumo e Próxima Aula

O que Aprendemos Hoje:

- Notação de Kendall e filas M/M/1
- Parâmetros: λ, μ, ρ e estabilidade
- Métricas de desempenho: L, Lq, W, Wq
- Lei de Little

Próximas aulas:

- Modelo M/M/c (múltiplos servidores)
- Modelo M/G/1 (serviço geral)
- Filas com capacidade limitada
- Sistemas com prioridades

Habilidades Práticas:

- Aplicar o modelo M/M/1
- Calcular métricas de desempenho
- Interpretar resultados para decisões
- Reconhecer zonas de utilização críticas

Fechamento — da Aula 07 para a Aula 08

Na Aula 07 vimos cenários reais com mais de um atendente (supermercado, banco, RU), mas ainda sem ‘batizar’ o modelo.

Hoje, na Aula 08, formalizamos o caso de 1 atendente (M/M/1).

Agora fechamos mostrando como reconhecer quando usar M/M/1

O que consolidamos hoje (M/M/1)

Hipóteses: chegadas ‘Markovianas’ (sem memória), serviço exponencial, FIFO, fila e população grandes.

Métricas principais: L, Lq, W, Wq, com checagem pela Lei de Little ($L=\lambda\cdot W$; $Lq=\lambda\cdot Wq$).

Leitura operacional: $\rho=\lambda/\mu$ indica quão ‘apertado’ está o sistema;
$\rho\to 1$ implica espera no limite.

Conexão com a Aula 07 (vários atendentes)

Quando há mais de um atendente, existem duas arquiteturas típicas:

• Fila única que alimenta vários atendentes equivalentes (modelo M/M/c).

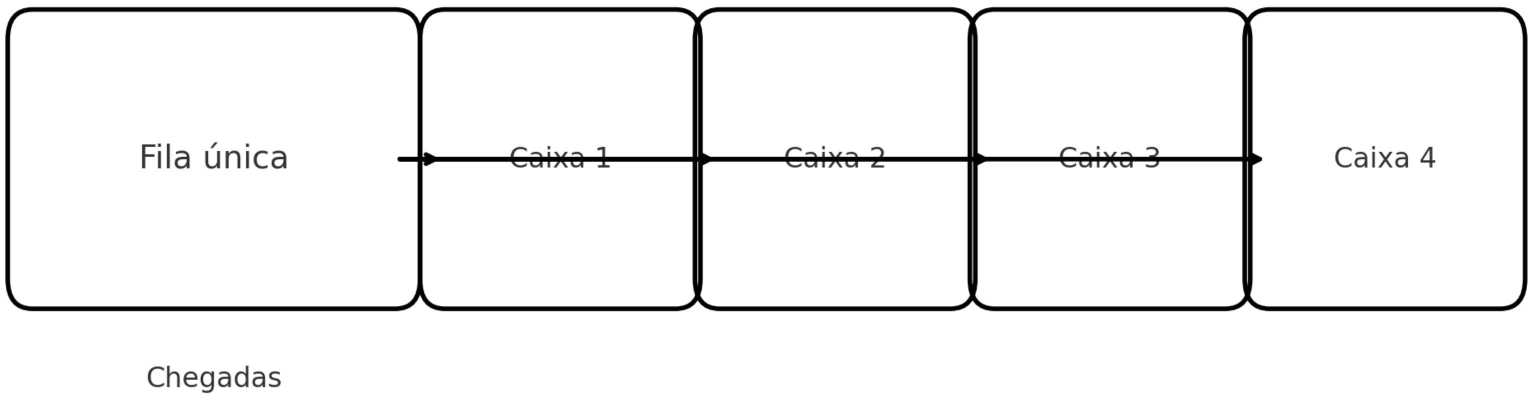
• Várias filas separadas, uma por atendente (vários M/M/1 independentes).

Saber qual é o layout real define qual modelo aplicar.

Figura — Uma fila única para vários caixas (M/M/c)

Uso típico: call center, guichês com senha única, atendimento com triagem ‘uma fila só’.

Vantagens: justiça (ninguém ‘pega a fila lenta’), menor variabilidade de espera, melhor nível de serviço por atendente.

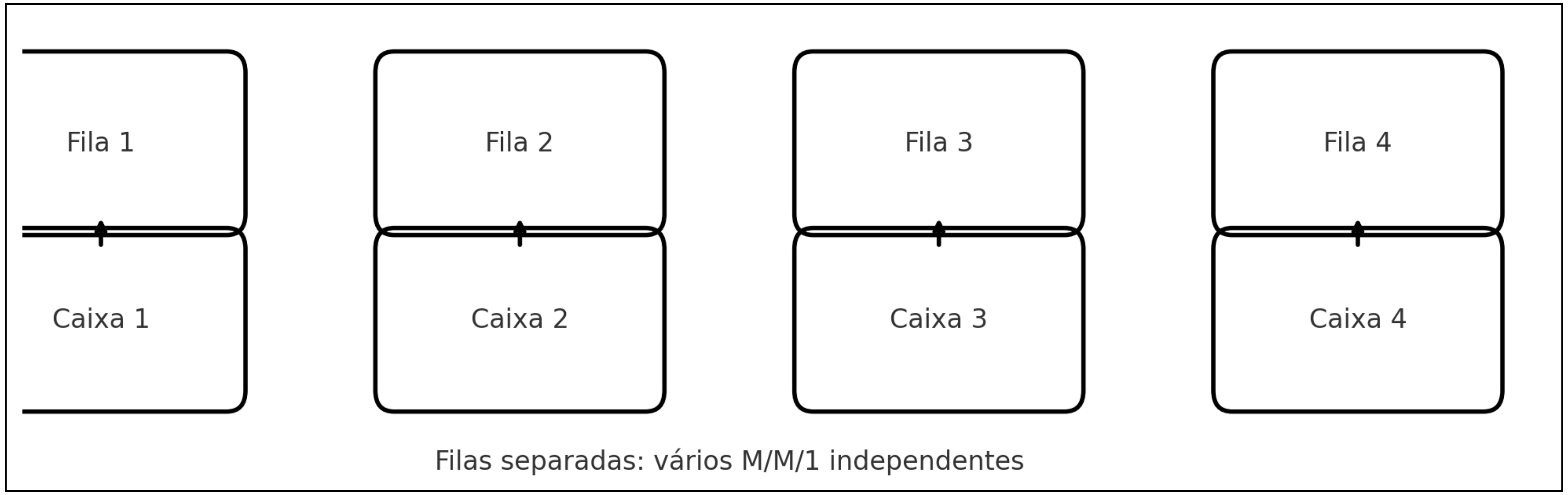


Fila única → próximo cliente vai para o primeiro caixa livre (M/M/c).

Figura — Várias filas, um caixa por fila (vários M/M/1)

Uso típico: supermercado com filas por caixa, eventos sem fila única.

Observação: clientes podem ‘escolher mal’ e pegar uma fila mais lenta;
A experiência varia de fila para fila.



Filas separadas → cada fila é um M/M/1 independente.

Quando é M/M/1 e quando é M/M/c?

- M/M/1: um único servidor; ou, mesmo com vários servidores, se cada um tem sua própria fila e você analisa uma fila específica.

- M/M/c: vários servidores idênticos atendendo a partir de uma fila única.

Regra prática para responder: ‘Há uma fila comum?
Então é M/M/c.

Há várias filas?
São vários M/M/1.’

Pequenas diferenças técnicas

Utilização:
M/M/1 → $\rho=\lambda/\mu$;   (deve ser <1).
M/M/c → $\rho=\lambda/(c\cdot\mu)$ (deve ser <1).

Probabilidade de esperar: em M/M/c usamos a fórmula de Erlang C para P(wait) (veremos na próxima aula).

A partir de P(wait), calculamos Wq, Lq, W, L.
A intuição é a mesma: quanto mais perto de 1 estiver 'ρ', maior a espera.

Como aplicar M/M/c na prática (4 passos)

1) Medir λ (chegadas/min), μ (serviço/min por atendente) e contar c (nº de atendentes).
2) Verificar estabilidade: $\rho=\lambda/(c\cdot\mu)<1$.
3) Calcular P(wait) por Erlang C (usaremos função pronta).
4) Derivar Wq, Lq, W e L e comparar com metas (ex.: Wq≤2 min ou P(wait)≤20%).

Exercícios:

Descrição no pdf em anexo.

Referências e material de apoio

PIDD, Michael. Modelagem empresarial: ferramentas para tomada de decisão. Porto Alegre: Artes Médicas: Bookman, 1998. 314 p.

PRADO, Darci; X PRADO, Darci Santos do. Teoria das filas e da simulação. Belo Horizonte, MG: Instituto de Desenvolvimento Gerencial - INDG, 1999. 122 p. (Pesquisa Operacional; 2).

Vicente Falconi Campos. Usando o Arena em simulação, 2000. (Biblioteca Digital)

KELTON, W. David; LAW, Averill M. Simulation modeling and analysis. 4. ed. Boston: Mc Graw Hill, 2007. 768 p.

BANKS, Catherine M., 1960-; SOKOLOWSKI, John A., 1953-. Principles of modeling and simulation: a multidisciplinary approach . New Jersey: Wiley, 2010. xiii, 259 p. : il. ISBN 978-0-470-28943-3

CHWIF, Leonardo; MEDINA, Afonso C. Modelagem e simulação de eventos discretos: teoria & aplicações. 2. ed. São Paulo, SP: Os Autores, c2007. 254 p.

ZEIGLER, Bernard P.; PRAEHOFER, Herbert; KIM, Tag Gon. Theory of modeling and simulation: integrating discrete event and continuous complex dynamic systems. 2nd ed. San Diego, Califórnia: Academic Press, 2010. xxi, 510 p. ISBN 9780127784557.

BARBETTA, Pedro Alberto; REIS, Marcelo Menezes; BORNIA, Antonio Cezar. Estatística para cursos de engenharia e informática. São Paulo, SP: Atlas, 2004. 410 p.

Obrigado pela sua atenção!!

Email: andre.flores@ufn.edu.br

Santa Maria – RS 2025