Análise de Desempenho de um Sistema de Filas com Servidores de Taxa de Atendimento Dinâmica

1st Alexandre de Araújo
Doutorado em Telecomunicação
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, Brasil
alexandre.araujo@dtel.inatel.br

2rd Daniel Malenga Moisés
Mestrado em Telecomunicação
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, Brasil
daniel.malenga@mtel.inatel.br

3st João Paulo Silva Dias
Doutorado em Telecomunicação
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, Brasil
joao.silva@dtel.inatel.br

***Abstract*—O artigo aborda a análise de desempenho de um sistema de filas com servidores de taxa de atendimento dinâmica. Destacamos a importância dos sistemas de filas nas telecomunicações. A abordagem principal do estudo é a avaliação do impacto da modificação dinâmica das taxas de atendimento dos servidores no sistema de filas. Comparação de resultados utilizando cenários com taxa de serviço constante. Além disso, investigamos diferentes configurações de tamanho de fila e distribuições de tempo de serviço para analisar como essas variáveis influenciam o desempenho global do sistema.**

***Index Terms*—Teoria de Filas, Python, Simulação**

I. Introdução

Os sistemas de filas desempenham um papel crucial em diversos setores, como telecomunicações, processamento de dados e operações industriais, onde o desempenho e a eficiência são fundamentais. Uma abordagem frequentemente utilizada para aprimorar o desempenho desses sistemas envolve o ajuste dinâmico das taxas de atendimento dos servidores com base na demanda atual.

Este estudo analisa o impacto dessa modificação dinâmica nas taxas de atendimento e compara os resultados com outros cenários, como taxas de atendimento constantes para ambos os servidores. Além disso, investiga diferentes configurações de tamanho de fila e distribuições de tempo de serviço para avaliar como essas variáveis influenciam o desempenho global do sistema.

Oliveira [1] apresentou uma técnica preditiva e preventiva, cuja aplicação auxilia na tomada de decisões, evitando futuras penalidades. O autor destaca ainda os benefícios da simulação de ambientes, que proporciona uma visão abrangente do sistema.

As filas resultam de um descompasso entre a capacidade de atendimento do serviço oferecido e a demanda dos usuários. Um sistema de filas pode ser descrito como clientes que chegam, esperam pelo serviço (caso não sejam atendidos imediatamente) e saem do sistema após o atendimento.

Segundo Moreira [2], a "teoria das filas é um corpo de conhecimentos matemáticos aplicado ao fenômeno das filas." Trata-se de um ramo da probabilidade que estuda a formação de filas por meio de análises matemáticas precisas e propriedades mensuráveis das filas.

Essa teoria fornece modelos que antecipam o comportamento de um sistema que oferece serviços cuja demanda cresce aleatoriamente, permitindo dimensioná-lo de forma a satisfazer os clientes e ser economicamente viável para o provedor do serviço, evitando desperdícios e gargalos Prado [3].

Observar uma quantidade de ocorrências em uma fila para descrever o modelo analítico com a realidade não é a abordagem mais recomendada. Portanto, o ideal é executar a simulação várias vezes e trabalhar com as médias dos resultados obtidos.

II. Estudo de Caso

Este estudo explora um sistema com dois servidores, S1 e S2, e uma única fila com *buffer* finito de tamanho J, onde os clientes chegam seguindo uma distribuição de Poisson com uma taxa média de λ pacotes por segundo. A particularidade deste sistema reside na capacidade dos servidores de alternar entre duas taxas de atendimento diferentes, μ_1 (mais lenta) e μ_2 (mais rápida), com base na quantidade de clientes presentes no sistema. A comutação da taxa de atendimento ocorre quando o número de elementos no sistema (fila + servidores) é igual ou superior a m.

III. Referencial Teórico

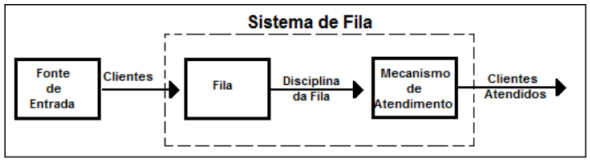
A. Teoria das Filas

De acordo com Moreira [2] e Moraes [4], a "teoria das filas é um corpo de conhecimentos matemáticos aplicado ao fenômeno das filas." Os sistemas de filas podem ser descritos, de maneira geral, como processos em que clientes (ou produtos) chegam a um sistema de atendimento (beneficiamento, produção) para receber um ou mais serviços executados por uma quantidade definida de servidores. As filas se formam quando a demanda pelo serviço supera a capacidade do sistema de atendimento.

Para avaliar o comportamento dos sistemas de filas, são associadas medidas de desempenho, como o tempo médio de espera dos clientes na fila, o tempo médio de chegada dos clientes, e a probabilidade de encontrar o sistema lotado, entre outras. Segundo Barbosa [5], os sistemas de filas são caracterizados por cinco componentes principais:

- **Modelo de chegada dos usuários:** distribuição de probabilidade dos intervalos de tempo entre as chegadas e as saídas dos usuários da fila.
- **Modelo de serviço:** distribuição de probabilidade dos tempos de serviço para cada usuário.
- **Número de atendentes:** quantidade de servidores disponíveis para realizar o atendimento.
- **Capacidade do sistema:** número máximo de usuários que podem permanecer ou entrar na fila, serem atendidos e saírem.
- **Disciplina da fila:** ordem na qual os usuários aguardam para acessar os serviços. A disciplina mais comum é FIFO (First In First Out), mas existem outras, como LIFO (Last In First Out), SIRO (Selection In Random Order), ou disciplinas baseadas em critérios de prioridade, como o mais novo, o mais antigo, o mais grave (em hospitais, por exemplo), entre outros.

Fig. 1. Representação esquemática de uma fila



Fonte: Barbosa [5]

A Figura 1 apresenta uma representação esquemática de uma fila, onde as equações desse modelo são baseadas nas seguintes características dos processos de chegada e de atendimento aos clientes:

- As chegadas seguem uma distribuição de Poisson com uma média de chegadas por unidade de tempo.
- Os tempos de atendimento seguem uma distribuição exponencial com média $1/mu$ (o número de atendimentos segue a distribuição de Poisson com média mu).
- O atendimento à fila é feito por ordem de chegada (FIFO).
- O número de clientes potenciais é suficientemente grande para que a população possa ser considerada infinita.

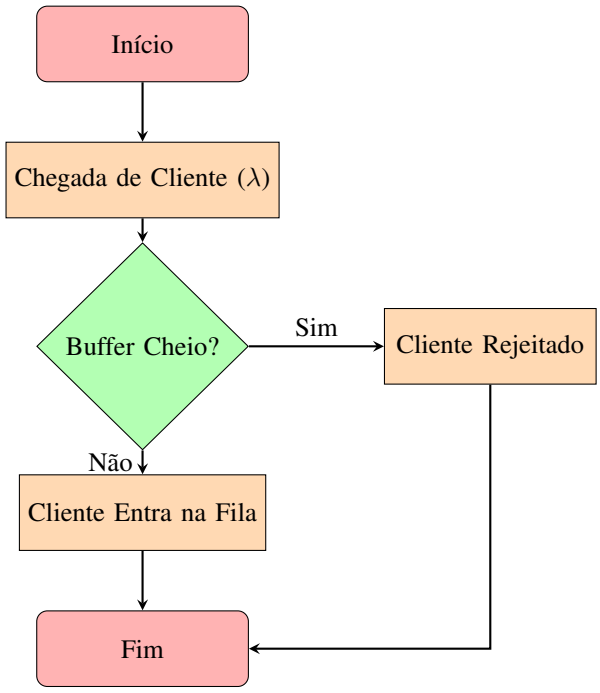
IV. Metodologia

A metodologia para análise deste sistema inclui as seguintes etapas:

1) **Definição dos Parâmetros:** Determinação dos valores para λ, μ_1, μ_2, J e m. Esses valores serão ajustados para simular diferentes cenários e avaliar o desempenho do sistema sob diversas condições.
2) **Simulação do Sistema:** Utilização de simulações estocásticas para modelar o comportamento do sistema. Cada simulação registrará métricas como o tempo médio de espera na fila, a utilização dos servidores e a taxa de perda de pacotes devido à capacidade finita da fila.
3) **Análise Comparativa:** Comparação do sistema dinâmico (com comutação de taxas) com outros sistemas:
   - **Taxas Constantes:** Ambos os servidores operando sempre na taxa μ_1 ou μ_2.
   - **Variação de Tamanho de Fila:** Avaliação do impacto de diferentes valores de J na performance do sistema.
   - **Distribuições de Tempo de Serviço:** Exploração de como distribuições diferentes do tempo de serviço (além da exponencial padrão) afetam o desempenho.
4) **Avaliação de Desempenho:** Análise das métricas de desempenho obtidas para cada configuração e comparação dos resultados. As métricas incluem:
   - **Tempo Médio de Espera:** Tempo médio que um cliente espera na fila antes de ser atendido.
   - **Utilização dos Servidores:** Proporção do tempo em que os servidores estão ocupados.
   - **Taxa de Perda de Pacotes:** Percentual de clientes que são rejeitados devido à fila cheia.

A. Chegada

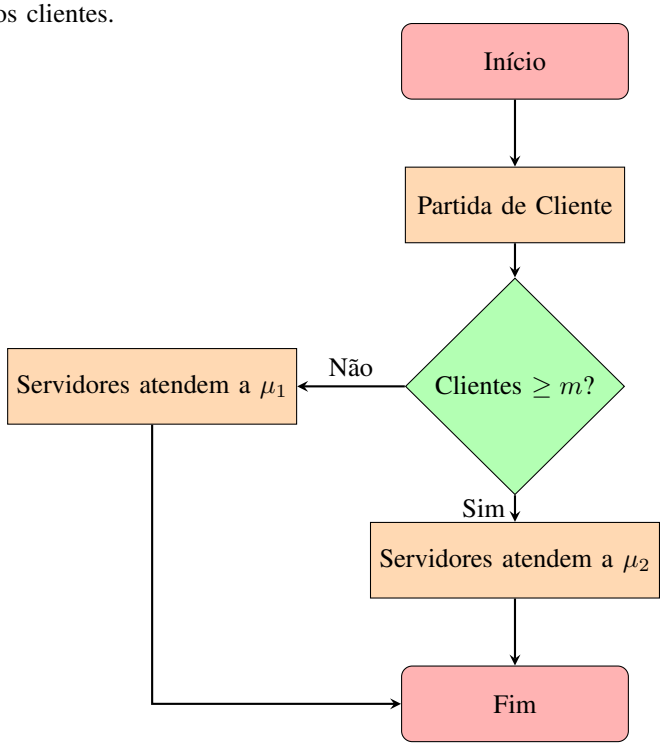
O fluxograma de chegada de clientes começa com a chegada de um cliente ao sistema, seguindo uma distribuição de Poisson com taxa λ. Em seguida, o sistema verifica se o *buffer* está cheio. Se o *buffer* estiver cheio, o cliente é rejeitado e o fluxo termina. Caso contrário, o cliente entra na fila, aguardando atendimento, e o fluxo termina.



B. Partida

O fluxograma de partida de clientes inicia quando um cliente sai do sistema após ser atendido. O sistema então verifica se o número de clientes presentes é maior ou igual a m. Se o número de clientes for maior ou igual a m, os servidores atendem os clientes a uma taxa mais rápida, μ_2. Se o número de clientes for menor que m, os servidores atendem os clientes

a uma taxa mais lenta, μ_1. O fluxo termina após o atendimento dos clientes.



V. Resultados

Para o desenvolvimento do código em Python, utilizou-se o Google Colab, uma plataforma baseada em nuvem que oferece um ambiente de desenvolvimento integrado (IDE) para a escrita e execução de código Python sem a necessidade de configurações complexas ou instalações locais. A estrutura do código foi desenvolvida seguindo a metodologia apresentada na Seção IV.

Os parâmetros iniciais de simulação são: $\lambda = 8$, $\mu_1 = 5$, $\mu_2 = 1$, $J = 16$ e $m = 8$. Foi importante definir um valor de λ superior aos valores de μ devido à característica do problema, e que μ_1 fosse significativamente maior que μ_2.

O objetivo da simulação é obter os valores do tempo médio no sistema (t_s), tempo médio de pacotes na fila (t_p), número médio de pacotes no sistema (N_{ps}) e número médio de pacotes na fila (N_{pf}). A Tabela 1 apresenta os resultados iniciais da simulação.

TABLE I
Resultados iniciais de simulação.

μ_1	μ_2	J	m	t_s	t_p	N_{ps}	N_{pf}
5	1	16	8	0.23362	0.03362	1.86899	0.53799

A Tabela 2 apresenta os valores de simulação, onde variamos os valores da taxa de serviço μ_2 de 2 a 5. A alteração significativa observada foi no número médio de pacotes na fila, especialmente quando $\mu_2 = 5$.

A Tabela 3 descreve os valores de simulação onde variamos os valores de J (8, 10, 12, 14). Com as taxas de serviço mantidas nos valores iniciais ($\mu_1 = 5$, $\mu_2 = 1$), a única alteração significativa observada foi no número médio de pacotes na fila.

TABLE II
Resultados de simulação alterando os valores de μ_2 .

μ_1	μ_2	J	m	t_s	t_p	N_{ps}	N_{pf}
5	5	16	8	0.23122	0.03122	1.84978	0.74936
5	4	16	8	0.24085	0.04085	1.92687	0.32687
5	3	16	8	0.23977	0.03977	1.91816	0.31816
5	2	16	8	0.23768	0.03768	1.90149	0.30149

TABLE III
Resultados de simulação alterando os valores de J.

μ_1	μ_2	J	m	t_s	t_p	N_{ps}	N_{pf}
5	1	8	8	0.23414	0.03414	1.87319	1.09279
5	1	10	8	0.24269	0.04269	1.94156	1.02469
5	1	12	8	0.23242	0.03242	1.85942	3.11307
5	1	14	8	0.23791	0.03791	1.90334	0.30334

A. Resultados das Simulação

1) **Parâmetros Iniciais:** Os parâmetros iniciais foram definidos como $\lambda = 8$, $\mu_1 = 5$, $\mu_2 = 1$, $J = 16$, e $m = 8$. Estes valores foram escolhidos para garantir que a taxa de chegada (λ) fosse superior à taxa de atendimento (μ), com μ_1 significativamente maior que μ_2, refletindo a natureza do problema em estudo.
2) **Análise dos Resultados Iniciais:** Com os parâmetros iniciais, os valores de desempenho do sistema foram: $t_s = 0.23362$, $t_p = 0.03362$, $N_{ps} = 1.86899$, e $N_{pf} = 0.53799$. Esses valores servem como referência para a avaliação das simulações subsequentes.
3) **Variação da Taxa de Serviço μ_2:** Alterando-se os valores de μ_2 de 2 a 5, observou-se que o aumento da taxa de serviço μ_2 resultou em uma diminuição do tempo médio de pacotes na fila (t_p) e no número médio de pacotes no sistema (N_{ps}). No entanto, o número médio de pacotes na fila (N_{pf}) apresentou um aumento significativo quando μ_2 foi igual a 5, indicando que o aumento na taxa de serviço pode impactar a fila de maneira não linear.
4) **Variação do Tamanho do *Buffer* (J):** Variando-se os valores de J (8, 10, 12, 14) e mantendo-se as taxas de serviço iniciais, notou-se que o número médio de pacotes na fila (N_{pf}) foi a métrica mais sensível às mudanças em J. Com valores menores de J, o sistema apresentou um número maior de pacotes na fila, sugerindo que a capacidade do sistema é um fator crucial para o desempenho do sistema de filas.

VI. Conclusão

A simulação demonstra que a definição adequada dos parâmetros λ, μ, J, e m é essencial para o desempenho eficiente do sistema de filas. Ajustes na taxa de serviço e no número de servidores impactam significativamente as métricas de desempenho. Portanto, a escolha desses parâmetros deve

considerar a natureza específica do sistema e a demanda esperada, visando minimizar o tempo de espera e o número de pacotes na fila para garantir um serviço mais eficiente e eficaz.

VII. Apêndice

```
import numpy as np
import matplotlib.pyplot as plt
lambda1 = 8
mu1 = 5
mu2 = 1
J = 16
m = 8
N = 10000
t = 0
nq = 0
ns = 0
k = 0
ta = np.random.exponential(1/lambda1)
td1 = float('inf')
td2 = float('inf')
tp = []
ts = []
proxima_chegada = t + ta

while k < N:
if proxima_chegada <= np.minimum(td1, td2):
t = proxima_chegada
if ns < 2:
ns += 1
if ns == m:
td1 = t + np.random.exponential(1/mu2)
else:
td1 = t + np.random.exponential(1/mu1)
else:
if nq < J:
nq += 1
k += 1
ta = np.random.exponential(1/lambda1)
proxima_chegada = t + ta
 else:
t = min(td1, td2)
if t == td1:
ns -= 1
tp.append(t)
ts.append(t - (t - ta))
if ns >= m:
td1 = t + np.random.exponential(1/mu2)
else:
if nq > 0:
nq -= 1
td1 = t + np.random.exponential(1/mu1)
else:
td1 = float('inf')
else:
ns -= 1
tp.append(t)
ts.append(t - (t - ta))
if ns >= m:
td2 = t + np.random.exponential(1/mu2)
else:
if nq > 0:
nq -= 1
td2 = t + np.random.exponential(1/mu1)
else:
 td2 = float('inf')
tempo_m_sistema = np.mean(ts)
tempo_m_fila = np.mean(ts) - (1/mu1)
num_m_pct_sistema = tempo_medio_sistema * lambda1
num_m_pcts_fila = tempo_medio_fila * lambda1 * nq
```

References

[1] OLIVEIRA, G. B. Simulação Computacional: Análise de um Sistema de Manufatura em Fase de Desenvolvimento. Dissertação de Mestrado. Universidade Federal de Engenharia de Itajubá: UNIFEI, 2007, 154p.

[2] MOREIRA, D. A. Pesquisa Operacional – Curso Introdutório. 2. ed. São Paulo: Thomson Learning, 2007.

[3] PRADO, D. S., Usando o ARENA em Simulação, Belo Horizonte, Editora de Desenvolvimento Gerencial, 2009.

[4] MORAES, F.G., SILVA, G.F., REZENDE, T.A. Introdução a Teoria das Filas. Universidade Federal do Mato Grosso, 2011.

[5] BARBOSA, R. A. Modelagem e análise do sistema de filas de caixas de pagamento em uma drogaria: uma aplicação da teoria das filas. In: XXIX Encontro Nacional De Engenharia De Produção: A Engenharia De Produção E O Desenvolvimento Sustentável Integrando Tecnologia E Gestão. Salvador, BA, Brasil, 06 a 09 de outubro de 2009.