Análise de Desempenho de Sistemas de Filas Baseado em dois Servidores e uma Fila única de Buffer Finito

Georgino Da Silva Baltazar, Eylen Jhuliana Mercado Ontiveros e Fábio Augusto Pereira

***Resumo*—Este trabalho apresenta uma análise de desempenho de um sistema de fila com dois servidores e um buffer de tamanho finito. A chegada das mensagens segue uma distribuição de Poisson com média λ, enquanto as durações de serviço são exponencialmente distribuídas com médias $1/\mu_1$ e $1/\mu_2$. A mensagem é encaminhada inicialmente ao servidor S1 e, se ocupado, redirecionada ao servidor S2. Foram analisadas três configurações distintas: 1) λ = 10, $\mu_1 = \mu_2 = 10$; 2) λ = 10\), $\mu_1 = 10$, $\mu_1 = 8$; 3) λ= 20, $\mu_1 = \mu_2 = 10$. Através de simulações, avaliamos a probabilidade de bloqueio, o número médio de elementos no sistema e o tempo médio no sistema em função do tamanho do buffer. Os resultados mostram que, para λ = 20, o sistema enfrenta maior sobrecarga, resultando em tempos médios no sistema e números médios de elementos significativamente maiores, além de uma maior probabilidade de bloqueio. Para λ = 10, o sistema demonstra um desempenho mais robusto com menores tempos médios e números médios de elementos no sistema, além de uma baixa probabilidade de bloqueio. Concluímos que o dimensionamento adequado do buffer e a configuração das taxas de serviço são cruciais para otimizar o desempenho do sistema. Estes achados podem ser aplicados em diversas áreas, como call centers, servidores web e sistemas de atendimento ao cliente, contribuindo para a melhoria da eficiência operacional e satisfação do usuário.**

Palavras Chaves*—Filas, M/M/2/J/J+2/∞/FCFS, M/M/2** ***e análises estatísticas.

I. INTRODUÇÃO

OS sistemas de filas são fundamentais para a análise de desempenho em diversas áreas, como redes de computadores, bancos, supermercados, aeroportos, centros de atendimento e muitos outros serviços onde há a necessidade de gerenciar o atendimento a clientes ou a processamentos de dados [1][2]. A teoria de filas, inicialmente desenvolvida para analisar sistemas telefônicos, tem sido amplamente aplicada para otimizar sistemas que envolvem espera e atendimento, visando reduzir o tempo de espera e melhorar a eficiência do sistema [3]

A crescente demanda por serviços rápidos e eficientes torna crucial a análise e otimização de sistemas de filas. Modelos de filas como M/M/1, M/M/2 e suas variações são frequentemente utilizados para prever o comportamento de sistemas reais e orientar decisões sobre a capacidade de atendimento necessária [4][5]. Além disso, a implementação de sistemas de fila com buffers finitos é comum em ambientes onde há limitação física ou lógica de armazenamento temporário de clientes ou dados, como em sistemas de comunicação e processamento de transações financeiras [6].

Neste trabalho, estudamos um sistema de filas com dois servidores e uma fila única de buffer finito. Os servidores operam com taxas de serviço fixas, e a fila possui capacidade limitada. A modelagem e simulação desse sistema têm como objetivo avaliar o impacto das taxas de chegada e dos tamanhos de buffer no desempenho geral do sistema, medido através de métricas como tempo médio no sistema, número médio de elementos no sistema e a probabilidade de bloqueio.

O artigo está estruturado da seguinte forma: na Seção II, apresentamos o modelo do sistema. A Seção III descreve a modelagem do Sistema, onde apresentamos o diagrama de transição de estados, os fluxogramas dos eventos de chegadas e partidas e algumas expressões matemáticas de interesse. Na Seção IV, apresenta os resultados obtidos e a análise comparativa entre diferentes configurações. A Seção V, mostra as aplicações. Finalmente, na Seção V, discutimos as conclusões e implicações dos resultados para a otimização de sistemas de filas.

II. MODELO DO SISTEMA

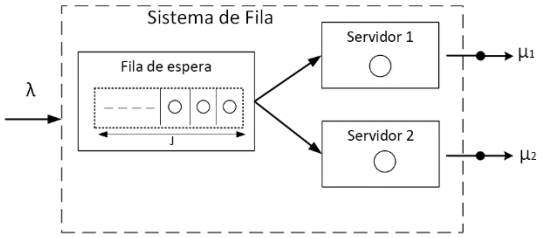
O modelo considera um sistema de fila com características *M/M/2/J/J+2/∞/FCFS*, onde: *M/M* indica que tanto os tempos entre chegadas quanto os tempos de serviço seguem distribuições exponenciais; *2* indica que há dois servidores no sistema; *J* representa o tamanho máximo do buffer; *J+2* representa o número total máximo de clientes no sistema (clientes em serviço + clientes na fila); ∞ indica que há um número infinito de possíveis chegadas de clientes; e *FCFS* (*First-Come, First-Served*) indica que os clientes são atendidos na ordem em que chegam. A chegada das mensagens obedece a uma distribuição Poissoniana de média λ e as durações de serviço têm distribuições genéricas com médias 1/μ1 e 1/μ2 como mostra a figura 1. A mensagem é sempre encaminhada inicialmente ao servidor S1. Se S1 estiver ocupado a mensagem é então encaminhada ao servidor S2. Se ambos os servidores estiverem ocupados, o cliente

G. S. Baltazar, E. J. M. Ontivero, e F. A. Pereira. Instituto Nacional de Telecomunicações - Inatel, Santa Rita do Sapucaí, MG, Brasil. E-mails: georgino.baltazar@dtel.inatel.br,eylen.ontiveros@mtel.inatel.br,fabio.pereira @mtel.inatel.br.

Relatório final de TP547, Princípios de Simulação de Sistemas de Comunicação. Um estudo de caso sobre sistemas de filas baseados em servidor duplo e Buffer Finito.

Julho de 2024.

entra na fila, desde que o buffer não esteja cheio, caso contrário, o cliente é bloqueado e não entra no sistema.



λ=Taxa de chegada; μ_1=Taxa de partida Servidor 1 μ_2=Taxa de partida Servidor 2
J=Tamanho de buffer.

Fig. 1. Modelo do Sistema de Fila

III. Modelagem do Sistema

Nesta seção, apresentamos o diagrama de transições de estados, algumas expressões matemáticas relacionadas às estatísticas de interesse no modelo M/M/2/J, bem como informações detalhadas sobre os componentes do sistema, fluxogramas de chegada e partida, e as metodologias de análise adotadas. A Tabela I mostra, simbolicamente, os parâmetros de gerais de interesse para as análises.

TABLE I: Simbologia de parâmetros estatísticos segundo a teoria de filas.

Parâmetros gerais de interesse	
Taxa de Chegada de clientes	E{n}=λ
Taxa de Partida de clientes	μ
Tempo médio de Serviço	E{Ts}=1/μ
Número médio de Clientes na fila	E{w}
Tempo médio na fila	E{Tw}
Número médio de clientes no sistema	E{q}
Tempo médio de permanência no sistema	E{Tq}
Fator de Utilização	ρ = λ / μ

A. Diagrama de Transições de Estados

O diagrama de estados que modela o sistema é mostrado na Figura 2. A cada instante, o estado atual muda para outro sempre que um pacote chega ou quando um pacote parte. Cada estado representa o número de elementos no sistema naquele instante. Como descrito anteriormente, λ é a taxa de chegada de clientes, $\mu 1$ é a taxa de partida do servidor *S1*, e $\mu 2$ é a taxa de partida do servidor *S2*. O tamanho do buffer é *J*.

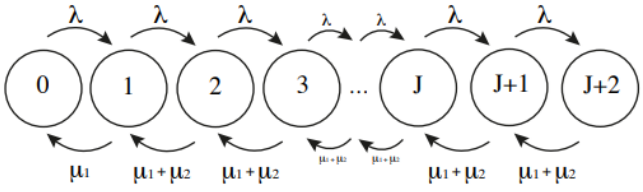


Fig. 3. Modelo do Sistema de Fila

B. Fluxogramas de Chegada e Partida

Nesta subseção, apresentamos o modelamento de diferentes blocos que permitem realizar os eventos de chegada e partida. As figuras 3 e 4 ilustram os fluxogramas básicos de chegadas e partidas de acordo com o código apresentado no Apêndice A. O fluxograma de chegada começa com a verificação se o tempo da próxima chegada (tc) é menor que o tempo das próximas partidas (tp1 e tp2). Se o tempo de chegada for menor, então o sistema atualiza o tempo atual (t) para o tempo de chegada (tc), incrementa o contador de pacotes (n), armazena o tempo de chegada e gera um novo tempo de chegada (tc). Se o servidor S1 estiver livre, o pacote é atendido por ele, gerando um novo tempo de partida tp1. Caso contrário, verifica-se se o servidor S2 está disponível. Se S2 estiver livre, o pacote é atendido por S2, gerando um novo tempo de partida tp2. Se ambos os servidores estiverem ocupados e houver espaço no buffer, o pacote é adicionado à fila. Se o buffer estiver cheio, o pacote é bloqueado, incrementando o contador de bloqueios (Block).

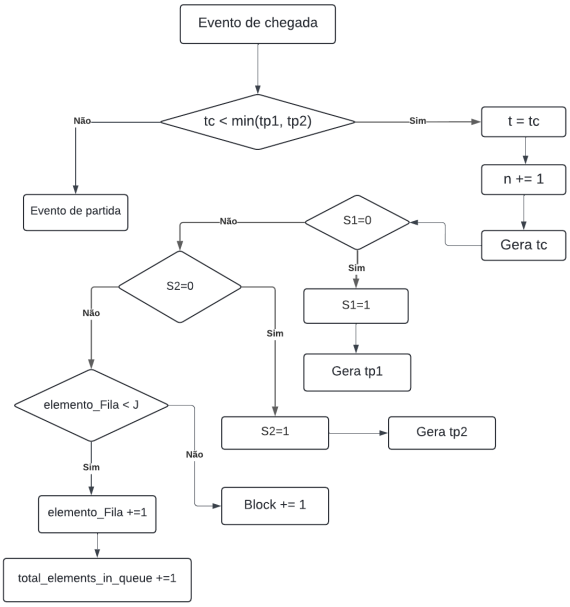


Fig. 3. Fluxograma de Chegada

O fluxograma de partida como ilustra a figura 4, inicia com a verificação se o tempo de partida do servidor S1 (tp1) é menor que o tempo de partida do servidor S2 (tp2). Se o tempo do servidor S1 for menor que o tempo do servidor S2, então o sistema atualiza o tempo atual (t) para o tempo da próxima partida no servidor S1 (tp1), incrementa o contador de partidas (k), e calcula o tempo no sistema (tq). Se houver pacotes na fila, um pacote é removido da fila e um novo tempo de partida é gerado para S1 (tp1). Se a fila estiver vazia, o servidor S1 é liberado e o tempo de partida é definido como infinito. Se o tempo do S2 for menor que o tempo do S1 então, o tempo atual é atualizado para tp2, o contador de partidas (k) é incrementado, e o tempo no sistema é calculado. Se houver pacotes na fila, um pacote é removido e um novo tempo de partida é gerado para S2 (tp2). Se a fila estiver vazia, o servidor S2 é liberado e o tempo de partida é definido como

infinito.

Em ambos os fluxogramas, após cada evento de chegada ou partida, o algoritmo coleta as estatísticas de interesse, atualiza o tempo de simulação (t) para o horário do próximo evento iminente e procede para o próximo evento. Este procedimento é repetido até que o número de partidas (k) atinja um valor predefinido de término da simulação.

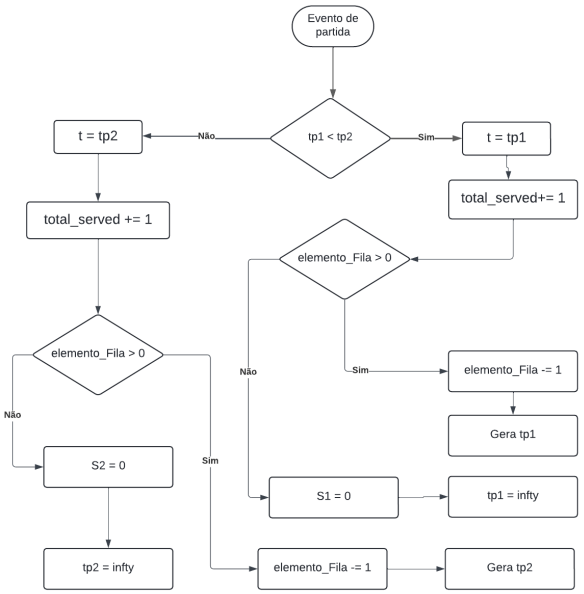


Fig. 4. Fluxograma de Partida

A. Expressões Matemáticas

Nesta subseção, descrevemos algumas expressões matemáticas utilizadas para calcular as estatísticas de interesse, como a probabilidade de cada estado, número médio de elementos no sistema, probabilidade de bloqueio e tempo médio no sistema são apresentadas. Estas fórmulas permitem uma análise analítica do desempenho do sistema.

A probabilidade de cada estado, pode ser escrita por:

$$\begin{cases} P_1 = \frac{\lambda}{\mu_1} \; ; \qquad Para\ (k=1) \\ P_k = \frac{(\frac{\lambda}{\mu_1+\mu_2})}{k!} \; ; \; Para\ (2 \geq k \geq J+2) \end{cases} \tag{1}$$

Onde, P_0 é a probabilidade do sistema estar vazio. A soma de as probabilidades dentro do sistema é igual a 1, e pode ser expresso por:

$$\sum_{k=0}^{1} P_k + \sum_{k=2}^{J+2} P_k = 1 \tag{2}$$

$$P_0 = \left(\sum_{K=0}^{1} \frac{(\lambda/\mu_1)^K}{K!} + \sum_{K=2}^{J+2} \frac{(\frac{\lambda}{\mu_1+\mu_2})^K}{K!} \right)^{-1} \tag{3}$$

B. Métricas de Desempenho

Para avaliar o desempenho do sistema, consideramos as seguintes métricas:

Probabilidade de bloqueio, P_b é a probabilidade de que um cliente que chega ao sistema encontre o buffer cheio e seja bloqueado, e pode ser obtido por:

$$P_b = P_{J+2} = \frac{\left(\frac{\lambda}{\mu_1+\mu_2}\right)^{J+2}}{(J+2)!} P_0 \tag{4}$$

Número médio de elementos no sistema *E[q],* é a quantidade média de clientes no sistema, incluindo aqueles em atendimento e na fila. Pode ser expressa por:

$$E[q] = \sum_{k=0}^{1} k \cdot P_k + \sum_{k=2}^{J+2} k \cdot P_k \tag{5}$$

Tempo médio no sistema $E[T_q]$, é o tempo total que um cliente passa no sistema, incluindo espera e atendimento. É dado por:

$$E[T_q] = \frac{E[q]}{\lambda \cdot (1 - P_b)} \tag{6}$$

IV. Resultados Numéricos

Nesta seção, apresentamos os resultados obtidos a partir das simulações realizadas para diferentes configurações de taxa de chegada λ e tamanhos de buffer J. Avaliamos métricas de desempenho como a probabilidade de bloqueio, o número médio de elementos no sistema *E[q]* e o tempo médio no sistema *E[Tq]*. Toda simulação foi implementada utilizando a linguagem python no ambiente de desenvolvimento colab. Os parâmetros utilizados nas simulações estão detalhados na Tabela II.

TABLE I: Simbologia de parâmetros estatísticos segundo a teoria de filas.

Parâmetro	Símbolo	Valor
Tamanho do Buffer	J	1 à 10
Taxa de Chegada	λ	10 e 20 Pacotes /s
Taxa de Partida no S1	μ_1	20; 10 Pacotes /s
Taxa de Partida no S2	μ_2	10; 8 Pacotes /s

As Figuras 5, 6 e 7 mostram a probabilidade de bloqueio, o número médio elementos no Sistema e o tempo médio no sistema em função do tamanho do buffer para três configurações: 1) λ=10, $\mu_1 = \mu_2$ =10; 2) λ=10, μ_1 =10, μ_2 =8; 3) λ=20, $\mu_1 = \mu_2$ =10;

Observa-se que na configuração 1, a probabilidade de bloqueio é baixa para todos os tamanhos de buffer e diminui à medida que o buffer aumenta. Isso indica que o sistema raramente enfrenta situações em que ambas as filas estão cheias, graças à capacidade de serviço adequada. Na configuração 2, o bloqueio é ligeiramente maior do que na Configuração 1 para buffers pequenos, devido à menor taxa de serviço do servidor 2. No entanto, a probabilidade de bloqueio

ainda é baixa e diminui significativamente com o aumento do buffer. Por lado, a configuração 3, A probabilidade de bloqueio é inicialmente alta devido à alta taxa de chegada, mas diminui à medida que o tamanho do buffer aumenta. Isso mostra que aumentar o buffer pode compensar parcialmente a alta taxa de chegada, reduzindo a frequência de bloqueios.

Como ilustra a figura 6, na configuração 1, O número médio de elementos no sistema aumenta levemente com o aumento do buffer, mas permanece baixo, indicando que a capacidade de serviço é adequada para a taxa de chegada de mensagens. No entanto, na configuração 3, cresce linearmente com o aumento do buffer. A alta taxa de chegada resulta em um acúmulo maior de mensagens no sistema, evidenciando a necessidade de um buffer maior para evitar bloqueios frequentes.

Finalmente, na figura 7, observa-se que o tempo médio no Sistema, na configuração 1 cresce levemente à medida que o tamanho do buffer aumenta, mas permanece relativamente baixo. Isso indica que os servidores são capazes de processar as mensagens recebidas de forma eficiente, mantendo o sistema com pouca espera. Por tanto, na configuração 2 é um pouco maior que na Configuração 1, especialmente para buffers pequenos. Isso se deve à menor taxa de serviço do servidor 2 (μ2=8). À medida que o buffer aumenta, o tempo médio no sistema se estabiliza, indicando que o impacto da menor taxa de serviço do servidor 2 é amortecido pelo maior buffer. Já na configuração 3, aumenta significativamente com o aumento do buffer. A alta taxa de chegada (λ=20\lambda = 20λ=20) sobrecarrega os servidores, resultando em mais mensagens acumuladas no sistema, o que eleva o tempo de espera.

Os resultados demonstram que a configuração dos parâmetros λ, μ1 e μ2 tem um impacto significativo no desempenho do sistema. Para uma taxa de chegada moderada λ=10, o sistema mantém um bom desempenho com baixos tempos médios e números médios de elementos no sistema, além de uma baixa probabilidade de bloqueio, mesmo com um buffer menor. No entanto, para uma alta taxa de chegada λ=20, o sistema enfrenta maiores desafios, com tempos médios e números médios de elementos no sistema significativamente mais altos.

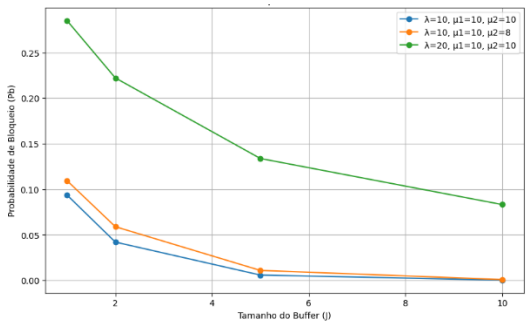


Fig. 5. Probabilidade de Bloqueio.

V. Aplicações

Os sistemas de filas com dois servidores e buffer finito (M/M/2/J) têm uma ampla gama de aplicações práticas. Um exemplo claro é na gestão de filas em supermercados e lojas de varejo. Quando há poucos clientes na fila, o atendimento pode ser realizado por um único caixa

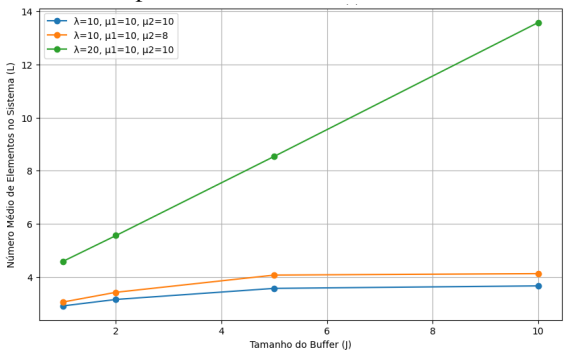


Fig. 6. Número médio de Elementos no sistema

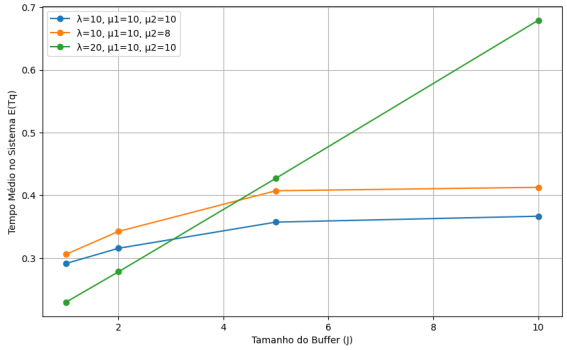


Fig. 7. Tempo médio no sistema de com μ1=10 e μ2=8

No entanto, à medida que a fila cresce, um segundo caixa pode ser aberto para agilizar o atendimento. Isso reduz o tempo de espera dos clientes, melhora a eficiência do serviço e evita que os clientes desistam da compra devido a longas filas. A análise e otimização desse processo garantem uma melhor experiência de compra e maior satisfação do cliente.

Outro exemplo de aplicação é nos aeroportos, especificamente na fila de check-in. Inicialmente, o atendimento pode ser feito por um único agente. Contudo, se o número de passageiros na fila aumentar, um segundo agente pode ser alocado para ajudar no processo de check-in. Isso não só reduz o tempo de espera dos passageiros, mas também melhora o fluxo de pessoas no aeroporto, evitando congestionamentos e garantindo que os passageiros embarquem a tempo. A aplicação do modelo M/M/2/J neste contexto é crucial para manter a eficiência e a satisfação dos passageiros em momentos de alta demanda.

VI. Conclusão

Neste trabalho, analisamos o desempenho de um sistema de fila com dois servidores (S1 e S2) e um buffer de tamanho finito J. A chegada das mensagens segue uma distribuição de Poisson com média λ, e as durações de serviço são exponencialmente distribuídas com médias 1/μ1 e 1/μ2. A análise comparativa foi realizada para diferentes tamanhos de buffer e distribuições para os tempos de serviço, considerando três configurações distintas de parâmetros.

Os Resultados das simulações revelaram importantes insights sobre o comportamento do sistema:

1. **Impacto da Taxa de Chegada λ\ e das Taxas de Serviço μ1, μ2**:

Quando a taxa de chegada λ é igual às taxas de serviço dos servidores μ1=μ2=10, o sistema demonstra um desempenho eficiente com baixos tempos médios no sistema e números médios de elementos no sistema. A probabilidade de bloqueio é mínima e diminui ainda mais com o aumento do tamanho do buffer.

Para uma configuração com λ=10 e μ2 reduzida para 8, o desempenho do sistema é levemente afetado. O tempo médio no sistema e o número médio de elementos aumentam um pouco, mas a probabilidade de bloqueio continua baixa, mostrando que o sistema ainda é capaz de lidar com a carga de trabalho, embora com menor eficiência.

Com uma alta taxa de chegada λ=20, mesmo quando as taxas de serviço são iguais a 10, o sistema enfrenta maiores desafios. O tempo médio no sistema e o número médio de elementos aumentam significativamente, indicando sobrecarga. A probabilidade de bloqueio é maior, mas pode ser mitigada com um aumento no tamanho do buffer.

2. **Influência do Tamanho do Buffer J**:

O aumento do tamanho do buffer contribui para a redução da probabilidade de bloqueio em todas as configurações. Isso mostra a importância de dimensionar adequadamente o buffer para acomodar picos na chegada de mensagens e evitar congestionamentos.

Embora o aumento do buffer reduza a probabilidade de bloqueio, o tempo médio no sistema e o número médio de elementos no sistema aumentam proporcionalmente com o tamanho do buffer, especialmente em sistemas com alta taxa de chegada. Isso indica um trade-off entre minimizar bloqueios e controlar os tempos de espera no sistema.

3. **Comparação de Desempenho entre Configurações**:

As configurações com λ=10\lambda = 10λ=10 apresentam um desempenho bastante robusto, com tempos médios e números médios de elementos no sistema relativamente baixos, e uma probabilidade de bloqueio que diminui rapidamente com o aumento do buffer.

A configuração com λ=20\lambda = 20λ=20 destaca a importância de ajustar as taxas de serviço e o tamanho do buffer para lidar com altas cargas de trabalho. Mesmo com taxas de serviço adequadas, a alta taxa de chegada demanda um buffer maior para manter o desempenho aceitável.

Este estudo sublinha a importância de uma cuidadosa configuração dos parâmetros de chegada e serviço, bem como do dimensionamento adequado do buffer, para garantir um desempenho eficiente de sistemas de fila com múltiplos servidores. Os resultados obtidos podem servir como guia para projetar sistemas que equilibram a necessidade de minimizar a probabilidade de bloqueio com a necessidade de manter baixos tempos de espera e números médios de elementos no sistema.

Futuras investigações poderiam explorar a variabilidade das distribuições de serviço e diferentes políticas de encaminhamento de mensagens, para aprofundar a compreensão do impacto de tais fatores no desempenho do sistema. Além disso, a implementação de algoritmos de controle adaptativo que ajustem dinamicamente as taxas de serviço e o tamanho do buffer com base na carga de trabalho poderia ser uma área promissora para melhorar ainda mais a eficiência do sistema.

ANEXO

https://colab.research.google.com/drive/1KXO_X1LPlcQiEwD8lpIzp-r8b1t0sr_5?usp=sharing.

AGRADECIMENTOS

Gostaríamos de agradecer ao Professor Samuel Bararldi Mafra pelas orientações prestadas na elaboração das simulações e todo conhecimento que nos passou que serviu de uma base sólida para confecção deste relatório.

REFERENCES

[1] D. Gross, “Fundamentals of queueing theory,” John Wiley & Sons, 2008.
[2] . B. K. Kumar and S. P. Madheswari, “An M/M/2 queueing system with heterogeneous servers and multiple vacations,” Mathematical and Computer Modelling, vol. 41, no. 13, pp. 1415-1429, 2005, doi: 10.1016/j.mcm.2004.09.004.
[3] M. Zukerman, “Introduction to queueing theory and stochastic teletraffic models,” arXiv preprint arXiv:1307.2968, 2013, doi: 10.48550/arXiv.1307.2968.
[4] L. Kleinrock, “Queueing Systems, Volume 1: Theory,” Wiley-Interscience, 1975.
[5] H. A. Taha, “Operations Research: An Introduction,” Prentice Hall, 2011.
[6] D. Bertsekas and R. Gallager, “Data Networks,” Prentice Hall, 1992.