Teoria de Filas - Estudo de caso: $M/M/2$

Eliseu Elias Cândido Moreira, Emmanuel Priestley Titus e Jonas Vilasboas Moreira

***Abstract*—O presente artigo aborda a análise de um sistema de filas M/M/2, onde dois servidores atendem a duas filas de chegadas Poissonianas com tempos de serviço exponencialmente distribuídos e idênticos. Um exemplo prático deste tipo de fila pode ser observado em um serviço de atendimento ao cliente com dois atendentes, onde clientes chegam aleatoriamente e o tempo de atendimento de cada cliente segue uma distribuição exponencial. A análise compara o desempenho do sistema M/M/2 com um sistema de dois servidores operando com uma fila única. Os resultados indicam que o sistema M/M/2, com duas filas alimentando dois servidores, apresenta maior eficiência em termos de tempo médio de espera e taxa de utilização dos servidores, destacando a superioridade do modelo M/M/2 na otimização do atendimento e redução de tempos de espera em cenários com características similares.**

***Index Terms*—Teoria de filas, teoria das esperas.**

I. Introdução

A Teoria de filas, também conhecida como teoria das esperas, é um campo da matemática aplicada que estuda o comportamento de filas em diversos tipos de sistemas. De modo geral, sistemas que empregam teoria de filas tem um funcionamento simples: clientes chegam a um servidor e podem ou não ser atendidos, a depender do estado de ocupação do servidor. Se estiver livre, os clientes são prontamente atendidos, mas se o servidor estiver ocupado, os clientes aguardam para serem atendidos em uma fila. Esta teoria ajuda a analisar e otimizar o desempenho de sistemas em várias áreas, como telecomunicações, computação, manufatura e serviços públicos, ao prever tempos de espera, comprimento das filas e outros parâmetros importantes [1].

A teoria de filas é essencial para a otimização de diversos sistemas operacionais e de serviços, pois permite a análise detalhada de processos onde há demanda por recursos limitados. Por exemplo, em um *call center*, a teoria de filas pode ser utilizada para determinar o número ideal de atendentes necessários para minimizar o tempo de espera dos clientes, garantindo um atendimento eficiente e reduzindo custos operacionais. Da mesma maneira, em hospitais, pode-se usar a teoria de filas para gerenciar a capacidade de leitos e salas de emergência, melhorando o atendimento ao paciente e otimizando o uso dos recursos hospitalares.

Existem diferentes modelos de filas, cada um adequado para situações específicas. Um dos mais comuns é o modelo $M/M/1$, que assume chegadas de clientes segundo um processo de *Poisson*, tempos de serviço exponenciais, e um único servidor. Outro modelo é o $M/M/c$, que é uma generalização do $M/M/1$ com múltiplos servidores. Há também o modelo $M/G/1$, onde o tempo de serviço segue uma distribuição geral. Além desses, existem modelos mais complexos como filas em *tandem* e redes de filas, que permitem a análise de sistemas com múltiplas etapas ou servidores interdependentes [2].

Na prática, a teoria de filas é aplicada em diversas indústrias. Em telecomunicações, é usada para dimensionar a capacidade de redes e prever congestionamentos, garantindo a qualidade do serviço. Na área de transportes, ajuda a modelar e otimizar o fluxo de veículos em estradas e sistemas de transporte público, reduzindo tempos de espera e melhorando a eficiência. Em serviços bancários, a teoria de filas auxilia no gerenciamento de filas de atendimento e caixas eletrônicos, proporcionando um melhor atendimento ao cliente. Esses exemplos ilustram a versatilidade e a importância da teoria de filas na melhoria de processos e na tomada de decisões estratégicas em diferentes setores.

Neste trabalho, será apresentado um estudo de caso da fila $M/M/2$, onde clientes chegam segundo um processo de *Poisson*, com tempo de serviço exponencial, e dois servidores.

II. Desenvolvimento

Para o estudo de caso proposto para este trabalho, apresenta-se o seguinte problema:

"Suponha que as chegadas sejam distribuídas de acordo com um processo de *Poisson* da taxa λ e, quando os clientes chegam, estes selecionam a fila mais curta; se um cliente chegar e encontrar as duas filas com o mesmo número de clientes em espera, ele precisa selecionar sua própria linha de espera aleatoriamente (considere uma escolha com distribuição de *Bernoulli* de média 0.5). Os tempo de serviço seguem uma distribuição genérica de média $1/\mu$."

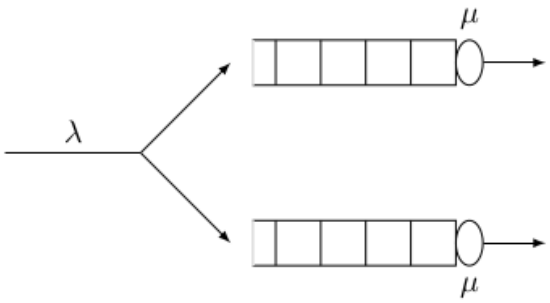
O principal objetivo do estudo de caso proposto é entender o funcionamento do sistema em questão e compará-lo com outros métodos de filas, julgando eficiência, custo benefício e complexidade de implementação.

Inicialmente o sistema descrito possui chegadas de clientes distribuídas segundo um processo de *Poisson* com taxa λ, onde os clientes escolhem a fila mais curta ao chegar. Se as filas tiverem o mesmo número de clientes, a escolha é feita aleatoriamente com probabilidade igual seguindo uma distribuição de *Bernoulli* de média 0.5. Os tempos de serviço seguem uma distribuição genérica com média $1/\mu$.

Este sistema de filas, pode ser exemplificado por uma rede de filas com roteamento dependente do estado, e pode ser classificado como um sistema de filas com múltiplos servidores, também conhecido como um sistema $M/M/2$. Deste modo, o mesmo pode ser visualizado como mostrado na Figura 1:

Desta maneira, observa-se os principais pontos:

Fig. 1. Modelo de fila para o problema apresentado



- As chegadas dos clientes são modeladas por um processo de Poisson com taxa λ;
- Os tempos de serviço seguem uma distribuição genérica com média $1/\mu$;
- Os clientes escolhem a fila mais curta ao chegar. Se as filas têm o mesmo comprimento, a escolha é feita de forma aleatória, representada por uma distribuição de Bernoulli com média 0.5;
- Caso haja saída e uma das filas tenha um comprimento menor, usuários já ingressos em outra fila não podem alterar de fila;

A análise e otimização desse tipo de sistema requerem métodos mais avançados da teoria de filas, considerando tanto a distribuição dos tempos de serviço quanto as políticas de roteamento.

Trazendo para este trabalho o contexto da notação de Kendall [3], utilizada para classificar diferentes tipos de fila, denotada por A/B/c/K/m/Z de tal forma que:

- **A**: descreve a distribuição das chegadas dos clientes. Usa-se M para representar processo chegada *Poissoniano*;
- **B:** descreve a distribuição do tempo de serviço;
- **c:** descreve o número de servidores;
- **K:** descreve a capacidade do sistema, ou seja, quantos clientes podem ser atendidos pelo sistema;
- **m:** tamanho da população;
- **Z:** descreve a disciplina de tratamento da fila, por exemplo, *First In, First Out* (FIFO). Neste caso em específico, a denotação pode ser ocultada;

Para o sistem M/M/2, tem-se o seguinte equacionamento [2], [4]:

Probabilidade de nenhum cliente no sistema:

$$P_0 = \left(\sum_{n=0}^{C-1} \left(\frac{n^n}{n!} \right) + \frac{c \cdot r^c}{c!(c-r)} \right)^{-1} \tag{1}$$

No qual:

$$r = \frac{\lambda}{\mu} \tag{2}$$

Probabilidade de n clientes no sistema:

$$P_n = P_0 \left(\frac{r^n}{n!} \right), \ Se \ 1 \leq n \leq c \tag{3}$$

e:

$$P_n = P_0 \left(\frac{r^n}{(c^{n-c}c!)} \right), \ Se \ n \geq c \tag{4}$$

Probabilidade média de clientes no sistema:

$$L = r + \left(\frac{r^{C+1} \cdot c}{c!(c-r)^2} \right) \cdot P_0 \tag{5}$$

Probabilidade média de clientes na fila:

$$L_q = \left(\frac{P_0 c r^{c+1}}{c!(c-r)^2} \right) \tag{6}$$

Tempo médio de espera do sistema:

$$w = \frac{1}{\mu} + \left(\frac{r^c \mu}{(c-1)!(c\mu-1)^2} \right) P_0 \tag{7}$$

Tempo médio de espera na fila:

$$w_q = \left(\frac{r^c \mu}{(c-1)!(c\mu-1)^2} \right) P_0 \tag{8}$$

Nesas configurações é possível realizar o estudo de caso de maneira mais abrangente, alterando parâmetros afim de se obter diferentes resultados.

Para uma experiência mais segura e precisa, implementou-se o sistema por meio de Python, realizando diversas simulações com diferentes parâmetros.

Analisando a Figura 1, é possível notar que as entidades são todas aquelas que estão submetidas aos eventos ocorridos, sendo eles a chegada, acomodação em sua respectiva fila, tempo de espera, tempo de serviço e por fim a saída [5]. Além disso observa-se também com o equacionamento as principais variáveis de estado do sistema, como a taxa de chegada (λ), a distribuição genérica de média $1/\mu$, probabilidade média de tempo e clientes tanto em filas quanto no sistema.

III. Conclusão

Para analisar o desempenho da fila $M/M/2$ foram realizadas uma série de simulações, com diferentes parâmetros, a fim de observar as diferenças de comportamento da fila, para parâmetros diversos.

A tabela I apresenta os resultados dos testes para o modelo $M/M/2$ com duas filas e tabela II apresenta os resultados dos testes para o modelo $M/M/2$ com uma fila.

Pelo código implementado, é possível notar como a fila estudada pode influenciar de acordo com os parâmetros, como visto na Tabala I e na Tabela II. Em comparação com outro tipo de fila que é parecido, ao alterar o fator de utilização ρ, os parâmetro mais importante que sofrem as consequências são justamente o tempo de espera, tanto na fila quanto o tempo de espera do serviço, alterando também o tempo médio total. Observa-se também o número médio de pacotes, nos quais são alterados tanto para uma fila quanto para duas filas, cujo valor com duas filas é menor do que a metade, visto que são divididos por 2 filas, consequentemente diminuindo também o número médio de pacotes no sistema completo. Desta forma, conclui-se que a eficiência utilizando duas filas e dois servidores, é consideravelmente maior caso fosse utilizado apenas uma fila e um servidor de saída.

TABLE I
TABELA RESULTADOS 2 SERVIDORES E 2 FILAS

Fator de Uso 1	Fator de Uso 2	Tempo no Sistema [ms]	Tempo na Fila [ms]	N. Médio de Pacotes no Sistema	N. Médio Pacotes na Fila
0.6667	0.6667	334.59 2	334.56	6691.74	6691.07
0.6667	0.333	183.36 2	183.53	3667.36	3666.69
0.333	0.6667	188.85 2	188.84	3777.16	3776.82
1.333	1.333	831.79 2	831.72	16635.88	16634.55
0.333	0.333	107.64 2	107.63	2152.94	2152.60

TABLE II
TABELA RESULTADOS 2 SERVIDORES E 1 FILA

Fator de Uso 1	Tempo no Sistema [ms]	Tempo na Fila [ms]	N. Médio de Pacotes no Sistema	N. Médio Pacotes na Fila
0.6667	1000.01	999.98	20000.36	19999.69
0.333	629.35	629.33	12587.05	12586.72
1.333	1421.25	1421.18	28425.05	28423.72
0.217	446.37	446.36	8927.47	8927.25

APPENDIX A
TABELAS DE RESULTADOS

APPENDIX B
GITHUB

https://github.com/LzuElias/MTEL-TP547.git

REFERENCES

[1] M. C. F. de Sinay, "Modelagens de filas a partir de diagramas de fluxo," *XXXVI - SBPO. O Impacto da Pesquisa Operacional nas Novas Tendências Multidisciplinares*, 11 2004.

[2] L. Kleinrock, *Queueing Systems: Theory*, ser. A Wiley-Interscience publication. Wiley, 1974. [Online]. Available: https://books.google.com.br/books?id=Q2ZRAAAAMAAJ

[3] T. Robertazzi, *Computer Networks and Systems: Queueing Theory and Performance Evaluation*. Springer New York, 2012. [Online]. Available: https://books.google.com.br/books?id=99DTBwAAQBAJ

[4] L. Kleinrock, *Queueing Systems, Volume 2: Computer Applications*. Wiley, 1976. [Online]. Available: https://books.google.com.br/books?id=$_2ZRAAAAMAAJ$

[5] G. Dattatreya, *Performance Analysis of Queuing and Computer Networks*, ser. Chapman and Hall/CRC Computer and Information Science Series. Taylor & Francis Limited (Sales), 2019. [Online]. Available: https://books.google.com.br/books?id=arShyAEACAAJ