



UNIVERSIDADE FEDERAL DO ABC

TRABALHO DE GRADUAÇÃO

Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas (CECS)

**“Estudo Observacional de Desempenho de Filas Aplicada a Postos de
Combustível”**

NOME: Heitor Ferreira Kulcsar

RA: 11201720062

Nome do professor orientador: Profº. Dr. Marco Aurélio Cazarotto Gomes

Santo André

2026

RESUMO

Sistemas de atendimento convivem com variações na demanda e limitações de capacidade, o que frequentemente resulta na formação de filas, aumento do tempo de espera e redução da eficiência operacional. Nesse contexto, a Teoria de Filas oferece ferramentas matemáticas capazes de descrever e analisar o comportamento desses sistemas, auxiliando no dimensionamento adequado de recursos.

Este trabalho teve como objetivo aplicar os fundamentos da Teoria de Filas na análise de desempenho e no dimensionamento da capacidade de um sistema de atendimento. A metodologia combinou modelagem analítica e simulação computacional em Python, permitindo calcular métricas como taxa média de chegada (λ), taxa média de atendimento (μ), número de servidores (S), taxa de utilização (ρ), tempo médio de espera, probabilidade de bloqueio ($P_{\square}$) e vazão efetiva.

Como aplicação, foi analisado o atendimento de um posto de combustível localizado no bairro do Sacomã, em São Paulo, utilizado como ambiente real para coleta de dados e validação dos modelos teóricos. Foram comparadas diferentes arquiteturas de filas para o sistema de abastecimento, incluindo uma fila única com seis servidores ($M/M/6$), arranjos intermediários com três filas e dois servidores cada ($3 \times M/M/2$) e filas independentes com um servidor por bomba ($6 \times M/M/1$).

Os resultados mostraram que o modelo $M/M/6$ apresentou maior vazão efetiva e menor probabilidade de bloqueio, porém com maior tempo médio de permanência no sistema. O arranjo $3 \times M/M/2$ apresentou desempenho intermediário, enquanto o modelo $6 \times M/M/1$ apresentou menor tempo médio para os clientes atendidos, mas maior probabilidade de bloqueio e maior variabilidade entre filas.

O estudo indicou que a configuração com quatro filas $M/M/2$ para o abastecimento e uma fila $M/M/3$ para o pagamento representa uma alternativa mais equilibrada entre capacidade operacional, tempo de atendimento e retorno econômico. Conclui-se que a integração entre modelagem matemática e a linguagem de programação constitui uma abordagem eficaz para apoiar decisões de dimensionamento e organização de sistemas de atendimento.

Palavras-chave: Teoria de Filas; Sistemas de Atendimento; Dimensionamento de Capacidade; Modelos $M/M/S/K$; Simulação em Python.

Área do conhecimento: Teoria de filas, Engenharia da informação e ciência da computação.

SUMÁRIO

1. INTRODUÇÃO	5
1.1. OBJETIVO DE ESTUDO	6
1.2. OBJETIVOS ESPECÍFICOS	7
1.3. JUSTIFICATIVA	8
2. TEORIA DE FILAS	9
2.1 SISTEMAS DE FILAS	9
2.2 SERVIDORES	10
2.3 NOTAÇÃO DE KENDALL	11
2.4 M/M/1	13
2.5 M/M/S/K	13
3. APLICAÇÃO EM SISTEMAS DE SERVIÇOS	16
3.1 PROCESSO DE ATENDIMENTO NO POSTO DE COMBUSTÍVEL	16
3.2 ETAPAS DO ATENDIMENTO	17
3.3 FLUXOGRAMA DAS ETAPAS DO ATENDIMENTO	17
3.4 ESTUDO DE CASO	18
3.5 PROGRAMAÇÃO PYTHON PARA ANÁLISES DE SIMULAÇÕES	21
3.6 CÓDIGO E SIMULAÇÕES	21
3.7 COMPARAÇÃO ENTRE RESULTADOS TEÓRICOS E SIMULADOS	24
4. RESULTADOS	26
4.1 CHEGADA DOS CLIENTES	27
4.2 ATENDIMENTO DOS CLIENTES	31
4.3 CÁLCULOS DE DESEMPENHO E SIMULAÇÕES DA FILA	38
4.3.1 PARÂMETROS E CONSIDERAÇÕES PRÁTICAS	39
4.3.2 ANÁLISE ECONÔMICA E CENÁRIOS DE EXPANSÃO	40
4.3.3 INTERPRETAÇÃO DOS RESULTADOS	47
4.3.4 PROPOSTA DE SISTEMA	49
4.3.5 DESEMPENHO ESPERADO DO SISTEMA PROPOSTO	50
4.3.6 AVALIAÇÃO ECONÔMICA	52
4.4 SÍNTESE E RECOMENDAÇÃO FINAL	54
5. DISCUSSÃO DOS RESULTADOS	56
6. CONCLUSÃO	58
REFERÊNCIAS	60
APÊNDICES	63

1. INTRODUÇÃO

Nos sistemas de serviços, a formação de filas é um problema recorrente e pode comprometer significativamente a qualidade do atendimento. Longos tempos de espera afetam a eficiência operacional, a satisfação dos usuários e, conseqüentemente, a competitividade das organizações, especialmente em ambientes onde a demanda é variável e a capacidade de atendimento é limitada. Estudos indicam que filas mal dimensionadas ou mal gerenciadas tendem a gerar perda de desempenho, aumento de custos e redução da percepção de valor por parte dos clientes (SILVA; SOUSA, 2020).

Nesse contexto, a Teoria de Filas oferece um conjunto de modelos matemáticos e estatísticos capazes de descrever e analisar o comportamento desses sistemas. Por meio desses modelos, é possível compreender a dinâmica estocástica das chegadas e dos atendimentos, estimar métricas de desempenho e apoiar decisões relacionadas à alocação de recursos e à melhoria do nível de serviço.

O objetivo deste trabalho é analisar o desempenho de diferentes configurações de sistemas de filas aplicáveis a um ambiente de serviço, utilizando modelos clássicos da literatura, como $M/M/1$, $M/M/s$ e $M/M/s/K$, para avaliar cenários operacionais e propor alternativas de otimização.

A linguagem Python foi utilizada como ferramenta de apoio às análises e simulações, permitindo o cálculo das principais métricas, a realização de testes paramétricos e a comparação entre diferentes configurações de atendimento.

Como aplicação prática, foi analisado o atendimento em um posto de combustível localizado no bairro do Sacomã, em São Paulo. O posto foi utilizado apenas como exemplo real para aplicação dos modelos teóricos, servindo como base para coleta de dados e validação dos métodos empregados, sem constituir o foco central do estudo.

A partir dessa aplicação, foi possível identificar gargalos, avaliar o desempenho do sistema e demonstrar, de forma concreta, como os modelos de filas podem apoiar a tomada de decisão em contextos reais.

Para apresentar essa construção de forma organizada, partindo da fundamentação teórica até a análise prática e as conclusões, o trabalho foi estruturado em seis capítulos, descritos na sequência.

O “**Capítulo 2 – Teoria de Filas**” apresenta a base conceitual que sustenta o estudo. Nele são discutidos os elementos que compõem um sistema de filas, como processos de chegada, servidores e capacidade, além da notação de Kendall e dos modelos M/M/1 e M/M/s/K, que servem como referência para as análises desenvolvidas posteriormente.

O **Capítulo 3 – “Estudo de Caso e Modelagem do Sistema”** descreve o processo de atendimento do posto de combustível utilizado como aplicação prática. São apresentadas as etapas do atendimento, o fluxograma do processo, os dados observados e a forma como o sistema foi modelado. Também é introduzido o uso da linguagem Python para a realização das simulações.

O **Capítulo 4 – “Análise e Simulações do Sistema”** concentra os cálculos de desempenho e os cenários avaliados. Neste capítulo são analisadas as taxas de chegada e atendimento, desenvolvidas simulações, comparadas configurações de filas e apresentada a proposta de sistema recomendada, acompanhada de avaliação econômica.

O **Capítulo 5 – “Discussão dos Resultados”** interpreta os resultados obtidos à luz da teoria apresentada, relacionando os achados do estudo de caso com os conceitos clássicos da Teoria de Filas.

Por fim, o “**Capítulo 6 – Conclusão**” reúne as principais conclusões do trabalho, destacando as contribuições da pesquisa e possíveis desdobramentos para estudos futuros.

1.1. OBJETIVO DE ESTUDO

Este trabalho tem como objetivo investigar e aplicar os fundamentos da Teoria de Filas em sistemas de atendimento do setor de serviços, com ênfase na análise de desempenho, no dimensionamento de capacidade e na avaliação dos impactos operacionais e econômicos decorrentes das decisões de alocação de recursos.

Além da abordagem analítica, o trabalho também utiliza simulação computacional como ferramenta complementar, com o objetivo de reproduzir o comportamento do sistema ao longo do tempo e validar os resultados obtidos pelos modelos teóricos.

Serão estudadas métricas clássicas como taxa média de chegada (λ), tempo médio de atendimento (μ), número de servidores (S), taxa de utilização, tempos

médios de espera e probabilidade de perda, permitindo caracterizar o comportamento estocástico do sistema e comparar diferentes configurações de atendimento.

Como aplicação prática, o estudo utiliza um posto de combustíveis como exemplo representativo de sistemas com restrição física de capacidade e sujeição à formação de filas. Esse estudo de caso tem caráter ilustrativo, servindo para validar os modelos teóricos, demonstrar sua aplicabilidade e evidenciar como alterações estruturais podem impactar indicadores de desempenho e resultados financeiros.

Dessa forma, o trabalho contribui para a construção de uma abordagem quantitativa de apoio à decisão, potencialmente extensível a diferentes tipos de serviços que enfrentam problemas similares de congestionamento, dimensionamento de recursos e otimização operacional.

1.2. OBJETIVOS ESPECÍFICOS

Os objetivos específicos deste trabalho consistem em:

- **Analisar e simular o desempenho de sistemas de filas em diferentes cenários operacionais**, utilizando programação em Python para as simulações das principais métricas (λ , μ , S , tempos médios de espera, taxa de utilização e probabilidade de perda).
- **Determinar configurações de capacidade capazes de atender à demanda com eficiência**, equilibrando nível de serviço, custos operacionais e restrições físicas do sistema.
- **Avaliar o impacto econômico das decisões de dimensionamento**, estimando ganhos ou perdas associados à variação da capacidade em sistemas de atendimento.
- **Aplicar a teoria de filas aos dados observados em um posto de combustível no bairro do Sacomã**, empregando-o como estudo de caso ilustrativo para validar o modelo proposto e demonstrar sua utilidade na identificação de gargalos e oportunidades de melhoria.
- **Demonstrar como a aplicação dos modelos teóricos pode melhorar a eficiência operacional e a utilização dos recursos**, fornecendo subsídios quantitativos para decisões gerenciais em diferentes tipos de serviços.

1.3. JUSTIFICATIVA

O gerenciamento inadequado de filas em sistemas de serviço figura entre os principais fatores de insatisfação dos usuários e perda de competitividade. Tempos excessivos de espera reduzem a qualidade percebida, prejudicam a experiência do cliente e podem ocasionar perdas operacionais e financeiras relevantes.

A Teoria de Filas, ao oferecer ferramentas matemáticas e estatísticas consolidadas, configura-se como abordagem fundamental para analisar o desempenho de sistemas de atendimento, identificar gargalos e propor configurações de capacidade mais eficientes.

A justificativa deste estudo baseia-se na crescente demanda por métodos quantitativos que auxiliem gestores a tomar decisões embasadas em dados, especialmente em ambientes sujeitos a alta variabilidade na demanda e limitações físicas de capacidade.

O estudo de caso aplicado a um posto de combustível no bairro do Sacomã, em São Paulo, tem caráter ilustrativo e demonstra como os modelos teóricos podem ser empregados para avaliar cenários reais de congestionamento e orientar ajustes operacionais. Entretanto, a relevância da pesquisa transcende esse caso específico, uma vez que a metodologia proposta apresenta potencial de aplicação em diversos segmentos do setor de serviços, como bancos, restaurantes, call centers e unidades de saúde, que enfrentam desafios semelhantes de dimensionamento e gestão de filas (GROSS; HARRIS, 2008; KLEINROCK, 1975).

2. TEORIA DE FILAS

As filas estão presentes em diversos contextos do cotidiano moderno, como postos de combustível, supermercados, restaurantes, hospitais, parques, estacionamentos e até sistemas computacionais. De modo geral, elas surgem quando a demanda por atendimento supera a capacidade disponível, formando uma sequência de pessoas, tarefas ou requisições que aguardam para serem atendidas (HILLIER; LIEBERMAN, 2013).

O estudo sistemático das filas teve início em 1908, com os trabalhos de Erlang (ERLANG, 1909). Ao analisar o congestionamento em sistemas telefônicos, Erlang desenvolveu modelos matemáticos capazes de estimar tempos médios de espera e probabilidades de bloqueio, sendo posteriormente reconhecido como o fundador da Teoria de Filas (PRADO, 2006). A partir dessas contribuições, a área evoluiu e passou a oferecer ferramentas para a análise de diferentes sistemas de atendimento.

Com o avanço tecnológico, muitos serviços passaram a adotar soluções digitais, como *check-in* antecipado, atendimento por *chatbot*, telemedicina e sistemas logísticos mais rápidos. Essas iniciativas buscam reduzir o tempo de espera e aumentar a eficiência. Ainda assim, a necessidade de dimensionamento adequado permanece, pois a demanda continua sendo variável e os recursos possuem capacidade limitada.

Nesse contexto, a Teoria de Filas fornece instrumentos matemáticos que auxiliam na definição do número de servidores, na capacidade da fila e na avaliação do desempenho do sistema. Para isso, considera-se que as chegadas e os tempos de serviço são variáveis aleatórias e que o funcionamento do sistema pode ser descrito por processos estocásticos, como os modelos de nascimento e morte (MARINS, 2011).

Nas seções seguintes, são apresentados os principais conceitos, métricas e modelos que fundamentam a teoria e orientam sua aplicação em sistemas de serviço.

2.1 SISTEMAS DE FILAS

Um sistema de filas pode ser entendido como uma estrutura formada por três elementos principais: clientes, servidores e uma regra de atendimento. Os usuários chegam ao sistema de acordo com um processo estocástico, aguardam caso todos

os servidores estejam ocupados e são atendidos conforme critérios previamente definidos, como ordem de chegada (FIFO, do inglês *First In First Out*), último a chegar primeiro a ser atendido (LIFO, do inglês *Last In Last Out*) ou ainda regras baseadas em prioridade (PINTO, 2011).

A fila pode ter capacidade ilimitada ou possuir um limite máximo de usuários. Quando existe essa limitação, a entrada de novos clientes depende do estado do sistema no momento da chegada, podendo ocorrer bloqueio ou perda, conforme a política adotada.

Existem diferentes formas de organizar um sistema de filas, entre elas:

- fila única com um servidor;
- fila única com múltiplos servidores em paralelo;
- múltiplas filas com servidores paralelos;
- fila única com servidores em série.

Cada estrutura apresenta características próprias de desempenho e exige modelagem adequada.

A eficiência de um sistema de filas pode ser avaliada por meio de métricas como o tempo médio de espera, o tempo médio de permanência no sistema, a taxa de atendimento, o nível de utilização dos servidores e a probabilidade de congestionamento. Esses indicadores permitem compreender como o sistema se comporta diante da demanda e se a capacidade disponível está adequada.

Nos modelos clássicos, costuma-se assumir que as chegadas seguem um processo de Poisson com taxa média λ e que os tempos de serviço obedecem a uma distribuição exponencial com média $1/\mu$. Essas hipóteses caracterizam os modelos Markovianos da família M/M/s e tornam possível a obtenção de expressões analíticas para estimar o número médio de usuários no sistema, os tempos médios de espera e o desempenho global da operação (BRANDÃO et al., 2012).

Segundo Oliveira (2006), quando aplicada de forma adequada, a Teoria de Filas contribui para a redução de atrasos, melhor aproveitamento dos recursos e apoio à tomada de decisão gerencial, especialmente em ambientes onde a qualidade do serviço depende do equilíbrio entre demanda e capacidade.

2.2 SERVIDORES

Em um sistema de filas, os servidores são os recursos responsáveis pela realização do atendimento ou pelo processamento da tarefa. Enquanto está sendo

atendida, a entidade permanece ocupando o servidor até a conclusão do serviço, momento em que o recurso é liberado para o próximo usuário (HILLIER; LIEBERMAN, 2013).

A estrutura dos servidores pode assumir diferentes configurações:

- Fila única para múltiplos servidores (pool de servidores): todos os clientes aguardam em uma única fila centralizada, sendo direcionados ao primeiro servidor disponível.
- Múltiplas filas paralelas: cada servidor possui sua própria fila, e os clientes escolhem ou são direcionados individualmente para cada uma delas.
- Servidores em série: o cliente passa por mais de um servidor em sequência, caracterizando um fluxo encadeado.

Cada uma dessas estruturas influencia o desempenho do sistema de maneira distinta, afetando o tempo de espera, o equilíbrio entre os servidores e o nível geral de eficiência. Por isso, a escolha do arranjo mais adequado depende das características da demanda, da capacidade disponível e da estratégia operacional adotada.

2.3 NOTAÇÃO DE KENDALL

A notação proposta por Kendall, em 1953, é amplamente utilizada para classificar sistemas de filas. Ela é representada pela forma geral $A/B/m/K/N/D$, em que cada elemento indica uma característica específica da estrutura do sistema (COSTA, 2020).

Os três primeiros termos, $A/B/m$, formam a base do modelo e descrevem o processo de chegada, a distribuição dos tempos de serviço e o número de servidores. Os demais termos acrescentam informações complementares, como a capacidade máxima do sistema, o tamanho da população de clientes e a disciplina de atendimento adotada.

A — Distribuição dos tempos de chegada

- M (exponencial): chegadas ocorrem segundo um processo Poisson, com tempos entre chegadas exponenciais.
- D (determinístico): os intervalos de chegada são constantes.

- E_k (k-Erlang): soma de k distribuições exponenciais, representando processos com múltiplas fases.
- H_k (mistura de exponenciais): combina diferentes taxas exponenciais.
- PH (fase): modelo geral composto por fases exponenciais.
- G (geral): qualquer distribuição arbitrária.

B — Distribuição dos tempos de atendimento

Segue a mesma classificação do item A, podendo ser M, D, E_k, H_k, PH ou G.

m — Número de servidores em paralelo

Ex.: $m = 1$ (um servidor); $m > 1$ (múltiplos servidores); $m \rightarrow \infty$ (capacidade teórica muito elevada).

K — Capacidade máxima do sistema

Inclui clientes em atendimento + fila.

Ex.: $K = 10$ (limite físico), $K = \infty$ (fila ilimitada).

N — População de clientes

Define o número total de possíveis usuários do sistema.

Ex.: N finito (manutenção de máquinas); $N = \infty$ (supermercado).

D — Disciplina de atendimento

- FCFS/FIFO: primeiro a chegar, primeiro a ser atendido.
- LCFS/LIFO: último a entrar, primeiro a ser atendido.
- RSS: seleção aleatória.
- PR: prioridades.
- GD: disciplina arbitrária ou específica.

A Tabela 1 apresenta os símbolos padronizados e suas descrições.

Tabela 1 - Tabela de símbolos padrões da notação de Kendall

Características	Símbolo	Explicação
Distribuição de Tempo de Interchegada (A) e Distribuição de Tempo de Serviço (B)	M	Exponencial
	D	Determinístico
	E_k	Tipo k-Erlang ($k = 1, 2, \dots$)
	H_k	Mistura de k exponenciais
	PH	Tipo Fase
	G	Geral
Número Paralelo de Servidores (m)	$1, 2, \dots, \infty$	
Restrição na capacidade do sistema (k)	$1, 2, \dots, \infty$	
Disciplina da fila (M)	FCFS	First Come First Served
	LCFS	Last Come First Served
	RSS	Seleção Aleatória por Serviço
	PR	Prioridade
	GD	Disciplina Geral

Fonte: Adaptado de Costa (2020)

A notação de Kendall permite representar de forma padronizada diferentes configurações de sistemas de filas. A partir dessa representação, é possível descrever diversos modelos utilizados para analisar o funcionamento de sistemas de atendimento. Entre os mais conhecidos estão os modelos M/M/1 e M/M/S/K, que consideram diferentes quantidades de servidores e possíveis limitações na capacidade do sistema. Nas seções seguintes, esses modelos são apresentados juntamente com suas principais métricas de desempenho.

2.4 M/M/1

De acordo com a notação de Kendall, o modelo M/M/1 representa um sistema de filas com chegadas e tempos de atendimento exponenciais e apenas um servidor. Nesse modelo, as chegadas seguem uma distribuição de Poisson com taxa λ , enquanto os tempos de serviço são representados por uma distribuição exponencial com taxa μ . Considera-se, em geral, que a capacidade do sistema é ilimitada e que o atendimento ocorre na ordem de chegada, seguindo a disciplina FIFO (HILLIER; LIEBERMAN, 2013).

As principais métricas analíticas do modelo incluem:

Tabela 2 - Fórmulas modelo M/M/1

Descrição	Fórmulas
Probabilidade de que o Sistema Esteja Ocupada	$\rho = \frac{\lambda}{\mu}$
Probabilidade de que n Clientes Encontrem-se no Sistema	$P_n = (1 - \rho)\rho^n$
Probabilidade de que o sistema esteja Desocupada	$P_0 = (1 - \rho)$
Numero Médio de Clientes no Sistema de Atendimento	$L = \frac{\lambda}{\lambda - \mu}$
Numero Médio de Clientes na Fila de Espera	$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)}$
Tempo Médio Gasto no Sistema pelo Cliente	$W = \frac{1}{\mu - \lambda}$
Tempo Médio de Espera na Fila por Cliente	$W_q = \frac{\lambda}{\mu(\mu - \lambda)}$

Fonte: Adaptado de Pinto (2011)

2.5 M/M/S/K

O modelo M/M/S/K pode ser entendido como uma generalização do M/M/1, pois considera a existência de S servidores atuando em paralelo e uma capacidade total limitada a K clientes no sistema. Assim como no modelo básico, as chegadas seguem uma distribuição de Poisson com taxa λ e os tempos de atendimento são representados por uma distribuição exponencial com taxa μ . A disciplina de atendimento adotada é, em geral, FIFO (HILLIER; LIEBERMAN, 2013).

O desempenho desse sistema depende diretamente da relação entre a demanda e a capacidade total de atendimento, representada por $S\mu$. A partir dessa relação, é possível caracterizar o comportamento do sistema por meio dos seguintes parâmetros principais:

Probabilidade de que o Sistema Esteja Ocupado (ρ): Representa a razão entre a taxa média de chegada (λ) e a capacidade total de atendimento do sistema ($S\mu$), por exemplo: se chegam 20 clientes por hora ($\lambda=20$) e o sistema possui dois servidores com taxa de atendimento de 15 clientes por hora cada ($S=2$, $\mu=15$), a probabilidade de ocupação será $\rho = \frac{\lambda}{S\mu} = \frac{20}{2 \cdot 15} = 0,67$.

Probabilidade de que o sistema esteja Desocupado (P_0): Calcula a probabilidade de o sistema estar vazio, ou seja, sem clientes no momento, utilizando a fórmula que combina somatório e fatoriais, por exemplo: em um sistema com três servidores ($S=3$) e taxa de ocupação calculada ($\rho=0,5$), a fórmula determinará o percentual de tempo em que todos os servidores estão ociosos.

Probabilidade de que n Clientes Encontrem-se no Sistema (P_n):

Determina a probabilidade de haver exatamente n clientes no sistema, onde n pode variar entre 0 e S para ocupação parcial e ser maior que S para filas, por exemplo: se $n=5$ em um sistema com dois servidores ($S=2$), a probabilidade será calculada separadamente conforme o estado de ocupação ou espera.

Número Médio de Clientes na Fila de Espera (L_q): Representa a quantidade média de clientes que aguardam atendimento na fila.

Assim, em um sistema com $S=2$ servidores e taxa de chegada $\lambda=10$, pode-se determinar que, em média, 3 clientes estão na fila de espera.

Número Médio de Clientes no Sistema de Atendimento (L): Calcula o número médio de clientes atendidos e na fila, considerando o tempo médio de permanência no sistema, por exemplo: em um sistema com taxa de chegada $\lambda=5$ e taxa de serviço $\mu=3$, o valor de L incluirá clientes na fila e sendo atendidos.

Tempo Médio de Espera na Fila por Cliente (W_q): Determina o tempo médio que um cliente passa aguardando na fila antes de ser atendido, por exemplo: em um sistema com taxa de chegada $\lambda=8$ clientes por hora e taxa de atendimento $\mu=4$ clientes por hora, o tempo de espera pode ser de 15 minutos por cliente.

Tempo Médio Gasto no Sistema pelo Cliente (W): Representa o tempo total que um cliente passa no sistema, incluindo fila e atendimento, por exemplo, se o tempo médio na fila é de 10 minutos (W_q) e o atendimento dura 5 minutos ($\frac{1}{\mu}$), o tempo total será 15 minutos por cliente.

As equações que calculam essas métricas podem ser vistas na Tabela 3:

Tabela 3 - Fórmulas modelo M/M/S

Descrição	Fórmulas
Probabilidade de que o Sistema Esteja Ocupada	$\rho = \frac{\lambda}{S\mu}$
Probabilidade de que o sistema esteja Desocupada	$P_0 = \left[\sum_{n=0}^{S-1} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \frac{\left(\frac{\lambda}{\mu}\right)^S}{S!} \frac{1}{1-\rho} \right]^{-1}$
Probabilidade de que n Clientes Encontrem-se no Sistema	$P_n = \begin{cases} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} P_0, & \text{se } 0 \leq n \leq S \\ \frac{\left(\frac{\lambda}{\mu}\right)^n}{S! S^{n-S}} P_0 & \text{se } n \geq S \end{cases}$
Numero Médio de Clientes na Fila de Espera	$L_q = \frac{P_0 \left(\frac{\lambda}{\mu}\right)^S \rho}{S! (1-\rho)^2}$
Numero Médio de Clientes no Sistema de Atendimento	$L = \frac{P_0 \left(\frac{\lambda}{\mu}\right)^S \rho}{S! (1-\rho)^2} + \frac{\lambda}{\mu}$
Tempo Médio de Espera na Fila por Cliente	$W_q = \frac{L_q}{\lambda}$
Tempo Médio Gasto no Sistema pelo Cliente	$W = W_q + \frac{1}{\mu}$

Fonte: Adaptado de Pinto (2011)

3. APLICAÇÃO EM SISTEMAS DE SERVIÇOS

Este trabalho combina análise teórica e simulações computacionais para investigar o desempenho de sistemas de filas em ambientes de serviço, tendo como aplicação prática um posto de combustível localizado no bairro do Sacomã, em São Paulo. Os dados utilizados foram coletados diretamente pelo autor por meio de observações estruturadas no local, com registro de tempos de atendimento, taxas de chegada e outras variáveis operacionais relevantes. Essa coleta permitiu trabalhar com informações alinhadas à realidade do sistema analisado.

As análises desenvolvidas tiveram como base a Teoria de Filas e simulações utilizando Python, que possibilitam representar o funcionamento do sistema e avaliar diferentes cenários operacionais. A utilização de simulação como abordagem metodológica justifica-se pela sua capacidade de lidar com sistemas dinâmicos e sujeitos à variabilidade, além de permitir a comparação entre alternativas de configuração e a identificação de oportunidades de melhoria (STEVENSON, 2018).

Para a realização das simulações e o cálculo das métricas de desempenho, como tempo médio de espera, nível de utilização e dimensionamento do número de servidores, foi utilizada a linguagem Python como ferramenta principal. A escolha se deu pela sua flexibilidade e pela disponibilidade de bibliotecas voltadas a cálculos científicos. O uso da ferramenta permitiu aproximar teoria e prática, reproduzindo o comportamento dos modelos analíticos e comparando-os com as situações observadas em campo, o que reforça a consistência dos resultados obtidos (TAHA, 2017).

3.1 PROCESSO DE ATENDIMENTO NO POSTO DE COMBUSTÍVEL

As simulações realizadas neste trabalho foram construídas a partir do fluxo real de atendimento observado em um posto de combustível, o que permitiu aplicar os conceitos da Teoria de Filas a uma situação concreta. O mapeamento do processo considerou desde a chegada do veículo até a finalização do abastecimento e do pagamento, possibilitando compreender a dinâmica do sistema e identificar pontos de maior sobrecarga.

A partir dessa observação, foi possível detectar momentos de acúmulo de clientes e aumento no tempo de espera, caracterizando gargalos operacionais. Essa análise serviu de base para a avaliação de alternativas de organização do sistema,

como ajustes no número de servidores, reorganização das filas e redistribuição das etapas do atendimento.

A simulação, conforme apontado por Shannon (1975), é uma ferramenta relevante para estudar o comportamento de sistemas complexos e testar hipóteses operacionais. Em ambientes de serviço, sua aplicação se mostra especialmente útil, pois permite analisar cenários hipotéticos, estimar os impactos de mudanças estruturais e antecipar possíveis resultados antes de qualquer implementação prática (BANKS et al., 2010).

3.2 ETAPAS DO ATENDIMENTO

Para a modelagem do sistema e parametrização das simulações, foram identificadas e estruturadas as principais etapas do fluxo de atendimento no posto de combustível analisado.

As etapas consideradas foram:

- 1) Chegada ao posto: o cliente chega ao estabelecimento e dirige-se à área de abastecimento, geralmente selecionando a bomba com menor fila ou a que estiver imediatamente disponível.
- 2) Fila de espera na bomba: quando todos os servidores (bombas) estão ocupados, os clientes aguardam em fila até que um frentista e uma bomba fiquem livres.
- 3) Atendimento e abastecimento: o frentista recebe o cliente, identifica o combustível solicitado, realiza o abastecimento e finaliza as operações necessárias relativas ao veículo.
- 4) Pagamento: concluído o abastecimento, o cliente efetua o pagamento, podendo utilizar diferentes meios (dinheiro, cartão ou aplicativos digitais), finalizando o processo de atendimento.

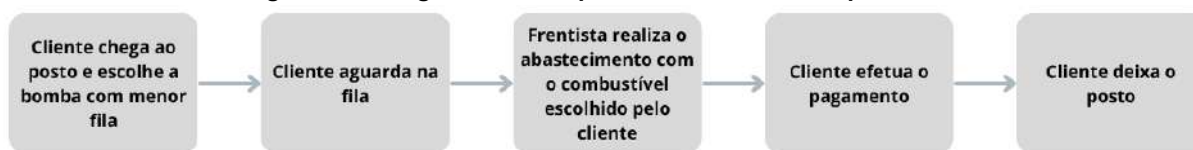
3.3 FLUXOGRAMA DAS ETAPAS DO ATENDIMENTO

Com base nas etapas identificadas, foi elaborado um fluxograma representando o fluxo operacional do sistema de atendimento, desde a entrada do cliente até a sua saída. Nele são destacados os momentos de chegada, espera, atendimento e conclusão do serviço.

Esse fluxograma contribui para a compreensão da dinâmica do processo e serviu como apoio para a modelagem dos cenários simulados. Além disso, facilitou

a interpretação dos resultados, ao permitir relacionar os indicadores obtidos com cada etapa do atendimento.

Figura 1: Fluxograma das etapas do atendimento no posto.



Fonte: Autor (2025)

3.4 ESTUDO DE CASO

O estudo de caso foi realizado em um posto de combustível localizado no bairro do Sacomã, na zona sul de São Paulo, com o objetivo de analisar o desempenho do sistema de atendimento e avaliar possíveis alternativas para melhorar sua eficiência operacional. O estudo concentrou-se nas etapas de chegada, espera, abastecimento e pagamento, conforme apresentado no fluxograma da seção anterior.

No processo observado, o cliente chega ao posto, posiciona o veículo em uma das filas de abastecimento e aguarda até ser atendido por um frentista. O profissional recebe o veículo, confirma o tipo e o valor do combustível desejado, realiza o abastecimento e, ao final, libera a bomba para o próximo cliente.

Embora o posto ofereça serviços complementares, como verificação de óleo, calibragem de pneus e venda de aditivos, o estudo considerou apenas as etapas de abastecimento e pagamento, por representarem o núcleo da operação e o principal ponto de formação de filas.

A coleta de dados foi realizada em três dias distintos, por meio da gravação em vídeo de períodos de aproximadamente 10 minutos das operações do posto, com o intuito de observar diferentes condições de funcionamento e variações na demanda ao longo do período analisado.

- Dia 1 (sexta-feira, 18h): horário de pico, com atendimento unificado (abastecimento + pagamento) diretamente na bomba.

Figura 2: Ilustração da fila registrada pelo autor durante a coleta de dados na sexta-feira (10/out) às 18h.



Fonte: Autor (2025).

- Dia 2 (quarta-feira, 11h): atendimento moderado, com pagamento centralizado em uma fila M/M/2.

Figura 3: Ilustração da fila registrada pelo autor durante a coleta de dados na quarta-feira (5/nov) às 11h.



Fonte: Autor (2025).

- Dia 3 (sábado, 8h): menor fluxo, com pagamento separado em fila M/M/1.

Figura 4: Ilustração da fila registrada pelo autor durante a coleta de dados no sábado (8/nov) às 8h.



Fonte: Autor (2025).

A coleta realizada também permitiu identificar diferenças relevantes na estrutura do atendimento. Na sexta-feira, o pagamento era efetuado diretamente na bomba, enquanto nos demais dias os clientes se dirigiam a uma fila única para pagar. Essa alteração muda a organização do sistema e influencia diretamente seu desempenho, pois afeta a taxa efetiva de chegada (λ), a taxa de atendimento (μ) e o nível de ocupação da capacidade disponível ($S \cdot \mu$).

Observou-se ainda que, nos horários de maior movimento, as filas ultrapassam os limites do pátio do posto e avançam para a via pública, interferindo no tráfego local. Esse comportamento evidencia um desequilíbrio entre demanda e capacidade, além de indicar possível perda de clientes que desistem do atendimento diante do tempo de espera.

Esses aspectos reforçam a importância de uma análise quantitativa baseada na Teoria de Filas, a fim de avaliar alternativas relacionadas ao número de servidores, à organização das filas e à estratégia operacional adotada em diferentes períodos do dia.

Conforme Andrade (2009), fatores como o padrão de chegada, a disciplina de atendimento e a capacidade do sistema exercem influência direta sobre o desempenho. No caso estudado, verificou-se que as chegadas dos veículos podem ser aproximadas por um processo de Poisson e que os tempos de atendimento

apresentam comportamento compatível com a distribuição exponencial, o que justifica a utilização dos modelos M/M/S e M/M/S/K.

Com base nesses fundamentos, o posto foi analisado em diferentes configurações operacionais, simulando cenários com 6 filas M/M/1, 3 filas M/M/2 e 1 fila M/M/6, além de modelagens em duas etapas, separando abastecimento e pagamento.

A configuração que apresentar melhor desempenho será utilizada para estimar o dimensionamento mais adequado do número de servidores, considerando custos de mão de obra, ganho potencial e redução de perdas operacionais.

3.5 PROGRAMAÇÃO PYTHON PARA ANÁLISES DE SIMULAÇÕES

Neste trabalho, utilizou-se a linguagem Python como principal ferramenta para análise e simulação dos sistemas de filas. Trata-se de uma linguagem de alto nível, bastante utilizada em aplicações científicas e de modelagem, principalmente pela sua simplicidade e facilidade de implementação.

Além dos cálculos analíticos, o uso do Python permitiu desenvolver um script capaz de simular o funcionamento do sistema ao longo do tempo, reproduzindo situações próximas da realidade.

Entre os benefícios observados destacam-se:

- facilidade na implementação de modelos matemáticos,
- precisão no cálculo das métricas,
- flexibilidade para simular diferentes cenários operacionais e
- integração eficiente entre dados coletados e modelos analíticos.

3.6 CÓDIGO E SIMULAÇÕES

Além da análise teórica, foi desenvolvido um script em Python para simular o comportamento do sistema em diferentes configurações. A ideia foi testar, de forma prática, como o sistema reage à variação de parâmetros como a taxa de chegada (λ), o tempo médio de atendimento, o número de servidores (S) e a capacidade máxima (K).

A simulação foi construída seguindo a lógica de eventos discretos, ou seja, o sistema não evolui continuamente no tempo, mas sim de evento em evento

(chegada ou saída de um cliente). Dessa forma, o tempo avança sempre para o próximo evento relevante, o que torna a simulação mais eficiente e mais próxima do funcionamento real de um sistema de filas.

Os tempos entre chegadas e os tempos de atendimento não são fixos. Em cada execução, esses valores são gerados aleatoriamente a partir de uma distribuição exponencial, utilizando a função *expovariate* da biblioteca *random*. Isso segue diretamente a hipótese dos modelos M/M/s, nos quais tanto o processo de chegada quanto o de atendimento apresentam comportamento probabilístico. A taxa de atendimento (μ) foi obtida como o inverso do tempo médio de serviço informado.

Durante a execução, o código mantém o controle das principais variáveis do sistema, como o número de clientes presentes (n), o número de clientes na fila (n_q) e a quantidade de servidores ocupados. Para organizar a fila de espera, foi utilizada uma estrutura do tipo FIFO, garantindo que os clientes sejam atendidos na ordem de chegada.

Quando ocorre uma chegada, o programa primeiro verifica se o sistema possui limite de capacidade. Caso exista um valor máximo K e ele já tenha sido atingido, o cliente é bloqueado e não entra no sistema. Caso contrário, ele é admitido. Se houver servidor livre, o atendimento começa imediatamente; se não houver, o cliente entra na fila.

Já nos eventos de saída, o cliente que termina o atendimento é removido do sistema, e seus tempos são utilizados para calcular as métricas de desempenho. Caso existam clientes aguardando, o próximo da fila inicia atendimento imediatamente. Caso contrário, o servidor fica ocioso até a próxima chegada.

Um ponto importante da implementação foi o uso de um período inicial de aquecimento (*warm-up*). Como a simulação começa com o sistema vazio, os primeiros instantes não representam bem o comportamento real. Por isso, os dados coletados nesse período inicial são descartados, e apenas os valores após esse intervalo são considerados no cálculo das métricas.

As métricas foram calculadas de duas formas. A primeira considera os clientes atendidos, permitindo obter o tempo médio no sistema (W) e o tempo médio de espera na fila (Wq). A segunda considera o comportamento do sistema ao longo do tempo, acumulando áreas para calcular o número médio de clientes no sistema (L), o número médio na fila (Lq), a utilização dos servidores e a probabilidade de o sistema estar vazio.

Além disso, o código também registra a quantidade de clientes que chegaram, foram atendidos, bloqueados e concluídos, o que permite calcular a taxa efetiva de atendimento e a probabilidade de bloqueio, especialmente relevantes nos cenários com capacidade limitada.

Por se tratar de uma simulação baseada em geração aleatória, os resultados variam ligeiramente a cada execução. Mesmo assim, para tempos de simulação mais longos, os valores tendem a se aproximar dos resultados teóricos, o que permite validar o comportamento do modelo. A utilização de uma semente fixa no gerador de números aleatórios também ajudou a manter a consistência entre diferentes testes.

O código completo utilizado nas simulações está apresentado no Apêndice A e pode ser adaptado para outros sistemas de filas com características semelhantes.

3.7 COMPARAÇÃO ENTRE RESULTADOS TEÓRICOS E SIMULADOS

A Tabela 4 apresenta a comparação entre os resultados teóricos e os obtidos por simulação para os cenários analisados.

Tabela 4 - Comparação entre os resultados teóricos e simulados.

Modelo	Métrica	Teórico	Simulado	Erro (%)
M/M/1	W (min)	13,33	13,06	2,03%
	Wq (min)	11,73	11,45	2,39%
	L	7,33	7,18	2,05%
	Lq	6,45	6,3	2,33%
M/M/2/10	W (min)	14,09	14,01	0,56%
	Wq (min)	10,29	10,2	0,88%
	L	7,24	7,17	1,03%
	Lq	5,29	5,22	1,36%
M/M/6/30	W (min)	16,65	16,52	0,80%
	Wq (min)	12,85	12,71	1,12%
	L	26,29	26	1,09%
	Lq	20,29	20	1,41%

Fonte: Autor (2025).

De forma geral, observa-se que os valores simulados ficaram bastante próximos dos valores teóricos, com diferenças pequenas em todas as métricas avaliadas. Considerando os três cenários, os erros percentuais médios ficaram em torno de **1% a 2%**, dependendo do indicador analisado.

As métricas relacionadas ao número de clientes no sistema (L e Lq) e aos tempos médios (W e Wq) apresentaram comportamento consistente, com variações reduzidas e dentro do esperado para esse tipo de abordagem. Mesmo nos casos com maior diferença, os desvios permaneceram baixos e não comprometem a análise dos resultados.

Essas diferenças ocorrem principalmente porque a simulação considera a aleatoriedade dos processos de chegada e atendimento, enquanto o modelo teórico

trabalha com valores médios. Dessa forma, é natural que os resultados não coincidam exatamente, mas se mantenham próximos.

Além disso, como a simulação foi realizada em um tempo finito, ainda existe um pequeno nível de variação nos resultados. Ainda assim, observa-se que, para simulações mais longas, os valores tendem a se aproximar ainda mais dos resultados teóricos.

De modo geral, a proximidade entre os resultados indica que o modelo implementado é adequado e consegue representar bem o comportamento do sistema, sendo uma ferramenta confiável para análise dos cenários propostos.

4. RESULTADOS

Este capítulo apresenta os resultados obtidos a partir da coleta de dados realizada em três dias e horários distintos em um posto de combustível localizado no bairro do Sacomã, na cidade de São Paulo.

A coleta foi realizada por meio da gravação em vídeo das atividades do posto durante períodos de aproximadamente 10 minutos. Posteriormente, as gravações foram analisadas manualmente, registrando-se os instantes de chegada dos veículos e o tempo necessário para a conclusão dos atendimentos.

A partir dessas observações foram calculados os intervalos entre chegadas consecutivas e os tempos de serviço, permitindo estimar a taxa média de chegada (λ) e a taxa média de atendimento (μ), parâmetros utilizados na modelagem do sistema de filas.

Com base nesses parâmetros, foram analisados diferentes cenários de operação do posto. As análises consideram as variações na demanda observadas ao longo desses períodos, bem como diferentes configurações operacionais, com pagamento realizado na própria bomba e com fila separada para pagamento. A partir dessa comparação, foi possível avaliar como essas mudanças influenciam o desempenho do sistema e de que forma diferentes arranjos de filas impactam a eficiência operacional e os resultados econômicos.

Os resultados estão organizados em quatro partes principais:

- (4.1) Chegada dos clientes, que descreve os padrões de fluxo e as taxas médias de chegada (λ);
- (4.2) Atendimento, que apresenta os tempos médios de serviço ($1/\mu$) e a dinâmica operacional observada;
- (4.3) Cálculos de desempenho e simulações, em que são aplicados modelos da Teoria das Filas (M/M/S, M/M/S/K e sistemas em série) para comparar os cenários de abastecimento e pagamento (6×M/M/1, 3×M/M/2, 1×M/M/6, M/M/2 e 2×M/M/1), incluindo análise econômica e cenários de expansão com oito bombas (4×M/M/2); e
- (4.4) Síntese e recomendação final, que integra os resultados operacionais e econômicos e apresenta a configuração recomendada para o posto, com base nas simulações realizadas.

4.1 CHEGADA DOS CLIENTES

Dia 1 – Sexta-feira às 18h

No primeiro dia de observação, registrou-se a chegada de 17 veículos em um intervalo de 10 minutos, conforme registrado na Tabela 5, resultando em uma taxa média de chegada de $\lambda = 1,70$ veículos/minuto.

Nesse período, o posto operava com o pagamento realizado diretamente na bomba, concentrando abastecimento e pagamento em uma única etapa. Como consequência, o tempo de ocupação de cada servidor tornou-se maior.

Durante o horário de pico, observou-se que as filas frequentemente ultrapassavam os limites do pátio e avançavam sobre a via pública (Figura 2). Esse comportamento indica que a taxa de chegada superou a capacidade total de atendimento, representada por $S\mu$, caracterizando um regime de operação supercrítico. Nessa situação, o sistema tende à saturação, o que pode gerar perda de clientes potenciais, tanto pelo tempo elevado de espera quanto pela limitação física do espaço disponível.

Tabela 5 - Informações de chegada de clientes (Dia 1)

Minuto	Clientes que chegaram
1	3
2	1
3	2
4	2
5	0
6	2
7	3
8	1
9	2
10	1
Total	17
Clientes por minuto	1,7

Fonte: Autor (2025).

Figura 5 – Gráfico de chegadas de veículos por minuto (Dia 1)



Fonte: Autor (2025).

Dia 2 – Quarta-feira às 11h

Na segunda coleta, foi registrada uma taxa média de chegada de $\lambda = 2,00$ veículos por minuto, conforme registrado na Tabela 6, valor ligeiramente superior ao observado na sexta-feira. Essa diferença pode ser explicada, em parte, pelo fato de que, na sexta-feira, as filas ultrapassavam os limites do pátio, o que possivelmente acabava restringindo a entrada de novos veículos devido à limitação física do espaço.

Diferentemente do primeiro cenário, o pagamento passou a ser realizado em uma fila separada após o abastecimento, configurando um sistema M/M/2 para essa etapa. Essa alteração reduziu o tempo de ocupação das bombas, aumentou o giro dos servidores e contribuiu para uma maior fluidez no pátio, com filas mais curtas e melhor organização do atendimento.

O padrão de chegada dos veículos manteve comportamento aleatório e aproximadamente independente, compatível com um processo de Poisson, o que reforça a adequação da parametrização adotada nos modelos de filas analisados nas seções seguintes.

Tabela 6 - Informações de chegada de clientes (Dia 2)

Minuto	Clientes que chegaram
1	2
2	3
3	1
4	4
5	2
6	3
7	0
8	2
9	2
10	1
Total	20
Clientes por minuto	2

Fonte: Autor (2025).

Figura 6 – Gráfico de chegadas de veículos por minuto (Dia 2)



Fonte: Autor (2025).

Dia 3 – Sábado às 8h

No terceiro dia de observação, registrou-se uma taxa média de chegada de $\lambda = 1,7$ veículos/minuto, conforme registrado na Tabela 7, valor semelhante ao observado na sexta-feira, porém com formação de filas substancialmente menor.

Essa diferença pode ser atribuída a dois fatores principais. Na sexta-feira, o sistema já operava próximo ao limite físico do pátio, o que possivelmente reduziu a taxa efetiva de entrada em razão do bloqueio causado pelas filas que se formavam fora do estabelecimento. No sábado, por sua vez, o processo foi organizado em duas etapas, com abastecimento seguido de pagamento em uma fila única M/M/1.

Essa divisão reduziu o tempo de ocupação das bombas e contribuiu para diminuir o nível de saturação do sistema.

Mesmo em condições mais estáveis, verificou-se que o padrão de chegada dos veículos permaneceu compatível com uma distribuição de Poisson, comportamento comum em sistemas de serviços urbanos e coerente com os modelos adotados neste estudo (ALMEIDA et al., 2019).

Tabela 7 - Informações de chegada de clientes (Dia 3)

Minuto	Clientes que chegaram
1	1
2	0
3	2
4	1
5	4
6	1
7	3
8	0
9	4
10	1
Total	17
Clientes por minuto	1,7

Fonte: Autor (2025).

Figura 7 – Gráfico de chegadas de veículos por minuto (Dia 3)



Fonte: Autor (2025).

Os padrões observados nos 3 dias são coerentes com um processo de chegadas do tipo Poisson, típico de demandas aleatórias em serviços urbanos. A distribuição de Poisson é dada por:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

em que o parâmetro λ é a taxa média do processo. (ALMEIDA, ARAUJO, BARROS, SAMPAIO, 2019).

4.2 ATENDIMENTO DOS CLIENTES

O processo de atendimento no posto de combustível foi analisado a partir das observações realizadas nos três dias avaliados, considerando diferentes níveis de fluxo e distintas formas de organização operacional.

O processo de atendimento é composto pelas etapas de escolha do combustível, abastecimento e pagamento. Dependendo da configuração operacional adotada em cada dia de coleta, essas etapas podem ocorrer de forma integrada ou sequencial. No primeiro dia, o pagamento foi realizado diretamente na bomba, caracterizando um sistema integrado. Já nos dias 2 e 3, o pagamento foi realizado em caixa separado após a conclusão do abastecimento, configurando um sistema sequencial.

Nos dias em que o pagamento ocorreu de forma sequencial, observou-se que veículos que realizaram pagamento em dinheiro puderam efetuar o pagamento diretamente na bomba. Dessa forma, nem todos os veículos que passaram pela etapa de abastecimento ingressaram na fila de pagamento, o que implica que o número de veículos atendidos no abastecimento não é necessariamente igual à taxa de chegada na etapa de pagamento.

Durante os cálculos e simulações, as etapas de abastecimento e pagamento foram analisadas separadamente, considerando-se uma nova taxa de chegada específica para a etapa de pagamento, ajustada de acordo com o fluxo real de veículos que efetivamente ingressaram nessa etapa do sistema.

Em todos os cenários analisados, o posto operava com seis bombas de abastecimento, cada uma com sua própria fila de espera. Dessa forma, a taxa total de chegada ao posto, λ total, foi distribuída entre as seis bombas, resultando em uma taxa média de chegada por servidor dada por:

$$\lambda_i = \frac{\lambda_{total}}{6}$$

Essa aproximação está de acordo com o comportamento observado dos motoristas, que, em geral, procuram a bomba com menor fila, promovendo um balanceamento informal entre os pontos de atendimento.

Assim, cada bomba pôde ser modelada como um sistema M/M/1 independente na etapa de abastecimento. Essa representação se mostrou adequada para descrever a dinâmica real do processo e possibilitou a aplicação dos modelos analíticos apresentados no Capítulo 2.

A seguir foi analisado o atendimento de cada dia de forma isolada:

Dia 1 — Sexta-feira às 18h

Configuração: pagamento realizado diretamente na bomba (sistema em uma única etapa).

Modelo de abastecimento: $6 \times M/M/1$.

A taxa total de chegada registrada foi de:

$$\lambda_{total} = 1,70 \text{ veículos/minuto}$$

Resultando em uma taxa média por bomba de

$$\lambda_i = 1,70/6 \approx 0,2833 \text{ veículos/minuto por bomba.}$$

O tempo médio de atendimento foi de 5,5 minutos, calculado a partir dos registros da Tabela 8, o que corresponde a:

$$\mu = 1/5,50 \approx 0,1818 \text{ veículos/minuto.}$$

Como o pagamento é realizado diretamente na bomba, o tempo total de serviço por cliente aumenta, prolongando a ocupação do servidor e reduzindo o giro operacional. Isso resulta em um fator de utilização por bomba de:

$$\rho_i = \frac{\lambda_i}{\mu} \approx 1,56$$

Valor significativamente superior a 1, caracterizando operação em regime supercrítico, com formação de filas controladas apenas pela capacidade física do pátio. Nessa condição, as filas ultrapassam frequentemente os limites da área interna, impactando o tráfego externo e indicando perda potencial de clientes.

Tabela 8 - Tempos de atendimento (Dia 1)

Cliente	Tempo de atendimento (min)
Astra	6
C4	8
Voyage	4
Meriva	4
Tempo médio	5,5

Fonte: Autor (2025).

Dia 2 — Quarta-feira às 11h

Configuração: pagamento separado, atendido por dois caixas.

Etapa 1 – Abastecimento:

Modelo: $6 \times M/M/1$.

A taxa total de chegada registrada foi de:

$$\lambda_1 = 2,00 \text{ veículos/minuto}$$

resultando em uma taxa média por bomba de

$$\lambda_i = 2,00/6 \approx 0,3333 \text{ veículos/minuto por bomba.}$$

O tempo médio de atendimento na etapa de abastecimento foi de 3,8 minutos, conforme registrado na Tabela 9, resultando em:

$$\mu_1 = 1/3,8 \approx 0,2632 \text{ veículos/minuto.}$$

Dessa forma, o fator de utilização das bombas foi de:

$$\rho_1 = \frac{\lambda_i}{\mu_1} \approx 1,27$$

valor superior a 1, indicando operação em regime supercrítico, ainda que parcialmente mitigada pela separação da etapa de pagamento.

Tabela 9 - Tempos de atendimento no abastecimento (Dia 2)

Cliente	Tempo de atendimento no abastecimento (min)
Cobalt	3
Sportage	3
HRV	6
Siena	3
T-Cross	4
Tempo médio	3,8

Fonte: Autor (2025).

Etapa 2 – Pagamento:**Modelo:** M/M/2.

Conforme registrado na Tabela 10, a taxa de chegada à fila de pagamento foi de

$$\lambda_2 = 1,10 \text{ veículos/minuto}$$

O tempo médio de atendimento foi de 1,6 minutos, conforme registrado na Tabela 11, implicando:

$$\mu_2 = 1/1,60 \approx 0,625 \text{ veículos/minuto.}$$

Com dois servidores, o fator de utilização da etapa de pagamento é:

$$\rho_2 = \frac{\lambda_2}{2\mu_2} \approx 0,88$$

A separação do pagamento da bomba reduziu o tempo de ocupação dos servidores da etapa 1, aumentando o giro das bombas e diminuindo a formação de filas no pátio.

Apesar da utilização elevada na etapa de pagamento, o sistema permaneceu estável, com tempos de espera moderados ($W_q \approx 5,16$ minutos), graças ao balanceamento automático proporcionado pela fila única M/M/2.

Tabela 10 - Informações de chegada de clientes no pagamento (Dia 2)

Minuto	Clientes que chegaram no pagamento
1	1
2	0
3	1
4	3
5	2
6	0
7	1
8	1
9	2
10	0
Total	11
Clientes por minuto	1,1

Fonte: Autor (2025).

Tabela 11 - Tempos de atendimento no pagamento (Dia 2)

Cliente	Tempo de atendimento no pagamento (min)
Fit	1
Kicks	1
Ka	2
Sportage	2
T-Cross	2
Tempo médio	1,6

Fonte: Autor (2025).

Dia 3 — Sábado às 8h

Configuração: pagamento separado, atendido por um caixa.

Etapas 1 – Abastecimento:

Modelo: $6 \times M/M/1$.

A taxa total de chegada observada foi

$$\lambda_{total} = 1,70 \text{ veículos/minuto}$$

Resultando em uma taxa média por bomba de:

$$\lambda_i = 1,70/6 \approx 0,2833 \text{ veículos/minuto por bomba.}$$

O tempo médio de atendimento foi de 3,8 minutos, conforme registrado na Tabela 12, levando a:

$$\mu_1 = 1/3,80 \approx 0,2632 \text{ veículos/minuto.}$$

O fator de utilização na etapa de abastecimento foi de:

$$\rho_1 = \frac{\lambda_i}{\mu_1} \approx 1,08$$

Esse valor é ligeiramente superior a 1, caracterizando operação em regime supercrítico, embora com menor intensidade que nos demais cenários.

Tabela 12 - Tempos de atendimento no abastecimento (Dia 3)

Cliente	Tempo de atendimento no abastecimento (min)
Sentra	3
Fit	3
Fox	4
Fusion	4
Siena	5
Tempo médio	3,8

Fonte: Autor (2025).

Etapas 2 – Pagamento:

Modelo: M/M/1.

Conforme registrado na Tabela 13, a taxa de chegada à etapa de pagamento foi de

$$\lambda_2 = 0,50 \text{ veículos/minuto}$$

O tempo médio de atendimento de 1,4 minutos, conforme registrado na Tabela 14, corresponde a:

$$\mu_2 = 1/1,40 \approx 0,7143 \text{ veículos/minuto.}$$

O fator de utilização do caixa é:

$$\rho_2 = \frac{\lambda_i}{\mu_2} \approx 0,7$$

A separação do pagamento da bomba reduziu o tempo de ocupação dos servidores da etapa 1, aumentando o giro das bombas e diminuindo a formação de filas no pátio.

Apesar da utilização elevada na etapa de pagamento, o sistema permaneceu estável, com tempos de espera moderados ($W_q \approx 3,27$ minutos).

Tabela 13 - Informações de chegada de clientes no pagamento (Dia 3)

Minuto	Cientes que chegaram
1	1
2	1
3	0
4	1
5	4
6	0
7	2
8	0
9	0
10	1
Total	10
Cientes por minuto	1

Fonte: Autor (2025).

Tabela 14 - Tempos de atendimento no pagamento (Dia 3)

Cliente	Tempo de atendimento no pagamento (min)
Pagero	1
Fox	2
Corsa	1
Palio	1
Etios	2
Tempo médio	1,4

Fonte: Autor (2025).

Síntese comparativa

A análise dos três dias de coleta mostra que a divisão da taxa total de chegada entre as seis bombas representa de forma adequada o comportamento operacional observado no posto. Essa distribuição permite modelar cada bomba

como um sistema M/M/1 na etapa de abastecimento, refletindo de maneira consistente a dinâmica real do atendimento.

Também se verificou que a separação da etapa de pagamento reduz o tempo de ocupação das bombas, aumentando o giro operacional e diminuindo a probabilidade de formação de filas extensas.

O Dia 1 apresentou o pior desempenho, com saturação evidente em razão da integração entre abastecimento e pagamento, resultando em $\rho > 1$ em todas as bombas. No Dia 2, observou-se o cenário mais equilibrado, no qual o pagamento organizado em fila única M/M/2 conseguiu absorver o fluxo gerado pelo abastecimento, evitando o acúmulo excessivo de veículos no pátio. Já no Dia 3, a etapa de pagamento, modelada como M/M/1, apresentou o melhor desempenho entre os cenários analisados, operando com taxa de utilização inferior a 1 ($\rho \approx 0,70$) e mantendo o sistema estável, com tempos de espera moderados ($Wq \approx 3,27$ minutos). Esse resultado, no entanto, está diretamente relacionado à menor taxa de chegada e ao menor tempo de atendimento em comparação ao cenário com M/M/2.

Essas evidências empíricas servem de base para as simulações apresentadas no próximo capítulo (4.3 – Cálculos de Desempenho), no qual os arranjos $6 \times M/M/1$, $3 \times M/M/2$ e $1 \times M/M/6$ serão comparados em termos de desempenho, eficiência operacional e impacto econômico.

4.3 CÁLCULOS DE DESEMPENHO E SIMULAÇÕES DA FILA

Este capítulo apresenta a modelagem e a análise dos sistemas de filas observados no posto de combustível, utilizando os dados coletados nos três dias de estudo para representar cenários operacionais próximos da realidade.

As simulações foram desenvolvidas com o objetivo de avaliar o desempenho do sistema nas etapas de abastecimento e pagamento, considerando diferentes estruturas de filas, regimes de operação e as limitações físicas do pátio. A partir dessa análise, busca-se identificar alternativas que contribuam para melhorar a eficiência operacional e o retorno financeiro do estabelecimento.

As observações realizadas em campo indicaram que o modelo com pagamento separado do abastecimento apresentou resultados mais favoráveis,

reduzindo o tempo médio total no sistema e diminuindo a formação de filas extensas no pátio.

Por esse motivo, as análises e simulações apresentadas na sequência adotam essa configuração como referência. Para garantir maior consistência nas conclusões, foram utilizadas as métricas mais conservadoras observadas em campo, conforme documentado na Tabela 15.

Tabela 15 - Métricas utilizadas nos cálculos.

Parâmetro	Símbolo	Valor	Unidade
Taxa média de chegada	λ	2	veículos/min
Tempo médio de abastecimento	$1/\mu$	3,8	min
Tempo médio de pagamento	$1/\mu$	1,6	min
Taxa média de abastecimento	μ	0,2632	veículos/min
Taxa média de pagamento	μ	0,625	veículos/min
Nº de bombas	S	6	—
Capacidade total do sistema	K	30 (5 por bomba)	veículos
Disciplina de atendimento	—	FIFO	—
Salário base do frentista	—	R\$ 1.800	mês
Custo total com encargos (~80%)	—	R\$ 3.240	mês

Fonte: Autor (2025).

4.3.1 PARÂMETROS E CONSIDERAÇÕES PRÁTICAS

Os parâmetros utilizados nas simulações foram definidos com base nos valores observados durante os três dias de coleta. Para a análise, foram selecionados os valores que mais exigem do sistema, como maiores taxas de chegada e maiores tempos de atendimento, de forma a representar cenários operacionais críticos. Essa abordagem permite avaliar o comportamento do sistema em condições próximas ao seu limite de capacidade, garantindo maior robustez nas conclusões obtidas.

Para a etapa de abastecimento, foram considerados os seguintes parâmetros básicos:

Tempo médio de serviço: 3,8 minutos, então temos:

$$\mu = \frac{1}{3,8} \approx 0,2632 \text{ veículos/minuto}$$

Taxa média de chegada utilizada como referência para os cenários comparativos: $\lambda = 2,0 \text{ veículos/minuto}$, correspondente ao período de maior demanda observado.

O sistema foi modelado em duas etapas em série, refletindo a operação real quando há separação entre abastecimento e pagamento:

Etapa 1 — Abastecimento

Clientes são atendidos por um dos servidores (bombas), três arquiteturas de filas foram simuladas:

- $6 \times M/M/1$ — seis filas independentes, uma por bomba;
- $3 \times M/M/2$ — três grupos com duas bombas cada;
- $1 \times M/M/6$ — fila única servindo seis bombas.

Essas arquiteturas permitem avaliar o impacto do balanceamento entre servidores e da disciplina de fila na utilização e no tempo médio do sistema.

Etapa 2 — Pagamento

Após o abastecimento, os veículos ingressam em uma segunda fila para pagamento, foram avaliadas duas alternativas operacionais:

- $M/M/2$ — fila única para dois caixas;
- $2 \times M/M/1$ — duas filas independentes, uma por caixa.

A comparação entre essas configurações permite estimar a sensibilidade do sistema ao subdimensionamento da etapa de pagamento, identificada empiricamente como possível gargalo.

4.3.2 ANÁLISE ECONÔMICA E CENÁRIOS DE EXPANSÃO

Para avaliar o desempenho do posto sob diferentes configurações, foram simulados três cenários de abastecimento e dois cenários de pagamento, todos parametrizados com a taxa de chegada crítica: $\lambda = 2,0 \text{ veículos/minuto}$.

O tempo médio de serviço de 3,8 minutos implica em $\mu = 0,2632$ veículos por bomba. Assim, para as seis bombas atualmente instaladas, a capacidade total é

$S \cdot \mu = 6 \cdot 0,2632 \approx 1,58$ *veículos/min*, valor inferior à taxa de chegada adotada, caracterizando operação em regime supercrítico ($\rho > 1$).

Nessa condição, a fila tenderia ao crescimento indefinido na ausência de restrição física; por isso, todos os cenários foram modelados com capacidade finita M/M/S/K, adotando-se $K = 30$ *veículos*, limite estimado do pátio do posto.

Essa abordagem permite avaliar probabilidades de bloqueio, *throughput* efetivo e impacto das restrições físicas do sistema — elementos essenciais para as decisões de expansão apresentadas na próxima seção.

Cenários de Abastecimento

1) Abastecimento – 1 fila M/M/6 (única fila com 6 servidores)

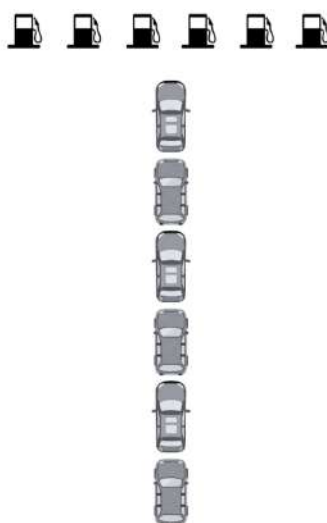
Nesse cenário, todos os veículos entram em uma única fila, sendo atendidos pela primeira bomba disponível. Esse modelo tende a apresentar melhor balanceamento entre as bombas e aumentar a eficiência operacional.

Parâmetros: $\lambda = 2,0$; $\mu = 0,2632$; $S = 6$; $K = 30$

Resultados simulações:

- Utilização: $\rho = 1,27$
- Probabilidade de bloqueio: $P_{\square} \approx 21\%$
- Taxa efetiva atendida: $\lambda_{ef} \approx 1,58$ *veíc/min*
- Tempo médio no sistema: $W \approx 16,52$ *minutos*

Figura 8 - Sistema de 1 fila M/M/6 para abastecimento



Fonte: Autor (2025).

Na prática, a configuração com fila única tenderia a ultrapassar o limite físico do pátio. No entanto, do ponto de vista teórico, desconsiderando essa restrição de espaço, o resultado esperado seria o aumento do número médio de clientes simultaneamente no sistema e, consequentemente, do tempo médio de permanência.

Apesar disso, a taxa de bloqueio observada nesse modelo é inferior a dos demais arranjos analisados, o que indica um aproveitamento mais eficiente da capacidade total das seis bombas.

2) Abastecimento – 3 filas M/M/2 (três áreas independentes)

Nesse cenário, o posto é dividido em três subsistemas independentes, cada qual composto por duas bombas e uma fila própria. Essa subdivisão busca facilitar o direcionamento dos veículos e reduzir a sensação de congestionamento, sem abdicar de uma distribuição relativamente equilibrada dos atendimentos.

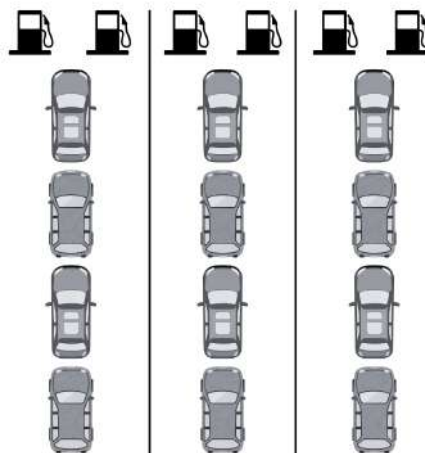
Parâmetros por grupo: $\lambda = 0,67$; $\mu = 0,2632$; $S = 2$; $K = 10$.

Resultados simulações:

- Utilização: $\rho = 1,27$
- Probabilidade de bloqueio por subsistema: $P_{\square} \approx 23\%$
- Taxa efetiva total: $\lambda_{ef} \approx 1,54$ veíc/min

- Tempo médio no sistema: **$W \approx 14,11$ minutos**

Figura 9 - Sistema de 3 filas M/M/2 para abastecimento



Fonte: Autor (2025).

Como cada subsistema possui capacidade física menor, o bloqueio tende a ocorrer mais cedo. Isso reduz o número médio de veículos simultaneamente no sistema e, conseqüentemente, diminui o tempo médio de permanência quando comparado ao arranjo M/M/6.

O resultado é um desempenho intermediário, que apresenta um equilíbrio razoável entre vazão do sistema e tempo médio de permanência.

3) Abastecimento – 6 filas M/M/1 (fila individual por bomba)

Este cenário reflete a estrutura mais comum em postos tradicionais, com uma fila por bomba. A independência entre filas acentua a variabilidade e dificulta o balanceamento entre servidores.

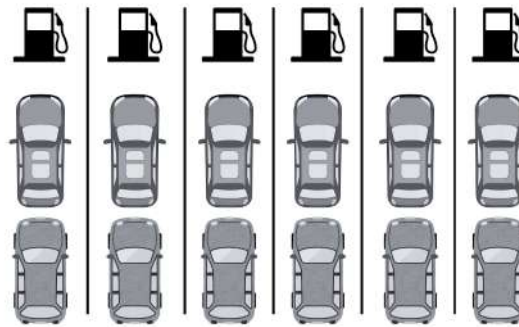
Parâmetros por bomba: $\lambda = 0,33$; $\mu = 0,2632$; $K = 5$.

Resultados simulações:

- Utilização: **$\rho \approx 1,25$**
- Probabilidade de bloqueio por bomba: **$P_{\square} \approx 27\%$**
- Taxa efetiva total: **$\lambda_e f \approx 1,44$ veíc/min**

- Tempo médio no sistema: **$W \approx 12,96$ minutos**

Figura 10 - Sistema de 6 filas M/M/1 para abastecimento



Fonte: Autor (2025).

A menor capacidade de cada fila aumenta a probabilidade de bloqueio, o que reduz o número médio de clientes atendidos, mesmo com um menor tempo médio entre os atendimentos. Além disso, a variabilidade entre as filas pode fazer com que alguns servidores fiquem ociosos enquanto outros operam sobrecarregados, comprometendo a eficiência global do sistema.

Síntese dos Cenários de Abastecimento

Os resultados mostram que existe um *trade-off* estrutural entre tempo médio no sistema, taxa de bloqueio e vazão efetiva.

No modelo M/M/6, observa-se a maior vazão efetiva (λ_{ef}) e menor taxa de bloqueio. Por outro lado, o tempo médio de permanência tende a ser maior, em razão do número mais elevado de clientes simultaneamente no sistema.

O arranjo 3×M/M/2 apresentou desempenho intermediário, com níveis moderados de bloqueio e tempo médio de permanência, configurando uma alternativa mais equilibrada.

Já o modelo 6×M/M/1 apresentou menor tempo médio de permanência entre os clientes efetivamente atendidos, porém com maior probabilidade de bloqueio e maior variabilidade entre as filas, o que pode comprometer a eficiência global.

Dessa forma, a escolha da arquitetura de filas deve estar alinhada à estratégia operacional adotada, seja ela voltada à maximização da vazão, à redução do tempo médio de permanência ou à minimização de perdas por bloqueio. Esses resultados fundamentam a simulação mais aprofundada e a análise econômica apresentadas nas próximas seções.

Cenários de Pagamento

1) Pagamento – 1 fila M/M/2 (fila única para dois caixas)

Após o abastecimento, os veículos ingressam em uma fila única atendida por dois caixas, configuração observada empiricamente em dois dos três dias de coleta.

Parâmetros por caixa: $\lambda = 1,10$; $\mu = 0,625$; $S = 2$.

Resultados simulações:

- Utilização: $\rho = 0,88$
- Tempo médio de espera na fila: **$W_q \approx 4,90$ minutos**
- Tempo médio total no sistema: **$W \approx 6,50$ minutos**

Figura 11 - Sistema de 1 fila M/M/2 para pagamento



Fonte: Autor (2025).

A adoção de fila única promove um balanceamento automático entre os servidores, reduzindo a variabilidade no atendimento e evitando a concentração de clientes em determinados pontos. Como resultado, os tempos de espera tornam-se significativamente menores.

Entre as alternativas analisadas, esse modelo apresentou o melhor desempenho para a etapa de pagamento, mostrando-se a configuração mais eficiente nesse contexto.

2) Pagamento – 2 filas M/M/1 (um caixa por fila)

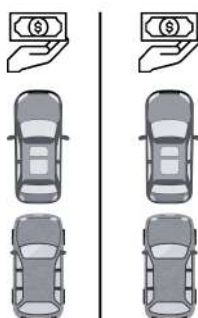
Cada caixa opera com sua própria fila, o que elimina o balanceamento entre servidores e amplifica a variabilidade dos tempos de espera.

Parâmetros por caixa: $\lambda = 0,55$; $\mu = 0,625$; $S = 1$.

Resultados simulações:

- Utilização: $\rho = 0,88$
- Tempo médio de espera na fila: **$Wq \approx 11,45$ minutos**
- Tempo médio total no sistema: **$W \approx 13,06$ minutos**

Figura 12 - Sistema de 2 filas M/M/1 para pagamento



Fonte: Autor (2025).

A separação das filas apresentou desempenho significativamente inferior, com tempos de espera superiores ao dobro dos observados no modelo M/M/2. Esse

resultado está relacionado ao fato de que cada fila individual passa a operar de forma independente, ficando mais exposta à variabilidade típica de sistemas M/M/1.

Essa dinâmica reduz a equidade entre os clientes e aumenta a probabilidade de desequilíbrio entre os caixas, fazendo com que alguns fiquem ociosos enquanto outros acumulam filas maiores.

4.3.3 INTERPRETAÇÃO DOS RESULTADOS

Os cenários simulados permitem compreender, de forma mais ampla, como a estrutura de atendimento e o número de servidores influenciam o desempenho operacional do posto de combustível. De maneira geral, os resultados indicam que, sob condições de demanda elevada ($\lambda = 2,00$ veículos por minuto), o sistema de abastecimento opera de forma recorrente acima de sua capacidade teórica, estimada em $S \cdot \mu \approx 1,58$ veículos por minuto, caracterizando um regime supercrítico.

Nessa situação, o crescimento das filas deixa de ser controlado pela própria dinâmica do atendimento e passa a ser limitado apenas pela restrição física do pátio, representada por $K = 30$ veículos. Como consequência, a ocorrência de perdas por bloqueio torna-se inevitável, uma vez que o sistema não consegue absorver integralmente a demanda.

Etapas de Abastecimento

Na etapa de abastecimento, ficou evidente que a organização das filas influencia diretamente tanto a taxa efetiva de atendimento quanto o tempo médio de permanência no sistema.

O modelo **M/M/6**, estruturado com fila única e acesso compartilhado às seis bombas, apresentou a maior taxa efetiva ($\lambda_{ef} \approx 1,58$ veículos por minuto) e a menor probabilidade de bloqueio (aproximadamente 21%). Em contrapartida, registrou o maior tempo médio de permanência ($W \approx 16,52$ minutos). Esse resultado está associado ao fato de que, em uma fila única, o sistema tende a manter um número maior de clientes simultaneamente em atendimento e em espera antes de atingir o limite físico do pátio.

O arranjo **3×M/M/2**, composto por três subsistemas paralelos, apresentou desempenho intermediário, com tempo médio próximo de 14,11 minutos e taxa efetiva total de aproximadamente 1,54 veículos por minuto. A divisão em subfilas reduz a competição global pelo espaço e antecipa o bloqueio em cada subsistema, diminuindo o número médio de clientes simultaneamente no sistema, o que contribui para a redução do tempo médio de permanência.

Por fim, o modelo **6×M/M/1**, caracterizado por filas independentes em cada bomba, apresentou o menor tempo médio de permanência entre os clientes atendidos ($W \approx 12,96$ minutos). Contudo, esse ganho ocorreu ao custo de maior probabilidade de bloqueio (cerca de 27%) e pior balanceamento entre servidores. Nessa configuração, as filas individuais atingem rapidamente sua capacidade máxima, impedindo a entrada de parte dos veículos. Assim, embora o tempo médio dos atendidos seja menor, a taxa efetiva total de atendimento também se reduz.

Etapas de Pagamento

Na etapa de pagamento, os resultados foram mais favoráveis. O modelo **M/M/2** com fila única apresentou desempenho equilibrado, com tempo médio total de aproximadamente 6,50 minutos e nível de utilização moderado ($\rho = 0,88$). Em contraste, o arranjo **2×M/M/1**, que distribui os clientes em duas filas independentes, registrou tempo médio total significativamente maior ($W \approx 13,06$ minutos) e tempo médio de espera de cerca de 11,45 minutos. Esse resultado evidencia a perda de eficiência decorrente da ausência de balanceamento automático entre os servidores.

Com base nessas análises, verifica-se que o principal gargalo operacional do posto está concentrado na etapa de abastecimento, cuja capacidade se mostra insuficiente para absorver o fluxo observado nos horários de pico. Já a etapa de pagamento, quando organizada em fila única com dois caixas, apresenta desempenho adequado e não contribui de forma relevante para a formação de filas excessivas.

Na seção seguinte (4.3.4 – Proposta de Sistema), são apresentadas simulações adicionais que consideram a ampliação do número de bombas para oito

unidades, permitindo estimar os efeitos dessas mudanças sobre o desempenho global e sobre a rentabilidade da operação.

4.3.4 PROPOSTA DE SISTEMA

Os resultados apresentados nas seções anteriores mostraram que o sistema atual de abastecimento, composto por seis bombas operando em arranjos equivalentes a $6 \times M/M/1$ e $3 \times M/M/2$, entra em regime de saturação nos horários de pico. As probabilidades de bloqueio variam entre 23% e 27%, o que implica perda direta de clientes, redução da taxa efetiva de atendimento e aumento do tempo médio de permanência no sistema. Esses indicadores apontam que a capacidade instalada é insuficiente para absorver o fluxo de demanda observado.

Diante desse cenário, torna-se necessária uma reconfiguração capaz de ampliar a capacidade de atendimento, reduzir a formação de filas e melhorar tanto a experiência do usuário quanto o desempenho econômico da operação.

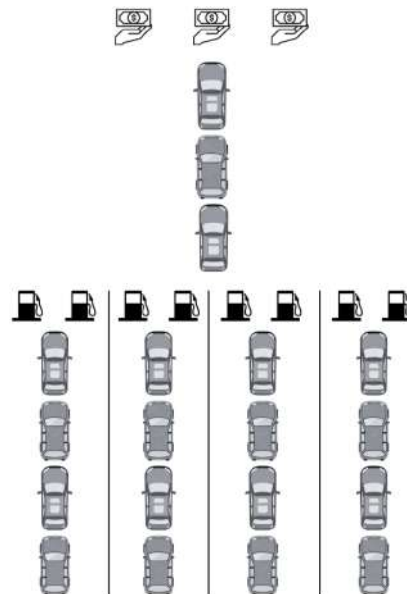
Considerando as limitações físicas do pátio, que inviabilizam a adoção de uma fila única centralizada ($M/M/6$) por questões de geometria e circulação interna, foi avaliada a implantação de um sistema com oito bombas, organizadas em quatro subsistemas $M/M/2$ ($4 \times M/M/2$). Esse arranjo é compatível com a dinâmica real do posto e permite aumentar a capacidade total de atendimento, reduzindo o risco de bloqueio precoce e favorecendo um melhor balanceamento entre os servidores.

Com a ampliação da capacidade de abastecimento de seis para oito bombas, a capacidade teórica do sistema passa de $S_{\mu} \approx 1,58$ para aproximadamente $S_{\mu} \approx 2,11$ veículos por minuto, representando um aumento de cerca de 26,6% no volume de veículos atendidos. Considerando que os veículos abastecidos seguem para a etapa de pagamento, foi adotado um aumento proporcional na taxa de chegada dessa etapa, que passa de $\lambda = 1,10$ veículos/min para aproximadamente $\lambda = 1,39$ veículos/min.

Com apenas dois caixas, o fator de utilização aumentaria de $\rho = 0,88$ para aproximadamente $\rho = 1,11$, caracterizando um novo gargalo no sistema. Dessa forma, foi proposta a ampliação para três caixas de pagamento, um sistema $M/M/3$, resultando em fator de utilização de aproximadamente $\rho = 0,74$ e tempo médio no sistema de cerca de 2,71 minutos, garantindo estabilidade operacional e

compatibilidade com a nova capacidade de abastecimento. Os cálculos detalhados dessa etapa serão apresentados no capítulo seguinte.

Figura 13 - Sistema proposto pelo autor



Fonte: Autor (2025).

4.3.5 DESEMPENHO ESPERADO DO SISTEMA PROPOSTO

Abastecimento 4×M/M/2:

A taxa de chegada no posto utilizada nas simulações corresponde ao cenário mais crítico observado em campo,

$$\lambda = 2,0 \text{ veículos/minuto.}$$

O tempo médio de serviço obtido experimentalmente foi de 3,8 minutos, o que implica:

$$\mu = \frac{1}{3,8} \approx 0,26 \text{ veículos/minuto}$$

Com oito bombas, a capacidade total do sistema torna-se:

$$S\mu = 8 \times 0,2632 \approx 2,11 \text{ veículos/minuto}$$

Assim, o índice de utilização para o sistema é:

$$\rho = \frac{\lambda}{s\mu} \approx \frac{2}{2,11} \approx 0,95$$

Indicando que o sistema passa a operar em regime estável, diferentemente dos cenários anteriores, nos quais $\rho > 1$. Ainda assim, devido à limitação física do pátio, o sistema pode atingir sua ocupação máxima antes de atender todos os clientes, resultando em uma probabilidade de bloqueio estimada em aproximadamente 10%.

$$P_k = 0,10$$

Assim o sistema atende 90% da demanda, ou seja, a taxa efetiva de atendimento é:

$$\lambda_{ef} = \lambda(1 - P_k) = 2,0 \times 0,9 = 1,8 \text{ veículos/min} = 108 \text{ veículos/hora}$$

Com grupos M/M/2 operando com $\rho = 0,95$, o tempo médio calculado com aproximação do modelo é:

$$W = 8,78 \text{ minutos}$$

Valor numericamente inferior aos obtidos nos cenários anteriores (12,96 e 14,11 min).

A divisão das bombas em módulos M/M/2 reduz a variabilidade das chegadas em cada subsistema, melhora o balanceamento entre as bombas e contribui para a redução do tempo médio de espera, tornando essa organização mais eficiente do que os arranjos anteriormente analisados.

Pagamento M/M/3:

A taxa de chegada na etapa de pagamento considera o cenário mais crítico do modelo sequencial, no qual abastecimento e pagamento ocorrem em etapas distintas. Nesse caso, tem-se inicialmente:

$$\lambda = 1,1 \text{ veículos/minuto.}$$

Com a ampliação da capacidade do sistema de abastecimento, a capacidade teórica passou de

$$S_{\mu} = 6 \times 0,2632 \approx 1,58 \text{ veículos/minuto}$$

para

$$S_{\mu} = 8 \times 0,2632 \approx 2,11 \text{ veículos/minuto}$$

Isso representa um aumento de aproximadamente 26,6% no volume de veículos atendidos. Assim, foi adotado um aumento proporcional na taxa de chegada da etapa de pagamento, que passa de $\lambda = 1,10$ veículos/min para aproximadamente:

$$\lambda \approx 1,39 \text{ veículos/min.}$$

Nessa condição, o sistema de pagamento com dois caixas (modelo M/M/2) passaria a operar com fator de utilização $\rho \approx 1,11$, caracterizando regime de saturação e formação de filas excessivas.

Dessa forma, foi proposto o aumento para três caixas, configurando um sistema M/M/3. Nesse novo arranjo, o fator de utilização passa a ser aproximadamente

$$\rho = \frac{\lambda}{S_{\mu}} \approx \frac{1,39}{3 \cdot 0,625} \approx 0,74$$

Com tempo médio de espera na fila de cerca de 1,11 minutos e tempo médio total no sistema de aproximadamente

$$W = 2,71 \text{ minutos.}$$

Somando as duas etapas, o tempo médio total no sistema torna-se aproximadamente 11,49 minutos,

$$W_{total} = 11,49 \text{ minutos.}$$

Dessa forma, o sistema de pagamento torna-se compatível com a nova capacidade do sistema de abastecimento, garantindo maior estabilidade operacional e redução do tempo total de atendimento dos clientes.

4.3.6 AVALIAÇÃO ECONÔMICA

O ticket médio de um abastecimento foi definido como:

$$TKM = R\$200,00$$

Com vazão efetiva de $\lambda_{ef} = 108$ veículos/hora a receita potencial por hora do sistema ampliado torna-se:

$$Receita/h = \lambda_{ef}/h \times TKM = 108 \times 200 = R\$21.600/h$$

O custo total de um funcionário de posto de combustível, considerando salário e encargos (INSS, FGTS, Sistema S, férias, 13º e provisões), foi estimado em:

$$Custo\ frentista = R\$14/h \text{ (fonte: FGV)}$$

De forma conservadora, foi considerado que cada bomba exige um frentista, e também foi adicionado um terceiro caixa, totalizando 3 funcionários adicionais, assim o custo horário da operação com oito bombas e 3 caixas é:

$$Custo/h = 3\text{ funcionários adicionais} \times 8\text{ horas} \times R\$14 = R\$336/h$$

O resultado operacional horário do sistema ampliado é:

$$Lucro\ Operacional/h = 21.600 - 336 = R\$21.264/h$$

Para avaliar o impacto da expansão, compara-se a receita por hora do cenário proposto (8 bombas) com a receita máxima possível no sistema atual (6 bombas), que atende: $\lambda_{ef} \approx 92,4$ veículos/hora resultando em receita aproximada de:

$$18.480 = 92,4 \times 200.$$

Assim, o incremento direto na receita por hora devido à expansão é:

$$\Delta Receita/h = 21.600 - 18.480 = R\$3.120/h$$

Esse aumento representa aumento do faturamento nos horários de pico, evidenciando a atratividade financeira da ampliação de seis para oito bombas.

4.4 SÍNTESE E RECOMENDAÇÃO FINAL

A comparação técnica e econômica dos três cenários analisados permite concluir que a configuração $4 \times M/M/2$, com oito bombas, representa a alternativa mais eficiente e operacionalmente robusta para o posto estudado. Trata-se do único arranjo capaz de:

- eliminar o bloqueio de entrada,
- atender integralmente a demanda de pico,
- reduzir substancialmente o tempo médio de permanência,
- maximizar a receita por hora,
- manter a operação financeiramente viável com acréscimo moderado de custos de mão de obra.

O modelo $3 \times M/M/2$ apresentou desempenho intermediário. Embora seja superior ao arranjo atual $6 \times M/M/1$ em termos de balanceamento e redução de variabilidade, ainda se mostra insuficiente para absorver a taxa de chegada observada nos horários de maior fluxo, operando com $\rho > 1$.

O cenário $6 \times M/M/1$, por sua vez, apesar de apresentar menor tempo médio entre os clientes efetivamente atendidos, registra elevada probabilidade de bloqueio (aproximadamente 27%), redução significativa da vazão efetiva (cerca de 1,44 veículos por minuto) e forte desequilíbrio entre os servidores. Esses fatores resultam em desempenho global inferior, tanto sob o ponto de vista operacional quanto econômico.

Diante desse conjunto de resultados, recomenda-se a expansão do sistema para oito bombas, organizadas em quatro subsistemas $M/M/2$. Inicialmente, considerou-se a manutenção do sistema de pagamento em fila única com dois caixas ($M/M/2$), porém, com o aumento da capacidade de abastecimento e, conseqüentemente, do número de veículos atendidos por minuto, verificou-se que essa configuração passaria a operar próxima da saturação.

Assim, para evitar que o gargalo do sistema fosse deslocado para a etapa de pagamento, propôs-se a ampliação para três caixas, configurando um sistema $M/M/3$ em fila única. Essa alteração permite manter o fator de utilização em níveis

estáveis, reduzir o tempo médio de espera e garantir compatibilidade entre a capacidade das etapas de abastecimento e pagamento.

Dessa forma, a configuração proposta, oito bombas organizadas em quatro subsistemas M/M/2 e pagamento em fila única M/M/3, apresenta maior robustez operacional, melhor aproveitamento da capacidade instalada e redução da formação de filas, contribuindo para a melhoria da experiência do cliente e para o aumento de aproximadamente 17% do desempenho econômico do posto nos horários de pico. Os resultados por sistema de abastecimento estão disponíveis na Tabela 16.

Tabela 16 - Comparação entre os sistemas de abastecimento.

Indicador	6×M/M/1	3×M/M/2	4×M/M/2 (Proposto)
Utilização (ρ)	1,25	1,27	0,95
Probabilidade de bloqueio (P_0)	0,27	0,23	0,10
Vazão efetiva (veíc/min)	1,44	1,54	1,80
Vazão efetiva (veíc/h)	86,40	92,40	108,00
Tempo médio W	12,96	14,11	8,78
Receita por hora (R\$)	17.280	18.480	21.600
Custo adicional/h (frentistas)			336
Resultado	Baixa eficiência	Intermediário	Melhor desempenho e maior rentabilidade

Fonte: Autor (2025).

5. DISCUSSÃO DOS RESULTADOS

Os resultados obtidos neste estudo reforçam a aplicabilidade da Teoria de Filas como ferramenta analítica para compreender o desempenho de sistemas de serviço sujeitos à variabilidade da demanda. Embora o posto de combustível tenha sido utilizado como estudo de caso, seu papel foi ilustrativo, servindo como ambiente real para aplicar e testar os modelos teóricos. Os comportamentos observados, como saturação, bloqueio, influência da capacidade instalada e impacto da estrutura das filas, são características próprias de sistemas M/M/S/K e não se restringem ao setor de combustíveis.

As simulações mostraram que, quando a taxa de chegada supera a capacidade agregada do sistema ($\lambda > S \cdot \mu$), a tendência é de saturação, com crescimento acelerado das filas e dependência do limite físico (K) para conter o aumento do número de clientes no sistema. Esse comportamento, amplamente descrito na literatura, foi claramente identificado nas configurações com seis servidores analisadas no estudo.

Outro ponto importante foi a influência da estrutura das filas no desempenho global. A comparação entre fila única (M/M/6), arranjos intermediários (3×M/M/2) e filas independentes (6×M/M/1) evidenciou que o compartilhamento das chegadas reduz a variabilidade entre servidores, melhora o aproveitamento da capacidade e diminui a probabilidade de bloqueio. Mesmo com a mesma capacidade total, o desempenho varia significativamente conforme a forma de organização das filas, aspecto discutido por autores como Gross e Harris (1998) e Fitzsimmons e Fitzsimmons (2014).

Na etapa de pagamento, a análise também evidenciou diferenças importantes entre as configurações de filas. Nos cenários iniciais, o modelo M/M/2 e o arranjo 2×M/M/1 apresentavam a mesma taxa de utilização, porém com tempos médios de espera bastante distintos. A divisão das chegadas em filas independentes aumentava a variabilidade e reduzia o equilíbrio do sistema, enquanto a fila única proporcionava melhor desempenho, confirmando um princípio clássico da teoria das filas.

Entretanto, com a ampliação da capacidade do sistema de abastecimento, houve aumento proporcional do fluxo de veículos direcionados à etapa de pagamento, tornando insuficiente a configuração com dois caixas. Dessa forma, foi necessária a ampliação para três caixas, configurando um sistema M/M/3 em fila única, capaz de manter o fator de utilização em níveis estáveis e evitar a formação de um novo gargalo operacional.

A análise econômica complementou esses resultados ao mostrar como o dimensionamento do sistema influencia diretamente a vazão efetiva (λ_{ef}), a probabilidade de bloqueio ($P_{\square}$) e o tempo médio de permanência (W). A ampliação para oito bombas, organizadas em $4 \times M/M/2$, associada ao pagamento em M/M/3, mostrou-se a alternativa mais equilibrada, capaz de estabilizar o sistema, reduzir perdas e melhorar o desempenho financeiro em até 17% nos horários de pico, sem aumento desproporcional de custos. Ainda que obtido no contexto específico do posto estudado, esse resultado ilustra um princípio geral: existe um ponto de dimensionamento que equilibra eficiência operacional e viabilidade econômica.

Por fim, o *script* desenvolvido em Python (Apêndice A) demonstrou ser uma ferramenta útil para apoiar a análise, permitindo simular diferentes combinações de λ , μ , S e K e comparar cenários de forma ágil. A integração entre modelagem teórica e simulação computacional proporcionou uma base quantitativa consistente para a tomada de decisão, evidenciando como o uso estruturado da Teoria de Filas pode contribuir para melhorar a eficiência e a rentabilidade de sistemas de serviço.

6. CONCLUSÃO

Este trabalho teve como objetivo demonstrar como a Teoria de Filas pode ser aplicada na análise e melhoria de sistemas de atendimento sujeitos à variabilidade da demanda. Para isso, utilizou-se um posto de combustível como estudo de caso, exclusivamente como ambiente real para coleta de dados e validação das métricas teóricas. O foco principal do estudo esteve nos fundamentos matemáticos e nos modelos que descrevem o comportamento de sistemas M/M/S/K, independentemente do setor analisado.

Os cálculos realizados permitiram identificar de forma clara as condições que levam à saturação, ao bloqueio e à perda de eficiência operacional. Verificou-se que, quando a capacidade agregada é inferior à taxa média de chegada, o sistema tende à instabilidade, como consequência direta da relação entre λ , μ e S. Também ficou evidente que a estrutura das filas exerce influência significativa sobre o desempenho. Modelos com compartilhamento de filas apresentaram tempos de espera menores e melhor aproveitamento da capacidade quando comparados a arranjos com filas independentes, confirmando resultados já consolidados na literatura.

Outro ponto importante foi o uso da linguagem de programação Python, que contribuiu tanto para simular os resultados dos modelos quanto para a organização metodológica do estudo. As rotinas implementadas permitiram automatizar cálculos, testar diferentes configurações e avaliar métricas como tempo médio, probabilidade de bloqueio e vazão efetiva com agilidade e precisão. Além disso, o uso da ferramenta facilita a reprodução dos cenários analisados e a ampliação do estudo para novas condições.

A aplicação aos dados coletados mostrou como a Teoria de Filas pode orientar decisões de dimensionamento e alocação de recursos. No caso analisado, a ampliação para oito servidores demonstrou ser uma alternativa capaz de reduzir bloqueios e melhorar o desempenho sem aumento desproporcional de custos.

Conclui-se que a Teoria de Filas constitui uma ferramenta consistente e aplicável para a análise de sistemas de serviço, especialmente quando integrada a sistemas que automatizam os resultados de diferentes cenários. Essa abordagem pode ser replicada em diferentes contextos, como call centers, clínicas, serviços públicos e operações logísticas.

Como continuidade do estudo, sugere-se a incorporação de modelos com variação temporal da demanda, sistemas com prioridades e distribuições mais gerais, além da integração com técnicas de simulação discreta e otimização estocástica, ampliando ainda mais as possibilidades de aplicação em ambientes mais complexos.

REFERÊNCIAS

ANDRADE, Eduardo Leopoldino. Introdução à pesquisa operacional: métodos e modelos para análise das decisões. 4. ed. Rio de Janeiro: LTC, 2009.

ALMEIDA, A.; ARAÚJO, J.; BARROS, M.; SAMPAIO, R. Cálculo de probabilidades. Belo Horizonte: Editora Poisson, 2019. Disponível em: [link](#). Acesso em: 9 fev. 2026.

ARAÚJO, J. A importância da estatística e suas aplicações. Fortaleza: Universidade Federal do Ceará, 2015. Disponível em: [link](#). Acesso em: 9 fev. 2026.

BANKS, Jerry; CARSON, John S.; NELSON, Barry L.; NICOL, David M. Discrete-event system simulation. 5. ed. Upper Saddle River: Pearson, 2010.

BHAT, U. Narayan. An introduction to queueing theory: modeling and analysis in applications. 2. ed. Boston: Birkhäuser, 2015.

BOCHAROV, P. P. et al. Modern probability and statistics: queueing theory. Boston: VSP, 2004.

BRANDÃO, J.; GOMES, L.; STELMASTCHUK, T.; SANTOS, R. Teoria de filas aplicada a panificadora de um supermercado. In: SIMPÓSIO DE ENGENHARIA DE PRODUÇÃO, XIX, 2012. Anais... 2012. Disponível em: [link](#). Acesso em: 9 fev. 2026.

COSTA, A. Teoria das filas. São Luís: Universidade Federal do Maranhão, [ano]. Disponível em: [link](#). Acesso em: 9 fev. 2026.

DUARTE, J. Uso das distribuições Poisson, Poisson-Gama, Poisson-Inversa Gaussiana e Poisson-Lindley generalizada para dados de contagem. Fortaleza: Universidade Federal do Ceará, 2021. Disponível em: [link](#). Acesso em: 9 fev. 2026.

ERLANG, A. K. The theory of probabilities and telephone conversations. Nyt Tidsskrift for Matematik, Copenhagen, v. 20, p. 33–39, 1909.

FITZSIMMONS, James A.; FITZSIMMONS, Mona J. Administração de serviços: operações, estratégia e tecnologia da informação. 7. ed. Porto Alegre: AMGH, 2014.

FUNDAÇÃO GETULIO VARGAS. Escola de Economia de São Paulo. Custo do trabalho no Brasil: relatório final. São Paulo: FGV/EESP, 2017. Disponível em: [link](#). Acesso em: 9 fev. 2026.

GROSS, Donald; HARRIS, Carl M. Fundamentals of queueing theory. 4. ed. Hoboken: John Wiley & Sons, 2008.

HILLIER, Frederick S.; LIEBERMAN, Gerald J. Introdução à pesquisa operacional. 9. ed. Porto Alegre: AMGH, 2013.

MARINS, Fernando A. S. Introdução à pesquisa operacional. São Paulo: Cultura Acadêmica, 2011.

PINHEIRO, R. As práticas do cotidiano na relação oferta e demanda dos serviços de saúde: um campo de estudo e construção da integralidade. Rio de Janeiro: CEPESC; IMS/UERJ; ABRASCO, 2009. Disponível em: [link](#). Acesso em: 9 fev. 2026.

PINTO, Ângelo Santos. Aplicação da teoria de filas na análise da capacidade operacional de um sistema: estudo de caso BCA Porto Novo. 2011. Trabalho de Conclusão de Curso. Disponível em: [link](#). Acesso em: 9 fev. 2026.

PRADO, Darci Santos. Teoria das filas e da simulação. Nova Lima: INDG, 2006.

SABBADINI, Fernando; GONÇALVES, A. A.; OLIVEIRA, M. J. F. Gestão da capacidade de atendimento e simulação computacional para a melhoria na alocação de recursos e no nível de serviço em hospitais. In: SEGET – Simpósio de Excelência em Gestão e Tecnologia, III, 2006.

SHANNON, Robert E. Simulation: a survey with research suggestions. AIIE Transactions, v. 7, n. 3, p. 289–301, 1975.

TAHA, Hamdy A. Pesquisa operacional: uma introdução. 9. ed. São Paulo: Pearson, 2017.

STEVENSON, William J. Operations management. 13. ed. New York: McGraw-Hill, 2018.

APÊNDICES

Código Python da função que faz as simulações de fila M/M/S/K:

```
import random
import math
from collections import deque

def simular_fila_mm_s(lambd, tempo_medio_atend, s, T=50000,
warmup=5000, K=None, seed=1):
    """
    Simula M/M/s (se K=None) ou M/M/s/K (se K=int).

    Entradas:
        lambd: taxa de chegada (clientes por unidade de tempo)
        tempo_medio_atend: tempo médio de atendimento (unidade de tempo)
        s: número de servidores
        T: tempo total de simulação
        warmup: tempo inicial ignorado nas métricas (aquecimento)
        K: capacidade total do sistema (fila + atendimento). None =
infinito
        seed: semente (reprodutibilidade)

    Saídas (dict):
        rho, lambda_eff, p_block, W, Wq, L, Lq, p0, contadores
    """
    if seed is not None:
        random.seed(seed)

    mu = 1.0 / tempo_medio_atend

    # Estado
    t = 0.0
    n = 0          # no sistema (fila + atendimento)
    nq = 0         # na fila
    busy = 0       # servidores ocupados

    # Próxima chegada
    t_arr = random.expovariate(lambd)

    # Saídas: lista com tempos de término de serviço
```

```

dep_times = []

# Filas de tempos (pra calcular W e Wq)
fila = deque()          # armazena arrival_time de quem espera
em_servico = []         # lista de tuplas (arrival, start, dep)

# Contadores
arrivals = admitted = blocked = completed = 0

# Acumuladores (médias no tempo) pós-warmup
last_t = 0.0
area_L = 0.0
area_Lq = 0.0
area_busy = 0.0
area_p0 = 0.0

# Tempos médios (amostrais) pós-warmup
sum_W = 0.0
sum_Wq = 0.0
dep_count = 0

def acumular_areas(t_novo):
    nonlocal last_t, area_L, area_Lq, area_busy, area_p0
    a = max(last_t, warmup)
    b = min(t_novo, T)
    if b > a:
        dt = b - a
        area_L += n * dt
        area_Lq += nq * dt
        area_busy += busy * dt
        area_p0 += (1 if n == 0 else 0) * dt
    last_t = t_novo

while t < T:
    t_dep = min(dep_times) if dep_times else math.inf
    t_event = min(t_arr, t_dep, T)

    acumular_areas(t_event)
    t = t_event

```

```

if t >= T:
    break

# Chegada
if t_arr <= t_dep:
    arrivals += 1
    t_arr = t + random.expovariate(lambd)

    # Bloqueio (capacidade finita)
    if K is not None and n >= K:
        blocked += 1
        continue

    admitted += 1
    n += 1

    # Entra direto em serviço
    if busy < s:
        busy += 1
        start = t
        dep = t + random.expovariate(mu)
        dep_times.append(dep)
        em_servico.append((t, start, dep))
    else:
        # Vai pra fila
        nq += 1
        fila.append(t)

# Saída
else:
    dep_times.remove(t_dep)
    completed += 1
    n -= 1

    # Pega o registro correspondente pra medir W e Wq
    idx = None
    for i, rec in enumerate(em_servico):
        if abs(rec[2] - t_dep) < 1e-12:
            idx = i

```

```

        break
    if idx is not None:
        a_time, s_time, _ = em_servico.pop(idx)
    else:
        a_time, s_time = t, t # fallback

    if t >= warmup:
        sum_W += (t - a_time)
        sum_Wq += (s_time - a_time)
        dep_count += 1

    # Se tem fila, puxa o próximo
    if nq > 0:
        nq -= 1
        a2 = fila.popleft()
        start2 = t
        dep2 = t + random.expovariate(mu)
        dep_times.append(dep2)
        em_servico.append((a2, start2, dep2))
    else:
        busy -= 1

tempo_util = T - warmup

L = area_L / tempo_util
Lq = area_Lq / tempo_util
busy_med = area_busy / tempo_util
rho = (busy_med / s)
p0 = area_p0 / tempo_util

lambda_eff = dep_count / tempo_util if tempo_util > 0 else 0.0

W = (sum_W / dep_count) if dep_count > 0 else float("nan")
Wq = (sum_Wq / dep_count) if dep_count > 0 else float("nan")

p_block = (blocked / arrivals) if arrivals > 0 else 0.0

rho_ofertado = lambda / (s * mu)

```

```

return {
    "modelo": f"M/M/{s}" if K is None else f"M/M/{s}/{K}",
    "lambda": lambda,
    "mu": mu,
    "s": s,
    "K": K,
    "rho": rho,
    "p0": p0,
    "p_block": p_block,
    "lambda_eff": lambda_eff,
    "W": W,
    "Wq": Wq,
    "L": L,
    "Lq": Lq,
    "arrivals": arrivals,
    "admitted": admitted,
    "blocked": blocked,
    "completed": completed,
    "rho_ofertado": rho_ofertado,
    "rho_realizado": rho,
}

def print_metricas(m):
    print(f"\n=== {m['modelo']} ===")
    print(f"lambda (chegadas): {m['lambda']:.4f}")
    print(f"mu (atend): {m['mu']:.4f} (tempo médio = {1/m['mu']:.4f})")
    print(f"servidores (s): {m['s']}")
    print(f"capacidade (K): {m['K']}")
    print(f"rho ofertado ( $\lambda/S_p$ ): {m['rho_ofertado']:.4f}")
    print(f"rho realizado: {m['rho_realizado']:.4f}")
    print(f"P0 (sist vazio): {m['p0']:.4f}")
    print(f"prob bloqueio: {m['p_block']:.4f}")
    print(f"lambda efetiva: {m['lambda_eff']:.4f}")
    print(f"W (tempo no sistema): {m['W']:.4f}")
    print(f"Wq (tempo na fila): {m['Wq']:.4f}")
    print(f"L (no sistema): {m['L']:.4f}")

```

```
print(f"Lq (na fila):          {m['Lq']:.4f}")  
    print(f"contadores:      arrivals={m['arrivals']}  
admitted={m['admitted']}      blocked={m['blocked']}  
completed={m['completed']}")  
  
.+-
```