Análise dos modelos de Filas M/G/1 e M/G/1/K com desestímulo

1st Diana Laura Fernández Duarte
Instituto Nacional de Telecomunicações (INATEL)
Santa Rita do Sapucaí, Brazil
diana.duarte@inatel.br

2nd Jose G. C. Sanchez
Instituto Nacional de Telecomunicações (INATEL)
Santa Rita do Sapucaí, Brazil
jose.sanchez@inatel.br

***Resumo*—Neste projeto, analisa-se o desempenho de sistemas de filas M/G/1 e M/G/1/K com chegadas desestimuladas. À medida que mais clientes entram na fila, a taxa de chegadas é reduzida, enquanto a taxa de saída permanece constante. Para modelar os tempos de serviço, consideram-se diferentes distribuições: exponencial, uniforme e determinística, com o objetivo de avaliar o impacto da variabilidade do tempo de serviço no desempenho das filas. Os resultados obtidos por meio de simulações demonstram que o desestímulo atua como um mecanismo de regularização que reduz a congestão do sistema. O código-fonte está disponível publicamente[1].**

I. Introdução

A teoria das filas é aplicada em uma ampla gama de modelos de negócios, incluindo redes de telecomunicações, servidores de dados, modelagem do tráfego e sistemas logísticos de produção. Dado que, na prática, os clientes costumam abandonar a fila quando a espera se prolonga excessivamente, torna-se necessário tomar ações corretivas que contribuam para reduzir a congestão do sistema [1]. Para isso, analisa-se seu comportamento dinâmico, avaliando métricas-chave como os tempos de espera, o número médio de clientes no sistema e o fator de utilização dos recursos, a fim de otimizar o desempenho e garantir um serviço de qualidade [2].

Na modelagem clássica de sistemas de filas, assume-se que a taxa de chegada é constante. No entanto, em situações reais, os clientes podem se desmotivar e não entrar na fila devido ao seu tamanho e à longa espera que isso implica. Esse fenômeno é conhecido como desestímulo, e se reflete em uma taxa de chegada decrescente com relação ao número de clientes no sistema [1], [3]. Por exemplo, em um sistema informático com processamento em lotes, as solicitações de trabalho são desincentivadas quando o sistema está sobrecarregado. No caso do tráfego veicular, quando uma via está altamente congestionada, os motoristas costumam desviar sua rota e buscar uma via alternativa menos saturada. De forma semelhante, em um restaurante, quando o local está completamente cheio, muitos clientes optam por procurar outro estabelecimento para comer em vez de esperar que uma mesa seja desocupada, especialmente quando há várias opções disponíveis na região. Em todos esses casos, o número de usuários no sistema influencia diretamente na taxa de novas chegadas, o que caracteriza um sistema de filas com desestímulo.

[1]https://github.com/Diana9908/TP547

O objetivo deste trabalho é analisar os modelos de fila M/G/1 e M/G/1/K com desestímulo, considerando diferentes distribuições para os tempos de serviço, especificamente as distribuições exponencial, uniforme e determinística. A Seção II deste projeto descreve o modelo do sistema proposto e a Seção III apresenta os principais resultados obtidos a partir de simulações. Por fim, a Seção IV apresenta as conclusões alcançadas.

II. Modelo do Sistema

A seguir, são descritos os modelos de sistemas de filas M/G/1 e M/G/1/K com chegadas desestimuladas em função do número de clientes presentes no sistema. Consideram-se as principais entidades envolvidas, cliente e servidor, e as variáveis de estado que permitem avaliar o desempenho do sistema em qualquer instante: fator de utilização ρ, tempo médio de permanência no sistema $E[tq]$, número médio de clientes no sistema $E[q]$, tempo médio de permanência na fila $E[tw]$, número médio de clientes na fila $E[w]$, e probabilidade de bloqueio P_b. Os tempos de serviço seguem diferentes distribuições: exponencial, uniforme e determinística.

A. M/G/1 com desestímulo

Em um sistema de filas M/G/1 com desestímulo, considera-se um único servidor, capacidade infinita e uma taxa de chegada variável dada por:

$$\lambda_k = \frac{\lambda_0}{k+1}, \quad k \geq 0 \tag{1}$$

onde k representa o número de clientes presentes no sistema e λ_0 é uma constante real positiva que corresponde à taxa máxima de chegada quando o sistema está completamente vazio.

O tempo de serviço de cada cliente segue uma distribuição geral. Dado que a taxa de serviço se mantém constante, pode-se considerar:

$$\mu_k = \mu, \quad k \geq 1 \tag{2}$$

Para descrever mais detalhadamente o funcionamento deste sistema, a Figura 1 modela a lógica de um evento de chegada. Se o servidor estiver livre, o cliente passa diretamente a ocupá-lo, sendo necessário gerar um novo tempo de serviço e programar uma nova partida. Caso contrário, o cliente se junta à fila. Dado que o número de clientes no sistema aumentou,

é preciso atualizar a taxa de chegada, gerar um novo tempo de chegada com base nessa nova taxa e agendar o próximo evento de chegada. Em seguida, são atualizadas as métricas relevantes e o relógio de simulação avança para o próximo evento.

A lógica do evento de partida é apresentada na Figura 3. Se houver clientes na fila, o primeiro da fila passa a ocupar o servidor e a fila é reduzida em uma unidade. Gera-se o tempo de serviço correspondente a esse novo cliente e programa-se sua partida. Caso não haja clientes na fila, o servidor passa a ficar desocupado. Como a saída de um cliente reduz a ocupação do sistema, estimulando a taxa de chegada, é necessário atualizar essa taxa e programar um novo evento de chegada. Por fim, as estatísticas do sistema são atualizadas e passa-se à execução do próximo evento.

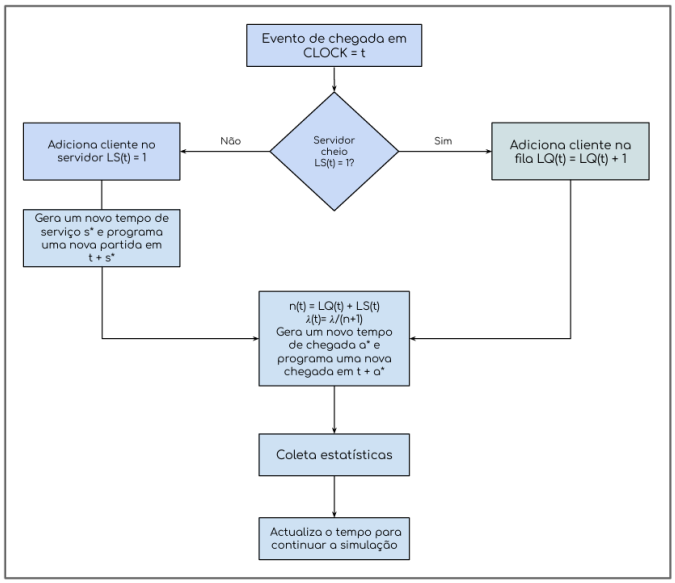

Fig. 1. Fluxograma de chegada para fila M/G/1.

1) M/M/1 com desestímulo: Um sistema com modelo de filas M/M/1 constitui um caso particular em que os tempos de serviço seguem uma distribuição exponencial. O diagrama de transição de estado desse sistema é apresentado na Figura 1.

A probabilidade estacionária de estar no estado *k* pode ser expressa como:

$$p_k = p_0 \left(\frac{\lambda_0}{\mu}\right)^k \frac{1}{k!} \tag{3}$$

Dado que $\sum_{k=0}^{\infty} p_k = 1$, a probabilidade estacionária do sistema estar vazio p_0, é dada por:

$$\begin{aligned} p_0 &= \frac{1}{1 + \sum_{k=1}^{\infty} \left(\frac{\lambda_0}{\mu}\right)^k \frac{1}{k!}} \\ p_0 &= e^{-\lambda_0/\mu} \end{aligned} \tag{4}$$

Consequentemente, o fator de utilização do servidor ρ é expresso como:

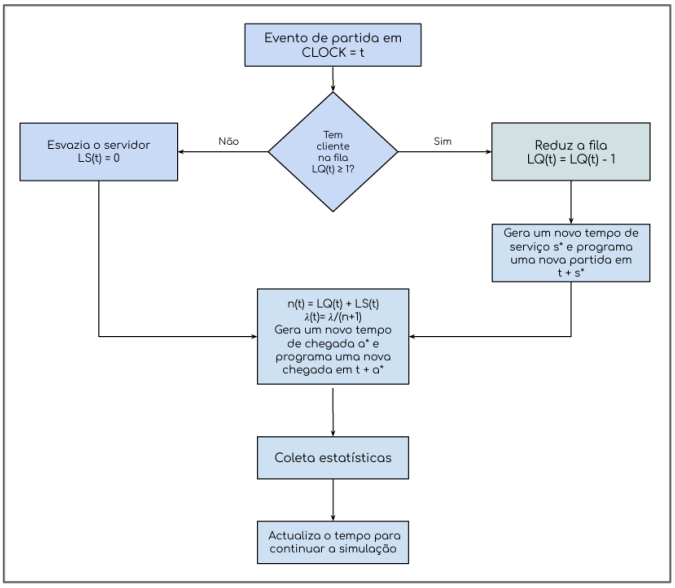

Fig. 2. Fluxograma de partida para fila M/G/1 e M/G/1/K.

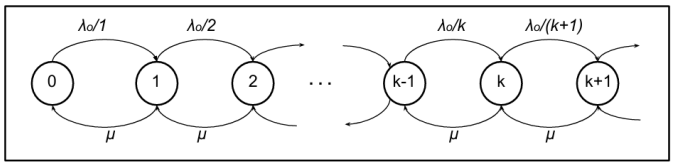

Fig. 3. Diagrama de taxa de transição de estado para fila M/M/1.

$$\rho = 1 - e^{-\lambda_0/\mu} \tag{5}$$

O número médio de clientes no sistema é dado por:

$$E[q] = \frac{\lambda_0}{\mu} \tag{6}$$

Aplicando a Lei de Little, o tempo médio que um cliente passa no sistema pode ser expresso como:

$$E[t_q] \ = \frac{\lambda_0}{\mu^2(1 - e^{-\lambda_0/\mu})} \tag{7}$$

B. M/G/1/K com desestímulo

Em uma fila M/M/1/K com desestímulo a capacidade do sistema é limitada de duas formas: por um lado, de maneira implícita, devido ao fato de que a taxa de chegada diminui progressivamente à medida que o número de clientes no sistema aumenta, desestimulando novas chegadas; por outro lado, de forma explícita, por uma capacidade máxima *K*, que impede a entrada de novos clientes assim que esse limite é atingido.

O tempo de serviço de cada cliente segue uma distribuição geral. As taxas de chegada e de saída continuam sendo descritas pelas Equações (1) e (2), respectivamente, com a diferença de que a taxa de chegada é definida para $0 \leq k \leq K$, enquanto a taxa de saída é válida apenas para $1 \leq k \leq K$.

Quando um novo cliente chega, deve-se primeiramente verificar se o sistema está cheio; se estiver, o cliente é bloqueado. Caso contrário, segue-se o mesmo fluxo descrito para o sistema sem limite de capacidade, conforme ilustrado na Figura 4. Os eventos de partida por sua vez, seguem a mesma lógica aplicada à fila M/G/1 com desestímulo, conforme ilustrado na Figura 2.

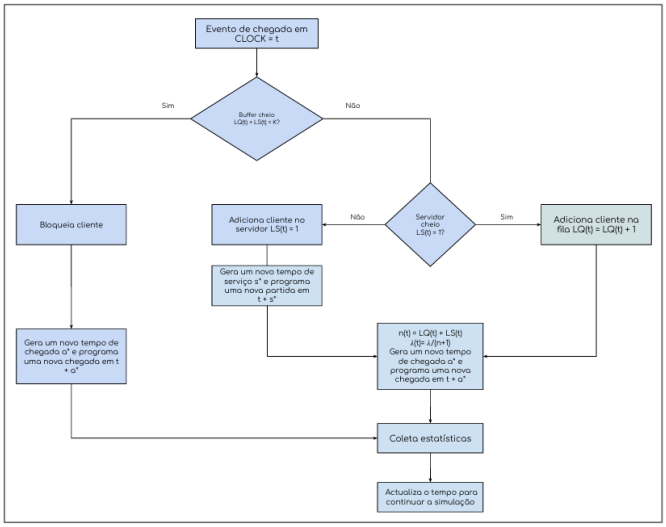

Fig. 4. Fluxograma de chegada para fila M/G/1/K.

1) M/M/1/K con desestímulo: Em particular, quando os tempos de serviço seguem uma distribuição exponencial, o sistema M/G/1/K com desestímulo se reduz a uma fila M/M/1/K com desestímulo. Neste caso, o processo pode ser descrito como uma cadeia de nascimento e morte apresentado na Figura 5.

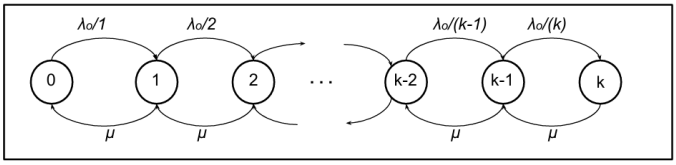

Fig. 5. Diagrama de taxa de transição de estado para fila M/M/1/K.

III. Resultados

A seguir, são apresentados os resultados obtidos por meio de simulações para avaliar o desempenho dos modelos de filas M/G/1 e M/G/1/K com chegadas desestimuladas. Para isso, analisam-se diferentes métricas, incluindo o fator de utilização, o tempo médio de permanência e o número médio de clientes, além da probabilidade de bloqueio no caso de sistemas com capacidade finita. No caso particular de uma fila M/M/1, os resultados simulados são comparados com a formulação teórica, a fim de validar a precisão do modelo de simulação. Todas as simulações consideram uma taxa de partida $\mu = 12$.

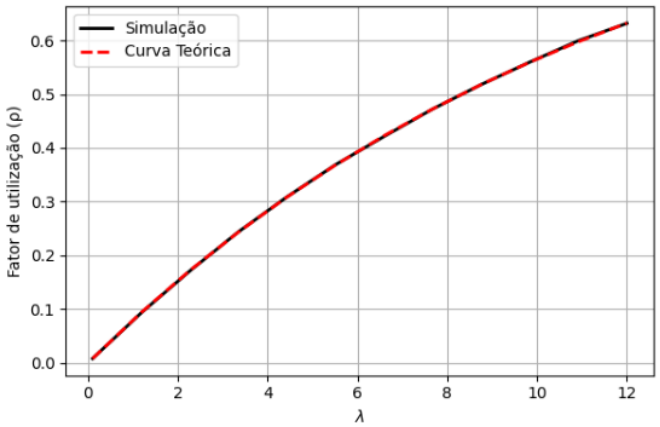

Fig. 6. Fator de utilização (ρ) em função de λ_0 em uma fila M/M/1 com desestímulo.

A. M/M/1 com desestímulo

Na Figura 6 é apresentado o comportamento do fator de utilização ρ em função da taxa máxima de chegadas λ_0. Observa-se que, à medida que λ_0 aumenta, o fator de utilização ρ também cresce.

No entanto, esse crescimento não é linear. Para valores baixos de λ_0, o sistema se comporta de maneira muito semelhante a uma fila M/M/1 convencional. Por outro lado, para valores altos de λ_0, o sistema tende a operar de forma mais congestionada, tornando mais evidentes os efeitos do desestímulo. Como consequência, a taxa efetiva de chegadas cresce mais lentamente. Isso gera uma forma de autorregulação que evita a saturação do servidor. Por exemplo, para $\lambda_0 = 10$, a taxa de ocupação do servidor é de 0,56, o que indica que o sistema ainda opera em regime estável. Os resultados obtidos por meio da simulação coincidem perfeitamente com a formulação teórica, o que valida a robustez da simulação.

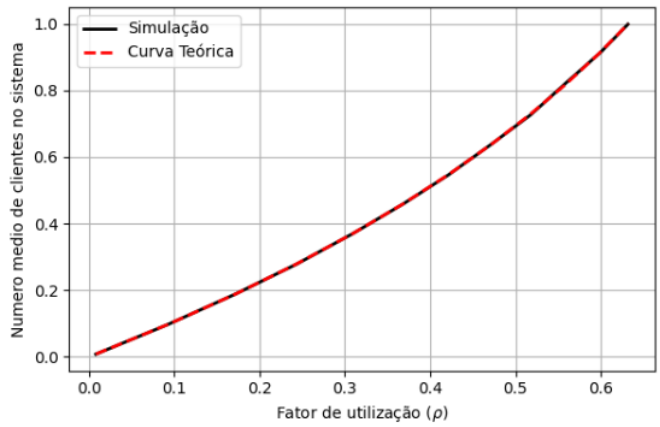

Fig. 7. Número médio de clientes no sistema $E(q)$ em função de ρ em uma fila M/M/1 com desestímulo.

A Figura 7 mostra a relação crescente entre o número médio de clientes no sistema $E(q)$ e o fator de utilização ρ. À medida que o servidor se torna mais ocupado, pequenos aumentos na

utilização geram uma acumulação mais rápida de clientes no sistema. Da mesma forma, o tempo médio de permanência no sistema aumenta à medida que o fator de utilização cresce, conforme mostrado na Figura 8.

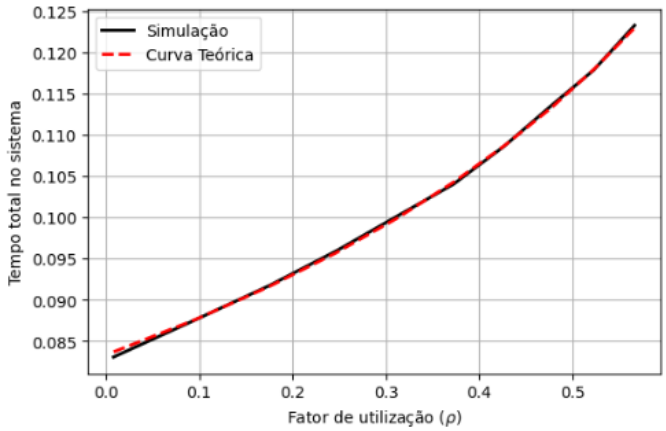

Fig. 8. Tempo médio de permanência no sistema $E[t_q]$ em função de ρ em uma fila M/M/1 com desestímulo.

A Figura 9 mostra uma comparação entre uma fila M/M/1 clássica e uma fila M/M/1 com desestímulo. Para isso, analisa-se o comportamento do fator de utilização em função da taxa de chegada. No sistema clássico, ρ cresce linearmente, de modo que o servidor permanece ocupado 100% do tempo quando $\lambda = \mu$. Por outro lado, no modelo com desestímulo, como a taxa de chegada diminui à medida que aumenta a congestão, o sistema se autorregula e o fator de utilização é aproximadamente 0,66 quando $\lambda = \mu$.

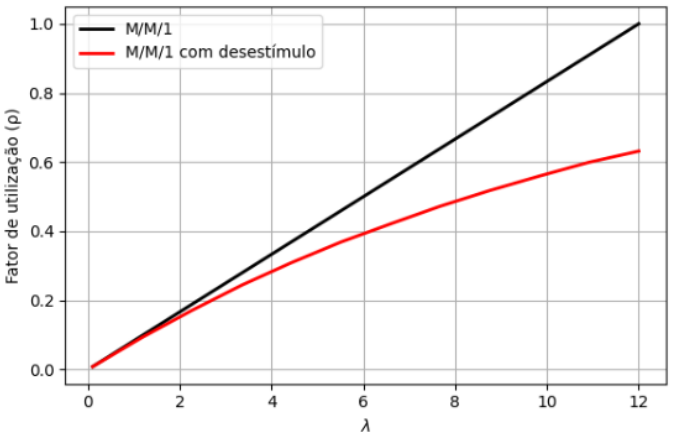

Fig. 9. Comparação do fator de utilização ρ em função de λ entre sistemas M/M/1 com e sem desestímulo.

A Tabela I resume os resultados das principais métricas utilizadas para avaliar o desempenho de ambos os sistemas, considerando $\lambda = 8$. Observa-se que o modelo com desestímulo apresenta uma redução significativa no número médio de clientes e no tempo de espera, tanto na fila quanto no sistema.

TABLE I
COMPARAÇÃO ENTRE M/M/1 E M/M/1 COM DESESTÍMULO PARA $\lambda = 8$

Parâmetros	M/M/1	M/M/1 com desestímulo
Fator de utilização (ρ)	0,665	0,487
Tempo médio de permanência no sistema $E[t_q]$	0,253	0,114
Número médio de Clientes no Sistema $E[q]$	2,027	0,667
Tempo médio de permanência na fila $E[t_w]$	0,17	0,031
Número médio de Clientes na Fila $E[w]$	1,36	0,181

B. M/G/1 com desestímulo

A Figura 10 apresenta o comportamento do tempo médio de espera na fila em função do fator de utilização, em sistemas de filas com desestímulo e com diferentes distribuições do tempo de serviço. Como era de se esperar, à medida que o desvio padrão do tempo de serviço diminui, o tempo de espera também diminui. A distribuição determinística representa o caso ideal, em que o tempo de serviço para cada cliente é constante, enquanto a distribuição exponencial é usada para modelar eventos com alta variabilidade, motivo pelo qual apresenta o maior tempo de espera.

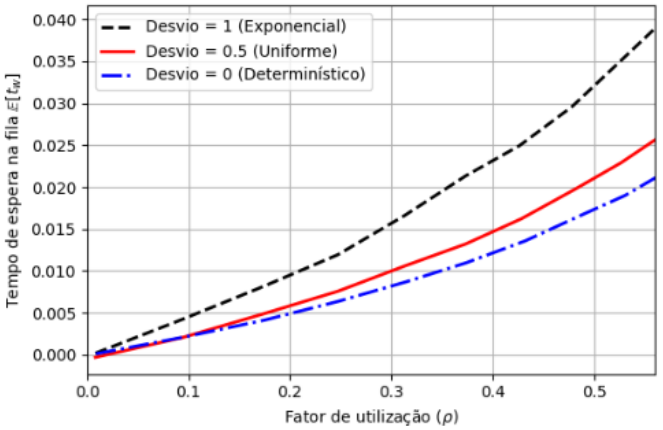

Fig. 10. Tempo médio de espera na fila $E[t_w]$ em função de ρ para diferentes distribuições de tempo de serviço em sistemas M/G/1 com desestímulo.

C. M/M/1/K com desestímulo

A Tabela 2 resume as principais métricas utilizadas para avaliar e comparar o desempenho de um sistema M/M/1/K e M/M/1/K com desestímulo, considerando $\lambda_0 = 8$ e $K = 4$. Observa-se que, ao aplicar o desestímulo, há uma redução significativa na probabilidade de bloqueio, na ocupação média do sistema e nos tempos médios de permanência.

TABLE II
COMPARAÇÃO ENTRE M/M/1/K E M/M/1/K COM DESESTÍMULO PARA $\lambda = 8$

Parâmetros	M/M/1/K	M/M/1/K com desestímulo
Probabilidade de bloqueio P_b	0,078	0,001
Fator de utilização ρ	0,619	0,488
Tempo médio de permanência no sistema $\mathbb{E}[t_q]$	0,17	0,114
Número médio de clientes no sistema $\mathbb{E}[q]$	1,256	0,666
Tempo médio de permanência na fila $\mathbb{E}[t_w]$	0,086	0,03
Número médio de clientes na fila $\mathbb{E}[w]$	0,641	0,179

D. M/G/1/K com desestímulo

Na Figura 11, é avaliado o impacto do desvio do tempo de serviço na probabilidade de bloqueio de uma fila M/M/1/K, considerando K = 4. Como a distribuição exponencial possui alta variabilidade, as filas tendem a ser mais longas, o que resulta em uma maior probabilidade de bloqueio. No caso determinístico, a formação de filas é reduzida devido ao comportamento previsível do sistema, o que diminui a probabilidade de bloqueio para um mesmo fator de utilização.

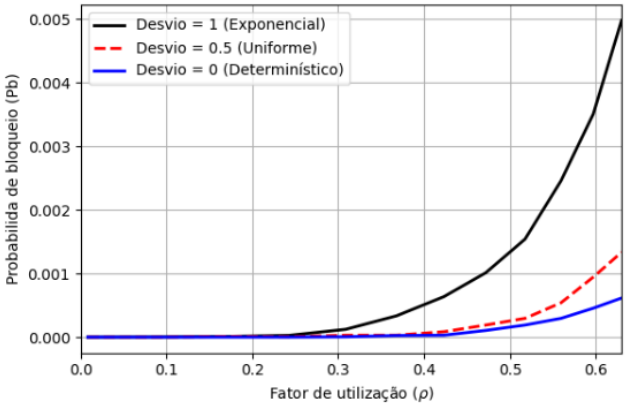

Fig. 11. Probabilidade de bloqueio P_b em função de ρ para diferentes distribuições de tempo de serviço em sistemas M/G/1/K com desestímulo e $K = 4$

IV. Conclusões

Neste projeto, avaliou-se o desempenho de sistemas de filas M/G/1 e M/G/1/K com desestímulo, considerando diferentes distribuições para modelar o tempo de serviço: exponencial, uniforme e determinística. No caso particular das filas M/M/1 com desestímulo, comparou-se seu desempenho com filas M/M/1 clássicas, o que permitiu observar que o desestímulo atua como um mecanismo de autorregulação que reduz efetivamente a congestão do sistema. Da mesma forma, foi realizada uma comparação entre um sistema de filas M/M/1/K com desestímulo e um sistema M/M/1/K clássico, a fim de avaliar o impacto do desestímulo sobre a probabilidade de bloqueio. Além disso, analisou-se o efeito da variabilidade do tempo de serviço no desempenho das filas.

Referências

[1] K. V. A. Rasheed and M. Manoharan, "Markovian queueing system with discouraged arrivals and self-regulatory servers," *Advances in Operations Research*, vol. 2016, pp. 1–11, 2016.

[2] B. A., "An infinite capacity single server markovian queueing system with discouraged arrivals retention of reneged customers and controllable arrival rates with feedback," *Indian Journal of Science and Technology*, vol. 17, pp. 3100–3108, 08 2024.

[3] C. Pravina, P. Kamala, S. Subbarayan, and S. Damodaran, "Analysis of a queue subject to discouraged arrivals, customer impatience, and self-switching server dynamics," *Contemporary Mathematics*, pp. 2738–2751, 04 2025.