# Análise de uma Rede de Sistemas de Fila com Dois Servidores em Série

*Projeto Final.

1nd Revisor: Adrian Alejandro Chavez Alanes, *Programa de Pós-Graduação (PPG) em Telecomunicações*
*Instituto Nacional de Telecomunicações*
Santa Rita do Sapucaí, Brasil
2rd Revisor: Everton Vilhena Cardoso, *Programa de Pós-Graduação (PPG) em Telecomunicações*
*Instituto Nacional de Telecomunicações*
Santa Rita do Sapucaí, Brasil

***Resumo*—This article presents the modeling and analysis of a tandem queueing network composed of two servers. The first server has a finite-capacity buffer, while the second has an infinite buffer. External arrivals follow a Poisson process, and the system is modeled using M/M/1/K and M/M/1 queueing disciplines, respectively. Analytical expressions are derived for key performance metrics such as *blocking probability*, *effective arrival rates*, and the *average number of customers in the queues*. Three operational scenarios are examined — an underdimensioned system, a balanced system, and a system with a bottleneck at the second server — using occupancy graphs and event flowcharts to illustrate the system's dynamics under different parameter configurations.**

## I. Descrição Geral do Sistema

O sistema em análise consiste em uma rede com dois servidores em série, como mostra o diagrama de blocos na Figura 1.

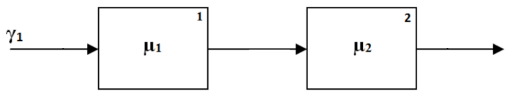


Figura 1. Diagrama de blocos do sistema com dois servidores em série.

As chegadas externas ocorrem a uma taxa de $\gamma_1$ mensagens/segundo [**ref2**]. O sistema é composto por:

- **Servidor 1:** Caracterizado por uma fila com buffer finito de tamanho J e uma taxa de atendimento $\mu_1$. A representação detalhada é vista na Figura 2.
- **Servidor 2:** Possui uma fila com buffer infinito e atende a uma taxa $\mu_2$. A saída do Servidor 1 alimenta diretamente esta fila. A Figura 3 ilustra este segundo estágio.

Quando a fila do Servidor 1 atinge sua capacidade máxima (J clientes aguardando e 1 em serviço), novas chegadas são bloqueadas e perdidas. A fila do Servidor 2, por ter buffer infinito, não apresenta bloqueios.

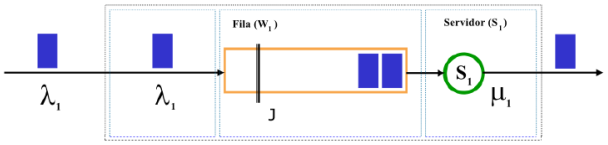


Figura 2. Primeiro estágio do sistema (Servidor 1 com fila de buffer finito J).

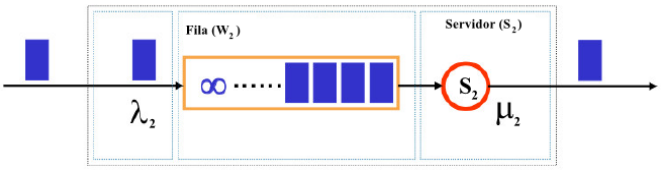


Figura 3. Segundo estágio do sistema (Servidor 2 com fila de buffer infinito).

## II. Aplicações

- **Título:** *"Modeling Two-Server Tandem Queues for QoS Optimization in Cloud Networks"*
  **Autores:** Gupta, P.; Zhang, H.
  **Publicação:** *Computer Communications Journal* (2019).
  **Contribuição:** Aplica o modelo de filas em série para balanceamento de carga e garantia de QoS em redes de data centers.
- **Título:** *"Performance Analysis of Tandem Queues with Active Queue Management in IP Networks"*
  **Autores:** Smith, J.; Li, Y.
  **Publicação:** *IEEE Transactions on Networking* (2020).
  **Contribuição:** Estuda um sistema de dois servidores em série com AQM (e.g., RED) para otimizar o atraso e a perda de pacotes em redes IP.
- **Título:** *"Delay and Throughput in Tandem Queues with AQM for IoT Traffic"*
  **Autores:** Silva, R.; Kumar, S.
  **Publicação:** *Springer Performance Evaluation* (2021).
  **Contribuição:** Investiga o impacto de políticas AQM em

redes IoT com servidores em série, focando em métricas de atraso e vazão.

## III. Síntese dos Trabalhos

Os artigos destacados, demonstram a versatilidade da modelagem de sistemas de filas em série, especialmente em cenários que envolvem gerenciadores ativos de fila (AQM) em ambientes de redes. A combinação de teoria das filas e otimização de redes permite melhorias significativas no desempenho de sistemas distribuídos modernos.

## IV. Notação do modelo

A notação de Kendall é utilizada para descrever o sistema de forma padronizada.

### A. Primeiro tramo

O primeiro estágio é um sistema M/M/1/K:

- **M**: Indica que as chegadas seguem um processo de Poisson com taxa $\gamma_1$.
- **M**: Refere-se a um tempo de serviço com distribuição exponencial, com taxa $\mu_1$.
- **1**: Significa que há apenas um servidor nesta etapa.
- **K**: Define a capacidade máxima do sistema (clientes em serviço + aguardando), onde $K = J + 1$.

### B. Segundo tramo

O segundo estágio é um sistema M/M/1:

- **M**: As chegadas correspondem às saídas do Servidor 1.
- **M**: Refere-se a um tempo de serviço com distribuição exponencial, com taxa $\mu_2$.
- **1**: Significa que há apenas um servidor nesta etapa.

O buffer é infinito, portanto não há o parâmetro K na notação. Além, a taxa de chegada ao Servidor 2 ($\lambda_2$) não é um parâmetro independente, mas uma consequência da dinâmica do primeiro estágio. Ela é calculada como:

$$\lambda_2 = \gamma_1 \times (1 - P_{bloq1})$$

onde $P_{bloq1}$ é a probabilidade de bloqueio na Fila 1.

## V. Modelagem

### A. Variáveis Principais

- $\gamma_1$: Taxa de chegadas externas.
- $K$: Capacidade total do primeiro nó ($K = J + 1$).
- $\mu_1$: Taxa de serviço do Servidor 1.
- $\mu_2$: Taxa de serviço do Servidor 2.
- $n_1(t)$: Número de clientes no nó 1 no instante t.
- $n_2(t)$: Número de clientes no nó 2 no instante t.

### B. Entidades e Eventos

As entidades são os clientes que chegam e transitam pelo sistema. Os eventos que alteram o estado do sistema são:

1) **Chegada de cliente à Fila 1:** Ocorre à taxa $\gamma_1$. Se a fila não está cheia ($i < K$), o cliente entra; caso contrário, é bloqueado.
2) **Término de atendimento no Servidor 1:** Ocorre à taxa $\mu_1$ (se $i > 0$). O cliente sai da Fila 1 e entra na Fila 2.
3) **Término de atendimento no Servidor 2:** Ocorre à taxa $\mu_2$ (se $j > 0$). O cliente sai definitivamente do sistema.

### C. Estados do sistema

O estado do sistema é descrito pelo par $(i, j)$, onde $i$ representa o número de clientes na fila 1, incluindo o que está em atendimento, e $j$ representa o número de clientes na fila 2, também incluindo o atendimento.

O valor de $i$ varia de 0 até $K$, que é a capacidade máxima do buffer da primeira fila, enquanto $j \geq 0$, já que a segunda fila possui buffer infinito.

O sistema evolui conforme chegadas externas e término dos atendimentos nos servidores 1 e 2. Uma chegada externa aumenta $i$ se $i < K$, caso contrário ocorre bloqueio. O término do atendimento no servidor 1 reduz $i$ e aumenta $j$, enquanto o término no servidor 2 reduz $j$.

## VI. Análise Analítica

### A. Etapa 1: M/M/1/K

- **Fator de Utilização ($\rho_1$):** $\rho_1 = \frac{\lambda_1}{\mu_1}$
- **Probabilidade de Bloqueio ($P_{bloq1}$):** $P_b = P_1(K) = \frac{(1-\rho_1)\rho_1^K}{1-\rho_1^{K+1}}$
- **Taxa de chegada na Fila 2 ($\lambda_2$):** $\lambda_2 = \lambda_1 \times (1 - P_{bloq1})$
- **Nº médio de clientes ($E[q]_1$):** $E[q]_1 = \frac{\rho_1}{1-\rho_1} - \frac{(K+1)\rho_1^{K+1}}{1-\rho_1^{K+1}}$
- **Tempo médio de permanência ($E[t_q]_1$):** $E[t_q]_1 = \frac{E[q]_1}{\lambda_1(1-P_{bloq1})}$

### B. Etapa 2: M/M/1

- **Fator de Utilização ($\rho_2$):** $\rho_2 = \frac{\lambda_2}{\mu_2}$
- **Nº médio de clientes na Fila 2 ($E[q]_2$):** $E[q]_2 = \frac{\rho_2}{1-\rho_2}$
- **Tempo médio de permanência no sistema 2 ($E[t_q]_2$):** $E[t_q]_2 = \frac{E[q]_2}{\lambda_2}$

## VII. Resultados e Cenários

Foram analisados três cenários distintos. Os parâmetros de entrada são apresentados na Tabela I.

Tabela I
Parâmetros dos Cenários Analisados

| Cenário | $\gamma_1$ | $\mu_1$ | $\mu_2$ | K |
|---|---|---|---|---|
| 1 | 4.5 | 4.0 | 4.5 | 3 |
| 2 | 1.5 | 2.0 | 2.0 | 5 |
| 3 | 1.5 | 2.0 | 1.0 | 5 |

Fazendo uso dos parâmetros do premer cenário na Tabela

I, obtemos os seguintes resultados:

$$\rho_1 = \frac{4.5}{4} = 1.125$$
$$P_b = \frac{(1-1.125)\cdot 1.125^3}{1-1.125^4} = 0.2957$$
$$\lambda_2 = 4.5 \cdot (1-0.2957) = 3.1694$$
$$E[q_1] = \frac{1.125}{1-1.125} \cdot \frac{4 \cdot 1.125^4}{1-1.125^4} = 1.6466$$
$$E[t_{q_1}] = \frac{1.6466}{(1-0.2957)\cdot 4.5} = 0.5195$$
$$\rho_2 = \frac{3.1694}{4.5} = 0.7043$$
$$E[q_2] = \frac{0.7043}{1-0.7043} = 2.3818$$
$$E[t_{q_2}] = \frac{0.7043}{4.5-3.1694} = 0.5293$$
$$E[q]_{total} = 1.6466 + 2.3818 = 4.0284$$
$$E[t_q]_{total} = 0.5195 + 0.5293 = 1.0488$$

A Tabela II sumariza os resultados analíticos para cada cenário de forma resumida.

Tabela II
RESULTADOS ANALÍTICOS DOS CENÁRIOS

| **Variável** | **Cenário 1** | | **Cenário 2** | | **Cenário 3** | |
|---|---|---|---|---|---|---|
| | Analítico | Simulado | Analítico | Simulado | Analítico | Simulado |
| $\rho_1$ | 1.125 | 1.125 | 0.75 | 0.75 | 0.75 | 0.75 |
| $P_b$ | 0.2957 | 0.2960 | 0.07217 | 0.0740 | 0.07217 | 0.0728 |
| $\lambda_2$ | 3.1694 | 3.1689 | 1.3917 | 1.3909 | 1.3917 | 1.3943 |
| $E[q]_1$ | 1.6466 | 1.6474 | 1.7009 | 1.711 | 1.7009 | 1.7119 |
| $E[tq]_1$ | 0.5195 | 0.5199 | 1.2221 | 1.2302 | 1.2221 | 1.2278 |
| $\rho_2$ | 0.7043 | 0.7037 | 0.6958 | 0.6954 | 1.3917 | 1.3987 |
| $E[q]_2$ | 2.3818 | 2.0371 | 2.2873 | 2.1057 | $\infty$ | 19896.9 |
| $E[tq]_2$ | 0.5293 | 0.6428 | 1.1438 | 1.5139 | $\infty$ | 14332.1 |
| $E[tq]_{total}$ | 1.0488 | 1.1627 | 2.3659 | 2.7441 | $\infty$ | 14333.4 |

Esta seção apresenta os resultados da simulação para três cenários distintos, que refletem diferentes condições do sistema de filas. O Cenário 1 representa um sistema subdimensionado com alta demanda e servidor lento.

O Cenário 2 mostra um sistema equilibrado com baixa taxa de bloqueio. O Cenário 3 destaca o impacto de um gargalo no servidor 2, onde a fila cresce indefinidamente devido à alta taxa de chegada.

### A. *Cenário 1: Sistema subdimensionado*

Objetivo: Demonstrar o comportamento com alta demanda e servidor lento. A Figura 4 ilustra a alta volatilidade na ocupação das filas.

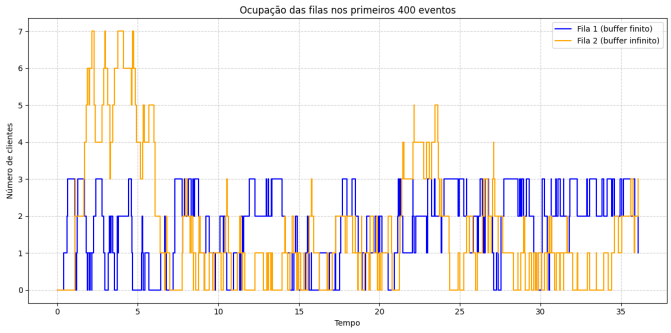


Figura 4. Ocupação das filas no Cenário 1.

### B. *Cenário 2: Sistema equilibrado*

Objetivo: Mostrar um sistema bem dimensionado com baixo bloqueio. A Figura 5 mostra uma operação mais estável.

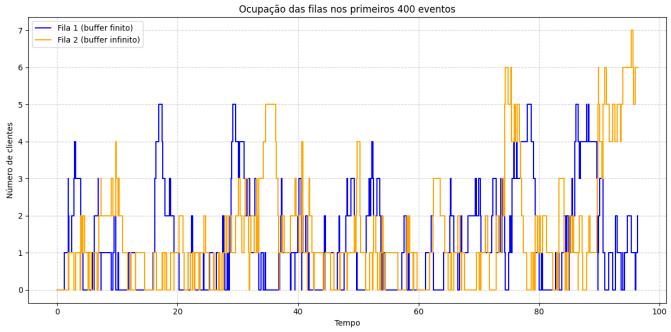


Figura 5. Ocupação das filas no Cenário 2.

### C. *Cenário 3: Servidor 2 limitado*

Objetivo: Evidenciar o impacto de um gargalo no Servidor 2. Com $\lambda_2 > \mu_2$, a Fila 2 cresce indefinidamente, indicando instabilidade, como visto na Figura 6.

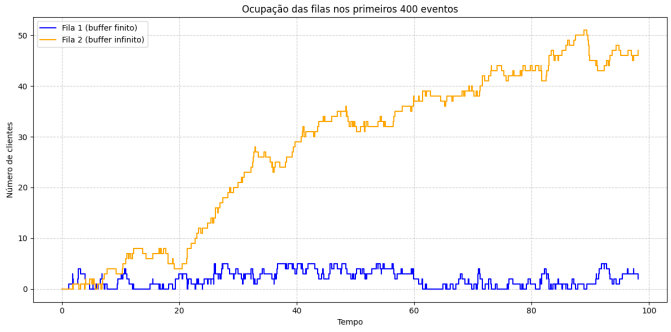


Figura 6. Ocupação das filas no Cenário 3.

## VIII. CONCLUSÕES

Este trabalho analisou uma rede de filas com dois servidores em série, utilizando os modelos M/M/1/K e M/M/1 para os estágios respectivamente.

Nos três cenários simulados, destacam-se:

- Cenário 1: alta taxa de bloqueio causada por alta demanda e capacidade limitada no primeiro servidor.
- Cenário 2: sistema equilibrado com baixa probabilidade de bloqueio e tempos de espera reduzidos.
- Cenário 3: gargalo no segundo servidor, gerando crescimento ilimitado da fila e instabilidade.

Os resultados analíticos e simulados mostraram boa concordância, validando o modelo e ressaltando a importância do dimensionamento correto para garantir estabilidade.

Sistemas com dois servidores em série, como o estudado, são comuns em linhas de produção industriais, atendimento em call centers com múltiplas etapas e redes de comunicação onde pacotes passam por diferentes estágios de processamento.

## REFERÊNCIAS

[1] L. Kleinrock. *Queueing Systems, Volume I: Theory*. Wiley-Interscience, 1975. ISBN: 0471491101. Disponível em: https://www.wiley.com/en-us/Queueing+Systems%2C+Volume+I-p-9780471491101.

[2] L. Kleinrock. *Queueing Systems, Volume II: Computer Applications*. Wiley-Interscience, 1976. ISBN: 047149111X. Disponível em: https://www.wiley.com/en-us/Queueing+Systems%2C+Volume+2%3A+Computer+Applications-p-9780471491118.

[3] L. Lakatos. *Queueing Theory: A Linear Algebraic Approach*. Wiley, 2013. Disponível em: https://onlinelibrary.wiley.com/doi/book/10.1002/9781118625651.

## IX. Anexos

Os fluxogramas nas Figuras 7, 8 e 9 descrevem a lógica de processamento dos eventos principais do sistema.

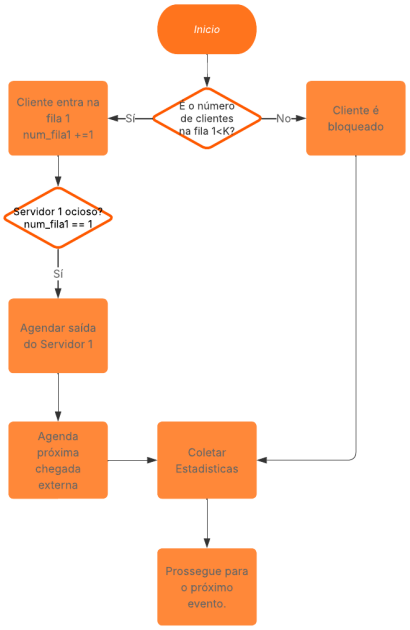


Figura 7. Fluxograma do evento de chegada de cliente externo.

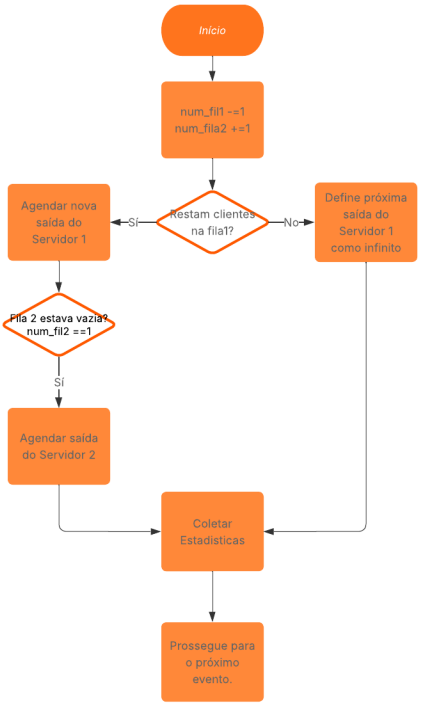


Figura 8. Fluxograma do evento de término de serviço no Servidor 1.

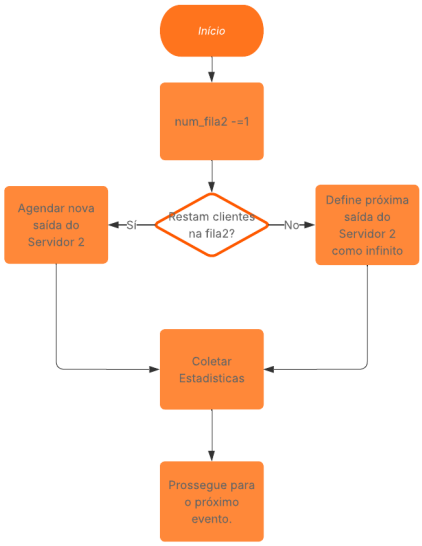


Figura 9. Fluxograma do evento de término de serviço no Servidor 2.