# Teoria de Filas

Prof. Fabrício B. Gonçalves

Instituto Federal Fluminense
Campus Bom Jesus do Itabapoana

# Agenda

1 Introdução

# Agenda


1. Introdução
2. Sistemas de Fluxo

# Agenda


1. Introdução
2. Sistemas de Fluxo
3. Processo Estocástico

# Agenda


1. Introdução
2. Sistemas de Fluxo
3. Processo Estocástico
4. Caracterização dos Sistemas de Fila Única

# Introdução

# Introdução

Todos nós enfrentamos filas em nosso dia a dia: seja esperando para ser atendido em um banco, aguardando a página de um website carregar ou até mesmo esperando por um documento a ser impresso. No núcleo dessas situações está a dinâmica de chegada, espera e atendimento.

As filas influenciam diretamente a experiência do usuário, a eficiência dos sistemas e os custos operacionais. A Teoria de Filas nos ajuda a entender, analisar e otimizar esses processos, garantindo que os recursos sejam utilizados da maneira mais eficaz possível.

# Sistemas de Fluxo

# Definição

# Sistemas de Fluxo

## Definição

### Definição

Um sistema de fluxo é um sistema no qual alguma mercadoria flui, se move ou é transferida através de um ou mais canais de capacidade limitada, de forma a ir de um ponto a outro no sistema.

# Exemplo

# Sistemas de Fluxo

## Exemplo

### Exemplo

Imagine uma rede de computadores, onde o computador $A$ transfere dados para o computador $B$, ou um processador transferindo bits entre a memória e o disco.

Nesses exemplos, as mecadorias seriam os pacotes e os bits; os canais seriam o link de rede e o barramento de dados.

# Classificação dos Sistemas de Fluxo

# Sistemas de Fluxo

## Classificação dos Sistemas de Fluxo

Os sistemas de fluxo são classificados em dois tipos:

- **Sistemas estáveis:** são aqueles nos quais o fluxo passa de forma previsível, isto é, a quantidade de fluxo que passa pelo canal é constante e conhecida durante o intervalo de tempo de observação do sistema. Também são constantes o tempo que o fluxo chega ao canal e quanto aquele fluxo demanda do canal.
- **Sistemas não estáveis:** são problemas de fluxos estocásticos ou randômicos, isto é, as chegadas de mercadorias ao canal e o tamanho da demanda por uso do canal são imprevisíveis.

# Sistemas de Fluxo

## Classificação dos Sistemas de Fluxo

Os sistemas de fluxo também podem ser classificados de acordo com a natureza dos canais:

- **Canal único:** as mercadorias entram no sistema de único canal para serem processadas (transferidas, executadas, servidas) pelo canal de capacidade (ou taxa de serviço) C. Considerando uma taxa de chegadas (número de mercadorias que chegam durante certo intervalo de tempo) $R$, temos que:
- **Redes de canais:** As mercadorias entram em uma rede de filas através de múltiplos pontos de acesso e são direcionadas a diferentes estações de serviço conforme as regras estabelecidas. Cada estação possui uma capacidade ou taxa de serviço própria, denotada por $C_i$ (onde $i$ representa a estação específica). A taxa de chegadas em cada estação, ou o número de mercadorias que chegam durante um determinado intervalo de tempo, é representada por $R_i$. Em uma rede, após serem atendidas em uma estação, as mercadorias podem ser transferidas para outra estação ou deixar o sistema, dependendo das probabilidades ou políticas definidas.

# Sistemas de Fluxo

## Classificação dos Sistemas de Fluxo

**Atenção**

- Se **C** > **R**, o sistema consegue processar as mercadorias sem que ocorra sobrecarga do canal;
- Se **R** > **C**, significa que o canal não consegue processar as mercadorias e a sobrecarga levará o canal a um nível de congestionamento, isto é, filas de mercadorias se formarão.

# Processo Estocástico

# Processo Estocástico

Para a análise dos sistemas de fluxo, incluindo as filas, é necessário identificar o processo estocástico que descreve o fluxo, a estrutura e a disciplina do canal. Considerando um sistema de filas, chamaremos mercadorias de clientes, e o canal de centro de serviço.

O processo de chegada de clientes no centro de serviço é descrito em termos da distribuição de probabilidade dos tempos de chegada dos clientes. Denominado por $A(t)$, temos:

$$\boldsymbol{A(t) = P\,[\text{tempo entre chegadas} \leq t]}$$

significando que a taxa de chegadas de clientes no sistema é determinada pela possibilidade de que o tempo entre chegadas seja menor que um tempo $\boldsymbol{t}$.

# Processo Estocástico

A segunda quatidade é a demanda das chegadas no canal, também chamada de tempo de serviço, isto é, a quantidade de tempo que um cliente demora sendo atendido dentro de um centro de serviço. A distribuição de probabilidade é dada por ***B(x)***, tal que:

$$\boldsymbol{B(x) = P\,[\text{tempo de serviço} \leq x]}$$

significando que o tempo de serviço para o atendimento de uma requisição do cliente no sistema é determinado pela probabilidade de que o tempo de serviço seja menor do que um tempo ***x***. O tempo entre chegadas e o tempo de serviço são variáveis independentes entre si.

# Caracterização dos Sistemas de Fila Única

# Caracterização dos Sistemas de Fila Única

Um sistema de fila é composto por:

- **Linha de entrada:** é onde chegam os clientes (requisições, pacotes *jobs* etc.) com o tempo arbitrário entre chegadas ***A(t)***, que se encaminham para o sistema de filas, onde são atendidos (servidos) em um tempo arbitrário de serviço **B(x)** e depois são liberados em uma linha de saída.
- **Sistema de fila:** encaminha os clientes para um ou mais centros de serviço. Dizemos que o sistema tem ***m*** servidores sendo $m \geq 1$.
- **Linha de saída:** é onde os clientes atendidos saem do sistema de fila.

# Variáveis Relacionadas a uma Fila

# Caracterização dos Sistemas de Fila Única

## Variáveis Relacionadas a uma Fila

Algumas variáveis são importantes e devem ser analisadas para cada sistema de fila:

- Número de clientes;
- Taxa de chegadas;
- Taxa de serviço;
- Tempo de espera;
- Tempo de resposta.

# Caracterização dos Sistemas de Fila Única

## Variáveis Relacionadas a uma Fila: Número de Clientes

- $C_n$: denota o $n$-ésimo cliente que entra no sistema;
- $n_w$: representa o número de usuários esperando para serem atendidos. Este número é sempre menor que do que $n$, dado que não inclui os usuários que estão sendo atendidos;
- $n_s$: representa o número de usuários em atendimento dentro do centro de serviço;
- $n$: denota o número de clientes no sistema em um determinado instante de tempo. Esta variável inclui os usuários que já estão sendo atendidos, assim como os que estão esperando na fila. A nédia do número total de clientes no sistema inclui a média dos clientes na fila e a média dos clientes em atendimento.

$$E[n] = E[n_w] + E[n_s]$$

# Caracterização dos Sistemas de Fila Única

## Variáveis Relacionadas a uma Fila: Taxa de Chegadas

A taxa de chegada $\lambda_n$ é o tempo de chegada de $C_n$ clientes no sistema.

$$\lambda = \frac{1}{A(x)}$$

Por exemplo, ser observam oito chegadas de clientes durante um tempo de observação de quatro minutos, então a taxa de chegadas é 8/4 = 2 clientes por minuto.

# Caracterização dos Sistemas de Fila Única

## Variáveis Relacionadas a uma Fila: Taxa de Serviço

Taxa de serviço $\mu$ é a taxa média de serviço por cliente no servidor, isto é, quantos clientes o sistema consegue atender em uma unidade de tempo. É o inverso do tempo de serviço:

$$\mu = \frac{1}{B(x)}$$

Por exemplo, se o servidor do sistema demora dois segundos para atender a uma requisição, então a taxa de serviço é 1/2 = 0,5 requisição por segundo.

Para $m$ servidores conectados a uma única fila, a taxa total de serviço é $m \times \mu$.

Então, se a fila receber mais um servidor para atender os clientes, a taxa de serviço será uma requisição por segundo.

# Caracterização dos Sistemas de Fila Única

## Variáveis Relacionadas a uma Fila: Tempo de Espera

O tempo de espera $w$ é o tempo que o cliente espera na fila antes de ser atendido pelo servidor.

# Caracterização dos Sistemas de Fila Única

## Variáveis Relacionadas a uma Fila: Tempo de Resposta

O tempo de resposta $s$ é o tempo médio total gasto por um cliente, contando desde a entrada do cliente no sistema, incluindo o tempo médio gasto na fila e o tempo médio que demorou em atendimento $x$, até sua saída.

$$E[s] = E[w] + E[x]$$

Por exemplo, o servidor demora em média um segundo para atender a uma requisição, o tempo médio na fila para cada requisição é de cinco segundos. Por tanto, o tempo médio de resposta para uma requisição seria de seis segundos.

# Medidas e Relacionamentos entre as Variáveis

# Caracterização dos Sistemas de Fila Única

## Medidas e Relacionamentos entre as Variáveis: Lei de Litle

Se os usuários não forem perdidos por causa limitações de capacidade do sistema, temos que o número médio de usuários no sistema é a taxa de chegadas multiplacada pelo tempo médio de resposta:

$$E[n] = \lambda \times E[s]$$

Da mesma forma, o número médio de usuários na fila é obtido pela taxa de chegadas multiplicada pelo tempo médio de espera:

$$E[n_w] = \lambda \times E[w]$$

# Caracterização dos Sistemas de Fila Única

## Medidas e Relacionamentos entre as Variáveis: Probabilidade do Número de Clientes no Sistema

Representada por $P_n$, esta probabilidade é utilizada para descrever o comportamento de um sistema de filas em termos do número de clientes, $n$, atualmente no sistema.

Para cada tipo de fila, é necessário calcular $P_n$, que pode ser usado para calcular outras informações sobre a fila.

# Caracterização dos Sistemas de Fila Única

## Medidas e Relacionamentos entre as Variáveis: Utilização

$U(t)$ denota a utilização do sistema, isto é, a fração de tempo na qual o servidor está ocupado. Quando $U(t) = 0$, o sitema está vazio. Quando $U(t) > 0$, o sistema está com alguma ocupação e em $U(t) = 1$, o sistema está 100% ocupado.

O servidor é dito ocioso quando não há nenhuma requisição no sistema; em todos os demais momentos ele é dito ocupado. O intervalo entre dois intervalos ociosos sucessivos é denominado período ocupado.

# Caracterização dos Sistemas de Fila Única

## Medidas e Relacionamentos entre as Variáveis: Condição de Estabilidade

A taxa média de chegadas deve ser menor do que a taxa média de atendimento, ou seja, $\lambda < m\mu$, com $m \geq 1$. Esta relação é válida apenas para filas infinitas. Com filas finitas, o sistema nunca é instável.

# Caracterização dos Sistemas de Fila Única

## Medidas e Relacionamentos entre as Variáveis: Intensidade de Tráfego

A taxa de utilização $\rho$ mede o congestionamento do sistema de fila. Se for perto de zero, há pouca fila, e em geral, quando a intensidade de tráfego aumenta, a fila também aumenta. Veremos que, para alguns sistemas de filas, a intensidade de tráfego é equivalente a utilização do sistema.

Definimos a taxa de utilização como:

$$\rho = \frac{\lambda}{\mu}$$

Esta variável também pode ser utilizada para formular a condição de estabilidade, que deve ser $\rho < 1$, a média dos números de requisições que chegam em uma unidade de tempo deve ser menor do que o número de requisições que podem ser processadas.

Para *m* servidores conectados a mesma fila, de forma geral,

$$\rho = \frac{\lambda}{m\mu}$$

.

# Caracterização dos Sistemas de Fila Única

## Medidas e Relacionamentos entre as Variáveis: Vazão

Em termos de filas, vazão é a média das requisições processadas em uma unidade de tempo, isto é, a taxa de saídas do sistema de filas. Para filas em equilíbrio, a taxa de saídas é igual a de chegadas, então podemos definir a vazão como:

$$\lambda = m\rho\mu$$

No entanto, é crucial observar que a vazão máxima do sistema é limitada pela taxa de serviço. Se a taxa de chegada excede a capacidade do sistema de atender a essas chegadas, então a vazão efetiva será igual à taxa máxima de serviço do sistema.

# Notação de Kendall

# Caracterização dos Sistemas de Fila Única

## Notação Kendall

Em 1953, D. G. Kendall criou uma notação composta por três parâmetros funcionais para caracterizar os sistemas de fila. Hoje em dia, a notação possui seis parâmetros e segue a forma A/S/m/B/K/SD, onde:

- **A:** é a distribuição do processo de chegadas;
- **S:** é a distribuição do tempo de serviço;
- **m:** é o número de servidores;
- **B:** é a capacidade do sistema;
- **K:** é o tamanho da população;
- **SD:** é a disciplina de atendimento da fila.

# Caracterização dos Sistemas de Fila Única

## Notação Kendall: Distribuição do Processo de Chegada (A)

É a distribuição do processo de chegadas de clientes no sistema. Algumas distribuições utilizadas:

- **M:** processo de chegadas de Poisson (ou randômico);
- **E$_k$:** distribuição de Erlang de parâmetro $k$;
- **D:** tempo entre chegadas fixo ou determinítico;
- **G:** distribuição genérica;
- **GI:** distribuição genérica com tempo entre chegadas independente.

# Caracterização dos Sistemas de Fila Única

## Notação Kendall: Distribuição do Tempo de Serviço (S)

É a distribuição do tempo de serviço para atendimento de um cliente em um servidor. As principais distribuições usadas são as mesmas para o processo de chegadas.

# Caracterização dos Sistemas de Fila Única

## Notação Kendall: Número de Servidores (m)

Representa o número de servidores (ou canais de serviço) disponíveis no sistema. No mínimo, o sistema de filas deve ter um servidor e, no máximmo, infinitos. Centros de serviço com infinitos servidores são chamados de centro de atrasos.

# Caracterização dos Sistemas de Fila Única

## Notação Kendall: Capacidade do Sistema (B)

A capacidade do sistema é o número máximo de clientes que são permitidos dentro do sistema em certo momento. Quando a capacidade chega ao limite máximo, novos clientes que chegam ao sistema são recusados. Na notação Kendall, se esse número é omitido, assume-se que a capacidade do sistema é ilimitada.

A capacidade inclui o número máximo de clientes que podem ser atendidos e o número de clientes que cabem na fila. Se esse valor é omitido, significa que a capacidade do sistema é infinita.

# Caracterização dos Sistemas de Fila Única

## Notação Kendall: Tamanho da População (K)

É o tamanho da população de clientes, representando o número de clientes interessados em usar o sistema ou possíveis usuários.

Por exemplo, o servidor de aplicações acadêmicas de uma universidade tem como possível população alunos, funcionários e professores. Já o servidor de uma aplicação Web tem uma população teoricamente infinita.

Se esse valor é omitido na definição da fila, significa que a população é considerada infinita.

# Caracterização dos Sistemas de Fila Única

## Notação Kendall: Disciplina de Atendimento da Fila (SD)

Representa a ordem em que os clientes são atendidos após chegar à fila. Algumas das principais disciplinas de atendimento são:

- **FIFO/FCFS (*First In First Out/First Come First Served*):** os clientes são atendidos (servidos) na ordem em que chegam;
- **LIFO/LCFS (*Last In First Out/Last Come First Served*):** os clientes são atendidos (servidos) na ordem reversa em que chegam;
- **SIRO (*Service In Randon Order)*:** os clientes são atendidos em ordem randômica, sem considerar a ordem em que chegam;
- **RR (*Round Robin*):** cada cliente tem uma fatia de tempo para ser atendido. Após usar sua fatia de tempo, sua requisição para de ser servida e retorna para o fim da fila. Isto repete até a requisição ter ter o serviço completado;
- **PS (*Processor Sharing*):** Equivale a RR com fatias de tempo infinitesimais. Aparenta ter todas as requisições atendidas simultaneamente, mas com o tempo de serviço aumentado;
- **IS (*Infinite Server*):** são tantos servidores que nunca se forma uma fila.