

# **TEORIA DAS FILAS**

---

Por : Daniel Csillag, Darlan, Iara Cristina e Janaina Neres

# PARTE TEÓRICA

Introdução, Conceitos e Propriedades

# TEORIA DAS FILAS

## INTRODUÇÃO

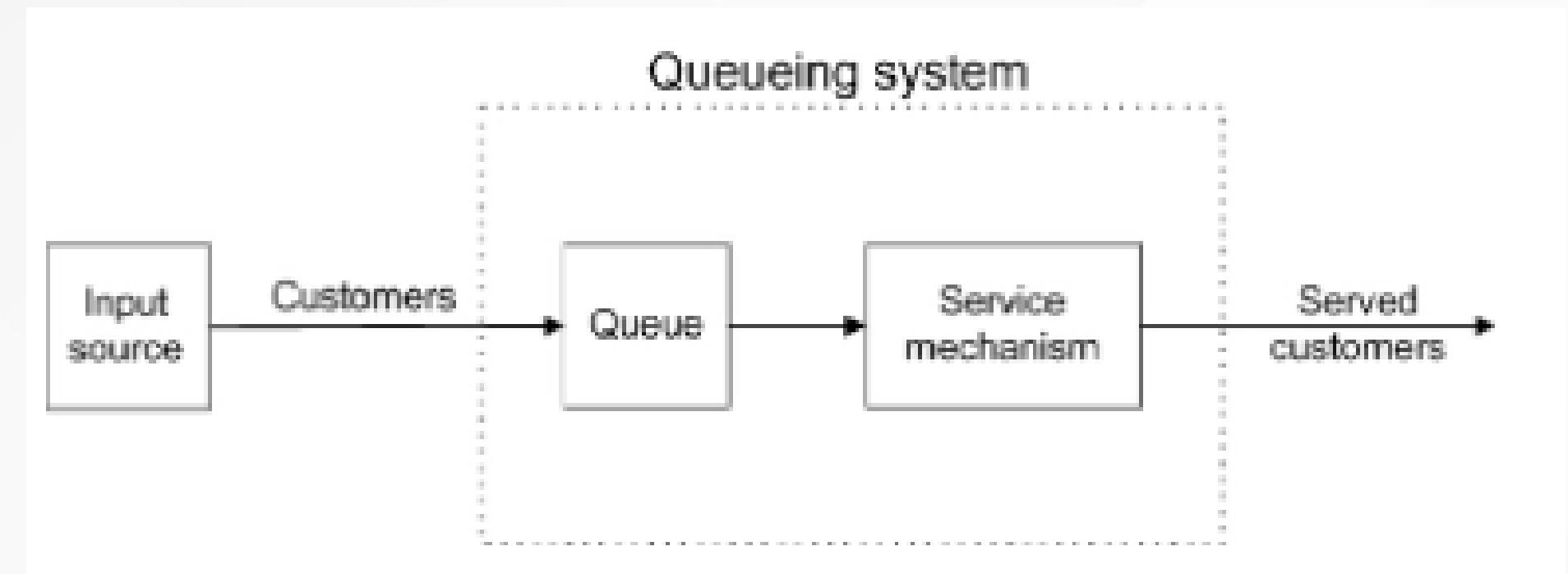
---

A teoria das filas é uma disciplina que se dedica a compreender a dinâmica de formação, operação e análise de filas e sistemas de espera.



# TEORIA DAS FILAS

## COMPONENTES DE UM SISTEMA DE FILAS



Um sistema de filas consiste no processo de chegada, da distribuição do tempo de serviço, do número de servidores, da capacidade do sistema, da população de usuários e da disciplina de atendimento.

# TEORIA DAS FILAS

## PROCESSO DE CHEGADA

---

O processo de chegada indica qual o padrão de chegada dos clientes no sistema.

Clientes podem chegar simultaneamente (chegada em batch\lote).

A reação do cliente na fila pode variar. Ele pode esperar independentemente do tamanho da fila, também pode decidir não entrar no sistema caso a fila esteja muito grande (cliente decepcionado), ele pode esperar na fila mas depois de um tempo desistir e sair do sistema, e também pode mudar de uma fila para outra em sistemas com servidores paralelos.

A distribuição mais comum é a de Poisson, ou seja, os tempos entre as chegadas são exponencialmente distribuídos. Entre outras distribuições, estão a de Erlang, hiperexponencial e arbitrária.



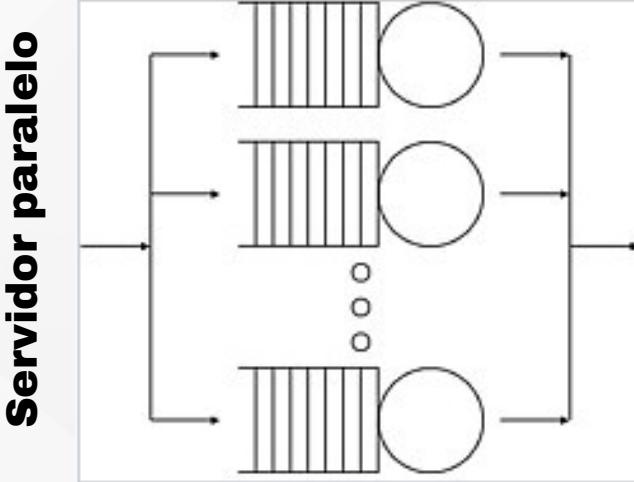
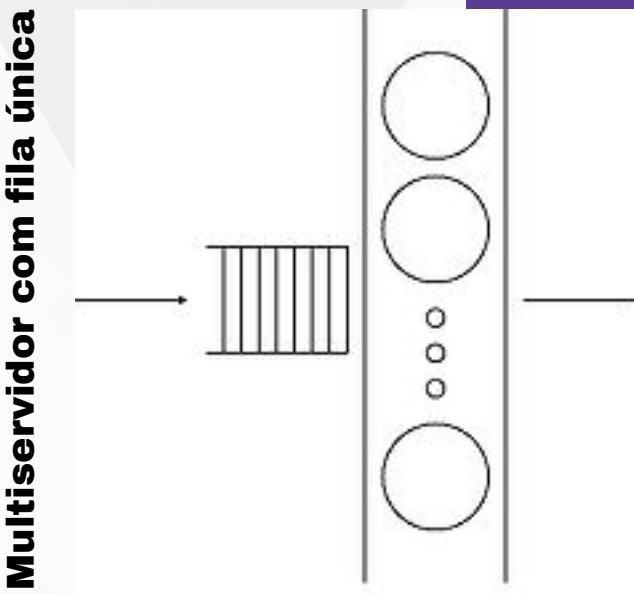
# TEORIA DAS FILAS

## DISTRIBUIÇÃO DO TEMPO DE SERVIÇO

Os serviços podem também ser simples ou lote.

O estado pode ser independente: o processo de atendimento não depende do número de clientes esperando pelo serviço.

Em contrapartida, em um estado dependente, o processo de atendimento muda de acordo com o número de clientes na fila. Por exemplo, um servidor pode trabalhar mais rápido quando a fila aumenta ou, ao contrário, ficar confuso e então mais lento



# TEORIA DAS FILAS

## DISCIPLINA DE ATENDIMENTO

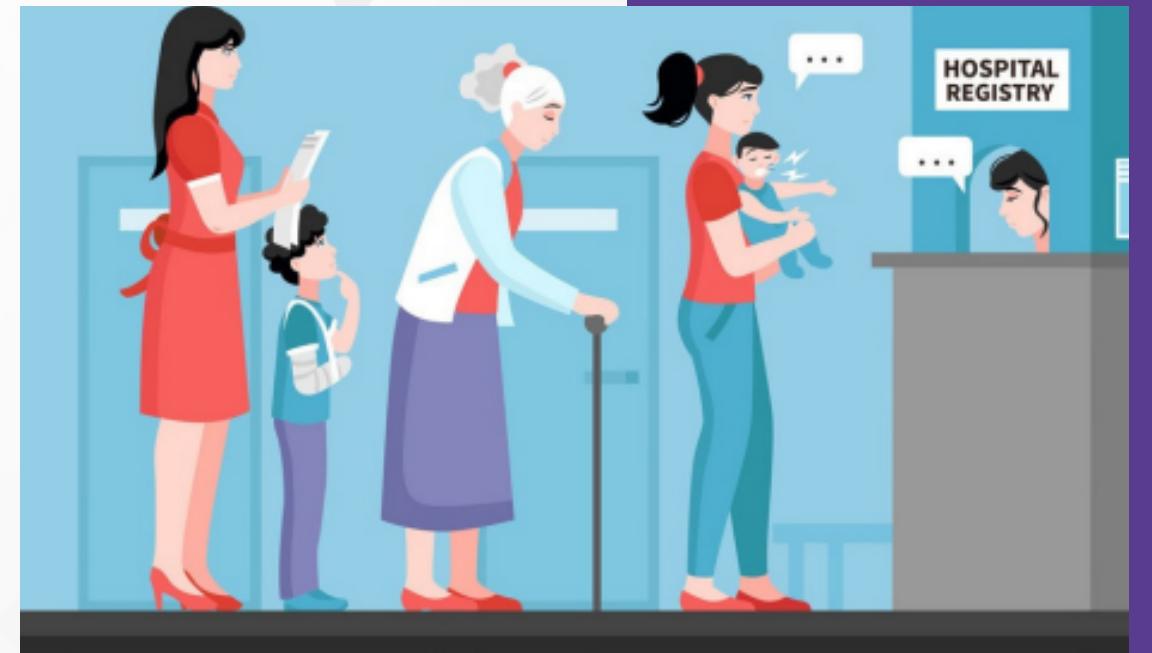
Descreve a forma como os clientes saem da fila de espera para serem atendidos. Algumas disciplinas são:

**FIFO** (First In, First Out): Primeiro a Entrar, Primeiro a Sair). Disciplina mais comum, inclusive na vida diária. [FIFO também é chamado de FCFS (First Come, First Served)]

**LIFO** (Last In, First Out): Último a Chegar, Primeiro a Sair. Aplicável em sistemas em que o item mais recente é mais fácil de ser recuperado, como por exemplo em sistemas de controle de estoque. [LIFO também é chamado de LCFS (Last Come, First Served)]

**Fila com prioridade**: a cada cliente é atribuída uma prioridade; clientes com maior prioridade têm preferência no atendimento.

**SIRO** (Serve In Random Order): Atendimento em Ordem Aleatória. Indenpendente de um item ser recente ou estar na fila há mais tempo, as chances de cada um são as mesmas



# TEORIA DAS FILAS

## NOTAÇÃO DE KENDALL E PARÂMETROS DA FILA

**A/S/m/K/N/Q**

Em que:

- **A:** Distribuição dos tempos entre as chegadas (Processo de chegada)
- **S:** Distribuição dos tempos de serviço
- **m:** Número de servidores
- **K:** Capacidade do sistema
- **N:** Tamanho da população
- **Q:** Disciplina de atendimento

Exemplos de sistemas de filas  
**M/G/4/50/2000/LCFS**

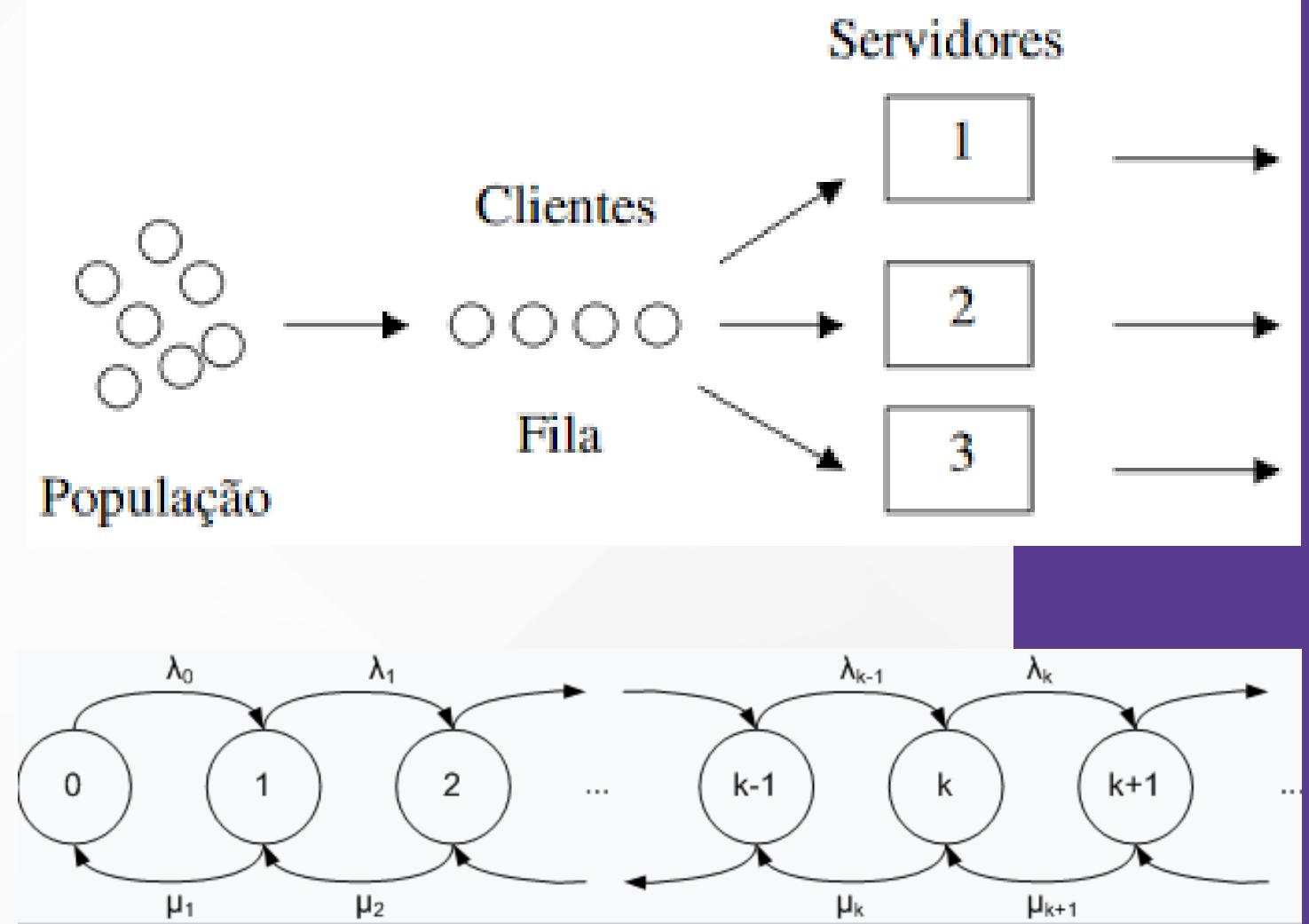
- Processo de chegada exponencial (Markoviano) ou de Poisson
- Distribuição dos tempos de serviço arbitrária (Geral)
- Quatro servidores
- Capacidade para cinquenta clientes
- População de dois mil clientes
- Disciplina de atendimento "Último a Chegar, Primeiro a ser Servido"

Obs: Muitas vezes, os três últimos símbolos são omitidos. Nestes casos, assume-se capacidade ilimitada, população infinita e disciplina de atendimento FCFS (M/M/1)

# TEORIA DAS FILAS

## FILAS M\textbackslash M\textbackslash C

- Cadeia de Markov em tempo contínuo
- Processo de nascimento e Morte em tempo contínuo
- Avançar para o próximo estado ( $i+1$ ) tem distribuição exponencial  $\lambda$
- Regredir para o estado anterior ( $i-1$ ) tem distribuição exponencial:
  - $i\mu$ , para  $i < c$
  - $c\mu$ , c.c



# TEORIA DAS FILAS

## LEI DE LITTLE

A Lei de Little relaciona o número de clientes no sistema com o tempo médio despendido no sistema.

E se aplica sempre que o número de chegadas é igual ao número de saídas (denominado sistema em equilíbrio).

$$L = \lambda W$$

L é o número médio de longo prazo de clientes no sistema.  
 $\lambda$  é a taxa média de chegada no longo prazo.  
W é o tempo médio que um cliente passa no sistema.

# TEORIA DAS FILAS

## EQUILÍBRIOS DAS FILAS

Qualquer fila que seja da configuração  $M/M/c$  terá distribuição estacionária se  $c\mu > \lambda$ .

$$\sum_{k=0}^{c-1} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} + \frac{(\lambda/\mu)^c}{c!} \sum_{k=c}^{\infty} \left(\frac{\lambda}{c\mu}\right)^{k-c}$$

# TEORIA DAS FILAS

## EQUILÍBRIOS DAS FILAS

Tendo isso em vista, caso o processo tenha distribuição estacionária, ele seria:

$$\pi_0 = \left( \sum_{k=0}^{c-1} \left( \frac{\lambda}{\mu} \right)^k \frac{1}{k!} + \frac{(\lambda/\mu)^c}{c!} \left( \frac{1}{1 - \lambda/c\mu} \right) \right)^{-1}$$

$$\pi_k = \begin{cases} \frac{\pi_0}{k!} \left( \frac{\lambda}{\mu} \right)^k, & \text{for } 0 \leq k < c \\ \frac{\pi_0}{c^{k-c} c!} \left( \frac{\lambda}{\mu} \right)^k, & \text{for } k \geq c. \end{cases}$$

# TEORIA DAS FILAS

## EQUILÍBRIO DAS FILAS

$$\sum_{k=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} = e^{\frac{\lambda}{\mu}} \Rightarrow \pi_0 = e^{-\frac{\lambda}{\mu}} \Rightarrow \pi_k = \frac{e^{-\frac{\lambda}{\mu}}}{k!} \left(\frac{\lambda}{\mu}\right)^k$$

$$\sum_{k=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^k = \frac{\mu}{\mu - \lambda} \Rightarrow \pi_0 = \frac{\mu - \lambda}{\mu} \Rightarrow \pi_k = \frac{\mu - \lambda}{\mu} \left(\frac{\lambda}{\mu}\right)^k$$

# TEORIA DAS FILAS

## FÓRMULAS DO SISTEMA M/M/1

Descrição	Fórmulas
Probabilidade de que o Sistema Esteja Ocupada	$\rho = \frac{\lambda}{\mu}$
Probabilidade de que $n$ Clientes Encontram-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	$L = \frac{\lambda}{\mu - \lambda}$
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	$W_q = \frac{\lambda}{\mu(\mu - \lambda)}$

# PARTE PRÁTICA

## APLICAÇÕES NO MUNDO REAL

# TEORIA DAS FILAS

## TIME-DEPENDENT QUEUE

Em uma fila "time-dependent", a taxa de chegada e/ou taxa de serviço muda ao longo do tempo. Isto pode modelar cenários como restaurantes ou lojas de varejo, onde o fluxo de clientes varia significativamente dependendo da hora do dia ou da semana.

Modelos Markovianos de tempo contínuo, como cadeias de Markov de tempo contínuo não-homogêneas, são utilizados para analisar tais filas.

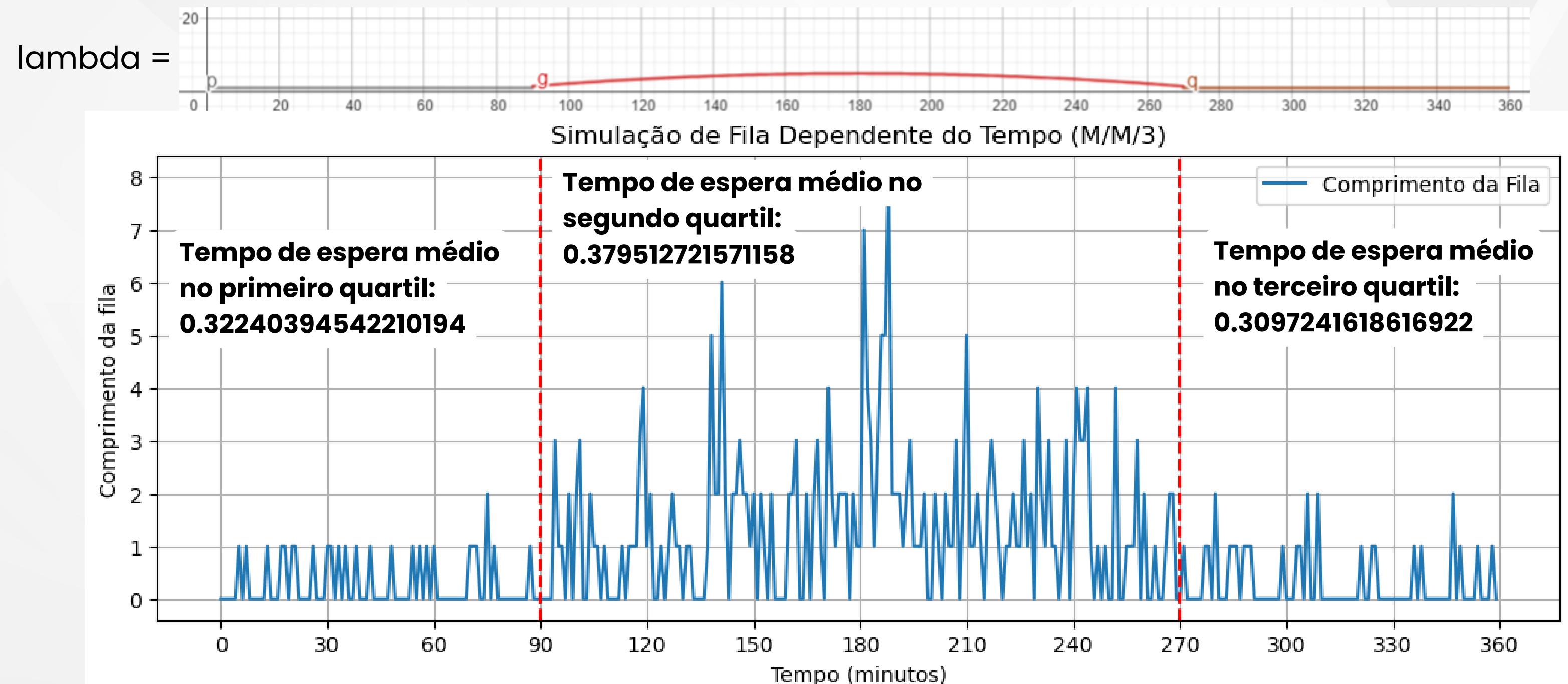
Parâmetros da Simulação:

- Servidores = 3
- Duração = 360 minutos (6H)
- Taxa de serviço ( $\mu$ ) = 3

# TEORIA DAS FILAS

## TIME-DEPENDENT QUEUE (M/M/3)

Resultados da Simulação:



# TEORIA DAS FILAS

## TIME-DEPENDENT QUEUE:

### CONCLUSÕES E INSIGHS

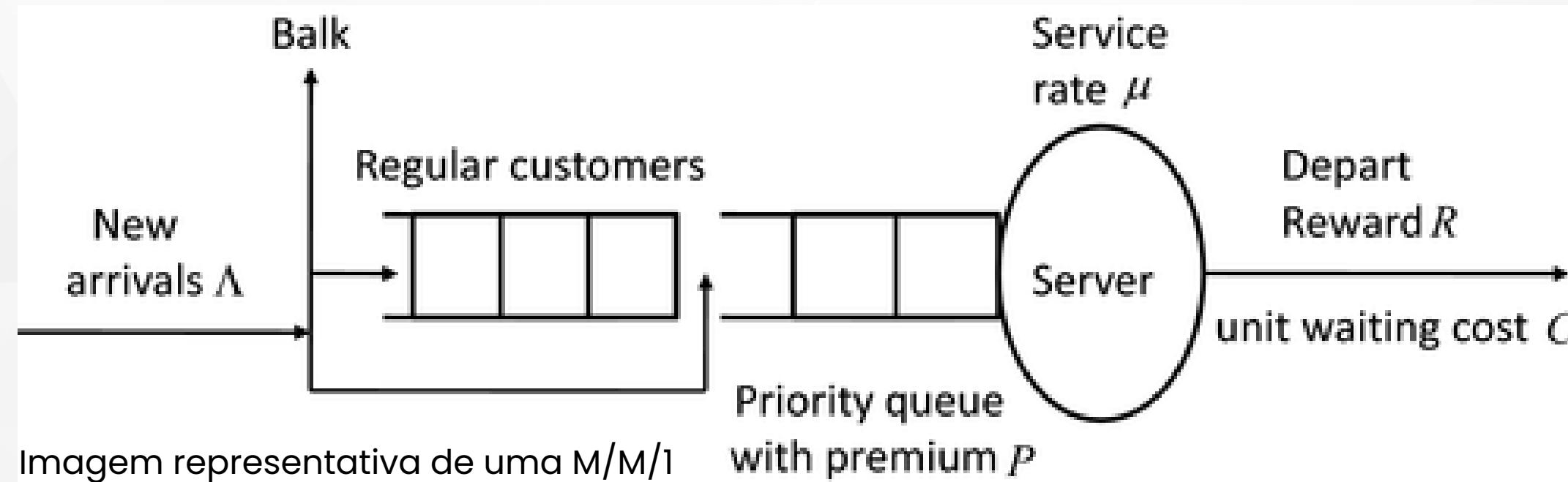
- Ausência de distribuição estacionária
- Análises em intervalos de tempo
- Métricas de performance que variam de acordo com o tempo
- Grande dificuldade em encontrar soluções analiticamente

# TEORIA DAS FILAS PRIORITY-QUEUE

- Uma fila de prioridade é semelhante uma fila regular, mas cada elemento possui uma prioridade associada.
- Clientes com alta prioridade são servidos antes de clientes com baixa prioridade.

Na maioria das implementações, se dois elementos tiverem a mesma prioridade, eles serão servidos na mesma ordem em que foram enfileirados.

- É um tipo de fila recorrente em sistemas operacionais, gerenciamento de tarefas, situações de planejamento e logística.



# TEORIA DAS FILAS

## PRIORITY-QUEUE (M/M/2)

Parâmetros da Simulação:

Servidores = 2

Duração = 360 minutos (6H)

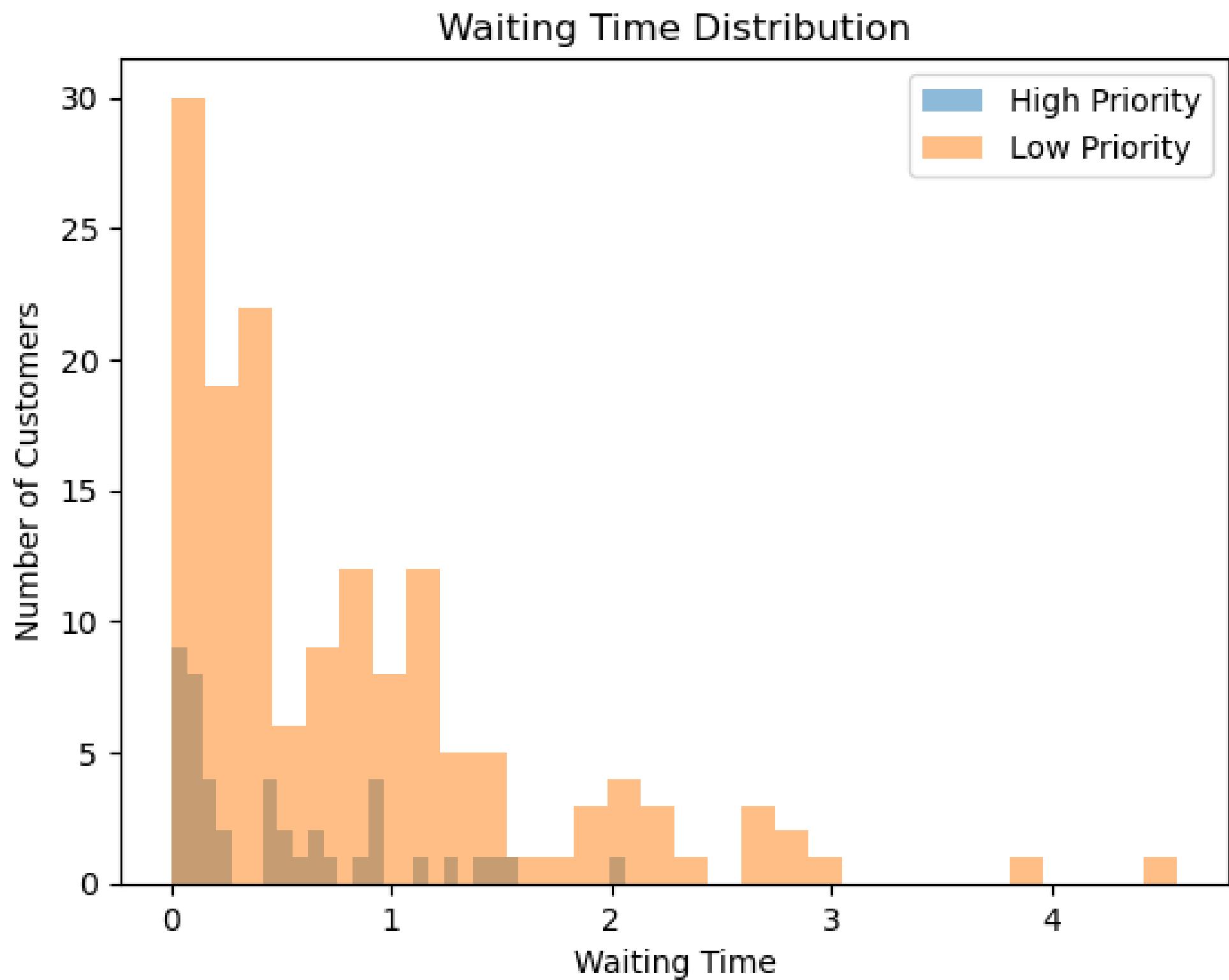
$\lambda_H = 0.6$

$\lambda_L = 1.2$

$\mu = 1$

Condições: Se houver um cliente de alta prioridade na fila ele sempre será o primeiro a ser atendido.

Para clientes de mesma prioridade, será servido quem chegou primeiro.



# TEORIA DAS FILAS **PRIORITY-QUEUE:** **CONCLUSÕES E INSIGHS**

- O Paradoxo do tempo de espera
- A diferença na intensidade do tráfego
- A importância do manejamento para clientes de baixa prioridade não serem excessivamente penalizados
- Custo x Benefício

# TEORIA DAS FILAS

## RETRIAL-QUEUE

Retrial Queues ocorrem em serviços onde cliente tentam novamente mais tarde, ao se deparar com servidor(es) ocupado(s). Existem muitas variações, mas as características principais são:

- Qualquer cliente, ao encontrar o servidor ocupado, se torna um cliente retrial (em órbita)
- Clientes em órbita não sabem o estado das filas e apenas saberão se a fila está vazia ou não quando chegarem ao servidor.
- Clientes entram e saem da órbita até serem atendidos ou desistirem.

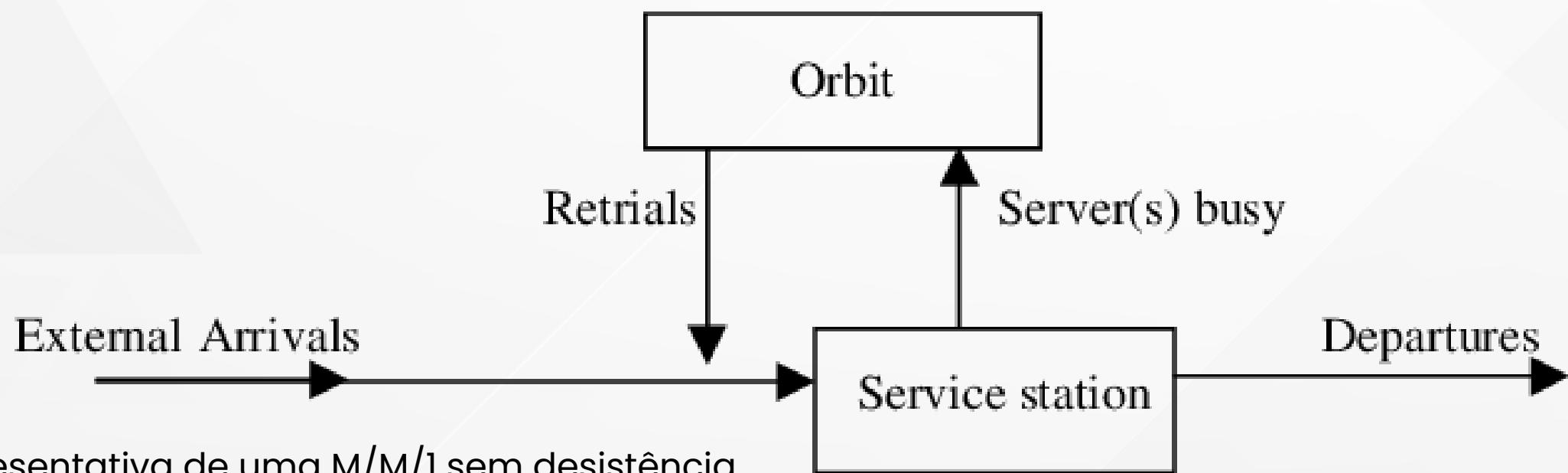
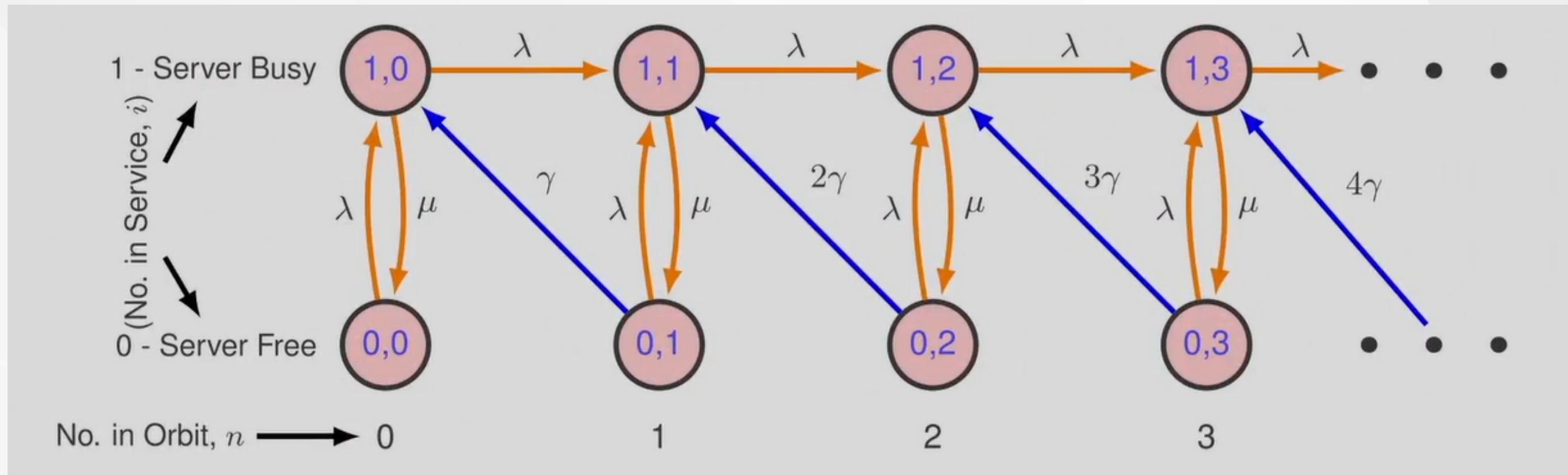


Imagen representativa de uma M/M/1 sem desistência

# TEORIA DAS FILAS RETRIAL-QUEUE (M/M/1) SEM DESISTÊNCIAS

- gamma = taxa de "retrial"



Seja  $p_{i,n}$  a probabilidade de que o sistema esteja no estado  $(i, n)$ . Então elas satisfazem:

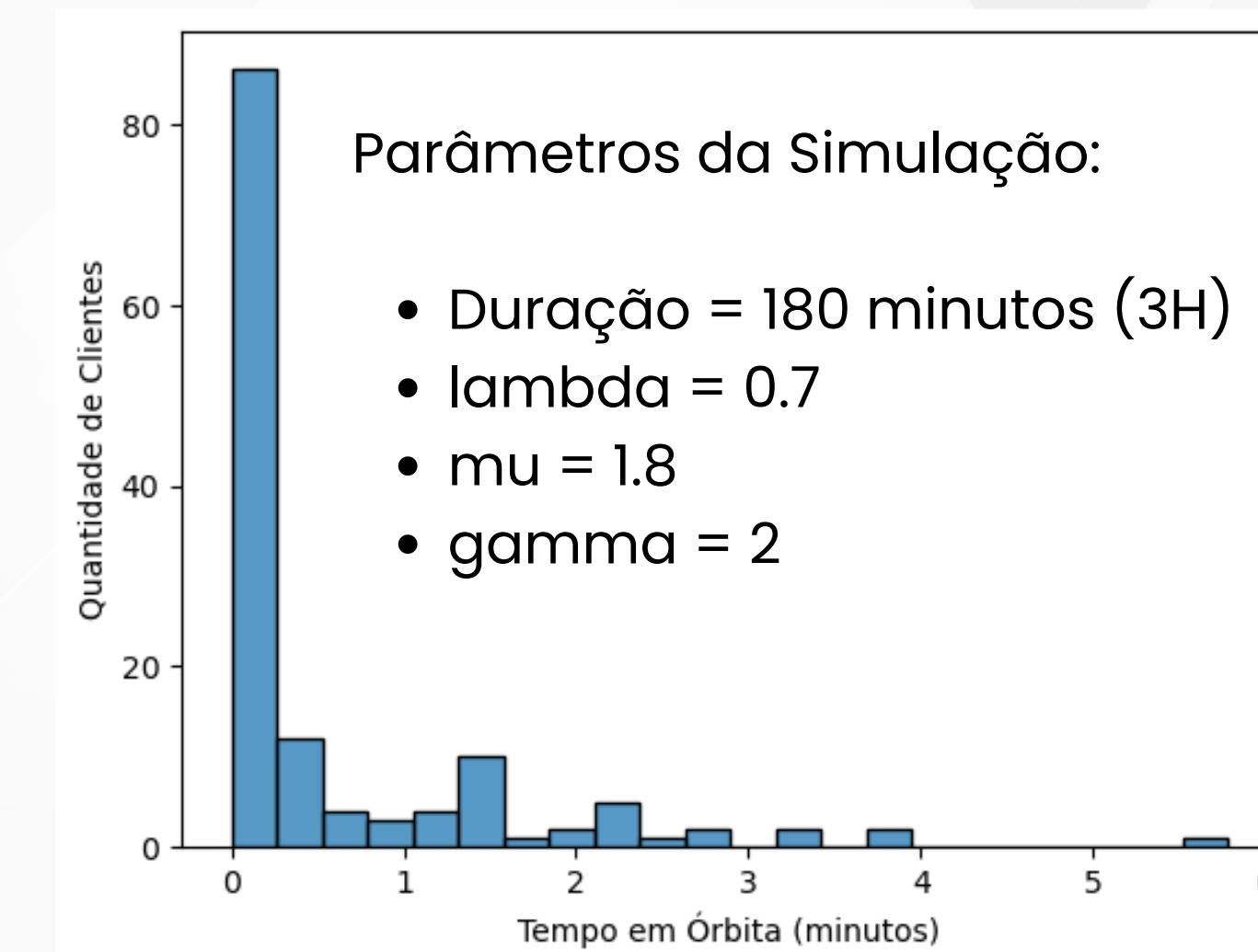
$$(\lambda + n\gamma)p_{0,n} = \mu p_{1,n}, \quad n \geq 0$$

$$(\lambda + \mu)p_{1,n} = \lambda p_{0,n} + (n+1)\gamma p_{0,n+1} + \lambda p_{1,n-1}, \quad n \geq 1$$

$$(\lambda + \mu)p_{1,0} = \lambda p_{0,0} + \gamma p_{0,1}.$$

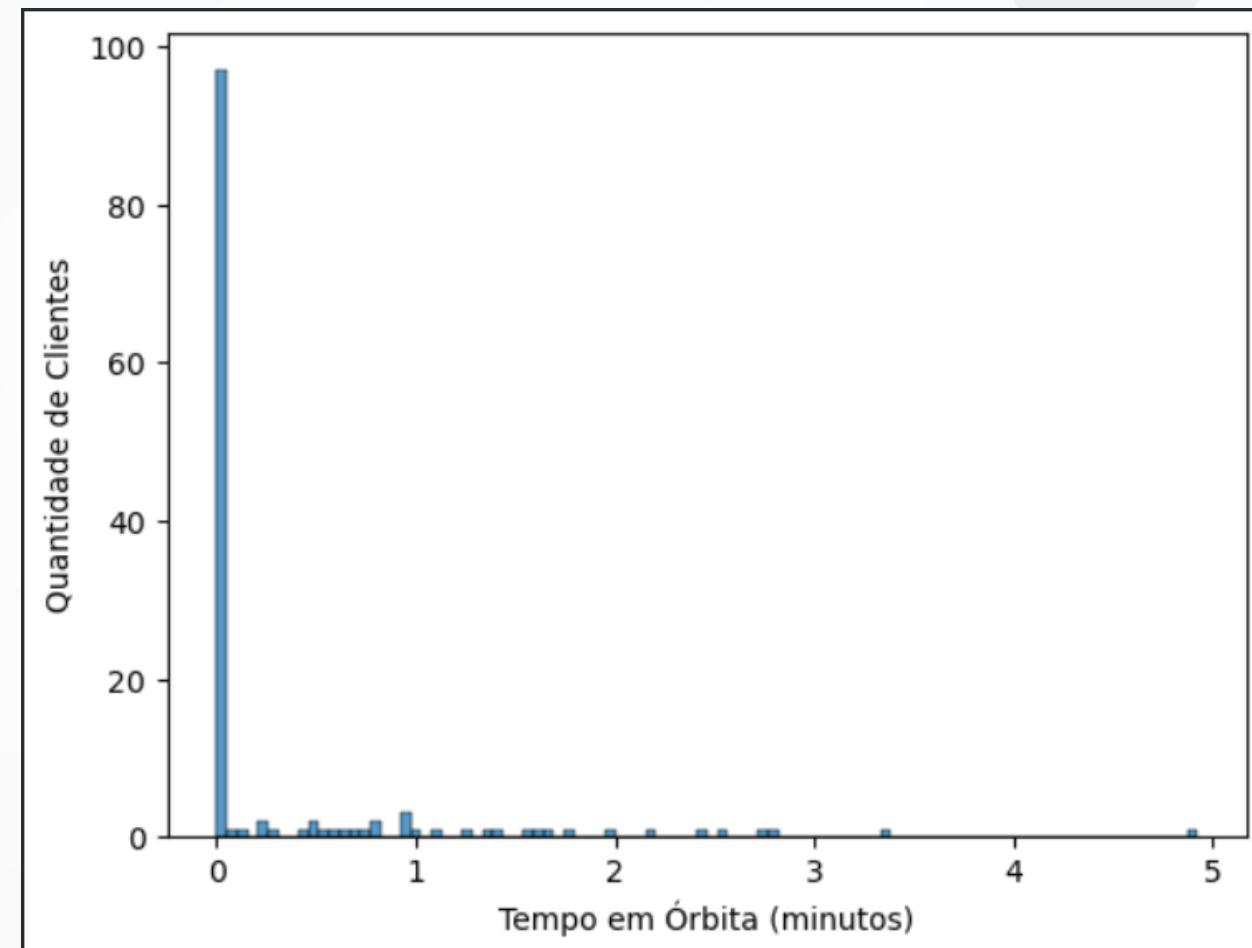
# TEORIA DAS FILAS RETRIAL-QUEUE (M/M/1) SEM DESISTÊNCIAS

Cliente 0 olhou para a fila em 3.434583719951421  
Fim de atendimento de cliente 0 em 4.626964219888548  
Cliente 1 olhou para a fila em 6.771997513986355  
Fim de atendimento de cliente 1 em 6.815216213835543  
Cliente 2 olhou para a fila em 7.17170364234576  
Fim de atendimento de cliente 2 em 7.876683989474547  
Cliente 3 olhou para a fila em 9.246932950497403  
Fim de atendimento de cliente 3 em 9.286580049089157  
Cliente 4 olhou para a fila em 13.710212888425257  
Fim de atendimento de cliente 4 em 13.745431041824162  
Cliente 5 olhou para a fila em 15.20529981205835  
Fim de atendimento de cliente 5 em 16.070963488500254  
Cliente 6 olhou para a fila em 16.761047852092577  
Fim de atendimento de cliente 6 em 17.376840683423364  
Cliente 7 olhou para a fila em 17.31375474436388  
Cliente 7 olhou para a fila em 17.359543378058977  
Cliente 7 olhou para a fila em 17.517132979125172  
Fim de atendimento de cliente 7 em 17.967732616810252  
Cliente 8 olhou para a fila em 17.53343082125849  
Cliente 9 olhou para a fila em 18.123409863742047  
Fim de atendimento de cliente 9 em 19.042394995814277  
Cliente 8 olhou para a fila em 18.157123796165305  
Cliente 10 olhou para a fila em 18.442804681188218  
Cliente 10 olhou para a fila em 18.651348034344124  
Cliente 8 olhou para a fila em 18.770666727772433



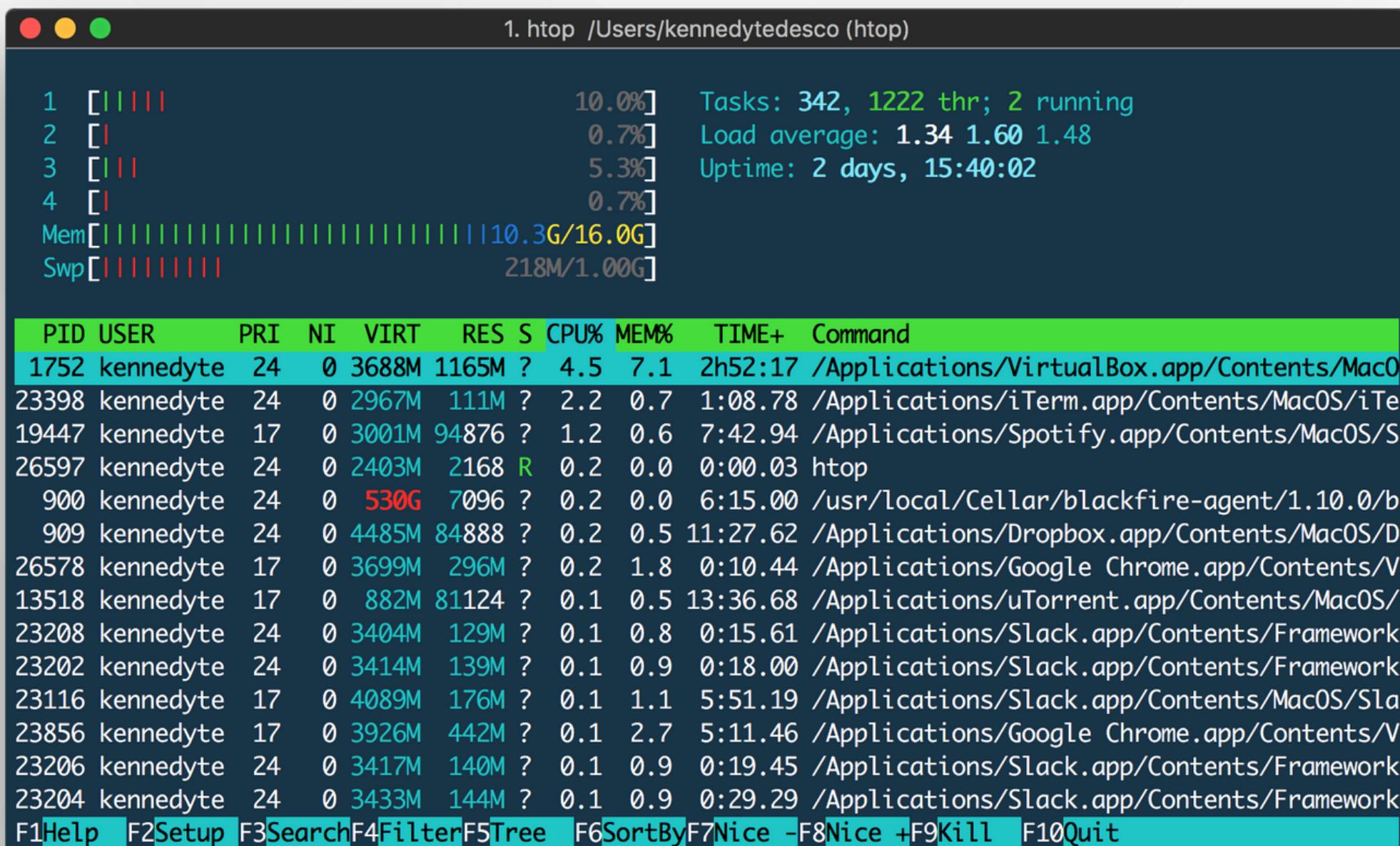
# TEORIA DAS FILAS RETRIAL-QUEUE: CONCLUSÕES E INSIGHS

- O impacto do parâmetro gamma: →  
Retrial Rate
- Possibilidade de uma órbita infinita
- A importância de manejamento de serviços e recursos para a satisfação do cliente



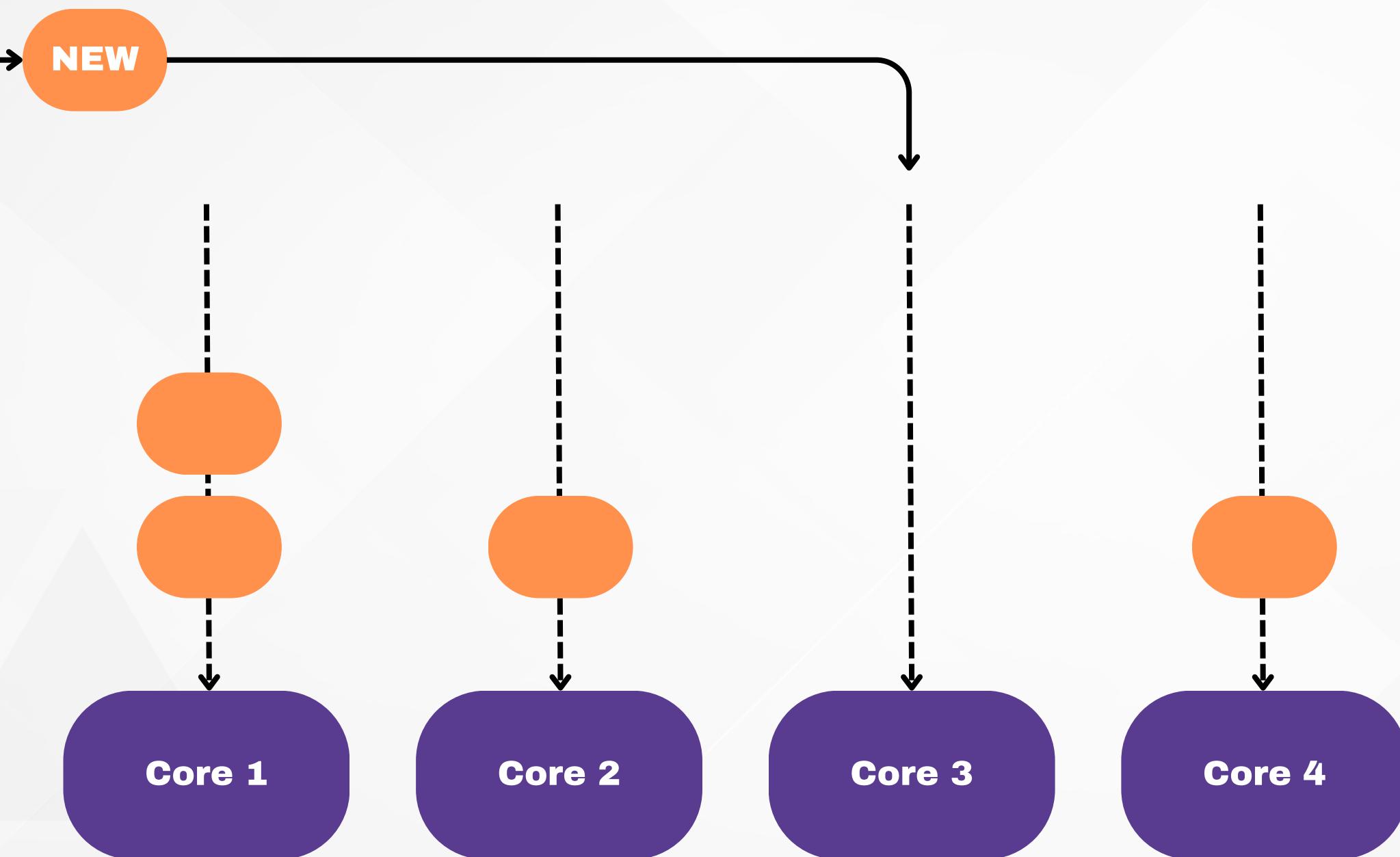
# TEORIA DAS FILAS

# CORE SCHEDULING



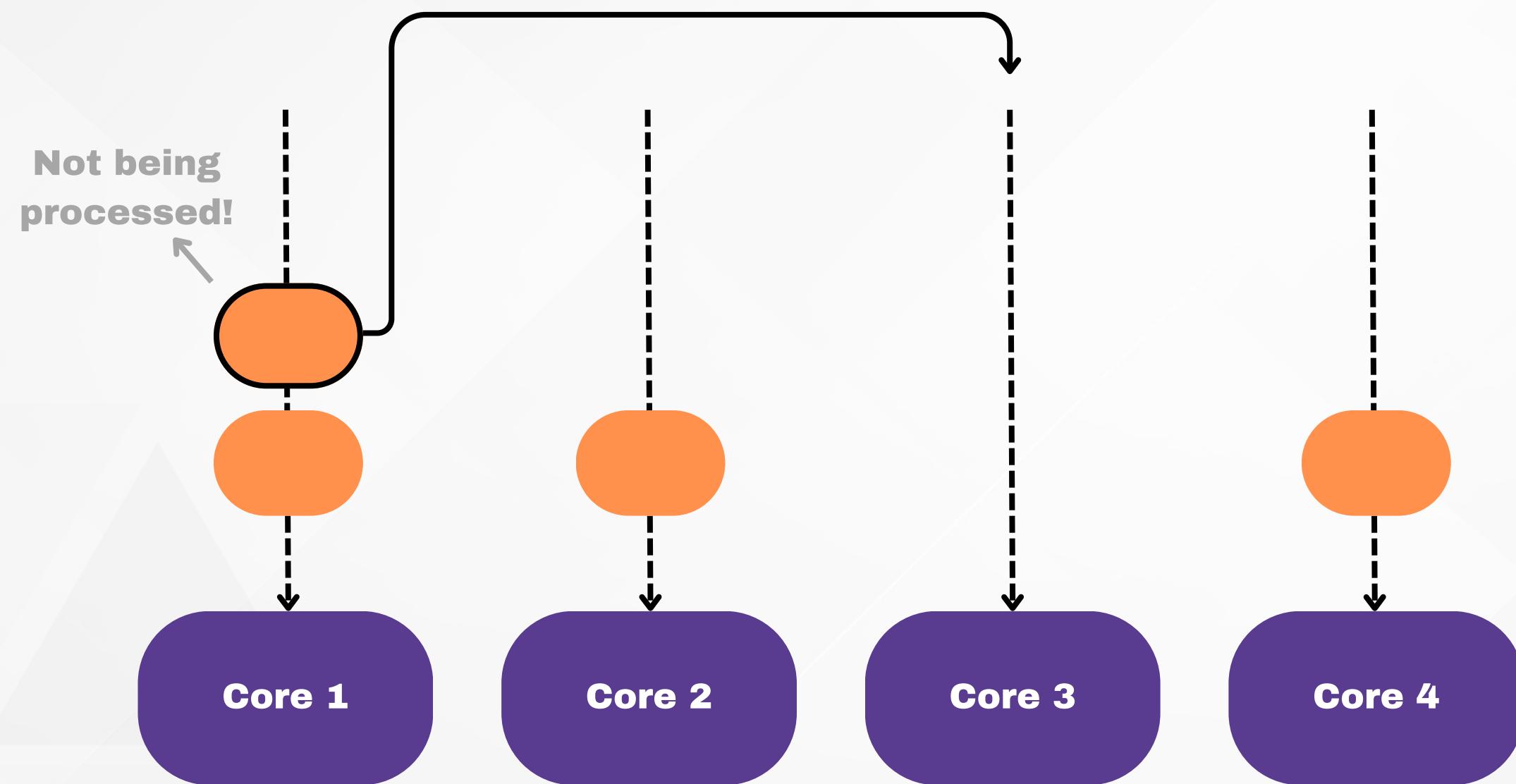
# TEORIA DAS FILAS

# WORK-SHARING SCHEDULING



# TEORIA DAS FILAS

## WORK-STEALING SCHEDULING



# TEORIA DAS FILAS NOSSAS SIMULAÇÕES

---

- Tarefas chegam com tempo entre chegadas Exponencial(1/3)
  - + 20 tarefas em backlog
- Tempo de execução de tarefas amostrado de uma Exponencial(1)  
*(quão realista é isso?)*
- 8 Processadores
  - Os processadores 0, 3 e 5 -- mas especialmente o 0 -- tem uma concentração maior de tarefas
- Work sharing vs. work stealing
  - Um caso de *jockeying!*

# TEORIA DAS FILAS

## SCHEDULING: TIME UNTIL DONE

---

Estimamos, ao longo de 1000 simulações, o tempo até a fila ficar vazia novamente -- i.e., até termos completado todas as tarefas.

### Work Sharing

**Média:** 14.08

**Desvio padrão:** 5.93

**Quantil 0.05 .. 0.95:** 9.13 .. 25.36

### Work Stealing

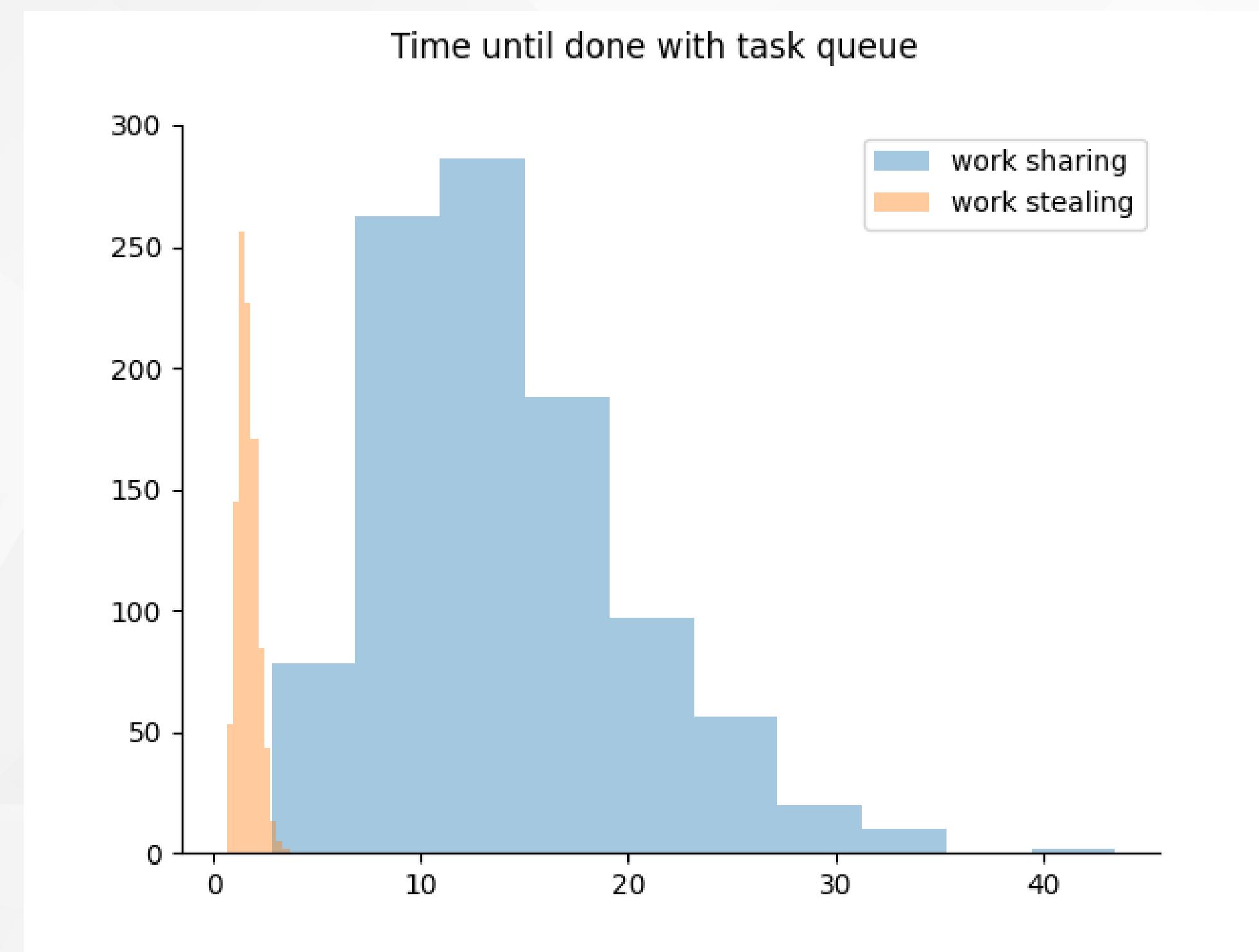
**Média:** 1.64

**Desvio Padrão:** 0.48

**Quantil 0.05 .. 0.95:** 0.91 .. 2.52

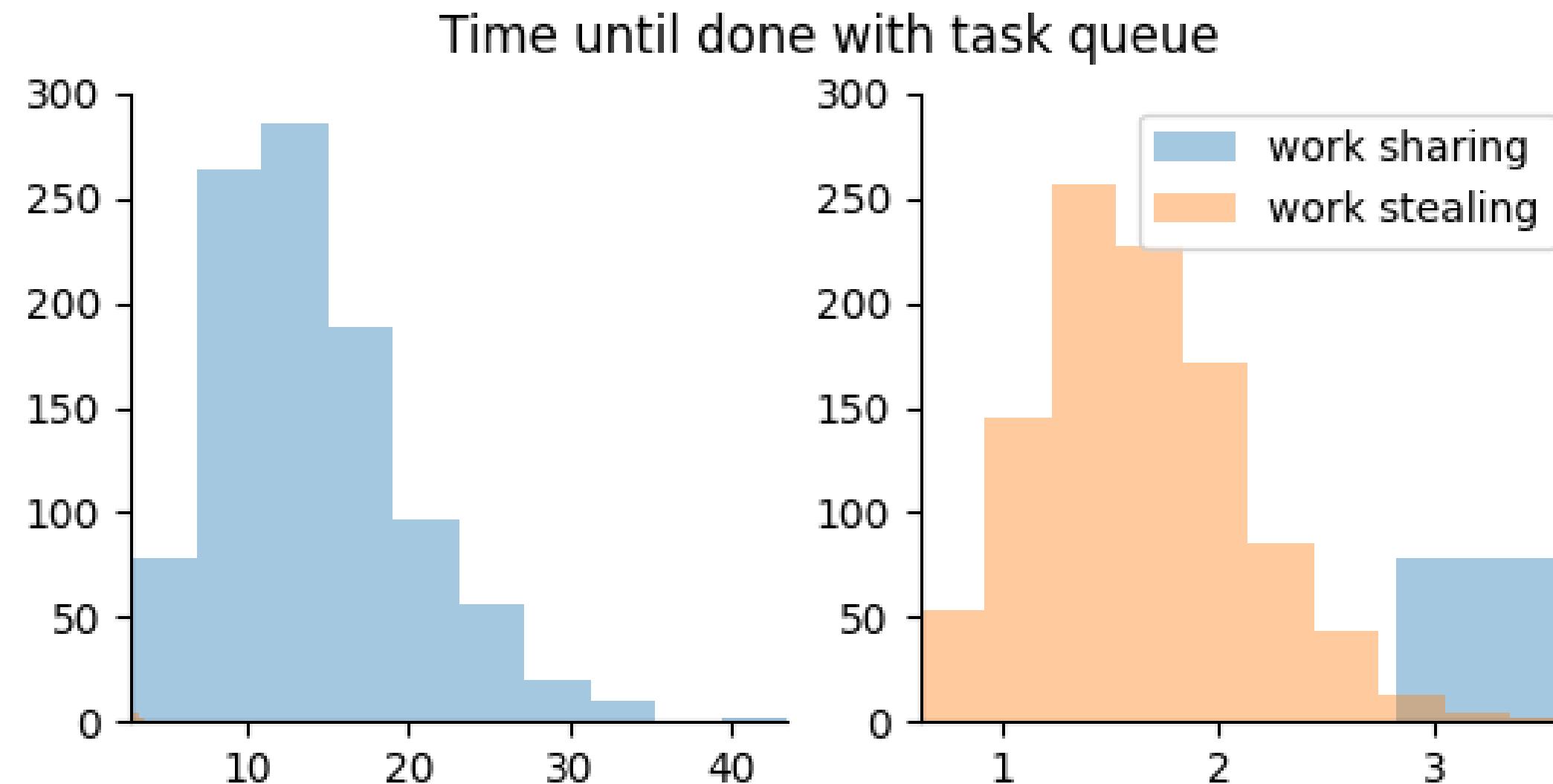
# TEORIA DAS FILAS

## SCHEDULING: TIME UNTIL DONE



# TEORIA DAS FILAS

## SCHEDULING: TIME UNTIL DONE



# TEORIA DAS FILAS

## SCHEDULING: TIME UNTIL PROCESSED

---

Estimamos também, ao longo de 1000 simulações, o tempo entre uma tarefa ser inserida e ser executada.

### Work Sharing

**Média:** 3.88

**Desvio padrão:** 4.53

**Quantil 0.05 .. 0.95:** 0.01 .. 13.02

### Work Stealing

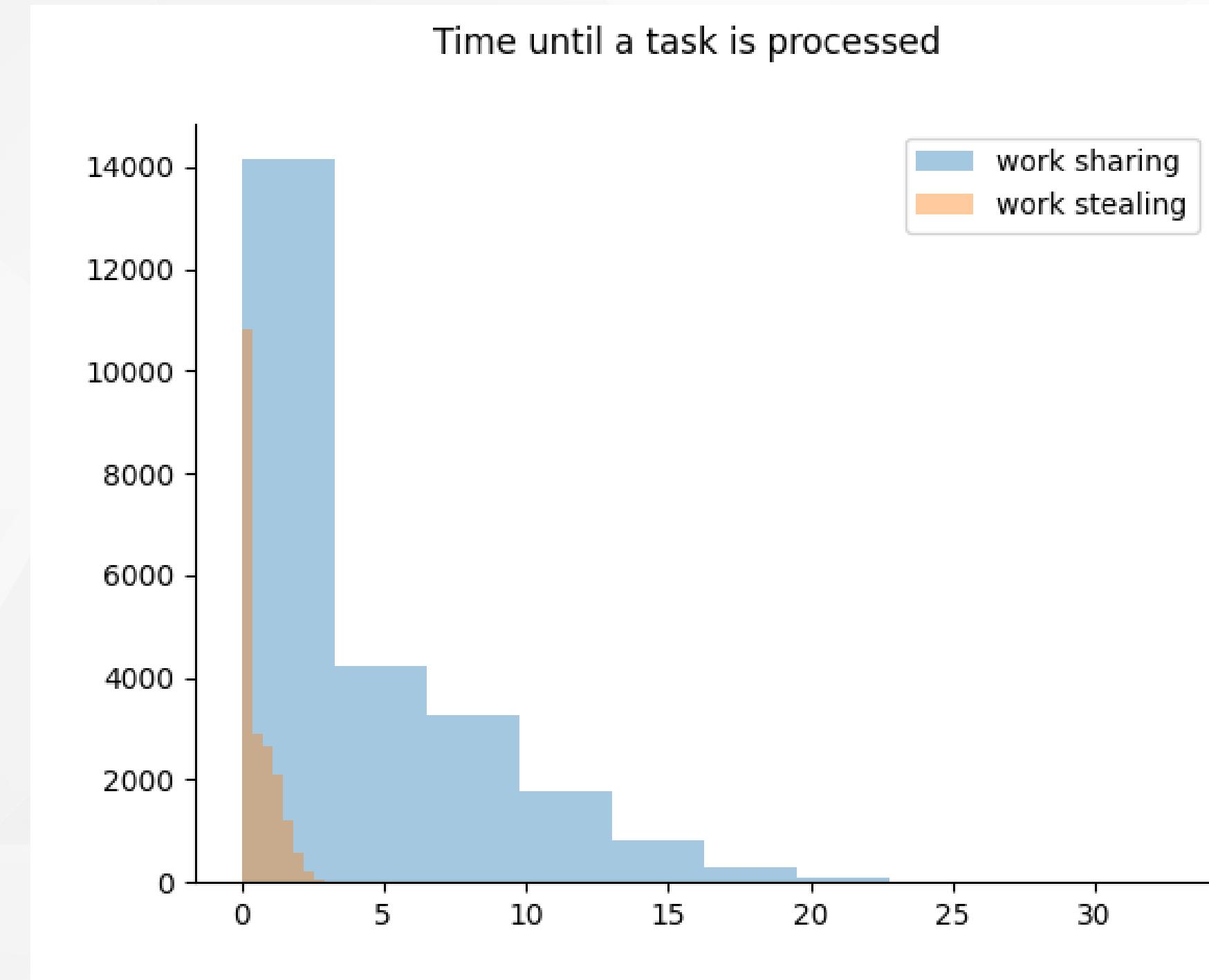
**Média:** 0.56

**Desvio Padrão:** 0.61

**Quantil 0.05 .. 0.95:** 0.02 .. 1.76

# TEORIA DAS FILAS

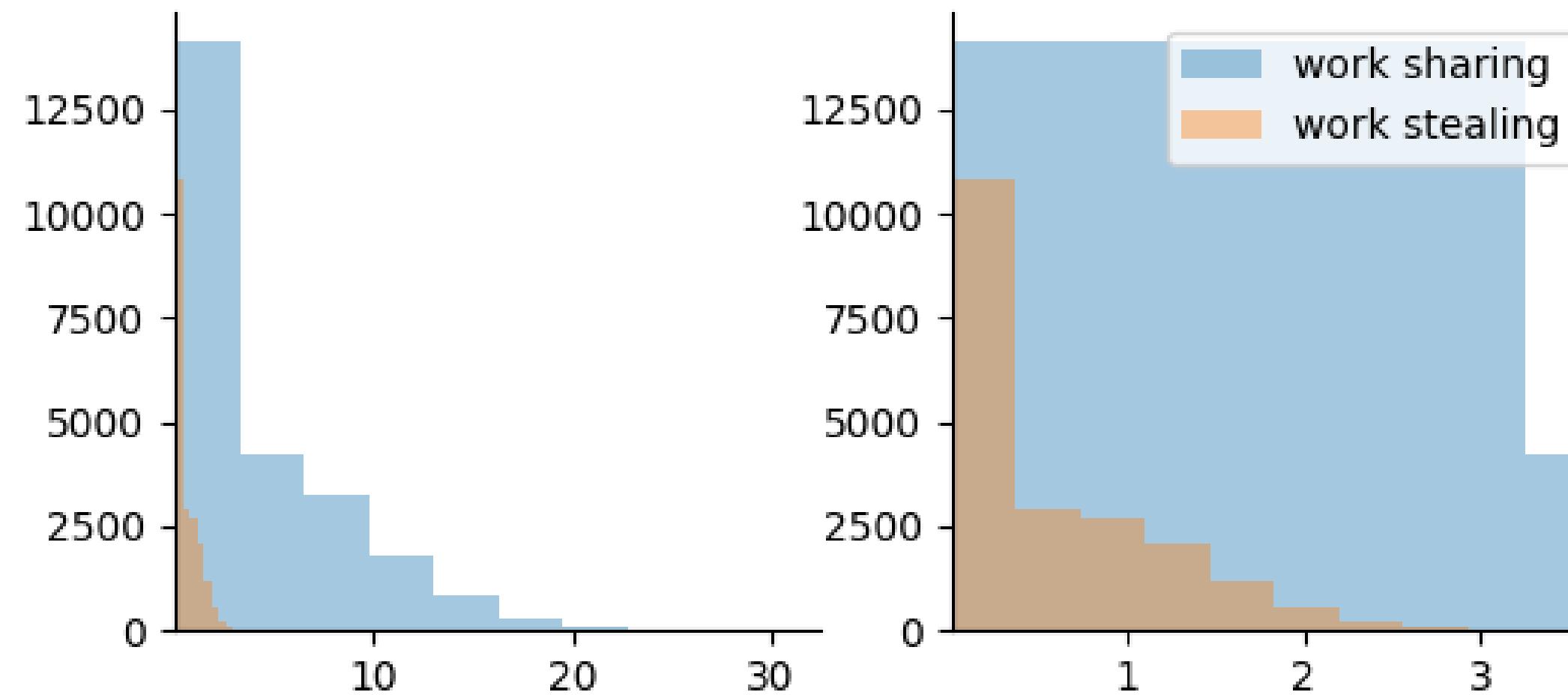
## SCHEDULING: TIME UNTIL PROCESSED



# TEORIA DAS FILAS

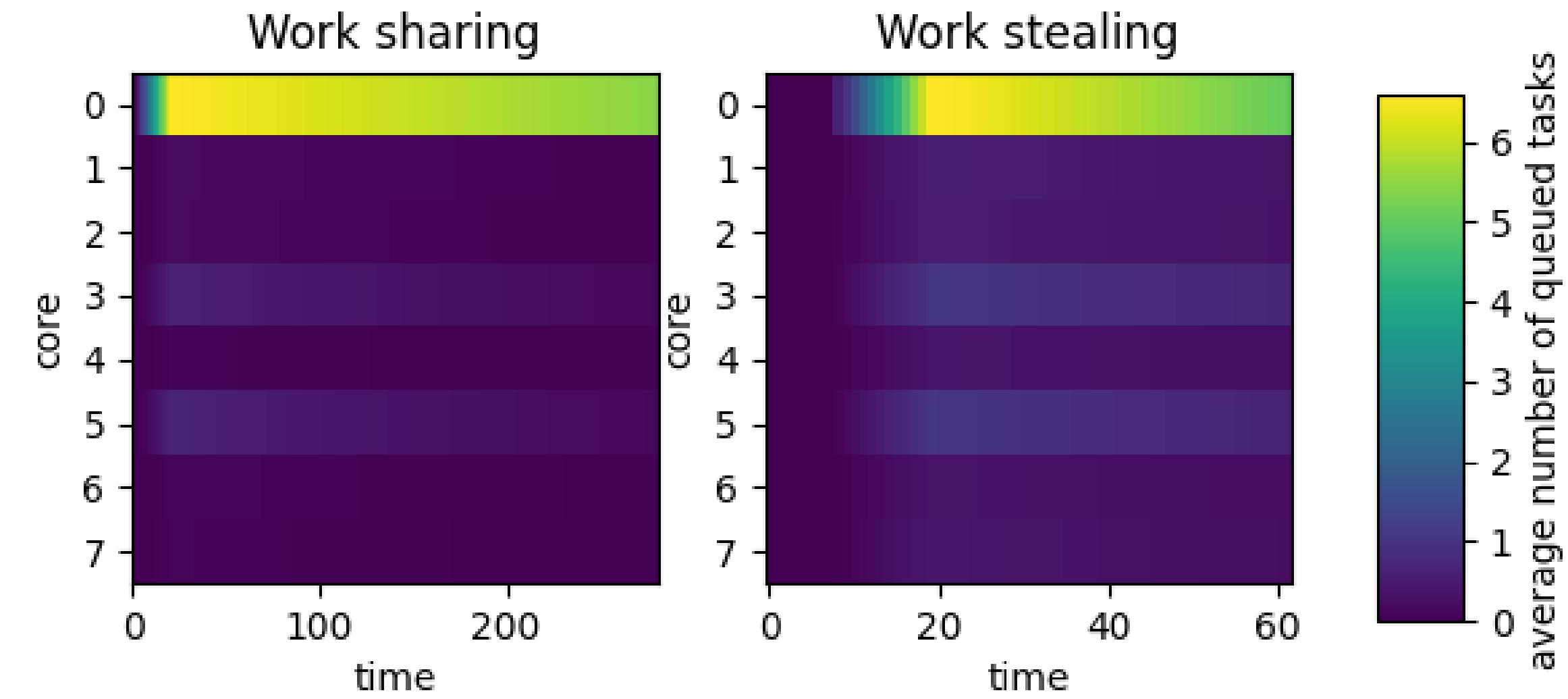
## SCHEDULING: TIME UNTIL PROCESSED

Time until a task is processed



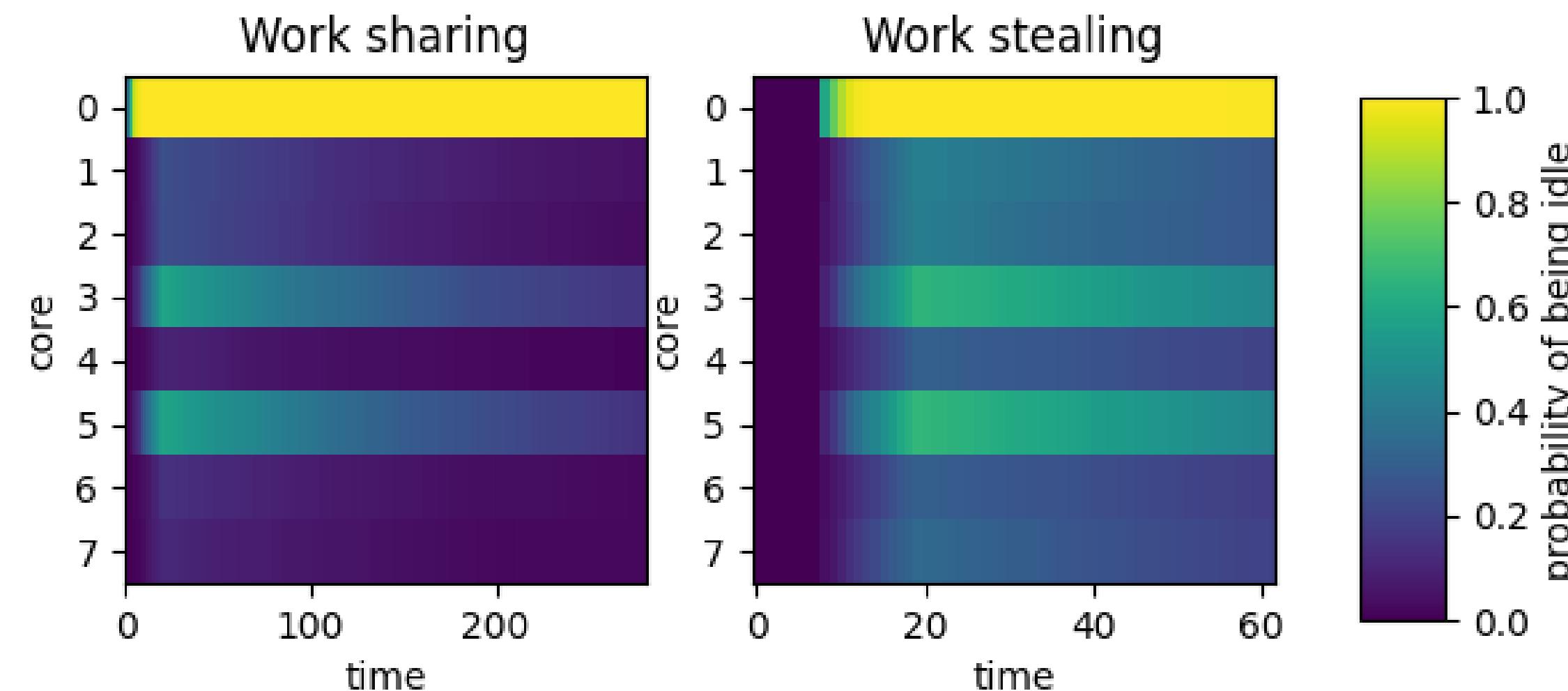
# TEORIA DAS FILAS

## SCHEDULING: QUEUE SIZE PER CORE



# TEORIA DAS FILAS

## SCHEDULING: IDLE CORES



# TEORIA DAS FILAS

## PRINCIPAIS REFERÊNCIAS

---

- ANÁLISE DE TEORIA DAS FILAS: SISTEMA DE FILAS DE UM SERVIÇO DE PRONTO ATENDIMENTO (Francieli de Fátima de Oliveira)
- Introduction to Queuing Theory (Robert B. Cooper)
- Scheduling Multithreaded Computations by Work Stealing (Robert D. Blumofe e Charles E. Leiserson)
- “M/M/1 Retrial Queues” <https://www.youtube.com/watch?v=xTkbnJ-Qupc>