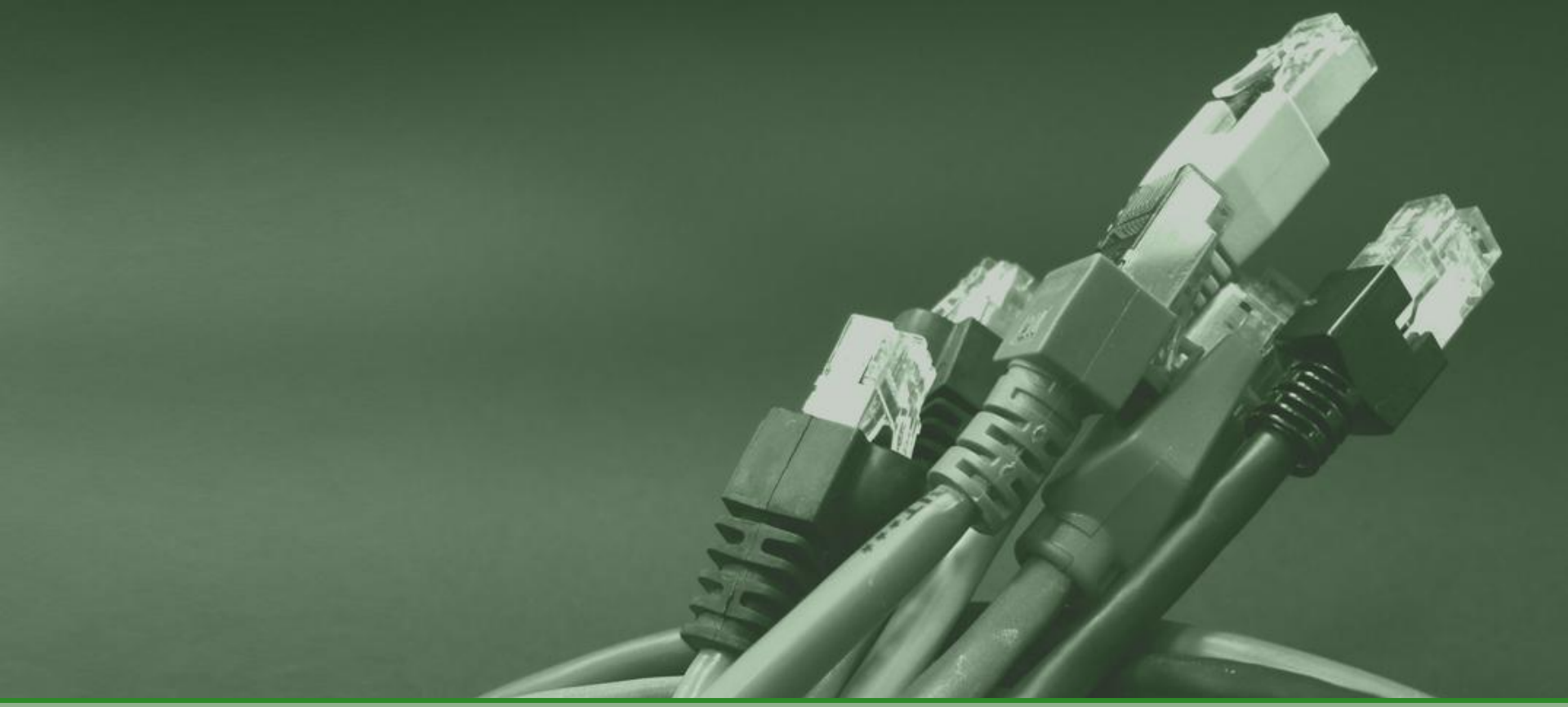


Teoria de Filas

Prof. Gustavo Leitão

INSTITUTO FEDERAL DE
EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
RIO GRANDE DO NORTE

Campus Natal Central.
Planejamento de Capacidade de Sistemas

Objetivo da Aula

Objetivo

Apresentar os conceitos de teoria de filas e suas aplicações

Introdução

Modelo

Um modelo é uma abstração de um sistema real

O nível de detalhe do modelo depende do propósito do modelo.

Por exemplo, se o objetivo é prever o que aconteceria se mais memória foram adicionados ao sistema, pode não ser necessário para o modelo (ou mesmo entender completamente) a estratégia de escalonamento do disco

Teoria de Filas

INTRODUÇÃO

Todas as pessoas já passaram pelo aborrecimento de ter que esperar em uma fila

Fila de ônibus, banco, padaria, trânsito, restaurante, etc.

Em sistemas computacionais há filas por toda parte:

- Acessar CPU
- O Dico
- A memória
- Impressora
- Rede

As filas surgem porque a demanda de serviço é maior que a capacidade de atendimento do sistema

O QUE É TEORIA DAS FILAS?

Ramo da probabilidade que estuda o fenômeno da formação de filas de solicitações de serviços

Permite estimar importantes medidas de desempenho de um sistema a partir de propriedades mensuráveis das filas

Dessa forma, pode-se dimensionar um determinado sistema segundo a demanda dos seus clientes, evitando desperdícios ou gargalos

Contudo, filas apresentam comportamento estocástico

Para que?

Provê modelos para prever o comportamento de sistemas que oferecem serviço para demandas com taxas de chegadas aleatórias

Utilizada para modelar sistemas onde:

- Clientes chegam para ser atendidos
- Esperam sua vez de ser atendidos
- São atendidos e vão embora

Resultados Possíveis

- Tempo de espera de um cliente
  - Quanto tempo um cliente espera no banco
  - Quanto tempo um pacote passa em um roteador
- Acúmulo de clientes na fila
  - Qual o tamanho médio da fila do banco
  - Como a fila do roteador se comporta
- Tempo ocioso/ocupado dos servidores
  - Quanto tempo o caixa fica livre
  - Qual a utilização do roteador
- Taxa de saída (vazão)
  - Quantos clientes são atendidos por hora
  - Quantos pacotes são encaminhados por segundo

Aplicações

Fluxo de tráfego

- Veículos
- Pessoas
- Redes de Comunicação

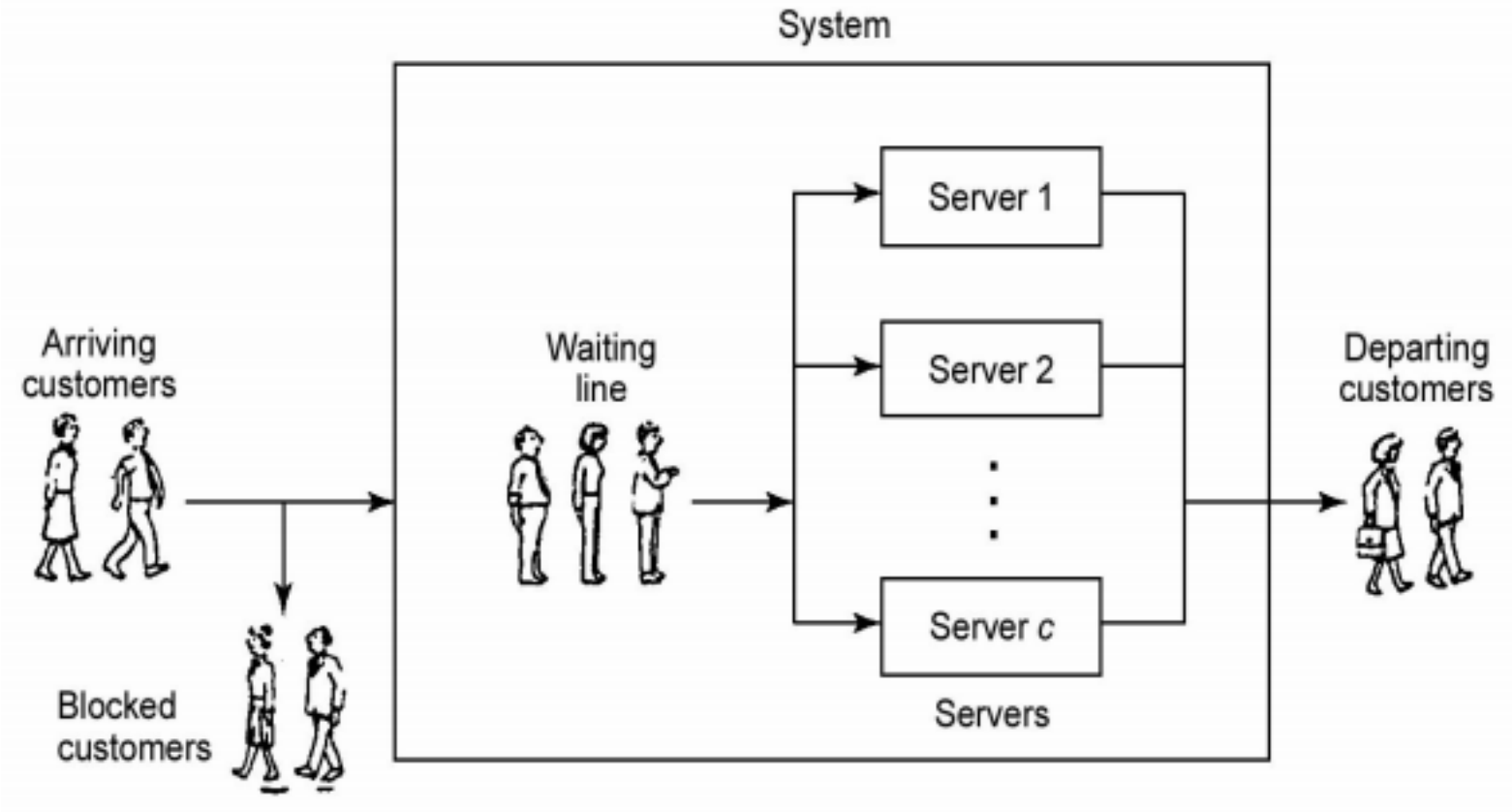
Escalonamento

- Paciente
- Tarefas
- Processos

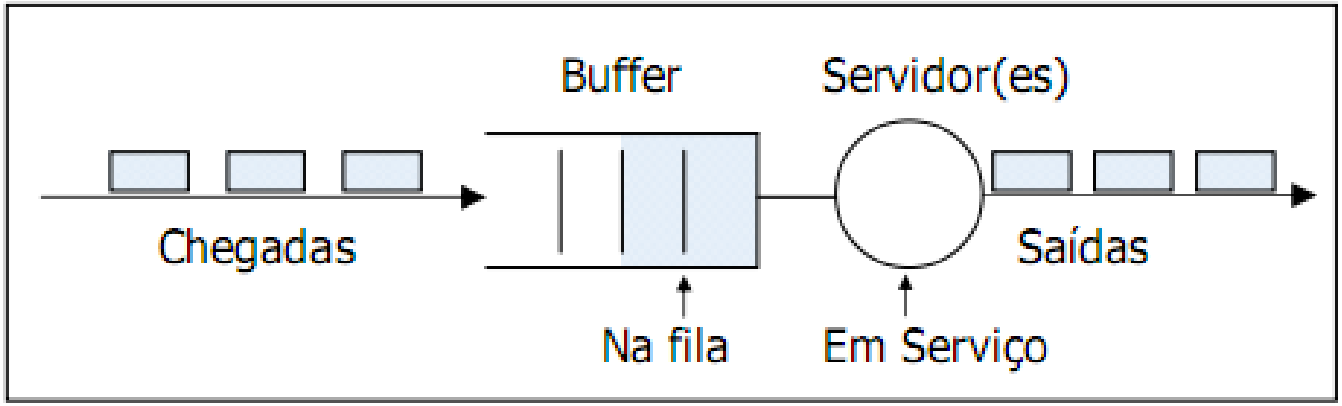
Serviço de Atendimento

- Banco
- Restaurante
- Servidores

Sistema de Filas



Modelo de Filas Básico



- Modela qualquer serviço com:
  - Um ou mais servidores
  - Uma área de espera (buffer)
- "Clientes" chegam para receber um "serviço"
- Um cliente que não encontra um servidor livre espera na fila (buffer)

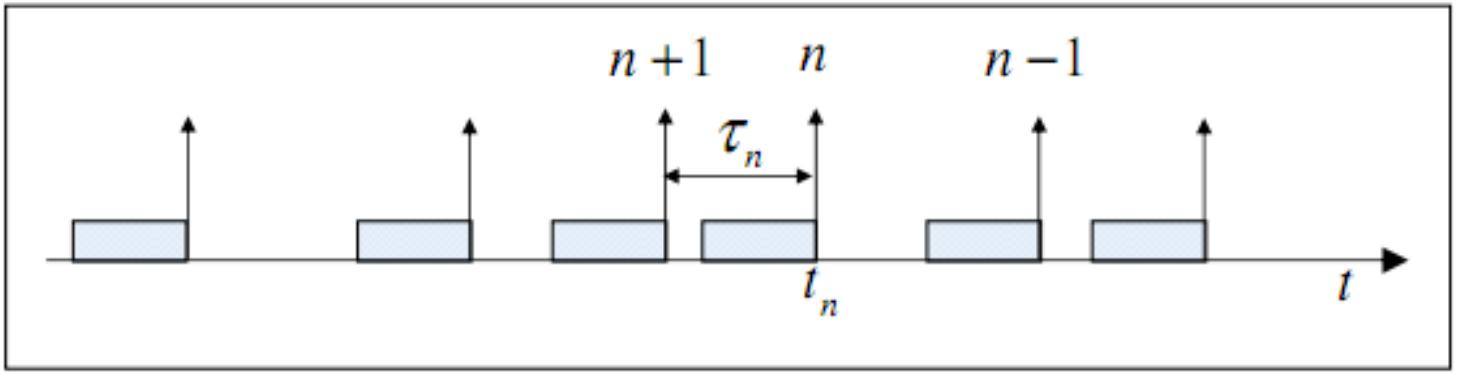
Característica das Filas

Processo de Chegada

- Normalmente é um processo estocástico
- Necessário saber a distribuição de probabilidade do tempo entre chegadas
  - Normalmente é Exponencial
- Processo Estacionário
  - A distribuição de probabilidade que descreve a chegada **não varia** com o tempo (é independente do tempo)
- Processo Não Estacionário
  - A distribuição **varia** com o tempo (depende do tempo)

Caracteristica das Filas

Processo de Chegada



- τ_n tempo decorrido entre as chegadas dos clientes *n* e *n+1*
- $\{\tau_n,\ n \geq 1\}$ é um processo estocástico
- Tempos entre chegadas são identicamente distribuídos e têm a mesma média
- Taxa de chegada = λ

$$E[\tau_n] = E[\tau] = 1/\lambda$$

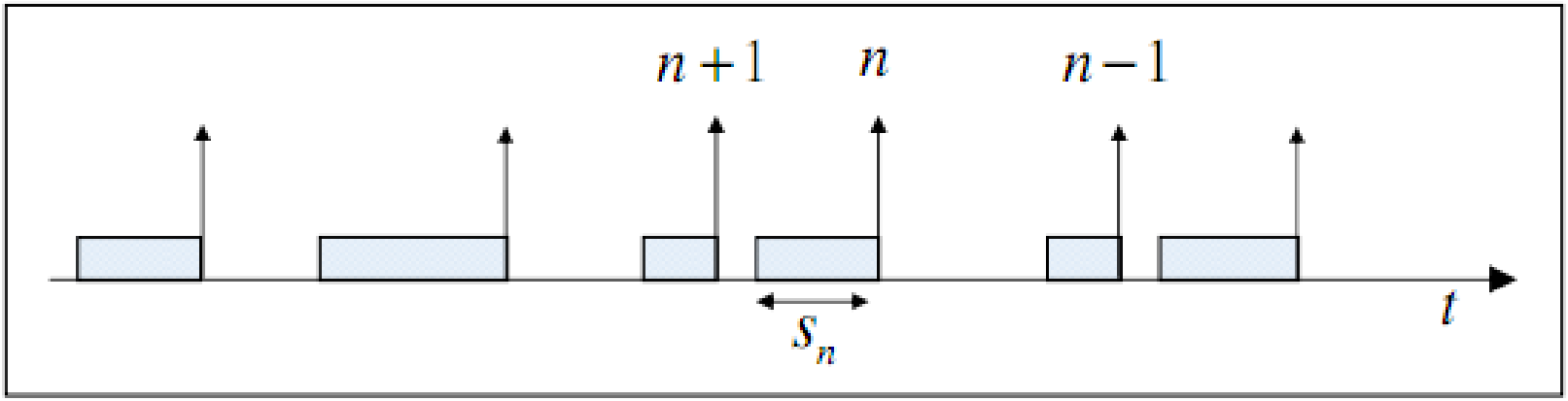
Caracteristica das Filas

Tempo de serviço

- Tempo que cada cliente leva para ser atendido
  - Ex: tempo que o cliente do banco passa no caixa
  - Ex: tempo para o roteador encaminhar um pacote
- Semelhante ao processo de chegada
- Distribuição de probabilidade para o **tamanho das filas** depende de:
  - o processo de chegada
  - o tempo de serviço

Caracteristica das Filas

Tempo de Serviço



- s_n é tempo que o cliente n passa no servidor
- $\{ s_n,\ n \geq 1 \}$ é um processo estocástico
- Tempos de serviço são identicamente distribuídos com uma média comum
- Taxa de serviço: μ $E[s_n] = E[s] = \mu$
- Os tempos de serviço são aleatórios para pacotes?

Característica das Filas

3. Número de Servidores: número de posições de atendimento disponíveis no sistema
   - Servidores idênticos / distintos
   - Fila única / por servidor / por grupo de servidores

4. Capacidade do Sistema: número máximo de clientes que podem permanecer no sistema, devido a restrições de espaço (buffers) ou de tempo de espera
   - Inclui clientes em serviço e esperando por serviço
   - Capacidade pode ser finita / infinita (mais fácil de analisar)

Característica das Filas

5. Tamanho da População (fonte): número potencial de clientes que podem chegar a um sistema
   - Tamanho finito / infinito

6. Disciplina de Atendimento (de fila): ordem na qual os clientes são atendidos
   Tipos:
   - FCFS (First-Come First-Served) ou FIFO
   - LCFS (Last-Come First-Served) ou LIFO
   - RR (Round Robin) / PS (Processor Sharing)
   - Prioridades (fila única / múltiplas filas)
   - Não-preemptivo / Preemptivo (resume/repeat) ...

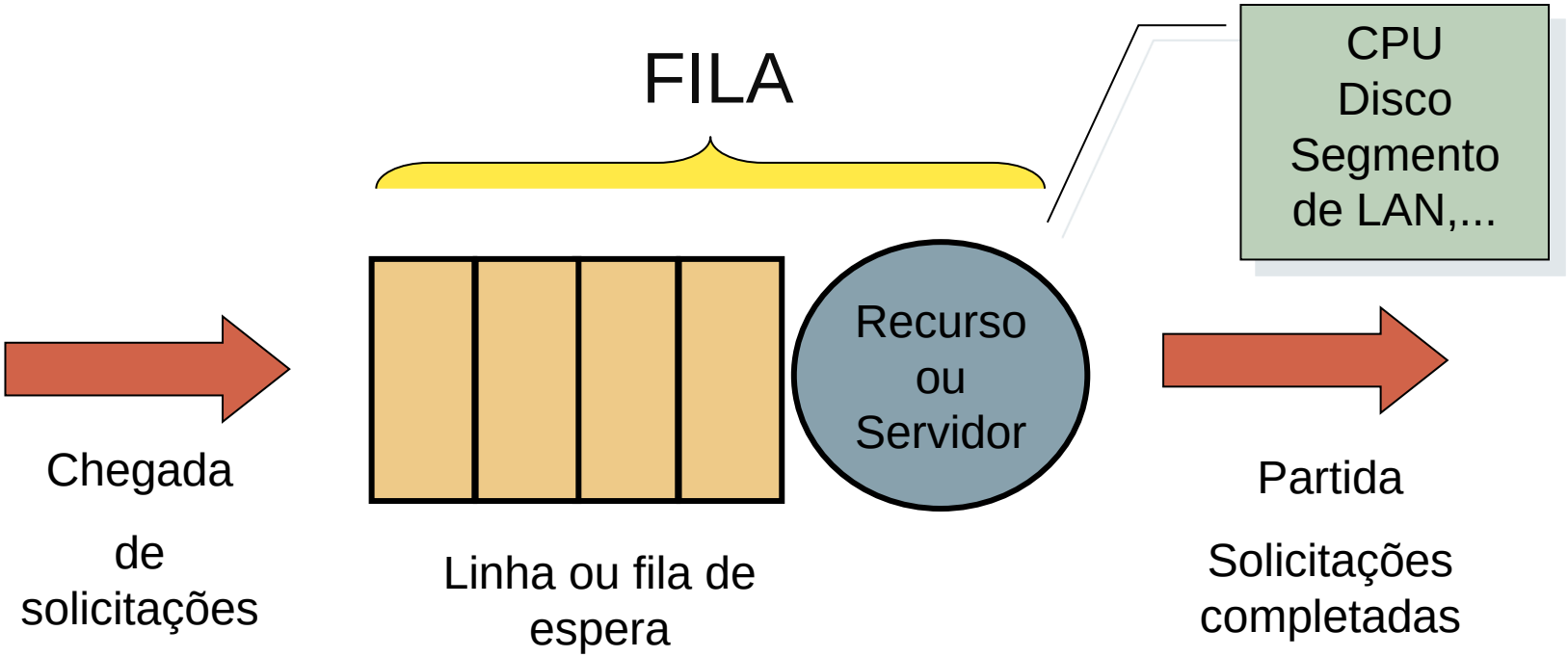
Notação Kendall

- A/S/NS/B/K/SD
  - A,S = Tempo entre chegadas, tempo de serviço
    - M = Exponencial (Markov, Memoryless)
    - Ek = Erlang
    - Hk = Hyperexponential
    - D = Determinístico
    - G = Geral (para todas as distribuições)
  - NS = Número de servidores
  - B = Número de buffers (lugares na fila)
  - K = Tamanho da população
  - SD = Disciplina de Serviço
    - FCFS,FCLS...
- Defaults B= ∞, K= ∞, SD=FCFS
- M/M/1 = M/M/1/∞/∞/FCFS

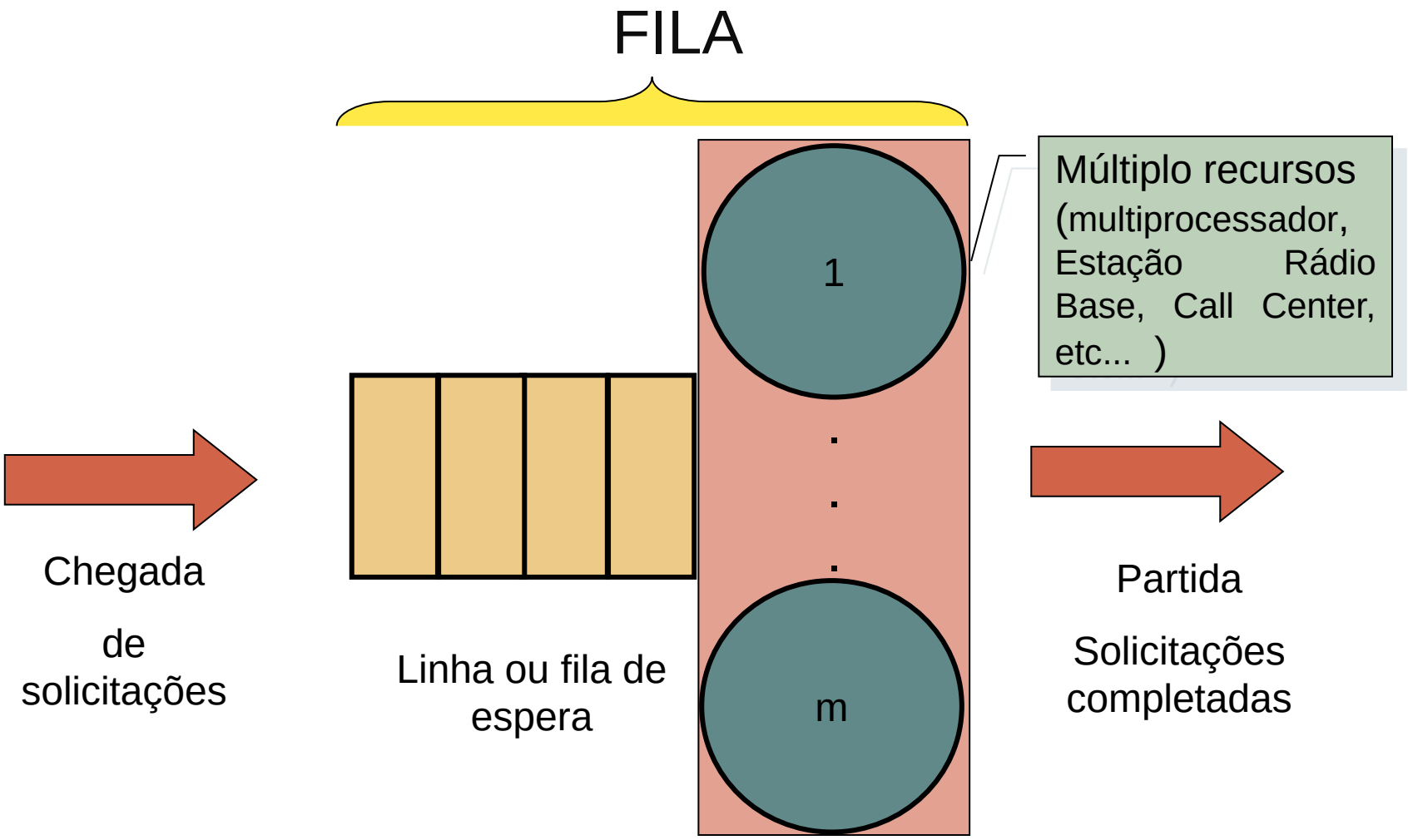
Notação Kendall

- M/M/1:
  - chegadas Poisson, tempo de serviço exponencial, 1 servidor, buffer infinito, FCFS
- M/M/m:
  - Igual ao anterior, com m servidores
- M/G/1:
  - chegadas Poisson, tempo de serviço geral, 1 servidor, buffer infinito

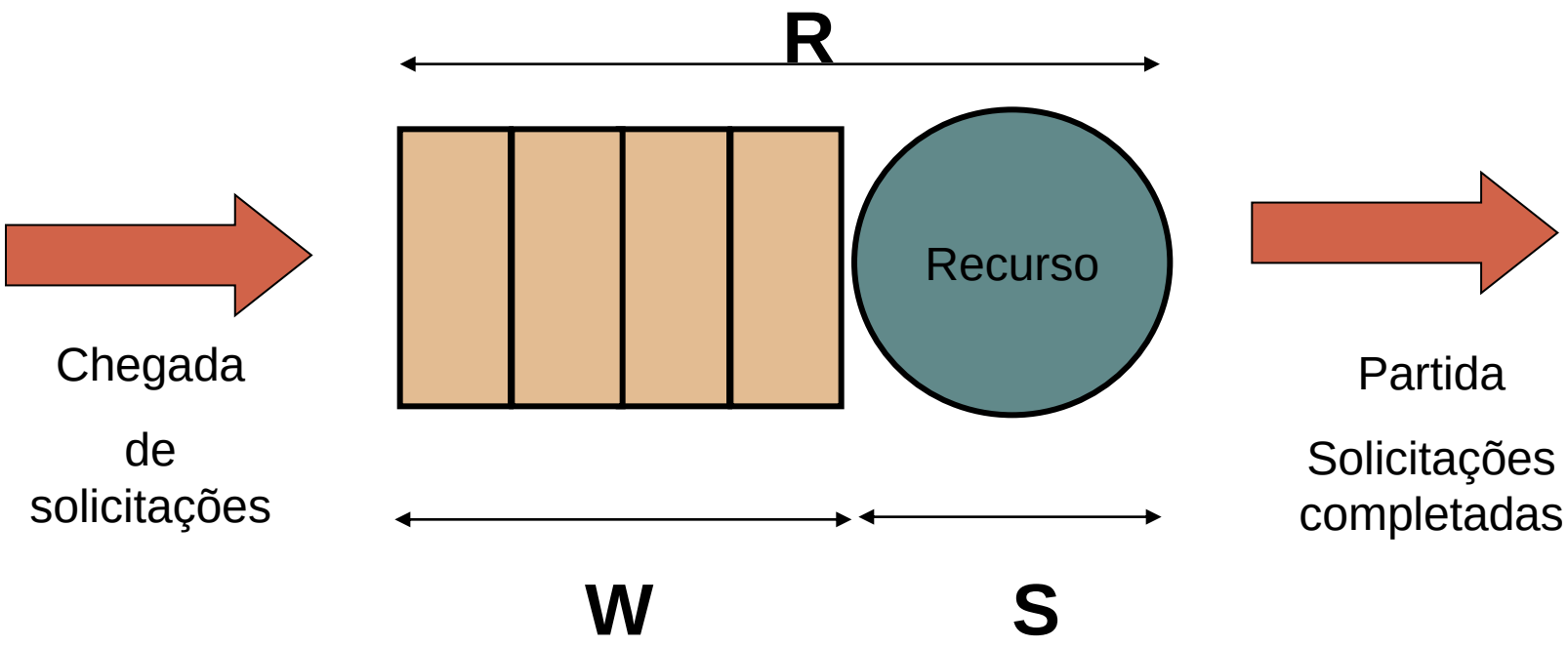
MODELAGEM


FILA
CPU
Disco
Segmento
de LAN,...
Recurso
ou
Servidor
Chegada
de
solicitações
Linha ou fila de
espera
Partida
Solicitações
completadas

MODELAGEM


FILA
1
m
Múltiplo recursos (multiprocessador, Estação Rádio Base, Call Center, etc...)
Chegada
de
solicitações
Linha ou fila de espera
Partida
Solicitações completadas

Modelagem



R-> Tempo de Reposta por visita ao recurso
W->Tempo de espera por visita ao recurso
S->Tempo de serviço por visita ao recurso
R=W+S

Modelagem

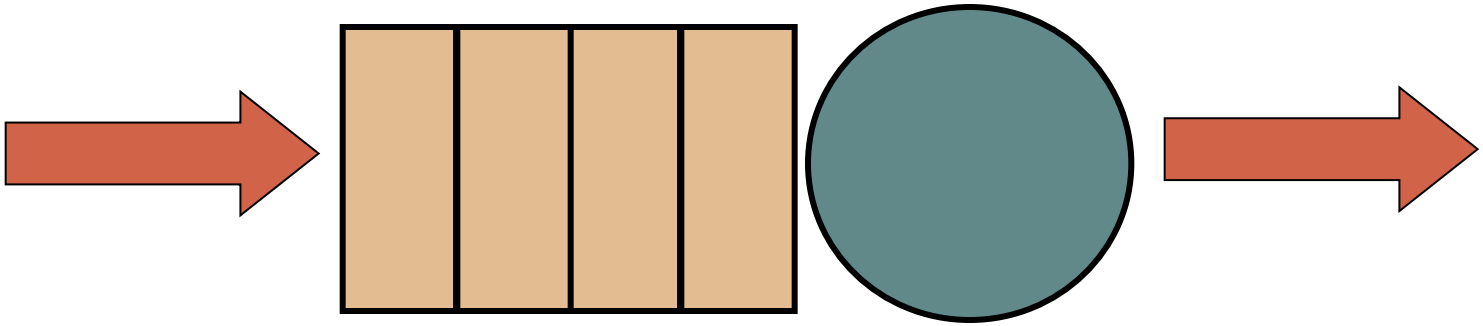

Questões a serem respondidas pelo modelo:
Qual o tempo de reposta (tempo de atendimento + tempo de fila) para um cliente?
Qual o tempo médio de atendimento?
Qual o tempo médio de fila?
Qual o número médio de clientes?
Qual a utilização dos recurso?

Tipos de recursos

Independente da carga (IC):

Centros de serviço cuja a taxa de serviço é constante (não depende da carga)

Ex.: CPU, disco

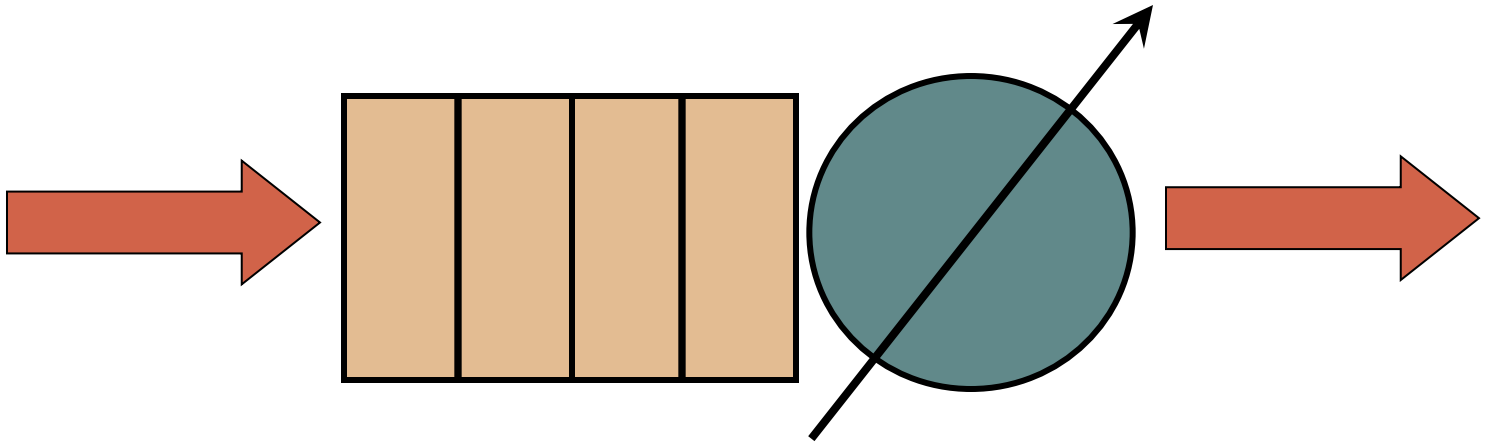


Tipos de recursos

Dependente da carga (DC):

Centros de serviço cuja a taxa de serviço é dependente do número de clientes na fila

Ex.: LAN

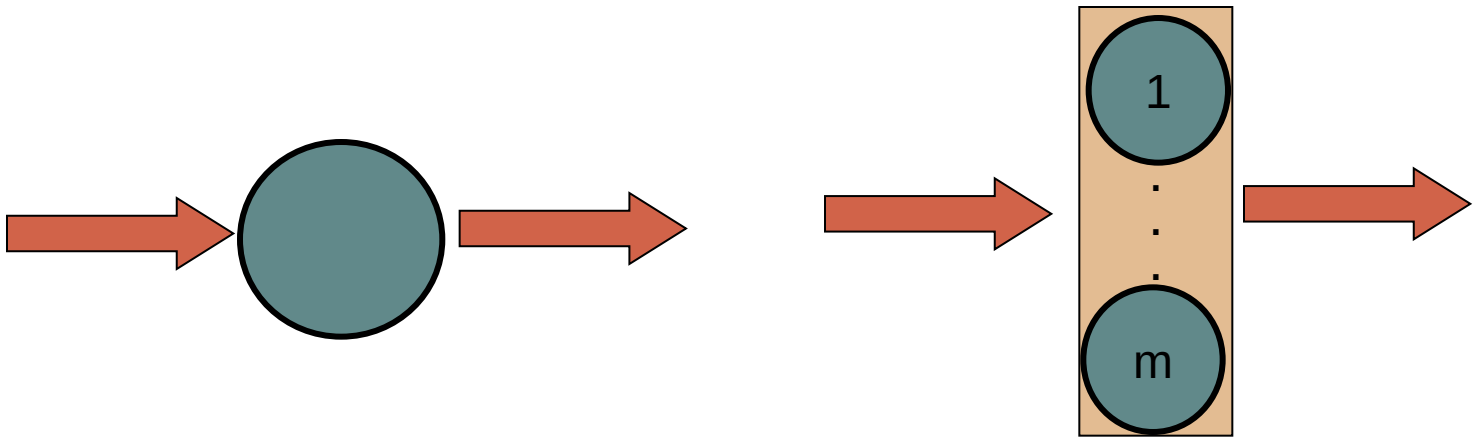


Tipos de recursos

Centro de atraso (A):

Não possuem fila, uma solicitação que chega é imediatamente atendida

Ex.: Recursos dedicados ou quando há mais recursos que solicitações



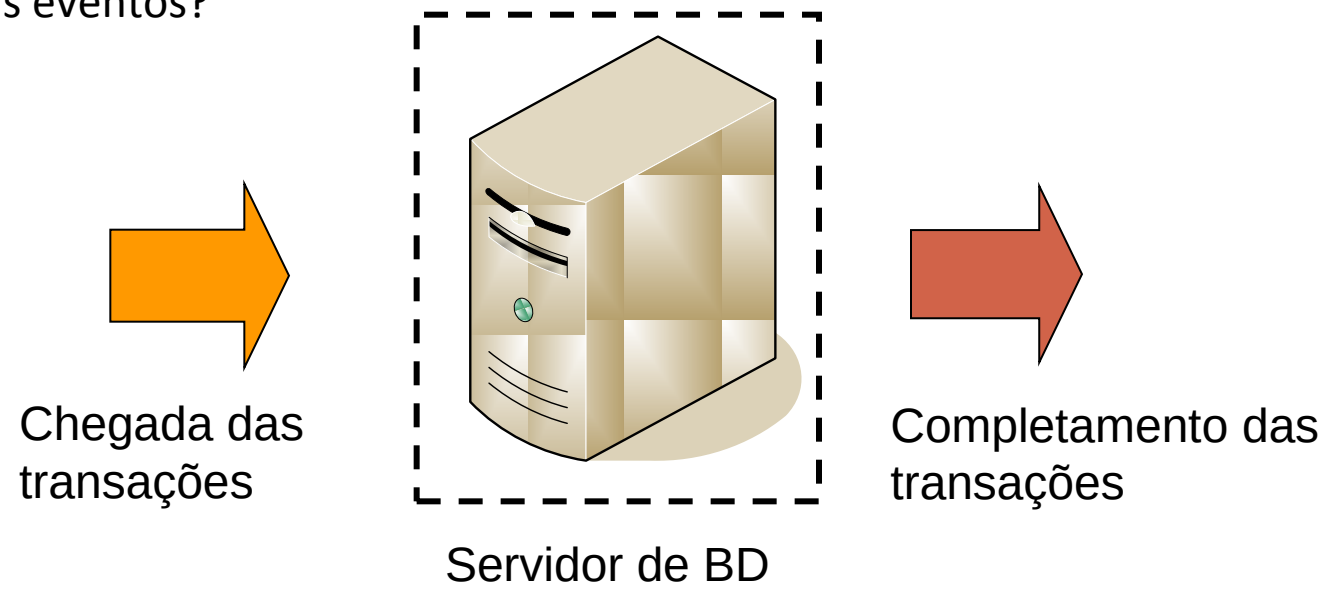
Modelando Sistemas com vários recursos

- Redes de filas
  - Coleção de filas (centros de serviço)
  - Permite uma avaliação analítica
  - Ex.: Servidor de banco de dados

MODELAGEM

Servidor de banco de dados

- Como modelar?
- Questões a serem respondidas
  - Nível de abstração?
  - Quais o elementos envolvidos (escopo) ?
  - Quais eventos?

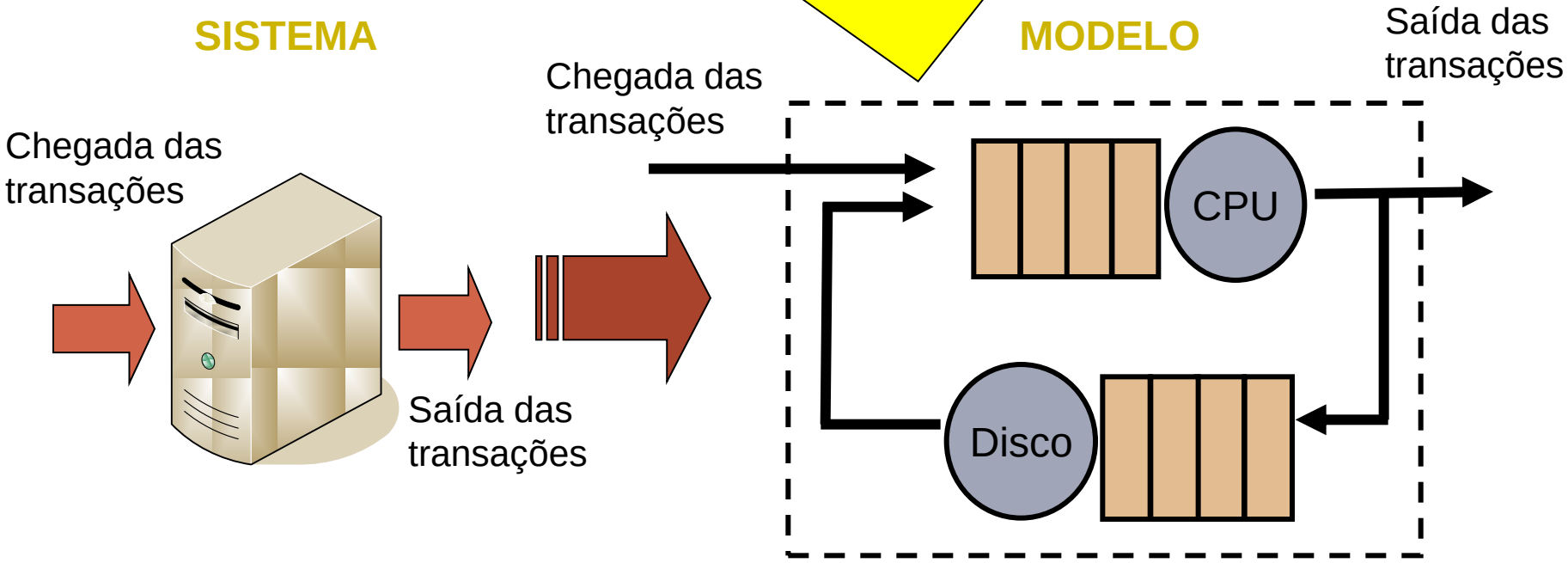


MODELAGEM

Servidor de banco de dados

Disco e CPU com um recursos

- Quem é o gargalo?
- O que aconteceria se a CPU fosse trocada por outra 2 vezes mais rápida?



Modelagem

Suponha que uma investigação do log do sistema de gestão de dados revela que as operações individuais apresentadas para o servidor de banco de dados com características significativamente diferentes.

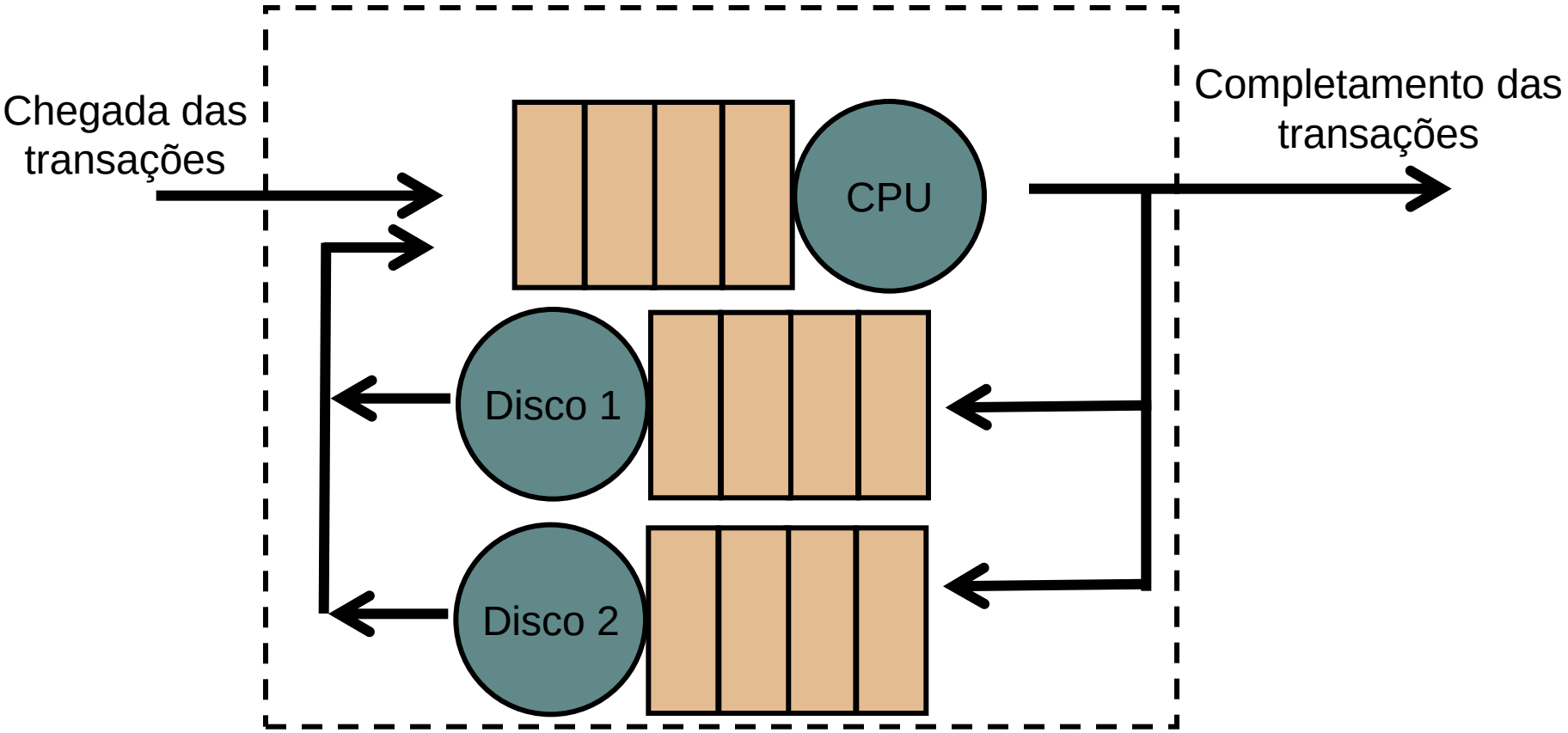
Supor que o analista observa que essas operações podem ser agrupadas em três grupos distintos de operações bastante similar : trivial, médio e complexo.

Se eles foram combinados em um único grupo o modelo resultante pode ser muito aproximativa e com grandes erros.

Assim, ao descrever um modelo de QN, tem de especificar as classes de clientes que usam os recursos da QN, a intensidade da carga de trabalho de cada classe, e as demandas de serviço em cada um dos recursos por grupo.

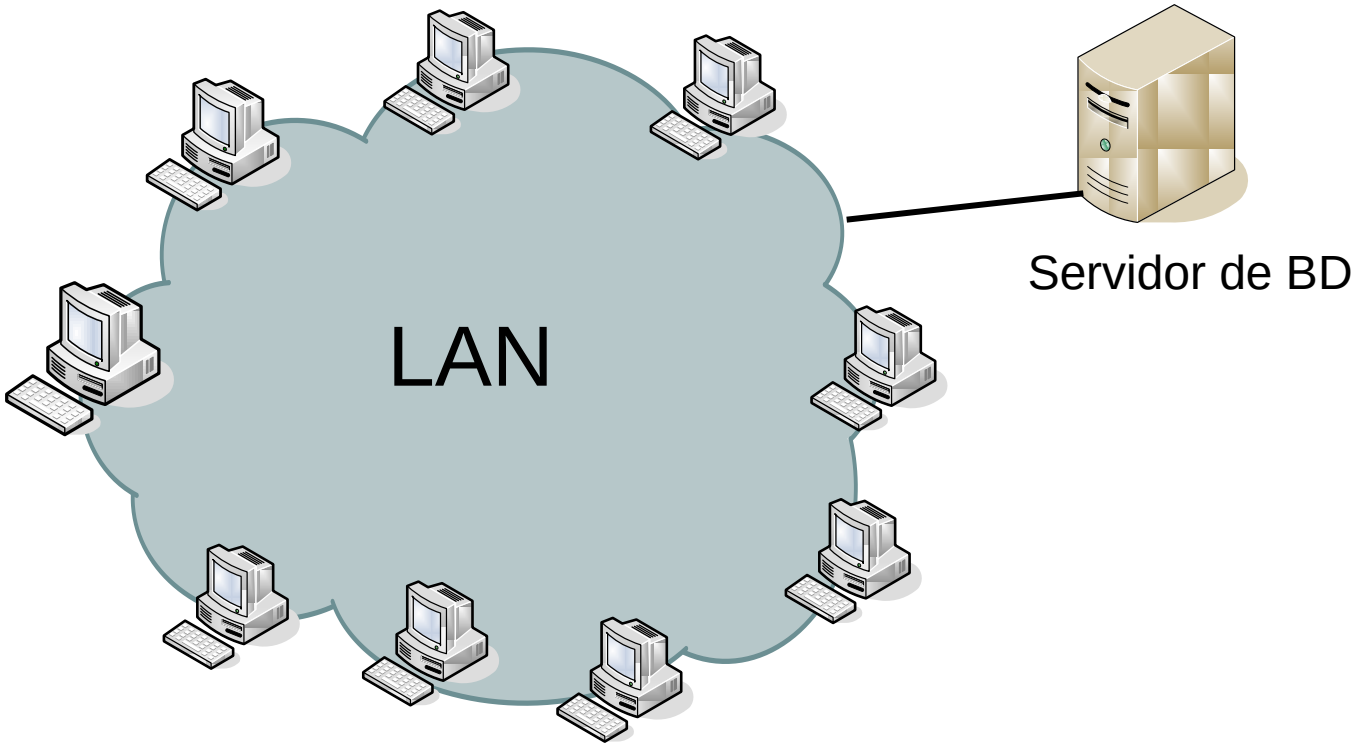
Transações Grupo	Porcentagem do Total	Média de tempo de CPU (s)	Avg. Número de I / Os
Trivial	45%	0,04	5,5
Médio	25%	0,18	28,9
Complexo	30%	1,20	85,0

Servidor de banco de dados com dois discos



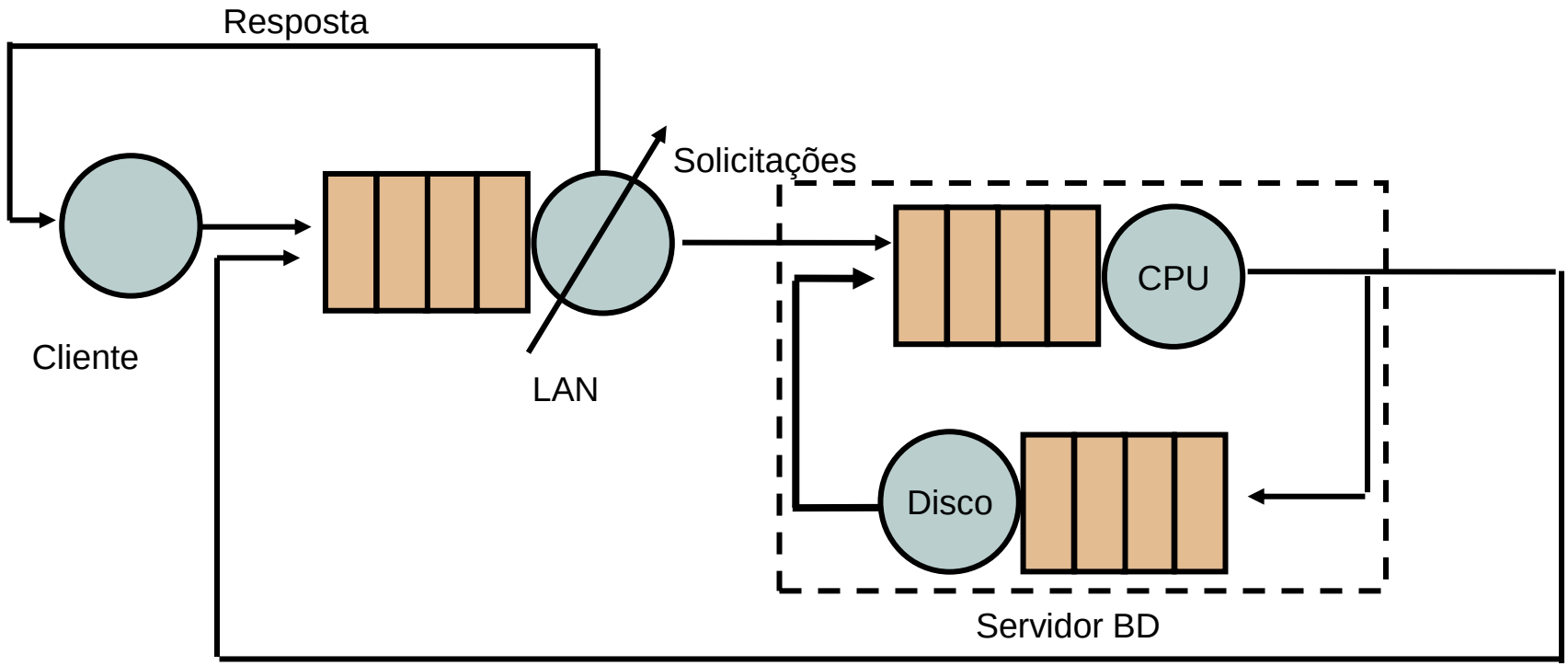
MODELAGEM

Servidor de banco de dados com clientes e LAN



MODELAGEM

Servidor de banco de dados com clientes e LAN

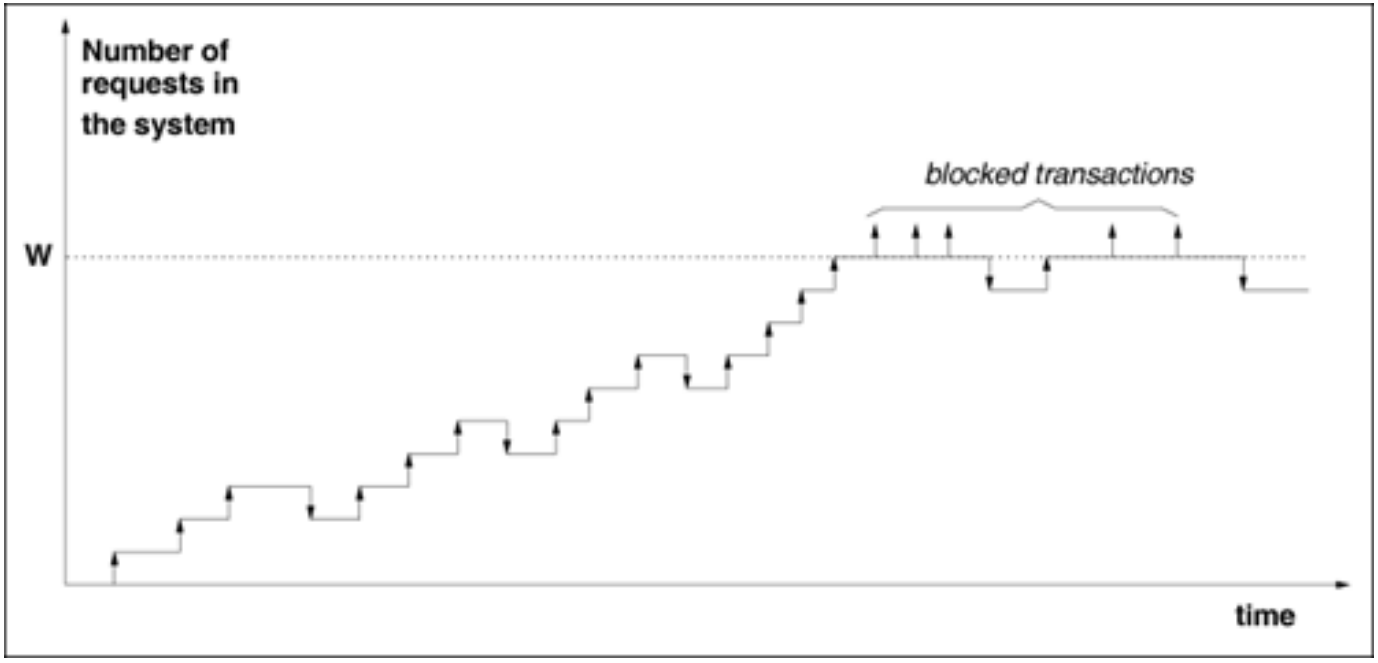


MODELAGEM

Servidor de banco de Dados - bloqueio

Manter garantia do tempo de resposta aos seus clientes.

A fim de fornecer essa garantia, independentemente da taxa de chegada dos pedidos, o número de operações simultâneas de dados tem de ser limitada.



MODELAGEM

Servidor de banco de Dados - bloqueio