

# FILAS

## *Conceitos Fundamentais*

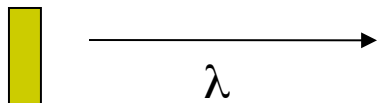
Pós Graduação em Engenharia Elétrica - PPGEE

Prof. Carlos Marcelo Pedroso

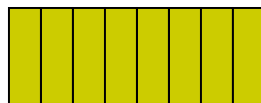
# Teoria de Filas

- É uma das abordagens mais utilizadas no estudo de desempenho e dimensionamento de redes;

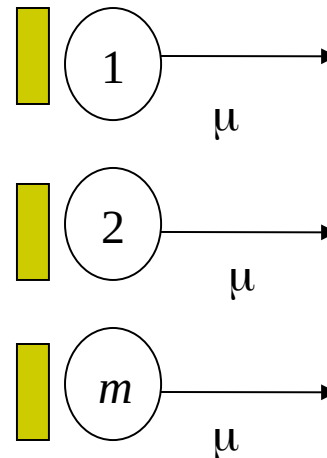
**Chegada**



**Fila**



**Serviço**



# Teoria de Filas

- Notação básica:
  - $\lambda \Rightarrow$  taxa média de chegada
    - $\lambda = 1/IC$ , IC é o intervalo médio entre pacotes
  - $\mu \Rightarrow$  taxa média de atendimento
    - $\mu = 1/TA$ , TA é o tempo médio de atendimento
  - TF, TS, TA: Tempo médio na fila, tempo médio no sistema e tempo médio de atendimento
    - $TS = TF + TA$

# Lei de Little

>> Veja anotações da aula

Foi observado o comportamento de um certo servidor de banco de dados.

Durante o período de 1 minuto de observação, o sistema ficou 10% do tempo livre. A taxa média de chegada foi de 100 requisições por segundo. Considere que o tempo de resposta aparente para o usuário foi de 0,1 segundo em média.

Qual o tempo gasto na fila e no processamento da requisição? Qual o número médio de elementos na fila?

# Exemplo

- Suponha um provedor de acessos a Internet que recebe uma média de 20 requisições/hora. O tempo médio de conexão é de 10 minutos. Qual a quantidade de modems para obter um desempenho adequado.
  - Um tempo adequado é o tempo limite que um cliente espera pela conexão para fazer um pagamento com um cartão, por exemplo.

# Exemplo

- Suponha um provedor de acessos a Internet que recebe uma média de 20 requisições/hora. O tempo médio de conexão é de 10 minutos. Qual a quantidade mínima de modems para atender este sistema?

# Teoria de Filas - Notação

- Para o estudo de filas, convencionou-se a seguinte notação:  $A/S/m/B/K/SD$ 
  - A indica a distribuição de probabilidade do intervalo entre chegadas;
  - S indica a distribuição de probabilidade do tempo de atendimento;
  - M é o número de servidores;
  - B é a capacidade do buffer;
  - K é o tamanho da população;
  - SD é a disciplina da fila (ex. Fifo).

# Teoria de Filas - Notação

- As distribuições são indicadas por:
  - M – Exponencial
  - E – Erlang
  - H – Hyperexponencial
  - D – Determinística
  - G – Geral

# M/M/1 com população infinita

$$NF = \frac{\lambda^2}{\mu(\mu - \lambda)}$$

$$TF = \frac{\lambda}{\mu(\mu - \lambda)}$$

$$NS = \frac{\lambda}{\mu - \lambda}$$

$$TS = \frac{1}{\mu - \lambda}$$

$$P_n = \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^n$$

*$P_n$  é a probabilidade de existirem  $n$  clientes no sistema*

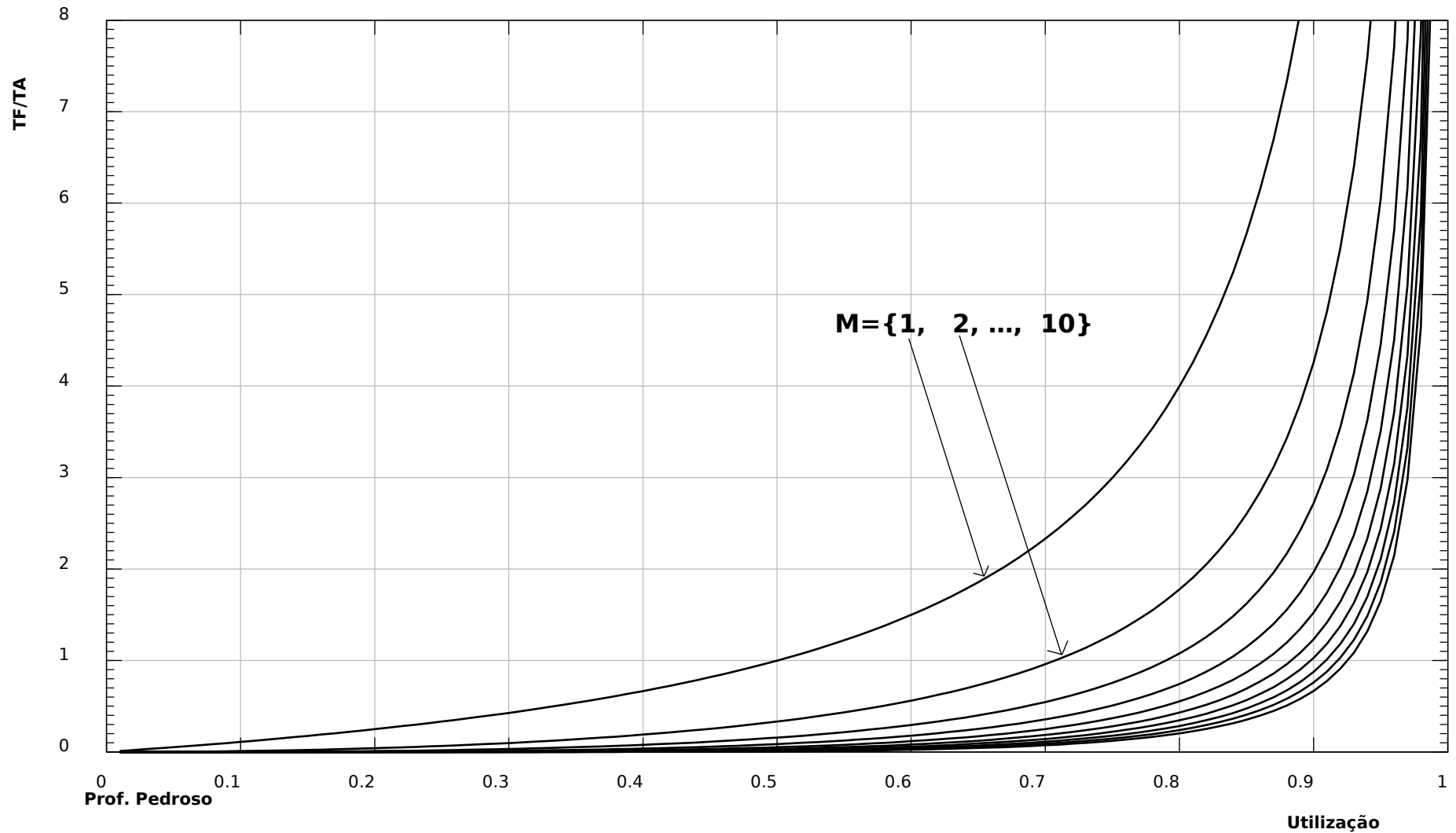
# M/M/c com população infinita

- As fórmulas para o modelo M/M/c são relativamente complexas;

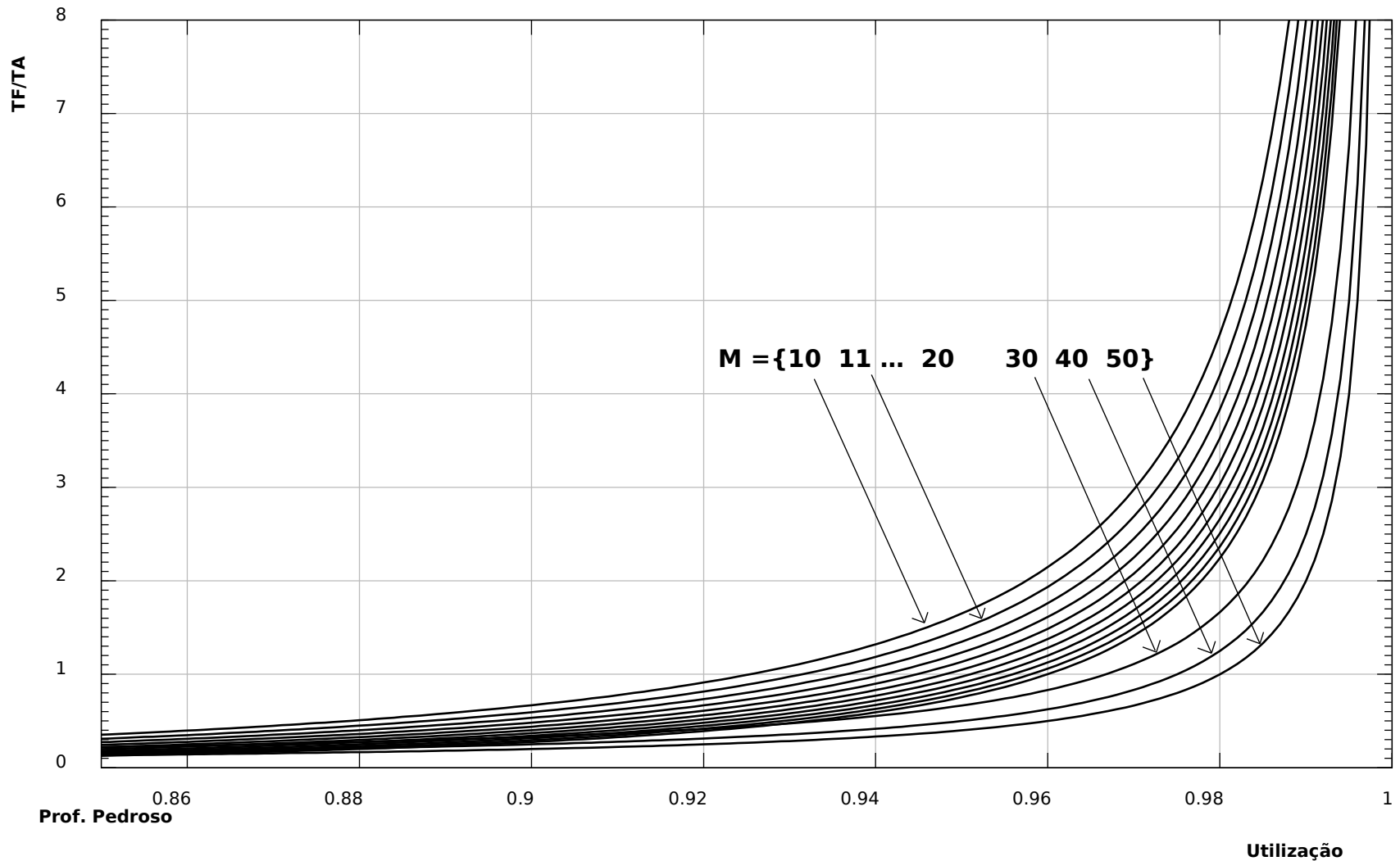
$$TF = TA \cdot \frac{1 - A}{(1 - \rho)(1 - \rho A)}, \quad A = \frac{\sum_{i=0}^{M-1} \frac{(M\rho)^i}{i!}}{\sum_{i=0}^{\infty} \frac{(M\rho)^i}{i!}}$$

- Normalmente, utiliza-se a abordagem gráfica para resolução de problemas.

# M/M/c com população infinita



# M/M/c com população infinita



# Exemplo

1. Suponha um provedor de acessos a Internet que recebe uma média de 20 requisições/hora. O tempo médio de conexão é de 10 minutos. Qual a quantidade de modems necessária para oferecer uma boa qualidade aos usuários?
2. Um sistema com uma base de dados consiste de 3 discos rígidos compartilhando uma fila única. O tempo de serviço para uma requisição de E/S é de 50m segundos. As requisições de E/S chegam ao sistema a uma taxa de 30 requisições por segundo. Utilizando o modelo M/M/3, determine o seguinte:
  - A) utilização média dos discos rígidos;
  - B) probabilidade do sistema esta ocioso;
  - C) número médio de requisições de acesso no sistema e o número médio de requisições esperando na fila;
  - D) O tempo médio de resposta.

# Exemplo

3. Considere um sistema com uma fila onde os clientes chegam de acordo com o processo de Poisson com uma taxa de 25 clientes por hora. Suponha que o tempo de atendimento possui distribuição exponencial.

Existem duas opções possíveis para o projeto do sistema. A primeira utiliza dois servidores, cada um atendendo clientes com um tempo médio de atendimento de 4 minutos. A segunda utiliza um único servidor com um tempo médio de atendimento 2 minutos.

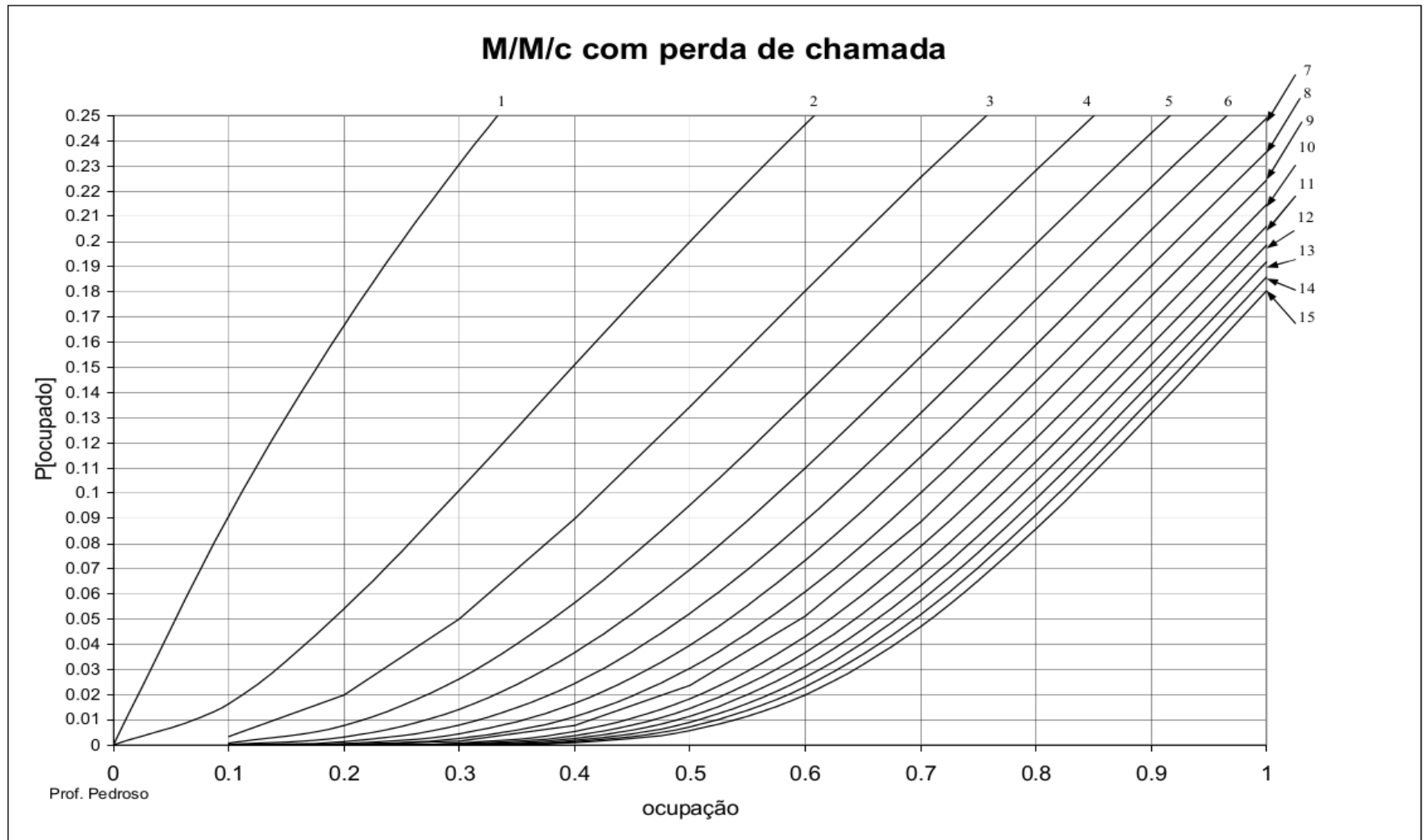
Calcule qual o tempo médio de espera na fila em cada um dos casos.

# M/M/c com perda de chamada

- Para o sistema M/M/c com perda de chamada temos:

$$P[\textit{ocupado}] = \frac{(M\rho)^M / M!}{\sum_{i=0}^M \frac{(M\rho)^i}{i!}}$$

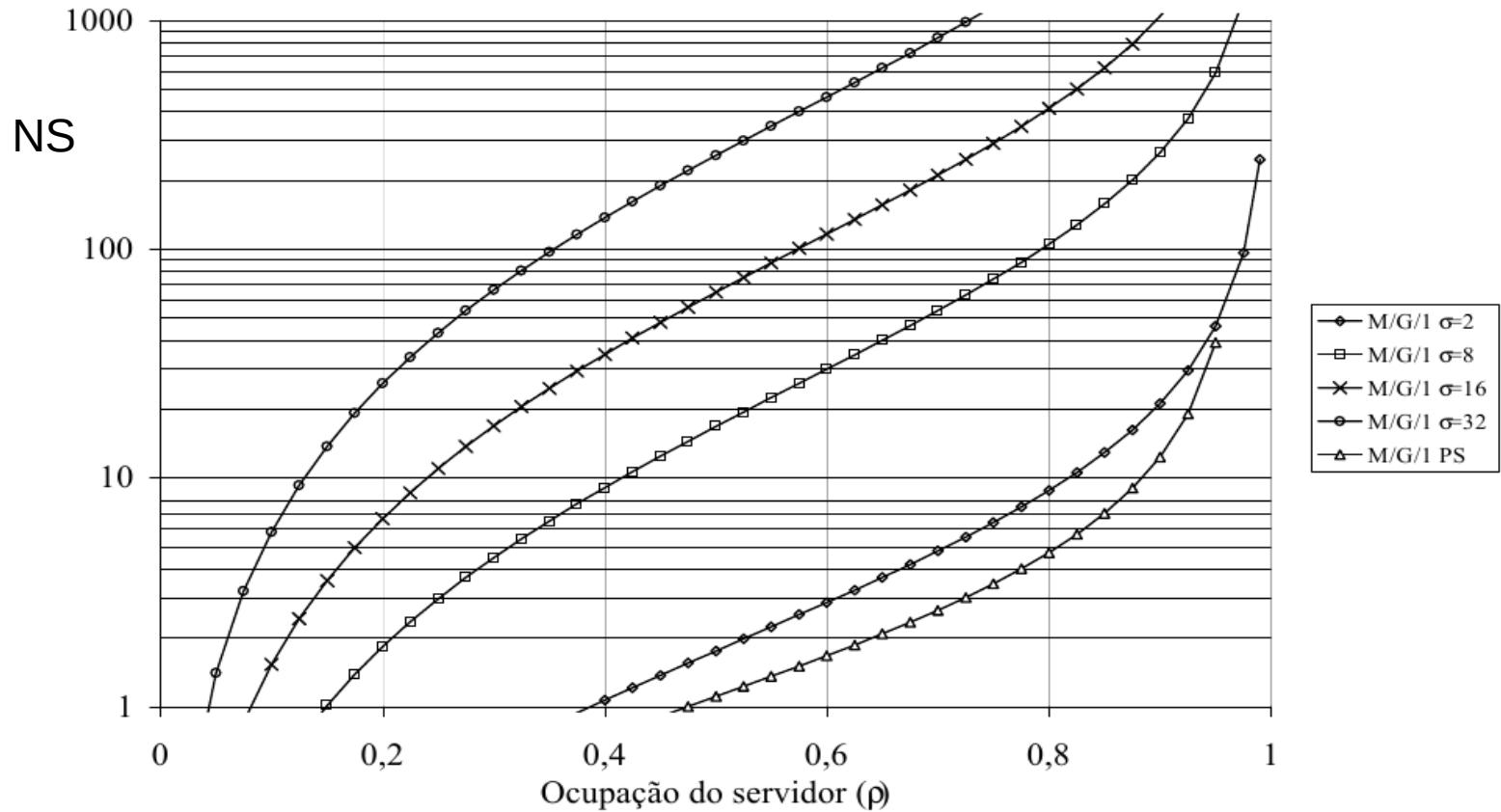
# M/M/c com perda



# Exemplo

- Suponha um provedor de acessos a Internet que recebe uma média de 20 requisições/hora. O tempo médio de conexão é de 10 minutos. Qual a quantidade de modems para obter um desempenho adequado.
  - Um tempo adequado é o tempo limite que um cliente espera pela conexão para fazer um pagamento com um cartão, por exemplo.

# M/G/1



$$NF = \frac{\rho^2}{2 \cdot (1 - \rho)} \cdot \left[ 1 + \frac{\sigma^2}{TA^2} \right], \quad \rho < 1$$

# Exemplo

- Uma empresa irá comprar uma máquina que faz a verificação da qualidade final de um produto através de processamento de imagem, e está comparando o desempenho de duas máquinas, chamadas de A e B. A máquina A é mais rápida, e é capaz de processar em média um item a cada 24 segundos, com desvio padrão de 20 segundos. A segunda máquina é mais lenta com um tempo de processamento médio de 25 segundos, mas com desvio padrão de 2 segundos. Se em média são enviados 2 produtos por minuto para verificação, determine qual máquina possui o melhor desempenho.

# Exemplo

- Uma oficina mecânica possui um almoxarifado que entrega ferramentas para os mecânicos. Considere que diversos atendentes são disponíveis, que a população de mecânicos é grande e que os tempos de atendimento e intervalos de chegada podem ser modelados com uma distribuição exponencial. A taxa de chegada é de 2 mecânicos por minuto e o tempo de serviço para atender uma solicitação por ferramenta é de 150 segundos em média.
  - Qual o número mínimo de atendentes?
  - Qual o tempo médio de espera do mecânico considerando o número médio de atendentes? Qual o tamanho médio da fila neste caso?
  - Qual o número de atendentes para que o tempo médio de espera do mecânico não seja maior que um minuto? Qual o tamanho da fila neste caso?

# M/G/c

- Não existe equacionamento conhecido para o sistema M/G/c para  $1 < c < \infty$ .
- Uma aproximação do modelo M/G/c pode ser obtido a partir das fórmulas do modelo M/M/c com a aplicação do fator de correção  $(1 + (\sigma/TA)^2)/2$ . A aproximação será melhor quando  $\sigma/TA$  estiver próximo de 1.

# Exemplo

- Suponha que no problema da oficina mecânica o tempo de atendimento não possa ser modelado por uma distribuição exponencial. Para este caso, determine o tempo médio de espera (aproximado) se o desvio padrão do tempo de serviço for dado por 50, 300 e 500 segundos.

# Outros modelos

A literatura apresenta uma série de modelos prontos:

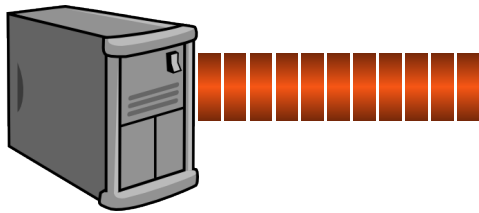
- Consultar:
  - “The Art of Computer Systems Performance Analysis”, Raj Jain.
  - “Discrete Event System Simulation”, Jerry Banks.

# Disciplina de Filas

- FIFO (First In First Out)
- PS (Process Sharing)
- PQ (Priority Queue)
- SFQ (Stochastic Fair Queue)
- RED (Rand Early Detection)
- Entre muitas outras ...

# DropTail

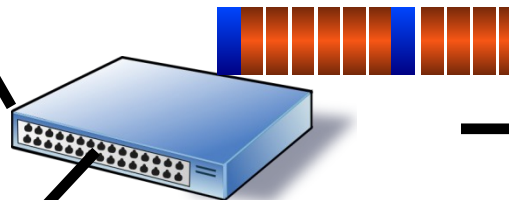
Transmissor 1



Perdas ocorrem devido à falta de espaço no buffer

- Padrão de fato

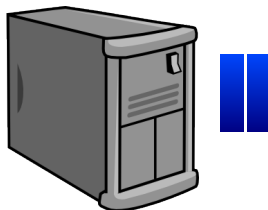
Receptor



+ Baixa complexidade de implementação

- Grande atraso
- Fluxos sincronizados
- **Bufferbloat:** atraso excessivo devido à grande capacidade do buffer causa problemas para muitas aplicações

Transmissor 2



# Random Early Detection (RED)

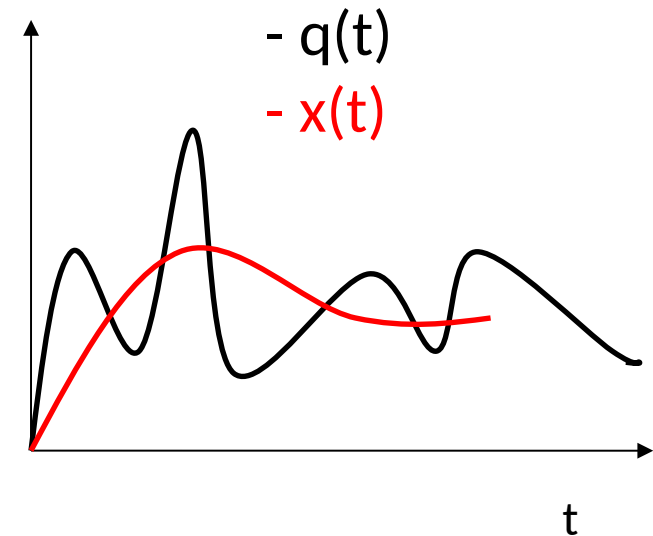
Proposto em 1993

Descarta pacotes (com uma probabilidade de descarte) para prevenir situações de congestionamento reagindo antecipadamente.

Remove a sincronização entre fluxos.

# RED

Drop based on *average* queue length  $x(t)$   
(EWMA algorithm used for averaging)



$$\text{Every } T: x(t) = (1 - w_q) * x(t - T) + w_q * q(t)$$

$x(t)$ : smoothed, time averaged  $q(t)$

# RED

## Muitos parâmetros

- $\min_{th}$ ,  $\max_{th}$ ,  $w_q$  (constante da EWMA)

## Parâmetros afetam desempenho

- Ajuste é difícil, vários métodos automáticos disponíveis
- Erros de ajuste podem levar a um desempenho inferior ao droptail

## Normalmente disponível em roteadores

# Escalonamento de pacotes

Selecionar qual a próxima fila a ser servida

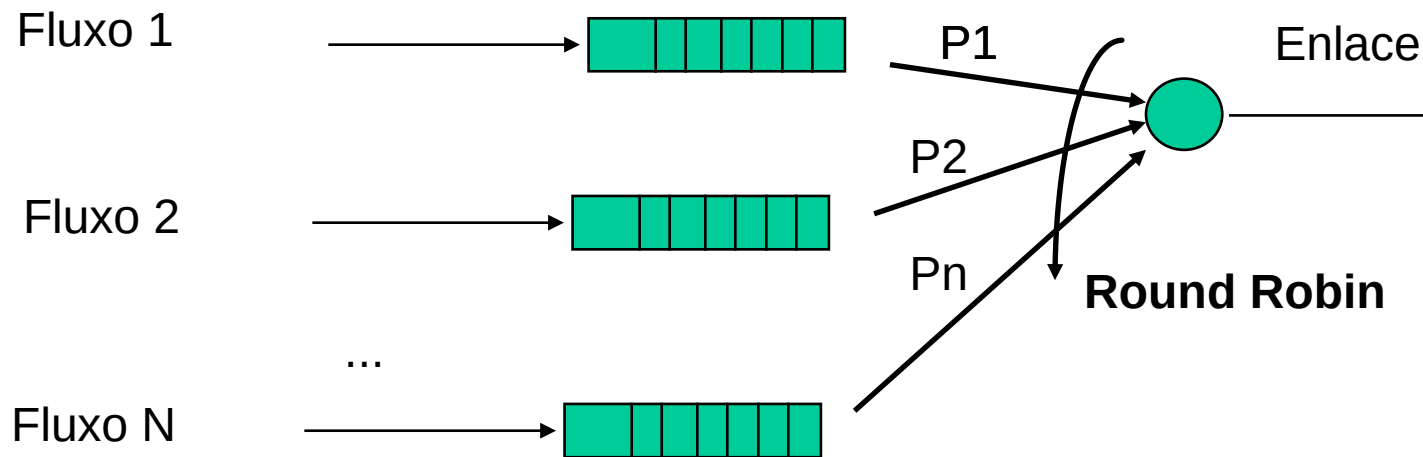
Cada fila possui uma disciplina específica

Principais métodos: GPS (Generalized Process Sharing), WRR (Weighted Round Robin), WFQ (Weighted Fair Queue).

# GPS

- Supõe que podem ser enviados pacotes divisíveis até o nível de bit.
- Usa Round Robin entre as diversas filas.
- Política de divisão ideal para distribuição de recursos.
- **GPS não é implementável!**
- Utilizado principalmente na comparação com as políticas implementáveis.

# WRR



*Anote explicação e exemplo durante a aula!*

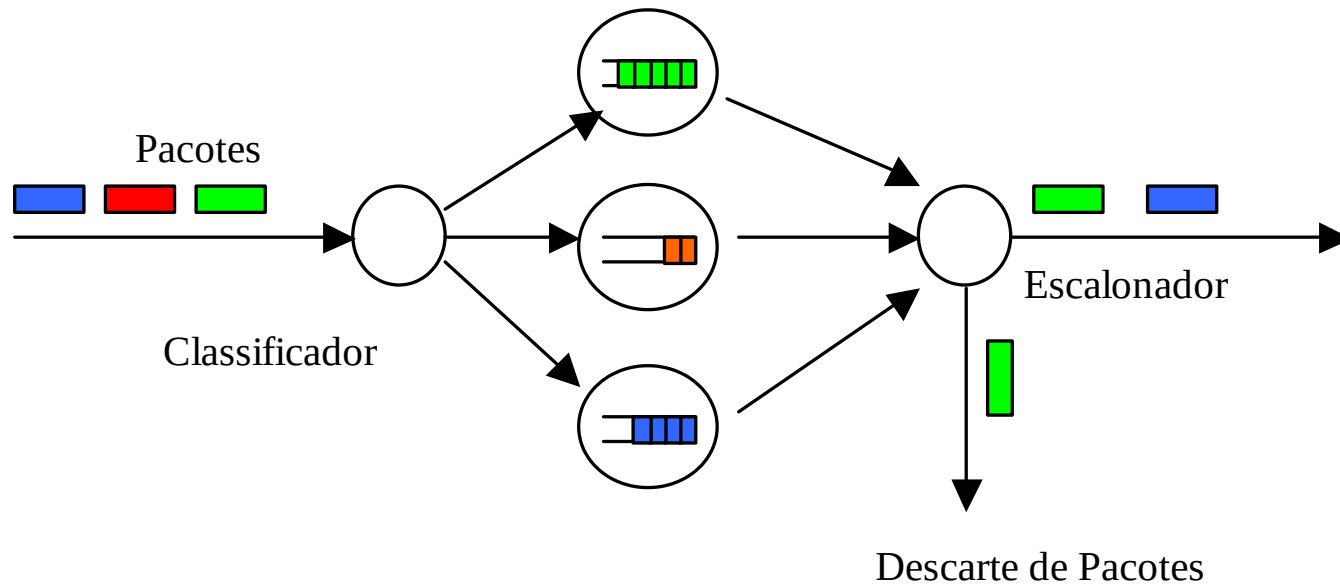
# WFQ - Weighted Fair Queuing

- Divisão justa da banda passante.
- Transmite os pacotes na ordem de término, do menor para o maior.
- Pesos determinam qual pacote será transmitido primeiro.
- Atraso pode ser calculado deterministicamente.

*Anote explicação e exemplo durante a aula!*

# QoS no IP

## Modelagem de um roteador



- 
- Controle de Admissão
  - Controle de Reservas (protocolo RSVP)
  - Roteamento

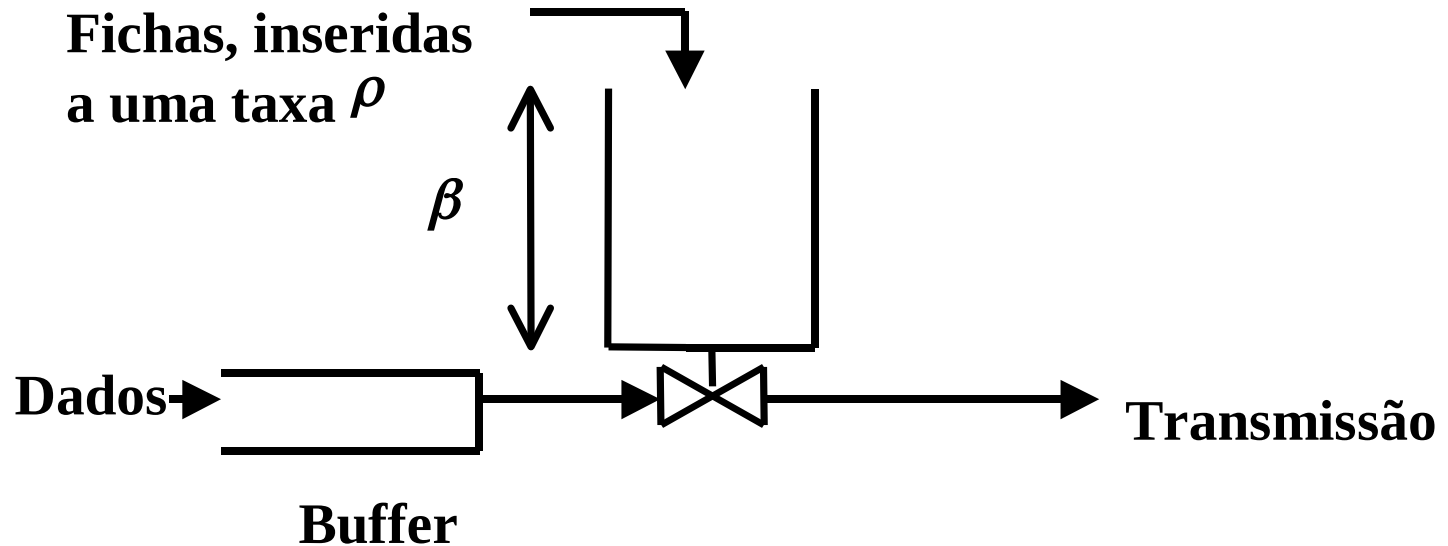
# QoS no IP

## Modelos:

- Serviço Integrado
  - Melhor Esforço
  - Carga Controlada
  - Serviço Garantido
- Serviço Diferenciado

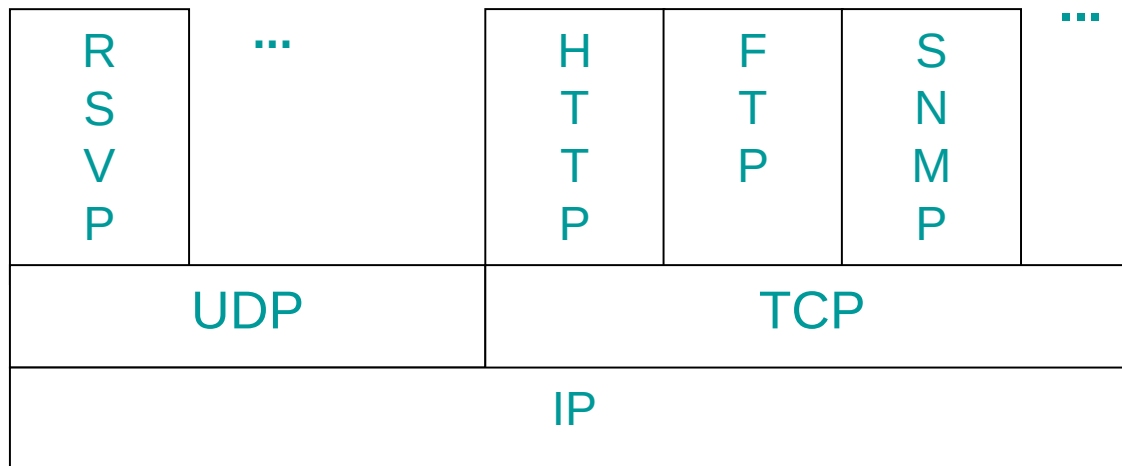
# QoS no IP

Caracterização de tráfego:  
algoritmo do balde de fichas.

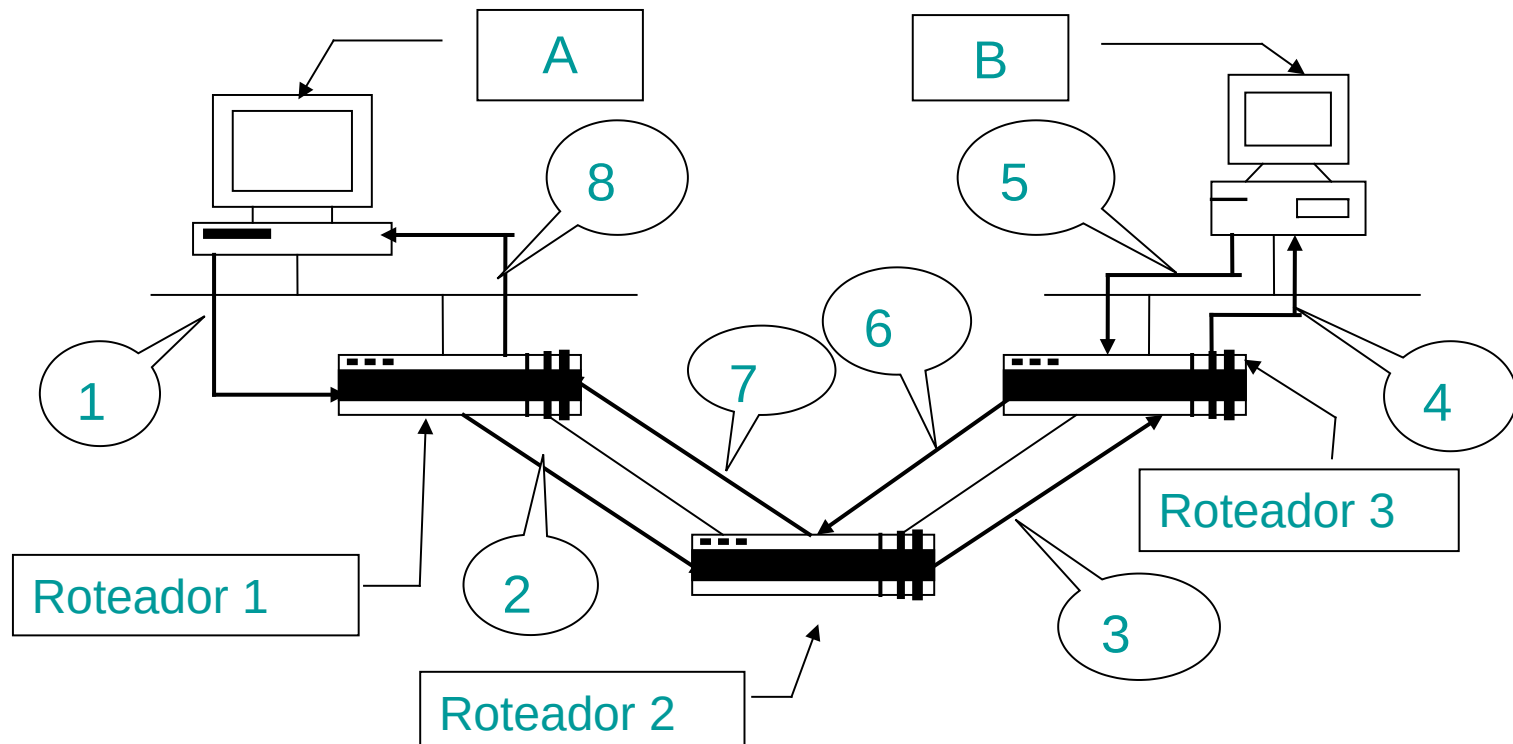


# RSVP - Resource Reservation Protocol

Protocolo de Reserva de Recursos sobre o TCP/IP



# *RSVP - Estabelecimento da Reserva*



# RSVP - Estilos de Reserva

Estilo Filtro Fixo

- $FF(S\{Q\})$
- $FF(S1\{Q1\}, S2\{Q2\}, \dots)$

Estilo Filtro Explícito Compartilhado

- $EC( (S1, S2, \dots)\{Q\} )$

Estilo Filtro de Curinga

- $FC(*\{Q\}),$

# QoS no IP: Carga Controlada

Modalidade do Serviço Integrado

Distribuição estatística da banda passante

— Parâmetros:

- $r$  : Taxa de inserção de fichas no balde
- $b$  : Tamanho do balde de fichas
- $p$  : Taxa de pico
- $m$  : Tamanho mínimo da unidade policiada
- $M$  : Tamanho máximo do pacote

# QoS no IP: Serviço Garantido

## Modalidade do Serviço Integrado

- Garante o atraso máximo de propagação
- Taxa garantida:  $R$
- Tamanho do balde:  $B$
- Quantidade de dados transmitidos não maior que  $[M + \min(p.T, r.T + b - M)]$
- Atraso deve ser menor que  $[b/R + C/R + D]$

# QoS no IP: Serviço Diferenciado

- Problema do Serviço Integrado. Complexidade devido a:
  - Grande número de filas.
  - Classificação do tráfego exige muitas vezes o processamento de nível 4 de protocolo.
- Estes problemas são endereçados pelo serviço diferenciado.

*Anote explicações e exemplo durante a aula*

# Algoritmos de Escalonamento de Pacotes

- Escolher o próximo pacote a ser transmitido
- Realiza o policiamento de tráfego no elemento de rede
- Algoritmo original do IP: FIFOQ
- Algoritmos existentes: SFQ, WFQ, CBQ, RED, etc.

# CBQ - Class Based Queuing

Divisão da banda em classes

Estimador de tráfego

