FILAS

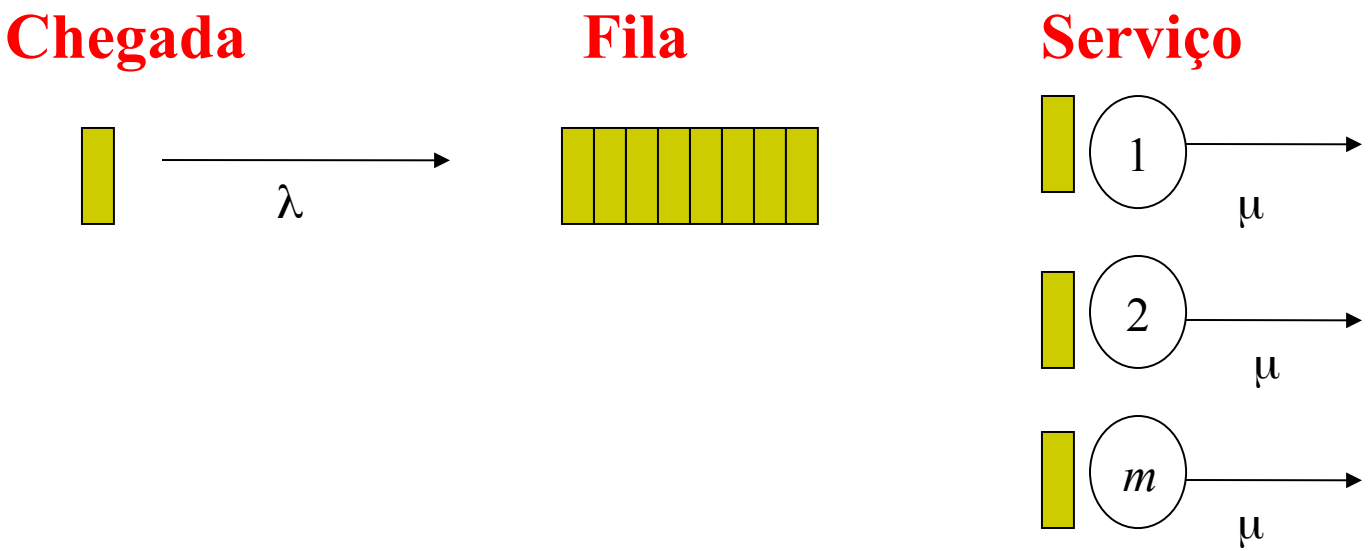
Conceitos Fundamentais

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;



Teoria de Filas

- Notação básica:
  - $\lambda \Rightarrow$ taxa de chegada de requisições
    - $\lambda = 1/IC$, IC é o intervalo médio entre requisições
  - $\mu \Rightarrow$ taxa de atendimento
    - $\mu = 1/T_a$, T_a é o tempo de atendimento
  - T_f, T_s, T_a: Tempo na fila, tempo no sistema e tempo de atendimento
    - $T_s = T_f + T_a$

Dinâmica da fila

- Exemplo. Suponha um sistema de pagamento de pedágio que recebe no momento de maior movimento 1000 carros/hora. O tempo médio de atendimento é de 30 segundos. Qual o número mínimo de guichês?
  - *Se o número for menor, o que deve ocorrer?*

Dinâmica da fila

- *Suponha o seguinte sistema com um servidor:*
  - *Determine:* λ, μ, ρ, T_f, T_s, T_a, N_f, N_s, N_a

Requisição	IC (s)	T_a (s)
1	2	4
2	1	3
3	3	2
4	1	1
5	3	1
6	2	2
7	3	1
8	1	1

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?

Previsões de Desempenho

- Suponha um sistema de pagamento de pedágio que recebe no momento de maior movimento 1000 carros/hora. O tempo médio de atendimento é de 30 segundos. Suponha que o sistema usa 10 guichês.
  - Qual o tamanho médio da fila?
  - Qual o tempo médio de espera?

  Existe alguma maneira de descobrir?

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

$$E[Nf] = \frac{\lambda^2}{\mu(\mu - \lambda)} \qquad E[Tf] = \frac{\lambda}{\mu(\mu - \lambda)}$$

$$E[Ns] = \frac{\lambda}{\mu - \lambda} \qquad E[Ts] = \frac{1}{\mu - \lambda}$$

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

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

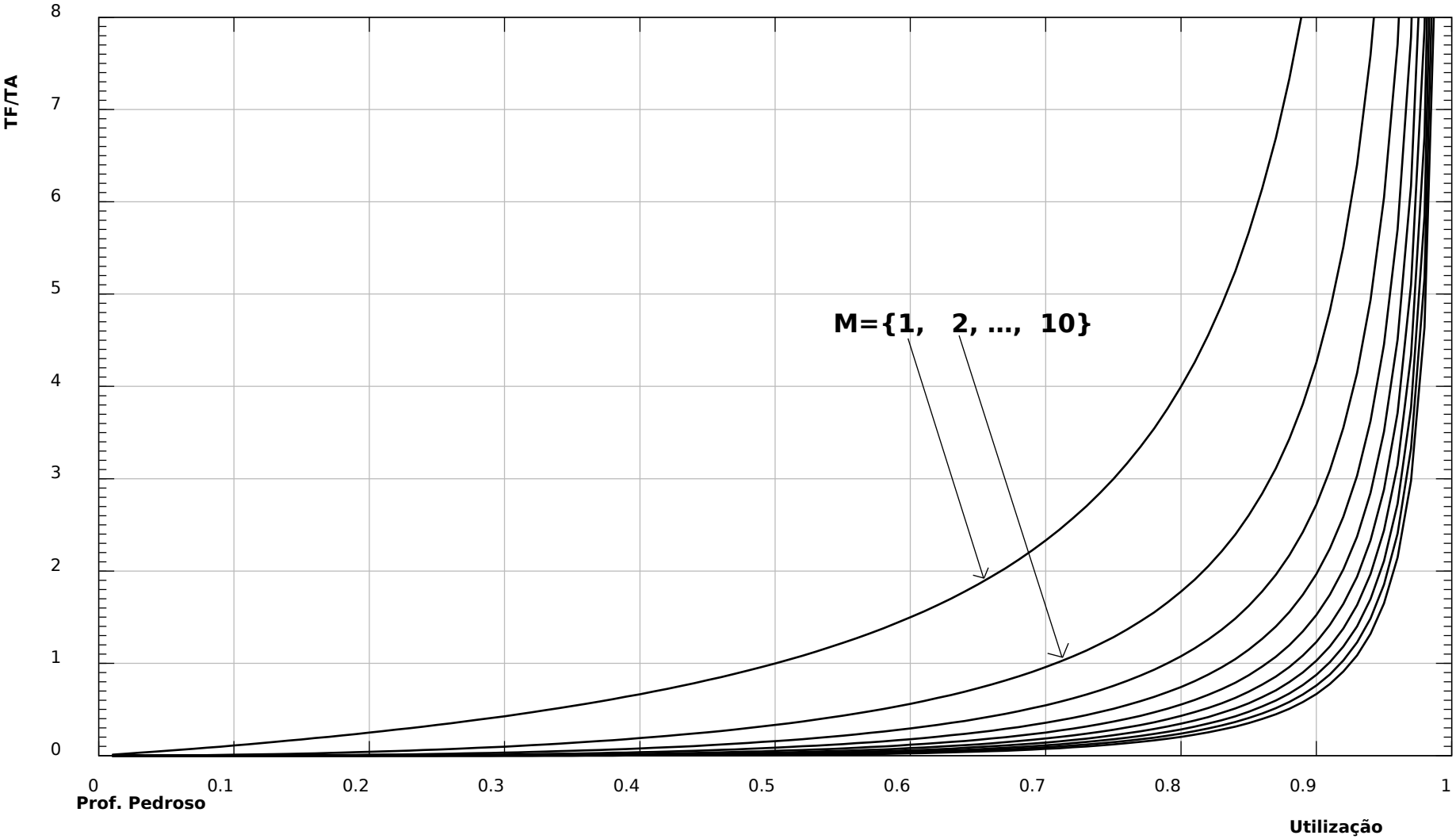
M/M/c com população infinita

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

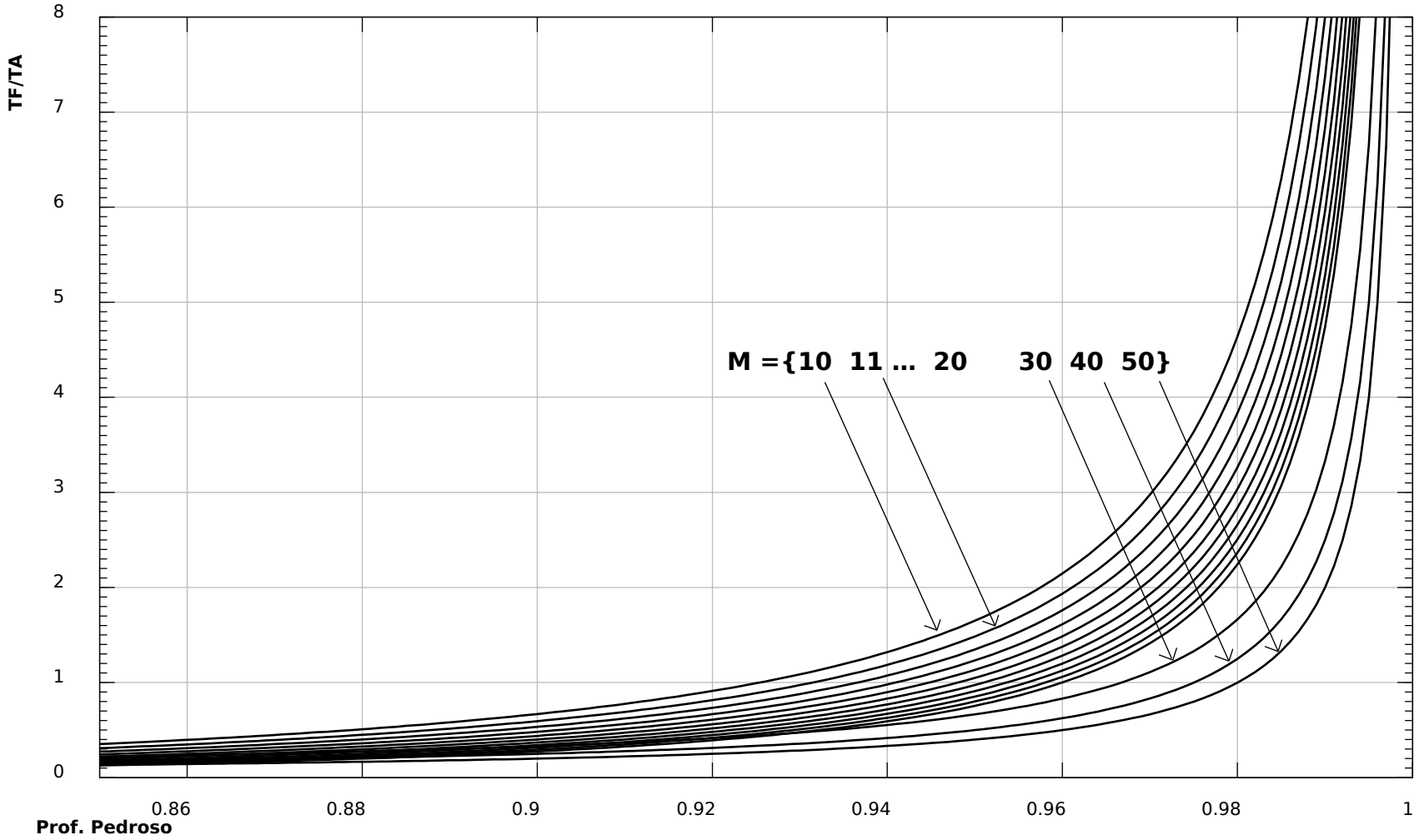
$$E[Tf] = E[Ta]\,\frac{1-A}{(1-\rho)(1-\rho A)},\quad A = \frac{\sum_{i=0}^{M-1}\frac{(M\rho)^i}{i!}}{\sum_{i=0}^{M}\frac{(M\rho)^i}{i!}}$$

- 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.

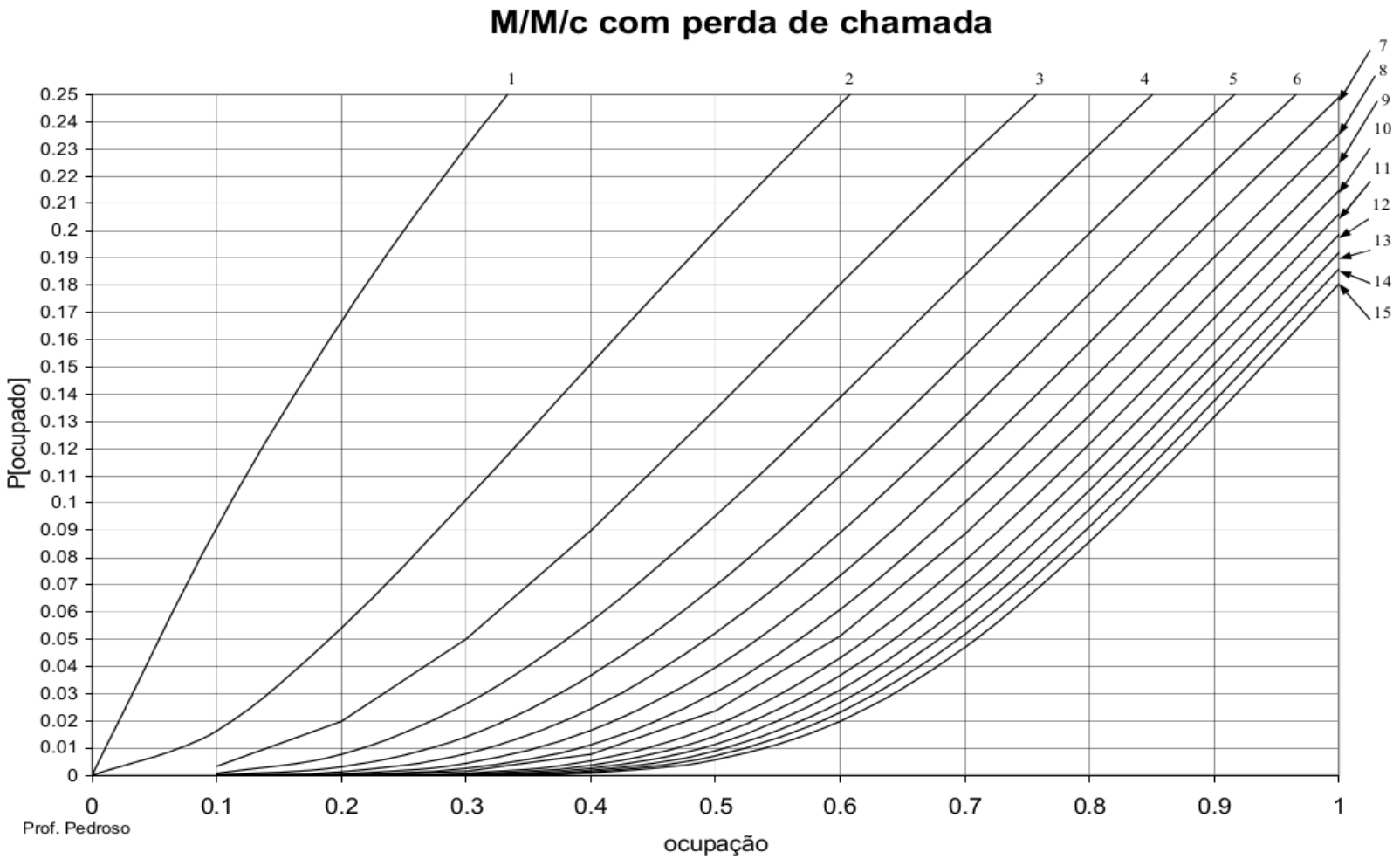
Calcule qual o tempo médio de espera na fila em cada um dos casos.

M/M/c com perda de chamada

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

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

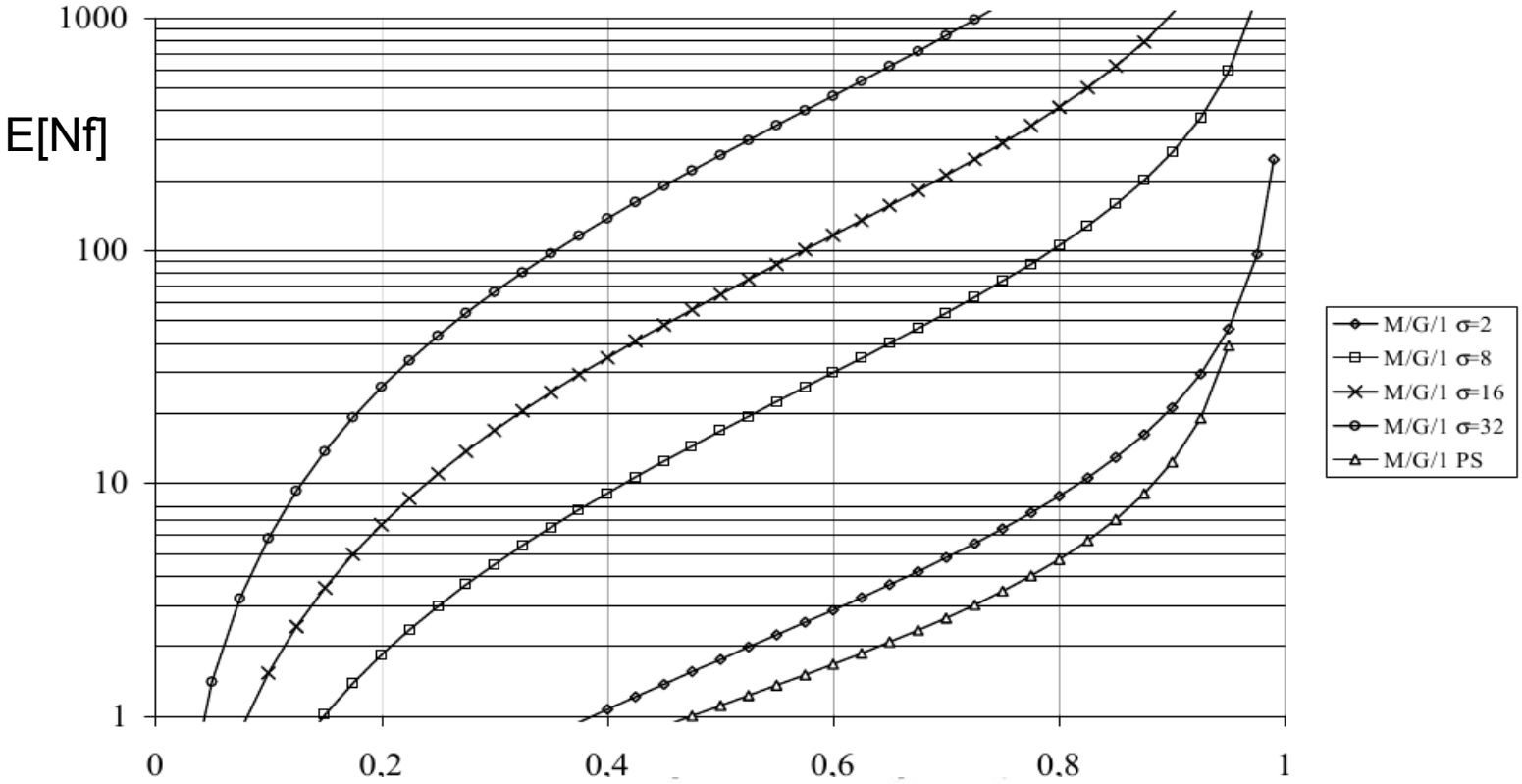
M/M/c com perda



Exemplo

- Considere um call center que recebe ligações de clientes. O sistema é modelado como M/M/c com perda de chamadas.
  - As chamadas chegam com média de 100 a cada hora.
  - O tempo médio de atendimento é de 4 minutos.
- Determine o número mínimo de agentes necessários para atender os requisitos de perda de chamada menor que:
  - 10%
  - 5%

M/G/1



$$E[Nf] = \frac{\rho^2}{2.(1-\rho)} \cdot \left[1 + \frac{\sigma^2}{E[Ta]^2}\right], \quad \rho < 1$$

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 σ/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

- M/D/1: $E[N_f] = \rho^2 / [2(1-\rho)]$
- D/M/1: $E[N_f] = \rho^2 / (1-\sigma)]$
  - σ (sigma) é uma variável crucial na análise do sistema de filas D/M/1. Ele representa a menor raiz positiva da equação transcendental que descreve o comportamento do sistema: $z = e^{-(1-z)/\rho}$, $0 < \sigma < 1$.
- G/M/1: $E[N_f] = \rho\sigma / (1-\sigma)$
  - σ é a menor raiz positiva da seguinte equação: $A^*(\mu(1-z))$
  - $A^*(s)$ é a Transformada de Laplace-Stieltjes da função de distribuição dos intervalos de chegada
  -

Redes de Filas: Sistemas Encadeados

- Uma rede de filas é um conjunto de filas individuais interconectadas. Os clientes, após serem atendidos em uma fila, podem se mover para outra, seguir para a saída ou retornar a uma fila anterior. É um modelo mais realista para sistemas complexos.
- Exemplos
  - Processos de produção em fábricas.
  - Fluxo de documentos em um escritório.
  - Pacotes de dados em uma rede de computadores.
  - Clientes em um hospital (recepção → triagem → médico).

Redes de Filas: Sistemas Encadeados

- Em uma fila M/M/1, as chegadas são Poisson (aleatórias) e o atendimento é Exponencial. A saída de uma fila M/M/1 tem uma propriedade crucial, descrita pelo Teorema de Burke: a taxa de saída de uma fila M/M/1 em estado estacionário também é Poisson.
  - Se as chegadas à segunda fila são Poisson, podemos tratá-la como um sistema M/M/1 independente. Isso simplifica drasticamente a análise de redes de filas.
  - Teorema de Jackson (generalização do Teorema de Burke): neste caso é possível analisar cada fila como se fosse um sistema independente, desde que a utilização de cada servidor seja menor que 1.
  - O Teorema de Jackson não se aplica se as distribuições não são exponenciais!
- Todos estes teoremas também são válidos para filas M/M/c, com múltiplos servidores.

Redes de Filas: Sistemas Encadeados

- Em um sistema encadeado, se o primeiro nó é, por exemplo, um M/D/1, a saída dele não será de Poisson. Em vez disso, a saída será um fluxo de clientes mais "regular" ou menos aleatório. Quando essa saída "não-Poisson" chega à próxima fila, a análise se complica porque as fórmulas para M/... (que pressupõem chegadas de Poisson) não são mais válidas. As filas não podem mais ser tratadas de forma independente.
- Como analisar filas encadeadas não M/M/… ?
  - Aproximações analíticas.
  - Simulações.

Exercício

- Uma empresas fabrica componentes metálicos em uma linha de produção com duas estações de trabalho em série. Os componentes chegam à linha seguindo uma distribuição de Poisson, com uma taxa de 10 componentes por hora, e o tempo de serviço em cada estação segue uma distribuição exponencial.
  - Estação 1: Corte a Laser: Esta estação possui um único operador e uma máquina de corte a laser, com capacidade tratar 12 componentes por hora. Os componentes cortados na Estação 1 seguem diretamente para a Estação 2.
  - Estação 2: Montagem: possui dois operadores, trabalhando em paralelo, que realizam a montagem final, com capacidade de 8 componentes por hora por operador.
- Calcule o tempo médio total que um componente gasta na linha de produção inteira.
- Se for adquirida uma estação de corte a laser com tempo de serviço determinístico (não segue a distribuição exponencial), isso tornaria o sistema mais eficiente?
- Se a estação 2 for automatizada, de forma que a montagem também seja determinística, qual seria o efeito?

Distribuições de Cauda Pesada

- E filas com processos de chegada baseados em distribuições de cauda pesada ou decaimento subexponencial, por exemplo, distribuição de Pareto (sistema P/P/1 ou P/P/c)?
  - Não existem expressões conhecidas.
  - Aproximações.
  - Existem expressões conhecidadas para os limites (superior, inferior) para o tamanho da fila.
  - Este problema é normalmente tratado com simulações computacionais, contornando a necessidade de solução analítica.
    - O problema ocorre para sistemas com chegadas modeladas por distribuições de Pareto, Lognormal, Weibull, Gamma e Log-Gamma, por exemplo.

Outros modelos

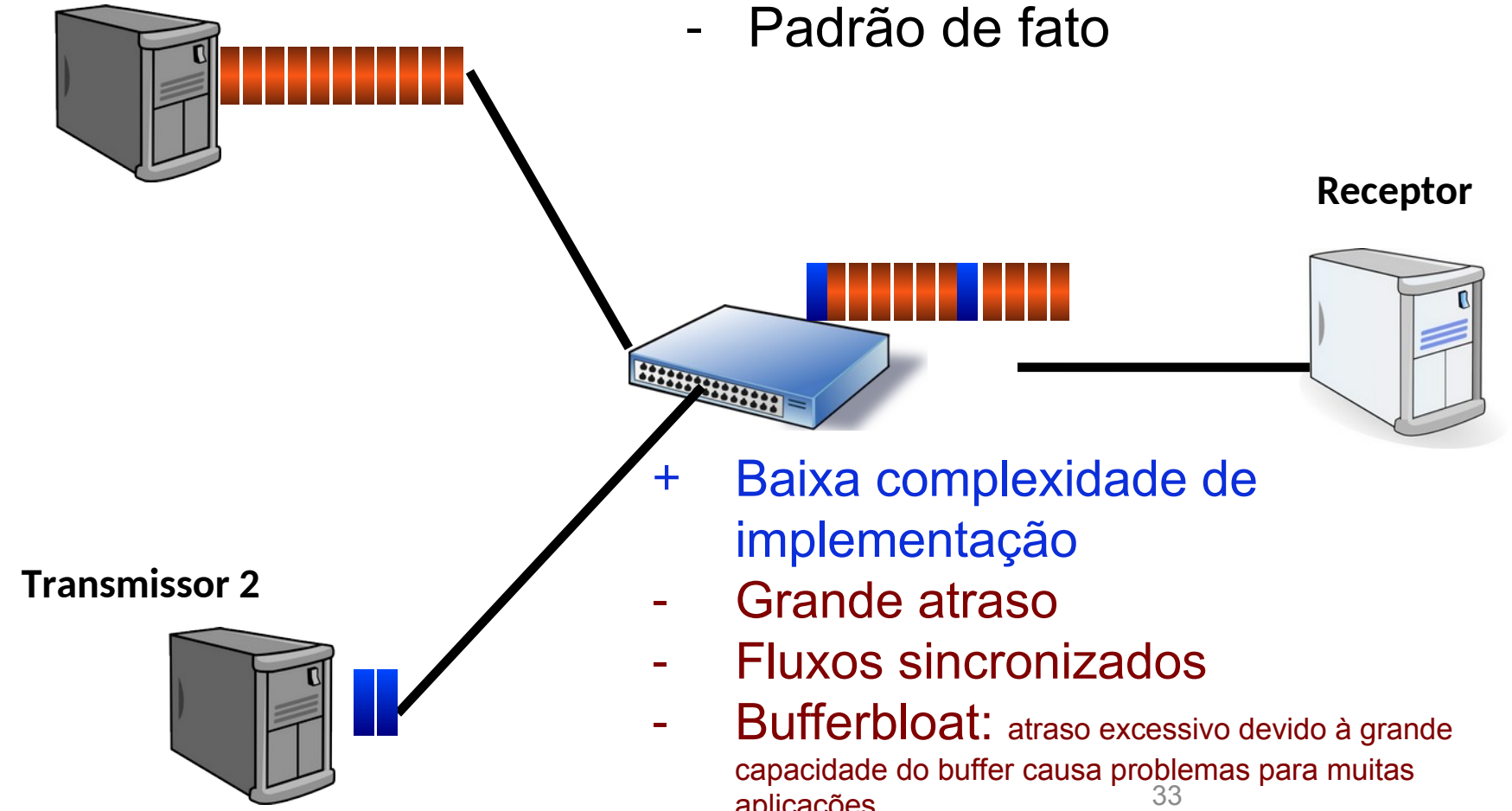
A literatura apresenta uma série de modelos prontos:

- 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



Perdas ocorrem devido à falta de espaço no buffer

- Padrão de fato

\+ Baixa complexidade de implementação

\- Grande atraso

\- Fluxos sincronizados

\- Bufferbloat: atraso excessivo devido à grande capacidade do buffer causa problemas para muitas aplicações

Random Early Detection (RED)

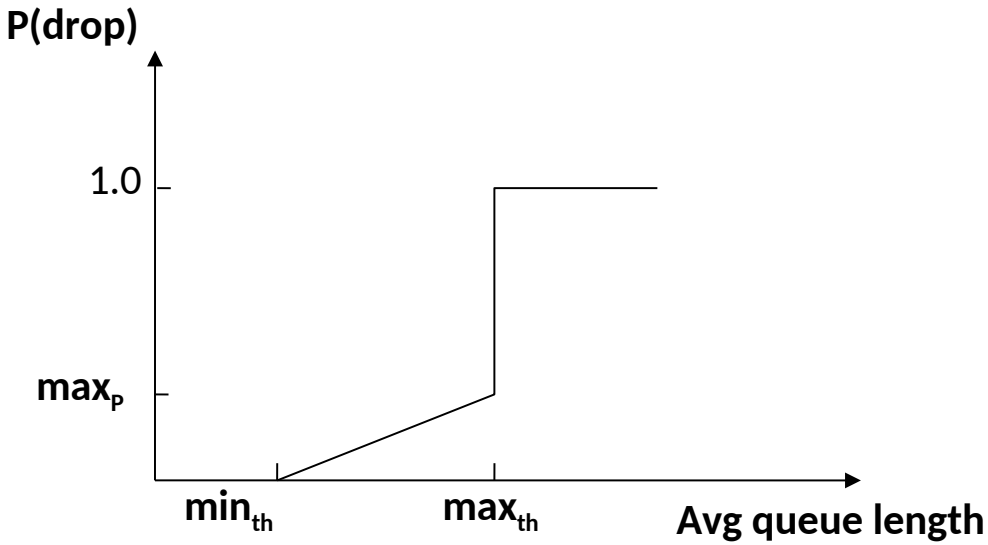
Proposto em 1993

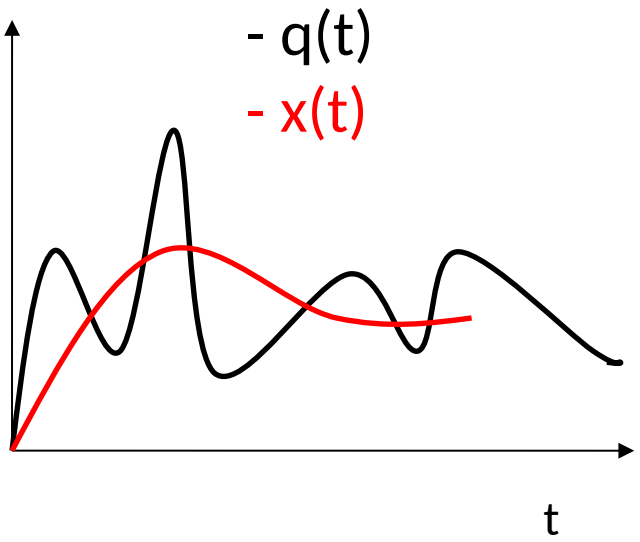
Descarta pacotes (com uma probabilidade de descarte) para prevenir situações de congestionamento reagindo antecipadamente.

Remove a sincronização entre fluxos.

RED

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





Every T: $x(t) = (1 - w_q) * x(t - T) + w_q * q(t)$

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

- Based on slide by Vishal Misra (Columbia)

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

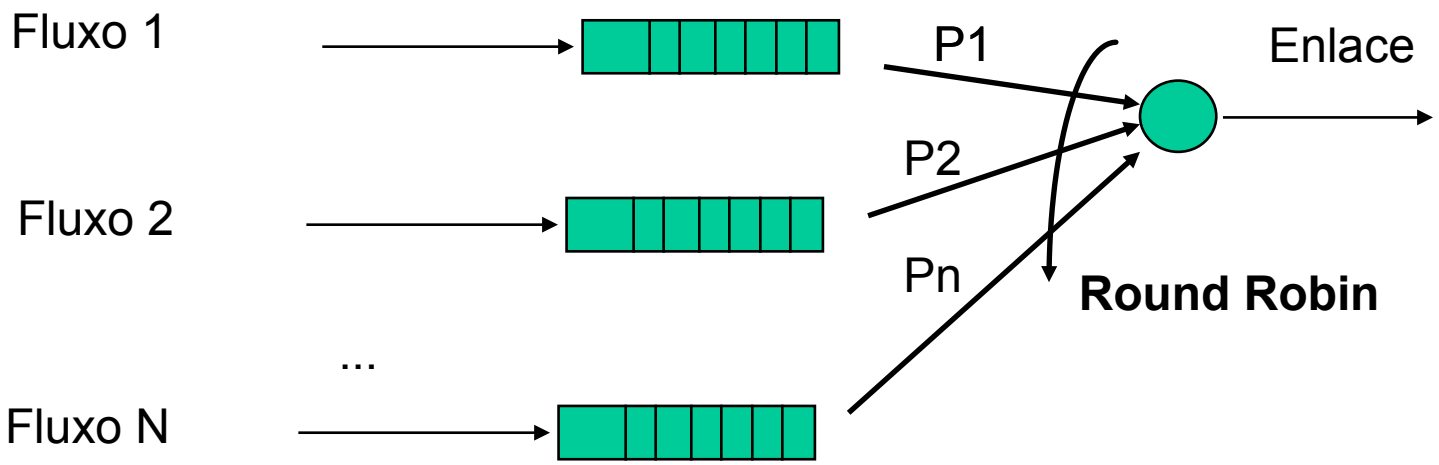
Cada fila possui uma disciplina específica

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!

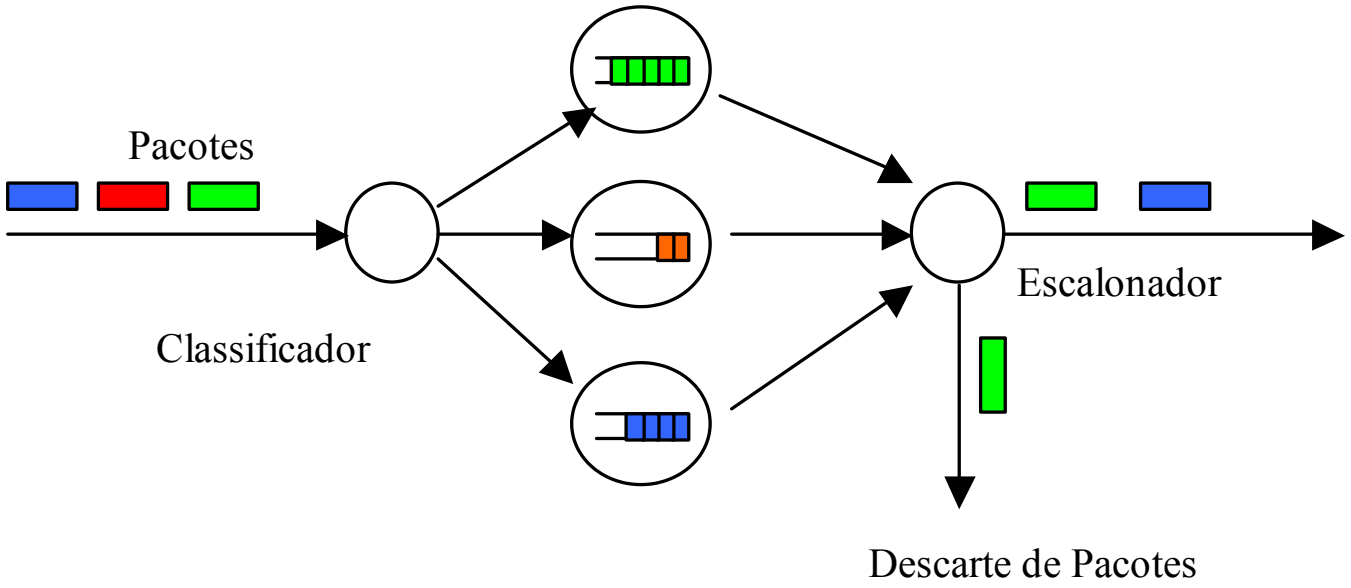
WFQ - Weighted Fair Queuing

- 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

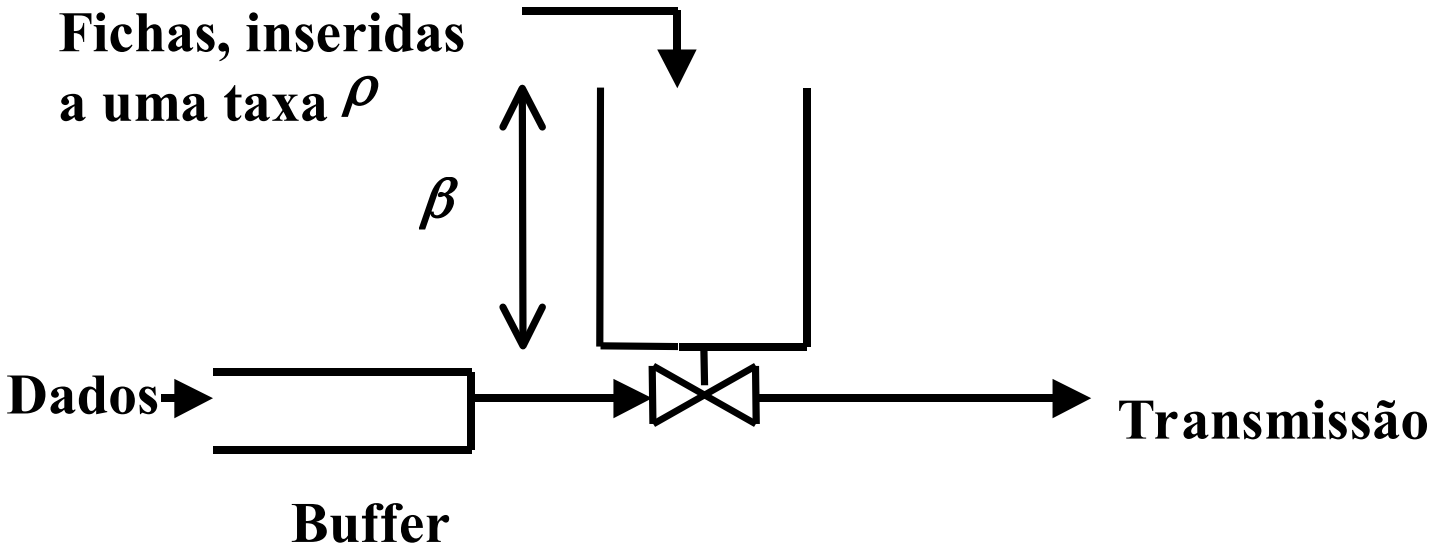
QoS no IP

Modelos:

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

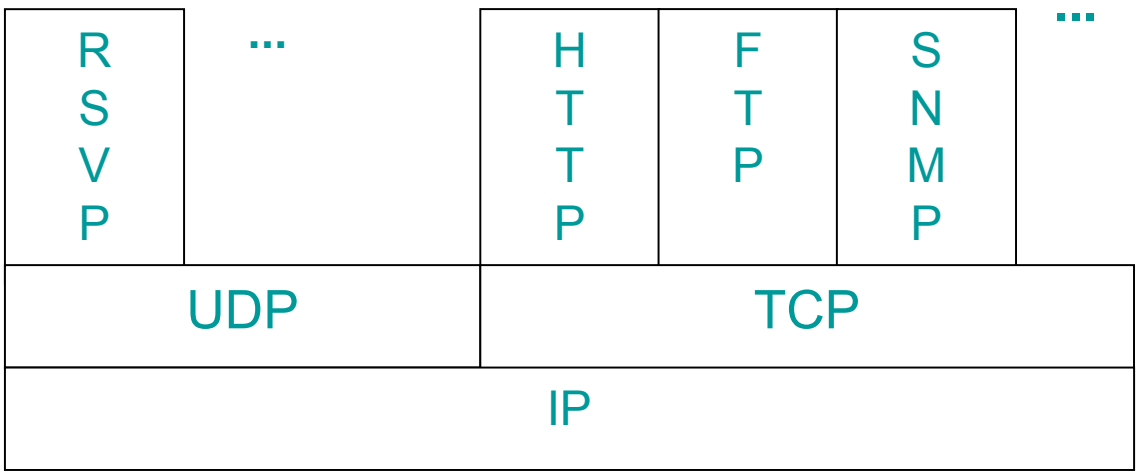
QoS no IP

Caracterização de tráfego:
algoritmo do balde de fichas.

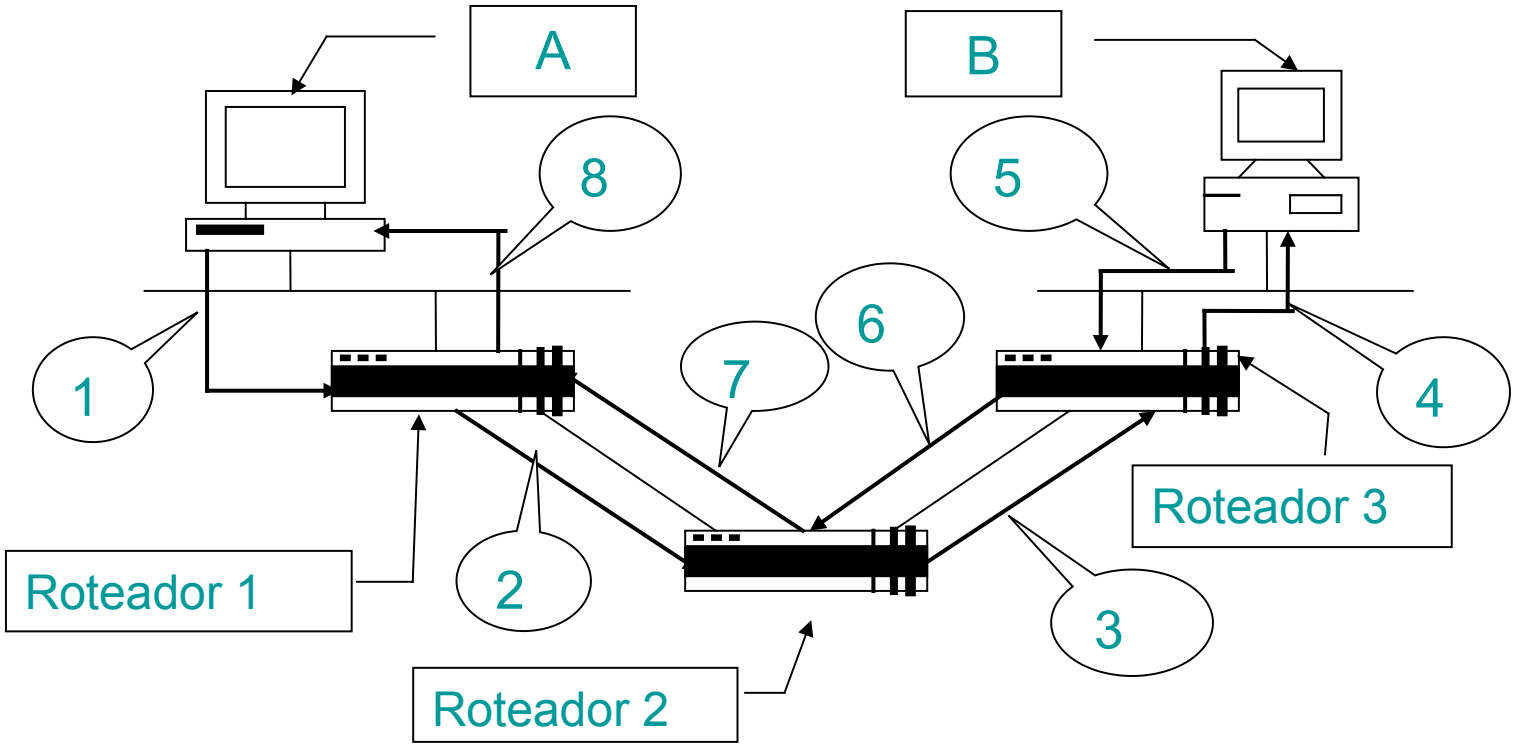


RSVP - Resource Reservation Protocol

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},...)*

Estilo Filtro Explícito Compartilhado

- *EC((S1,S2,...){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