Qualidade de Serviço

Redes Multimídia

Prof. Emerson Ribeiro de Mello

Instituto Federal de Santa Catarina – IFSC
campus São José
mello@ifsc.edu.br

17 de agosto de 2011

Sumário

Sumário

Introdução

- Existe uma competição pelos recursos (rede, ciclo do processador)
  - Aplicações convencionais *vs* aplicações multimídia
  - Fluxos de mídias de diferentes aplicações multimídia
- As redes foram projetadas para permitir o envio de mensagems de diferentes origens de forma intercalada
  - diversos canais virtuais sobre um mesmo canal físico

Introdução

- O padrão **Ethernet** é a tecnologia predominante nas redes locais
  - uso do melhor esforço para gerência do acesso ao meio compartilhado
- Qualquer nó pode usar o meio quando este estiver em silêncio
  - caso ocorra uma colisão, os nós emissores deverão aguardar um período aleatório até tentarem enviar os dados novamente
- Pontos chaves
  - não é possível garantir o tempo de entrega para todos os casos
  - consegue tratar com o aumento da demanda através da distribuição dos recursos disponíveis por todas as tarefas que estão competindo

Introdução

- O algoritmo de escalonamento varredura cíclica (*round-robin*) mostra-se como uma boa solução para compartilhamento de recursos, mas não atendem por completo aplicações multimídia
  - Os dados devem se fazer presentes em um determinado instante, após este os dados não possuem qualquer valor

Introdução

- O algoritmo de escalonamento varredura cíclica (*round-robin*) mostra-se como uma boa solução para compartilhamento de recursos, mas não atendem por completo aplicações multimídia
  - Os dados devem se fazer presentes em um determinado instante, após este os dados não possuem qualquer valor
- A **Qualidade de Serviço** (*Quality of Service* – QoS) é a grande área da computação que ataca este problema
  - Parte do pressuposto que nem todas as aplicações necessitam do mesmo desempenho em suas execuções
  - As aplicações são caracterizadas pelas necessidades de recursos em termos de requisitos de QoS

Como medir a qualidade de serviço de uma rede?

- Disponibilidade?
- Largura de banda
- Latência (atraso fim a fim)
- Variação de atraso
- Perda de pacotes

Como medir a qualidade de serviço de uma rede?

- Disponibilidade?
- Largura de banda
- Latência (atraso fim a fim)
- Variação de atraso
- Perda de pacotes

Aplicação	Confiabilidade	Atraso	Jitter	Largura de banda
WWW	Alta	Média	Baixa	Média
e-mail	Alta	Baixa	Baixa	Baixa
FTP	Alta	Baixa	Baixa	Média
SSH	Alta	Média	Média	Baixa
áudio/vídeo	Baixa	Alta	Alta	Média
VoIP	Baixa	Alta	Alta	Baixa

Qualidade de serviço

Objetivo das pesquisas com QoS

Permitir que os desenvolvedores das aplicações multimídia possam solicitar a infra-estrutura subjacente (S.O., rede) níveis de *QoS* apropriados para sua aplicação de forma que sejam mantidos desde a fonte até o destino (garantia fim a fim)

Qualidade de serviço

Objetivo das pesquisas com QoS

Permitir que os desenvolvedores das aplicações multimídia possam solicitar a infra-estrutura subjacente (S.O., rede) níveis de *QoS* apropriados para sua aplicação de forma que sejam mantidos desde a fonte até o destino (garantia fim a fim)

- **Gerenciamento de Qualidade de Serviço**
  - Responsável por determinar como e quais recursos serão disponibilizados para cada aplicação
    - **Negociação de QoS**
    - **Controle de admissão**

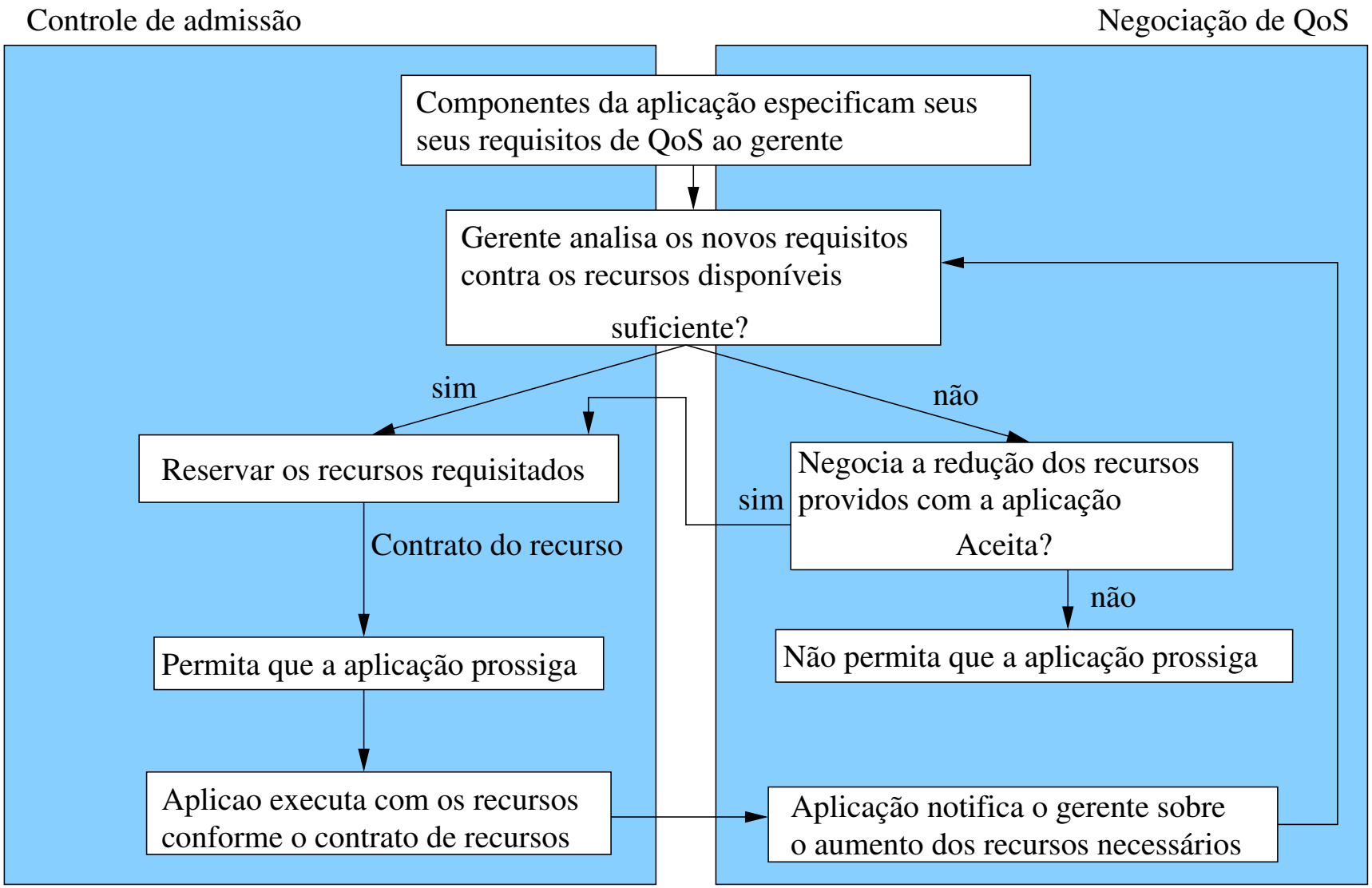
Negociação de QoS – funcionamento

1. A aplicação indica os recursos necessários
2. Gerente verifica a possibilidade em aceitar tal proposta
   - Recursos disponíveis *vs* atual comprometimento dos recursos
3. Apresenta uma resposta a aplicação: sim ou não
   - Se for **negativa** a aplicação pode ser reconfigurada para usar uma quantidade reduzida de recursos e submete um novo pedido
   - Se for **positiva** aplica-se o **controle de admissão**.

Controle de admissão

- Reserva os recursos e retorna um contrato (com tempo de expiração) à aplicação indicando quais recursos foram reservados de fato
- Se a aplicação mudar seus requisitos, esta deve notificar o gerente
  - Se os requisitos diminuírem, a "sobra" voltará para a base de recursos
  - Se aumentarem, então será necessária uma nova rodada de negociações

Tarefas do gerente de QoS



Negociação de QoS

- A negociação entre a aplicação e o gerente se faz através de uma **especificação de requisitos**
- Para as aplicações multimídia especifica-se três parâmetros
  - **Largura de banda.** Taxa que o fluxo será transportado
  - **Latência.** Indica o tempo necessário para o transporte.
  - **Taxa de perda.** Em um gerenciamento de QoS perfeito nunca teremos perda de pacotes
    - Prover tal garantia geralmente não é aceitável, uma vez que recai sobre uma reserva excessiva de recursos para tratar eventuais picos na rede
    - Saída comum: aceitar uma certa taxa de perdas

Negociação de QoS

Largura de banda, latência e taxa de perdas podem ser usadas para:

- **Descrever as características de um fluxo multimídia**

- **Descrever as habilidade dos recursos para transportar o fluxo**

Negociação de QoS

Largura de banda, latência e taxa de perdas podem ser usadas para:

- **Descrever as características de um fluxo multimídia**
  - Um fluxo de mídia requer uma **largura de banda** média de 1.5Mpbs e o atraso máximo tolerado é de 150ms. O algoritmo de descompressão tolera perder 1 quadro a cada 100
- **Descrever as habilidade dos recursos para transportar o fluxo**

Negociação de QoS

Largura de banda, latência e taxa de perdas podem ser usadas para:

- **Descrever as características de um fluxo multimídia**
  - Um fluxo de mídia requer uma **largura de banda** média de 1.5Mpbs e o atraso máximo tolerado é de 150ms. O algoritmo de descompressão tolera perder 1 quadro a cada 100
- **Descrever as habilidade dos recursos para transportar o fluxo**
  - Uma rede pode prover conexões de 64kbps e seu algoritmo de enfileiramento de pacotes garante atrasos menores que 10ms. Já a transmissão garante uma taxa de perdas menor que 1 em 10^6

RFC 1363 - Especificação do Fluxo

	Versão do protocolo
Largura de banda	Unidade máxima de transmissão Taxa de geração de fichas Tamanho do balde Taxa máxima de transmissão
Atraso	Atraso mínimo Variação máximo de atraso
Perda	Sensibilidade a perda Sensibilidade a rajadas de perda Intervalo de perdas
	Garantia da qualidade de serviço

Dificuldades com os parâmetros de QoS

- Algumas técnicas de compressão de vídeo produzem um fluxo de vídeo com uma taxa de *bits* variável (VBR)
  - Uma grande largura de banda será necessária quando o conteúdo mudar rapidamente
    - cenas de ação
  - Rajadas de tráfego podem fazer que elementos de mídia cheguem mais cedo que o esperado (de acordo com a taxa de chegada)

Sumário

Cenário 1

- Usuário de uma aplicação FTP contrata um serviço mais caro que o usuário de uma aplicação de áudio
  - Deveria o tráfego de áudio ter prioridade sobre o FTP?

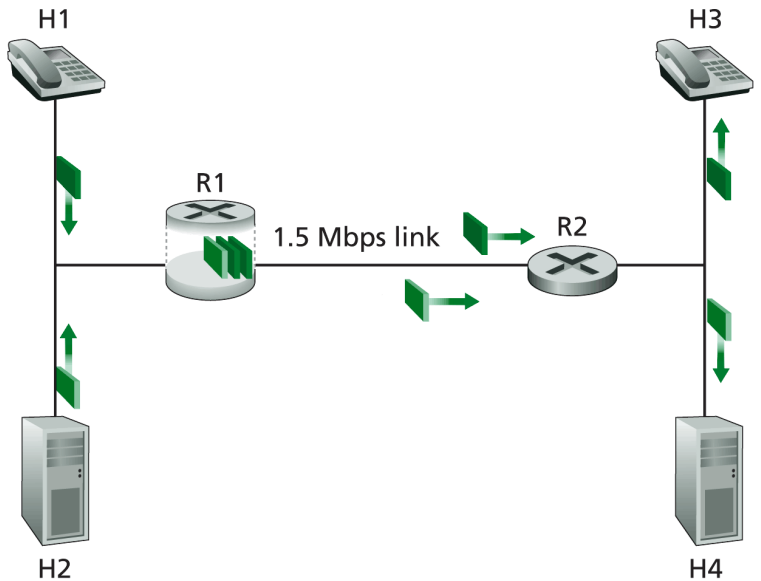


Figure 7.19 ♦ Competing audio and FTP Applications

Cenário 1

- Usuário de uma aplicação FTP contrata um serviço mais caro que o usuário de uma aplicação de áudio
  - Deveria o tráfego de áudio ter prioridade sobre o FTP? **Não!**

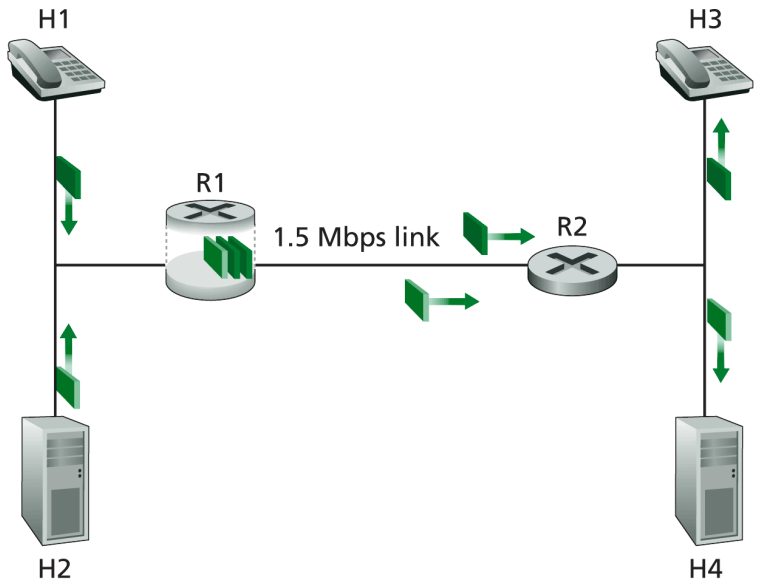


Figure 7.19 ♦ Competing audio and FTP Applications

- Deve haver alguma classificação dos pacotes para permitir aos roteadores distinguir os pacotes de diferentes classes

Cenário 1: Classificação de pacotes

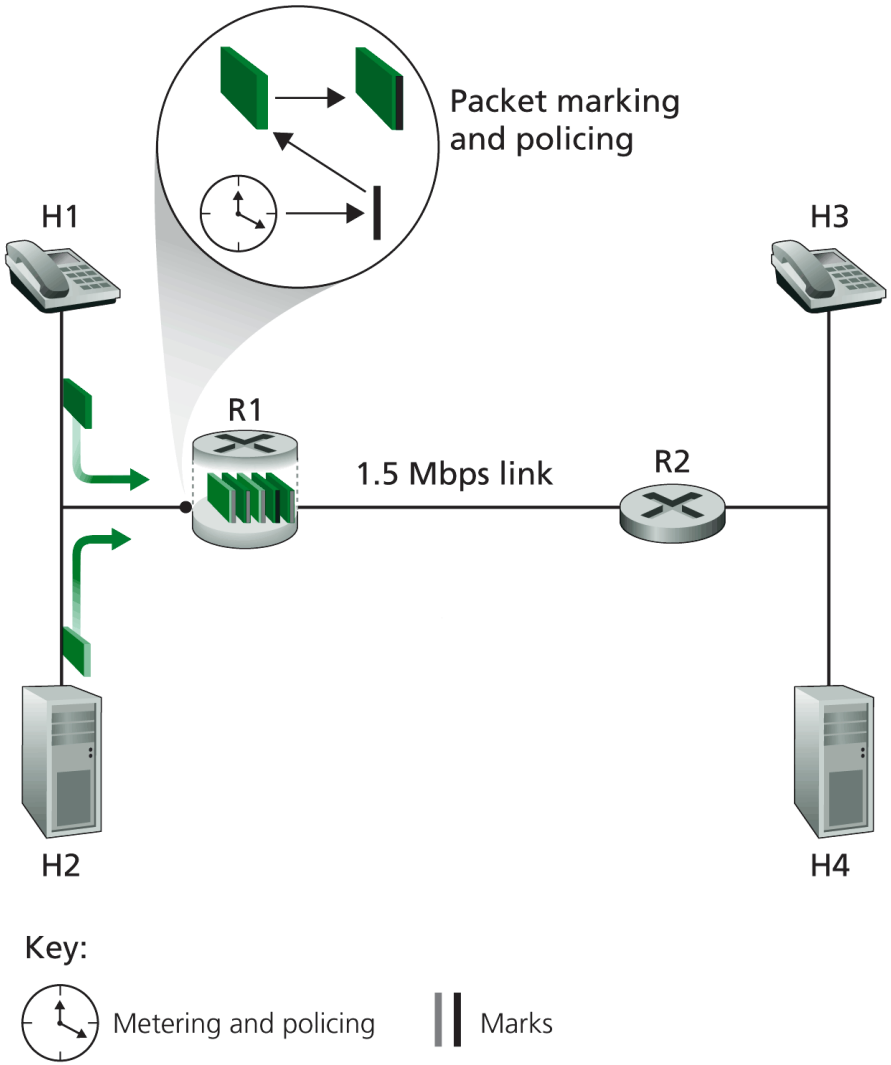


Figure 7.20 ♦ Policing (and marking) the audio and FTP traffic flows

Cenário 2

- O roteador sabe que deve dar $1 Mbps$ a aplicação de áudio, pois essa é a taxa de transmissão do áudio e o restante é destinado ao FTP. Porém o que acontece se a aplicação de áudio começar a transmitir na taxa $1.5 Mbps$?

Cenário 2

- O roteador sabe que deve dar $1Mbps$ a aplicação de áudio, pois essa é a taxa de transmissão do áudio e o restante é destinado ao FTP. Porém o que acontece se a aplicação de áudio começar a transmitir na taxa $1.5Mbps$?
  - Os pacotes FTP nunca serão transmitidos.

Cenário 2

- O roteador sabe que deve dar $1Mbps$ a aplicação de áudio, pois essa é a taxa de transmissão do áudio e o restante é destinado ao FTP. Porém o que acontece se a aplicação de áudio começar a transmitir na taxa $1.5Mbps$?
  - Os pacotes FTP nunca serão transmitidos.
- Deve haver um isolamento entre as classes de tráfego e entre os fluxos
- Criação de enlaces lógicos

Cenário 2: Criação de enlaces lógicos

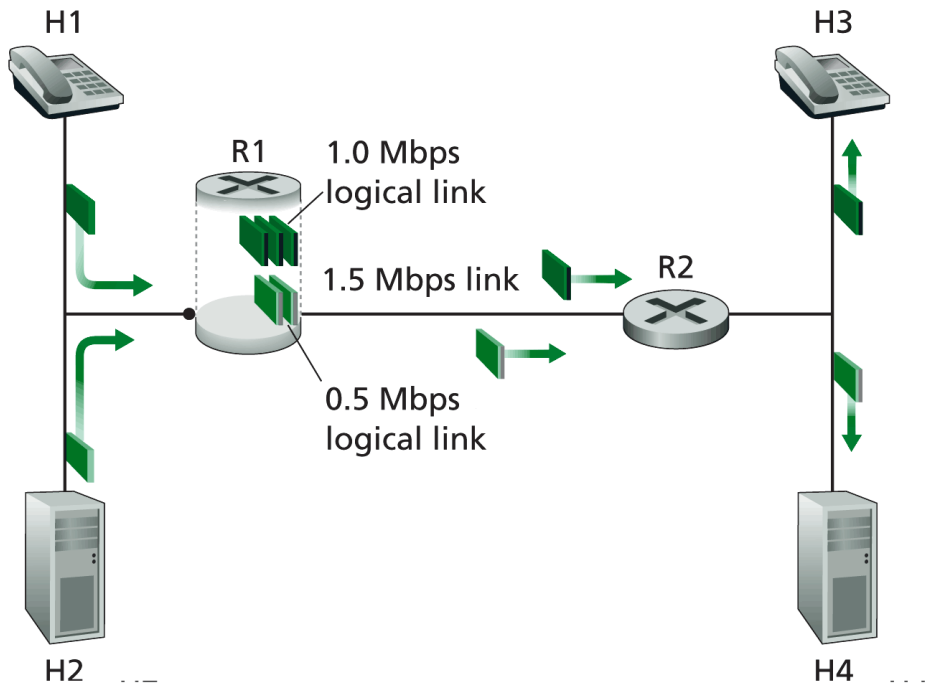


Figure 7.21 ♦ Logical isolation of audio and FTP application flows

Cenário 3

- A aplicação de áudio tem um enlace lógico de $1Mbps$. Ambas as aplicações estão transmitindo até que a aplicação de áudio para de transmitir (pararam de falar). Assim, existe uma largura de banda ociosa e que a aplicação FTP não pode usar.

Cenário 3

- A aplicação de áudio tem um enlace lógico de $1Mbps$. Ambas as aplicações estão transmitindo até que a aplicação de áudio para de transmitir (pararam de falar). Assim, existe uma largura de banda ociosa e que a aplicação FTP não pode usar.
- Ao prover o isolamento de classes e de fluxos é desejado utilizar os recursos da melhor maneira possível

Sumário

Escalonamento de filas

- Pacotes chegam de diferentes fluxos de rede e são multiplexados em uma fila para a sua transmissão
- Existem diferentes maneiras para selecionar os pacotes que estão na fila para serem transmitidos, isto chama-se **disciplina de escalonamento de filas**
  - FIFO – *Firs-In-First-Out*
  - Enfileiramente com prioridades
  - Varredura cíclica (*Round robin*)
  - Varredura cíclica com enfileiramento justo ponderado
    - *Round Robin* + *Weighted Fair Queueing* – RR WFQ)

FIFO

- FIFO – O primeiro pacote a chegar será o primeiro pacote a sair
- Pacotes que chegam na fila de saída aguardam por sua transmissão se o meio estiver ocupado
- Se não houver espaço suficiente na fila para armazenar pacotes que estão chegando, então a **política de descarte** irá determinar quais pacotes deverão ser descartados

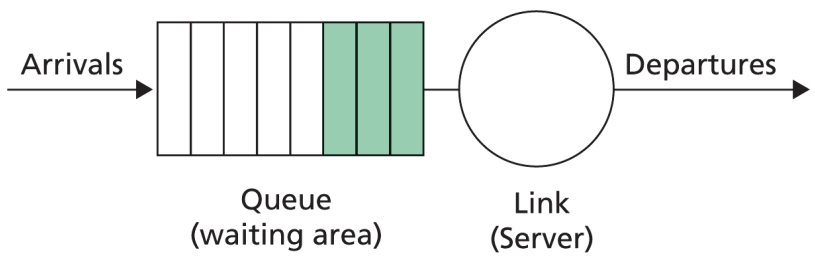


Figure 7.23 ♦ FIFO queuing abstraction

FIFO em operação

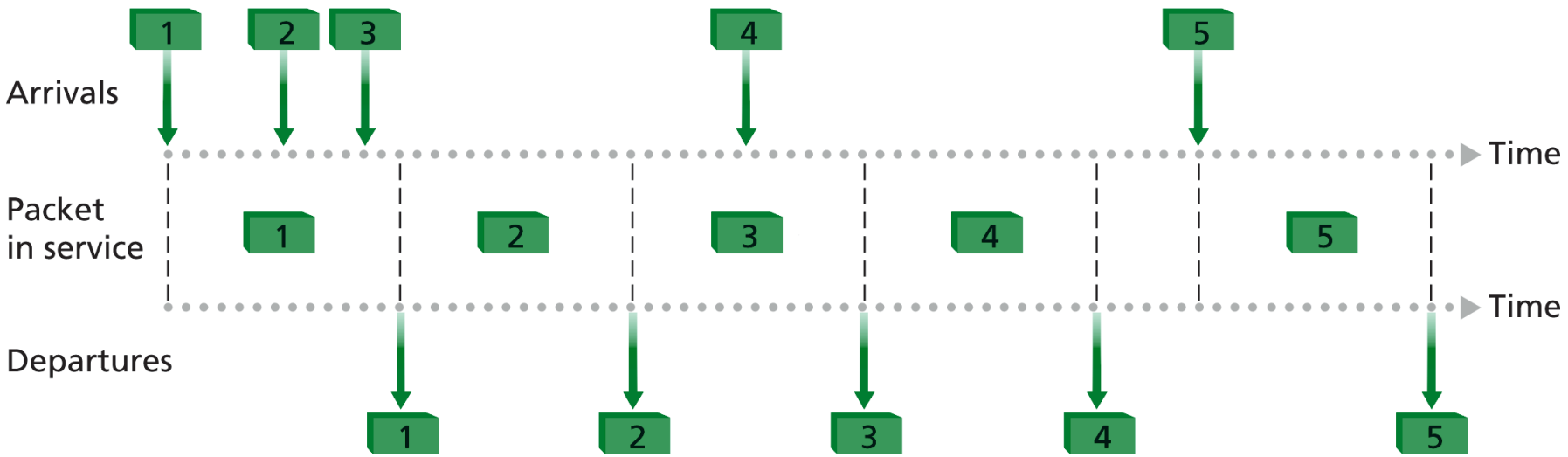


Figure 7.24 ♦ The FIFO queue in operation

Enfileiramento com prioridades

- Pacotes que chegam são classificados e dispostos em diferentes filas de acordo com a prioridade que possuem
  - Classificação pode usar o cabeçalho IP do pacote
- Sempre irá transmitir o pacote que está na fila de mais alta prioridade
  - Se existirem duas filas com a mesma prioridade, o FIFO é empregado

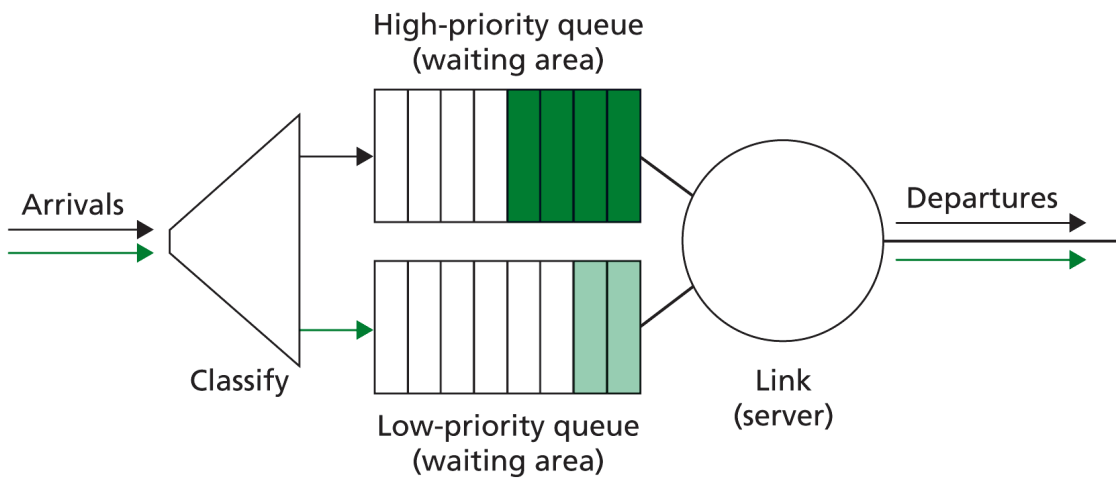


Figure 7.25 ♦ Priority queuing model

Enfileiramento com prioridades em operação

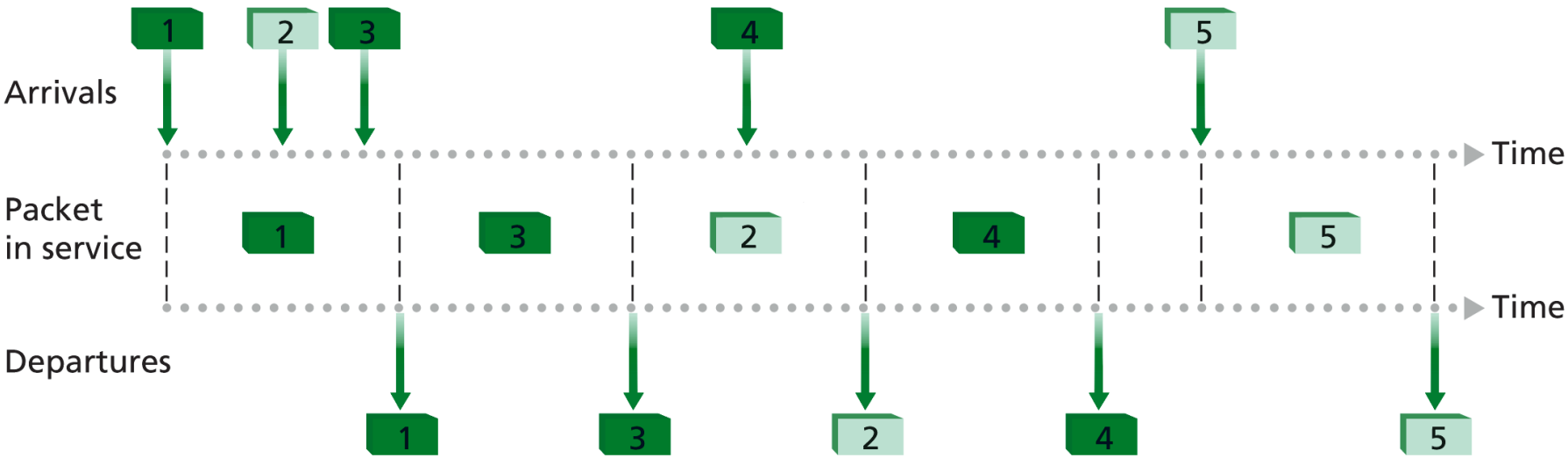
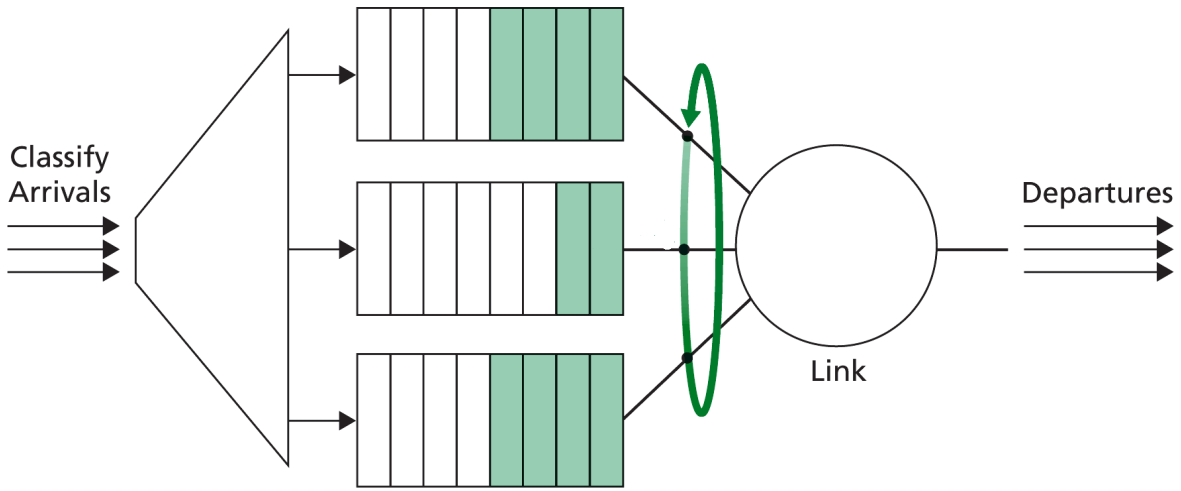


Figure 7.26 ♦ Operation of the priority queue

Varredura cíclica (Round robin)

- Classifica os pacotes que entram e os dispõe em diferentes filas
- As filas são servidas igualmente



Varredura cíclica (Round robin)

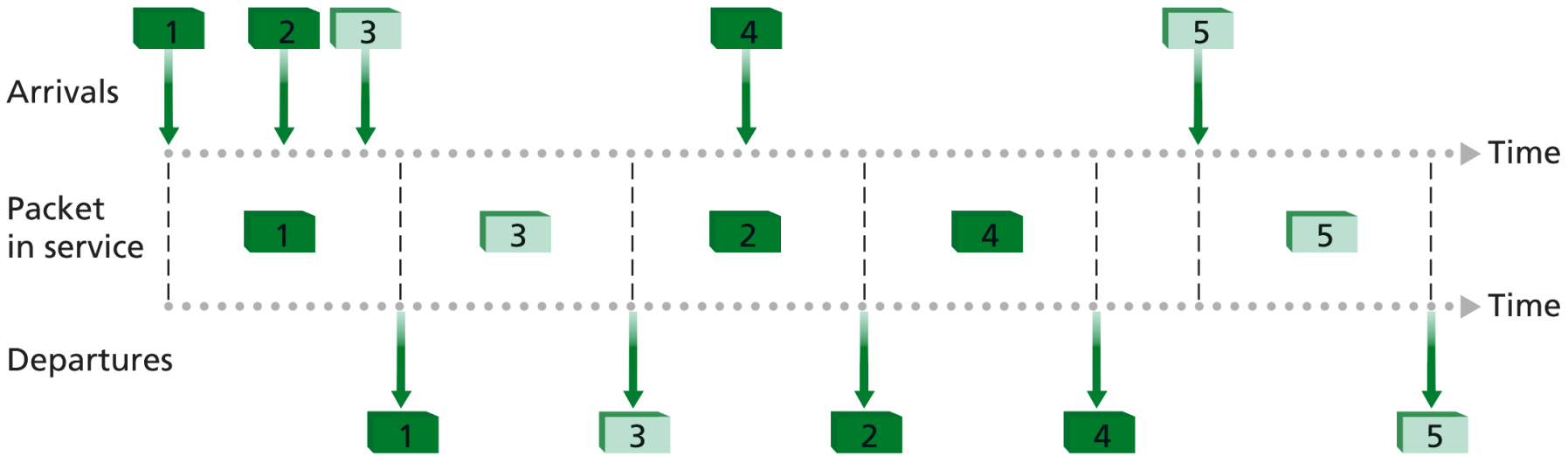


Figure 7.27 ♦ Operation of the two-class round robin queue

Varredura cíclica com enfileiramento justo ponderado

- Cada classe recebe um tempo de serviço diferenciado em cada ciclo
- Para cada classe i está associado um peso w_i. Durante um intervalo de tempo em que a classe i tem pacotes para enviar, a esta será concedida um fração do serviço igual a $\frac{w_i}{\sum w_j}$,
- A soma do denominador é obtida através de todas as classes que também possuem pacotes para transmitir.
- Assim, com uma taxa de transmissão R, a classe i sempre irá obter um vazão mínima de $R * \frac{w_i}{\sum w_j}$

Varredura cíclica com enfileiramento justo ponderado

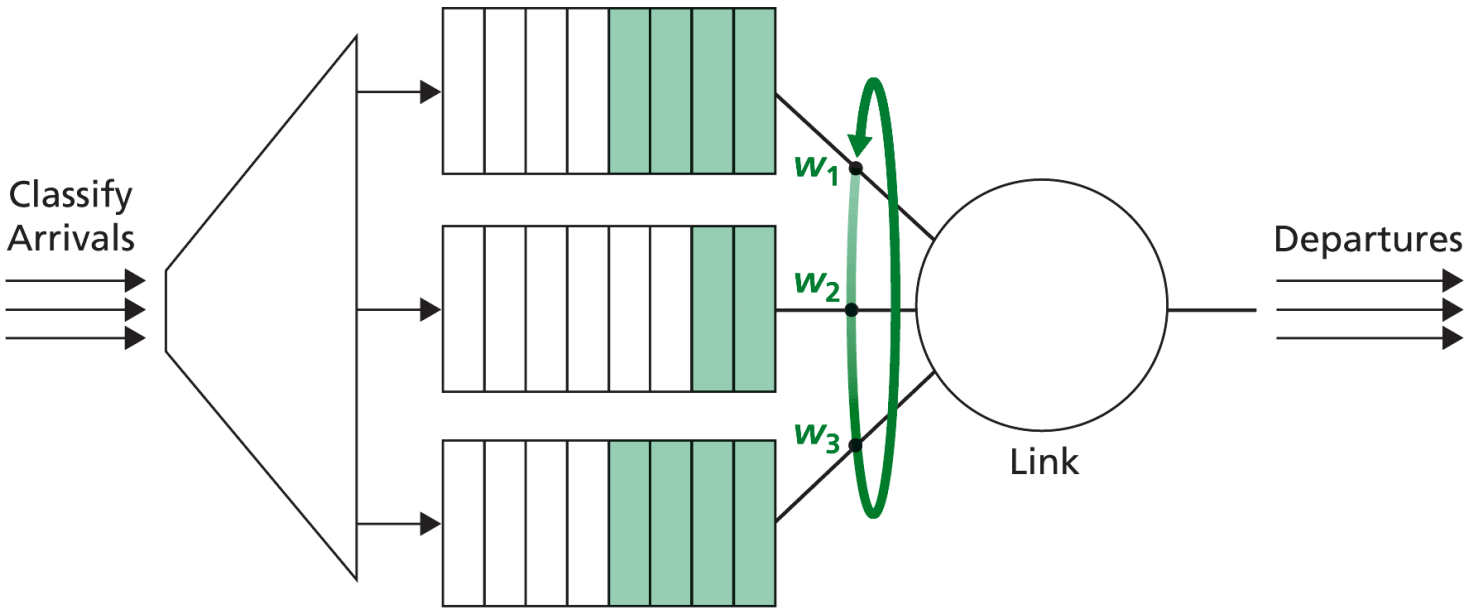


Figure 7.28 ♦ Weighted fair queuing (WFQ)

Sumário

Condicionamento de tráfego

Condicionamento de tráfego – *Traffic Shapping*

Termo usado para descrever o uso de uma área de armazenamento temporário (*buffer*) para suavizar o fluxo de dados de forma que a **saída dos dados seja contínua**

Condicionamento de tráfego

Condicionamento de tráfego – *Traffic Shapping*

Termo usado para descrever o uso de uma área de armazenamento temporário (*buffer*) para suavizar o fluxo de dados de forma que a **saída dos dados seja contínua**

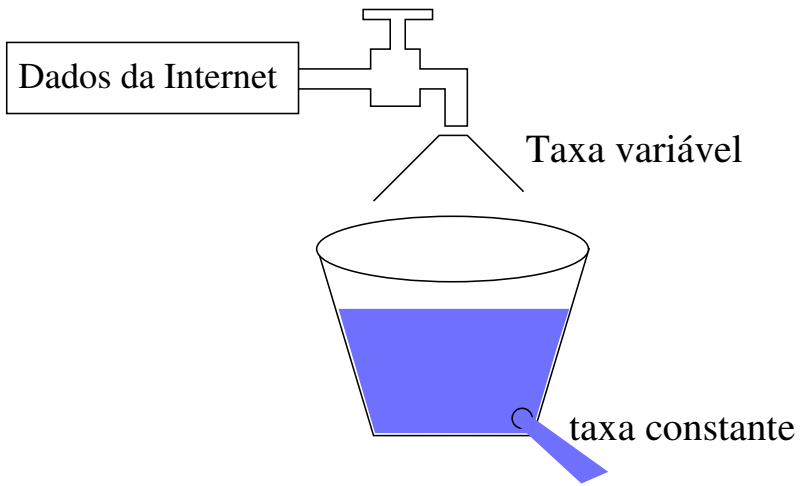
- Técnicas para fazer a suavização do tráfego
  - Balde Furado
  - Balde de fichas

Critérios para o condicionamento de tráfego

- **Taxa média**
  - Pode ser desejado limitar a taxa média de tráfego (pacotes) sobre um intervalo de tempo longo
  - Quantidade de tráfego que será injetada na rede
  - Ex: 100 pacotes por segundo é uma taxa mais restritiva que 6.000 pacotes por minuto
- **Taxa de pico**
  - Restringe o número máximo de pacotes em um intervalo de tempo mais curto
  - Ex: taxa média de 6.000 pacotes por minuto e uma taxa de pico de 1.500 pacotes por segundo
- **Tamanho da rajada**
  - Número máximo de pacotes injetados na rede em um período de tempo extremamente curto

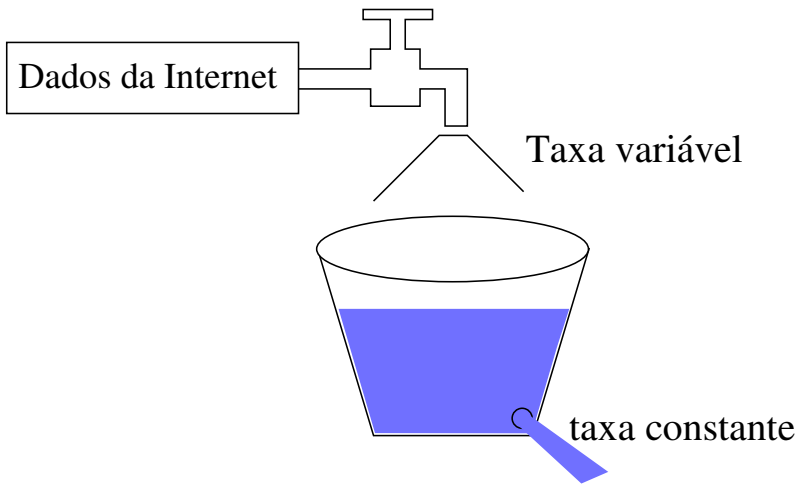
Balde furado

- O tráfego chega a uma taxa variável suscetível a rajadas enquanto que a saída é linear



Balde furado

- O tráfego chega a uma taxa variável suscetível a rajadas enquanto que a saída é linear



- **Problema:** A eliminação de rajadas nem sempre é necessária
  - exemplo: Quando se tem uma largura de banda disponível sobre um certo intervalo de tempo

Balde de fichas

- Permite que grande rajadas ocorram quando o fluxo tenha ficado **ocioso** por um tempo
- Balde de tamanho B é preenchido com fichas sendo geradas a uma taxa constante R
- Um dado de tamanho s só poderá ser enviado sem houver pelo menos s fichas no balde
- Isto garante que sobre qualquer intervalo de tempo t a quantidade de dados enviados não será superior a $Rt + B$

Balde de fichas

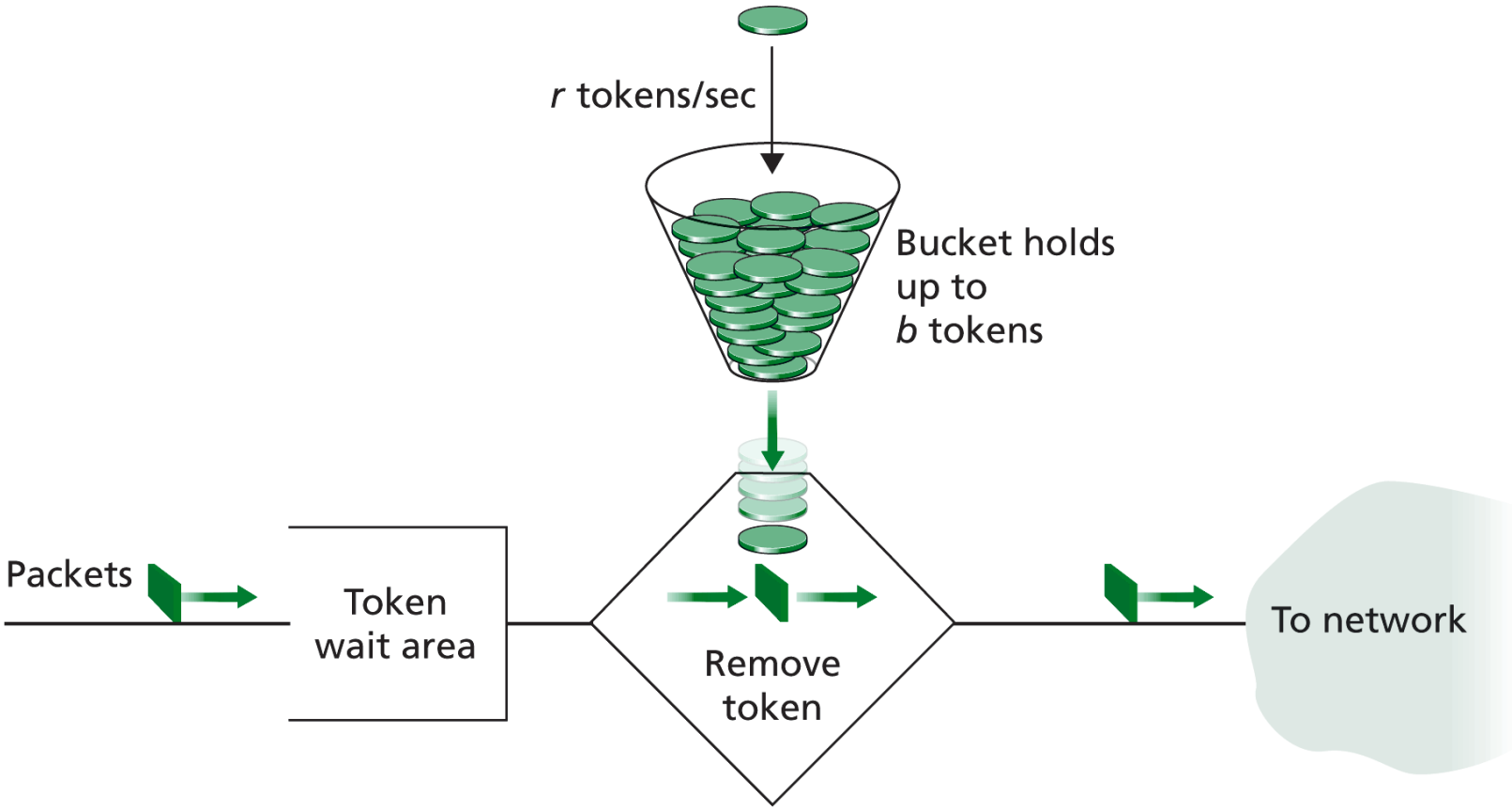


Figure 7.29 ♦ The leaky bucket policer

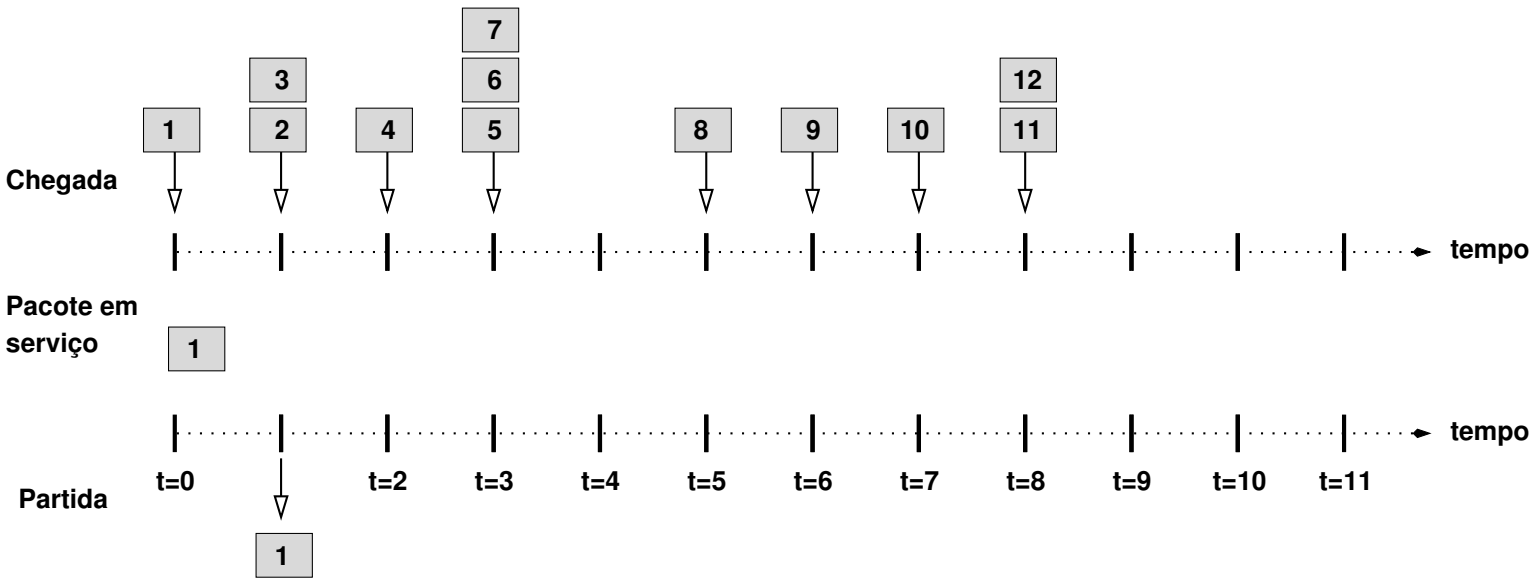
Exercícios

1. Ao usar uma política de fila RR + WFQ em uma fila que provê suporte a três classes de tráfego, sendo os pesos: 0.3, 0.5 e 0.2, em que sequência devem ser servidas estas classes, considerando que cada classe possui um grande número de pacotes na fila.
    - E no caso de não haver pacotes na classe 2, qual seria a sequência?

Exercícios

Analise a figura abaixo e para cada uma das seguintes disciplinas para escalonamento de filas (FIFO, enfileiramento prioritário, varredura cíclica e enfileiramento justo ponderado) responda:

- O instante de saída de cada um dos pacotes de 2 a 12.
- O atraso de cada pacote desde a sua chegada até o instante de sua transmissão. Indique ainda o atraso médio para os 12 pacotes.



Exercícios

Assuma:

Enfileiramento prioritário. Os pacotes ímpares são considerados de alta prioridade. Os pacotes pares são considerados de baixa prioridade;

Varredura cíclica. Os pacotes (`1, 2, 3, 6, 11, 12`) fazem parte da classe 1 e os pacotes (`4, 5, 7, 8, 9, 10`) fazem parte da classe 2.

Enfileiramento justo ponderado. Os pacotes ímpares fazem parte da classe 1. Os pacotes pares fazem parte da classe 2. A classe 1 possui peso 2 e a classe 2 possui peso 1.

Bibliografia


George Couloris, Jean Dollimore and Tim Kindberg
Distributed Systems: Concepts and Design - 4th edition
Addison-wesley, 2005


James F. Kurose and Keith W. Ross
Redes de computadores e a Internet: Uma abordagem top-down, 3ª edição
Addison-Wesley, 2005.