Comparando política de tráfego e modelagem de tráfego para o limite de largura de banda

Índice

Introdução

Este documento esclarece as diferenças funcionais entre modelagem e vigilância, que limitam a taxa de saída. Mesmo que os dois mecanismos utilizem um token bucket como medidor de tráfego para medir a taxa de pacote, eles possuem diferenças funcionais importantes. (a seção What Is a Token Bucket? (O que é um Token Bucket?) contém a descrição de um token bucket).

Antes de Começar

Convenções

Consulte as Convenções de Dicas Técnicas da Cisco para obter mais informações sobre convenções de documentos.

Pré-requisitos

Não existem requisitos específicos para este documento.

Componentes Utilizados

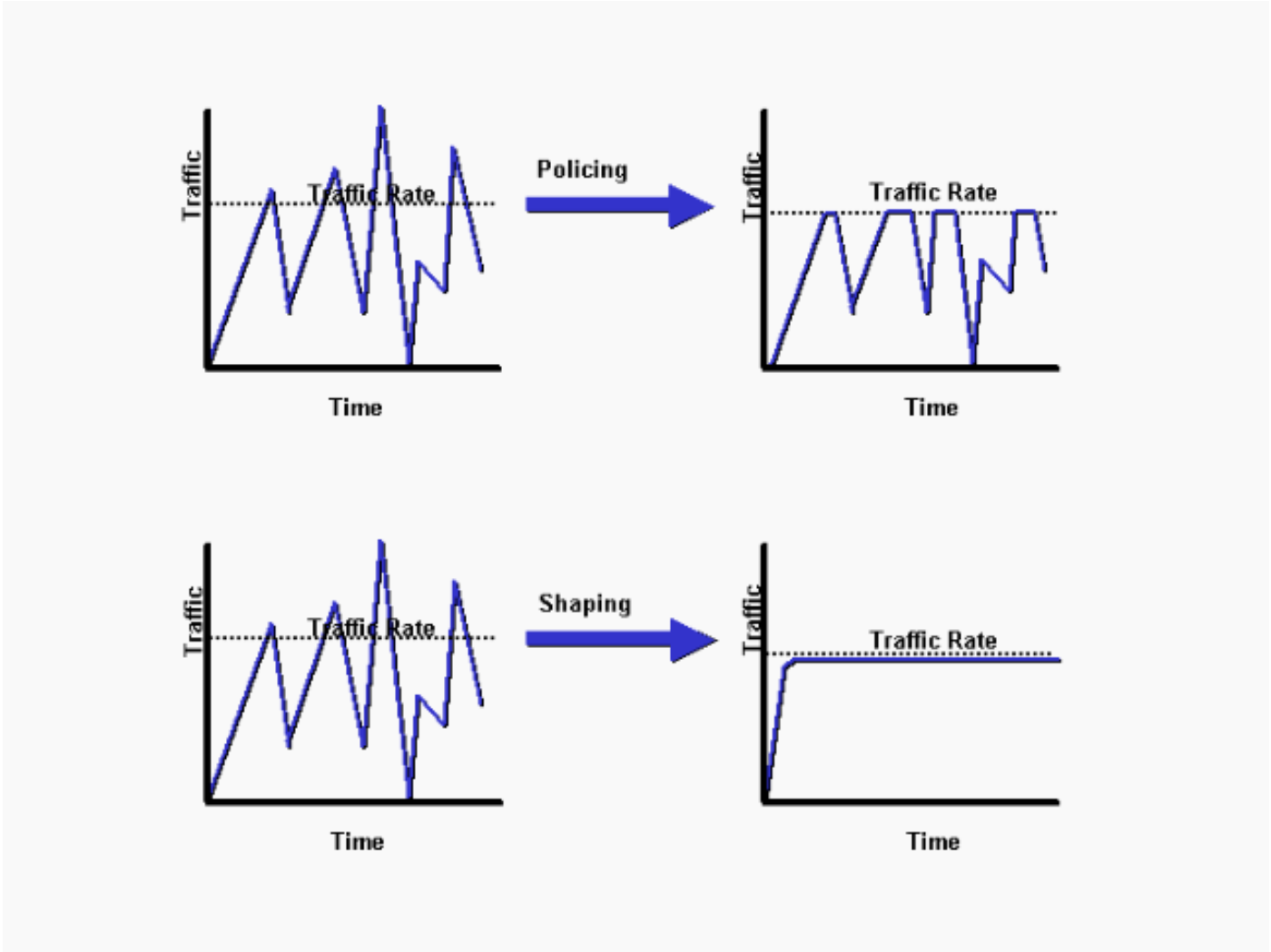
Este documento não se restringe a versões de software e hardware específicas.

As informações neste documento foram criadas a partir de dispositivos em um ambiente de laboratório específico. Todos os dispositivos utilizados neste documento foram iniciados com uma

configuração (padrão) inicial. Se você estiver trabalhando em uma rede ativa, certifique-se de que entende o impacto potencial de qualquer comando antes de utilizá-lo.

Vigilância versus modelagem

O diagrama a seguir ilustra a diferença principal: A política de tráfego propaga os bursts. Quando a taxa de tráfego atinge a taxa máxima configurada, o tráfego de excesso é liberado (ou remarcado). O resultado é uma taxa de saída que aparece como um dente de serra com picos e depressões. Em contraste com a vigilância, a modelagem de tráfego retém pacotes em excesso em uma fila e agenda o excesso para transmissão posterior de acordo com incrementos de tempo. O resultado da modelagem de tráfego é uma taxa de saída de pacote facilitada.



A modelagem implica na existência de uma fila e de memória suficiente para armazenar os pacotes com retardo em buffer, diferente da vigilância. Queueing é um conceito de saída; os pacotes que saem de uma relação são enfileirados e podem ser modelados. Somente a política pode ser aplicada ao tráfego de entrada em uma interface. Certifique-se de que você tenha memória suficiente ao habilitar a modelagem. Além disso, a modelagem requer uma função de agendamento para transmissão posterior dos pacotes retardados. Essa função de programação permite que você organize a fila de modelagem em diferentes filas. Exemplos de funções de programação são o Enfileiramento justo ponderado com base em classe (CBWFQ, Class Based Weighted Fair Queuing) e o Enfileiramento de latência baixa (LLQ, Low Latency Queuing).

Critérios de seleção

A tabela a seguir lista as diferenças entre modelagem e política para ajudá-lo a escolher a melhor solução.

	Modelagem	Vigilância
Objetivo	Pacotes excedentes de fila e buffer acima das taxas comprometidas.	Desconecte (ou marque) os pacotes em excesso acima das taxas comprometidas. Não é possível armazenar em buffer.*
Taxa de atualização de token	Incrementada no início de um intervalo de tempo. (É necessário um número mínimo de intervalos.)	Contínua com base em fórmula: `1 / committed information rate`
Valores de token	Configurado em bits por segundo.	Configurado em bytes.
Opções de configuração	• **shape** comando na Interface de linha de comando de qualidade de serviço modular (MQC, Modular Quality of Service Command-Line Interface) para implementar a class-based shaping. • Comando frame-relay traffic-shape para implementar modelagem de tráfego de Frame Relay (FRTS). • comando traffic-shape para implementar o GTS (Generic Traffic Shaping).	• Comando police no MQC para implementar a vigilância baseada em classe. • comando rate-limit para implementar a Taxa de acesso consolidada (CAR).
Aplicável na Entrada	Não	Sim
Aplicável na Saída	Sim	Sim

Intermitências	Controla as intermitências suavizando a taxa de saída ao longo de pelo menos oito intervalos de tempo. Utiliza um vazamento de bucket para retardar tráfego, o que causa um efeito facilitador.	Propaga intermitências. Não faz suavização.
Vantagens	Menor probabilidade de descartar pacotes em excesso, visto que estes sofrem buffer. (Pacotes de buffer até o comprimento da fila. Descartes podem ocorrer se o tráfego em excesso for mantido a taxas altas.) Normalmente evita retransmissões devido a pacotes descartados.	Controla a taxa de saída por meio do cancelamento de pacotes. Evita atrasos devido ao enfileiramento.
Desvantagens	Pode introduzir um retardo devido ao enfileiramento, especialmente em caso de filas grandes.	Descarta pacotes em excesso (quando configurada) controlando os tamanhos das janelas TCP e diminuindo a taxa de saída total de fluxos de tráfego afetados. Tamanhos de burst excessivamente agressivos podem levar a quedas excessivas de pacotes e acelerar a taxa global de saída, especificamente com fluxos baseados em TCP.
Remarcação de Pacote Opcional	Não	Sim (com recurso de CAR legado).

*Embora a vigilância não se aplique ao buffer, um mecanismo de enfileiramento configurado se aplica a pacotes adaptados que podem precisar ser enfileirados enquanto esperam ser serializados na interface física.

Taxa de atualização de token

Uma diferença importante entre modelagem e política é a taxa à qual os tokens são recarregados. Esta seção examina a diferença.

Em termos simples, a modelagem e a política usam a metáfora do token bucket. Um token bucket propriamente dito não possui política de descarte ou prioridade. Vamos observar como a metáfora de bucket de token funciona:

- Os tokens são colocados em um bucket a uma certa taxa.
- Cada token é uma permissão para que a origem envie um determinado número de bits para a rede.
- Para enviar um pacote, o regulador de tráfego deve estar apto a remover do bucket um número de símbolos que seja igual em representação ao tamanho do bucket.
- Se não houver tokens suficientes no bucket para enviar um pacote, o pacote esperará até que o bucket tenha tokens suficientes (no caso de um formador) ou o pacote será descartado ou marcado (no caso de um vigilante).
- O bucket propriamente dito possui uma capacidade específica. Se o bucket for totalmente preenchido, os tokens recém-chegados serão descartados e não estarão disponíveis para pacotes futuros. Assim, a qualquer momento, o maior surto que uma origem pode enviar na rede é proporcional ao tamanho do pacote. Um token bucket permite intermitência, mas a limita.

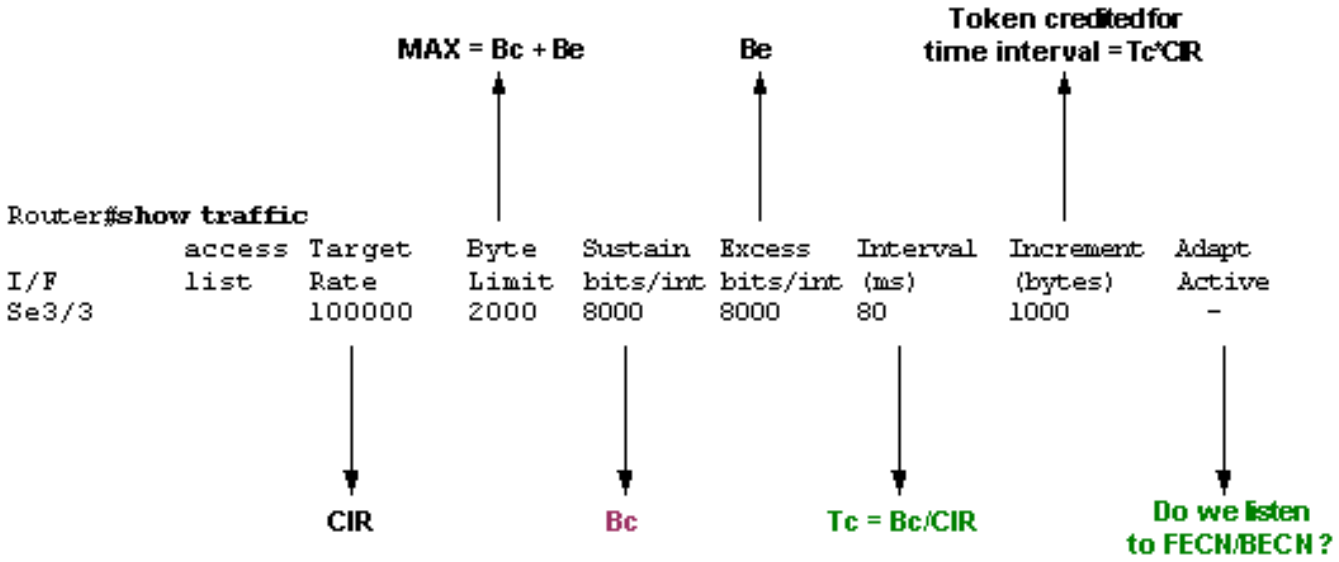
Como a metáfora do token bucket em mente, vamos observar como a modelagem e a vigilância adiciona tokens ao bucket.

A modelagem incrementa o token bucket a intervalos de tempo determinados usando um valor de bits por segundo (bps). Um modelador usa a seguinte fórmula:

```
Tc = Bc/CIR (in seconds)
```

Nessa equação, Bc representa o burst comprometido, e CIR quer dizer Taxa de informação comprometida (Committed Information Rate). (Consulte Configuração da modelagem de tráfego do frame relay para obter mais informações.) O valor de Tc define o intervalo de tempo durante o qual você envia os bits Bc para manter a taxa média da CIR em segundos.

O intervalo de Tc é entre 10 ms e 125 ms. Com Modelagem de tráfego distribuído (DTS, Distributed Traffic Shaping) no Cisco 7500 Series, o Tc mínimo é de 4 ms. O roteador calcula internamente esse valor com base nos valores da CIR e do Bc. Se Bc/CIR for menor que 125 ms, ele utiliza o Tc calculado a partir daquela equação. Se o Bc/CIR for maior que ou igual a 125 ms, ele utiliza um valor Tc interno, caso o Cisco IOS® determine que o fluxo de tráfego será mais estável com um intervalo menor. Utilize o comando show traffic-shape para determinar se o roteador está utilizando um valor interno para Tc ou o valor configurado na linha de comando. O exemplo de saída a seguir do comando show traffic-shape é explicado em Comandos show para modelagem de tráfego de Frame Relay.



Quando o excesso de burst (Be, excess burst) é configurado para um valor diferente de 0, o modelador permite que os tokens sejam armazenados no bucket até Bc + Be. O maior valor que o token bucket pode alcançar é Bc + Be, e os tokens de sobrefluxo são descartados. A única maneira de ter mais tokens Bc no bucket é não utilizar todos os tokens Bc durante um ou mais Tcs. Como o token bucket é repovoado a cada Tc com tokens Bc, é possível acumular tokens não utilizados para uso posterior em Bc + Be.

Em contraste, a class-based policing e a taxa limite adicionam tokens ao bucket continuamente. Especificamente, a taxa de chegada do token é calculada da seguinte maneira:

```
(time between packets<which is equal to t-t1>* policer rate)/8 bits per byte
```

Em outras palavras, se a chegada anterior do pacote é t1 e o horário atual é t, o bucket é atualizado com o valor de bytes de t-t1 com base na taxa de chegada do token. Note que um policer de tráfego usa os valores de burst especificados nos bytes, e a fórmula acima converte de bits para bytes.

Observe um exemplo usando uma CIR (ou a taxa do policer) de 8000 bps e um burst normal de 1000 bytes.

```
Router(config)# policy-map police-setting Router(config-pmap)# class access-match Router(config-pmap-c)# police 8000 1000 conform-action transmit exceed-action drop
```

Os token buckets são iniciados completos a 1000 bytes. Se um pacote com 450 bytes chegar, ele é aceito, pois há bytes suficientes disponíveis no token bucket. A ação de ajuste (transmissão) é realizada pelo pacote e os 450 bytes são removidos do token bucket (deixando 550 bytes). Se o próximo pacote chegar 0,25 segundo depois, 250 bytes serão adicionados ao token bucket, conforme a seguinte fórmula:

```
(0.25 * 8000)/8
```

O cálculo deixa 700 bytes no token bucket. Se o próximo pacote tiver 800 bytes, o pacote implicará excesso e a ação exceder (cancelamento) será executada. Nenhum byte foi retirado do token bucket.

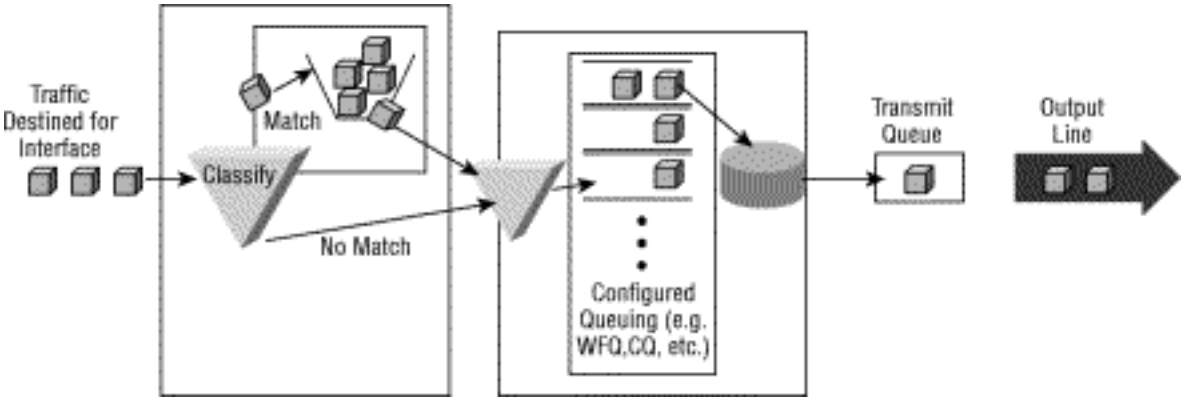
Modelagem de tráfego

O Cisco IOS suporta os seguintes métodos de modelagem de tráfego:

- Modelagem de tráfego genérico
- Modelagem de tráfego de Frame Relay
- Modelagem com base em classes e Modelagem Distribuída com base em classes

Todos os métodos de modelagem são semelhantes em implementação, embora suas interfaces de linha de comandos (CLIs) sejam um pouco diferentes e usem tipos diferentes de filas para conter e modelar o tráfego adiado. A Cisco recomenda modelagem com base em classes e modelagem distribuída, que são configuradas com o uso da CLI de QoS modular.

O diagrama a seguir ilustra como uma política de QoS separa o tráfego em classes e enfileira os pacotes que excedem as taxas de modelagem configuradas.



Vigilância de tráfego

O Cisco IOS suporta os seguintes métodos de vigilância de tráfego:

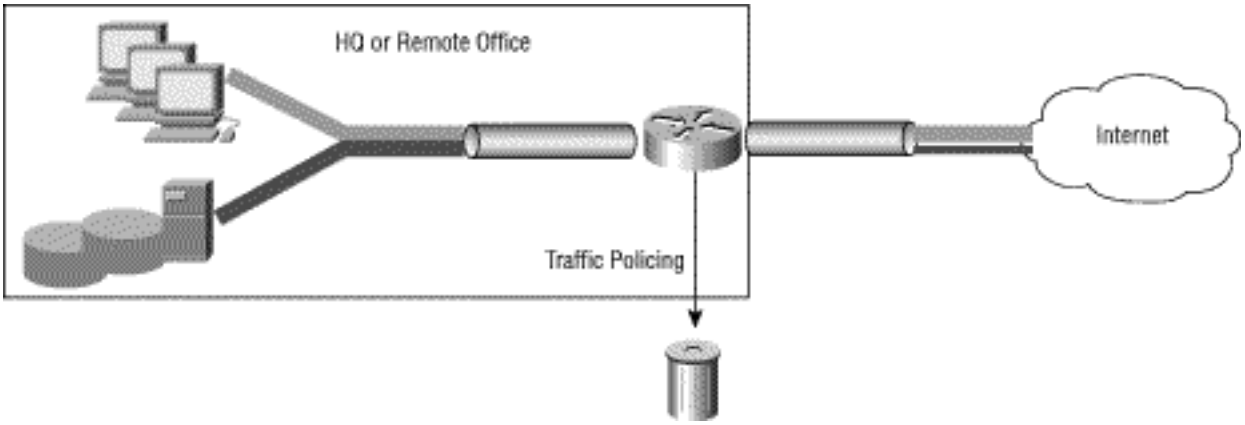
- Taxa de acesso comprometida
- Vigilância baseada em classe

Os dois mecanismos têm importantes diferenças funcionais, conforme explicado em Comparação de Vigilância Baseada em Classe e Taxa de Acesso Comprometida. A Cisco recomenda a class-based policing e outros recursos da CLI de QoS modular ao aplicar políticas de QoS.

Use o comando **police** para especificar que uma classe de tráfego deve ter uma taxa máxima imposta a ela e, se essa taxa for excedida, uma ação imediata deverá ser tomada. Ou seja, com o comando de vigia, não é possível armazenar o pacote no buffer e, posteriormente, enviá-lo, como é o caso do comando de modelagem.

Além disso, com a vigilância, o token bucket determina se um pacote excede ou corresponde à taxa aplicada. Em todo caso, a vigilância implementa uma ação configurável, que inclui a definição da precedência de IP ou DSCP (Ponto de código de serviços diferenciados).

O diagrama a seguir ilustra um aplicativo comum de vigilância de tráfego em um ponto de congestionamento, onde os recursos de QoS geralmente se aplicam.



Controles de largura de banda mínima versus máxima

Os comandos shape e police restringem a taxa de saída a um valor máximo em kbps. O mais importante é que nenhum dos mecanismos fornece uma garantia de largura de banda mínima durante períodos de congestionamento. Use o comando bandwidth ou priority para fornecer essas garantias.

Uma política hierárquica usa duas políticas de serviços – uma política principal para aplicar um mecanismo de QoS a um agregado de tráfego e uma política secundária para aplicar um mecanismo de QoS a um fluxo ou a um subconjunto do agregado. Interfaces lógicas, como subinterfaces e interfaces de túnel, exigem uma política hierárquica com o recurso de limite de tráfego no nível principal e enfileiramento em níveis inferiores. O recurso de limite de tráfego reduz a taxa de saída e (supostamente) cria um congestionamento, conforme acontece com pacotes em excesso enfileirados.

A configuração a seguir não é a ideal e é exibida para ilustrar a diferença entre os comandos **police** e shape ao limitar um agregado de tráfego – neste caso, padrão de classe – a uma taxa máxima. Nesta configuração, o comando **police** envia pacotes das classes secundárias com base no tamanho do pacote e do número de bytes que restam nos token buckets de conformidade e exceção. (Consulte Política de tráfego.) O resultado é que as taxas atribuídas às classes de Voz sobre IP (VoIP, Voice over IP) e Protocolo de Internet (IP, Internet Protocol) não podem ser garantidas, pois o recurso **police** está cancelando as garantias dadas pelo recurso priority.

Entretanto, se o comando shape for utilizado, o resultado será um sistema de enfileiramento hierárquico, e todas as garantias serão implementadas. Em outras palavras, quando a carga oferecida excede a taxa de forma, as classes VoIP e IP têm garantia de sua taxa, e o tráfego padrão da classe (no nível infantil) incorre em qualquer queda.

Cuidado: Essa configuração não é recomendada e é mostrada para ilustrar a diferença entre o comando police e o comando shape durante a limitação de um agregado de tráfego.

```
class-map match-all IP
   match ip precedence 3
class-map match-all VoIP
   match ip precedence 5

policy-map child
  class VoIP
     priority 128
   class IP
     priority 1000
```

```
 policy-map parent
   class class-default
     police 3300000 103000 103000 conform-action transmit exceed-action drop
     service-policy child
```

Para que a configuração acima faça sentido, a política deve ser substituída pela modelagem. Por exemplo:

```
policy-map parent
    class class-default
     shape average 3300000 103000 0
      service-policy child
```

A fim aprender mais sobre o pai e as políticas infantil, refira por favor a política de serviços da criança de QoS para a classe de prioridade.

Informações Relacionadas

- Suporte da tecnologia de QoS
- Suporte Técnico - Cisco Systems