**Construindo Tecidos Sem Perdas**

Introdução

Você sabia que uma única perda de pacote em uma rede de inteligência artificial e aprendizado de máquina (IA/ML) de alto desempenho pode levar a uma degradação significativa do desempenho e potencialmente atrasar cálculos críticos em vários minutos ou até horas? No mundo atual, impulsionado por dados, onde as aplicações de IA/ML estão na vanguarda da inovação tecnológica, garantir uma transmissão de dados contínua e sem perdas não é apenas um luxo — é uma necessidade.

Bem-vindo ao treinamento "Construindo Fabrics sem Perdas". Este treinamento foi desenvolvido especificamente para profissionais de TI e engenheiros de rede comprometidos em dominar as complexidades das fabrics Ethernet sem perdas e é essencial para quem busca aprimorar suas habilidades na criação de redes de alta taxa de transferência e baixa latência, cruciais para aplicações de IA/ML. Ao compreender e implementar mecanismos avançados de qualidade de serviço (QoS), você poderá garantir que sua rede possa lidar com as tarefas mais exigentes sem comprometer o desempenho.

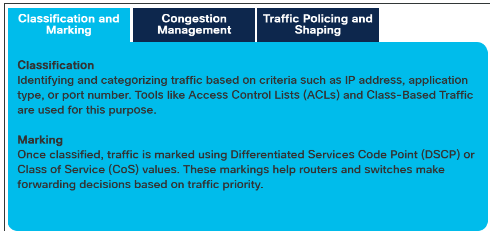
Ao longo deste treinamento, você explorará os principais componentes de uma malha Ethernet sem perdas, incluindo controle de fluxo prioritário (PFC), notificação explícita de congestionamento (ECN) e seleção aprimorada de transmissão (ETS). Você também examinará os recursos inovadores dos switches Cisco Nexus Série 9000, como gerenciamento inteligente de buffer e priorização dinâmica de pacotes (DPP). Ao final deste treinamento, você terá o conhecimento e a confiança necessários para construir e gerenciar uma rede que atenda aos mais altos padrões de desempenho e confiabilidade.

Imagine a satisfação de saber que você pode evitar perdas de pacotes, gerenciar o congestionamento da rede com eficácia e garantir que seus aplicativos de IA/ML funcionem de forma eficiente e sem problemas. Com as habilidades e o conhecimento adquiridos neste treinamento, você estará bem equipado para enfrentar os desafios das redes de data centers modernas, tornando-se um trunfo para qualquer organização.

Junte-se à jornada para dominar as estruturas Ethernet sem perdas e dê o primeiro passo para se tornar um especialista na construção de redes sem perdas de alto desempenho. Para aproveitar ao máximo este treinamento, sugerimos que você leia o texto a seguir, visualize os gráficos que o acompanham e assista aos vídeos. Em seguida, serão feitas perguntas de revisão para avaliar seu conhecimento sobre a construção de estruturas sem perdas antes de passar para a próxima seção

**Explore os mecanismos tradicionais do conjunto de ferramentas de QoS da Cisco**

Clique em cada mecanismo ou pressione os números de 1 a 3 no seu teclado para saber mais sobre cada um.



Interface gráfica do usuário, Texto

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

Sem QoS, um dispositivo oferece o serviço de melhor esforço para cada pacote, independentemente do conteúdo ou tamanho do pacote. O dispositivo envia os pacotes sem garantir confiabilidade, latência ou taxa de transferência. Com a configuração de QoS, você pode habilitar o tratamento de prioridades e reservar recursos para tráfego de aplicativos específicos. Uma configuração de QoS de ponta a ponta adequada oferece os seguintes benefícios:

* **Desempenho aprimorado:** aplicativos críticos recebem os recursos necessários, o que melhora o desempenho e a confiabilidade.
* **Experiência de usuário aprimorada:** a latência e a perda de pacotes são reduzidas para aplicativos em tempo real, como voz e vídeo.
* **Utilização eficiente da largura de banda:** o congestionamento da rede é evitado e faz melhor uso da largura de banda disponível.

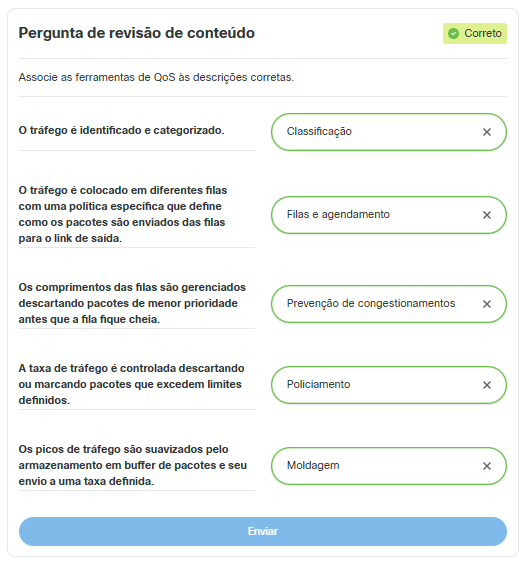
A implementação de mecanismos de QoS e o suporte a recursos específicos dependem da plataforma e do modelo. A figura a seguir apresenta uma visão geral da implementação dos recursos de QoS suportados para switches de data center Cisco Nexus 9000

Carta

O conteúdo gerado por IA pode estar incorreto.

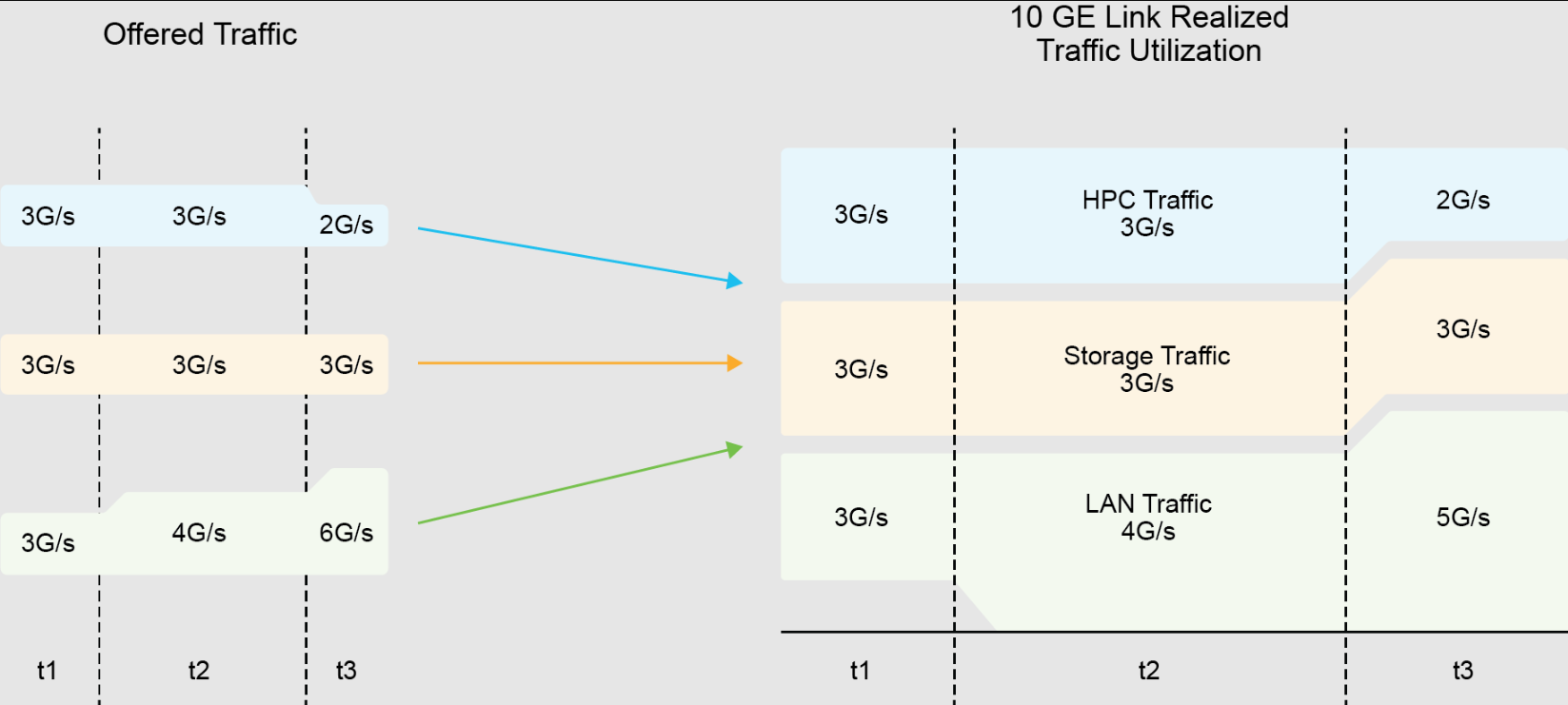
**Observação**

*O algoritmo de agendamento round robin ponderado por déficit (DWRR) garante a distribuição de largura de banda de acordo com um parâmetro de peso, que define a parte da largura de banda do link que está disponível para essa fila.*

**

**Seleção de transmissão aprimorada**

ETS é uma técnica de gerenciamento de largura de banda do conjunto Cisco Data Center Bridging (DCB) e está descrita no padrão IEEE 802.1Qaz. O ETS permite o gerenciamento ideal da largura de banda em diferentes links virtuais e prioriza classes de tráfego específicas, ao mesmo tempo em que fornece uma alocação justa de largura de banda para todas as classes de tráfego. O ETS gerencia a largura de banda permitindo acesso justo à largura de banda para todas as classes de tráfego, evitando congestionamentos e melhorando a eficiência geral da rede. Além disso, você pode fornecer largura de banda mínima garantida para determinadas classes de tráfego, como armazenamento ou computação de alto desempenho (HPC). Quando a carga de tráfego não utiliza totalmente a largura de banda alocada, a largura de banda restante fica disponível para outras classes, o que ajuda a acomodar classes com tráfego em rajadas.

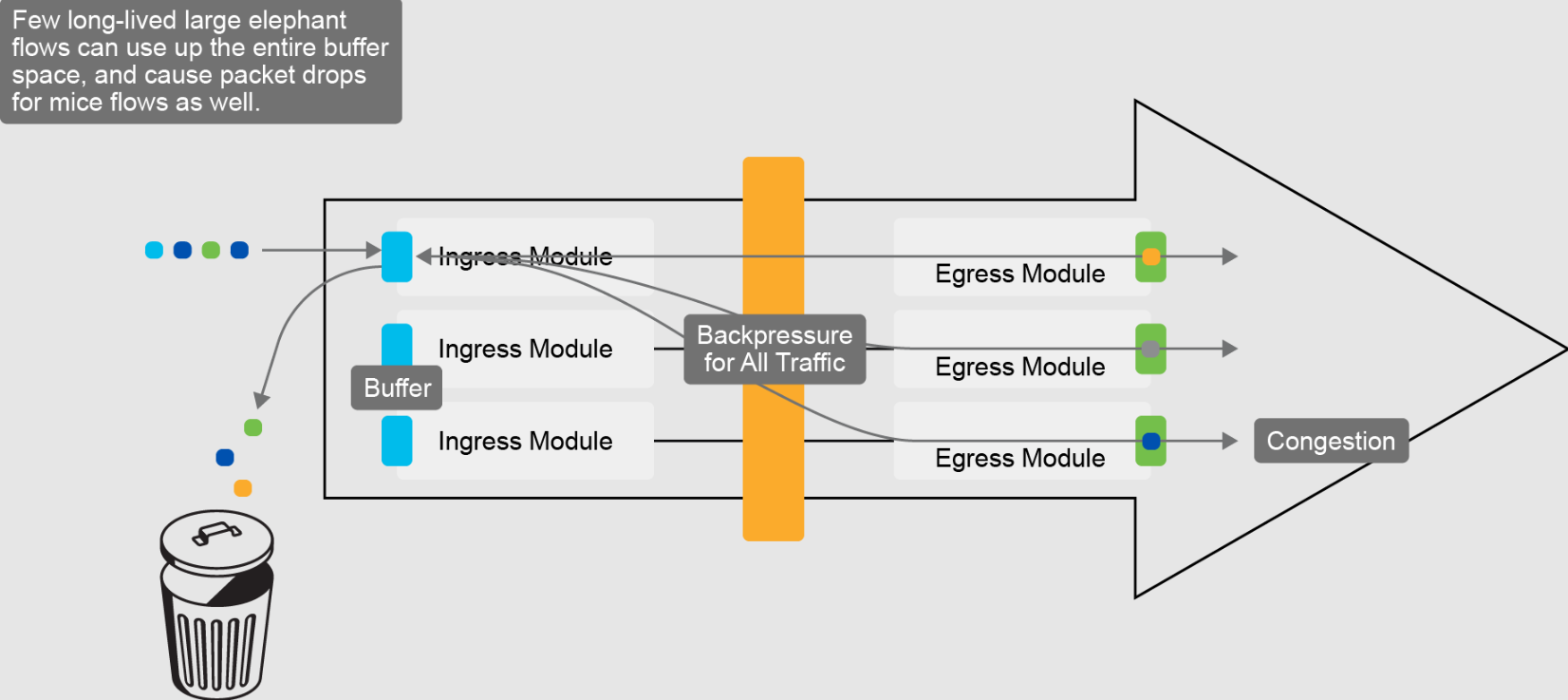


Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

**Gerenciamento Inteligente de Buffer em Switches Cisco Nexus Série 9000**

O tráfego moderno de data centers consiste em fluxos curtos, frequentemente chamados de fluxos de mouse, e fluxos longos, frequentemente chamados de fluxos de elefante. Em uma fila, os switches devem tratar os fluxos de mouse e elefante de forma diferente. No entanto, switches tradicionais de data centers com gerenciamento de buffer simples não conseguem diferenciar os fluxos de mouse e elefante. O escalonamento de filas do tipo "primeiro a entrar, primeiro a sair" (FIFO) não impede que os fluxos de mouse experimentem longa latência de enfileiramento durante congestionamentos ou quedas de cauda em caso de estouro de buffer.



Detecção Antecipada Aleatória (RED) e WRED são mecanismos ativos de gerenciamento de filas. Eles são usados ​​para evitar congestionamentos, utilizando descartes proativos de pacotes para acionar o controle de congestionamento TCP no remetente, o que, por sua vez, reduz a taxa de transmissão do tráfego. No entanto, RED e WRED descartam pacotes aleatoriamente em uma fila quando o comprimento da fila excede o limite de congestionamento. Esse processo também pode afetar fluxos de mensagens que transportam tráfego de controle importante, o que leva à degradação do desempenho.

**Observação**

*A distribuição dos comprimentos de fluxo em redes de data centers é tal que alguns fluxos elefante transportam a maior parte do tráfego de dados. Os fluxos ratos adicionam picos à carga geral de tráfego. Os fluxos elefante e ratos executam tarefas de aplicação diferentes e, portanto, têm requisitos diferentes de largura de banda, perda de pacotes e latência de rede.*

Os switches tradicionais precisam armazenar em buffer todos os pacotes em um link congestionado para evitar perdas de pacotes em fluxos de mouse, pois não conseguem armazenar em buffer seletivamente esses fluxos. Por esse motivo, algumas plataformas de switches mais recentes são construídas com buffers profundos para ajudar a garantir amplo espaço de buffer. No entanto, essa abordagem apresenta desvantagens significativas:

* O custo de um switch com maior espaço de buffer é significativamente mais alto.
* A profundidade da fila estendida causa latência de enfileiramento mais longa. A latência aumentará o tempo de conclusão de todos os fluxos de dados, degradando assim o desempenho geral do aplicativo.
* **Observação**
* O TCP tenta consumir a máxima largura de banda disponível em um caminho de tráfego ponta a ponta, o que pode consumir todos os buffers dos enlaces. A menos que o ECN esteja habilitado, o TCP usa quedas de pacotes como notificações implícitas de congestionamento para acionar seu mecanismo de controle de congestionamento no remetente e reduzir a transmissão de pacotes. Se nenhuma queda de pacote for detectada, um remetente TCP pode aumentar continuamente sua taxa de transmissão, o que pode eventualmente causar estouro de buffer no enlace congestionado. Quando essa situação ocorre, o novo fluxo de mensagens que chega à fila estará sujeito à queda de cauda

Os switches Cisco Nexus Série 9000 com ASICs Cisco em escala de nuvem são construídos com uma quantidade moderada de espaço de buffer no chip para atingir a máxima taxa de transferência e funções inteligentes de gerenciamento de buffer para lidar com fluxos de mouse e elefante de forma eficiente. O conceito de gerenciamento de buffer inteligente e inovador da Cisco é fundamental e se baseia na capacidade de distinguir fluxos de mouse e elefante e aplicar diferentes técnicas de gerenciamento de filas com base nos requisitos de encaminhamento de rede em caso de congestionamento de link. Esse processo permite que os fluxos de elefante e de ratos alcancem o melhor desempenho, aprimorando o desempenho geral da aplicação.

Diagrama

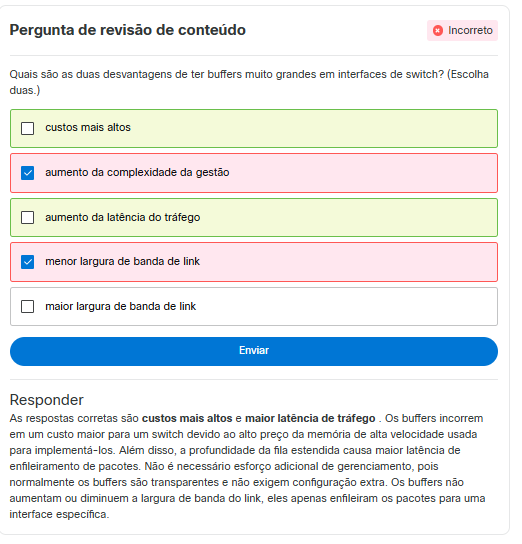
O conteúdo gerado por IA pode estar incorreto.

**Observação**

*Os recursos de gerenciamento de buffer inteligente do Cisco Nexus permitem que você tenha buffers inteligentes e de tamanho apropriado.*

Os recursos de gerenciamento inteligente de buffer são incorporados aos ASICs em escala de nuvem da Cisco para desempenho acelerado por hardware e incluem:

* Queda justa aproximada (AFD) com armadilha para elefantes (ETRAP). A AFD preserva o espaço de buffer para absorver fluxos de ratos, particularmente microbursts e fluxos de ratos agregados, limitando o uso de buffer por fluxos de elefantes com uso intensivo de largura de banda. Também garante a equidade de largura de banda entre os fluxos de elefantes. A ETRAP é uma técnica que detecta fluxos de elefantes.
* O DPP permite a separação dos fluxos de mouse e elefante em duas filas diferentes, permitindo a distribuição independente do espaço de buffer e a aplicação de diferentes agendamentos de fila. Por exemplo, os fluxos de mouse podem ser mapeados para uma fila de baixa latência, enquanto os fluxos de elefante podem ser enviados para uma fila justa ponderada.
* **Observação**
* O AFD e o DPP podem ser implantados separadamente ou em conjunto. Eles se complementam para oferecer o melhor desempenho da aplicação. Além disso, um algoritmo de admissão de buffer chamado Dynamic Buffer Protection (DBP) garante acesso justo ao buffer disponível para todas as filas.



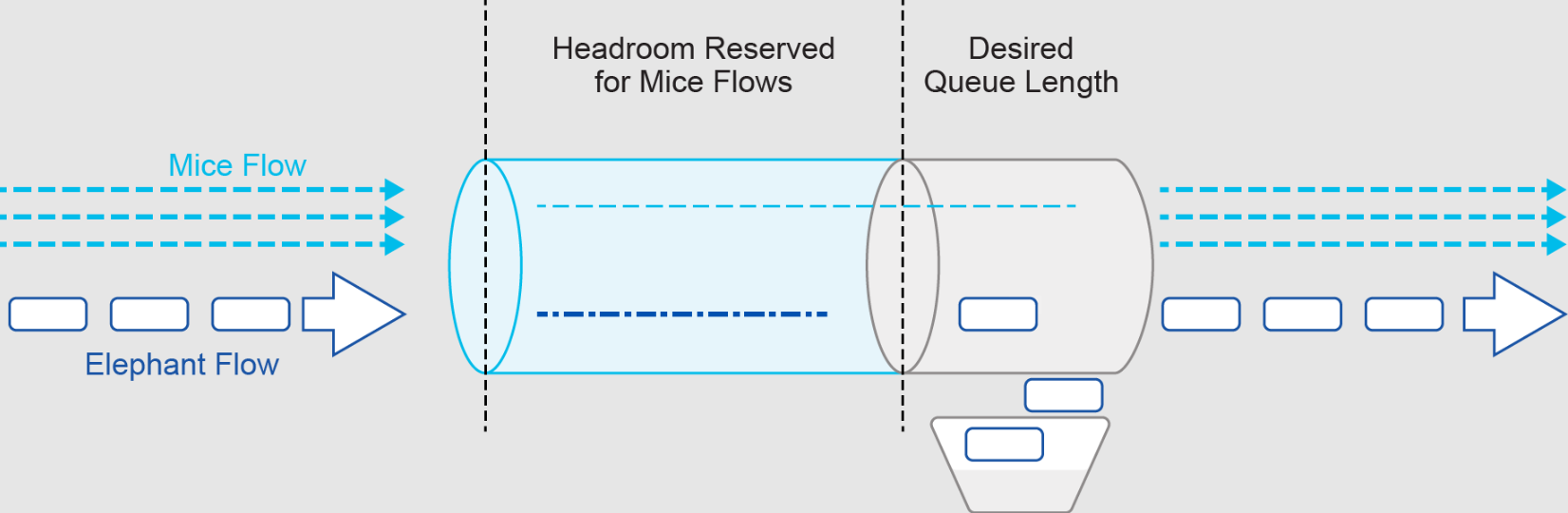
**AFD com ETRAP**

O AFD é um esquema de gerenciamento de filas ativo que garante alocação justa de largura de banda entre fluxos.

A justiça tem dois aspectos:

* O AFD preserva o espaço do buffer para lidar com fluxos de mouse, limitando o uso do buffer por fluxos de elefante que consomem muita largura de banda. Os fluxos de elefante geralmente têm uma duração de sessão muito mais longa do que os fluxos de mouse.
* O AFD rastreia fluxos de elefantes e aplica o algoritmo AFD na fila de saída para alocar uma parcela justa de largura de banda a eles.

O AFD utiliza o ETRAP para distinguir fluxos de elefantes de longa duração de fluxos de ratos de curta duração. Os fluxos de ratos são excluídos do algoritmo de descarte, de modo que recebem uma parcela justa da largura de banda, e os fluxos de elefantes com uso intensivo de largura de banda não consomem toda a largura de banda. Este conceito é ilustrado na figura a seguir.



AFD com ETRAP usa estes princípios:

* Um fluxo pode ser definido usando vários parâmetros, mas normalmente é usada a tupla quíntupla (endereço IP de origem e destino, número da porta de origem e destino, protocolo de transporte).
* O ETRAP opera no lado de entrada de um switch. Ele mede a contagem de bytes dos fluxos de entrada e compara essa contagem com o limite do ETRAP.
* Fluxos com uma contagem de bytes menor que o limite são fluxos de mouse.
* Um fluxo com uma contagem de bytes maior que o limite ETRAP se torna um fluxo elefante e é movido para a tabela elefante para rastreamento.

Você pode configurar o limite ETRAP para definir o que é um fluxo elefante no seu ambiente de data center.

A tabela Elephant armazena e rastreia a taxa de chegada e a atividade do fluxo Elephant. As taxas de dados medidas são passadas para o mecanismo de gerenciamento de buffer nas filas de saída, onde o algoritmo AFD usa as taxas para calcular a probabilidade de quedas para cada fluxo. Os fluxos Elephant são desativados por tempo limite se não permanecerem em um determinado nível de atividade. Quando a largura de banda média de um fluxo Elephant durante o período for inferior ao limite de largura de banda configurado, ele será considerado inativo, desativado por tempo limite e removido da tabela Elephant Flow.

**Observação**

*Um temporizador de período de idade configurado pelo usuário e um limite de largura de banda são usados ​​para avaliar a atividade de um fluxo de elefantes.*

Antes do AFD, o WRED era a principal tecnologia usada para os mesmos propósitos ou propósitos semelhantes. O WRED aplica descarte antecipado aleatório ponderado a filas baseadas em classes, mas não possui reconhecimento de fluxo dentro de uma classe. O WRED tem as seguintes desvantagens:

* Todos os pacotes, incluindo fluxos de mouse sensíveis à perda de pacotes, estão sujeitos à mesma probabilidade de queda.
* Embora os fluxos de elefantes possam usar gotas como sinais de congestionamento para diminuir a taxa de envio de tráfego, as gotas podem afetar negativamente os fluxos de ratos.
* A mesma probabilidade de queda faz com que fluxos elefantes com uma taxa mais alta (devido ao curto tempo de ida e volta [RTT]) obtenham mais largura de banda. Portanto, a largura de banda de saída não é dividida igualmente entre os fluxos elefantes que atravessam o mesmo link congestionado.

Como resultado, o tempo de conclusão do fluxo para fluxos de mice aumenta, e os fluxos de elephant não têm acesso justo à largura de banda do link e aos recursos de buffer. Em contraste, o AFD considera os tamanhos dos fluxos e as taxas de chegada de dados antes de tomar uma decisão de descarte. O algoritmo de descarte é projetado para proteger os fluxos de mice e garantir a equidade entre os fluxos de elephant.

O AFD não distingue entre tipos de protocolos de transporte, como TCP e UDP. Ao contrário do TCP, o UDP não possui um algoritmo de gerenciamento de congestionamento integrado. O AFD com marcação ECN só conseguirá evitar congestionamento se a aplicação baseada em UDP estiver ciente de ECN, o que geralmente não é o caso. Portanto, o AFD não deve ser habilitado em filas que transportam tráfego não TCP

**Observação**

*Como o AFD é configurável por fila, é melhor classificar o tráfego por protocolo e garantir que o tráfego de aplicativos baseados em UDP de alta largura de banda sempre use uma fila não habilitada para AFD.*

*Interface gráfica do usuário, Texto, Aplicativo

O conteúdo gerado por IA pode estar incorreto.*

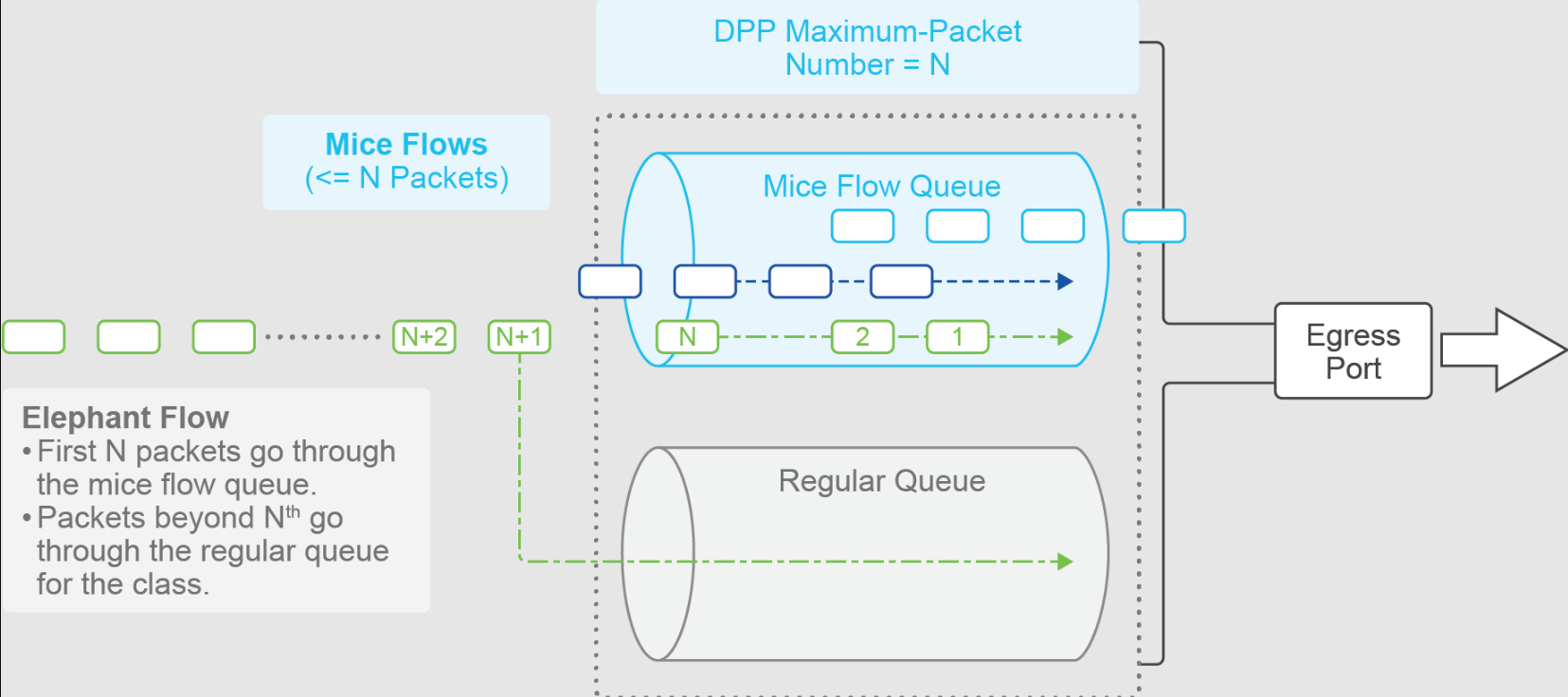
**Priorização dinâmica de pacotes**

O DPP enfileira fluxos de ratos e fluxos de elefantes na mesma classe de tráfego separadamente. Ele os isola em duas filas separadas, mesmo que pertençam à mesma classe de tráfego. Essa separação é impossível em tecnologias tradicionais de gerenciamento de filas, pois elas não têm reconhecimento de fluxo dentro de uma classe.

O DPP rastreia fluxos para distinguir fluxos de ratos de fluxos de elefantes na porta de entrada e, em seguida, aplica políticas de enfileiramento separadas para fluxos de ratos e elefantes na porta de saída. O DPP opera com base nos seguintes princípios:

* Na porta de entrada, os fluxos são identificados com base em suas cinco tuplas, e seus contadores de pacotes iniciais são monitorados.
* Os pacotes em um fluxo são então classificados no grupo QoS de mouse ou no grupo QoS original com base no tamanho do fluxo.
* Dentro do mecanismo DPP, um contador de pacotes máximos é usado para determinar a classificação.
* Se o contador máximo de pacotes do DPP for definido como N, os primeiros N pacotes em cada fluxo serão classificados no grupo QoS de ratos.
* O pacote N+1 e os seguintes são classificados no grupo QoS original da classe de tráfego à qual o fluxo pertence.

Essencialmente, um novo fluxo sempre começa como um fluxo de ratos até enviar mais pacotes do que o valor do contador máximo de pacotes do DPP. A figura a seguir ilustra como o mecanismo do DPP distingue os fluxos de ratos dos fluxos de elefantes.



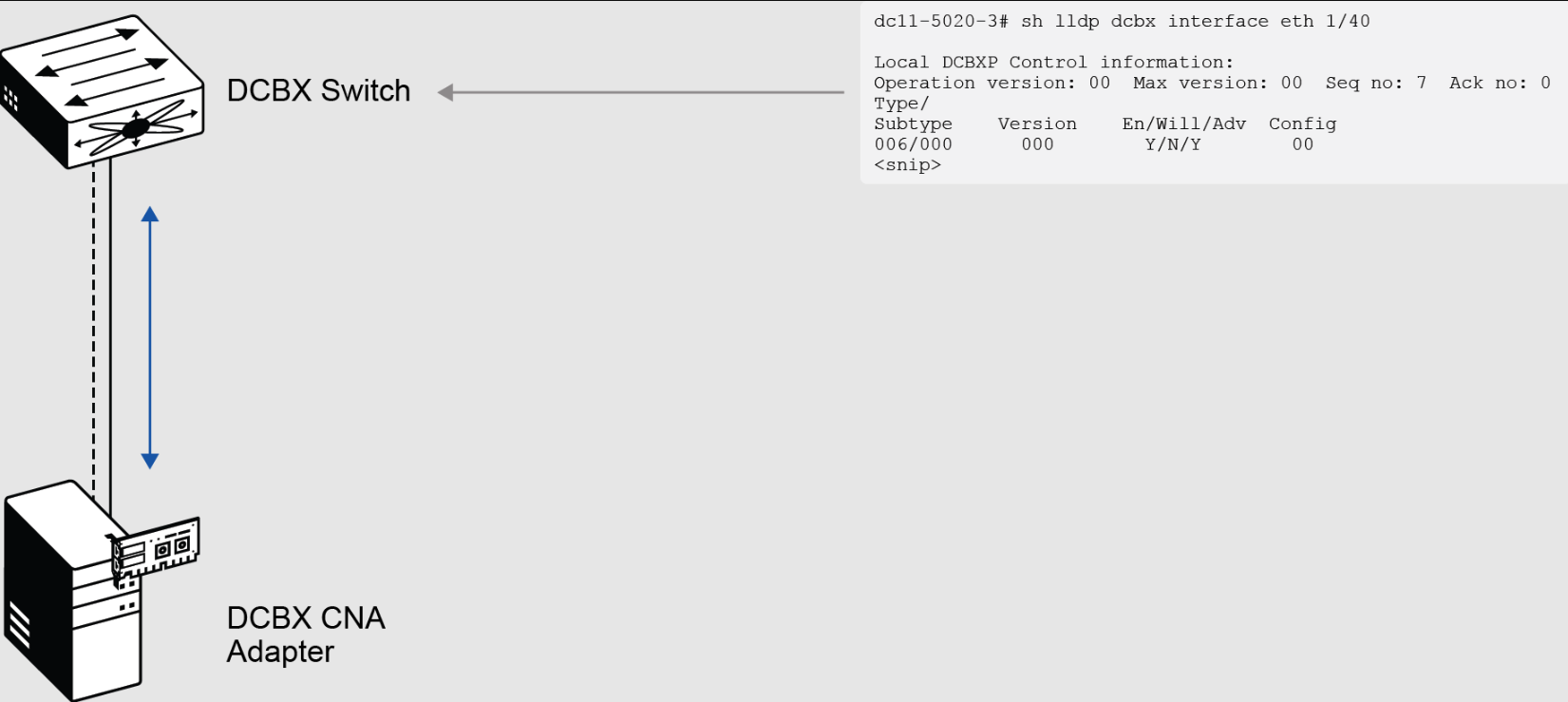
Após o DPP diferenciar os fluxos de mouse e elefante em grupos de QoS separados na porta de entrada, diferentes controles de enfileiramento podem ser aplicados a eles na porta de saída para atender aos diferentes requisitos de encaminhamento para fluxos de mouse e elefante. A fila de fluxo de mouse deve ser uma fila prioritária, enquanto a fila regular para fluxos de elefante é uma fila ponderada com largura de banda alocada para esses fluxos.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

**Troca de ponte de data center**

O protocolo Data Center Bridging Exchange (DCBX) pode descobrir e trocar informações de prioridade e largura de banda entre pontos de extremidade. O DCBX simplifica o gerenciamento, permitindo a configuração e a distribuição de parâmetros de um nó para outro, garantindo que ambas as extremidades de um link tenham uma configuração consistente.



O DCBX é implementado como uma extensão do Protocolo de Dados da Camada de Enlace (LLDP) com novos campos de tipo, comprimento e valor (TLV). Os seguintes parâmetros dos recursos Ethernet do data center podem ser trocados e sincronizados entre os dois nós:

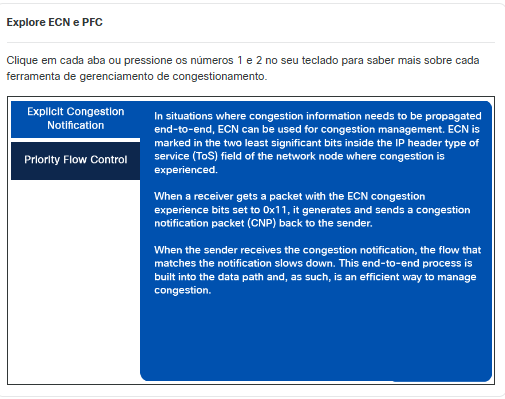
* ETS
* PFC
* Valores de CoS
* Notificação de congestionamento
* Link lógico inativo
* **Observação**
* TLV é um esquema de codificação de dados genérico e flexível, usado em protocolos de comunicação para estender facilmente recursos específicos de protocolos. Consiste em três partes: tipo, comprimento e valor.

**Interface gráfica do usuário, Texto, Aplicativo

O conteúdo gerado por IA pode estar incorreto.**

**Estrutura Ethernet sem perdas usando RoCEv2**

Para transporte RDMA sobre Ethernet Convergente versão 2 (RoCEv2), a rede deve fornecer alta taxa de transferência e baixa latência, evitando quedas de tráfego em situações de congestionamento. Como as quedas de pacotes tornam o treinamento de aplicações de IA/ML mais lento, o RoCEv2 requer transporte sem perdas. A rede sem perdas pode ser alcançada usando algoritmos de prevenção de congestionamento ECN e PFC. Os switches Cisco Nexus Série 9000 suportam gerenciamento de congestionamento PFC e marcação ECN com WRED ou AFD para indicar congestionamento no nó da rede.



Interface gráfica do usuário, Texto, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

**Crie redes Ethernet sem perdas com ECN e PFC**

Tanto o ECN quanto o PFC conseguem gerenciar o congestionamento muito bem sozinhos. Trabalhando juntos, podem ser ainda mais eficazes. O ECN pode reagir primeiro para mitigar o congestionamento. Se o ECN não reagir com rapidez suficiente e a utilização do buffer continuar a aumentar, o PFC atua como um sistema de segurança contra falhas, prevenindo quedas de tráfego. Essa técnica é a maneira mais eficiente de gerenciar o congestionamento e construir redes Ethernet sem perdas. Esse processo colaborativo entre o PFC e o ECN é chamado de Notificação Quantizada de Congestionamento em Data Center (DCQCN) e foi desenvolvido para redes RoCE.

Juntos, PFC e ECN fornecem gerenciamento de congestionamento eficiente de ponta a ponta:

* Quando o sistema sofre congestionamento leve com uso moderado de buffer, o WRED com ECN gerencia o congestionamento perfeitamente.
* Em casos em que o congestionamento é mais grave ou causado por microbursts que produzem alto uso de buffer, o PFC é acionado para gerenciar o congestionamento.

**Observação**

*ECN e PFC devem ser configurados de ponta a ponta em toda a rede do data center.*

*Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.*

**Gerenciamento avançado de congestionamento com AFD**

A abordagem recomendada para redes RoCE é usar ECN e PFC juntos. Você também pode aprimorar o gerenciamento de congestionamento usando algoritmos avançados de QoS, como o AFD, em switches Cisco Nexus Série 9000 para distinguir fluxos de alta largura de banda (fluxos elefantes) de fluxos de curta duração e baixa largura de banda (fluxos ratos).

O AFD pode ser combinado com o ECN como alternativa ao WRED. Por padrão, o AFD aplica descartes antecipados de pacotes aos fluxos "elephant" com base na possibilidade de descarte, calculada pelo algoritmo do AFD. Os descartes de pacotes servem como notificações implícitas de congestionamento para os hosts de origem. Se você deseja evitar descartes de pacotes ou está lidando com UDP (que não causa lentidão), o AFD pode trabalhar com o ECN para enviar notificações sem descartar pacotes. Essa colaboração entre AFD e ECN funciona da seguinte forma:

* Com base na possibilidade de marcação calculada pelo algoritmo AFD, o switch pode marcar o bit de congestionamento ECN (CE) nos pacotes.
* Essa marcação é feita apenas para fluxos de alta largura de banda. Um número diferente de pacotes é marcado com ECN com base na largura de banda utilizada pelo fluxo.
* Depois disso, os pacotes continuarão a ser transmitidos em direção ao host de destino em vez de serem descartados.
* O host de destino retornará essa indicação de congestionamento ao remetente gerando um CNP.
* Quando o host de origem recebe o CNP, sua janela de transmissão será reduzida.

Esse processo permite que o AFD e o ECN forneçam as funções pretendidas, incluindo alocação justa de largura de banda e mitigação de congestionamento de rede, o que otimiza o desempenho para a menor latência sem descartar pacotes.

O WRED marca todo o tráfego em uma fila igualmente. A vantagem do AFD em relação ao WRED é sua capacidade de distinguir e reduzir a velocidade dos fluxos que causam mais congestionamento. O AFD é mais granular e marca apenas os fluxos "elefante" com maior largura de banda, deixando os fluxos "rato" sem marcação. Essa abordagem garante que os fluxos "rato" não sejam penalizados, o que os tornaria mais lentos

**Observação**

*Em um cluster de IA, é melhor deixar as comunicações de curta duração serem concluídas, evitando que uma longa transferência de dados e qualquer congestionamento resultante as tornem mais lentas. Perdas de pacotes ainda são evitadas, e muitas transações são concluídas mais rapidamente porque o sistema consegue diferenciar os fluxos de dados "elephant" e apenas os torna mais lentos.*

*Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.*

**Resumo**

Agora você deve ter um conhecimento sólido de técnicas de rede avançadas para construir estruturas RoCE de alto desempenho e sem perdas, além de ferramentas e recursos de QoS que aumentam a eficiência e a confiabilidade das estruturas RoCE.

Agora que você concluiu o treinamento, reflita sobre as seguintes perguntas:

* Como você pode implementar mecanismos de QoS em sua infraestrutura de rede atual para priorizar aplicativos críticos e melhorar o desempenho geral?
* Que estratégias você usaria para gerenciar o congestionamento em um ambiente de data center que lida com fluxos de tráfego de curta duração (mice) e longa duração (elephant)?
* Como a combinação de ECN e PFC pode ser usada em sua rede para criar uma estrutura Ethernet sem perdas e quais são os benefícios potenciais para seu caso de uso específico?