**Visibilidade de congestionamento**

**Introdução**

Bem-vindo ao treinamento de Visibilidade de Congestionamento, um passo vital para dominar o gerenciamento de redes de alto desempenho. No mundo digital de hoje, o gerenciamento eficiente de congestionamento é essencial. Você sabia que o congestionamento de rede pode causar uma perda de 20% na produtividade das empresas? Este treinamento visa equipá-lo com as habilidades necessárias para prevenir essas perdas e garantir que sua rede opere com a máxima eficiência.

Imagine os aplicativos de IA da sua empresa paralisados ​​devido ao congestionamento da rede. A frustração e as perdas financeiras podem ser significativas. Este treinamento ensinará como evitar tais cenários usando mecanismos de notificação explícita de congestionamento (ECN) e controle de fluxo prioritário (PFC). Essas ferramentas são cruciais para a construção de uma rede Ethernet sem perdas e fornecem a visibilidade necessária para identificar e lidar com pontos críticos de congestionamento.

Seja você um administrador de rede, um profissional de TI ou um especialista em inteligência artificial e aprendizado de máquina (IA/ML), entender a visibilidade do congestionamento é crucial. O Cisco Nexus 9000 Series, com seus recursos avançados de telemetria, oferece suporte incomparável para a manutenção de uma rede de alto desempenho. Este treinamento o guiará pelas complexidades de ECN e PFC e como monitorar métricas relacionadas ao congestionamento usando o Cisco Nexus Dashboard Insights. Ao final deste treinamento, você estará bem equipado para otimizar sua rede para aplicações de IA/ML e garantir um fluxo de dados contínuo e tempo de inatividade mínimo.

**Notificação explícita de congestionamento**

ECN e PFC desempenham papéis importantes na construção de uma rede Ethernet sem perdas e de alto desempenho. Neste tópico, você explorará como o mecanismo ECN funciona.

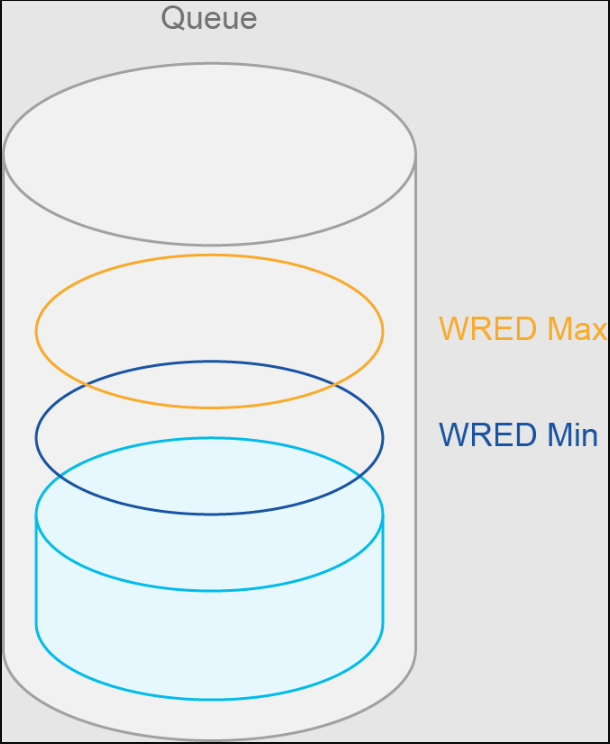
O recurso ECN é usado entre dois endpoints habilitados para ECN para sinalizar possível congestionamento, de modo que o remetente reduza a taxa de transmissão do tráfego. Um nó de rede com capacidade para ECN utiliza um algoritmo de prevenção de congestionamento para verificar a quantidade de fila utilizada. Após atingir um limite especificado, o tráfego que contribui para o congestionamento será marcado. Neste exemplo, a detecção antecipada aleatória ponderada (WRED) indica congestionamento e marca o tráfego com bits ECN.

**Observação**

*Em vez de WRED, a técnica de queda justa aproximada (AFD) pode ser usada para permitir granularidade em nível de fluxo.*

Os switches Cisco Nexus Série 9000 podem marcar pacotes com bits ECN em caso de congestionamento de rede. O WRED nos switches Cisco Nexus Série 9000 é feito por fila. Na fila, dois limites são definidos. O limite mínimo do WRED é definido como um nível mais baixo de utilização do buffer e indica um evento de congestionamento menor. Se a utilização do buffer atingir o limite mínimo, o WRED marca vários pacotes de saída que saem da fila. O número de pacotes depende do valor da probabilidade de queda na configuração do WRED no switch Cisco Nexus Série 9000, que é representado como uma porcentagem de todos os pacotes de saída.

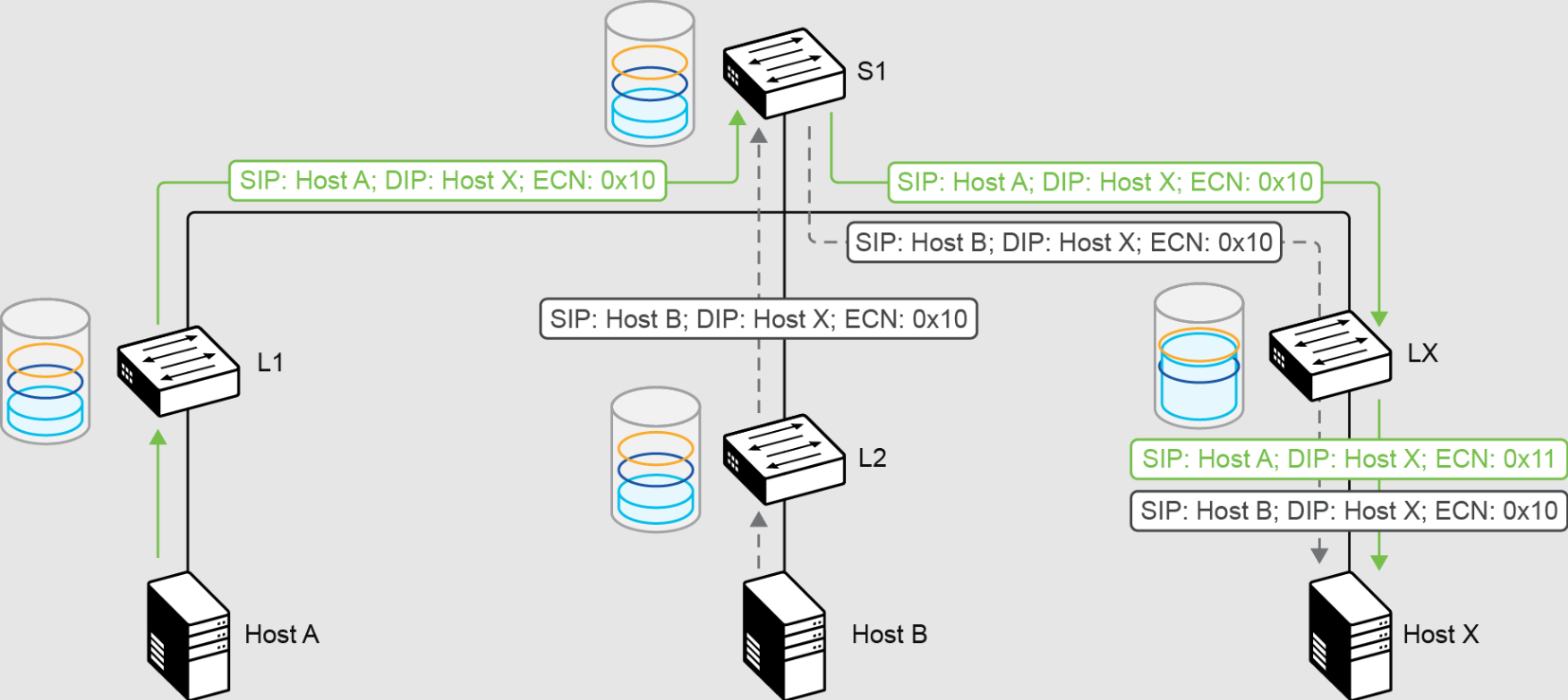
Por exemplo, se os parâmetros de probabilidade de queda forem definidos como 10, isso significa que 10% de todos os pacotes de saída serão marcados. Se essa ação não reduzir o congestionamento, a utilização do buffer da fila continuará a crescer e atingirá o limite máximo do WRED. Após ultrapassar o limite máximo do WRED, o switch marca bits ECN em cada pacote de saída.



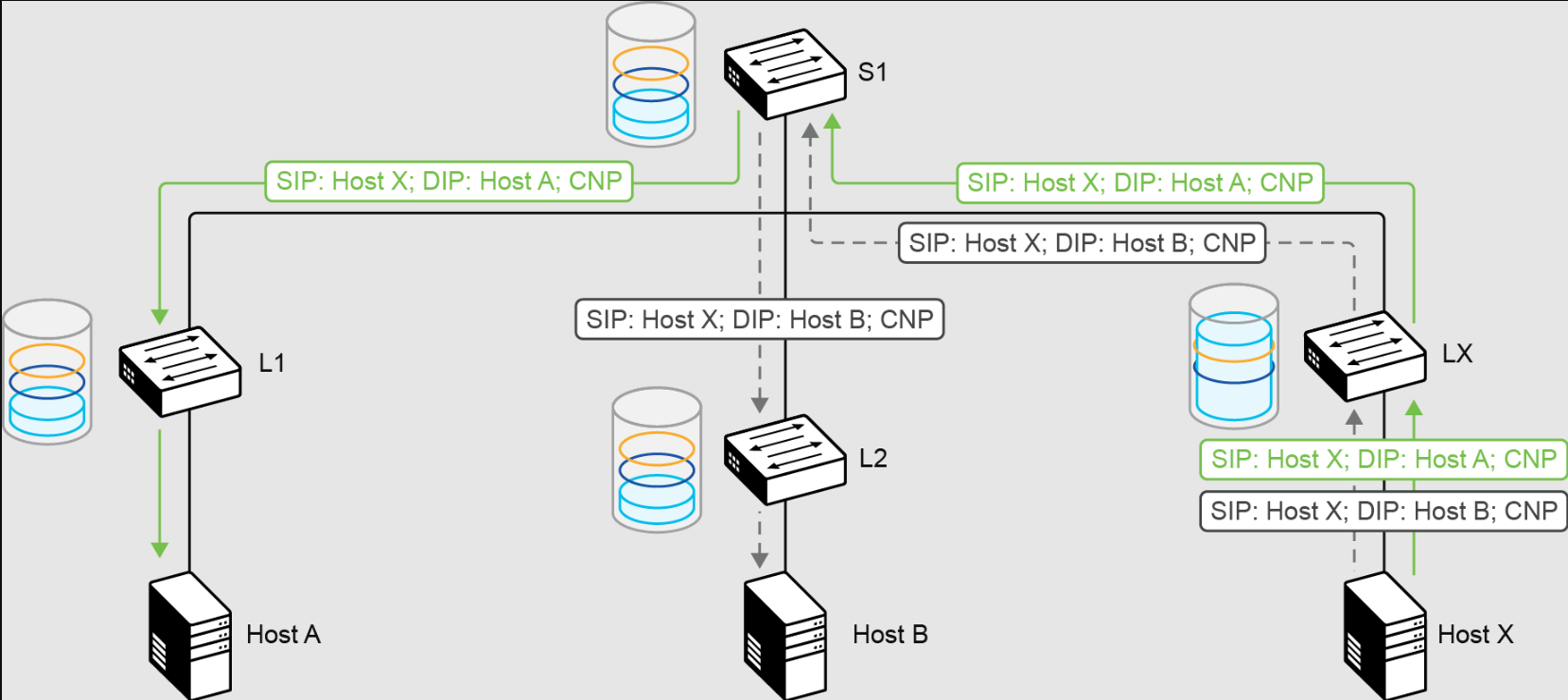
O ECN é marcado no nó da rede onde ocorre o congestionamento, utilizando os dois bits menos significativos do campo Tipo de Serviço (ToS) no cabeçalho IP. Quando um receptor recebe um pacote com os bits de experiência de congestionamento do ECN definidos como 0x11, ele gera e envia um pacote de notificação de congestionamento (CNP) de volta ao remetente.

| Bits ECN | Comportamento ECN |
| --- | --- |
| 0x00 | Não compatível com ECN |
| 0x10 | Transporte com capacidade ECN (0) |
| 0x01 | Transporte com capacidade ECN (1) |
| 0x11 | Congestionamento encontrado |

Este mecanismo WRED ECN informa os endpoints sobre congestionamento. Os endpoints podem usar essas informações para reduzir a taxa de envio de tráfego. No exemplo a seguir, você tem uma rede de duas camadas. O Host A e o Host B estão enviando dados para o Host X.

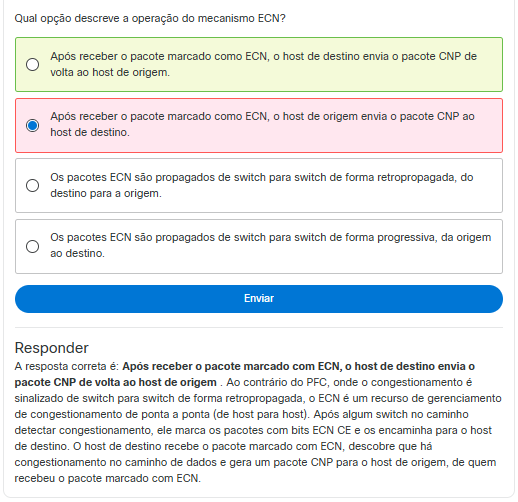


Ocorre congestionamento na Folha X, pois há mais largura de banda no link spine-leaf do que no link leaf-host, e a porta conectada ao Host X está sobrecarregada. A utilização do buffer começará a aumentar na Folha X. Ao atingir o limite mínimo de WRED, a Folha X começa a marcar os bits ECN de vários pacotes com um valor 0x11 para indicar congestionamento no caminho de dados.



Quando o destino (Host X) recebe o pacote marcado, ele entende que há congestionamento no caminho de dados e envia um pacote CNP ao host de origem. O pacote CNP informa o host de origem sobre o congestionamento para que ele possa reduzir a taxa de transmissão do tráfego.

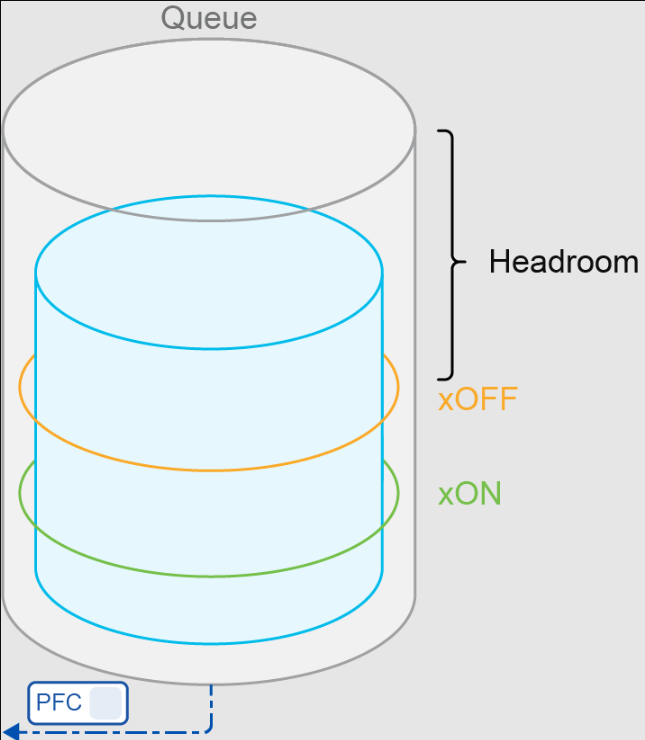
Como apenas alguns pacotes foram marcados com bits de congestionamento, a origem reduz a taxa de transferência de tráfego para esse fluxo e continua enviando pacotes. Se o congestionamento continuar e a utilização do buffer ultrapassar o limite máximo do WRED, o switch começa a marcar cada pacote com bits ECN. O remetente começa a receber muitos pacotes CNP e, com base em seu algoritmo, deve reduzir drasticamente a taxa de transmissão de dados para o destino. Essa ação atenua o congestionamento e a utilização do buffer deve começar a diminuir. Após algum tempo, o remetente aumentará gradualmente a taxa de tráfego até que o congestionamento ocorra novamente e todo o processo se repita.



**Controle de Fluxo Prioritário**

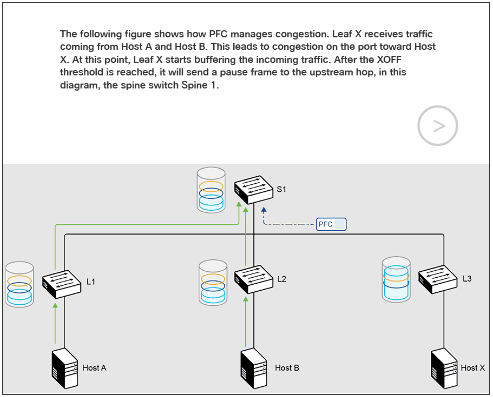
O PFC é um mecanismo que previne a perda de pacotes em caso de congestionamento. Quando o PFC está habilitado no switch Cisco Nexus Série 9000, uma classe de serviço (CoS) é reservada para transporte sem perdas. O tráfego nessa classe é tratado de forma diferente do tráfego em outras classes. Essa abordagem permite o controle de fluxo por prioridade, permitindo simultaneamente prioridades de tráfego sem perdas e com perdas no link. Qualquer porta no switch Cisco Nexus Série 9000 configurada com PFC recebe uma fila dedicada de no-drop e um buffer dedicado para essa fila.

Buffers de tamanho adequado são importantes para o funcionamento ideal de uma rede sem perdas. A reserva de buffer é necessária para acomodar os pacotes atualmente na rede e as flutuações no tráfego de rede, como microbursts. Para fornecer recursos sem perdas, a fila possui dois limites. O limite XOFF é definido em um nível mais alto de utilização do buffer, onde um quadro PFC é gerado e enviado para a origem do tráfego. Quando a utilização do buffer está abaixo do limite XON, os quadros de pausa são interrompidos e não são mais enviados para os remetentes. Nesse ponto, não há mais congestionamento.



**Explore o PFC Gerenciando o Congestionamento**

Clique nos botões próximo ou anterior para navegar pelo processo de gerenciamento de congestionamento do PFC.



Diagrama

O conteúdo gerado por IA pode estar incorreto.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

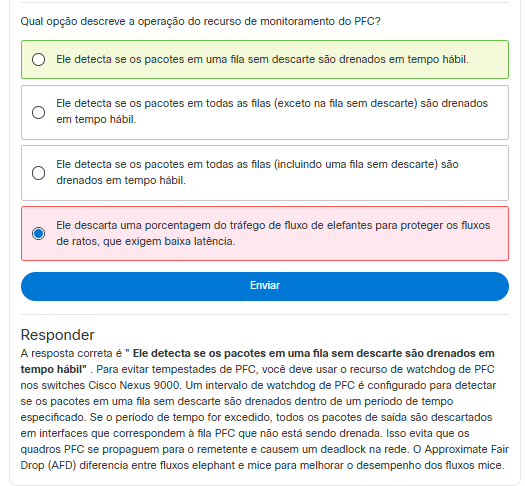
Diagrama

O conteúdo gerado por IA pode estar incorreto.

Em raras ocasiões, uma tempestade de PFC pode ocorrer quando um host com defeito transmite continuamente quadros PFC. Esse comportamento pode causar o estouro dos buffers em todos os nós da rede e, após atingir todos os hosts finais, pode interromper completamente a rede. O recurso de monitoramento de PFC nos switches Cisco Nexus série 9000 ajudará você a evitar tempestades de PFC. Um intervalo de monitoramento de PFC é configurado para detectar se os pacotes em uma fila sem descarte são drenados dentro de um período de tempo especificado. Se o período de tempo for excedido e os pacotes não forem drenados, todos os pacotes de saída serão descartados nas interfaces que correspondem à fila PFC que não está sendo drenada. Essa abordagem impede que os quadros PFC contínuos cheguem ao remetente e causem uma paralisação da rede.

Diagrama

O conteúdo gerado por IA pode estar incorreto.



**Visibilidade de congestionamento em redes de cluster de IA/ML usando o Cisco Nexus Dashboard Insights**

Os switches Cisco Nexus Série 9000 possuem os recursos de hardware e software necessários para fornecer mecanismos de gerenciamento de latência e congestionamento para atender aos requisitos de aplicações de IA/ML. Eles também contam com recursos avançados de monitoramento e telemetria para permitir visibilidade do desempenho da rede. Os recursos e funcionalidades do Cisco Nexus Dashboard Insights ajudam você a monitorar e otimizar o congestionamento e o desempenho.

**Insights do painel do Cisco Nexus**

Para ajudar os administradores de rede a otimizar suas redes de IA/ML para obter o melhor desempenho e prever problemas antes que eles afetem o serviço, o sistema de rede deve fornecer um nível profundo de visibilidade dos algoritmos de gerenciamento de congestionamento e da integridade geral do sistema de rede. Os switches Cisco Nexus Série 9000 contam com poderosos recursos de telemetria integrados que podem ser usados ​​para correlacionar problemas na rede e ajudar a otimizá-la para o transporte Remote Direct Memory Access (RDMA) sobre Ethernet Convergente (RoCE) versão 2 (RoCEv2).

O poder da telemetria e dos insights permite monitorar e otimizar estruturas de IA/ML em tempo real. Com o Cisco Nexus Dashboard Insights, você pode monitorar o seguinte:

* Desempenho do tecido (latência, utilização, descarte de pacotes)
* Estatísticas Ethernet sem perdas (ECN, PFC)
* Pontuação de congestionamento
* Pontuação de latência
* Execute solução de problemas em tempo real

Os switches Cisco Nexus Série 9000 fornecem informações de telemetria de fluxo de hardware usando uma tabela de fluxo e eventos de tabela de fluxo. A tabela de fluxo coleta informações completas de fluxo e metadados, como os seguintes:

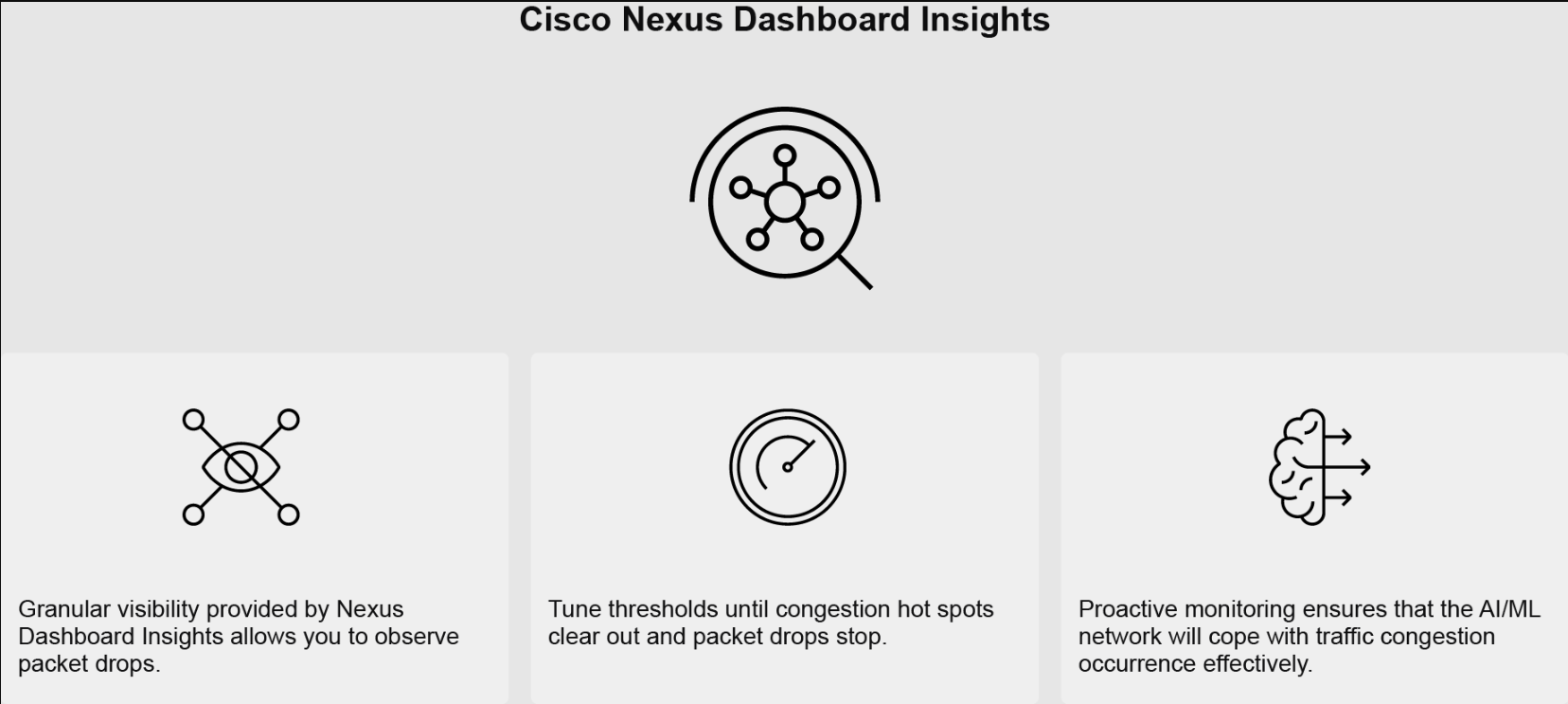
* Informações de fluxo de 5 tuplas
* Informações de interface e fila
* O fluxo inicia e para o tempo
* Indicadores de queda de pacotes
* Medição de explosão

Com esses recursos, cada pacote que passa pelo switch pode ser contabilizado, observado e correlacionado com comportamentos como microbursts ou descartes de pacotes. Você pode exportar esses dados para o Cisco Nexus Dashboard Insights e exibir os dados por dispositivo, por interface e em um nível de fluxo.

**Monitoramento de congestionamento e desempenho com o Cisco Nexus Dashboard Insights**

O Cisco Nexus Dashboard Insights utiliza um conjunto de algoritmos avançados de alerta, linha de base, correlação e previsão para fornecer insights aprofundados sobre o comportamento da rede, utilizando dados de telemetria obtidos de componentes de rede e computação. O Cisco Nexus Dashboard Insights automatiza a solução de problemas e auxilia na análise rápida da causa raiz e na correção antecipada. Além disso, simplifica as auditorias e garante a conformidade, utilizando um repositório de rede unificado e conformidade.

O Cisco Nexus Dashboard Insights monitora as propriedades da malha Ethernet sem perdas e fornece contadores de marcas ECN por dispositivo e interface, e em nível de fluxo. Ele também pode reportar informações sobre pacotes PFC emitidos ou recebidos por um switch em nível de CoS. Com essas informações, um administrador de rede pode observar estatísticas de congestionamento da rede em tempo real e usá-las para ajustar a rede e responder melhor ao congestionamento.



Com a visibilidade granular fornecida pelo Cisco Nexus Dashboard Insights, o administrador de rede pode observar quedas e ajustar os limites de WRED ou AFD até que as quedas parem em condições normais de tráfego. Esta etapa é a primeira e mais crucial para garantir que a rede de IA/ML lide com ocorrências regulares de congestionamento de tráfego de forma eficaz. Além disso, juntamente com o ajuste dos limites, os administradores de rede podem habilitar o recurso PFC para atingir um comportamento completamente sem perdas. Após a prevenção das quedas, você pode usar marcações ECN e relatórios de contagem de recepção e transmissão de PFC para ajustar o sistema e obter o melhor desempenho.

Interface gráfica do usuário, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

O recurso Análise de Tráfego permite monitorar a latência, o congestionamento e as quedas de tráfego da sua rede. O recurso Análise de Tráfego descobre automaticamente os serviços na sua rede, comparando portas conhecidas da Camada 4 com suas respectivas categorias de endpoint de serviço.

O Cisco Nexus Dashboard Insights avalia o desempenho do serviço com base nos limites definidos para as seguintes métricas:

* **Latência:** esta métrica mede o tempo total (microssegundos) que um pacote leva para ir de um lugar para outro.
* **Congestionamento:** esta métrica mede a utilização da largura de banda da rede e os mecanismos de ativação da qualidade de serviço (QoS) para determinar se um serviço está enfrentando congestionamento de rede.
* **Quedas:** esta métrica mede a porcentagem de pacotes descartados em comparação aos transmitidos, considerando fatores como erros de verificação de redundância cíclica (CRC), cabos defeituosos e outros dispositivos.

Uma anomalia é gerada se as métricas de desempenho, como latência, congestionamento ou quedas, se desviarem do normal. A pontuação de desempenho é calculada para cada conversa e agregada ao endpoint de serviço ou ao nível do endpoint para gerar anomalias.

A pontuação de desempenho é calculada com base no seguinte:

* **Congestionamento:** esta pontuação é um cálculo da prevenção consistente de congestionamento entre pontos finais.
* **Latência:** esta pontuação indica desvio da linha de base medida.
* **Quedas:** esta pontuação indica o número de quedas de pacotes.

Interface gráfica do usuário, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

O recurso Análise de Tráfego permite que você realize as seguintes tarefas:

* Monitore o tráfego extensivamente, incluindo congestionamento e pontuação de latência.
* Relate problemas de desempenho usando anomalias geradas para métricas de desempenho.
* Classifique os principais serviços e clientes ativos e determine os endpoints mais ativos no sistema.
* Determine as contagens de sincronização (sinalizador SYN no cabeçalho TCP) ou redefinição (sinalizador RST no cabeçalho TCP) por serviço.
* Solucione problemas de conversas ou fluxos sob demanda.

Interface gráfica do usuário, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

* Considerações sobre o pipeline
* Aplicações de IA/ML são executadas no ambiente de produção como fluxos de trabalho compostos por múltiplas etapas. Essas etapas têm diferentes requisitos de recursos e desempenho para a infraestrutura do data center. Portanto, os recursos e a capacidade da infraestrutura do data center impactam significativamente o desempenho dos fluxos de trabalho de IA/ML. Vejamos uma interação entre os requisitos de aplicações de IA/ML e a infraestrutura do data center.
* Visão geral do pipeline de IA/ML
* O fluxo de trabalho de IA/ML tem duas etapas principais: treinamento e inferência. Primeiro, você precisa construir e treinar seu modelo e ajustá-lo, e depois implantá-lo em hardware de produção para tomar decisões ou fazer previsões com base em novos dados. Você também deve avaliar e retreinar seu modelo continuamente.

Texto

O conteúdo gerado por IA pode estar incorreto.

O treinamento do modelo se concentra no aprendizado a partir de grandes conjuntos de dados e no ajuste de parâmetros para minimizar erros. A fase de treinamento envolve algumas etapas:

* Coleta de dados
* Pré-processamento
* Engenharia de recursos
* Seleção de modelo
* Treinamento
* Avaliação
* Ajuste de hiperparâmetros
* Treinamento do modelo final

A inferência envolve as seguintes etapas:

* Implantação do modelo
* Entrada de dados
* Previsão
* Pós-processamento
* Avaliação
* Monitoramento
* Estabelecendo um ciclo de feedback

Modelos de grande linguagem (LLMs) de IA generativa exigem recursos significativos que geralmente estão disponíveis apenas para usuários de hiperescala, portanto, você não desenvolverá nem treinará seus próprios modelos. Em vez disso, você usará modelos externos pré-desenvolvidos e os ajustará para o seu caso de uso específico.

Interface gráfica do usuário, Texto

O conteúdo gerado por IA pode estar incorreto.

**Impacto da infraestrutura no desempenho de IA/ML**

Clusters de IA/ML geralmente compreendem muitos nós especializados, com unidades de processamento gráfico (GPUs) e outros aceleradores de hardware interligados em uma rede especializada. Os algoritmos executados nessas GPUs exigem muita computação e realizam cálculos em conjuntos de dados enormes, geralmente maiores do que a memória disponível em uma única GPU. O trabalho é dividido entre várias GPUs para distribuir a carga, e o cluster executa um conjunto iterativo de cálculos no conjunto de dados. Cada GPU realiza uma parte menor do cálculo e envia os resultados a todos os seus pares em um processo de transmissão conhecido como coletivo "Todos para Todos". Após a transmissão, ocorre uma operação de barreira (sincronização), que paralisa todas as GPUs enquanto aguarda o recebimento de todos os dados.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

Essa operação de barreira torna todo o processo extremamente sensível ao desempenho da rede. Se houver apenas um caminho lento na rede, todas as GPUs ficarão paralisadas, aguardando a conclusão dessa transmissão. Essa situação é conhecida como latência de cauda da tarefa.

**Observação**

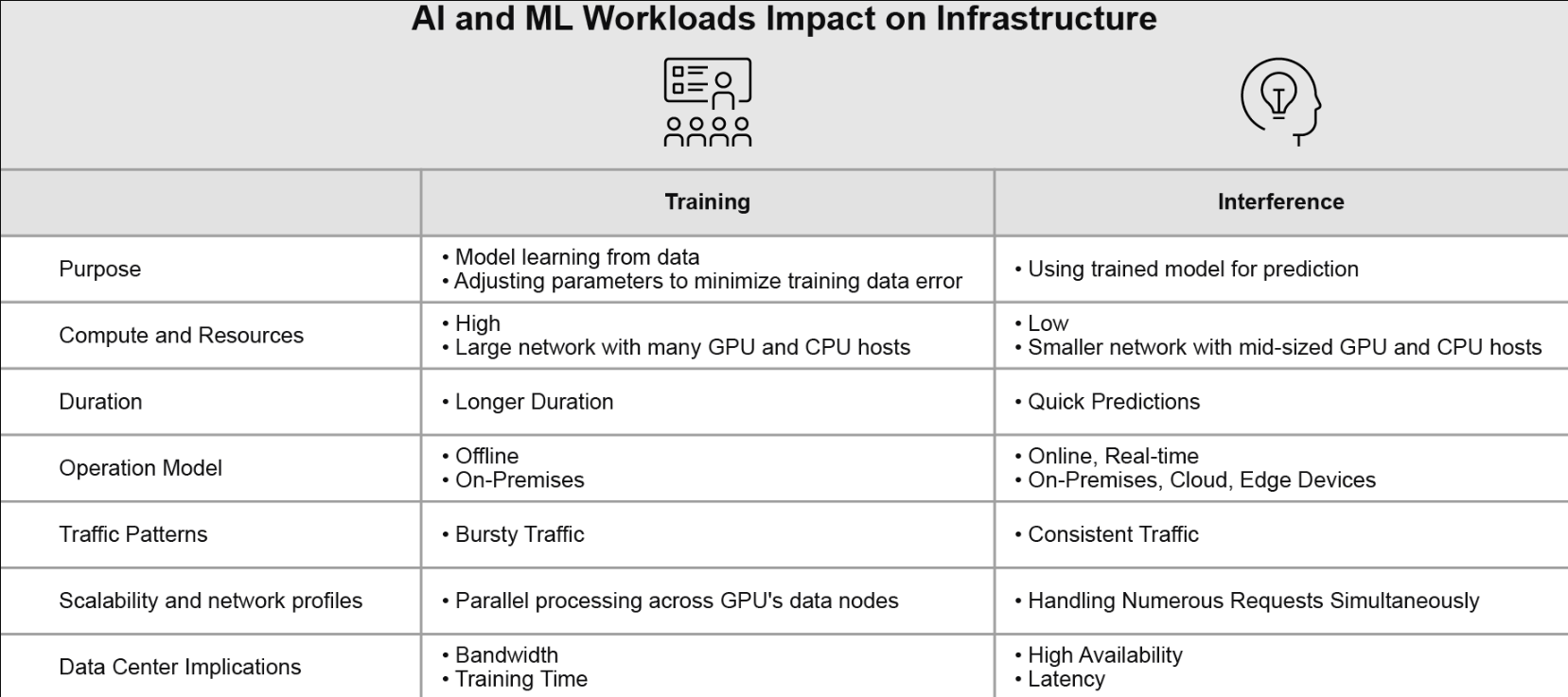
*O tempo que decorre entre o início da transmissão e o momento em que todas as GPUs recebem os seus resultados é o tempo de conclusão da tarefa (JCT). O JCT é utilizado como uma medida crítica do desempenho da IA. Balanceamento de carga deficiente e perdas de pacotes podem causar latências de cauda elevadas, resultando em um JCT deficiente.*

Na vida real, clusters de IA/ML executam muitas tarefas simultâneas e independentes na mesma rede. À medida que mais tarefas são executadas independentemente, a interferência entre elas aumenta. À medida que o congestionamento da rede aumenta, a latência de cauda aumenta. Essa progressão é um evento normal em redes tradicionais. Infelizmente, em redes de IA/ML, o componente de sincronização torna o impacto dessa latência de cauda significativamente maior.

**Impacto do pipeline de IA/ML na infraestrutura**

O treinamento é a etapa mais exigente em termos de recursos de hardware necessários. Requer um ambiente de alta taxa de transferência para E/S e rede. Também requer altos recursos computacionais, frequentemente envolvendo grandes redes com muitos hosts de GPU ou CPU. O processo costuma ser mais longo e opera offline, geralmente no local. Os padrões de tráfego de treinamento são intermitentes e exigem processamento paralelo em vários nós de GPU.

A inferência utiliza o modelo treinado para previsões. Ela requer menos recursos computacionais com redes menores e hosts de GPU ou CPU de médio porte. As previsões são feitas rapidamente, o que permite operações online em tempo real em dispositivos locais, na nuvem e na borda. Portanto, a inferência é mais sensível à latência, pois toma decisões em tempo real com base em consultas ou solicitações do usuário.



Os principais requisitos do data center para suportar a fase de treinamento são alta capacidade de largura de banda e disponibilidade de rede por períodos mais longos. A fase de inferência requer alta disponibilidade e baixa latência para garantir um processamento eficiente em tempo real. Essas diferenças destacam as características das cargas de trabalho de IA/ML e os requisitos de infraestrutura, além de enfatizar a importância de estratégias personalizadas de alocação e gerenciamento de recursos em data centers.

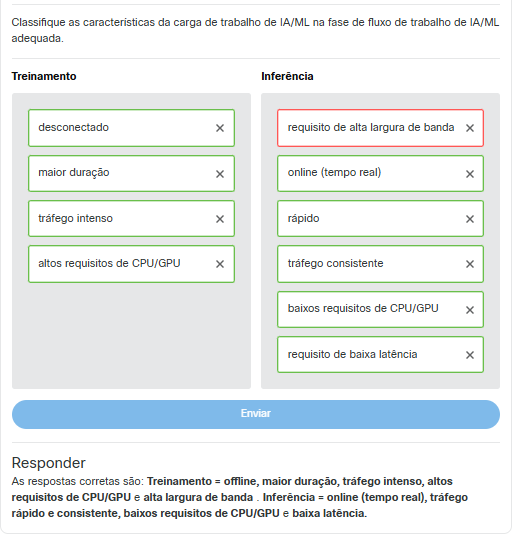
Carta

O conteúdo gerado por IA pode estar incorreto.

**Observação**

*Modelos de IA/ML de classificação são usados ​​para ordenar itens de acordo com critérios definidos pelo usuário. Aplicações comuns são mecanismos de busca e sistemas de recomendação.*

*Os LLMs conseguem compreender e gerar textos com linguagem humana. Aplicações comuns incluem tradução de idiomas, sumarização de textos e muito mais.*



**Resumo**

Neste treinamento, você explorou os mecanismos e a importância da visibilidade do congestionamento em ambientes de rede, com foco especial em redes de cluster de IA/ML. Você aprendeu sobre ECN e PFC como técnicas essenciais para gerenciar e mitigar o congestionamento da rede. A ECN ajuda a sinalizar possíveis congestionamentos para os endpoints, levando-os a reduzir a taxa de transmissão do tráfego, enquanto a PFC garante o transporte sem perdas, pausando o tráfego quando o congestionamento é detectado.

Você também explorou os recursos do Cisco Nexus Dashboard Insights, que oferece recursos abrangentes de monitoramento e telemetria para otimizar o desempenho da rede. Essa ferramenta permite solução de problemas em tempo real, monitoramento de desempenho e gerenciamento de congestionamento, cruciais para manter a eficiência de aplicações de IA/ML.

Por fim, você analisou o impacto dos fluxos de trabalho de IA/ML na infraestrutura do data center para entender os requisitos distintos das fases de treinamento e inferência. O treinamento exige alta largura de banda e recursos computacionais, enquanto a inferência requer baixa latência e recursos de processamento em tempo real.

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

* Como você pode implementar ECN e PFC na sua configuração de rede atual para melhorar o gerenciamento de congestionamento e garantir transporte sem perdas?
* Quais métricas específicas do Cisco Nexus Dashboard Insights seriam mais benéficas para monitorar o desempenho de seus aplicativos de IA/ML e como você usaria esses dados para otimizar sua rede?
* Considerando os requisitos distintos das fases de treinamento e inferência em fluxos de trabalho de IA/ML, como você alocaria recursos em seu data center para equilibrar as necessidades de alto rendimento e baixa latência de forma eficaz?