**Rede de IA**

Introdução

Imagine que você é um engenheiro de rede de TI encarregado de garantir o processamento de dados perfeito e em tempo real para os aplicativos de inteligência artificial (IA) da sua empresa. Você enfrenta o desafio de gerenciar grandes volumes de dados, manter baixa latência e garantir a escalabilidade e a confiabilidade da rede. Este treinamento é voltado para profissionais que estão na vanguarda da integração de cargas de trabalho de IA em ambientes de rede complexos.

Como engenheiro de rede de TI, você precisa garantir que sua infraestrutura de rede possa atender às demandas exclusivas de aplicações de IA. Este treinamento ajudará você a enfrentar esses desafios, aprofundando-se em protocolos essenciais de Camada 2 e Camada 3, como aprendizado de endereço MAC, VLANs, Redes Locais Extensíveis Virtuais (VXLANs) e Protocolo de Árvore de Abrangência (STP). Você também explorará opções avançadas de roteamento, como Protocolo de Gateway de Borda (BGP) e VPN Ethernet (EVPN), e aprenderá como a computação em névoa pode otimizar o processamento distribuído de IA.

Como todas essas informações ajudarão você? Ao final deste treinamento, você será capaz de projetar e implementar redes de IA eficientes, escaláveis ​​e confiáveis. Você adquirirá as habilidades para gerenciar tráfego de broadcast e multicast, agregar largura de banda com EtherChannel e garantir topologias sem loop com STP. Além disso, você aprenderá a usar a computação em névoa para processamento de dados em tempo real e baixa latência, aprimorando o desempenho e a capacidade de resposta de seus aplicativos de IA.

Você poderá aplicar esses conceitos para criar infraestruturas de rede de IA robustas, avaliar e otimizar o desempenho da rede e implementar soluções escaláveis ​​que atendam às crescentes demandas de cargas de trabalho de IA. Esse conhecimento é crucial para manter alto desempenho e confiabilidade em seus projetos baseados em IA, contribuindo, em última análise, para o sucesso da sua organização.

**Protocolos da Camada 2**

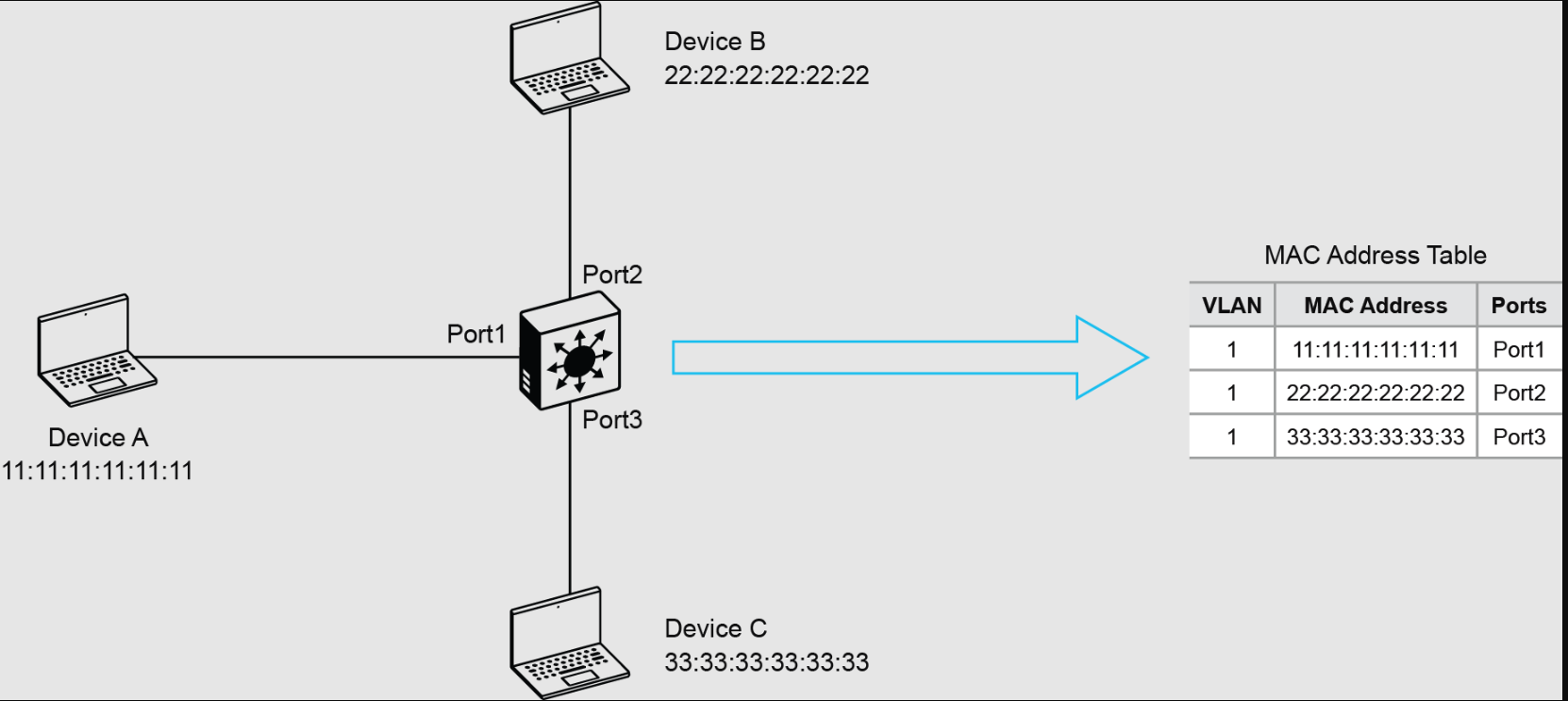
A rede de Camada 2 é fundamental para o transporte de dados em uma rede local. Sua operação se baseia no aprendizado de endereços MAC, na comutação de quadros Ethernet, no gerenciamento de tráfego de broadcast e multicast e na implementação de STP.

A Ethernet estabelece protocolos que permitem que dispositivos em uma rede se comuniquem com eficiência, previnam colisões de dados e garantam uma troca de informações estruturada e organizada. Embora a Ethernet permita a inundação de dados para descobrir destinos desconhecidos, sua função principal é usar endereços aprendidos para direcionar o tráfego apenas para os destinatários pretendidos, minimizando o congestionamento desnecessário da rede. Para evitar esses problemas, protocolos como o STP são essenciais para garantir a estabilidade da rede e evitar loops prejudiciais.

Esses protocolos são cruciais para garantir baixa latência, alto desempenho e estabilidade em redes de IA. Compreender esses aspectos ajuda a projetar redes de Camada 2 eficientes, adaptadas às necessidades dos ambientes de IA.

**Aprendizado de endereço MAC**

O aprendizado de endereços MAC é um mecanismo fundamental da comunicação Ethernet. Ele permite a comutação de quadros Ethernet, o fluxo de dados eficiente e mapeia os endereços MAC dos dispositivos para as portas, permitindo a comutação eficiente de dados dentro de uma rede. Esse mapeamento é salvo na tabela de endereços MAC. A figura a seguir ilustra um exemplo de uma tabela de endereços MAC.



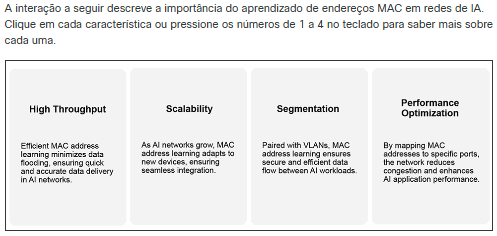
O processo de aprendizagem do endereço MAC consiste nas seguintes etapas:

1. **Chegada do quadro:** Um quadro chega a uma porta do switch.
2. **Verificação de tabela:** O switch verifica se o endereço MAC de origem está em sua tabela de endereços MAC.
3. **Atualização da tabela:** se o endereço MAC não estiver presente, ele mapeia o endereço para a porta atual, incluindo as seguintes informações:
   * **Endereço MAC:** O endereço MAC de origem do quadro.
   * **Porta:** O número da porta na qual o quadro foi recebido.
   * **VLAN:** O ID da VLAN associado ao quadro se as VLANs estiverem configuradas na rede.
4. **Encaminhamento de quadros:** o switch encaminha o quadro para a porta de destino ou o inunda se o destino for desconhecido.
5. **Envelhecimento:** o switch remove periodicamente entradas antigas de endereços MAC com base em um temporizador.

Em redes de IA, o aprendizado de endereços MAC ajuda a garantir uma comunicação estável, sem perdas e com baixa latência entre componentes como servidores, sistemas de armazenamento e unidades de processamento de IA. Esse tipo de comunicação é crucial para aplicações de IA, que dependem do processamento de grandes conjuntos de dados em tempo real.

Interface gráfica do usuário

O conteúdo gerado por IA pode estar incorreto.



**Tráfego de transmissão e multidifusão**

O tráfego de broadcast e multicast otimiza o desempenho das redes de IA. O tráfego de broadcast envia dados para todos os dispositivos em um segmento de rede, independentemente de cada dispositivo precisar ou não das informações. O tráfego multicast envia dados de uma única fonte para um grupo específico de dispositivos que optaram por recebê-los. Ambos os tipos de tráfego atendem a propósitos distintos, e seu gerenciamento é fundamental para manter uma rede eficiente e de alto desempenho, especialmente em ambientes onde o processamento e a distribuição de dados em tempo real são cruciais.

No nível da Camada 2, a Ethernet lida com o tráfego de broadcast e multicast usando inundação, onde o switch envia pacotes para todas as portas, exceto a porta de onde os pacotes foram recebidos. Essa abordagem garante que o tráfego chegue aos seus destinos pretendidos, mas pode levar ao congestionamento da rede se não for gerenciada adequadamente.

Existem dois tipos de transmissão na Camada 3: transmissão local (limitada) e transmissão direcionada (direcionada). A transmissão local é representada como 255.255.255.255 e é limitada pela sub-rede local à qual o nó está conectado. Essas transmissões não podem ser enviadas através do dispositivo da Camada 3. E este pacote de transmissão será recebido por todos os nós nesta sub-rede. A transmissão direcionada tem um endereço de destino, que é o último endereço IP na sub-rede IP, mas se origina de um nó que não faz parte dessa sub-rede de destino. Um dispositivo que não está diretamente conectado à sua sub-rede de destino encaminha uma transmissão direcionada da mesma forma que encaminharia pacotes IP unicast destinados a um host nessa sub-rede. Quando um pacote de transmissão direcionada atinge um dispositivo que está diretamente conectado à sua sub-rede de destino, esse pacote é transmitido na sub-rede de destino. O endereço de destino no cabeçalho IP do pacote é reescrito no endereço de transmissão IP configurado para a sub-rede, e o pacote é enviado como uma transmissão da camada de enlace. Por padrão, a transmissão direcionada por IP é desabilitada nas interfaces do roteador e todas as transmissões direcionadas por IP são descartadas.

O tráfego de broadcast é normalmente local e é usado por protocolos como o Protocolo de Resolução de Endereços (ARP) ou DHCP para descobrir dispositivos ou atribuir endereços IP. Em redes de IA, onde grandes volumes de dados podem ser transmitidos, o tráfego de broadcast excessivo pode levar ao congestionamento da rede e reduzir o desempenho e a eficiência gerais. Em cenários que envolvem implantações de IA em larga escala, como cidades inteligentes ou redes de veículos autônomos, o gerenciamento do tráfego de broadcast torna-se ainda mais crítico, pois pode saturar a rede e levar a atrasos e potencial perda de dados.

A figura a seguir mostra o fluxo de um tráfego de broadcast direcionado originado do servidor principal, passando pelo switch e pelo roteador e alcançando todos os dispositivos finais conectados. Todos os dispositivos finais recebem a mensagem de broadcast.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

Em redes de IA, o tráfego multicast é importante para distribuir com eficiência dados em tempo real, como atualizações de modelos em aprendizagem colaborativa ou feeds de sensores em sistemas autônomos, para vários nós simultaneamente. Essa abordagem reduz o uso de largura de banda e garante atualizações sincronizadas, tornando-a ideal para aplicações de IA em larga escala, nas quais vários dispositivos ou nós precisam processar os mesmos dados simultaneamente.

A figura a seguir exibe o fluxo de tráfego multicast.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

A seguir estão alguns exemplos de como o tráfego de transmissão e multidifusão é usado em redes relacionadas à IA.

Interface gráfica do usuário, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

**Protocolo de Árvore de Abrangência**

STP é um protocolo de rede que garante uma topologia sem loops em redes Ethernet com caminhos redundantes. Um loop de rede ocorre quando existem vários caminhos entre dispositivos, fazendo com que os pacotes de dados circulem indefinidamente. Esse problema leva a tempestades de transmissão, congestionamento e interrupções na rede.

O STP bloqueia algumas portas do switch para evitar loops e ainda permitir redundância de rede.

**Observação**

*Uma tempestade de transmissão é um número anormalmente alto de pacotes de transmissão em um curto período.*

A figura a seguir ilustra o processo STP.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

 processo STP garante que exista apenas um caminho ativo entre dois dispositivos de rede, eliminando loops e ainda fornecendo redundância de rede.

As etapas a seguir descrevem o processo STP, que começa quando o PC A envia uma mensagem de broadcast para o PC B. A mensagem passa pelo Switch B e é então encaminhada para o Switch C e o Switch A. Cada switch encaminha a mensagem para todas as portas, exceto aquela de onde foi recebida. Esse comportamento pode criar um loop, pois as mensagens trafegam entre os Switches A, B e C e não saem do fluxo.

Ao bloquear uma porta específica, o loop é evitado, conforme detalhado no processo a seguir:

1. Eleição da ponte raiz
   1. Todos os switches inicialmente se consideram a ponte raiz.
   2. O switch com o menor ID de ponte (prioridade) torna-se a ponte raiz. Se vários switches tiverem a mesma prioridade, o switch com o menor endereço MAC é selecionado. Na figura anterior, o switch com o menor endereço MAC é o Switch A, com endereço MAC AA:AA:AA:AA:AA:AA.
2. Seleção da porta raiz
   1. Cada switch não raiz seleciona uma porta raiz — a porta com o caminho de menor custo para a ponte raiz.
   2. A porta raiz está no estado de encaminhamento e é o caminho principal do switch para a bridge raiz.
3. Seleção de porta designada
   1. Para cada segmento de rede, uma porta designada é selecionada no switch com o menor custo de caminho para a bridge raiz.
   2. A porta designada também está no estado de encaminhamento e manipula o tráfego para esse segmento.
4. Bloqueio de portas não designadas
   1. Portas que não são portas raiz ou portas designadas são colocadas em um estado de bloqueio para evitar loops.

Manter uma rede estável e sem loops em ambientes de IA é importante para evitar interrupções no fluxo de dados, o que pode afetar o processamento em tempo real e o desempenho do aplicativo.

A tabela a seguir descreve possíveis problemas que podem ocorrer se o STP não for usado e o impacto no ambiente de IA, dado o cenário específico.

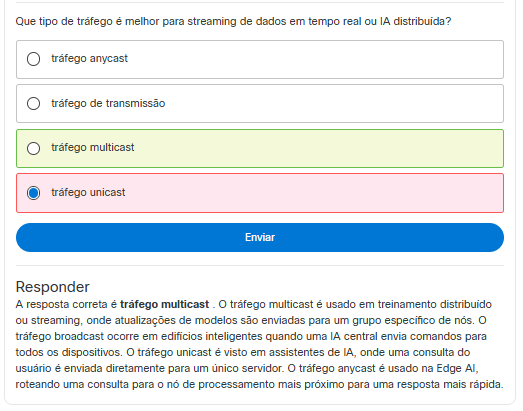
| Cenário | Possíveis problemas sem STP | Impacto no ambiente de IA |
| --- | --- | --- |
| Data centers que hospedam cargas de trabalho de IA | Loops de rede podem causar tempestades de transmissão e sobrecarregar a rede. | O desempenho degradado nas transferências de dados entre GPUs e armazenamento pode tornar o treinamento e a inferência do modelo mais lentos. |
| Sistemas autônomos baseados em IA | Loops de rede podem causar atrasos na transmissão de dados. | A tomada de decisões atrasada ou incorreta por veículos autônomos ou drones afeta a segurança. |
| Implantações de IA de ponta | Loops de rede podem interromper a conectividade entre dispositivos de ponta e modelos e bancos de dados de IA centralizados. | A perda de conectividade pode interromper operações como manutenção preditiva ou gerenciamento de tráfego. |

Uma imagem contendo Diagrama

O conteúdo gerado por IA pode estar incorreto.

Tabela

O conteúdo gerado por IA pode estar incorreto.



**Protocolos da Camada 3**

A rede de Camada 3 é essencial para gerenciar o tráfego de dados em diferentes ambientes de rede. BGP e EVPN são as principais tecnologias que otimizam o roteamento de Camada 3, especialmente em cenários de rede complexos e de grande escala.

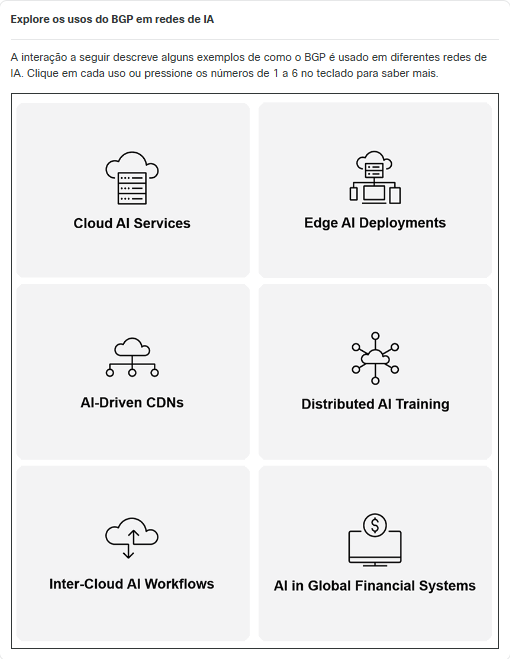
Redes de IA utilizam protocolos avançados, como BGP e EVPN, para ajustar dinamicamente os caminhos de roteamento, otimizar o fluxo de tráfego e prever o congestionamento da rede antes que ele ocorra. A integração da IA ​​com tecnologias de Camada 3 aumenta a eficiência e a resiliência das redes, permitindo que elas se adaptem a mudanças em tempo real e garantam o desempenho ideal em infraestruturas cada vez mais complexas.

**Protocolos BGP e EVPN**

O BGP é usado para trocar informações de roteamento entre diferentes sistemas autônomos na internet. Ele determina o melhor caminho para os dados trafegarem pelas redes interconectadas.

O BGP inclui dois recursos principais:

* **Protocolo de vetor de caminho** : o BGP mantém as informações de caminho, que são atualizadas dinamicamente conforme a topologia da rede muda.
* **Roteamento baseado em políticas** : o BGP permite que os administradores definam políticas de roteamento com base em vários atributos, como caminho do sistema autônomo e comprimento do prefixo.



Tabela

O conteúdo gerado por IA pode estar incorreto.

A figura a seguir mostra uma rede BGP envolvendo quatro roteadores distribuídos em três sistemas autônomos. O sistema autônomo 100 possui dois roteadores: RTA e RTB. O sistema autônomo 300 contém o roteador RTC e o sistema autônomo 400 inclui o roteador RTD.

O RTA no sistema autônomo 100 está conectado ao RTC no sistema autônomo 300 usando o intervalo de endereços IP 10.2.2.x. Da mesma forma, o RTB no sistema autônomo 100 está conectado ao RTD no sistema autônomo 400 usando o intervalo de endereços IP 10.1.1.x. As redes internas são identificadas como 172.31.202.2 para o RTA, 172.16.10.0 para o RTC e 172.16.175.0 para o RTD.

Esses roteadores usam BGP para trocar informações de roteamento, transportar tráfego de forma eficiente entre sistemas autônomos, gerenciar loops de roteamento e otimizar a seleção de caminhos.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

A EVPN estende os serviços Ethernet da Camada 2 sobre redes IP da Camada 3, permitindo a integração perfeita entre redes Ethernet e IP. Ela utiliza o BGP como plano de controle para gerenciar a distribuição de endereços MAC e informações de roteamento IP.

Há dois recursos principais do EVPN:

* **Aprendizado de endereço MAC** : o EVPN fornece um método escalável para distribuir informações de endereço MAC em uma rede de Camada 3.
* **Roteamento e bridging integrados** : o EVPN oferece suporte a serviços de Camada 2 e Camada 3, permitindo projetos de rede flexíveis e operações de rede simplificadas.

Uma imagem contendo Tabela

O conteúdo gerado por IA pode estar incorreto.

Tabela

O conteúdo gerado por IA pode estar incorreto.

A lista a seguir inclui os principais recursos do EVPN que atendem às crescentes necessidades das operadoras de rede:

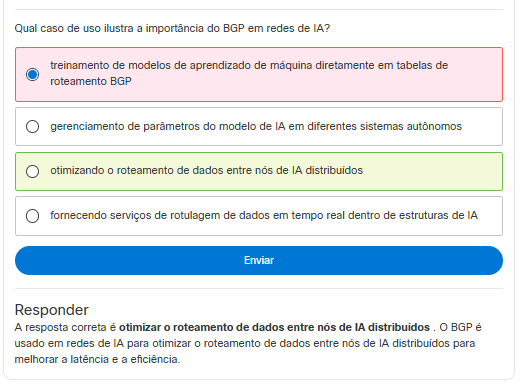
* **Operações de interconexão de data center (DCI):**
  1. **Escalabilidade:** O tráfego de dados aumenta e os data centers distribuídos geograficamente são gerenciados.
  2. **Eficiência:** conectividade perfeita é fornecida entre vários data centers.
* **Virtualização de nuvem e serviços:**
  1. **Alocação dinâmica:** há suporte para provisionamento flexível de recursos e serviços.
  2. **Integração:** os serviços da Camada 2 e da Camada 3 são combinados em uma VPN unificada.
* **Simplificação de redes:**
  1. **Facilidade operacional:** Vários protocolos são consolidados para reduzir a complexidade.
  2. **Menor sobrecarga:** o gerenciamento e a solução de problemas são simplificados.
* **Integração perfeita das Camadas 2 e 3:**
  1. **Convergência:** Vários serviços de rede são integrados em uma única VPN.
  2. **Flexibilidade:** Diversas necessidades e ambientes de rede são atendidos.
* **Implantação de serviço flexível e posicionamento de carga de trabalho:**
  1. **Gerenciamento dinâmico:** o posicionamento e a movimentação de serviços e cargas de trabalho são gerenciados com eficiência.
  2. **Otimização de recursos:** as cargas de rede são balanceadas e a utilização de recursos é otimizada.
* **Suporte para multilocação:**
  1. **Isolamento:** a segregação eficaz de dados e serviços é garantida para vários locatários.
  2. **Segurança:** A segurança e a conformidade são aprimoradas por meio da segmentação adequada.
* **Encaminhamento ideal e mobilidade de carga de trabalho:**
  1. **Otimização de tráfego:** caminhos de encaminhamento eficientes são fornecidos e as alterações de rede são processadas sem problemas.
  2. **Mobilidade:** a movimentação da carga de trabalho é simplificada e minimamente perturbadora.
* **Convergência rápida e utilização de largura de banda:**
  1. **Confiabilidade:** O tempo de inatividade é reduzido com a rápida adaptação às mudanças na rede.
  2. **Eficiência:** a utilização eficaz da largura de banda mantém o desempenho e reduz custos.

**Otimização do Protocolo de Roteamento**

BGP e EVPN otimizam protocolos de roteamento de Camada 3 em diversos ambientes de rede. Juntos, eles oferecem melhorias significativas para DCI, ambientes multilocatários e redes de provedores de serviços, além de garantir alto desempenho, escalabilidade e confiabilidade em infraestruturas de rede modernas. A tabela a seguir mostra como eles oferecem essas melhorias.

| Categoria | Visão geral | Otimização de BGP e EVPN |
| --- | --- | --- |
| DCI | A DCI envolve a conexão de vários data centers para permitir mobilidade de dados e aplicativos, recuperação de desastres e alta disponibilidade. | * A EVPN fornece conectividade escalável de Camada 2 entre data centers, mantendo o roteamento de Camada 3 usando BGP. Essa abordagem permite a migração perfeita de máquinas virtuais (VMs) e o balanceamento de carga entre data centers. * A capacidade do BGP de manipular grandes tabelas de roteamento e roteamento baseado em políticas aumenta a eficiência e a flexibilidade do DCI. |
| Ambientes multilocatários | Ambientes multilocatários, como provedores de serviços de nuvem, exigem isolamento e gerenciamento eficiente de múltiplas redes de clientes. | * O EVPN oferece suporte ao isolamento de locatários por meio de serviços de LAN privada virtual (VPLS) para garantir segmentação de rede segura e eficiente. * O BGP gerencia o roteamento através desses segmentos isolados para fornecer interconectividade escalável e flexível entre locatários ou redes virtuais. |
| Redes de provedores de serviços | Os provedores de serviços devem oferecer soluções de rede de alto desempenho, escaláveis ​​e resilientes para dar suporte a muitos clientes e serviços. | * O EVPN permite serviços VPN de Camada 2 eficientes em um backbone de Camada 3 para dar suporte a uma ampla gama de serviços e, ao mesmo tempo, simplificar o gerenciamento de rede. * O BGP fornece recursos robustos de roteamento entre domínios e suporta recursos avançados como engenharia de tráfego e balanceamento de carga para otimizar os recursos de rede. |

A integração da IA ​​em ambientes BGP e EVPN otimiza ainda mais o desempenho da rede. Em cenários complexos de DCI, multilocatários e provedores de serviços, a IA aprimora a capacidade de manter o roteamento ideal, reduzir a latência e melhorar a eficiência geral da rede, resultando em infraestruturas de rede mais inteligentes e adaptáveis.



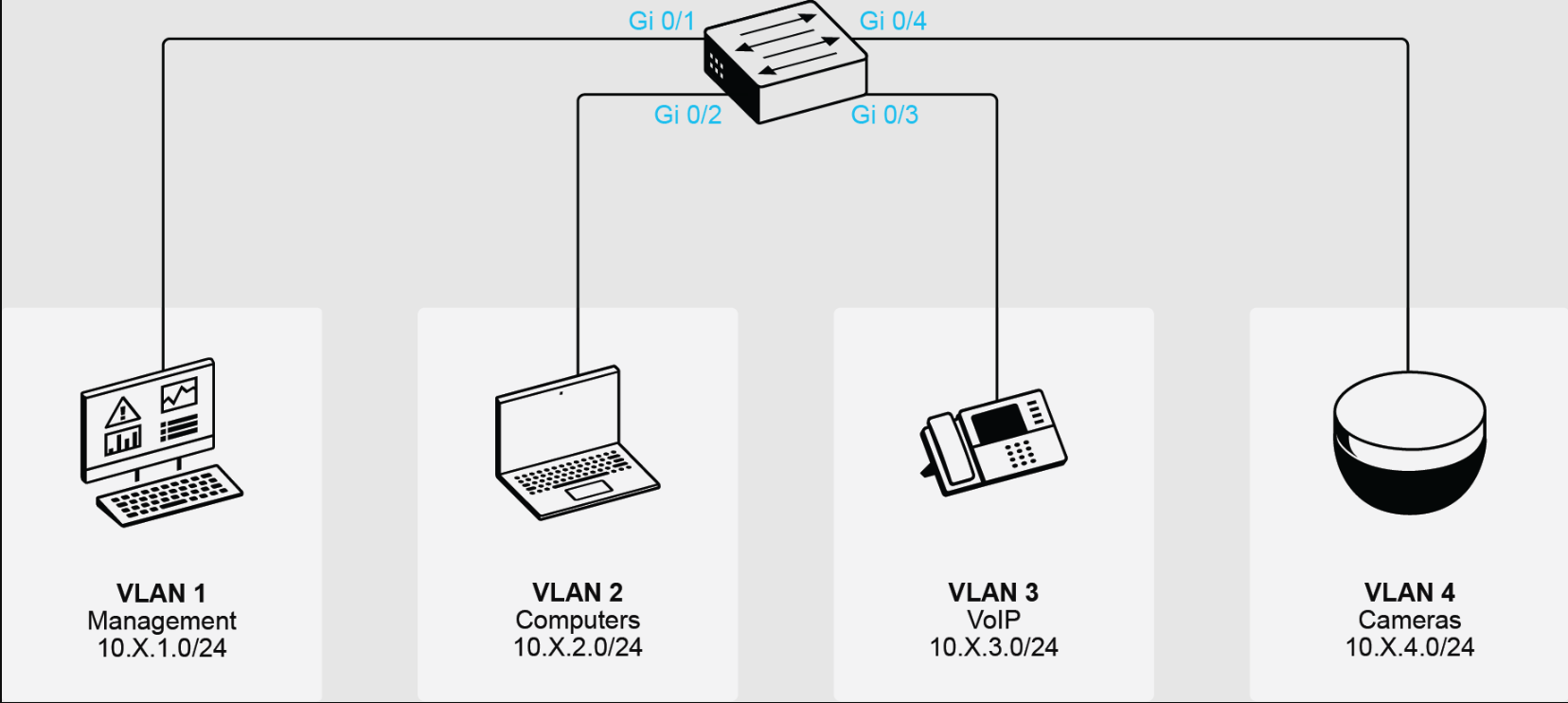
**Considerações de escalabilidade para implantação de cargas de trabalho de IA**

Em ambientes de IA modernos, a escalabilidade e a eficiência da rede são fatores críticos que influenciam diretamente o desempenho das cargas de trabalho de IA. À medida que cresce a demanda por processamento em tempo real, análise de dados em larga escala e serviços de IA baseados em nuvem, garantir que a infraestrutura de rede subjacente possa suportar essas demandas torna-se cada vez mais importante.

Diversas tecnologias de rede avançadas, como VLANs, VXLAN e EtherChannel, desempenham papéis cruciais na otimização e no dimensionamento de redes de IA. Essas tecnologias melhoram a gerenciabilidade, a segurança e a confiabilidade, essenciais para manter a operação ininterrupta dos sistemas de IA em ambientes diversos e complexos.

**VLAN**

Conforme mostrado na figura a seguir, as VLANs segmentam uma rede física em várias redes lógicas para melhorar o desempenho e a capacidade de gerenciamento.



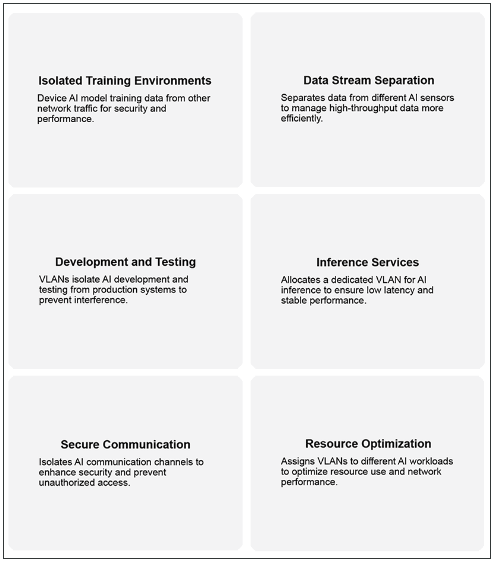
As VLANs podem ser usadas para isolar diferentes cargas de trabalho de IA ou segmentos da rede, reduzir o tráfego de broadcast e aumentar a segurança. Essa segmentação permite um melhor gerenciamento de recursos e o isolamento de tarefas de IA de alta prioridade.

**Explore casos de uso de VLAN em IA**

A interação a seguir fornece alguns exemplos de casos de uso de VLANs em redes de IA. Clique em cada caso de uso ou pressione os números de 1 a 6 no teclado para saber mais.

Uma imagem contendo Tabela

O conteúdo gerado por IA pode estar incorreto.



**VXLAN**

VXLAN é uma tecnologia de virtualização de redes que estende redes da Camada 2 por toda a infraestrutura da Camada 3. Ela encapsula quadros Ethernet em pacotes UDP, permitindo a criação de redes virtuais que podem abranger múltiplas redes físicas. Essa abordagem facilita a implantação de muitas redes virtuais isoladas dentro da mesma infraestrutura física.

A VXLAN é benéfica em ambientes de data center onde as cargas de trabalho de IA podem abranger vários servidores e localizações geográficas. Ela permite a virtualização de rede escalável e atende às altas demandas de implantações de IA em larga escala e serviços de IA baseados em nuvem.

A figura a seguir mostra um diagrama de rede com dois pods de data center conectados por switches Cisco Nexus Série 9300. Esses switches atuam como o núcleo da rede e gerenciam a comunicação entre os pods usando a tecnologia VXLAN. A topologia exibe as camadas de rede e garante alta disponibilidade e redundância.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

**Observação**

*A sobreposição VXLAN é uma tecnologia que permite a criação de uma rede virtualizada de Camada 2 sobre uma infraestrutura de rede de Camada 3 (IP) existente.*



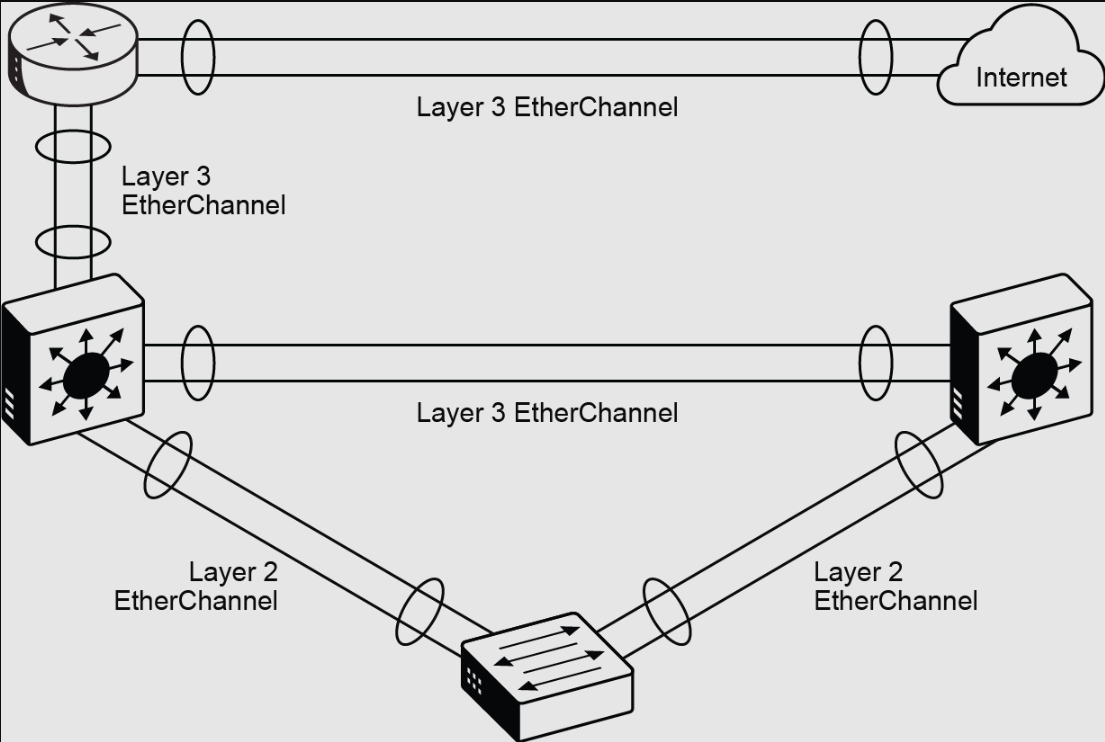
Tabela

O conteúdo gerado por IA pode estar incorreto.

**Canal Ether**

O EtherChannel agrega múltiplos links físicos em um único link lógico. Existem dois tipos de EtherChannel: EtherChannel de Camada 2 e EtherChannel de Camada 3. O EtherChannel de Camada 2 agrega múltiplos links físicos em um único link lógico na camada de enlace de dados (Camada 2). Ele é usado para aumentar a largura de banda e fornecer redundância entre switches. A Camada 3 agrega múltiplos links físicos em um único link lógico na camada de rede (Camada 3). Ele é usado entre roteadores ou switches da Camada 3 para balanceamento de carga e failover.

A figura a seguir mostra como a agregação de links funciona com múltiplos links físicos. A agregação de links pode ser feita como EtherChannel de Camada 2 ou Camada 3.



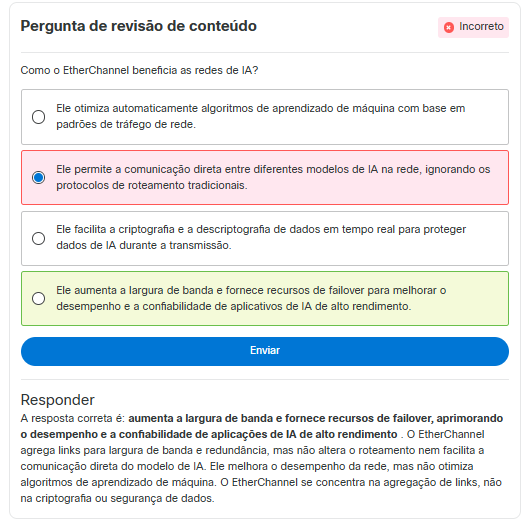
O EtherChannel melhora o desempenho e a confiabilidade das redes de IA, aumentando a largura de banda e fornecendo recursos de failover. Essas melhorias são cruciais para aplicações de IA de alto rendimento que exigem conexões de rede consistentes e confiáveis.

Uma imagem contendo Tabela

O conteúdo gerado por IA pode estar incorreto.

Texto, Tabela

O conteúdo gerado por IA pode estar incorreto.



**Computação em Névoa para Processamento Distribuído de IA**

A computação em névoa é uma arquitetura de computação descentralizada que estende os recursos da computação em nuvem para mais perto da borda da rede e dos locais onde os dados são produzidos ou coletados. Ao contrário da computação em nuvem tradicional, que depende de data centers centralizados para processamento de dados, a computação em névoa distribui recursos de processamento, armazenamento e rede entre vários dispositivos, incluindo sensores, dispositivos de borda, servidores locais e recursos de nuvem.

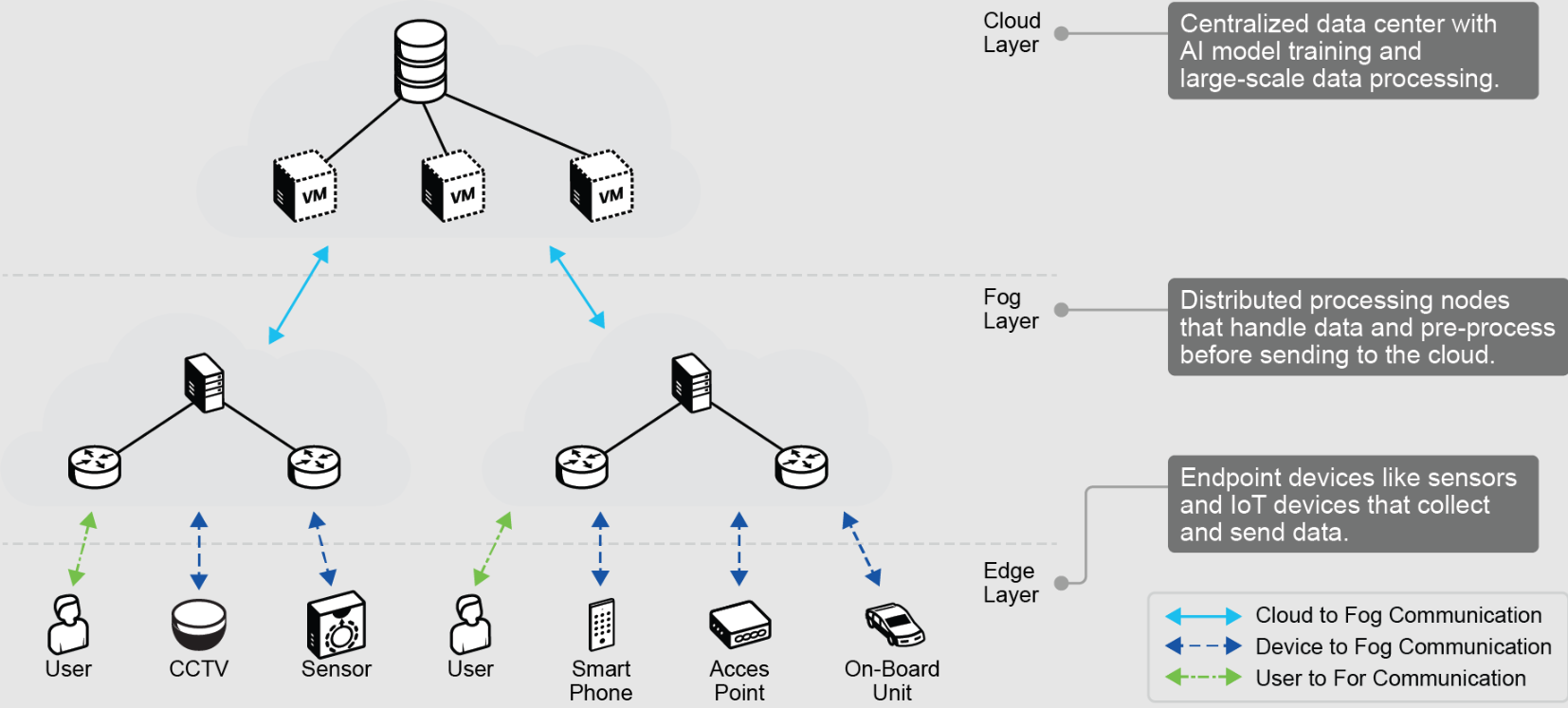
Essa arquitetura é útil no processamento distribuído de IA, onde a computação e a comunicação de modelos de IA são distribuídas entre vários nós ou dispositivos. A IA distribuída utiliza processamento paralelo e computação colaborativa para aprimorar a escalabilidade, melhorar a estabilidade e lidar com tarefas complexas com eficiência. No entanto, apresenta desafios como gerenciar a sobrecarga de comunicação e garantir a sincronização.

O sistema de computação em neblina é composto por três camadas: camada de nuvem, camada de neblina e camada de borda. Cada camada possui funções e componentes específicos que interagem para otimizar o processamento e o gerenciamento de dados, especialmente no contexto de aplicações de IA e Internet das Coisas (IoT).

A tabela a seguir descreve os componentes de um sistema de computação em neblina.

| Camada | Componentes | Funções |
| --- | --- | --- |
| Camada de nuvens | Servidores em nuvem | Infraestrutura centralizada para processamento de alto nível, armazenamento de longo prazo e análises complexas. Comunica-se com servidores Fog. |
| Camada de neblina | Servidores de neblina | Atua como uma ponte entre as camadas de nuvem e de borda, fornece processamento descentralizado mais próximo das fontes de dados e reduz a latência e o uso de largura de banda. |
| Camada de borda | Dispositivos de usuário final e fontes de dados (usuários, câmeras, sensores, smartphones, pontos de acesso, unidades de bordo) | Mais próximo das fontes de dados; processa os dados localmente ou os envia para a camada de neblina para processamento posterior. |

A figura a seguir exibe os componentes do sistema de computação em neblina e as características de cada camada em relação aos processos distribuídos de IA



A computação em névoa permite o processamento em tempo real e com baixa latência de dados gerados por dispositivos de IoT, veículos autônomos e sistemas industriais. Ao trazer o poder computacional para mais perto da borda, a computação em névoa permite que modelos de IA sejam treinados, inferidos e implantados com eficiência na fonte de geração de dados ou próximo a ela.

**Observação**

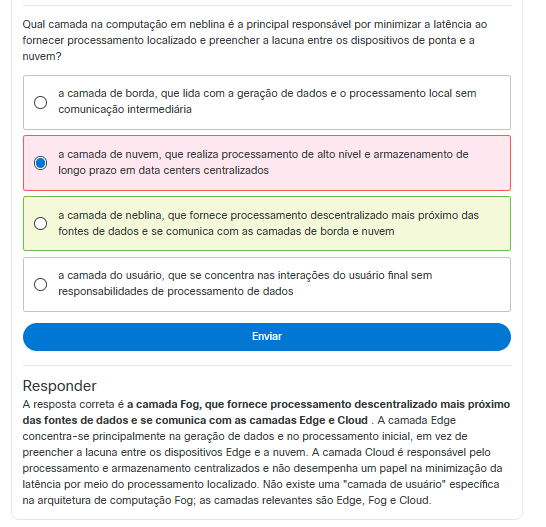
*Embora a IA tradicional opere de maneira centralizada, a IA distribuída distribui a computação, melhora a escalabilidade e melhora a tolerância a falhas.*

A lista a seguir descreve os principais benefícios da computação em névoa para IA distribuída:

* **Latência reduzida:** os dados são processados ​​localmente, o que é crucial para aplicações em tempo real, como direção autônoma e automação industrial.
* **Eficiência de largura de banda:** os dados enviados para a nuvem são limitados para conservar a largura de banda e reduzir custos em cenários com restrição de largura de banda.
* **Segurança aprimorada** : dados confidenciais são processados ​​na borda, o que melhora a privacidade e a segurança antes da transmissão na nuvem.
* **Escalabilidade:** o processamento de IA distribuído ocorre entre nós, o que dimensiona recursos dinamicamente para cargas de trabalho variadas.
* **Confiabilidade aprimorada:** as tarefas são distribuídas entre os nós de neblina, garantindo operação contínua mesmo se um nó falhar.
* **Tomada de decisão em tempo real:** ações imediatas são suportadas em aplicativos sensíveis ao tempo porque os dados são processados ​​onde são gerados.



A computação em névoa é um importante facilitador para a IA, fornecendo a infraestrutura necessária para processar grandes volumes de dados localmente, tomar decisões em tempo real e dar suporte a sistemas inteligentes em uma ampla gama de aplicações. À medida que a IA evolui, a computação em névoa desempenhará um papel mais significativo no processamento distribuído e levará a sistemas de rede mais inteligentes e eficientes.



**Resumo**

Parabéns por concluir o treinamento *em Redes de IA* !

Neste treinamento, você aprendeu sobre protocolos e tecnologias de rede essenciais para ambientes de IA. O treinamento começou com protocolos de Camada 2, enfatizando o aprendizado de endereços MAC, o gerenciamento de tráfego de broadcast e multicast e o STP para garantir baixa latência e estabilidade da rede. Você examinou protocolos de Camada 3, como BGP e EVPN, para otimizar o roteamento e aprimorar o desempenho da rede em redes de IA complexas. Considerações de escalabilidade, incluindo VLANs, VXLAN e EtherChannel, que melhoram a capacidade de gerenciamento, a segurança e a confiabilidade, foram discutidas. Você também aprendeu sobre a computação em névoa como uma arquitetura descentralizada que aproxima o processamento das fontes de dados, reduz a latência e aprimora a tomada de decisões em tempo real.

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

* Como o EVPN estende os serviços da Camada 2 sobre redes da Camada 3 e quais são seus benefícios para redes de IA?
* Como você pode implementar VLANs ou VXLANs em sua rede para melhorar o desempenho e a segurança das cargas de trabalho de IA?
* Como o tráfego de transmissão e multidifusão diferem e quais são suas funções em ambientes de IA?
* Como a computação em neblina pode ser integrada à sua infraestrutura atual para melhorar o processamento de dados e a tomada de decisões em tempo real?