**Desempenho de dados de carga de trabalho de IA/ML**

Cargas de trabalho de IA/ML exigem hardware de computação de alto desempenho (HPC), o que representa um investimento significativo. Para maximizar a lucratividade, é crucial que esse hardware seja utilizado continuamente em plena capacidade. No entanto, o congestionamento da rede pode fazer com que os nós de HPC fiquem ociosos e subutilizados, levando a ineficiências de custo. Obter insights detalhados sobre o desempenho da rede é fundamental para detectar e lidar preventivamente com o congestionamento. Ao otimizar parâmetros de configuração, como limites de buffer, e empregar mecanismos como Controle de Fluxo Prioritário (PFC) e Notificação Explícita de Congestionamento (ECN), esses problemas podem ser mitigados. Neste treinamento, você aprenderá a usar o Cisco Nexus Dashboard Insights (NDI) e seu recurso de Análise de Tráfego para monitorar o congestionamento da rede com eficácia.

Ao final deste treinamento, você será capaz de:

* Adicione uma estrutura ao Cisco Nexus Dashboard Insights (NDI).
* Habilite o recurso Análise de Tráfego.
* Interpretar relatórios de anomalias relacionados ao congestionamento da rede.
* Analise métricas de congestionamento de rede ao longo do tempo no NDI.



Desempenho de dados de carga de trabalho de IA/ML

Você é engenheiro de rede sênior em uma empresa especializada no desenvolvimento de assistentes de IA. Entre suas responsabilidades está garantir que as redes que conectam os servidores GPU funcionem com eficiência e estejam livres de congestionamento. Recentemente, você implantou uma rede de topologia spine-leaf de pequena escala, composta por dois switches leaf e um switch spine, usando switches Cisco Nexus 9300. Você configurou uma malha BGP com uma rede sobreposta EVPN por meio do Cisco Nexus Dashboard Fabric Controller (NDFC). Você utilizou seus modelos integrados para implementar Qualidade de Serviço (QoS), Controle de Fluxo Prioritário (PFC) e Notificação Explícita de Congestionamento (ECN) com Detecção Antecipada Aleatória Ponderada (WRED), com o objetivo de criar uma rede sem perdas pronta para RoCEv2. Ao configurar o PFC e o ECN, você optou pelos valores de limite de buffer padrão, mas agora se preocupa se essas configurações serão suficientes para as cargas de trabalho de IA. Você projetou a rede com uma arquitetura não bloqueante, conectando os switches leaf à espinha com uplinks de 40G, com portas de acesso de 1G em mente. Embora a equipe de IA planeje inicialmente implantar apenas três servidores GPU, existe a possibilidade de expansão para até 40 servidores. O projeto da sua rede considera essa expansão, garantindo um desempenho sem bloqueios em toda a rede. A etapa final da implantação é testar se as configurações padrão lidam adequadamente com o congestionamento dos três servidores iniciais.

O Cisco Nexus Dashboard Insights (NDI) é uma ferramenta ideal para monitorar sua rede, graças ao seu recurso de Análise de Tráfego. Este recurso é capaz de detectar congestionamentos e emitir alertas. Ele também fornece métricas detalhadas relacionadas ao congestionamento, incluindo tail-drops, fluxos de pacotes de Controle de Fluxo Prioritário (PFC) e Notificação Explícita de Congestionamento (ECN), além de taxas de utilização da interface. Você reconhece o valor de continuar usando o NDI mesmo quando a equipe de IA expande sua rede de servidores. Este recurso permite monitorar o congestionamento e ajustar as configurações de PFC e ECN para evitar futuros problemas de congestionamento.

**Observação**

*Esta é uma simulação de laboratório e não reflete inteiramente o comportamento de dispositivos reais. Preste atenção às notas sobre as diferenças entre a simulação de laboratório e os dispositivos reais ao longo deste exercício.*

*Adicionar um site ao Cisco Nexus Dashboard Insights e habilitar a análise de tráfego*

*Você já possui um cluster Cisco Nexus Dashboard Insights (NDI) implantado com conectividade com o Cisco NDFC, gerenciando a rede de IA habilitada. Agora, você precisa incluir essa rede de IA como um site no Cisco NDI e habilitar o recurso de Análise de Tráfego para o site, para poder começar a monitorá-lo.*