**Migração de arquitetura para rede de IA/ML**

**Introdução**

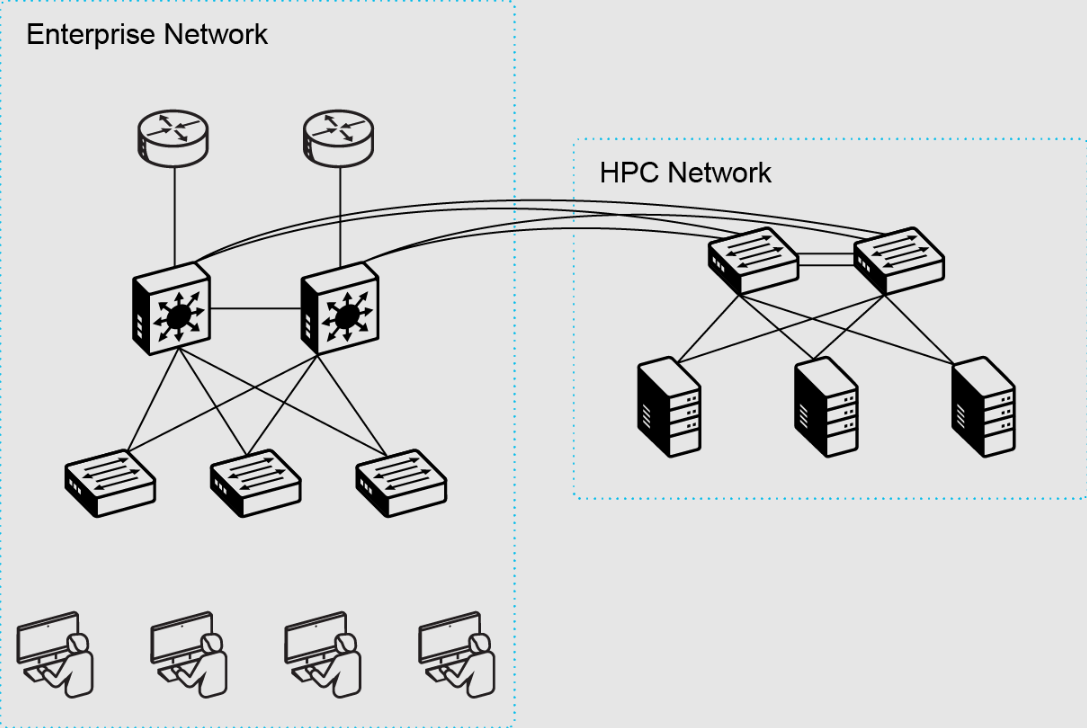
Esta atividade examina os aspectos de rede da infraestrutura de IA/ML. Ela detalha a configuração da infraestrutura necessária com Unidades de Processamento Gráfico (GPUs) avançadas para um projeto de Modelo de Linguagem Grande (LLM) e os desafios de integração da combinação de outras cargas de trabalho com cargas de trabalho de IA em um data center compartilhado.

**Descrição do Projeto**

A empresa de trading para a qual você trabalha auxilia traders, investidores e consultores em suas necessidades. A empresa é bem conhecida no setor de trading e atende a diversos clientes.

Uma pequena equipe dentro da sua empresa de trading quer experimentar novas maneiras de aprimorar seu trabalho. Eles acreditam que os LLMs podem ajudá-los a redigir ofertas aos clientes com mais rapidez e gerar um feedback mais preciso. Esta equipe está animada para montar um laboratório para testar frameworks de IA, na esperança de tornar a comunicação mais eficiente, melhorar a satisfação do cliente e se manter à frente no mercado.

A empresa possui uma rede corporativa projetada para dar suporte às suas diversas atividades de negociação e operações internas. Essa rede garante conectividade segura e de alta velocidade para todas as funções da empresa e facilita a comunicação e a troca de dados ininterruptas em toda a organização. Sua empresa de negociação também mantém um pequeno data center dedicado exclusivamente à execução de cargas de trabalho de Computação de Alto Desempenho (HPC). Essas cargas de trabalho são essenciais para a execução de algoritmos e análises de negociação complexos. Todos os outros serviços utilizados pela empresa são fornecidos por meio de soluções de Software como Serviço (SaaS), que oferecem flexibilidade, escalabilidade e redução de custos de manutenção.



**Seu papel**

Você é membro da equipe de TI da empresa comercial. Sua responsabilidade é configurar a infraestrutura necessária para o novo projeto de LLM. Essa tarefa envolve projetar e implementar uma estrutura de TI robusta para dar suporte ao uso de IA para aprimorar a comunicação com o cliente.

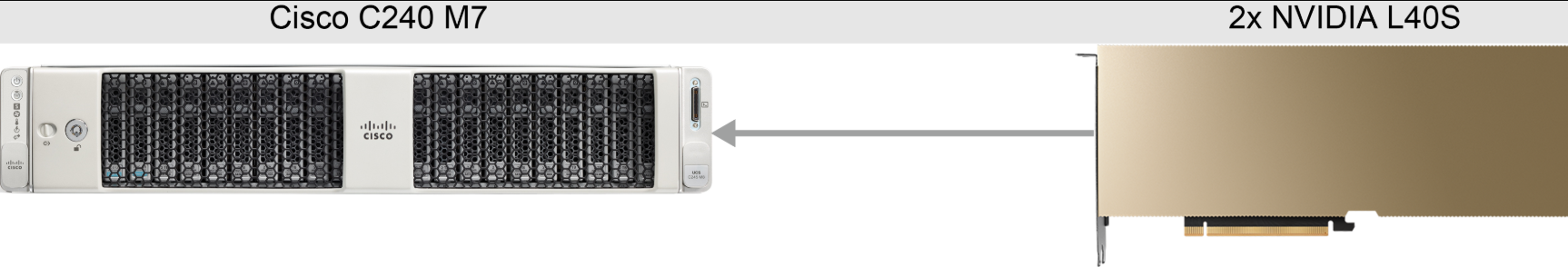
Você recomendará e supervisionará a integração da infraestrutura de rede necessária, garantindo que ela atenda aos requisitos do projeto. Sua função inclui avaliar diferentes opções de infraestrutura, identificar investimentos essenciais e propor estratégias para maximizar a eficiência e minimizar custos.

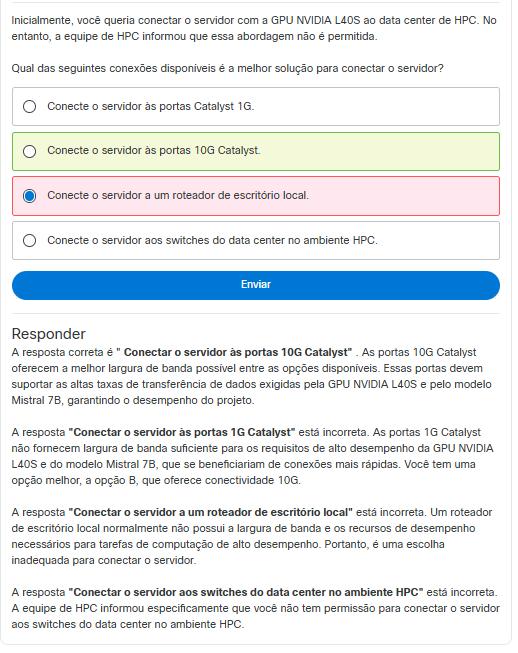
Sua experiência será essencial para ajudar sua empresa comercial a implementar com sucesso essa tecnologia avançada de IA, melhorando os processos de comunicação, aumentando a satisfação do cliente e mantendo uma vantagem competitiva no mercado.

**Atividade 1: Começando pequeno**

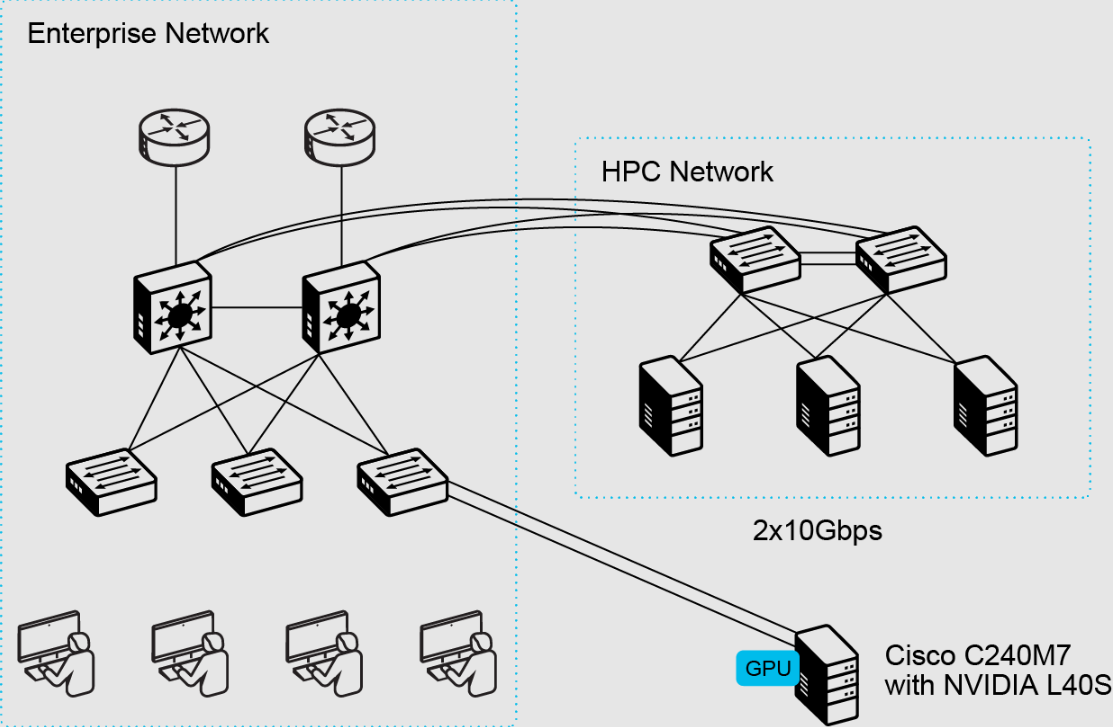
Sua equipe optou por usar o Mistral 7B Large Language Model para o projeto. Este modelo precisa de cerca de 16 GB de memória de GPU para funcionar bem em tarefas básicas. No entanto, para obter o melhor desempenho, especialmente em tarefas maiores, mais memória é útil. Se a equipe planeja treinar ou ajustar o modelo, pode precisar de cerca de 32 GB ou mais.

A equipe solicitou um servidor com GPUs NVIDIA L40S. A NVIDIA L40S possui 48 GB de memória de GPU, o que é mais do que suficiente para executar o modelo Mistral 7B sem problemas. A memória extra garantirá que o modelo funcione bem tanto para tarefas básicas quanto para tarefas mais complexas, como geração aumentada de recuperação (RAG) e ajuste fino. Portanto, a L40S é uma ótima opção para as necessidades atuais e futuras da equipe. Para hospedar essas GPUs, vocês escolheram o servidor de rack Cisco UCS C240 ​​M7.





Os usuários têm notado que o aplicativo LLM não está funcionando conforme o esperado. Os tempos de processamento são longos, mas, ao mesmo tempo, há subutilização dos recursos da GPU. Esse problema surge porque a GPU NVIDIA L40S, que é uma placa gráfica bastante potente, recomenda no mínimo uma Placa de Interface de Rede (NIC) de 2x25 Gbps para funcionar de forma ideal. No entanto, a conexão atual (como você deve se lembrar da pergunta anterior) é feita apenas por meio de um link de 2x10 Gbps, o que é insuficiente.



A boa notícia é que o servidor C240M7 está equipado com um Cisco VIC 15427, que possui quatro portas de 10/25/50 Gbps. Infelizmente, o pequeno rack onde o servidor está localizado possui apenas switches Catalyst com portas de 10 Gbps. Você identificou uma solução para conectar-se a portas de servidor de 25 Gbps em um pequeno rack localizado em uma sala vizinha, a 20 metros (aproximadamente 66 pés) de distância.

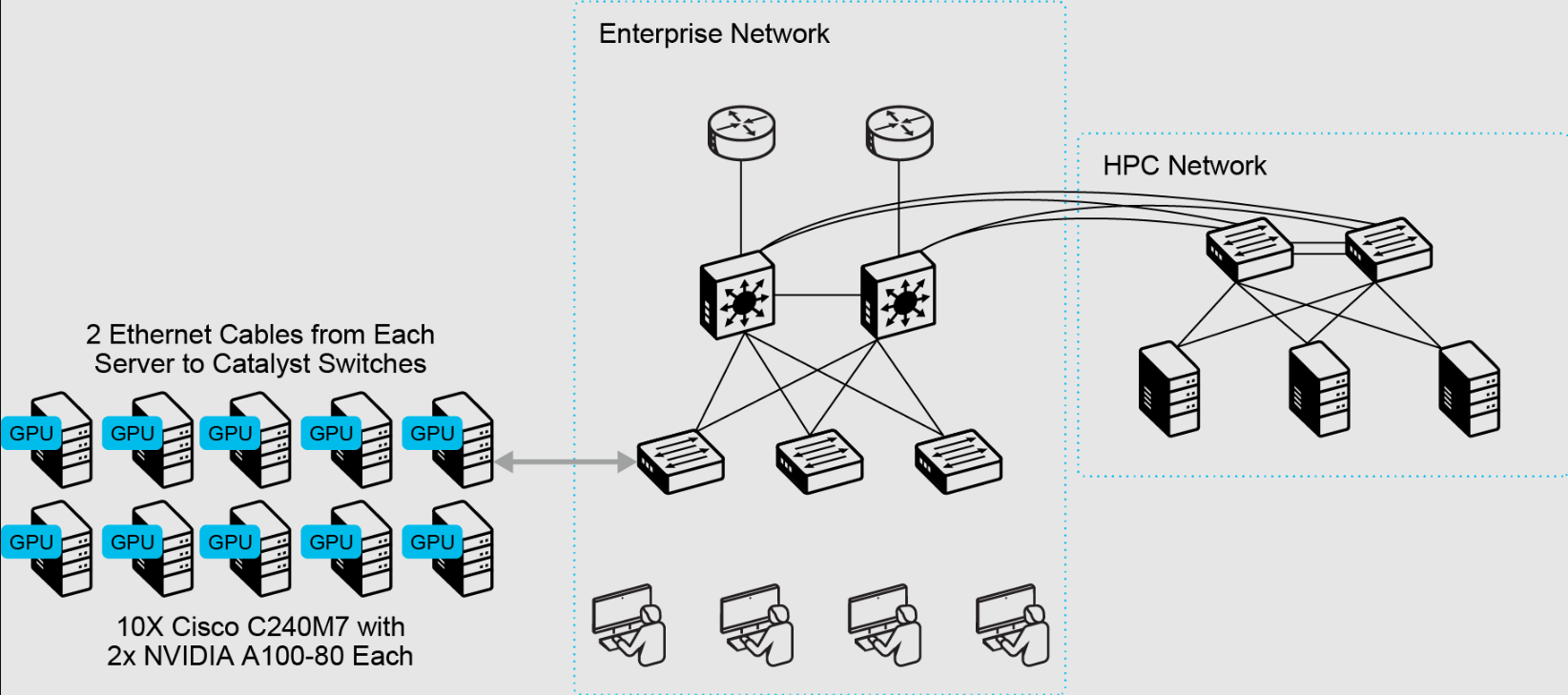
Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

**Atividade 2: Indo além de um servidor**

O projeto de IA tem sido um grande sucesso para sua empresa de trading. Inicialmente, a empresa utilizava LLMs para escrever e-mails e ofertas aos clientes, o que melhorou a eficiência e o atendimento ao cliente. Agora, eles identificaram casos de uso ainda maiores para o uso do LLM. Um novo uso importante está em suas operações de trading. O LLM analisa grandes quantidades de dados de mercado e fornece insights que ajudam os traders a tomar decisões mais inteligentes e rápidas, impulsionando suas estratégias de trading e lucros.

Graças a esse sucesso e aos novos casos de uso, sua empresa comercial está aumentando seu poder computacional. Sua tarefa é substituir o servidor existente por 10 servidores Cisco UCS C240 ​​M7 Rack, cada um com 2 potentes GPUs NVIDIA A100-80. Esses novos servidores fornecerão o poder de processamento necessário para lidar com mais tarefas e suportar o crescimento contínuo.



Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

O projeto está crescendo, e a equipe está até mesmo experimentando treinar LLMs na nova infraestrutura. Todo esse projeto de IA evoluiu rapidamente de uma pequena ferramenta útil para um sistema crítico para os negócios. No entanto, os usuários não estão satisfeitos com os problemas frequentes relacionados a timeouts e conectividade. Acontece que conectar todos esses servidores com GPUs está sobrecarregando a rede corporativa.

Uma rede de back-end separada e adequada é necessária para cargas de trabalho de IA. Seus colegas estão confusos sobre o que são essas redes de front-end e back-end.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

Texto

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário, Texto, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário, Texto, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

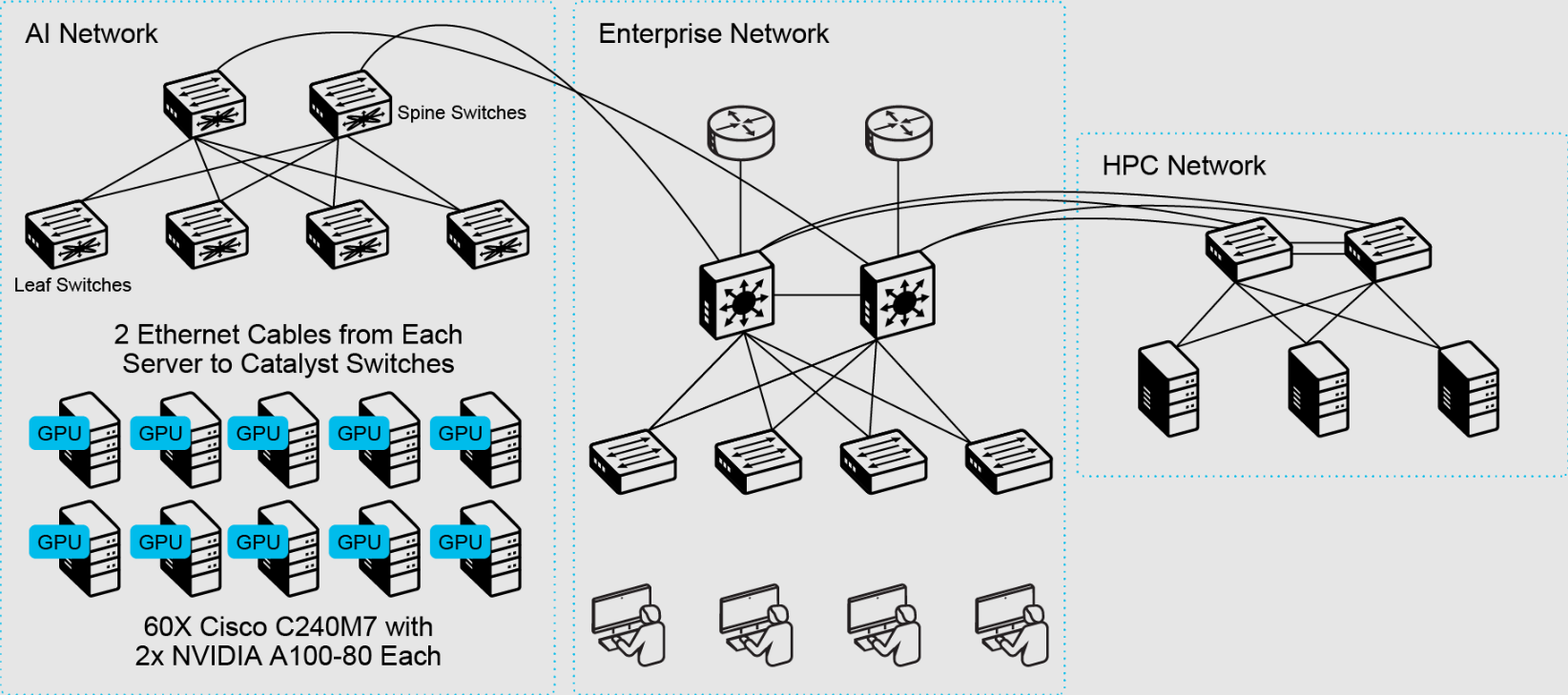
Interface gráfica do usuário, Texto, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário, Texto, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

Sua equipe expandiu recentemente seus esforços de treinamento em IA, exigindo a adição de 60 novos servidores à rede de back-end que gerencia as cargas de trabalho de IA. Apesar dessa expansão significativa, a equipe está enfrentando longos tempos de processamento para suas solicitações de treinamento.



Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

Texto

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário, Texto, Aplicativo

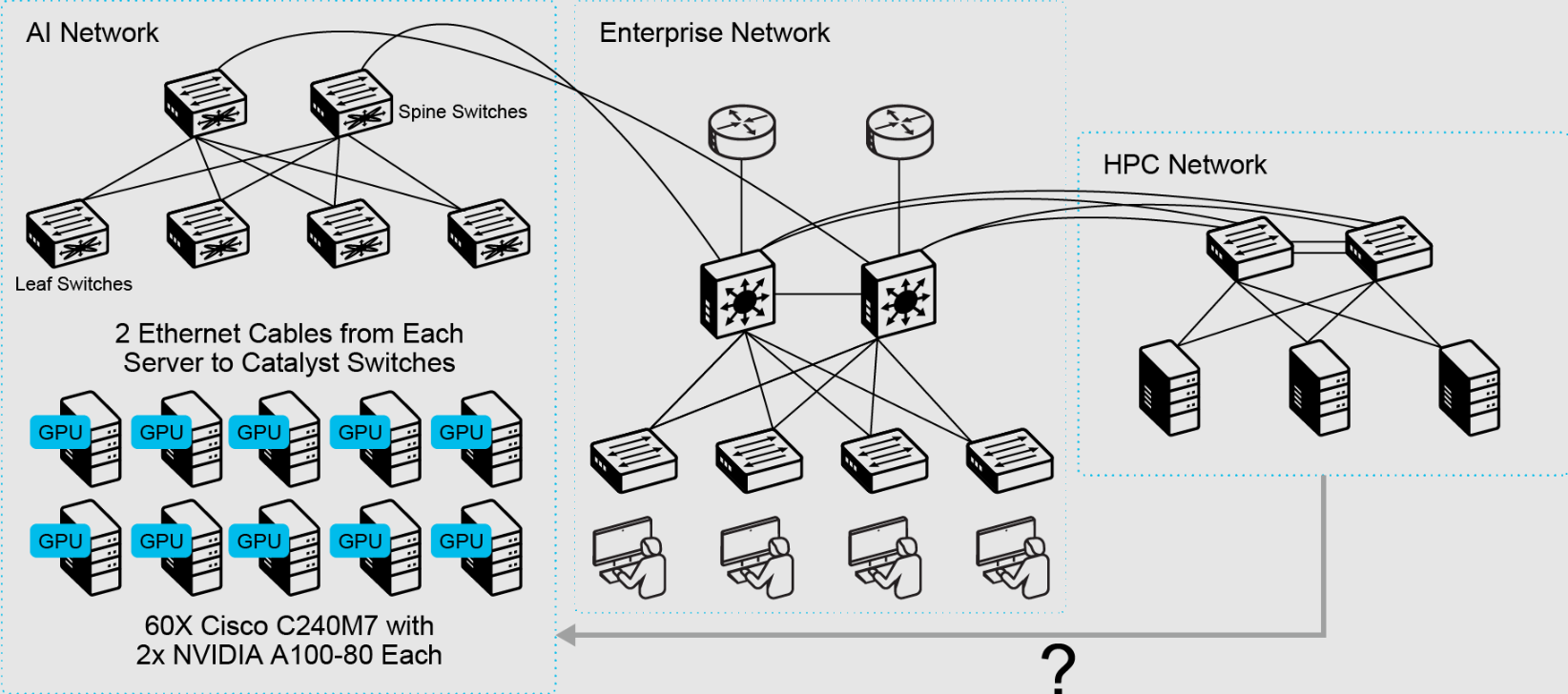
O conteúdo gerado por IA pode estar incorreto.

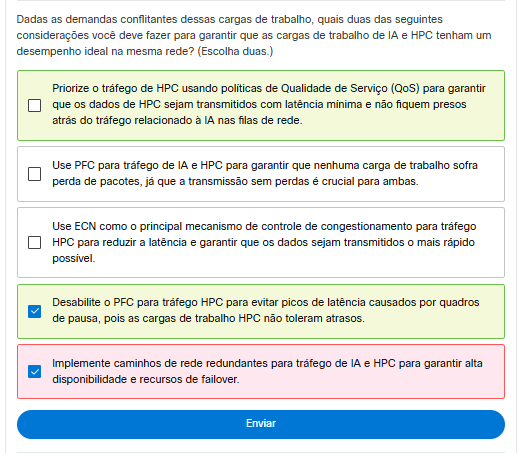
Texto

O conteúdo gerado por IA pode estar incorreto.

**Atividade 3: Considerações sobre o trânsito**

Sua empresa de trading agora conta com uma rede back-end totalmente funcional, otimizada para cargas de trabalho de IA, e tudo está funcionando perfeitamente. De fato, o desempenho tem sido tão impressionante que a equipe de HPC percebeu. Sua empresa observou que as velocidades e a taxa de transferência de dados na rede de IA são significativamente melhores do que as do próprio data center, que é bastante antigo e precisa de uma reforma. Impressionada com esse desempenho, a equipe de HPC está considerando migrar seus servidores para o data center de IA para aproveitar sua eficiência e recursos. No entanto, eles também estão preocupados com os potenciais riscos e desafios da integração de cargas de trabalho de HPC com cargas de trabalho de IA.





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.

Texto

O conteúdo gerado por IA pode estar incorreto.

Resumo

Os principais pontos de aprendizado desta atividade incluem a compreensão da configuração de rede para a transição para uma arquitetura de IA/ML, particularmente com GPUs avançadas para um projeto de LLM. Você aprenderá a importância de começar pequeno e escalar, usando as conexões apropriadas. A atividade enfatiza a necessidade de uma rede de IA dedicada para evitar sobrecarregar a rede corporativa e os benefícios de separar as redes front-end e back-end. Você explorará as compensações entre Ethernet e InfiniBand para cargas de trabalho de IA, com foco em largura de banda, custo e latência. Além disso, este laboratório de artigos aborda o uso de ECN e PFC em ambientes RoCEv2 para gerenciar congestionamentos e garantir comunicação sem perdas. Por fim, a integração de IA com cargas de trabalho não relacionadas a IA requer priorização cuidadosa da rede e gerenciamento de latência.