**Protocolos de nível de aplicação**

*Introdução*

Neste treinamento, você aprenderá sobre os mecanismos e operações de Remote Direct Memory Access (RDMA) e RDMA over Converged Ethernet (RoCE) e como as infraestruturas de inteligência artificial (IA) são construídas e mantidas.

Imagine um laboratório de pesquisa movimentado onde cientistas de dados trabalham no desenvolvimento de um modelo avançado de IA para prever mudanças climáticas. O modelo requer a análise de terabytes de dados meteorológicos coletados por sensores ao redor do mundo. Inicialmente, a equipe enfrenta atrasos e gargalos significativos devido ao grande volume de dados que precisa ser processado e transferido entre servidores. Os métodos tradicionais de transferência de dados não conseguem acompanhar o ritmo, fazendo com que o tempo de treinamento se estenda por dias ou semanas.

O RDMA permite que os dados sejam transferidos diretamente da memória de um computador para outro sem envolver a CPU, o que reduz significativamente a latência e melhora a taxa de transferência. O RoCE estende esses recursos às redes Ethernet, tornando-o ideal para ambientes de data center modernos.

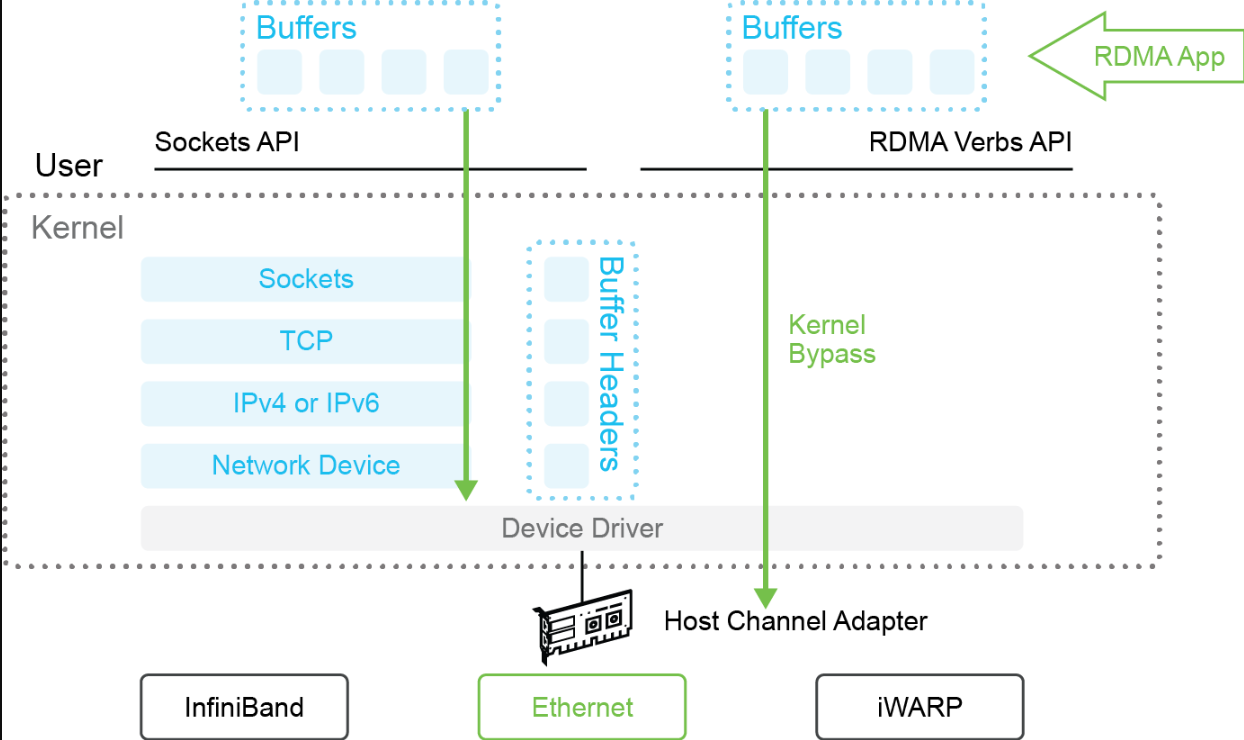
Ao implementar os protocolos RDMA e RoCE, o laboratório de pesquisa transforma seu pipeline de processamento de dados. Os dados agora fluem suavemente entre os nós, reduzindo o tempo de transferência de horas para minutos. Essa melhoria drástica permite que a equipe treine seu modelo de IA com muito mais rapidez, acelerando sua pesquisa e permitindo atualizações mais frequentes do modelo com novos dados. A eficiência obtida economiza tempo e reduz os custos computacionais, tornando o projeto mais sustentável.

Este exemplo destaca como os protocolos RDMA e RoCE são vitais para otimizar o processamento e o manuseio de dados em sistemas distribuídos, o que é essencial para treinar com eficiência modelos de IA em larga escala.

**Fundamentos do RDMA**

O RDMA tem sido tradicionalmente utilizado em ambientes de computação de alto desempenho (HPC) e redes de armazenamento. As operações RDMA são úteis em clusters de computadores extremamente paralelos, como clusters de inteligência artificial e aprendizado de máquina (IA/ML), onde as unidades de processamento gráfico (GPUs) nos nós remotos podem acessar diretamente a memória umas das outras (GPU Direct RDMA). O RDMA tradicionalmente requer uma rede InfiniBand dedicada e específica, com seus próprios adaptadores de rede e switches. No entanto, também existem implementações de RDMA sobre Ethernet e redes de transporte IP.

Em um caminho de dados de rede tradicional, o sistema operacional controla totalmente o hardware, e o aplicativo utiliza serviços do sistema operacional, como chamadas de sistema, para acessar o hardware. Especificamente, o processo do aplicativo no espaço do usuário utiliza buffers de memória por meio de chamadas de interface de programação de aplicativos (API) de soquetes. No kernel, o caminho de dados inclui a pilha TCP/IP até o driver do dispositivo do adaptador de rede e, eventualmente, a malha da rede. Esse processamento de movimentação de dados orientado por software consome muitos ciclos de CPU, o que aumenta significativamente a latência e não atinge alta taxa de transferência.



Com o RDMA, o kernel é ignorado, o que evita a dispendiosa sobrecarga da CPU associada a trocas de contexto e cópia de dados entre o usuário e o kernel (operação de cópia zero). O RDMA permite que os aplicativos realizem transferências diretas de dados de memória para memória entre nós remotos na rede sem sobrecarregar a CPU. A função de transferência de dados é transferida para um adaptador de rede especializado, o que permite que as operações regulares da CPU continuem em paralelo. Assim, o aplicativo acessa diretamente os buffers na placa de interface de rede (NIC) compatível com RDMA e troca dados com os buffers no host remoto sem envolver os sistemas operacionais de ambos os lados.

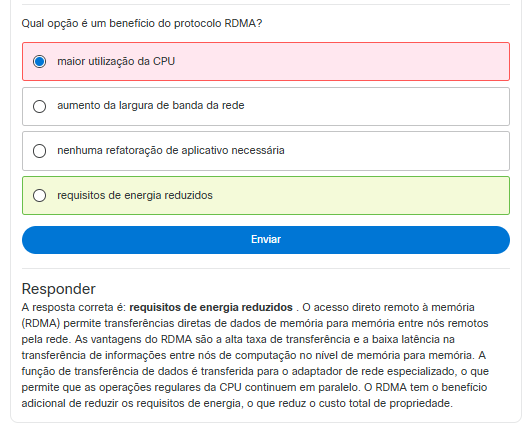
**Observação**

*Neste treinamento, o foco está em aplicações de IA/ML executadas no espaço do usuário e que trocam dados diretamente pela rede. Outras aplicações RDMA comuns são processos no espaço do kernel, como um sistema de arquivos trocando dados com um sistema de armazenamento remoto.*

Diagrama

O conteúdo gerado por IA pode estar incorreto.

A transferência de dados permite a transferência de informações de alta taxa de transferência e baixa latência entre nós de computação no nível de memória para memória. O RDMA tem o benefício adicional de reduzir os requisitos de energia, o que reduz o custo total de propriedade.



**Arquitetura RDMA**

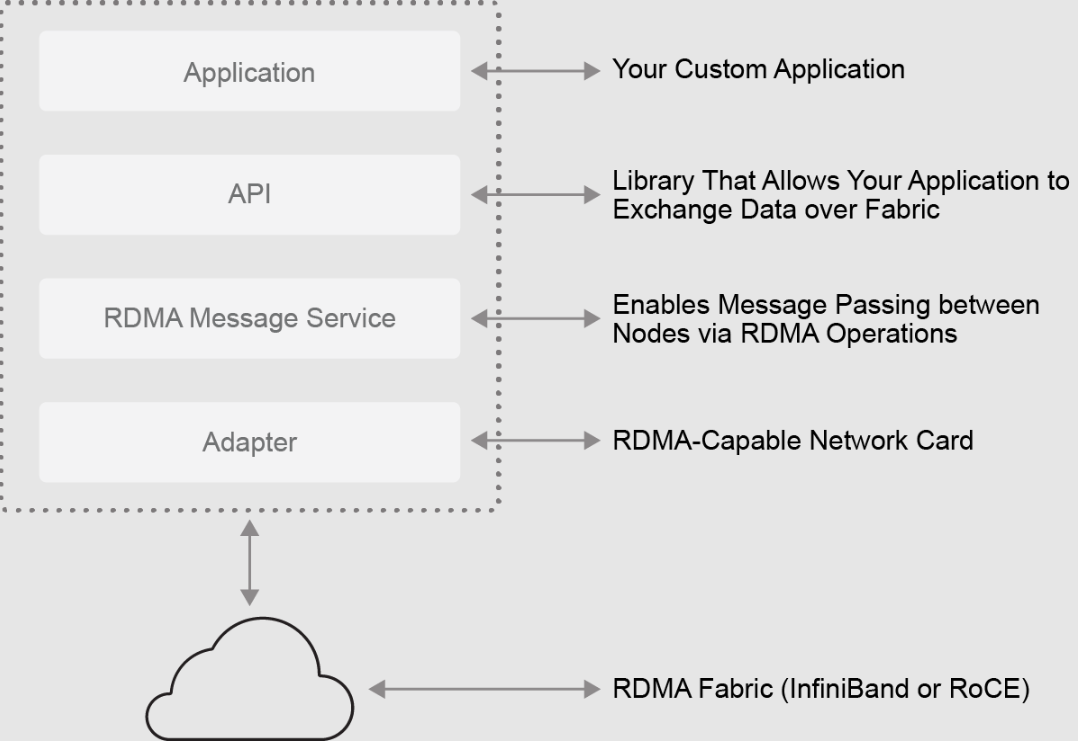
O RDMA tem um objetivo simples: permitir que os aplicativos se comuniquem diretamente pela rede. Esse objetivo é alcançado por aplicativos que utilizam o serviço de mensagens RDMA por meio de uma API. O resultado é que um canal virtual é estabelecido internamente entre os dois aplicativos remotos — entre dois espaços de endereço de memória de aplicativo separados. Cada extremidade do canal virtual possui um par de filas de trabalho, que é mapeado para o espaço de memória do aplicativo. Um par de filas de trabalho consiste em dois tipos de filas de comunicação: uma fila de envio e uma fila de recebimento. Além disso, há uma fila de conclusão que rastreia a conclusão das solicitações de trabalho. Essas filas permitem que os aplicativos troquem mensagens diretamente, enviando solicitações de trabalho, que são efetivamente operações RDMA.

**Observação**

*Cada aplicativo pode configurar várias conexões, e cada conexão terá seu próprio par de filas de trabalho dedicado.*

Para habilitar esse mecanismo de comunicação, a pilha de arquitetura RDMA possui os seguintes componentes, ilustrados na figura a seguir:

* **API** : A API permite que um aplicativo utilize RDMA por meio do serviço de mensagens RDMA. A API RDMA implementa os métodos e operações apropriados que os aplicativos usam para acessar serviços RDMA. Os métodos especificam os comportamentos necessários das APIs, conforme definido pelo padrão InfiniBand. Essa abordagem garante a compatibilidade entre diferentes sistemas operacionais e fornecedores de hardware.
* **Serviço de mensagens RDMA** : O serviço de mensagens fornece acesso ao hardware RDMA por meio de pares de filas de trabalho e operações RDMA realizadas sobre as filas. Em vez de usar serviços do sistema operacional, os aplicativos utilizam o serviço de mensagens RDMA para trocar dados diretamente. Na prática, o serviço de mensagens abstrai o funcionamento interno do RDMA e oculta as complexidades do RDMA de um aplicativo.
* **Pares de filas de trabalho RDMA** : Cada par de filas consiste em uma fila de envio (para enviar dados a um nó remoto) e uma fila de recebimento (para receber dados de um nó remoto). Dependendo do tipo de operação necessária, o aplicativo envia a solicitação de trabalho apropriada para uma fila de envio ou recebimento. Essa ação aciona o serviço de mensagens RDMA para transportar a mensagem no canal. Após a conclusão do processamento, uma entrada é colocada na fila de conclusão, que notifica o aplicativo de que a operação foi concluída.
* **Solicitações de trabalho RDMA (operações)** : A operação RDMA é um método de transferência de mensagens. Duas semânticas diferentes são suportadas — semântica de memória e semântica de canal — com seu próprio conjunto de operações suportadas.
* **Adaptador de rede compatível com RDMA** : O adaptador de rede InfiniBand ou Ethernet implementa RDMA. O adaptador de rede executa as operações solicitadas transmitindo a mensagem para um host remoto. O adaptador de rede também executa outras tarefas de controle necessárias para estabelecer e manter um canal de comunicação físico.
* **Interconexão de rede** : esta rede compreende switches e cabeamento (InfiniBand ou Ethernet)



Interface gráfica do usuário, Aplicativo

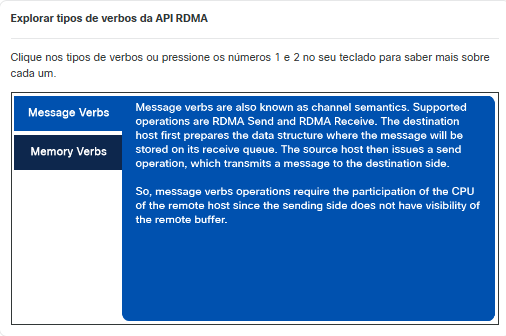
O conteúdo gerado por IA pode estar incorreto.

**Operações RDMA**

Uma aplicação RDMA se comunica diretamente com o adaptador de rede compatível com RDMA usando a API baseada em verbos RDMA. Diferentemente do processamento de fluxo baseado em bytes usado pelo TCP, o RDMA utiliza transações baseadas em mensagens para transferir dados entre diferentes hosts. Cada mensagem tem um tamanho variável e é uma unidade básica de trabalho em um domínio RDMA.

**Operações RDMA**

Uma API RDMA tem dois tipos de verbos, que podem ser considerados estilos de comunicação, ilustrados na figura a seguir.



Interface gráfica do usuário, Texto, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

Tabela

O conteúdo gerado por IA pode estar incorreto.

Diferenças importantes entre as duas semânticas de comunicação são mostradas na tabela a seguir.

| Característica | Semântica de Canal (Verbos de Mensagem) | Semântica da Memória (Verbos de Memória) |
| --- | --- | --- |
| Confiabilidade | Há suporte para serviços de transporte confiáveis ​​e não confiáveis. | Somente transporte confiável |
| Modo síncrono ou assíncrono | Síncrono | Assíncrono |
| Caso de uso | Transferir mensagens de controle curtas | Transferências de dados em massa |

**Observação**

*As características de transferência confiável do RDMA incluem ordem correta de entrega, transmissão sem erros e integridade de dados.*

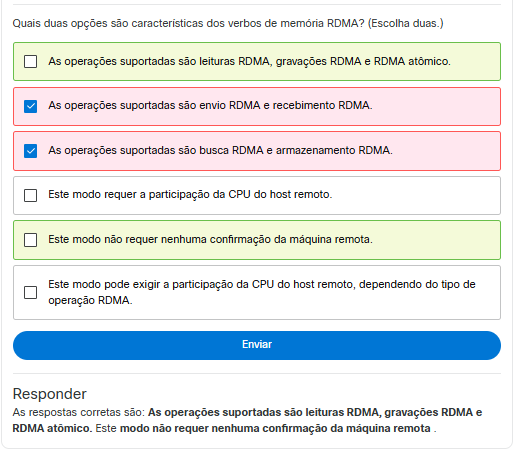
Os dois estilos de comunicação se complementam e são usados ​​em conjunto para diferentes casos de uso. Para cargas de trabalho de IA/ML, operações de leitura e gravação RDMA semânticas de memória são usadas para permitir transferências de dados de baixa latência e alto rendimento entre GPUs remotas no cluster de IA/ML. Os principais recursos e benefícios que tornam o modelo de operações baseado em mensagens RDMA altamente eficiente para cargas de trabalho de IA/ML e HPC incluem o seguinte:

* Suporte para milhares a milhões de pares de filas de trabalho.
* Registro de memória de buffers de espaço do usuário com o adaptador de rede.
* Capacidade de associar um buffer registrado a vários pares de filas, como entradas de fila de trabalho que podem fazer referência a quaisquer buffers registrados.
* Suporte a entradas de dispersão ou coleta; o RDMA funciona nativamente com múltiplas entradas de dispersão ou coleta, como ler vários buffers de memória e enviá-los como um bloco ou obter um bloco e gravá-lo em vários buffers de memória.
* Operações adicionais para InfiniBand, como Atomic Fetch & Add e Compare & Swap, e um mecanismo para sinalizar e recuperar os resultados de conclusão de operações de fila de trabalho enviadas anteriormente.

Em resumo, todas as características do modelo RDMA mencionadas e a eficiência das operações RDMA permitem que os aplicativos de IA/ML sejam dimensionados e aproveitem os recursos de cluster de IA/ML de alta largura de banda de maneira compartilhada e multilocatária.

**Observação**

*Operações de dispersão e coleta manipulam com eficiência dados descontínuos distribuídos por diferentes regiões da memória. Esse recurso é muito útil para transferir estruturas de dados complexas ou grandes conjuntos de dados.*

**

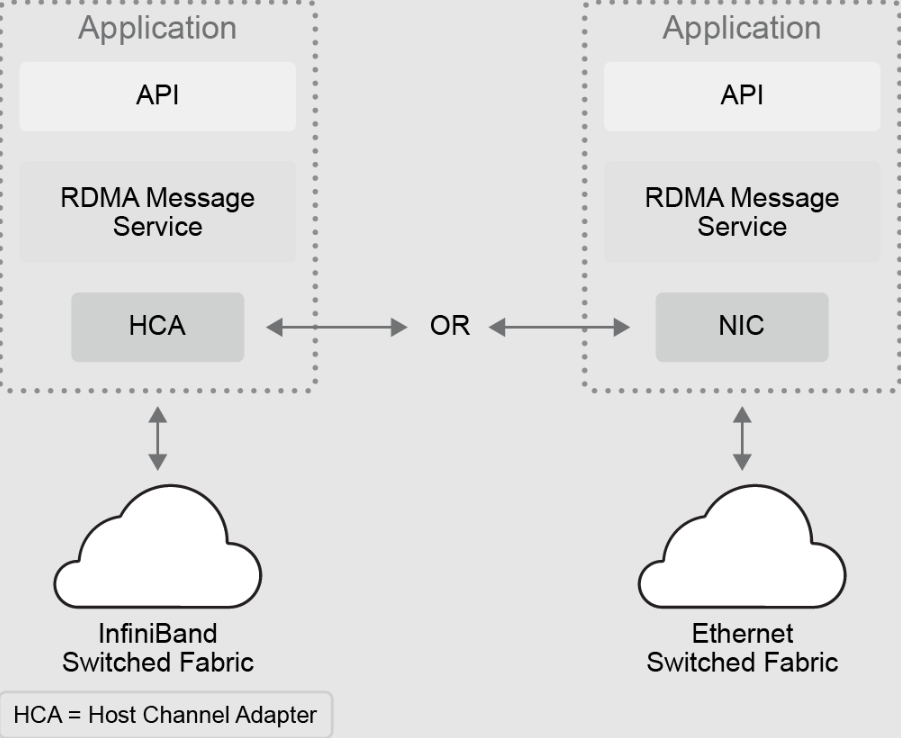
**RDMA sobre Ethernet convergente**

Em sua primeira implementação, a InfiniBand Trade Association (IBTA) trouxe todos os benefícios do RDMA ao mercado, utilizando a rede InfiniBand para fornecer alta taxa de transferência, bypass de CPU e menor latência. O InfiniBand também incorporou o gerenciamento de congestionamento ao protocolo. Embora esses benefícios tenham tornado o InfiniBand o transporte HPC preferido, ele exigia uma rede InfiniBand personalizada e dedicada. Essas redes, especialmente projetadas, trouxeram custos e complexidade adicionais para a empresa.

Como alternativa ao InfiniBand, a especificação IBTA descreve o uso de uma rede Ethernet para transporte, conhecida como RDMA sobre Ethernet Convergente (RoCE). Isso efetivamente sobrepõe os protocolos RDMA e o serviço de mensagens ao transporte Ethernet, permitindo que você aproveite todos os benefícios da arquitetura RDMA nos data centers onde a Ethernet já está presente.

A Ethernet é projetada com perdas, o que não é ideal para executar cargas de trabalho de IA/ML e HPC, pois a perda de pacotes causa degradação do desempenho. Aproveitando as inovações Ethernet mais recentes, como extensões de ponte de data center (DCB) que incluem controle de fluxo prioritário (PFC) e notificação explícita de congestionamento (ECN), você pode ter as mesmas garantias de entrega que o InfiniBand nativo.

Essas inovações efetivamente oferecem os mesmos benefícios do RDMA, como transferências de dados RDMA, bypass eficiente da pilha do kernel e outros recursos que não são possíveis no TCP sobre Ethernet tradicional. O RoCE oferece a maioria dos benefícios da arquitetura InfiniBand, mas o faz sobre Ethernet sem exigir uma atualização significativa.



O RoCE versão 1 (RoCEv1) é um padrão IBTA introduzido em 2010 e funciona apenas no mesmo domínio de transmissão da Camada 2. O protocolo RoCEv1 é um protocolo da camada de enlace Ethernet com EtherType 0x8915. O RoCE versão 2 (RoCEv2) foi introduzido em 2014 e permite o roteamento de tráfego por estruturas IP. O RoCEv2 encapsula o transporte InfiniBand em cabeçalhos Ethernet, IP e UDP para roteamento por redes Ethernet. O protocolo RoCEv2 existe sobre o protocolo UDP-IPv4 ou UDP-IPv6. A porta de destino UDP número 4791 foi reservada para o RoCEv2. Portanto, a estrutura interna do pacote entre os dois protocolos é diferente, conforme ilustrado na figura a seguir.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

Portanto, embora RoCEv1 e RoCEv2 ofereçam os mesmos benefícios do ponto de vista da aplicação, RoCEv2 permite projetos de rede muito mais flexíveis. RoCEv2 é considerado um padrão nos data centers atuais, portanto, sempre que RoCE é mencionado, geralmente se refere a RoCEv2. RoCEv1 e RoCEv2 são comparados na tabela a seguir.

| RoCEv1 | RoCEv2 |
| --- | --- |
| Opera na Camada 2 (Ethernet) e requer uma estrutura Ethernet sem perdas. | Opera na Camada 3 (IP), permitindo que funcione em redes roteadas. |
| Alcança transmissão sem perdas usando técnicas DCB. | Usa encapsulamento UDP, permitindo que o tráfego RoCEv2 atravesse redes IP. |
| Normalmente limitado a um único domínio de transmissão Ethernet devido à sua natureza de Camada 2. | Mais escalável que o RoCEv1 devido à sua capacidade de usar roteamento IP. |

O RoCEv1 é compatível com os adaptadores Cisco UCS Virtual Interface Card (VIC) série 1300 mais antigos. O RoCEv2 é compatível com os adaptadores Cisco VIC séries 1400 e 15000 mais recentes. O RoCEv2 não é compatível com versões anteriores do RoCEv1.

**Observação**

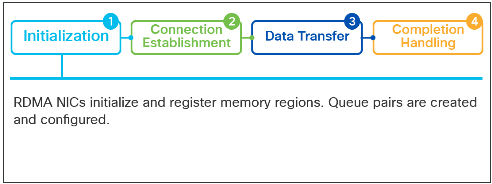
*Para que o RoCE funcione corretamente, é crucial configurar as NICs e o sistema operacional corretamente.*

**Fluxo de trabalho RoCE**

Primeiro, a malha Ethernet precisa ser configurada corretamente, incluindo a configuração de uma VLAN dedicada, a configuração da qualidade de serviço (QoS), os recursos do DCB nos switches e a instalação e o ajuste do driver e das bibliotecas do sistema operacional da placa de rede. Somente depois que a malha estiver totalmente operacional, os fluxos de trabalho RoCE poderão ser estabelecidos. A configuração do fluxo de trabalho RoCE é transparente porque é realizada automaticamente, mas envolve muitas etapas, como registrar a memória, estabelecer conexões, lidar com as conclusões e, eventualmente, encerrar as conexões.

**Explore o fluxo de trabalho do RoCE**

O fluxo típico de comunicação do RoCE tem vários estágios: clique no estágio para descobrir o que acontece durante cada um.



Uma imagem contendo Texto

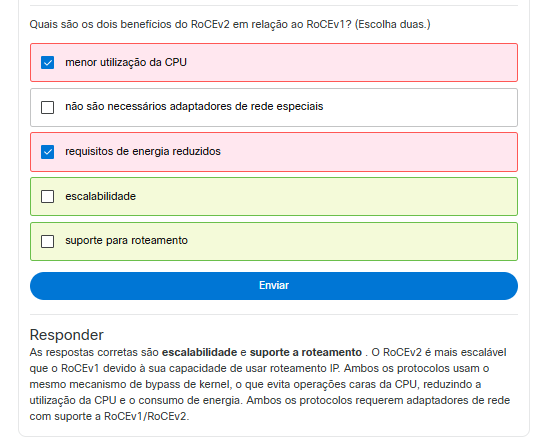
O conteúdo gerado por IA pode estar incorreto.

Texto

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.



**Resumo**

Parabéns por concluir o treinamento *em Protocolos em Nível de Aplicação* ! Agora você entende a tecnologia RDMA, incluindo sua arquitetura e operações, e aprecia as melhorias de desempenho que o RDMA oferece ao ignorar o kernel do sistema operacional e permitir trocas diretas de memória entre aplicativos. Além disso, você tem um sólido domínio das diversas operações da API RDMA, o que o ajuda a adaptar essas tecnologias para atender às necessidades específicas do seu ambiente de aplicativos e maximizar as vantagens do RDMA.

Além disso, você explorou como aproveitar a tecnologia Ethernet amplamente disponível para construir estruturas RoCE, que tradicionalmente dependem de redes InfiniBand. Ao entender como integrar RDMA com Ethernet, você estará bem equipado para implementar essas tecnologias avançadas em diversos ambientes de rede.

Agora que você concluiu este treinamento, considere as seguintes questões:

* Quais cenários específicos em seu ambiente poderiam se beneficiar do RDMA?
* Como você pode usar o RoCE para melhorar a eficiência?