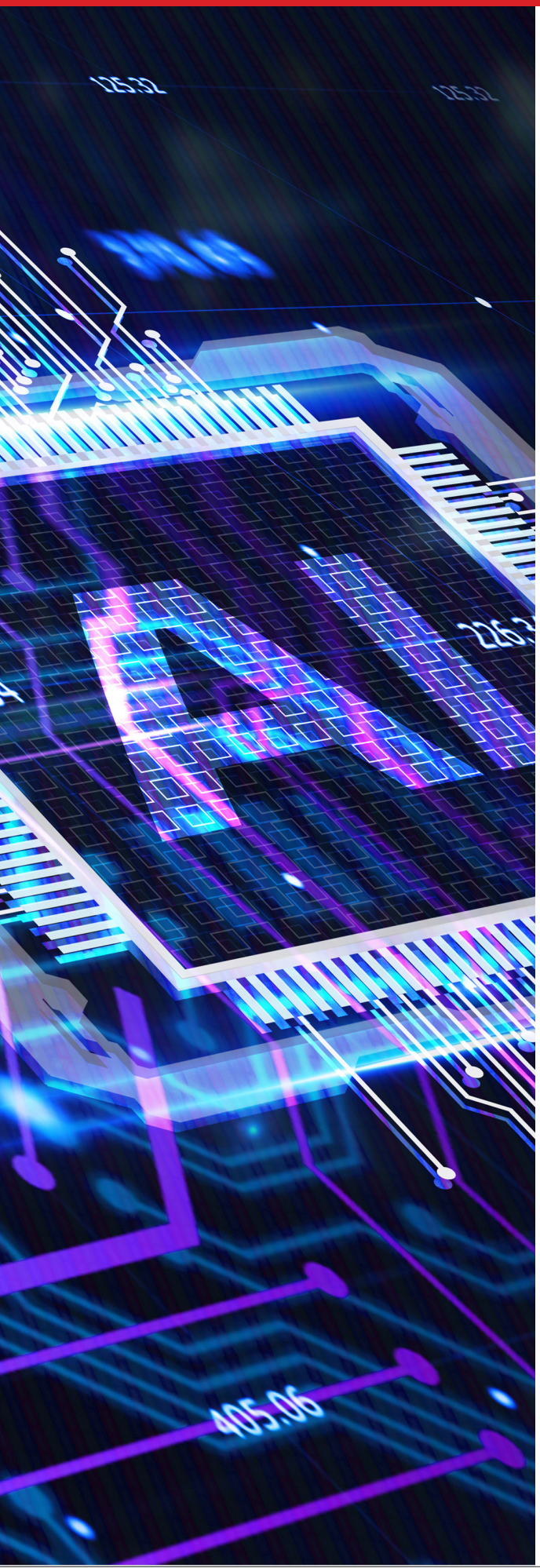


Enfrentando os desafios das cargas de trabalho de IA com o portfólio de IA da Dell

Uma comparação do portfólio de IA da Dell com ofertas semelhantes da HPE

Não há dúvida de que a inteligência artificial (IA) transformou o ambiente empresarial e permitiu que setores e organizações de todos os portes obtivessem insights mais profundos dos dados, automatizassem processos de negócios, entregassem experiências personalizadas para os clientes e usuários e competissem melhor no próprio setor. Para aproveitar de maneira eficaz o poder da IA, as organizações precisam de um provedor de infraestrutura que possa oferecer um portfólio abrangente de soluções integradas que englobe todo o ciclo de vida da IA.

Para ajudar os clientes a lidar com as crescentes demandas de IA e navegar pelas complexidades inerentes, existem os fornecedores de infraestrutura como a Dell Technologies e a Hewlett Packard Enterprise (HPE). Com portfólios prontos para IA, esses fornecedores oferecem diferentes níveis de soluções de IA, que agregam soluções de infraestrutura de alto desempenho no local e na nuvem, com parcerias estratégicas e um menu de serviços de suporte e consultoria.

Este relatório analisa informações disponíveis publicamente sobre os portfólios de IA da Dell e da HPE, com o objetivo de destacar as vantagens específicas de arquitetura, desempenho e suporte que podem beneficiar os clientes que escolherem a Dell Technologies para suprir as necessidades de IA. Comparamos informações detalhadas sobre os servidores criados pela Dell para suportar implementações de IA e consultamos os resultados de testes de referência de desempenho do setor da ML Commons®. Também exploramos outras ofertas de software e serviço que atendem aos clientes em cada estágio da jornada de IA.

*Nota: todas as pesquisas da PT foram concluídas até o dia 5 de dezembro de 2023. Portanto, este documento não refletirá ofertas ou alterações nas versões da Dell ou da HPE após essa data.

Desafios da adoção da IA

A adoção de uma estratégia de IA apresenta muitos desafios novos para os datacenters e para os profissionais de TI que os gerenciam, incluindo:

- Lidar com a falta de habilidades da equipe atual por meio de treinamento interno em IA ou contratação externa.
- Compreender as necessidades de preparação de dados da IA, incluindo a qualidade, quantidade, localização e estado atual dos dados dos negócios.
- Avaliar as metas específicas de IA dos negócios para determinar melhor quais modelos e implementações de IA apresentarão benefícios.
- Avaliar as necessidades computacionais, de rede e de armazenamento dos sistemas planejados de IA e determinar um plano de aquisição.

Esses são apenas alguns exemplos e, muitas vezes, obstáculos significativos que uma empresa enfrentará ao buscar extrair os benefícios da implementação da IA nos datacenters.

O portfólio de IA da Dell busca ajudar os clientes a enfrentar esses desafios por meio de serviços profissionais e de consultoria. Esses serviços auxiliam os clientes na criação de roteiros de implementação e na preparação dos dados para modelos de IA.[1] O portfólio também inclui cursos de treinamento que abrangem conceitos de aprendizado de máquina (ML) e outros tópicos educacionais, além de oferecer designs validados para IA que ajudam a garantir o sucesso da implementação.[2] Além disso, a Dell tem parceria com terceiros para oferecer aos clientes ferramentas adicionais de IA, como um portal personalizado da Dell dentro da comunidade Hugging Face com scripts e contêineres dedicados para implementação de modelos de IA de código aberto[3] e fácil implementação do grande modelo de linguagem (LLM) Meta Llama 2.[4] Juntamente com uma grande seleção de ofertas de computação e PC, desde workstations móveis a servidores que permitem até oito GPUs NVIDIA de ponta, a Dell também disponibiliza o armazenamento de dados não estruturados que a IA requer, por meio de um portfólio de storage arrays em arquivo e em objeto de alto desempenho. Essas ofertas de armazenamento, incluindo o Dell PowerScale, o ObjectScale, o ECS e o armazenamento integrado, podem lidar com os dados não estruturados que as cargas de trabalho de IA empregam com frequência.[5] A Dell também fez uma parceria com a Snowflake para oferecer uma solução de armazenamento em nuvem híbrida para os clientes.[6] De acordo com a análise da Dell, de agosto de 2023, a empresa oferece "o portfólio de IA generativa mais amplo", avançando para além de servidores e armazenamento e disponibilizando recursos durante toda a jornada de implementação da IA.[7]

Desempenho de IA e opções de computação acelerada: Dell x HPE

As cargas de trabalho de IA podem usar CPUs, GPUs ou ambas como recursos computacionais, dependendo do tamanho ou tipo de carga de trabalho. Algumas CPUs oferecem aceleradores específicos de IA, como Intel Advanced Matrix Extensions (Intel AMX) nos mais recentes processadores escaláveis Intel Xeon.[8] As GPUs geralmente são melhores para cargas de trabalho maiores e/ou mais complexas, mas o formato da GPU pode afetar os níveis de desempenho. Por exemplo, algumas GPUs dos modelos NVIDIA A100 e H100 vêm no formato PCIe universal ou SXM exclusivo. Esse último usa a arquitetura NVIDIA SXM de alto desempenho.[9] Grandes capacidades de memória e recursos de design de servidor, como arquitetura de resfriamento e eficiência no consumo de energia, também afetam o desempenho. A maioria dos datacenters ainda usa resfriamento a ar, ou seja, as cargas de trabalho de computação com alto desempenho (HPC) precisam de servidores projetados para resfriar com ar da maneira mais eficaz possível. Abaixo, destacamos as ofertas de servidor PowerEdge em termos de componentes, opções de resfriamento, entre outras, além das pontuações publicadas pelo MLPerf® da MLCommons®.

Referência de desempenho do modelo de IA: comparação de resultados do MLPerf

O MLPerf® é um conjunto de referências de desempenho que testa o desempenho de IA, tanto para treinamento quanto para inferência. Para que uma organização publique resultados oficiais do MLPerf®, os resultados precisam estar em conformidade com condições específicas, definidas pela desenvolvedora da referência de desempenho, a MLCommons®.[10] Essas diretrizes de conformidade apresentam padrões que facilitam a comparação do desempenho. Para os testes de inferência, o MLPerf® usa os conjuntos de dados "Datacenter", "Edge", "Mobile" e "Tiny" e relata as pontuações de IA e os watts de energia consumida durante os testes. O conjunto de referências de desempenho de inferência inclui testes para diversos modelos comuns de IA, ML e DL. Consulte a Tabela 1.

Tabela 1: Modelos de IA, ML e DL incluídos nos testes do MLPerf® e casos de uso típicos para cada um. Fonte: Principled Technologies.

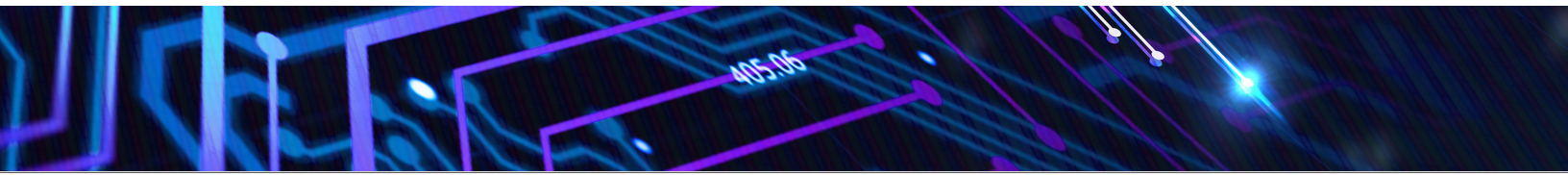
Modelos comuns de IA	Casos de uso comuns
ResNet	Um modelo de classificação de imagens que ajuda os computadores a aprender, lembrar e identificar imagens diferentes para casos de uso como imagiologia médica, moderação de conteúdo de mídias sociais e reconhecimento facial
RetinaNet	Um tipo de detecção de objetos que pode lidar com complexidades adicionais, em comparação com o ResNet. Ele ajuda os computadores a identificar e localizar objetos dentro de imagens ou frames de vídeo e pode classificá-los por importância. Para casos de uso como direção autônoma, tecnologia de assistência automática para veículos, monitoramento de segurança e reconhecimento facial
3D-UNet	Específico para segmentação de imagens médicas
RNN-T	Reconhecimento de voz para casos de uso como tradução automática de idiomas
BERT	Processamento de linguagem natural para casos de uso como resumo de texto, tradução de idiomas e realização automática de tarefas
DLRM-v2-99.9	Modelo de recomendação para casos de uso como anúncios direcionados e recomendações personalizadas de produtos
GPTJ-99 e 99.9	LLM para processamento de linguagem natural, que se destaca na geração de texto. Para casos de uso como chatbots e ferramentas de IA baseadas em chat

MLPerf

Os resultados do MLPerf® incluem vários parâmetros, além dos próprios modelos de IA, o que pode fazer com que muitos dados sejam analisados em um só gráfico ou em uma só tabela. Veja abaixo uma referência rápida sobre esses parâmetros:

- "99.0" e "99.9": esses números referem-se à precisão com a qual o modelo foi treinado. Quanto maior a necessidade de precisão do resultado, mais complexo será o modelo e mais tempo ele poderá levar para processar os dados.
- "Offline samples/sec": modo em que a referência de desempenho envia todas as consultas no início do teste, simulando dados já presentes no sistema.
- "Server queries/sec": modo em que a referência de desempenho envia consultas durante todo o teste, simulando a análise de um fluxo de dados em tempo real.

Para obter mais informações sobre a MLCommons® e os resultados do MLPerf®, consulte https://mlcommons.org/benchmarks/inference-datacenter/.



As informações neste relatório são provenientes dos resultados de novembro de 2023 do MLPerf® v3.1 Inference Datacenter, publicados no site da MLCommons®.[11] Esses resultados incluem envios de fabricantes de tecnologia e provedores de serviços em nuvem e abrangem diversas configurações. Em comparação com os envios publicamente disponíveis da HPE, os servidores Dell produziram melhores resultados em determinados modelos de IA. (Nota: diferentes configurações de GPU nos servidores podem dificultar comparações diretas.) Consulte a Tabela 2 para obter detalhes.

Tabela 2: Servidores Dell e HPE incluídos nos resultados do MLPerf® 3.1 da MLCommons®, publicados em 29/11/23. Fonte: Principled Technologies.

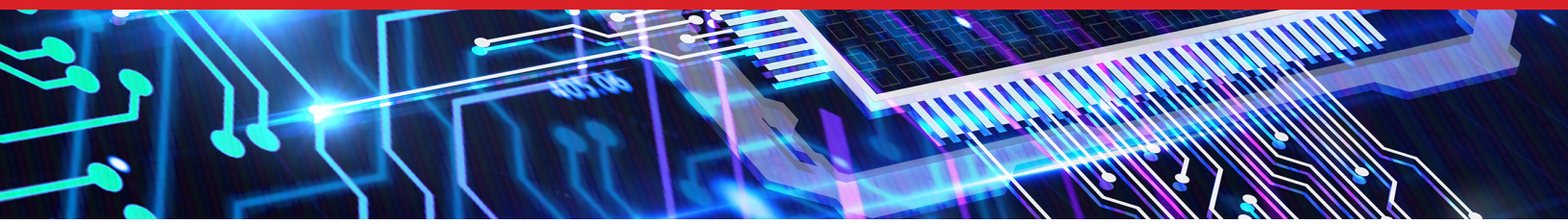
Remetente	Modelo de servidor	N° e modelo de GPUs	Descrição
Dell[12]	PowerEdge XE9680	8 NVIDIA H100 SXM	Para inferência e treinamento de IA com amplas cargas de trabalho, como grandes modelos de linguagem
	PowerEdge XE9640	4 NVIDIA H100 SXM	Para treinamento de grandes modelos de IA em datacenters de alta densidade e resfriados por líquido
	PowerEdge XE8640	4 NVIDIA H100 SXM	Para orientar aplicativos tradicionais de treinamento de IA, HPC e lógica analítica de dados, em um formato 4U, para datacenters resfriados a ar
	PowerEdge R760xa	4 NVIDIA H100 PCIe	Para uma ampla variedade de cargas de trabalho com uso computacional intenso, incluindo inferência e treinamento em IA-ML/DL, que não exijam GPUs de alto desempenho
HPE[13,14]	ProLiant XL675d Gen10 Plus	8 NVIDIA A100 SXM	Para computação com alto desempenho e IA
	ProLiant DL380a Gen11	4 NVIDIA H100 PCIe	Servidor 2U para cargas de trabalho moderadas de IA

Comparação direta entre servidores Dell e HPE

Embora em uma estratégia abrangente de IA haja mais do que apenas hardware, garantir o melhor desempenho do hardware é um dos fatores mais vitais para o sucesso das cargas de trabalho de IA. Conforme novas GPUs e outras tecnologias são disponibilizadas, os recursos de carga de trabalho de IA também evoluem. Na época em que os resultados do MLPerf® v3.1 foram publicados pela primeira vez, a melhor GPU da NVIDIA disponível era a H100 Tensor Core, com a qual a Dell publicou resultados do MLPerf® em vários de seus servidores, tanto nos formatos PCIe quanto SXM5.[15] Os resultados publicados da HPE incluíram apenas um envio da H100 e somente com o formato PCIe. Nossa pesquisa mostrou que poucos dos servidores HPE disponíveis com capacidade para GPU eram compatíveis com o formato SXM5 da H100 para entregar o melhor desempenho da GPU NVIDIA, e que nenhum dos servidores HPE ProLiant era compatível.[16] Conforme mostrado abaixo, ter GPUs melhores geralmente aprimora o desempenho das cargas de trabalho de IA.

Resultados do MLPerf para oito GPUs

O Dell PowerEdge XE9680 oferece suporte para até oito GPUs NVIDIA H100 SXM5 para aceleração de IA e para até dois processadores escaláveis Intel® Xeon® de 4ª geração. A família de produtos PowerEdge XE tem uma arquitetura modular, compatível com as GPUs NVIDIA SXM4 ou SXM5 ou com os conjuntos de GPU Open Compute Project Accelerator Module (OAM) da AMD, que podem aumentar o desempenho em comparação com uma GPU PCIe padrão.[17] Ocupando apenas 6U de espaço em rack, o PowerEdge XE9680 é um compacto servidor NVIDIA H100 SXM5 de oito vias. Atualmente, os mais recentes servidores HPE ProLiant Gen11 não são compatíveis com o formato SXM da H100,[18] embora alguns servidores de supercomputação HPE Cray sejam.[19] Como a HPE não enviou qualquer resultado do MLPerf® com os servidores Cray e destaca apenas os servidores ProLiant na página do portfólio de IA, nós abordaremos somente os servidores ProLiant neste documento. (Consulte a Figura 1.)



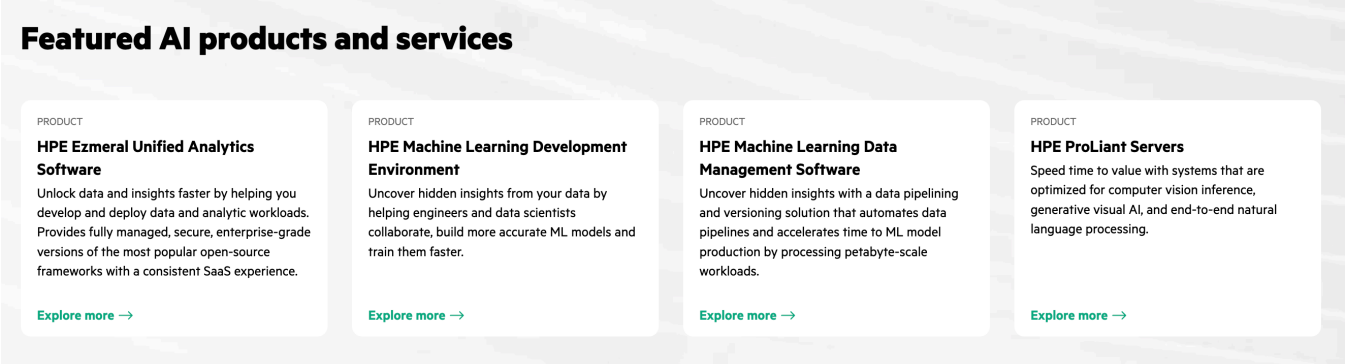


Figura 1: Captura de tela dos produtos e serviços de IA em destaque, em https://www.hpe.com/us/en/solutions/ai-artificial-intelligence.html, realçando os servidores HPE ProLiant em 05/12/2023.

Nos resultados do MLPerf® v3.1 publicados pela primeira vez em novembro de 2023 para servidores com oito GPUs, o Dell PowerEdge XE9680 com GPUs NVIDIA H100 SXM5 superou em cerca de 4,25x o HPE ProLiant XL675d Gen10 Plus com GPUs NVIDIA A100 SXM4 (consulte a Figura 2).

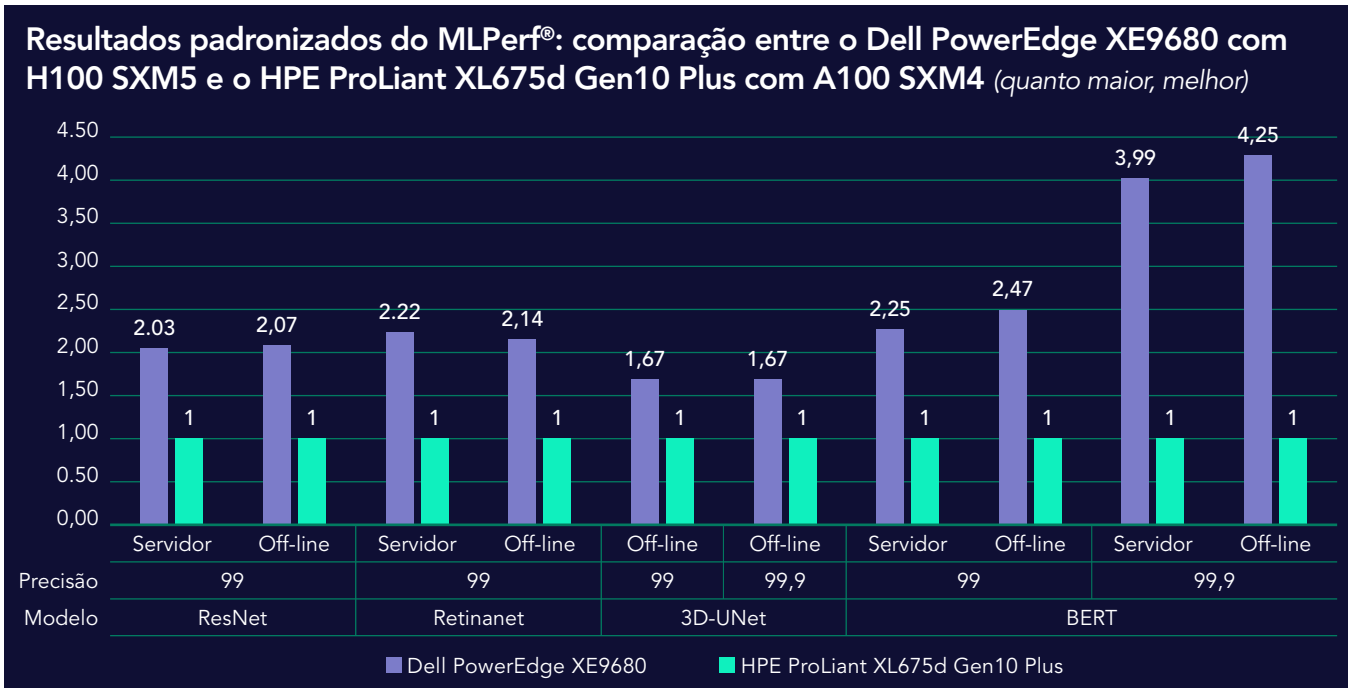


Figura 2: Resultados do MLPerf® publicados para o Dell PowerEdge XE9680 e o HPE ProLiant XL675d Gen10 Plus em 29/11/23. O sistema Dell usa GPUs NVIDIA H100, enquanto o sistema HPE usa GPUs da geração anterior. Fonte: Principled Technologies com dados da MLCommons®.[20,21]

Para facilitar a comparação, padronizamos os resultados dos testes nas Figuras 2 a 5. Isso significa que atribuímos o valor de 1 para cada resultado do HPE ProLiant DL380a Gen 11 e mostramos o resultado correspondente do Dell PowerEdge R760xa em relação a ele. Como esses resultados mostram, até mesmo uma geração de diferença entre os modelos de GPU pode mudar significativamente o desempenho esperado em uma infinidade de cargas de trabalho de IA.

Resultados do MLPerf para quatro GPUs

Quando a principal preocupação é a economia de energia ou a economia de espaço do datacenter, o Dell PowerEdge XE9640 2U pode ser a solução. Com até quatro GPUs NVIDIA H100 SXM, o PowerEdgeXE9640 oferece metade da potência computacional da GPU do XE9680, em dois terços do espaço.[22] O Dell PowerEdge XE9640, densamente compactado, incorpora a tecnologia Dell Smart Cooling, disponibilizando uma variedade de tecnologias térmicas, incluindo o resfriamento líquido direto para CPUs e GPUs.[23]

O chassi 2U do PowerEdge XE9640 acomoda mecanismos aprimorados de fluxo de ar, incluindo dissipadores de calor e ventiladores maiores, para ajudar a resfriar os outros componentes vitais, como placas PCIe e memória.[24] Atualmente, o PowerEdge XE9640 é a única oferta, seja da Dell ou da HPE, que vem com GPUs H100 HGX de quatro vias em 2U. O portfólio de IA da HPE oferece servidores ProLiant Gen11 em 1U e 2U, mas eles se limitam a GPUs de formato PCIe.[25]

O servidor Dell PowerEdge XE9640 também é compatível com GPUs OAM Intel Max série 1550, uma opção de GPU de baixa potência e alta densidade que inclui uma placa PCIe e um Open Compute Accelerator Module (OAM).[26] Embora não possamos confirmar que, em 05/12/23, a HPE oferecesse um servidor com essas GPUs, podemos confirmar que ela oferece o HPE ProLiant DL380 Gen11 e o DL380a Gen11 com GPUs Intel Data Center Max 1100.[27] Isso significa que o Dell PowerEdge XE9640 pode ser a única oferta atual com quatro GPUs OAM Intel Max 1550 em um servidor 2U. Para empresas que se preocupam com o espaço do datacenter e a eficiência no consumo de energia, um servidor 2U com quatro GPUs Intel Max 1550 oferece uma solução que combina computação com alto desempenho e eficiência no consumo de energia, sem sacrificar o espaço do data center.

O Dell PowerEdge XE9640 com quatro GPUs H100 HGX superou em cerca de 1,99x o HPE ProLiant DL380a com quatro GPUs H100 PCIe, nos resultados publicados do MLPerf® 3.1 (consulte a Figura 3).

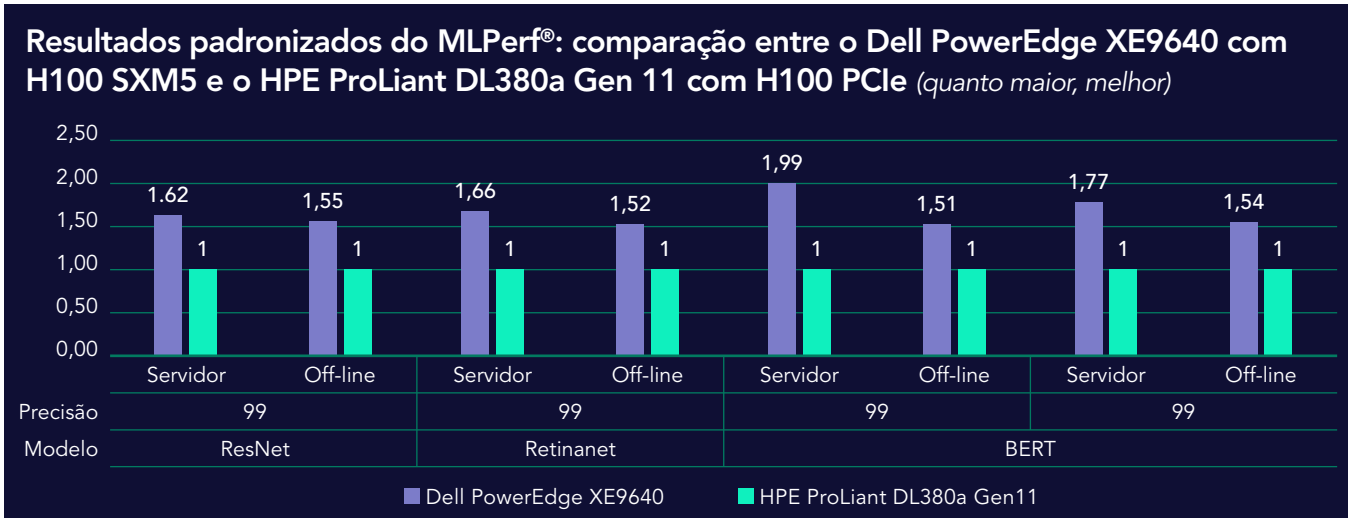


Figura 3: Resultados do MLPerf® publicados para o Dell PowerEdge XE9680 e o HPE ProLiant XL675d Gen10 Plus em 29/11/23. O sistema Dell usa GPUs NVIDIA H100, enquanto o sistema HPE usa GPUs da geração anterior. Fonte: Principled Technologies com dados da MLCommons®.[28,29]

O PowerEdge XE8640 oferece uma configuração de GPU de quatro vias com resfriamento a ar para processadores e um radiador de resfriamento a ar auxiliado por líquido para GPUs, que não exige disponibilidade de água para rack nas instalações. Para aqueles que não usam ou não podem usar refrigerante externo,[30] o Dell PowerEdge XE8640 4U permite quatro GPUs NVIDIA H100 SXM5, entregando a mesma potência computacional que o PowerEdge XE9640, sem a necessidade de resfriamento líquido direto.[31]

O Dell PowerEdge XE8640 apresenta os mais recentes processadores escaláveis Intel Xeon de 4ª geração e até 4 TB de memória[32] para lidar com grandes conjuntos de dados e cálculos complexos, comuns em IA e lógica analítica de dados. Novamente, a HPE oferece as GPUs NVIDIA H100 SXM5 em sistemas HPE Cray, mas os servidores HPE ProLiant habilitados para GPU não são compatíveis com elas.

Na comparação dos dados do MLPerf® publicados em novembro de 2023, o servidor PowerEdge XE8640 com quatro GPUs NVIDIA H100 SXM5 atingiu a maior taxa de transferência de IA entre todos os envios de quatro GPUs, em nove categorias diferentes. Conforme mostrado na Figura 4, em comparação com o servidor HPE ProLiant DL380a, o servidor Dell atingiu uma pontuação até 2,07 vezes maior.

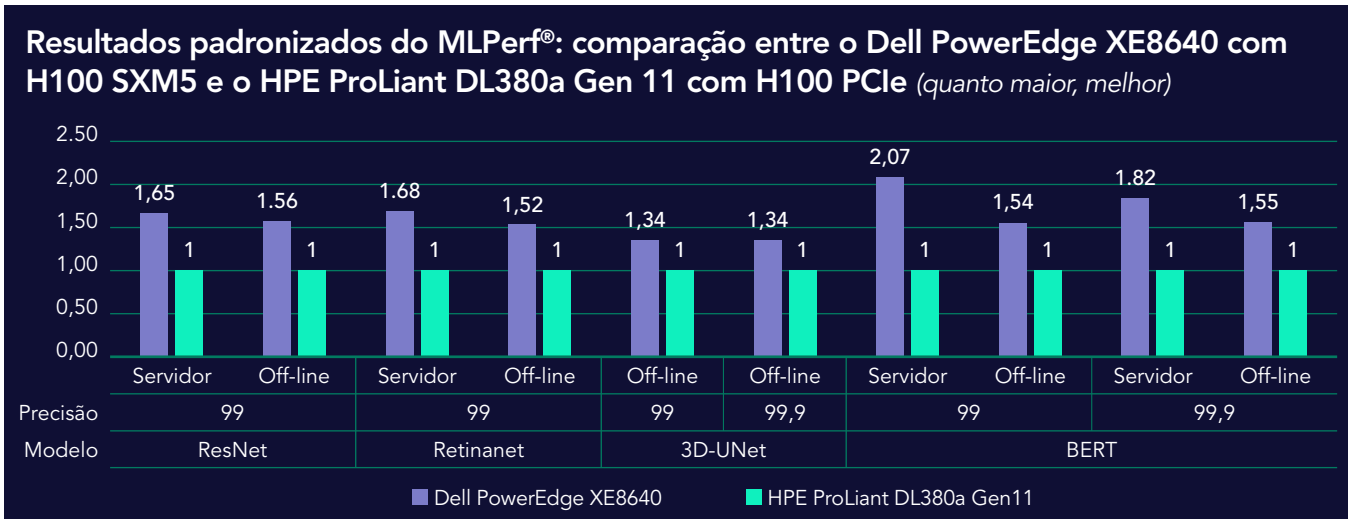


Figura 4: Resultados do MLPerf® publicados para o Dell PowerEdge XE8640 e o HPE ProLiant Dl380a Genn11 em 29/11/23. O sistema Dell usa o formato NVIDIA H100 SXM, enquanto o sistema HPE usa o formato PCIe, menos potente. Fonte: Principled Technologies com dados da MLCommons®.[33,34]

Por fim, para organizações que desejam começar com pouco e ampliar conforme necessário, o servidor Dell PowerEdge R760xa 2U aceita uma variedade de GPUs NVIDIA, AMD e Intel, com suporte para até quatro GPUs PCIe de 5ª geração de largura dupla ou 12 GPUs PCIe de largura única.[35] Ele apresenta 32 slots DIMM, um compartimento de oito unidades para discos de 2,5 polegadas e 12 slots PCIe, oferecendo armazenamento escalável que pode ser ampliado de acordo com as necessidades de dados de IA cada vez maiores e suporte para até 12 GPUs PCIe de largura única ou quatro GPUs PCIe de largura dupla, como a NVIDIA H100 ou L40S.[36] Essa escalabilidade significa que o servidor pode se adaptar às tarefas de IA em evolução, desde o treinamento de modelos de aprendizado de máquina até o processamento avançado de dados.

O sistema de resfriamento a ar do PowerEdge R760xa é compatível com ambientes computacionais de alta densidade e pode acomodar aceleradores com maior potência de design térmico (TDP) de até 350 W,[37] uma capacidade que pode ajudar o servidor a manter o desempenho sob cargas intensas de computação. Nos resultados dos testes ResNet, RetinaNet e BERT do MLPerf® usando o modo "Servidor", publicados em novembro de 2023, o PowerEdge R760xa com quatro GPUs NVIDIA H100 PCIe superou o HPE ProLiant Dl380a Gen 11, também equipado com quatro GPUs H100 PCIe (consulte a Figura 5).

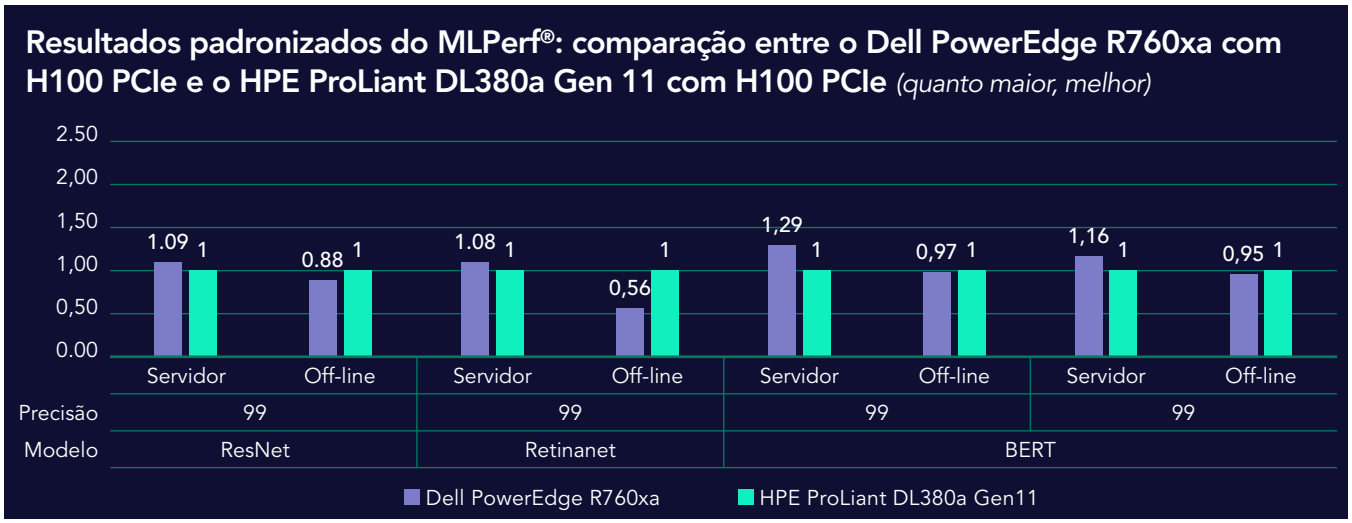


Figura 5: Resultados do MLPerf® publicados para o Dell PowerEdge R760xa e o HPE ProLiant Dl380a Genn11 em 29/11/23. Os dois sistemas usam GPUs NVIDIA H100 no formato PCIe. Fonte: Principled Technologies com dados da MLCommons®.[38,39]

No geral, os resultados do MLPerf® mostram que o desempenho varia amplamente entre servidores e componentes e, portanto, é essencial selecionar as opções certas para oferecer suporte às suas cargas de trabalho e às demandas de desempenho delas. Os servidores Dell PowerEdge para cargas de trabalho de IA oferecem várias opções de resfriamento e densidade para atender a quaisquer necessidades de datacenter que uma empresa possa ter e, ao mesmo tempo, possibilitar um desempenho sólido no MLPerf®.

Cobertura mais detalhada do portfólio de IA da Dell

Embora crucial, o desempenho computacional é apenas uma das considerações durante o planejamento de cargas de trabalho de IA. Você também precisa considerar o restante do portfólio de IA de um fornecedor ao iniciar uma implementação de IA. A seguir, discutiremos categorias adicionais essenciais para esses portfólios de IA, incluindo workstations client, produtos nativos da nuvem, armazenamento e muito mais. Também destacaremos áreas em que as ofertas da Dell podem proporcionar uma vantagem em comparação com as da HPE.

Workstations

Para desenvolvedores de IA e cientistas de dados, as workstations Dell Precision Data Science oferecem GPUs NVIDIA RTX™ e CPUs Intel Xeon®, além de um conjunto de ferramentas de ciência de dados.[40] Esses sistemas aproveitam as opções de computação de nível profissional com as GPUs NVIDIA certificadas para mais de 100 aplicativos profissionais[41] e os aceleradores do processador escalável Intel Xeon, como o Intel DL Boost.[42] As workstations Precision estão disponíveis nos formatos móvel, torre e rack para atender a necessidades que vão desde análises maiores e estacionárias de dados até modelagens de campo científico em movimento.

As ofertas de workstations da HPE são mais limitadas, apresentando principalmente workstations individuais em torre equipadas com NVIDIA L4; a HPE não oferece opções de workstation móvel.[43] Embora adequadas para muitas tarefas, as ofertas de workstations em torre não disponibilizam a mesma flexibilidade e cobertura de carga de trabalho que a gama mais ampla oferecida pela Dell. A variedade de opções de tamanho e portabilidade das workstations Dell Precision permite soluções mais personalizadas, acomodando as diferentes necessidades em ambientes como laboratórios, escritórios e operações em campo.

Armazenamento

O armazenamento pode ser tão vital quanto a computação para executar cargas de trabalho de IA. Quanto maior a quantidade de dados, melhor é a precisão dos modelos de IA. No entanto, o armazenamento e o gerenciamento de grandes conjuntos de dados podem ser um desafio para grande parte dos recursos dos datacenters. Além disso, como os modelos geralmente são treinados usando dados não estruturados, os sistemas de armazenamento prontos para IA precisam manusear muitos tipos de dados diferentes com facilidade.[44] Para oferecer capacidade e dimensionamento para conjuntos de dados de IA, ML e DL, a Dell oferece a série PowerScale™ para armazenamento em arquivo e o Elastic Cloud Storage (ECS) ou ObjectScale definido por software para armazenamento em objeto.

O portfólio NAS All-Flash do PowerScale oferece opções de capacidade bruta por nó de 3,84 TB até 720 TB e opções All-Flash em cluster com capacidade bruta de até 186 PB. A flexibilidade e a escala do PowerScale podem aceitar uma ampla variedade de clientes e casos de uso de IA.[45] Quando implementado em cluster, o PowerScale F900 pode atingir até 186 PB de armazenamento bruto total.[46] Todos os três modelos do PowerScale All-Flash (F200, F600 e F900) incluem desduplicação e compactação de dados em linha para melhorar a eficiência de armazenamento.[47] Cada modelo de armazenamento do PowerScale usa o file system Dell OneFS™, que emprega políticas para classificar o armazenamento em níveis a fim de priorizar os dados mais importantes nos níveis mais rápidos para otimização da carga de trabalho.[48] A Dell também oferece o software OneFS no AWS Marketplace com o Dell APEX File Storage for AWS. Os clientes podem aproveitar o OneFS com suas instâncias de computação da AWS para obter uma experiência do usuário consistente com os mesmos recursos disponíveis nos arrays locais do OneFS.[49] Embora a HPE ofereça integração de nuvem pública para soluções de armazenamento híbrido, nós não encontramos uma opção nativa da nuvem, como o Dell APEX File Storage for AWS, entre as ofertas da empresa.

As opções de armazenamento em objeto da Dell incluem o Dell Enterprise Object Storage (ECS), que foi "criado especificamente para armazenar dados não estruturados em escala de nuvem pública".[50] Juntamente com a compatibilidade integrada com o armazenamento em objeto do Amazon S3 para funcionalidade de nuvem híbrida, os nós de armazenamento do ECS oferecem capacidades de até 14 PB por rack.[51] A HPE também oferece armazenamento não estruturado com opções de armazenamento em arquivo e em objeto, embora sua oferta de armazenamento em objeto seja por meio de uma parceria com a Scality. Os clientes podem comprar soluções HPE para Scality com a HPE.[52]

Serviços profissionais

A Dell disponibiliza uma ampla gama de serviços profissionais, incluindo consultoria, preparação de dados, implementação e suporte, além de serviços gerenciados para oferecer suporte às implementações de IA. No caso de organizações que buscam arquiteturas e soluções validadas, a Dell oferece designs validados para IA, que visam a casos de uso específicos para eliminar as suposições durante o desenvolvimento e a implementação de recursos de IA. Essas soluções de IA validadas pela Dell incluem pacotes de hardware e software, modelos de IA conversacional, operações de aprendizado de máquina e muito mais. Ao combinar soluções pré-configuradas e desenvolvidas para fins específicos com serviços relacionados à IA, a Dell oferece uma solução de IA abrangente para atender a todas as necessidades de IA. Essas ofertas podem viabilizar um caminho mais rápido e fácil para o sucesso da IA, em comparação com a criação de soluções pontuais.

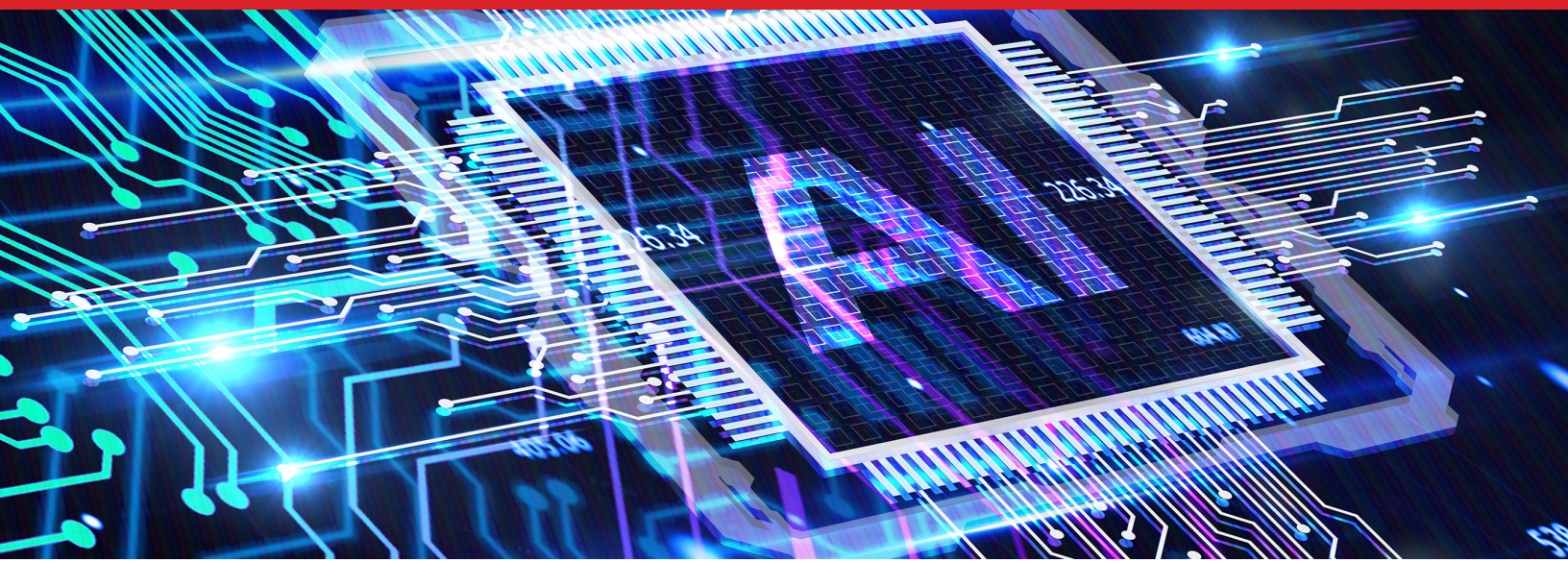
Os serviços Dell também podem orientar sua jornada de IA, desde a consultoria até a implementação. A Dell ProConsult Advisory Services ajuda os clientes a identificar onde os usuários podem obter benefícios com a adoção dos processos de GenAI e a criar um roteiro que inclua as soluções e as habilidades de TI necessárias. Os serviços Dell podem preparar dados para a integração de grandes modelos de linguagem e treinar as equipes de TI para que elas tenham conhecimento sobre IA. Para a adoção completa da GenAI, as equipes da Dell revisam seus casos de uso específicos e determinam, implementam e configuram o melhor modelo de IA para atender às suas necessidades. A HPE também oferece serviços profissionais para dar suporte às empresas em seus esforços de IA.[53,54]

Considerações sobre gerenciamento

Os servidores exigem gerenciamento contínuo, o que ocupa o tempo dos administradores. Os firmwares, softwares e drivers exigem atualizações periódicas, e a equipe de TI precisa otimizar, manter o desempenho, controlar as temperaturas e muito mais. Em testes anteriores da Principled Technologies (PT), nós avaliamos os recursos de gerenciamento dos servidores Dell com o Integrated Dell Remote Access Controller 9 (iDRAC9).[55] Ao usar o OpenManage™ Enterprise (OME) do iDRAC9 com agendamento configurável, os administradores podem contar com atualizações on-line automatizadas. Dessa forma, eles podem manter os servidores atualizados e usar perfis para integrar novos servidores com rapidez e facilidade à medida que as cargas de trabalho vão aumentando. Com o iDRAC e o OME, os clientes da Dell podem acessar mais recursos de gerenciamento remoto, implementar servidores com mais facilidade e atualizar o firmware de maneira mais simples do que se estivessem usando o HPE OneView e o HPE iLO. Os servidores Dell PowerEdge acompanham o gerenciamento e os serviços da Dell que podem ajudar as organizações "reduzindo o tempo e o esforço em tarefas como monitorar a integridade do sistema ou atualizar o firmware", o que libera ciclos de TI para inovação e outras tarefas.[56]

Tabela 3: Resumo da comparação entre as ferramentas de gerenciamento da Dell e da HPE, em um relatório da PT de novembro de 2022.[57] Fonte: Principled Technologies.

	Qual é o diferencial das ferramentas de gerenciamento da Dell?	O quanto é melhor?
Mais recursos de gerenciamento remoto iDRAC x iLO	Mais recursos de configuração do BIOS e do console HTML5 para uma funcionalidade mais remota no iDRAC	2,5x os recursos do console HTML5 e 13x os recursos do BIOS
Implementação mais fácil de servidores OME x OneView	Implementação de perfis de um para muitos com o OME	52% menos tempo para implementar um servidor do que com o OneView
Atualizações mais fáceis de firmware OME x OneView	Atualizações on-line automáticas com o OME	Conecte-se à página Dell.com para atualizar vários servidores e economize o tempo que você levaria para atualizar servidores carregando pacotes manualmente com o OneView
Alertas mais fáceis OME x OneView	Configure políticas de alerta no OME e tome medidas automatizadas com base em alertas	A automação desse processo proporciona economia de tempo e reduz a possibilidade de erros, em comparação com a realização manual de ações sempre que você recebe um alerta no OneView
Recursos de segurança mais fáceis de usar (bloqueio do sistema e USB dinâmico) iDRAC x iLO	Menos etapas, menos tempo e sem reinicializações com o iDRAC	¼ das etapas e 91% menos tempo de bloqueio do sistema
Lógica analítica mais sólida CloudIQ para PowerEdge x InfoSight	Relatórios personalizáveis e mais métricas de integridade para melhorar o controle administrativo com o CloudIQ para PowerEdge	15x mais opções de métricas, em comparação com o InfoSight



Conclusão

Aproveitar o poder da IA para simplificar e melhorar as operações empresariais pode ser uma tarefa desafiadora, com consequências significativas para os negócios. Com o avanço tecnológico mais rápido do que nunca, a parceria com o fornecedor certo de IA é fundamental. Ao escolher uma empresa como a Dell, que não só oferece um portfólio de IA abrangente, como também pode disponibilizar serviços de planejamento, preparação, implementação e gerenciamento, os clientes podem enfrentar esses desafios de frente. Os testes de referência de desempenho do MLPerf® mostram que as ofertas do portfólio de IA da Dell entregam desempenho sólido e consistente para cargas de trabalho de IA. Com opções de servidor flexíveis e de alto desempenho, juntamente com várias alternativas de armazenamento, soluções validadas e serviços profissionais, especificamente adaptados para IA, a Dell pode ajudar as empresas a adotar a IA e obter os benefícios associados.

---

1. Dell, "Increasing Your Data Value with Dell Generative AI Solutions", acessado em 19 de dezembro de 2023, https://www.dell.com/en-us/blog/increasing-your-data-value-with-dell-generative-ai-solutions/.
2. Dell, "Soluções de IA da Dell", acessado em 12 de dezembro de 2023, https://www.dell.com/pt-br/dt/solutions/artificial-intelligence/index.htm#accordion0&tab0=0&tab1=0.
3. Dell, "Dell Technologies and Hugging Face to Simplify Generative AI with On-Premises IT", acessado em 12 de dezembro de 2023, https://www.dell.com/pt-br/dt/corporate/newsroom/announcements/detailpage.press-releases~usa~2023~11~20231114-dell-technologies-and-hugging-face-to-simplify-generative-ai-with-on-premises-it.htm#/filter-on/Country:pt-br.
4. Dell, "Dell and Meta Collaborate to Drive Generative AI Innovation", acessado em 12 de dezembro de 2023, https://www.dell.com/en-us/blog/dell-and-meta-collaborate-to-drive-generative-ai-innovation/.
5. Dell, "Plataforma de dados pronta para IA da Dell", acessado em 12 de dezembro de 2023, https://www.dell.com/pt-br/dt/solutions/artificial-intelligence/storage-for-ai.htm?hve=explore+unstructured+storage#tab0=0.
6. Dell, "Snowflake and Dell Partnership Gains Momentum", acessado em 19 de dezembro de 2023, https://www.dell.com/en-us/blog/snowflake-and-dell-partnership-gains-momentum/.
7. Robert McNeal, "Dell, VMware and NVIDIA Bring AI to Your Data", acessado em 17 de janeiro de 2024, https://www.dell.com/en-us/blog/dell-vmware-and-nvidia-bring-ai-to-your-data/. De acordo com o link acima: "Com base em uma análise da Dell, de agosto de 2023. A Dell Technologies disponibiliza soluções projetadas para oferecer suporte a cargas de trabalho com IA, desde workstations (móveis e fixas) até servidores para computação com alto desempenho, armazenamento de dados, infraestrutura definida por software nativa na nuvem, switches de rede, proteção de dados, HCI e serviços".
8. Intel, "Acelere cargas de trabalho de inteligência artificial com Intel Advanced Matrix Extensions", acessado em 12 de dezembro de 2023, https://www.intel.com.br/content/www/br/pt/content-details/785250/accelerate-artificial-intelligence-workloads-with-intel-advanced-matrix-extensions.html.

9. Vipera, "NVIDIA's H100 and A100 GPU Cards: Exploring the Intricacies of SXM and PCI-E Connections", acessado em 12 de dezembro de 2023, https://www.viperatech.com/unraveling-the-mysteries-sxm-vs-pci-e-connections-in-nvidias-high-end-h100-and-a100-gpus/.
10. GitHub, "MLPerf® Results Messaging Guidelines", acessado em 16 de janeiro de 2024, https://github.com/mlcommons/policies/blob/master/MLPerf_Results_Messaging_Guidelines.adoc.
11. MLCommons®, "MLPerf® Inference: Datacenter Benchmark Suite Results", acessado em 12 de dezembro de 2023, https://mlcommons.org/en/inference-datacenter-31/.
12. Dell, "Servidores PowerEdge AI", acessado em 12 de dezembro de 2023, https://www.dell.com/pt-br/dt/servers/specialty-servers/poweredge-xe-servers.htm?hve=explore+poweredge+xe#tab0=0&accordion0.
13. HPE, "HPE ProLiant XL675d Gen10 Plus Configure-to-order Server", acessado em 12 de dezembro de 2023, https://www.hpe.com/us/en/product-catalog/compute/proliant-servers/pip.1013142988.html.
14. HPE, "HPE ProLiant DL380a Gen11", acessado em 12 de dezembro de 2023, https://www.hpe.com/us/en/product-catalog/compute/proliant-servers/pip.proliant-dl380-server.1014696168.html.
15. MLCommons®, "MLPerf® Inference: Datacenter Benchmark Suite Results v 3.1", acessado em 12 de dezembro de 2023, https://mlcommons.org/benchmarks/inference-datacenter/.
16. HPE, "NVIDIA Accelerators for HPE ProLiant Servers", acessado em 12 de dezembro de 2023, https://www.hpe.com/psnow/doc/c04123180.html?jumpid=in_pdp-psnow-qs.
17. Dell, "Specification Sheet do PowerEdge XE9680", acessado em 19 de janeiro de 2024, https://www.delltechnologies.com/asset/pt-br/products/servers/technical-support/poweredge-xe9680-spec-sheet.pdf.
18. HPE, "HPE & NVIDIA financial services solution sets new records in performance", acessado em 12 de dezembro de 2023, https://community.hpe.com/t5/alliances/hpe-amp-nvidia-financial-services-solution-sets-new-records-in/ba-p/7197388.
19. HPE, "QuickSpecs: HPE Cray Supercomputing XD670", acessado em 12 de dezembro de 2023, https://www.hpe.com/psnow/doc/a50004292enw.
20. Pontuação encerrada da inferência v3.1 verificada pelo MLPerf®. Recuperada de https://mlcommons.org/benchmarks/inference-datacenter/, 5 de dezembro de 2023, entrada 3.1-0069. O nome e o logotipo MLPerf® são marcas registradas e não registradas da MLCommons® Association, nos Estados Unidos e em outros países. Todos os direitos reservados. Uso não autorizado estritamente proibido. Acesse www.mlcommons.org para obter mais informações.
21. Pontuação encerrada da inferência v3.1 verificada pelo MLPerf®. Recuperada de https://mlcommons.org/benchmarks/inference-datacenter/, 5 de dezembro de 2023, entrada 3.1-0085. O nome e o logotipo MLPerf® são marcas registradas e não registradas da MLCommons® Association, nos Estados Unidos e em outros países. Todos os direitos reservados. Uso não autorizado estritamente proibido. Acesse www.mlcommons.org para obter mais informações.
22. Dell, "Servidor em rack PowerEdge XE9640", acessado em 12 de dezembro de 2023, https://www.dell.com/pt-br/shop/ipovw/poweredge-xe9640.
23. Accelsius, "Enabling the AI Revolution with Liquid Cooling", acessado em 12 de dezembro de 2023, https://www.accelsius.com/blog/enabling-the-ai-revolution-with-liquid-cooling.
24. Dell, "Dell PowerEdge XE9640 Technical Guide", acessado em 12 de dezembro de 2023, https://www.delltechnologies.com/asset/pt-br/products/servers/technical-support/poweredge-xe9640-technical-guide.pdf.
25. HPE, "HPE ProLiant DL380a Gen11", acessado em 12 de dezembro de 2023, https://www.hpe.com/psnow/doc/PSN1014696168WWEN.pdf?jumpid=in_pdp-psnow-dds.
26. Intel, "Intel® Data Center GPU Max Series Technical Overview", acessado em 12 de dezembro de 2023, https://www.intel.com/content/www/us/en/developer/articles/technical/intel-data-center-gpu-max-series-overview.html#gs.08874l.
27. HPE, "Intel Data Center GPU Max 1100 48GB Accelerator for HPE Data sheet", acessado em 12 de dezembro de 2023, https://www.hpe.com/psnow/doc/PSN1014779728WWEN.
28. Pontuação encerrada da inferência v3.1 verificada pelo MLPerf®. Recuperada de https://mlcommons.org/benchmarks/inference-datacenter/, 5 de dezembro de 2023, entrada 3.1-0066. O nome e o logotipo MLPerf® são marcas registradas e não registradas da MLCommons® Association, nos Estados Unidos e em outros países. Todos os direitos reservados. Uso não autorizado estritamente proibido. Acesse www.mlcommons.org para obter mais informações.
29. Pontuação encerrada da inferência v3.1 verificada pelo MLPerf®. Recuperada de https://mlcommons.org/benchmarks/inference-datacenter/, 5 de dezembro de 2023, entrada 3.1-0084. O nome e o logotipo MLPerf® são marcas registradas e não registradas da MLCommons® Association, nos Estados Unidos e em outros países. Todos os direitos reservados. Uso não autorizado estritamente proibido. Acesse www.mlcommons.org para obter mais informações.

30. Dell, "AI and HPC —With Air or Liquid Cooling", acessado em 12 de dezembro de 2023, https://www.delltechnologies.com/asset/pt-br/products/servers/briefs-summaries/poweredge-xe9640-and-xe8640-infographic.pdf.
31. Dell, "PowerEdge XE8640: Drive AI, HPC modeling and simulation workloads with superior performance", acessado em 12 de dezembro de 2023, https://www.delltechnologies.com/asset/pt-br/products/servers/technical-support/poweredge-xe8640-spec-sheet.pdf.
32. Dell, "Servidor em rack PowerEdge XE8640", acessado em 12 de dezembro de 2023, https://www.dell.com/pt-br/shop/ipovw/poweredge-xe8640.
33. Pontuação encerrada da inferência v3.1 verificada pelo MLPerf®. Recuperada de https://mlcommons.org/benchmarks/inference-datacenter/, 5 de dezembro de 2023, entrada 3.1-0067. O nome e o logotipo MLPerf® são marcas registradas e não registradas da MLCommons® Association, nos Estados Unidos e em outros países. Todos os direitos reservados. Uso não autorizado estritamente proibido. Acesse www.mlcommons.org para obter mais informações.
34. Pontuação encerrada da inferência v3.1 verificada pelo MLPerf®. Recuperada de https://mlcommons.org/benchmarks/inference-datacenter/, 5 de dezembro de 2023, entrada 3.1-0084. O nome e o logotipo MLPerf® são marcas registradas e não registradas da MLCommons® Association, nos Estados Unidos e em outros países. Todos os direitos reservados. Uso não autorizado estritamente proibido. Acesse www.mlcommons.org para obter mais informações.
35. Dell, "Servidor em rack PowerEdge R760xa", acessado em 12 de dezembro de 2023, https://www.dell.com/pt-br/shop/ipovw/poweredge-r760xa#features_section.
36. SANStorageWorks, "Dell EMC PowerEdge R760xa: Powerful and scalable for GPU workloads", acessado em 12 de dezembro de 2023, https://www.sanstorageworks.com/PowerEdge-R760xa.asp.
37. Dell, "Dell PowerEdge Servers and NVIDIA GPUs", acessado em 12 de dezembro de 2023, https://infohub.delltechnologies.com/l/design-guide-generative-ai-in-the-enterprise-inferencing/dell-poweredge-servers-and-nvidia-gpus-1/.
38. Pontuação encerrada da inferência v3.1 verificada pelo MLPerf®. Recuperada de https://mlcommons.org/benchmarks/inference-datacenter/, 5 de dezembro de 2023, entrada 3.1-0064. O nome e o logotipo MLPerf® são marcas registradas e não registradas da MLCommons® Association, nos Estados Unidos e em outros países. Todos os direitos reservados. Uso não autorizado estritamente proibido. Acesse www.mlcommons.org para obter mais informações.
39. Pontuação encerrada da inferência v3.1 verificada pelo MLPerf®. Recuperada de https://mlcommons.org/benchmarks/inference-datacenter/, 5 de dezembro de 2023, entrada 3.1-0084. O nome e o logotipo MLPerf® são marcas registradas e não registradas da MLCommons® Association, nos Estados Unidos e em outros países. Todos os direitos reservados. Uso não autorizado estritamente proibido. Acesse www.mlcommons.org para obter mais informações.
40. Dell, "Workstations for AI", acessado em 12 de dezembro de 2023, https://www.dell.com/pt-br/lp/dt/ai-technologies?utm_source=AISearchTools&utm_medium=youtube&utm_campaign=precisionai#pdf-overlay=//www.delltechnologies.com/asset/en-us/products/workstations/briefs-summaries/ai-industry-brochure.pdf.
41. NVIDIA, "NVIDIA RTX em workstations profissionais", acessado em 12 de dezembro de 2023, https://www.nvidia.com/pt-br/design-visualization/desktop-graphics/.
42. Intel, "Intel® Deep Learning Boost (Intel® DL Boost)", acessado em 12 de dezembro de 2023, https://www.intel.com.br/content/www/br/pt/artificial-intelligence/deep-learning-boost.html.
43. HPE, "HPE ProLiant ML350 Gen11", acessado em 12 de dezembro de 2023, https://buy.hpe.com/br/pt/compute/tower-servers/proliant-ml300-servers/proliant-ml350-server/hpe-proliant-ml350-gen11/p/1014696172.
44. ComputerWeekly.com, "Storage requirements for AI, ML and analytics in 2022", acessado em 12 de dezembro de 2023, https://www.computerweekly.com/feature/Storage-requirements-for-AI-ML-and-analytics-in-2022.
45. Dell, "Plataforma de dados pronta para IA PowerScale", acessado em 12 de dezembro de 2023, https://www.dell.com/pt-br/shop/powerscale-family/sf/powerscale.
46. Dell, "Compare o PowerScale", acessado em 12 de dezembro de 2023, https://www.dell.com/pt-br/shop/powerscale-family/sf/powerscale#compare-module.
47. Dell, "Dell PowerScale All-Flash", acessado em 12 de dezembro de 2023, https://www.delltechnologies.com/asset/pt-br/products/storage/technical-support/h15963-ss-powerscale-all-flash-nodes.pdf.
48. Dell, "Recursos de software do Dell PowerScale OneFS", acessado em 12 de dezembro de 2023, https://www.delltechnologies.com/asset/pt-br/products/storage/technical-support/h18275-onefs-software-features-data-sheet.pdf.
49. Dell, "Dell APEX File Storage for AWS", acessado em 12 de dezembro de 2023, https://www.delltechnologies.com/asset/pt-br/products/storage/briefs-summaries/h19575-so-apex-file-storage-for-aws.pdf.

50. Dell, "Armazenamento empresarial em objeto do Dell ECS", acessado em 12 de dezembro de 2023, https://www.dell.com/pt-br/dt/storage/ecs/index.htm?hve=explore+ecs#tab0=0&tab1=0.
51. Dell, "Armazenamento empresarial em objeto do Dell ECS", acessado em 12 de dezembro de 2023, https://www.dell.com/pt-br/dt/storage/ecs/index.htm#tab0=0&tab1=0&accordion0.
52. HPE, "Storage Solutions for Scality", acessado em 12 de dezembro de 2023, https://www.hpe.com/br/en/storage/file-object/scality.html.
53. HPE, "Make AI work for you", acessado em 16 de janeiro de 2024, https://www.hpe.com/br/en/solutions/ai-artificial-intelligence.html.
54. HPE, "HPE AI Services – Generative AI Implementation", acessado em 16 de janeiro de 2024, https://www.hpe.com/br/en/services/generative-ai-implementation-service.html.
55. Principled Technologies, "Simplify administrator tasks and improve security and health monitoring with tools from the Dell management portfolio vs. comparable tools from HPE", acessado em 12 de dezembro de 2023, https://www.principledtechnologies.com/Dell/Management-tools-vs-HPE-1122.pdf.
56. Principled Technologies, "Simplify administrator tasks and improve security and health monitoring with tools from the Dell management portfolio vs. comparable tools from HPE", acessado em 12 de dezembro de 2023, https://www.principledtechnologies.com/Dell/Management-tools-vs-HPE-1122.pdf.
57. Principled Technologies, "Simplify administrator tasks and improve security and health monitoring with tools from the Dell management portfolio vs. comparable tools from HPE".

Consulte a versão original do relatório, em inglês, em https://facts.pt/zPmSx4c

Este projeto foi encomendado por Dell Technologies.

Principled Technologies®

Facts matter.®