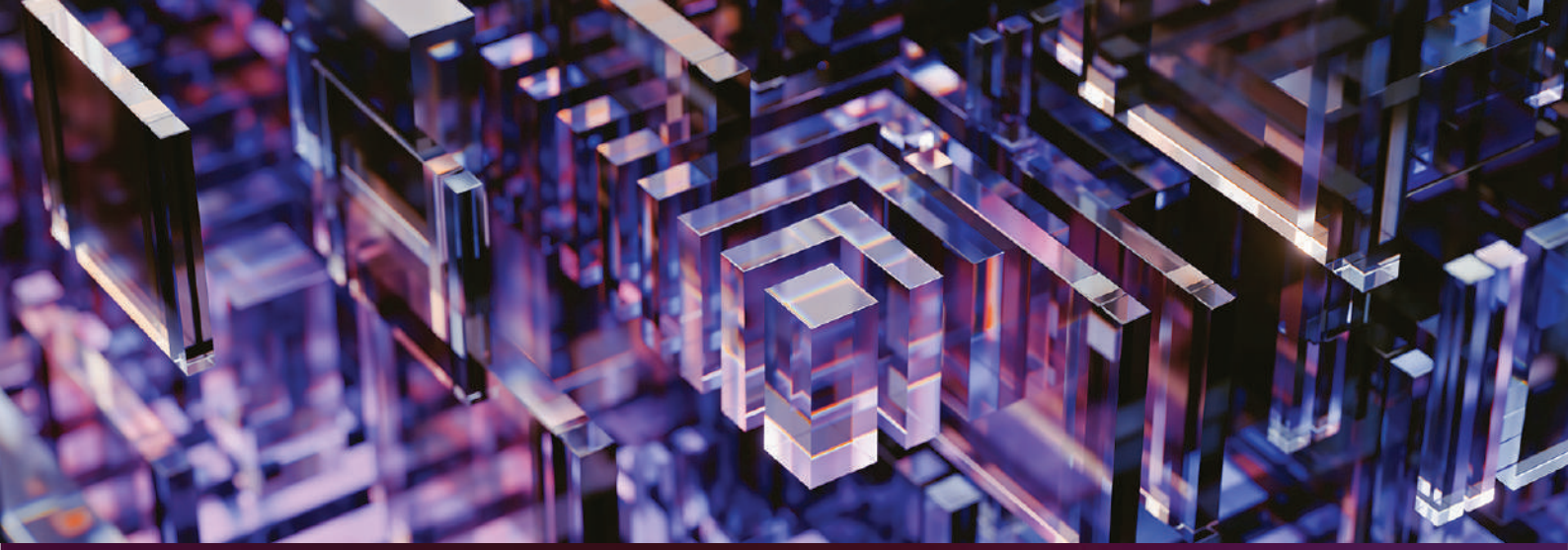


Fornecendo Infraestrutura de Alto Desempenho para Inteligência Artificial Generativa e Empresarial

Soluções da Lenovo e NVIDIA® para Aprimorar Produtividade, Inovação e Tempo de Chegada ao Mercado

Resumo Executivo

Em diversos setores, a Inteligência Artificial (IA) e a Inteligência Artificial Generativa (GenAI) podem acelerar a inovação e aprimorar a posição competitiva de uma empresa, a qualidade de produtos/serviços, operações e o engajamento do cliente. No entanto, existem numerosos desafios de implementação ao implantar casos de uso da vida real.

A Lenovo auxilia as empresas a superarem esses obstáculos, fornecendo um conjunto abrangente de serviços com as melhores práticas e uma infraestrutura otimizada para IA, composta por servidores, armazenamento, estações de trabalho, dispositivos móveis e software, desde o edge até o data center e a nuvem. Por exemplo, os servidores Lenovo ThinkSystem e ThinkEdge, alimentados por unidades de processamento gráfico (GPUs) da NVIDIA e software, podem acelerar a jornada de IA e GenAI do cliente, proporcionando:

- Resultados mais rápidos no tempo de treinamento e workloads de inferência em texto, vídeo, imagem e outras modalidades de dados.
- Mais flexibilidade para personalizar e otimizar diversas workloads de IA, GenAI e outras relacionadas, do edge ao data center e à nuvem.
- Melhor eficiência energética e menor custo total de propriedade (TCO).
- Diversas ofertas e serviços imersivos complementares para facilitar a transformação digital da empresa com IA e GenAI, incluindo acesso ao Centro de Excelência Lenovo AI Discover (AIDiscover@lenovo.com).

Introdução

Soluções de IA e GenAI estão crescendo rapidamente no meio empresarial, proporcionando muitos benefícios em várias indústrias. Empresas estão implementando IA e GenAI para acelerar a inovação e aprimorar sua posição competitiva, a qualidade de produtos/serviços, operações e o engajamento do cliente.

Apesar da promessa e do valor econômico imenso da IA, os desafios de implementação são igualmente grandes, dadas as grandes quantidades de dados e a necessidade de armazenar, analisar e proteger efetivamente todos os dados valiosos ao longo de seu ciclo de vida. Com GenAI, essas questões se tornam ainda mais agudas.

Este white paper discute como a Lenovo e a NVIDIA se associam com suas tecnologias únicas e respectivas para fornecer a arquitetura ideal para oferecer IA e GenAI para empresas. Com base no envolvimento da Lenovo e da NVIDIA com clientes e parceiros, este documento oferece orientações valiosas para selecionar configurações otimizadas para o desempenho em vários casos de uso de IA e GenAI em diversas indústrias para obter uma vantagem competitiva. A Lenovo continua investindo[1] em parcerias de IA, incluindo a NVIDIA, para acelerar a implementação de IA para empresas em todo o mundo e ajudar os clientes a iniciar sua jornada de IA e GenAI hoje!

IA e GenAI Impulsionam o Valor Empresarial em Diversas Indústrias

IA, que inclui Aprendizado de Máquina (ML) e Aprendizado Profundo (DL), está crescendo rapidamente e transformando uma ampla gama de indústrias e aplicações. O mercado global de IA deverá atingir US $241,80 bilhões em 2023 e prevê-se que cresça a uma taxa de crescimento anual composta (CAGR) de 17,30% de 2023 a 2030, resultando em um volume de mercado de US $738,80 bilhões até 2030.[2]

GenAI, incluindo Modelos de Linguagem Grandes (LLMs), é um novo tipo poderoso de DL que pode criar conteúdo, como texto, imagens, áudio e vídeo. Ele faz isso aprendendo padrões a partir de dados existentes e usando esse conhecimento para gerar saídas novas e únicas. GenAI pode produzir conteúdo altamente realista e complexo que imita a criatividade humana. Ele está crescendo ainda mais rápido que a IA[3] (mais de 58%), tornando-se uma ferramenta valiosa para muitas indústrias, como serviços financeiros, saúde, manufatura, varejo, telecomunicações/mídia etc.

A Figura 1 representa vários casos de uso proeminentes de GenAI resumidos de um estudo recente da Deloitte[4], que adicionam valor significativo aos negócios em diversas indústrias, extraindo insights profundos e acionáveis em muitas modalidades de dados (Texto, Áudio, Imagem, Vídeo, Código e Artefatos 3D/Especializados). As cores sólidas e intensas no gráfico de pizza representam as modalidades de dados geralmente proeminentes para cada caso de uso. A cor branca no gráfico de pizza é para modalidades de dados que geralmente não são significativas.

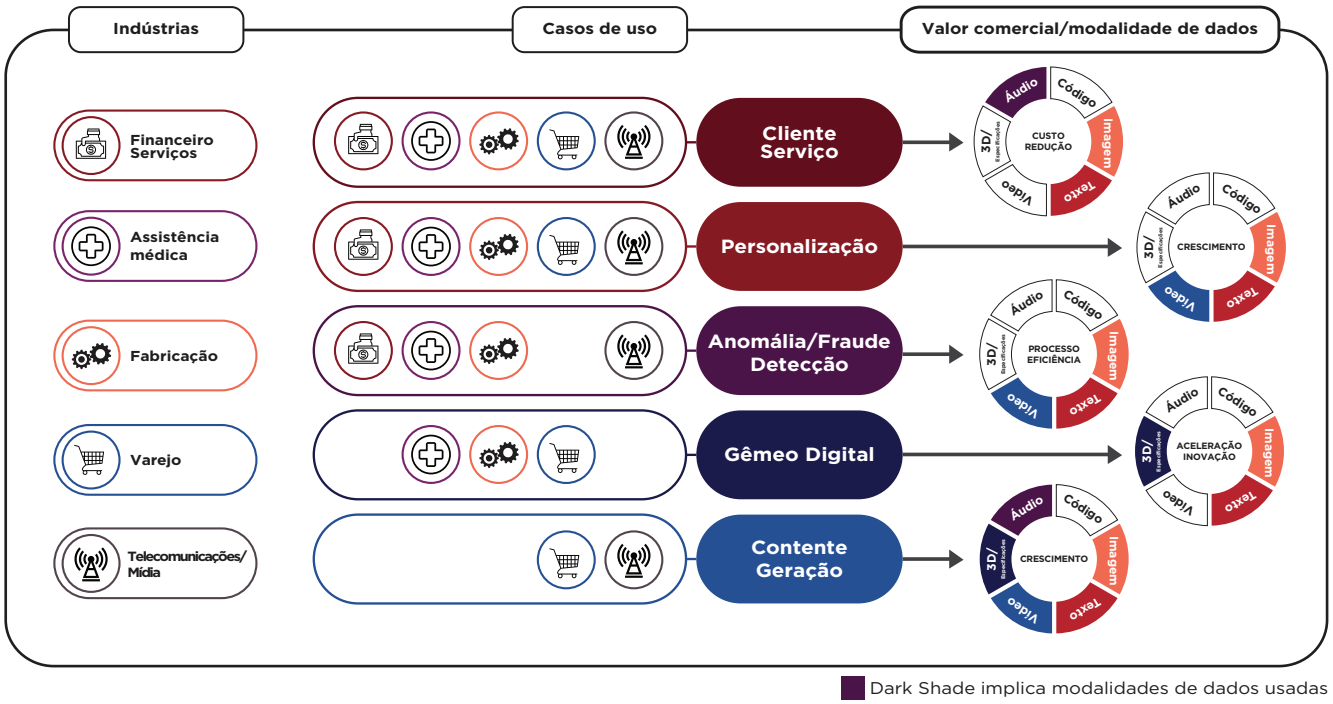


Figura 1: Casos de Uso de Alto Valor de GenAI em Diversas Indústrias e Modalidades de Dados[5]

Serviços Financeiros: Bancos e seguradoras estão adicionando GenAI aos seus processos intensivos em dados para aprimorar:

- **Atendimento ao Cliente:** Interface de avatar digital alimentada por GenAI com opções de texto, áudio e imagens que aprimoram o suporte ao cliente 24/7, respondem a consultas e auxiliam em tarefas financeiras para melhorar a eficiência do processo e o engajamento do cliente.
- **Personalização:** Entregar materiais de marketing em conformidade com as regulamentações, promoções de produtos e engajamento de vendas com texto, imagens e vídeos personalizados em diferentes geografias para impulsionar o crescimento e adquirir novos clientes.
- **Detecção de Fraudes:** Identificar transações fraudulentas em tempo real, analisando padrões e anomalias em dados reais e sintéticos em várias modalidades, ajudando a aprimorar processos e prevenir perdas financeiras.

Assistência médica: Pagadores, provedores, organizações farmacêuticas e de biotecnologia estão incorporando GenAI para:

- **Atendimento ao Cliente:** Acelerar a autorização prévia para pacientes e gerar respostas a perguntas sobre o processo de reivindicação, cobertura de seguro e outros detalhes do plano para aprimorar o processo. Possibilitar monitoramento contínuo e proativo 24/7 de pacientes com dispositivos IoT e análises alimentadas por IA para monitorar os sinais vitais do paciente, alertar os profissionais de saúde para desvios de valores típicos e tomar medidas corretivas.
- **Personalização:** Descobrir e adaptar tratamentos e medicamentos para pacientes individuais com base em sua composição genética e histórico médico para aprimorar a eficácia do cuidado e impulsionar os negócios e a vantagem competitiva.
- **Detecção de Fraude/Anomalia:** Identificar reivindicações fraudulentas em tempo real, analisando padrões e anomalias em dados de várias modalidades, ajudando a aprimorar processos e prevenir perdas financeiras. Analisar imagens médicas como raios-X, ressonâncias magnéticas e tomografias computadorizadas para auxiliar na detecção precoce de doenças e anomalias.
- **Gêmeo Digital:** Construir réplicas digitais centradas no paciente de ponta a ponta para analisar dados do paciente, incluindo registros médicos/imagens e sintomas, para aprimorar o diagnóstico de doenças e recomendar planos de tratamento. Identificar possíveis candidatos a medicamentos, prever sua eficácia e otimizar estruturas moleculares.

Prever surtos de doenças, readmissão de pacientes e utilização de recursos de saúde, auxiliando hospitais e clínicas a alocar recursos de maneira eficiente. Tudo isso impulsiona mais inovação em todo o ecossistema de saúde.

Manufatura: Fabricantes automotivos, aeroespaciais e de semicondutores estão incorporando GenAI para:

- **Manutenção:** Analisar dados de sensores de máquinas em várias modalidades para prever quando podem falhar. Isso ajuda a realizar serviços preventivos e evitar tempo de inatividade e interrupções custosas.
- **Personalização:** Possibilitar a personalização em massa analisando dados e adaptando eficientemente os processos de fabricação para produzir produtos personalizados. Isso impulsiona maior apelo ao cliente e crescimento nos negócios.
- **Detecção de Anomalias:** Sistemas de visão computacional alimentados por IA podem inspecionar rapidamente e com precisão produtos em busca de defeitos, reduzindo o número de itens defeituosos na linha de produção para impulsionar a eficiência do processo.
- **Gêmeo Digital:** Construir uma réplica digital de ponta a ponta de todo o ciclo de vida do produto, desde o desenvolvimento até a fabricação e o serviço. Isso ajuda a gerar e avaliar novos designs de produtos, otimizando-os para desempenho, custo e fabricabilidade. Otimizar os processos de fabricação analisando dados de sensores e linhas de produção para melhorar a eficiência e a qualidade do produto. Fornece insights em tempo real sobre a cadeia de suprimentos e a operação do cliente, ajudando os fabricantes a rastrear matérias-primas, monitorar o progresso da produção, responder a interrupções, rastrear o uso real do cliente de seus produtos e garantir manutenção oportuna. Tudo isso melhora drasticamente a inovação de produtos e processos e a qualidade.

Varejo: Empresas voltadas para o consumidor, como grandes varejistas, pequenos comerciantes e varejistas especializados, estão utilizando GenAI para:

- **Atendimento ao Cliente:** Opções de interface de avatar digital alimentadas por GenAI com texto, áudio e imagens aprimoram o suporte ao cliente 24/7, respondem a consultas e auxiliam nas recomendações de produtos para melhorar a eficiência do processo e o engajamento empático com o cliente, construindo lealdade e equidade da marca. Também libera recursos humanos caros para lidar com questões mais complexas do cliente.
- **Personalização:** Entregar materiais de marketing, promoções de produtos e engajamento de vendas com texto, imagens e vídeos personalizados em diferentes geografias para impulsionar o crescimento e adquirir novos clientes. Gerar recomendações mais específicas e direcionadas em muitas modalidades do que os motores de busca para tornar a compra mais personalizada e conveniente.
- **Gêmeo Digital:** Criar showrooms virtuais, demonstrações de produtos e planogramas, personalizar a experiência do cliente, prever a demanda, simular o layout e as operações de uma loja e identificar áreas para melhoria. Tudo isso pode ajudar os varejistas a inovar mais.
- **Geração de Conteúdo:** Criar descrições de produtos, imagens, vídeos e muito mais de maneira mais rápida e consistente do que ferramentas e processos tradicionais. Para impulsionar os negócios, personalizar esse conteúdo por geografia, idioma, nuances culturais e regulamentações locais.

Talco/Mídia: Com o GenAI, essas empresas estão acelerando a transformação digital com:

- **Atendimento ao Cliente:** Opções de interface de avatar digital e assistência virtual alimentadas por GenAI com texto, áudio e imagens aprimoram o suporte ao cliente 24/7, respondem a consultas e auxiliam com recomendações de serviços para melhorar a eficiência do processo e o engajamento empático do cliente, construindo lealdade e conten-do custos de troca. Também libera recursos huma-nos caros para lidar com questões mais complexas do cliente. Otimizar o desempenho da rede e reduzir a congestão, melhorando a experiência do cliente.
- **Personalização:** Organizar e gerenciar tipos complexos de arquivos, analisar conteúdo antes da tradução para otimizar a localização e integrar outras ferramentas de idiomas no fluxo de trabalho para aumentar as conversões e o engajamento para construir lealdade. O reconhecimento de fala ajuda a transcrever conteúdo de vídeo e áudio em texto e traduzir conteúdo falado para outros idiomas para impulsionar os negócios.
- **Detecção de Fraude:** Usar dados reais e sintéticos para aprimorar a eficiência do processo e detectar atividades fraudulentas em redes de telecomunicações, como troca de SIM e acesso não autorizado.
- **Geração de Conteúdo:** Imitar o estilo dos materiais de marketing da empresa e gerar novas e diversas versões de conteúdo rapidamente e sob demanda adaptadas a diferentes audiências. Aprimorar a qualidade da linguagem de materiais de marketing com fraseologia, gramática, estilo da empresa e aderência aos valores da empresa. Criar rapidamente inúmeras versões de conteúdo em diversos estilos para identificar a melhor opção para impulsionar os negócios.

Embora o retorno sobre o investimento (ROI) do GenAI possa ser substancial para a empresa, implementar aprendizado profundo (DL) e a infraestrutura de tecnologia da informação (TI) de alto desempenho associada pode ser complexo e dispendioso. Existem inúmeros desafios de implementação.

Para discutir seu caso de uso específico, entre em contato com o Laboratório de Descoberta de IA da Lenovo, enviando um e-mail para AIDiscover@lenovo.com.

Desafios de Implementação de IA e GenAI

A implementação de fluxos de trabalho de DL e GenAI em produção normalmente passa por quatro estágios (Figura 2)[6]:

1. **Gerenciamento de Dados** para preparar os dados necessários para construir o modelo DL e **GenAI.**
2. **Aprendizado do Modelo (Treinamento)** para definir, selecionar e treinar o modelo DL e **GenAI.**
3. **Verificação do Modelo (Treinamento)** para garantir que o modelo atenda a requisitos específicos de funcionalidade e desempenho.
4. **Implantação do Modelo (Inferência)** para integrar o modelo treinado na infraestrutura de TI e executar, manter e atualizar o modelo conforme necessário.

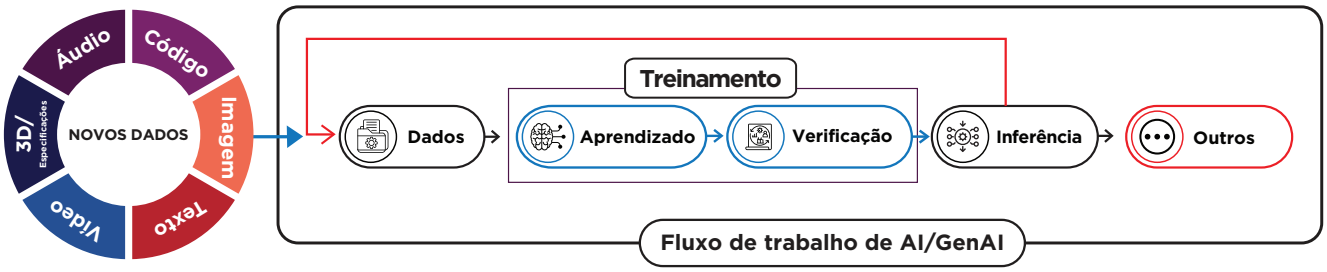


Figura 2: Principais Fases no Fluxo de Trabalho de DL e GenAI

Estes estágios têm etapas menores (Figura 2) que podem ser executadas em paralelo e com feedback. Além disso, há outras considerações éticas, legais, de confiança e segurança. Tudo isso torna a implementação do GenAI muito desafiadora. A Figura 3 representa esses desafios de Dados, Processos, Negócios, Infraestrutura e Outros:

Figura 3: Desafios de Implementação de GenAI

Dados

Qualidade e Quantidade dos Dados

- Disponibilidade de Dados: Modelos GenAI frequentemente requerem grandes quantidades de dados de alta qualidade, o que pode ser desafiador de coletar e curar.
- Diversidade de Dados: Garantir que os dados de treinamento representem uma ampla variedade de cenários e demografias pode ser complexo.

Viés e Equidade

- Viés nos Dados: Modelos GenAI podem herdar viés nos dados de treinamento, resultando em saídas tendenciosas ou injustas.
- Equidade: Garantir equidade no conteúdo gerado, especialmente em domínios sensíveis como finanças e saúde, é um desafio significativo.

Interpretabilidade e Explicabilidade

- Modelos Caixa-Preta: Muitos modelos GenAI são como "caixas-pretas", tornando difícil entender seus processos de tomada de decisão. Isso pode ser problemático para aplicações onde a transparência é crucial.

Processos

Gestão de Mudanças

- Cultura Organizacional: Implementar GenAI pode exigir mudanças significativas na cultura, processos e fluxos de trabalho de uma organização, o que pode exigir superar resistência organizacional.

Colaboração Humano-AI

- Treinamento e Monitoramento: Empresas precisam estabelecer processos de colaboração entre operadores humanos e sistemas GenAI, incluindo monitoramento contínuo e manutenção.

Aceitação e Confiança do Usuário

- Ceticismo do Usuário: Os usuários podem ser céticos em relação ao conteúdo gerado por IA, afetando as taxas de adoção.
- Construção de Confiança: Construir confiança no conteúdo gerado por IA é crucial para a aceitação do usuário.

Negócios

Avaliação do ROI

- Mensuração do Impacto: Avaliar o retorno sobre o investimento (ROI) da implementação do GenAI pode ser desafiador, especialmente na quantificação do valor gerado por soluções de IA.

Habilidades e Talentos

- Escassez de Talentos: Pode haver escassez de especialistas em IA e cientistas de dados com as habilidades necessárias para implementar e manter sistemas GenAI de forma eficaz.

Insfraestrutura

Recursos Computacionais

- Infraestrutura de Alto Desempenho: Treinar e implantar modelos GenAI em grande escala requer recursos computacionais substanciais, resultando em custos de infraestrutura elevados.
- Escalabilidade: Garantir que a infraestrutura possa se expandir para lidar com aumentos nas demandas computacionais à medida que os modelos GenAI evoluem é um desafio contínuo.
- Eficiência Energética: A infraestrutura deve ser eficiente em termos de energia. Análises mostram que o treinamento de um LLM, um modelo GenAI com 200 bilhões de parâmetros, produz aproximadamente 75.000 kg de emissões de CO_2, em comparação com apenas 900 kg de emissões de CO_2 para um voo de Nova York a San Francisco[7].

Integração com Sistemas Existentes

- Sistemas Legados: Integrar GenAI com a infraestrutura de TI existente e sistemas legados pode ser complexo e exigir esforço substancial.

Treinamento e Ajuste do Modelo

- Tempo de Treinamento: Treinar modelos GenAI complexos pode ser demorado, atrasando a implantação de soluções de IA.
- Ajuste de Hiperparâmetros: Ajustar modelos para tarefas específicas e otimizar seu desempenho pode exigir esforço significativo.

Outros

Preocupações Éticas

- Uso Malicioso: Há preocupações sobre o uso indevido do GenAI para gerar conteúdo falso, deep fakes ou outros fins maliciosos.
- Privacidade: Gerar conteúdo altamente personalizado pode levantar preocupações com a privacidade, exigindo medidas robustas de proteção de dados.
- Alucinações: São saídas do modelo que são ou sem sentido ou completamente falsas.

Conformidade Regulatória

- Privacidade de Dados: Cumprir regulamentos de privacidade de dados, como o GDPR ou o HIPAA, pode ser complexo ao lidar com dados gerados pelo usuário.
- Regulamentações de Conteúdo: Algumas indústrias, como farmacêutica, bancária e financeira, têm regulamentações rigorosas que governam o conteúdo que produzem e compartilham.

Segurança

- Vulnerabilidades: Modelos GenAI podem ser vulneráveis a ataques adversários, comprometendo potencialmente sua confiabilidade e segurança.
- Propriedade Intelectual: Modelos GenAI e os processos usados para construí-los são frequentemente os "tesouros" de uma organização e devem ser protegidos.

A Lenovo espera que a IA seja desenvolvida e usada consistentemente com seus valores fundamentais. O Comitê de IA Responsável da Lenovo garante que todas as soluções, incluindo aquelas dos parceiros Inovadores de IA, atendam a requisitos que protejam os usuários finais e garantam que o uso da IA seja justo, ético e responsável, com foco em:

- Diversidade e Inclusão
- Privacidade e Segurança
- Responsabilidade e Confiabilidade
- Explicabilidade
- Transparência
- Impacto Ambiental e Social

A Lenovo e a NVIDIA estão trabalhando com um ecossistema amplo e crescente de parceiros e clientes para desenvolver as melhores práticas e soluções que ajudem

as empresas a superar esses desafios de implementação. A Lenovo também criou uma arquitetura de referência[8] para GenAI baseada em GPUs e software da NVIDIA.

Arquitetura em Alto Nível das Soluções da Lenovo Impulsionadas pela NVIDIA

A Lenovo simplifica a implementação de IA e GenAI com infraestrutura otimizada, pronta para implantação (hardware, software e serviços), expertise comprovada, soluções pré-validadas de ISVs e parceiros projetados para qualquer tamanho ou escala. Na base desta arquitetura em alto nível (Figura 4), estão os sistemas líderes engenhados esperta-mente para alto desempenho e armazenamento para IA e GenAI, desde estações de trabalho até o edge, passando pelo data center até a nuvem.

Figura 4: Arquitetura em Alto Nível de IA e GenAI

Alguns componentes (não abordados anteriormente) desta arquitetura em alto nível, começando na camada de infraes-trutura, incluem:

- **Servidores Lenovo ThinkSystem Otimizados para Desempenho:** Servidores altamente confiáveis, escaláveis e de alto desempenho para acelerar significativamente a IA e GenAI. Este portfólio de servidores da Lenovo inclui o Lenovo ThinkSystem SR675 V3 rico em GPU. Aproveitando as tecnologias de resfriamento líquido da Lenovo, alguns sistemas vão desde o resfriamento direto à água para CPUs e GPUs, até sistemas aprimorados com líquido, em que o líquido aumenta o resfriamento padrão a ar.
- **Servidores Lenovo ThinkEdge:** Entregam plataformas específicas e seguras, adequadas para aplicações intensivas em computação e sensíveis à latência, como o Lenovo ThinkEdge SE455 V3, implantado fora dos data centers tradicionais.
- **Armazenamento Lenovo:** JBODs de armazenamento diretamente conectados e unidades de expansão oferecem armazenamento flexível, econômico e de alta capacidade, ideal para ambientes com restrições de espaço e clientes sensíveis a custos. As matrizes totalmente em flash da série DE ThinkSystem da Lenovo são projetadas para desempenho extremo com até 2,0 milhões de IOPS e latência inferior a 100 microssegundos, incluindo recursos comprovados de disponibilidade e segurança em nível empresarial.
- **Estações de Trabalho Lenovo:** Estações de trabalho da série P ThinkStation com GPUs da NVIDIA proporcionam desempenho poderoso e são certifi-cadas por ISVs, eficientes em energia e altamente versáteis.
- **Opções de Implantação:** Oferece autonomia para adaptar a abordagem de implantação, escolhendo entre uma configuração robusta de metal nu ou uma implantação virtual versátil.
- **NVIDIA AI Enterprise:** Como uma pilha de software completa de IA, o NVIDIA AI Enterprise (com componentes-chave como NVIDIA NeMoTM, NVIDIA Riva e NVIDIA TritonTM) acelera os pipelines de IA e simplifica o desenvolvimento e a implemen-tação de IA de produção para uma ampla gama de casos de uso, desde visão computacional até GenAI, incluindo LLMs.
- **NVIDIA OmniverseTM Enterprise:** É uma plataforma de software nativa OpenUSD para conectar pipelines 3D complexos e desenvolver aplicativos para digita-lização industrial. Unifique suas ferramentas e dados 3D para quebrar silos de informações, minimizar a preparação tediosa de dados e potencializar a colaboração em equipes empresariais. Aproveite ferramentas de desenvolvedor fáceis de usar para construir aplicativos 3D avançados em tempo real que permitam visualizar e simular produtos, ativos e instalações em plena fidelidade de design. Implante a

plataforma em seu ambiente preferido, seja em estações de trabalho móveis profissionais NVIDIA RTXTM, estações de trabalho e servidores certificados pela NVIDIA, ou NVIDIA OVXTM.

- **Aplicação:** Componentes principais nesta camada incluem:
  - **Databricks MLflowTM:** Fornece uma plataforma unificada para gerenciar o ciclo de vida de aprendizado de máquina, desde o rastreamento de experimentos e registro de modelo até a implantação e monitoramento do modelo.
  - **Lenovo XClarity:** É uma família de software que simplifica e automatiza a implantação e gestão da infraestrutura da Lenovo, permitindo que os clientes se concentrem em projetos de alto valor.
  - **Lenovo Intelligent Computing Orchestration (LiCO):** Reduz a complexidade de usar um cluster HPC massivo e simplifica o deployment, operação e aceleração de aplicativos.
  - **Run:ai:** É um escalonador que gerencia tarefas em lotes usando múltiplas filas no topo do Kubernetes®, permitindo que os administradores do sistema definam diferentes regras, políticas e requisitos para cada fila com base nas prioridades de negócios.
- **Lenovo Remote Visualization:** Fornece acesso confiável e seguro a aplicativos intensivos em gráficos a qualquer momento, em qualquer lugar. Em vez de fornecer novas estações de trabalho caras para toda a equipe de design, a TI pode implantar computadores pessoais empresariais ou de classe consumidor menos caros. Além disso, os departamentos de TI podem manter a segurança e reduzir os custos usando virtualização remota hospedada em um data center interno ou na nuvem. A visualização remota realiza operações gráficas intensivas em um servidor gráfico de alta qualidade gera uma versão de pixel 2D que os usuários podem receber rapidamente. Além disso, a renderização do lado do servidor acelera consideravelmente o processo de usar gráficos em sessões remotas.
- **Lenovo ou Serviços Certificados por Parceiros:** A Lenovo e seu ecossistema global de parceiros altamente especializados em software e serviços de IA podem fornecer toda ou partes da pilha integrada da Lenovo representada na Figura 4. Eles também podem fornecer serviços de instalação e inicialização no local para integrar isso ao ambiente de trabalho do cliente, incluindo a instalação de aplicativos de IA e GenAI em várias indústrias.
- **"Como um Serviço":** Assine a inovação que cresce com você com o Lenovo TruScale, que fornece serviços de entrega, gerenciamento e atualização de ponta a ponta, o que significa que suas equipes de TI não precisam levantar um dedo ao implantar novos dispositivos e escalar sua infraestrutura de TI.

No cerne desta arquitetura de alto nível estão os servidores da Lenovo com software e GPUs da NVIDIA, que oferecem excelente desempenho para IA e GenAI.

Software e GPUs de Alto Valor da NVIDIA para IA e GenAI

O software de alto valor da NVIDIA representado nesta arquitetura de alto nível inclui:

- **NVIDIA AI Enterprise** é uma plataforma de software AI de alto desempenho, segura e nativa da nuvem, com segurança, estabilidade, gerenciabilidade e suporte de nível empresarial para criar e implantar modelos de IA. Acelera pipelines de IA e simplifica o desenvolvimento e a implementação de IA em produção, cobrindo uma variedade de casos de uso, desde visão por computador até IA e GenAI. NVIDIA AI Enterprise inclui:
  - **NVIDIA NeMoTM (Neural Models)** é um framework abrangente e nativo da nuvem para construir, personalizar e implantar modelos de IA e GenAI. Ele vem com um conjunto abrangente de ferramentas e recursos, incluindo:
    - Uma biblioteca de modelos pré-treinados para várias tarefas, como geração de texto, tradução, reconhecimento de fala e geração de imagens.
    - Um conjunto de ferramentas para personalizar e treinar modelos.
    - Uma plataforma baseada na nuvem para implantar e gerenciar modelos em escala.
    - NeMo Guardrails ajuda as empresas a manterem aplicativos construídos em LLM alinhados com seus requisitos de segurança.
  - **NVIDIA Riva** é um SDK de IA acelerado por GPU para construir e implantar pipelines de IA conversacional totalmente personalizáveis e em tempo real para:
    - Reconhecimento automático de fala (ASR).
    - Avatares digitais de IA conversacional.
    - Sistemas interativos de resposta por voz (IVR).
    - Tradução neural de máquina (NMT).
    - Texto para fala (TTS).
    - Assistentes de voz.
  - **NVIDIA Triton™ Inference Server** é um software de código aberto que padroniza a implantação e execução de modelos de IA em todas as workloads.Triton acelera e otimiza a implantação e execução de modelos de IA em nuvem, data center e dispositivos de edge.
- **NVIDIA Omniverse™** é uma plataforma poderosa em GPUs da NVIDIA que facilita a integração de tecnologias de realidade aumentada (AR) e realidade virtual (VR) em empresas. Fornece um ambiente colaborativo onde equipes podem criar, simular e visualizar mundos virtuais e aprimorar vários aspectos de seus fluxos de trabalho.

A NVIDIA oferece várias GPUs de alto desempenho para ajudar os clientes a implantar workloads de IA, GenAI e outras.

Aqui estão algumas GPUs acessíveis com base na arquitetura NVIDIA Ada Lovelace que são adequadas para essas workloads:

- **NVIDIA L40S** (2U) proporciona aceleração de ponta a ponta para a próxima geração de aplicativos habilitados para IA, desde o treinamento e inferência de modelos GenAI até gráficos 3D e aceleração de mídia. As poderosas capacidades de inferência do L40S, combinadas com ray tracing acelerado por NVIDIA RTX e motores dedicados de codificação e decodificação, aceleram áudio habilitado para IA, reconhecimento de fala, GenAI 2D, vídeo e 3D.
- **NVIDIA L40** (2U) oferece gráficos neurais revolucionários, virtualização, computação e capacidades de IA para workloads de data center acelerados por GPU.
- **NVIDIA L4** (1U) é um acelerador universal, econômico e eficiente em termos energéticos, projetado para atender às necessidades de IA em vídeo, computação visual, gráficos, virtualização e inúmeras aplicações, incluindo jogos em nuvem, simulação e ciência de dados. Fornece alto throughput e baixa latência em qualquer servidor, do edge ao data center à nuvem.

* Mostrado com escassez. As especificações são metade mais baixas sem dispersões.
** Com escassez.

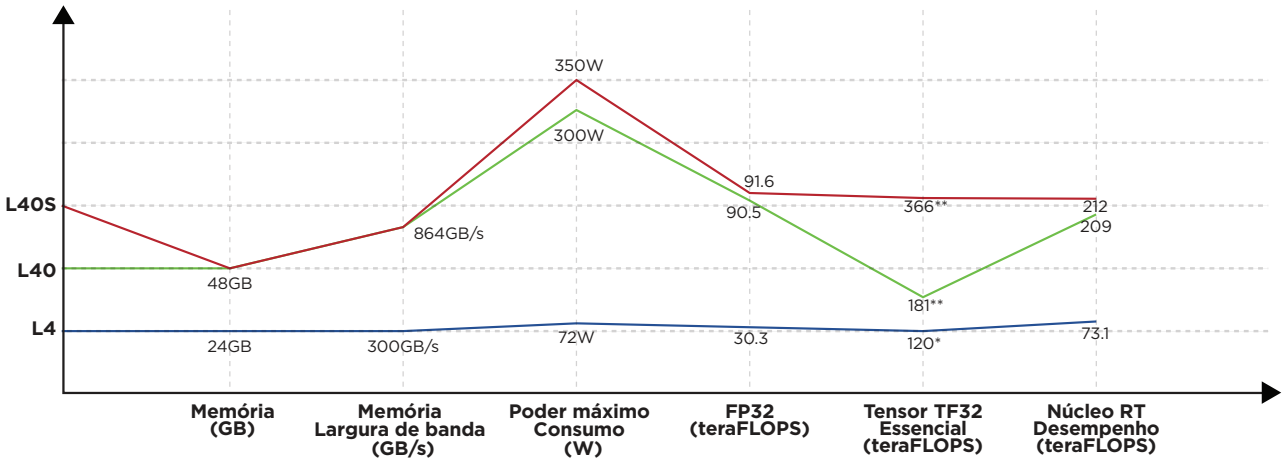


Figura 5: Comparativo de características GPU

A Figura 5 mostra as características-chave dessas três GPUs. A Tabela 1 apresenta um mapa sugerido de afinidade de workloads (O Melhor, Melhor e Bom) por GPU, embora isso dependa das necessidades específicas do cliente.

As células em branco na Tabela 1 significam que a GPU correspondente é excessiva ou insuficiente para aquela workload específica. Por exemplo, para desktops virtuais (VDI) e aceleração de edge distante, o L4 é a única GPU sugerida. O L40 e o L40S são excessivos, mais caros e ocupam mais slots, já que ambos têm 2U. Para treinamento de DL e computação de alto desempenho (HPC), o L40S é a única GPU sugerida devido ao seu desempenho significativamente melhor no Tensor Core TF32. O L40 é o melhor para renderização, com seu excelente desempenho no RT Core e melhor acessibilidade do que o L40S.

Portfólio de GPU NVIDIA e Afinidade de Workload

GPU	Treinamento DL	Inferência DL	HPC/AI	Renderização	Wkstn. Virtual	Desktop Virtual (VDI)	Vídeo AI	Aceleração de Edge Distante
L40S	◎	◎	◎	◎	◎		◎	
L40				●	●			
L4		○		○	●	●	●	●

● O Melhor ◎ Melhor ○ Bom

Tabela 1: Portfólio de GPU NVIDIA para Afinidade de Workload

As Tabelas 2 e 3 apresentam as GPUs sugeridas para treinamento de IA e GenAI (apenas L40S) e workloads de inferência, incluindo LLMs.

Portfólio de Treinamento de GPU NVIDIA						
GPU	NLP/LLM				Imagem/ Vídeo Gen AI	Recsys
	Up to 5B	6B to 65B	66B to 175B	>175B		
L40S	◎	◎	○		◎	○

● O Melhor ◎ Melhor ○ Bom

Tabela 2: Portfólio de Treinamento de GPU NVIDIA

Portfólio de Inferência de GPU NVIDIA								
GPU	NLP/LLM				Imagem/ Vídeo Gen AI	Recsys	Visão Computacional	Vídeo AI
	Up to 5B	6B to 65B	66B to 175B	>175B				
L40S	◎	◎	○		●		◎	◎
L4	○				○		●	●

● O Melhor ◎ Melhor ○ Bom

Tabela 3: Portfólio de Inferência de GPU NVIDIA

Com base nessas GPUs e software da NVIDIA, a Lenovo fornece aos clientes, em muitas indústrias, soluções validadas e otimizadas para desempenho, com escolha e flexibilidade para personalizar com base em casos de uso específicos, workloads, orçamentos e outros requisitos.

Lenovo Oferece a Arquitetura Ideal com a NVIDIA para IA e GenAI

As Figuras 6 e 7 mostram um mapa de alto nível dos servidores Lenovo ThinkSystem com GPUs específicas da NVIDIA. Esses sistemas são projetados e construídos desde o início para atender e exceder os rigorosos requisitos de desempenho de aplicativos e fluxos de trabalho de IA e GenAI da indústria.

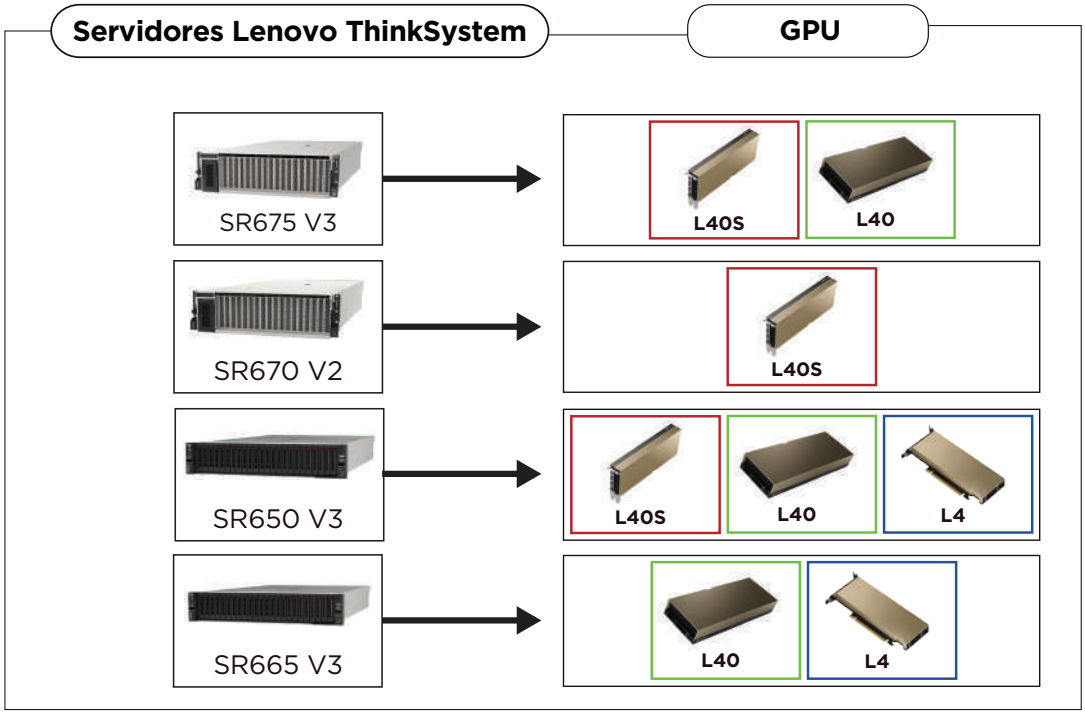


Figura 6: Servidores Lenovo ThinkSystem com GPUs Relevantes Suportadas para IA e GenAI

- O Lenovo ThinkSystem SR675 V3 Server e o Lenovo ThinkSystem SR670 V2 Server são servidores versáteis de rack 3U ricos em GPU que suportam oito GPUs de largura dupla, incluindo as GPUs L40S Tensor Core, com NVLink e refrigeração híbrida Lenovo Neptune líquido-ar. Esses servidores oferecem desempenho ideal em muitas indústrias para IA, GenAI, HPC e workloads gráficas.
- O Lenovo ThinkSystem SR665 V3 Server oferece o máximo desempenho de servidor de dois soquetes em um formato de 2U. É ideal para workloads densos que utilizam processamento de GPU e unidades NVMe de alto desempenho.
- O Lenovo ThinkSystem SR650 V3 Server é um servidor de rack 2U e 2 soquetes ideal para confiabilidade líder do setor, gerenciamento e segurança, maximizando desempenho e flexibilidade para crescimento futuro. Pode lidar com várias workloads empresariais, como bancos de dados, virtualização, computação em nuvem, streaming de mídia etc.

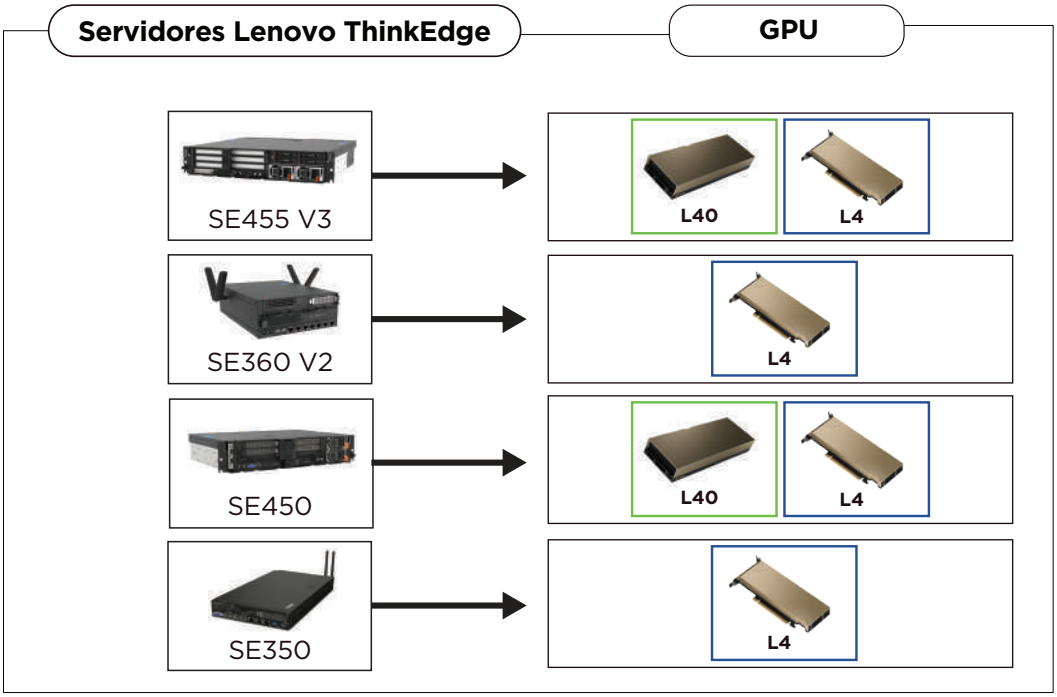


Figura 7: Servidores Lenovo ThinkEdge com GPUs Relevantes Suportadas para IA e GenAI

- O ThinkEdge SE455 V3 Edge Server é destinado a soluções específicas de IA e Telco e suporta estratégias emergentes de consolidação de edge workloads, com grande contagem de núcleos em uma pegada menor.
- O Lenovo ThinkEdge SE450 Edge Server é um servidor de soquete único com altura de 2U e gabinete de profundidade curta que pode ir quase a qualquer lugar, operar silenciosamente em uma ampla faixa de temperaturas e tolerar poeira e vibração.
- O Lenovo ThinkEdge SE360 V2 Edge Server e o Lenovo ThinkEdge SE350 V2 Edge Server têm metade da largura e são significativamente mais curtos do que um servidor tradicional, ideal para implantação em espaços apertados. Eles oferecem aumento de poder de processamento, armazenamento e rede mais próximo à fonte de geração de dados para workloads em tempo real, como AR/VR, vigilância, IA etc.

A Lenovo também fornece valor adicional e várias soluções complementares para ajudar os clientes em sua jornada de IA e GenAI com:

- A Arquitetura de Referência da Lenovo para GenAI baseada em LLMs: A Lenovo criou recentemente esta arquitetura de referência para ajudar os clientes em sua jornada de IA e GenAI.
- Lenovo Inovador em Resfriamento Eficiente de Energia: À medida que as frequências dos processadores e o número de núcleos aumentam, e as GPUs se tornam mais poderosas para fornecer o melhor desempenho, é crucial resfriar esses sistemas de maneira eficiente para evitar problemas de superaquecimento que causam desligamentos, desempenho mais lento e possíveis perdas de dados. Por mais de uma década, a Lenovo tem liderado em tecnologia de energia e resfriamento de data centers e possui várias soluções inovadoras e únicas com dissipadores de calor Especializados ou Líquido para Ar (L2A) e ventiladores de alta velocidade com baixa impedância. Se o resfriamento a ar não for viável, os clientes podem utilizar outras tecnologias de resfriamento líquido no portfólio Lenovo Neptune.

- Aceleração na Descoberta e Adoção de IA: Muitas empresas enfrentam desafios de implementação devido a limitações de recursos e complexidades de infraestrutura, interrompendo o lançamento de iniciativas de IA e GenAI. O programa Lenovo AI Innovators inclui um ecossistema de parceiros de software líderes que colaboram com a Lenovo para fornecer aos clientes soluções de IA e GenAI personalizadas, comprovadas e prontas para implantação para seus casos de uso.
- Laboratório de Descoberta de IA: Trabalhe com especialistas em IA da Lenovo e NVIDIA para obter o máximo valor, reduzindo os riscos do projeto. A Lenovo tem estado na vanguarda da IA há quase uma década. Beneficie-se do Laboratório de Descoberta de IA da Lenovo, workshops de avaliação de IA e um comitê de IA impulsionando a adoção de IA para clientes em todos os continentes.

O Laboratório de Descoberta de IA pode oferecer os seguintes serviços:

  - Acesso a Cientistas de Dados, Arquitetos de Soluções e Engenheiros de Desempenho de GPU.
  - Ajuda na identificação e entrega de soluções de IA que atendam ou superem os KPIs estabelecidos por sua empresa. Identificará e entregará soluções de IA que gerem retorno sobre o investimento, não apenas projetos de IA, por causa dos projetos de IA.
  - Foco em casos de uso em manufatura, varejo, saúde e finanças, mas também realizou muitos projetos em várias outras indústrias.
  - Auxílio na determinação da estratégia de IA e adaptação a novas tecnologias GenAI.
  - Entrega de casos de uso de implantação de visão computacional, como fizemos em muitos projetos, desde a NASCAR até a Island Conservation, até detecção de defeitos de fabricação.
  - Entrega de soluções GenAI para implantações locais que preservam privacidade e segurança, como fizemos com muitos LLMs de código aberto.

- **Prática de Serviços Profissionais de IA da Lenovo:** Oferecendo uma variedade de serviços, soluções e plataformas, a Prática de Serviços Profissionais de IA da Lenovo ajuda empresas de todos os tamanhos a navegar pelo cenário de IA, encontrar as soluções certas e colocar a IA para funcionar em suas organizações de maneira rápida, eficaz e em escala. Ajuda a levar a IA do conceito à realidade - desde a elaboração de mapas de IA até a implantação de plataformas e fornecimento de transparência na utilização de tecnologia com o Lenovo TruScale Hub.
- **Soluções Complementares Inovadoras:** A Lenovo está entregando muitas tecnologias de ponta em estações de trabalho, laptops, tablets, dispositivos móveis, AR/VR (ThinkReality) e computação em nuvem (TruScale), que atendem às necessidades de integração, flexibilidade e experiência imersiva de clientes em várias indústrias.
- **Do Edge ao Data Center até uma Plataforma de Nuvem:** O modelo de computação de IA e GenAI é híbrido, com treinamento feito no data center com servidores ThinkSystem e inferência feita no edge com servidores ThinkEdge (Figura 8).

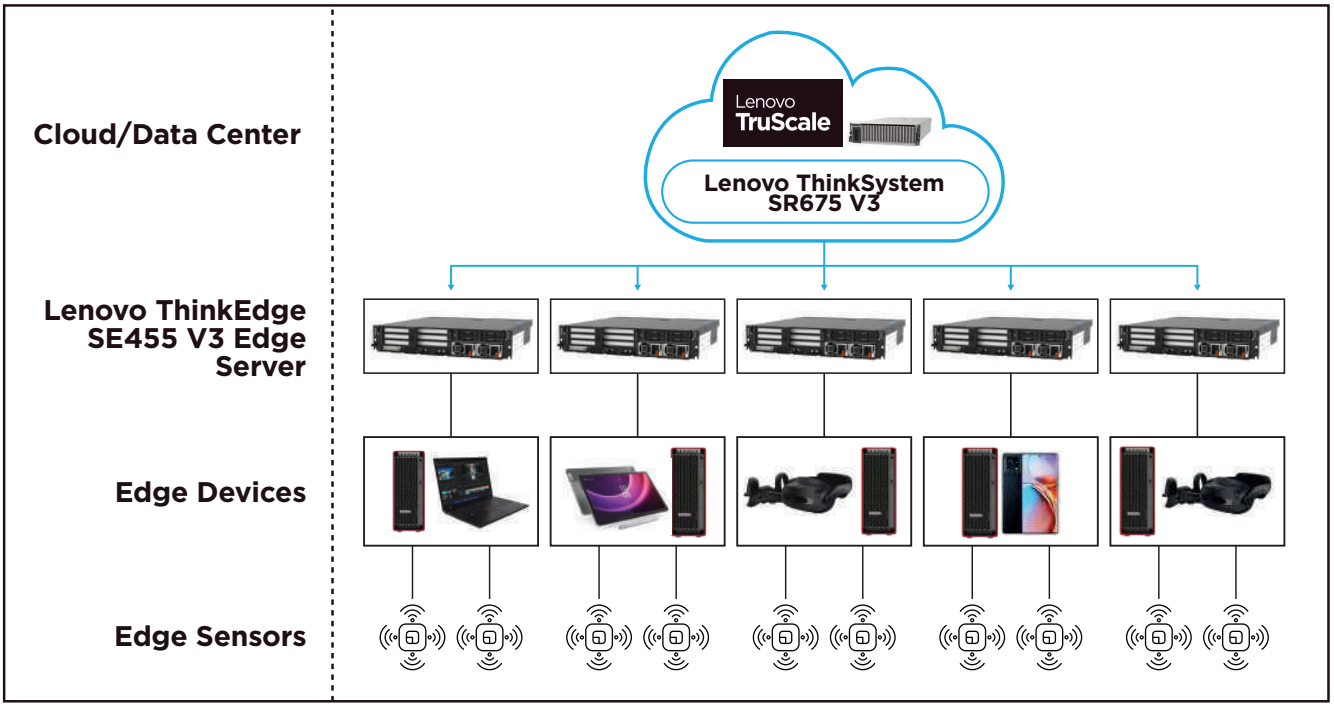


Figura 8: Soluções Lenovo do Edge ao Data Center para IA e GenAI

Casos de Uso do Edge ao Data Center Específicos por Indústria

Aqui estão alguns exemplos específicos por indústria:

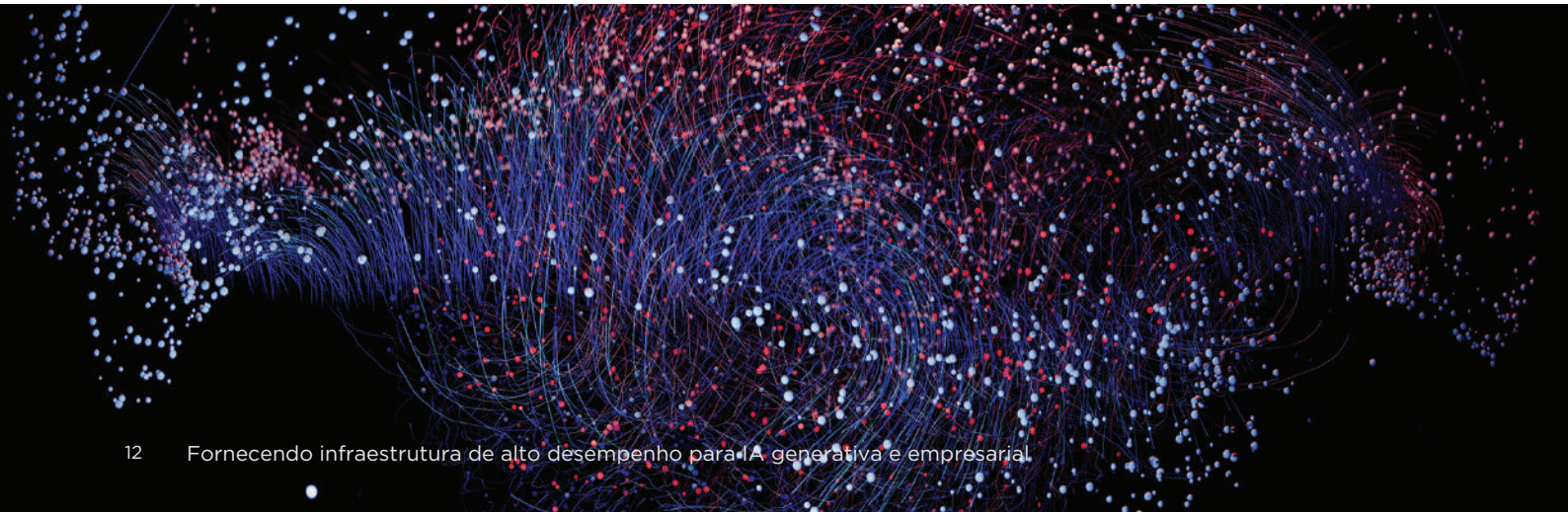
- **Empresas de Serviços Financeiros** podem usar servidores Lenovo ThinkEdge para detecção de fraude em tempo real. Esses servidores agora têm o poder de executar autenticação biométrica (inferência) usando modelos treinados no data center que são periodicamente atualizados com dados reais e sintéticos para melhorar a precisão.
- **Provedores de Saúde** podem usar servidores Lenovo ThinkEdge para monitorar sinais vitais de pacientes e outros dados de saúde em tempo real de seus dispositivos vestíveis. Ao analisar esses dados (inferência) usando servidores ThinkEdge em seus consultórios, os provedores podem identificar problemas de saúde potenciais precocemente e fornecer recomendações de saúde personalizadas aos pacientes. Os provedores podem compartilhar esses dados com organizações de saúde afiliadas, que podem usar esses dados de pacientes anonimizados para construir modelos de treinamento de IA melhores e fazer previsões mais precisas no futuro.
- **Varejistas** podem usar GenAI para criar recomendações personalizadas para clientes em suas lojas. O varejista treina um modelo GenAI com dados de vendas para aprender quais produtos os clientes provavelmente comprarão juntos. Um servidor ThinkEdge em uma loja específica pode executar esse modelo para gerar (inferência) recomendações personalizadas quando um cliente entra na loja.
- **Fabricantes** podem inspecionar produtos com análise de imagem (inferência) em servidores Lenovo ThinkEdge para defeitos na linha de montagem, reduzindo o desperdício e melhorando a qualidade, o design e a fabricabilidade do produto. Esses insights influenciam os novos designs de produtos analisados em servidores Lenovo ThinkSystem de alto desempenho no data center.
- **Empresas de Telecomunicações/Mídia** podem personalizar a experiência de TV para seus clientes. Ao treinar um modelo de IA com dados do cliente em um servidor Lenovo ThinkSystem, uma empresa pode aprender que tipos de programas e filmes os clientes provavelmente gostarão. Esse modelo pode então ser implantado em dispositivos de edge, como set-top boxes, para gerar recomendações personalizadas.

Inicie sua jornada de IA e GenAI com Lenovo e NVIDIA

À medida que as empresas incorporam IA e GenAI como parte de seus processos de negócios essenciais, não podem se dar ao luxo de ter problemas de desempenho, atrasos ou tempo de inatividade. Portanto, o suporte deve ser proativo, realizado por especialistas técnicos que trabalham em estreita colaboração com o cliente e compreendem profundamente seu ambiente.

Como parte do contrato com a Lenovo, as empresas podem contar com um gerente de contas técnico dedicado ou administrador de sistema como ponto de contato único. Seja no local, trabalhando remotamente ou uma combinação de ambos, os profissionais de suporte podem identificar e resolver rapidamente quaisquer problemas, garantindo que o ambiente de IA funcione de maneira otimizada 24/7.

No entanto, a Lenovo vai muito além do suporte técnico especializado. O serviço de ponta a ponta da Lenovo para IA e GenAI inclui consultas iniciais, workshops, análises e configuração do ambiente adequado, passando por avaliação contínua de resfriamento e serviços de monitoramento/manutenção até faturamento e administração. Esses serviços abrangentes podem ajudar os clientes a maximizar o retorno sobre o investimento em suas iniciativas de IA.



A Vantagem Lenovo e NVIDIA

À medida que a IA, especialmente a GenAI, se torna parte integrante dos processos de negócios essenciais de uma empresa, ela precisa superar vários desafios de implementação. A Lenovo e a NVIDIA ajudam as empresas a maximizar o retorno sobre o investimento em suas iniciativas de IA, acelerar o tempo de valor e impulsionar a inovação e produtividade, oferecendo:

- **Sistemas Otimizados para Desempenho:** Servidores ThinkSystem e ThinkEdge alimentados por GPUs e software NVIDIA oferecem excelente desempenho para workloads exigentes de treinamento e inferência em texto, vídeo e modalidades de dados de imagem para casos de uso em várias indústrias.
- **Serviços e Software de Alto Valor:** O programa Lenovo AI Innovators inclui um ecossistema líder em software e parceiros de serviços para fornecer aos clientes soluções personalizadas, comprovadas e prontas para implementação de IA e GenAI, desde consultas iniciais, workshops, análises até configuração do ambiente adequado.
- **Liderança em Eficiência Energética:** A Lenovo lidera em tecnologia de energia e resfriamento de data centers e oferece várias soluções inovadoras e únicas de resfriamento a ar e líquido, incluindo as tecnologias de resfriamento líquido Neptune™.
- **Suporte de Nível Empresarial:** Os sistemas são testados, validados e otimizados para desempenho, gerenciabilidade, segurança e escalabilidade. A Lenovo, ou um parceiro de negócios certificado, oferece instalação no local, inicialização, integração e monitoramento e resolução proativa de quaisquer problemas operacionais.
- **Portfólio Completo de Soluções:** Com a Lenovo, os clientes podem implementar soluções de IA de ponta a ponta usando um amplo portfólio de dispositivos móveis inteligentes, estações de trabalho até servidores ThinkEdge e os servidores ThinkSystem mais escaláveis. Esses sistemas vêm com uma ampla gama de armazenamento, software e serviços abrangentes que proporcionam excelente desempenho, confiabilidade e segurança para o ambiente de TI do cliente, desde o edge até o data center e a nuvem.
- **Roteiro Sólido com Inovação Contínua:** A NVIDIA continua liderando o mercado de GPUs, fornecendo consistentemente um portfólio de GPUs e software de alto desempenho para acelerar as workloads de treinamento e inferência mais exigentes da GenAI, reduzindo o TCO. Da mesma forma, a Lenovo entrega servidores de data center e edge que integram rapidamente essas GPUs NVIDIA com outras tecnologias de ponta em computação em nuvem (TruScale) e AR/VR (ThinkReality), que abordam as necessidades futuras de desempenho, acessibilidade, eficiência energética e experiência imersiva para empresas e seus clientes.

Maximize o Retorno sobre o Investimento em Sua Iniciativa de IA

Por favor, entre em contato com seu representante Lenovo ou envie um e-mail para AIDiscover@lenovo.com para agendar uma consulta inicial com um Especialista em IA da Lenovo ou solicitar um workshop personalizado de IA.

[1] A receita da Infraestrutura de IA da Lenovo ultrapassa US$ 2 bilhões e traz a IA para os Dados com o Portfólio mais Abrangente da Indústria - Lenovo StoryHub

[2] "Inteligência Artificial - Em todo o mundo", Statista

[3] O mercado de software de IA generativa deve ultrapassar US$ 36 bilhões em receitas agregadas até 2028, com uma taxa de crescimento anual composta de 58% entre 2023 e 2028 | S&P Global Market Intelligence (spglobal.com)

[4 e 5] Instituto de IA da Deloitte, "Dossiê de IA Generativa: Uma seleção de casos de uso de alto impacto em seis grandes setores", 2023.

[6] Desafios na Implementação de Aprendizado de Máquina: Uma Pesquisa de Estudos de Caso, 2011.09926v2.pdf (arxiv.org), Jan 2021

[7] Como os Transformadores funcionam? - Curso de NLP Hugging Face

[8] https://lenovopress.lenovo.com/lp1798-reference-architecture-for-generative-ai-based-on-large-language-models#authors.