Primeiros passos com IA

Como Lenovo e Intel estão potencializando aplicações práticas para IA hoje

Lenovo ThinkSystem SR650 V3 servidores construídos com a 5ª geração dos processadores escaláveis Intel Xeon, projetados para IA.

Smarter technology for all

Lenovo

Sumário

3 O rápido crescimento da IA

4 Habilitando a IA em todo lugar

5 O básico

6 Inferência de IA

7 Desbloqueando insights

8 Expertise em transformação por IA

10 Acelerando a implantação

11 Estudo de caso: Experiências do espectador

12 Flexibilidade para escalar sem problemas

13 De olho na sustentabilidade

14 Uma abordagem mais inteligente

O rápido crescimento da IA

A inteligência artificial (IA) fez tremendos avanços desde seus dias pioneiros na década de 1950. As primeiras e rigorosas abordagens de design para aprendizado estatístico e análise preditiva executadas em computadores deram lugar a instâncias iniciais de aprendizado de máquina na década de 1980, quando algoritmos foram ensinados a reconhecer relações e construir modelos de sistemas complexos.

O advento de grandes redes neurais nos anos 2000 abriu caminho para expansões massivas de capacidade computacional e introduziu a capacidade de gerenciar e analisar grandes quantidades de dados complexos e padrões abstratos.

Do ponto de vista empresarial, o potencial para obter insights, reduzir cargas de trabalho e acelerar a produtividade parece quase ilimitado – e as empresas estão investigando ativamente maneiras de colocar a IA para funcionar.

80%

dos CIOs hoje têm a tarefa de pesquisar e avaliar possíveis implementações de IA às suas soluções de tecnologia.[1]

Habilitando a IA em todo lugar

Abordar iniciativas de IA pode ser assustador. Historicamente, a IA tem estado apenas no domínio dos motores de busca, instituições financeiras e pesquisa científica. Além do custo de adquirir o hardware necessário, em muitos casos os centros de dados existentes não suportam o poder adicional e os requisitos de resfriamento, o que necessita de mais investimentos de capital e tempo.

A boa notícia é que a introdução de modelos baseados em IA treinados em dados públicos reduziu as barreiras para as organizações implementarem soluções de IA.

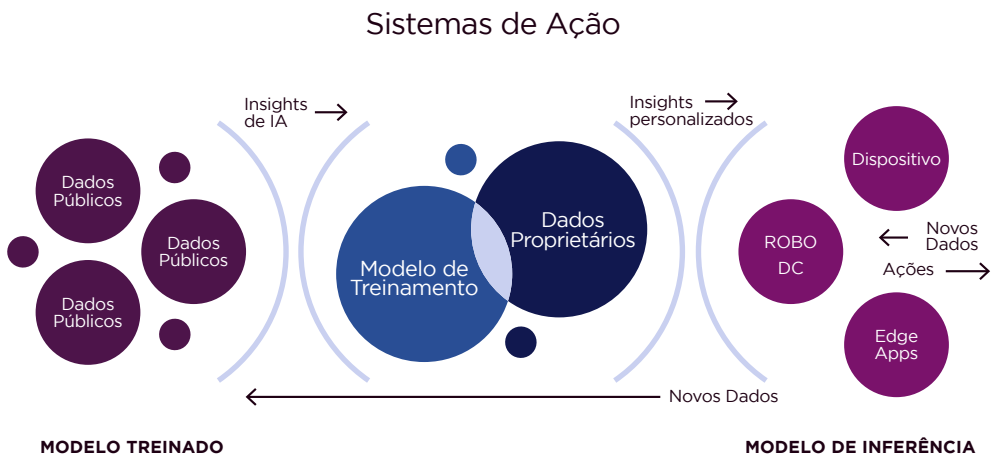
Lenovo e Intel estão utilizando sua longa parceria para trabalhar, entregando soluções que permitem às empresas alavancar toda a grande obra que está sendo feita de forma rápida e prática em maneiras que entregam resultados mensuráveis.

Vamos começar com alguns conceitos básicos

Nos termos mais simples, a IA é definida como qualquer sistema de automação que simula a inteligência humana aprendendo no trabalho. A IA é implementada em duas fases:

1 **Treinamento ou desenvolvimento de modelo:** Esta é a fase onde cientistas de dados desenvolvem e otimizam modelos fundamentais com um conjunto de dados selecionado.

2 **Inferência:** Aplicar novos dados a um modelo treinado para derivar novas informações e acelerar a automação.



Desenvolvimento do modelo de treinamento

O treinamento da IA é alcançado através de um processo chamado Machine Learning (ML), no qual um modelo é treinado com base em parâmetros específicos que definem a tarefa (por exemplo, cor, formas e bordas) e usa técnicas como agrupamento, regressão e redes neurais para extrair esses elementos de quantidades enormes de dados para desenvolver previsões. A partir daí, o modelo continua a consumir e analisar dados enquanto melhora suas capacidades de tomada de decisões futuras.

Os conjuntos de dados usados para o treinamento fundamental cresceram a uma escala que requer grandes quantidades de poder computacional especializado, funcionando em paralelo em milhares de processadores, o que historicamente esteve sob o domínio exclusivo de organizações acadêmicas, financeiras e governamentais.

A quantidade de poder computacional necessário para treinar os maiores modelos de IA é de

3 a 10 meses.[2]

Introduzindo a inferência de IA

A inferência de IA envolve a utilização de modelos treinados existentes e sua aplicação a novos conjuntos de dados proprietários para tarefas específicas de aplicação. Os resultados e insights são então adaptados para novas aplicações que são projetadas para oferecer experiências mais precisas e relevantes.

A inferência expande o aprendizado já realizado, portanto, as demandas de processamento para gerar previsões e insights são significativamente menores do que aquelas necessárias durante o treinamento inicial.

Com as demandas reduzidas de processamento, está se abrindo portas para empresas e organizações de todos os tamanhos aproveitarem o poder da IA para uma ampla gama de aplicações.

Veja como a Lenovo e a Intel estão acelerando a Indústria 4.0 com inspeções visuais assistidas por IA e análises de dados mais rápidas.
Saiba mais.

O número de empresas usando IA cresceu

300% em **5 anos.**[3]

As organizações não precisam desenvolver modelos de treinamento fundamentais, isso acelera dramaticamente o desenvolvimento e as ajudam a mover para aplicações reais mais rapidamente. Além disso, como a abordagem necessita apenas de novas informações a serem aplicadas ao modelo, as demandas de dados e processamento podem se estender além do centro de dados. Isso significa que a inferência pode ocorrer onde os dados são coletados, inclusive na borda.

Isso é importante porque permite funções críticas em tempo real sem penalidades de latência ao enviar dados para frente e para trás até a nuvem, como nos sistemas autônomos encontrados em veículos autodirigidos ou em fábricas automatizadas.

O modelo de inferência de IA treinado funciona apenas com os dados necessários para tomar decisões, o que acelera o processo de decisão e reduz a necessidade de mover grandes quantidades de dados através das redes.

Por exemplo, em um ambiente de manufatura, os servidores de borda que executam modelos de inferência de IA podem usar visão computacional (com base em modelos treinados existentes) para identificar defeitos, tomar decisões e tomar as medidas apropriadas (usando dados locais proprietários) para resolver o defeito mantendo a linha de produção.

Desbloqueie insights nos seus dados mais rapidamente

À medida que as aplicações de IA evoluem, a tecnologia que as suporta também está evoluindo para se adaptar a essas novas expectativas, acelerando e possibilitando o desdobramento da IA em cada etapa, do edge à nuvem. Lenovo e Intel uniram forças para entregar soluções construídas especificamente para aplicações de inferência de IA.

A mais recente geração de servidores ThinkSystem, como o ThinkSystem SR650 V3, são construídos sobre os processadores escaláveis Intel® Xeon® de 5ª geração, desenhados para IA. A aceleração integrada oferece um aumento de desempenho para tarefas de inferência de IA e reduz os requisitos de energia e resfriamento, o que significa que os servidores ThinkSystem SR650 V3 podem ser implementados em data centers existentes em vez de construir novos centros.

Até
2.7x mais desempenho de IA
em qualquer outro CPU[4] com processadores escaláveis Intel® Xeon® de 5ª geração com Intel® AI Engines.

Até
14x mais desempenho de inferência em tempo real para detecção de objetos comparado com os processadores Intel® Xeon® de 3ª geração.[5]

Além disso, Lenovo oferece um portfólio líder de soluções de edge AI, como o **ThinkEdge SE350 V2** e **ThinkEdge SE360 V2** usando processadores Intel® Xeon® D para fornecer insights em tempo real. Os aprimorados recursos de computação e designs flexíveis de implantação suportam múltiplos tipos de cargas de trabalho de AI com desempenho avançado e designs eficientes. Com a IA na borda, as organizações podem capitalizar em informações dinâmicas em tempo real e entregar automação, remediação e insights onde são mais acionáveis — diretamente no front.

Alavanque a expertise em transformação de IA

Projetar e implementar modelos de inferência de IA que entreguem insights confiáveis e acionáveis requer um conjunto muito específico de habilidades e extrema atenção aos detalhes.

O **AI Discover Center of Excellence da Lenovo** reúne especialistas da Lenovo e da Intel para ajudar seus desenvolvedores a criar e acelerar a entrega de aplicações de IA e modelos de inferência.

Nossos especialistas em IA conduzem uma ampla série de workshops para fornecer avaliações de negócios abrangentes, avaliações de TI e blueprints de design documentados.

Engenheiros técnicos, parceiros e cientistas de dados otimizam seus códigos de IA usando frameworks de código aberto para funcionar em servidores ThinkSystem com hardware e software Intel.

Nós podemos te ajudar a aproveitar o conjunto completo de recursos da Intel, como a ferramenta OpenVINO™ e a oneAPI Deep Neural Network Library (oneDNN) para simplificar a implantação de inferência de aprendizado profundo para centenas de modelos pré-treinados.

A Lenovo também oferece uma ampla série de workshops de Serviços Profissionais Lenovo para acelerar sua jornada de transformação de IA.

intel. XEON PLATINUM

Lenovo

O kit de ferramentas OpenVINO™

As barreiras para a adoção de IA geralmente incluem a necessidade de modelos grandes, otimizados, diversificados, uma ampla gama de arquiteturas de XPU (muitas vezes implementadas juntas), e um vasto ecossistema de frameworks de software API para escolher. Implementar IA pode ser um processo difícil e que consome tempo, envolvendo muitas escolhas de fornecedores.

Com todas essas complexidades, os conceitos comprovados muitas vezes não chegam à produção, criando um "cemitério de POCs."

Essas barreiras precisam ser quebradas para criar oportunidades, e é isso que o OpenVINO™ faz ao oferecer um kit de ferramentas de código aberto que suporta uma ampla gama de arquiteturas de XPU e frameworks de software de IA.

intel. XEON PLATINUM

OpenVINO™

Benefícios:

1. Ampla acessibilidade para múltiplas arquiteturas de XPU através de um modelo de código aberto.

2. Uma solução de inferência de IA acessível e eficiente que reduz os custos de adoção e aplicação da tecnologia de IA desde a nuvem até PCs locais.

3. Uma arquitetura aberta que permite a colaboração em todo o ecossistema – desde cientistas de dados criando modelos até desenvolvedores aplicando frameworks de aprendizado profundo em uma variedade de mercados verticais, aproveitando funções de IA múltiplas como processamento de linguagem natural, sistemas de recomendação e IA generativa.

Emparelhado com a **Plataforma Edge da Intel**, soluções nativas de edge podem ser construídas para acelerar iniciativas de edge IA com recursos de otimização de modelo, treinamento e desenvolvimento de aplicativos.

As empresas também podem embarcar e gerenciar de forma segura uma frota de nós de edge, aproveitando os componentes mais adequados e custo-efetivos, seja em ambientes novos ou já existentes, em parceria com nosso ecossistema inigualável para um custo total de propriedade mais baixo.

Veja como a Lenovo e a Intel estão agilizando a adoção de IA com o OpenVINO™. **Saiba mais.**

Acelere sua jornada com soluções comprovadas de implantação

Quando chegar a hora de implantar sua solução de inferência de IA, o programa Lenovo AI Innovators simplifica o processo com soluções comprovadas usando software ISV de melhor categoria em infraestrutura otimizada para IA da Lenovo e da Intel.

A Lenovo e a Intel constroem, testam e validam soluções de inferência de IA com um parceiro ecossistema de Inovadores em IA para garantir implementações suaves e ótimas que mantenham você dentro do cronograma e do orçamento.

- Solução de gerenciamento remoto da **Nybl**
- Solução de inspeção visual assistida por IA da **byteLAKE**
- Soluções de visão computacional, manutenção preditiva e detecção de anomalias da **Guise AI**
- Solução de Análise de Filas e Multidões da **WaitTime**
- Solução da **Sunlight.io** que acelera a transformação digital de restaurantes e drive-thrus
- Solução de inteligência industrial da **Smartia** que conecta e transforma dados em insights acionáveis

Continuamos a monitorar, avaliar e construir relacionamentos com parceiros ISV à medida que suas soluções evoluem.

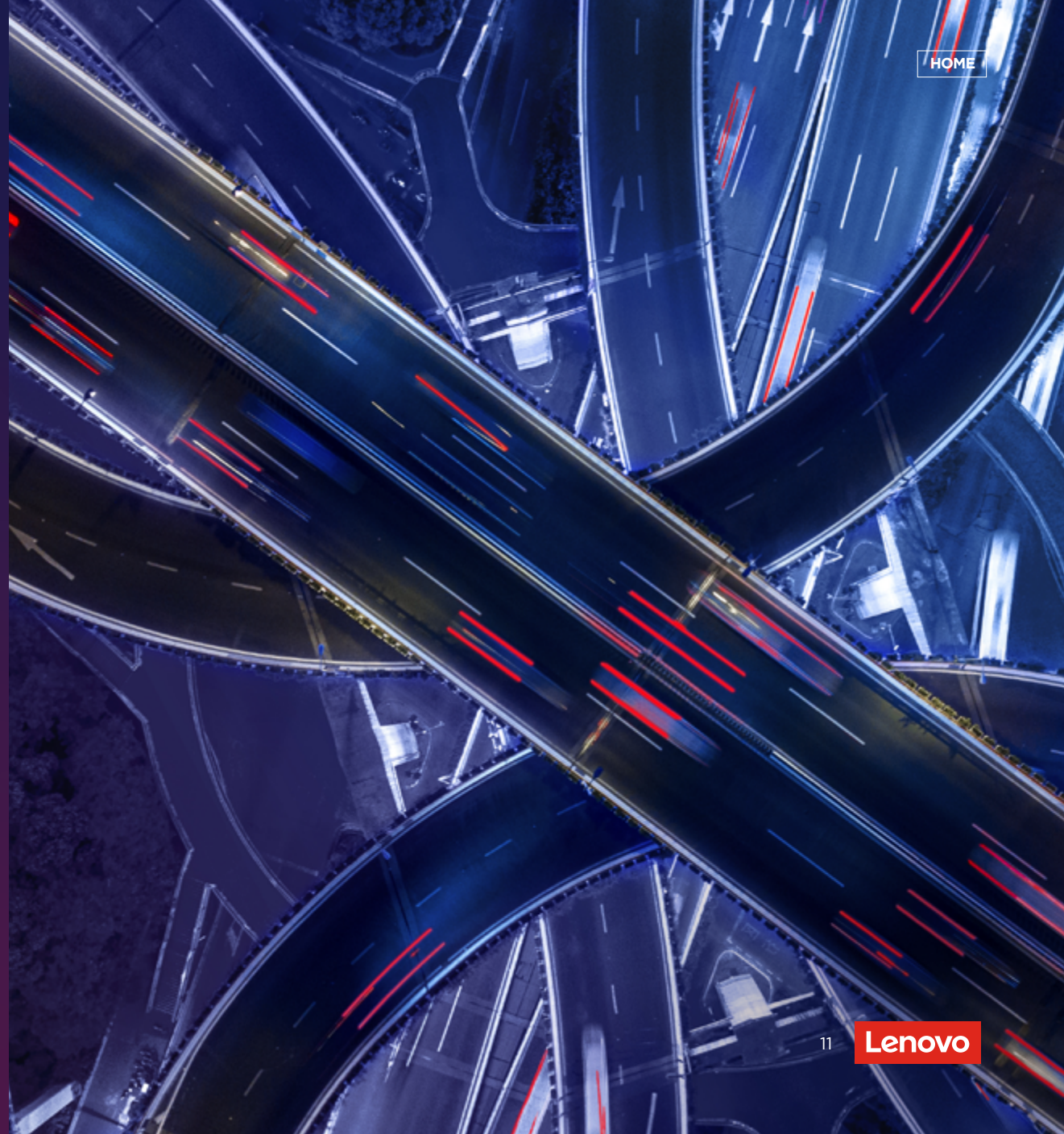
Estudo de caso:
IA está transformando as experiências dos espectadores

Lenovo e WaitTime apresentaram uma solução inovadora para locais de eventos, transformando a experiência dos espectadores da Fórmula 1® com tecnologia de ponta. Ao combinar 18 câmeras estrategicamente instaladas no autódromo Circuit of The Americas (COTA), com a tecnologia patenteada de IA da WaitTime em servidores Lenovo ThinkEdge, alimentados por processadores Intel® Xeon®, os operadores do COTA podem monitorar meticulosamente as multidões e filas de pessoas.

“Esta plataforma de análise de dados em tempo real fornece insights valiosos, permitindo que os operadores compreendam dinamicamente como as multidões estão crescendo e mudando,” disse Zachary Klima, fundador e CEO da WaitTime.

“Essas informações instantâneas empoderam-nos a fazer ajustes em tempo real nas operações e estratégias de receita, garantindo uma experiência ótima e contínua para os espectadores, ao mesmo tempo maximizando a eficiência e a receita para o evento.”

Você pode ler mais sobre a solução **aqui**.



Obtenha a flexibilidade para escalar sem problemas

Implementar inferência de IA requer muito menos despesas iniciais comparado à construção e treinamento de modelos fundamentais do zero, mas ainda existem custos a serem considerados para hardware, software e serviços.

Lenovo TruScale oferece a flexibilidade de um modelo de pagamento escalável conforme o uso para suas iniciativas de inferência de IA — fornecendo acesso a expertise que acelera suas iniciativas.

O modelo OpEx reduz o investimento inicial e escala com as necessidades de negócios em mudança, permitindo que você leve projetos do conceito à implantação e além.

Implementação mais rápida

Ao substituir os requisitos de aprovação de CapEx e mudar para um modelo OpEx, TruScale pode aumentar a flexibilidade e acelerar os tempos de aquisição e implantação.

Opções escaláveis

Escolha entre um contrato fixo ou consumo medido para atender às necessidades de sua organização.

Expertise e serviços integrados de IA

Aproveite os serviços especializados da Lenovo para fechar lacunas de habilidades e recursos e garantir o sucesso da implementação. Além disso, gerentes de sucesso do cliente dedicados da Lenovo podem ajudar a facilitar e coordenar com os recursos da Lenovo.

Essa flexibilidade não só torna mais fácil para uma gama mais ampla de organizações aproveitar a IA, mas também antecipa o futuro da tecnologia e elimina o risco de obsolescência conforme a tecnologia evolui.

IA com um olhar para a sustentabilidade

O aumento da capacidade computacional necessária para treinar e operar modelos de IA resulta em maior consumo de energia e geração de calor, o que continua sendo uma fonte de preocupação global.

À medida que a IA se integra mais nos aspectos do cotidiano, o aumento no poder de computação necessário só tende a crescer.

Por exemplo, uma pesquisa típica no Google consome **menos de 0,3 watt-horas (Wh)** por solicitação. Adicionar um grande modelo de linguagem que reaja à solicitação pode aumentar esse consumo para algo entre **7Wh e 9Wh** por pedido. Considerando o volume atual de buscas do Google, se cada pesquisa incluísse um componente de IA, o Google sozinho poderia consumir cerca de **30 terawatt-horas (TWh)** por ano, aproximadamente o equivalente ao consumo do país da Irlanda.[7]

Treinar um único modelo de IA pode produzir

626,000

libras de CO_2 equivalente.[6]

Lenovo e Intel estão comprometidos com soluções de inferência de IA sustentáveis, eficientes em termos energéticos e ambientalmente responsáveis.

Os processadores escaláveis Intel® Xeon® de 5ª geração são os mais sustentáveis já oferecidos pela Intel para centros de dados, entregando até 10 vezes mais desempenho por watt com aceleradores direcionados para cargas de trabalho específicas.[8] E podem ser implantados em centros de dados existentes sem requisitos adicionais de energia ou refrigeração.

No centro de dados, a tecnologia de medição TruScale pode ajudar você a monitorar consumo de energia, utilização e temperatura para gerenciar o uso e custos de energia de forma mais eficiente. Além disso, nosso Runtime Energy Aware (EAR) software e xClarity Energy Manager ajudam a otimizar o desempenho com um nível baixo de consumo de energia, otimizando estados de energia, desligando componentes não usados e direcionando cargas de trabalho para os recursos mais eficientes.

Otimizar seu centro de dados com a Lenovo TruScale pode ajudar a reduzir emissões de CO_2 até 20%.[9]

Pesquisa Google <0,3Wh

Pesquisa do Google com tecnologia de IA 7–9Wh

Toda a IA do Google pesquisa 30TWh por ano

Uma abordagem mais inteligente para a inferência de IA em qualquer lugar

A inferência de IA possui um tremendo potencial para acelerar o crescimento dos negócios, reduzir cargas de trabalho e otimizar a eficiência para empresas em todas as indústrias.

Não importa onde você esteja na jornada para implementar soluções de IA em sua organização, a Lenovo e a Intel estão prontas para ajudar com soluções feitas sob medida, expertise líder do setor e parceiros da melhor classe.

Visite a **página da Aliança Intel AI** para saber mais.

Fontes

1. Foundry, "State of the CIO Survey 2024"
2. Accenture, "Technology Vision 2023", Março de 2023
3. Tidio, "Os 10 dados essenciais de estatísticas de IA que você precisa saber para 2023", Outubro de 2023
4. Baseado em ganhos de desempenho de 119% a 269% com as extensões Intel® Advanced Matrix Extensions (Intel® AMX) para inferência em GPU-T, LLAMA-2 128, DL RM, DiscBERT, BERT-Large, e ResNet50v1.5 comparado a AMD EPYC 9654 e 9754. See A201, A202, A208-A211 at intel.com/processors/claims. Os processadores escaláveis da 5ª geração Intel Xeon Scalable apresentam resultados que podem variar.
5. Veja A20 at intel.com/processors/claims. Os processadores escaláveis da 5ª geração Intel Xeon Scalable fornecem até 1.4x (BFI6) e 1.3x (INT8) vs. 4th Gen e até 1.4x (BFI6) e 6.7x (INT8) vs. 3rd Gen Intel® Xeon® processors. Resultados podem variar.
6. Universidade de Massachusetts, "Energy and Policy Considerations for Deep Learning in NLP", Junho de 2019
7. Be Davis, "The growing energy footprint of artificial intelligence", Outubro de 2023
8. Baseado no desempenho por watt que vai de 1.46x a 10.6x com aceleradores embutidos em uma gama de cargas de trabalho de AI, banco de dados e redes, veja Site A9-A15, D2, D5, D25, N6 em intel.com/processors/claims: Os processadores escaláveis da 5ª geração Intel Xeon Scalable apresentam resultados que podem variar.
9. TruScale mede de maneira precisa apenas o desempenho e a capacidade, permitindo que infraestruturas gerenciadas sejam projetadas, implementadas e sintonizadas não apenas para desempenho e capacidade, mas também para emissões de CO_2. O monitoramento contínuo do sistema usando Lenovo xClarity Power Monitor e sistemas de desempenho permite otimizar o consumo de energia pela infraestrutura. As emissões de CO_2 são calculadas com base na pegada de carbono localizada da fonte de energia utilizada.

intel. XEON PLATINUM

Lenovo ThinkSystem SR650 V3 servers construídos sobre processadores escaláveis da 5ª geração Intel® Xeon® projetados para IA.