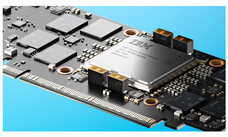
IBM **Power**

Cinco principais razões para executar cargas de trabalho de IA no IBM® Power

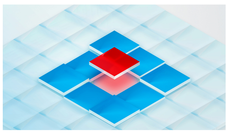
Uma base confiável para potencializar sua estratégia de IA

O Power é uma transformação total de negócios graças à integração perfeita da IA. Seja incorporando a IA em transações, escalando modelos generativos ou implementando aplicações em ambientes híbridos, os servidores IBM Power oferecem o desempenho, a segurança e a flexibilidade para fazer tudo isso, sem interrupção. Vamos analisar melhor.



1 O Power está colocando a IA onde seus dados residem

Em um mundo onde os dados estão por toda parte (no local, na nuvem ou na edge), você precisa de soluções de IA que não sejam apenas inteligentes, mas também estratégicas. O Power leva a inferência de IA para onde seus dados são gerados e residem. Chega de transferir dados sigilosos entre redes ou esperar por GPUs caras e complexas. O Power permite que você incorpore a inferência de IA diretamente onde suas aplicações de missão crítica estão, ajudando a reduzir a latência, os riscos de segurança e custos adicionais. Com o IBM Spyre Accelerator e o IBM watsonx.data on Power, você libera insights mais rápidos integrando um data lakehouse moderno com aceleração avançada fora do chip, criando uma ponte entre dados confiáveis e recursos eficientes de inferência.



2 O Power é segurança 24 horas por dia, 7 dias por semana

Cargas de trabalho de IA são valiosas, então não as deixe vulneráveis a ameaças. Os servidores Power protegem suas cargas de trabalho de IA com segurança integrada em cada camada do stack e resguardam insights sem afetar o desempenho usando criptografia de memória transparente. Para tarefas complexas, como IA generativa, escale a inferência com confiança (e sem preocupações com o desempenho). Com os servidores Power, você também tem detecção de ameaças de ransomware garantida em menos de um minuto, graças ao IBM Power Cyber Vault[1], e até 99,999% de tempo de atividade[2] para sua IA sempre ativa, resiliente e pronta para tudo.

< 1 minuto
para detecção garantida de ameaças[1]

99,9999%
de confiabilidade do tempo de atividade[2]

3 O Power é flexibilidade híbrida, sem atrito

A flexibilidade de TI híbrida é crítica ao implementar cargas de trabalho de IA. Os servidores Power oferecem software corporativo totalmente otimizado para uma experiência de nuvem híbrida sem atritos, permitindo que você migre as cargas de trabalho entre sistemas locais e o IBM Power Virtual Server com facilidade. Seja treinando modelos na nuvem ou fazendo inferência local, o Power oferece a flexibilidade para fazer mais. Além disso, sendo compatível com mais de 130 ferramentas e pacotes de IA de código aberto, a plataforma ajuda suas equipes a criar, implementar e escalar sem encontrar obstáculos.

4 Power é desempenho sustentável sob demanda

Quem disse que os requisitos de sustentabilidade não podem ser atendidos em cargas de trabalho de IA? O Power11 entrega 2x mais desempenho por watt em comparação com servidores baseados em processadores x86, permitindo que você execute a mesma carga de trabalho com menor consumo de energia.[3] E com o novo modo de eficiência energética nos servidores Power11, você pode atingir até 28% mais eficiência energética em comparação com o modo de desempenho máximo.[4]

↑ 28%
mais eficiência energética
do que com servidores x86[4]

5 Power é aceleração sem comprometer a qualidade

Procurando executar IA em escala, em tempo real e sem interromper os fluxos de trabalho existentes? Os servidores IBM Power foram criados justamente para isso. Com alto paralelismo, memória massiva e aceleração integrada ao chip, eles são projetados para incorporar a IA diretamente nos seus fluxos de trabalho. Seus cientistas de dados não precisam refatorar o código para usar a plataforma; eles podem executar cargas de trabalho de IA *no modo em que se encontram*. Além disso, se você executa grandes modelos de linguagem, pode processar até 42% mais consultas em lote por segundo [5] no Power S1022 em comparação com servidores baseados em processadores x86 durante o pico de carga de 40 usuários simultâneos, desfrutando de latência de inferência abaixo de um segundo.[6] Isso significa insights mais rápidos, operações mais fluidas e velocidade de inferência inferior a um segundo.

↑ 42%
mais consultas em lote por segundo no Power S1022
em comparação com servidores x86[5]

Conheça a fundo a IA e a tecnologia IBM® Power →

1. Esta garantia cobre somente a exibição de um alerta em menos de um minuto. A remediação consiste na substituição da unidade, limitada ao valor do produto coberto. Sujeitos aos termos e condições, cujos detalhes estão disponíveis aqui.
2. Com base no downtime não planejado de um único sistema Power E1180, conforme calculado no RAS dos sistemas baseados no processador Power11 (ver seção 99.999% de tempo de atividade) https://www.ibm.com/downloads/documents/br-pt/10a99803d9af7f7e
3. Com base nos dados do índice de desempenho quantitativo (QPI) em 15 de maio de 2025, da IDC, disponíveis em https://www.idc.com/about/qpi, e na utilização. O IBM Power E1150 (4×Power11 de 30 núcleos a 3,0–4,1 GHz) apresentou um QPI de 241.000E, em comparação com o HPE Compute Scale-up Server 3200 (4×Intel de 60 núcleos a 1,9 GHz) com um QPI de 208.898 e taxas de utilização de 75% para o E1150, com base na garantia de utilização de desempenho do IBM Power e 40% para x86.
– O consumo de energia é baseado na potência máxima de entrada: IBM Power E1050 com potência máxima de 5.200 W https://www.redbooks.ibm.com/redpapers/pdfs/redp5684.pdf
– HPE Compute Scale Up Server 3200 com potência máxima de 4.740 W https://www.hpe.com/psnow/doc/a50004268enw.html?jumpid=in_pdp-psnow-qs
4. Com base em medições da IBM sobre desempenho por watt em servidores, comparando o modo de desempenho máximo com o modo de eficiência energética durante a execução de cargas de trabalho baseadas em computação, disco e memória em sistemas Power11 com soquetes e memória totalmente configurados, conforme segue: E1180 com 4x10c / 64x64 GB DDIMM, E1150 com 4x16c / 64x32 GB DDIMM, S1124 com 2x16c / 32x32 GB DDIMM, S1122 com 2x16c / 32x32 GB DDIMM
5. Comparação baseada em testes internos da IBM de inferência de perguntas e respostas usando o modelo PrimeQA (https://github.com/primeqa, com base nos modelos Dr. Decr e ColBERT). Resultados válidos em 22 de agosto de 2023 e conduzidos em condições de laboratório. Os resultados individuais podem variar com base no tamanho da carga de trabalho, no uso de subsistemas de armazenamento e em outras condições. A comparação é baseada na taxa de transferência total em pontuação (inferências) por segundo no IBM Power S1022 (1x20 núcleos/512 GB) executando SMT 8 em comparação com sistemas baseados no Intel Xeon Platinum 8468V (1x48 núcleos/512 GB). O teste foi realizado com ambientes Python e Anaconda, incluindo pacotes de Python 3.10 e PyTorch 2.0. As bibliotecas Python utilizadas são otimizadas para as plataformas Power e Intel. Configuração: tamanho do lote = 60 com 40 usuários simultâneos. O torch.set_num_threads(int) otimizado em uma variedade de níveis de carga.
– IBM Power S1022 (https://www.redbooks.ibm.com/abstracts/redp5675.html): 6,26 consultas com inferência por segundo com 40 usuários simultâneos.
– Sistema x86 comparado: Supermicro SYS-221H-TNR (https://www.supermicro.com/en/products/system/hyper/2u/sys-221h-tnr): 4,4 consultas com inferência por segundo com 40 usuários simultâneos.
– Modelos ajustados pela IBM em um acervo de dados internos da empresa.
6. Com base em testes internos de perguntas e respostas da IBM com inferência usando modelos PrimeQA (baseados nos modelos Dr. Decr e ColBERT). Resultados válidos em 31 de agosto de 2023 e conduzidos em condições de laboratório. Os resultados individuais podem variar com base no tamanho da carga de trabalho, no uso de subsistemas de armazenamento e em outras condições. Com base em resultados para um IBM Power S1022 (2x20 núcleos a 2,9–4 GHz/512 GB) usando uma LPAR de 10 núcleos alinhada a chip NUMA. Os testes foram realizados com ambientes Python e Anaconda, incluindo pacotes de Python 3.10 e PyTorch 2.0. As bibliotecas Python utilizadas são otimizadas para a plataforma Power.
Configuração: SMT 2, torch.set_num_threads(16); tamanho de lote = 1.
–IBM Power S1022 (https://www.redbooks.ibm.com/abstracts/redp5675.html)
–Modelos PrimeQA: (https://github.com/primeqa)