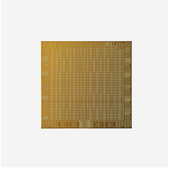
Top 5 reasons to run AI workloads on IBM Power

A trusted foundation to empower your AI strategy



↑42%

more batch queries per second on IBM Power S1022

1 Accelerate efficiently

Leverage IBM® Power® servers—with their AI-optimized hardware, large memory, high parallelism, on-chip acceleration and AI-optimized software—to provide superior performance for your AI workloads. Data scientists can fully use IBM Power platform capabilities without requiring any change to their code. For large language AI models, process up to 42% more batch queries per second on IBM Power S1022 servers than compared x86 servers during peak load of 40 concurrent users[1] and enjoy inferencing latency below one second.[2]

2 Converge AI with data

Deploy AI with enterprise mission-critical processes, data and transactions that resides on IBM Power servers. This convergence allows you to:

- Streamline IT operations with simplified architectures.
- Minimize exposure and risks by keeping the data within regulatory compliant boundaries.
- Reduce latency by bringing AI to data.

4x

crypto engines in every core

60x

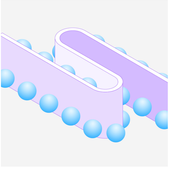
more secure than unbranded commodity servers

99.999999%

uptime for best-in-class reliability

3 Safeguard insights

With IBM Power servers, safeguard AI insights without impacting performance using transparent memory encryption and protect AI workloads with security at every layer of the stack. Scale AI inferencing for complex tasks such as generative AI with reliable performance. IBM® Power10 has 4 times cryptography engines in every core, and IBM Power is 60 times more secure than unbranded commodity servers[3] and provides up to 99.999999% uptime for best-in-class reliability.[4]



4 Hybrid flexibility

Hybrid flexibility is critical when it comes to deploying AI workloads. IBM Power provides that flexibility, enabling enterprises to harness the power of AI both on-premises and in the cloud with IBM Power Virtual Server. In addition to environment flexibility, choice matters for higher levels of the AI solution stack. IBM Power supports multiple AI-optimized software options including:

- Enterprise
- Open-source, community supported
- Open-source, enterprise supported

↓50%

less energy at maximum input power

5 Sustainable and on-demand infrastructure

Meeting sustainability requirements combined with cost-optimized infrastructure to deploy evolutionary AI workloads is a challenge. IBM Power E1050 provides comparable performance and uses 50% less energy at maximum input power than the compared x86-based server[5], allowing clients to run the same work with lower energy usage. At the same time dynamic capacity on IBM Power servers helps clients reduce capital expenditures and procurement costs that can help contribute to lower TCO. Dynamic consumption provides many of the attributes that clients like about public cloud in an on-premises, private cloud with better control and security.

Dive deeper into AI and IBM Power technology →

1. Comparison based on IBM internal testing of question and answer inferencing using PrimeQA model (https://github.com/primeqa, based on Dr. Decr and ColBERT models). Results valid as of 22 August 2023 and conducted under laboratory conditions. Individual results can vary based on workload size, use of storage subsystems and other conditions. Comparison is based on total throughput in score (inferences) per second on IBM Power S1022 (1x20-core/512 GB) running SMT 4 versus Intel Xeon Platinum 8468V-based (1x48-core/512 GB) systems. Test was run with Python and Anaconda environments, including packages of Python 3.10 and PyTorch 2.0. The Python libraries used are platform-optimized for both Power and Intel. Configuration: batch size = 60 with 40 concurrent users. The torch.set_num_threads(int) optimized across a variety of load levels.

   IBM Power S1022 (https://www.redbooks.ibm.com/abstracts/redp5675.html): 6.26 batch queries inferenced per second with 40 concurrent users.

   Compared x86 system: Supermicro SYS-221H-TNR (https://www.supermicro.com/en/products/system/hyper/2u/sys-221h-tnr): 4.4 batch queries inferenced per second with 40 concurrent users.

   Models fine-tuned by IBM on a corpus of IBM internal data: https://github.ibm.com/systems-cto-innovation/ai-on-ibm-systems/tree/master/primeqa/inference

2. Based on IBM internal testing of question and answer inferencing using PrimeQA models (based on Dr. Decr and ColBERT models). Results valid as of 31 August 2023, and conducted under laboratory conditions. Individual results can vary based on workload size, use of storage subsystems and other conditions. Based on results for an IBM Power S1022 (2x20-core 2.9–4 GHz/512 GB) using a chip NUMA aligned 10-core LPAR. Tests were run with Python and Anaconda environments, including packages of Python 3.10 and PyTorch 2.0. The Python libraries used are platform-optimized libraries for Power. Configuration: SMT 2, torch.set_num_threads(16); batch size = 60.

   IBM Power S1022 (https://www.redbooks.ibm.com/abstracts/redp5675.html)

   PrimeQA models: https://github.com/primeqa

   Models fine-tuned by IBM on a corpus of IBM-internal data

3. Information Technology Consulting (ITIC), "ITIC 2022 Global Server Hardware, Server OS Security Report", ITIC, August and September 2022, p.15 (https://www.ibm.com/account/reg/us-en/signup?formid=urx-50800)
4. ITIC 2023 Global Server Hardware, Server OS Reliability Report, August 2023 (https://www.ibm.com/account/reg/us-en/signup?formid=urx-39584)
5. 5 Performance is based on Quantitative Performance Index (QPI) data as of 18 July 2022 from IDC available at https://www.idc.com/about/qpi. IBM Power E1050 (4x24c Power10) QPI of 192,831 versus HPE Superdome Flex 280 (8x28-core Xeon 8280M) QPI of 187,005. Energy consumption is based on maximum input power: IBM Power E1050 with maximum power of 5,200 W (https://www.redbooks.ibm.com/redpapers/pdfs/redp5684.pdf Superdome Flex 280 with maximum power of 10,540 W https://www.hpe.com/psnow/doc/a00059783enw?jumpid=in_lit-psnow-red