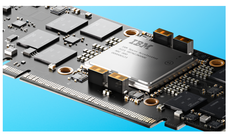
IBM **Power**

# Top 5 reasons to run AI workloads on IBM Power

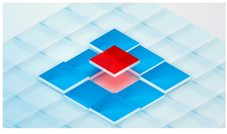
A trusted foundation
to empower your AI strategy

Power is total business transformation thanks to seamless AI integration. Whether you're embedding AI into transactions, scaling generative models or deploying applications across hybrid environments, IBM® Power® servers give you the performance, security and flexibility to do it all—without disruption. Let's take a closer look.



## 1 Power is putting AI where your data lives

In a world where data lives everywhere—on prem, in the cloud, at the edge—you need AI solutions that aren't just smart but strategic. Power brings AI inferencing to where your data is generated and lives. No more moving sensitive data across networks or waiting on costly and complex GPUs. Power lets you embed AI inferencing right where your mission-critical applications are, helping to reduce latency, security risks and additional costs. With IBM Spyre™ Accelerator and IBM watsonx.data® on Power, you can unlock faster insights by integrating a modern data lakehouse with advanced off-chip acceleration, creating a bridge between trusted data and powerful inferencing capabilities.



## 2 Power is 24x7 security

AI workloads are valuable, so don't let them be vulnerable to threats. Power servers protect your AI workloads with security baked into every layer of the stack and safeguard insights without impacting performance using transparent memory encryption. For complex tasks such as generative AI, scale inferencing with confidence—and without worry about performance. With Power servers, you also get less than one-minute guaranteed ransomware threat detection, thanks to IBM Power Cyber Vault[1], and up to 99.9999% uptime[2] for your always-on AI that is resilient and ready for anything.

**<1 minute**
guaranteed threat detection[1]

**99.9999%**
uptime reliability[2]

## 3 Power is hybrid flexibility, no friction

Hybrid IT flexibility is critical when it comes to deploying AI workloads. Power servers offer fully optimized enterprise software for a frictionless hybrid cloud experience, letting you move workloads between on-prem systems and IBM Power Virtual Server with ease. Whether you're training models in the cloud or inferencing on-prem, Power gives you the flexibility to do more. Plus, with support for over 130 open-source AI tools and packages, the platform can help your teams build, deploy and scale without hitting roadblocks.

## 4 Power is sustainable performance on demand

Who says sustainability requirements can't be met on AI workloads? Power11 delivers 2x the performance per watt versus compared x86 processor-based servers, so you can run the same workload with lower energy usage.[3] And with the new Energy Efficient mode on Power11 servers, you can clock up to 28% better energy efficiency compared to the Maximum Performance mode.[4]

**↑28%**
better energy efficiency
versus compared x86 servers[4]

## 5 Power is acceleration without compromise

Looking to run AI at scale, in real time and without disrupting existing workflows? IBM Power servers are built to do just that. With high parallelism, massive memory and on-chip acceleration, they're engineered to embed AI directly into your workflows. Your data scientists don't need to refactor their code to use the platform; they can run AI workloads *as is*. Additionally, if you are running large language models, you can process up to 42% more batch queries per second [5] on Power S1022—compared to x86 processor-based servers during peak load of 40 concurrent users—and enjoy inferencing latency below one second.[6] That means faster insights, smoother operations and sub-second inferencing speed.

**↑42%**
more batch queries
per second on Power S1022
versus compared x86 servers[5]

Dive deeper into AI and IBM Power technology →



1. This guarantee covers only the displaying of an alert in less than one minute. Remediation is in the form of drive replacement up to the cost of the Covered Product. Terms and conditions apply, full details can be found here.
2. Based upon unplanned downtime of a single Power E1180 system as calculated in the Power11 Processor-Based Systems RAS (see section: 99.9999% uptime) https://www.ibm.com/downloads/documents/us-en/10a99803d9afd77fc
3. Based upon Quantitative Performance Index (QPI) data as of 15 May 2025 from IDC available at https://www.idc.com/about/qpi and utilization. IBM Power E1150 (4x30c Power11 at 3.0-4.1 GHz) QPI of 241,000E versus HPE Compute Scale-up Server 3200 (4x60-core Intel cores at 1.9 GHz) QPI of 208,898 and utilizations of 75% for E1150 based on IBM Power Performance Utilization Guarantee and 40% for x86.
– 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
– HPE Compute Scale Up Server 3200 with maximum power of 4,740 W https://www.hpe.com/psnow/doc/a50004268enw.html?jumpid=in_pdp-psnow-qs
4. Based upon IBM measurements of performance per watt on servers comparing Maximum Performance Mode to Energy-Efficient Mode while running compute-, disk-, and memory-based workloads on Power11 systems with fully configured sockets and memory as follows: E1180 with 4x10c / 4x4x64GB DDIMM, E1150 with 4x16c / 64x32GB DDIMM, S1124 with 2x16c / 32x32GB DDIMM, S1122 with 2x16c / 32x32GB DDIMM
5. 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 Aug 22, 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/512GB) running SMT 8 versus Intel Xeon Platinum 8468V-based (1x48-core/512GB) 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 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 queries inferenced per second with 40 concurrent users.
– Models fine-tuned by IBM on a corpus of IBM-internal data.
6. Based on IBM internal testing of question and answer inferencing using PrimeQA models (based on Dr. Decr and ColBERT models). Results valid as of Aug 31, 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 2.9-4GHz/512GB) 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 = 1.
– IBM Power S1022 (https://www.redbooks.ibm.com/abstracts/redp5675.html)
– PrimeQA models (https://github.com/primeqa)