AI on IBM Power

The platform built for enterprise AI

Highlights

Accelerate AI Efficiently: Run AI models with high performance, simplify solution architectures and achieve economies of scale.

Orchestrate AI Flexibly: Consume hybrid cloud infrastructure seamlessly, benefit from elastic consumption of resources, combine enterprise and open-source AI software.

Safeguard AI and Data: Minimize exposure and risks by converging AI with data, secure AI workloads at all layers, and protect data through accelerated encryption.

The use of Artificial Intelligence (AI) is projected to unlock nearly $16 trillion in productivity by 2030[1]. Customers today expect seamless experiences and timely answers to their questions, and companies that fail to meet these experiences risk falling behind. Investment in generative AI is expected to grow 4X over the next 2 to 3 years, but it remains a small fraction of total AI spend[2], and 89% of enterprise decision makers agree that scaling AI leads to competitive differentiation[3].

With AI, especially generative AI, moving from ideation to operationalization, enterprises are looking to choose the right infrastructure that is reliable, provides hybrid flexibility and trusted insights. IBM® Power® provides an accelerated, flexible, and safeguarded platform designed for enterprise AI workloads. Additionally, IBM Power clients have valuable data residing on their IBM Power systems, which helps them to derive trusted insights from their enterprise data and reap the benefits that AI offers.

42%

More batch queries per second on IBM Power S1022 servers than compared x86 servers during peak load of 40 concurrent users when using LLMs

51%

Lower total cost of ownership over a 3-year period running parallel inferencing in Cloud Pak for Data on IBM Power S1022 vs. compared x86 server

Why AI on IBM Power?

Accelerate AI Efficiency
AI-optimized hardware and software empower clients to accelerate AI workloads efficiently without requiring data scientists to alter their code, creating optimization directly out-of-the-box.

- **Improved Performance**: IBM Power10 hardware comes with features optimally suited for AI workloads including an in-core accelerator called Matrix Math Accelerator (MMA). Together with the large memory capacity of IBM Power10 and high parallelism, these differentiators offer efficient and cost-effective acceleration for AI workloads. For large language models (LLMs), clients can 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[4] and **enjoy inferencing latency below 1 second**[5].
- **Run AI on a highly performant, sustainable platform**: IBM Power10 improves the sustainability posture by providing **39% more inferencing per watt** than the compared Intel-based servers[6].
- **Improved Economics**: Clients can leverage the parallel inferencing capabilities and higher utilization on the IBM Power platform to gain **51% lower total cost of ownership over a 3-year period** running parallel inferencing in Cloud Pak for Data on IBM Power S1022 vs. compared x86 server[7].
- **Scale AI solutions** with growing ecosystem.

Orchestrate AI Flexibly
IBM Power provides clients the choice to create and run their AI workloads where and how needed by providing:

- **A frictionless hybrid infrastructure** that is built to be consistent at all layers – infrastructure, operating-system, virtualization, and software - whether on-premises, in a private/managed cloud or in the public cloud.
- **A flexible consumption model** with pay-as-you-use licensing for infrastructure and platform software regardless of where the workload is being executed.
- **A combination of enterprise and/or open-source software for AI** providing the choice of building blocks for creating best fit AI workloads to serve their business needs.

Safeguard AI and Data
Enterprise clients are concerned about safety, risk, vulnerabilities, and compliance. These are all growing areas of concern. AI models may process sensitive data at large scales and, hence, data must be safeguarded by appropriate data governance and security mechanisms.

- Simplify encryption and support end-to-end security with **transparent memory encryption capabilities** on IBM Power without affecting performance by using hardware features for a seamless user experience.
- Minimize latency and consolidate cryptography without having to send data to off-device accelerators **with on-chip cryptographic algorithm acceleration**, which allows algorithms, such as Advanced Encryption Standard (AES), SHA2, and SHA3 to run fast on IBM Power10 servers.
- Protect your applications and data with **secure virtual machine (VM) isolation** with orders of magnitude lower Common Vulnerability Exposures (CVEs) than hypervisors related to x86 processor-based servers.
- Security compliance profiles & real-time updates: Capabilities of **PowerSC** help clients centrally manage, monitor, report, and visualize security and compliance to help support compliance audits, including GDPR.

Flexible choice of best-of-breed enterprise and open-source software combined with a friction-less hybrid platform (on-prem and in the cloud).

AI capabilities on IBM Power

IBM Cloud Pak® for Data
IBM enterprise AI solutions for IBM Power include IBM Cloud Pak for Data. IBM Cloud Pak for Data is a modular set of integrated software components for data analysis, organization and management. IBM Cloud Pak for Data on IBM Power contains a wide range of Watson, Apache, Db2 and Red Hat components which help accelerate data analytics tasks within Cloud Pak for Data. As we continue to grow, additional capabilities and services within Cloud Pak for Data will be made available on IBM Power.

Open-Source Solutions
IBM Power offers community and enterprise supported open-source AI capabilities. Open-source AI solutions on IBM Power are provided through RocketCE and the Rocket AI Hub for IBM Power. RocketCE is a packaging of open-source AI tools that are optimized for IBM Power10, leveraging the IBM Power10 on-chip acceleration; available via Rocket Software's public Anaconda channel (https://anaconda.org/rocketce/repo). Rocket AI Hub for IBM Power is an integrated and freely available set of best-of-breed open-source AI platform tools all optimized for IBM Power such as Katib, Kubeflow, Kubeflow Pipelines, KServe and RocketCE. All tools are delivered as container images that are operated within Kubernetes-based environments such as Red Hat OpenShift. All tools are integrated via Kubeflow and optimized to leverage unique AI hardware capabilities of the IBM Power platform.

watsonx™
As the market continues to adopt foundation models for generative AI use cases, IBM Power is aligned to offer generative AI capabilities with watsonx and well positioned to deliver inferencing capabilities of foundation models. These capabilities will allow clients to deploy generative AI uses cases to improve customer experiences, increase productivity, and optimize business processes. Generative AI takes advantage of IBM Power10 on-chip acceleration to provide a differentiated experience for IBM clients. As the market continues to evolve and compute requirements change, IBM Power's mission is to provide clients with a platform that can meet the demands in a cost-effective, sustainable, resilient, and secure way.

Red Hat OpenShift
In addition to the workloads that support AI initiatives for IBM clients, advancements in Red Hat OpenShift will impact the AI solution architecture. An example of this is the introduction of OpenShift capabilities like Multi-Architecture Cluster (MAC) support. MAC enables clients to have multi-architecture OpenShift cluster with both x86 and IBM Power compute nodes. This capability allows IBM clients to deploy workloads where it makes sense, leveraging strengths and benefits of the different architectures and benefit from optimized workload deployment flexibility.

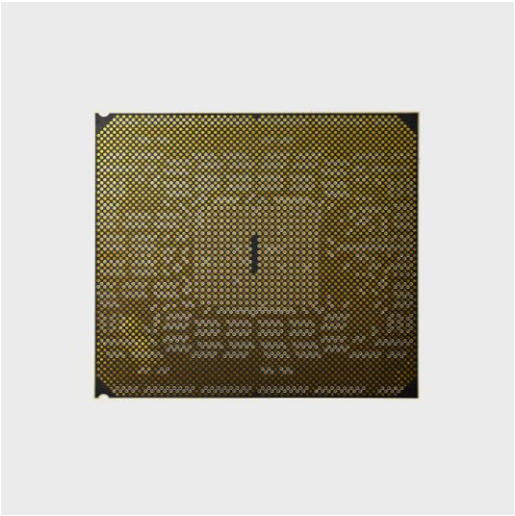

Figure 1. IBM Power10 processor

Conclusion

IBM Power clients now have access to a suite of AI capabilities that leverage IBM Power10's on-chip acceleration, can address the need for both enterprise and open-source solutions, and target key AI market drivers today. These capabilities allow IBM clients to tackle the most salient business challenges today by deriving actionable insights from their ever-growing multimodal data.

IBM believes that AI, especially generative AI, must be tailored for the enterprise. This is espoused via the following tenets:

- **Open:** provide a foundation that is based on the best open-source technologies allowing clients to innovate rapidly with access to an open community and multiple models.
- **Trusted**: provide security and data protection via strict governance and ethics to match ever increasing regulatory and compliance demands.
- **Targeted**: design for enterprise-specific use cases bringing new business value to the client.
- **Empowering**: allow clients to bring their own models and data, build their AI solutions and scale across the enterprise for maximum adoption.

These tenets are delivered via capabilities that encompass the full lifecycle of AI:

- watsonx – the AI and data platform
- Infrastructure for AI – in the cloud or on-premises – IBM Cloud®, IBM Power, IBM Z®, and IBM Storage.

For more information

To learn more about IBM Power, contact your IBM representative or IBM Business Partner or visit www.ibm.com/power.

1. Fortune, April 20, 2023: IBM CEO: 'Today's workforce should prepare to work hand in hand with A.I.'
2. IBM Institute for Business Value, Generative AI: The state of the market
3. Forrester Consulting Thought Leadership Paper: Overcome Obstacles To Get To AI At Scale
4. 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 4 versus Intel Xeon Platinum 8468V-based (1x48-core/512GB) systems. Test was run with Python and Anaconda environments including packages of Python 3.9 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/inferenc
5. 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 3.9 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: https://github.ibm.com/systems-cto-innovation/ai-on-ibm-systems/tree/master/primeqa/inferenc
6. Based on IBM internal testing of data science components, (WML, WSL, Analytic Engine) of Cloud Pak for Data version 4.8 in OpenShift 4.12. Results valid as of 11/17/2023 and conducted under laboratory condition. Individual results can vary based on workload size, use of storage subsystems & other conditions.
7. 1.Based on IBM internal testing of data science components, (WML, WSL, Analytic Engine) of Cloud Pak for Data version 4.8 in OpenShift 4.12. Results valid as of 11/17/2023 and conducted under laboratory condition. Individual results can vary based on workload size, use of storage subsystems & other conditions. 2.The workload mimics a real-time fraud detection logic flow. JMeter is used to submit credit card transactions for different user id and card number combinations. The inferencing application running as microservices in Cloud Pak for Data deployment space extracts the user id and credit card number and uses them to look up 6 previous transactions of the same user and card combination from the Db2 database which is also running within the Cloud Pak for Data cluster. The data retrieved from the database is then combined with the new entry and pass to the LSTM model to determine whether the latest transaction is fraud or not. The score (value between 0 to 1) is returned to the JMeter client as an indicator of whether that transaction is likely a fraud or not. 3.The measurement used for both Power and Intel systems is the throughput result (score/second) reported by JMeter, when running 192 current threads (1 thread representing 1 user) against 96 inferencing end points. 4.Power10 S1022 has a total of 40 physical cores and 2 TB RAM (machine type 9105-22A). There are 7 LPAR on this system including 3 master nodes of 2 cores and 32 GB RAM each, 3 worker nodes of 10 cores and 490 GB RAM each, and a bastion node of 4 cores 128 GB RAM. A local 800 GB NVME drives are used as boot drives for each node, and one 1.6TB NVMe used for NFS server storage running on the bastion node. There is one 100G Ethernet adapters virtualized through SRIOV, with each LPAR taken 10% of network bandwidth. Each LPAR ran with CPU frequency range 3.20GHz to 4.0GHz. All 3 worker nodes ran in SMT 4 mode, while master and bastion nodes ran in SMT 8 mode. 5.The Intel system is Xeon Platinum 8468V with 96 physical cores and 2 TB RAM. The KVM host takes 2 core and 32 GB RAM, which supports 7 KVM guests on this system, including 3 master nodes of 4 cores and 32 GB RAM each, 3 worker nodes of 24 cores and 490 GB RAM each, and a bastion node of 4 cores 128 GB RAM. Local 1.6 GB NVME drives are used as boot drives for these nodes, and one 1.6TB NVMe used for NFS storage on the bastion node. There is one 100G Ethernet adapters virtualized through SRIOV. Each KVM guest ran with CPU frequency range from 2.40GHz to 3.8GHz. All nodes are RHEL CoreOS KVM guests running on the server with hyperthreading enabled. Pricing is based on: Power S1022 (see page 4). Typical industry standard Intel x86 (example on page 5) pricing https://www.synnexcorp.com/us/govsolv/pricing/ and IBM software pricing available at https://www.ibm.com/downloads/cas/DLBOWBPK

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
February 2024