Building Next Generation AI Systems

Accelerating AI, Machine Learning and Deep Learning Initiatives

IBM Corporation
6710 Rockledge Drive
Bethesda, MD 20817

Enterprise AI: A New Technical Challenge

In February 2019, the White House issued an Executive Order on "Maintaining American Leadership in Artificial Intelligence." For many federal agencies, the enterprise-wide use of artificial intelligence (AI) to complete specified tasks is now an essential strategy. In order to develop AI algorithms capable of executing these tasks, teams of data scientists, business experts, and developers must collaborate across an multi-stage process which typically includes the following:

1 **Data Preparation:** Generating, establishing governance, cleansing and formatting data for model training, testing, and production.

2 **AI Model Building:** Developing Machine Learning / Deep Learning models, often with the assistance of open-source frameworks.

3 **AI Model Training:** Train Machine Learning / Deep Learning models, often in high-performance computing environments.

4 **AI Model Deployment & Management:** Save model versions and deploy select versions in the cloud, on devices, or on premise.

5 **AI Model Performance Monitoring:** Evaluate AI model performance while in production to detect drift or modeling errors and trigger model retraining efforts

Yet many factors frequently prevent organizations today from realizing their full potential in AI, Machine Learning (ML) and other areas of data science. Among these factors are:

Compute- and Data-intensive Workloads. The development of accurate AI models requires compute and storage systems capable of processing many thousands of ML training iterations performed on extremely large data sets.

Fragmented Enterprise Data. Data stores are often decentralized, and the movement and duplication of data needed for AI and ML development efforts can be costly, risky, and slow.

Skills Shortage. As diverse analytical frameworks, open-source technologies, and user applications advance rapidly within the competitive AI community, it has become increasingly difficult to employ data scientists possessing the skills necessary to build, train, tune, and deploy best-in-class AI models. Gaps in talent can adversely affect an organization's ability to execute their AI strategy.

Complexity of Model Management. Like living organisms, AI models are designed to update their behaviors (or outputs) dynamically to reflect trends in recent data from external sources. This training process often occurs continuously, meaning that models are frequently changing and must be organized and monitored for accuracy in a continuous manner.

Why IBM?

For decades, IBM has been engaged in the successful design and deployment of innovative high-performance compute and storage environments. For example, during the early 2000s when it became apparent that the advancement of high-performance computing (HPC) could no longer rely on increasingly sophisticated CPUs running at higher and higher frequencies, IBM led the world by introducing two revolutionary systems that not only broke records in both energy efficiency and compute capability, but anticipated the two major approaches to high-end supercomputer architecture employed today: ultra-scalability (as demonstrated first by the Blue Gene/L system at Lawrence Livermore National Laboratory) and the use of accelerators (as demonstrated first by the Roadrunner system at Los Alamos

National Laboratory). Currently, every high-end HPC system in the world employs one or the other (or both) of these approaches.

More recently, IBM embarked on a ‘data-centric’ architecture strategy that has made IBM Power Systems the premier CPU for AI and Machine Learning workloads. The data-centric design entails embedding compute everywhere data resides in order to minimize data movement and achieve faster performance. A key feature of the architecture includes NVLink—a technology developed in conjunction with NVIDIA and Mellanox—which offers the only CPU-to-GPU interconnect in the marketplace, and which was designed to mitigate CPU-GPU throughput bottlenecks often encountered with AI workloads. The Power System AC922 comprises the backbone for accelerated compute and AI clusters built as part of the U. S. Department of Energy’s CORAL program—Summit and Sierra—which were named the world’s fastest computers in 2018 (see inset ‘IBM Power System AC922’ for more detail).

In February 2019, with the support of New York State (NYS), SUNY Polytechnic Institute, and other founding partnership members, IBM announced the creation of a global research hub for the development of next-generation AI hardware among collaborating research and commercial partners. IBM’s acquisition of Red Hat, Inc. in July 2019 facilitates the deployment of AI workloads—primarily built on open source frameworks—in hybrid multi-cloud environments. This means that IT organizations will be able to store and run data and applications for AI workloads in an environment most suitable to deliver the best performance at the lowest cost.

AI Solutions from IBM Systems

As organizations move beyond single-user data science projects toward a more widespread adoption of AI across the enterprise, IBM is committed to developing the technologies that will make such growth possible. As utilization of AI environments continues to expand, AI systems will need to achieve or exceed required performance levels while maintaining ease of use, enabling ease of management, and minimizing workload disruption. IBM’s Systems portfolio addresses the following key capabilities critical for growing enterprise-scale AI at the lowest total cost of ownership:

Key Capabilities

Compute power capable of processing many thousands of machine learning training iterations involving hundreds of combinations of data features and model parameters—with high accuracy and within actionable timeframes.

Scalability of compute and storage resources to accommodate an increasing number of data science users and/or increasing volumes of AI training data, with minimal-to-no user disruption.

Shared compute and storage environments, which makes it possible for organizations to eliminate data silos and reduce the total cost of IT supporting AI workloads.

Flexibility of use, so that data scientists can program ML algorithms using PyTorch, TensorFlow, Caffe, and other common ML/Deep Learning (DL) frameworks.

Simplified model management, such that dynamically changing, re-trained ML models are easily organized and continuously monitored for accuracy and bias.

Explainability sufficient to build user transparency and trust in AI/ML systems—where results generated by AI/ML algorithms are readily traceable to their origin and easily validated.

Simplified application management, which facilitates the installation, configuration, and maintenance and/or upgrade of rapidly advancing data science tools (i. e., notebooks, programming languages, application frameworks).

Up-to-date security protocols across the AI development lifecycle that have been subjected to extensive security scanning and penetration testing, thereby allowing them to be deployed into regulated organizations including major financial and governmental institutions.

Data governance to ensure that data used to train AI models are consistent and trustworthy across the organization.

To address these requirements, IBM offers hardware and software solutions designed to accelerate AI/ML workloads across all stages of development. The heart of the solution is the **IBM Power System Accelerated Compute Server**, which is described in the inset on page 7. In addition, IBM offers the following software user applications to facilitate the development and deployment of AI/ML workloads:

Prepare / Build

IBM Watson Studio. A secure multi-tenant platform that simplifies the process by which data scientists across an enterprise can prepare data for AI models, as well as collaborate on the building, training and deployment of these models in on-premise, cloud, and hybrid cloud environments. Watson Studio makes it possible for data scientists to build AI models using popular open source tools of their choice, including the integrated development environments Jupyter, Zeppelin, and RStudio; analytic environments Apache Spark, sci-kit learn, and XGBoost; as well as DL frameworks such as TensorFlow, PyTorch and Caffe.

Train / Deploy

IBM Watson Machine Learning Accelerator (WMLA). Software that facilitates the training, deployment, and monitoring of AI models developed using Watson Studio and/or other data science tools. With WMLA, common open source ML/DL frameworks (e. g., TensorFlow, PyTorch, Caffe) that have been optimized for superior performance on Power Systems can be run at enterprise scale across 100s of servers in a multi-tenant distributed environment. Refer to the inset on page 8 for additional benefits. Organizations interested in taking advantage of ML/DL frameworks optimized on Power Systems can access Watson Machine Learning - Community Edition (WML-CE) free of charge; however, WML-CE does not extend to enterprise scale. See FAQ on page 9 for more detail.

Monitor

IBM Watson OpenScale. A software application that continuously measures AI model performance for accuracy and bias, and triggers retraining in response to changing business situations to improve model results. OpenScale helps organizations maintain regulatory compliance by tracing and explaining AI decisions across workflows.

IBM software solutions for computer vision (**PowerAI Vision**), video analytics (**IBM Video Analytics**), and automated machine learning (**H2O Driverless AI**) are also available.[†] These solutions have been designed for users without expertise in the development of ML/DL algorithms, and are thus ideal for organizations interested in AI/ML initiatives where this expertise is limited or unavailable.

PowerAI Vision	IBM Video Analytics	H2O Driverless AI
Create DL models for object detection and classification in static images and video without ML/DL expertise. Augment small data sets for better accuracy. Auto-label images to reduce overall training time. Train models on a central server and deploy remotely.	Analyze live and post-event video using configurable, customizable DL models capable of continuous learning and system improvement over time. Track people, objects, and patterns of movement. Analyze and identify established and emerging scenarios without the need for ML/DL experts *Integrated with PowerAI Vision.*	Automate the development of ML models that fit your organization's data. Iterate across 1000s of possible models using best practice model recipes, feature engineering, and parameter tuning. Deploy best models to production with low-latency automatic scoring.

[†] Achieves optimal performance on Power System AC922.

IBM Complete Starter Environments

For organizations just beginning their AI journey, IBM offers complete high-performance analytics environments to meet the computational needs of AI data scientists and software developers. The **Watson Machine Learning Accelerator AI Starter Kit** and **Accelerated Computing Platform** ("**Mini Coral**") are small installations that readily scale upward and outward to support expanding AI workloads, as needed. To simplify the client experience, these Systems solutions are assembled and configured to client specifications at an IBM facility, and then delivered to the client location for fast and easy implementation.

WMLA AI Starter Kit. Includes the following:

- Management Server: One (1) LC922 with 128 GB memory, 5 X 10 TB drives, NFS
- Compute Server: Two (2) AC922, each with 512 GB memory and four (4) NVIDIA V100 GPUs
- Software: Two (2) WMLA Licenses with RHEL 7.6
- Implementation: Approximately 5 days onsite stand-up with IBM Systems Lab Services

Accelerated Computing Platform ("Mini CORAL"). An HPC cluster modeled after the fastest computers in the world in 2018—i. e., Summit and Sierra from the U. S. Department of Energy's CORAL program—at a scale suitable for AI/ML workloads in smaller organizations. These implementations are comprised of the following building blocks, which together comprise a complete turnkey solution for 'AI as a Service:' AC922 compute nodes, LC922 management/login nodes, Mellanox switches, IBM Spectrum Storage software, IBM Elastic Storage Server (optional), and AI/ML software licenses (optional). IBM and/or IBM business partners will work with individual clients to determine which Mini CORAL configurations best address their functional, technical, and budget requirements.

IBM Power System AC922

Power System AC922 currently serves as the backbone of the world's fastest supercomputer—the U. S. Department of Energy's Summit system at Oakridge National Laboratory—which delivers 200+ PetaFlops of HPC and 3 ExaFlops of 'AI as a Service' performance.

- ✓ Specifically designed for enterprise-scale AI
- ✓ Massively parallel multi-threaded cores, large memory bandwidth, low latency, and high I/O throughput to accelerate the training of AI models
- ✓ Supports AI models approximately 60x larger than that supported by x86 servers via sharing of GPU and system memory
- ✓ Easily scalable from a single server to 1000s of nodes with near linear scaling efficiency

- Two (2) IBM Power CPUs
- Two (2) to six (6) NVIDIA Tesla V100 GPUs

- **PCIe Gen4** with 2X the data bandwidth of the PCIe Gen3 currently found in x86 servers
- **OpenCAPI** for low-latency, high-speed, direct memory connections of the CPU to FPGAs and other accelerators
- **NVLink 2.0**, which provides the only CPU-to-GPU interconnect in the marketplace. Enables 5.6X faster data sharing than the PCIe Gen3 found in x86 systems, and it mitigates workflow bottlenecks at the PCIe Gen3 that often occur between CPUs and GPUs when AI workloads are run in x86 servers.

Support & Maintenance. All components specified in this proposal that are acquired from IBM—including compute servers, storage, and software (with AI frameworks)—are fully backed by IBM Levels 1-3 support. IBM Systems Lab Services can provide pre-configured or custom services, professional skills transfer, off-the-shelf training, and online and classroom courses to meet the specific needs of your organization.

Financing. IBM Global Financing provides numerous payment options to help organizations acquire the technologies they need. Detailed information about these payment options can be found at *www.ibm.com/financing*.

IBM Watson Machine Learning Accelerator (WMLA)

Software that enables the distributed processing of ML/DL workloads in an elastic, scalable, and secure multi-tenant environment. Modeling workflows are developed, managed and launched with Watson Studio & Watson Machine Learning Accelerator in on-premise, cloud, or hybrid cloud architectures. WMLA achieves optimal performance with AI workloads on GPU-enabled Power System AC922.[††]

Enhance the user experience for open-source frameworks including:

TensorFlow PyTorch Caffe Caffe2 Chainer

Key Capabilities

- ✓ **Distributed Deep Learning**: Distribute the training of DL models across a multi-tenant, shared, high-performance cluster, with runtime isolation of workloads and near linear scaling to 100s of GPUs.
- ✓ **Large Model Support**: Combine system memory with GPUs to support more complex models and higher resolution data.
- ✓ **Auto-Hyperparameter Optimization**: Continuously refine DL models by testing 10s of hyperparameter values in parallel and selecting only those yielding the best results.
- ✓ **Elastic Distributed Training:** Dynamically schedule ML/DL workloads across compute servers—at the level of the GPU—based on resource requirements and job priority.
- ✓ **Elastic Distributed Inference**: Run multiple inference models on shared production infrastructure at scale, with runtime isolation of inference workloads.
- ✓ **Security:** Employ the latest security protocols—subjected to extensive vulnerability and penetration testing—across the entire AI workflow. Suitable for regulated organizations including major US, Canadian, European, and Chinese financial and government institutions.
- ✓ **Accounting**: Monitor cluster usage for detailed chargeback reporting.

Key Benefits

Facilitates AI/ML development initiatives by supporting popular open-source AI frameworks preferred by today's leading data scientists

Accelerates the training of AI models, thereby shortening time to model deployment

Reduces the total cost of IT infrastructure through the scaling and distribution of AI workloads from different users across shared compute and storage servers

[††] Runs on x86 servers; however, does not achieve the performance levels possible with Power AC922.

FAQ: Beginning an AI Journey with IBM

Whether an organization is experimenting with AI on a single server, ramping up AI efforts across the department, or trying to optimize the performance of existing AI workloads across the enterprise, IBM can provide software and server infrastructure that meets the organization's needs at any stage of AI development.

How can my organization learn more about AI and IBM's AI offerings?

Reach out to your IBM representative via the contact information located on the first page of this document. Your representative will schedule an **onsite briefing** for your team with AI and IT experts from IBM and/or IBM business partners. During the briefing, which will be tailored to your organization's stage of AI development and business objectives, IBM will discuss best practices in AI technology implementations and relevant IBM offerings.

My team has been asked to expand the utilization of AI within our organization. How can IBM help us get started?

Your IBM client representative can schedule an onsite 4-hour **AI Discovery Workshop** for your team. This activity is sponsored by IBM and IBM business partners, and it will be provided to your team free of charge. During the workshop, your team will work alongside AI and IT technical experts from IBM and/or IBM business partners to uncover use cases for AI that will help your organization achieve business and/or research objectives. Our team will also help your organization prioritize the implementation of these use cases based on the value each would bring to your organization; the capabilities of your existing IT infrastructure; and an understanding of your staff, time, and budget constraints. The AI Discovery Workshop can be held at any stage of AI development. Use cases requiring limited technical resources may result in a Proof of Concept implementation, free of charge, if there is agreement that the result would mutually benefit your organization and IBM.

Does IBM offer any trial software evaluations?

To help minimize the risk of AI investments, IBM offers **trial evaluations** of 30 days (or more with extension) for Watson Studio, Watson OpenScale, Watson Machine Learning/Accelerator, PowerAI Vision, and H2O Driverless AI. These applications can be tested in your local IT infrastructure and/or hosted in an IBM data center. If you are interested in a trial evaluation, please contact your IBM representative. He/she will connect you with a technical specialist who can help your team access and deploy the trial software.

Why should my organization invest in Watson Machine Learning Accelerator (WMLA) when the Community Edition (WML-CE) can be accessed at no cost?

If your organization is planning to expand AI initiatives over time, then an investment in WMLA is highly recommended. WMLA was designed specifically to enhance user experience and system performance during the development and deployment of AI models in enterprise environments. The table below compares features of WML-CE and WMLA.

	WML-CE	WMLA
Distributed Deep Learning	Up to 4 server nodes	More than 4 server nodes
Job Scheduling & Cluster Management	-	IBM Spectrum Conductor
Large Model Support	Version 2	Version 2 (same as CE)
Elastic Distributed Training	-	Available
Elastic Distributed Inference	-	Available
Auto-Hyperparameter Optimization	-	Available

My organization currently runs high-performance analytics workloads on Intel x86 processors, yet we are currently unable to invest in new hardware. Can we still take advantage of the enterprise capabilities of WMLA?

Yes. WMLA runs on both IBM Power and x86 architecture servers, allowing clients to select the platform that best meets their requirements. However, WMLA on Power offers greater performance, scalability, and large model support. A comparison of WMLA on Power AC922 and x86 architectures is summarized below.

	Power AC922	x86
Distributed Deep Learning	More than 4 server nodes (up to 100s of nodes)	Up to 2 server nodes
Large Model Support	Version 2 (offers 2x-3x the image support of Version 1)	Version 1
Job Scheduling & Cluster Management	IBM Spectrum Conductor	IBM Spectrum Conductor
Elastic Distributed Training	Available with RDMA	Available without RDMA
Auto-hyperparameter Optimization	Available	Available

My organization plans to move forward with our AI/ML efforts in a cloud environment. Can we utilize IBM AI/ML software offerings in cloud environments?

Yes. IBM supports AI/ML initiatives in on-premise, private cloud, public cloud, and hybrid cloud environments. It should be noted, however, that movement to cloud-based platforms does not always guarantee the lowest total cost of ownership and/or better workload performance. IBM technical specialists will work with your organization to determine which environment best meets your workload, technical and budget requirements. The acquisition of Red Hat, Inc. in July 2019 facilitates the deployment of your AI/ML/DL workloads in hybrid multi-cloud environments (see ‘Why IBM?’).