IDC ANALYZE THE FUTURE

IDC TECHNOLOGY SPOTLIGHT

Sponsored by: IBM

It is increasingly clear that artificial intelligence (AI) requires unique hardware and software combinations that are co-optimized in order to provide data scientists with the best solution possible for developing high-quality AI models fast, while IT is assured of the soundness of the hardware-software package.

Scaling AI Requires Hardware and Software That Work Well Together

July 2019

Written by: Peter Rutten, Research Director, Infrastructure Systems, Platforms, and Technologies Group

Introduction

Around the world and across industries, businesses and other organizations are starting to leverage artificial intelligence (AI) at scale. Of the many use cases being developed, the five most common are as follows: automated customer service agents, sales process recommendation engines, automated threat intelligence systems, fraud analysis, and automated preventive maintenance. IDC has identified more than a dozen additional AI use cases that are being developed in such industries as transportation, manufacturing, education, and healthcare. These use cases leverage a variety of advanced software platforms for AI such as conversational AI, predictive analytics, text analytics, voice and speech analytics, and image and video analytics.

Over the past year or two, among early adopter organizations, lines of business (LOBs), IT staff, data scientists, and developers have journeyed along a learning curve to define an AI strategy for their business, launch initial AI initiatives, and develop and test AI applications. They are now preparing to scale these initiatives, and initial experiences with running their AI models have taught them that standard, multipurpose infrastructure will not suffice for scaling their AI applications.

AT A GLANCE

KEY TAKEAWAYS

The right solution for AI must provide co-optimized hardware and software with:

- Acceleration
- CPU performance
- Fast interconnects
- Sufficient memory
- I/O bandwidth
- Security
- Power requirements
- Scalability
- Heat dissipation
- Open source APIs, libraries, software development kits, toolkits, frameworks, programming languages, etc.
- Accelerator software and tools
- Cluster management software
- Antibias software tools
- Hybrid cloud deployment software

AI parses vast amounts of data and requires powerful parallel processing capabilities based on many more cores than CPUs can deliver. Parallel processing is best achieved with clustered servers that have multithreaded CPUs combined with multicore co-processors such as graphics processing units (GPUs), fast interconnects, large amounts of memory, and advanced I/O capabilities. Furthermore, the three fundamental AI stages of preparing the data, training the model, and inferencing on the AI model are extremely sensitive to how well hardware and software work together. A well-tuned combination of hardware and software can make a world of difference.

Definitions

IDC defines various AI-related terms used in this document as follows:

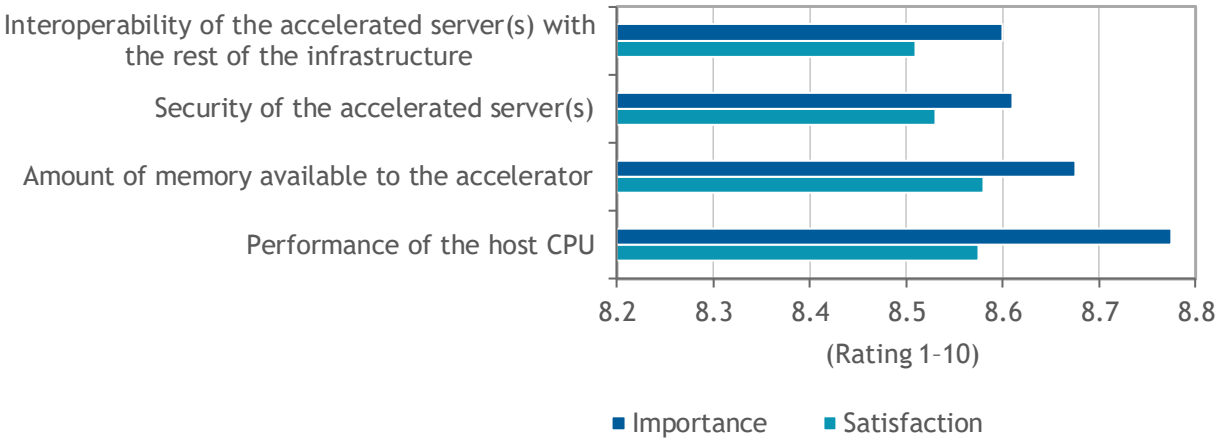
- Machine learning (ML) is a subset of AI techniques that enables computer systems to learn from previous experience (i.e., data observations) and improve their behavior for a given task. It is the process of creating a statistical model from various types of data that performs various functions without having to be programmed by a human. Machine learning models are "trained" by various types of data (often, lots of data). ML techniques include support vector machines (SVMs), decision trees, Bayes learning, k-means clustering, association rule learning, regression, neural networks (NNs), and many more.
- In machine learning, "training" usually refers to the process of preparing a machine learning model to be useful by feeding it data from which it can learn. "Training" may refer to the specific task of feeding that model with the expectation that the resulting model will be evaluated independently (e.g., on a separate "test" set), or it might refer to the general process of feeding it data with the intention of using it for something.
- With regard to machine learning in general, "inference" refers to the process of taking a model that's already been trained (refer back to the previous bullet) and using that trained model to make useful predictions, or it refers to the process of inferring things about the world by applying a model to new data.
- Neural networks or artificial NNs are a subset of ML techniques, loosely inspired by biological neural networks. They are usually described as a collection of connected units, called artificial neurons, organized in layers.
- Deep learning (DL) is a subset of NNs that makes the computational multilayer NN feasible. Typical DL architectures are deep neural networks (DNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), generative adversarial networks (GANs), and many more.
- Deep learning training is the phase in which a deep neural network tries to learn from the data provided to it. In training, each layer of data is assigned some random weights and the neural network classifier runs a forward pass through the data, predicting the class labels and scores using those weights. The class scores are then compared against the actual labels, and an error is computed via a loss function. This error is then back propagated through the network, and weights are updated accordingly via a weight update algorithm. This occurs repeatedly until the model is sufficiently trained to predict accurately at a level decided upon by the developer. Training is usually compute intensive and can take billions and potentially trillions of math operations to develop what will become the production model. Training also requires significant amounts of data that need to be analyzed.
- Deep learning inference is the stage in which a trained model is used to infer/predict the testing samples and comprises a forward pass similar to that of training to predict the values. Unlike training, it doesn't include a backward pass to compute the error and update weights. It's usually a production phase where the deep learning neural network model is deployed to predict real-world data. The trained deep learning neural network is deployed using what it has learned to recognize images, spoken words, and so on, typically as part of an application. This production model of a neural network infers things about new data with which it's presented based on its training.

Hardware and Software Co-Optimization

Data scientists and IT staff have different perspectives on the requirements for AI. Data scientists need the best tools with which they are familiar to be as productive as possible and build high-quality AI models fast. For IT staff, the environment in which data scientists work needs to be easy to integrate into the overall datacenter, manageable, affordable, secure, and extremely high performing in order for them to satisfy the needs of their data science colleagues.

IDC has found that when it comes to accelerated servers for AI, the top 4 characteristics deemed most important by survey respondents have lower satisfaction in terms of infrastructure performance. Figure 1 demonstrates that interoperability, security, accelerator memory, and host CPU performance are all deemed highly important, but respondents — to varying degrees — are not satisfied with how well their current solutions are performing on these characteristics. In other words, end users are essentially asking vendors to come up with innovations to improve these AI infrastructure characteristics.

FIGURE 1: ***Importance Versus Satisfaction Levels for the Top 4 Most Important AI Infrastructure Characteristics***



Source: IDC, 2019

Performance, in this context, should be looked at as a combination of hardware and software. Sometimes the hardware and software are available together from a server vendor in an optimized fashion; sometimes this is true only for a portion of the environment (e.g., NVIDIA GPUs and CUDA are highly optimized together); and sometimes it's up to the end user to make it all work together. AI deep learning hardware and software considerations are covered in the next two sections.

The Right Server for AI Must Provide:

» **Acceleration.** Training AI models require co-processors that facilitate extensive parallelization and that can process massive volumes of data. Three types of co-processors enable this: GPUs, FPGAs, and ASICs. GPUs are the most flexible and have the most advanced software ecosystem. FPGAs can be faster than GPUs and save energy, but they are somewhat harder to program. ASICs can achieve the highest performance with the least amount of energy, but they require very long and costly development times. FPGAs and ASICs are generally used for very large-scale AI implementations.

- **CPU performance.** Even with one or more GPUs, the CPUs still play an important role in terms of maximizing the utilization of those GPUs. The CPUs perform all the typical CPU functions, plus they initiate GPU function calls. If the CPUs execute data pre-processing in an AI deep learning training session, their ability to parallelize is also important in terms of both number of cores and number of threads per core. The CPUs' clock speed plays a role too, even though most of the processing is done by the GPUs — lower clock speeds decrease performance.
- **Fast interconnects.** On a system with multiple GPUs, having sufficient numbers of PCIe lanes makes a difference in terms of performance. For example, on a server with 4 GPUs, 8 PCIe lanes per GPU will help improve AI training performance. Using NVIDIA's NVLink instead of PCIe will increase performance significantly.
- **Sufficient memory.** Memory is important in a GPU-accelerated server, although the bottleneck for AI training is the maximum available GPU memory, which is currently fixed at 32GB per GPU — not huge, therefore. Data scientists complain quite often about GPU memory limitations, and scaling up on GPUs helps. NVIDIA enables scaling up to as many as 16 GPUs using its NVSwitch switching technology to enable these GPUs to communicate with each other. Having sufficient system RAM plays a role if that RAM can be accessed by the GPUs, which is a solution that IBM has implemented: It's called Large Model Support on POWER9 with integrated NVLink 2.0 (see the Large Model Support section for more details). Memory coherency between the processors, co-processors, and memory is very important as well.
- **I/O bandwidth.** Pulling data from storage during AI training requires fast storage and high-bandwidth I/O. If data is preloaded, this plays less of a role, but with training models getting larger and larger, preloading is not always possible. NVMe SSDs with high-bandwidth interconnects will be critical.
- **Security.** For privacy and compliance reasons, the data that is used to train a model needs to be secure. Increasingly, AI applications are being built with enterprises' core data. Hardware, middleware, and software security solutions are critical on the system that executes AI training models using this data.
- **Power requirements.** GPUs are power hungry. NVIDIA's Tesla V100 can use up to 300 watts per unit. The power supply unit (PSU) of a server that will be performing AI training needs to be able to feed energy to the CPU and GPUs and various other power-consuming components as efficiently as possible.
- **Scalability.** Scalability is a critical requirement on multiple levels. An AI platform needs to be able to scale up with high linearity (e.g., to 4 or 8 GPUs within a single server) and then scale out to a cluster of nodes with high efficiency. Software is required to orchestrate the data flow on a cluster.
- **Heat dissipation.** GPUs consume a lot of power and dissipate a lot of heat. A single standalone server with one or two GPUs can be air cooled, but in clustered systems with high compute density and multiple GPUs per node, air cooling is not sufficient, and liquid cooling needs to be used. Liquid cooling technology is highly efficient, and its use in datacenters is increasing.

Software for AI Should Include:

- **Open source APIs, libraries, software development kits, toolkits, frameworks, programming languages, and so forth.** The number of tools, libraries, and frameworks available for AI deep learning training and inferencing is staggering and too extensive too list here. There are more than 17,000 repositories on GitHub under the topic "deep learning," which is why it is helpful when server vendors pre-package the most popular and useful open source software with their AI-targeting server offerings — luckily, most vendors do this. It's even better when they package open source software with a security layer so that enterprises don't have to worry when their business-critical data is being leveraged for AI model training. Popular AI frameworks include:
  - Accord.NET
  - Amazon Machine Learning
  - Apache Mahout
  - Caffe
  - Keras
  - Microsoft Cognitive Toolkit
  - Scikit-learn
  - TensorFlow
  - Theano
  - Torch
- **GPU management software and tools.** There is an extensive ecosystem of software for managing GPUs, developing on GPUs, executing AI training and inferencing on GPUs, and even running use case–specific containerized applications on GPUs. Many of these tools are open source; others are proprietary but not monetized by NVIDIA and downloadable for free. For example:
  - CUDA is a parallel computing platform and API model that allows developers to use GPUs for nongraphical, general-purpose workloads.
  - Data Center GPU Manager helps with the administration of GPUs that are running in a server cluster with configuration, monitoring, diagnostics, and other functions.
  - RAPIDS is a suite of open source software libraries and APIs for executing end-to-end data science and analytics pipelines on GPUs, including data preparation.
- **Cluster management software.** Scalability is essential for AI deep learning, which can be achieved with a high-performing cluster of GPU-accelerated server nodes and various approaches for attaching storage devices to the nodes. Optimally running such a cluster requires software to manage the servers and the storage as well as software to orchestrate the workloads. A cluster file system (CFS) is essential to ensure proper file access by multiple nodes without data corruption or loss. There are open source solutions as well as commercial solutions. Software to optimize GPU utilization is useful to extract maximum value from the investment in expensive GPUs and to allow multiple data scientists or developers to use the GPUs simultaneously and get results faster.

- **Antibias software tools.** Tools to prevent algorithmic bias have quickly emerged after several egregious cases of bias in AI algorithms against specific demographics were published. The damage that AI algorithms can inflict on unsuspecting individuals is severe, and organizations should avoid using algorithms that have not been cleared, not only to prevent individuals from being disadvantaged but also to avoid the public relations consequences when such bias is discovered and publicized. Antibias tools are widely available as open source and as commercial software.
- **Hybrid cloud deployment software.** There are advantages to on-premise AI model development (e.g., data security, performance, and cost for large-scale deployments) and AI model development in the cloud (instant scalability, opex model). While developers and LOBs gravitate toward the cloud, IT and data scientists are increasingly opting for a hybrid cloud model to get the best of both worlds. The open source software that is available for hybrid cloud is rich and plentiful, with multiple vendors providing enterprise-grade open source cloud deployment solutions.

Market Trends

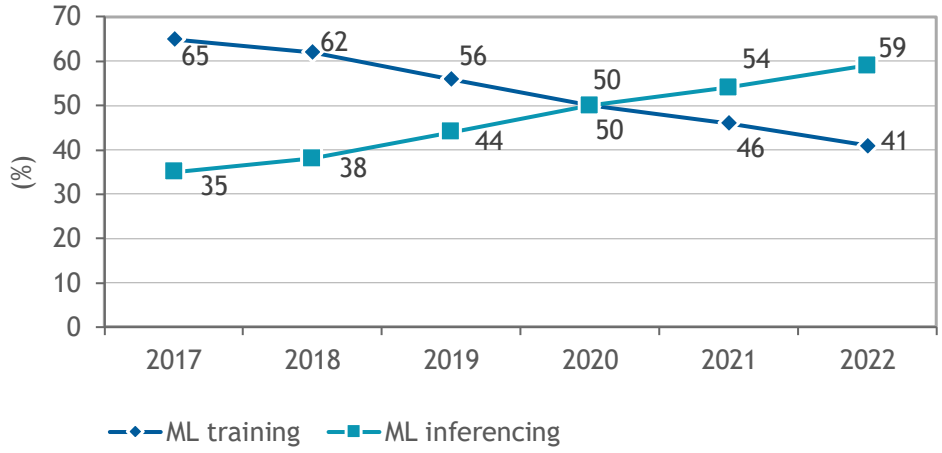
Massively Parallel Computing

The demand for computational power is driven by an insatiable appetite to reduce the time to value from data sets whose volume is increasing exponentially. Computational approaches for AI deep learning training and inferencing are evolving toward massive parallelization, leveraging thousands of cores within co-processors such as GPUs as well as leveraging from half a dozen to thousands of server nodes in a cluster, made possible by low-latency, high-bandwidth fabric to ensure coherency between systems and subsystems. IDC labels this technology massively parallel computing (MPC) and is seeing increasing use of MPC in the datacenter and the cloud for AI, big data and analytics (BDA), and simulation and modeling.

AI Training Versus AI Inferencing

In the past few years, there's been tremendous focus on machine learning training, but IDC expects that by 2020, inferencing will become equally prevalent. From there on, inferencing will become larger than training, with an expected ratio of around 41% for training and 59% for inferencing by 2022 (see Figure 2). The reasons are obvious: As more and more models are completed and go into production, the inferencing on those models starts to take off.

FIGURE 2: ***Percentage of Worldwide Server Revenue from Servers That Run ML Training Versus ML Inferencing, 2017–2022***



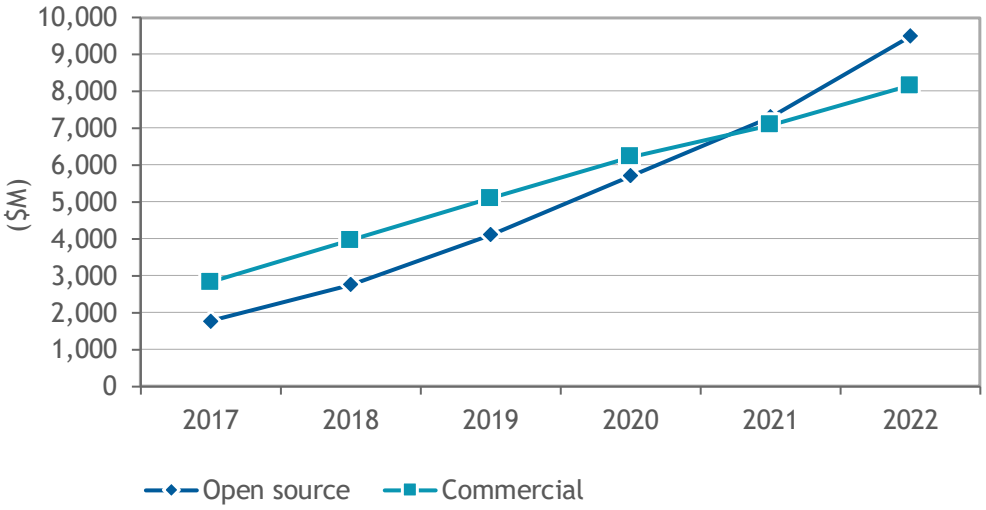
Source: IDC, 2019

Organizations that are starting to inference on trained models will find that the infrastructure requirements are somewhat less demanding in terms of the level of MPC they need. Smaller GPUs, or even beefed-up CPUs, can support less data-intense inferencing tasks. Large-scale inferencing and real-time inferencing, however, especially on images and video, are still extremely compute intensive and may require MPC infrastructure.

Commercial Versus Open Source Software for AI

As mentioned previously, open source software plays a crucial role in AI, to the point where IDC expects that, by 2021, server vendor revenue from servers that run open source AI software will exceed server vendor revenue from servers that run commercial AI software (see Figure 3). The wealth of open source software available can be daunting, and not every business will have the resources to select the right open source tools straight from the communities that develop them. Also, as mentioned, even though open source software tends to be built with security in mind, for some organizations, an extra security layer is required. They can take advantage of vendors that pre-package their server solutions with open source software that has been verified to be secure or — if vulnerabilities are found — that they have fixed.

FIGURE 3: ***Worldwide Revenue from Servers That Run Open Source Versus Commercial Software, 2017–2022***

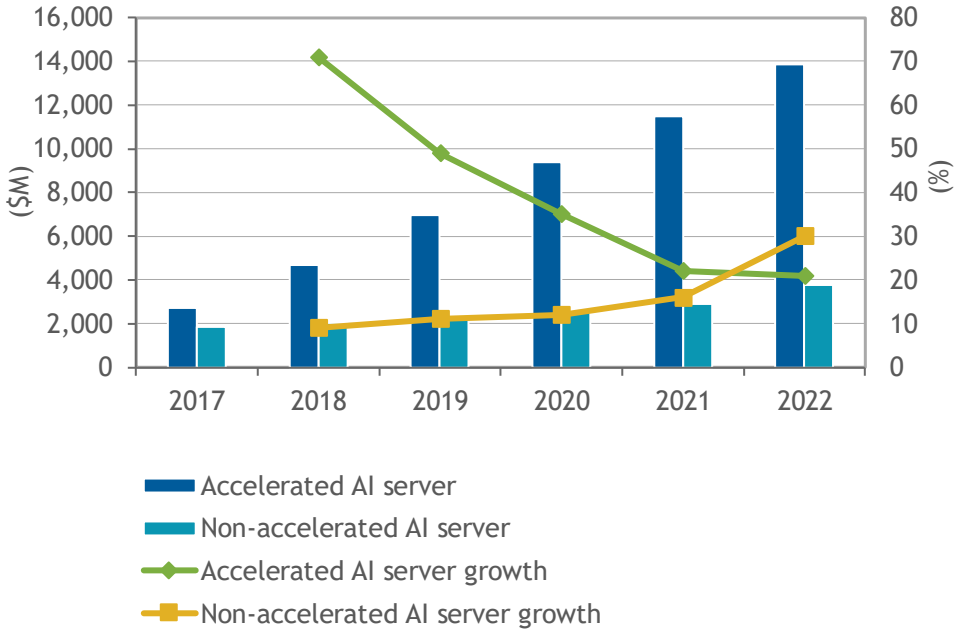


Source: IDC, 2019

Acceleration Trends

Server acceleration has become a common approach to improving workload performance. IDC expects that by 2022, accelerated servers will account for about a quarter of server revenue worldwide (see Figure 4). The end of Moore's law has been widely accepted by processor manufacturers, and acceleration, software optimization, faster interconnects, faster storage, and various other technology approaches are being introduced to continue boosting server performance for the foreseeable future. The increasing demand for acceleration with GPUs, FPGAs, ASICs, or even SmartNICs is driven by about a dozen workloads, including AI.

FIGURE 4: ***Worldwide Revenue from Servers That Run AI Platforms and Applications — Accelerated and Non-Accelerated, 2017–2022***



Source: IDC, 2019

Vendor Profile: The IBM Watson Portfolio

IBM offers an AI development approach based on the three pillars of data, train, and inference, in which hardware and software are co-optimized. At the foundation of the portfolio are the IBM Power AC922, IBM Spectrum Scale Storage, and IBM Cloud Paks. Layered on top of this this foundation are:

- **Watson Studio,** a product that came from IBM Watson (It is a collaborative, self-service platform for building AI models.)
- **Watson Machine Learning Accelerator** and **Watson Machine Learning Community Edition (CE),** formerly known as PowerAI Enterprise and PowerAI (This is a comprehensive IBM solution for model training, life-cycle management, and deployment.)
- **Watson OpenScale,** an IBM product for model monitoring and transparency

With this approach, IBM aims to offer businesses an accurate, fast, and relatively simple way to deploy machine learning and deep learning training and inferencing within their organizations.

Continuous Workflow

IBM's enterprise AI workflow solution has been designed to allow for a continuous workflow across the data, train, and inference pillars — without the need for any kind of interpolation between each stage. Organizations can bring in a model directly from Watson Studio, or they can perform data ingest, data preparation, data sanitation, and model development in Watson Studio.

Next, they can move the model into Watson ML Accelerator or Watson ML CE for training and for managing the model, running on an IBM Power AC922 node or a cluster of nodes. This is where model life-cycle management happens as well as deployment of ML/DL models built from open source or IBM tools. The idea behind Watson ML CE and Accelerator is that both data scientists and IT will want to use the software, giving scientists the tools they like to work with and IT staff the tools they need for interoperability, security, and performance.

Watson ML Accelerator provides multitenancy at scale for distributed deep learning training, experimentation, hyper parameter optimization, and elastic distributed inference. It allows data scientists to train and deploy many models, across many users, and at large scale.

The next stage is a deployment stage, which is still being developed by IBM (in terms of both software and hardware). Until this stage is finalized, deployment takes place on Watson ML Accelerator and Watson ML CE.

Watson OpenScale is focused on explainability of the model, on business KPIs, and on avoiding algorithmic bias. OpenScale monitors models at runtime to assess whether they can be trusted and to allow for full model transparency for the technical team and for IT.

Watson ML Community Edition allows users to do almost everything they can do with Watson ML Accelerator, except that Accelerator is targeting environments with four or more nodes, optimizing across a large number of IBM Power AC922s or non-Power hardware. Watson ML Accelerator comes with Spectrum Scale Storage, IBM's cluster file system that is widely used on supercomputers, including the 4,000-node Summit built by IBM and NVIDIA and currently the fastest supercomputer in the world.

Hardware

The hardware consists of the IBM Power Systems AC922, a two-socket server with POWER9 processors that has been developed for data-intensive workloads such as AI and big data analytics. POWER9 processors have four or eight threads, depending on the Power Systems model. The Power Systems AC922 comes with four threads, doubling the thread-based parallelization capabilities compared with standard architectures.

POWER9 processors have extremely high per-core performance, and the Power Systems AC922 features large memory and extra large caches. Built in are two or four NVIDIA Tesla V100 GPUs as well as NVLink 2.0, which IBM has integrated into the POWER processor to enable high-speed communications between the GPUs and the CPU, giving the GPUs full access to system memory.

Software

In terms of software, the solution features IBM's software for machine and deep learning that enterprises have been using for several years under the Watson brand. It also includes a host of open source frameworks for deep learning and various productivity tools to speed up development. Table 1 shows the various software elements that Watson ML Accelerator and Watson ML Community Edition support.

TABLE 1: ***Supported Software in Watson ML Accelerator and Watson ML CE***

	Watson Machine Learning Accelerator	**Watson Machine Learning** Community Edition
	Available for a licensing fee for more than four nodes	*Downloadable software at no charge for four or fewer nodes*
Secure wrapper around open source frameworks	√	√
Hyper parameter optimization with auto-tuner	√	√
Large Model Support	√	√
Elastic distributed training using Spectrum Scale	√	√
SnapML	√	
IBM Support	√	

Source: IDC, 2019

Elastic Distributed Training

With Watson ML (both CE and Accelerator), businesses have the ability to do elastic distributed training using Spectrum Scale, which ensures maximum utilization of the server resources. This is job scheduling software that maximizes the utilization of all the available GPUs by running multiple jobs simultaneously and dynamically allocating GPU resources to them. The immediate impact is that multiple jobs can run at the same time, and scientists are not waiting for their job to get a turn. But at the same time, costly GPUs are being fully leveraged rather than sitting idle or being only partially used — which is important for IT.

Hyper Parameter Optimization

Another software feature is hyper parameter optimization, which allows data scientists to execute training runs in parallel to automatically determine the best parameters to put on the model. Built into Watson ML Accelerator and called the auto-tuner, this tool can launch tens of thousands of parallel job iterations of a training model to determine the best parameters for that model. Instead of data scientists manually tweaking the parameters, the auto-tuner provides them with initial values that are very close to where they ultimately need to be. Data scientists will need to do only some final tweaking.

Large Model Support

Watson Machine Learning Accelerator has a built-in feature called Large Model Support, which enables tensors to be moved from GPU memory into system memory in order to free up GPU memory space for a training run. AI models are built with tensors that are placed into GPU memory. With a large model, a data scientist can quickly run out of space to place tensors into the GPU memory. They may get an "out-of-memory" error and will have to end the training run. Often, these errors do not end the training run gracefully. Subsequently, the data scientists will have to either reduce the amount of data or pare down the model so that all the layers of the network fit into GPU memory. This reduces accuracy.

With Large Model Support in Watson ML Accelerator, tensors are added to GPU memory and when memory space becomes too low, tensors that are already in GPU memory are moved to system memory over the NVLink 2.0 connection. This allows data scientists to continue iterating and adding tensors to the GPU memory and then move them over to system memory. As a result, they can use models that are significantly larger per tensor, or they can have more layers and therefore more tensors. They can bring the tensors back as needed to propagate through the model, from system memory into GPU memory.

Security

Data scientists like to work with open source frameworks, but open source frameworks trigger security concerns among IT staff. Even open source frameworks that are backed by large companies or communities require scrutiny in enterprises with large amounts of sensitive data and mission-critical applications. To give IT peace of mind, IBM has put what it calls a "secure wrapper" around these open source frameworks. IBM validates that the frameworks are secure by running tests on them and ensuring that there are no vulnerabilities.

SnapML

Also part of the Watson ML Accelerator package is SnapML, which is software from IBM Research that greatly speeds up running classical machine learning, such as Logistic Regression, on GPUs.

Watson Machine Learning Community Edition Versus Accelerator

Watson ML Community Edition is provided to organizations with a Power Systems AC922. They can download the software and start using it at no charge. The distributed deep learning features as well as resource utilization optimization are less relevant with fewer than four nodes — the features and optimization are available only for Watson ML Accelerator. Watson ML Accelerator also gives users the ability to do resource pooling, run user groups, and perform various other administrative IT tasks that are important when dealing with a large number of servers in a cluster. Watson ML Accelerator is available for a licensing fee and comes with IBM support.

Both products are available for x86-based servers, although Large Model Support cannot run on x86 because NVLink 2.0 between the GPUs and the CPU does not exist for x86.

Challenges

IBM has developed a comprehensive package of co-optimized hardware and software for AI deep learning training, with additions for deep learning inferencing on the road map. The philosophy behind this offering is sound: bring hardware and software together to achieve the greatest possible performance, usability, security, and resource utilization for developing AI, and do so in a way that satisfies both the data scientist and IT. By looking at the customer's

needs in three stages — data, train, inference — IDC has tailored the package to the unique requirements of each stage. And by making much of the software free for users of three or fewer Power Systems AC922 units, IBM has created a nice ramp for organizations to begin leveraging these tools without incurring costs. IBM has also brought important innovation to market with the integrated NVLink 2.0 into the POWER9 processor, allowing for tensors to be placed in main memory rather than in the limited GPU memory and giving scientists dramatically more memory to work with.

IDC believes that the greatest challenges that IBM faces in the market for AI deep learning hardware and software are market immaturity, market unawareness, and vendor bias. Market immaturity points to the fact that many organizations are still unclear about what AI can do for them and how they might get started with it. Identifying the right hardware and software therefore is not yet top of mind. Market unawareness refers to the problem that IBM has with explaining to the market that its offerings do not represent a hurdle for IT or the data scientist because they don't run on x86 processors. Another part of market unawareness refers to how performant yet affordable the Power Systems single- or dual-socket portfolio is — the market is still learning that IBM is no longer just producing large systems. Vendor bias is about the tendency among organizations to stay with a current vendor for new solutions, often without thoroughly investigating alternatives.

Ultimately, this means that IBM needs to continue evolving how it presents itself to customers. Typically, the company is research driven, matching user needs to innovations and assuming that the customer will grasp the importance of the innovations. In reality, the message needs to be presented to the user in clear, simple, compelling, and well-packaged ways. As IBM further develops its position in this market, IDC hopes to see IBM finding the right tone to speak to future customers.

Conclusion

AI is a specialized workload that requires unique hardware and software solutions that work extremely well together. What's more, it has become apparent that AI deep learning demands massively parallel compute and the software that optimizes that compute paradigm. For data scientists, this is an important consideration because it means that they can develop higher-quality, more accurate AI models faster. For IT organizations, the hardware and software requirements can present a major challenge if they choose to combine and optimize the many components themselves. IT will achieve better results and waste a lot less time with a fully supported co-optimized hardware-software solution for AI. IDC believes that IBM has convincingly put together such a comprehensive package for AI deep learning.

About the Analyst

Peter Rutten, *Research Director, Infrastructure Systems, Platforms, and Technologies Group*

Mr. Rutten focuses on high-end, accelerated, and heterogeneous infrastructure and use cases, which include supercomputing, massively parallel computing, artificial intelligence (AI) and analytics, and in-memory computing. He also covers compute for various workloads such as SAP HANA.

MESSAGE FROM THE SPONSOR

IBM Power Systems provides industry-leading enterprise AI infrastructure for machine learning, deep learning and inference to fuel new thinking and capabilities across organizations. Drive greater confidence in business decisions at scale with a solution designed to grow with organizations and make the best use of people, processes and processors. Find meaningful results faster with the industry's highest data throughput and IBM research keeping teams on the cutting edge of AI technology. All of this is provided on top of the proven security of Power and using open-source frameworks secured by IBM.

At each stage of the AI process, from data to training to inference, IBM has a solution. Learn more about the enterprise AI workflow at ***ibm.com***.

IDC Custom Solutions

The content in this paper was adapted from existing IDC research published on www.idc.com.

IDC Corporate USA
5 Speen Street
Framingham, MA 01701, USA
T 508.872.8200
F 508.935.4015
Twitter @IDC
idc-insights-community.com
www.idc.com