Why Your Current IT infrastructure Won't Work for AI

Sponsored by IBM

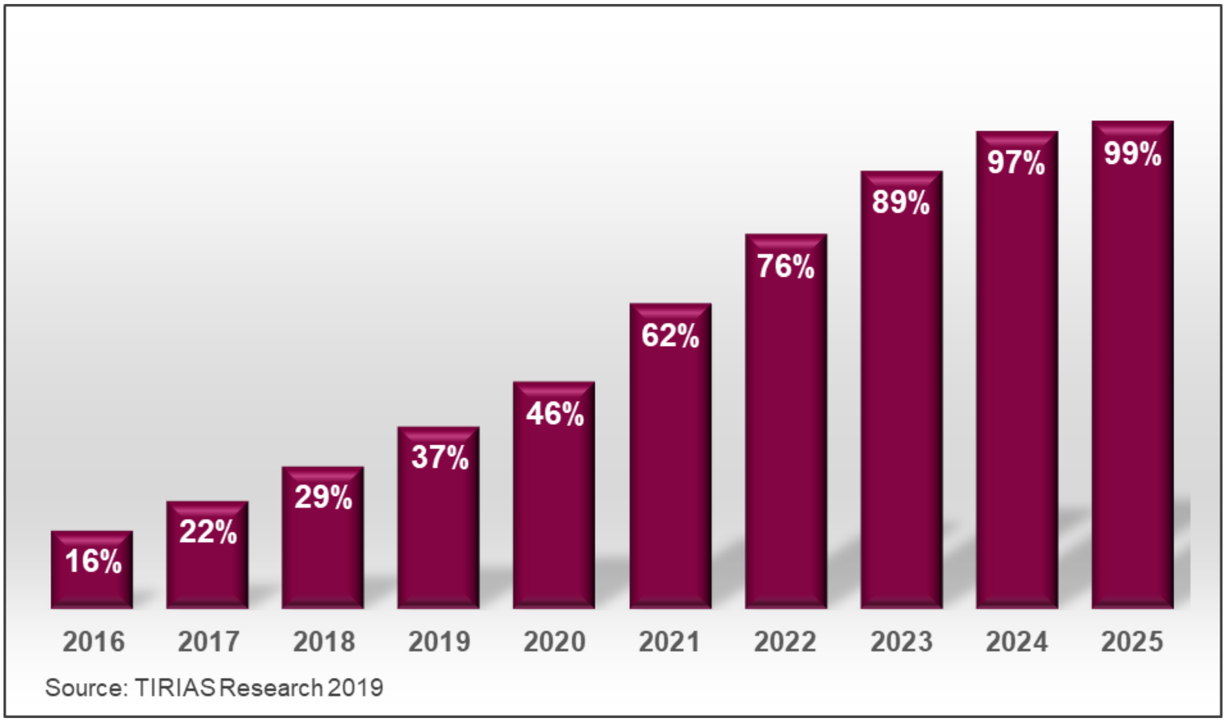
This is the first of two papers focused on how AI is changing on-prem and cloud data centers requirements. This paper focuses on the different requirements for AI processing compared with more traditional data center workloads. The second paper focuses on the different between AI training and inference processing.

Introduction

The concept of deep learning, the process of using layers of algorithms to process data, was first introduced in the 1940s, but recent advances in hardware technology and software frameworks have enabled a revolution in deep learning and artificial intelligence. The introduction of CUDA, a parallel computing platform introduced by NVIDIA in 2006, combined with the massive computing capabilities of modern GPUs, marked the beginning of artificial neural networks capable of processing massive amounts of data with high speeds and accuracy.

During the same period, the technology industry has shifted to a world of connected devices referred to as the Internet of Things or IoT, which combined with modern sensor technology, is increasing the amount of data generated exponentially. Artificial neural networks are using this data to create intelligent applications in almost every industry. Most enterprises are now using or evaluating how they can use AI to increase quality, improve operational efficiency, and/or solve design and engineering challenges. In addition, TIRIAS Research estimates that by the end of 2025, 99% of all new devices will leverage some form of AI.

Figure 1. Forecast for New Platforms Using Artificial Intelligence/Machine Learning



Source: TIRIAS Research 2019

The Challenge of AI

AI is typically broken into two types of workloads or processes–training and inference. Deep learning is the development, training, and optimization of the neural network based on the subject data and desired outcome (example: training a neural net using pictures of various dogs). Inference is the use of a trained neural network model (example: identify that an animal is a dog).

Training a neural network requires both large data sets to ensure a high degree of accuracy and many compute engines to process the data quickly and efficiently. This requires a computer with high-bandwidth I/O, large amounts of memory and storage, and highly parallel processing engines working in parallel on a single workload. Similarly, inference processing requires high throughput, which also stresses the throughput and compute performance of the data center. Inference processing also tends to be more time sensitive and may require being located close to the point of use. Unfortunately, most enterprise and cloud data centers are not architected for AI requirements.

Data centers, both enterprise and cloud services, generally fall into two camps – applications specific or IT (information technology) services. The application specific data centers are typically dedicated to applications like web hosting, video services, etc. The IT data centers are designed to handle a broad range of applications and workloads. In both cases, the data center is designed to process tasks from many users or application instances. Handling many users or workloads drives everything from the network topology, storage requirements, I/O, and operating environments. These data centers also leverage virtualization and containers for further abstraction, flexibility, and operational efficiencies.

AI workloads share some similarities to other data center workloads, such as the high I/O and throughput bandwidth of communications, the need for large amounts of storage of web hosting, and the high memory bandwidth and compute performance of HPC (High-Performance Computing). However, the processing requirements for AI are unique. Rather than being tied to one or several processing engines, AI training requires hundreds or thousands of matrix multiplication engines operating in parallel on the same data set. As a result, AI workloads run most efficiently close to the metal (no virtualization or abstraction).

Modern GPUs are well suited for both AI training and inference tasks. The NVIDIA Volta V100 GPU has 5120 32-bit floating point CUDA cores (or 2560 operating at 64-bit double-precision floating point) for training large data sets and 640 Tensor cores for accelerating inference processing.

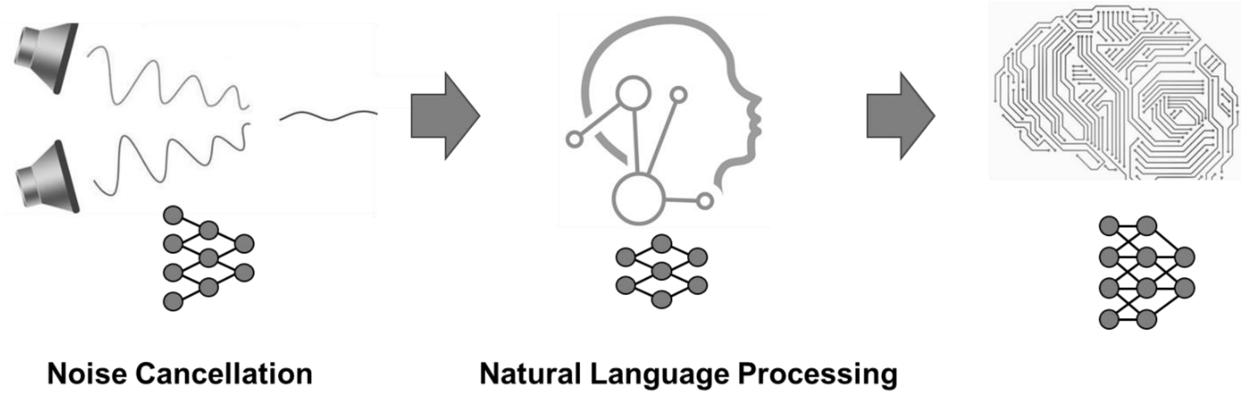
Reasons AI Requires Dedicated Resources

AI is a Unique Function

In addition to the processing requirements, AI functions are neither static nor singular. AI models change over time with the addition of more data and new data. As a result, AI models go through training and retraining on a continuous basis, which improves accuracy by further optimizing the neural net model.

AI workloads are also not singular. A single AI tasks, such as object or voice recognition, is often broken down into multiple AI tasks. As an example, just listening for key control words may require noise cancellation, natural language processing, and the control function itself, each of which can benefit from neural network models, but each of those tasks require a different type of neural net model.

Figure 2. Potential AI Workloads for Key Word Recognition



Source: TIRIAS Research

Unique Performance Requirements

Training a neural network is a repetitive task with the performance and accuracy predicated on the size and quality of the data set. However, the efficiency of performing this task can have a significant result on the cost of training as well as the value to an organization. While a neural network can be trained on any form of processing engine, time and costs are critical. Just shifting a workload from CPUs to accelerators like GPUs can reduce the training requirements by an order of magnitude or more, such as months to days, hours, or in some cases even minutes. This reduction in time can result in thousands of dollars or more of savings in operational costs, as well as accelerating the deployment of the new model.

AI inference processing often requires real-time (<10ms) or near real-time (10ms to100ms). As a result, execution latency is critical to the quality-of-service (QoS) for many applications, such as cloud applications like search engines, digital assistants, and language translation, and on-prem applications like security monitoring, fraud detection, and manufacturing quality control.

As a result, architecting systems and data centers to reduce or eliminate latency and performance bottlenecks requires optimizing:

> Storage & Memory – While the data center storage for AI must be enough to support the workloads, the memory becomes the critical factor for operational efficiency. To ensure efficient execution, the model should not be waiting of data calls from storage, everything should be executed from high-speed memory. As a result, the memory must be sized according to the potential neural network models and memory bandwidth must be high to minimize latency.
>
> Network topology – With computes nodes organized in layers according to the algorithms/neural network model, execution between cores, chips, and even systems must be both configurable and high-speed to increase workload throughput.
>
> Processing engines – As mentioned earlier, AI computing requires dedicated accelerators for both training and inference capable of parallel processing large data sets. Training requires complex matrix multiplication of large data sets, while interference requires fast processing of a limited data sets.
>
> I/O – The QoS requirements of the platform may also require very high-speed I/O to reduce the latency time from when the data is received to when a determinization or decision is returned. AI is pushing the limits of I/O bandwidth and often requiring the latest in I/O technology of 40Gb/s or higher.
>
> Software Framework – The software framework also plays a key role in the overall performance of an AI platform. Optimizing the resulting neural network model is crucial to ensuring the highest performance, lowest latency, and efficient use of resources.

AI Stresses Traditional Data Center Resources

Power and Thermal Requirements

Systems designed for AI workloads are optimized for performance. The average NVIDIA Volta V100 has a Thermal Design Power (TDP) of 300W. When you include several of these cards, along with dual high-performance CPUs, several terabytes of memory, and a high-speed network, a single system can consume 5kW of power, or more. This places high demands on both the power for the systems, and system thermal management and data center environmental systems to maintain operational temperatures. Traditional data centers can balance workloads across different systems for power leveling and thermal management. To execute AI workloads in an efficient manner, the resources need to be configured close together to limit the time and distance between compute nodes and the system must operate at or near maximum performance. This requires special consideration in the data center design to handle these power and thermal requirements.

I/O

Traditional data centers also typically handle data in batches. With AI applications, especially inference, leveraging a constant stream of data from a wide array of sources ranging from simple sensors to complex autonomous systems, I/O demands are high and the density of the data may vary greatly. Once again, this requires designing the communications system handling data to and from the data center and the network within the data center around the challenging requirements of current and future AI workloads.

Data Security

AI applications may use data from a wide variety of sensors and resources, as a result, they may not only be subject to IT security standards, but also that of specific industries requirements such as HIPAA (Health Insurance Portability and Accountability Act) in the U.S. for medical information, PCI-SSC (Payment Card Industry Security Standards Council) for credit card information, and ISO 27000 for financial information. AI systems must also abide by new privacy regulations, such as the European GDPR (General Data Protection Regulation).

In addition, neural networks may be making decisions that could impact human life, especially in the case of transportation or healthcare applications. As a result, AI data centers require the highest levels of security both for the data being used and the results from the data.

Adaptability

The most challenging aspect of AI is planning for the fast pace of innovation in the field. On the software side, new libraries, frameworks, and neural network models are being developed very rapidly. This is pushing the performance requirements of processors, accelerators, memory, networks, and I/O technology higher. As a result, AI data centers need to adapt to the latest technologies without knowing the future specifications.

Even when a neural network is trained and being deployed, it can change with the data, which may result in changes to the overall system requirements, such as more I/O bandwidth for inference processing, new nodes or layers to the neural network model for higher accuracy or to accommodate additional data, or new AI tasks/neural networks for additional actions. The system demands of AI are increasing and will continue to for the foreseeable future.

Most traditional data centers are designed around a specific set of specifications, especially power and thermal management, not in anticipation of where the technology may be in one, three, or five years. Upgrading a traditional data center is typically done in cycles that are planned and budgeted for long in advance. An AI data center must be more flexible in upgrading to the latest technology.

Resources

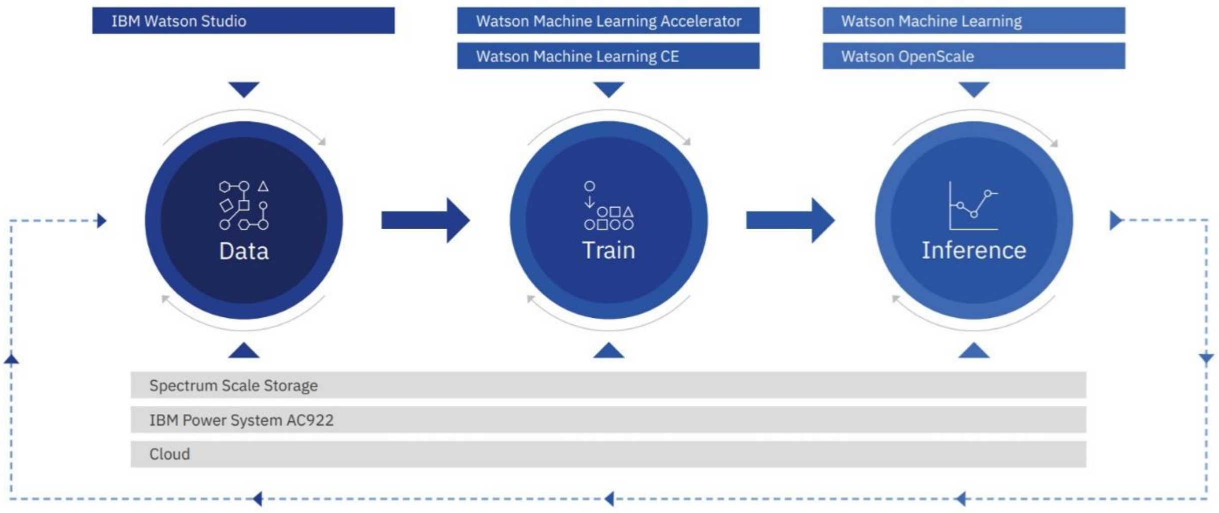
Another issue is intellectual resources. There are limited numbers of data scientists and analytics experts that can effectively develop efficient AI solutions. These resources are better allocated to developing and working with dedicated AI data centers than general resources or traditional data centers.

Watson Machine Learning Accelerator

IBM has taken all these requirements into consideration in the development of a scalable AI environment that enables data scientists to develop and deploy enterprise AI solutions using co-optimized AI hardware and software. The Watson Machine Learning Accelerator combines popular open source deep learning frameworks, efficient AI development tools, and accelerated IBM Power Systems servers. The platform combines accelerated IBM Power Systems with enterprise software containing open-source frameworks, software libraries and tools.

Watson Machine Learning Accelerator is designed to provide lifecycle management from installation and configuration; data ingest and preparation; building, optimizing, and distributing the training model; to moving the model into production.

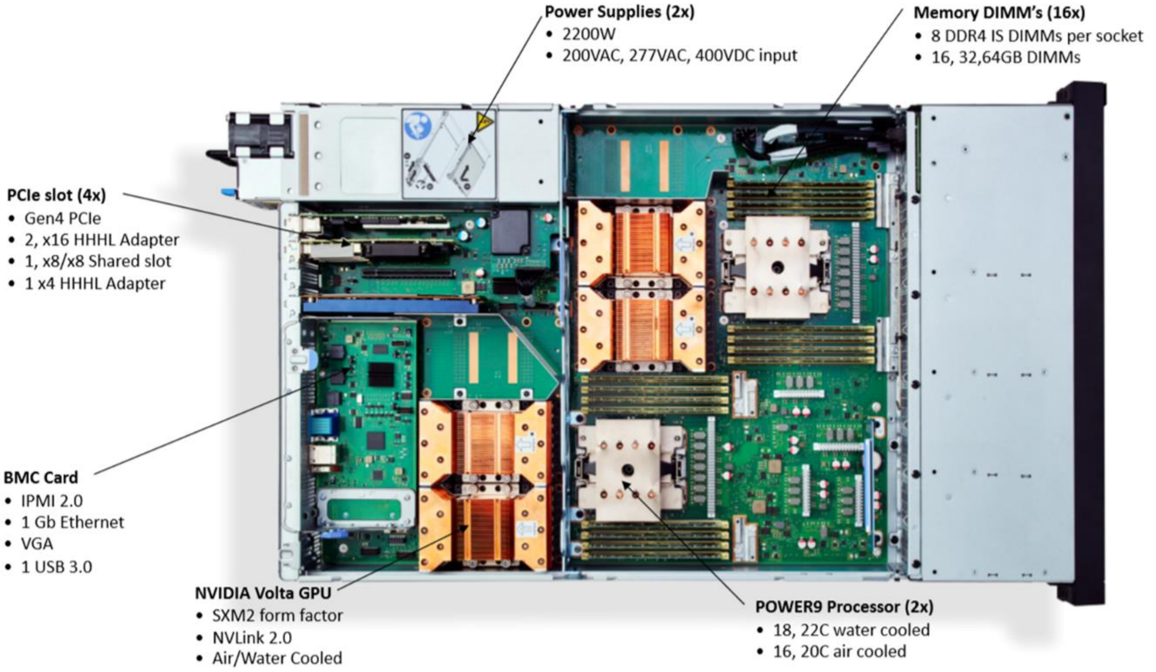
Figure 3. IBM Watson Machine Learning Accelerator



Source: IBM

At the core of the environment is the IBM Power System AC922. Purpose built for training, the Power AC922 uses the latest 20-core POWER9 processor in conjunction with up to six NVIDIA Tesla V100 GPUs. The POWER9 provides the system management with a high-speed NVIDIA NVLink to each GPUs, the only platform to leverage the capabilities of NVLink between the host processors and GPU accelerators. The NVIDIA V100 GPUs provide a combination of efficient highly parallel CUDA and Tensor cores for neural network training and inference processing. In addition, CPUs and GPUs share 16 DDR4 memory lanes for up to 32TB. The system also leverages the latest PCIe Gen4 interconnect for system I/O. When combined with the SnapML machine learning library, the platform provides an unrivaled level of performance. (https://www.ibm.com/blogs/systems/power-snapml-watson-machine-learning/).

Figure 4. IBM Power System AC922



Source: IBM

Conclusion

One could argue that any IT workload provides some form or monetary or efficiency value for an organization. However, AI is unique in that processing data can often find unknown or hidden efficiencies for an organization and/or produce information that could enhance the value or the organization's products or services or be a product or service within itself. Most organizations can benefit from the use of AI and should consider either developing their own on-prem AI platform or leveraging such resources from a third-party cloud service provider like IBM.

AI will change everything from how we use technology to the technology itself. However, it has different requirements than traditional IT processing. Key requirements are:

- Massive parallel processing through dedicated accelerators
- High-memory and I/O bandwidth
- Challenging power and thermal requirements
- System flexibility for training and retraining
- A highly secure environment

As a result, AI requires dedicated resources that interact with existing IT resources.

IBM is unique in that it provides everything from the business services to assist in identifying where and how to use AI to supplying the on-premises and cloud resources. IBM is also a leader in next generation technology like Quantum Computing that may provide further enhancements to AI computing in the future.

The information contained in this report was believed to be reliable when written but is not guaranteed as to its accuracy or completeness.

The contents of this report represent the interpretation and analysis of statistics and information that is either generally available to the public or released by responsible agencies or individuals.