White Paper

Ready to Scale AI? Don't Suffer from Core Starvation

Sponsored by: IBM

Peter Rutten
May 2019

IDC OPINION

The business opportunities that can be achieved with AI are exceptionally promising. Businesses and other organizations know that not acting on AI could potentially be a business disaster as competitors gain a wealth of previously unavailable data and capabilities to grow and delight their customer base. Few, if any, businesses today believe that "AI is not for us" or that "AI is mostly hype." Rather, serious AI initiatives are being undertaken worldwide, across industries, and across company sizes.

Many organizations' lines of business (LOBs), IT staff, data scientists, and developers have been working to learn about AI, understand the use cases, define an AI strategy for their business, launch initial AI initiatives, and develop and test the resulting AI applications that deliver new insights and capabilities using machine learning (ML) algorithms, especially deep learning (DL).

Organizations are now ready to scale these initiatives and new questions that emerge. They know — indeed, they may have first-hand experience — that they cannot use standard, multipurpose infrastructure. Also, they have established that AI training (the training of the AI model) and AI inferencing (the use of the trained model to understand or predict an event) require different types of compute. But what is that different compute? Also, should they deploy on-premise, in the cloud, or a hybrid cloud model?

AI applications and especially deep learning systems, which parse exponentially greater amounts of data, are extremely demanding and require powerful parallel processing capabilities based on large numbers of cores, and standard CPUs cannot sufficiently execute these AI tasks. IDC research shows that, in terms of core capacity, a large gap between actual and required CPU capability will develop in the next several years. To overcome this gap, AI users that have experimented with existing infrastructure and that are ready to scale need to overhaul their infrastructure to obtain the required parallel processing performance, which is achieved with multithreaded CPUs combined with GPUs, fast interconnects, large amounts of memory, and advanced I/O capabilities.

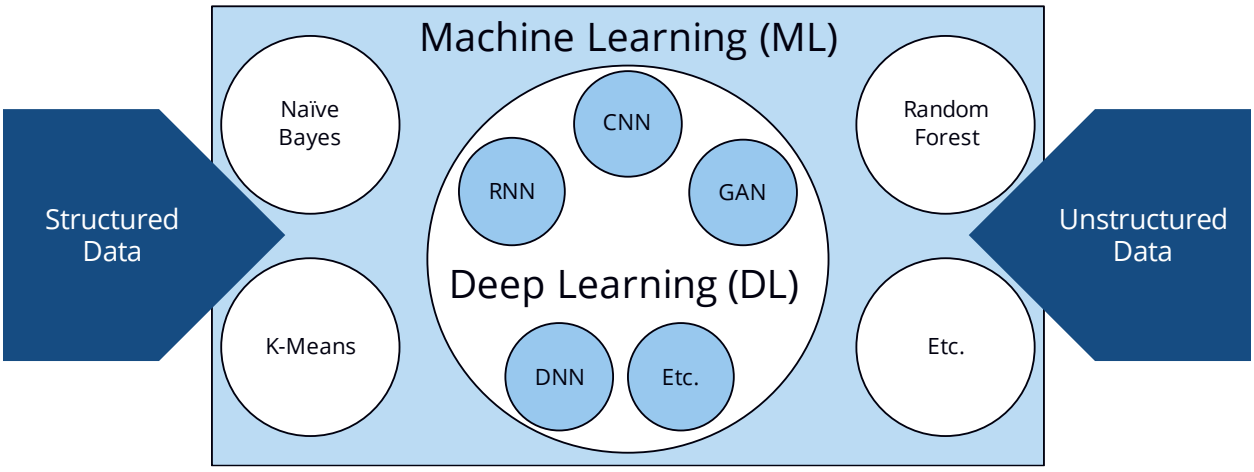
SITUATION OVERVIEW

Businesses around the world are responding vigorously to the new opportunities offered by AI workloads. IDC defines AI as a set of technologies that use natural language processing (NLP), image/video analytics, machine learning, knowledge graphs, and other technologies to answer questions, discover insights, and provide recommendations. These systems hypothesize and formulate possible answers based on available evidence, can be trained through the ingestion of vast amounts of content, and adapt and learn from their mistakes and failures through retraining or human supervision.

Machine learning is a subset of AI techniques that enable computer systems to learn and improve their behavior for a given task without having to be programmed by a human. Machine learning models are algorithms that can improve over time by testing themselves over and over again using large amounts of structured and/or unstructured data until they are deemed to have "learned" a task (e.g., recognizing a human face). Figure 1 illustrates how deep learning is a subset of ML. Typical DL architectures are deep neural networks (DNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), generative adversarial networks (GAN), and many more.

FIGURE 1

Machine Learning and Deep Learning



Source: IDC, 2019

AI software platforms include:

- Conversational AI software (e.g., digital assistants)
- Predictive analytics to discover hidden relationships in data and make predictions
- Text analytics and natural language for recognizing, understanding, and extracting value from text
- Voice/speech analytics for recognizing, identifying, and extracting information from audio, voice, and speech
- Image and video analytics for recognizing, identifying, and extracting information from images and video, including pattern recognition, objects, colors, and other attributes such as people, faces, emotion, cars, and scenery

Many businesses are well on their way with AI initiatives and have reached a stage where they are ready to start deploying AI at production scale. Others are still experimenting with AI, while a third group is currently at the stage of evaluating what AI applications can mean for its organization.

With regard to the first group (ready to deploy), IDC is seeing a range of AI use cases that businesses, governments, and other organizations have begun to implement. The five most common use cases today are (ranked by the amount that businesses spend on them in terms of hardware, software, and services):

- **Automated customer service agents.** In the banking industry, for example, these AI applications provide customer service via a learning program that understands customer needs and problems and helps a bank reduce the time and resources needed for resolving customer issues. These agents are becoming widely used across industries.
- **Sales process recommendation and automation.** Used in in various industries, these are AI applications that work with customer relationship management (CRM) systems to understand customer context in real time and recommend relevant actions to sales agents
- **Automated threat intelligence and prevention systems.** Becoming a critical part of threat prevention across governments and industries, these AI applications process intelligence reports, extract information from them, establish relationships between diverse pieces of information, and then identify threats to databases, systems, websites, and so forth.
- **Fraud analysis and investigation.** In the insurance industry, but used widely elsewhere as well, these AI applications use rule-based learning to identify fraudulent transactions, and they automatically learn to identify various insurance-related fraud schemes.
- **Automated preventative maintenance.** In the manufacturing industries, these AI applications are based on machine learning algorithms that build an accurate predictive model of potential plant and machinery failures, reducing downtime and maintenance cost.

Additional AI use cases that have gained traction in enterprises are (ranked in order of spending on hardware, software, and services):

- Program advisors and recommendation systems
- Diagnosis and treatment systems
- Intelligent processing automation
- Quality management investigation and recommendation systems
- IT automation
- Digital assistants for enterprise knowledge workers
- Expert shopping advisors and product recommendations
- Supply and logistics
- Regulatory intelligence
- Asset/fleet management
- Automated claims processing
- Digital twin/advanced digital simulation
- Public safety and emergency response
- Adaptive learning
- Smart networking
- Freight management
- Pharmaceutical research and discovery

Cloud Versus On-Premise

The applications that address these use cases may be custom developed by an organization, may be based on commercial AI software, or may be delivered as AI SaaS. Deployment considerations for the custom developed and commercial software are on-premise, in the cloud on IaaS, or as a hybrid cloud, wherein the on-premise environment interacts with a public cloud environment.

For the various deployment scenarios, solutions must be considered for:

- Securely processing the volume of data that is required for training AI models with extremely high performance. The performance requirements for deep learning training involve the ability to execute massively parallel processing using GPUs combined with high-bandwidth data ingestion.
- Securely processing the volume of data that the AI model will perform inferencing on with extremely high performance. Performance with regard to inferencing means the ability to process incoming data through the trained AI model and deliver near real-time AI insights or decisions.

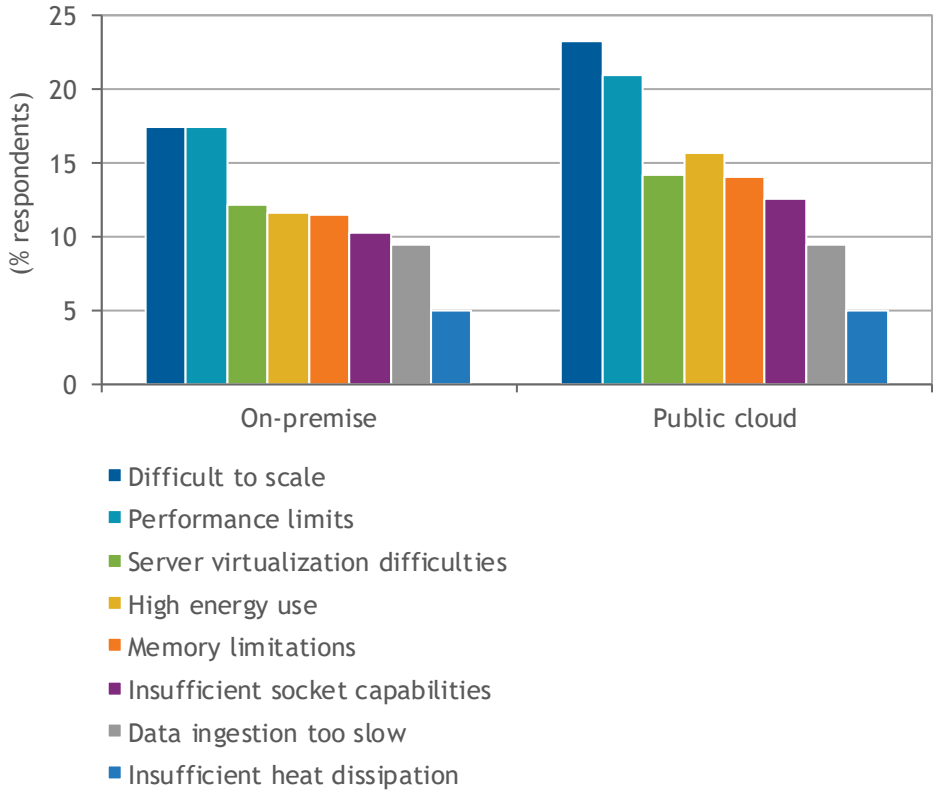
For data scientists and developers, it can sometimes be easier to start an AI initiative in the cloud, saving them from having to arrange on-premise compute that, for deep learning, typically needs to be accelerated. Accelerated AI cloud instances are available on most public clouds, usually with open source AI stacks. Of course, with accelerated cloud instances for AI training, the cloud SP dictates what's available to the end user in terms of processors, coprocessors, interconnects, memory sizes, I/O bandwidth, and so forth. Not all cloud SPs offer the most optimized combinations of these components, which ultimately determine the speed and quality with which data scientists can develop training models. As a result, many organizations opt for on-premise deployment.

During their AI experiments in the past few years, many organizations found themselves "hitting the wall" with their standard infrastructure or with the basic cloud instances. Training models took too long, and inferencing was too slow. IDC research shows that 77.1% of respondents say they ran into one or more limitations with their AI infrastructure on-premise and 90.3% ran into compute limitations in the cloud.

Figure 2 shows the kinds of hardware limitations that are being encountered most often, on-premise and in the cloud. Most organizations experience a combination of these hurdles. The responses have been ranked by most often cited on-premise hurdles.

FIGURE 2

Top Hardware Limitations Encountered with Server Infrastructure for Running AI Use Cases



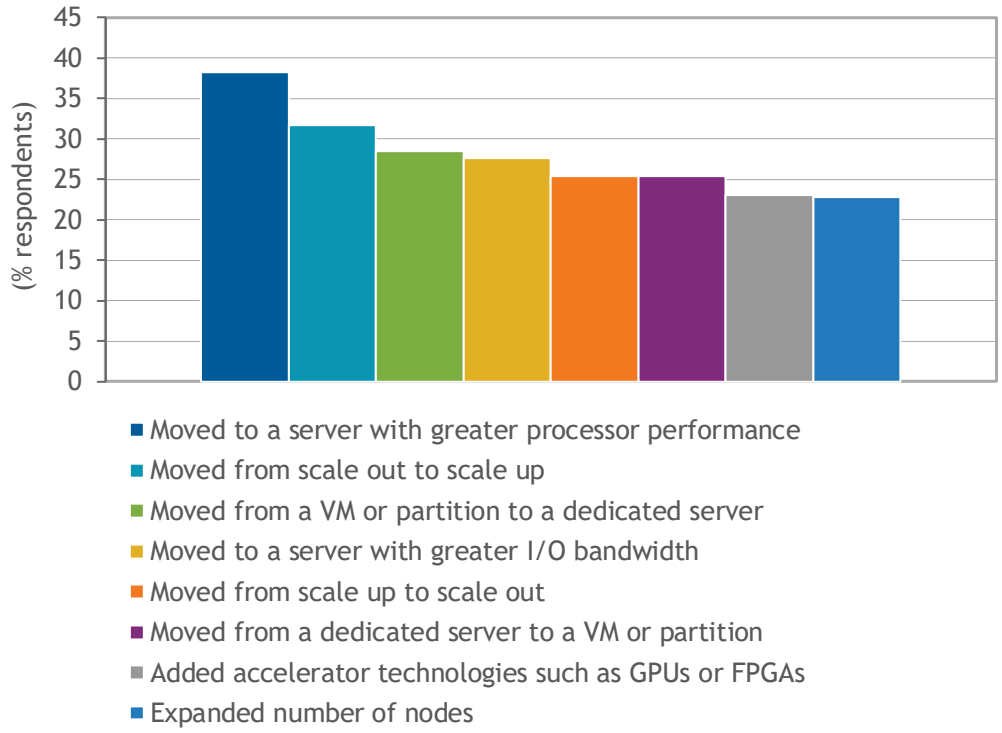
Source: IDC 2019

As a result of the hardware-related limitations, IDC has seen businesses completely overhaul their AI infrastructure, sometimes twice in just a few years. Figure 3 shows the nature of these infrastructure shifts and also illustrates how contradictory some of them sometimes are in the sense that some organizations went in the exact opposite direction that others take.

Figure 3 also shows that the most often performed infrastructure change to improve AI performance is the move to greater processor performance, as depicted in the far-left column. The fourth most common change is improving the I/O performance to speed up the data ingestion for AI. Adding accelerators has become more common, as is expanding the number of nodes. Note that the shifts depicted in the chart are not mutually exclusive – for example, some respondents shifted their infrastructure to both scale-out and acceleration. There are also a few diametric infrastructure overhauls (scale up to scale out as well as scale out to scale up), which is indicative of some of the experimentation that has been going on in the first years of AI's entry into the business world.

FIGURE 3

Nature of Generational Shifts with AI Infrastructure



Source: IDC, 2019

Core Starvation

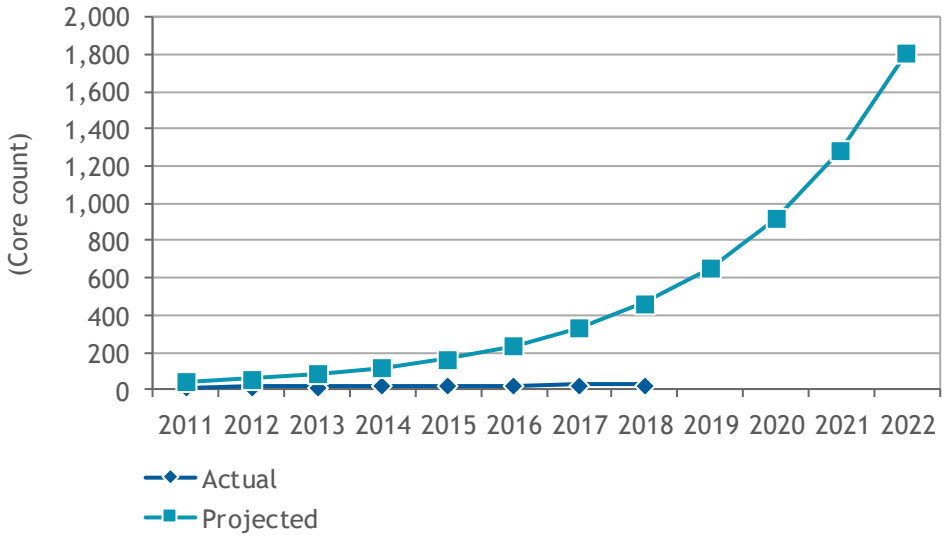
The fundamental reason for the limitations that businesses encounter is what IDC has termed "core starvation." AI is based on sophisticated mathematical and statistical computations. Take, for example, image and video analytics. Images are converted to matrices, with each pixel being represented by a number. Millions of matrices plus their classifications are fed into a neural network for correlation. The matrices are then multiplied with each other to find the right result (e.g., a "dog" or a "soda can").

To speed this process up, it must be done in parallel on many more cores than CPUs can provide. CPUs are designed for serial processing, and they are close to reaching their maximum potential due to the limitations of material physics. Today, all processor manufacturers have acknowledged that we have reached the end of Moore's law and that supplementary approaches outside of CPU improvements are needed to continue on the path of performance increases that had been the norm for decades. The primary reason is that CPUs have a limited number of cores (dozens rather than thousands) due to the size and cost of these cores.

Figure 4 shows historical CPU capability growth for 2011-2018 as well as a logarithmically modeled CPU capability curve for 2011-2022 as if CPU capability were unrestrained by physical limitations. The modeled CPU capability assumes (counter to current reality) that there is no physical limit to CPU capability, thus providing insight into how the need for CPU capability will develop. Figure 4 also illustrates the resulting gap between CPU capability need and actual CPU capability available.

FIGURE 4

Actual and Projected Worldwide Core Count Requirement for All Workloads



Source: IDC, 2019

Hence the rise of GPUs (with thousands of cores) and custom designed processors (ASIC, FPGAs) fills the gap between actual and projected CPU capability need. These accelerators have massively parallel architectures with hundreds or even thousands of cores on a die that affordably deliver the parallel compute performance needed. The impact of these coprocessors has been a distinct performance boost combined with other benefits based on the type of coprocessor. At the same time, technologies to feed these coprocessors' enormous volumes of data gained in importance, such as interconnects between coprocessors and between CPUs and coprocessors; increased memory sizes; and fast storage.

IDC is seeing the worldwide market for accelerated servers grow to $25.6 billion in 2022, with a 31.6% CAGR. These are servers that are accelerated by either GPUs, FPGAs, or ASICs, both on-premise and in the cloud. Indeed, this market is growing so fast that IDC is forecasting that accelerated compute will start eating away at nonaccelerated compute in the market to the point that by 2021, 12% of worldwide server value is from accelerated compute.

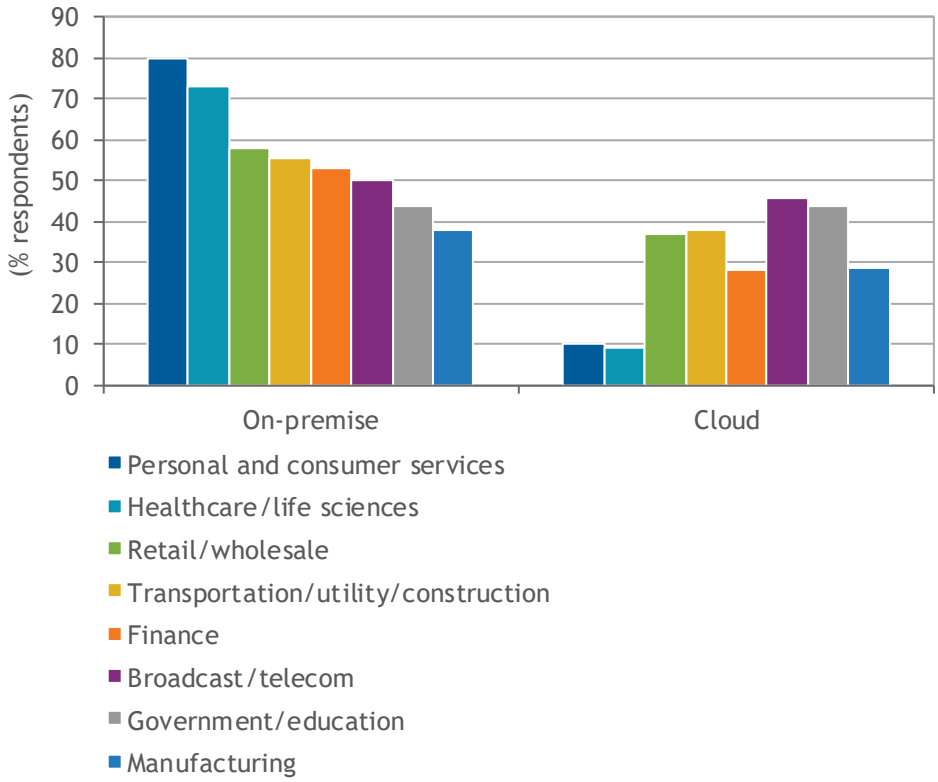
Yet acceleration is a relatively young technology. IDC research has found that on average, organizations that leverage accelerators to boost the performance of their infrastructure have been doing so for about two to three years. IDC expects this to become a permanent aspect of computing, with acceleration for multiple workloads becoming a standard feature until entirely new types of computing have become mainstream, such as quantum computing. Even then, accelerated classical computing will continue to be the norm for most workloads.

Apart from such workloads as networking, encryption, security, and compression, which have the highest level of acceleration, the workloads that are increasing their penetration of coprocessors for acceleration are real-time analytics; AI deep learning; Hadoop and databases; video, image, and voice recognition; simulation and modeling; and AI inferencing.

To date, on-premise has been the preferred deployment for accelerated servers. Figure 5 shows the on-premise versus cloud deployment of GPU-accelerated servers for select industries. Figure 5 also shows that the greatest on-premise GPU use happens in the personal and consumer services industry (80% of respondents), whereas the greatest cloud use occurs in the broadcast/telecom industry (45.5%).

FIGURE 5

On-Premise Versus Cloud Deployment of GPU-Accelerated Servers by Industry



Source: IDC, 2019

Interestingly, on-premise deployments will increase in the next 12 months, and cloud SPs should be aware of the challenge this represents. For example, 50.8% businesses that use GPUs for acceleration say they do so on-premise today, while this percentage goes up to 54.2 in the next 12 months. Similarly, on-premise use of ASICs and FPGAs will increase by about 5 percentage points.

Even more alarming for cloud SPs should be the fact that quite a few businesses are moving certain accelerated workloads that they were running in the public cloud back to on-premise (also referred to as "repatriation"). The reasons these organizations cite for bringing workloads back on-premise are issues with cost, security, scaling, and performance that they experienced in the cloud. 66% of organizations say that they have, in the past, started a workload that requires accelerators in the cloud but have since moved that workload to their on-premise datacenter.

The costs versus performance gains from acceleration vary by industry; in some industries performance is – on average – doubled, while the on-premise capex (cost of acquiring the accelerated server) or

cloud opex (cost of running an accelerated instance) increases vary between 26–33% depending on industry. Much depends on the actual infrastructure — with some accelerated servers, performance gains can be higher. Regardless, organizations should consider this to be a strong ROI, even as AI will — over time — start to demand more than a doubling of performance — hence the exceptionally large number of new processor and coprocessor research that is in progress at hardware start-ups as well as the likes of IBM, Intel, AMD, Xilinx, and NVIDIA. IDC research has also found that most organizations prefer to have the server vendor install the acceleration technology, fewer use a systems integrator or VAR, and even fewer have their IT team install the accelerators.

Programming an accelerator, for example, using CUDA, OpenACC, or OpenMP for GPUs, requires skilled staff, and organizations said they needed on average between 2.4 and 3.7 FTE staff for this purpose, depending on their industry (regardless of accelerator type). GPUs tend to be easiest to program, FPGAs somewhat harder, and ASICs have a long development phase. Most businesses say they are fairly to very satisfied with the speed with which accelerators can be programmed. In other words, this should not be seen as an impediment.

Ultimately, businesses say that gaining more performance is the most important goal for choosing an accelerated server. Performance translates directly into how long a data scientist has to wait for an AI model to be trained, how deep and accurate that AI model is, and how fast inferencing can be performed on the trained AI model. In other words, greater performance means faster and more accurate results. Cost is secondary while the need for new skill sets is deemed much less important. By all accounts, businesses want AI horsepower and are willing to pay for it while worrying less about the cost or the need to acquire new skills.

Inferencing

There is some debate as to the differences in requirements for training an AI model versus inferencing on that model. IDC research shows that, over time, server infrastructure used for inferencing will grow faster than server infrastructure for training. This includes infrastructure for all types of machine learning, including "classical" machine learning (e.g., nearest neighbors or Naïve Bayes classifier), as well as deep learning (e.g., for voice/speech or image and video analytics). In terms of worldwide server value, inferencing will exceed training by 2020. Inferencing too will require higher-performing compute and, in many cases, accelerated compute.

There are various factors that make AI inferencing a different workload than AI training. Inferencing applies incoming data to the trained AI model. In most applications, this needs to happen with minimal latency so as to deliver a near-real-time result. The data may be small (e.g., a one-time image recognition task) or huge (e.g., a constant, real-time facial recognition task on a public street). It is generally agreed that the overall server infrastructure requirements for inferencing may differ from those for training predominantly in the required type of acceleration (e.g., less powerful GPUs) or for very large-scale deployments that perform highly repetitive inferencing tasks (e.g., FPGAs or ASICs).

In some cases, where inferencing is light enough to be executed on powerful host processors, acceleration using coprocessors may not be required. This may be the case for AI-enabled applications (versus fully AI applications). In AI-enabled applications, only a small portion of the application performs an AI function — for example, a procurement function in business software that uses a conversational interface. With powerful CPUs, such applications may not need to offload the AI processing onto a coprocessor such as a GPU for acceleration. AI inferencing can benefit from clustering as well but can also be executed on multisocket scale-up platforms.

AI Infrastructure Considerations

In the past two to three years, the right infrastructure for AI applications was subject to much experimentation. Businesses tried everything, from hyperconverged to scale up to scale out. Since then, as AI implementations have matured and are starting to scale, IDC has seen significant consensus being reached around the concept of AI benefitting first and foremost from clustering. Because of the parallel nature of AI processing, the ability to leverage hundreds of cores in an accelerator, multiple accelerators in a server node, and multiple servers in a server cluster provides a performance advantage.

Businesses that are ready to start scaling their AI infrastructure typically go through a list of considerations that can be fulfilled with an on-premise deployment as well as in the cloud, albeit to different degrees. They take time, therefore, to determine what matters most, before deciding whether they want to scale their accelerated AI infrastructure on-premise, in the cloud, or in a hybrid cloud.

For cloud SPs that want to provide their customers with the best possible instances for AI, these considerations are important, especially as we are seeing increases in on-premise deployment of applications that require acceleration. The acceleration-as-a-service offerings from cloud SPs need to therefore be on par with the best possible on-premise offerings if they wish to be seen as an AI cloud provider of choice.

Table 1 provides an overview of multiple hardware, software, and datacenter considerations for accelerated system. They can be viewed in light of their importance (which has been based on an IDC survey of accelerated systems users) and with regard to their attainability in either an on-premise deployment or at one of the major cloud SPs.

Performance, memory, security, high availability, virtualization, and interconnect bandwidth are server characteristics that are deemed important and, currently, more attainable on-premise than in the cloud, which points at a hybrid cloud model as an ideal approach.

TABLE 1

Importance of Accelerated Server Considerations and Attainability On-Premise and in the Cloud

	Importance	Attainability on-premise	Attainability in the cloud
Hardware			
Performance of the host CPU	●	●	◕
Amount of memory available to the accelerator	●	●	◕
Security of the accelerated servers	●	●	◕
High availability of the accelerated servers	●	●	◕
Server virtualization of the accelerated system	●	◕	●
Scaling up accelerators within a server node	◕	●	◕
Bandwidth of the interconnect between the accelerator and the host CPU	◕	●	◒
Performance improvements from the accelerator	◕	●	●
Power requirements of the accelerated servers	◕	◒	●
Scaling out of accelerated server nodes	◕	●	●
Heat dissipation from the accelerated servers	◕	◒	●
Software			
Ease of programming the accelerator	●	●	●
Diagnostics on the accelerated servers	●	●	◒
Availability of APIs, libraries, software development kits, toolkits, frameworks, programming languages, and so forth	◕	●	◕
Time required to program the accelerator	◕	◕	◕
Cost required to program the accelerator	◕	◕	◕
Support for open source such as OpenCL, OpenMP, and OpenACC	◕	●	◕
Datacenter			
Interoperability of the accelerated servers with the rest of the infrastructure	●	◕	◕
Skill levels required for the operating environment	●	◕	◕
Manageability of the accelerated servers	◕	●	◕

High: ●; Less high: ◕; Average: ◒

Source: IDC, 2019

THE IBM POWER SYSTEM AC922

The IBM Power System AC922 is the building block of the fastest supercomputer on earth, Summit, at Oak Ridge National Laboratory. IDC believes that IBM has been too modest about this achievement, announced in mid-2018. Summit is capable of 200 petaflops, and it is the first supercomputer ever to reach exaops, meaning 10^{18} operations per second. Summit is also the third-ranking greenest supercomputer on the planet. Interestingly, Summit was built for AI unlike many other supercomputers that are built for simulation and modeling. IBM offers accelerated compute platforms — you might call them "mini-Summits" — for businesses that are scaling very large AI workloads.

For organizations that are not operating at such scale, they can leverage the remarkable performance of a single Power System AC922 or a small to medium-sized cluster of Power System AC922s. IBM claims that, thanks to its dense design and the combination of PCIe Gen4 and InfiniBand, organizations can start with a single Power System AC922 node and then scale to a rack or even thousands of nodes with near linear scaling efficiency.

When IBM started building Linux-based scale-out Power Systems several years ago, it was laser focused on building systems for extremely data-intensive computing and as such ahead of the market with that strategy — AI and analytics were still emerging workloads then. At the same time, IBM made a point of extending the Power Systems' reputation as a secure and reliable platform for mission-critical data to these new Linux-based scale-out systems.

The result of these efforts has been a broad portfolio of 1- and 2-socket Linux servers that show some of the highest core performance in the industry. The Power System AC922 is the beast in this lineup — a 2-socket 2U system that ships with either 4 or 6 NVIDIA Tesla V100 GPUs plus, uniquely integrated with the POWER9 processor, NVIDIA's interconnect NVLink2 (the second generation of NVLink) that enables seamless CPU-GPU interaction and coherent memory sharing. The coherence allows the system to treat system memory just like GPU memory, enabling a simplification of programing and much larger AI model sizes. No other server platform today has NVLink directly built into the processor for extremely fast CPU-GPU connectivity, allowing the GPUs to gain high-bandwidth access to the DRAM over that link.

There are several other noteworthy details about the Power System AC922. Level 2 cache per core is 512KB, Level 3 cache is no less than 10MB, and system memory is available between 256 and 2,048GB. With eight threads, the POWER cores provide four times as many threads as x86 solutions — this is an important aspect for parallelized workloads such as AI. The Power System AC922 has a high-throughput on-chip fabric, including an on-chip switch that can move data at 7TBps. Moving data in and out of each core happens at 256GBps. Apart from NVLink2 also included are CAPI 2.0 for coherent and extremely high-bandwidth attachment to ASICs, FPGAs, or external flash storage and — a first in the industry — PCIe Gen 4 to connect to PCIe devices.

Cooling can be either air or water for the 4 GPU configured Power System AC922 and water only for the 6 GPU version. Here too, IBM is ahead of the market with a few other vendors. Water cooling is an emerging — or reemerging — technology in today's era of accelerated computers, allowing for higher densities — more GPUs per node and more nodes in a rack.

The benefits of the Power System AC922 with regard to AI are that the system allows for much faster AI training times. IDC sees training times as a significant impediment for organizations to go to production with their AI solutions. Often data scientists have to wait for days or even weeks before a

training model has completed, at which point they may have to make tweaks and start over again. The Power System AC922 allows them to iterate models much faster. What's more, thanks to the near direct access from the GPU to system memory, they can train using IBM's Large Model Support (LMS), meaning that they can use much larger and more complex models and/or higher precision levels than when relying on GPU memory only.

To enable scaling, IBM has developed a library called Distributed Deep Learning (DDL), which allows any single job to be nearly linearly scaled out across hundreds of servers in a way that is completely seamless for data scientists. All they need to do is make a simple call and request the number of GPUs they want. DDL hooks into ML frameworks such as TensorFlow, Caffe, Torch, and Chainer and enables these frameworks to scale to multiple GPUs. This makes it easy for a data scientist to scale their training project across dozens of GPUs, something that can otherwise be difficult to achieve, and have the system dynamically manage it for them.

Table 2 checks the Power System AC922 against the aforementioned hardware features that end users deemed important.

TABLE 2

IBM Power System AC922 on Important Hardware Features of an Accelerated Server

Hardware	
Performance of the host CPU	✓
Amount of memory available to the accelerator	✓
Security of the accelerated servers	✓
High availability of the accelerated servers	✓
Virtualization of the accelerated system	✓
Scaling up accelerators within a server node	✓
Bandwidth of the interconnect between the accelerator and the host CPU	✓
Performance improvements from the accelerator	✓
Power requirements of the accelerated servers	✓
Scaling out of accelerated server nodes	✓
Heat dissipation from the accelerated servers	✓

Source: IDC, 2019

The Power System AC922, whether on-premise or in the IBM Cloud, represents the hardware foundation of IBM's new open source-based AI stack (although it should be mentioned that IBM has made most of this stack available for x86-based hardware as well). For the Power System AC922, the stack is highly optimized. Deep learning is very dependent on the accelerators, and IBM has optimized the software to take advantage of the 4 or 6 GPUs in the Power System AC922, NVLink, and of clusters of Power System AC922 servers. The stack is available on a private cloud as well as in public clouds, including the IBM Cloud. It consists of:

- **Watson Studio for data preparation and model development, including Jupyter and RStudio.** Watson Studio now also includes Watson Machine Learning Community Edition (WML CE), essentially PowerAI integrated into Watson Studio. This is a free offering.
- **Watson Machine Learning, which is a runtime environment to train, deploy, and manage AI models both in private and public clouds.** This differentiates IBM from some of the public cloud solutions for model management, with the private cloud options being either IBM Cloud Private or a Kubernetes-based approach. Watson Machine Learning includes, among other things, Spark, TensorFlow, PyTorch, Chainer, Keras, and IBM's new "traditional" machine learning performance booster SnapML that has shown to be very popular for such data science methods as Logistic Regression, Decision Trees, and Random Forests. Also included here are WML Community Edition and what is now called Watson ML Accelerator. Watson ML Accelerator used to be branded as PowerAI Enterprise, IBM's software for training deep learning models. Watson ML Accelerator is focused on resource management, when multiple data scientists are trying to use the same infrastructure, and on scaling a single job elastically across the infrastructure.
- **Watson OpenScale, which provides AI model metrics as well as bias and fairness monitoring of the AI model.** OpenScale is akin to application performance management but for AI model performance monitoring. It can track accuracy of the model or implement a given set of metrics around bias and fairness and check the model on them.

On top of this stack, organizations can run a wide range of AI applications. One interesting example is IBM PowerAI Vision, which allows organizations to very easily and quickly develop a neural network model for various kinds of image classification and detection, essentially computer vision deep learning. PowerAI Vision provides comprehensive workflow support, including complete life-cycle management of installation and configuration, data labeling, model training, inferencing, and moving models into production. PowerAI Vision can be used for drone surveillance, safety regulation enforcement at work sites and plants, manufacturing quality inspections, city traffic management, and many other use cases.

FUTURE OUTLOOK

IDC has done extensive research into the future of AI, publishing predictions, drivers, impact on IT, and recommendations in so-called IDC FutureScapes and in other document types. This space is wholly insufficient to capture even a small part of these predictions. Suffice it to say that by 2024, AI will have dramatically transformed how we live our lives, conduct business, or run a datacenter.

In the short term, the most critical step will be to enable data scientists with the software and hardware tools that allow them to build better, higher-quality AI solutions faster. The amount of compute that data scientists require to fulfill the promise of AI is truly astounding. Accelerated compute is expected to become the norm rather than the exception for data-driven workloads. Even as GPUs (and FPGAs and ASICs) are delivering heretofore unimaginable processing performance, new AI processors are being invented at both start-up tech companies and large incumbent tech corporations, including IBM.

These new processors claim to deliver tenfold, even hundredfold, greater AI processing performance. At the same time, new software models are being invented to facilitate the infusion of AI in everything.

AI model training will continue unabated, and the models will become larger, more complex, and will require greater accuracy, for example, to rule out bias. Inferencing on those models will soon become the largest workload at the edge, and already retraining is taking place at the edge. What all this boils down to is that the AI momentum in terms of use case design and model development is unstoppable (i.e., unless we fail to build and operate the right infrastructure to support it). The days of experimentally "hitting the wall" with AI infrastructure are over. For many organizations that are starting to scale business-critical AI applications, infrastructure should now be a top priority.

CHALLENGES/OPPORTUNITIES

For Organizations

This document has discussed a range of infrastructure challenges organizations face when they are ready to scale their AI applications for production. From data preparation to model development to runtime environments to training, deploying, and managing AI models, the requirements for the underlying infrastructure defy the old models of general-purpose hardware. Only infrastructure that is designed for data-intensive workloads, with superior core performance, multiple GPUs, fast interconnects, large amounts of coherent memory, and exceptional I/O bandwidth can execute deep learning training workloads fast enough. Organizations will need to make decisions about replacing existing general-purpose hardware or supplementing it with hardware dedicated to AI-specific processing tasks. Obviously, the associated opportunity is the availability of processing power that will enable them to develop and run cutting-edge AI applications.

For IBM Power Systems

The challenge for IBM Power Systems is always one of market recognition. IBM offers exceptional AI infrastructure solutions that are packaged with well thought-out AI software stacks, but that potential customers — incorrectly — perceive as "different" or more costly. The subsequent knee-jerk reaction to go with one of the large commodity hardware vendors for extremely data-intense workloads is depriving these organizations of AI infrastructure solutions that they could truly benefit from. The Power System AC922, for example, is a supercomputing building block for every datacenter. IBM is streamlining its AI infrastructure and AI software branding under the Watson brand. In the long term, this makes sense and will benefit both IBM and its customers. In the short term, more must be done to clarify the new subbrands and eliminate confusion among customers and industry analysts. The opportunity for IBM lies in its technical prowess. Now that new AI workloads are starting to seriously challenge the on-premise and cloud infrastructure they run on, this is the moment for IBM to take the stage and woo — and wow! — its potential customers.

CONCLUSION

In the past several years, IDC has witnessed how many organizations have started to develop a wide range of AI capabilities. Initially launched as experiments by relatively inexperienced staff and executed on whatever infrastructure was available, these initiatives have now started to gain critical mass. Many organizations have developed extensive AI expertise, and they are having first-hand experience with the speed with which their AI capabilities are becoming a critical aspect of their business.

At the same time, IT too has journeyed through a learning curve with regard to the infrastructure to run AI on. There is today much greater clarity as to the infrastructure requirements for deep learning training or inferencing and how to scale those environments for production. That deep learning training requires different infrastructure than other workloads is pretty much a given. Deep learning training wants clustered nodes with strong processors, powerful coprocessors, fast interconnects, large I/O bandwidth, and plenty of memory.

Today, the biggest decision that IT has to make is how much per-node performance to make available for an AI workload, taking into account these aforementioned components and how they are connected and optimized together, because that's how performance gains are achieved — not with just a bunch of GPUs but with a platform that maximizes the utilization of those GPUs. IDC believes that IBM's Power System AC922 is an excellent choice to achieve maximum per-node performance for AI training. After all, the Power System AC922 is the building block of the world's fastest supercomputer.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

Global Headquarters

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
Twitter: @IDC
idc-community.com
www.idc.com