Intersect360 Research White Paper: IBM AND NVIDIA® SOLUTIONS POWER INSIGHTS WITH THE NEW AI

MARKET DYNAMICS

Machine Learning Gets More Intelligent

The first thing to remember about artificial intelligence (AI) is that we've seen intense interest in it before. Alan Turing, a famed pioneer of computer science, explored it in 1950 in his hallmark paper, "Computing Machinery and Intelligence," including the notion of the "Turing test," which set forward the proposition of a computer imitating a human in conversation. In the 1990s, the IBM computer Deep Blue used a combination of AI and brute-force computation to become the first computer to win a chess game and match against a reigning world champion, Garry Kasparov, bringing public speculation as to how far AI can go.

Today AI is again the subject of intense interest, promising to revolutionize broad aspects of human existence, addressing lofty topics such as limiting world hunger (by optimizing agriculture), finding energy (reading seismic geologic maps), avoiding deaths (with self-driving cars), and curing disease (with personalized medicine). Meanwhile, chatbots and virtual personal assistants are getting closer to acing that Turing test. If these pie-in-the-sky plans sound too good to be true, there is reason to be optimistic. Thanks to a confluence of factors, AI has come a long way.

To understand why, it helps to peel back some of the magic. Traditionally, most computer programs have been *deterministic*; that is, they calculate answers to math questions based on a set of input values and formulas they are programmed with. This approach works, though it can be computationally expensive, and it is typically as accurate as the formulas describing the model. The computer can tell you how a hurricane will advance toward the coast, provided you have perfect formulas describing how a hurricane behaves, perfect atmospheric data to feed into the model, and a handy supercomputer for running the simulation. Another type of application is *probabilistic*; rather than calculating an exact answer, it iterates scenarios to determine what is likely. Known as Monte Carlo simulations, these are deployed in special situations when experiment parameters are well-known and controlled.

AI, or more technically, *machine learning*, presents a third category of application—one that is *experiential*. Based on a repository of past data, it is programmed to make inferences about new data, based on pattern recognition. We refer to this type of application as "artificial intelligence" because it mimics how we as humans typically learn. I may have never seen this

P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

particular fish before, but I am confident it is a fish, and I have expectations of its fishy characteristics and behaviors, based on my previous history of experiences regarding fish. I may have never driven in this particular neighborhood before, but I am reasonably confident driving here, because I know how to drive in general, and it is similar to other neighborhoods I have driven in before.

Note that even human learning has limitations. Seeing a fish, can I tell you what type of fish it is, or whether it is typically found in these waters? That depends on whether I have acquired special fish-based knowledge or experience. I had to be taught specifically that dolphins are mammals, not fish; they look and act more like fish than like gophers. And if I found a fish—even a familiar goldfish—doing something unusual, like sitting at my kitchen table reading the newspaper, I may have a crisis in reconciling my past knowledge with my current experience. And I might find that no matter how much experience I have driving under my own normal conditions, I have less confidence extrapolating to other conditions, like driving in a snowstorm or in a foreign country.

Nevertheless, human existence suggests to us that the more experience we acquire, the more confidently we can apply it to new situations. Here we have the revolution in AI. Through the internet, cloud-scale computing, and worldwide digital connectedness, we are now able to train machine learning algorithms with vastly more data than ever before, leveraging advanced AI platforms and accelerated computing.

Machine learning takes place in two phases—*training* and *inference*. In the training phase, a machine learning model is given access to troves of information, generally together with a notion of success or affirmation (e.g., this image is or is not a cat; this sound is or is not the word "cat"). The data is fed through a "deep neural network," a hierarchical architecture designed to eliminate superfluous information to focus on the question at hand, and to identify the key elements of data that best correlate with the yes-or-no decision.

In the inference phase, new data can be compared to the previous model. How well does it match the previous data? Can a decision be made based on the previous data, based on how similar this data is? How confident is that decision? Here it is important to realize that mistakes can be made, but humans make these mistakes in learning too. (The dolphin has fish-like qualities but is not a fish.) The more training, and the broader the training, the more accurate the model becomes. When a machine learning algorithm is giving ongoing, recursive feedback to refine and improve the model, it is known as *deep learning*.

In theory, AI can have long-term advantages over human intelligence. Computers can be trained faster and with more data than humans can, such that a machine learning model has access to more practical experience than all but the most expert humans. As such, there is a wide variety of applications in which AI shows promise for large-scale, revolutionary change.

AI is promising to revolutionize broad aspects of human existence, addressing lofty topics such as limiting world hunger (by optimizing agriculture), finding energy (reading seismic geologic maps), avoiding deaths (with self-driving cars), and curing disease (with personalized medicine).

P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

Applications for Machine Learning

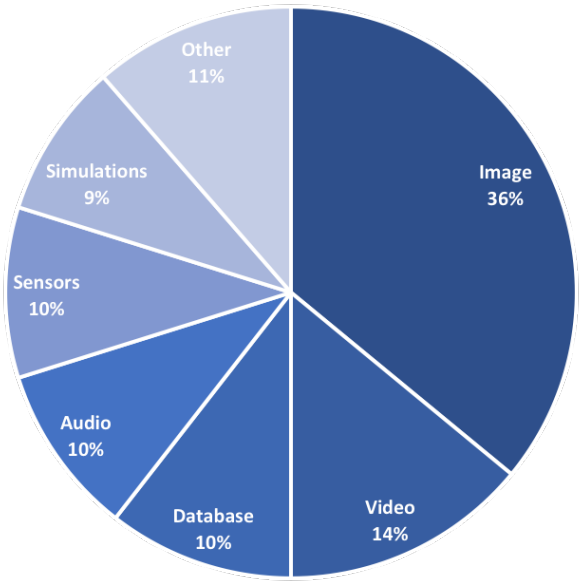
Part of the need for AI comes from the vast increase in data that is stored and accessible. The big data analytics revolution highlighted this need, to look for valuable information and insights from large amounts of unstructured data. Much of this data falls outside of traditional values, such as numbers and texts, and instead in modern, multimedia forms, such as images, audio, and video streams.

This rich media falls into a category that humans are able to process but has defied computers' ability to search. Given a stack of photos, a person might be able to pick out which ones have Brad Pitt wearing a hat, but unless they were properly tagged with metadata, the computer search function could not do the same. Machine learning has the potential to unlock this limitation.

A 2018 Intersect360 Research survey of machine learning usage revealed that for 36% of applications, the source data for analysis was image-based. (This did not include certain specialty types of images, such as medical imagery captured by medical devices.) Another 14% was video, and 10% was audio. The figure below shows how these rich data sources are primary targets for machine learning. Overall, 73% of applications in the study involved image, video, audio, or sensor data.

Data Types for Machine Learning Applications

Intersect360 Research, 2019



When broken into constituent applications, these top-level categories fragment into many specialties, showing that the same types of analysis capabilities can be applied to different types of data. Medical imaging and autonomous driving were the most commonly named domains. This reflects a combination of dynamics: the applicability of machine learning to

these domains; the marketability of these applications as wide-reaching and beneficial; the corresponding interest in research in these areas; and the relative secrecy of other fields, such as financial services, with respect to their own applications.

25 different vertical markets were represented in the survey, not including a general AI category for companies developing machine learning capabilities across multiple vertical markets. 18 of the 25 verticals were mentioned by only one or two respondents, indicating the "long tail" of general applicability to a wide range of applications. Machine learning can therefore augment other computing areas, leading to incremental growth across the entire computing market spectrum.

Technologies for Machine Learning

The boom in AI is not only due to the availability of data, but also to the power of accelerated computing resources that can be tied to it. In many cases, these technologies were already in development and use for traditional high performance computing (HPC) and big data applications. Many of these have gotten an additional boost from the applicability to machine learning algorithms. In addition, the upswing in AI development has led to the creation of targeted products aimed directly at specific portions of the machine learning spectrum.

Across all these technologies, the greatest focus is on the AI platform, inclusive of the processing elements and an associated software stack with optimized deep neural network models. Most machine learning training deployments currently leverage graphics processing units (GPUs) as computational accelerators. As machine learning applications have become more common, so have server configurations that include GPUs.[1]

GPUs emerged as HPC accelerators in the mid-2000s and have seen a steady increase in adoption since that time. A GPU-accelerated application makes use of the GPU as a co-processor, breaking out computationally intensive portions of algorithms to be run faster on the GPU than on the host microprocessor, which carries overhead such as inter-node communication, job management, and running the operating environment. The availability of GPUs as accelerated computational elements has helped to bring about the new era of modern AI.

Naturally, productivity requires more than hardware. Real metrics of time-to-solution involve bringing AI solutions from proof-of-concept to production deployment quickly. A customized software stack—with optimized deep neural network models and GPU acceleration libraries—ensures the highest throughput on GPUs in a seamless, unified AI platform. This brings solutions to production faster and maximizes their utility for data scientists.

The boom in AI is not only due to the availability of data, but also to the power of accelerated computing resources that can be tied to it.

A customized software stack—with optimized deep neural network models and GPU acceleration libraries—ensures the highest throughput on GPUs in a seamless, unified AI platform.

[1] Intersect360 Research, "Worldwide HPC 2018 Total Market Model and 2019–2023 Forecast: Products and Services," May 2019.

Furthermore, a server needs more than GPU computing capability to be considered ideal for machine learning. For starters, there is still the question of data movement and data management. The bandwidth in transferring data in any accelerated computing architecture can be critical for success. Furthermore, most organizations still have workloads beyond AI, and as such, the management of complete enterprise workloads, often in hybrid cloud environments, can be paramount in any enterprise technology evaluation.

IBM POWER SYSTEMS SOLUTIONS FOR MACHINE LEARNING

With so much potential for breakthrough advancements, AI has captivated the collective imagination of consumers, researchers, and enterprises. As such, leading providers of accelerated computing solutions have turned their attention to what configurations will support machine learning in the context of high-performance enterprise environments.

Among technology vendors, IBM has been a first mover in embracing forward-looking concepts such as big data and AI. In 2011, IBM's Watson AI platform captivated audiences with its natural-language comprehension abilities by beating human champions on the popular trivia game show *Jeopardy!*; that technology is now expanded to be a central core of enterprise AI solutions.

On the hardware front, IBM has introduced its newest line of servers based on the latest IBM POWER9 processors. The IBM Power System AC922 model specifically targets enterprise AI workloads. The Power AC922 has dual POWER9 processors and up to four NVIDIA Tesla V100 GPUs. NVIDIA has been the dominant leader in GPU computing—over 90% of accelerator-based HPC systems incorporate NVIDIA GPUs for computation[2]—and NVIDIA has been expanding its presence with chips optimized for AI workloads.

What sets the IBM solution apart is its ability to move data quickly, CPU to CPU, GPU to GPU, and CPU to GPU. With POWER9, IBM has chosen to focus its processors on data throughput, from memory, between processors, and with GPU accelerators, keeping its computational engines working.

- CPU-to-CPU and CPU-to-memory: The current POWER9 "SU" (Scale Up) model runs a third-generation OpenCAPI (Coherent Accelerator Processor Interface) 3.0 interconnect, which provides a secure, native 25 Gigabit-per-second transfer protocol on-chip, compared to the 8 Gbps available on PCIe Gen3. POWER9 SU also introduced buffered memory, with 210 GB/sec of sustained memory bandwidth and support for alternate memory architectures.
- *CPU-to-GPU:* Unlike servers based on x86 microprocessors, the POWER9-based Power AC922 supports NVIDIA's second-generation NVLink 2.0 in moving data between CPU and GPU. In internal benchmarks, IBM claims a 5.6x advantage in

IBM has been a first mover in embracing forward-looking concepts such as big data and AI.

The IBM Power System AC922 model specifically targets enterprise AI workloads.

What sets the IBM solution apart is its ability to move data quickly, CPU to CPU, GPU to GPU, and CPU to GPU.

[2] Intersect360 Research HPC User Site Census survey data, 2019.

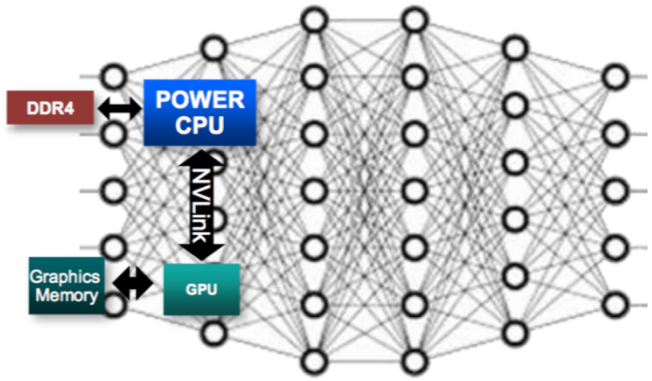
data throughput versus a competing x86 system using PCIe Gen3.[3] NVLink 2.0 also connects GPU to GPU for high-bandwidth transfer between accelerators.

- *I/O Bandwidth:* In addition to OpenCAPI, POWER9 has 48 lanes of PCIe Gen4, for a total of 192 GB/sec of duplex bandwidth. (Because it is an I/O protocol, PCIe base units are typically expressed in bytes, as opposed to networking speeds, which are typically expressed in bits. It is worth noting that 192 GB/sec full duplex equates to 768 Gigabits per second in each direction, over three-quarters of a Terabit.) POWER9 supports 25 GigaTransfers/sec and 300 GB/sec of advanced I/O signaling.

These efficiencies in data movement translate into real-world advantages for AI workloads, according to IBM. Based on internal measurements, IBM claims a 3.8x advantage in model training time over competing x86 architecture–based systems.[4] In addition to the time savings, the IBM Power AC922 solution allows the combination of memory components across both CPU and GPU to support larger models in memory. (See Figure below.)

IBM POWER9 Plus GPU Memory Architecture for Machine Learning Applications

Source: IBM



As for scale, IBM has two powerful case studies. The IBM Power System AC922 POWER9-based server with NVIDIA GPUs forms the backbone for the two most powerful computers in the world, according to the latest semiannual TOP500 ranking of supercomputers worldwide[5]: Summit, installed at Oak Ridge National Laboratory, and Sierra, at Lawrence Livermore National Laboratory. Both supercomputers are run by the U.S. Department of Energy, combining AI with traditional deterministic models to advance the forefront of scientific discovery.

---

[3] Source: IBM. CUDA H2D bandwidth test conducted by NVIDIA: Intel Xeon E5-2640 V4 with NVIDIA P100 versus POWER9 with NVIDIA V100. https://www.ibm.com/downloads/cas/6PRDKRJ0.

[4] Source: IBM. 1,000 iterations of Enlarged GoogleNet model (mini-batch size=5) on Enlarged Imagenet Dataset (2240x2240), test run by IBM. Power AC922, 40 cores (2 x 20c chips), POWER9 with NVLink 2.0, 4x NVIDIA Tesla V100 GPU, versus 2x Intel Xeon E5-2640 v4, 20 cores (2 x 10c chips), 40 threads, 2.4 GHz, 4x NVIDIA Tesla V100 GPU. https://www-03.ibm.com/press/us/en/pressrelease/53452.wss.

[5] http://www.top500.org.

INTERSECT360 RESEARCH ANALYSIS

The second thing to remember about AI is that the emphasis should be on *artificial*, not *intelligence*. AI is a powerful new tool in the computational arsenal for scientific and business computing. It mimics intelligence. It can be programmed to find solutions in ways that the programmers may not have expected.

As advanced as AI is becoming, it is most exciting to think about the ways it can be used in combination with traditional supercomputing techniques, not merely in replacement of them. Consider: If machine learning algorithms can be trained to play a game such as poker or go, then they might be taught to play different games, such as predict-the-hurricane, optimize-the-airplane-wing, or discover-the-drug. Such a deployment would effectively place AI into the role of computational steering, not replacing the traditional models, but running them iteratively, as a scientist or engineer might, trying different approaches, perhaps millions of them, and highlighting the ones that seem the most promising.

IBM's approach embraces the current direction of machine learning. By ceding the AI platform—both accelerator and software—to its partner NVIDIA, IBM can focus itself on system architecture, data bandwidth, and software solutions, where it has strengths. That said, IBM should embrace and take credit for HPC—not only AI—where it can, including with world-leading systems like Summit and Sierra. NVIDIA, meanwhile, has complete solutions that go beyond the GPUs themselves, with software stacks and frameworks that are optimized for AI.

AI is often best deployed in combination with high-performance technologies, as well as with enterprise data management, analytics, and IT trends such as cloud and Internet-of-Things (IoT). With the breadth of its solutions, its traditional strength in enterprise IT, and its forward-looking views on evolving trends, IBM is well-positioned to help organizations incorporate high-performance solutions for AI into the enterprise landscape.

If machine learning algorithms can be trained to play a game such as poker or go, then they might be taught to play different games, such as predict-the-hurricane, optimize-the-airplane-wing, or discover-the-drug.

For more information about the IBM Power System AC922 solution for AI, visit: https://www.ibm.com/us-en/marketplace/power-systems-ac922/.

For more information about Nvidia solutions for AI, visit: https://www.nvidia.com/en-us/deep-learning-ai/.

P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com