ACCELERATE DEEP LEARNING WITH A MODERN DATA PLATFORM

PRODUCED BY TABOR CUSTOM PUBLISHING IN CONJUNCTION WITH:

datanami

• BIG DATA • BIG ANALYTICS • BIG INSIGHTS •

SPONSORED BY:

PURESTORAGE®

SOCIETAL IMPACT OF ARTIFICIAL INTELLIGENCE

Massive amounts of data are being created driven by billions of sensors all around us such as cameras, smart phones, cars as well as the large amounts of data across enterprises, education systems and organizations. In the age of big data, artificial intelligence (AI), machine learning and deep learning deliver unprecedented insights in the massive amounts of data.

Amazon CEO Jeff Bezos spoke about the potential of artificial intelligence and machine learning at the 2017 Internet Association's annual gala in Washington, D.C., "It is a renaissance, it is a golden age," Bezos said. "We are solving problems with machine learning and artificial intelligence that were in the realm of science fiction for the last several decades. Natural language understanding, machine vision problems, it really is an amazing renaissance." Machine learning and AI is a horizontal enabling layer. It will empower and improve every business, every government organization, every philanthropy — basically there's no institution in the world that cannot be improved with machine learning."

Technology companies such as Amazon, Apple, Baidu, Facebook, Google (Alphabet), Microsoft and NVIDIA have dedicated teams working on AI projects in areas such as image recognition, natural language understanding, visual search, robotics, self-driving cars and text-to-speech. Examples of their innovative AI, machine and deep learning projects include:

- **Amazon:** Amazon uses AI and complex learning algorithms that continuously assess the market dynamics to determine product recommendations and which products are selected for the Amazon Buy Box.
- **Apple:** Apple's Siri virtual assistant on iPhones and other Apple hardware uses deep learning to do searches as well as provide relevant answers interactively using a voice interface.
- **Baidu:** A speech-recognition system called Deep Speech 2 developed by Baidu easily recognizes English or Mandarin Chinese speech and, in some cases, can translate more accurately than humans.
- **Facebook:** Facebook's DeepMask and SharpMask software works in conjunction with its MultiPathNet neural networks allowing Facebook to understand an image based on each pixel it contains.
- **Google (Alphabet):** Google CEO Sundar Pichai indicates that when users tap on Google Maps that the Google StreetView product uses AI to automatically recognize street signs or business signs to help define the location.
- **Microsoft:** Microsoft's AI uses a cognitive vision system in PowerPoint that analyzes photos and auto-generates Alt-Text or suggests diagrams that illustrate that process.
- **NVIDIA:** NVIDIA DRIVE™ PX is the open AI car computing platform that enables automakers and tier 1 suppliers to accelerate production of automated and autonomous vehicles.

IMPORTANCE OF AI

Research by Forrester called "Artificial Intelligence: What's Possible for Enterprises in 2017" indicates that AI is finally becoming a reality and that more organizations, researchers and educational institutions are looking at its potential. The report found that "only 12% of the 391 business and technology professionals it polled are currently using AI systems. However, 58% are researching AI technologies and what it takes to support their use in the enterprise, and 39% are identifying and designing AI capacities to deploy. The report, published in November 2016, also found that 36% of respondents are educating the business or building the business case about AI's potential."

BIG BANG OF ARTIFICIAL INTELLIGENCE

The emergence of AI started when three key technologies came together like a perfect storm, known as the big bang of AI. The three key drivers are deep learning algorithms, parallel processors based on graphics processing units (GPUs) and the availability of big data.

DEEP LEARNING—NEW COMPUTING MODEL THAT WRITES ITS OWN SOFTWARE

Traditionally, programs were designed to sequentially process data and to use specific code instructions in the processing. Deep learning allows computer systems to analyze data to provide insights and predictions about the data. Machine learning refers to any type of computer program that can learn by itself without being programmed by a human. Deep learning, also called deep structured learning or hierarchical learning, is an element of machine learning that uses artificial neural networks. Deep learning systems can be supervised, partially supervised or unsupervised.

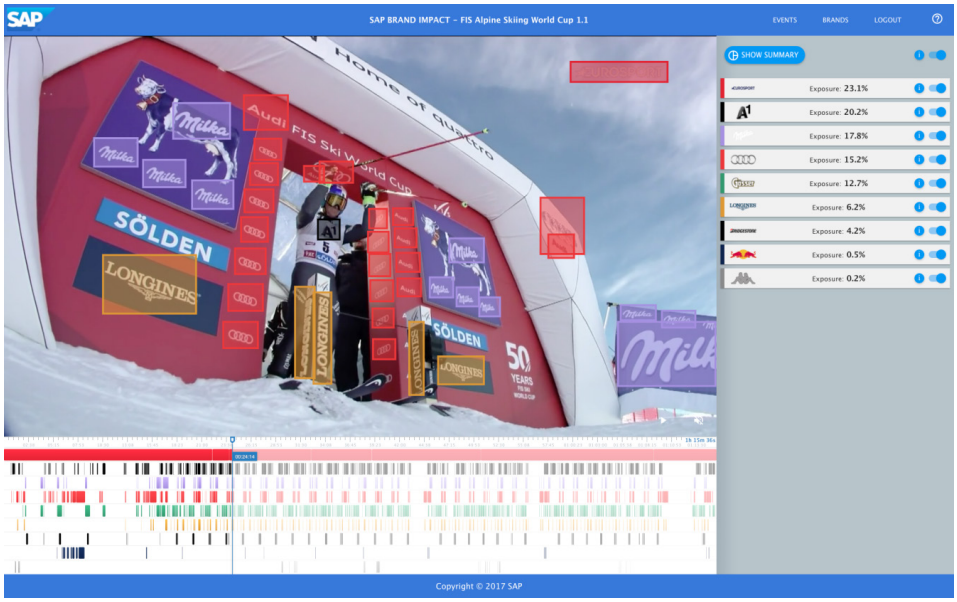

Figure 1. SAP Brand Impact — capturing brand logo placement in near real time. (Source: SAP).

AI IN ACTION: MEASURING BRAND IMPACT

The SAP Brand Impact product is an example of how deep learning is used to analyze a company's brand exposure. The SAP application was shown thousands of images or videos and trained on how to recognize logos and other brand information from the images, without the need to explicitly program the software. Many brands rely on sponsoring televised events and they typically use a manual process that takes up to six weeks after an event to report brand impact return or investment from their logos shown during the event. It can take an entire quarter to adjust brand marketing expenditures.

SAP Brand Impact software uses deep neural networks trained on an NVIDIA DGX-1 system. SAP's deep learning analysis[i] provides immediate and accurate results of the logos in the video. With the SAP software, auditable results are delivered in a day. Figure 1 shows an example of video analysis.

GPU: THE MODERN PARALLEL PROCESSOR

Modern compute typically consists of multi-core CPUs or GPUs. It's not uncommon for CPUs to have up to 20 cores and GPUs have thousands of cores as shown in Figure 2. Both CPUs and GPUs are parallel processors capable of performing parallel computing on more than one task.

In 1997, NVIDIA pioneered GPU-accelerated computing, a new computing model that accelerates

compute-intensive portions of the application to the GPU, while the remainder of the code still runs on the CPU. Powered by NVIDIA Volta™, the latest GPU architecture, NVIDIA introduced the Tesla® V100 which offers the performance of 100 CPUs in a single GPU.

Today, both multi-core CPUs and GPUs are used to accelerate deep learning, analytics, and engineering applications—enabling data scientists, researchers, and engineers to tackle challenges that were once impossible. New deep learning algorithms leverage massively parallel neural networks inspired by the human brain. Instead of experts handcrafting software, a deep learning model writes its own software by learning from many examples, delivering super-human accuracy for common tasks like image, video, and text processing.

Figure 2 provides an example of how the combination of using parallel computing and running new deep learning massively parallel algorithms provide a superhuman accuracy rate in identifying images.

BIG DATA

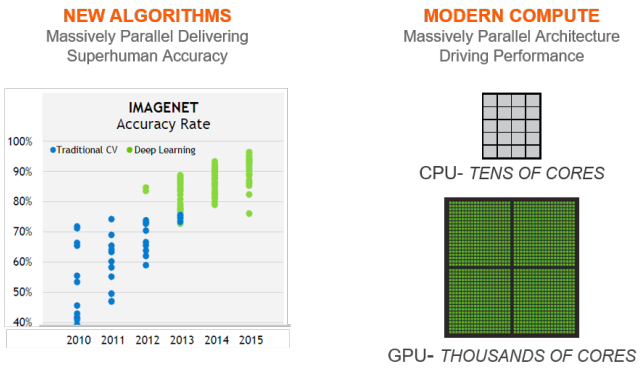

Figure 2. The big bang of intelligence fueled by parallel computing, new algorithms and big data. Courtesy of Pure Storage.

Data is the most important asset in an organization. In fact, the Economist in May 2017 claimed that data has become more valuable than oil. And data continues to grow. According to IDC's The Digital Universe in 2020 report, "The Digital Universe is the total collection of newly created data including streaming video and stored digital data which is growing at an incredible pace. It doubles in size every two years, in 2013 it was 4.4 Zettabytes, and is expected to exceed 50 Zettabytes in 2020."

Deep learning technology and neural networks have been around for a long time. So why is deep learning now starting to peak and what is the value of big data? Andrew Ng, a luminary in the field of AI, described the evolution of big data and deep learning at the 2016 Spark Conference[ii]. Ng indicated that if you take an older traditional learning algorithm such as logistic regression and feed it more data, the system performance plateaus because the algorithm cannot squeeze any more insight with more data. Ng observed that deep neural networks are different. The more training data is fed into the neural network, the more accurate it becomes, as shown in Figure 3. The adoption of deep learning is rapidly growing because of algorithm innovation, performance leaps of GPU-based computing systems, and the constant growth of big data.

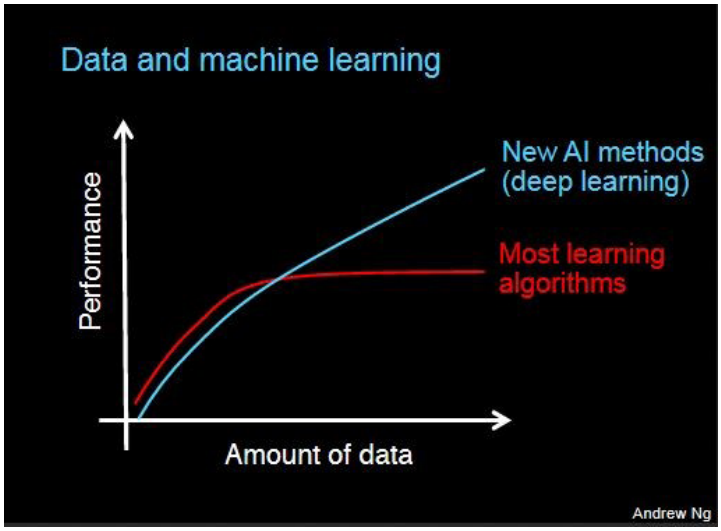

Figure 3. Deep learning performance grows with more data. (Source: Andrew Ng).

WHY TRADITIONAL STORAGE CAN'T MEET DEEP LEARNING NEEDS

There's been significant advancement in parallel computing and algorithms, but the technology that stores and delivers big data has largely been built on legacy building blocks, designed in the serial era. A new type of storage system is required to deliver the massive amounts of data for these new computing paradigms.

In the past several years alone, the amount of compute required by deep learning and the amount of compute delivered by GPUs jumped more than 10 times. Meanwhile, disks and SSDs have not increased in performance during the same period. While the volume of unstructured data is exploding, legacy storage struggles to handle the storage performance needs of the emerging big data drivers.

Most deployments today use direct-attached storage (DAS) or distributed direct attached storage (DDAS) where datasets are spread across disks in each server. Use of DDAS allowed data scientists to use commodity off-the-shelf systems/components for their analytics pipeline, like X86 processors and standard hard disk drives, but the approach is full of potential problems. At the time that modern data analytics technologies were being developed, there wasn't a storage platform big enough for such large amounts of data nor fast enough to meet the high bandwidth requirements from big data software.

A goal of modern analytics is to analyze data and extract insights from that data. This information is often unstructured data from sources such as logs and Internet of Things (IoT) devices. The older systems may only work for highly normalized data and are not optimized to analyze semi-structured and unstructured data often used in AI and deep learning. As a result, legacy storage has become the primary bottleneck for applications—performance is bottlenecked by decades-old, serial technologies in the stack that's not optimized for unstructured data. If data is the new currency for the fourth industrial revolution, why is the storage industry still stuck in the serial era?

REAL BUSINESS BENEFITS FOR LEVERAGING DEEP LEARNING

Companies and institutions that have already invested in AI or deep learning saw these benefits:

Forrester results: 25% said they achieved business process efficiency, 21% improvement in customer satisfaction and 18% cost savings.

Health care institution: Deployed agile predictive analytics to discover that 10 % of the covered employees were consuming 70% of the resources due to chronic conditions and other issues with care management.

Data center: DeepMind AI reduces data center cooling bill by 40%.

Cost reduction for testing computer chips: A major chip manufacturer saved a total of $3 million in manufacturing costs using predictive analytics for chip testing.

Machine learning at Tour de France: In 2017, the Tour de France used machine learning for predictions and merging historical data with 2017 race content providing more detailed information than previously available.

Enabling the smart electric grid: Research by the McKinsey Global Institute found that AI can make the electrical grid smarter by using sensors and machine learning to allow by-the-minute adjustments to maximize electric generation efficiency.

Improving job performance: Factories use the UpSkill Skylight platform augmented reality smart glasses that connect hands-on workers to the information they need to do their jobs with greater efficiency and fewer errors. Average worker performance has improved up to 32%.

FLASHBLADE BUILT FOR DEEP LEARNING

In the new age of big data, applications are leveraging large farms of powerful servers and extremely fast networks to access petabytes of data served for everything from data analytics to scientific discovery to movie rendering. These new applications demand fast and efficient storage, which legacy solutions are no longer capable of providing.

What's needed is a new, innovative storage architecture to support advanced applications while providing best-of-breed performance in all dimensions of concurrency – including input/output operations per second (IOPs), throughput, latency, and capacity – while offering breakthrough levels of density. The new FlashBlade™ flash-based storage by Pure Storage® meets all these needs. FlashBlade can handle big data and concurrent workloads that will drive tomorrow's discoveries, insights and creations.

PURE STORAGE

For the past four years, Gartner has rated Pure Storage as a Leader in the Magic Quadrant of Solid State Arrays for their innovations in all-flash data storage. Since Pure Storage unveiled its FlashBlade scale-out storage platform, the company has made significant inroads in providing storage for real-time and big data analytics, financial analysis and manufacturing.

The FlashBlade architecture is designed from the ground-up for modern analytics workloads, delivering high performance, cost-effective, simple-to-own-and-operate scale-out storage for petabytes of operational data. It is specifically designed for flash media and the architecture contains no provision for mechanical disks. **FlashBlade is purpose-built for massively parallel workloads that are required for deep learning processing.**

The key property of FlashBlade is to deliver elastic performance at scale – the ability to increase performance, capacity, connectivity, and functionality as customer requirements dictate. This is possible because at its core, from software to flash, FlashBlade is massively parallel. With its unique Evergreen™ business model, customers never rebuy any terabytes (TBs) they already own and can upgrade technologies as they evolve without disrupting service or diminishing performance or data integrity. See Figure 4 for an example of a FlashBlade Chassis.

Figure 4. FlashBlade Chassis: 1.6 Petabytes in 4U. Courtesy of Pure Storage.

DELIVERING DATA THROUGHPUT FOR AI

Deep learning systems often use mostly small files to keep the training computers busy. In the example shown in Figure 5, the deep learning training is running on NVIDIA DGX-1 servers and the FlashBlade data storage platform. In the example, each DGX-1 is processing 13k images per second through AlexNet using Microsoft CNTK framework. The training model uses small files with random access, which older legacy systems do not handle efficiently. In this example, a FlashBlade can deliver enough ingest throughput to maximize training performance on multiple DGX-1 systems.

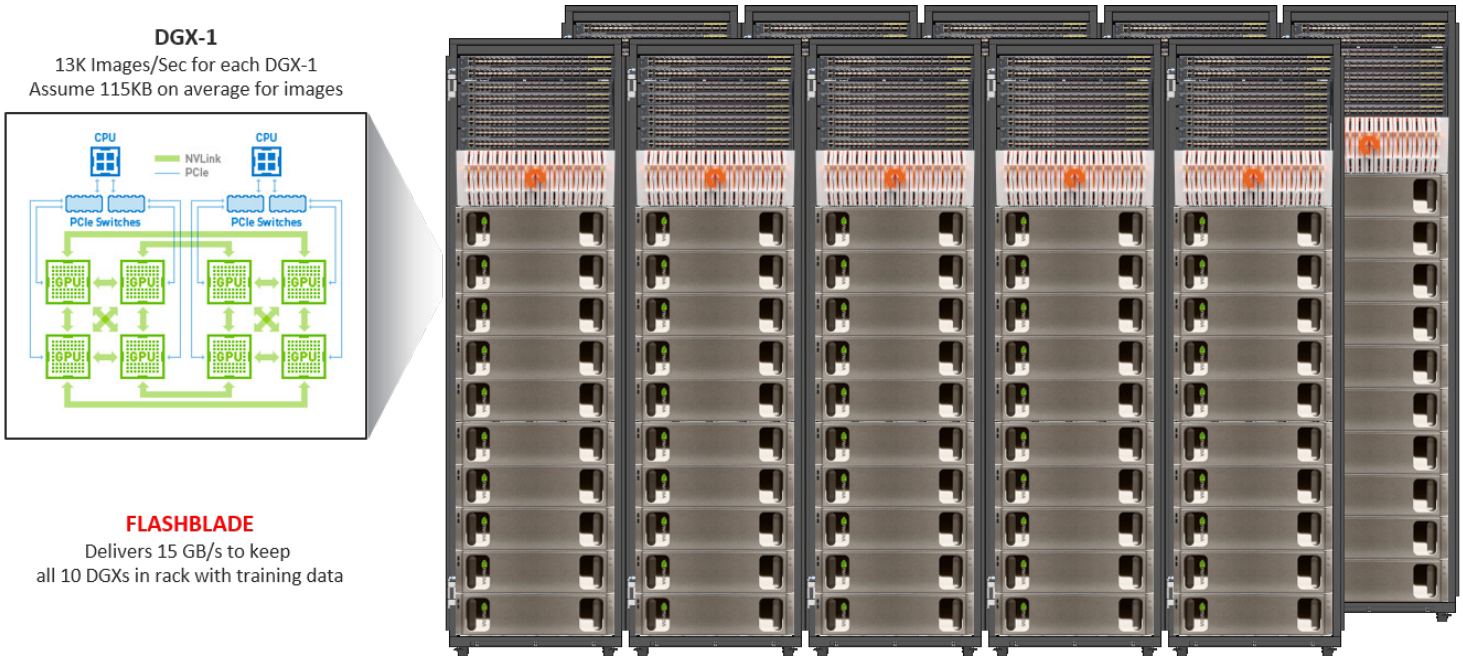

Figure 5. Example of how FlashBlade delivers required AI throughput. Courtesy of Pure Storage.

SUMMARY

Data is growing at amazing rates and will continue this rapid rate of growth. New techniques in data processing and analytics including AI, machine and deep learning allow specially designed applications to not only analyze data but learn from the analysis and make predictions.

Computer systems consisting of multi-core CPUs or GPUs using parallel processing and extremely fast networks are required to process the data. However, legacy storage solutions are based on architectures that are decades old, un-scalable and not well suited for the massive concurrency required by machine learning. Legacy storage is becoming a bottleneck in processing big data and a new storage technology is needed to meet data analytics performance needs.

The FlashBlade all-flash storage array from Pure Storage is designed to meet these needs. FlashBlade performance improves linearly with more data. Whether files are small or large, FlashBlade delivers true linear scaling of capacity and performance, and as a result, is well-suited to modern analytics workloads for AI and deep learning.

"Modern computing frameworks have given rise to increasingly complex, high-performance analytics and valuable data," said Par Botes, VP of Engineering, Pure Storage. "With FlashBlade, our mission is to make big data into fast data with an all-flash platform that is big, fast, and simple to deploy – and one that provides value across all industries and segments."

ABOUT PURE STORAGE

Pure Storage[iii] (NYSE:PSTG) helps companies push the boundaries of what's possible. Pure's end-to-end data platform - including FlashArray, FlashBlade and our converged offering with Cisco, FlashStack – is powered by innovative software that's cloud-connected for management from anywhere on a mobile device and supported by the Evergreen business model. The company's all-flash based technology, combined with its customer-friendly business model, drives business and IT transformation with solutions that are effortless, efficient and evergreen. With Pure's industry leading Satmetrix-certified NPS score of 83.5, Pure customers are some of the happiest in the world, and include organizations of all sizes, across an ever-expanding range of industries.

i NVIDIA and SAP Partner to Create a New Wave of AI Business Applications, https://blogs.nvidia.com/blog/2017/05/10/nvidia-sap-partner/.

ii AI: The New Electricity, Andrew Ng, Spark 2016 Summit, https://www.youtube.com/watch?v=4eJhcxfYR4I.

iii Pure Storage, FlashBlade and the "P" logo are trademarks or registered trademarks of Pure Storage in the U.S. and other countries. All other trademarks are the property of their respective owners.