IDC PERSPECTIVE

AI and Deep Learning on GPUs Everywhere: Impressions from NVIDIA's GPU Technology Conference 2017

David Schubmehl
Lloyd Cohen
Linn Huang
Hiroshi Shibutani
Lydie Virollet
Chwee Kan Chua
Shawn Fitzgerald
Peter Rutten
Vernon Turner

EXECUTIVE SNAPSHOT

FIGURE 1

Executive Snapshot: AI and Deep Learning Everywhere: GTC 2017

This IDC Perspective examines and provides an overview of the NVIDIA GTC Conference held in San Jose, California, which highlighted a range of new cognitive/AI and deep learning platforms and services by NVIDIA and its partners to help companies develop AI-based solutions and tools. Specifically, NVIDIA announced its Tesla V100 and personal DGX Station, which provide an order of magnitude processing performance over last year's Pascal using neural networks and other types of deep learning.

Key Takeaways

- NVIDIA and its partners have made tremendous strides in just the past year to "democratize AI and deep learning."
- One of NVIDIA's major goals is to accelerate the adoption of deep learning through a number of initiatives, including the creation of a Deep Learning Institute to educate developers and enterprises.
- NVIDIA sees the future as accelerated computing and deep learning on enterprise data within every firm and is working with partners like SAP, Microsoft, IBM, Dell, Amazon, and others to achieve this vision.
- NVIDIA is working with all of the major cloud vendors to provide GPU-based deep learning infrastructure.

Recommended Actions

- Develop a cognitive/AI systems strategy for your organization and begin implementing it.
- Examine and proactively invest in the development and supervision of AI and deep learning over the next 24-36 months.
- Incorporate cognitive/AI capabilities into enterprise collaboration and decision making – creating intelligent enterprise applications using machine learning and deep learning tools and frameworks.
- Actively evaluate AI and deep learning tools like NVIDIA's DGX-1 for suitability in areas such as digital transformation, recommendation systems, accelerated research, and improved collaboration.

Source: IDC, 2017

June 2017, IDC #US42710717

SITUATION OVERVIEW

NVIDIA held its annual GPU Technology Conference (GTC) in San Jose, California, May 8-11, 2017. With over 7,000 attendees, the conference has grown by 3x in the past five years. At the same time, NVIDIA estimates that the number of GPU developers has grown 11x in five years to well over half a million strong. This year, NVIDIA celebrated the rise of the GPU from an esoteric gaming and graphics device to becoming a core component advancing deep learning and artificial intelligence (AI) applications into the enterprise and announced several products and programs intended to help the company establish its presence as a key supplier in the world of deep learning and artificial intelligence infrastructure.

Introduction

NVIDIA is a relatively new company in the computer marketplace when you think of core foundation companies like Intel and Microsoft. Founded in 1993 by Jensen Huang, Chris Malachowsky, and Curtis Priem, NVIDIA launched its first product in 1995, the NV1 — which was the first microprocessor to integrate GPU acceleration, 3D rendering, video acceleration, and wave-table synthesis into a single chip. In 2006, the CUDA architecture (Compute Unified Device Architecture) was invented. CUDA is a parallel computing platform and application programming interface (API) that allows software developers to use a graphics processing unit (GPU) for general-purpose computing. The GPU was introduced to the market in 1999 with the launch of the NVIDIA GeForce 256. The first GPU for general-purpose computing was launched in 2007, and with a CPU, it can accelerate analytics, deep learning, high-performance computing (HPC), and scientific simulations. The GPU is designed as a specialized computer processor used in situations where processing of large blocks of data in parallel is required.

With its current technology, the CPU is reaching a point where performance is impacted. For years, the premise of Moore's law was that the number of transistors in a dense integrated circuit would double approximately every two years. We are now at a point where the size of the transistor and performance is reaching its limit. GPU computing recognizes that CPUs are very good at single-threaded operations. The goal of GPU computing is to offload parallelized workloads from the CPU and to act as a companion to the GPU where compute can be better optimized.

With the transformation now under way with the movement to the 3rd and 4th Platform, which incorporates accelerated computing, deep learning, artificial intelligence, and cloud computing, to name a few, the stage is now set for GPU computing to expand from the HPC market and more limited workloads to the mainstream of computing.

Digital Transformation

For those organizations with cognitive/AI among their digital transformation's strategic priorities, NVIDIA offers a rich set of technical capabilities. They are a design and engineering culture at their core, which makes them very good in their domain. Because they are so strong in what they do, it is easy to get enamored by the technology and forget that they are part of a solution set for enabling DX. We advise business users to understand their partner ecosystem, as it will be necessary to effectively translate their technical capabilities into a meaningful business process competency. The business needs to own the broader DX and cognitive/AI business strategies while using the NVIDIA partner to effectively craft a winning technical strategy in the organization's ecosystem against a clear set of tactical goals and longer-term capabilities.

Today, workers spend the great majority of their time hunting and gathering data from across the organization, which they then analyze to support business and operational decisions. IDC expects NVIDIA's deep learning platform to provide companies with agile methods and operations for exposing the data, categorizing and organizing data in a semantic data model, and automated data mining. This is a tremendous way to shift the people from gathering data to providing those higher-value insights.

To this end, one of the efforts NVIDIA showcased is a $70,000 desktop AI machine, which is short money since enterprises can connect it to their data sources and pilot data discovery within their organization using only one data scientist at the helm in program pilot mode. While this may seem like a high cost for a desktop unit, it replaces entire rows of rack space and the computing power of a much costlier datacenter investment.

IoT Considerations

Over the past two years, much of the IoT discussion has centered around compute at the edge of the network and how to create meaningful business outcomes once the IoT sensors generated the data. In many ways, this forced IT suppliers into creating powerful IoT gateways whose primary functions were to aggregate the data from the IoT sensors that were hanging off the gateways. This assumed that the sensors could never be powerful enough to perform any workloads that required significant compute with low power consumption. However, that model may be on the verge of meeting a challenge as IoT applications built on NVIDIA's Jetson microprocessor platform allow developers to write applications that can perform both meaningful analytics and artificial intelligence. This also creates new opportunities on the IoT connectivity/network model. By running IoT sensors on the Jetson platform, batched updates can be transmitted as needed and machine/deep learning updates can be applied to the IoT sensors. To date, this model has been very well received in IoT deployments in agriculture, factory floor monitoring, and even in external functions of a connected car. In addition, with the announcement of the NVIDIA cloud strategy, cloud service providers (CSPs) will quickly look to create new service for IoT devices connected at the edge of the network. In doing so, the CSPs will accelerate the software support model even more through API calls directly onto a Jetson-based sensor. Overall, this business model (i.e., combining compute, AI, analytics, and software over-the-air upgrades) could be very disruptive to current "hardware" IoT gateway/IT vendor solutions for specific IoT applications and use cases. Such a solution won't remove the need for an IoT gateway nor high-quality network connectivity (such as 5G), but it does enable higher-value IoT solutions to be economically created, thus spurring faster and wider deployments.

Finally, at a more strategic level, it is always important to build "systems" that are in balance – in this case, while not a true integrated IT system, NVIDIA has worked hard at making sure that latency is minimized both at edge of the network and within the enterprise. In doing so, IoT outcomes can quickly be generated and acted upon across the "system" thus ensuring the right actions happen at the right time.

Barriers to Deep Learning Continue to Drop with GPU-Accelerated SDKs and One-Stop Platforms

The challenge for most organizations in accessing and adopting deep learning applications is the barrier to entry due to the high technical expertise requirements, retraining effort of existing development resources, and time to market for initiatives. As announced at GTC 2017, GPU-accelerated deep learning SDKs and solutions aim to tackle this issue.

Various start-ups have jumped into the AI "Wild West" to provide a path for organizations to quickly apply and implement deep learning capabilities for innovation and automation. Credit is also given to NVIDIA's Inception Program that has provided the catalyst to bring these new technological innovations to the market in a very short time frame. Example vendor start-ups:

- **H20.ai:** H2O's open source AI platform is rapidly gaining traction as a deep learning and machine learning platform. It has garnered real-world uses cases in various industries ranging from energy to financials, while integrating with enterprise-scale Hadoop, Spark, and NoSQL and commercial technology such as SAP HANA and S3. H2O curates the best-of-breed open source technology and empowers users with a user-friendly interface and APIs to quickly extract insights from their data. The platform is further accelerated with NVIDIA's GPUs under the Deep Water umbrella.
- **Neurala:** Neurala's Brain for Bots SDK provides a real-time lightweight API for deep learning. It drastically lowers the barrier to entry for existing developers to adopt deep learning by providing user-friendly C++ APIs with the reduction of many lines of code. In short, while not providing the full flexibility of developing directly on the deep learning frameworks, it enables deep learning applications to be quickly developed. The SDK supports a broad range of cross-platform GPU enable systems, meaning it can be applied quickly for interference and intelligence at the edge.
- **Bitfusion:** Bitfusion's Flex platform seeks to provide an efficient end-to-end solution for deep learning by simplifying IT operations such as resource management and development/deployment. The solution comes prepackaged with portable containerization, GPU virtualization, deep learning frameworks and libraries to allow organizations to quickly kick-start and hit the ground running without worrying about the tedious tasks of setting up the environment.

Other notable GPU-enabled AI innovations include:

- SigOpt's optimized machine learning as a service, with a focus on the financial services industry (banking, insurance, and trading)
- Skymind's deep learning operating system, which serves as an enterprise-grade integrated platform for deep learning libraries (e.g., D4J) and connectors to big data components such as Spark and Hadoop

Adoption of deep learning applications in organizations will accelerate as the ease and speed of use improves at the current pace of innovations.

Planning for Tomorrow with the NVIDIA Inception Program

In June 2016, NVIDIA launched a global program to support the innovation and growth of start-ups that are driving new breakthroughs in artificial intelligence and data science. The NVIDIA Inception Program provides unique tools, resources, and opportunities to the waves of entrepreneurs starting new companies, so they can develop products and services with a first-mover advantage, through access to:

- Latest NVIDIA deep learning technologies – early access to the latest GPU hardware as well as the NVIDIA Deep Learning SDK and more
- NVIDIA's deep learning experts and world-class engineering teams
- NVIDIA's global network – customers, partners, and suppliers, and NVIDIA marketing reach
- Courses via the NVIDIA Deep Learning Institute to improve technical knowledge

- Funding – winners of the competition, announced at the GTC conference, received a combined $1.5 million funding through NVIDIA's GPU Ventures Program

Since the launch of the Inception Program, over 1,300 start-ups have applied to the program, from which NVIDIA presented 15 finalists ranging across various industries. A panel composed of six recognized individuals in the industry (including NVIDIA's founder Jensen Huang, investment gurus, and IT professionals) selected six start-ups across three categories to receive a combined $1.5 million funding:

- **Best social innovation AI start-up:** Over 100 companies were considered across industry verticals ranging from healthcare to agriculture to smart cities. Nominees for this category were focusing on cardiovascular imaging, chest pain triage imaging, drug discovery medical data analysis and interpretation, and aging research.
  - The winner, Genetesis, is a medical device company that has developed CardioFlux to reinvent how emergency rooms diagnose chest pains. The runner-up, Bay Labs, aims at fighting heart disease by making ultrasound scanners inexpensive and generally available.
- **Most disruptive AI start-up:** Over 250 companies were considered across industry verticals ranging from manufacturing to insurance to agriculture. Nominees focused on geospatial imagery for real estate, cybersecurity, customer service, railway systems, and engineering and construction.
  - The winner is an Israeli preventive security start-up DeepInstinct; it uses deep learning to detect malware in real time. The runner-up, Smartvid.io, aims at making construction work safer.
- **Hottest emerging AI start-up:** Over 600 companies were considered across industry verticals ranging from retail to healthcare to automotive. Nominees aspire to bring innovation around "at-home" blood analysis, data governance workflows, phone calls and video footage analytics, and brick-and-mortar retail stock management.
  - The winning start-up, Athelas, created a portable device that lets users get an analysis of their blood anytime and anywhere. The runner-up, Focal Systems, helps shoppers find products, spot sales, and pay for groceries without stopping at the checkout counter of grocery stores.

The Inception Program enables NVIDIA to widen its market reach. By providing high-end computing technologies, the vendor risks limiting the reach of its products to its base clientele: gamers and large enterprises with high compute needs. Opening its platform to start-ups will make the products more accessible for organizations regardless of their sizes.

In an era where cognitive and deep learning are slowly being adopted by organizations as a competitive or a disruptive tool, having a strong mindshare in the market will be critical for success. Making the platform available to the new generation and having them trained on the platform will create strong advocates for their brand and systems. This is the case especially since most end users have to go through NVIDIA's partner ecosystem to customize their solutions. Having those types of advocates will make the difference between using NVIDIA or a competitor's products.

Accelerating Virtual Computing via NVIDIA GRID: The Honda Use Case

One of the lesser known capabilities of NVIDIA GPUs is their ability to significantly contribute to a virtual computing environment using what is known as virtual desktops. Specifically, NVIDIA's GRID vGPU is a graphics acceleration technology that allows a single GPU to be shared among multiple virtual desktops. When NVIDIA GRID cards (installed in an Intel x86 host) are used in a desktop

virtualization solution running on VMware vSphere 6.0, graphics can be rendered with superior performance compared with non-hardware-accelerated environments. This capability is useful for graphics-intensive use cases such as designers in a manufacturing setting and architects as well as power users who need access to rich 2D and 3D graphical interfaces.

At this year's conference, one of the GRID topics was "GPU Accelerated VDI for Car Design Environments." Dassault Systèmes, HPE, NVIDIA, and VMware announced the support of 3D experience NVIDIA GRID Tesla M60 in their design environment. IDC has learned that Honda has adopted vGPU technology in its car design activities using CAD to solve several challenges, specifically around environmental issues and resource optimization issues that the company was facing.

Honda's datacenters were having issues with the heat, power, and floor space for 1,000 units of physical engineering workstations (EWSs) in their facilities. With the help of NVIDIA GRID, the datacenters were able to reduce that to seven racks of 1,000 virtual machines (VMs). This was equivalent to an area reduction of 94%, which enabled them to achieve the objective of EWS datacenter transition. Because of this, car designers at Honda were released from heat problems, noise problems, maintenance workloads, data loss risk, and fixed CAD environments. However, the designers at Honda still had a challenge since this did not provide for flexible resources management on virtual desktop engineering workstations (VDI EWS) by GPU pass-through since VM:GPU are assigned as 1:1.

Therefore, to solve this second challenge, resource optimization has been done with virtual desktop integration (VDI) resource consolidation and flexible adaptation of VM/GPU specifications depending on the need of each CAD user. By adopting NVIDIA Tesla M60, Honda was able to achieve up to 200% of performance of a standard engineering workstation, provide one-day reconstruction of resource flexibility, and a 20–40% improvement in resource efficiency.

Now, 4,000 car designers at Honda are using CAD (CATIA) in a mixed real/virtual environment (physical EWS, VDI EWS by GPU pass-through and VDI by GRID vGPU on VMware Horizon View). IDC believes that this is an excellent use case that CAD users in manufacturing, BIM (Building Information Modeling) users in construction, and CIM (Construction Information Modeling) users in public/engineering should consider.

Finally, IDC believes that NVIDIA is focusing on not only the aforementioned area for virtual computing but also all Windows 10 virtual desktop (VDI) environments. Since Windows 10 uses a lot of graphic resources, NVIDIA believes that GPU computing in a VDI environment will be able to offload significant amounts of CPU resources.

The Client View

GTC 2017 was a relatively quiet show for NVIDIA on the client side of the business. NVIDIA unveiled major product announcements at GDC only a couple of months prior, as well at CES at the beginning of the year, electing to focus its own show on its work in AI. Still, two key observations on the company can be made from the client lens.

The first observation applies to the broader PC industry. Over the past few years, traditional titans of the PC industry have been ramping up their transformation stories. The starting point and destination of this journey is different for each company, but a similar plotline permeates all: strengthen efficiencies in the old PC business to catalyze growth in newer ones. Top PC OEMs have increasingly

pivoted to become more full-fledged service providers, while Microsoft builds up to a full sprint in the cloud. Intel has been reorganizing to shift its focus on datacenters and even cancelled IDF earlier this year.

The trend of large-scale corporate transformation in the PC industry has been necessitated by growth of the 3d Platform. Empires that were built in the PC domain in the era of the 2nd Platform must reassess and adapt to the new IT world, which deemphasizes PCs in favor of mobile, cloud, big data, and social technologies. NVIDIA built an esteemed brand by pushing outward the boundaries of the PC experience. GTC 2017 showed that pushing boundaries is still in its corporate DNA, but like with the broader PC industry, the PC is getting increasingly deemphasized in the story.

This leads to the second observation. As aforementioned, the destination of transformation varies for all companies. Each company must decide what position it wants to carve out in a 3rd Platform-dominated world and where to build inroads. At the show, NVIDIA made it clear where it wants to be and how it wants to get there. NVIDIA sees a future where accelerated computing and AI converge in several practical applications ranging from automotive to robotics to IoT. The company sees the same future being driven by doing what it does best: pushing boundaries outward. NVIDIA helped shape the PC industry in the past two decades by focusing on accelerated computing, and it believes that it can drastically shape the broader IT industry by transferring that functional knowledge into new areas. In each new area, NVIDIA's destination is clear: be at the leading edge of performance.

Consequently, the absence of major client presence on the mainstage at GTC 2017 could be summed up as such: the DNA hasn't changed, but the eyes are looking strictly ahead and not back.

The Datacenter and Cloud Perspective

Arguably, the datacenter as we know it may well be disrupted and transformed by what IDC calls "accelerated compute" – infrastructure that has been beefed up with GPGPUs, FPGAs, ASICs, or Intel Phi, among others. While the use of GPUs from NVIDIA and AMD as well as Intel Phi have been common in specialized high-performance computing, the compute-intensive nature of today's enterprise workloads is increasingly causing performance issues that can no longer be resolved by adding an extra box to the cluster, upgrading to a stronger CPU, or bringing in Flash. Especially massive parallel tasks don't perform very well on CPUs, which are much better at linear processing.

While cognitive/AI workloads, not just the training component but inferencing as well, demand such parallel processing, driving the shift to acceleration technologies such as GPUs into the datacenter (or the cloud, for that matter), IDC is seeing other workloads in the datacenter hitting infrastructure performance limitations as well. Media streaming, web serving, compute (such as video transcoding), structured data management and analytics, security, business applications, and even collaboration and application development are all expected to increasingly require accelerated infrastructure. Businesses that are early adopters in this space expect anywhere from 5% to 20% of these common workloads to run on accelerated compute in the next few years.

In this emerging reshuffle of the stakes between Intel and alternative processor manufacturers, NVIDIA is, by all accounts, the dominant (and most eye-catching) player. Not just dominant, but rising and apparently unstoppable. Much of that is because of a combination of smart strategy plus a little luck. As a graphics processor unit developer, high-performance computing, then cognitive/AI, and other data-intensive workloads more or less fell into NVIDIA's lap. But as GTC 2017 demonstrated, the company didn't waste a millisecond to subsequently build on this opportunity, investing billions of dollars into new technologies from processors to operating system to deep learning frameworks.

At GTC 2017, NVIDIA announced the Tesla V100 based on the GV100 – or "Volta" – GPU, which by now most readers will have been introduced to (for product details, see **www.nvidia.com**), and Tesla V100 represents a dramatic performance increase compared with the Tesla P100 based on the GP100 – or "Pascal." Volta is equipped with tensor cores, which allow each of the 5,120 cores to run multiple operations in parallel, making the GPU act much more like a neural network rather than a, well, GPU (NVIDIA's CEO, Jensen Huang, quipped that he had been toying with new acronyms for the processor, rattling off multiple improbable three-letter combinations ending with "PU"). According to NVIDIA, the Tesla V100 will be available in 3Q17, first as part of NVIDIA's DGX-1 V, a 4U server that holds eight V100s plus two CPUs for bootstrapping. The interconnect – either NVLink or PCIe – enables 800GBps bandwidth.

The DGX-1 was already described as a monster by many observers, but the V100-based DGX-1 V is said to contain the equivalent of hundreds of CPU-based servers. It is aimed at the most demanding deep learning tasks, in terms of both training and inferencing. Also announced was the DGX Station, a liquid-cooled workstation version of the DGX-1 V with four V100 processors, that is designed to help developers achieve greatly reduced training times compared with Pascal-based systems. Plus, the company announced the NVIDIA GPU Cloud (NGC), which is a bit of a misnomer since it's not a cloud – rather, it's a cloud-based platform that aims to make deep learning software even easier with a preintegrated, optimized container approach for the entire deep learning stack allowing developers to easily deploy deep learning workloads on the desktop (DGX Station), in the datacenter (DGX-1), or in the cloud (AWS, Azure, etc.).

NVIDIA also announced that it is working together with SAP to infuse SAP's enterprise software with cognitive/AI that runs on GPU-based systems, including DGX-1. The first product from the collaboration will be SAP Brand Impact, which uses NVIDIA deep learning and measures brand attributes such as logos in near real time. SAP is also training its accounts payable application to understand and process invoices without human support. Finally, NVIDIA has been extremely proactive at partnering with cloud providers to enable them to offer GPU-based instances. At GTC 2017, NVIDIA announced that Microsoft Azure had been added and that today NVIDIA GPU-based instances can be rented on any of the world's top cloud providers.

It should be noted that in the months leading up to GTC 2017, a slew of announcements had already indicated that NVIDIA is capitalizing on the accelerated compute opportunity at neck-breaking speed. IBM Cloud started to offer Pascal-based computing. Fujitsu announced it will build a supercomputer based on 24 DGX-1 servers, The Tesla P100 became available on Google Cloud Platform for Google Compute Engine and Google Cloud Machine Learning users. AWS started offering NVIDIA-based GPU instances. IBM launched its Power System S822LC for high-performance computing powered by NVIDIA Tesla P100 GPUs and NVLink. And various other server OEMs and ODMs launched new servers with NVIDIA GPUs built in.

The momentum with which GPUs are entering the datacenter appears to be significant. Significant enough that some observers have started to try and calculate the impact this phenomenon may have on Intel. But Intel, too, aims to take a share in this market for accelerated compute with Nervana's accelerator Lake Crest, and with Phi. AMD is not sitting still, as are a host of processor start-ups such as Graphcore, which NVIDIA said at GTC 2017 they are keeping a close watch on. The bottom line seems to be that 2017 may become the beginning of the end of the Intel CPU hegemony in the datacenter. GTC 2017 had the feel of a datacenter revolution about to happen.

ADVICE FOR THE TECHNOLOGY BUYER

Cognitive/AI systems present significant challenges and opportunities, in part, because the collective body of technologies is evolving so rapidly. This not only makes selection and adoption difficult but also presents a competitive advantage to those that master that competency. Some cognitive systems technologies are ready for prime time, and others should be monitored and/or vetted in small, low-risk pilots. Most important for CEOs, CIOs, and CTOs is to begin building a cognitive systems business strategy now that considers the technologies that are commercially available.

The key to deep learning technologies is that the algorithms self-program based on the data provided and significant compute resources, typically based on GPUs today. Using these technologies, organizations can develop solutions that tune themselves automatically at a speed that human programmers cannot hope to emulate. In addition, since it can be based strictly on data, the quality of solution is improved as more data is provided. The inclusion of deep learning capabilities into consumer and enterprise applications and solutions will make it possible for an organization to become more responsive to changes in the market.

Organizations should already be deploying or experimenting with deep learning technologies to embark upon digital transformation and augment or improve processes using deep learning. To do this, the organization needs large amounts of well-curated and accurate data as fuel for these deep learning algorithms. Identifying, locating, and organizing this data is the first step in making use of deep learning tools such as NVIDIA offers. The second step is to make sure that the organization's staff has the education and training needed to plan and develop deep learning solutions. NVIDIA's Deep Learning Institute makes this easier for organizations to accomplish.

Deep learning is here to stay and is a crucial component of artificial intelligence for the foreseeable future. NVIDIA intends to be a major force in artificial intelligence and is doing everything in its power to make it accessible and usable by every enterprise.

As to the hardware, buyers have many choices for procuring accelerated compute for cognitive/AI and other workloads. Most businesses seem to prefer procuring servers with accelerators built in, which are available from all the large, brand-name OEMs as well as several ODMs. Many businesses also prefer procuring from their main server vendor, but it can pay off to shop around and compare. Some important considerations are:

- **Type of acceleration technology.** The type of acceleration technology (GPU, FPGA, ASIC, Phi, etc.) required is important as each has a different purpose, with FPGAs being more programmable for specific tasks and ASICs being even more specialized, while GPUs and Phi can run general workloads.
- **Core performance of the CPU.** GPU-enabled systems come with various types of processors, including AMD, ARM, IBM Power, and Intel, with varying core performance metrics.
- **Vendor of the acceleration technology.** Brand awareness of acceleration technologies is still nascent, but road maps matter as does the vendor's focus on the entire stack, from processor to operating environment to frameworks and other software.
- **Total wattage (energy use and cooling technology).** Total wattage starts to matter when the footprint of accelerated compute in the datacenter increases.
- **Ease and speed of deployment.** Ease and speed of deployment is a major consideration in this rapidly evolving space where developers are clamoring to get the right hardware fast to build new applications.

- **Capex and opex.** Capex and opex always matter for servers, but currently many businesses appear to be eager to get into accelerated compute, with cost being secondary.
- **Included software.** Some vendors offer open source software, including deep learning frameworks, packaged with their accelerated server platform.
- **Service and support.** Some vendors will go very far to help their customers with getting into accelerated compute, especially for cognitive/AI.

LEARN MORE

Related Research

- *Executive Guide to Assessing Tangible and Intangible Impacts of Cognitive Computing and Artificial Intelligence* (IDC #US42348117, March 2017)
- *Worldwide Cognitive Server Infrastructure Forecast, 2016-2021* (IDC #US42294414, February 2017)
- *Japan Virtual Client Computing 2017 Top 10 Predictions* (IDC #JPJ41772916, January 2017)
- *Market Analysis Perspective: Worldwide Cognitive Systems and Content Analytics Software, 2016* (IDC #US40797116, September 2016)
- *Japan Virtual Client Computing Market Shares, 2015: Prologue of Virtual Client Computing Fourth Generation* (IDC #JPE40933216, August 2016)
- *Worldwide Cognitive Systems, Content Analytics, and Discovery Software Forecast, 2016-2020* (IDC #US40305316, June 2016)
- *IDC PlanScape: Implementation of Cognitive Systems* (IDC #US41477516, June 2016)
- *IDC PeerScape: Digital Transformation – Practices for Strategically Leveraging Cognitive Systems* (IDC #US41191916, April 2016)
- *IDC TechScape: Cognitive Systems Technologies, 2016* (IDC #US41005816, February 2016)

Synopsis

This IDC Perspective examines and provides an overview of the NVIDIA 2017 GPU Technology Conference held in San Jose, California, which highlighted a range of new cognitive/AI and deep learning platforms and services by NVIDIA and its partners to help companies develop AI-enabled solutions and tools.

According to Dave Schubmehl, research director and lead analyst for IDC's Cognitive Systems and Content Analytics research, "Intelligent applications based on cognitive computing, artificial intelligence, and deep learning are the next wave of technology transforming how consumers and enterprises work, learn, and play. NVIDIA's 2017 GPU Technology Conference highlighted the ways that deep learning is being used to solve an ever-wider range of challenges. Going forward, deep learning will be an integral component of almost every enterprise and consumer application and will affect everyone daily."

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