IDC
Analyze the Future

IDC TECHNOLOGY SPOTLIGHT

Accelerated Workstation: Run Deep Learning Workloads at Your Desk

December 2017

Adapted from *IDC's Worldwide Accelerated Compute Taxonomy, 2017* by Peter Rutten and Ashish Nadkarni, IDC #US42878517*, and Worldwide Accelerated Server Infrastructure Forecast, 2017-2021* by Peter Rutten, IDC #US42998817

Sponsored by NVIDIA

This Technology Spotlight showcases the need for desktop-based "supercomputers" such as NVIDIA's DGX Station – a graphics processing unit (GPU)-based accelerated workstation designed for computationally intensive deep learning workloads.

Introduction

While GPU technology has been around for some time, the ever-increasing computational requirements for running deep learning workloads are driving the need for advanced workstations and servers that make use of hardware accelerators to turbocharge the parallelized algorithms that are core to these workloads. Over time, increasing the return on investment for firms – which are measured not only in personnel costs but also as a derivative of increased scientific breakthroughs – will no doubt require the right IT infrastructure that shortens job runs with on-demand, high-performance computing that improves professional productivity.

Researchers, developers, and creative professionals at enterprises all face the same hurdle: They are handcuffed by IT resources that are made available to them to perform leading-edge research. Artificial intelligence, deep learning, and data analytics are just a few of the examples that require accelerated computing resources, which are not always readily available to these professionals. This often requires them to wait in a queue for compute cycles on a datacenter server. NVIDIA recently introduced the DGX Station, a type of workstation specifically with such professionals in mind. It is designed to be placed in the professional's office and used as a deskside-based accelerated computing system that utilizes GPUs to perform data-intensive analysis at a far greater speed than most similarly configured traditional-based workstations.

Workstations: Purpose Built for Specific Workloads

A workstation is a server-grade computer that is designed primarily for individual use, with limited shared resource capabilities. Workstations are computers in which the hardware and software is designed to work in harmony (i.e., the hardware and software are often certified to be compatible with each other). With workstations, buyers expect a high level of hardware and software integration, reliability, and a minimal amount of IT intervention, much of it being nondisruptive to the workload being executed.

Workstations allow an individual to have complete control over a specific set of compute-intensive applications and are optimized specifically for such workloads. In general, workstations are designed to be turnkey systems working with specific application systems. They are intended for users who do not have the time to optimize their software code, nor do they have the expertise to troubleshoot hardware-software incompatibility, and therefore expect their workstation to run in a reliable manner.

Workstations are designed for desk use but, unlike many inexpensive desktops and personal computers intended for a single user, are equipped with high-end features designed for performance, availability, and reliability. Workstations are not designed to replace high-end servers that are specifically designed with a datacenter-wide shared use in mind but can offer a cost-effective middle ground with levels of performance traditionally only available in a server form factor.

Market Trends: The Adoption of Accelerated Workstations

Accelerated computing is the ability to accelerate applications by offloading a portion of the processing onto adjacent silicon subsystems such as GPUs. Accelerated computing is gaining traction in enterprises, as businesses seek solutions for overcoming the limitations of central processing units (CPUs) for workloads that require data processing acceleration. Accelerated computing is increasingly used to reduce the "time to value" for data science workloads including analytics, cognitive, deep learning, and machine learning applications.

Accelerated computing is a fast-growing market segment in the worldwide server market. It is growing far quicker than the overall single-digit growth forecast for the server market out to 2021, according to IDC. IDC is forecasting the worldwide accelerated computing infrastructure revenue will grow from $2.5 billion in 2016 to $6.8 billion in 2021, a CAGR of 21.9%. The on-premises portion of this market is forecast to grow from $1.6 billion in 2016 to $3.4 billion in 2021, at a CAGR of 16.3%. IDC expects an increasing portion of this market to be served by new form factors, including desktops and workstations.

Major scientific and technological breakthroughs have been made in recent years with the use of supercomputers in either an academic or enterprise setting. In higher education and research, getting time on these machines requires not only being affiliated with a university or government institution but the ability to get your project approved by a committee. In a private enterprise setting, IT must have GPU-accelerated servers in the datacenter or secure cloud-hosted capacity co-resident with their data sets.

Datacenter infrastructure can cost hundreds of thousands to millions of dollars, and cloud-hosted compute costs can escalate as developers iterate repeatedly on their models during the experimentation phase. Deep learning and artificial intelligence (AI) are necessitating the reimagining of the datacenter, since the CPU-only technology of yesterday is not keeping pace with the IT demands of today. However, many organizations are just now discovering the importance of GPU-accelerated computing and are only beginning the effort to update their datacenters accordingly.

Accelerated computing is becoming very attractive to organizations that want increased performance while not increasing their budget. One trend that is just beginning to be seen is providing GPU-based systems at a price below the $100,000 level that can be located at a researcher's or developer's office. IDC expects this trend to continue. AI practitioners want to

control their own workflow and not be dependent on machine time being available on a centralized, shared resource, especially as an organization first sets out in the "productive experimentation" phase of the AI development. They want a system that is fast to install, is easy to manage, and does not require a high learning curve to become proficient in its operation. This dream is now a reality, and once IT departments realize the benefits of a GPU-based system, more researchers and developers will demand the use of this equipment. The previously mentioned forecast of on-premises accelerating computing supports this scenario.

AI practitioners are aware of the benefits provided by GPU-accelerated servers and workstations, and IT organizations have begun to include it in their infrastructure for specific workloads.

NVIDIA and the Growth of Accelerating Computing

NVIDIA was founded in 1993 by Jen-Hsun Huang, Chris Malachowsky, and Curtis Priem. Since its founding, it has provided disruptive technology products to the computer industry competing with companies like Advanced Micro Devices and Intel. The firm designs and manufactures computer graphic processors for consumers and enterprises, which NVIDIA invented in 1999. In 2006, NVIDIA released Compute Unified Device Architecture (CUDA), which is a parallel computing platform and programming model to work with the GPU processors NVIDIA invented. Parallel computing improves the speed in which complex compute-intensive problems can be solved.

At first, NVIDIA's target markets included gaming and professional markets such as workstations along with systems on a chip (SoCs) targeted at the automotive and mobile computing industries. From 2014 forward, as the company has grown, it has increased the breadth of its primary market to include professional visualization, gaming, virtual reality, datacenters, and the automotive industry.

As the computer industry continued to evolve, so too has NVIDIA, with a notable key difference: GPU's have now outstripped the pace of Moore's law scaling constraining the CPU marketplace. The massively parallel architecture of the GPU with its thousands of cores working in parallel allows complex problems to be solved in a shorter period by offloading the CPU and allowing each processor to do what it does best. This has enabled GPU technology to become the de facto platform for the computationally intensive, parallelized nature of the calculations employed in AI and deep learning. GPUs are used by engineers, scientists, and researchers in support of mission-critical projects and commercial enterprise applications. NVIDIA is now totally immersed in AI, accelerated analytics, and deep learning, to name a few.

The compute demands for AI and deep learning are immense, and this will only increase as time goes on. In April 2016, NVIDIA introduced the DGX-1 server. This is an eight-GPU configuration targeted at deep learning and accelerated analytics, offering an integrated platform that combines the next-generation Volta architecture GPU – NVIDIA Tesla V100 – and popular deep learning frameworks, each optimized for maximum GPU-accelerated performance. The platform offers a total of 40,960 CUDA cores with a total of 128GB of total GPU memory and delivers 1 petaflops of performance.

Considering Accelerated Workstations: The NVIDIA DGX Station

NVIDIA has also introduced the DGX Station – designed to be used by a single user or as a workgroup server supporting a small team of users, usually scientists, researchers, or AI developers that require dedicated and on-demand access to high-performance accelerated

compute resources for a specific set of cognitive, AI, and deep learning workloads that exceed the computational capacity of a traditional workstation or CPU server.

The DGX Station represents an economical midway point to acquire performance-optimized accelerated compute workstations at half the price and half the performance of its server form factor sibling (the DGX-1). It is ergonomically designed to be used in an office, as a professional-grade workstation.

An additional use case for DGX Station is its role as a workgroup server. While some organizations will have individual developers who can sufficiently consume the entire compute capacity available in a dedicated AI workstation, some environments will have multiple team members, each running their own experiments in a workgroup setting either simultaneously or at varying schedules. DGX Station can provide a more-than-adequate solution that can multiplex its GPU computing power to several users, thereby improving the utilization and overall economic benefit of the platform.

The DGX Station is designed to be deployed in an individual's office or in a lab. It is coined as "the world's first personal supercomputer for leading-edge AI development." The DGX Station is built on the same NVIDIA GPU Cloud Deep Learning stack that is incorporated into all DGX systems. As such, it has a single unified stack for deep learning frameworks with predictable execution across platforms. A user can seamlessly transfer deep learning models if necessary to the DGX-1 server located in the datacenter or to the NVIDIA GPU cloud and then back to the DGX Station without any need for recompilation of code or other software concern. This workstation incorporates an integrated hardware and software solution that allows the system to operate with greater reliability and predictability, a feature a scientist who is not an IT professional appreciates. These features allow researchers to spend more time doing their real job and less time playing "systems integrator" or trying to troubleshoot or fine-tune a system to their workload requirements.

DGX Station is based on the NVIDIA Volta GPU, incorporating four Tesla V100 GPUs offering 20,480 CUDA cores and 500 teraflops of AI performance. It comes configured with 256GB of RDIMM DDR4 system memory, 3 x 1.92TB SSDs for data storage configured in a RAID 0, and support for up to 3 video displays with 4K resolution, with a peak power consumption of 1500W, which is dramatically less than the equivalent power draw of 400 x86 CPUs, which, according to NVIDIA, is the equivalent of what DGX Station offers in terms of AI computing power. It's worthwhile to note that this is the world's first workstation leveraging NVIDIA's NVLink GPU-interconnect technology. NVLink delivers ultra-high bandwidth (aggregate 200GBps) between GPUs within the Station, with very low latency. This is critical for ensuring deep learning jobs can scale across multiple GPUs with better linearity of performance than GPUs that are connected via standard PCIe bus.

Like many traditional workstations, the DGX Station is designed to operate with limited noise. At first, this might be viewed as a feature that many would place at the bottom of a feature list. However, for end users — and data scientists are no different in this regard — this is one of the top sought-after features. Namely, the ability to seamlessly insert the platform into a normal workflow and have it used as a natural extension of the developer/researcher, where they're most comfortable. In addition, many traditional workstations are air cooled, which can create an uncomfortable operating environment when the system is deployed close to the user. This issue becomes more pronounced when several workstations are used by a single user to operate high-performance workloads. The DGX Station is water cooled, as most GPU systems operate at a

higher temperature. This feature alleviates the temperature concern and creates a much quieter, comfortable operating environment.

From a deployment perspective, the DGX Station uses a standard 115–240 VAC outlet and can draw up to 1500W. The workstation can therefore be deployed quickly and simply enabling a plug-and-play setup from power-on to working on computationally intensive projects, in just a few hours. In addition, important to any scientist or developer is the access and support offered by NVIDIA along with the latest software updates. It is now easier for data scientists to be more productive, work on larger data sets in real time, and find actionable results that can help justify the ROI on the solution.

Challenges and Opportunities for NVIDIA

Accelerated computing is rapidly gaining traction in both the private sector and the public sector as organizations embrace these technologies to overcome the limitations of CPUs. NVIDIA has been at the forefront of GPU technology and innovation for over 10 years now. It has been an uphill battle in selling the technology but now that CPU speed increases are slowing down, IDC believes that NVIDIA is in a prime position to sell its GPU technology to users that are influencing the IDC 3rd Platform era and beyond. With AI and deep leaning workload requirements now being hindered by systems just having CPU technology, NVIDIA provides a valid story to break this roadblock.

IDC believes that NVIDIA's message to CIOs and IT managers – that a GPU-based solution is superior to a CPU-only solution from a performance perspective – has started to resonate with its IT buyers and that in turn is increasing the adoption of accelerated technologies for computationally intensive workloads. IDC believes that NVIDIA's message will be further strengthened when it is backed up with ROI results that incorporate worker productivity in addition to IT infrastructure.

IT Infrastructure

IT managers generally like to standardize on a minimum set of machine architectures to reduce downtime and maximize staff efficiency. They do not want employees within a company to own unique or one-off servers, workstations, and desktops. They look for solutions that can reduce bottom-line expenditures for hardware and software. Typically, such decisions are not taken with workloads and applications in mind. This is a major roadblock that scientists, researchers, or any users that have compute-intensive workloads face each day. They are concerned with their own productively by having a machine that can run such workloads on demand and in a timely manner. They require a machine that provides high performance, reliability, and business agility. The key to achieving harmony between IT and end users is the awareness of various options available in the market. The opportunity for NVIDIA is to help end users to make a business case for the DGX Station to their IT department, and accordingly, it would benefit IT managers in knowing how standardizing on the DGX Station would help its end users.

Talent Retention

One message heard repeatedly is the desire of scientists to do the job they were hired for – which is to provide beneficial research results to the company and to society in a timely manner. They were not hired to be systems integrators, programmers, or IT administrators! Business leaders need to view the DGX Station as a means to attract and retain the best talent, especially in a

domain where professionals with valuable AI and deep learning experience are so few in number and desire to do their life's work on the best tools available.

Ambitious organizations that seek to build a data science team that will take them into the future will carefully consider the importance of putting tools such as DGX Station in the hands of these highly compensated professionals. As with DGX-1, these kinds of assets can differentiate organizations competing for a limited pool of talent.

Optimizing End-User Productivity

Procuring and maintaining the IT infrastructure is always a challenge for any enterprise. IT managers are fully aware of the compromises that must be made due to budget constraints, which then impact equipment support decisions. The DGX Station allows the user to be operational in hours, not days or weeks, once receiving the server. It provides one of the fasted paths to deep learning as the system enables the user to be more productive, running experiments instead of managing infrastructure or engineering a software stack. For some organizations, the software engineering costs alone can amount to hundreds of thousands of dollars in operating expenses annually, in contrast with the pre-optimized, ready-to-use, integrated stack offered on DGX Station.

The challenge for leaders who envision AI enabling their business is to convince their IT managers that the DGX Station should be part of an onboarding plan that attracts and retains the best AI talent but also saves money through increased productivity for researchers and developers. Deep learning requires uncompromising system performance and reliability and, crucially, the ability to run the workload in hours, not days.

The Build-Versus-Buy Decision

Some organizations that are accustomed to using consumer-grade GPUs (like those NVIDIA offers for gaming) will consider mimicking the DGX Station by building their own AI workstations leveraging commonly available hardware. While this might seem more economical from an initial capital outlay perspective, such efforts should be weighed against the merits of DGX Station, along with total cost of ownership view, including:

- The benefit of having a standard, "solutionized" platform that's already integrated and offers a simple procurement approach and turn-up approach
- The cost of taking on the systems integrator role in terms of lost time and productivity spent sourcing componentry and software from multiple places
- The software engineering effort required to adapt deep learning frameworks, along with supporting drivers, libraries, and primitives to achieve optimal performance
- The cost of lost time and productivity associated with troubleshooting and isolating system issues
- The benefit of having a single point of contact for enterprise-grade customer support

For many mission-critical environments, the total life-cycle costs associated with being a systems integrator, software engineer, and support desk will quickly eclipse the initial capital savings that come with commodity/consumer-grade hardware, especially when weighed against the lost productivity of valued researchers, developers, and innovators.

Conclusion

IDC believes the growth of hardware accelerators such as GPUs will continue to outpace the overall server market. While accelerated systems were historically limited to very large supercomputers, their form factors and acquisition costs are dropping to the point where high performance and enterprise-grade reliability can now be found in an integrated system that can fit in a developer's office for an MSRP of $69,000. As NVIDIA seeks to create this new market segment, the success of DGX Station will reside in its ability to help businesses see the benefit of solutionized GPU technology. Business leaders will need to embrace the value of saving their scientists and developers from losing productivity on efforts spent integrating systems, engineering software, and performing IT support.

Global Headquarters: 5 Speen Street, Framingham, MA 01701 USA P.508.872.8200 F.508.935.4015 idc.com