IDC ANALYZE THE FUTURE

White Paper

Why Developing and Deploying AI on Workstations Makes Sense

Sponsored by: Dell Technologies

Dave McCarthy Peter Rutten
February 2021

IDC OPINION

AI has taken off as an important, differentiating capability in all industries, and the hardware required to run AI is rapidly evolving. The technology industry is often very focused on the exponential growth in size that the most advanced AI models are going through. The discussions are about tens of billions of parameters, reducing precision, expanding memory, high-performance computing (HPC)-like needs for AI training and inferencing, and racks of accelerated servers. In reality, this extraordinary scale of AI computing is the exception, especially in the enterprise.

Today many businesses are working hard on AI initiatives that do not require a supercomputer. Indeed, a lot of AI development — and increasingly AI deployment, notably at the edge — is actually taking place on powerful workstations. Workstations have numerous advantages for AI development and deployment. They liberate the AI scientist or developer from having to negotiate server time, they provide GPU acceleration even as server-based GPUs are still not easily available in the datacenter, they are extremely affordable vis-à-vis servers, and they represent a smaller, one-time expense rather than a rapidly accumulating bill for a cloud instance. That way, they also free the scientists or developer from the anxiety of racking up costs while merely experimenting on AI models.

IDC is seeing the edge grow faster than on premises or cloud as an AI deployment scenario. Here too, workstations play an increasingly vital role as AI inferencing platforms, often not even requiring GPUs but performing inferencing on software-optimized CPUs. The use cases for AI inferencing at the edge on workstations are growing rapidly and include AIOps, disaster response, radiology, oil and gas exploration, land management, telehealth, traffic management, manufacturing plant monitoring, and drones.

This white paper looks at the increasing role that workstations play in AI development and deployment and briefly discusses Dell's portfolio of workstations for AI.

SITUATION OVERVIEW

The AI Explosion and the Infrastructure Impact

The number of AI projects that organizations worldwide are engaged in is growing rapidly. Already, across all industries, many tasks are performed by software that is partially or entirely driven by an AI model. IDC tracks AI on many levels, and one metric that is useful to consider is the amount that businesses and cloud service providers are forecast to spend on servers to develop and run AI. By 2024, this will be $23.9 billion, representing more than 20% of the total worldwide server spend. It is reasonable to assume that the workstation market will have a similar AI penetration, which would mean that the worldwide AI workstation market will be approximately $2.5 billion by 2024.

But servers do not make up the entire picture. A lot of AI preparation, development, prototyping and, increasingly, *deployment* is happening on workstations. As organizations, small and large, discover that new business opportunities can be realized by infusing their applications with some amount of AI functionality, experimentation with AI models has skyrocketed, and robust workstations are ideal for this purpose.

How did AI suddenly become so prevalent, given that AI algorithms have been deployed for decades? That's primarily because two quintessential conditions for powering a specifically successful type of AI algorithm, the neural network, have been realized in the past few years: the easy availability of vast, cheap, and diverse types of data, such as unstructured and semistructured data, and the augmentation of linear compute with a parallel model to process those neural networks within an acceptable time frame. With those two basic conditions met, data scientists have made tremendous advances with developing neural networks that automatically learn how to execute increasingly impressive tasks. While traditional machine learning (ML) remains relevant for textual or numeric data, deep learning (DL) is more effective for video, audio, languages, and so on.

Traditional machine learning models can typically be developed on the CPUs of a workstation, which have at most several dozen cores, but neural networks require coprocessors to parallelize their processing across thousands of cores. The main reason for this is: in ML, feature extraction and classification is a manual process, while that in DL is automated requiring the model to be trained through constant repetition using large data sets. Currently, the most common coprocessor is the GPU, but new AI-specific processors developed by start-ups are becoming available as well. This type of acceleration, using a discrete coprocessor for parallel processing, has revolutionized the server and workstation markets, giving rise to what IDC calls massively parallel compute.

In 2020, accelerated servers constituted a $12 billion worldwide market, growing to $28 billion by 2024, with 63% of that total representing accelerated servers for running AI. The number of workstations grew from 3.9 million to 5.6 million from 2016 to 2019, while the number of GPUs sold for use in workstations grew to 6 million in 2019; in other words, workstations have a nearly one-to-one GPU attach rate for all workloads. IDC projects that the market for workstations used for scientific or software engineering purposes, which are increasingly driven by AI development, will increase from $1.2 billion in 2020 to $1.5 billion in 2024.

AI Development Stages

As mentioned previously, neural networks became feasible because of expanding data types and volumes and new approaches to compute. The first part of this equation, data volumes and types, is not a trivial portion — by some accounts, as much as 80% of the effort in a deep learning AI initiative goes into data management and preparation. Data needs to be ingested, managed, and prepared before model design and training can commence. According to IDC, the following are the AI development stages (see Figure 1):

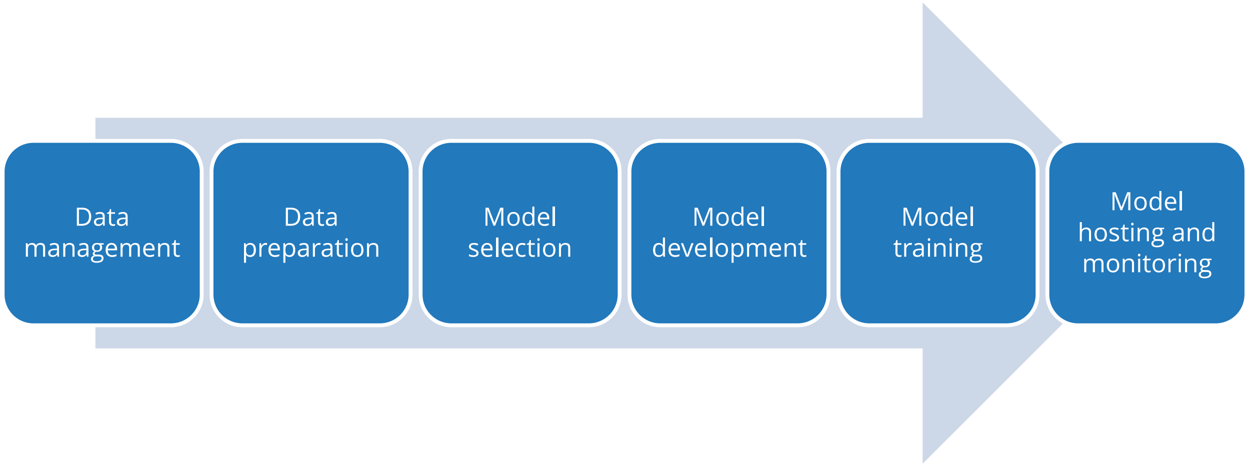
- **Data management:** Identifying and managing relevant data for the AI model from the vast volumes of data across the datacenter, the edge, and the cloud that an organization ingests, generates, and/or acquires (This data can be of any type, event driven or streaming, and much of it may require some type of governance.)

- **Data preparation:** Storing of data (file, block, or object) in a data warehouse or data lake, cleaning it, making sure it is complete and of high quality, and then transforming it into a form that will make it usable for the AI model – for example, with Spark or tools like pandas
- **Model selection:** Deciding which model is optimally performing the AI task for which it is programmed in terms of error rate and/or performance
- **Model development:** Designing the AI model using frameworks such as XGBoost, LightGBM, GLM, Keras, TensorFlow, PyTorch, Caffe, RuleFit, FTRL, Snap ML, scikit-learn, or H2O
- **Model training:** Training the model on compute infrastructure with sufficient processor and/or coprocessor cores for parallelization (increasingly also including the ability to explain, validate, and document the decisions of a model to ensure fairness, accountability, and transparency); this includes prototyping – testing the trained model by running inferencing on it.
- **Model hosting and monitoring:** Deploying the model in a production environment to execute the task for which it has been designed, typically referred to as "AI inferencing," and monitoring its performance

Workstations can play an important role in any of these six stages in combination with datacenter, cloud, or edge infrastructure.

FIGURE 1

AI Development Stages



Source: IDC, 2021

DEVELOPING AI MODELS ON WORKSTATIONS

Workstations Vis-à-Vis Personal Computers

It is generally well understood that personal computers (PCs) are not powerful enough for AI development. Data scientists and AI developers are typically involved in strategically important projects for their organizations, and unimpeded productivity is of the utmost importance. Workstations tend to perform more predictably than PCs, as they are usually built with higher-performance components and optimized for the software that is running on them. These components include:

- **High-grade processors:** One example is Intel Xeon Scalable processors.
- **Powerful GPUs:** One example is NVIDIA's Quadro RTX. The Ampere generation of Quadro RTX will be branded as NVIDIA RTX A6000.
- **More storage:** Some workstations can deliver as much as 48TB, and I/O speeds tend to be significantly higher than those of PCs.
- **More memory:** Workstations are now available with as much as 4TB of memory.
- **New memory types:** One example is Intel Optane Memory, which is Intel's persistent memory technology based on 3D XPoint technology that enables storing data and programs closer to the processor, thus reducing latency.
- **Error-correcting code (ECC) memory:** ECC detects and corrects the most common kinds of internal data corruption, preventing blue screens during a long AI training run from either a hard error (bad bit) or a soft error (flipped bit, causing bad values).
- **Specialized silicon:** One example is Intel Movidius vision processing units (VPUs), which are parallel processing coprocessors for computer vision and edge AI applications that are used in settings such as retail, security, and industrial automation. FPGAs are also used in workstations, for example, for financial applications.
- **Optimization software:** Examples include OneAPI, which is Intel's standards-based programming model to simplify development and deployment of data-centric workloads across CPUs, GPUs, FPGAs, and other accelerators, or CUDA, which is NVIDIA's parallel computing platform and application programming interface for running general workloads on GPUs.

CPUs Versus GPUs for AI

As mentioned previously, workstations can be used in various stages of AI development, and they are typically equipped for a variety of capabilities. Despite the emphasis on GPUs for parallel processing, the CPUs play a critical role when developing an AI model on a workstation. They are essential for data manipulation and, of course, for developing traditional ML models. CPUs are also used for data exploration — the process of using visual representations of a data set to understand the characteristics of the data.

In DL training, the host CPUs' role is somewhat reduced as the GPUs take over during the actual training process, but even then, the CPUs continue to serve as the processing layer for critical software such as the OS or CUDA and for orchestrating processes among the GPUs or with other silicon. Furthermore, the CPUs have increasingly taken on a new-way role as AI inferencing engines in cases where a workstation is used for running an AI model in production. IDC expects that, by 2024, spending on infrastructure for AI inferencing will exceed spending on AI infrastructure for AI training and that a significant portion (39%) of that inferencing will take place on the host CPUs.

Workstations Vis-à-Vis Servers: A Symbiotic Relationship

For most organizations, pragmatism is the rule of thumb as to when a workstation, an on-premises server, a cloud instance, or any combination of these three are deployed for AI development. There is a symbiotic relationship between workstations, servers, and cloud instances for the different development stages of an AI project.

The advantage of workstations versus datacenter servers is that data scientists can work anywhere they want — an important factor in the current pandemic but also under normal circumstances. They can also experiment freely on their AI models, iterating as often as they deem necessary, without having to request access to servers or running into other datacenter restrictions. And workstations provide them with the flexibility to move the compute closer to the data rather than the other way around, saving bandwidth, reducing network congestion, and increasing throughput. What is more, workstations can be configured for different needs: traditional ML tasks, for example, or more DL-intensive work.

Also, even though there is significant growth in the accelerated server market, accelerated servers are still not widely available in enterprise datacenters. At the time this white paper was written, on average, 4% of servers in enterprise datacenters were accelerated, meaning that many organizations do not have the means to develop or run AI on readily available on-premises GPUs. For this reason, too, accelerated workstations are a useful alternative for AI development.

Highly accelerated workstations are now powerful enough that they can perform DL training as long as the AI model is not excessively large, eliminating the need for training on servers. And models trained on workstations with GPUs can be deployed on either workstations or on servers without GPUs, leveraging inference capabilities in the CPUs. Software technologies such as Intel's DL Boost and oneAPI can power AI inferencing on the CPU, enabling non-accelerated servers already deployed in datacenters to support AI applications.

Workstations Vis-à-Vis the Cloud

Cloud computing has revolutionized how organizations think about infrastructure, data, and applications. With the promise of near-limitless scalability, the cloud allows developers to provision resources on demand, potentially accelerating the pace of innovation with fewer constraints. At face value, the cloud appears to be the perfect paradigm for AI development.

However, this is not always the case. In fact, IDC's survey data (see *Infrastructure for Edge Survey,* June 2020) revealed that 15% of organizations have repatriated workloads from the public cloud to on-premises infrastructure in the past 12 months. This is driven by several factors:

- **Cloud availability:** Anyone who has relied on cloud services has experienced an outage, whether due to problems within the cloud provider itself or due to a lapse in network connectivity somewhere between the hyperscale datacenter and the end user. In these situations, users are at the mercy of the service provider to resolve the issue, while productivity grinds to a halt.
- **Security and compliance:** In many industries, corporate governance policies dictate where data can be communicated and stored, which limits the usage of cloud services. Government regulations such as GDPR in Europe and the California Consumer Privacy Act also enforce rules on data sovereignty.

- **Cost:** It is common for organizations to underestimate how quickly cloud service fees can grow, especially for workloads that require high-performance compute capabilities. Cloud economics are based on metering all types of resource consumption, including egress of data back to onsite infrastructure.
- **Trial-and-error pressure:** Most AI initiatives start with a significant amount of experimentation, in which models that fail are part and parcel of the development process; in this process, there is a psychological tax that AI scientists and developers pay when cloud billing accumulates without them being able to show executable results yet.

Workstations can address these limitations while still leveraging cloud-native technologies such as microservices-based architectures and API-driven automation. This enables some of the same benefits as when comparing workstations with datacenter servers:

- **Work anywhere:** By removing dependence on the public cloud, disconnect scenarios are now possible. Many high-security environments are air gapped from public networks, and AI workstations can uniquely address this need. Local resources also reduce demand for expensive network connectivity.
- **Data locality:** The proliferation of IoT devices and other connected equipment is contributing to an exponential growth of data in edge locations. In many situations, it makes sense to colocate the computing resources with a dedicated workstation. This also solves many compliance requirements by limiting the movement of data.
- **Experiment freely:** The training and optimization of AI models is an iterative process, which often includes some element of trial and error. Developers need the freedom to conduct experiments without compromising because of the potential of additional service fees. Workstations also provide more flexibility for custom tooling.

Regarding the latter point, comparing the price of a workstation with a cloud deployment is relatively easy as most cloud service providers will give instant cost estimates of any configuration that an end user wishes to deploy. For example, the cost of a single regular virtual machine (VM) with one NVIDIA T4 and one instance of 375GiB SSD storage that is used eight hours a day, five days a week, is $140 on one major cloud provider. Double the VMs, T4s, and SSDs, and the cost will go up to $365 per month. Stay on two VMs but double the T4s to four and the storage to 4 x 375GiB and do a full-time training run on the environment, and the cost goes up to $2,700 per month. So it is fair to say that cloud costs for AI development can easily spike to tens of thousands of dollars per year, substantially more than the annual depreciation of a high-end workstation.

PROTOTYPING AI ON WORKSTATIONS

Compared with both on-premises servers and the cloud, workstations provide a distinct advantage when it comes to prototyping AI models. Servers in the datacenter may be at full utilization or be too mission critical for AI prototyping and testing, and as discussed previously, cloud instances can quickly lead to cost overruns when used liberally as a test environment. Workstations release the AI scientist or developer from having to negotiate server access or from the nagging concern of racking up cloud bills during the prototyping stage. Their low one-time costs provide complete freedom to prototype anywhere and at any time without additional costs.

DEPLOYING AI MODELS ON WORKSTATIONS

While developing AI models on a workstation has been a common strategy for years, IDC is seeing increasing use cases for *deploying* an AI model on a workstation, typically at the edge — in other words, putting the AI model in production on the workstation by having it run inferencing on the AI model. The edge is growing fast as an AI deployment location for servers — more than tripling from 2020 to 2024 in terms of annual hardware spending — and workstations are not far behind, as end users discover their advantages at the edge.

IDC defines edge as a distributed computing paradigm that includes the deployment of infrastructure and applications outside of centralized cloud and on-premises datacenters as close as necessary to where data is generated and consumed. This includes remote and branch offices as well as industry-specific locations such as factories, warehouses, hospitals, and retail stores.

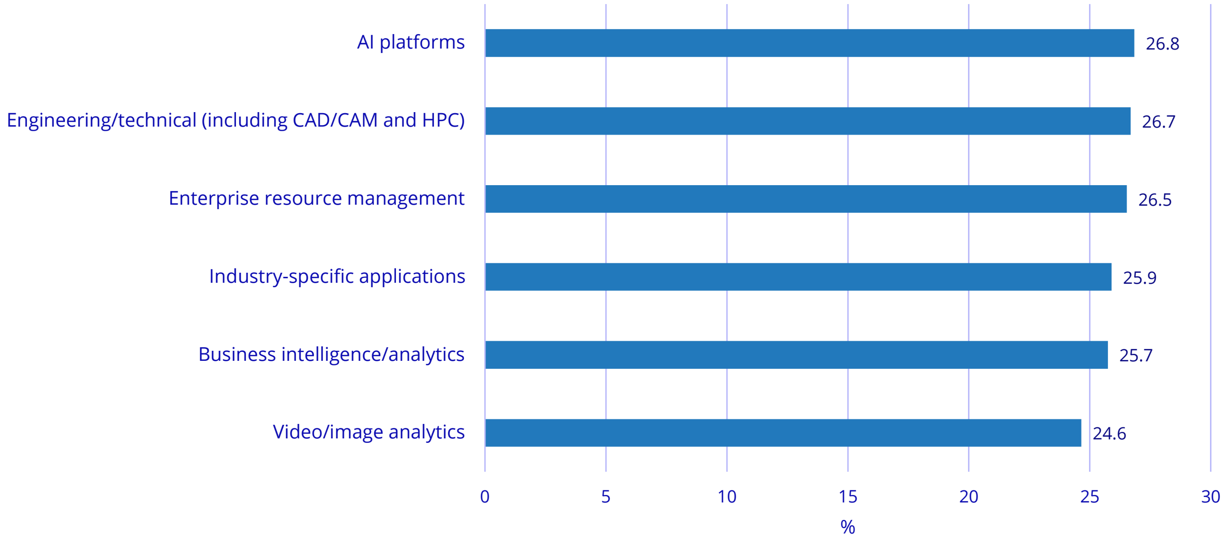
Data- and compute-intensive workloads are increasingly deployed either on premises or in edge locations. This is done to mitigate limitations inherent in public clouds such as the time it takes to upload large data sets and the variable costs of conducting AI training, especially in situations that require a significant amount of data science experimentation.

IDC research shows that 26.8% of organizations plan on deploying AI platforms on premises or in edge locations in the next 12–24 months (see Figure 2). In the same study, 26.7% are following suit with other engineering and technical workloads. This is an area where it makes sense to deploy an AI workstation.

FIGURE 2

Outlook for Workloads Deployed On Premises and in Edge Locations

Q. Which of the following workloads used by your organization will be increasingly deployed at the edge and/or to the private/hybrid cloud in the next 12-24 months?



n = 637

Base = all respondents

Notes:

This survey is managed by IDC's Quantitative Research Group.

Data is not weighted.

Multiple responses were allowed.

Use caution when interpreting small sample sizes.

Source: IDC's *Infrastructure for Infrastructure Deployment Survey*, June 2020

When deploying an AI model on a workstation at the edge, there is not always the need for high-end GPUs as is the case for AI development. Lighter GPUs can perform the AI inferencing, and in quite a few cases, GPUs are not needed at all. In those instances, CPUs can adequately perform the inferencing task, especially when used with optimizations such as Intel DL Boost, a set of instruction set features on Intel microprocessors designed to accelerate AI workloads, including AI inferencing. With Intel DL Boost and 2nd Gen Intel Xeon Scalable processors, Intel says it saw up to 7.7x performance for a FP32 model and up to 19.5x performance for an INT8 model when running ResNet50 inference. This also helps making a workstation more suitable for being deployed at the edge, where considerations such as power, mobility, and thermal management demand smaller wattages. Intel Movidius Myriad (M2) fits well in this power envelope, thanks to its small energy footprint at 12W.

Use Cases for Deploying AI on Workstations

There are several situations that naturally lend themselves to deploying AI on locally deployed workstations. Common characteristics are large volumes of machine-generated time-series data and unstructured data like video streams and images. There are also instances where subject matter experts must augment AI models with human interpretation.

Examples include:

- **AIOps:** As IT systems grow in scale and complexity, there is an increasing need to move from reactive incident management to a proactive monitoring. This is especially true as infrastructure and applications are distributed to edge locations where there is little to no technical staff. By modeling a baseline of normal performance, it is possible to identify anomalies and automate remediation steps.
- **Disaster response:** In an emergency, first responders must quickly assess a situation, track critical equipment, and deploy resources to help those most in need. This often must take place in an environment without network connectivity, necessitating a local workstation that can aggregate data feeds, infer against AI models, and automate communications to key personnel.
- **Radiology:** Advances in imaging technology has led to an increase in the size of data generated from a single scan, requiring it to remain onsite for it to be analyzed in a timely manner. AI models trained from millions of previous examples can identify patterns more accurately than the human eye, increasing accuracy rates.
- **Oil and gas exploration:** Upstream oil and gas companies use a combination of telemetry, seismic, and imaging data to locate reserves of natural resources, select drilling locations, and optimize the performance of equipment in the production process. This often requires analysis of information in areas where only expensive satellite communication is available.
- **Land management:** Satellite images in conjunction with object recognition techniques can be employed to assess changes in forests, water bodies, and cities. This not only monitors climate change in the environment but also identifies misuse of natural resources and violations of government regulations.
- **Insurance claims assessments:** Manual claim processing is labor intensive and prone to human error. AI that can evaluate claim validity reduces costs by allowing insurance adjustors to focus on the cases that require more investigation. This increases the overall throughput of the operation without sacrificing accuracy.
- **Telehealth:** AI is improving patient recovery rates by tailoring individual treatment plans based on real-time vital signs from wearable devices. This information is combined with historical patient records and a knowledge base of similar cases. This is particularly important in rural areas that have a higher reliance on remote healthcare.
- **Retail security (anti-theft):** Real-time analytics applied to video streams are being used to predict human behavior that may lead to criminal activity. This typically requires stitching together multiple video feeds to track an individual's movements within a store. Given the time-sensitive nature of identifying a material event, this is a process that is best run locally.
- **Traffic management:** Government entities that are responsible for transportation operations are increasing using AI to provide coordination of traffic lights and digital signage to improve the flow of vehicles and keep citizens safe. This requires a combination of inputs including videocameras and telemetry from road sensors to optimize traffic patterns.

- **Manufacturing plant monitoring:** For a plant manager, ensuring uptime of critical processes and meeting production schedules are paramount. This translates to predictive maintenance of key equipment, automated detection of defects, and optimization both in and out of the site's supply chain. This is an area where AI can assist human operators to increase performance while maintaining safety standards.
- **Drones:** Automated analysis of images captured by drones are enabling the ability to monitor a wide range of conditions at a scale not previously possible. This is having a significant impact on the inspection of gas and electric utility infrastructure, insurance surveys, search and rescue efforts, precision agriculture, and maintenance of fisheries and wildlife preserves.

DELL WORKSTATIONS FOR AI

Dell offers a wide range of workstations for various levels of AI development and/or deployment, all under the umbrella of the Precision brand. We will briefly provide the specs and then discuss where Dell stands out.

Dell Precision Workstation Specs

Dell Precision 5820 Tower

This is Dell's entry-level single-socket system based on the Intel Xeon W Processor family (up to 18 cores), which is a processor platform specifically for professional creators, designed for performance, security, and reliability as well as capabilities for visual effects (VFX), 3D rendering, complex 3D CAD, and AI development and edge deployments. The system is accelerated for deep learning training and inferencing and comes with up to 512GB ECC memory and 1TB of NVMe storage. It is "NGC-Ready," meaning that it has been certified by NVIDIA as a platform that can take advantage of NVIDIA GPU Cloud (NGC), a hub of GPU-accelerated containers for DL, traditional ML, and HPC that are optimized, tested, and ready to run on workstations with NVIDIA GPUs.

Dell Precision 7920 Tower

This is Dell's highest-end workstation targeting ML and DL workloads and running on one Intel Xeon Scalable Gold processor (up to 28 cores), up to three NVIDIA GPUs with NVLink, NVIDIA's extremely fast CPU-GPU and GPU-GPU interconnect technology, 3TB of ECC memory, and two NVMe devices of 1TB each. This workstation is also NGC-Ready.

Designed for heavy-duty AI training tasks, this workstation is also available with two Xeon Scalable processors with up to 28 cores, up to three NVIDIA GPU cards (with 32GB memory each) with NVLink, 3TB of ECC memory, and two NVMe devices of 2TB each.

Dell Precision 7550

This is Dell's high-end 15in. mobile workstation, featuring a 10th generation Intel Xeon processor (up to eight cores), NVIDIA or AMD GPUs, 128GB memory, and up to four 2TB NVMe SSDs. The workstation also comes with Dell Optimizer for Precision (DOP) (see the Dell Optimizer for Precision section) and is NGC-Ready.

Dell Precision 7920 Rack

This is Dell's 2U rack-based workstation that is available with one or two Intel Xeon Silver Scalable processors (up to 28 cores per processor), three NVIDIA GPUs with 40GB memory each for enabling larger data sets, and 3TB of ECC memory. It is NGC-Ready.

How Dell's Workstations Stand Out

Three important areas where Dell workstations stand out because of Dell's technologies that are not available on other vendors' workstations are discussed in the sections that follow.

Reliable Memory Technology

Dell provides a technology on top of ECC that is called Reliable Memory Technology Pro (RMT Pro), which is designed to help maximize uptime. It works in conjunction with ECC memory to detect and correct memory errors in real time. According to Dell, RTM Pro virtually eliminates memory errors by preventing bad memory from being revisited again even as the DIMM remains in full use. After a system reboot, RTM Pro will isolate the defective memory area and hide it from the OS. As a result, AI data scientists and developers will not run into the issue of ongoing crashes because the bad memory continues to be addressable — a major productivity boost.

Dell Optimizer for Precision

Dell also includes Dell Optimizer for Precision in most of its workstations, which automatically adjusts system settings so that the workstation will run various popular commercial applications at the fastest speed possible. This improves a data scientist's or developer's productivity. The tool also creates real-time performance reports for IT on processor, storage, memory, and graphics utilization. DOP does not run on Linux yet and is therefore mostly useful for deploying AI, as developing AI tends to be done with Linux-based open source software. Dell Optimizer for Precision also provides ExpressSign-in, ExpressCharge (on mobiles), Intelligent Audio, and reporting and analytic tools to help fine-tune the workstation.

Optane Persistent Memory

Some of Dell's workstations offer Intel Optane Persistent Memory, which Intel made available for workstations in late 2019, with capacities up to 512GB per module, totaling up to 3TB per CPU or 6TB per workstation for systems with two processors. Optane Persistent Memory consists of nonvolatile media that is placed on the DIMM and installed on the memory bus. It exists alongside the workstation's DRAM and works in conjunction with the DRAM for higher system memory capacity or better performance through DRAM caching. This is an important capability for AI scientists and developers as it allows them to run bigger AI models, a critical capability as neural networks are becoming larger and larger.

CHALLENGES/OPPORTUNITIES

For Businesses

IDC is seeing a bifurcation in the market for AI. On the one hand, businesses are compelled to get onboard with AI in a big way to remain competitive. By way of example, they are presented with peers that have done extraordinary work, using enterprise AI infrastructure offerings that are actually registered in the top 100 supercomputers. On the other hand, businesses see the daily reality of small AI initiatives being tried out on available servers in the datacenter or in the cloud, often with insufficient budget and underperforming hardware.

For many businesses, the first scenario is not relevant and the second one all too real. For them, the challenge is to give their AI data scientists and/or developers the right tools to perform AI training in a timely fashion without spending vast amounts of money on cloud instances or GPU-accelerated datacenter servers. IDC believes that these businesses are well served by supplying their scientists and developers with powerful GPU-accelerated workstations.

For Dell

There is a misunderstanding in the market that AI development and deployment requires expensive, accelerated server hardware, often even in a cluster. This may be true for the largest of AI algorithms, with billions of parameters, but most businesses are not developing such massive algorithms. They are doing something useful, impactful, and manageable with their AI initiative, and many businesses do not realize that such common-scale AI models can be developed – and deployed – on workstations. Dell's challenge is to break through the preconception and educate the market about the possibilities with its workstation portfolio.

At the same time, Dell must make sure that its workstations do deliver and do not become technology bottlenecks over time. This means rapid ongoing innovation so as to never disappoint end users who are using the workstations appropriately (in other words, who are not trying to run a multibillion parameters algorithm). It also means that, for customers that are suddenly starting to scale very fast or whose algorithms are indeed becoming very large, there is a seamless transition from the workstation to Dell's AI server line. Therein, of course, also lies the opportunity for Dell – to have the right solution for every customer, no matter what size AI initiative they are working on.

CONCLUSION

IDC believes that workstations are currently underappreciated as the workhorses of AI development and deployment for many use cases. They provide AI scientists and developers with a powerful GPU-accelerated platform that represents lower capex than servers, dramatically lower opex than cloud instances, and much greater freedom to experiment with AI models. Businesses that are developing AI initiatives that do not require billion-parameter algorithms (as most do not) should consider empowering their AI teams with workstations for unconstrained AI development and for easy edge-based deployment.

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

Copyright Notice