DELLTechnologies

NVIDIA

AI APPS:

Why developers are turning to ultra-powerful workstations for more creative freedom at less cost

Research shows that large and small companies alike are using powerful workstations with even more powerful graphic processing units (GPUs) as integral parts of their artificial intelligence infrastructure. Are you among them? If not, you should be.

Artificial intelligence (AI) offers the biggest and most transformational business opportunities for app developers and OEMs in a generation. Already AI applications are reshaping the products, services and business models of companies, small and large, across just about every industry. Manufacturing, logistics, financial services, energy, healthcare, marketing, customer service—and the list goes on. Even McDonald's has set up an AI center in Silicon Valley—McD Tech Labs—to develop ways AI can help boost same-store sales.[1]

The market for AI applications is immense and their economic value even greater. Forbes magazine published a roundup of 2019 predictions from leading market researchers as follows:[2]

- **IDC:** Worldwide spending on cognitive and AI systems will reach $77.6 billion in 2022, more than three times its $24-billion forecast for 2018.
- **McKinsey:** AI and machine learning have the potential to create an additional $2.6 trillion in value by 2020 in marketing and sales, and up to $2 trillion in manufacturing and supply chain planning.
- **Gartner:** The business value created by AI will reach $3.9 trillion in 2022.

AI and its subsets of machine learning and deep learning offer big opportunities for application developers and OEMs seeking to enhance the performance, monitoring and diagnostic capabilities of their products. But to achieve success, developers and OEMs need to make critical hardware and software choices for their development platforms—decisions that can impact their business strategies and models. Highly scalable HPC infrastructure and cloud platforms are no longer the only option.

Workstations are proving their worth in AI development and workflows

To find out more about this trend, Dell commissioned Forrester to research the extent to which workstations are being used by companies to lower the cost, increase the security and accelerate workflows of their AI infrastructure.

Its survey of 210 respondents across a wide range of U.S.-based industries revealed that the addition of workstations to a firm's AI workflow allows servers and cloud platforms to be tasked with business cases that require more robust computing while workstations take on tasks with longer timeframes and smaller budgets.[3]

In fact, Forrester found that the use cases among respondents deploying workstations as part of their AI infrastructure included:

33%
to develop AI applications

32%
to train or process algorithms

26%
to run core AI applications

Cloud AI development takes a backseat to powerful workstations

Notably, among those firms surveyed, only 24 percent were using cloud environments for AI application development. According to the research, many companies developing or running AI applications have found workstations to be more cost-effective, especially in running the many thousands of iterations required to train AI algorithms in their work purposes compared to using cloud or HPC data center environments.

A 4x performance boost and three months time savings?

How an AI app developer is using powerful tower workstations to accelerate time to market and help doctors mitigate blindness

Voxeleron has developed an ophthalmic AI app that uses a deep-learning, convolutional neural network (CNN) model to enhance retinal images. This lets doctors predict which patients suffering age-related macular degeneration will convert from its dry form to its blinding wet form.

In effect, the application is a triage tool, providing ophthalmologists with predictive powers they lacked before. The company needed to develop a CNN-based, deep-learning model that could be rapidly trained via machine learning to process 3D imaging data sets drawn from optical coherence tomography.

To achieve this, Voxeleron acquired Dell Precision 7920 Tower models and added three NVIDIA Quadro GV100 GPUs to each one. The NVIDIA GV100 GPU features more than 5,000 cores, including 640 Tensor cores, and is capable of 118.5 teraflops of Tensor performance.

"We saved as much as three months by running our AI models on the Dell Precision 7920 Tower workstation with NVIDIA Quadro GV100 GPUs versus our previous setup," says Jonathan Oakley, a cofounder. "That's about a fourfold improvement, helping us get our AI app to market much sooner and, most importantly, helping doctors better mitigate blindness caused by an incurable disease."

Dell Precision Workstations for AI

Dell offers three families of AI-ready Precision workstations. All are engineered and built to accommodate NVIDIA GPUs for massive parallel processing on the desktop or on the go:

- **5000 Series** of compact tower workstations for extreme reliability

- **7000 Series** of tower and rack models for scalability

- **7000 Series** of mobile workstations for powerful performance in the office, at home or in transit

Many of these Dell Precision workstations models can accommodate up to three NVIDIA GPUs, adding as many as 15,000 cores to a Precision workstation's ultra-baseline horsepower, which can feature multiple 22-core Intel® Xeon® CPUs, with up to 256GB of RAM. Other unique features include:

- Reliable Memory Technology (RMT) Pro, a patented Dell feature for all Precision workstations with ECC memory that provides higher availability and reliability—especially important for mission-critical applications.

- Dell Precision Optimizer (DPO) is the industry's only AI-based performance optimizer software to automatically tune Precision workstations via machine learning, for up to 521% better application performance.

More creative freedom, less cost worries

In addition, they are finding more creative freedom to run investigations without worrying about the cost of running those investigations in the cloud. This can be especially advantageous to smaller firms with limited resources and startups with investors concerned about the latter's cash-burn rates. That's not to mention, of course, the ability to evaluate many thousands of AI models in-house using workstations can vastly expand their frontiers of discovery.

Five ways workstations can be deployed for AI applications most efficiently and effectively

In the course of its research into the use of workstations as AI infrastructure, Forrester was able to identify five best practices in use among the companies surveyed. These are:

1. Take a focused, practical approach to AI

In addition, they are finding more creative freedom to run investigations without worrying about the cost of running those investigations in the cloud. This can be especially advantageous to smaller firms with limited resources and startups with investors concerned about the latter's cash-burn rates. That's not to mention, of course, the ability to evaluate many thousands of AI models in-house using workstations can vastly expand their frontiers of discovery.

2. Develop a portfolio of infrastructure systems to power AI

Focused AI should involve local parallel-processing capabilities as well as the same in cloud and data centers to provide flexibility in AI development, training, workflows and operations. Highly scaled servers and cloud platforms can be required to run AI applications at top speed and where cost is not a barrier. Conversely, workstations can provide excellent support for applications where data security is a priority but where timelines are more flexible, and costs are not a major consideration.

3. Balance performance, cost and security

Firms designing an infrastructure to run AI applications must balance performance against costs while always keeping security in mind.

Think of it as a scale: Security keeps everything in balance, and you must make trade-offs between performance and costs. Performance speaks to the speed at which focused AI can actually solve problems, while costs involve highly skilled AI expertise, the technology to run AI models and the associated overhead for both. Costs must also align with an AI project's business case and justification.
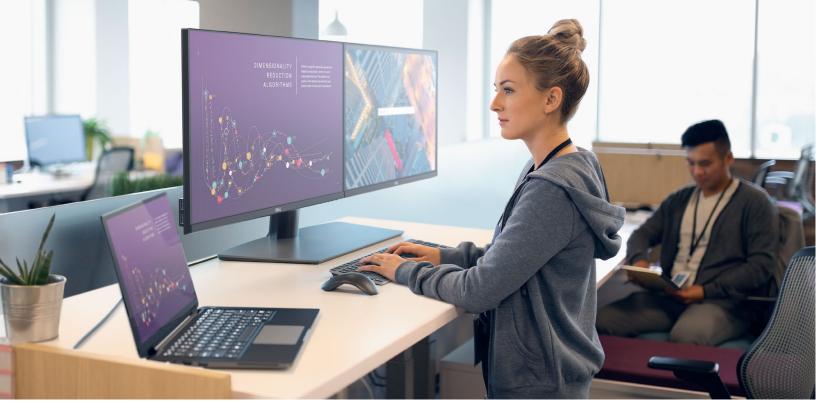
Security is more crucial than ever, especially with data sets that involve customer, patient or other sensitive data that can be subject to regulatory compliance. To meet increased security requirements, companies might need to trade off the time it takes to train and run algorithms.

Data is fundamentally changing the way companies do business, driving demand for data scientists and increasing the complexity in their workflows.

Get the performance you need to transform massive amounts of data into insights and create amazing customer experiences with NVIDIA-powered data science workstations. Built by Dell to combine the power of Quadro RTX GPUs with accelerated CUDA-X AI data science software to deliver a new breed of fully-integrated Precision desktop and mobile workstations for data science.

4. Deploy workstations as platforms for AI application development

Nearly a third of firms today see developing AI applications, training or processing algorithms, and R&D ideation or iteration as key workloads for workstations. The workflows of developers and data scientists can benefit from the flexibility, security and speed of workstations for creating focused models, even if these are later run as cloud or data center workloads. Today, they can evaluate or expand their use of workstations to achieve secure, cost-managed solutions that meet the needs of many AI applications.
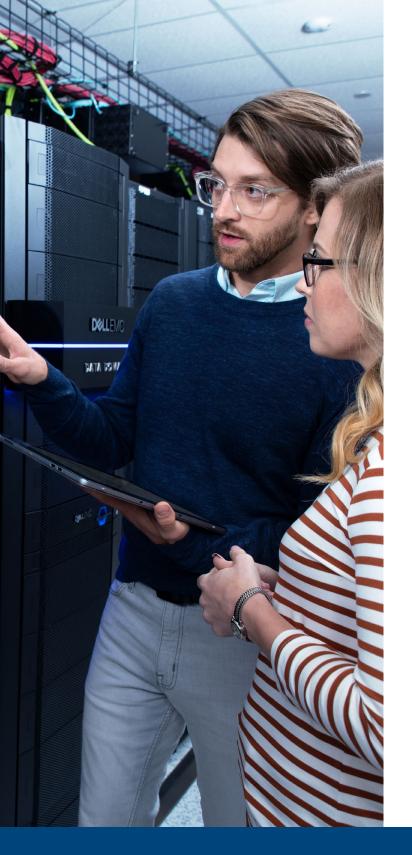
5. Weave workstations into an end-to-end process for AI

Today, nearly a quarter of workstation decision-makers say they already use workstations to run core AI applications in pretrained vertical solutions, intelligent research solutions, image and video analysis, and speech analytics.

Even for scenarios requiring cloud scale, workstations can still play a key role. For example, workstations can be used to explore AI models, train algorithms and manage data that will later be scaled out to the cloud.

Open source algorithms are another example. They can be run on workstations for evaluation, modified as needed and then deployed to the cloud once the AI applications are ready for launch.

Workstations can then be used in combination with data center and cloud deployments during the development and training phases. They may offer a better overall solution in cases when time to train an AI algorithm or run a problem is a smaller consideration than budget.

The time to get started building out AI development and operational infrastructures is now

Powerful AI-ready workstations with GPUs are proven workhorses in AI infrastructure deployments. As the Forrester research has shown, companies of all sizes across a wide range of different industries are using them as on-premises AI infrastructure, either as standalones or as complements to data center and cloud HPC parallel-processing capabilities.

These firms are finding that workstations can provide more efficient, secure and cost-effective AI development, while expanding the investigative horizons of their development teams.

For developers and OEMs, the time to get started building out an AI development and operational infrastructure is now. You'll open doors to tremendous business opportunities while gaining extreme competitive advantages. If you wait, you may be ceding the high ground on those opportunities to new market leaders.

To learn more about stepping up your AI game fast, contact your Dell or NVIDIA representative today. Or visit delltechnologies.com/precision.

Not only can Dell Technologies and NVIDIA provide the powerful workstation platforms to save your AI development efforts cost and time, but they can provide strategic partnerships—especially for small and medium-size businesses. They can also open doors to their respective ecosystems of AI partners, leaders in their particular fields.

1 David Yaffe-Bellamy. "Would You Like Fries With That? McDonald's Already Knows the Answer." The New York Times. October 22, 2019.
https://www.nytimes.com/2019/10/22/business/mcdonalds-tech-artificial-intelligence-machine-learning-fast-food.html

2 Louis Columbus. "Roundup Of Machine Learning Forecasts And Market Estimates For 2019." Forbes. March 27, 2019.
https://www.forbes.com/sites/louiscolumbus/2019/03/27/roundup-of-machine-learning-forecasts-and-market-estimates-2019/#4323fd357695

3 "Using Workstations to Reshape Your Artificial Intelligence Infrastructure." Forrester. May 2018.
https://tools.marketimpacttools.com/go/dell/workstations/

DELLTechnologies

NVIDIA