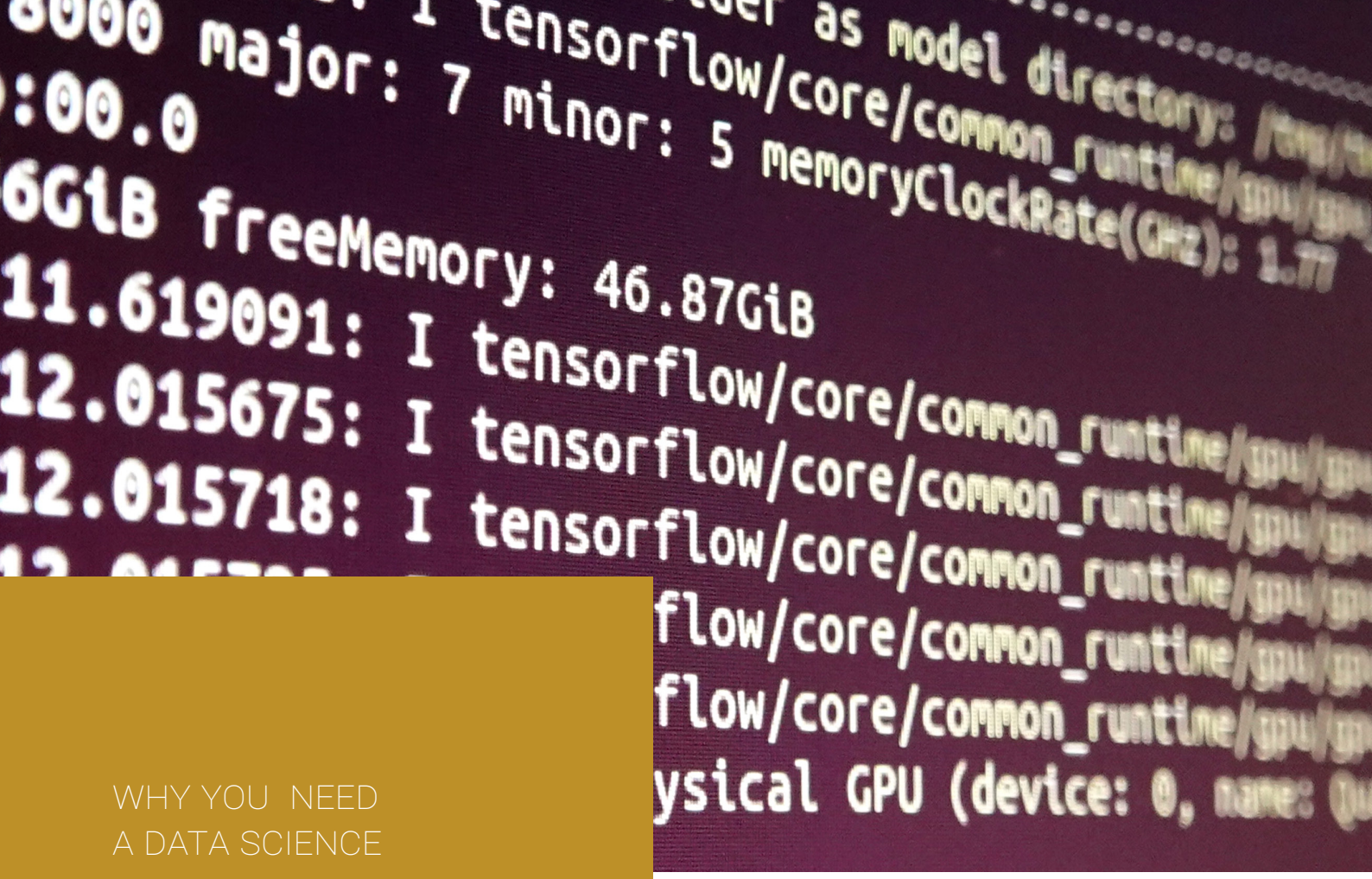


WHY YOU NEED
A DATA SCIENCE

WORKSTATION

For Artificial Intelligence and Big Data

DATA IS NOT INFORMATION

- The world produces enormous amounts of data every 24 hours.
- Decision making requires actionable information, not data.
- Artificial intelligence and data analytics technologies transform data into information.
- Success hinges on giving your team high performance, reliable tools.
- AI and data science make businesses more intelligent, more competitive, and more efficient.

THE DATA SCIENCE WORKSTATION: PROBLEM SOLVED.

How do you get the right computing hardware and software tools into your project fast?

The world produces enormous amounts of data every 24 hours. And the volume of data is growing. Valuable data flows through your company every day.

Data is not information. For companies to make decisions, data must become information. Until recently, processing this vast quantity of data has been very difficult if not impossible. Today, converging hardware and software technologies make the impossible, possible.

AI and big data analytics are key technologies in the world of data science. GPU computing is critical to both. It is no surprise that Data Science Workstations are built on massive GPU computing performance wrapped with enterprise-class workstation hardware, a complete software stack for AI, and professional support.

Together, these hardware and software technologies can deliver AI and data science tools to make almost any business activity more intelligent, more competitive, and more efficient.

ARTIFICIAL INTELLIGENCE

Artificial intelligence today means neural networks. The concept of computer-based neural networks has been part of AI research for well over half a century. So why have neural networks, and with them AI, only taken off in the last ten years?

The answer is simple: computing power. Neural network research had good models and methods which remained impractical due to the time and computing resources needed to train a network. Then researchers applied GPU computing to neural network training.

It turns out that the extreme parallelism of GPU computing is a perfect match for AI challenges. The application of GPU computing to neural network training created an explosion in AI applications.

THE TRIAD OF PERFORMANCE, PRODUCTIVITY, & PROFESSIONALISM

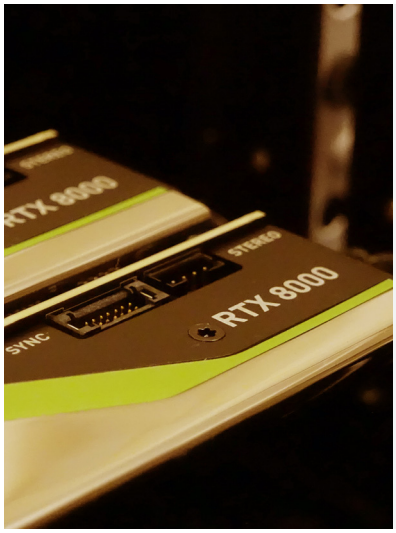
A Data Science Workstation is designed, configured, and supported with performance, productivity, and enterprise support. NVIDIA Data Science Workstations are built with professional graphics, a complete, pre-installed software stack, and dedicated support.

Performance is critical for AI, data science, and HPC. The performance challenge is tackled in my system using dual Quadro GPUs with a high-speed interconnect device, NVLink.

The workstation is pre-configured with a thoroughly tested, GPU-accelerated, software stack. The pre-configured software is tailored to data science and AI.

The benefit of this software stack configuration is simple: push the power button and run workloads in fewer than 15 minutes. Compare 15 minutes to the hours, if not the days, needed to properly install, configure, and test Python, Docker, CUDA, RAPIDS, Dask, and the AI frameworks.

The solution is completed wth expert, enterprise-level support and a well-developed user community.



THE HEART OF A DATA SCIENCE WORKSTATION IS THE GPU

In most computing infrastructures, the heart of the system is the “central” processing unit, the CPU. That is not true for a data science workstation. It’s heart is the GPU, or more accurately, multiple GPUs.

System performance depends on balance. A typical data science workstation configuration includes dual GPUs, dual CPUs, extensive memory, and fast storage. The secret sauce is an optimized software stack that uses every ounce of available computing power.

FASTER TRAINING, FASTER DISCOVERY, FASTER ANALYTICS: EACH DEPENDS ON FASTER GPU COMPUTING

A fast GPU is essential but not sufficient. A Data Science Workstation needs multiple GPUs, lots of graphics memory, and a smart, fast interconnect.

The hardware specification is critical. Whether in AI or in analytics, data science problems are accelerated with GPU computing. Choosing graphics is arguably the most important decision.

At the heart of this Data Science Workstation are two Quadro RTX 8000 graphics boards. Each hosts 48 GB of GDDR6 graphics memory, and the boards are connected with NVLink.

Performance also scales well with multiple GPUs. NVLink is a high-speed interconnect device based on Mellanox technology which allows the GPUs to communicate and share data efficiently and is critical to multi-GPU performance.The board requires two slots and a lot of power: 295 W total dissipated power (TDP).

The GPU sports 18.6 billion transistors and is built on a 12 nm process. All of those transistors deliver impressive specs. The GPU has 4,608 CUDA cores and 576 Tensor cores.

The board is manufactured with 48 GB of GDDR6 memory. The large memory capacity allows for larger data sets to be kept on the graphics card.

Tensor cores are designed to accelerate Deep Learning. They accelerate specific mathematical operations which are commonly found in Deep Learning applications. One example includes a combined matrix multiply plus accumulate function.

RT cores are specific to casting rays in raytracing. The accelerated raytracing pipeline is combined with NVIDIA's AI-based de-noising function. The combination uses both RT cores for graphics and Tensor cores for AI to enable real-time raytracing. It is one example in which AI is able to accelerate processes by orders of magnitude compared to alternative methods.

MULTI-GPU PERFORMANCE DEPENDS ON NVLINK

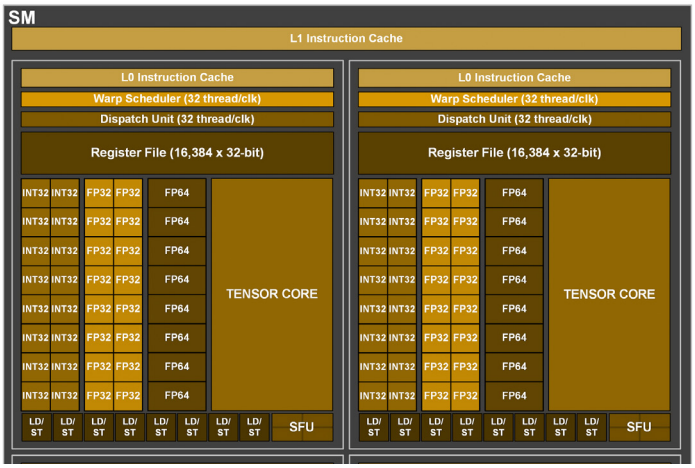
The two boards are connected via NVLink. NVLink provides a high-speed interconnect between the GPUs and their memory which makes on-board calculations more efficient.

NVLink allows for transfers of 100 GB/s between the GPUs. NVLink also creates a flat memory space so that both the GPU and the CPU can access system and graphics memory directly.

THE GPU IS A HIGHLY PARALLEL COMPUTING ENGINE

Parallelism is exactly the nature of graphics. It's natural to apply GPU power to parallel processing problems.

In AI, applications are driven by neural networks and Deep Learning. Training a neural network to perform its task is an inherently parallel computing problem which requires large data sets and a large number of training iterations.



This massively parallel, computationally intensive problem was unrealistic until GPU-computing could be applied.

Parallelism is fundamental to graphics processing. It's natural to apply GPU power to other computationally intensive, parallel processing problems.

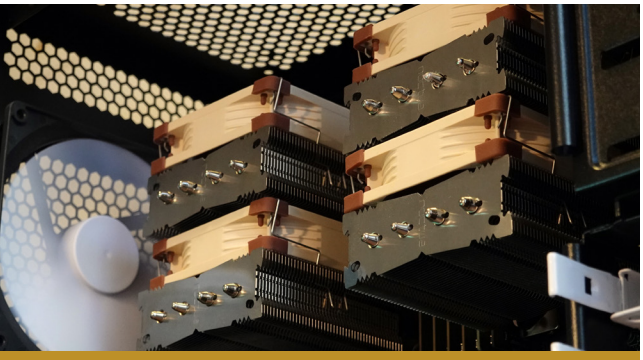
These graphics processors were developed and grew in the shadows. In the 80's and early 90's graphics was accelerated by ASICs. These ASICs were not programmable, and they were referred to simply as graphics chips.

Graphics chips executed very low-level 2D drawing functions: line drawing and later basic 3D functions like triangle shading and texture mapping. The computationally intensive graphics functions were still processed on the CPU.

Transistor counts increased and more of the 3D computations were absorbed by the graphics chip. Further acceleration required the graphics chip to become programmable. Initially, only simple programmability with limited flexibility was possible.

But the quest for faster, more realistic graphics meant that the ability to program a graphics chip was unstoppable. The graphics chip became a complex parallel computing processor. Initially, developers loaded and ran general purpose programs on the GPU via graphics APIs such as OpenGL. When NVIDIA married the graphics processor with the C-like programming language, CUDA, the GPU was finally born.

GPUs are the key computing technology driving AI training applications today. The GPU sits at the heart of Data Science Workstation acceleration.

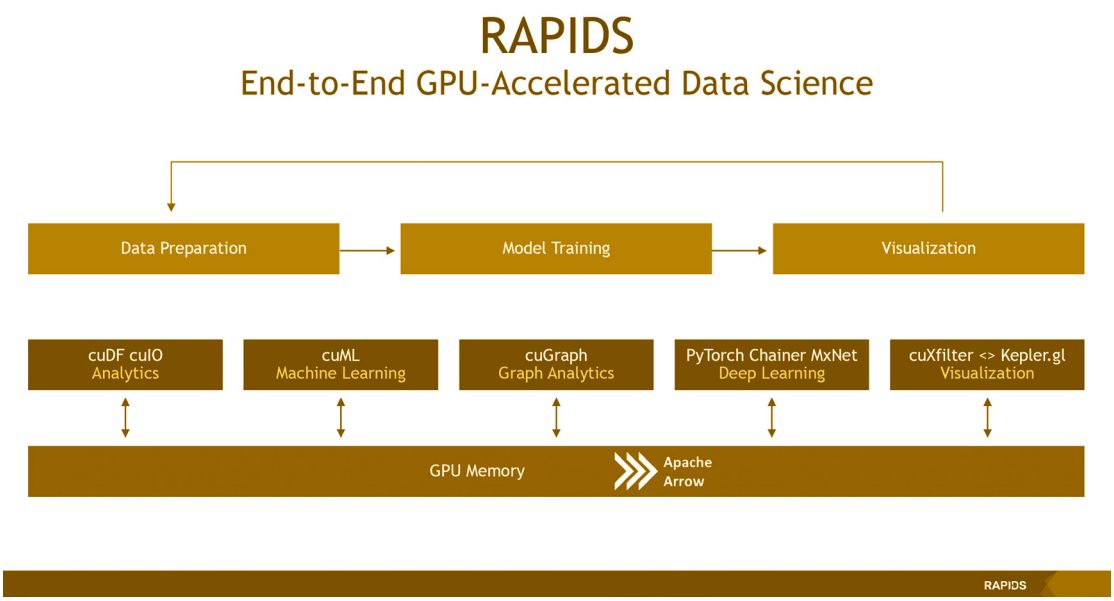


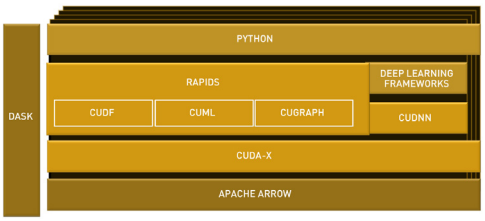
INTEL XEON PROCESSORS PROVIDE DL BOOST

Research in AI training and inference has shown that computing at lower precision can yield results that are equally accurate to higher precision calculations.

This research lies at the heart of Intel's DL Boost technology. Reducing the computing precision for deep learning has two benefits. One benefit is to reduce memory bandwidth requirements. A second benefit is to reduce the transistors and power needed to perform operations. In other words, the same number of transistors and amount power can process more data. This optimization is implemented in Intel's fused-multiply add (FMA) units.

To make this feature available, Intel developed a new instruction set, Intel Advanced Vector Extensions 512 (Intel AVX-512) for Xeon Phi co-processors. Today, the Intel AVX-512 instruction set is also available on Intel Xeon Scalable processors.





AN END-TO-END GPU ACCELERATED WORKFLOW WITH RAPIDS

RAPIDS interfaces to Python above it and uses the CUDA libraries below it. This combination delivers the familiar Python environment for data scientists on top of the GPU-optimized CUDA platform.

RAPIDS integrates with Apache Arrow which allows the data from RAPIDS to be used seamlessly by many deep learning frameworks such as Chainer, PyTorch, DLPack, TensorFlow, and MXNet.

Because RAPIDS works well with Python, it also works well with many data science visualization libraries. When these libraries leverage the native GPU in-memory formats, they can achieve high-performance rendering even with large data sets.

RAPIDS IS THE HEART OF THE GPU ACCELERATED SOFTWARE

The NVIDIA Data Science Workstation is pre-configured with a thoroughly tested, GPU-accelerated software stack. The pre-configured software is tailored to data science and AI.

The core of the GPU-acceleration for AI and data science comes from RAPIDS. Altogether, a properly configured AI & data science workstation needs RAPIDS, CUDA-X, Python, multiple deep learning frameworks, and GPU-optimized libraries. NVIDIA Docker containerization, which simplifies development and deployment with GPU acceleration, is included in the software stack as well.

The software stack for the data science workstation is comprehensive. Data scientists can access tools and interfaces like Python as well as many deep learning frameworks with which they are already familiar.

Given the completeness of the pre-installed software stack, it is very likely that you will be running existing projects on a new Data Science Workstation in minutes instead of hours or days. This makes injecting the performance of a Data Science Workstation into running projects very simple - essentially seamless.

ACCELERATE. SCALE. DEPLOY.

RAPIDS, Dask, and Docker accelerate, scale, and deliver data science at scale.

RAPIDS provides GPU accelerated libraries. Dask provides scaling from one workstation to a data center filled with GPU accelerated servers. NVIDIA Docker provides GPU accelerated portability. This combination is critical for data science workloads.

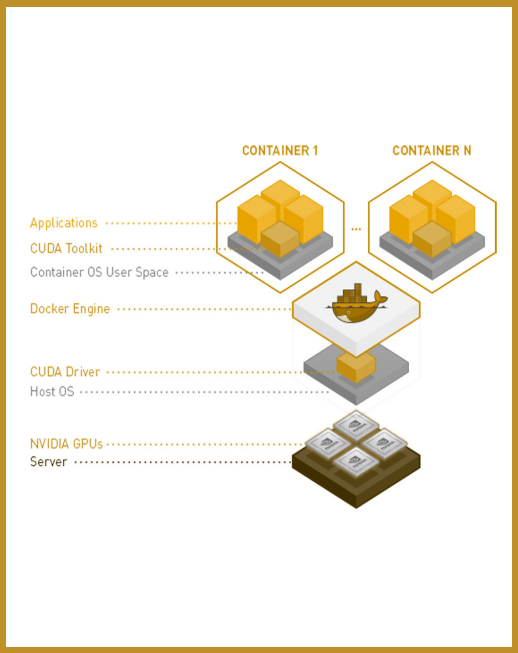
Dask scales Python workloads. It scales from a single notebook computer to data center clusters. The software schedules individual Python workloads distributing them across the available computing resources. That could be multiple cores in a single CPU to 1000 nodes in a data center super computer.

Dask is an open source project. The project coordinates with other relevant projects like Numpy, Pandas, and Scikit-Learn.

Without Dask, data scientists often discovered that their initial Python work would not scale. Dask integrates easily into Python so that data scientists do not need to rewrite their code when they want to scale-up a project.

Containerisation is used across the IT industry. It facilitates the development and deployment of any application across an infrastructure of heterogeneous computing systems. Docker is the most widely used container.

NVIDIA added enhancements to integrate GPU acceleration into portable Docker containers. creating a wrapper around the Docker container that initializes the system. This enables GPU computing in the Docker container.



DOCKER CONTAINERS: NVIDIA DOCKER

A problem every company faces when developing and running projects on a Data Science Workstation is deploying those projects on other workstations or on AI and HPC servers. Docker containers can help. Docker delivers portability.

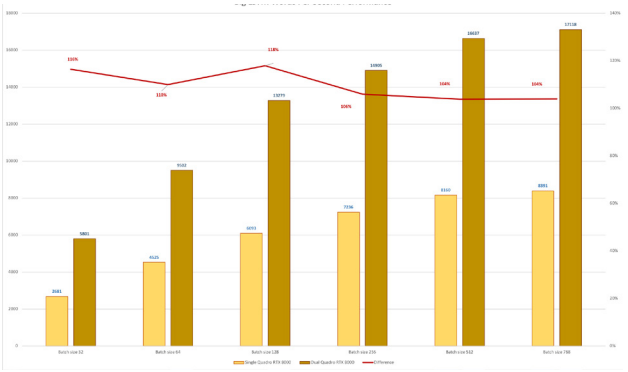
Therefore, NVIDIA Docker is a critical part of the Data Science Workstation's software stack. In order to gain the value in development and deployment that Docker containers provide, it is necessary to use GPU accelerated containers.

Making containers run seamlessly with GPU acceleration, however, is not so obvious. And the configuration for GPU containers is a problem that every customer needs to solve.

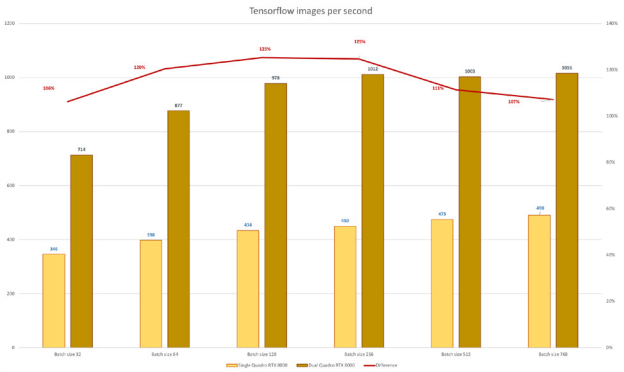
NVIDIA Docker adds two items to standard Docker containers. It includes driver independent CUDA images in the container. And it loads the necessary driver components into the container at launch time.

By doing so, NVIDIA Docker enables portable, GPU accelerated containers. With NVIDIA Docker, you can develop on a Data Science Workstation and deploy across a range of GPU accelerated computing resources.

MORE GRAPHICS MEMORY MEANS MORE PERFORMANCE



DUAL GPU CONFIGURATION WITH LINEAR SCALING



The most interesting performance result is performance scaling with multiple GPUs.

Each of the Quadro RTX 8000 boards has 48 GB of memory. A GPU with a much larger set of memory, like the Quadro RTX 8000, can run workloads with larger batch sizes. A GPU with less memory requires a smaller batch size to run successfully.

Six tests were run changing the batch size parameter from 32 to 768. The result? As the batch size value increased, so did performance. The Big LSTM workload from the NVIDIA examples increased performance three-fold. It showed performance gains for each incremental increase in the batch size parameter.

This is not a head-to-head comparison of memory capacity, but it supports the choice of GPUs with large memory capacity.

For multi-GPU tests, a simple parameter changes allows running tests with a single GPU or with two GPUs. A good scaling result would be around 90%.

In fact, the scaling in this system consistently showed a performance gain above 100%. This can likely be attributed to the dual GPU configuration using NVLink and NVLink's ability to create a larger, flat memory model as well as a high-speed interconnect.

The flat memory model creates one memory space including the 96 GB of graphics memory and the 196 GB of main memory. Both are directly addressasble by the GPU and CPU. Directly addressing a single set of memory with both GPUs and both CPUs creates a more efficient computing environment.

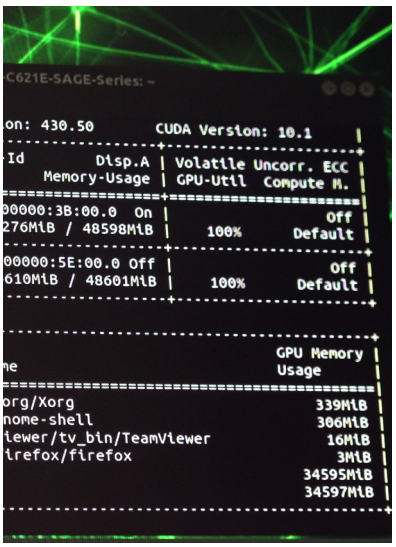
A DATA SCIENCE WORKSTATION MAKES THE IMPOSSIBLE COME TO LIFE

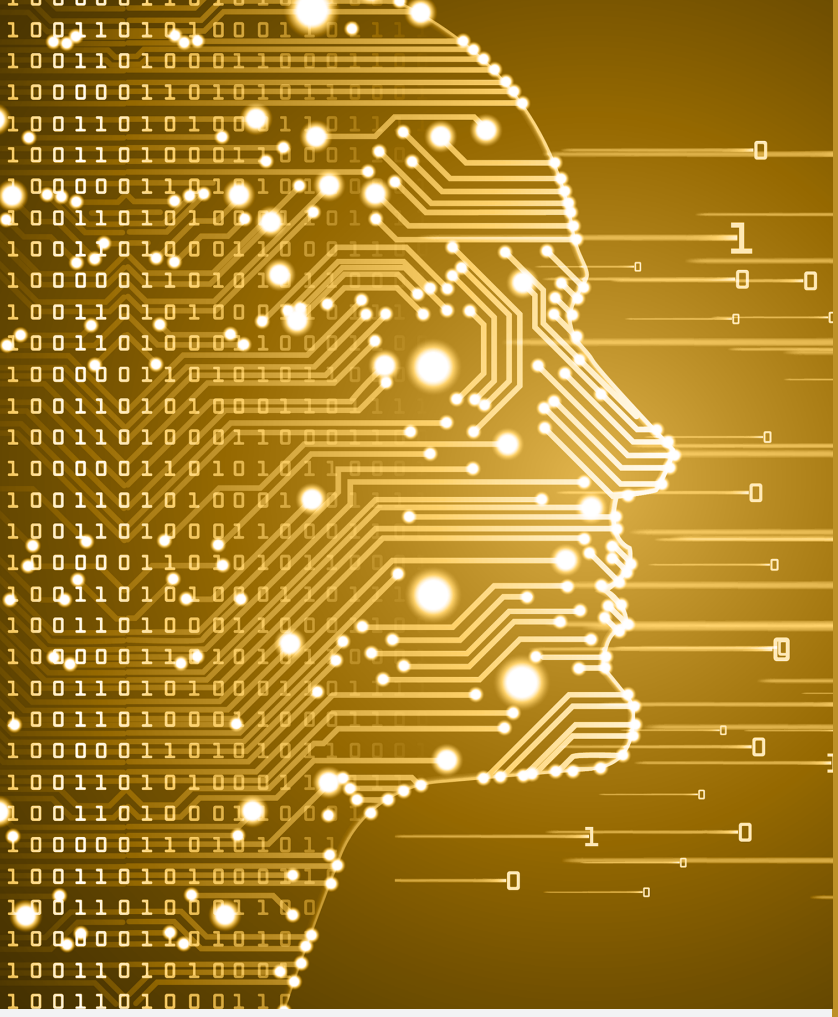
GPU computing allows applications for interactive data analytics with large data bases to manage queries interactively.

OmniSci is a data analytics software that allows users to query databases with billions of lines of data. The ability to tackle such problems with the highly parallel computing of a GPU makes the impossible, possible.

Analyzing a data base with millions of lines runs easily on an NVIDIA Data Science Workstation. The system performance is smooth with only a single GPU being loaded at 25 to 30 percent.

Clearly a Data Science Workstation can handle much more complex workloads.





WHY DO YOU NEED A DATA SCIENCE WORKSTATION?

Data Science and AI technologies are permeating many business sectors and changing the competitive relationships between companies. These technologies impact domains as diverse as image recognition to natural language processing, visualization, video, smart cities, telecommunications, finance, and transportation.

The NVIDIA Data Science Workstation is designed with optimized, enterprise-grade hardware. The technology is balanced to deliver performance in machine learning and big data analytics.

The software stack includes the major tools needed for AI and data science. It is pre-installed and tested so that your team can be productive minutes after booting the system.

A WORKSTATION WITH A PROFESSIONAL PARTNER

The Data Science Workstation used in this analysis was provided by Scan Computers International.

Scan Computers has been developing high-end workstations for data science and AI for years. The company has invested and built a team to support customers at the IT level and also at the project level.

Scan was the first certified NVIDIA Elite Solution Provider in Europe. The company's dedication to this scientific market includes a team with multiple data scientists.

Scan supports customers as their computing needs grow. When projects scale up and you need GPU-servers, then Scan can help. When you want cloud workstation, then Scan can help. When you need HPC servers, then Scan can help. When you need specialized applications, then Scan can help.

They have the years of experience, the in-house expertise, and the product lines that you might need as your computing requirements in AI and Data Science grow.

With the experienced team at Scan, the company can accompany your projects as you progress through each stage of your AI journey.

This report has been written by WSM.

WSM would like to acknowledge the generous support from Scan Computers International. The company provided a Scan 3XS Data Science Workstation as well as technical assistance for testing and investigation.

More information is published on Professional Workstation at www.professional-workstation.com/AI

WSM Professional Marketing