White Paper

XENON AND NVIDIA GPUS ACCELERATE DATA SCIENCE

Graphics Processing Units have transformed data science

Fast, flexible GPU power is critical for successful modern data-science work. The latest solutions enable researchers and students alike to access best-in-class data processing power.

EXECUTIVE SUMMARY

Data-science workflows have traditionally been slow and cumbersome, relying on central processing units (CPUs) to load, filter, and manipulate data, and train and deploy models. Graphics processing units (GPUs) substantially reduce infrastructure costs and provide superior performance for end-to-end data-science workflows using RAPIDS™ open-source software libraries. GPU-accelerated data science can be performed on a laptop, in a data centre, and in the cloud. A single GPU server node can take on the workload of 100 CPU server nodes, which means that replacing CPU-based clusters with GPU-based clusters can accelerate data-science workloads by more than 100x while reducing operational costs and infrastructure complexity.

WHEN GPUS ARRIVE AT THE PROCESSING PARTY, THE DATA TELLS BIGGER STORIES

The past decade has seen data explode as the world has moved beyond plain old 'big data' to data warehouses and deep data lakes. The analysis of the petabytes of data points being collected at an exponential pace is its own fast-growing discipline, data science. In the past five years alone, the average value of businesses deemed to be 'data driven' has increased four-fold as companies devise new ways to apply data-science techniques to derive myriad benefits from the insights that are multiplying in those data lakes.

Just as big data is driving the enterprise, data science is becoming increasingly important for higher education and research institutes. Researchers and students use scientific methods, processes and algorithms to extract knowledge and insights from data. Modern computing casts the multidisciplinary field of data science as an applied branch of statistics, with the ability to apply rigorous analysis to quantitative and qualitative data, using both structured and unstructured data sets.

Scientists use data to better understand the world around us and unlock next-generation discoveries. As one CSIRO scientist said as his own organisation took delivery of a powerful NVIDIA GPU cluster from XENON: "We're drowning in data, and the challenge is to do the analysis with all that data. Having the GPU cluster available gives us the extra horsepower that we need to tackle really significant bioscience problems."

The other point the CSIRO scientist made was that GPU upgrades unlock not just much faster processing power, but also new ways of thinking. "A lot of the time we're not using graphics at all … these GPUs provide a cheap and cheerful speedup that give us a way to tackle problems that we wouldn't have dreamt of tackling. The thing that most excites me about the GPU cluster are all the ideas we haven't thought of yet. By having this cluster around, I know that there will be moments down the track when CSIRO scientists and other researchers will be saying, 'Maybe we could do this...' And that kind of a-ha moment, that 'let's try this', that's where the real excitement comes," he said.

When time is short or budgets need more proof points before they're beefed up, simply upgrading a GPU – even in a laptop – can lead to massive improvements in performance.

Prior to GPU's, data science applications focused on the central processing unit (CPU) architecture to do all the compute. Accelerated computing really gathered speed when people realised that the graphics processors, which were originally designed for parallel processing of graphics outputs, could be used for parallel computing of other datasets.

The fact that GPUs are anchored in parallel computing is what makes them so powerful. CPUs remain essential to computing as they race through a series of tasks requiring lots of interactivity, such as calling up information from a hard drive in response to users' keystrokes. Architecturally, the CPU is composed of just a few cores with lots of cache memory that can handle a few software threads at a time. A GPU is composed of hundreds of cores that can handle thousands of threads simultaneously, enabling GPUs to break complex problems into thousands or millions of separate tasks and work them out at once. That makes them ideal for graphics, where textures, lighting and the rendering of shapes have to be done at once to keep images flying across the screen.

GPUs were designed to render graphics through rapid mathematical calculations, and it's that high-performance processing that makes them such powerful workhorses for data science. GPUs enable AI to learn from images and sound data, processing massive mixed media data sets for these deep learning processes.

THE DATA EXPLOSION

1.7MB of data was created every second by every person during 2020.

90% of the world's data has been created in the past two years alone.

2.5 10^{18} (quintillion) bytes of data are produced by humans every day.

463 EB (exabytes) of data will be generated each day by humans as of 2025.

By the end of 2020,

44 ZB (zettabytes) made up the entire digital universe.

— TechJury.net

A SHORT HISTORY OF GPU COMPUTING FOR DATA SCIENCE

XENON has a long history in GPU computing. A couple of decades ago, XENON worked with NVIDIA to build graphics workstations for customers in Australia, which was what XENON became known for back then.

For more than a decade, XENON has also worked with the NVIDIA community on using GPUs to accelerate scientific workloads. Over that time, NVIDIA has developed products such as the A100, the V100 and the T4. All specialise in accelerating data-science workloads, accelerating AI and deep-learning. These GPUs have been developed with one aim: to accelerate all aspects of data science.

When these GPUs became available in Australia, XENON worked closely with NVIDIA to deliver workshops at universities around the country. Today XENON continues to deliver AI and deep-learning workshops for higher education and research partners.

Among the many customers benefiting from the combination of NVIDIA's technology and XENON's local expertise for GPU upgrades and clusters are CSIRO and CSL.

In 2008, XENON delivered Australia's first GPU HPC cluster for CSIRO. Three subsequent GPU upgrades over the following five years made the BRAGG cluster, in Canberra, the 10th most energy-efficient supercomputer in the world. XENON rolled out progressive upgrades to increase capacity and minimise environmental impact and energy costs.

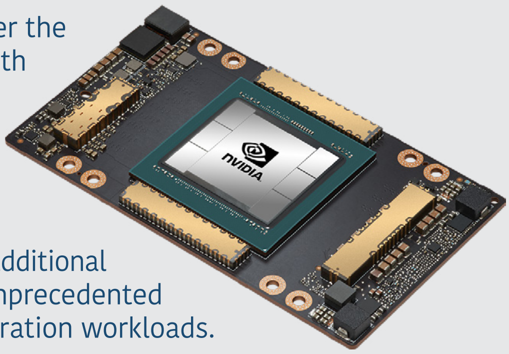
The productivity boost launched as soon as it was installed: The GPU HPC cluster allowed CSIRO data scientists to perform computations in a single morning that previously used to take them weeks.

Biotech giant CSL has also turned to XENON for a HPC solution to support their world-leading research. The brief from CSL was to improve the speed and capabilities of research projects, eliminate processing bottlenecks even as data sets grew, enable faster data analytics for projects such as drug trials, and to build a long-term technology platform to accommodate the increasing demand from burgeoning genomics and biotech data.

XENON's solution was a new HPC cluster integrated with the existing environment, and among other hardware included a GPU server loaded with four NVIDIA V100 GPUs. The solution delivered traditional processing and data analytics processing that ran 50% faster than the previous set-up, taking runs from days and weeks to hours. The increased compute and processing power quickly opened new research opportunities for CSL scientists.

POWER, SPEED, PRECISION

NVIDIA A100 Tensor Core GPUs have double precision Tensor Cores to deliver the biggest leap in HPC performance since the introduction of GPUs. Combined with 80GB of the fastest GPU memory, researchers can reduce a 10-hour, double-precision simulation to under four hours on an A100. HPC applications can also leverage TF32 to achieve up to 11X higher throughput for singleprecision, dense matrix-multiply operations.



For the HPC applications with the largest data sets, A100 80GB's models pack additional memory which delivers up to a 2X throughput. This massive memory and unprecedented memory bandwidth makes the A100 80GB the ideal platform for next-generation workloads.

THE THRIVING ECOSYSTEM AROUND NVIDIA GPUS

Knowledge and success build on each other, and the growth of applications that are being accelerated by NVIDIA GPUs has created a powerful ecosystem. NVIDIA has created a catalogue of GPU-accelerated applications which span every industry.

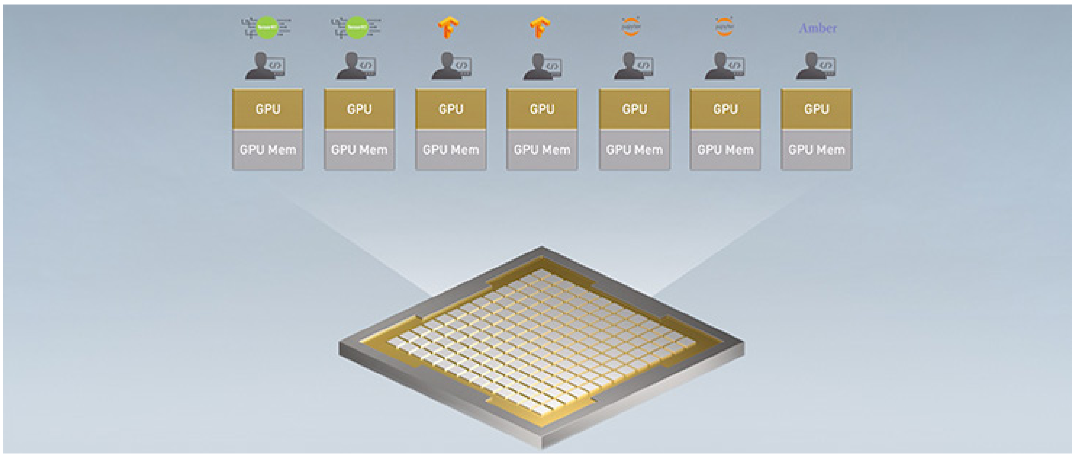
GPUs are now the accelerator of choice for a wide variety of applications, including scientific workloads, deep learning and AI. The strength of the NVIDIA suite of GPUs is that a relatively simple upgrade with one of them can give an existing system an instant and enormous boost to data-processing power, transforming the work of users.

A single GPU server node can take on the workload of 100 CPU server nodes. Replacing CPU-based clusters with GPUbased clusters can accelerate data-science analysis by more than 100x. That kind of leap in processing speed is a game-changer for researchers. Conversely, those without access to this superior resource are genuinely hampered in their work. More and more, the availability of HPC is a key factor for both students and post-grads as they choose the university where they will study.

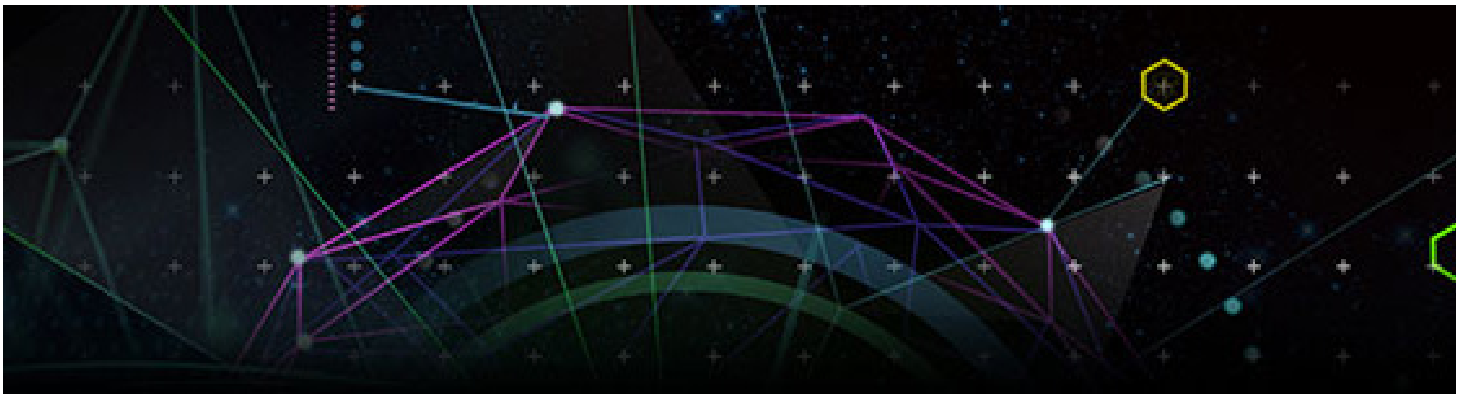
The family of NVIDIA's data science focused GPUs comprises the NVIDIA Ampere Architecture series (A100, A40, A30, A10, T4 and A16). Each provides powerful, flexible GPU computational power to accelerate data-science work across discrete and AI workloads and enable researchers and students alike to quickly realise the benefits of faster data analytics and visualisations. Each specific model is tuned to specific requirements such as highest performance compute (A100), smallest footprint and lowest power consumption (T4), or VDI and virtual or remote GPUs (A16).

The NVIDIA A100 is built with the latest third generation Tensor Cores, and has the unique capability to create up to seven logically discrete GPU instances with NVIDIA's Multi-Instance GPU (MIG). MIG allows administrators to partition the A100 GPU into as many as seven instances per A100, each fully isolated with their own highbandwidth memory, cache and compute cores. MIG can also span A100's, allowing data scientists the ability to combine MIG instances to create a massive GPU with the power of 8 A100's or 56 total MIG instances. MIG enables agility, where right-sized GPU capacity can be assigned for any data science or AI workload, with guaranteed quality of service. This capability also ensures full optimisation of valuable resources and extends the reach of accelerated computing to all users. The availability of such precise allocation means that this valuable resource should never be sitting idle. The A100 MIG feature enables an organisation to share this powerful GPU compute among discrete workloads or individual researchers, or even between different faculties or institutions.

The NVIDIA T4 GPU accelerates diverse data science workloads, including HPC, deeplearning training and inference, machine learning, data analytics, and graphics. Based on the NVIDIA Turing™ architecture and packaged in an energy-efficient 70-watt, small PCIe form factor, T4 is optimised for mainstream computing environments and features multi-precision Turing Tensor Cores and latest RT Cores.



◄NVIDIA's Multi-Instance GPU (MIG)



DIVE INTO THE RAPIDS SOFTWARE STACK

Data science is booming, but the expertise that can help drive faster breakthroughs requires students to have a foundation in various languages and libraries. RAPIDS is an open-source ecosystem of GPU-accelerated libraries which can integrate into existing codes and frameworks allowing existing programs to leverage GPU parallel processing power. In essence, it's a GPU-accelerated platform for data science, and NVIDIA has numerous how-to handbooks, webinars and video tutorials to help beginner students and experienced researchers, plus forums on numerous platforms – Google Groups, Docker Hub, Slack Channel and Stack Overflow – where RAPIDS users can connect with developers and experts to find answers to all manner of technical questions.

The RAPIDS software stack creates a simple way to get data science work done. Using Python and other high level languages, existing code bases can call on behind-the-scenes communication protocols within the data centre to exploit GPU processing power with minimal code changes.

RAPIDS delivers GPU-accelerated machine-learning and data-analytics libraries, deployed on NVIDIA GPU platforms, for maximised data-science productivity, performance and insights. Using RAPIDS helps students and researchers to bring the power of GPU acceleration to their work, and decreases the time to new discoveries.

- RAPIDS accelerates the Python data-science toolchain with minimal code changes and no new tools to learn.
- RAPIDS can accelerate model training and time to deployment.
- The wealth of accelerated apps available to maximise data-centre throughput, utilisation, and efficiency.

NVIDIA has also created the Data Science Starter Kit for Higher Education, designed for university students to offer hands-on training from the NVIDIA Deep Learning Institute.

A GPU upgrade with the RAPIDS tool kit is the quickest path to transforming the performance of computing systems, and in turn the data scientists using them.

RAPIDS

CPU VS GPU

CPU	GPU
Central Processing Unit	Graphics Processing Unit
Several cores	Many cores
Low latency	High throughput
Good for serial processing	Good for parallel processing
Can do a handful of operations at once	Can do thousands of operations at once

XENON – ENABLING DATA SCIENTISTS TO DO GREAT NEW THINGS

XENON delivers for data science teams looking to iterate faster, collaborate more, and develop solutions for tomorrow's problems today. As XENON has for over 25 years.

Explore the resources below, or get in touch with the XENON team to explore how NVIDIA GPU technology can assist your team.

Contact XENON:

- 1300 888 030 (Australia)
- +61 3 9549 1111 (International)
- info@xenon.com.au
- www.xenon.com.au

Talk to a Solutions Architect

RESOURCES

- ***RAPIDS Resources and Starter Kit for Higher Education***
  https://www.nvidia.com/en-us/industries/higher-education-research/data-science/
- ***RAPIDS Info***
  https://www.nvidia.com/en-us/deep-learning-ai/software/rapids/
- ***RAPIDS On-Demand Webinar***
  https://info.nvidia.com/accelerating-data-science-workflows-with-rapids-reg-page

REFERENCES AND RELATED READING

- ***The Data Explosion***
  https://techjury.net/blog/how-much-data-is-created-every-day/#gref
- ***CPU v GPU***
  https://blogs.nvidia.com/blog/2009/12/16/whats-the-difference-between-a-cpu-and-a-gpu/
- ***CSIRO Interview***
  https://youtu.be/x56Wy4cfT00
- ***CSIRO Bragg Story***
  https://xenon.com.au/case-studies/csiro-chooses-xenon-for-supercomputing-upgrade/
- ***GPUs and Data Science***
  https://www.dataversity.net/what-are-gpus-and-why-do-data-scientists-love-them/