THE POWER TO PUSH LIMITS

Accelerated GPUs

Featuring contributions from leading thinkers including:

Todd Mostak
CEO and Co-founder,
OmniSci

Darren Seymour-Russell
Head of Data Science,
Mudano

Amit Marathe
Director of AI & ML,
Inseego Corp

Jared Dame
Director of AI and Data Science,
Z by HP

A NEW GENERATION OF GPUs AND GPU-ACCELERATED SOFTWARE IS SET TO RELEASE A WAVE OF PRODUCTIVITY AND INNOVATION

Many organizations are fighting an uphill battle to handle the scale and velocity of data they are harvesting.

Whether training machine learning (ML) models or crunching large volumes of data, analysts and data scientists on the front line often experience frustration as their effectiveness is hindered by the limitations of compute power.

Unfortunately, these challenges are becoming more acute. On one hand, the pressure to extract real-time value from connected devices and big data is mounting. On the other, the sheer volume of data being created is burgeoning. According to IDC, worldwide data creation will increase to 175 zettabytes by 2025—a tenfold increase on that produced during 2017.[1]

Furthermore, as data scientists increasingly use ML models to capture insights, many data tasks are growing increasingly complex, making training and deployment more difficult. This is putting a strain on the traditional approach to analytics led by CPUs (central processing units) and fueling demand for hardware that will perform ahead of the pace of innovation.

In short, data scientists need greater compute power. And this is driving a major shift to a new era of accelerated hardware—one that is will be defined by the emergence of a new generation of ultra-powerful GPUs (graphics processing units), and GPU accelerated software. This move opens up a new realm of possibilities for data scientists by super-charging productivity, speeding up workflows and unlocking the full potential of ML to accelerate analytics.

The shift to this new era is already well underway. The GPU market alone is predicted to surpass revenue of over $80 billion by 2024, an increase of more than 31 percent from 2018—by which time worldwide GPU industry shipments are anticipated to reach 121,000 thousand units, according to a recent Global Market Insights report.[2] As the rate of adoption increases, so too will the advances in what can be achieved. →

$80 BILLION

The value that the GPU market is predicted to exceed by 2024 according to Global Market Insights.[2]

16,000

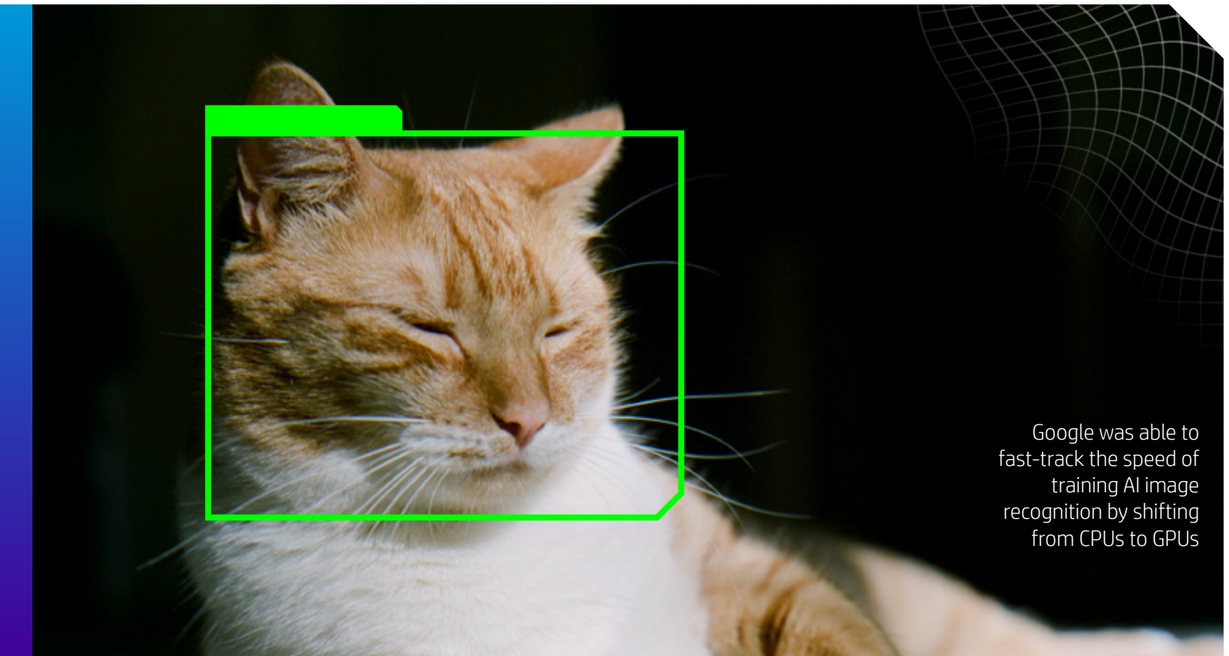
The number of CPUs NVIDIA was able to match in performance with just 48 GPUs for Google's image recognition system. [4]

FAST-TRACKING DATA WORKFLOWS

Much of this trend is being largely driven by the adoption of AI (artificial intelligence), ML innovations in sectors such as healthcare and automotive, and the rise of internet of things (IoT). This is because GPU-accelerated analytics offer teams a way of dealing with the volume and velocity of data associated with big data, enabling data scientists to uncover critical insights faster.

"In the early days of this revolution, people were talking about managing your data at the speed of business," says Kirk Borne, Principal Data Scientist and Executive Advisor at Booz Allen Hamilton. "I think that expression needs to be inverted now. We need business at the speed of data."

GPUs are a response to this challenge.



Google was able to fast-track the speed of training AI image recognition by shifting from CPUs to GPUs

Having started out as accelerators to offload the burden of graphics processing from the CPU in games, GPU architecture proved equally effective for accelerating data science workloads.

They have relatively simple cores, optimized for a floating-point throughput. But there are a large number of them—thousands in high-end chips—so they can process a large set of identical computations in parallel.[3] In contrast, CPUs are general purpose and have just a few complex cores optimized for sequential processing of application logic. This means GPUs have a clear advantage when analyzing huge data sets, where the same calculations need to be performed on all of the data.

For example, when Google started improving its AI systems for image recognition it was using 16,000 CPUs to train an AI to recognize photos of cats. By working with GPU maker NVIDIA, it achieved roughly the same performance with just 48 GPUs.[4]

A big reason for the shift to GPUs is that the needs of data scientists and the types of data they are handling is changing. This presents particular challenges for CPUs. For instance, according to software firm OmniSci, around 80 percent of the data created today contains location-time (spatiotemporal) data which is compute-intensive because it requires rapid analyzing at a granular level. This is difficult to achieve using traditional indexing and pre-aggregation techniques, so most mainstream CPU-led business intelligence and analytics systems struggle to cope with spatiotemporal datasets above low volumes.

> "The productivity yields and gains for, not just the data scientist, but also the organization as a whole, are going to be immense."
>
> **Todd Mostak**, CEO and Co-founder, OmniSci

Also, when handling large volumes of data, traditional CPU architectures require large hardware footprints and consume considerable resources and time in 'wrangling' of data. In contrast, GPU-accelerated data analytics avoid lower-value tasks such as down-sampling, indexing, and cubing by 'ingesting' the entire dataset. The parallelism provided by GPUs means queries can be evaluated in real time without relying on pre-computation.

According to Todd Mostak, CEO and Co-founder of OmniSci, this means data scientists will be able to work "at the speed of curiosity."

"I'm most excited about the agility and the iteration possibilities: where you have →

1000x

Increase in GPU performance that NVIDIA expects to see by 2025.[11]

Logistics firm, UPS, has been using GPU-accelerated analytics to improve delivery efficiency

an idea you don't have to go to IT and wait two weeks to get some rolled up version of your data. You can use the GPUs to query the data, perhaps even churn the records, pull back, interactively visualize and find what you're looking for—the good data, the bad data—and feed it into the ML pipeline, which instead of taking hours or days comes back in minutes. The productivity yields and gains for, not just the data scientist, but also the organization as a whole, are going to be immense."

ACCELERATING INNOVATION

By enabling real-time analytics and freeing up data scientists to spend more time on high-value tasks, advances in GPU technology are also expected to be a catalyst for innovation.

"I think using GPUs for more advanced data science is going to be really big," states Jake Gardner, Enterprise Account Executive at Domo. "It enables more and more people to have reliable access to the hardware that you need to be able to do more of these advanced types of applications like neural networks."

Already, there are lots of early wins. For instance, in logistics, companies like UPS are using GPU-accelerated analytics to extract value from big data supplied by customers, drivers, and vehicles. The firm has created a proprietary tool that uses advanced algorithms to find the optimal routes for delivery trucks in real time, responding to weather conditions, and accidents. The impact was hundreds of millions of dollars saved and an improved customer experience.

Many sectors are ripe for change. In finance, capital market trading firms face challenges around market risk reporting, compounded by new regulatory drivers such as Fundamental Review of the Trading Book (FRTB) reporting. GPU-accelerated database technology could vastly improve processing times for their interactive data and big data challenges. A recent trial by CitiHub Consulting with a tier one investment bank found that GPUs outperformed a CPU configuration from 2.5x for simple queries to more than 400x with complex ones.[5]

Similarly, in insurance, companies currently spend hundreds of millions of dollars per year to process data using CPUs to model actuarial decisions. There are easy wins to be had shifting to a GPU-led approach. Elsewhere in the sector, start-ups like Ravin AI is already automating the time-consuming process of vehicle inspections so that car insurers, dealers, and rental agencies can use drivers' mobile phone images, via GPU-powered algorithms, to evaluate any damage in real time.

Weather forecasting is an example of a notoriously difficult area due to the huge amount of processing power required to model future conditions from weather data. GPU acceleration will play a major role in boosting processing to deliver more accurate forecasting, enabling power companies to more accurately predict electricity demand in different areas, and farmers to better prepare for dramatic shifts in weather.[6]

There are numerous areas in healthcare where GPU acceleration of data science reaps dividends. Analysts looking to uncover hidden patterns and correlations between biological information and effectiveness of pharmaceuticals are seeking more powerful tools beyond legacy medical analysis software.[7] →

In oil and gas exploration, traditional geophysical analysis software struggles to analyze and visualize the large volumes of data needed to determine borehole viability. GPU-accelerated analytics could use a wealth of real-time petrophysical data to enable better visualization capabilities to inform of new drilling opportunities.[8]

There are also ways to disrupt data science processes themselves. According to data scientist Gregory Piatetsky-Shapiro, Founder and President of KDnuggets, GPU-accelerated analytics will enable data scientists to crunch through huge amounts of data and enable vastly more sophisticated AIs that may eventually completely automate many data science processes.[9]

OPTIMIZING FOR AI

NVIDIA in particular is driving innovation in GPU acceleration. The firm effectively pre-empted the current boom by more than a decade through the development of software to allow its GPUs to process the millions of minuscule computations that data science workloads require. The company has also been optimizing its GPUs for deep learning, adding functions to make training and inferencing deep neural networks faster and more energy efficient.

To take advantage of the power of its GPU technology, NVIDIA is driving the data science software ecosystem with NVIDIA RAPIDS, a collection of NVIDIA GPU-accelerated open source libraries and APIs for accelerating end-to-end data science including deep learning, machine learning, and data analytics. The software ecosystem is critical to unlocking the power of GPUs.

This is an ongoing process, since datasets and ML models are going to get larger and more complex over time, which means that GPUs need to keep getting faster and more powerful to keep up.

"Next-generation AI hardware solutions will need to be both more powerful and more cost efficient to meet the needs of the sophisticated training models that are increasingly being used in edge applications," says Chris Nicol, Co-founder and Chief Technology Officer of Wave Computing.[10]

NVIDIA only sees progress in GPUs ramping up. It predicts that new GPU architectures for graphics and AI will continue to increase in compute power such that we can expect to see a thousand times greater performance by 2025.[11]

STRIKING A HARDWARE BALANCE

It is also important to realize that AI and ML accelerators such as GPUs do not replace CPUs. In the vast majority of cases, CPUs are still required to handle the application logic, while the GPU provides the heavy lifting where required. For this reason, the combination of CPUs with GPUs will deliver the best value of system performance, price, and power.

"Ideally, it would be a combination of both—CPU cores as well as GPU cores," says Amit Marathe, Director of AI & ML at Inseego Corp. "Almost every team I know of does data science and machine learning in parallel, you would want to combine the two, put CPU and GPU accelerators in one machine, and provide that as a hybrid solution for all the teams."

Mostak agrees, saying that it should not be seen as a GPU →

> "Next-generation AI hardware solutions will need to be both more powerful and more cost efficient for the models being used in edge applications."
>
> **Chris Nicol,** Co-founder and Chief Technology Officer, Wave Computing

EXPERT VIEW:

Accelerating business decisions with GPUs

Darren Seymour-Russell
Head of Data Science, Mudano

"Computing power is key to deriving insights, and hence advantage, not available to a competitor. So, from a financial services analytics perspective, we see the demand for GPU-accelerated deep learning platforms increasing ever upwards."

Todd Mostak
CEO and Co-founder, OmniSci

"I think that GPUs are going to dramatically transform the field of data science because people will actually come to expect it can be done interactively in real time, and when the data scientist doesn't have to go get a cup of coffee or even sleep before getting to the next step in the machine learning feedback loop, then I think we'll be able to move dramatically faster."

> "The GPU will evolve, because right now it's just good for doing everything generally. In five years, we'll have a specific design for finance and another will be biomedical research."
>
> **Jared Dame,** Director of AI and Data Science, Z by HP

vs CPU choice. "I think you'll see a convergence in compute where more and more people are going to embrace hybrid compute scenarios, where you're going to have some of the workload running on GPUs and some of it running on CPUs," he says.

NVIDIA has introduced a new class of professional workstation—the data science workstation. This platform combines the latest Quadro GPUs, which now include Tensor cores for accelerating AI workloads, with a complete GPU-accelerated software stack to provide an integrated hardware and software solution for data science. The data science workstation gets data science projects up and running quickly, eliminating the time-consuming task of building, and maintaining the multi-application and multi-library software installations required for data science workflows.[12]

According to NVIDIA's Director of Global Business Development, Geoffrey Levene, having such a "personal sandbox" for data work is a real boon for data scientists. "They are finding they can do a week's work in one day with GPU-accelerated workflows."

POWERING THE FUTURE

Next-generation GPUs are likely to be the dominant accelerator for the near future. In the next few years, we may see the GPU market diversify into models optimized for specific markets. "I think what we will see in the future is actually an increase in specialized hardware and software for specialization of AI technologies, streamlining it for various segments, and verticals," says Jared Dame, Director of AI and Data Science at Z by HP. "Finance will have its own hardware/software combination that will streamline its particular workflows and then biomedical sciences will have a slightly different one, and so on between all different segments.

"Take the GPU for example. It will evolve because right now it's just good for doing everything generally. And in five years we'll look at that piece of tech and it will actually have 10 iterations of it. One will be a specific design for finance and then another will be biomedical research, while another will be security, visual image recognition, and natural language processing."

As businesses compete to attract and retain data scientists and analysts, the need to equip talent with the best hardware has never been clearer. And as analytics capabilities continue to improve, so will competition to get the right insights faster.

By using GPUs, these same companies can save millions on computational costs and achieve more accurate results. Keeping up the pace of accelerator innovation will lead to even greater advances.

OUTLOOK

GPU technology has already given a turbo boost for data scientists looking to develop ML models to find better solutions to problems. In the meantime, businesses that want to stay ahead of rivals need their teams to have the right hardware to make the right decisions.

KEY TAKE AWAYS

- GPUs are optimized for the kind of calculations used in ML workloads
- Advances in GPUs will drive performance gains for the next five years
- Businesses need to ensure they have the right hardware to drive decision-making
- Having sufficiently powerful hardware will help retain talented data scientists

Find out more about the benefits of Accelerated GPUs.

LEARN MORE

Sources

1 https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

2 https://www.globenewswire.com/news-release/2019/01/29/1706699/0/en/Graphic-Processing-Unit-GPU-Market-to-cross-80bn-by-2024-Global-Market-Insights-Inc.html

3 https://www.datascience.com/blog/cpu-gpu-machine-learning

4 https://www.wired.co.uk/article/nvidia-artificial-intelligence-gpu

5 https://www.citihub.com/insights/whitepapers/gpu-accelerated-databases-addressing-frtb-risk-results-reporting-and-other-performance-at-scale-challenges-in-financial-services/

6 https://www.hpcwire.com/2019/01/09/ibm-global-weather-forecasting-system-gpus/

7 https://www.omnisci.com/solutions/use-case/clinical-trial-analysis

8 https://www.omnisci.com/solutions/use-case/well-logging-formation-evaluation

9 https://blogs.thomsonreuters.com/answerson/future-of-data-science

10 https://emerj.com/partner-content/artificial-intelligence-hardware-adopt-first/

11 https://www.nextbigfuture.com/2017/06/moore-law-is-dead-but-gpu-will-get-1000x-faster-by-2025.html

12 https://www.digitalengineering247.com/article/the-rise-of-data-science-workstations

4AA7-7927ENW, July 2020