accenture technology

Accenture Labs

GPU Acceleration to Boost Business Outcomes

Achieve growth through next-generation data analytics

High performance. Delivered.

Strategy | Consulting | Digital | Technology | Operations

Introduction

Enterprise growth is fueled in large part by big data. To keep up with the ever-growing volume and velocity, businesses need scalable solutions that can ingest information and produce useful insights. One answer lies in graphic processing units (GPUs), which can process vast amounts of data—both quickly and extremely cost effectively—to power data analytics such as graph analytics, image processing, deep learning and other machine learning techniques.

What problems are GPUs best suited to address? How do GPUs speed artificial intelligence (AI) technology adoption? Should companies access GPU capability from the cloud or upgrade their data centers? Accenture Labs provides some answers to these questions to help businesses take advantage of GPU acceleration.

History of a processing powerhouse

First, a definition: A GPU is a computer hardware chip that performs rapid mathematical calculations. Launched in 2007 for general purpose computation and growing at least 10x or more in performance during the last decade, today's GPUs provide superior processing power, memory bandwidth and efficiency over their central processing unit (CPU) counterparts—up to 50 to 100 times faster in practice on parallel problems such as image processing and machine learning—at a fraction of the cost.

Originally, GPUs were designed and built to render complex images away from the CPU, essentially offloading tasks to a different processing unit. Companies in the gaming industry were the first to harness the power of GPU chips to run visually intensive video games on computers, game consoles and mobile phones. The US military also recognized the potential of GPU capacity and in 2010 linked together more than 1,700 Sony PlayStation 3TM systems in order to more quickly process high-resolution satellite imagery and kick-start AI research.

In response to growing market demand, GPU chip manufacturers began thinking enterprise first, expanding upstream from gaming and downstream from supercomputing. At the same time, businesses increasingly began using the processing speed and capacity of GPUs to accelerate computationally expensive analytics such as simulating outcomes and training models. Today, GPUs are the hardware backbone of nearly all intensive computational applications, especially those that drive AI-related technologies.

Facebook, for example, is invested in GPUs both on the hardware and software fronts. Recently, the company released an open source AI computing solution to more quickly build models for natural language processing and image recognition. Baidu, a search engine provider based in China, uses GPUs to fast-track its work around deep learning, including visual search, speech recognition, language translation, text search and click-through-rate estimation. The company uses GPU clusters on the back-end to train complex models, which can then be sent to an individual's mobile device so that image identification or classification can be run locally. And SAP recently announced it will turn to GPUs to bring AI capabilities to its global customer base.

Meet today's GPU

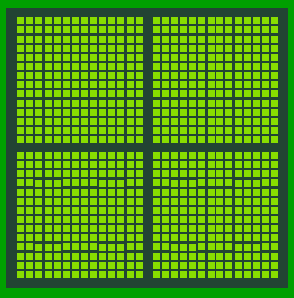
Processing power—GPUs have more cores than CPUs to complete calculations so they can be performed in significantly less time, which improves the ability to build better analytical models and discover insights.

Memory bandwidth—Currently, flagship GPUs (732 GB/s) have significantly more memory bandwidth than CPUs (102 GB/s), enabling GPUs to access and process data at a much faster rate.

Efficiency—GPUs are claimed to be up to 10 times more efficient than CPUs in both performance versus power consumption and performance versus cost.

CPU
Mulitiple Cores

+



GPU
Thousands of Cores

GPU manufacturers have continued to push the boundaries of visual computing, offering platforms that can be used to process data in the enterprise, improve data visualization, power virtual reality applications and, most recently, support self-driving cars.

GPU sweet spots

The problem solving capability of CPUs differs from GPUs so companies will need to continue using both depending on the nature of the task to be performed. Obtaining the average of a very large data set, for instance, is best handled by a CPU, whereas running a complex Monte Carlo simulation for a forecast would be more suited for a GPU.

In terms of enterprise use, GPUs can accelerate operations and reduce costs in areas such as cybersecurity anomaly detection, risk modeling for insurance, shipping route optimization for freight and logistics, and routing decisions for physical infrastructure such as city traffic or water flows. Case in point: In the insurance industry with its domain-specific language, companies currently spend hundreds of millions of dollars per year to process data on CPUs and model actuarial decisions. By using GPUs, these same companies could save millions on computational costs and achieve more accurate modeling results by performing simulations hundreds of times instead of current industry norms.

GPUs are also a win-win solution across enterprise functions. The IT department can save physical infrastructure costs by porting computationally heavy models to GPU, which saves rack space and energy costs for cooling. Data scientists can build more complex data analytics models and visualize massive amounts of data at scale by using GPUs for insight generation. Finally, executives charged with innovation initiatives can use GPUs to introduce AI technologies and other GPU-accelerated applications across the enterprise.

Acquiring GPU capability

Companies ready for GPU acceleration have two options: cloud or on-premise. Every major cloud provider has recognized the demand for GPUs and responded with cost-effective options. Amazon Web Services (AWS) recently announced its EC2 P2 Instances with up to 16 GPUs for "machine learning, high performance databases... and other workloads requiring massive parallel floating point processing power." Microsoft added GPU resources in mid-2016 with the Azure N-Series, a virtualized offering and options to choose from various grades of GPU depending on the task to be completed. IBM Cloud offers GPU resources that are not virtualized while IBM Softlayer has multiple GPU options available in bare metal servers. Finally, Google, though last to offer GPUs in the cloud, will be the first to offer both AMD and NVIDIA GPUs in the cloud starting in early 2017.

For on-premise solutions, companies such as HP and Dell are ramping up offerings for GPU-accelerated servers, which the IT department can purchase to supplement CPU capability in the data warehouse. Another option is to modify existing hardware and add GPUs to a sub-set of servers. Finally, Nvidia is marketing the Tesla P100 and DGX-1 to build very dense servers for high performance computing and deep learning. These "supercomputers in a box" represent a whole new league in GPU capability.

Regardless of how GPUs are sourced, companies can get started by identifying complex and time-consuming data operations and processes that could be GPU accelerated. It will be important to price out how and where to access GPU processing capacity and to determine the return on investment for each approach (i.e., using existing software, building new software, upgrading hardware or enhancing capabilities via the cloud).

GPU vs. CPU

CPUs are required to enable the data processing workhorses of GPUs. However, digital businesses can optimize performance by redirecting CPU-heavy tasks to GPUs. The main differences between the two hardware chips are:

GPU	CPU
Throughput optimized	Latency optimized
Faster at answering complex questions	Faster at answering simple questions
Designed to maximize the performance of the entire job	Designed to maximize the performance of a single task within a job
Process multiple tasks at once	Process tasks one at a time
Currently, a server can have 8 GPUs with ~5,000 cores per GPU for a total of up to 40,000 GPU cores	High-end CPUs can have up to 24 cores; high end servers can have up to 4 CPUs for a total of 96 CPU cores

Conclusion

Businesses looking for that next competitive edge should consider GPU acceleration. Whether used to speed big data processing or prepare for a future of AI-related technologies, the opportunities for generating better outcomes through GPUs are a smart bet.

GPUs power technology advancements

 **Self-driving cars**	Ingest images and make inferences to enable split-second decisions about acceleration, braking and turning in order to maneuver vehicle safely.
 **Graph analytics**	Generate advanced insights and representations of connected networks such as cyber security, telcos, financial trading patterns, social media (Facebook, Twitter), and more.
 **Image processing**	Accurately process millions of images for use in industries such as border control/security, medical x-ray processing, oil field seismic modeling.
 **Deep learning**	Computers draw insights from large datasets using methods similar to human brain's neural network. Applications include image analysis, speech recognition, policy learning, and multimodal data analysis.
 **Simulations**	Conduct longer runs of simulations at a rapid pace to generate more accurate results.
 **Virtual reality**	Use parallel processing capability to quickly render and maintain realistic images (with proper lighting and shading) in order to deliver a believable experience.
 **Machine learning**	Rapidly and cost effectively complete more iterations to land on desired result (expected success rate) and try more types of models to verify results.
 **Advanced Visualization**	Process big data dynamically, depict as interactive visualization and integrate with other datasets in order to explore volume and velocity of data much faster.
 **Molecular modeling, genomics**	Power gene mapping by processing data and analyzing co-variances to understand relationship between different combinations of genes.

Contributors

Keith Kraus
Consultant, Cybersecurity R&D

Josh Patterson
Manager, Cyber Security R&D

Lisa O'Connor
Managing Director Cybersecurity R&D
cybersecuritylab@accenture.com

References

1 http://www.karlrupp.net/2013/06/cpu-gpu-and-mic-hardware-characteristics-over-time/

2 http://www.nvidia.com/object/tesla-servers.html

3 http://www.popsci.com/technology/article/2010-12/air-forces-new-supercomputer-made-1760-playstation-3s

4 https://code.facebook.com/posts/1687861518126048/facebook-to-open-source-ai-hardware-design/

5 http://www.itpro.co.uk/business-intelligence/27314/sap-teams-with-nvidia-to-bring-ai-capabilities-to-its-customers

6 https://aws.amazon.com/ec2/instance-types/p2/

7 https://azure.microsoft.com/en-us/documentation/videos/build-2016-introduction-to-nvidia-gpus-in-azure/; https://azure.microsoft.com/en-us/blog/azure-n-series-preview-availability/

8 http://www.forbes.com/sites/moorinsights/2016/06/20/nvidias-new-gpus-set-a-high-bar-for-hpc-and-deep-learning/#76c6b55a2620

9 http://www.nvidia.com/object/deep-learning-system.html

About Accenture Labs

Accenture Labs invents the future for Accenture, our clients and the market. Focused on solving critical business problems with advanced technology, Accenture Labs brings fresh insights and innovations to our clients, helping them capitalize on dramatic changes in technology, business and society. Our dedicated team of technologists and researchers work with leaders across the company to invest in, incubate and deliver breakthrough ideas and solutions that help our clients create new sources of business advantage. Accenture Labs is located in seven key research hubs around the world: Silicon Valley, CA; Sophia Antipolis, France; Arlington, Virginia; Beijing, China; Bangalore, India; Herzliya, Israel and Dublin, Ireland. The Labs collaborates extensively with Accenture's network of nearly 400 innovation centers, studios and centers of excellence located in 92 cities and 35 countries globally to deliver cutting-edge research, insights and solutions to clients where they operate and live.

About Accenture

Accenture is a leading global professional services company, providing a broad range of services and solutions in strategy, consulting, digital, technology and operations. Combining unmatched experience and specialized skills across more than 40 industries and all business functions—underpinned by the world's largest delivery network—Accenture works at the intersection of business and technology to help clients improve their performance and create sustainable value for their stakeholders. With more than 394,000 people serving clients in more than 120 countries, Accenture drives innovation to improve the way the world works and lives. Visit us at www.accenture.com.