COMPLIMENTS OF kinetica

HOW GPUS ARE DEFINING THE FUTURE OF DATA ANALYTICS

What Every Technology Executive Should Know

ERIC MIZELL AND ROGER BIERY

CONTENTS

INTRODUCTION

After decades of achieving steady gains in price/performance, Moore's Law has finally run its course for CPUs. The reason is: The number of x86 cores that can be placed cost-effectively on a single chip has reached a practical limit, and the smaller geometries needed to reach higher densities are expected to remain prohibitively expensive for most applications.

This limit has given rise to the use of server farms and clusters to scale both private and public cloud infrastructures. But such brute force scaling is also expensive, and threatens to exhaust the finite space, power and cooling resources available in data centers.

Fortunately, for database and big data analytics applications, there is now a more capable and cost-effective alternative for scaling compute performance: the graphics processing unit. GPUs are proven in practice in a wide variety of applications, and advances in their design have now made them ideal for keeping pace with the relentless growth in the volume, variety and velocity of data confronting organizations today.

The purpose of this eBook is to provide an educational overview of how advances in high-performance computing technology are being put to use addressing current and future database and big data analytics challenges. The content is intended for technology executives and professionals, but is also suitable for business analysts and data scientists.

The eBook is organized into 8 chapters:

1. **The Evolution of Data Analytics** provides historical context leading to today's biggest challenge: the shifting bottleneck from memory I/O to compute
2. **GPUs: A Breakthrough Technology** describes how graphics processing units overcome the compute-bound limitation to enable continued price/performance gains
3. **New Possibilities** highlights the many database and data analytics applications that stand to benefit from GPU acceleration
4. **Advanced In-Database Analytics** explains how GPU-accelerated user defined functions (UDFs) now converge AI and BI on one solution for business analysts

5. **Real-time Data Analytics** describes how GPU-accelerated databases can process streaming data in real time, including in on-line analytical processing (OLAP) applications
6. **Interactive Location-Based Analytics** explores the performance advantage GPU databases afford in demanding geospatial applications
7. **Cognitive Computing:** The Future of Analytics provides a vision of how even this, the most compute-intensive application currently imaginable, is now within reach using GPUs
8. **Getting Started** outlines how organizations can begin implementing GPU-accelerated solutions in public, private and/or hybrid cloud architectures

1. The Evolution of Data Analytics

The diagram below shows four distinct stages in the evolution of data analytics since 1990.

Just as CPUs evolved to deliver constant improvements in price/performance under Moore's Law, so too have data analytics architectures.

In 1990, the ***Data Warehouse*** and relational database management system (RDBMS) technologies enabled organizations to store and analyze data on servers cost-effectively with satisfactory performance. Storage Area Networks (SANs) and Network-Attached Storage (NAS) were common in these applications. But as data volumes continued to grow, the performance of this architecture became too expensive to scale.

Circa 2005, the distributed ***Server Cluster*** that utilized direct-attached storage (DAS) for better I/O performance offered a more affordable way to scale data analytics applications. Hadoop and MapReduce, which were specifically designed to take advantage of the parallel processing power available in clusters of servers, became increasingly popular. While this architecture continues to be cost-effective for batch-oriented data analytics applications, it lacks the performance needed to process data streams in real time.

By 2010, the ***In-Memory Database*** became affordable owing to the ability to configure servers with terabytes of low-cost random access memory (RAM). Given the dramatic increase in read/write access to RAM (100 nanoseconds vs. 10 milliseconds for DAS), the improvement in performance was dramatic. But as with virtually all advances in performance, the bottleneck shifted—this time from I/O to compute for a growing number of applications.

This performance bottleneck has been overcome with the recent advent of ***GPU-accelerated*** Compute. As will be explained in Chapter 2, GPUs provide massively parallel processing power that can be scaled both up and out to achieve unprecedented levels of performance and major improvements in price/performance in most database and data analytics applications.

TODAY'S DATA ANALYTICS CHALLENGES

Performance issues are impacting business users

- In-memory database query response times degrade significantly with high cardinality datasets
- Systems struggle to ingest and query simultaneously, making it difficult to deliver acceptable response times with live streaming data

Price/performance gains are difficult to achieve

- Commercial RDBMS solutions fail to scale out cost-effectively
- x86-based compute can become cost prohibitive as data volumes and velocities explode

Solution complexity remains an impediment to new applications

- Frequent changes are often needed to data integration, data models/ schemas and hardware/software optimizations to achieve satisfactory performance
- Hiring and retaining staff with all the necessary skillsets is increasingly difficult—and costly

2. GPUs: A Breakthrough Technology

The foundation for affordable and scalable high-performance data analytics already exists based on steady advances in CPU, memory, storage and networking technologies. As described in Chapter 1, these evolutionary changes have shifted the performance bottleneck from memory I/O to compute.

In an attempt to address the need for faster processing at scale, CPUs now contain as many as 32 cores. But even the use of multi-core CPUs deployed in large clusters of servers can make sophisticated analytical applications unaffordable for all but a few organizations.

A far more cost-effective way to address the compute performance bottleneck today is the graphics processing unit. GPUs are capable of processing data up to 100 times faster than configurations containing CPUs alone. The reason for such a dramatic improvement is their massively parallel processing capabilities, with some GPUs containing nearly 5,000 cores—upwards of 200 times more than the 16-32 cores found in today's most powerful CPUs.

The GPU's small, efficient cores are also better suited to performing similar, repeated instructions in parallel, making it ideal for accelerating the processing-intensive workloads common in today's data analysis applications.

The Evolution of the GPU

As the name implies, GPUs were initially used to process graphics. The first-generation GPU was installed on a separate video interface card with its own memory (video RAM or VRAM). The configuration was especially popular with gamers who wanted high-quality real-time graphics. Over time, both the processing power and the programmability of the GPU advanced, making it suitable for additional applications.

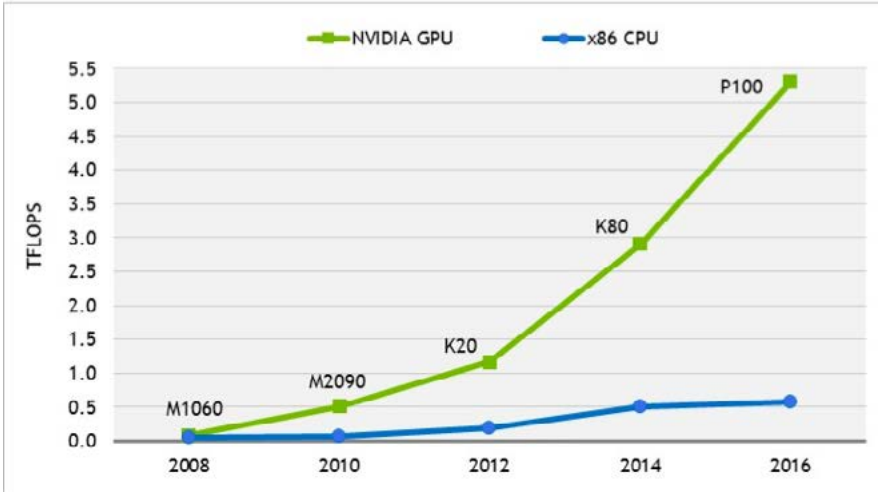
SCALING PERFORMANCE MORE AFFORDABLY

In one application, a simple two-node cluster was able to query a database of 15 billion Tweets and render a visualization in less than a second. Each server was equipped with two 12-core Xeon E5 processors running at 2.6 GHz and two NVIDIA K80 cards for a total of four CPUs and four GPUs.

GPU architectures designed for high-performance computing applications were initially categorized as General-Purpose GPUs. But the rather awkward GPGPU moniker soon fell out of favor once the industry came to realize that both graphics and data analysis applications share the same fundamental requirement for fast floating point processing.

Subsequent generations of these fully programmable GPUs increased performance in two ways: more cores and faster I/O with the host server's CPU and memory. NVIDIA®'s K80 GPU, for example, contains 4,992 cores. Most GPU accelerator cards today utilize the PCI Express bus with a bi-directional bandwidth of 32 gigabytes per second (GB/s) for a 16 lane PCIe interconnect. While this throughput is adequate for most applications, others stand to benefit from NVIDIA's NVLink™ technology, which provides 5 times the bandwidth (160 GB/s) between the CPU and GPU, and among GPUs.

For the latest generation of GPU cards, the memory bandwidth is significantly higher, with rates up to 732 GB/s. Compare this bandwidth to the 68 GB/s in a Xeon E5 CPU at just over twice that of a PCIe x16 bus. The combination of such fast I/O serving several thousand cores enables a GPU card equipped with 16 GB of memory to achieve single-precision performance of over 9 TeraFLOPS (floating point operations per second).



The latest generation of GPUs from NVIDIA contain upwards of 5,000 cores and deliver double-precision processing performance of 5 TeraFLOPS. Note also the relatively minor performance improvement over time for multi-core x86 CPUs, and how it is now flattening. (Source: NVIDIA)

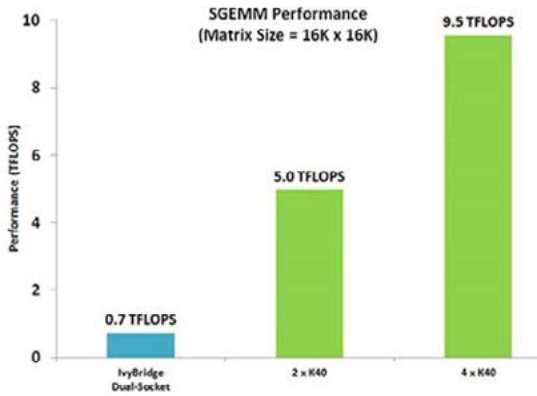
"Small" vs. "Big" Data Analytics

The relatively small amount of memory on a GPU card compared to the few terabytes now supported in servers has led some to believe that GPU acceleration is limited to "small data" applications. But that belief ignores two practices common in "big data" applications.

The first is that it is rarely necessary to process an entire dataset at once to achieve the desired results. Data management in tiers across GPU VRAM, system RAM and storage (DAS, SAN, NAS, etc.) is capable of delivering virtually unlimited scale for big data workloads. For machine learning, for example, the training data can be streamed from memory or storage as needed. Live streams of data coming from the Internet of Things (IoT) or other applications, such as Kafka or Spark, can also be ingested in a similar, "piecemeal continuous" manner.

The second practice is the ability to scale GPU-accelerated configurations both up and out. Multiple GPU cards can be placed in a single server, and multiple servers can be configured in a cluster. Such scaling results in more cores and more memory all working simultaneously and massively in parallel to process data at unprecedented speed. The only real limit to potential processing power of GPU acceleration is, therefore, the budget.

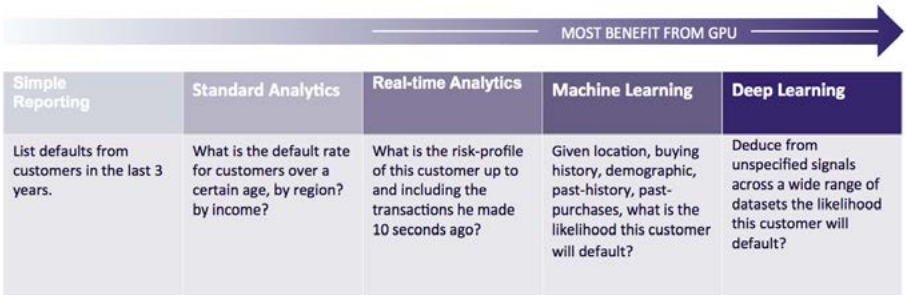
But whatever the available budget, a GPU-accelerated configuration will always be able to deliver more FLOPS per dollar. CPUs are expensive—far more expensive than GPUs. So whether in a single server or a cluster, the GPU delivers a clear and potentially substantial price/performance advantage.



GPUs are able to scale up performance in a nearly linear manner, as shown by these single-precision floating general matrix multiply (SGEMM) benchmark tests. (Source: NVIDIA)

3. New Possibilities

As shown in the diagram below, the benefit from the boost afforded by GPU acceleration is different for different applications. In general, the more processing-intensive the application, the greater the benefit.



While most data analytics applications stand to benefit from the GPU's price/performance advantage, those requiring the most processing stand to benefit the most.

This chapter describes how GPU acceleration can be used to improve both the performance and price/performance of a wide variety of database and data analytics applications. The next three chapters focus on those applications that stand to benefit the most.

FAST/FULL TEXT ANALYTICS AND NATURAL LANGUAGE PROCESSING

A common requirement in many data analytics applications is text analytics and NLP, and this need serves as a good example of the complementary nature of GPU acceleration. Its massively-parallel processing enables the GPU to perform these and other text analytics in real-time on large datasets:

- Exact Phrases
- AND/OR
- Wildcards
- Grouping
- Fuzzy Search
- Proximity Search
- Ranges of Numbers

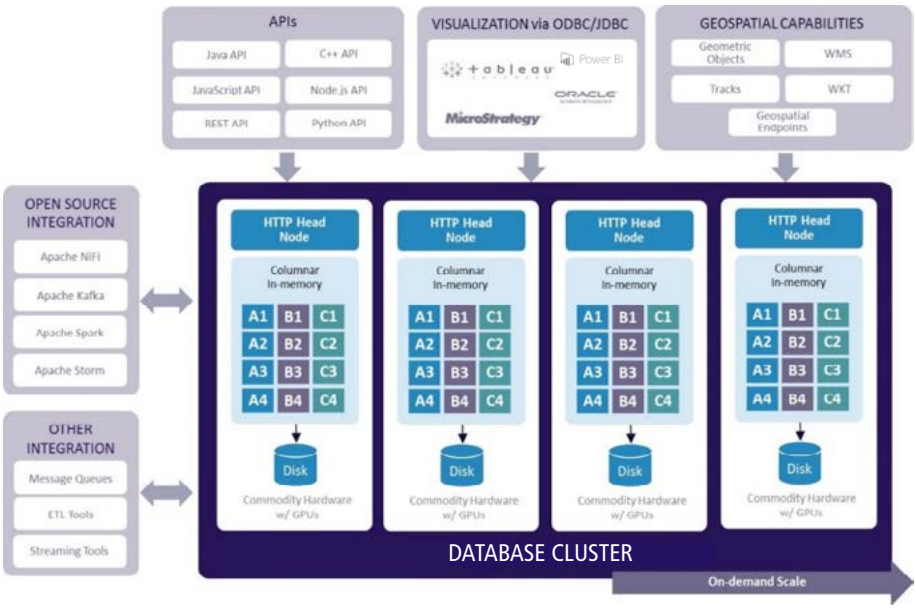
Designed for Interoperability and Integration

Although different GPU-based database and data analytics solutions offer different capabilities, all are designed to be complementary to and/or to be integrated with existing applications and platforms. Some of the more common techniques used are outlined here.

Beginning with the hardware, virtually all GPU-based solutions operate on commonly-used industry-standard servers equipped with x86 CPUs, enabling the configuration to be scaled cost-effectively both up and out to achieve the desired performance.

Scaling up usually involves adding more/faster GPUs and/or video RAM. Performance in servers containing multiple GPU cards can be scaled up even further using NVLink (described in Chapter 2), which offers 5x the bandwidth available in a 16 lane PCIe bus.

Scaling out involves simply adding more servers in a cluster, which can also be done in a distributed configuration to enhance reliability.



GPU-accelerated solutions have open architectures, enabling them to be integrated easily into a wide variety of analytical applications.

For the software, most GPU-based solutions employ open architectures to facilitate integration with virtually any application that stands to benefit from higher and/or more cost-effective performance. Potential applications range from traditional relational databases and artificial intelligence, including machine learning and deep learning, to those requiring real-time analysis of streaming data or complex event processing—increasing common with the Internet of Things.

GPU solutions often serve in a complementary role, for example, as a fast query for batch mode MapReduce jobs. The ultra-low-latency performance also makes GPU-accelerated solutions ideal for those for those applications that require simultaneous ingest and analysis of a high volume and velocity of streaming and large, complex data.

Recognizing that GPUs are certain to be utilized in many mission-critical applications, many solutions are also designed for both high availability and robust security. High availability capabilities may include data replication with automatic failover in clusters of two or more servers, with data integrity being provided by disk-based persistence on individual servers.

For security, support for user authentication, and role- and group-based authorization help make GPU acceleration suitable for applications that must comply with security regulations, including those requiring personal privacy protections. These enhanced capabilities virtually eliminate any risk of adoption for organizations in both public and private cloud infrastructures.

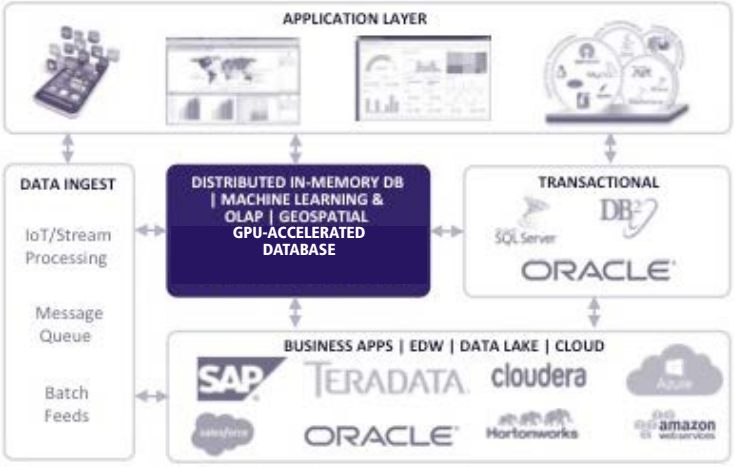
Some GPU-based solutions are implemented as in-memory databases, making them similar in functionality to other databases that operate in memory. What makes the GPU-accelerated database different is how it manages the storage and processing of data for peak performance in a massively parallel configuration.

OPEN FOR BUSINESS

Most GPU-accelerated databases have open designs, enabling them to support a broad range of data analytics applications, environments and needs. Examples of open design elements include:

- Connectors to simplify integration with the most popular open-source frameworks, including Accumulo, H2O, HBase, Kibana, Kafka, MapReduce, NiFi, Spark and Storm
- Drivers for ODBC/JDBC to afford seamless integration with existing visualization and business intelligence tools, such as Caravel, PowerBI and Tableau
- APIs to enable binding with commonly-used programming languages, including SQL, C++, Java, JavaScript, Node.js and Python
- Support for the Web Map Service (WMS) protocol for integrating the georeferenced map images used in geospatial visualization applications

Data is usually stored in system memory in vectorized columns to optimize processing across all available GPUs. Data is then moved as needed to GPU VRAM for all calculations, both mathematical and spatial, and the results are returned to system memory. With smaller data sets and live streams the data can be stored directly in the GPU's VRAM to enable faster processing. Whether stored in system memory or VRAM, all data can be persisted to hard disks or solid state drives to ensure no data loss.



The GPU-accelerated in-memory database becomes a "speed layer" capable of providing higher performance for data analytics application.

4. Advanced In-Database Analytics

Most organizations believe that data analytics is not currently adding as much value it should. Part of the reason is that data scientists often do not understand the business as well as the business analysts do. The flip side of that reason, of course, is that business analysts are normally ill-equipped to conduct their own analyses—at least to the degree needed to bring out the value that is hidden deep within the data.

Ideally, teams consisting of business analysts and data analysts (or scientists) would be able to mine the data residing in existing databases without the need to extract, transform and load it somewhere else. They should also be able to make both simple and complex queries, run simulations, apply different models, create advanced calculations and more—all without requiring much or any programming. And recognizing that many attempts fail to produce meaningful results, they should be able to tweak existing analyses continuously and run new ones frequently, which both require getting results quickly.

> *"With so much raw compute power, GPU-accelerated databases offer extraordinary performance without needing to define schemas around pre-determined questions or spending weeks tuning."*
>
> **— James Curtis, Senior Analyst, Data Platforms and Analytics at 451 Research**

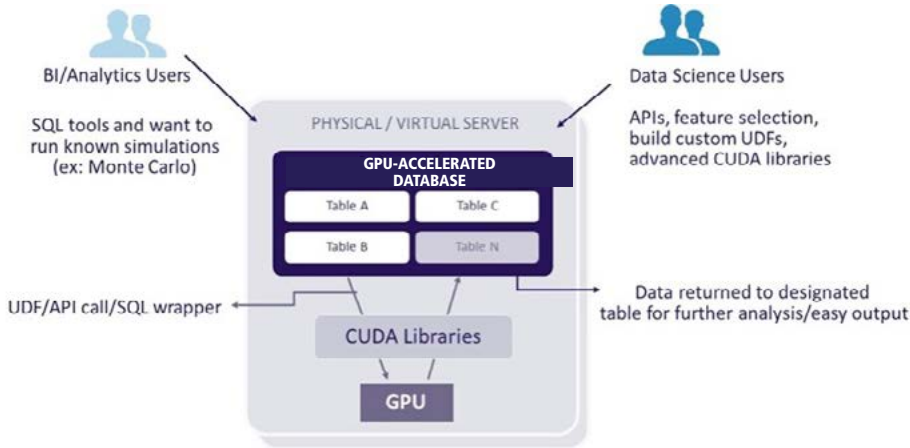
A few years ago, this ideal would have been impossible owing to the requirement for and complexity of programming GPUs. But that all changed with the advent of solutions supporting user-defined functions (UDFs) that make the massively parallel processing power of GPUs readily accessible to the business analysts who best understand the data and its potential value.

Support for user-defined functions in database and data analytics applications gives business analysts and developers open access to the GPU's full power, enabling them to:

- Converge machine learning, deep learning and business intelligence into a single solution providing fast, easy and rich data insights
- Create and modify custom algorithms and libraries
- Leverage third-party code (e.g. NVIDIA CUDA® or other) by implementing orchestration hooks deployed via REST APIs

- Utilize native API bindings in C/C++, Java and Python
- Call the endpoints via the exposed API and specify input table/output table
- Implement a true compute-to-grid application that minimizes or even eliminates the need to move data
- Integrate user-defined processes with the in-memory database's high-speed inter-process communications (IPC) layer
- Use arbitrary binaries to receive table data, perform arbitrary computations, and save the output to a global table in a distributed manner

The open access afforded by user-defined functions could be used, for example, to enable machine learning/artificial intelligence libraries, such as TensorFlow, BIDMach, Caffe and Torch, to run in-database alongside and converged with business intelligence workloads. The result is, in effect, the democratization of data science.

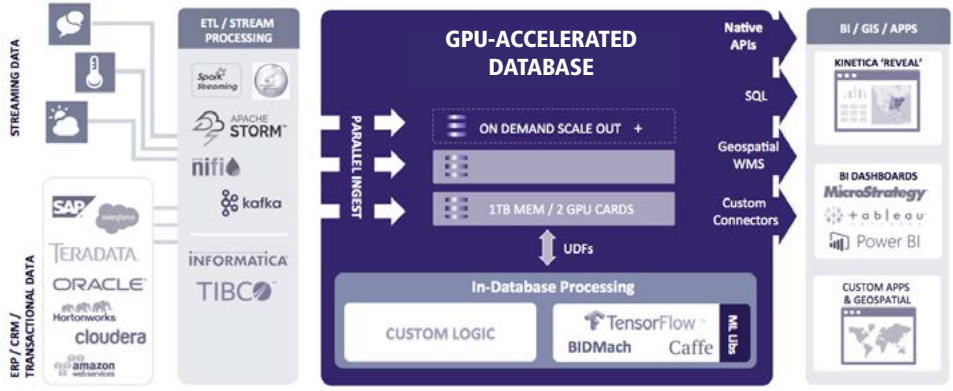


User-defined functions give business analysts easy access to high-performance data analytics, and also give data scientists more options for creating advanced algorithms and libraries.

5. Real-time Data Analytics

Live data can have enormous value, but only if it can be processed quickly enough. Its potential value has led to creation of both proprietary and open source software purpose-built for analyzing streaming data. Without the processing power required to ingest and analyze these streams in real time, however, organizations risk missing out on this opportunity in two ways. One is that the solution will be limited to a relatively low volume and velocity of data. The second is that the results will come too late to have real value.

The need to analyze live data in real time, often coupled with data at rest, has become universal. Some organizations have rather obvious sources of streaming data involving mobile assets, financial transactions, point of sale systems, etc. But every organization has a data network, a website, inbound and outbound phone calls, heating and lighting controls, machine logs, a building security system, and other infrastructure—all of which are continuously generating data that holds potential value. And with the Internet of Things, or as some pundits claim, the Internet of Everything (IoE), the sources of streaming data are destined to proliferate.



A GPU-accelerated database can not only ingest, process, and query streaming data in real time but also converge AI and BI workloads.

Peak performance is achieved by "pinning" the entire dataset in the GPU's ultra-low latency video VRAM. Such a configuration can be scaled in two ways:

1. by adding more GPU cards in the server(s); and
2. by adding more servers to a cluster.

The "Small" vs. "Big" Data Analytics section in Chapter 2 also noted that it is rarely necessary to process an entire dataset all at once to achieve the desired results. For example, data can be streamed directly to VRAM, and the output can then be moved into system RAM and/or be persisted to disk.

The ability to perform advanced data analytics in real time is a common requirement in on-line analytical processing (OLAP) applications. For this reason, some GPU-accelerated databases also support standards like SQL-92 and BI tools, as well as the high availability and robust security often required in such applications.

THE MANY DIMENSIONS OF GEOSPATIAL DATA

GPU-accelerated databases are ideal for processing geospatial data in real time, which like the universe itself, exists in space-time with four dimensions. The three spatial dimensions can utilize native object types based on vector data (points, lines and polygons/shapes) and/or raster imagery data. The latter is typically utilized by BaseMap providers to generate the map overlay imagery used in interactive location-based applications.

The many different functions used to manipulate geospatial data, many of which operate in all four dimensions, create additional processing workloads ideally fitted to GPU-accelerated solutions. Examples of these functions include:

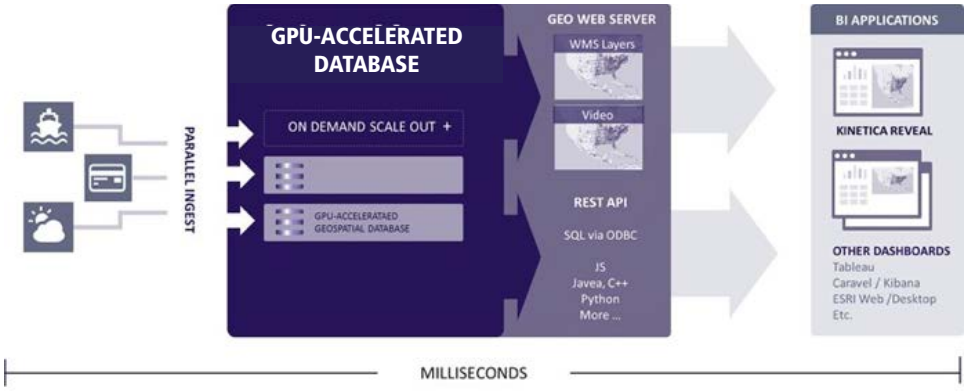
- Filtering by area, attribute, series, geometry, etc.
- Aggregation, potentially in histograms
- Geo-fencing based on triggers
- Generating videos of events
- Creation of heat maps

6. Interactive Location-Based Analytics

Just as most organizations now have a need to process at least some data in real time, most also have a growing need to somehow integrate location into data analytics applications.

Location-based analytics normally utilize a geospatial database. These databases evolved in an environment where the datasets were relatively small, the data was fairly static, and the applications were rarely required to render results in real time. Today, geospatial data sources produce quite detailed datasets that are constantly being updated, and users are demanding analytics that are fully interactive.

Given its roots in graphics processing, it should come as no surprise that the GPU is especially well-suited to handling the real-time geospatial computations required for interactive location-based analytics. With its massively parallel processing, the GPU makes it possible for databases to ingest, analyze and render billions of geospatial data points in human-interactive time.



The GPU-accelerated database is ideally-suited for the interactive location-based analytics that are becoming increasingly common.

The ability to interact with geospatial data in real time gives business analysts the power to make better decisions faster. Support for industry standards makes it easy to integrate data from major mapping providers, including Google, Bing, Esri and MapBox. And with some solutions, users can now simply drag and drop analytical applets, data tables and other "widgets" to create and modify customized dashboards.

Companies in a Variety of Industries Benefit from GPU- Accelerated Solutions

Here are just a few use cases.

GLAXOSMITHKLINE (GSK) finds that during the drug development process GPU-accelerated database can speed up simulations of chemical reactions. The chemical reaction data is distributed over a large number of nodes. By subdividing the reactions, GSK can perform the simulation much faster and significantly reduce the time to develop new drugs. Researchers can use a traditional language such as SQL, making it possible to run an analysis in a traditional relational environment first, and then in the GPU-accelerated database if the workload requires a more computationally-intensive environment, all without having to do a lot of tooling and rework.

MAJOR HEALTHCARE PROVIDER is using a GPU-accelerated data warehouse and analytics for reducing pharmacy fraud, as well as for improving Patient 360 with dynamic geospatial analysis and healthcare IoT data.

A BIG UTILITY is using a GPU-accelerated database for Predictive Infrastructure Management. The GPU database operates as an agile layer to monitor, manage, and predict infrastructure health. GPU acceleration allows the utility to analyze, model and ingest multiple data feeds, including location data for their field deployed assets, into a single centralized data "store."

GLOBAL RETAILER One of the world's largest retailers is using a GPU-accelerated analytics database to optimize their supply chain and inventory. The in-memory database accelerated by GPUs is used to consolidate information about their customers, including sentiment analysis from social media, buying behavior, and online and in-store purchases. The retailer's analysts can now achieve sub-second results on queries that used to take hours. They use the database to correlate that data with weather, point-of-sale systems, and wearable devices in order to build the most accurate view of their customers.

THE UNITED STATES POSTAL SERVICE (USPS) is the single largest logistic entity in the country, moving more individual items in four hours that the combination of UPS, FedEx, and DHL move all year, making daily deliveries to more than 154 million addresses using hundreds of thousands of vehicles. To gain better visibility into operations, every mail carrier now carries a device that scans packages and emits exact geographic location every minute. USPS needed to be able to capture and use this data to improve various aspects of its massive operation, including improving carriers' route efficiency.

USPS selected a GPU-accelerated analytics database to support 15,000 concurrent users and analyze data from over 200,000 scanning devices to reallocate resources based upon personnel, environmental, and seasonal data. The in-memory database provides USPS managers and analysts with the capability to instantly analyze their areas of responsibility via dashboards and to query it as if it were a relational database. As a result, the USPS has improved their end-to-end business process performance while reducing costs.

7. Cognitive Computing: The Future of Analytics

Cognitive computing, which seeks to simulate human thought and reasoning in real time, could be considered the ultimate goal of information technology, and IBM's Watson supercomputer has demonstrated that this goal can indeed be achieved with existing technology.

The real question is: When will cognitive computing become practical and affordable for most organizations?

With the advent of the GPU, the Cognitive Era of computing is now upon us. Converging business intelligence with artificial intelligence and other analytical processes in various ways that makes real-time, human-like intelligence a reality. Such "speed of thought" analyses would not be practical—or even possible—were it not for the unprecedented performance afforded by massively parallel processing of in-memory data stores.

THE OPEN GEOSPATIAL CONSORTIUM

Most graphical information system (GIS) databases support the standards being advanced by the Open Geospatial Consortium (OGC), and a growing number of GPU-based databases are also now supporting these standards.

Some standards specify how GIS images are converted to a common format, such as PNG, and are then wrapped and transported via Web Services that utilize JSON and/or XML. Three such standards currently exist: Web Mapping Service (WMS), Keyhole Markup Language (KML) and Web Feature Service (WFS).

Others standards specify a common format for importing, storing and exporting GIS data. These include Well Known Text (WKT) and Well Known Binary (WKB), which are now used in virtually all GIS databases, and the file-based Shapefile (SHP).

The GPU's Role in Cognitive Computing

If cognitive computing is not real-time, it's not really cognitive computing. After all, without the ability to chime in on Jeopardy! before its opponents did (sometimes before the answer was read fully), Watson could not have scored a single point, let alone win. And the most cost-effective way to make cognitive computing real-time today is to use GPU acceleration.

Cognitive computing applications will need to utilize the full spectrum of analytical processes-business intelligence, artificial intelligence, machine learning, deep learning, expert systems, natural language processing, text search and analytics, pattern recognition, and more. Every one of these processes can be accelerated using GPUs. In fact, its thousands of small, efficient cores make GPUs particularly well-suited to parallel processing of the repeated similar instructions found in virtually all of these compute-intensive workloads.

Cognitive computing servers and clusters can be scaled up and/or out as needed to deliver whatever real-time performance might be required—from sub-second to a few minutes. Performance can be further improved by using algorithms and libraries optimized for GPUs.

By breaking through the cost and other barriers to achieving performance on the scale of a Watson supercomputer, GPU acceleration is ushering in the cognitive computing for many organizations.

8. Getting Started

GPU acceleration delivers both performance and price/performance advantages over configurations containing only CPUs in most database and data analytics applications.

From a performance perspective, GPU acceleration makes it possible to ingest, analyze, and visualize large, complex, and streaming data in real time. In both benchmark tests and real-world applications, GPU-accelerated solutions have proven their ability to ingest billions of streaming records per minute, and perform complex calculations and visualizations in mere milliseconds. Such an unprecedented level of performance will help make even the most sophisticated applications, including cognitive computing, a practical reality. And the ability to scale up and/or out enables performance to be increased incrementally and predictably—and affordably—as needed.

From a purely cost perspective, GPU acceleration is equally impressive. The GPU's massively parallel processing can deliver performance equivalent to a CPU-only configuration at 1/10th the hardware cost, and 1/20th the power and cooling costs. For example, the U.S. Army's Intelligence & Security Command (INSCOM) unit was able to replace a cluster of 42 servers with a single GPU-accelerated server in an application with over 200 sources of streaming data that produces over 100 billion records per day.

But of equal importance is that the GPU's performance and price/performance advantages are now within reach of any organization. Open designs make it easy to incorporate GPU-based solutions into virtually any existing data architecture, where they can integrate with both open source and commercial data analytics frameworks.

GPUS IN THE PUBLIC CLOUD

The availability of GPUs in the public cloud makes it even more affordable and easier than ever to get started. As of this writing, Amazon and Nimbix have begun deploying GPUs, Microsoft's offering is in preview, and Google will soon equip its Cloud Platform with GPUs. Such pervasive availability of GPU acceleration in the public cloud is welcomed news for those organizations who want to get started without having to invest in hardware.

With purpose-built GPU solutions, the potential gain can be quite literally without the pain normally associated with the techniques traditionally used to achieve satisfactory performance. This means no more need for indexing or redefining schemas or tuning/tweaking algorithms, and no more need to ever again pre-determine queries in order to be able to ingest and analyze data in real time, however the organization's data analytics requirements might change.

As with anything new, of course, it is best to research your options and choose a solution specifically designed to take advantage of the GPU, meets enterprise requirements, and can scale as you need. You can start with a pilot project to gain familiarity with the technology and assess its potential, and experience first-hand the processing power of the GPU. You should then be able to fully appreciate the raw power and real potential of a GPU-accelerated database.

About the Authors

ERIC MIZELL

Eric is the VP of Global Solution Engineering at Kinetica. Prior to Kinetica, Eric was the Director of Solution Engineering for Hortonworks, a distributor of Apache Hadoop. Earlier in his career, Eric was both a Director of Field Engineering and a Solutions Architect for Terracotta, a provider of in-memory data management and big data solutions for the enterprise. He started his career in systems and software engineering roles at both McCamish Systems and E/W Group. Eric holds a B.S. in Information Systems from DeVry University.

ROGER BIERY

Roger is President of Sierra Communications, a consultancy firm specializing in computer networking. Prior to founding Sierra Communications, Roger was Vice President of Marketing at Luxcom and a Product Line Manager at Ungermann-Bass, where he was accountable for nearly one third of UB's total revenue and had systems-level strategic planning responsibility for the entire Net/One family of products. Roger began his career as computer systems Sales Representative for Hewlett-Packard after graduating Magna Cum Laude from the University of Cincinnati with a B.S. in Electrical Engineering.

Notes

DON'T SEND A CPU TO DO A GPU'S JOB

Complex analytics are hard work for a CPU. That's why Kinetica was designed from the ground up to leverage the parallel processing power of GPUs. Kinetica delivers **100x performance improvements** over leading in-memory analytics databases and is ideal for advanced analytics, machine learning and deep learning workloads.

Find out more at kinetica.com

kinetica