O'REILLY®

Introduction to GPUs for Data Analytics

Advances and Applications for Accelerated Computing

Eric Mizell & Roger Biery

Introduction to GPUs for Data Analytics

Advances and Applications for Accelerated Computing

Eric Mizell and Roger Biery

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Introduction to GPUs for Data Analytics

by Eric Mizell and Roger Biery

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://oreilly.com/safari*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editor: Shannon Cutt
Production Editor: Justin Billing
Copyeditor: Octal Publishing, Inc.

Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Rebecca Demarest

September 2017: First Edition

Revision History for the First Edition

2017-08-29: First Release

See *http://oreilly.com/catalog/errata.csp?isbn=9781491998038* for release details.

While the publisher and the authors have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the authors disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

978-1-491-99801-4

[LSI]

Table of Contents

Introduction

After decades of achieving steady gains in price and performance, Moore's Law has finally run its course for CPUs. The reason is simple: 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 it threatens to exhaust the finite space, power, and cooling resources available in data centers.

Fortunately, for database, big data analytics, and machine learning applications, there is now a more capable and cost-effective alternative for scaling compute performance: the *graphics processing unit*, or GPU. 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 book is to provide an educational overview of how advances in accelerated 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 it is also suitable for business analysts and data scientists.

The ebook is organized into eight chapters:

- Chapter 1, *The Evolution of Data Analytics* provides historical context leading to today's biggest challenge: the shifting bottleneck from memory I/O to compute.
- Chapter 2, *GPUs: A Breakthrough Technology* describes how graphics processing units overcome the compute-bound limitation to enable continued price and performance gains.
- Chapter 3, *New Possibilities* highlights the many database and data analytics applications that stand to benefit from GPU acceleration.
- Chapter 4, *Machine Learning and Deep Learning* explains how GPU databases with user-defined functions simplify and accelerate the machine learning/deep learning pipeline.
- Chapter 5, *The Internet of Things and Real-Time Data Analytics* describes how GPU-accelerated databases can process streaming data from the Internet of Things and other sources in real time.
- Chapter 6, *Interactive Location-Based Intelligence* explores the performance advantage GPU databases afford in demanding geospatial applications.
- Chapter 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.
- Chapter 8, *Getting Started* outlines how organizations can begin implementing GPU-accelerated solutions on-premise and in public, private, and hybrid cloud architectures.

CHAPTER 1
The Evolution of Data Analytics

Data processing has evolved continuously and considerably since its origins in mainframe computers. Figure 1-1 shows four distinct stages in the evolution of data analytics since 1990.

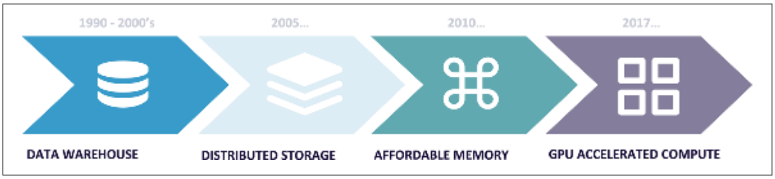


Figure 1-1. Just as CPUs evolved to deliver constant improvements in price/performance under Moore's Law, so too have data analytics architectures

In the 1990s, *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. Although 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 versus 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 is explained in Chapter 2, GPUs provide massively parallel processing power that we can scale both up and out to achieve unprecedented levels of performance and major improvements in price and performance in most database and data analytics applications.

Today's Data Analytics Challenges

Performance issues are affecting 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 of the necessary skillsets is increasingly difficult—and costly

CHAPTER 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 noted 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 multicore CPUs deployed in large clusters of servers can make sophisticated analytical applications unaffordable for all but a handful of organizations.

A far more cost-effective way to address the compute performance bottleneck today is the graphics processing unit (GPU). 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 6,000 cores—upwards of 200 times more than the 16 to 32 cores found in today's most powerful CPUs. For example, the Tesla V100—powered by the latest NVIDIA Volta GPU architecture, and equipped with 5,120 NVIDIA CUDA cores and 640 NVIDIA Tensor cores—offers the performance of up to 100 CPUs in a single GPU.

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.

Scaling Performance More Affordably

In one application, a simple two-node cluster was able to query a GPU database containing 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.

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.

GPU architectures designed for high-performance computing applications were initially categorized as General-Purpose GPUs (GPGPUs). But the rather awkward GPGPU moniker soon fell out of favor when 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 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. And most GPU accelerator cards now utilize the PCI Express bus with a bidirectional bandwidth of 32 GBps for a 16-lane PCIe interconnect. Although this throughput is adequate for most applications, others stand to benefit from NVIDIA's NVLink technology, which provides five times the bandwidth (160 GBps) between the CPU and GPU, and among GPUs.

For the latest generation of GPU cards, the memory bandwidth is significantly higher, as illustrated in Figure 2-1, with rates up to 732 GBps. Compare this bandwidth to the 68 GBps 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 VRAM to achieve single-precision performance of over 9 teraFLOPS (floating-point operations per second).

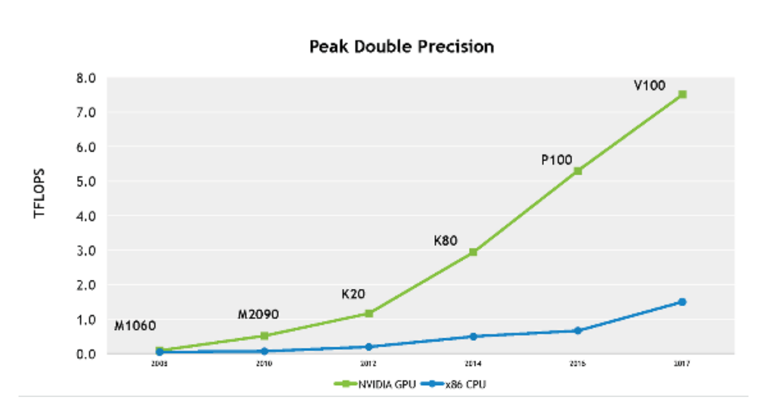


Figure 2-1. The latest generation of GPUs from NVIDIA contain upwards of nearly 6,000 cores and deliver peak double-precision processing performance of 7.5 TFLOPS; note also the relatively minor performance improvement over time for multicore x86 CPUs (source: NVIDIA)

"Small" Versus "Big" Data Analytics

The relatively small amount of VRAM on a GPU card compared to the few terabytes of RAM 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 (direct-attached storage [DAS], Storage Area Networks [SAN], Network-Attached Storage [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.

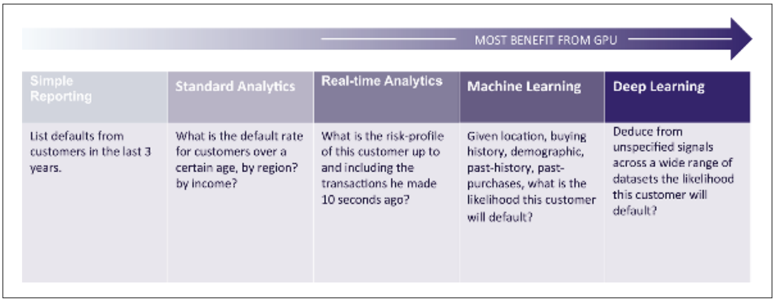
The second practice is the ability to scale GPU-accelerated configurations both up and out. Multiple GPU cards can be placed in a single server, and multiple servers can be configured in a cluster. Such scaling results in more cores and more memory all working simultaneously and massively in parallel 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 because CPUs are and will remain far more expensive than GPUs. So, whether in a single server or a cluster, the GPU database delivers a clear and potentially substantial price/performance advantage.

CHAPTER 3

New Possibilities

The benefit from the performance boost afforded by GPU acceleration is different for different applications. In general, the more processing-intensive the application, the greater the benefit, as shown in Figure 3-1



MOST BENEFIT FROM GPU

Simple Reporting	Standard Analytics	Real-time Analytics	Machine Learning	Deep Learning
List defaults from customers in the last 3 years.	What is the default rate for customers over a certain age, by region? by income?	What is the risk-profile of this customer up to and including the transactions he made 10 seconds ago?	Given location, buying history, demographic, past-history, past-purchases, what is the likelihood this customer will default?	Deduce from unspecified signals across a wide range of datasets the likelihood this customer will default?

Figure 3-1. Although 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 you can use GPU acceleration to improve both the performance and price of a wide variety of database, data analytics, and business intelligence (BI) applications. The next three chapters focus on the three applications or use cases that stand to benefit the most:

- Machine learning and deep learning (Chapter 4)
- Internet of Things (IoT) and real-time data analytics (Chapter 5)

- Interactive location-based intelligence (Chapter 6)

Fast/Full Text Analytics and Natural-Language Processing

A common requirement in many data analytics applications is text analytics and natural-language processing (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 the following (and other) 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 or integrated with existing applications and platforms. Some of the more common techniques 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 or faster GPUs or VRAM. Performance in servers containing multiple GPU cards can be scaled-up even further using NVLink (described in Chapter 2), which offers five times the bandwidth available in a 16-lane PCIe bus.

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

For flexibility, you can deploy GPU solutions on-premise or in public cloud.

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 (see Figure 3-2). 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—increasingly common with the Internet of Things.

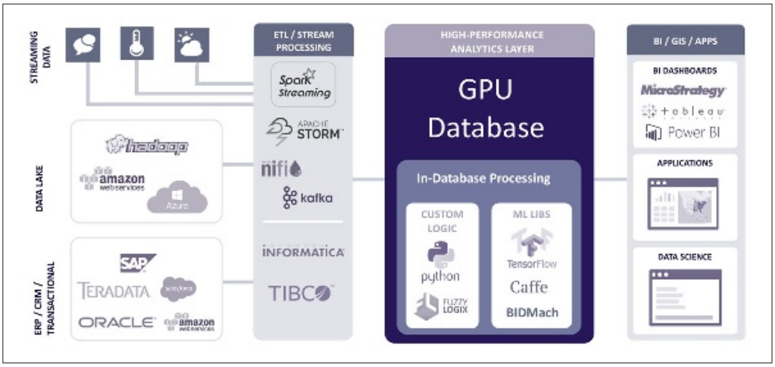


Figure 3-2. GPU databases have open architectures, enabling them to be integrated easily into a wide variety of analytical and BI applications

GPU databases can also serve in a complementary role; for example, as a fast query layer for Hadoop. The ultra-low-latency performance makes GPU-accelerated solutions ideal for those applications that require simultaneous ingestion and analysis of a high volume and velocity of streaming or large, complex data.

Open for Business

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

- Connectors to simplify integration with the most popular open source frameworks, including Accumulo, H2O, HBase, Kibana, Kafka, Hadoop, NiFi, Spark, and Storm
- Drivers for Open Database Connectivity (ODBC) and Java Database Connectivity (JDBC) that enable seamless integration with existing visualization and BI tools such as Tableau, Power BI, and Spotfire
- APIs to enable bindings 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

Recognizing that GPUs are certain to be utilized in mission-critical applications, many solutions are now designed for both high availability and robust security. High-availability capabilities can include data replication with automatic failover in clusters of two or more servers, with data integrity being provided by saving the data to hard disks or solid-state storage on individual servers.

For security, support for user authentication, as well as role- and group-based authorization, help make GPU acceleration suitable for applications that must comply with government 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.

As Figure 3-3 shows, in GPU databases, 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. For smaller datasets 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 saved to hard disks or solid-state drives to ensure no data loss.

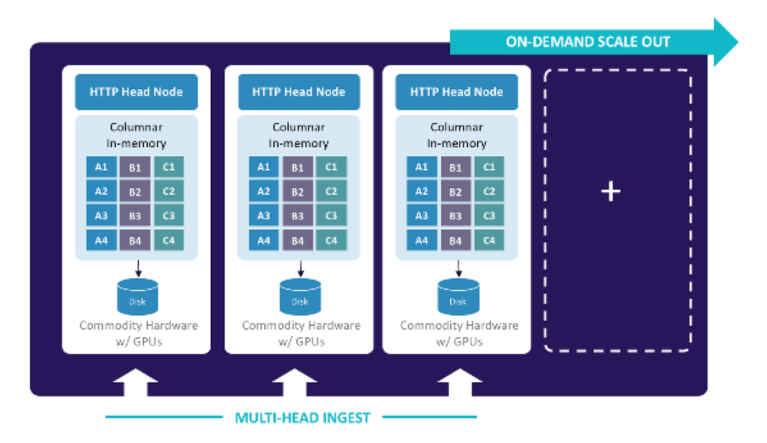


Figure 3-3. The GPU-accelerated in-memory database becomes a "speed layer" capable of providing higher performance for many data analytics and business intelligence applications

CHAPTER 4
Machine Learning and Deep Learning

Machine learning (ML) and deep learning (DL) have emerged as viable technologies for helping organizations progress from deep analytics to predictive analytics by discovering actionable insights in the data. ML/DL models crunch massive datasets and automatically uncover the patterns, anomalies, and relationships needed to make more impactful, data-driven decisions.

But deploying ML within the enterprise presents some challenges. To help overcome these and achieve the full promise of ML, we can use technologies like GPU databases with hardware acceleration, in-memory data management, distributed computing, and integrated open source ML frameworks such as TensorFlow to deliver simpler, converged, and more turnkey solutions.

ML applications have become even easier to implement with the advent of user-defined functions (UDFs). UDFs are able to receive filtered data, perform arbitrary computations, and save the output to a separate table—all in real time on a single database platform. This simplifies and accelerates the entire ML pipeline by unifying the three key processes—data generation, model training, and model serving (as shown in Figure 4-1)—in a single solution that takes advantage of the GPU's massively parallel processing power to deliver the performance needed.

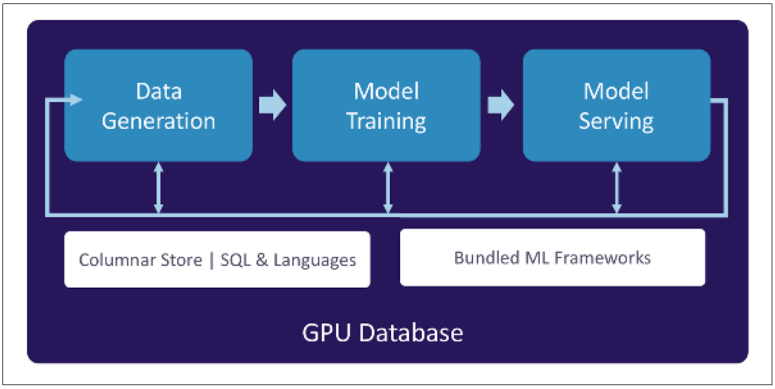


Figure 4-1. GPU databases accelerate the ML pipeline for faster model development and deployment

Data Generation involves acquiring, saving, and preparing datasets to train machine learning models. GPU databases offer advantages in all three data generation tasks:

- For data acquisition, connectors for data-in-motion and at-rest with high-speed ingest make it easier to acquire millions of rows of data across disparate systems in seconds
- For data persistence, the ability to store and manage multistructured data types in a single GPU database makes all text, images, spatial and time–series data easily accessible to ML/DL applications
- For data preparation, the ability to achieve millisecond response times using popular languages like SQL, C++, Java, and Python makes it easier to explore even the most massive datasets

Model training is the most resource-intensive step in the ML pipeline, making it the biggest potential bottleneck. GPU databases with UDFs offer the performance needed to support plug-in custom code and open source ML libraries, such as TensorFlow, Caffe, Torch, and MXNet, for in-line model training. GPU databases maximize performance in three ways:

- Acceleration—Massively parallel processing makes GPUs well-suited for compute-intensive model training workloads on large datasets; this eliminates the need for data sampling and expensive, resource-intensive tuning, and makes it possible to achieve a performance improvement of 100 times on commodity hardware

- Distributed, scale-out architecture—Clustered GPU databases distribute data across multiple database shards enabling parallelized model training for better performance; a scale-out architecture makes it easy to add nodes on demand to improve performance and capacity
- Vector and matrix operations—GPU databases use purpose-built, in-memory data structures and processing optimization to take full advantage of the parallelization available in modern GPUs to deliver an order of magnitude performance improvement on the vector and matrix operations that are common in ML workloads

Model Serving benefits from the ability to operationalize ML by bundling the ML framework(s) and deploying the models in the same GPU database used for data generation and model training, as depicted in Figure 4-2. With such unification, models can be assessed in-line for faster scoring and more accurate predictions.

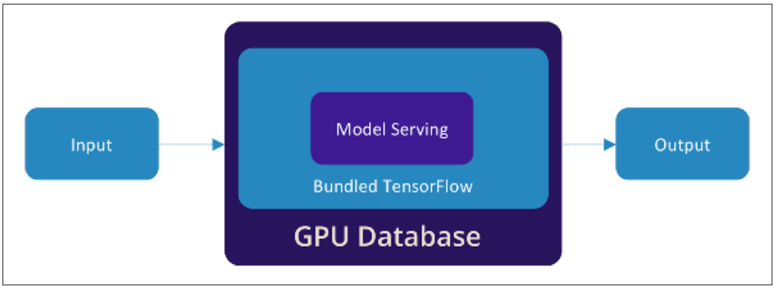


Figure 4-2. Bundling the ML framework and deploying the models in the GPU database makes model serving an in-line process to help operationalize ML

GPU databases unify data with compute and model management to enable data exploration and preparation at any scale. GPU databases also accelerate model training, facilitate model deployment in production and make it easier to manage the model lifecycle to simplify the core workflows of any ML endeavor.

CHAPTER 5
The Internet of Things and Real-Time Data Analytics

Live data can have enormous value, but only if it can be processed as it streams in. Without the processing power required to ingest and analyze these streams in real time, however, organizations risk missing out on the opportunities in two ways: the applications will be limited to a relatively low volume and velocity of data, and the results will come too late to have real value.

This need for speed is particularly true for the Internet of Things (IoT). The IoT offers tremendous opportunities to derive actionable insights from connected devices, both stationary and mobile, and to make these devices operate more intelligently and, therefore, more effectively.

Even before the advent of the IoT, the need to analyze live data in real time, often coupled with data at rest, had become almost universal. Although some organizations have industry-specific sources of streaming data, nearly 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 continuously generates data that holds potential—and perishable—value.

Today, with the IoT, or as some pundits call it, the Internet of Everything, the number of devices streaming data is destined to proliferate to 30 *billion* or more by 2020, according to various estimates.

Only the GPU database has the processing power and other capabilities needed to take full advantage of the IoT. In particular, the ability to perform repeated, similar instructions in parallel across a massive number of small, efficient cores makes the GPU ideal for IoT applications. Because many "Things" generate both time- and location-dependent data, the GPU's geospatial functionality enables support for even the most demanding IoT applications.

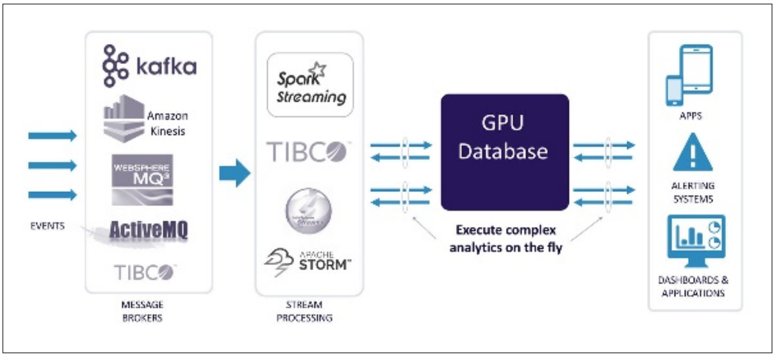


Figure 5-1. A GPU database is able to ingest, analyze, and act on streaming data in real time, making it ideal for IoT applications

For these and other reasons, Ovum declared GPU databases a breakout success story in its *2017 Trends to Watch* based on the GPU's ability to "push real-time streaming use cases to the front burner" for IoT use cases.

The ability to ingest, analyze, and act on streaming IoT data in real time makes the GPU database suitable for virtually any IoT use case. Even though these use cases vary substantially across different organizations in different industries, here are three examples that help demonstrate the power and potential of the GPU.

- Customer experience—GPU databases can ingest information about customers from a variety of sources, including their devices and online accounts, to monitor and analyze buying behavior in real time; this is particularly valuable for retailers with "Customer 360" applications that correlate data from point-of-sale systems, social media streams, weather forecasts, and other sources
- Supply-chain optimization—You can use GPU databases to provide real-time, location-based insights across the entire supply

chain, including suppliers, distributors, logistics, transportation, warehouses, and retail locations, enabling businesses to better understand demand and manage supply

- Fleet management—Public sector agencies and businesses that own and operate vehicles can use GPU databases to integrate live data into their fleet management systems; IoT applications that track location in real time can benefit even more with the geospatial processing power of the GPU

The IoT era is here and growing relentlessly, and only a GPU database can enable organizations to take full advantage of the many possibilities. For those online analytical processing and other business intelligence (BI) applications that stand to benefit from IoT insights, some GPU-accelerated databases now support standards like SQL-92 and BI tools, as well as the high availability and robust security often required in such applications.

CHAPTER 6

Interactive Location-Based Intelligence

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

As more data becomes available from mobile sources like vehicles and smartphones, there are more opportunities to benefit from analyzing and visualizing the geospatial aspects of this data. But traditional geospatial mapping tools, which were designed primarily for creating static maps, are hardly up to the task.

Analyzing large datasets with any sort of interactivity requires overcoming two fundamental challenges: the lack of sufficient computational power in even today's most powerful CPUs to handle large-scale geospatial analytics in anything near real time; and the inability of browsers to render the resulting points, lines and polygons in all but the simplest visualizations.

Given its roots in graphics processing, it should come as no surprise that the GPU is especially well-suited to processing geospatial algorithms on large datasets in real time, and rendering the results in map-based graphics that display almost instantly on ordinary browsers (see Figure 6-1). The GPU-accelerated database also makes it possible to ingest, analyze, and render results on a single platform, thereby eliminating the need to move data among different layers or technologies to get the desired results.

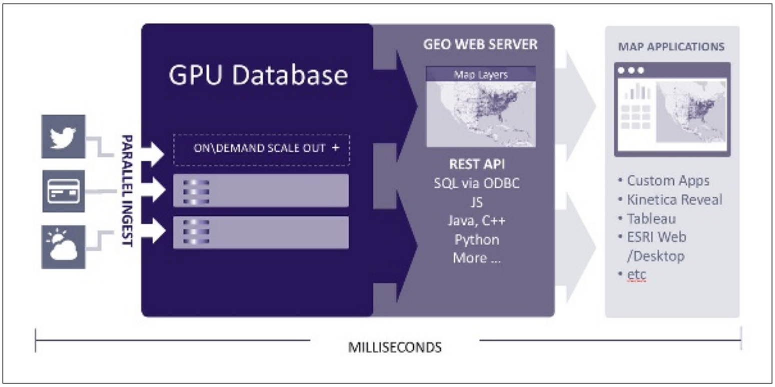


Figure 6-1. The GPU-accelerated database is ideally suited for the interactive location-based analytics that are becoming increasingly desirable

The massively parallel processing power of GPUs makes it possible to support both geospatial objects and operations in their native formats. The ability to perform geospatial operations, such as filtering by area, track, custom shapes, geometry, or other variables, directly on the database assures achieving the best possible performance. Support for geospatial objects, such as points, lines, polygons, tracks, vectors, and labels, in their standard formats also makes it easier to ingest raw data from and export results to other systems.

Standards are critical, as well, to ensuring a quality user experience when the results are rendered on browsers in various visualizations, including heatmaps, histograms, and scatter plots. Most graphical information system (GIS) databases support standards being advanced by the Open Geospatial Consortium, and a growing number of GPU databases now support these standards. OGC standards specify how GIS images are converted to common graphics formats, and also how the graphics are transported via standard web services software that can be incorporated directly into the GPU database.

This approach makes it easy to integrate data from major mapping providers, including Google, Bing, ESRI and MapBox, and facilitates the means for users to interact with the visualizations and change the way the results are displayed. With some solutions, users can now simply drag and drop analytical applets, data tables, and other “widgets” to create completely customized dashboards.

You can further extend geospatial analyses through user-defined functions (UDFs) that enable custom code to be executed directly on the GPU database. By bringing the analysis to the data, this approach eliminates the need to ever extract any data to a separate system.

These forms of customization open a world of possibilities, including using machine learning libraries such as TensorFlow for advanced geospatial predictions. Machine learning makes it possible, for example, to flag deliveries that are unlikely to arrive on time based on traffic, predict which drivers are most likely to be involved in an accident based on driving behavior, or calculate insurance risk for assets based on weather models.

The ability to interact with geospatial data in real time gives business analysts the power to make better decisions faster. With the breakthrough price and performance afforded by GPU databases, that ability is now within reach of almost every organization.

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

Real-World Use Cases

Here are just a few examples of how organizations in different industries are benefiting from GPU-accelerated solutions.

A large pharmaceutical company finds that during the drug development process, the GPU database accelerates simulations of chemical reactions. By distributing the chemical reaction data over multiple nodes, the company can perform simulations much faster and significantly reduce the time to develop new drugs. Researchers can use a traditional language, such as SQL, to run an analysis in a traditional RDBMS environment first, and then as needed, run the same analysis in the GPU-accelerated database.

A major healthcare provider is using a real-time GPU-accelerated data warehouse to reduce pharmacy fraud as well as to enhance its Patient 360 application with dynamic geospatial analysis and healthcare Internet of Things (IoT) data.

A big utility is using a GPU-accelerated database for predictive infrastructure management (PIM). The GPU database operates as an agile layer to monitor, manage, and predict infrastructure health. GPU acceleration enables the utility to simultaneously ingest, analyze, and model multiple data feeds, including location data for field-deployed assets, into a single centralized datastore.

A large global bank is using a GPU-accelerated database to make counterparty customer risk analytics—previously an overnight process—a real-time application for use by traders, auditors, and management. The change was motivated by new regulations requiring the bank to determine the fair value of its trading book as certain trades were being processed. With valuation adjustments needing to be projected years into the future, the risk algorithms had become too complicated and computationally intensive for CPU-only configurations.

One of the world's largest retailers is using a GPU-accelerated database to optimize its supply chain and inventory. The GPU database consolidates information about customers, including sentiment analysis from social media, buying behavior, and online and brick-and-mortar purchases, enabling the retailer's analysts to achieve subsecond results on queries that used to take hours. The application was later enhanced to add data about weather and wearable devices to build an even more accurate view of customer behavior.

The United States Postal Service (USPS) is the single largest logistic entity in the country, moving more individual items in four hours than UPS, FedEx, and DHL combined move all year, and 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 uses a device for scanning packages that also emits precise geographic location every minute to improve various aspects of its massive operation, including maximizing the efficiency of all carrier routes. In total, the GPU database supports 15,000 concurrent sessions analyzing the data streaming in from more than 200,000 scanning devices.

CHAPTER 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 business intelligence (BI), and IBM's Watson supercomputer has demonstrated that this goal can indeed be achieved with existing technology.

The real question is this: 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 streaming analytics with artificial intelligence (AI) and other analytical processes in various ways holds the potential to make real-time, human-like cognition a reality. Such "speed of thought" analyses would not be practical—or even possible—were it not for the unprecedented price and performance afforded by massively parallel processing of the GPU.

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, AI, machine learning, deep learning, 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 or out as needed to deliver whatever real-time performance might be required —from subsecond to a few minutes. We can further improve performance 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 will indeed usher in the Cognitive Era of computing.

CHAPTER 8
Getting Started

GPU acceleration delivers both performance and price 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 or out enables performance to be increased incrementally and predictably—and affordably—as needed.

From a purely financial perspective, GPU acceleration is equally impressive. The GPU's massively parallel processing can deliver performance equivalent to a CPU-only configuration at one-tenth the hardware cost, and one-twentieth the power and cooling costs. The US Army's Intelligence & Security Command (INSCOM) unit, for example, was able to replace a cluster of 42 servers with a single GPU-accelerated server in an application with more than 200 sources of streaming data that produce more than 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 GPU-based solutions even more affordable and easier than ever to access. All of the major cloud service providers, including Amazon Web Services, Microsoft Azure, and Google, now offer GPU instances. Such pervasive availability of GPU acceleration in the public cloud is particularly welcome news for those organizations who want to get started without having to invest in hardware.

With purpose-built GPU solutions, the potential gain can quite literally be 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 predetermine queries in order to be able to ingest and analyze data in real time, regardless of how the organization's data analytics requirements might change over time.

As with anything new, of course, it is best to research your options and choose a solution that can meet all of your analytical needs, scale as you require, and, most important, be purpose-built to take full advantage of the GPU. So start with a pilot project to gain familiarity with the technology, because you will not be able to fully appreciate the raw power and potential of a GPU-accelerated database until you experience it for yourself.

About the Authors

Eric Mizell is the Vice President 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 began 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 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 that company's total revenue and had systems-level strategic planning responsibility for the entire Net/One family of products. Roger began his career as a computer systems sales representative for Hewlett-Packard after graduating Magna Cum Laude from the University of Cincinnati with a B.S. in Electrical Engineering.