WHITE PAPER

kinetica

Maximizing Data Analytics Price/Performance

WITH GPU ACCELERATION

kinetica

Maximizing Data Analytics Price/Performance
WITH GPU ACCELERATION

After 50 years of achieving steady gains in price/performance, Moore's Law has finally run its course for CPUs, where the number of x86 cores that can be placed on a single IC has reached a practical limit. This limit has given rise to the use of server farms or clusters to scale both private and public cloud infrastructures.

But such brute force scaling is expensive, and threatens to exhaust the finite space, power and cooling resources available in many data centers.

Fortunately, for database and big data analytics applications there is now a more capable and cost-effective alternative for scaling 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 and velocity of data confronting organizations today.

This white paper, intended for both technical and business decision-makers, is organized into three sections followed by a brief conclusion. The first section, GPUs 101, is a primer on Graphics Processing Units.

Readers familiar with GPUs can skim or even skip this section. The second section offers an introductory overview of Kinetica's GPU-accelerated architecture for in-memory databases, with emphasis on its suitability for various applications. The third section characterizes Kinetica's performance by providing highlights from both benchmark testing and real-world experience.

GPUs 101

The foundation for affordable and scalable real-time data analytics exists based on steady advances in CPU, memory, storage and networking technologies. Major changes in database price/performance occurred with the advent of solid state storage and more affordable random access memory (RAM). For very large data sets, performance can be accelerated using RAM or flash cache, and/or solid state drives (SSDs). And the ability to configure servers with terabytes of RAM now makes in-memory databases increasingly common.

These changes have shifted the performance bottleneck from input/output to processing. 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.

The most cost-effective way to address this 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—over 100 times more than the 16-32 cores found in today's most powerful CPUs. The GPU's small, efficient cores are

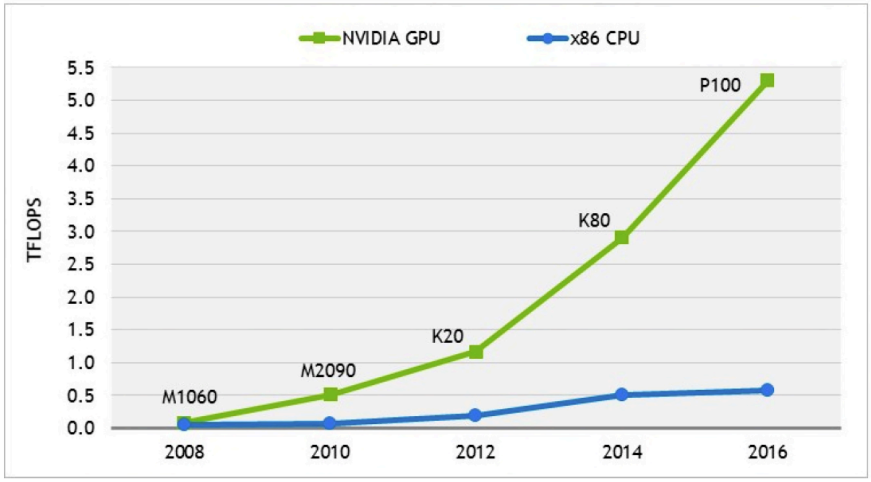
also better suited to performing similar, repeated instructions in parallel, making it ideal for accelerating the processing-intensive workloads common in data analysis.

As the name implies, the GPU was initially designed to process graphics. The first-generation GPU was installed on a separate card with its own memory (video RAM) as the interface to the PC's monitor. The configuration was especially popular with gamers who wanted superior 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. 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. For example, NVIDIA®'s K80 GPU contains 4,992 cores. The typical GPU accelerator card today utilizes 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 that provides 5 times the bandwidth (160 GB/s) between the CPU and GPU, and among GPUs.

Within the latest generation of GPU cards, the memory bandwidth is significantly higher at rates up to 732 GB/s. Compare this bandwidth to the 68 GB/s in a Xeon E5 CPU—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).



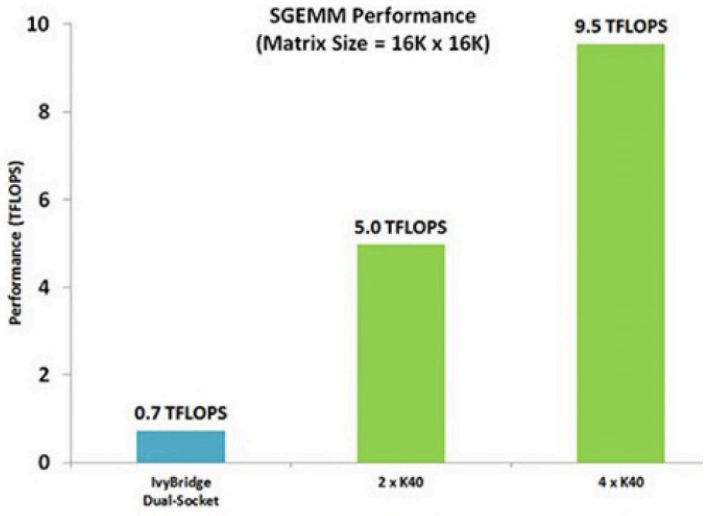
The latest generation of GPUs from NVIDIA contain upwards of 5,000 cores and deliver doubleprecision processing performance of 5 TeraFLOPS. Note also the relatively minor performance improvement over time for multi-core x86 CPUs, and how it is now flattening. (Source: NVIDIA)

The relatively small amount of memory on a GPU card compared to the hundreds of GB or few TB 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 data set at once to achieve the desired results. 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, 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 clusters. 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 singleprecision floating general matrix multiply (SGEMM) benchmark tests. (Source: NVIDIA)

Kinetica's GPU-Accelerated Architecture for In-Memory Databases

Kinetica is an in-memory, GPU-accelerated distributed database designed for both high-performance and ease of integration into a wide range of data analytics applications. The massively parallel processing at the core of Kinetica's architecture also makes the solution more scalable and affordable than other solutions.

The Kinetica database operates on commodity servers equipped with both x86 CPUs and GPUs that can be configured to scale both up and out—predictably and linearly as needed—to achieve the desired performance.

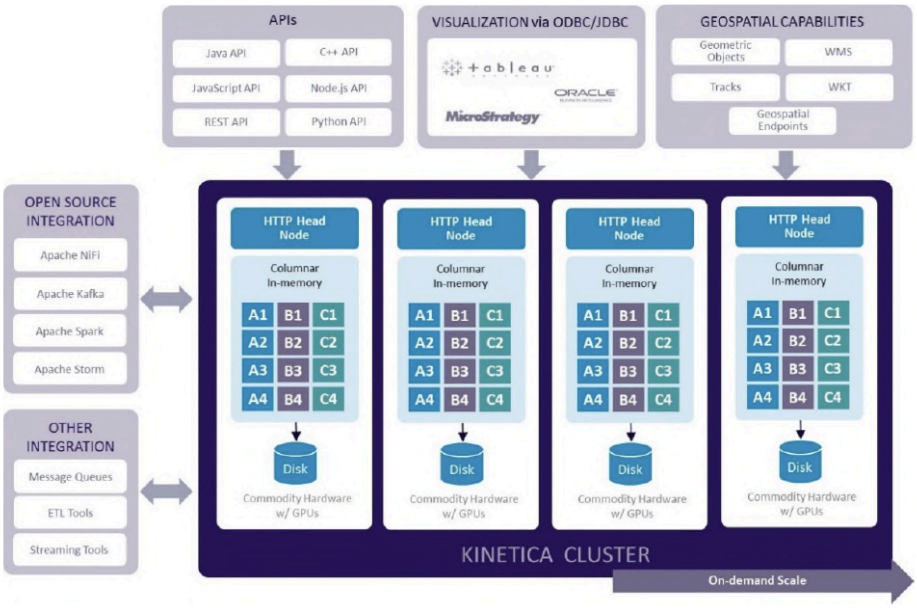
Scaling up involves adding more/faster GPUs and/or system RAM. Performance in servers containing multiple GPU cards can be scaled up even further using NVLink, which offers 5x the bandwidth available in a 16 lane PCIe bus. Scaling out involves simply adding more servers in a cluster, which can also be done in a distributed configuration to enhance reliability.

The open Kinetica architecture makes it suitable for virtually any data analytics application that might benefit from higher and/or more cost-effective performance. Potential applications range from traditional relational databases to those requiring real-time analysis of streaming data or complex event processing, with the latter two becoming increasing common with the Internet of Things. Its ultra-low-latency performance makes Kinetica suitable even for those applications that require simultaneous ingest and analysis of a high volume and velocity of streaming data. More information about how Kinetica's architecture facilitates integration with a wide variety of applications can be found in the Open for Business sidebar.

Open for Business

To enable support for a broad range of data analytics environments and needs, Kinetica has an "application-agnostic" architecture that includes:

- Built-in connectors to simplify integration with the most popular open-source frameworks, including (in alphabetical order) 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 and Tableau
- Application Programming Interfaces (APIs) to enable binding with commonly-used programming languages, including REST, 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

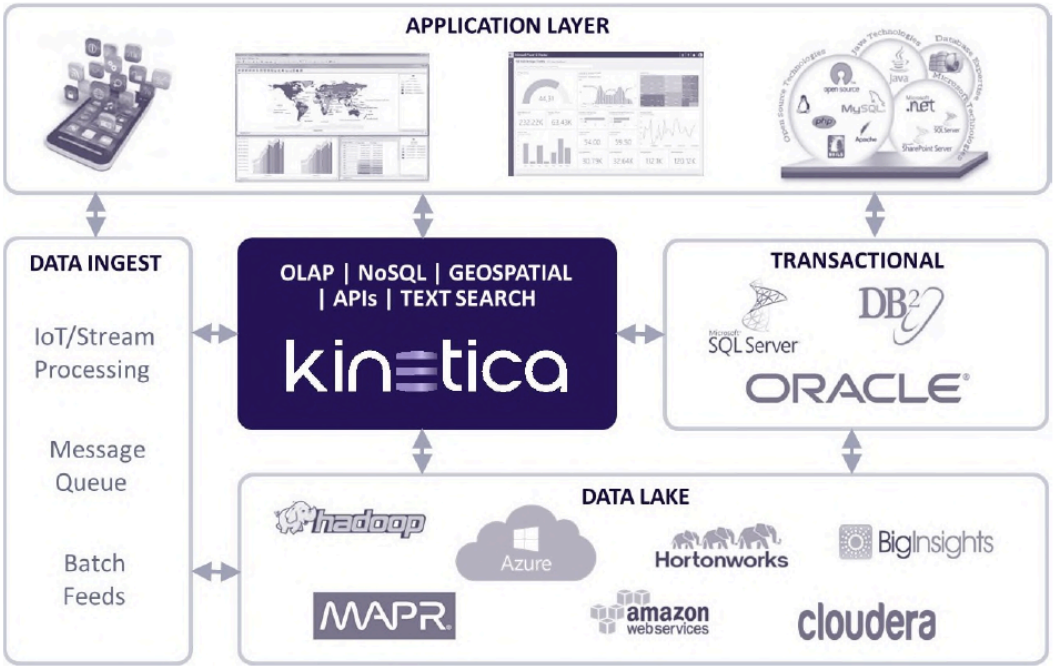


The Kinetica architecture is designed to be open, enabling it to be integrated easily into a wide variety of analytical applications.

Recognizing that the Kinetica database is certain to be utilized in many missioncritical applications, the architecture has been designed for both high availability and robust security. High availability while data integrity is assured with disk-based persistence on individual servers. Security is provided by rigorous user authentication and authorization.

As a database, Kinetica is similar in its functionality to other databases, including those that operate in memory. What makes Kinetica different is how it manages the storage and processing of data for peak performance in massively parallel configurations.

Data is stored in system memory in vectorized columns to optimize processing across all available GPUs. Data is moved 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.



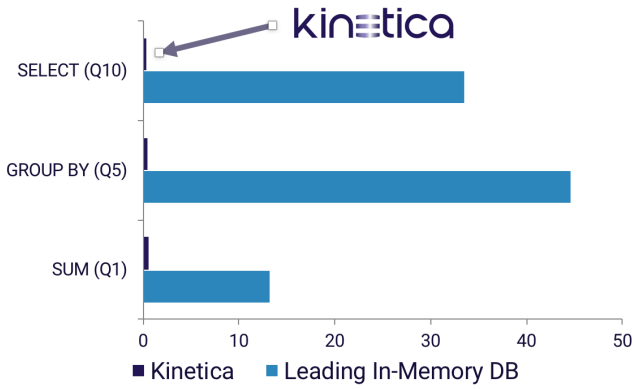
Kinetica is a "speed layer" capable of providing higher and/or more cost-effective performance for virtually any data analytics application, especially those requiring real-time response.

Putting GPU Acceleration to the Test

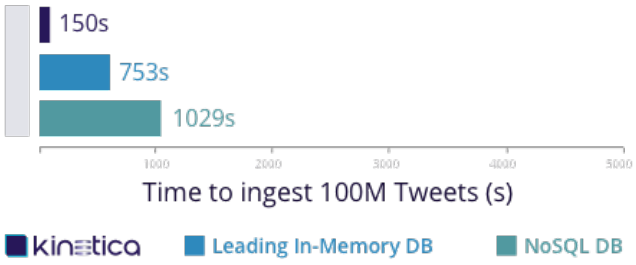
Virtually all applications, algorithms, libraries and processes achieve better performance when executed in a GPU without any modification. One example of impressive price/performance can be found in a two-node/four-GPU cluster that was able to query a database of 15 billion Tweets and render a visualization in less than a second.

Even better performance can be achieved when the application software is optimized to take advantage of the GPU's massive parallel processing. And that is precisely what Kinetica has done: fully optimize an in-memory application for processing in both CPUs and GPUs to achieve industry-leading performance.

The graphs below provide summary results from two different sets of benchmark tests. In each graph, the purple bar is Kinetica; the blue bar is an in-memory databased with no GPU acceleration; and the green bar is a NoSQL database.



Although less than one-sixth the size of the in-memory cluster without GPU acceleration, the Kinetica configuration was able to process all three of these advanced analytical queries on 150 billion rows of data in less than a second.



With a single GPU on a single server, Kinetica is able to out-perform both the un-accelerated in-memory and traditional NoSQL databases by factors of 5 and 7, respectively.

The real-world experience of Kinetica customers confirms the findings of superior performance revealed in these and other benchmark tests. Here are summaries of the results experienced by two different customers.

The ***U.S. Army Intelligence & Security Command*** (INSCOM) has a need to analyze geospatial and temporal data to track assets and identify terrorist threats in real-time as part of its overall threat intelligence responsibilities. With over 200 sources of streaming data producing over 100 billion new records per day, the application is particularly demanding.

INSCOM chose Kinetica based on its ability to ingest and analyze all of the data streams in real-time. The original cluster of 42 servers running Oracle 10gR2, which required 92 minutes to complete one geospatial query, was replaced by a single server running Kinetica that is now able to perform the same query in less than a second. The results are truly impressive: superior performance with 28 times lower cost and 38 times less power consumption.

The ***U.S Postal Service*** (USPS) moves more individual items in four hours than UPS, FedEx and DHL combined move all year, making it the single largest logistics entity in the nation. The USPS tracks over 200,000 devices that are emitting location once per minute, resulting in more than a quarter-billion events that need to be ingested and analyzed every day.

The USPS chose Kinetica based on its industry-leading price/performance. The Kinetica cluster is able to serve up to 15,000 sessions daily, providing USPS managers and analysts with real-time dashboard views of "breadcrumb" and sensor data to track where delivery vehicles and carriers are at any moment, including at collection and delivery points. By analyzing this data, the USPS is able to:

- Reduce costs by streamlining deliveries and minimizing inefficiencies, such as overlapping coverage of assigned areas, uncovered areas and distribution bottlenecks
- Enhance decision-making capabilities through a better understanding of where investments would achieve the best results
- Improve customer service through contingency planning whenever any carrier is unable to complete an assigned route

Conclusion

GPUs deliver a substantial price/performance advantage over CPUs in many applications, and Kinetica has now brought that advantage to database and data analytics applications.

From a performance perspective, Kinetica is able to ingest and analyze large volumes of high-velocity data in real time. In both benchmark tests and real-world applications, Kinetica has proven its 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, such as cognitive computing, a practical reality.

From a cost perspective, the GPU-accelerated Kinetica database is equally impressive. The GPU's massively parallel processing results in significant savings at 1/10TH the hardware costs and 1/20TH the power and cooling costs. Indeed, as the U.S. Army's INSCOM experience shows, Kinetica is able to replace large clusters with a single server. And the ability to scale up and/or out enables performance to be increased incrementally and predictably—and affordably—as needed.

But just as important is that Kinetica's price/performance advantage is easily within reach of any IT organization. Its open architecture makes it easy to plug Kinetica into virtually any existing data architecture, and to integrate with both open source and commercial data analytics frameworks. The gain is quite literally without the pain normally associated with indexing or redefining schemas or tuning/tweaking algorithms, and without the need to ever again pre-determine queries in order to be able to ingest and analyze data in realtime, however your needs might change.

To learn more about or get a demo too see how your organization can benefit from GPU acceleration in your database and data analytics applications, please visit us at **kinetica.com** *or call us at* **(415) 604-3444**.

About Kinetica

Kinetica addresses today's data paradigm by bringing Graphics Processing Units (GPUs) to the datacenter. Built from the ground up to scale linearly, Kinetica's distributed, in-memory database accelerated by GPUs delivers truly real-time actionable intelligence on large, complex and streaming data sets: 100x faster performance at 1/10 of the hardware of traditional databases. Kinetica makes real time a reality. Organizations use Kinetica to simultaneously ingest, explore, analyze and visualize streaming data within milliseconds to make critical decisions and find efficiencies, lower cost, generate new revenue, and improve customer experience. Learn more at kinetica.com.