kinetica

Bringing the Power of GPUs to Analytics

WITH CISCO UCS, NVIDIA, AND KINETICA

Harness the power of graphics processing unit (GPU) acceleration using the Cisco UCS® with NVIDIA GPUs and Kinetica to ingest, analyze, and visualize data in real time.

cisco Solution Partner

NVIDIA

HIGHLIGHTS

Integrated Infrastructure Built on Cisco UCS Advantages

Cisco UCS Integrated Infrastructure for Big Data and Analytics is a proven platform for enterprise analytics applications with capabilities for powering GPU-accelerated applications.

High Performance with Linear Scalability

Cisco UCS Integrated Infrastructure for Big Data is a simplified, intelligent infrastructure with high performance and scalability to meet growing business demands.

Ease of Deployment

Cisco UCS Manager simplifies infrastructure deployment with an automated, policy-based mechanism that helps reduce configuration errors and system downtime. It offers proven, high-performance linear scalability and easy scaling of the architecture with single- and multiple-rack deployments.

Unparalleled Speed

NVIDIA's Tesla Accelerated-Computing Platform accelerates the most demanding high-performance data analytics and scientific computing applications

Applying the Power of GPUs for Real-Time Analytics

Kinetica harnesses the power of GPUs for unprecedented performance to ingest, explore, and visualize streaming data in real time. Kinetica's parallelized processing architecture not only enables predictable and near-linear scalability, but also reduces analytical processing times compared to leading in-memory and analytical databases.

Advanced Analytics with In-Database Processing

User-defined functions (UDFs) enable compute as well as data-processing, within the database. Kinetica uniquely offers such in-database functionality that fully utilizes the parallel compute power of the GPU to bring AI and BI together.

Extreme Performance for Analytics

The initial analytics breakthrough enabled the use of many servers to work on a single problem simultaneously and brought parallel processing into the mainstream. This approach meant large numbers of servers could work together harmoniously but was constrained by disk I/O. The obvious way to deal with this limitation was to move the processing off of disk and into memory. This capability enabled the creation of new solutions with low-latency requirements such as interactive analysis and real-time processing of streaming data.

The current generation of in-memory data processing frameworks has dramatically improved performance and response times over its predecessors but is constrained by the number of processing cores operating on the data. The next logical step is to increase the number of simultaneous processing threads on each server. Hyperthreading, more cores per CPU, and more CPUs per server are ways to achieve this.

There is another kind of processor, a graphics processing unit, or GPU, which has traditionally been used to drive graphical user interfaces. This processor has a massively parallel architecture consisting of thousands of smaller cores. These cores can do only a small subset of what a CPU can do, but what they can do, they do far more efficiently, and they are specifically designed to handle multiple tasks simultaneously.

Recent advances in GPU technology have extended their use beyond graphics into the realm of data analytics and scientific computing. With GPUs installed, the amount of raw processing power per server is orders of magnitude greater than what can be achieved with CPUs alone. GPUs are not general-purpose processors like CPUs. They work along with CPUs to offload specific jobs (video analytics, deep learning, and so on) to GPUs to accelerate the performance of these tasks and free up CPUs for more general-purpose processing.

Kinetica has created a database from the ground up that harnesses the power of GPUs to ingest, explore, analyze, and visualize data, both at rest and in motion. Kinetica is a GPU-accelerated, in-memory, distributed database with SQL-style query capabilities. It is ODBC-compliant, supports ANSI SQL-92-compliant syntax, and presents a familiar traditional RDBMS interface to users and developers.

By bringing together the Cisco UCS Integrated Infrastructure for Big Data, NVIDIA's GPU-accelerated hardware, and Kinetica's in-memory distributed database, you can achieve truly extreme performance for data analytics.

Cisco UCS Integrated Infrastructure for Big Data and Analytics

Organizations today must help ensure that the underlying physical infrastructure can be deployed, scaled, and managed in a way that is agile enough to change as workloads and business requirements change. The Cisco Unified Computing System™ (Cisco UCS) has redefined the potential of the data center with its revolutionary approach to integrated infrastructure to meet the business needs of IT innovation and acceleration. The Cisco UCS Integrated Infrastructure for Big Data and Analytics provides an end-to-end architecture for processing high volumes of real-time or archived data, both structured and unstructured. At the same time, it transparently integrates relevant complex capabilities to deliver an enterprise-class offering with high performance and scalability as applications demand.

Cisco UCS Manager

Cisco UCS Manager enables rapid and consistent server configuration using service profiles and automates ongoing system maintenance activities such as firmware updates across the entire cluster as a single operation. It enables fast and accurate configuration of computing, network, and storage resources. Cisco UCS Manager also offers advanced monitoring with options to raise alarms and send notifications about the health of the entire cluster.

Cisco UCS 6200 and 6300 Series Fabric Interconnects

Cisco UCS 6200 Series Fabric Interconnects provide high-bandwidth, low-latency connectivity for servers, with integrated, unified management provided for all connected devices by Cisco UCS Manager. The Cisco UCS 6300 Series is the latest version of this technology. The Cisco UCS 6300 Series Fabric Interconnects are a core part of Cisco UCS, providing low-latency, lossless, 10 and 40 Gigabit Ethernet, Fibre Channel over Ethernet (FCoE), and Fibre Channel functions with management capabilities for systems deployed in redundant pairs. Cisco® fabric interconnects offer the full active-active redundancy, performance, and exceptional scalability needed to support the large number of nodes that are typical in clusters serving big data applications. Cisco UCS Manager enables rapid and consistent server configuration using service profiles and automates ongoing system maintenance activities such as firmware updates across the entire cluster as a single operation. Cisco UCS Manager also offers advanced monitoring with options to raise alarms and send notifications about the health of the entire cluster.

Cisco UCS C-Series Rack Servers

Cisco UCS C240 M4 Rack Servers support a wide range of computing, I/O, and storage-capacity demands in a high-density, compact design. The server uses dual Intel Xeon processor E5-2600 v4 series CPUs and supports up to 1.5 TB of main memory and a range of hard-disk drive (HDD) and solid-state disk (SSD) drive options. The performance-optimized option supports 24 small-form-factor (SFF) disk drives, and the capacity-optimized option supports 12 large-form-factor (LFF) disk drives. This server can be used with the Cisco UCS Virtual Interface Card (VIC) 1227 or 1387, depending on the fabric interconnect that is being used. The VIC 1227 is designed to optimize high-bandwidth and low-latency cluster connectivity. The VIC 1387 offers dual-port Enhanced Quad Small Form-Factor Pluggable (QSFP+) 40 Gigabit Ethernet and FCoE in a modular-LAN-on-motherboard (mLOM) form factor.

NVIDIA Tesla Accelerated Computing Platform

NVIDIA makes some of the world's fastest accelerators. Part of the NVIDIA Tesla Accelerated Computing Platform, Tesla GPU accelerators are built on the NVIDIA Kepler compute architecture and powered by CUDA, a widely used parallel computing model. This makes them ideal for delivering record acceleration and compute performance efficiency for applications in fields including:

- Machine learning and data analytics
- Seismic processing
- Computational biology and chemistry
- Weather and climate Modeling
- Image, video, and signal processing
- Computational finance/physics
- CAE and CFD

Key features include:

- Hyper-Q allows multiple CPU cores to simultaneously use the CUDA cores on single or multiple Kepler-based GPUs. This dramatically increases GPU utilization, simplifies programming, and slashes CPU idle times.
- Memory error protection meets a critical requirement for computing accuracy and reliability in data centers and supercomputing centers.
- Asynchronous transfer with dual DMA engines dramatically increases system performance by transferring data over the PCIe bus while the computing cores are crunching other data.
- GPU boost enables the end user to convert power headroom to higher clocks and achieve even greater acceleration for various HPC workloads.
- Zero-power idle increases data center energy efficiency by powering down idle GPUs when running legacy nonaccelerated workloads.

Kinetica: Reinventing the Distributed Database

Kinetica is a GPU-accelerated, in-memory, distributed database with SQL-style query capability. It has been designed from the ground up with thousands of processing cores in mind. Kinetica presents a familiar, traditional RDBMS interface to users and developers and therefore does not require that they understand the intricacies of the underlying distributed nature of the database.

Traditional database design requires complex data structures aimed at reducing the computational workload at query time. It also requires very specific application-level insight into how the data will be queried and the creation of indexes needed to achieve reasonable response times. This approach was mandatory when the hardware had a single or very few threads of control.

Kinetica uses GPU-based technology to provide a system that merges query needs of traditional databases with the scalability and performance demands of today's big data systems.

The GPU parallelized processing architecture not only enables predictable and near-linear scalability, but also reduces analytical processing times for multibillion-row datasets by orders of magnitude compared to leading in-memory and analytical databases.

The availability of in-database analytics via user-defined functions (UDFs–an industry-first capability–makes the parallel processing power of the GPU accessible to custom analytics functions deployed within Kinetica. This opens the opportunity for machine learning/artificial intelligence libraries such as TensorFlow, BIDMach, Caffe, and Torch to run in-database alongside, and converged with, BI workloads.

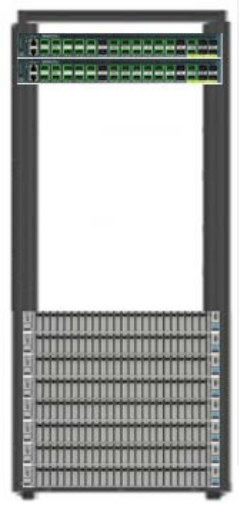
Kinetica's advanced in-database analytics make it possible for organizations to affordably converge Artificial Intelligence, Business Intelligence, Machine Learning, natural language processing, and other data analytics into one powerful platform. It exposes advanced analytics to business users who understand the data resulting in better business value. By democratizing data science workloads, businesses get more efficient and effective business process outcomes, faster time to market, and net new business value.

Kinetica's extensible and flexible visualization framework, Reveal, enables interactive, real-time data exploration with GPU-accelerated rendering of maps and accompanying dashboards. Business analysts can make faster decisions by visualizing and interacting with billions of data elements in real-time. Reveal also connects with Kinetica's geospatial pipeline for advanced mapping and interactive location-based analytics. Kinetica can also be connected to open-source tools such as Kibana and Caravel or to business intelligence reports and dashboards using ODBC/JDBC.

Recommended Configuration

8 Cisco UCS C240 M4 (Dual Socket) servers each having:

CPU:	Intel Xeon E5-2680v4 (14 cores)
Memory:	384 GB RAM
GPU:	2 x K80s each server (1 per socket)
Drives:	4 x 1.6TB SSD (enterprise values) 2x240G SSD (for boot)
Connectivity:	2 x Cisco UCS 6332 UP 32-Port Fabric Interconnect (40G links)



Performance Tests and Results

The testing methodology for this evaluation focused on performance across the following three major dimensions: cluster size, data ingestion speed, and bounding box calculations. Bounding box helps identify how many objects within the given table lie in a rectangular box. These calculations showcase the advantage over traditional database solutions by running GPU-intensive calculations massively in parallel. The tests included the below steps:

Data Ingestion:

Ingest ~4Billion tweets (static data) to bring the data set on the cluster to production size. These tweets were collected in advance prior to the test.

Data query:

Run the small to large bounding box queries to test the GPU intensive calculations on this static data set (on x,y co-ordinates simultaneously)

The real benefit of the GPU-enabled cluster manifests in the performance of queries that require brute-force scanning of huge volumes of unindexed data. The data is split up and sent to the thousands of GPU cores for parallel scanning, and the aggregated result is returned to the user.

Figure 2 highlights the near-linear performance of the cluster with increasing numbers of nodes.

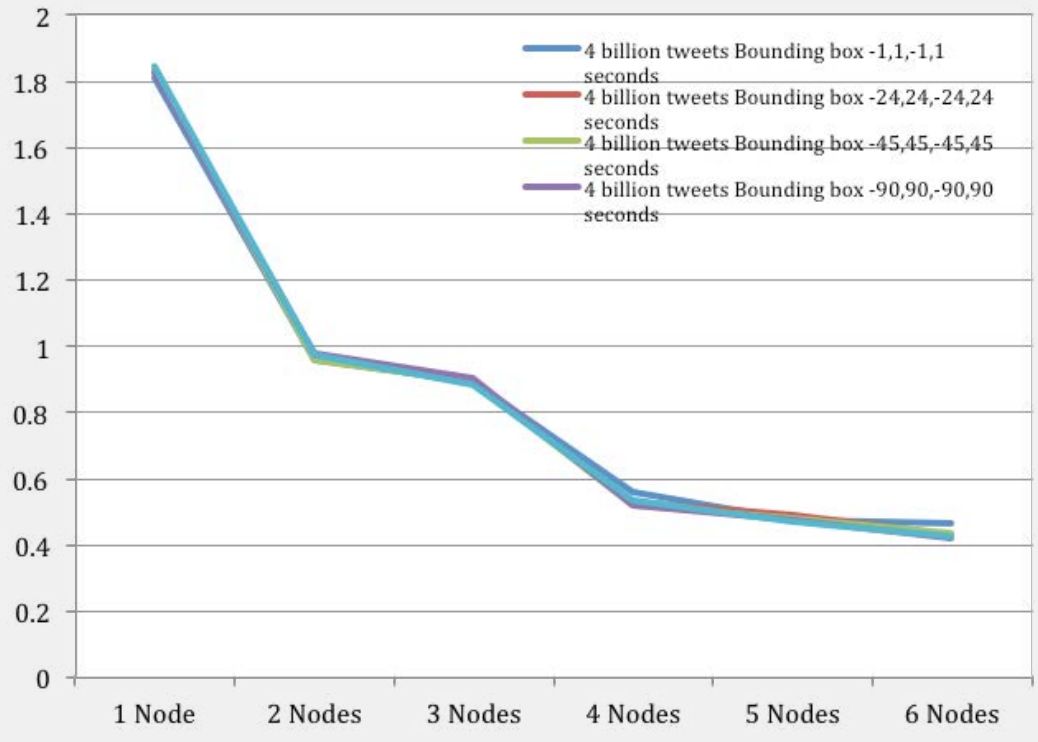


Figure 2. Results of Running CPU-Intensive Bounding Box Queries (explained above) with Increasing Number of Nodes

Conclusion

The current generation of in-memory processing frameworks achieves orders of magnitude greater performance than its I/O-constrained predecessors. Kinetica, by using the power of GPU acceleration and in-memory capabilities, achieves further orders of magnitude acceleration compared to in-memory analytical engines. The availability of in-database analytics via user-defined functions (UDFs) enables compute as well as data-processing, within the database, and opens the way for converged AI and BI workloads accelerated by GPUs.

The Cisco UCS Integrated Infrastructure for Big Data and Analytics provides a simplified intelligent infrastructure with the scalability to meet growing business demands. By bringing together the Cisco UCS Integrated Infrastructure for Big Data and Analytics, NVIDIA's GPU-accelerated hardware, and Kinetica's in-memory distributed database, you can achieve truly extreme performance for real-time data analytics.

This solution provides a GPU parallelized processing architecture accomplishing predictable and near-linear scalability and reduces analytical processing times for multibillion-row datasets.

For More Information

- **For more information about Cisco UCS big data solutions, see www.cisco.com/go/bigdata_design**
- **For more information about the Cisco UCS Integrated Infrastructure for Big Data, see http://blogs.cisco.com/datacenter/cpav4/**
- **For more information about NVIDIA's Tesla K80 accelerator, see http://www.nvidia.com/object/tesla-k80.html**
- **For more information about Kinetica, see http://www.kinetica.com/**

kinetica

NVIDIA

cisco Solution Partner

Kinetica and the Kinetica logo are trademarks of Kinetica and its subsidiaries in the United States and other countries. Other marks and brands may be claimed as the property of others. The product plans, specifications, and descriptions herein are provided for information only and subject to change without notice, and are provided without warranty of any kind, express or implied. Copyright © 2017 Kinetica