Bring the Power of the GPU to Analytics with Cisco UCS, NVIDIA, and Kinetica

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

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 graphics processing unit (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.

Exceptional speed

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

The power of GPUs for real-time analytics

Kinetica harnesses the power of GPUs for outstanding performance in ingesting, exploring, and visualizing streaming data in real time. Kinetica's parallelized processing architecture enables predictable and near-linear scalability and reduces analytical processing times compared to leading in-memory and analytical databases.

Advanced analytics with in-database processing

User-defined functions (UDFs) enable both computing and data processing within the database. Kinetica uniquely offers such in-database capabilities, making full use of the parallel computing power of the GPU to bring artificial intelligence and business intelligence 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 that large numbers of servers could work together harmoniously, but performance was constrained by disk I/O. The obvious way to deal with this limitation was to move the processing off the disk and into memory. This capability enabled the creation of new solutions for applications 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 (GUIs). 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 the use of GPUs beyond graphics into the realms of data analytics and scientific computing. With GPUs installed, the amount of raw processing power per server is orders of magnitude greater than that offered 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 CPUs for more general-purpose processing.

Kinetica has created a database from the foundation 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 Open Database Connectivity (ODBC) compliant, supports syntax compliant with ANSI SQL-92, and presents a familiar traditional relational database management system (RDBMS) interface to users and developers.

By bringing together 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 data analytics.

Cisco UCS Integrated Infrastructure for Big Data and Analytics

Organizations today must be sure 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. Cisco UCS Integrated Infrastructure for Big Data and Analytics has redefined the potential of the data center with its revolutionary approach to managing computing, network, and storage resources to successfully address the business needs of IT innovation and acceleration. Cisco UCS Integrated Infrastructure for Big Data and Analytics provides an end-to-end architecture for processing high volumes of structured and unstructured data for both real-time processing and archival purposes.

Cisco UCS 6300 Series Fabric Interconnects

Cisco UCS 6300 Series Fabric Interconnects provide high-bandwidth, low-latency connectivity for servers, with Cisco UCS Manager providing integrated, unified management for all connected devices. The Cisco UCS 6300 Series Fabric Interconnects are a core part of Cisco UCS, providing low-latency, lossless 40 Gigabit Ethernet, Fibre Channel over Ethernet (FCoE), and Fibre Channel functions.

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 C240 M5 Rack Server

The Cisco UCS M5 Rack Server is a dual-socket, 2-rack-unit (2RU) server offering industry-leading performance and expandability for a wide range of storage and I/O-intensive infrastructure workloads, for big data and analytics. This server uses the latest Intel® Xeon® Processor Scalable Family with up to 28 cores per socket. It supports up to 24 double-data-rate 4 (DDR4) dual in-line memory modules (DIMMs) for improved performance and lower power consumption. The DIMM slots are 3D XPoint ready, supporting next-generation nonvolatile memory technology. Depending on the server type, Cisco UCS rack servers offer a range of storage options. The Cisco UCS C240 M5 supports up to 24 small form-factor (SFF) 2.5-inch drives (with support for up to 10 Non-Volatile Memory Express [NVMe] Peripheral Component Interconnect Express [PCIe] solid-state drives [SSDs] on the NVMe-optimized chassis version) or 12 large-form-factor (LFF) 3.5-inch drives plus 2 rear hot-swappable SFF drives with a Cisco 12-Gbps SAS Module RAID Controller. A modular LAN-on-motherboard (mLOM) slot supports dual 40 Gigabit Ethernet network connectivity with the Cisco UCS Virtual Interface Card (VIC) 1387.

NVIDIA Tesla Accelerated Computing Platform

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

- Machine learning and data analytics
- Seismic processing
- Computational biology and chemistry
- Weather and climate modeling
- Image, video, and signal processing
- Computational finance and physics
- Computer-aided design (CAE) and computational fluid dynamics (CFD)

The main features of the Tesla Accelerated Computing Platform include the following:

- Hyper-Q allows multiple CPU cores to simultaneously use the CUDA cores on single or multiple Kepler-based GPUs. This feature dramatically increases GPU utilization, simplifies programming, and decreases the amount of CPU idle time.
- Memory error protection meets a critical requirement for computing accuracy and reliability in data centers and supercomputing centers.
- Asynchronous transfer with dual direct-memory-access (DMA) engines dramatically increases system performance by transferring data over the PCIe bus while the computing cores are processing other data.
- GPU Boost technology enables the end user to convert power headroom to higher clock speeds and achieve even greater acceleration for various high-performance computing (HPC) workloads.
- The zero-power-idle feature increases data center energy efficiency by powering down idle GPUs when the system is running traditional nonaccelerated workloads.

Kinetica: Reinventing the Distributed Database

Kinetica is a GPU-accelerated, in-memory, distributed database with SQL-style query capabilities. It is designed from the foundation to handle thousands of processing cores. 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 the way that the data will be queried and the creation of indexes needed to achieve reasonable response times. This approach was mandatory when the hardware had just one or very few threads of control.

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

The GPU parallelized processing architecture both enables predictable and near-linear scalability and reduces analytical processing times for multibillion-row data sets by orders of magnitude compared to leading in-memory and analytical databases.

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

Kinetica's advanced in-database analytics enable organizations to affordably converge artificial intelligence, business intelligence, machine learning, natural language processing, and other data analytics into one powerful platform. Kinetica exposes advanced analytics to business users who understand the data, resulting in better business value. By democratizing data science workloads, Kinetica helps businesses achieve more efficient and effective business process outcomes, faster time to market, and 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 and to business intelligence reports and dashboards using ODBC and Java Database Connectivity (JDBC).

Reference Architecture

The Cisco UCS Integrated Infrastructure for Big Data and Analytics for Kinetica includes eight or more C240 M5 servers, each with dual Intel Xeon Processor Scalable Family 6132 CPUs (2 x 14 cores , 2.6 GHz) 384 GB of RAM, dual 40-Gbps network connectivity, and 8 (or 16) SSDs. These servers are connected to Cisco UCS 6332 Fabric Interconnects.

Note: 2 x P100 GPUs per server.

Figure 1 shows the reference architecture for Kinetica.

Performance Tests and Results

Note: The tests described here were performed using Cisco UCS C240 M4 Rack Servers. These servers are used here to highlight the linear scalability of the solution. Cisco UCS C240 M5 Rack Servers are expected to provide even higher performance with the latest GPUs.

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 calculations help identify the number of objects within the given table that reside in a rectangular box. These calculations show the advantage of the tested solution compared to traditional database solutions by running GPU-intensive calculations massively in parallel.

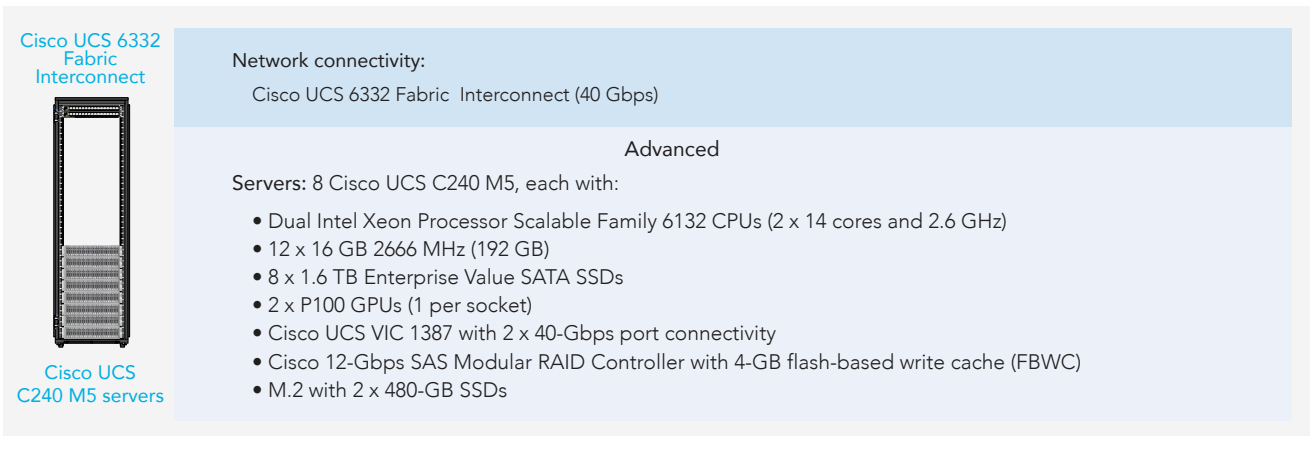

Figure 1. Cisco UCS Reference Architecture for Kinetica

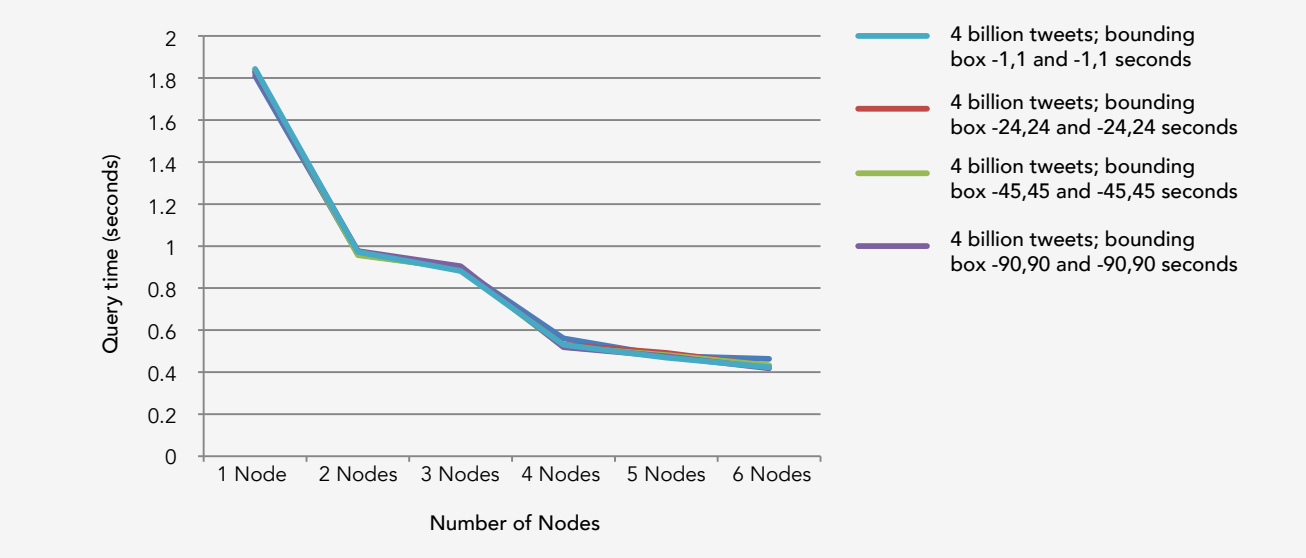

Figure 2. Results of Running CPU-Intensive Bounding-Box Queries with Increasing Numbers of Nodes

The tests included the following processes:

- Data ingestion: Ingest about 4 billion 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 small to large bounding-box queries to test the GPU-intensive calculations on this static data set (on x,y coordinates simultaneously).

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

Figure 2 shows the near-linear performance of the cluster with increasing numbers 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 through UDFs enables both computing and data processing within the database and opens the way for converged artificial intelligence and business intelligence workloads accelerated by GPUs.

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 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 that offers predictable and near-linear scalability and reduces analytical processing times for multibillion-row data sets.

Reference

- For more information about Cisco UCS big data solutions, see https://www.cisco.com/go/bigdata_design.
- For more information about Cisco UCS Integrated Infrastructure for Big Data, see https://blogs.cisco.com/datacenter/cpav5/.
- For more information about the NVIDIA Tesla K80 accelerator, see http://www.nvidia.com.
- For more information about Kinetica, see http://www.kinetica.com/.