Data Services on Ampere® Altra® Processors

March 2023

Big Data – Spark on Ampere® Processors

Ampere® Empowering the Future

Ampere® Altra® supports up to 128 cores based on the AArch64 (Arm) architecture. In addition to delivering a large number of high-performance cores, its innovative architecture delivers predictability, linear scalability, and power efficiency.

Apache Spark is an open source, distributed processing system used for big data workloads. It utilizes in-memory caching and optimized query execution for fast analytic queries against data of any size. It provides APIs in Java, Scala, and Python and supports multiple real-time analytic workloads, batch processing, interactive queries, and machine learning. Spark addresses the limitations of Hadoop by performing in-memory processing using Resilient Distributed Dataset (RDD) and reusing data across multiple parallel operations. It relies on other storage systems like HDFS, Couchbase, Cassandra and others.

Spark can run in standalone cluster mode or can run on a Cluster Management system like Yarn, Kubernetes, or Docker.

The Spark architecture comprises a Driver, Cluster Manager, and Executor. The driver is the controller of the Spark execution engine and maintains the state of the cluster. It interfaces with the cluster manager to allocate physical resources like vCPU and memory, and it launches the executors. The executors run the tasks and report back their results and state to the driver. The cluster manager is responsible for maintaining the cluster of nodes that run the Spark application.

Key Benefits

Cloud Native: Designed from the ground up for 'born in the cloud' workloads, Ampere Altra can deliver much higher price-performance over its x86 peers.

Consistency and Predictability: Ampere Altra processors that are designed for cloud native usage provide consistent and predictable performance of Spark and specifically for bursting workloads.

Scalable: Predictable performance under high utilization. Ampere Altra processors have high core counts with compelling single-threaded performance combined with consistent frequency for all cores. This makes Spark scale up and scale out efficiently.

Power Efficient: Industry-leading energy efficiency allows Ampere Altra processors to hit competitive levels of raw performance while consuming much lower power than the competition.

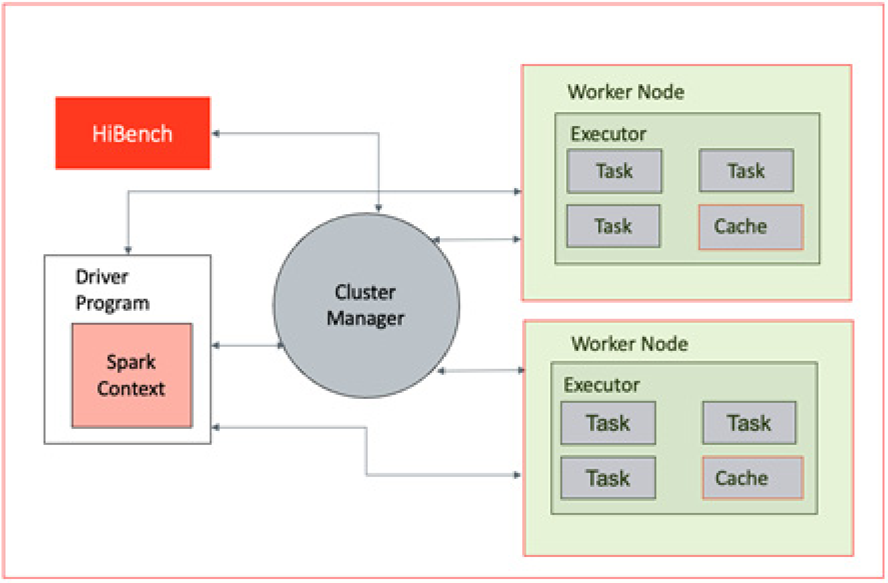
Demo Configuration

We use the HiBench benchmarking tool and run the Spark TeraSort benchmark. HiBench is a big data benchmark suite that helps evaluate different big data frameworks like WordCount, TeraSort, and so on, in terms of speed, throughput, and system resource utilization.

TeraGen is used to generate a dataset of 500 GB, and then the data is sorted using TeraSort capturing throughput in MB/s.

Cluster Details

ITEM	DESCRIPTION
Nodes	3 x Supermicro ARS-210
CPU	Ampere Altra® Max M128-30
CPU Cores	128 per node (50 used for Spark)
Memory	64 GB x 8 DIMMS per node
Disks	4 x 2 TB Samsung drives per node (two drives used for Spark)
Operating System	Ubuntu 22.04
Spark	Spark 3.3.1 with HDFS



For additional information, visit the Ampere Solutions Portal.