WHITE PAPER

Fusion ioMemory™ PCIe Solutions from SanDisk® and Sqrll make Accumulo Hypersonic

SanDisk®
a Western Digital brand

Western Digital Technologies, Inc.
951 SanDisk Drive, Milpitas, CA 95035

www.SanDisk.com

Table of Contents

Executive Summary **3**

Introduction 3

Accumulo Technical Overview **4**

About The Technologies **5**

Fusion ioMemory PCIe Solutions 5

Sqrrl 5

Accumulo on Fusion ioMemory PCIe Devices 5

About The Test **6**

Environment 6

Accumulo-Optimized Yahoo! Cloud Serving Benchmark (YCSB) Test 6

Test Configuration 6

Test Results 7

Summary **9**

Executive Summary

Replacing hard disk drives (HDDs) in Accumulo servers with the persistent flash memory of Fusion ioMemory™ PCIe Application Accelerators enables Accumulo architects to increase the performance density of cluster nodes up to 10x over disk-based nodes, while reducing average latency by nearly 90%. This enables architects to shrink system footprint and reduce costs and maintenance, while increasing performance to improve service levels.

Introduction

Most people think of Hadoop as the de facto tool for processing large volumes of data (Big Data). While Hadoop has many valuable properties, such as its predictable scalability and availability, government agencies require access controls and data management features that make deploying Hadoop challenging. Accumulo, originally built by the NSA and now managed by the Apache Foundation, addresses these unique processing needs and is now being adopted by companies in financial and healthcare industries that share similar requirements.

Accumulo is a distributed key/value store based on the Google BigTable design. Accumulo runs on top of Hadoop and provides unique cell-level access control that is not commonly available in other NoSQL databases. In addition, Accumulo provides data management features such as iterators, which provide key data retrieval functions. Accumulo has a significantly different performance profile than the traditional MapReduce workloads commonly found in Hadoop. Unlike Hadoop MapReduce that often drives high bandwidth sequential access workloads, Accumulo has a more random workload. This random access typically requires conventional Accumulo database architects to overprovision DRAM to cache the hot data set and maintain high performance. This strategy works well when the hot dataset fits entirely in DRAM, but as soon as the system is forced to access cold blocks from disk, performance can slow.

Fusion ioMemory PCIe solutions offer an alternative to the conventional Accumulo DRAM-heavy cluster architecture. Fusion ioMemory PCIe solutions replace the performance-poor, disk-based Hadoop Distribute File Systems (HDFS) with high-performance, low-latency persistent flash. This allows Accumulo architects to more effectively design clusters based on application requirements instead of DRAM capacity. Using Fusion ioMemory PCIe solutions to provide predictable and consistently high performance across the entire database enables more efficient Accumulo architectures that require fewer nodes and less DRAM, which reduces power and cooling costs. This paper describes the results of tests of an Accumulo cluster built by Sqrrl and SanDisk, designed to achieve the best performance density possible.

Accumulo Technical Overview

Accumulo is a key value store based on Google's BigTable design. Accumulo's architecture consists of the following components:[1]

- Tablets: Partitions of tables consisting of sorted key/value pairs.
- Tablet servers: Manage the tablets, including receiving writes from clients, persisting writes to a write-ahead log, sorting new key-value pairs in memory, periodically flushing sorted key-value pairs to new files in HDFS and responding to reads from clients. During a read the tablet servers provide a merge-sorted view of all keys and values from the files it has created and the sorted in-memory store.
- Master: Responsible for detecting and responding to tablet server failure. The Master tries to balance the load across Tablet Servers by assigning the tablets carefully and instructing Tablet Servers to migrate the tablets when necessary. The Master ensures each tablet is assigned to exactly one Tablet Server, and handles many miscellaneous database administration requests. The master also coordinates startup, graceful shutdown and recovery of write-ahead logs when the tablet servers fail.
- ZooKeeper: Distributed locking mechanism with no single point of failure. Zookeeper is responsible for maintaining configuration information, naming, and providing distributed synchronization. (http://www.sqrrl.com/whitepaper/)

At the heart of Accumulo is the Tablet mechanism, which simultaneously optimizes for low latency between random writes and sorted reads (real-time query support) and efficient use of disk-based storage.

Accumulo accomplishes this through a mechanism derived from the Log Structure Merge Tree, in which data is first buffered and sorted in memory and later flushed and merged through a series of background compaction operations. These compaction operations buffer random write operations so that they become sequential operations on disk, boosting the I/O efficiency of a scalable storage solution.

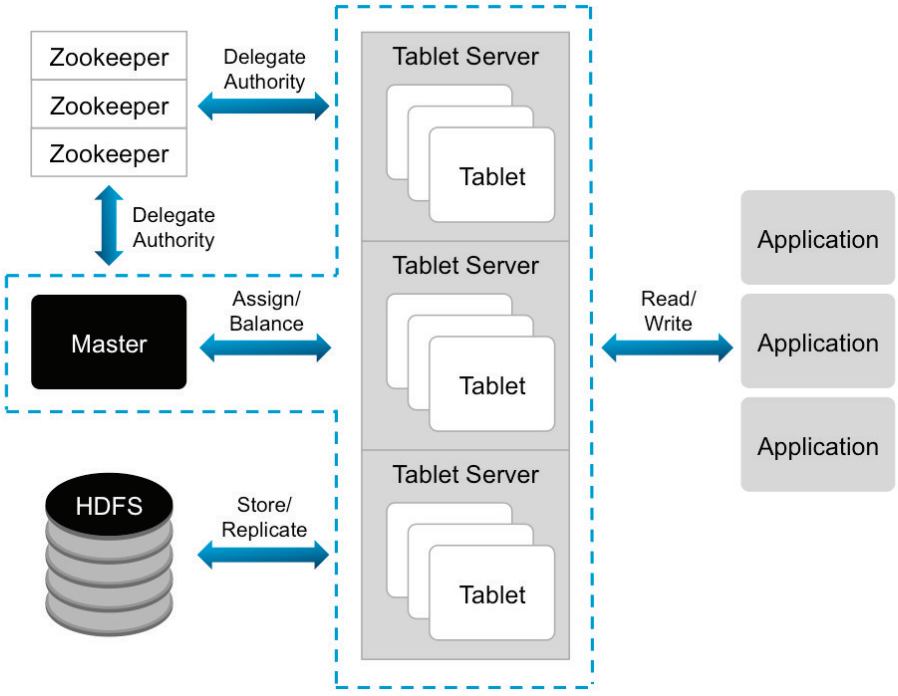

Figure 1. Accumulo Architecture

Major compactions run in the background to merge multiple files into one. The tablet servers determine which tablets to compact and which files within a tablet to compact. During this operation the entire data set is read and valid records are re-written. This generates significant file I/O if the working set is not entirely in DRAM. In clusters that use HDDs, resource contention between clients and the compaction process typically results in a drop in overall system performance.

In a conventional Accumulo cluster, the critical component for performance is DRAM. There must be sufficient memory available across the cluster to hold the working set of records so that reads from HDDs are minimized. Unfortunately servers configured with very large DRAM configurations quickly exceed most budgets. Despite being a commodity component, DRAM modules at high densities approach $35-$45/GB, limiting the cost-effective range for DRAM in a server to 48GB to 128GB per node. Organizations simply don't have sufficient rack-space to continue to scale at that relatively low-density per GB. Also, at some point, there is a physical limit to how many nodes can be added to a cluster.

About The Technologies

Fusion ioMemory PCIe Solutions

The Fusion ioMemory PCIe tier balances capacity and performance, operating at near DRAM speeds but with the persistence required for database storage. SanDisk provides direct PCIe bus access to NAND flash memory that does not rely on legacy block-storage protocols and interfaces. It is a persistent memory tier that is exposed to Hadoop as block storage via the operating system, thereby achieving the lowest access latency possible in a persistent-storage medium.

Sqrrl

Sqrrl is a Big Data software company whose employees have dealt with the world's largest, most complex, and most sensitive datasets for the last decade. Sqrrl's software product, Sqrrl Enterprise, is the most secure and scalable Big Data platform for building real-time applications and is powered by Apache Accumulo™ and Hadoop. Sqrrl Enterprise extends the capabilities of Accumulo with additional data ingest, security, and real-time analytical features that unlock the power of big data.

Accumulo on Fusion ioMemory PCIe Devices

By outfitting Accumulo with Fusion ioMemory devices, architects can change the traditional scaling properties of an Accumulo cluster from DRAM to Fusion ioMemory PCIe flash capacity, increasing per server density up to 10x. Fusion ioMemory PCIe devices replace locally attached HDDs asprimary storage, but offers near-DRAM speeds. Since all data stored is microseconds away, no penalties are paid for cold reads from primary storage. While traditional Accumulo nodes may utilize 128GB of DRAM per node to cache the working set, a Fusion ioMemory PCIe-equipped server can support 12.8TB or more. This ensures that the entire dataset is microseconds away from the CPU and performance is reliable and predictable. In addition compaction operations no longer slow normal operations. Each Fusion ioMemory PCIe module can support 1.5GB/s of bandwidth and hundreds of thousands of transactions per second. This ensures there is enough I/O capacity to support normal operations and compactions.

About The Test

Environment

Tests were run on a 5-node cluster. Each node was configured with a 16-core Xeon 2690 @ 2.90GHz processor, and 64GB of DRAM. Tests compared the performance density of placing primary storage on a 12-disk JBOD to a Fusion ioMemory ioScale PCIe card at 1650GB. The diagram below illustrates the test systems:

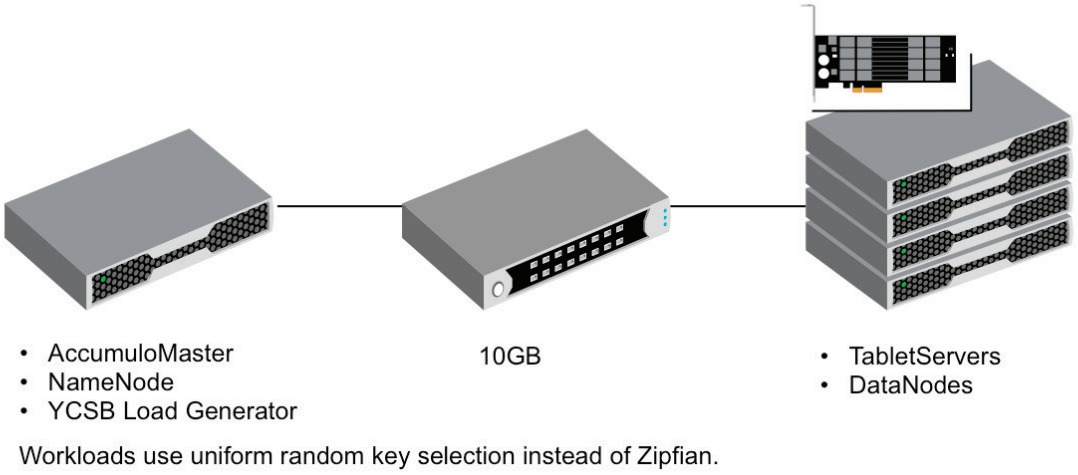

Figure 2. Fusion ioMemory PCIe-based System is Simpler and Easier to Manage

Accumulo-Optimized Yahoo! Cloud Serving Benchmark (YCSB) Test

Sqrrl provided a modified version of YCSB with an Accumulo client implementation. The client stores each field of the YCSB record in a separate Accumulo key/value entry. Record updates are organized into one Accumulo Mutation per record, and are sent to Accumulo with the BatchWriter.

Test Configuration

Using YCSB, a 500 million record dataset was loaded into Accumulo (~2TB) and tests were run to measure record update and retrieval times. Three primary workloads were tested:

- Workload A: 50/50 Read Write Mix
- Workload B: 95/5 Read Write Mix
- Workload C: 100% Read

Accumulo was configured with a standard replication factor of three. During the test a random distribution was used to ensure that the blocks were accessed from primary storage devices rather than the 320GB of DRAM in the cluster. Figure 3 shows a 50/50 read write mix @ 256 client threads, where the majority of transactions were served from primary storage, either 12 HDDs or a single Fusion ioMemory ioScale PCIe card at 1650GB.

Test Results

As the graph below illustrates, ...the Fusion ioMemory PCIe-based system maintained and average of 48,000 transactions a second while the traditionally configured DRAM/HDD based system averaged just 4,759 transactions per second. This shows that under the exact same conditions the Fusion ioMemory PCIe-based system can deliver 10x the transactions per second. Architects can use this increase in performance density to maintain performance on as little as 1/10th the hardware. Most rack-mounted servers can be configured with 12TB or more of Fusion ioMemory PCIe cards.

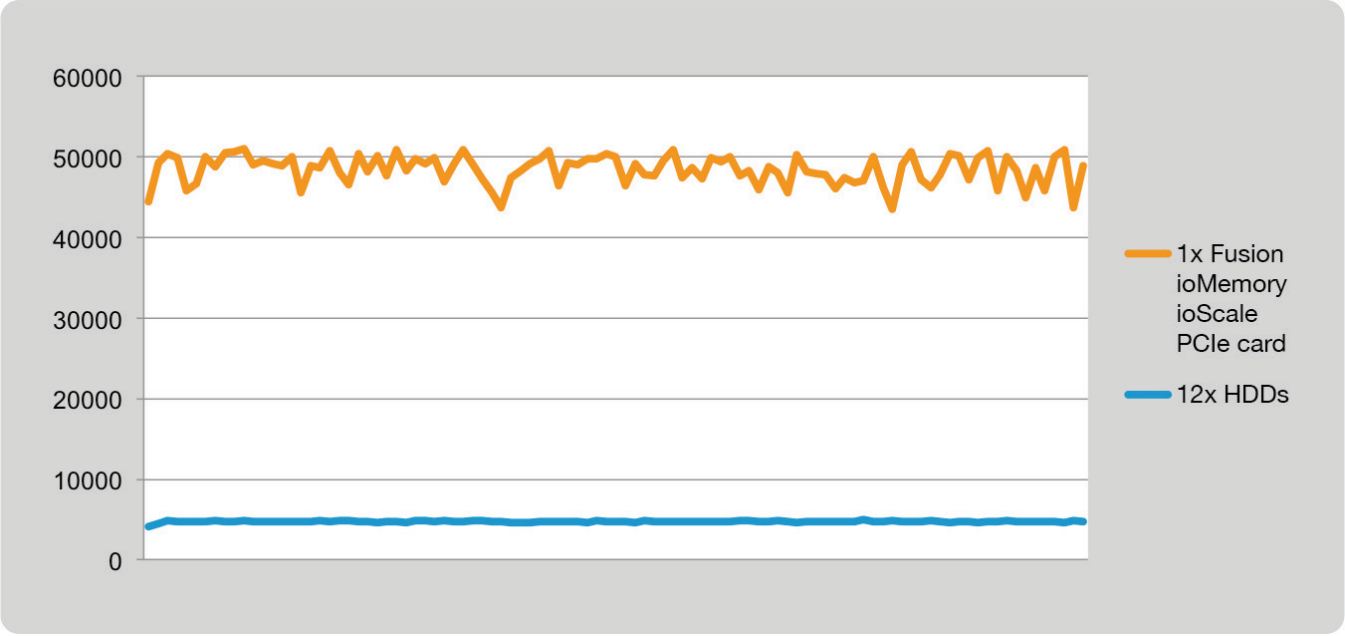

Figure 3. Fusion ioMemory PCIe-based System Delivers 10x Performance[2]

Latency plays a key role in application responsiveness. When CPUs have to wait for storage I/O, processors begin context switching, which can make the CPUs less efficient. Figure 4 shows the average latency across all workloads. As the graph in Figure 4 below illustrates, the HDD-based system has over 8x higher latency on average across all workloads than the Fusion ioMemory PCIe-based system.

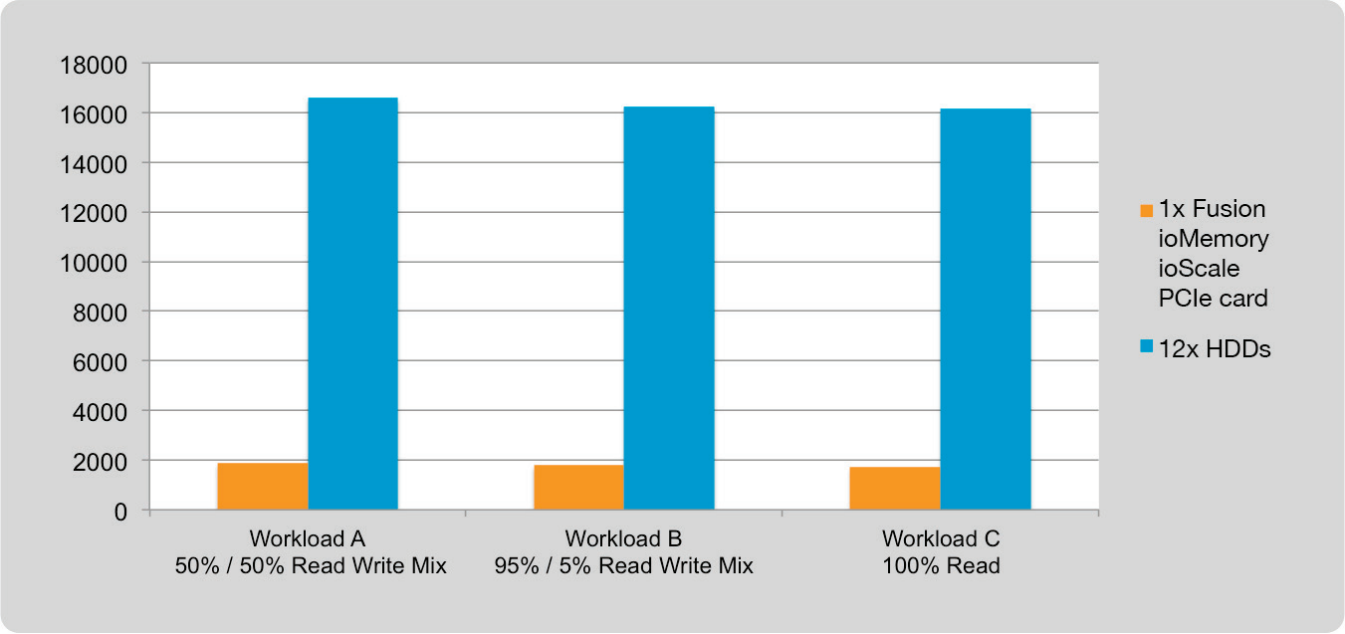

Figure 4. Fusion ioMemory PCIe-based System 1/8th Latency for Faster Response Times[3]

While the average latency of Fusion ioMemory PCIe devices is under 2 milliseconds at 64 client threads, the HDD-based system leaves the CPUs waiting for 16 milliseconds, on average, for every transaction. As the number of client threads increases, the workload to serve more threads becomes increasingly random. These results show that the Fusion ioMemory PCIe-based systems are much more resilient than HDD-based systems.

Summary

Fusion ioMemory PCIe solutions give architects a compelling alternative to traditional DRAM-based Accumulo clusters. With up to 10x DRAM capacity per node, entire working sets can be placed in consistent, high-performance flash to eliminate latency and even consolidate nodes.

The comparison between Fusion ioMemory PCIe devices and HDD-based clusters shows compelling results. Our testing with Accumulo demonstrates these benefits:

- Provides microsecond access times for random workloads.
- Eliminates the performance impact of compaction processes.
- Delivers the required performance at a fraction of the cost per gigabyte of high-density DRAM modules.
- Consumes significantly less power per GB.
- Reduces costly scale out.

Finally, for developers with limited engineering staff, Fusion ioMemory PCIe solutions allow technical staff to focus on solving problem rather than spend time optimizing system I/O between DRAM and HDD.

FOR MORE INFORMATION

Contact a SanDisk representative, 1-800-578-6007 or fusion-sales@sandisk.com.

1 Visit http://www.sqrrl.com/whitepaper/ for a technical overview of Accumulo.
2 50/50 Read/Write Mix @ 256 threads
3 Average latency in microseconds for 64 threads

The performance results and cost savings discussed herein are based on internal testing and use of Fusion ioMemory products. Results and performance may vary according to configurations and systems, including drive capacity, system architecture and applications.