Aerospike and ScaleFlux's ultra-high-performance solutions help Tongdun's ultra-large-scale core database system

Abstract

Tongdun has investigated several hardware and configuration options for supporting its high-transaction-volume deployment of the Aerospike database to support the rapidly growing use of Tongdun's services. While performance of the database deployment is important, total cost of the deployment is also critical.

This paper discusses the overall needs for the deployment, the various Aerospike configurations, and hardware options assessed to best meet Tongdun's needs to balance performance, cost, complexity, and scalability. Through this assessment, Tongdun determined that using Aerospike Enterprise Edition in combination with the ScaleFlux Computational Storage Drives met their performance and capacity needs while cutting the number of servers needed in half.

TONGDUN
www.tongdun.cn

AEROSPIKE

CSD 2000 Series

ScaleFlux CSD 2000 with Aerospike Enterprise Edition

10x Latency Improvement

50% infrastructure reduction

2x-3x Transactions Per Second

About Tongdun

Tongdun Technology is a leading company in the field of intelligent analysis and decision-making in China. Tongdun has developed a deep understanding of generating insight from data, utilizing the three core technology systems of artificial intelligence, cloud computing, and big data. , It combines advanced technologies such as deep learning and federal learning with in-depth business scenarios to provide intelligent analysis and decision-making services for industries such as finance, insurance, Internet, government affairs, retail, logistics, etc. Tongdun's services empower and inspire its customers, to make better decisions. Up to now, more than 10,000 corporate customers have chosen Tongdun's products and services. These customers cover a wide variety of vertical markets, spanning across 22 major industries and 118 subdivisions.

At present, the average daily API call volume of Tongdun exceeds 100 million, the peak value exceeds 200 million. The daily fraud intelligence monitoring identifies over 1 million potential risks per day, in addition to the average daily interception of over 1.5 million fraudulent activities (e.g. IP, mobile phone, equipment, and other related activities). Every year, it helps cooperative customers protect trillions of dollars of funds by providing security for their accounts and for over 20 billion transactions.

The Challenge – Meeting Tongdun's high performance requirements and rapidly scaling OLTP databases

Current OLTP systems usually require very high performance-millions of transactions per second (TPS) at single millisecond, or even sub-millisecond latency. Key examples include financial transactions and advertisement placements. In bank payment, credit card anti-fraud and risk control projects, the hardware infrastructure budget for the database is limited, and hundreds

of rules need to be executed within 100 milliseconds. In the precise placement of advertisements, matching users with advertisements and then delivering the content is often completed within 10 milliseconds. There are some applications that require such high performance under a large data scale, such as balance inquiry in telecommunications.

Traditional databases, such as Oracle, MySQL, and PostgreSQL, require a lot of hardware equipment and many optimizations to achieve this performance. Obviously, the cost of hardware, secondary development and maintenance will be relatively high.

Another commonly used solution is to add a cache layer to the production database or use DRAM / in-memory database. The disadvantage of this two-level architecture is that when there are many write or update transactions, the synchronization between the cache layer and the production database becomes a bottleneck, which often causes significant latency spikes, and prevents the system from achieving the performance needs. Timely synchronization between the the cache and production layers is crucial to avoid data loss when there is a failure in the system. The synchronization challenges also come into play with the in-memory database option since the volatile memory needs to be synchronized with non-volatile storage to prevent data loss.

As a domestic first-line risk control enterprise, Tongdun has tight requirements for database response time (i.e. transaction latency) and reliability. At the same time, because of frequent business changes, system operation and maintenance need to be changed agilely without causing cost increase. Tongdun has been using Aerospike Community Edition and ScaleFlux's CSS 1000 to support the stable operation of its core business.

With the rapid increase in business scale, Tongdun needs the Aerospike system to support larger-scale concurrent access. The amount of data has increased from 10s of TB to 100s of TB.

The Solution – Aerospike Enterprise Edition and ScaleFlux CSD 2000

Both Aerospike and ScaleFlux have products to help meet these needs for massive scaling and high performance. Aerospike Enterprise Edition's All-Flash solution can use Flash instead of DRAM to store the Key-Value pairs. This model improves system robustness without adding latency. ScaleFlux has launched a new generation of product, CSD 2000, with and integrated transparent compression engine. When the compression ratio of the data is at least 2:1, random write performance with CSD 2000 can even reach the level of Intel Optane. ***Tongdun tested the combination of Aerospike Enterprise Edition with CSD 2000 to assess (1) how well the combination could support Tongdun's large-scale database applications running on all-Flash memory and (2) if this combination could replace expensive and difficult-to-manage all-memory solutions.***

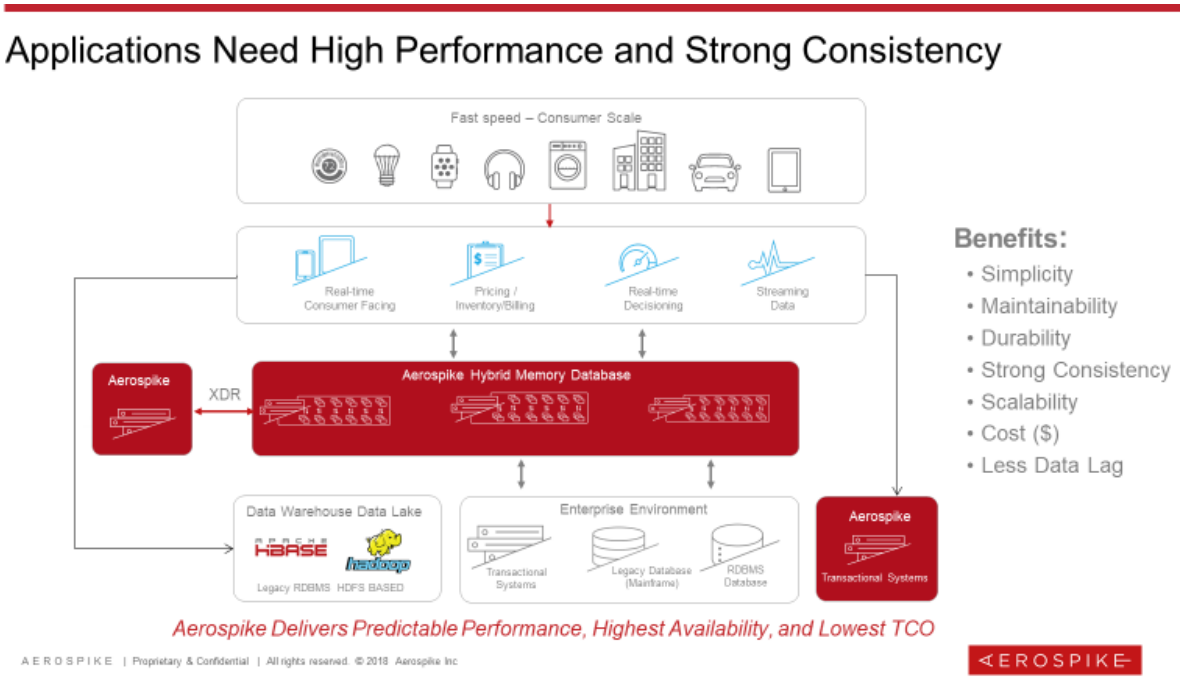
In Tongdun's test, 4 servers were used, each with 32 cores and 128GB memory, plus ScaleFlux CSD 2000 6.4TB SSD. The test compared query performance on Cassandra with the performance on Aerospike Enterprise Edition. Queries which completed in an average of 10 milliseconds on Cassandra (indexes and data), all completed in an average of 1 millisecond with the configuration of ScaleFlux CSD 2000 and Aerospike Enterprise Edition – ***a 10x improvement in latency!*** While CSD 2000 greatly improved performance, it also delivered significant cost savings. The CSD 2000 achieved a compression ratio of 5:1 with the test data. Through the built-in transparent compression feature of CSD 2000, ***data that originally required 5 SSDs to store, now only needs 1 drive!*** Additionally, the compression function is completely transparent to the upper application, without any code change.

Therefore, the Aerospike Enterprise Edition database, coupled with a high-performance SSD, such as ScaleFlux CSD 2000, can achieve very high-performance requirements. At the same time, it can achieve this single-

millisecond latency and millions of TPS of large data sets (hundreds of terabytes or even petabytes). Even the throughput of tens of millions of TPS. More importantly, the Aerospike + CSD 2000 solution gets rid of the two-level architecture of cache + production database and can use a much smaller cluster to achieve the same performance***, resulting in an order of magnitude savings in application costs.***

About the Aerospike Architecture Advantage

First, let us understand the advantages of Aerospike one-layer architecture. In this architecture, the data can be in memory or Flash memory (SSD) and there is only one copy of data (in comparison to the two-level architecture mentioned earlier which has a copy of data *both* in memory *and* in Flash memory). When the data is on the SSD, the written data will be persistent immediately after the transaction is completed (commit). When there is a network or server failure, data will not be lost. Therefore, while achieving high throughput and low latency performance, it also provides persistence and strong consistency. This greatly simplifies the database development, maintenance, launch and production, which reduces the total cost of ownership (TCO). This architecture also shows that Aerospike and traditional relational databases like Oracle/MySQL are complementary. Many customers use Aerospike to implement highly repetitive and high-performance operations that originally run in Oracle/MySQL/PostgreSQL while still using Oracle/MySQL/PostgreSQL to run complex, ad hoc operations.



Some of Aerospike's technical characteristics related to performance:

- Shared Nothing architecture, no master and slave nodes, no hot spots.
- Multi-threaded concurrency, NUMA support, full use of all server resources.
- Primary key consistency provides strong consistency guarantee, linearization and session consistency.
- Smart client, a data jump, without load balancer.
- Intelligent cluster management, zero manual intervention.

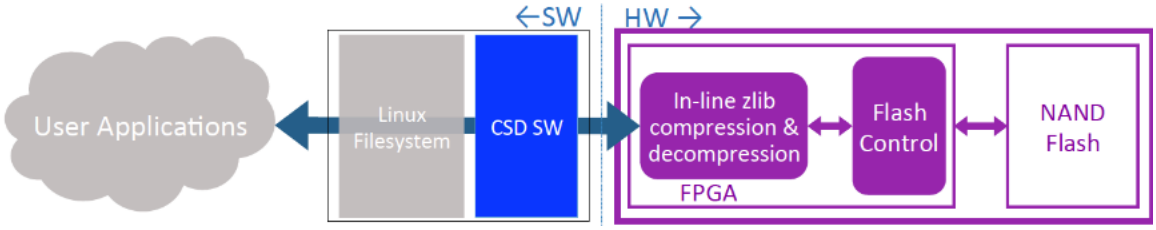
ScaleFlux CSD 2000 with Transparent Compression

The ScaleFlux CSD 2000 series is the only enterprise-level PCIe SSD product with transparent compression and decompression features, bringing excellent performance, high scalability, and cost savings to the deployment of mainstream Flash memory. The CSD 2000 series achieves data path compression by combining up to 8TB of the latest 3D NAND Flash memory technology with a hardware-accelerated computing engine. There is no need to modify the system kernel or application code to utilize the compression. The compression function does not consume the system's CPU and memory resources and has no penalty on performance. CSD 2000 achieves stable and high-speed data read/write speed and consistent low latency. In terms of mixed read and write workloads, it is 40%-70% higher than the industry-leading NVMe SSDs, which improves application performance and saves 50%-80% Of Flash memory space (based on customers' testing).

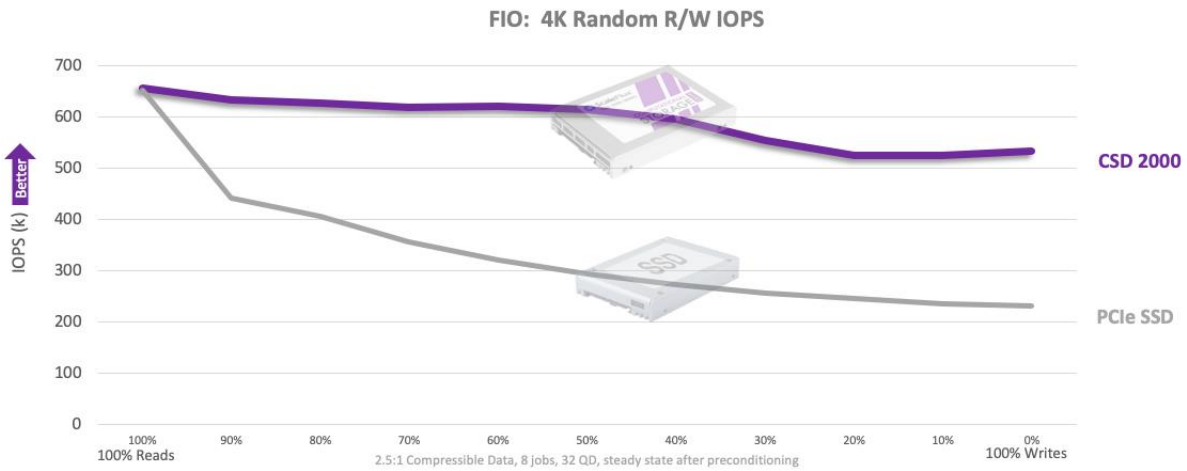
Transparent compression can improve SSD write I/O performance, improve long-tail latency, and also improve SSD write life. Compression is the first step in writing data to SSD. Through compression, the amount of data finally written to the physical medium is reduced, which not only reduces the bandwidth occupation of write IO, but also reduces the bandwidth and endurance consumed for background processes such as garbage collection , thereby reducing IO latency, increasing IO performance, and reducing write amplification. Reduced write amp consequently improves the life of the CSD.

Transparent compression is a built-in function of CSD 2000. It does not require any additional operations from the CPU and does not require additional equipment for the server.

The architecture design of CSD 2000 is shown in the figure below. CSD 2000 uses FPGA to perform lossless data compression on the data IO path, which is completely transparent to the upper software stack (including the operating system and front-end applications). CSD 2000 includes two parts: host driver and physical SSD. The latter uses the same FPGA to perform Flash control functions and the compression/decompression functions.



In the actual IO test, compared with the ordinary SSD, with a certain data compression ratio (2.5:1), CSD 2000 can achieve 4KB random write performance that is over twice as high as that of ordinary SSD. The performance is stable and excellent across the span of read/write ratios, as shown in the figure below.



The perfect combination of Aerospike hybrid storage architecture and ScaleFlux high performance

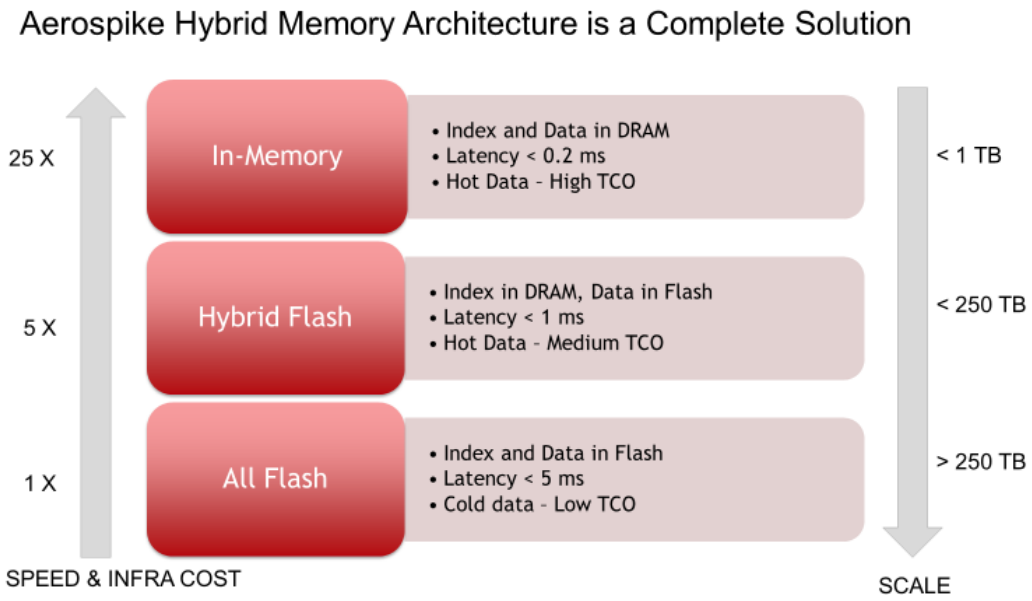
One of Aerospike's core technical advantages is SSD optimization, including:

- The hash of the data is distributed across all nodes and SSD
- Directly read and write to the original device beyond the file system
- Large block size read and write
- Optimized read and write distribution to avoid fast aging of SSD hot spots

Aerospike takes full advantage of the high performance provided by SSDs, such as ScaleFlux's CSD 2000. The optimization between them provides a complete solution. Aerospike can be performed in three main modes:

- Full memory àindex and data are in memory
- Hybrid Flash memory àindex in memory, data in Flash
- All-Flash à-index and data are in Flash

The following table lists the performance in the three modes:



How do we choose which mode to use?

From the above figure, we can see that the full memory model provides the fastest performance, but its overall cost of ownership is also the highest, and it does not provide durability.

The index of each piece of data in Aerospike occupies 64 bytes of memory, so when each piece of data is relatively small, such as a few KB, the memory is more likely to become a bottleneck of storage capacity. When each piece of data is relatively large, memory generally does not become a bottleneck. This is just a rough description. The accurate calculation needs to be performed based on your data and server configuration using Aerospike's capacity planning guide. According to the above analysis, the hybrid Flash mode is more suitable when each piece of data is relatively large. Its throughput will be about 10% larger, and the write latency will be about 1 millisecond, which is 3-5x faster than all Flash memory. If each piece of data is relatively small, All-Flash may provide a better price-performance ratio. From the perspective of system restart, Aerospike can quickly restart in a few seconds or tens of seconds after the server restarts in the All-Flash mode because there is no need to rebuild the index. In the hybrid Flash mode, when the Aerospike instance restarts, the index in the shared memory can be directly read to achieve a fast restart; but the index needs to be rebuilt after the server restarts.

As an example to illustrate the above analysis:

- Database Parameters:
  - Cluster Capacity: 100TB Aerospike database, and each piece of data is 2KB in size, that is, a total of about 50 billion pieces of data.
  - Compressibility: Assume that the compression ratio of data is 3:1.
  - Cluster Performance: 10 million TPS with 95% of Reads completed within 1 millisecond.
- Server Configuration:
  - Each server has 256GB of memory and a high-performance disk with an effective capacity of 4TB.
- Challenge:
  - With this database, it is almost impossible to store all in memory.
- Solution options:
  - Hybrid-mode:
    - Requires about 40 nodes due to the memory and storage constraints.
  - All-Flash mode with CSD 2000:
    - Requires only 20 nodes to achieve the storage & performance targets
- **Result:**
  - **50% reduction in the infrastructure needed to service the 100TB deployment**

Note: The above estimate is based on a single Aerospike node running on an ACT with a high-performance single SSD disk (such as ScaleFlux CSD 2000) that can reach 500k TPS.

Summary

In general, Aerospike provides various solutions you need, especially when the amount of data is relatively large, many customers will choose hybrid Flash or All-Flash mode. At present, the technology of SSD is advancing by leaps and bounds. ScaleFlux innovatively implements transparent compression in the storage controller, allowing the random write performance of applications to leap forward. Aerospike's cutting-edge technology-based optimization for these high-performance SSDs can provide very high performance-million or even tens of millions of TPS and single-millisecond ultra-low latency. Tongdun has a relatively tight total budget for hardware / infrastructure. With the combination of ScaleFlux CSD 2000 and Aerospike Enterprise Edition, Tongdun can make full use of its big data, reduce the complexity of product development and maintenance, and meet the production system's requirements for latency and TPS all while staying under budget.

"By using the all-Flash deployment of Aerospike Enterprise Edition with ScaleFlux Computational Storage Drive, we can achieve a 110-224% improvement in our deployment's performance, while providing storage with high IOPS and continuous low latency high-quality performance comparable to memory "-Tongdun Technology

(End of Document)