WHITE PAPER

Couchbase Server: Accelerating Database Workloads with NVM Express*

Database Performance
Intel® Solid-State Drives
Intel® Xeon® Processor E5-2600 v3 Product Family

Couchbase Server 4.0, a NoSQL database, shows dramatic performance improvements across workloads when upgrading local storage based on solid-state drives from SATA connectivity to Non-Volatile Memory Express* (NVMe*).[1] These performance improvements are generated in part by the architecture of ForestDB*, a key-value-pair storage engine used as the default for the indexing service in Couchbase Server 4.0.

NoSQL databases play an increasingly significant role in meeting the needs of simplicity, speed, elasticity, and reliability for big data workloads. Couchbase Server is an advanced NoSQL database that is optimized for interactive web applications, designed explicitly to meet changing requirements by distributing data and I/O efficiently across clusters of servers.

The ForestDB storage engine introduced with Couchbase Server 4.0 enhances the operation of Couchbase Server significantly. Its architecture is specifically built for use with flash memory, in contrast to the limitations that are common with other database platforms that were created for mechanical hard-disk drives (HDDs).

The platform architecture of the Intel® Xeon® processor E5 v3 product family provides advantages that contribute to the performance of Couchbase Server implementations. Non-Volatile Memory Express* (NVMe*) architecture, which is the basis of the Intel® Solid-State Drive (Intel® SSD) Data Center P3700 Series, provides additional benefits, by overcoming the potential bottlenecks associated with disk-access latency in SAS and SATA-based solid-state drives (SSDs).

This paper reports on testing results that show the value of NVMe drives relative to SATA SSDs across different scenarios using ForestDB.

Reducing Disk-Access Latency for SSDs with NVMe

SSDs provide for dramatically reduced disk-access latency compared to conventional mechanical HDDs, by avoiding the latency limitations associated with the physical movement of the read/write head. In addition, the NAND memory used in SSDs inherently supports parallel data access. While replacing the magnetic media of HDDs with non-volatile memory offers substantial benefits, however, first-generation SSDs typically relied on the SATA interface bus for system connectivity. The associated translation mechanism limits performance on systems that use SATA-based SSDs.

PCI Express* (PCIe*) solutions for SSD system connectivity improved on SATA solutions, bringing storage closer to the CPU. Furthering that concept, an industry consortium led by Intel introduced the NVMe specification, which replaces the proprietary PCIe solutions that existed before, standardizing the register set, feature set, and command set, with a solution that is designed to scale from client systems to the largest enterprise systems.

Table of Contents

Reducing Disk-Access Latency for SSDs with NVMe 1

Intel® Platform Innovations that Drive Up Database Performance 2

Introducing Couchbase Server 4.0: Next-Generation NoSQL Performance 3

ForestDB. 3

Multi-Dimensional Scaling (MDS) 3

Test Results: ForestDB with SATA SSDs versus NVMe 4

Key/Value Store Testing 4

Index Service Testing 5

Parallel Benchmark Throughput Testing 6

Conclusion. 7

NVMe is a specification for accessing SSDs over PCIe. Unlike the Advanced Host Controller Interface (AHCI) protocol it replaces, NVMe is engineered specifically for use with non-volatile memory such as SSDs. Therefore, NVMe delivers significant throughput and latency benefits by eliminating inefficiencies that were inherent to the use of AHCI with SSDs.

Using the Intel SSD Data Center Family for PCIe, an off-the-shelf server can currently accommodate up to 10 hot-swappable drives with up to 2 terabyte capacity each, taking advantage of the NVMe standard. Moreover, NVMe-compliant drives are available from a growing number of vendors in addition to Intel, helping deliver a truly open-standards solution architecture.

NVMe helps the most demanding big data workloads take full advantage of the parallelism offered by the Intel Xeon processor E5-2600 v3 product family, which provides up to 36 physical cores and 80 PCIe lanes in a two-socket server. Using a single physical interface, the same connector also provides the flexibility, for example, to swap in a lower-cost SATA boot drive. Benefits afforded by the Intel SSD Data Center P3700 Series include the following:

- **Dramatically improved data-transfer speed** compared to SAS and SATA-based SSDs.
- **Optimization for multi-core processors**, with features such as deeper command queues, parallel interrupt processing, and lockless thread synchronization.
- **Advanced software support**, with drivers incorporated in the Linux* kernel since March 2012.
- **Enhanced reliability, availability**, and serviceability, with rigorous qualification and compatibility testing, plus business-critical dependability.

Intel® Platform Innovations that Drive Up Database Performance

As the execution engine at the foundation of systems that power Couchbase Server database solutions, the Intel Xeon processor E5-2600 v3 product family provides innovations that help deliver advanced results. High-level comparisons with the previous generation are summarized in Table 1.

Table 1. Generation-to-generation comparison of the Intel® Xeon® processor E5-2600 product family.

	INTEL® XEON® PROCESSOR E5-2600 V3 PRODUCT FAMILY	INTEL® XEON® PROCESSOR E5-2600 V2 PRODUCT FAMILY
Cores/Threads per Socket (max)	18/36	12/24
Last-Level Cache (max)	45 MB	30 MB
Intel® QuickPath Interconnect Speed (max)	9.6 GT/s	8.0 GT/s
Memory Speed (max)	DDR4-2133 MHz	DDR3-1866 MHz

Architectural features and capabilities that are particularly beneficial to the testing reported on in this paper include the following:

- **Increased compute density** compared to previous generations is provided by up to 18 cores and 36 threads per socket, plus 45 MB of last-level cache that enables large amounts of frequently used data to be kept available for high-speed access to the processor.
- **Support for DDR4 memory** running at up to 2133 MHz improves performance on memory-intensive workloads, with up to 1.4x higher bandwidth compared to predecessors.[2] Each two-socket server supports up to 24 DIMMs, helping satisfy the memory demands of big data workloads.
- **Intel® Data Direct I/O Technology (Intel® DDIO)** increases the efficiency of system-level data flows by enabling the network controller to communicate directly with the processor cache, which acts as the primary destination and source of I/O, instead of relatively slower system memory.
- **Intel® Turbo Boost Technology 2.0** dynamically increases processor frequency to respond to workload peaks, taking advantage of power and thermal headroom as conditions permit.

Introducing Couchbase Server 4.0: Next-Generation NoSQL Performance

Couchbase Server is a document-oriented database, designed for streamlined elasticity that allows nodes to be added or removed with automatic rebalancing on a live cluster. It responds automatically to changing workloads, distributing data and I/O resources across both physical and virtual nodes as needed. Key capabilities and benefits of Couchbase Server include the following:

- **Simple**. There is no need to create and manage schemas, or to normalize, shard, or tune the database.
- **Fast**. Low latency and high throughput are maintained as workloads scale up and out.
- **Elastic**. Data and I/O are automatically distributed as changing application needs warrant.
- **Reliable**. Robust monitoring capabilities and simplified maintenance help ensure continuous uptime.

ForestDB

Although the industry is in the process of transitioning to flash memory as the primary data-storage medium of choice, database mechanisms have evolved over the course of decades for use with mechanical hard-disk drives. That legacy often leads to suboptimal performance for established software when using it with flash storage. The next-generation key-value engine used for the indexing service in Couchbase Server 4.0, ForestDB, is explicitly designed for use with flash storage, to overcome those limitations. Key areas where ForestDB is optimized for use with SSDs include the following:

- **Adaptation of the SSD Flash Translation Layer**. By adapting the SSD Flash Translation layer, ForestDB is able to read and write data more efficiently.
- **Integration of asynchronous I/O library (libaio)**. This advance helps ensure that ForestDB read and write operations do not interfere with each other, increasing the efficiency of both.

ForestDB is an open-source project, the source code for which can be downloaded at https://github.com/couchbase/forestdb.

Multi-Dimensional Scaling (MDS)

Couchbase Server has redefined the way enterprises scale distributed databases with the option of MDS. It separates, isolates, and scales individual services—query, index, and data—to improve application performance and increase resource utilization. Couchbase Server 4.0 is the first and only distributed database capable of scaling with the speed and precision required by enterprise applications with variable workloads.

MDS enables enterprises to optimize hardware by allocating resources based on the workload of a specific service, and to avoid resource contention by performing queries, maintaining indexes, and writing data with different nodes. It is inefficient to require participation from every node to perform a query or maintain an index. Couchbase Server 4.0 solves this problem by scaling data independent of queries and indexes. The benefits of MDS include the following:

- Improve performance by separating services to avoid resource contention
- Improve performance by scaling the data service, not the query and index services
- Improve resource utilization by separating services to optimize the hardware

MDS allows data, indexing, and query workloads to scale independently within a cluster. Conventional scaling is homogeneous, meaning that each node in the cluster participates in index, query, and data operations, with the workloads for each distributed equally across the nodes, as illustrated in Figure 1.

This model has the inherent limitation that the services compete for resources and interfere with each other. In addition, services must scale equally across nodes, regardless of the actual scaling requirements of each, potentially preventing the system from making optimal use of hardware resources. Couchbase Server supports this approach, but it also offers an alternative enabled by MDS, as illustrated in Figure 2.

Each service is deployed to an independent zone within the cluster, represented by the blue portions of nodes 1–8 in Figure 2. The green portions represent the sample scalability potential for each service, with query and index services scaling up on a smaller number of more powerful servers, and the data service scaling out onto additional nodes. MDS improves performance and throughput by combining the advantages of scale-out and scale-up models.

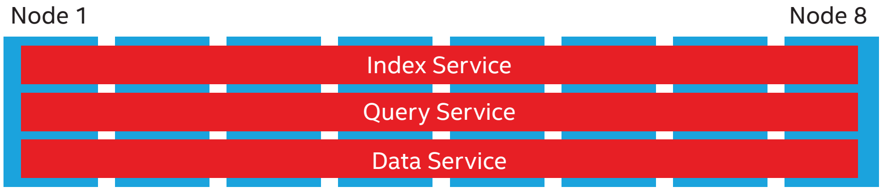

Figure 1. Homogenous scaling of services.

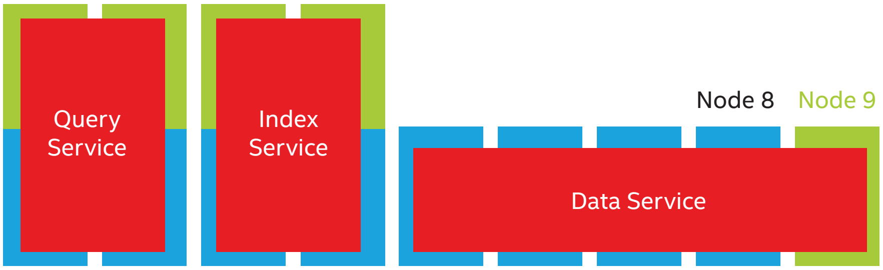

Figure 2. Independent scaling of services, enabled by Multi-Dimensional Scaling.

Test Results: ForestDB with SATA SSDs versus NVMe

To quantify the value of storage innovations in terms of performance, Couchbase undertook three bodies of testing, on hardware that included conventional and mechanical HDDs, as well as both SATA-based SSDs and NVMe drives.[1] These testing scenarios were chosen to be applicable to a wide range of customer organizations and usages:

- **Key/value store testing**. Implementing ForestDB as a read/write caching layer.
- **Index service testing**. Index simulation using a global secondary index.
- **Throughput testing**. Pushing database performance limits using a parallel benchmark.

NOTE: Couchbase Server 4.0 and ForestDB are still in development, and the final product versions may not provide identical results to those discussed in this document.

Key/Value Store Testing

The first body of testing, key/value store, is analogous to a real-world shopping cart application on an e-commerce website. In this scenario, high throughput and low latency enable users to interact with the site efficiently, which facilitates fast completion of shopping tasks and helps drive large amounts of transactions for the site, enhancing profitability.

For the key/value store testing, four reader threads and one writer thread were allocated, with a corresponding traffic mix of 80 percent read operations and 20 percent write operations. This testing involved a single ForestDB instance, in a single file/single benchmark (synchronous write) scenario. The workload applied had the following characteristics:

- **Average key size**: 48 bytes
- **Working data size**: 100 GB (100 million documents at 1 KB each)
- **Buffer cache allocated**: 30 GB

The test results are illustrated in Figure 3. In terms of both read and write throughput, results with NVMe were better than 1.5x that of the SATA case. As a reference point, the two SSD throughput results were both orders of magnitude better than the HDD results. To provide a more complete comparison between SATA and NVMe, latency comparisons show even more significant results, given that the buffer cache is only 30 GB compared to the 100 GB working data size, meaning that data access needs to go to disk. In this scenario, NVMe is nevertheless able to deliver approximately a 40 percent reduction or better, in both read and write latency.

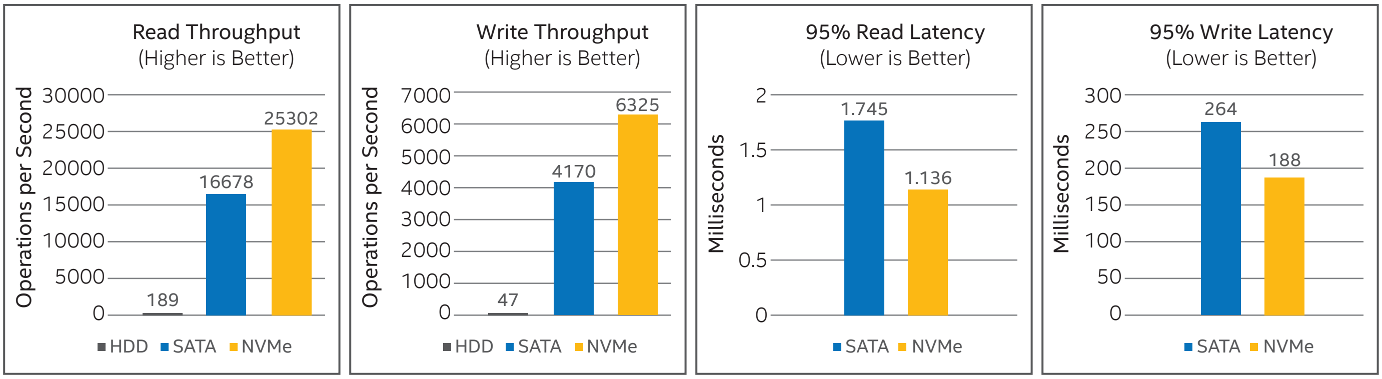

Figure 3. Key/value store testing results.

This testing shows that Couchbase Server using the ForestDB data store delivers a 50 percent increase in read/write throughput and approximately a one-third reduction in 95-percent read/write latency using SSD drives with NVMe connectivity, compared with SSD drives that use SATA.

Index Service Testing

Testing of the index service is representative of user-profile management functionality for web or mobile applications, enabling functionality such as online transactions, user preferences, and user authentication. As the number and complexity of profiles increases, relational databases often bog down under indexing tasks such as referencing users by specific profile information. The higher throughput and lower latency afforded by NoSQL scale-out architecture accelerates index creation, which supports faster queries.

Extending the challenges of the key/value test scenario, index service testing used a much larger key size, creating stresses in terms of tree size and compaction. In general, this test scenario is similar to the key/value store testing. Again, four reader threads and one writer thread were allocated, with a corresponding traffic mix of 80 percent read operations and 20 percent write operations. Important workload characteristics included the following:

- **Key size**: 1 KB
- **Number of documents**: 100 million
- **Buffer cache allocated**: 30 GB

The test results are illustrated in Figure 4.

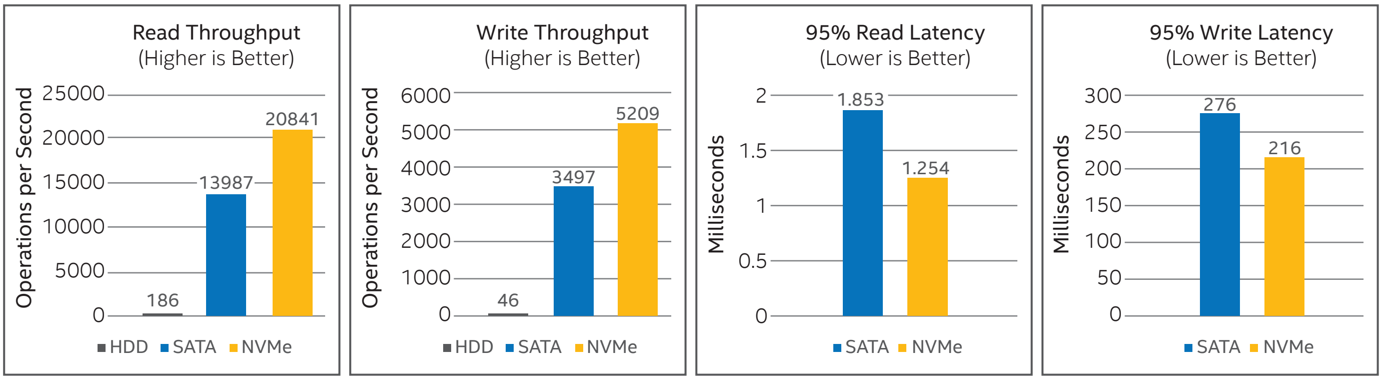

Figure 4. Index service testing results.

Similar to the key/value store tests, the index service results showed nearly a 50 percent increase in read and write throughput using NVMe drives, compared to SATA-based SSDs, with both showing order-of-magnitude advantages in comparison to HDDs. The latency results show that ForestDB handles the larger key size well, delivering a 95 percent read-latency reduction of about one third and a 95 percent write-latency reduction of approximately 22 percent using NVMe-based drives, compared with SSD drives that use SATA connectivity.

Parallel Benchmark Throughput Testing

Having established the substantial general benefit of using NVMe drives in comparison to SATA-based SSDs, the next testing stage targeted a greater level of stress on the systems. The thrust of this testing was to have four simultaneous instances of ForestDB running in parallel, with each thread running at maximum capacity and reading from/writing to its own file.

Test parameters for this testing were similar to those for the initial key/value store testing scenario, again with four reader threads and one writer thread allocated for each benchmark program. One significant change is that the buffer cache for each instance was reduced to avoid over-allocation relative to overall available system resources:

- **Average key siz**e: 48 bytes
- **Working data size**: 100 GB (100 million documents at 1 KB each)
- **Buffer cache allocated**: 10 GB

The test results are illustrated in Figure 5.

While the reduction of buffer cache allocated per instance could be expected to impact read performance, read results are nearly identical to the previous two test cases, with slightly better than a 50 percent increase using NVMe compared to NVMe-based drives. Even more compelling, write throughput increased by a factor of nearly 9x.

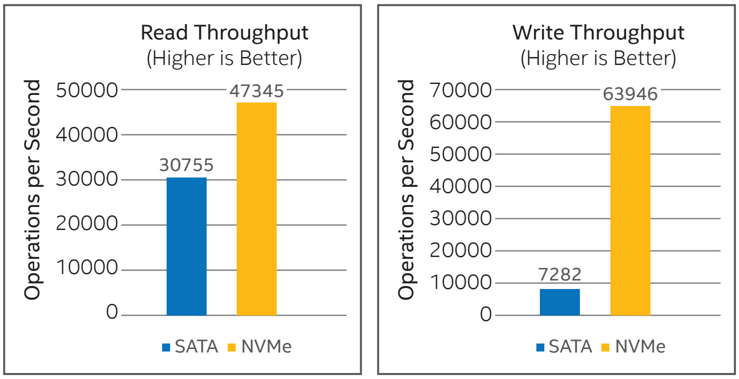

Figure 5. Parallel benchmark throughput testing results.

On the whole, these results demonstrate that NVMe provides far better parallel I/O support than SATA for ForestDB, when used as the indexing service in Couchbase Server 4.0. As a result, users can dramatically improve throughput and latency across workloads and usages by upgrading to NVMe-based local storage.

As a result, businesses can realize efficiencies such as handling more transactions per second per server, as well as performance gains such as more rapid analytics to drive more timely business decisions. Those organizations can also handle increased usage and more complex queries without bogging down as data stores continue to grow. Supporting larger workloads per server can also help lower capital requirements by reducing the number of machines that must be added to handle emerging business needs.

Ongoing work is being done by Intel and Couchbase to quantify the potential for further performance improvements using the Intel® Xeon® processor E7 product family. In particular, performance teams are investigating the performance potential of higher core counts on the processor-intensive index service. One area of interest is deploying index data and compute operations on a single scale-up host, avoiding network overhead.

Conclusion

Couchbase Server 4.0 provides the foundation for supporting big data workloads across clusters of Intel® architecture-based servers. Taking excellent advantage of the innovations in the Intel Xeon processor E5-2600 v3 product family, Couchbase Server 4.0 also provides simplicity, speed, elasticity, and reliability that make it an excellent choice for demanding applications. With innovations such as ForestDB and MDS, Couchbase Server showcases the advantages of NVMe storage, positioning businesses for successful, forward-looking implementations.

The Intel processor Xeon E5-2600 v3 product family is the foundation for the hardware that delivers next-generation results from Couchbase Server 4.0. This compute engine provides higher core counts, larger cache, faster Intel QuickPath Interconnects, and a more advanced memory subsystem than its predecessors, optimized for the needs of Couchbase Server's demanding NoSQL workloads. ForestDB shines on servers that utilize this latest Intel architecture, with a balanced platform that also includes NVMe-based Intel SSDs and Intel® Ethernet network interfaces, for an advanced solution stack of components built to bring out the best in each other.

By moving to solutions using Couchbase Server 4.0 on servers based on the Intel Xeon processor E5-2600 product family and complementary components engineered by Intel, organizations of all types and sizes can attain the performance and scalability that will help them get the most out of future opportunities.

For more information, visit intel.com/ssd and couchbase.com

[1] Testing performed by Couchbase.
Test configuration: Intel® Xeon® processor E5-2697 v3 @ 2.60 GHz, CentOS* 6.5, 65 GB RAM, 7200 RPM hard-disk drive (1 TB), Intel® Solid-State Drive Data Center S3710 (1.2 TB), Intel® Solid-State Drive Data Center P3700 (1.6 TB), ForestDB*.
ForestDB benchmark: https://github.com/couchbaselabs/ForestDB-Benchmark.

[2] Source as of August 2014 TR#3044 on STREAM (triad): Intel® Server Board S2600CP with two Intel® Xeon® processor E5-2697 v2, 24x16 GB DDR3-1866 @1066 MHz DR-RDIMM, score: 58.9 GB/sec.
New configuration: Intel® Server System R2208WTTYS with two Intel® Xeon® processor E5-2699 v3, 24x16 GB DR4-2133 @ 1600 MHz DR-RDIMM, score: 85.2 GB/sec.

Software and workloads used in performance tests may have been optimized for performance only on Intel® microprocessors.

Performance tests, such as SYSmark* and MobileMark*, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/performance.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document. You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a non-exclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

Printed in USA 0815/MW/MESH/PDF ♻ Please Recycle 332622-001US