MongoDB on Lenovo ThinkAgile VX V4 and FX V4 with VMware Cloud Foundation 9.0 and vSAN ESA

Planning / Implementation

A Modern Data Platform for Private Cloud, AI/ML, and Analytics

Today's enterprise applications increasingly rely on dynamic, fast-changing, and diverse data sources from customer profiles and product catalogs to clickstream events, IoT telemetry, application logs, and operational metrics. These workloads rarely fit neatly into rigid relational schemas. As a result, organizations are turning to document-based data models that align more naturally with the way modern applications evolve. MongoDB has emerged as a leading operational database platform for these scenarios because it enables agility, scalability, and simplicity across the full application lifecycle.

MongoDB's document data model is flexible, semi-structured or hierarchical data model to support rapid application development cycle without constant schema redesign and refactoring resulting faster time to market. MongoDB's horizontal scalability across clusters, high availability with built in replication and resilience for mission critical workload makes ideal for ingest-heavy environments such as IoT platforms, streaming analytics, and large-scale operational systems. MongoDB as a vector database combines vector search with full text search to retrieve most relevant results to power AI applications and agentic systems with retrieval augmented generation (RAG).

Lenovo ThinkAgile VX V4 and FX V4 with Intel Xeon 6 Processors

The Lenovo ThinkAgile VX650 V4 and FX650 V4 are 2-socket, 2U systems and The Lenovo ThinkAgile VX630 V4 and FX630 V4 are 2-socket, 1U systems. These hyperconverged systems features Intel Xeon 6 processors and powered by vSAN ESA and VMware Cloud Foundation and the servers support up to 86 cores, higher memory bandwidth and capacity, PCIe 5.0 I/O, and performance optimized all NVMe storage configuration to meet various workloads including databases, virtualization, VDI, analytics, AI/ML and ERP.

ThinkAgile FX offers a unique, industry first flexibility for software-defined approach to hyper convergence, leveraging the ability to move between hypervisors of your choice to deliver compute, storage and management in a tightly integrated software stack and future-proof your investment with seamless HCI software transitions.

MongoDB paired with a vSAN ESA and VMware Cloud Foundation provide cloud-like experience with on-prem cost efficiency, ensuring consolidated deployments, simplified scaling and unified management for computer, storage, container and database services. The ThinkAgile VX V4 and FX V4 with vSAN ESA enables a "Data-First" infrastructure with MongoDB which handles Transactional (OLTP), Analytics (OLAP), and Vector (AI) requirements simultaneously. Modern MongoDB versions are engineered to exploit high core processors to achieve maximum parallelism and allow insert-heavy ingest and mixed read/write activity to achieve maximum throughput.

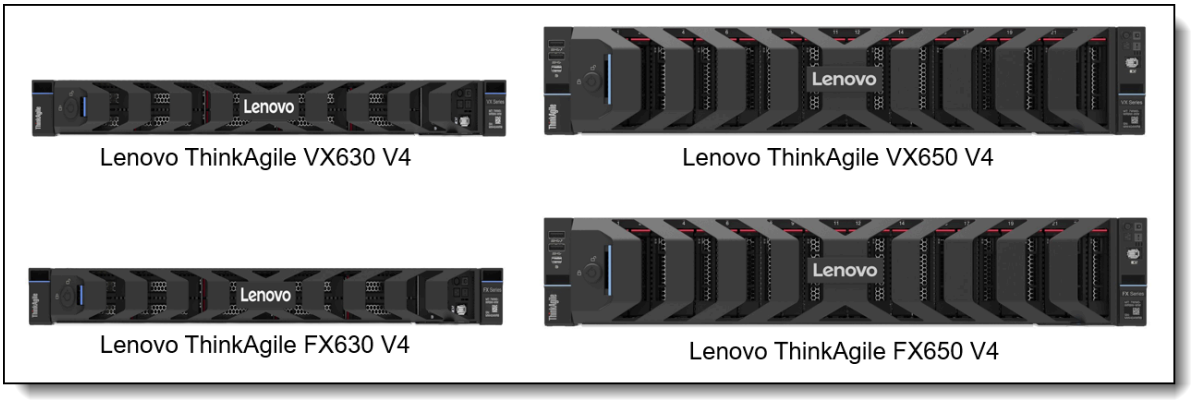

Figure 1. Lenovo ThinkAgile VX and FX V4 Systems

MongoDB Features

Transactions, Analytics and Vector database
MongoDB's combination of flexible modeling, distributed architecture, and AI-ready search and vector capabilities makes it a powerful choice for enterprises modernizing their private cloud and data platforms. MongoDB is a Document-based NoSQL database that uses BSON documents to map and store dynamic and semi-structured data which enables to build adaptable applications without rigid relational schemas.

MongoDB architecture consists of:

- **mongod** – primary database engine responsible for storage and querying
- **mongos** – a query router for sharded clusters, routing client requests to appropriate shards
- **Config servers (CSRS)** – store cluster metadata and sharding configuration

Replica Sets ensure high availability and fault tolerance by maintaining multiple copies of data across nodes. Primary node handles writes; secondary nodes replicate via the oplog and supports globally distributed deployments.

Sharding enables horizontal scaling across many machines for large datasets and high-throughput workloads by using shards (Replica set), partitioned data subsets.
MongoDB's aggregation framework supports Time-series optimizations, complex transformations and real time analytics.

MongoDB Ops Manager: The "Command Center" for enterprise deployments. It automates backups, point-in-time recovery, and rolling upgrades.

MongoDB supports AI-native capabilities bringing full-text, semantic, and vector search to self managed and on prem deployment to build RAG (Retrieval Augmented Generation) systems directly where your data lives and eliminate external vector DBs and search engines, reducing complexity and latency.

MongoDB provides 8.2 provide significant performance improvement for unindexed queries and time series bulk inserts. Queryable Encryption enhances secure operations without compromising query expressiveness.
VMware Aria Operations for MongoDB provides monitoring dashboards and metrics across performance, capacity and KPIs.

YCSB overview
Yahoo Cloud Serving Benchmark (YCSB) is a benchmarking framework that is used to evaluate NoSQL and cloud-serving databases. It is designed to simulate application-access patterns and allows for a repeatable means of measuring latency and the throughput of a database under various load conditions.

There are two phases for the YCSB benchmarks:

- **Load** - the dataset used during testing is loaded into the database.
- **Run** - transactional workload (either read, update, insert or scan) defined by the mix of transactions is executed against the database. This makes it possible to test both the pure ingest of data as well as operational workloads randomly using a mixture of transaction types.

In this study, YCSB was used to evaluate MongoDB using only insert workloads and mixed read-write workloads while varying the size of the documents inserted into the database. By changing the size of the documents being added to the database and keeping the client-side concurrency the same, the benchmark demonstrates how the architecture of the database responds to different payload sizes and how throughput and latency react to different patterns of workloads.

Configurations

Hardware Configuration

The testing was performed with 4 node Lenovo ThinkAgile VX650 V4 cluster equipped with two Intel Xeon 6747P 48C processors with Hyper-Threading enabled. This provided a high-density compute platform for virtualized database workloads.

Table 1. Lenovo ThinkAgile VX650 V4 configuration

Item	Description
Server platform	Lenovo ThinkAgile VX650 V4
Number of hosts	4
Processor per host	2 × Intel Xeon 6747P 48C 330W 2.7GHz Processor
Hyper-Threading	Enabled
Memory	1.5TB (16 x ThinkSystem 96GB TruDDR5 6400MHz (2Rx4) RDIMM)
Storage	• 2 x ThinkSystem M.2 7450 PRO 480GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD • 8 x ThinkSystem 2.5" U.2 PM9D3a 1.92TB Read Intensive NVMe PCIe 5.0 x4 HS SSD
Network	2 x ThinkSystem Mellanox ConnectX-6 Lx 10/25GbE SFP28 2-port OCP Ethernet Adapter
VMware Software	VMware ESXi 9.0.0.0.24755229, VMware Cloud Foundation 9.0

MongoDB configuration

MongoDB was configured to use the WiredTiger storage engine with a 64 GB WiredTiger cache. To improve storage throughput, the MongoDB data volume was built using eight VMDKs attached to the virtual machine. These virtual disks were combined in the guest operating system using LVM and formatted with XFS.

Table 2. MongoDB configuration

Item	Description
Database	MongoDB 7.0.3
OS	Ubuntu 22.04.5 LTS
VM Configuration	128 VCPU, 256 GM Memory, 2 TB disk, NUMA Enabled
Storage engine	WiredTiger
WiredTiger cache size	64 GB
Data path	/mongodata
Port	27017
MongoDB data filesystem	XFS
MongoDB logical volume	/dev/mapper/mongovg-mongolv
Mount point	/mongodata
Number of MongoDB data VMDKs	8
Backing virtual disks	sdc, sdd, sde, sdf, sdg, sdh, sdi, sdj
Volume management method	LVM

Testing

Testing Methodology

The goal of testing was to evaluate how changing document sizes, and the number of concurrent clients would affect throughput and latency.
For this testing, the YCSB dataset contained a record count of 10,000,000, operation count of 10,000,000 and a field count of 1. The document sizes were chosen to represent three practical application profiles, respectively, rather than a single synthetic point.

- **256 bytes** represents small records, such as lightweight metadata, compact session objects, short key-value style application data, or simple event entries.
- **2000 bytes** represents a medium-size operational document, which is a realistic size for many business applications using MongoDB for customer, catalog, profile, or transactional data.
- **8192 bytes** represents a larger document profile, useful for understanding how the platform behaves when records contain richer payloads, embedded structures, or more application context.

The testing performed with 64, 96, and 128 concurrent threads to stress the database and observe not only the effects of document size, but also how well the platform performed as client concurrency is increased.

The following workload profiles tested, as follows:

Table 3. Workload profiles

Insert	**100% write**
50/50 READ/WRITE	Reasonably balanced mix of transactional activity
80/20 READ/WRITE	Read intensive workloads

The data was loaded into the database at the selected document sizes at each client thread level for insert testing. The workload data was loaded first into the database and then tested with concurrent threads for the mixed workloads.

Test Results

Below are the YCSB results for MongoDB from the Lenovo ThinkAgile VX650 V4 running on VMware vSAN 9.0. The data included insert only operations as well as two mixed load profiles using a document size of 256 bytes, 2000 bytes, or 8192 bytes and 64, 96 or 128 concurrent client threads respectively.

Throughput Performance

The following table summarizes results for the three workload profiles tested. This dataset gives a very clear look into how MongoDB scales under vSAN ESA with increasing concurrency and IO sizes.

Throughput increases from 64-96 threads across most cases and throughput drops sharply from 96-128 threads for nearly all combinations.

Small documents (256B) and medium documents(2000B) achieved highest throughput and scaling overall.

Large documents (8192B) resulted significantly lower ops/sec at all the thread counts. The throughput performance can be mapped to vSAN ESA which chunk I/O for write heavy or mixed workloads with large documents which also increases network traffic between the nodes.

The read heavy workloads achieved higher threads for 96 threads and medium size document, and it can be mapped to MongoDB WiredTiger cache and fewer writes during the test.

It is recommended to evaluate different NVMe drive configurations to get optimal performance by exploiting vSAN ESA features NVMe-optimized data paths, Log-structured writes and Parallel IO processing.

Table 4. Throughput results

Document size	Threads #	Load (ops/sec)	50 / 50 Read/Write (ops/sec)	80 / 20 Read/Write (ops/sec)
256 B	64	87,898	80,459	86,314
256 B	96	100,196	101,581	93,538
256 B	128	48,724	43,352	49,440
2000 B	64	85,832	52,041	85,337
2000 B	96	108,193	48,361	117,729
2000 B	128	52,860	53,939	52,062
8192 B	64	61,646	19,261	45,341
8192 B	96	68,867	21,130	47,478
8192 B	128	49,862	20,811	52,006

Latency Performance

The latency increases as threads increase (especially 96 → 128) and it correlates to the drop in throughput. The larger the document size, higher latency especially for write operations. The larger documents possess more overhead for MongoDB (journaling, cache, compression) and vSAN ESA (write amplification, parity and distribution across nodes). Mixed and write heavy workloads throughput saturates quickly and more drastic increase in latency. The mixed workload 50/50 throughput is lower and it could be mapped to read and write path contention. The MongoDB lock contention increases at high concurrency and creates CPU overhead resulting in increase in latency.

The following table shows the latency results.

Table 5. Latency results

Document size	Threads #	Load (ms)	50 / 50 Read/Write (ms)	50 / 50 Read/Write (ms)	80 / 20 Read/Write (ms)	80 / 20 Read/Write (ms)
256 B	64	0.716	0.733	0.834	0.699	0.862
256 B	96	0.941	0.873	0.991	0.981	1.151
256 B	128	2.609	2.819	3.036	2.512	2.785
2000 B	64	0.729	0.919	1.517	0.703	0.892
2000 B	96	0.865	1.240	2.687	0.766	0.928
2000 B	128	2.401	2.217	2.489	2.378	2.679
8192 B	64	1.026	1.847	4.766	1.132	2.454
8192 B	96	1.366	2.215	6.828	1.432	4.307
8192 B	128	2.541	4.125	8.119	2.354	2.801

Conclusion

The results show that Lenovo ThinkAgile VX V4 systems running VMware Cloud Foundation 9.0 with vSphere 9.0 and vSAN ESA provide an ideal platform for MongoDB in a virtualized environment. Overall, the performance aligns with expectations on vSAN ESA, which excels in read-heavy and moderate-size IO but sees increased latency under large, mixed read/write workloads. The latency trends mirror throughput results, confirming that write amplification and mixed IO patterns drive higher latency, especially with large documents. The solution demonstrated strong support for MongoDB workloads across insert-only and mixed access patterns, with the most favorable results occurring at the medium document size and midrange concurrency level.

Performance testing and results

The testing was performed under the following conditions:

- The Azure Local environment hosted two SQL Server virtual machines per node, resulting in a total of four SQL Server 2025 VMs across the cluster.
- Each virtual machine ran SQL Server 2025 and hosted a TPC-C database configured with 800 warehouses.
- Each VM was allocated with 16 vCPUs and 128 GB RAM to ensure consistent resource availability during testing.
- HammerDB was used as the workload generator, specifically running the TPC-C benchmark to simulate an OLTP workload

Across the four SQL Server 2025 virtual machines, the test configuration delivered a combined throughput of 7,715,373 TPM and 1,676,634 NOPM. These values were obtained on a instance built with Intel Xeon 6505P processors, a mid-range part with 12 performance cores per socket.

The results show that the platform maintains consistent OLTP throughput even when core counts are lower. This is relevant for customers who prioritize predictable latency and stable performance rather than maximum core density.

The Xeon 6505P operates at a lower TDP compared to higher-end models in the same generation. Because of this, the system shows favorable performance-per-watt behaviour under OLTP load.

SQL Server 2025 benefits from improved intelligent query processing and metadata engine efficiencies, which help sustain throughput without requiring high-frequency turbo states for long periods. Lower power draw reduces thermal load, keeps fan speeds stable and contributes to overall efficiency.

About the benchmarks:

- HammerDB was used as the workload generation tool for performance validation. HammerDB is an open-source benchmarking framework that implements industry-standard workloads and can automate large-scale transactional testing. More information about HammerDB is available at: https://www.hammerdb.com
- The TPC-C workload used in this test is defined by the Transaction Processing Performance Council and represents a classic OLTP profile composed of order entry, payment, delivery and stock-level transactions. Full details on the TPC-C benchmark specification can be found at: https://www.tpc.org

Bill of Materials

The following table lists the feature codes for the lab configuration.

Table 6. ThinkAgile VX650 V4

Feature code	Description	Quantity
7DG6CTO1WW	Server: ThinkAgile VX650 V4	1
C68E	ThinkAgile VX650 V4 24x2.5" Chassis	1
BVGL	Data Center Environment 30 Degree Celsius / 86 Degree Fahrenheit	1
B0W3	XClarity Pro	1
C5QX	Intel Xeon 6737P 32C 270W 2.9GHz Processor	2
BYTJ	ThinkSystem 32GB TruDDR5 6400MHz (2Rx8) RDIMM	16
BT2G	vSAN ESA	1
BYRM	vSAN-HCI-SM	1
C2BS	ThinkSystem 2.5" U.3 7500 PRO 3.84TB Read Intensive NVMe PCIe 4.0 x4 HS SSD	3
C46P	ThinkSystem 2U V4 8x2.5" NVMe Backplane	1
C0JJ	ThinkSystem M.2 RAID B540p-2HS SATA/NVMe Adapter	1
BQ1Y	ThinkSystem M.2 5400 PRO 480GB Read Intensive SATA 6Gb NHS SSD	2
C1YK	ThinkSystem SR650 V4/SR630 V4 x16 OCP Cable Kit	1
BE4T	ThinkSystem Mellanox ConnectX-6 Lx 10/25GbE SFP28 2-port OCP Ethernet Adapter	1
C4S2	ThinkSystem SR650 V4 Processor Board	1
AURS	Lenovo ThinkSystem Memory Dummy	16
BPP5	OCP3.0 Filler with screw	1
C7Y8	ThinkSystem SR650 V4 System I/O Board	1
B8K8	ThinkSystem 2U MS 24x2.5" NVMe HDD Type Label1	1
AUTQ	ThinkSystem small Lenovo Label for 24x2.5"/12x3.5"/10x2.5"	1
BZ7F	ThinkSystem WW Lenovo LPK, Birch Stream	1
5641PX3	XClarity Pro, Per Endpoint w/3 Yr SW S&S	1
1340	Lenovo XClarity Pro, Per Managed Endpoint w/3 Yr SW S&S	1

Author

Cristian Ghetau is an Advisory Engineer for Lenovo in Romania and has experience in Cloud Infrastructure technologies. He has had more than 13 years of experience working with virtual environments from VMware, Microsoft, Oracle, Linux.

Related product families

Product families related to this document are the following:

- ThinkAgile FX Series
- ThinkAgile VX Series for VMware
- VMware vSphere

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service. Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
8001 Development Drive
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION ”AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary. Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk. Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This document, LP2411, was created or updated on March 30, 2026.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2411
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2411.

Trademarks

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. A current list of Lenovo trademarks is available on the Web at https://www.lenovo.com/us/en/legal/copytrade/.

The following terms are trademarks of Lenovo in the United States, other countries, or both:
Lenovo®
ThinkAgile®
ThinkSystem®
XClarity®

The following terms are trademarks of other companies:

Intel®, the Intel logo and Xeon® are trademarks of Intel Corporation or its subsidiaries.

Azure® and SQL Server® are trademarks of Microsoft Corporation in the United States, other countries, or both.

TPC® is a trademark of Transaction Processing Performance Council.

Other company, product, or service names may be trademarks or service marks of others.