A Principled Technologies report: Hands-on testing. Real-world results.

Ingesting data for use with a large language model for AI: Latest-generation Dell PowerEdge servers powered by 5th Generation AMD EPYC processors offer a range of strong options

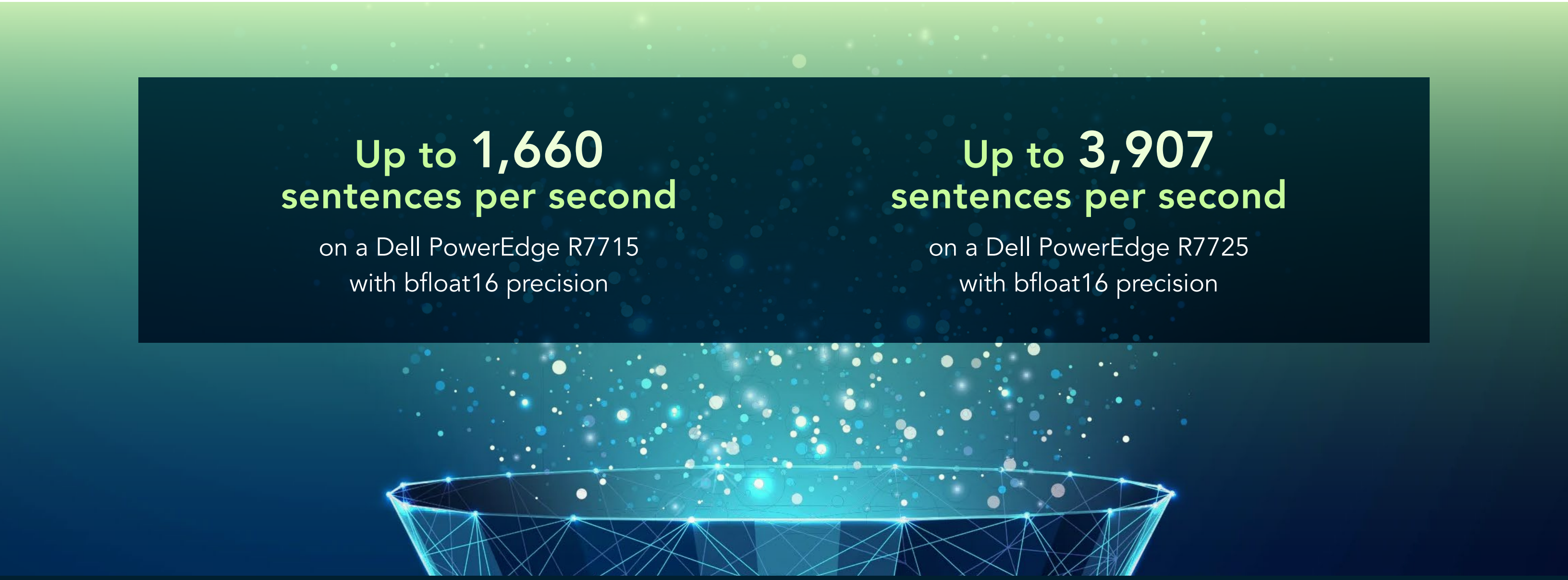
We measured the performance of multiple disaggregated infrastructure server configurations to help decision-makers choose the right one for their needs

Organizations across industries are rapidly adopting internal AI platforms to boost employee productivity, support customers more cost-effectively, and remain competitive while keeping their proprietary data secure. Getting started with these applications typically involves the ingestion of a great deal of information into a vector-searchable database that a large language model (LLM) will use to answer questions. Because this work is by nature resource-intensive, it's essential to select gear that is up to the task. At the same time, no one wants to spend more than is necessary to achieve their company's goals.

The different components within a server can have an enormous impact on how efficiently it can execute a complex task such as data ingestion. Understanding these factors is critical to selecting a server solution that hits the sweet spot of handling your demands while avoiding expensive overprovisioning.

Latest-generation Dell™ PowerEdge™ servers, which offer the flexibility of a disaggregated infrastructure, are an excellent choice for AI ingestion. But how do you determine which model and which processor can provide the right amount of power for your specific needs?

To help answer this question, we conducted a series of tests of the ingestion capabilities of two latest-generation Dell servers—the PowerEdge R7715 and the PowerEdge R7725—with a variety of different AMD EPYC™ processors. Our findings will help you determine which option will deliver the "just right" capabilities for you.



Up to 1,660 sentences per second

on a Dell PowerEdge R7715 with bfloat16 precision

Up to 3,907 sentences per second

on a Dell PowerEdge R7725 with bfloat16 precision

For AI ingestion, power and flexibility are paramount

Internal AI applications typically use retrieval-augmented generation (RAG), a process where LLMs refine their ability to answer user questions using an internal body of information. The critical first stage of establishing any such system is ingesting your organization's proprietary data into a vector-searchable database that the LLM will access.

A server that can perform this task quickly enables your AI application to deliver value sooner, and also expedites the process when the time comes to add more of your company's internal information. A server that can perform this task quickly enables your AI application to deliver value sooner, and also expedites the process when the time comes to add more of your company's internal information. A popular way to quantify this speed is using a metric of sentences per second—the number of text inputs an LLM can convert into embeddings on a given system. For instance, a data set of 100 product descriptions ingested at 10 sentences per second, would finish ingestion in 10 seconds. The processors in a server are a critical factor in its ingestion capabilities and other server specifications, such as memory bandwidth and cache size, also play a role.

The type of infrastructure is another important consideration. In a hyperconverged infrastructure (HCI), where multiple workloads run simultaneously on a server, the particularly intensive work of AI ingestion can slow down—and be slowed down by—other workloads. A more efficient approach is using a disaggregated infrastructure, where a server temporarily dedicated to ingestion can enable an organization to quickly finish the job while other critical business operations run on other servers.

When you select a server for AI workloads, flexibility matters in addition to performance. Committing to HCI can risk vendor lock-in, limiting future choices, and increasing spend on licensing. But with a disaggregated infrastructure, companies pay for exactly what they need, scaling compute and storage independently and customizing resources to boost server efficiency and utilization.

The role of embedding in AI ingestion

Embeddings capture semantic meaning and relationships within data—such as text, images, or audio—using numerical representations of data. Embeddings represent data using long lists of numbers called vectors. Our testing focused on the embedding phase of ingesting data into vector-searchable databases for RAG because it can be the most time-consuming.

Our testing approach

The goal of our testing was to glean data that companies can use to determine the optimal configurations and settings for their unique AI ingestion workload requirements. To this end, we performed a series of tests on the following five server configurations:

Dell PowerEdge R7715

Configuration 1

- One 5th Generation AMD EPYC 9575F processor
- Total core count: 64
- Max boost clock: 5 GHz
- All-core boost speed: 4.5 GHz
- Base clock: 3.3 GHz
- Total L3 cache: 256 MB
- System mem BW: 500 GB/s

Configuration 2

- One 5th Generation AMD EPYC 9745 processor
- Total core count: 128
- Max boost clock: 3.7 GHz
- All-core boost speed: 3.45 GHz
- Base clock: 2.4 GHz
- Total L3 cache: 256 MB
- System mem BW: 500 GB/s

Configuration 3

- One 5th Generation AMD EPYC 9845 processor
- Total core count: 160
- Max boost clock: 3.7 GHz
- All-core boost speed: 3.25 GHz
- Base clock: 2.1 GHz
- Total L3 cache: 320 MB
- System mem BW: 500 GB/s

Dell PowerEdge R7725

Configuration 1

- Two 5th Generation AMD EPYC 9575F processors
- Total core count: 128
- Max boost clock: 5 GHz
- All-core boost speed: 4.5 GHz
- Base clock: 3.3 GHz
- Total L3 cache: 512 MB
- System mem BW: 1,228 GB/s

Configuration 2

- Two 5th Generation AMD EPYC 9965 processors
- Total core count: 384
- Max boost clock: 3.7 GHz
- All-core boost speed: 3.35 GHz
- Base clock: 2.25 GHz
- Total L3 cache: 768 MB
- System mem BW: 1,228 GB/s

We tested at two precision levels, **float32** and **bfloat16**, and performed tuning to identify the best settings to use for comparison across configurations. We used the msmarco-distilbert-base-v4 Sentence Transformer model and collected a range of metrics, including time to complete an ingestion task and utilization of resources, such as CPU and memory, to determine best performance.

Test findings and their implications for companies preparing internal data for vector-searchable databases

Our results show that whether organizations select bfloat16 or float32 precision, configurations with a higher processor core count achieved a higher rate of sentences per second. The dual-socket PowerEdge R7725 configurations also ingested at faster rates than the single-socket PowerEdge R7715. For teams choosing solutions for AI ingestion, opting for a greater number of cores or processors directly improves performance. Especially in use cases with large datasets, these faster rates could significantly speed gaining access to your data, enabling you to use your LLM sooner.

Note: The graphs in this report use different scales to keep a consistent size. Please be mindful of each graph's data range as you compare.

Dell PowerEdge R7715

Figure 1 shows the number of sentences per second ingested by the three configurations of the Dell PowerEdge R7715 with float32 precision.

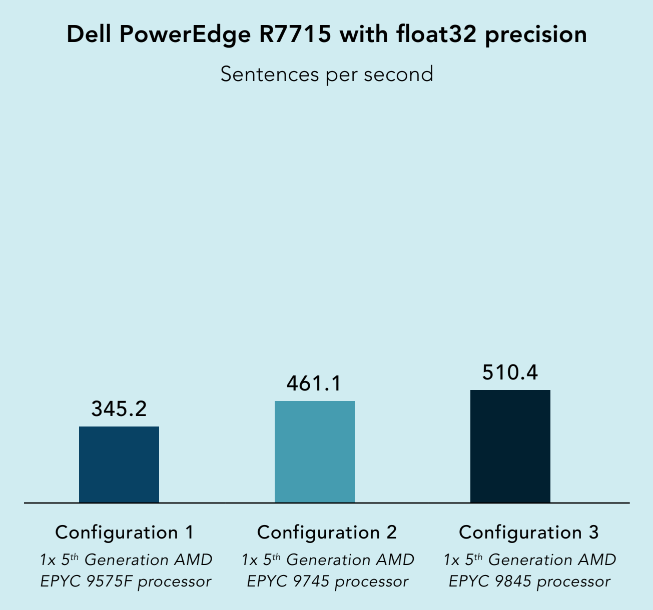


Figure 1: Dell PowerEdge R7715 float32 sentences per second. Higher is better. Source: PT.

Figure 2 shows the number of sentences per second ingested by the three configurations of the Dell PowerEdge R7715 with bfloat16 precision.

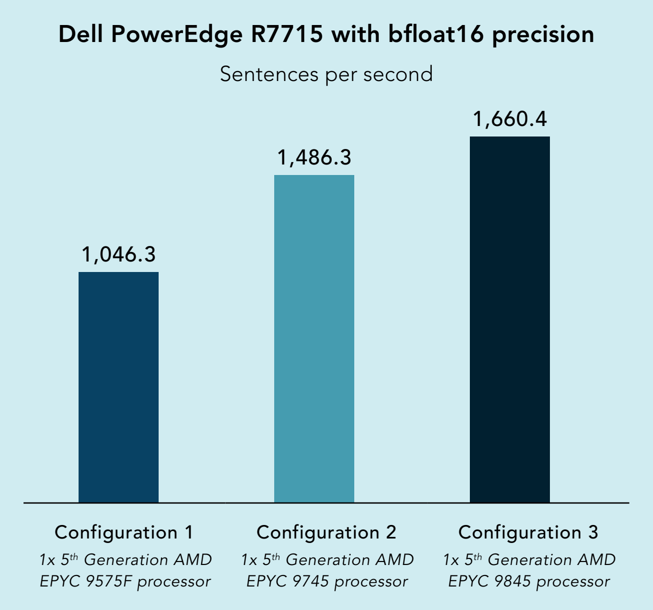


Figure 2: Dell PowerEdge R7715 bfloat16 sentences per second. Higher is better. Source: PT.

About the latest-generation Dell PowerEdge servers we tested

The Dell PowerEdge R7715 is a single-socket, 2U rack server powered by a single new 5th Generation AMD EPYC processor. The PowerEdge R7715 supports:

- Up to 24 DDR5 DIMM slots for 6 TB max memory
- Up to 8 PCIe® Gen5 slots
- Smart Cooling configuration that enhances cooling to allow most configurations to be air-cooled[1]

▸ Learn more

The Dell PowerEdge R7725 is a dual-socket, 2U rack server powered by two new 5th Generation AMD EPYC processors. It offers the same memory and PCIe capacity as the R7715 and uses the same Smart Cooling design. According to Dell, the server delivers "breakthrough performance that scales for traditional and emerging workloads, including big data analytics, AI/ML and high-performance compute (HPC), using the latest performance and density with optional acceleration."[2]

▸ Learn more

Dell PowerEdge R7725

Figure 3 shows the number of sentences per second ingested by the two configurations of the Dell PowerEdge R7725 with float32 precision.

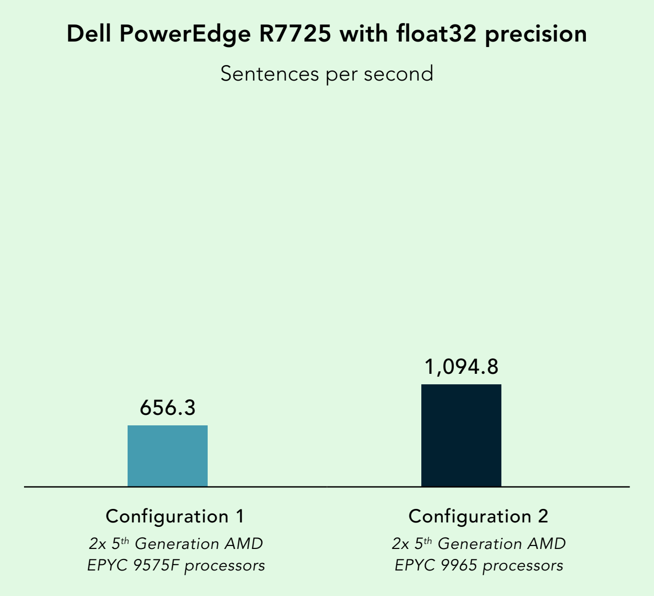


Figure 3: Dell PowerEdge R7725 float32 sentences per second. Higher is better. Source: PT.

Figure 4 shows the number of sentences per second ingested by the two configurations of the Dell PowerEdge R7725 with bfloat16 precision.

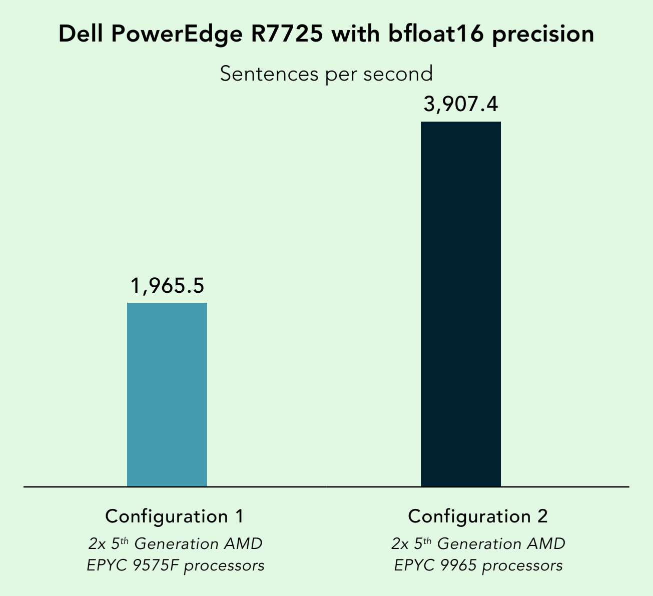


Figure 4: Dell PowerEdge R7725 bfloat16 sentences per second. Higher is better. Source: PT.

Analysis: How hardware configurations affected performance

Based on our tests, results, and tuning, we found that several hardware factors affected the throughput we saw:

- The number of **processor cores** had the greatest impact on performance. In our tests, we saw anywhere from 47 percent to 89 percent scaling for every CPU core the configuration added. Having more processor cores sometimes increases licensing requirements, which can become cost prohibitive. With this particular workload being built on open-source software, purchasing higher-core processors can be much more affordable.
- **Processor cache** was the next most important factor, including both L2 and L3 caches.
    - The **L2 cache** in each processor was 1 MB per core, meaning that the L2 cache scaled with the number of cores. This is another reason the number of cores had the greatest impact on performance.
    - **L3 cache** also played an important role. The configurations we tested used processors with L3 caches with anywhere from 256 to 384 MB in total capacity and eight to 12 packages per processor. (See the science behind the report for more information.) While we didn't optimize our tests for latency, our early tuning experiments showed that latency especially depended on optimizing process affinity with L3 locality.
- Higher **CPU frequencies** had a small impact on performance, especially the all-core turbo frequency. Greater maximum boost frequency for single cores also helped speed short single-threaded tasks, such as parsing the data and converting it to NumPy format.
- Increased **memory bandwidth** also had a small impact on performance, though much less than we typically see on other machine learning (ML) workloads, such as running an LLM on CPU. This is because the sentence-transformers model we used is relatively small, running mostly from CPU cache rather than system memory.

About 5th Generation AMD EPYC Processors

The Dell PowerEdge servers we tested were all equipped with 5th Generation AMD EPYC processors from the 9005 family, specifically, the 9575F, 9745, 9845, and 9965. According to AMD, the 5th Generation family of AMD EPYC 9005 processors is "The Leading CPU for AI," and is designed to maximize per-server performance while delivering leadership in AI inference performance.[3]

AMD states that this family of processors can deliver the same integer performance of older processors while "dramatically reducing physical footprint, power consumption, and the number of software licenses needed – freeing up space for new or expanded AI workloads."[4]

▶ Learn more

Conclusion

In our testing, latest-generation Dell PowerEdge R7725 and R7715 servers powered by 5th Generation AMD EPYC processors demonstrated strong performance for ingesting information into vector-searchable databases for use by LLMs in AI applications. Configurations leveraging bfloat16 precision significantly boosted sentence processing rates, with the dual-socket PowerEdge R7725 models delivering up to 3,907 sentences per second, highlighting their suitability for demanding AI applications. A disaggregated architecture using these servers allows organizations to independently scale compute and storage resources, optimizing infrastructure efficiency and cost-effectiveness. By carefully selecting the appropriate server model and processor configuration based on workload needs, companies can achieve a balanced solution that accelerates AI ingestion while avoiding unnecessary overprovisioning, enabling faster deployment and expansion of internal AI platforms.

[1] Dell, "New PowerEdge R7715 Rack Server," accessed July 17, 2025, https://www.dell.com/en-us/shop/dell-poweredge-servers/new-poweredge-r7715-rack-server/spd/poweredge-r7715/.

[2] Dell, "PowerEdge R7725 Specification Sheet," accessed July 17, 2025, https://www.delltechnologies.com/asset/en-us/products/servers/technical-support/poweredge-r7725-spec-sheet.pdf.

[3] AMD, "5th Generation AMD EPYC™ Processors," accessed July 23, 2025, https://www.amd.com/en/products/processors/server/epyc/9005-series.html.

[4] AMD, "5th Generation AMD EPYC™ Processors."

Read the science behind this report ▶

Principled Technologies®

Facts matter.®

For additional information, review the science behind this report.

This project was commissioned by Dell Technologies.