November 2026

Put agentic AI to work on familiar, general purpose servers

A Dell PowerEdge R770 with Intel Xeon 6 processors completed 110 financial-services agentic AI tasks per minute through the Intel OpenVINO toolkit with no GPU

Executive summary

For many organizations that want to deploy agentic AI, an early decision is the platform on which to run it. Specialized accelerator platforms can carry long lead times, significant power and cooling demands, and a learning curve for IT admins. For a team that wants to move a model into production quickly, those obstacles can compound.

The Dell™ PowerEdge™ R770 offers a compelling alternative. It is a two-socket, 2U rack server built on Intel® Xeon® 6 processors, the same class of server that already runs databases, virtual desktops, and line-of-business applications in many data centers. We used OpenVINO Model Server (OVMS), powered by the OpenVINO toolkit. OVMS serves AI/ML models, optimizing inference execution on Intel platforms your staff likely already knows, enabling you to run agentic AI work, including the language model itself.

Working with Dell, we selected a financial-services agentic AI workload, deployed it on a PowerEdge R770 running OVMS, and measured two things that determine whether a server is up to the task: how much work it completes per minute inside a realistic service target and how long each agent task takes.

In our tests, the PowerEdge R770 completed 110 agentic AI tasks per minute (about 6,650 per hour) on its two Intel Xeon 6767P processors and kept the 95th-percentile task time at 28.06 seconds, inside our 30-second service level.

Dell™ PowerEdge™ R770

- Intel® Xeon® 6 processors: two Intel Xeon 6767P, 64 cores per socket
- Memory: 1 TB of DDR5-6400 (16 x 64GB)
- OpenVINO™ Model Server 2026.4.1 (Intel® Distribution of OpenVINO toolkit)
- Model: Qwen3-8B at INT4 weight precision (OpenVINO/Qwen3-8B-int4-ov)



Product images provided by Dell Technologies

Serve more work per server
110 agent tasks per minute on a financial-services workload, no GPU required

Hit SLA targets
110 tasks/minute rate with 30 concurrent tasks under 30-second target

Key takeaways

from this Principled Technologies report

"A server that posts a large aggregate number by queuing work deeply will feel slow to everyone waiting in that queue."



1 A mainstream CPU-powered server can carry agentic AI work.

On our financial-services agentic AI workload, it sustained 110 agentic AI tasks per minute with a 95th-percentile task time of 28.06 seconds, proving that with some workloads, you can run your agentic AI with the hardware refresh you might already have been planning.

2 OpenVINO puts Intel AMX to work.

OVMS runs models through the OpenVINO toolkit, which uses the features it finds on the processor. On the Intel Xeon 6767P, those features included Intel AMX. For this study, we used a version of Qwen3-8B that Intel publishes ready for OpenVINO, with weights compressed to INT4, which brings the model to 4.6 GB.

3 Throughput and response time trade off against each other.

Up to a point, adding work raises total output. Past that point, extra work mostly waits in line. On our configuration, throughput held between 105.3 and 111.7 tasks per minute from 28 to 34 concurrent tasks, while 95th-percentile task time climbed from 26.57 to 32.04 seconds. At 31 concurrent tasks, task time crossed our 30-second target.

4 Usable capacity depends on your latency target.

We raised concurrency until 95th-percentile task time passed our 30-second target. The highest level under the target was 30 concurrent tasks. Past that point, task time rose while throughput did not.

Why you can use a general-purpose server as your first AI server

Media coverage of enterprise AI often focuses on large implementations: racks of accelerators, liquid cooling loops, and power budgets measured in tens or even hundreds of kilowatts. That picture suits organizations doing intense training of models or serving thousands of users, but doesn't make sense for smaller organizations that simply may want to run a document classifier, a quality-inspection model, a forecasting model, or a modest language model against their own data.

Running inference, rather than training models, is what those organizations need, and it is where a large portion of AI spending is heading. Gartner projects that inference will account for 55 percent of spending on AI-optimized infrastructure as a service in 2026, passing training for the first time, and will exceed 65 percent by 2029.[1] Those figures cover cloud infrastructure rather than on-premises systems, but the direction is the same in both places: The dominant AI workload is no longer building models, but running them.

Inference workloads rely on pre-trained models, and prompts can vary from steady streams to unpredictable bursts. Optimizing throughput and latency requires careful selection of hardware, model, memory bandwidth, and serving engine.

Agentic AI adds another consideration. An AI agent does much more than call a model: it queries databases, runs analysis code, searches documents, and checks its results, and typically all of that work runs on CPUs. In our financial workload, most agent tasks made between zero and 30 tool calls, while transaction reconciliation tasks, which check a batch of 200 accounts each, averaged about 220. Every task made at least one model call, and fraud alert triage tasks, which work in a loop, averaged about six.

For an organization putting its first AI workload into production, that CPU-heavy profile makes a general-purpose server a practical starting point. It can run the model, the agents, and the data they work with on the same processors, without the power, cooling, and skills demands that specialized accelerator platforms can bring. And if demand grows, a server with open expansion slots can take accelerators later.



What we set out to learn

We set out to quantify exactly what organizations can expect by running their agentic AI workloads on a Dell PowerEdge R770 with OVMS.

Working with Dell, we chose a financial-services agentic AI workload because it exercises both halves of an agentic system: the language model and the data work around it, such as database queries, analysis code, and document search. We ran it on Qwen3-8B, an open-weight large language model (LLM), served through OpenVINO Model Server 2026.4.1, at INT4 weight precision.

We collected data across five concurrency levels (28, 29, 30, 31, and 34 concurrent tasks), ran every level three times, and report the run with the median tasks per minute. We generated the testing load in a closed loop: At each concurrency level, we ran a fixed number of tasks continuously, and as tasks finished, new tasks were started.

How we measured performance

Quantifying an agentic AI deployment involves three metrics, each of which answers a different business question.

- **Throughput**, measured in agent tasks completed per minute, indicates the volume of work the server finishes in a given period. It governs how many reconciliations, valuations, or reports you can finish in a certain amount of time, and it sets the ceiling on what one server can do.
- **Latency**, measured in seconds per agent task, indicates how long a person or an upstream process waits. For interactive work, this is the number users feel. We report the 95th percentile: the time within which 95 percent of tasks finished.
- **Concurrency** indicates how many agentic AI tasks the server supports simultaneously while holding latency under an agreed threshold.

We set our latency target at 30 seconds, which reflects financial work in which a person is likely waiting on the result, such as an analyst checking a filing. We then raised concurrency step by step and recorded throughput at the highest level at which the 95th-percentile task time stayed under the target. We also recorded output tokens per second.

Each test held a fixed number of tasks in flight (the concurrency level) for a 420-second measurement window,

About the Dell PowerEdge R770

The PowerEdge R770 is a two-socket, 2U rack server built for Intel Xeon 6 processors with P-cores or E-cores, supporting up to 86 P-cores or 144 E-cores per socket.[2]

It carries 32 DDR5 DIMM slots for as much as 8 TB of memory at speeds up to 6400 MT/s, and it takes either air cooling or direct liquid cooling.

For acceleration, the chassis holds up to six 75 W single-width cards or up to two 450 W double-width cards, which gives a buyer room to add GPUs later without changing platforms.

Storage scales to 40 front-mounted E3.S Gen5 NVMe drives, and the server manages through iDRAC and the Dell OpenManage portfolio like the rest of the PowerEdge line.

We configured our system with two 64-core Intel Xeon 6767P processors, 1 TB of DDR5-6400 memory (16 x 64GB RDIMMs), a Dell BOSS-N1 boot module with two M.2 NVMe drives, and four 1.92TB Dell DC NVMe CD8 U.2 drives in RAID 10 for data, running Ubuntu 26.04.1 LTS. Learn more about the Dell PowerEdge R770.

starting a new task the moment one finished. We discarded a warm-up period before each window (600 seconds before the first test), so the model server's prompt cache was warm. When the window closed, we kept the load on until every task started in the window had finished, so slow tasks were timed under full load. We ran each level three times and report the run with the median tasks per minute. We then repeated the entire set of 15 runs twice more as separate sweeps, and all three sweeps agreed within about 2 percent on throughput and 95th-percentile task time at every level.

The workload we ran

We used a PT-developed agentic AI benchmark running its financial-services workload, which simulates the back-office work of a mid-sized financial firm. Each AI agent, built with the open-source LangGraph framework, takes one piece of work; gathers the data it needs from a PostgreSQL database, PDF filings, or a library of public SEC filings; runs analysis code; and calls the LLM to generate the response, which is a narrative, a list of findings, or a decision. The benchmark then grades each answer automatically against a known answer key.

The workload mixes eight task types (Figure 1):

- Regulatory ratios, where an agent computes compliance ratios for 200 institutions and reports which ratios breach their limits
- Portfolio valuation, where an agent prices a portfolio at market and reports its net asset value and largest sector exposures
- Filing analysis, where an agent reads key figures from annual-report PDFs and flags those that differ from the database by 5 percent or more
- Transaction reconciliation, where an agent finds accounts whose recorded balance does not match their transactions and explains the exceptions
- Fraud alert triage, where an agent investigates one fraud alert step by step with database tools and decides whether to escalate or clear it
- Peer benchmarking, where an agent ranks the companies in a sector by return and writes sector commentary
- Filing Q&A, where an agent answers 10 questions about company filings by searching about 270,000 passages of real SEC filings

About the Intel OpenVINO toolkit

OpenVINO is the Intel open-source toolkit for optimizing and deploying trained models. It takes a model from PyTorch, TensorFlow, ONNX, or Hugging Face; converts it to an intermediate representation; and compiles it for the device that will run it.[3]

It targets Intel CPUs, integrated and discrete Intel GPUs, and Intel NPUs from one codebase, so the same model can move between devices without a rewrite.

Its optimization work matters as much as its runtime. Weight compression and post-training quantization shrink a model's memory footprint and speed up arithmetic, often with little effect on accuracy.

Recent releases broaden generative AI coverage, including a GenAI API for language-model pipelines and OpenVINO Model Server for serving models at scale behind a standard endpoint.

For this study, we used OpenVINO Model Server 2026.4.1 with an Intel-published INT4 version of Qwen3-8B (4-bit integer weights), and prompt (prefix) caching.

- Board-pack narrative, where an agent writes a five-section board report for a company from a briefing

Most task types do their data work and then make one model call. The fraud alert triage agent works in a loop instead, calling database tools and the model up to nine times before it decides.



Figure 1. The eight task types in our financial-services agentic AI workload and and each type's share of the 776 tasks completed in the 420-second run we report at 30 concurrent tasks. Source: PT.

The data is synthetic and models a mid-sized firm: 500 companies, 50,000 accounts across 500 branches, 1 million transactions over 7 days, 5 years of daily prices for 1,000 stock tickers, and 500 synthetic annual-report PDFs. For filing Q&A, the agent searches 270,806 passages: 90 hold the answers in synthetic filings, and the rest come from 949 real public filings from the SEC EDGAR database, which act as distractors.

Although these tasks are financial, their shape is common to agentic AI: An agent pulls records, runs analysis code, searches documents, and asks a language model to explain or decide. If your agents do similar data-heavy work, these results offer a starting point for sizing. The throughput a server delivers within a given service level depends on the model, the prompts, the task mix, and the latency target you choose.

About Intel AMX

Intel Advanced Matrix Extensions (Intel AMX) is an accelerator built into each core of Intel Xeon 6 processors. It adds dedicated matrix-math hardware for the kind of arithmetic that dominates AI inference.[4]

The Intel Xeon 6767P processors in our PowerEdge R770 include Intel AMX, and AI software such as OpenVINO can use it on processors that offer it. Because the feature is built into every core, it comes with the server rather than as a separate card to buy and manage.

What the Dell PowerEdge R770 delivered

Throughput: How much work did the server complete?

Throughput tells you how much agent work the server finishes per minute. We measured it in steady state: the tasks completed inside each 420-second window, divided by 7 minutes. The R770 sustained 110 agent tasks per minute at 30 concurrent tasks, with a 95th-percentile task time under our 30-second target, as Figure 2 shows. That works out to about 6,650 agent tasks per hour.

Throughput held nearly flat, between 105.3 and 111.7 tasks per minute, from 28 to 34 agents. To keep memory bandwidth optimized in our configuration, we chose to run OVMS on socket 0 and run our agentic tasks, PostgreSQL, other code, and benchmark harness on socket 1. At these load levels, socket 0 ran at nearly 100 percent. In our tuning tests, most of that inference work was reading the prompts that carry each task's data.



Figure 2. Steady-state agent tasks per minute the Dell PowerEdge R770 delivered running our financial-services workload through OpenVINO Model Server, with Qwen3-8B at INT4 weight precision, by number of concurrent tasks. Each bar is the run with the median tasks per minute of three runs at that level. Higher numbers are better. Source: PT.

Latency: How long did each agent task take?

Looking at a throughput number in isolation can mislead. A server that posts a large aggregate number by queuing work deeply will feel slow to everyone waiting in that queue, so we tracked the 95th-percentile task time at every concurrency level we tested.

At 30 concurrent tasks, the 95th-percentile task time was 28.06 seconds, and the typical (median) task took 14.20 seconds. At 31 concurrent tasks, the 95th-percentile time rose to 30.39 seconds, just past our target, and at 34 tasks it reached 32.04 seconds. The three runs' 95th-percentile times at each level landed within 0.6 seconds of each other, and two more independent sweeps found the same 30-task result. Figure 3 plots that relationship.

The 30-second level applies to the mix of tasks as a whole. Individual task types varied: At 30 concurrent tasks, 95 percent of regulatory ratios tasks finished within about 15 seconds, while board-pack narrative tasks, which write five report sections each, took about 114 seconds (Table 2).

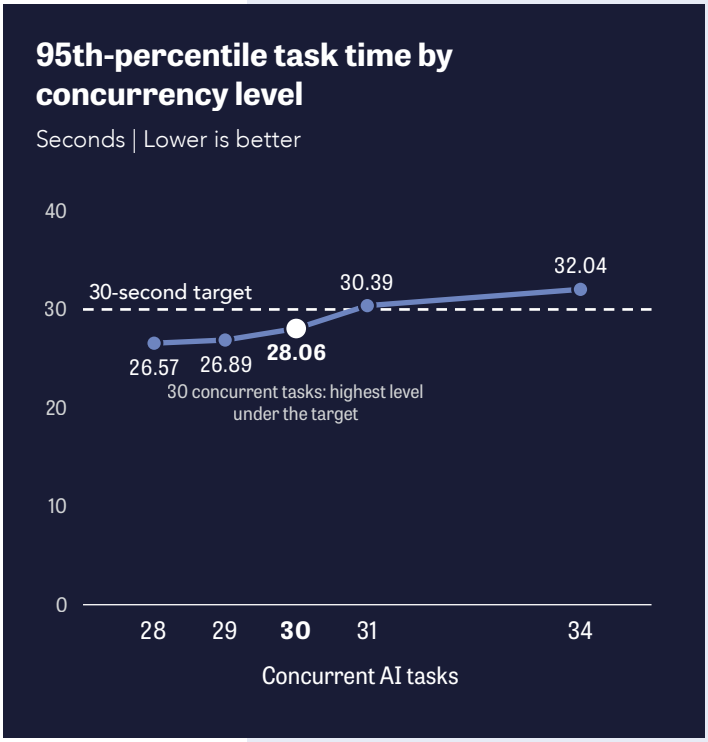


Figure 3. 95th-percentile task time in seconds on the Dell PowerEdge R770 by number of concurrent tasks, running our financial-services agentic workload through OpenVINO Model Server. Each point is the run with the median tasks per minute of three runs at that level. Source: PT.

Capacity at the service level: How much work did the server finish inside the target?

Holding the 30-second target, the PowerEdge R770 completed 110 tasks per minute at its highest compliant load, 30 concurrent tasks. At 31 concurrent tasks, the 95th-percentile time crossed the target while throughput did not rise, so added load led to queueing rather than output. The compliant concurrency level depends on the target: With a 35-second target, every level we tested, up to 34 concurrent tasks, would have qualified at about the same throughput. Table 1 shows the results at every level we tested.

At 110 tasks per minute, an 8-hour working day would yield roughly 53,000 agent tasks for this task mix. To size for your own needs, divide the agent tasks you need per hour by the hourly rate we measured (about 6,650), and adjust for your own models and tasks.

Concurrent tasks	Agent tasks per minute (steady state)	95th-percentile task time (seconds)	Meets 30-second target	Output tokens per second
28	111.71	26.57	Yes	213.8
29	111.43	26.89	Yes	214.4
30	110.86	28.06	Yes	215.5
31	105.29	30.39	No	203.8
34	107.86	32.04	No	215.1

Table 1. Results by concurrency level on the Dell PowerEdge R770: 95th-percentile task time, steady-state agent tasks per minute, and output tokens per second, each from the run with the median tasks per minute of three 420-second runs at that level. Lower task times and higher tasks per minute are better. Source: PT.

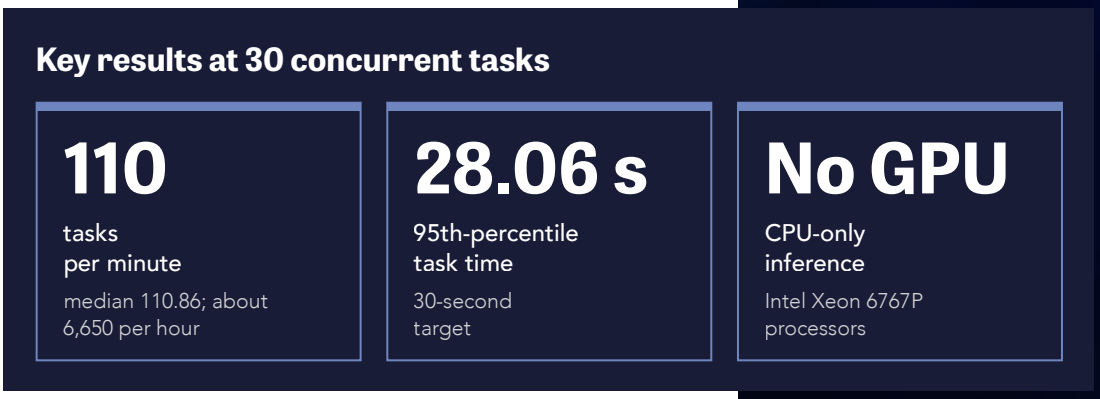


Figure 4. Key results for the Dell PowerEdge R770 at 30 concurrent tasks, the highest load we tested that held the 30-second target, running our financial-services workload through OpenVINO Model Server. From the run with the median tasks per minute of three runs. Source: PT.

Task type	Tasks completed	95th-percentile task time (seconds)
Regulatory ratios	260	14.61
Portfolio valuation	115	14.87
Filing analysis	103	22.41
Transaction reconciliation	98	21.46
Fraud alert triage	77	28.10
Peer benchmarking	72	23.79
Filing Q&A	36	31.70
Board-pack narrative	15	113.54
All task types	**776**	**28.06**

Table 2. Results by task type at 30 concurrent tasks on the Dell PowerEdge R770: tasks completed and 95th-percentile task time in the run with the median tasks per minute of three 420-second runs. Lower task times are better. Source: PT.

Where the work happens inside the server

Understanding which component does what can help you decide where to spend on the next refresh.

The Intel Xeon 6 processors carry each agent task end to end. They receive the call, prepare the input, run the model when inference stays on the CPU, and assemble the response. On a general-purpose server running mixed duties, that breadth matters, because the same processors that run the model also serve other applications.

In our configuration, we divided that work between the two processors (Figure 5). Processor 0 ran OpenVINO Model Server and the Qwen3-8B model on all 64 of its cores, one thread per core. Processor 1 ran everything else: the agents, their analysis code, document search, and the PostgreSQL database. We used no GPU. If demand outgrows the processors, the R770 has room to add up to two 450W double-width GPUs or six 75W single-width cards.

OpenVINO sits between the model and the hardware. It prepares the model for the processor it runs on, chooses math routines that suit that processor (on the Intel Xeon 6767P, these included Intel AMX), and runs the work on whichever device the deployment targets. The model we used already had its weights compressed to INT4. OpenVINO also supports Intel GPUs, but we tested only the processors.

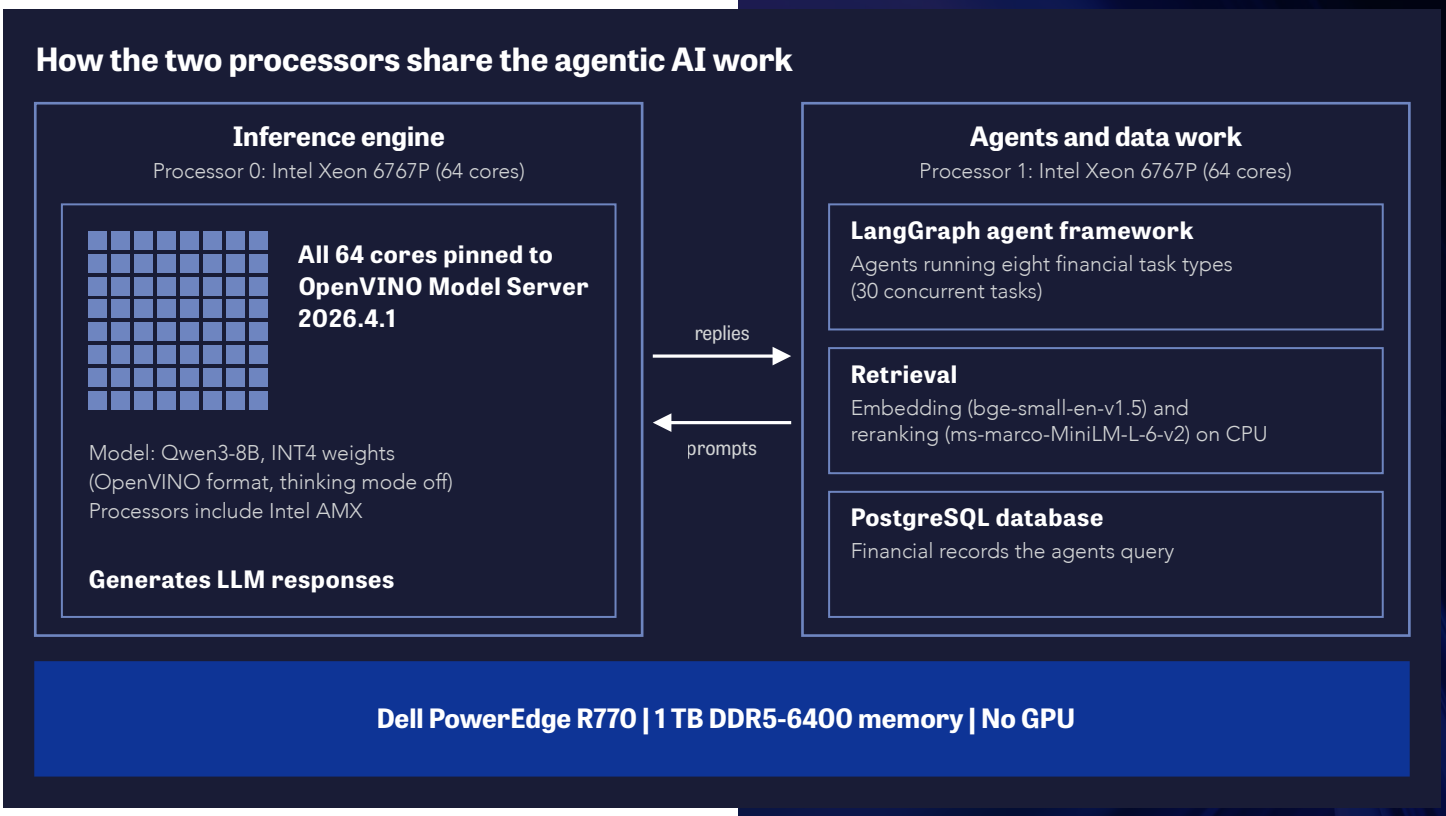


Figure 5. How we divided the agentic AI workload between the two Intel Xeon 6767P processors in the Dell PowerEdge R770. Source: PT.

Conclusion

Organizations that want to move AI from discussion into production face a choice about where to run it. Specialized accelerator platforms answer that question at one end of the market. For many workloads, the practical answer could be a general-purpose server.

Running a financial-services agentic AI workload on a Dell PowerEdge R770 with Intel Xeon 6 processors and OVMS, we recorded 110 agentic AI tasks per minute of steady-state throughput with a 95th-percentile task time of 28.06 seconds—within our 30-second target—at 30 concurrent tasks and without using GPUs. Those numbers came from a standard two-socket, 2U rack server that admins can manage with the same Dell tools they likely use to manage other PowerEdge servers in your data center.

That makes the R770 a reasonable place to start your agentic AI journey on Intel Xeon 6 processors with Intel AMX and a reasonable place to stay for workloads of this shape. When demand outgrows what the processors deliver, the chassis takes accelerators, and OpenVINO supports Intel GPUs as well as CPUs.

Learn more about the Dell PowerEdge R770 and the Intel OpenVINO toolkit.

Resources

1. Gartner, "Gartner says AI-optimized IaaS is poised to become the next growth engine for AI infrastructure," accessed October 15, 2026, https://www.gartner.com/en/newsroom/press-releases/2025-10-15-gartner-says-artificial-intelligence-optimized-iaas-is-poised-to-become-the-next-growth-engine-for-artificial-intelligence-infrastructure.
2. Dell, "PowerEdge R770 spec sheet," accessed October 15, 2026, https://www.delltechnologies.com/asset/en-us/products/servers/technical-support/poweredge-r770-spec-sheet.pdf.
3. Intel, "OpenVINO toolkit," accessed October 15, 2026, https://www.intel.com/content/www/us/en/developer/tools/openvino-toolkit/overview.html.
4. Intel, "What Is Intel® Advanced Matrix Extensions (Intel® AMX)?" accessed October 15, 2026, https://www.intel.com/content/www/us/en/products/docs/accelerator-engines/what-is-intel-amx.html.

Commissioned by Dell Technologies

In the science behind the report, we provide a detailed breakdown of the results, hardware configurations, methodologies, and any scripts we used to produce the results in this report.

Read the science behind the report ►

Principled Technologies®

Primary contributors

Tech: Corey B.

Writing: Laura W.

Design: Laura K.

PM: Greg C.

How we created this report

A PT team, which includes the contributors we've listed and others, created this report and performed the technical work behind it. In addition to our use of AI in testing, we used AI to help create charts, flow content into layout, and draft and proofread sections of this report.

Principled Technologies is a registered trademark of Principled Technologies, Inc. All other product names are the trademarks of their respective owners. For additional information, review the science behind this report.