From CPU Sizing to Token Economics: A Business-First Tool to Enterprise AI Inference on Intel Xeon 6

Planning / Implementation

As generative AI moves from experimentation into production, infrastructure planning becomes more than a benchmark comparison. Enterprises need to understand how model choice, workload characteristics, expected user demand, latency requirements, infrastructure cost, and energy consumption come together before deciding what to deploy.

Traditional benchmarking usually starts with the hardware: select a processor, run a workload, and report performance such as tokens per second. Those measurements are important, but they do not directly answer the questions customers typically ask.

The CPU AI Sizing tool was developed to connect two views, AI Sizing and Token Economics. It includes Intel Xeon 6 P-core processors ranging from 8 to 128 cores per processor in Lenovo ThinkSystem V4 servers and combines measured benchmark results with workload sizing, power modeling, server economics, and cloud comparisons.

The overall workflow of the tool is shown in the following figure.

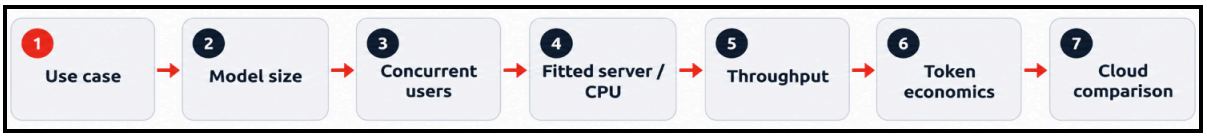


Figure 1. Workflow of the tool

The four questions the CPU Sizing Tool is designed to answer:

1. What size of model fits the use case?
2. How many concurrent users can one server support while meeting the latency SLA?
3. What throughput can the customer expect?
4. What does that translate to in $/1M output tokens, power efficiency, upfront investment, monthly cost, and cloud economics?

Accessing the CPU Sizing Tool

To access the tool:

1. Go to https://lenovopress.lenovo.com/llm-sizing-tool
2. Click the **CPU Sizing** tab to display the tool.

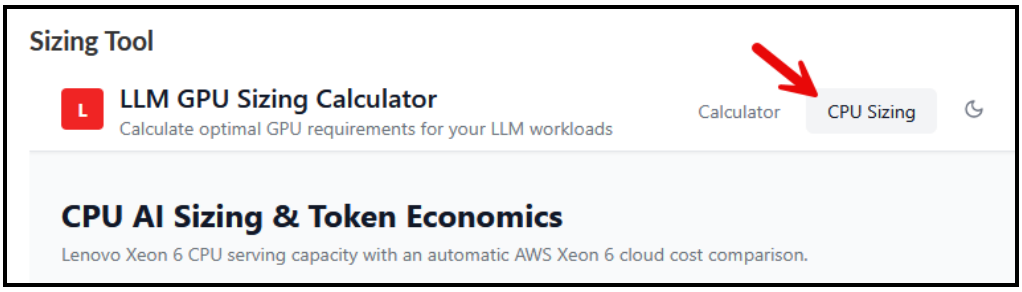


Figure 2. Click the **CPU Sizing** tab to display the tool

The tool starts with three business inputs, as shown in the figure below, rather than asking users to interpret a large matrix of benchmark results:

- Business use case
- Model size
- The number of concurrent users that is required

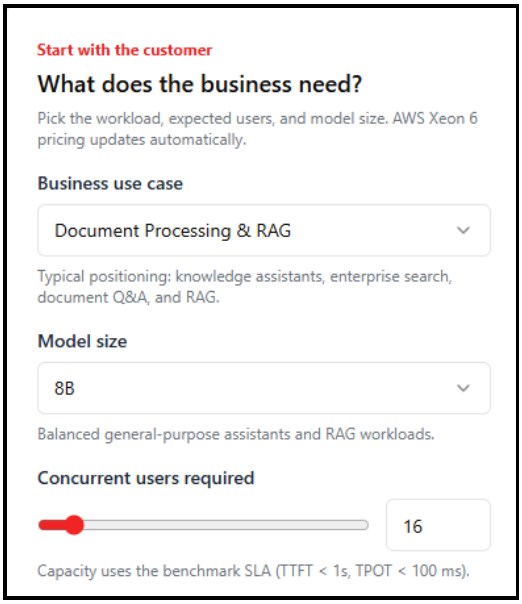


Figure 3. Three customer inputs from the CPU sizing tool

The sizing tool identifies server and CPU configurations that can meet the workload requirement and allows users to compare throughput, capacity headroom, server acquisition cost, cost per million output tokens, power efficiency, cloud economics, break even, and lifecycle savings.

The objective is not to identify one universally “best” CPU. It is to help identify the right-sized infrastructure for a particular workload and business requirement, and to translate AI benchmark results into metrics that customers, solution architects, sales teams, and business decision-makers can use.

For further instructions, see the How to Use the Sizing Tool section.

Workloads and model sizes covered

The sizing tool groups representative enterprise inference workloads into three business-oriented categories:

- Short Query & Classification
- Document Processing & RAG
- Full Document Analysis & Reporting

These categories provide a practical way to group workloads with similar performance and infrastructure characteristics for sizing purposes.

The following table summarizes the three workload categories, representative applications, and the primary infrastructure consideration associated with each workload type.

Table 1. Representative enterprise AI inference workload categories and infrastructure priorities

Business workload	Representative applications	Infrastructure priority
Short Query & Classification	FAQ, routing, extraction, classification, short interactions	High user density and throughput
Document Processing & RAG	Knowledge assistants, enterprise search, document Q&A, RAG	Balance responsiveness and capacity
Full Document Analysis & Reporting	Summarization, report generation, longer-context document workflows	Higher compute demand per request

In addition to workload type, model size is a key sizing input because it directly affects compute requirements, throughput, concurrency, and infrastructure economics. The tool groups models into three size ranges to support practical capacity planning and configuration selection, as listed in the following table.

Table 2. Model-size categories and typical infrastructure positioning (1B = 1 billion parameters)

Model size (B = billion)	Typical infrastructure positioning
1B–3B	Task-focused, higher-volume workloads where user density and economics are priorities
8B	Balanced general-purpose assistants, enterprise chat, and RAG
14B–20B	Higher-capability workloads where additional compute demand and lower concurrency may be acceptable

Note: Model size is an infrastructure-planning dimension, not a quality ranking. Application teams should separately evaluate task accuracy, reasoning quality, domain knowledge, safety, retrieval quality, and other use-case requirements.

Intel Xeon 6 and Lenovo platform coverage

The CPU sizing tool focuses on Intel Xeon 6 P-core processors. The benchmark matrix is intentionally not all-to-all: CPUs were tested on the Lenovo platforms where hardware was available, and additional P-core SKUs are estimated from a measured anchor on the same server platform.

Platforms covered in the tool are listed below:

- SR630 V4: a 2-socket 1U server supporting Xeon 6500/6700-series processors
- SR650 V4: a 2-socket 2U server supporting Xeon 6500/6700-series P-core processors
- SC750 V4: a 2-processor Neptune direct-water-cooled platform supporting Xeon 6900-series processors

The CPU sizing tool spans a broad range of Intel Xeon 6 P-core processors, allowing the tool to evaluate configurations from lower-core-count systems for lighter workloads through high-core-count platforms for greater throughput and concurrency. The following table summarizes the processor SKUs included in the sizing tool, their core counts, and the Lenovo ThinkSystem platforms on which they are supported.

Table 3. Intel Xeon 6 P-core processors and their supported Lenovo ThinkSystem platforms

Intel Xeon 6 processor	Cores / processor	Platform Eligibility
6714P	8	SR630 V4 & SR650 V4
6505P	12	SR630 V4 & SR650 V4
6724P	16	SR630 V4 & SR650 V4
6527P	24	SR630 V4 & SR650 V4
6732P	32	SR630 V4 & SR650 V4
6747P	48	SR630 V4 & SR650 V4
6767P	64	SR630 V4 & SR650 V4
6960P	72	SC750 V4
6787P	86	SR630 V4 & SR650 V4
6972P	96	SC750 V4
6980P	128	SC750 V4

The following table summarizes the Lenovo ThinkSystem platforms used for benchmarking, the Xeon 6 CPUs tested, and each platform's profile.

Table 4. Benchmark platforms and tested Intel Xeon 6 CPU configurations

Lenovo platform	Benchmark CPU(s)	Platform profile
ThinkSystem SR630 V4	Xeon 6732P	2-socket, 1U; density-oriented rack server
ThinkSystem SR650 V4	Xeon 6787P	2-socket, 2U; broader expansion and configuration flexibility
ThinkSystem SC750 V4 Neptune	Xeon 6960P, 6972P, 6980P	Two-processor node with direct-water cooling and high-core-count 6900-series P-core CPUs

Benchmark methodology: Throughput and interactive capacity

The benchmark evaluates two complementary performance modes to capture both maximum throughput and interactive serving capacity:

- Offline throughput
  Offline throughput measures maximum output-token generation when interactive latency is not the primary constraint. It is useful for batch document processing, asynchronous generation, summarization pipelines, and other throughput-oriented workloads.
- Interactive serving
  Interactive serving measures throughput and concurrency while maintaining a defined latency target. The sizing tool uses the following serving service level agreement (SLA):

  Serving SLA:

  - Time to First Token (TTFT) < 1 second
  - Time per Output Token (TPOT) < 100 milliseconds

  For each workload, the largest tested concurrency level that remains inside both thresholds becomes the measured Max Concurrent Users value. This translates engineering performance into a customer-facing capacity question: how many simultaneous users can this server support while maintaining the target response experience?

Estimating untested CPU SKUs

Testing every model, workload, CPU SKU, and server combination would create a very large matrix. To extend measured results into a practical sizing tool, the tool uses a compute-scaling estimation for untested P-core CPUs on the same server platform.

Throughput for an untested CPU is estimated by scaling the measured baseline throughput according to relative core count and base frequency:

$$Target\ Throughput\ =\ Baseline\ Throughput\ \times\ \frac{Target\ Cores\ \times\ Target\ Base\ Frequency}{Baseline\ Cores\ \times\ Baseline\ Base\ Frequency}$$

The estimation assumes the same socket count, comparable memory and software configuration, no binding bandwidth ceiling, and scaling efficiency = 1. Offline and serving throughput are estimated independently of the appropriate platform anchor.

Example: SR650 V4 with Xeon 6724P, 8B RAG

The following table shows how the measured Xeon 6787P configuration is used as the baseline to estimate throughput for the Xeon 6724P.

Table 5. Baseline and target CPU parameters for SR650 V4 throughput estimation

Parameter	**Baseline: 6787P**	**Target: 6724P**
Cores/socket	86	16
Base frequency	2.0 GHz	3.6 GHz
Serving throughput	609.89 tok/s	To estimate
Offline throughput	1,377.30 tok/s	To estimate

Scaling factor = (16 × 3.6) ÷ (86 × 2.0) = 0.3349. Applying that factor gives approximately 204.24 serving output tok/s and 461.24 offline output tok/s. These values match the planning data used by the tool.

Approximation only: This is a capacity-planning approximation. It should not be interpreted as a substitute for a measured benchmark on the target CPU, especially when memory bandwidth, NUMA behavior, frequency behavior under sustained load, or software optimization changes materially.

Estimating concurrent users

For untested P-core configurations, concurrent-user capacity is estimated from the measured SLA baseline using the formula below, which scales baseline Max Users by the ratio of target to baseline serving throughput.

$$Estimated\ Max\ Users\ =\ ROUND(Baseline\ Max\ Users\ \times\ \frac{Target\ Serving\ Throughput}{Baseline\ Serving\ Throughput})$$

For the same SR650 V4 8B RAG example, the baseline is 32 users at 609.89 tok/s and the estimated target is 204.24 tok/s. Therefore, 32 × 204.24 ÷ 609.89 = 10.72, which rounds to 11 users.

The HTML application uses Max Users as a capacity filter: a configuration “fits” only when its modeled Max Users is at least the customer’s requested concurrent users.

Production SLA caveat: Estimated Max Users values are planning estimates, not measured SLA results. Validate the target configuration on hardware before presenting an estimated concurrency value as a formal production commitment.

Modeling whole-server power and cost

Power-supply capacity is not the same as server power consumption. A 2,000 W or 3,200 W PSU rating defines available power-delivery capacity; it does not mean the server continuously consumes that amount.

For token economics, the relevant quantity is whole-server AC input power while processing the workload. For unmeasured SR630 V4 and SR650 V4 P-core configurations, the tool scales the dynamic power above the measured idle baseline by CPU TDP:

```
Target Load Power = Reference Idle Power + (Reference Load Power − Reference Idle Power) × (Target CPU TDP / Reference CPU TDP)
```

Example using the SR650 V4 reference: idle power = 543.5 W, reference workload power = 1,086.8 W, reference 6787P TDP = 350 W, and target 6724P TDP = 210 W. The modeled target load power is approximately 869.5 W.

For SC750 V4, the current economics layer uses a conservative modeled direct-water-cooled node power until workload-level node telemetry is available. All power estimates should be replaced with measured platform telemetry whenever formal customer claims are required.

Turning performance into Token Economics

Performance becomes a business metric only after it is combined with acquisition cost, utilization, electricity, data-center efficiency, and lifecycle assumptions.

The following table listed out the default assumptions used in the token-economics calculation, which can be adjusted for customer-specific lifecycle, utilization, electricity cost, and PUE.

Table 6. Default financial and operating assumptions used in the token-economics calculation

Input	Default
Server lifecycle	5 years
Active utilization	24 hours/day
Days/month	30.4375
Electricity	$0.10/kWh
PUE	1.40
Additional monthly OPEX	$0 placeholder

The core financial calculations used here are as follows:

```
Monthly Amortized CAPEX = Server Package Cost / (Lifecycle Years × 12)
```

```
Monthly Energy + Cooling = Load Power kW × Active Hours/Day × Days/Month × Electricity Rate × PUE
```

```
Monthly All-In Cost = Monthly Amortized CAPEX + Monthly Energy + Cooling + Other Monthly OPEX
```

Token economics metrics are listed in the following table.

Table 7. Token economics metrics

Business metric	Formula
Serving Full TCO ($/1M output tokens)	Monthly all-in cost ÷ monthly serving output (million tokens)
Serving CAPEX-only ($/1M output tokens)	Monthly amortized CAPEX ÷ monthly serving output (million tokens)
Offline Full TCO ($/1M output tokens)	Monthly all-in cost ÷ monthly offline output (million tokens)
Serving Efficiency (M tok/s/MW)	Serving output tok/s ÷ load power W
Serving Yield (M output tok/MWh)	Serving output tok/s × 3,600 ÷ load power W
Monthly SLA Output (B tokens)	Serving tok/s × active h/day × days/month × 3,600 ÷ 1B

TCO metric distinction: Serving Full TCO uses a monthly all-in cost. Serving CAPEX-only isolates monthly amortized server CAPEX as a separate $/1M output-token metric.

Adding the cloud dimension

Cloud infrastructure is priced in dollars per instance-hour, while AI consumption is increasingly discussed in dollars per token. The tool converts hourly rental into the same token-economic unit used for on-prem inference:

```
Cloud $/1M Output Tokens = Cloud Rental $/Hour ÷ (Cloud Output tok/s × 3,600 ÷ 1,000,000)
```

The current comparison uses Intel Xeon 6–based AWS memory-optimized instance families, US East (N. Virginia), Linux, shared tenancy, 1 TiB RAM per instance, and both On-Demand and 3-Year EC2 Instance Savings Plan No Upfront purchasing models. The commercial comparison currently uses two AWS instances per Lenovo server as a planning assumption.

When measured AWS inference throughput is unavailable, cloud throughput is normalized to the selected Lenovo configuration. This creates a capacity-normalized planning comparison; it is not a measured cloud benchmark. Enter measured aggregate AWS throughput in Advanced assumptions when available.

Lifecycle savings and rank

Lifecycle savings compares the modeled cost of cloud rental with the total on-premises cost over the selected period. This value is also used to rank configurations by their long-term economic benefit.

```
5-Year Net Savings = 5-Year Cloud Rental Cost − (Upfront Server Package Cost + 5-Year On-Prem Cash OPEX)
```

The sizing tool ranks only configurations that meet the requested concurrent-user capacity. Rank #1 is the configuration with the highest modeled lifecycle savings versus AWS On-Demand; ties are broken by higher savings versus the 3-Year No-Upfront plan, then by lower Lenovo package price. This is an economic rank—not a model-quality or benchmark-performance rank.

How to use the Sizing Tool

The tool can be accessed at https://lenovopress.lenovo.com/llm-sizing-tool and then clicking the **CPU Sizing** tab as described in the Accessing the CPU Sizing Tool section.

The interface is designed for sales, marketing, solution architects, and customers who do not want to interpret a raw benchmark matrix. The recommended workflow is as follows:

1. Select the business use case: Short Query & Classification, Document Processing & RAG, or Full Document Analysis & Reporting.

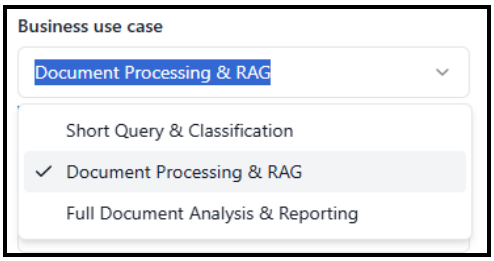


Figure 4. Select the business use case

2. Select the model size: 1B–3B, 8B, or 14B–20B.

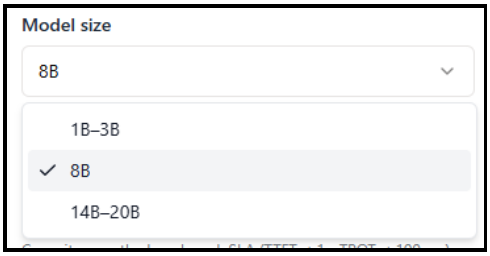


Figure 5. Select the model size

3. Enter the required concurrent users. This becomes the capacity threshold used to filter the server/CPU options.

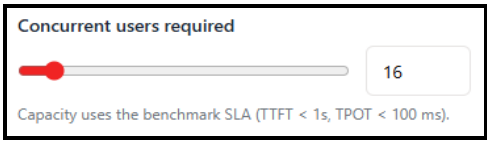


Figure 6. Enter the required concurrent users

4. Review the server/CPU dropdown. The servers listed means the configurations that meet the requested user target under the sizing model; options are ordered by the current 5-year savings rank.

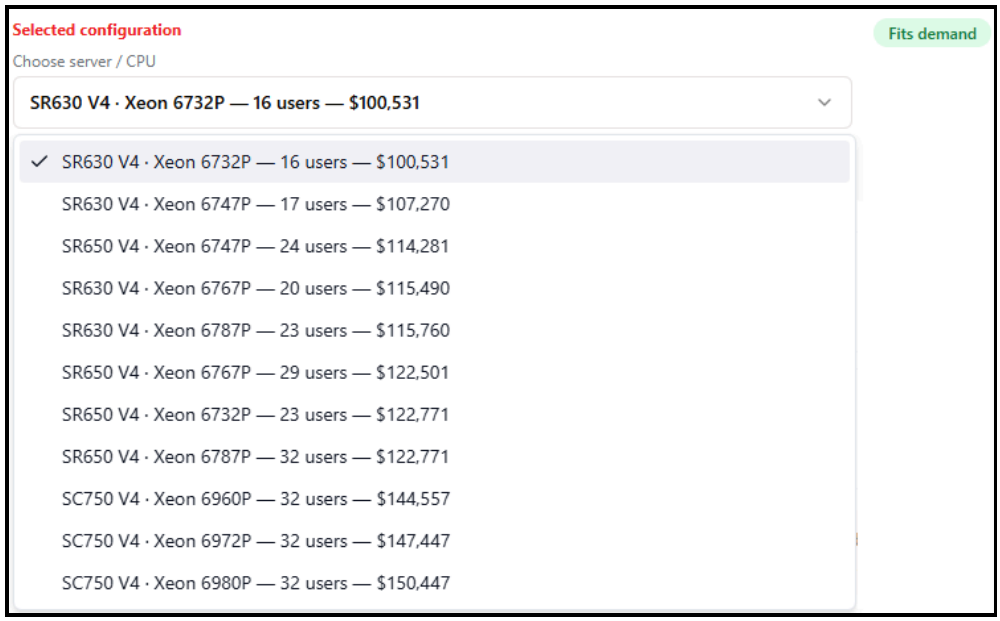


Figure 7. Review the server/CPU dropdown

5. Select any configuration to update capacity headroom, server package price, on-prem $/1M tokens, cloud $/1M, cash breakeven, lifecycle savings, and charts.

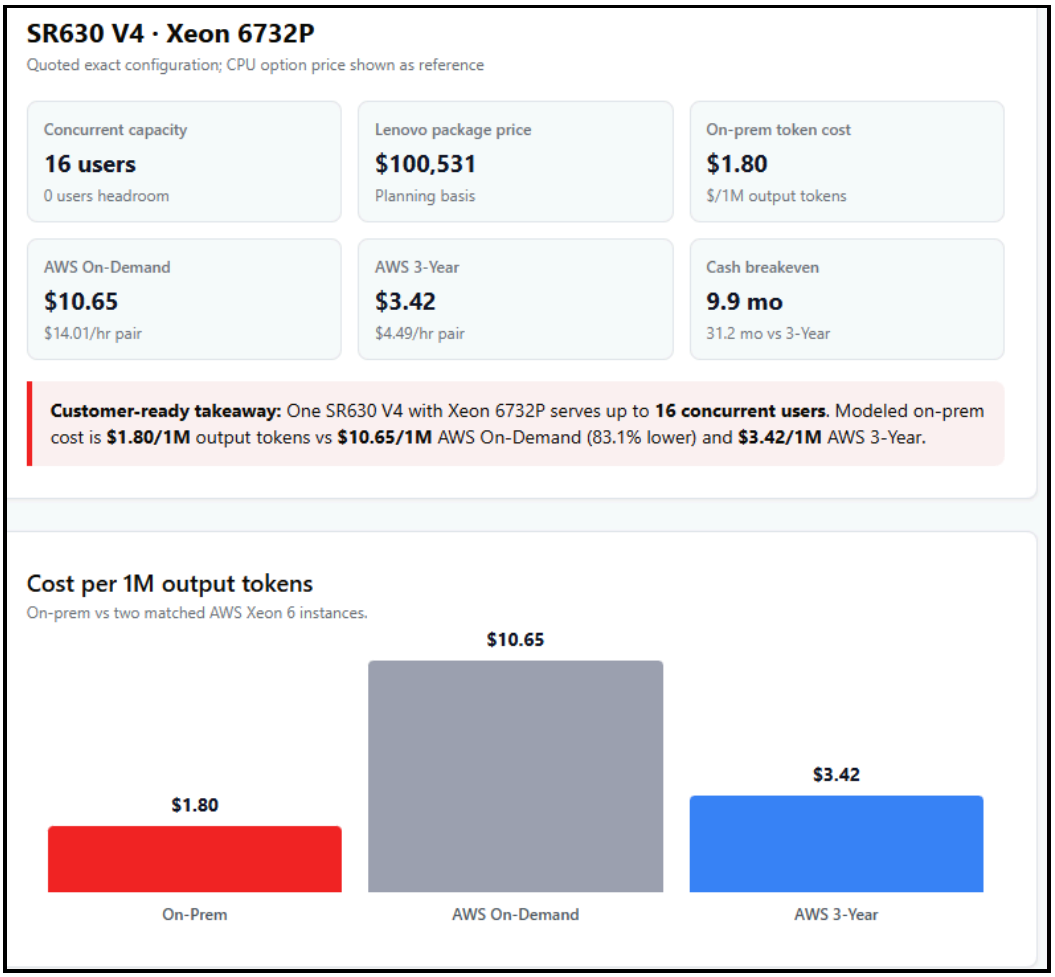


Figure 8. Select the configuration

6. Use Advanced assumptions to change active hours/day, lifecycle, electricity rate, PUE, or to enter measured aggregate cloud throughput.

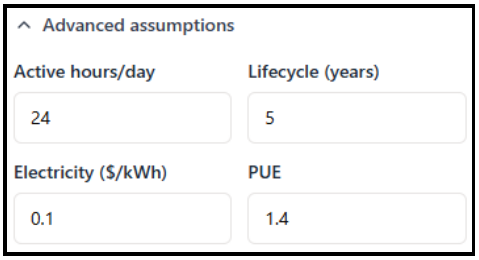
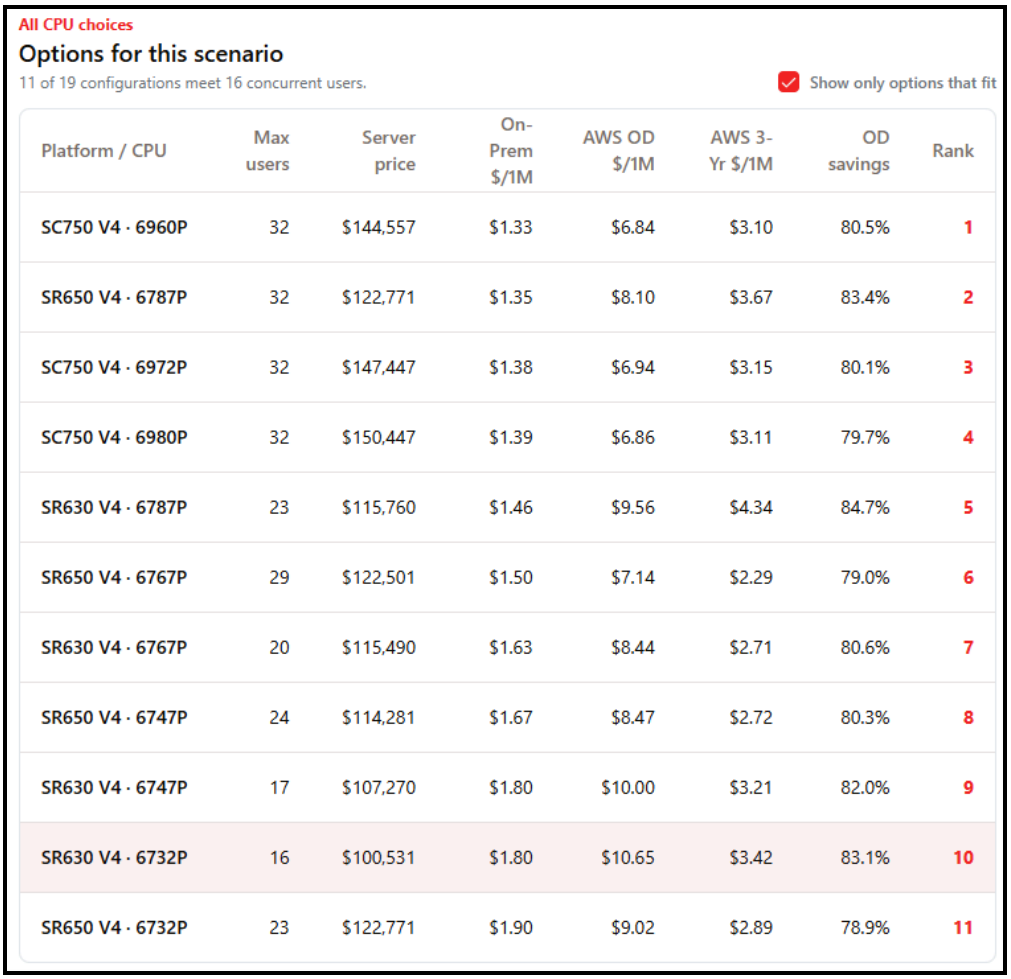


Figure 9. Advanced assumptions

7. Use the options table to compare all CPU choices and the Print / Save PDF action to capture a customer-ready view.



All CPU choices

Options for this scenario

11 of 19 configurations meet 16 concurrent users.

Show only options that fit

Platform / CPU	Max users	Server price	On-Prem $/1M	AWS OD $/1M	AWS 3-Yr $/1M	OD savings	Rank
SC750 V4 · 6960P	32	$144,557	$1.33	$6.84	$3.10	80.5%	1
SR650 V4 · 6787P	32	$122,771	$1.35	$8.10	$3.67	83.4%	2
SC750 V4 · 6972P	32	$147,447	$1.38	$6.94	$3.15	80.1%	3
SC750 V4 · 6980P	32	$150,447	$1.39	$6.86	$3.11	79.7%	4
SR630 V4 · 6787P	23	$115,760	$1.46	$9.56	$4.34	84.7%	5
SR650 V4 · 6767P	29	$122,501	$1.50	$7.14	$2.29	79.0%	6
SR630 V4 · 6767P	20	$115,490	$1.63	$8.44	$2.71	80.6%	7
SR650 V4 · 6747P	24	$114,281	$1.67	$8.47	$2.72	80.3%	8
SR630 V4 · 6747P	17	$107,270	$1.80	$10.00	$3.21	82.0%	9
SR630 V4 · 6732P	16	$100,531	$1.80	$10.65	$3.42	83.1%	10
SR650 V4 · 6732P	23	$122,771	$1.90	$9.02	$2.89	78.9%	11

Figure 10. Server+processor choices for this scenario

Worked example: 8B RAG for 20 concurrent users

To illustrate the user flow, assume an enterprise knowledge assistant with an 8B model, Document Processing & RAG workload, and a requirement for 20 concurrent users. Under the current default financial assumptions, the tool first removes configurations whose Max Users is below 20, then ranks the remaining options by modeled 5-year savings versus AWS On-Demand.

The following table shows the server configurations for this scenario and compares their capacity, throughput, package cost, on-prem token cost, and modeled 5-year savings versus AWS On-Demand.

Table 8. Server configurations for the 8B RAG, 20-concurrent-user scenario

Savings rank	Server / CPU	Max users	Serving tok/s	Package price	On-prem $/1M	5-Yr savings vs OD
1	SR630 V4 · 6787P	23	516.5	$115,760	$1.46	$660K
2	SR650 V4 · 6787P	32	609.9	$122,771	$1.35	$650K

Savings rank	Server / CPU	Max users	Serving tok/s	Package price	On-prem $/1M	5-Yr savings vs OD
3	SC750 V4 · 6960P	32	722.0	$144,557	$1.33	$627K
4	SC750 V4 · 6972P	32	711.4	$147,447	$1.38	$625K
5	SC750 V4 · 6980P	32	720.4	$150,447	$1.39	$622K

This example demonstrates why the ranking is not simply “highest throughput” or “lowest purchase price.” The rank combines the scenario’s capacity with the modeled lifecycle economics. A different user target, model size, utilization level, or cloud throughput assumption can change the order.

How to consider the result: You may prefer Rank #1 for modeled lifecycle savings, but another option can still be the better design when more headroom, measured rather than estimated performance, rack density, expansion, cooling strategy, or standardization matters. The tool supports comparison; it does not replace architecture judgment.

Interpreting the sizing outputs

The sizing tool produces both technical and economic outputs. The following table summarizes what each metric represents and how it can be interpreted when evaluating infrastructure options.

Table 9. Interpretation of sizing and token-economics outputs

Metrics	What the metric represents	How to interpret it
Max Users	Single-server concurrency capacity under the sizing SLA	Capacity fit and growth headroom
Serving tok/s	Interactive output throughput for the selected workload	Expected aggregate generation rate
Server price	Planning/quoted package acquisition cost	Upfront investment
On-prem $/1M	Amortized CAPEX + modeled energy/facility cost per 1M output tokens	Unit economics
Power efficiency / yield	Output normalized by server load power	Energy and data-center planning
AWS $/1M	Hourly cloud rental translated to token economics	Common-unit cloud comparison
Cash breakeven	When cumulative cloud rental exceeds upfront on-prem investment plus cash OPEX	Time-to-economic-advantage
Lifecycle savings	Modeled cloud cost minus on-prem cash cost over the selected lifecycle	Longer-term business value
Savings rank	Economic ordering of server configurations	Shortlist, not an absolute “best CPU” claim

Together, these metrics provide a more complete view than throughput alone by connecting workload capacity and latency requirements with acquisition cost, operating cost, power efficiency, and cloud economics.

The bigger picture: Workload-centric AI infrastructure

There is no single infrastructure platform that is optimal for every AI workload. GPU infrastructure remains important for training, very large models, and the most demanding inference. Cloud infrastructure offers rapid access and flexible capacity. CPU-based infrastructure can be a practical production option for

smaller and mid-sized models, predictable workloads, moderate concurrency, governance-sensitive applications, and sustained inference demand.

The better infrastructure question: Not “Can CPUs replace GPUs?” but “What infrastructure delivers the required application experience at the right business cost?”

That is the philosophy behind CPU AI Sizing & Token Economics: connect the business use case, model size, concurrency, latency, throughput, processor choice, server CAPEX, energy consumption, and cost per token so infrastructure planning becomes workload-driven and economically transparent.

Validation guardrails and known limitations

The following guardrails clarify where the sizing results are measured, modeled, or assumption-based, and where additional validation is recommended before using them for production planning or formal commitments:

- The benchmark matrix is not all-to-all. Measured configurations and planning estimates coexist in the underlying model.
- Throughput scaling assumes comparable platform configuration and compute-limited behavior. Memory bandwidth, NUMA topology, frequency under load, software versions, quantization, and kernel/framework optimizations can change real scaling.
- Estimated Max Users is derived from throughput scaling and must be validated before being used as a formal SLA commitment.
- Load power for unmeasured SKUs is modeled. Measured XCC/SMM/platform telemetry is preferred for formal power and TCO claims.
- Additional monthly OPEX is currently a placeholder. Colocation, maintenance, networking, administration, financing, and other local costs should be added when relevant.
- Cloud pricing changes over time, and cloud throughput should be measured for a formal apples-to-apples comparison. The current normalized-throughput mode is a planning assumption.
- The economic rank is scenario-specific and should not be interpreted as a model-quality ranking or universal hardware recommendation.

References

For more information, see the following resources:

- Lenovo ThinkSystem SR630 V4 Product Guide
  https://lenovopress.lenovo.com/lp1971-thinksystem-sr630-v4-server
- Lenovo ThinkSystem SR650 V4 Product Guide
  https://lenovopress.lenovo.com/lp2127-thinksystem-sr650-v4-server
- Lenovo ThinkSystem SC750 V4 Neptune Product Guide
  https://lenovopress.lenovo.com/lp2009-thinksystem-sc750-v4-neptune-server
- Intel Xeon Processors
  https://www.intel.com/content/www/us/en/products/details/processors/xeon.html
- Amazon EC2 X8i Instances — Intel Xeon 6
  https://aws.amazon.com/ec2/instance-types/x8i/
- Amazon EC2 R8i Instances — Intel Xeon 6
  https://aws.amazon.com/ec2/instance-types/r8i/
- AWS EC2 On-Demand Pricing
  https://aws.amazon.com/ec2/pricing/on-demand/

- AWS Savings Plans — What are Savings Plans?
  https://docs.aws.amazon.com/savingsplans/latest/userguide/what-is-savings-plans.html

Benchmark and estimation disclaimer

Performance results were obtained under specific benchmark configurations and workloads. Directly tested results represent those environments. Values calculated for untested CPU SKUs—including throughput, concurrent-user capacity, load power, token economics, cloud comparisons, breakeven, and lifecycle savings—are capacity-planning estimates derived from the methodology described in this article.

Actual results can vary based on processor configuration, memory, software stack, model implementation, quantization, workload characteristics, system utilization, power configuration, cloud instance performance, pricing, and other environmental factors. Estimated configurations should be validated on the target hardware before being used for formal performance, power, cost, or SLA commitments.

Author

Kelvin He is an AI Data Scientist at Lenovo. He is a seasoned AI and data science professional specializing in building machine learning frameworks and AI-driven solutions. Kelvin is experienced in leading end-to-end model development, with a focus on turning business challenges into data-driven strategies. He is passionate about AI benchmarks, optimization techniques, and LLM applications, enabling businesses to make informed technology decisions.

Related product families

Product families related to this document are the following:

- Intel Alliance
- Processors

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.

© Copyright Lenovo 2026. All rights reserved.

This document, LP2530, was created or updated on October 9, 2026.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2530
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2530.

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®
Neptune™
ThinkSystem®

The following terms are trademarks of other companies:

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

Linux® is the trademark of Linus Torvalds in the U.S. and other countries.

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