Analyzing the Performance of Intel Xeon CPUs for Enterprise Automated Speech Recognition

Planning / Implementation

Automated Speech Recognition (ASR) converts spoken audio into written text for applications such as transcription, captioning, contact-center analytics, and voice assistants. Modern ASR systems increasingly use transformer-based deep learning models trained on large speech datasets, enabling stronger accuracy across accents, domains, and speaking conditions. For enterprise deployment, accuracy must be balanced with serving efficiency, latency, concurrency, and infrastructure cost.

Automated Speech Recognition deployment should be evaluated in two stages:

1. Whether optimization changes recognition quality
2. How much serving capacity the optimized system can deliver under load

In this paper, WER (Word Error Rate) is used primarily to confirm that moving from FP16 to INT4 does not materially degrade Whisper Large v3 accuracy on English ASR benchmarks. This establishes INT4 as a viable efficiency-oriented deployment option rather than a compromise that significantly harms transcription quality.

After validating accuracy stability, the paper uses a vLLM-based LibriSpeech serving benchmark to measure full system capacity on two Intel Xeon CPU SKUs. This serving benchmark evaluates BF16 and W8A8 configurations to characterize practical CPU inference performance under production-style request load.

By adding the vLLM LibriSpeech serving benchmark, this paper expands the evaluation from model/runtime accuracy-latency trade-offs to CPU serving capacity. The comparison highlights how 5th Gen and 6th Gen Intel Xeon processors can support scalable ASR serving, with AVX-512 vector acceleration and Intel AMX matrix acceleration contributing to transformer inference performance on modern CPU platforms.

The resulting methodology provides a more complete deployment view:

- WER measures recognition quality
- OpenVINO results measure optimized ASR runtime efficiency
- vLLM serving metrics measure the concurrency and tail-latency behavior needed for production transcription services

Experimental Scope and Methodology

This section defines the model variants, benchmark workloads, evaluation metrics, and comparison controls used to measure both transcription quality and CPU serving performance.

Topics in this section:

Model, Precision, Backend Under Test

The benchmark compares five deployment configurations based on the same Whisper Large v3 model family, as listed in the following table.

The first three configurations are used to evaluate recognition accuracy and runtime efficiency across FP16 and INT4 variants. The vLLM BF16 and vLLM W8A8 configurations are added to evaluate CPU serving capacity under production-style request load.

Table 1. Model, Precision, and Backend Under Test

Model	Runtime / Precision	Role in Study
openai/whisper-large-v3	Hugging Face FP16	Baseline implementation of the original checkpoint.
OpenVINO/whisper-large-v3-fp16-ov	OpenVINO FP16	Accuracy-preserving optimized runtime variant.
OpenVINO/whisper-large-v3-int4-ov	OpenVINO INT4	Aggressively compressed deployment variant for efficiency-sensitive serving.
openai/whisper-large-v3	vLLM BF16	CPU serving-capacity configuration used under BF16 inference
openai/whisper-large-v3	vLLM W8A8	CPU serving-capacity configuration with weight-and-activation INT8 quantization

Because the configurations share the same base model family, the comparison is framed as a deployment and serving study rather than an ASR architecture comparison. WER is used first to verify that lower-precision deployment does not materially degrade recognition quality. After that accuracy check, vLLM BF16 and W8A8 benchmarks are used to measure full CPU serving capacity on the selected Intel Xeon platforms.

Benchmark Workload

The benchmark uses two English ASR datasets to evaluate both clean-speech accuracy and serving capacity under controlled CPU inference conditions.

Table 2. Benchmark Workload

Dataset	Domain	Speech Style	Acoustic Difficulty	Usage
LibriSpeech	Audiobooks	Read speech	Low	WER evaluation and vLLM serving-capacity benchmarking
TED-LIUM Release 3	TED talks	Prepared speech	Medium	WER evaluation only

LibriSpeech provides a clean-speech reference point for both recognition accuracy and serving-capacity measurement. TED-LIUM adds more natural speaking variability and is used to confirm that accuracy remains stable beyond clean audiobook speech.

The vLLM benchmark is run only on LibriSpeech because the MLPerf Inference Whisper benchmark from MLCommons uses Whisper Large v3 with a modified LibriSpeech audio dataset. This keeps the serving-capacity test aligned with the MLCommons/MLPerf benchmark direction while separating it from the broader WER comparison across LibriSpeech and TED-LIUM.

Metrics

The evaluation uses accuracy metrics to measure transcription quality and serving metrics to measure deployment capacity under load:

- **WER:** Word Error Rate, defined as , where substitutions, deletions, and insertions are divided by the number of reference words. Lower WER indicates better recognition accuracy.
- **Throughput tok/s:** vLLM serving throughput, measured as the number of generated tokens produced per second across active requests. Higher tok/s indicates greater serving capacity.
- **Throughput xRT:** Real-time throughput factor, measuring how many times faster than real time the system processes the audio workload. For example, 10xRT means one second of wall-clock time processes ten seconds of audio.
- **Maximum concurrent users:** The highest number of simultaneous users or requests the serving system can sustain while meeting the selected latency or quality-of-service target.
- **P99 TTFT:** The 99th-percentile Time to First Token. This captures worst-case first-token responsiveness under serving load.
- **P99 TPOT:** The 99th-percentile Time Per Output Token. This captures worst-case sustained decoding latency after the first token is generated.

Fair-Comparison Controls

The WER comparison should be run on the same platform and CPU SKU for all three variants: Hugging Face FP16, OpenVINO FP16, and OpenVINO INT4. Hardware, decoding configuration, language setting, preprocessing, timestamp policy, and transcript normalization should be matched so that differences reflect deployment variant behavior rather than test-condition changes.

The vLLM serving-capacity benchmark is evaluated separately on two Intel Xeon CPU SKUs. This part intentionally varies the CPU platform to measure capacity scaling with BF16 and W8A8, including throughput, xRT, concurrency, P99 TTFT, and P99 TPOT.

Results

This section summarizes the benchmark findings across accuracy, runtime efficiency, and CPU serving capacity, showing how each deployment configuration performs under the selected ASR workloads.

Topics in this section:

Accuracy and Serving Efficiency

The following results were measured on the 5th Gen Intel Xeon processor 8568 platform. This comparison evaluates whether OpenVINO FP16 and INT4 optimizations improve runtime efficiency while maintaining recognition accuracy close to the Hugging Face FP16 baseline.

Table 3. Accuracy and Serving Efficiency Results

Model Variant	Runtime / Precision	LibriSpeech WER	TED-LIUM WER	Throughput (audio sec/s)	TTFT (ms)	TPOT (ms/token)
openai/whisper-large-v3	HF FP16	2.00%	3.88%	4.02	925.2	33.27
OpenVINO/whisper-large-v3-fp16-ov	OpenVINO FP16	1.98%	3.89%	10.42	139.8	20.23
OpenVINO/whisper-large-v3-int4-ov	OpenVINO INT4	2.01%	3.93%	11.55	150.9	17.37

The latency and throughput trend is clear. The Hugging Face FP16 baseline processes 4.02 audio seconds per wall-clock second, while OpenVINO FP16 reaches 10.42 and OpenVINO INT4 reaches 11.55. TTFT falls from 925.2 ms to approximately 140 ms, and TPOT falls from 33.27 ms/token to 20.23 ms/token for FP16 and 17.37 ms/token for INT4.

Relative Change Versus Hugging Face FP16 Baseline

The following relative-change results are based on the same 5th Gen Intel Xeon processor platform used in the accuracy and serving-efficiency comparison above. Values are shown relative to the Hugging Face FP16 baseline to highlight the percentage change in accuracy, throughput, and latency for each optimized deployment variant.

Table 4. Relative Change Versus Hugging Face FP16 Baseline

Model Variant	Δ Avg. WER	Δ Throughput	Δ TTFT	Δ TPOT
OpenVINO/whisper-large-v3-fp16-ov	~0%	+159.20%	-84.89%	-39.19%
OpenVINO/whisper-large-v3-int4-ov	+1.02%	+187.31%	-83.69%	-47.79%

OpenVINO FP16 is the safest optimized deployment option in this result set because it closely preserves baseline accuracy while delivering a large runtime improvement. OpenVINO INT4 provides the strongest observed efficiency result, with the highest throughput and the lowest TPOT, while maintaining comparable average WER across LibriSpeech and TED-LIUM.

vLLM Serving-Capacity Results for Whisper Large v3

The vLLM benchmark extends the evaluation from offline accuracy and runtime efficiency to production-style CPU serving capacity. These results compare BF16 and W8A8 inference across the 5th Gen and 6th Gen Intel Xeon platforms, measuring throughput, concurrency, and tail-latency behavior under offline, batch, and single-request modes.

Table 5. vLLM Serving-Capacity Results for Whisper Large v3

Mode	CPU	Precision	Concurrent Users	Throughput xRT	Throughput tok/s	P99 TTFT (ms)	P99 TPOT (ms/token)
Offline	6980P_128C	BF16	-	556.84	1,737.79	-	-
	8568_48C	BF16	-	231.29	722.19	-	-
		W8A8	-	266.30	836.30	-	-
Batch	6980P_128C	BF16	72	-	-	2,147.54	94.65
	8568_48C	BF16	48	-	-	2791.86	77.06
		W8A8	96	-	-	2498.33	100.9
Single	6980P_128C	BF16	-	-	-	278.98	19.42
	8568_48C	BF16	-	-	-	1,181.28	60.05
		W8A8	-	-	-	832.28	34.84

The vLLM serving benchmark shows Whisper Large v3 outputs across offline, batch, and single-request modes. In offline mode, the reported outputs include 556.84xRT / 1,737.79 tok/s on 6980P_128C BF16, 231.29xRT / 722.19 tok/s on 8568_48C BF16, and 266.30xRT / 836.30 tok/s on 8568_48C W8A8. In batch and single modes, the benchmark reports tail latency and serving load, including 72 concurrent users with 2,147.54 ms P99 TTFT and 94.65 ms P99 TPOT on 6980P_128C BF16 batch mode, and single-request latency as low as 278.98 ms P99 TTFT and 19.42 ms P99 TPOT on 6980P_128C BF16.

Output Interpretation and Business Value

The results below show that Whisper Large v3 can be optimized for CPU-based ASR serving without materially changing recognition quality. The WER results indicate that OpenVINO FP16 and INT4 remain close to the Hugging Face FP16 baseline, meaning lower-precision deployment can be considered after accuracy validation.

In addition, the 5th Gen Intel Xeon 8568 (48 cores) and the 6th Gen Xeon 6980P (128 cores) can each handle the two dominant speech workloads on CPU alone, no GPU required.

Node A: 6th Gen Xeon 6980P (128 cores)

Node A represents the high-throughput 6th Gen Intel Xeon platform used to evaluate maximum CPU-based Whisper Large v3 serving capacity. Its 128-core configuration is positioned for large-scale offline transcription and dense real-time ASR workloads where per-node throughput and concurrency are primary deployment priorities.

Topics in this section:

Offline Audio Processing — Throughput to Expect

For offline transcription, Node A is evaluated as a batch-processing engine where the main objective is to maximize the amount of recorded audio processed per unit of compute time.

Table 6. Node A Offline Audio Processing Throughput

Metric	Value & What It Means
Best config	Whisper Large v3, BF16, Offline mode
Real-time factor	556.84 xRT (1,737.79 tok/s) — ~556 seconds of audio transcribed per second of compute.
Per minute	≈ 9.3 hours of audio processed per minute of compute.
Per hour	≈ 557 hours of audio per compute-hour.
Per 24h node-day	≈ 13,000 hours of audio cleared by a single node in one day.
Worked example	100,000-hour archive (e.g. multi-year call recordings) cleared in ≈ 8 days on one node.

Real-Time Voice Captioning — Max Concurrent Users

For real-time captioning, Node A is evaluated by the number of simultaneous audio streams it can sustain while keeping first-token response time and per-token decoding latency within practical serving limits.

Table 7. Node A Real-Time Voice Captioning Capacity

Metric	Value & What It Means
Tested config	Whisper Large v3, BF16, Batch mode
Sustained users	72 concurrent streams at P99 TTFT ≈ 2,147 ms, P99 TPOT ≈ 94.65 ms.

Business Value — Node A

For enterprise ASR deployments, Node A provides the strongest value when the priority is clearing large transcription backlogs quickly or maximizing real-time stream density on a single CPU server.

- **Backlog clearance:** one node replaces weeks of cloud-API processing and per-minute fees; CPU-only means no GPU procurement or supply constraint.
- **Fewer nodes, less orchestration:** highest per-node throughput minimizes node count, networking, and management overhead for large pipelines.
- **Deploy as:** the offline transcription workhorse; also the real-time choice where maximum stream density per node is required.

Node B: 5th Gen Xeon 8568 (48 cores)

Node B represents the 5th Gen Intel Xeon platform for efficient CPU-based Whisper Large v3 serving, balancing strong throughput, concurrency, and deployment cost.

Topics in this section:

Offline Audio Processing — Throughput to Expect

For offline transcription, Node B measures how quickly recorded audio can be processed in batch mode.

Table 8. Node B Offline Audio Processing Throughput

Metric	Value & What It Means
Best config	Whisper Large v3, W8A8, Offline mode (BF16 figures in parentheses)
Real-time factor	266.30 xRT (836.30 tok/s) with W8A8; 231.29 xRT in BF16 — quantization adds ~15%.
Per minute	≈ 4.4 hours of audio processed per minute of compute (W8A8).
Per hour	≈ 267 hours of audio per compute-hour (W8A8).
Per 24h node-day	≈ 6,400 hours of audio cleared by a single node in one day.
Worked example	100,000-hour archive cleared in ≈ 16 days on one node (W8A8).

Real-Time Voice Captioning — Max Concurrent Users

For real-time captioning, Node B measures the number of concurrent audio streams it can support with acceptable latency.

Table 9. Node B Real-Time Voice Captioning Capacity

Metric	Value & What It Means
Best config	Whisper Large v3, W8A8, Batch mode
Sustained users	96 concurrent streams at P99 TTFT ≈ 2,498 ms, P99 TPOT ≈ 100.9 ms (W8A8).
BF16 baseline	48 concurrent streams at TPOT ≈ 77.06 ms — quantization doubles user density on the same hardware.
Planning guidance	Plan ≈ 96 streams/node with W8A8; scale horizontally by adding nodes as demand grows.

Business Value — Node B

Node B provides strong business value as an efficient scale-out option for real-time ASR, especially when W8A8 quantization is used to increase stream capacity on the same hardware.

- **W8A8 = capacity for free:** raise streams per node (48 → 96) on identical hardware halves the node count for a given user base — a direct ~50% cut in servers, power, rack space, and licensing.
- **Efficient unit of scale:** size the fleet to peak concurrent users; add nodes linearly as demand grows.

Conclusion

The results show that Whisper Large v3 can be optimized for CPU-based ASR deployment while preserving practical transcription quality and delivering measurable serving capacity. The WER comparison confirms that FP16-to-INT4 optimization keeps accuracy close to the baseline, while the vLLM serving benchmark translates system performance into business capacity through xRT, tok/s, concurrent users, P99 TTFT, and P99 TPOT.

From a deployment perspective, OpenVINO FP16 is the safest accuracy-preserving path, INT4 and W8A8 are strong options for efficiency and concurrency, and high-core Xeon CPU serving provides the best fit for large-scale offline transcription and low-latency real-time workloads.

Overall, the study demonstrates that modern CPU platforms can support scalable Whisper Large v3 transcription, helping enterprises reduce GPU dependency, size infrastructure more predictably, and deploy ASR across batch, real-time captioning, meeting assistant, and voice translation use cases.

System Configuration

The experiments were conducted on a CPU-based server platform with the hardware and software configuration shown in the following tables.

Table 10. System Hardware Configuration

Component	**5th Gen Xeon System Details**	**6th Gen Xeon System Details**
System	Lenovo ThinkSystem SR680a V3	Lenovo ThinkSystem SC750 V4
CPU	2× Intel Xeon Platinum 8568Y+	2x Intel 6th Xeon 6980P
CPU topology	48 cores/socket, 2 threads/core, 192 total threads	128 cores/socket, 2 threads/core, 512 total threads
CPU frequency	Up to 4.0 GHz	Up to 3.9 GHz
Memory	2.0 TB 6400 DDR	1.0 TB 6400 DDR

Table 11. Software Configuration

Software	**Version**
System Software Configuration	
OS	Ubuntu 24.04.3 LTS (Noble Numbat)
Kernel	6.8.0-94-generic
Python	3.12
Key Python Package Versions	
torch	2.9.1+cu130
transformers	4.57.6
sentence-transformers	5.2.2
openvino	2026.1.0
optimum-intel	1.27.0
pandas	2.3.3
huggingface_hub	0.36.2
vLLM	0.21.1 + cpu

Resources

For more information, see these web pages:

- Lenovo ThinkSystem SR680a V3 Server:
  https://lenovopress.lenovo.com/lp1909-thinksystem-sr680a-v3-server
- Lenovo ThinkSystem SC750 V4 Server:
  https://lenovopress.lenovo.com/lp2009-thinksystem-sc750-v4-neptune-server
- Whisper Large v3 model card:
  https://huggingface.co/openai/whisper-large-v3
- OpenVINO Whisper Large v3 FP16 model card:
  https://huggingface.co/OpenVINO/whisper-large-v3-fp16-ov
- OpenVINO Whisper Large v3 INT4 model card:
  https://huggingface.co/OpenVINO/whisper-large-v3-int4-ov
- Hugging Face ASR evaluation overview:
  https://huggingface.co/learn/audio-course/chapter5/evaluation
- LibriSpeech dataset page:
  https://www.openslr.org/12
- TED-LIUM dataset page:
  https://huggingface.co/datasets/kfajdsl/tedlium
- OpenAI Whisper overview:
  https://openai.com/index/whisper/
- OpenAI Whisper paper:
  https://cdn.openai.com/papers/whisper.pdf

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:

- Artificial Intelligence
- 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.

This document, LP2448, was created or updated on June 11, 2026.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2448
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2448.

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®
ThinkSystem®

The following terms are trademarks of other companies:

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

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