Solution Brief

Lablup + Tenstorrent

Powerful Infra Management On Powerful AI Accelerator

Powerful Infra Management with Backend.AI® on Powerful Tenstorrent Wormhole™ AI Accelerator

Lablup Backend.AI links up with Tenstorrent Wormhole

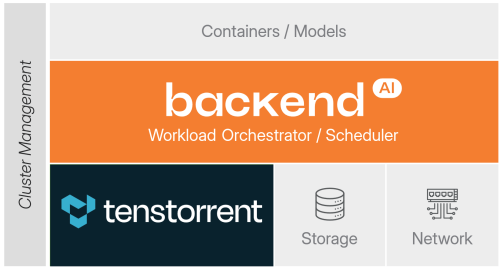
Meet the combination of Lablup Backend.AI® and Tenstorrent Wormhole™, enabling enterprises to seamlessly develop, deploy, manage, and scale generative AI solutions.

Why enterprises need a chip dedicated for AI

Modern AI workloads demand massive parallel processing, high memory bandwidth, and low latency that only dedicated AI chips can deliver. For enterprises, adopting these chips is not just about performance; it is essential for remaining competitive and driving innovation.

Businesses should focus on selecting AI solutions that align with their specific operational requirements. AI accelerators must strike the right balance between performance and efficiency to support targeted workloads, enabling organizations to deploy advanced models, reduce operational costs, and manage growing business complexity effectively.

Powerful software backing up the hardware is critical

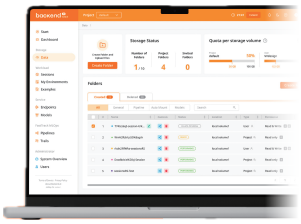


The full value of an AI accelerator can only be realized with a robust and intelligent orchestration layer. Such orchestrators allocate GPU resources to maximize utilization, prevent resource contention, and ensure fair scheduling across multiple users and projects. These platforms also enable organizations to efficiently manage multi-node, multi-tenant clusters and adapt to changing workload requirements in real time.

By doing so, they not only accelerate time-to-insight and model deployment but also reduce operational complexity and risk, empowering enterprises to innovate at scale.

Lablup Backend.AI®

Transforming GPU complexity into operational simplicity

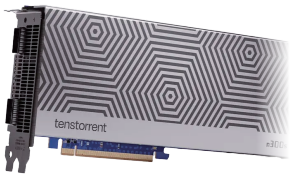


Backend.AI is an accelerated workload hosting platform designed for heterogeneous AI accelerators. Powered by our home-grown orchestration and job scheduling engine, Sokovan, it operates seamlessly across on-premises, cloud-native, hybrid, or fully air-gapped environments, efficiently scaling training and inference workloads across thousands of nodes and GPUs.

Built on the Sokovan core, Backend.AI can integrate into a customer's AI infrastructure as a container-based multi-node scheduler, a bare-metal resource manager, or a full-fledged control plane SDK to accelerate the time-to-market for customer's offerings including GPUaaS, cluster operation solutions, and more.

Tenstorrent Wormhole™

Flexible, scalable processors built with Tensix Cores™



Tenstorrent Wormhole™ ASIC delivers superior performance-per-cost compared to traditional GPUs. The Wormhole PCIe cards deliver a flexible and scalable processing solution, powered by advanced Tensix Cores™. Each card integrates a compute unit, network-on-chip, local cache, and "baby RISC-V" cores-enabling efficient, high-throughput data movement across the chip. Wormhole provides superior performance-to-cost efficiency compared to traditional GPUs, while supporting a wide range of data precision formats.

The builder of Wormhole, Tenstorrent is a full stack AI company, building AI computers, software, IP, silicon, systems and cloud eco-system. Tenstorrent's AI computers run inference, training and HPC models, open sourcing their software and license the IP (NPUs & RISC-V CPUs).

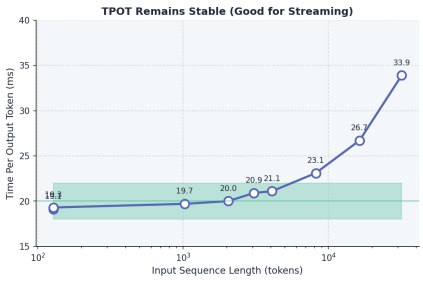
© 2025 Lablup Inc. All rights reserved. December17

Performance benchmark

We benchmarked Llama-3.1-8B-Instruct on Wormhole LoudBox. All benchmark data collected under a controlled environment. Installing Backend.AI on Wormhole LoudBox leverages the ultimate option to utilize from its Tensix Processors to storage, and its network connectivity.

Use case recommendations

Interactive chat, real-time assistants

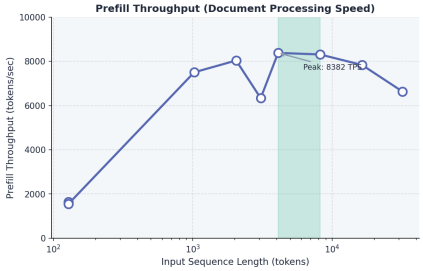


In interactive environments such as real-time chatbots, copilots, and customer support assistants, low TTFT is a key performance indicator since instant response is the top priority. Benchmarks show that with low concurrency (1–4), TTFT remains between 77 and 255 ms, while TPOT stays stable at around 19–20 ms. This enables users to perceive the initial response quickly and experience smooth, continuous text streaming.

Balanced setup:

- Low concurrency, moderate context (ISL ≤2K)
- Medium-length responses (OSL 128–512)

RAG, Document-based Q&As

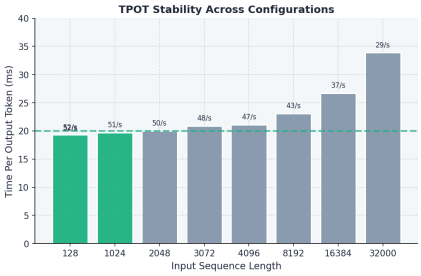


RAG-based document Q&A needs to balance long context handling with interactive responsiveness. In this type of workload, retrieved documents are included in the prompt, increasing the input sequence length (ISL) to around 4K–8K tokens. Benchmarks show that with concurrency set to 1, the time to first token (TTFT) remains within one second at 489–986 ms, and prefill throughput reaches 8,300–8,400 TPS. This indicates that Tenstorrent hardware performs efficiently during the prefill when processing long contexts.

Balanced setup:

- Limiting the output sequence length (OSL) to 128–256 tokens
- Stabilize responsiveness at low concurrency, then refine caching or retrieval to reduce context and improve user experience.

Long-form content generation



For workloads that generate long-form content with output sequence lengths (OSL) exceeding 1K tokens, perceived quality depends more on consistent generation speed and stability (TPOT) than on initial response time. Benchmarks show that with concurrency set to 1, TTFT remains between 78 and 137 ms, and TPOT stays steady at around 19–20 ms. The end-to-end time for producing 1,024 tokens is roughly 20 seconds.

Balanced setup:

- Better to maintain low concurrency to stabilize TPOT
- Combine policies like maximum output length and mid-generation saving for more predictable SLA design

Recommendation matrix

Use case	Concurrency	ISL	OSL	Primary Metric
Interactive Chat	1	≤ 2k	128-512	TTFT < 300ms
RAG / Q&A	1	4k-8k	≤ 256	TTFT <1s
Long-form Gen	1	≤ 1k	1k+	Stable TPOT

Overall Verdict: Broad LLM inference scalability from real-time to batch workloads

LoudBox, powered by the Tenstorrent Wormhole architecture, shows that Tenstorrent hardware can deliver both predictable low latency and high scalability across a wide range of inference workloads. In interactive scenarios with low concurrency, TTFT stays between 77–255 ms and TPOT around 19–20 ms, ensuring responsive user experiences for real-time chat or agent applications. Even in long-context workloads like RAG, it sustains high Prefill Throughput (8,300–8,400 TPS) within 4K–8K tokens while keeping TTFT under one second, highlighting its balanced performance for both context-intensive and streaming-based applications.

Start your journey to build AI services with Tenstorrent and Backend.AI.