SOLUTION BRIEF

Supermicro Sets New Standard: Industry-First, Independently Audited STAC-AI™ LANG6 LLM Inference Benchmarks on Red Hat® OpenShift®

Faster Inference. Greater Efficiency. Independently Proven Results on Red Hat OpenShift

SEPTEMBER 2026

Supermicro SYS-222C-TN and NVIDIA RTX PRO™ 6000

Table of Contents

Executive Summary 2

AI Inference Solution Architecture Overview 2

Configuration & Performance 4

NVIDIA RTX PRO 6000™ Blackwell Server Edition Overview 6

Red Hat OpenShift Platform Overview 7

Supermicro CloudDC Server Overview 7

Summary 8

For More Information 8

EXECUTIVE SUMMARY

Financial services organizations are rapidly adopting AI for document analysis, regulatory compliance, knowledge extraction, and retrieval-augmented generation (RAG) workflows built on real-world SEC EDGAR filings. These demanding LLM inference workloads require high throughput, low latency, enterprise-grade operational simplicity, and exceptional density within compact, power-efficient data center footprints.

The Supermicro CloudDC SuperServer SYS-222C-TN, configured with dual Intel® Xeon® 6730P processors and 2x NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs, delivers exceptional inference performance in a 2U form factor purpose-built for modern cloud data centers. In collaboration with NVIDIA and Red Hat, this platform achieved the industry's first independently audited STAC-AI LANG6 (Inference-Only) results on a fully containerized Kubernetes environment using Red Hat OpenShift 4.20.

The validated configuration demonstrates outstanding inference throughput, near-bare-metal GPU efficiency, and class-leading space efficiency — enabling financial institutions to accelerate AI-driven initiatives without compromising density, power, or operational control.

AI INFERENCE SOLUTION ARCHITECTURE OVERVIEW

The STAC-AI LANG6 benchmark focuses on the sixth step of a RAG (Retrieval-Augmented Generation) pipeline — LLM inference. The benchmark uses real-world SEC EDGAR 10-K filing datasets (EDGAR4 for targeted summarization and EDGAR5 for full-filing, multi-question analysis) and measures infrastructure performance in both batch and interactive request modes. This scope explains the benchmark name: STAC-AI LANG6 (Inference-Only).

Solution Architecture Layers

The tested solution is organized across three primary layers:

- Hardware Layer — Supermicro SYS-222C-TN with 2x NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs, dual Intel Xeon 6730P, and 2 TiB DDR5 memory
- Platform Layer — Red Hat OpenShift Container Platform 4.20 (Single Node OpenShift) on Red Hat Enterprise Linux CoreOS 9.6, providing container orchestration, GPU lifecycle management, and enterprise security
- Inference Layer — NVIDIA TensorRT-LLM 1.2.0rc2 with NVFP4 quantization, automated end-to-end by the Red Hat STAC-AI Kubernetes Operator.

Benchmark Request Modes

STAC-AI supports two inference request modes representing the most common production use cases:

- Batch Mode — all inference requests are submitted in a single API call; results returned together. Measures maximum achievable throughput.
- Interactive Mode — requests arrive following a Poisson arrival process (mean rate λ, measured in inferences/second), simulating real-world concurrent user traffic. Measures latency distributions (Reaction Time and Response Time) alongside throughput at each arrival rate.

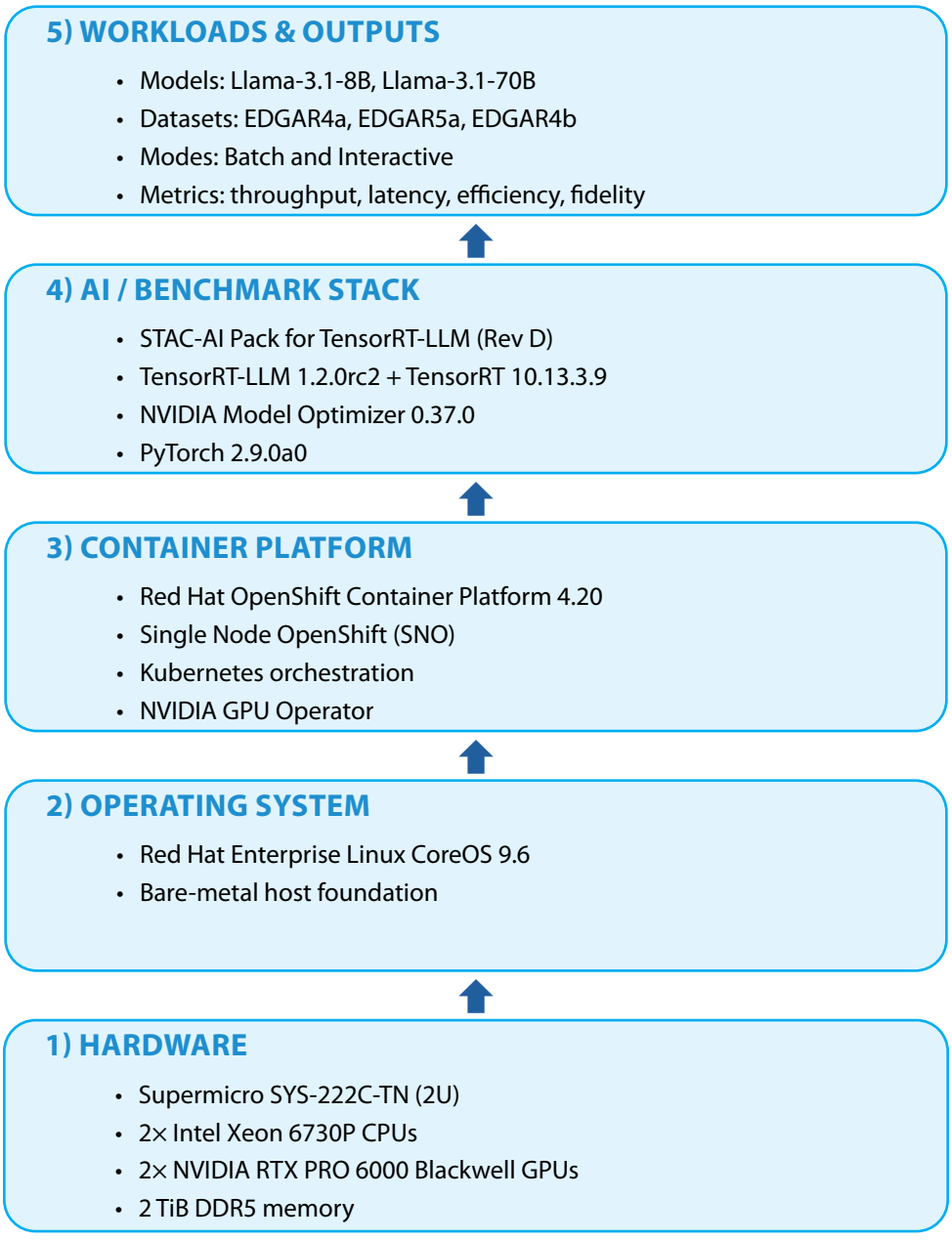
Models and Datasets

Two open-source Llama models were benchmarked against two financial-domain datasets:

- Llama-3.1-8B-Instruct (8.03B parameters, BF16 reference, NVFP4 quantized) with EDGAR4a and EDGAR5a datasets
- Llama-3.1-70B-Instruct (70.6B parameters, BF16 reference, NVFP4 quantized) with EDGAR4b dataset
- EDGAR4 — summarization of a company's relationship to physical/financial concepts from 10-K paragraphs (median prompt: ~1,197 words; median response: ~167 words)
- EDGAR5 — multi-question analysis of a complete 10-K filing (median prompt: ~44,029 words; median response: ~370 words)

Technical Architecture

Hardware: Red Hat Enterprise Linux® CoreOS 9.6, OpenShift 4.20, STAC-AI LANG6 Workloads



CONFIGURATION AND PERFORMANCE

Validated Configuration:

Component	Specification
Server	Supermicro CloudDC SuperServer SYS-222C-TN (2U, DC-MHS Compliant)
Processors	2x Intel Xeon 6730P
Memory	32x 64 GiB ECC DDR5 RDIMM @ 5200 MT/s (2 TiB total)
GPUs	2x NVIDIA RTX PRO 6000 Blackwell Server Edition — 96 GiB GDDR7 ECC each
Storage	8 hot-swap 2.5" NVMe/SATA/SAS bays + dual M.2 PCIe 5.0
Networking	6x PCIe 5.0 x16/x8 FHFL slots + 2x PCIe 5.0 x16 AIOM slots
Power	Redundant 3200W Titanium-level PSUs
OS / Platform	Red Hat Enterprise Linux CoreOS 9.6 + OpenShift Container Platform 4.20 (SNO)
Inference SW	NVIDIA TensorRT-LLM 1.2.0rc2 with NVFP4 Quantization, TensorRT 10.13.3.9
Benchmark Pack	STAC-AI LANG6 (Inference-Only) — Pack for TensorRT-LLM Rev D

STAC-AI LANG6 Benchmark Results

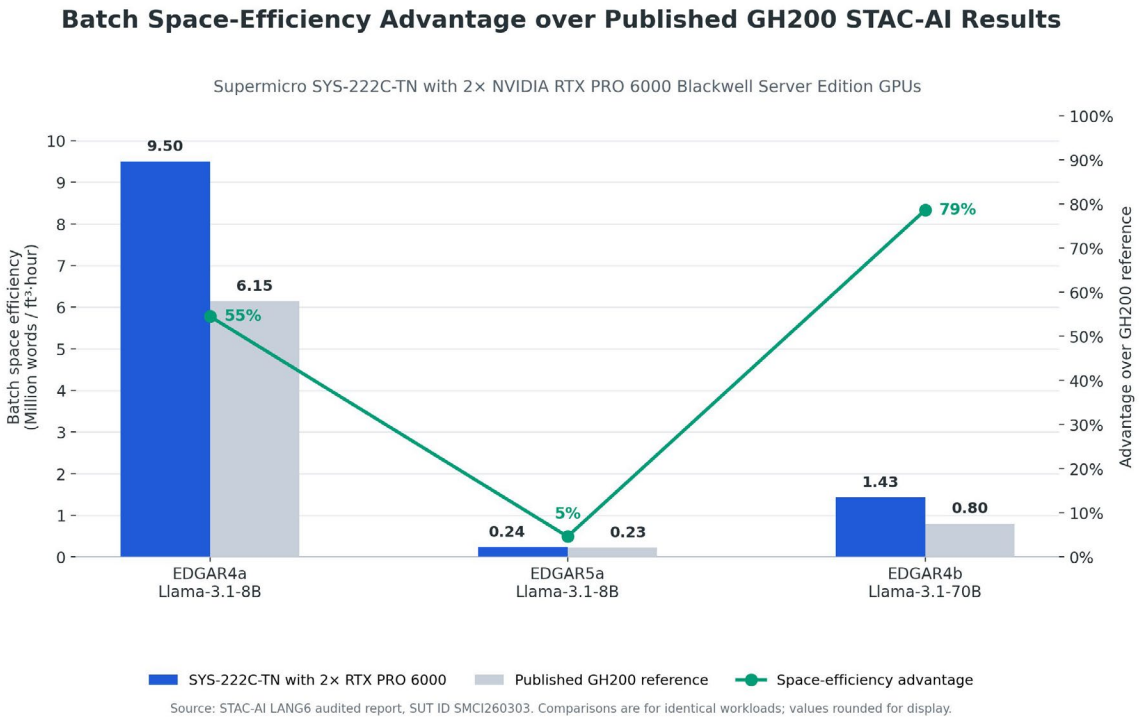
Independent audited benchmarks using the STAC-AI Pack for TensorRT-LLM (Rev D) on real-world EDGAR financial datasets. Interactive-mode values shown as: 95p Reaction Time | 95p Response Time.

Workload	Batch Mode	Interactive Mode (4x arrival rate)	Space Efficiency
Model / Dataset	**Inference Rate / Throughput**	**Peak λ | 95p Reaction | 95p Response**	**words / ft³·hour**
EDGAR4a Llama-3.1-8B	32.9 infs/s | 5,549 words/s	30.0 infs/s (91% of batch) | 0.320 s | 14.6 s	9.501M
EDGAR5a Llama-3.1-8B (long-context)	0.345 infs/s | 139 words/s	0.320 infs/s (93% of batch) | 29.1 s | 126 s	237.4K
EDGAR4b Llama-3.1-70B	5.28 infs/s | 834 words/s	5.00 infs/s (95% of batch) | 2.26 s | 44.8 s	1.428M

Fidelity scores: 89.55% (EDGAR4a) · 89.48% (EDGAR5a) · 91.29% (EDGAR4b). Word Count Ratio within 98–106% of Fidelity Reference across all workloads. All models quantized at NVFP4.

Space Efficiency vs. GH200 Reference

The 2U SYS-222C-TN achieves industry-leading Space Efficiency (words/ft^3·hour) compared to the GH200 STAC-AI reference results on identical workloads:

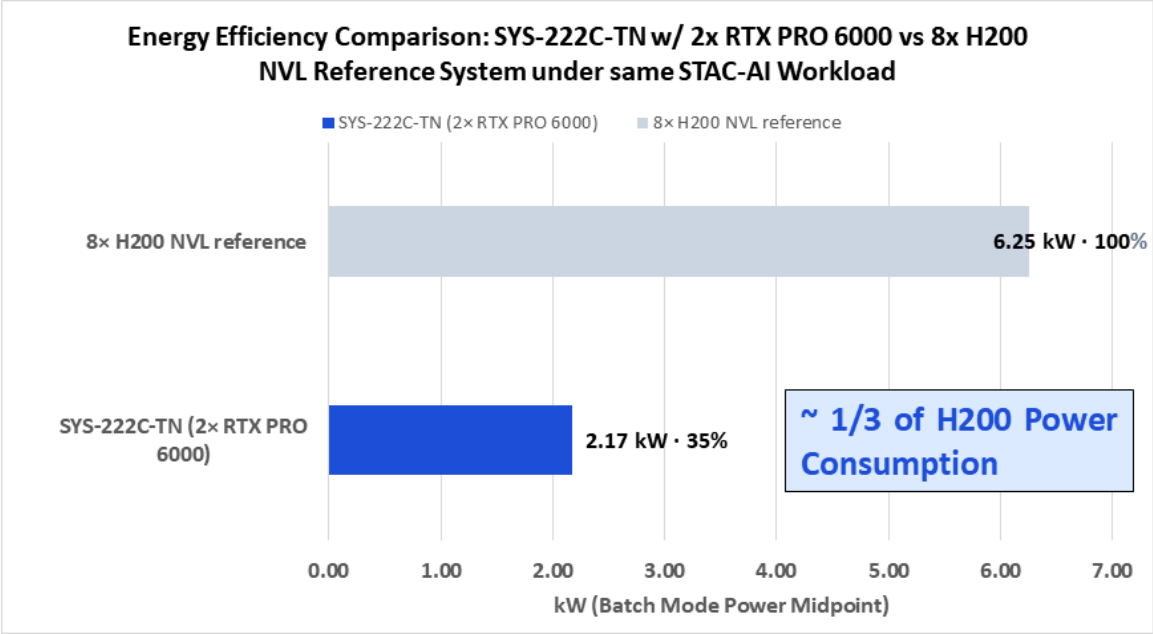


The GH200 reference results are sourced from publicly available STAC-AI benchmark submissions. Comparisons apply to identical workloads only. Differences in quantization and Word Count Ratio should be considered per STAC guidelines.

Energy Efficiency

Batch mode setup power draw averaged 2.13–2.21 kW across all workloads. Resulting energy efficiency:

- EDGAR4a (Llama-3.1-8B): 9.320 M words/kWh
- EDGAR5a (Llama-3.1-8B): 234.6 K words/kWh
- EDGAR4b (Llama-3.1-70B): 1.358 M words/kWh



Reference Comparison: For context, a previous 8× H200 NVL reference system consumed approximately 6.1–6.4 kW under the same STAC-AI workloads and methodology. The SYS-222C-TN therefore delivers strong energy efficiency at roughly one-third the power draw of that reference configuration — a major power advantage and cooling-constrained data centers.

NVIDIA RTX PRO 6000 BLACKWELL SERVER EDITION OVERVIEW

The NVIDIA RTX PRO 6000 Blackwell Server Edition brings professional-grade AI acceleration to enterprise servers. The STAC-AI Pack for TensorRT-LLM, developed by NVIDIA, is a comprehensive benchmarking suite designed to evaluate performance, efficiency, and scalability of AI inference workloads — particularly large language models — on modern accelerator architectures.

Key GPU Specifications:

- 96 GiB high-bandwidth GDDR7 ECC memory per card — enabling full Llama-3.1-70B NVFP4 inference across just two GPUs
- Blackwell Tensor Cores with Transformer Engine optimizations purpose-built for LLM inference
- NVFP4 (4-bit floating-point) precision support via NVIDIA Model Optimizer (nvidia-modelopt 0.37.0)
- PCIe Gen 5 host interface delivering full bandwidth on the SYS-222C-TN's DC-MHS interconnect fabric

Software Stack:

- NVIDIA TensorRT-LLM 1.2.0rc2 with PyTorch backend — high-performance deep learning inference library optimized for transformer-based models
- NVIDIA TensorRT 10.13.3.9
- NVIDIA Model Optimizer 0.37.0 for NVFP4 quantization
- PyTorch 2.9.0a0 (NVIDIA PyTorch container 25.10)

By standardizing test methodologies across hardware and software stacks, the STAC-AI Pack for TensorRT-LLM delivers a trusted framework for evaluating GPU and CPU configurations, memory architectures, and containerized AI deployment environments — helping organizations identify optimal system designs for production-grade AI inference.

RED HAT OPENSHIFT PLATFORM OVERVIEW

The SUT ran on Red Hat OpenShift Container Platform 4.20, deployed as Single Node OpenShift (SNO) on Red Hat Enterprise Linux CoreOS 9.6. This architecture delivered bare-metal-equivalent performance while enabling full Kubernetes operations — consistent with prior audited STAC benchmark submissions (STAC-N1 and STAC-A2) from Red Hat.

- Container orchestration and lifecycle management — the container and orchestration layers introduced no material performance limitations
- Red Hat STAC-AI Operator automates the full benchmark lifecycle: workspace provisioning, model quantization, benchmark execution, and results collection
- NVIDIA GPU Operator and Node Feature Discovery (NFD) for streamlined GPU resource management within OpenShift
- Unified deployment model applicable equally to workstation-class edge GPUs and datacenter-class GPU clusters
- Enterprise security: RBAC, audit logging, network policies, and multi-tenancy out of the box

SUPERMICRO CLOUDDC SERVER OVERVIEW

The Supermicro CloudDC SuperServer SYS-222C-TN delivers high-performance computing, empowering organizations to validate AI inference and analytics workloads with speed and efficiency. Built on a robust 2U rackmount platform with DC-MHS compliance, it features dual Intel Xeon 6700/6500 series processors, massive memory scalability, and high-bandwidth PCIe Gen 5 storage and I/O.

SYS-222C-TN Key Specifications

- Dual Socket E2 (LGA-4710): Intel Xeon 6700/6500 series — up to 86 cores/172 threads (P-cores) or 144 cores/144 threads (E-cores)
- Up to 32 DIMM slots — up to 4 TB ECC DDR5-6400 for large model serving and multi-tenant workloads
- Up to 24 hot-swap 2.5" NVMe/SATA/SAS bays + dual M.2 PCIe 5.0 slots for high-bandwidth storage
- Up to 6x PCIe 5.0 x16 FHFL slots + 2x PCIe 5.0 x16 AIOM slots — supports NVIDIA H200 NVL, L40S, and Blackwell Server Editions
- Up to 3200W Titanium-level PSUs with OpenBMC-based management and full server automation
- DC-MHS compliant — simplifies CSP deployments, serviceability, and supply chain
- Server management flexibility with OpenBMC

Fully integrated, pre-tested, and ready for rapid deployment, the SYS-222C-TN minimizes time-to-value for AI infrastructure initiatives from edge deployments to large-scale data centers.

SUMMARY

The Supermicro CloudDC SuperServer SYS-222C-TN configured with 2x NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs (SMCI260303) represents the optimal balance of performance, density, efficiency, and enterprise readiness for today's AI inference demands in financial services.

The independently audited STAC-AI LANG6 results confirm its leadership in high-throughput LLM serving on containerized Kubernetes platforms — achieving class-leading space efficiency that exceeds the GH200 reference by up to 79% while running Llama-3.1-70B with just two GPUs in a 2U chassis. The system maintained 91–95% of batch-mode throughput under realistic interactive load conditions across all tested workloads.

Organizations can deploy this solution today to accelerate AI workloads, reduce infrastructure footprint, and achieve superior performance-per-watt. Contact Supermicro for fully integrated, racked-and-stacked solutions tailored to your environment.

FOR MORE INFORMATION

Supermicro CloudDC System:
https://www.supermicro.com/en/products/system/clouddc/2u/sys-222c-tn

STAC AI LANG 6 report:
https://www.supermicro.com/thought-leadership/STAC-AI-Audited-Report-SMCI260303.pdf

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements.

www.supermicro.com

RED HAT

Red Hat is the world's leading provider of enterprise open source software solutions, using a community-powered approach to deliver reliable and high-performing cloud, developer, AI, Linux, automation and application platform technologies. Red Hat helps customers develop cloud-native applications, integrate existing and new IT applications, and automate and manage complex environments. A trusted adviser to the Fortune 500, Red Hat provides award-winning support, training, and consulting services that bring the benefits of open innovation to any industry. Red Hat is a connective hub in a global network of enterprises, partners, and communities, helping organizations grow, transform, and prepare for the digital future.

Visit: www.redhat.com

©Super Micro Computer, Inc. Specifications subject to change without notice. All other brands and names are property of their respective owners. All logos, brand names, campaign statements and product images contained herein are copyrighted and may not be reprinted and/or reproduced, in whole or in part, without express written permission by Supermicro Corporate Marketing.

Printed in USA ♻ Please Recycle STAC-Solution-Brief-2026-08-R07 SEPTEMBER 2026