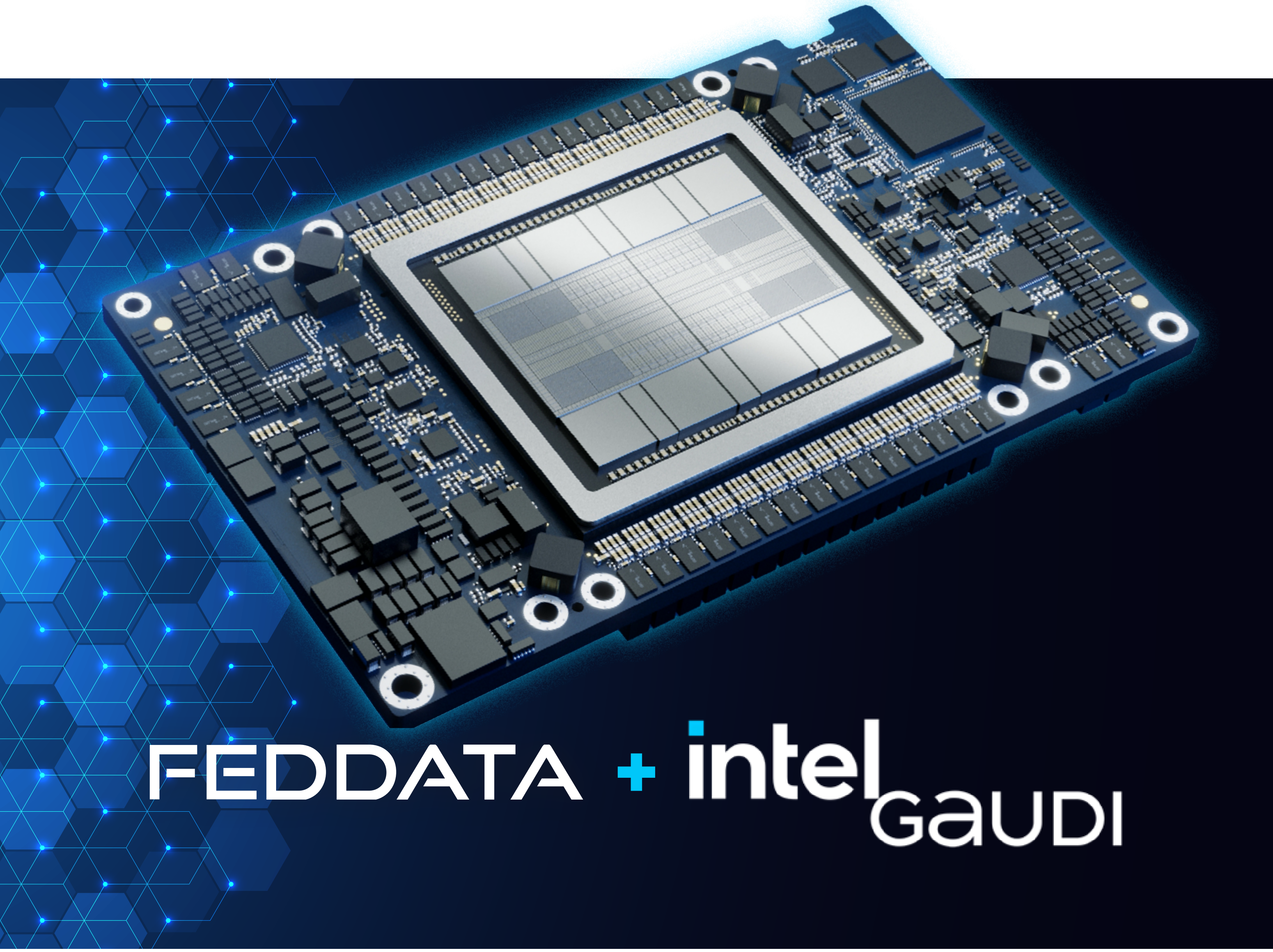


SOLUTION BRIEF AND PERFORMANCE ANALYSIS

CONTRIBUTORS:

Intel
Shreejan Mistry
Shawna Meyer-Ravelli
Taylor Mcnearney
Rabail Abbas

FedData
Dr. Gianluca Longoni
Jerry Whitacre

Contents

INTRODUCTION

Reshaping AI Strategy: Smarter Infrastructure for Enterprise & Government

Artificial intelligence is central to digital transformation for businesses and government agencies. Whether the goal is operational efficiency, enhanced public services, or national security, the urgency to deploy AI solutions is growing rapidly.

But IT leaders across sectors face the same hard truth: the path to scalable AI is blocked by limited infrastructure, high costs, and constrained power and space.

That's why leading organizations are taking a more sustainable and cost-effective **approach: building with Intel® Xeon® CPUs and scaling AI capabilities with Intel® Gaudi® AI accelerators**—instead of jumping straight into high-cost GPU deployments. Using AI accelerators like Intel® Gaudi® provides significant benefits over porting applications to general-purpose GPUs. AI accelerators are purpose-built to optimize deep learning performance, offering better efficiency, lower power consumption, and improved scalability across diverse workloads. Unlike traditional GPU porting—which can require complex rewrites and tuning—AI accelerators like Intel Gaudi are designed with native support for popular frameworks such as PyTorch, reducing development overhead. Intel has demonstrated strong interoperability with PyTorch, delivering optimized performance through open software stacks like the Intel® Extension for PyTorch. This seamless integration enables developers to accelerate AI training and inference while maintaining flexibility across CPU and accelerator resources, streamlining infrastructure development for enterprise and cloud-scale AI deployments.

The Challenge: Modern AI, Legacy Infrastructure

Government agencies and enterprises alike are facing unprecedented demand to modernize:

- **Federal mandates for AI-driven services** require faster data processing and real-time analytics.
- **AI governance and transparency** place a premium on open, auditable systems.
- **Budgets and procurement cycles** limit the ability to overspend on infrastructure.
- **Legacy data centers** weren't built for dense, high-power GPU clusters.
- **Sustainability requirements** are rising, especially in public sector operations.

Intel + FedData: An AI Infrastructure Built for Government Needs

To address these challenges, Int**el® and FedData** have partnered to deliver an alternative: an open, cost-effective AI infrastructure powered by **Intel® Gaudi® AI accelerators**, supported by FedData's proven government integration capabilities. This infrastructure enables agencies to deploy, scale, and sustain mission-ready AI solutions with security, flexibility, and performance at the core. Instead of diving headfirst into high-cost, high-risk GPU infrastructure, FedData and Intel are building a smarter path with better ROI.

SOLUTION OVERVIEW

As artificial intelligence moves from experimentation to operational scale, U.S. government agencies are under growing pressure to deploy AI that is not only powerful and secure but also cost-effective, sustainable, and compliant with federal mandates. Whether supporting real-time intelligence, automating document workflows, enhancing citizen engagement, or accelerating scientific research, the need for scalable and efficient AI infrastructure is clear. However, traditional GPU-based solutions—while performant—come with significant drawbacks for public-sector environments: high acquisition costs, excessive power consumption, limited supply availability, and often proprietary software stacks that restrict long-term flexibility.

Intel® Gaudi® 3 Accelerator

Intel Gaudi 3 AI accelerators offer a purpose-built strategic alternative to meet the technical and operational demands of government AI workloads. Technically, Intel Gaudi 3 delivers impressive performance for large language model (LLM) training and inference, outperforming or matching the industry-standard Nvidia H100 in many benchmarks. It supports up to 128 GB of HBM2e high-bandwidth memory with 3.7 TB/s of bandwidth, ensuring rapid access to large model parameters. With up to 8 accelerators per server node and an open, Ethernet-based RoCE (RDMA over Converged Ethernet) interconnect fabric offering 24 x 200 Gbps ports per chip, Intel Gaudi 3 scales easily without the need for proprietary switches or fabric, reducing both hardware cost and complexity.[1]

For federal IT leaders and mission owners, the business value of Intel Gaudi 3 is equally compelling. Signal65's 2025 benchmarking shows Intel Gaudi 3 **delivers up to 2.5x better price-performance than Nvidia H100 on LLMinference workloads**[2], enabling significant cost savings across large-scale deployments. This is particularly important in public-sector environments where budget constraints, long procurement cycles, and lifecycle sustainment are major operational considerations. Furthermore, Intel Gaudi 3's architecture is optimized for energy efficiency—an essential benefit for agencies working to meet executive sustainability directives or operating within legacy data centers with limited power and cooling headroom.

The software ecosystem is a key strength of Intel Gaudi 3, especially for developers who have already built Python applications optimized for CUDA and NVIDIA GPUs. With Intel Gaudi 3's native support for open frameworks like PyTorch and seamless integration with Hugging Face's Optimum library, migrating an existing model often requires only changing the device ID in the initial PyTorch call where the model is instantiated. This ease of transition protects development investments while enabling broader hardware flexibility. Intel Gaudi 3 also provides an open developer SDK that promotes interoperability, transparency, and long-term control over AI pipelines—an essential advantage for government and regulated sectors that demand traceability, auditability, and compliance with evolving AI governance standards. Paired with Intel's hardware-based security features, including Intel® Software Guard Extensions (Intel® SGX) and Intel® Total Memory Encryption (Intel® TME), Intel Gaudi platforms are ideally suited for high-security, mission-critical environments such as classified or air-gapped deployments via trusted partners like FedData.

Architecture and Configuration

The benchmark was executed on a high-performance server platform equipped with **Intel Gaudi 3 AI accelerators**, optimized to reflect a production-realistic configuration for enterprise-scale generative AI serving. A single node was provisioned with **8x Intel Gaudi 3 accelerators**, interconnected via Intel Gaudi's native **RoCE-based 200 Gbps Ethernet fabric**, enabling linear scaling without reliance on proprietary interconnects like NVLink or InfiniBand.

System Stack Overview:

Accelerators:	8× Intel Gaudi 3 (8 x HL-325L)
CPU:	Dual-socket Intel® Xeon® 6980P (6th Generation) processors, 256-core configuration – hyperthreaded (2 threads per core), 3.9 GHz Frequency
Memory:	2.3 TB DDR5 RAM
Interconnect:	24× 200 Gbps Ethernet per Intel® Gaudi® 3 via integrated NICs
Networking:	Top-of-rack Arista switch, RoCEv2 optimized, jumbo frames enabled
Storage:	High-throughput NVMe RAID array for model loading
OS/Drivers:	Ubuntu 22.04 LTS
Intel Gaudi Software Version:	1.19.2
Intel Gaudi Driver Version:	1.19.2-ff37fea
Intel Gaudi SPI Firmware Version:	1.19.2-fw-57.2.4
Serving Framework:	Habana vLLM Fork (v0.6.4.post2+ Intel Gaudi-1.19.2)
Model:	mistralai/Mistral-7B-Instruct-v0.3, meta-llama/Llama-3.1-8B, meta-llama/Llama-3.3-70B-Instruct
System Part Number:	SuperServer SYS-822GA-NGR3

The system was configured to stress **concurrent inference throughput**, simulating typical use cases like RAG-based chat systems or multi-user LLM inference services. Intel Gaudi 3 architecture allowed us to scale horizontally across accelerators, taking advantage of its distributed memory and compute pipeline for low-latency token generation.

Benchmarking Methodology

To validate the suitability of Intel Gaudi 3 AI accelerators for real-world government AI workloads, FedData used a highly controlled environment and mission-relevant testing methodology. The primary objective was to assess performance, scalability, and power efficiency under the types of concurrent-query workloads typical in generative AI deployments, such as those supporting Retrieval-Augmented Generation (RAG), secure chat interfaces, and multi-session LLM agents.

The test environment consisted of a full-stack AI server configured with Intel Gaudi 3 accelerators, using firmware and drivers optimized in close coordination with Intel. We leveraged an Intel optimized fork of the vLLM inference framework, modified to run efficiently on Intel Gaudi 3. We executed the benchmark with **LLAMA 3 8B, Mistral & LLAMA 70B** - a few of the most widely adopted open-weight LLMs. Our workload simulated a realistic inference pressure scenario with multiple concurrent queries, designed to simulate real-time production-ready conditions encountered in secure federal environments, such as chatbots for citizen service, multi-agent collaboration for mission planning, and tactical document summarization at the edge.

The goal of the benchmarks was to capture a holistic set of performance and operational metrics. The study focused on metrics that directly affect operational effectiveness in federal deployments:

Metric	What It Means for Gov / DoD Operations
Queries per Second (QPS)	How many simultaneous chats or requests one node can handle.
Time to First Token (TTFT)	Delay before the first word appears — key to snappy, real time interactions.
Time Per Output Token (TPOT)	Token generation speed per user; faster feels smoother.
Throughput (Tokens/sec)	Aggregate tokens generated per second across all users — overall capacity.
Performance per Watt (Perf/W)	Useful work per watt — critical in power and cooling limited sites.
Thermal Profile	Heat output — determines suitability for sealed or rugged enclosures
Accelerator Utilization	How busy the Gaudi chips are —guides capacity planning and spend.

Benchmark Specifications

The benchmarks simulate varying levels of concurrent user requests, effectively functioning as a stress test for AI systems. These benchmarks evaluate key performance metrics—including latency, throughput, and system stability—under different load conditions to reflect the dynamic nature of real deployment scenarios. By focusing on workloads such as chatbots and retrieval-augmented generation (RAG) services, the benchmarks ensure relevance to common use cases where responsiveness and scalability are critical. This approach provides a robust framework for assessing how well a system can maintain performance as demand scales, offering insights into both immediate response quality and long-term operational resilience.

Benchmark Specifications Used	
vLLM Version:	**0.6.3.dev2224 + g3457022d1**, pulled from the Habana maintained fork https://github.com/HabanaAI/vllm-fork
Benchmark Scripts:	**Serving Latency** – benchmark_serving.py **Throughput** – benchmark_throughput.py **Example Command** – vLLM Serving Benchmark (benchmark_serving.py) : • python3 benchmark_serving.py --num-prompts <#prompts> --dataset ./ShareGPT_V3_unfiltered_cleaned_split.json --port 9050 --model <#cached LLM> • #prompts = [50, 100, 200, 500, 1000] • #cached LLM = [Mistral-7B-Instruct-v0.3, Llama-3.1- 8B, Llama-3.3-70B-Instruct] • Dataset: ShareGPT_V3_unfiltered_cleaned_split, about 53K chat conversations

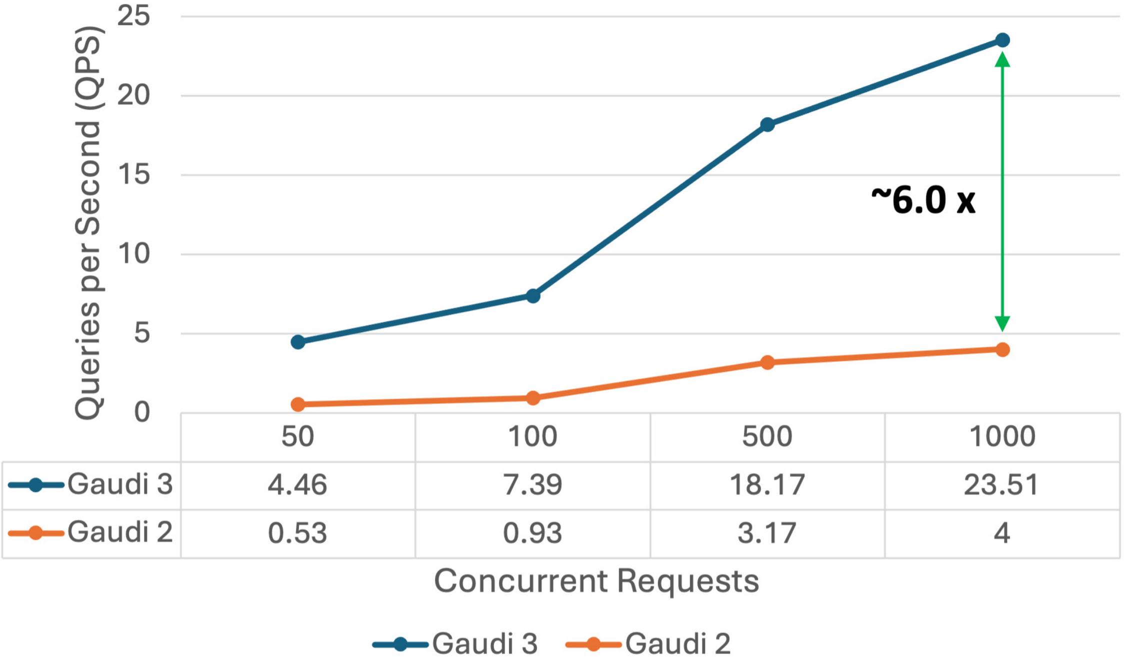
Results & Analysis

Intel Gaudi 3 vs Intel Gaudi 2

Our results showed Intel Gaudi 3 demonstrated **superior scalability and responsiveness under concurrent load**, outperforming previous-generation Intel Gaudi 2 accelerators with an improvement of approximately **6x more queries per second (QPS)**. (See Chart 1) and an approximate **6x average speed-up** in time per output token (See Chart 2).

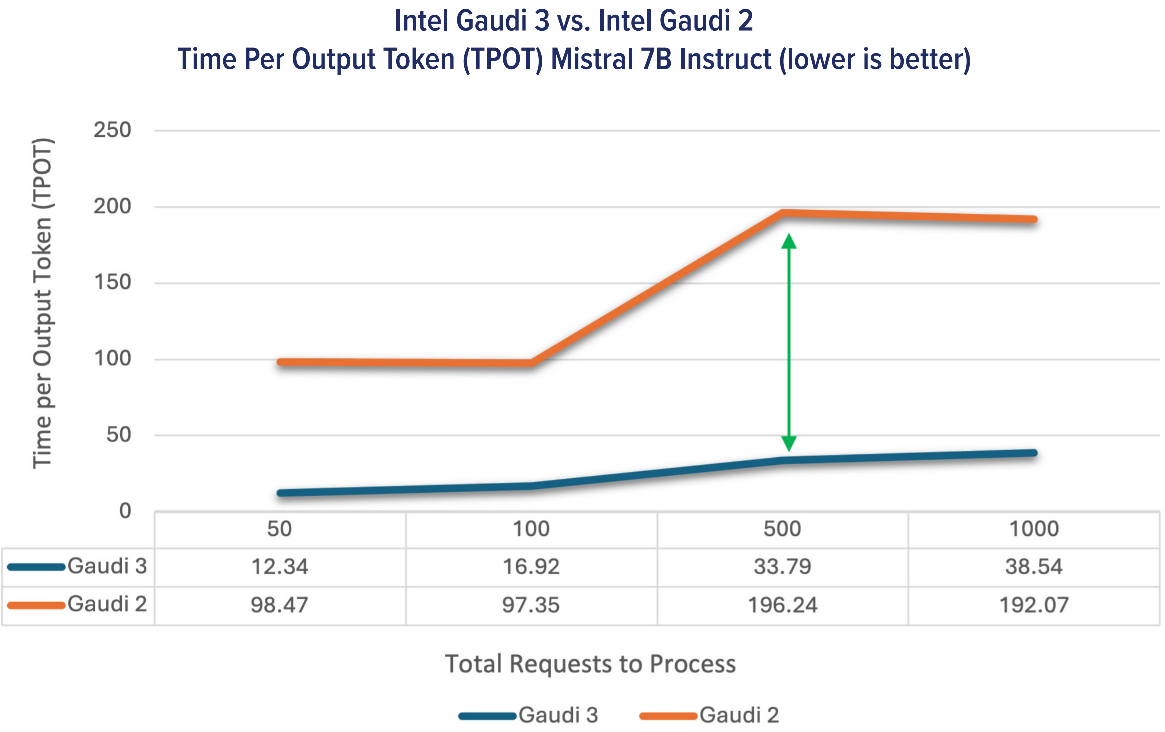
Using the Mistral 7B v0.3 Instruct model benchmark under a 1,000-concurrent-request load reveals a compelling value proposition for customers adopting Intel Gaudi 3 AI accelerators. By measuring the number of requests served per second relative to average power draw, Intel Gaudi 3 **delivers a 6x better performance uplift over Intel Gaudi 2, and it does so with 2x the power efficiency**.[3]

Intel Gaudi 3 vs. Intel Gaudi 2
Queries Per Second (QPS) Mistral 7B Instruct (higher is better)



	50	100	500	1000
Gaudi 3	4.46	7.39	18.17	23.51
Gaudi 2	0.53	0.93	3.17	4

Chart 1: Intel Gaudi 3 Gen over Gen Comparison – Queries per Second (QPS)



	50	100	500	1000
Gaudi 3	12.34	16.92	33.79	38.54
Gaudi 2	98.47	97.35	196.24	192.07

Chart 2: Intel Gaudi 3 vs. Intel Gaudi 2 Gen over Gen Comparison Time per Output Token (TPOT)

Intel Gaudi 3 vs. NVIDIA H100

Under a 1,000-concurrency load—typical of production-grade RAG Pipeline and high-traffic chatbots—Gaudi 3 sustains ~10.4 k tokens/sec, only 5 % shy of NVIDIA's single-H100 result (~10.9 k tps) (See Chart 3). Because Habana's architecture scales almost linearly with concurrency, engineers can meet identical service-level targets **while operating in a smaller power envelope and at a materially lower $/tps**[4]. In practice, that means:

- **Higher density per rack:** lower board power lets you deploy more accelerators before hitting facility power or cooling ceilings.
- **Reduced run-rate OPEX:** fewer watts translate directly into smaller energy and cooling bills, especially significant for 24 × 7 conversational workloads.
- **Improved TCO:** when both CapEx and OpEx are normalized, Gaudi 3 delivers superior through-put-per-dollar across the system's life cycle.

For organizations scaling large-context RAG or always-on conversational AI, the marginal 5 % performance delta is outweighed by Gaudi 3's efficiency gains, offering a more economical and greener path to high-concurrency inference.

Llama 3.1 8B Inference Throughout vs. Concurrency Single Accelerator

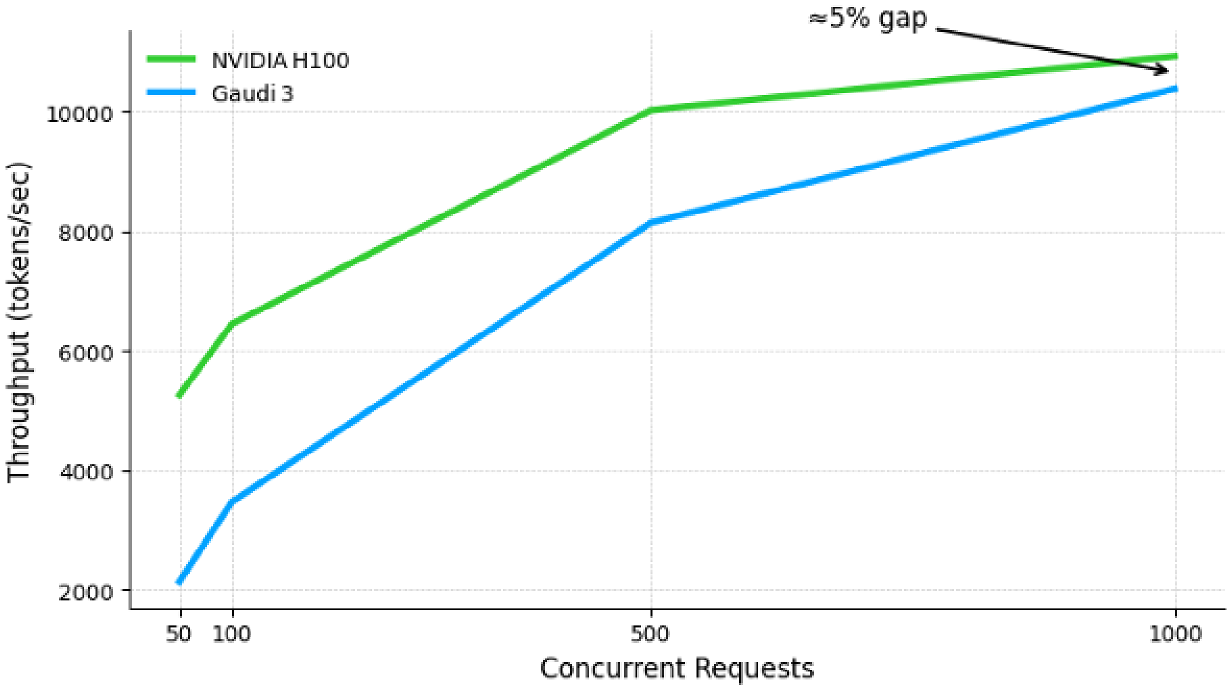


Chart 3: Llama 3.1 8B Inference Throughput vs. Concurrency Single Accelerator.

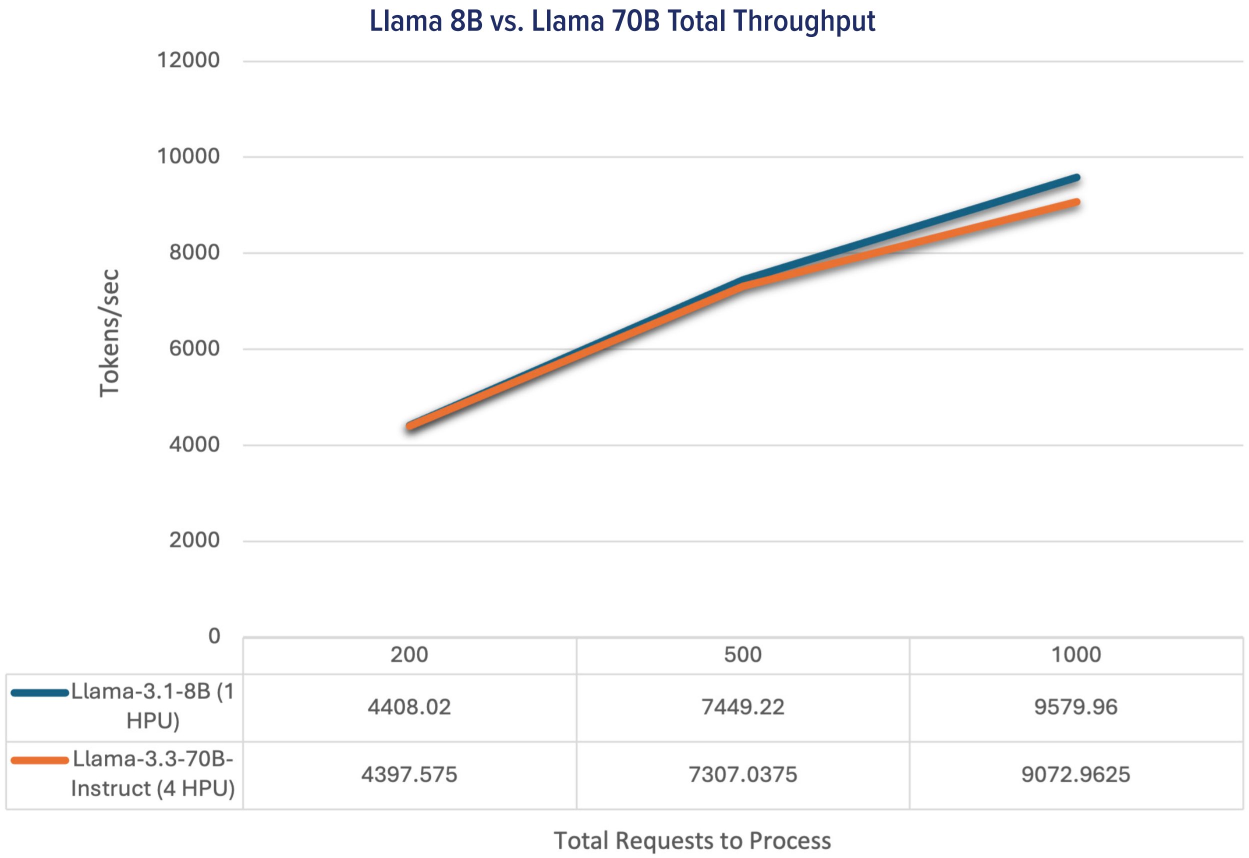
Scalability on Large Models

4.4 LargeModel Scaling & Efficiency (Llama3.370B on 4 HPUs)

To test Gaudi 3 on larger LLMs, we sharded **Llama3.370BInstruct** across **four HPUs** and compared its serving throughput to **Llama3.18B** on a single HPU. We observed an operational benefit in that agencies can deploy richer 70 B-class models without a large capacity penalty, keeping rack density and power budgets in check.

Key Takeaways on Scalability:

- **Nearlinear scaling:** Four HPUs deliver almost the same raw throughput on a LLAMA 3.3 70B model as one HPU on an LLAMA 3.1 8B model—indicating that Intel Gaudi 3's ROCEv2 interconnect and compiler stack scale efficiently. (Chart 4)
- **Perparameter efficiency:** Despite having ≈ 8.75 × more parameters, the 70 B model achieves ≈ 95 % of the LLAMA 3.1 8B model's perparameter throughput across varying concurrent prompts. (Chart 4)



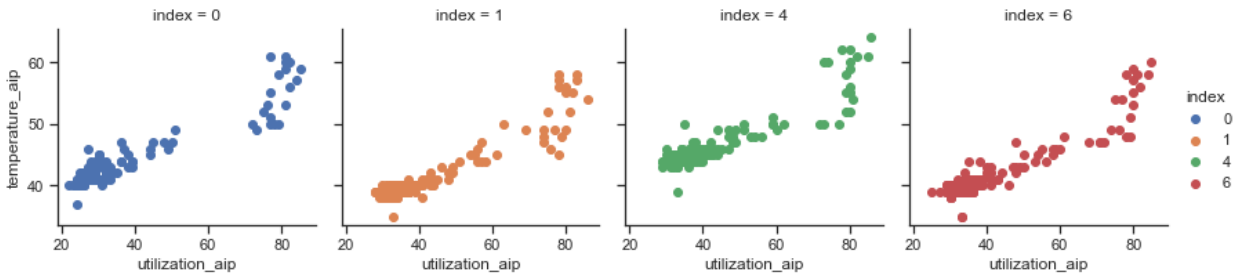
	200	500	1000
Llama-3.1-8B (1 HPU)	4408.02	7449.22	9579.96
Llama-3.3-70B-Instruct (4 HPU)	4397.575	7307.0375	9072.9625

Chart 4: Llama 8B vs. Llama 70B Total Throughput in Tokens/Second.

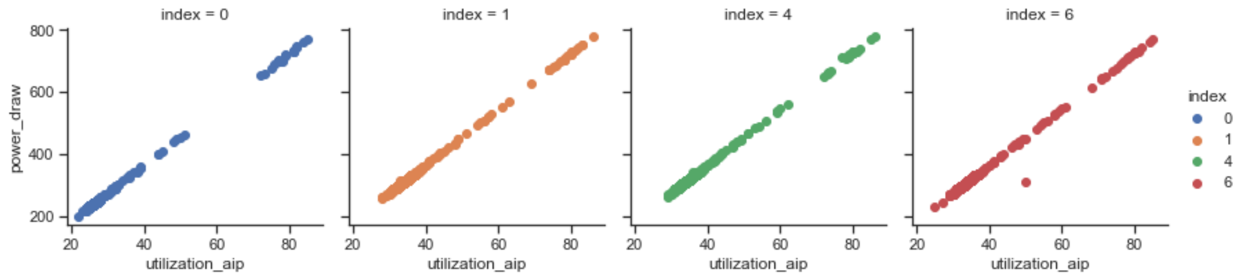
Key Takeaways for Power-Conscious Enterprises:

- **Thermal stability you can trust:** Even at > 90 % HPU utilization, every Gaudi HPU stays locked in the 60 – 65 °C range. Uniform temperatures mean balanced cooling, no throttling, and longer component life.
- **Truly linear power draw:** Power consumption scales in step with workload. Push the card to 20 % and it draws ~20 % of peak watts; push to 90 % and power rises proportionally—no hidden overhead, no hotspots, just predictable usage.
- **Only pay for what you use:** Because energy scales linearly, you're never charged (in utility dollars or data-center capacity) for idle silicon. Whether you're bursting a single workload or driving the cluster at full throttle, the platform matches your power budget in real time.
- **Scalable, efficient performance:** The combination of tight thermal control and proportional power lets you densify racks without overcooling, meet sustainability goals, and reinvest saved watts into more compute—turning power efficiency into a competitive advantage.

Temperature (C) vs. HPU Utilization (%) for 4-HPU LLM Inferencing Case



Power Draw (W) vs. HPU Utilization (%) for 4-HPU LLM Inferencing Case



For enterprises and government agencies managing constrained data center resources, this translates directly to lower operating costs, reduced cooling demand, and increased compute density. Intel Gaudi 3's performance gains are driven by architectural improvements that allow more work to be completed per watt, making it ideal for scalable, energy-aware AI deployments. Whether optimizing for TCO or expanding capacity within limited power envelopes, Intel Gaudi 3 offers a future-proof solution tailored to high-demand inference and generative AI workloads.

Summary

This study affirms Intel Gaudi 3 as a highly capable, power-efficient, and mission-aligned AI accelerator. It offers a scalable alternative to GPU-based systems while maintaining openness, control, and cost predictability—essential for public-sector AI deployments.

Intel and FedData bring a unique value proposition to government AI modernization efforts. Intel provides an open, scalable, high-performance hardware foundation, while FedData delivers secure deployment expertise, compliance alignment, and lifecycle support tailored to federal missions. Whether agencies are just starting to build AI capabilities using existing Intel® Xeon® infrastructure or scaling to large inferencing clusters in support of next-gen analytics, the Intel Gaudi 3 platform provides a future-ready, economically rational, and secure foundation for public-sector AI success.

SPECIAL THANKS

A special thanks to Patrick Fallon, Technology Enablement at Supermicro, for his support and overnight shipping of their latest server with Intel Gaudi 3 accelerators. Without this impressive and timely support, this testing by FedData Technology Solutions would not have been possible to provide the results presented here.

[1] *https://www.intel.com/content/www/us/en/content-details/817486/intel-gaudi-3-ai-accelerator-white-paper.html*

[2] *https://www.intel.com/content/www/us/en/content-details/849557/intel-gaudi-3-ai-accelerator-performance-and-economic-analysis-white-paper.html*

[3] *This metric is the number of requests server per second ratio vs. the average power draw (W): Intel Gaudi 3: 0.05 reqs/sec/W; Intel Gaudi 2: 0.025 reqs/sec/W.*

[4] *https://infohub.delltechnologies.com/static/media/client/7phukh/DAM_4139cb1e-746d-4064-820d-4d17c20ef93b.pdf*