SOVEREIGN: How does the inference latency and throughput of AnyExperts' on-demand routing strategy compare to fixed routi

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Modern datacenters increasingly rely on low-power, single-slot inference accelerators to balance performance, energy efficiency, and rack density constraints. The NVIDIA T4 GPU has become widely deployed due to strong performance per watt and mature software support. Its successor, the NVIDIA L4 GPU, introduces improvements in Tensor Core throughput, cache capacity, memory bandwidth, and parallel execution capability. However, limited empirical evidence quantifies the practical inference performance gap between these two generations under controlled and reproducible conditions. This work int

1 Introduction

Analysis of: DEEP-GAP: Deep-learning Evaluation of Execution Parallelism in GPU Architectural Performance. Research goal: How does the inference latency and throughput of AnyExperts' on-demand routing strategy compare to fixed routing baselines (e.g., top-k routing) across varying batch sizes on standard multimodal benchmarks like VQAv2 and Visual7W?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

9 papers retrieved. 9 claims extracted, 3 verified. Tribunal: 6.5/10 → REVISE (revision_round=1). Policy: ESCALATE_TO_OWNER.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
INT8 achieves up to 58× throughput improvement over CPU baselines.	×	0.12
L4 achieves up to 4.4× higher throughput than T4 while reaching peak efficiency at smaller batch sizes (B=16–32).	✓	0.26
T4 remains competitive for large-batch, throughput-oriented workloads and environments where cost efficiency, power cons	×	0.14
L4 maintains methodological continuity with previous CPU benchmarking to understand the transition from CPU-bound infere	×	0.06
DEEP-GAP provides practical guidance for selecting precision modes, batch sizes, and accelerator generations in accordan	✓	0.25
ResNet-18 with INT8 TensorRT on L4 achieves 38,932 images/sec throughput.	×	0.06
FP16 achieves up to 3.83× speedup over CPU baseline for ResNet-18 on T4.	×	0.04
FP16 achieves up to 8.84× speedup over CPU baseline for ResNet-18 on L4.	×	0.04
L4 reaches peak efficiency at smaller batch sizes (B=16–32), improving latency–throughput tradeoff characteristics for l	✓	0.24

References

- http://arxiv.org/abs/2604.14552v2
- http://arxiv.org/abs/1007.0409v1

- http://arxiv.org/abs/2107.12246v2