LLM Workload Characterization Over Edge-Cloud Systems

Işıl Su Karakuzu

Computer Engineering, Boğaziçi University

Advisors: Cem Ersoy, Bahri Atay Özgövde

Introduction

LLMs are widely used by students for writing, coding, and research. This project benchmarks their GPU-backed deployments and simulates them on EdgeCloudSim to gain insights.

Real Workload Characterization

Model Selection	LLaMA3–8B (Meta), Gemma–7B (Google), DeepSeek-Coder–6.7B — evaluated for writing, Q&A, and code generation.
Deployment	Dockerized containers on Cloud Run with 1, 3, and 5 × NVIDIA L4 GPUs (24GB VRAM).
Test Setup	Load tested via Locust under 100–2000 users; prompt types: general vs. code.
Measured Metrics	Time to First Token (TTFT), latency (p50–p95), throughput, 4xx/5xx error rates, GPU/CPU utilization.
Platform Constraints	Max 1000 concurrent requests, 300s timeout, 8vCPU, 32GiB RAM per instance.

General Prompt	Code Prompt
How do I write a professional follow-up email after a job interview?	Design a simple URL shortener with encode/decode functions.

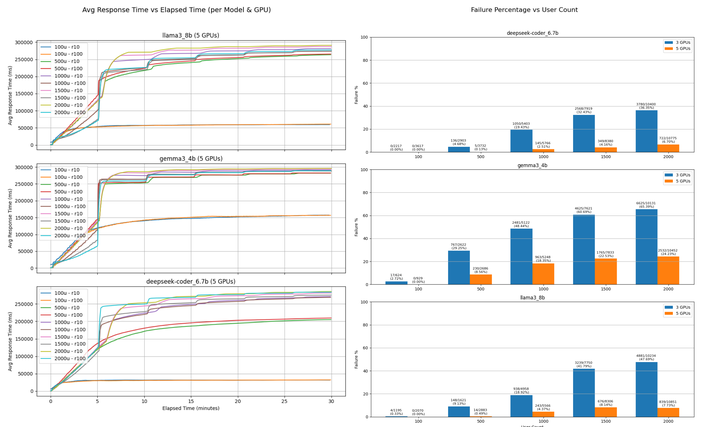

Table: Average Latency and Resource Utilization at 1000 Users

Model	Latency p50 (ms)	GPU Util.	CPU Util.
DeepSeek	276099.78	0.885	0.16
Gemma	294497.60	0.80	0.20
LLaMA	279071.78	0.71	0.16

Table: Average Metrics at 100 Users and 5 GPUs

Model	2xx Req. Count	Avg. Resp. Time (ms)	Tokens/sec
DeepSeek	1.98	31,887.16	712.76
Gemma	0.48	156,832.82	501.64
LLaMA	1.12	60,056.80	597.72

EdgeCloudSim Integration

Part 1 results were mapped into EdgeCloudSim, a simulator developed based on Boğaziçi University to model and evaluate local LLM deployments in campus network settings.

Hardware and Network Configuration

- **Cloud:** Each VM was assigned 30M MIPS, based on 5×NVIDIA L4 (150TFLOPS FP32) [1], approximating an A100 GPU (312TFLOPS FP16) [2].
- **Edge:** VMs limited to 10,000 MIPS to preserve CPU-only edge characteristics [3]. Throughput scaled by increasing VM count, matching 3×L4 performance seen in local tests.
- **WAN:** 50Mbps bandwidth and 0.1s latency were chosen based on our 2000-user DeepSeek run on GCP, with peak upload 23.1Mbps and p99 RTTs between 38–85ms.
- **WLAN:** Set to 300Mbps to avoid last-hop bottlenecks—3× the average Eduroam bandwidth (100Mbps) measured at Boğaziçi University.

System Configuration and Workload Parameters

Property	Original	Modified Configuration
Number of Edge Datacenters	14	14
Hosts per Datacenter	1	1
Host Core/MIPS	4 cores, 2000 MIPS	320 cores, 10000 MIPS
VMs per Host	2	16
VM Core Count	2	20
VM MIPS	1000	10000
VM RAM	2GiB	4GiB

Table: Edge Device Configuration Changes (VM-Scaled for LLM Workload)

Parameter	Code Generation	General Use	Heavy Computation	Non-LLM
usage_percentage	21	59	5	15
prob_cloud_selection	40	40	40	10
task_length	500000	1000000	5000000	15000
vm_utilization_on_cloud	3	3	4	1
vm_utilization_edge	15	30	40	10

Table: Application Parameters

Key Finding: Edge CPU Saturates Early for LLMs

Despite aggressively scaling CPU-only edge resources (224 VMs at 10,000MIPS each), our simulations reveal that LLM workloads saturate edge compute capacity at merely **300–400 concurrent users**.

Why? Even under ideal networking (WLAN=300Mbps, WAN=50Mbps, negligible latency), failure rates remained high due to **VM capacity**, not network I/O.

Implication: A single LLM inference requires millions of instructions. Without GPUs or specialized accelerators, CPU-only edge nodes cannot keep up—even with 2-tier offloading strategies.

Conclusion: *LLMs are currently infeasible at the edge without GPU-class acceleration or extreme model compression.*

[1] NVIDIA L4 Tensor Core GPU Specifications, TechPowerUp. https://www.techpowerup.com/gpu-specs/l4.c4091

[2] NVIDIA A100 Tensor Core GPU Datasheet. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-us-nvidia-1758950-r4-web.pdf

[3] Tango of Edge and Cloud Execution for Reliability. https://engineering.purdue.edu/dcsl/publications/papers/2019/dependability_edge_mecc19_cameraready.pdf

Simulations

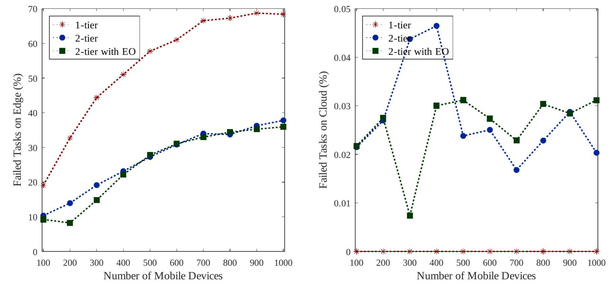

Figure: Comparison of Failed Tasks on Cloud and Edge Across Architectures

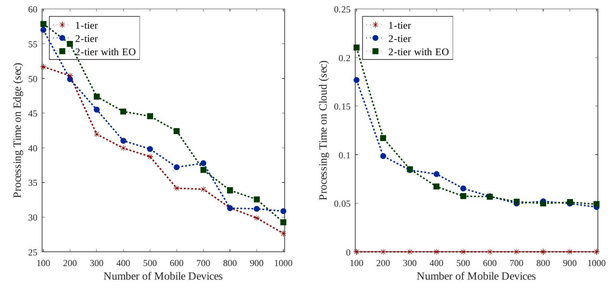

Figure: Processing Time on Edge and Cloud vs. Number of Devices

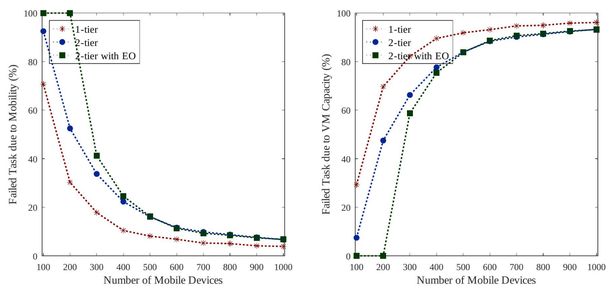

Figure: Comparison of Failure Reasons: Mobility vs VM Capacity

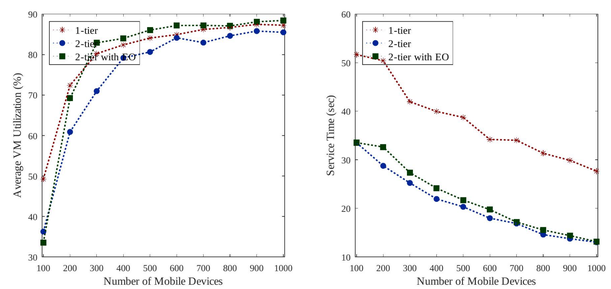

Figure: Average VM Utilization and Service Time vs. Number of Mobile Devices