Benchmarking AI Factories on MeluXina supercomputer

Luca Leonzio, Jonah Holtmann, and Mihkel Tiks

EUMaster4HPC double degree program.

Motivation and Objectives

AI Factories require benchmarking across heterogeneous workloads. The framework provides a modular, reproducible solution for benchmarking inference servers, vector databases, and parallel file systems using SLURM orchestration.

Our Approach: A modular Python framework for the MeluXina supercomputer that orchestrates standardized benchmarks via SLURM, instruments services with Prometheus/Grafana, and provides interactive CLI workflow management.

System Architecture

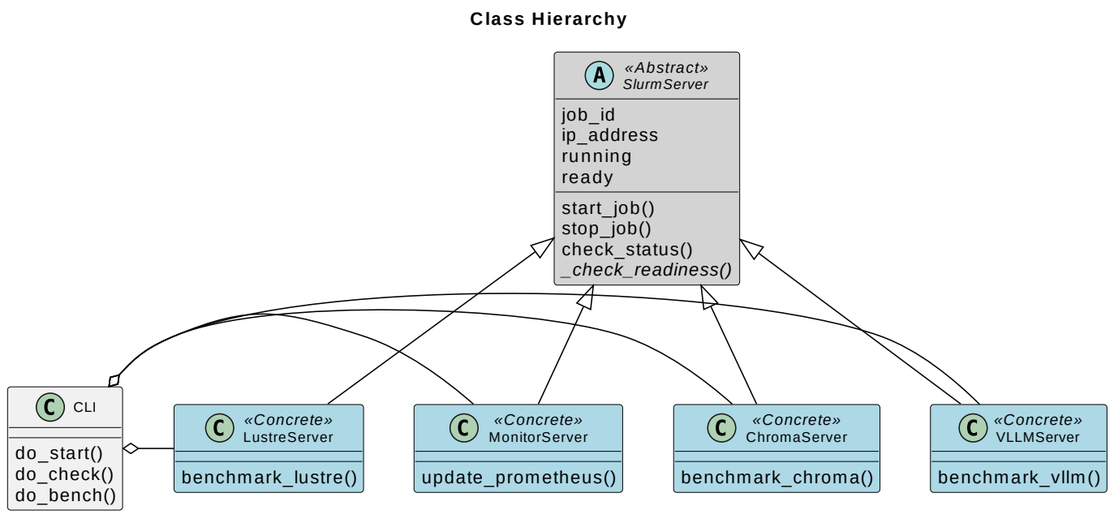
Developed a two-layer architecture, clean separation between UI and infrasturcure.
Object-Oriented Design: Abstract SlurmServer base class. Concrete services: VLLMServer, ChromaServer, LustreServer
Unified Monitoring Stack

1. Prometheus (metrics collection)
2. Grafana (visualization)
3. OpenTelemetry (instrumentation)

Benchmarking capabilities

1. vLLM Inference Server
2. ChromaDB Vector Database
3. Lustre Parallel File System

Interactive CLI Interface taking care of SLURM job management ,real-time status tracking and dynamic configurations.



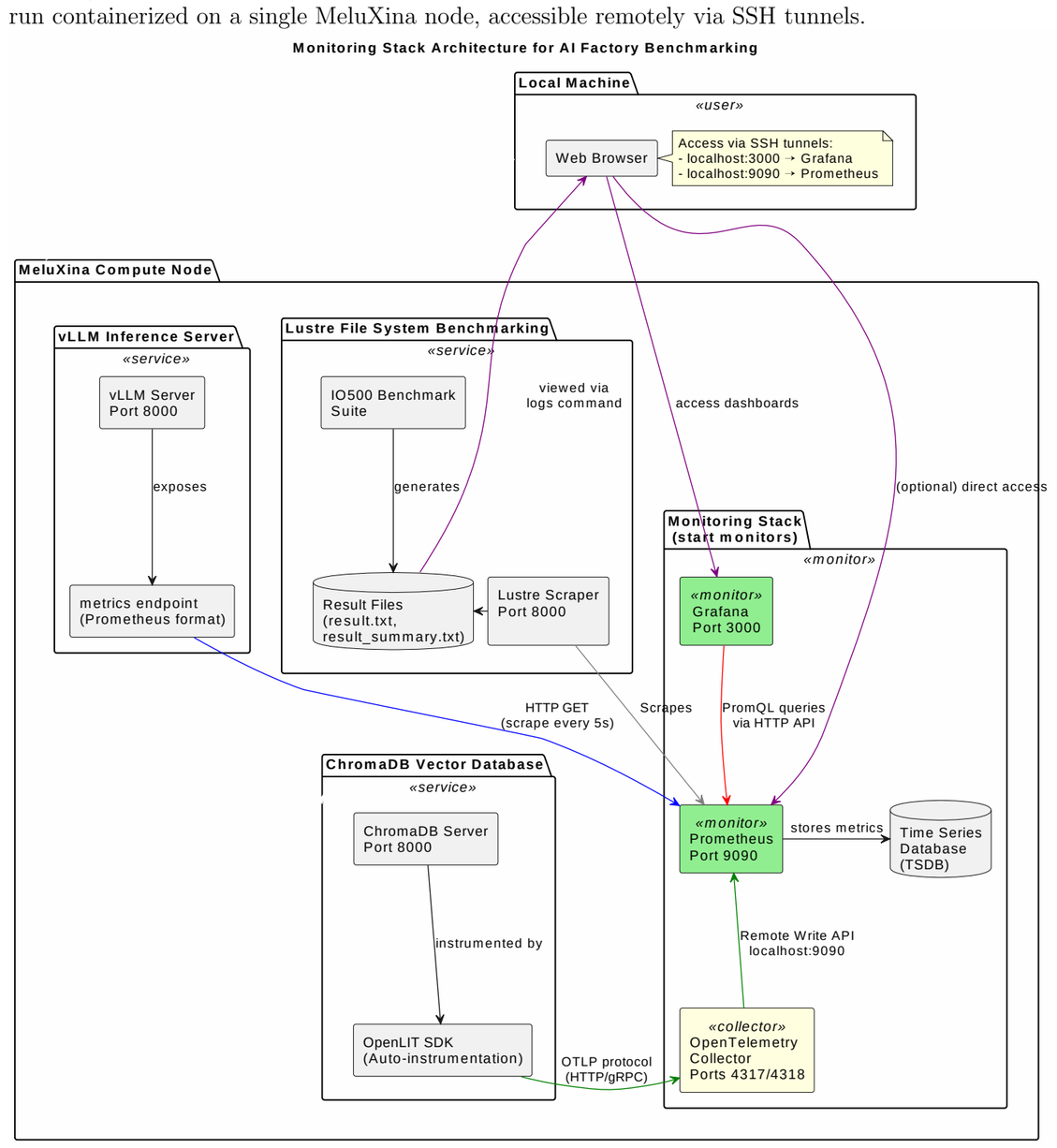
CLI Workflow

An easy-to-use command-line interface manages the entire benchmarking workflow. Users start monitoring services, launch AI Factory workloads (vLLM, ChromaDB, Lustre), and retrieve real-time performance metrics through intuitive commands with automatic service discovery and Prometheus integration.
The available commands are the following:

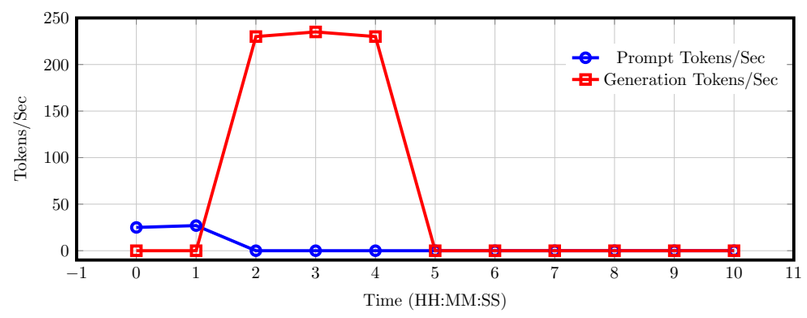
```
# Start/Stop/Check services
bench> start [vllm|chroma|lustre|monitors]
bench> stop [vllm|chroma|lustre|monitors]
bench> check [vllm|chroma|lustre|monitors]

# Run benchmarks
bench> bench [vllm|chroma|lustre]

# View logs
bench> logs [vllm|chroma|lustre|monitors]

# Save logs to archive
bench> save [vllm|chroma|lustre|monitors] [filename.zip]
```

Monitoring Workflow

A unified monitoring solution where Prometheus aggregates metrics from vLLM (scraped) and ChromaDB (via OpenTelemetry), with Grafana providing real-time visualization dashboards. All services run containerized on a single MeluXina node, accessible remotely via SSH tunnels.



Token Throughput and temperature (°C) during vLLM structured output benchmarking. Stable thermal profile validates cooling under sustained inference workload.



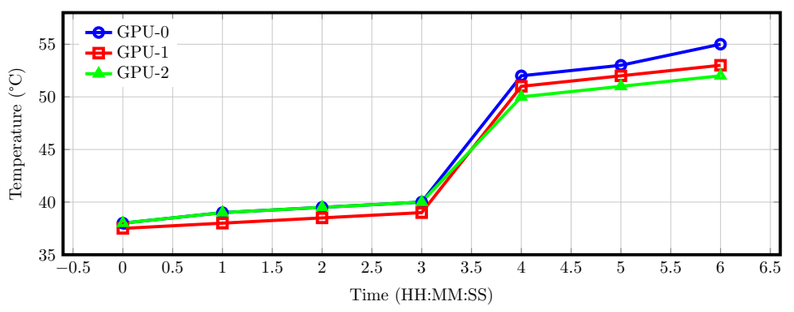


Figure 1: Model used for the graphs: meta-llama/Llama-3.1-8B-Instruct

Key results and capabilities

Metrics

What we measure. Prometheus collects vLLM metrics and the OpenTelemetry Collector forwards ChromaDB metrics (push, OTLP). IO500/Lustre results are ingested from log files for post-analysis. We report throughput, latency percentiles, resource utilization, and dataset/index growth.

- **vLLM:** Throughput (tokens/s), requests/s, latency p50/p95/p99, GPU utilization and VRAM usage (GiB).
- **ChromaDB:** Query QPS, insert QPS, latency p50/p95/p99, batch embedding throughput (vectors/s), collection size (rows), index size (MiB).
- **Lustre/IO500:** IOR read/write bandwidth (GB/s), mdtest metadata ops/s (create/stat/remove), find phase time (s).

Granularity normalization. Metrics are aggregated over fixed scrape windows and normalized to wall-clock time; latency percentiles use distribution histograms for stable p95/p99 estimates.

Performaces

Scaling behaviors (single MeluXina node, containerized). We evaluate how workload throughput and latency evolve with batch size, concurrency, and dataset growth.

Metric	GPU-0	GPU-1	GPU-2	Unit
Peak Utilization	80	81	79	%
Peak Memory	38	37	36	GiB
Peak Temperature	55	53	52	°C
Memory Util.	100	100	100	%

Conclusions

This work presents a **comprehensive, modular benchmarking framework** for evaluating AI Factory workloads on HPC system¡ s. We provide researchers and practitioners with reproducible performance evaluation tools for next-generation AI systems.

- **Reproducible at Scale:** Container-based deployment (Apptainer) ensures portability across HPC systems; SLURM orchestration enables distributed benchmarking
- **Operational Intelligence:** Real-time monitoring via Prometheus/Grafana dashboards with automatic service discovery and dynamic configuration
- **Production-Ready:** 15+ unit tests, end-to-end validation, and pre-built Grafana dashboards for vLLM and ChromaDB