HPRIE: A NUMA-Aware Lock-Free Inference Engine for Low-Latency AI Serving

Huy Ngo
Independent Researcher (High School Senior)

December 2025

Abstract

AI inference serving in production environments is constrained by systems-level bottlenecks such as lock contention and non-uniform memory access (NUMA) effects. This paper presents HPRIE, a high-performance inference engine integrating lock-free concurrency, explicit memory ordering, and NUMA-aware scheduling. HPRIE provides predictable low-latency behavior under contention. We analyze the architecture, implementation rationale, and progress guarantees of its core components.

1 Introduction

Recent advances in machine learning have shifted performance bottlenecks from model computation to inference serving infrastructure. In latency-sensitive domains, tail latency and predictability are critical. HPRIE treats inference serving as a systems-level problem, applying explicit resource management and lock-free data structures to reduce contention.

2 Design Goals

- **Low Latency:** Minimize average and tail latency by avoiding global locks.
- **NUMA Awareness:** Preserve memory locality on multi-socket systems.
- **Scalability:** Scale efficiently with increasing core counts.
- **Resilience:** Maintain stable performance under load spikes.

3 System Architecture

HPRIE follows a modular pipeline consisting of:

1. **Request Ingress Layer:** Accepts requests from multiple producers.
2. **Task Scheduling Layer:** Distributes tasks using lock-free MPMC queues.
3. **Execution Layer:** Performs inference using NUMA-local resources.

4 Concurrency and Memory Model

At its core, HPRIE utilizes lock-free MPMC queues implemented with C++ atomic primitives. By using `memory_order_acquire` and `memory_order_release`, we ensure cache-line alignment and eliminate kernel transitions, which are major contributors to tail latency (P99).

5 Formal Analysis

The MPMC queue satisfies **Linearizability** and **Lock-Freedom**, ensuring that at least one thread makes progress regardless of system-wide delays. This avoids convoy effects common in mutex-protected systems.

6 Conclusion

HPRIE demonstrates that applying rigorous systems programming principles to AI inference yields substantial benefits in scalability and predictability. This work lays the groundwork for treating inference as a first-class systems challenge.