International Journal of Computer Science and Engineering (IJCSE)
ISSN (P): 2278–9960; ISSN (E): 2278–9979
Vol. 12, Issue 2, Jul – Dec 2023; 493–516
© IASET

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

SCALABLE SOLUTIONS FOR REAL-TIME MACHINE LEARNING INFERENCE IN MULTI-TENANT PLATFORMS

Abhishek Das[1], Abhijeet Bajaj[2], Priyank Mohan[3], Prof.(Dr) Punit Goel[4], Dr Satendra Pal Singh[5] & Prof.(Dr.) Arpit Jain[6]

[1]Researcher, Texas A&M University, North Bend, WA -98045

[2]Scholar, Columbia University, Aurangabad, Maharashtra India

[3]Scholar, Seattle University, Dwarka, New Delhi, 110077, India

[4]Research Supervisor, Maharaja Agrasen Himalayan Garhwal University, Uttarakhand, India

[5]Ex-Dean, Gurukul Kangri University Haridwar, Uttarakhand, India

[6]Department of CSE, KL University, Guntur, Andhra Pradesh, India

ABSTRACT

Real-time machine learning inference is a critical capability for modern multi-tenant platforms serving industries such as finance, healthcare, and e-commerce, where timely predictions directly impact user experience and business outcomes. However, scaling real-time inference for multiple tenants introduces unique challenges, such as managing resource allocation, maintaining low latency, ensuring system stability, and handling dynamic workloads. This paper presents a comprehensive exploration of scalable solutions for real-time inference in multi-tenant environments, addressing these challenges by proposing an architecture that leverages dynamic resource scaling, tenant isolation, model optimization techniques, and distributed computing frameworks.

The proposed architecture incorporates a microservices-based approach with container orchestration to enable dynamic scaling and efficient resource utilization. By utilizing Kubernetes and serverless computing techniques, the system dynamically allocates resources to each tenant based on real-time demand, thereby minimizing idle resource usage while maintaining high availability and performance. In addition, a multi-level load balancing strategy is employed to distribute inference requests across nodes, reducing latency spikes during peak loads and ensuring consistent response times.

To address the specific challenges of multi-tenancy, the architecture integrates robust tenant isolation mechanisms through namespace segregation and resource quotas. This prevents resource contention among tenants and enables secure model deployment for different users. Furthermore, to optimize inference speed, model compression techniques such as pruning, quantization, and knowledge distillation are applied to reduce model size without sacrificing accuracy, allowing for faster inference times even under resource-constrained environments.

The paper also explores the use of specialized hardware accelerators such as GPUs, TPUs, and FPGAs for high-throughput inference, analyzing the trade-offs between cost and performance. Dynamic hardware allocation strategies are proposed to ensure that compute-intensive workloads are directed to appropriate accelerators based on real-time demand. Additionally, caching and pre-computation strategies are used to eliminate redundant inference calculations for repeated or similar inputs, further enhancing throughput.

Experimental results demonstrate that the proposed architecture achieves significant improvements in both latency and throughput, with a 30% reduction in response times and a 50% increase in request handling capacity compared to traditional approaches. The system's ability to dynamically scale and isolate tenant workloads also results in more predictable performance, making it ideal for real-time applications that require stringent service-level agreements (SLAs).

Finally, the paper highlights key challenges such as managing model drift, ensuring data privacy, and handling cross-tenant data dependencies, suggesting future research directions in areas like federated learning and multi-tenant optimization frameworks. By addressing these aspects, the proposed solution provides a robust foundation for scalable, real-time machine learning inference in complex multi-tenant platforms.

KEYWORDS: *Real-Time Inference, Multi-Tenant Platforms, Scalable Architecture, Dynamic Scaling, Tenant Isolation, Model Optimization, Kubernetes, Microservices, Serverless Computing, Hardware Accelerators, Model Compression, Caching Strategies, Machine Learning, Latency Reduction, Throughput Optimization, Resource Allocation, Data Privacy, Cross-Tenant Learning.*

Article History

Received: 18 Nov 2023 | Revised: 23 Nov 2023 | Accepted: 27 Nov 2023