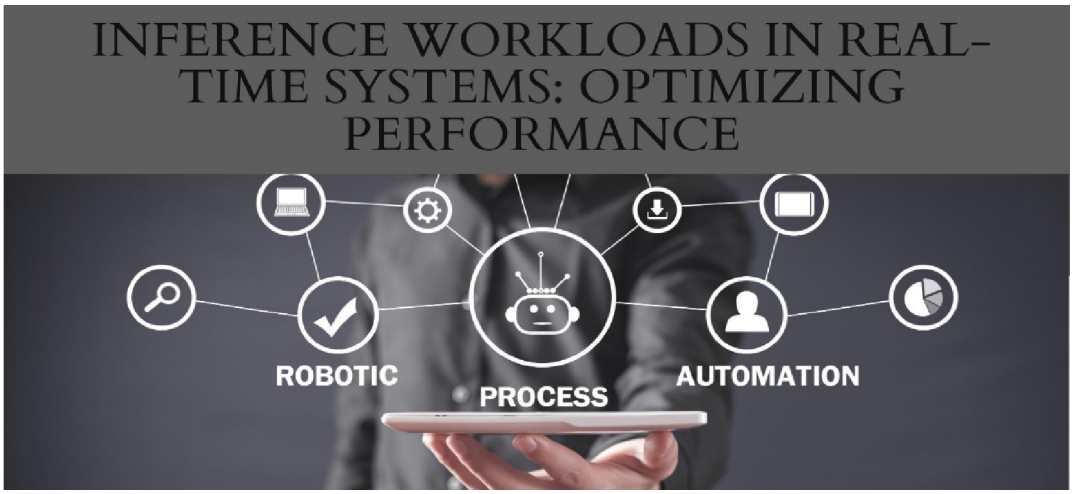
IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 12, April 2025
Impact Factor: 7.67

Inference Workloads in Real-Time Systems: Optimizing Performance

Deepika Bhatia
San Jose State University, USA



Abstract: *This article examines the optimization strategies for inference workloads in real-time systems across various performance-critical applications. As artificial intelligence becomes increasingly embedded in time-sensitive domains, the need for efficient execution of inference tasks under strict latency constraints has become paramount. Unlike training processes prioritizing accuracy regardless of computational cost, inference workloads must balance precision with performance constraints, particularly in resource-limited environments. The article explores three key optimization approaches: reduced precision computing techniques that preserve accuracy while decreasing computational demands, resource allocation and workload management strategies that adapt to fluctuating conditions, and specialized cache architectures that minimize memory access latencies for tensor operations. Through case studies in autonomous vehicles, industrial automation, and financial transaction monitoring, the article demonstrates how these optimizations enable mission-critical AI systems to meet stringent real-time requirements. Additionally, emerging directions, including hardware-software co-design, neural architecture search for efficiency, and sparse computation, are explored as promising frontiers for future optimization efforts*

Keywords: Real-time inference, reduced precision computing, resource allocation, memory optimization, hardware-software co-design

DOI: 10.48175/IJARSCT-25987