CSCE 5610 Computer System Architecture
Term Project – Survey

Survey Title: Architectural Innovations for Real-Time Machine Learning in Healthcare

Uday Bhaskar Valapadasu - 11696364
Tulasi Sai Pechetti -11663410
Jayakrishna Amathi - 11697680

Motivation and Background:

The integration of real-time machine learning (ML) applications into healthcare is transforming the industry by enabling personalized medicine, diagnostics, and medical imaging. For example, machine learning models are increasingly being used for tasks such as analyzing medical images (e.g., MRI, CT scans), predicting patient outcomes, and providing real-time diagnostic support during surgeries. However, these applications demand highly efficient and dependable architectures that can deliver low-latency inference, high computational performance, and energy efficiency, all while ensuring reliability and fault tolerance given the critical nature of healthcare systems.

Healthcare applications present unique computational challenges due to strict constraints on latency, accuracy, and power consumption—especially in real-time scenarios such as medical imaging analysis during surgeries or diagnostic predictions in intensive care units. Traditional general-purpose architectures are often insufficient to meet these requirements, which has led to the development of specialized hardware and architectural innovations tailored for real-time machine learning in healthcare.

Project Objective:

This survey aims to explore recent architectural innovations that have been designed to support real-time machine learning applications in healthcare. The survey will focus on both the training and inference phases of ML models, analyzing how various architectural optimizations are addressing the specific needs of healthcare applications. The key areas of focus will include architectures for real-time medical image processing, low- latency inference for personalized medicine, hardware accelerators for AI applications in healthcare, and fault- tolerant architectures that ensure reliability in critical healthcare systems.

Key Topics:

1. Architectures for Real-Time Medical Image Processing (e.g., MRI, CT scans):
- Explore hardware and architectural solutions that enable real-time processing of medical images, which is crucial for diagnostic accuracy and timely interventions. This section will cover GPU-based solutions, FPGA-based accelerators, and application-specific integrated circuits (ASICs) designed to handle high-dimensional imaging data efficiently.

2. Low-Latency Inference for Personalized Medicine:
- Examine architectural techniques that minimize inference latency for applications like personalized medicine, where patient-specific data must be processed in real-time to provide treatment recommendations or adjust therapies dynamically. This will include techniques such as model partitioning, dynamic voltage scaling, and edge computing architectures.

3. Hardware Accelerators for Healthcare AI Applications:
- Survey the use of specialized hardware accelerators, such as Google's Tensor Processing Units (TPUs), NVIDIA's AI GPUs, and custom ASICs, in the context of healthcare. This section will focus on how these accelerators are optimized for healthcare-specific workloads, including deep learning for diagnostics, decision support systems, and predictive analytics for patient outcomes.

4. Fault-Tolerant and Dependable Architectures for Critical Healthcare Systems:
- Investigate architectural innovations aimed at ensuring reliability and fault tolerance in healthcare systems. Given the life-critical nature of healthcare applications, these systems must operate without fail. Techniques such as error correction, redundancy, checkpointing, and failover mechanisms will be explored in the context of both hardware and software to ensure dependable operation.

Approach:

The project will involve an extensive literature review of recent research papers, articles, and industry reports on architectural innovations for real-time machine learning in healthcare. We will categorize different architectural approaches based on the types of healthcare applications they target (e.g., diagnostics, medical imaging, personalized medicine) and evaluate each based on specific metrics such as latency, energy efficiency, fault tolerance, and scalability. Additionally, we will highlight case studies of real-world implementations of these architectures in healthcare, such as AI-powered diagnostic tools in hospitals or edge devices for remote patient monitoring.

Expected Outcome:

By the end of this survey, we aim to provide a comprehensive overview of the state-of-the-art architectural advancements that are enabling real-time ML applications in healthcare. The survey will summarize key challenges in the field, highlight current solutions, and identify potential areas for future research and development. This work will serve as a valuable resource for researchers and engineers interested in designing next-generation architectures for healthcare AI applications