FPGA-based CNN Accelerator

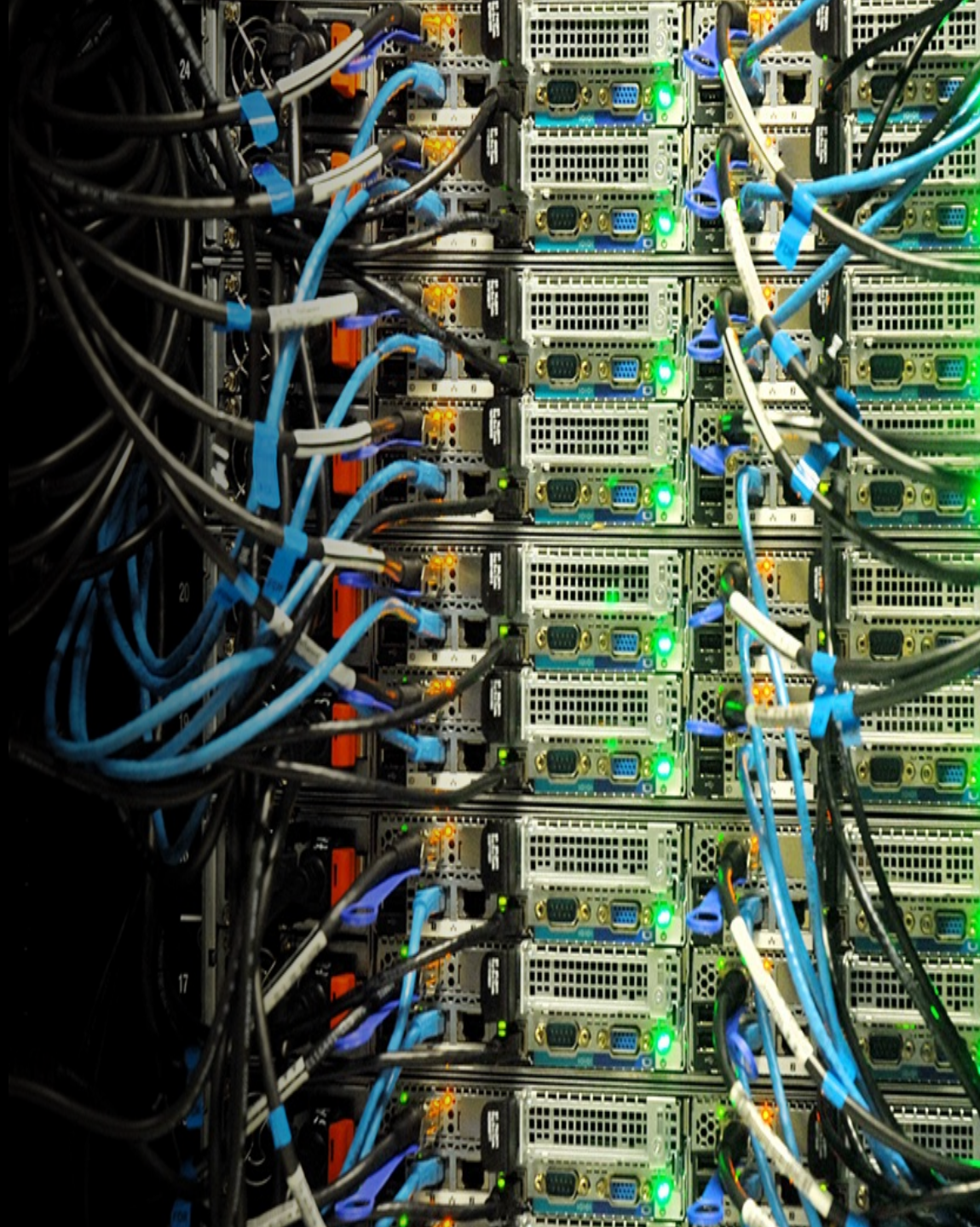
Group Members

- Syed Muhammad Ashhar Shah 2020478
- Hasaan Noor 2020546
- Khizar Ali Shah 2020196

Supervisor: Dr. Muhammad Taj

Problem Statement

- Convolutional Neural Networks (CNNs) have become a fundamental tool in the field of computer vision and deep learning. However, their computational requirements can be demanding, especially for large-scale models or real-time applications. To achieve high detection accuracy, the large and complex CNN models are used, but this leads to expensive computational cost, which is hard to process in real time in resource-constrained devices with strict latency and energy requirements.
- CPUs, and GPUs may not provide the necessary real-time or energy efficiency required for these tasks. Secondly, the physical size and energy efficiency of GPUs is not feasible to be implemented in many such scenarios.
- Problem: High computation, physical size, energy efficiency, and other constraints cause to CPU and GPU based CNNs be unusable or not being fully utilized in real-time applications involving resource-constraint devices.

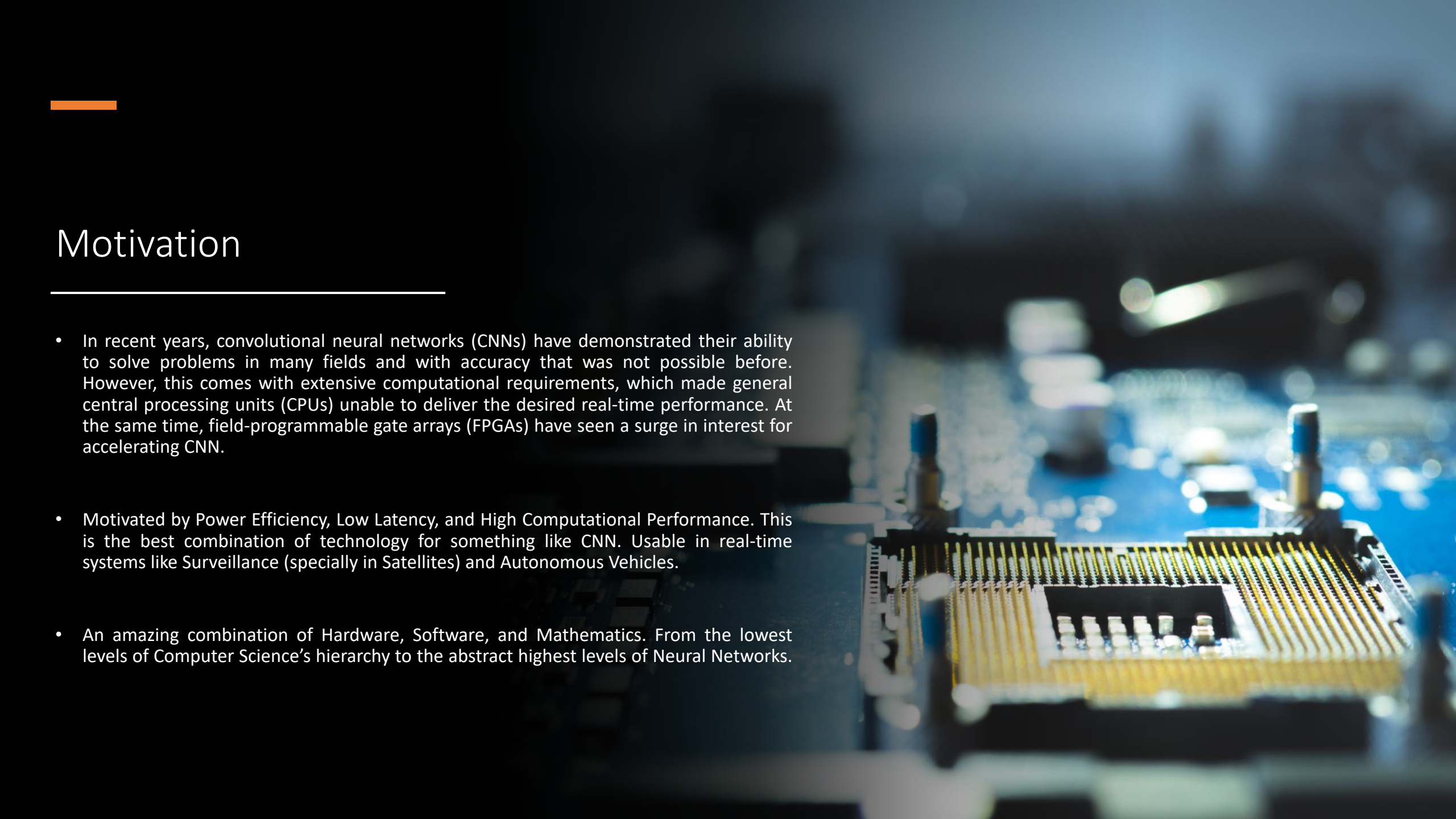


Motivation

- In recent years, convolutional neural networks (CNNs) have demonstrated their ability to solve problems in many fields and with accuracy that was not possible before. However, this comes with extensive computational requirements, which made general central processing units (CPUs) unable to deliver the desired real-time performance. At the same time, field-programmable gate arrays (FPGAs) have seen a surge in interest for accelerating CNN.

- Motivated by Power Efficiency, Low Latency, and High Computational Performance. This is the best combination of technology for something like CNN. Usable in real-time systems like Surveillance (specially in Satellites) and Autonomous Vehicles.

- An amazing combination of Hardware, Software, and Mathematics. From the lowest levels of Computer Science's hierarchy to the abstract highest levels of Neural Networks.



Justification

- Limitation of CPU and GPU architectures
- Better performance per watt than GPU
- Low latency
- Energy efficiency
- On-chip memory
- Physical Size
- Parallelism

Look at the combined effect

Literature Review

- Real-time Object Tracking:
- https://www.researchgate.net/publication/359411914_AlgorithmHardware_Co-Design_for_Real-Time_On-Satellite_CNN_based_Ship_Detection_in_SAR_Imagery

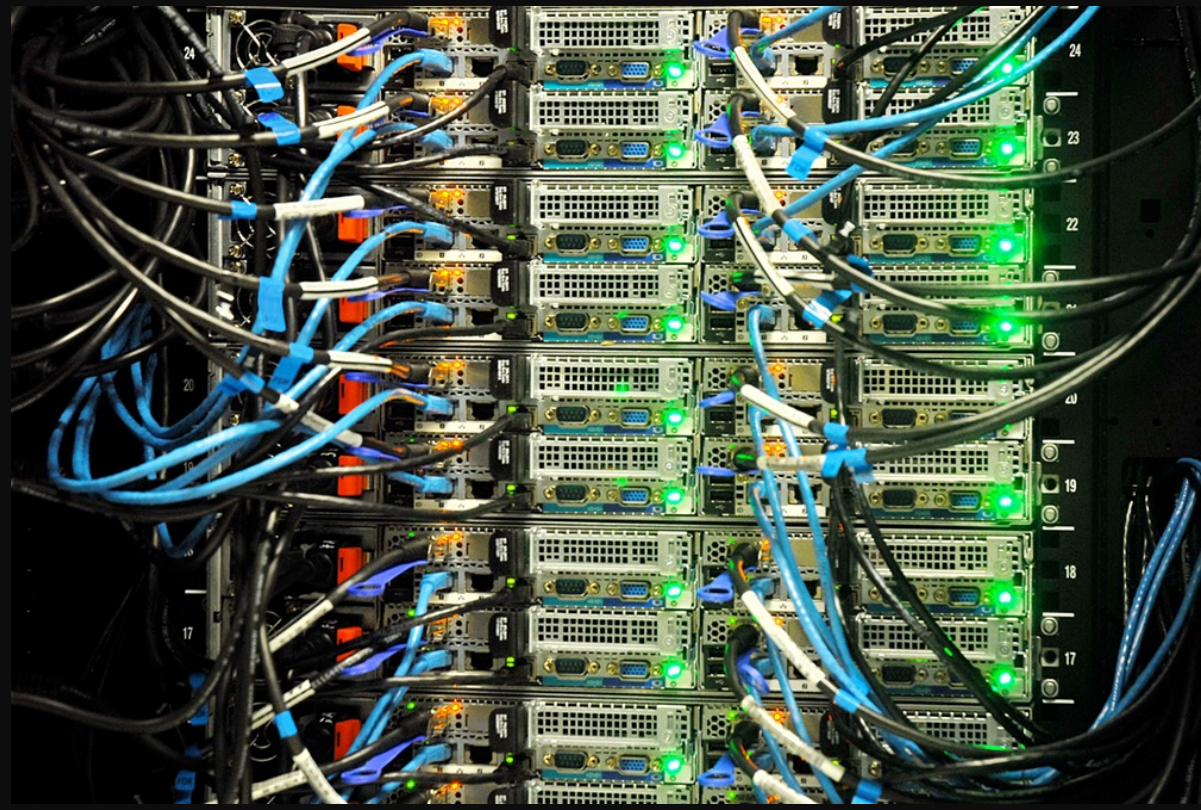
- Achieving Parallelism:
- https://www.researchgate.net/publication/368529697_Model_Parallelism_Optimization_for_CNN_FPGA_Accelerator

Engineering Challenges

- Optimizing the resources available on the target FPGA device to deliver the optimal pe rformance

- Writing RISC-V Assembly Language Code of CNN

- Due to growing complexity of CNN architectures, it is hard to meet performance metrics such as latency and throughput while optimi zing power



Attributes Included

- Wide-ranging/conflicting technical issues
  - FPGAs have finite resources and the CNNs, especially deep and complex models, require significant computational resources and memory to operate efficiently. Conflicts can arise when trying to strike a balance between performance and power consumption.
- Component Parts / Sub Problems
  - When implementing a design on an FPGA (Field-Programmable Gate Array), several component parts or subproblems need to be considered. This includes but not limited to FPGA Fabric, Clocking, Memory, Arithmetic Units, etc.
- Diverse Group of Stakeholders
  - In this project there are diverse groups of stakeholders involved, each with their own interests, roles, and responsibilities such as Hardware Engineers, Software Engineers, Data Scientists and Artificial Intelligence Engineers.

Work Plan

Learn CNN

Verilog

Explore some CNN Achitectures

FPGA Tools

FPGA Design Flow

- Verilog -> Simulation -> RTL Schematic -> Burn

Understanding CNN Working:

- ML/DL -> Python -> C -> RISC-V

Measurement: Keep comparing it with software models and ‘repeat’

Deliverables (Tentative)

- Workable first design by the 3rd week of 7th semester
- Debugged and Optimizeed design by the end of 7th semester