# ECE 554 Group 3 Lightning Talk

## Custom RISC-V Processor and Speech-to-Text Model

25th March 2025

# Team Introduction

**Rohan Rao** **Asish Das** **Aditi Shah** **M Sadman Sakib**

# Use-Case Definition

## Custom RISC-V & Speech-to-Text

- **Custom RISC-V co-processor on FPGA fabric:** with a CPU core, SIMD Engine / Custom Accelerator

- **Deep Learning-based Speech-to-Text Model:** Use processor to accelerate a STT model like DeepSpeech or Wav2Vec

# Motivation

- **Altera Cyclone V and Acceleration:** FPGAs provide custom hardware acceleration, which is good to experiment with such a computationally intensive task.

- **Customizability of RISC-V:** RISC-V's open-source nature promotes customization and extension with custom instructions - like SIMD - for applications. RISC-V's Toolchain allows use of high-level languages as needed.

- **Pertinence of Speech-to-Text:** Speech-to-text is a critical technology for applications like voice assistants, transcription, and accessibility tools, and can be further used for real-time language translation.

# Approach

- **RISC-V Co-Processor Design**:
  - Implement a RISC-V core (5-stage pipelined CPU or an existing core like PicoRV32) on the FPGA.
  - Extend the ISA with custom instructions for SIMD operations and DeepSpeech acceleration.
- **SIMD Engine and Custom Accelerator**:
  - Design a SIMD engine for parallel processing (matrix multiplications).
  - Implement a custom accelerator for LSTM layers and other DeepSpeech operations.
- **Integration with ARM HPS/SDRAM**:
  - Use shared memory or DMA for data transfer between the ARM HPS or SDRAM with the RISC-V co-processor.
  - Write a SDRAM controller to interface with the memory and memory-map it with the co-processor.
  - Develop firmware to manage data flow and control.
- **DeepSpeech Model Execution**:
  - Preprocess audio data and store in the SDRAM or HPS memory.
  - Offload model execution to the RISC-V co-processor.
  - Postprocess the output and display the text.

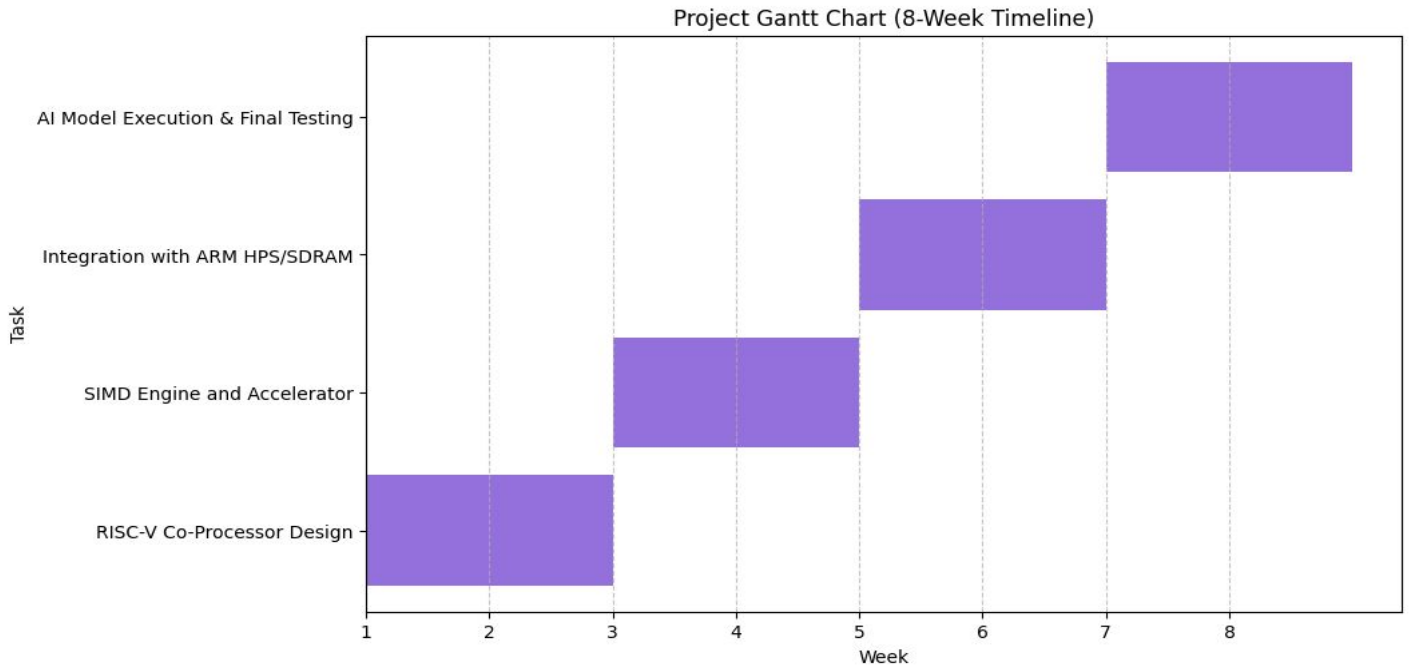
# Anticipated Challenges

- **Resource Constraints**
  - FPGAs have limited resources which can restrict the size of the RISC-V core and accelerators.
- **Model Complexity**
  - DeepSpeech is computationally intensive, and fitting it onto the FPGA may require significant optimization (either quantization of the model or correct loading).
- **Toolchain Management**
  - Ensuring the toolchain is correctly configured to support our extensions and use them correctly to compile programs could be tricky.
- **Data Transfer Bottlenecks**
  - Transferring data between the ARM HPS or SDRAM and RISC-V co-processor may become a bottleneck if not optimized.
- **Debugging and Verification**
  - Debugging a system with multiple integrated components (SDRAM, ARM-HPS, FPGA) can be challenging.

# Risk Management

- **Resource Constraints**
  - Optimize the design for resource usage (use fixed-point arithmetic, reduce model size).
- **Model Complexity**
  - Start with a smaller model or a subset of DeepSpeech with a simpler accelerator (only the LSTM layers).
- **Toolchain Configuration**
  - Setup the extensions in a clear and correct way and configure the toolchain early to avoid issues.
- **Data Transfer Bottlenecks**
  - Use DMA for efficient data transfer and optimize memory access patterns.
- **Debugging and Verification**
  - Use simulation tools (ModelSim) for early verification and GDB for debugging of each and every module before integration.

# Milestones and Evaluation




**Checkpoint 1 (3-4 weeks from now)**: RISC-V Co-processor and and SIMD engine integration. Successfully passing simulations.

**Final Checkpoint**: Integrating AI model with designed hardware and successful speech to text conversion.

# Team Responsibilities

**Aditi: RISC-V Co-Processor Design Lead**

- Implement RISC-V core (Custom Risc-V 5-Stage Pipeline or Open Source Core (PicoRV32)).
- Add custom instructions for AI acceleration (Instructions to be used for the SIMD/Accelerator).

**Rohan: SIMD Engine/Custom AI Accelerator Lead**

- Design and implement SIMD engine for parallel processing.
- Develop custom AI accelerator for specific workloads (model dependent).
- Optimize accelerator for performance and resource usage.

**Sadman: Memory and ARM HPS Integration Lead**

- Integrate external SDRAM and implement memory controller.
- Develop interface between ARM HPS and RISC-V co-processor.
- Write firmware for data transfer and control.

**Asish: AI Workload and System Integration Lead**

- Load and run AI models (DeepSpeech) on RISC-V co-processor.
- Optimize system for end-to-end performance.
- Test system with real-world data and verify correctness.

# Thank you!

# Questions?