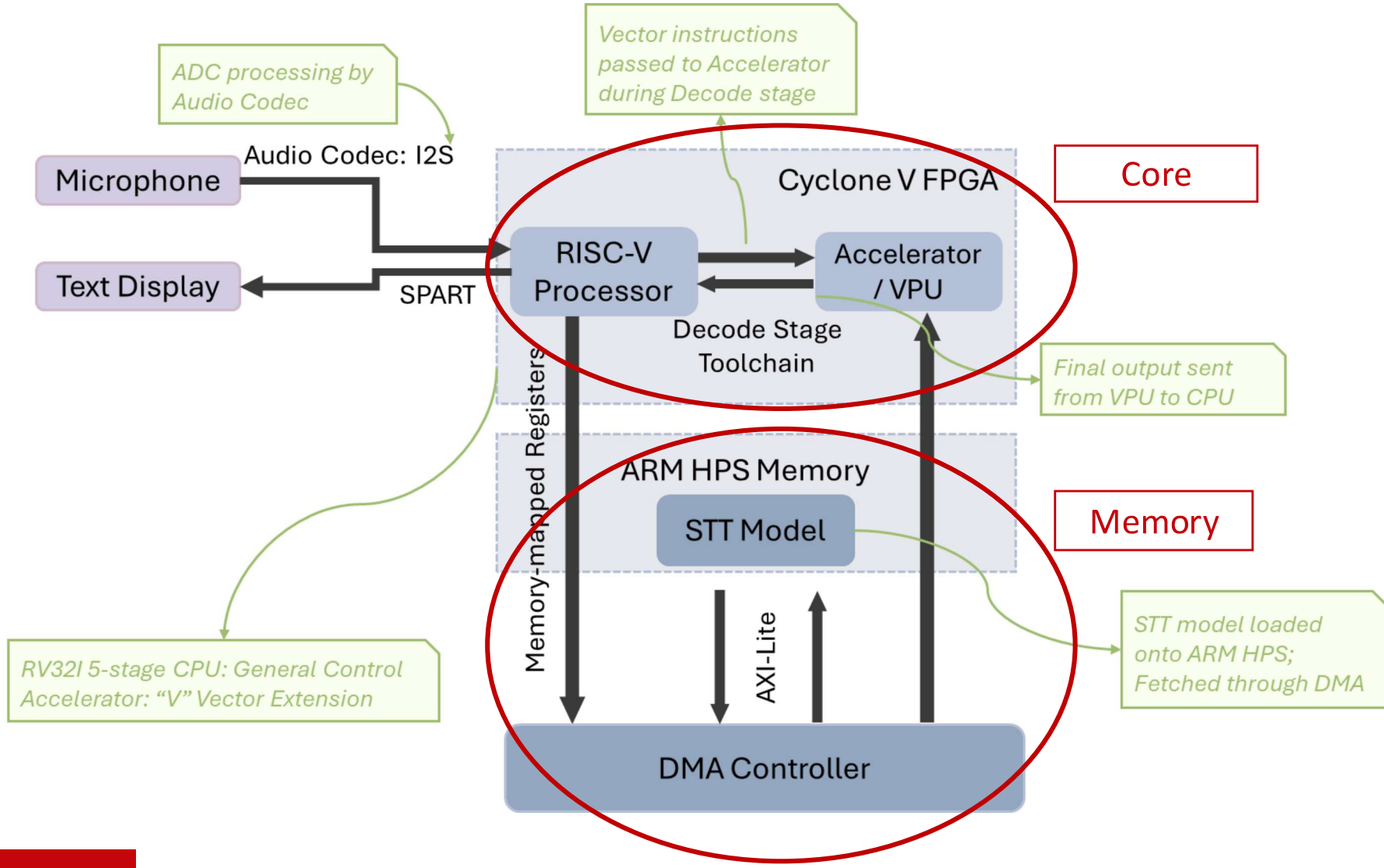
SpeakEasy: Microarchitectural Review

Asish Das, Rohan Rao, Sadman Sakib, Aditi Shah
11th March 2025

High-Level Block Diagram



CPU-VPU Pipeline Diagram

Ibex

- **2-stage** with stalling in the second stage
- 32-bit instructions
- Not PicoRV32 or VescRiscv
- Ibex is in System Verilog and works with the Toolchain

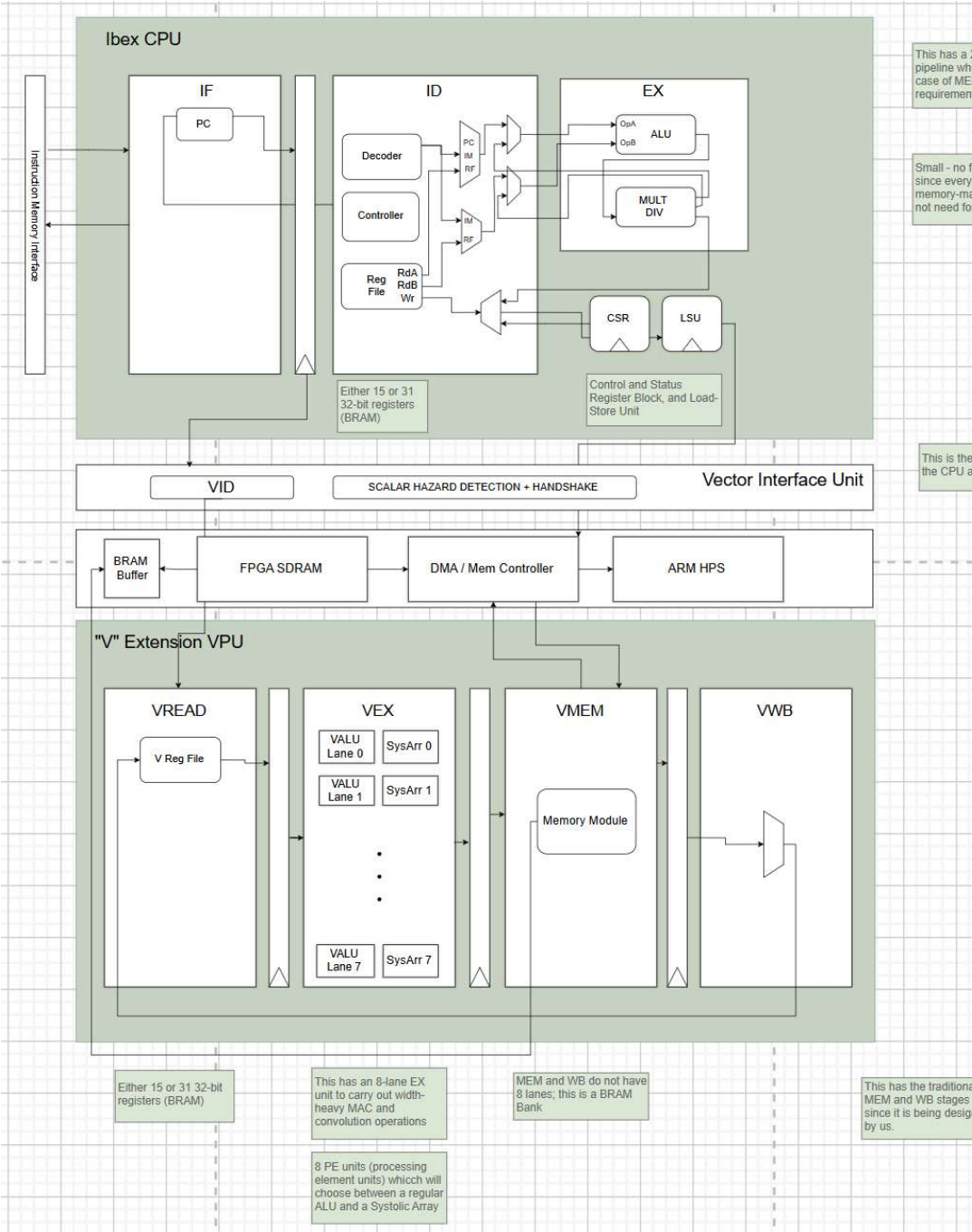
VIU

- For Vector Decode and Handshake

"V" Vector Extension

- **4-stage** with a connection to CPU IF

- **GNU Toolchain-compatible**



VPU Pipeline

- 4-Stage Pipelined Vector Processing Unit (Read, Execute, Memory, Writeback)
- VPU writes to BRAM Buffers (SPRAM) to store outputs and results of different operations (Matrix Mult

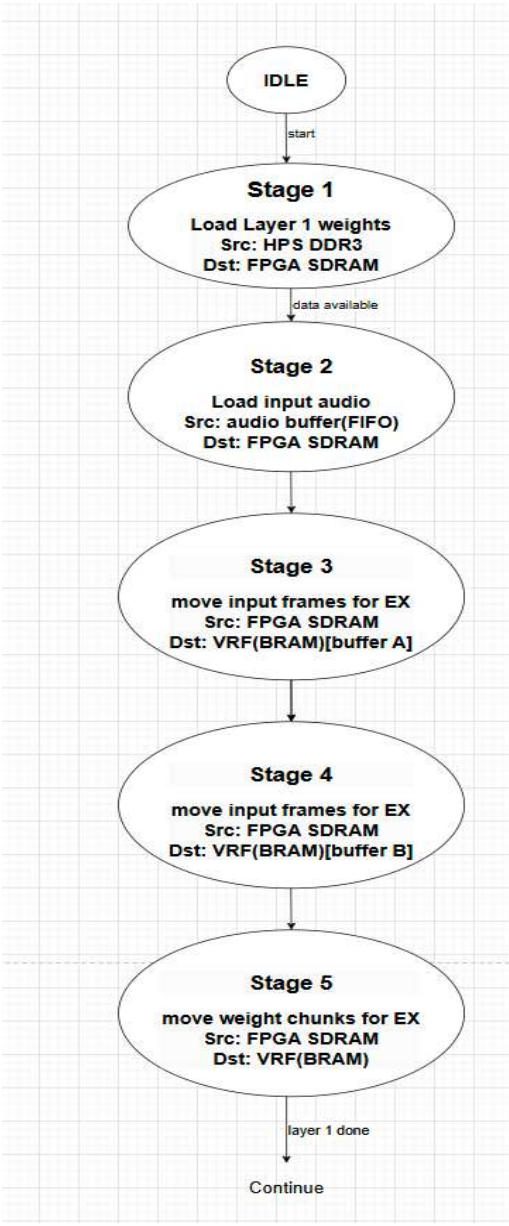
  CNN, RNN, Activations)
- VPU interfaces to the processor through a Vector Interface Unit/Bus which handles the mutual exclusi

  execution of either the processor or the VPU (one is stalled when the other is running), Systolic Array

  Sequencing, and the Vector Decoder Unit for Vector instructions.

Memory Controller

- Async DMA controller
- Descriptor based DMA unit
- None/minimal CPU overhead
- Includes SDRAM controller and FGPA-HPS bridge controller (AXI Interconnect, AXI Lite to APB Bus, AXI-DMA Controller IP Cores)
- Memory map defined for VRF(BRAM), SDRAM and HPS DDR3
- Enables prefetching to hide memory latency



STT: Why Vosk?

Vosk

- **Architecture**: Based on **Kaldi**, supports multiple languages.
- **Pros**:
  - Small model size.
  - Offline, low-latency inference.
  - Easy to integrate with custom hardware.
- **Cons**:
  - Slightly lower accuracy compared to DeepSpeech.

DeepSpeech (Mozilla)

- **Architecture**: Based on Baidu's Deep Sp
  uses **RNNs (LSTM)**.
- **Pros**:
  - Open-source and well-documente
  - High accuracy for general-purpose
    recognition.
- **Cons**:
  - Large model size (~300 MB–1 GB)
  - Computationally intensive (LSTM l
    hard to accelerate).

Thank you! Any Questions?