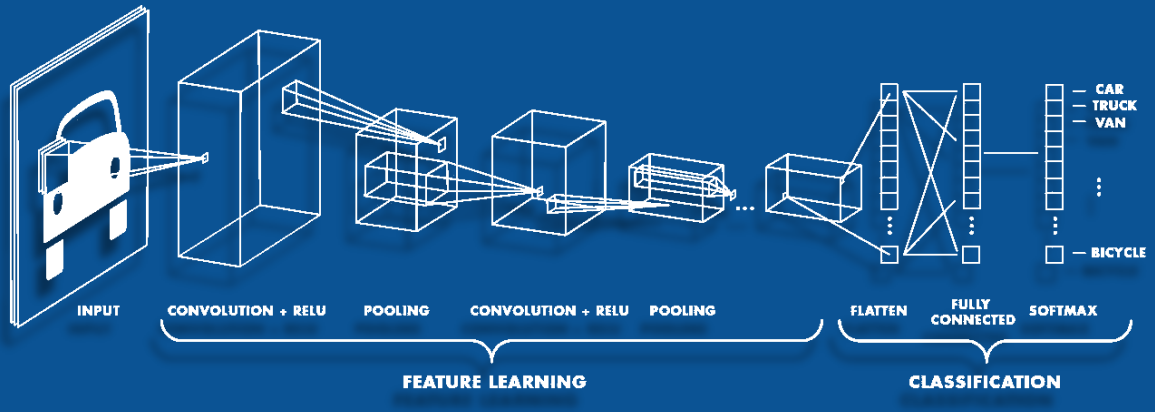
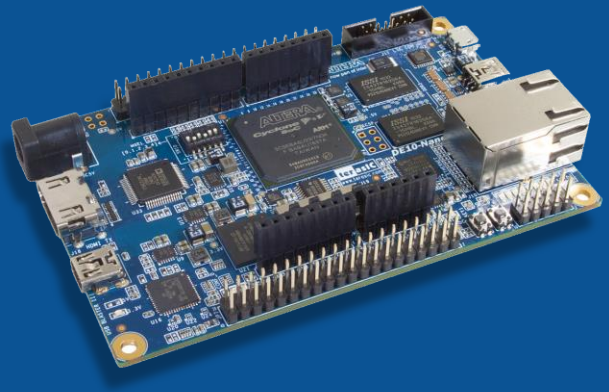
AP085 - Embedded Neural Coprocessor

Ratnasegar Natheesan | Kamalakkannan Kamalavasan | Kathirgamaraja Pradeep

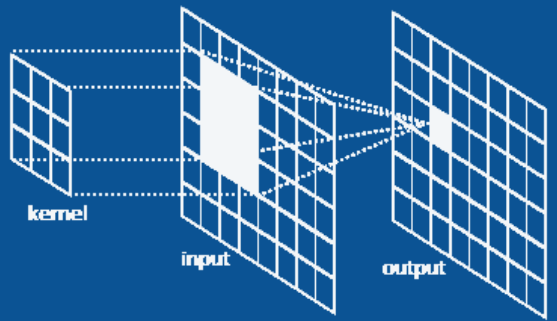
Supervisor - Dr. Ajith Pasqul

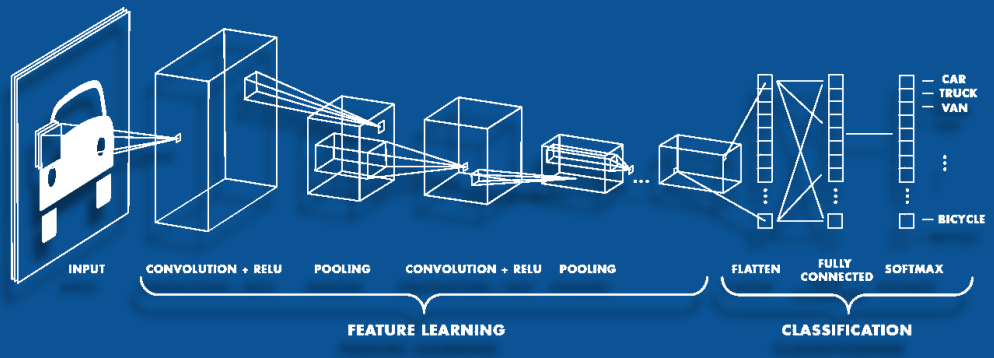




Introduction

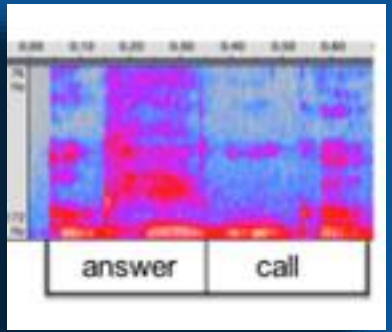
- Convolution Neural Network(ConvNet)
  - Popularity of ConvNets in computer vision
  - Large size - computation bounded, memory bounded
- Small ConvNet - SqueezeNet
  - Small convolution kernel
  - Removed Fully connected layer
- ConvNet in Embedded System
  - Novel Architecture for SoC FPGA
  - SqueezeNet like architecture





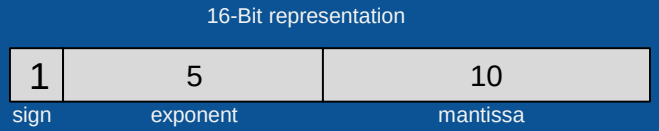
Applications

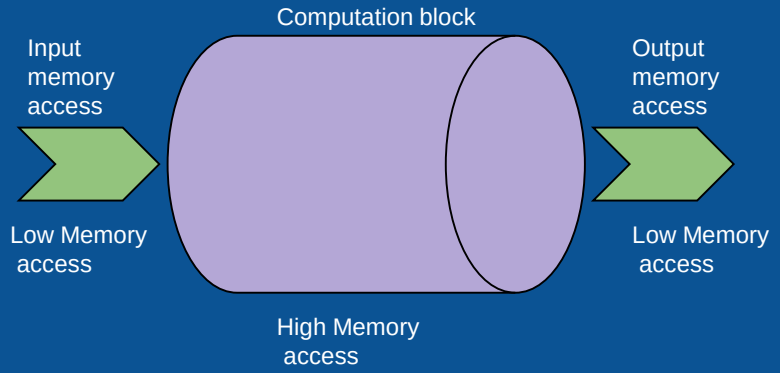
- Computer vision tasks - classification, detection, segmentation
  - Autonomous vehicle
  - Drones
  - Robots
  - Face recognition
  - Surveillance camera
- Signal processing
  - Voice command recognition
  - Sensor data processing



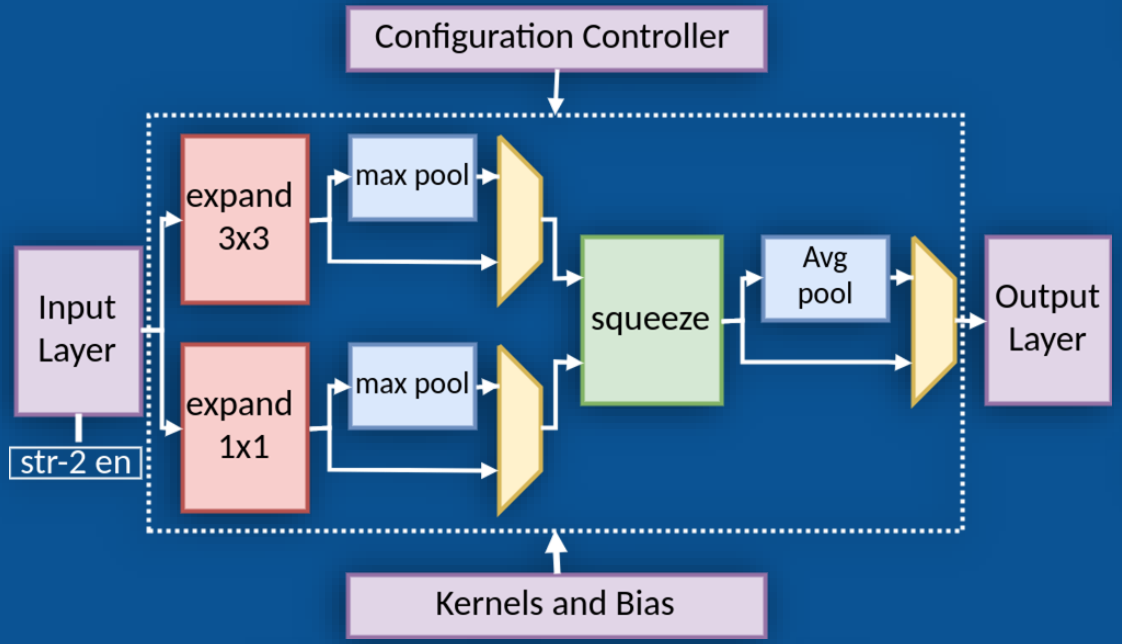
Design Strategies

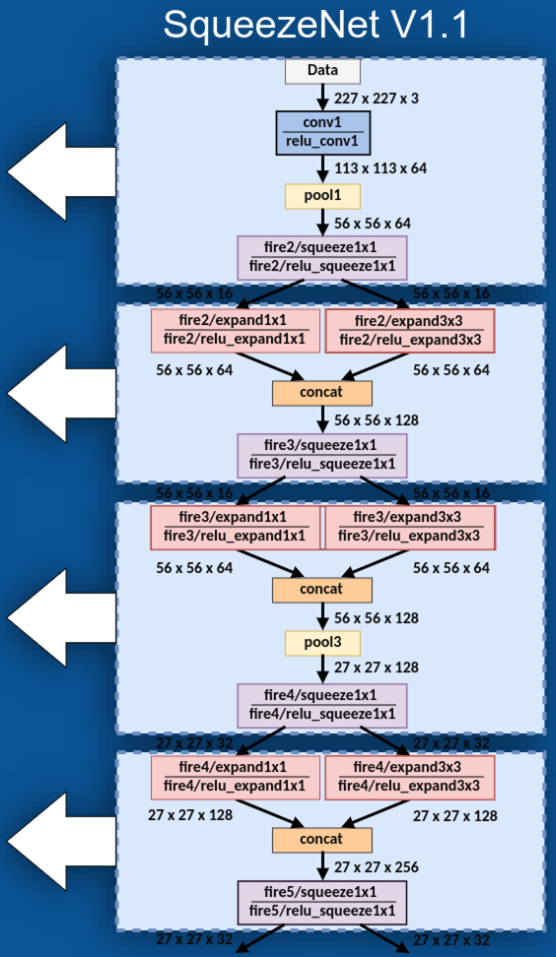
- Resource Constraints
  - 16 bit float number
  - Fixed point operation
  - Blockwise execution Computation blocks
- Memory access constrain
  - Intermediate output stored in block ram
  - Low memory access for computation block



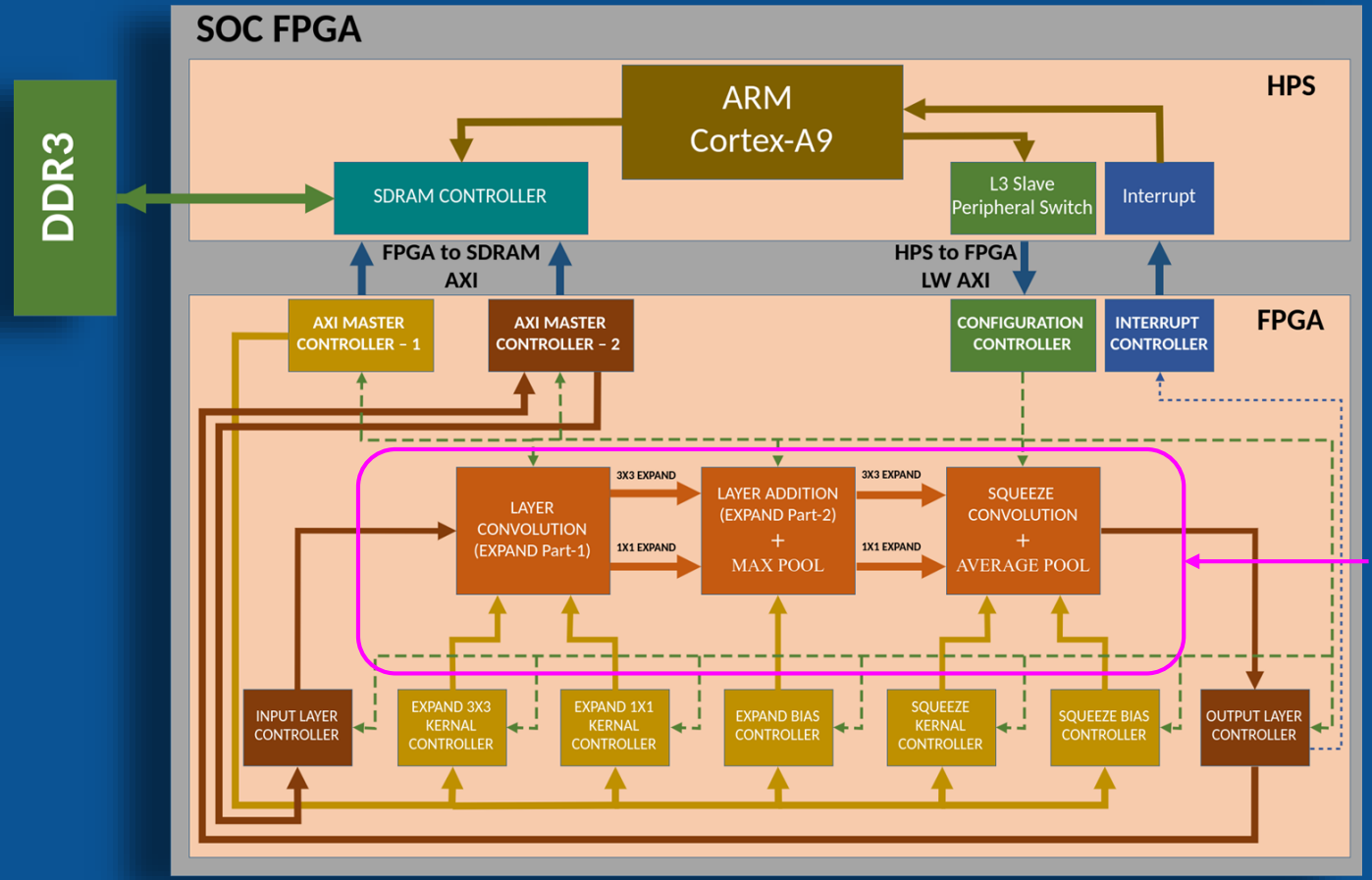


Computation Block





Overall Architecture

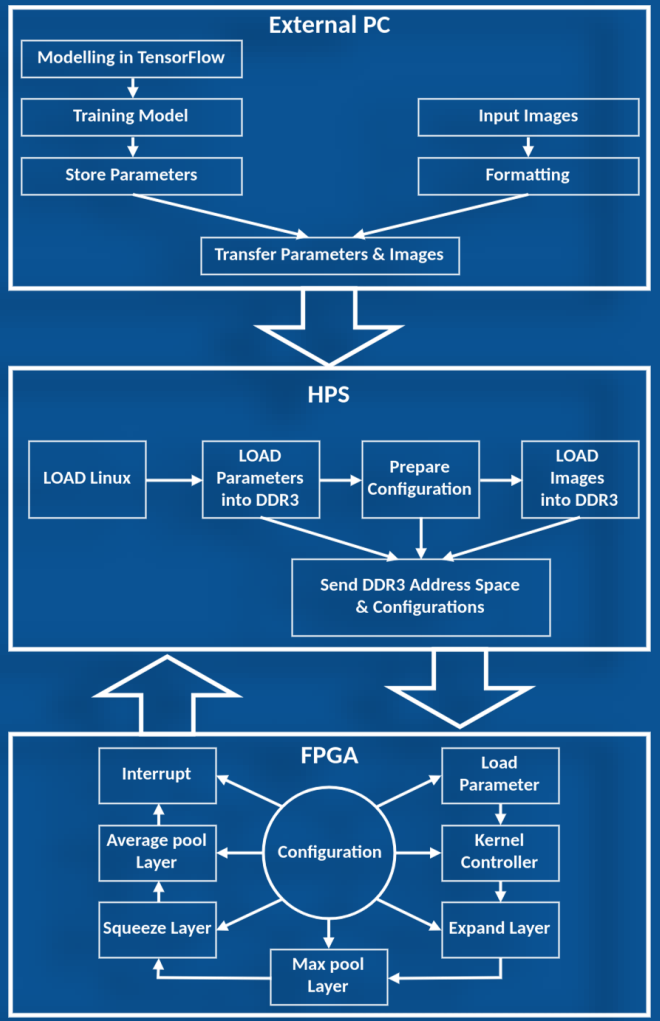


Key Features of Architecture

- Compact size
- Scalable and Extendable Design
- Higher performance
- Cost effective
- Power efficient
- Data privacy
- Low network bandwidth

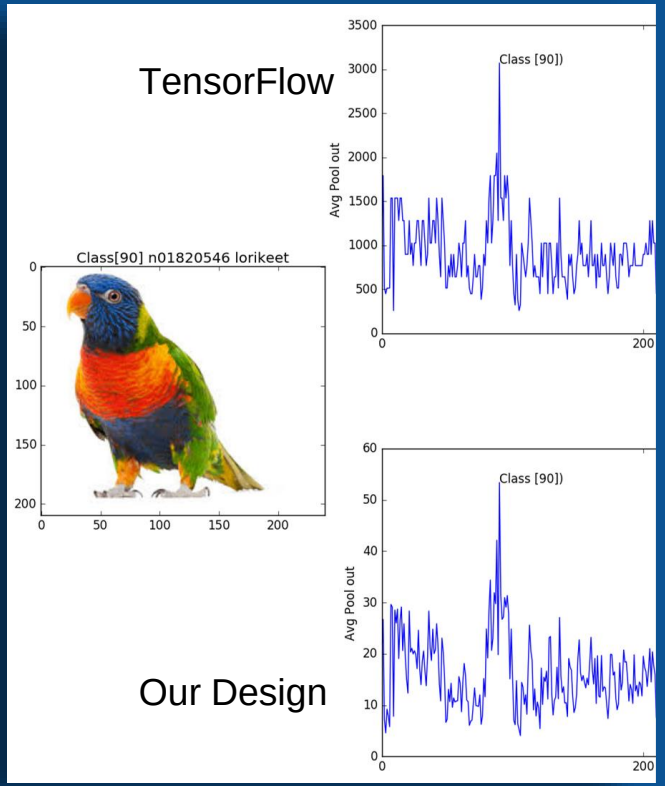
Functional Flow

- External PC
  - Training model
  - Converting parameters to custom representation
  - Storing parameters in HPS system
- HPS system
  - Loading parameters in memory
  - Generate configuration
  - Configure hardware
  - Loading input
- FPGA system
  - Running Inference
  - Writing output to memory
  - Interrupt generation



Implementation SqueezeNet v1.1

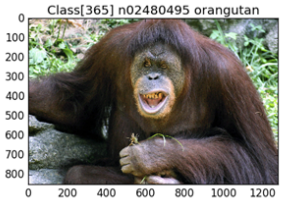
- Performance
  - Runtime : 9 FPS @ 100MHz
  - Power : 2W
  - Memory Bandwidth: 142 MB/s
- Resource Utilization
  - Logics : 14K
  - Registers : 19K
  - DSP : 56 Blocks
  - Block RAM : 1.3 MBit

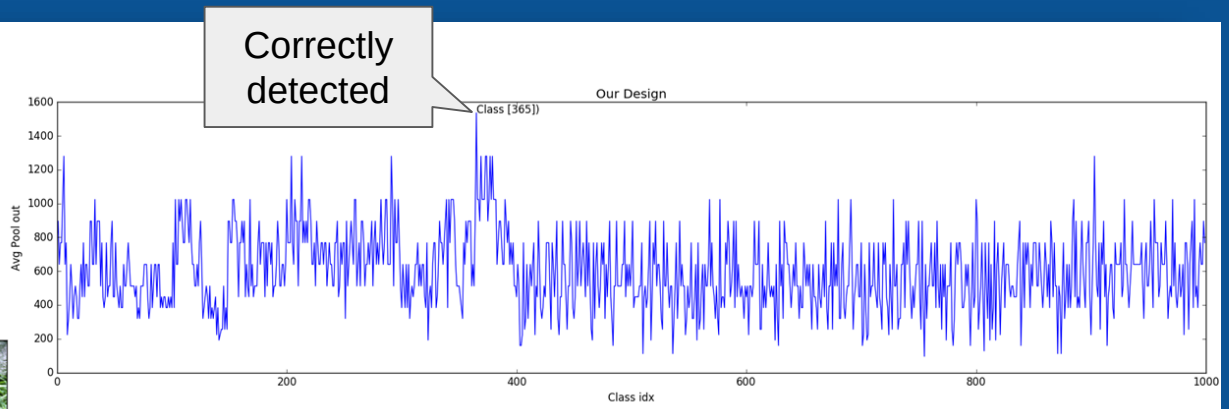


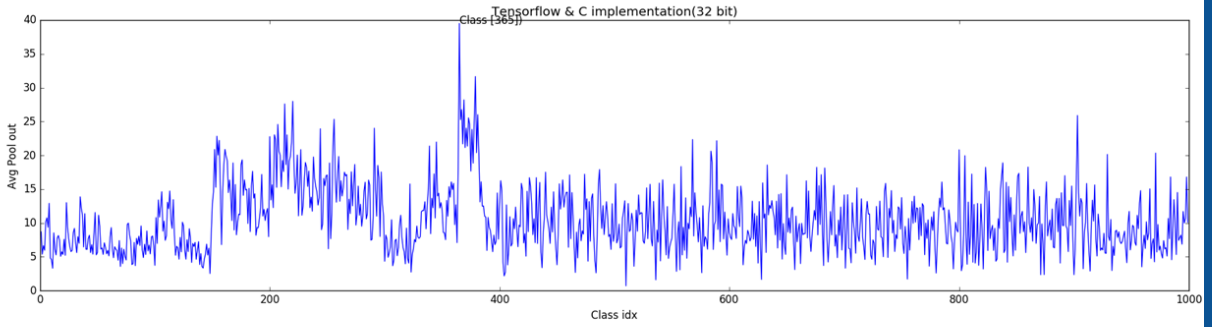
Implementation SqueezeNet v1.1

Our Design

TensorFlow







Comparison

Design	Performance	Power	Feature
Our Design	9 FPS	2W	100MHz clock
Raspberry Pi 3 (Caffe)	3-5 FPS	3.5W	~1GHz, VideoCore IV GPU(1GB LPDDR2 @900MHz)
GTX 1080 Ti (Caffe)	~100 FPS	280W	11.3 TFLOPS
Mobile: Snapdragon 820 (without GPU)	~13 FPS	N/A	Quad core ~ (2x 2.15GHz and 2 x 1.67GHz)
Mobile: Snapdragon 820 (with GPU)	~21 FPS	2.5W	Quad core ~ (2x 2.15GHz and 2 x 1.67GHz)