Bespoke:

A Framework for Rapid Development and Deployment of ML Stream Processors on FPGA Systems

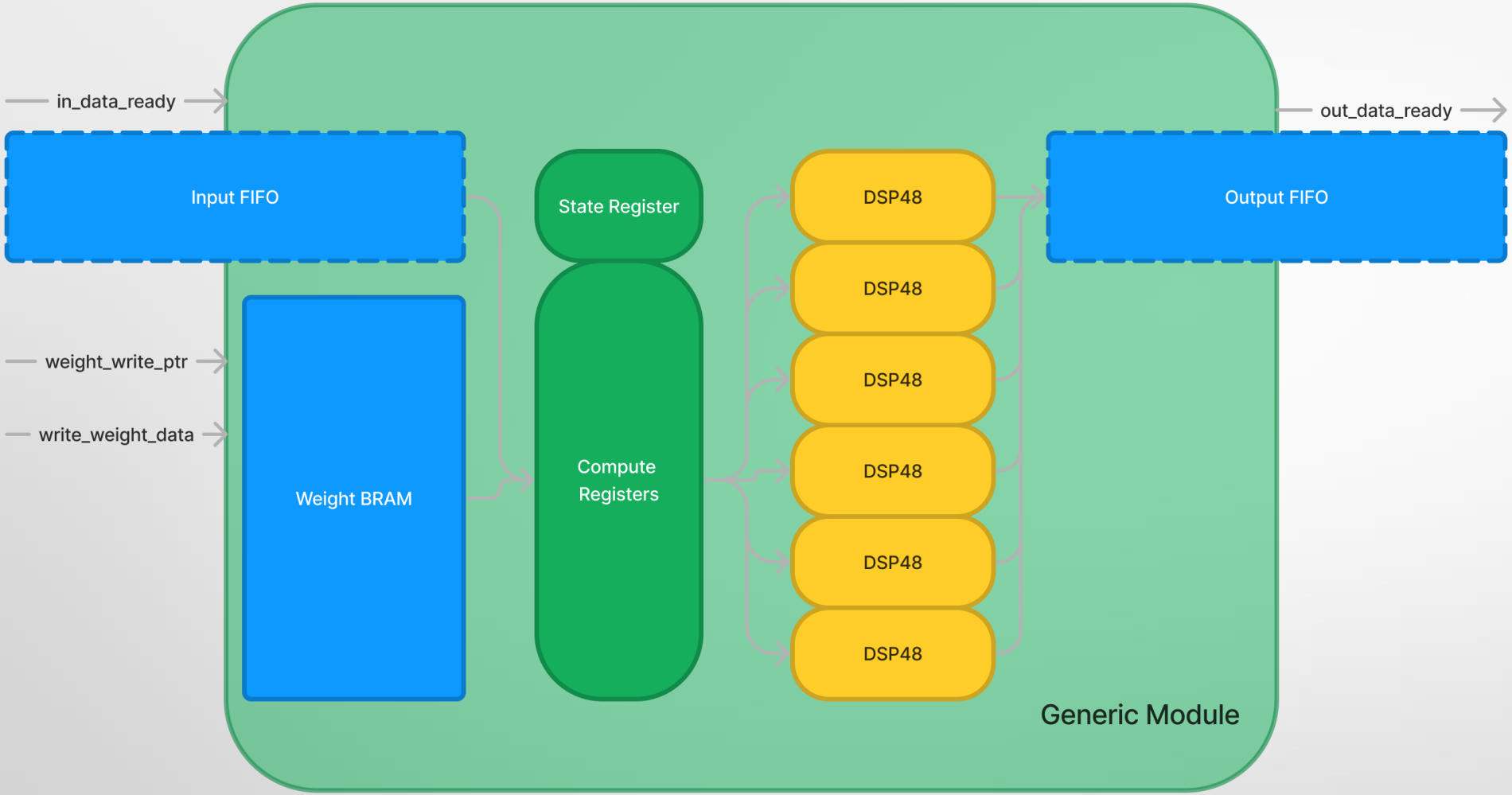
6.205 Final Project Proposal Presentation by Thelonious Cooper

(Team 34)

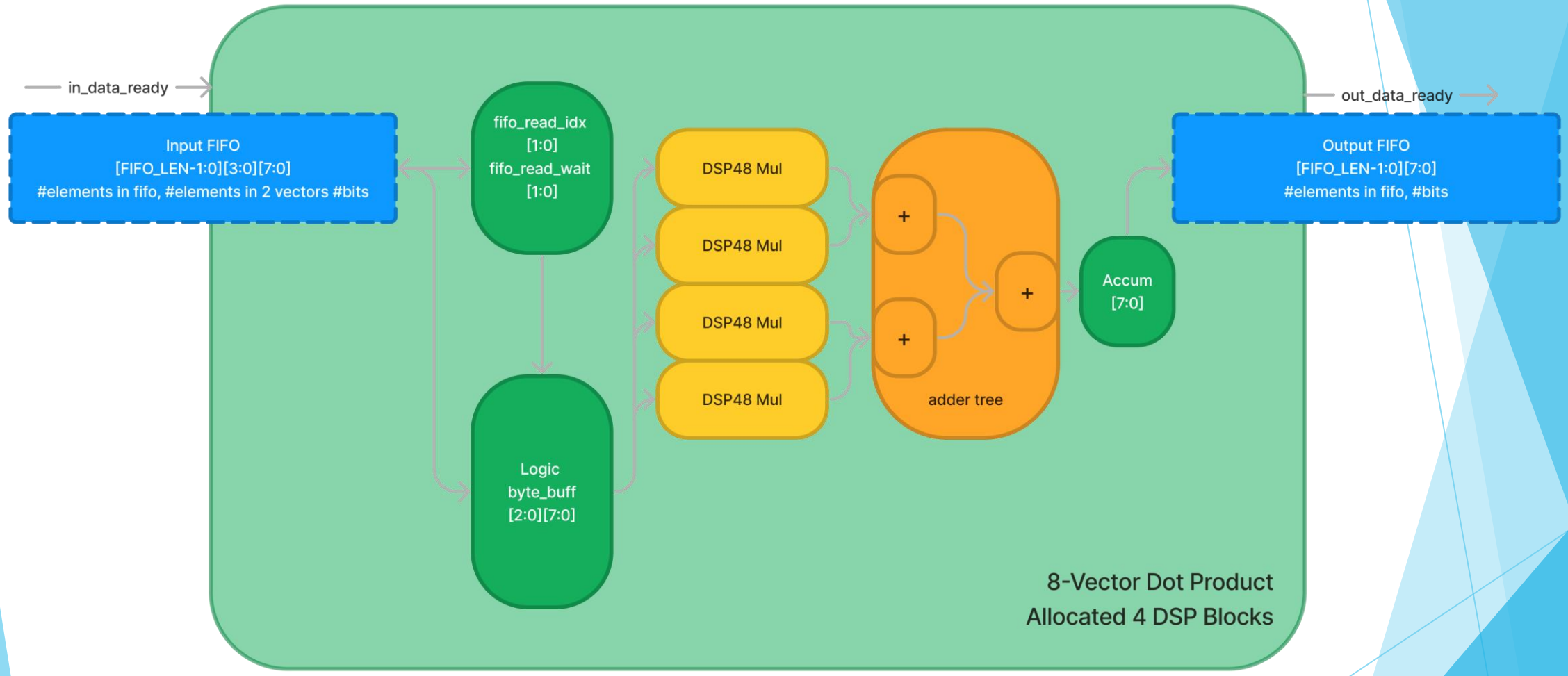
Project Goals:

- ML Library
  - Vector Operations: Vector-Dot Product, Vector Addition, Scalar Multiplication
  - Matrix Operations: Matrix-Vector Product, 1d/2d convolution
  - Activation Functions: Tanh, ReLU, ReLU6, LeakyReLU
- Dynamically link ML functions according to specification formats (ONNX)
- Demo of MLP stream processor on MEMS Flow Sensor

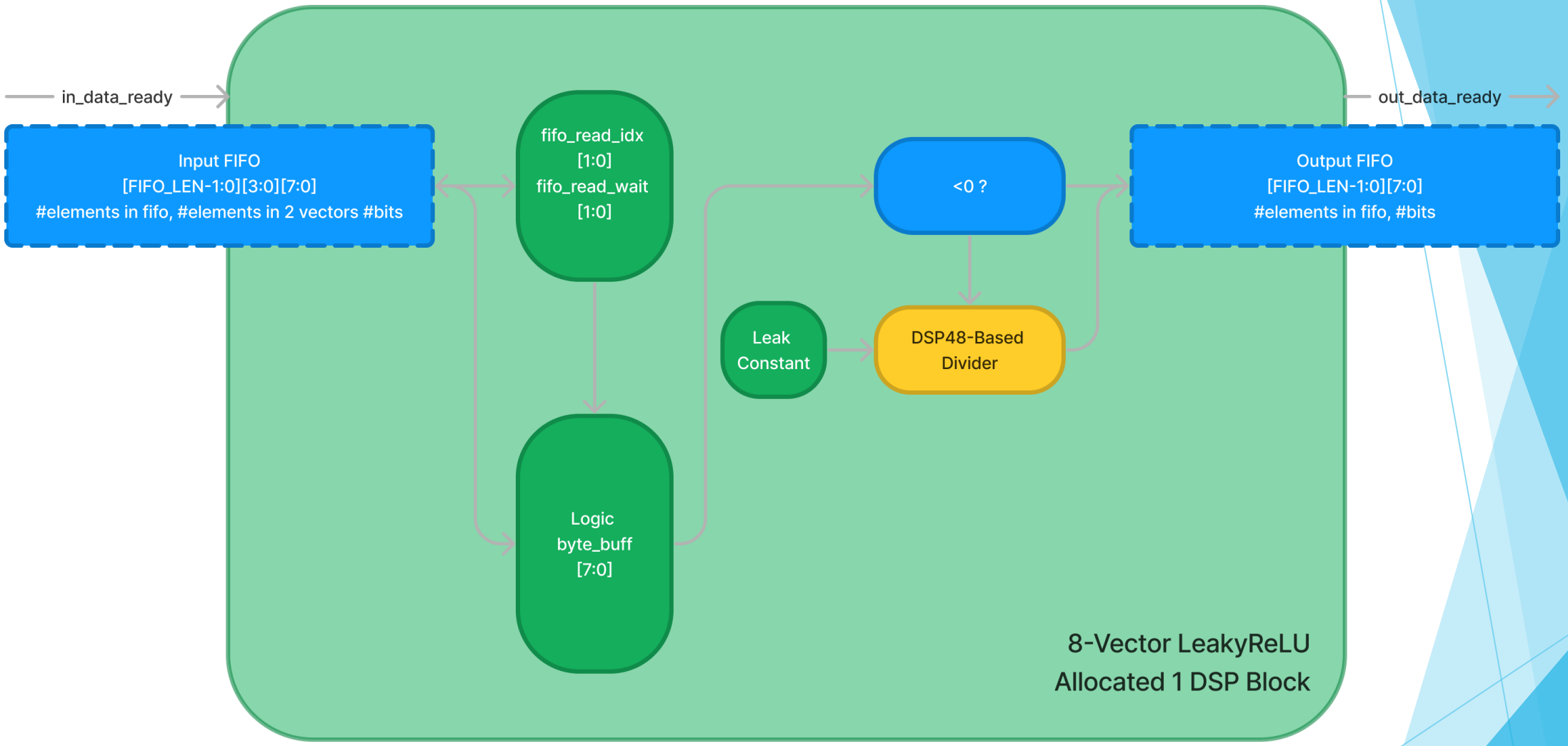
ML Stream Processor Library Component Architecture



Generic ML Module Architecture

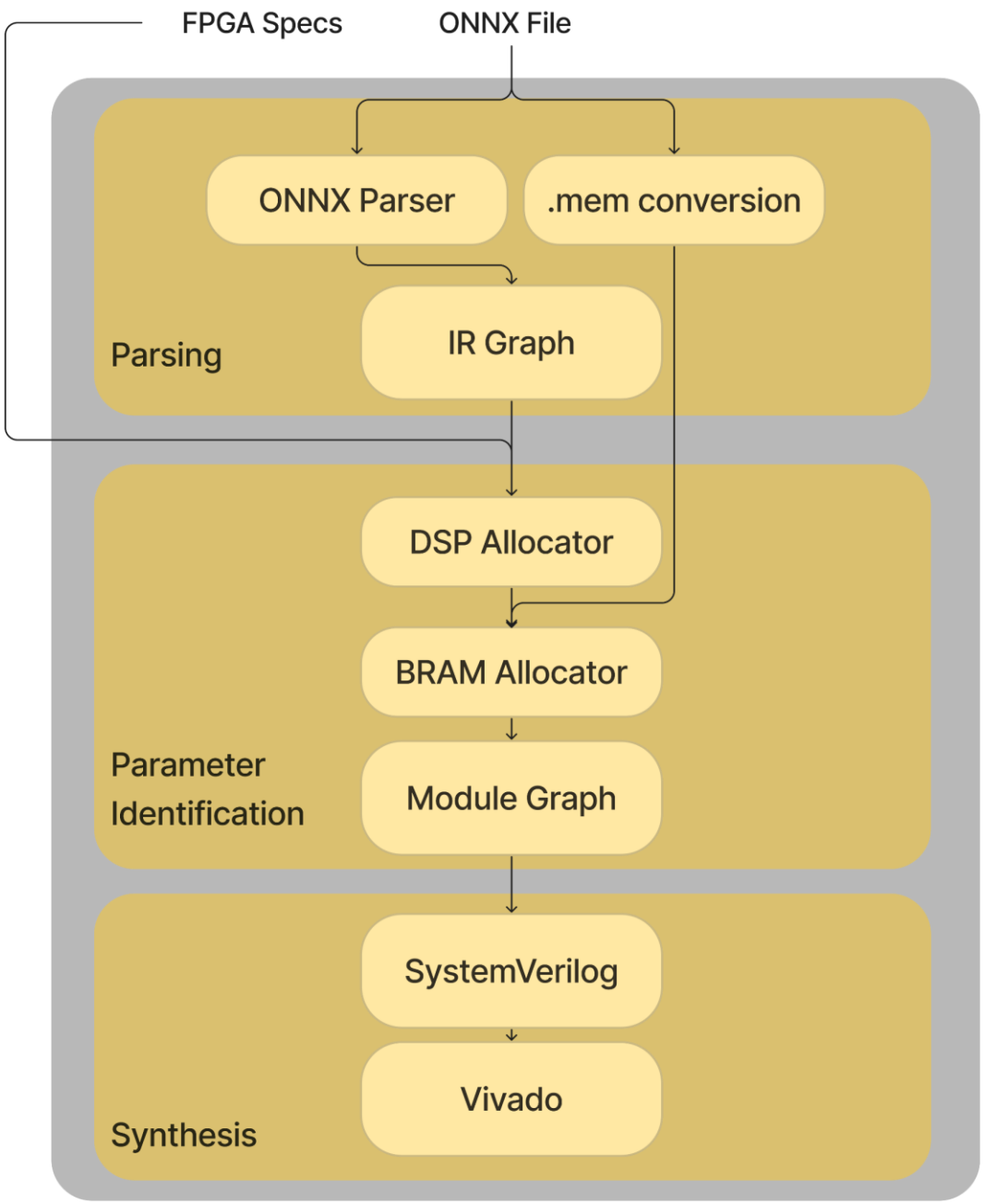


Example particular module architecture
8-Vector Dot Product with 4 DSPs



Example particular module architecture
Vector LeakyReLU 1 DSP Block

Dynamic Model Router and Compiler Architecture



Compiler python script architecture

Project Evaluation

- Test bench and in-situ comparison to PyTorch reference implementation
- Network 8-bit quantization efficacy evaluation
- Multi-scale efficiency reports
  - Critical path delay
  - Throughput

Timeline

Nov 13

Basic Operations:
- Matrix product, ReLU
- .sv unit tests

Nov 20

Top-level .sv manual integration test
- .sv test
- In-situ test

Nov 27

Compiler:
- ONNX Parser
- Router
- BRAM Allocator
- File Generator

Dec 4

.sv generated Integration Test

Dec 11

In-Situ integration test

Dec 18

Real-world model implementation tests

Stretch goals

- Larger library of components
  - Convolution
  - Tanh activation
- Support for Recurrent neural networks
  - LSTM
  - IIR-Filter

Deliverables

- Minimal:
  - MAST network implementation: matrix-vector mult and ReLU
- Expected:
  - Dynamic model parsing and routing, demonstrated on several networks of varying sizes in simulation
- Stretch:
  - Convolutional networks
  - Recurrent networks
- A+:
  - Recurrent conv-nets for image stream-processing