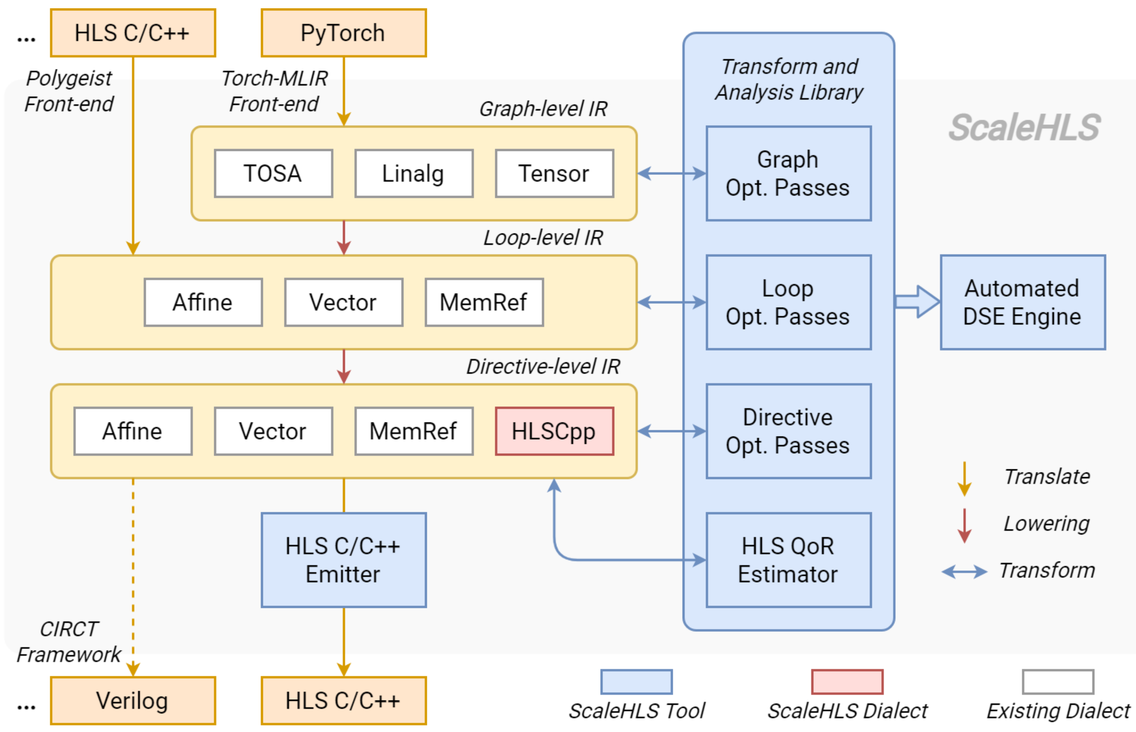
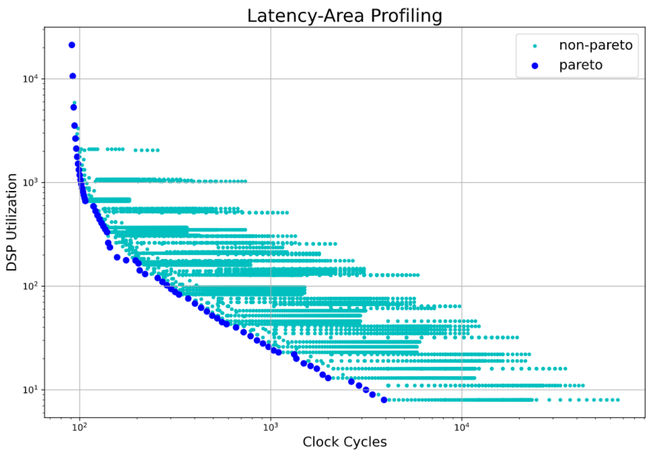
StreamTensor: A Compiler from PyTorch to FPGA for AI/ML Applications

Hanchen Ye[12] (hanchen8@illinois.edu) and Deming Chen[12] (dchen@illinois.edu)
[1]*University of Illinois Urbana-Champaign* [2]*Inspirit IoT*

ScaleHLS: Single-kernel Optimization (2022)

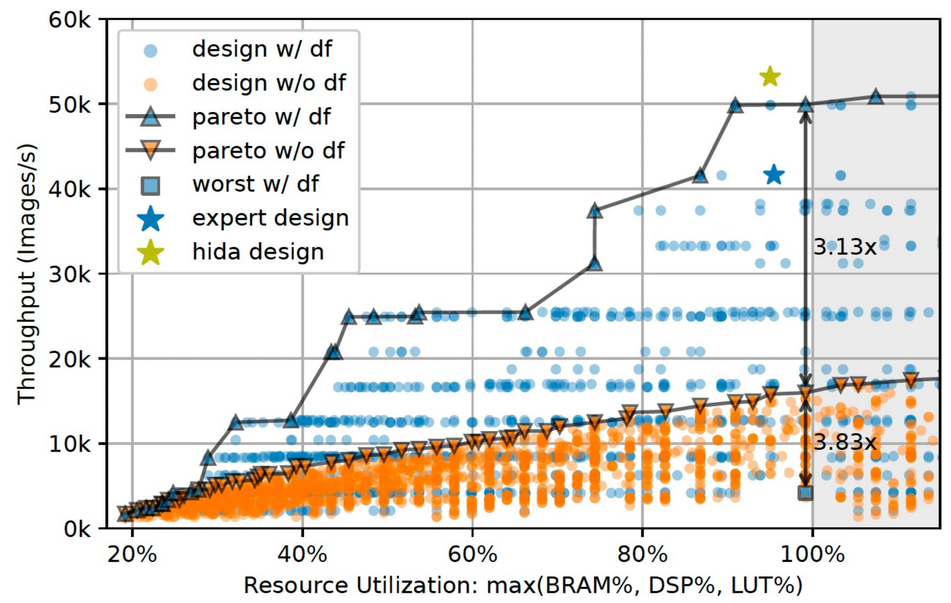




ScaleHLS Features

- C++/PyTorch frontend
- MLIR-based compilation
- Multi-objective single kernel optimization
- Generate near-optimal single kernel HLS design given resource constraints

HIDA: Multi-kernel Optimization (2024)



Multi-kernel Dataflow* Optimization Matters! 🤔⬇️

Inter-kernel X Intra-kernel Optimization

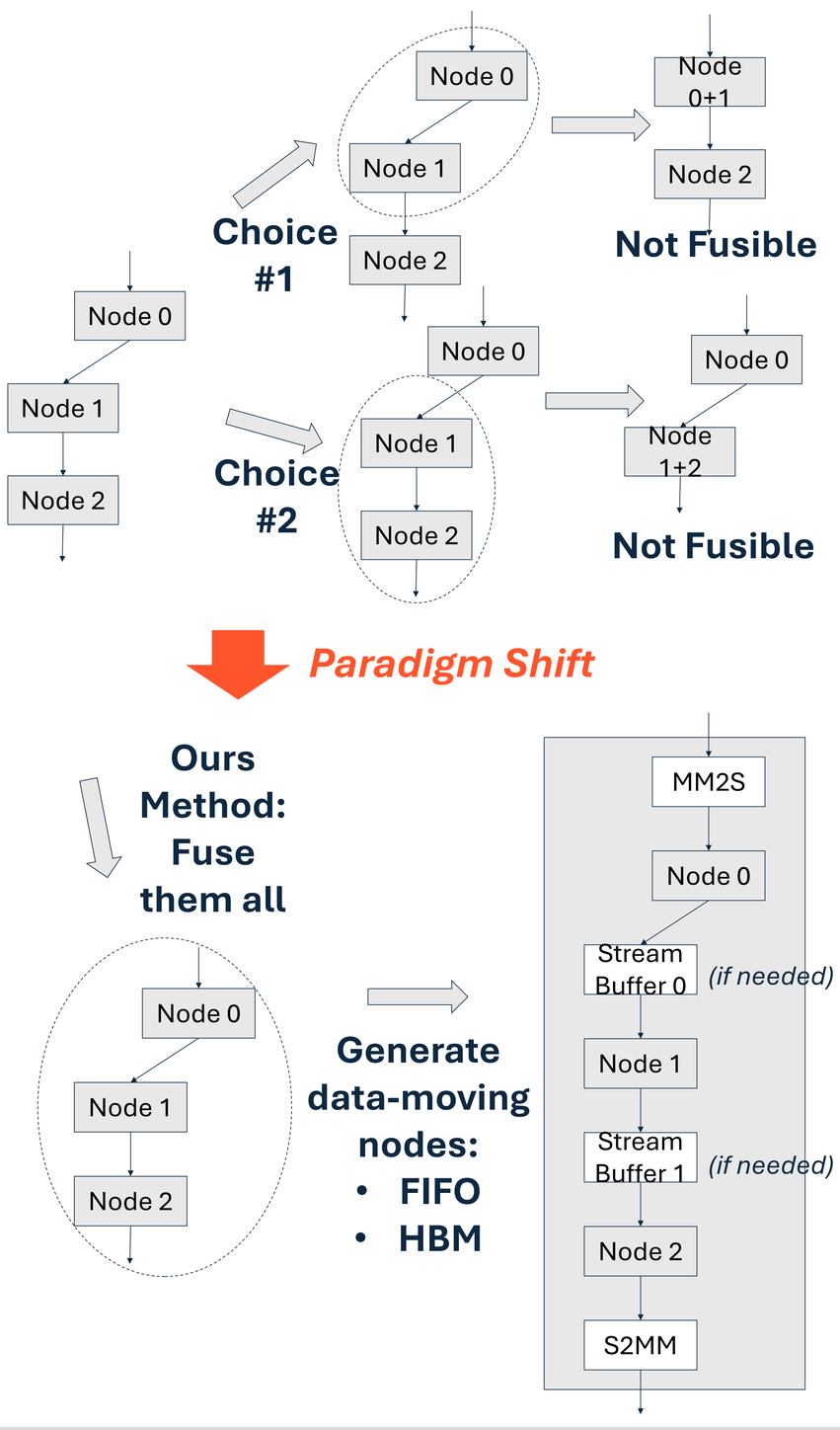
*Dataflow = Coarse-grained Pipeline between Kernels

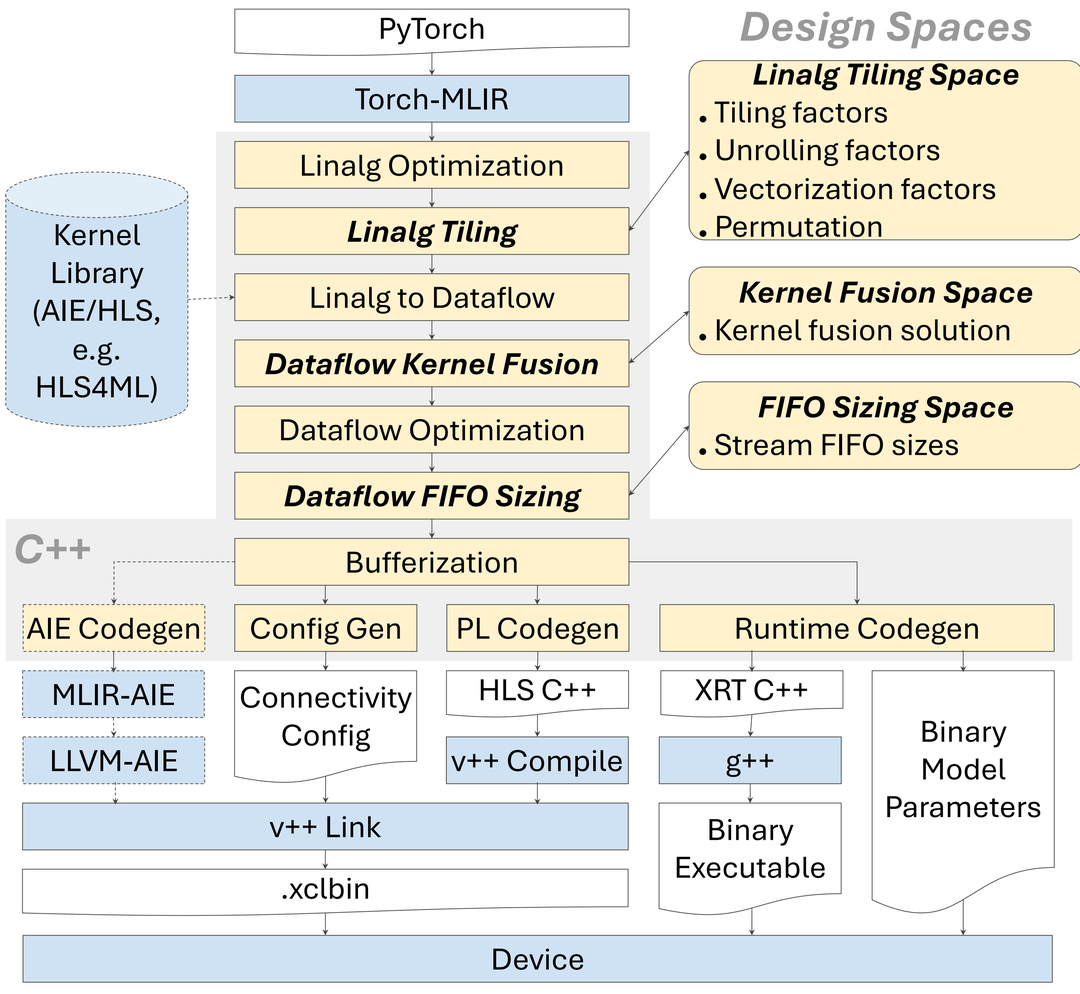
Model	Throughput (Samples/s)*		
	HIDA	DNNBuilder [75]	ScaleHLS [68]
ResNet-18	45.4	-	3.3 (13.88×)
MobileNet	137.4	-	15.4 (8.90×)
ZFNet	90.4	112.2 (0.81×)	-
VGG-16	48.3	27.7 (1.74×)	6.9 (6.99×)
YOLO	33.7	22.1 (1.52×)	-
MLP	938.9	-	152.6 (6.15×)
Geo. Mean		1.29×	8.54×

HIDA Features

- C++/PyTorch frontend
- Two-level dataflow IR:
  - Functional dataflow
  - Structural dataflow
- Intensity and connectivity aware global multi-kernel optimization
- Better performance than RTL accelerators

StreamTensor: Streaming-based Multi-kernel Dataflow (Now)





A new solution for FPGA users:
Software-like compile experience for AI/ML applications

- **End-to-end PyTorch-to-device flow validated on FPGA BOARD**
- **Aggressive kernel fusion with streaming**
  - Single kernel for a single 🤗 Mistral-7b or Llama-8b decoder layer on U55C FPGA
- **Automated data movement generation**
  - DMA, data packing, and interface widening for HBMs
  - Data vectorization, data layout conversion, and FIFO sizing for streaming FIFOs
- **Pythonic user-level API**
  - Python API for passes and design spaces; auto-tuning-ready