Automated and Holistic Co-design of Smart Readout ASICs With Embedded Machine Learning

Brookhaven National Laboratory

U.S. Department of ENERGY

Presenter: Shubha Raj Kharel

Team:

Dr Shubha Raj Kharel, Mr Dominik Gorni, Dr Grzegorz Deptuch, Dr Piotr Maj , Dr Prashansa Mukim, Dr Shinjae Woo, Dr Yihui Ren, Dr Soumyajit Mandal,

@BrookhavenLab

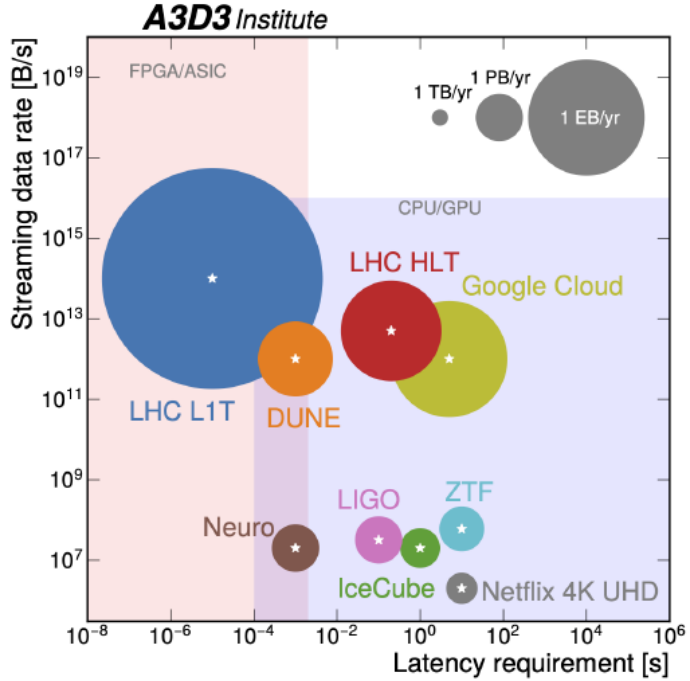
Date: 11/21/2024

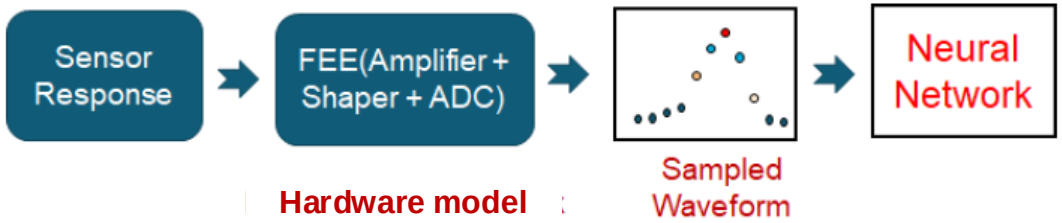
Introduction

- Overall goal: Explore non-linear digital signal processing (DSP) and machine learning (ML) algorithms as an efficient yet high performance alternative to analog circuits for real-time on-chip waveform processing.

- These algorithms will be implemented on highly-digital readout ASICs for a new generation of “smart” detectors with built-in amplitude estimation, pile-up detection/correction, and pulse shape classification capabilities for X-ray spectroscopy and other applications.

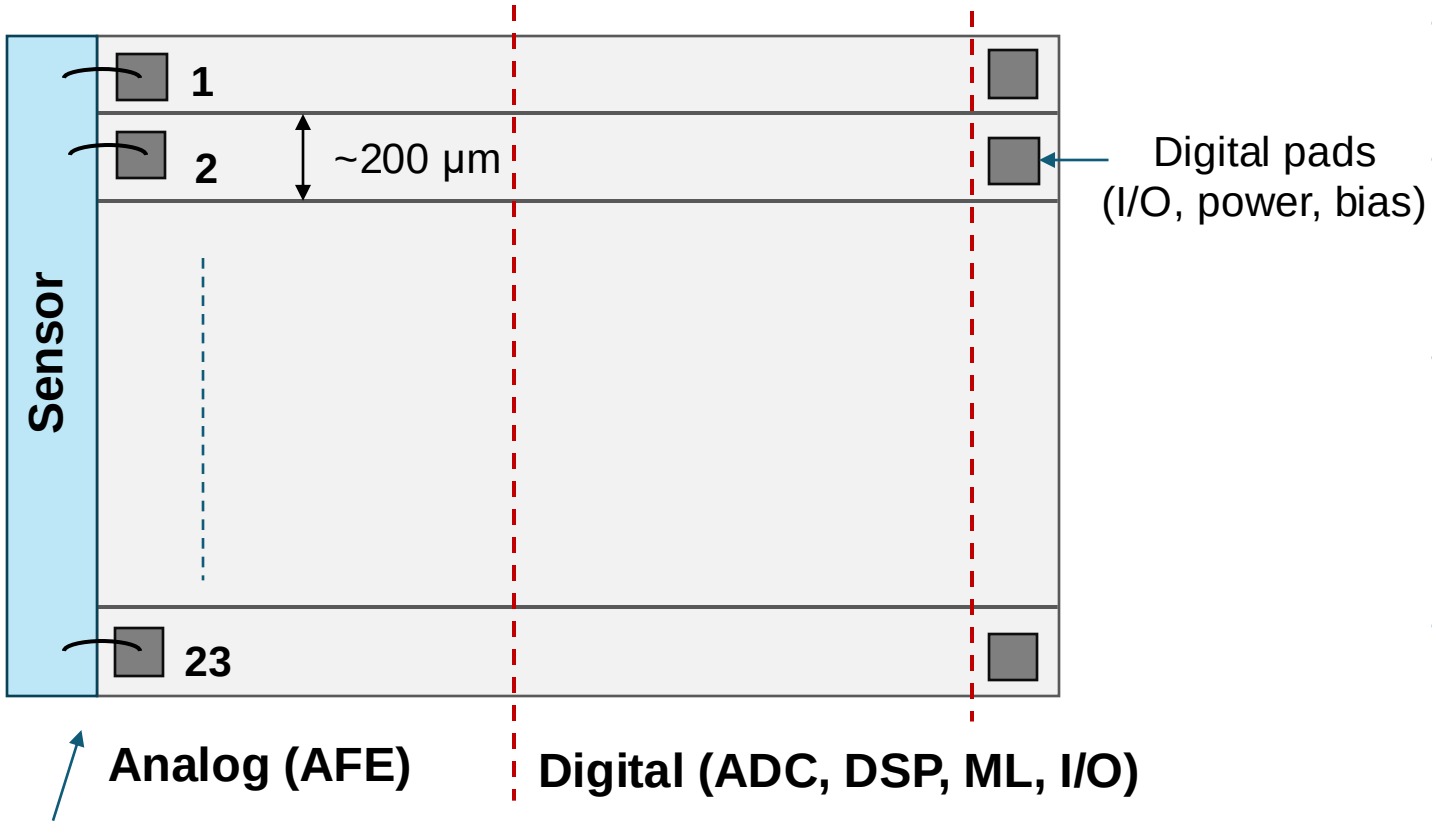
- These capabilities can be utilized to reduce output data rates, enable real-time adaptation of the detector, and provide distributed “edge intelligence’’ for improved physics outcomes [0].





[0] Carini, Gabriella, et al. "Smart sensors using artificial intelligence for on-detector electronics and ASICs.“ *arXiv preprint arXiv:2204.13223* (2022).

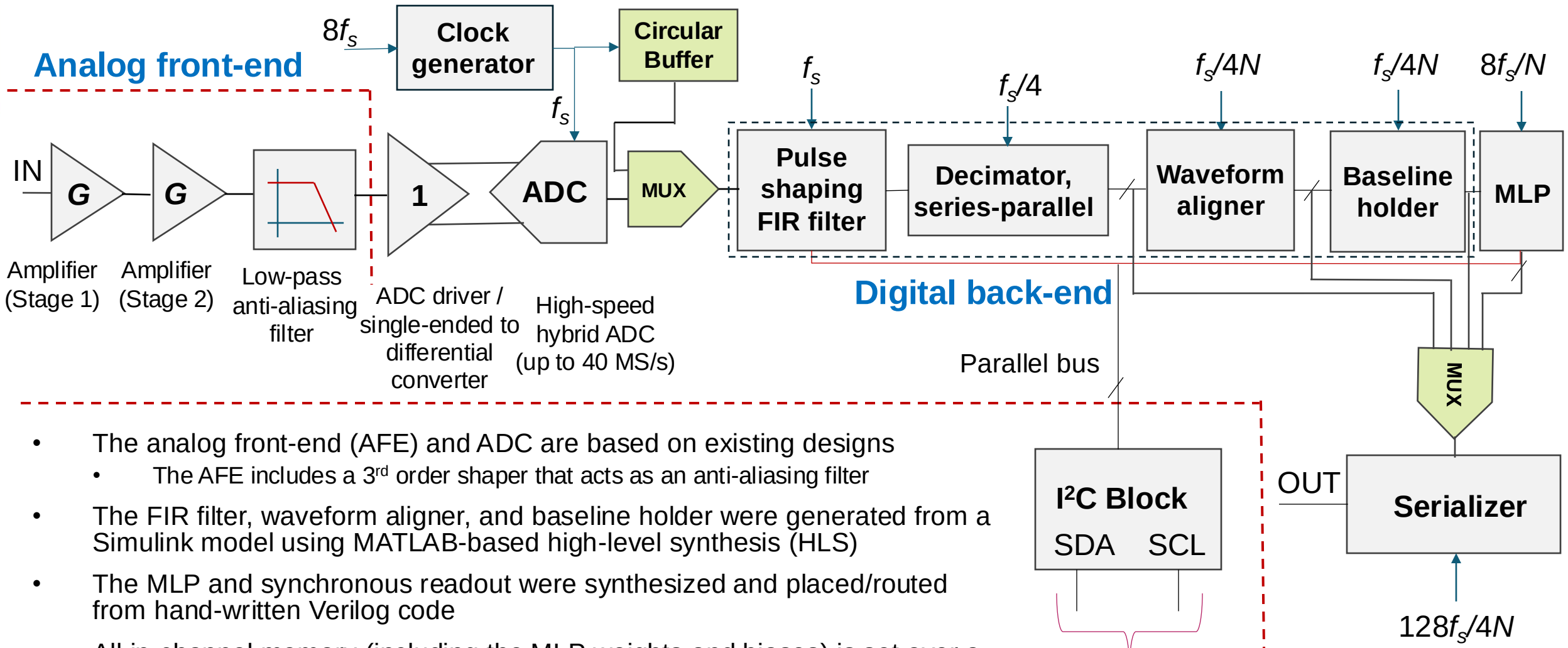
Overview of the proposed readout ASIC



Area: 5 x 5 mm^2
Channel height: 200 µm

- Simplified AFE design (charge-sensitive amp + anti-aliasing filter)
- Per-channel 12-bit ADC generates inputs for on-chip DSP and ML
  - Fully-differential, includes calibration
- Per-channel DSP:
  - Programmable shaping filter (FIR)
  - Waveform alignment / snippeting
  - Baseline removal
- Per-channel machine learning (ML):
  - Artificial neural network (ANN): multi-layer perceptron (MLP) with programmable weights
  - Performs classification or regression
- Additional blocks:
  - Low-speed I^{2}C programming and testability interface
  - Output serializer

Simplified view of the channel design

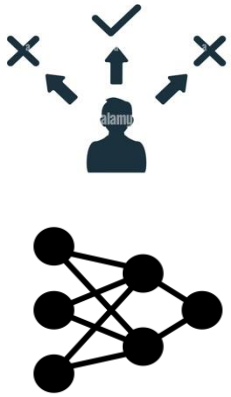


- The analog front-end (AFE) and ADC are based on existing designs
  - The AFE includes a 3rd order shaper that acts as an anti-aliasing filter
- The FIR filter, waveform aligner, and baseline holder were generated from a Simulink model using MATLAB-based high-level synthesis (HLS)
- The MLP and synchronous readout were synthesized and placed/routed from hand-written Verilog code
- All in-channel memory (including the MLP weights and biases) is set over a standard two-wire I²C interface

Brookhaven National Laboratory

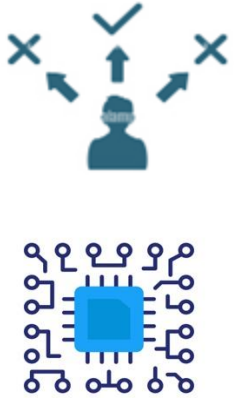
Motivation: Finding best for both big worlds

NN design choices



Effective AI

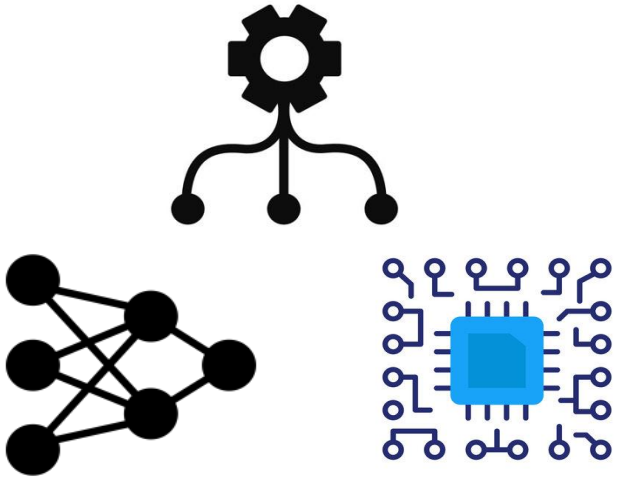
ASIC design choices



Efficient ASIC

Design Problem

1. Competing Goals: accuracy, area, power, delay
2. Choices available > Atoms in universe

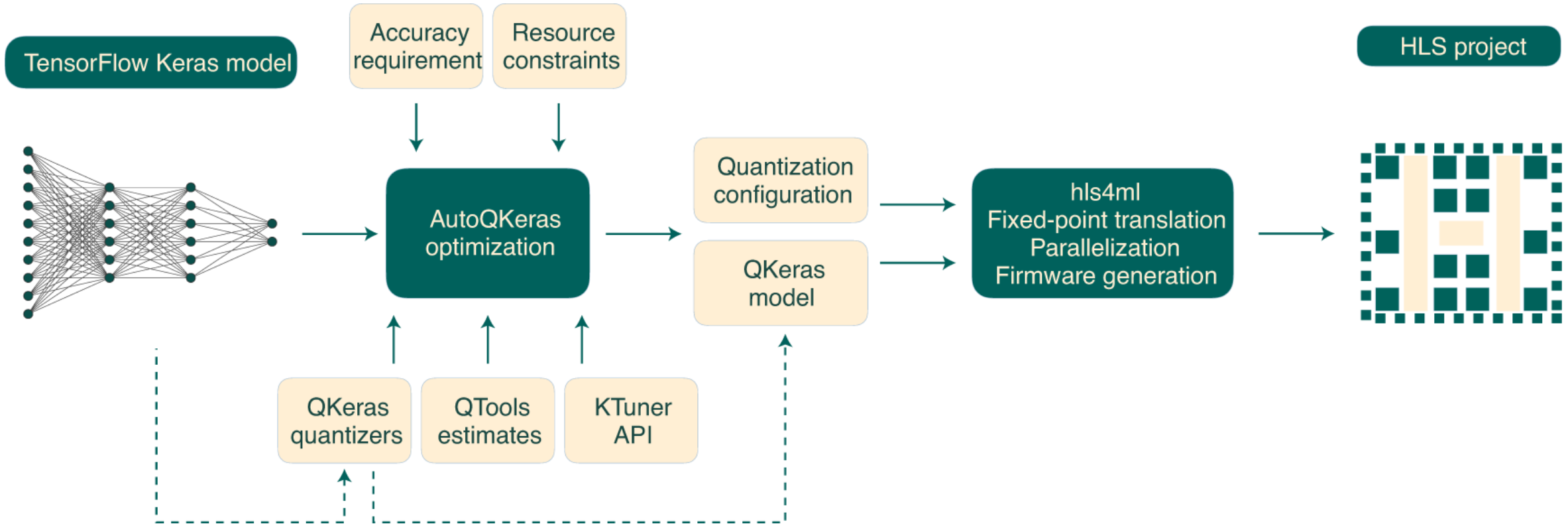


Automation Problem

1. Costly to compute the impact of design choice
2. Need to consider impacts of every choices so far to make new choice

Brookhaven National Laboratory

Existing Method



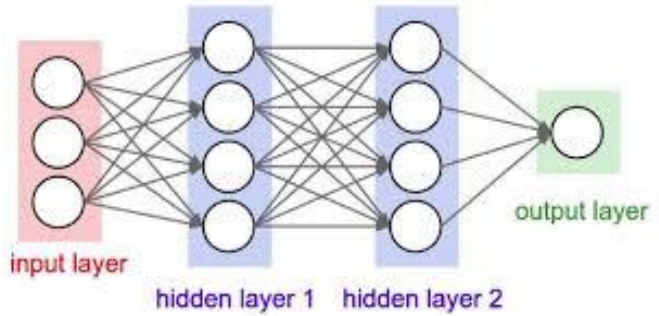
Automated neural architecture search with heterogenous quantization using theoretical energy estimate for

Brookhaven National Laboratory

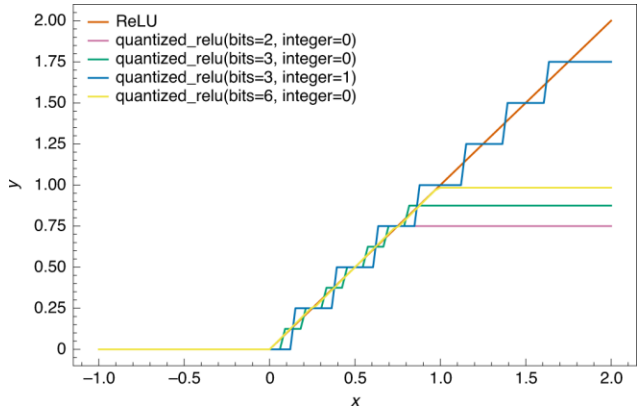
Coelho, C.N., Kuusela, A., Li, S. *et al.* Automatic heterogeneous quantization of deep neural networks for low-latency inference on the edge for particle detectors. *Nat Mach Intell* **3**, 675–686 (2021). https://doi.org/10.1038/s42256-021-00356-5

Methodology

Expanded Design Space



Neural Architecture



Heterogenous Quantization

```
synth_config:
  design_name: ${synthesis.design}
  file: ${synthesis.files.dir}/config.json
  json:
    PDK: ${synthesis.pdk}
    DESIGN_NAME: ${synthesis.design}
    VERILOG_FILES:
      - dir::${synthesis.files.name}
    SYNTH_STRATEGY: ${synthesis.openlane.stra
    CLOCK_PORT: clk
    CLOCK_PERIOD: ${synthesis.clock_period} #
    PL_RANDOM_GLB_PLACEMENT: true
    FP_SIZING: absolute
    DIE_AREA: 0 0 34.5 57.12
    FP_PDN_AUTO_ADJUST: false
    FP_PDN_VPITCH: 25
    FP_PDN_HPITCH: 25
    FP_PDN_VOFFSET: 5
    FP_PDN_HOFFSET: 5
    DIODE_INSERTION_STRATEGY: 3
    PL_TARGET_DENSITY: 0.75 # this is jus pla
    SYNTH_BUFFERING: ${synthesis.openlane.syn
    SYNTH_SIZING: ${synthesis.openlane.synth_
```

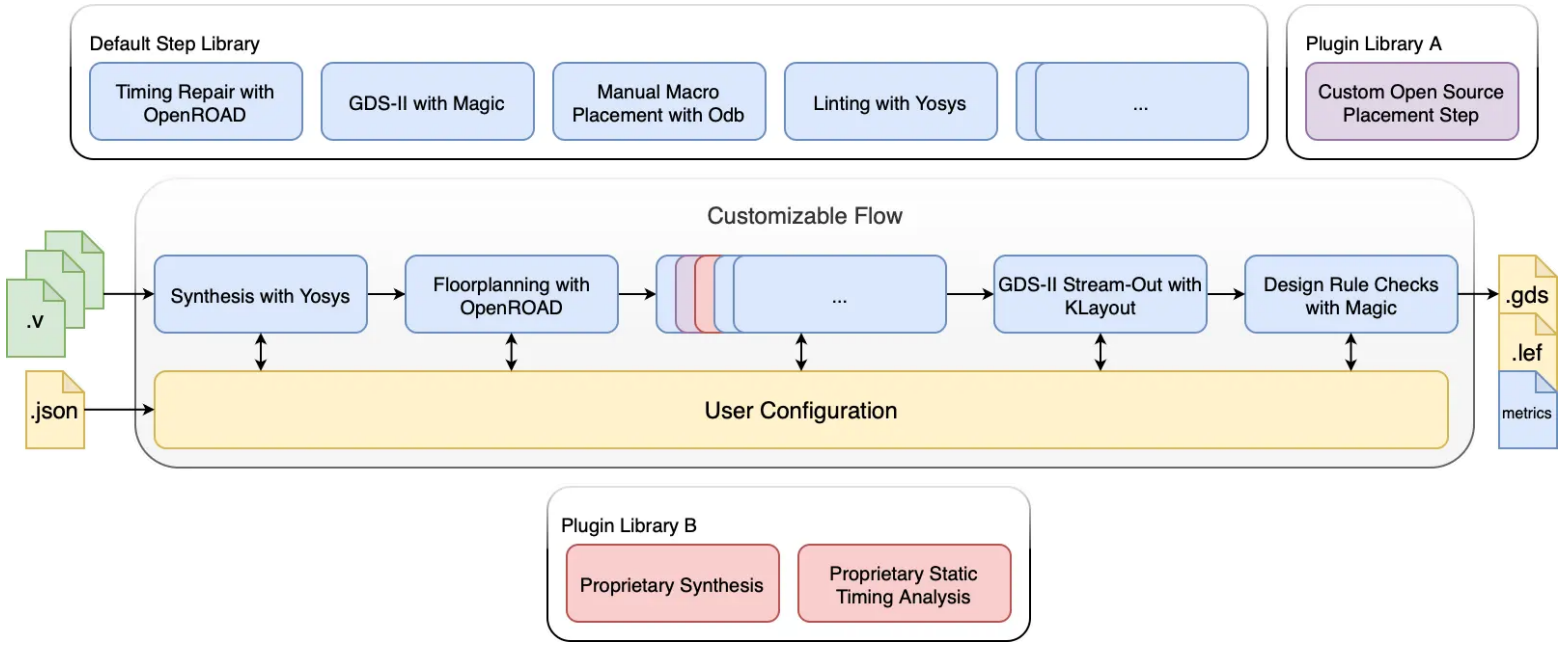
Synthesis Configuration

Co-design Choices	Range
Hidden Layers	1-3
Perceptrons in a layer	2-18
Weights/Biases Quantizations	2-16
Input/Output Quantizations	2-16
ASIC Synthesis Strategies	9 choices
Dimension of co-design space:	2,247,298,425

Exponentially increasing design space

Brookhaven National Laboratory

Synthesis informed objective



Openlane Pipeline

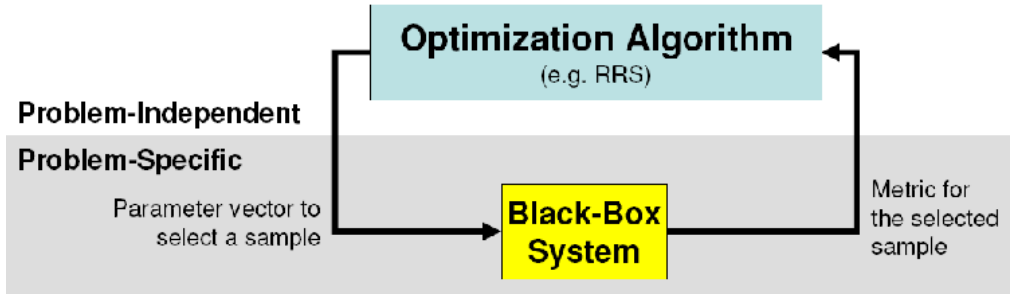
PDKs

FOSS 130nm Production PDK
github.com/google/skywater-pdk

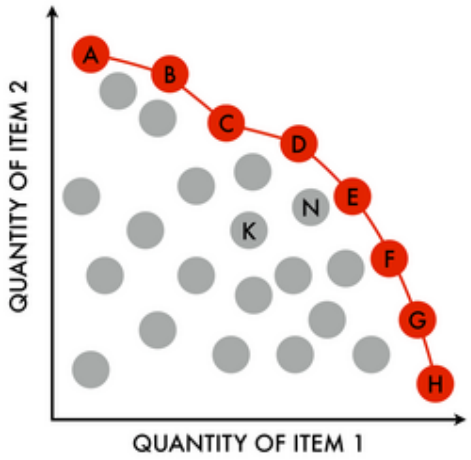
FOSS 180nm Production PDK
github.com/google/gf180mcu-pdk

Brookhaven National Laboratory

Muti-objective Bayesian Optimization



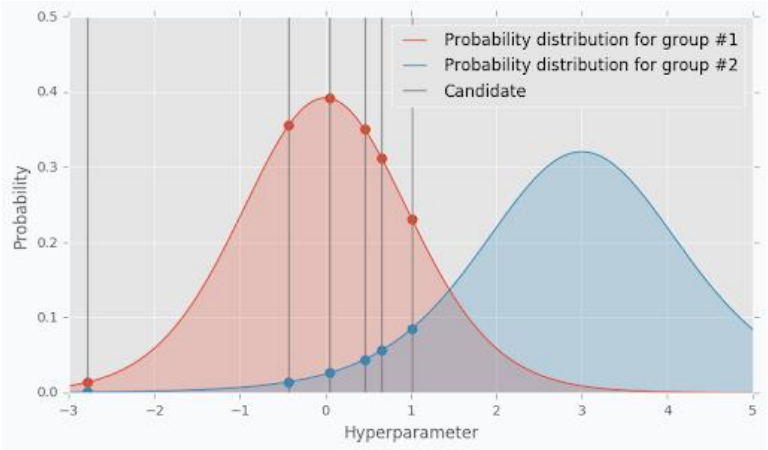
Black-box optimization



Pareto Front

All equally good solutions

Tree Parzen Estimators



How to get new promising candidates?

Idea: a promising candidate is likely to

- Have low probability under the bad distribution $P(x|\text{bad})$
- Have high probability under the good distribution $P(x|\text{good})$

⇒ **"Promisingness"** $\propto \frac{P(x|\text{good})}{P(x|\text{bad})}$

Tool Integration and automation

Goals

- Orchestrator design
- Containerization
- Configurability
- Distributed
- Flexible design space change
- Extensible

Brookhaven National Laboratory

Pyverilog

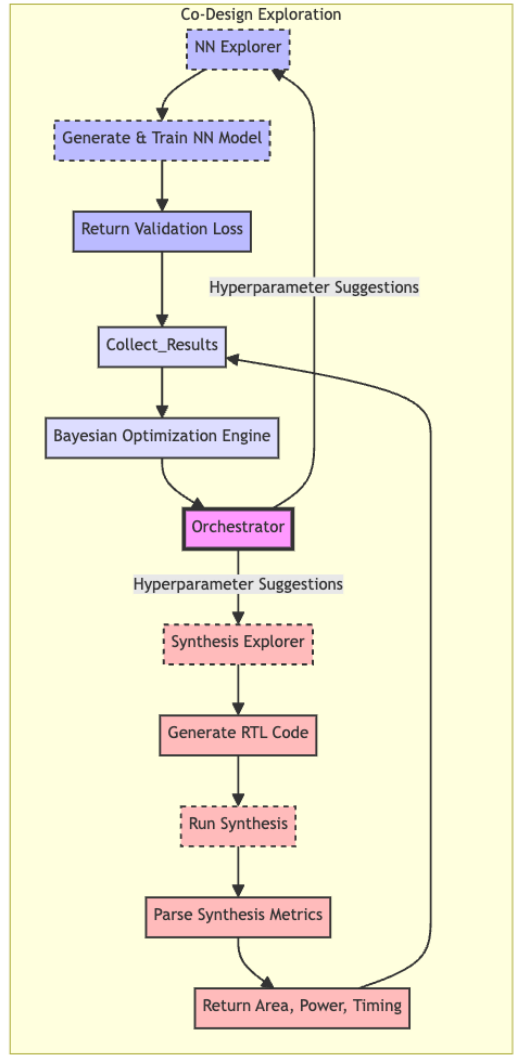
CI no status | build passing

Google + skywater

FOSS 130nm Production PDK
github.com/google/skywater-pdk

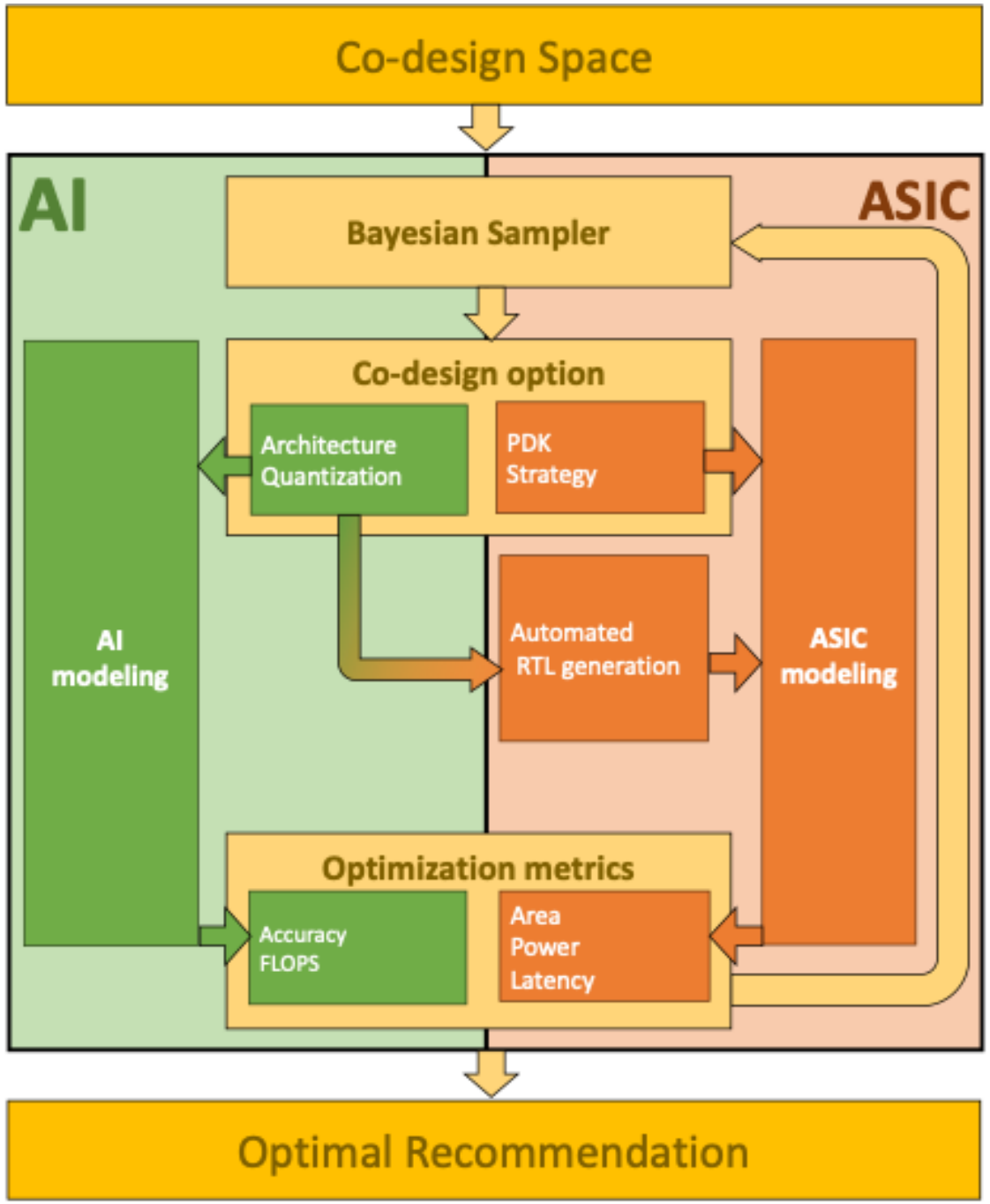
OPTUNA

Tools



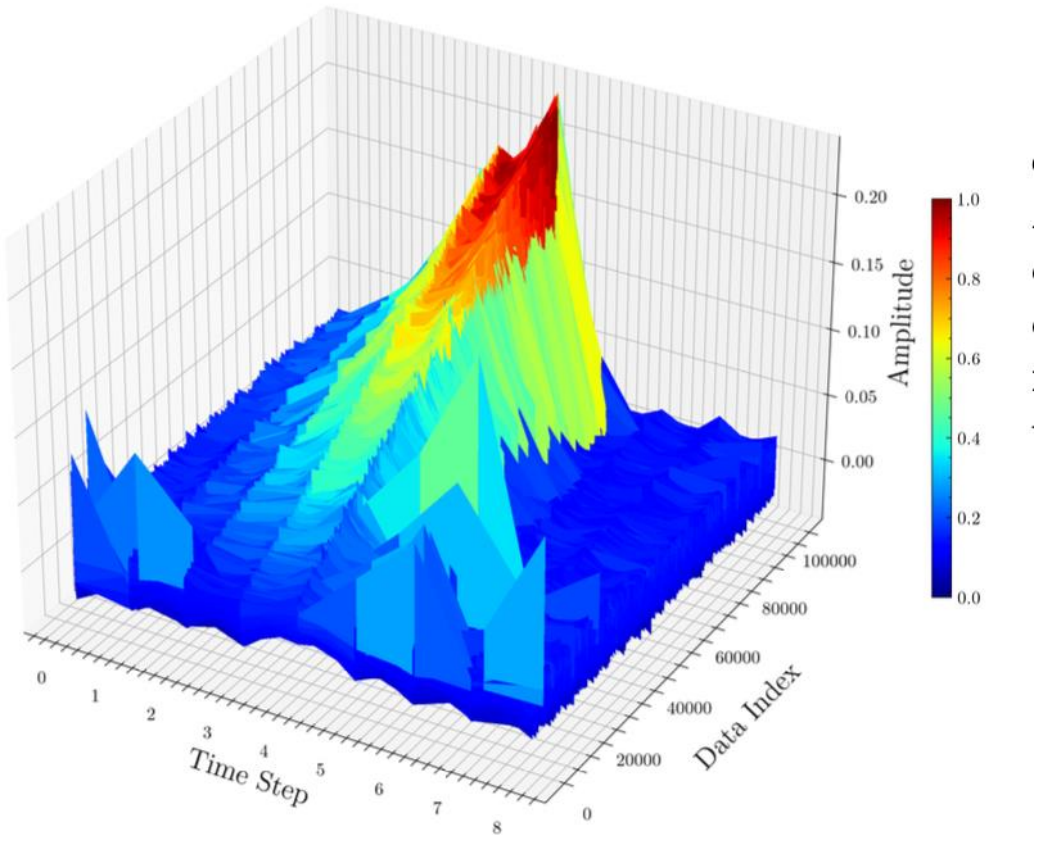
Integration

Full Pipeline



Experiment

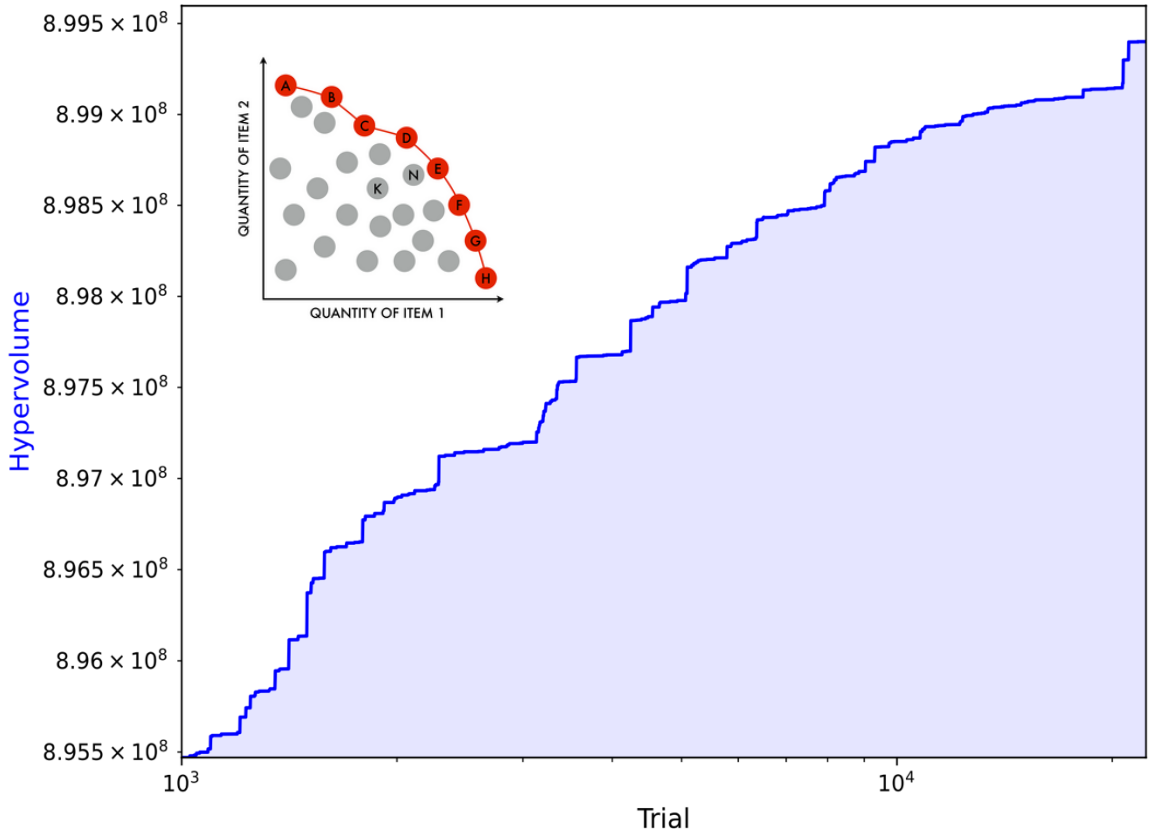
Waveform Data



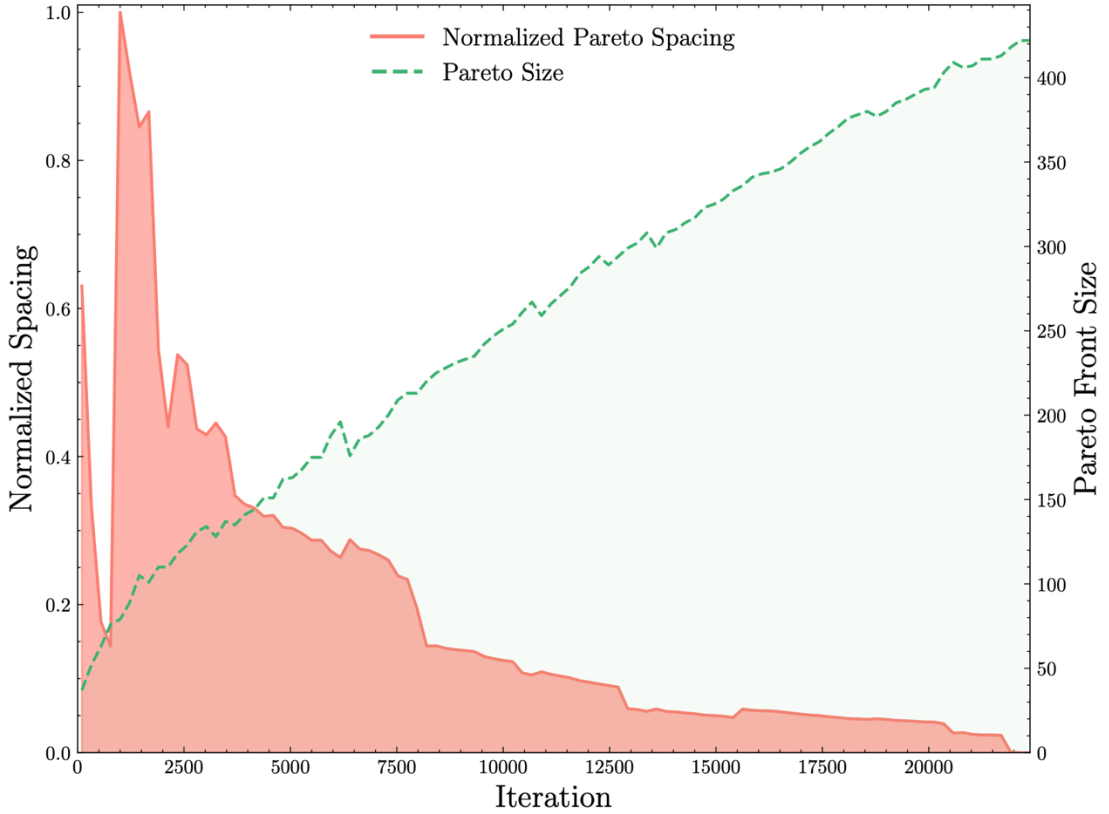
Simulated waveforms data (98,583)
down-sampled to just 9 data points

Brookhaven National Laboratory

Optimization Quality



Hypervolume

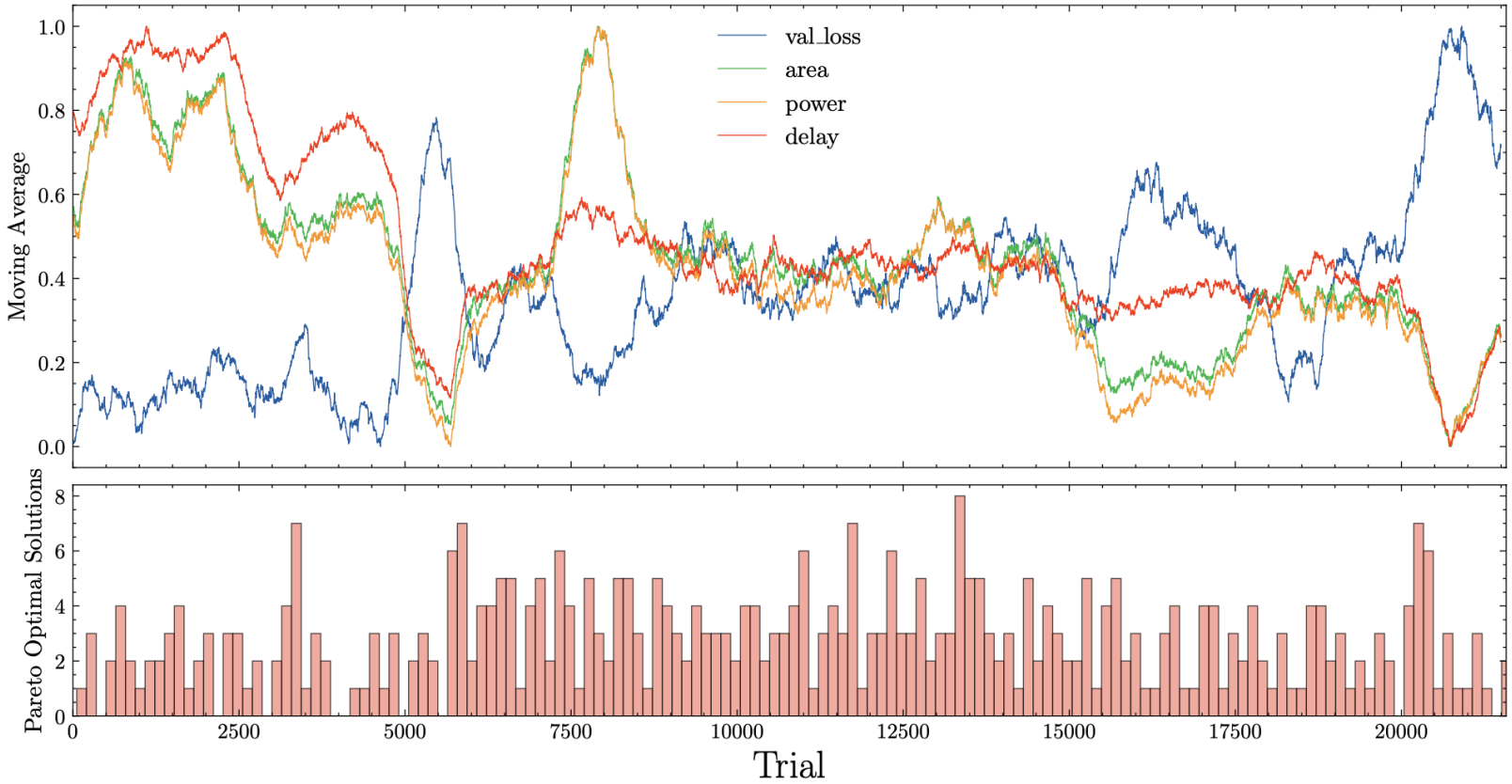


Pareto Metrics

~22,000 Synthesis and Training

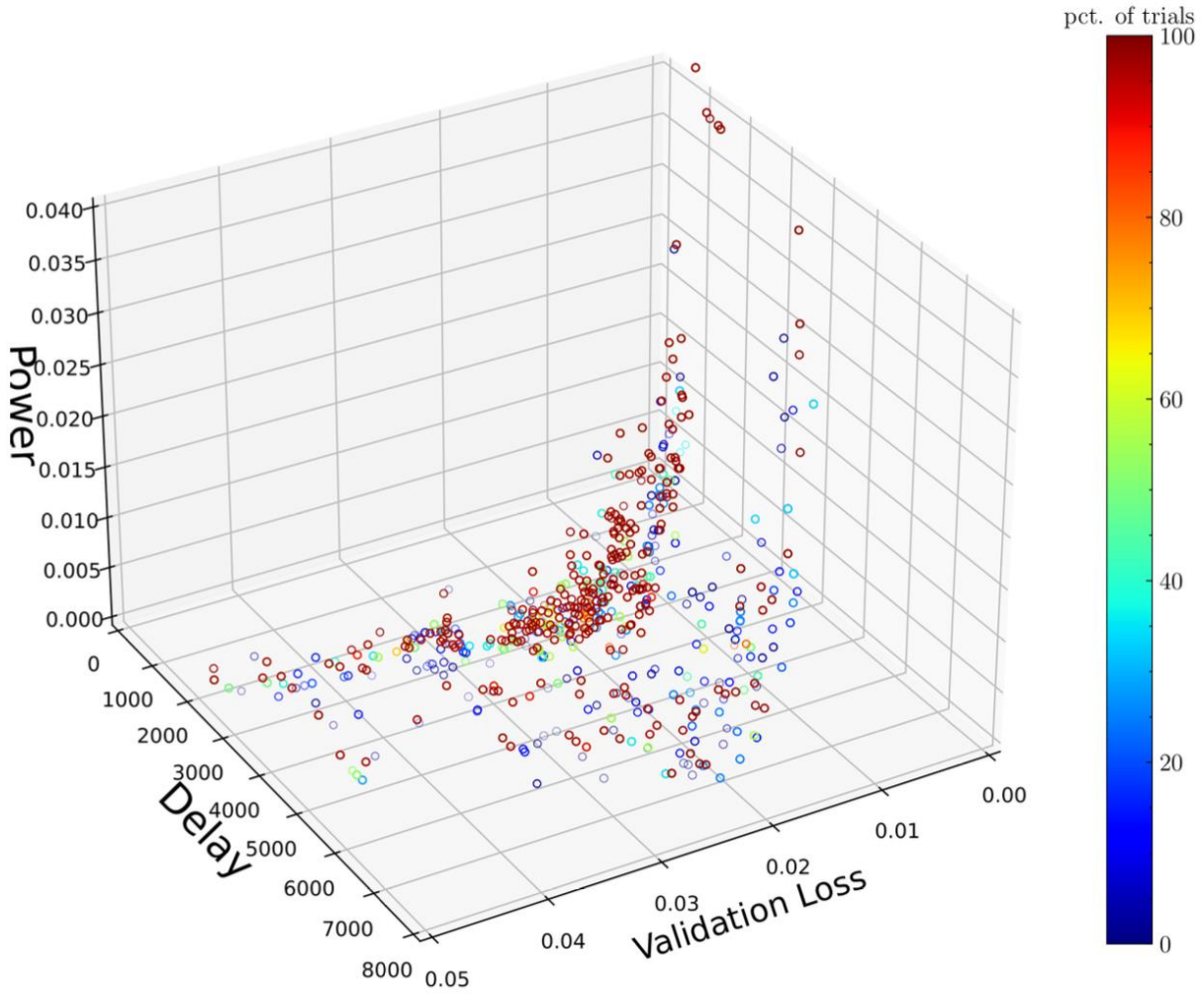
Brookhaven National Laboratory

Optimization Dynamics



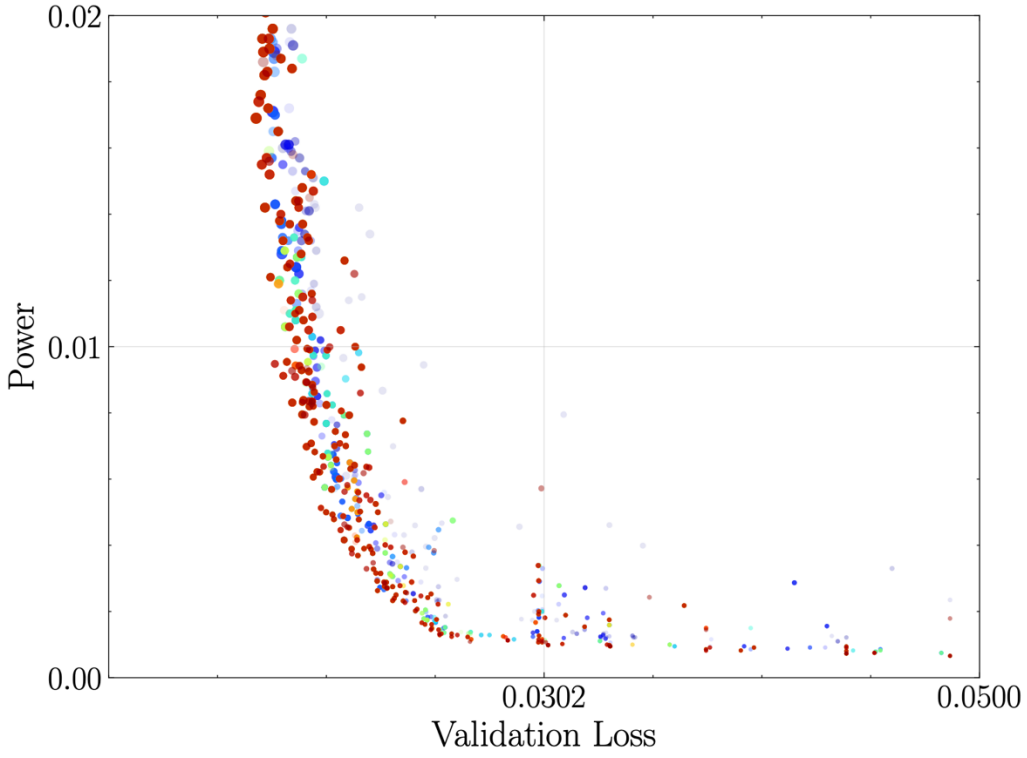
Competing Objectives and pareto discovery

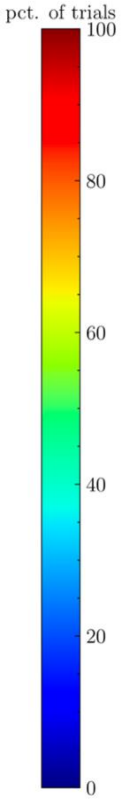
Pareto Front Dynamics

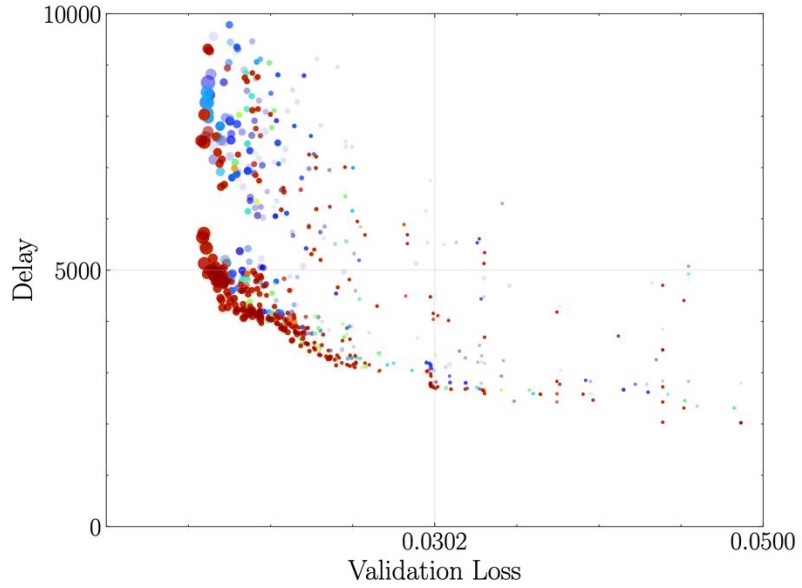


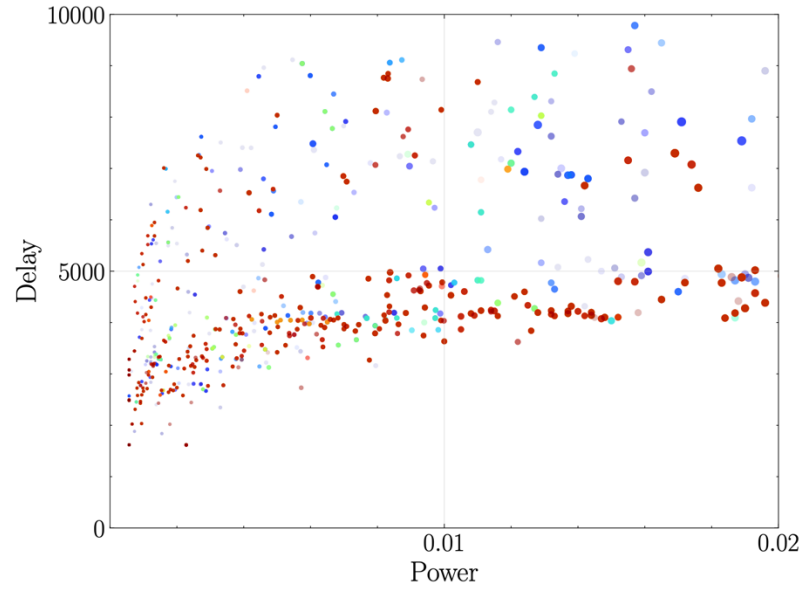
Brookhaven National Laboratory

Pareto Slices



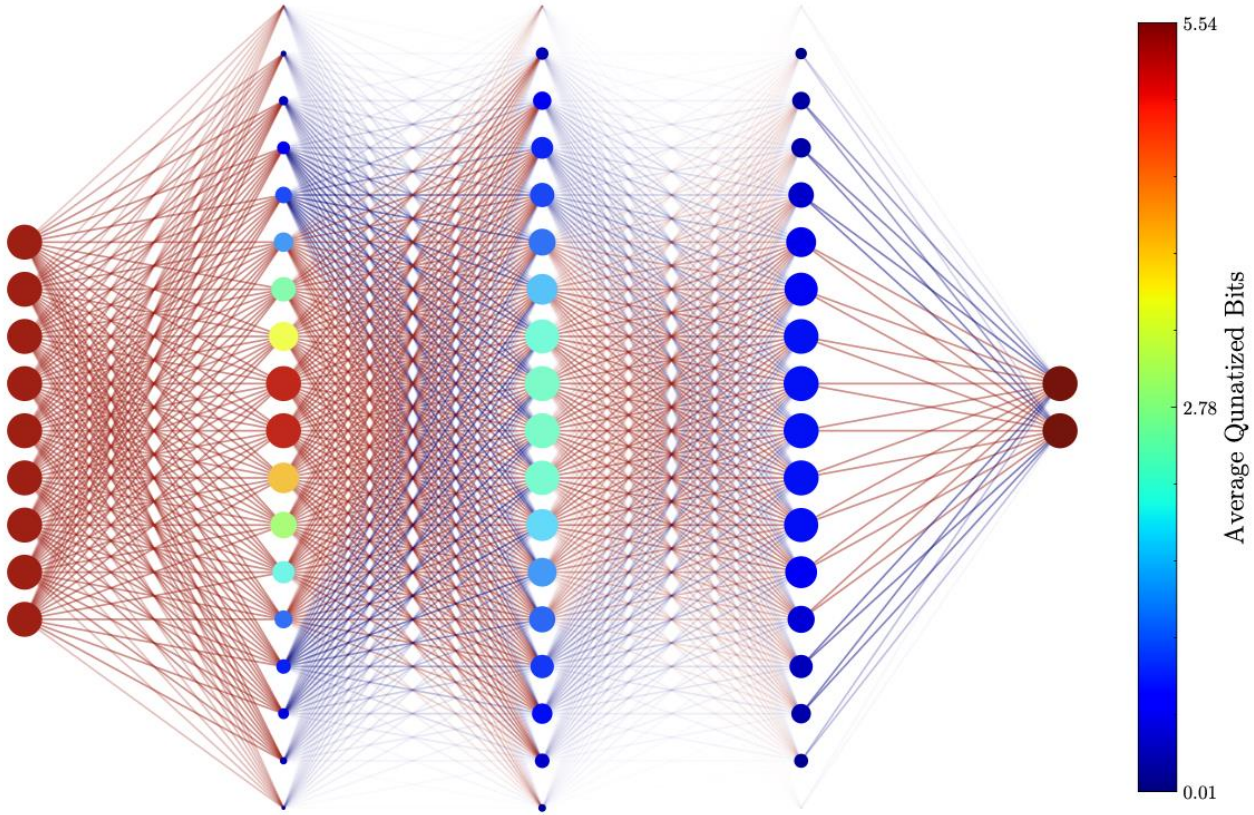






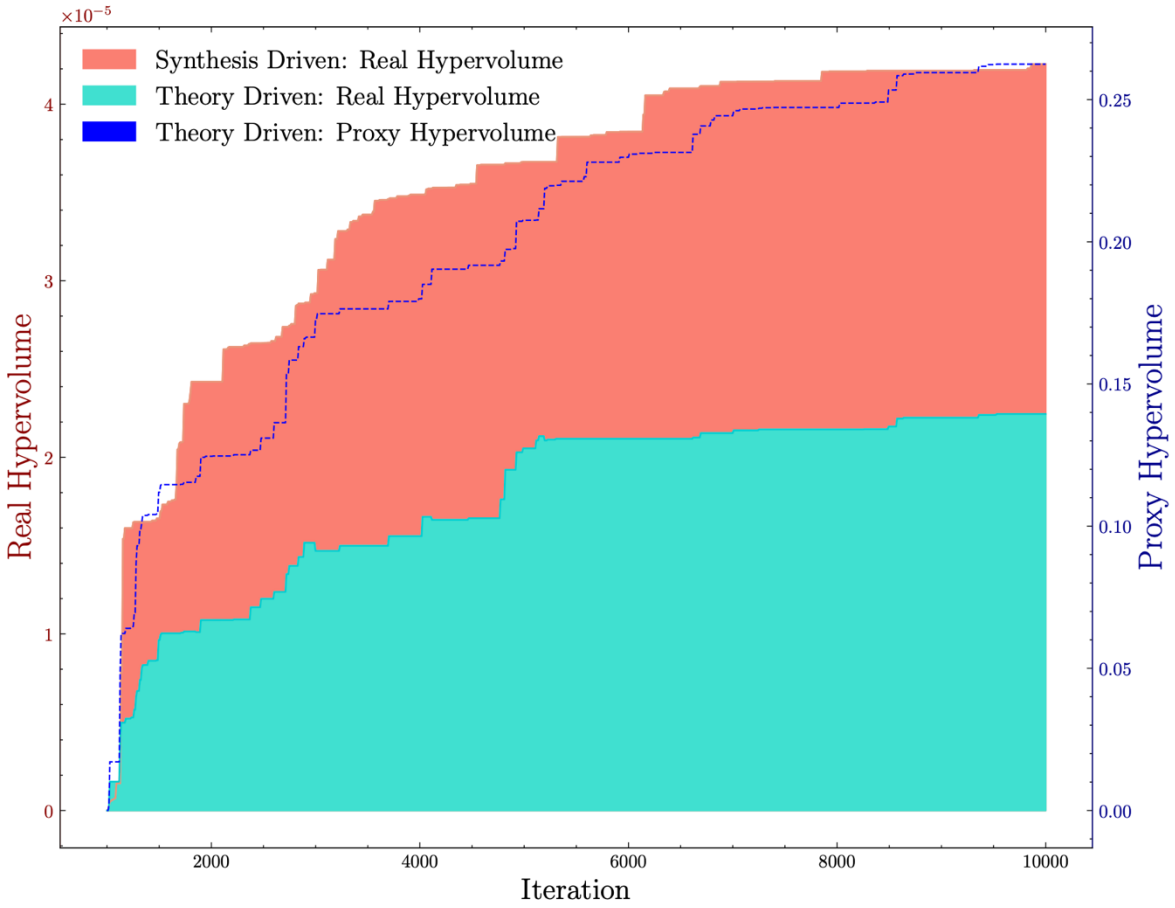
Brookhaven National Laboratory

Pareto-optimal Neural networks

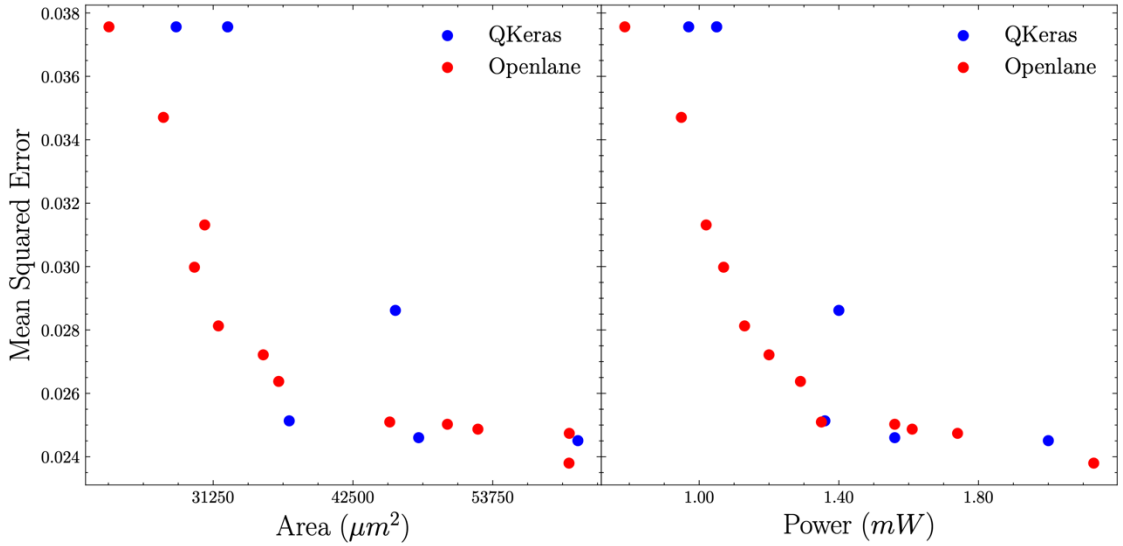


Weighted view of neural network architectures

Comparison with theory driven search



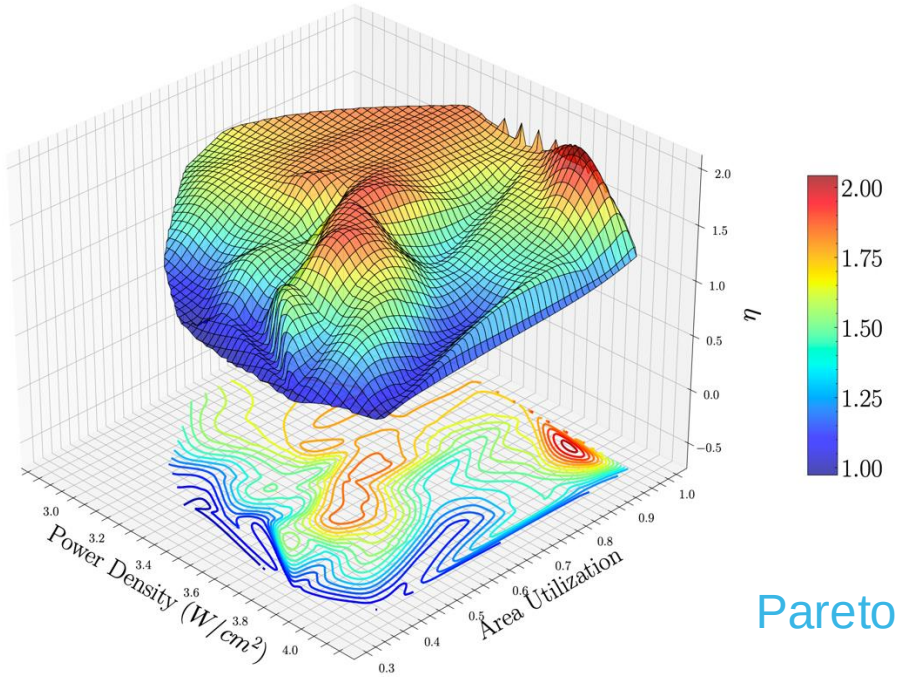
Theory-driven search is not optimal

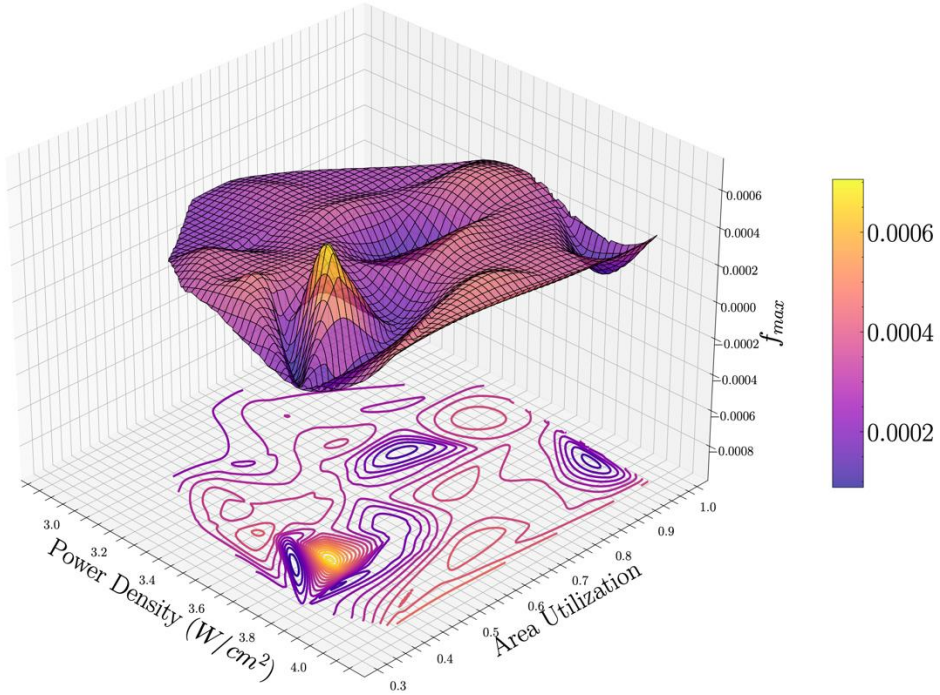


More pareto-optimal soultions

Brookhaven National Laboratory

In-pixel Constraints

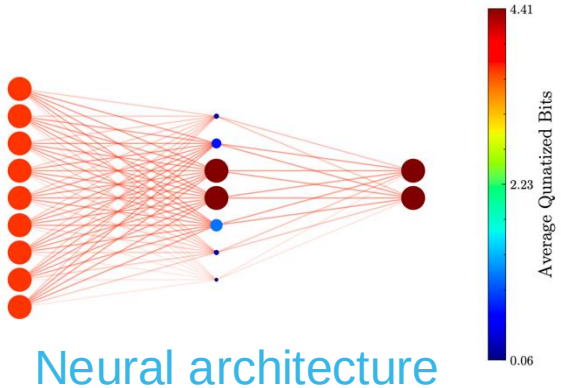




Pareto Optimality

Parameter	Maximum Constraint
Area	$\leq$ 250 x 250 μm^2
Power Density	$\leq$ 5 W/cm^2
Delay	$\leq$ 20 ps
Val. MSE	$\leq$ 0.044837

Constraints



Neural architecture

Brookhaven National Laboratory

Questions?