Real-Time Toroidal Equilibrium Reconstruction for RFX-mod2 via Quantized Neural Network in FPGA

L. Saccaro, A. Rigoni[1], P. Zanca[1], D. Terranova[1,2], R. Cavazzana[1]

[1]Consorzio RFX, Corso Stati Uniti 4, 35127 Padova, Italy
[2]Istituto per la Scienza e la Tecnologia dei Plasmi, CNR, Padova, Italy

Lorenzo Saccaro · Consorzio RFX / Università di Padova

25th IEEE Real Time Conference · La Biodola, Elba, Italy · May 25-29, 2026

RFX-mod2

Flexible device:
- Tokamak
- Reversed Field Pinch (RFP)
- Ultra-low q

Unique control capabilities:
- 192 independently driven saddle coils
- Advanced MHD modes amplitude and phase feedback control

Diagnostics enhancements and additions:
- Magnetics, e^- and ion temperature, plasma density, flow and electrostatic potential, neutron and x-ray production, impurity behaviour and more…
- Higher spatial and temporal resolution

Revamped EM acquisition:
- 1500 probes
- Up to 1 MHz for data acquisition
- 10 kHz for control

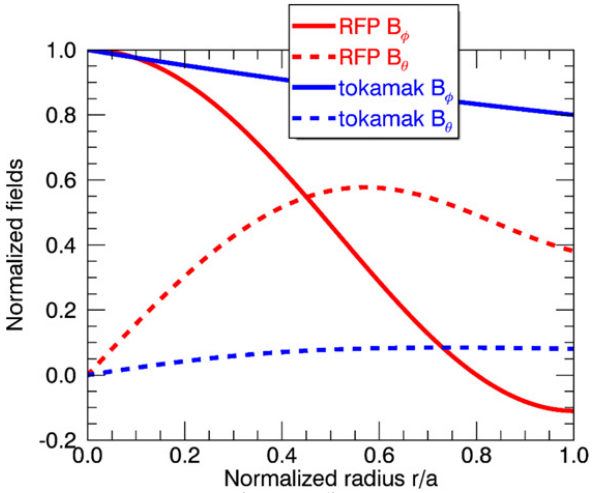


image credits:
https://doi.org/10.1088/1741-4326/abc06c

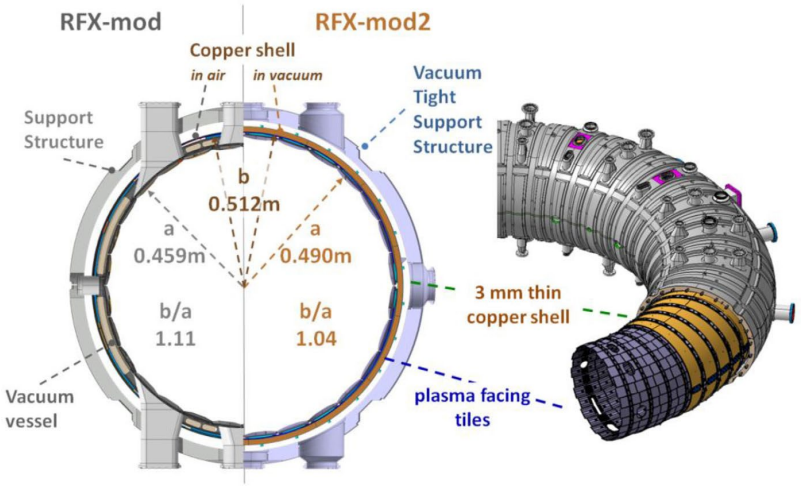
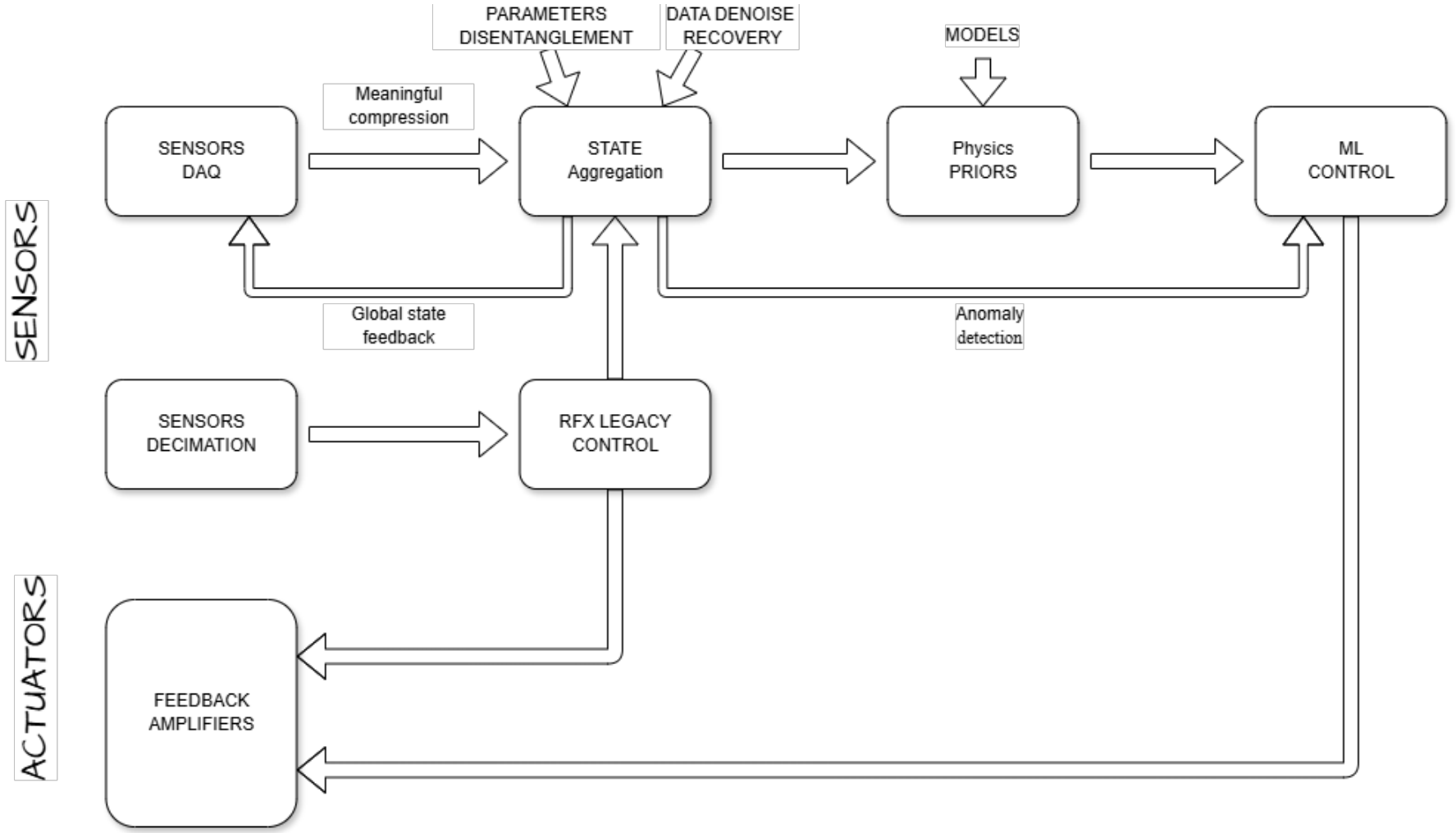
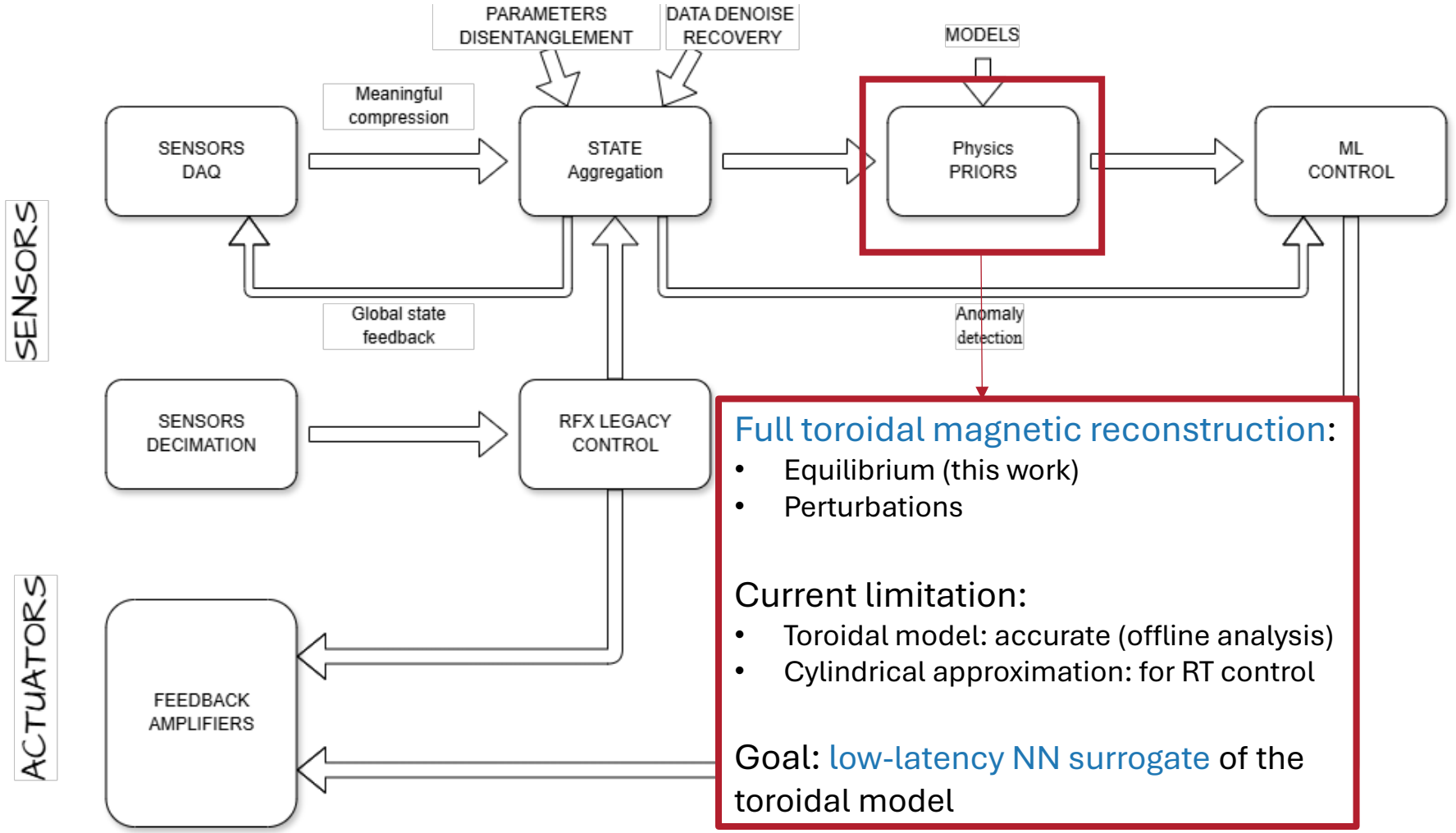


image credits:
https://doi.org/10.1088/1741-4326/ab1c6a

Towards Intelligent Control



Towards Intelligent Control



Full toroidal magnetic reconstruction:

- Equilibrium (this work)
- Perturbations

Current limitation:

- Toroidal model: accurate (offline analysis)
- Cylindrical approximation: for RT control

Goal: low-latency NN surrogate of the toroidal model

Toroidal Equilibrium Surrogate Model

Equilibrium Model

Assumptions:

- Plasma with circular cross section
- Force-free (e.g. low-β RFP → neglect pressure effect)
- Ideal shell as magnetic boundary

Using flux co-ordinates:

$$\mu_o \mathrm{J}_0 = \sigma(r)\mathrm{B}_0$$

$$\sigma(r) = \frac{2\Theta_0}{a}\left(1 - \frac{r}{a}\right)^{\alpha}$$

Parametric system of ODEs:

$$\frac{d\mathbf{y}}{dr} = \mathbf{f}(\mathbf{y}, r, \boldsymbol{\lambda}), \qquad \mathbf{y}(r_0) = \mathbf{y}_0(\boldsymbol{\lambda})$$

$$\mathbf{y} = [\hat{B}_1^{\phi}, \hat{B}_1^{\vartheta}, \Delta, \Delta', \hat{B}_2^{\phi}, \hat{B}_2^{\vartheta}] \qquad \boldsymbol{\lambda} = [\alpha, \Theta_0, \Delta_H]$$

Shift of plasma surface

EDGE MEASUREMENTS

Neural Surrogate

Approximate the non-trivial part of the flow map and hard-constrain the I.C.:

$$\mathbf{y}_{\mathrm{NN}}(r; \boldsymbol{\lambda}) = \mathbf{y}_0(\boldsymbol{\lambda}) + (r - r_0)\mathbf{N}_{\theta}(r, \boldsymbol{\lambda})$$

Learnable parameters

Simple MLP architecture w/ ReLU activations

Database of experimentally relevant equilibria:

- $\alpha \in [3,10]$
- $\Theta_0 \in [1.3,1.6]$
- $\Delta_{\mathrm{H}} \in [-0.01, 0.01]\mathrm{m}$

Training set:

- Parameters sampled on regular grid (101x101x11)
- 100 equally spaced points $\mathrm{r} \in [\mathrm{r}_0, \mathrm{r}_{\mathrm{max}}]$

Validation set:

- 100k params triplets uniformly sampled
- 250 uniformly sampled points $\mathrm{r} \in [\mathrm{r}_0, \mathrm{r}_{\mathrm{max}}]$

Inputs to NN mapped to [0, 1]

Architecture Optimization

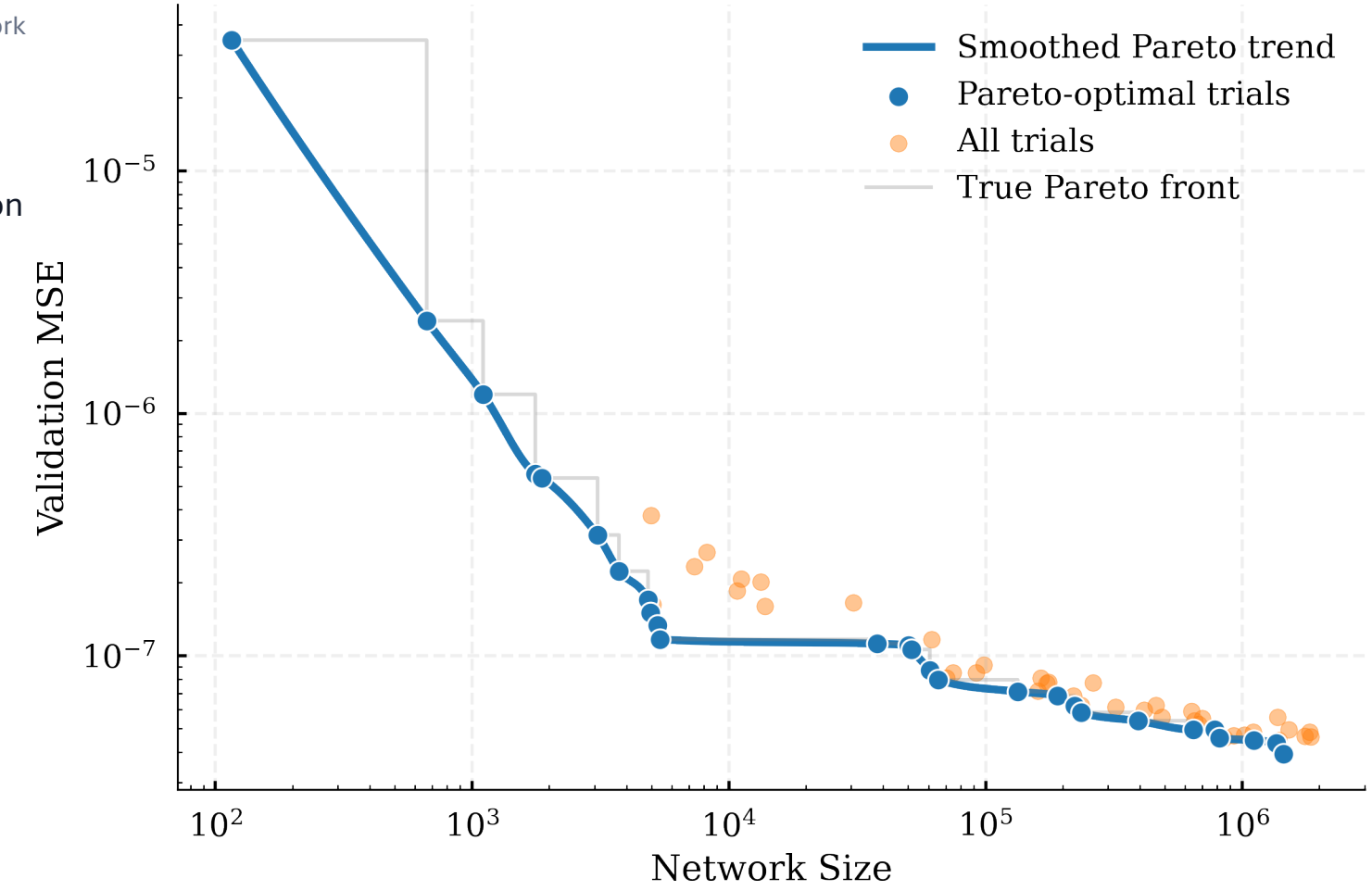
OPTUNA hyperparameter optimization framework

Multi-Objective Optimization (MOO):

- Trade-off network size (# of parameters) vs Validation Mean Squared Error (MSE)
- NSGA-II sampler: well suited for MOO
- Depth (1–5), Width (10–500)
- 1000 trials budget

Takeaways:

- Accuracy plateau in the 6k to 40k params region
- Diminishing returns for larger models
- Shallow, wide architecture dominates low parameters region



Architecture Optimization

OPTUNA hyperparameter optimization framework

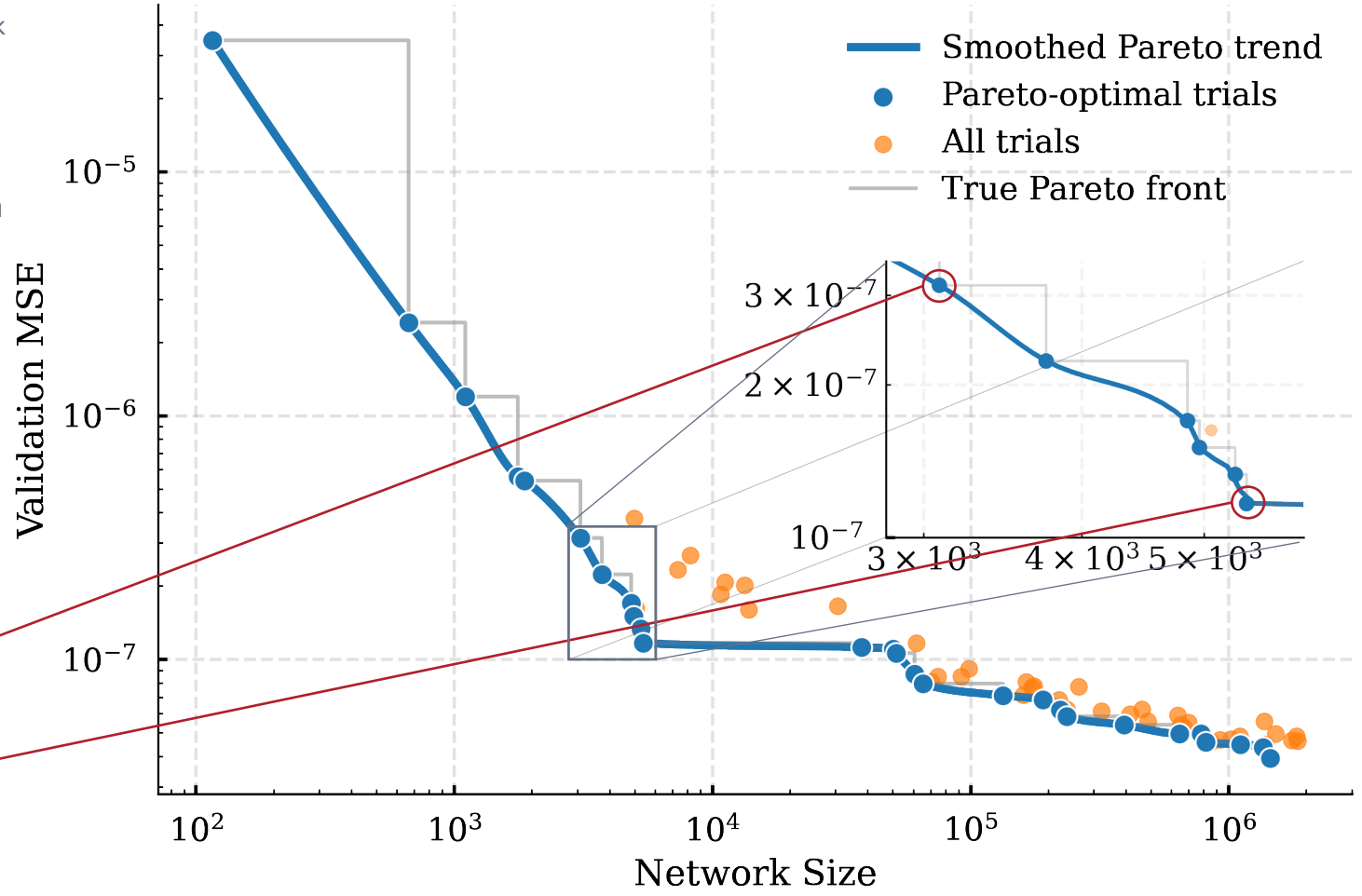
Multi-Objective Optimization (MOO):

- Trade-off network size (# of parameters) vs Validation Mean Squared Error (MSE)
- NSGA-II sampler: well suited for MOO
- Depth (1–5), Width (10–500)
- 1000 trials budget

Takeaways:

- Accuracy plateau in the 6k to 40k params region
- Diminishing returns for larger models
- Shallow, wide architecture dominates low parameters region

Selected candidates: one-hidden-layer networks with 280 and 490 units



Quantization-Aware Training

32-bit floating point (FP32)

↓

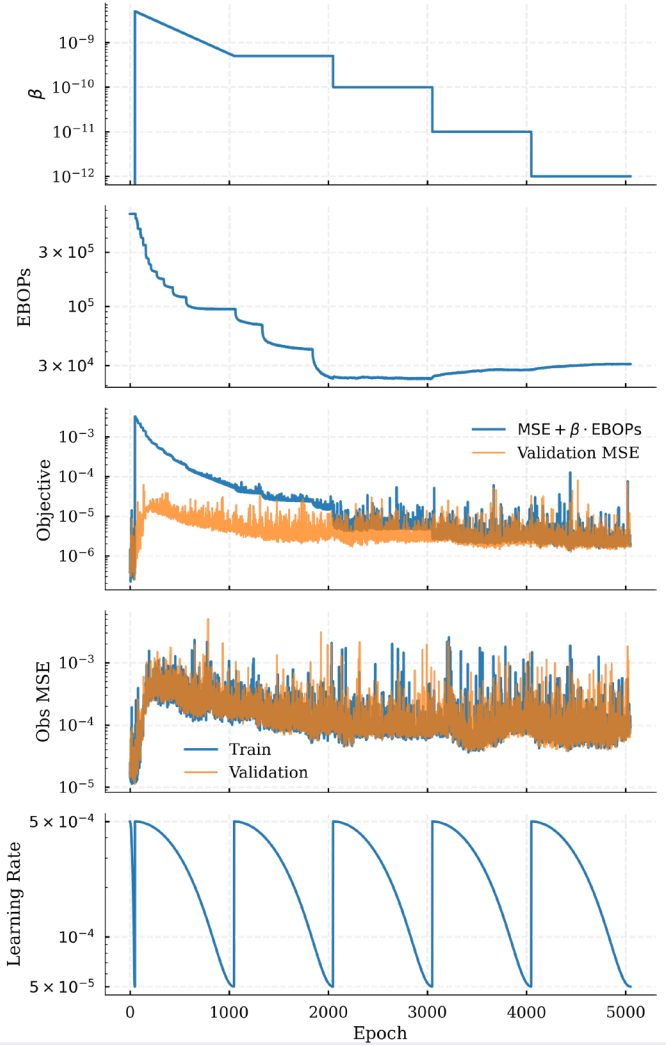
n-bit fixed-point precision

High Granularity Quantization:

- Heterogeneous quantization (down to individual weights / activations)
- Hardware-cost proxy: Effective Bit Operations (EBOPs)
- Bit-widths optimized via gradient descent
- Training loss: MSE + β · EBOPs

Training Strategy:

- Start with pretrained FP32 network
- Fix inputs and outputs precisions
- Vary β during training via scheduler
- Use cosine annealing with warm restarts for the learning rate



Quantization-Aware Training

HGQ II

32-bit floating point (FP32)

↓

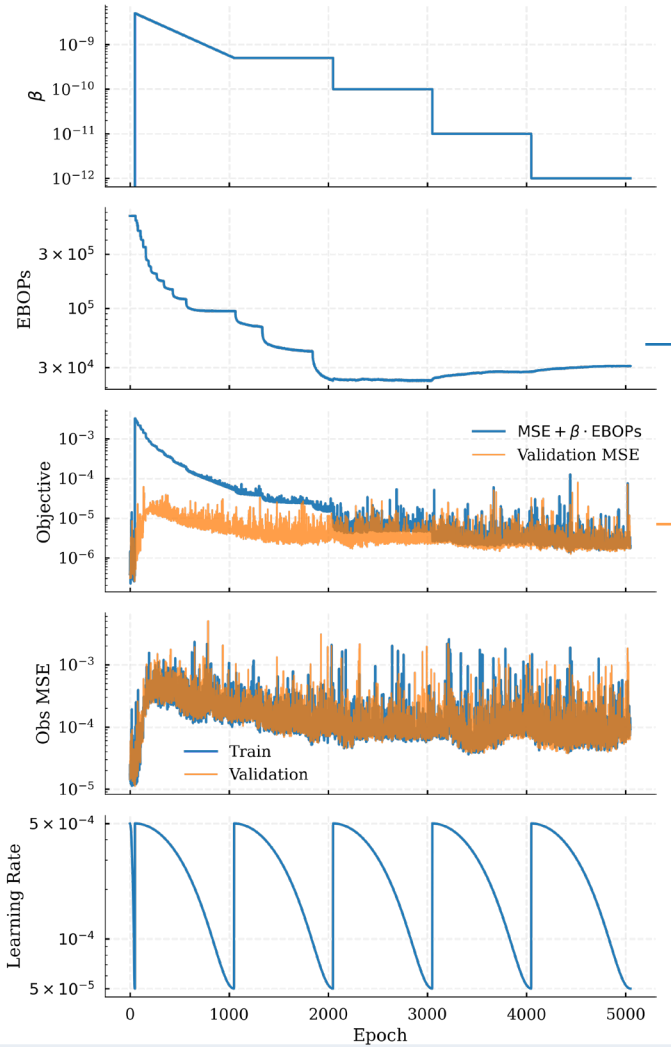
n-bit fixed-point precision

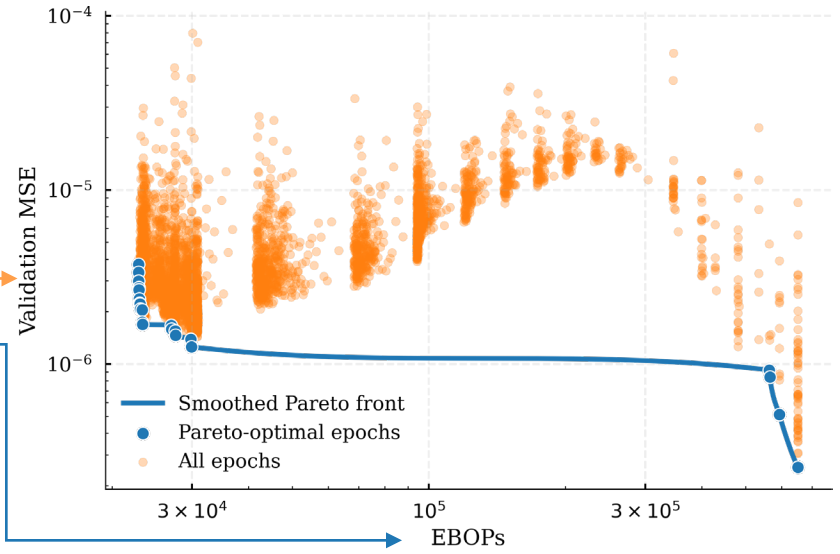
High Granularity Quantization:

- Heterogeneous quantization (down to individual weights / activations)
- Hardware-cost proxy: Effective Bit Operations (EBOPs)
- Bit-widths optimized via gradient descent
- Training loss: MSE + β · EBOPs

Training Strategy:

- Start with pretrained FP32 network
- Fix inputs and outputs precisions
- Vary β during training via scheduler
- Use cosine annealing with warm restarts for the learning rate





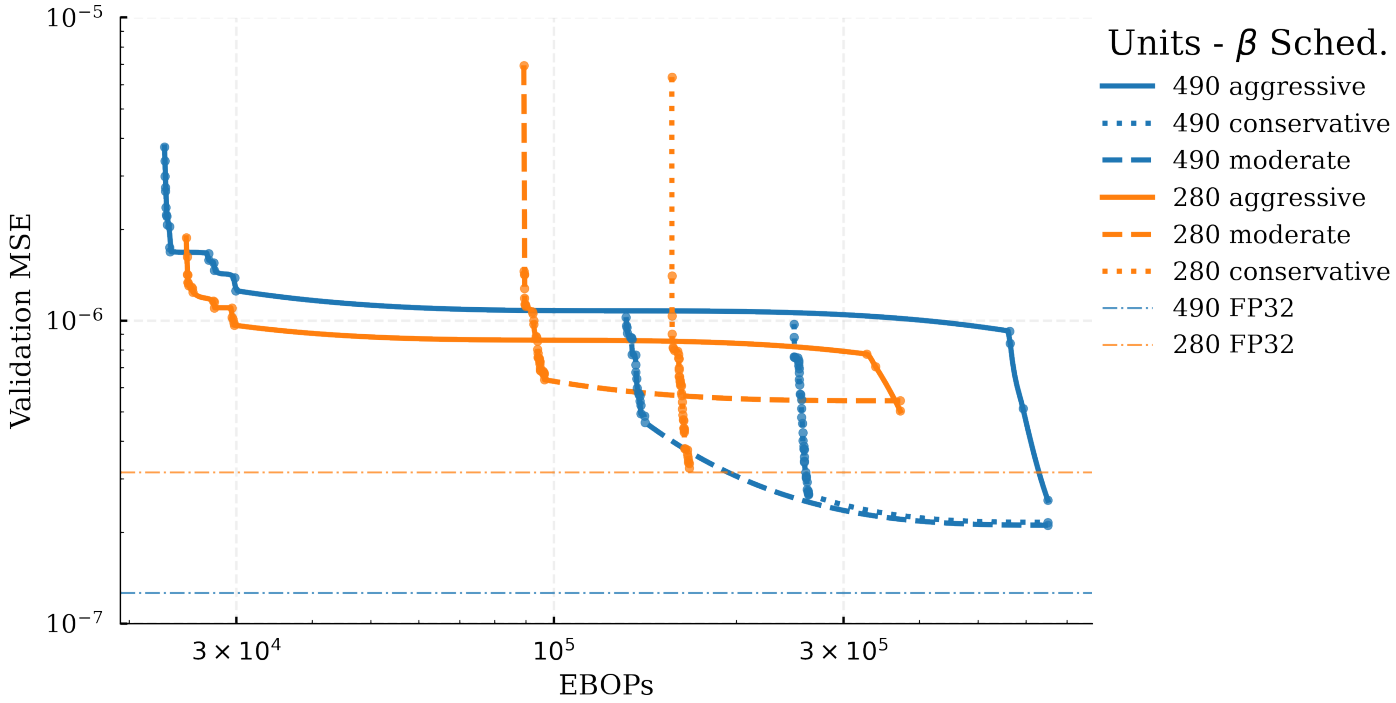
QAT run traces a Pareto trajectory:

- Initial EBOPs reduction and increased loss
- Gradual recovery of accuracy with limited EBOP increase

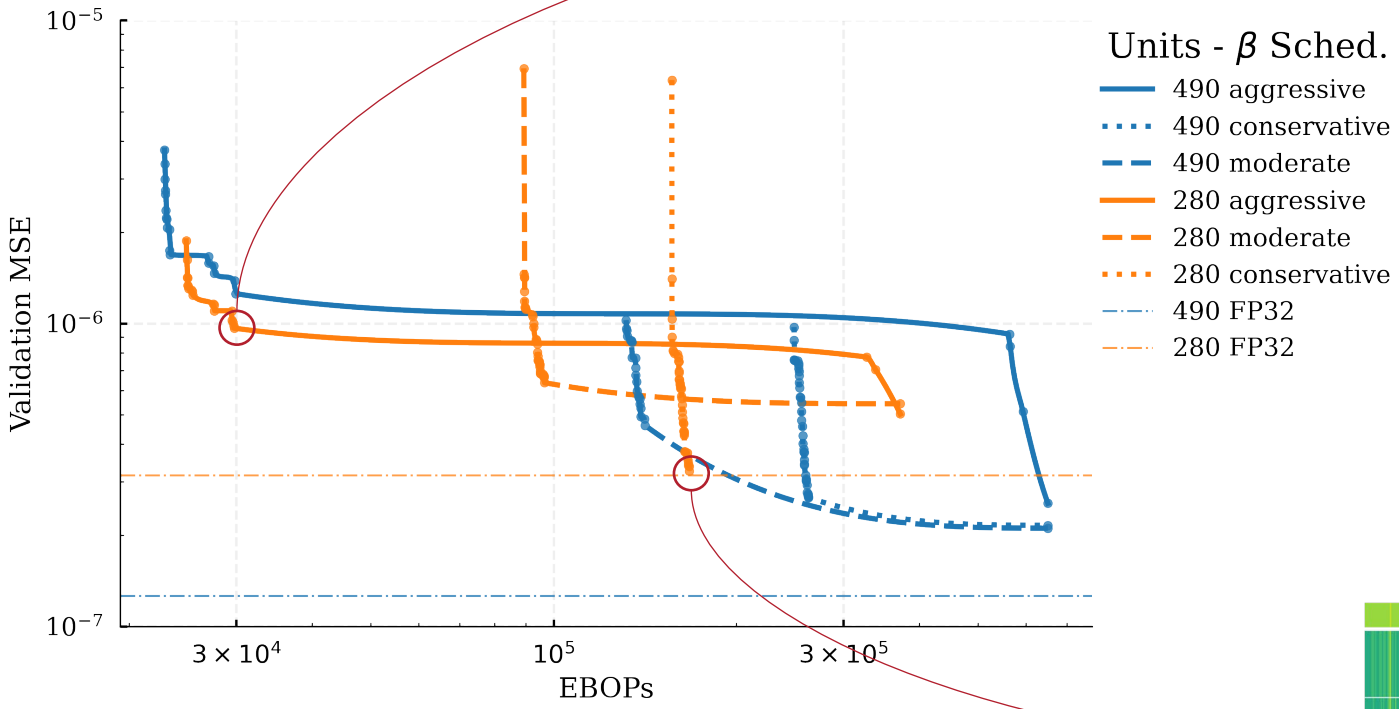
Explore new regions with different β schedulers:

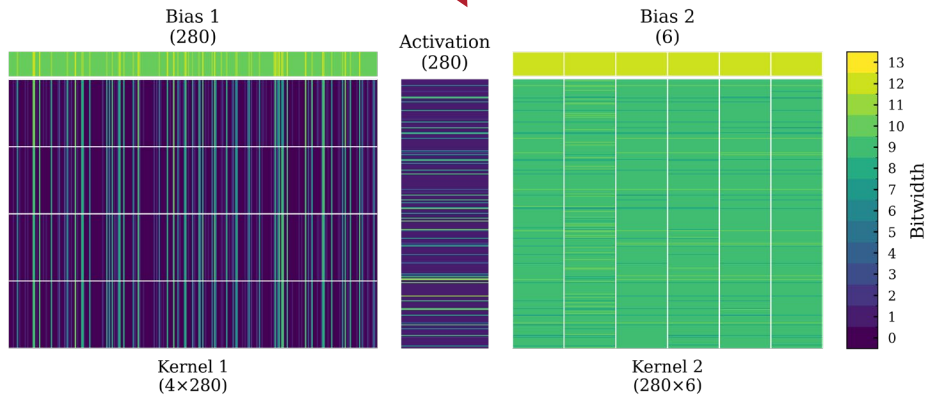
- Conservative (focus on precision)
- Aggressive (focus on EBOPs)
- Moderate (middle ground)

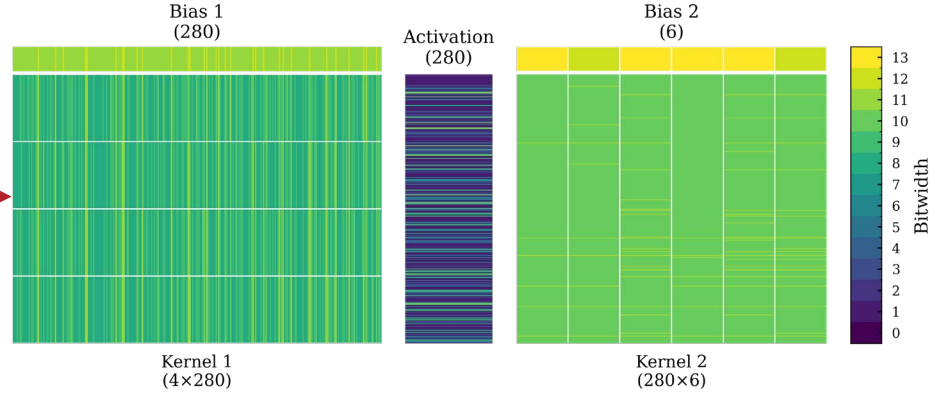
Heterogeneous quantization results



Heterogeneous quantization results







- Pruning recovered
- Inner weight matrix more compressed
- Outer weight matrix retains higher precision
- FP32-level performance achieved with a 2× EBOP reduction

HLS Conversion: Resource and Latency Estimates

hls4ml

Firmware implementation of ML algorithms using high-level synthesis (HLS)

Automatic precision inference:

- Bit-exact with HGQ model
- No need for manual tuning

Target minimum achievable latency:

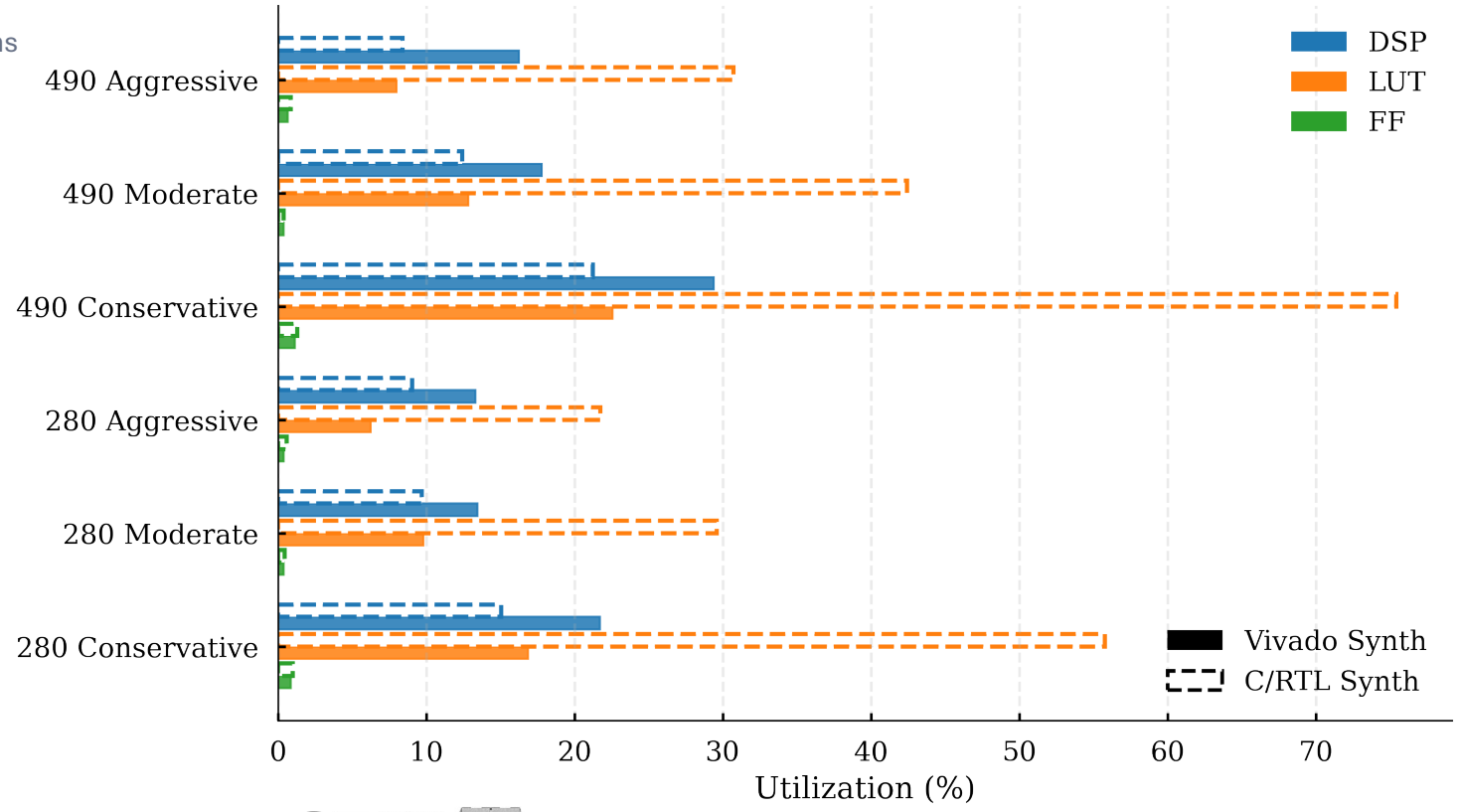
- Latency strategy
- io_parallel
- 10 ns clock period
- May increase resource usage

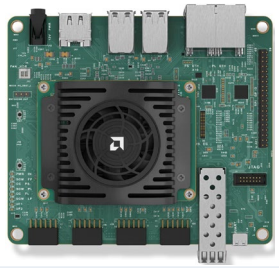
Achieved latency (single inference):

- 280/490 cons./aggr. 30 ns (3 clock cycles)
- 280/490 mod. 20 ns (2 clock cycles)

Fully pipelined in all cases (II = 1)

Selected NN for hw implementation: 280 conservative

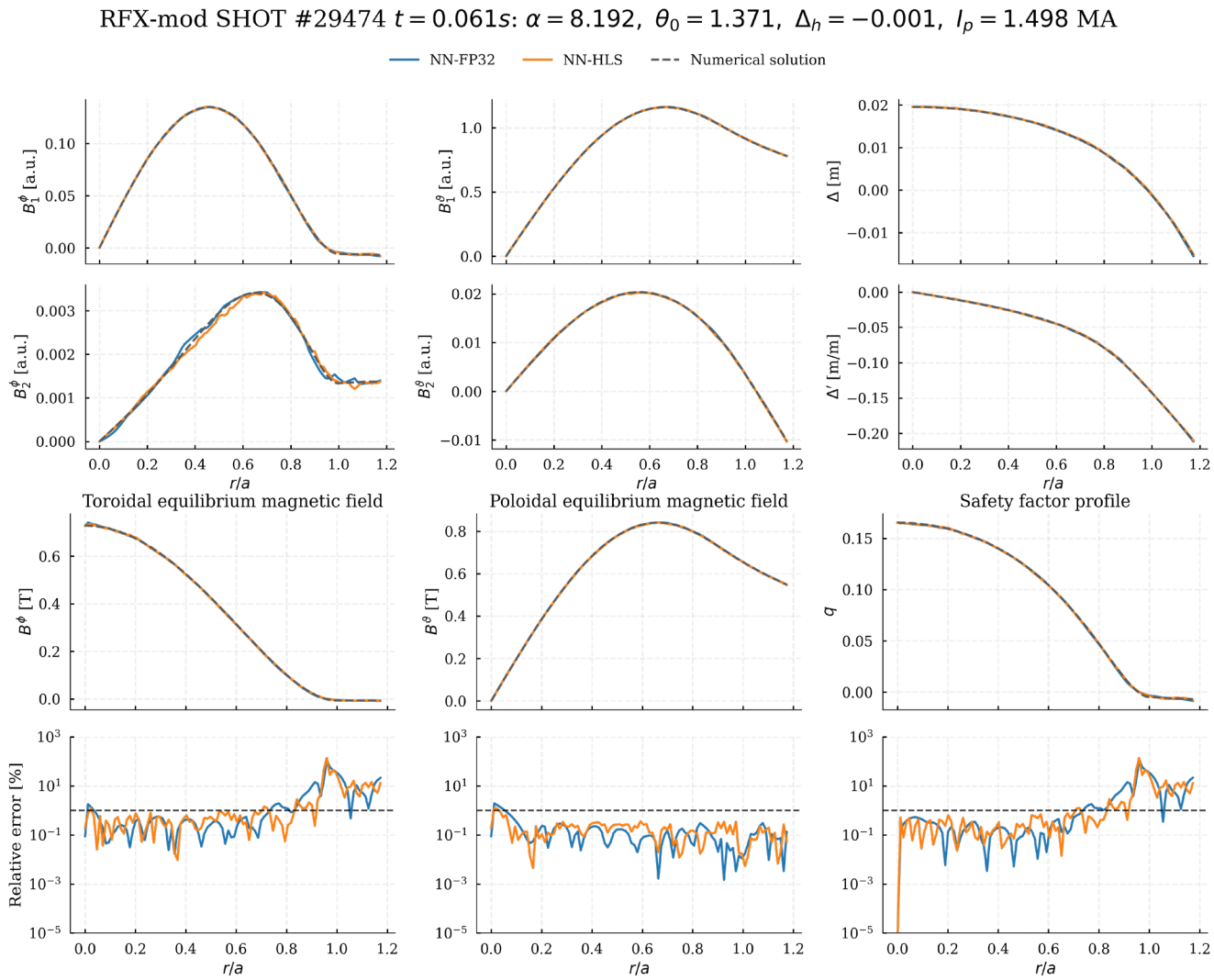




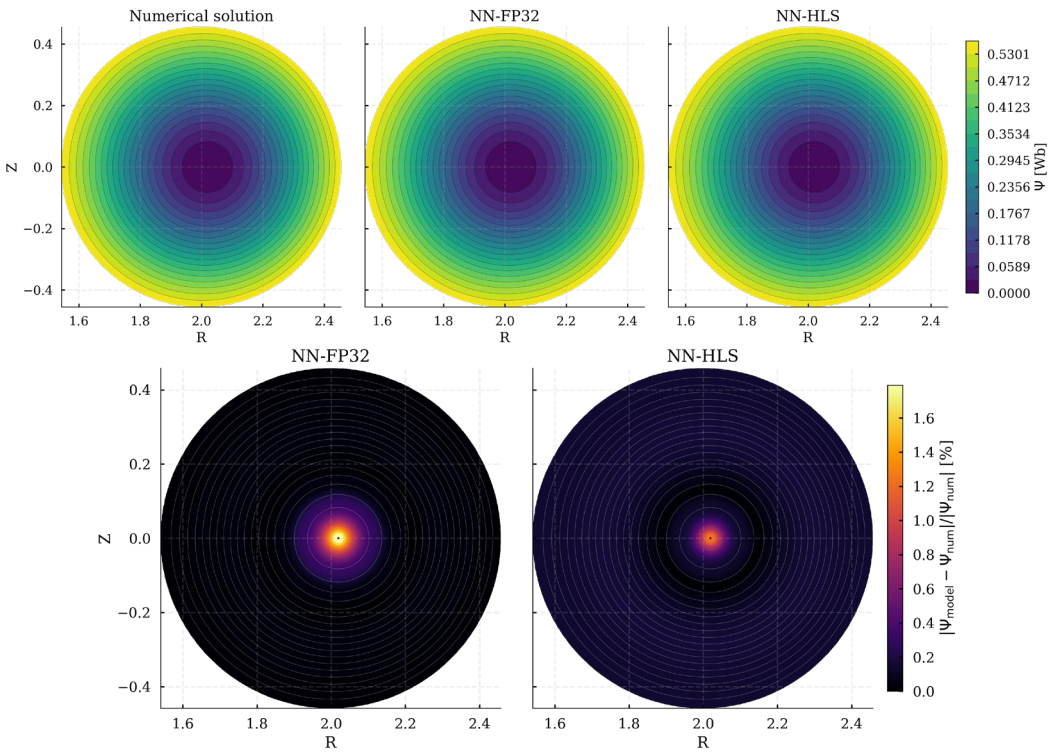
Target device: Kria K26 SOM

- DSP: 1248
- LUT: 117k
- FF: 234k

Profiles Reconstruction



Equilibrium Poloidal Flux Map Ψ(R, Z)



Average profile error < 1%

Only spikes when reference quantity → 0, or outside plasma radius

Largest flux-map deviations occur in the core

Radial Position of Rational Surfaces

Rational surfaces: $q(r) = \dfrac{rB_\varphi}{RB_\theta} = \dfrac{m}{n}$

- Resonant (mainly tearing) modes inside the plasma
- Accurate positioning key for perturbation reconstruction and control purposes
- Error distributions computed over 500 RFP equilibria for the most relevant modes

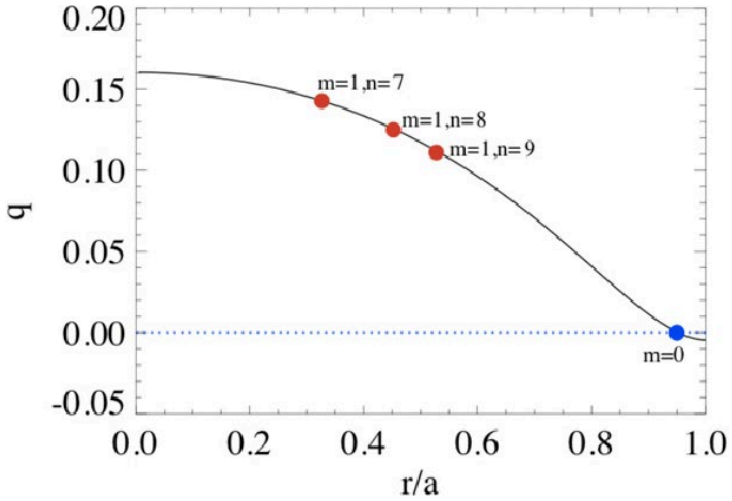
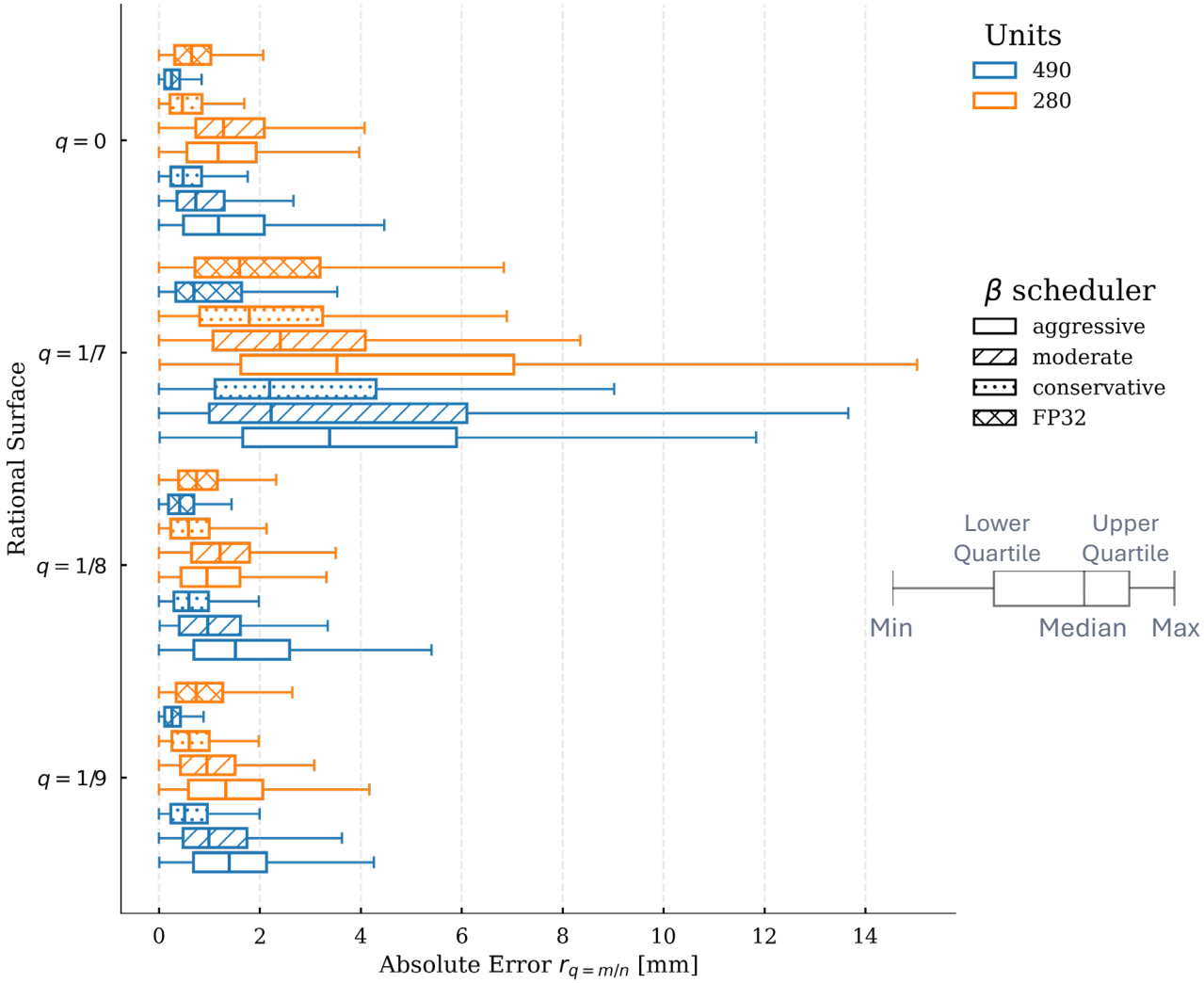


image credits: https://doi.org/10.1088/0029-5515/51/9/094023



Final Hardware Implementation

Realistic wrapper implementation:

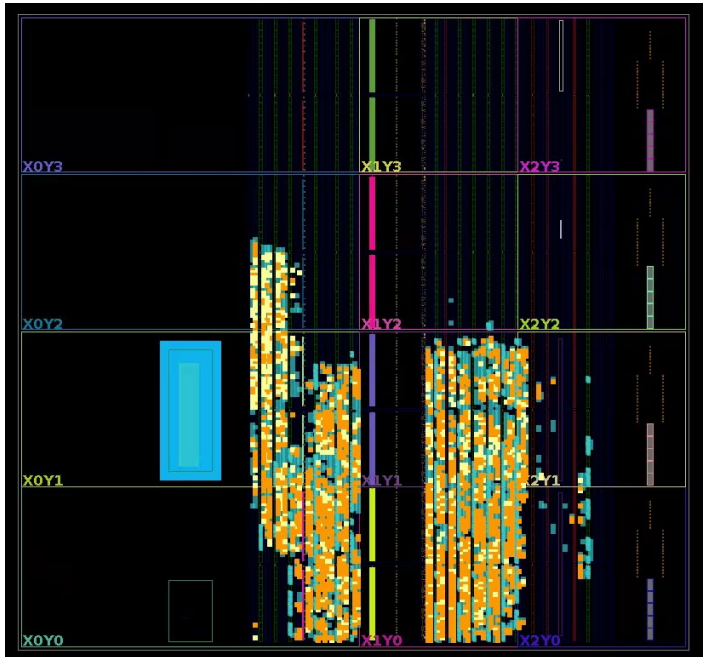
- Slowly varying inputs (α, Θ_0, Δ_H) received via AXI4-Lite
- 21-point radial scan from internal fixed LUT
- Outputs streamed via AXI4-Stream to downstream NN HLS cores or back to the processing system via DMA

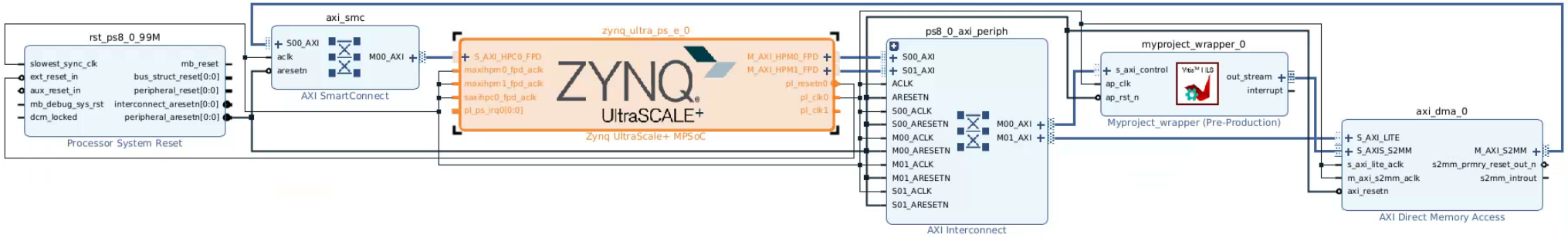
Post-implementation timing simulation:

- Full radial scan latency: ~300 ns
- ~500 ns including input readout

FP32 model:

- Optimized implementation + JIT compilation (numba)
- Average latency: 86.2 µs (5.0 µs std, 105.6 µs 99th pct)
- ≈280× speedup





Conclusions

Summary

Compact quantized NN surrogate for toroidal equilibrium reconstruction:

- Accurate reconstruction of magnetic field profiles and Shafranov shift
- Preservation of key physical quantities

FPGA implementation meets real-time requirements:

- Full radial scan completed in the sub-microsecond range
- Large latency and resource margin to accommodate larger and/or additional models

Demonstrated an end-to-end workflow that transforms a computationally expensive physics-based model into a low-latency FPGA accelerator

Future work

Complete full magnetic reconstruction chain:

- Extend surrogate approach to perturbations
- Integrate with magnetic measurements

Enhance physical consistency of NN predictions:

- Improve smoothness and derivative consistency of reconstructed profiles
- Add PINN-like regularization terms: penalize ODE residuals using automatic differentiation

Improve deployment workflow:

- Combine architecture optimization and quantization
- Constraint bit-widths to boost hardware efficiency (LUT/DSP usage)

Thank you

Questions?

lorenzo.saccaro@igi.cnr.it