Analysis of Deployment Challenges for Machine Learning Signal Processing Algorithms

V. Rajesh, C. C. Hall, J. P. Edelen

15 Oct 2025

LLRF 2025: Newport News, VA

Industrial Accelerators

Security and Defense
- Directed energy
- Single effects testing

Imaging
- X-ray sources
- Gamma-ray sources
- electron microscopy

Manufacturing
- Polymer treatment
- Industrial Curing
- Welding
- Ion Implantation

Medical
- Proton Therapy
- E-beam therapy
- X-ray therapy

Sterilization
- Medical device sterilization
- Food irradiation
- Wastewater treatment

- Legacy systems lack complexity, automation is straightforward
  - Single RF structure controlled with a PLL or similar
  - loose beam tolerances
- Next generation of industrial systems are increasing in complexity
  - Synchronization of multiple structures for higher energy applications
  - Tighter tolerances on output beams for emerging applications

Opportunities for Industrial Accelerators

- Focus areas for improving controls
  - Improvement of feedback systems for beam stabilization
  - Automation of startup routines (calibrations and synchronization)
  - Improvement of signal quality for RF systems

Opportunities for Industrial Accelerators

- Focus areas for improving controls
  - Improvement of feedback systems for beam stabilization
  - Automation of startup routines (calibrations and synchronization)
  - **Improvement of signal quality for RF systems**

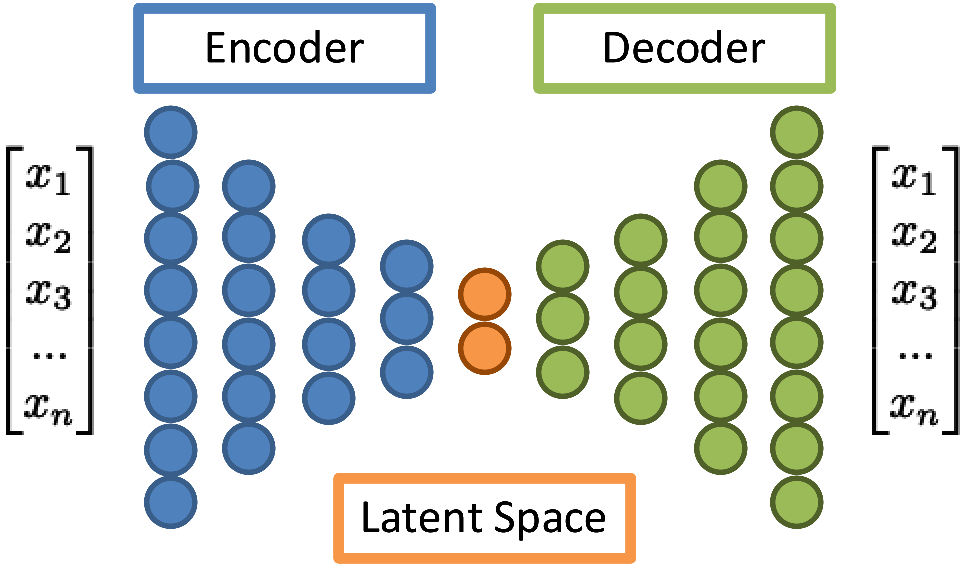
Opportunities for Industrial Accelerators

- Focus areas for improving controls
  - Improvement of feedback systems for beam stabilization
  - Automation of startup routines (calibrations and synchronization)
  - **Improvement of signal quality for RF systems**

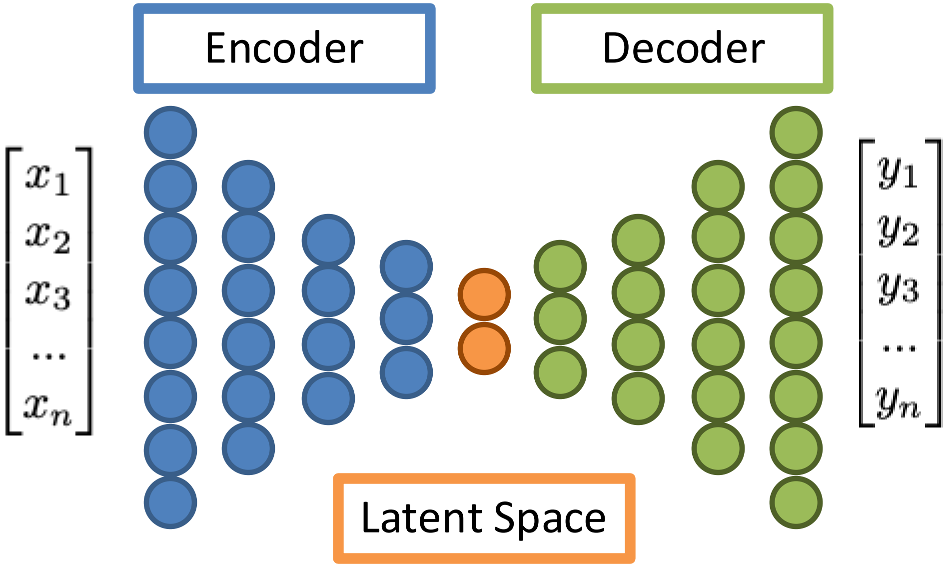
Autoencoders for Noise Reduction in RF Signals

What is an Autoencoder?

- Autoencoder
  - Type of neural network
  - Transforms data into a latent space and performs a reconstruction
  - Inputs and Outputs are the same: i.e. it is an identity transformation for a given dataset

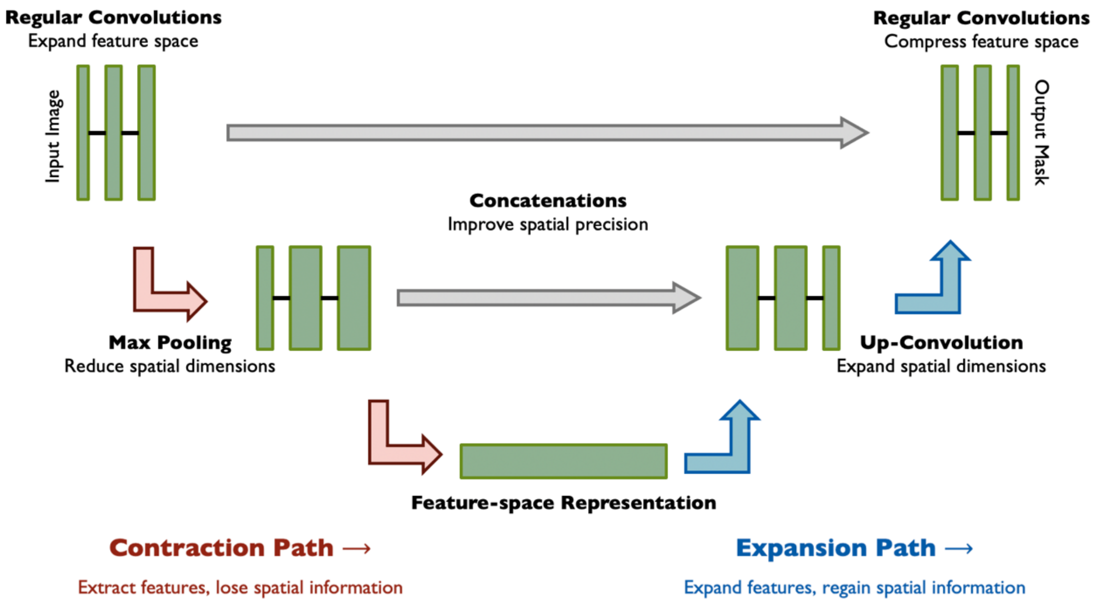


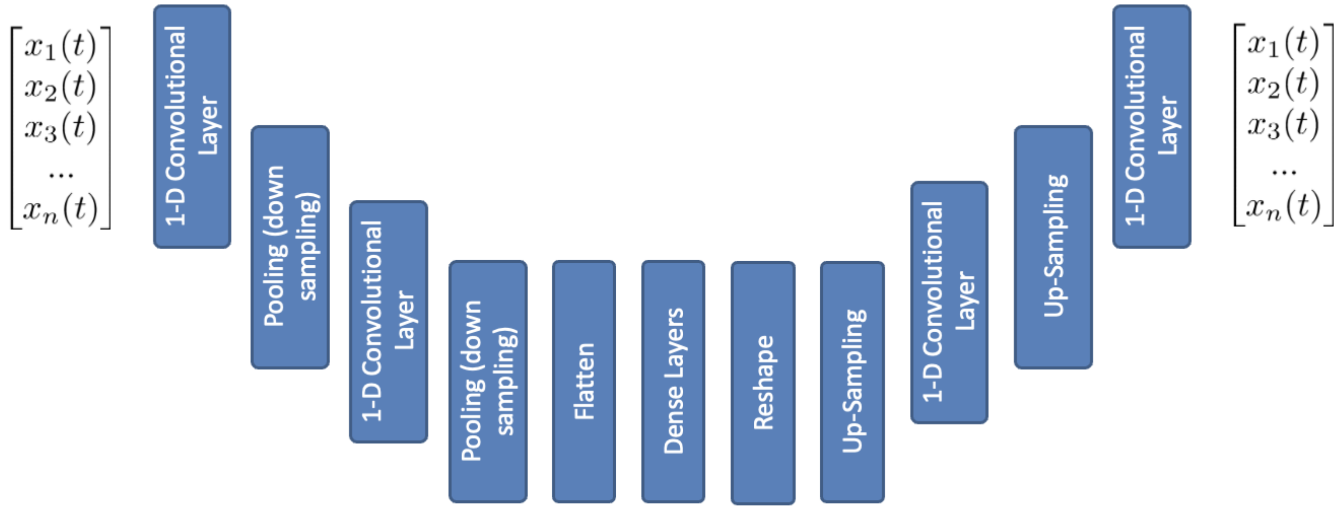
- Encoder-Decoder network
  - Transforms data into a latent space which is mapped to an output space



Convolutional Autoencoders

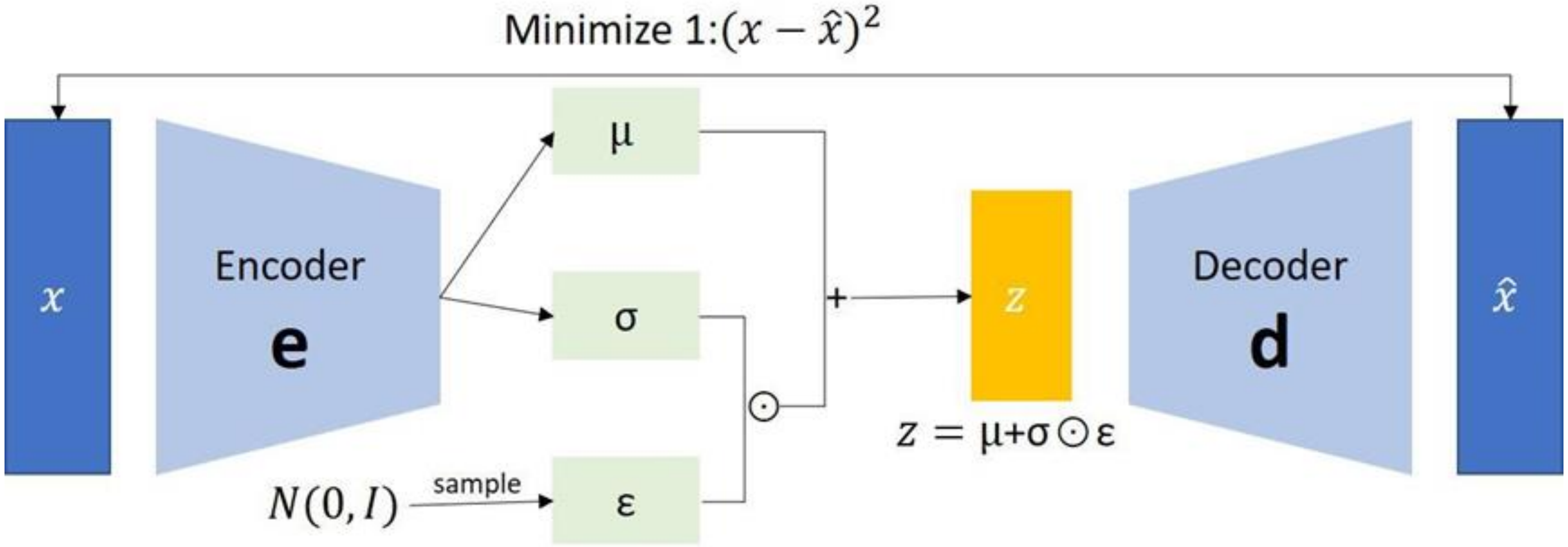
- Neural network that converts 1-D sequence into a latent-space
  - Filters learn translation invariant features much like an image based U-net
  - Pooling layers for downsampling
  - Signal is upsampled and filtered to reconstruct the original signal
  - Deconvolutional layers can also be used





Variational Autoencoders

- Variational autoencoders enforce smoothness condition in the latent space
- Dimensionality reduction removes complexity of noise
- Tests done using simulated BPM data
- Logically extended to RF data
- Could implement the autoencoder on a FPGA for near-real-time noise reduction



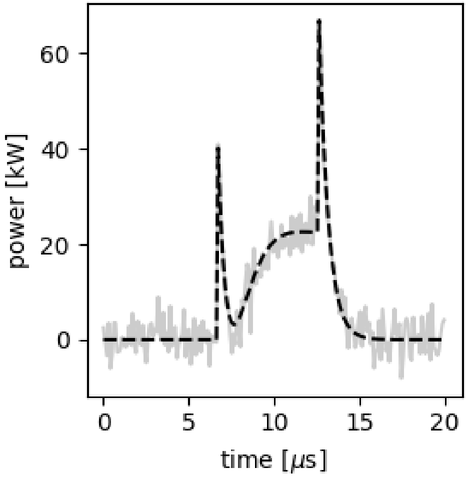
Cavity simulator

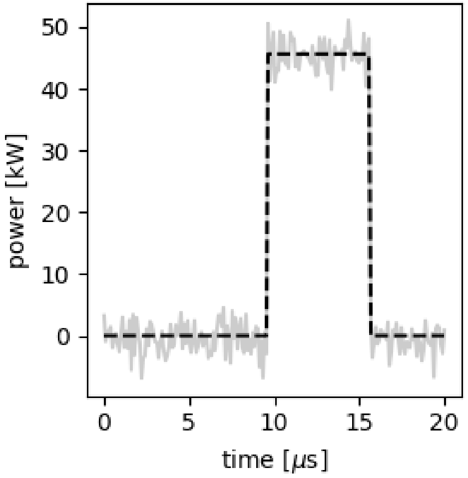
- Based on an equivalent RLC circuit model
  - Transmitted voltage differential equation:

$$\frac{d}{dt}\begin{bmatrix}\mathrm{Re}(V_t)\\ \mathrm{Im}(V_t)\end{bmatrix} = \begin{bmatrix}-\omega_{1/2} & -\Delta\omega\\ \Delta\omega & -\omega_{1/2}\end{bmatrix}\begin{bmatrix}\mathrm{Re}(V_t)\\ \mathrm{Im}(V_t)\end{bmatrix} + \frac{R_L\omega_{1/2}}{m}\begin{bmatrix}\mathrm{Re}(I_{fwd})\\ \mathrm{Im}(I_{fwd})\end{bmatrix}$$

  - Reflected voltage computed from transmitted:

$$V_r = \frac{1}{m}V_t - \frac{Z_0}{2}I_{fwd}$$



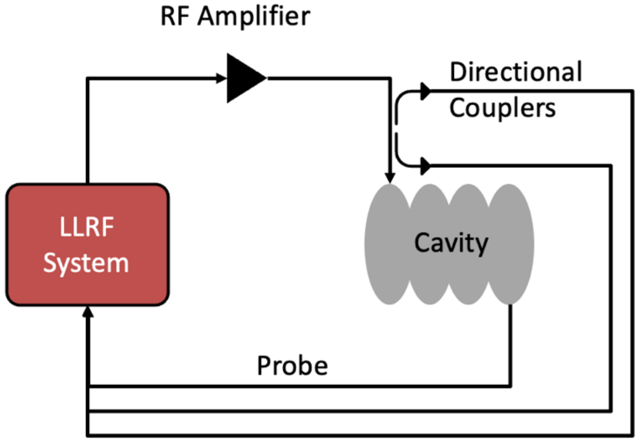


V_t: transmitted voltage | *R_L: loaded "shunt" resistance*

V_r: reflected voltage | *m: cavity/waveguide coupling ratio*

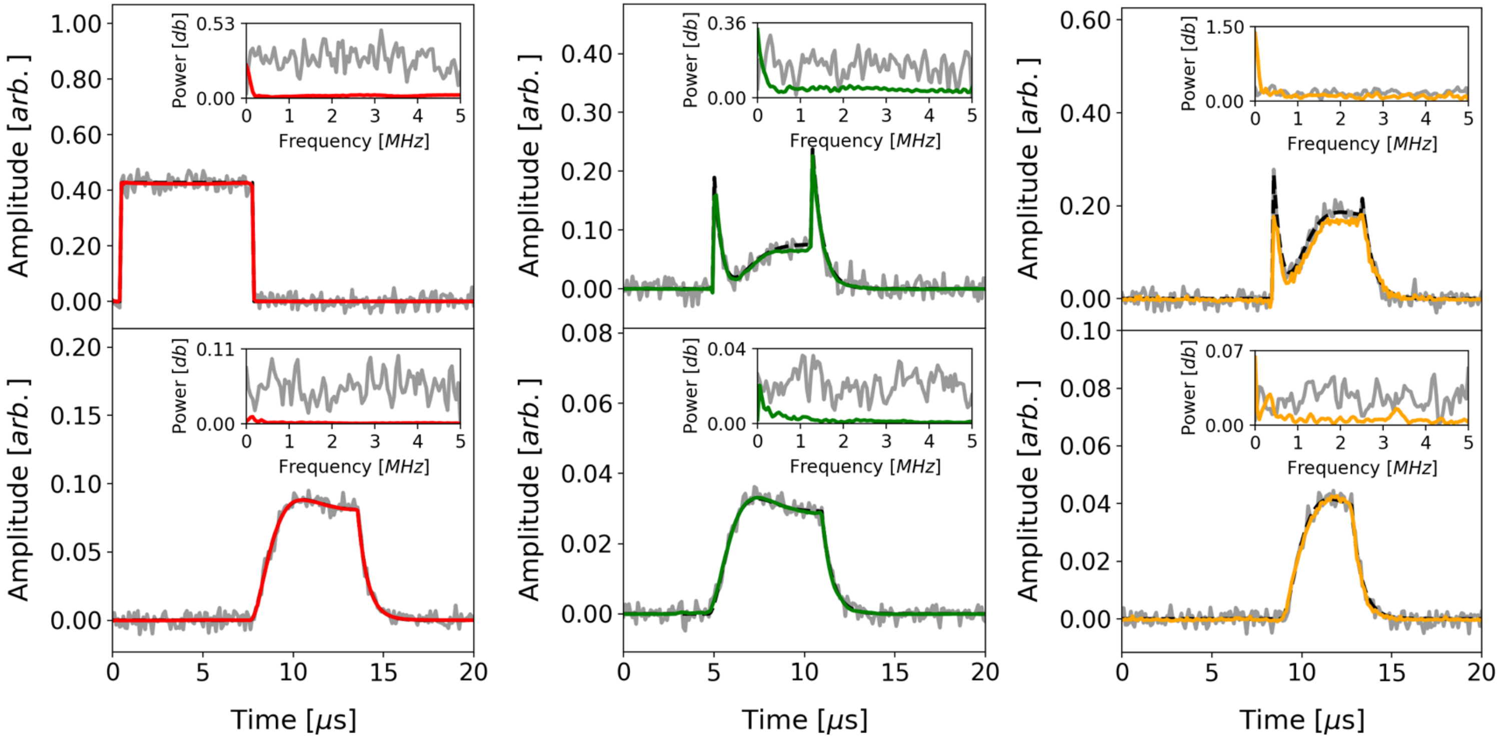
$\omega_{1/2}$: half band-width | *I_{fwd}: forward current*

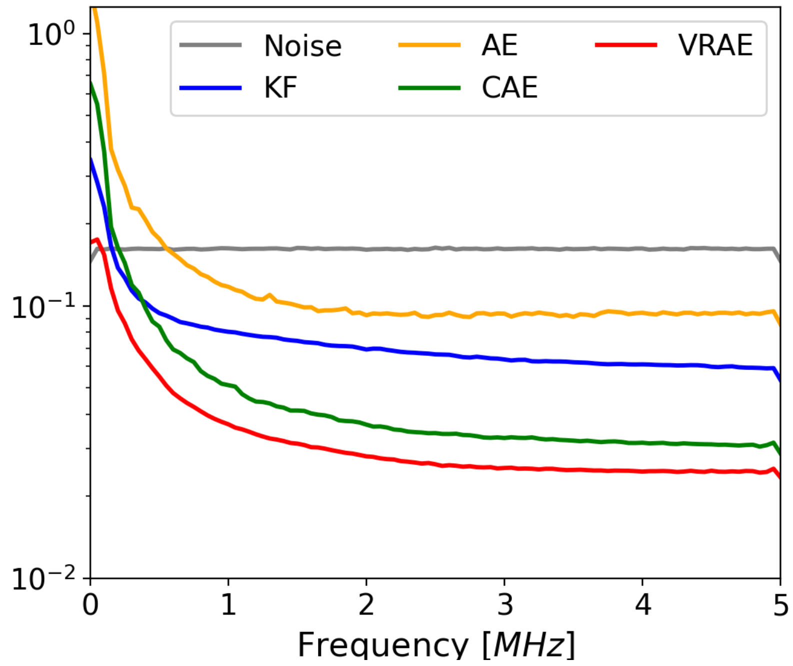
$\Delta\omega$: frequency detuning | *Z_0: reference impedance*



ML Based Noise Reduction

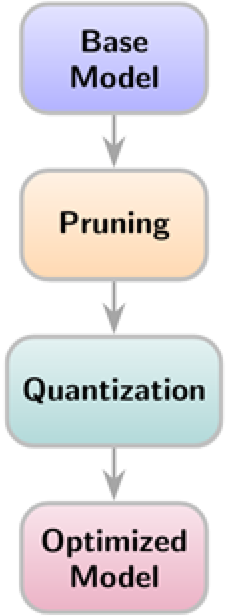
- Studying the efficacy of autoencoders for noise reduction in RF signals
  - Initial studies focused on amplitude data
  - Compared feed forward, convolutional, and variational architectures with conventional Kalman filtering





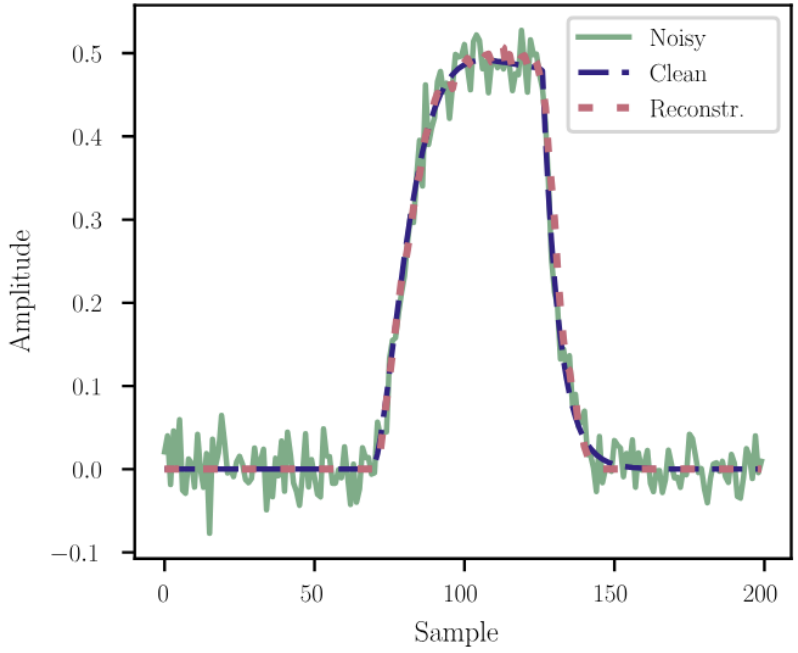
Assessment of FPGA Deployment Pipeline

- Targeting use in a real time pulsed feedback system
  - Model pruning - 75% / 80% / 87.5% reduction in weights
  - Quantization – int8 / 12-bit / 16-bit / 16x8 quantization schemes
  - Optimization of resource usage – test deployment on chosen architecture



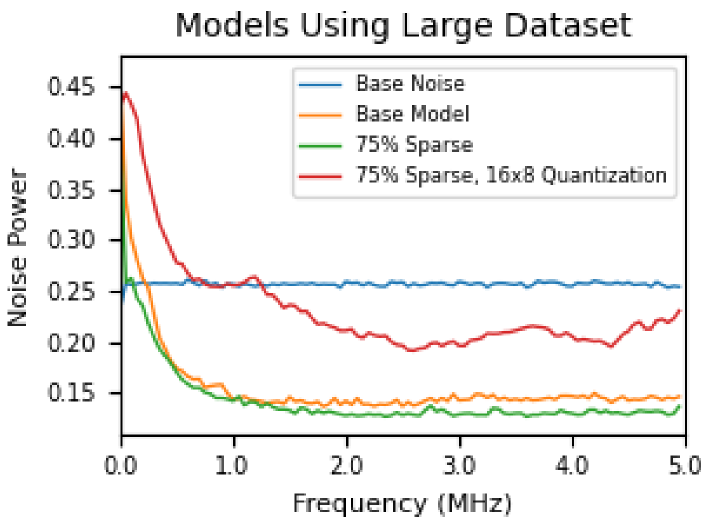
Assessment of FPGA Deployment Pipeline

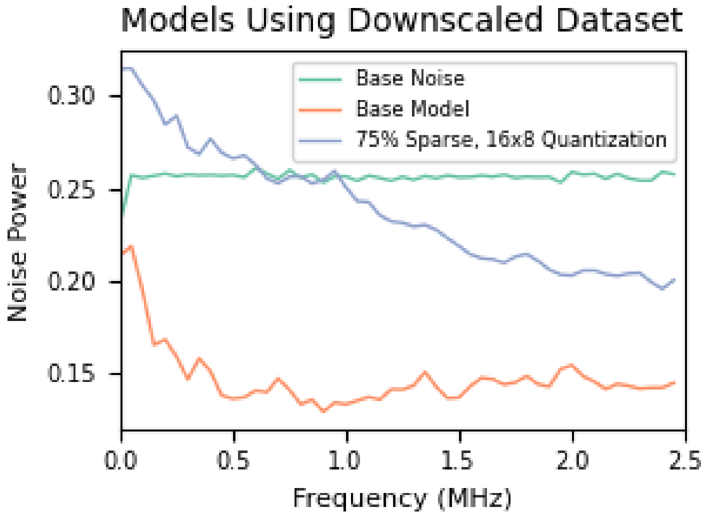
- Pruning and quantization show some reduction in performance on bulk figures of merit
  - Right: Residual noise power spectra for pruned and quantized models
  - Data down sampling required for the models to fit on the FPGA



Sparsity (%)	MSE Value	IMSE (W · Hz)
75	0.000175	707000
80	0.000178	714000
87.5	0.000208	775000

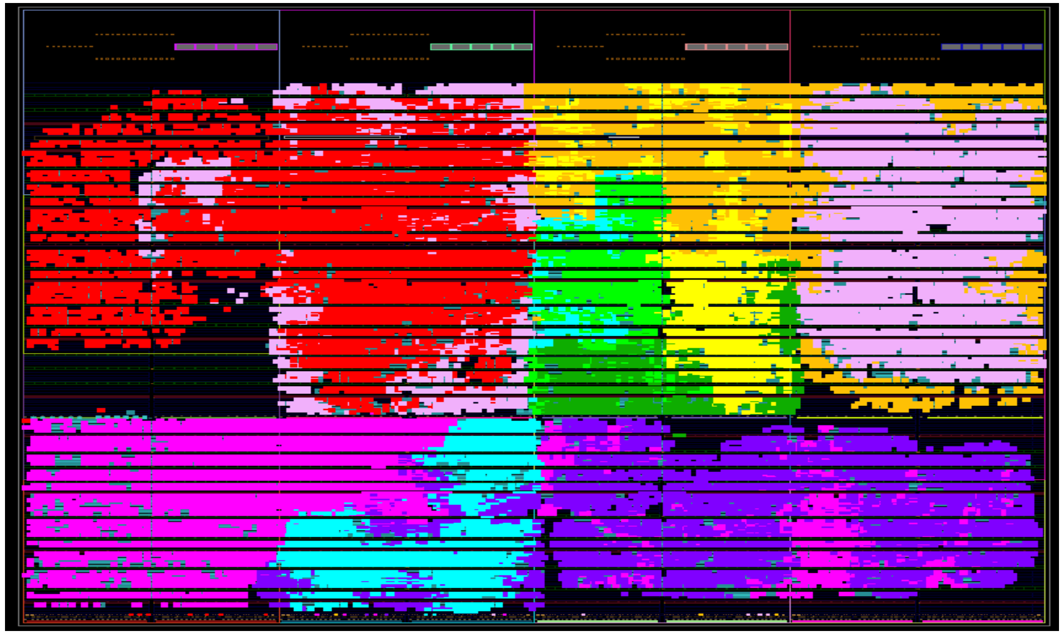
Quantization Scheme	MSE Value	IMSE (W · Hz)
int8	0.00120	1790000
12-bit	0.00122	1730000
16-bit	0.00121	1780000
16x8	0.000544	1160000





Assessment of FPGA Deployment Pipeline

- Targeting use in a real time pulsed feedback system
  - Left: FPGA floorplan for the hls4ml 'Resource' strategy for an Int16xInt8 network, with each layer colored. From input to output of network: red, orange, yellow, green, blue, dark purple, light purple, brown.

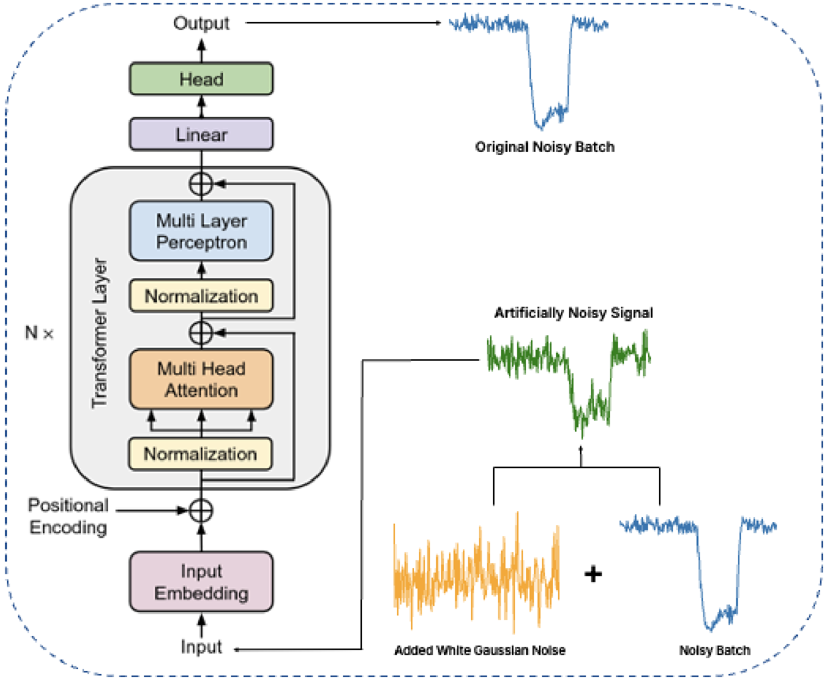


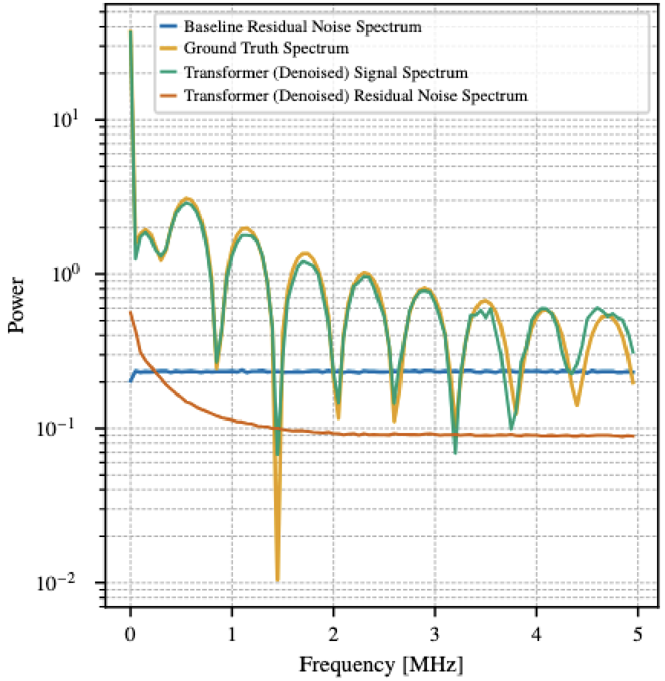
Resource usage for 16x8 model as synthesized by Vitis HLS 2024.2 and hls4ml

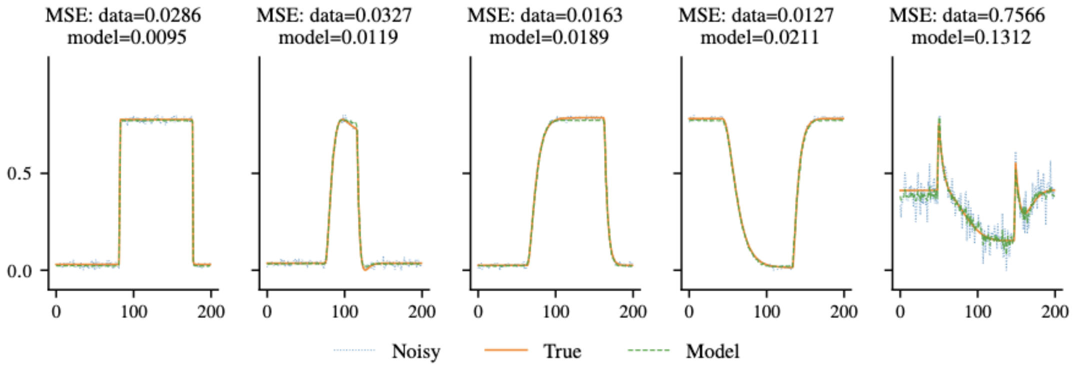
	Latency	Resource	ZCU104
Latency (Cycles)	1042	10004	
Look-Up Tables	245503	106357	274,080
Flip-Flops	238734	108378	548,160
DSP48E Slices	106	9	728
BRAM (36Kb)	0	23.50	440

Model based noise removal from RF signals

- Many studies using different architectures
  - CNNs autoencoders
  - Vanilla autoencoders
  - Kalman filters
  - Variational autoencoders
- Difficulty when transferring to I/Q data
- Moved to transformer models and modified training schema
- Variance in test data optimistic for implementation in pulse-to-pulse feedback

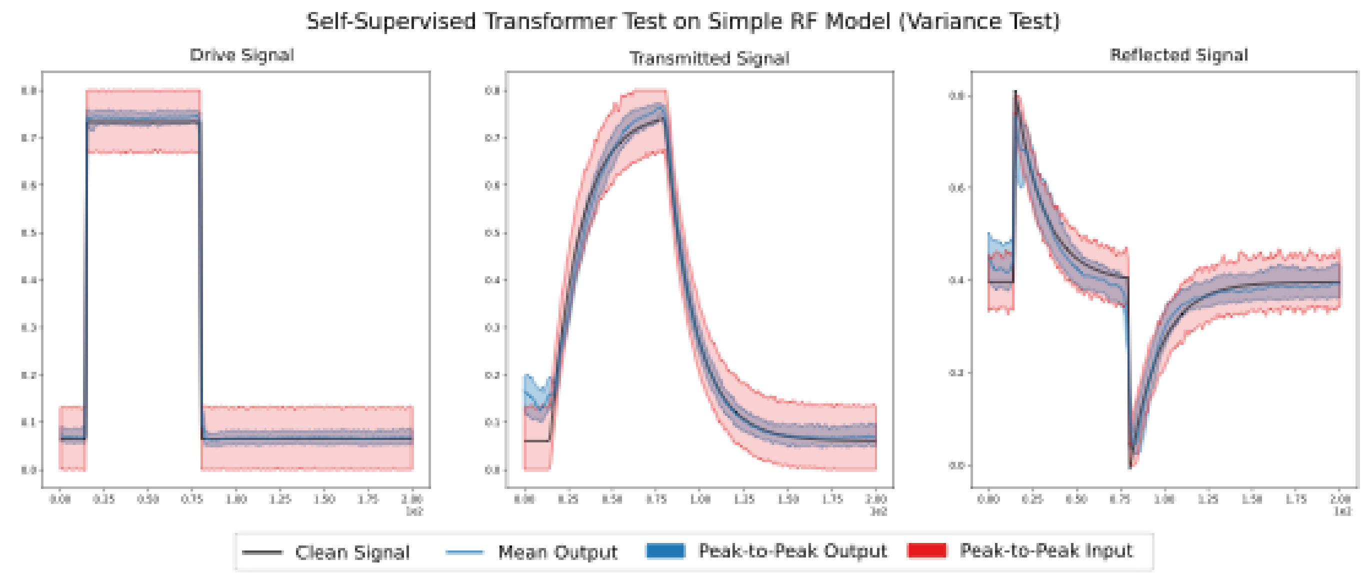


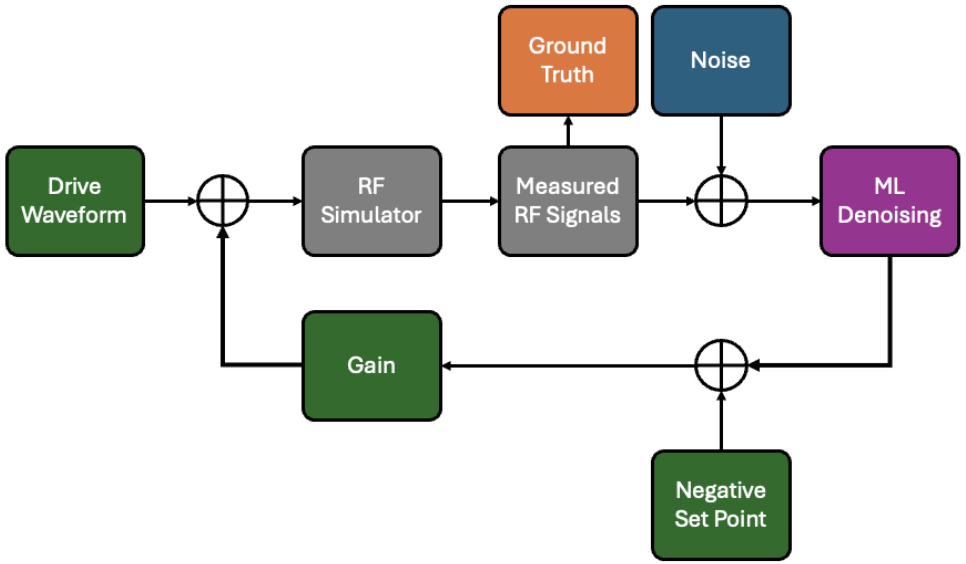




Evaluation of Feedback Performance

- Transformer model tested on simplified RF model – example is outside the training distribution
  - 100 waveforms were sent to the transformer with different noise signatures
  - Red: peak-to-peak of the input signal
  - Blue: peak-to-peak of the transformer output
- PID feedback used to regulate the RF pulse under slow drift

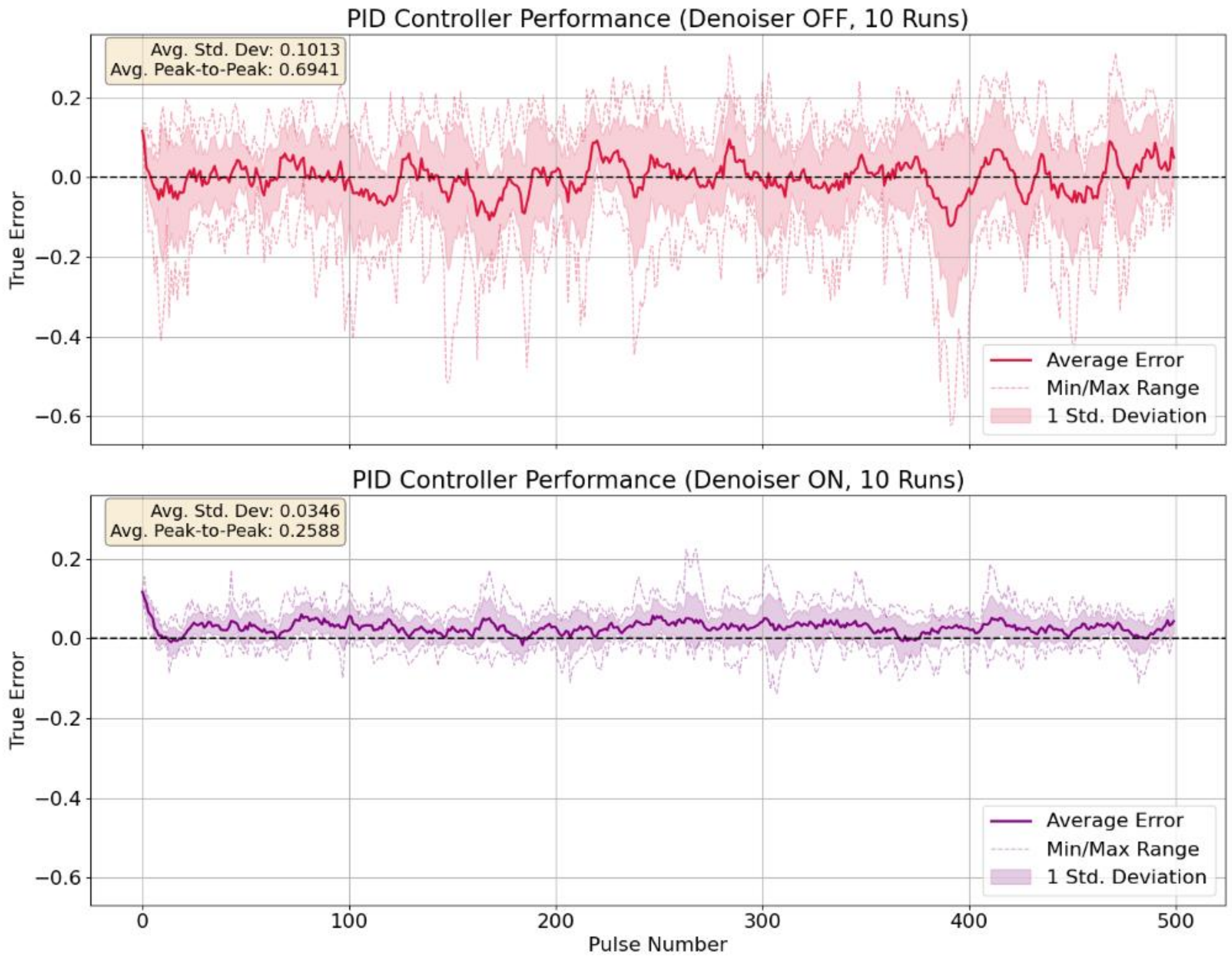




Conclusions

- Industrial accelerators have a large landscape of applications
  - growing demand for industrial systems
  - complexity of industrial accelerators is increasing
  - automation is critical when operating outside the laboratory environment

- Developing ML tools for automation
  - Initial studies focused on noise reduction
  - Various ML methods show promise for this application
  - Deployment path is taking shape
  - Simulated use in feedback systems shows promise ~65% reduction in error (both peak-to-peak and standard deviation)



Disclaimer

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.