Smartpixels

On-Pixel Featurization for Single-Layer Silicon Tracking

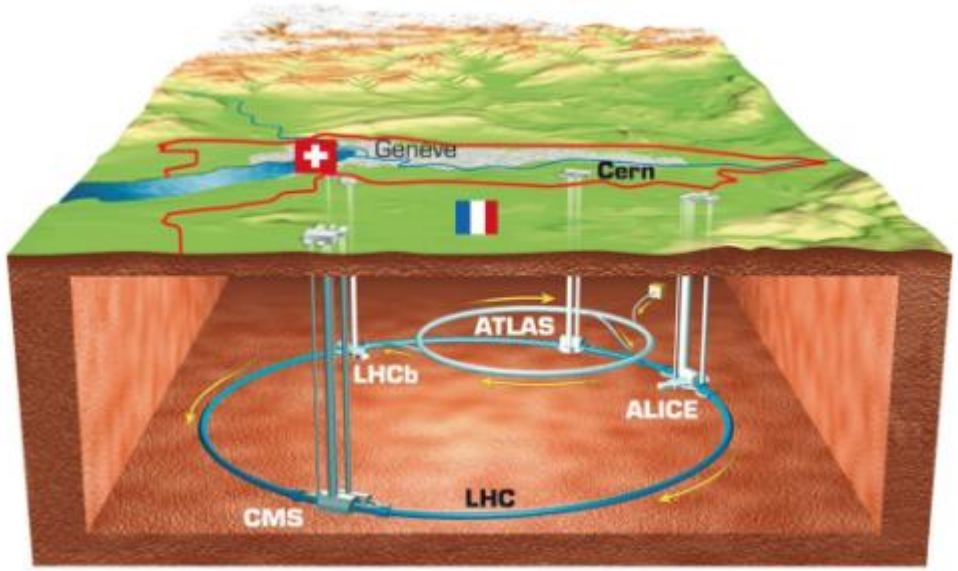
FastML – September 25, 2023

Rachel Kovach-Fuentes (University of Chicago)

with Douglas Berry, Jennet Dickinson, Giuseppe Di Guglielmo, Karri DiPetrillo, Farah Fahim, Lindsey Gray, Jim Hirschauer, Shruti Kulkarni, Ronald Lipton, Petar Maksimovic, Corrinne Mills, Benjamin Parpillon, Gauri Pradhan, Morris Swartz, Nhan Tran & Jieun Yoo

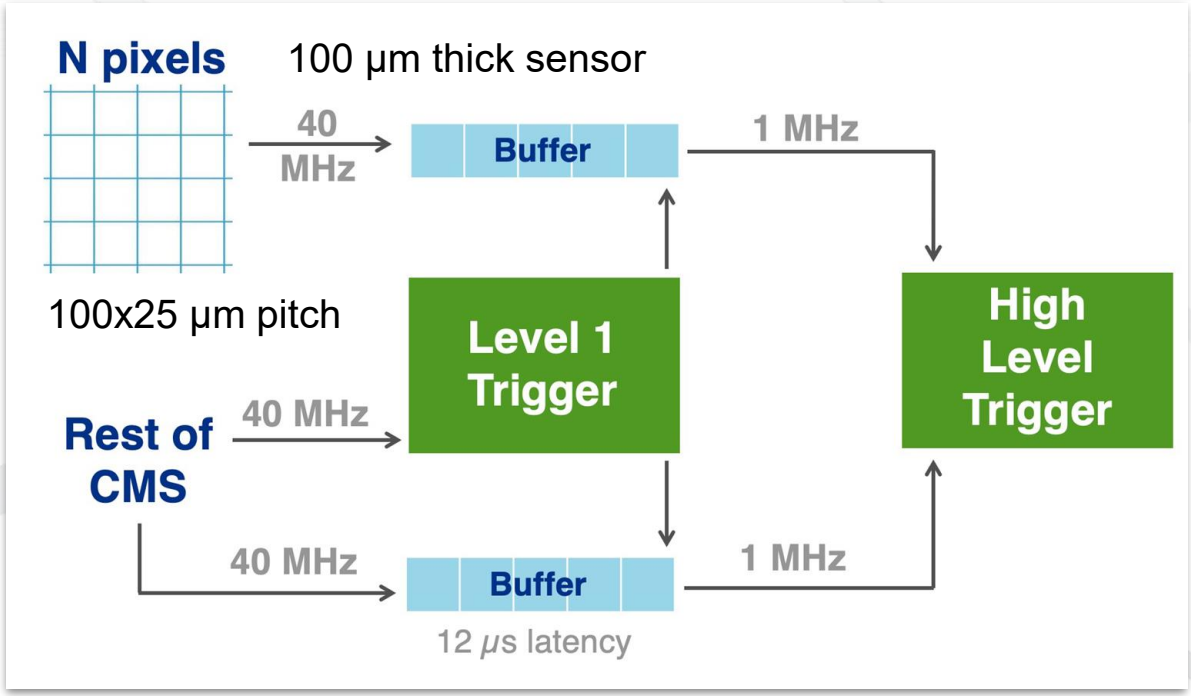
What is the LHC (Large Hadron Collider)?

- Proton-proton (pp) collider at CERN, on the French-Swiss Border
  - Collision energies are the highest ever reached
- Four experiments located where the proton beams intersect
- So far LHC has produced ~ 10^{16} pp collisions
  - 2 x 10^{17} collisions planned by 2040

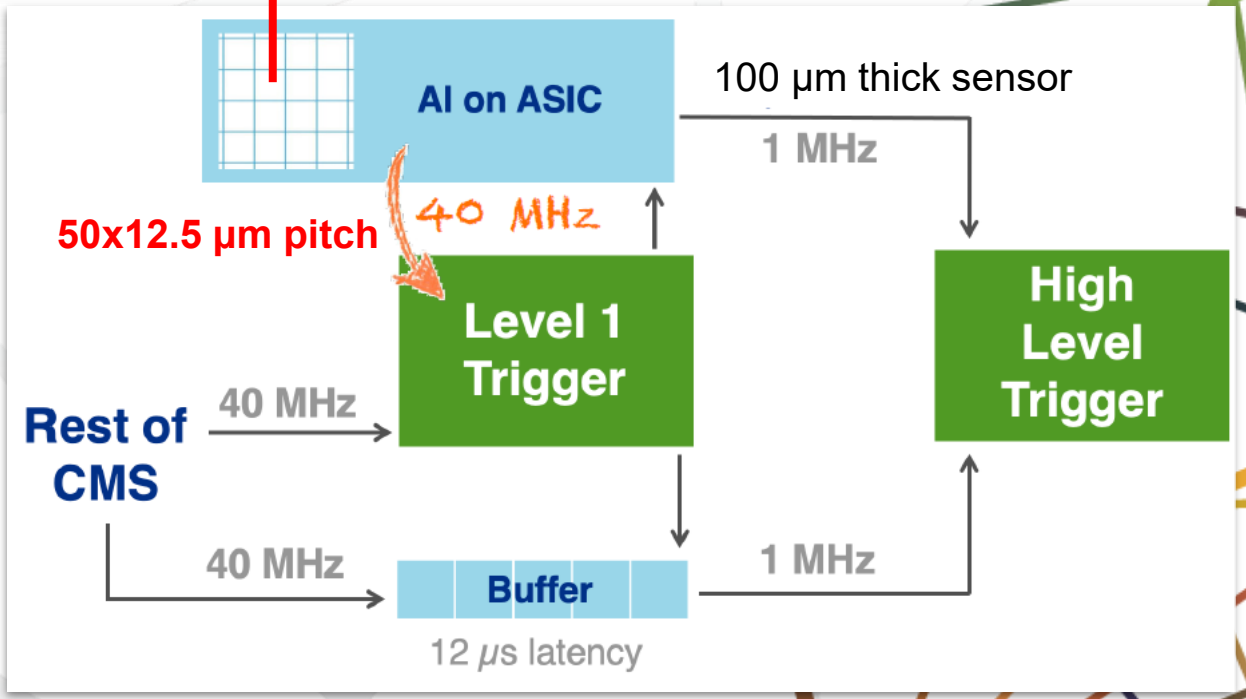


Motivation: Futuristic Detector

Current pixel readout chain:



Futuristic detector:



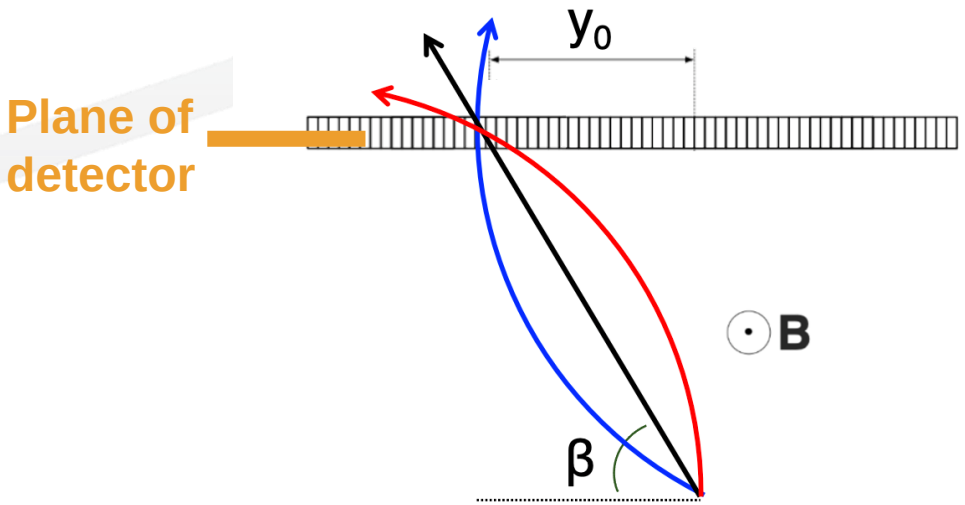
Problem:
Need to transfer 4-160x more data

Solution:
Implement AI in the detector readout electronics for fast data reduction

Goals for the algorithm

Extract useful properties from incident particle (hit position, incident angle)

- Predict means, free of bias
- Predict uncertainties, need to describe residuals
- Must be compact for FPGA or ASIC implementation



Process:

Design keras network

↓

Convert to qkeras

↓

hls4ml

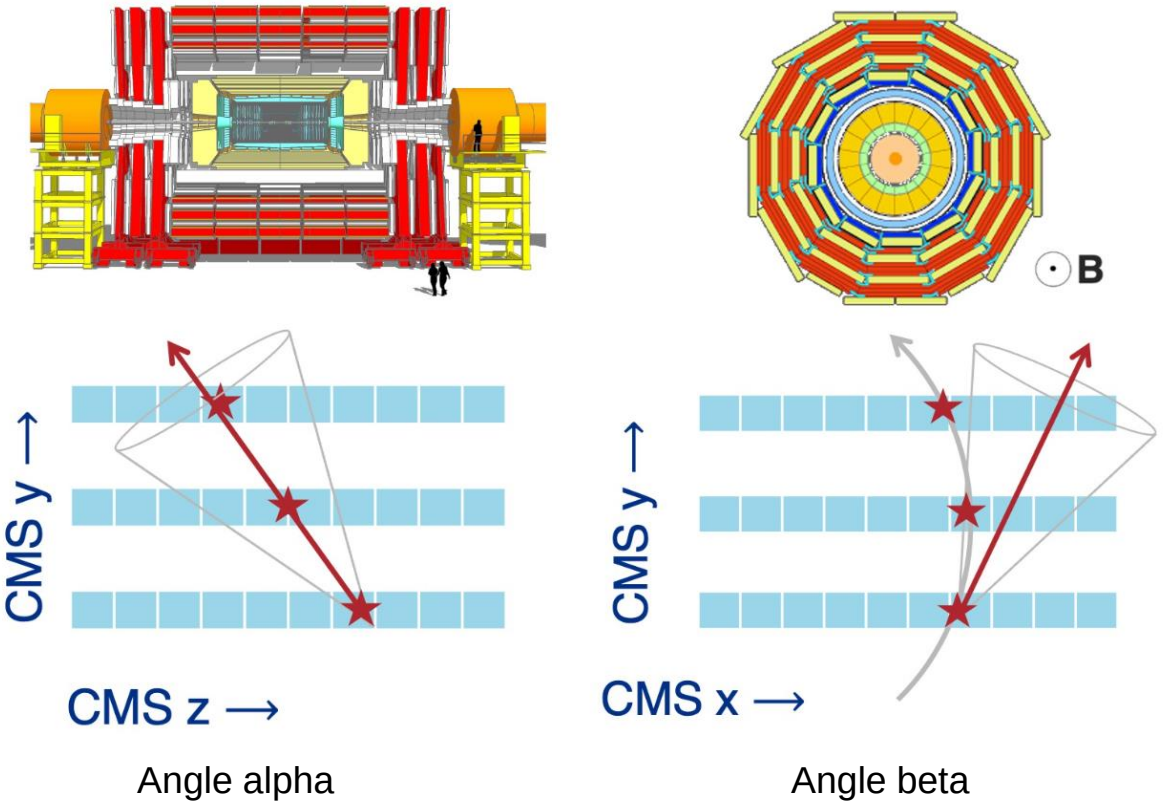
↓

FPGA/ASIC implementation

Note: Other methods of data reduction are also being explored, including filtering by transverse momentum.

See Jennet Dickinson's talk from MODE workshop 2023: https://indico.cern.ch/event/1242538/timetable/#96-smart-pixels-with-data-redu

Towards a pixel track trigger



Angle alpha

Angle beta

Problem: More complex final states → more hits → more hit combinations for track seeding

- Expensive, slow

Solution: Predicted angle + uncertainty → region of expected hits in the next layer

- Small uncertainty → small region
- Ignore the rest!

Fast tracking and vertexing

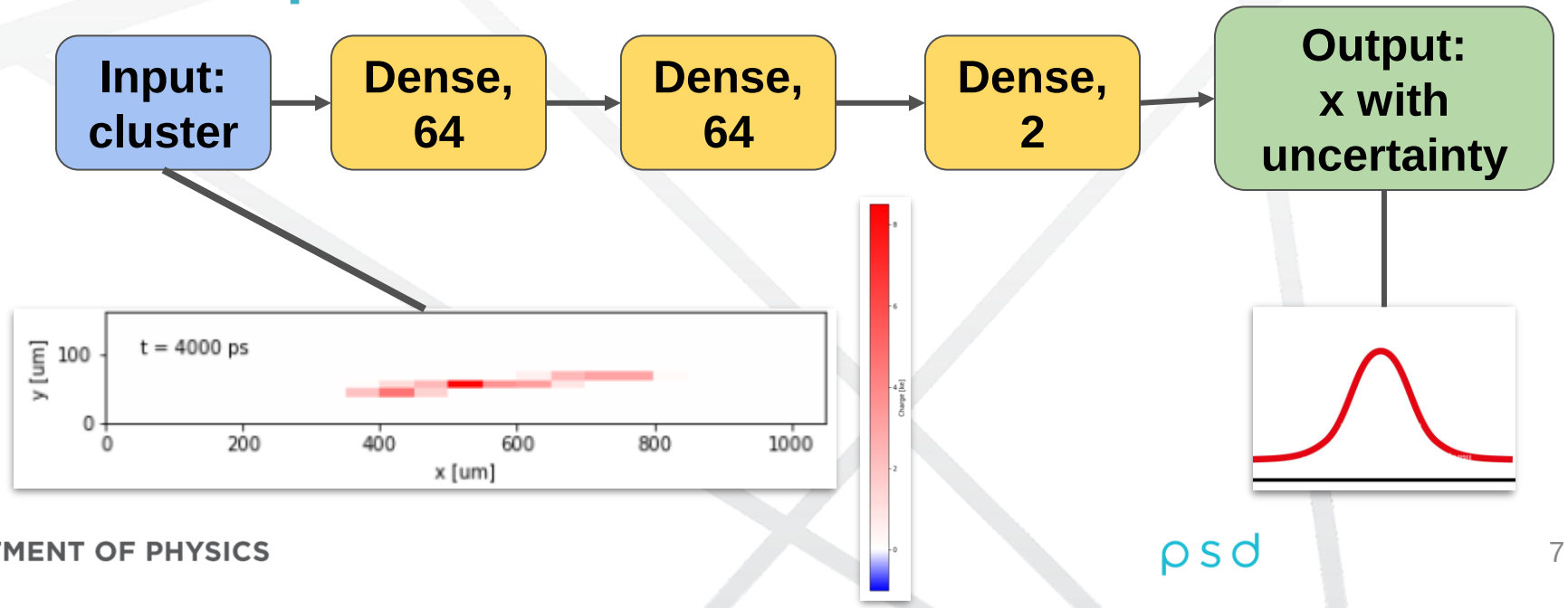
- Valuable for hh, e+e-, μμ
- At HL-LHC: makes L1 pixel trigger feasible?

Initial Network Architectures

Training Dataset:
Simulated MIP interactions in 21x13 array of pixels
Located at radius of 30 mm
3.8 T magnetic field
Time steps of 200 picoseconds

- Mostly dense layers
- Predicted 1, at most 2, parameters given an input cluster
- Did not include time information

Example: 1DX



Mixture Density Network

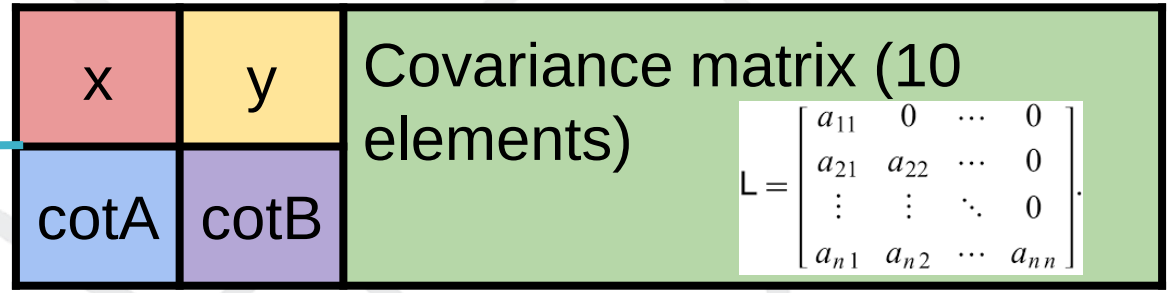
Use Mixture Density Network (MDN) to simultaneously fit the position, angles, and parameter errors of the incident tracks of all the clusters

- Predict the parameters of a multidimensional gaussian likelihood distribution
- Loss is a sum of these likelihoods over all clusters
- MDN presents interesting challenges for implementation on FPGA

Network should have:

- Greatest possible precision
- Extract largest possible amount of info
- Smallest possible network size
- Time information

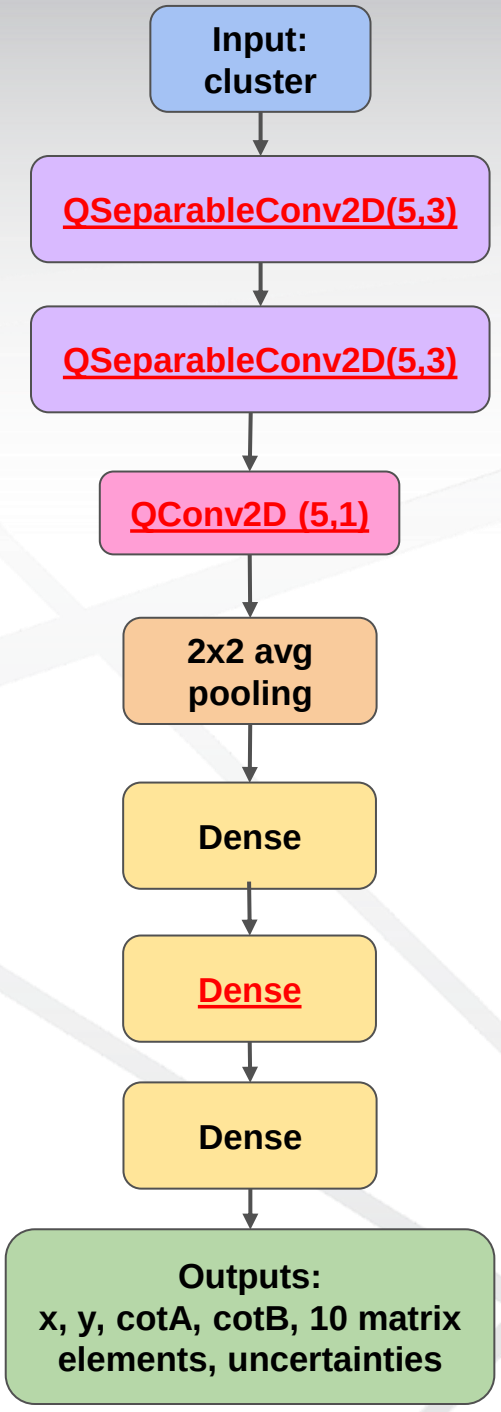
Total network outputs: 14



final network - qkeras

Use L1 regularization to force neurons to specialize.

Activations:
quantized_tanh(8,0,1) for convolutions,
quantized_relu(16, 4) for dense layers except last



quantized_bits(8,2,1, alpha=1)

quantized_bits(8,2,1, alpha=1)

quantized_bits(8,2, alpha=1)

quantized_bits(8,2, alpha=1)

quantized_bits(8,2, alpha=1)

quantized_bits(8,2, alpha=1)

quantized_bits(16,4, alpha=1) (wider b/c of output dynamic range)

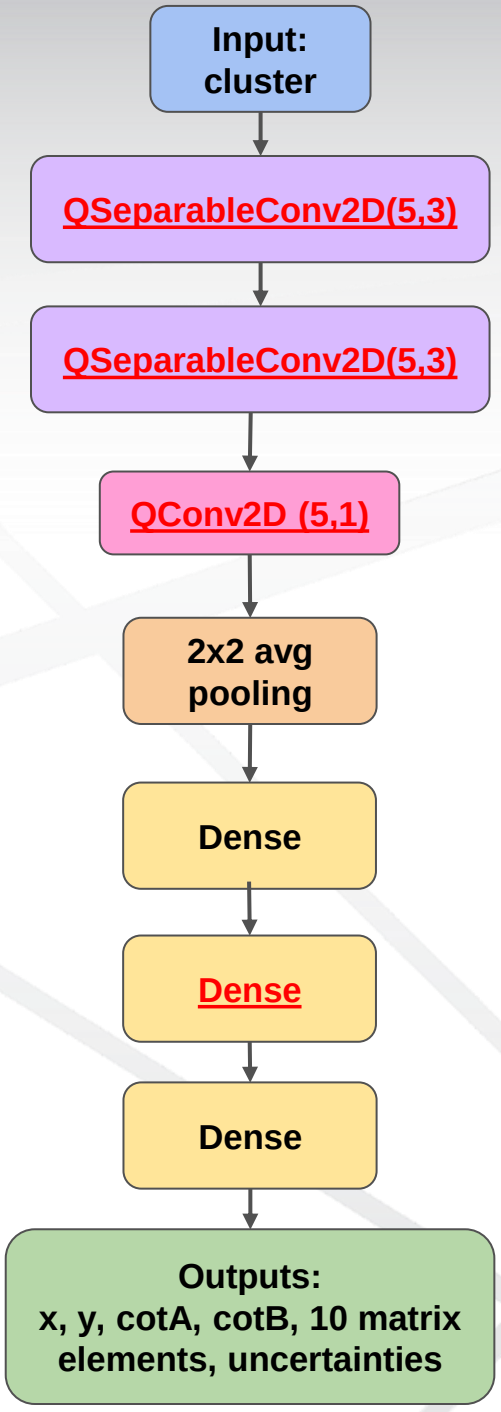
Params: 3,476

final network - hls4ml

Activations:
quantized_tanh(8,0,1) for convolutions,
quantized_relu(16, 4) for dense layers except last

Final precisions:
default_precision='fixed<23,7>'

Last layer requires 'fixed<25,9>' for result and accumulator to retain necessary output range.



quantized_bits(8,2,1, alpha=1)

quantized_bits(8,2,1, alpha=1)

quantized_bits(8,2, alpha=1)

quantized_bits(8,2, alpha=1)

quantized_bits(8,2, alpha=1)

quantized_bits(8,2, alpha=1)

quantized_bits(16,4, alpha=1) (wider b/c of output dynamic range)

Params: 3,476

Full Synthesis Resource Usage & Latency

Synthesis for:
Alveo U250
accelerator card
Includes fifo depth optimization.
Clock frequency 200MHz

```
VivadoSynthReport:
    LUT: 66307
    FF : 27153
    BRAM_18K: 12.5
    URAM: 0
    DSP48E: 341
```

```
CosimReport:
    RTL: Verilog
    Status: Pass
    LatencyMin: 299
    LatencyMax: 299
    IntervalMin: 276,
    IntervalMax: 276,
    LatencyAvg: 299.0
    IntervalAvg: 276.0
```

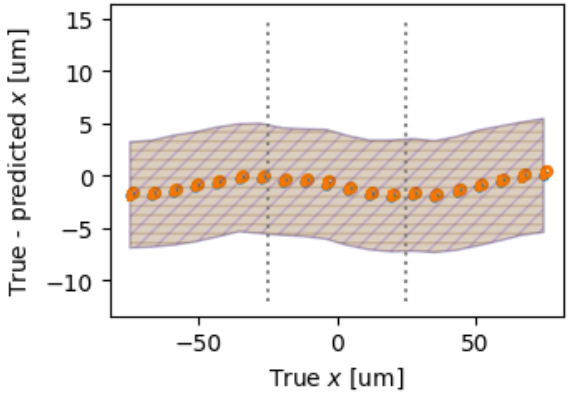
Not yet there for an ASIC, but sufficient as proof of concept.

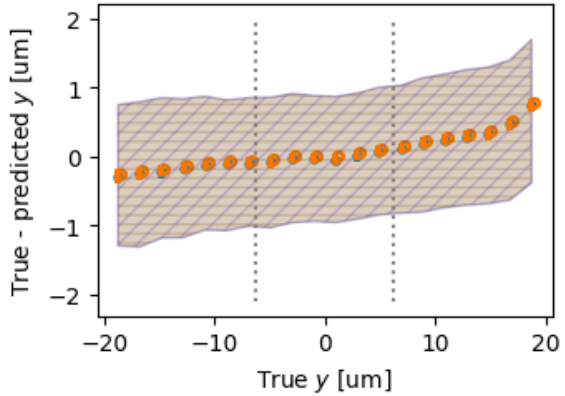
Directions for improvement:

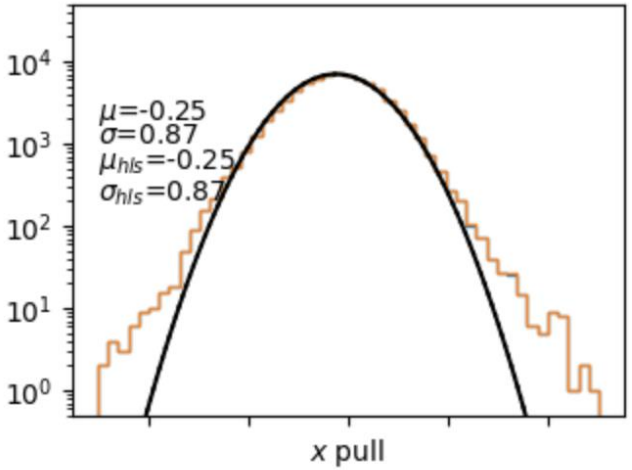
- Initiation interval must be 25ns (one interference starts every LHC clock)
- DSP Usage - complex multipliers too much floor space on chip
- Accumulators must be reduced to much smaller bitwidth

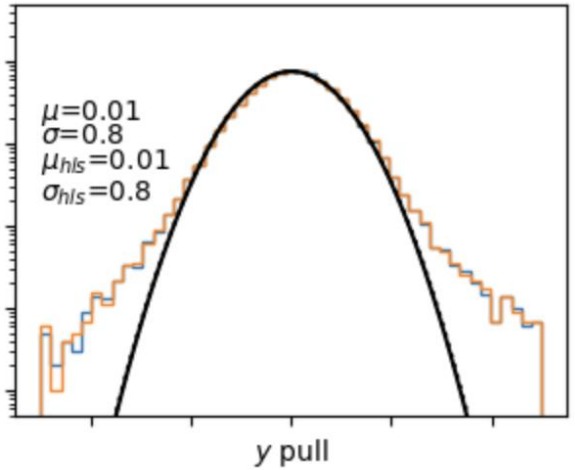
Bitwise agreement is not a major concern – achievable once above are addressed.

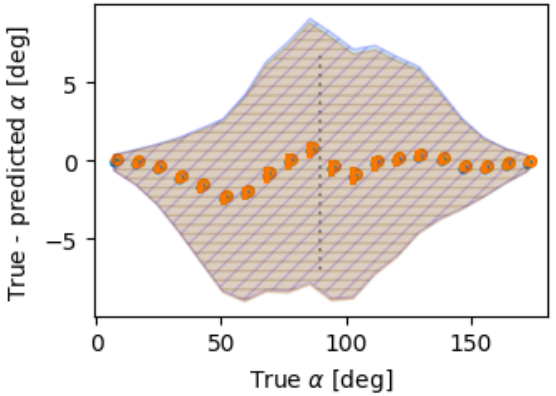
Final Network Performance

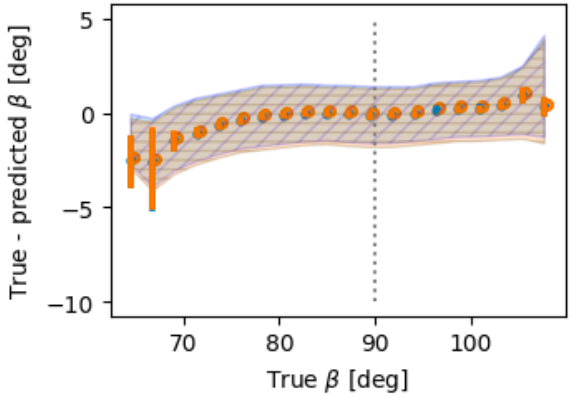


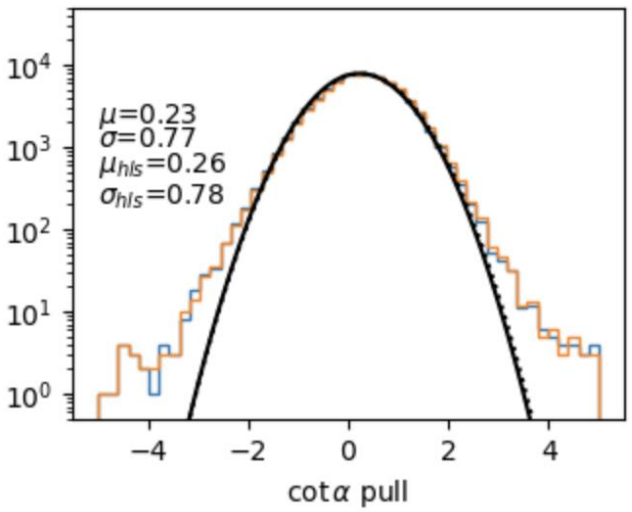


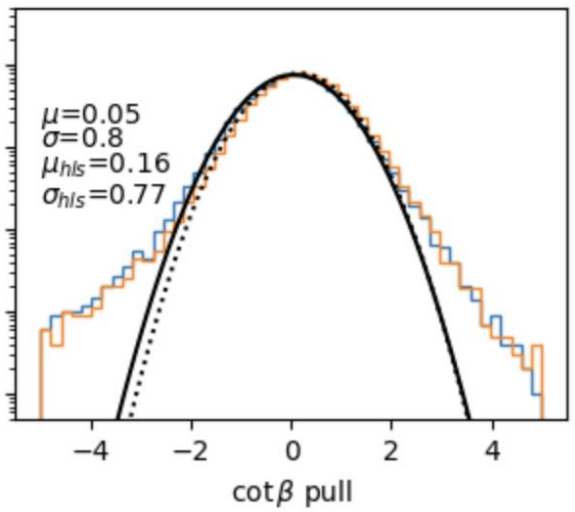






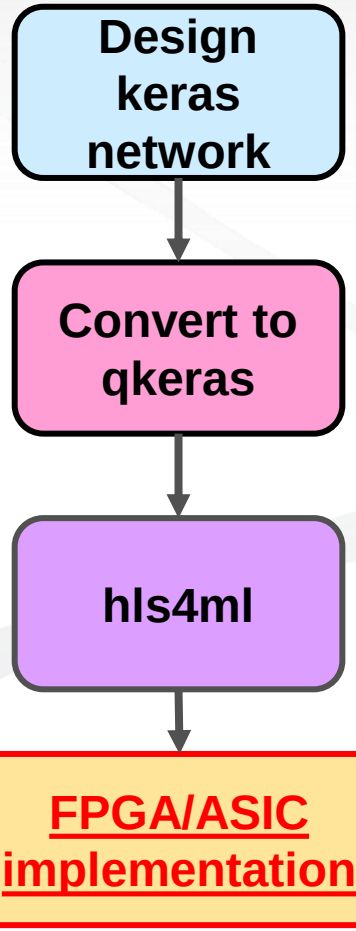






= qkeras = hls4ml

Next steps…



- **Deploy proof-of-concept on FPGA**
  - Demonstrate real time inference on hardware and validate against software network
- **Bring realism to proof-of-concept**
  - More accurate input data (200ps ADC not realistic)
  - Smaller bit widths for weights and accumulators
  - Improve initiation interval to be consistent with LHC
  - Estimate power consumption per inference
- **Find more use cases**
  - Real-time direction reconstruction has many uses

Acknowledgements

This work was completed using computing resources at the Fermilab Elastic Analysis Facility (EAF). We thank Burt Holzman for computing support.

We acknowledge the Fast Machine Learning collective as an open community of multi-domain experts and collaborators. This community, Javier Duarte and Vladimir Loncar in particular,were important for the development of this project.

DB, JD, GDG, FF, LG, JH, RL, BP, GP, CS and NT are supported by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the Department of Energy (DOE), Office of Science, Office of High Energy Physics. JD, FF, GDG, BP, GP, and NT are also supported by the DOE Early Career Research Program. NT is also supported by the DOE Office of Science, Office of Advanced Scientific Computing Research under the “Real-time Data Reduction Codesign at the Extreme Edge for Science” Project (DE-FOA-0002501). AB is supported through NSF-PHY award 2013007. MS is supported by NSF-PHY award 2012584. CM is supported by NSF-PHY award 2208803.

KD is supported in part by the Neubauer Family Foundation Program for Assistant Professors and the University of Chicago. RK is supported by the Metcalf Fellowship program of the University of Chicago. AY and SK are supported by the DOE Office of Science Research Program for Microelectronics Codesign (sponsored by ASCR, BES, HEP, NP, and FES) through the Abisko Project. MSN is supported through NSF cooperative agreement OAC-2117997, the DOE Office of Science, Office of High Energy Physics, under Contract No. DE-SC0023365, REFERENCES 24 and the Discovery Partners Institute under the ”Democratizing AI Hardware with an Open-Source AI-Chip Design Toolkit” Project.

Estimated Resource Usage & Latency

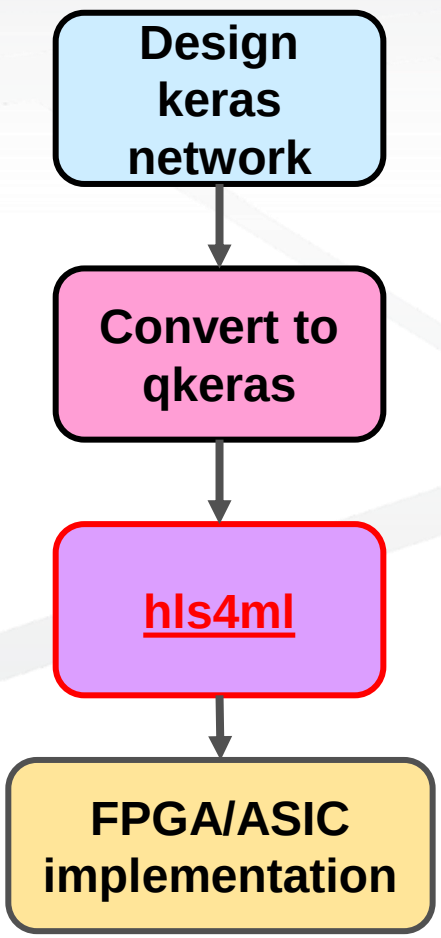
Estimates for:

Alveo U250 accelerator card

```
================================================================
== Utilization Estimates
================================================================
* Summary:
+---------------------+---------+-------+---------+---------+------+
|         Name        | BRAM_18K| DSP48E|    FF   |   LUT   | URAM |
+---------------------+---------+-------+---------+---------+------+
|DSP                  |        -|      -|        -|        -|     -|
|Expression           |        -|      -|        0|        2|     -|
|FIFO                 |       55|      -|     2555|     4336|     -|
|Instance             |       50|    339|    38729|   132161|     -|
|Memory               |        -|      -|        -|        -|     -|
|Multiplexer          |        -|      -|        -|        -|     -|
|Register             |        -|      -|        -|        -|     -|
+---------------------+---------+-------+---------+---------+------+
|Total                |      105|    339|    41284|   136499|     0|
+---------------------+---------+-------+---------+---------+------+
|Available SLR        |     1344|   3072|   864000|   432000|   320|
+---------------------+---------+-------+---------+---------+------+
|Utilization SLR (%)  |        7|     11|        4|       31|     0|
+---------------------+---------+-------+---------+---------+------+
|Available            |     5376|  12288|  3456000|  1728000|  1280|
+---------------------+---------+-------+---------+---------+------+
|Utilization (%)      |        1|      2|        1|        7|     0|
+---------------------+---------+-------+---------+---------+------+

+ Latency:
    * Summary:
    +---------+---------+----------+----------+-----+-----+----------+
    |  Latency (cycles) |  Latency (absolute) |  Interval | Pipeline |
    |   min   |   max   |    min   |    max   | min | max |   Type   |
    +---------+---------+----------+----------+-----+-----+----------+
    |      278|      278| 1.390 us | 1.390 us |  276|  276| dataflow |
    +---------+---------+----------+----------+-----+-----+----------+
```

hls4ml Conversion

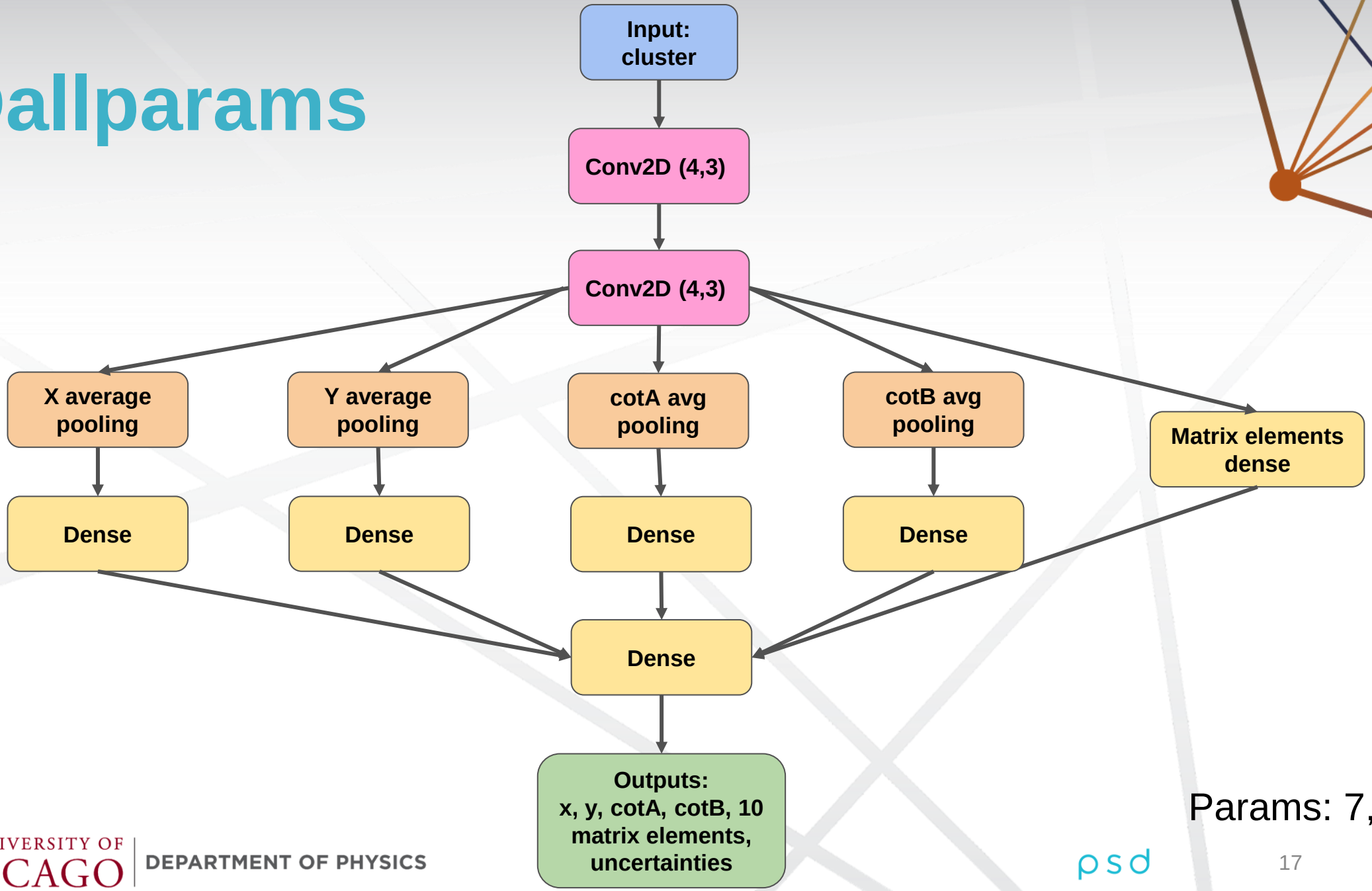



```
alid", data_format=None)(x_conv[...,:1])
valid", data_format=None)(x_conv[...,1:2])
"valid", data_format=None)(x_conv[...,2:3])
"valid", data_format=None)(x_conv[...,3:4])
valid", data_format=None)(x_conv[...,4:5])
```

Problem: slicing in our model couldn't be converted to hls

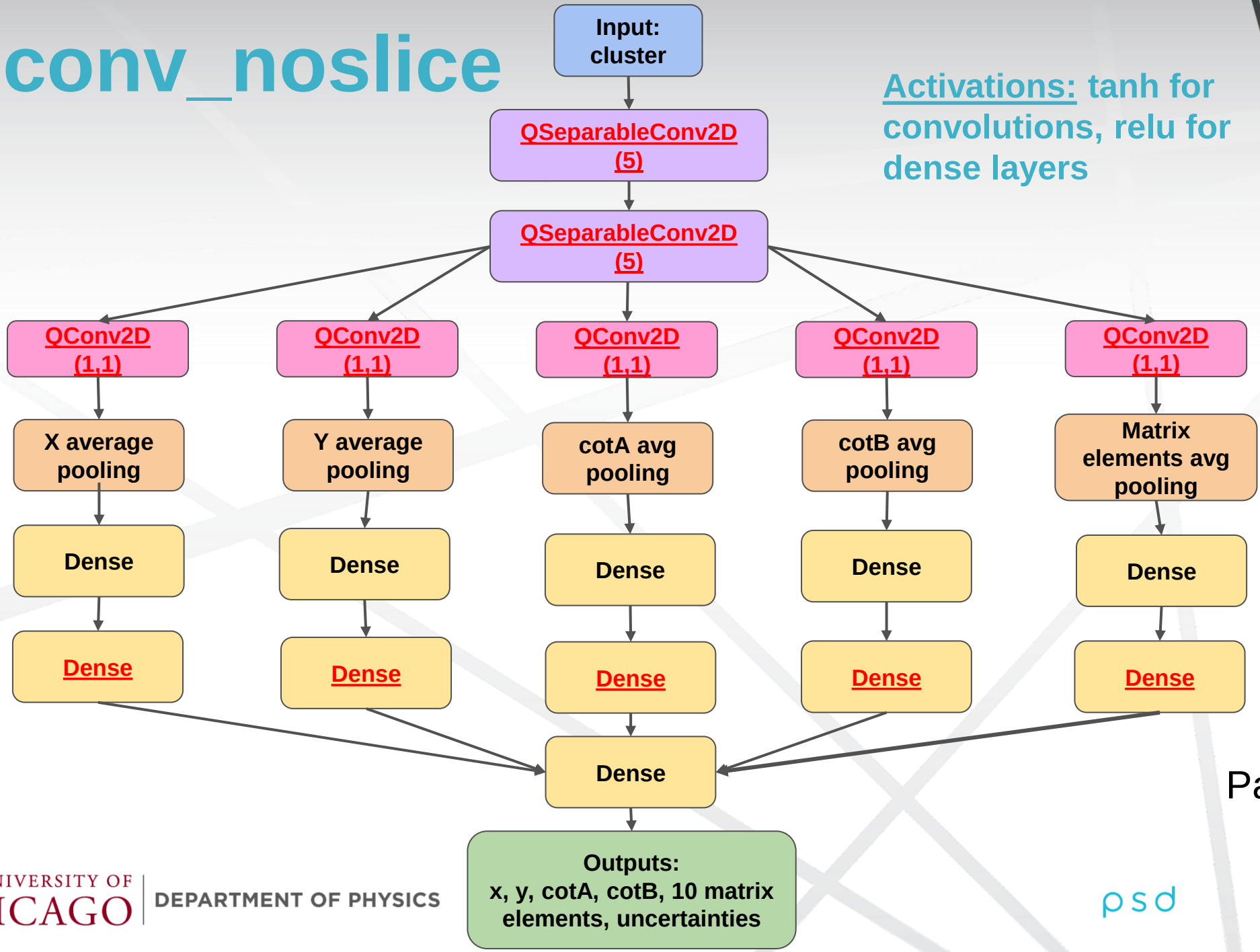
Solution: a new model architecture without slicing

3Dallparams



Params: 7,891

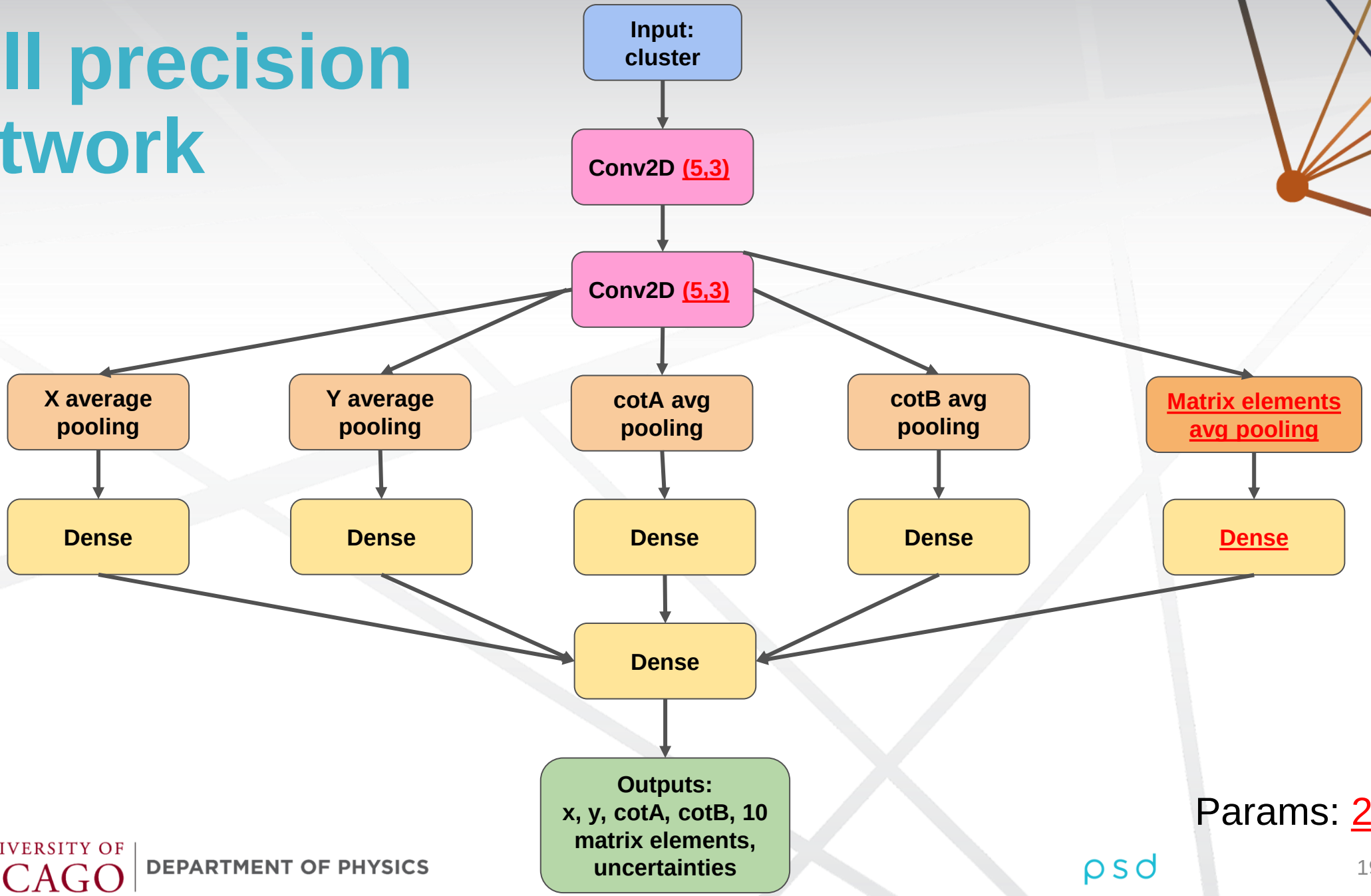
tanhconv_noslice



Activations: tanh for convolutions, relu for dense layers

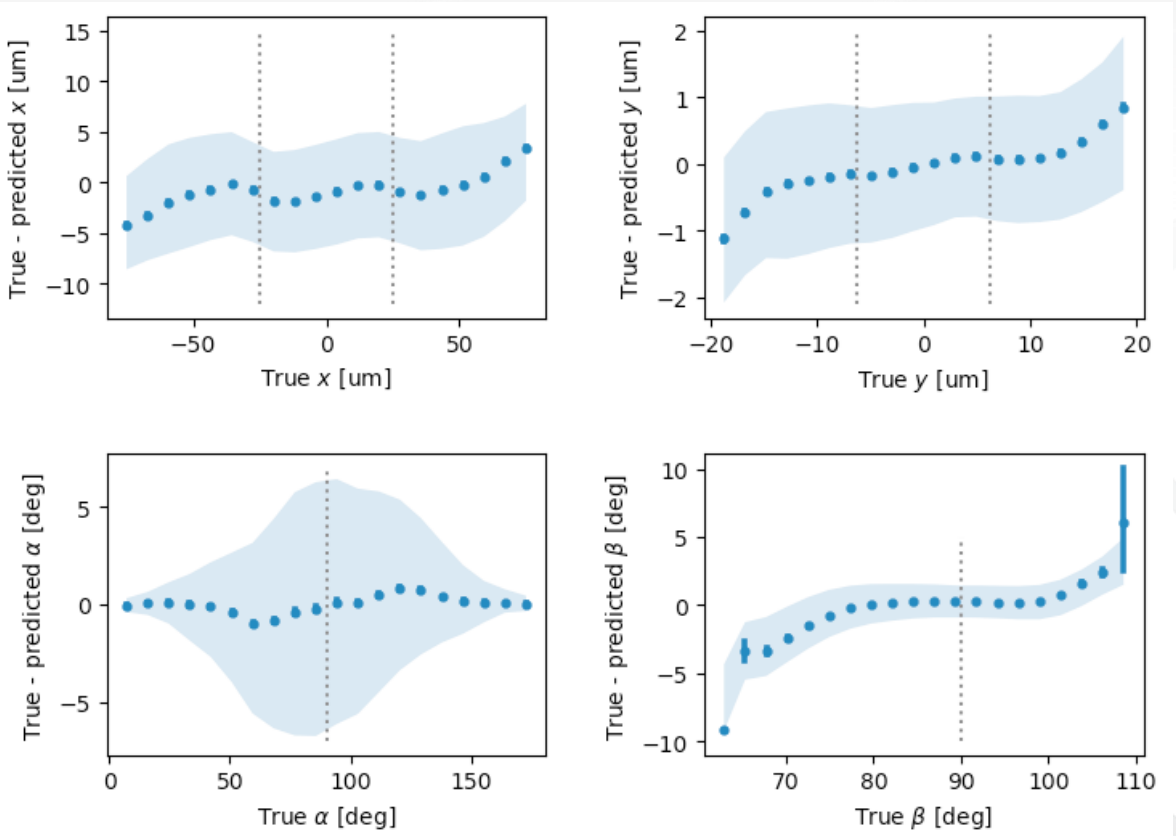
Params: 2,184

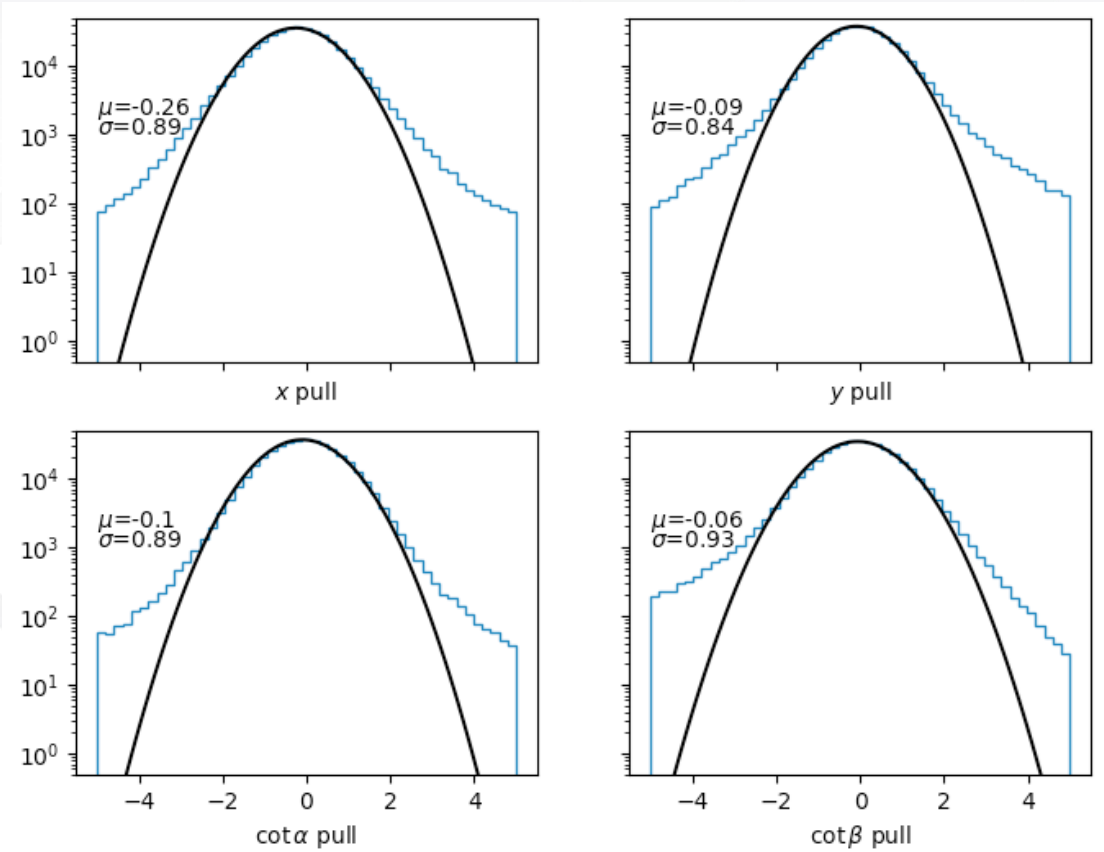
Full precision network



Params: 2,181

tanhconv_noslice Performance





Full precision network performance