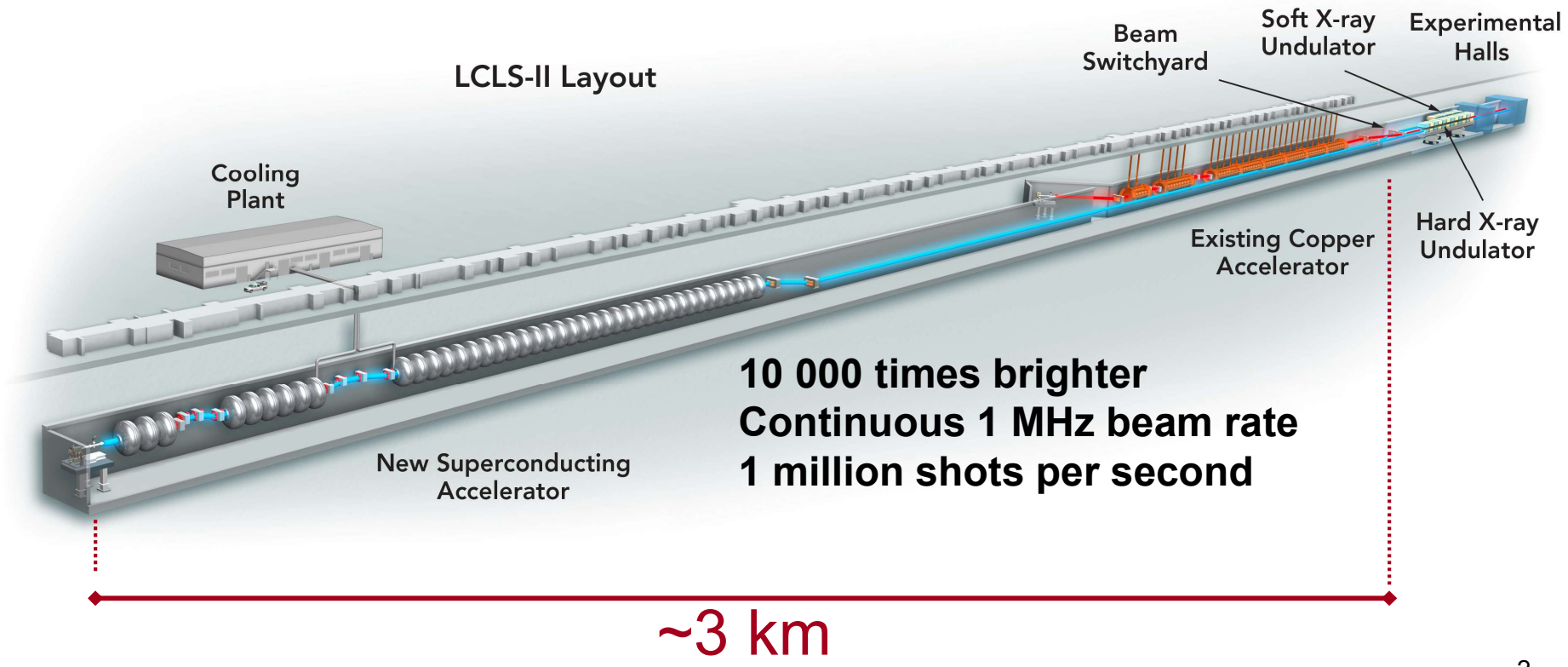
Machine Learning at the Edge: Intelligent Data Triage in Real Time

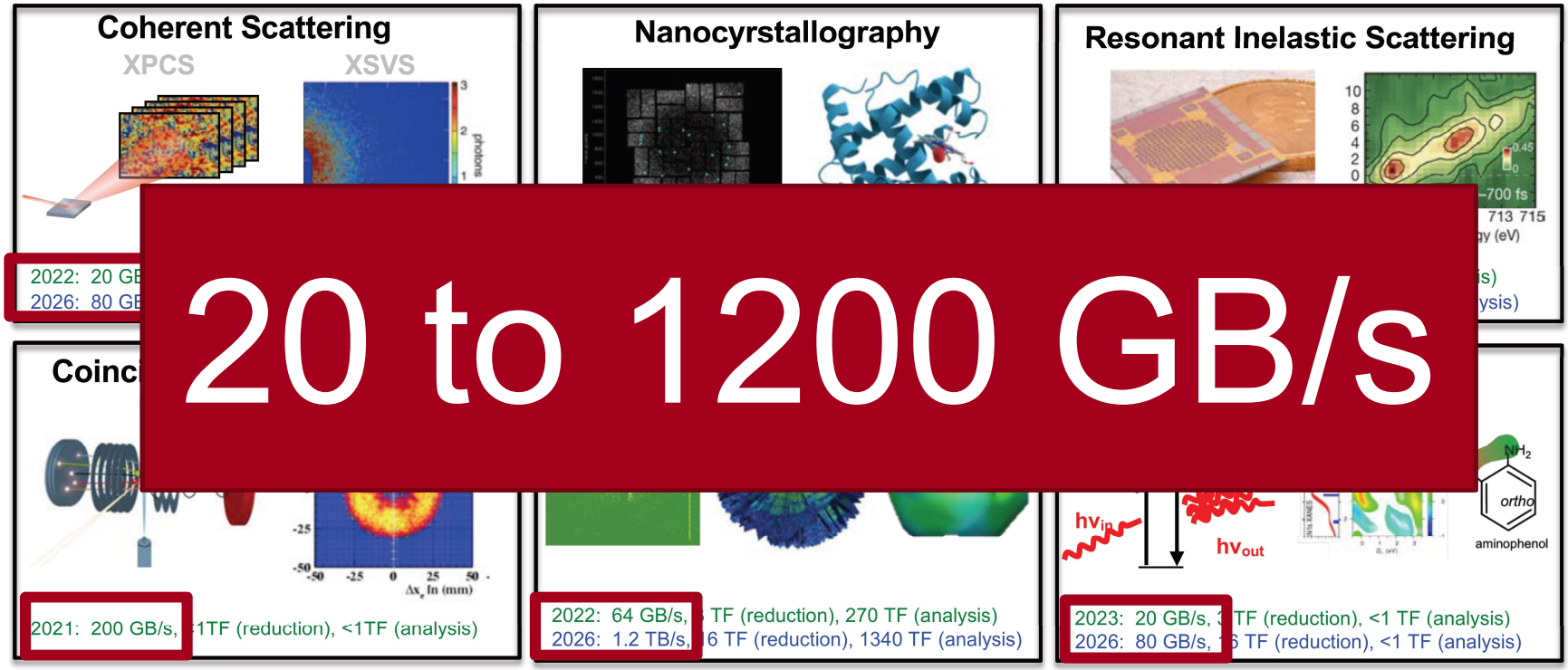
Audrey Corbeil Therrien

21.02.2020
DESY Hamburg
Joint Instrumentation Seminar

LINAC Coherent Light Source - II



LCLS-II



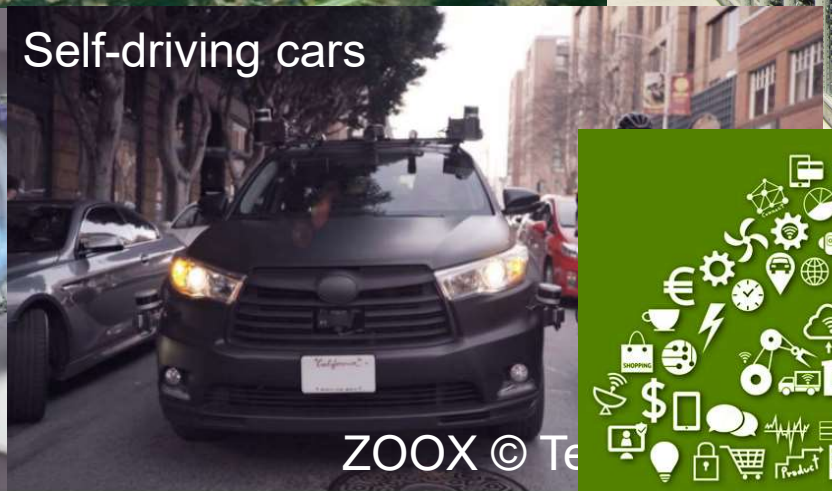
20 to 1200 GB/s

Image courtesy of Jana Thayer, Mike Dunne

Increased data production

SLAC

Price et al. 2014



ZOOX © Te

Increased data production

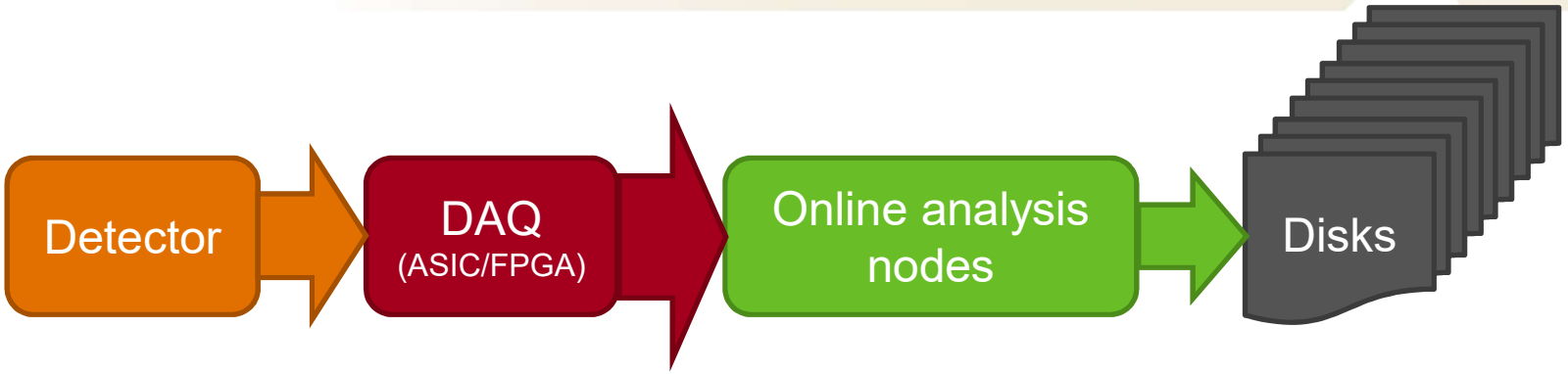
SLAC

- Economical and environmental impact - More data means:
  - More fibers to the DAQ system and storage
  - More power to transmit (FCC projected 2 MW for links alone*)
  - More storage hardware (disks, tape, etc.)
  - More power to data centers hosting hardware
  - More people to manage data centers
  - More data mining - more people needed to do that data mining
  - More power needed to compute for data mining

- Drowning in data – **focus on meaningful information**

*Projections by Dr Bortoletto

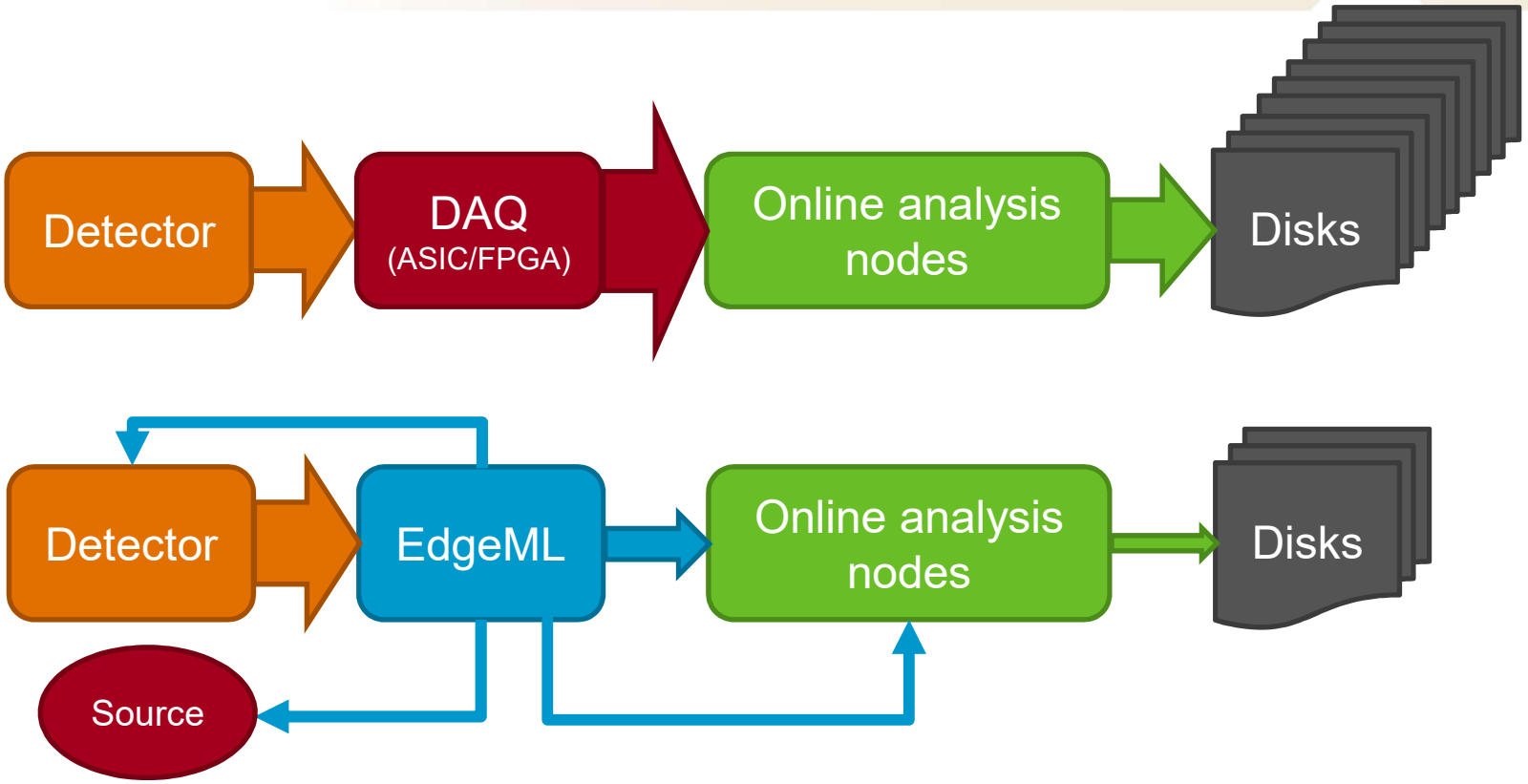
Real-Time Data Reduction at the Edge



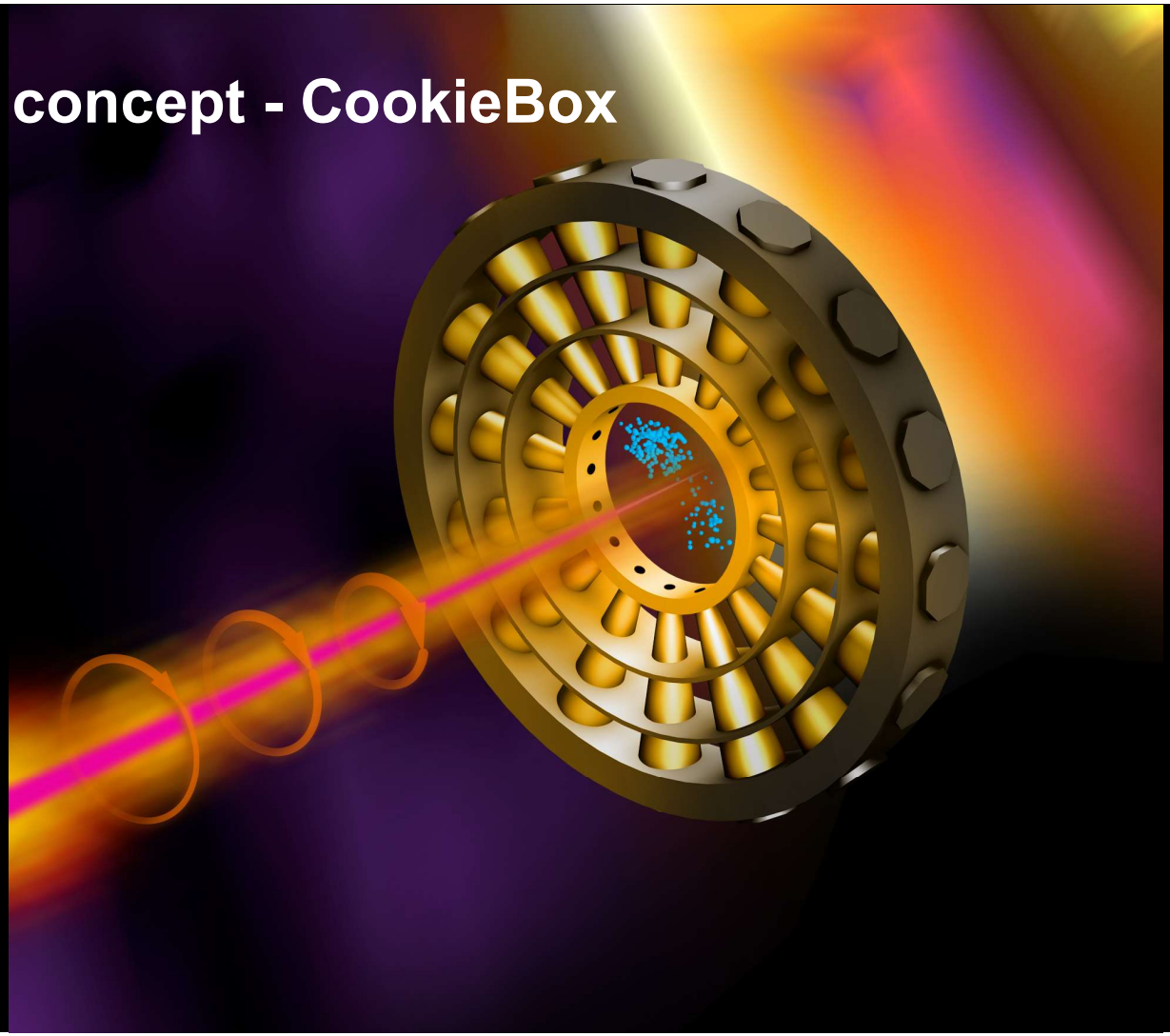
- Assuming 1 TB/s, 12 hour shift, nonstop
- 43 200 TB per shift – 56 years of 4K movies
- 1.3 million$/month of storage costs created every shift

Real-Time Data Reduction at the Edge

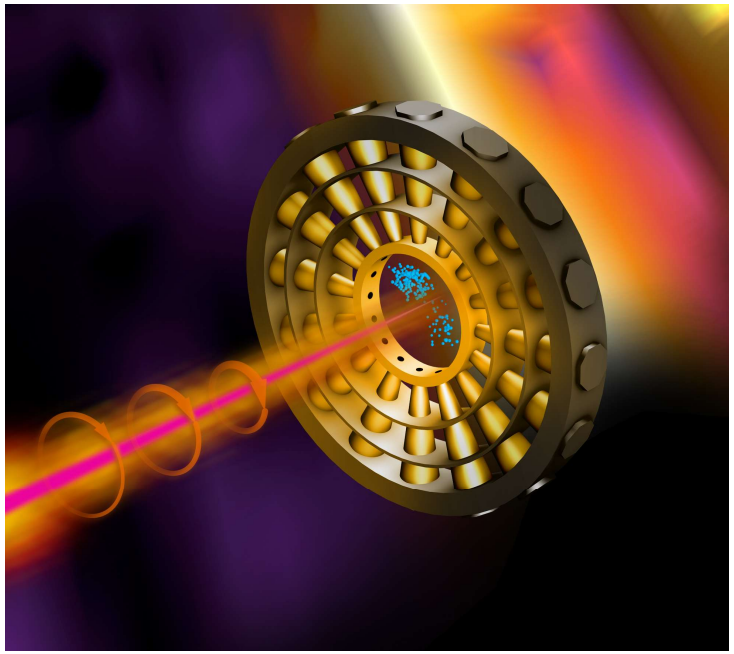
SLAC

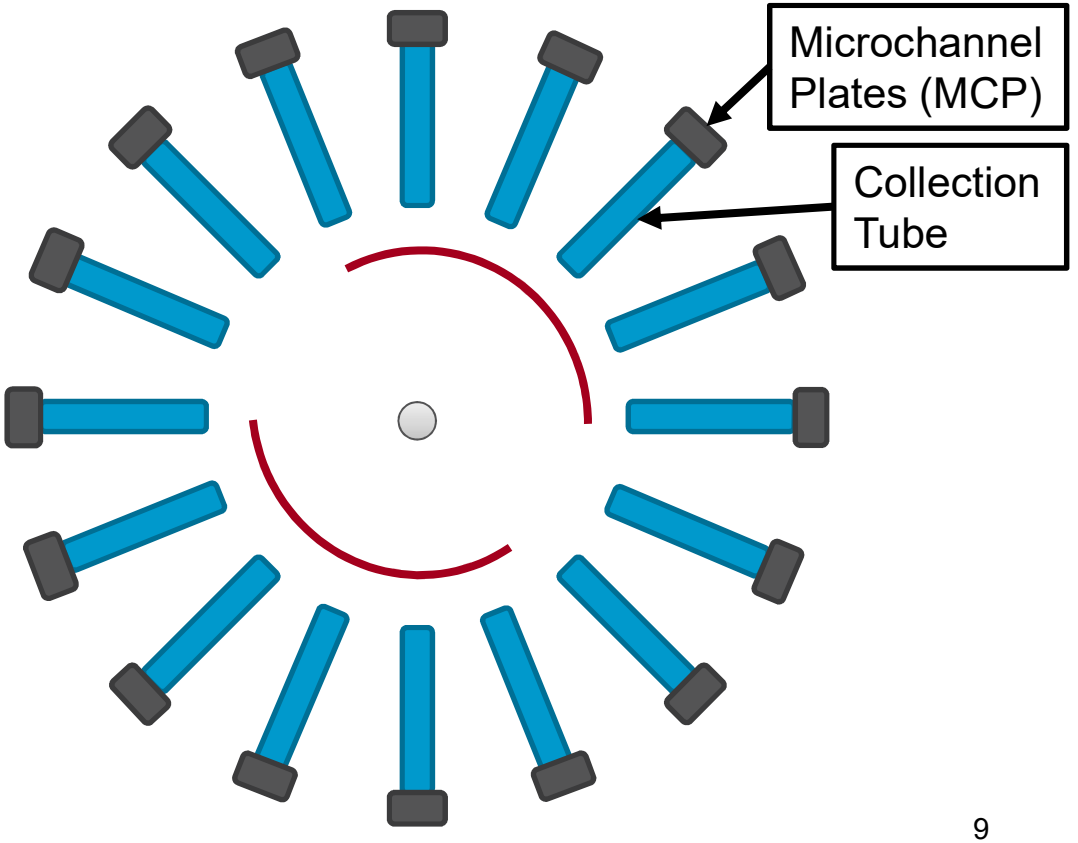


Proof of concept - CookieBox



CookieBox – Angular Streaking Detector



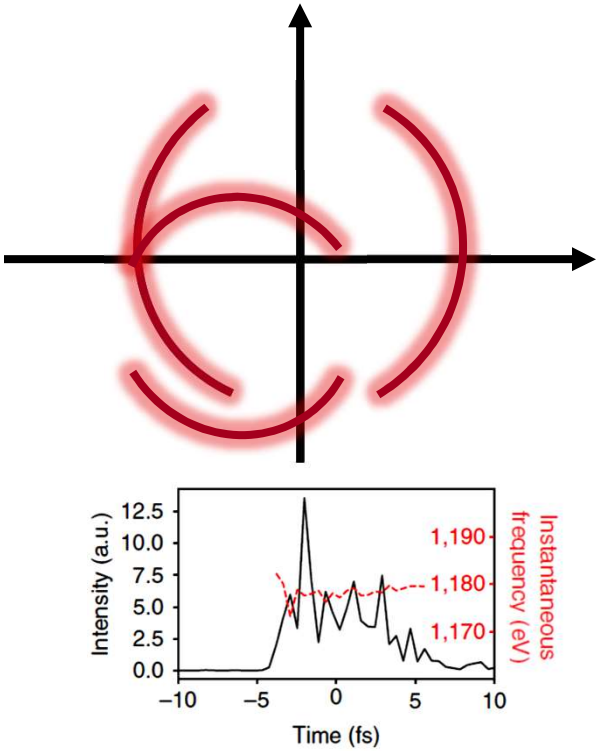


Hartmann, N. et al., Nature photonics, 2018
Siqi, Li et al. Optics express, 2018

CookieBox

The momentum of the electrons in the cloud give us information about :

- Location of the origin
- Polarization of the x-ray shot
- Number of pulses
- Energy spectrum of the x-ray shot
- Relative time spectrum of the x-ray shot
  - Using a circularly polarized laser



Goals

SLAC

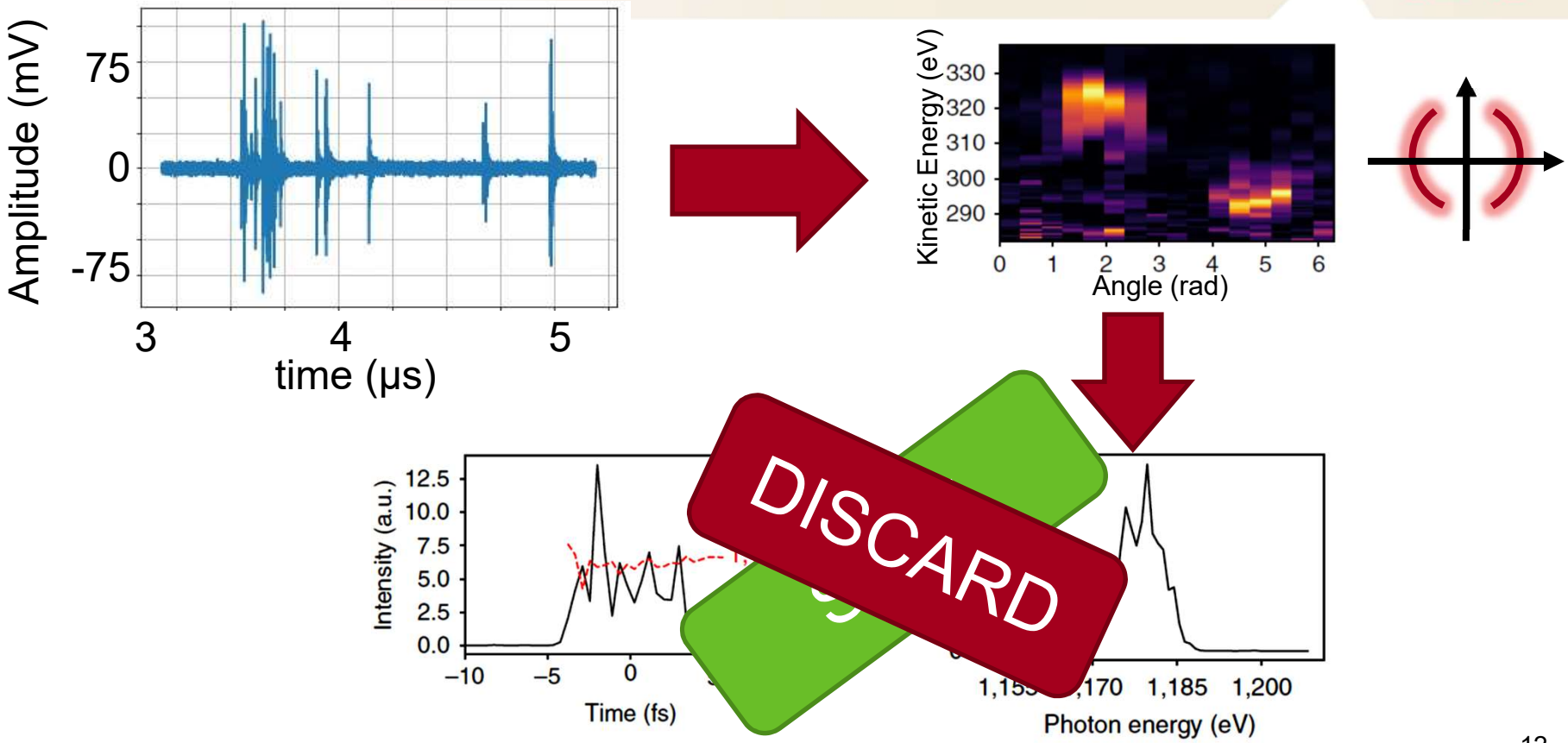
Use the CookieBox to veto LCLS-II shots

in **less than 100 µs**

at the rate of ~~**10 kHz for 2020**~~ **120 Hz for 2020**

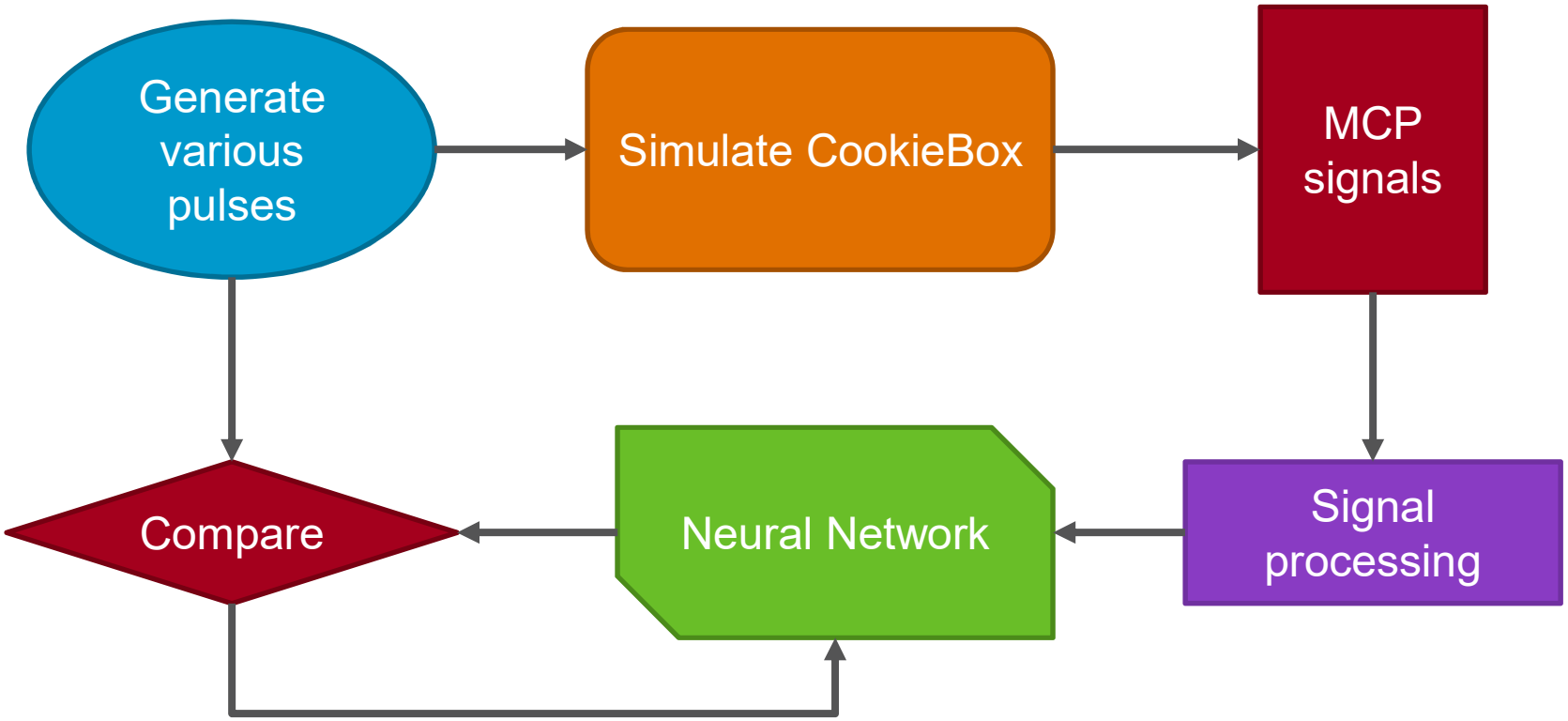
(eventually 1 MHz)

The Reconstruction Problem

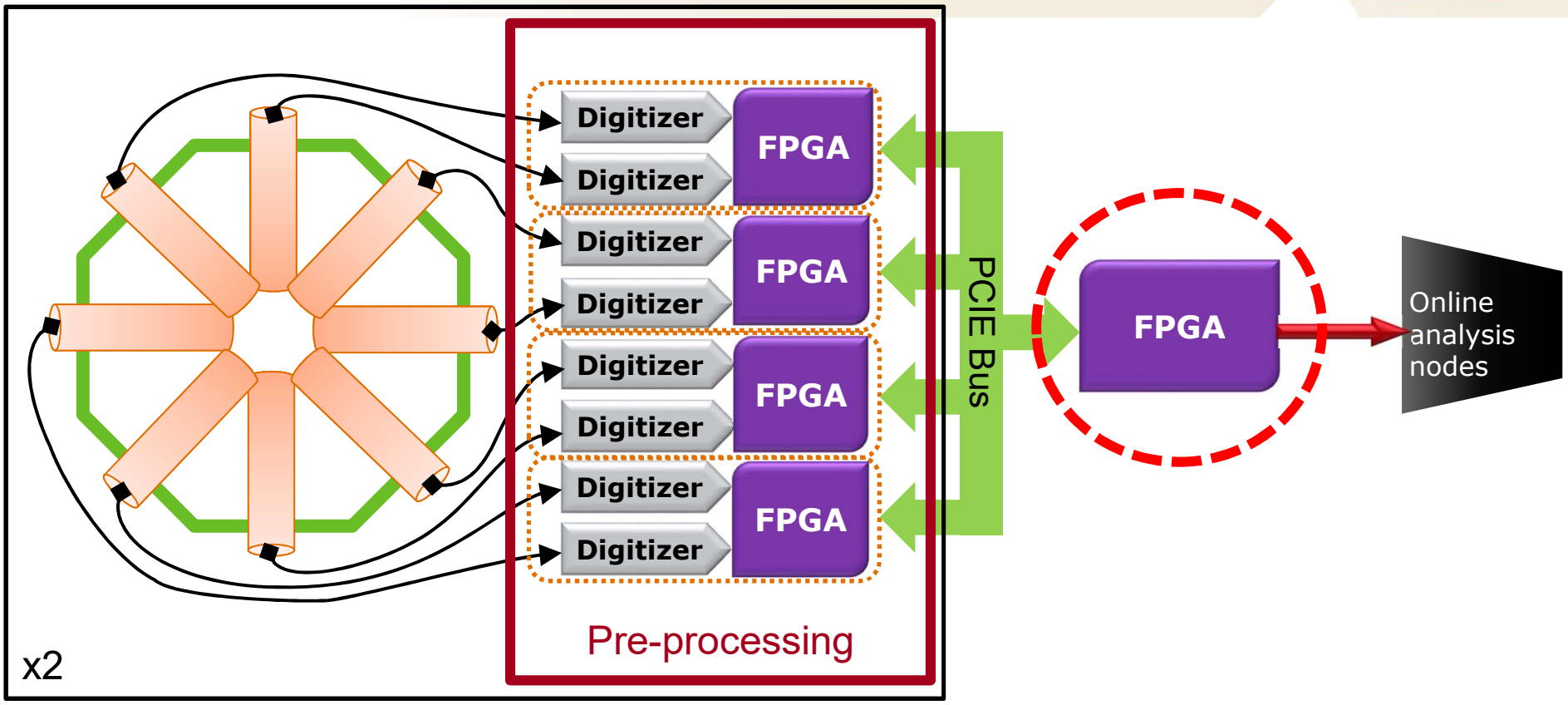


N. Hartmann et al., Nature Photonics, 2018

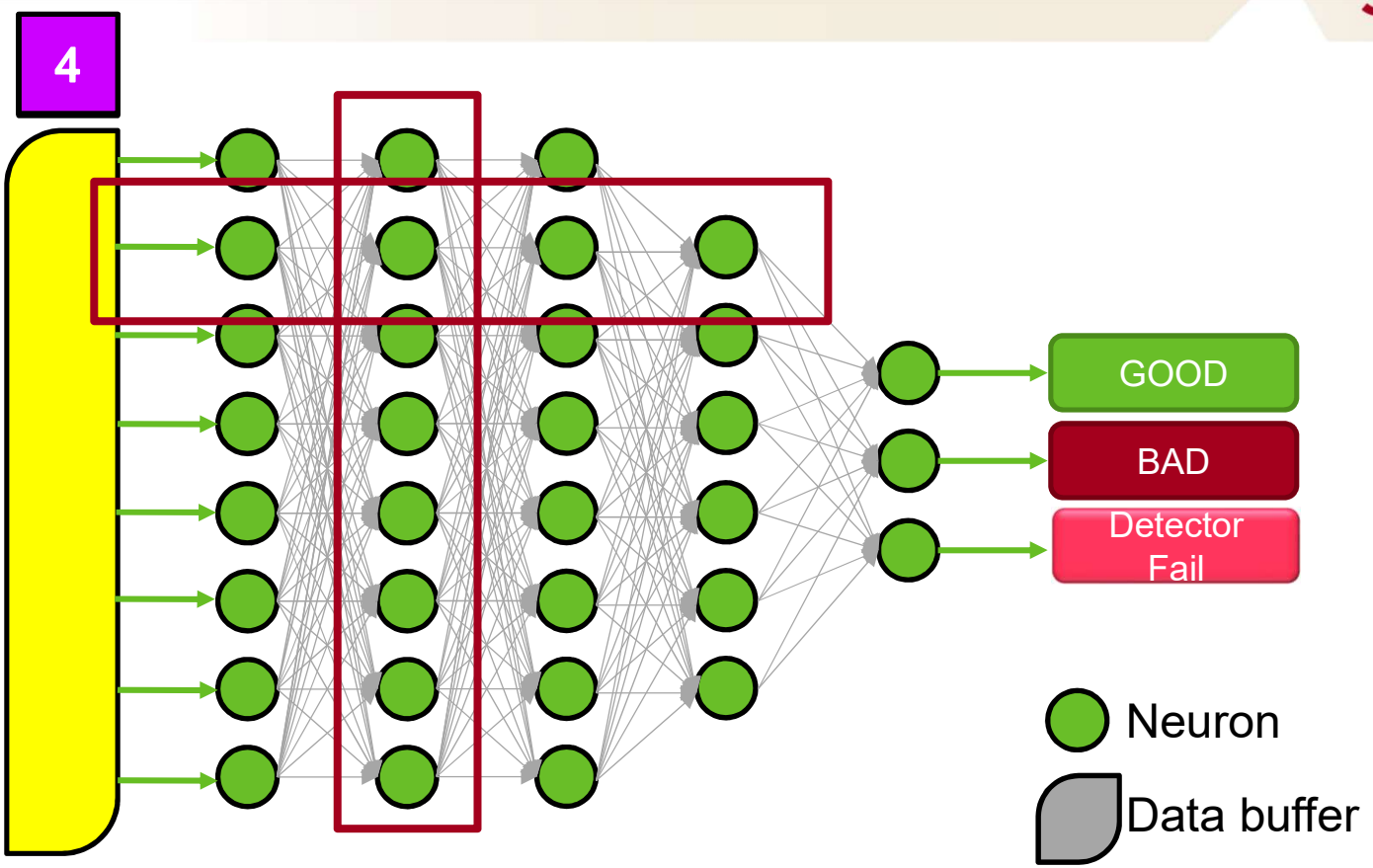
Inference Neural Network



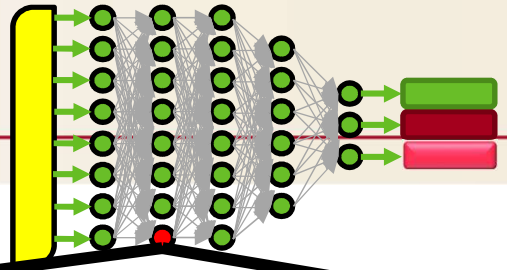
DAQ Chain Overview



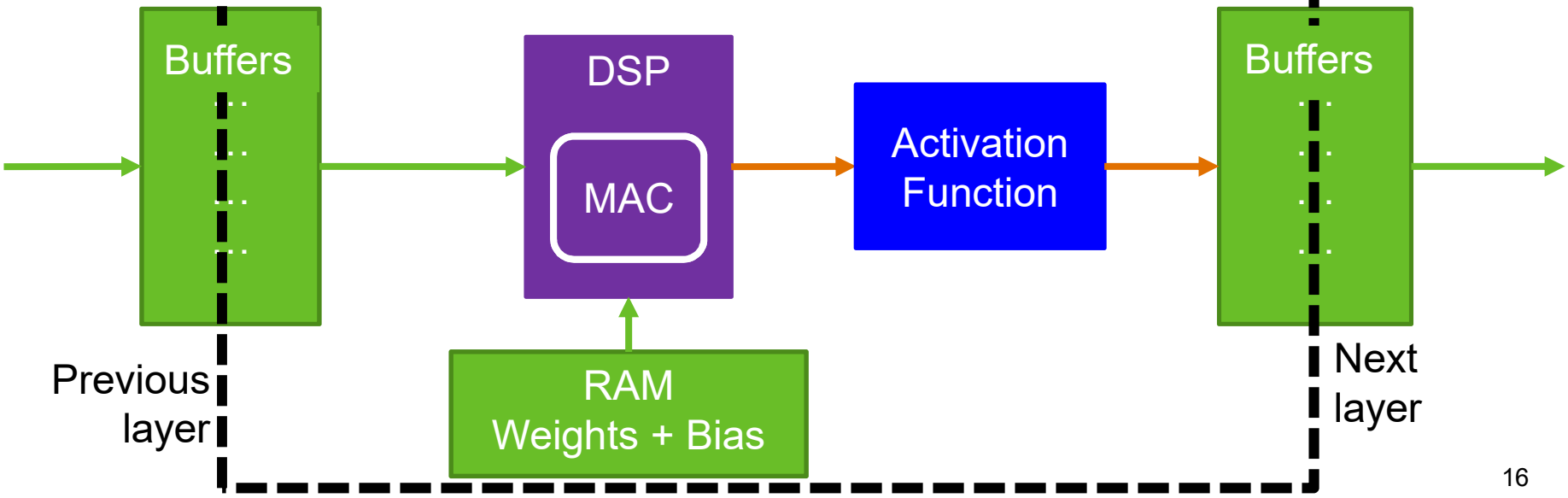
Neural Networks - Parallel and Pipelined



FPGA Implementation

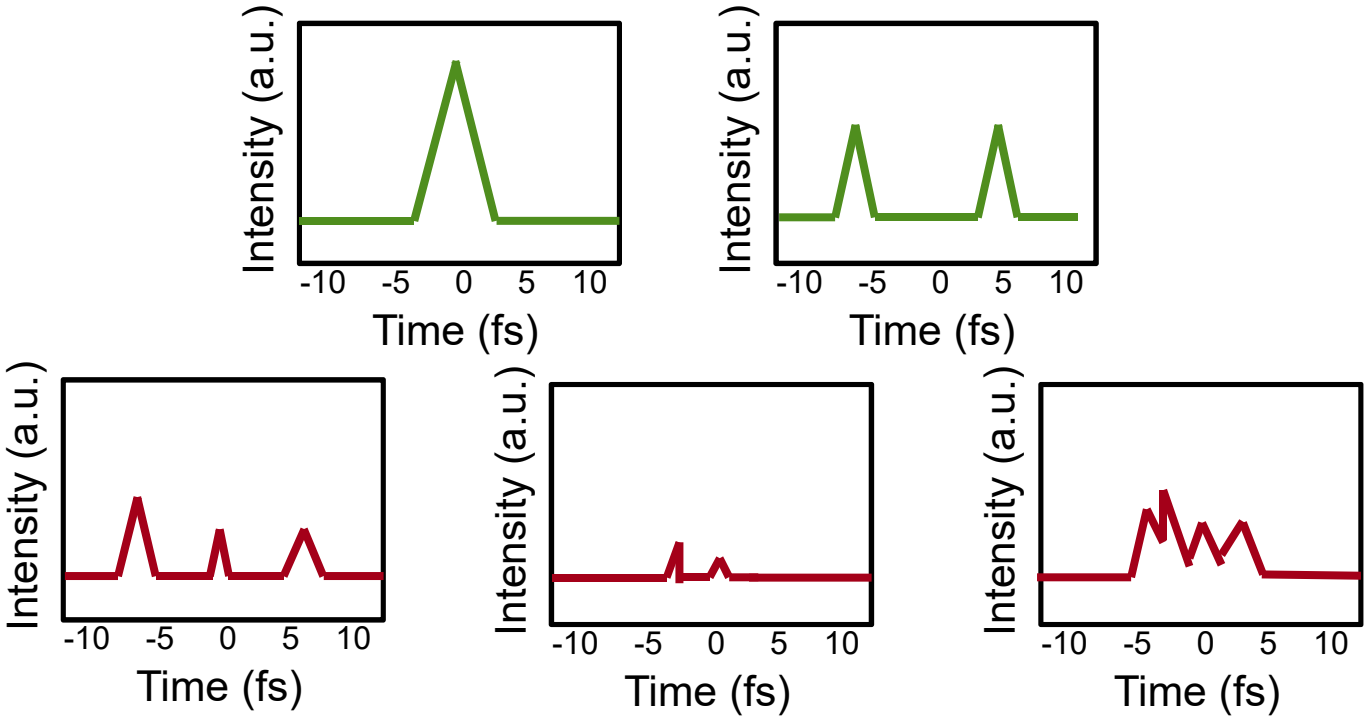


SLAC



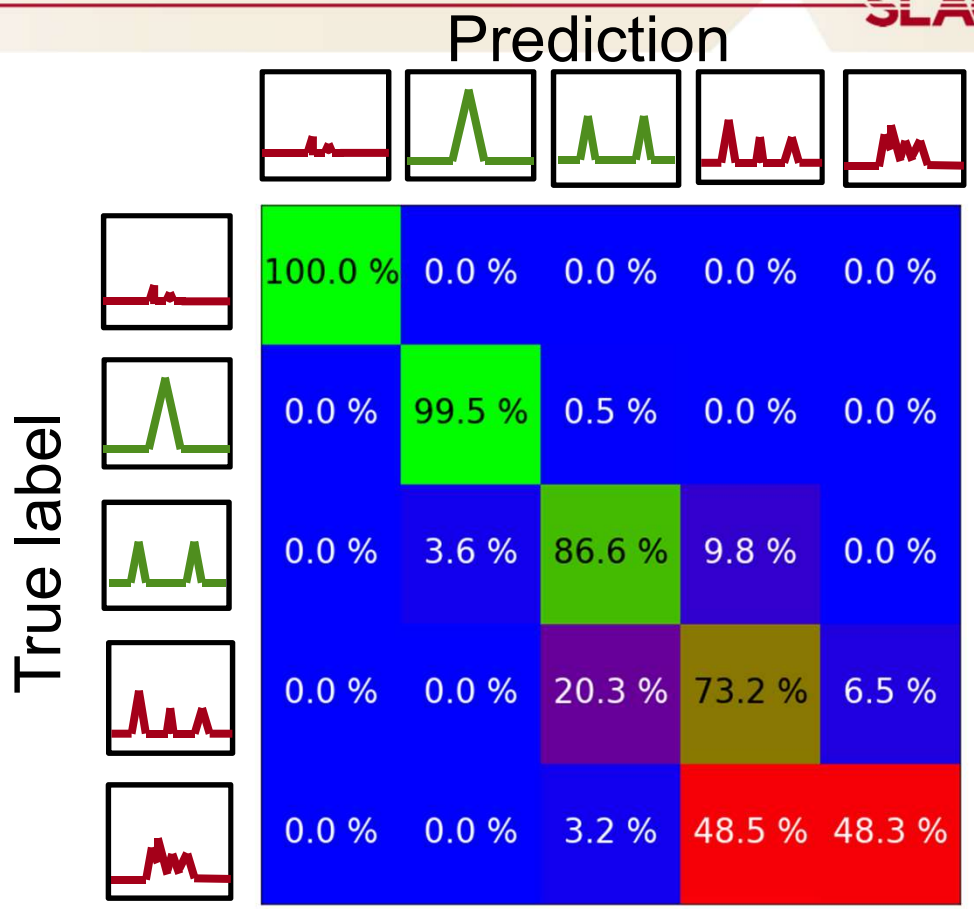
Proof of Concept

How many pulses in the shot?

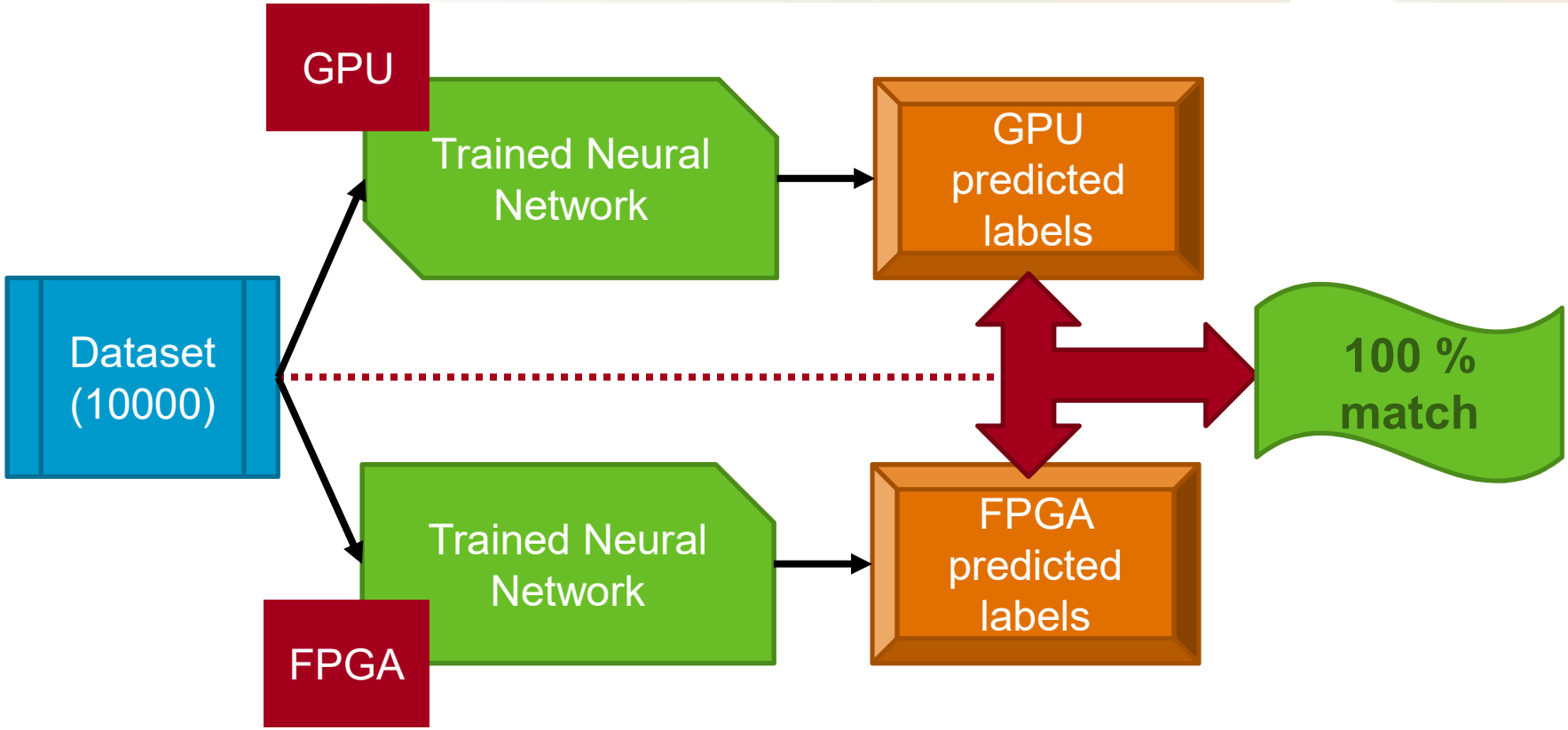


Neural Network Confusion Matrix

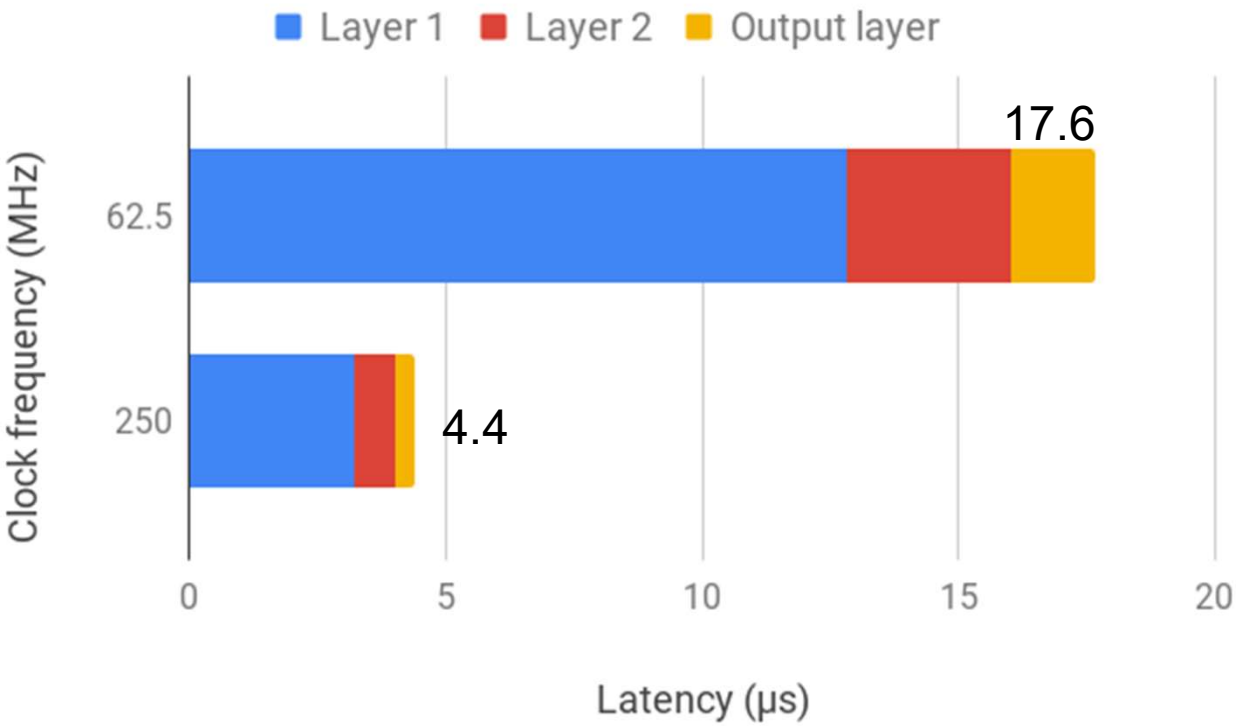
Parameter	Value
Layers	3 full
Activation	ReLu
Optimizer	RMSProp
Training set	10 000
Testing set	2000
Epochs	50
Accuracy	80.9 %



Functionality Test



Latency – Theoretical



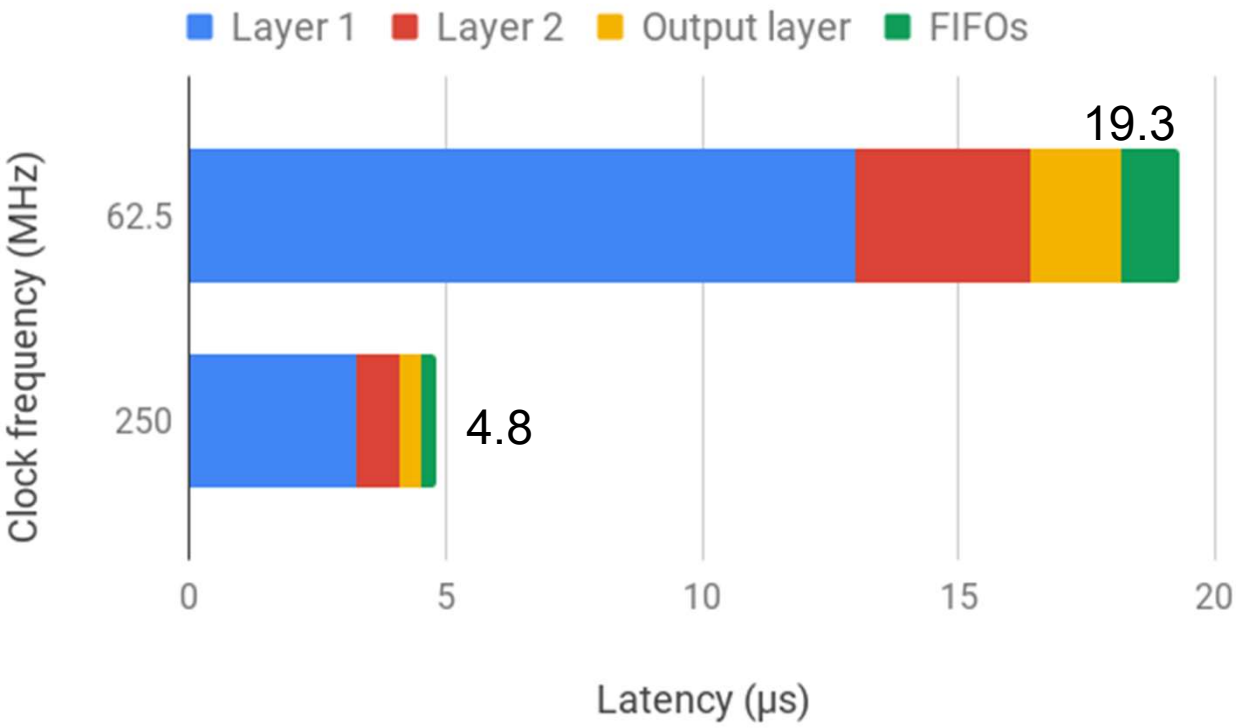
Layer 1 : 800 inputs
Layer 2 : 200 inputs
Output Layer : 100 inputs

Maximum theoretical throughput R :

$$R = \frac{1}{MAX\ (layer\ latency)}$$

R (62,5 MHz) = 78 kHz
R (250 MHz) = 312 kHz

Latency – Measured



Layer 1 : 800 inputs
Layer 2 : 200 inputs
Output Layer : 100 inputs

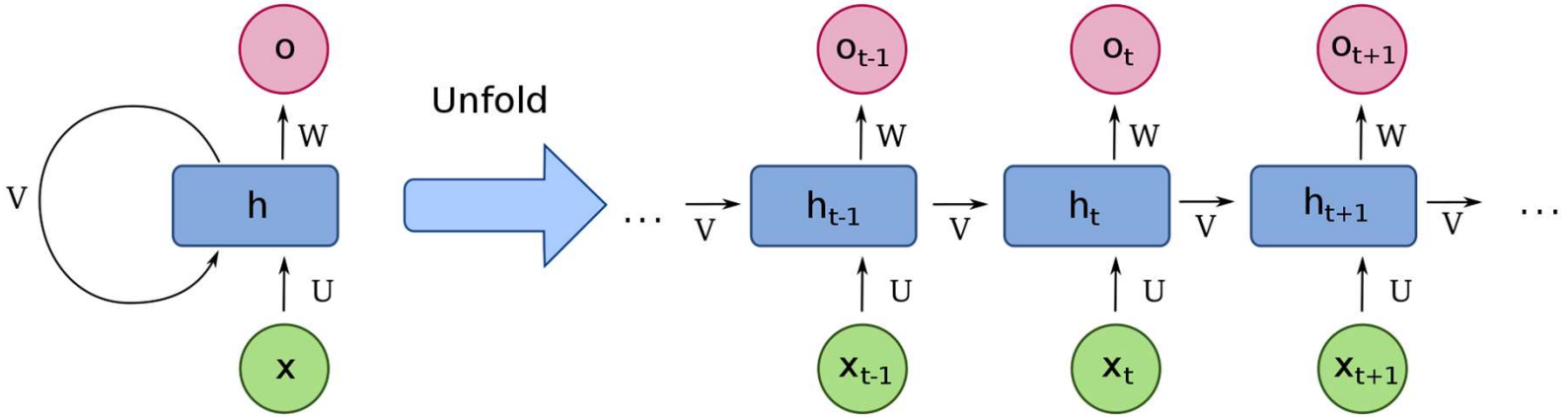
Maximum theoretical throughput R :

$$R = \frac{1}{MAX\ (layer\ latency)}$$

R (62,5 MHz) = 77 kHz
R (250 MHz) = 308 kHz

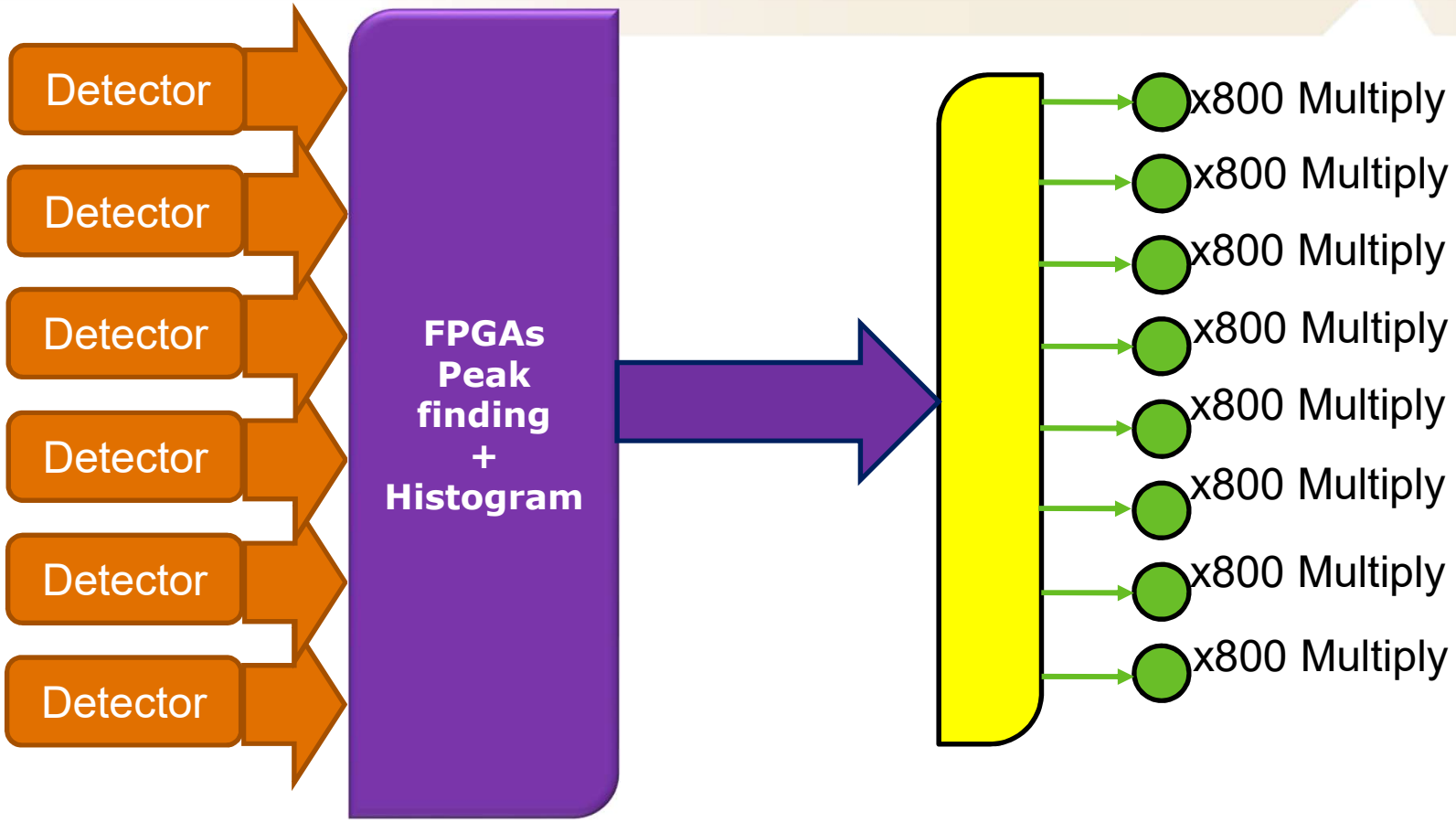
Ongoing work

- Ran new simulations
  - Removed a source of bias
- Designing a recurrent neural network

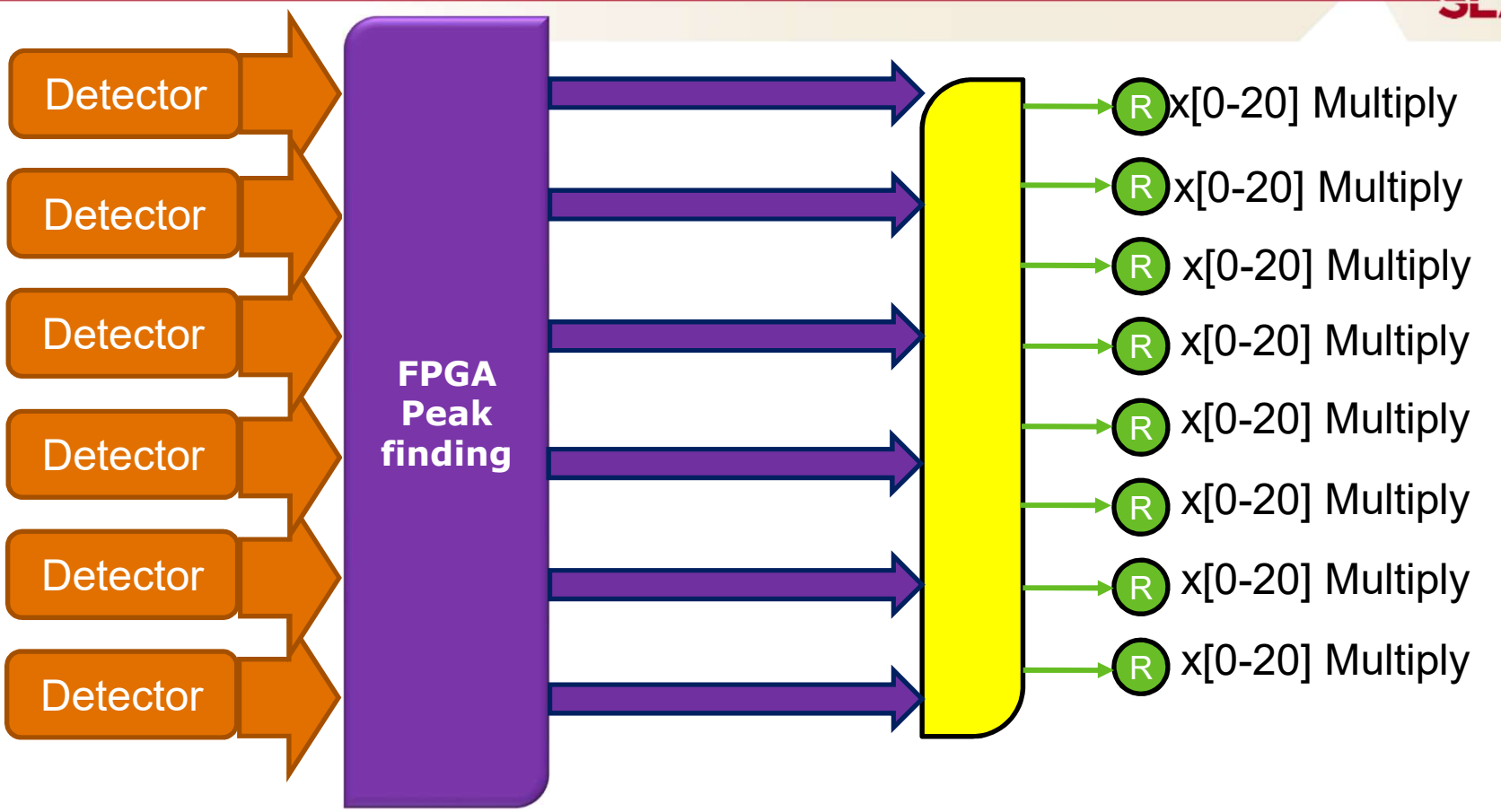


By fdeloche - Own work, CC BY-SA 4.0,
https://commons.wikimedia.org/w/index.php?curid=60109157

Recurrent Neural Net



Recurrent Neural Net

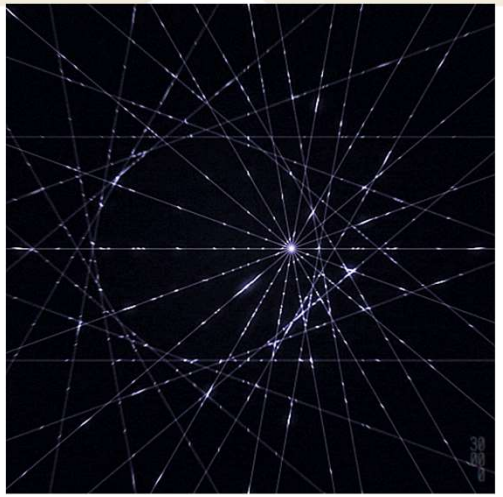


Next Steps

- Increase the network complexity and performance
  - Polarization
  - Average energy/peak energy
  - Time between two pulses in a single shot
  - Add an uncertainty metric – Anomaly detection

- Deploy for LCLS-II CookieBox in 2020 – ~~10 kHz~~120 Hz
- Create tools for our users to deploy their own models on FPGA inference engines

But my data!?!



- Prohibitively expensive to save all data
- Silly to save all data
  - Lower beam rate
  - Involves months to years of data mining
  - Storage costs
- Algorithms will be logged in the metadata
- More science opportunities
  - Anomaly detection – need a human, please!

Summary

SLAC

New photon sources and detectors will require new approaches towards data acquisition to achieve data reduction targets

Implemented a fast inference model on FPGA as proof of concept

- The current FPGA inference model achieves good performance:
  - 100 % functional
  - Latency – 19.3 μs
  - Throughput – 77 kHz
- New model being designed:
  - New simulation data
  - Recurrent model architecture

- Customized AI for every experiment
- Concept transferable to other high data rate applications

UNIVERSITÉ DE SHERBROOKE

Audrey.corbeil.therrien@usherbrooke.ca

Thank you!